AWARD NUMBER: W81XWH-20-1-0778

TITLE: Epigenetic Modifications of Cytosines in Clear Cell Kidney Carcinogenesis and Survival

PRINCIPAL INVESTIGATOR: Lucas A Salas Diaz

CONTRACTING ORGANIZATION: Dartmouth College, Hanover, NH

REPORT DATE: October 2022

TYPE OF REPORT: Annual

PREPARED FOR: US Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Limited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved

OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE	2. REPORT TYPE ANNUAL	3. DATES COVERED
October 2022	Annual	01Sep2021-31Aug2022

4. TITLE AND SUBTITLE

Epigenetic Modifications of Cytosines in Clear Cell Kidney Carcinogenesis and Survival

5a. CONTRACT NUMBER W81XWH-20-1-0778

5b. GRANT NUMBER KC190133

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Lucas A Salas Diaz

E-Mail: lucas.a.salas@dartmouth.edu

5d. PROJECT NUMBER W81XWH-20-1-0778

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

TRUSTEES OF DARTMOUTH COLLEGE 1 HINMAN DARTMOUTH COLLEGE
DARTMOUTH COLLEGE HANOVER NH
03755-4099

8. PERFORMING ORGANIZATION REPORT NUMBER
Dartmouth College

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Limited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project aims to characterize the specific genome-wide DNA methylation and hydroxymethylation patterns in clear cell renal cell carcinoma-ccRCC, evaluate whether specific changes in DNA hydroxymethylation are related to changes in cancer survival, and examine whether the observed changes in DNA hydroxymethylation and DNA methylation alter the expression of mRNA and miRNA in the ccRCC. Kidney samples from the Renal Cancer Bank at Dartmouth will be used in this project. Tandem DNA OxyBs conversion will be performed and used to measure the levels of these features through Illumina HumanMethylationEPIC microarrays. mRNA-seq and miRNA-seq information will be obtained for the samples

15. SUBJECT TERMS
None listed.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
				221	USAMRDC
a. REPORT	b. ABSTRACT	c. THIS PAGE		221	19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified	Unclassified		

TABLE OF CONTENTS

	Page
1. Introduction	**4**
2. Keywords	**4**
3. Accomplishments	**5**
4. Impact	**12**
5. Changes/Problems	**14**
6. Products	**14**
7. Participants & Other Collaborating Organizations	**16**
8. Special Reporting Requirements	**17**
9. Appendices	**17**

1. Introduction

Narrative

Around 73,820 new cases of renal cell cancer will be diagnosed in the USA in 2019; it is the 6th-leading cause of newly diagnosed cancer cases in men and the 8th in women. Clear cell renal carcinoma (ccRCC) accounts for 70–75% of these tumors. Epigenetic dysregulation has been established as one of the factors involved in tumors with global hypermethylation. The alteration of critical pathways, such as the hypoxia-inducible factor-1 (HIF-1) signaling cascade, could regulate downstream epigenetic control by modulating Ten to eleven translocation (TET) enzymes, ultimately affecting the regulation of DNA methylation/hydroxymethylation. In preliminary data, hydroxymethylation is decreased in tumor samples compared to normal kidney samples; some of these alterations were found to be concentrated on the gene bodies of the tumors. Some in vitro and in vivo models have shown that these alterations could be reverted. Here, we aim to characterize the specific genome-wide DNA methylation and hydroxymethylation patterns in ccRCC, evaluate whether specific changes in DNA hydroxymethylation are related to changes in cancer survival, and examine whether the observed changes in DNA hydroxymethylation and DNA methylation alter the expression of mRNA and miRNA in the ccRCC. 200 ccRCC and 50 normal adjacent kidney samples from the Renal Cancer Bank at Dartmouth will be used in this project. Tandem DNA OxyBs conversion will be performed and used to measure the levels of these features through Illumina HumanMethylationEPIC microarrays. mRNA-seq and miRNA-seq information will be obtained for the samples. Specific pipelines for analyzing DNA hydroxymethylation will be employed for statistical analyses. To reduce the multiple comparison burden, specific strategies will be used to deal with data sparsity and low variability. The importance of epigenetic dysregulation in cancer is well established, and epigenetic therapies are progressing rapidly in several cancer types. This project will fill several gaps in the knowledge of the epigenetic dysregulation occurring in ccRCC. Understanding the specific changes occurring at the loci or regional level could help to explain how these events occur in these tumors. Further, this work will open the door to future developments in targeted or untargeted therapies on these mechanisms. Finally, this project will explore if some of these changes are related to cancer prognosis, which could have clinical applications for patients suffering from this disease.

2. Keywords

DNA methylation

DNA hydroxymethylation

Clear cell renal cell carcinoma

Epigenetics

Survival

Cytosine

CpG

Methylation quantitative trait loci

Gene expression

Small RNA

Long noncoding RNA

3. Accomplishments

What were the major goals of the project?

The project comprises three major aims.

Aim 1 "To characterize the specific DNA methylation and DNA hydroxymethylation patterns between 200 ccRCC and 50 normal adjacent kidney samples from the Renal Tumor Tissue Bank at Dartmouth."

Aim 2 "To evaluate whether specific changes in DNA hydroxymethylation are related to changes in cancer survival."

Aim 3 "To examine whether the observed changes in DNA methylation and DNA hydroxymethylation alter the levels of mRNA and miRNA in the kidney cancer samples."

What was accomplished under these goals?

According to the approved Statement of Work for year 2 of the project, three major research-specific tasks (3 - **Pilot DNA and RNA quality control and preliminary analyses**, 4- **DNA and RNA extractions and processing**, and 5- **Integrated data analysis DNA and RNA quality**) must be completed. In the following lines, I will provide details of those task accomplishments and observed delays.

As a reminder for the reader, in year 1, major tasks 1 (**sample selection**) and 2 (**pilot DNA and RNA extractions and processing**) were achieved. The Dartmouth Cancer Center IRB approved the study protocols, and DNA and RNA processing protocols for eight pilot samples were obtained.

Major task 3 - Pilot DNA and RNA quality control and preliminary analyses

The second year of the grant aimed to delineate further all the protocols for retrieving the cytosine modifications in the tumor samples. The activities were centered on improving the protocols for nucleic acid extraction (DNA and RNA) and performing the nucleic acid extractions according to the pilot protocol results.

Subtask 1- Quality control: Our pilot observed an abnormal distribution in the DNA hydroxymethylation signals during the tandem oxidative-bisulfite conversion (OxBS) to separate the 5-methylcytosines from the 5-hydroxymethylcytosines signals (see appendix, page 87 (manuscript Figure 2)). The results and subsequent analyses were published in the manuscript entitled “Navigating the hydroxymethylome: experimental biases

and quality control tools for the tandem bisulfite and oxidative bisulfite Illumina microarrays" Zhang….Salas LA, 2022 (see Appendices pages 81-94 of the current report). The troubleshooting required additional experimentation to understand the potential causes of DNA degradation's signal loss. We compared our pilot results and added information totaling 417 pairs from 12 independent datasets (see pages 83 and 86 of the report).

Several technical gaps were analyzed and solved: 1) we optimized the oxidative-bisulfite protocol on the DNA samples, trying to reduce any unintended nucleic acid degradation due to the aggressive oxidative treatment (following the Tecan protocols), 2) technical problems arising from the tandem nucleic acid protocol extraction (DNA and RNA extraction) were identified as the critical factors affecting the results relative to the tandem oxidative-bisulfite conversion, but not when the protocol only involved sodium bisulfite conversion (Zhang et al, 2022, see page 87), 3) RNA protocols were also optimized for increasing the yield of material obtained from the samples, the protocol improvement aims to reduce any impact in the assays related to RNA degradation biases in particular for the small-RNA experiment (this according to the Qiagen protocols). See alignment report below

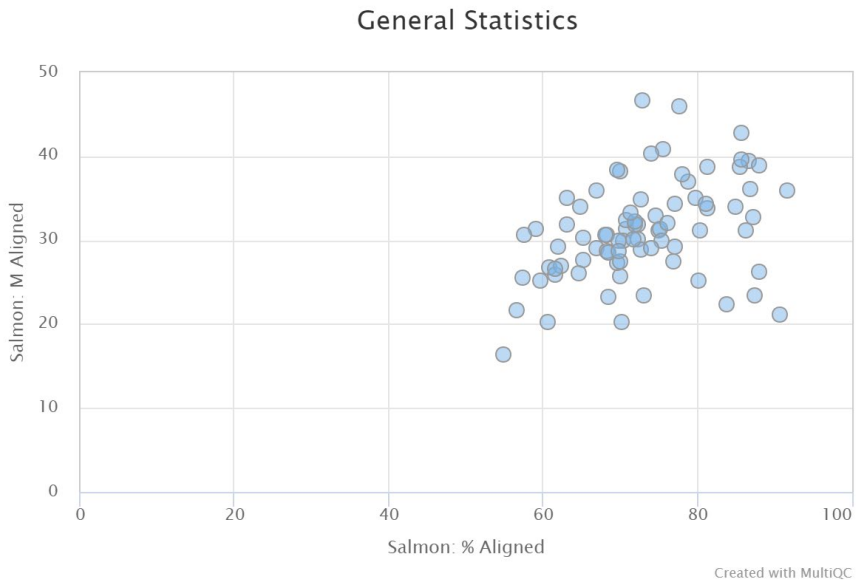


Figure 1. Percentage of transcripts aligned using Salmon (DOI: 10.1038/nmeth.4197)

Subtask 2- Data analyses, evaluate data distributions in the pilot data for selecting analytical methods

In addition to standard analytical pipelines developed for DNA methylation, RNA sequencing, and small-RNA sequencing, this subtask evaluated the potential impact of the limit of detection and the presence of potentially

unreliable measurements. A pipeline for selecting a cut-off for data reliability was established using a dynamic threshold. The software is publicly available (https://github.com/SalasLab/oxBSQC), and the results are also published in Zhang et al. 2022 (see pages 83, 84, 88, and 89 of this report). The pilot data is also publicly available on the Gene Expression Omnibus (GEO) GSE182919.

Major Task 4 Main DNA and RNA extractions and processing

Subtask 1 Sample thawing, RNA (mRNA, miRNA), DNA extractions: The samples selected for the study were located and accounted for. To the day of the report (October/2022), 70 samples had been processed (including 20 RNA duplicates due to low nucleic acid yields) reported in Figure 1 above (20% progress). The remaining 200 samples are currently being extracted with personnel dedicated exclusively to this project, all the aliquots are expected to be available on March 2023.

Subtask 2 DNA tandem conversion (hydroxymethylation, methylation): In total 50 DNA samples have been processed (20% progress). The process was optimized according to **Major task 3**, the conversions will be processed once all the DNA aliquots are available to reduce potential technical effects, the task is expected to be completed May 2023.

Subtask 3 Aliquoting of the samples and submission to the Genomic Core Facility. Once all the extractions/conversions are completed, the remaining 200 aliquots will be submitted on June 2023.

Subtask 4 DNA processing EPIC and sequencing of mRNA and miRNA: Microarrays and sequencing will be performed at the Dartmouth Cancer Center Genomics Core, the expected completion date is July 2023.

Milestones: Currently 20% of the idat files and fastq files have been generated and delivered.

Major Task 5 Integrated data analysis DNA and RNA quality

Subtask 1 Data QC. As mentioned on Major Task 3 these tasks are interconnected. We developed a quality control pipeline for the samples available. The procedures are established and will be deployed for the remainder of the samples as they become available.

Subtask 2 Data analyses, implementation of the established pipeline : We have a data analysis pipeline for DNA methylation, DNA Hydroxymethylation, RNA-sequencing, small RNA-sequencing, including multiomic integration. We have established and implemented tumor DNA methylation Epigenome Wide Association

Studies (EWAS) analytical pipelines in other cancer types (pages 55-66 of the appendix). In addition to the main pipeline we established a tumor deconvolution method to control for cell heterogeneity in carcinoma samples was optimized presented at AACR 2022 (pages 52-53 of this report) and IKCS 2022, preliminary analyses of the samples in this study are shown on Figure 2, below. We also established a protocol to control for copy number variations using DNA methylation microarray information was added to identify potential interferences arising from genetic DNA alterations when specific areas of the genome are deleted or amplified (the comparison of different protocols were published in Mariani et al. 2022, see pages 113-121, we emphasized the application to clear cell renal cell carcinomas, in page 115).We established the utility of using cell deconvolution in the 50 samples currently available, the results are showed in Figure 3, below. This major task is critical for the major task 7 (**multiomic analysis)** that is set to be completed on year 3 of this grant. Finally, for major task 6 (**survival data analysis**), we have accrued all the survival data for the subjects in this study using information from the NH cancer registry and the National Vital Statistics System (NVSS). In parallel, we had gained experience using complex survival modeling in collaboration with several DNA methylation laboratories leveraging both molecular features and cell heterogeneity of the samples (see pages 67-81 (gliomas), 147-159 (lung cancer), and 161-176 (bladder cancer) in the appendix). Additional tools for the more complex analysis were also developed and published including use of the Fetal cell of Origin in normal adult kidney (pages 123-147 of the current report), previously our group had published the application of this biomarker for Kidney cancer. Several tools for immune cell deconvolution (pages 193-207), solid tissue cell deconvolution (pages 20-40 of the appendix), integration of RNA sequencing with histologic images (pages 40-52 of the appendix), evaluation of unknown cells present in the DNA methylation mixture (RESET method pages 96-112 of the appendix), establishing myeloid-derived suppressor cells biomarkers (pages 177- 192), and a biomarker of exposure to glucocorticoids (pages 208-223). All these tools have potential applications to the current project analysis.

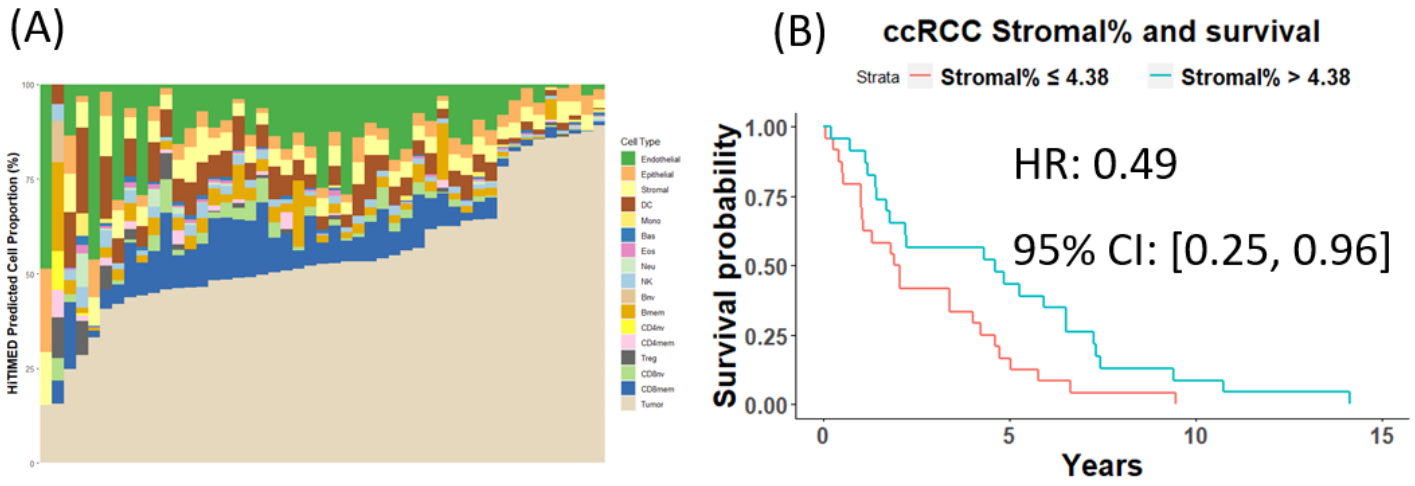


Figure 2. Understanding the tumor cell microenvironment of the ccRCC samples. Panel A shows a barplot with each column representing a sample and the colored bars representing the percentage of specific cells in the sample, Panel B, shows the survival trajectories for those samples separated by median stromal content, favoring those with higher stromal content with 51% higher survival in this subset of the data.

<u>Subtask 3 Repeat experiments as needed if failures during QC, establish potential issues and troubleshoot problems</u>. As observed in figure 1 above, a subset of the samples showed a lower alignment relative to lower RNA yields even with adequate RNA quality. 20 samples were reextracted and repeated increasing the yields and the alignment eficiency.

Milestones achieved: pipeline established, experimental bioinformatic pipeline has been established for the analyses. 50 samples have been succesfully been analyzed using this pipeline.

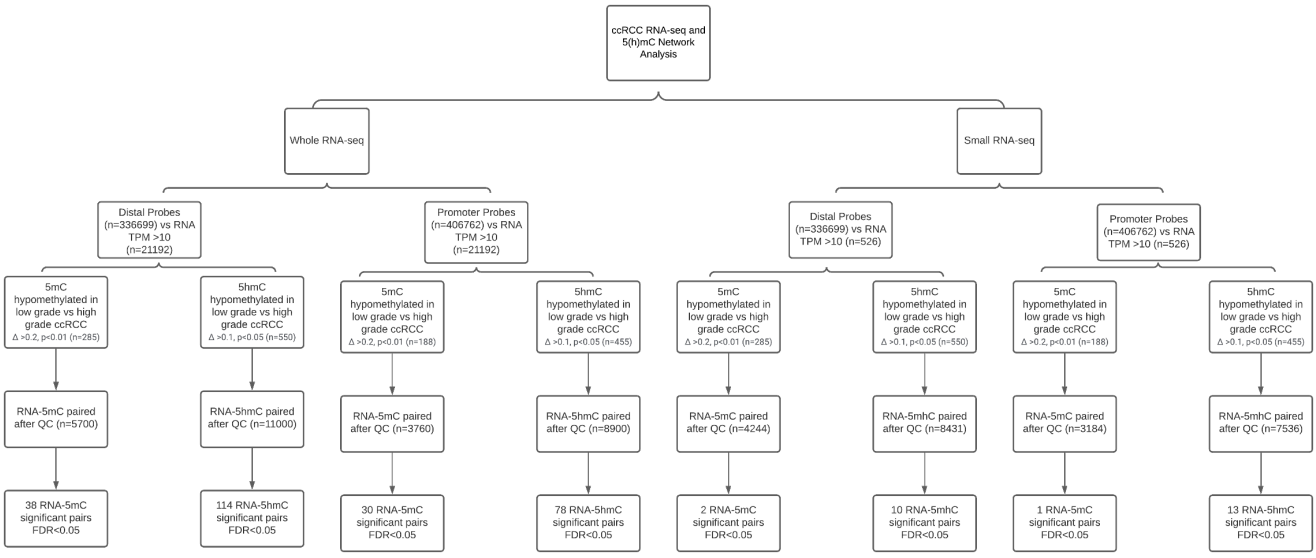


Figure 3. Preliminary data analysis using DNA methylation, DNA hydroxymethylation, RNA-sequencing and small RNA-sequencing (presented on November 2021 at the IKCS North America Meeting (https://www.kcameetings.org/wp-content/uploads/2021/12/IKCSNA21_N18_Zhang.pdf).

Accomplished research-specific Tasks

Major task 3. A QC and analytical method was generated. In total, 70 samples (including repeats) were assayed, and 20 samples were analyzed in the pilot study using tandem nucleic acid extraction. Due to the low yield in the results, they were repeated, and 50 samples have been completed (20% of the total samples). The findings related to the problems and opportunities observed in the pilot study were published in the manuscript by Zhang et al. included in the appendix (pages 81-94) and entitled "Navigating the hydroxymethylome: experimental biases and quality control tools for the tandem bisulfite and oxidative bisulfite Illumina microarrays." Briefly, we optimized the nucleic acid extraction protocol and found that tandem DNA/RNA interfered with the tandem oxidative/bisulfite (OxBS) conversion required to discern between methylation and hydroxymethylation cytosine modifications. A bioinformatic quality control protocol for OxBS was developed (OxBSQC available here: https://github.com/SalasLab/oxBSQC) to classify samples and check potential mislabeling using assay-specific internal control probes. A dynamic cut-off was used to classify lower bound false positives for 5-hydroxymethylation signals. Simultaneously, we worked on protocols to optimize RNA extraction and reduce false signals due to low yields in RNA extraction.

Using public and in-house data, bioinformatic protocols were developed for the preprocessing and analysis of the: 1) tandem OxBS converted DNA samples using the IlluminaHumanMethylation EPIC microarray, 2) mRNA and small RNA data, 3) Integration of cytosine modifications and RNA sequencing data. These results were presented on November 2021 at the IKCS North America Meeting (https://www.kcameetings.org/wp-content/uploads/2021/12/IKCSNA21_N18_Zhang.pdf). An updated interim analysis was presented at IKCS 2022.

Major task 4 Currently, according to our initial timeline, we would have extracted 100% of the nucleic acids (DNA and RNA) for 124 additional tumor samples and 46 normal-adjacent kidney samples. They would have been undergoing quality control before OxBS conversion and sequencing. Unfortunately, due to the labor shortage in the country, this critical step has been delayed. Although this may represent a delay in the chronogram of the project, this has offered a potential opportunity for the project scope. The microarray

manufacturer is ready to launch a new generation of microarrays adding new targets, and we are discussing with them the possibility of testing the new extended technology in these samples using the newer Illumina EPIC v2 or the EPIC+ (custom HLA probes). This delay offers an attractive scientific opportunity, as the additional content will allow exploring certain genomic areas that are immune-specific targets (in which we are actively collaborating) and may be critical to discover additional information for clear cell renal cell carcinomas.

Major Task 5 Integrated data analysis DNA and RNA quality

We have tested the current pipeline with te 50 samples available, a pipeline integrating QC and Another methodological paper to address the potential interference of copy number variations evaluated several available pipelines, and it was published in Frontiers in Bioinformatics. The bioinformatics software is called MethylMasteR (https://doi.org/10.5281/zenodo.6360630), and its manuscript is entitled "MethylMasteR: A Comparison and Customization of Methylation-Based Copy Number Variation Calling Software in Cancers Harboring Large Scale Chromosomal Deletions." The manuscript is available in the appendices (Mariani et al. 2022, pages 113-121).

A manuscript assessing cell heterogeneity in carcinoma samples is currently under review. The preliminary results of this study (that are needed for the final analysis of the samples in this project were presented at the American Association of Cancer Research (AACR) 2022 in New Orleans, and the abstract entitled "Abstract 1212: ExTIME: Extended tumor immune micro-environment cell mixture deconvolution using DNA methylation and a novel tumor-site-specific hierarchical approach" is published on Cancer Research (http://dx.doi.org/10.1158/1538-7445.AM2022-1212), and included in the appendices. Briefly, seventeen cell types are deconvolved using bioinformatic methods separating tumor cells, normal adjacent epithelial cells, endothelium, stroma, and immune cells. We used TCGA data to observe differential survival related to the endothelial tumor content in clear cell renal cell carcinomas. We applied the algorithm with good results as observed in the previous section, and we expect to apply the algorithm to this project's complete set of samples to expand on these observations and evaluate different molecular pathways involved in differential survival.

What opportunities for training and professional development has the project provided?

Accomplished training specific Tasks
As part of the Academy of Kidney Cancer Investigators (AKCI), I have actively participated in the monthly webinars and our annual meeting organized by Dr. Brian Rini. The calendar and talks are summarized below:

- 11/4/2021: 2nd Annual AKCI workshop, Austin, TX.
- 11/11/2021 Didactic Lecture Nima Sharifi “Running a successful lab / obtaining peer-review funding.”
- 12/15/2021 Didactic Lecture Dr. Rakesh Jain “Tumor Microenvironment.”
- 1/19/2022 Didactic Lecture Dena Battle and Sally McAdoo.
- 2/4/2022 Quarterly meeting AKCI
- 2/9/2022 Didactic Lecture Samra Turajlic.
- 3/24/2022 Didactic Lecture Jeff Rathmell.
- 4/28/2022 Didactic Lecture Brian Such.
- 5/6/2022 Quarterly meeting AKCI.
- 5/23/2022 Didactic Lecture Ralph DeBerardinis
- 6/24/2022 Didactic Lecture Kimryn Rathmell
- 7/20/2022 Didactic Lecture Tim Chan
- 9/14/2022 Didactic Lecture Scott Welford.

In addition to the AKCI regular meetings, I also attended the American Association of Cancer Research general meeting on April 2022, where I participated in a lecture by my mentor Dr. James Brugarolas. Additional training opportunities arise from learning from several groups working with DNA methylation and single-cell technologies during that meeting.

I have met regularly with my mentor and co-mentor to set the project and brainstorm potential protocol issues with my pilot. They supported the practical problems and engaged in mentoring sessions to solve other practical lab problems.

Additional refreshers in topics related to machine learning and deep learning have been taken at my institution.

How were the results disseminated to communities of interest?

Nothing to report

What do you plan to do during the next reporting period to accomplish the goals?

For the next report, I aim to have completed the assays in the project's remining samples, including quality control and preliminary analyses. Focusing on completing the major tasks 3 to 7 of this award. We will have the full pipeline for major task 6 (**survival analyses**) and for major task 7 (**multiomics analyses**) and evaluate how the proposed modeling approaches are applied to resolve the hypotheses in this project. I will continue my training in kidney cancer and expand my working network to allow further expansion of my laboratory.

4. Impact
What was the impact on the development of the principal discipline(s) of the project?

We have now reported several valuable technical manuscripts for future researchers and additional expansion of the current scope of research. The knowledge about the potential problems during nucleic acid extraction

and how that impact the OxBS conversion to discern between 5-hydroxymethylcytosines and 5-methylcytosines is critical for the success of this project. As mentioned above, we also developed a quality control bioinformatic OxyBSQC pipeline to evaluate potential mislabeling and minimize false positives in the lower bound of the distribution (see Zhang et al. 2022). Using DNA methylation microarrays, we have also critically evaluated bioinformatic pipelines for Copy Number Variation inference. Establishing a pipeline for these analyses is critical for tumor samples to reduce genomic variation interference when considering cytosine modifications customizable to specific conditions in the samples and fully Dockerized to reduce version-to-version changes (MethylMasteR, see Mariani et al, 2022). These products will impact future research as they allow it reuse in additional future research.

Among the products included, I have detailed several manuscripts related to this project's goals. Salas et al. 2022 Nat Comm (pages 193-207) developed a new method for blood cell deconvolution; see Figure 1 of the manuscript (page 195 of the appendix) for details of the method pipeline. This project is part of my lab work on cell heterogeneity, and it will help in some of the methodological aspects for the analysis of DNA methylation in the tumor samples of this project. The new cell deconvolution algorithms (HiTIMED included as an abstract) is poised to critically impact DNA methylation research in the future.

What was the impact on other disciplines?

Our methodological paper establishing limits of detection and quality control for DNA hydroxymethylation has been cited for manuscripts working on neurosciences (Parkinson Disease) and microbiome research. Peripheral blood cell deconvolution methods have been used for different areas of research including biological anthropology, neurosciences, bioinformatics, pediatrics with 31 citations in less than one year of publication.

What was the impact on technology transfer?

We have developed a protocol for studying cell heterogeneity in carcinomas using DNA methylation. A provisional patent application has been filed for the tumor deconvolution. Two distinct software libraries were developed (OxyBSQC and MethylMasteR) which are on the public domain for researchers reuse.

What was the impact on society beyond science and technology?

Nothing to report

5. Changes/Problems

The CoVID-19 pandemic negatively impacted the development of the project. In the rural community of Dartmouth, we are facing a significant labor shortage. This year has allowed more in-person interactions to troubleshoot nucleic acid extraction and cytosine OxyBS tandem conversion. The pool of graduate students is lower than in previous years, and the postdoctoral fellow hired in 2021 decided to move to a different job. A new postdoctoral fellow started on August 2022 and will be critical to augment the laboratory capacities to manage the data influx in the following months. The following two months are critical to deciding between using the current microarray or updated technology. For the following report, I aim to have completed the quality control and preliminary analyses for the remaining samples in the project. If the labor shortage continues to impede the project's advancement, I have contacted several providers that can offer externalized services for the OxBS conversion. This problem was solved with a hiring on November 2022, and currently this person is dedicated 100% of the time to this project. As an alternative the externalization of nucleic acid extractions and processing was discarded due to the extremely negative impact the total budget of the project, but it is still a solution if the labor shortage affects the project again. The proposed chronogram is delayed six months versus the original SOW. The expectation is that once the assays (microarrays and sequencing) are completed, the effort used in the pipeline optimization during the project's first two years will allow us to catch up with the initial schedule quickly.

6. Products

The products listed below correspond to methodological developments related to the current grant, and collaborations with researchers in other areas. Some methods have been associated with patent applications (Salas et al. 2022- reported in the previous progress report, and Zhang et al. 2022). In kidney cancer specific research (highlighted with a *), we have published (Zhang et al., Mariani et al and Zhang et al.). Additional scholarly work are related to my area of expertise in DNA methylation and cancer, and immune responses and cancer. All the products mentioned below are directly related to my overall goal of becoming a leader in cancer epigenomics with emphasis in kidney cancer research.

Published peer-reviewed journal publications:

1. Wiencke JK, Molinaro AM, Warrier G, Rice T, Clarke JL, Taylor JW, Wrensch M, Hansen H, McCoy L, Tamaki S, Bracci PM, **Salas LA**, Koestler DC, Christensen BC, Zhang Z, Kelsey KT. DNA methylation as pharmacodynamic markers of glucocorticoid response and glioma survival. Nat Commun. 2022 Sep 20;13(1):5505. PubMed PMID: 36127421.

2. Muse ME, Bergman DT, **Salas LA**, Tom LN, Tan JM, Laino A, Lambie D, Sturm RA, Schaider H, Soyer HP, Christensen BC, Stark MS. Genome-scale DNA methylation analysis identifies repeat element alterations that modulate the genomic stability of melanocytic nevi. J Invest Dermatol. 2022 Jul;142(7):1893-1902.e7. PubMed PMID: 34871578.
3. *Mariani MP, Chen JA, Zhang Z, Pike SC, **Salas LA**. MethylMasteR: A Comparison and Customization of Methylation-Based Copy Number Variation Calling Software in Cancers Harboring Large Scale Chromosomal Deletions. Front Bioinform. 2022;2. Epub 2022 Apr 12. PubMed PMID: 35573871.
4. Bell-Glenn S, **Salas LA**, Thompson JA, Koestler DC. A Novel Framework for the Identification of Reference DNA Methylation Libraries for Reference-Based Deconvolution of Cellular Mixtures. Front Bioinformatics. 2022. 2. Epub 2022 Mar 21. PubMed PMID: 35419567
5. Molinaro AM, Wiencke JK, Warrier G, Lee J, Koestler DC, Chunduru P, Hansen H, Rice T, Bracci PM, McCoy L, **Salas LA**, Christensen BC, Wrensch M, Kelsey KT, Taylor JW, Clarke JL. Interactions of Age and Blood Immune Factors and Non-Invasive Prediction of Glioma Survival. JNCI, 2022. Mar 8;114(3):446-457.PMID: 34597382.
6. Wiencke JK, Zhang Z, Koestler DC, **Salas LA**, Molinaro AM, Christensen BC, Kelsey KT. Identification of a foetal epigenetic compartment in adult human kidney.Epigenetics. 2022 Mar 30:1-21. PMID: 33783321.
7. Berglund-Brown I, Nissen E, Koestler DC, Butler RA, Eliot MN, Padbury JF, **Salas LA**, Molinaro AM, Christensen BC, Wiencke JK, Kelsey KT. A core of differentially methylated CpG loci in gMDSCs isolated from neonatal and adult sources. Clin Epigenetics. 2022 Feb 21;14(1):27. PubMed PMID: 35189960.
8. **Salas LA,** Zhang Z, Koestler DC, Butler RA, Hansen HM, Molinaro AM, Wiencke JK, Kelsey KT, Christensen BC. Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. Nat Comm Nat Comm. 2022 13(1):761. PMID*:* 35140201
9. *Zhang Z, Lee MK, Perreard L, Kelsey KT, Christensen BC, **Salas LA**. Navigating the hydroxymethylome: experimental biases and quality control tools for the tandem bisulfite and oxidative bisulfite Illumina microarrays. Epigenomics. 2022. Feb;14(3):139-152. PMID: 35029129.
10. Chen JQ, **Salas LA,** Wiencke JK, Koestler DC, Molinaro AM, Andrew AS, Seigne JD, Karagas MK, Kelsey KT, Christensen BC. Immune profiles and DNA methylation alterations related with non-muscle-invasive bladder cancer outcomes. Clin Epigenetics. 2022 Jan 21;14(1):14. PMID: 35063012.
11. Levy JJ, Bobak CA, Nasir-Moin M, Veziroglu EM, Palisoul SM, Barney RE, **Salas LA**, Christensen BC, Tsongalis GJ, Vaickus LJ. Mixed Effects Machine Learning Models for Colon Cancer Metastasis Prediction using Spatially Localized Immuno-Oncology Markers. Pac Symp Biocomput. 2022;27:175-186. PMID: 34890147.
12. **Salas LA,** Stewart TG, Mobley B, Peng C, Loganathan S, Wang J, Berger MS, Absher D, Moots PL, Christensen BC, Clark SW. Phase I Study of High-Dose l-Methylfolate in Combination with Temozolomide and Bevacizumab in Recurrent IDH Wild-Type High-Grade Glioma. Canc Res Comm. (2022) 2 (1): 1–9. PMID: 35392283.
13. Zhao N, Ruan M, Koestler DC, Lu J, **Salas LA**, Kelsey KT, Platz EA, Michaud DS. Methylation-derived inflammatory measures and lung cancer risk and survival. Clin Epigenetics. 2021 Dec 16;13(1):222. doi: 10.1186/s13148-021-01214-2. PubMed PMID: 34915912.

Conference publications:

1. *Zhang Z, Wiencke JK, Kelsey KT ,Koestler DC**,** Christensen BC, **Salas LA**. HiTIMED: Hierarchical Tumor Immune Microenvironment Epigenetic Deconvolution for accurate cell type resolution in the tumor microenvironment using tumor-type-specific DNA methylation data. AACR 2022. Doi: 10.1158/1538-7445.AM2022-1212#
2. A Molinaro, J Wiencke, G Warrier, JY Lee, D Koestler, H Hansen, T Rice, Bracci P, McCoy L, **Salas LA**, Christensen BC, Wrench M, Kelsey KT, Taylor J, Clarke J. EPCO-25. An Immunomethylomic

Platform Integrating Systemic Immune Profiles And Epigenetic Age In Neuro-Oncology. Neuro-Oncology 22 (Supplement_2), ii74-ii74. Doi: 10.1093/neuonc/noaa215.304.#

3. J Wiencke, A Molinaro, G Warrier, T Rice, J Clarke, J Taylor, M Wrensch, Tamaki S, Lee J, Hansen H, McCoy L, Bracci P, **Salas LA**, Koestler DC, Christensen BC, Kelsey KT. BIOM-13. DNA methylation marks glucocorticoid pathway response in dexamethasone-treated brain tumor patients. Neuro-Oncology 22 (Supplement_2), ii4-ii4. Doi: 10.1093/neuonc/noaa215.013.#

Papers currently under review:

4. Zhang Z, Wiencke JK, Kelsey KT ,Koestler DC, Christensen BC, Salas LA. HiTIMED: Hierarchical Tumor Immune Microenvironment Epigenetic Deconvolution for accurate cell type resolution in the tumor microenvironment using tumor-type-specific DNA methylation data.
5. Moroishi Y, Salas LA, Zhou J, Baker ER, Hoen AG, Everson T, Marsit C, Madan J, Gui J, Karagas MR. Umbilical Cord Blood Immune Cell Profiles in Relation to the Infant Gut Microbiome.
6. Zhang Z, Butler R, Koestler DC, Bell-Green S, Warrier G, Molinaro AM, Christensen BC, Wiencke JK, Kelsey KT , Salas LA. Comparative analysis of the DNA methylation landscape in CD4, CD8, and B memory lineages.

Presentations:

Date	Topic/Title	Organization	Location
8-12-2022	Phase I study of #methylfolate in recurrent high-grade glioma	AACR August Journal Club: Cancer Research Communications	Virtual
9-13-2022	Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling.	**PACE Consortium**	Virtual
9-7-2022	Epigenetic, transcriptional, and compositional shifts in clear cell renal cell carcinomas a molecular epidemiologist perspective”	Centers of Biomedical Research Research in Progress	Lebanon, NH

Patent applications:

US Provisional Patent: “HiTIMED Hierarchical Tumor Immune Microenvironment Epigenetic Deconvolution for Highly Accurate Cell Type Resolution in the Tumor Microenvironment using Tumor-Type-Specific Hierarchical Deconvolution and DNA Methylation Data“ Inventors/Affiliation: Christensen/Dartmouth; Salas/Dartmouth; Zhang/Dartmouth; Koestler/Univ of Kansas Med Center; Kelsey/Brown; Wiencke/UCSF

7. Participants & Other Collaborating Organizations

PI: Lucas A Salas Diaz

Research group:

Laurent Perreard, PhD

Ze Zhang

Prasoona Karra, PhD

Jason Pettus, MD PhD

Michael Mariani, PhD (former postdoctoral fellow).

8. Special Reporting Requirements

Not applicable

9. Appendices

Genome-Scale DNA Methylation Analysis Identifies Repeat Element Alterations that Modulate the Genomic Stability of Melanocytic Nevi

Meghan E. Muse[1], Drew T. Bergman[1], Lucas A. Salas[1], Lisa N. Tom[2], Jean-Marie Tan[2], Antonia Laino[2], Duncan Lambie[3,4], Richard A. Sturm[2], Helmut Schaider[2,5], H. Peter Soyer[2,6], Brock C. Christensen[1,7,8,9] and Mitchell S. Stark[2,9]

Acquired melanocytic nevi grow and persist in a stable form into adulthood. Using genome-wide methylation profiling, we evaluated 32 histopathologically and dermoscopically characterized nevi to identify the key epigenetic regulatory mechanisms involved in nevogenesis. Benign (69% globular and 31% nonspecific dermoscopic pattern) and dysplastic (95% reticular/nonspecific dermoscopic pattern) nevi were dissimilar, with only two shared differentially methylated loci. Benign nevi showed an increase in both genome-scale methylation and methylation of Alu/LINE-1 retrotransposable elements, a marker of genomic stability, as well as global methylation. In contrast, dysplastic nevi showed evidence for genomic instability through the hypomethylation of Alu/LINE-1 (Alu: P ¼ 0.00019; LINE-1: P ¼ 0.000035). Using dermoscopic classifications, reticular/nonspecific patterned nevi had 59,572 5'-C-phosphate-G-3' differentially methylated loci (Q < 0.05), whereas globular nevi had no significant differentially methylated loci. In reticular/nonspecific patterned nevi, the tumor suppressor PTEN had the greatest proportion of hypermethylated 5'-C-phosphate-G-3' loci in its promoter region than all other assayed gene promoters. The relative activity of reticular/nonspecific nevi was evidenced by 50,720 hypomethylated loci being enriched for accessible chromatin and 8,852 hypermethylated loci strongly enriched, for example, marks of active gene promoters, which suggests that gain of DNA methylation observed in these nevus types plays a role in gene regulation.

Journal of Investigative Dermatology (2022) 142, 1893e1902; doi:10.1016/j.jid.2021.11.025

INTRODUCTION

Acquired melanocytic nevi (AMN) share many phenotypic characteristics of early cutaneous melanoma and may be excised as a precautionary measure to be diagnosed through histopathology. The histopathologic classification of an AMN is dependent on the location of the melanocytic proliferations or nests (junctional, intradermal, or compound). Broadly speaking, if histopathologic features of melanoma are not identified, AMN are diagnosed as benign, or if architectural disorder or atypia are observed, then a classification of dysplastic may be provided (Gerami and Busam, 2018). There are varying degrees of dysplasia (e.g., mild/moderate/ severe or low/high grade) that may be observed (Gerami and

1 Department of Epidemiology, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA; [2]The University of Queensland Diamantina Institute, The University of Queensland, Dermatology Research Centre, Brisbane, Australia; [3]IQ Pathology, Brisbane, Australia; [4]Pathology Queensland, Princess Alexandra Hospital, Brisbane, Australia; [5]Department of Dermatology, Sunshine Coast Hospital and Health Service, Birtinya, Australia; [6]Department of Dermatology, Princess Alexandra Hospital, Brisbane, Australia; [7]Department of Molecular & Systems Biology, Dartmouth Geisel School of Medicine, Hanover, New Hampshire, USA; and [8]Department of Community & Family Medicine, Dartmouth Geisel School of Medicine, Hanover, New Hampshire, USA 9 These authors contributed equally to this work.

Correspondence: Mitchell S. Stark, Dermatology Research Centre, The University of Queensland Diamantina Institute, Level 5, Translational Research Institute, 37 Kent Street, Woolloongabba, Brisbane 4102, Australia. E-mail: m.stark@uq.edu.au

Abbreviations: AMN, acquired melanocytic nevi; CNA, copy number aberration; CpG, 5^0-C-phosphate-G-3^0; FDR, false discovery rate

Received 30 June 2021; revised 16 November 2021; accepted 18 November

2021; accepted manuscript published online 3 December 2021; corrected proof published online 15 March 2022

Busam, 2018); however, this nomenclature is difficult to have a consensus among pathologists (Elmore et al., 2017). The decision to excise a melanocytic lesion is based on phenotypical characteristics, and because AMN are a heterogeneous collection of lesions, dermoscopy is used as the primary method to aid in clinical decision making. AMN with a reticular or nonspecific pattern are often flat, corresponding to junctional or compound melanocytic nevi. AMN with a globular pattern are frequently raised, or are papillomatous, and primarily have intradermal melanocytic nests. Globular nevi that are raised tend to be stable throughout life, and clinically, they are very different from reticular/nonspecific nevi, which are acquired during puberty and adulthood (reviewed in Tan et al. [2019]). Nonspecific nevi with a peripheral rim of globules have been determined to be growing symmetrically, with apparent growth cessation occurring during a span of 4–5 years (Bajaj et al., 2015; Kittler et al., 2000). Melanoma has been found to arise within both benign and dysplastic nevi, which may be flat or raised;

ª 2021 The Authors. Published by Elsevier, Inc. on behalf of the Society for Investigative Dermatology.

however, there is a lack of evidence of which type of nevus is more prone for melanoma development. Along with common phenotypic characteristics, nevi share numerous molecular features with cutaneous melanoma. The BRAF oncogene is well-known to be mutated in nevi (Pollock et al., 2003), with the mutation found in up to 100% of nevi exhibiting growth hallmarks (Tan et al., 2018), indicating that this mutation is involved in proliferation. Because most AMN harbor this oncogenic mutation, this event does not constitute progression toward malignancy. We and others have investigated nevi using whole-genome and exome sequencing (Colebatch et al., 2019; Stark et al., 2018) to reveal the common UVR mutation signatures between nevi and melanoma. Importantly, some melanomas have been observed to arise from a pre-existing nevus, with molecular studies (Shain et al., 2015) identifying some early genomic events (e.g., TERT mutation and copy number loss of tumor suppressor genes) that occur before malignant transformation.

Despite the similarities described earlier, most nevi remain stable, and very few will transform into melanoma, so it is important to understand and recognize early molecular changes in AMN. DNA methylation profiling using a panel of genes has established that differential methylation profiles of only 22 genes can accurately determine (area under the curve > 0.95) a melanoma from an AMN (Conway et al., 2011). Moreover, genome-wide methylation profiling of healthy skin revealed a clear methylation signature associated with a high nevus count (Roos et al., 2017). We have previously determined that raised, papillomatous nevi with a globular dermoscopic pattern are genomically silent, with no copy number aberrations (CNAs), whereas reticular/nonspecific nevi had significant ($P < 0.0001$) CNAs (Stark et al., 2018), often observed in melanoma. In this study, we performed genome-wide methylation profiling of AMN with globular and reticular/nonspecific patterns to reveal distinct methylation profiles.

RESULTS

Study population and mutation status

A total of 32 paired nevus and adjacent normal samples were collected from 24 study participants, with eight study participants contributing two nevus-adjacent normal pairs (Table 1). All excised nevi were dermoscopically classified (Supplementary Materials and Methods) before histopathologic diagnosis. It was confirmed that all specimens were melanocytic nevi and could be classified as benign (n ¼ 13) or dysplastic (n ¼ 19) with junctional, compound, or intradermal histopathologic architecture (Table 1 and Supplementary Table S1). No statistically significant differences were observedfortheage,sex,andUVexposurestatusofthebenign and dysplastic nevi (Supplementary Table S2). Nevi histopathologically diagnosed as benign either had a globular dermoscopic pattern (69% or 9 of 13), or a nonspecific pattern (31% or 4 of 13) (Supplementary Tables S1 and S2). Dysplastic nevi from this collection revealed primarily reticular/nonspecific pattern, or were nevi with a peripheral rim of globules (95% or 18 of 19) (Supplementary Tables S1 and S2), with only one globular nevus exhibiting dysplastic histopathologic features. The BRAF and NRAS mutation status were confirmed in 30 of 32 nevi using droplet digital PCR (Tan et al., 2018) or

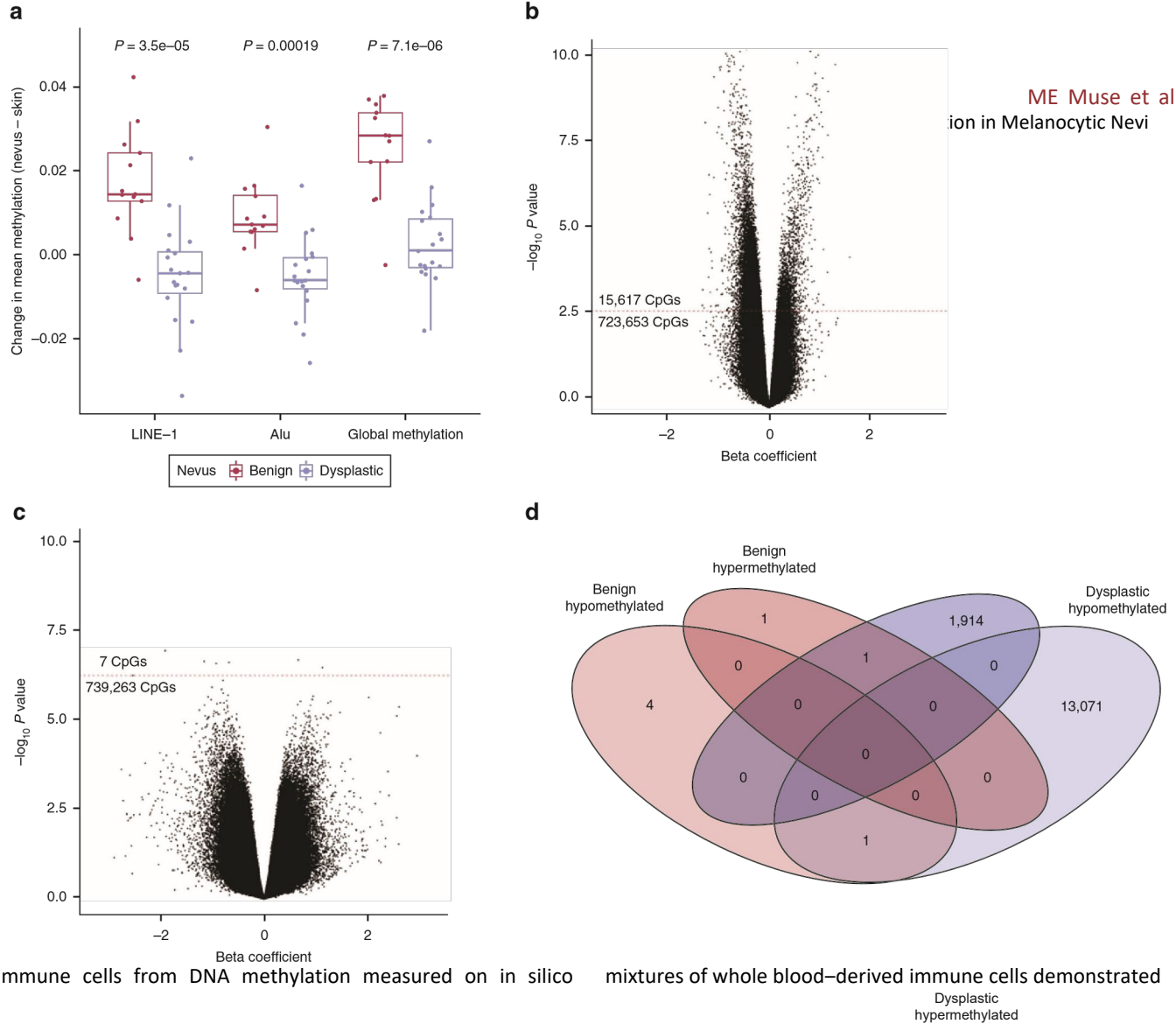
Table 1. Subject and Sample Characteristics

Characteristics	Subjects n [24	Samples n [32
Sex, n (%)		
Female	7 (29.2)	
Male	17 (70.8)	
Age, mean (SD)	51.88 (14.36)	
Status, n (%)		
Benign		13 (40.6)
Dysplastic		19 (59.4)
Histopath, n (%)		
Compound		13 (40.6)
Intradermal		12 (37.5)
Junctional		7 (21.9)
Dermoscopic pattern, n (%)		
Globular		10 (31.2)
Reticular		4 (12.5)
Nonspecific (PRG)		10 (31.2)
Nonspecific (other)		8 (25.0)
Sun exposure, n (%)		
Exposed		25 (78.1)
Shielded		7 (21.9)
		2 (6.7)
BRAF/NRAS, n (%)*		
BRAF V600E		23 (76.7)
BRAF V600E c. 1799_1800delTGinsAA		
BRAF V600K		3 (10)
BRAF V600R		1 (3.3)
NRAS G12S		1 (3.3)
unknown		2 (-)

Abbreviation: PRG, peripheral rim of globules.
*Mutation frequency is determined in 30 nevi with available DNA.

whole-exome sequencing (Table 1 and Supplementary Table S1). Insufficient DNA was available for 2 of 32 nevi to perform BRAF/NRAS mutation detection. BRAF was confirmed to be frequently mutated at codon 600 in 29 of 30 nevi (96.66%), with BRAF V600E being the most frequent (83.33%), followed by V600K (10%) and V600R (3.33%). In BRAF wild-type nevi, the NRAS G12S mutation (3.33%) was present. Taken together, this nevus cohort was 100% mutated for either BRAF or NRAS, which is consistent with our previous observations (Tan et al., 2018).

Inferred proportions of melanocytes differed in histopathologically diagnosed benign and dysplastic nevi We created a reference library for DNA methylation–based cell-type deconvolution of skin biopsies using publicly available cell line data for epithelial cells, fibroblasts, and melanocytes (GSE74877); and purified B cells, $CD4^b$ T cells, $CD8^b$ T cells, NK cells, monocytes, and neutrophils (GSE110554). The optimized reference library for deconvoluting epidermal tissue cell types included 609 5^0-C-phosphate-G-3^0 (CpG) probes for estimating the proportions of melanocytes, epithelial cells, fibroblasts, B cells, $CD4^b$ T cells, $CD8^b$ T cells, monocytes neutrophils, and NK cells (Supplementary Table S3 and Supplementary Figure S1a-c). Using the generated library, estimated relative proportions of



immune cells from DNA methylation measured on in silico mixtures of whole blood–derived immune cells demonstrated

e

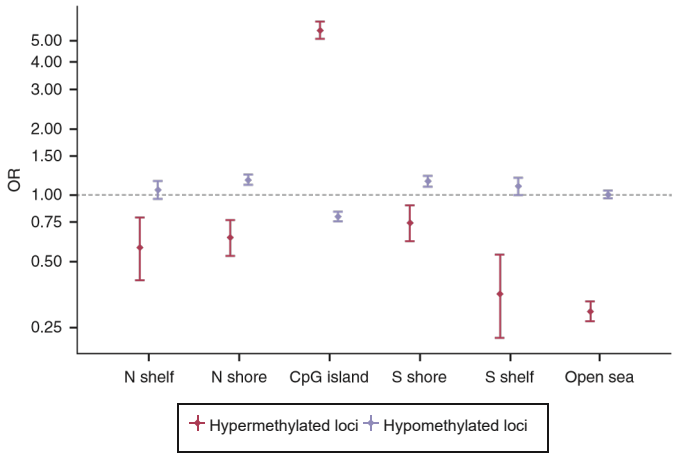


Figure 1. DNA methylation differences in benign and dysplastic nevi relative to those in adjacent normal skin. (a) Distribution of the change in mean methylation between nevi and adjacent normal skin for LINE-1 and Alu repeat elements and overall methylation status. Epigenome-wide association analyses identifying CpG sites that are significantly differentially methylated in (b) dysplastic nevi and (c) benign nevi relative to those in adjacent normal skin, adjusted for estimated proportions of epithelial cells, fibroblasts, and melanocytes. Dashed lines in (b) and (c) indicated a FDR significance threshold of Q < 0.05. (d) Overlap of identified hypermethylated and hypomethylated (Q < 0.05) loci in benign and dysplastic nevi relative to those in adjacent normal skin. (e) Enrichment analysis of all differentially methylated CpG loci in dysplastic nevi relative to those in adjacent normal skin. CpG, 5^0-C-phosphate-G-3^0; FDR, false discovery rate.

a strong correlation (R^2 > 0.99) with the true relative proportions of included cell types (Supplementary Figure S2). In addition, estimated relative proportions of all cell types included in the library from computationally generated in silico mixtures demonstrated a strong correlation (R^2 > 0.99) with the true relative proportions (Supplementary Figure S3).

After projecting the library on to the nevi and adjacent skin samples to estimate cellular composition, median melanocyte proportions were found to be higher in benign nevi (46.9%) than in dysplastic nevi (14.0%) (Supplementary Figure S1b). The benign nevi in this collection were primarily nevi that had a raised clinical presentation, with a globular dermoscopic pattern (69%). Because these globular nevi all have an intradermal histopathologic architecture, corresponding to high nevus cell cellularity, the DNA methylation–based reference library was congruent with nevus biology. Furthermore, the dysplastic nevi, which primarily had a flat clinical presentation with a reticular/nonspecific dermoscopic pattern (95%), are known to have low nevus cell cellularity. In fact, the proportion of melanocytes in the nevus samples estimated from deconvolution were strongly correlated (R^2 ¼ 0.82) with a dermatopathologist-provided (HPS) semiquantitative assessment of the relative proportion of melanocytes in each nevus sample on the basis of H&E staining (Supplementary Figure S1c).

Global methylation and inferred methylation of repeat element regions of the genome in nevi

We compared the difference in DNA methylation between nevi and paired adjacent normal skin at LINE-1 and Alu repeat elements as well as the mean methylation of all measured probes on the array. Genome-scale methylation and repeat element measures were higher in our collection of benign nevi and lower in dysplastic nevi than in adjacent normal skin. The mean methylation change differences between nevi types were statistically significant for all the three measures (see Figure 1a). In addition, assessing the inferred repeat element methylation in repeat elements tracking to the chr9p21 region, a region known to be rich with repeat elements, as well as having well-known association with melanoma, inferred methylation status of LINE-1 repeat elements tracking to tumor suppressor gene CDKN2A was increased in benign nevi relative to that in paired adjacent normal skin, whereas it was decreased in dysplastic nevi relative to that in paired adjacent normal skin (Supplementary Figure S4).

Hypermethylated loci in dysplastic nevi show strong enrichment for CpG dense CpG island regions

When comparing nevi classified as dysplastic (n ¼ 19) with paired adjacent normal skin, 15,617 loci were identified as differentially methylated between the two tissue types (false discovery rate [FDR] < 0.05) (Figure 1b and Supplementary Table S4), and only seven loci were differentially methylated in benign nevi (n ¼ 13) relative to those in paired adjacent normal skin (FDR < 0.05) (Figure 1c and Supplementary Table S4). Of the significantly differentially methylated loci, only one hypermethylated (cg11264521) and one hypomethylated (cg02233071) loci were shared between benign and dysplastic nevi relative to those in adjacent normal skin (Figure 1d). Among the 15,617 differentially methylated loci in dysplastic nevi, 1,915 loci were hypermethylated relative to those in adjacent normal skin, and these 1,915 hypermethylated loci showed 5.6-fold enrichment (95% confidence interval: 5.1–6.1, P ¼ 2.4 10^{-290}) for CpG island regions (Figure 1e). Among the five most significantly hypermethylated loci were CpG loci mapping to the promoter regions of TTC21A, ASAP1, HES5, PREPL, and the documented melanoma tumor suppressor RASSF5 (Aoyama et al., 2004) (Supplementary Table S5a). Conversely, the most

hypomethylated loci predominantly mapped to gene body regions (Supplementary Table S5b), including GALNT15, SSH1, COBL, and CYGB.

Candidate oncogenes and tumor suppressor genes in melanoma demonstrate altered DNA methylation profiles in benign and dysplastic nevi relative to those in adjacent normal skin

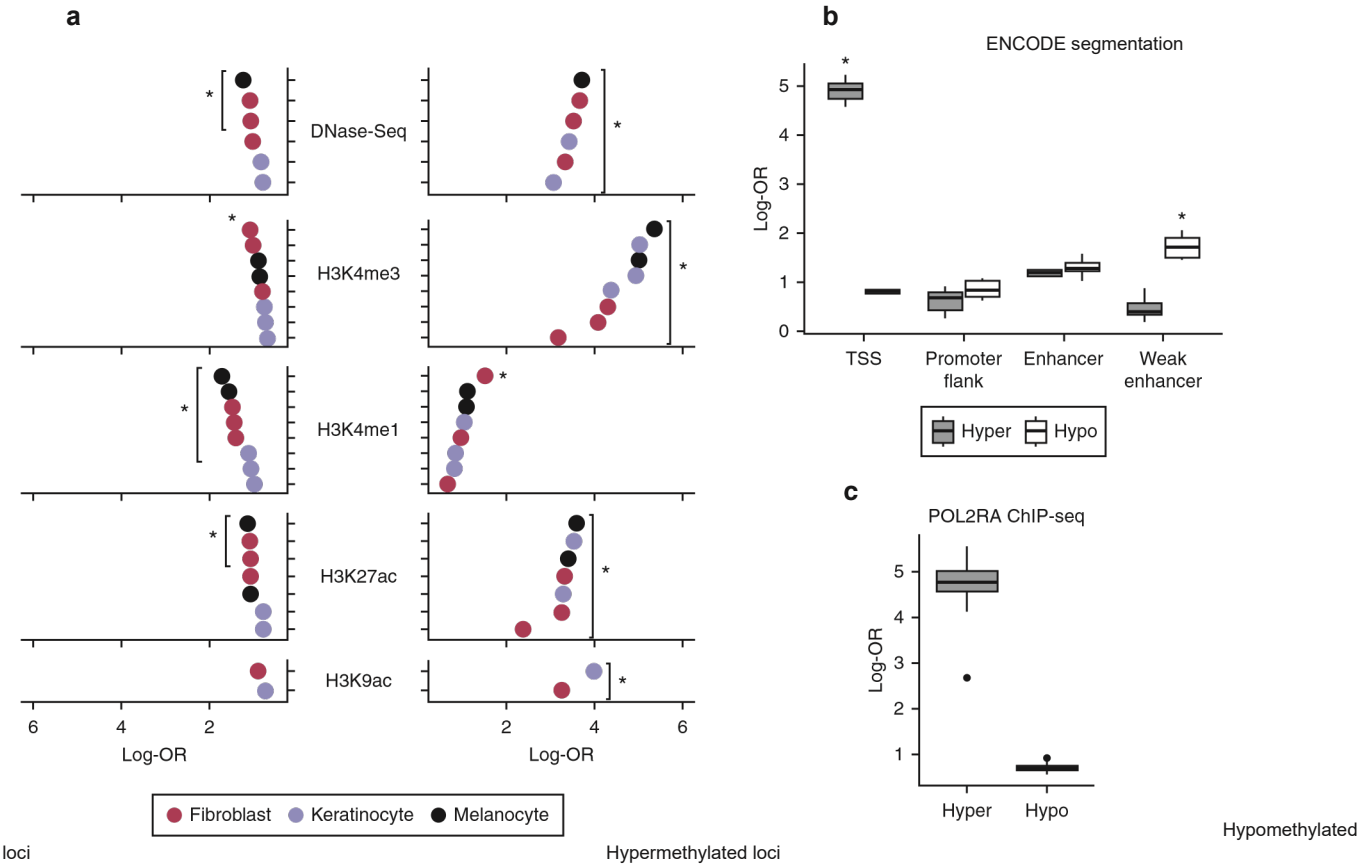


Figure 2. Enrichment of differentially methylated loci in dysplastic nevi in functional genomic regions. (a) Log-OR enrichment of hypermethylated and hypomethylated CpG loci in fibroblast, keratinocyte, and melanocyte Roadmap epigenomic datasets (* ¼ Q < 0.05, Fisher's exact test). (b) Log-OR enrichment of differentially methylated loci in ENCODE genome segments in six cell types (* ¼ Q < 0.05 across all cell types, Fisher's exact test). (c) Log-OR enrichment of differentially methylated loci in POL2RA ChIP-Seq peaks in 52 cell types from ENCODE. ChIP-Seq, chromatin immunoprecipitation sequencing; CpG, 5^0-Cphosphate-G-3^0; Dnase-Seq, DNase I hypersensitive sites sequencing; ENCODE, Encyclopedia of DNA Elements; TSS, transcription start site.

Differentially methylated loci in dysplastic nevi are enriched for gene regulatory regions and active gene promoters The 1,915 hypermethylated loci were strongly enriched for marks of active promoters in all skin cell types from the Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010), including chromatin accessibility assessed by DNase I hypersensitive sites sequencing (log-OR up to 3.71 in foreskin melanocytes) and chromatin immunoprecipitation sequencing for histone marks, including H3K4me3, H3K27ac, H3K9ac, as well as the active enhancer mark H3K4me1 (Figure 2a). Of these skin-cell–type histone mark chromatin immunoprecipitation sequencing datasets, the hypermethylated loci were most enriched consistently in foreskin melanocytes. Using the Encyclopedia of DNA Elements resource, the hypermethylated loci were also enriched for transcription start site genomic segments on the basis of predictions of regulatory activity in six cell types (Figure 2b) and were strongly enriched for the binding of POL2RA across 52 cell types and contexts (ENCODE Project Consortium, 2012) (Figure 2c).

Mean methylation was assessed in candidate oncogenes and tumor suppressor genes. Please refer to Supplementary Materials and Methods for details of these analyses. After adjusting for cellular composition, although the effect size was considered small, MAP2K1 and CDKN2A demonstrated statistically significant (FDR < 0.05) hypomethylation of the gene body in dysplastic nevi relative to those in adjacent normal skin (Supplementary Table S5a and Supplementary Table S7). Next, PIK3CA demonstrated statistically significant promoter hypomethylation, whereas MDM2 demonstrated statistically significant promoter hypermethylation (Supplementary Table S6a and Supplementary Table S7) in dysplastic nevi relative to that in adjacent normal skin. In addition, MYC showed statistically significant hypomethylation of the gene body in benign nevi relative to that in adjacent normal skin (Supplementary Table S6b).

Nevi with a reticular/nonspecific pattern have distinct DNA methylation profiles from adjacent normal skin

Independent of histopathologic diagnosis, we next sought to determine whether there were any global methylation differences on the basis of dermoscopic classifications. Unsupervised clustering of the top 5,000 most variable loci, across all samples, showed no distinct clustering patterns of reticular and nonspecific nevi on the basis of dermoscopic pattern (Supplementary Figure S5). Therefore, for subsequent analyses, dermoscopic patterns were grouped as globular versus nonglobular. Interestingly, globular nevi (n ¼ 10) had no identified differentially methylated CpG loci relative to paired adjacent normal skin at an FDR threshold of Q < 0.05 (Figure 3a). However, when assessing the nevi classified as nonglobular (reticular/nonspecific; n ¼ 22), there were 59,572 loci identified as differentially methylated relative to those in paired adjacent normal skin at an FDR threshold of Q < 0.05 (Figure 3b). Of these loci, 8,852 were hypermethylated in nevi relative to those in adjacent normal skin, which demonstrated a 1.9-fold enrichment (95% confidence interval: 1.8–2.0; P ¼ 3.2 10^{-146}) for CpG dense CpG island regions (Figure 3c). A total of 1,458 of these hypermethylated loci were also identified as hypermethylated in dysplastic nevi relative to those in adjacent normal skin (Supplementary Figure S6).

Differentially methylated loci in nevi with a reticular/nonspecific pattern are enriched for gene regulatory regions and active gene promoters

To further investigate the location and function of the broad methylation changes observed in reticular/nonspecific patterned nevi compared with those in adjacent normal skin, we tested the differentially methylated loci for enrichment in functional genomic regions using publicly available datasets (Locus Overlap Analysis) (Sheffield and Bock, 2016) as described earlier. Notably, the 8,852 hypermethylated loci (compared with 1,915 loci when grouped as dysplastic) were strongly enriched for marks of active promoters in all skin cell types from the Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010), including chromatin accessibility assessed by DNase I hypersensitive sites sequencing (log-OR up to 2.76 in adult dermal fibroblasts) and chromatin immunoprecipitation sequencing for histone marks including H3K4me3, H3K27ac, H3K9ac, as well as the active enhancer mark H3K4me1 (Supplementary Figure S7a). Of these skincell–type histone mark chromatin immunoprecipitation sequencing datasets, the hypermethylated loci were consistently most enriched in epidermal keratinocytes. Using the Encyclopedia of DNA Elements resource, the 8,852 hypermethylated loci were also enriched for transcription start sites, promoter flanking, and enhancer genomic segments on the basis of predictions of regulatory activity in six cell types (Supplementary Figure S7b) and were enriched for the binding of POL2RA across 52 cell types and contexts from the Encyclopedia of DNA Elements consortium (ENCODE Project Consortium, 2012) (Supplementary Figure S7c).

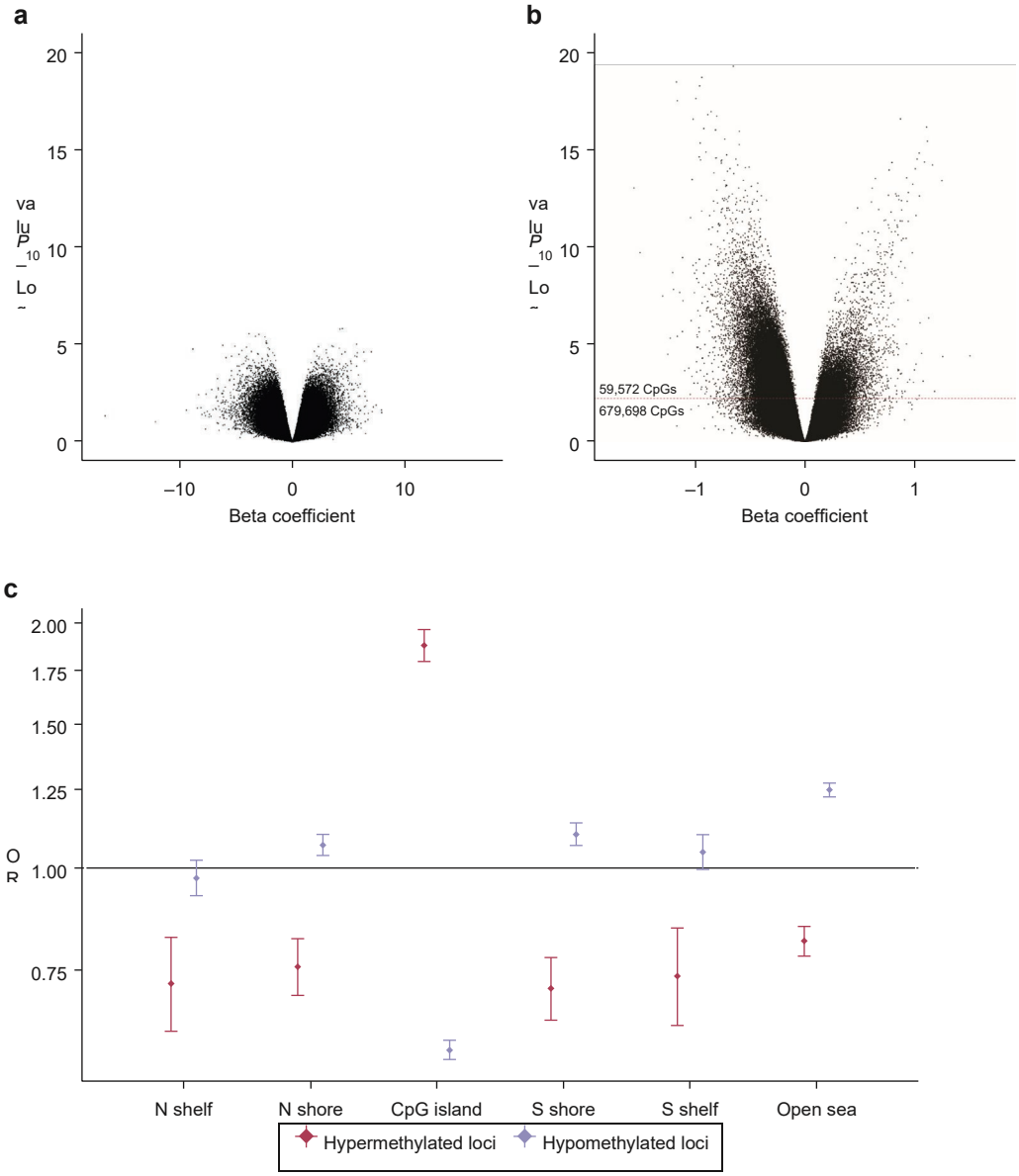


Figure 3. Epigenome-wide association analyses identifying CpG sites that are significantly differentially methylated in globular nevi and nonglobular nevi relative to those in adjacent normal skin. The data represented in panels a and b illustrates CpG sites in (a) globular and (b) nonglobular nevi that have been adjusted for the estimated proportions of epithelial cells, fibroblasts, and melanocytes. Dashed lines in (b) indicate a FDR significance threshold of Q < 0.05. (c) Enrichment analysis of all differentially methylated CpG loci in nonglobular nevi relative to those in adjacent normal skin. CpG, 5^0-C-phosphate-G-3^0; FDR, false discovery rate.

Investigation of the 50,720 hypomethylated loci revealed that these were not enriched for promoter-specific features. Instead, they were enriched for accessible chromatin in melanocytes (log-OR of 1.22) and fibroblasts derived from adult foreskins (log-OR up to 1.09) as well as H3K4me1 in all Roadmap skin cell types (log-OR up to 1.89 in foreskin melanocytes) (Supplementary Figure S7a). In addition, H3K27ac histone marks were significantly enriched in adult foreskin melanocytes (log-OR of 1.06) and adult dermal fibroblasts (log-OR of 1.03), suggesting that this set of hypomethylated loci may contain enhancer elements (Supplementary Figure S7a). Further analysis indicated that the hypomethylated loci were enriched for Encyclopedia of DNA Elements enhancer segments and were strongly enriched for weak enhancer segments (Supplementary Figure S7b).

PTEN demonstrates promoter hypermethylation in nevi with a reticular/nonspecific pattern relative to adjacent normal skin

Of the hypermethylated loci in nonglobular nevi relative to those in adjacent normal skin that occur in gene promoter regions, PTEN had the highest proportion of hypermethylated promoter CpG loci among all genes (Table 2). Further

investigation of the pattern of promoter hypermethylation demonstrated that the nine significantly hypermethylated loci (Q < 0.05) in nonglobular nevi, relative to those in adjacent normal skin were grouped at the start of the promoter region (Figure 4).

Table 2. Genes with the Top 10 Most Differentially Methylated Promoter Regions

Gene	Hypermethylated Loci	Total Loci	Proportion Hypermethylated
PTEN	9	42	0.21
PRR15	8	9	0.89
SPTBN1	6	17	0.35
PAMR1	5	9	0.56
NFIX	5	10	0.50
MIR375	5	12	0.42
LOC100132215	5	15	0.33
ADM	5	17	0.29
HES5	4	9	0.44
ATOH8	4	10	0.40

Abbreviation: CpG, 5^0-C-phosphate-G-3^0.

The genes are sorted first by the total number of hypermethylated promoter loci (Q < 0.05) and then by the proportion of CpG loci in the promoter that was hypermethylated.

DISCUSSION

Using genome-scale methylation profiling, we have presented evidence that the benign nevi assessed in this study were significantly different from the dysplastic nevi, sharing only two differentially methylated loci relative to adjacent normal skin. As previously stated, the benign nevi in our study primarily were raised, or were papillomatous, and exhibited a globular (69%) dermoscopic pattern, whereas the dysplastic nevi were flat, and had a reticular/nonspecific (95%) pattern. We have previously illustrated that globular nevi are genomically stable, with no CNAs present, which contrasts with reticular/nonspecific nevi, which have numerous CNAs (Stark et al., 2018).

To provide further support of their stability, we have now confirmed that our collection of benign nevi has significantly higher mean methylation at LINE-1 and Alu repeat elements (which are commonly methylated across the genome in normal cells) than adjacent normal skin, which contrasts with dysplastic nevi that had lower mean methylation. LINE-1 and Alu repeat elements are commonly found in the intergenic regions of our genome. These retrotransposable elements are related to genomic stability through methylation regulation. Furthermore, hypomethylation of these elements has been associated with genomic instability (Ayarpadikannan and Kim, 2014; Su et al., 2012). These data confirm our previous CNA analysis of globular nevi having no CNAs and provide reasoning for the increased CNA observed in reticular/nonspecific nevi (Stark et al., 2018). In addition, we focused specifically on the chr9p21 region (containing the CDKN2A gene locus), to identify any differential methylation patterns in intergenic regions, as this region has been previously associated with flat nevi using quantitative-trait locus analysis (Zhu et al., 1999). In the study by Zhu et al. (1999), flat nevi were classed as junctional or compound (i.e., primarily reticular/nonspecific). The authors concluded that because germline mutations were rare in CDKN2A, noncoding regions of this locus were likely responsible for high (flat) nevus count (Zhu et al., 1999). We provide evidence that hypomethylation of LINE-1 repeat elements was significantly associated with our collection of dysplastic nevi, which were 95% flat or reticular/nonspecific. Interestingly, LINE-1 elements are located within w0.5 mb from D9S942, the specific polymorphic marker described in the Zhu et al. (1999) study.

A strength of our approach is the development of a library for the estimation of the cellular composition of the samples. Although in our study, only modest differences are observed between dysplastic nevi and adjacent normal skin; after adjusting for cellular composition and accounting for multiple comparisons, loci that were identified as being hypermethylated in dysplastic nevi demonstrated strong enrichment for CpG dense CpG island regions. Importantly, although our cell-type library performed well in validation analyses, there remains the potential for error in the estimation of cellular composition, thus possibly introducing bias into our findings because these results were heavily altered after adjustment for cellular composition. CpG islands are often hypermethylated in tumor tissue relative to those in adjacent normal skin and commonly overlap with gene promoter regions (Esteller, 2002). Therefore, the hypermethylation events observed in our set of dysplastic nevi may play a role in early carcinogenic processes in melanoma.

In addition, when assessing oncogenes and tumor suppressor genes previously implicated in melanoma, we observed promoter hypomethylation, associated with increased transcription (Jones and Baylin, 2002), of the documented oncogene PIK3CA and promoter hypermethylation, associated with transcriptional repression (Suzuki and Bird, 2008), in MDM2. Because MDM2 generally acts in an oncogenic manner, precisely why MDM2 would have increased promoter methylation remains to be determined. We also observed gene body hypomethylation, associated with decreased gene transcription (Suzuki and Bird, 2008), in the oncogene MAP2K1 as well as in the tumor suppressor gene CDKN2A, in dysplastic nevi relative to that in adjacent normal skin.

Although significant differences were observed between dysplastic nevi and adjacent normal skin, vastly more differentially methylated loci were observed between reticular/nonspecific nevi and adjacent normal skin. Similar to the analysis of the dysplastic nevus group, hypermethylated loci in reticular/nonspecific nevi relative to those in adjacent normal skin were enriched for CpG island regions, which commonly

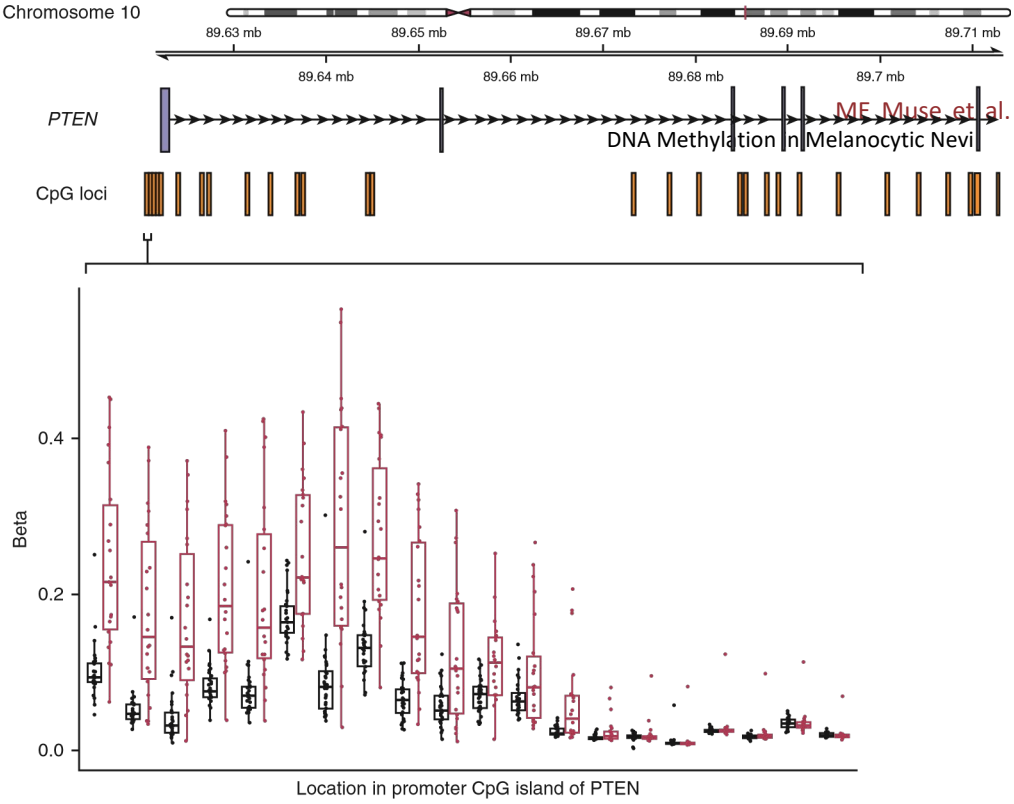


Figure 4. Distribution of methylation across the *PTEN* promoter in nonglobular and adjacent normal skin. The zoomed-in section of the *PTEN* promoter region highlights the first 15 loci of each sample that contains the nine hypermethylated ($Q < 0.05$) loci in nonglobular nevi relative to those in adjacent normal skin. CpG, 5′-C-phosphate-G-3′.

overlap with gene promoter regions (Esteller, 2002). Importantly, these hypermethylated loci were also enriched for gene promoter regions, enhancer regions, and regions enriched for RNA polymerase II binding, suggesting that the gain of DNA methylation observed in these nevi plays a role in the regulation of gene expression.

In gene-level assessments of DNA methylation alterations, we identified PTEN as having the most significantly hypermethylated loci in its promoter region of any tested gene. This unbiased assessment confirms a recent study that found that the PTEN promoter is commonly methylated in nevi, albeit significantly lower than in melanoma (Salgado et al., 2020). Given that promoter hypermethylation is associated with gene silencing (Suzuki and Bird, 2008) and that PTEN is a well-documented tumor suppressor gene across multiple tumor types, including melanoma (Wu et al., 2003), this suggests that alterations to the regulation of PTEN occur early on in the development in melanoma, and can be detected in flat acquired nevi, potentially priming these nevi for subsequent carcinogenesis. The MAPK pathway in nevi is constitutively activated, commonly by mutations in BRAF, leading to increased proliferation. Some studies have shown that this overactivation leads to oncogene-induced senescence (Michaloglou et al., 2005) and, in turn, the stability of nevi. This assumption cannot be applied to all nevus types, and only partially explains their relative stability. We have previously documented that BRAF was 100% mutated in nevi that are actively growing (the presence of globules around the periphery of the nevus) (Tan et al., 2018). Our observation of PTEN promoter hypermethylation in nevi with a reticular/ nonspecific pattern (which includes nevi that are actively growing), provides further support for the activation of the phosphoinositide 3-kinase pathway, being a crucial step to overcome BRAF V600E–induced senescence. This notion is supported by an elegant study using a BRAF V600E–driven mouse model, which induced nevus growth (Vredeveld et al., 2012). The authors showed that depletion of PTEN using short-hairpin RNA, led to tumor progression through the activation of phosphoinositide 3-kinase. Melanoma is commonly mutated for BRAF and NRAS and rarely for PIK3CA or its subunits (Curtin et al., 2006), which is mirrored in AMN. The growth of a nevus is often a very slow process, with active growth spanning years before growth arrest occurs. PTEN promoter hypermethylation likely plays a role in this process, but other mechanisms warrant further investigation.

The limitations of this study are the number of nevi diagnosed as benign that have a reticular/nonspecific dermoscopic pattern. All our nevi were clinically benign and were selected on the basis of their dermoscopic pattern. Because 86% (19 of 22) of the nevi with a reticular/nonspecific pattern were diagnosed as dysplastic by histopathology, it would be important to determine whether this trend occurs in a larger sample collection. The benign nevi collection, analyzed in this study, has evidence of UVR damage after mutation signature analysis (Stark et al., 2020). Despite the continued exposure to UVR, known to contribute to hypomethylation in photoaged skin (Vandiver et al., 2015), as observed in our collection of dysplastic nevi, the methylation status of our benign nevi did not fit this model, which is likely due to phenotypic differences as described. An additional limitation of this study is that by only measuring bisulfite converted DNA, and not tandem bisulfite, and oxybisulfite converted DNA, all measures of DNA methylation reflect the relative proportion of both methylated and hydroxymethylated DNA. Therefore, we are unable to draw specific conclusions as to whether the alterations observed reflect alterations in methylation or hydroxymethylation of DNA. Overall, the increased global methylation, and

methylation of repeat elements, are likely one of the driving mechanisms that contribute to nevus stability and the rarity of malignant transformation.

MATERIALS AND METHODS

Study population

This study was approved by the Metro South Human Research Ethics Committee (Brisbane, Australia) (HREC/09/QPAH/162) and was carried out in accordance with the Declaration of Helsinki. With written informed consent, 32 acquired nevi with globular (n ¼ 10) and reticular/nonspecific (n ¼ 12) dermoscopic patterns, and nonspecific nevi with a peripheral rim of globules (n ¼ 10) dermoscopic pattern, identified from a database of prospectively imaged naevi (Daley et al., 2017), were shave excised from 24 participants (17 males and 7 females, mean age of 51.88 years). Of note, 8 of the 24 participants provided consent for two nevi to be shaved excised (n ¼ 16). Please refer to the Supplementary Materials and Methods for more details.

Bisulfite modification for Illumina HumanMethylationEPIC array and data processing

DNA extracted from each nevus and matching adjacent skin (n ¼ 64) was sent to the fee-for-service provider University of Southern California Norris Comprehensive Cancer Center Molecular Genomics Core. Please refer to the Supplementary Materials and Methods for more details.

Reference library development and cell-type estimation We developed a reference library for DNA methylation–based celltype deconvolution of skin biopsies using publicly available cell line data for epithelial cells, fibroblasts, and melanocytes (GSE74877). The library also included primary purified B cells, $CD4^{þ}$ T cells, $CD8^{þ}$ T cells, NK cells, monocytes, and neutrophils from our previous work (GSE110554). Please refer to the Supplementary Materials and Methods for more details.

Repeat element estimation

The methylation status of Alu and LINE-1 repetitive elements was predicted using the REMP package in R (Zheng et al., 2017). The median values of Alu and LINE-1 were calculated across all estimated regions for each sample. Please refer to the Supplementary Materials and Methods for more details.

Statistical analysis

Differential methylation was assessed using linear mixed-effects models, modeling the relationship between logit-transformed beta values (M-values) in nevi samples relative to those in paired adjacent normal samples, controlling for the estimated relative proportions of epithelial cells, fibroblasts, and melanocytes in each sample. Please refer to the Supplementary Materials and Methods for more details.

Data availability statement

The datasets generated and analyzed during this study are available in the Gene Expression Omnibus under the accession number GSE188593 (https://www.ncbi.nlm.nih.gov/geo/).

ORCIDs
Meghan E. Muse: http://orcid.org/0000-0002-4775-2133
Drew T. Bergman: http://orcid.org/0000-0002-8314-7088 Lucas A. Salas: http://orcid.org/0000-0002-2279-4097
Lisa N. Tom: http://orcid.org/0000-0002-9068-1107
Jean-Marie Tan: http://orcid.org/0000-0003-1160-0282
Antonia Laino: http://orcid.org/0000-0002-1101-8762 Duncan Lambie: http://orcid.org/0000-0002-0102-6638
Richard A. Sturm: http://orcid.org/0000-0003-1301-0294
Helmut Schaider: http://orcid.org/0000-0003-3195-2517
H. Peter Soyer: http://orcid.org/0000-0002-4770-561X
Brock C. Christensen: http://orcid.org/0000-0003-3022-426X Mitchell S. Stark: http://orcid.org/0000-0002-4510-2161

CONFLICT OF INTEREST

HPS is a shareholder of MoleMap NZ and e-derm consult GmbH and undertakes regular teledermatological reporting for both companies. HPS is a Medical Consultant for Canfield Scientific and MoleMap Australia Pty and is a medical advisor for First Derm.

ACKNOWLEDGMENTS

The authors would like to thank the study participants and Harald Oey for his technical assistance. This work was funded by a grant from the Centre of Research Excellence for the Study of Nevi (APP1099021) from the National Health and Medical Research Council (Australia) and the Merchant Charitable Foundation. This work was additionally supported by grants R01CA216265 and R01CA253976 to BCC. LAS was supported by Congressionally Directed Medical Research Programs/Department of Defense (W81XWH-20-1-0778) and the National Institute of General Medical Sciences (P20GM104416-09 8299). MSS holds a fellowship (APP1106491) from the National Health and Medical Research Council. HPS holds a National Health and Medical Research Council Medical Research Future Fund Next Generation Clinical Researchers Program Practitioner Fellowship (APP1137127).

AUTHOR CONTRIBUTIONS

Conceptualization: BCC, HS, MSS; Formal Analysis: MEM, DTB, BCC, HPS, MSS; Funding Acquisition: HPS, MSS, RAS, HS; Visualization: MEM, DTB; Writing - Original Draft Preparation: MEM, MSS; Writing - Review and Editing: MEM, DTB, LAS, BCC, LNT, JMT, DL, AL, RAS, HS, HPS, MSS

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at www.jidonline.org, and at https://doi.org/10.1016/j.jid.2021.11.025.

REFERENCES

Aoyama Y, Avruch J, Zhang XF. Nore1 inhibits tumor cell growth independent of Ras or the MST1/2 kinases. Oncogene 2004;23:3426e33.

Ayarpadikannan S, Kim HS. The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. Genomics Inform 2014;12:98e104.

Bajaj S, Dusza SW, Marchetti MA, Wu X, Fonseca M, Kose K, et al. Growthcurve modeling of nevi with a peripheral globular pattern. JAMA Dermatol 2015;151:1338e45.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 2010;28:1045e8.

Colebatch AJ, Ferguson P, Newell F, Kazakoff SH, Witkowski T, Dobrovic A, et al. Molecular genomic profiling of melanocytic nevi. J Invest Dermatol 2019;139:1762e8.

Conway K, Edmiston SN, Khondker ZS, Groben PA, Zhou X, Chu H, et al. DNA-methylation profiling distinguishes malignant melanomas from benign nevi. Pigment Cell Melanoma Res 2011;24:352e60.

Curtin JA, Stark MS, Pinkel D, Hayward NK, Bastian BC. PI3-kinase subunits are infrequent somatic targets in melanoma. J Invest Dermatol 2006;126: 1660e3.

Daley GM, Duffy DL, Pflugfelder A, Jagirdar K, Lee KJ, Yong XLH, et al. GSTP1 does not modify MC1R effects on melanoma risk. Exp Dermatol 2017;26:730e3.

Elmore JG, Barnhill RL, Elder DE, Longton GM, Pepe MS, Reisch LM, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study [published correction appears in BMJ 2017;358:j3798]. BMJ 2017;357:j2813.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57e74.

Esteller M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. Oncogene 2002;21:5427e40.

Gerami P, Busam KJ. Acquired melanocytic nevi. In: Busam KJ, Germani P, Scolyer RA, editors. Pathology of melanocytic tumors. Philadelphia, PA: Elsevier; 2018. p. 8e25.

Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet 2002;3:415e28.

Kittler H, Seltenheim M, Dawid M, Pehamberger H, Wolff K, Binder M. Frequency and characteristics of enlarging common melanocytic nevi. Arch Dermatol 2000;136:316e20.

Michaloglou C, Vredeveld LC, Soengas MS, Denoyelle C, Kuilman T, van der Horst CM, et al. BRAFE600-associated senescence-like cell cycle arrest of human naevi. Nature 2005;436:720e4.

Pollock PM, Harper UL, Hansen KS, Yudt LM, Stark M, Robbins CM, et al. High frequency of BRAF mutations in nevi. Nat Genet 2003;33:19e20.

Roos L, Sandling JK, Bell CG, Glass D, Mangino M, Spector TD, et al. Higher nevus count exhibits a distinct DNA methylation signature in healthy human skin: implications for melanoma. J Invest Dermatol 2017;137: 910e20.

Salgado C, Oosting J, Janssen B, Kumar R, Gruis N, van Doorn R. Genomewide characterization of 5-hydoxymethylcytosine in melanoma reveals major differences with nevus. Genes Chromosomes Cancer 2020;59: 366e74.

Shain AH, Yeh I, Kovalyshyn I, Sriharan A, Talevich E, Gagnon A, et al. The genetic evolution of melanoma from precursor lesions. N Engl J Med 2015;373:1926e36.

Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and bioconductor. Bioinformatics 2016;32: 587e9.

Stark MS, Denisova E, Kays TA, Heidenreich B, Rachakonda S, Requena C, et al. Mutation signatures in melanocytic nevi reveal characteristics of defective DNA repair. J Invest Dermatol 2020;140:2093e6.e2.

Stark MS, Tan JM, Tom L, Jagirdar K, Lambie D, Schaider H, et al. Whole-exome sequencing of acquired nevi identifies mechanisms for development and maintenance of benign neoplasms [published correction apears in J Invest Dermatol 2018;138:2085]. J Invest Dermatol 2018;138:1636e44.

Su J, Shao X, Liu H, Liu S, Wu Q, Zhang Y. Genome-wide dynamic changes of DNA methylation of repetitive elements in human embryonic stem cells and fetal fibroblasts. Genomics 2012;99:10e7.

Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 2008;9:465e76.

Tan JM, Tom LN, Jagirdar K, Lambie D, Schaider H, Sturm RA, et al. The BRAF and NRAS mutation prevalence in dermoscopic subtypes of acquired naevi reveals constitutive mitogen-activated protein kinase pathway activation. Br J Dermatol 2018;178:191e7.

Tan JM, Tom LN, Soyer HP, Stark MS. Defining the molecular genetics of dermoscopic naevus patterns. Dermatology 2019;235:19e34.

Vandiver AR, Irizarry RA, Hansen KD, Garza LA, Runarsson A, Li X, et al. Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. Genome Biol 2015;16:80.

Vredeveld LC, Possik PA, Smit MA, Meissl K, Michaloglou C, Horlings HM, et al. Abrogation of BRAFV600E-induced senescence by PI3K pathway activation contributes to melanomagenesis. Genes Dev 2012;26:1055e69.

Wu H, Goel V, Haluska FG. PTEN signaling pathways in melanoma. Oncogene 2003;22:3113e22.

Zheng Y, Joyce BT, Liu L, Zhang Z, Kibbe WA, Zhang W, et al. Prediction of genome-wide DNA methylation in repetitive elements. Nucleic Acids Res 2017;45:8697e711.

Zhu G, Duffy DL, Eldridge A, Grace M, Mayne C, O'Gorman L, et al. A major quantitative-trait locus for mole density is linked to the familial melanoma gene CDKN2A: a maximum-likelihood combined linkage and association analysis in twins and their sibs. Am J Hum Genet 1999;65: 483e92.

SUPPLEMENTARY MATERIALS AND METHODS

Tissue processing

As described previously (Tan et al., 2018), each nevus was bisected, and one half of the tissue was formalin fixed and histopathologically diagnosed by a board-certified pathologist (DL), and was assessed for solar elastosis and semiquantitative cellularity estimates by a qualified dermatopathologist (HPS). The second half of the tissue was dissected to isolate nevus from adjacent nonlesional skin and was either flash frozen at –80 C or preserved in RNAlater (Thermo Fisher Scientific, Waltham, MA) before storage at – 80 C. DNA extraction was performed either using the
QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) or the AllPrep DNA/RNA/miRNA Universal Kit (Qiagen) according to manufacturer's instructions. Briefly, the tissue specimens were placed in Buffer RLT Plus (Qiagen) and b-mercaptoethanol and then disrupted and homogenized using a bullet blender (Next Advance, Troy, New York). The debris was pelleted, and the supernatant was collected for coextraction.

Bisulfite modification for Illumina HumanMethylationEPIC array

A total of four specimens had low DNA yield (<250 ng) and had to undergo restoration using the Infinium HD FFPE DNA Restore Kit (Illumina, San Diego, CA). All DNA samples were then bisulfite converted and assessed for quality using custom real-time PCR probes (University of Southern California, Los Angeles) before undergoing amplification, fragmentation, and hybridization to the Illumina HumanMethylationEPIC array, following standard procedures according to the Infinium MethylationEPIC BeadChip Kit (Illumina). BeadChips were scanned using the iScan system (Illumina) according to the manufacturer's instructions.

Data processing

Sample intensity data (IDAT) files were processed using the functional normalization method from the R package minfi (Aryee et al., 2014). Probes with a detection P > 1.0 10^{-6} in more than 5% of the samples, 5^0-C-phosphate-G-3^0s (CpGs) with high-frequency SNP(s) in the probe that are within five base pairs of the target CpG dinucleotide, probes previously described to be potentially cross-hybridizing, and sex-specific probes, were excluded from subsequent analysis (Zhou et al., 2017). From the original set of 865,859 probes included on the Illumina HumanMethylationEPIC array, a total of 126,589 probes were removed during filtering steps, leaving a total of 739,270 CpG loci included for downstream analyses.

Reference library development and cell-type estimation

Computationally derived artificial mixtures were generated by first calculating the mean of randomly selected subsets of reference samples for a given cell type. Cell type mean methylation profiles were then multiplied by a matrix of fixed relative proportions of component cell types included in the mixture. For in silico mixtures of immune cells (GSE110554), computationally derived artificial mixtures, and DNA methylation profiles of nevi and paired adjacent normal skin, relative proportions of all cell types were estimated using the epidish function (Teschendorff et al., 2017) and specifying the Cell-type identification by Estimating Relative Subsets of RNA Transcripts (CIBERSORT) method (Newman et al., 2015). A semiquantitative assessment of the relative proportion of melanocytes in each nevus sample was made by a pathologist on the basis of H&E staining. Verification of library performance was conducted by calculating the Pearson correlation coefficient between the true relative proportions of cells present in in silico mixtures of immune cells, computationally derived artificial mixtures, and the semiquantitative estimate of melanocyte proportions in nevus samples with estimated proportions using the reference library.

Repeat element estimation

The REMP package in R (Zheng et al., 2017) was utilized for the estimation of the methylation status of Alu and LINE-1 repeat elements. Briefly, the REMP method utilizes random forest to predict the methylation status of individual repeat elements throughout the genome from array-based measures of DNA methylation. To generate summary statistics across all inferred loci, the median Alu and LINE-1 methylation status were calculated, respectively.

Candidate oncogenes and tumor suppressor genes methylation analysis

Mean methylation was assessed in candidate oncogenes (Akt1, BRAF, CDK4, HRAS, IDH1, KRAS, MAP2K1, MDM2, MITF, MTOR, MYC, NRAS, PARP1, PIK3CA, RAC1, SETDBI, SKI, and TERT) and tumor suppressor genes (ARID2, CDKN2A, CDKN2B, NFI, PPP6C, PTEN, RB1, and TP53). All significant differential methylation was collated in the following regions: 1,500 and 200 nucleotides from the transcriptional start site (TSS1500) and (TSS200), which encompasses the putative gene promoter region, the 5^0- and 3^0untranslated regions, the gene body (body), as well as methylation sites mapping to the exon boundary or the regions at the intron/exon boundary.

Statistical analysis

The annotation data for the Illumina HumanMethylationEPIC array was used to identify the proximity to a CpG island for each identified differentially methylated CpG locus. Cochran–Mantel–Haenszel tests were used to assess the enrichment of CpG island contexts among the identified significantly hypomethylated and hypermethylated CpG loci (false discovery rate < 0.05) in nevi relative to those in adjacent normal skin, compared with all other CpG loci included in the analysis.

Locus overlap analysis (LOLA) was used to test the enrichment of genomic regions among differentially methylated CpG loci (Sheffield and Bock, 2016). Briefly, the core and extended locus overlap analysis databases were used to test for the enrichment of genomic regions using a Fisher's exact test among identified differentially methylated CpG loci (false discovery rate < 0.05) relative to those in the background set of all 739,270 CpG loci included in the analysis. Enrichment tests were limited to Roadmap Epigenomics datasets in skin cell types (E055–E059, E061, E126, E127) and segmentation and

chromatin immunoprecipitation sequencing datasets from the Encyclopedia of DNA Elements consortium (ENCODE Project Consortium, 2012).

SUPPLEMENTARY REFERENCES

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. Bioinformatics 2014;30:1363e9.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57e74.

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453e7.

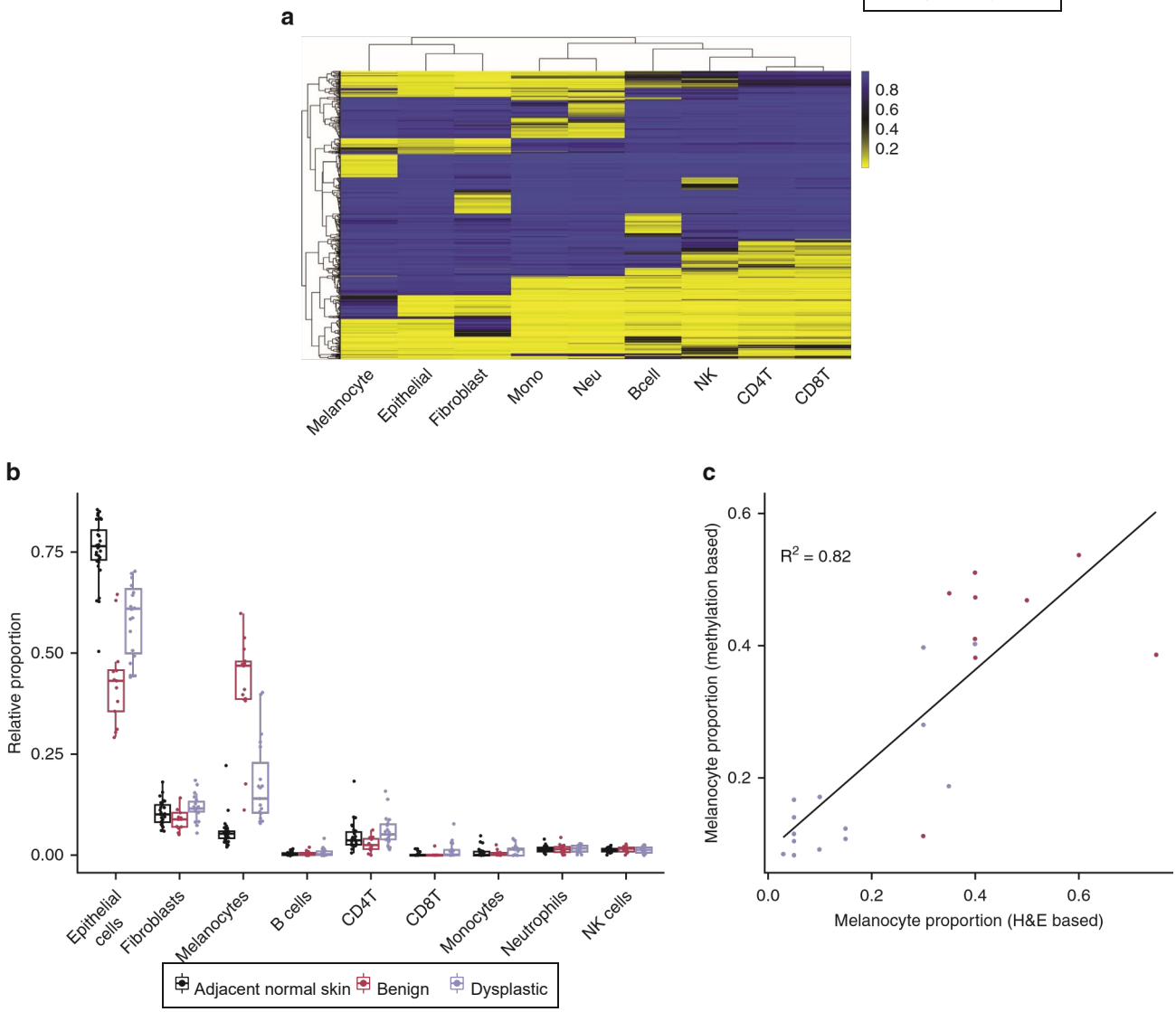
Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics 2016;32:587e9.

Tan JM, Tom LN, Jagirdar K, Lambie D, Schaider H, Sturm RA, et al. The BRAF and NRAS mutation prevalence in dermoscopic subtypes of acquired naevi reveals constitutive mitogen-activated protein kinase pathway activation. Br J Dermatol 2018;178:191e7.

Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of referencebased algorithms for correcting cell-type heterogeneity in epigenomewide association studies. BMC Bioinformatics 2017;18:105.

Zheng Y, Joyce BT, Liu L, Zhang Z, Kibbe WA, Zhang W, et al. Prediction of genome-wide DNA methylation in repetitive elements. Nucleic Acids Res 2017;45:8697e711.

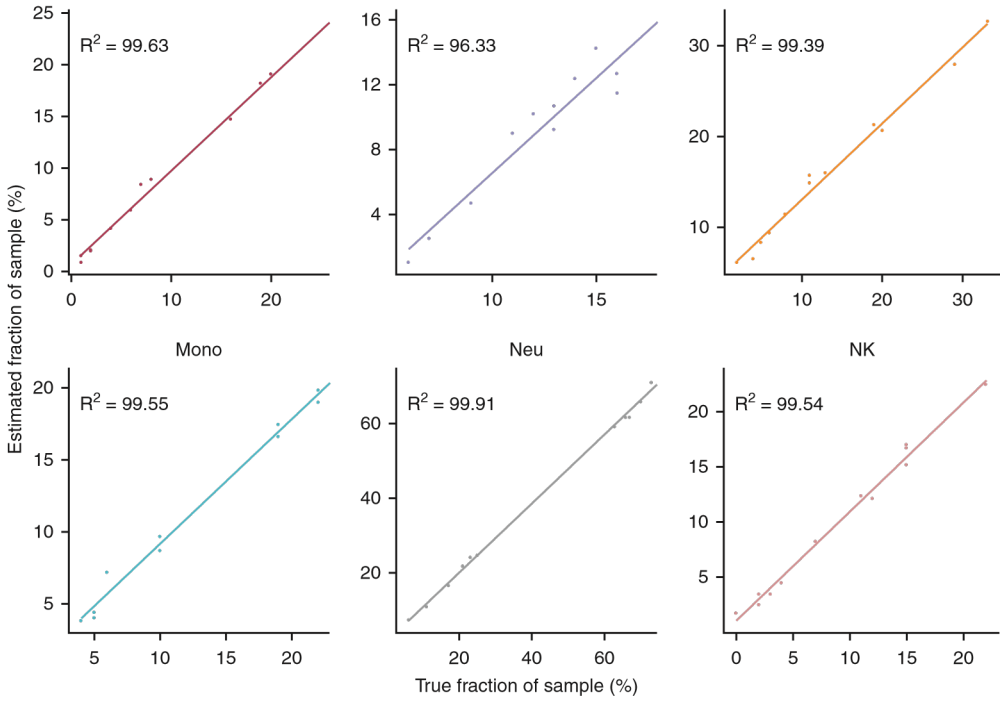
Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res 2017;45:e22.

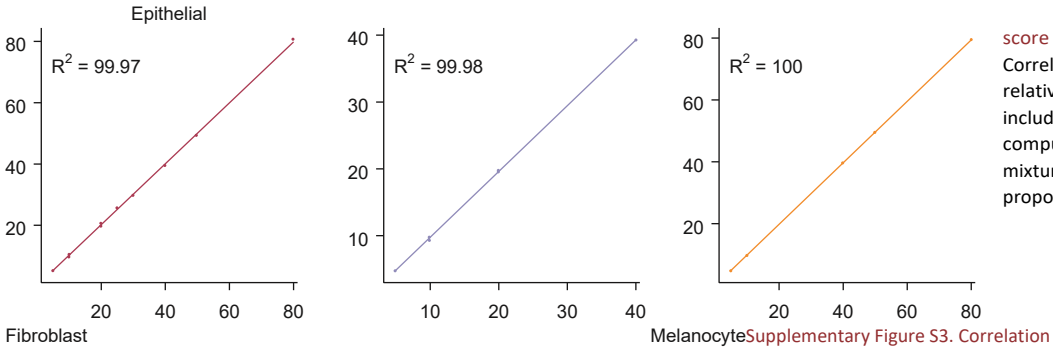


Supplementary Figure S1. Cell-type estimation in benign and dysplastic nevi and matching adjacent normal skin. (a) Heatmap of the CpG weights assigned to each cell type for the 609 probes included in the skin-specific reference library. Cell types included in the library include epithelial cells, fibroblasts, melanocytes, B cells, $CD4^{b}$ T cells, $CD8^{b}$ T cells, NK cells, monocytes, and neutrophils. Unsupervised hierarchical clustering of CpG loci and cell types was conducted utilizing Manhattan distance. (b) Distribution of the estimated relative proportion of each cell type represented in the reference library across adjacent normal skin, benign nevi, and dysplastic nevi, respectively. (c) Correlation between estimated relative proportion of melanocytes using H&E staining and using the reference library in nevi samples. CpG, 5^{0}-C-phosphate-G-3^{0}.

Supplementary Figure S2. Correlation Bcell CD4T CD8T

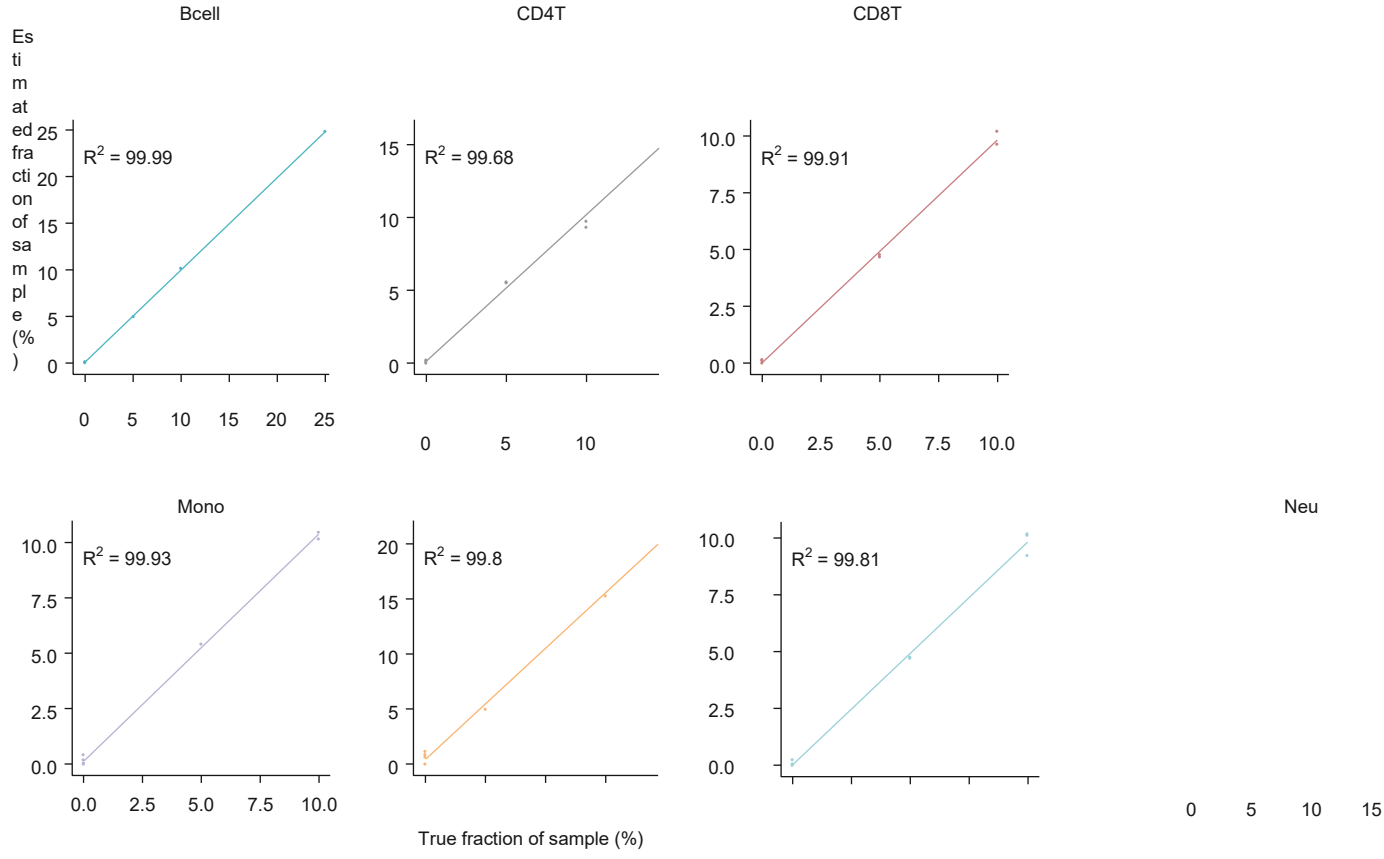
score of estimated immune cells present. Correlation between the estimated relative proportions of immune cells included in the reference library from in silico mixtures with the true relative proportions.

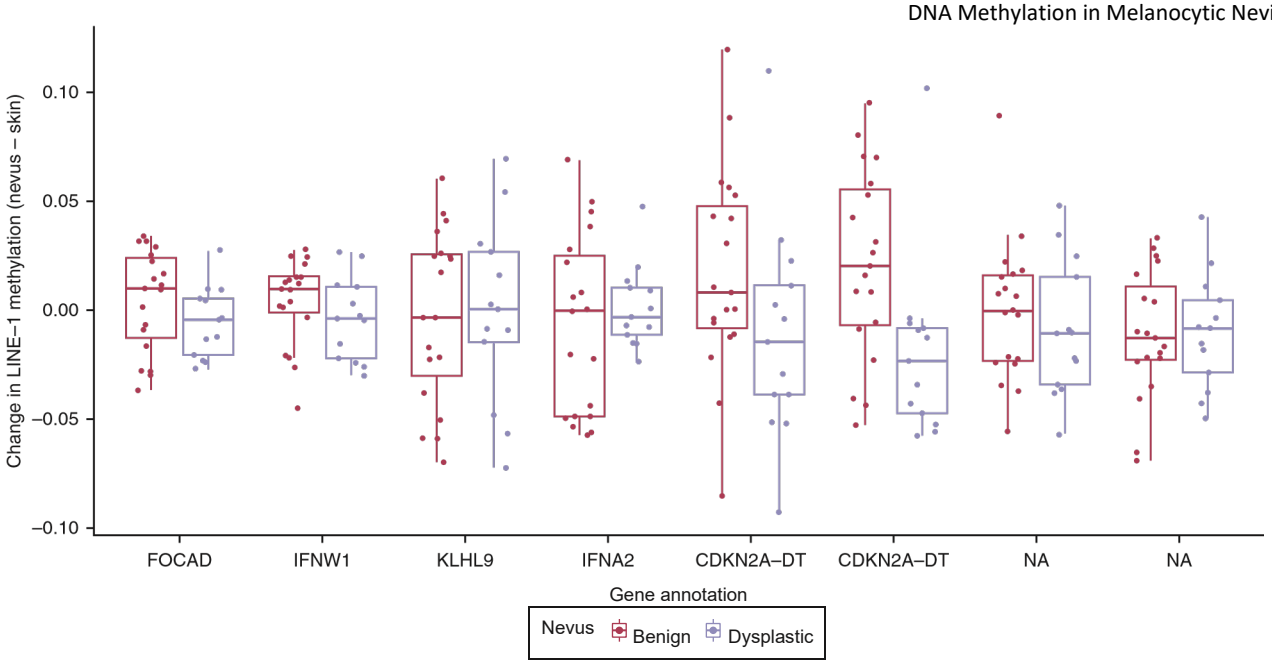




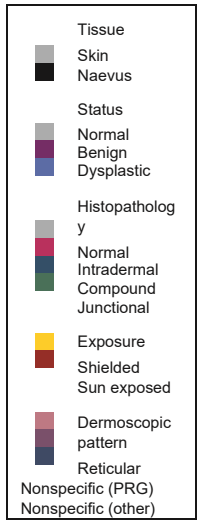
score of all cell types present.
Correlation between estimated relative proportions of all cell types included in the reference library from computationally generated artificial mixtures with the true relative proportions.

Supplementary Figure S3. Correlation



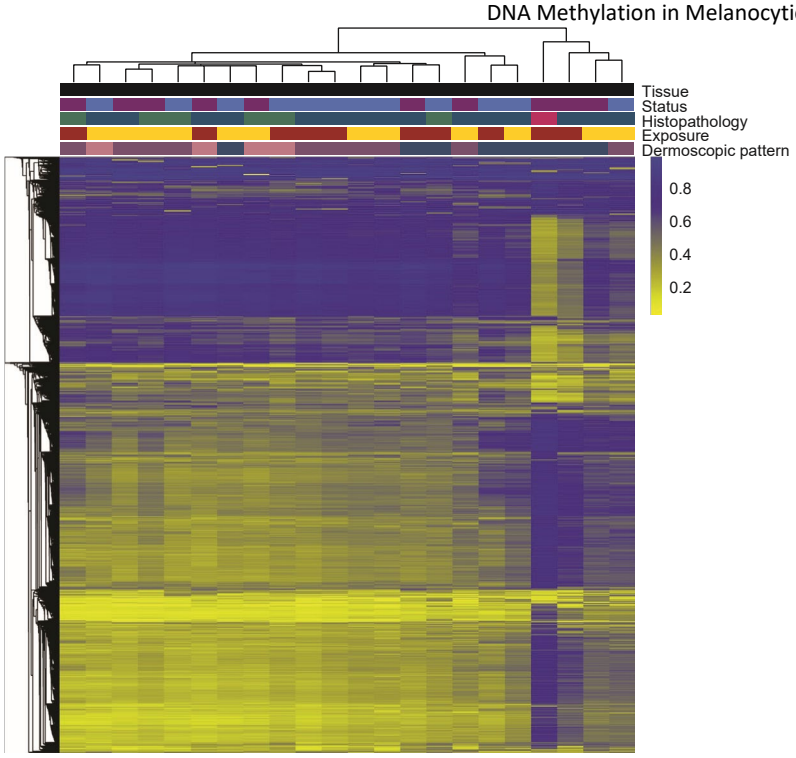


Supplementary Figure S4. Difference between inferred LINE-1 element methylation in the Chr9p21 region in nevi relative to that in paired adjacent normal skin. CDKN2A-DT refers to CDKN2A divergent transcript. NA refers to an inferred repeat element that does not map to a gene.



of the top 5,000 most variable loci across all samples for nevi classified as reticular and nonspecific on the

peripheral rim of globules.

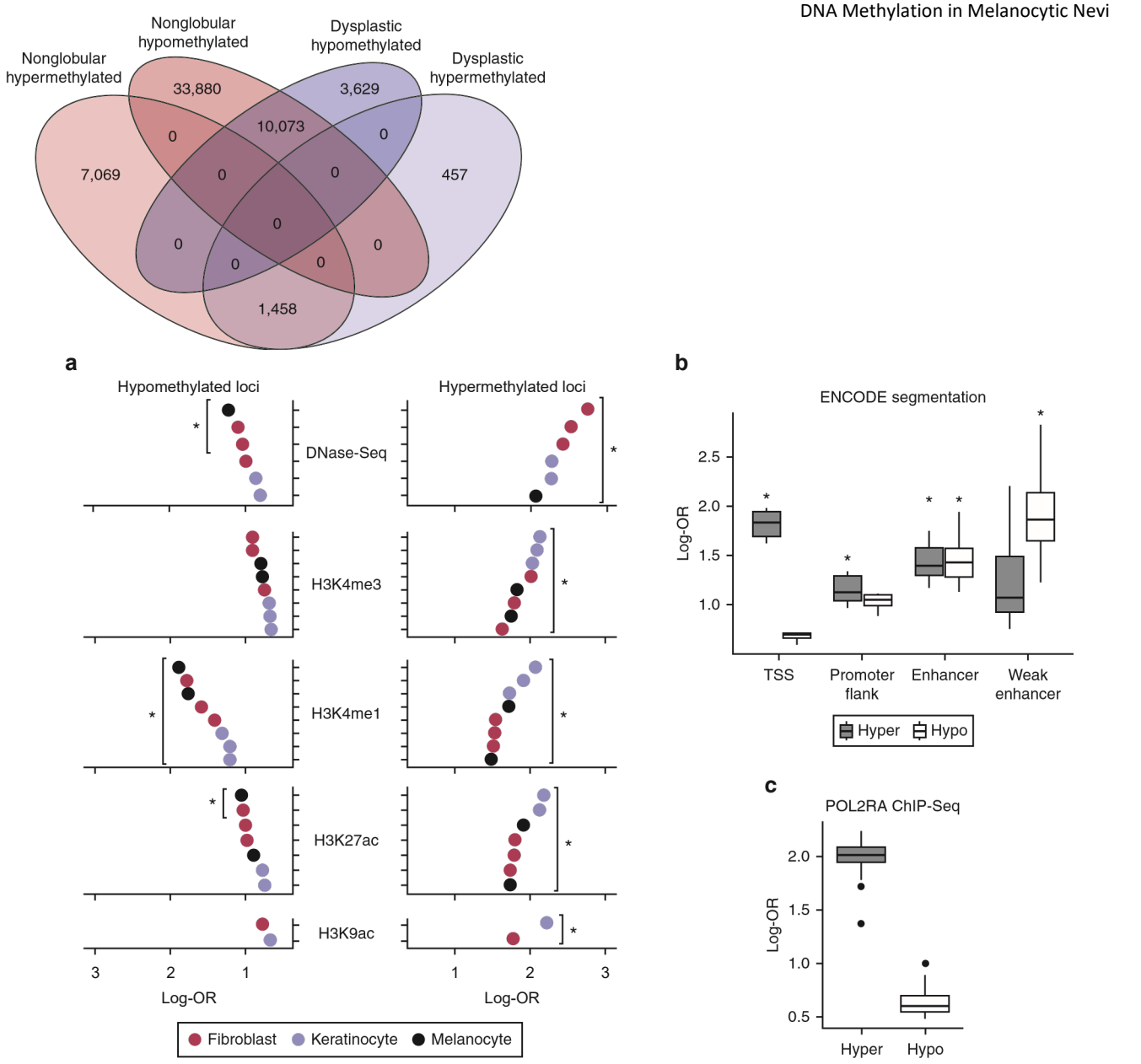


Supplementary Figure S5.
Unsupervised hierarchical clustering

basis of dermoscopic pattern. PRG,

6 19 8 8 23 11 7 22 1 4 4 3 24 9 15 2 13 13 14 10 12 23

Supplementary Figure S6. Overlap of identified differentially methylated loci (FDR < 0.05) in nonglobular nevi relative to those in adjacent normal skin and dysplastic nevi relative to those in adjacent normal skin. FDR, false discovery rate.



Supplementary Figure S7. Enrichment of differentially methylated loci in nonglobular nevi in functional genomic regions. (a) Log-OR enrichment of hypermethylated and hypomethylated CpG loci in fibroblast, keratinocyte, and melanocyte Roadmap epigenomic datasets (* ¼ Q < 0.05, Fisher's exact test). (b) Log-OR enrichment of differentially methylated loci in ENCODE genome segments in six cell types: GM12878, H1-hESC, HeLa-S3, HepG2, Huvec, and K562 (* ¼ Q < 0.05 across all cell types, Fisher's exact test). (c) Log-OR enrichment of differentially methylated loci in POL2RA ChIP-Seq peaks in 52 cell types from ENCODE. ChIP-Seq, chromatin immunoprecipitation sequencing; CpG, 5^0-C-phosphate-G-3^0; Dnase-Seq, DNase I hypersensitive sites sequencing; ENCODE, Encyclopedia of DNA Elements; TSS, transcription start site.

by DARTMOUTH COLLEGE
Bioinoc09o /2m 9/pu22tin. g Re20-22usDeo anw dnl disoatridebud tiofrom m isw strw ictw.lyw noorltdspeciermntiittficed , exceptforOpenAcce

Mixed Effects Machine Learning Models for Colon Cancer Metastasis Prediction using Spatially Localized Immuno-Oncology Markers

Joshua J. Levy[1], Carly A. Bobak

Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth Lebanon, NH 03756, USA
Email: joshua.j.levy@dartmouth.edu, carly.a.bobak.gr@dartmouth.edu

Mustafa Nasir-Moin, Eren M. Veziroglu, Scott M. Palisoul, Rachael E. Barney

EDIT, Department of Pathology and Laboratory Medicine, Dartmouth Hitchcock Medical Center Lebanon, NH 03756, USA
Email: mustafa.nasir-moin@dartmouth.edu, eren.m.veziroglu.med@dartmouth.edu, scott.m.palisoul@hitchcock.org, rachael.e.barney@hitchcock.org

Lucas A. Salas, Brock C. Christensen

Department of Epidemiology, Geisel School of Medicine at Dartmouth Lebanon, NH 03756, USA
Email: lucas.a.salas@dartmouth.edu, brock.c.christensen@dartmouth.edu

Gregory J. Tsongalis, Louis J. Vaickus

EDIT, Department of Pathology and Laboratory Medicine, Dartmouth Hitchcock Medical Center Lebanon, NH 03756, USA
Email: gregory.j.tsongalis@hitchcock.org, louis.j.vaickus@hitchcock.org

Spatially resolved characterization of the transcriptome and proteome promises to provide further clarity on cancer pathogenesis and etiology, which may inform future clinical practice through classifier development for clinical outcomes. However, batch effects may potentially obscure the ability of machine learning methods to derive complex associations within spatial omics data. Profiling thirty-five stage three colon cancer patients using the GeoMX Digital Spatial Profiler, we found that mixed-effects machine learning (MEML) methods[2] may provide utility for overcoming significant batch effects to communicate key and complex disease associations from spatial information. These results point to further exploration and application of MEML methods within the spatial omics algorithm development life cycle for clinical deployment.

[1] To whom correspondence should be addressed.

[2] Code available on GitHub at the following URL: https://github.com/jlevy44/MEML_Colon_DSP_METS. Data available on reasonable request due to privacy/ethical restrictions.

© 2021 Joshua Levy, Carly Bobak, Mustafa Nasir-Moin, Eren Veziroglu, Scott Palisoul, Rachael Barney, Lucas Salas, Brock Christensen, Gregory Tsongalis, Louis Vaickus. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

by DARTMOUTH COLLEGE Biooc09o /2m9/pu22tin. g Re20-22usDeo anwdnl disoatridebud tiofrnmiswstrwictw.lywnoorltdspecierntiittficed, except for Open Ace

Keywords: GeoMx Digital Spatial Profiler; Spatial Omics; Mixed Effects; Machine Learning; Boosting Methods; Tumor Metastasis; Colon Cancer

1. Introduction

Practicing pathologists routinely order immunohistochemical stains to assess the spatial distribution of important biomarkers for disease prognostication. Recently, higher resolution spatially resolved technologies have emerged. Nature Methods declared "Spatially Resolved Transcriptomics" technologies as the method of the year in 2020 [1]. Although multiplexed immunohistochemistry and laser capture microdissection approaches once served as popular methods for assessing spatially resolved omics information, these approaches are largely intractable for high throughput analysis because quantitative analysis is hindered by significant tissue distortion from chemical restaining processes, insufficient multiplexing, and prohibitive costs. New spatial omics technologies can multiplex far more markers through targeted cleavage of oligonucleotide tags [2–8]. For instance, the Nanostring GeoMX Digital Spatial Profiler (DSP) first utilizes multiple immunofluorescent (IF) antibody stains to highlight various cellular lineages, then uses UV-light to cleave attached oligos for quantification via the nCounter instrument or next-generation sequencing (NGS)[9,10]. A particular novelty of this technology is the ability to highlight regions of interest (ROI) through segmentation of the IF stains or selection of specific locations of interest. Informatics techniques developed for spatial data include estimation of: 1) which genes are spatially variable, 2) how spatial clustering patterns of expression may differ between cases and controls, 3) colocalization and coordination between cellular populations, accomplished using canonical protein markers or cellular deconvolution, and 4) integration with other imaging (e.g., morphology, architecture via hematoxylin and eosin stain, H&E) and single-cell multi-omics (e.g., RNASeq, ATACSeq; models of cell fate) which can further shed light on diagnosis and prognostication [11]. Prospective deployment of clinical decision aids can use low-cost tissue staining techniques (e.g., IHC) to profile protein markers deemed relevant through spatial omics analysis. However, existing approaches (e.g., differential expression and spatial autocorrelation) suffer from technical factors, including: reagent application, batch effects, and other forms of repeated measurements. These confounders have the potential to bias disease associations and preclude prospective classifier deployment. Compared to traditional penalized, generalized linear modeling approaches, algorithms that can use the complete multiplexed set of markers and their interactions (e.g., Random Forest models) may be best positioned to capture the disease pathogenesis yet are likely less resistant to batch effects. Thus, we need machine learning algorithms that can fully utilize spatial multiplexed information while being immune to batch effect corruption. In this study, we assess the potential for mixed-effects machine learning (MEML) methods to address this issue, capturing unintuitive spatial and multiplexed complexities, while respecting patient and batch level differences.

We assessed colorectal adenocarcinoma (CRC) cases, with or without nodal and/or distant metastasis using spatial omics technologies (**Figure 1A-B**). Prognosis / metastatic potential is well characterized based on presence of tumor invasion into distinct tissue layers (e.g., epithelium, lamina propria, submucosa, *muscularis propria*, pericolic fat, and serosa). Previous research has demonstrated the importance of specific somatic alterations (e.g., APC, MMR) and tumorinfiltrating lymphocytes (and their spatial coordination) as prognostic indicators [12–15]. Here,

by DARTMOUTH COLLEGE Bioinformatics on 09/20/22 using Redistribution subject to debut.ly/w/nl/disclaimer, except for Open Access

we used the DSP to collect spatial expression of 39 immune markers across 840 ROIs and applied MEML to characterize factors relating to different macro-architectural contexts (e.g., markers related to tissue interface with tumor, *inter*; within tumor, *intra*; away from tumor, *away*) and metastasis. This study will serve as the backbone of an expanded investigation of spatial immune markers descriptive of nodal and distant metastasis for Colon cancer patients.

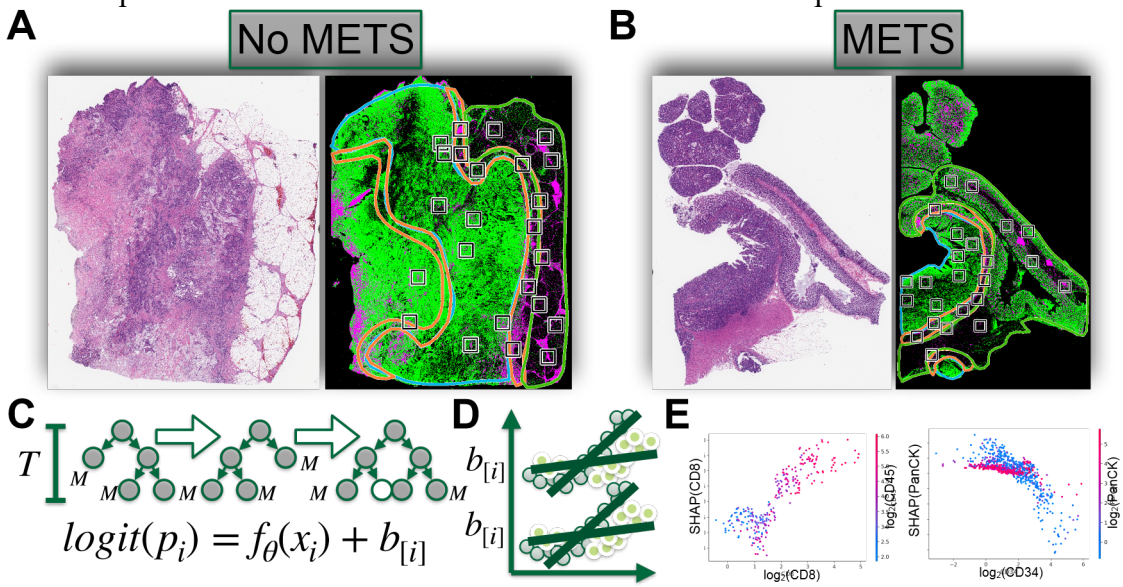


Figure 1: Overview of Experiment: Paired H&E and IF stains (PanCK stain, green; CD45 stain, pink) used to help select ROI (black squares) contained in macro-architectural contexts (outlined: blue for intra-tumoral, red for tumor interface, green for away from tumor) for: A) Non-metastatic patient, B) Metastatic patient; C) Tree boosting methods !"($%) combined with mixed effects modeling to adjust for D) patient/batch-level effects '[%] (e.g., interactions within nested observations, continuous scale) to yield E) disease associated interactions; SHAP dependence plots demonstrate how predictor (x-axis) covaries with another (colors) and impact on predictor importance, y-axis

2. Motivation for Comparison Study

First, we will provide a broad overview of common analytical techniques for spatial omics data, then discuss potential statistical oversights (e.g., modeling of statistical interactions and nonlinearities, batch effects/repeat measurements), to highlight the advantages of MEML methods.

3. Review of Prior Spatial Omics Analysis Methods

Within a slide, there is spatial variation in marker abundance, such that regions of interest (ROI) that are closer together may exhibit similar gene/protein expression [5,8,16]. Spatial autocorrelation is typically assessed using a Gaussian Process (GP), a mixed-effects model with a parameterized kernel. Included is an example of this model for multivariate normally distributed expression data:

$$*_{\%} \sim ,\text{-./0123}(4_{\%}, 6^{7})$$

$$4_{\%} = 9\$;;;^{\rightarrow}{:} + '_{[\%]} \qquad (1)\ ';^{\rightarrow} \sim ,\text{-./0123}{>}0;^{\rightarrow}, \Sigma A$$

Where ';$^{\rightarrow}$ is a vector of random intercepts, here observed at locations B$^{\rightarrow}$. The covariance matrix of the random effects, Σ, is a kernel used to test spatial hypotheses; for instance, that the correlation in expression between two spatially dependent markers drops off with distance C, $\Sigma_{\%,D} = 2^{7}E\$F$

by DARTMOUTH COLLEGE Bioc09o /2m9/pu22tin. g Re20-22usDe o an w d nl dis oatri debu d tio fron m is w str w ict w.ly w no orlt dspe cierm ntiitt ficed , ex ce pt for O pe n Ac ce

G−I7C>B%, BDA7J. The target of inference would be I, the degree to which the correlation in expression is retained as a function of the distance (C; e.g., L2-norm) between the ROIs K and L. Less parametric forms of this hypothesis can be tested through network autocorrelation and, more recently, through prediction models such as graph neural networks (GNN), where the latter is useful for increasing the resolution of the spatial expression patterns– e.g., predicting the current ROI's expression from that of neighboring (neighborhood .) cells/ROIs, L [17,18]: *% = 6>∑D∈0 2D*DA.

Other analyses expand the aforementioned GP models to estimate characteristic clustering patterns, enabling inter-slide comparison. Less parametric forms that take into account regional dependencies include graph neural networks, which can utilize aggregation/pooling mechanisms across nested observations to predict a single outcome while learning representative patterns. Here, establishing differential expression for specific markers and applying penalized regression procedures across a set of markers are simpler analytic approaches that compare spatially variable markers across slides and differentiate disease status. The latter is formalized below, where *% represents whether the case underwent metastasis, predicted from markers $;;;⃗:

$$*_{\%} \sim \mathrm{PKQ/1K23{>}1},\ F_{\%,\mathrm{STUV}}A$$

$$3/\mathrm{WKX{>}}F_{\%,\mathrm{STUV}}A = 9\$;;;\vec{}: \quad (2)$$

In this simplified scenario, spatial clustering patterns can be incorporated into this analysis by estimating stratified or interaction effects by assigned macro-architecture (e.g., *intra, inter, away*).

4. Motivation for Mixed Effects Machine Learning Approaches

The goal of this preliminary comparative study is to explore methods for classification of tissue macro-architecture/metastasis from spatially resolved expression patterns and to discuss opportunities for classifier development. In this section, we discuss deficiencies of existing methods. Machine learning (ML) techniques search for the ideal specification of interactions and nonlinear transformations that capture the relationship between the input and target of inference. For instance, classification and regression models (CART; e.g., Random Forest) utilize conditional decision splits to extract these associations [19]. Interpretation methods (e.g., Shapley Additive Feature Explanations, SHAP) can derive disease-associated predictors and interactions [20]. As omics datasets are usually high-dimensional and multi-collinear, CART methods excel in spotting relevant interactions. However, these algorithms assume that observations are independent and identically distributed (*i.i.d.*), which is invalid in the presence of repeated measurements, where statistical dependencies exist between observations. In the case of spatial omics modalities, an ROI may exist within several levels of nested dependencies (e.g., ROIs per section, sections per block and/or profiling batch). Multiple blocks may be extracted from different biopsies across space (sampling site) and time. These factors may potentially exacerbate the negative impact of technical effects on statistical inference. Even after documenting and controlling for batch effects, the degradation in performance resulting from the correlation between batch and outcome has not been thoroughly studied in this context [21,22]. Hierarchical modeling can leverage information across clusters of observations for efficient modeling of the fixed effects while respecting the dependency structure to avoid potentially biased inference. A random-intercept model, modeling different baseline measurements per cluster, follows the form of a linear mixed-effects model, which is a special case of the GP model [23]:

$$* \sim ,\text{-}./0123(4, 6^7) \qquad i \qquad i$$

$$\mu_i = \beta\vec{x_i} + b_{[i]}$$

by DARTMOUTH COLLEGE Biooc09o /2m9/pu22tin.g Re20-22usDeo anwdnldisoatridebud tiofrnm iswstrwictw.lywnoorltdspecierntiittficed, exceptforOpenAcce

(3) $';^{\rightarrow} \sim ,\text{-./0123}>;0^{\rightarrow}, \tau^{7}\text{IA}$

Given that spatial omics data may require algorithms that leverage nonlinear, complex high-order interactions offered by machine learning while respecting the dependency structure of the data, practitioners of spatial omics analyses, specifically those developing classifiers, should consider machine learning models which utilize mixed-effects modeling. Several such methods have been developed in the previous decade, being usually formulated as [16,24–26]:

(4) $*\ \sim ,\text{-./0123}(4\,,\,6^{7})$ $\quad i$ $\quad ';^{\rightarrow} \sim$ $\quad \mu_i = f_\theta(x_i) + b_{[i]}$ $\quad i$ $\quad ,\text{VNormal}>0;^{\rightarrow}, \Sigma\text{A}$

This modeling approach learns to estimate a random-effect component, $;b^{\rightarrow}$, that allows the estimation of a fixed-effect component robust to clustering by replacing a linear fixed-effects component, $9\$_{\%}$, with the fixed-effects machine learning model, $!_c(\$_{\%})$ (**Figure 1C-D**). Model fitting is typically accomplished by first removing the random effects from the outcome before fitting the fixed effects machine learning model, $Y_i^* = Y_i - b_{[i]}$, and then estimating or drawing fixed effects components through frequentist or Bayesian procedures. This procedure is repeated until convergence of both the fixed and random effects components. Potential drawbacks of separating the fixed and random effect components include having to omit the random-effect component during inference, where in many cases out of cluster prediction yields similar results to the fixed effects counterpart, and complexities in the random effects are not well captured (only random intercepts and slopes). Alternatively, clustering can be directly included in the model decision splits. For instance, for a given split, one can fit a generalized estimating equation model with a custom dependency structure to assess the ability to partition variable e with cutpoint 2: $\text{logit}>F_{\%Di}A = \beta_k + \beta_l m_{nopqr\,t} + \epsilon_{\%D}$, where $v_{\%D} \in .(0, \Sigma)$ and the correspondent Wald statistic on 9_l may inform the variable and value to split on [27]. Failing to consider the dependency structure of omics data for ML risks introducing bias by learning properties of the batch rather than the inference target. As omics data becomes increasingly available at the single-cell and spatial resolutions, mixed-effect modeling approaches may be optimally positioned to capture potential dependency structures/repeat sampling in the data.

5. Materials and Methods

6. Data Acquisition and Preprocessing

Under IRB approval, pathology reports and slides were reviewed for colon adenocarcinoma cases at Dartmouth Hitchcock Medical Center (DHMC) from 2016 to 2019. A total of 36 biopsies were initially selected for spatial profiling, half of them exhibiting local invasion but no nodal or distant metastasis (no METS) and the other half of them having undergone nodal and/or distant metastasis (METS). The cohort was restricted to stage 3 assignments under the pTNM staging system, which balances the impacts of local invasion, nodal and distant metastasis for prognostication. Cases were matched between the non-metastatic and metastatic groups based on tissue size (measured through connected component analysis of whole slide image), tumor grade, mismatch repair (MMR) dysregulation status, site of the tumor (e.g., left or right colon), age, and sex. Matching was achieved through conducting fisher's exact tests and two-sample t-tests after iterative resampling. For instance, average ages for the no METS and METS groups were 71±13 and 66±16

respectively (p=0.34) and stratified by sex, average ages for females and males were 68±18 and 69±13 respectively (p=0.96). A similar percentage of males comprised the METS and no METS groups (56% and 58% respectively, p=1.0). Tissue blocks were sectioned into 5-μm sections. Sections were stained with fluorescent-labeled antibodies (highlighting tumor (PanCK), immune cells (CD45), and nuclei (SYTO13)) which are covalently linked to photocleavable oligonucleotide tags targeting a panel of Immune Cell Profiling and Tumor Immune Environment markers. These sections were visualized using immunofluorescent (IF) images provided by the digital spatial profiling (DSP) instrument. Subsequent sections were stained with hematoxylin and eosin (H&E) and scanned using the Leica Aperio-AT2 scanner at 20x. IF whole slide images (WSI) were stored in TIFF format (16bit unsigned color channels; one channel per stain) and H&E images were stored in SVS format (8bit color channels). IF and H&E images were viewed simultaneously by the pathologist using the ASAP annotation software, which allowed the pathologist to delineate immune populations within macroarchitectural regions of interest (*intra* or within tumor, *inter* or at interface with tumor, *away* from tumor) on the IF images using polygonal/spline annotations. For each slide, a semi-automated workflow placed eight regions of interest (ROI; square grids of maximal spatial dimensions allowable by the GeoMx DSP instrument) within each annotation region (24 ROI per slide) after removing ROI with low immune cell density as quantified by the presence of CD45 stain. ROIs were mapped onto IF slides, uploaded to the DSP, and manually registered to the unmanipulated DSP IF images. ROI were placed onto DSP slides corresponding to the semi-autonomous placements and manually adjusted to the nearest appropriate region by a practicing pathologist if any ROI were misplaced (e.g., incorrect region or low immune cell density). Within each ROI, immune cells were segmented by using color thresholding and connected component analysis to deduce which areas in the ROI had overlapping nucleus and immune stains while excluding the tumor stain. The oligo tags within the immune cell areas were cleaved using ultraviolet light, collected and counted with nCounter to quantify protein expression within the immune cell regions of all ROI. Four slides were assessed at a time due to the DSP's batching mechanism. Additional cases were included/excluded due to tissue lifting after cover-slipping procedures (leading to case removal) or to balance subsequent batches, leading to inclusion of 35 cases (840 ROIs).

A total of 28 ROIs were removed after quality control procedures, retaining 812 ROIs. Within each batch, we separately performed and compared expression after ERCC normalization, normalization based on nuclei area and count, and normalization to housekeepers and IgG isotype controls. Using the raw data, we compared log2-transformed counts of housekeeper markers to each other (S6, Histone H3, GAPDH) and separately compared counts of IgG isotope controls (Rb IgG, Ms IgG1, Ms IgG2a), also looking for significant batch effects. We used this information to subselect markers that were consistent with other control markers and exhibited the lowest potential for batch effects (Housekeepers: GAPDH, Histone H3; IgG: Rb-IgG, Ms-IgG1), and normalized the expression of the other markers based on the geometric mean of these respective markers. A preliminary analysis demonstrated the lowest impact from technical factors when using IgG isotopes, so we normalized expression based on IgG isotope controls for the following experiments. ROIs were further labeled with positional coordinates, nuclei count and total area, MMR alteration status, age and sex, site of origin/metastasis, tumor grade, nodal and distant metastasis status, and macro-architectural region (*intra*, *inter*, *away*). A total of 36 protein markers were selected for analysis, not including housekeeping genes. We emphasize that the collected

by DARTMOUTH COLLEGE Bioc09o /2m 9/pu22tin. g Re20-22usD e o an w d nl dis oatri debud tio fron m is w str w ict w.ly w no orlt ds pe cier m ntiitt ficed , ex ce pt for O pe n Ac ce

data represents a preliminary analysis and does not represent the data for final classifier development. Nonetheless, the data collected may prove helpful for the comparison of different classification algorithms.

7. Experimental Design: Prediction Tasks and Modeling Approaches

We compared predictive performance of several mixed-effects machine learning (MEML), two fixed effects techniques (Random Forest, XGBoost) [28], and classical statistical methods. As input to the algorithms are 36 markers (ROI-level fixed effects), age and sex (patient-level fixed effects), batch (random effects), and ROI coordinates (kernel). First, we treated the patient as a random intercept and predicted whether ROI could be localized to the tumor interface (*macro*), estimating: a) out of cluster predictive accuracy (*macro-ooc*; tested on held out set of patients) and b) within batch accuracy (*macro-ws*; samples from the same batch were partitioned for testing). Then, using the profiling batch as a random intercept, we sought to establish whether, based on assigned macroarchitectural region, colon cancer metastasis could be predicted (*mets*); first, predicting *mets* regardless of assigned macro-architecture (*mets-overall*; including whether ROI was intra- or inter-tumoral as additional predictors). We then built classifiers to predict whether there was nodal or distant metastasis based on localization of spatial markers within distinct macro-architecture (*metsintra, mets-inter, mets-away*). Our primary comparison was to compare out-of-cluster sampling performance with within-cluster sampling for the macro tasks. Since predicting out-of-cluster leads to the exclusion of random effects, we expect differences between fixed effects machine learning models and their mixed effects counterparts to be marginal. Thus, for the *mets* tasks, we focused on comparison of predictions within batch, as we wanted to shift focus from development of a classifier which can be used on external datasets to one where interrogation of important fixed effects predictors could potentially yield metastasis-related factors worth follow up. We provide brief overviews of selected modeling approaches and their implementations (R v3.6, Python v3.7):

1. **Gaussian Process Boosters (GPBoost):** tree boosting algorithms are additive models which fit weak learners to the residual between the outcome and sum of previous trees. Covariance parameters that account for random intercepts and arbitrary covariance functionals are updated at each boosting step via iterative minimization of the model likelihood and risk functional by gradient/newton descent. GPBoost models can generalize to classification tasks by regressing on a continuous latent outcome, dichotomized to yield binary outcomes. We model patients/batches as grouped random intercepts [24]:

$$*_{\%} \sim \mathrm{PKQ/1K23}(1, \mathrm{F}_{\%})$$
$$3/\mathrm{WKX}(\mathrm{F}_{\%}) = !^{"}(\$_{\%}) + '_{[\%]}$$
$$!^{"}(\$_{\%}) = \sum^{x_D} !_D(\$_{\%}) \qquad (5) \qquad '_{[\%]} \sim .(0, \tau^7 z)$$

2. **GPBoost with Nested Coordinates (GPBoost-Coords):** Similar to GPBoost, with a random intercept for patient/tissue section and grouped within-patient, a spatial Gaussian Process with covariance function of exponential form which captures spatial dependence between ROI:

$$3/\mathrm{WKX}(\mathrm{F}_{\%}) = !^{"}(\$_{\%}) + '_{(l)[\%]} + '_{(7)[\%]}$$
$$'_{(l)[\%]} \sim .(0, \{7z)$$

by DARTMOUTH COLLEGE Bioinc09 o /2 m 9/ pu22 tin. g Re 20- 22us D e o an w d nl dis oatri debu d tio fron m is w str w ict w.ly w no orlt ds pe cierm ntiitt fice , ex ce pt for O pe n Ac ce

;';';;[′cluster k′;;;;;;;;;;;;;;;;;];(;;7;⃗) ∼ ,-./0123>0;⃗, Σ[cluster k]A

(6)

$$\Sigma_{[\text{cluster } k]} \quad \%,D = 2 * \text{Exp } G-I * C>B\%, BDAJ$$

3. **Semi-Parametric Mixed Effects Bayesian Additive Regression Trees (SP-BART):** Bayesian additive regression trees are similar to boosting in that decision trees are being successively estimated, this time in the form of particle Gibbs sampling of decision trees, with priors set over tree parameters (splitting structure, predictor selection, tree depth, residual variance) as regularization and Bayesian backfitting techniques to draw new trees based on the residual. This semi-parametric approach uses Markov Chain Monte Carlo procedures to draw additive trees as the fixed effects model components (*dbarts* [29]), with subsequent Hamiltonian Monte Carlo (HMC) draws via Stan, to update patient/batch-level random intercepts [30].
4. **Bayesian generalized linear mixed-effects model (BGLMM) with Horseshoe Prior**: Implementation of Bayesian logistic regression using the *brms* package [31], where model parameters/random intercepts are drawn using HMC, and horseshoe shrinkage prior with documented performance over Lasso priors to effectively remove irrelevant/collinear genes [32].
5. **Regularized BGLMM with Interactions (BGLMM-Int):** A hybrid approach between mixedeffects CART methods and generalized linear procedures utilizes SHAP to extract the top 10 salient interactions from the GPBoost model and adds proposed interactions (robust to clustering; **Figure 1E**) as terms in the BGLMM Horseshoe model, under the premise that these interactions help the GLMM recover any performance gap for their MEML counterparts [33].
6. **BGLMM with Gaussian Process (BGLMM-GP):** Similar to GPBoost-Coords, replacing the gradient boosting decision trees with a linear functional: $3/WKX(F\%) = \beta\$\% + '^{(l)}[\%] + '^{(7)}[\%]$.
7. **Random Forest (RF) and Extreme Gradient Boosting (XGBoost):** Fixed effects counterparts (FEML) to aforementioned MEML methods, fit using *scikit-learn* and *xgboost* packages [28,34]. To assess the predictive performance of each modeling approach, we performed ten-fold crossvalidation and bootstrapped performance statistics (C-Statistic and Log-Loss) through iterative resampling of the held-out validation sets of each of the cross-validation folds before averaging. We focused on sensible/coarse selection of hyperparameters. Salient predictors for the CART methods were extracted using SHAP to estimate the local importance of each predictor and summed to form global attributions [35]. Interaction effects were extracted using SHAP's *TreeExplainer*[20]. After selecting predictors via Horseshoe Lasso variable selection and standardizing, posterior draws of unpenalized BGLMM model parameters were used to assess and visualize significant predictors.

by DARTMOUTH COLLEGE Bioinformatics on 09/20-22using Re Dedo and wdnl distributiondebud tio from is wstr wict w.ly w no orlt dspe cierm ntiitt ficed , ex ce pt for O pe n Ac ce

8. Results

9. Macro: Inter-Tumoral Prediction

The ability to characterize ROI out-of-batch/sample (*OOS*) indicates the potential to assess new cases prospectively for macro-architectural characteristics. Here, we expected within-sample (*WS*) performance to exceed predictions made on held-out patients. As expected, the prediction of whether an immune cell ROI was embedded in the interface with tumor exhibited greater performance within-sample across all approaches. MEML methods, taken as a group, slightly outperformed fixed effects approaches for macro-architectural prediction (**Table 1**). Surprisingly, generalized linear mixed effects approaches outperformed both the fixed effects and MEML methods, with slightly greater performance assigned to the BGLMM model, which utilized interactions detected from the GPBoost model (**Table 1**). Adding spatial dependence to the model in the form of Gaussian Processes did not improve predictive performance for GPBoost and BGLMM.

Top predictors for tumor interface macro-architecture between the GPBoost model and BGLMM model were concordant (e.g., PanCk. CD127, CTLA4, CD44) and of similar directionality (**Figure 2A,C**). IgG isotope technical factor Ms-IgG2A was found to be in the top half of predictor for GPBoost, though effects were largely negligible and undetected by the BGLMM. Some of the interactions extracted from the GPBoost model were found to be some of the top predictors for the BGLMM-Int model (e.g., PanCk:CD68, PanCk:CD34) (**Figure 2B-C; Table 2**), indicating the relevance of MEML extracted interactions in capturing macro-architectural associations.

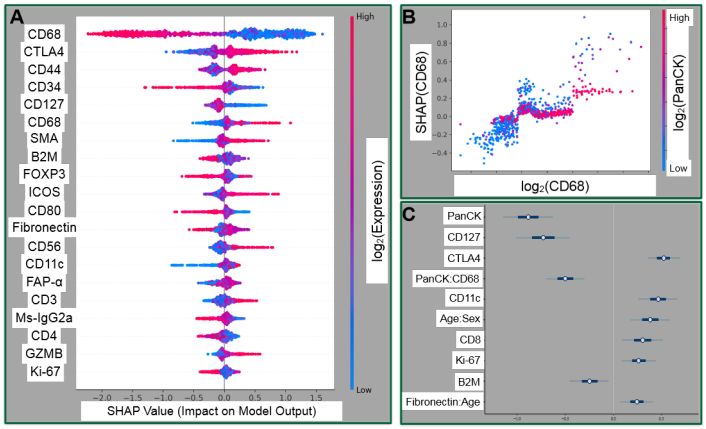


Figure 2: Macro Predictions: A) SHAP summary plot of top GPBoost terms; each point represents value of specific predictor for ROI; color indicates predictor's value; positive SHAP value (x-axis) indicates how related to tumor interface; B) GPBoost SHAP partial dependency of CD68 interaction with PanCK; C) Posterior distributions of unpenalized BGLMM-Int predictors (thick and thin bars represent 50% and 90% credible intervals respectively)

Table 1: Bootstrapped model performance comparison, using C-Statistic/Area Under the Curve (AUC) and Log-Loss

		Fixed Effects		MEML			Bayesian Generalized Linear Mixed Models	
Task	**AUC ± SE**	**RF**	**XGBoost**	**GPBoost**	**GPBoostCoords**	**SP-BART**	**BGLMM**	**BGLMM-BGLMMInt GP**

by DARTMOUTH COLLEGE Bioinoc09 o /2 m 9/ pu22 tin. g Re 20- 22us D e o an w d nl dis oatri debu d tio fron m is w str w ict w.ly w no orlt dspe cierm ntiitt fied , ex ce pt for O pe n Ac ce

Task		Log-Loss ± SE	RF	XGBoost	GPBoost	GPBoost-Coords	SP-BART	BGLMM	BGLMM-Int	BGLMM-GP
Macro		OOS	0.759±0.01	0.747±0.01	0.752±0.01	0.747±0.01	0.772±0.01	0.778±0.01	**0.785±0.009**	0.777±0.01
		WS	0.781±0.009	0.773±0.01	0.782±0.009	0.78±0.009	0.784±0.009	0.788±0.009	**0.802±0.009**	0.786±0.01
METS		Overall	0.909±0.006	0.951±0.004	**0.971±0.003**	n/a	0.897±0.006	0.852±0.008	0.896±0.006	n/a
		Intra	0.834±0.014	0.849±0.014	**0.89±0.011**	n/a	0.867±0.012	0.848±0.013	0.877±0.012	n/a
		Inter	0.881±0.013	0.866±0.013	**0.899±0.012**	n/a	0.874±0.012	0.836±0.014	0.881±0.012	n/a
		Away	0.827±0.014	0.842±0.014	**0.895±0.011**	n/a	0.887±0.011	0.885±0.012	0.885±0.011	n/a

Task		Log-Loss ± SE	RF	XGBoost	GPBoost	GPBoost-Coords	SP-BART	BGLMM	BGLMM-Int	BGLMM-GP
Macro		OOS	0.546±0.008	0.747±0.025	0.572±0.013	0.578±0.014	0.547±0.013	0.535±0.01	**0.528±0.011**	0.536±0.01
		WS	0.542±0.017	0.701±0.025	0.535±0.013	0.538±0.013	0.526±0.011	0.522±0.01	**0.508±0.01**	0.524±0.01
METS		Overall	0.485±0.005	0.284±0.013	**0.248±0.007**	n/a	0.409±0.014	0.455±0.01	0.393±0.01	n/a
		Intra	0.544±0.009	0.509±0.027	**0.422±0.014**	n/a	0.454±0.021	0.479±0.017	0.444±0.017	n/a
		Inter	0.507±0.009	0.499±0.032	**0.412±0.017**	n/a	0.457±0.026	0.484±0.017	0.429±0.019	n/a
		Away	0.559±0.01	0.559±0.031	**0.428±0.017**	n/a	0.442±0.023	0.441±0.015	0.434±0.015	n/a

10. METS: Nodal and Distant Metastasis Prediction

The greatest performance differences between fixed effects methods and MEML methods were noticed across all metastasis prediction tasks, where the GPBoost method obtained as little as a 2% boost in C-statistic to as much as a 9% performance gain versus RF and XGBoost. Within batch, it was easier to detect metastasis when all macro-architectural regions were utilized, while performance was roughly equivalent for each of the subregions (**Table 1**). Since performance of GPBoost exceeded that of all other approaches, the model may have learned significant interactions between genes that were invariant to batch differences. Using the detected interactions to supplement the BGLMM model allowed the BGLMM interaction model to vastly exceed the performance of its non-interaction counterpart and obtain performance competitive with GPBoost (**Table 1**). Across all tasks, the BGLMM model mostly outperformed its CART counterpart BART. GPBoost and BGLMM derived some concordant predictors (e.g., CD20, CD14), while others (e.g., FOXP3) were not found in the BGLMM model (**Figure 3A,C**). IgG isotope technical factor Ms-IgG1 was found to be in the top half of GPBoost predictors; effects were largely negligible and undetected by the BGLMM. The GPBoost model was also able to detect important interactions, (e.g., effect modification of markers CD20 and Fibronectin by age) (**Figure 3B-C; Table 2**) [36].

Table 2: Top terms from BGLMM-Int models; GPBoost-extracted interactions are emphasized in bold

Top 7 Ranked BGLMM-Int Terms					
Macro		METS			
OOS	WS	Overall	Intra	Inter	Away
PanCk	PanCk	**CD20:Age**	Beta-2-microglobulin (B2M)	**CD68:CTLA4**	CD25
CTLA4	CD127	CD14	CD45	CD11c	CD44
CD68:PanCk	CTLA4	Age	CD8	CTLA4	FAP-α
CD34	**PanCk:CD68**	CD68	FOXP3	CD14	Fibronectin
PanCk:CD27	CD44	CD11c	**B2M:FOXP3**	GZMB	CD68
CD68	**PanCk:CD34**	**Fibronectin:Age**	Age	CD68	Age
CD27	CD68	Ki-67	**CD8:CD45**	**CD14:CD66b**	**FAP-α:CD163**
	CD56	**Age:Sex**		Age	CD163
	CD34	CD8		**CD14:CTLA4**	PanCk
	PanCk:CD56	Beta-2-microglobulin (B2M)		CD66b	**PanCk:Age**
		CD20		**GZMB:Age**	
		B2M:Age			

Fibronectin
Sex

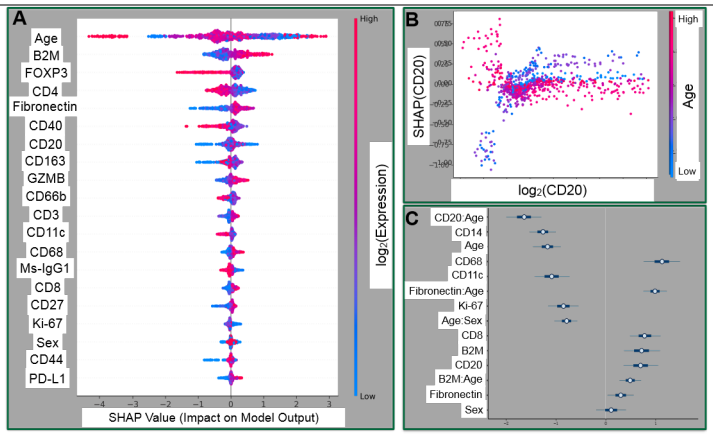


Figure 3: METS Predictions: A) SHAP summary plot of top GPBoost terms; each point represents value of specific predictor for ROI; color indicates predictor's value; positive SHAP value (x-axis) indicates how related to tumor interface; B) GPBoost SHAP partial dependency of CD20 interaction with age; C) Posterior distributions of unpenalized BGLMM-Int predictors (thick and thin bars represent 50% and 90% credible intervals respectively)

11. Discussion

Spatial localization of gene and protein markers provides additional resolution on disease pathogenesis and etiology, which may soon lead to more informative and promising diagnostic and prognostic decision aids. While failing to account for repeated measurements in this setting may threaten to preclude classifier deployment, incorporating mixed-effects modeling into machine learning model development may derive potential additional clinically relevant disease predictors due to nonlinear interaction modeling while accounting for batch effects. In this study, we performed a preliminary assessment and comparison of different generalized linear and CART models for deployment as classifiers to assess macro-architectural characteristics and potential for undergoing nodal or distant metastasis. Results from our study indicate that classifier development for spatial omics technologies such as GeoMX DSP or Visium's Spatial Transcriptomics may benefit by taking into account repeated measurements on the case and batch level. At minimum, mixed effects classification algorithms in spatial omics studies may detect complex gene-gene and genephenotype interactions that can be applied towards clinical findings (e.g., reportable risk estimates). For instance, the BGLMM-Int model indicated that the presence of CD20 (B-cells) was more predictive of metastatic potential in older individuals compared to younger individuals, while the presence of Fibronectin indicated the opposite effect (more predictive of metastasis in older individuals). We plan to further investigate the clinical utility of such identified interaction effects.

The added advantage of employing an MEML model is the ability to extract an interaction that the generalized linear approach can utilize, which may outperform generalized linear approaches without interactions, often unfairly reported in algorithm comparison studies. In this study, using this hybrid approach of informing interactions through MEML methods yielded BGLMM models that could outperform fixed effect and MEML methods due to simpler decision boundaries (line instead of stepwise function) and less biased sampling of the posterior of the parameters.

by DARTMOUTH COLLEGE Biooc09o /2m9/pu22tin. g Re20-22usDeo anwdnl disoatridebud tiofrom mis wstrw ictw.lyw noorltdspecierm ntiittficed, except for Open Acce

We acknowledge a few study limitations. Namely, we performed very coarse hyperparameter searches and other notable MEML methods [27,29,37–40] were omitted due to overlapping design. Additional algorithmic fine-tuning may demonstrate superior performance that is different from that reported in this work. The role of batch imbalance (association of batch with an outcome) has been well explored as it degrades the model performance in bulk expression datasets. However, its effects, particularly on intra-batch normalization and estimation of random effects, are unclear with the presented data and demand further exploration. We chose not to include batch effect correction using ComBat to limit the number of comparisons[41]. Regarding the batch effects, other than modeling Gaussian Processes nested within random intercepts, we did not include any random slopes in any of the MEML and BGLMM models, nor did we consider multiple hierarchy levels simultaneously (random intercept for batch and patient/biopsy). METS predictions were made within-batch and are not an accurate reflection of deployment on new samples. We did not incorporate time from initial biopsy to development of local or distant METS, though we plan to comment on the temporality of such associations in follow-up clinical findings. Projective predictive feature selection can help select more relevant BGLMM predictors [42,43]. We neither performed nor compared findings from univariable analyses with multivariable analyses, which is outside the study scope. We also noticed that there was no gain in performance from using the ROI spatial coordinates, which may have been an artifact of implementational difficulties, the low sample size of ROIs and/or selection/learning of improper kernels. Spatial autocorrelation is often studied in spatial omics analyses, and, in the future, we expect to include an in-depth assessment of spatial effects.

In the future, we aim to refocus these findings for a more clinical setting after further algorithmic fine-tuning and orthogonally validation of our research results using immunohistochemistry (IHC). For prospective deployment of these technologies, we envision IHC serving as a low-cost alternative and potentially less encumbered by batch effects to rapidly spatially assess markers that have demonstrated utility for identification of colon metastasis. Such data modalities require standardization of collection/analysis processes (e.g., automated segmentation of macroarchitecture (*intra, inter, away*) with deep learning algorithms, assessment of potential for bias from semi-subjective/automated selection of ROI) [44]. Application of chemical reagents demands further adjustment to ensure meaningful deployment. Regarding further methods development, aggregation across macro-architectural regions may be explored using graph neural networks, which can pool spatial omics information across these regions to form a succinct vectorial representation of patient's overall profile that can be integrated with H&E imaging modalities. Simulation studies may provide further validation of the utility of MEML methods for the assessment of spatial omics technologies.

12. Conclusion

In this work, we demonstrated the potential utility of MEML methods for the assessment of spatial protein markers collected using the Nanostring GeoMx DSP. We illustrated that MEML methods obtain superior within-batch performance for Colon metastasis prediction versus fixed effects and generalized linear methods. Furthermore, MEML methods, when combined with generalized linear modeling, may lead to clearer communication of significant spatial disease associations in clinical research studies through the extraction of key complex interactions. Further standardization of data

by DARTMOUTH COLLEGE Biooc09o /2m 9/pu22tin. g Re20-22usDeo andwdnl disoatridebud tiofrmonm iswstrwictw.lywnoorltdspecierntittficed, exceptforOpenAcce

collection, normalization, and analysis are potential opportunities to consider spatial omics machine learning technologies for meaningful clinical deployment.

13. Acknowledgements

We acknowledge James O'Malley, Robert Frost, Prajan Divakar, and Christian Haudenschild for leadership, support and discussion, and support of the Laboratory for Clinical Genomics and Advanced Technology in the Department of Pathology and Laboratory Medicine of Dartmouth Hitchcock Health System and Pathology Shared Resource, at the Norris Cotton Cancer Center at Dartmouth with NCI Cancer Center Support Grant 5P30 CA023108-37. Additional funding support from Burroughs Wellcome Fund Big Data in the Life Sciences training grant, Dartmouth College Neukom Institute for Computational Science CompX awards, and NIH grant R01CA216265.

14. References

1. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat Methods* **18**, 9–14 (2021).
2. Ziemys, A. *et al.* Integration of Digital Pathologic and Transcriptomic Analyses Connects Tumor-Infiltrating Lymphocyte Spatial Density With Clinical Response to BRAF Inhibitors. *Front. Oncol.* **10**, (2020).
3. Jakubowska, K., Koda, M., Kisielewski, W., Kańczuga-Koda, L. & Famulski, W. Tumor-infiltrating lymphocytes in primary tumors of colorectal cancer and their metastases. *Exp Ther Med* **18**, 4904–4912 (2019).
4. P, D. *et al.* Advances in mixed cell deconvolution enable quantification of cell types in spatially-resolved gene expression data. (2020) doi:10.1101/2020.08.04.235168.
5. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods* **17**, 193–200 (2020).
6. Cang, Z. & Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications* **11**, 2084 (2020).
7. de Vries, N. L., Mahfouz, A., Koning, F. & de Miranda, N. F. C. C. Unraveling the Complexity of the Cancer Microenvironment With Multidimensional Genomic and Cytometric Technologies. *Front. Oncol.* **10**, (2020).
8. Zhang, M. *et al.* Spatial molecular profiling: platforms, applications and analysis tools. *Brief Bioinform* doi:10.1093/bib/bbaa145.
9. Delorey, T. M. *et al.* COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature* **595**, 107–113 (2021).
10. McNamara, K. L. *et al.* Spatial proteomic characterization of HER2-positive breast tumors through neoadjuvant therapy predicts response. *Nat Cancer* **2**, 400–413 (2021).
11. Foster, D. S. *et al.* Integrated spatial multi-omics reveals fibroblast fate during tissue repair. *bioRxiv* 2021.04.02.437928 (2021) doi:10.1101/2021.04.02.437928.
12. Idos, G. E. *et al.* The Prognostic Implications of Tumor Infiltrating Lymphocytes in Colorectal Cancer: A Systematic Review and Meta-Analysis. *Scientific Reports* **10**, 3360 (2020).
13. Lazarus, J. *et al.* Spatial and phenotypic immune profiling of metastatic colon cancer. *JCI Insight* **3**,.
14. Nearchou, I. P. *et al.* Automated Analysis of Lymphocytic Infiltration, Tumor Budding, and Their Spatial Relationship Improves Prognostic Accuracy in Colorectal Cancer. *Cancer Immunol Res* **7**, 609–620 (2019).
15. Nearchou, I. P. *et al.* Spatial immune profiling of the colorectal tumor microenvironment predicts good outcome in stage II patients. *npj Digital Medicine* **3**, 1–10 (2020).
16. Sigrist, F. Gaussian Process Boosting. *arXiv:2004.02653 [cs, stat]* (2021).
17. Song, Q. & Su, J. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in Bioinformatics* (2021) doi:10.1093/bib/bbaa414.
18. Yuan, Y. & Bar-Joseph, Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biology* **21**, 300 (2020).
19. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
20. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56–67 (2020).
21. Parker, H. S. & Leek, J. T. The practical effect of batch on genomic prediction. *Stat Appl Genet Mol Biol* **11**, Article-10 (2012).
22. Soneson, C., Gerster, S. & Delorenzi, M. Batch Effect Confounding Leads to Strong Bias in Performance Estimates Obtained by Cross-Validation. *PLOS ONE* **9**, e100335 (2014).
23. McElreath, R. *Statistical rethinking: A Bayesian course with examples in R and Stan*. (CRC press, 2020).
24. Sigrist, F. Latent Gaussian Model Boosting. *arXiv:2105.08966 [cs, stat]* (2021).
25. Hajjem, A., Bellavance, F. & Larocque, D. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* **84**, 1313–1328 (2014).
26. Chipman, H. A., George, E. I. & McCulloch, R. E. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **4**, 266–298 (2010).
27. Repeated measures random forests (RMRF): Identifying factors associated with nocturnal hypoglycemia. doi:10.1111/biom.13284.
28. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). doi:10.1145/2939672.2939785.
29. Kapelner, A. & Bleich, J. bartMachine: Machine Learning with Bayesian Additive Regression Trees. *arXiv:1312.2171 [cs, stat]* (2014).
30. Carpenter, B. *et al. Stan: A Probabilistic Programming Language*. vol. 76 (2017).
31. Bürkner, P.-C. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* **80**, 1–28 (2017).
32. Carvalho, C. M., Polson, N. G. & Scott, J. G. Handling Sparsity via the Horseshoe. in *Artificial Intelligence and Statistics* 73–80 (PMLR, 2009).
33. Levy, J. J. & O'Malley, A. J. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC Medical Research Methodology* **20**, 171 (2020).
34. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
35. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
36. Levy, J., Lebeaux, R., Christensen, B., Tosteson, T. D. & Bryan, Y. Journey across epidemiology's third variables: an anesthesiologist's guide for successfully navigating confounding, mediation, and effect modification. *Regional Anesthesia & Pain Medicine* (2021) doi:rapm-2020-101984.
37. Speiser, J. L. *et al.* BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes. *Chemometrics and Intelligent Laboratory Systems* **185**, 122–134 (2019).
38. Pellagatti, M., Masci, C., Ieva, F. & Paganoni, A. M. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **14**, 241–257 (2021).
39. Lin, S. & Luo, W. A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes. *Multivariate Behavioral Research* **54**, 578–592 (2019).
40. Capitaine, L., Genuer, R. & Thiébaut, R. Random forests for high-dimensional longitudinal data. *Stat Methods Med Res* **30**, 166–184 (2021).
41. Chen, C. *et al.* Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLOS ONE* **6**, e17238 (2011).
42. Piironen, J., Paasiniemi, M. & Vehtari, A. Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics* **14**, 2155–2197 (2020).
43. Bartonicek, A., Wickham, S. R., Pat, N. & Conner, T. S. The value of Bayesian predictive projection for variable selection: an example of selecting lifestyle predictors of young adult well-being. *BMC Public Health* **21**, 695 (2021).
44. Levy, J., Haudenschild, C., Barwick, C., Christensen, B. & Vaickus, L. Topological Feature Extraction and Visualization of Whole Slide Images using Graph Neural Networks. *Pac Symp Biocomput* 285–296 (2021) doi:10.1101/2020.08.01.231639.

ARTICLE NAVIGATION

POSTER PRESENTATIONS - PROFFERED ABSTRACTS | JUNE 15 2022

Abstract 1212: ExTIME: Extended tumor immune microenvironment cell mixture deconvolution using DNA methylation and a novel tumor-site-speci c hierarchical approach

Ze Zhang; Brock C. Christensen; Lucas A. Salas

Check for updates

Author & Article Information

Cancer Res (2022) 82 (12_Supplement): 1212. https://doi.org/10.1158/1538-7445.AM2022-1212

Abstract

Background: The solid tumor microenvironment is heterogeneous and varies in composition by tumor type. Previous gene expression and DNA methylation deconvolution approaches for tumor micro-environment have had some success for major cell types. However, existing methods lack specificity to tumor type and detailed cell types. We developed 21 tumor-specific DNA methylationbased libraries. We employed a novel hierarchical approach in 3 major tumor microenvironment components (tumor, angiogenic, immune) to profile 17 cell types (see methods below).

Methods: DNA methylation data on tumor samples (n=6183) and normal control samples (n=689) for 21 tumor sites were downloaded from GEO and TCGA to develop tumor-type-specific libraries. The top 1000 most informative differentially methylated CpG (DMC) sites were identified using InfiniumPurify for 21 tumor types to project tumor cell proportion. Epithelial, endothelial, and stromal cell samples were used to identify DMCs to profile the cells in the angiogenic environment. Basophil, eosinophil, neutrophil, dendritic cell, monocyte, B naïve, B memory, CD4T naïve, CD4T memory, CD8T naïve, CD8T memory, T regulatory, and natural killer cells were used to identify DMCs to deconvolve the immune environment. In conjunction with the constrained projection/quadratic programming approach, a novel hierarchical approach was employed with six layers and 12 libraries per tumor type to project cell proportions in first, tumor, second, angiogenic, and third, immune micro-environments. The method was validated using purified samples and experimental artificial mixtures.

Results: 12 libraries were developed per tumor site to deconvolve 17 cell types in 21 tumors. A preliminary application of the method on TCGA data investigating the association between angiogenic cells and survival revealed worse survival outcomes with a higher proportion of angiogenic cell proportions in BLCA ($p<0.01$) and HNSC ($p=0.02$), a higher endothelial cell proportion in CESC ($p=0.04$), a higher epithelial cell proportion in COAD ($p=0.02$), a lower endothelial proportion in KIRC ($p<0.01$), and a lower epithelial proportion in LUAD ($p=0.04$). Further analyses will be done to investigate the angiogenic and immune microenvironments with prognosis across tumor sites.

Conclusion: We developed a DNA methylation-based algorithm, ExTIME, to estimate cell proportions in the tumor microenvironments. This novel approach increased the specificity and accuracy of cell projection by employing a tumor-site-specific hierarchical model. Furthermore, the ExTIME profiles the tumor microenvironment to the most granular level compared to the existing methods. ExTIME's capability of depicting the cellular composition in tumors promises a better understanding of the cell heterogeneity and its relationship with prognosis across cancers.

Citation Format: Ze Zhang, Brock C. Christensen, Lucas A. Salas. ExTIME: Extended tumor immune micro-environment cell mixture deconvolution using DNA methylation and a novel tumorsite-specific hierarchical approach [abstract]. In: Proceedings of the American Association for Cancer Research Annual Meeting 2022; 2022 Apr 8-13. Philadelphia (PA): AACR; Cancer Res 2022;82(12_Suppl):Abstract nr 1212.

©2022 American Association for Cancer Research

CANCER RESEARCH COMMUNICATIONS

AACR American Association for Cancer Research

RESEARCH ARTICLE OPEN https://doi.org/10.1158/2767-9764.CRC-21-0088 ACCESS

Phase I Study of High-Dose L-Methylfolate in Combination with Temozolomide and Bevacizumab in Recurrent *IDH* Wild-Type High-Grade Glioma

Lucas A. Salas[1], Thomas G. Stewart[2], Bret C. Mobley[3], Chengwei Peng[4], Jing Liu[5], Sudan N. Loganathan[5], Jialiang Wang[5], Yanjun Ma[6], Mitchel S. Berger[7], Devin Absher[8], Yang Hu[9], Paul L. Moots[10], Brock C. Christensen[1,12], and Stephen W. Clark[10,11,13,14]

ABSTRACT

Isocitrate dehydrogenase (*IDH*) mutations in low-grade gliomas (LGG) result in improved survival and DNA hypermethylation compared with *IDH* wild-type LGGs. *IDH*-mutant LGGs become hypomethylated during progression. It is uncertain whether methylation changes occur during *IDH* wild-type GBM progression and whether the methylome can be reprogrammed. This phase I study evaluated the safety, tolerability, efficacy, and methylome changes after l-methylfolate (LMF) treatment, in combination with temozolomide and bevacizumab in patients with recurrent high-grade glioma. Fourteen patients total, 13 with GBM, one with anaplastic astrocytoma, all *IDH* wild-type were enrolled in the study. All patients received LMF at either 15, 30, 60, or 90 mg daily plus temozolomide (75 mg/m^2 5 days per month) and bevacizumab (10 mg/kg every two weeks).No MTD was identified. LMF-treated patients had median overall survival of 9.5 months [95% confidence interval (CI), 9.1–35.4] comparable with bevacizumab historical control 8.6 months (95% CI, 6.8–10.8). Six patients treated with LMF survived more than 650 days. Across all treatment doses, the most adverse events were diarrhea (7%, 1 patient, grade 2), reflux (7%, 1 patient, grade 2), and dysgeusia (7%, 1 patient, grade 2). In the six brains donated at death, there was a 25% increase in DNA methylated CpGs compared with the paired initial tumor. LMF in combination with temozolomide and bevacizumab was well tolerated in patients with recurrent *IDH* wild-type high-grade glioma. This small study did not establish a superior efficacy with addition of LMF compared with standard bevacizumab therapy; however, this study did show methylome reprogramming in high-grade glioma.

Significance: Glioblastoma (GBM) is a primary brain tumor with a poor prognosis. Therapies to date have failed to improve survival. LGGs, with IDH mutations, have increased global DNA methylation and increased survival compared with GBMs. GBMs lack this mutation and have less DNA methylation. Here we show that the DNA methylome can be modified in GBM with LMF. Such treatment might be useful in methylome priming prior to immunotherapy.

Introduction

In solid tumors, half of the genome contains large blocks of hypomethylated DNA (1). These hypomethylated blocks are found early in tumorigenesis and correspond to large organized chromatin lysine modifications and lamin-associated domains (2–5). Within these large hypomethylated blocks, at promoter regions of specific genes, there is increased methylation of sites called CpG island (CGI) regions. Gliomas with hypermethylated CGIs are associated with improved prognosis and are classified as having a CpG island methylator phenotype (CIMP; refs. 6, 7). In primary brain tumors, glioma CIMP (G-CIMP) is highly associated with mutations in genes encoding isocitrate dehydrogenase (*IDH*[1]/[2]), and the *IDH* mutation, when overexpressed

[1]Department of Epidemiology, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire. [2]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee. [3]Department of Pathology, Vanderbilt University Medical Center, Nashville, Tennessee. [4]Department of Medicine, Yale Medical School, New Haven, Connecticut. [5]Department of Neurosurgery, Vanderbilt University Medical Center, Nashville, Tennessee. [6]Tennessee Oncology PLLC, Nashville, Tennessee. [7]Department of Neurosurgery, UCSF, San Francisco, California. [8]HudsonAlpha, Huntsville, Alabama. [9]CD Genomics, Shirley, New York. [10]Department of Neurology, Vanderbilt University Medical Center. [11]Division of Neuro-Oncology, Vanderbilt-Ingram Cancer Center, Nashville, Tennessee. [12]Department of Molecular and Systems Biology, Geisel School of Medicine at

Downloaded from http://aacrjournals.org/cancerrescommun/article-pdf/2/1/1/3206298/crc-21-0088.pdf by Dartmouth College user on 29 September 2022

Dartmouth, Hanover, New Hampshire. [13]Epiphany Biosciences, San Francisco, California. [14]Sir Galahad Labs, Nashville, Tennessee.

S.W. Clark and B.C. Christensen jointly supervised this article.

Corresponding Author: Stephen W. Clark, Department of Neurology, Division of Neuro-Oncology, Vanderbilt University Medical Center, 1161 21st Avenue South, Nashville, TN 37232. Phone: 615-936-0060; E-mail: stephenwclark@gmail.com **doi:** 10.1158/2767-9764.CRC-21-0088

This open access article is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

© 2022 The Authors; Published by the American Association for Cancer Research

in normal astrocytes, can recapitulate the G-CIMP hypermethylation pattern (6–8). G-CIMP tumors have improved prognosis, irrespective of tumor grade, but *IDH* mutations and the G-CIMP epitype are rare in the most common and aggressive glioma, glioblastoma (GBM; ref. 6). In addition to the large blocks of DNA hypomethylation found in the initial tumor, there is additional loss of DNA methylation during tumor recurrence or tumor progression in many cancer types (1, 2). Similarly, hypomethylation occurs during the progression of *IDH*-mutant low-grade glioma (*IDH*-mutant LGG) to secondary (*IDH*-mutant) GBM (9), and during *IDH*-mutant LGG (grade II/III) recurrence, where G-CIMP–high tumors progress to G-CIMP–low tumors with hypomethylation in the CIMP genes (9, 10). Moreover, DNA hypomethylation occurs during *IDH*-mutant secondary GBM recurrence (9, 11). Together, this implies that recurrent or progressing gliomas have loss of DNA methylation (hypomethylation) compared with the primary tumor.

Recent data by Zhou and colleagues have shown that DNA hypomethylation in cancer is mainly found in lamina-associated domains that are located within late-replicating regions of the genome (12). Moreover, it was found that DNA hypomethylation increased with age and tracked the accumulation of cell divisions, and was associated with somatic mutation density. Thus, they proposed a remethylation window model, where replication late in the S-phase provides less time for remethylation of newly synthesized daughter strands during DNA replication, contributing to hypomethylation during tumor growth (12).

It has been shown that treatment with the methyl donor, folate, increased global DNA methylation in cultured glioma cells. Moreover, in a xenograft-induced glioma mouse model, folate increased global DNA methylation and reduced tumor size (13, 14). It is unknown whether DNA hypomethylation occurs with primary GBM progression (*IDH* wild-type) and it is unknown if the epigenome of gliomas can be reprogrammed.

These data provide a rationale for exploring the use of a methyl donor in recurrent high-grade glioma. In this study, we evaluated the safety, tolerability, preliminary efficacy, and DNA methylome dynamics after treatment with the methyl donor, l-methylfolate, combined with temozolomide and bevacizumab in recurrent *IDH* wild-type high-grade gliomas.

Materials and Methods

Protocol Objectives

The primary objective of phase I was to determine the MTD of l-methylfolate (LMF) in combination with bevacizumab and temozolomide in patients with *IDH* wild-type recurrent (first recurrence) malignant glioma. The secondary objectives were to assess the objective response and safety profile of LMF in combination with bevacizumab and temozolomide in patients with *IDH* wildtype recurrent MG.

Patient Eligibility

This protocol (ClinicalTrials.gov # NCT01891747, using the CONSORT guideline) was institutional review board (IRB) approved, and all study patients signed written informed consent. Study patients, 18 years of age or older, must have had histologically proven malignant glioma. Patients must have had genetically confirmed *IDH* wild-type tumor. Patients must have had measurable contrast-enhancing recurrent malignant glioma by MRI within two weeks of starting the treatment. Measurable disease is defined by at least one enhancing lesion accurately measured in at least one dimension greater or equal to 5 mm and patients could have nonmeasurable disease if they had recent surgery for radiographic progression but no patients had recent surgery. They must have recovered from the severe toxicity of prior therapy. Study patients must have had a Karnofsky Performance Status greater or equal to 60. They must have had adequate bone marrow and organ function. Women of childbearing potential were required to have a negative pregnancy test within 10–14 days.

Exclusion criteria include prior treatment within two weeks of entering the study, genetically confirmed *IDH*[illegible] mutation, HIV positive, or pregnancy. The remaining exclusion criteria were similar to those of previous brain tumor therapeutic clinical trials.

Trial Design

In this nonrandomized prospective phase I trial with standard 3 + 3 design, study patients received LMF (Alfasigma) orally twice daily in a 28-day cycle and bevacizumab 10 mg/kg every two weeks by intravenous infusion in an outpatient setting on days 1 and 15. In addition, patients received a 5-day regimen of temozolomide at 150 mg/m^2/day each cycle, and daily 250 mg tablet of vitamin C. Dose escalation involved 3 patients treated at each dose level of LMF (15 mg, 30 mg, 60 mg, or 90 mg) and dose escalation occurred in a stepwise fashion. The 15-mg dose level was a once-a-day dose of 15 mg, whereas subsequent dose levels were twice daily (i.e., 30-mg level, was 15 mg twice daily). MTD was not reached. Patients were continued on therapy until there was disease progression or significant toxicities occurred.

This phase I study was conducted in accordance with the Declaration of Helsinki, International Ethical Guidelines for Biomedical Research Involving Human Subjects (CIOMS), Belmont Report, and the U.S. Common Rule.

Toxicity Assessment

At each patient visit, both safety and toxicity were evaluated. The Common Terminology Criteria of Adverse Events (CTCAE version 4.02) was used to grade adverse patient events. If patients experienced a grade 4 nonhematologic doselimiting toxicity (DLT), LMF, and/or bevacizumab and/or temozolomide was stopped (this did not occur). We defined hematologic DLTs as those where platelets were equal or less than 25,000 or if a second occurrence where the platelets fell below 50,0000; the latter would be considered a grade 3. Other conditions that were considered as

Downloaded from http://aacrjournals.org/cancerrescommun/article-pdf/2/1/1/3206298/crc-21-0088.pdf by Dartmouth College user on 29 September 2022

hematologic DLTs were any event in which there is clinically significant bleeding and a platelet count below 50,000 or an ANC equal or below 1,000/μL or febrile neutropenia.

Response Assessment

Patients were evaluated every 8 weeks with a contrasted brain MRI. We used the criteria outlined in the Response Assessment in Neuro-Oncology (RANO) for evaluation of changes in tumor (15). Patients were evaluated whether they had a complete response, partial response, stable disease, or progressive disease as outlined by the RANO criteria (15).

Bevacizumab Cohort

To generate a bevacizumab control group, we performed a retrospective chart review for all patients with recurrent GBM treated at Vanderbilt Medical Center (Nashville, TN). After approval by the Vanderbilt IRB, we identified 50 patients treated between the January 1, 2005 and December 31, 2014. To avoid any selection bias, we had the following inclusion criteria: (i) GBM diagnosis; (ii) treated with only with standard therapy including chemoradiation with temozolomide, followed by adjuvant monthly temozolomide; (iii) first recurrence; (iv) only one surgical resection for initial diagnosis and no further surgery; (v) recurrence had to be greater than 8 weeks from the completion of radiation; (vi) patient had to have signed consent for database participation; (vii) patients had to be treated with bevacizumab therapy at recurrence.

Survival Statistical Analysis

We defined progression-free survival (PFS) as the time between starting therapy and progression on MRI or death. We defined overall survival (OS) as the time from clinical trial start date to time of death. The Kaplan–Meier method was used to estimate PFS and OS. A median (range) was used to summarize continuous variables and frequencies and percentages were used to describe categorical variables. Any patients that remained alive at the end of the study were censored in the OS calculation.

O^6-methylguanine-DNA Methyltransferase Gene Promoter Methylation and IDH Mutation Analysis

This protocol is for O^6-methylguanine-DNA Methyltransferase Gene (*MGMT*) methylation and *IDH* mutation results used on all tumors seen at the Vanderbilt University Medical Center (Nashville, TN) and are not to be confused with the methylation studies involving CpGs described later.

DNAwasisolatedfromformalin-fixedparaffin-embedded(FFPE)patientspecimens. Using our AB17900 PCR (Applied Biosystems), methylation-specific PCR analysis was carried out. Methylated MGMT was defined as those tumors with a methylation score greater than 2. β-*actin* was used as a control for copy number. We used multiplex PCR and primer extension to evaluate for *IDH1* and *IDH2* mutations. Products of PCR were fluorescently labeled and analyzed with capillary electrophoresis. The following mutations were tested for *IDH1*: R132G, R132S, R132C, R132H, R132 L, and R132P. For *IDH2*, the following mutations were tested: R140P, R140L, R140Q, R140G, R140W, R172S, R172T, R172M, R172K, R172G, and R172W.

Surgical and Autopsy Sample Acquisition

All patients provided informed written consent. The use of trial (LMF) patient's initial tumor and autopsy tumor was approved by the Vanderbilt IRB. Controlled deidentified paired tumor samples were obtained after approval from Vanderbilt and UCSF IRBs. In the latter case, deidentified paired GBM tissue was obtained from the UCSF Brain Tumor SPORE Tissue Core (P50CA097257). Where possible initial, recurrent, and autopsy samples were collected during surgical resection (or at autopsy) and were snap-frozen in liquid nitrogen and stored at −80°C. All autopsy samples were freshfrozen.

DNA Isolation

Genomic DNA was isolated from fresh-frozen or FFPE samples. DNA was extractedfromthesamplesusingtheQIAGENDNeasykit(QiagenInc.)following the manufacturer's protocol at Hudson Alpha. The FFPE samples required the use of a DNA Restoration Kit (Agencourt FormaPure, Beckman Coulter, and Illumina's FFPE Restore Kit, Illumina Inc.) prior to processing.

The isolated genomic DNA was stored at −80°C until use. For every sample, 1 μg of DNA was bisulfite-converted using the Zymo EZ DNA Methylation Kit (ZYMO Research Corporation) according to the manufacturer's instructions. Converted DNA was eluted in a 22 mL elution buffer. DNA methylation level was assessed using the Infinium Human Methylation EPIC Beadchip (Illumina Inc.) at Hudson Alpha according to the manufacturer's instructions. Bisulfite-converted DNA was amplified and then enzymatically fragmented and precipitated. The DNA was then dissolved in hybridization buffer and placed on the Human Methylation EPIC Beadchips. To reduce batch effects, samples were distributed in random blocks and hybridization was allowed to occur for 20 hours at 48°C in Illumina's hybridization oven. Free DNA was then washed away, then the beadchips were processed through a single nucleotide extension following IHC staining (dideoxynucleotides triphosphates; ddNTP) using capillary flow through chambers (Tecan GenePaint automated slide processor). Using the Illumina iScan software and iScan system, a fluorescent signal was captured. Using Illumina GenomeStudio software, background subtraction was completed and intensity data (idat) files were generated, one per channel. A similar process was performed in the The Cancer Genome Atlas (TCGA) datasets, although they only used fresh-frozen samples not requiring the use of the restoration kit; the samples of the TCGA were run using Infinium HumanMethylation450K BeadChips.

Bioinformatics Analyses

The idat files generated from the Infinium Human Methylation EPIC Beadchip and the Infinium Human Methylation 450 Beadchip (Illumina, Inc.) were imported and preprocessed using the minfi and sesame packages in R. DNA methylation β-values were estimated based on the measured intensities of the two-paired channels (i.e., red and green) and computed as the ratio of the methylated probe intensity, divided by the sum of the unmethylated plus the methylated intensities signals plus an offset of 100 (16, 17). β-values range between 0 and 1 and can be interpreted as a proxy to the proportion of methylated alleles at a specific CpG site. β-values were

Downloaded from http://aacrjournals.org/cancerrescommun/article-pdf/2/1/32062 98/crc-21-0088.pdf by Dartmouth College user on 29 September 2022

background corrected using noob and a nonlinear dye correction using sesame. As two different generations of microarrays were combined, we included the microarray slide as a fixed effect to control potential batch effects. Low-quality samples were considered whether the probes showed no statistical difference using an out-of-band sesame array hybridization p-OOBAH detection value ($P_{detection} > 0.05$), or whether the mean bisulfite conversion probes of a specific sample were below 3 SD of the mean bisulfite conversion of all the probes. Probes that were marked as low quality in more than 5% across samples were masked for all the samples, and samples that showed more than 30% of probes with low-quality detection were eliminated from the analyses. Probes marked as CpG loci on the X and Y chromosomes, and those previously documented as polymorphic or cross-reactive were excluded from subsequent analyses (18).

Genome-wide DNA Methylation Analyses

We implemented a locus-by-locus analysis aimed at identifying differentially methylated CpG sites based on the treatment effect (using the package limma; ref. 19). Briefly, linear mixed effect models were fit to each CpG site separately and modeled β-values of the DNA methylation as the response against the treatment versus the controls. Models were adjusted for chronologic age (years), sex, sample method processing (fresh-frozen or FFPE), microarray slide, and subject as a random effect. Although our examination was exploratory in nature, *P* values were adjusted for multiple comparisons by computing the Benjamini– Hochberg Q values also known as false discovery rate (FDR). As we expected small effect sizes, we ranked the CpG loci according to the FDR, prioritizing those with the biggest magnitude of the DNA methylation differences (absolute β>0.2) for further exploration. All analyses were carried out using the R statistical package, version 4.0.2 (www.r-project.org/).

DNA Methylation Changes During IDH wild-type GBM Recurrence

We identified 15 total matched samples, five from our institutions (Vanderbilt/UCSF, 850K chip), and ten from the TCGA (450K chip) of *IDH* wild-type GBMs with initial tumor and recurrent tumor data. To determine the extent to which these tumors had altered methylomes, we calculated the change in methylation (β value, the methylated fraction at a CpG site) from initial GBM to recurrent GBM at each CpG site in each patient, and identified CpG sites with consistent methylation changes upon recurrence across all patients. We restricted our analysis to shared CpG sites between the two methylation array platforms (450K and 850k) and fitted linear mixed effect models to test the relation of methylation with GBM recurrence.

Genome-wide DNA Methylation Dysregulation Index

To evaluate global methylation changes between paired samples, we used a modification of the methylation dysregulation index (MDI; ref. 20). Briefly, this modified MDI measure represents the cumulative change from the primary tumor DNA methylation levels in a CpG locus–specific manner calculated by summing the absolute difference in DNA methylation β-values at each CpG between each recurrent tumor sample and the matched primary tumor sample β-value and then divided by the total number of interrogated CpGs. To better reflect scale, MDI was multiplied by 100. The MDI here represents the average change in β-value per CpG in the recurrent tumor sample compared with the primary sample. Therefore, an MDI value close to 0 suggests a similar methylation profile to the matched primary tumor while increasing levels of MDI indicate that the DNA methylome has been deregulated to a greater extent in the recurrent tumor.

Enrichment Analyses

For the DNA methylation analyses, a test for enrichment was performed using the missMethyl package (21). This algorithm maps the CpG probes to gene IDs and test for Gene Ontology (GO) and KEGG pathways using a hypergeometric

TABLE 1 Patient characteristics

test. As in the design of the Illumina arrays, multiple CpGs could be present in the same locus, the test for enrichment is corrected taking into account the bias generated for the number of CpGs per gene in the EPIC or Illumina Human Methylation 450K arrays. The terms obtained in the GO analyses were grouped using REVIGO into hierarchical groups (22).

Statistical Analysis

All analysis was performed using R version 4.0.2 or higher, using R packages for statistical analysis, DNA methylation, and gene expression preprocessing and analysis including Table 1 (general descriptive statistics), survival (survival analysis), minfi and sesame (DNA methylation preprocessing, quality control, filtering, background correction, and normalization), limma (linear models and linear mixed effect models), q value (FDR adjustments), and missMethyl (hypergeometric test for enrichment). Graphs were generated using R plots, ggplot2, treemap (REVIGO), and heatmap.

Data Availability Statement

All high-throughput data mentioned in the article are publicly available from GEO under accession GSE111627. All data that support the findings of this study are available from the corresponding author upon reasonable request.

Results

We evaluated the DNA methylome changes of *IDH* wild-type GBM at recurrence and compared baseline methylation status between GBM and LGG and especially *IDH* wild-type GBM and *IDH*-mutant LGG. First, we compared the DNA methylomes between matched samples (initial tumor at diagnosis to the recurrent tumor from the same patient) of primary (*IDH* wild-type) GBMs, using Illumina's Infinium Human Methylation 450K and Methylation EPIC(850K) Beadchips. We identified 15 total matched samples, five from our institutions (Vanderbilt/UCSF, 850K chip), and ten from the TCGA (450K chip). The clinical, histopathologic, and molecular characteristics of these patients are summarized in Table 1 (third column "TCGA Cohort" and fourth

Downloaded from http://aacrjournals.org/cancerrescommun/article-pdf/2/1/1/320629

	LMF + bevacizumab + temozolomide Cohort	TCGA Cohort	Bevacizumab Cohort	Vanderbilt/UCSF Characteristics	
n (pathology)	14 (13 GBM, 1 AA)	50 (all GBM)	10 (all GBM)	5 (all GBM)	
Age, median (range), years	59 (41–71)	58 (18–75)	60 (36–72)	56 (38–72)	
Gender (%)					
Male	86	52	67	25	
Female	14	48	33	75	
KPS, median (range)	80 (60–90)	80 (50–100)	d	80 (70–90)	
Received chemoradiation with temozolomide (%)	100	100	100	100	
MGMT methylated (%)[a]	33	b	b	b	
IDH mutated (%)	0	b	0	0	
Median exposure to LMF (range), days [c]	192 (46–701)	–	–	–	

[a]Nine of the 14 samples had *MGMT* methylation data and all were IDH wild-type.

[b]In the majority of these patients the *MGMT* and *IDH* status were not known.

[c]Days to progression. [d]KPS is not known for TCGA samples.

Abbreviation: KPS = Karnosfsky performance status.

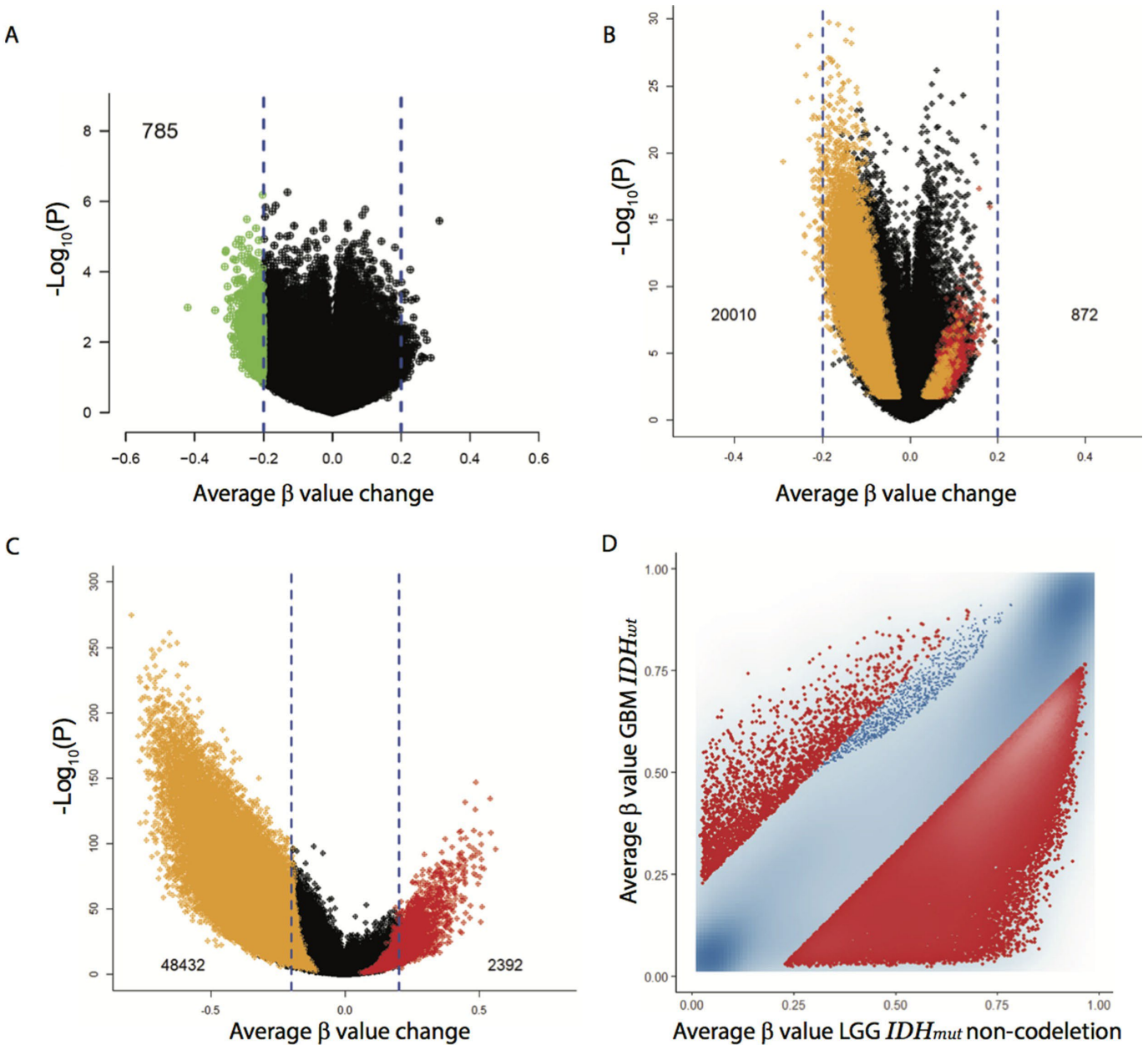


FIGURE 1 DNA methylation changes during IDH-wild-type (IDH-wt) GBM recurrence. **A,** Volcano plot comparing the average methylation change from primary GBM to recurrent GBM using paired samples from TCGA and from our institutions (450K methylation platform). 785 probes with reduced methylation in recurrent GBM of >0.2 on the average beta scale are shown in green. **B,** Volcano plot comparing overlapping CpGs from GBM to LGG (not specific for IDH) using TCGA methylation data (450K methylation platform). Colored dots represent CpG sites that show significant hypermethylation, FDR<0.05 and β> 0.2, (red dots, total count provided) or hypomethylation, FDR<0.05 and β<−0.2, (orange dots, total count provided) including those that were inconsistent between the model coefficient and the β direction. **C,** Volcano plot comparing overlapping CpGs from GBM (*IDH wt*) to LGG (*IDH mutant*). Colored dots represent as in **B**. **D,** GBM (*IDH wt*) –LGG (*IDH mutant*) scatterplot, each axis representing the average β-value per interrogated CpG, GBM *IDH-wt* on the *y*-axis and LGG *IDH-mutant* on the *x*-axis.

column "Vanderbilt/UCSF Cohort"). We identified 785 CpG sites that were specifically hypomethylated in recurrent GBM (|β|>0.2, or >20% decrease in methylation), more than six times the 124 CpGs with an equivalent increase in methylation (Fig. 1A, green dots: hypomethylated CpGs; ref. 9). Due to the small sample size, the methylation β values did not reach statistical significance after correcting for multiple comparisons.

Downloaded from http://aacrjournals.org/cancerrescommun/article-pdf/2/1/32/3206298/crc-21-0088.pdf by Dartmouth College user on 29 September 2022

Using TCGA methylation data, we compared GBM to LGG and found significant hypomethylation in GBM compared with LGG (Fig. 1B, orange dots: hypomethylated CpGs from GBM compared with LGG; red dots: hypermethylated CpGs from GBM compared with LGG). This difference is most apparent when comparing *IDH* wild-type GBMs to *IDH*-mutant LGGs (Fig. 1C and D, orange dots: hypomethylated CpGs from GBM *IDH* wild-type compared with LGG *IDH*-mutant; red dots: hypermethylated CpGs from GBM *IDH* wild-type compared with LGG *IDH*-mutant).

TABLE 2 Toxicity results of LMF treatment in phase 1

Patient#		Cycle	Total dose (LMF)mg/d	Toxicity Category	Toxicity	Grade	Clinically significant	Overall relation	Invest Tx
5	4	30 GI	GI reflux 2	No	Invest Tx LMF				
6	2	30 GI	Diarrhea 2	No	Invest Tx Vit C				
7	1	30 CNS	Dysgeusia 2	No	Invest Tx LMF				
7		3	30	CNS	Dysgeusia	1	No	Invest Tx	LMF
9 1 60 GI Dyspepsia 1 No Invest Tx LMF 9 2 60 Derm Pruritus 1 No Invest Tx LMF 9 2 60 GI Dyspepsia 1 No Invest Tx LMF 9 5 60 GI Bloating 1 No Invest Tx LMF									

We next investigated whether LMF could reprogram the DNA methylome in recurrent high-grade glioma patients by increasing DNA methylation. Patients with recurrent primary (*IDH*-wild type), high-grade glioma were enrolled in a phase I study and treated with LMF (NCT01891747; patient characteristics, Table 1, first column LMF + bevacizumab + temozolomide). Fourteen patients with recurrent high-grade glioma (13 patients with GBM and one patient with anaplastic astrocytoma) received twice-daily LMF along with standard therapy bevacizumab and temozolomide. The median age of the study population was 59 (range, 41–71) years. Most patients were men. Six patients decided to donate their brains at death.

Overall, the most common AEs were diarrhea in 1 (7%) of 14 patients, reflux in 1 (7%) patient, and dysgeusia in 1 (7%) patient (Table 2). There was no grade 3 or higher toxicity. No laboratory toxicities were appreciated and no MTD was reached.

Out of 14 patients treated with LMF, 2 patients had progressive disease (14%), 6 patients with partial response (43%), and 6 with stable disease (43%). LMFtreated patients had a median OS (mOS) of 9.5 months (95% CI, 9.1–35.4) and PFS of 6 months (95% CI, 3.6–10.1) and these were not statistically different from the bevacizumab control cohort (patient characteristics, Table 1, second column, Bevacizumab cohort) with a mOS, of 8.6 (95% CI, 6.8–10.8), and PFS of 4.1 (95% CI, 2.8–6; Fig. 2A; Supplementary Fig. S1). Of the 14 patients treated with LMF, one patient is alive, and 6 patients survived longer than 650 days (656, 658, 689, 739, 1,080 days), whereas two survived less than 73 days (72 and 69 days). Six of the trial participants (all GBM) donated their brains at death and were available for additional studies; time on LMF of these 6 patients: 46, 76, 108, 365, 459, and 620 days. (Supplementary Table S1).

Using the Illumina 850k Methylation platform, we compared DNA methylation between paired samples, initial tumor to the autopsied tumor from the same patient in the 6 patients that donated thier brains after LMF treatment. We observed increased DNA methylation in the autopsied tumors of LMF-treated patients, with 1,239 CpGs demonstrating significantly increased DNA methylation compared with the initial tumor (Fig. 2B, right volcano plot; black and orange dots). Next, we broadly analyzed the genomic location of the DNA hypermethylation (1,239 CpGs) and found these methylated CpGs associated with the DNA shelf, shore, and open sea, but not within the CGIs (Supplementary Fig. S2). When we restricted our analysis to the most hypermethylated CpGs (645 hypermethylated > 0.2β), we found increased methylation across the genome, including CGIs (Fig. 2C). Furthermore, LMF treatment reversed the DNA hypomethylation/hypermethylation ratio in LMF-treated patients (515 hypomethylated to 645 hypermethylated CpGs) compared with that found in LGG transition to secondary (*IDH*-mutant) GBM (4,343 hypomethylated to 311 hypermethylated CpGs) and primary (*IDH*-wild-type) GBM recurrence (785 hypomethylated to 124 hypermethylated CpGs; ref. 9).

To further highlight the methylation dynamics imparted by LMF treatment, we compared control recurrent GBM groups (TCGA and Institutional samples) to the LMF-treated group using an aggregate measure of methylation alteration, MDI (20); this is a top-down approach, where MDI represents the average departure of DNA methylation between initial tumor and recurrent tumor. Using this method, we found a significant MDI difference between the LMF-treated group and the TCGA control group (Fig. 2D).

We evaluated targets of LMF-induced DNA hypermethylation by gene ontology enrichment and REVIGO. We found that the most hypermethylated CpGs (Fig. 2B, orange dots) were associated with diverse pathways, including cell projection organization, neuronal differentiation, cell adhesion, and cell proliferation (Fig. 2E; Supplementary Figs. S3 and S4).

Finally, to further evaluate the clinical implications of using LMF + temozolomide, we studied changes in methylation of CpGs associated with MGMT after LMF treatment. A total of 163 CpGs were annotated to the MGMT locus out of a total of 207 contained within the EPIC array. In two of the six patient's tumors, we found increased DNA methylation in MGMT CpGs after treatment with LMF compared with their initial tumor (Supplementary Fig. S5).

Downloaded from http://aacrjournals.org/cancerrescommun/article-pdf/2/1/1/3206298/crc-21-0088.pdf by Dartmouth College user on 29 September 2022

Discussion

Here we show that recurrent GBMs are epigenetically stable in comparison to the initial tumor with minimal loss of DNA methylation. Some DNA hypomethylation was appreciated but due to the small sample size this was not significant and this data is consistent with a larger study of 112 matched primary (*IDH* wild-type) GBM patients comparing initial tumor to recurrent tumor by Klughammer and colleagues (23); this group, using a different technique, reduced representation bisulfite sequencing, found additional loss of DNA methylation in some promoters of recurrent tumors, but no significant global hypomethylation in recurrent GBM. These results contrast what is seen in progression from *IDH*-mutant LGG to *IDH*-mutant GBM, where substantial DNA hypomethylation occurs in the *IDH*-mutant GBMs

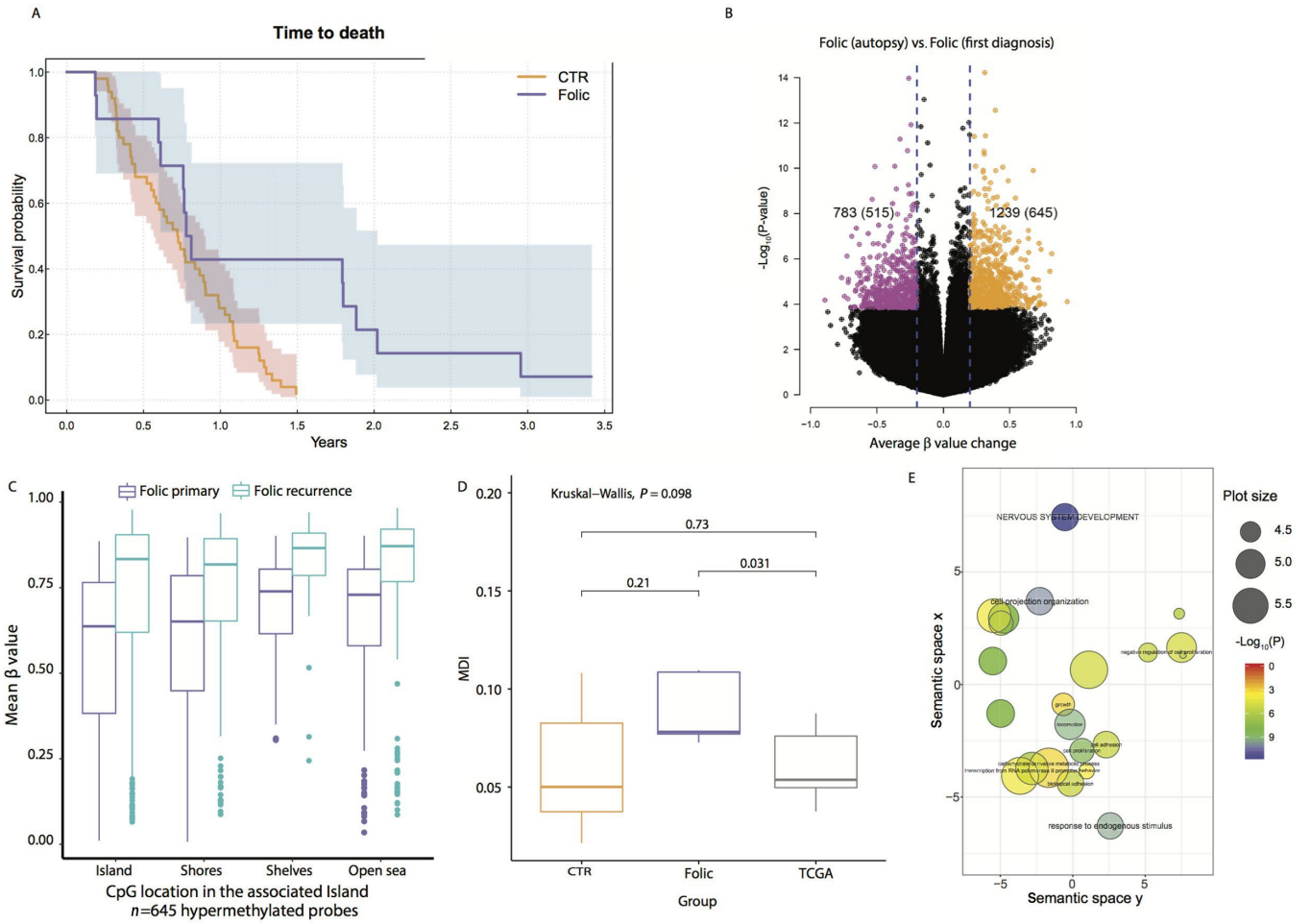


FIGURE 2 DNA methylation changes in recurrent *IDH*-wt GBM after treatment with LMF. Note: "Folic" is LMF. **A,** Kaplan–Meier curve of OS in LMF-treated patients (blue line) versus historical control group treated with bevacizumab alone (red line). The 95% CI is represented by the blue and red shaded areas. See Supplementary Table S1 for patient characteristics and n value. CTR, bevacizumab treatment; folic is LMF-treated patients. **B,** Volcano plot comparing the average methylation change from initial GBM (at diagnosis) to autopsied GBM (after exposure to LMF at recurrence, measurements were taken from distinct paired samples). Colored dots represent CpG sites that show significant hypermethylation, FDR<0.05 and β > 0.2, (orange dots, total count provided in the parentheses) or hypomethylation, FDR<0.05 and β<−0.2, (purple dots, total count provided in the parentheses) at autopsy. The total CpGs that were FDR<0.05 at any level of β are summarized outside the parentheses. The volcano plot was generated using the 850K-EPIC platform. Folic, LMF treated. **C,** The genomic distribution of the 645 hypermethylated CpGs [red box plot, initial GBM (at diagnosis); blue box plot, autopsied GBM tumor (after exposure to LMF at recurrence)]. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers. Folic, LMF treated. **D,** Comparison of the DNA MDI between LMF-treated patients (Folic), TCGA paired samples (TCGA), and our institutional paired controls (CTR). Box plot elements as in **C**. **E,** Using gene ontology (GO) analysis, we show pathways impacted by the 645 hypermethylated CpGs identified in **B**. GO terms were summarized using REVIGO (http://revigo.irb.hr/). For each class, terms remaining after the redundancy reduction are represented in a two-dimensional space, which summarizes GO terms' semantic similarities.

Semantically similar terms are close together in the plot, but the semantic space units have no intrinsic meaning. Bubble color indicates the *P* value (all terms included had a $P < 0.05$); circle size indicates the frequency of the GO term in the underlying GO database (more general terms have larger plot size). CpG Island: 300 to 2,000 bp in length close to promoters with high GC content; CpG Shores: <2 kb from CpG Island; CpG Shelves: 2–4 kb from CpG Island; Open Sea: >4 kb from CpG Island.

(9). We speculated the reason for this difference is a lower baseline global methylation in *IDH* wild-type GBM compared with *IDH*-mutant LGG. We show here that *IDH* wild-type GBM when compared with *IDH*-mutant LGG have significantly lower DNA methylation (Fig. 1C and D). Thus, we hypothesize that hypomethylated tumors, such as *IDH* wild-type GBMs can more easily remethylate during replication and thus less DNA hypomethylation occurs in late-replicating regions (12).

Downloaded from http://aacrjournals.org/cancerrescommun/article-pdf/2/1/32/3206298/crc-21-0088.pdf by Dartmouth College user on 29 September 2022

On the basis of the data that *IDH*-mutant gliomas have both increased global DNA methylation and survival compared with *IDH* wild-type gliomas, and that there is *in vivo* and *in vitro* data showing folate treatment can reduce growth of glioma cells, we tested whether the DNA methylome of *IDH* wild-type gliomas could be reprogrammed. We used LMF, for several reasons: first, folate or folic acid when taken by mouth is not efficiently transported into the brain; however,

tetrahydrofolate forms such as LMF are (24). Moreover, LMF has shown clinical activity in two CNS clinical trials including a phase IV study of selective serotonin reuptake inhibitor–resistant depression (25, 26). Even though all patients in phase I had failed standard therapy with temozolomide, temozolomide was used because preclinical data showed folate treatment increased the sensitivity of temozolomide-induced apoptosis in glioma cells (13, 14).

We used recurrent *IDH* wild-type GBMs instead of recurrent *IDH*-mutant tumors in this phase I study for several reasons. Because *IDH* wild-type GBMs have worse prognosis than *IDH* mutants, our primary goal was to evaluate whether remethylation of the wild-type tumors could improve their survival and mimic the survival found in *IDH*-mutant GBMs. In addition, as we showed in Fig 1B and C, IDH wild-type GBMs are more hypomethylated at baseline than*IDH*-mutants,andthustoincreaseourabilitytodetectmethylationchange after remethylation, we used the more hypomethylated *IDH* wild-type gliomas. Finally, although the *IDH* mutation is the driver of hypermethylation, *IDH* mutants that retain this mutation still lose DNA methylation with progression from *IDH*-mutant LGG to *IDH*-mutant GBM, implying other defects in the methylation machinery or other factors in the methylation pathway go awry with progression to *IDH*-mutant GBM (9). We were concerned that such defects could interfere with LMF remethylation in *IDH* mutants, and thus, *IDH* mutants were excluded.

Our trial data shows that LMF was well tolerated and is safe. While we found no significant impact on survival in recurrent high-grade gliomas, we do show for the first time that the DNA methylome of *IDH* wild-type high-grade gliomas can be reprogrammed.

The limitation of this study is the sample size, as the phase I trial was not powered to detect a survival advantage. Likewise, the paucity of paired recurrent/autopsy GBM samples treated with LMF, limited our ability to identify the effects of tumor heterogeneity on LMFs ability to reprogram the DNA methylome (9, 11, 27–31).

In conclusion, we show here for the first time in a human study that the DNA methylome of high-grade glioma can be reprogrammed. Given the emerging role of epigenetic treatments in immuno-oncology, understanding the dynamics of epigenetic reprogramming will be essential and our work is a first step toward this goal (32).

Authors' Disclosures

No disclosures were reported.

Authors' Contributions

Lucas A. Salas[1], Thomas G. Stewart[2], Bret C. Mobley[3], Chengwei Peng[4], Jing Liu[5], Sudan N. Loganathan[5], Jialiang Wang[5], Yanjun Ma[6], Mitchel S. Berger[7], Devin Absher[8], Yang Hu[9], Paul L. Moots[10], Brock C. Christensen[1,11], and Stephen W. Clark[10,12,*]

L.A. Salas: Data curation, formal analysis, methodology, writing-review and editing. **T.G. Stewart:** Formal analysis, methodology, writing-review and editing. **B.C. Mobley:** Data curation, writing-review and editing. **C. Peng:** Data curation, formal analysis, writing-review and editing. **J. Liu:** Data curation, writing-review and editing. **S.N. Loganathan:** Data curation, writing-review and editing. **J. Wang:** Data curation, formal analysis, writing-review and editing. **Y. Ma:** Formal analysis, writing-review and editing. **M.S. Berger:** Resources, writing-review and editing. **D. Absher:** Data curation, formal analysis, writing-review and editing. **Y. Hu:** Formal analysis, writing-review and editing. **P.L. Moots:** Data curation, investigation, writing-review and editing. **B.C. Christensen:** Formal analysis, supervision, methodology, writing-review and editing. **S.W. Clark:** Conceptualization, resources, data curation, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing-original draft, project administration, writing-review and editing.

Acknowledgments

We thank the patients and families who selflessly contributed to this study through tissue donations obtained from surgery and/or autopsy. We also thank the nursing and research staff for their help with clinical trial and patientrelated issues: Katie Lyons, Pam Dawson, Linda Leo, Elizabeth Kieffer, Laura Jean, Kristine Leenders, and Allen Redmond. We are also grateful for Dr. Mark Anderson, Department of Neurology, University of Mississippi, and Dr. Anna Mathew, Department of Pathology, the University of Southern California, for their help obtaining autopsy tissue. We also acknowledge the UCSF Brain Tumor SPORE Tissue Core (P50CA097257), Vanderbilt Brain Tumor Center Molecular Neurosurgical Tissue Bank, and Vanderbilt REDCap database (1 UL1 RR024975) for providing de-identified paired GBM tissue. This research is supported by The Barbara Russell Family Trust, in memory of Barbara Russell, who participated in this study, and the Voland Fund, in memory of Stephen Voland. Additional funding support was provided by the NIH grants (R01DE022772, R01CA216265, P20GM104416/6369, to B.C. Christensen; and P20GM104416–09/8299, to L.A. Salas), and the Congressionally Directed Medical Research Programs/Department of Defense (W81XWH-20–1-0778, to L.A. Salas).

Note

Supplementary data for this article are available at Cancer Research Communications Online (https://aacrjournals.org/cancerrescommun/).

Received October 19, 2021; revised October 31, 2021; accepted November 18, 2021; published first January 05, 2022.

Downloaded from http://aacrjournals.org/cancerrescommun/article-pdf/2/1/32/3206298/crc-21-0088.pdf by Dartmouth College user on 29 September 2022

References

1. Timp W, Bravo HC, McDonald OG, Goggins M, Umbricht C, Zeiger M, et al. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. Genome Med 2014;6: 61-71.
2. Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 1983;301: 89-92.

8

3. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, et al. Increased methylation variation in epigenetic domains across cancer types. Nat Genet 2011;43: 768-75.
4. Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet 2012;44: 40-6.
5. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP, et al. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci USA 1999;96: 8681-6.
6. Noushmehr H, Weissenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 2010;17: 510-22.
7. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell 2013;155: 462-77.
8. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature 2012;483: 479-83.
9. Mazor T, Pankov A, Johnson BE, Hong C, Hamilton EG, Bell RJ, et al. DNA methylation and somatic mutations converge on the cell cycle and define similar evolutionary histories in brain tumors. Cancer Cell 2015;28: 307-17.
10. de Souza CF, Sabedot TS, Malta TM, Stetson L, Morozova O, Sokolov A, et al. A distinct DNA Methylation shift in a subset of Glioma CpG Island Methylator phenotypes during tumor recurrence. Cell Rep 2018;23: 637-51.
11. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell 2016;164: 550-63.
12. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet 2018;50: 591-602.
13. Hervouet E, Debien E, Campion L, Charbord J, Menanteau J, Vallette FM, et al. Folate supplementation limits the aggressiveness of glioma via the remethylation of DNA repeats element and genes governing apoptosis and proliferation. Clin Cancer Res 2009;15: 3519-29.
14. Cartron PF, Hervouet E, Debien E, Olivier C, Pouliquen D, Menanteau J, et al. Folate supplementation limits the tumourigenesis in rodent models of gliomagenesis. Eur J Cancer 2012;48: 2431-41.
15. Quant EC, Wen PY. Response assessment in neuro-oncology. Curr Oncol Rep 2011;13: 50-56.
16. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: A flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014;30: 1363-9.
17. Zhou W, Triche TJ, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res 2018;46: 1-15.
18. Zhou W, Laird PW, Shen H. Comprehensive characterization annotation and innovative use of Infinium DNA methylation BeadChips probes. Nucleic Acids Res 2017;45: e22.
19. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43: e47.
20. Salas LA, Johnson KC, Koestler DC, O'Sullivan DE, Christensen BC. Integrative epigenetic and genetic pan-cancer somatic alteration portraits. Epigenetics 2017;12: 561-74.
21. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's humanmethylation 450 platform. Bioinformatics 2016;32: 286-8.
22. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 2011;6: e21800.
23. Klughammer J, Kiesel B, Roetzer T, Fortelny N, Nemc A, Nenning KH, et al. The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. Nat Med 2018;24: 1611-24.
24. Mattson RH, Gallagher BB, Reynolds EH, Glass D. Folate therapy in epilepsy. A controlled study. Arch Neurol 1973;29: 78-81.
25. Zajecka JM, Fava M, Shelton RC, Barrentine LW, Young P, Papakostas GI. Long-term efficacy, safety, and tolerability of L-methylfolate calcium 15 mg as adjunctive therapy with selective serotonin reuptake inhibitors: a 12-month, open-label study following a placebo-controlled acute study. J Clin Psychiatry 2016;77: 654-60.
26. Roffman JL, Petruzzi LJ, Tanner AS, Brown HE, Eryilmaz H, Ho NF, et al. Biochemical, physiological and clinical effects of l-methylfolate in schizophrenia: a randomized controlled trial. Mol Psychiatry 2018;23: 316-22.
27. Wang Q, Hu B, Hu X, Kim H, Squatrito M, Scarpace L, et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. Cancer Cell 2017;32: 42-56.
28. Kim J, Lee IH, Cho HJ, Park CK, Jung YS, Kim Y, et al. Spatiotemporal evolution of the primary glioblastoma genome. Cancer Cell 2015;28: 318-28.
29. Consortium G. Glioma through the looking GLASS: molecular evolution of diffuse gliomas and the Glioma Longitudinal Analysis Consortium. Neuro Oncol 2018;20: 873-84.
30. deCarvalho AC, Kim H, Poisson LM, Winn ME, Mueller C, Cherba D, et al. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. Nat Genet 2018;50: 708-17.
31. Chen X, Wen Q, Stucky A, Zeng Y, Gao S, Loudon WG, et al. Relapse pathway of glioblastoma revealed by single-cell molecular analysis. Carcinogenesis 2018;39: 931-36.
32. Topper MJ, Vaz M, Marrone KA, Brahmer JR, Baylin SB. The emerging role of epigenetic therapeutics in immuno-oncology. Nat Rev Clin Oncology 2020;17: 75-90.

Downloaded from http://aacrjournals.org/cancerrescommun/article-pdf/2/1/1/3206298/crc-21-0088.pdf by Dartmouth College user on 29 September 2022

OXFORD

JNCI J Natl Cancer Inst (2022) 114(3): djab195

https://doi.org/10.1093/jnci/djab195

First published online October 1, 2021 Article

Downloaded from https://academic.oup.com/jnci/article/114/3/463/6379711 by Dartmouth College Library user on 29 September 2022

Interactions of Age and Blood Immune Factors and Noninvasive Prediction of Glioma Survival

Annette M. Molinaro, PhD [1,2,*] John K. Wiencke, PhD,[1] Gayathri Warrier, MS [1] Devin C. Koestler, PhD,[3] Pranathi Chunduru, MS,[1] Ji Yoon Lee, BA,[1] Helen M. Hansen, BA,[1] Sean Lee, BS [1] Joaquin Anguiano, BA,[1] Terri Rice, MPH,[1] Paige M. Bracci, PhD [2] Lucie McCoy, MPH [1] Lucas A. Salas, MD, PhD [4] Brock C. Christensen, PhD [4,5] Margaret Wrensch, PhD,[1] Karl T. Kelsey, MD [6] Jennie W. Taylor, MD [1,7] Jennifer L. Clarke, MD [1,7]

[1]Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA; [2]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA; [3]Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, KS, USA; [4]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA; [5]Departments of Molecular and Systems Biology and Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA; [6]Departments of Epidemiology and Pathology and Laboratory Medicine, Brown University, Providence, RI, USA; and [7]Department of Neurology, University of California San Francisco, San Francisco, CA, USA

*Correspondence to: Annette M. Molinaro, PhD, Department of Neurological Surgery, University of California San Francisco, 1450 3rd St, San Francisco, CA 94158 USA (e-mail: annette.molinaro@ucsf.edu).

Abstract

Background: Tumor-based classification of human glioma portends patient prognosis, but considerable unexplained survival variability remains. Host factors (eg, age) also strongly influence survival times, partly reflecting a compromised immune system. How blood epigenetic measures of immune characteristics and age augment molecular classifications in glioma survival has not been investigated. We assess the prognostic impact of immune cell fractions and epigenetic age in archived blood across glioma molecular subtypes for the first time. Methods: We evaluated immune cell fractions and epigenetic age in archived blood from the University of California San Francisco Adult Glioma Study, which included a training set of 197 patients with IDH-wild type, 1p19q intact, TERT wild type (IDH/1p19q/TERT-WT) glioma, an evaluation set of 350 patients with other subtypes of glioma, and 454 patients without glioma. Results: IDH/1p19q/TERT-WT patients had lower lymphocyte fractions ($CD4^{þ}$ T, $CD8^{þ}$ T, natural killer, and B cells) and higher neutrophil fractions than people without glioma. Recursive partitioning analysis delineated 4 statistically significantly different survival groups for patients with IDH/1p19q/TERT-WT based on an interaction between chronological age and 2 blood immune factors, $CD4^{þ}$ T cells, and neutrophils. Median overall survival ranged from 0.76years (95% confidence interval¼0.55-0.99) for the worst survival group (n¼28) to 9.72years (95% confidence interval¼6.18 to not available) for the best (n¼33). The recursive partitioning analysis also statistically significantly delineated 4 risk groups in patients with other glioma subtypes. Conclusions: The delineation of different survival groups in the training and evaluation sets based on an interaction between chronological age and blood immune characteristics suggests that common host immune factors among different glioma types may affect survival. The ability of DNA methylation-based markers of immune status to capture diverse, clinically relevant information may facilitate noninvasive, personalized patient evaluation in the neuro-oncology clinic.

ARTICLE

Identification of molecular subtypes of human glioma has helped refine these cancers' classification beyond traditional histopathologic and grade criteria (1-3). The current World Health Organization (WHO) 2016 glioma classification scheme incorporates tumor grade, isocitrate dehydrogenase (IDH1/2) mutational status, and 1p/19q co-deletion status (3). Glioblastomas (GBMs), the most common type of glioma, are either IDH-mutant (9%) and occur at a median age of 38years (with a median overall survival [OS] of 3.6years) or are IDH-wild type (WT) (91%) and occur at a median age of 59years (with a median OS of 1.2years) (1,4). Non-GBM astrocytomas are either IDH-mutant and occur in younger patients (median age¼36years), with a median OS of 9.3years, or IDH-WT and occur in older patients (median age¼52years), with a median OS of 1.9years (1,4). Oligodendrogliomas have a median age of 44years and a median OS of 17.5years and are defined by both IDH mutation and co-deletion of 1 copy of the 1p and 19q chromosomal

Received: April 19, 2021; Revised: June 30, 2021; Accepted: September 23, 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com

arms. Tumor telomerase promoter mutation (TERT) can further define subtypes (Molecular Subtypes 2015) (2,5). Up to 20% of patients with IDH-WT glioblastomas and 35% of those with IDH-WT astrocytomas lack 1p/19q co-deletion and TERT mutation (MT) (1); they are known as IDH/1p19q/TERT-WT gliomas, or "triple-negatives." We previously showed that patients with IDH/1p19q/TERT-WT astrocytoma had statistically significantly longer and more variable survival than patients with IDH-WT/1p19q-WT/TERT-MT astrocytomas (4). The median OS was 3.3years (95% confidence interval [CI]¼2.1 to not available [NA]) for patients with triple-negative astrocytoma vs 1.7years (95% CI¼1.4 to 2.0) for those with IDH-WT/ 1p19q-WT/TERT-MT astrocytoma (4), which nonmolecular characteristics may further elucidate.

After the WHO 2016 classification, the single most important prognostic factor in adult glioma is age at diagnosis (6). Among patients with GBM, being older or younger than 50years of age distinguishes shorter- and longer-term survival groups (7-11). Unknown mechanisms contribute to age as a dominant factor in glioma survival, but likely age-related immune system changes play a vital role (12). Most immune response components are affected during the aging process, which supports the concept of immunosenescence (13) and its corollary that a person's immunologic "age" may diverge from their chronological age (14). Immune aging phenotypes include the diminished primary immune responses of older individuals and lower rates of successful vaccination (15) as well as age-related accumulation of memory effector T cells (16), deficits of naı¨ve T cells (17), and central changes in bone marrow stem cell populations (18,19) leading to shifts in the balance of myeloid and lymphoid compartments. Thus, survival from glioma may depend on both patient age and specific immune characteristics. Measures of immune status that reflect individual immune aging may help disentangle the contribution of these interacting factors (20,21).

Both immune cell identification and understanding biological age have been areas of intensive epigenetic research that implicate the importance of DNA methylation. Immunomethylomics is an approach to immune cell profiling based on DNA methylation cytometry, which returns quantitative immune profile data using DNA methylation signatures specific for different immune cells (22-24). DNA-based immune profiling is reproducible; has a standardized workflow insensitive to batch and operator; and does not require intact cells, obviating the need for time-sensitive blood processing (which can include subjective gating procedures). Another line of epigenetic research has focused on DNA methylation-based epigenetic clocks as indicators of biological age (14). The initial bloodbased models of Hannum (25) and the multitissue algorithm by Horvath (26) were trained to predict chronological age and, as such, strongly correlate with chronological age. Subsequently, more clinically relevant biomarkers of aging have been explored by replacing prediction of chronological age with the prediction of a surrogate measure of phenotypic age to differentiate morbidity and mortality risk among same-age individuals (27). One such novel epigenetic age metric, DNAmPhenoAge, is based on 513 CpG sites and is a statistically significant predictor of morbidity and mortality; it is also correlated with immune profiles signaling immunosenescence (27). Epigenetic age based on blood has not been examined in glioma survival, although 1 study assessed it in glioma tumor tissue (28).

Here, we integrated cutting-edge bioinformatic techniques to profile immune phenotype and epigenetic age using genomescale DNA methylation measures. Our goal was to assess whether noninvasive epigenetic measures, such as blood immune profiles and epigenetic age, and interactions of such measures with clinical variables delineate survival risk groups among IDH/1p19q/TERT-WT glioma patients and whether such risk group characterizations could be applied to other glioma subtypes.

Methods

Patient Samples

The San Francisco Adult Glioma Study (AGS) is a case-control study that includes 3164 patients newly diagnosed with glioma between 1991 and 2012 who were residents of the San Francisco Bay Area or patients of the University of California San Francisco (UCSF) neuro-oncology clinic and 2140 people without glioma who were residents of the San Francisco Bay Area or seen in the UCSF phlebotomy clinic (29). This study was approved by the institutional review board of the UCSF Human Research Protection Program in the Office of Ethics and Compliance under UCSF federal-wide assurance 00000068. Informed consent was obtained from all study participants.

The study population consisted of 2 independent sets of AGS patients with glioma (ie, the training and evaluation sets) plus a group of 454 AGS participants without glioma (Figure 1, A). Among patients with blood samples and molecular tumor marker data, methylation arrays were run for 197 patients with IDH/1p19q/ TERT-WT (the training set) and 350 patients with other glioma subtypes (evaluation set: 41 IDH-MT gliomas [32 astrocytomas and 9 GBMs]; 130 IDH-WT gliomas [78 astrocytomas and 52 GBMs]; 141 patients with IDH-MT/1p19q co-deleted oligodendrogliomas; and 38 participants who could not be categorized according to WHO 2016) (Figure 1, A) (1,30). The 454 participants without glioma were partially matched to patients with IDH/1p19q/TERT-WT by age, sex, race, and AGS study phase. Blood samples were collected approximately 3 months after diagnosis.

DNA Methylation Array

Frozen anticoagulated whole blood was processed for DNA isolation and bisulfite conversion, as previously described (24). The Illumina 850K EPIC DNA methylation platform (Illumina, Inc, San Diego, CA) was used except for a subset of the evaluation set samples on the Illumina 450K array. Preprocessing and quality control of methylation array data are described in the Supplementary Methods (available online).

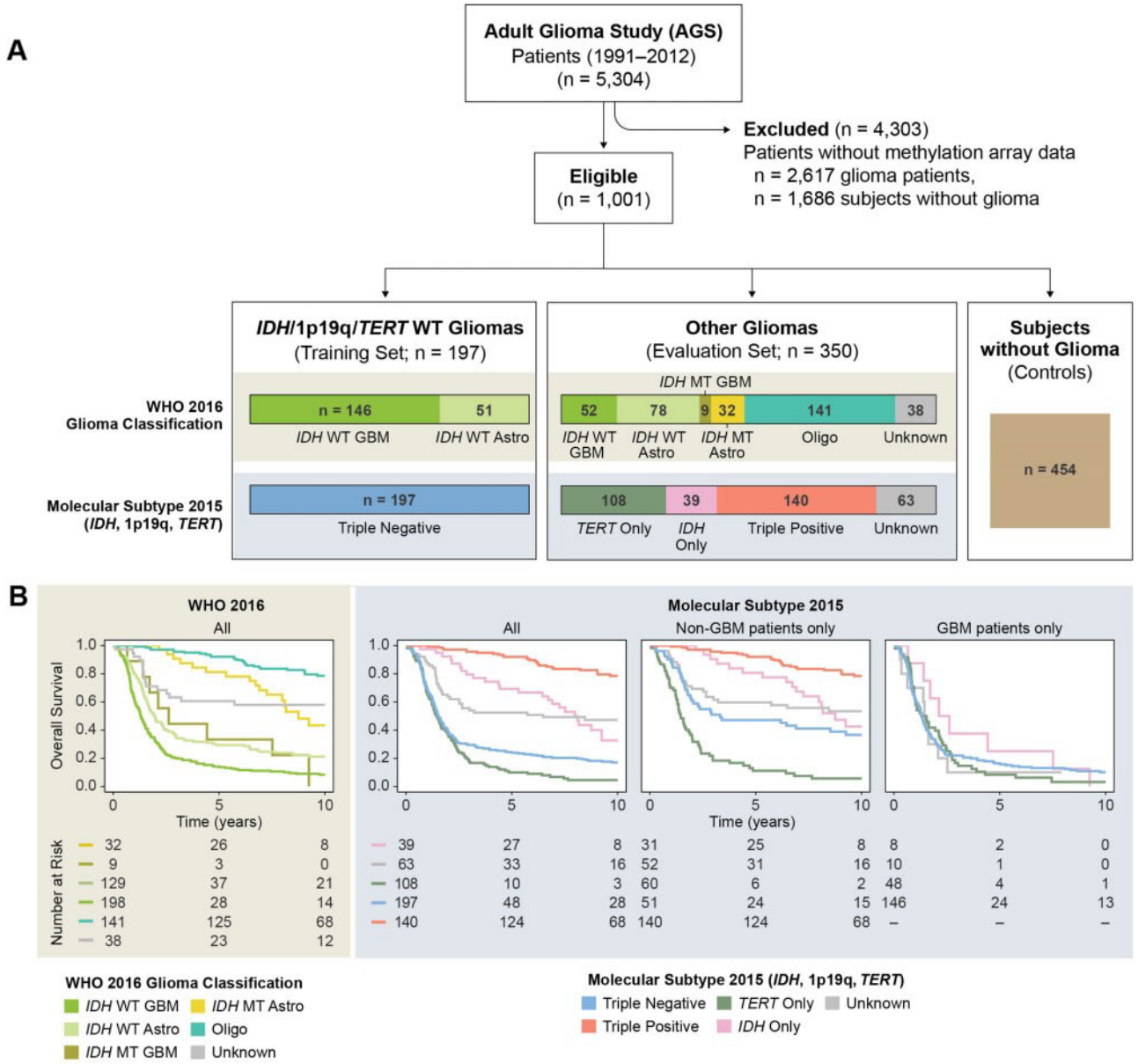
Immunomethylomic Assay

Using the preprocessed and normalized methylation data, we applied an optimized reference-based cell mixture deconvolution methodology (31) to the 3 cohorts to estimate the proportions of CD4þ T cells, CD8þ T cells, B cells, natural killer cells, monocytes, neutrophils, and 3 cell ratios (neutrophil-to-lymphocyte ratio [NLR], lymphocyte-to-monocyte ratio [LMR], and CD4þ T/CD8þ T cells).

Accuracy and reproducibility of cell deconvolution are previously described (23,24,32). A comparison of cell deconvolution using the 450K and 850K platforms run on the same 12-evaluationset patients indicated close correspondence of cell proportion estimates across the 2 platforms (Supplementary Figure 1, available online), and comparison of CD4þ T cell estimates by flow

Figure 1. Adult Glioma Study (AGS) glioma cases and participants without glioma. A) Data flow diagram for the AGS participants, with World Health Organization (WHO) 2016 glioma classification and molecular subtypes 2015[2], for IDH/1p19q/TERT-wild type (WT) gliomas (training set) vs other glioma groups (evaluation set). B) Survival curves for the AGS glioma cases by WHO 2016 glioma classification and molecular subtypes 2015[2] classification. Astro¼astrocytoma; GBM¼glioblastoma; Oligo¼oligodendroglioma.

Downloaded from https://academic.oup.com/jnci/article/114/3/463/6379711 by guest on 29 September 2022



cytometry vs cell deconvolution in 72 AGS patients illustrates their similarity (Supplementary Figure 2, available online).

DNAmAge, DNAmPhenoAge, and HannumAge Estimation

DNA methylation-based epigenetic clocks can be indicators of biological age (14). The 3 clocks evaluated were the initial bloodbased models of Hannum (25) (HannumAge), a multitissue algorithm to predict chronological age (26) (DNAmAge), and a subsequent algorithm (27) to predict a surrogate measure of phenotypic age to differentiate morbidity and mortality risk among same-age individuals (DNAmPhenoAge).

Statistical Analysis

OS was defined as the time from the date of diagnosis until the date of death (ie, event) or date of last follow-up (ie, censored). Cox proportional hazards models were used to evaluate associations of overall patient survival with baseline clinical variables, immune profiles, and WHO 2016 glioma classification (Table 1). Because of violations of the linearity and proportional hazards assumptions and our primary goal to examine interactions, we assessed combinations of clinical and epigenetic markers via partDSA (33,34), a recursive partitioning algorithm that accommodates censored data (see the Supplementary Methods, available online, for details). We built models of increasing complexity in the training set: first,

Table 1. Demographic and clinical characteristics of patients in the adult glioma study: training set of 197 IDH/1p/19q/TERT wild type patients, evaluation set of 350 patients with another glioma, and 454 participants without glioma

ARTICLE
ARTICLE
Downloaded from https://academic.oup.com/jnci/article/114/3/463/6379711 by Dartmouth College Library user on 29 September 2022

Variable	Training IDH/1p19q/TERT WT patients (n¼197)	Evaluation set Patients with another glioma (n¼350)	Control participants without glioma (n¼454)	Training vs control, P value	Training vs evaluation, P value	Evaltrols vs controls, P value
Clinical variables						
Chronological age, median (IQR), y	52.0 (41-64)	48.0 (40-56)	52.0 (41-63)	.91[a]	.002[a]	<.001[a]
Days between diagnosis and blood draw, median (IQR)	103 (72-161)	104.5 (33-182)	—	—	.63[a]	—
Sex, No. (%)				.20[b]	.71[b]	.27[b]
Female	79 (40.1)	146 (41.7)	207 (45.6)			
Male	118 (59.9)	204 (58.3)	247 (54.4)			
Smoking status, No. (%)				<.001[b]	.19[b]	.001[b]
Missing	0	2	1			
Current	26 (13.2)	35 (10.1)	61 (13.5)			
Ever/past	84 (42.6)	126 (36.2)	164 (36.2)			
Never	87 (44.2)	187 (53.7)	228 (50.3)			
Alcohol drinking status, No. (%)				<.001[b]	.003[b]	<.001[b]
Missing	0	2	1			
Current	116 (58.9)	169 (48.6)	226 (49.9)			
Ever/past	65 (33.0)	152 (43.7)	160 (35.4)			
Never	16 (8.1)	27 (7.8)	67 (14.8)			
BMI, median (IQR), kg/m^2	25.7 (23.3-29.1)	26.1 (23.2-29.5)	25.8 (23.3-29.8)	.16[a]	.41[a]	.44[a]
Race, No. (%)				.06[b]	.02[b]	<.001[a]
Asian	18 (9.1)	20 (5.7)	32 (7.0)			
Black	7 (3.6)	1 (0.3)	50 (11.0)			
Native American	1 (0.5)	1 (0.3)	2 (0.4)			
Other	9 (4.6)	11 (3.1)	17 (3.7)			
Pacific Islander	0 (0.0)	1 (0.3)	1 (0.2)			
White	162 (82.2)	316 (90.3)	352 (77.5)			
Vital status, No. (%)				—	<.001[b]	—

Alive	28 (14.2)	145 (41.4)	–			
Dead	169 (85.8)	205 (58.6)	–			
Follow-up, median (95% CI), y	15 (12.9 to 19.0)	12.3 (10.6 to 13.7)	–			
WHO 2016 Glioma Classification, No. (%)				–	<.001[b]	–
Missing	0	38	–			
IDH-WT glioblastoma	146 (74.1)	52 (16.7)	–			
IDH-WT astrocytoma	51 (25.9)	78 (25.0)	–			
IDH mutant glioblastoma	0 (0.0)	9 (2.9)	–			
IDH mutant astrocytoma	0 (0.0)	32 (10.3)	–			
IDH mutant/1p19q co-deleted oligodendroglioma	0 (0.0)	141 (45.2)	–			
Glioma diagnosis grade, No. (%)				–	<.001[b]	
Missing	0	1	–			
Grade 2	18 (9.1)	163 (46.7)	–			
Grade 3	33 (16.8)	120 (34.4)	–			
Grade 4	146 (74.1)	66 (18.9)	–			
Molecular subtype (based on IDH, 1p19q, and TERT)[c]*, No. (%)				–	<.001[b]	–
Missing	0 (0.0)	63 (18.0)	–			
IDH only	0 (0.0)	39 (11.1)	–			
TERT only	0 (0.0)	108 (30.9)	–			
Triple negative	197 (100.0)	0 (0.0)	–			
Triple positive	0 (0.0)	140 (40.0)	–			
Surgery, No. (%)				–	.48[b]	–
Biopsy only	26 (13.2)	39 (11.1)	–			
Surgery	171 (86.8)	311 (88.9)	–			
Chemotherapy, No. (%)				–	.001[b]	–
Missing	0	3	–			

No	41 (20.8)	117 (33.7)	—			
Yes	156 (79.2)	230 (66.3)	—			

(continued)

Table 1. (continued)

Variable	Training IDH/1p19q/TERT WT patients (n¼197)	Evaluation set Patients with another glioma (n¼350)	Control participants without glioma (n¼454)	Training vs control, P value	Training vs evaluation, P value	Evaluation vs controls, P value
Radiation, No. (%)				—	<.001[b]	—
No	15 (7.6)	141 (40.3)	—			
Yes	182 (92.4)	209 (59.7)	—			
Treatment, No. (%)				—	<.001[b]	—
Missing	0	3	—			
Both chemotherapy and radiation	152 (77.2)	160 (46.1)	—			
Chemotherapy only	4 (2.0)	70 (20.2)	—			
Radiation only	30 (15.2)	47 (13.5)	—			
Neither	11 (5.6)	70 (20.2)	—			
Taking dexamethasone at blood draw, No. (%)				<001[b]	<.001[b]	<.001[b]
Missing	1	6	6			
No	97 (49.5)	229 (66.6)	447 (99.8)			
Yes	99 (50.5)	115 (33.4)	1 (0.2)			
Epigenetic age variable DNAmAge, median (IQR), y	57.4 (47.0-66.0)	54.8 (46.6-63.0)	56.7 (44.9-66.4)	.51[a]	.17[a]	.46[a]
DNAmPhenoAge, median (IQR), y	48.7 (36.9-58.1)	42.9 (34.0-51.8)	43.3 (31.5-54.9)	<.001[a]	<.001[a]	.55[a]
HannumAge, median (IQR), y	45.8 (35.9-53.7)	45.0 (35.5-53.2)	42.9 (32.9-52.7)	.04[a]	.44[a]	.10[a]
DNAmPhenoAge acceleration with IPs, median (IQR), y	0.3 (-3.1-3.7)	0.9 (3.5 to 3.1)	0.0 (3.6 to 3.7)	.37[a]	.08[a]	.32[a]
DNAmPhenoAge acceleration without IPs, median (IQR), y	2.6 (1.8 to 6.5)	0.8 (3.0 to 5.1)	2.0 (6.4 to 2.4)	<.001[a]	.003[a]	<.001[a]
HannumAge acceleration with IPs, median (IQR), y	1.0 (4.2 to 1.3)	0.3 (2.4 to 3.2)	0.1 (2.6 to 2.5)	.003[a]	<.001[a]	.09[a]
HannumAge acceleration without IPs, median (IQR), y	0.4 (2.4 to 2.5)	1.5 (1.8 to 5.9)	2.1 (4.8 to 0.4)	<.001[a]	<.001[a]	<.001[a]
DNAmAge acceleration with IPs, median (IQR), y	1.1 (3.7 to 1.9)	1.0 (2.4 to 4.1)	0.6 (3.7 to 2.8)	.89[a]	.002[a]	<.001[a]
DNAmAge acceleration without IPs, median (IQR), y	0.8 (3.5 to 1.9)	1.3 (1.8 to 4.8)	1.4 (4.2 to 2.2)	.15[a]	<.001[a]	<.001[a]

ARTICLE

Downloaded from https://academic.oup.com/jnci/article/114/3/446/6379711 by Dartmouth College Library user on 29 September 2022

IP variable						
B cell, median (IQR)	2.1 (1.3-3.4)	3.3 (2.0-4.9)	5.3 (4.0-7.0)	<.001[a]	<.001[a]	<.001[a]
Monocyte, median (IQR)	7.5 (6.0-9.5)	7.7 (5.9-10.0)	7.4 (6.0-9.4)	.28[a]	.74[a]	.07[a]
Neutrophils, median (IQR)	70.4 (61.7-80.5)	64.5 (55.4-71.9)	57.9 (51.0-65.5)	<.001[a]	<.001[a]	<.001[a]
NK, median (IQR)	2.9 (1.9-4.5)	3.3 (1.9-5.1)	4.7 (3.4-6.2)	<.001[a]	.37[a]	<.001[a]
CD4-positive T, median (IQR)	8.9 (5.2-14.2)	12.5 (8.2-17.8)	15.7 (11.8-19.7)	<.001[a]	<.001[a]	<.001[a]
CD8-positive T, median (IQR)	6.3 (3.4-9.6)	7.3 (4.4-11.4)	8.9 (6.1-12.6)	<.001[a]	.002[a]	<.001[a]
CD4-positive T/CD8-positive T, median (IQR)	1.5 (0.9-2.6)	1.6 (1.1-2.4)	1.7 (1.2-2.6)	.16[a]	.08[a]	.19[a]
LMR, median (IQR)	2.8 (1.9-4.2)	3.8 (2.3-5.3)	4.8 (3.8-6.4)	<.001[a]	<.001[a]	<.001[a]
NLR, median (IQR)	3.2 (2.0-5.9)	2.3 (1.4-3.7)	1.6 (1.2-2.2)	<.001[a]	.002[a]	<.001[a]
Total lymphocytes, median (IQR)	21.3 (13.8-30.6)	28.3 (19.7-37.8)	37.1 (29.4-43.9)	<.001[a]	<.001[a]	<.001[a]

[a]Linear model ANOVA. All tests were 2-sided. ANOVA¼analysis of variance; BMI¼body mass index; CI¼confidence interval; IP¼immune profile; IQR¼interquartile range; LMR[b] 2¼lymphocyte-to-monocyte ratio; NK¼natural killer; NLR¼neutrophil-to-lymphocyte ratio; WHO¼World Health Organization; WT¼wild type. Pearson v test. All tests were 2-sided.

[c]Patients were categorized into 5 glioma groups based on 3 tumor molecular markers: IDH, 1p19q, and TERT. Details of how these were categorized can be found in Eckel-Passow et al. (2).

Clinical Variables Only; then, Clinical and Immune Profiles; finally, Clinical, Immune Profiles, and Epigenetic Ages. Subsequently, the risk groups' generalizability was assessed by the evaluation set—see the Supplementary Methods (available online) for details. For training and evaluation sets, within- and between-group summary measures were based on the restricted mean survival time (RMST), which does not require the proportional hazards assumption (35). Two-sided tests of statistical significance were used, and P<.05 was selected as the cutoff for statistical significance. All analyses were performed in R, version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria) (36).

Results

Clinical and Demographic Characteristics of Patients With Glioma and Controls

Figure 1 shows the 3 sets of participants and molecular classifications for patients with glioma. Clinical and demographic characteristics are in Table 1. Because of frequency matching for choosing participants without glioma, age, race, and sex were similar to those of IDH/1p19q/TERT-WT glioma patients. Fewer participants without glioma than IDH/1p19q/TERT-WT glioma patients consumed alcohol, smoked, or took dexamethasone at blood draw. Evaluation set patients were younger, less likely to consume alcohol at blood draw, predominantly White, less likely to receive chemotherapy or radiation, and less likely to be on dexamethasone than the training set patients. Compared with the participants without glioma, the other patients with glioma were younger, less likely to be current smokers, more likely to have consumed alcohol, and more likely to be White.

Blood Immune Profiles Alterations in IDH/1p19q/TERT-WT Patients

The IDH/1p19q/TERT-WT patients had increased fractions of neutrophils and lower proportions of B cells, CD4^{þ} T, CD8^{þ} T, natural killer cells, and total lymphocytes compared with participants without glioma (all P<.001; Figure 2 and Table 1). The NLR was higher among IDH/1p19q/TERT-WT patients compared with participants without glioma (median [interquartile range (IQR)]¼3.2 [2.0-5.9] vs 1.6 [1.2-2.2], P<.001), while LMR was lower (median [IQR]¼2.8 [1.9-4.2] vs 4.8 [3.8-6.4], P<.001), and the CD4^{þ} T–to–CD8^{þ} T cell ratio was not statistically significantly different (median [IQR]¼1.5 [0.9-2.6] vs 1.7 [1.2-2.6], P¼.16).

There were consistent differences in leukocyte proportions for grade 4 vs grade 2 to 3 IDH/1p19q/TERT-WT tumors (Supplementary Table 1, available online). Further stratification of the 196 IDH/1p19q/TERT-WT patients with dexamethasone status known at blood draw revealed a greater neutrophil proportion; higher NLR; and lower CD4^{þ} T, CD8^{þ} T, natural killer– cell, and B-cell proportions among dexamethasone-exposed vs nonexposed patients (Supplementary Table 2, available online). Similar relationships were seen in the other patients with glioma (Supplementary Tables 3 and 4, available online).

The IDH/1p19q/TERT-WT patients also had an increased fraction of neutrophils and higher NLR than the other patients with glioma and decreased B cells, CD4^{þ} T cells, CD8^{þ} T cells, total lymphocytes, and LMR (Table 1). The distributions of immune profiles were statistically different between the other patients with glioma and the participants without glioma, except for monocytes (Table 1).

Correlations Between Epigenetic Age and Immune Profiles in IDH/1p19q/TERT-WT Patients and Participants Without Glioma

Among the 3 DNA methylation age estimates, DNAmPhenoAge and HannumAge were higher in patients with IDH/1p19q/TERTWT than in participants without glioma (Table 1). The DNAmPhenoAge and HannumAge age acceleration were also increased. Once adjusted for leukocyte cell type composition, only HannumAge age acceleration remained statistically significant (median [IQR]¼–1.0 [–4.2 to 1.3] vs 0.1 [–2.6 to 2.5], P¼.003). Both in IDH/1p19q/TERT-WT patients and in participants without glioma, chronological age was modestly negatively correlated with lymphocyte subsets and positively associated with neutrophils (Supplementary Figure 3, available online). Stronger associations were observed for the HannumAge and DNAmPhenoAge measures.

Survival in IDH/1p19q/TERT-WT Patients and Variations in Blood Immune Profiles

The median OS for the 197 IDH/1p19q/TERT-WT patients was 1.5years (95% CI¼1.4 to 1.8) (Figure 1, B and Supplementary Table 5, available

Downloaded from https://academic.oup.com/jnci/article/114/3/463/6379711 by [illegible] College Library user on 29 September 2022

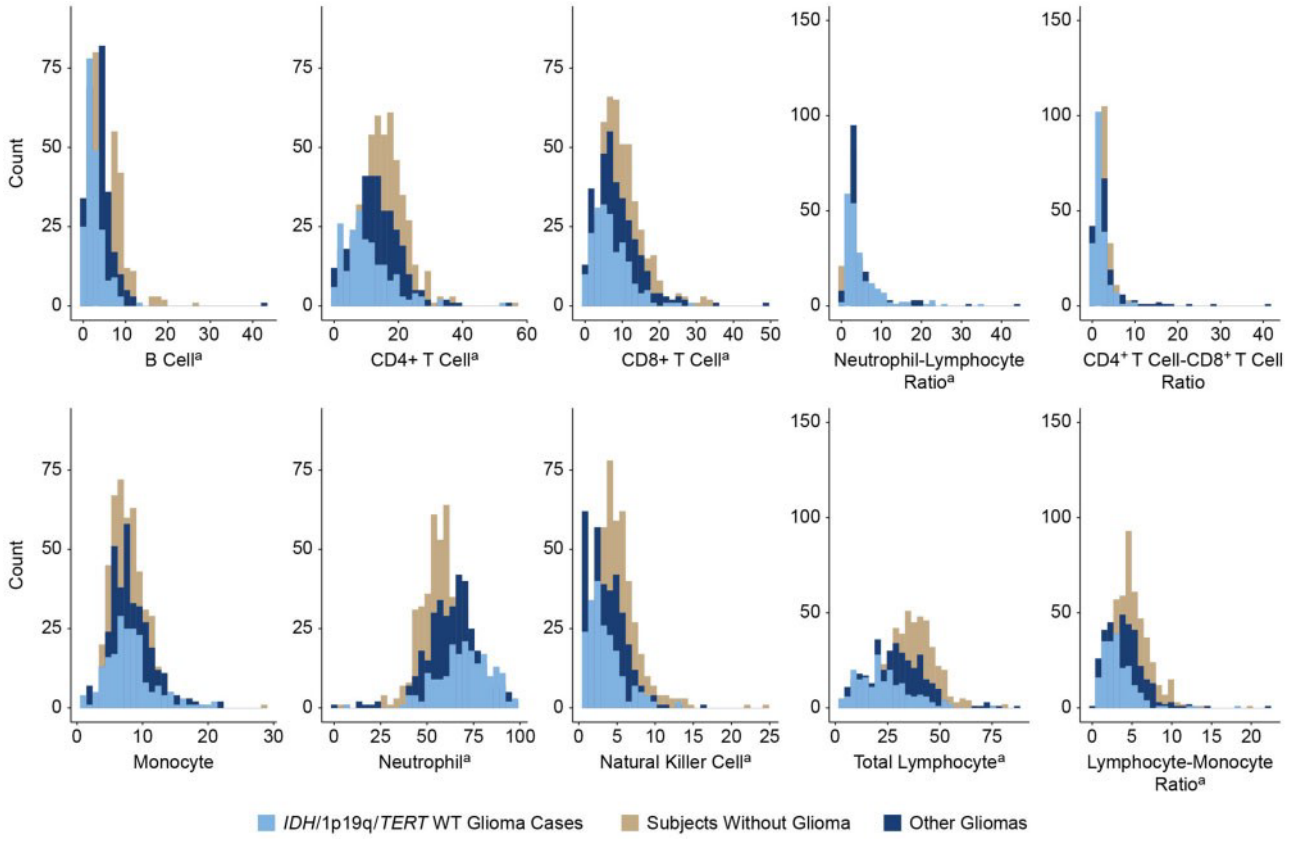


Figure 2. Comparisons of immune cell fractions in blood from the training set of 197 IDH/1p19q/TERT-wild type (WT) patients, evaluation set of 350 other glioma groups, and 454 participants without glioma.

[a] Indicates a statistically significant difference in the specified immune cell fractions via analysis of variance.

online). Split by the WHO 2016 classification, the IDH-wildtype GBMs had a median OS of 1.3years (95% CI¼1.2 to 1.52), and the IDH-WT astrocytomas had 3.3years (95% CI¼1.97 to NA). Univariate survival models are in Figure 3, A and Supplementary Table 6 (available online).

Clinical Variables Only. Interactions were explored by constructing survival recursive partitioning analyses (RPAs), including univariately statistically significant clinical variables (Figure 3, B). Important interactions among age, dexamethasone use, and body mass index (BMI) were elucidated and separated the patients into 4 mutually exclusive risk groups (Supplementary Figure 4, A, available online). The patients with the best survival were younger, not taking dexamethasone, and had a lower BMI (group 1, gold: n¼34, median OS¼9.19years [95% CI¼6.18 to NA]). The second-best survival risk group included the younger patients not taking dexamethasone but with higher BMI (group 2, tan: n¼26, median OS¼3.33years [95% CI¼1.74 to NA]). The second-worst survival group included 2 subgroups: younger patients taking dexamethasone and older patients with a low BMI (group 3, gray: n¼89, median OS¼1.47years [95% CI¼1.31 to 1.69]). The patients who had the worst survival were older, with a higher BMI (group 4, blue: n¼47, median OS¼0.78years [95% CI¼0.67 to 0.99]). RMST estimates for the group assignment are given in Table 2 (Model A: Clinical Variables).

Clinical Variables and Immune Profiles. We then added immune profile variables to the clinical variables (Figure 3, B). Interactions among age, CD4þ T cell, and neutrophil proportions separated the patients into 4 mutually exclusive survival groups (Figure 4, A). The patients with the best survival were younger, with a high CD4þ T-cell proportion (group 1, gold: n¼33; median OS¼9.72years [95% CI¼6.18 to NA]). The second-best survival risk group included young patients with a low CD4þ T-cell proportion (group 2, tan: n¼93; median OS¼1.75years [95% CI¼1.50 to 2.06]). The second-worst survival was experienced by those who were older and had a low neutrophil proportion (group 3, gray: n¼43, median OS¼1.27years [95% CI¼0.83 to 1.64]). The patients who had the worst survival included those who were older and had high neutrophil proportions (group 4, blue: n¼28, median OS¼0.76years [95% CI¼0.55 to 0.99]). Kaplan-Meier survival curves show the discrimination of survival groups (Figure 4, B). IDH-WT GBM comprises 96.4% of group 4, 81.4% of group 3, 74.2% of group 2, and 45.5% of group 1 (Figure 4, A; Supplementary Table 7, available online). KaplanMeier curves are in Figure 4, B, and RMST estimates are in Table 2 (Model B: Clinical þ Immune Profiles). Blood draw
treatment time points were included in the survival RPA, but this variable did not statistically significantly affect the survival results (Supplementary Table 8, available online).

Clinical Variables, Immune Profiles, and Epigenetic Ages. Next, we included HannumAge and DNAmPhenoAge and the clinical and immune profile data (Figure 3, B). Important interactions among HannumAge, CD4þ T cells, and neutrophil proportions were identified (Supplementary Figure 5, A, available online). Patients fell into 1 of 4 risk groups. The patients with the best survival had a younger HannumAge and a high CD4þ T-cell proportion (group 1, gold: n¼31, median OS¼14.0years [95% CI¼8.33 to NA]). The second-best survival risk group included younger HannumAge, with a low

ARTICLE

Downloaded from https://academic.oup.com/jnci/article/114/3/463/6379711 by Dartmouth College Library user on 29 September 2022

CD4þ T-cell proportion (group 2, tan: n¼77, median OS¼1.97years [95% CI¼1.54 to 2.91]). The second-worst survival was experienced by those who were older by HannumAge and had a lower neutrophil proportion (group 3, gray: n¼52, median OS¼1.27years [95% CI¼0.89 to 1.69]). The older patients by HannumAge who had a higher neutrophil proportion had the worst survival (group 4, blue: n¼37, median OS¼0.79years [95% CI¼0.63 to 0.99]). Kaplan-Meier curves are shown in Supplementary Figure 5, B (available online), and RMST estimates are in Table 2 (Model C: Clinical þ Immune Profiles þ HannumAge). An interaction between DNAmPhenoAge and years of education was identified (Supplementary Figure 6, A, available online). Kaplan-Meier curves are shown in Supplementary Figure 6, B (available online) and RMST estimates are in Table 2 (Model C: Clinical þ Immune Profiles þ DNAmPhenoAge).

Survival Among Other Patients With Glioma. The 350 other patients with glioma had a median OS of 7.5years (95% CI¼6.1 to 10.6). Survival by WHO 2016 and Molecular Subtype 2015 guidelines are shown in Figure 1, B and Supplementary Table 5

(available online). We assessed whether the same variables and cut points built by the 4 RPA models also distinguished survival among the other glioma cases. Only the RPA from the Clinical Variable and Immune Profiles (Figure 4, C and D) distinguished 4 survival groups based on the same cut points generated by the training set, whereas in the other 3, 2 of the 4 groups collapsed (Supplementary Figures 4, C; 5, C; and 6, C, available online; Table 2). The Kaplan-Meier curves in Figure 4, D, estimate the median OS for group 4 as 0.7years (95% CI¼0.4 to 1.4), for group 3 as 2.5years (95% CI¼1.6 to 5.7), for group 2 as 8.2years (95% CI¼6.2 to NA), and for group 1 as 13.6years (95% CI¼10.6 to 18.2). The RMST estimates are in Table 2. Oligodendrogliomas comprise 0% of group 4, 30.8% of group 3, 35.9% of group 2, and 56.1% of group 1 (Figure 4, C; Supplementary Table 9, available online). Groups by Molecular Subtype 2015 are in Supplementary Figure 7 and Supplementary Table 9 (available online).

Discussion

In this study, we explored the utility of blood DNA methylationbased measures of age and immune cells to create prognostic patient subgroups that may augment traditional WHO classification. Consistent with past studies in the IDH-WT gliomas, age at diagnosis (chronological or epigenetic) figured prominently in

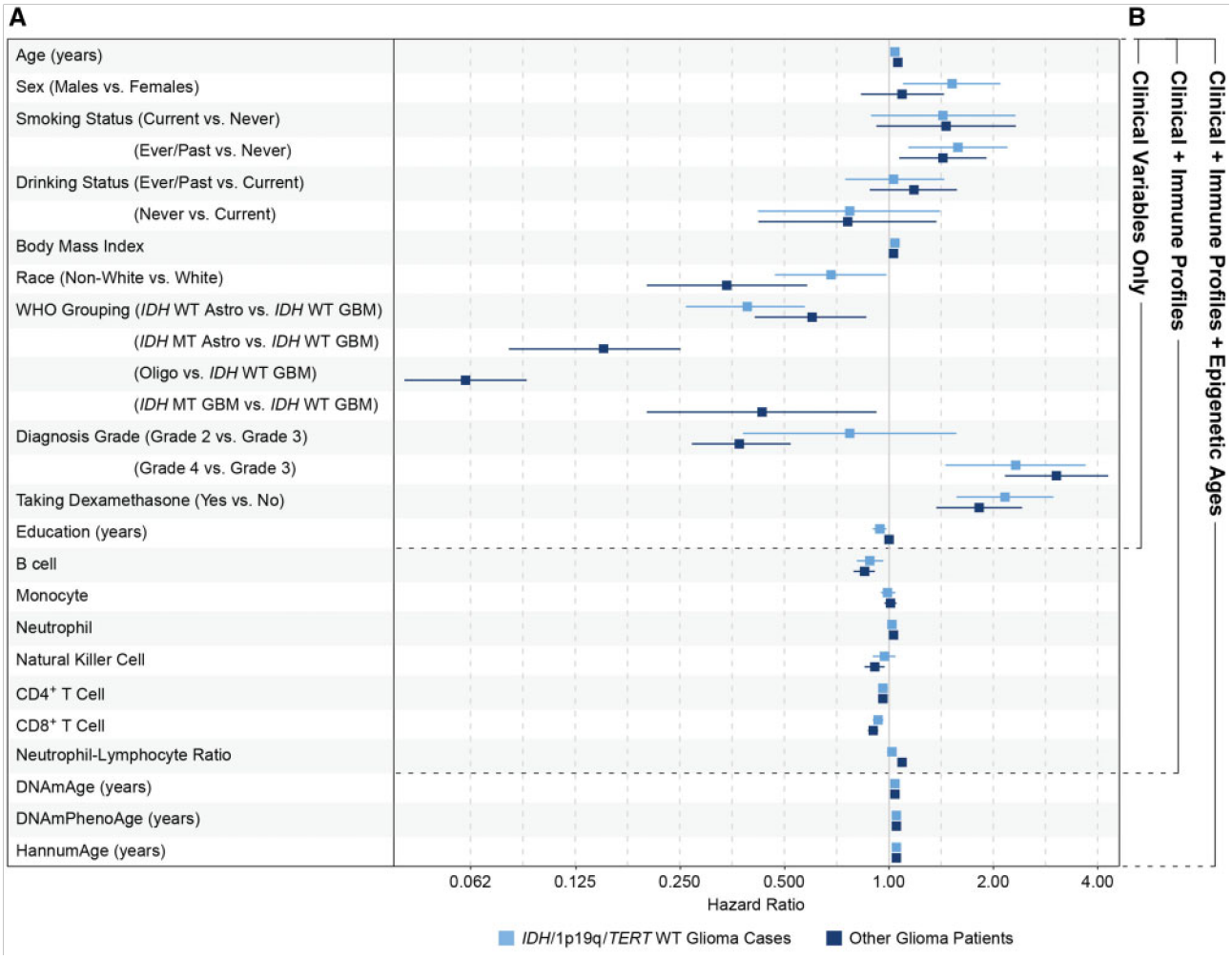


Figure 3. Forest plot of hazard ratios for univariate survival models and recursive partitioning analysis (RPA) model inclusion. A) The hazard ratios (squares) and 95% confidence intervals (error bars) for the training set of 197 IDH/1p19q/TERT-wild type (WT) gliomas in light blue and the 350 other glioma groups in dark blue. For the IDH/1p19q/TERT-WT gliomas, violations of the proportional hazards assumption occurred in the univariate models for chronological age, monocytes, DNAmAge, DNAmPhenoAge, and HannumAge. For the other gliomas, violations of the proportional hazards assumption occurred in the univariate models for chronological age, World Health Organization (WHO) grouping, diagnosis grade, taking dexamethasone, monocytes, neutrophils, natural killer cells, CD4-positive T cells, CD8-positive T cells, DNAmAge, DNAmPhenoAge, and HannumAge. B) Variable groupings depict which variables are included in each of the 3 RPA models: clinical variables only; clinical variables þ immune profiles (IPs); clinical variables, IPs, and epigenetic ages. Astro¼astrocytoma; GBM¼glioblastoma; Oligo¼oligodendroglioma.

survival models (11). Although strongly correlated with chronological age, epigenetic-based age algorithms are designed to represent multisystem proxies of physiologic dysregulation. Such markers could convey patient-specific information about disease outcomes, but these epigenetic age measures were not consistently informative for glioma survival because neither the HannumAge nor the DNAmPhenoAge RPAs generalized to the other glioma patients. Nonetheless, in these models, immune cell proportions were associated with survival. Given the correlation between immune cell proportions and epigenetic age, epigenetic age appears to be a proxy of immune status, the latter being better captured by the immune cell proportions themselves.

Specifically, we showed that proportions of CD4 T cells and neutrophils and patient chronological age divided both sets of patients into 4 groups with statistically significantly different survival experiences. In both the training and evaluation sets, younger patients (aged 58years) with higher $CD4^{þ}$ T-cell proportions and older patients (aged >58years) with lower neutrophil proportions both had better survival. The reciprocal increase in myeloid cells and decreased lymphocytes associated with worse survival are a recurring theme in many cancers (37).

Although the same factors distinguished survival in the 2 independent sets of patients, the groups' median survival times were different, reflecting the different underlying survival of participants according to glioma subtype. Interestingly, because of the inclusion of subgroups with better survival, there were somewhat greater differences in the survival times of the patients in the evaluation set than in the training set. Notably, among the IDH/1p19q/TERT-WT patients, although a large majority of the older patients with high neutrophils (ie, the highest-risk group), were IDH-WT GBM, as expected; almost half of the lowest risk group (ie, the younger patients with high $CD4^{þ}$ T cells) were also IDH-WT GBM. Similarly, in the other patients with glioma, although just over half of the lowest-risk group had oligodendrogliomas, one-third of the third-highest risk group (ie, the older patients with lower neutrophil proportion) also had oligodendrogliomas. Finally, patients whose tumors could not be assigned to a WHO subtype still were distributed across each of the 4 risk groups. Therefore, an

Downloaded from https://academic.oup.com/jnci/article/114/3/463/6379711 by guest on 29 September 2022

Table 2 Recursive partitioning analysis models: training set of 197 IDH/1p19q/TERT wild type patients and evaluation set of 350 patients with another glioma from the adult glioma study

Model description	Group[a]	Training set (n¼197) RMST (95% CI)	Training set (n¼197) P value[b]	Evaluation set (n¼350) RMST (95% CI)	Evaluation set (n¼350) P value[b]
Model A: clinical variables[c]	Clinical-only tree				
	Best survival (group 1, gold)	9.14 (7.17 to 11.12)	.12	10.36 (9.12 to 11.61)	.68
	Second-best survival (group 2, tan)	6.66 (4.26 to 9.07)	(Referent)	10.01 (8.89 to 11.13)	(Referent)
	Second-worst survival (group 3, gray)	2.86 (2.08 to 3.65)	.003	6.82 (5.70 to 7.94)	<.001
	Worst survival (group 4, blue)	1.40 (0.76 to 2.05)	<.001	3.91 (2.51 to 5.30)	<.001
Model B: clinical þ IPs[d]	Clinical þ IPs tree				
	Best survival (group 1, gold)	9.33 (7.32 to 11.34)	<.001	10.09 (9.08 to 11.11)	.03
	Second-best survival (group 2, tan)	4.07 (3.08 to 5.07)	(Referent)	8.50 (7.53 to 9.48)	(Referent)
	Second-worst survival (group 3, gray)	2.07 (1.16 to 2.98)	.004	5.18 (3.68 to 6.68)	<.001
	Worst survival (group 4, blue)	1.15 (0.36 to 1.95)	<.001	0.83 (0.54 to 1.12)	<.001
Model C: clinical[e] þ IPs þ HannumAge	Clinical þ IPs þ HannumAge tree				
	Best survival (group 1, gold)	9.74 (7.68 to 11.80)	<.001	10.18 (9.10 to 11.27)	.90
	Second-best survival (group 2, tan)	4.69 (3.54 to 5.85)	(Referent)	10.08 (8.87 to 11.30)	(Referent)
	Second-worst survival (group 3, gray)	1.98 (1.22 to 2.73)	<.001	6.44 (5.32 to 7.56)	<.001
	Worst survival (group 4, blue)	1.13 (0.52 to 1.73)	<.001	2.66 (1.63 to 3.69)	<.001
Model D: clinical þ IPs[f] þ DNAmPhenoAge	Clinical þ IPs þ DNAmPhenoAge tree				
	Best survival (group 1, gold)	8.07 (6.50 to 9.64)	<.001	10.65 (9.73 to 11.58)	<.001
	Second-best survival (group 2, tan)	3.58 (2.54 to 4.63)	(Referent)	6.72 (5.66 to 7.78)	(Referent)
	Second-worst survival (group 3, gray)	1.64 (0.96 to 2.33)	.003	6.98 (5.50 to 8.46)	.78
	Worst survival (group 4, blue)	1.29 (0.30 to 2.28)	.002	2.80 (0.99 to 4.61)	.002

[a] Model A: Group 1 (gold) were 58years of age, were not taking dexamethasone at the time of blood draw, and had a BMI $_2$ 26kg/m . Group 2 (tan) were 58$_2$years of age, were not taking dexamethasone at the time of blood draw, and had a BMI >26kg/m . Group 3 (gray) were either >58years of age and had a BMI 24kg/m or were $_2$ 58years of age and took dexamethasone at the time of blood draw. Group 4 (blue) were >58 years of age and had a BMI >24kg/m . Model B: Group 1 (gold) were 58years of age and had a CD4-positive T-cell count >14. Group 2 (tan) were 58years of age and had CD4-positive T-cell count 14. Group 3 (gray) were >58years of age and had a neutrophil count 77. Group 4 (blue) were >58years of age and had a neutrophil count >77. Model C: Group 1 (gold) consisted of the 31 patients who were 47years of age by HannumAge and had CD4-positive T-cell count >14. Group 2 (tan) were 47years of age by HannumAge and had CD4-positive T-cell count 14. Group 3 (gray) were >47years of age by HannumAge and had a neutrophil count 74. Group 4 (blue) were >47years of age by HannumAge and had a neutrophil count >74. Model D: Group 1 (gold) were 39.5years of age by DNAmPhenoAge. Group 2 (tan) were between 39.5 and 63.2years of age by DNAmPhenoAge and had >15years of education. Group 3 (gray) were between 39.5 and 63.2years of age by DNAmPhenoAge and had 15years of education. Group 4 (blue) were >63.2years of age by DNAmPhenoAge. BMI¼body mass index; CI¼confidence interval; IP¼immune profile; RMST¼restricted mean survival time; RPA¼recursive partitioning analysis.

[b] Two-sided P values were calculated using the Wald test.

[c] RMST for RPA in Supplementary Figure 4, A (available online).

[d] RMST for RPA in Figure 4, A.

[e] RMST for RPA in Supplementary Figure 5, A (available online).

[f] RMST for RPA in Supplementary Figure 6, A (available online).

interaction between chronological age and immune cell composition may fine-tune the current prognostic molecular classification systems and, if validated in other studies, may be critical when molecular data are insufficient for classification.

One of this study's limitations was that blood samples were obtained after diagnostic surgery (on average within 3 months of diagnosis), not at uniform times relative to other treatments. Blood draw timing did not affect our

ARTICLE

Downloaded from https://academic.oup.com/jnci/article/114/3/446/6379711 by Dartmouth College Library user on 29 September 2022

analyses' outcome (Supplementary Table 8, available online). Other limitations are that the 6 immune factors are cell proportions rather than absolute cell counts and do not capture additional subtypes. Finally, we had incomplete data on some tumor markers, such as EGFR, TP53, CDKN2A/B, and MGMT, but our evaluation set of patients with glioma represented the various WHO 2016 glioma subtypes not represented in the training set. We also acknowledge that some of the variability in the survival data could have been the result of potential diagnostic misclassification despite the use of several relevant tumor markers. For example, studies have shown that classification by tumor methylation arrays can change diagnoses in up to 12% of patients with glioma and is becoming a more standard diagnostic approach where available
(38). It will be many years, however, until cohorts of patients with glioma diagnosed using today's methylation arrays or other, new diagnostic paradigms have sufficient events for rigorous outcomes studies. Thus, the use of the available markers on this historic, well-characterized cohort is of value now and in future glioma research.

Abnormalities in peripheral blood immune profiles have long been observed in patients with glioma (39-42) and associated with glioma outcomes (30,43). The current studies illustrate how peripheral blood is a rich source of immune biomarker information, reflecting the dynamic interactions of tumor and host and the beneficial or deleterious effects of therapeutic interventions (44,45). The critical impact of the systemic immune compartment on antitumor immunity has been highlighted in recent studies of checkpoint inhibition therapy in animal models (46) and humans (47-49). These observations' potential relevance for patients with GBM was illustrated in a recent neoadjuvant programmed death-ligand 1 blockade trial that found response rates linked with pretherapy (baseline) blood immune profiles (50). DNA methylation has significant potential to contribute to these efforts and provide noninvasive

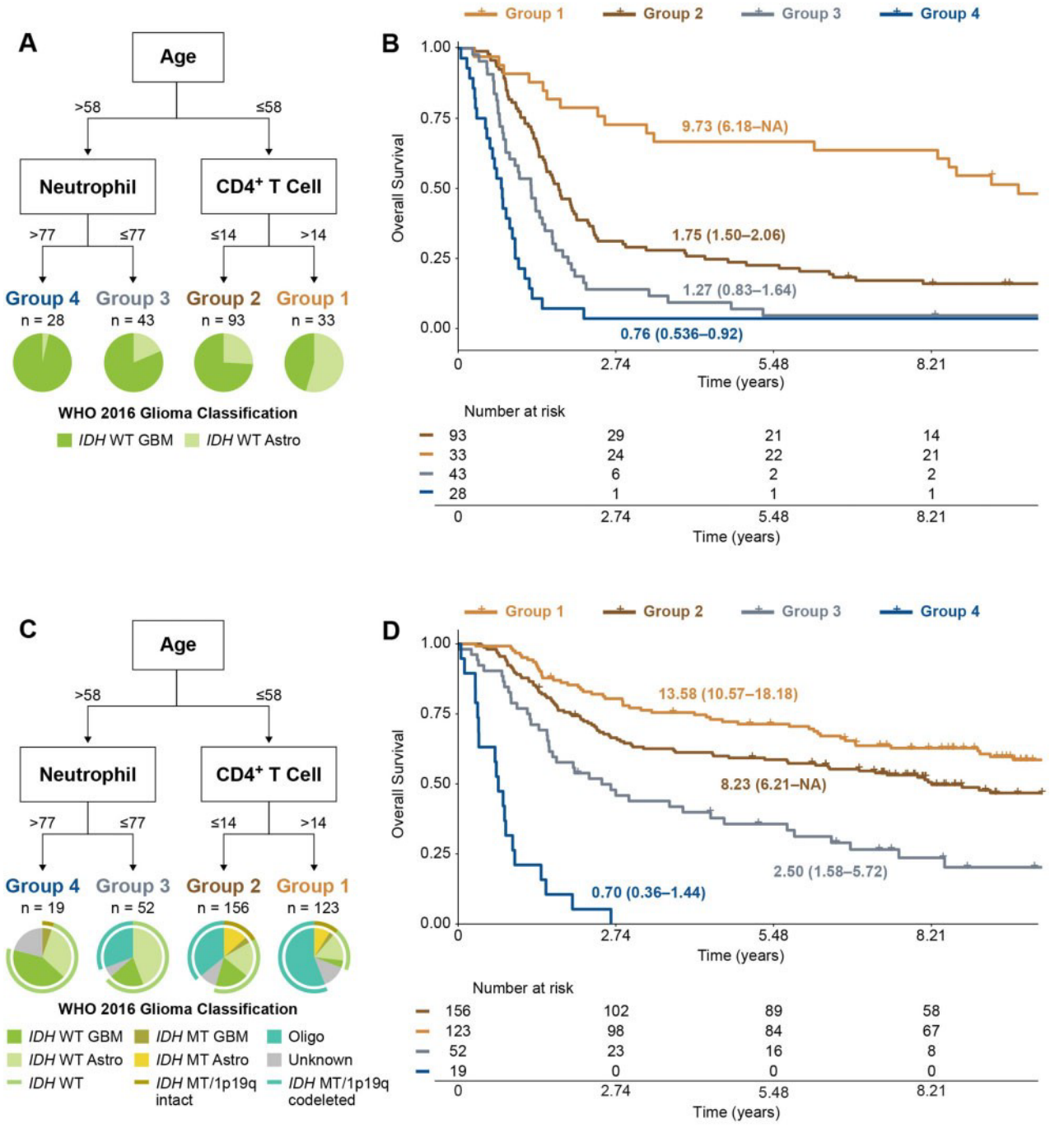


Figure 4. Clinical 1 immune profiles (IPs) recursive partitioning analysis (RPA), and Kaplan-Meier curves in the training and evaluation sets A) For the clinical þ IPs RPA in the training set of 197IDH/1p19q/TERT-wild type (WT) patients, chronological age was the RPA's primary node, with neutrophil counts and CD4 T-cell counts as secondary nodes. Patients fell into 1 of 4 risk groups: Group 1 (gold) consisted of the 33 patients who were 58 years of age and had CD4 T-cell counts>14, with a median overall survival (OS) of 9.73years (95% confidence interval [CI]¼6.18 to NA). Group 2 (tan) consisted of the 93 patients who were 58 years of age and had CD4-positive T-cell counts 14, with a median OS of 1.75years (95% CI¼1.50 to 2.06). Group 3 (gray) consisted of the 43 patients who were>58 years of age and had a neutrophil count 77, with a median OS of 1.27years (95% CI¼0.83 to 1.64). Group 4 (blue) consisted of the 28 patients >58years of age with a neutrophil count >77 and a median OS of 0.76years (95% CI¼0.55 to

Downloaded from https://academic.oup.com/jnci/article/114/3/463/6379711 by Dartmouth College Library user on 29 September 2022

0.99). World Health Organization (WHO) 2016 glioma classifications for 4 groups are shown in pie charts. B) Kaplan-Meier curves are shown for the training set based on the clinical þ IPs RPA. Kaplan-Meier curves and risk table in the training set of 197 IDH/1p19q/TERT-WT patients for risk groups 1-4 are defined in Figure 4, A. C) The evaluation set of 350 patients with another glioma for risk groups 1-4 are defined by the training set RPA in Figure 4, A. Evaluation set patients fell into 1 of 4 risk groups defined by the training set. Group 1 (gold) consisted of the 123 patients who were 58years of age and had CD4 T-cell count >14, with a median OS of 13.58years (95% CI¼10.57 to 18.18). Group 2 (tan) consisted of the 156 patients who were 58years of age and had a CD4 T-cell count 14, with a median OS of 8.26years (95% CI¼6.21 to NA). Group 3 (gray) consisted of the 52 patients who were >58years of age and had a neutrophil count 77, with a median OS of 2.496years (95% CI¼1.58 to 5.72). Group 4 (blue) consisted of the 19 patients >58years of age, with a neutrophil count >77 and a median OS of 0.70years (95% CI¼0.36 to 1.44). WHO 2016 glioma classifications for the 4 groups are shown in pie charts. D) Kaplan-Meier curves are shown for the evaluation set based on the training set clinical þ IPs RPA. Kaplan-Meier curves and risk table in the evaluation set of 350 patients with another glioma for risk groups 1-4 are defined by the training set in Figure 4, A. Astro¼astrocytoma; GBM¼glioblastoma; Oligo¼oligodendroglioma.

blood-based immune markers as an additional tool for glioma survival risk prediction.

Funding

This work was supported by the National Institutes of Health (grant numbers R01CA52689, P50CA097257, R01CA126831, R01CA139020, R25CA112355, and
R01CA207360), the loglio Collective, and the National Brain Tumor Foundation and by donations from families and friends of John Berardi, Helen Glaser, Elvera Olsen, Raymond E. Cooper, and William Martinusen. JKW is supported by the Robert Magnin Newman Endowed Chair in Neuro-oncology and MW by the Stanley D. Lewis and Virginia S. Lewis Endowment. DCK was supported by NIH grants P30 CA168524 and P20 GM130423, and P20GM103428. LAS is supported by NIGMS grant P20GM1044168299 and CDMRP/DoD grant W81XWH-20–1-0778. KTK was supported by a 2018 AACR-Johnson & Johnson Lung Cancer Innovation Science grant (18-90-52-MICH).

Notes

Role of the funder: The funder had no role in the design of the study, collection of data, analysis and interpretation of the data, the writing of the manuscript, or the decision to submit the manuscript for publication.

Disclosures: JKW and KTK are co-founders of Cellintec, which had no role in this research. JLC receives research funding from Agios and Merck and is a consultant for Agios (not directly related to this study). JWT receives research support from AbbVie, Agios, BMS, and Navio. MW is a member of the board of directors of the Central Brain Tumor Registry of the United States (unpaid position). BC is a paid consultant on R01CA207360 (Immune Epigenetic Biomarkers of Survival in Glioma
Epidemiology). All other authors have no disclosures.

Author contributions: Conceptualization: AMM, JKW, DCK, LAS, BCC, MW, KK, JWT, JLC. Methodology: AMM, GW, JYL, DCK, PC, HH, LAS, BCC, MW, KTK. Writing—Original draft: AMM, JKW, GW. Writing—Reviewing and editing: AMM, JKW, GW, JYL, DCK, PC, HH, SL, JA, TR, PB, LM, LAS, BCC, MW, KTK, JWT, JLC.

All authors were involved in the writing of the manuscript and have read and approved the final version.

Acknowledgements: The authors wish to acknowledge study participants, the clinicians, and research staff at the participating medical centers; the UCSF Helen Diller Family Comprehensive Cancer Center Genome Analysis Core, which is supported by a National Cancer Institute Cancer Center Support Grant (5P30CA082103); the UCSF Cancer Registry (for updating UCSF glioma case survival and vital status); and the UCSF Neurosurgery Tissue Bank. We thank William Karlon, MD, PhD, for his flow cytometry expertise.

Disclaimers: This publication was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through UCSF-CTSI Grant Number UL1 RR024131. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health. The collection of cancer incidence data used in this study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; Centers for Disease Control and Prevention's National Program of Cancer Registries, under cooperative agreement 5NU58DP006344; the National Cancer Institute's Surveillance, Epidemiology, and End Results Program under contract HHSN261201800032I awarded to UCSF, contract
HHSN261201800015I awarded to the University of Southern California, and contract HHSN261201800009I awarded to the Public Health Institute, Cancer Registry of Greater California. The ideas and opinions expressed herein are those of the authors and do not necessarily reflect the opinions of the state of California, Department of Public Health, the National Cancer Institute, and the Centers for Disease Control and Prevention or their contractors and subcontractors. All analyses, interpretations, and conclusions reached in this manuscript from the mortality data are those of the authors and not the state of California Department of Public Health.

Prior presentations: This research was presented as an ondemand prerecorded oral presentation at the Society for Neurooncology Annual Meeting (November 19-22, 2020).

Data Availability

Methylation and phenotype data used in this manuscript will be available through dbGaP Study Accession phs001497.v2.p1. Participant-level information about subtype, gender, age, DNAmAge, HannumAge, and DNAmPhenoAge, with a sample ID link to the deposited dbGaP data, can be found in
Supplementary Table 10 (available online). Requests to access additional deidentified data used in this manuscript will be considered through a request to the corresponding author.

References

1. Molinaro AM, Taylor JW, Wiencke JK, Wrensch MR. Genetic and molecular epidemiology of adult diffuse glioma. Nat Rev Neurol. 2019;15(7): 405–417.
2. Eckel-Passow JE, Lachance DH, Molinaro AM, et al. Glioma groups based on1p/19q, IDH, and TERT promoter mutations in tumors. N Engl J Med. 2015; 372(26):2499–2508.
3. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organizationclassification of tumors of the central nervous system: a summary. Acta Neuropathol. 2016;131(6):803–820.
4. Pekmezci M, Rice T, Molinaro AM, et al. Adult infiltrating gliomas with WHO2016 integrated diagnosis: additional prognostic roles of ATRX and TERT. Acta Neuropathol. 2017;133(6):1001–1016.

ARTICLE

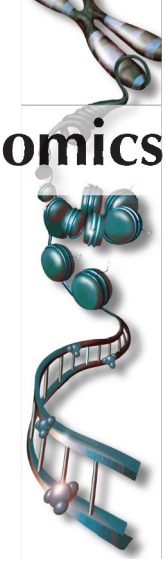
Downloaded from https://academic.oup.com/jnci/article/114/3/463/6379711 by Dartmouth College Library user on 29 September 2022

5. The Cancer Genome Atlas Research Network; Brat DJ, Verhaak RG, Aldape KD,et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. New Engl J Med. 2015;372(26):2481–2498. [10.1056/NEJMoa1402121] [26061751]
6. Delgado-Lopez PD, Corrales-Garc ıa EM. Survival in glioblastoma: a review on the impact of treatment modalities. Clin Transl Oncol. 2016;18(11): 1062–1071.
7. Curran WJ Jr, Scott CB, Horton J, et al. Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. J Natl Cancer Inst. 1993;85(9):704–710.
8. Scott CB, Scarantino C, Urtasun R, et al. Validation and predictive power ofRadiation Therapy Oncology Group (RTOG) recursive partitioning analysis classes for malignant glioma patients: a report using RTOG 90-06. Int J Radiat Oncol Biol Phys . 1998;40(1):51–55.
9. Paravati AJ, Heron DE, Landsittel D, et al. Radiotherapy and temozolomide fornewly diagnosed glioblastoma and anaplastic astrocytoma: validation of Radiation Therapy Oncology Group-Recursive Partitioning Analysis in the IMRT and temozolomide era. J Neurooncol. 2011;104(1):339–349.
10. Bell EH, Zhang P, Fisher BJ, et al. Association of MGMT promoter methylationstatus with survival outcomes in patients with high-risk glioma treated with radiotherapy and temozolomide: an analysis from the NRG oncology/RTOG 0424 trial. JAMA Oncol. 2018;4(10):1405–1409.
11. Molinaro AM, Hervey-Jumper S, Morshed RA, et al. Association of maximalextent of resection of contrast-enhanced and non-contrast-enhanced tumor with survival within molecular subgroups of patients with newly diagnosed glioblastoma. JAMA Oncol. 2020;6(4):495–503.
12. Ladomersky E, Scholtens DM, Kocherginsky M, et al. The coincidence between increasing age, immunosuppression, and the incidence of patients with glioblastoma. Front Pharmacol. 2019;10:200.
13. Pawelec G. Age and immunity: what is “immunosenescence”? Exp Gerontol. 2018;105:4–9.
14. Field AE, Robertson NA, Wang T, Havas A, Ideker T, Adams PD. DNA methylation clocks in aging: categories, causes, and consequences. Mol Cell. 2018; 71(6):882–895.
15. Goronzy JJ, Weyand CM. T cell development and receptor diversity during aging. Curr Opin Immunol. 2005;17(5):468–475.
16. Goronzy JJ, Fang F, Cavanagh MM, Qi Q, Weyand CM. Naive T cell maintenance and function in human aging. J Immunol. 2015;194(9):4073–4080.
17. Silva SL, Sousa AE. Establishment and maintenance of the human naı¨ve CD4(þ) T-cell compartment. Front Pediatr. 2016;4:119.
18. de Haan G, Lazare SS. Aging of hematopoietic stem cells. Blood. 2018;131(5): 479–487.
19. Geiger H, de Haan G, Florian MC. The ageing haematopoietic stem cell compartment. Nat Rev Immunol. 2013;13(5):376–389.
20. Aiello A, Farzaneh F, Candore G, et al. Immunosenescence and its hallmarks: how to oppose aging strategically? A review of potential options for therapeutic intervention. Front Immunol. 2019;10:2247.
21. Alpert A, Pickman Y, Leipold M, et al. A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. Nat Med. 2019;25(3):487–495.
22. Salas LA, Wiencke JK, Koestler DC, Zhang Z, Christensen BC, Kelsey KT.Tracing human stem cell lineage during development using DNA methylation. Genome Res. 2018;28(9):1285–1295.
23. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays assurrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.
24. Accomando WP, Wiencke JK, Houseman EA, Nelson HH, Kelsey KT.Quantitative reconstruction of leukocyte subsets using DNA methylation. Genome Biol. 2014;15(3):R50.
25. Hannum G, Guinney J, Zhao L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–367.
26. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115.
27. Levine ME, Lu AT, Quach A, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018;10(4):573–591.
28. Liao P, Ostrom QT, Stetson L, Barnholtz-Sloan JS. Models of epigenetic agecapture patterns of DNA methylation in glioma associated with molecular subtype, survival, and recurrence. Neuro Oncol. 2018;20(7):942–953.
29. Wrensch M, Rice T, Miike R, et al. Diagnostic, treatment, and demographicfactors influencing survival in a population-based study of adult glioma patients in the San Francisco Bay Area. Neuro Oncol. 2006;8(1):12–26.
30. Wiencke JK, Koestler DC, Salas LA, et al. Immunomethylomic approach to explore the blood neutrophil lymphocyte ratio (NLR) in glioma survival. Clin Epigenetics. 2017;9:10.
31. Koestler DC, Jones MJ, Usset J, et al. Improving cell mixture deconvolution byidentifying optimal DNA methylation libraries (IDOL). BMC Bioinformatics. 2016;17(1):120.
32. Salas LA, Koestler DC, Butler RA, et al. An optimized library for referencebased deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19(1):64.
33. Lostritto K, Strawderman RL, Molinaro AM. A partitioning deletion/substitution/addition algorithm for creating survival risk groups. Biometrics. 2012; 68(4):1146–1156.
34. Molinaro AM, Lostritto K, van der Laan M. partDSA: deletion/substitution/addition algorithm for partitioning the covariate space in prediction. Bioinformatics. 2010;26(10):1357–1363.
35. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifyingthe between-group difference in survival analysis. J Clin Oncol. 2014;32(22): 2380–2385.
36. R: A Language and Environment for Statistical Computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www. R-project.org/.
37. Templeton AJ, McNamara MG, Seruga B, et al. Prognostic role of neutrophil- to-lymphocyte ratio in solid tumors: a systematic review and meta-analysis. J Natl Cancer Inst. 2014;106(6):dju124.
38. Capper D, Jones DTW, Sill M, et al. DNA methylation-based classification ofcentral nervous system tumours. Nature. 2018;555(7697):469–474.
39. Grossman SA, Ye X, Lesser G, et al.; NABTT CNS Consortium. Immunosuppression in patients with high-grade gliomas treated with radiation and temozolomide. Clin Cancer Res. 2011;17(16):5473–5480.
40. Hughes MA, Parisi M, Grossman S, Kleinberg L. Primary brain tumors treatedwith steroids and radiotherapy: low CD4 counts and risk of infection. Int J Radiat Oncol Biol Phys. 2005;62(5):1423–1426.
41. Dix AR, Brooks WH, Roszman TL, Morford LA. Immune defects observed inpatients with primary malignant brain tumors. J Neuroimmunol. 1999;100(1-2): 216–232.
42. Chongsathidkiet P, Jackson C, Koyama S, et al. Sequestration of T cells inbone marrow in the setting of glioblastoma and other intracranial tumors [published correction appears in Nat Med. 2019;25(3):529]. Nat Med. 2018;24(9): 1459–1468.
43. Mason M, Maurice C, McNamara MG, et al. Neutrophil-lymphocyte ratio dynamics during concurrent chemo-radiotherapy for glioblastoma is an independent predictor for overall survival. J Neurooncol. 2017;132(3): 463–471.
44. Lesterhuis WJ, Bosco A, Millward MJ, Small M, Nowak AK, Lake RA. Dynamicversus static biomarkers in cancer immune checkpoint blockade: unravelling complexity. Nat Rev Drug Discov. 2017;16(4):264–272.
45. Hartmann FJ, Bendall SC. Immune monitoring using mass cytometry and related high-dimensional imaging approaches. Nat Rev Rheumatol. 2020;16(2): 87–99.
46. Spitzer MH, Carmi Y, Reticker-Flynn NE, et al. Systemic immunity is requiredfor effective cancer immunotherapy. Cell. 2017;168(3):487–502.e15.
47. Yost KE, Satpathy AT, Wells DK, et al. Clonal replacement of tumor-specific Tcells following PD-1 blockade. Nat Med. 2019;25(8):1251–1259.
48. Krieg C, Nowicka M, Guglietta S, et al. High-dimensional single-cellanalysis predicts response to anti-PD-1 immunotherapy [published correction appears in Nat Med. 2018;24(11):1773-1775]. Nat Med. 2018;24(2):144–153.
49. Iwahori K, Shintani Y, Funaki S, et al. Peripheral T cell cytotoxicity predicts Tcell function in the tumor microenvironment. Sci Rep. 2019;9(1):2636.
50. Cloughesy TF, Mochizuki AY, Orpilla JR, et al. Neoadjuvant anti-PD-1 immunotherapy promotes a survival benefit with intratumoral and systemic immune responses in recurrent glioblastoma. Nat Med. 2019;25(3): 477–486.

Downloaded from https://academic.oup.com/jnci/article/114/3/446/6379711 by [illegible] College Library user on 29 September 2022

Research Article

For reprint orders, please contact reprints@futuremedicine.com



Navigating the hydroxymethylome: experimental biases and quality control tools for the tandem bisulfite and oxidative bisulfite Illumina microarrays

Ze Zhang[1], Min Kyung Lee[1], Laurent Perreard[2], Karl T Kelsey[3], Brock C Christensen[1,2] & Lucas A Salas*[,1,2]

[1]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Lebanon, 03756 NH, USA

[2]Department of Molecular & Systems Biology, Geisel School of Medicine, Dartmouth College, Lebanon, 03756 NH, USA

[3]Department of Epidemiology, Department of Pathology & Laboratory Medicine, Brown University School of Public Health, Providence, 02912 RI, USA

*Author for correspondence: Tel.: 603 646 5496; lucas.a.salas@dartmouth.edu

Aim: Tandem bisulfite (BS) and oxidative bisulfite (oxBS) conversion on DNA followed by hybridization to Infinium HumanMethylation BeadChips allows nucleotide resolution of 5-hydroxymethylcytosine genome-wide. Here, the authors compared data quality acquired from BS-treated and oxBS-treated samples. **Materials & methods:** Raw BeadArray data from 417 pairs of samples across 12 independent datasets were included in the study. Probe call rates were compared between paired BS and oxBS treatments controlling for technical variables. **Results:** oxBS-treated samples had a significantly lower call rate. Among technical variables, DNA-specific extraction kits performed better with higher call rates after oxBS conversion. **Conclusion:** The authors emphasize the importance of quality control during oxBS conversion to minimize information loss and recommend using a DNA-specific extraction kit for DNA extraction and an oxBSQC package for data preprocessing.

Lay abstract: DNA hydroxymethylation (5-hydroxymethylcytosine [5hmC]) is a chemical modification of the cytosines in the DNA that affects gene transcription. 5hmC has been used as a biomarker for early cancer detection and survival prediction in recent years. 5hmC is measured using tandem bisulfite (BS) and oxidative bisulfite (oxBS) conversion of DNA followed by quantification through DNA methylation microarrays. This study observed a consistent loss of high-quality data in oxBS-treated samples compared with BS-treated samples. The authors offer a bioinformatic tool to evaluate potential quality issues in the process and some technical advice to reduce false signals in the data. Thus, they emphasize the importance of preserving DNA integrity when using tandem BS- and oxBS-treated DNA to measure 5-methylcytosine and 5hmC.

First draft submitted: 25 November 2021; Accepted for publication: 6 January 2022; Published online: 14 January 2022

Keywords: bisulfite conversion • DNA hydroxymethylation • DNA methylation • epigenetics • Infinium HumanMethylation450K BeadChip • Infinium MethylationEPIC BeadChip • oxidative bisulfite conversion

DNA methylation is a type of epigenetic modification where a methyl group replaces one hydrogen in the fifth position of the cytosine nucleotide in DNA (5-methylcytosine [5mC]). Specifically, DNA methylation occurs most frequently in the context of CpG dinucleotides. 5mC is well-studied in regulating gene transcription, inhibiting retrotransposonactivationandmaintaininggenedosagecompensationinnaturallyimprintedregionsandisessential for establishing and preserving cellular identity [1–4]. Although 5mC is a stable DNA chemical modification, 5mC demethylation can be achieved by either passive dilution through DNA replication or active demethylation by the TETfamilyofenzymes.Duringactivedemethylation,theTETenzymesoxidize5mCinto5-hydroxymethylcytosine

10.2217/epi-2021-0490 ©2022 Z Zhang, MK Lee, L Perreard, KT Kelsey, BC Christensen, LA Salas *Epigenomics* (2022) 14(3), 139–152ISSN 1750-1911

(5hmC). TET further oxidizes 5hmC to 5-formylcytosine (5fC) and then 5-carboxylcytosine (5caC) [5–7]. Finally, the TDG and the base excision repair (BER) pathway convert 5fC and 5caC into an unmethylated cytosine [8].

5hmC is frequently used as a marker to investigate the dynamic DNA demethylation process, as an intermediate state between 5mC and an unmethylated cytosine [5]. Most tissues have ~4% of 5mC, while 5hmC varies under 1% by different tissue types [9]. 5fC and 5caC are ten- to 1000-fold less abundant than 5hmC and considered committed demethylation intermediates as they are effectively removed to complete the demethylation process [10]. Unlike 5fC and 5caC, 5hmC modifications are relatively stable and have multiple independent functions beyond serving as an intermediate oxidized form of 5mC in the active demethylation process [5]. Recent research has demonstrated that 5hmC can serve as an informative biomarker in mammalian development and diseases [6,11]. In the human fetal brain, 5hmC modification plays a critical role in neuronal function, learning and memory during neurodevelopment [12]. In the immune system, the dynamic change of 5hmC is responsible for T-cell development and differentiation [13]. In cancer, the 5hmC signal in circulating cell-free DNA (cfDNA) is a promising biomarker for early cancer detection and indicates multiple tumor types and stages [14–17].

The Infinium HumanMethylation450K and MethylationEPIC BeadChip arrays designed by Illumina (CA, USA) have been most commonly used after sodium bisulfite (BS) treatment for genome-wide methylation studies in recent years [18]. When DNA is treated with sodium BS, 5mC and 5hmC are preserved as cytosines, while unmethylated cytosines are deaminated to uracil. In other words, DNA methylation measured by methylation microarrays following BS treatment is an aggregation of 5mC and 5hmC. Most differentiated tissues and cell types have a very low level of 5hmC, with the central nervous system tissue containing the highest level of 5hmC (0.3–0.7%) in DNA [9]. The kidney, nasal epithelium, bladder, heart, skeletal muscle and lung are considered to have a medium level of 5hmC abundance (0.15–0.17%) [9]. The aggregation of 5mC and 5hmC measured using BS treatment is acceptable for most tissues, as the 5hmC level is generally low. However, for tissues with a relatively high 5hmC abundance, it is necessary to distinguish 5hmC from 5mC to obtain the true signals.

Tandem BS and oxidative bisulfite (oxBS) treatment is one of the most commonly used methods to discern 5hmC from 5mC at the single-nucleotide level on the same sample, followed by hybridization to the methylation microarrays [19,20]. Separate aliquots of each DNA sample are processed in parallel where one aliquot undergoes BS treatment only and the other first undergoes oxidation and then BS treatment. When DNA is oxidized, 5mC remains as 5mC, while 5hmC is converted into 5fC. The 5fC is susceptible to BS treatment, and it is deaminated into uracil (equivalent to an unmethylated cytosine), while 5mC is preserved as a cytosine upon BS treatment [21]. Thus, oxBS conversion allows the measurement of nucleotide-level 5mC specifically. The level of 5hmC at a single-nucleotide resolution is then estimated by subtracting the oxBS measure (5mC) from the BS measure (5hmC + 5mC) at each probe [19].

Although it is considered the gold standard for measuring DNA methylation, BS treatment results in some DNA damage by fragmentation [22,23]. By first using an oxidant to convert 5hmC to 5fC in oxBS treatment, additional damage may be incurred [24]. However, the second-generation oxBS-seq protocol demonstrated a comparable DNA recovery from the oxBS conversion workflow to that from the BS conversion [21]. Importantly, the oxidant is also sensitive to several contaminants if present. Special cautions must be taken to prevent interactions with chemicals such as alcohols (ethanol, isopropanol, phenol), alcohol-containing compounds (Tris, EB buffer, TE buffer, glycerol, surfactants), solutions < pH 9.0 and CO_2, both in the air and in the solution, which may interfere with the oxBS reaction [25].

Probe call rate is usually evaluated using detection p-values derived from comparing negative control probe fluorescence signal intensity to intensities at measured cytosines. Background signals are primarily based on negative technical controls included in the array and used in multiple methylation data processing pipelines as part of their quality control [26,27]. However, more recently, some evidence has shown that the use of negative controls (even using extremely stringent detection p-value limits) cannot filter out artifacts caused by failed hybridization due to compromised DNA integrity [28,29]. Those unfiltered, low-quality probes often manifest falsified intermediate methylation. In one study, p-value with out-of-band (OOB) array hybridization (pOOBAH) was designed and tested to address those artifacts [28]. Unlike a negative detection p-value, pOOBHA evaluates probe call rate using OOB signals of the type I probes [28,30]. For type I probes, only one out of two color channels was used to measure the methylation signal. The intensities in the unused color channel were utilized as a reference to calculate pOOBHA and assess probe call rate [28].

This study aims to evaluate the call rate of the probes on Illumina microarrays in paired samples exposed to tandem BS and oxBS conversion and identify potential sources of technical variability contributing to the altered

Table 1. Baseline characteristics of discovery datasets.

Study			Tissue	n (%) array	Size (pairs)	Methylation	Age mean (SD)	Male,	DNA extraction kit type
ccRCC	ccRCC	22	EPIC		60.5 (11.8)	18 (40.9)			Tandem Nucleic Acid Extraction Kit/DNA-specific Extraction Kit
HFO	HFO	18	EPIC		0 (0)	36 (100)			DNA-specific Extraction Kit
DWB	Glioma patients treated with	16	EPIC		35.3 (14.6)	24 (75)			DNA-specific Extraction Kit dexamethasone and controls
PedCNS	Pediatric brain tumor	40	EPIC		27.1 (19)†	39 (48.8)†			DNA-specific Extraction Kit
GSE63179	Human cerebellum	4	450K		25 (0)	8 (100)			DNA-specific Extraction Kit
GSE71328	Kagami–Ogata syndrome blood and	10	450K		6 (8.8)†	12 (60)			DNA-specific Extraction Kit control, normal brain
GSE71719	Human placenta	23	450K		0 (0)	24 (52.1)			DNA-specific Extraction Kit
GSE73895	Glioblastoma	30	450K		66 (12)	36 (60)			DNA-specific Extraction Kit
GSE104723	High-grade human glioma	21	EPIC		58.4 (18.3)†	24 (57.1)†			DNA-specific Extraction Kit
GSE105109	Alzheimer's disease and normal	192	450K		81.4 (9.3)	216 (56.3)			DNA-specific Extraction Kit control brain samples
GSE116402	Human bronchial epithelial cells	6	EPIC		42.4 (4.1)†	0 (0)†			Tandem Nucleic Acid Extraction Kit
GSE133062	Bronchoalveolar lavage samples in	35	EPIC		26.9 (6.7)	26 (37.1)			Tandem Nucleic Acid Extraction Kit smokers and nonsmokers

†Inferred age and sex, based on methylation using SeSAMe.

ccRCC: Clear cell renal cell carcinoma; DWB: Dexamethasone-treated glioma patient whole blood; HFO: Human fetal oligodendrocyte; SD: Standard deviation.

signals. As a secondary aim, the authors developed a new bioinformatic tool to perform quality control in an effort to eliminate unstable 5hmC signals and identify potentially mislabeled samples undergoing tandem oxBS using control probes in the Illumina microarrays.

Materials & methods

Datasets

Twelve independent datasets were included in this study, four data sets from the authors' laboratories available through Gene Expression Omnibus (GEO; GSE182919) and eight publicly available datasets on GEO that have Illumina Human Methylation450K or MethylationEPIC array data on paired BS and oxBS-treated samples (Table 1). In total, 417 pairs of samples were accessed, including clear cell renal cell carcinoma (ccRCC; n = 22), human fetal oligodendrocyte (HFO; n = 18), dexamethasone-treated glioma patient whole blood (DWB; n = 16), pediatric brain tumors (n = 40), human cerebellum (n = 4) [19], Kagami–Ogata syndrome patient peripheral blood (n = 10) [31], human placenta (n = 23) [32], glioblastoma (n = 30) [33], high-grade glioma (n = 21), Alzheimer's disease patients' brains and controls (n = 192) [34], human bronchial epithelial cells (n = 6) [35] and smokers' bronchoalveolar lavage (n = 35) [36]. The pOOBHA value was computed using the SeSAMe pipeline.

Probe call-rate calculation

The probe call rate was defined as the proportion of probes with a pOOBHA <0.05 over the total number of probes contained on the microarray. It was calculated as $\#_{\textit{probes with pOOBHA}} / \#_{\textit{total probes on array}} < 0.05 \times 100\%$. A one-sided pairwise Wilcoxon rank-sum test was used to compare the probe call rate between paired BS-treated samples and oxBS-treated samples across datasets.

Statistical analyses on technical variables & signal variation

Next, the authors conducted principal component regression analyses, with extraction kit, array ID, single-stranded DNAamount(ng/μl),BSandoxBStreatmentbatch,totalDNAamount,amountofsampletouseperoxBSreaction and subject included in the model to investigate the variable attribution to the variation of unfiltered methylation signal in BS-converted and oxBS-converted samples, respectively. Multiple types of commercial DNA extraction kits are available on the market and the performance of kits may vary by the purpose of the experiment [37,38]. While a DNA-specific extraction kit targets DNA extraction solely, tandem nucleic acid extraction kits are designed for both DNA and RNA extraction. In this study, the DNA-specific extraction kits used included Qiagen Blood & Tissue Kit, CelProgen and Zymo Quick-DNA Miniprep Plus Kit, and the tandem nucleic acid extraction

kit used was Qiagen AllPrep Kit. To test whether the probe call rate is significantly different between using a DNA-specific extraction kit and using a tandem nucleic acid extraction kit, for DNA extraction in BS and oxBS samples, the authors employed two different models: a multivariable linear regression model with sample source adjusted as a fixed-effect (Equation 1) and a linear mixed-effect model with sample source included as a random intercept (Equation 2).

$$Y_{ij} = \beta_{0j} + \beta_{1j} Kit_i + \beta_{2j} Study_i + \varepsilon_{ij} \quad \text{(Equation 1)}$$

$$Y_{ij} = \beta_{0j} + \beta_{1j} Kit_i + b_{0j} Study_i + \varepsilon_{ij} \quad \text{(Equation 2)}$$

In the equations above, the dependent variable Y is the probe call rate. i represents sample (i = 1, 2, . . . , n). j representswhetherthesamplesareBS-convertedoroxBS-converted(j=BSoroxBS).$b$representsrandomeffect,and ε is random error. p-value < 0.05 is the cutoff for testing the hypothesis that the mean probe call rate is equivalent between using a tandem nucleic acid extraction kit and using a DNA-specific extraction kit (e.g., H_0:β_{1j} = 0).

Statistical analyses on 5mC & 5hmC stability

To further investigate the consistency and reproducibility of 5mC and 5hmC signals, the authors examined the correlation of 5mC and 5hmC beta values across 18 replicate HFO samples. The Pearson's correlation coefficient was computed for 5mC and 5hmC signals in each pair in the replicates. Next, to explore the stability of 5mC and 5hmC signals across all probes, the coefficient of variation was calculated and plotted against the median beta value for each probe in the replicate samples. The replicate samples were prepared in four different batches. The intraclass correlation coefficient (ICC) with different batches defined as groups was calculated and regressed against the median beta value for each probe across the replicate samples. The ICC indicates how stable the signal is per probe across the batches.

Making recommendations to eliminate unstable 5hmC signals & evaluate the information loss with low-quality probes

To eliminate the probes with unstable 5hmC signals, a cutoff 5hmC level is recommended based on the coefficient of variation of the 5hmC signal across probes. The cutoff point was selected using the optimality criteria modified by Perkins and Schisterman [39]. The unstable probes were defined as probes with a median 5hmC value smaller than the optimized cutoff value. For probes that passed the recommended cutoff (stable probes), a low-quality probe was identified as a probe with a pOOBHA <0.05 for less than 20% of the samples. The low-quality probe fraction is calculated as $\#_{pOOBHA^{5hmc}>cutoff \text{ for } <0.05} \frac{probes\ with\ median\ 5\#hmC\ probes\ with\ median>cutoff\ and}{20\%\ samples} \times 100\%$.

To evaluate the loss of information for low-quality probes, a density plot was created to visualize the distribution of 5hmC signal between the low-quality probes and the probes that passed quality control. A genomic location enrichment analysis was conducted to assess if a certain CpG location (island, shore, shelf and open sea) was enriched for low-quality probes. The *oxBScut* function in the oxBSQC package was created to make the 5hmC cutoff recommendation, identify low-quality probes and evaluate the information loss by excluding the low-quality probes from the study. The package is available on https://github.com/SalasLab/oxBSQC.

The package was tested on the authors' ccRCC datasets. The ccRCC dataset is comprised of two batches of sample processing. The first batch contains eight pairs of BS and oxBS-treated DNA that are extracted using only the tandem nucleic acid extraction kit. The second batch contains 16 pairs of BS- and oxBS-treated DNA that are extracted using the tandem nucleic acid extraction kit and the DNA-specific extraction kit. More operational effort was made toward maintaining the integrity of DNA in the second batch.

Development of using internal normalization control probes to check BS & oxBS sample labeling

The Illumina HumanMethylation BeadChip contains internal normalization control probes (NORM A/G/C/T) to assess the array performance in two color channels (red A/T and green G/C) and can further be used for correcting intensity-related dye biases [26,40]. As the oxBS procedure involves additional oxidative treatment on DNA compared with the BS procedure, the authors hypothesized that the signal intensities in the internal control probes are lower in oxBS-converted samples than in the BS-converted samples. Such differences would allow discrimination between two samples in a BS and an oxBS-treated pair. To test the hypothesis, the authors applied the one-sided pairwise Wilcoxon rank-sum test to the paired ccRCC samples to investigate if the signal intensity in oxBS-treated samples is lower than that in BS-treated samples in NORM A, NORM G, NORM C and NORM D probes, respectively. The p-value <0.05 is defined as the cutoff for statistical significance.

Furthermore, to test if the same standard can be used to correctly categorize BS and oxBS-converted samples in another independent dataset, the authors applied the one-sided pairwise Wilcoxon rank-sum test with the direction defined as 'greater' to the paired HFO samples with their BS and oxBS information masked and randomized. In a BS and an oxBS-treated HFO pair, if the returned p-value <0.05, the first sample is categorized as 'BS' whereas the other one is classified as 'oxBS'. If the returned p-value ≥0.05, the first sample is categorized as 'oxBS' while the other is 'BS' The sensitivity and specificity of using the method to predict BS and oxBS-converted samples were calculated for NORM A, NORM G, NORM C and NORM D probes, respectively. The *oxBSlabel* function was created to predict BS and oxBS labels using the NORM A/G/C/T internal control probes and included in the oxBSQC package.

Results

Probe call rates in BS- & oxBS-treated samples

Twelve independent datasets with 417 paired BS- and oxBS-treated samples across 14 tissue types were investigated for probe call rate (Table 1). A consistent decrease of probe call rate was observed in paired oxBS-treated samples compared with BS-treated samples across all datasets (Figure 1). The probe call rate decreased, on average, by 11.1% (ranges from 3.4 to 22.9%) in oxBS samples compared with BS samples (Supplementary Figure 1). All datasets with sample sizes of five or more (11 out of the 12) had significantly lower probe call rates in oxBS-treated samples ($p < 0.05$). The magnitude of the probe call rate decrease varied by datasets (CV = 0.47).

Probe quality differences between BS- & oxBS-treated samples based on DNA extraction kit used

To understand the attribution of different factors to the variation of methylation signal in BS-treated samples and oxBS-treated samples, the authors conducted principal component analyses on the ccRCC samples. For the BS-treated samples, the only variable identified as significant in the first principal component is the subject, which explains 22.1% of the data variation (Supplementary Figure 2A). For the oxBS-treated samples, the first component identified the extraction kit, array ID, oxBS treatment batch, total DNA amount and subject as significant, which in total explains 29.3% of the data variation (Supplementary Figure 2B). The results revealed more variables that greatly contribute to the variation of methylation signal after the oxBS treatment compared with the BS treatment. To further explore how the DNA extraction kit can impact the data quality, the authors compared the probe call rate between samples that used a DNA-specific extraction kit and samples that used a tandem nucleic acid extraction kit (Figure 2). For BS-converted samples, no significant difference in probe call rate was observed between samples that used a DNA-specific extraction kit and samples that used a tandem nucleic acid extraction kit in the multivariable linear regression model after adjusting for sample source (= -0.4%; $p > 0.1$; Supplementary Table 1) and the linear mixed-effect model with sample source included as a random intercept (= -1.7%; $p > 0.1$; Supplementary Table 2). For oxBS-converted samples, significantly lower probe call rates were observed in samples that used the tandem nucleic acid extraction kit compared with samples that used the DNA-specific extraction kit in the multivariable linear regression model after adjusting for sample source (= -19.5%; $p < 0.01$; Supplementary Table 1) and linear mixed-effect model with sample source as a random intercept (= -15.6%; $p < 0.01$; Supplementary Table 2). As a sensitivity analysis, the authors excluded the ccRCC samples, as they appeared to drive the insignificant results in the BS-converted samples. The authors observed significantly lower probe call rates in samples that used the tandem nucleic acid extraction kit compared with samples that used the DNA-specific extraction kit for both BS (= -12.2%; $p = 0.038$) and oxBS (= -12.5%; $p = 0.039$) conversions. Cumulatively, the data demonstrated that for oxBS conversion, using a DNA-specific extraction kit as opposed to a tandem nucleic acid extraction kit yields higher-quality data.

5mC & 5hmC signal stability

With 18 pairs of replicate HFO samples, the authors assessed the stability of the 5mC and 5hmC signals. 5mC signal is highly stable in the replicates with the Pearson's r >0.99 for any pair (Figure 3A), whereas the Pearson's r ranges from 0.70 to 0.88 for 5hmC signal (Figure 3B). For instance, the Pearson's r between HFO001 and HFO002 is 0.998 for 5mC (Figure 3A) and 0.820 for 5hmC (Figure 3B). Next, the authors computed the coefficient of variation for each probe across the replicate samples and regressed against median beta values for 5mC and 5hmC to explore the stability of the signal at the probe level. For 5hmC, the signal is unstable for probes with a low 5hmC

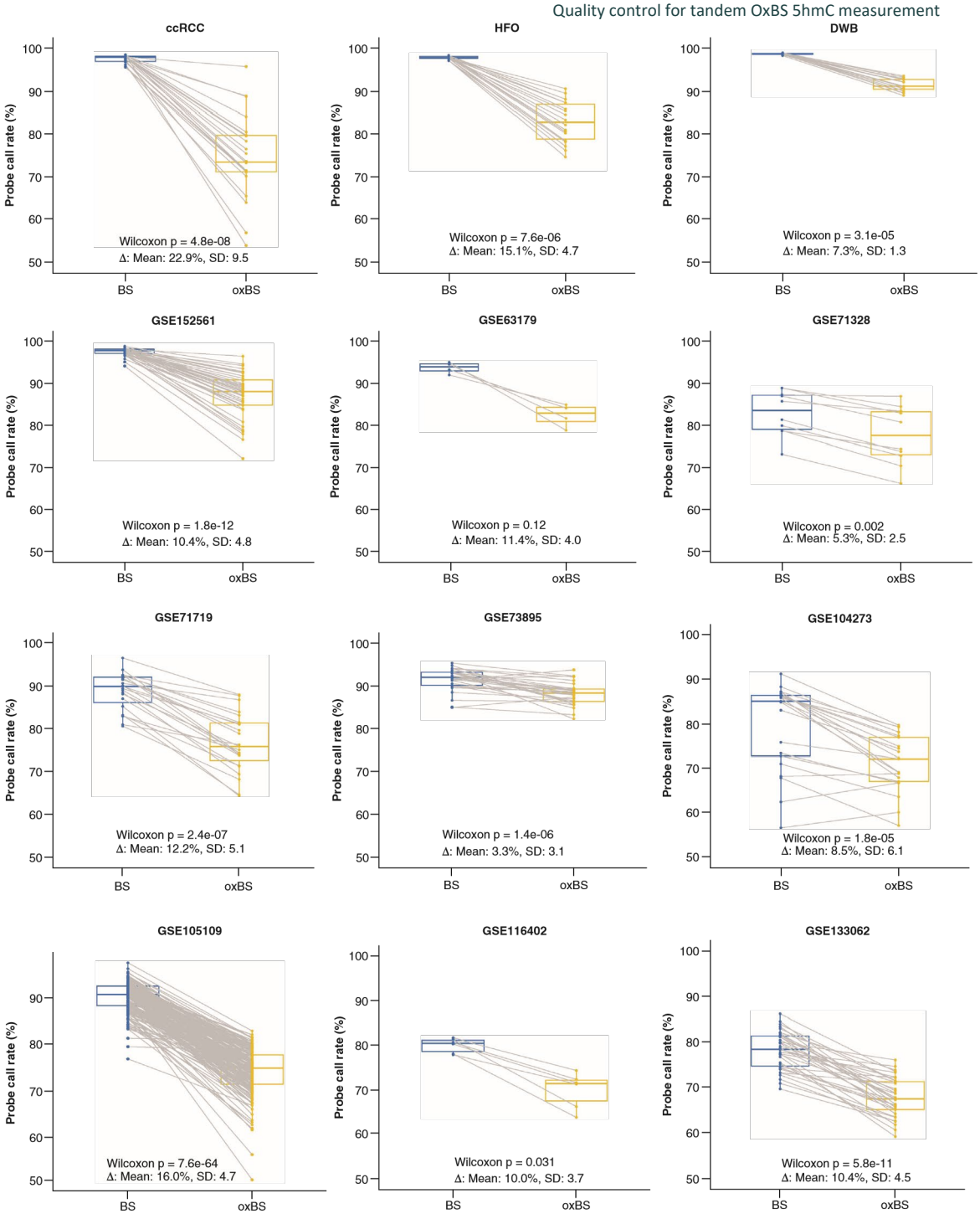


Figure 1. Probe call rates in bisulfite- and oxidative bisulfite-treated samples. A consistent decrease of probe call rates was observed in paired oxidative bisulfite-treated samples compared with bisulfite-treated samples across all datasets. p-values were calculated using one-sided pairwise Wilcoxon rank-sum tests.
ccRCC: Clear cell renal cell carcinoma; DWB: Dexamethasone-treated glioma patient whole blood; HFO: Human fetal oligodendrocyte.

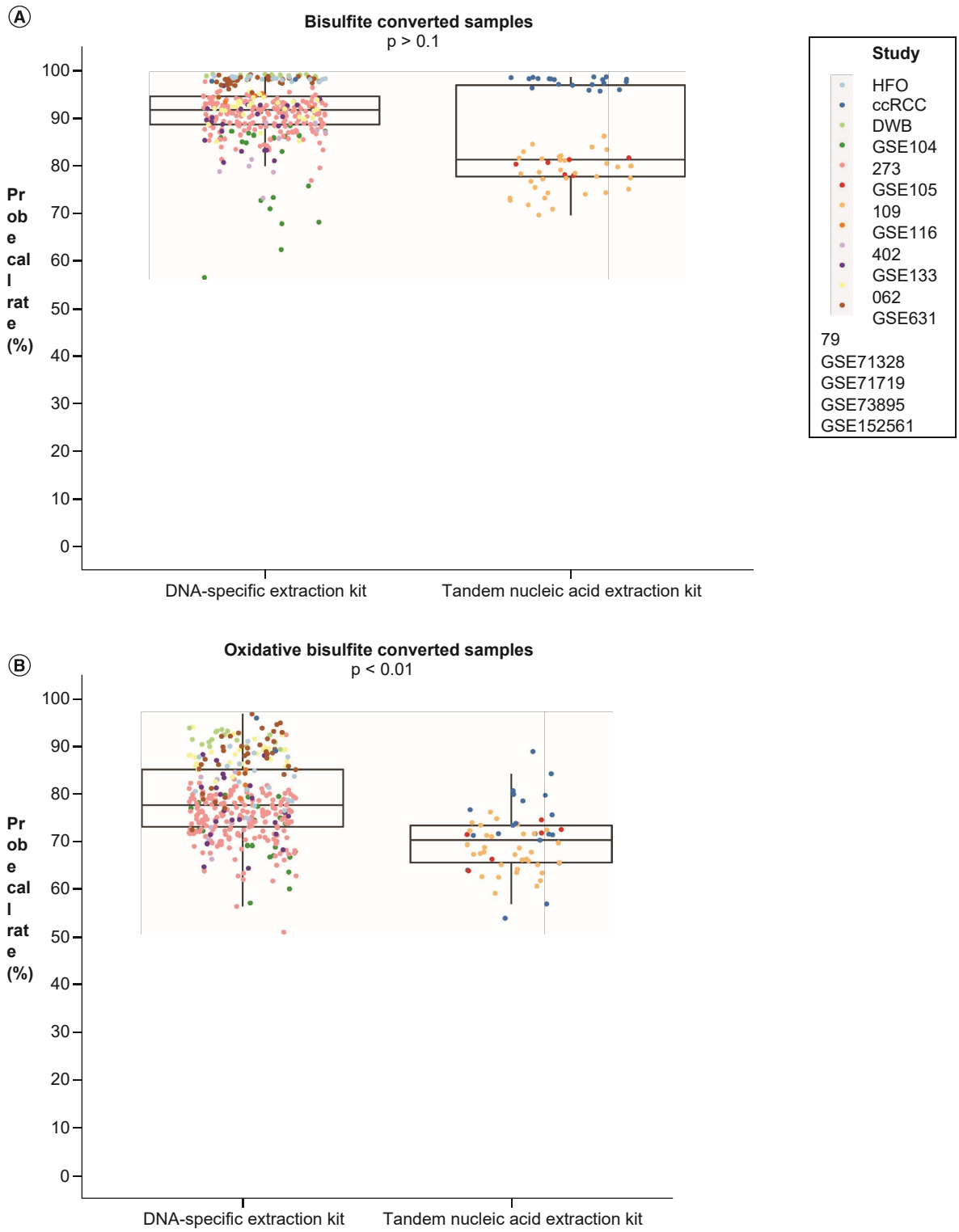


Figure 2. Probe quality differences between bisulfite- and oxidative bisulfite-treated samples based on the DNA extraction kit used. (A) For bisulfite-converted samples, no significant difference was observed in the probe call rates between samples that used DNA-specific extraction kits and samples that used tandem nucleic acid extraction kits ($p > 0.1$). **(B)** For oxidative bisulfite-converted samples, significantly lower probe call rates were observed in samples that used the tandem nucleic acid extraction kit compared with samples that used the DNA-specific extraction kit ($p < 0.01$). p-values were calculated using multivariable linear regression models and linear mixed-effect models. ccRCC: Clear cell renal cell carcinoma; DWB: Dexamethasone-treated glioma patient whole blood; HFO: Human fetal oligodendrocyte.

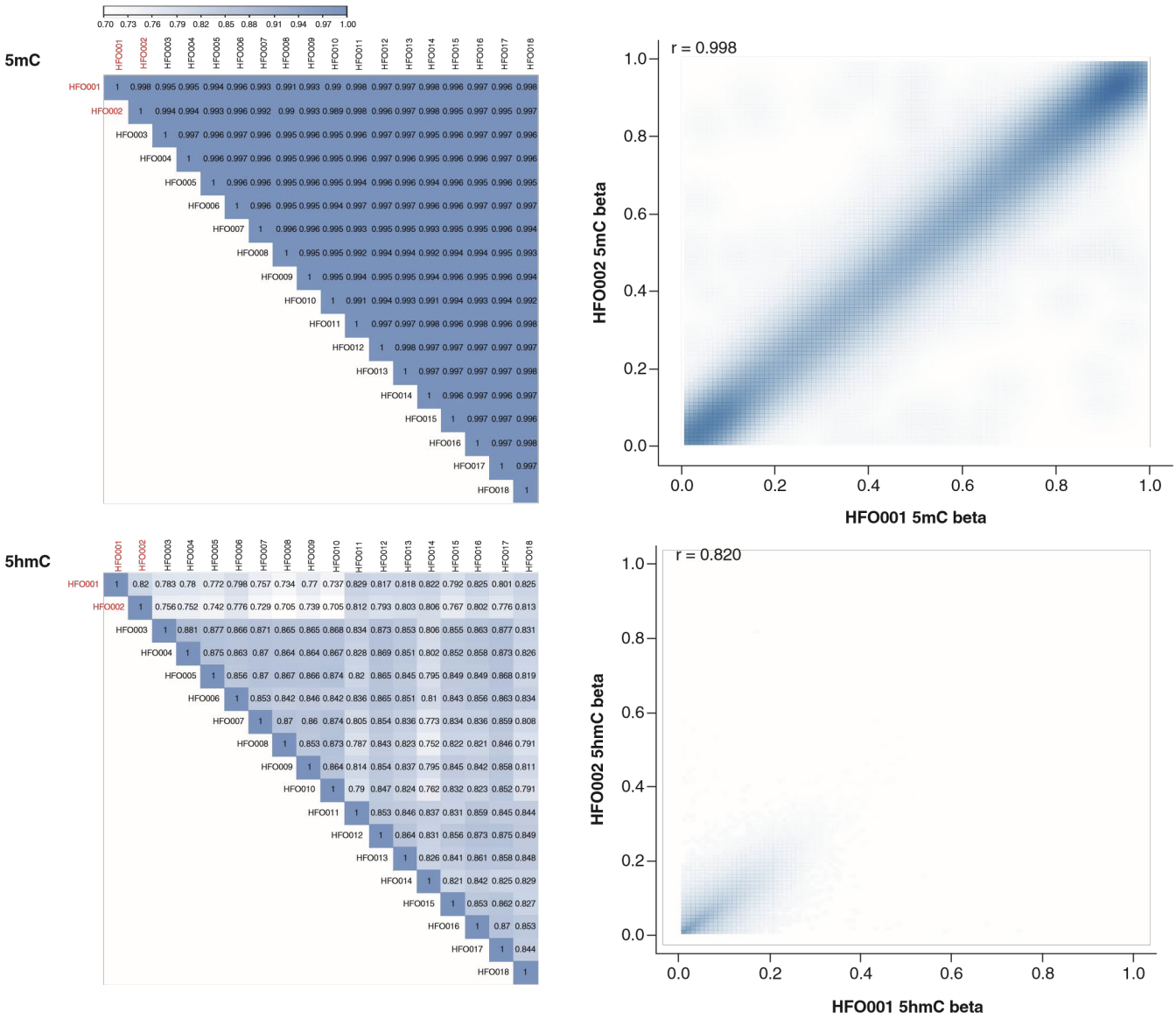


Figure 3. Evaluation of the stability of the 5-methylcytosine and 5-hydroxymethylcytosine signals using 18 pairs of replicate human fetal oligodendrocyte samples. (A) The 5mC signal is highly stable in the replicates with the Pearson's r >0.99 for any pair. For instance, the Pearson's r between HFO001 and HFO002 is 0.998 for 5mC. **(B)** The Pearson's r ranges from 0.70 to 0.88 for the 5hmC signal. The Pearson's r between HFO001 and HFO002 is 0.820 for 5hmC.
5hmC: 5-hydroxymethylcytosine; 5mc: 5-methylcytosine; HFO: Human fetal oligodendrocyte.

Ⓐ

level, especially for the ones approaching zero (Supplementary Figure 3A). Although the signal is also relatively unstable for low 5mC level probes, the CV does not reach as high as that of low 5hmC level probes (Supplementary Figure 3B). Next, the authors investigated how the BS and oxBS treatment batch can impact 5mC and 5hmC stability at the probe level by exploring the relationship between the ICC with different batches defined as groups and the median beta values for each probe across the replicate samples. The authors observed that ICC is lower in low 5hmC level probes ($r = 0.45$; Supplementary Figure 3C). Although a positive r ($r = 0.30$) is also observed between ICC and 5mC level, the correlation is not as strong as that in 5hmC (Supplementary Figure 3D). In summary, the authors observed more instability for the 5hmC signal than for the 5mC signal in replicate samples, especially for probes with low levels of 5hmC.

Making a recommendation for a 5hmC cutoff value

As the authors observed instability with low-level 5hmC signals, they next developed a bioinformatic tool, oxBSQC, to recommend a 5hmC cutoff value that excludes the probes with an unstable 5hmC signal. The oxBSQC was applied to the ccRCC datasets. The recommended 5hmC cutoff is 0.09 for batch 1 samples and 0.08 for batch 2 samples (Supplementary Figure 4). The low-quality probe fractions are 71.84% in batch 1 samples and 40.82% in batch 2 samples (Supplementary Figure 5). By evaluating the CpG genomic location enrichment for low-quality probes, the authors observed that the open sea area is significantly enriched for the low-quality probes in both batch 1 and batch 2 samples (Supplementary Figure 6). In summary, the oxBSQC is a practical tool that

Ⓑ

makes recommendations for 5hmC cutoff and evaluates the information loss with the unusable probes accommodating the variability of datasets.

Using internal normalization controls for a quality check on BS & oxBS labeling

In the next step, the authors assessed whether the signal intensities in the internal normalization control probes (NORM A/G/C/T) were distinguishable between BS-converted and oxBS-converted ccRCC samples. Among 22 pairs of BS- and oxBS-converted ccRCC samples, all pairs demonstrated discernible signal intensities in NORM A and NORM T probes in the red channel and NORM C in the green channel ($p < 0.05$). For NORM G probes in the green channel, one pair out of 22 did not reach the statistical significance threshold but was marginally significant ($p = 0.053$; Supplementary Figure 7 & Supplementary Table 3). Signal intensity in the internal normalization control probes to distinguish BS- and oxBS-converted samples in 18 pairs of HFO replicate samples reached 100% sensitivity and 100% specificity. Applying the same standard to the publicly available datasets included in this study, the authors identified mislabeled samples and reached out to the researchers for correction on public repositories.

Discussion

Although DNA hydroxymethylation serves as an intermediate during the dynamic DNA demethylation process, it is involved in normal tissue development and disease pathogenesis, such as cancer [41–43]. Studies have identified stable 5hmC biomarkers for the early diagnosis, prognostic evaluation and prediction of therapeutic efficacy for various cancer types [11,42]. 5hmC signatures in cfDNA using non-invasive blood tests have shown promising early cancer detection results [14–17]. Tandem oxBS and BS treatments are among the most commonly used methods for measuring 5mC and 5hmC at the genome-wide scale due to cost–effectiveness and efficiency [19,20].

The present study observed a consistent decrease in the probe call rate after oxBS treatment on Illumina HumanMethylation microarrays, although the magnitude of the loss varied across datasets. The reduction in the probe call rates in oxBS-treated samples was consistent across all datasets. Although one of the most commonly used methods for measuring 5hmC at the base resolution, paired BS conversion and oxBS conversion have been criticized for their harsh chemical treatment that could significantly degrade the DNA [22,23]. BS-free methods such as TET-assisted pyridine borane sequencing (TAPS) have been developed to overcome the issue [44]. However, such methods are costly and laborious to conduct and have not been validated for use with microarrays. Although a certain level of DNA degradation can be caused by BS treatment, the authors have shown that obtaining a probe call rate above 80–90% is achievable. Although the second-generation oxBS-seq protocol illustrated that a similar level of DNA recovery between BS and oxBS workflow is feasible [21], the authors observed variability in the extent of high-quality probe loss after oxBS treatment. Two major components that could explain the variation of DNA quality after BS and oxBS treatments are the biological nature of the sample used and the artifacts engendered during the experimental operation. DNA fragility is dependent on the tissue or organ from which the cell originated [45,46]. Inherently more fragile DNA would be more susceptible to the damage caused by BS and oxBS treatment. Due to the biological nature of tissue-specific DNA fragility, the possible effort is limited to

addressing the vulnerability of specific DNA during BS and oxBS conversion. However, interventions are possible during the process to reduce artificial damage to DNA to preserve DNA integrity to the fullest extent. For instance, in the present study, in the two batches of ccRCC samples, the data quality of the second batch outperformed the first batch immensely, with the low-quality probe fraction decreased by 31%. The authors believe that such improvement is greatly attributed to the operational attention paid to maintaining the integrity of DNA, with DNA fragmentation being monitored in the second batch.

The present study identified multiple factors that contribute to the variation in methylation and hydroxymethylation data. Apart from the variability between subjects, the data variation is substantially attributed to the DNA extraction kit used, the total DNA amount used and the batches of Infinium methylation arrays. The authors identified 5hmC as more susceptible to technical effects other than the between-subject effect compared with the 5mC signals. DNA extraction is a key step preceding BS and oxBS treatment. The two major types of DNA extraction kits on the market are DNA-specific extraction kits designed specifically for DNA purification and extraction and tandem nucleic acid extraction kits designed for simultaneous purification of DNA and RNA. Previous evidence has demonstrated the variations in extracted DNA quality across different commercial extraction kits [37,38]. An 83–406% increase in DNA yields was observed in a DNA-specific kit compared with tandem nucleic acid extraction kits [38]. From these results, although there were no significant differences between using the two types of kits in methylation data integrity in BS-treated samples, the authors observed a better quality of data retained in oxBS-treated samples by using the DNA-specific extraction kit. Compared with the BS conversion of DNA, the additional oxidative chemical treatment in oxBS conversion likely added more burden to DNA fragments [24]. Larger yields of DNA by using a DNA-specific extraction kit may help prepare the samples to survive the harsh chemical treatments. DNA-specific extraction is the authors' recommended method for preparing DNA for tandem BS and oxBS treatment. Further studies are still required to identify potential contaminants in the process.

The high consistency and reproducibility of methylation measurements using Infinium HumanMethylation BeadChips across replicate samples observed in the present study are consistent with previous findings [47]. Furthermore, with the 5mC and 5hmC signals distinguished, the authors observed better stability of the 5mC signal than the 5hmC signal in replicate samples. The instability of the 5hmC signal mainly came from CpGs with low levels of 5hmC. The probes with a generally lower measure of 5hmC demonstrated more variation and more susceptibility to batch effect in this study. Especially for the probes approaching zero levels of 5hmC, the variation increased exponentially. The instability of 5hmC measurements in those probes may result in false positive or negative findings in downstream epigenome-wide association study analyses. To eliminate the probes with an unstable 5hmC signal, the authors developed the *oxBScut* function in the package to make a recommendation for a 5hmC cutoff; the recommended cutoff is derived from minimizing the coefficient of variation of 5hmC signals across samples. As 5hmC level and DNA susceptibility to oxBS treatment vary by tissue and cell type [9,45,48], the authors recommend applying the *oxBScut* for 5hmC cutoff selection with specificity to tissue and cell type.

On Illumina HumanMethylation BeadChips, internal normalization control probes are included for quality control (NORM A/G/C/T). 93 and 85 pairs of internal normalization control probes are contained in the 450K array and the EPIC array, respectively [40,49]. In each pair, one probe will incorporate a base A or T in the red channel, and the other probe will include a base G or C in the green channel. The probes are used to evaluate the array performance in two color channels and can further be used for correcting intensity-related dye biases [26]. On publicly available datasets, sample mislabeling can impede research communication and data reusability and even produce spurious results [50]. With the heterogeneity and abundance of publicly available methylation data, mislabeling can often happen [51]. Especially for paired samples, incorrect labels are a common mistake. For BS and oxBS paired data, one can identify mislabeled samples when calculating the 5mC and 5hmC levels, as the 5hmC are measured as BS methylation minus oxBS methylation. If the 5hmC signal is estimated to be under zero, the sample may be mislabeled. However, random assay noise and artifacts can generate negative 5hmC estimation with correct labels [20]. To the authors' knowledge, no other standards have been developed to check the labeling of BS and oxBS paired samples for quality control. The authors compared the signal intensities of internal normalization control probes between BS-treated and oxBS-treated samples. They developed a method to distinguish BS-converted samples and oxBS-converted samples in a tandem treated pair. The consistent lower intensities of oxBS-converted samples in internal normalization control probes allowed a quality check for BS and oxBS sample labeling in the data preprocessing pipeline. By applying the method to publicly available datasets, the authors identified multiple mislabeled BS and oxBS samples. They reached out to the curators of the mislabeled samples and asked for reinspection of the sample labels. The authors believe that inserting the method into the tandem treated BS and oxBS methylation data preprocessing pipeline would benefit the scientific community by eliminating the relatively common BS and oxBS mislabeling issue.

While this analysis generally points to valid results on high-quality probe loss in oxBS-treated samples compared with BS-treated samples and 5hmC signal instability compared with 5mC signal, there are some limitations. First, although the loss of high-quality probes is consistent in oxBS-treated samples, the variation of the magnitude of the problem by different tissue used is not elucidated. The problem might be more prevalent in tissue types with more fragile DNA [45,46]. Additional research is needed to

investigate the susceptibility of the DNA from specific tissues to damage by oxBS treatment. Second, although multiple technical variables are included in this study, further investigation on how other technical effects impact the DNA reaction to oxBS treatment is needed. For instance, technical variation caused by using different protocols on sample freezing, thawing and storage may also significantly impact how DNA reacts to oxBS conversion [52,53]. Third, 5mC and 5hmC stability was investigated in only one cell type, HFOs. Future research on validation using multiple tissue types is necessary. Also, how technical variables can impact signal stability needs further investigation. Finally, when comparing data quality between the tandem nucleic acid extraction kit samples and the DNA-specific extraction kit samples, multiple studies with different sample types and protocols were pooled. For a better comparison, the same samples with DNA extracted using the tandem nucleic acid extraction kits and the DNA-specific extraction kits are needed. Also, only one type of commercial tandem nucleic acid extraction kit was included in the authors' datasets, and they can only speculate about the reasons behind those alterations, such as lower DNA yield or the presence of kit contaminants interfering with the oxidation step. Modifications in the DNA extraction protocols to reduce these problems are also needed in future analyses.

Conclusion

Our study demonstrated a consistent decrease in the probe call rate in oxBS-treated samples compared with BStreated samples. Here, we emphasized preserving DNA integrity during the oxBS conversion operation to minimize information loss. We recommend using DNA-specific extraction kits instead of tandem nucleic acid extraction kits on DNA purification and extraction. To mitigate the instability of 5hmC signals, we recommend using oxBSQC to remove highly unstable probes. We recommend a quality check using the oxBSQC on possible mislabeling on BS and oxBS samples using the internal normalization control probes when using publicly available datasets.

Future perspective

DNA hydroxymethylation has been found to serve more than just an intermediate role in the demethylation process. While 5mC is a well-studied biomarker for cancer diagnosis and prognosis, 5hmC biomarkers have been developed for early cancer detection and survival in recent years, especially in cfDNA. As more effort will be put into studies using 5hmC to predict cancer and other diseases, our study emphasizes the importance of preserving DNA integrity when using tandem BS and oxBS-treated DNA to measure 5mC and 5hmC. A reasonable evaluation of DNA quality is required when using these approaches. Future studies should pay attention to preserving DNA integrity during BS and oxBS conversion and develop methods to more efficiently measure 5hmC.

Summary points

- A consistent decrease in probe call rates was observed in oxidative bisulfite samples compared with bisulfite samples.
- Concerns are raised regarding reduced power in 5-hydroxymethylcytosine (5hmC) epigenome-wide association studies with a decreased level of high-quality probes.
- Multiple technical components can contribute to compromised DNA integrity in oxidative bisulfite conversion, especially DNA extraction kits.
- DNA-specific extraction kits are recommended for DNA extraction for oxidative bisulfite treatment compared with tandem nucleic acid extraction kits.
- 5hmC signal is relatively unstable compared with 5-methylcytosine signal, especially for probes with low levels of 5hmC.
- Using oxBSQC to eliminate unstable probes with 5hmC signal is recommended.
- Using internal normalization controls for quality checks on bisulfite and oxidative bisulfite labeling is recommended.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/epi-2021-0490

Author contributions

Z Zhang, MK Lee and LA Salas designed the study. Z Zhang acquired data and performed data analyses of the paper. Z Zhang, MK Lee, L Perreard, BC Christensen and LA Salas participated in the interpretation of data for the work. Z Zhang and MK Lee were responsible for the initial draft of the work. Z Zhang, MK Lee, L Perreard, BC Christensen, KT Kelsey and LA Salas participated in the final drafting and critical revision for important intellectual content. Z Zhang, MK Lee, L Perreard, BC Christensen, KT Kelsey and LA Salas read and approved the final manuscript.

Financial & competing interests disclosure

This study is funded by United States Congressionally Directed Medical Research Programs Funding (number: W81XWH-20-10778), United States National Institute of General Medical Sciences (number: P20GM104416/8299) and National Institutes of Health (number: R01CA216265, R01CA253976, R01CA207360). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

For local human samples, the authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Data sharing statement

The datasets analyzed during the current study are publicly available on the Gene Expression Omnibus data repository: https://www. ncbi.nlm.nih.gov/geo/ (accession numbers: GSE182919, GSE152561, GSE63179, GSE71328, GSE71719, GSE73895, GSE104723, GSE105109, GSE116402, GSE133062). For publicly available data, the original data sources are referenced in the manuscript methods.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit www.creativecommons.org/licenses/by-nc-nd/4.0

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Bogdanovic O, Lister R. DNA methylation and the preservation of cell identity. *Curr. Opin. Genet. Dev.* 46, 9–14 (2017).
2. Houseman EA, Accomando WP, Koestler DC *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13, 86 (2012).
3. Sharp AJ, Stathaki E, Migliavacca E *et al.* DNA methylation profiles of human active and inactive X chromosomes. *Genome Res.* 21(10), 1592–1600 (2011).
4. Ponomaryova AA, Rykova EY, Gervas PA, Cherdyntseva NV, Mamedov IZ, Azhikina TL. Aberrant methylation of LINE-1 transposable elements: a search for cancer biomarkers. *Cells* 9(9), doi:org/10.3390/cells9092017 (2020).
5. Cui XL, Nie J, Ku J *et al.* A human tissue map of 5-hydroxymethylcytosines exhibits tissue specificity through gene and enhancer modulation. *Nat. Commun.* 11(1), 6161 (2020).
6. Tahiliani M, Koh KP, Shen Y *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 324(5929), 930–935 (2009).
7. Ito S, Shen L, Dai Q *et al.* TET proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* 333(6047), 1300–1303 (2011).
8. He YF, Li BZ, Li Z *et al.* TET-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* 333(6047), 1303–1307 (2011).
9. Globisch D, Munzel M, Muller M *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS ONE* 5(12), e15367 (2010).
10. Song CX, He C. Potential functional roles of DNA demethylation intermediates. *Trends Biochem. Sci.* 38(10), 480–484 (2013).
11. Xu T, Gao H. Hydroxymethylation and tumors: can 5-hydroxymethylation be used as a marker for tumor diagnosis and treatment? *Hum. Genomics* 14(1), 15 (2020).
12. Spiers H, Hannon E, Schalkwyk LC, Bray NJ, Mill J. 5-hydroxymethylcytosine is highly dynamic across human fetal brain development. *BMC Genomics* 18(1), 738 (2017).
13. Tsagaratou A, Aijo T, Lio CW *et al.* Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation. *Proc. Natl Acad. Sci. USA* 111(32), e3306–e3315 (2014).
14. Guler GD, Ning Y, Ku CJ *et al.* Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA. *Nat. Commun.* 11(1), 5270 (2020).

• **A recent research paper on 5-hydroxymethylcytosine (5hmC) in cfDNA for pancreatic cancer detection showed the potential of 5hmC for being a reliable biomarker.**

15. Song CX, Yin S, Ma L *et al.* 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Res.* 27(10), 1231–1242 (2017).

16. Zeng C, Stroup EK, Zhang Z, Chiu BC, Zhang W. Towards precision medicine: advances in 5-hydroxymethylcytosine cancer biomarker discovery in liquid biopsy. *Cancer Commun. (Lond.)* 39(1), 12 (2019).

17. Cao F, Wei A, Hu X *et al.* Integrated epigenetic biomarkers in circulating cell-free DNA as a robust classifier for pancreatic cancer. *Clin. Epigenetics* 12(1), 112 (2020).

18. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* 19(3), 129–147 (2018).

19. Field SF, Beraldi D, Bachman M, Stewart SK, Beck S, Balasubramanian S. Accurate measurement of 5-methylcytosine and 5-hydroxymethylcytosine in human cerebellum DNA by oxidative bisulfite on an array (OxBS-array). *PLoS ONE* 10(2), e0118202 (2015).

• **An early paper that demonstrated accurate measurement of 5-methylcytosine and 5hmC using tandem bisulfite and oxidative bisulfite treatment.**

20. Houseman EA, Johnson KC, Christensen BC. OxyBS: estimation of 5-methylcytosine and 5-hydroxymethylcytosine from tandem-treated oxidative bisulfite and bisulfite DNA. *Bioinformatics* 32(16), 2505–2507 (2016).

21. Booth MJ, Ost TW, Beraldi D *et al.* Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat. Protoc.* 8(10), 1841–1851 (2013).

•• **The second-generation protocol for BS and oxBS-seq, which demonstrated important technical variables in the pipeline that impact the results.**

22. Tanaka K, Okamoto A. Degradation of DNA by bisulfite treatment. *Bioorg. Med. Chem. Lett.* 17(7), 1912–1915 (2007).

23. Olova N, Krueger F, Andrews S *et al.* Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.* 19(1), 33 (2018).

24. Cadet J, Davies KJA. Oxidative DNA damage & repair: an introduction. *Free Radic. Biol. Med.* 107, 2–12 (2017).

25. Nugen ATGC. TrueMethyl® oxBS module user guide. (2019).
M01481 v3 User Guide: TrueMethyl oxBS module 4977.pdf(nugen.com)

26. Wilhelm-Benartzi CS, Koestler DC, Karagas MR *et al.* Review of processing and analysis methods for DNA methylation array data. *Br. J. Cancer* 109(6), 1394–1402 (2013).

27. Aryee MJ, Jaffe AE, Corrada-Bravo H *et al.* Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30(10), 1363–1369 (2014).

28. Zhou W, Triche TJ Jr, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res.* 46(20), e123 (2018).

•• **The paper demonstrated that p-value with out-of-band array hybridization is better to avoid false methylation signal compared to negative detection p-value.**

29. Heiss JA, Just AC. Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses. *Clin. Epigenetics* 11(1), 15 (2019).

30. Triche TJ Jr, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* 41(7), e90 (2013).

31. Matsubara K, Kagami M, Nakabayashi K *et al.* Exploration of hydroxymethylation in Kagami–Ogata syndrome caused by hypermethylation of imprinting control regions. *Clin. Epigenetics* 7, 90 (2015).

32. Green BB, Houseman EA, Johnson KC *et al.* Hydroxymethylation is uniquely distributed within term placenta, and is associated with gene expression. *FASEB J.* 30(8), 2874–2884 (2016).

33. Johnson KC, Houseman EA, King JE, Von Herrmann KM, Fadul CE, Christensen BC. 5-Hydroxymethylcytosine localizes to enhancer elements and is associated with survival in glioblastoma patients. *Nat. Commun.* 7, 13177 (2016).

34. Smith AR, Smith RG, Pishva E *et al.* Parallel profiling of DNA methylation and hydroxymethylation highlights neuropathology-associated epigenetic variation in Alzheimer's disease. *Clin. Epigenetics* 11(1), 52 (2019).

35. Zhang X, Chen X, Weirauch MT *et al.* Diesel exhaust and house dust mite allergen lead to common changes in the airway methylome and hydroxymethylome. *Environ. Epigenet.* 4(3), dvy020 (2018).

36. Ringh MV, Hagemann-Jensen M, Needhamsen M *et al.* Tobacco smoking induces changes in true DNA methylation, hydroxymethylation and gene expression in bronchoalveolar lavage cells. *EBioMedicine* 46, 290–304 (2019).

37. Patel PG, Selvarajah S, Guerard KP *et al.* Reliability and performance of commercial RNA and DNA extraction kits for FFPE tissue cores. *PLoS ONE* 12(6), e0179732 (2017).

38. Mathieson W, Thomas GA. Simultaneously extracting DNA, RNA, and protein using kits: is sample quantity or quality prejudiced?*Anal. Biochem.* 433(1), 10–18 (2013).

• **A research paper on the performance variation of different DNA/RNA extraction kits used.**

39. Perkins NJ, Schisterman EF. The inconsistency of 'optimal' cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am. J. Epidemiol.* 163(7), 670–675 (2006).

40. Xu Z, Langie SA, De Boever P, Taylor JA, Niu L. RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip. *BMC Genomics* 18(1), 4 (2017).

41. Richa R, Sinha RP. Hydroxymethylation of DNA: an epigenetic marker. *EXCLI J.* 13, 592–610 (2014).

42. Pfeifer GP, Xiong W, Hahn MA, Jin SG. The role of 5-hydroxymethylcytosine in human cancer. *Cell Tissue Res.* 356(3), 631–641 (2014).

43. Ficz G, Gribben JG. Loss of 5-hydroxymethylcytosine in cancer: cause or consequence? *Genomics* 104(5), 352–357 (2014).

44. Liu Y, Siejka-Zielinska P, Velikova G *et al.* Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* 37(4), 424–429 (2019).

45. Huebner K. Molecular biology: DNA fragility put into context. *Nature* 470(7332), 46–47 (2011).

46. Letessier A, Millot GA, Koundrioukoff S *et al.* Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. *Nature* 470(7332), 120–123 (2011).

47. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 8(3), 389–399 (2016).

48. Li W, Liu M. Distribution of 5-hydroxymethylcytosine in different human tissues. *J. Nucleic Acids* 2011, 870726 (2011).

    - **An early research paper demonstrated the distribution of 5hmC across different human tissues.**

49. Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res.* 44(3), e20 (2016).

50. Toker L, Feng M, Pavlidis P. Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies. *F1000Res* 5, 2103 (2016).

51. Heiss JA, Just AC. Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. *Clin. Epigenetics* 10, 73 (2018).

    - **A research paper raised the problem of mislabeling in publicly available datasets.**

52. Shao W, Khin S, Kopp WC. Characterization of effect of repeated freeze and thaw cycles on stability of genomic DNA using pulsed field gel electrophoresis. *Biopreserv. Biobank.* 10(1), 4–11 (2012).

53. Huang LH, Lin PH, Tsai KW *et al.* The effects of storage temperature and duration of blood samples on DNA and RNA qualities. *PLoS ONE* 12(9), e0184692 (2017).

frontiers in Bioinformatics

ORIGINAL RESEARCH
published: 21 March 2022
doi: 10.3389/fbinf.2022.835591

Check for updates

OPEN ACCESS

Edited by:
Christine Nardini,
National Research Council (CNR), Italy

Reviewed by:

Jin-Xing Liu,
Qufu Normal University, China
Yongchun Zuo,
Inner Mongolia University, China

*Correspondence:

Devin C. Koestler dkoestler@kumc.edu

Specialty section:
This article was submitted to Genomic Analysis, a section of the journal Frontiers in Bioinformatics

Received: 14 December 2021
Accepted: 15 February 2022
Published: 21 March 2022

Citation:
Bell-Glenn S, Thompson JA, Salas LA and Koestler DC (2022) A Novel Framework for the Identification of Cellular Mixtures. Front. Bioinform. 2:835591. doi: 10.3389/fbinf.2022.835591

Abbreviations: CMD, Cell Mixture Deconvolution; DML, Differentially Methylated Loci; DNAm, DNA methylation; DSC, Dispersion Separability Criterion; FDR, False Discovery Rate; EWAS, Epigenome Wide Association Studies; RESET, Reference Library Selection without a Training Dataset; RMSE, Root Mean Squared Error.

Reference DNA Methylation Libraries for Reference-Based Deconvolution of Cellular Mixtures

Keywords: reference-based deconvolution, IDOL, cell heterogeneity, cell proportion estimation, DNA methylation, EWAS

A Novel Framework for the Identification of Reference DNA Methylation Libraries for Reference-Based Deconvolution of Cellular Mixtures

Shelby Bell-Glenn[1], Jeffrey A. Thompson[1], Lucas A. Salas[2] and Devin C. Koestler[1]*

[1]Department of Biostatistics and Data Science, University of Kansas Medical Center, Kansas City, KS, United States,
[2]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, NH, United States

Reference-based deconvolution methods use reference libraries of cell-specific DNA methylation (DNAm) measurements as a means toward deconvoluting cell proportions in heterogeneous biospecimens (e.g., whole-blood). As the accuracy of such methods depends highly on the CpG loci comprising the reference library, recent research efforts have focused on the selection of libraries to optimize deconvolution accuracy. While existing approaches for library selection work extremely well, the best performing approaches require a training data set consisting of both DNAm profiles over a heterogeneous cell population and gold-standard measurements of cell composition (e.g., flow cytometry) in the same samples. Here, we present a framework for reference library selection without a training dataset (RESET) and benchmark it against the Legacy method (minfi:pickCompProbes), where libraries are constructed based on a pre-specified number of cell-specific differentially methylated loci (DML). RESET uses a modified version of the Dispersion Separability Criteria (DSC) for comparing different libraries and has four main steps: 1) identify a candidate set of cell-specific DMLs, 2) randomly sample DMLs from the candidate set, 3) compute the Modified DSC of the selected DMLs, and 4) update the selection probabilities of DMLs based on their contribution to the Modified DSC. Steps 2–4 are repeated many times and the library with the largest Modified DSC is selected for subsequent reference-based deconvolution. We evaluated RESET using several publicly available datasets consisting of whole-blood DNAm measurements with corresponding measurements of cell composition. We computed the RMSE and R^2 between the predicted cell proportions and their measured values. RESET outperformed the Legacy approach in selecting libraries that improve the accuracy of deconvolution estimates. Additionally, reference libraries constructed using RESET resulted in cellular composition estimates that explained more variation in DNAm as compared to the Legacy approach when evaluated in the context of epigenome-wide association studies (EWAS) of several publicly available data sets. This finding has implications for the statistical power of EWAS. RESET combats potential challenges associated with existing approaches for reference library assembly and thus, may serve as a viable strategy for library construction in the absence of a training data set.

INTRODUCTION

Epigenome-wide association studies (EWAS) explore epigenetic variation, specifically in DNA methylation (DNAm). These studies provide insight into how environmental factors can influence disease, as well as indirectly inform potential novel therapeutics to treat human diseases (Rakyan et al., 2011; Michels et al., 2013; Flanagan, 2015). A well-recognized challenge in the statistical analysis and interpretation of EWAS, particularly EWAS that involve DNAm profiling of heterogeneous tissue types (e.g., whole blood, peripheral blood mononuclear cells, etc.), arises from the cell specificity of DNA methylation. It has been well documented that DNAm analyses of heterogeneous cell populations are prone to a decreased statistical power for detecting CpG-specific methylation effects or, worse, confounding and misleading results (Reinius et al., 2012; Jaffe and Irizarry, 2014; Liang and Cookson, 2014; Houseman et al., 2015). One obvious way to deal with the potential for confounding due to cellular heterogeneity is to adjust downstream statistical models for cellular composition (e.g., by adding terms to statistical models that reflect the proportion of each cell type within the heterogeneous sample used for methylation profiling). When available, cell composition is measured using complete blood cell counts (CBC) with differential or flow cytometry in the same biospecimens used for DNA methylation profiling (Houseman et al., 2015). However, these measurements are not routinely collected for most EWAS due to added cost to the study and technical limitations of these approaches (e.g., both CBC and flow cytometry require fresh samples and are not feasible for stored specimens).

To address these issues, methods for DNAm-based cell mixture deconvolution (CMD) have been developed. CMD relies on the assumption that methylation signatures for a heterogeneous sample can be thought of as a weighted mixture of the unique methylation signature of each cell type making up the sample (Houseman et al., 2012; Newman et al., 2015; Houseman et al., 2016; Koestler et al., 2016; Teschendorff et al., 2017; Titus et al., 2017; Decamps et al., 2020; Scherer et al., 2020). Thus, CMD makes use of the cell-specificity of DNAm to estimate the cell proportions in samples with heterogeneous cell composition. CMD includes both reference-free and referencebased approaches. Reference-free deconvolution obtains cell type-specific proportions when DNAm profiles in purified cell types are not available (Houseman et al., 2016; Decamps et al., 2020; Scherer et al., 2020). While broadly applicable across a range of different tissue types, a major limitation of reference-free deconvolution methods is that they are incapable of resolving the specific identity of the individual cell types for which deconvolution estimates are obtained. For this reason and because the identities of individual cell types are often of interest themselves (Koestler et al., 2017; Wiencke et al., 2017; Grieshober et al., 2018; Grieshober et al., 2021), reference-based deconvolution methods are often preferred over reference-free approaches when cell-specific methylation signatures are available for the cell types expected to be present in the tissue source being used for DNAm assessment. Reference-based methods use reference signatures of DNAm for the underlying cell types that make up a sample (reference libraries) (Houseman et al., 2012; Newman et al., 2015; Koestler et al., 2016; Teschendorff et al., 2017). Several reference-based methods for deconvoluting DNAm data have been proposed in recent years, including support vector regression (CIBERSORT), robust partial correlation (EpiDish), and constrained projection/quadratic programming (Houseman et al., 2012; Newman et al., 2015; Teschendorff et al., 2017). While the specific statistical method used for CMD has some impact on the accuracy of cell proportion estimates, the major factor underlying the performance of these methods is the reference library, or collection of cell-specific differentially methylated loci (DML), used as the basis for deconvolution (Koestler et al., 2016; Teschendorff et al., 2017; Titus et al., 2017). Reference libraries that result in accurate deconvolution estimates are comprised of cell-specific methylation biomarkers that exhibit distinct patterns of methylation across cell types.

To date, there have been several approaches for constructing reference libraries for DNAm-based CMD. The first approach assembles libraries using the $J^P \ll J$ CpGs that have the largest F-statistic computed from fitting a series ANOVA models independently to each CpG based on DNA methylation data of purified isolated cell types (Houseman et al., 2012; Koestler et al., 2016). For a given CpG, the ANOVA-based approach tests the null hypothesis that the mean beta value is the same across all K cell types versus the alternative hypothesis that the mean beta differs between at least two of the K cell types. Reference libraries assembled using this approach are limited in that they tend to be comprised of CpGs that distinguish some cell types well (e.g., neutrophils and lymphocyte subtypes, in the context of white blood cell types) but not cell types that share a common lineage (e.g., CD4T and CD8T) (Koestler et al., 2016). Improvements were made to the ANOVA approach by instead using the top hypermethylated and hypomethylated CpGs for each cell type, hereafter referred to as the Legacy approach. When using the Legacy approach, CpGs are selected by their t-statistic, computed from a series of two-sample t-tests fit independently to each CpG that test whether the mean beta value of each cell type is the same as the mean beta value across the K-1 remaining cell types. CpGs are rank-ordered based on their t-statistic, and those with the largest and smallest statistics are used to assemble a reference library for CMD (Aryee et al., 2014; Jaffe and Irizarry, 2014; Koestler et al., 2016). The Legacy approach is implemented in the Bioconductor package minfi using the function pickCompProbes (Aryee et al., 2014). While the Legacy method is an improvement over the ANOVA approach (Koestler et al., 2016), it is limited because it requires users to make arbitrary decisions about the number of hypermethylated and hypomethylated CpGs for each cell type used in assembling reference libraries. Further, the assumption of an equal number of DMLs for each cell type in the creation of reference libraries may hinder the accuracy of CMD-based estimates of cell proportions. Yet another approach, proposed

by Koestler et al., called IDOL, iteratively searches CpGs in the construction of reference libraries that maximize the prediction accuracy of CMD-based estimates of cell proportions (Koestler et al., 2016). IDOL improves upon both the ANOVA and Legacy method in prediction performance but requires a training dataset consisting of whole-blood DNA methylation data and corresponding cellular composition measurements (e.g., flow cytometry) (Wiencke et al., 2017). This is not ideal because, in order to use IDOL, a new training data set needs to be created every time a reference signature is added for a new cell type. Further, IDOL also requires users to choose the size of the reference library in advance, which introduces arbitrary decisions in choosing a reference library.

Motivated by the shortcomings of these existing approaches, we propose an algorithm for reference library selection without a training dataset (RESET) that does not require users to choose the size of the reference library in advance, nor does it require a training data set like IDOL. Our approach is further driven by the desire to save researchers time and money; since a training data set does not need to be acquired or updated every time the reference signature for a new cell type becomes available. Further, choosing a reference library size in advance is an arbitrary decision that does not guarantee the quality of a library. RESET is similar in principle to IDOL but utilizes a modified version of the Dispersion Separability Criteria (DSC) as a metric for evaluating and ranking potential reference libraries for reference-based deconvolution. The Cancer Genome Atlas originally developed the DSC for detecting batch effects as a measure of the between vs. within batch dispersion (Akbani et al., 2020). We hypothesize that higher values of this metric correspond to better discrimination across cell types and, as a consequence, better prediction performance when used for reference-based deconvolution. In what follows, we first describe our Modified DSC metric and the proposed RESET framework for identifying reference libraries. We benchmark the performance of RESET compared to the Legacy method using three publicly available data sets across a range of different parameter settings. Next, we explore the impact that different libraries have on the operating characteristics of EWAS, using both simulations and data applications involving two publicly available data sets. We finish with a discussion of our findings, limitations of our study, and opportunities for future work.

MATERIALS AND METHODS

In what follows, we describe the materials and methods used in this research. Figure 1 provides a workflow that summarizes the data sets used and analyses that were performed.

Cell Mixture Deconvolution

Before we introduce our proposed method, we first provide a brief description of cell mixture deconvolution. Let Y_i $[Y_{i1}, Y_{i2}, \ldots, Y_{iJ^p}]$, where $0 \le Y_{ij} \le 1$, be the methylation beta-values for CpGs $j \in \{1, 2, \ldots, J^p\}$ for sample $i \in \{1, 2, \ldots, N\}$. Assume that for sample i, DNA methylation is profiled in a heterogeneous tissue type (e.g., whole-blood), which has K underlying cell types and whose proportions in sample i are denoted w_i $[w_{i1}, w_{i2}, \ldots, w_{iK}]$. As previously described in Houseman et al. (2012), the methylation signature of Y_i is assumed to be a weighted mixture of the DNA methylation signature of each of the K underlying cell types (Houseman et al., 2012). That is:

$$E[Y_i] \; w_i \mu^T, \; 0 \le w_{ik} \le 1 \text{ and } \sum_{k=1}^{K} w_{ik} \le 1$$

where μ is a J^p x K matrix of mean methylation beta values. The rows of μ correspond to the same ordering of the J^p CpGs as in Y_i and the columns of μ are the K distinct cell types. Following the above equation, the Houseman approach to CMD proceeds by finding the w_i that minimizes the squared-error loss function between Y_i and $w_i \mu^T$,

$$argmin_{w_i} Y_i - w_i \mu^T$$

subject to the above-mentioned constraints on w_{ik}. Since μ is unknown, the sample mean, M, is used and can be estimated using a reference methylation data set consisting of DNA methylation data on isolated, purified cell populations (Houseman et al., 2012; Reinius et al., 2012).

CMD is made possible by leveraging the unique methylation signature of each cell type, which is reflected in the columns of M. However, as previously described, the accuracy of cell proportion estimates, $\hat{w}_i$, rely heavily on the J^p CpGs are used as the basis of CMD. Strong reference libraries for CMD are comprised of CpGs whose methylation signature is distinct across each of the K cell types.

Modified Dispersion Separability Criteria

To be able to perform CMD with high accuracy when the true cell proportions are not known, we need a metric that identifies J^p CpGs that discriminate between cell types well. We believe that the Dispersion Separability Criteria (DSC) could be one such metric since it was designed by the Cancer Genome Atlas (TCGA) to identify batch effects in high-dimensional molecular metric in selecting reference libraries for deconvolution found that this DSC and deconvolution accuracy is more complicated. Some cell types are more alike than others due to shared lineages. For example, the methylation pattern of CD8T and CD4T cells is more related than, say, CD8T cells and monocytes because of their shared lymphoid lineage. The

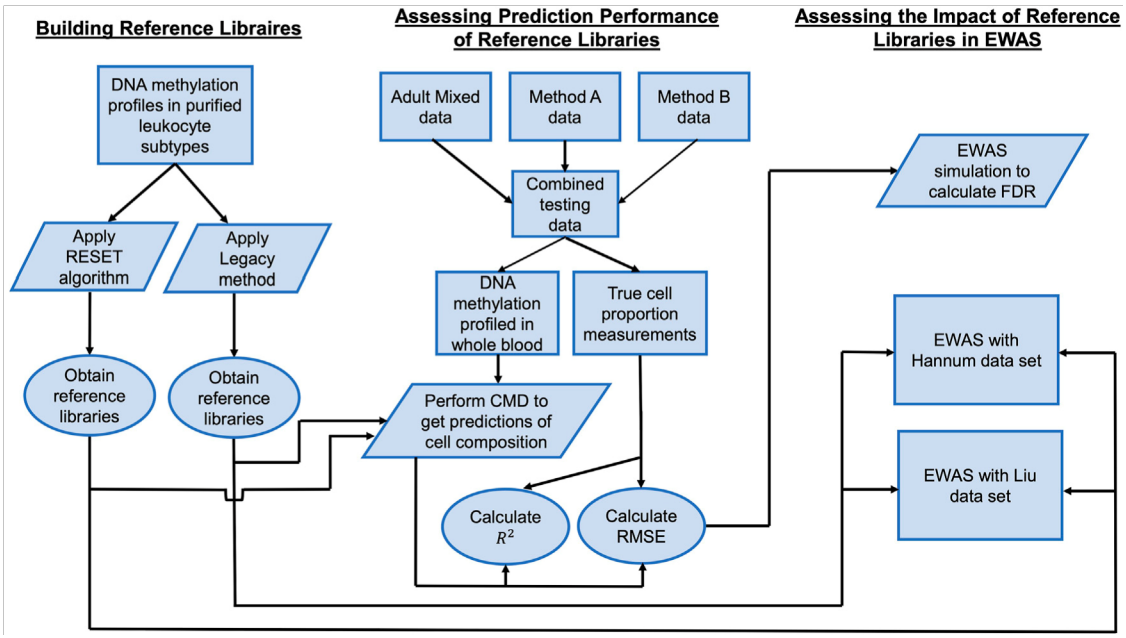


FIGURE 1 | Analytical workflow. A workflow for describing the data sets used and analyses that were performed.

$$\frac{D_b}{D_w}$$

data sets (Akbani et al., 2020). Defined as follows, *DSC*

D_b is a measure of dispersion between batches and D_w is a measure of dispersion within batches. D_b can be thought of as the mean Euclidean distance between batch centroids and the overall centroid, while D_w can be thought of as the mean Euclidean distance between samples of a given batch and the centroid for that batch. More specifically D_b and D_w are defined as:

$$D_b \; trace((S_b))$$

$$D_w \qquad trace\, S_w$$

where S_b is the between batch scatter matrix and S_w is the within batch scatter matrix as described previously by Dy and Brodley (2004) (Dy, 2004). Higher DSC values indicate that there is a greater dispersion between batches than within batches and indicate the presence of batch effects. Thus, we hypothesized that for a given set of CpGs (e.g., library), high values of the DSC would indicate greater dispersion between cell types and that libraries with large DSC values correspond to improved deconvolution accuracy. However, our early exploration of this

To address such challenges, we propose a metric for above-mentioned formulation of the DSC was ill-suited to address such challenges, and poor deconvolution performance ensued.

assembling reference libraries for CMD that involves a slight modification of the DSC. We define the Modified DSC as follows:

$$\underline{Min}D_{pb}{}^p \; \underline{Min}D_w$$

Modified DSC

$D^p{}_b$ represents a vector of the Euclidean distances between every pair of cell type centroids and $D^p{}_w$ is a vector containing the mean Euclidean distances between samples of a given cell type and the centroid for that cell type. More precisely:

$$D_b{}^p \; d_{b1}, d_{b2}, \ldots, d_{bK2}$$

$$D_{wp} \; [d_{w1}, d_{w2}, \ldots, d_{wK}]$$

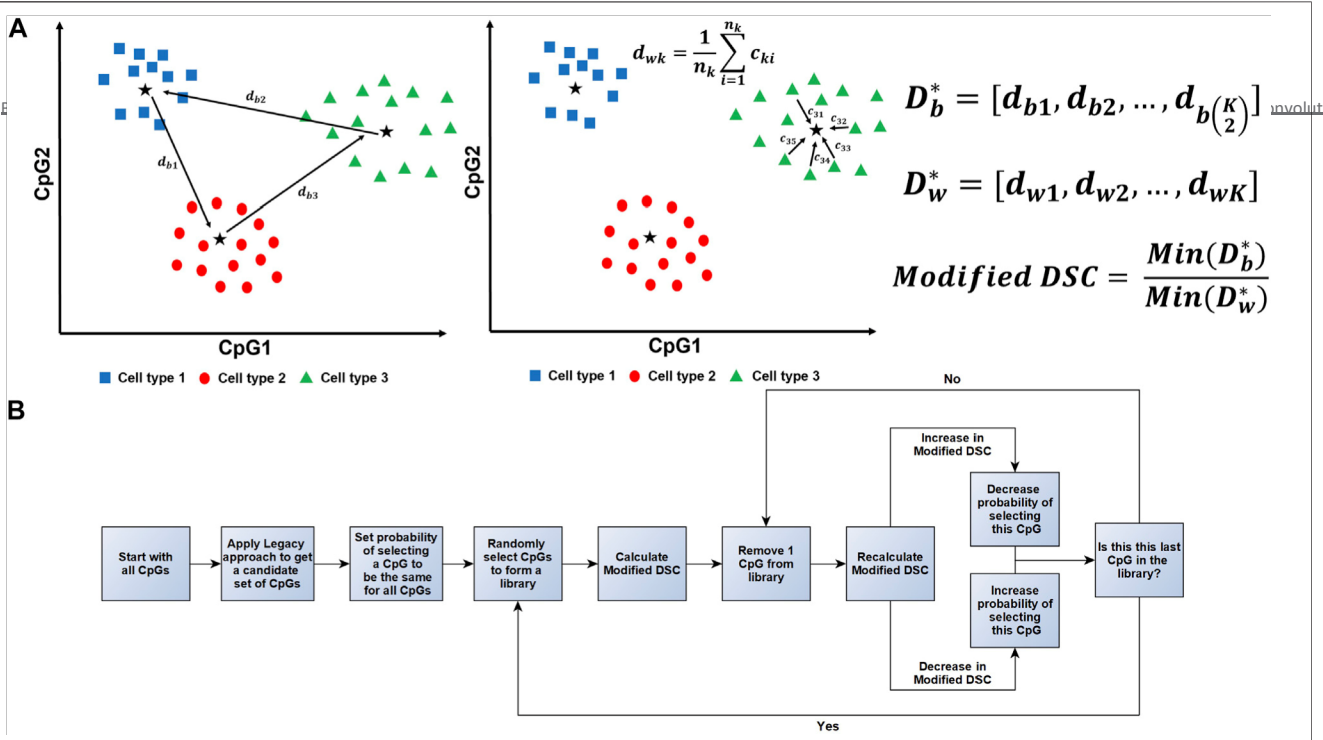


FIGURE 2 | Illustration of the Modified DSC Calculation and Workflow. (A) A conceptual illustration of how the Modified DSC is calculated assuming three different cell types and two CpGs. (A, Left) d_{b1}, d_{b2} and d_{b3} represent the Euclidean distances between pairs of cell-type-specific centroids (denoted as stars). These distances are elements of the vector, D_b^p, which represents the between-cell type dispersion index. (A, Right) c_{ik}, $k = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, n_k$, represent the Euclidean distances between each sample of a given cell type and the centroid (denoted as star) for that cell type where k is an index for cell type, and i is an index for sample. The average within-cell type distance is represented by d_{wk}, where $d_{wk} = \frac{1}{n_k}\sum_{i=1}^{n_k} c_{ik}$, and n_k represents the number of samples for cell type k. These distances are the elements of the vector, D_w^p, which represents the within-cell type dispersion index. The modified DSC is then calculated as the ratio of the minimum between-cell type dispersion and the minimum within-cell type dispersion $\frac{Min(D_b^p)}{Min(D_w^p)}$. (B) Workflow for RESET algorithm to find the optimal library using the Modified DSC metric as described in Algorithm for RESET Utilizing the Modified DSC Metric.

K

where reflects the number of cell-type pairs. Although

2 reference libraries for CMD are typically comprised of hundreds or thousands of CpGs, Figure 2A provides a simplified, conceptual illustration of the calculation of the Modified DSC, assuming three cell types and two CpGs. We chose to utilize the minimum of these distances after exploring various combinations of summary statistics (data not shown). The idea is that high values of the Modified DSC correspond to better discrimination between cell types since it is based on minimum distances. More specifically, if the minimum distance between any two cell type centroids is large and the minimum within-cluster distance is small, this will yield a large Modified DSC and indicate large differences between cell types with respect to their DNAm signature.

Algorithm for RESET Utilizing the Modified DSC Metric

The below algorithm was taken from Koestler et al. (2016) and adjusted to utilize the Modified DSC when a training data set is not available (Koestler et al., 2016). A further illustration of this algorithm can be found in Figure 2B.

Step 0: Assembling a candidate reference library:

a. Fit a series of two-sample t-tests to the J CpGs to compare mean methylation (beta-values) between each cell type against the mean beta-values across the remaining K-1 cell types.

b. Identify L_2 CpGs with the largest t-statistics and the L_2 CpGs with the smallest t-statistics for each of the K cell types. Here L represents the number of cell-specific DMLs we desire.

c. Create a set of CpGs, Q, identified in (b). Specifically, Q is comprised of P = LK DMLs and makes up the candidate search space for the subsequent steps.

It should be noted that the choice of L is completely arbitrary, and trade-offs must be made depending on how large L is chosen to be. Large values of L broaden the search space for possible cellspecific DMLs, but also increase the computational burden in downstream steps. Small values of L correspond to less computational burden but risk excluding viable cell-specific DMLs.

Step 1: Random assembly of reference libraries:

a. At iteration ℓ, J^p CpGs are randomly chosen from Q with probability π_j^ℓ, $j \in \{1, \ldots, P\}$. At the first iteration, $\ell = 0$, every CpG among the P candidate DMLs has an equal probability of being chosen. That is, $\forall\, j \in \{1, \ldots, P\}$, $\pi_j^0 = \frac{1}{P}$.

b. Let $Q^\ell \subset Q$ denote the randomly assembled reference library, comprised of J^p randomly selected CpGs at iteration ℓ.

Note that J^p, the number of CpGs that make up the library Q^ℓ, is also randomly selected from a uniform distribution with a minimum of $j_{min} \geq k$ and a maximum of $j_{max} \leq P$.

Step 2: Calculation of the Modified DSC using the randomly assembled library:

a. Using the random library Q^{ℓ}, calculate the Modified DSC. We will denote this $MDSC^{\ell}$.

Step 3: Assessing the contribution of each CpG to the Modified DSC using leave-one-out procedure:

a. Each of the J^{p} CpGs that make up Q^{ℓ} are iteratively removed to obtain sets of CpGs, Q^{ℓ}_{j}, which includes ℓ - all CpGs in Q, except CpG j.

b. Step 2 is repeated for each reduced library Q^{l}_{-j} so that we have J^{p} new Modified DSCs, denoted as $MDSC_{-}^{\ell}{}_{1}, MDSC_{-}^{\ell}{}_{2}, \ldots, MDSC_{-}^{\ell}{}_{Jp}$.

In subsequent iterations, we seek to retain CpGs whose inclusion in the reference library results in large values of the Modified DSC. Calculating $MDSC^{\ell}{}_{j}$ gives us a framework for ℓ updating selection probabilities of each CpG in. Q.

Step 4: Updating selection probabilities:

a. For each of the J^{p} CpGs in the reduced library, Q^{ℓ}, we update the probability of selecting CpG j as follows:

$$\frac{MDSC^{\ell} r}{MDSC_{-j}\,\pi_{j}^{\ell+1}}$$

$$\pi_{j}^{\ell} r$$

Thus, the greater the contribution CpG j has on the value of the Modified DSC, the more influence it has on the probability of being selected in future iterations. More specifically, if removing CpG j from Q^{ℓ} results in a smaller value of the Modified DSC, r will be greater than 1 and thus $\pi_{j}^{\ell+1}$ will increase. If removing CpG j from Q^{ℓ} results in a larger value of the modified DSC, r will be less than 1 and thus $\pi_{j}^{\ell+1}$ will decrease. How much π_{j}^{ℓ} changes from iteration to iteration depends on the contribution of CpG j to the Modified DSC.

b. After the probabilities for the J^{p} CpGs in the random library, Q^{ℓ}, have been updated, the probabilities of all P CpGs in Q are then scaled to sum to 1.

Step 5: Continue iteration: Using the updated probabilities, $\pi_{j}^{\ell+1}\ j \in \{1, \ldots, P\}$, repeat steps 1–4. The final solution is the library made up of the J^{p} CpGs, which yielded the largest value of the Modified DSC.

Datasets

We now describe the DNA methylation array data sets used for this research. For more information about how the data was preprocessed and what measures were taken to ensure quality control, we refer readers to Koestler et al. (2016).

Cell Mixture Reconstruction Experiment

Purified cells taken from normal human subjects were purchased from AllCells LLC (Emeryville, CA). Namely, granulocytes, monocytes, CD4T, CD8T, natural killer, and B cells. DNA was extracted from the purified blood leukocyte subtypes, dsDNA was quantified, and then DNA extracted from purified leukocyte subtypes were mixed in preplanned proportions to reconstruct two different sets with n = 6 samples in each. One set of samples was reconstructed to contain similar proportions of the purified leukocyte subtypes. In contrast, the other was reconstructed to resemble the proportions in normal human adults′ peripheral blood. We will refer to these data sets as MethodA and MethodB, respectively. These DNA samples were bisulfite modified, and epigenome-wide DNA methylation assessment was done using the Illumina HumanMethylation450 array platform (Koestler et al., 2016). These data are publicly available in the Gene Expression Omnibus (GEO) repository (GEO Accession ID: GSE77797).

Adult Whole Blood Samples

Another n = 6 whole blood (WB) samples were taken from healthy adult donors. Immune cell profiling data for these samples, gathered from flow cytometry, were procured from AlCells LLC. We will refer to this data set as AdultMixed. This data set followed the same protocol as above and was assayed using the Illumina HumanMethylation450 array platform (Koestler et al., 2016). These data are publicly available (GEO Accession ID: GSE77797).

Reference DNA Methylomes for Isolated Leukocyte Subtypes

A publicly available dataset (GEO Accession ID: GSE35069) was used to identify cell-type-specific DMLs and for the construction of reference libraries for CMD. This data set is comprised of epigenome-wide DNA methylation profiles in purified leukocyte subtypes (the same six leukocytes described above) in n = 6 healthy non-diseased subjects. We will refer to this data set as the Reference data set. Further details about this data set can be found elsewhere (Reinius et al., 2012).

Additional DNA Methylation Data Sets

We also used two large publicly available blood-derived DNA methylation data sets (GEO Accession IDs: GSE42861 and GSE40279). Together, these two data sets are comprised of whole-blood DNA methylation data on more than 1,200 adult subjects. The purpose of these data sets was to better understand the implications of the method/technique used for assembling reference libraries for CMD in terms of an EWAS. The Liu dataset (GEO Accession ID: GSE42861) consists of blood-derived DNA methylation data on 689 human subjects, of which n = 354 were rheumatoid arthritis cases, and n = 335 were non-diseased control patients (Liu et al., 2013). The Hannum data set (GEO Accession ID: GSE40279) consists of blood-derived DNA methylation on 656 non-diseased patients (Hannum et al., 2013). Epigenomewide DNA methylation

assessment was done using the Illumina HumanMethylation450 array platform for both data sets.

Application and Assessment of the Modified DSC Metric

Proof-Of-Principle Example to Illustrate the Modified DSC Metric

To examine our hypothesis that larger values of the Modified DSC reflect libraries that better discriminate cell types, we conducted a proof-of-principle simulation study where we simulated data under different circumstances and calculated the Modified DSC. We simulated 100 features for three different cell types with six samples of each cell type. We allowed the percentage of features that exhibit a difference across the cell types to vary between 0% and 100%. We further allowed for the magnitude of the difference in the mean feature values to vary between 1 and 10 for each cell type. Let Y_{jk}, be the value for feature j, $j \in \{1, 2, \ldots, 100\}$ for cell type k, $k \in \{1, 2, 3\}$. For the features that were randomly selected to exhibit a difference across cell types, we randomly chose one cell type to have mean 0, one cell type to have mean Δ, and one cell type to have mean −Δ, where Δ reflects the magnitude of the difference between the mean feature levels. Thus, these features were distributed as $Y_{i1} \sim Normal(0, 1)$, $Y_{i2} \sim Normal(\Delta, 1)$, $Y_{i3} \sim Normal(-\Delta, 1)$. All other features were distributed as $Y_{ik} \sim Normal(0, 1)$ across the three cell types. While DNA methylation data is not normally distributed, the design implemented here nevertheless emulates the behavior of cellspecific methylation and provides a useful framework for understanding the behavior of the Modified DSC as a function of the number/fraction of distinct features and the magnitude of their difference across cell types. After each dataset was simulated, the Modified DSC was calculated. Finally, a heatmap of simulated data and a plot of the first two principal components (PC1 vs. PC2) were generated to visualize the separation of the cell types for a library with no differences, a library with moderate cellspecific differences, and a library with large cell-specific differences.

Training the Algorithm

To examine the performance of RESET, we conducted an analysis of several publicly available data sets and benchmarked the accuracy of deconvolution estimates obtained using RESET against the Legacy approach. To do this, we first applied our approach using the Reference data set to identify "optimal" reference libraries for deconvoluting WB mixtures. As previously described, the reference data set is comprised of DNA methylation profiles for granulocytes, monocytes, CD4T, CD8T, natural killer, and B cells in n = 6 subjects. To get a candidate list of CpGs, we first applied the Legacy approach to the broader reference data set. We specified L to be 50, giving us the top 50 hyper and hypomethylated CpGs for each cell type for a total of 1800 CpGs (J 1800). After this, we ran RESET for 100,000 iterations. The library that yielded the largest value of the Modified DSC was taken to be the "optimal" reference library and used for subsequent CMD. A workflow for this analysis can be found in Figure 2B. We also applied the Legacy approach, as previously described, to the same reference data set. To compare the prediction performance across a range of library sizes, we assumed different values for J^p. That is, we fixed J^p 72, 120, 180, 240, 300, 360, 540 for both RESET and the Legacy method. Further, since the Legacy approach and IDOL require users to specify the size of the reference library, we also analyzed prediction performance for RESET when J^p was treated as random. Specifically, instead of fixing J^p to be one of the values above, we let J^p be a random number between 50 (j_{min}) and 1,500 (j_{max}). As before, the library that resulted in the largest value of Modified DSC was taken to be the "optimal" reference library. It should be noted that, currently, the only other method for building reference libraries, which does not require a training data set, is the ANOVA approach. As this method has been shown to perform poorly compared to the Legacy approach in Koestler et al. (2016), we only benchmark RESET against the Legacy approach (Koestler et al., 2016).

Validation of the Identified Reference Libraries To validate RESET, we applied CMD to three independent testing sets (MethodA, MethodB, AdultMixed) using the libraries identified using the RESET algorithm. These data sets were selected since the underlying leukocyte fractions are known for each sample. Figure 4A illustrates the differing celltype proportions across the three testing data sets.

To assess the performance of our method and the Legacy approach, we first estimated the proportion of variation of the known mixture fractions explained by the cell type predictions (R^2). We also assessed our performance by calculating the root mean squared error (RMSE), which reflects the average deviation between the true mixture proportions and our predictions. These metrics were computed for each cell type individually as well as their averages across all cell types for each reference library we obtained.

Simulation Study Comparing the FDR

To explore the implications of prediction error in cell proportion estimates for EWAS, we conducted a simulation study in which the false discovery rate (FDR) was calculated for cell composition adjusted EWAS, where cell composition estimates were based on differing methods/approaches for library selection. We implemented a study design that is common to many EWAS, in which the goal is to identify differently methylated CpGs between two groups of samples (n $n_1 + n_2$). We conducted this simulation considering a moderately sized study (n_1, n_2 100) and assuming a large study (n_1, n_2 500). Further, we introduced varying dissimilarities in the true, simulated cellular distribution between the cases and controls.

For this simulation study, we assumed that the methylation beta-value for CpG j for sample i, that is Y_{ij}, follows a beta-

distribution with mean $w_i\mu_{Tj}$ and variance $(1-w_i\mu_{Tj}+\mu_{Tj}\varphi_j)w_i\mu_{Tj}$. Recall that w_i is a vector that represents the true cell proportions for sample i, μ_j is a vector of mean methylation beta-values for CpG j across the K cell types, and φ_j is the unobserved precision parameter for CpG j where $\varphi_j > 0$. Often, EWAS fit regression models as follows:

$$Y_{ij} = \alpha_{0j} + \alpha_{1j}X_i + \sum_{k=1}^{K-1} \gamma_{kj}w_{ik} + \epsilon_{ij},$$

$$E(\epsilon_{ij}) = 0 \text{ and } Var(\epsilon_{ij}) = \sigma^2_j$$

Here, $\kappa_k = \gamma_{1kj}w_{ik}$ controls for cell composition across subjects, $X_i \in (0, 1)$ denotes the group subject i belongs to, and ϵ_{ij} captures any remaining variation in methylation after controlling for group and cell proportions. Usually, interest lies in tests of no difference in DNA methylation between groups. That is, testing the null hypothesis that $\alpha_{1j} = 0$. In practice w_{ik} is unknown and is estimated by $\hat{w}_{ik}$, which is obtained through CMD (Houseman et al., 2012). However, $\hat{w}_{ik}$ is an estimate, so any hypothesis tests based on the above regression model can be susceptible to Type 1 and Type 2 errors (Koestler et al., 2016).

To check how cell proportion prediction errors estimated using the RESET and Legacy libraries affect the FDR of hypothesis tests (H_0: $\alpha_{1j} = 0$), we first estimated the uncertainty of cell proportion predictions for each method. To do this, we squared the RMSEs across the three testing data sets, across the different library sizes, to get mean squared prediction errors (MSPEs). We did this for both the Legacy approach and RESET. We will denote the MSPEs as $\hat{\tau}^2_{kl}$, where l represents the type of library used (either Legacy or RESET), and k is the index for cell type. To estimate the mean and precision parameters, we fit Beta regression models to each CpG. We then conducted our simulation study by randomly sampling 1,000 CpGs, generating random cell distributions for the two groups, simulating beta-values for each group, randomly sampling cell type predictions for each sample, fitting the above regression model to each of the CpGs, and calculating the FDR for each method assuming a nominal p-value cut off of 0.05. More specific details about the algorithm for this simulation can be found in Supplementary Material S1. Recall that the methylation beta-values for the two groups were simulated assuming no difference in the methylation profile (e.g., under the null hypothesis $\alpha_{1g} = 0$). Thus, the only difference between the two groups lies in the dissimilarity in cell composition. So, any rejections of the hypothesis of $\alpha_{1g} = 0$ reflect a Type 1 error.

Data Application and Implications for EWAS We finally used the Liu and Hannum data sets described above to understand the implications of the Legacy approach and the RESET algorithm on cell type prediction for EWAS (Hannum et al., 2013; Liu et al., 2013). Specifically, an analysis of these blood-derived DNA methylation data sets was done to see which method performed better in explaining variation in DNA methylation. To do this, we started by using CMD for estimating cell proportions in the two data sets using the reference libraries identified using the RESET method, as well as the libraries identified using the Legacy approach. For each data set, linear regression models were fit to the J total CpGs independently, modeling beta-values as the response against the predicted cell proportions. Using the fitted models, we found the proportion of variation in methylation explained by the cell mixture proportion estimates using both methodologies $(R^2_j,\ j \in \{1, 2, \ldots, J\})$.

The difference in R^2 between models, adjusted for cell composition estimates using RESET and the Legacy approach were then calculated for each of the J CpGs. That is, $\Delta_j = R^2_{j,RESET} - R^2_{j,Legacy}$. Finally, to answer our question, we computed the proportion of CpGs where RESET explained more variation in the methylation explained as compared to the Legacy approach $(\frac{1}{J}\sum_{j=1}^{J} I(\Delta_j > 0))$ for each library size.

RESULTS Proof-Of-Principle Example to Illustrate the Modified DSC Metric

To rationalize the use of the Modified DSC as a metric for identifying sets of CpGs which discriminate well between cell types, we simulated cell-specific methylation data with varying degrees of differences and calculated the Modified DSC. Recall that we allowed the percentage of features that exhibit a difference across the cell types to vary as well as allowing the magnitude of the difference in mean methylation to vary across cell types. Each of these differences gave us a new cell-specific methylation data set. We hypothesized that higher values of the Modified DSC result in reference libraries that better discriminate the cell types. Figure 3A shows the results for the 210 different simulated data sets. As the number of features/CpGs that exhibit a difference between cell types increases, and the magnitude of this difference increases, the value of the Modified DSC gets larger. The minimum Modified DSC value of 0.580 occurred when 0% of the 100 features were simulated to exhibit a difference across the three cell types. The maximum DSC value of 15.252 occurred when 90% of the 100 features were simulated to exhibit a difference across the cell types with a magnitude of 10, the largest possible magnitude difference considered in our simulation study. To further demonstrate the utility of the Modified DSC for building reference libraries for cell mixture deconvolution, we looked at heat maps of the simulated data and plots of the first two principal components (PCs) calculated from a principal components analysis (PCA) of the simulated data at various

degrees of differences. Figure 3B depicts the data that were simulated to exhibit no difference in feature/CpG values across the three cell types. Not surprisingly, the heat map shows no discernible difference between cell signatures. Further, no discernable clustering samples by cell type are evident in the plot of PC1 by PC2. The value of the Modified DSC for this data set is 0.580. Figure 3C depicts the data set where 50% of the features/ CpGs were simulated to exhibit a magnitude difference of 5 between the three cell types. The heat map shows some differences in the signatures across cell types, and the plot of PC1 by PC2 shows a clear separation of samples by cell type. The value of the Modified DSC for this data is 5.443. Finally, Figure 3D depicts the data that were simulated to exhibit a difference in 100% of the features/CpGs, with that difference being 10. The heatmap shows clear cell-type-specific signatures, and the plot of PC1 by PC2 shows even more separation in the clusters of cell types (take note of the scale of the x- and y-axis in the PC-plots between Figures 3C,D).

Validation of the Optimal Reference Libraries

As mentioned previously, to validate the reference libraries identified by RESET, we examined CMD prediction performance by calculating R^2 and RMSE using three independent test sets: Method A, Method B, and Adult Mixed, and benchmarked the performance of RESET

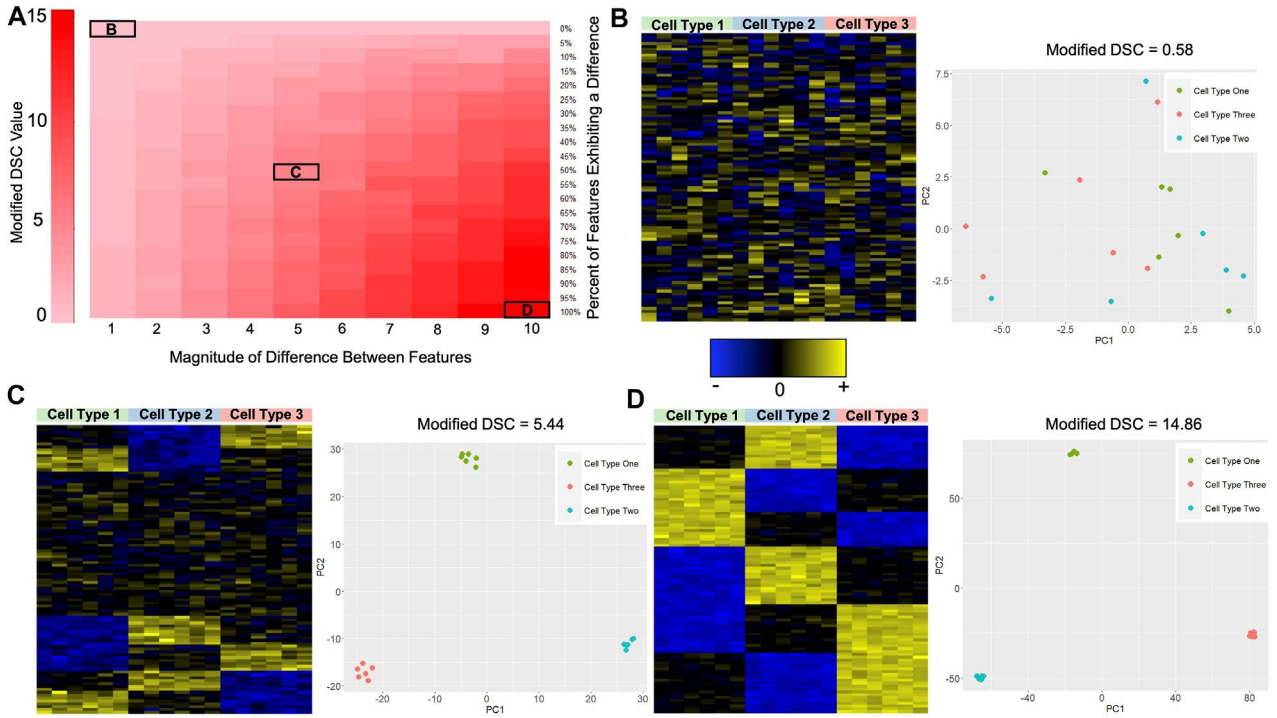


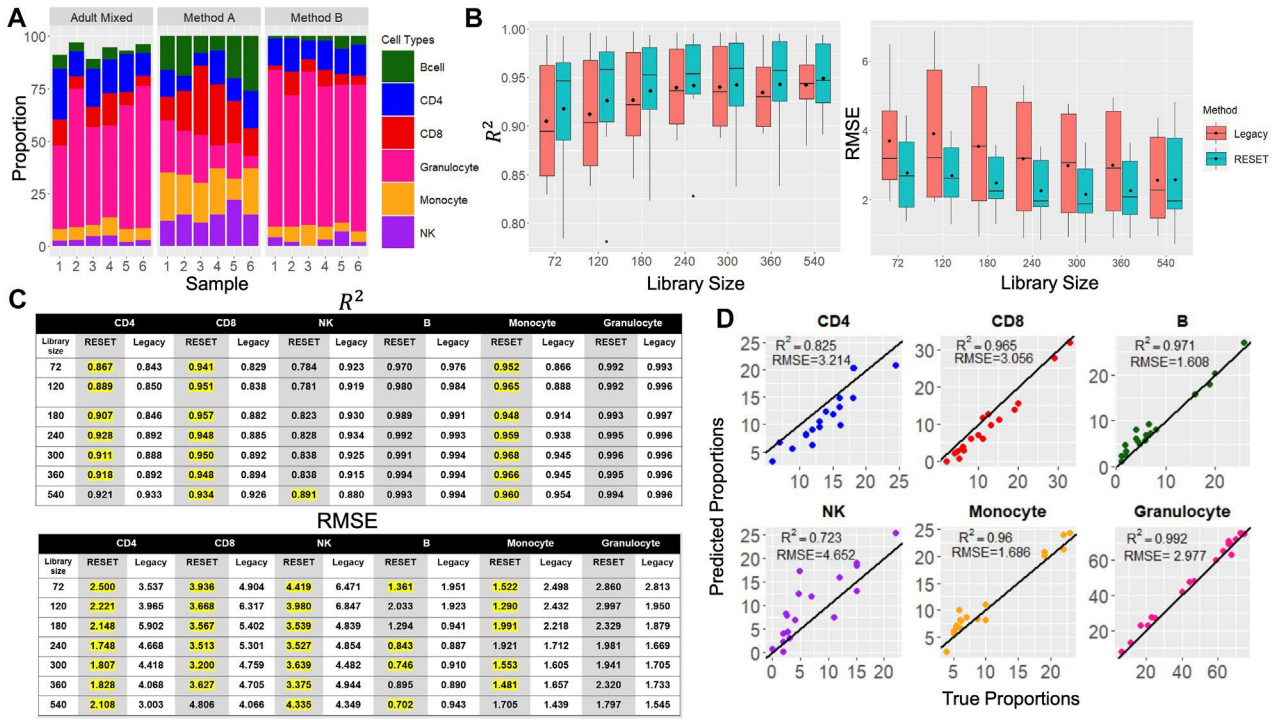
FIGURE 3 | Proof-of-Principle Illustration of the Modified DSC Metric. (A) Heat map showing the value of the Modified DSC under the different data scenarios as described in section 2.5.1. The y-axis shows the percent of features that were simulated to exhibit a difference in the mean feature(s) value across cell types, and the x-axis reflects the magnitude of the difference between the mean feature(s) values across cell types. Darker red indicates higher values of the Modified DSC. (B) Heat map and PCA plot for the simulated data where no features exhibited a difference in the mean feature(s) values across cell types. The value of the modified DSC for this simulated data is also provided. (C) Heat map and PCA plot for the simulated data where 50% of the features exhibited a difference in mean feature(s) values across cell types with a magnitude of 5 (i.e., Δ 5). The value of the modified DSC for this simulated data is also provided. (D) Heat map and PCA plot for the simulated data where 100% of the features exhibited a difference in the mean feature(s) values across cell types with a magnitude of 10 (i.e. Δ 10). The value of the modified DSC for this simulated data is also provided.

against the Legacy approach for building reference libraries. As noted in Figure 4A, the different testing data sets are made up of varying proportions of each cell type. The six samples in the Method A testing set have approximately equal proportions of each cell type with the mean proportion as follows: CD4T = 11.8%, CD8T = 20.8%, NK = 15%, B = 16%, Monocyte = 19.2%, Granulocyte = 17.2% (Wiencke et al., 2017). The Method B samples more closely resemble the immune cell landscape in human adults with mean proportions as follows: CD4T = 13.2%, CD8T = 6%, NK = 3%, B = 2.7%, Monocyte = 6.2%, Granulocyte = 69% (Wiencke et al., 2017). Like the Method B samples, the Adult mixed samples had mean proportions as follows: CD4T = 17.9%, CD8T = 9.7%, NK = 3.5%, B = 4.9%, Monocyte = 6.7%, Granulocyte = 57.4% (Wiencke et al., 2017).

Prediction performance for RESET and the Legacy method was evaluated for several different library sizes (J^{o} 72, 120, 180, 240, 300, 360, 540) using the three mentioned testing data sets. RESET resulted in categorically better prediction performance over the Legacy approach. The left plot of Figure 4B shows the R^2 across the six cell types for each library size. On average RESET only did marginally better than the Legacy method. However, the median R^2 for RESET was higher than Legacy for all library sizes, except for 540.

Improvements in prediction performance were even more apparent when looking at the RMSE. The right plot of Figure 4B shows the RMSE across the six cell types for each library size. On average, the RMSE was smaller when using our approach for all library sizes except for 540. Further, median RMSE was smaller for all library sizes when using RESET.

Cell-type-specific R^2 and RMSE values for both RESET and the Legacy method can be found in Figure 4C. Interestingly, for CD4T, CD8T, and monocytes the R^2 was higher than what was observed in the Legacy approach across nearly all reference library sizes considered. The R^2 for granulocytes and B cells was very close to 1 for both our method and the Legacy approach. However, the R^2 was higher for the Legacy approach in NK cells. Additionally, the RMSE was lower for nearly all reference library sizes using RESET for CD4T, CD8T, NK, B, and monocytes. The only cell type that did not have lower RMSE when comparing the libraries constructed using the



R^2

Library size	CD4 RESET	CD4 Legacy	CD8 RESET	CD8 Legacy	NK RESET	NK Legacy	B RESET	B Legacy	Monocyte RESET	Monocyte Legacy	Granulocyte RESET	Granulocyte Legacy
72	0.867	0.843	0.941	0.829	0.784	0.923	0.970	0.976	0.952	0.866	0.992	0.993
120	0.889	0.850	0.951	0.838	0.781	0.919	0.980	0.984	0.965	0.888	0.992	0.996
180	0.907	0.846	0.957	0.882	0.823	0.930	0.989	0.991	0.948	0.914	0.993	0.997
240	0.928	0.892	0.948	0.885	0.828	0.934	0.992	0.993	0.959	0.938	0.995	0.996
300	0.911	0.888	0.950	0.892	0.838	0.925	0.991	0.994	0.968	0.945	0.996	0.996
360	0.918	0.892	0.948	0.894	0.838	0.915	0.994	0.994	0.966	0.945	0.995	0.996
540	0.921	0.933	0.934	0.926	0.891	0.880	0.993	0.994	0.960	0.954	0.994	0.996

RMSE

Library size	CD4 RESET	CD4 Legacy	CD8 RESET	CD8 Legacy	NK RESET	NK Legacy	B RESET	B Legacy	Monocyte RESET	Monocyte Legacy	Granulocyte RESET	Granulocyte Legacy
72	2.500	3.537	3.936	4.904	4.419	6.471	1.361	1.951	1.522	2.498	2.860	2.813
120	2.221	3.965	3.668	6.317	3.980	6.847	2.033	1.923	1.290	2.432	2.997	1.950
180	2.148	5.902	3.567	5.402	3.539	4.839	1.294	0.941	1.991	2.218	2.329	1.879
240	1.748	4.668	3.513	5.301	3.527	4.854	0.843	0.887	1.921	1.712	1.981	1.669
300	1.807	4.418	3.200	4.759	3.639	4.482	0.746	0.910	1.553	1.605	1.941	1.705
360	1.828	4.068	3.627	4.705	3.375	4.944	0.895	0.890	1.481	1.657	2.320	1.733
540	2.108	3.003	4.806	4.066	4.335	4.349	0.702	0.943	1.705	1.439	1.797	1.545

FIGURE 4 | Results from RESET and Legacy Approach. (A) True cell-type proportions for the three independent data sets used to generate predictions for the RESET and Legacy approaches. Each data set is comprised of six samples and the same six leucocyte subtypes. (B) Box plots showing the R^2 and RMSE across all cell types for each assumed reference library size (recall J^p { 72, 120, 180, 240, 300, 360, 540}) for RESET compared to the Legacy approach. (C) Table of results comparing the prediction performance of the RESET algorithm and the Legacy approach for each cell type and across different assumed reference library sizes (rows of the table). The top table shows the R^2 while the bottom table shows the RMSE, calculated by comparing the predicted/estimated versus observed cell proportions. Yellow highlighted text denotes the conditions where the use of RESET resulted in better prediction performance than the Legacy approach. (D) Cell-specific scatter plots of the observed (x-axes) versus estimated (y-axes) cell proportions using the reference library yielded the largest value of the Modified DSC when the number of DMLs that comprise the library was allowed to be random. Additionally, the R^2 and RMSE for these predictions are provided.

Modified DSC approach to the Legacy method was granulocytes.

In addition to comparing the R^2 and RMSE for the fixed libraries, we also calculated the Modified DSC value across the RESET land Legacy libraries. The average difference in the Modified DSC across the fixed libraries for the two methods was 14.82, indicating that on average, the Modified DSC of libraries obtained via RESET is higher than the Legacy method.

When the reference library size was chosen randomly, RESET identified a library with 54 CpGs. That is, the library size that resulted in the largest value of the Modified DSC was comprised of 54 loci. The prediction performance was evaluated using the same three previous testing data sets. Even with so few CpGs, this library still had an average R^2 of 0.906 across all six cell types and an average RMSE of 2.865. CD8T, B, monocyte, and granulocyte predictions all had an R^2 higher than 0.950. The predictions for NK and CD4T cells had an R^2 of 0.723 and 0.825, respectively. Figure 4D depicts the predicted cell proportions versus the true cell proportions for each of the six cell types based on this reference library.

Simulation Study Comparing FDR

The validation results showed that the libraries identified by RESET were overall only marginally better than those identified by the Legacy approach; however, prediction performance is often not an end in itself but rather a means to an end. Specifically, the cell proportion estimates identified from these libraries are often used for cell-type adjustment in downstream statistical analysis. As such, we performed a simulation to explore the consequences of prediction error in cell proportion estimates for EWAS (Figure 5A, top two plots). We conducted a simulation study where methylation beta values were simulated for two groups, 1,000 CpGs, and assuming both moderate and large within-group sample size (n = 100, 500, respectively). Additionally, the cellular composition varied across groups while the cell types were simulated to have no difference between groups. When adjustments for cell composition were made using the true simulated cell proportions, the FDR was controlled at 5% as expected. When no adjustments for cell composition were made, there was large inflation in the FDR, which increased as a function of increasing cell composition dissimilarity between groups. When n1 = n2 = 100, the FDR was controlled at 5%

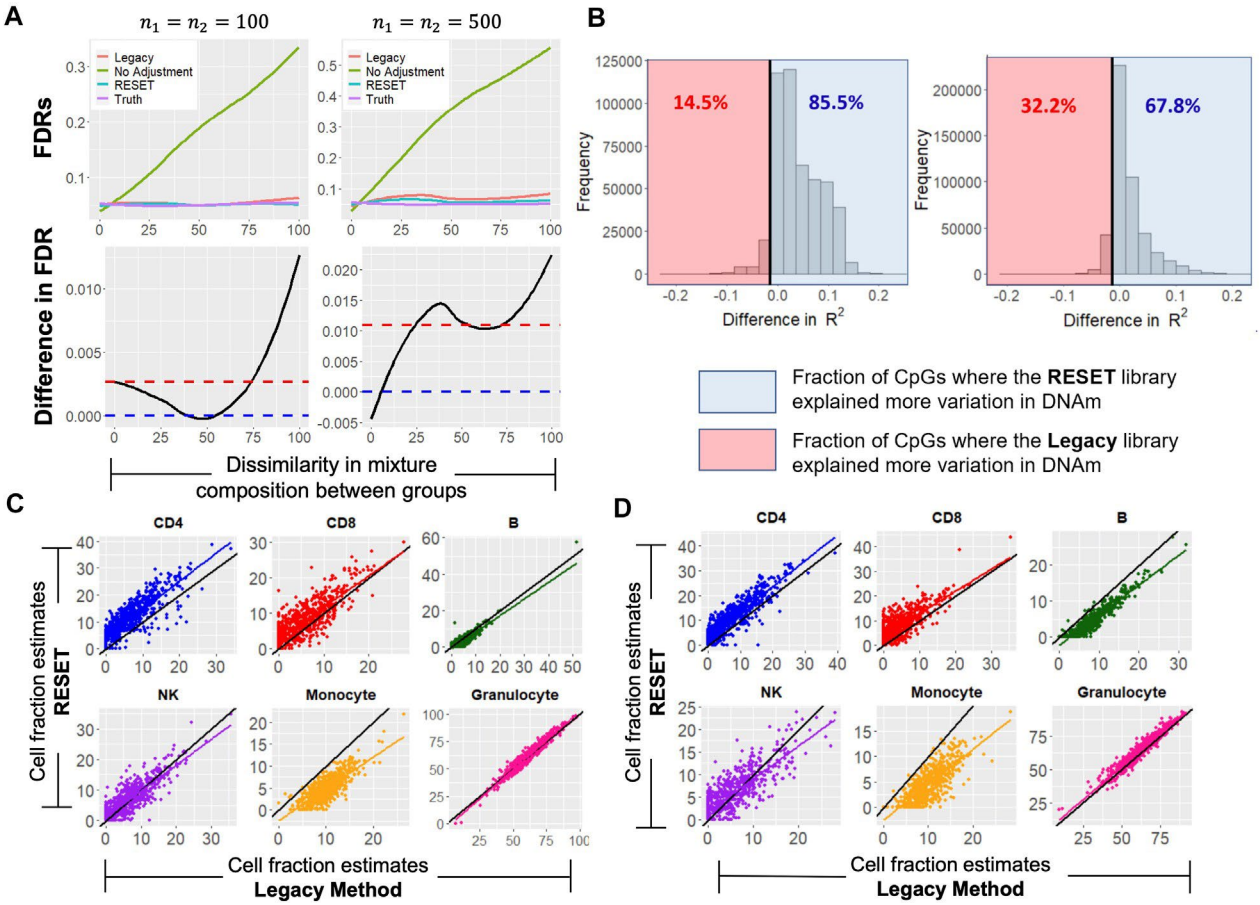


FIGURE 5 | FDR Simulation Results and Data Application Results for the DML Reference Library of Size 120.(A) Results from the FDR simulation to explore implications of prediction error in cell proportion estimates for EWAS as described in section 2.5.4. The top panel shows the estimates of the FDR for the two-group comparison of methylation as a function of the dissimilarity in the distribution of cell-type composition between the two groups. The FDRs were estimated under four methods/approaches for cell composition adjustment. The bottom panel shows the difference in FDR estimates when cell type estimation was undertaken using the Legacy library versus the RESET library. Points that lie above the blue dotted line signify that the Legacy library resulted in more false-positive results as compared to the RESET. The red dotted line shows the mean difference in FDR between the Legacy library and the RESET library. A loess smoother was used to generate these curves. (B) Distribution of the difference in R^2 obtained from the RESET and Legacy libraries of size 120 CpGs applied to the Liu (left panel) and Hannum (right panel) data sets described in section 2.5.5. (C) Scatter plots of the predicted cell-type proportions using the Legacy library of 120 CpGs (x-axis) and the RESET library of 120 CpGs (y-axis) as applied to the Liu data set.(D) Scatter plots of the predicted cell-type proportions using the Legacy library of 120 CpGs (x-axis) and the RESET library of120 CpGs (y-axis) as applied to the Hannum data set.

when adjustments for cell composition were made using both RESET and the Legacy approach. However, when n1 = n2 = 500, there was slight inflation in the FDR using the Legacy approach while it remained controlled at 5% when using RESET. Additionally, the bottom two plots of Figure 5A show the difference in FDR from the Legacy approach and our approach. RESET, on average, resulted in fewer false positives compared to the Legacy approach across the different simulation conditions.

Data Application and Implications for EWAS

In addition to using the above simulation study, we also made use of two publicly available data sets to better understand the consequences of the approach used for reference library construction in the context of EWAS. All fixed reference libraries that were previously identified were applied to the Lui and Hannum data sets (Hannum et al., 2013; Liu et al., 2013) to see which libraries led to a better explanation in the variation in DNA methylation (R^2). When using the Liu data set, the proportion of CpGs where RESET explained more variation in the methylation, as compared to the Legacy approach, was high for most of the library sizes. These proportions are 0.654, 0.855, 0.892, 0.871, 0.870, 0.882, and 0.774 for the library sizes of 72, 120.180.240, 300, 360, and 540, respectively. When using the Hannum data set, the proportion of CpGs where RESET explained more variation in the methylation, as compared to the Legacy approach, while not as high, was still greater than 0.5 for all libraries. These proportions are 0.537, 0.678, 0.728, 0.678, 0.680, 0.677, and 0.590 for the library sizes of 72, 120.180.240, 300, 360, and 540, respectively. A visualization of the results when using the reference libraries of size 120 is shown in Figure 5B. Scatter plots of the cell

proportion estimates of the RESET library of size 120 compared to the estimates of the Legacy library of size 120 can also be found in Figure 5C (using the Liu data set) and Figure 5D (Using the Hannum data set). RESET tended to result in higher predicted proportions for CD4T and CD8T cells for both data sets while tending toward lower predicted proportions for B cells and monocytes. Similar visualizations for results using the other libraries can be found in Supplementary Material S2.

DISCUSSION

This manuscript described and evaluated an iterative algorithm for assembling libraries for reference-based deconvolution in the absence of a training dataset. The motivation for this research is three-fold. First, the performance of reference-based deconvolution depends heavily on the reference library used as the basis of deconvolution (Koestler et al., 2016; Teschendorff et al., 2017; Titus et al., 2017). Second, in the absence of a training data set, current approaches require users to make arbitrary decisions about the number of DMLs and the library size even though performance is sensitive to these decisions (Houseman et al., 2012; Koestler et al., 2016). Third, emergence and interest in new cell types means that not only will reference libraries need to be updated but training data sets also. If DNAm along with measured cell-type proportions for these new cell types are not available, then IDOL is not possible (Koestler et al., 2016).

Recall that RESET builds DNAm reference libraries by utilizing the Modified DSC. This metric increases in value as clusters of cell-type-specific DNAm data get farther apart from one another and makes it a useful metric for evaluating the quality of a reference library. RESET works by randomly selecting a reference library from a candidate list of CpGs, calculating the Modified DSC, and then iteratively updating the probability of selecting a CpG in future iterations depending on its contribution to the modified DSC. The library which yields the largest value of the Modified DSC after many iterations is the reference library used for deconvolution. As previously described, while the Legacy approach has been previously shown to outperform the ANOVAbased method for library construction (Koestler et al., 2016), its primary shortcoming is that the user needs to prespecify both the total library size and the number of cell-specific DMLs to include in the library. As shown here (Figure 3) and previously (Koestler et al., 2016), the accuracy of corresponding deconvolution estimates is sensitive to these selections, with the R^2 of some cell types varying by nearly 0.10 units across the limited range of library sizes considered here. This observation is not surprising as leukocytes share common linages, and thus varying numbers of DMLs may be needed for different leukocyte subtypes (Koestler et al., 2016). While the IDOL algorithm is an improvement on both the ANOVA-based and Legacy approaches, it requires a training data set consisting of both DNAm data—profiled in whole blood or some other heterogeneous biospecimen—and gold standard measurements of cellular composition in the same biospecimen used for methylation profiling (Koestler et al., 2016). Although such data presently exist for whole blood (Reinius et al., 2012; Koestler et al., 2016), they typically only include relative fractions of the major leukocyte components of whole blood (e.g., CD4T, CD8T, B cells, NK cells, monocytes, and granulocytes) and lack measurements of more exotic cell types (e.g., monocytic and/or granulocytic myeloid-derived suppressor cells) and cells with specific activation states (e.g., CD8T naïve, central memory, and effector memory). In contrast to the ANOVA-based, Legacy, and IDOL approaches, RESET does not require a training data set, the number of DMLs for each cell type need not be the same, and the total number of loci that make up the reference library does not need to be specifically selected in advance.

While our results showed a modest improvement in the average R^2 and RMSE resulting from the libraries constructed using RESET compared to the Legacy approach, they showed considerable improvement in the accuracy for specific cell types (CD4T, CD8T, and monocytes). For these three cell types, the libraries identified using RESET had a larger R^2 than those identified by the Legacy approach in all but one library (Figure 4C). Additionally, the RMSE for CD4T, CD8T, and monocytes were smaller using RESET libraries in all but two cases (Figure 4C). Our algorithm seemed to favor choosing small libraries for the library identified using a random number of loci. That is, the random-sized reference library with the largest value of the Modified DSC was comprised of only 54 CpGs. The R^2 averaged across all six cell types for this library was 0.906. This was smaller than the average R^2 for most of the Legacy libraries, but within 0.04 units. Interestingly, the CD8T and monocytes cell-specific R^2 were higher than any of the cell-specific R^2 for the Legacy libraries. Further, the average RMSE for this random library was 2.865, which was better than all but one of the Legacy libraries. Also interesting is the performance of this library when compared to the performance of the library identified by IDOL. The IDOL library was comprised of 300 loci, had an average R^2 of 0.993 across the six cell types, and an average RMSE of 1.15 (Koestler et al., 2016). While, on average, IDOL outperforms our algorithm, we note that IDOL should be expected to be slightly better because it utilizes a training data set. The slightly lower prediction performance of RESET is made up by the fact that our algorithm does not need a training data set. In addition to the favorable performance of the randomly sized library, there are also downstream benefits to having a small reference library for deconvolution. First is the fact that a small library comprised of high-quality DMLs may better resolve the cellular landscape and avoid the inevitable noise that would result from a larger library comprised of lower quality DMLs; second, the design of custom methylation arrays with potential clinical use will be simplified using a small number of precise markers. Finally, having a smaller reference library would be more technically feasible and less costly for researchers, given that fewer CpGs would need to be profiled. Our simulation study results showed that even small improvements in prediction performance for RESET resulted in lower FDR compared to the Legacy approach for all differences in the underlying cellular composition when the sample size of each group was 500 (Figure 5A). Since EWAS typically entails statistical testing of hundreds of thousands to millions of CpGs, the potential consequence is that thousands of CpGs are incorrectly classified as differentially methylated when the Legacy versus RESET is used to construct reference libraries for deconvolution. Finally, in both data applications, cell proportion estimates obtained by using RESET resulted in an improved ability to explain variation in whole-blood-derived DNA methylation signatures.

The consequence of this is improved statistical power for EWAS adjusted for cell-type composition when cell proportion estimates are obtained using the RESET versus the Legacy approach.

While our method resulted in improved accuracy of referencebased deconvolution estimates, lower FDR, and more variability in CpG-specific methylation explained in the context of EWAS adjusted for cell composition, RESET, and this study are not without limitations. The data sets used to build the reference libraries and benchmark the performance of our approach against the Legacy method only consisted of relative fractions of the six major leukocyte components of whole blood (e.g., CD4T, CD8T, B cells, NK cells, monocytes, and granulocytes), despite there being a myriad of other cell types and specific states of cells present in whole blood. In DNA methylation data, the main source of variability between isolated cell types is captured in the first two dimensions of the data. The Legacy approach exploits this fact using a fixed number of differentially methylated sites (usually 100 per cell-type) to capture that variability. However, as the cell lineages diverge, cell specialization drifts the epigenetic marks in a smaller proportion of CpGs. This translates into cell states/subtypes that are challenging to discern using surface/nuclear markers or even functional essays. As such, these states/subtypes fall closer to each other, rendering the assumption of a fixed number of sites untenable. Our algorithm makes no such assumption, and we hypothesize that our method will further outperform the Legacy approach in these situations. Additionally, while five independent data sets were used to evaluate and benchmark the performance of RESET, we note that data sets containing both DNA methylation data and gold-standard measurements of cell composition (e.g., flow cytometry) in the same set of samples are relatively rare, limiting the number of viable data sets to use in our comparisons. Further, only seven different fixed reference library sizes were considered in our evaluation (J^p 72, 120, 180, 240, 300, 360, 540). While it is unclear what the performance of our approach would look like for larger library sizes, the results presented by Koestler et al. suggest diminishing returns on performance for libraries with more than 500 CpGs, motivating our decision to consider libraries with 540 or fewer CpGs (Koestler et al., 2016). A potential limitation of the Modified DSC is that this metric seems to favor smaller library sizes, specifically reference libraries with less than 100 loci. This could be due to the phenomenon known as the curse of high dimensionality, which is an occurrence when calculating Euclidean distances in high dimensional settings (Mirkes et al., 2020; Sarkar and Ghosh, 2020). In high dimensions, data can become noisy, so clusters of data can become more saturated, making the space between them (distances) harder to uncover. One possible way to overcome this would be to use the same algorithm but use a different distance metric (e.g., Manhattan distance) (Mirkes et al., 2020). It is also worth mentioning that the R^2 for NK cells using RESET was lower across all libraries compared to the Legacy approach; however, the lowest performance still had an R^2 of 0.78. The reasons for this lowered performance could be explained as the original Reinius dataset used for training the data had a known technical problem derived from potential cross-contamination of CD8T cells into their NK reference (Salas et al., 2018). As such, the reference cannot fully resolve between NK, CD4T, and CD8T cells and the difficulty that arises in distinguishing between these cell types. An additional limitation of RESET is its computational burden compared to other methods. This issue could be resolved by utilizing a lowerlevel programming language or the use of more parallelization of computations. Moreover, our algorithm also was only performed on whole-blood DNA methylation data; however, in principle, it can easily be extended to other solid tissue types where goldstandard measurements of the cell proportions are not as feasible to obtain. Finally, while this research utilized data from the Illumina HumanMethylation450 array, we note that this method is generalizable to other platforms (whole-genome bisulfite sequencing, Illumina HumanMethylationEPIC BeadArray, etc.).

This research was motivated by the need for accounting for cellular landscape when DNA methylation is analyzed in heterogeneous tissue, and further, by the need to remove requirements and arbitrary decisions other methods need in order to build reference libraries for the purpose of referencebased cell mixture deconvolution. Future work includes improving the accuracy of cell proportion estimates in the absence of a training data set by considering different distance metrics, building reference libraries for tissue types, and extending this algorithm to explore performance in samples with more than six cell types. The utilization of RESET will allow researchers to save money and remove arbitrary decisions in the reference library selection process while maintaining accuracy. RESET introduces a new way of thinking for the future of reference library selection in the absence of a training data set and is a useful approach for library construction for EWAS.

DATA AVAILABILITY STATEMENT

The R scripts for the analyses presented here are available on GitHub at https://github.com/ShelbyBellGlenn/RESET_Code. The data sets used are all publicly available through GEO.

AUTHOR CONTRIBUTIONS

SB-G. conceived the statistical framework for selecting reference libraries for reference-based cell mixture deconvolution, performed the statistical analyses to evaluate and assess the proposed framework, and participated in the drafting and review of the manuscript. JT and LS helped direct the statistical analyses, contributed to the interpretation of study findings, and assisted in manuscript writing and development. DK helped to conceive the statistical framework the statistical framework for selecting reference libraries for reference-based cell mixture deconvolution, provided R code for some of the statistical analyses, provided guidance and direction in the

evaluation and assessment of the method, and assisted in manuscript writing and development.

FUNDING

Research reported was supported by: the National Cancer Institute (NCI) Cancer Center Support Grant P30 CA168524; the Kansas IDeA Network of Biomedical Research Excellence Bioinformatics Core, supported by the National Institute of General Medical Science award P20 GM103418; the Kansas Institute for Precision Medicine COBRE, supported by the National Institute of General Medical Science award P20 GM130423. LS. was supported by CDMRP/Department of Defense (W81XWH-20-1-0778) and NIGMS (P20GM104416-09/8299).

REFERENCES

Akbani, R., Bradley, N. Z., Broom, M., John, N., Weinstein Tod, D., Casasent James, M., et al. (2020). Allen T. Chang. TCGA Batch Effects Viewer. Available at: https://bioinformatics.mdanderson.org/public-software/tcga-batch-effects/.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: a Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays. Bioinformatics 30 (10), 1363–1369. doi:10.1093/bioinformatics/btu049

Decamps, C., Privé, F., Bacher, R., Jost, D., Waguet, A., et al. (2020). Guidelines for Cell-type Heterogeneity Quantification Based on a Comparative Analysis of Reference-free DNA Methylation Deconvolution Software. BMC Bioinformatics 21 (1), 16. doi:10.1186/s12859-019-3307-2

Dy, J. G. C. E. B. (2004). Feature Selection for Unsupervised Learning. J. machine Learn. Res. 5, 845–899.

Flanagan, J. M. (2015). Epigenome-wide Association Studies (EWAS): Past, Present, and Future. Methods Mol. Biol. 1238, 51–63. doi:10.1007/978-14939-1804-1_3

Grieshober, L., Graw, S., Barnett, M. J., Goodman, G. E., Chen, C., Koestler, D. C., et al. (2021). Pre-diagnosis Neutrophil-To-Lymphocyte Ratio and Mortality in Individuals Who Develop Lung Cancer. Cancer Causes Control 32 (11), 1227–1236. doi:10.1007/s10552-021-01469-3

Grieshober, L., Graw, S., Barnett, M. J., Thornquist, M. D., Goodman, G. E., Chen, C., et al. (2018). Methylation-derived Neutrophil-To-Lymphocyte Ratio and Lung Cancer Risk in Heavy Smokers. Cancer Prev. Res. (Phila) 11 (11), 727–734. doi:10.1158/1940-6207.CAPR-18-0111

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., et al. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Mol. Cel 49 (2), 359–367. doi:10.1016/j.molcel.2012.10.016

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., et al. (2012). DNA Methylation Arrays as Surrogate Measures of Cell Mixture Distribution. BMC Bioinformatics 13, 86. doi:10.1186/14712105-13-86

Houseman, E. A., Kile, M. L., Christiani, D. C., Ince, T. A., Kelsey, K. T., and Marsit, C. J. (2016). Reference-free Deconvolution of DNA Methylation Data and Mediation by Cell Composition Effects. BMC Bioinformatics 17, 259. doi:10. 1186/s12859-016-1140-4

Houseman, E. A., Kim, S., Kelsey, K. T., and Wiencke, J. K. (2015). DNA Methylation in Whole Blood: Uses and Challenges. Curr. Environ. Health Rep. 2 (2), 145–154. doi:10.1007/s40572-015-0050-3

Jaffe, A. E., and Irizarry, R. A. (2014). Accounting for Cellular Heterogeneity Is Critical in Epigenome-wide Association Studies. Genome Biol. 15 (2), R31. p. R31. doi:10.1186/gb-2014-15-2-r31

ACKNOWLEDGMENTS

In addition, we would like to extend our gratitude to members of the Statistical Omics Working Group in the Department of Biostatistics and Data Science at the University of Kansas Medical Center for their constructive feedback on this manuscript. In particular, a special thanks to: Alexander Alsup, Lisa Neums, Shachi Patel, Qing Xia, Whitney Shae, Bo Zhang, Samuel Boyd, Emily Nissen, Jonah Amponsah, Dr. Prabhakar Chalise, Dr. Jinxiang Hu, Dr. Lynn Chollet-Hinton, Dr. Nanda Yellapu, Dr. Dong Pei, and Dr. Mihaela Sardiu.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at:

https://www.frontiersin.org/articles/10.3389/fbinf.2022.835591/full#supplementary-material

Koestler, D. C., Jones, M. J., Usset, J., Christensen, B. C., Butler, R. A., Kobor, M. S., et al. (2016). Improving Cell Mixture Deconvolution by Identifying Optimal DNA Methylation Libraries (IDOL). BMC Bioinformatics 17, 120. doi:10.1186/ s12859-016-0943-7

Koestler, D. C., Usset, J., Christensen, B. C., Marsit, C. J., Karagas, M. R., Kelsey, K. T., et al. (2017). DNA Methylation-Derived Neutrophil-To-Lymphocyte Ratio: An Epigenetic Tool to Explore Cancer Inflammation and Outcomes. Cancer Epidemiol. Biomarkers Prev. 26 (3), 328–338. doi:10.1158/1055-9965.EPI-160461

Liang, L., and Cookson, W. O. (2014). Grasping Nettles: Cellular Heterogeneity and Other Confounders in Epigenome-wide Association Studies. Hum. Mol. Genet. 23, R83–R88. (R1): p. R83-8. doi:10.1093/hmg/ddu284

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., et al. (2013). Epigenome-wide Association Data Implicate DNA Methylation as an Intermediary of Genetic Risk in Rheumatoid Arthritis. Nat. Biotechnol. 31 (2), 142–147. doi:10.1038/nbt.2487

Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Greally, J. M., Gut, I., et al. (2013). Recommendations for the Design and Analysis of Epigenomewide Association Studies. Nat. Methods 10 (10), 949–955. doi:10.1038/nmeth. 2632

Mirkes, E. M., Allohibi, J., and Gorban, A. (2020). Fractional Norms and Quasinorms Do Not Help to Overcome the Curse of Dimensionality. Entropy (Basel) 22. Basel22(10). doi:10.3390/e22101105

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. Nat. Methods 12 (5), 453–457. doi:10.1038/nmeth.3337

Rakyan, V. K., Down, T. A., Balding, D. J., and Beck, S. (2011). Epigenome-wide Association Studies for Common Human Diseases. Nat. Rev. Genet. 12 (8), 529–541. doi:10.1038/nrg3000

Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S. E., Greco, D., et al. (2012). Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. PLoS One 7, e41361. p. e41361. doi:10.1371/journal.pone.0041361

Salas, L. A., Koestler, D. C., Butler, R. A., Hansen, H. M., Wiencke, J. K., Kelsey, K. T., et al. (2018). An Optimized Library for Reference-Based Deconvolution of Whole-Blood Biospecimens Assayed Using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 19 (1), 64. doi:10.1186/ s13059-018-1448-7

Sarkar, S., and Ghosh, A. K. (2020). On Perfect Clustering of High Dimension, Low Sample Size Data. IEEE Trans. Pattern Anal. Mach Intell. 42 (9), 2257–2272. doi:10.1109/TPAMI.2019.2912599

Scherer, M., Nazarov, P. V., Toth, R., Sahay, S., Kaoma, T., Maurer, V., et al. (2020). Reference-free Deconvolution, Visualization and Interpretation of Complex DNA Methylation Data Using DecompPipeline, MeDeCom and FactorViz. Nat. Protoc. 15 (10), 3240–3263. doi:10.1038/s41596-020-0369-6

Teschendorff, A. E., Breeze, C. E., Zheng, S. C., and Beck, S. (2017). A Comparison of Reference-Based Algorithms for Correcting Cell-type Heterogeneity in Epigenome-wide Association Studies. BMC Bioinformatics 18 (1), 105. doi:10.1186/s12859-017-1511-5

Titus, A. J., Gallimore, R. M., Salas, L. A., and Christensen, B. C. (2017). Cell-type Deconvolution from DNA Methylation: a Review of Recent Applications. Hum. Mol. Genet. 26, R216. (R2): p. R216-R224. doi:10.1093/hmg/ddx275

Wiencke, J. K., Koestler, D. C., Salas, L. A., Wiemels, J. L., Roy, R. P., Hansen, H. M., et al. (2017). Immunomethylomic Approach to Explore the Blood Neutrophil Lymphocyte Ratio (NLR) in Glioma Survival. Clin. Epigenetics 9, 10. doi:10. 1186/s13148-017-0316-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Bell-Glenn, Thompson, Salas and Koestler. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Bioinformatics

METHODS
published: 12 April 2022
doi: 10.3389/fbinf.2022.859828

Check for updates

OPEN ACCESS

Edited by:
Matteo Pellegrini,
University of California, Los Angeles,
United States Reviewed

by:

Yuanyuan Zhang,
Qingdao University of Technology,
China
Yiming Bao,
Beijing Institute of Genomics (CAS),
China *Correspondence:

Lucas A. Salas
lucas.a.salas@dartmouth.edu

Specialty section:
This article was submitted to
Genomic Analysis, a
section of the journal
Frontiers in Bioinformatics

Received: 21 January 2022
Accepted: 08 March 2022
Published: 12 April 2022 Citation:

Mariani MP, Chen JA, Zhang Z,
Pike SC and Salas LA (2022)
MethylMasteR: A Comparison and
Customization of Methylation-Based
Copy Number Variation Calling
Software in Cancers Harboring Large
Scale Chromosomal Deletions. Front.
Bioinform. 2:859828. doi:
10.3389/fbinf.2022.859828

MethylMasteR: A Comparison and Customization of Methylation-Based Copy Number Variation Calling Software in Cancers Harboring Large Scale Chromosomal Deletions

Michael P. Mariani[1,2], Jennifer A. Chen[1], Ze Zhang[1,3], Steven C. Pike[1,4] and Lucas A. Salas[1]*

[1]Geisel School of Medicine, Department of Epidemiology, Dartmouth College, Hanover, NH, United States, [2]Geisel School of Medicine, Department of Biomedical Data Science, Dartmouth College, Hanover, NH, United States, [3]Guarini School of Graduate and Advanced Studies, Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH, United States, [4]Guarini School of Graduate and Advanced Studies, Integrative Neuroscience at Dartmouth, Dartmouth College, Hanover, NH, United States

DNA methylation-based copy number variation (CNV) calling software offers the advantages of providing both genetic (copy-number) and epigenetic (methylation) state information from a single genomic library. This method is advantageous when looking at large-scale chromosomal rearrangements such as the loss of the short arm of chromosome 3 (3p) in renal cell carcinoma and the codeletion of the short arm of chromosome 1 and the long arm of chromosome 19 (1p/19q) commonly seen in histologically defined oligodendrogliomas. Herein, we present MethylMasteR: a software framework that facilitates the standardization and customization of methylation-based CNV calling algorithms in a single R package deployed using the Docker software framework. This framework allows for the easy comparison of the performance and the large-scale CNV event identification capability of four common methylation-based CNV callers. Additionally, we incorporated our custom routine, which was among the best performing routines. We employed the Affymetrix 6.0 SNP Chip results as a gold standard against which to compare large-scale event recall. As there are disparities within the software calling algorithms themselves, no single software is likely to perform best for all samples and all combinations of parameters. The employment of a standardized software framework via creating a Docker image and its subsequent deployment as a Docker container allows researchers to efficiently compare algorithms and lends itself to the development of modified workflows such as the custom workflow we have developed. Researchers can now use the MethylMasteR software for their methylation-based CNV calling needs and follow our software deployment framework. We will continue to refine our methodology in the future with a specific focus on identifying large-scale chromosomal rearrangements in cancer methylation data.

Keywords: methylmaster, copy number variation, DNA methylation, kidney cancer, clear cell renal cell carcinoma, epigenetics, genomics, multiomics

INTRODUCTION

CNV calling from DNA methylation data remains an attractive prospect as both genetic (copy-number) and epigenetic (methylation) data can be obtained from the same experimental results. There are, thanks to Illumina, a pedigree of methylation analysis platforms available to the genomicist. Beginning with the Illumina Infinium HumanMethylation27 or "27k" BeadArray platform, followed by the widely used Illumina Infinium HumanMethylation450 or "450k"

BeadChip (Bibikova et al., 2011), and finally, the Illumina Infinium MethylationEPIC or "EPIC" platform, released in 2015, which sports over 850,000 methylation probes (Sandoval et al., 2011; Triche et al., 2013; Moran et al., 2016). With each generation, the researcher is availed of more comprehensive genome coverage from the added probe density (Kilaru et al., 2020). DNA methylation arrays measure bisulfite converted unmethylated cytosines as thymines. In contrast, methylated cytosines are protected from the bisulfite conversion and remain as cytosines (Aryee et al., 2014). The CNV status can then be inferred from the relative ratio of converted thymines to remaining cytosines (red and green channel intensities) (Feber et al., 2014).

SNP array-based comparative genome hybridization retains the greatest resolution and is still used as the gold standard in several public databases (Koike et al., 2011; Kilaru et al., 2020). Some time after this technology was developed, researchers later sought to identify CNVs from DNA methylation data, the need to process this data was first answered with software such as DNAcopy (Seshan and Olshen, 2021) and Minfi (Aryee et al.). Shortly thereafter, Feber et al. built off these initial software with their release of ChAMP (Morris et al., 2014; Tian et al., 2017). After ChAMP followed the CopyNumber450kCancer (Marzouka et al., 2016) and cnAnalysis450k (Knoll et al., 2017) routines which were specifically aimed at analyzing the 450k platform data.

These arrays have been validated in a variety of studies since their inception, distribution, and utilization for cancer studies, among others: e.g., 450k with colorectal cancer (Sandoval et al., 2011), again the 450k BeadChip for acute lymphoblastic leukemia (Nordlund et al., 2013), or combined with other copy-number and transcriptomic data for multi-omics studies in glioblastoma (Sturm et al., 2012). Researchers have pioneered improvement at each stage of technology development in their application, including software and wet-lab improvements that have provided the genomics community with additional accuracy and inferential capabilities—such as normalization improvements to signals generated from the 450k platform (Fortin et al., 2014).

Feber et al. described several technical challenges in the processing and analysis of DNA methylation array data. However, they also identified how CNV calling from methylation data could potentially identify translocations and inversions while the earlier genotyping arrays cannot. Additionally while, the genotyping arrays have a very high resolution, methylation-based CNV callers have an easier time identifying CNVs in genes relative to intergenic regions (Feuk et al., 2006; Feber et al., 2014; Kilaru et al., 2020).

450k platform-specific software like CopyNumber450kCancer and cnAnalysis450k incorporated computational and statistical methodology in novel ways to address these unique technical challenges. For example, the latter borrowed normalization routines from other analytical routines such as the dasen family of preprocessing methods (wateRmelon) (Pidsley et al., 2013), and ssNoob, Quantile, Funnorm, and SWAN—all part of minfi (Aryee et al., 2014). Anotherkey finding was that the normalization distributions were skewed after comparing these procedures (Knoll et al., 2017). This skewing meant differing gain/loss cutoffs would have to be applied depending on the normalization method. During the DNA methylation array processing, the copy-number inference is based on the red (unmethylated signal) and green (methylated signal) intensity channels. Technical variations in these two signal channels need to be accounted for as the relative differences between these signals allow the CNV state to be inferred. Thus, changes in intensity normalization may affect CNV calling, but not necessarily the methylation calling (Knoll et al., 2017). To alleviate both the baseline offset and distribution skew issues, the cnAnalysis450k authors chose a normalization procedure that implements a z-transformation that scales data sets relative to one another and is purported to alleviate the above issues (Knoll et al., 2017).

Even so, analytical challenges persisted, as regions that harbored genomic deletions were still producing false-positive artefactual spurious signals (Zhou et al., 2018). Zhou et al. attempted to resolve this issue by developing a statistical approach to masking deleted or "hyperpolymorphic" genomic regions (Zhou et al., 2018). Their software, SeSAMe, implemented a technique called pOOBAH or p-value with out-of-band (OOB) array hybridization. They report reducing the number of false-positive epigenetic silencing regions reported. The focus was on the tumor suppressor genes CDKN2A and RB1, often deleted in tumors (somatic deletions). The authors also claimed that their method decreased technical variation and retained biological variation across 450k and EPIC platform samples. Finally, they claim that the SeSAMe package is suitable for efficiently analyzing thousands of samples, such as those from The Cancer Genome Atlas (TCGA), which we analyze in this paper (Triche et al., 2013; Zhou et al., 2018).

Most recently, in 2019, the authors of Epicopy (Cho et al., 2019) sought to use existing tools to streamline the analysis and improve accuracy for TCGA datasets. They used functional normalization (Fortin et al., 2014) and like the other methods mentioned above, built off of the methodology of, or incorporated, DNAcopy to perform methylation signal meancentering, circular binary segmentation, and copy number estimation (Olshen et al., 2004; Feber et al., 2014; Cho et al., 2019). They applied the GISTIC 2.0 (Mermel et al., 2011) software to get copy numbers for individual genes and samples as well as to find "focal" and "arm-level" events in each tumor type (Cho et al., 2019). Our software, MethylMasteR, integrates the four different routines described above: SeSAMe, our own version of cnAnalysis450k (Called "HM450" by us), ChAMP, and Epicopy. Other CNV calling methods that have been developed along the way—such as conumee (Hovestadt and Zapatka, 2017), as well as newer methods that continue to be developed to help address the ongoing analytical challenges discussed above. See Kilaru et al., 2020 for a review (Kilaru et al., 2020). More recently, methodology is also being developed to explore methylation and copy number variation within the context of single cell Multiomics as well (Hou et al., 2016; Bian et al., 2018).

While CNV calling can be performed from DNA methylation arrays, the gold standard remains high-density SNP arrays (Koike et al., 2011; Feber et al., 2014; Cho et al., 2019). It is desirable to perform genetic (copy number variation) and epigenetic (methylation) analyses at the same time via methylation data (rather than using separate SNP Chips to identify CNV's to reduce costs and materials used (Feber et al., 2014), but researchers need to be sure that the methods available to call CNVs from methylation data are accurate and performant under various conditions. Studies need to be performed to assess said

performance and accuracy of the pipelines available to accomplish this task.

Herein, we present MethylMasteR, a software that provides a stable and user-friendly platform for running DNA methylationbased CNV calling algorithms. MethylMasteR allows for the comparison between the performance (total runtime and peak memory usage), and the ability to identify large-scale CNVs (recall) in cancer samples, across four popular algorithms that call CNVs from methylation data. We also introduce our custom routine incorporating SeSAMe and parts of the CopyNumber450kCancer CNV calling routines. The MethylMasteR software package combines the DNA methylation CNV callers using a common Sample Sheet and raw IDAT input files. Individual algorithms are then run, and CNV segments are transformed into a common data. frame format, which facilitates visualizations and comparison across algorithms. Finally, and perhaps most importantly, we have implemented our software within the Docker software architecture (Merkel, 2014) which allows us, the developer to control software dependency harmonization and versioning without hassle to the end-user.

METHODS Sample Selection

A total of 31 kidney cancer samples (KIRC) from the Firehose database on GDC′s TCGA were selected based on the status of their VHL-coding region on the short arm of chromosome 3 (3p). All 31 samples were selected to contain deep 3p deletions associated with clear cell renal cell carcinomas. The corresponding Illumina 450k BeadChip IDAT files were downloaded from TCGA. A further 50 low-grade glioma (LGG), subtype oligodendroglioma, samples were also downloaded from the Firehose database. Twenty-five samples are histologically defined oligodendrogliomas and contained codeletion of the short arm of chromosome 1 and the long arm of chromosome 19 (1p/19q), and the other twenty-five samples were astrocyte-like oligodendrogliomas that were copy-number neutral in these regions (Li et al., 2016). The Affymetrix SNP Array 6.0 (“gold standard”) copy number segmentation data corresponding to the above data was also downloaded from TCGA.

All testing was performed on a Dell Precision 5,820 Tower X-series workstation running Windows 10 Pro 64bit and R 4.1.2. 128 GB of useable RAM and an Intel(R) Core(TM) i9-10900X CPU x64 processor @ 3.70 GHz. All analyses were performed only using a single core as some packages set multiple cores for the analysis; we wanted to ensure equal computational resources were allocated for each routine.

Copy Number Variation Calling Methods

We ran four main workflows. All workflows use the same general formula to calculate the copy number state from the original methylation signal intensities that we find in the raw IDAT files: a log ratio is first calculated between the methylation signal of a test set (I_T) over that of a reference set (I_R). This could be a ratio of tumor sample intensities over normal sample intensities or over a reference intensity set such as the Epic.5. normal default samples used by SeSAMe, or the median signal intensity used by default in Epicopy. Thus $LRR \quad log_2(\frac{I_T}{I_R})$. Then a threshold is applied to the LRR values to determine the copy number state of a particular region. We chose a more liberal threshold of −0.2 and 0.2 than what is often seen in literature (Feber et al., 2014) for copy number losses and gains respectively. Regions with LRRs within this range were consider neutral.

The first workflow that we ran is the Sesame workflow. SeSAMe and the sesameData R package were used to load in the “Epic.5. normal” reference (prostate-derived cell lines and consequently male-only samples) and the tumor. IDAT files from TCGA. The openSesame () function was used to read in the files, perform background correction using the normal-out-of-band (noob) algorithm (Fortin et al., 2014), a nonlinear dye bias correction step to account for differences in the red and green fluorescent dyes used to measure methylation (Zhou et al., 2018), and pOOBAH masking to identify and correct for aberrant methylation signal stemming from hybridization failure (Zhou et al., 2018) and output data as SeSAMe signal sets. Segmentation analysis was then performed with SeSAMe using the “Epic.5. normal” reference data set.

The MethylMasteR “HM450” workflow is based on the cnAnalysis450k routine. With this routine, SeSAMe was again used to read in and preprocess the control Epic.5. normal and tumor samples from IDAT files as described above. The preprocessed signal sets were then converted to an RGChannelSet object and extracted into a GenomicRatioSet object using the Minfi preprocessRaw() and getCN() functions for both the control and treatment samples (Aryee et al., 2014). The data in this object was then z-transformed for both the tumor and control data, and the median intensity values of all control samples were also taken. The CNV segments were then calculated from the above-transformed data and the median control intensity using cnAnalysis450k functionality. The final CNV calls were plotted as heatmaps following the style of CNAclinic (Chandrananda, 2017) and output as commaseparated-values files (CSVs).

ChAMP functionality was adjusted slightly and applied to the 31 KIRC deep 3p deletion samples. The function was modified to accept a modified Illumina-style sample sheet—the MethylMasteR Style sample sheet and the sample sheet path can be specified independently from the IDAT files path. The entire ChAMP routine was run with the modified champ. process () functionality, we disabled components that were not strictly necessary for CNV calling such as differential methylation region detection, as it would increase the overall run time, and possibly the peak memory usage as well. Again, heatmaps and CSVs for the CNVs were generated as outlined above.

For the Epicopy routine, the authors were unable to get the normal data object to load with a “magic number” error, suggesting that it may need to be regenerated for the newest version of Epicopy. In place of the built-in “thyroid,” “breast,” and “lung” normal (TCGA-derived normal-adjacent samples), the median was specified as the internal control to use for these samples. Heatmaps and CSVs were again generated. Again, as with ChAMP, some of the Epicopy (and Minfi) internal functionality was modified to use the MethylMasteR sample sheet and workflow.

Our custom routine mostly follows the SeSAMe methodology, but we added additional peak correction functionality to correct baseline offsets in the segmentation calls. This was done by employing the AutoCorrectPeak () functionality from CopyNumber450kCancer, which is meant to adjust the Log R Ratios (LRRs) baseline and improve the accuracy of CNV

segment calls (Marzouka et al., 2016).

In each case, the final segmentation state was calculated by setting a segment mean LRR threshold to ≤−0.2 and ≥0.2 to copy number losses and gains, respectively, for visualization. Any segment with a mean between these two values was considered neutral. The same threshold was used across all algorithms to keep the comparisons as equal as possible.

Finally, the gold standard 31 KIRC, and 50 LGG, samples from the firehose legacy portal SNP CNV calls from the Affymetrix SNP Array 6.0, were loaded into MethylMasteR using a custom routine. Heatmaps and CSVs were again generated against which the CNVs could be compared from the above four routines.

Comparison

The peakRAM() function from the peakRAM R package (Quinn, 2017) was used in R to identify peak memory usage during each routine, and the base R Sys. time () function was used to calculate the total time elapsed.

For comparison of renal cell carcinoma CNV segments identified across routines, we chose to measure the recall of each routine against a "gold standard" reference data set: the hg19 SNP6 CNV segment data corresponding to our 31 KIRC firehose legacy test samples. The countOverlaps () function from the GenomicRanges R package (Lawrence et al., 2013) was used to calculate overlaps between CNV segments identified by the MethylMasteR routines and the gold standard segments (both tumor and normal adjacent) from the Affymetrix SNP6 chips. Here the recall was calculated as the percent of CNV reference segments correctly identified, $R \frac{(TN+TP)}{TR} = \frac{ON+(OT-ONT)}{N \cup T}$.
Overlaps normal (ON) is the true negative number (TN) and is calculated as the number of normal reference (N) (SNP6 gold standard) CNV segments that overlapped at least one routine CNV call by one or more bases. Overlaps tumor (OT) is calculated the same as ON but for the tumor reference CNV segments (T). Because true negatives were prioritized over true positives, segments that overlapped between the normal standard and tumor standard (ONT) were subtracted from OT to get the true positive (TP) number of reference segments. The total number of reference segments (TR) is the total number of segments that comprises the union of the normal reference segments (N) and tumor reference segments (T) calculated with the GenomicRanges union () function. ONT was calculated using the seqsetvis R package (Boyd, 2021).

For the oligodendroglioma samples, the recall was calculated, and comparisons were performed similarly to the above with a couple of adjustments. Because there was no tissue adjacent Affymetrix SNP6 data, we used the 25 histologically determined oligodendroglioma samples, harboring the 1p/19q codeletions as the test group and the 25 astrocyte-like oligodendroglioma samples that were neutral for these deletions as the normal group. To determine which samples contained the deletions or were neutral, the mean LRR values were retrieved from cBioPortal. Any CNV with mean intensity ≤ −0.2 was labeled a deletion (copy number = 1), ≥0.2 as an amplification (copy number = 3), and in-between values were labelled as neutral (copy number = 2). If a sample harbored no CNVs with a copy number value = 1 and met a detection threshold to ensure a large-scale event (total marks for that CNV >10,000) in both 1p and 19q, it was classified as a tumor (T) reference sample. If instead, the state was equal to 2 with the other parameters being the same, it was classified as a normal (N) reference sample.

RESULTS

Overall, the analysis provided insightful information about individual routines performance within the standardized framework provided by our software and outlined in Figure 1. Our framework also facilitates the implementation of custom functionality, such as peak correction, exemplified in Figure 2. The baseline-corrected segments (Figure 2B) have fewer false positives at the -0.2,0.2 LRR level vs the uncorrected segments (Figure 2A). The MethylMasteR program completed each subroutine successfully, and the results of the individual routines relative to one another can be seen in Figure 3.

ChAMP took over 1 h to complete the renal cell carcinoma data with the 31 test samples. Many features were disabled, including batch correction, differential methylation region calling, and gene set enrichment analysis. In addition, ChAMP was pretty memory intensive, requiring over 2 Gigabytes of RAM at its peak. The HM450 routine required the most RAM at its peak but completed the fastest by far at well under 10 min . Epicopy took the middle ground regarding time and memory usage, while SeSAMe and our closely related custom routine had the best combined time and memory

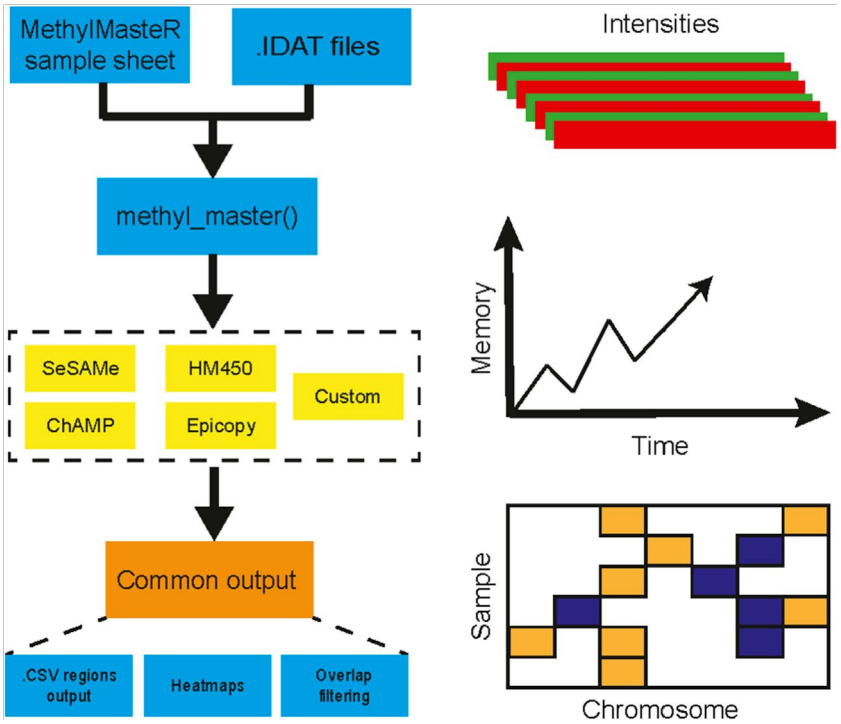


FIGURE 1 | MethylMasteR workflow overview. The MethylMasteR software package combines common methylation CNV callers using a common Sample Sheet and raw IDAT files and is implemented as a Docker image which can be run as a Docker container on any operating system that has Docker installed. Individual algorithms are then run, and CNV segments are formatted into a common data. frame format, which facilitates visualizations and comparison across algorithms. In addition to the generation of CNV heatmaps and tables, time and memory are also recorded for comparison across algorithms.

performance (Figures 2A,B). The above performance results can also be found in Supplementary Table S1.1. These patterns were observed when running the routines on the oligodendroglioma samples as well with the exception that ChAMP, Epicopy, and our custom routine required more memory than expected for the astrocyte-like (copy number neutral) oligodendroglioma samples compared to the other two categories: renal cell carcinoma and histologically determined oligodendroglioma (codeletion) samples (Supplementary Table S1.1).

We can see that Epicopy, and ChAMP performed the worst in terms of recall for copy number gains and losses for the kidney cancer samples but had improved performance, most notably with ChAMP in terms of identifying gains, in the oligodendroglioma samples. Both routines identified many neutral regions in both tumor types (Figures 3C,D). SeSAMe was able to recognize the gains and losses very well but had some trouble identifying neutral regions against the gold standard SNP6 Chip CNV calls. The HM450 routine and our custom pipeline performed the best in terms of recall, but HM450 used a high level of total peak RAM. In contrast, both SeSAMe and our custom routine (which builds on SeSAMe results) were slower than the HM450 routine (Figures 3A–D). A comparison of the final CNV states of our custom routine to the SNP6 standard (both tumor and normal) can be seen in Supplementary Figure S1. All recall results are also contained in Supplementary Table S1.2 (renal cell carcinoma) and Supplementary Table S1.3 (lowgrade glioma: oligodendroglioma).

DISCUSSION

Overall, our custom routine, SeSAMe, and HM450 performed the best in execution time, memory, and overall recall. Our custom workflow, which builds off SeSAMe, had an advantage in identifying gains and copy number neutral regions over SeSAMe alone. HM450 was best overall at identifying losses an important finding when considering that copy number losses in the short arm of chromosome 3 are essential drivers in many kidney cancers. In addition, the large-scale loss of 1p and 19q essential for the molecular diagnosis of classical oligodendrogliomas was most easily identified using HM450 as well, which thus demonstrated decent recall and run time but required the most peak RAM usage of any of the routines. This is an essential factor to consider when analyzing many samples. ChAMP and Epicopy appeared to perform the worst overall for the kidney cancer results but performed better for the oligodendroglioma samples. The many dependencies required for both of these routines may lead to a greater number of parameters that need to be fine-tuned for individual tumor and may be responsible for the variability in recall across tumor types. ChAMP held a slight edge in recall relative to Epicopy but required a significantly longer runtime and more peak RAM

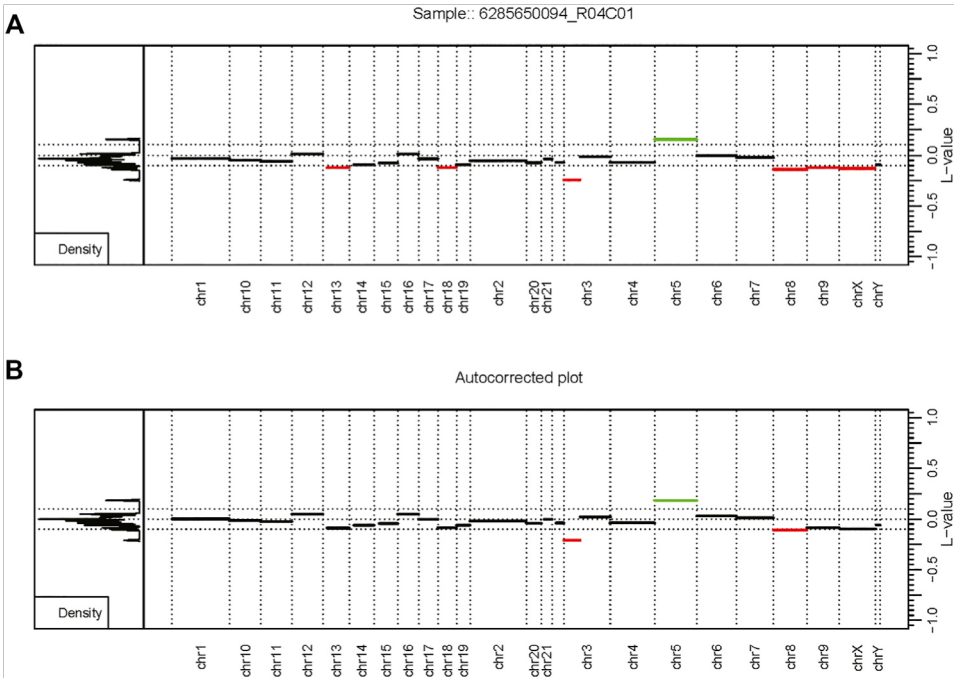


FIGURE 2 | Autocorrection as part of the custom workflow. The MethylMasteR software framework enables the easy deployment of custom workflows such as results from the SeSAMe and CnAnalysis450kCancer workflows above. (A) An uncorrected CopyNumber450kCancer-style plot.(B) The corrected CNV segmentation values after processing with SeSAMe segmentation values with the AutoCorrectPeak () function from the CopyNumber450kCancer R package. Abbreviations: L-value - the LRR value or log2 ratio of tumor methyla.

usage. There was also some divergence in expected memory usage in the copy-number neutral oligodendroglioma samples relative to the other two categories in ChAMP, Epicopy, and our custom routine as well. This demonstrates that while general patterns were observed, one must be cautious when analyzing and interpreting memory requirements for complex software that builds upon many dependencies, which are not all necessarily designed with memory optimization in hand. In addition, many of the available ChAMP features that were not essential for CNV segmentation calling were not run; thus, in many user cases, the runtime for ChAMP would be even longer. It should be noted that ChAMP offers parallel processing support that would speed up computation time, but it is not clear that this feature is supported or desirable on all operating systems due to the high RAM consumption during the serial processing.

Overall, it is crucial to note that the recall for calling CNV's using methylation data is low relative to the gold standard SNP6 results. Yet, to tumor biologists, large-scale chromosomal aberrations are of great biological interest, and we have demonstrated that some routines perform better than others in this regard. Indeed, the identification of such events is mentioned as an advantage to CNV calling with methylation data in some cases (Feber et al., 2014). At the same time, the resolution of the SNP-based CNV calling methods ensures that they will remain the gold standard for now. As we have reproduced herein, CNV calling with DNA methylation data can identify large-scale CNV events and has the advantage of providing DNA methylation information in parallel. Thus, the researcher will have to decide which cases are best suited for CNV calling with methylation data and which, or how many, analysis routines to run (Kilaru et al., 2020).

The overall formatting of the software in a standardized manner allowed us to quickly implement our custom routine involving the primary SeSAMe analysis pipeline with a modified version of the AutoCorrectPeak () function originally from the CopyNumber450kCancer pipeline. The organization of the MethylMasteR framework allowed us to push the overall gain and loss recall almost to the level of the HM450 routine, which was the best at identifying large-scale copy number losses such as the chromosome 3p deletion commonly seen in renal carcinoma.

Similarly, by modifying the Ilumina-style sample sheet that is used by minfi, ChAMP, and Epicopy, to our MethylMasteR sample sheet, we have incorporated a single input sample file that can easily be used across all the routines in our software. This feature added to our ability to streamline comparisons across the various algorithms and parameters; similarly, we can create a more efficient software platform to facilitate downstream analyses across the multiple routines by using the same internal data formats.

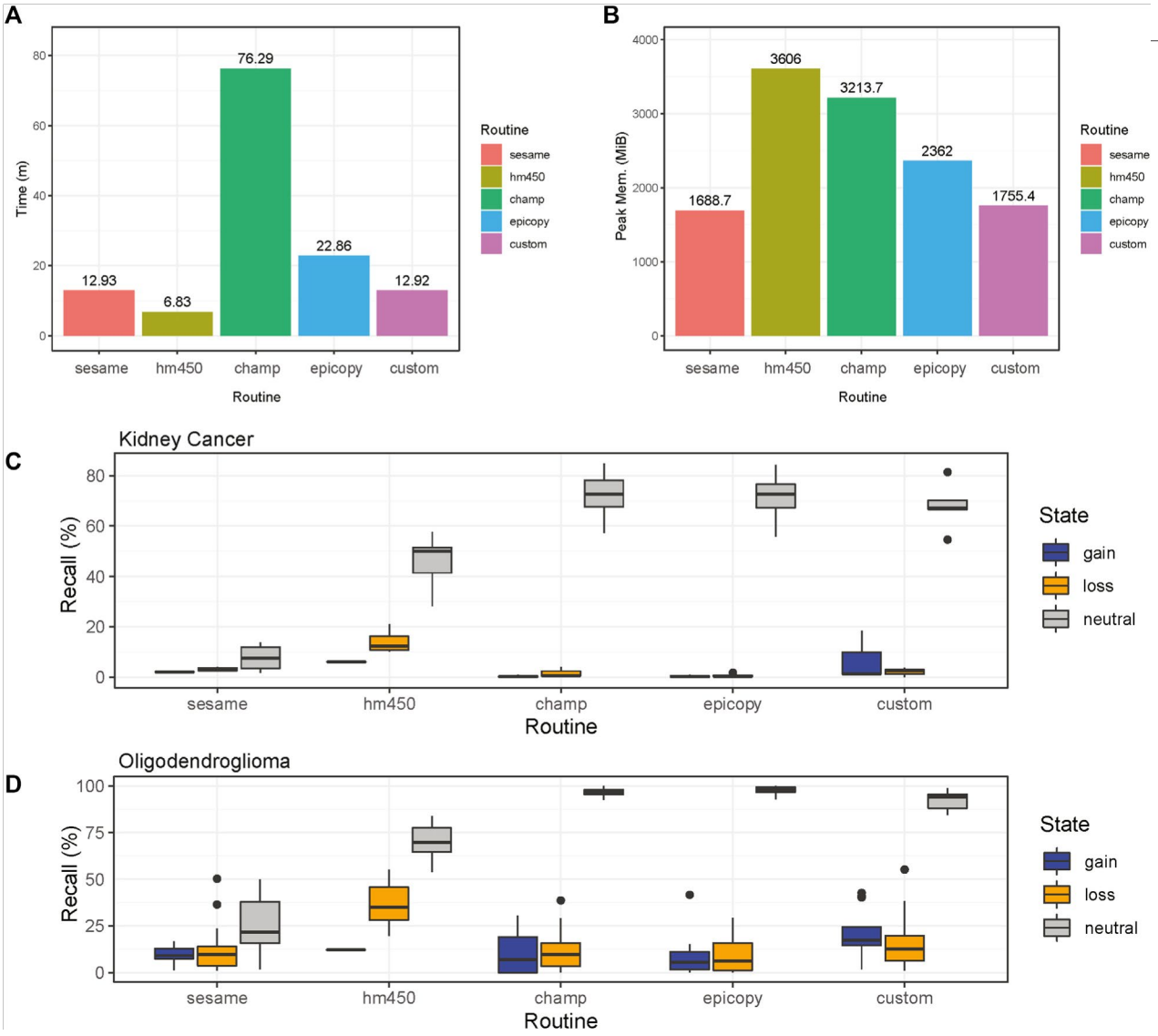


FIGURE 3 | Comparisons of CNV calling routines. Results are shown for the four common software routines incorporated into the MethylMasteR framework and the MethylMasteR custom analysis routine (A–B). ChAMP required an extended runtime and peak memory for successful completion, while HM450 was fast but memory intensive. SeSAMe and the custom routine were optimal for speed and memory usage, with Epicopy a close second (C–D). Overlaps with the Affymetrix SNP6 legacy data as the gold standard showed that Epicopy was the least accurate overall, with ChAMP having slightly better recall for identification of losses in kidney cancer but better recall for gains in oligodendroglioma. HM450 and our custom routine were the most accurate overall, and HM450 required less run time but used the highest amount of peak RAM of all the routines (except the copy number neutral astrocyte-like oligodendroglioma runtime results—see Supplementary Table S1.1). SeSAMe was not as accurate in identifying losses as our custom routine and had similar run time performance. This routine was also on par with ChAMP and Epicopy in terms of recall for the oligodendroglioma samples. In general, similar recall results to those obtained for the renal cell carcinoma samples in Figure 3C can be seen for the oligodendroglioma samples as well in Figure 3D.

The software also has additional features that allow the user to customize analyses. For instance, specifying a normal control within the sample sheet (instead of using the internal references) and the gathering of a final set of consensus regions (also output in. CSV format) using a modified version of population_ranges () function from CNVRanger (da Silva et al., 2020) are two such features. The population_ranges () function was called with default parameters: density = 0.1, rho = 0.5, and est. recur = TRUE. Another example of analysis customization is that MethylMasteR allows the user to change the LRR signal threshold parameter that is used identify copy number states. In addition to thresholds, the R equation: seg.state round($2^{seg.means}*2$) can be used by setting this parameter to "NULL" in R. Such intelligent parameter deployment is another example of the analytical versatility provided by MethylMasteR. Finally, the standardized visualization of CNV states across routines via heatmaps is useful and novel and allows for the fast and easy interpretation of data.

Finally, and perhaps most importantly, we decided to build our software into the Docker architecture, creating an image that can be downloaded and run as a container on any operating system that has Docker installed. Because Docker is now available for so many operating systems, this provides an

excellent, novel way to regulate software version control and compatibility on our end as developers and release the final product as a standalone executable for the user to run within a Docker environment. This approach reduces the added difficulty that many users experience when downloading, installing, or integrating various genomics software - all of which may not be actively maintained.

Overall, we have demonstrated how multiple DNA methylation software can be combined in a comprehensive yet efficient framework for accurately calling large-scale CNVs from raw DNA methylation signals and comparing their performance across algorithms. In the future, we plan to add additional functionality to the custom routine to improve MethylMasteR′s recall and overall accuracy further. We also will continue to refine our Docker-based approach, allowing users ease of use, and developers maximum control over, the development of genomics software requiring numerous versioned dependencies. Such a framework will enable researchers to explore the genetic and epigenetic cancer biology that can be gleaned from DNA methylation data with greater facility and accuracy.

CODE AVAILABILITY

The code used for this manuscript has been deposited in Zenodo doi: 10.5281/zenodo.6360630 (Mariani, 2022).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. Data from TCGA are available via the Genomic Data Commons (GDC).

AUTHOR CONTRIBUTIONS

MM wrote the MethylMasteR software and built the R package and Docker image. JC and LS contributed to the individual R functions. MM, JC, and LS devised algorithm (routine) comparison schema. MM and JC performed data analysis. MM, JC, ZZ, SP, and LS interpreted results. MM and LS wrote the paper. All authors read and approved the final article.

FUNDING

This work was supported by: the United States National Institute of General Medical Sciences (Number: P20GM104416/8299) and the United States Congressionally Directed Medical Research Programs Funding (Number: W81XWH-20-1-0778) to LS; the Ruth L. Kirschstein National Research Service Award (NRSA) Institutional Research Training Grant (T32) to MM; the Burroughs Wellcome Fund Big Data in the Life Sciences Fellowship to SP; and the Paul K. Richter and Evalyn E. Cook Richter Memorial Fund in the Sophomore Research Scholarship and the Presidential Scholarship Program to JC.

ACKNOWLEDGMENTS

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer. gov/tcga.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at:
https://www.frontiersin.org/articles/10.3389/fbinf.2022.859828/full#supplementary-material

REFERENCES

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: a Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays. Bioinformatics 30 (10), 1363–1369. doi:10.1093/bioinformatics/btu049

Bian, S., Hou, Y., Zhou, X., Li, X., Yong, J., Wang, Y., et al. (2018). Single-cell Multiomics Sequencing and Analyses of Human Colorectal Cancer. Science 362 (6418), 1060–1063. doi:10.1126/science.aao3791

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., et al. (2011). High Density DNA Methylation Array with Single CpG Site Resolution. Genomics 98 (4), 288–295. doi:10.1016/j.ygeno.2011.07.007

Boyd, J. R. (2021). Seqsetvis: Set Based Visualizations for Next-Gen Sequencing Data. R package version 1.14.1. doi:10.18129/B9.bioc.seqsetvis

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio Cancer Genomics portal: an Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov. 2 (5), 401–404. doi:10.1158/2159-8290.CD-12-0095

Chandrananda, D. (2017). CNAclinic: A Software Suite for Shallow Sequencing Copy Number Analysis. (R package version 1.0.). Available at: https://github.com/sdchandra/CNAclinic (Accessed March 15, 2022).

Cho, S., Kim, H. S., Zeiger, M. A., Umbricht, C. B., and Cope, L. M. (2019). Measuring DNA Copy Number Variation Using High-Density Methylation Microarrays. J. Comput. Biol. 26 (4), 295–304. doi:10.1089/cmb.2018.0143

da Silva, V., Ramos, M., Groenen, M., Crooijmans, R., Johansson, A., Regitano, L., et al. (2020). CNVRanger: Association Analysis of CNVs with Gene Expression and Quantitative Phenotypes. Bioinformatics 36 (3), 972–973. doi:10.1093/bioinformatics/btz632

Feber, A., Guilhamon, P., Lechner, M., Fenton, T., Wilson, G. A., Thirlwell, C., et al. (2014). Using High-Density DNA Methylation Arrays to Profile Copy Number Alterations. Genome Biol. 15 (2), R30. doi:10.1186/gb-2014-15-2-r30

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural Variation in the Human Genome. Nat. Rev. Genet. 7 (2), 85–97. doi:10.1038/nrg1767 Fortin, J. P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., et al. (2014). Functional Normalization of 450k Methylation Array Data Improves

Replication in Large Cancer Studies. Genome Biol. 15 (12), 503. doi:10.1186/s13059-014-0503-2

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci. Signal. 6 (269), pl1. doi:10.1126/scisignal.2004088

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., et al. (2016). Single-cell Triple Omics Sequencing Reveals Genetic, Epigenetic, and Transcriptomic Heterogeneity in Hepatocellular Carcinomas. Cell Res 26 (3), 304–319. doi:10.1038/cr.2016.23

Hovestadt, V., and Zapatka, M. (2017). Conumee: Enhanced Copy-Number Variation Analysis Using Illumina DNA Methylation Arrays. R package version 1.9.0. doi:10.18129/B9.bioc.conumee

Kilaru, V., Knight, A. K., Katrinli, S., Cobb, D., Lori, A., Gillespie, C. F., et al. (2020). Critical Evaluation of Copy Number Variant Calling Methods Using DNA Methylation. Genet. Epidemiol. 44 (2), 148–158. doi:10.1002/gepi.22269

Knoll, M., Debus, J., and Abdollahi, A. (2017). cnAnalysis450k: an R Package for Comparative Analysis of 450k/EPIC Illumina Methylation Array Derived Copy Number Data. Bioinformatics 33 (15), 2266–2272. doi:10.1093/bioinformatics/ btx156

Koike, A., Nishida, N., Yamashita, D., and Tokunaga, K. (2011). Comparative Analysis of Copy Number Variation Detection Methods and Database Construction. BMC Genet. 12, 29. doi:10.1186/1471-2156-12-29 Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for Computing and Annotating Genomic Ranges. Plos Comput. Biol. 9 (8), e1003118. doi:10.1371/journal.pcbi.1003118

Li, Y. X., Shi, Z., Aibaidula, A., Chen, H., Tang, Q., Li, K. K., et al. (2016). Not all 1p/ 19q Non-codeleted Oligodendroglial Tumors Are Astrocytic. Oncotarget 7 (40), 64615–64630. doi:10.18632/oncotarget.11378

Marzouka, N. A., Nordlund, J., Bäcklin, C. L., Lönnerholm, G., Syvänen, A. C., and Carlsson Almlöf, J. (2016). CopyNumber450kCancer: Baseline Correction for Accurate Copy Number Calling from the 450k Methylation Array.
Bioinformatics 32 (7), 1080–1082. doi:10.1093/bioinformatics/btv652

Merkel, D. (2014). Docker: Lightweight Linux Containers for Consistent Development and Deployment. Linux J. 2014 (239), 2. https://www.docker.com (Accessed March 15, 2022).

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 Facilitates Sensitive and Confident Localization of the Targets of Focal Somatic Copy-Number Alteration in Human Cancers. Genome Biol. 12 (4), R41. doi:10.1186/gb-2011-12-4-r41

Moran, S., Arribas, C., and Esteller, M. (2016). Validation of a DNA Methylation Microarray for 850,000 CpG Sites of the Human Genome Enriched in Enhancer Sequences. Epigenomics 8 (3), 389–399. doi:10.2217/epi.15.114

Morris, T. J., Butcher, L. M., Feber, A., Teschendorff, A. E., Chakravarthy, A. R., Wojdacz, T. K., et al. (2014). ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics 30 (3), 428–430. doi:10.1093/bioinformatics/btt684 Nordlund, J., Bäcklin, C. L., Wahlberg, P., Busche, S., Berglund, E. C., Eloranta, M. L., et al. (2013). Genome-wide Signatures of Differential DNA Methylation in Pediatric Acute Lymphoblastic Leukemia. Genome Biol. 14 (9), r105. doi:10.1186/gb-2013-14-9-r105

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data. Biostatistics 5 (4), 557–572. doi:10.1093/biostatistics/kxh008

Pidsley, R., Wong, C. C. Y., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C. (2013). A Data-Driven Approach to Preprocessing Illumina 450K Methylation Array Data. BMC Genomics 14, 293. doi:10.1186/1471-2164-14-293

Quinn, T. (2017). peakRAM: Monitor the Total and Peak RAM Used by an Expression or Function. R package version 1.0.2. Available at: https://github.com/tpq/peakRAM (Accessed March 15, 2022).

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., et al. (2011). Validation of a DNA Methylation Microarray for 450,000 CpG Sites in the Human Genome. Epigenetics 6 (6), 692–702. doi:10.4161/epi.6.6.16196 Seshan, V. E., and Olshen, A. (2021). DNAcopy: DNA Copy Number Data Analysis. R package version 1.66.0. doi:10.18129/B9.bioc.DNAcopy

Sturm, D., Witt, H., Hovestadt, V., Khuong-Quang, D. A., Jones, D. T., Konermann, C., et al. (2012). Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. Cancer Cell 22 (4), 425–437. doi:10.1016/j.ccr.2012.08.024

Tian, Y., Morris, T. J., Webster, A. P., Yang, Z., Beck, S., Feber, A., et al. (2017). ChAMP: Updated Methylation Analysis Pipeline for Illumina BeadChips. Bioinformatics 33 (24), 3982–3984. doi:10.1093/bioinformatics/btx513

Triche, T. J., Jr., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., and Siegmund, K. D. (2013). Low-level Processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 41 (7), e90. doi:10.1093/nar/gkt090

Zhou, W., Triche, T. J., Jr., Laird, P. W., and Shen, H. (2018). SeSAMe: Reducing Artifactual Detection of DNA Methylation by Infinium BeadChips in Genomic Deletions. Nucleic Acids Res. 46 (20), e123. doi:10.1093/nar/gky691

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Mariani, Chen, Zhang, Pike and Salas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

EPIGENETICS
2022, VOL. 17, NO. 3, 335–355
https://doi.org/10.1080/15592294.2021.1900027

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

Check for updates

Identification of a foetal epigenetic compartment in adult human kidney

John K. Wiencke[a,*], Ze Zhang[b,*], Devin C. Koestler[c], Lucas A. Salas[d], Annette M. Molinaro[a], Brock C. Christensen[d], and Karl T. Kelsey[b]

[a]Department of Neurological Surgery, Institute for Human Genetics, University of California, San Francisco, CA, USA; [b]Department of Epidemiology, Department of Pathology and Laboratory Medicine, Brown University School of Public Health, Providence, RI, USA; [c]Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS, USA; [d]Department of Epidemiology, Department of Molecular and Systems Biology, Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA

ABSTRACT

The mammalian kidney has extensive repair capacity; however, identifying adult renal stem cells has proven elusive. We applied an epigenetic marker of foetal cell origin (FCO) in diverse human tissues as a probe for developmental cell persistence, finding a 5.4-fold greater FCO proportion in kidney. Normal kidney FCO proportions averaged 49% with extensive interindividual variation. FCO proportions were significantly negatively correlated with immune-related gene expression and positively correlated with genes expressed in the renal medulla, including those involved in renal organogenesis (e.g., *FGF2, PAX8, and HOXB7)*. FCO associated genes also mapped to medullary nephron segments in mouse and rat, suggesting evolutionary conservation of this cellular compartment. Renal cancer patients whose tumours contained non-zero FCO scores survived longer. The kidney appears unique in possessing substantial foetal epigenetic features. Further study of FCO-related gene methylation may elucidate regenerative regulatory programmes in tissues without apparent discrete stem cell compartments.

ARTICLE HISTORY
Received 19 July 2020
Revised 7 January 2021
Accepted 15 January 2021

KEYWORDS
Kidney; DNA methylation; epigenetics; foetal stem cells; stem cell niche

Introduction

Our understanding of epithelial homoeostasis and regeneration draws heavily on the concept of the stem cell niche first proposed in the haematopoietic system[1]. Anatomically confined niche microenvironments promote the maintenance of stem cells, regulate their function, and there is abundant evidence of heterogeneity in these specialized environments among tissues [2,3]. While metabolic reprogramming favours glycolysis over mitochondrial oxidative phosphorylation and a dependence on oxygen-deficient conditions is a feature of the stem cell niche [4], in some epithelial tissues with hypoxic regions such as the kidney, a localized stem cell compartment has proven difficult to establish. Instead, fully differentiated adult kidney tubule epithelial cells appear to dedifferentiate and redifferentiate following injury in the absence of a restricted stem cell compartment [5–8]. The robust proliferative and reparative activity of renal epithelium exists across its segments [5,9,10]. Further, background epithelial cell division without injury suggesting programmed homoeostatic restoration in response to wear and tear damage [11]. To account for the kidney's observed regenerative capacity researchers have proposed multiple pools of regionally specified progenitors [12–14]. These observations challenge the prevailing view that epithelial homoeostasis and repair is a one-way process from undifferentiated progenitor cells to fully differentiated mature cells, and points towards mature epithelial cells acquiring a stem cell-like state upon injury [15]. The regulatory programmes that direct these processes are unknown. Based on the re-expression of embryonic genes following injury [16], we hypothesize that retention of foetal epigenetic programmes in the renal epithelium could play a role in this regenerative strategy and we further posit that this could be regulated

CONTACT John K. Wiencke john.wiencke@ucsf.edu Department of Neurological Surgery, Institute for Human Genetics, University of California, San Francisco, CA 94158, USA.

*Co-first authors

Supplemental data for this article can be accessed here.

© 2021 Informa UK Limited, trading as Taylor & Francis Group

through DNA methylation.

DNA methylation tags can be powerful tools for defining the dynamic contributions of different stem cell populations in epithelial homoeostasis [17]. Methylation tags allow fate-mapping because their patterns are somatically inherited. Numerous investigators have shown how DNA methylation signatures can form a memory trace of stem cell origin [18–27]. Recently, we used DNA methylation data from foetal and adult haematopoietic cells to devise a computational tool that estimates the proportions of cells (either haematopoietic or non-haematopoietic) derived from foetal cells [28]. All foetal tissues, including foetal kidney, contained DNA methylation sites whose status is indicative of embryonic stem cells. Based on the output of this DNA methylation-based algorithm we created an index we have termed the foetal cell origin (FCO) proportion. Among haematopoietic cells we found the FCO proportion marks a developmental transition from foetal to adult stem cells as development proceeds from foetal to mature forms; a process called stem cell maturation [29–31]. Similarly, in non-haematopoietic tissues the FCO proportion changes during developmental age. A recent survey revealed variable but generally low FCO levels in multiple adult solid tissues [32]. This latter observation suggests that transitions in some embryonic methylation features during haematopoietic cell development also occur in many diverse tissues. In this study, we explore the FCO proportion in kidney tissue in relationship to its complex anatomical and functional architecture.

The functional unit of the kidney is the nephron (Figure 1a), with multiple distinct compartments: the glomerulus, proximal tubule, loop of Henle, distal tubule, connecting tubule and collecting ducts. Decades long study has revealed how these segments perform different aspects of the filtration process [33]. Gene expression along the proximal to distal axis of the nephron is tied to its anatomic and functional segmentation [34–36]. In addition, the organization of the cortico-medullary compartments of the kidney places medullary epithelial cells in a hyperosmolar and hypoxic environment [37] that contains a paucity of lymphatics [38]. Adaptation to low oxygen entails constitutive signalling of hypoxia-inducible factor 1 alpha (*HIF- 1α*), which distinguishes the renal medulla from the cortex as well as from other parenchymal organs in the human body [39]. The low oxygen concentrations, continuous *HIF-1α* signalling, and as the greater reliance of the medulla on glycolytic energy metabolism [40] are reminiscent of conditions extant during embryonic development
[41,42].

The FCO signature represents a phenotypic block of CpG sites that display foetal methylation status and is associated with at least 928 genes. *A priori*, we recognized genes encoding transcription and growth factors fundamental in nephrogenesis within this FCO gene subset (Suppl Table 1) (e.g., *SIX2, PAX8, IRX1, HOXB7, FGF2*) [43]. *SIX2* is one of the earliest mesenchymal to epithelial morphogens required in nephrogenesis [44], and is not expressed in adult tissue [45]. *PAX8* and its homologue *PAX2* are central regulators of kidney development [46]. *IRX1* plays a crucial role in the formation intermediate segments of the embryonic nephron [47–49] and is expressed at low levels in the adult loop of Henle [34]. Another FCO related gene, *HOXB7*, is considered a specific cell lineage determining factor of collecting duct cells [50] and is expressed in the adult nephric epithelium in the loop of Henle and collecting ducts [51]. *FGF2* (basic fibroblast growth factor 2) has been implicated in early metanephric mesenchymal patterning [52], and although not expressed in adult rat nephrons [34] it is induced in acute and chronic kidney damage, where it reduces functional and structural damage [53–55].

The mammalian kidney has been shown to contain unique physiological and regenerative properties [6–8]. Importantly, a discrete stem cell compartment in the kidney has been elusive using existing markers [7], which prompted our studies of the FCO marker in kidney. Here we examined expression data from different cortical and medullary regions of the human kidney obtained

by localized biopsies at different depths of the organ [56]. We explored FCO-related gene expression and its relationship to established co-expression modules previously mapped to structural and functional regions of the human nephron [57]. Meticulous microdissection and application of bulk and single-cell RNA-Seq have revealed a deep homology of renal cell identity and function across human, mouse and rat species. Here we integrated DNA methylation and multi-species gene expression data to investigate links between the FCO proportion and renal tissue organization. We corroborated our findings in mouse using kidney single-cell RNA-Seq data [58]. In the rat, we studied comprehensive nephron segment-specific data, which catalogues more than 8000 transcripts in 14 structurally definable tubular subunits [34]. Because foetal and embryonic features are associated with renal cancer stem cells, whose frequencies increase in higher grade and stage tumours with shorter survival times [59–62] we examined the FCO fraction and its relationship with tumour characteristics and patient survival times. Our results implicate a heretofore unrecognized foetal programme as a potential controller of medullary gene expression in adult human kidney and point to the existence

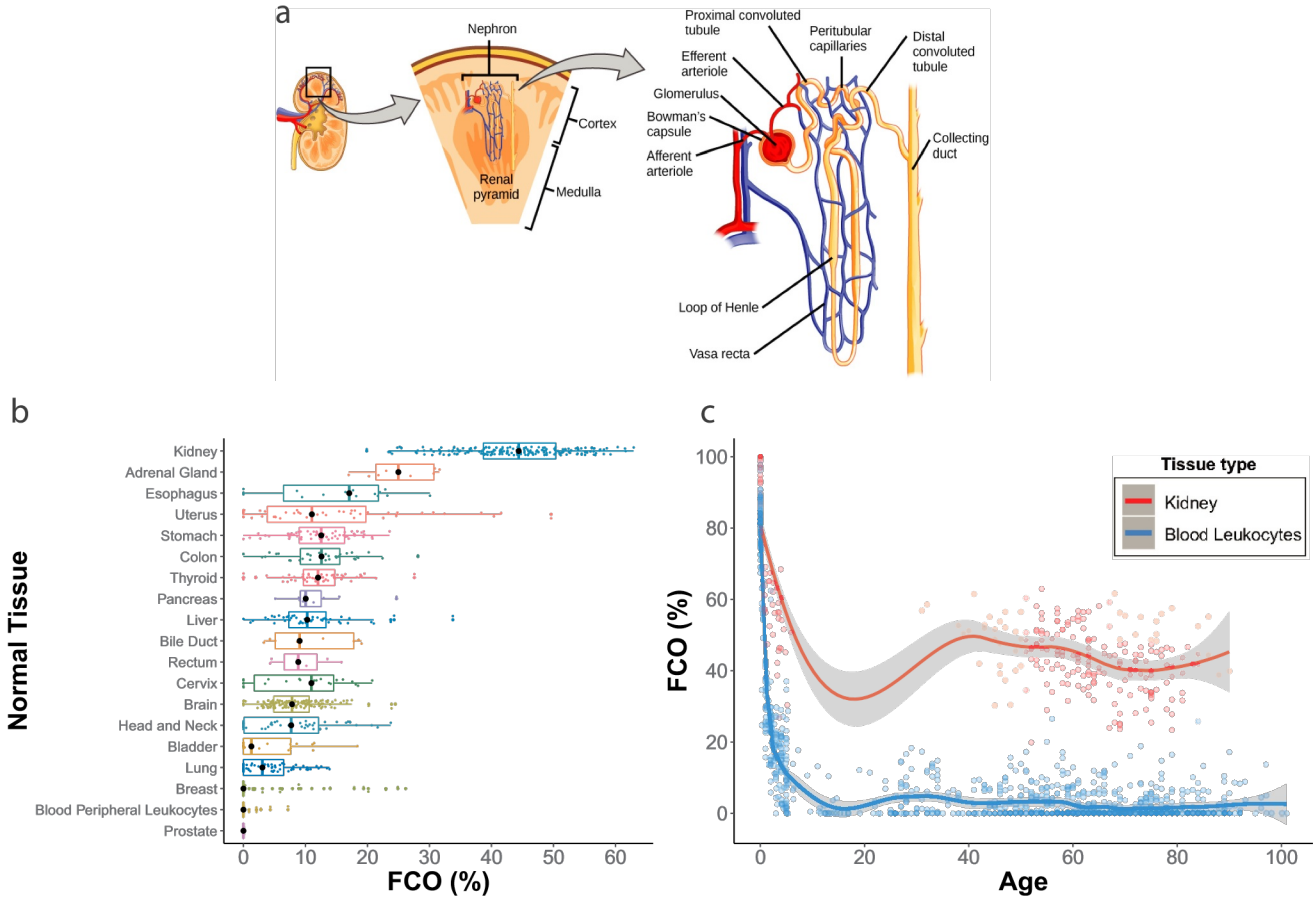


Figure 1. A-C. FCO proportions are associated with development stage, subject age and are elevated in adult kidney compared to non-renal tissues. Panel A. Schematic image of nephron (https://commons.wikimedia.org/wiki/File:Figure_41_03_03.png; permitted to use) Panel B. Horizontal box plots of median FCO proportions among adult tissues showing that adult kidney FCO proportions are higher than other solid tissues or haematopoietic cells. FCO proportions estimated in 964 normal tissue samples from 20 TCGA studies and 4 GEO studies. **Panel C**. Cross-sectional plot of the FCO proportions of human foetal, paediatric and adult kidney specimens and blood leukocytes by tissue donor age. Loess smoothing curves and 95% confidence intervals across different ages showing the relationship of donor age with FCO proportions. Data sources supplied in methods.

of a widespread niche of epigenetically distinct progenitors.

Methods

DNA methylation studies

We analysed DNA methylation data from 348 kidney specimens in 6 studies and 1812 other tissue specimens in 31 studies. Specifically, non-tumour adult kidney and other solid tissues from TCGA Level 3 Illumina Infinium HumanMethylation450 BeadChip array data were collected from 739 normal tissue samples from 20 TCGA studies (Suppl Table 2). Additional normal adult tissue samples (N = 225) of cervix, brain, adrenal gland and stomach from GEO data sets were downloaded from GSE46306 [63], GSE80970 [64], GSE77871 [65] and GSE103186 [66]. Foetal normal kidney sample data were downloaded from GSE69502 [67] and GSE76641 [18] (N = 22) and paediatric normal kidney sample data were downloaded from GSE59157 [68] (N = 36). We accessed additional datasets of microdissected renal cortex from diabetes-related chronic kidney disease (CKD) (N = 21) and CKD free controls (N = 64), GSE50874 [69]. DNA methylation data of 1027 blood leukocyte samples from birth to adulthood were downloaded from GSE62219 [70], GSE83334 [71], GSE36054 [72], GSE40279 [73], GSE35069 [74], GSE59065 [75] and GSE30870 [76] and 26 cord blood samples from packages FlowSorted.CordBlood.450 K[77] and FlowSorted.CordBloodNorway.450 K [77] on Bioconductor (Suppl Table 3).

Gene expression studies

Human kidney (RNA-Seq) expression data were downloaded from normal TCGA kidney samples [78] and biopsy depth microarray expression from [56]. RNA-Seq data from microdissected rat nephron [34] and single-cell mouse kidney [58] were accessed.

Transcriptome-wide expression correlations with FCO proportions

We also examined genome-wide kidney gene expression to identify transcripts associated with the FCO proportions of specimens, in contrast to the metagene described above, which are restricted to FCO genes that are associated with DMRs in either the extended (N = 1218 CpG) or algorithmic subset (N = 27 CpG). Specifically, we computed the Pearson and Spearman correlation between the FCO proportion and the expression levels of each gene among samples in the TCGA normal kidney data for whom both DNA methylation (necessary for FCO estimation) and gene expression data were available (N = 47). Suppl Table 4 contains the gene-specific Spearman and Pearson correlation coefficients with the FCO.

Gene expression metagenes associated with FCO proportions

FCO-related kidney metagenes for positively and negatively correlated genes were created using Level 3 RNA-seq gene-level HTSeq-FPKM data from 104 normal kidney samples from the KIRC and KIRP TCGA cohorts. The extended FCO subset of N = 1218 CpGs is associated with 928 genes, herein referred to as FCO subset genes. We restricted our analysis to transcripts with zFPKM > −3 as recommended [79]. Within normal kidney samples, 630 of these FCO subset genes were identified as significantly expressed and of these 187 were judged positively and 189 negatively correlated with the FCO proportion. Significant correlations were assigned if the expression Spearman's coefficient FDR was <0.1, after adjusting for age and gender (FCOgeneExpression = FCO + age +gender). Metagenes were created for positively and negatively correlated genes separately by taking the log2 transformed FPKM value of each gene after adding an offset of 1 and scaling the mean value for each gene to 0. Finally, samples were assigned an FCO score by calculating the mean expression of all differentially expressed FCO

genes. We also created metagenes for genes associated with the 27 CpGs contained in the computational probe set which is used in the algorithm for estimating the FCO proportion [28]; this restricted computational subset contains 33 genes. Among them, the expression levels of eight were identified as positively and four negatively correlated with the FCO proportion. Suppl Table 5 lists the metagene descriptions.

Biopsy depth gene expression analyses

We correlated FCO proportions and FCO metagene expression scores with previously published expression signatures that distinguish four different kidney locations (outer cortex, corticomedullary junction, outer medulla and inner medulla) in histologically normal adult kidney [56]. Samples were procured using fine-needle cores from approximately 10-mm-thick sections obtained from five nephrectomized patients in surgery for renal neoplasms. Tissue farthest from the tumour was selected. Gene expression data were collected using the HumanHT-12 4.0 expression BeadChip (Illumina Inc. San Diego CA) platform. A total of 102 samples were available to correlate biopsy depth expression and FCO metagene expression data. A total of 47 samples with biopsy depth expression also had DNA methylation data to allow correlations with FCO proportion.

Mouse kidney single cell-derived metagenes

Mouse renal cell expression metagene signatures were derived from a catalogue of mouse kidney cell types obtained in an unbiased manner using droplet-based single-cell RNA sequencing [58]. Clustering analyses based on 43,745 single cells identified 16 distinct cell clusters, which were assigned a cell identity through multiple corroborating sources including: bulk RNA sequencing from microdissected rat nephron [34], microarray data on human immune cell, and transgenic mouse models. Cell-specific marker genes for mouse kidney ascending loop of Henle, collecting duct principal cell, proximal tubule, distal convoluted tubule, B lymphocyte and T lymphocyte were downloaded and used as input for creating metagene signatures (Suppl Table 5). Each mouse cell metagene was correlated with the FCO proportion and FCO related expression scores of normal human kidney samples of the TCGA.

Mapping FCO differentially expressed genes to anatomic segments of the rat nephron

RNA-Seq data from rat microdissected glomerulus and 14 nephron segments were retrieved using the publicly accessible portal of the Epithelial Systems Biology Laboratory [34] https://hpcwebapps.cit.nih.gov/ESBL/Database/NephronRNAseq/All_transcripts.html. Rat renal tubule segments manually dissected under microscope were lysed in cell lysis buffer containing reverse transcriptase and oligo-dT primers. After reverse transcription of poly(A)-tailed mRNA transcripts, cDNAs were amplified and sequenced using Illumina HiSeq 2000 sequencer. FASTQ sequences were mapped to rat reference genome and gene expression levels of all genes were determined by calculating the reads per kilobase of exon model per million mapped reads (RPKM). To explore the representation of human FCO related gene expression within anatomic segments of the rat kidney, we first identified all human kidney genes (TCGA normal kidney) whose expression exhibited a Spearman correlation of at least > |0.6| with the corresponding FCO proportion of each sample. From this gene set, we identified G = 458 high-confidence human/rat orthologs, determined using the Ensembl database (http://www.ensembl.org/bio mart/martview), that were expressed in at least 5% of the samples from different anatomical segments of the rat kidney. The RPKM values for these 458 were log2-transformed (with an offset of 1 to account for RPKM values of zero), and used as input for a t-distributed Stochastic Neighbour Embedding (t-SNE) analysis to visualize the high-dimensional structure of the data in a lower-dimensional space (two-dimensions, in this case) [80]. The R-package

Rtsne [81] was used for the t-SNE analysis using default settings for all parameters, except number of dimensions (ndim = 2) and perplexity = 8. To gain insight into the genes driving the t-SNE clusters, *G* = 30 genes (out of the 458 used in t-SNE) were used to construct a hierarchical clustering heatmap to visualize influential genes across different renal tubule segments. These 30 genes were selected such that they exhibited the smallest p-values based on a Kruskal-Wallis test used to compare log2- transformed RPKM values between the different renal tubule segments (e.g., G, S1, S2, S3, SDL, etc.).

Data processing and quality control

Level 3 Illumina Infinium HumanMethylation450 BeadChip array data on TCGA contain beta values calculated from background-corrected methylated (M) and unmethylated (U) array intensities as Beta = M/(M + U). In these data, probes having a common SNP within 10bp of the interrogated CpG site or having overlaps with a repetitive element within 15bp from the interrogated CpG site are masked as 'NA' across all samples, as were probes with a non-detection probability ($P > 0.01$) in a given sample [32]. Level 3 TCGA RNA-seq gene-level expression data used are HTSeq-FPKM (Fragments Per Kilobase of transcript per Million mapped reads) data that normalize read count based on gene length and the total number of mapped reads [82]. For GEO DNA methylation data used, GSE46306 [63], GSE69502 [67], GSE59157 [68] and GSE40279 [73] contain normalized average beta values processed by BeadStudio software; GSE80970 [64] contains quantile-normalized beta values by using Dasen function within the wateRmelon package in Bioconductor; GSE77871 [65], GSE50874 [69], GSE36054 [72], GSE35069 [74] and GSE30870 [76] contain average beta values processed by the GenomeStudio software; GSE103186 [66] contains normalized beta values by using the minfi package in Bioconductor; GSE76641 [18] contains functional normalized beta values by using R; GSE62219 [70] contains quantile normalized M-values by using the GenomeStudio software; GSE83334 [71] and GSE59065 [75] contain normalized beta values by using the Subset-Quantile Within Array Normalization (SWAN) in the *minfi* Bioconductor package. In addition, raw DNA methylation data on FlowSorted.CordBlood.450 K [83] and FlowSorted.CordBloodNorway.450 K [77] were processed by using the *minfi* package in Bioconductor.

Calculation of the foetal cell origin (FCO) proportion

The derivation of the FCO algorithm was estimated as previously described [28]. Briefly, this method achieves a deconvolution of foetal versus adult cell fraction using constrained projection quadratic programming (CP/QP) proposed by Houseman [84], substituting the default reference library with the library identified based on the pipeline developed to discriminate foetal and adult blood leukocytes [28]. For analyses using GEO data sets, no additional normalization steps were employed to the already preprocessed β- values. β-value distributions were, however, inspected for irregularities, and where relevant, *k* nearest neighbours were performed for missing value imputation.

Metascape gene enrichment analysis

We assessed genes whose expression was positively and negatively correlated with the FCO proportion of kidney samples. Four sets of genes were characterized by gene enrichment analysis, including top 250 genome-wide genes whose expression are negatively correlated with FCO proportion based on the Spearman's correlation coefficient ranking, 2996 genome-wide genes whose expression are positively correlated with FCO proportion at the Spearman's correlation coefficient cut-off 0.45, 187 FCO genes whose expression were judged as positively correlated with FCO proportion and 189 FCO genes whose expression were judged as negatively correlated with FCO proportion (FDR < 0.1). The gene enrichment analyses were conducted with Metascape software [85], database updated on

2020–09-16. For each given gene list, pathway and process enrichment analysis has been carried out with the following ontology sources: KEGG Pathway, GO Biological Processes, Reactome Gene Sets, Canonical Pathways and CORUM. All genes in the human genome have been used as the enrichment background [85]. Terms with a p-value <0.01, a minimum count of 3, and an enrichment factor >1.5 (the enrichment factor is the ratio between the observed counts and the counts expected by chance) are collected and grouped into clusters based on their membership similarities [85]. More specifically, p-values are calculated based on the accumulative hypergeometric distribution, and q-values are calculated using the Benjamini- Hochberg procedure to account for multiple testing [85].

Additional kidney metagene scores

A total number of 823 genes in 25 quality threshold (QT) clusters representing kidney glomerulus, cortex, loop of Henle, and medulla formed co-expressed gene modules [56]. These co-expression modules were compared with the FCO proportion and FCO related expression metagenes. In addition, since the level of renal tissue fibrosis and tissue inflammation increases with age and often differs between individuals, we included metagenes representing different aspects of kidney damage [86,87]. A number of gene signatures were used to calculate the relative sample-level gene signature scores. Two gene signatures were downloaded from the Broad Institute Molecular Signatures Database (MSigDB; http://www.broadinstitute.org/gsea/msigdb/index.jsp): 97 genes found upregulated in transplant biopsies following acute kidney rejection as compared to well-functioning transplants [86], and 261 genes associated with ageing kidney histology [87], respectively (Suppl Table 5).

Epigenetic mutations in clear cell

We examined the distribution of previously identified missense and truncating mutations in genes that encode for histone-modifying enzymes in FCO-related survival subgroups. These genes include PBRM1, BAP1, SETD2 [78].

Statistical analysis of survival times TCGA survival data

Tree-based models of renal cancer overall survival were constructed using recursive partitioning via the partDSA algorithm [88,89] in an effort to create homogeneous survival risk groups. Models incorporated relevant patient variables, including patient age, gender, TMN class, tumour laterality, tumour FCO score. FCO score was treated as a continuous variable. partDSA models were built and chosen using the Brier loss function and fivefold cross- validation.

Statistical analysis to estimate the proportion of variation in FCO explained by kidney depth measurements

The function rsq.partial in the R-package *rsq* was used to compute the proportion of variation in FCO explained by depth measurements (e.g., values of cortex, cortex-medulla, outer medulla, and medulla papillae) not explained by subject age, gender, and race. Calculation of the partial R[2] was carried out using sums of squares estimates from two nested linear regression models, each fit to FCO as the dependent variable. Model 1 included subject age, gender, race as covariates, and Model 2 included all of covariates in Model 1, plus depth measurements of the cortex, cortex-medulla, outer medulla, and medulla papillae.

Database and software

All data sets from TCGA (https://portal.gdc.cancer.gov/), GEO (https://www.ncbi.nlm.nih.gov/geo/), ESBL (https://hpcwebapps.cit.nih.gov/ESBL/Database/NephronRNAseq/All_transcripts.html.) and the Ensemble database (http://www.ensembl.org/biomart/martview) are publicly

available. The software R studio with R version 3.6.3 and Metascape (https://metascape.org/gp/index.html#/main/step1) [85] are used to analyse the data and create figures and graphs.

Results

An overview of the analyses is shown in Suppl Figure 1 and a summary of the data sources is presented in Suppl Table 2 and Suppl Table 3.

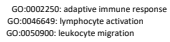
Human adult kidney displays high FCO proportions compared to other tissues and wide inter-individual variation

Comparing FCO proportions between adult haematopoietic cells and 18 solid tissues revealed that the kidney contained the highest FCO proportions (Figure 1b, N = 964). Mean FCO proportions of other solid organs averaged 8.2%, whereas adult kidney was 43.9%; a 5.4-fold higher FCO proportion ($p < 0.001$; t test). The second highest FCO proportions were observed in adrenal tissue (25.3%). Adult leukocytes have an average FCO of zero, although some outliers showed low to modest frequencies. To explore the influence of age on the renal FCO proportion, we plotted the FCO proportion by age in kidney (N = 263) and blood leukocytes (N = 1053) (Figure 1c). The FCO of paediatric kidneys (N = 36, average age = 2.8 yrs.) was lower than foetal samples (N = 22, 55.1% versus 96.5%, $p < 0.001$; t-test), and adult kidney (N = 205,18–90 yrs.), FCO proportions were lower compared with paediatric samples (43.9% vs 55.1%, $p < 0.001$; t-test). However, both paediatric and adult scores were markedly elevated compared to adult blood leukocytes. Wide individual variation in adult renal FCO proportions was observed with a subtle trend towards lower scores as age increased (Pearson's $r = -0.32$, $p < 0.001$). We studied FCO proportions in an additional adult dataset (GSE50874) consisting of micro-dissected kidney cortex enriched for proximal tubular epithelium (N = 85) in contrast to the normal bulk biopsy specimens of the TCGA (N = 205); including control samples (N = 64), and those with evidence of chronic kidney disease (CKD) (N = 21) [69]. Both micro-dissected sample sets contained lower FCO proportions compared to normal bulk kidney samples from TCGA (35.2% non-CKD versus 43.9% TCGA normal, $P < 0.001$; 23.2% CKD versus 43.9%TCGA normal, $P < 0.001$; t test) see Suppl Figure 2. The FCO proportions of CKD affected samples was lower than the control microdissected specimens (35.2% non-CKD versus 23.2% CKD, $P < 0.001$; t test).

Transcriptome-wide gene expression correlations with FCO proportions and enriched biological pathways

Two analyses identified differential gene expression (DGE) patterns associated with the FCO proportion. The first used all transcripts (N = 56,716) expressed in the TCGA normal kidney samples, and the second was a supervised analysis restricted to FCO genes (N = 928) containing DMRs (Suppl Table 6–9). Results from the genome-wide gene set (N = 20,851) enrichment analysis using the top 250 genes negatively correlated with FCO proportions (ranked by Spearman's r, Suppl Table 6) is summarized in Figure 2a. The ontology classes enriched among negatively correlated genes were strongly concentrated in immune functions including B cell, T cell activation and natural killer cell functions. For example, nine B cell immunoglobulin heavy-chain loci (IGHM, IGHA1, IGHA1, IGHG1, IGHG2, IGHG3, IGHG4, IGHE, IGHD) and the constant region of

a



b



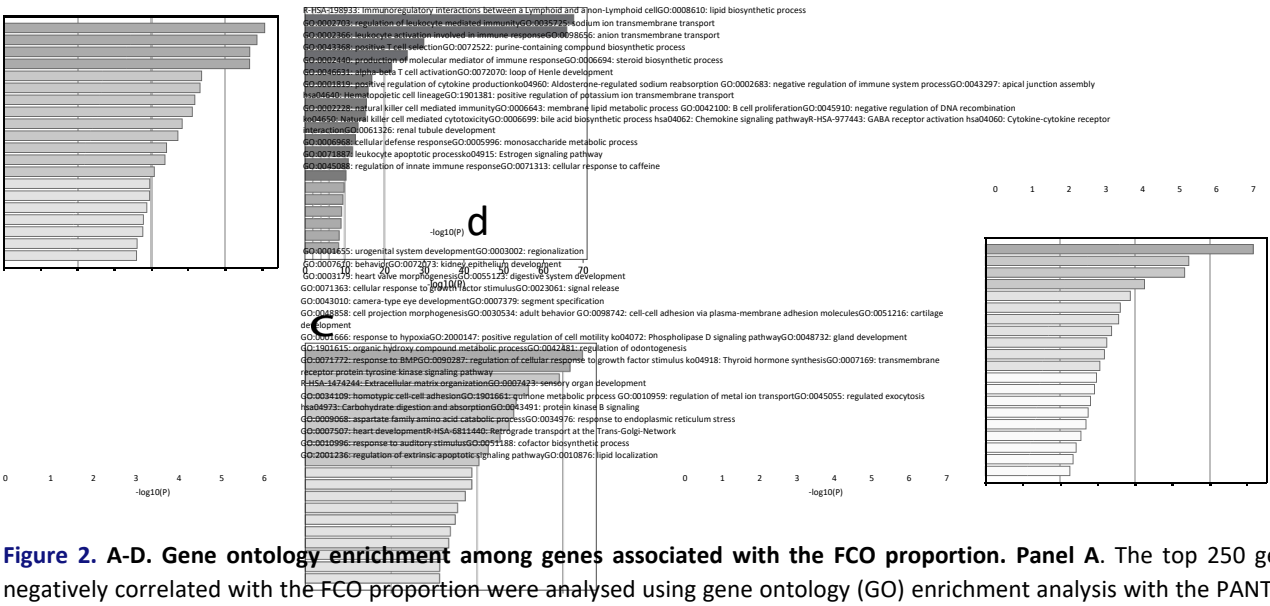


Figure 2. A-D. Gene ontology enrichment among genes associated with the FCO proportion. Panel A. The top 250 genes negatively correlated with the FCO proportion were analysed using gene ontology (GO) enrichment analysis with the PANTHER Overrepresentation Test. We present the top 20 GO biological process pathways in the table based on FDR ranking. **Panel B**. Top 20 GO terms enriched with genes exhibiting a positive correlation with FCO proportion based on a transcriptome-wide association analysis. **Panel C**. Top 20 GO terms enriched with genes exhibiting a negative correlation with FCO proportions based on restricted subset of genes containing FCO DMRs. **Panel D**. Top 20 GO terms enriched with genes exhibiting a positive correlation with FCO proportions based on restricted subset of genes containing FCO DMRs.

immunoglobulin light chains IGLC2, were represented. A recent study [90] reported B cells to be exclusively located in adult kidney cortex. Consistent with this result, we found a strong inverse correlation of the nine heavy-chain metagene expression score with the adult FCO proportion (Pearson's r = −0.82; p < .001). CD3D and CD3E in the list are part of the T-cell receptor/ CD3 complex (TCR/CD3 complex) and involved in T-cell development and signal transduction [91].

Results from the genome-wide gene set (N = 20,851) enrichment analysis using 2996 genes positively correlated with FCO proportions (Spearman's r > 0.45, Suppl Table 7) are summarized in (Figure 2b). Prominent categories observed were potassium ion transport, loop of Henle, and renal tubule development. Among the genes included are KCNJ family potassium channels and Na^+/K^+- ATPases localized in distal tubule [92]. Genes positively correlated with the FCO are predominately epithelial in origin compared with negatively correlated genes, which are haematological in origin.

FCO subset gene expression correlations with FCO proportions and enriched biological pathways

In the second set of gene enrichment analyses, we restricted our DGE pattern scans to the 928 genes whose methylation is used for FCO inference, the FCO DMRs [28]. Of the 630 genes expressed in kidney, 187 (FDR < 0.1) were positively correlated with FCO proportion, and 189 (FDR < 0.1) FCO genes were negatively correlated; we refer to these gene sets as positive FCO expression subsets, and negative FCO expression subsets, respectively. Among 189 genes containing DMRs and whose expression was negatively correlated with the FCO proportion (Suppl Table 8), we observed enrichment of genes involve in urogenital systems development (Figure 2c). For instance, *WNT6* is expressed in the branching ureteric bud during the early stages of organogenesis [93] and the

inactivation of NFIA causes both ureteropelvic junction and ureterovesical junction anomalies in murine models and humans [94]. Among 187 positive FCO subset genes (Suppl Table 9), we observed enrichment of GO terms for regionalization, kidney epithelial differentiation and segment specification Figure 2d. Regionalization is a pattern specification process resulting in the subdivision of an axis or axes in space to define an area or volume that specific cell differentiation patterns will take place. For example, we identified *HOXA2, HOXA11, HOXB3, HOXB6, HOXB7* and *HOXD9* in our positive FCO subset gene list and the HOX genes are proven to be instrumental in establishing the positional identity of cells along the embryonic anteroposterior (A-P) axis [95]. Also, *PAX8* in the list has been proved to regulate branching morphogenesis and nephron differentiation in the developing kidney [96]. Furthermore, *FGF2* in the list was implicated in early metanephric mesenchymal patterning [52].

Differentially methylated regions (DMRs) in FCO genes FGF2, PAX8 and HOXB7 across tissues

To provide three examples of FCO-related DMRs in foetal kidney, adult kidney, and adult blood leukocytes, we plotted the methylation beta values of *FGF2, PAX8* and *HOXB7* which have established roles in embryonic kidney development (Figure 3 upper panels). The errors represent standard deviations for the methylation proportions of 22 foetal kidneys [18,67], 205 normal adult kidneys (TCGA renal clear cell KIRC and papillary KIPR) and 36 adult leukocytes samples [74]. Genomic regions containing FCO DMRs displayed lower methylation values in adult kidney than adult leukocytes, indicating that adult kidney does not mature as much as adult leukocytes. Foetal kidney methylation values were typically lowest compared to adult kidney and white blood cells. Individual CpG values for each gene were correlated with each estimated tissue FCO proportion (Figure 3 Middle panel). We also found significant correlations of gene expression with methylation at the CpG level (Figure 3 Lower Panels) indicating that these differentially expressed genes in adult tissues contain methylation marks shared with embryonic stem cells. Furthermore, we added 151 foetal leukocytes samples [77,83], 12
embryonic stem cell-line samples [97] (Suppl

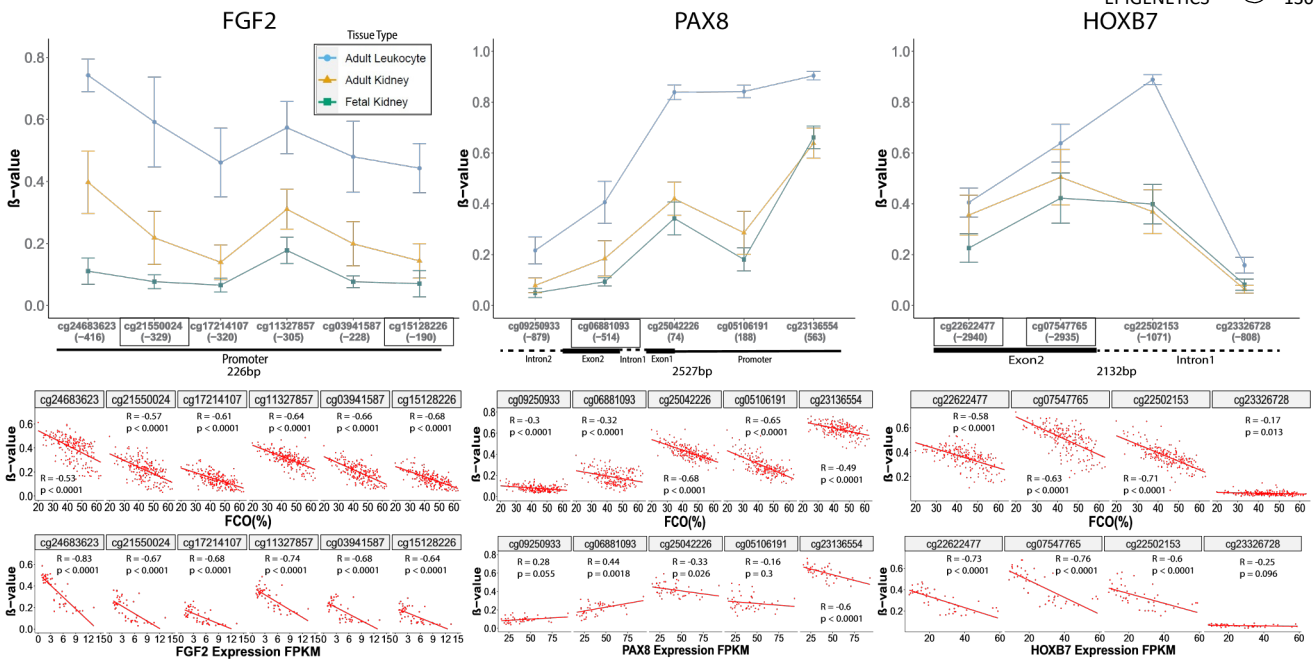


Figure 3. DNA methylation levels of FCO related genes depend on tissue type and are correlated with gene expression in adult kidney. Three examples of FCO-related DMRs in foetal kidney, adult kidney and adult blood leukocytes are shown; *FGF2, PAX8*, and *HOXB7* genes. The upper panels show that adult leukocytes, which have a non-detectable FCO score, demonstrate higher beta values (Y-axis) compared to foetal kidney and that adult kidney methylation tracks with foetal tissues (boxed are the FCO related CpGs). Error bars represent standard deviations for 22 foetal kidneys (GSE69502 and GSE76641), 205 normal adult kidneys (TCGA), and 23 adult blood leukocytes. The lower panels depict the correlations of each CpG site methylation value with the corresponding sample FCO proportion (middle panel) and the corresponding RNA-Seq expression score (lower panel) in adult kidney.

Figure 3a) and paediatric kidney samples (Suppl Figure 3b) to the methylation error bar plot [68].

Renal depth biopsy reveals that FCO related gene expression localizes to the medulla

Using the 187 genes positively correlated with FCO proportions (Suppl Table 9) we created a metagene expression signature to serve as a transcriptional correlate of the foetal methylation signature. We reasoned that associating FCO gene expression with biopsy depth-specific gene expression could help localize the cells whose expression is affected by the FCO. Four different regions representing nephron biopsy depth were previously differentiated by their distinctive gene expression [56]. We regressed these established biopsy depth expression values (X-axis) with each sample's FCO expression metagene score (Y-axis) at each of four different biopsy depths (Figure 4a). The resulting plots revealed clear linear trends of FCO with biopsy depth expression scores within each kidney region: cortex, cortico-medullary, outer medulla, inner medulla. However, the slopes of the regressions were positive for outer and inner medulla but were negative for the cortex and cortico-medullary depth samples. Similarly, the 189 genes identified as negatively correlated with the FCO proportion were used to create a negative FCO expression score (Suppl Table 8). These negative FCO parameters were also regressed against regional biopsy expression scores. The resultant pattern was a mirror image of the relationships observed for the positive expression scores (Suppl Figure 4). Metagene scores for negatively correlated

genes were positively associated with cortex and cortico-medullary depths and negatively associated with outer and inner medulla. Taken together, these observations show that genes whose expression increases with increasing FCO proportion are most strongly associated with medullar depth biopsy. In contrast, gene expression inversely related to the FCO proportion is related to the renal cortex. Using the depth metagene score for each sample in conjunction with donor age we modelled the FCO proportion and estimated that 33% of the variation in the normal kidney proportions was associated with variation in depth metagene score and hence potentially attributable to cell composition.

Co-expression modules implicate FCO proportions with distal tubule and medullar segment gene expression

To further understand FCO expression metagenes' relationship with the segmented structure of the human nephron, we aligned FCO expression parameters with a previously validated model of nephron co-expression 16 epithelial modules and additional inflammation-related signatures (Figure 4b) [56]. Both the FCO positive (n = 8) and negative (n = 4) expression metagenes are displayed in their relationship to co-expression module scores and show a coherent pattern of association between positive FCO metagenes

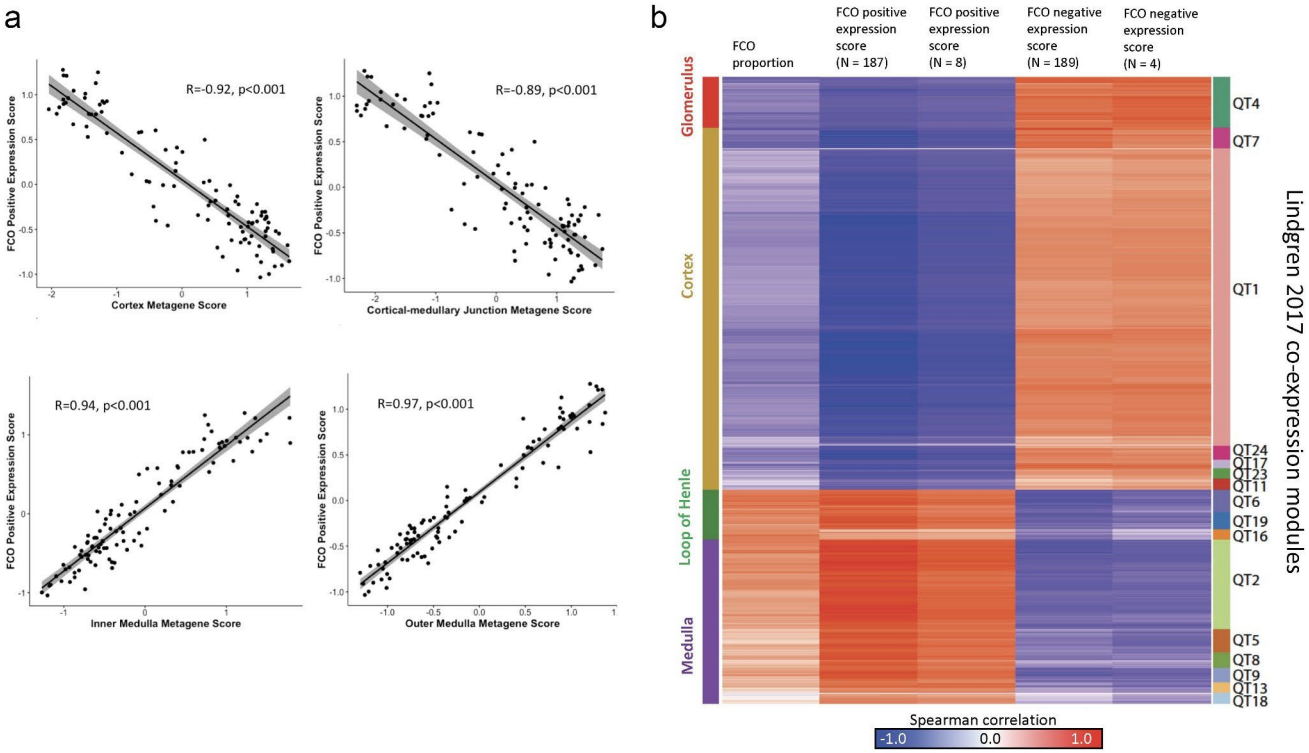


Figure 4. A-B. Gene expression signatures positively correlated with FCO proportions localize to renal medulla in adult human kidney. Panel A. The expression of genes positively correlated with the FCO proportion are highly correlated with differentially expressed genes of the renal medulla. Metagene signature scores for FCO positively correlated genes are plotted against four published metagene expression scores of cortical and medullar biopsy depth biopsies [56] (95% confidence interval indicated by grey shades). **Panel B**. Heatmap showing correspondence of FCO positive and negatively correlated gene metagenes with the co-expression modules of Lindgren DA 2017[56]. Co-expression modules correspond to published quality threshold (QT) module designation and are presented in the anatomic order as presented therein.

and distal nephron modules. In contrast, the metagenes negatively associated with the FCO proportions displayed a cohesive association pattern with glomerulus and proximal tubular location. These observations are reminiscent of the biopsy depth correlations, indicating robustness, this pattern was consistent even using a limited number of FCO related genes. Immune-related gene expression was inversely associated with the FCO proportion, consistent with results of other inflammation modules. Immune-related signatures were in general inversely associated with the FCO proportion and the positive FCO metagene score (Suppl Figure 5). *HIF-1* signalling pathway metagenes were significantly positively correlated with the FCO proportion and the positive FCO metagene score ($p < 0.01$) (Suppl Figure 6).

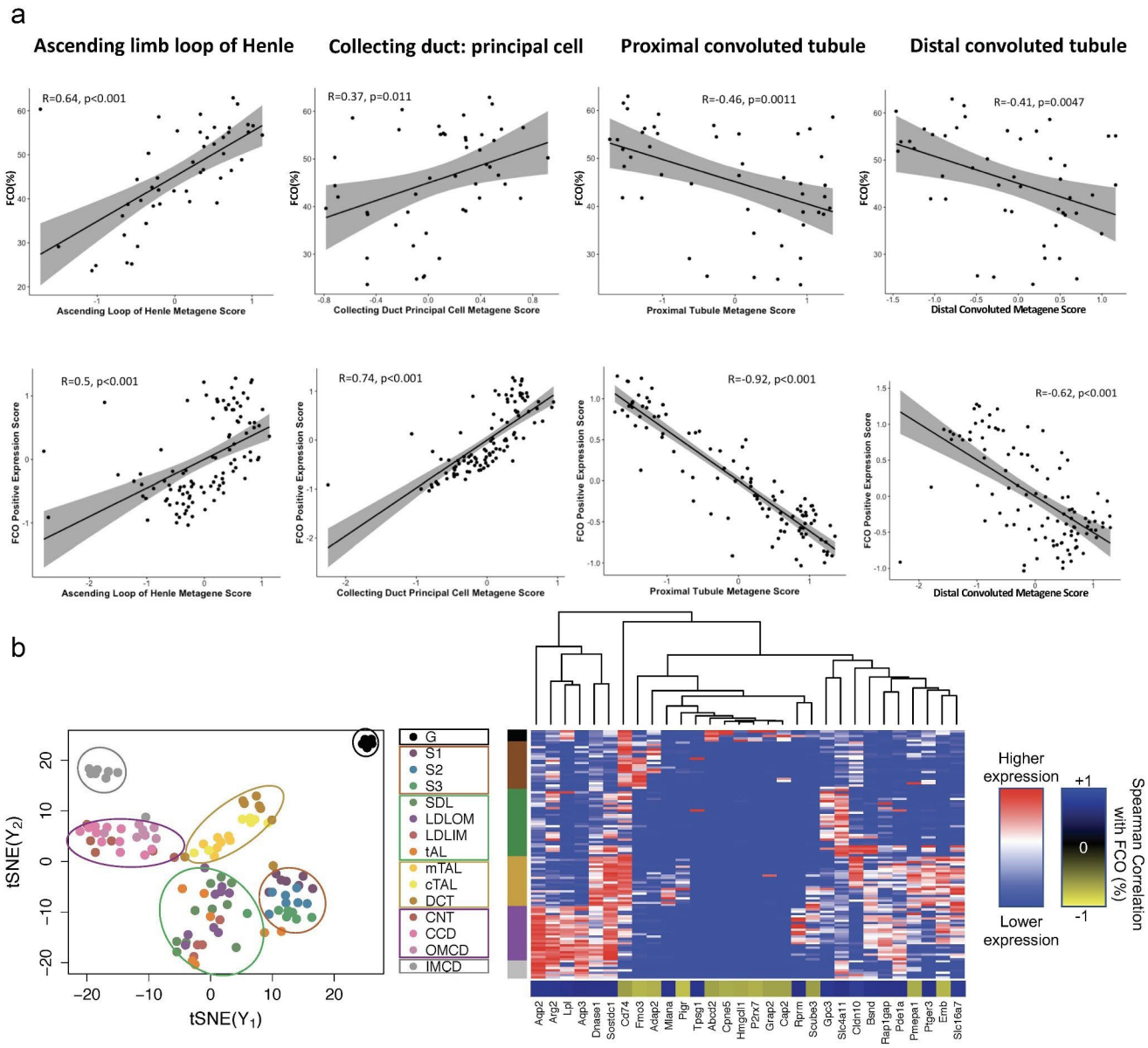
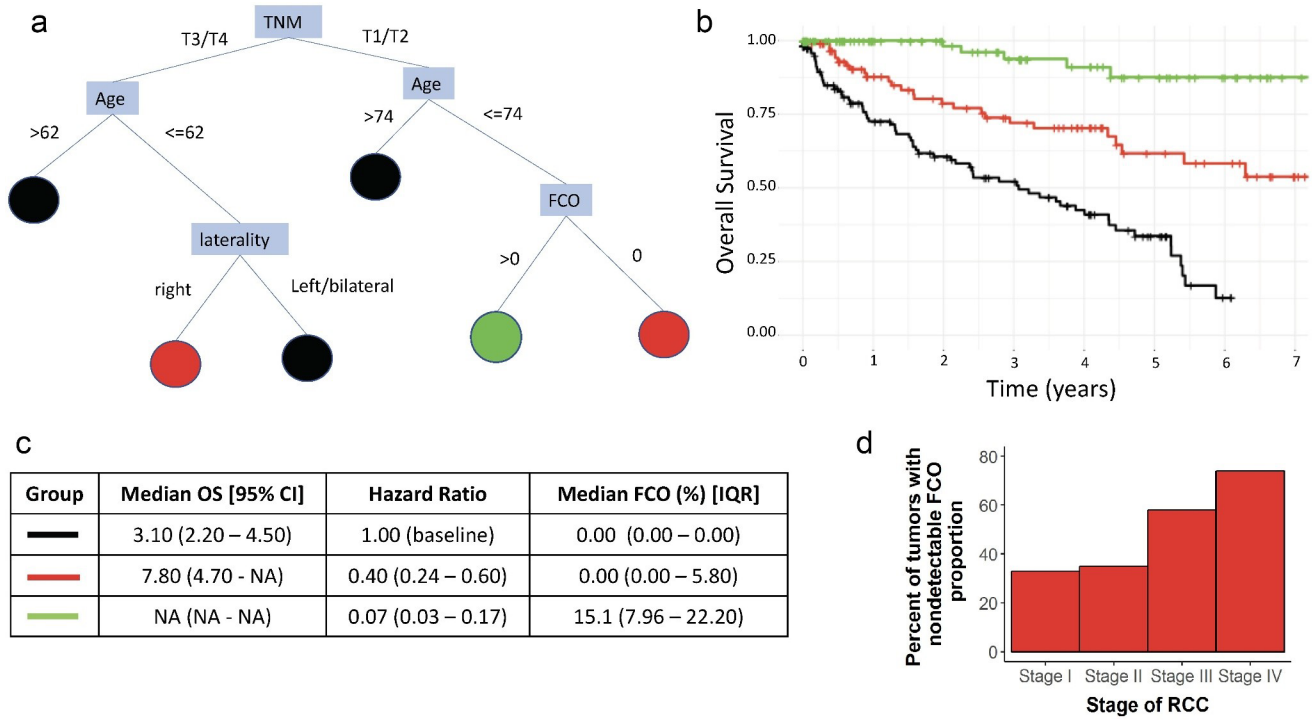


Figure 5. Corroboration of human FCO-related gene expression using orthologous renal genes in mouse (a) **and rat** (b). **Panel A**. The expression of human orthologs of mouse kidney cell gene signatures for ascending limb of the loop of Henle cells, and the collecting duct's principal cells are strongly correlated with FCO related gene expression. Expression of human orthologs of mouse proximal and distal convoluted cells are inversely correlated with FCO related gene expression (95% confidence interval indicated by grey shades). **Panel B, left**. t-SNE clustering using rat orthologs of human FCO related genes that discriminates six kidney regions in the rat. Human genes were chosen whose expression exhibited a Spearman correlation with FCO proportions of at least > |0.6|. **Panel B, right**. Hierarchical clustering using the top 30 influential genes in the t-SNE clusters showing association of medullary segments with genes positively correlated with FCO proportions whereas cortical nephron segments and glomerulus are influenced by gene expression negatively associated with FCO proportions.

Mouse kidney cell orthologs in human tissues aligned with the FCO proportion

We utilized mouse kidney single-cell RNA-Seq data to explore corroboration with our human data implicating medullar gene expression and FCO proportions [58]. We identified human orthologs of mouse renal cell-specific genes to evaluate whether the FCO proportion of human kidneys would align with results for proximal and distal tubule, the loop of Henle, and principal collecting duct cells (Figure 5a). The FCO proportion and metagene expression scores associated



Group	Median OS [95% CI]	Hazard Ratio	Median FCO (%) [IQR]
▬	3.10 (2.20 – 4.50)	1.00 (baseline)	0.00 (0.00 – 0.00)
▬	7.80 (4.70 - NA)	0.40 (0.24 – 0.60)	0.00 (0.00 – 5.80)
▬	NA (NA - NA)	0.07 (0.03 – 0.17)	15.1 (7.96 – 22.20)

Figure 6. Non-detectable FCO proportions within tumour tissue predict longer survival times in human clear cell renal cancer. Survival analysis was conducted using the data from TCGA. We applied a recursive partitioning algorithm, partDSA, in an effort to create homogeneous survival risk groups [88]. **Panel A**. partDSA survival tree for human renal clear cell carcinoma patients on TCGA. The models incorporated relevant patient variables, including patient age, gender, TMN class, tumour laterality, tumour FCO score. **Panel B**. Kaplan-Meier survival curves for the part DSA-identified subgroups. **Panel C**. Relevant survival statistics for the part DSA-identified subgroups. **Panel D**. Association of nondetectable FCO (FCO = 0) proportions with higher stage clear cell renal cancer tumours.

with FCO proportion were positively correlated with the expression of orthologs of ascending limb of the loop of Henle and the principal cell collecting duct, which are both medullar structures. By contrast, the FCO proportion and associated metagenes were both inversely related to the scores based on the orthologs of proximal and distal tubular cells located in the renal cortex. These data indicate that human orthologs of mouse renal cell-specific genes display features consistent with human kidney in pointing to FCO related expression within medullar anatomic regions.

Corroboration of renal medullary expression of FCO related genes using microdissected rat RNA-Seq nephron data

An orthogonal approach to our analysis of mouse expression data was undertaken to map 458 human FCO-related genes (Suppl Table 10) to orthologous genes within the extensive database of rat glomerulus and 16 nephron segments. Then, we used t-SNE to plot FCO associated genes to test whether human FCO related gene expression maps to specific regions of the rat nephron, which revealed six regional nephric clusters as

shown in Figure 5b. We then tested whether influential genes that clustered cortical and medullar regions of the kidney were associated with the FCO proportion. In Figure 5b, 30 of the most significant genes (Suppl Table 11) and their Spearman's correlation with the FCO proportion were aligned with renal anatomic position. Glomeruli and proximal tubular clusters were clustered by genes negatively associated with the FCO whereas collecting ducts and distal tubules were predominately clustered by genes positively associated with the FCO proportion. These results indicate that human genes associated with the FCO proportion are highly influential in the rat nephron as differentiating glomerulus and proximal tubules from distal tubule and collecting ducts, supporting the generalizability of the FCO as a strong correlative of medullar gene expression as compared to cortical gene expression.

FCO proportion is a prognostic factor in clear cell renal cancer

Recursive partitioning analysis was employed to investigate if FCO proportion was associated with survival in clear cell renal cancer using the TCGA database. Three distinct risk groups were identified. The high-risk group (n = 121; black group in Figure 6, median overall survival (OS) 3.1 years) included patients over 62 years old whose tumours were T3/T4 stage, or patients under 62 with left/ bilateral tumour location; or patients over 74 with T1/T2 stage tumours (median OS 3.1 years). The Intermediate risk group (n = 93; red group in Figure 6, median OS 7.8 years) included patients under 62 with T3/T4 right tumour location; or patients under 74 with T1/T2 and FCO equal to 0. The low-risk group (n = 103; green group in Figure 6, >75% surviving past 5 years) included patients under 74 with T1/T2 tumours and an FCO greater than 0 (median OS NA). These low- risk patients were at an 80% reduced risk of death (hazard ratio: 0.19, 95% CI (0.09, 0.41)) compared to those with similar age and stage whose FCO equal 0 (red group to right of green group, Figure 6). Of 319 TCGA tumours, 162 (50.8%) demonstrated non-detectable FCO levels that were most prevalent among T3/T4 stage tumours (Figure 6d). In patients under 74 with T1/T2 tumours, an FCO proportion >0 was associated with an 80% reduction in the risk of death.

Discussion

Widespread foetal methylation in adult kidney and age dependence of FCO proportions

The observed plasticity of adult renal epithelial cells prompted our studies of DMRs in embryonic and foetal genes. The kidney had the highest FCO proportions of any adult tissue examined, indicating a large portion of the kidney may have an ontogeny traceable to foetal cells. In comparison, leukocytes and most other solid organs in adults display non-detectable or low FCO proportions. Kidney samples from adult chronic kidney disease patients had lower FCO proportions than non- diseased samples, implicating clinical pathology as one factor influencing the FCO signature. Non- diseased and micro-dissected samples enriched in the renal cortex's proximal tubules displayed lower FCO scores compared to TCGA normal samples, which were all sampled without regard to anatomic depth. We interpret these latter observations as cell composition and renal pathology rather than donor age, explaining individual variation in FCO proportions after

childhood. Analysis of FCO methylation within individual kidney cells could directly test this hypothesis, but data quality and zero inflation limitations from single-cell methylation measures are currently prohibitive. In addition, comprehensive DNA methylation analysis of defined kidney segments does not exist in any species. We hypothesized that recent sequence-based gene expression studies could bridge the kidney structure and the FCO signature. Utilizing gene expression data from biopsies taken at different depths indicated that overall, 33% of individual variation in the adult FCO proportion might be explained by variations in the cell composition of kidney samples, which has critical implications for the study of kidney pathophysiology.

Gene expression correlated with FCO proportions in human kidney

Our genome-wide scans of nondissected TCGA normal samples revealed a dramatic negative association of the FCO proportion with immune response genes. Transcripts encoding lineage differentiation factors and effector molecules of adaptive and innate immune cells (e.g., B-cell antibody synthesis) were prominent among the genes most negatively correlated with FCO proportion. This result is consistent with the idea that immune cells in the kidney are of haematopoietic origin, which are derived from adult stem cells and do not contain foetal methylation marks, and consistent with gene expression data from non-dissected mouse kidney [58]. Moreover, inflammatory signatures were also inversely related to FCO proportions.

Focusing on gene expression positively correlated with the FCO proportion identified epithelial cell-derived transcripts prominent in regions of the distal nephron as compared with the glomerulus and proximal tubules, which are contained within the renal cortex. This conclusion is supported by three sets of empirical observations using human data. First, genes whose expression levels were positively correlated with the FCO proportion were enriched for gene ontology terms including loop of Henle and distal nephron development. Second, expression signatures based on these genes demonstrated strong associations with genes signalling medullar and cortico- medullary location in biopsy samples of different depth. Also, strong inverse correlations of FCO related gene expression were observed with metagenes of renal cortex and cortico-medullary depth biopsy. Finally, previously established co- expression modules of the distal nephron were most strongly associated with the expression of FCO related genes. As observed with the biopsy depth correlations, the co-expression modules that mapped to the glomerulus and proximal segments were inversely associated with FCO metagenes. FCO proportions and gene expression positively correlated with those proportions localized preferentially to medullary regions of the human kidney.

Mouse and rat expression data corroborate medulla associated DGE patterns

Though the FCO signature has only been studied in humans, we reasoned that postnatal maintenance of foetal epigenetic features is fundamental in mammalian renal biology. While extensive conservation in gene expression is observed between human and rodent kidney, there is reported inter-species divergence in quantitative gene expression, especially at the nephron substructure level [98,99]. To minimize potential bias in our analyses, we normalized

expression levels in each species and employed metagene signatures that averaged relative expression over groups of co- expressed genes to de-emphasize individual genes. Human orthologs of cell discriminating genes identified in the mouse kidney [58] had a linear relationship between their expression levels tested and FCO proportion in human samples for the loop of Henle and principal cells of the collecting ducts. In sharp contrast, we found strong inverse relationships of expression with FCO for orthologs of mouse proximal and distal convoluted cell discriminating genes, which are restricted to the renal cortex. Thus, mouse renal cell expression signatures based on single-cell analyses provide a picture consistent with human co-expression models and renal biopsy depth results. This observation strengthens the link between FCO proportions and DGE patterns within medulla in contrast to renal cortical regions.

We also applied an orthogonal approach to our mouse studies by mapping human FCO related genes to their corresponding segments in the rat nephron. We employed t-SNE and clustering of human FCO related genes that have expressed rat homologues and mapped their expression levels across the glomerulus and rat nephron tubular segments. The resulting clusters delineated six regions corresponding to glomeruli and portions of the proximal and distal tubules and collecting ducts. While the FCO related clusters did not differentiate the full substructure of the nephron, they did achieve broad regional patterning with good correspondence between influential members of each cluster to an anatomic region and the directionality of the input gene's correlation with the FCO proportion in humans. For example, medullary collecting duct segments were clustered by genes positively associated with the FCO proportion, whereas glomerular and proximal segments were clustered by genes negatively correlated with the human FCO proportion. We further note that clear discrimination of all nephron substructures has not been achieved even when using the full complement of expressed rat transcripts. Therefore, we judge that our results, using only 458 transcripts, are highly informative. The FCO proportion demonstrates biological relevance by ordering DGE patterns that contribute to regional transcriptional identity.

Repair hypothesis and renal stem cells

Considering the possible role of foetal epigenetic features leads to consideration of their potential relationship to stem cells. The role of stem cells in tissue homoeostasis and repair in the adult kidney has been extensively studied [7,16]. In contrast to tissues with high cell turnover such as the gut and bone marrow that have well-defined stem cell compartments, the kidney tubular epithelium, with its low rate of homoeostatic cell division, and highly segmented architecture, appears to have evolved a more distributive regenerative capacity. Early studies pointed to the extensive capacity of renal epithelial cells to dedifferentiate and redifferentiate during neoepithelialization after acute injuries [10]. Fully differentiated epithelia were shown to regenerate the proximal tubule without any contribution from a pre-existing intratubular stem cell population [100]. Anatomic-specific progenitor cell niches have been proposed for glomerulus, proximal tubules and medullary papillae [101]. In the mouse, developmental stage affected the distribution of putative stem cells across

medullar and cortical regions [102]. Response to renal artery occlusion in the medullary papilla indicated that repair is regionally specified [14]. The renal papilla itself, with its low oxygen content, has been proposed to be a good candidate for a stem cell niche environment [101]. In the adult, hypoxic cell niches provide favourable conditions for maintenance of stable long-lived dormant stem cells [103,104]. Hypoxia and associated glycolytic metabolism may play a protective role and provide signals to stem cells that help maintain them in a prolonged undifferentiated state [105,106]. Consistent with this, we observed that metagene expression signatures of both hypoxia and glycolysis were positively correlated with FCO proportions in normal adult kidney.

The epigenetic mechanisms underlying different epithelial maintenance and repair strategies are unclear, and we believe our results provide new insights into these processes. Following insult or damage, increased transcription of nephrogenic genes is well documented, and among these are FCO genes with specific foetal methylation features. Prior studies established the role of *PAX8, HOXB7*, and *FGF2* in the early stages of nephrogenesis and in response to injury in the adult organ. For each of these genes, we found DMRs exhibiting foetal patterns of unmethylated DNA in adult kidneys. Notably, these DMRs were correlated with differential gene expression levels. These observations suggest that the CpG methylation status of FCO-related genes plays a regulatory role in the kidney; they likely do not just act as biomarkers of cell ontogeny. In the normal kidney samples we studied, the differences in gene expression could be contributing to homoeostasis. The maintenance and regeneration of a highly segmented epithelium like the nephron with its distinct cell compartments may require an epigenetic programme that retains capacity to direct regionalization and segment specialization. This idea fits with the enrichment of regionalization, kidney epithelial differentiation, and segment specification ontology terms that we found among DGE patterns associated with FCO proportions. The widespread foetal methylation signature we report here may help programme the hypothesized multiple pools of progenitor cells of the kidney [14].

Renal cancer survival and the FCO proportion in tumour cells

Contrary to expectations, the FCO proportion of tumours was strongly associated with lower stage tumours and longer survival times. Renal cancer stem cells track with higher grade and stage tumours and shorter patient survival times. These observations point to a clear distinction of the cancer stem cell and foetal cell origin concepts with respect to the clinical behaviour of kidney cancer. The non-detectable FCO levels were most prevalent among T3/T4 stage tumours. In patients under 74 with T1/T2 tumours, an FCO proportion >0 was associated with a significant reduction in the risk of death. This suggests the FCO proportion may help discriminate against similarly staged patients. Although FCO negative tumours may have a different cell of origin, it is also possible that non-detectable FCO proportions represent a loss of foetal methylation features resulting from epigenomic instability. Preserving the methylation status of the block of CpG sites in the FCO gene set during proliferative expansion may require efficient mechanisms to prevent hypermethylation. It is not

surprising that malignant transformation could lead to a loss of the FCO signature due to disruption of the epigenetic machinery. A universal feature of solid tumours is that they display significantly lower FCO proportions than their respective normal tissues [32]. This may result in part from mutations or other acquired disruptions in epigenetic regulation, common in clear cell renal cancer and other tumours. However, we did not observe any associations of TCGA recorded mutations, specific DNA methylation signatures, or tumour purity estimates with the FCO scores of tumours. Therefore, we view the FCO proportion as a potentially unique tumour marker representing epigenome stability and improved renal cancer survival.

Limitations and conclusions

The current study is the first to suggest a functional role of conserved DNA methylation marks present in embryonic and foetal cells in differentiated adult cells' biology. However, our observations' correlative nature precludes us from definitively assigning any direct regulatory function to the DMRs within the FCO. Another limitation is that we do not have FCO measurements within isolated cells of the human kidney. This level of analysis is essentially prohibitive in any species. Given this limitation, our current view of a widespread epigenetic compartment in adult kidneys must await significant technical advances in renal epigenomics before we are able to move beyond correlative observations. Even though, we used the most comprehensive human samples available, they still constitute macrodissected biopsies from nephrectomy samples for renal cancer. Previous studies have indicated that gene expression from nephrectomy samples are not biased to normal kidney tissues [87]. However, more baseline data in the non- diseased kidney is needed. These limitations reinforce the importance of animal models that do provide high-resolution data. To corroborate our human data, we mapped FCO correlated gene expression with nephron segment-specific data in rodents. FCO-related genes positively correlated with medulla and distal tubular location in mouse and rat nephron were strongly correlated with medulla metagene signatures obtained from human biopsies. We acknowledge that divergence in substructure gene expression may limit the translation of these results across species.

Our study's broader implication is that the kidney appears to represent an archetypal example of a widespread niche of foetal derived cells in an adult epithelial tissue. Previous researchers have drawn attention to the renal medulla's hypoxic conditions and its resemblance to conditions in embryonic and foetal life. Such hypoxic regions, which support stem cell survival, occur throughout the body [107]. The development of computational tools and cell ontogeny methylation markers applicable to human tissues provide a path for future research to elucidate the role of conserved foetal epigenetic programs in tissue homoeostasis and regeneration.

Acknowledgments

Work was supported by the National Institutes of Health with grants R01CA207360, and P50CA097257 to J.K.W. and A.M.M., R01CA207110 to K.T.K., and R01DE022772 and R01CA216265 to B.C.C. L.A.S. is supported by the National
Institute of General Medical Science award P20GM1044168299 and the Congressionally Directed Medical Research Programs/Department of Defense W81XWH-20-1-0778. Support to J.K.W. and A.M.M. was also provided by the loglio Collective and to J.K.W.

by the Robert Magnin Newman Endowed Chair in Neurooncology. D.C.K. was supported by the National Cancer Institute (NCI) Cancer Center Support Grant P30 CA168524 and by the Kansas IDeA Network of Biomedical Research Excellence (K-INBRE) Bioinformatics Core, supported in part by the National Institute of General Medical Science award P20GM103418.

Author contributions

Conceptualization, J.K.W., Z.Z., K.T.K.; Methodology, J.K. W., Z.Z.; Software, Z.Z, DCK, L.A.S.

Formal Analysis, Z.Z, DCK, L.A.S.; Investigation, J.K.W., Z.Z., Z.Z, DCK, L.A.S., A.M.M., B.C.C., K.T.K.; Data Curation, Z.Z, DCK, L.A.S.; Writing – Original Draft, J.K. W., Z.Z.; Writing-Review & Editing, J.K.W., Z.Z., Z.Z, DCK, L.A.S., A.M.M., B.C.C., K.T.K.; Visualization, Z.Z, DCK; Supervision, J.K.W., K.T.K.; Project Administration, J.K.W., K.T.K.; Funding Acquisition, J.K.W.

Disclosure statement

J.K.W. and K.T.K. are cofounders of Cellintec.

Funding

This work was supported by the National Cancer Institute [P30 CA168524]; National Institutes of Health [P50CA097257]; National Institutes of Health [R01CA207360]; National Institutes of Health [R01DE022772]; National Institutes of Health [R01CA216265]; National Institutes of Health (US) [R01CA207110]; Kansas IDeA Network of Biomedical Research Excellence [P20GM103418].

Abbreviations

DGE: differential gene expression;
DMR: differentially methylated region;
FGF2: basic fibroblast growth factor 2;
FCO: foetal cell origin;
IHC: immunohistochemistry;
OS: overall survival;
PAX8: paired box transcription factor 8;
HOXB7: homeobox B7;
TCGA: Tumour Cell Genome Atlas;
GEO: Gene Expression Omnibus;
ESBL: Epithelial Systems Biology Laboratory.

ORCID

Devin C. Koestler http://orcid.org/0000-0002-0598-0146
Lucas A. Salas http://orcid.org/0000-0002-2279-4097
Brock C. Christensen http://orcid.org/0000-0003-3022-426X

References

[1] Schofield R. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. Blood Cells. 1978;4(1–2):7–25.
[2] Crane GM, Jeffery E, Morrison SJ. Adult haematopoietic stem cell niches. Nat Rev Immunol. 2017;17 (9):573–590.
[3] Tan DW, Barker N. Intestinal stem cells and their defining niche. Curr Top Dev Biol. 2014;107:77–107.
[4] Mohyeldin A, Garzon-Muvdi T, Quinones-Hinojosa A. Oxygen in stem cell biology: a critical component of the stem cell niche. Cell Stem Cell. 2010;7 (2):150–161.
[5] Witzgall R, Brown D, Schwarz C, et al. Localization of proliferating cell nuclear antigen, vimentin, c-Fos, and clusterin in the postischemic kidney. Evidence for a heterogenous genetic response among nephron segments, and a large pool of mitotically active and dedifferentiated cells. J Clin Invest. 1994;93(5):2175–2188.
[6] Bonventre JV. Dedifferentiation and proliferation of surviving epithelial cells in acute renal failure. J Am Soc Nephrol. 2003;14(Suppl 1):S55–61.
[7] Humphreys BD, Czerniak S, DiRocco DP, et al. Repair of injured proximal tubule does not involve specialized progenitors. Proc Natl Acad Sci U S A. 2011;108 (22):9226–9231.
[8] Angelotti ML, Ronconi E, Ballerini L, et al. Characterization of renal progenitors committed toward tubular lineage and their regenerative potential in renal tubular injury. Stem Cells. 2012;30:1714–1725.
[9] Villanueva S, Cespedes C, Vio CP. Ischemic acute renal failure induces the expression of a wide

range of nephrogenic proteins. Am J Physiol Regul Integr Comp Physiol. 2006;290(4):R861–870.

[10] Oliver JA, Maarouf O, Cheema FH, et al. The renal papilla is a niche for adult kidney stem cells. J Clin Invest. 2004;114(6):795–804

[11] Vogetseder A, Karadeniz A, Kaissling B, et al. Tubular cell proliferation in the healthy rat kidney. Histochem Cell Biol. 2005;124(2):97–104.

[12] Oliver JA. Adult renal stem cells and renal repair. Curr Opin Nephrol Hypertens. 2004;13(1):17–22.

[13] Oliver JA, Maarouf O, Cheema FH, et al. The renal papilla is a niche for adult kidney stem cells. J Clin Invest. 2004;114(6):795–804.

[14] Oliver JA, Sampogna RV, Jalal S, et al. A subpopulation of label-retaining cells of the kidney papilla regenerates injured kidney medullary tubules. Stem Cell Reports. 2016;6(5):757–771.

[15] Kramann R, Kusaba T, Humphreys BD. Who regenerates the kidney tubule? Nephrol Dial Transplant. 2015;30(6):903–910.

[16] Little MH, Kairath P. Does renal repair recapitulate kidney development? J Am Soc Nephrol. 2017;28(1):34–46.

[17] Yatabe Y, Tavare S, Shibata D. Investigating stem cells in human colon by using methylation patterns. Proc Natl Acad Sci U S A. 2001;98(19):10839–10844.

[18] Roost MS, Slieker RC, Bialecka M, et al. DNA methylation and transcriptional trajectories during human development and reprogramming of isogenic pluripotent stem cells. Nat Commun. 2017;8(1):908.

[19] Hiler D, Chen X, Hazen J, et al. Quantification of retinogenesis in 3D cultures reveals epigenetic memory and higher efficiency in iPSCs derived from rod photoreceptors. Cell Stem Cell. 2015;17(1):101–115.

[20] Hu Q, Friedrich AM, Johnson LV, et al. Memory in induced pluripotent stem cells: reprogrammed human retinal-pigmented epithelial cells show tendency for spontaneous redifferentiation. Stem Cells. 2010;28 (11):1981–1991.

[21] Hargus G, Ehrlich M, Arauzo-Bravo MJ, et al. Origin- dependent neural cell identities in differentiated human iPSCs in vitro and after transplantation into the mouse brain. Cell Rep. 2014;8(6):1697–1703.

[22] Bar-Nur O, Russ HA, Efrat S, et al. Epigenetic memory and preferential lineage-specific differentiation in induced pluripotent stem cells derived from human pancreatic islet beta cells. Cell Stem Cell. 2011;9 (1):17–23.

[23] Quattrocelli M, Swinnen M, Giacomazzi G, et al. Mesodermal iPSC-derived progenitor cells functionally regenerate cardiac and skeletal muscle. J Clin Invest. 2015;125(12):4463–4482.

[24] Kim K, Zhao R, Doi A, et al. Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. Nat Biotechnol. 2011;29(12):1117–1119.

[25] Vitaloni M, Pulecio J, Bilic J, et al. MicroRNAs contribute to induced pluripotent stem cell somatic donor memory. J Biol Chem. 2014;289(4):2084–2098.

[26] Polo JM, Liu S, Figueroa ME, et al. Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. Nat Biotechnol. 2010;28(8):848–855.

[27] Ohi Y, Qin H, Hong C, et al. Incomplete DNA methylation underlies a transcriptional memory of somatic cells in human iPS cells. Nat Cell Biol. 2011;13 (5):541–549.

[28] Salas LA, Wiencke JK, Koestler DC, et al. Tracing human stem cell lineage during development using DNA methylation. Genome Res. 2018;28 (9):1285–1295.

[29] Orkin SH, Zon LI. Hematopoiesis: an evolving paradigm for stem cell biology. Cell. 2008;132(4):631–644.

[30] Copley MR, Eaves CJ. Developmental changes in hematopoietic stem cell properties. Exp Mol Med. 2013;45(11):e55.

[31] Manesia JK, Franch M, Tabas-Madrid D, et al. Distinct molecular signature of murine fetal liver and adult hematopoietic stem cells identify novel regulators of hematopoietic stem cell function. Stem Cells Dev. 2017;26(8):573–584.

[32] Zhang Z, Wiencke JK, Koestler DC, et al. Absence of an embryonic stem cell DNA methylation signature in human cancer. BMC Cancer. 2019;19(1):711.

[33] McMahon AP. Development of the mammalian kidney. Curr Top Dev Biol. 2016;117:31–64.

[34] Lee JW, Chou CL, Knepper MA. Deep sequencing in microdissected renal tubules identifies nephron segment-specific transcriptomes. J Am Soc Nephrol. 2015;26(11):2669–2677.

[35] Higgins JP, Wang L, Kambham N, et al. Gene expression in the normal adult human kidney assessed by complementary DNA microarray. Mol Biol Cell. 2004;15(2):649–656.

[36] Papadopoulos T, Krochmal M, Cisek K, et al. Omics databases on kidney disease: where they can be found and how to benefit from them. Clin Kidney J. 2016;9 (3):343–352.

[37] Brezis M, Rosen S. Hypoxia of the renal medulla–its implications for disease. N Engl J Med. 1995;332 (10):647–655.

[38] Russell PS, Hong J, Windsor JA, et al. Renal lymphatics: anatomy, physiology, and clinical implications. Front Physiol. 2019;10:251.

[39] Neuhofer W, Beck FX. Survival in hostile environments: strategies of renal medullary cells. Physiology (Bethesda). 2006;21:171–180.

[40] Chen Y, Fry BC, Layton AT. Modeling glucose metabolism and lactate production in the kidney. Math Biosci. 2017;289:116–129.

[41] Jauniaux E, Gulbis B, Burton GJ. Physiological implications of the materno-fetal oxygen gradient in human early pregnancy. Reprod Biomed Online. 2003;7 (2):250–253.

[42] Dunwoodie SL. The role of hypoxia in development of the mammalian embryo. Dev Cell. 2009;17 (6):755–773.

[43] Patel SR, Ranghini E, Dressler GR. Mechanisms of gene activation and repression by pax proteins in the developing kidney. Pediatr Nephrol. 2014;29 (4):589–595.

[44] Kobayashi A, Valerius MT, Mugford JW, et al. Six2 defines and regulates a multipotent self-renewing nephron progenitor population throughout mammalian kidney development. Cell Stem Cell. 2008;3 (2):169–181.

[45] Humphreys BD, Valerius MT, Kobayashi A, et al. Intrinsic epithelial cells repair the kidney after injury. Cell Stem Cell. 2008;2(3):284–291.

[46] Sharma R, Sanchez-Ferras O, Bouchard M. Pax genes in renal development, disease and regeneration. Semin Cell Dev Biol. 2015;44:97–106.

[47] Alarcon P, Rodriguez-Seguel E, Fernandez-Gonzalez A, et al. A dual requirement for Iroquois genes during xenopus kidney development. Dev. 2008;135 (19):3197–3207.

[48] Massa F, Garbay S, Bouvier R, et al. Hepatocyte nuclear factor 1beta controls nephron tubular development. Dev. 2013;140(4):886–896.

[49] Heliot C, Desgrange A, Buisson I, et al. HNF1B controls proximal-intermediate nephron segment identity in vertebrates by regulating notch signalling components and Irx1/2. Dev. 2013;140(4):873–885.

[50] Patterson LT, Potter SS. Atlas of Hox gene expression in the developing kidney. Dev Dyn. 2004;229 (4):771–779.

[51] Pickering CM, Grady C, Medvar B, et al. Proteomic profiling of nuclear fractions from native renal inner medullary collecting duct cells. Physiol Genomics. 2016;48(2):154–166.

[52] Walker KA, Sims-Lucas S, Bates CM. Fibroblast growth factor receptor signaling in kidney and lower urinary tract development. Pediatr Nephrol. 2016;31 (6):885–895.

[53] Villanueva S, Cespedes C, Gonzalez AA, et al. Inhibition of bFGF-receptor type 2 increases kidney damage and suppresses nephrogenic protein expression after ischemic acute renal failure. Am J Physiol Regul Integr Comp Physiol. 2008;294:R819–828.

[54] Villanueva S, Cespedes C, Gonzalez A, et al. bFGF induces an earlier expression of nephrogenic proteins after ischemic acute renal failure. Am J Physiol Regul Integr Comp Physiol. 2006;291:R1677–1687.

[55] Villanueva S, Glavic A, Ruiz P, et al. Posteriorization by FGF, Wnt, and retinoic acid is required for neural crest induction. Dev Biol. 2002;241(2):289–301.

[56] Lindgren D, Eriksson P, Krawczyk K, et al. Cell-type- specific gene programs of the normal human nephron define kidney cancer subtypes. Cell Rep. 2017;20 (6):1476–1489.

[57] Lindgren D, Sjolund J, Axelson H. tracing renal cell carcinomas back to the nephron. Trends Cancer. 2018;4:472–484.

[58] Park J, Shrestha R, Qiu C, et al. Single-cell transcriptomics of the mouse kidney reveals

potential cellular targets of kidney disease. Sci. 2018;360(6390):758–763.

[59] Saroufim A, Messai Y, Hasmim M, et al. Tumoral CD105 is a novel independent prognostic marker for prognosis in clear-cell renal cell carcinoma. Br J Cancer. 2014;110(7):1778–1784.

[60] Mikami S, Mizuno R, Kosaka T, et al. Expression of TNF-alpha and CD44 is implicated in poor prognosis, cancer cell invasion, metastasis and resistance to the sunitinib treatment in clear cell renal cell carcinomas. Int J Cancer. 2015;136(7):1504–1514.

[61] Li X, Ma X, Chen L, et al. Prognostic value of CD44 expression in renal cell carcinoma: a systematic review and meta-analysis. Sci Rep. 2015;5(1):13157.

[62] Peired AJ, Sisti A, Romagnani P. Mesenchymal stem cell-based therapy for kidney disease: a review of clinical evidence. Stem Cells Int. 2016;2016:4798639.

[63] Farkas SA, Milutin-Gasperov N, Grce M, et al. Genome-wide DNA methylation assay reveals novel candidate biomarker genes in cervical cancer. Epigenet. 2013;8(11):1213–1225.

[64] Smith RG, Hannon E, De Jager PL, et al. Elevated DNA methylation across a 48-kb region spanning the
HOXA gene cluster is associated with Alzheimer's disease neuropathology. Alzheimers Dement. 2018;14(12):1580–1588.

[65] Legendre CR, Demeure MJ, Whitsett TG, et al. Pathway implications of aberrant global methylation in adrenocortical cancer. PLoS One. 2016;11(3): e0150629.

[66] Huang KK, Ramnarayanan K, Zhu F, et al. Genomic and epigenomic profiling of high-risk intestinal metaplasia reveals molecular determinants of progression to gastric cancer. Cancer Cell. 2018;33(1):137–150 e135.

[67] Price EM, Penaherrera MS, Portales-Casamar E, et al. Profiling placental and fetal DNA methylation in human neural tube defects. Epigenet Chromatin. 2016;9(1):6.

[68] Charlton J, Williams RD, Weeks M, et al. Methylome analysis identifies a wilms tumor epigenetic biomarker detectable in blood. Genome Biol. 2014;15(8):434.

[69] Ko YA, Mohtat D, Suzuki M, et al. Cytosine methylation changes in enhancer regions of core pro-fibrotic genes characterize kidney fibrosis development. Genome Biol. 2013;14(10):R108.

[70] Acevedo N, Reinius LE, Vitezic M, et al. Age-associated DNA methylation changes in immune genes, histone modifiers and chromatin remodeling factors within 5 years after birth in human blood leukocytes. Clin Epigenetics. 2015;7(1):34.

[71] Urdinguio RG, Torro MI, Bayon GF, et al. Longitudinal study of DNA methylation during the first 5 years of life. J Transl Med. 2016;14:160.

[72] Alisch RS, arwick BG, Chopra P, et al. Age-associated DNA methylation in pediatric populations. Genome Res. 2012;22(4):623–632.

[73] Hannum G, Guinney J, Zhao L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–367.

[74] Reinius LE, Acevedo N, Joerink M, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7(7):e41361.

[75] Tserel L, Kolde R, Limbach M, et al. Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. Sci Rep. 2015;5(1):13107.

[76] Heyn H, Li N, Ferreira HJ, et al. Distinct DNA methylomes of newborns and centenarians. Proc Natl Acad Sci U S A. 2012;109(26):10522–10527.

[77] Bakulski KM, Feinberg JI, Andrews SV, et al. DNA methylation of cord blood cell types: Applications for mixed cell birth studies. Epigenetics. 2016;11(5):354– 362

[78] Gervin K, Page CM, Aass HC, et al. Cell type specific DNA methylation in cord blood: A 450K-reference data set and cell count-based validation of estimated cell type composition. Epigenetics. 2016;11(9):690–698.

[79] Hart T, Komori HK, LaMere S, et al. Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics. 2013;14(1):778.

[80] Laurens Van Der Maaten GH. visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579-2605.

[81] Krijthe JH Rtsne: t-distributed stochastic neighbor embedding using a barnes-hut implementation.

(2015).

[82] Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinf. 2015;31:166–169.

[83] Bakulski KM, Feinberg JI, Andrews SV, et al. DNA methylation of cord blood cell types: applications for mixed cell birth studies. Epigenet. 2016;11(5):354–362.

[84] Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinf. 2012;13:86.

[85] Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523.

[86] Flechner SM, Kurian SM, Head SR, et al. Kidney transplant rejection and tissue injury by gene profiling of biopsies and peripheral blood lymphocytes. Am J Transplant. 2004;4(9):1475–1489.

[87] Rodwell GE, Sonu R, Zahn JM, et al. A transcriptional profile of aging in the human kidney. PLoS Biol. 2004;2(12):e427.

[88] Molinaro AM, Lostritto K, Van Der Laan M. partDSA: deletion/substitution/addition algorithm for partitioning the covariate space in prediction. Bioinf. 2010;26 (10):1357–1363.

[89] Lostritto K, Strawderman RL, Molinaro AM. A partitioning deletion/substitution/addition algorithm for creating survival risk groups. Biom. 2012;68 (4):1146–1156.

[90] Stewart BJ, Ferdinand JR, Young MD, et al. Spatiotemporal immune zonation of the human kidney. Sci. 2019;365(6460):1461–1466.

[91] Wegener AM, Letourneur F, Hoeveler A, et al. The T cell receptor/CD3 complex is composed of at least two autonomous transduction modules. Cell. 1992;68 (1):83–95.

[92] Hamilton KL, Devor DC. Basolateral membrane K+ channels in renal epithelial cells. Am J Physiol Renal Physiol. 2012;302(9):F1069–1081.

[93] Pulkkinen K, Murugan S, Vainio S. Wnt signaling in kidney development and disease. Organogenesis. 2008;4(2):55–59.

[94] Chen F. Genetic and developmental basis for urinary tract obstruction. Pediatr Nephrol. 2009;24(9):1621–1632.

[95] Vogels R, Charite J, De Graaff W, et al. Proximal cis-acting elements cooperate to set Hoxb-7 (Hox-2.3) expression boundaries in transgenic mice. Dev. 1993;118:71–82.

[96] Narlis M, Grote D, Gaitan Y, et al. Pax2 and Pax8 regulate branching morphogenesis and nephron differentiation in the developing kidney. J Am Soc Nephrol. 2007;18(4):1121–1129.

[97] Hamilton KL, Devor DC. Basolateral membrane K+ channels in renal epithelial cells. Am J Physiol Renal Physiol. 2012;302(9):F1069–F1081.

[98] Lindstrom NO, Tran T, Guo J, et al. Conserved and divergent molecular and anatomic features of human and mouse nephron patterning. J Am Soc Nephrol. 2018;29(3):825–840.

[99] Cheval L, Pierrat F, Rajerison R, et al. Of mice and men: divergence of gene expression patterns in kidney. PLoS One. 2012;7(10):e46876.

[100] Kusaba T, Lalli M, Kramann R, et al. Differentiated kidney epithelial cells repair injured proximal tubule. Proc Natl Acad Sci U S A. 2014;111(4):1527–1532.

[101] Huling J, Yoo JJ. Comparing adult renal stem cell identification, characterization and applications. J Biomed Sci. 2017;24(1):32.

[102] Rangarajan S, Sunil B, Fan C, et al. Distinct populations of label-retaining cells in the adult kidney are defined temporally and exhibit divergent regional distributions. Am J Physiol Renal Physiol. 2014;307 (11):F1274–1282.

[103] Nombela-Arrieta C, Pivarnik G, Winkel B, et al. Quantitative imaging of haematopoietic stem and progenitor cell localization and hypoxic status in the bone marrow microenvironment. Nat Cell Biol. 2013;15 (5):533–543.

[104] Ito K, Suda T. Metabolic requirements for the maintenance of self-renewing stem cells. Nat Rev Mol Cell Biol. 2014;15(4):243–256.

[105] Suda T, Takubo K, Semenza GL. Metabolic regulation of hematopoietic stem cells in the hypoxic niche. Cell Stem Cell. 2011;9(4):298–310.

[106] Morikawa T, Takubo K. Hypoxia regulates the hematopoietic stem cell niche. Pflugers Arch. 2016;468(1):13–22.

[107] Redondo PA, Pavlou M, Loizidou M, et al. Elements of the niche for adult stem cell expansion. J Tissue Eng. 2017;8:2041731417725464.

Zhao et al. Clinical Epigenetics (2021) 13:222
https://doi.org/10.1186/s13148-021-01214-2

Clinical Epigenetics

RESEARCH Open Access

Methylation-derived inflammatory measures and lung cancer risk and survival

Naisi Zhao[1], Mengyuan Ruan[1], Devin C. Koestler[2,3], Jiayun Lu[4], Lucas A. Salas[6], Karl T. Kelsey[7,8], Elizabeth A. Platz[4,5] and Dominique S. Michaud[1,7*]

Abstract

Background: Examining immunity-related DNA methylation alterations in blood could help elucidate the role of the immune response in lung cancer etiology and aid in discovering factors that are key to lung cancer development and progression. In a nested, matched case–control study, we estimated methylation-derived NLR (mdNLR) and quantified DNA methylation levels at loci previously linked with circulating concentrations of C-reactive protein (CRP). We examined associations between these measures and lung cancer risk and survival.

Results: Using conditional logistic regression and further adjusting for BMI, batch effects, and a smoking-based methylation score, we observed a 47% increased risk of non-small cell lung cancer (NSCLC) for one standard deviation (SD) increase in mdNLR (*n* = 150 pairs; OR: 1.47, 95% CI 1.08, 2.02). Using a similar model, the estimated CRP Scores were inversely associated with risk of NSCLC (e.g., Score 1 OR: 0.57, 95% CI: 0.40, 0.81). Using Cox proportional hazards models adjusting for age, sex, smoking status, methylation-predicted pack-years, BMI, batch effect, and stage, we observed a 28% increased risk of dying from lung cancer (*n* = 145 deaths in 205 cases; HR: 1.28, 95% CI: 1.09, 1.50) for one SD increase in mdNLR.

Conclusions: Our study demonstrates that immunity status measured with DNA methylation markers is associated with lung cancer a decade or more prior to cancer diagnosis. A better understanding of immunity-associated methylation-based biomarkers in lung cancer development could provide insight into critical pathways.

Keywords: Lung cancer, DNA methylation, Methylation-based inflammation measures, C-reactive protein, mdNLR

Background

Lung cancer is the leading cause of cancer death in the USA, projected to account for 21.7% of all cancer deaths in 2021 [1]. A large percentage of lung cancer patients are diagnosed at an advanced stage [2] and five-year relative survival rates for those patients are between 3 and 6% [3]. Thus, early detection remains a key strategy to improve survival. However, the currently recommended strategy for lung cancer screening—low-dose computed tomography (LDCT) for persons 50 to 80 years old with at least a 20 pack-year smoking history and currently smoke or have quit within the past 15 years—is expensive and has a high false positive rate [4, 5]. Modifying the current lung cancer screening strategy by performing risk stratification could help prioritize LDCT screening and optimize secondary prevention. We propose that immune system markers could be incorporated into such risk stratification tools to help identify persons at higher risk of lung cancer to target for screening.

While smoking is the most important risk factor for lung cancer in the population, there is growing evidence that the immune system, in response to or independent of smoking, plays an important role in lung cancer development, acting potentially through the genesis of

*Correspondence: Dominique.Michaud@tufts.edu
[1] Department of Public Health and Community Medicine, Tufts University School of Medicine, Tufts University, 136 Harrison Avenue, Boston, MA 02111, USA
Full list of author information is available at the end of the article

© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://

creat iveco mmons. org/ licen ses/ by/4. 0/. The Creative Commons Public Domain Dedication waiver (http:// creat iveco mmons. org/ publi cdoma in/ zero/1. 0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

chronic inflammation [6]. For instance, an aggregated genome-wide association studies (GWASs) analysis of lung cancer risk found a direct causal effect of BMI on small cell lung cancer and an inverse effect on lung adenocarcinoma, suggesting the complexity of the role BMI and chronic inflammation plays in lung cancer subtypes [7]. Furthermore, it is plausible that inflammatory profiles prior to lung cancer diagnosis are associated with lung cancer-specific survival. Markers of systemic inflammation, including elevated levels of C-reactive protein (CRP) and the peripheral blood neutrophil-to-lymphocyte ratio (NLR), also have been identified as robust markers of cancer-associated inflammation [8, 9]. Elevated CRP levels [8], elevated serum levels of pro-inflammatory cytokines [10–12], increased neutrophil counts and decreased lymphocyte counts [13, 14], and polymorphisms in inflammation-related genes [15–18] have been associated with increased lung cancer risk. These inflammatory measures have also been associated with poor survival of lung cancer patients in several retrospective and a few prospective studies [19–21]. In addition, both experimental and epidemiologic studies support a role for chronic inflammation as a hallmark of cancer development and progression [8, 22–25]. We posit that a better understanding of the role of inflammation in lung cancer etiology could be gained by examining DNA methylation alterations in blood that are associated with the systemic immune response.

In the current study, we first predicted peripheral blood leukocyte composition and a neutrophil to lymphocyte index using validated DNA methylation markers (mdNLR), then quantified DNA methylation levels at loci previously linked with circulating concentrations of CRP, and calculated methylation-derived immune cell ratios by using an expanded deconvolution library. We evaluated the associations of these potential markers with lung cancer risk and lung cancer-specific survival. To address this question, we used pre-diagnostic blood samples of cases and controls obtained from the CLUE I/II cohorts. Our analyses controlled for self-reported smoking and methylation-predicted cumulative smoking in order to better focus our examinations on the DNA methylation marks that are informative of the immune response profile [26].

Results

Population characteristics

Characteristics of the 208 lung cancer cases and their 208 matched controls included in this analysis are presented in Table 1. Over 99% of the majority of participants were White. The median time between blood draw and lung cancer diagnosis was 14 years. The median age at blood draw in 1989 was 59 and 57 years in cases and controls, respectively. Overall, 55% of cases and controls were women and 11% were never smokers (Table 1).

Methylation-derived mdNLR index, leukocyte proportions, and lung cancer risk

We observed a 47% increased risk of non-small cell lung cancer (NSCLC) for one standard deviation increase in mdNLR (*n* = 150 pairs; OR: 1.47 [1.08, 2.02]). However, higher mdNLR values were not statistically associated with overall risk of lung cancer in our study. This association was comparable for NSCLC cases diagnosed within 10 years and beyond 10 years after blood draw. No stable associations could be estimated for small cell lung cancer (SCLC). After multiple comparison adjustments, monocyte/lymphocyte ratio showed a borderline significant 65% increased risk of NSCLC for each standard deviation increase (*n* = 150 pairs; OR: 1.65, adjusted CI: [0.99, 2.76]). In addition, immune cell ratios for CD4/CD8, NLR, B cell/lymphocyte, T cell/lymphocyte, Neu + Mono/lymphocyte, Eos/lymphocyte, CD4nv/ lymphocyte, B cell/CD8, CD8/Treg, Bnv/Bmem, CD4nv/ CD4mem, CD8nv/CD8mem, and Treg > 0 vs. Treg = 0 were not statistically significantly associated with lung cancer risk overall or by histologic types (Table 2).

Methylation-derived CRP scores and lung cancer risk CRP Score 1 was built using 54 CpG sites that were previously associated with inflammatory markers, while CRP Score 2 and 3 were each built with a subset of these 54 CpGs that were putative cell-specific or cell type invariant, respectively. Using data from a previously published pancreatic cancer dataset [27], all three scores were moderately correlated with log CRP and log IL-6 levels (Table 3). In this nested case–control study, we found all three CRP Scores inversely associated with risk of NSCLC after additionally adjusting for methylation-predicted pack-years (*n* = 150 pairs; Score 1 OR: 0.57 [0.40, 0.81]; Score 2 OR: 0.62 [0.45, 0.84]; Score 3 OR: 0.65 [0.44, 0.95]). We also found statistically significant

Table 1 Baseline characteristics of lung cancer cases and matched controls nested in CLUE I/II

	Controls	Cases		
		All lung cancers	Non-small cell lung cancer	Small cell lung cancer
N	208	208	150	29
Median age (range, years)	56 [28, 81]	59 [30, 83]	58 [36, 83]	53 [30, 71]
Median time before diagnosis (range, years)		14 [0, 29]	14 [0, 28]	12 [2, 29]
Sex				
Male	95 (45.7%)	95 (45.7%)	61 (40.7%)	16 (55.2%)
Female	113 (54.3%)	113 (54.3%)	89 (59.3%)	13 (44.8%)
Race				
White	208 (100%)	205 (98.6%)	149 (99.3%)	28 (96.6%)
Black	0 (0%)	3 (1.4%)	1 (0.7%)	1 (3.4%)
Cigarette smoking status				
Never smoker	22 (10.6%)	22 (10.6%)	15 (10.0%)	1 (3.4%)
Ever smoker	80 (38.5%)	80 (38.5%)	56 (37.3%)	10 (34.5%)
Current smoker	106 (51.0%)	106 (51.0%)	79 (52.7%)	18 (62.1%)
Current	4 (1.9%)	4 (1.9%)	3 (2.0%)	0 (0%)
Median BMI, kg/m^2 (range)	25.8 [18.0, 40.9]	25.6 [15.5, 42.6]	25.4 [15.5, 42.6]	25.6 [18.3, 34.7]
mdNLR mean (SD)	1.74 (0.976)	1.86 (1.32)	1.85 (1.19)	1.65 (1.07)
mdNLR median (range)	1.58 [0.13, 6.42]	1.59 [0.28, 11.90]	1.56 [0.28, 9.45]	1.44 [0.39, 4.81]
CRP score 1 mean (SD)	0.02 (1.01)	− 0.02 (0.99)	− 0.04 (0.97)	0.47 (1.08)
CRP score 1 median (range)	0.06 [− 2.42, 2.99]	− 0.03 [− 3.26, 2.94]	− 0.03 [− 3.26, 2.94]	0.41 [− 1.93, 2.16]
CRP score 2 mean (SD)	0.04 (1.03)	− 0.04 (0.98)	− 0.06 (0.95)	0.36 (1.17)
CRP score 2 median (range)	0.01 [− 3.35, 3.93]	− 0.10 [− 2.68, 3.10]	− 0.11 [− 2.68, 3.10]	0.24 [− 2.34, 3.10]
CRP score 3 mean (SD)	− 0.06 (0.98)	0.06 (1.02)	0.09 (1.03)	0.32 (0.97)
CRP score 3 median (range)	− 0.09 [− 2.43, 2.51]	0.08 [− 2.34, 2.95]	0.10 [− 2.34, 2.95]	0.38 [− 2.19, 1.97]
Median cigarette smoking Intensity (range, cig/day)	20 [0, 80]	20 [0, 80]	20 [0, 70]	20 [0, 80]
Cigar or pipe smoking				
Never	176 (84.6%)	178 (85.6%)	132 (88.0%)	25 (86.2%)
Ever	28 (13.5%)	26 (12.5%)	15 (10.0%)	4 (13.8%)

inverse association between CRP Score 1 and risk of NSCLC among cases diagnosed within 10 years and beyond 10 years, and between CRP Score 2 for NSCLC cases diagnosed within 10 years of blood draw (Table 4). CRP Scores 1, 2, and 3 were not associated with lung cancer risk when taking into account the matching factors and only adjusting for BMI and four surrogate variables for batch effects (*n* = 208 pairs; Score 1 OR: 0.96 [0.77, 1.21]; Score 2 OR: 0.89 [0.71, 1.11]; Score 3 OR: 1.11 [0.89, 1.40]). However, when additionally adjusting for methylation-predicted pack-years, inverse associations with total lung cancer risk were observed for Score 1 (OR: 0.76 [0.59, 0.99]) and Score 2 (OR: 0.77 [0.61, 0.98]). We also observed a 33% decreased risk of lung cancer for one standard deviation increase in CRP Score 1 (OR: 0.67 [0.47, 0.97]) among those with time to diagnosis over 10 years.

Survival analysis

We examined whether the mdNLR, methylation-derived immune cell ratios, and CRP Scores were associated with risk of dying of lung cancer among lung cancer cases (Table 5, Fig. 1).

We observed a 47% increased risk of dying for one standard deviation of mdNLR for NSCLC cases (*n* = 149 cases; HR: 1.47 [1.20, 1.81]). Among the NSCLC cases whose mdNLR was from < = 10 years before their

diagnosis, we found a 73% increased risk of dying for a one standard deviation increase in mdNLR (HR: 1.73 [1.19, 2.51]). In comparison, the risk of dying for a one standard deviation increase in mdNLR was lower among the NSCLC cases whose mdNLR was from 10 to 25 years prior to diagnosis (HR: 1.39 [1.05, 1.85]). Lastly, we observed a 28% increased risk of dying from lung cancer for one standard deviation increase in mdNLR (n = 205 cases deleted 3 cases with personyear = 0 or > 25 years; HR: 1.28 [1.09, 1.50]).

Immune cell ratios for CD4/CD8, NLR, B cell/lymphocyte, T cell/lymphocyte, Mono/lymphocyte, Eos/ lymphocyte, CD4nv/lymphocyte, B cell/CD8, CD8/ Treg, Bnv/Bmem, CD4nv/CD4mem, CD8nv/CD8mem, and Treg (> 0 vs = 0) were not associated with lung cancer-specific death, except for a 48% increased risk for one standard deviation increase in Neu + Mono/ lymphocyte ratio among the NSCLC cases (HR: 1.48 [1.04, 2.11]) and a borderline significant 29% increased risk of dying from lung cancer (HR: 1.29, adjusted CI: [1.00, 1.67]) for one standard deviation increase in Neu + Mono/lymphocyte ratio after multiple comparison adjustments. Furthermore, the three CRP Scores were not associated with lung cancer-specific death.

Table 2 Association between methylation-predicted immune cell profiles and risk of total lung cancer and NSCLC risk, overall and stratified by time to diagnosis, case–control study nested in the CLUE I/II cohort

	All lung cancers[a] OR (95% CI)	**Non-small cell lung cancer[b] OR (95% CI)**
mdNLR[c]	1.11 (0.89, 1.38)	**1.47 (1.08, 2.02)**
Time to diagnosis ≤ 10 years	0.97 (0.69, 1.35)	1.41 (0.90, 2.20)
Time to diagnosis > 10 years	1.22 (0.91, 1.64)	1.57 (0.98, 2.49)
CD4/CD8 Ratio[c]	1.17 (0.80, 1.72)	1.16 (0.74, 1.82)
Time to diagnosis ≤ 10 years	0.94 (0.56, 1.57)	0.79 (0.38, 1.66)
Time to diagnosis > 10 years	1.47 (0.82, 2.62)	1.72 (0.77, 3.85)
B cell/lymphocyte r atio[c]	1.07 (0.73, 1.57)	0.94 (0.58, 1.52)
Time to diagnosis ≤ 10 years	1.16 (0.63, 2.14)	0.91 (0.42, 1.97)
Time to diagnosis > 10 years	0.98 (0.60, 1.59)	0.93 (0.50, 1.71)
T cell/lymphocyte ratio[c]	1.06 (0.72, 1.56)	0.94 (0.60, 1.48)
Time to diagnosis ≤ 10 years	1.11 (0.62, 1.98)	1.03 (0.49, 2.16)
Time to diagnosis > 10 years	1.05 (0.63, 1.76)	0.88 (0.48, 1.62)
Monocyte/lymphocyte ratio[c]	1.15 (0.78, 1.69)	1.65 (0.99, 2.76)
Time to diagnosis ≤ 10 years	1.18 (0.66, 2.11)	1.62 (0.75, 3.51)
Time to diagnosis > 10 years	1.16 (0.69, 1.94)	1.81 (0.86, 3.80)
(Neutrophil + monocyte)/lymphocyte ratio[c]	1.13 (0.79, 1.61)	1.46 (0.87, 2.44)
Time to diagnosis ≤ 10 years	0.96 (0.59, 1.56)	1.32 (0.67, 2.60)
Time to diagnosis > 10 years	1.26 (0.78, 2.04)	1.62 (0.80, 3.29)
Eosinophil/lymphocyte ratio[c]	0.94 (0.64, 1.38)	0.89 (0.53, 1.49)
Time to diagnosis ≤ 10 years	0.81 (0.44, 1.49)	1.21 (0.48, 3.08)
Time to diagnosis > 10 years	1.01 (0.64, 1.59)	0.73 (0.36, 1.48)
CD4 naïve/lymphocyte ratio[c]	1.22 (0.83, 1.80)	1.36 (0.84, 2.20)
Time to diagnosis ≤ 10 years	1.47 (0.72, 2.99)	1.70 (0.74, 3.93)
Time to diagnosis > 10 years	1.11 (0.68, 1.80)	1.19 (0.67, 2.12)
B cell/CD8 ratio[c]	0.97 (0.68, 1.38)	1.00 (0.66, 1.52)
Time to diagnosis ≤ 10 years	1.06 (0.61, 1.83)	1.11 (0.58, 2.11)
Time to diagnosis > 10 years	0.85 (0.45, 1.62)	0.90 (0.44, 1.83)
CD8/treg ratio[c]	0.81 (0.55, 1.19)	0.73 (0.45, 1.18)
Time to diagnosis ≤ 10 years	0.86 (0.47, 1.59)	0.72 (0.33, 1.56)
Time to diagnosis > 10 years	0.78 (0.45, 1.35)	0.68 (0.33, 1.38)
Treg > 0 versus Treg = 0[c]	1.05 (0.74, 1.50)	1.07 (0.70, 1.63)
Time to diagnosis ≤ 10 years	0.95 (0.53, 1.70)	1.12 (0.55, 2.27)
Time to diagnosis > 10 years	1.10 (0.70, 1.73)	1.06 (0.61, 1.83)
B cell naïve/B cell memory r atio[c]	0.85 (0.58, 1.25)	0.87 (0.57, 1.32)
Time to diagnosis ≤ 10 years	0.65 (0.24, 1.76)	0.65 (0.21, 2.01)

Time to diagnosis > 10 years	0.93 (0.56, 1.56)	0.96 (0.56, 1.66)
CD4 naïve/CD4 memory r atio[c]	1.04 (0.73, 1.48)	1.59 (0.81, 3.13)
Time to diagnosis ≤ 10 years	0.98 (0.59, 1.64)	2.67 (0.65, 11.01)
Time to diagnosis > 10 years	1.18 (0.60, 2.32)	1.34 (0.58, 3.10)
CD8 naïve/CD8 memory r atio[c]	0.99 (0.72, 1.37)	1.01 (0.53, 1.92)
Time to diagnosis ≤ 10 years	0.66 (0.15, 2.81)	0.74 (0.17, 3.15)
Time to diagnosis > 10 years	1.04 (0.71, 1.53)	1.73 (0.36, 8.38)

Bold OR and CI values indicate statistical significance. Underline OR and CI values indicate borderline statistical significance after multiple comparison adjustment [a] 208 cases and 208 controls matched on age at blood draw, sex, smoking status, and model further adjusting for BMI, surrogate variables for batch effects, and a methylation-predicted pack-years smoked

[b] 150 cases and 150 controls matched on age at blood draw, sex, smoking status, and model further adjusting for BMI, surrogate variables for batch effects, and a methylation-predicted pack-years smoked

[c] OR results reported per 1 unit SD increase (mdNLR SD: 1.16; CD4/CD8 ratio SD: 2.07; BL ratio SD: 0.0526; TL ratio SD: 0.0802; ML ratio SD: 0.118; NM/L ratio SD: 1.22; EL ratio SD: 0.0722; CD4nv/L ratio SD: 0.0915; B/CD8 ratio SD: 3.07; CD8/Treg ratio SD: 115; Bnv/Bmem ratio SD: 109; CD4nv/CD4mem ratio SD: 0.429; CD8nv/CD8mem ratio SD: 2.05) and OR results reported for Treg > 0 versus 0; Bonferroni-adjusted CI reported for all except mdNLR (a prior hypothesis)

Table 3 Correlations between methylation-based CRP scores and circulating log-CRP level, log-IL6 level, peripheral blood leukocyte types, BMI, and smoking score residual among controls only

	Score 1[b] Spearman (*p* value)	**Score 2[c] Spearman (*p* value)**	**Score 3[d] Spearman (*p* value)**
B cell	0.103 (1.38e−01)	0.069 (3.23e−01)	− 0.027 (6.99e−01)
Neutrophils − 0.467 (1.19e−12) Monocytes	− 0.292 (1.92e−05)	− 0.422 (2.24e−10)	− 0.264 (1.17e−04)
		− 0.330 (1.17e−06)	− 0.288 (2.38e−05)
mdNLR	− 0.488 (7.66e−14)	− 0.449 (1.01e− 11)	− 0.281 (3.96e− 05)
BMI	0.025 (7.22e−01)	0.070 (3.12e−01)	− 0.063 (3.68e−01)
	−13)	0.375 (2.43e−08)	0.632 (1.42e−24)
Pancreatic cancer study (controls only)			
Log-CRP l evel[a]	0.284 (4.27e−06)	0.247 (6.96e−05)	0.297 (1.56e−06)
Log-IL6 level[a]	0.213 (7.53e−04)	0.158 (1.32e−02)	0.203 (1.33e−03)
CLUE I/II (controls only)			
CD4T	0.425 (1.65e−10)	0.359 (1.00e−07)	0.212 (2.13e−03)
CD8T	0.504 (8.78e−15)	0.461 (2.34e−12)	0.390 (5.74e−09)
NK	0.062 (3.70e−01)	0.175 (1.17e−02)	0.026 (7.09e−01)
Methylation-predicted pack-years residual[e]	0.475 (4.46e		

[a] Correlations with log-CRP level and log-IL6 level were tested with a pancreatic cancer dataset[43] [b] CpG Score 1 is built using 54 CpG sites [c] CpG Score 2 is built using the top 10 highly cell-specific CpG sites [d] CpG Score 3 is built using the 10 modestly cell-specific CpG sites [e] We calculated a pack-years methylation score to represent pack-years smoked associated methylation alterations. This score correlates with gene expression changes that are affected by smoking

Table 4 Association between methylation-based CRP scores and risk of total lung cancer and NSCLC risk, overall and stratified by time to diagnosis, CLUE I/II cohort

	All lung c ancers[a] OR (95% CI)	**Non-small cell lung cancer[b] OR (95% CI)**
CRP Score 1 [c]	**0.76 (0.59, 0.99)**	**0.57 (0.40, 0.81)**
Time to diagnosis ≤ 10 years	0.84 (0.56, 1.26)	**0.57 (0.33, 0.97)**
Time to diagnosis > 10 years	**0.67 (0.47, 0.97)**	**0.53 (0.32, 0.88)**
CRP Score 2 [d]	**0.77 (0.61, 0.98)**	**0.62 (0.45, 0.84)**
Time to diagnosis ≤ 10 years	0.79 (0.54, 1.17)	**0.49 (0.28, 0.85)**
Time to diagnosis > 10 years	0.73 (0.53, 1.02)	0.68 (0.46, 1.00)
CRP Score 3 [e]	0.79 (0.58, 1.07)	**0.65 (0.44, 0.95)**

Time to diagnosis ≤ 10 years	0.87 (0.53, 1.42)	0.64 (0.35, 1.18)
Time to diagnosis > 10 years	0.73 (0.49, 1.09)	0.62 (0.36, 1.05)

Bold OR and CI values indicate statistical significance

[a] 208 cases and 208 controls matched for age at blood draw, sex, smoking status, and model further adjusting for BMI, surrogate variables for batch effects, and a methylation-predicted pack-years smoked [b] 150 cases and 150 controls matched for age at blood draw, sex, smoking status, and model further adjusting for BMI, surrogate variables for batch effects, and a methylation-predicted pack-years smoked. The matched pairs were kept within each stratum [c] CpG Score 1 is built using 54 CpG sites and OR results reported per 1 unit of SD of CpG Score 1 (SD 1) [d] CpG Score 2 is built using the top 10 highly cell-specific CpG sites and OR results reported per 1 unit of SD of CpG Score 2 (SD 1) [e] CpG Score 3 is built using the 10 modestly cell-specific CpG sites and OR results reported per 1 unit of SD of CpG Score 3 (SD 1)

Discussion

Our study prospectively assessed predicted immune cell profiles using DNA methylation markers and examined associations between previously identified DNA methylation markers of inflammation and lung cancer risk and lung cancer survival. Using pre-diagnostic blood samples of lung cancer cases and controls who participated in the CLUE I/II cohorts [23], pre-diagnosis mdNLR was associated with increased risk of NSCLC, and among cases, with total lung cancer and NSCLC lung cancer-specific death. In

Table 5 Association between immune cell ratios and methylation-based CRP scores and lung cancer-specific mortality among lung cancer cases, CLUE I/II cohort

	All lung cancers[a] HR (95% CI)	**Non-small cell lung cancer[b] HR (95% CI)**
mdNLR[c]	**1.28 (1.09, 1.50)**	**1.47 (1.20, 1.81)**
Time to diagnosis ≤ 10 years	**1.34 (1.01, 1.76)**	**1.73 (1.19, 2.51)**
Time to diagnosis > 10 years	1.20 (0.98, 1.48)	**1.39 (1.05, 1.85)**
CD4/CD8 ratio[c]	1.02 (0.76, 1.36)	1.07 (0.78, 1.48)
Time to diagnosis ≤ 10 years	1.28 (0.79, 2.07)	1.29 (0.70, 2.38)
Time to diagnosis > 10 years	0.96 (0.65, 1.41)	1.03 (0.66, 1.62)
B cell/lymphocyte ratio[c]	0.97 (0.73, 1.30)	1.06 (0.74, 1.51)
Time to diagnosis ≤ 10 years	0.87 (0.50, 1.50)	0.99 (0.52, 1.89)
Time to diagnosis > 10 years	1.11 (0.75, 1.63)	1.18 (0.75, 1.85)
T cell/lymphocyte ratio[c]	1.01 (0.78, 1.31)	0.95 (0.69, 1.31)
Time to diagnosis ≤ 10 years	1.01 (0.62, 1.64)	0.94 (0.53, 1.68)
Time to diagnosis > 10 years	1.05 (0.69, 1.60)	0.91 (0.54, 1.52)
Monocyte/Lymphocyte Ratio[c]	1.16 (0.87, 1.55)	1.21 (0.85, 1.72)
Time to diagnosis ≤ 10 years	1.45 (0.87, 2.43)	1.52 (0.80, 2.89)
Time to diagnosis > 10 years	1.25 (0.80, 1.96)	1.54 (0.84, 2.84)
(Neutrophil + monocyte)/lymphocyte ratio[c]	1.29 (1.00, 1.67)	**1.48 (1.04, 2.11)**
Time to diagnosis ≤ 10 years	1.36 (0.87, 2.13)	1.76 (0.95, 3.24)
Time to diagnosis > 10 years	1.22 (0.86, 1.74)	1.41 (0.87, 2.29)
Eosinophil/lymphocyte ratio[c]	1.01 (0.76, 1.35)	0.86 (0.60, 1.23)
Time to diagnosis ≤ 10 years	1.55 (0.93, 2.59)	1.40 (0.74, 2.67)
Time to diagnosis > 10 years	0.98 (0.67, 1.44)	0.77 (0.49, 1.21)
CD4 naïve/lymphocyte ratio[c]	0.92 (0.67, 1.27)	0.95 (0.65, 1.40)
Time to diagnosis ≤ 10 years	0.92 (0.53, 1.59)	0.73 (0.35, 1.53)
Time to diagnosis > 10 years	0.91 (0.60, 1.38)	0.99 (0.61, 1.60)
B cell/CD8 ratio[c]	1.03 (0.82, 1.29)	1.07 (0.83, 1.38)
Time to diagnosis ≤ 10 years	1.01 (0.69, 1.49)	1.04 (0.64, 1.69)
Time to diagnosis > 10 years	1.05 (0.71, 1.55)	1.20 (0.74, 1.95)

CD8/treg ratio[c]	0.92 (0.67, 1.27)	0.92 (0.63, 1.35)
Time to diagnosis ≤ 10 years	0.73 (0.42, 1.26)	0.63 (0.30, 1.32)
Time to diagnosis > 10 years	1.03 (0.66, 1.62)	1.09 (0.63, 1.88)
Treg > 0 versus treg = 0[c]	0.96 (0.72, 1.28)	0.91 (0.64, 1.30)
Time to diagnosis ≤ 10 years	1.27 (0.76, 2.13)	1.37 (0.70, 2.69)
Time to diagnosis > 10 years	0.85 (0.58, 1.25)	0.75 (0.45, 1.26)
B cell naïve/B cell memory Ratio[c]	1.09 (0.84, 1.41)	1.23 (0.89, 1.70)
Time to diagnosis ≤ 10 years	1.31 (0.86, 1.99)	1.66 (0.74, 3.71)
Time to diagnosis > 10 years	1.03 (0.75, 1.42)	1.19 (0.73, 1.93)
CD4 naïve/CD4 memory ratio[c]	0.97 (0.70, 1.34)	1.03 (0.70, 1.52)
Time to diagnosis ≤ 10 years	0.91 (0.53, 1.57)	0.71 (0.32, 1.59)
Time to diagnosis > 10 years	1.00 (0.64, 1.57)	1.17 (0.70, 1.96)
CD8 naïve/CD8 memory ratio[c]	0.96 (0.70, 1.32)	1.10 (0.77, 1.57)
Time to diagnosis ≤ 10 years	1.09 (0.63, 1.88)	1.19 (0.59, 2.42)
Time to diagnosis > 10 years	0.92 (0.61, 1.40)	1.05 (0.67, 1.65)
CRP Score 1 [cd]	0.92 (0.67, 1.26)	0.88 (0.60, 1.30)
Time to diagnosis ≤ 10 years	1.00 (0.57, 1.75)	1.22 (0.67, 2.22)
Time to diagnosis > 10 years	1.04 (0.67, 1.63)	0.82 (0.46, 1.46)
CRP Score 2 [cd]	0.86 (0.65, 1.13)	0.98 (0.69, 1.41)
Time to diagnosis ≤ 10 years	0.76 (0.47, 1.23)	1.11 (0.61, 2.03)
Time to diagnosis > 10 years	1.10 (0.73, 1.66)	0.93 (0.55, 1.56)

Table 5 (continued)

	All lung c ancers[a] HR (95% CI)	**Non-small cell lung cancer[b] HR (95% CI)**
CRP Score 3 [cd]	1.05 (0.76, 1.47)	1.07 (0.73, 1.59)
Time to diagnosis ≤ 10 years	1.01 (0.56, 1.81)	1.01 (0.46, 2.22)
Time to diagnosis > 10 years	1.16 (0.74, 1.83)	1.22 (0.69, 2.16)

Bold HR and CI values indicate statistical significance. Underline HR and CI values indicate borderline statistical significance after multiple comparison adjustment [a] 205 cases (3 cases with person-year = 0 or > 25 years were removed from the analytical dataset) Model adjusted for age at blood draw, sex, smoking status, BMI, surrogate variables for batch effects, and a pack-year-based smoking methylation score. The results for Inflammation Scores additional adjust for cell proportions. BMI was removed from model when stratified by BMI. Group person-year: all lung cancer cases diagnosis > 10 years = 316.9; cases with BMI < 25 kg/m^2 = 184.3; cases with BMI ≥ 25 kg/m^2 == 236.9 421.3; cases with time to diagnosis ≤ 10 years = 104.3; cases with time to

[b] 149 cases (1 case with person-year = >25 years were removed from the analytical dataset) Model adjusted for age at blood draw, sex, smoking status, BMI, surrogate variables for batch effects, and methylation-predicted pack-years smoked. The results for Inflammation Scores additional adjust for cell proportions. BMI was removed from model when stratified by BMI. Group person-year for all non-small cell lung cancer (NSCLC) cases = 337.6; NSCLC cases with time to diagnosis 25 kg/m_2 = 196.2 ≤ 10 years = 84.4; NSCLC cases with time to diagnosis > 10 years = 253.2; NSCLC cases with BMI < 25 kg/m_2 = 141.4; NSCLC cases with BMI ≥

[c] HR results reported per 1 unit of SD increase among all cases (mdNLR SD: 1.86; CD4/CD8 ratio SD: 2.14; BL ratio SD: 0.0534; TL ratio SD: 0.0844; ML ratio SD: 0.125; NM/L ratio SD: 1.38; EL ratio SD: 0.079; CD4nv/L ratio SD: 0.0920; B/CD8 ratio SD: 2.21; CD8/Treg ratio SD: 112; Bnv/Bmem ratio SD: 81.3; CD4nv/CD4mem ratio SD: 0.335; CD8nv/CD8mem ratio SD: 1.64; CpG Score 1 SD: 0.979; CpG Score 2 SD: 0.945; CpG Score 3 SD: 1.02) and HR results reported for Treg > 0 versus Treg = 0 among all cases; Bonferroni-adjusted CI reported for all except mdNLR and CRP Scores 1–3 (a prior hypotheses) [d] CpG Scores 1, 2, and 3 are built using 54 CpG sites, the top 10 highly cell-specific CpG sites, and the 10 modestly cell-specific CpG sites, respectively

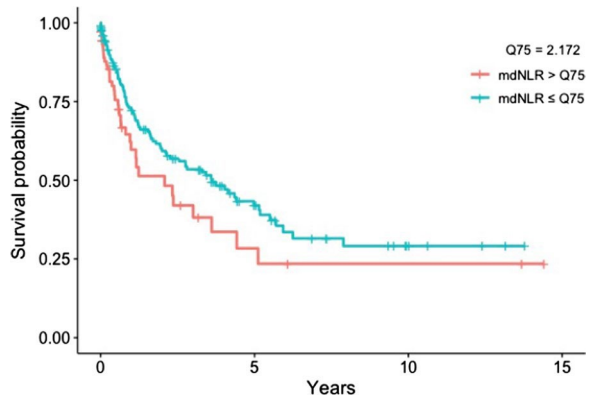


Fig. 1 Survival curves for lung cancer-specific mortality among lung cancer cases in the mdNLR high and low groups (> or ≤ 75% quartiles). Plot adjusted for age, sex, smoking status, methylation-predicted pack-years smoked, BMI, stage, and batch effects

addition, we built a series of methylation-derived CRP scores to capture individual systemic inflammatory profiles years before lung cancer diagnosis; these scores were inversely associated with risk of lung cancer, especially for NSCLC after adjusting for methylation-predicted pack-years smoked, but not with lung cancer-specific mortality.

Studies on NLR (calculated from measured WBC differentials) and lung cancer risk and survival typically measure pre-treatment NLR at diagnosis or up to 30 days prior to treatment [28–30]. Unlike prior studies, we were able to assess individual systemic inflammation profiles many years prior to diagnosis by using methylation markers of inflammation. Our study is not directly comparable to prior studies since we measured mdNLR using blood samples from subjects with a median of 14 years prior to lung cancer diagnosis. In addition, most cases in our study were diagnosed before the widespread use of immunotherapy. To our knowledge, only one other cohort, the multicenter β-Carotene and Retinol Efficacy Trial (CARET), examined pre-diagnosis mdNLR and lung cancer risk and survival using blood drawn years prior to diagnosis (median 4.7 years) [31, 32]. CARET, a study of heavy smokers, reported a 21% increased risk of lung cancer per one unit increase in mdNLR (OR: 1.21 [1.01, 1.45]), a 30% increased risk of NSCLC for one unit increase in mdNLR (OR: 1.30 [1.03, 1.63], and no association between higher pre-diagnosis mdNLR and risk of developing SCLC (OR: 1.06 [0.77, 1.47]) [31]. Like in CARET, in CLUE I/II we observed a 47% increased risk of NSCLC for a one standard deviation increase in mdNLR (n = 150 pairs; OR: 1.47 [1.08, 2.02]), but in contrast to CARET, we found no statistically significant association for overall lung cancer risk.

CARET researchers recently reported that pre-diagnosis mdNLR was positively associated with increased mortality for SCLC cases, but not for other case types [32]. In comparison, we observed a positive association between pre-diagnosis mdNLR and lung cancer-specific and NSCLC-specific mortality. In the case of SCLC, the number of cases was too limited for us to estimate stable associations (N = 29). Taken together, the CLUE and CARET results suggest that a systemic inflammatory profile marked by elevated NLR could indicate a lesser ability to mount a robust immune response to a developing lung cancer and/or a more favorable environment for cancer progression. Differences in findings between the two studies could stem from differences in study populations. The CARET cohort is exclusively heavy smokers, including a subgroup exposed to asbestos. In comparison, our analysis in the CLUE I/II cohorts included never, ever, and current smokers. Furthermore, our study population had a lower mdNLR in the lung cancer cases (mean 1.86 and SD 1.32) than in CARET (mdNLR mean 2.18 and SD 1.46).

Using a newly expanded deconvolution library, we were able to parse apart the granulocyte subtypes (neutrophils, eosinophils, and basophils) and investigate the balance between naïve and memory cell compartments for lung cancer. Previous research has identified the monocyte/lymphocyte (or lymphocyte/monocyte) ratio as an independent prognostic factor in NSCLC, demonstrating significant association with overall survival in patients with NSCLC [33–35]. In comparison, our exploratory analyses of immune cell ratios suggest that one standard deviation increase in the monocyte/lymphocyte ratio could potentially indicate increased risk of NSCLC after additionally adjusting for methylation-predicted pack-years. In addition, we found an increased risk of dying from lung cancer associated with an increase in Neu + Mono/ lymphocyte ratio among the NSCLC cases after multiple comparison adjustments.

We also investigated three CRP Scores that we built from 54 CpG sites that had been strongly associated with CRP in previous studies. We found these methylationpredicted CRP Scores to be moderately correlated with log-CRP and log-IL6 in the controls of a previously published pancreatic cancer dataset [27]. CRP is a systemic marker of chronic inflammation and has been reported as a risk factor for cancer development [36]. Previous studies of pre-diagnostic circulating CRP concentration and lung cancer risk (7 cohorts [10, 11, 19,

37–39] and 3 nested case–control studies [8, 12, 40]) have consistently found a moderate positive association between pre-diagnostic CRP concentrations and lung cancer risk. In our study, CRP Scores were not associated with lung risk when taking into account the matching factors, BMI, and batch effects. However, we observed an inverse association when additionally adjusting for methylationpredicted pack-year. Our results suggest that when strict control of smoking is applied, our CRP Score is likely capturing the unique individual immune response that is not driven by smoking.

Furthermore, these results provide preliminary evidence supporting the hypothesis that systemic inflammation not driven by smoking could have a protective effect on individuals. While smoking is by far the most important risk factor for lung cancer, our DNA methylationbased CRP Scores provide the opportunity to examine inflammatory measures not related to smoking that could play a role in modulating cancer risk years prior to diagnosis. Lastly, our experience with the CRP Scores suggests that measuring methylation-derived inflammatory responses using pre-diagnostic samples provides the opportunity to capture informative individual systemic inflammatory profiles years prior to diagnosis, potentially shedding light on risk factors key to lung cancer development and progression, e.g., underlying genetics, exposure to environmental risk factors, and behavior risk factors.

Like other observational studies, our study included a limited number of NSCLC and SCLC cases. The relatively small sample size of SCLC cases (N = 29) impacted our ability to observe associations for this subtype (SCLC comprises about 15% of lung cancer cases in the USA). In our survival analysis, we adjusted for stage and restricted our analysis to samples whose time between blood draw and date of lung cancer diagnosis was less than 25 years; however, our survival analysis did not have access to post-diagnosis smoking status information. Our study is also limited by a lack of replication dataset and reduced generalizability. (Study population is mainly White and with very few cases in never smokers.) The CRP Scores we built should be investigated in other populations to ensure that what we observed did not arise due to chance.

Conclusions

Our study suggests that elevated pre-diagnosis mdNLR and a lower non-smoking-related systemic inflammatory profile before diagnosis are associated with higher cancer risk and poorer lung cancer-specific survival. These relationships were especially evident for NSCLC. As the most common subtype of lung cancer, most NSCLC cases are diagnosed with locally advanced or metastatic disease. Our prospective results support future evaluation of whether DNA methylation-based inflammatory measures could enhance lung cancer risk stratification to improve targeted lung cancer screening.

Methods

Study Population

This nested case–control study selected cases and controls from individuals who participated and provided blood in both CLUE I and CLUE II [26]. The CLUE I cohort was developed to identify serologic precursors of cancer and was conducted in Washington County, Maryland, in the fall of 1974. A blood sample was collected from 25,620 volunteers at the time of participation [41, 42]. The CLUE II cohort was conducted from May through October 1989. During this time, 32,894 participants donated a blood sample which was collected in tubes containing heparin and kept chilled until centrifuged, aliquoted into plasma, erythrocytes, and buffy coat, and frozen at 70 °C [43]. In CLUE II, the baseline for this study, health information was collected at the time of blood draw, including attained education, cigarette smoking status, cigarette smoking dose, cigar/pipe smoking status, and self-reported weight and height.

Incident lung cancer cases were ascertained from linkage to the Washington County cancer registry (before 1992 to the present) and the Maryland Cancer Registry (since 1992 when it began to the present). We ascertained 241 incident lung cancer cases who participated in CLUE I and were diagnosed after the day of blood draw in CLUE II through January 2018. Cases were characterized with respect to histology. We used incidence density sampling to select one control matched to each case on age, sex, smoking status and intensity (cig/day), and cigar/pipe smoking status. Death from lung cancer as the underlying cause was obtained from death certificates. The Institutional Review Board at the Johns Hopkins Bloomberg School of Public Health and the Tufts University Health Sciences Campus Institutional Review Board approved this study.

DNA methylation measurements

Extracted DNA was bisulfite-treated using the EZ DNA Methylation Kit (Zymo), and DNA methylation was measured with the 850 K Illumina Infinium MethylationEPIC BeadChip Arrays (Illumina, Inc., CA, USA). All samples and all array experiments were performed blinded to case–control status. Details on DNA methylation measurements, data preprocessing

processing, and quality control assessment/screening are provided in the Additional file 1. The 850 K methylation microarray has been validated from a biological and technical standpoint. Reproducibility of results from 850 K Illumina array has been previously shown to be very high (r = 0.997) [44]. DNA volume and quality were sufficient for 208 of the cases and 222 controls totaling 208 matched pairs.

Estimation of peripheral blood leukocyte composition

Peripheral blood leukocyte subtypes proportions, including myeloid lineage sub-types [neutrophils (Neu), eosinophils (Eos), basophils (Bas), and monocytes (Mono)] and lymphoid lineage subtypes [B lymphocytes naïve (Bnv), B lymphocytes memory (Bmem), T helper lymphocytes naïve (CD4nv), T helper lymphocytes memory (CD4mem), T regulatory cells (Treg), T cytotoxic lymphocytes naïve (CD8nv), T cytotoxic lymphocytes memory (CD8mem), and natural killer lymphocytes (NK)], were estimated using a newly expanded reference-based deconvolution library EPIC IDOL-Ext [45]. This library used the IDOL methodology [46] to optimize the currently available six-cell reference library [47] in order to deconvolve the proportions of 12 leukocyte subtypes in peripheral blood. This EPIC IDOL-Ext library (Bioconductor package FlowSorted.BloodExtended.EPIC) was validated using flow cytometry gold standard data and substantiated by including publicly available data from > 100,000 samples [45].

Methylation-Derived Neutrophil Lymphocyte Ratio (mdNLR)

The peripheral blood neutrophil-to-lymphocyte ratio (NLR) is a cytological marker of both inflammation and poor outcomes in cancer patients [48–52]. We used a DNA methylation-derived NLR (mdNLR) index to predict the common clinical NLR parameter using a previously described approach [9]. This index is based on normal isolated leukocyte reference DNA methylation libraries and established reference-based cell mixture deconvolution algorithms [9, 53].

Inflammation-associated CpG score

We used 54 CpG sites that have been strongly associated with C-reactive protein (CRP) [54, 55] to build three CRP Scores. We selected these 54 CpGs (remaining 4 were not on the 850 K array that we used) from the 58 CpGs identified by Ligthart and colleagues [54] for their association with serum CRP level (listed in Table 3) using 450 K DNA methylation data. Forty-five of these 58 CpG sites were validated to have the same direction of protein–methylation associations by Myte et al. [55]. These CpGs, while identified based on their CRP association, have also been shown to be associated with other inflammatory mediators [54–56]. To compute CRP Score 1, we multiplied the beta value at each selected CpG site with the effect size estimates reported by Ligthart et al. These estimated beta coefficients represented the change in DNA methylation per one unit increase in log CRP. In the CRP Score 1 formula, we weighted the beta coefficients estimated by Ligthart et al. with their corresponding standard errors.

$$\text{CRP Score}_i = B_{ij} \times \text{SE}__{}_{jj}$$

B_{ij} is the beta value for the ith participant at the jth CpG site. Δ_j is the beta coefficients reported by Ligthart et al. for the jth CpG site. SE_j is the SE reported by Ligthart et al. for the jth CpG site.

Since most of the estimated beta coefficients are negative, CRP Score 1 ranged between − 0.059 and -0.026 in these participants. A score closer to zero indicated higher CRP levels. Based on CRP Score 1, we computed two additional CRP Scores, one cell (leukocyte)-type invariant (CRP Score 2) and one cell-specific (CRP Score 3). Among the 54 inflammation (CRP)-associated CpGs, we identified putative cell-type invariant and cell-specific CpGs by conducting ANOVA using the dataset described in Salas and Koestler et al. [47] and publicly available on the Gene Expression Omnibus (GSE110555). The dataset used for this ANOVA consisted of EPIC methylation data profiled in purified leukocyte cell population isolated from different healthy adults. Specifically, methylation signatures were available for CD4 + T cells, CD8 + T cells, NK cells, B cells, monocytes, and neutrophils. One-way ANOVA models were fit independently to each of the 54 CRP-associated CpGs treating methylation as the dependent variable and cell type as the independent variable. We tested the null hypothesis that the mean methylation beta-value is the same across the cell types. The F-statistic, corresponding p value, and maximum absolute pairwise difference in the mean methylation beta value across cell types were calculated for each of the 54 CpGs. We then selected subgroups of CpG sites that had the top 10 smallest or top 10 largest F-statistic value to build the two additional CRP Scores. CRP Score 2 consists of putative cell-specific CpGs with high F-statistics, e.g., those exhibiting a difference in mean

methylation beta-values between at least two of the six cell types. CRP Score 3 is made of cell-type invariant CpGs with low *F*-statistics, e.g., CpGs for which there did not appear to be a substantial difference in mean methylation beta-values across the normal six leukocyte subtypes. Score 2 ranged between − 0.0002 and 0.0046, while Score 3 ranged between − 0.025 and − 0.016. In the regression analyses, we used a standardized version of CRP Scores 1, 2, and 3 (mean = 0, sd = 1) for easier interpretation of results and allowing us to compare the results for each of the scores.

Statistical analyses

All statistical analyses were performed in R (version 3.5.1). We estimated mdNLR as described above, used an independent pancreatic cancer dataset [27] to estimate the correlation between estimated values of CRP Scores 1–3 with the log CRP and log IL-6 levels, and tested a series of a priori hypotheses concerning the mdNLR and CRP Scores. In addition, we also conducted exploratory analyses to generate novel hypotheses regarding the role of methylation-derived leukocyte proportions in lung cancer. Immune cell ratios (e.g., CD4/CD8, Neu/lymphocyte, B cell/lymphocyte, T cell/lymphocyte, Mono/ lymphocyte, Neu + Mono/lymphocyte, Eos/lymphocyte, CD4nv/lymphocyte, B cell/CD8, CD8/Treg, Bnv/ Bmem, CD4nv/CD4mem, and CD8nv/CDmem) were
calculated for each sample by taking the ratio of its predicted cell proportions described above and tested as continuous variables. The presence of Treg was tested as a dichotomous variable. Given the need for multiple comparison adjustment, Bonferroni adjustment (familywise error rate = 0.0013) was conducted for all exploratory analyses.

We used conditional logistic regression to examine the association between DNA methylation-based inflammatory measures (CRP Scores 1–3 and continuous mdNLR) and lung cancer risk. Models were fit with age, sex, and smoking status (never, former, current) as matching factors and were adjusted for potential confounding factors, including body mass index (BMI), batch effect, and previously described methylation-predicted pack-years smoked [57]. These analyses did not additionally adjust for methylation-derived cell proportions given how these proportions correlated with methylation-based inflammatory measures (Table 3). We repeated these analyses by lung cancer histology (NSCLC, SCLC), length of time between blood draw and diagnosis (< ^{2}). = 10, > 10 years), and BMI (< 25, ≥ 25 kg/m

Among the lung cancer cases, we examined the association between these same pre-diagnostic DNA methylation-based inflammatory measures (CRP Scores 1–3 and continuous mdNLR) and risk of lung cancer-specific death using a series of multivariable Cox proportional hazard regression adjusting for age, gender, smoking status, BMI, stage at diagnosis (three strata: stage 1 & 2, stage 3 & 4, and missing), cell proportion, batch effects, and methylation-predicted pack-years smoked. The proportional hazards assumption was checked by conducting global tests of correlating the set of scaled Schoenfeld residuals with time for each covariate. We excluded three lung cancer cases whose date of diagnosis and date of death were the same, or whose time between blood draw and date of lung cancer diagnosis was longer than 25 years. Cases were followed until their date of death from lung cancer, death from another cause, or the end of follow up in 2018, whichever came first.

Supplementary Information

The online version contains supplementary material available at https:// doi. org/ 10. 1186/ s13148- 021- 01214-2.

Additional file 1 Supplementary Methods.

Acknowledgements

Cancer data were provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health, with funding from the State of Maryland and the Maryland Cigarette Restitution Fund. The collection and availability of cancer registry data are also supported by the Cooperative Agreement NU58DP006333, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

Authors' contributions

DSM, KTK, and EAP designed the study, obtained funding, and acquired the data. JL assisted with identifying the study samples and preparation of dataset.
LAS assisted with cell deconvolution. DSM supervised all research activities. MR and NZ conducted the statistical analyses. NZ drafted the manuscript. DSM, EAP, KTK, and DCK interpreted the data and provided critical revisions of the manuscript. All authors read and approved the final version of the manuscript.

Funding

This work was supported by 2018 American Association for Cancer Research (AACR)-Johnson and Johnson Lung Cancer Innovation Science (18-90-52MICH). Note: The funders had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication.

Availability of data and materials

The datasets generated during the current study are available from the corresponding author on reasonable request and will be deposited into dbGaP by publication.

Declarations

Ethics approval and consent to participate

The Institutional Review Board at the Johns Hopkins Bloomberg School of Public Health and the Tufts University Health Sciences Campus Institutional Review Board approved this study.

Consent for publication Not applicable.

Competing interests

Dr. Kelsey is a founder and scientific advisor for Cellintec, which had no role in this research.

Author details

[1] Department of Public Health and Community Medicine, Tufts University School of Medicine, Tufts University, 136 Harrison Avenue, Boston, MA 02111, USA. [2] Department of Biostatistics and Data Science, Medical Center, University of Kansas, Kansas City, KS, USA. [3] University of Kansas Cancer Center, Kansas City, KS, USA. [4] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [5] The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD, USA. [6] Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA. [7] Department of Epidemiology, Brown University, Providence, RI, USA. [8] Department of Pathology and Laboratory Medicine, Brown University, Providence, RI, USA.

Received: 19 August 2021 Accepted: 9 December 2021

Published online: 16 December 2021

References

1. ACS. Cancer Facts & Figures 2021. Atlanta: American Cancer Society, Inc. 2021.
2. Howlader N, Forjaz G, Mooradian MJ, Meza R, Kong CY, Cronin KA, et al. The effect of advances in lung-cancer treatment on population mortality. N Engl J Med. 2020;383(7):640–9.
3. Howlader N, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, et al. SEER cancer statistics review, 1975–2017, vol. 21. Bethesda, MD: National Cancer Institute; 2020. p. 12.
4. Fabrikant MS, Wisnivesky JP, Marron T, Taioli E, Veluswamy RR. Benefits and challenges of lung cancer screening in older adults. Clin Ther. 2018;40(4):526–34.
5. Force UPST. Screening for lung cancer: US preventive services task force recommendation statement. JAMA. 2021;325(10):962–70.
6. Crusz SM, Balkwill FR. Inflammation and cancer: advances and new agents. Nat Rev Clin Oncol. 2015;12(10):584–96.
7. Zhou W, Liu G, Hung RJ, Haycock PC, Aldrich MC, Andrew AS, et al. Causal relationships between body mass index, smoking and lung cancer: univariable and multivariable Mendelian randomization. Int J Cancer. 2021;148(5):1077–86.
8. Chaturvedi AK, Caporaso NE, Katki HA, Wong H-L, Chatterjee N, Pine SR, et al. C-reactive protein and risk of lung cancer. J Clin Oncol. 2010;28(16):2719.
9. Koestler DC, Usset J, Christensen BC, Marsit CJ, Karagas MR, Kelsey KT, et al. DNA methylation-derived neutrophil-to-lymphocyte ratio: an epigenetic tool to explore cancer inflammation and outcomes. Cancer Epidemiol Prev Biomark. 2017;26(3):328–38.
10. Il'yasova D, Colbert LH, Harris TB, Newman AB, Bauer DC, Satterfield S, et al. Circulating levels of inflammatory markers and cancer risk in the health aging and body composition cohort. Cancer Epidemiol Prev Biomark. 2005;14(10):2413–8.
11. Siemes C, Visser LE, Coebergh J-WW, Splinter TA, Witteman JC, Uitterlinden AG, et al. C-reactive protein levels, variation in the C-reactive protein gene, and cancer risk: the Rotterdam Study. J Clin Oncol. 2006;24(33):5216–22.
12. Trichopoulos D, Psaltopoulou T, Orfanos P, Trichopoulou A, Boffetta P. Plasma C-reactive protein and risk of cancer: a prospective study from Greece. Cancer Epidemiol Prev Biomark. 2006;15(2):381–4.
13. Guthrie GJ, Charles KA, Roxburgh CS, Horgan PG, McMillan DC, Clarke SJ. The systemic inflammation-based neutrophil–lymphocyte ratio: experience in patients with cancer. Crit Rev Oncol Hematol. 2013;88(1):218–30.
14. Templeton AJ, McNamara MG, Šeruga B, Vera-Badillo FE, Aneja P, Ocaña A, et al. Prognostic role of neutrophil-to-lymphocyte ratio in solid tumors: a systematic review and meta-analysis. J Natl Cancer Inst. 2014;106(6):dju124. https:// doi. org/ 10. 1093/ jnci/ dju124.
15. Campa D, Zienolddiny S, Maggini V, Skaug V, Haugen A, Canzian F. Association of a common polymorphism in the cyclooxygenase 2 gene with risk of non-small cell lung cancer. Carcinogenesis. 2004;25(2):229–35.
16. Engels EA, Wu X, Gu J, Dong Q, Liu J, Spitz MR. Systematic evaluation of genetic variants in the inflammation pathway and risk of lung cancer. Can Res. 2007;67(13):6520–7.
17. Seifart C, Plagens A, Dempfle A, Clostermann U, Vogelmeier C, von Wichert P, et al. TNF-α, TNF-β, IL-6, and IL-10 polymorphisms in patients with lung cancer. Dis Mark. 2005;21(3):157–65.
18. Shih C-M, Lee Y-L, Chiou H-L, Hsu W-F, Chen W-E, Chou M-C, et al. The involvement of genetic polymorphism of IL-10 promoter in non-small cell lung cancer. Lung Cancer. 2005;50(3):291–7.
19. Allin KH, Bojesen SE, Nordestgaard BG. Baseline C-reactive protein is associated with incident cancer and survival in patients with cancer. J Clin Oncol. 2009;27(13):2217–24.
20. Heikkilä K, Ebrahim S, Lawlor DA. A systematic review of the association between circulating concentrations of C reactive protein and cancer. J Epidemiol Community Health. 2007;61(9):824–33.
21. Qu Z, Sun F, Zhou J, Li L, Shapiro SD, Xiao G. Interleukin-6 prevents the initiation but enhances the progression of lung cancer. Cancer Res. 2015;75(16):3209–15.
22. Ballaz S, Mulshine JL. The potential contributions of chronic inflammation to lung carcinogenesis. Clin Lung Cancer. 2003;5(1):46–62.
23. Bauer AK, Dwyer-Nield LD, Keil K, Koski K, Malkinson AM. Butylated hydroxytoluene (BHT) induction of pulmonary inflammation: a role in tumor promotion. Exp Lung Res. 2001;27(3):197–216.
24. Engels EA. Inflammation in the development of lung cancer: epidemiological evidence. Expert Rev Anticancer Ther. 2008;8(4):605–15.
25. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
26. Genkinger JM, Platz EA, Hoffman SC, Comstock GW, Helzlsouer KJ. Fruit, vegetable, and antioxidant intake and all-cause, cancer, and cardiovascular disease mortality in a community-dwelling population in Washington County, Maryland. Am J Epidemiol. 2004;160(12):1223–33.
27. Michaud DS, Ruan M, Koestler DC, Alonso L, Molina-Montes E, Pei D, et al. DNA Methylation–derived immune cell profiles, CpG markers of inflammation, and pancreatic cancer risk. Cancer Epidemiol Prev Biomark. 2020;29(8):1577–85.
28. Lohinai Z, Bonanno L, Aksarin A, Pavan A, Megyesfalvi Z, Santa B, et al. Neutrophil–lymphocyte ratio is prognostic in early stage resected smallcell lung cancer. PeerJ. 2019;7:e7232.
29. Galvano A, Peri M, Guarini AA, Castiglia M, Grassadonia A, De Tursi M, et al. Analysis of systemic inflammatory biomarkers in neuroendocrine carcinomas of the lung: prognostic and predictive significance of NLR,

LDH, ALI, and LIPI score. Ther Adv Med Oncol. 2020;12:1758835920942378.
30. Ozyurek BA, Ozdemirel TS, Ozden SB, Erdogan Y, Kaplan B, Kaplan T. Prognostic value of the neutrophil to lymphocyte ratio (NLR) in lung cancer cases. Asian Pac J Cancer Prev. 2017;18(5):1417.
31. Grieshober L, Graw S, Barnett MJ, Thornquist MD, Goodman GE, Chen C, et al. Methylation-derived neutrophil-to-lymphocyte ratio and lung cancer risk in heavy smokers. Cancer Prev Res. 2018;11(11):727–34.
32. Grieshober L, Graw S, Barnett MJ, Goodman GE, Chen C, Koestler DC, et al. Pre-diagnosis neutrophil-to-lymphocyte ratio and mortality in individuals who develop lung cancer. Res Sq 2021.
33. Mandaliya H, Jones M, Oldmeadow C, Nordman II. Prognostic biomarkers in stage IV non-small cell lung cancer (NSCLC): neutrophil to lymphocyte ratio (NLR), lymphocyte to monocyte ratio (LMR), platelet to lymphocyte ratio (PLR) and advanced lung cancer inflammation index (ALI). Transl Lung Cancer Res. 2019;8(6):886.
34. Hu P, Shen H, Wang G, Zhang P, Liu Q, Du J. Prognostic significance of systemic inflammation-based lymphocyte-monocyte ratio in patients with lung cancer: based on a large cohort study. PLoS ONE. 2014;9(10):e108062.
35. Chen X, Wu J, Zhang F, Ying L, Chen Y. Prognostic significance of preoperative monocyte-to-lymphocyte ratio in lung cancer patients undergoing radical surgery. Lab Med. 2018;49(2):e29–39.
36. Tsilidis KK, Branchini C, Guallar E, Helzlsouer KJ, Erlinger TP, Platz EA. C-reactive protein and colorectal cancer risk: a systematic review of prospective studies. Int J Cancer. 2008;123(5):1133–40.
37. Heikkilä K, Harris R, Lowe G, Rumley A, Yarnell J, Gallacher J, et al. Associations of circulating C-reactive protein and interleukin-6 with cancer risk: findings from two prospective cohorts and a meta-analysis. Cancer Causes Control. 2009;20(1):15–26.
38. dos Santos SI, De Stavola BL, Pizzi C, Meade TW. Circulating levels of coagulation and inflammation markers and cancer risks: individual participant analysis of data from three long-term cohorts. Int J Epidemiol. 2010;39(3):699–709.
39. Van Hemelrijck M, Holmberg L, Garmo H, Hammar N, Walldius G, Binda E, et al. Association between levels of C-reactive protein and leukocytes and cancer: three repeated measurements in the Swedish AMORIS study. Cancer Epidemiol Prev Biomark. 2011;20(3):428–37.
40. Suzuki K, Ito Y, Wakai K, Kawado M, Hashimoto S, Seki N, et al. Serum heat shock protein 70 levels and lung cancer risk: a case-control study nested in a large cohort study. Cancer Epidemiol Prev Biomark. 2006;15(9):1733–7.
41. Braun MM, Helzlsouer KJ, Hollis BW, Comstock GW. Colon cancer and serum vitamin D metabolite levels 10–17 years prior to diagnosis. Am J Epidemiol. 1995;142(6):608.
42. Comstock G, Helzlsouer KJ, Bush TL. Prediagnostic serum levels of carotenoids and vitamin E as related to subsequent cancer in Washington County, Maryland. Am J Clin Nutr. 1991;53(1):260S-S264.
43. Kakourou A, Koutsioumpa C, Lopez DS, Hoffman-Bolton J, Bradwin G, Rifai N, et al. Interleukin-6 and risk of colorectal cancer: results from the CLUE II cohort and a meta-analysis of prospective studies. Cancer Causes Control. 2015;26(10):1449–60.
44. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;8(3):389–99.
45. Salas LA, Zhang Z, Koestler DC, Butler RA, Hansen HM, Molinaro AM, et al. Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. bioRxiv. 2021.
46. Koestler DC, Jones MJ, Usset J, Christensen BC, Butler RA, Kobor MS, et al. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinform. 2016;17(1):120.
47. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19(1):1–14.
48. Wang Y, Liu P, Xu Y, Zhang W, Tong L, Guo Z, et al. Preoperative neutrophilto-lymphocyte ratio predicts response to first-line platinum-based chemotherapy and prognosis in serous ovarian cancer. Cancer Chemother Pharmacol. 2015;75(2):255–62.
49. Salim DK, Mutlu H, Eryılmaz MK, Salim O, Musri FY, Tural D, et al. Neutrophil to lymphocyte ratio is an independent prognostic factor in patients with recurrent or metastatic head and neck squamous cell cancer. Mol Clin Oncol. 2015;3(4):839–42.
50. Yu Y, Wang H, Yan A, Wang H, Li X, Liu J, et al. Pretreatment neutrophil to lymphocyte ratio in determining the prognosis of head and neck cancer: a meta-analysis. BMC Cancer. 2018;18(1):1–9.
51. Ethier J-L, Desautels DN, Templeton AJ, Oza A, Amir E, Lheureux S. Is the neutrophil-to-lymphocyte ratio prognostic of survival outcomes in gynecologic cancers? A systematic review and meta-analysis. Gynecol Oncol. 2017;145(3):584–94.
52. Diem S, Schmid S, Krapf M, Flatz L, Born D, Jochum W, et al. Neutrophilto-Lymphocyte ratio (NLR) and Platelet-to-Lymphocyte ratio (PLR) as prognostic markers in patients with non-small cell lung cancer (NSCLC) treated with nivolumab. Lung Cancer. 2017;111:176–81.
53. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform. 2012;13(1):1–16.
54. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biol. 2016;17(1):1–15.
55. Myte R, Sundkvist A, Van Guelpen B, Harlid S. Circulating levels of inflammatory markers and DNA methylation, an analysis of repeated samples from a population based cohort. Epigenetics. 2019;14(7):649–59.
56. Ahsan M, Ek WE, Rask-Andersen M, Karlsson T, Lind-Thomsen A, Enroth S, et al. The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. PLoS Genet. 2017;13(9):e1007005.
57. Sugden K, Hannon EJ, Arseneault L, Belsky DW, Broadbent JM, Corcoran DL, et al. Establishing a generalized polyepigenetic biomarker for tobacco smoking. Transl Psychiatry. 2019;9(1):1–12.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Chen et al. Clinical Epigenetics (2022) 14:14
https://doi.org/10.1186/s13148-022-01234-6

Clinical Epigenetics

RESEARCH Open Access

Immune profiles and DNA methylation alterations related with non-muscle-invasive bladder cancer outcomes

Ji-Qing Chen[1], Lucas A. Salas[1], John K. Wiencke[2], Devin C. Koestler[3], Annette M. Molinaro[2], Angeline S. Andrew[4], John D. Seigne[5], Margaret R. Karagas[1], Karl T. Kelsey[6] and Brock C. Christensen[1,7,8*]

Abstract

Background: Non-muscle-invasive bladder cancer (NMIBC) patients receive frequent monitoring because ≥ 70% will have recurrent disease. However, screening is invasive, expensive, and associated with significant morbidity making bladder cancer the most expensive cancer to treat per capita. There is an urgent need to expand the understanding of markers related to recurrence and survival outcomes of NMIBC.

Methods and results: We used the Illumina HumanMethylationEPIC array to measure peripheral blood DNA methyl-

ation profiles of NMIBC patients (*N* = 603) enrolled in a population-based cohort study in New Hampshire and applied *Cox* proportional hazard models, we identicell type deconvolution to estimate immune cell-type proportions. Using fied that increasing CD4T and CD8T cell proportions were associated with a statistically significant decreased hazard of tumor recurrence or death (CD4T: HR = 0.98, 95% CI = 0.97–1.00; CD8T: HR = 0.97, 95% CI = 0.95–1.00), whereas increasing monocyte proportion and methylation-derived neutrophil-to-lymphocyte ratio (mdNLR) were associated with the increased hazard of tumor recurrence or death (monocyte: HR = 1.04, 95%

CI = 1.00–1.07; mdNLR: HR = 1.12, 95% CI = 1.04–1.20). Then, using an epigenome-wide association study (EWAS) approach adjusting for age, sex, smoking status, BCG treatment status, and immune cell profiles, we identified 2528 CpGs associated with the hazard of tumor recurrence or death ($P < 0.005$). Among these CpGs, the 1572 were associated with an increased hazard and were significantly enriched in open sea regions; the 956 remaining CpGs were associated with a decreased hazard and were significantly enriched in enhancer regions and DNase hypersensitive sites.

Conclusions: Our results expand on the knowledge of immune profiles and methylation alteration associated with NMIBC outcomes and represent a first step toward the development of DNA methylation-based biomarkers of tumor recurrence.

Keywords: DNA methylation, Non-muscle-invasive bladder cancer, Immune profile, Immunomethylomic, Recurrence, Survival

Background

In 2021, the estimated number of bladder cancer deaths is projected to be 17,200, with an estimated number of new cases of 83,730 in the USA. Bladder cancer is the fourth most common cancer among men and twelfth most common among women, which may in part be due to smoking prevalence rates being higher in men than women as cigarette smoking accounts for half of all cases

*Correspondence: Brock.Christensen@Dartmouth.edu
[8] Dartmouth Hitchcock Medical Center, 1 Medical Center Dr, 660 Williamson Translation Research Building, Lebanon, NH 03756, USA

Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http:// creat iveco mmons. org/ licen ses/ by/4. 0/. The Creative Commons Public Domain Dedication waiver (http:// creat iveco mmons. org/ publi cdoma in/ zero/1. 0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(47%) in the USA. [1]. Seventy-five percent of bladder cancers are diagnosed as low-grade non-muscle invasive tumors, NMIBC [2]. Cystoscopy is used for diagnosis (biopsy) with transurethral excision of localized tumors as the primary treatment [3–5]. Although transurethral excisions can successfully control the disease and mortality from bladder cancer among patients with localized tumors is low, 45% of NMIBC cases have recurrences within 12 months of surgery [6]. In addition, frequent invasive follow-up via cystoscopy without prognostic markers leads to significant patient morbidity and comes with a considerable cost burden to the health care system, estimated at approximately $4 billion dollars annually in the USA [7]. To control the patient and healthcare burdens associated with NMIBC, there is an imminent need for biomarkers to identify those at the highest risk of tumor recurrence.

Peripheral blood immune profiles have been associated with different outcomes in bladder cancer patients and may have clinical utility for NMIBC prognosis [8–13]. For instance, in NMIBC, patients with elevated neutrophil-to-lymphocyte ratio (NLR) had poorer cancerspecific survival than patients with lower NLR [9–11]. Thus, elevated NLR could be a potential predictor of overall survival and cancer-specific survival for this disease. Other studies have indicated that bladder cancer patients with increased lymphocyte-to-monocyte ratio had poorer overall survival and cancer-specific survival [12, 13]. In addition, Bacillus Calmette–Guérin (BCG), a commonly used intravesical immunotherapy for NMIBC administered post-surgery, has been reported to reduce the proportion of natural killer T cells, memory CD4T, CD8T, and regulatory T cells in the peripheral blood of NMIBC patients [8]. Together, circulating immune profiles might be promising markers for reducing adverse outcomes in NMIBC patients. Previous studies have relied on complete blood count differential (CBC) tests to determine immune profile variables [14, 15]. The CBC test relies on fresh blood samples and is incapable of providing proportions of specific lymphocyte subtypes [16].

Our prior work has established methods to infer the immune profiles in archival samples using immune celltype-specific DNA methylation [17, 18]. DNA methylation plays an essential role in gene regulation for cell lineage specification [19, 20]. Differentially methylated regions (DMRs) have been used to distinguish cell types, including leukocyte subtypes, and form the basis of reference-based deconvolution methods for estimating specific immune cell-type proportions [18, 21, 22]. Compared with cytological methods for determining cell type abundances and proportions, such as flow cytometry, DNA methylation-based cell-type deconvolution does not require a fresh substrate, intact cells, or batch-sensitive reagents, is reproducible and cost-effective relative to time-sensitive blood processing [23, 24]. With cell-type deconvolution approaches, it is possible to identify immune cell-type profiles and test their relationship with cancer outcomes in archived samples [25]. This study used the archival blood samples of NMIBC patients to test the association between the immune profiles and outcomes in bladder cancer patients.

In the present study, we sought to identify immune profiles and epigenetic features associated with disease recurrence in the hope that such information might help improve the management of NMIBC. Here, we hypothesized that CpG-specific DNA methylation and DNA methylation-derived immune cell profiles are associated with recurrence-free survival in NMIBC patients. We used archival blood samples from a population-based case–control study to obtain genome-scale DNA methylation profiles. We then investigated the association between the methylation-derived immune profiles and outcomes in NMIBC patients. Preliminary work from our group observed an association between methylationderived NLR (mdNLR) and survival in bladder cancer patients using a smaller sample size (223 cases) and an early genome-scale methylation array (HumanMethylation27K array) [26]. In this study, we

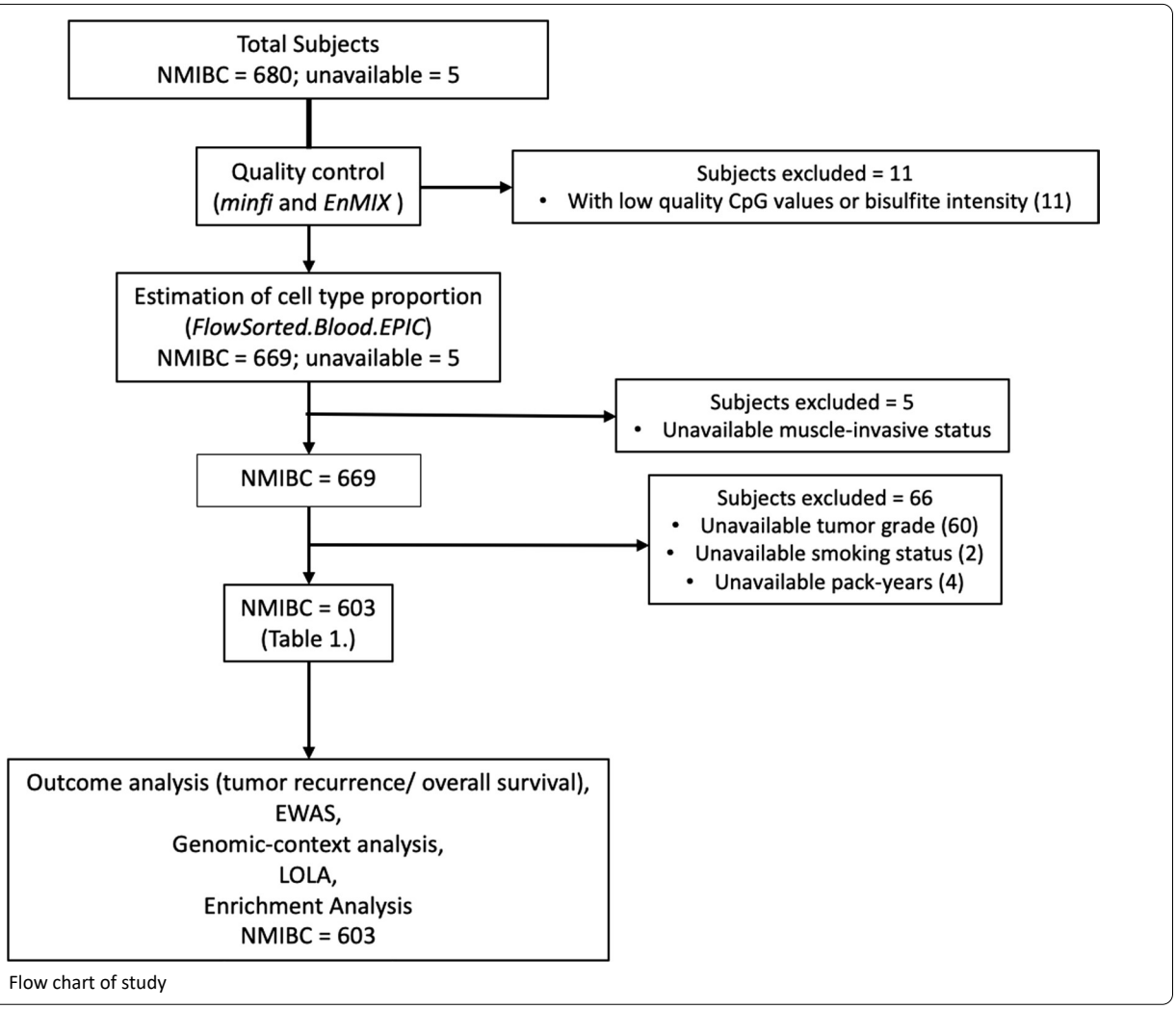


Fig. 1 Flow chart of study

increased the sample size (603 cases), used a new more comprehensive array (HumanMethylationEPIC array) with 30 times as many measured features, and a new cell type deconvolution library [18] to estimate immune cell-type proportions. An epigenome-wide association study (EWAS) and enrichment analyses were used to determine possible CpG sites and gene sets associated with recurrence-free survival.

Results

Profiles of DNA methylation were obtained from 685 peripheral blood samples using the Human MethylationEPIC array. Eighty-two subjects were excluded due to low-quality CpG value or bisulfite intensity (n = 11), or without muscle-invasive status, histopathology re-review, tumor grade, smoking status, and pack-years (n = 71) (Fig. 1). The remaining subjects (N = 603) included in the study group were 75.8% men (n = 457), 82.9% eversmokers (n = 500), and had a median age of 66 (Table 1). We estimated the cell-type proportions for each patient using methylation values by performing *FlowSorted. Blood.EPIC* (see Additional file 7: Figure S1 for the distribution). Neutrophil-to-lymphocyte ratio (NLR) was then calculated according to the ratio of neutrophil proportion to lymphocyte proportion (B cell + CD4T cell + CD8T cell + NK cell), and the median methylation-derived NLR (mdNLR) was 1.97. Further details of study population characteristics are described in Table 1.

To characterize 10-year recurrence-free survival (RFS), we generated Kaplan–Meier curves for each covariate and fit a Cox proportional hazard regression model for univariate analyses and multivariable analyses, respectively. In the Kaplan–Meier analysis, the NMIBC patients aged > 65 had a worse probability of RFS compared with the NMIBC patients with age ≤ 65 (P = 0.0006). Females had a greater probability of RFS compared with males (P = 0.002), and the NMIBC patients with highgrade tumors (grade 3 + 4) had a lesser probability of

RFS than those with low-grade tumors (grade 1 -[5]). Ever-smokers had a worse probability + 2)
(P = 8.0 × 1 0

of RFS than the NMIBC patients who were never smokers ($P = 3.0 \times 1\ 0\text{-}^{4}$). NMIBC patients with low mdNLR had a greater probability of RFS than the patients with high mdNLR ($P = 0.002$) (Fig. 2). Consistent with the Kaplan–Meier results, in a multivariable Cox model, age > 65 (HR = 1.01, 95% CI = 1.00–1.03), high tumor grade (HR = 1.48, 95% CI = 1.17–1.87), ever-smoking (HR = 1.65, 95% CI = 1.22–2.25), and mdNLR (HR = 1.12, 95% CI = 1.04–1.20) were significantly associated with an increased hazard of RFS (Table 2).

We next investigated the association between immune cell-type proportions and RFS in multivariable models. CD4T (HR = 0.98, 95% CI = 0.97–1.00) and CD8T cell proportion (HR = 0.97, 95% CI = 0.95– 1.00) were significantly associated with the decreased hazard of RFS. Monocyte cell proportion (HR = 1.04, 95% CI = 1.00–1.07) was significantly associated with the increased hazard of RFS. Both B cell and NK cell proportion hazard estimates were < 1 but not statistically significant (Additional file 1). We also examined the 10-year overall survival (OS) in univariate models and the multivariable models. Observed associations with RFS were consistent for OS for: age, male, high tumor grade, ever-smoking, and mdNLR (Table 3 and Additional file 7: Figure S2). In addition, CD4T, CD8T, B cell, and NK cell proportion were significantly

Table 1 Characteristics of subjects after excluding subjects with missing values

	NMIBC (*n* = 603)
Age	
Median (Q1, Q3)	66 (57,71)
Sex	
Male	457 (75.8%)
Female	146 (24.2%)
Tumor grade	
Grade 1	290 (48.1%)
Grade 2	162 (26.9%)
Grade 3	125 (20.7%)
Grade 4	26 (4.3%)
Smoking status	
Never	103 (17.1%)
Ever	500 (82.9%)
BCG: immunotherapy	
No	514 (85.2%)
Yes	89 (14.8%)
NLR	
Median (Q1, Q3)	1.97 (1.40, 2.93)
10-year dead status	
Alive	423 (70.1%)
Deceased	180 (29.9%)
10-year survival	
Median (Q1, Q3)	120.0 (105, 120)
10-year recurrence status	
No	193 (32.0%)
Yes	295 (48.9%)
Missing	115 (19.1%)
10-year recurrence	
Median (Q1, Q3)	17.7 (6.9, 67.1)
Missing	115

Exclude subjects with missing values in muscle-invasive status, pathology reviewing status, tumor grade, smoking status, or pack-years

associated with the decreased hazard of death; neutrophil cell proportion was significantly associated with the increased hazard of death (Additional file 2).

Next, we assessed the relation of NMIBC patient outcomes with DNA methylation. First, Cox proportional hazards models were fit for each CpG controlling for age, stratified sex, tumor grade, smoking status, and stratified BCG receiving status. Without adjustment for immune cell proportions, we identified 27,575 CpGs whose methylation was associated with a significant ($P < 0.005$) difference in hazard of RFS (Fig. 3A, Additional file 3). We then conducted similar analyses controlling for age, sex, tumor grade, smoking status, BCG status, and winsorized immune cell proportions. As expected and demonstrating the importance of adjusting for cell-type proportions, the fully adjusted models were attenuated and identified 2528 CpGs whose methylation was associated with a significant difference ($P < 0.005$) in the hazard of RFS. The 10 CpGs most strongly (with smallest *P*-value) associated with hazard of RFS corresponded to 10 genes: *TMCO4* (cg04738197), *LENG9* (cg12057190), *CDC42EP5* (cg12057190), *LNP1* (cg02540094), *TOMM70A* (cg02540094), *RUNX2* (cg08012149), *TBXAS1* (cg01584377), *SSH1* (cg16237760), *SFXN2* (cg08609163), and *COG3* (cg06172950). 1572 CpGs were associated with the increased hazard of RFS, and 956 CpGs were associated with the decreased hazard of RFS (Fig. 3B). The complete list of RFS-associated CpGs is provided in Additional files 4 and 5.

To gain a better understanding of the regions where the hazard-associated CpGs were located, we tested for enrichment of CpG island region context among CpG loci associated with a significant change in the hazard of RFS. We found that 1572 CpGs associated with the increased

hazard of RFS were significantly enriched in the open sea (OR = 1.14, 95% CI = 1.03–1.27) and were significantly depleted in CpG island S Shore regions (OR = 0.82, 95% CI = 0.67–0.99). The 956 CpGs associated with a decreased hazard of RFS were significantly enriched in CpG island N Shore regions (OR = 1.30, 95% CI = 1.06–1.57) and were significantly depleted in CpG island (OR = 0.60, 95% CI = 0.46–0.78) (Fig. 4A). We also tested for enrichment of other gene regulatory regions among CpG loci associated with a significant change in the hazard of RFS.

Location of CpGs in regulatory regions also was investigated and the CpGs associated with the increased hazard of RFS were significantly enriched in enhancer regions (OR = 1.78, 95% CI = 1.43–2.18), DNase hypersensitive sites (DHS) (OR = 1.25, 95% CI = 1.13–1.40), 5'UTR regions (OR = 1.37, 95% CI = 1.17–1.60) and gene body (OR = 1.15, 95% CI = 1.04–1.27); however, these CpGs were significantly depleted in regions 200–1500 bps upstream of the transcription start site (TSS1500) (OR = 0.85, 95% CI = 0.73–0.99). In addition, while the CpGs associated with a decreased hazard were strongly enriched in enhancer regions (OR = 4.18, 95% CI = 3.44–5.05) and DHS (OR = 2.46, 95% CI = 2.13–2.86), they were depleted for TSS200 (OR = 0.70, 95% CI = 0.48–0.99), gene body (OR = 0.85, 95% CI = 0.74–0.97) and transcription factor binding sites (TFBS) (OR = 0.68, 95% CI = 0.55–0.83) (Fig. 4B). Using the CpGs associated with hazard of RFS from cell-type unadjusted models in tests for enrichment gave results that were largely consistent with those above based on CpGs from fully adjusted models (Additional file 7: Figure S3).

To further understand the biological function of the hazard-associated CpGs, Gene Set Enrichment Analysis (GSEA) Molecular Signature Database (MSigDB) was used to explore the potential gene sets which might associate with the tumor recurrence or death of NMIBC patients. The input was 2528 RFS associated CpGs from the Cox model EWAS adjusting for immune cell composition. In the top 10 hazard-associated gene sets in gene ontology (GO) terms, some gene sets were related to neurological system processing (Additional file 7: Figure S4A), and most related genes were associated with the increased hazard of RFS in NMIBC (Additional file 7: Figure S4B). Since mdNLR was associated with the hazard of recurrence or death of NMIBC, we were interested

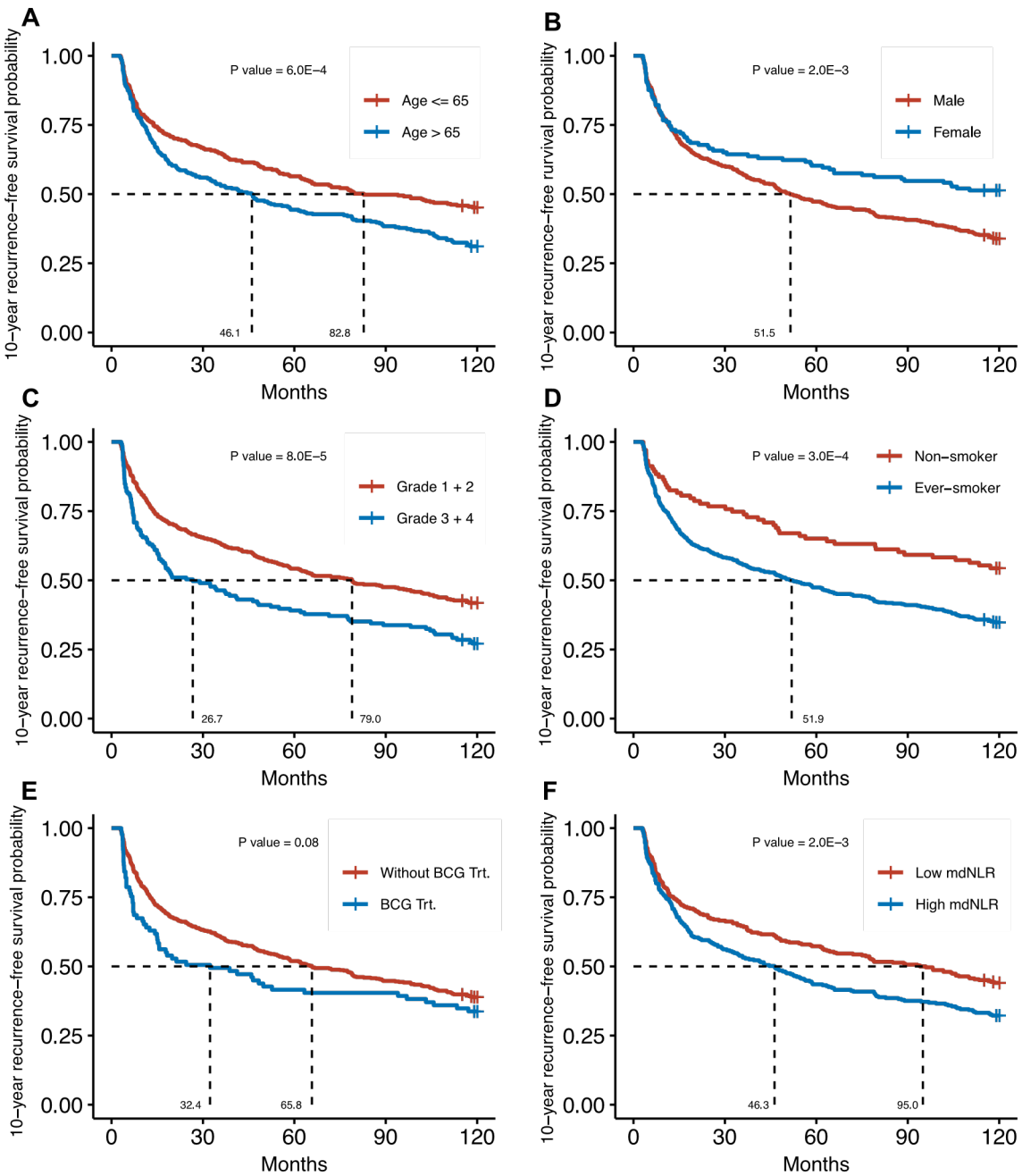


Fig. 2 Kaplan–Meier analysis of 10-year recurrence-free survival (RFS). 10-year RFS curves stratified by **A** age, **B** sex, **C** tumor grade, **D** smoking status, **E** BCG treatment status or **F** mdNLR level. *P*-values for log-rank tests are shown

Table 2 Cox proportional hazards 10-year recurrence-free survival models

	n (%) event[#]	*n* (%) no-events[#]	Event occurrence months		Multivariable[$] model	
			Mean	Median	HR (95% CI)	*P* value
Age	373 (61.9)	230 (38.1)	65.6	62.0	1.01 (1.00–1.03)	0.024
Tumor grade						1.0E−3
1 + 2	263 (58.2)	189 (41.8)	69.9	79.0	Referent group	
3 + 4	110 (72.8)	41 (27.2)	52.8	26.7	1.48 (1.17–1.87)	
Smoking status		56 (54.4)				
		174 (34.8)				
		230 (38.1)				
		lation-derived neutrophil to ly				
			iphocyte ratio			

HR: hazard ratio, CI: confidence interval, mdNLR: meth

Stratification was used on sex and BCG treatment status

[$] The model controlling for age, sex, tumor grade, smoki

[*] Winsorization was used on the top 2% values for fitting

[#] Initial recurrence or the death whose cause was unkno

Table 3 Cox proportional hazards 10-year o
models

	n (%) deceased	*n* (%) alive	Survival months	Multivariable model	
			Mean[#]		
				HR (95% CI)	*P* value
Age	180 (29.9)	423 (70.1)	104.3	1.07 (1.04–1.08)	8.7E−9
Sex					
Male	156 (34.1)	301 (65.9)	102.2	Referent group	
Female	24 (16.4)	122 (83.6)	111.1	0.62 (0.40–0.96)	0.032
Tumor grade					
1 + 2	118 (26.1)	334 (73.9)	106.6	Referent group	
3 + 4	62 (41.1)	89 (58.9)	97.7	1.54 (1.12–2.12)	8.0E−3
Smoking status					
Non-smoker	19 (18.4)	84 (81.6)	110.8	Referent group	
Ever-smoker	161 (32.2)	339 (67.8)	103.0	1.66 (1.03–2.67)	0.039
BCG treatment					
No	152 (29.6)	362 (70.4)	104.8	Referent group	
Yes	28 (31.5)	61 (68.5)	101.9	1.04 (0.68–1.60)	0.842
mdNLR*	180 (29.9)	423 (70.1)	104.3	1.46 (1.32–1.60)	6.0E−15

HR: hazard ratio, CI: confidence interval, mdNLR: methylation-derived neutrophil to lymphocyte ratio

[*] Winsorization was used on the top 2% values for fitting linearity assumption

All covariates modeled met proportionality assumptions [#]: the median survival month for each variable is 120

in immune-related gene sets. We checked the immunologic signature gene set, and the top 10 gene sets and their genes were related to immune cell regulation. However, only one gene set (BCELL VS MDC UP: up-regulated genes in B cells compared with myeloid dendritic cells after vaccination for influenza) consisting of 41 genes was significantly associated with RFS (FDR < 0.05) (Additional file 7: Figure S4C-D). Results from GO term analyses using 27,575 CpGs from the cell-type unadjusted models identified 5 pathways related to immunologic regulation among the top 10 pathways. Further, in the immunologic signature gene set specifically, we observed the top 10 pathways were associated with monocytes and lymphocytes (Additional file 7: Figure S5).

Locus overlap analysis (*LOLA*) was used to test the enrichment of CpGs in genomic regions. As our analysis was conducted on blood samples, results focus on the genomic regions within hematopoietic stem cells. When controlling for immune cell profiles, the 2528 CpGs associated with the hazard of RFS in NMIBC patients were most significantly enriched in Histone H3 acetylated at lysine 9 and 14 (H3K9K14ac) (Q-value = 1.9 × 1 0^{-20}) (Additional file 7: Figure S6A). *LOLA* results for the 27,575 CpGs associated with the hazard of RFS from cell type unadjusted models were most significantly enriched in cistrome of the promyelocytic leukemia protein (PML) (Q-value < 0.05) (Additional file 7: Figure S6B).

We also ran a differentially methylated regions analysis using *DMRcate*. In this study, we found 11 CpGs in a specific genome region overlapping with the gene *BLCAP*. NMIBC patients without tumor recurrence or death within 10 years had a higher mean of methylation levels for this region compared with NMIBC patients with tumor recurrence or death within 10 years, and one CpG in this region was significant (FDR = 2.63 × 10^{-14}) (Additional files 6, 7: Figure S7).

Discussion

In this study, we tested whether immune profiles and epigenetic features are associated with NMIBC recurrence. Although previous work observed the association between NLR and overall survival in NMIBC patients, it was in a smaller study sample and used DNA methylation data from a dated (second generation) array platform with ~ 27,000 CpGs. In this study, our sample size was nearly three times larger and DNA methylation data was collected using the current genome-scale platform (fourth generation), measuring ~ 860,000 CpGs for which an optimized cell-type deconvolution library

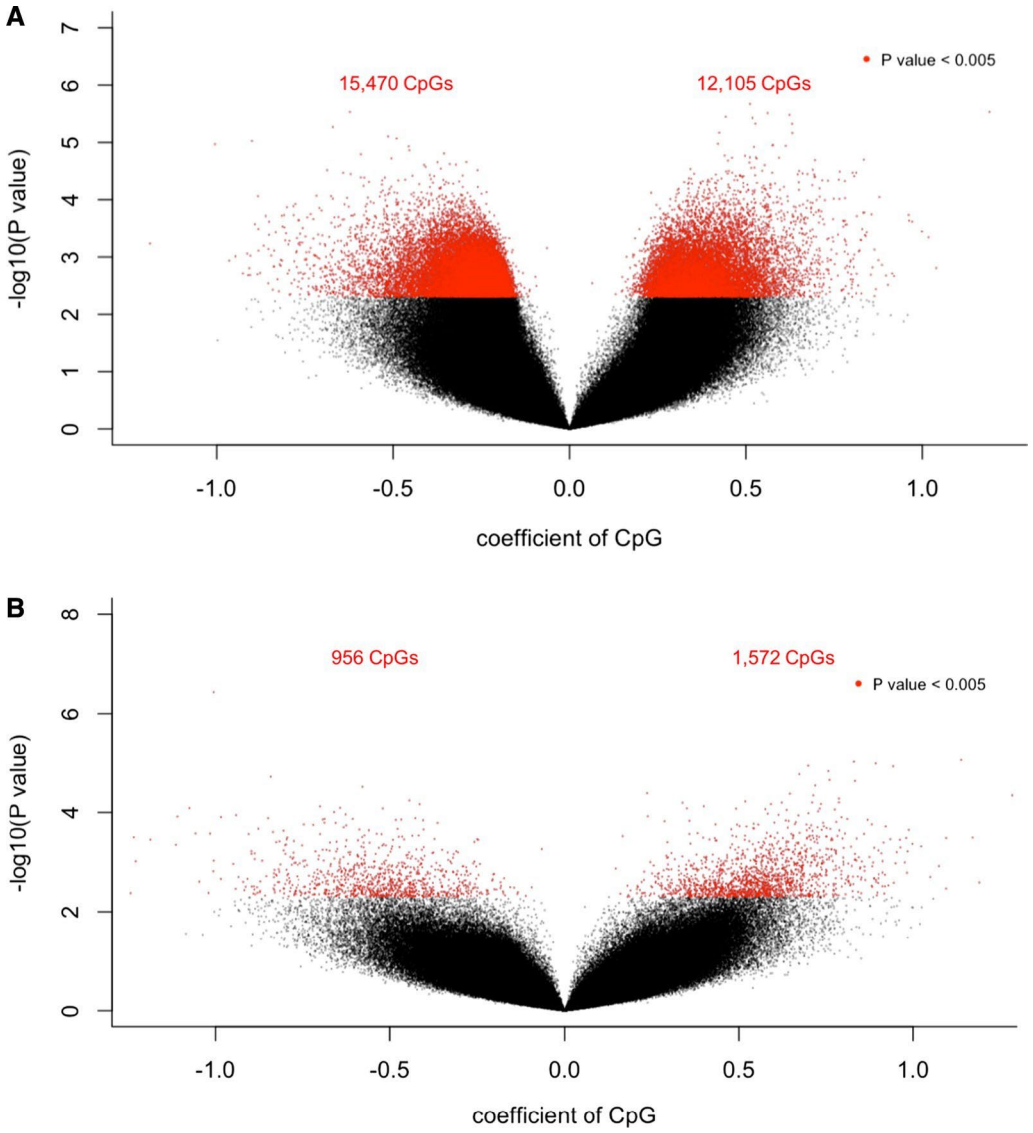


Fig. 3 Volcano plots of recurrence-free survival (RFS) associated CpGs from the epigenome-wide association study (EWAS) analyses. The Cox multivariable model that was fitted in EWAS was shown in each plot. CpGs are colored in red (**A**) 12,105 CpGs were associated with the increased hazard of NMIBC 10-year RFS, and 15,470 CpGs were associated with the decreased hazard of NMIBC 10-year RFS. (**B**) 1572 CpGs were associated with the increased hazard of NMIBC 10-year RFS, and 956 CpGs were associated with the decreased hazard of NMIBC 10-year RFS

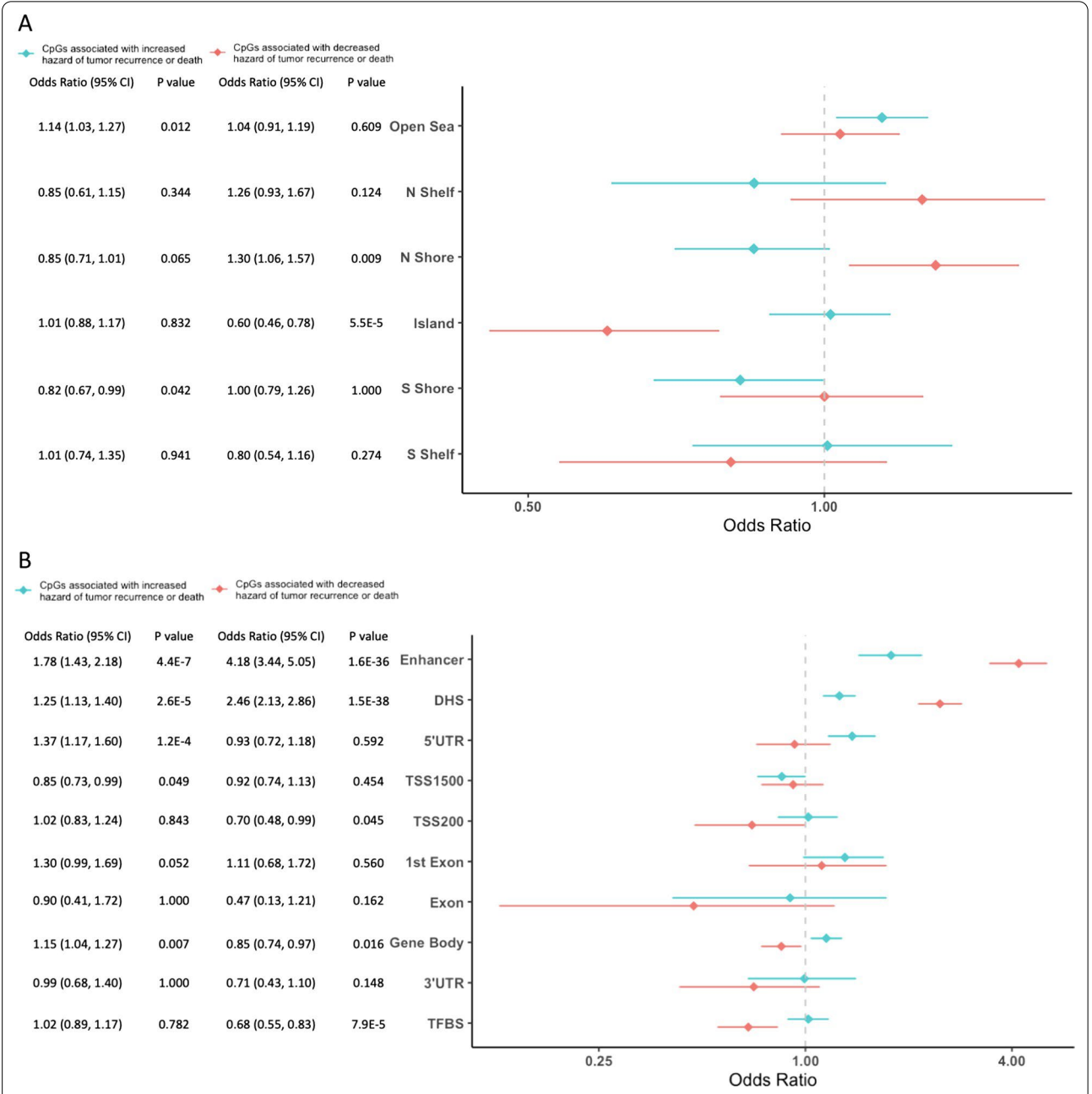


Fig. 4 Genomic context enrichment analysis of CpG sites whose methylation state is significantly associated with recurrence-free survival. Enrichment analysis of (**A**) relation to CpG island and (**B**) genomic context of NMIBC recurrence-free survival associated CpGs. The 2528 CpGs from EWAS ($P < 0.005$) were tested for enrichment versus all modeled CpGs. The bar represents the 95% confidence interval. Mantel–Haenszel was used to test RFS-associated CpGs enrichment of CpG island-related gnome context. An odds ratio larger than 1 means enrichment, and an odds ratio smaller than 1 indicates depletion

exists to determine highly accurate immune cell-type proportions.

We extended tests of association with the patient outcomes beyond the methylation-derived neutrophil-tolymphocyte ratio (mdNLR) to include leukocyte-specific cell-type proportions. Our findings suggest that elevated mdNLR increased the hazard of RFS in NMIBC patients. These findings are consistent with previous studies demonstrating that NLR was significantly higher in highrisk NMIBC patients [27], and increased NLR was positively associated with poor prognosis [28, 29]. While past studies have shown a significant association between NLR and outcomes in NMIBC patients using the conventional method of CBC tests [30–32], this study uses NLR derived from blood methylation profiles. Compared with flow cytometry, estimation using the differentially methylated regions (DMR) library has several advantages: it does not require long sample processing time, large volume of blood, or intact cells as DNA and 5-methylcytosine are both stable [33]. The advantage of our approach is the ability to use archived samples that many investigators may already have. Moreover, utilizing blood samples to monitor patient outcomes is less invasive compared with a cystoscopy, the routine screening method. For NMIBC patients with a high risk of tumor recurrence, BCG is the standard intravesical immunotherapy to induce immune system eliminating bladder cancer cells that might be left after surgery [8, 34, 35], and therefore, blood immune profile is a potential prognostic factor. Immune profiling with DNA

methylation data is a promising avenue for assessing NMIBC prognosis.

As NLR is composed of lymphocyte and neutrophil proportions, we also considered the association between each methylation-derived immune cell type proportion and patient outcomes. Interestingly, increased CD4T and CD8T cell proportions were associated with decreased NMIBC recurrence-free survival and overall survival. Prior work has shown that NMIBC patients with a high CD4T cell count in BCG pretreatment microenvironment have a significantly prolonged recurrence-free survival compared to patients with a low CD4T cell count [35]. In addition to the immune cell types explored in this study, other cell types have been shown to affect NMIBC development, such as GATA3 + T cells, regulatory T cells, and tumor-associated macrophages [35]. Other work showed an increased C D8+ILT2+ T cell proportion was associated with a significantly increased hazard of NMIBC recurrence [36]. Further, peripheral neutrophil x platelet/lymphocyte was inversely correlated with highrisk NMIBC recurrence-free survival [37]. These results demonstrate the potential of immune cell profiles in evaluating the prognosis of NMIBC patients, and future work mapping DNA methylation profiles of additional immune cell types and states could add detail to investigations of immune profiles and bladder cancer outcomes. In addition, peripheral immune cell distribution is affected by potentially residual confounding factors such as, infection [38], inflammation [39], lifestyles, treatments [40, 41], obesity [42], chronic alcohol consumption [43], and type 2 diabetes [44]. Since our study includes hundreds of subjects and has a decade or more of follow-up time, in the future, we will re-explore the effect of these residual confounding factors on the association of circulating immune cell distribution and NMIBC outcomes. In the future, we will perform higher resolution methylation cell mixture deconvolution to resolve additional immune cell types and employ blood count on prospectively collected samples. Together, these data will allow us to understand if specific subsets of neutrophils or lymphocytes contribute to high NLR in patients with poor outcomes.

Bacillus Calmette–Guérin (BCG) is the most commonly used immunotherapy for high-risk NMIBC. After transurethral resection of bladder tumors, NMIBC patients may receive BCG intravesical therapy to induce an immune response in the bladder to attack cancer cells. Previous studies have shown that tumor immune environments may interact with BCG and interfere with the efficacy of this therapy. For instance, IL-12 secreted by BCG-induced monocytes was increased in NMIBC patients without tumor recurrence, a phenomenon that may involve the innate immune memory of circulating monocytes [45]. Nevertheless, patients receiving BCG in our study did not have peripheral immune cell profiles that differed from patients who were BCG naïve, and BCG was not significantly associated with NMIBC outcomes. As the number of patients with BCG treatment was relatively low (N = 89; 14.8%) and limited power to observe possible therapy induced changes in peripheral immune profiles, future work is needed to address the potential associations of BCG with peripheral immune profiles and patient outcomes. Most BCG-treated patients were already high-risk. In addition, blood samples were collected only after surgery or initial BCG treatment. To assess the changes in immune profiles over time, having multiple blood draws is necessary for future work.

This EWAS identified several CpGs associated with NMIBC recurrence-free survival when controlling for stratified sex, age, tumor grade, smoking status, stratified BCG receiving status, and immune cell profiles. With a P-value < 0.005, 2528 CpGs were found to be associated with the hazard of RFS.

Among the top CpGs whose methylation was associated with NMIBC recurrence or death in the fully adjusted models, some genes have been previously associated with bladder cancer. Slingshot homolog-1 (*SSH1*) had a positive association with tumor grade, tumor invasion, and tumor recurrence of bladder cancer patients [46]. Runt-related transcription factor 2 (*RUNX2*) is a key factor of osteoblast differentiation and has been reported to be associated with epithelial-mesenchymal transition in bladder tumors. Furthermore, *RUNX2* could predict early recurrence in bladder cancer patients with high accuracy [47, 48]. Similar to past studies, *SSH1* and *RUNX2* were associated with the increased risk of tumor recurrence in our study. In addition, NMIBC patients with tumor recurrence within ten years had significantly higher methylation levels in the gene body of these two genes compared with patients without tumor recurrence. When not adjusting for immune cell profiles, 27,575 CpGs were associated with the decreased hazard of RFS. The top 10 CpGs with the most significant P-value corresponded to 6 genes: *BCL11A* (cg24361098), *TMCO4* (cg04738197), *MICALCL* (cg01518090), *GRAP2* (cg21012238), *TRAM2* (cg15085626), and *KIRREL* (cg10570484). While these genes have not been reported to be associated with bladder cancer outcomes, they have been reported to be involved in mechanisms promoting cancer development in tumors [49–52] or blood [53, 54]. Although we

measured blood methylation in this study, we plan to measure tumor methylation and will explore the association of CpGs in these genes with NMIBC outcomes. What was more interesting is that the model, with or without controlling for immune cell profiles, led to different results. While immune cell profiles are usually not adjusted in the Cox model, we controlled for immune cell profiles since the immune system plays a key role in tumor development. This study presents a new perspective to demonstrate the difference between models with or without adjusting for immune profiles, indicating the need for further investigation on the involvement of immune profiles in outcome analyses.

Through DMR analysis, we found NMIBC patients with tumor recurrence or death within 10 years had a lower methylation level in the *BLCAP* region compared with patients without tumor recurrence or death. The bladder cancer-associated protein (*BLCAP*) gene encodes a protein that stimulates apoptosis. It has been reported that loss of protein expression is associated with bladder tumor progression, and the application of staining patterns for this protein could be a potential biomarker in bladder cancer [55]. In functional analysis, strong nuclear expression of *BLCAP* was associated with expression of p-*STAT3* and overall poor disease outcome. Additionally, *BLCAP* was discovered to interact with *STAT3* physically and may involve the *STAT3*-mediated progression of precancerous lesions to invasive bladder tumors [56]. These results were consistent with our finding that NMIBC patients with poor outcomes had lower methylation levels in the *BLCAP* region. Since the model we used for DMRs analysis was adjusted for immune cell profiles, we are curious whether immune cells may play roles in the interaction between *BLCAP* and *STAT3* and will investigate this in the future. The CpG site with significantly lower methylation (cg10642330) is located in the 5' UTR of *BLCAP* and the gene body of *NNAT*. The methylation beta value of this CpG site in *NNAT* had been reported significantly higher in prostate cancer tissue relative to adjacent normal tissues [57]. Though no CpG site in the *BLCAP* region was significantly associated with the hazard of tumor recurrence or death in the adjusted EWAS, five *BLCAP* CpGs (cg26083330, cg23757721, cg13790727, cg03061677, and cg04489586) were associated in the model not controlling for immune cell proportions (not including the DMR analysis site cg10642330). In the future, we will investigate the relation of *BLCAP* tumor methylation with survival outcomes.

Our results reveal several features of peripheral blood immune profiles that are associated with outcomes in NMIBC patients. We showed that higher CD4 or CD8 proportions were associated with decreased hazard of recurrence or death and further established that high NLR is associated with an increased hazard of RFS. The EWAS portion of our study also points to epigenetic reprogramming within the immune compartment being involved in tumor recurrence of NMIBC patients. Future study in a prospective setting will assess the clinical utility of incorporating methylation in predicting hazard of recurrence and shaping recommendations for disease surveillance. In addition to immune cells, future work examining cell-type proportions in tumor microenvironments of NMIBC patients is needed to understand the relationship between peripheral immune profiles with tumor-infiltrating immune profiles and patient outcomes. This work contributes to our understanding of associations of methylation-derived immune profiles and NMIBC patient outcomes and could further contribute to developments in epigenetic biomarkers of cancer.

Conclusions

Here we demonstrate the associations of non-muscleinvasive bladder cancer outcomes with immune profiles. In addition, we identify preliminary evidence of discrete and regional CpG methylation associations with bladder cancer outcomes. The findings could contribute to developing epigenetic biomarkers for recurrence-free survival in non-muscle-invasive bladder cancer.

Methods

Study subjects and samples

The subjects and data used in this study are described in more detail in prior publications [58–60]. Briefly, subjects were recruited from all three phases of a New Hampshire population-based bladder cancer case–control study [61]. The first wave of this study (phase 1) collected blood samples from 331 individuals diagnosed with incident bladder cancer between July 1994 and June 1998. The second study wave (phase 2) collected blood samples from 243 individuals diagnosed between July 1998 and December 2001. Finally, the third study wave (phase 3) obtained blood samples from 194 individuals recruited and diagnosed between July 2002 and December 2004. All the subjects were identified using the New Hampshire State Cancer Registry, hospital pathology departments, and hospital cancer registries, and all blood samples were collected after the time of diagnosis (time range: 20–1790

days). Among patients, 40 patients received BCG treatment in phase 1, 29 patients received BCG in phase 2, and 19 patients received BCG in phase 3. All patients with BCG treatment had blood drawn after receiving BCG (time range: 7–1542 days). An outline of data filtering and inclusion/exclusion criteria applied to these data are shown in Fig. 1. Briefly, subjects without muscle-invasive status, histopathology re-review, tumor grade, smoking status, or pack-years were removed from the study. Subjects that withstood the aforementioned exclusion criteria were retained and used in downstream statistical analyses.

DNA extraction, quantification, and bisulfite modification

Each blood sample was maintained at 4C and frozen within 24 h of blood draw. A hundred µl buffy coat was used to extract DNA. The QIAMP DNA blood & Tissue kit was used to extract DNA from blood samples according to the manufacturer's protocol. Extracted DNA was quantified by using the Qubit 3.0 Fluorometer. After bisulfite modification, an established 5 mC microarray protocol optimized for Illumina methylation arrays was used to determine the genome-wide 5mC profile. DNA samples were subjected to bisulfite conversion (according to the manufacturer's protocol of the Zymo EZ DNA methylation Kit) with an input of 750 ng per sample and whole-genome amplified prior to array hybridization. Recovered substrate ssDNA were submitted for DNA methylation array processing.

DNA methylation data

Bisulfite-modified DNA samples were measured for their DNA methylation status using the MethylationEPIC array, which interrogates > 860,000 CpG sites. Probe intensity data (iDAT files) from the EPIC methylation array were processed for quality control via the R package *minfi* [62] and *ENmix* [63] in R version 3.6. After quality control, 11 samples with low-quality CpG values or bisulfite intensity (threshold: 7000) were excluded from the study. The data were then normalized and conducted background correction through preprocessNoob procedure from *minfi*. The *ComBat* [64] was used to adjust for potential batch effect. Probes with a detection $P > 1.0 \times 10^{-6}$ in more than 10% of the samples were excluded (32,414). Also, 98,826 probes, which are crossreactive, SNP-associated, and non-CpG (CpH) methylation [65], as well as 17,120 probes on sex chromosomes, were excluded. In total, 726,856 probes were used in downstream statistical analyses in downstream statistical analyses. *IlluminaHumanMethylationEPICanno.ilm10b4.hg19* [66] was used to annotate CpG sites. Relation to CpG island was defined by the "*Relation_to_Island*" as in the Illumina annotation used in the genomic context analysis. '5'UTR,' 'Exon,' 'Gene Body,' and '3'UTR' contexts were defined by having '5UTR,' 'ExonBnd,' 'Body' and '3UTR' in *UCSC_RefGene_Group*. 'Enhancer' context was defined by having a record in the *Phantom5_Enhancers*. 'DHS' context was defined by finding a record in the *DNase_Hypersensitivity_NAME*. 'TFBS' context was defined by having a record in the *TFBS_NAME*.

Statistical analysis

The estimation of cell-type proportions was processed through the *estimateCellCounts2* from the *FlowSorted.Blood.EPIC* package in Bioconductor (version 3.9; [18]). Methylation-derived neutrophil-to-lymphocyte ratio (mdNLR) was calculated by performing cell-mixture deconvolution to estimate the proportion of leukocyte subtypes and the ratio of neutrophil proportion to lymphocyte proportion was then computed. Individual leukocyte cell-type proportions and mdNLR were included in outcome analyses as continuous variables for Cox proportional hazard regression. In addition, mdNLR was dichotomized based on the median mdNLR for the Kaplan–Meier method.

Ten-year overall survival was defined as the time interval from the date of initial diagnosis to death within 10 years (all deaths were related with bladder cancer). Subjects who were alive or lost to follow-up were censored at the last follow-up. Ten-year recurrence-free survival was defined as the time interval from the date of initial diagnosis to the first tumor recurrence or death (all causes), whichever occurred first within 10 years. Patients alive and free of the disease or lost to followup were censored at the last follow-up. For both overall and recurrence-free survival, patient survival times over 10 years were truncated at 10 years, and patients were censored if the first tumor recurrence or death occurred after 10 years. The median survival times for the two survival outcomes were estimated using the Kaplan–Meier method. In multivariable analyses, Cox proportional hazard regression models were used to examine the association of each variable on bladder cancer outcomes and were fit via *coxph* in the *survival* R package. The proportional hazards assumption was tested by using *cox.zph* from the *survival* R package. The *cox.zph* function tests the proportionality of all the predictors in Cox models by creating interactions with time. As sex and BCG treatment status violated the proportional hazards assumption, stratification on both variables was included. The linearity assumption was

examined via *ggcoxfunctional* from the R *survminer* package, and methylation-derived immune cell profiles were found to violate the linearity assumption. Hence, winsorization was used on methylation-derived immune cell profiles. The winsorization cutpoint of each immune cell profile is shown in Additional file 7: Figure S1. A *P*-value of < 0.05 was the significance threshold on multivariable analysis. Cox model results were presented using the *stargazer* R package.

Epigenome-wide association study (EWAS), enrichment analysis, and differentially methylated regions (DMRs) An epigenome-wide association study (EWAS) was performed using ewaff R package (https:// github. com/ peris hky/ ewaff) to investigate the association of CpG-specific DNA methylation and bladder cancer recurrence. We fit Cox proportional hazards models independently to each CpG, controlling for age, sex, tumor grade, smoking status, Bacillus Calmette–Guérin (BCG) receiving status, and immune cell profiles. Since all *P*-values adjusted for false discovery rate from EWAS results are higher than 0.05, we relaxed the threshold for genomic context and enrichment analyses using a *P*-value of < 0.005.

For CpG sites associated ($P < 0.005$) with bladder tumor recurrence, we examined whether those CpGs were enriched in CpG island-related genomic context or regulatory regions via using Mantel–Haenszel tests, adjusted for Illumina probe type to eliminate the difference in distributions on genomic context. CpG island-related genomic context includes open sea, north shelves, north shores, islands, south shores, and south shelves. Regulatory regions include enhancers, DNase hypersensitivity sites, 5'UTR, TSS1500, TSS200, 1st Exon, Exon, gene body, 3'UTR, and transcription factor binding sites. Next, the Locus Overlap analysis (*LOLA*) R package in Bioconductor (version 1.20.0; [67]) was used to investigate enrichment of genomic regions limited to tissue equal to "hematopoietic stem cell." Finally, *gometh* and *gsameth* from the *missMethyl* package in Bioconductor (version 1.6.2; [68]) were used to test the enrichment of gene sets for Gene Ontology (GO) terms and the C7: immunologic signature gene set in the Gene Set Enrichment Analysis (GSEA) Molecular Signature Database (MSigDB) *cnetplot* from the *enrichplot* package in Bioconductor (version 1.10.1; https:// yulab- smu. top/ biome dical- knowl edge- mining- book/) was used to plot gene-concept network plots.

Differentially methylated regions were identified and extracted through the *dmrcate* and *extractRanges* from the *DMRcate* R package [69]. The inputs were logit-transform of beta values (M-values). The phenotype of interest for comparison was "NMIBC patients with tumor recurrence within ten years or not" in our designed model. In addition, the designed model was adjusted for sex, age, tumor grade, smoking status, BCG receiving status, and immune cell profiles. We relaxed the threshold for DMRs analysis using an FDR-corrected *P*-value of < 0.1. Then, the *visualizeGene* from the *sesame* package [70] was used for observing the methylation levels of CpGs in regions identified by *DMRcate*.

Abbreviations

NMIBC: Non-muscle-invasive bladder cancer; NLR: Neutrophil-to-lymphocyte ratio; mdNLR: Methylation-derived neutrophil-to-lymphocyte ratio; EWAS: Epigenome-wide association study; BCG: Bacillus Calmette–Guérin; CBC: Complete blood count differential; DMRs: Differentially methylated regions; RFS: Recurrence-free survival; OS: Overall survival; HR: Hazard ratio; OR: Odds ratio; CI: Confidence interval; CpG: 5'-Cytosine-phosphate-guanine-3'; DHS: DNase hypersensitive sites; TFBS: Transcription factor binding sites; FDR: False discovery rate; LOLA: Locus overlap analysis; GO: Gene ontology; GSEA: Gene set enrichment analysis; MSigDB: Molecular signature database.

Supplementary Information

The online version contains supplementary material available at https:// doi. org/ 10. 1186/ s13148-0 22-0 1234-6.

Additional file 1: Cox proportional hazards 10-year recurrence-free survival. Created via using *stargazer()* function

Additional file 2: Cox proportional hazards 10-year overall survival. Created via using *stargazer()* function

Additional file 3: The 27,575 CpGs associated with the hazard of 10-year RFS when the model without adjusting for immune cell profiles. The list of the 27,575 CpGs which were associated with the hazard of 10-year RFS from the EWAS with Cox model without adjusting for immune cell profiles.

Additional file 4: The 1,572 CpGs associated with the increased hazard of 10-year RFS. The list of the 1,572 CpGs which were associated with the increased hazard of 10-year RFS from the EWAS with Cox model adjusting for immune cell profiles.

Additional file 5: The 956 CpGs associated with the decreased hazard of 10-year RFS. The list of the 956 CpGs which were associated with the decreased hazard of 10-year RFS from the EWAS with Cox model adjusting for immune cell profiles.

Additional file 6: The results of differentially methylated regions. **no.cpgs**: Number of CpG sites constituting the significant region.

Additional file 7: Supplemental tables and figures.

Acknowledgements
Not applicable

Authors' contributions
JC, BCC, and LAS contributed to the analysis. JC, BCC, LAS, DCK, KTK, JKW, AMM, and MRK contributed to the conception, design, and writing of the manuscript. ASA and JDS contributed to the data collection. All authors read and approved the final manuscript.

Funding
Work at Dartmouth College was supported by the National Institutes of Health Grant Numbers R01CA216265, P20GM104416, R01CA057494, R01CA253976, and CDMRP/Department of Defense W81XWH-20–1-0778. Work at the University of California, San Francisco was supported by R01CA207360, P50CA097257, JKW is supported by the Robert Magnin Newman Endowed Chair in Neuro-oncology.

Availability of data and materials
The datasets generated and analyzed during this current study are available in the gene expression omnibus repository at GSE183920.

Declarations

Ethics approval and consent to participate This case study was approved by the Dartmouth IRB.

Consent for publication Not applicable.

Competing interests

The authors declare that they have no competing interests. JKW and KTK are co-founders of Cellintec which had no role in this work.

Author details

[1] Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Lebanon, NH 03766, USA. [2] Department of Neurological Surgery, University of California San Francisco, San Francisco, CA 94143, USA. [3] Department of Biostatistics and Data Science, University of Kansas Medical Center, Kansas City, KS 66160, USA. [4] Department of Neurology, Geisel School of Medicine, Dartmouth College, Lebanon, NH 03766, USA. [5] Department of Surgery, Section of Urology, Geisel School of Medicine, Dartmouth College, Lebanon, NH 03766, USA. [6] Departments of Epidemiology and Pathology and Laboratory Medicine, Brown University, Providence, RI 02912, USA. [7] Departments of Molecular and Systems Biology, and Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH 03766, USA. [8] Dartmouth Hitchcock Medical Center, 1 Medical Center Dr, 660 Williamson Translation Research Building, Lebanon, NH 03756, USA.

Received: 15 September 2021 Accepted: 12 January 2022
Published online: 21 January 2022

References

1. Atlanta: American Cancer Society. Cancer Facts & Figures 2020. Am. Cancer Soc. 2020.
2. Martinez Rodriguez RH, Buisan Rueda O, Ibarz L. Bladder cancer: present and future. Med Clín. 2017;149:449–55. https:// doi. org/ 10. 1016/j. medcle. 2017. 10. 005.
3. Svatek RS, Lotan Y. Is there a rationale for bladder cancer screening? Curr Urol Rep. 2008;9:339–41.
4. Larré S, Catto JWF, Cookson MS, Messing EM, Shariat SF, Soloway MS, et al. Screening for bladder cancer: rationale, limitations, whom to target, and perspectives. Eur Urol. 2013;63:1049–58.
5. Degeorge KC, Holt HR, Hodges SC. Bladder cancer: diagnosis and treatment. 2017.
6. Hall MC, Chang SS, Dalbagni G, Pruthi RS, Seigne JD, Skinner EC, et al. Guideline for the management of nonmuscle invasive bladder cancer (stages Ta, T1, and Tis): 2007 update. J Urol. 2007;178:2314–30.
7. Mossanen M, Gore JL. The burden of bladder cancer care: direct and indirect costs. Curr Opin Urol. 2014;24:487–91.
8. Lim CJ, Nguyen PHD, Wasser M, Kumar P, Lee YH, Nasir NJM, et al. Immunological hallmarks for clinical response to BCG in bladder cancer. Front Immunol. 2021;11:1–13.
9. Ogihara K, Kikuchi E, Yuge K, Yanai Y, Matsumoto K, Miyajima A, et al. The preoperative neutrophil-to-lymphocyte ratio is a novel biomarker for predicting worse clinical outcomes in non-muscle invasive bladder cancer patients with a previous history of smoking. Ann Surg Oncol. 2016;23:1039–47.
10. Tan YG, Eu E, Lau Kam On W, Huang HH. Pretreatment neutrophil-tolymphocyte ratio predicts worse survival outcomes and advanced tumor staging in patients undergoing radical cystectomy for bladder cancer. Asian J Urol. 2017;4:239–46. https:// doi. org/ 10. 1016/j. ajur. 2017. 01. 004.
11. Kang M, Jeong CW, Kwak C, Kim HH, Ku JH. Preoperative neutrophillymphocyte ratio can significantly predict mortality outcomes in patients with non-muscle invasive bladder cancer undergoing transurethral resection of bladder tumor. Oncotarget. 2017;8:12891–901.
12. Yoshida T, Kinoshita H, Yoshida K, Mishima T, Yanishi M, Inui H, et al. Prognostic impact of perioperative lymphocyte–monocyte ratio in patients with bladder cancer undergoing radical cystectomy. Tumor Biol. 2016;37:10067–74. https:// doi. org/ 10. 1007/ s13277-0 16-4 874-8.
13. Ma JY, Hu G, Liu Q. Prognostic significance of the lymphocyte-to-monocyte ratio in bladder cancer undergoing radical cystectomy: a metaanalysis of 5638 individuals. Dis Mark. 2019;2019.
14. Bhindi B, Hermanns T, Wei Y, Yu J, Richard PO, Wettstein MS, et al. Identification of the best complete blood count-based predictors for bladder cancer outcomes in patients undergoing radical cystectomy. Br J Cancer. 2016;114:207–12. https:// doi. org/ 10. 1038/ bjc. 2015. 432.
15. Ojerholm E, Smith A, Hwang W, Baumann BC, Tucker KN, Lerner SP, et al. Neutrophil-to-lymphocyte ratio as a bladder cancer biomarker: assessing prognostic and predictive value in SWOG 8710. Cancer. 2018;123:794–801.
16. Dixon LR. The complete blood count: physiologic basis and clinical usage. J Perinat Neonatal Nurs. 1997;11:1–18.
17. Accomando WP, Wiencke JK, Houseman EA, Nelson HH, Kelsey KT. Quantitative reconstruction of leukocyte subsets using DNA methylation. Genome Biol. 2014;15:1–12.
18. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19:64.
19. Suelves M, Carrió E, Núñez-Álvarez Y, Peinado MA. DNA methylation dynamics in cellular commitment and differentiation. Brief Funct Genom. 2016;15:443–53.
20. Salas LA, Wiencke JK, Koestler DC, Zhang Z, Christensen BC, Kelsey KT. Tracing human stem cell lineage during development using DNA methylation. Genome Res. 2018;28:1285–95.
21. Baron U, Türbachova I, Hellwag A, Eckhardt F, Berlin K, Hoffmuller U, et al. DNA methylation analysis as a tool for cell typing. Epigenetics. 2006;1:55–60.
22. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform. 2012;13.
23. Li Y, Pan X, Roberts ML, Liu P, Kotchen TA, Cowley AW, et al. Stability of global methylation profiles of whole blood and extracted DNA under different storage durations and conditions. Epigenomics. 2018;10:797–811.
24. Titus AJ, Gallimore RM, Salas LA, Christensen BC. Cell-type deconvolution from DNA methylation: a review of recent applications. Hum Mol Genet. 2017;26:R216–24.
25. Wiencke JK, Koestler DC, Salas LA, Wiemels JL, Roy RP, Hansen HM, et al. Immunomethylomic approach to explore the blood neutrophil lymphocyte ratio (NLR) in glioma survival. Clin Epigenet. 2017;9:1–11.
26. Koestler DC, Usset J, Christensen BC, Marsit CJ, Karagas MR, Kelsey KT, et al. DNA methylation-derived neutrophil-tolymphocyte ratio: an

epigenetic tool to explore cancer inflammation and outcomes. Cancer Epidemiol Biomarkers Prev. 2017;26:328–38.
27. Aydın M, Bitkin A, Kadıhasanoğlu M, İrkılata L, Akgüneş E, Keleş M, et al. Correlation of neutrophil-lymphocyte ratio and risk scores in non-muscle invasive bladder cancer. Actas Urol Españolas. 2019;43:503–8.
28. Yuk HD, Jeong CW, Kwak C, Kim HH, Ku JH. Elevated neutrophil to lymphocyte ratio predicts poor prognosis in non-muscle invasive bladder cancer patients: initial intravesical bacillus Calmette-Guerin treatment after transurethral resection of bladder tumor setting. Front Oncol. 2019;9:1–8.
29. Getzler I, Bahouth Z, Nativ O, Rubinstein J, Halachmi S. Preoperative neutrophil to lymphocyte ratio improves recurrence prediction of nonmuscle invasive bladder cancer. BMC Urol BMC Urology. 2018;18:1–10.
30. Vartolomei MD, Porav-Hodade D, Ferro M, Mathieu R, Abufaraj M, Foerster B, et al. Prognostic role of pretreatment neutrophil-to-lymphocyte ratio (NLR) in patients with non–muscle-invasive bladder cancer (NMIBC): a systematic review and meta-analysis. Urol Oncol Semin Orig Investig. 2018;36:389–99. https:// doi. org/ 10. 1016/j. urolo nc. 2018. 05. 014.
31. Zhang Q, Lai Q, Wang S, Meng Q, Mo Z. Clinical value of postoperative neutrophil-tolymphocyte ratio change as a detection marker of bladder cancer recurrence. Cancer Manag Res. 2021;13:849–60.
32. Cantiello F, Russo GI, Vartolomei MD, Farhan ARA, Terracciano D, Musi G, et al. Systemic inflammatory markers and oncologic outcomes in patients with high-risk non–muscle-invasive urothelial bladder cancer. Eur Urol Oncol. 2018;1:403–10.
33. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenomewide association studies: a review and recommendations. Epigenomics. 2017;9:757–68.
34. Witjes JA, Hendricksen K. Intravesical pharmacotherapy for non-muscleinvasive bladder cancer: a critical analysis of currently available drugs, treatment schedules, and long-term results. Eur Urol. 2008;53:45–52.
35. Pichler R, Fritz J, Zavadil C, Schäfer G, Culig Z, Brunner A. Tumor-infiltrating immune cell subpopulations influence the oncologic outcome after intravesical bacillus calmette-guérin therapy in bladder cancer. Oncotarget. 2016;7:39916–30.
36. Desgrandchamps F, LeMaoult J, Goujon A, Riviere A, Rivero-Juarez A, Djouadou M, et al. Prediction of non-muscle-invasive bladder cancer recurrence by measurement of checkpoint HLAG's receptor ILT2 on peripheral CD8+ T cells. Oncotarget. 2018;9:33160–9.
37. Akan S, Ediz C, Sahin A, Tavukcu HH, Urkmez A, Horasan A, et al. Can the systemic immune inflammation index be a predictor of BCG response in patients with high-risk non-muscle invasive bladder cancer? Int J Clin Pract. 2021;75:1–9.
38. Douglas RG, Alford RH, Cate TR, Couch RB. The leukocyte response during viral respiratory illness in man. Ann Intern Med. 1966;64:521–30.
39. Woodell-May JE, Sommerfeld SD. Role of inflammation and the immune system in the progression of osteoarthritis. J Orthop Res. 2020;38:253–7. https:// doi. org/ 10. 1002/ jor. 24457.
40. Bedel C, Korkut M, Armag˘an HH. NLR, d-NLR and PLR can be affected by many factors. Int Immunopharmacol. 2021;90:107154.
41. Lin BD, Hottenga JJ, Abdellaoui A, Dolan CV, De Geus EJC, Kluft C, et al. Causes of variation in the neutrophil-lymphocyte and platelet-lymphocyte ratios: a twin-family study. Biomark Med. 2016;10:1061–72.
42. Elisia I, Lam V, Cho B, Hay M, Li MY, Kapeluto J, et al. Exploratory examination of inflammation state, immune response and blood cell composition in a human obese cohort to identify potential markers predicting cancer risk. PLoS ONE. 2020;15:1–21.
43. Laso FJ, Vaquero JM, Almeida J, Marcos M, Orfao A. Chronic alcohol consumption is associated with changes in the distribution, immunophenotype, and the inflammatory cytokine secretion profile of circulating dendritic cells. Alcohol Clin Exp Res. 2007;31:846–54.
44. Grossmannm V, Schmitt VH, Zeller T, Panova-Noeva M, Schulz A, Laubert-Reh D, et al. Profile of the immune and inflammatory response in individuals with prediabetes and type 2 diabetes. Diabetes Care. 2015;38:1356–64.
45. Graham CH, Paré JF, Cotechini T, Hopman W, Hindmarch CCT, Ghaffari A, et al. Innate immune memory is associated with increased disease-free survival in bladder cancer patients treated with bacillus Calmette-Guérin. Can Urol Assoc J. 2021;15:1–13.
46. Luo Q, Liu Y, Zhao H, Guo P, Wang Q, Li W, et al. Slingshot homolog-1 expression is a poor prognostic factor of pT1 bladder urothelial carcinoma after transurethral resection. World J Urol. 2020;38:2849–56. https:// doi. org/ 10. 1007/ s00345- 020- 03092-4.
47. Abdelzaher E, Kotb AF. High coexpression of runt-related transcription factor. Appl Immunohistochem Mol Morphol. 2016;2:24.
48. Liu B, Pan S, Liu J, Kong C. Cancer-associated fibroblasts and the related Runt-related transcription factor 2 (RUNX2) promote bladder cancer progression. Gene. 2021;775:145451. https:// doi. org/ 10. 1016/j. gene. 2021. 145451.
49. Li L, Ugalde AP, Scheele CLGJ, Dieter SM, Nagel R, Ma J, et al. A comprehensive enhancer screen identifies TRAM2 as a key and novel mediator of YAP oncogenesis. Genome Biol. 2021;22:1–28.
50. Deng W, Wang Y, Gu L, Duan B, Cui J, Zhang Y, et al. MICAL1 controls cell invasive phenotype via regulating oxidative stress in breast cancer cells. BMC Cancer. 2016;16:1–11. https:// doi. org/ 10. 1186/ s12885-016-2 553-1.
51. Chen K, Zhao R, Yao G, Liu Z, Shi R, Geng J. Overexpression of kin of IRRELike protein 1 (KIRREL) as a prognostic biomarker for breast cancer. Pathol Res Pract. 2020;216:153000. https:// doi. org/ 10. 1016/j. prp. 2020. 153000.
52. Lee I, Yeom SY, Lee SJ, Kang WK, Park C. A novel senescence-evasion mechanism involving Grap2 and cyclin D interacting protein inactivation by Ras associated with diabetes in cancer cells under doxorubicin treatment. Cancer Res. 2010;70:4357–65.
53. Weniger MA, Pulford K, Gesk S, Ehrlich S, Banham AH, Lyne L, et al. Gains of the proto-oncogene BCL11A and nuclear accumulation of BCL11AXL protein are frequent in primary mediastinal B-cell lymphoma [1]. Leukemia. 2006;20:1880–2.
54. Satterwhite E, Sonoki T, Willis TG, Harder L, Nowak R, Arriola EL, et al. The BCL11 gene family: involvement of BCL11A in lymphoid malignancies. Blood. 2001;98:3413–20.
55. Moreira JMA, Ohlsson G, Gromov P, Simon R, Sauter G, Celis JE, et al. Bladder cancer-associated protein, a potential prognostic biomarker in human bladder cancer. Mol Cell Proteom. 2010;9:161–77.
56. Gromova I, Svensson S, Gromov P, Moreira JMA. Identification of BLCAP as a novel STAT3 interaction partner in bladder cancer. PLoS ONE. 2017;12:1–18.
57. Jacobs DI, Mao Y, Fu A, Kelly WK, Zhu Y. Dysregulated methylation at imprinted genes in prostate tumor tissue detected by methylation microarray. BMC Urol. 2013;13:1.
58. Baris D, Karagas MR, Verrill C, Johnson A, Andrew AS, Marsit CJ, et al. A case-control study of smoking and bladder cancer risk: emergent patterns over time. J Natl Cancer Inst. 2009;101:1553–61.
59. Schned AR, Andrew AS, Marsit CJ, Zens MS, Kelsey KT, Karagas MR. Survival following the diagnosis of noninvasive bladder cancer: WHO/International Society of Urological pathology versus WHO classification systems. J Urol. 2007;178:1196–200.
60. Kelsey KT, Hirao T, Schned A, Hirao S, Devi-Ashok T, Nelson HH, et al. A population-based study of immunohistochemical detection of p53 alteration in bladder cancer. Br J Cancer. 2004;90:1572–6.
61. Karagas MR, Tosteson TD, Blum J, Morris JS, Baron JA, Klaue B. Design of an epidemiologic study of drinking water arsenic exposure and skin and bladder cancer risk in a U.S. population. Environ Health Perspect. 1998;106:1047–50.
62. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.

63. Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. 2016;44:1–6.
64. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
65. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucl Acids Res. 2017;45:e22.
66. KD H. IlluminaHumanMethylationEPICanno.ilm10b4.hg19: annotation for Illumina's EPIC methylation arrays. R Packag version 060. 2017;
67. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and bioconductor. Bioinformatics. 2016;32:587–9.
68. Phipson B, Maksimovic J, Oshlack A. MissMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2016;32:286–8.
69. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, Lord VR, et al. De novo identification of differentially methylated regions in the human genome. Epigenet Chromatin. 2015;8:1–16.
70. Zhou W, Triche TJ, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucl Acids Res. 2018;46:1–15.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Berglund-Brown *et al. Clinical Epigenetics* (2022) 14:27
https://doi.org/10.1186/s13148-022-01247-1

Clinical Epigenetics

RESEARCH Open Access

A core of differentially methylated CpG loci in gMDSCs isolated from neonatal and adult sources

Isabella Berglund-Brown[1], Emily Nissen[2], Devin C. Koestler[2], Rondi A. Butler[3], Melissa N. Eliot[3], James F. Padbury[1], Lucas A. Salas[4], Annette M. Molinaro[6], Brock C. Christensen[4,5], John K. Wiencke[6] and Karl T. Kelsey[3*]

Abstract

Background: Myeloid-derived suppressor cells (MDSCs), which include monocytic (mMDSCs) and granulocytic (gMDSCs) cells, are an immunosuppressive, heterogeneous population of cells upregulated in cancer and other pathologic conditions, in addition to normal conditions of stress. The origin of MDSCs is debated, and the regulatory pattern responsible for gMDSC differentiation remains unknown. Since DNA methylation (DNAm) contributes to lineage differentiation, we have investigated whether it contributes to the acquisition of the gMDSC phenotype.

Results: Using the Illumina EPIC array to measure DNAm of gMDSCs and neutrophils from diverse neonatal and adult blood sources, we found 189 differentially methylated CpGs between gMDSCs and neutrophils with a core of ten differentially methylated CpGs that were consistent across both sources of cells. Genes associated with these loci that are involved in immune responses include *VCL, FATS, YAP1, KREMEN2, UBTF*, *MCC-1*, and *EFCC1*. In two cancer patient groups that reflected those used to develop the methylation markers (head and neck squamous cell carcinoma (HNSCC) and glioma), all of the CpG loci were differentially methylated, reaching statistical significance in glioma cases and controls, while one was significantly different in the smaller HNSCC group.

Conclusions: Our findings indicate that gMDSCs have a core of distinct DNAm alterations, informing future research on gMDSC differentiation and function.

Keywords: DNA methylation, Immunomethylomics, MDSCs, gMDSCs, Cancer

Introduction

A heterogeneous population of immune cells, termed myeloid-derived suppressor cells (MDSCs) [1], are defined phenotypically by their immunosuppressive properties. These cells arise due to various normal stressful states and other pathological conditions that induce a condition reminiscent of emergency myelopoiesis [2]. For example, neonates have an under-developed anti-microbial host defense. It is now recognized that MDSCs play an important role in regulating the earliest immune response [3], as they are quite prevalent in cord blood. Similarly, pregnant women have been reported to have expanded populations of gMDSCs, presumably playing a significant role in producing maternal–fetal tolerance [4]. Finally, MDSCs are commonly produced in numerous cancers [5]. In these conditions, the normal myeloid differentiation pathway is altered; the mature pool of myeloid cells becomes depleted, prompting the genesis of MDSCs that are detectable in peripheral blood [6]. MDSCs suppress the proliferation and function of

*Correspondence: karl_kelsey@brown.edu
[3] Departments of Epidemiology, and Pathology and Laboratory Medicine, Brown University, 70 Ship Street, Providence, RI 02912, USA
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were

made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

T cells and have impaired migratory properties [7]. In cancers, this is thought to create a favorable microenvironment for tumors to grow while evading the immune system [6].

Since the initial identification of immunosuppressive MDSCs, researchers have focused on further characterizing these cells, including how they are induced in these many stressful states. MDSCs can be divided into monocytic (mMDSCs), and granulocytic (gMDSCs) populations based on surface phenotype [8]. Because of their granulocytic origin, it is commonly thought that gMDSCs harbor an immunosuppressive phenotype that is derivative of the neutrophil lineage [9]. Despite numerous studies to develop specific markers of gMDSCs, there is no consensus immunophenotype. The most common feature used to distinguish these cells is their lower density than granulocytic cells, thus co-purification with peripheral blood mononuclear cells following density gradient centrifugation [10].

The precise mechanisms of differentiation or activation that give rise to MDSCs remain unclear. Since DNA methylation (DNAm) has been explored extensively in studies seeking to define the molecular basis of hematopoietic lineage differentiation [11, 12], as well as in lineage activation (e.g., NK cells) [13], DNAm is an excellent candidate for playing an essential role in generating gMDSCs. Patterns of methylation in promoters, enhancers, and gene bodies can strongly correlate with gene expression [14] and define differentially methylated regions (DMRs) used to predict cell lineage [15, 16]. For example, stable and heritable changes in DNAm, regardless of environmental influences, define lineage in T-cell differentiation [17]. Regulatory T cells have been shown to have stable and invariant methylation marks that define their phenotype [18, 19]. Taken together, these data suggest that DNA methylation analysis could provide powerful markers of differentiation states of polymorphonuclear leukocytes (PMNs) that may help define the nature of gMDSCs.

The few DNAm analyses that have been conducted point to the value of this approach, though prior studies have not explored cells isolated from diverse biologic sources. For example, when comparing methylation of inhibitory/suppressive molecules in MDSCs to antigenpresenting cells, CpG islands in the promoter regions of *TGF-β1*, *TIM-3*, and *ARG1* were highly unmethylated [20]. Here, researchers hypothesize that epigenetic mechanisms could control the genesis of suppressive molecules in the tumor microenvironment that contribute to immune tolerance [20]. In another study looking at infiltrating MDSCs in the context of colorectal cancer, genes associated with DNAm-mediated transcriptional silencing and suppression pathways like WNT were upregulated in these cells [21]. Further, some data suggest that DNAm may play a vital role in the genesis of MDSCs, in addition to function. Researchers have also found that demethylated CpG sites characteristic of dendritic cell differentiation were not demethylated in MDSCs [22]. MDSCs show hypermethylation and gene repression in immunogenic regions important for their immunosuppressive properties, which could specifically be involved in the "switch" from immunogenic to tolerant in a tumor environment [22].

While DNAm of MDSCs has been studied in the context of specific cancer types and immunosuppression, epigenetic characterization of multiple sources of gMDSCs has not yet been done. Given the extreme heterogeneity of neutrophils in different clinical in vivo environments, our central objective was to directly examine the DNA methylome of gMDSCs arising in different biologic contexts to epigenetically phenotype these cells. We followed the recommended phenotypic features of gMDSCs as described in Bronte et al., 2016 to guide the isolation of putative gMDSCs from diverse sources [23]. The approach focused on identifying consistent alterations, allowing us to uncover a core set of DNAm alterations that are associated with the genesis of immune suppression and thus potential markers of this phenotype. To implement this, the DNAm of gMDSCs and neutrophils from two distinct age populations were compared: adult blood, where gMDSCs were collected from cancer patients and a pregnant mother, and from umbilical cord blood. These analyses revealed several robust differences in DNAm, consistent across both the adult and cord blood datasets, where some of the alterations associated with genes that have known immune functions. We then compared the whole blood methylation value for the identified loci in cancer patient studies to test whether they were altered in cancer.

Materials and methods

Patient samples (discovery)

The Institutional Review Boards approved biological sample collections from their respective institutions. Deidentified whole blood samples were obtained from adults following written informed consent; head and neck cancer patient samples were collected from the Dana Farber Cancer Institute (Boston, MA) from newly diagnosed patients before initiation of treatment. Whole blood was obtained from a newly diagnosed glioma patient at the Rhode Island Hospital (Providence, RI), following all applicable IRB guidelines. Cord blood and maternal blood samples were obtained at Women and Infants Hospital (Providence, RI) from discarded cords and consenting pregnant women, following the institutional guidelines. Hospital personnel collected all samples during regular standard of care visits, except for cord blood samples obtained from discarded umbilical cords. Samples were drawn and transported to the laboratory at Brown University within 0.5–3 h. The list of samples included in the analyses is summarized in Fig. 1A.

Discovery sample processing

Blood was diluted threefold in PBS (pH 7.4, without calcium or magnesium; ThermoFisher cat no 10010023) supplemented with 2 mM EDTA (Fluka cat no 03690100 ml), and run over a Ficoll density gradient (density 1.077 g/ml; GE Healthcare cat no 17-1440-02). The PBMC fraction, which also contains low-density granulocytic cells, was collected and washed with HBSS (no calcium or magnesium; ThermoFisher cat no 14170112) supplemented with 2 mM EDTA, to remove platelets. The following cell-surface markers were labeled for sorting granulocytes: CD11b-PE (clone REA592; Miltenyi), CD14-eVolve605 (clone 61D3, eBiosciences) or CD14Superbright600 (also clone 61D3, Thermo), CD15APC (clone VIMC6, Miltenyi), CD33-PEVIO770 (clone AC104.3E3, Miltenyi), CD66b-FITC (clone REA306, Miltenyi), HLA-DR-PerCPVio700 (clone AC122, Miltenyi), and viability dye efluor450 (eBiosciences). All antibodies (including new batches) were titrated for optimal signal prior to initiation of cell sorting experiments. Gating was optimized using FMO controls before the initiation of sorting experiments. Each sorting session began by setting compensation with fresh compensation controls using beads, either AbC Anti-Mouse Bead Kit (ThermoFisher cat no A10344) or MACS anti-human Igκ Comp Bead kit (Miltenyi cat no 130-104-187). In cases with HNC samples, antibody labeled cells were 'lightly fixed' with IC Fixation Buffer (ThermoFisher cat no 00-8222-49) at 4 °C for 10 min, washed, and stored overnight at 4 °C before cell sorting. Cell sorting was carried out at the Brown University Flow Cytometry and Sorting Facility on a 5-laser, 20-parameter BD FACSAria IIIu.

Samples were gated for live, singlet cells, and, based on forward scatter vs. side scatter, the granulocytic fraction containing gMDSCs was gated for CD11b+, CD33+, CD14−, CD15+, HLA-DR- and CD66b+. Neutrophils were collected from 2 ml of whole blood using Magnetic Activated Cell Sorting methods (MACSxpress human neutrophil isolation kit, Mitenyi cat no 130-104-434) with some modification: cells were processed in HBSS (no calcium or magnesium) supplemented with 2 mM EDTA. Neutrophils were negatively selected using two rounds of the kit antibody cocktail, resuspended in supplemented HBSS, checked for viability (trypan blue), and counted. Aliquots of cells were frozen at − 80 °C for later downstream processing and analyses. One aliquot of fresh cells was reserved for antibody labeling and FACS phenotype verification. **Suppression assay**

The suppression assay was carried out for representative gMDSC discovery samples using heterologous gMDSCs and PBMCs (containing target T cells). PBMC aliquots were generated from healthy donor leukoreduction filters obtained from the local blood donation center (Rhode Island Blood Center). White blood cells were backflushed with 250 ml PBS (no calcium or magnesium) at a rate of approximately 25 ml/min to avoid excessive pressure. The PBMC fraction was collected from a Ficoll gradient as described above (see Sample Processing) and cryopreserved in single experiment aliquots. Just before use, the PBMCs were thawed, washed, labeled with CellTrace Far Red Cell Proliferation Kit (ThermoFisher cat no C34572) *per* manufacturer's instructions, and stimulated with 30U human IL-2 IS premium grade (Miltenyi cat no 130-097-745) *per* ml of cell suspension and 1ul of Dynabead Human CD3/CD28 T Cell Activation Kit (ThermoFisher cat no 11161D) *per* 1×10^5 (100 ul) PBMCs.

Live gMDSCs were sorted into complete RPMI: RPMI 1640 (ThermoFisher cat no 11875168), supplemented with 10% FBS (ThermoFisher cat no 16000044), 1 × pen/strep/l-glutamine (ThermoFisher cat no 10378016), washed and resuspended at 1×10^5 gMDSCs/100ul medium. Granulocytic MDSCs were co-cultured with activated PBMCs in 96-well tissue culture plates in a 1:1

ratio and cultured under standard conditions (37 °C with 5% CO_2). After 4–5 days, cells were harvested, activation beads removed, and cells labeled for CD66b to determine gMDSCs, and CD3 (clone BW264/56; Miltenyi), CD4 (clone SK3; BioLegend), and CD8 (clone BW135/80; Miltenyi) to determine T-cell populations. The signal from Far Red dye was used to determine the proliferation of CD4+ and CD8+ cells in the presence or absence of gMDSCs. Proliferation metrics were calculated using the FlowJo 10 Proliferation Tool. Both Division Index (the average number of cell divisions across all cells in the original population, regardless of whether they have divided or not) and the Proliferation index (the number of cell divisions/the number of cells that divided).

DNA isolation and methylation array of discovery and testing samples

Fresh frozen cell pellets from training samples were extracted for DNA using the Zymo Quick-DNA MiniPrep Plus kit (cat no D4068) per manufacturer's instructions. To maximize DNA recovery from lightly fixed frozen cells, pellets were subjected to a high heat/high pH method of Campos and Gilbert [24]. Briefly, fixed cells are incubated for 40 min in 0.1 M sodium hydroxide,

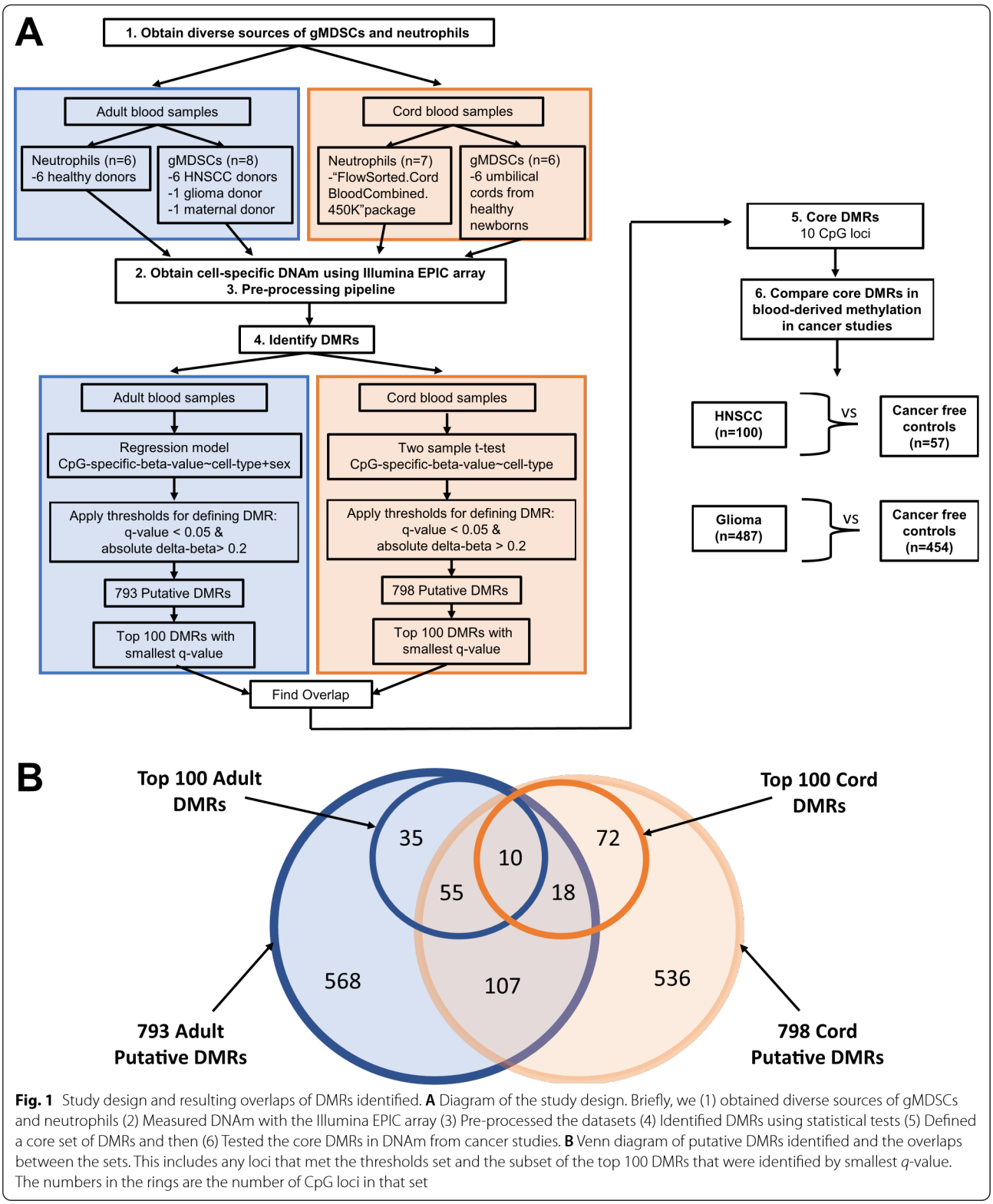


Fig. 1 Study design and resulting overlaps of DMRs identified. **A** Diagram of the study design. Briefly, we (1) obtained diverse sources of gMDSCs and neutrophils (2) Measured DNAm with the Illumina EPIC array (3) Pre-processed the datasets (4) Identified DMRs using statistical tests (5) Defined a core set of DMRs and then (6) Tested the core DMRs in DNAm from cancer studies. **B** Venn diagram of putative DMRs identified and the overlaps between the sets. This includes any loci that met the thresholds set and the subset of the top 100 DMRs that were identified by smallest *q*-value. The numbers in the rings are the number of CpG loci in that set

1% SDS solution at 99 °C, allowed to cool for 5 min, and organically extracted. Samples were precipitated with sodium acetate (pH 5.s; ThermoFisher cat no R1181) and isopropanol in the presence of GlycoBlue carrier (ThermoFisher cat no AM9516). Testing sample DNA was extracted from whole blood using either Qiagen DNEasy Blood & Tissue Kit (cat no 69504) or Zymo

Quick-DNA MiniPrep Plus kit (cat no D4068) per manufacturer's instructions.

DNA was sent to either the USC Molecular Genomics Core facility or the Avera Institute for Human Genetics to process samples for the Illumina EPIC Human DNA Methylation Beadchip. Both facilities bisulfite convert using Zymo Research products. The DNA from fixed samples was processed by the USC Core and treated with the Illumina FFPE QC and Infinium HD FFPE Restore Solution kits.

The neutrophil adult blood methylation data are from Salas et al. [25], and the neutrophil cord blood methylation data are from publicly available Illumina array data on umbilical blood cells (FlowSorted.CordBloodCombined.450 k) [26].

Statistical analysis

Identifying DMRs

To identify putative DMRs between neutrophils and gMDSCs from adult blood, a series of linear regression models were fit independently to each CpG by modeling CpG-specific methylation beta-values as the dependent variable and cell identity as the independent variable (n = 8 gMDSC, n = 6 neutrophil) (Fig. 1A). Linear regression models were adjusted for sex. The regression coefficient for cell-type identity and its corresponding p-value were recorded. To correct for multiple testing, the resulting p-values were adjusted to control the false discovery rate (FDR) using the Benjamini–Hochberg method [24]. As a filtering step, the recommended general-purpose masking from Zhou et al. [27] was used, along with removing CpG loci located on the X and Y chromosomes and removing "ch" and "rs" probes. Putative neutrophil/ gMDSC DMRs were defined as the 100 CpG loci with the smallest adjusted p-value and an absolute delta-beta (difference in mean beta-value between gMDSCs and neutrophils) greater than 0.2.

Similarly, to identify putative DMRs between neutrophils and gMDSCs from cord blood, a series of twosample t-tests were conducted to test the difference in CpG-specific DNAm between neutrophils and gMDSCs (n = 6 gMDSC, n = 7 neutrophil) (Fig. 1A). The mean difference between neutrophil methylation and gMDSC methylation was recorded along with the corresponding p-value from the t-test. Raw p-values were adjusted using the Benjamini–Hochberg method [24], followed by filtering using the same process described above, and putative DMRs were defined similarly as for the adult blood. The CpGs in common between these two lists of DMRs from adult and cord bloods, were identified, and these signature core loci formed the basis of subsequent statistical analyses (Fig. 1B).

Case–control study populations *Head and neck squamous cell carcinoma (HNSCC)*

We interrogated methylation using the Illumina HumanMethylationEPIC array for whole blood samples from 100 cases diagnosed with incident oropharyngeal squamous cell carcinoma that were randomly selected from a population-based case–control study of head and neck cancer in the greater Boston area that has been described elsewhere [28]. Briefly, 533 incident cases of HNSCC (phase 1, December 1999 to December 2003) and an additional 509 incident cases (phase 2, October 2006 to June 2011) were recruited through major teaching hospitals located in Boston, Massachusetts. Adult cases (> 18 years old) included residents of greater Boston with a confirmed incident diagnosis of HNSCC. Cases were excluded if tumors originated in the lip, salivary gland, nasopharynx, nasal sinus, or cavity. Cases were also excluded if their initial primary tumor was diagnosed more than 6 months prior to study contact. Based on the American Joint Committee on Cancer (AJCC) recommendations, tumors were classified as oral cavity, oropharynx, or larynx. We randomly selected 100 participants with oropharyngeal cancer for the current study (ICD-9 codes 146, 148, 149). Cancer-free controls were frequency-matched to cases on age and gender. Controls were selected by matching their town of residence, age, and sex, as has been previously described [28]. ***Glioma***

The case–control San Francisco Adult Glioma Study (AGS) includes 3,164 glioma patients newly diagnosed between 1991 and 2012 who were residents of the SF Bay Area or patients of the UCSF Neuro-oncology clinic and 2,140 subjects without glioma who were residents of the SF Bay Area or patients seen in the UCSF phlebotomy clinic [29]. Controls were frequency-matched to cases by age, race/ethnicity, and gender. Blood samples were collected from glioma patients a median of 100 days after they were histologically diagnosed. For this paper, we included only AGS subjects who had 850 K methylation data available (487 glioma patients and 454 controls).

Evaluating the gMDSC signature loci in cancer

To test whether the identified gMDSC core loci were altered in cancer, the whole blood methylation value for these loci were compared in two independent cancer patient studies: HNSCC and glioma (Fig. 1A). For each study, a series of linear regression models were fit

independently to each CpG by modeling the CpG-specific whole blood methylation beta-value as the dependent variable and cancer status as the independent variable (n = 100 HNSCC cases and n = 57 controls; n = 487 glioma cases and n = 454 controls). The models for comparing case versus control from the HNSCC study also adjusted for cell-type, age, sex, smoking status, drinking status and race. The models for comparing case versus control from the glioma study also adjusted for cell-type, age, sex, smoking status, steroid use, and race. Cell-type was included as the predicted proportions of CD4T cells, CD8T cells, monocytes, neutrophils, and B-cells, obtained via reference-based cell mixture deconvolution [11, 30]. The regression coefficient for cancer status and its corresponding p-values was recorded for each model. A CpG was considered to have significantly different whole-blood methylation between cancer cases and cancer-free controls if the p-value for cancer status was less than 0.05.

Statistical analysis software

All analyses were done in R version 4.1.0.

Results

The overall study design is shown in Fig. 1A. Briefly, we isolated neutrophils and gMDSCs from diverse sources, profiled their DNA methylation status using the Illumina EPIC array and compared the profiles to discover the changes in methylation that characterized the gMDSCs. After the initial gMDSC isolation we assessed the lineage of cells and Fig. 2 shows the predicted proportions of each of the12 cell-types obtained by reference-based cell mixture deconvolution for each of the gMDSC samples, using the method of Salas et al. [25]. For all isolated gMDSC samples, as expected, the most abundant celltype was neutrophils (median of 96.6%) (Additional file 1: Table S1). This (presumably) shows that the DMRs used to predict neutrophil lineage are retained in gMDSCs. To ensure our isolation methods captured phenotypically active gMDSCs, representative samples from HNSCC, cord and maternal sources were assessed for their ability to suppress T-cell activity. The suppression assay was carried out using heterologous gMDSCs and PBMCs (containing target T cells). The Proliferation Index (PI) and Division Index (DI) were calculated for CD4 and CD8 T cells in the presence and absence of our isolated gMDSCs from each representative cell source. Both PI and DI were significantly diminished in the presence of gMDSCs, indicating that these cells could suppress normal T cell responses, consistent with the expected gMDSC phenotype (Table 1).

Having confirmed the protocols employing density gradients and FACS to isolate gMDSCs with an immunosuppressive phenotype, we sought to identify differentially methylated regions (DMRs) that distinguish between gMDSCs and neutrophils. Initially, both cell subtypes were compared in blood from adult donors, where DNA methylation (DNAm) was measured using the Illumina HumanMethylationEPIC platform. A series of linear models were fit independently to each CpG and used to test differences in DNAm between gMDSCs and neutrophils. The DNA methylation signals were compared at each CpG site and initially sorted based on the magnitude of the difference in methylation beta-values (commonly referred to delta-beta, which ranges from -1 to 1). Figure 3A shows the difference in mean methylation

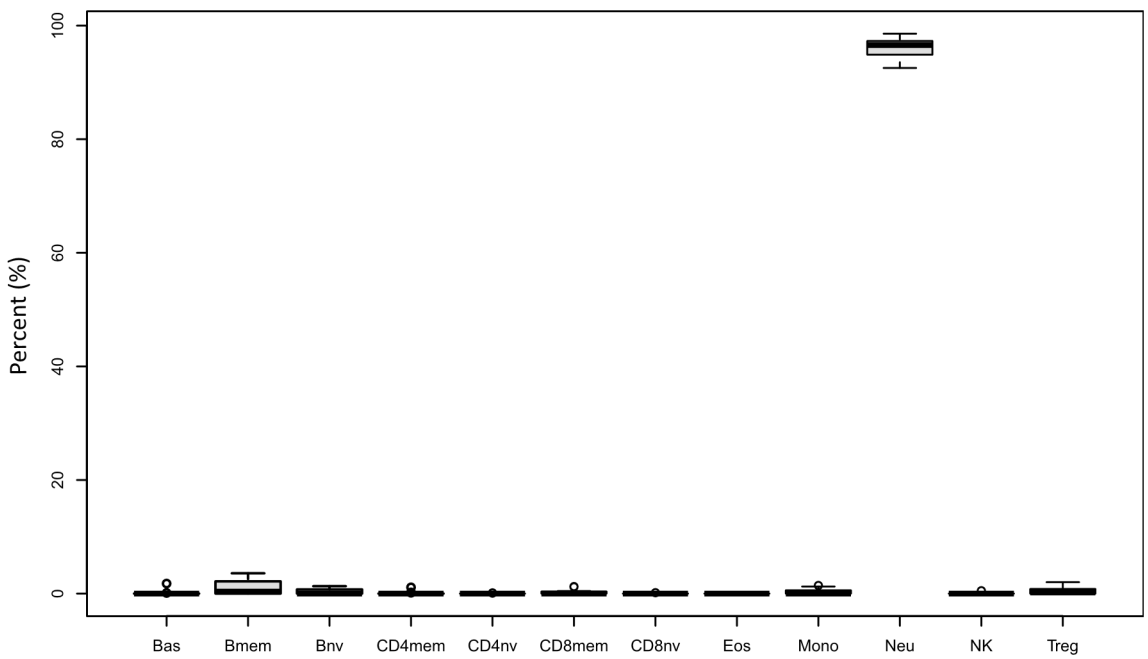


Fig. 2 gMDSC sample deconvolution. The relative prevalence of each cell type in the low-density neutrophil fraction. Isolated gMDSCs were arrayed and the resulting data combined and assessed by deconvolution to generate relative proportions for each cell type. The means for each cell type from all donors are shown

beta-values between neutrophils and gMDSCs, adjusted for sex, for all CpGs in adult blood samples. There were 793 CpGs that met our criteria for differential methylation (q-value (FDR) < 0.05 and absolute adjusted delta-beta > 0.2) (Fig. 1B and Additional file 2: Table S2). From these 793 putative DMRs, the top 100 neutrophil/ gMDSC DMRs were defined as the 100 CpG loci with the smallest adjusted p-value and neutrophil and gMDSC samples were compared using the methylation beta values in an unsupervised clustering heatmap (Additional file 3: Table S3 and Fig. 3B). DNA methylation in leukocytes varies with age, and the age range of the patients who provided the adult blood for gMDSC and neutrophils differed statistically: neutrophil donor mean age was 26.8 years and gMDSC donor mean age, 68.1 years (Additional file 4: Table S4). To mitigate the potential false discovery of gMDSC DMRs attributable to the effects of age and to enhance the sample heterogeneity, we next compared DNAm between gMDSCs and neutrophils in cord blood with the same analytic approach. Figure 3C shows the difference in mean methylation beta-value between neutrophils and gMDSCs from cord blood. There were 798 CpGs that met our criteria of being differentially methylated (q-value (FDR) < 0.05 and absolute delta-beta > 0.2) (Fig. 1B and Additional file 5:
Table S5). Figure 3D shows a heatmap of the methylation beta-values in gMDSC and neutrophils of the top 100 neutrophil/gMDSC DMRs for each cell type (Additional file 6: Table S6). The presence of non-overlapping putative DMRs illustrates the heterogeneity of these cells. There were 603 non-overlapping putative DMRs identified only in adult blood and 608 non-overlapping putative DMRs identified only in cord blood. (Fig. 1B). The genomic and functional context of these DMRs are also heterogeneous (Additional file 7: Table S7).

Table 1 %Reduction in Division Index (DI) and Proliferation Index (PI) in the presence of gMDSCs from different sources

		CD4		CD8	
PBMC: gMDSC	**gMDSC source**	**% ↓ DI**	**% ↓ PI**	**% ↓ DI**	**% ↓ PI**
1:1	Cord blood	58.7	8.6	58.7	11.3
	Maternal				
	HNC				
1:1	HNC	30.2	5.3	58.0	21.4
1:1		43.6	1.8	10.8	13.6
1:1					58.0
			20.7	63.8	10.3

because we do not have leukocyte counts done at blood draw and numerous other cell types likely will change in

Despite the great heterogeneity of these cells, there was a core of DNAm alterations specific to gMDSCs. When we compared adult and cord blood gMDSC DMRs, 190 CpGs overlapped (Fig. 1B and Additional file 8: Table S8). Since these 190 DMRs were derived from independent comparisons, we further analyzed these DMRs to assess whether the direction of the change in methylation in the gMDSCs (when compared with neutrophils) was consistent; for 189 of the 190 DMRs, the direction of the change was the same in adult and cord blood. To discover the loci reflecting the most prominent, generalizable differences in DNAm between neutrophils and gMDSC across lifespan, we compared the top 100 DMRs in adult blood with the top 100 DMRs in cord blood and identified an overlap of ten loci (Fig. 1B). These ten loci represent a core of differentially methylated CpG loci across diverse sources. A heatmap generated of these 10 CpGs is shown in Fig. 4. In both adult and cord blood, gMDSCs cluster together, and the neutrophils cluster together across these ten loci. As displayed in Table 2, the directionality of the change in the beta values of these ten loci also is preserved between adult and cord blood. Of the 10 CpG sites in the gMDSC signature loci, nine are associated with specific genes (Table 2).

Since numerous prior studies have demonstrated that gMDSC proportions are increased in the peripheral circulation of cancer patients, we reasoned that if alterations in the DNAm of the 189 CpGs represent gMDSCs, one might expect whole blood leukocyte DNAm of patients with cancers to change together, in a direction commensurate with the relatively small change in the prevalence of the MDSCs in whole blood. Importantly, we cannot interpret the meaning of the directionality of any change
concert with occurrence of the cancer. That is, since the relative numbers of leukocytes systematically change in cancer patients, the *proportional* change direction of any subtype (especially small subtypes) does not necessarily reflect the *numerical* change, although the direction would be hypothesized to be uniform. Hence, we compared the case–control delta-beta in the 189 consistent CpG loci signature from whole blood DNAm gMDSC in two cancer patient studies. In the HNSCC study, 170 CpG loci had a case–control delta-beta in the opposite direction of the corresponding gMDSC-

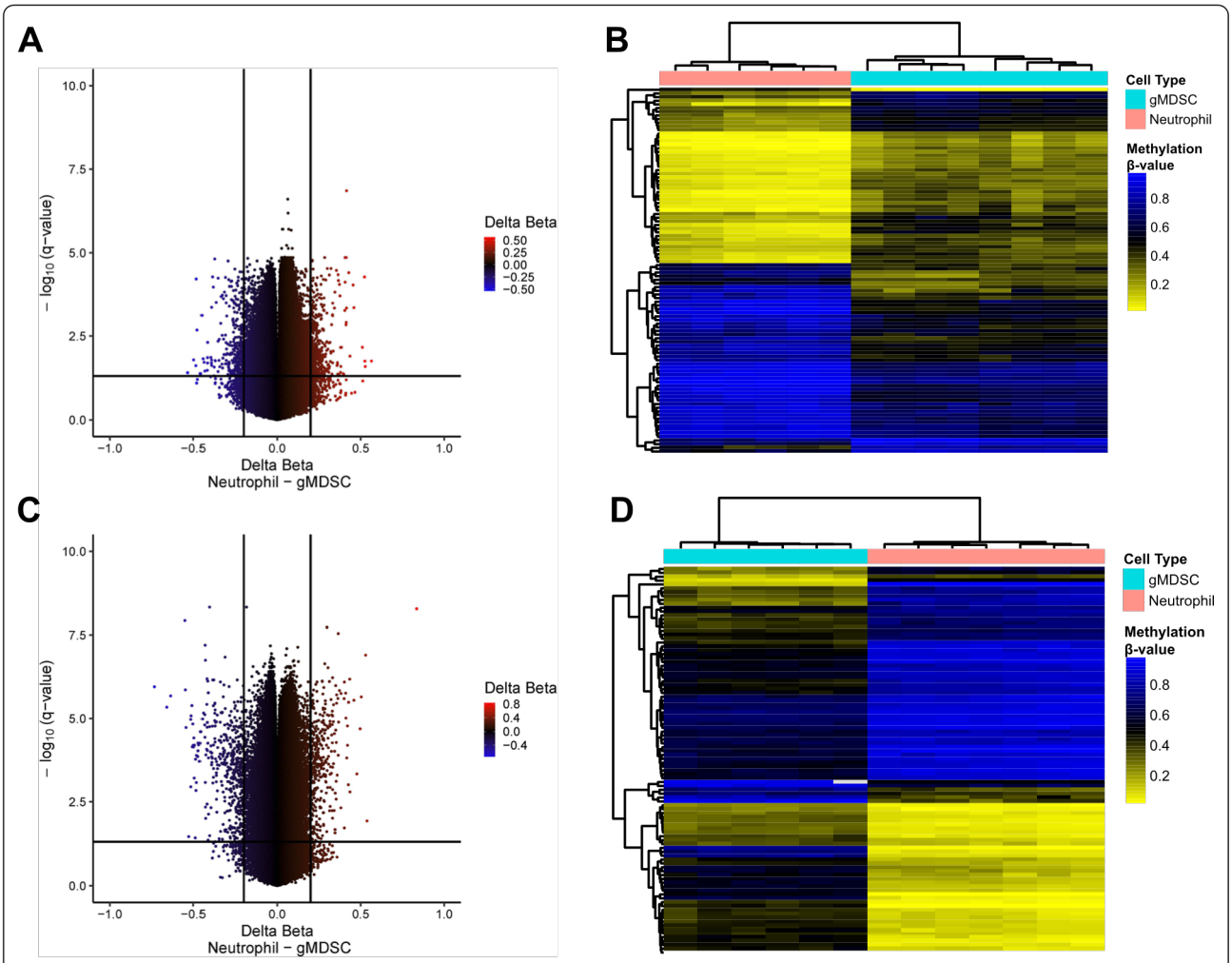


Fig. 3 Identifying putative neutrophil/gMDSC DMRs from adult blood samples and cord blood samples. **A** Volcano plot of − log 10(*q*-value) against delta-beta, which represents the difference in mean methylation beta-value between neutrophils and gMDSCs, adjusted for sex, in adult blood samples. The horizontal black bar represents the threshold of significance (*q*-value (FDR) = 0.05) and the two vertical black bars represent the threshold for delta-beta (delta-beta = · 0.2). **B** Unsupervised clustering heatmap of the top 100 gMDSC DMRs identified in adult blood samples, defined as the 100 CpG loci with the smallest adjusted *p*-value and absolute adjusted delta-beta greater than 0.2. Each column represents a sample, and each row represents a CpG locus. Above the heatmap, color indicates cell-type (blue = gMDSC and pink = neutrophil). Within the heatmap, color indicates methylation beta-value (blue = β of 1, or methylated, and yellow = β of 0, or unmethylated). **C** Volcano plot of − log10(*q*-value) against delta-beta, which represents the difference in mean methylation beta-value between neutrophils and gMDSCs in cord blood samples. The horizontal black bar represents the threshold of significance (*q*-value (FDR) = 0.05) and the two vertical black bars represent the threshold for delta-beta (delta-beta = · 0.2). **D** Unsupervised clustering heatmap of the top 100 gMDSC DMRs identified in cord blood samples, defined as the 100 CpG loci with the smallest adjusted *p*-value and absolute delta-beta greater than 0.2. Each column represents a sample, and each row represents a CpG locus. Above the heatmap, color indicates cell-type (blue = gMDSC and pink = neutrophil). Within the heatmap, color indicates methylation beta-value (blue = β of 1, or methylated, and yellow = β of 0, or unmethylated)

Neutrophil delta-beta. In the glioma study, 134 CpG loci

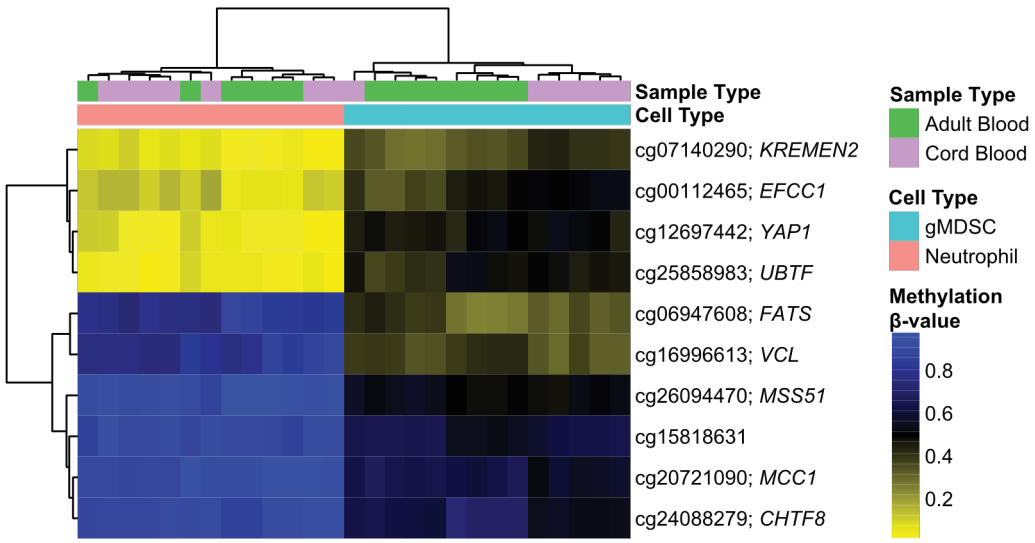


Fig. 4 Core of differentially methylated CpG loci between gMDSCs and neutrophils from diverse adult and neonatal sources. Unsupervised clustering heatmap of the core ten gMDSC loci. Each column represents a sample, and each row represents a CpG locus. Above the heatmap, colors indicate sample type (green = adult blood, purple = cord blood) and cell-type (blue = gMDSC and pink = neutrophil). Within the heatmap, color indicates methylation beta-value (blue = β of 1, or methylated, and yellow = β of 0, or unmethylated)

had a case–control delta-beta in the opposite direction of the corresponding gMDSC-Neutrophil delta-beta. Further, there was an overlap of 122 loci between these groups of CpG loci, meaning there were 122 loci that had a consistent case–control delta-beta in the two cancer studies that were opposite of the respective gMDSCNeutrophil delta-beta found from adult and cord blood samples (Fig. 5A). To assess the chance of observing 122 loci with consistent delta-betas in these cancer studies that are also in the opposite direction of their respective gMDSC-Neutrophil delta-betas, we derived the distribution of the number of loci under the null hypothesis.

Table 2 10 CpG overlap between adult blood and cord blood

CpG	Chr	Gene	Genomic context	Delta beta AB				Delta beta CB	*P*-value AB	*P*
cg07140290	chr16	*KREMEN2*	CpG Island	−0.26	−0.32	1.91E−08	2.77E−10	TSS200		
cg00112465	chr3	*EFCC1*	CpG Island	−0.30	−0.33	3.01E−06	3.43E−08	TSS1500		
cg12697442	chr11	*YAP1*	CpG Island	−0.38	−0.41	2.05E−09	1.35E−09	TSS200		
cg25858983	chr17	*UBTF*	North shore	0.37	0.41	1.24E 07	6.96E 16	–	–	
– cg06947608	chr10	*FATS*	Open sea	0.21	0.33	7.20E−06	4.86E−08	0.41	5.6E−07	1
sea 0.45 cg16996613	chr10	*VCL*	Open sea							
sea 0.39 cg26094470	chr10	*MSS51*	Open sea							
sea 0.41 cg15818631	chr12	Open sea	0.26							
cg20721090	chr5	*MCC1*	Open sea	0.28						
cg24088279	chr16	*CHTF8*	Open sea							

0.43		6.11E−10	7.00E−09	Body	0.31	4.51E−10	5.24E−08	Body
0.38	3.77E−10	1.52E−09	Body	0.27	2.58E−06			Body
	1.57E−10							

This was done by: (1) creating a set of CpG loci with consistent gMDSC-Neutrophil delta-betas in adult and cord samples (n = 536,320 loci) (2) randomly sample 189 CpGs from this set (3) count the number of loci where the case–control delta-beta for both the HNSCC and glioma studies are the opposite direction of the respective gMDSC-Neutrophil delta-beta. This process was repeated 10,000 times and a histogram of the results are shown in Fig. 5B. The mean of this distribution is 82 and our observed value of 122 falls far outside the range. Of the core ten gMDSC signature loci, all ten CpGs were statistically significantly differentially methylated between cancer type and healthy controls in the larger glioma study, while for the less well-powered study of HNSCC patients, one loci was significantly different in cases compared with controls (Fig. 5C). When the difference in the estimate of DNA methylation (mean beta-value) between the control and cancer case was compared, the direction of the change was consistent in HNSCC and glioma for nine of ten of the loci (Fig. 5D).

Discussion

In this study, we have directly compared the DNA methylation features of putative gMDSCs isolated from different sources and have identified the great heterogeneity of these cells at the epigenetic level, but also a core of common DNAm alterations. The isolation procedures used for the discovery samples captured a highly purified low-density population of cells with an immunophenotype and methylation signature consistent with neutrophil lineage. This result is in line with previous studies [10].

We have identified CpG loci that are differentially methylated between gMDSCs and neutrophils. Using two age-distinct populations of gMDSCs and neutrophils (adult and cord blood samples derived from diverse sources), we identified a set of 189 common CpG loci, with 10 CpG loci displayed consistent differential methylation between the cell subtypes. A number of the genes associated with the identified loci are known to have roles in immune responses, including some immune processes specific to neutrophils.

VCL, for example, codes for the scaffolding protein vinculin, which is involved in the maturation of integrinbased focal adhesions in a fashion previously termed 'mechanosensitive' [31]. Recent work has shown that this protein has a role in leukocyte trafficking, including a mechanosensitive role for vinculin in neutrophil adhesion and spreading [32]. Our work suggests, then, that vinculin may play an essential role in the genesis of particular features associated with gMDSCs. Several of the other core loci identified in this

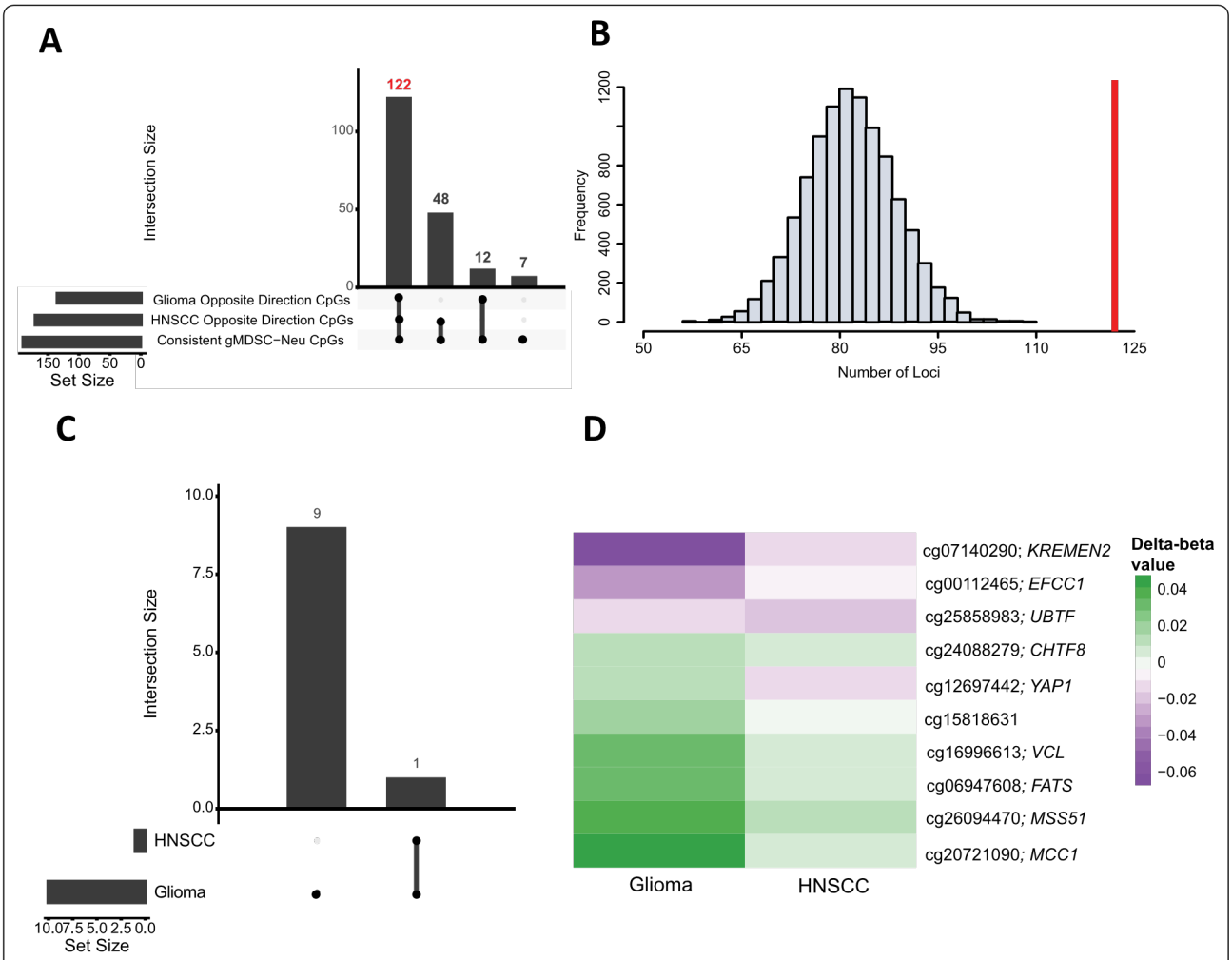


Fig. 5 Assessing uniform change of whole blood DNAm in cancer studies for the 189 consistent loci and the core 10 loci. **A** Upset plot showing the overlap of the 189 DMRs with consistent gMDSC-Neu delta-betas in adult and cord blood samples and the respective CpG loci in the HNSCC and glioma studies that had case–control delta-betas in the opposite direction (termed "opposite direction CpGs"). **B** Histogram of the distribution under the null hypothesis of the number of loci with case–control delta-betas that are in the opposite direction of the respective gMDSC-Neu delta-beta in both the HNSCC and glioma study. Distribution was created by randomly drawing 189 CpG loci from the set of all loci with consistent gMDSC-Neu delta-betas and the counting the number of loci that have case–control delta-betas in the opposite direction of gMDSC-Neu delta-betas in both the glioma and HNSCC studies. The horizontal red line is at our observed value of 122 loci, far outside the range of the distribution. **C** Upset plot of CpGs that were significantly different between cases and cancer-free controls from testing the association between whole blood methylation beta-value and cancer status for each of the gMDSC core loci in two cancer studies. **D** Heatmap of the whole blood methylation delta-beta for cancer versus control in the glioma study population and the HNSCC study population. Each row represents one of the core ten loci. The color in the heatmap indicates the delta-beta value

study have been shown to influence immune activity, through evidence of an association with the genesis of MDSCs. Deleting *FATS*, Fragile Site Tumor Suppressor, (one of the 10 common loci) in mice, has been shown to decrease the prevalence of gMDSCs in tumor tissues [33].

Expression of *YAP1*, or Yes-Associated Transcriptional Regulator, is associated with MDSC expansion [34]. Hyperactivated YAP1 signaling in tumors leads to MDSC recruitment [35], while blockage of *YAP1* leads to a decreased induction of MDSCs [34]. This suggests that *YAP1* expression is heightened in tumors and plays a role in recruiting and expanding immunosuppressive MDSCs. *YAP1* also seems to have a similar function as MDSCs in propagating tumor growth—it suppresses T-cell function and infiltration to the tumor microenvironment [36]. Though expression of *YAP1* by tumor cells propagates tumor growth, recent research has shown that there are binary pan-cancer classes where *YAP1* displays either pro- or anti-cancer activity depending on whether it is expressed or silenced in different types of cancer [37]. Hence, the role of *YAP1* in cancer seems to be complex, with our data suggesting that this gene has a direct role in gMDSC genesis.

KREMEN2 (cg07140290), encoding Kringle-Containing TransmembraneProtein 2, is a protein implicated in cancer immunosurveillance in relation to WNT signaling [38]. Dysregulated WNT signaling culminates in cancer progression, malignant transformation, and resistance to treatment, undermining cancer immunosurveillance [38]. In the context of MDSCs, WNT displays anti-cancer effects as WNT/β-catenin signaling limits the tumorpromoting role of MDSCs, with downstream inhibition of β-catenin resulting in MDSC expansion and tumor infiltration [38, 39]. WNT/β-catenin within MDSCs themselves limits their ability to expand and infiltrate tumors. *KREMEN2* has potential involvement in these pathways because it hinders WNT activation [40].

Some of the identified loci are immunologically relevant, but their association with gMDSCs has not previously been recognized. Upstream Binding Transcription Factor (UBTF) plays a role in innate antiviral immunity as part of the pattern recognition receptors, where it partners with Interferon Gamma Inducible Protein 16 (IFI16) to restrict herpes simplex virus replication [41]. Interestingly, this process may also have a role in selftolerance, as IFI16 forms oligomers spontaneously in Sjogren's syndrome [42]. UBTF also contains a gamma activating sequence, suggesting it may operate as part of the interferon-gamma pathway [43]. The gene product of MCC Regulator of WNT Signaling Pathway (*MCC-1*) is similarly involved in the innate immune response [44], although its role in the genesis of gMDSCs is also unclear.

Finally, EF-Hand And Coiled-Coil Domain-Containing Protein 1 (*EFCC1*) has been studied in lung adenocarcinoma and shown, using immunohistochemical as well as RNA expression analysis, to be less prevalent in lung cancers and to predict poorer outcome [45]. Our data might suggest that some of the cells contributing to this assessment are infiltrating gMDSCs, with these having more methylated *EFCC1* and thus less protein and less RNA present. Regardless, our result, combined with the work of Xia et al. [45] does point to the potential utility of studies of both RNA and protein expression in tumors for study of MDSC infiltration. In terms of the other candidate loci (Table 2), cg15818631 does not have any associations with known genes, and *CHTF8* has no known role in the immune response.

After identifying these the common loci, we investigated their whole blood methylation status in cancer cases and healthy controls. The overlapping set of CpGs that represented gMDSCs in the cancer patients exhibited the same change in the direction of the methylation value compared with controls, suggesting that they reflect a co-methylation module, perhaps associated with coordinate differentiation. All of the core ten loci were significantly different between glioma cases and controls while 1 was significantly different in the smaller HNSCC study (possibly attributable to issues with power to detect small changes in the subtype prevalence); this could indicate that the pattern of altered methylation is partially attributable to an altered number of gMDSCs, which are known to be produced in cancer [6].

Strengths of this research include replication of our analyses in two distinct populations representative across the life-course—adult and cord blood. This work is particularly relevant to cancer, as most gMDSC samples came from cancer patients, and the results from discovery applied to cancer cases vs. cancer-free controls.

The use of heterogeneous populations might be a limitation, as some have suggested that low density cells in cord blood do not suppress lymphocyte mitogenesis, but rather this finding is an artifact of active macrophages that eliminate beads that stimulate T cell mitosis [46]. Regardless, the similar isolation procedures for both populations produced 190 CpGs with altered methylation that overlapped between these cell subtypes. Other limitations of this research include a relatively small sample size, although this allowed for the identification of ten loci with significantly altered methylation between

gMDSCs and neutrophils in both adult and cord blood. While the scope of our research was to identify DNAm markers of gMDSCs, opportunities for future work include further investigation of these candidate loci. For example, lineage differentiation has long been known to be associated with microRNA expression [47] and further evaluation of the mechanistic importance of our findings should include detailed integration of these data with an assessment of altered microRNA. We've further investigated the whole blood DNAm of these loci in two different cancer types and found that nine of the ten loci were consistently different in cancer patients than controls. Together, our data indicate that gMDSCs have a core of distinct DNAm changes from neutrophils that inform future research on gMDSC differentiation and function.

Supplementary Information

The online version contains supplementary material available at https:// doi. org/ 10. 1186/ s13148- 022- 01247-1.

Additional file 1: Table S1. gMDSC deconvolution.

Additional file 2: Table S2. CpGs meeting criteria in adult blood samples.

Additional file 3: Table S3. Top 100 CpGs in adult samples.

Additional file 4: Table S4. Adult sample demograph.

Additional file 5: Table S5. CpGs meeting criteria in cord blood samples.

Additional file 6: Table S6. Top 100 CpGs in cord blood samples.

Additional file 7: Table S7. CpG characteristics.

Additional file 8: Table S8. Overlapping CpGs between adult and cord.

Acknowledgements

We appreciate the contributions of the San Francisco Adult Glioma study investigators and staff in providing clinical epidemiologic data.

Authors' contributions

Experimental design: KTK, JKW, DCK, IBB, EN, BCC, LAS, RB. Implementation: KTK, JKW, DCK, IBB, EN, BCC, LAS, RB, ME, JP. Analysis and interpretation of the data: KTK, JKW, DCK, IBB, EN, BCC, LAS, AMM. All authors were involved in the writing of the manuscript and have read and approved the final version.

Funding

Research was supported by R01CA216265, R01CA207360, R01CA253976, UCSF Brain Tumor SPORE (P50CA097257), the National Cancer Institute (NCI) Cancer Center Support Grant P30 CA168524, the Kansas Institute for Precision
Medicine COBRE (Supported by the National Institute of General Medical Science award P20 GM130423), the Kansas IDEA Network Of Biomedical Research Excellence (P20 GM103418), the CDMRP/Department of Defense (W81XWH-20-1-0778) and NIGMS (P20GM104416-09/8299). JKW is supported by the Robert Magnin Newman Endowed Chair in Neuro-oncology.

Availability of data and materials

The data generated in this study are available in the GEO database, upon acceptance of the paper in the journal.

Declarations

Ethics approval and consent to participate

The Institutional Review Boards from respective institutions approved biological sample collections. Written informed consent was received from all of the participating adults for all of the studies involved in this work. The analysis presented here was conducted on deidentified whole blood samples.

Consent for publication Not applicable.

Competing interests

JKW and KTK are co-founders of Cellintec, which had no role in this research. The other authors declare no potential conflicts of interest.

Author details

[1] Warren Alpert Medical School, Brown University, Providence, RI, USA. [2] Department of Biostatistics and Data Science, University of Kansas Medical Center, Kansas City, KS, USA. [3] Departments of Epidemiology, and Pathology and Laboratory Medicine, Brown University, 70 Ship Street, Providence, RI 02912, USA. [4] Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA. [5] Departments of Molecular and Systems Biology, and Community and Family Medicine, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA. [6] Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA.

Received: 2 December 2021 Accepted: 11 February 2022

Published online: 21 February 2022

References

1. Gabrilovich DI, et al. The terminology issue for myeloid-derived suppressor cells. Cancer Res. 2007;67(1):425 (**author reply 426**).
2. Brusa D, et al. Circulating immunosuppressive cells of prostate cancer patients before and after radical prostatectomy: profile comparison. Int J Urol. 2013;20(10):971–8.
3. Rieber N, et al. Neutrophilic myeloid-derived suppressor cells in cord blood modulate innate and adaptive immune responses. Clin Exp Immunol. 2013;174(1):45–52.
4. Köstlin N, et al. Granulocytic myeloid derived suppressor cells expand in human pregnancy and modulate T-cell responses. Eur J Immunol. 2014;44(9):2582–91.
5. Almand B, et al. Increased production of immature myeloid cells in cancer patients: a mechanism of immunosuppression in cancer. J Immunol. 2001;166(1):678–89.
6. Gabrilovich DI, Nagaraj S. Myeloid-derived suppressor cells as regulators of the immune system. Nat Rev Immunol. 2009;9(3):162–74.
7. Brandau S, et al. Myeloid-derived suppressor cells in the peripheral blood of cancer patients contain a subset of immature neutrophils with impaired migratory properties. J Leukoc Biol. 2011;89(2):311–7.
8. Millrud CR, Bergenfelz C, Leandersson K. On the origin of myeloid-derived suppressor cells. Oncotarget. 2017;8(2):3649–65.
9. Pillay J, et al. Immune suppression by neutrophils and granulocytic myeloid-derived suppressor cells: similarities and differences. Cell Mol Life Sci. 2013;70(20):3813–27.

10. Hardisty GR, et al. High purity isolation of low density neutrophils casts doubt on their exceptionality in health and disease. Front Immunol. 2021;12:2057.
11. Houseman EA, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform. 2012;13(1):1–16.
12. Salas LA, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19(1):64.
13. Wiencke JK, et al. The DNA methylation profile of activated human natural killer cells. Epigenetics. 2016;11(5):363–80.
14. Laurent L, et al. Dynamic changes in the human methylome during differentiation. Genome Res. 2010;20(3):320–31.
15. Nazor KL, et al. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. Cell Stem Cell. 2012;10(5):620–34.
16. Boland MJ, Nazor KL, Loring JF. Epigenetic regulation of pluripotency and differentiation. Circ Res. 2014;115(2):311–24.
17. Naito T, Taniuchi I. Roles of repressive epigenetic machinery in lineage decision of T cells. Immunology. 2013;139(2):151–7.
18. Toker A, et al. Active demethylation of the Foxp3 locus leads to the generation of stable regulatory T cells within the thymus. J Immunol. 2013;190(7):3180–8.
19. Li X, Zheng Y. Regulatory T cell identity: formation and maintenance. Trends Immunol. 2015;36(6):344–53.
20. Saleh R, et al. DNA methylation in the promoters of PD-L1, MMP9, ARG1, galectin-9, TIM-3, VISTA and TGF-β genes in HLA-DR(-) myeloid cells, compared with HLA-DR(+) antigen-presenting cells. Epigenetics. 2020;15(12):1275–88.
21. Sasidharan Nair V, et al. Transcriptomic profiling disclosed the role of DNA methylation and histone modifications in tumor-infiltrating myeloidderived suppressor cell subsets in colorectal cancer. Clin Epigenet. 2020;12(1):13.
22. Rodriguez-Ubreva J, et al. Prostaglandin E2 leads to the acquisition of DNMT3A-dependent tolerogenic functions in human myeloid-derived suppressor cells. Cell Rep. 2017;21(1):154–67.
23. Bronte V, et al. Recommendations for myeloid-derived suppressor cell nomenclature and characterization standards. Nat Commun. 2016;7(1):1–10.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol). 1995;57(1):289–300.
25. Salas, L.A., et al., Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. bioRxiv, 2021. p. 2021.04.11.439377.
26. Salas LA, et al. R topics documented. 2019.
27. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucl Acids Res. 2017;45(4):e22–e22.
28. Applebaum KM, et al. Lack of association of alcohol and tobacco with HPV16-associated head and neck cancer. J Natl Cancer Inst. 2007;99(23):1801–10.
29. Wrensch M, et al. Diagnostic, treatment, and demographic factors influencing survival in a population-based study of adult glioma patients in the San Francisco Bay Area. Neuro-oncol. 2006;8(1):12–26.
30. Koestler DC, et al. Improving cell mixture deconvolution by id entifying o ptimal dna methylation l ibraries (idol). BMC Bioinform. 2016;17(1):1–21.
31. Atherton P, et al. Vinculin controls talin engagement with the actomyosin machinery. Nat Commun. 2015;6(1):1–12.
32. Wilson ZS, et al. Context-dependent role of vinculin in neutrophil adhesion, motility and trafficking. Sci Rep. 2020;10(1):1–14.
33. Zhang R, et al. FATS deficiency in mice limits melanoma progression via converting the tumor immune microenvironment to an anti-tumor state. Am Assoc Immnol. 2016.
34. Yang R, et al. Tumour YAP1 and PTEN expression correlates with tumourassociated myeloid suppressor cell expansion and reduced survival in colorectal cancer. Immunology. 2018;155(2):263–72.
35. Wang G, et al. Targeting YAP-dependent MDSC infiltration impairs tumor progression. Cancer Discov. 2016;6(1):80–95.
36. Stampouloglou E, et al. Yap suppresses T-cell function and infiltration in the tumor microenvironment. PLoS Biol. 2020;18(1):e3000591.
37. Pearson JD, et al. Binary pan-cancer classes with distinct vulnerabilities defined by pro- or anti-cancer YAP/TEAD activity. Cancer Cell. 2021;39:1115–34.
38. Galluzzi L, et al. WNT Signaling in cancer immunosurveillance. Trends Cell Biol. 2019;29(1):44–65.
39. Di Piazza M, et al. Loss of cutaneous TSLP-dependent immune responses skews the balance of inflammation from tumor protective to tumor promoting. Cancer Cell. 2012;22(4):479–93.
40. Mao B, et al. Kremen proteins are Dickkopf receptors that regulate Wnt/βcatenin signalling. Nature. 2002;417(6889):664–7.
41. Lum KK, et al. Charge-mediated pyrin oligomerization nucleates antiviral IFI16 sensing of herpesvirus DNA. MBio. 2019;10(4):e01428-e1519.
42. Antiochos B, et al. IFI16 filament formation in salivary epithelial cells shapes the anti-IFI16 immune response in Sjögren's syndrome. JCI Insight. 2018;3(18):e120179.
43. Saha B, et al. Gene modulation and immunoregulatory roles of interferonγ. Cytokine. 2010;50(1):1–14.
44. Alcami A, Koszinowski UH. Viral mechanisms of immune evasion. Trends Microbiol. 2000;8(9):410–8.
45. Xia L, et al. Decreased expression of EFCC1 and its prognostic value in lung adenocarcinoma. Ann Transl Med. 2019;7(22):672.
46. Weinhage T, et al. Cord blood low-density granulocytes correspond to an immature granulocytic subset with low expression of S100A12. J Immunol. 2020;205(1):56–66.
47. Chen C-Z, et al. MicroRNAs modulate hematopoietic lineage differentiation. Science. 2004;303(5654):83–6.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



nature COMMUNICATIONS

https://doi.org/10.1038/s41467-021-27864-7 ARTICLE OPEN

Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling

Lucas A. Salas[1], Ze Zhang[1], Devin C. Koestler[2], Rondi A. Butler[3], Helen M. Hansen[4], Annette M. Molinaro[4], John K. Wiencke[4,5], Karl T. Kelsey[3]✉ & Brock C. Christensen[1,6,7]✉

DNA methylation microarrays can be employed to interrogate cell-type composition in complex tissues. Here, we expand reference-based deconvolution of blood DNA methylation to include 12 leukocyte subtypes (neutrophils, eosinophils, basophils, monocytes, naïve and memory B cells, naïve and memory CD4 + and CD8 + T cells, natural killer, and T regulatory cells). Including derived variables, our method provides 56 immune profile variables. The IDOL (IDentifying Optimal Libraries) algorithm was used to identify libraries for deconvolution of DNA methylation data for current and previous platforms. The accuracy of deconvolution estimates obtained using our enhanced libraries was validated using artificial mixtures and whole-blood DNA methylation with known cellular composition from flow cytometry. We applied our libraries to deconvolve cancer, aging, and autoimmune disease datasets. In conclusion, these libraries enable a detailed representation of immune-cell profiles in blood using only DNA and facilitate a standardized, thorough investigation of immune profiles in human health and disease.

[1] Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA. [2] Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, KS, USA. [3] Departments of Epidemiology and Pathology and Laboratory Medicine, Brown University, Providence, RI, USA. [4] Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA. [5] Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA. [6] Department of Molecular and Systems Biology, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA. [7] Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA. ✉email: Karl_Kelsey@brown.edu; Brock.C.Christensen@dartmouth.edu

Advances in DNA methylation microarrays have allowed a greater understanding of how DNA methylation is affected by environmental exposures and altered in chronic diseases[1,2]. Peripheral blood is one of the most common biological matrices for those investigations. Blood DNA methylation profiles include information from multiple cell lineages and, in some cases, cell states. Every cell lineage has unique DNA methylation patterning regulating cell-specific gene expression, and some methods leverage DNA methylation to understand underlying cell heterogeneity[1–7]. The reference-based approach assumes that the principal source of signal variability in a heterogeneous sample (such as blood) reflects the signals' proportions in the different cell components[8]. Constrained projection/ quadratic programming (CP/QP) employs purified cell types as reference samples to generate a "reference library," a matrix of differentially methylated sites among cell types, and yields highly accurate estimates of the underlying cell composition in mixed cell populations (e.g., peripheral blood)[9]. Previously established statistical deconvolution frameworks such as CP/QP, support vector regression (CIBERSORT), and robust partial regression (EpiDISH) have similar accuracy and precision in deconvolution estimates[10]. Marker selection methods for library creation use automatic procedures to discern library markers[11] or iterative approaches for selecting sets of markers (IDOL—IDentifying Optimal Libraries) that maximize the accuracy of deconvolution estimates[12]. To enhance the utility of cell-type deconvolution, reference library improvements and expansions of existing libraries to include additional cell types are needed to broaden the scope of DNA methylation-based immune phenotyping[13–16].

Continued methodological advancements are highly dependent on the quality and genome coverage of a reference library. In the original description of CP/QP for methylation-based deconvolution, Houseman et al. developed a library based on an early microarray platform, the Illumina HumanMethylation27k microarray[9]. When the Illumina-HumanMethylation450k technology was released, Jaffe et al. applied the Houseman method with the reference data developed by Reinius et al.[11,17]. They accurately discriminated CD8 and CD4T cells, but NK and granulocytes (neutrophils and eosinophils) discrimination performance showed room for improvement[11,17]. Potentially limiting generalizability, the reference cell populations were purified solely from males of Northern European (Swedish) origin[17,18]. We recently developed a deconvolution library to discriminate better six major cell types (CD4(+) T cells, CD8(+) T cells, NK, B cells, monocytes, and neutrophils) using the Illumina HumanMethylationEPIC technology, hereafter named EPIC IDOL-6[19]. A distinct advantage of this library is the inclusion of more ethnically diverse male and female subjects.

Beyond the six major leukocyte cell types in peripheral blood, there have been further attempts to deconvolve memory and naïve cells and other granulocytes[13,14]. However, they have not been widely adopted or tested as some of the references are not publicly available or involved a combination of different technologies. Some algorithms include other rarer cell subpopulations [e.g., plasmablasts, exhausted CD8(+) T cells] as linearly related scores, but they do not represent the sample's cell-type proportions[15]. Several newborn umbilical cord blood-specific libraries also have been developed[20–24].

Here, we augment reference-based deconvolution of adult peripheral blood DNA methylation data to include memory and naïve cells from cytotoxic and helper T cells and B cells and parse the granulocyte subtypes into neutrophils, eosinophils, and basophils (Fig. 1). Our comprehensive library provides information across 12 different cell subtypes; depending on the hypothesis, categories could be collapsed at in seven additional higher branches (T cells, B cells, CD8T, CD4T, granulocytes, lymphoid, and myeloid), Fig. 1 panel c, resulting in 19 relative cell-type proportions and 19 cell counts (derived data using the complete cell blood counts or flow cytometry). In addition, the library includes multiple cell-derived ratios and proportions, such as the neutrophil to lymphocyte ratio or naïve to memory ratios (see Fig. 1 for 18 known examples), totaling more than 56 total immune profile variables. This library, hereafter named EPIC IDOL-Ext, will find wide application to the study of immune profiles in health and disease. Additionally, a second extended library using probes present on the legacy DNA HumanMethylation450k array, henceforth named 450k IDOL-Ext, was created for application to existing 450 K datasets to expand cell-type representation.

Results

Deconvolution library development. To define a novel deconvolution library from 12 purified cell types, we measured DNA methylation and performed rigorous quality assessment and control for all samples. The final reference dataset included the following cell types: neutrophils (Neu, n = 6), eosinophils (Eos, n = 4), basophils (Bas, n = 6), monocytes (Mono, n = 5), B naïve cells (Bnv, n = 4), B memory cells (Bmem, n = 6), T-helper CD4 + naïve cells (CD4nv, n = 5), T-helper CD4 + memory cells (CD4mem, n = 4), T regulatory cells (Treg, n = 3), T-cytotoxic CD8 + naïve cells (CD8nv, n = 5), T-cytotoxic memory CD8 + cells (CD8mem, n = 4), and natural killer cells (NK, n = 4). The estimated purity of reference samples is based on commonly accepted CD marker definitions (Supplementary Table 1). The cells represented a wide variety of adult donors, including both sexes and different genetic ancestries, a more granular demographic information is included in Supplementary Table 2. The mean purity obtained from the flow cytometry confirmation step (after antibody-linked magnetic bead sorting) was 93% (range 85–99%), with the lowest purity observed for the CD8mem samples (85%). We first used the minfi pickCompProbes function to select an automatic library and estimated cell-type proportions using this library with methylation data from purified cells. This library represents the expected average signal of some extreme hypo and hypermethylated markers per cell type. Pure samples would approximate the average signal of the specific cell type. However, when a sample is contaminated, the signal from another cell type(s) will differ from the average and indicate the potential contaminant's proportion. This procedure for reverse cell-type estimation is denoted as "DNA methylation purity." We previously used this technique to corroborate cell identity and estimate potential residual cross-contamination during the flow cytometry procedures[24]. The mean DNA methylation purity for these samples was 97.6% (range: 85.7–100%), where a target celltype purity of 85% was required to include the sample in the

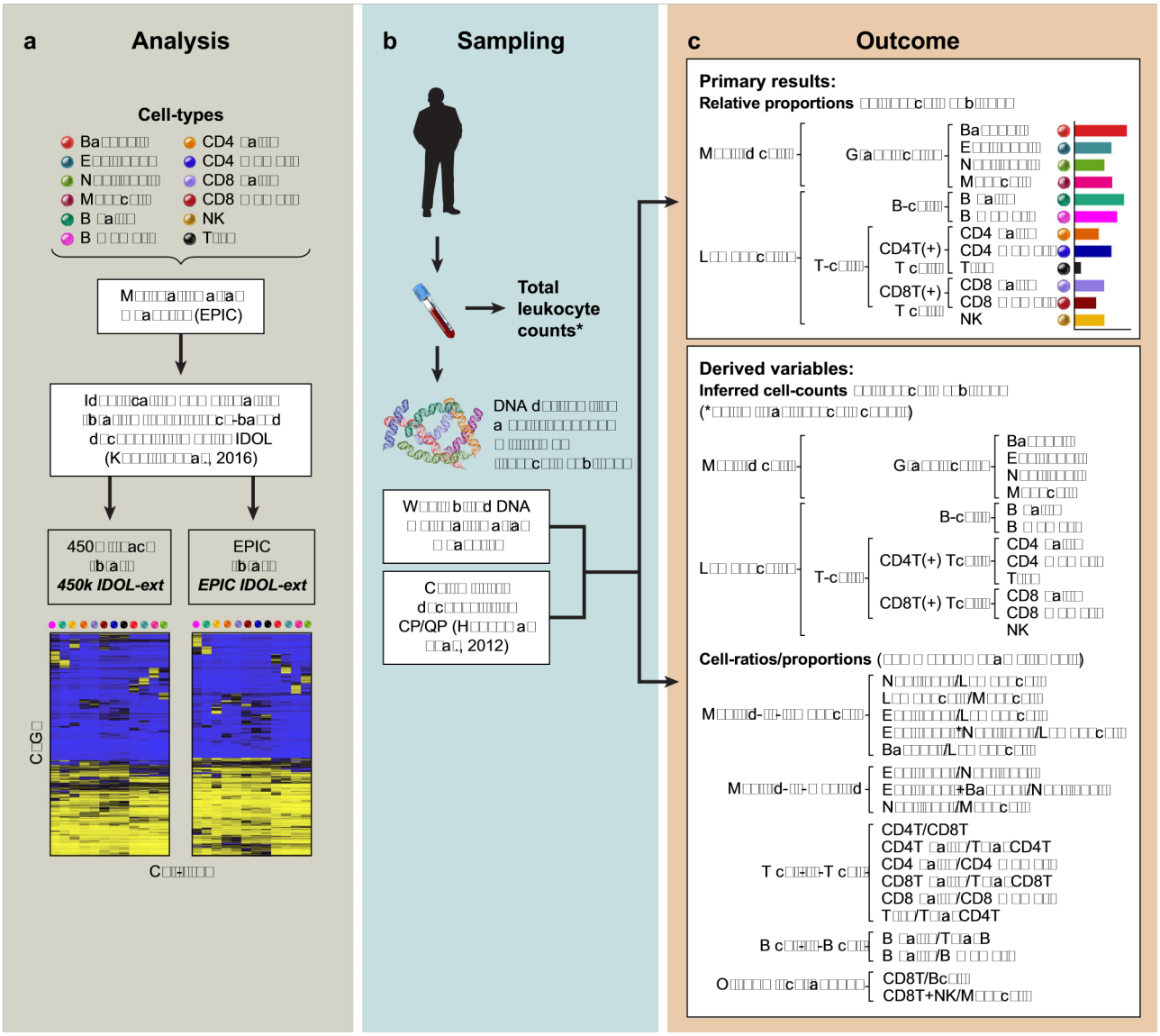
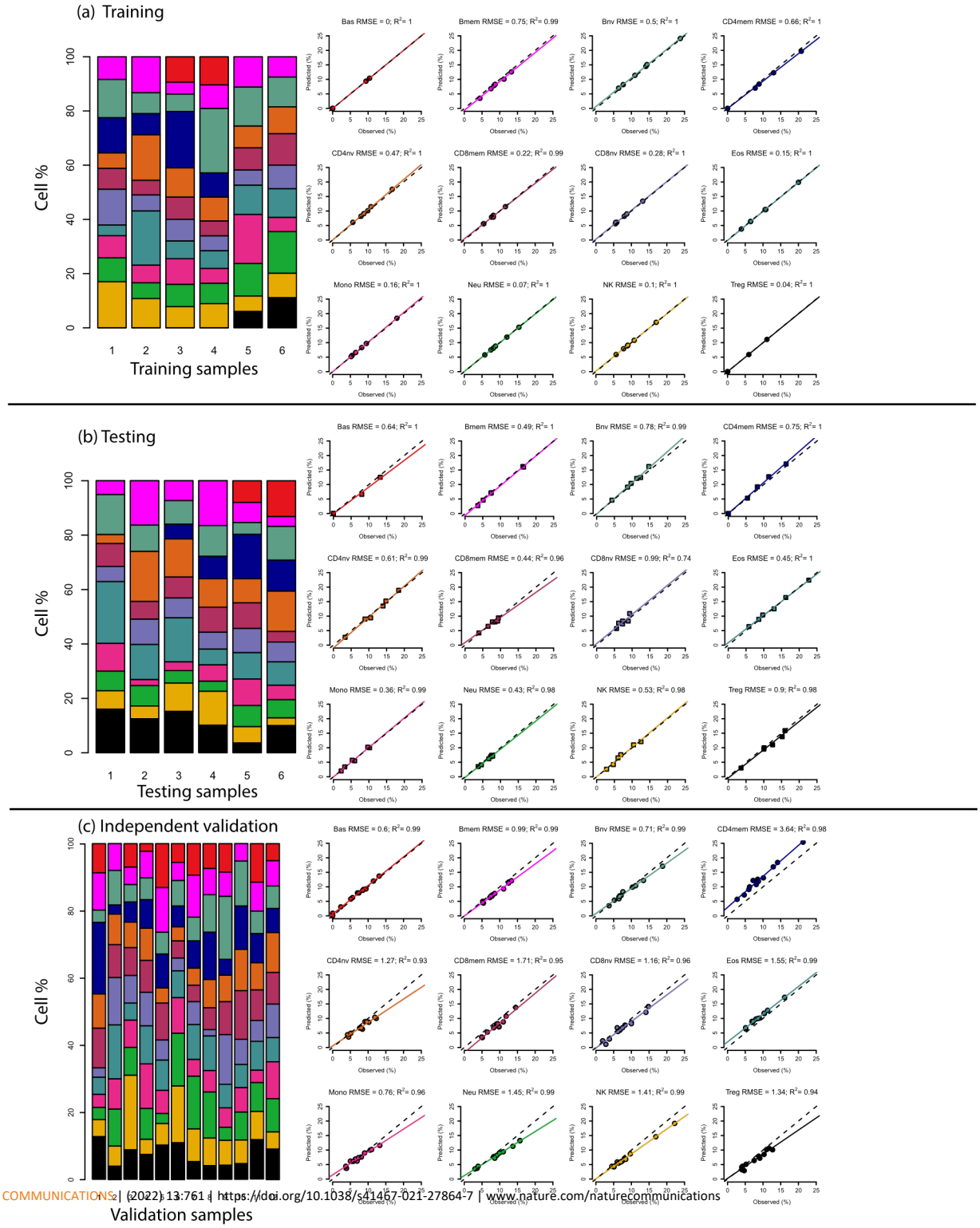


Fig. 1 Extended library development, samples intended analysis and expected and optional measures from the library application. a Twelve cell types were acquired commercially, their DNA was isolated, and DNA methylation was measured using the Illumina HumanMethylationEPIC (EPIC) microarray. Using artificial mixtures as ground truth, two libraries were identified using an iterative process named IDOL (IDentifying Optimal Libraries, Koestler, et al. 2016). The two extended (ext) libraries were designed for microarray data derived from the EPIC array (EPIC IDOL-ext) or legacy data derived from the previous Illumina-HumanMethylation450k array (450k IDOL-ext). b Samples with variable amounts of leukocytes are arrayed using any of the two validated microarray technologies. Using the appropriate library for the microarray, a cell mixture deconvolution is performed using the constrained projection/quadratic programming (CP/QP, Houseman, et al. 2012). *Optionally, leukocyte counts can be collected for downstream analyses. c The primary results of the deconvolution are the 12 cell types of the library. These results could be aggregated at different levels for different hypotheses. Two sets of derived results are possible: (1) if total leukocyte counts are available cell-type-specific counts may be inferred, or (2) cell ratios and proportions are used to evaluate immune-cell shifting between the different cell-type subpopulations (only a few examples illustrated here).

dataset for library construction. Of the cell subtypes included, CD4nv had the highest estimated cell purity (range: 95.2–100%, median DNA methylation purity: 100%, interquartile range-IQR: 96.5–100%), and CD8mem cells had the lowest DNA methylation purity (range: 85.7–98.5%, median DNA methylation purity: 93%, IQR: 88.3–97.8%). The remaining cell types had median DNA methylation purity that ranged between 97.5 and 99.4%, IQR: 97.3–99.4%, (Supplementary Fig. 1). Potential genetic sources of variability were assessed, including known SNPs tracing genetic ancestry (Supplementary Fig. 2). We excluded probes potentially tracking to polymorphisms, cross-reactive areas, or CpHs, probes tracking to sex chromosomes, and those whose signal intensities were equal or below to the background probes (see Methods for details). After filtering, 675,992 high-quality probes were retained for analysis and deconvolution library construction. The first 20 principal components showed that the main sources of variability corresponded to the cell-type identity and the slide beadchip, not to sex, age, or other phenotype variables (Supplementary Fig. 3). Due to the distribution of the 12 cell types across 26 different slide chips, some residual experimental variance cannot be corrected in the modeling. As such, we strived for the highest quality data to eliminate most of the additional technical variability in the experiment.

We next applied the IDOL algorithm to the 56 samples collected across all 12 cell types using the 675,992 high-quality CpGs to select the optimal library for accurate deconvolution of these 12 cell types. To test our library's performance, we used artificial mixtures (n = 24) of DNA from purified cell types representing varying proportions of the 12 cell types in the IDOL analysis and measured DNA methylation in these samples (Fig. 2, Supplementary Table 3). Using a discovery selection pool totaling 3535 CpGs representing candidate markers of differentially methylated CpGs across the interrogated cell types, we implemented the IDOL algorithm with 500 iterations to select a set of libraries ranging in size from 250 to 3000 CpGs, in increments of 50 CpGs (detailed in the methods). The coefficient of determination (R^2) and the root mean square error (RMSE) were calculated based on a comparison of deconvolution estimates obtained using each library versus the known proportions of the 12 cell types in the artificial mixture samples (Supplementary Table 4). The optimal library, EPIC IDOL-Ext, consisted of 1200 CpGs and was observed to have an average R^2 of 1 across all cell types and the lowest average RMSE (0.226) among the interrogated library sizes (Fig. 2a). The testing dataset showed high R^2 and low RMSE across the different interrogated



cell types (Fig. 2b). The library was further validated using a set of 12 artificial mixtures obtained from independently isolated samples (purity > 90%) not used in the library construction (Fig. 2c), with a high coefficient of determination and low RMSE. We observed some slight increase in the error estimations for CD4mem (RMSE = 3.64) with compensatory decreases in the Treg, CD4nv, and CD8mem compartments.

Probes in the EPIC IDOL-Ext library tracked mostly to open sea regions of low CpG density (76%), followed by CpG island shore regions (14%) (see Table 1), and were enriched for open sea regions (OR: 2.19 95% CI:1.93–2.49), and depleted for probes in CpG Islands and Shores (See Supplementary Table 5 for details). Probes were enriched for DNAse hypersensitivity sites-DHS and Phantom5 enhancers. Notably, the library probes were also highly distinct as only 4% overlapped with the EPIC IDOL-6[19].

As only 459 (38%) of the 1200 probes in the EPIC IDOL-Ext library are common to both the EPIC and 450k array platforms, we developed a library for the legacy IlluminaHumanMethylation450k array platform, repeating the above-described optimization process after constraining the selection pool for the candidate list of CpGs to those only present on the 450k array. The resulting library, 450k IDOL-Ext, contains 1500 probes; the library details are included in Table 1 and Supplementary Table 5. One thousand five hundred contained a lower proportion of open sea (52%) and enhancer-related probes (5%) than the EPIC IDOLExt library, although Phantom5-enhancer probes and DHS probes were enriched compared to the total background probes in the microarray. In total, 330 probes were shared between the 450k IDOL-Ext and EPIC IDOL-Ext libraries.

Finally, we compared the EPIC IDOL-Ext, and 450k IDOL-Ext versus the pickCompProbes EPIC library obtained using functions in the minfi Bioconductor package[11]. The pickCompProbes automatic selection method builds the library picking the top 50 most hyper- and hypomethylated CpGs per cell type, totaling 1200 probes (same size as our EPIC IDOL-Ext library), summarized in Table 1 and Supplementary Table 5. The probes overlap from the pickCompProbes library with the EPIC IDOLExt library was only 147 (12%). However, the probes' genomic context distribution was similar between libraries, with similar enrichment of open sea and larger Phantom5-enhancer regions than the EPIC IDOL-Ext. When analyzing the cell type by celltype estimation performance across all three libraries, there was consistency in monocytes, neutrophils, and NK cells. However, more variability was observed for estimates obtained from the library derived by pickCompProbes. When assessing the deconvolution accuracy of the pickCompProbes library, the

Fig. 2 Comparison of estimate cell proportions using constrained projection/quadratic programming (CP/QP) versus the reconstructed (true) DNA fraction in the ar mixtures using the EPIC IDOL-Ext method. a Cell-specific DNA proportions per sample included in the training set, R^2 and Root Mean Square Error-RMSE using the EP method per cell type. b Cell-specific DNA proportions per sample included in the testing set, R^2 and RMSE using the EPIC IDOL-Ext method per cell type. c Cell-s proportions per sample included in an independent validation set (using isolated cells not included in the testing or training), R^2, and RMSE using the EPIC IDOL-Ext met type. Source data are provided as a Source Data file. Bas Basophils, Eos eosinophils, Neu neutrophils, Bnv B-cells naïve, Bmem B-cells memory, CD4nv helper CD4(+) T CD4mem helper CD4(+) T-cells memory, Treg CD4(+) T regulatory cells, CD8nv cytotoxic CD8(+) T-cells naïve, CD8mem cytotoxic CD8(+) T-cells memory, Mono monocytes killer cells.

Table 1 Characteristics of the context of the different blood cell-type deconvolution libraries.

	EPIC IDOL-Ext n = 1200 N (%)	450k IDOL-Ext n = 1500 N (%)	EPIC IDOL-6 n = 450 N (%)	pickCompProbes n = 1200 N (%)
Genomic context				
CpG Island	50 (4)	122 (8)	10 (2)	62 (5)
Shores	189 (16)	388 (26)	63 (14)	146 (12)
Shelves	95 (8)	213 (14)	35 (8)	90 (8)
Open Sea	866 (72)	777 (52)	342 (76)	902 (75)
Functional context				
Promoter	276 (23)	558 (37)	104 (23)	271 (23)
Exon	88 (7)	144 (10)	25 (6)	100 (8)
Intron	551 (46)	556 (37)	201 (45)	548 (46)
Intergenic	285 (24)	242 (16)	120 (27)	281 (23)
Enhancers	139 (12)	76 (5)	70 (16)	194 (16)
DHS	856 (71)	1001 (67)	328 (73)	868 (72)
Open chromatin	135 (11)	174 (12)	43 (10)	113 (9)
TFBS	166 (14)	193 (13)	59 (13)	147 (12)
Contained in 450 K	459 (38)	1500 (100)	149 (33)	517 (43)
Total overlap vs. Ref				
vs. EPIC IDOL-Ext	Ref	330 (22)	43 (10)	147 (12)
vs. 450k IDOL-Ext	330 (28)	Ref	27 (6)	89 (7)
vs. EPIC IDOL-6	43 (4)	27 (2)	Ref	57 (5)
vs pickCompProbes	147 (12)	89 (6)	57 (13)	Ref

DHS DNase Hypersensitive site, TFBS transcription factor binding site, Ref reference for comparison in the column, Enhancers Phantom5 enhancers. Genomic context, Enhancers, DHS, and open chromatin information were extracted from the Illumina EPIC annotation file "IlluminaHumanMethylationEPICanno.ilm10b5.hg38". Functional context information was extracted from the UCSC reference genome file "UCSC_hg19_refGene.bed".

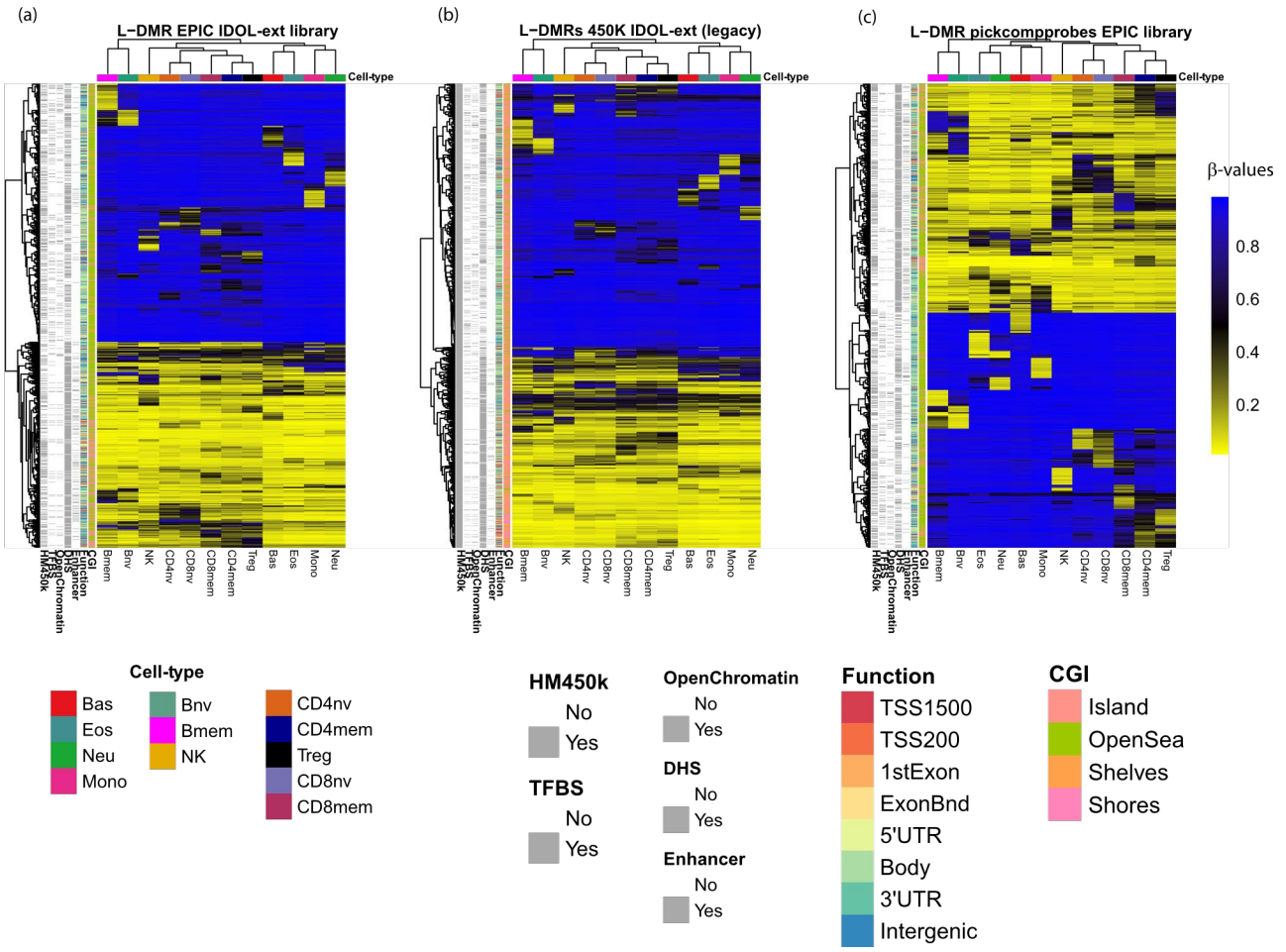


Fig. 3 Comparison of the selected CpG among the EPIC IDOL-Ext, 450k IDOL-Ext (legacy), and pickComProbes. a DNA methylation of the 1200 CpG probe EPIC IDOL-Ext library with average methylation for cell-type samples shown in columns as labeled and tracking bars for CpG probes information in the rows, including: present on 450k array (HM450k), tracking to transcription factor binding site (TFBS), DNA hypersensitivity site (DHS), enhancer region from Phantom5 annotation, and genomic context relative to gene (Function) and CpG island (CGI). b DNA methylation of the 1500 CpG probe 450k IDOL-Ext (legacy) library and c DNA methylation of the 1200 CpG pickComProbes library. The average DNA methylation levels (here as beta values) are represented per each of the 12 cell types. Information about CpG island gene context, Phantom5-enhancer information, DNase Hypersensitivity sites (DHS), Open chromatin, and annotated transcription factor binding sites (TFBS) from ENCODE included in the Illumina annotation file are summarized using the row ribbons. Source data are provided as a Source Data file. Bas Basophils, Eos eosinophils, Neu neutrophils, Bnv B-cells naïve, Bmem B-cells memory, CD4nv helper CD4(+) T-cells naïve, CD4mem helper CD4(+) T-cells memory, Treg CD4(+) T regulatory cells, CD8nv cytotoxic CD8(+) T-cells naïve, CD8mem cytotoxic CD8(+) T-cells memory, Mono monocytes, NK natural killer cells.

RMSE was severely biased for eosinophils and T-cell subtypes. Specifically, CD4 and CD8 naïve T-cell distributions were biased with an underrepresentation of CD8nv (RMSE: 5.81%) and the overestimation of CD4nv (RMSE: 9.25%). The pickCompProbes library also had unreliable results for CD4mem vs. Treg compartments and Eos, Supplementary Fig. 4. In contrast, both the EPIC IDOL-Ext and 450k IDOL-Ext libraries were highly accurate in the training and testing datasets, Fig. 2 and Supplementary Fig. 4. The heatmaps summarizing the markers in the three libraries are shown in Fig. 3. The complete set of markers information is available as Supplementary Data Files 1 (EPIC IDOL-Ext), 2 (450k IDOL-Ext), and 3 (pickCompProbes).

Libraries validation. The EPIC IDOL-Ext library was validated using samples with blood cell counts from flow cytometry (FCM) and by using an independent set of artificial mixtures from the Gene Expression Omnibus (GEO) (Fig. 4). In five samples (Fig. 4a, GSE110530) from a healthy male subject in his forties, who were followed longitudinally for 400 days, we observed a strong correlation between the deconvolution estimates and FCM measurements, with a maximum root mean square error- RMSE of 2.15% for CD4T cells[19]. Using independent artificial mixtures from the previously published FlowSorted.Blood.EPIC library with six known cell types (Neu, CD8T, CD4T, B cell, Mono, NK; Fig. 4b, GSE110554), the correlation was close to 1, and the maximum RMSE was 3.39 for the CD8T. The 450k IDOL-Ext library was also validated using the GSE77797 dataset with six anonymous adult volunteers [5 males, epigenetic age mean (sd): 34.1(11.4) years] with FCM information (Fig. 4c) and six artificial mixtures with six known cell types (Neu, CD8T, CD4T, B cell, Mono, NK). The information from this dataset was used in the original IDOL manuscript (Fig. 4d)[12]. Of note, the coefficient of determination was lower for monocytes (0.19) and

NK (0.66), Fig. 4a, c, respectively, using the EPIC IDOL-Ext library. This decrease is driven by the narrower range of distribution of these cells in the FCM samples. Consequently, any minor deviation in the estimation is magnified even with the high precision reflected by the RMSE.

As a proof-of-concept, we explored how the cell proportion estimates from the EPIC IDOL-Ext library could recapitulate cell counts using FCM cell count information from the GSE112618 dataset (Supplementary Fig. 5)[19]. This dataset included six subjects [five males: mean (sd) age 42.8 (10.6) years] with counts for nine cell types (Bas, B cell, CD4T, CD8T, Eos, Neu, Mono, NK, and Treg). Additional validation sets were analyzed, including glioma patient blood and the cytometric information for the Reinius dataset[17]. A group of 76 glioma patients (GSE180683) was used to evaluate the information from T-cell subsets. The patients were 51% male with a mean (sd) epigenetic age (Horvath) of 54.7 (14.1) years. For this study, cytometric information was obtained using a two-stage characterization of T-cell subtypes. Briefly, T cells were measured and separated into CD4 + and CD8 + . Next, using a second tube, samples were characterized as naïve or memory. Although samples were measured from the same subjects, the proportions were estimated based on two independent measurements and were mathematically derived based on both experiments' average counts. We observed the second-largest difference for the CD8mem with an

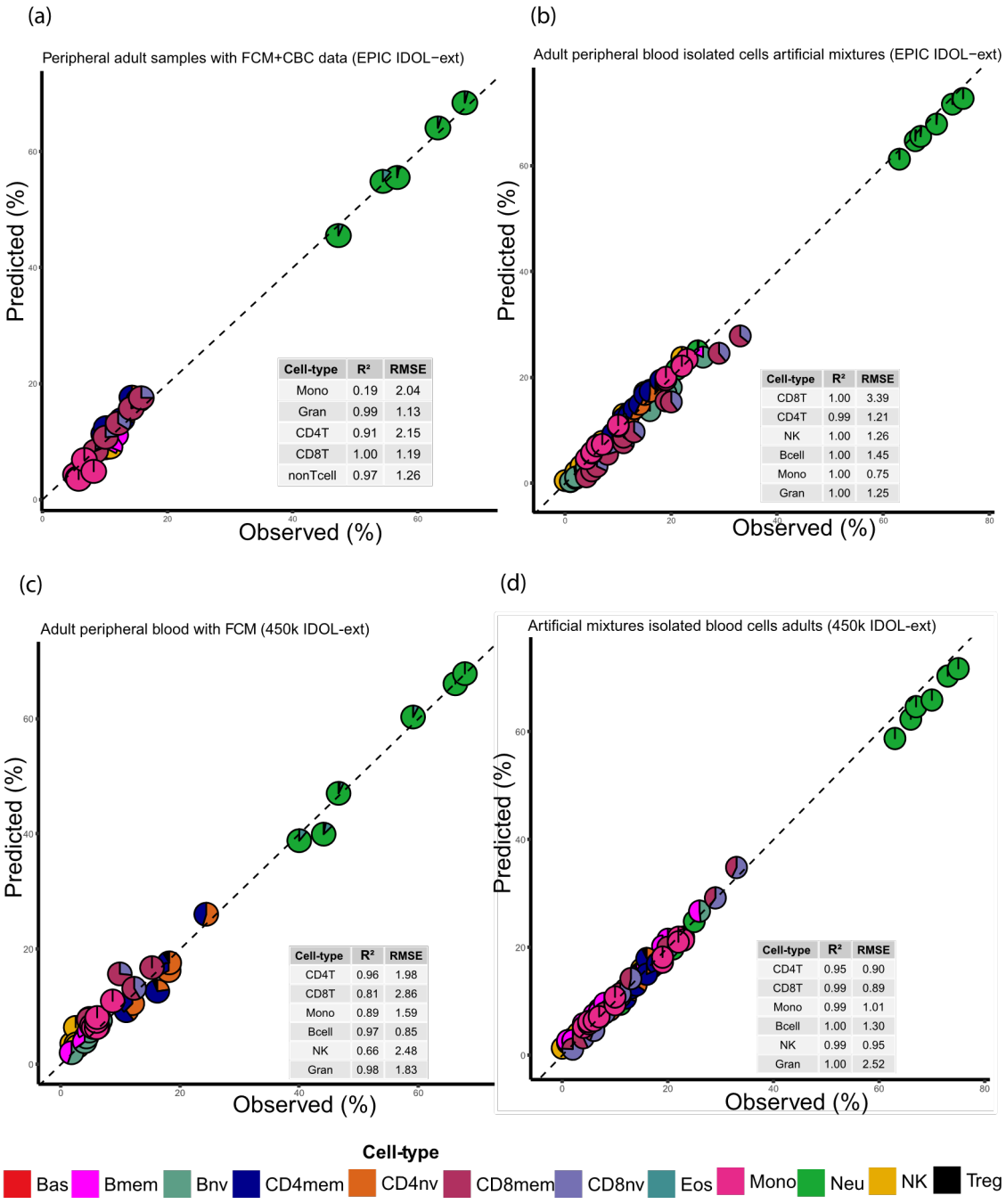


Fig. 4 Validation of the library using flow cytometry-FCM and independent artificial mixtures for adult peripheral blood. a and b Vaalidation of the EPIC IDOL-Ext. c and d Validation of the 450k IDOL-Ext. The area of each pieplot corresponds to the estimated proportion of the cell types within each group. Gran (Granulocytes) corresponds to the sum of Neu-neutrophils, Eos-eosinophils, and Bas-basophils. CD4T corresponds to the sum of CD4 + T cells naïve-CD4nv, memory-CD4mem and Treg. CD8T corresponds to the sum of CD8 + T-cells naïve-CD8nv and memory-CD8mem. B cell to the sum of the naïve-Bnv and memory-Bmem. No basophils were detected in the samples illustrated in a. Additionally, in a, "nonTcell" (Bnv, Bmem, and NK) were not experimentally measured. Source data are provided as a Source Data file.

RSME of 4.34% and the largest with the unmeasured remainder cells (RMSE of 12.6%) (Supplementary Fig. 6a). The 450k IDOLExt was also validated using the GSE35069 datasets (Reinius) with FCM information (Supplementary Fig. 6b). The Reinius dataset was composed of six male subjects in their thirties. FCM information was obtained for whole blood, peripheral blood mononuclear cells, and granulocytes for six cell types (granulocytes, CD8T, CD4T, Mono, NK, and B cell). The FCM information is abstracted from the supplementary material in the original manuscript by Reinius et al. We observed considerable variation (~10%) for the granulocytes, monocytes, and CD4T cells in this dataset[17]. Differences were more prominent for the PBMC samples' observed values and in two of the six WBC samples, particularly for the granulocytes reported.

Finally, we explored whether these libraries could be applied for the deconvolution of umbilical cord blood samples. We used the Jones et al. GSE127824 dataset (450k) with known FCM information for seven cell types (erythroblasts-nucleated red

blood cells-nRBC, granulocytes, CD8T, CD4T, Mono, NK, and B cell)[24]. As erythroblasts were not present in our libraries, these cells are only briefly present in newborns' blood but not at any other age, the fraction was added to the myeloid granulocyte components for comparison. The rationale for adding the erythroblasts to the granulocytes is based on the known biology that establishes the closest lineage to the erythroid-myeloid precursor[25]. This is reinforced by clustering of nucleated red blood cells with basophils and eosinophils using the markers in EPIC IDOL-Ext (Supplementary Fig. 7a). The nRBC fraction showed the most considerable variability in this dataset, with an RMSE of 16.02% (Supplementary Fig. 7b). In contrast, in an experiment of 12 umbilical cord blood artificial mixtures (GSE180970) from healthy anonymous newborns acquired from commercial vendors with only six cell types (granulocytes, CD8T, CD4T, Mono, NK, and B cell), in which the largest RMSE observed for granulocytes was less than 3% (Supplementary Fig. 7c). Of importance, the B cells and T cells clustered with the purified adult naïve cells, while NK, monocytes, and neutrophils clustered perfectly with their adult equivalents.

Biological interpretation of the libraries' components. The different libraries' components were examined for how the included genes were related to specific cell types and immune pathways. First, we used the molecular signatures database (MSigDB v. 7.2), curated by the Broad Institute, for gene set enrichment analyses. Most of the pathways were immune-related and are summarized in Supplementary Data File 4. Some specific genes showing differential hypomethylation for specific cell types are outlined in Fig. 5. We also used eForge (https://eforge.altiusinstitute.org/) to identify whether the components of the library were related to particular primed enhancer histone marks (H3K4me1). When analyzing probes specific for the 12 cell types, all subsets were enriched for blood components, but the top significant results were consistent with the specific cell type: E029 Monocytes with monocytes (Supplementary Fig. 8 panel f), E032 B cells with Bnv, and Bmem (panels b and a), E033 T cells with CD4nv, CD4mem, Treg, CD8nv and CD8mem (panels i, h, j, l, and k, respectively) and E046 Natural killers with NK (panel g). There were no specific datasets to compare the granulocytes (Bas, Eos, Neu), where the signal was less straightforward (panels c, d, and e).

Applications of the libraries. We next applied the EPIC IDOLExt library and the 450k IDOL-Ext libraries to several publicly available datasets from GEO and ArrayExpress to identify potential variation in immune-cell proportions in multiple sclerosis, rheumatoid arthritis, breast cancer patients, and COVID-19 infection (Supplementary Table 6). In multiple sclerosis patients, we observed significantly increased basophil and naïve B cell proportions in cases ($n = 13$) compared to controls ($n = 14$) (Wilcoxon rank-sum $P < 0.01$, Supplementary Fig. 9). In rheumatoid arthritis, significant increases in neutrophil and regulatory T-cell proportions and decreases in memory B cell, memory CD4 cell, naïve CD4 cell, memory CD8 cell, naïve CD8 cell, eosinophil, monocyte, and NK cell proportions were observed in case blood samples ($n = 354$) compared to control blood samples ($n = 355$) (Wilcoxon rank-sum $P < 0.01$, Supplementary Fig. 10). Predicted immune-cell proportions in breast cancer patients before and after receiving chemotherapy or a combination of chemo/radiation therapy were explored. Patients receiving radiation therapy only ($n = 74$) showed a significant relative increase of the neutrophil proportion and a mirror decrease of several lymphoid lineages (memory B cell, naïve B cell, naïve CD4 cell, and NK) after treatment (Wilcoxon ranksum $P < 0.01$, Supplementary Fig. 11a). Patients receiving radiation therapy and chemotherapy ($n = 70$) exhibited significant increases in eosinophil, monocyte, and regulatory T-cell proportions and decreases of memory B cells after treatment (Wilcoxon rank-sum $P < 0.01$, Supplementary Fig. 11b). Finally, we evaluated the changes in immune-cell proportions in six COVID-19 patients with and without remission versus six healthy controls in Supplementary Fig. 12. Because of the limited sample size, no statistically significant differences were observed, but, as expected, the median of neutrophils was higher in patients versus controls. In contrast, all the median lymphocyte subpopulations were lower in the infected patient than those who recovered from the disease. The median monocytes were lower in those that remitted.

We investigated blood methylation data associated with subject-to-subject variation in non-pathological conditions using data from monozygotic versus dizygotic twins and subjects at different ages (Supplementary Table 6). To characterize immune cell variation between twins, predicted immune cell proportions in monozygotic ($n = 852$) and dizygotic twins ($n = 612$) were estimated. Significant differences in all the immune cells were observed in monozygotic twins and dizygotic twins (Paired t-test $P < 0.01$). More considerable differences between dizygotic twins and monozygotic twins were seen in memory B cell, naïve B cell, memory CD4 cell, naïve CD4 cell, memory CD8 cell, naïve CD8 cell, eosinophil, monocyte, and NK (Wilcoxon rank-sum $P < 0.01$, Supplementary Fig. 13). Next, the impacts of aging on predicted immune cell proportions in blood samples from newborn to nonagenarian ($n = 2504$) were investigated. Extensive literature has explored the changes in immune proportions with aging[26,27], but it has not been systematically explored using DNA methylation[28–30]. Several subpopulation cell ratios across different ages were calculated (Supplementary Fig. 14). The complete granular data and trajectories of the different cell subpopulations are represented in Supplementary Fig. 15. Longitudinal changes of predicted immune cell proportions within five years after birth in human blood leukocytes from 10 healthy girls are shown in Supplementary Fig. 16.

We included some sensitivity analyses to compare the 450k IDOL-Ext vs. the EPIC IDOL-Ext libraries when using EPIC or a combination of EPIC and 450k legacy arrays. As shown in Supplementary Fig. 4, both libraries provide an accurate estimation for the 12 cell types. There are minor variations in Bmem and Bnv estimations, with more precise estimates using the EPIC IDOL-Ext (absolute mean difference 0.08%-EPIC vs. 0.48%–450k, and 0.23-EPIC% vs. 0.52%–450k, respectively), and CD4mem and Treg with more precise estimations using the 450k IDOL-Ext (absolute mean difference 0.62%-EPIC vs. 0.07%–450k, and 0.6-EPIC% vs. 0.34%–450k, respectively), see Supplementary Fig. 17 for additional information. Results using other statistical procedures for cell deconvolution, CIBERSORT,

and robust partial correlations, were compared and are illustrated in Supplementary Fig. 18. Results were generally similar using the different methods, though slightly lower accuracy was observed for CIBERSORT compared to CP/QP or robust partial correlations.

Discussion

We established compact and reliable libraries to deconvolve the proportions of 12 different cell types in peripheral blood, including closely related cell types such as T-cell subtypes and various types of granulocytes. Importantly, the libraries' derived variables offer a detailed immune profile from peripheral blood, providing more than 56 cell-type and ratio components. Our libraries are designed both for current (EPIC) and legacy (450k)

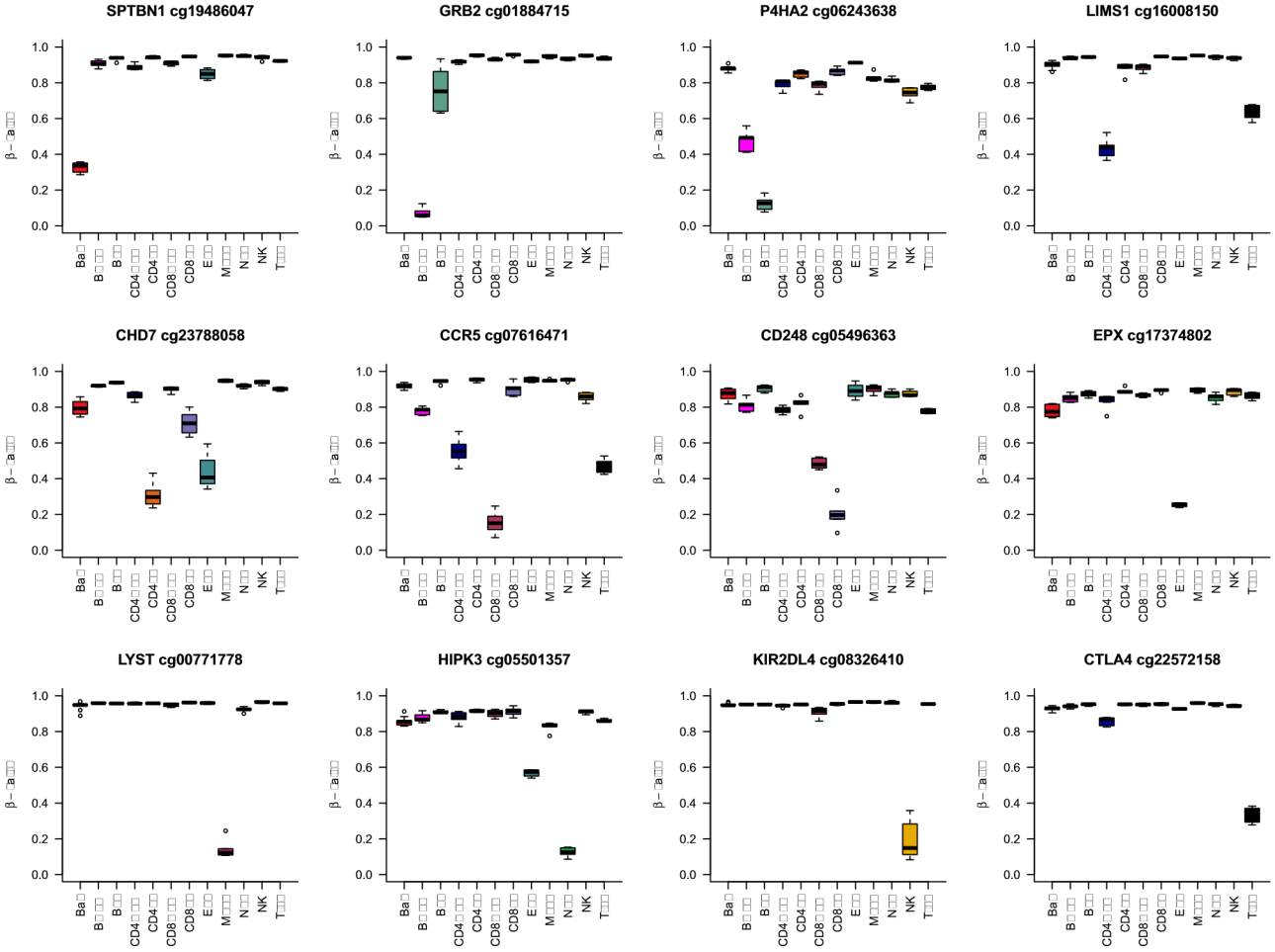


Fig. 5 Examples of differentially methylated sites demarcating the different cell lineages. The boxplots include the following information: (1) The box shows the interquartile range (IQR), (2) the whiskers show the inner fences (1.5 × IQR out of the box), (3) the bolded line shows the median of the data. The color inside the box corresponds to the cell type. Source data are provided as a Source Data file. Bas Basophils, Eos eosinophils, Neu neutrophils, Bnv B-cells naïve, Bmem B-cells memory, CD4nv helper CD4(+) T-cells naïve, CD4mem helper CD4(+) T-cells memory, Treg CD4(+) T regulatory cells, CD8nv cytotoxic CD8(+) T-cells naïve, CD8mem cytotoxic CD8(+) T-cells memory, Mono monocytes, NK natural killer cells.

DNA methylation measurement platforms. The Gene Expression Omnibus repository, as of July 2021, includes publicly available data from ~148,325 samples for both platforms[31]. Our libraries have high accuracy and minimal bias compared to flow cytometry, considered the "gold standard" data (Fig. 4). For most cell types, that accuracy represents a difference of less than 3% in the estimations. However, more considerable differences (>4%) are observed in some datasets for the most abundant cell types (neutrophils and total CD4T, Supplementary Figs. 6 and 7) and for CD8Tmem (Supplementary Fig. 6). This error results from limited purity in this cell component and possibly differences in gating processes during the count splitting[32]. For the CD8mem compartment, it is also possible that unaccounted residual heterogeneity is present, TEMRA (T effector memory cell expressing CD45RA isoform), effector memory, central memory, NKT cells, and other less characterized CD8Tmem subclasses[33]. The new libraries will be important where specific memory and naïve cell compartments are interrogated for disease, pathogenesis, or exposure. Both EPIC IDOL-Ext and 450k IDOL-Ext are accurate, with the former requiring fewer probes for the estimation. In complex problems, such as EWAS using EPIC and 450k technologies simultaneously, we recommend using the more straightforward approach for the data using the 450k IDOL-Ext given the minimal variation observed (<1%) in the estimates. We advise the researchers to evaluate their data quality before using any deconvolution approach to avoid potential technical biases (e.g., examining the library probe call rate and possible batch effects in the data).

We consistently observed biologically relevant genes and groups of genes in related pathways for the different cell types,

such as the SPTBN1, which is more expressed in basophils[34]; GRB2, involved in Bmem formation[35,36]; P4HA2, related to immature B cells[37]; CHD7, associated with thymic function[38]; CCR5, related to CD8mem recruitment and trafficking[39]; CD248 a CD8 regulator and marker of naïve states[40]; EPX, the eosinophil peroxidase marker of eosinophils[41]; LYST a lysosome maturation in monocytes[42]; or KIR2DL4 a natural killer cell receptor[43]; and the CTLA4 an essential marker in Treg tumor suppression[44].

Our EPIC IDOL-Ext and 450k IDOL-Ext libraries provide enhanced features for researchers using DNA methylation data derived from blood. When the research question requires precise resolution, the newly extended libraries provide that extra information, with a minimal trade-off in precision compared with the earlier EPIC IDOL-6 library. Previous attempts to generate extended libraries for blood deconvolution encountered several problems. The discrimination of scarcer cells has been problematic, leading to inconsistent estimates; even when candidate references were available, this library has largely overcome those challenges. Some markers delineating particular cell types are dependent on sex chromosome dosage (e.g., FOXP3 located in the X chromosome for Tregs). To avoid that potential problem here, we only used autosomal probes. Distinguishing hierarchically close cell types such as CD4mem and Tregs is challenging as the pool of potential specific markers is limited. These markers may even show some dependency when transitioning from one state to another. For example, CD248 is mostly unmethylated in CD8nv, but some cells in the CD8mem compartment are not completely methylated at CD248, while it is completely methylated and repressed in other cell types. In such case, using multiple markers attenuates the effect of using single markers to define cell identity. Finally, the statistical process may favor markers that do not necessarily reflect biologically relevant genes when using libraries derived from an automatic selection. For example, four (EPIC) and five (450k) EPX probes were selected in the extended libraries, in contrast to none by the minfi pickCompProbes procedure. EPX is the gene containing the information for the eosinophil peroxidase protein, which is essential for eosinophil function and solely expressed in this blood cell type[41,45].

Including six additional cell types and, in particular, the naïve compartments is a significant step towards a universal deconvolution library for application to cord or adult blood data. Particularly intriguing for umbilical cord blood, nucleated red blood cells interfered with more precise estimations than available FCM information. Yet artificial mixtures showed the library capturing leukocyte components accurately. As expected, most (if not all) of the B and T cells were naïve cells as memory populations are derived after antigen exposures, most commonly after birth.

Another advance of this library is the most accurate eosinophil predictions compared to previous libraries and now basophil estimations. These cells are involved in multiple pathologies, including allergies and asthma, and through these libraries, they can now be estimated. As with previous libraries, some residual confounding from unaccounted cell types may be observed. In particular, we observed inflation in several datasets of the basophil estimation. Based on single-cell murine data, one hypothesis for this estimate inflation is common progenitors' relation with triple differentiation to basophil/mast cells and erythroid cells[46,47]. According to those potential common markers, it is possible the basophil signal could interfere with the signal derived from circulating immature nucleated erythroid cells. At least in umbilical cord blood samples, the signal obtained from basophils is highly correlated with the predicted fraction of nucleated red blood cells. As such, a careful assessment of any high basophil signal is warranted, as it could be driven by interference with nucleated red blood cells in newborns and potentially some pathological or adaptative conditions at other ages. Thus, careful interpretation based on those estimates alone is prudent as they may signal residual cell signals of other cell types not included in the library.

We applied this library to several public datasets, comparing cases and controls for several pathologies. Our results are concordant with known pathophysiology and immune responses for these diseases: increased proportions of abnormal naive B cell populations in the blood in multiple sclerosis patients[48], pancytopenias with a predominance of neutrophils in radiotherapy plus chemotherapy in breast cancer[49], predominancies of neutrophilia, and lymphopenia in COVID-19 patients that did not show remission[50]. However, the limited sample sizes of these data preclude more generalizable conclusions for most of the findings. We will only discuss the most extensive dataset available of rheumatoid arthritis patients vs. controls. This dataset, published by Liu et al[51]., has explored how some immune cell components (monocytes) are associated with specific DNA methylation changes in the cases who were not receiving any treatment at the time of blood collection. We observed a reduction in monocyte proportions in rheumatoid arthritis patients compared to controls (P-value = 1.1 E-5). Paradoxically in our analysis, we found a non-statistically significant increase in Tregs in cases compared to controls (P-value = 0.06), which are related to disease activity or response to medications in other studies[52]. These differences are essential to understand the activity of the disease and potentially for the classification of the heterogeneous subjects when compared to healthy controls. It is difficult to predict the extent of discoveries made possible with the extended capability for immune profiling we provide. Our method gives higher resolution than clinical approaches allows for more scalable, efficient, standardized immune profiling compared with flow cytometry, and has the added benefit of applying to archival specimens. Although we have only begun to demonstrate the potential of our approach, we do not doubt that the research community will leverage this method to provide new insights into the relation of immune profiles in human health and disease in exciting and meaningful ways that impact public health and translational medicine.

In summary, the new reference libraries greatly enhance the detail of immune cell profiling with DNA methylation. These enhanced libraries can be applied directly for immune profiling and adjustment of cell-type proportions in EWAS. Future work includes validating the use of this library in methylation data from children and umbilical cord blood and expanding the libraries to additional cell subsets critical for some pathologies (e.g., dendritic cells). The increased detail in describing leukocyte subtypes using the extended libraries will be crucial for controlling for aging-related DNA methylation changes; age-associated changes in specific subpopulations of memory T and

B lymphocytes are unaccounted for in previously established libraries. Indeed, the IDOL-Ext library applications developed here greatly enhance the detail of immune characterization for existing immune profiles data and open opportunities for future studies using DNA methylation data derived from human blood samples.

Methods

This work extended the available six-cell reference library using twelve cell subtypes for deconvolution of blood cell proportions using the EPIC array, as well as a legacy library for the 450k platform. Using cytometric and magnetic-sorted, flow confirmed neutrophils, eosinophils, basophils, B cells (naïve and memory), monocytes, NK cells, CD4 + T cells (naïve, memory and T regulatory cells), and CD8 + T cells (naïve and memory), DNA methylation was measured with the 850 K/EPIC DNA methylation array. We applied the IDOL method to identify optimal Leukocyte -Differentially Methylated Regions (L-DMR) libraries using a testing set of six artificial mixtures containing the 12 cell types. Artificial mixtures (also referred to as reconstructions) consist of DNA from purified isolated cell types, representing mock blood samples of known, predefined cell proportions. Six additional testing artificial mixtures were employed to corroborate the performance, and 12 independent artificial mixtures derived from a set of 12 isolated cell types not utilized in the training or testing, were used to validate the results. We compared the performance of cell estimates obtained applying our previously developed six model cell type (available in Bioconductor as"FlowSorted. Blood.EPIC") and optimized an additional L-DMR IDOL library limiting the probes to those available in the older 450 K array, and again compared the performance using the training, testing, and validation datasets.

The DNA used to generate mixtures were derived from four MACS-isolated and FACS-verified purity cell subtypes from the myeloid lineage [neutrophils (Neu), eosinophils (Eos), basophils (Bas), and monocytes (Mono)], and eight MACSisolated and FACS-verified purity cell subtypes from the lymphoid lineage [B lymphocytes naïve (Bnv), B lymphocytes memory (Bmem), T-helper lymphocytes naïve (CD4nv), T-helper lymphocytes memory (CD4mem), T regulatory cells (Treg), T-cytotoxic lymphocytes naïve (CD8nv), T-cytotoxic lymphocytes memory
(CD8mem), and natural killer lymphocytes (NK) cells] were purchased from AllCells® corporation (Alameda, CA, USA), StemExpress (Folsom, CA), and STEMCELL Technologies (Vancouver, BC, Canada). Cells were isolated from 41 males and 15 females, all anonymous healthy donors. The donors had a mean age of 32.2 years (sd = 12.2, range 19–58 years) and an average weight of 85.1 kg (range 57–136 Kg). Donors identified themselves from multiple ethnicities, including mixed ethnicities, and were categorized broadly into four groups (African-Americans, East-Asian, Indo-European, multiple/admixed). They were negative for Human Immunodeficiency Virus-HIV, Hepatitis B Virus-HBV, and Hepatitis C Virus-HCV. Women were not pregnant at the time of sample collection, and samples were collected from donors with no history of heart, lung, kidney disease, asthma, blood disorders, autoimmune disorders, cancer, or diabetes. All donors provided written informed consent before donation. The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (Salas et al., 2021) and are accessible through GEO Series accession number GSE167998. Isolation protocols are available through the commercial websites of AllCells, StemExpress, and STEMCELL Technologies. In brief, cells were selected using immunomagnetic labeling through the vendors' specific protocols (see, Supplementary Table 1 for details). Recovered cells were confirmed using flow-sorting. Twenty-four artificial mixtures were determined by randomly generating proportions from a twelve-component Dirichlet distribution. Each mixture of 1.2 µg total DNA was generated from isolated cell DNA using the proportions in Supplementary Table 3. The isolated cell DNA and those of the artificial mixtures were bisulfite converted and processed according to the Illumina protocols at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, Avera, or Diagenode. Samples were randomized prior to loading onto microarray chips. The EPIC methylation array raw idat files were pre-processed using minfi, EnMIX, and SeSaMe for quality control using R v.4.0.2 and 4.1.0[53–55]. To assess data quality, we used an out-of-band detection P-value of 0.05, three standard deviations of the mean bisulfite conversion control probe fluorescence signal intensity, and a minimum of three beads per probe. To ascertain the highest purity of the samples included in our library, in addition to the information obtained through the FCM confirmation, we projected back the proportions, thus "purity," of the cells, using Jaffe's procedure[11]. The corresponding cell-type proportion was retrieved and designated as "DNA methylation purity." Only samples with DNA methylation purity levels higher than 85% (range 85.7 to 100%) were included in the library. Subsequently, a stringent out-of-band p-detection value (pOOBHA) > 0.05 was applied and set those that could not be distinguished from the background probes as "missing values." Only those probes with complete information for all the samples were selected for the library. No imputation was performed in this context as the signals could be dependent on the specific cell type. Additionally, all nonCpG (referred to as CpH) probes were filtered in light of the minimal variation and all CpH beta values were under 0.08. Finally, as both female and male samples were present, we discarded probes tracking the X and Y chromosomes. According to Zhou et al., those that showed known polymorphisms or cross-reactivity were also excluded[56]. Our set for library discovery included 675,992 complete high-quality probes. The EPIC IDOL-Ext L-DMR library is available as an R library "FlowSorted.BloodExtended.EPIC" please contact the Office of Technology Transfer Technology.Transfer@dartmouth.edu for a free Academic license (for license instructions, please refer to https://github.com/immunomethylomics/FlowSorted.BloodExtended.EPIC). The extended blood deconvolution can be performed using the FlowSorted.Blood.EPIC Bioconductor library and we recommend using the minfi noob background correction for the target dataset. The package contains an RGChannelSet R object processed using SeSaMe in which probes showing channel switching were corrected and SNPs derived from Infinium Type I probes were added, using the total signal intensities, to the control for genetic ancestry. The object is unfiltered and contains 56 samples and the 12 artificial mixtures information on 1,008,711 probes corresponding to 866,091 sites (CpGs and CpHs) using the latest annotation released by Illumina (MethylationEPIC_v-1-0_B5). The reader needs to note that the cells were purified using an immunomagnetic procedure; the name "FlowSorted" is kept for historical reasons and downstream integration with previous minfi pipelines and similar algorithms.

IDOL algorithm. For a complete description of the IDOL algorithm, please refer to the original application in Koestler et al.[12]. In brief, the IDOL algorithm utilizes a training dataset (ground truth) consisting of samples with DNA methylation data in which the measured fraction of each of the underlying cell types is known (here corresponding to artificial mixtures with prespecified proportions) as a means to identify a set of probes confirming an optimal reference library for cell mixture deconvolution. A series of t-tests compared the mean CpG-specific methylation between each leukocyte cell type vs. the mean methylation across all the other cell types identified the probes discriminating CpGs (e.g., leukocyte differential methylated regions or L-DMRs) for each specific cell type of the 12 included in this application. CpGs were then rank-ordered using their t-statistics, and the L/2 CpGs with the largest and smallest t-statistic for each K cell type were identified and pooled. Our application set the tuning parameter L to 150 in, consistent with Koestler et al.[12]. A discovery L-DMR library containing the total L*K unique L-DMRs for each cell type forms the IDOL algorithm search space. L-DMR subsets of size <L*K are sequentially selected and examined for their prediction accuracy in deconvolving the training dataset samples. The user needs to preselect the library size to balance the accuracy and precision of cell-composition estimates. For the current application of IDOL presented here, we considered libraries ranging from 250 to 3000 CpGs. We initially set increments of 50 CpGs until a size of 1100; then we increased by 100 CpGs until a size of 2000 and tested libraries of 2500 and 3000 CpGs to corroborate the elbow in the error distribution. In the first iteration of the IDOL algorithm, all L*K CpGs constituting the candidate library has an equal probability of being selected to be included in the L-DMR library. We applied the constrained projection/quadratic programming approach[9] to obtain the cellcomposition estimates for each sample in the training dataset. These predictions then allow us to calculate the R^2 and RMSE (root mean square error) for each cell type, contrasting the cell estimates versus the known proportion in each sample. Then one-by-one CpGs are removed from the randomly selected DMR library, followed by computation of R^2 and RMSE based on cell-composition estimates obtained using the new library. Through this, we can assess the contribution of each CpG in the library in terms of its impact on the accuracy of cell-composition estimates and, then the algorithm modifies the probability of each CpG being selected in subsequent IDOL iterations. The process is repeated at each of the 500 iterations, with the algorithm eventually converging on an "optimal" library for deconvolution (showing the lowest error and highest precision).

Enrichment analysis. Gene set enrichment analyses used missMethyl to control for multiple probes bias and the Molecular Signatures Database (MSigDB) version 7.2[57,58]. Hypergeometric tests were used to evaluate enrichment of specific genomic context and functions compared to the background probes used for selecting the candidates in the library. Additionally, eForge was used to evaluate the enrichment of cell-specific probes in the library versus the presence of primed enhancer histone marks (H3K4me1)[59].

Additional validation sets. We used five datasets for validation: a set of samples obtained from a longitudinal analysis (GSE110530) with six observations. A set of independent artificial mixtures (GSE110554) derived from six cell types. A set of samples from 72 glioma patients in different treatment stages (GSE180683), including FCM information for T cells CD4 + and CD8 + naïve and memory. A set of 20 umbilical cord blood (GSE68456) with FCM information. A set of 12 umbilical cord blood artificial mixtures including six cell types (GSE180970). A set of samples using the 450k technology (GSE77797) with FCM and artificial mixture information derived from six cell types. Finally, the Reinius et al. GSE35069 dataset using the FCM information for the whole-blood cells, peripheral blood mononuclear cells, and granulocytes as reported in the original manuscript's supplementary material.

Potential applications. We identified 12 publicly available datasets from GEO and ArrayExpress that contained DNA methylation data on two different normal health conditions (twins, aging) and four diseases (Supplementary Table 6). The application datasets included whole-blood samples from 426 pairs of monozygotic twins and 306 dizygotic twins, 2504 umbilical cord and peripheral blood samples from newborn to nonagenarian, 13 multiple sclerosis whole-blood case samples and 14 controls, 354 rheumatoid arthritis peripheral blood leukocyte samples, and 355 controls, 144 peripheral blood samples from breast cancer patients before and after receiving chemotherapy and radiation or isolated radiation therapy treatment, and six COVID-19 patients (with 18 samples) and six healthy controls. Illumina Infinium DNA methylation IDAT files were retrieved from GEO and ArrayExpress for application datasets. minfi package from Bioconductor was used to process the data using noob. Four thousand eight hundred and seventy-two samples in total were eventually contained in the application datasets. We estimated the proportion of immune cells in those datasets using the appropriate extended immune cell deconvolution library for the array (EPIC or 450k).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The main source dataset (12 cell types, testing, and training artificial mixtures) generated in this study have been deposited in GEO under accession code GSE167998. This Superseries GSE181034 is composed of the following series GSE180683 (glioma samples, validations for T-cell memory subsets), GSE180970 (umbilical cord artificial mixtures), GSE182379 (independent validation 12 cell types artificial mixtures). Additional public datasets analyzed in this manuscript are available in GEO and ArrayExpress with accession numbers: GSE110554, GSE77797, GSE110530, GSE35069, GSE68456, GSE88824, GSE42861, GSE140038, GSE161778, GSE105018, E-MTAB-7069, GSE85042, GSE103189, GSE104778, GSE62219, GSE87571, E-MTAB-7309, GSE87571, GSE12163. Source data are provided with this paper.

Code availability

The code used for this manuscript has been deposited in Zenodo doi: 10.5281/zenodo.5338513[60]. Instructions to obtain a license for FlowSorted.BloodExtended.EPIC R library are available in GitHub (https://github.com/immunomethylomics/FlowSorted.BloodExtended.EPIC) or through the Technology Transfer Office Technology.Transfer@dartmouth.edu. Free licenses are available solely for NonCommercial Entities to conduct academic research. All other parties will require a specific license. Most of the licenses are granted less than two business days after the agreement signature.

Received: 20 March 2021; Accepted: 20 December 2021;
Published online: 09 February 2022

References

1. Christensen, B. C. et al. Aging and environmental exposures alter tissuespecific DNA methylation dependent upon CpG island context. PLoS Genet. 5, e1000602 (2009).
2. You, C. et al. A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes. Nat. Commun. 11, 4779 (2020).
3. Grieshober, L. et al. Methylation-derived neutrophil-to-lymphocyte ratio and lung cancer risk in heavy smokers. Cancer Prev. Res.11, 727–734 (2018).
4. Koestler, D. C. et al. DNA methylation-derived neutrophil-to-lymphocyte ratio: an epigenetic tool to explore cancer inflammation and outcomes. Cancer Epidemiol. Biomark. Prev. 26, 328–338 (2017).
5. Kelsey, K. T. & Wiencke, J. K. Immunomethylomics: a novel cancer risk prediction tool. Ann. Am. Thorac. Soc. 15, S76–S80 (2018).
6. Wiencke, J. K. et al. Immunomethylomic approach to explore the blood neutrophil lymphocyte ratio (NLR) in glioma survival. Clin. Epigenetics 9, 10 (2017).
7. Teschendorff, A. E. & Zheng, S. C. Cell-type deconvolution in epigenomewide association studies: a review and recommendations. Epigenomics 9, 757–768 (2017).
8. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics 18, 105 (2017).
9. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13, 86 (2012).
10. Zheng, S. C., Breeze, C. E., Beck, S. & Teschendorff, A. E. Identification of differentially methylated cell types in epigenome-wide association studies. Nat. Methods 15, 1059–1066 (2018).
11. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 15, R31 (2014).
12. Koestler, D. C. et al. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinformatics 17, 120 (2016).
13. Kim, S. et al. Enlarged leukocyte referent libraries can explain additional variance in blood-based epigenome-wide association studies. Epigenomics 8, 1185–1192 (2016).
14. Waite, L. L. et al. Estimation of cell-type composition including T and B cell subtypes for whole blood methylation microarray data. Front. Genet. 7, 23 (2016).
15. Horvath, S. DNA methylation age of human tissues and cell types. Genome Biol. 14, R115 (2013).
16. Heiss, J. A. et al. Training a model for estimating leukocyte composition using whole-blood DNA methylation and cell counts as reference. Epigenomics 9, 13–20 (2017).
17. Reinius, L. E. et al. Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. PLoS ONE 7, e41361 (2012).
18. Galanter, J. M. et al. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. eLife 6, 1–24 (2017).
19. Salas, L. A. et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 19, 64 (2018).
20. Bakulski, K. M. et al. DNA methylation of cord blood cell types: applications for mixed cell birth studies. Epigenetics 11, 354–362 (2016).
21. Gervin, K. et al. Cell type specific DNA methylation in cord blood: a 450Kreference data set and cell count-based validation of estimated cell type composition. Epigenetics 11, 690–698 (2016).
22. de Goede, O. M., Lavoie, P. M. & Robinson, W. P. Characterizing the hypomethylated DNA methylation profile of nucleated red blood cells from cord blood. Epigenomics 8, 1481–1494 (2016).
23. Lin, X. et al. Cell type-specific DNA methylation in neonatal cord tissue and cord blood: a 850K-reference panel and comparison of cell types. Epigenetics 13, 941–958 (2018).
24. Gervin, K. et al. Systematic evaluation and validation of reference and library selection methods for deconvolution of cord blood DNA methylation data. Clin. Epigenetics 11, 125 (2019).
25. Farlik, M. et al. DNA methylation dynamics of human hematopoietic stem cell differentiation. Cell Stem Cell 19, 808–822 (2016).
26. Zheng, Y. et al. A human circulating immune cell landscape in aging and COVID-19. Protein Cell 11, 740–770 (2020).

27. Simon, A. K., Hollander, G. A. & McMichael, A. Evolution of the immune system in humans from infancy to old age. Proc. R. Soc. B Biol. Sci. 282, 20143085 (2015).
28. Zhao, M. et al. Distinct epigenomes in CD4+ T cells of newborns, middle-ages and centenarians. Sci. Rep. 6, 38411 (2016).
29. Heyn, H. et al. Distinct DNA methylomes of newborns and centenarians. Proc. Natl Acad. Sci. USA 109, 10522–10527 (2012).
30. Yousefi, P. et al. Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies. Environ. Mol. Mutagenesis 56, 751–758 (2015).
31. Maden, S. K., Thompson, R. F., Hansen, K. D. & Nellore, A. Human methylome variation across Infinium 450K data on the Gene Expression Omnibus. NAR Genomics Bioinformatics 3, lqab025 (2021).
32. Staats, J., Divekar, A., McCoy, J. P. & Maecker, H. T. Guidelines for gating flow cytometry data for immunological assays. Methods Mol. Biol. 2032, 81–104 (2019).
33. Abdelsamed, H. A. et al. Human memory CD8 T cell effector potential is epigenetically preserved during in vivo homeostasis. J. Exp. Med. 214, 1593–1606 (2017).
34. Monaco, G. et al. RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. Cell Rep. 26, 1627–1640.e7 (2019).
35. Ackermann, J. A., Radtke, D., Maurberger, A., Winkler, T. H. & Nitschke, L. Grb2 regulates B-cell maturation, B-cell memory responses and inhibits B-cell Ca2+ signalling. EMBO J. 30, 1621–1633 (2011).
36. Engels, N. et al. The immunoglobulin tail tyrosine motif upgrades memorytype BCRs by incorporating a Grb2-Btk signalling module. Nat. Commun. 5, 5456 (2014).
37. Jiang, W. et al. Prolyl 4-hydroxylase 2 promotes B-cell lymphoma progression via hydroxylation of Carabin. Blood 131, 1325–1336 (2018).
38. Gennery, A. R. et al. Mutations in CHD7 in patients with CHARGE syndrome cause T-B + natural killer cell + severe combined immune deficiency and may cause Omenn-like syndrome. Clin. Exp. Immunol. 153, 75–80 (2008).
39. Fukada, K., Sobao, Y., Tomiyama, H., Oka, S. & Takiguchi, M. Functional expression of the chemokine receptor CCR5 on virus epitope-specific memory and effector CD8 + T cells. J. Immunol. 168, 2225–2232 (2002).
40. Hardie, D. L. et al. The stromal cell antigen CD248 (endosialin) is expressed on naive CD8+ human T cells and regulates proliferation. Immunology 133, 288–295 (2011).
41. Mitra, S. N., Slungaard, A. & Hazen, S. L. Role of eosinophil peroxidase in the origins of protein oxidation in asthma. Redox Rep. 5, 215–224 (2000).
42. Westphal, A. et al. Lysosomal trafficking regulator Lyst links membrane trafficking to toll-like receptor-mediated inflammatory responses. J. Exp. Med. 214, 227–244 (2017).
43. Goodridge, J. P., Witt, C. S., Christiansen, F. T. & Warren, H. S. KIR2DL4 (CD158d) genotype influences expression and function in NK cells. J. Immunol. 171, 1768–1774 (2003).
44. Pedros, C. et al. Requirement of Treg-intrinsic CTLA4/PKCη signaling pathway for suppressing tumor immunity. JCI Insight 2, e95692 (2017).
45. Uhlen, M. et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. Science 366, eaax9198 (2019).
46. Grootens, J., Ungerstedt, J. S., Nilsson, G. & Dahlin, J. S. Deciphering the differentiation trajectory from hematopoietic stem cells to mast cells. Blood Adv. 2, 2273–2281 (2018).
47. Dahlin, J. S. et al. A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. Blood 131, e1–e11 (2018).
48. van Langelaar, J., Rijvers, L., Smolders, J. & van Luijn, M. M. B and T cells driving multiple sclerosis: identity, mechanisms and potential triggers. Front. Immunol. 11, 760 (2020).
49. Verma, R. et al. Lymphocyte depletion and repopulation after chemotherapy for primary breast cancer. Breast Cancer Res. 18, 10 (2016).
50. Zhang, B. et al. The dynamics of immune response in COVID-19 patients with different illness severity. J. Med. Virol. 93, 1070–1077 (2021).
51. Liu, Y. et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat. Biotechnol. 31, 142–147 (2013).
52. Avdeeva, A., Rubtsov, Y., Dyikanov, D., Popkova, T. & Nasonov, E. Regulatory T cells in patients with early untreated rheumatoid arthritis: phenotypic changes in the course of methotrexate treatment. Biochimie 174, 9–17 (2020).
53. Aryee, M. J. et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30, 1363–1369 (2014).
54. Xu, Z., Niu, L., Li, L. & Taylor, J. A. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. Nucleic Acids Res. 44, e20 (2016).
55. Zhou, W., Triche, T. J., Laird, P. W. & Shen, H. SeSAMe: Reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res. 46, e123 (2018).
56. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 45, e22 (2017).
57. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics 32, 286–288 (2016).
58. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740 (2011).
59. Breeze, C. E. et al. eFORGE: A tool for identifying cell type-specific signal in epigenomic data. Cell Rep. 17, 2137–2150 (2016).
60. Salas, L. A. & Zhang, Z. Zenodo: v1.1 immunomethylomics/analysis_FlowSorted.BloodExtended.EPIC. https://doi.org/10.5281/zenodo.5338513 (2021).

Acknowledgements

L.A.S. was supported by CDMRP/Department of Defense (W81XWH-20-1-0778) and NIGMS (P20 GM104416-09/8299). D.C.K. was supported by NIH grants P30 CA168524, P20 GM130423, and P20 GM103428. J.K.W. and A.M.M. were supported by NIH grants R01 CA207360 and P50 CA097257. K.T.K. was supported by a 2018 AACR-Johnson & Johnson Lung Cancer Innovation Science (18-90-52-MICH) and NIH grant
R01CA253976. B.C.C. was supported by NIH grants R01 CA216265 R01 CA253976, P20 GM130454, and P30 DK117469. Research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number
P20 GM104416. We acknowledge the following Shared Resources facilities at the Norris Cotton Cancer Center at Dartmouth with NCI Cancer Center Support Grant P30 CA023108: Immune Monitoring & Flow Cytometry.

Author contributions

L.A.S., D.C.K., J.K.W., K.T.K., and B.C.C. conceived the project and designed the experiments. L.A.S. and Z.Z. performed the bioinformatic quality control and analyzed the data. H.M.H. and R.A.B. carried out the wet-lab work, including the spiking of the artificial mixtures used in this experiment (R.A.B.). D.C.K. and A.M.M. supported the statistical analysis. L.A.S. wrote the manuscript with input from all the coauthors. The final version of the manuscript has been reviewed and approved by all the authors.

Competing interests

J.K.W. and K.T.K. are founders of Cellintec, which had no role in this research. A U.S. Provisional Patent Application entitled: ENHANCED DNA METHYLATION LIBRARY
FOR DECONVOLUTING PERIPHERAL BLOOD was filed on February 12, 2021 by Dartmouth College for the IDOL deconvolution libraries invention included in this manuscript. Inventors: Christensen, BC, Salas LA, Kelsey KT, Wiencke JW, and Koestler DC. Application number: 63/148,695. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-27864-7.

Correspondence and requests for materials should be addressed to Karl T. Kelsey or Brock C. Christensen.

Peer review information Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. This article has been peer reviewed as part of Springer Nature's Guided Open Access initiative.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

nature communications

Article

https://doi.org/10.1038/s41467-022-33215-x

DNA methylation as a pharmacodynamic marker of glucocorticoid response and glioma survival

Received: 17 August 2021

Accepted: 8 September 2022

Published online: 20 September 2022

Check for updates

J. K. Wiencke [1], Annette M. Molinaro [1], Gayathri Warrier [1], Terri Rice[1], Jennifer Clarke [1,2] ✉, Jennie W. Taylor [1,2], Margaret Wrensch[1], Helen Hansen [1], Lucie McCoy [3], Emily Tang [4], Stan J. Tamaki [4], Courtney M. Tamaki [4], Emily Nissen [5], Paige Bracci[3], Lucas A. Salas[6], Devin C. Koestler[5], Brock C. Christensen[6,7,8], Ze Zhang [6] & Karl T. Kelsey [9,10]

Assessing individual responses to glucocorticoid drug therapies that compromise immune status and affect survival outcomes in neuro-oncology is a great challenge. Here we introduce a blood-based neutrophil dexamethasone methylation index (NDMI) that provides a measure of the epigenetic response of subjects to dexamethasone. This marker outperforms conventional approaches based on leukocyte composition as a marker of glucocorticoid response.TheNDMIisassociatedwithlowCD4Tcellsandtheaccumulationof monocytic myeloid-derived suppressor cells and also serves as prognostic factor in glioma survival. In a non-glioma population, the NDMI increases with a history of prednisone use. Therefore, it may also be informative in other conditions where glucocorticoids are employed. We conclude that DNA methylation remodeling within the peripheral immune compartment is a rich source of clinically relevant markers of glucocorticoid response.

Cortisol and synthetic glucocorticoid analogs activate intracellular glucocorticoid receptors (GR) and modulate many physiologic processes. Processes affected include carbohydrate and lipid metabolism, stress, and immune responses[1]. Because of their potent effects on immune cells, glucocorticoids have long been the most widely prescribed class of anti-inflammatory and immunosuppressive drugs[2]. However, glucocorticoid-treated patients may exhibit resistance or experience a wide range of adverse side effects. Our understanding of these risks is incomplete[3,4]. To account for individual variability, biomarkers have been developed that integrate variations in glucocorticoid exposure and response[5]. Several well-established safety concerns

of glucocorticoids have been addressed with markers bridged to clinical outcomes, namely: adrenal suppression, insulinresistance, and accelerated bone turnover[6–9]. In the management of brain tumor patients, the synthetic glucocorticoid dexamethasone (DEX) is a mainstay for control of peritumoral edema, helping to alleviate symptoms due to high intracranial pressure or presence of tumor/ edema in eloquent areas of brain[10]. No randomized controlled trials have specifically addressed DEX. However, its' use in glioma patients has been associated with poor survival[11–13] and compromised response to immunotherapy[14–16], There are no established markers to assess these risks, although altered blood leukocyte composition has been

used to mark glucocorticoid-related immunosuppression[17–19]. Depletion of circulating lymphocyte numbers[20–23] and expansion of neutrophils and immunosuppressive myeloid cell populations[24,25], including monocytic myeloid-derived suppressor cells (mMDSC)[26,27], have been documented in DEX exposed brain tumor patients. Although blood leukocyte compositions are sensitive to glucocorticoid exposures, they lack specificity, as tumor inflammation and chemoradiation also directly affect them[28]. In our search to uncover more specific markers of glucocorticoid response, we turned to recent epigeneticstudiesshowingthatcell-specificchromatinmodificationsplay a pivotal role in determining the genomic effects of glucocorticoids[29,30]. Cell-specific actions of glucocorticoids have been linked to highly accessible chromatin domains located preferentially within distal enhancer elements that are intrinsic to cell lineage[31]. Glucocorticoid alterations of DNA methylationmaybe integralto their genomic effects[32,33]. Because DNA methylation is a chemically stable endpoint and easy to access in peripheral blood, it is an attractive

[1] Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA. [2] Department of Neurology, University of California San Francisco, San Francisco, CA,USA. [3]Department of Epidemiology and Biostatistics, University ofCaliforniaSan Francisco, San Francisco,CA, USA.[4]Parnassus Flow Cytometry CoLab, University of California San Francisco, San Francisco, CA, USA. [5]Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, KS, USA. [6]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA. [7]Department of Molecular and Systems Biology, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA. [8]Department of Community and Family Medicine, Geisel School of Medicine, DartmouthCollege, Lebanon, NH, USA. [9]Department of Epidemiology,Brown University,Providence, RI, USA.[10]Department of Pathology and Laboratory Medicine, Brown University, Providence, RI, USA. ✉e-mail: John.Wiencke@ucsf.edu

potential marker for glucocorticoid pathway activation. Previous studies of DNA methylation and glucocorticoids focused on a limited number of genes and cytosine guanine dinucleotide (CpG) loci in the context of neurodevelopmental and stress-related outcomes[34,35]. Furthermore, prior research did not take advantage of the cell-selective natureofglucocorticoid-mediatedgenomiceffects.Comparativegene expression studies showed that, among different hematopoietic cell types, neutrophils exhibited the most pronounced transcriptional response to glucocorticoid treatments[36]. In severe COVID-19 subjects DEX induced profound transcriptional remodeling and induction of immunosuppressive neutrophil states[37].

In this work, we explore DNA methylation as pharmacodynamic markersofDEXexposureinadultgliomapatientswhohadandhadnot been given the drug as part of their routine care. We employ a bioinformatic approach to discover and assess DEX-related neutrophilspecific DNA methylation as a marker of in vivo glucocorticoid exposure and response, the neutrophil dexamethasone methylation index (NDMI). Our results indicate that blood DNA methylation can accurately discriminate DEX exposed and non-exposed subjects and that a methylation marker targeting blood neutrophils performs better than a marker based on altered blood leukocyte composition in characterizing individual glucocorticoid response. We also explore the association of NDMI scores with blood counts of CD4 T cells and mMDSC and show that neutrophil-specific methylation markers of DEX exposure serve as prognostic factors in brain tumor survival.

Results

Neutrophil-specific and non-cell-specific DNA methylation induced by DEX

Clinical and demographic characteristics of our study groups and study design are presented (Fig. 1a; Supplementary Table 1). After modeling all cell types using CellDMC, we identified 2621 CpG probes that displayed statistically significant ($p < 0.01$) interactions with neutrophil proportions and one CpG that indicated significant interaction ($p < 0.01$) with B-cell proportions (Figs. 1b, 2a, Supplementary Data File 1). The predominant effect of DEX was hypomethylation. Another subset of 17,733 loci was affected by DEX but not in a cell-specific fashion (i.e., non-cell-specific methylation).

Genomic and functional features of neutrophilic DEX responsive loci

To test whether neutrophil-specific CpGs follow pre-established and lineage-specific DNA methylation patterns, we explored the methylation levels of DEX-responsive loci in isolated non-exposed neutrophils and other immune cell types. DEX-related neutrophil loci displayed intermediate levels of methylation. These loci were partially demethylated in unexposed neutrophils and other myeloid cells (basophils, eosinophils, monocytes), whereas, in lymphocytic cell types (CD4 T, CD8 T, NK cell), they were hypermethylated (Fig. 2b). Interestingly, some B-cell loci were partially demethylated, which appeared related to their naïve and memory differentiation status. The average methylation beta values of DEX loci in neutrophils were markedly lower than lymphoid cells. Similarly, other myeloid cells (eosinophils, basophils, and monocytes) demonstrated lower beta values than lymphocytes (Fig. 2c).

Mapping the genomic locations of neutrophil-specific loci revealed overlap atGR binding sites and occurred preferentially within introns, enhancers, and in a non-CpG island context (Fig. 3a). Gene ontology (GO) analyses indicate enrichment among 100 genes for neutrophil activation and granule biology (Fig. 3b). Comparing these enriched GO-associated genes with published 11 RNA co-expression modules (EM1-11) in human bone marrow myelopoiesis[38] revealed a close correspondence of 21 differentially methylated genes with one co-expression module (designated EM6 Fig. 3c) that was maximally transcribed at the metamyelocyte and band stages (Fig. 3d).

Machine learning approach to DEX predictors using leukocyte composition and DNA methylation

DEX associated alterations in blood leukocyte composition in the training dataset followed expectations based on historical experience with glucocorticoids: increases in neutrophil proportions and decreases in lymphocytes were observed in DEX exposed subjects (Supplementary Table 3). Although absolute monocyte counts were marginally increased, monocyte proportions were not significantly higher inDEX exposed subjects.Elasticnet regularized modelswerefit using all six cell proportions and the ratiosofneutrophil/lymphocytes, CD4/CD8 T cells, and lymphocytes/monocytes for DEX exposed and non-exposed subjects. The resulting blood leukocyte composition index contains five parameters (neutrophil, monocyte, NK and CD4 T-cell proportions, and neutrophil/lymphocyte ratio), along with their regression coefficients. Elastic net models were also constructed using the 2621 neutrophil-specific and 17,733 non-cell-specific methylation loci. The elastic net model fit to the 2621 neutrophil-specific DEX associated loci resulted in 28 CpGs with non-zero coefficients. A classifier, termed the neutrophil dexamethasone methylation index (NDMI), was created based on these 28 CpGs and is defined as follows:

$$\text{Probability of being a Dex User} = \frac{1}{1 + e^{-\text{NDMI}}}$$

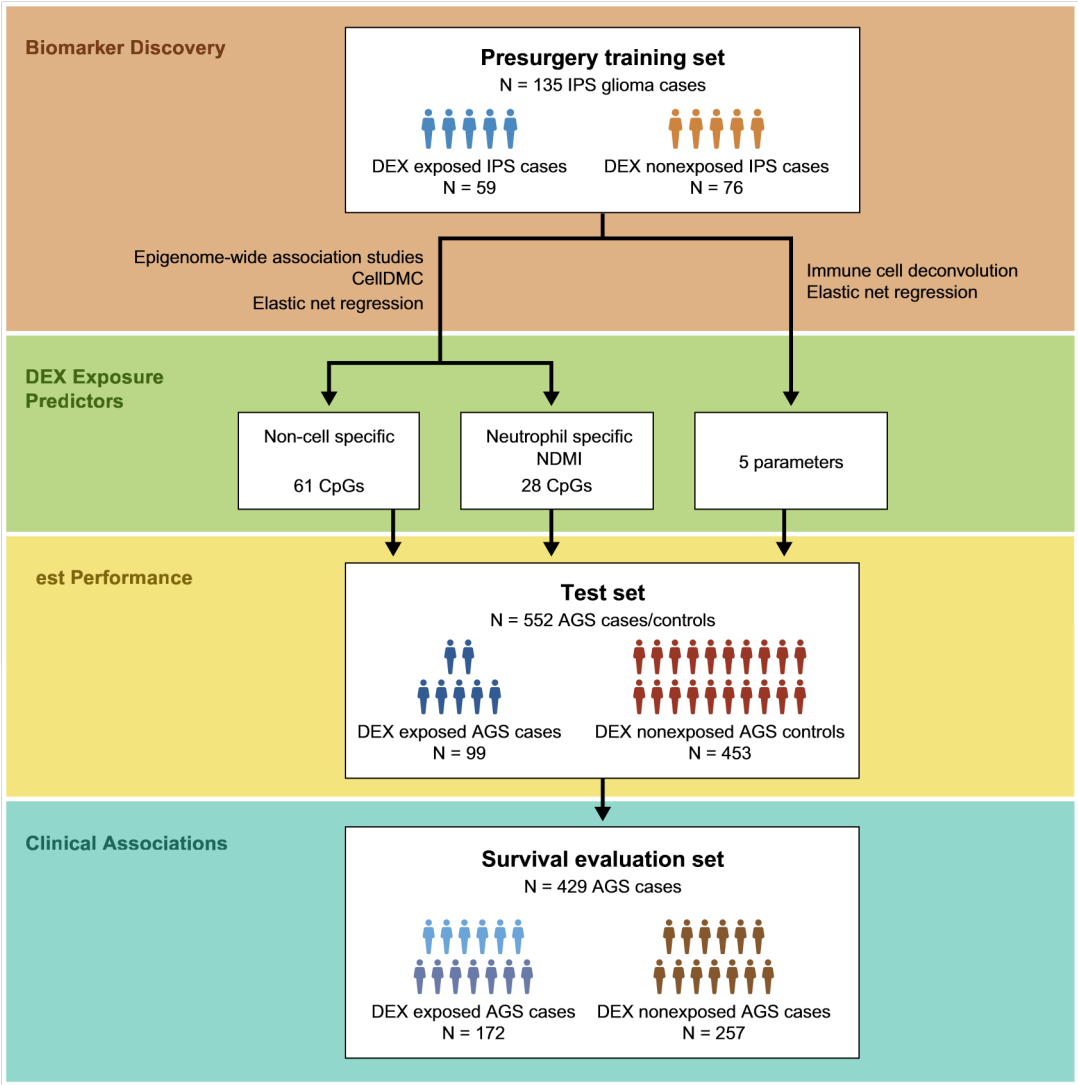
$$\text{NDMI} = \beta_0 + \beta_1 * \text{cg00052684} + \beta_2 * \text{cg01994208} + \ldots + \beta_{28} * \text{cg27094376}$$

The NDMI contains well-known glucocorticoid targets of demethylationsuchasFKBP5andZBTB16andotherswithlesswell-described rolesinglucocorticoidresponse(SupplementaryTable4).Enteringthe 28 sites into the eFORGE program revealed a significant overrepresentation of the enhancer chromatin mark H3K1me3 in primary human monocytes but not in T or B cells or other hematopoietic cells; neutrophils are not represented in the eFORGE reference database (Supplementary Fig. 1). NDMI loci were often sites (16/28) for transcription factor binding, including Ikaros (IKZF1, IKZF2), MYC, and MAX, which can form multi-protein repressive complexes (Supplementary Table 7, Supplementary Data File 2). In contrast to the larger set of neutrophil genes identified in GO analysis, the 16 genes represented in the NDMI do not overlap with the EM6 bone marrow coexpression module and instead display maximal transcription during the earliest and late stages of neutrophil maturation (Fig. 3e). The predictor based on 61 non-cell-specific CpGs is termed the non-cellspecific predictor. The components of each predictor are summarized (Supplementary Table 5).

Sources of variation in NDMI scores include age, a history of prednisone use and baseline NDMI status

We estimated potential technical error in NDMI measurements using replicate methylation measures from 17 subjects (coefficient of variation <18% see Supplementary Table 7). We found age to be positively associated with NDMI scores of the demographic factors. The current use of glucocorticoid medications was associated with higher NDMI scores (Supplementary Table 8, Supplementary Fig. 2). Oral prednisone users had higher NDMI scores than non-users ($P = 0.003$) or subjects using inhaled glucocorticoids ($P = 0.03$). NDMI scores of inhaled glucocorticoid users were not statistically significantly different from non-glucocorticoid users. Medical conditions recorded among the 18 non-glioma controls receiving oral prednisone included rheumatoid arthritis, solid organ transplants, kidney disease, lupus,

a


Biomarker Discovery
Presurgery training set
N = 135 IPS glioma cases
DEX exposed IPS cases
N = 59
DEX nonexposed IPS cases
N = 76
Epigenome-wide association studies
CellDMC
Elastic net regression
Immune cell deconvolution
Elastic net regression
DEX Exposure Predictors
Non-cell specific
61 CpGs
Neutrophil specific
NDMI
28 CpGs
5 parameters
est Performance
Test set
N = 552 AGS cases/controls
DEX exposed AGS cases
N = 99
DEX nonexposed AGS controls
N = 453
Clinical Associations
Survival evaluation set
N = 429 AGS cases
DEX exposed AGS cases
N = 172
DEX nonexposed AGS cases
N = 257

b

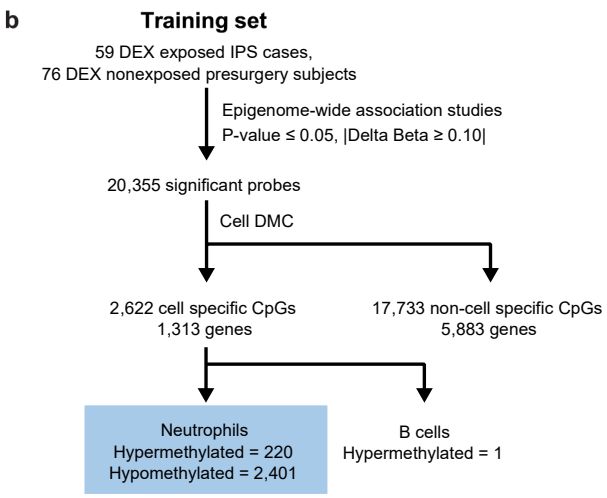

Training set
59 DEX exposed IPS cases,
76 DEX nonexposed presurgery subjects
Epigenome-wide association studies
P-value ≤ 0.05, |Delta Beta ≥ 0.10|
20,355 significant probes
Cell DMC
2,622 cell specific CpGs
1,313 genes
17,733 non-cell specific CpGs
5,883 genes
Neutrophils
Hypermethylated = 220
Hypomethylated = 2,401
B cells
Hypermethylated = 1

Fig. 1 | Study design and results. a There were four phases to the design. First, in Biomarker Discovery two pipelines (EWAS/CellDMC (see b) and immune cell deconvolution) were applied to a cohort of 135 immune profile study (IPS) glioma cases of which 59 were DEX exposed and 76 were not. Second, three different elastic net regression models (DEX Exposure Predictors) were built using different CpGs(Non-cell-specific(1stmodel)andNeutrophilspecific(NDMI:2ndmodel))and 5 immune parameters (3rd model) on the same 135 IPS glioma cases. Third, the three models were evaluated by an independent cohort of 552 Adult Glioma Study methylationbeta values of 2621DEX (AGS) cases and controls. Fourth, the NDMI model was evaluated as a predictor for survival in 429 AGS glioma cases of which 172 were exposed to DEX and 257 were not. b Overview of CpGs selected from EWAS/CellDMC pipeline in Biomarker Discovery (1st phase of a). Nominally significant loci ($p < 0.05$) were identified and entered into the CellDMC program resulting in loci with significant interaction terms (FDR < 0.05). Two subsets of loci were identified non-cell specific and Neutrophil specific (1st and 2nd models in a). Source data are provided as a Source Data file.

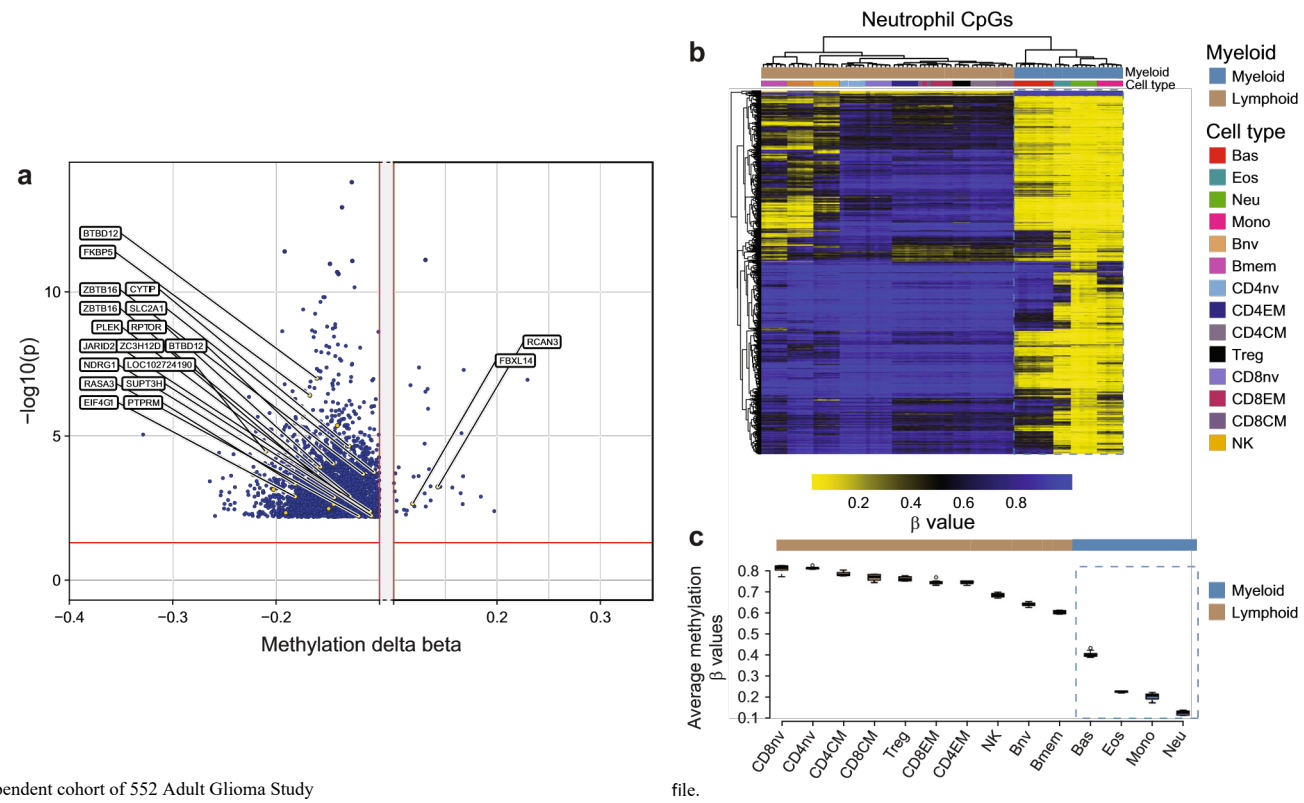
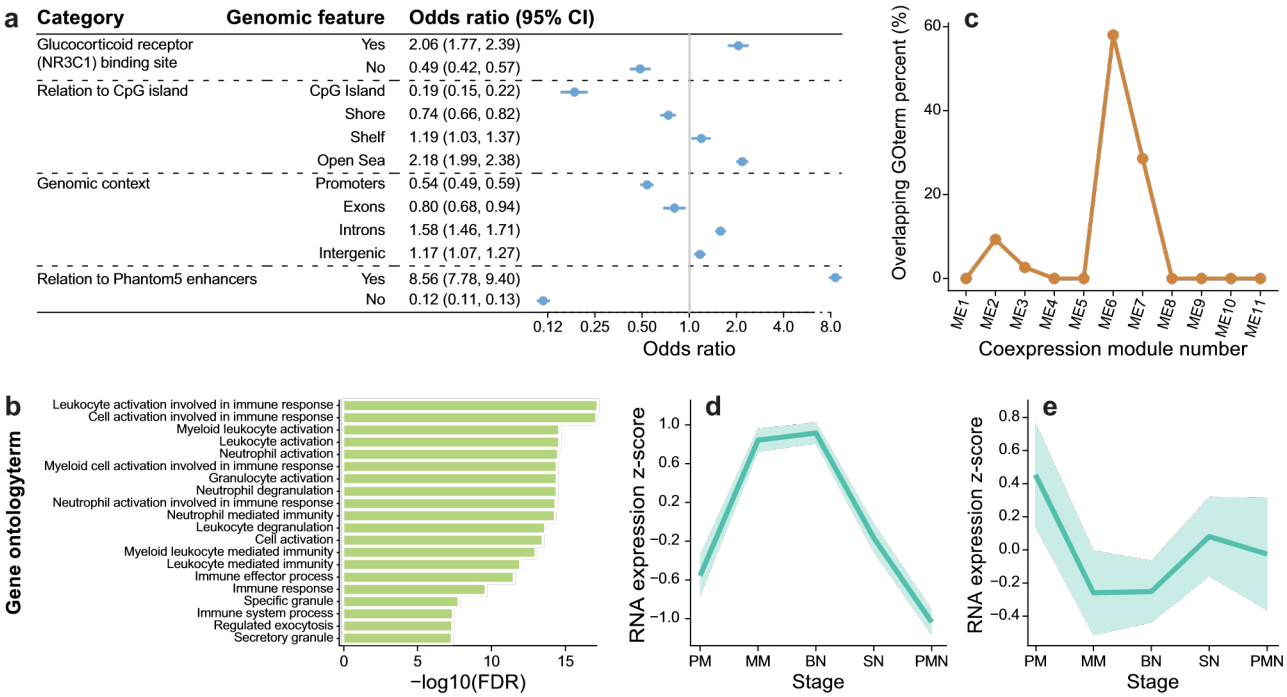


Fig. 2 | DEX-related neutrophil-specific DNA methylation. a Volcano plot showing the output of CellDMC analysis of 2621 DEX associated CpG loci with nominally significant (FDR < 0.05) interactions with neutrophil proportions. Plotted are the average difference in methylation beta value (delta beta) between DEX exposed and DEX non-exposed subjects (x-axis) and the p value for the interaction of methylation and neutrophilproportion (y-axis).Hypomethylated locihave lower methylation beta values in exposed versus non-exposed (left side of volcano plot). The plot is truncated at |delta beta|> 0.1. CpGs associated with genes included in the NDMI scoreare highlighted.b Heatmapof associated neutrophil probes (output of CellDMC) across 14 different isolated nonexposed immune cell types. Median beta values and Tukey's interfence (interquartile range;Q3–Q1) are depicted;6 individuals weresampled for CD4nv, CD8nv, CD4CM, CD8CM, CD4EM, CD8EM, NK, Bnv, Bmem, mono, neutrophils, and 9 basophil, 4 Treg, 4 Eos, 3 monocytes. Myeloid cells are demarcated with a dashed line. c The median beta values and Tukey's interfence range (interquartile range) are depicted for each of 14 immune cell types contained in b; note cellsare ordered from highest to lowest beta value. Source data are provided as a Source Data file.

scleroderma, multiple sclerosis, uveitis, hepatitis, and connective tissuedisorder.Thechangesinimmunecellproportionincontrolstaking prednisone were similar to those observed among glioma patients treated with DEX, Supplementary Table 9. Among the DEX exposed glioma subjects in the training data, cumulative and average daily mg DEX exposure was associated with variations in the NDMI; dose–response plots suggested a non-linear and saturating relationship between NDMI score and DEX dose (Fig. 4a, b). Several DEX users had lower scores overlapping non-exposed cases. These outliers included 7 DEX users with low cumulative doses and recent (within 30 days) but not current DEX exposure, suggesting a time-dependent changeinNDMIfollowingcessationofDEXtreatment.Inalongitudinal analysis of subjects who completed DEX treatments before the blood draw, we observed four subjects with no exposure for seven or more days before the blood draw had NDMI scores identical to their pretreatment scores (Supplementary Fig. 3). In another longitudinal cohort of 31 DEX naïve subjects subsequently treated with DEX following surgery, the NDMI scores increased during DEX exposure (i.e., at blood draw) and were inversely correlated with each subject's baseline NDMI scores, P =0.008 (Supplementary Fig. 4). Because low neutrophil proportions may affect NDMI performance, we simulated the effects of different neutrophil proportions on the probability of detecting DEX exposures (Supplementary Fig. 5). Variation of immune proportions among samples and the sample size of exposed and nonexposed subjects were found to be essential factors affecting NDMI performance. Critical values of less than ~20–30% neutrophils could influence the performance of the NDMI as an indicator of DEX exposure. Among 457 glioma subjects in our study, only 2.3% had <40% neutrophil proportions.



DEX treatments modify the regression of NDMI scores on neutrophil proportions

A fundamental premise of the CellDMC program is that differential methylation occurring within neutrophils will alter the regression of methylation on neutrophil proportions in DEX exposed subjects. We estimatedregressionparametersforNDMIandneutrophilproportions

Fig. 3 | Functional features and gene ontology analyses of genes containing neutrophil-specific DEX loci. a Genomic context enrichment of neutrophilspecific CpGs (n = 2621) with all tested CpGs (n = 830,277) as background.

b Enriched GO terms associated with 1360 genes displaying differentially methylated DEX loci and interaction with neutrophil proportions. c Percent of GO terms in the myeloid expression (ME) modules overlapping with the GO terms enriched for 1360 neutrophil-specific loci associated genes. ME6 displayed the highest overlapping percent (58%) with 29 out of 50 overlapping with neutrophil-specific GO terms. d Plot of maximal expression of 21 overlapping differentially methylated neutrophil genes during myelopoiesis; these genes include CASP10, ALDDH3B1, GAS7, NLRC4,APMAP, OLFM4, EHD4, S100A8, ACAD8, ANKRD22, CNKSR3, FCER1G, UGP2, CLEC12A, OLR1, LRRC57, MAFG, PAPSS2, CD177, LTBR4, LIMS1. e Plot of maximal expression of 16 genes contained in the NDMI algorithm (see Supplementary Table 4) showing maximal expression in promyelocyte and PMN stages of myelopoiesis.PMpromyelocyte, MM metamyelocyte,BNbandneutrophil, SNsegmented neutrophil, PMN polymorphonuclear neutrophil. Source data are provided as a Source Data file.

in training (N =135) and an extended case/control dataset (N =602). In both, the ANOVA linear regression interaction terms of NDMI and DEX exposure were highly statistically significant (P <0.0001). DEX treatments increased the slope gradient of NDMI regressed on neutrophil proportions among DEX exposed subjects (Fig. 4c, d).

NDMI is a highly accurate predictor of DEX exposure in an independent dataset

Weassessedthequalityofourpredictivemodelsinanindependentset of DEX exposed cases and non-glioma controls (Fig. 5a). Receiver operating characteristic (ROC) analyses in the test population indicated high sensitivity and specificity of the NDMI score as a binary predictor of current DEX exposure (Fig. 5b). ROC performance was highest for the NDMI score (0.91; 95% CI 0.871, 0.95) although leukocyte composition performed well (AUC 0.71; 95% CI 0.649, 0.77). The non-cell-specific methylation predictor appeared superior to the leukocyte composition model (AUC 0.84; 95% CI 0.793, 0.896). We also built a version of the NDMI using only presurgery glioblastoma (GBM) (N = 71) subjects. The performance of the GBM only NDMI (AUC 0.88; 95% CI 0.83, 0.92) was greater than the leukocyte composition model but less than the NDMI created using all grades of glioma.

NDMI scores associated with depressed CD4 T cell counts and increased mMDSC levels

Depressed CD4 T cell counts signal immunosuppression; among the 135 training set subjects, 47 (34.8%) had CD4 counts <500cells/μl and 16 (11.8%) had counts <200 cells/μl (Table 1). Current DEX use was recorded in 39 (83%) and 15 (94%) of these subjects, respectively. NDMI, total white cell count, and their interaction were significantly associated with CD4 T-cell levels in whole blood and explained substantial variation in CD4 counts among glioma subjects (r^2 =0.64; p <0.001) (Table 2). CD4 counts were positively associated with total white cells, whereas NDMI scores were inversely associated. In these models, DEX exposure variables were not statistically significant, although daily DEX dose was inversely associated with CD4 counts in univariate tests (r^2 = 0.21; p =0.004). A subset of subjects' bloods was run in parallel using flow cytometry (FCM) and methylation deconvolution. Bland-Altman's analysis of these paired data showed DNA methylation deconvolution to be highly accurate in estimating CD4 T cell concentrations (Supplementary Fig 6).

We estimated blood mMDSC levels both as proportions of monocytes and as concentrations in 38 consecutive glioma volunteers using FCM (Fig. 6). Subjects taking DEX had greater proportions (p = 0.002) and concentrations (p =0.005) of $CD14^+HLA\text{-}DR^{neg/low}$ mMDSCs compared to non-exposed subjects. We further subdivided themMDSCpopulationbyCD16expressionandfoundsignificantDEXrelated increases in $CD14^+HLA\text{-}DR^{neg/low}CD16^-$ cells/μl (p= 0.005). NDMI scores were positively correlated with total mMDSCs proportions (r^2 = 0.30; p =0.003) and concentrations (r^2 = 0.33; p = 0.001) as well as $CD16^-$ mMDSC populations (proportions r^2 =0.31; p = 0.002; concentrations r^2 =0.34; p < 0.0001). The most robust model predicting mMDSC concentrations included monocyte counts. NDMI and monocyte counts together were positively associated with mMDSC levels in whole blood and together explained 76% of the variation in mMDSC levels and 77% of the variation of $CD16^-$ mMDSCs (p <0.001). Average daily DEX exposure, although univariately associated with mMDSCs (r^2 = 0.24; p =0.04), was not significant in models containing monocytes and NDMI. Neither cumulative mg of DEX nor duration of use were significantly associated with variations in mMDSC levels (P > 0.05).

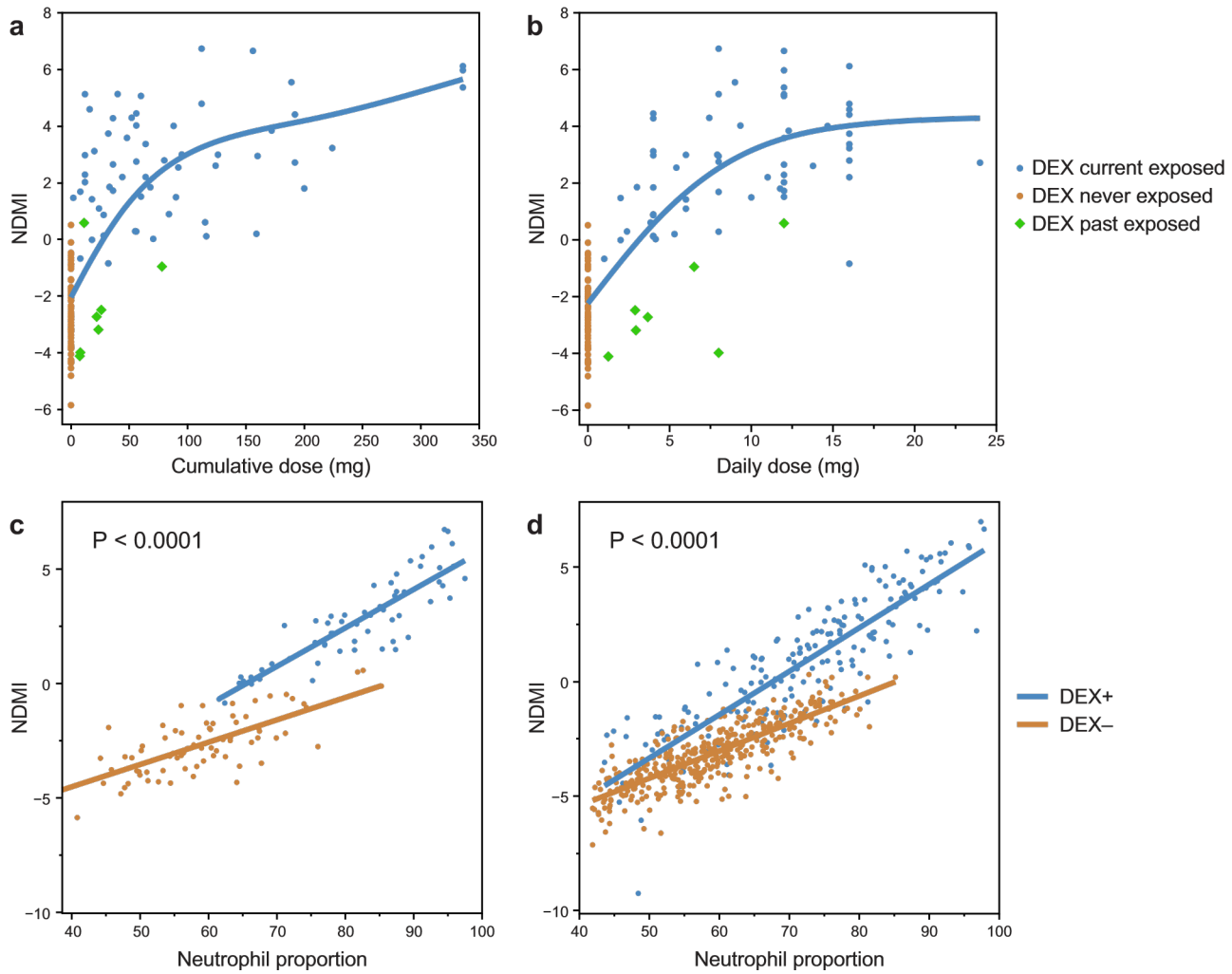


Fig. 4 | Effects of DEX on NDMI response and modification of NDMI regression on neutrophil proportions. a Dose–response of cumulative mg of DEX versus blood NDMI scores. b Dose–response of average daily mg DEX versus blood NDMI scores. Smooth curves are shown. In a, b, the 7 past exposed subjects received DEX 4, 7, 8, 10, 17, 18, and 22 days prior to blood draw. c Scatterplot of NDMI (y-axis) by the predicted relative fraction of neutrophils (x-axis) between 135 DEX+/DEX− subjects (different colored points) in the training dataset. The lines in the plot reflect the fitted regression line for DEX+ and DEX− subjects. The p value resulted from a test of equality of slopes between the DEX+ and DEX− regression lines. d Scatterplot of NDMI (y-axis) by the predicted relative fraction of neutrophils (xaxis) between DEX+/DEX− subjects (different colored points) in an independent dataset of 99 exposed glioma subjects and 453 control subjects. The lines in the plot reflect the fitted regression line for DEX+ and DEX− subjects. The p value resulted from a test of equality of slopes between the DEX+ and DEX− regression lines. Source data are provided as a Source Data file.

Elevated NDMI scores in isolated DEX exposed monocytes, mMDSCs and neutrophils

To provide a reference for comparing NDMI scores in different immune subsets we compared DEX exposed cells and 12 different DEX naïvecellsubtypes.InDEXnaïvecellsweobservedsignificantlygreater NDMI scores among myeloid-derived cells compared to lymphoid cells; Fig. 6d. In vivo DEX exposed monocytes and mMDSCs contained elevated NDMI scores compared to non-exposed monocytes (p = 0.001) and compared to paired DEX exposed monocytes from the same individuals (p =0.03). Monocytes from DEX exposed subjects had higher NDMI scores compared to non-exposed monocytes (p = 0.002) (Fig. 6e). To test whether isolated neutrophils exposed in vivo to DEX contain DNA methylation changes consistent with the CellDMCcomputationalmethod,weisolatedcellsfrom5DEXexposed glioma patients. The purity of these neutrophil isolates was estimated to be 99.9% (range 99.3–100%) using an extensive methylation deconvolution method[39]. DEX exposed isolated neutrophils exhibited NDMIscoresmarkedlyhigherthanunexposedcellsandhighestamong all isolated cells examined (Fig. 6e). We compared the NDMI scores of isolated neutrophils exposed to DEX to the respective whole-blood NDMI scores from the same individuals; in each of the subjects the neutrophil NDMI score was significantly greater than the whole-blood score (paired t test p =0.01; Supplementary Table 10). Further, of the 2621 neutrophil-specific DEX associated CpG sites identified in the CellDMC analysis, 1903 (72.6%) exhibited the same direction of association in isolated cells and 823 probes were nominally statistically significant (p <0.05) between exposed and non-exposed neutrophils. Of the 823 CpG, 683 exhibited the same direction of association (83%) as the CellDMC analysis. Of the 28 CpGs in the NDMI algorithm 9 (32%) exhibited a nominally statistically significant p value (p <0.05) and a consistent direction of association as compared with the CellDMC analysis.

NDMI score is a prognostic marker of glioma survival To evaluate the NDMI score as a prognostic marker in glioma, we entered patients' NDMI scores into a survival elastic net model that included conventional risk factors of age, gender, tumor classification, tumor location, surgery and chemoradiation treatments, and DEX use

https://doi.org/10.1038/s41467-022-33215-x

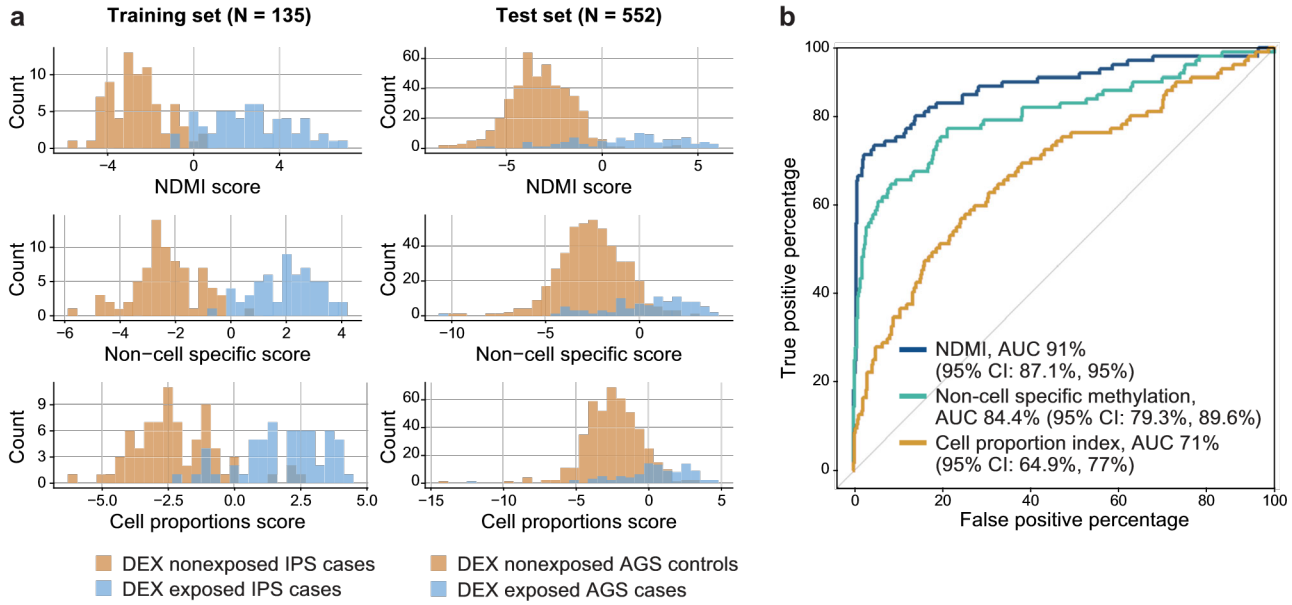


Fig. 5 | Comparison of DEX predictors in training and test datasets. and two other DEX predictor scores in training and test predictors of DEX exposure (NDMI, non-cell-specific methylation, blood leukocyte composition). a Distributions of NDMI Source data are provided as a Source Data file. datasets. b ROC analysis comparing the performance of three

Table 1 | Immune parameters for 38 Immune Profiles Study glioma patients and their neutrophil dexamethasone methylation index (NDMI) stratified by dexamethasone use

Variables	Dexamethasone at Blood Draw								P value[a]
	Yes (n = 19)				No (n = 19)				
	Mean		StDev	Range	Mean	Median	StDev	Range	Median
Average daily DEX dose[b] (mg/day)	8.85	8	5.21	(2, 16)	NA	NA	NA	NA	NA
Age at Surgery (yrs)	54.8	57	17.43	(21, 79)	47.84	49	13.28	(28, 72)	0.18
NDMI Score	2.36	2.7	1.55	(0.02, 5.37)	−1.73	−1.72	1.14	(−3.84, 0.51)	<0.0001
Absolute B cells (cells/μl)	310.4	279.8	215.7	(47.5, 846.1)	281.4	272.2	123.9	(70.2, 545.6)	0.62
Absolute CD4 (cells/μl)	517.6	440.5	443.8	(46.8, 1744.8)	930.5	848.7	466.9	(275.4, 1825.9)	0.008
Absolute CD8 (cells/μl)	374	398.3	223.1	(43.6, 818.1)	468.1	447	230.1	(130.1, 1032.2)	0.21
Absolute Monocytes (cells/μl)	536.7	524.8	357.2	(96.5, 1238.2)	489.2	474.8	140.2	(288.9, 754.6)	0.59
Absolute NK Cells (cells/μl)	246.8	222	129.5	(114.5, 671.6)	393.6	365.7	204.3	(146.4, 871.3)	0.01
Absolute Neutrophils (cells/μl)	8651.2	7475.3	4527	(1332.3, 17802)	4180	3335.6	2369.1	(1472.4, 9402)	0.0007
mMDSC CD16− (cells/μl)[c]	95.9	72.3	84.7	(2.63, 249.9)	24.5	17.2	28.8	(0, 123.5)	0.002
mMDSC CD16+ (cells/μl)[c]	28.6	15.5	31.1	(0.53, 94.4)	13.1	1.8	27.3	(0, 98.5)	0.11
mMDSC as a % of Monocytes[c]	22.1	20	15.4	(1.48, 52.2)	8.7	7	8.3	(0, 28.8)	0.002
mMDSC counts (cells/μl)[c]	124.4	101.2	112.1	(3.16, 339.5)	37.7	22.2	55.3	(0, 221.9)	0.005
Categorical variables	Number		%		Number		%		P value[a]
Gender									
Male	14		0.74		14		74%		1
Female	5		0.26		5		26%		
Tumor Grade									
Grade 4	17		0.895		7		36.80%		0.003
Grade 3	1		5.30%		4		21.10%		
Grade 2	1		5.30%		8		42.10%		

[a]T-test for continuous variables, chi-squared for categorical variables [b]Estimated by dividing the cumulative mg of DEX reported by the number of days on DEX [c]Measured by FCM other immune parameters estimated by methylation deconvolution

atblooddrawaswellasneutrophilandotherimmunecellproportions. NDMI score, but not DEX use at blood draw or neutrophil proportion, remained a significant predictor of survival time in a final model that included age, surgery, tumor location, WHO subtype, body mass index, CD4 T cell proportion, and an interaction between WHO subtype and NDMI score (Table 3). Figure 7 shows the Kaplan–Meier survival curves for NDMI score split at the median in the test set's isocitrate dehydrogenase (IDH) wildtype GBM and IDH-wildtype

Table 2 | Neutrophil dexamethasone methylation index (NDMI) scores predict CD4 T and mMDSC cell counts in Immune Profiles Study glioma patients with presurgery samples[a]

Model 1 (CD4 T cell)	Parameter estimates (dependent variable=absolute CD4 counts)				
	Predictor	Estimate	Std Error	t Ratio	P value
	Intercept	−58.27	75.84	−0.77	0.44
	NDMI Score	−158.78	11.21	−14.17	<0.0001
	Total WBC Count	0.10	0.009	11.62	<0.0001
	Interaction (NDMI*Total WBC)	−0.01	0.002	−5.69	<0.0001
	Summary of Fit				
	R^2 = 0.64, Adj R^2 = 0.63 Observations = 135 F-Ratio= 76.23 (p < 0.001)				
Model 2 (mMDSC)	Parameter estimates (dependent variable=mMDSCs)				
	Predictor	Estimate	Std Error	t Ratio	P value
	Intercept	−37.78	14.30	−2.64	0.013
	NDMI Score	14.79	2.80	5.28	<0.0001
	Absolute_Monocytes	0.20	0.025	8.09	<0.0001
	Summary of Fit				
	R^2 = 0.76, Adj R^2 = 0.74 Observations = 36[b] F-Ratio= 52.56 (p <0.001)				
Model 3 ($CD16^-$ mMDSC)	Parameter estimates (dependent variable=$CD16^-$ mMDSCs)				
	Predictor	Estimate	Std Error	t Ratio	P value
	Intercept	−26.88	10.50	−2.56	0.015
	NDMI Score	11.99	2.06	5.83	<0.0001
	Absolute_Monocytes	0.15	0.018	8.09	<0.0001
	Summary of Fit				
	R^2 = 0.77, Adj R^2 = 0.76 Observations = 36[b] F-Ratio= 56.17 (p < 0.001)				

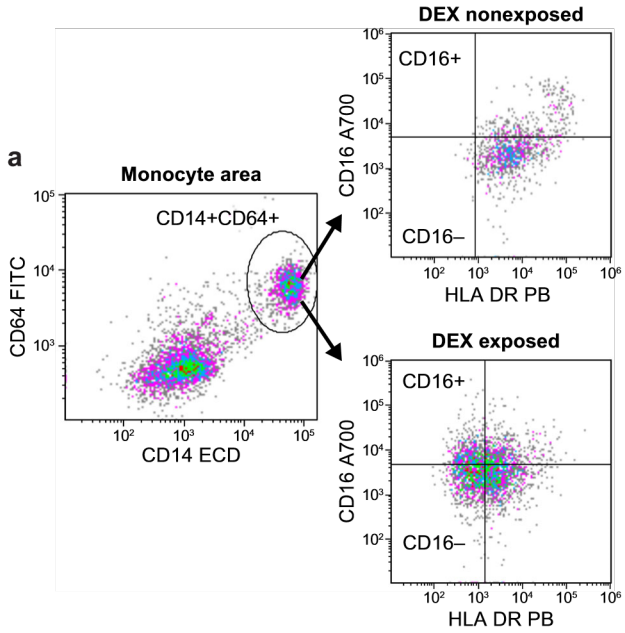
[a]Model 1 included all 135 IPS presurgery samples, Models 2 and 3 included 36 presurgery samples with FCM
[b]Two outliers were removed from the model

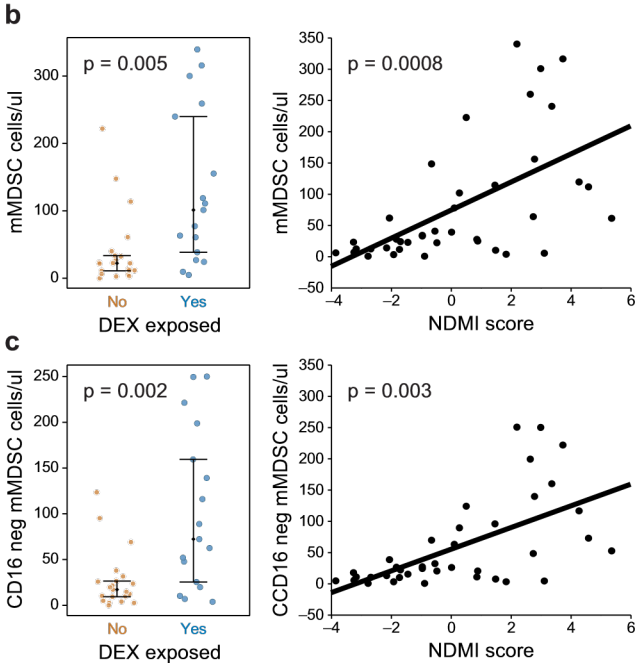
astrocytoma patients, in each subtype survival times were statistically significantly different. Survival among patients with oligodendrogliomas was not significantly different by NDMI status. In a subset of74gliomasubjectsweevaluatedbloodNDMI,immunecellestimates and clinical variables in multivariate Cox proportional hazards survival models with extent of surgical resection with and without preoperative tumor volume (Supplementary Tables 11, 12). Importantly, in both models, the NDMI score remained statistically significant as a predictor of survival. In both models, volumetric measurements were significantaswereageandWHOtumorclassification.DEXuseatblood draw was not statistically significant in either model (Supplementary Tables 11, 12).

Discussion

The present work shows that the DNA methylation-based NDMI measured in whole blood is a robust marker of DEX response that outperforms an approach that is based on the drug's effects on immune cell proportions. Thus, the NDMI could be a potentially useful tool for evaluating glucocorticoid treatments in neuro-oncology. Though further study is needed to demonstrate the full value of this approach, several areas of translational application are anticipated. The NDMI is an objective measure of glucocorticoid response that does not rely on medication logs and associated assumptions about patient adherence. The NDMI appears to capture individual variability in DEX response, as indicated by two measures of immunosuppression (low CD4 and elevated mMDSC

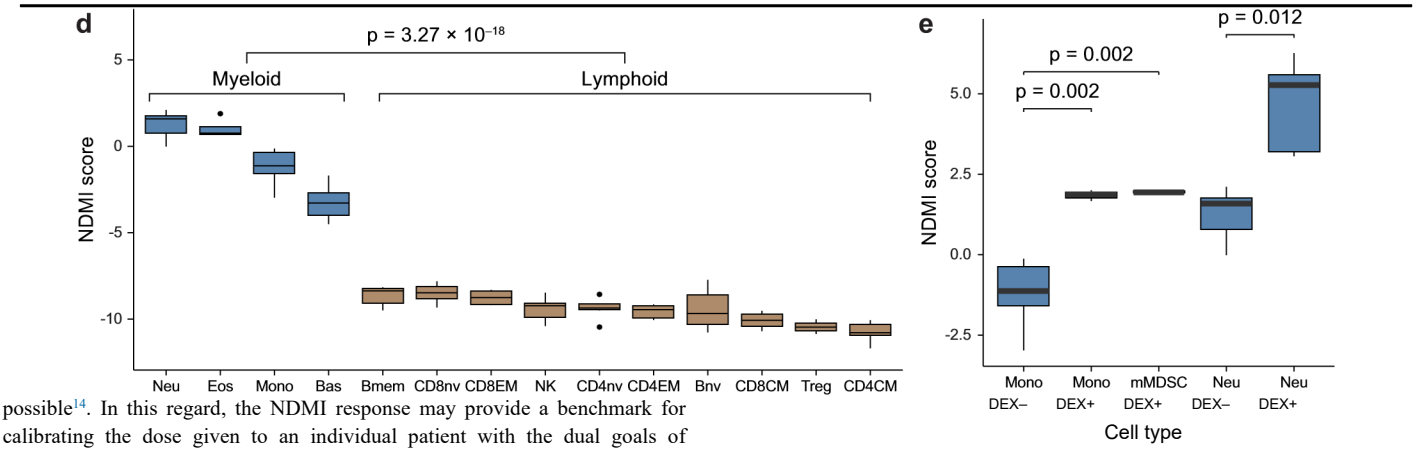
counts). The prevalence of both these cell types was best predicted by NDMI scores and not by either daily or cumulative drug doses.





These results provide another datapoint that adds to historical observations[11–13], which have raised concerns about the widespread use of glucocorticoids. Early in diagnostic workup, even the suspicion of GBM can trigger the administration of high-dose DEX in routine clinical practice, regardless of neurological symptoms, tumor size, or extent of cerebral edema. Some of these exposures may be unnecessary and may lead to worsening outcomes. There is a lack of data and consensus in the field concerning optimal glucocorticoid dosing and schedules to guide clinical decision making. An ongoing trial (NCT04266977) is the first of its kind to formally test whether restrictive use of DEX in suspected GBM subjects is an acceptable and safe alternative to current practice, though in symptomatic subjects, DEX may be unavoidable. In these instances, our observation of increased survival risk associated with the NDMI score reinforces the need to titrate to the lowest dose possible[14]. In this regard, the NDMI response may provide a benchmark for calibrating the dose given to an individual patient with the dual goals of providing adequate clinical efficacy and minimizing toxicity, in particular if there is a need to re-start steroids during treatment. Such a precision approach has the potential to significantly reduce steroid-related morbidity in the GBM patient population.

Given the convincing evidence that DEX compromises immunotherapies[14–16] the most important potential translational value of the NDMI score could be as a biomarker of response in immunotherapy trials. As an objective measure of extent of steroid exposure, the NDMI score could be utilized as a stratification factor in such trials to minimize confounding from steroid interference with therapies. Beyond these applications we observed that the NDMI score, but not DEX use, was a significant predictor of survival in IDH-wildtype GBM and astrocytoma, though not in oligodendroglioma. Thus, the



Cell type

Fig. 6 | NDMI scores in isolated DEX exposed and non-exposed myeloid and lymphoid blood immune cell subtypes. a FCM dot plots showing the gating procedure to quantify HLA-DR⁻ CD14⁺ CD16⁻ monocytes (putative mMDSCs). b Elevated total mMDSC concentrations in glioma subjects exposed to DEX (n = 38 (19 DEX exposed and 19 non-DEX exposed)) and correlation of total mMDSCs with NDMI scores. c Elevated blood concentrations of CD16⁻ mMDSC in glioma subjects exposed to DEX(n = 38(19 DEX exposed and 19 non-DEX exposed)) and correlation of CD16⁻ mMDSC concentrations with NDMI scores. For b, c; p values for difference in mean MDSC values, standard deviations and standard linear regression coefficientswerecalculatedfrom2-sidedStudent's ttests. d NDMIscores ofisolatednonDEX exposed myeloid cells are statistically significantly higher compared to nonDEX exposed lymphoid cell types. e Boxplots showing statistically significant elevations in the NDMI scores of monocytes, mMDSCs and neutrophils isolated by FACS from DEX exposed subjects compared to non-exposed cells. For d, e median and interquartile ranges are shown, two-sided Wilcoxon test used to calculate p values. Source data are provided as a Source Data file.

NDMI score may be a more relevant prognostic factor infuture studies of IDH-wildtype GBM.

Given the widespread availability of DNA methylation arrays and the open access Web-based tool that we have created, the NDMI approach is within reach for researchers wishing to assess clinical correlates with this measure. The small volumes of blood samples (0.5 mL) required, and the lack of any labor-intensive special processing needed for DNA methylation-based analyses make it an accessible technique for archival samples, and greatly enhances the value of samples inclinical biobanks. Finally, our observation of elevated NDMI scores among prednisone users in non-glioma controls highlights the applicability of this method to assessment of glucocorticoid use in a range of clinical indications beyond neuro-oncology (e.g., rheumatoid arthritis, lupus, multiple sclerosis, solid organ transplantation). We also note that, consistent with the low systemic absorption associated with inhaled glucocorticoids[40], we did not observe altered NDMI scores in controls reporting inhaled glucocorticoid use. This is consistent with a recent study in pediatric asthma[41] that failed to find consistent alterations in DNA methylation among users of inhaled glucocorticoids.

The acute effects of DEX and other glucocorticoids on peripheral blood myeloid and lymphoid cells are well described and reflect the intimate linkage of the glucocorticoid pathway with hematopoiesis and immune regulation[42,43]. We add to this knowledge that blood neutrophils dominate the leukocyte methylomic profile of DEX exposure. This result is reminiscent of earlier studies of glucocorticoidinduced gene transcription showing that the numbers of glucocorticoid-responsive genes were highest in hematopoietic compared to nonhematopoietic cells and that neutrophils had the strongest transcriptional response of all cell types studied[36]. Moreover, a heightened response to glucocorticoids by neutrophils is striking given their terminally differentiated state and short lifespan as well as their limited transcriptomic capacity compared with other leukocytes[44]. We found that DEX-responsive methylation in neutrophils was enriched in distal regulatory elements and glucocorticoid receptor binding sites. Consistent with previous work, we also showed that DEX-related differential methylation occurs at preexisting accessible chromatin locations as evidenced by their partially demethylated status in unexposed cells and occurring in a lineage-specific fashion. Amonglymphocytes,thesamesitesweretypicallydenselymethylated. A robust association of differentially methylated genes encoding granule proteins and neutrophil activation functions was found in gene ontology analysis. Twenty-one of the neutrophil enriched genes overlapped with a previously described co-expression module in human bone marrow whose peak transcriptional expression occurs during the metamyelocyte and band neutrophil stages of myelopoiesis. In contrast, the group of 16 genes that comprise the NDMI predictor were distinct in that they did not belong to the same coexpression group and instead showed peak expression in the promyelocytic and PMN stages of myelopoiesis. Taken together, these results indicate functional heterogeneity among genes targeted by DEX differential methylation.

Isolated neutrophils from DEX exposed subjects displayed the highest NDMI scores of all isolated cells and in each case the scores were elevated

Table 3 | Neutrophil dexamethasone methylation index (NDMI) is a prognostic factor in glioma survival: Cox multivariate model of NDMI and glioma survival in 429 UCSF Adult Glioma Study patients

Variable	Hazard ratio	95% CI	P value**
NDMI score	1.43	1.29–1.59	8.5×10^{-12}
Age (yrs, continuous)	1.04	1.03–1.05	2×10^{-16}
World Health Organization 2016 classification			
IDH-WT GBM	1.89	1.44–2.51	6.8×10^{-6}
IDH-MT astrocytoma	0.33	0.06–1.95	0.2
Oligodendroglioma	0.18	0.10–0.33	2.3×10^{-8}
IDH-WT astrocytoma (baseline)	1.0	NA	NA
Interactionof NDMIscorewith WorldHealthOrganization 2016classification			
IDH-WT GBM	0.85	0.76–0.94	2×10^{-3}
IDH-MT astrocytoma	0.56	0.31–1.03	0.06
Oligodendroglioma	0.74	0.59–0.92	7.7×10^{-3}
IDH-WT astrocytoma (baseline[a])	1.0	NA	NA
Surgery (biopsy versus resection)	0.52	0.37–0.73	1.5×10^{-4}
CD4 T cells (proportions)	1.04	1.01–1.07	2×10^{-3}
BMI (basal metabolic index)	1.05	1.02–1.07	1.2×10^{-4}
Race (white vs. non-white)	1.47	1.07–2.03	0.02
Tumor location			
Occipital lobe	3.34	1.73–6.47	3.4×10^{-4}
Overlapping sites	1.56	1.14–2.14	5.8×10^{-3}
Parietal lobe	0.65	0.45–0.95	0.03
Frontal/cerebrum/temporal/other (baseline)	1.0	NA	NA

**P values were calculated using the Wald Test [a]IDH-WT Astrocytoma chosen as comparator baseline for the model

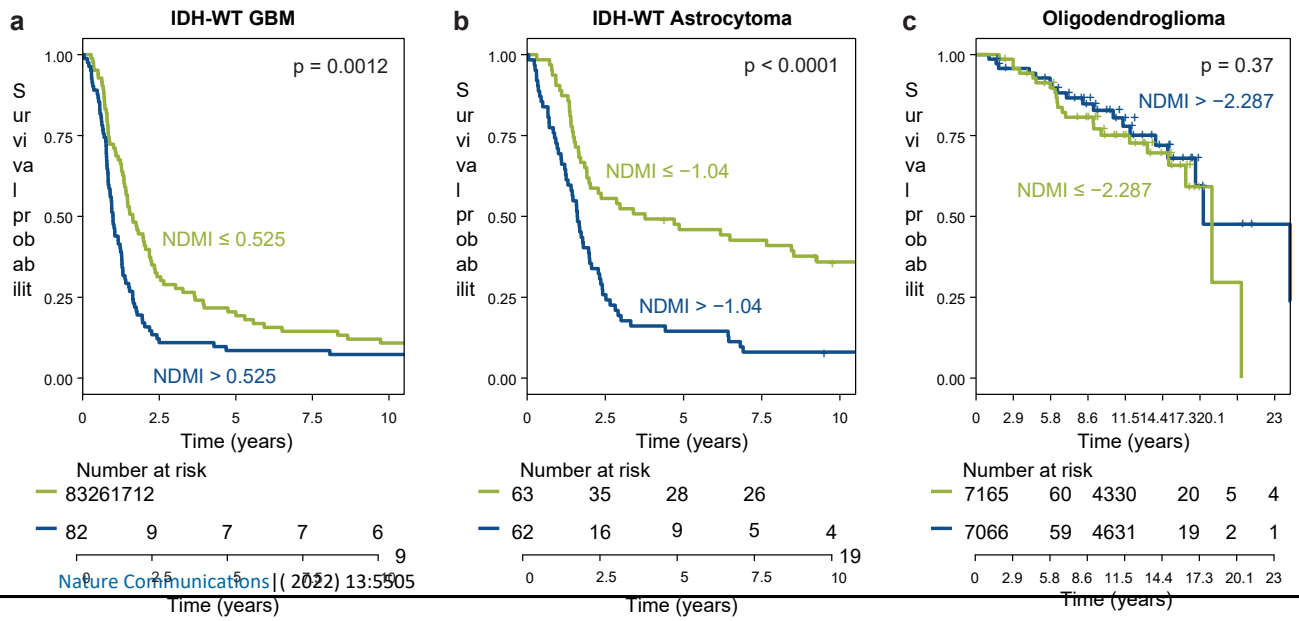


Fig. 7 | Association of NDMI scores and glioma survival. a The figure shows the survival curves for NDMI score split at the median in the test set's patients. b The figure shows the Kaplan–Meier survival curves for NDMI score split at the median in the test set's IDH-wildtype Astrocytoma patients. c Survival among patients with oligodendrogliomas was not significantly different by NDMI status. Source data are provided as a Source Data file.

above paired whole-blood values from the same donors. These observations are consistent with our bioinformatic predictions. However, despite the lack of statistically significant interactions of methylation values and monocyte proportions, our methylation and FCM studies of mMDSCs also point to DEX-mediated epigenetic modification in subpopulations of $CD14^+$ monocytes. We confirmed previously reported increased levels of mMDSCs among current DEX users and further those variations in mMDSCs were significantly correlated with individual NDMI scores. We also provide evidence that the monocyte count in conjunction with the NDMI score may improve our understanding of individual variations in mMDSCs. $CD16^-$ mMDSCs also demonstrated consistent associations with DEX exposure. Earlier studies reported that $CD16^-$ mMDSC populations display superior immune-suppressive activity compared with $CD16^+$ fractions[45]. Significantly, of the three DEX exposure measures, only the average daily dose was associated with mMDSC concentrations. In models that contained monocytes and NDMI, daily DEX dose became insignificant, suggesting that the NDMI score captures individual variation in the epigenetic response to DEX treatments. Furthermore, the NDMI scores of isolated mMDSC and paired $CD14^+$ $HLA\text{-}DR^+$ monocytes from DEX exposed subjects were elevated compared to non-exposed monocytes. Based on these observations, we hypothesize that DEX influences both the neutrophil and monocyte cell methylomes, perhaps reflecting the shared DNA methylation landscape of DEX-responsive regions in both lineages. We have confirmed previously reported deficits in circulating CD4 T cell levels among glioma subjects prior to chemoradiation; 11.8% of our training population demonstrated <200 CD4 T cells/μl. A model incorporating an interaction between the NDMI, and total white cell count provided the strongest predictor of variations in CD4 levels. DEX exposure or clinicopathological variables did not improve model performance. Thus, although excess mMDSCs and low CD4 T cells often occur together in cancer patients, our data suggest that different features of the immune profile in conjunction with the NDMI can be useful to describe variation in these two important immunological parameters.

Thecurrentworkhasseverallimitations.First,oursurvivalstudies were not a randomized controlled study and in the biomarker development phase detailed longitudinal studies were not available to estimate the kineticsof change inDNA methylationduring initiation of drugtherapyortherecoverytimefollowingcessation.Ouranectodical observations indicate that NDMI score can return to pre-therapy levels following ~7 days. Second, while the empirical associations of DNA methylation with DEX treatments are robust, a mechanistic understanding of these effects is lacking. At the tissue level, it is unknown whether DEX-related changes in DNA methylation arise through redistribution of preexisting subpopulations of cells in the bone marrow or marginated pools or instead reflect CpG level modification within stem/progenitor cells during myelopoiesis. At the molecular level, glucocorticoid ligand-triggered demethylation of single CpGs at de novo sites are rapid and replication independent, suggesting an active enzymatic mechanism[32]. In rodent embryonic neural stem cells, Tet3, a member of the ten-eleven translocation family of methylcytosine dioxygenases, was implicated in mediating the effects of DEX on CpG methylation[33]. Whether similar mechanisms apply in adult human myelopoiesis is unknown. Finally, our studies have not assessed gene expression in DEX exposed cells. Recent studies in severe COVID-19 have shown that in vivo DEX exposures induce dramatic changes in peripheral blood neutrophil states leading to immunosuppressive cell subtypes not observed in unexposed subjects[37]. The relationship of different transcriptional states with DNA methylation could provide a mechanisticlinkbetweenDEXexposures,immuneprofiles,andglioma survival. In conclusion, we find that alterations in DNA methylation of peripheral whole blood are sensitive indicators of DEX response and are potential pharmacodynamic markers of individual sensitivity to epigenetic modification that are bridged to glioma survival, elevated immunosuppressive myeloid cells and depletion of critical T cell populations.

Methods

Patient and control samples

These glioma studies were approved by the Institutional Review Board of the University of California, San Francisco, Human Research Protection Program in the Office of Ethics and Compliance under UCSF Federal wide Assurance 00000068 and met all relevant ethical regulations. Informed consent was obtained from all study participants. The study design utilized a training dataset to develop predictors of DEX exposure and a test dataset to evaluate their performance Fig. 1a. The training dataset included 135 glioma subjects (59 exposed to DEX, 76 not exposed to DEX) sampled before surgery, radiation, or temozolomide treatments. To discover DEX-responsive CpG sites, we carried out an EWAS and then applied those results to the CellDMC program, a method to detect DNA methylation within a specific cell type, Fig. 1b. CellDMC is a linear modeling program that tests each locus associated with DEX to interact with the relative abundance of 6 leukocyte cell types.[46] Separate models were evaluated for each CpG probe and its interaction with CD4 T, CD8 T, B-cell, NK, monocyte, or neutrophilproportions.ThetrainingdatasetwasdrawnfromtheUCSF Immune Profiles Study (IPS), a prospective neuro-oncology clinicbased collection of blood samples, imaging, and other clinical data from adult glioma patients. Presurgery blood samples were typically taken the day before surgery; none were obtained during or after exposure to anesthesia. Blood samples were transferred the same day as drawn for determining total cell counts of specific immune cell types, further processing, and FCM study. At time of blood draws, we also collected a presurgery questionnaire to document daily/cumulative DEX exposure. Additional post-surgery blood samples were collected on some subjects to explore changes in immune parameters following resection. The test dataset (n =99 glioma cases, n= 454 controls) and the survival evaluation set (n= 429 glioma cases) were drawn from the San Francisco Adult Glioma Study (AGS). AGS was a case-control study of glioma patients newly diagnosed between 1991 and 2012 who were residents of the San Francisco Bay Area or patients of the UCSF Neuro-oncology clinic and age and gender-matched controls without any history of glioma recruited through random digit dialing or from the UCSF blood draw lab[47]. Medicalhistoryand current medications were available for all the controls. Patients were selected who had both archival blood and tumor marker data[47]. AGS participants represent newly diagnosed glioma patients; no recurrent or secondary GBM cases are included. In both IPS and AGS cohorts, we classified cases according to the WHO 2016 classification scheme that delineates IDH-mutant and wildtype grade IV GBM, grades II and III astrocytoma, and IDH-mutant 1p/19q co-deleted oligodendrogliomas. In the AGS, blood samples were collected from patients a median of 100 days after they were histologically diagnosed. Clinical information was collected on patient treatments, including temozolomide chemotherapy, radiation therapy, the extent of surgery, and steroid useat the time of blood sampling. Information on current medication use at blood draw (including any steroid use and names of the types of steroid drugs taken) was obtained from an in-person blood draw questionnaire which collected information on treatments and medications taken at the time of blood draw; medication logs for DEX dosages were not available.

DNA methylation array

Frozen (−80°C) anticoagulated whole blood or isolated cells were processed, DNA isolated, and bisulfite converted[48]. All samples and array experiments were performed blinded to clinicopathologic variables. Approximately 200–500ng ofDNA was applied to Illumina DNA methylation EPIC 850K beach chip arrays. Preprocessing and quality control of fluorescence data were accomplished using the minfi Bioconductor package[49]. CpG loci having a sizable fraction (>25%) of detection p values above a predetermined threshold were filtered (detection P $>10E^{-5}$)[50]. A "noob" background correction method was used to account for background fluorescence and dye bias[51]. Betamixture quantile normalization (BMIQ) was performed to correct probe design bias[52]. The presence of technical sources of variability induced by plate or BeadChip was examined using principal components analysis (PCA), the top K principal components[53] were examined in terms of their association with plate and

BeadChip. If plate or BeadChip were found to be significantly associated with any of the top K principal components, we applied ComBat method[54] for normalization using the sva Bioconductor package. The final dataset contained 830,277 probes.

Immunomethylomic assay

Using the preprocessed and normalized methylation data, we applied an optimized reference-based cell mixture deconvolution[55]. This immunomethylomic method provides the proportions of 6 major cell types (CD4 T, CD8 T, B-cell, NK, monocytes, neutrophils) and the neutrophil/lymphocyte, CD4/CD8, and lymphocyte/monocyte ratios. DNA methylation-based immune profiles are highly accurate and reproducible[48,55,56] (Supplementary Table7,Supplementary Fig.6).The cellfractionsweremultipliedbythesampletotalwhitecellcountwhen these data were available (IPS samples) to estimate absolute cell counts. Historical AGS cases did not have concurrent cell count data. To characterize the purity of isolated neutrophils and explore the NDMIscores withindifferent immune cells, we employed anexpanded deconvolution technique that captures 12 different cell types[39].

Elastic net regression and creation of three DEX exposure predictors

We built three predictors of DEX exposure using elastic net regression procedures on the following data:(i) blood leukocyte composition, (ii) neutrophil-specific methylation, and iii) non-cell-specific methylation. Developing these markers required blood leukocyte composition, which we derived from a validated DNA methylation deconvolution method[48,55,56]. We then created a leukocyte composition index as input for DEX exposure prediction (i) we utilized 6 cell proportions and ratios of neutrophil/lymphocyte, CD4/CD8 T cell, and lymphocyte/ monocytes. To discover in vivo DEX-responsive CpGs in the training dataset for predictors (ii) and (iii), we performed an epigenome-wide association study (EWAS) using the "lmfit" function in the limma package[57] comparing methylation beta values for patients on DEX at the time of blood draw and patients not taking DEX at the time of blood draw. The EWAS yielded a list of probes nominally significantly differentially methylated $p < 0.05$ with an absolute delta beta value (≥ 0.10). CpGs identified in EWAS were further analyzed the CellDMC[46] function of the EpiDISH package using R with the default threshold parameters (FDR ≤ 0.05)[58]. Neutrophil-specific CpG loci (N =2621) and non-cell-specific CpG loci (N = 17,733) were identified and used to develop the methylation DEX-responsive predictors. Elastic net regression was used to build the DEX exposure classifiers using (i) leukocyte composition, the (ii) neutrophil-specific, and (iii) non-cellspecific CpGs identified from CellDMC[46]. The R glmnet was used to build the elastic net models and cross-validation to identify the optimum lambda value[59]. Each of the three predictors was tested using an external dataset consisting of 99 DEX^+ glioma samples and 453 DEX^- control samples. An ROC curve was built to compare the predictive capability of the three models on the same test set[60].

Genomic features of DEX-responsive CpG methylation

Odds ratio and p values are calculated to test for NR3C1 binding site enrichment in differentially methylated neutrophil-specific CpGs (n =2621)usingthetotalmeasuredasbackground(EWASn= 830,277). Genomic information for transcription factor NR3C1 binding site is extracted from the ENCODE database "Transcription Factor ChIP-seq Clusters", version wgEncodeRegTfbsClusteredV3. The Illumina HumanMethylationEPIC annotation file and the UCSC Genome Browser UCSC_hg19_refGene file were used to test neutrophil-specific CpGs (n =2621) for genomic context enrichment among all tested CpGs (n =830,277).Therelationto CpGislandsandenhancerswasidentified from the Illumina HumanMethylationEPIC annotation file for each probe. To define the genomic regions as promoters, introns, exons, or intergenic for each probe, the annotateWithGeneParts function from the R-package genomation and the UCSC_hg19_refGene file were used to map the regions to all CpG loci on the Illumina HumanMethylationEPIC array. If a probe mapped to more than a single genomic region, theprobewasassignedpreferentiallywiththeorderpromoters,exons, introns, and intergenic. Fisher's exact tests of 2 × 2 tables were conductedtocalculateoddsratios(ORs),pvalues,and95%confidence intervalforCpGisland,enhancers, andgenomicregionsenrichmentof neutrophil-specific CpGs compared to all other DEX differentially methylated loci. eFORGE 2.0 was used to assign histone chromatin marks to differentially methylated probes[61] in the 28 CpG NDMI.

FCM detection of blood mMDSC and CD4 T cells

MDSCs are a heterogeneous population of immature myeloid cells, and the definition of these subsets has not yet reached a consensus. We defined mMDSC immunophenotype consistent with commonly employed markers[62,63], substituting the myeloid marker CD64 for CD33. Relative and absolute quantities of total ($CD14^+CD64^+CD11b^+HLA\text{-}DR^{neg/low}$) and CD16 negative ($CD14^+CD64^+CD11b^+HLA\text{-}DR^{neg/low}CD16^-$) mMDSC populations were estimated. FCM was performed by the UCSF clinical immunology core using established clinical protocols and analyticspecific reagents. Blood samples (3 ml EDTA preserved whole blood) were collected, and total white blood cell (WBC) counts were recorded. Following red blood cell lysis, cells were mixed with a monoclonal antibody mixture including CD45, CD64, CD16, CD11b, HLA-DR (Supplementary Table 2) in a single tube, incubated in the dark for 20 min, washed, and fixed with 0.1% formaldehyde prior to acquisition on the Navios EX Flow Cytometer (Beckman Coulter Life Sciences). Data were analyzed using Kaluza Analysis Software 2.1 (Beckman Coulter Life Sciences). Forward and side scatter were employed to discriminate the non-doublet, non-debris, $CD45^+$ leukocyte population and cells counted (Supplementary Fig. 8). Further $CD64^+CD14^+$ gating identified the monocyte population. The HLA-DR $^{neg/low}$ gate in the monocyte population was set in each sample according to the HLA-DR $^{neg/low}$ granulocyte population in the same tube. These HLA-DR $^{neg/low}$ monocytes were further gated for $CD16^{+/-}$ expression. The proportion of each cell type present in the sample was then calculated based on the number of $CD45^+$ leukocytes. The mMDSC proportion of $CD45^+$ leukocytes counted in the acquisition tube was calculated. Total WBC counts were then used to calculate the final cell/µl concentrations of mMDSCs in the blood. Antibodies were used to stain CD45, CD3, and CD4 to identify the CD4 T cells. Within the CD45 population, the low side scatter intensity $CD3^+$ cluster identified the $CD3^+$ population. These $CD3^+$ cells were further gated for $CD4^{+/-}$ expression. The CD4 T-cell proportion of CD45+ leukocytes counted in the acquisition tube was calculated. Total WBC counts were then used to calculate the final cell/µl concentrations of CD4 T cells in the blood. Linear regression models were constructed with NDMI and other factors as predictors to explore CD4 T cell and mMDSC parameters as outcomes. Examination of residuals led to dropping two outliers from mMDSC models; these subjects were of advanced age (i.e., 72, 75 yrs.) and were sampled soon after receiving high DEX doses (i.e., 16 mg/day).

FACS isolation of mMDSCs, monocytes and neutrophils

Fresh anticoagulated blood was processed within 24h. For mMDSCs and monocytes: blood mononuclear cells were isolated with 1.077 Histopaque gradients, stained with a cocktail of fluorescently labeled antibodies (CD3, CD56, CD19, CD14, CD11b, CD16, HLA-DR, CD33, and CD15 (Supplementary Table 6) treated with PE/Cyanine7 Streptavidin andresuspendedat1:5000dilutionofSYTOXTMGreen.Cellswerethen run directly on a BD FACSAriaTM Fusion cell sorter. Forward scatter hi $CD3^-$, $CD19^-$, $CD11b^+$, $CD33^+$, $CD14^+$, and $CD15^-$ cells were gated and plotted for HLA-DR expression. CD3 $HLA\text{-}DR^-$ neg cells and CD19 $HLADR^+$ positive B cells were used to set the sorting gate for mMDSC cells lacking HLA-DR expression (i.e., $HLA\text{-}DR^{neg/low}$). $HLA\text{-}DR^{high}$ cells (normal monocytes) were collected from the same individuals. For neutrophil isolations,aHistopaquetwo-stepgradient (1.119layeredwith1.077)was used. Granulocytes were collected on the 1.119 layer above the red blood cells and below monocytes and lymphocytes; residual red cells were lysed with ammonium chloride. Cells were FcBlocked and stained with CD3, CD56, CD123, and CD49d in a negative depletion channel. Neutrophils were then sorted as $CD19^-$, $CD33^-$,

and CD66b+ cells (Supplementary Table 13). Gated cells were plotted against FSC and SSC to select cells with high forward and side scatter characteristic of granulocytes. All isolated cell pellets were stored at −80°C until DNA methylation analysis. The purity of isolates was checked using a highdefinition immune cell methylation deconvolution method[39].

Survival analysis

Overall survival is defined as the time from the initial diagnosis to death or last follow-up. Those patients without a date of death were censored at the date of the last follow-up. Multivariate survival models were initially built via regularized cox regression[55] using the glmnet package in R. Cross-validation was performed to identify the optimum lambda value[50]. Subsequently, forward/backward stepwise feature selection was employed, and interactions were investigated. Covariates with p values< 0.05 were retained. For each of the three most common WHO 2016 subtypes (IDH-wildtype GBM, IDH-wildtype astrocytoma, and oligodendroglioma), the 1st and 3rd quartile of NDMI scores were estimated, and plots of the corresponding Kaplan–Meier curves were drawn.

Simulation study of NDMI performance and neutrophil proportions

We conducted a series of simulation studies to assess the statistical power for detecting a difference in the mean value of the NDMI between DEX exposed and non-exposed individuals (Supplementary Fig. 7). Specifically, our interest centered on determining the nature by which varying fractions of neutrophils in the whole-blood samples used as the basis for calculating NDMI impacts the statistical power for detecting a difference in the mean value of the NDMI between DEX exposed and non-exposed individuals. To achieve this goal, we simulated whole-blood DNA methylation data for DEX exposed and non-exposed individuals, varying the fraction of neutrophils across the study samples from 20% to 80% in increments of 5%. For each simulated dataset, the NDMI was calculated for each sample and used to test the null hypothesis of no difference in the mean NDMI between DEX exposed and non-exposed individuals. For a fixed fraction of neutrophils across the study samples (e.g., 20%), the above process was repeated one-thousand times, and statistical power was calculated as the proportion of times the null hypothesis was correctly rejected.

To simulate in-silico whole-blood methylation data for DEX exposedandnon-exposedindividuals,wefirstusedapubliclyavailable dataset consisting of cell-specific DNA methylation data profiled using the Illumina HumanMethylationEPIC array for various isolated leukocyte subtypes, including: CD4 T cells, CD8T cells, B cells, natural killer cells, monocytes, and neutrophils[55] GEO: GSE110554) to estimate the cell-specific means and standard deviations of the 28 CpGs that comprisetheNDMI.Thecell-specificmeansandstandarddeviationsforthe 28 NDMI CpGs were then used to calculate the two shape parameters of the beta-distribution using methods of moments estimation. Using the beta-distribution shape parameters, we next generated cellspecific methylation data, $X^{\delta 0 \flat} = X^{\delta}_{1}{}^{0\flat}, X_2^{\delta 0\flat}, \ldots, X^{\delta}_{n_0}{}^{0\flat}$, for n_0 nonexposed individuals by randomly sampling methylation beta values for the 28 NDMI CpGs using the previously estimated shape parameters. Here, $X^{\delta}_{j}{}^{0\flat}$ is the cell-specific methylation data for individual j, and is a 28 ×6 matrix whose rows represent the 28 NDMI CpGs and whose columns are the 6 leukocyte subtypes. The same process was performed for generating cell-specific methylation for n_1 DEX exposed individuals, $X^{\delta 1\flat} = X^{\delta}_{1}{}^{1\flat}, X_2^{\delta 1\flat}, \ldots, X^{\delta}_{n_1}{}^{1\flat}$, except that the beta values simulated for neutrophils across the 28 NDMI CpGs weregenerated by imposing a difference in mean methylation consistent with that calculated in a separate dataset consisting of monocyte-specific methylation data in n= 6 non-exposed individuals and n=3 DEX exposed individuals. Using the cell-specific methylation data for the n_0 nonexposed and n_1 DEX exposed individuals, we next generated cell proportions for each of the samples $W = \delta w_1, w_2, \ldots, w_N \flat$ by simulating data from a Dirichlet distribution with total concentration parameter (α_0) equal to 18, 73, and 127. Here, w_j is a 1 ×6 vector representing the cell proportions for subject j and $N = n_0 + n_1$ is the total number of samples, which ranged from N =20 to N = 80 in increments of 20. The totalconcentrationparameterisrelatedtothevariabilityinthecellular landscape across samples wherein larger values are associated with less variability inthe cellularlandscape acrosssamples and conversely, smaller values, more variability in the cellular landscape across samples (Supplementary Fig. 7). The rationale for the choice of the total concentration parameters is described elsewhere[64]. In simulating cell proportions for each of the samples, the Dirichlet parameter for neutrophils was varied such that the mean fraction of neutrophils ranged from 0.20 to 0.80 in increments of 0.05. Having generated the cellspecific methylation data $\delta X^{\delta\ \flat 0}, X^{\delta\ \flat 1}\ \flat$ and cell proportions W, we next computedwhole-bloodmethylationsignaturesforeachofthesamples $Y = \delta Y_1, Y_2, \ldots, Y_N \flat$ by taking the matrix product of the cell-specific methylation data for a given sample and their simulated cell proportion:

$$\text{non exposed} : Y_j = X^{\delta 0\flat}_{j} w^{T}_{j} \qquad \delta^{3}\flat$$

$$\text{Dex exposed} : Y_k = X_k{}^{\delta 1\flat} w^{T}_{k} \qquad \delta^{4}\flat$$

Where j=1, 2,...,n_0 and k = 1, 2,...,n_1. Using the whole-blood methylation signatures, Y= (Y_1, Y_2,...,Y_N) we next calculated the NDMI for each of the N samples and performed a two-sample t test to test the null hypothesis of no difference in the mean NDMI between DEX exposed and non-exposed individuals assuming a type 1-error rate of 5%. For a fixed proportion of neutrophils in whole blood, overall sample size N and the Dirichlet concentration parameter, the process described above wasrepeated 1000 times and statistical power was calculated as the proportion of times the null hypothesis was correctly rejected.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Methylation and phenotype data used in this paper are available through dbGaP controlled access. Methylation and phenotype data from the Adult Glioma Study are available through dbGaP Study Accession phs001497.v2.p1. Methylation and phenotype data from the Immune Profiles Study are available through dbGaP Study Accession phs002998.v1.p1. Please note that we do not have IRB approval to releaseindividual level data for 15 out of457AGS controls (fromSeries 1), so these controls are not included in the source files. The remaining data are available within the Article, Supplementary Information. Source data are provided with this paper.

Code availability

The online application we developed using the R Shiny application to calculate neutrophil dexamethasone methylation index (NDMI) scores can be found at the following URL: https://btcdb.shinyapps.io/ndmi/

References

1. Ramamoorthy, S. & Cidlowski, J. A. Corticosteroids: Mechanisms of Action in Health and Disease. Rheum. Dis. Clin. North Am. 42, 15–31 (2016).
2. Pelt, A. C. (ed.) Glucocorticoids: effects, actions, mechanisms and therapeutic uses. (Nova Sciences Publishers, 2010).
3. Quax, R. A. et al. Glucocorticoid sensitivity in health and disease. Nat. Rev. Endocrinol. 9, 670–686 (2013).
4. Schijvens, A. M., Ter Heine, R., de Wildt, S. N. & Schreuder, M. F. Pharmacology and pharmacogenetics of prednisone and prednisolone in patients with nephrotic syndrome. Pediatr. Nephrol. 34, 389–403 (2019).
5. Czock, D., Keller, F., Rasche, F. M. & Häussler, U. Pharmacokinetics and pharmacodynamics of systemically administered glucocorticoids. Clin. Pharmacokinet. 44, 61–98 (2005).

6. Cooper, M. S., Seibel, M. J. & Zhou, H. Glucocorticoids, bone and energy metabolism. Bone 82, 64–68 (2016).
7. Hathout, Y. et al. Serum pharmacodynamic biomarkers for chronic corticosteroid treatment of children. Sci. Rep. 6, 31727 (2016).
8. Kauh, E. et al. Prednisone affects inflammation, glucose tolerance, and bone turnover within hours of treatment in healthy individuals. Eur. J. Endocrinol. 166, 459–467 (2012).
9. van Raalte, D. H. & Diamant, M. Steroid diabetes: from mechanism to treatment? Neth. J. Med. 72, 62–72 (2014).
10. Phillips, K. A., Fadul, C. E. & Schiff, D. Neurologic and medical management of brain tumors. Neurol. Clin. 36, 449–466 (2018).
11. Petrelli, F. et al. Steroids use and survival in patients with glioblastoma multiforme: a pooled analysis. J. Neurol. 268, 440–447 (2021).
12. Pitter, K. L. et al. Corticosteroids compromise survival in glioblastoma. Brain 139, 1458–1471 (2016).
13. Shields, L. B. et al. Dexamethasone administration during definitive radiation and temozolomide renders a poor prognosis in a retrospective analysis of newly diagnosed glioblastoma patients. Radiat. Oncol. 10, 222 (2015).
14. Iorgulescu, J. B. et al. Concurrent dexamethasone limits the clinical benefit of immune checkpoint blockade in glioblastoma. Clin. Cancer Res. 27, 276–287 (2021).
15. Keskin, D. B. et al. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. Nature 565, 234–239 (2019).
16. Reardon, D. A. et al. Effect of nivolumab vs bevacizumab in patients with recurrent glioblastoma: the CheckMate 143 phase 3 randomized clinical trial. JAMA Oncol. 6, 1003–1010 (2020).
17. Fauci, A. S. Mechanisms of corticosteroid action on lymphocyte subpopulations.I.RedistributionofcirculatingTandblymphocytes to the bone marrow. Immunology 28, 669–680 (1975).
18. Fauci, A. S., Dale, D. C. & Balow, J. E. Glucocorticosteroid therapy: mechanisms of action and clinical considerations. Ann. Intern. Med. 84, 304–315 (1976).
19. Mager, D. E., Lin, S. X., Blum, R. A., Lates, C. D. & Jusko, W. J. Dose equivalency evaluation of major corticosteroids: pharmacokinetics and cell trafficking and cortisol dynamics. J. Clin. Pharm. 43, 1216–1227 (2003).
20. Chitadze, G. et al. In-depth immunophenotyping of patients withglioblastoma multiforme: Impact of steroid treatment. Oncoimmunology 6, e1358839 (2017).
21. Dubinski, D. et al. Controversial roles for dexamethasone in glioblastoma—opportunities for novel vascular targeting therapies. J. Cereb. Blood Flow. Metab. 39, 1460–1468 (2019).
22. Hughes, M. A., Parisi, M., Grossman, S. & Kleinberg, L. Primary brain tumors treated with steroids and radiotherapy: low CD4 counts and risk of infection. Int J. Radiat. Oncol. Biol. Phys. 62,1423–1426 (2005).
23. Hui, C. Y., Rudra, S., Ma, S., Campian, J. L. & Huang, J. Impact ofoverall corticosteroid exposure during chemoradiotherapy on lymphopenia and survival of glioblastoma patients. J. Neurooncol. 143, 129–136 (2019).
24. Dubinski, D. et al. CD4+ T effector memory cell dysfunction isassociated with the accumulation of granulocytic myeloid-derived suppressor cells in glioblastoma patients. Neuro Oncol. 18, 807–818 (2016).
25. Dubinski, D. et al. Dexamethasone-induced leukocytosis is associated with poor survival in newly diagnosed glioblastoma. J. Neurooncol. 137, 503–510 (2018).
26. Gustafson, M. P. et al. Systemic immune suppression in glioblastoma: the interplay between CD14+HLA-DRlo/neg monocytes, tumor factors, and dexamethasone. Neuro Oncol. 12, 631–644 (2010).
27. Moyes, K. W. et al. Effects of tumor grade and dexamethasone on myeloid cells in patients with glioma. Oncoimmunology 7, e1507668 (2018).
28. Kleinberg, L., Sloan, L., Grossman, S. & Lim, M. Radiotherapy, lymphopenia, and host immune capacity in glioblastoma: a potentially actionable toxicity associated with reduced efficacy of radiotherapy. Neurosurgery 85, 441–453 (2019).
29. Burd, C. J. & Archer, T. K. Chromatin architecture defines the glucocorticoid response. Mol. Cell Endocrinol. 380, 25–31 (2013).
30. John, S. et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. 43, 264–268 (2011).
31. Love, M. I. et al. Role of the chromatin landscape and sequence in determining cell type-specific genomic glucocorticoid receptor binding and gene regulation. Nucleic Acids Res. 45, 1805–1819 (2017).
32. Wiench, M. et al. DNA methylation status predicts cell type-specific enhancer activity. Embo J. 30, 3028–3039 (2011).
33. Bose, R. et al. Tet3 mediates stable glucocorticoid-induced alterations in DNA methylation and Dnmt3a/Dkk1 expression in neural progenitors. Cell Death Dis. 6, e1793 (2015).
34. Zannas, A. S. Epigenetics as a key link between psychosocial stressand aging: concepts, evidence, mechanisms. Dialogues Clin. Neurosci. 21, 389–396 (2019).
35. Berretta, E., Guida, E., Forni, D. & Provenzi, L. Glucocorticoidreceptor gene (NR3C1) methylation during the first thousand days: Environmental exposures and developmental outcomes. Neurosci. Biobehav Rev. 125, 493–502 (2021).
36. Franco, L. M. et al. Immune regulation by glucocorticoids can belinked to cell type-dependent transcriptional responses. J. Exp. Med. 216, 384–406 (2019).
37. Sinha, S. et al. Dexamethasone modulates immature neutrophils and interferon programming in severe COVID-19. Nat. Med. 28, 201–211 (2022).
38. Hoogendijk, A. J. et al. Dynamic transcriptome-proteome correlation networks reveal human myeloid differentiation and neutrophilspecific programming. Cell Rep. 29, 2505–2519.e2504 (2019).
39. Salas, L. A. et al. Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. Nat. Commun. 13, 761 (2022).
40. Lipworth, B. J. Systemic adverse effects of inhaled corticosteroidtherapy: a systematic review and meta-analysis. Arch. Intern. Med. 159, 941–955 (1999).
41. Kere,M. et al. Effects of inhaled corticosteroids on DNA methylation in peripheral blood cells in children with asthma. Allergy 75, 688–691 (2020).
42. Ince, L. M., Weber, J. & Scheiermann, C. Control of leukocyte trafficking by stress-associated hormones. Front. Immunol. 9, 3143 (2018).
43. Shimba, A. & Ikuta, K. Control of immunity by glucocorticoids inhealth and disease. Semin. Immunopathol. 42, 669–680 (2020).
44. Ecker, S. et al. Genome-wide analysis of differential transcriptionaland epigenetic variability across human immune cell types. Genome Biol. 18, 18 (2017).
45. Wang, K. et al. Early myeloid-derived suppressor cells (HLA-DR()/(low)CD33(+)CD16(-)) expanded by granulocyte colonystimulating factor prevent acute graft-versus-host disease (GVHD) in humanized mouse and might contribute to lower GVHD in patients post allo-HSCT. J. Hematol. Oncol. 12, 31 (2019).
46. Zheng, S. C., Breeze, C. E., Beck, S. & Teschendorff, A. E. Identification of differentially methylated cell types in epigenome-wide association studies. Nat. Methods 15, 1059–1066 (2018).
47. Wrensch, M. et al. Diagnostic, treatment, and demographic factors influencing survival in a population-based study of adult glioma patients in the San Francisco Bay Area. Neuro Oncol. 8, 12–26 (2006).
48. Accomando, W. P., Wiencke, J. K., Houseman, E. A., Nelson, H. H. &Kelsey, K. T. Quantitative reconstruction of leukocyte subsets using DNA methylation. Genome Biol. 15, R50 (2014).

49. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30, 1363–1369 (2014).
50. Wilhelm-Benartzi, C. S. et al. Review of processing and analysismethods for DNA methylation array data. Br. J. Cancer 109, 1394–1402 (2013).
51. Triche, T. J. Jr., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina infinium DNA methylation BeadArrays. Nucleic Acids Res. 41, e90 (2013).
52. Teschendorff, A. E. et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 29, 189–196 (2013).
53. Teschendorff, A. E., Zhuang, J. & Widschwendter, M. Independentsurrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics 27, 1496–1505 (2011).
54. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects inmicroarray expression data using empirical Bayes methods. Biostatistics 8, 118–127 (2007).
55. Salas, L. A. et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 19, 64 (2018).
56. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform. 13, 86 (2012).
57. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015).
58. Zou, H. & Hastie, T. Regularization and variable selection via theelastic net. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 67, 301–320 (2005).
59. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths forgeneralized linear models via coordinate descent. J. Stat.Softw.33, 1–22 (2010).
60. Robin, X. et al. pROC: an open-source package for R and S+ toanalyze and compare ROC curves. BMC Bioinform. 12, 77 (2011).
61. Breeze, C. E. et al. eFORGE v2.0: updated analysis of cell typespecific signal in epigenomic data. Bioinformatics 35, 4767–4769 (2019).
62. Bronte, V. et al. Recommendations for myeloid-derived suppressorcell nomenclatureandcharacterization standards.Nat.Commun.7, 12150 (2016).
63. Gustafson, M. P. et al. Intratumoral CD14+ cells and circulatingCD14+HLA-DRlo/neg monocytes correlate with decreased survival in patients with clear cell renal cell carcinoma. Clin. Cancer Res. 21, 4224–4233 (2015).
64. Meier, R., Nissen, E. & Koestler, D. C. Low variability in the underlying cellular landscape adversely affects the performance of interaction-based approaches for conducting cell-specific analyses of DNA methylation in bulk samples. Statistical applications in genetics and molecular biology, 20, 73–84. https://doi.org/10.1515/ sagmb-2021-0004 (2021).
65. Harris, P. A. et al. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inf. 42, 377–381 (2009).
66. Harris, P. A. et al. The REDCap consortium: Building an internationalcommunity of software platform partners. J. Biomed. Inf. 95, 103208 (2019).

Acknowledgements

Work at University of California, San Francisco was supported by the National Institutes of Health (grant numbers R01CA52689 (M.W., J.K.W.), P50CA097257 (J.K.W., A.M.M., J.C., M.W.),
R01CA126831 (J.K.W.), R01CA139020 (M.W., J.K.W.), and R01CA207360 (J.K.W.), the loglio Collective, the National Brain Tumor Foundation, and by donations from families and friends of John Berardi, Helen Glaser, Elvera Olsen, Raymond E. Cooper, and William Martinusen. J.K.W. is supported by the Robert Magnin Newman Endowed Chair in Neuro-oncology, MW by the Stanley D. Lewis and Virginia S. Lewis Endowment. UCSF Parnassus Flow Core RRID:SCR_018206 and DRC Center Grant NIH P30 DK063720 (S.J.T.). The authors wish to acknowledge study participants, the clinicians, and research staff at the participating medical centers, the UCSF Helen Diller Family Comprehensive Cancer Center Genome Analysis Core, which is supported by a National Cancer Institute Cancer Center Support Grant (5P30CA082103), the UCSF Cancer Registry (for updating UCSF glioma case survival and vital status), and the UCSF Neurosurgery Tissue Bank. This publication was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through UCSF-CTSI Grant Number UL1 RR024131. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The collection of cancer incidence data used in this study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; Centers for Disease Control and Prevention's (CDC) National Program of Cancer Registries, under cooperative agreement 5NU58DP006344; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201800032I awarded to the UCSF, contract HHSN261201800015I awarded to the University of Southern California, and contract HHSN261201800009I awarded to the Public Health Institute, Cancer Registry of Greater California. The ideas and opinions expressed herein are those of the author(s) and do not necessarily reflect the opinions of the State of California, Department of Public Health, the National Cancer Institute, and the Centers for Disease Control and Prevention or their Contractors and Subcontractors. All analyses, interpretations, and conclusions reached in this manuscript from the mortality data are those of the author(s) and not the State of California Department of Public Health. Study data were collected and managed using REDCap electronic data capture tools hosted at the University of California, San Francisco[65,66]. REDCap (Research Electronic Data Capture) is a secure, web-based software platform designed to support data capture for research studies, providing (1) an intuitive interface for validated data capture; (2) audit trails for tracking data manipulation and export procedures; (3) automated export procedures for seamless data downloads to common statistical packages; and (4) procedures for data integration and interoperability with external sources.

Author contributions

Conceptualization: J.K.W., A.M.M. Writing original draft: J.K.W.
Experimental design: A.M.M., J.K.W., D.C.K., L.A.S., H.H., S.J.T., Z.Z., M.W., K.T.K., J.W.T., J.C. Implementation: A.M.M., G.W., L.M., T.R., H.H., E.T., S.J.T., C.M.T., Z.Z., D.C.K., L.A.S., P.B., B.C.C., K.T.K. Analysis and interpretation of the data: J.K.W., A.M.M., G.W., D.C.K., E.N., P.B., H.H., L.A.S., B.C.C., M.W., K.T.K,. E.T. Review and editing: All authors were involved in the writing of the paper and have read and approved the final version.

Competing interests

J.K.W. and K.T.K. are cofoundersof Cellintec, which played no role in the current study. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-33215-x.

CorrespondenceandrequestsformaterialsshouldbeaddressedtoJ.K. Wiencke.

Peer review information Nature Communications thanks Antoinette Perry and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.
Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022