

AWARD NUMBER: W81XWH-20-1-0522

TITLE: Bridging Epigenomics and Patient Data to Detect Enhanceropathies in Cancer

PRINCIPAL INVESTIGATOR: Emily Hodges

CONTRACTING ORGANIZATION: Vanderbilt University, Nashville, TN

REPORT DATE: February 2023

TYPE OF REPORT: Final

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

DISTRIBUTION STATEMENT: Approved for Public Release; Distribution Unlimited

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 February 2023		2. REPORT TYPE Final		3. DATES COVERED 01Jul2020-31Oct2022	
4. TITLE AND SUBTITLE Bridging Epigenomics and Patient Data to Detect Enhanceropathies in Cancer				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER W81XWH-20-1-0522	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Emily Hodges E-Mail: emily.hodges@vanderbilt.edu				5d. PROJECT NUMBER 0011482279-0001	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Vanderbilt University 110 21st Avenue South, Suite 800, Nashville, TN, 37203-2408				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. 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 Unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT The objectives of this proposal are to leverage DNA methylation data and large-scale patient Electronic Health Record (EHR) data to identify and characterize novel, non-coding genetic variants associated with lymphoid and other blood cancers. During the two years of this award, we made substantial progress towards the stated goals of the project, despite the early delays brought on by COVID19. Specifically, we identified new DNA methylation-defined lymphocyte enhancers that regulate key lymphocyte genes and contain non-coding single nucleotide polymorphisms (SNPs) associated with risk for immune disease phenotypes including cancer. We have developed new protocols for functionally analyzing non-coding SNPs and their effects on enhancer function using ATAC-STARR-seq, a massively parallel reporter assay that simultaneously measures chromatin accessibility, transcription factor occupancy and regulatory activity. Integrating these data allowed us to prioritize a high value set of lymphocyte enhancers for further mechanistic studies. Finally, we demonstrate how DNA methylation data can be combined with genotype-phenotype data to understand the function of gene regulatory elements and the genes they target. Our work provides a roadmap for using epigenetic data to bridge the genome-phenome knowledge gap in cancer.					
15. SUBJECT TERMS Electronic health records, DNA methylation, Non-coding SNPs					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Unclassified	18. NUMBER OF PAGES 80	19a. NAME OF RESPONSIBLE PERSON USAMRDC
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (include area code)

TABLE OF CONTENTS

	<u>Page</u>
1. Introduction	4
2. Keywords	4
3. Accomplishments	4
4. Impact	9
5. Changes/Problems	10
6. Products	10
7. Participants & Other Collaborating Organizations	11
8. Special Reporting Requirements	12
9. Appendices	12

1. INTRODUCTION:

Genome-wide association studies have reported numerous non-coding single nucleotide polymorphisms (SNPs) that are associated with lymphoid leukemia. These ncSNPs are often found in and around enhancers of genes that control immune cell development and function. Enhanceropathies, diseases associated with germline variants within transcriptional enhancers, are driven by perturbation or co-option of gene regulatory circuitries. There is a substantial lack of approaches that identify enhanceropathies in cancer and understanding mechanistic links between ncSNPs and associated cancer phenotypes remains challenging. The objective of this proposal is to use DNA methylome and patient data to determine cell-type specific gene regulatory patterns that pinpoint SNP contributions to disease risk. Interpreting ncSNPs is challenging because genetically driven gene expression changes are context-dependent, with many only exhibiting disease-relevant effects in specific cell-types. Studying DNA methylation in normal blood cells, our lab has shown that cell-type specific DNA hypomethylation (hypomethylated regions, HMRs) identifies cell-type specific gene enhancers. Current work from our lab also shows that immune cell specific HMRs are enriched for reported SNPs linked to immunological traits, including cancers. Consequently, we hypothesized that immune cell specific HMRs are likely to harbor previously unidentified SNPs associated with lymphoid and other leukemias. Thus, we proposed to leverage DNA methylation data to identify “onco-enhanceropathies” in immune cells using genotyping and patient data from the electronic health record (EHR). We further proposed to mechanistically interrogate genotype effects on enhancer function for ncSNPs associated with lymphoid cancers using large-scale, sequencing-based reporter assays and CRISPR-Cas9 genome editing.

2. KEYWORDS:

DNA methylation, Non-coding Single Nucleotide Polymorphism, Enhanceropathies, Massively Parallel Reporter Assays, Lymphoid Cancers, Electronic Health Record, PheWAS, CRISPR-Cas9

3. ACCOMPLISHMENTS:

What were the major goals of the project?

The following tasks were included in Years 1-2 of the Statement of Work.

Major Task 1: Identify lymphoid “onco-enhanceropathies” in B cell enhancers using genotyping and patient data	Month	% Complete
Collect whole genome bisulfite sequencing (WGBS) datasets	1	100
Determine enhancer HMRs from existing WGBS	1-3	100
Perform comparative analysis to isolate lymphoid enhancer HMRs	1-3	100
Apply for access to Vanderbilt EHR (BioVU) data	1	100
QC genotype data and impute LD (linkage disequilibrium) SNPs	1-3	100
Prioritize SNPs for PheWAS	3	100
Perform PheWAS for limited set of immune related ICD10 patient codes	3-6	100
Widen association tests to any PheWAS categories that are related to hematopoietic disorders and blood cancers (69 categories)	3-6	100
Perform test for phenotype enrichment associated with HMR variants	6-9	100
Validate and Perform replication studies in UK Biobank cohort	6-9	50
Prioritize HMRs for functional studies with orthogonal data	9-12	100

<i>Milestone Achieved: significant set of HMR variants associated with blood cancers identified and validated, methodology paper published</i>	12	75
Major Task 2: Mechanistically interrogate genotype effects on enhancer function for ncSNPs associated with lymphoid cancers.		
Design oligos for cloning HMR regions into reporter constructs	9-12	100
Obtain DNA from BioVU or genotype matched HapMap samples	9-12	100
Clone HMR fragments and sequence verify clones	12-18	50
Perform reporter assays	12-18	75
Design and order CRISPR guide RNAs for select HMRs	9-12	75
Assemble RNPs and nucleofect LCLs	12-18	25
Sequence-validate insertions/deletions	12-18	0
Perform RNA-seq and analyze data	18-24	0
<i>Milestone Achieved: Report findings from overall study</i>	18-24	50

What was accomplished under these goals?

Major Activities:

Major Task 1: Identify lymphoid “onco-enhanceropathies” in B cell enhancers using genotyping and patient data

A major objective of this proposal was to use DNA methylation data and patient genotype-phenotype data to uncover gene regulatory elements that contribute to cancer risk. During the first year of this award, we focused on the original goals described in the proposal and associated Statement of Work. Specifically, we performed comparative methylation profiling across an extensive set of cell types to identify cell-type specific hypomethylated regions (HMRs) in normal cellular contexts. Using this approach, we defined cell-type and lineage specific gene regulatory regions in hematopoietic cell genomes, including B and T cells. These activities provided the framework for executing the specific objectives detailed below.

Specific Objectives and Key Outcomes:

Collect whole genome bisulfite sequencing (WGBS) datasets:

We collected WGBS data for diverse cell types and tissues representing distinct organ systems and developmental states. These datasets included H1 embryonic stem cells (H1ESCs), fetal spinal cord, fetal heart, liver, adrenal gland, hematopoietic stem and progenitor cells (HSPCs), B cells, T cells and macrophages.

Determine enhancer hypomethylated regions from existing WGBS:

For each dataset, we determined a high-quality set of HMRs across the genome using a computational model devised to detect HMRs based on clustering of unmethylated CpG sites, the major substrate of DNA methyltransferases. Specifically, HMRs were defined using a 2-state hidden Markov model (HMM) with Beta-Binomial emission distributions to determine both high and low methylation levels.

Perform comparative analysis to isolate lymphoid enhancer HMRs:

We identified HMRs across all nine cell-types, merging overlapping HMRs to create a universal set of 102,390 regions. For each dataset, we measured methylation levels for all 102,390 HMRs. Unsupervised clustering of 102,390 distinct HMRs revealed that levels of HMR specificity across cell-types reflects a developmental hierarchy supported by enrichment of stage-specific transcription

factors (TFs) and gene ontologies (**Figure 1**). This produced a set of approximately ~20k lymphoid specific HMRs (shared between B and T cells), of which ~6k were exclusive to B cells. These regions were enriched for functional genomic annotations associated with enhancers. This work is described in our new pre-print, Scott et al. *BioRxiv*, 2022 (see Products and Appendices).

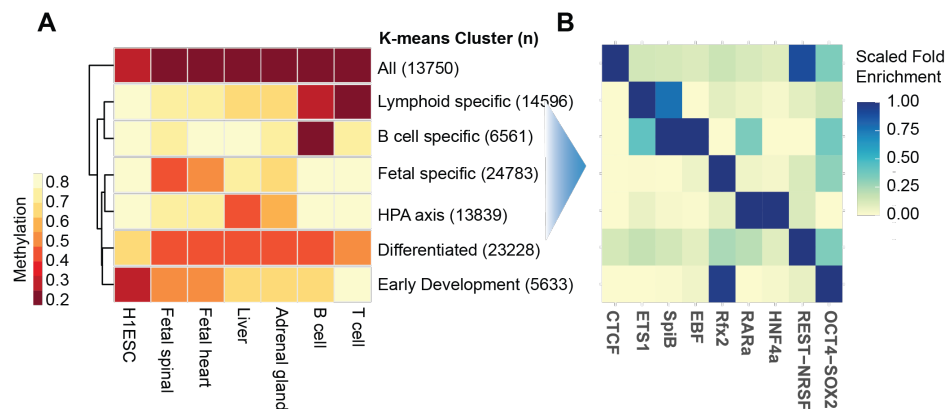


Figure 1. HMRs reconstruct developmental relationships. **A.** Non-coding HMRs were k-means clustered based on their average CpG methylation values across 7 cell types and tissues. The k-means generated HMRs groups that are consistent with the biological relationships of their cell types. **B.** The TF motif enrichment of each k-means group reflects biological relationships captured in (A). Representative TFs were selected from top hits filtered by Benjamini-Hochberg adjusted p-value. Fold enrichment values are normalized from 0 to 1 across TFs. The background comparison file comprises HMRs across all represented cell types. HPA=hypothalamus-pituitary-adrenal axis.

Prioritize SNPs and perform PheWAS:

We performed a phenome-wide association study (PheWAS) and a lab-wide association scan (LabWAS) for selected genotyped SNPs located within or proximal to B cell specific HMRs. A total of 94,474 patient DNAs from Vanderbilt's BioVU repository were genotyped on the Illumina MEAGEx array, and an additional 5,500 were genotyped on the Illumina OMNI-Quad array – covering more than 2 million individual SNP markers. Selected SNPs were tested for associations with 1,814 phenotype codes and 487 clinical laboratory measurements (e.g., cholesterol levels, white blood cells, glucose, etc.) using multivariable logistic or linear regression models, respectively. Our approach revealed previously uncharacterized, high-ranking enhancer HMRs with suggested links to immune phenotypes or lab values. We are currently replicating these findings with independent, larger genetic cohorts with better power for specific phenotypes.

Prioritize HMRs for functional studies with orthogonal data:

To provide an example of how HMRs were selected for mechanistic studies, our analysis revealed a cluster of B cell specific HMRs that harbor SNPs (e.g., rs7976678) significantly linked to multiple immune cell traits including CD8/CD3/CD45 ratios ($p=4.79 \times 10^{-10}$) and acute lymphoid leukemia ($p=3.54 \times 10^{-4}$). In a multi-ethnic GWAS of 15 hematological traits in 746,667 individuals, rs7976678 was also significantly associated with lymphocyte counts ($p=8.1 \times 10^{-6}$, Chen MH et al, *Cell*, 2020).

Despite CD27 being its nearest neighbor, rs7976678 is an eQTL for three other genes: TAPBPL ($p=0.00043$), ZNF384 ($p=0.000017$), and LTBR located ~40 kb upstream of the CD27 locus (**Figures 2-3**). ZNF384 itself is frequently mutated in acute lymphoblastic leukemia, and LTBR is involved in lymphotoxin signaling and lymphoid organ development. Together these data provide convergent evidence that this B cell HMR regulates ZNF384 and LTBR, and that the HMR is significant for normal B cell development function.

We have prioritized this locus for downstream functional dissections and are building tools for this analysis as described in Aim 2 of our original proposal. These tools include a) generation of luciferase reporter constructs to test the functional consequences of genetic variants on enhancer regulatory activity, and b) the selection of guide RNAs for CRISPR-Cas9-mediated deletions of individual HMRs within the locus to test the effect on transcription of associated genes (e.g., TAPBPL, ZNF384 and LTBR).

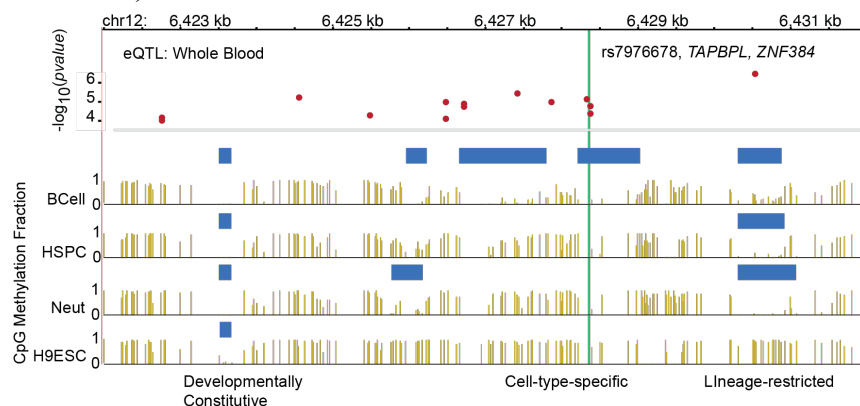


Figure 2. Genome browser tracks of whole genome bisulfite sequencing data and HMRs. Methylation fractions for individual CpG sites are displayed (vertical yellow lines) for three hematopoietic cell types (B cells, Neutrophils, and hematopoietic stem and progenitor cells) and H9 embryonic stem cells. Genomic intervals representing HMRs are shown in blue boxes. HMRs falling into different “sharedness” categories are labeled below methylation tracks. Locations of eQTLs (red dots) identified in a single tissue (whole blood) are provided (GTEx version 8). eQTLs cluster over cell-type specific HMRs. P-values indicate significance of association with transcript variance. Green line highlights location of rs7976678.

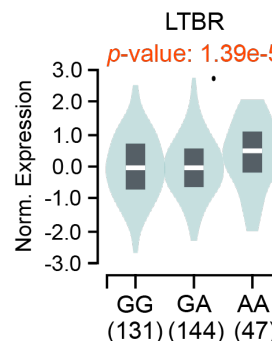


Figure 3. LTBR eQTL from whole blood. Boxplot depicts rank normalized gene expression levels of individuals homozygous (G/G), heterozygous (A/G), or homozygous for the minor allele (A/A, MAF = 0.35). Data were obtained from GTEx.

Milestones Achieved:

We completed a manuscript that describes the use of comparative methylation profiling to identify cell-type specific hypomethylated regions in B cells, as well as other blood cell types, with detailed characterization of these regions including patterns of genetic heritability within these regions (Scott et al. *BioRxiv*, 2022). This work demonstrates how combining DNA methylation with patient genotype and EHR data can identify risk loci for multiple immune traits (see Appendices). We have another manuscript *in prep* describing our mechanistic characterization of the TAPBPL, ZNF384 and LTBR enhancer HMR locus and its link to lymphocyte development and cancer (see Products).

Major Task 2: Mechanistically interrogate genotype effects on enhancer function for ncSNPs associated with lymphoid cancers.

Major Activities:

In Aim 2 we proposed to mechanistically interrogate a set of lead enhancer HMRs identified in Aim 1. Our objective was to validate and characterize the transcriptional activity of our cell-type restricted sets of B cell enhancer HMRs. We hypothesized that cell-type specific HMRs contain core sequence variants that drive differential enhancer activity.

Specific Objectives and Key Outcomes:

To probe the functional activities of HMR-defined genomic sequences, we developed a powerful, multi-omic approach to isolate “driver” HMRs in lymphoblastoid cells. Our ATAC-STARR-seq approach enables simultaneous profiling of chromatin accessibility, transcription factor occupancy, and regulatory activity all from a single DNA fragment library (Hansen and Hodges, *Genome*

Research, 2022). ATAC-STARR-seq uses transposase-assisted cloning to pre-select accessible DNA from native chromatin. This is important for two reasons: 1) we narrow the scope from 100% of the genome to just ~2%, which allows for much deeper sequencing coverage of HMRs, and 2) we only assay HMRs capable of driving transcription endogenously, because they are derived from open chromatin. Together, our approach takes advantage of STARR-seq's unbiased cloning procedure and dynamic range while leveraging ATAC-seq to enrich for cell type specific *cis*-regulatory elements, allowing integrated functional assessment and prioritization of “driver” HMRs.

Summary from the manuscript abstract: We tested ~50 million unique DNA fragments tiling ~101,000 accessible chromatin regions in human lymphoblastoid cells. We discovered that 30% of all accessible regions contain an activator, a silencer or both. Although few MPRA studies have explored silencing activity, we demonstrate silencers occur at similar frequencies to activators, and they represent a distinct functional group enriched for unique transcription factor motifs and repressive histone modifications. We further show that Tn5 cut-site frequencies are retained in the ATAC-STARR plasmid library compared to standard ATAC-seq, enabling TF occupancy to be ascertained from ATAC-STARR data. With this approach, we found that activators and silencers cluster by distinct TF footprint combinations and these groups of activity represent different gene regulatory networks of immune cell function. Altogether, these data highlight the multi-layered capabilities of ATAC-STARR-seq to comprehensively investigate the regulatory landscape of the human genome all from a single DNA fragment source.

ATAC-STARR-seq applied to HMRs (Perform reporter assays):

In our new manuscript (Scott et al.), we comprehensively characterize HMR relationships both within and between developmentally diverse cell types to understand the functional significance of complex HMR patterns. Focusing primarily on cell types within the hematopoietic lineage, we show that levels of HMR specificity across cell-types captures time-point specific branchpoints of development. Our detailed analysis in B cells further reveals that HMRs form clusters in proximity to B cell identity genes. We measured the proportion of B cell HMRs that contain an activator or silencer (**Figure 4**). We show that HMRs in general contain a substantial portion of activators or silencers, and that clustered HMRs contain a significantly higher proportion of both activators and silencers ($p = 2.39 \times 10^{-13}$ and 0.0106 , respectively), than unclustered HMRs. This is a critical observation considering the need to understand how SNPs effect the transcriptional output of enhancer HMRs.

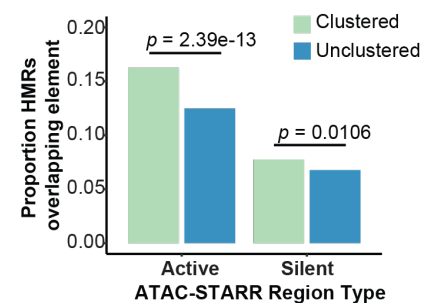


Figure 4. Proportion HMRs containing an activator or silencer. HMRs are categorized as clustered (3+ HMRs within 6kb of each other) or unclustered.

Milestones Achieved:

We developed and published a manuscript describing ATAC-STARR-seq in lymphoblastoid cells. We used data generated by this study to evaluate the transcriptional competency of HMR patterns. This data is described in our new pre-print (Scott et al., *BioRxiv*, 2022).

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

The graduate students listed in the Statement of Work, Timothy Scott and Tyler Hansen, present their work related to this project and receive feedback on a monthly basis during Hodges lab meetings. Mr. Scott and Mr. Hansen have presented their research to the broader Vanderbilt community at the Genetics Interest Group, Biochemistry Student Association and Stem Cell and Progenitor Interest

Group meetings. Mr. Hansen presented a poster describing ATAC-STARR-seq at the Cold Spring Harbor Systems Biology: Global Regulation of Gene Expression meeting which took place March 9 - March 12, 2022. Graduate student Timothy Scott presented a poster and a poster talk at the American Society of Human Genetics Annual meeting in October 25-29th, 2022. His poster received the “Reviewers Choice” award and was selected for a highly competitive poster talk (see Appendices). Both students have prepared manuscripts for publication (see Products and Appendices) and have had opportunities to mentor rotation students throughout the year. Additionally, the project has provided opportunities to expand their training in genomic and genetic data analysis – both Mr. Scott and Mr. Hansen have sought online workshops and tutorials, including those available through our ongoing lab subscription to DataCamp.

During the first year of this award, there were limited opportunities to attend international meetings due to COVID. However, Mr. Hansen presented a poster at the Cold Spring Harbor Virtual Epigenetics and Chromatin Meeting in September of 2020. Mr. Scott attended the American Society for Human Genetics Virtual meeting held in October of 2020.

How were the results disseminated to communities of interest?

Over the last two years, Dr. Hodges has presented data related to this project to the Vanderbilt community through invited seminars as well as seminars held within the Department of Biochemistry and Vanderbilt Ingram Cancer Center. A manuscript describing the ATAC-STARR-seq method, a novel reporter assay developed as part of aim 2, was deposited in *BioRxiv* (bioRxiv.2022.03.25.485870) and has been published as an open access research article in *Genome Research*. We finalized and submitted a manuscript describing the functional and genetic characterization of lymphoid specific HMRs (see products and appendices). This manuscript is currently under revision and has been posted on *BioRxiv* (bioRxiv 2022.12.15.520535).

As previously reported, a poster presentation describing the results of these studies was presented by graduate student Tyler Hansen at the Cold Spring Harbor Systems Biology: Global Regulation of Gene Expression meeting which took place March 9 - March 12, 2022. Graduate student Timothy Scott presented a poster and a poster talk at the American Society of Human Genetics Annual meeting in October 25-29th, 2022.

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

Nothing to Report

4. IMPACT:

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

These studies establish a fundamentally new way to understand how DNA methylation bridges the connection between genomes and phenomes, revealing important gene regulatory principles that are essential to understanding why epigenetic instability leads to specific disease outcomes.

What was the impact on other disciplines?

The new approaches developed to execute this project - including 1) ATAC-STARR-seq, and 2) combining DNA methylation with patient genotype and EHR data to identify cancer risk loci - are broadly applicable to many different types of cancer. Additionally, these methods provide a novel route to identify enhancer-gene relationships, which will be of broad interest to the gene regulation field.

What was the impact on technology transfer?

Nothing to Report

What was the impact on society beyond science and technology?

The goals of this work are to establish methods for detecting and characterizing new, non-coding genetic markers associated with lymphoma and other blood cell cancers. By identifying new genetic risk mutations associated with lymphoma and blood cell cancers, this proposal addresses military relevant environmental risk factors associated with cancer (e.g., ionizing radiation, chemicals, infectious agents, environmental carcinogens, and stress). This approach will be broadly applicable to identify previously undetected gene regulatory contributions, not only to lymphoma and blood cell cancers, but other types of cancer as well. Ultimately, we expect our approach to accelerate the discovery and characterization of gene regulatory contributions to the molecular pathology of lymphoma and leukemia, uncovering new druggable gene targets, leading to a wider range of potential treatment options and better outcomes for patients.

5. CHANGES/PROBLEMS:

Nothing to Report

Actual or anticipated problems or delays and actions or plans to resolve them

Nothing to Report

Changes that had a significant impact on expenditures

Nothing to Report

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

Nothing to Report

6. PRODUCTS:

- **Publications, conference papers, and presentations**

Journal publications

T. Hansen and E. Hodges; ATAC-STARR-seq reveals transcription factor bound activators and silencers across the chromatin accessible human genome; *Genome Research*; 2022; *published*; yes

Scott T, Hansen TJ, McArthur E, and Hodges E. Cross-tissue patterns of DNA hypomethylation reveal genetically distinct histories of cell development. *bioRxiv* 2022.12.15.520535; 2022; *Under revision*; yes

Hansen TJ, Scott, TJ, Hodges E. Identification of a novel B cell enhancer driving risk for lymphoid leukemia; 2023; *In prep*; yes

Other publications, conference papers and presentations

Scott T, Hansen TJ, McArthur E, Hodges E. Cross-tissue patterns of DNA hypomethylation reveal genetically distinct histories of cell development. American Society of Human Genetics Annual Meeting; October 2022; yes (see appendix for conference abstract)

- **Website(s) or other Internet site(s)**

Scripts for ATAC-STARR data analysis are posted on GitHub: <https://github.com/HodgesGenomicsLab/ATAC-STARR-seq>.

An interactive version of the ATAC-STARR protocol is posted on protocols.io: [dx.doi.org/10.17504/protocols.io.bggijtue](https://doi.org/10.17504/protocols.io.bggijtue)

- **Technologies or techniques**

Protocols for ATAC-STARR-seq has been made publicly available on web portals mentioned above and have been submitted to *BioRxiv* (bioRxiv. 2022.03.25.485870). The ATAC-STARR paper has also been made “open access” as an advanced online publication in *Genome Research*.

- **Inventions, patent applications, and/or licenses**

Nothing to Report

- **Other Products**

Data resulting from Hansen and Hodges, *Genome Research* 2022, has been deposited in the Gene expression omnibus under accession number GSE181317.

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name: Emily Hodges

Project Role: PD/PI

Researcher Identifier (ORCID ID): 0000-0001-6513-610X

Nearest person month worked NCE: 1

Contribution to Project: Dr. Hodges is the project PI and supervised all aims pertaining to this award.

Funding Support:

Name: Timothy Scott

Project Role: Graduate Student

Researcher Identifier (ORCID ID): 0000-0002-9454-5700

Nearest person month worked NCE: 1

Contribution to Project: Tim has coordinated the project, aided in manuscript preparation, and provided essential effort in running genetic association studies as well as analysis and integration of largescale functional genomic datasets including whole genome bisulfite sequencing.

Funding Support:

Name: Tyler Hansen

Project Role: Graduate Student

Researcher Identifier (ORCID ID): 0000-0003-3328-783X

Nearest person month worked NCE: 1

Contribution to Project: Tyler performed functional genomics assays and next-generation sequencing analysis. He developed and performed the massively parallel reporter assays and experiments described in aim 2 of this proposal.

Funding Support:

Name: Anthony Capra

Project Role: Collaborator

Researcher Identifier (ORCID ID): 0000-0001-9743-1795

Nearest person month worked NCE: 1

Contribution to Project: Dr. Capra provided expertise and supervision of the integration of electronic health record and other phenotypic data from the Vanderbilt biobank with genetic analyses.

Funding Support:

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Changes in active support and completed support for all key personnel since the JIT are annotated as an appendix to this document.

What other organizations were involved as partners?

Nothing to Report

8. SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS: *Not applicable*

QUAD CHARTS: *Not applicable*

9. APPENDICES: *The following items are attached for clarification.*

1. Annotated changes to Other Support for all key personnel.
2. Scott T, Hansen TJ, McArthur E, Hodges E. Cross-tissue patterns of DNA hypomethylation reveal genetically distinct histories of cell development. **American Society of Human Genetics Annual Meeting**; October 2022.
3. Scott T, Hansen TJ, Dorans E, McArthur E, Singh K and Hodges E. Cross-tissue patterns of DNA hypomethylation reveal genetically distinct histories of cell development. **bioRxiv** 2022.12.15.520535; doi: <https://doi.org/10.1101/2022.12.15.520535>. *Under revision.*

Name of Individual: Hodges, Emily C.

Positions/Scientific Appointments

Assistant Professor, Department of Biochemistry, Vanderbilt University School of Medicine

Other Support – W81XWH-20-1-0522

ACTIVE:

*Title: The DNA methylome-based regulation of functional beta-cell mass

Major Goals: The goal of this study is to investigate how modulating the DNA methylome in pancreatic progenitors impact postnatal functional beta-cell mass. The significance is that humans need to maintain certain levels of this mass to avoid debilitating hypoglycemia or diabetes. We expect to uncover basic molecular and cellular mechanisms to help improve the regulation of this mass to avoid these diseases.

*Status of Support: Active - **New**

Project Number: **5R01 DK125696-03**

Name of PD/PI: Gu, Guoqiang

*Source of Support: NIH/NIDDK

*Primary Place of Performance: Vanderbilt University

Role: Collaborator

Procuring Contracting/Grants Officer and Address: Albert J. Hwa albert.hwa@nih.gov

Specific Aims: 1) Fitness of M+N+ vs. M-N+ progenitor-derived β -cell subsets under metabolic stress; 2) Molecular underpinning for different fitness of M+N+ and M-N+ progenitor-derived β cells; 3) Establish how maternal low-protein diet (LPD) affects postnatal functional β -cell mass

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 07/2020-06/2025

Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
1. 2021	1.20 Calendar
2. 2022	3.00 Calendar
3. 2023	1.20 Calendar
4. 2024	1.20 Calendar
5. 2025	1.20 Calendar

*Title: The Evolution of Gene Regulation and Human Disease

Major Goals: The objectives of this proposal are to develop computational models that leverage machine learning, evolutionary patterns, and large-scale clinical biobanks to identify mutations that disrupt proper control of genes and lead to disease.

*Status of Support: Active

Project Number: **5R35 GM127087-06**

Name of PD/PI: Capra, John A.

*Source of Support: NIH/NIGMS (62423 UCSF)

*Primary Place of Performance: Vanderbilt University

Role: Collaborator

Procuring Contracting/Grants Officer and Address: Albert J. Hwa albert.hwa@nih.gov

Specific Aims: N/A

Statement of Potential Overlap: None

Name of Individual: Hodges, Emily C.

Project/Proposal Start and End Date: (MM/YYYY) (if available): 09/2020–03/2023

* Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
4. 2021	0.21 Calendar
5. 2022	0.36 Calendar
6. 2023	0.36 Calendar

*Title: Decoding genome function with DNA methylation and human phenome data

Major Goals: Despite having near-identical genome sequences, all human cell-types have unique DNA methylation profiles that reflect their individual programs of gene regulatory activity. This research will leverage human population genetics to link DNA methylation patterns to traits recorded in the electronic health record and will couple these data with mechanistic studies to uncover the relationship between DNA methylation, the genome sequence and phenotypic outcomes. These studies will reveal how DNA methylation mediates genome function to establish normal and pathological cellular identities.

*Status of Support: Active - **New**

Project Number: **1R01 GM147078-01**

Name of PD/PI: Hodges, Emily C.

*Source of Support: NIH/NIGMS

*Primary Place of Performance: Vanderbilt University

Role: PI

Procuring Contracting/Grants Officer and Address: Anthony D. Carter cartera@nigms.nih.gov

Specific Aims: 1) Identify cell-type essential HMRs with human phenome data; 2) Define the transcriptional potential of DNA methylation-defined genomic sequences.

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 08/2022-05/2027

Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
1. 2023	5.80 Calendar
2. 2024	4.00 Calendar
3. 2025	4.00 Calendar
4. 2026	4.00 Calendar
5. 2027	4.00 Calendar

PENDING:

*Title: Multiomics at Vanderbilt to Enhance Novel Discovery (MAVEN)

Major Goals: As part of the MAVEN Discovery Center, the Hodges lab will provide expertise in generating functional genomic data to characterize epigenetic profiles across disease states. We will lend expertise and provide training in various epigenomic techniques including ATAC-seq, whole genome bisulfite sequencing, ChIP-seq and Cut&Tag. The Hodges lab will also participate in data generation and bioinformatic data analysis and assist in interpretation of results. Dr. Hodges will also work with the MAVEN team to provide overall project oversight and integration of aims, to facilitate communication between all personnel engaged in this project, and

Name of Individual: Hodges, Emily C.

to report results and progress to both NIH and to the broader community through publications and speaking engagements.

*Status of Support: Pending

Project Number: **1U01 HG-01**

Name of PD/PI: Cox, Nancy

*Source of Support: NIH/NHGRI (VUMC / ISA)

*Primary Place of Performance: Vanderbilt University

Role: Co-Investigator

Procuring Contracting/Grants Officer and Address: Joannella Morales joannella.morales@nih.gov

Specific Aims: 1: To apply state-of-the-art, high-throughput methodologies to quantify multiple domains of molecular phenotypes in circulation and tissue across disease states, focusing on best practices to reduce variability and improve reproducibility; 2: To use a multi-layer integration approach across phenotypes and 'omics; 3: To establish standard workflows that span the entire lifespan of multi-omic analyses, from specimen collection, data generation, processing, analysis, and visualization.

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 07/2023-06/2028

Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
1. 2024	0.60 Calendar
2. 2025	0.60 Calendar
3. 2026	0.60 Calendar
4. 2027	0.60 Calendar
5. 2028	0.60 Calendar

*Title: The Role of Mutated Immune Cells in Human Clonal Hematopoiesis

Major Goals: The objectives of this proposal are to elucidate the impact of TET2-deficiency on blood cell production, to understand its impact on methylation patterns in specific developmental stages of hematopoiesis and to define and target the link between inflammatory signaling, TET2 and clonal hematopoiesis. The Hodges lab will aid in generating functional genomic data to characterize epigenetic mechanisms underlying TET2-deficient human cells lines and TET2-deficient murine-derived cells.

*Status of Support: Pending

Project Number: **1R01 -01**

Name of PD/PI: Ferrell, P. Brent

*Source of Support: NIH/Unknown (VUMC)

*Primary Place of Performance: Vanderbilt University

Role: Co-Investigator

Procuring Contracting/Grants Officer and Address: Not yet assigned

Specific Aims: Aim 1: Identify and dissect mutated and non-mutated cells in human TET2-mutant CHIP; Aim 2: Investigate altered signaling and function in monocytes from TET2-deficient CHIP; Aim 3: Investigate the role of epigenetic reprogramming in TET2-mutant related monocyte activation in CHIP.

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 12/2023-11/2028

Person Months (Calendar/Academic/Summer) per budget period.

Name of Individual: Hodges, Emily C.

Year Ending (YYYY)	Person Months (##.##)
1. 2024	0.60 Calendar
2. 2025	0.60 Calendar
3. 2026	0.60 Calendar
4. 2027	0.60 Calendar
5. 2028	0.60 Calendar

COMPLETED (last 5 years):

Title: Dissecting Genetic and Epigenetic variation in the Cancer Regulome

Goals: These approaches will establish an essential foundation to study the contribution of gene regulatory variation to cancer development and progression. The proposed research will complement current efforts to better understand the molecular basis of cancer and may provide alternative routes to identify cancer susceptibility loci and targets for drug intervention.

Agency Number: 5K22 CA184308-03

Principal Investigator: Emily Hodges, Ph.D.

Time Commitments/Role: 75%, 9.00 Calendar Months/PI

Supporting Agency: NIH/NCI

Procuring Contracting/Grants Officer: Sonia B. Jakowlew jakowles@mail.nih.gov

Performance Period: 01/01/2015 – 05/31/2018

Level of Funding:

Specific Aims:

- 1) Identify cell-type specific regulatory (enhancer) regions through comparative methylation analysis.
- 2) Develop a “regulome” capture method to study genotype and epitype variation at enhancers.
- 3) Perform reverse genetics screens measuring the effect of regulatory variation on enhancer activity.
- 4) Test the competence of patient-derived regulatory DNA to drive transcription.

*Title: Inhibition of DNA methyltransferase within Somatic Mutations and the Effect on Clinical Response

Major Goals: The Hodges lab will help select cases, prepare samples and collaborate in the interpretation of whole genome methylation sequencing done on the patient samples.

*Status of Support: Completed

Project Number: VUMC53996

Name of PD/PI: Savona, Michael

Time Commitments/Role: 10%, 1.20 Calendar Months / PI

*Source of Support: Astex Pharmaceuticals, Inc. via Consortium from VUMC (UNIV59445/VUMC61152)

Procuring Contracting/Grants Officer: Astex Pharmaceuticals, Inc., VP-Corporate Counsel, 4220 Rosewood Drive, Suite 200, Pleasanton, CA, 94404

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date: MM/YYYYY (if available): 03/2017-06/2019 (NCE)

Level of Funding:

Specific Aims: Not Applicable to this Contract

*Title: Phenotype Heterogeneity and Dynamics in SCLC – Pilot: Linking epigenomics and large-scale biobanks to identify enhanceropathies in lung cancer

Major Goals: The objective of this Pilot Project is to determine whether we can identify onco-enhanceropathies in DNA methylation defined lung enhancers using high-density genotyping and patient data.

*Status of Support: Completed

Name of Individual: Hodges, Emily C.

Project Number: 5U54 CA217450-02-Pilot

Name of PD/PI: Emily Hodges (Prime PI Vito Quaranta)

Time Commitments/Role: 8.33%, 1.00 Calendar Months / Pilot Project PI

*Source of Support: NIH/NCI

Procuring Contracting/Grants Officer: Shannon K. Hughes shannon.hughes@nih.gov

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date: MM/YYYY (if available): 07/2019-03/2020

Level of Funding:

Specific Aims:

1) Identify “onco-enhanceropathies” in SCLC enhancers using genotyping and patient data.

*Title: Pilot - Bridging Epigenomic and Patient Data to Detect Enhanceropathies in Cancer

Major Goals: The proposed research will address understudied relationships between non-coding DNA and lymphoid cancer by characterizing a growing number of “onco-enhanceropathies”, where disease risk is mediated by malignant changes in gene regulatory activities of normal cell function. If successful, this approach will be broadly applicable to detect genome-phenome relationships in other types of cancer.

*Status of Support: Completed - **Ended**

Project Number: ACS-IRG-15-169-56-Pilot

Name of PD/PI: Emily Hodges (Prime PI Scott Hiebert)

Time Commitments/Role: 1%, 0.12 Calendar Months (No Salary Support) / Pilot Project PI

*Source of Support: American Cancer Society via consortium from VUMC

Procuring Contracting/Grants Officer: William C. Phelps, PhD grants@cancer.org

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date: MM/YYYY (if available): 07/01/2019-06/30/2020

Level of Funding:

Specific Aims:

1) Identify lymphoid “onco-enhanceropathies” in B-cell enhancers using genotyping and patient data.

2) Mechanistically interrogate enhancer regions nominated by ncSNPs associated with lymphoid cancers.

Overlap: There was scientific overlap between the ACS-IRG-15-169-59 Pilot Project and the DoD Idea Award application. This pilot project purpose was to assist with costs for preliminary research studies and ended 06/30/2020 (non-renewable).

*Title: Bridging Epigenomics and Patient Data to Detect Enhanceropathies in Cancer

Major Goals: The objective of this proposal is to identify the genetic determinants of disease susceptibility and treatment response that are frequently found in the non-coding regulatory regions of the genome. We will integrate gene regulation data with large-scale patient genome data and Electronic Medical Record (EMR) data to identify and characterize novel genetic markers of lymphoma and blood cancers.

*Status of Support: Completed – **This Grant Report - Ended**

Project Number: W81XWH-20-1-0522

Name of PD/PI: Hodges, Emily

*Source of Support: Department of Defense

*Primary Place of Performance: Vanderbilt University

Role: PD/PI

Procuring Contracting/Grants Officer and Address: Jamie A. Shortall Jamie.a.shortall.civ@mail.mil

Specific Aims: 1) Identify lymphoid “onco-enhanceropathies” in iHMR-defined B cell enhancers using genotyping and patient data.

2) Mechanistically interrogate enhancer regions nominated by ncSNPs associated with lymphoid cancers.

Statement of Potential Overlap: None

Name of Individual: Hodges, Emily C.

Project/Proposal Start and End Date: (MM/YYYY) (if available): 07/2020-10/2022 (NCE)

* Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
1. 2021	3.00 Calendar
2. 2022	3.00 Calendar
2. 2022	1.00 Calendar

IN-KIND:

*Summary of In-Kind Contribution: Translational Research - This In-Kind contribution is a supplement to start-up funds provided by the VUMC Cancer Center

Overall Objectives: In FY2020, Vanderbilt University established the Translational Research-Emily Hodges-Biochemistry fund as the result of a gift from Vanderbilt University Medical Center. The fund will support translational research under the direction of Emily Hodges, Ph.D. in the Department of Biochemistry at Vanderbilt University.

Statement of Potential Overlap: None

*Source of Support: Vanderbilt University Medical Center

*Status of Support: Active

Project Number: **GET_300042**

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date: (MM/YYYY) (if available): 02/2021–Continuous

*Person Months (Calendar/Academic/Summer) per budget period.

Period:	Person Months (##.##)
2022	0.66 Calendar
2023	0.00 Calendar
2024	0.00 Calendar

*Summary of In-Kind Contribution: Stanley Cohen Innovation Fund - decoding DNA methylation during cellular differentiation

Overall Objectives: Hodges' project will provide a deeper understanding of DNA methylation by investigating its presence throughout different stages of cellular differentiation and different regions of the genome. Her lab will monitor methylation patterns on chromatin, along with the accessibility of the chromatin, under conditions of dynamic gene regulation in both single cells and across different cellular systems. Through this work, they hope to define the role of DNA methylation in determining cellular identity and function.

Statement of Potential Overlap: None

*Source of Support: Vanderbilt University

*Status of Support: Active

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date (MM/YYYY) (if available): 11/2022-10/2023

*

Name of Individual: Hiebert, Scott W.

Positions/Scientific Appointments

Member, Frederick National Laboratory Advisory Board
Member, National Cancer Advisory Board
Chair, Scientific Advisory Board, Edward P. Evans Foundation
Professor, Department of Biochemistry, Vanderbilt University School of Medicine

Other Support – W81XWH-20-1-0522

ACTIVE:

*Title: Establishing that Hdac3 is a therapeutic target in a subset of B cell Lymphoma

Major Goals: This project will directly test if inhibiting this enzyme with a drug is a viable therapeutic strategy for a deadly type of lymphoma that contains a specific mutation. If so, this approach could lead to cures for this lymphoma, which is resistant to normal chemotherapy.

*Status of Support: Active

Project Number: **5R01 CA164605-10**

Name of PD/PI: Hiebert, Scott

*Source of Support: NIH/NCI

*Primary Place of Performance: Vanderbilt University

Role: PI

Procuring Contracting/Grants Officer and Address: Chamelli Jhappan, jhappanc@mail.nih.gov

Specific Aims: 1) Define the function of Hdac3 in FOXO1-mediated transcriptional repression

2) Determine if HDAC3 is a therapeutic target for FOXO1 mutant DLBCL *in vivo*

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 05/2018-04/2023

* Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
6. 2019	2.04 Calendar
7. 2020	2.04 Calendar
8. 2021	2.04 Calendar
9. 2022	2.04 Calendar
10. 2023	2.64 Calendar

*Title: Biochemical and Chemical Training for Cancer Research

Major Goals: This training grant provides support for 4 pre-doctoral and 3 post-doctoral candidates.

*Status of Support: Active

Project Number: **5T32 CA009582-35**

Name of PD/PI: Hiebert, Scott

*Source of Support: NIH/NCI

*Primary Place of Performance: Vanderbilt University

Role: PI

Procuring Contracting/Grants Officer and Address: Susan E. Lim, lims@mail.nih.gov

Specific Aims: 1) To provide integrated training in biochemistry, chemistry, structural biology, and cancer biology which is critical to the molecular understanding of cancer and the development of new approaches to molecularly targeted therapeutics.

Name of Individual: Hiebert, Scott W.

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 06/2018-05/2023

* Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
31. 2019	0.12 Calendar
32. 2020	0.60 Calendar
33. 2021	0.60 Calendar
34. 2022	0.60 Calendar
35. 2023	0.60 Calendar

*Title: SPORE in Breast Cancer

Major Goals: The overall goal of this project is to conduct multidisciplinary, mechanism-based, translational research of the highest possible impact that will contribute meaningfully to measurable progress in breast cancer.

*Status of Support: Active

Project Number: **5P50 CA098131-20**

Name of PD/PI: Pietenpol, Jennifer

*Source of Support: NIH/NCI (VUMC59469 / 58822)

*Primary Place of Performance: Vanderbilt University

Role: DRP Director

Procuring Contracting/Grants Officer and Address: Tamara Walton waltont@mail.nih.gov

Specific Aims: 1) Project 1: Inhibition of phosphatidylinositol-S kinase (PISK) as a strategy to abrogate antiestrogen resistance in breast cancer

2) Project 2: Strategies to improve outcomes for triple negative breast cancer patients involving subtype-specific targeted therapies and genomic discovery

3) Project 3: Mcl-1 inhibitors for the treatment of breast cancer

4) Project 4: The obesity-metabolic biomarker axis and breast cancer risk

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 08/2019 - 07/2024

* Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
17. 2020	0.28 Calendar
18. 2021	0.30 Calendar
19. 2022	0.30 Calendar
20. 2023	0.30 Calendar
21. 2024	0.30 Calendar

*Title: Cancer Center Support Grant

Major Goals: The goal of this grant is to coordinate and integrate the cancer-related activities of Vanderbilt; Conduct, support and enhance cancer research and integrate cancer-related research throughout the university and to coordinate and integrate the care of cancer patients at VUMC and VA.

*Status of Support: Active – **Competitively Renewed**

Project Number: **5P30 CA068485-27**

Name of PD/PI: Park, Ben

Name of Individual: Hiebert, Scott W.

*Source of Support: NIH/NCI (VUMC59496 / 58778)

*Primary Place of Performance: Vanderbilt University

Role: Project Lead

Procuring Contracting/Grants Officer and Address: Sonya Roberson robersos@mail.nih.gov

Specific Aims: 1) To coordinate and integrate the cancer and cancer-related activities of Vanderbilt University Medical Center, The Vanderbilt University, and the Veteran's Administration Medical Center

2) To conduct, support and enhance cancer research and to integrate cancer-related activities throughout these collaborating institutions

3) To integrate, develop and conduct cancer education programs

4) To coordinate and integrate the care of cancer patients at Vanderbilt University Medical Center and the Veteran's Administration Medical Center

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 09/2020-08/2025

* Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
25. 2021	2.52 Calendar
26. 2022	3.12 Calendar
27. 2023	2.52 Calendar
28. 2024	2.52 Calendar
29. 2025	2.52 Calendar

*Title: Role of PRC1 in RUNX1-ETO-mediated transcriptional control

Major Goals: We must understand how a mutation causes cancer before we can develop targeted therapeutics or design drug combinations that eliminate the leukemic cells, leaving normal cells to flourish. This proposal seeks to define how AML1-ETO alters epigenetic circuits to establish a permissive environment that allows secondary mutations to transform these cells into AML. We will focus on the recruitment of HDAC3 and PRC1.5 by AML1-ETO and test how recruitment of these histone-modifying complexes is used to set the stage for the development of AML.

*Status of Support: Active - **New**

Project Number: **5R01 CA255446-02**

Name of PD/PI: Hiebert, Scott

*Source of Support: NIH/NCI

*Primary Place of Performance: Vanderbilt University

Role: PI

Procuring Contracting/Grants Officer and Address: Malgorzata Klauzinska, klauzing@mail.nih.gov

Specific Aims: Aim 1: Define the mechanism of AE transcriptional control using chemical-genetics

A. Define the temporal order of histone modifications and nucleosome accessibility at AE-genes

B. Genetic complementation analysis of AE protein contacts after degradation of endogenous AE

C. Define how degrading AE affects "preleukemic" CD34+ human HSPCs

Aim 2: Determine the contribution of PRC1.5 and HDAC3 to AE-dependent transcription control

A. Define the transcriptional circuits directly regulated by PRC1, and HDAC3 using PROTACs

B. Determine how AML1-ETO cooperates with ASXL1 mutations

C. Identify the effects of inhibiting PRC1 and HDAC3 on primary t(8;21) AML

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 07/2021-06/2026

Name of Individual: Hiebert, Scott W.

* Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
1. 2022	3.00 Calendar
2. 2023	3.00 Calendar
3. 2024	3.00 Calendar
4. 2025	3.00 Calendar
5. 2026	3.00 Calendar

*Title: Epigenetic modulation of immune suppressed mesenchymal triple-negative breast cancer

Major Goals: Validate MHC-I, H3K27me3 and DNA methylation in mesenchymal TNBC patients and association with response to anti-PD-L1 response. Identify PRC2-dependent transcriptional changes. Identification of PRC2 binding sites and H3K27me3 changes with EZH2 inhibition. Identification of PRC2 binding sites and H3K27me3 changes with EZH2 inhibition [CUT&RUN-seq]. Evaluate efficacy of PRC inhibitors in mice.

*Status of Support: Active - **New**

Project Number: **W81XWH-21-10765**

Name of PD/PI: Lehmann, Brian

*Source of Support: Department of Defense (VUMC94472/62450)

*Primary Place of Performance: Vanderbilt University

Role: Collaborator

Procuring Contracting/Grants Officer and Address: Jamie A. Shortall Jamie.a.shortall.civ@mail.mil

Specific Aims: The Specific Aims of this proposal are designed to address the overarching challenges of revolutionizing treatment by replacing standard of care with regimens that are more effective, less toxic and increase survival and eliminating the mortality associated with metastatic breast cancer by restoring immune function through epigenetic modulation.

Specific Aim 1: To determine the mechanism of PRC2-mediated immune suppression in mesenchymal TNBC.

Specific Aim 2: To evaluate the effect of PRC2 inhibition in combination with immune checkpoint therapy or chemotherapy in metastatic mesenchymal TNBC murine models.

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 08/2021-08/2024

Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
1. 2022	0.36 Calendar
2. 2023	0.36 Calendar
3. 2024	0.36 Calendar

*Title: Toward precision therapy for PAX3-FOXO1-dependent rhabdomyosarcoma

Major Goals: Define the therapeutic value of degrading PAX3-FOXO1, inhibiting its direct transcriptional targets, and inhibiting the epigenetic enzymes that it recruits. Use CRISPR screens to identify therapeutic targets to trigger aRMS cell differentiation or death. Given our ability to align these data with the Broad Dependency Map, we are very confident that many of the direct targets and associating factors represent critical dependencies that can be exploited for targeted therapy for rhabdomyosarcoma. We fully expect our analysis to lead to rational clinical trials for one of the hardest to treat childhood cancers.

*Status of Support: Active - **New**

Name of Individual: Hiebert, Scott W.

Project Number: **T2021-005 Pediatric Call (Translational Research)**

Name of PD/PI: Hiebert, Scott

*Source of Support: V Foundation for Cancer Research

*Primary Place of Performance: Vanderbilt University

Role: PI

Procuring Contracting/Grants Officer and Address: Carole C. Wegner, Ph.D. Grants@v.org

Specific Aims: 1) Define the therapeutic value of degrading PAX3-FOXO1, inhibiting its direct transcriptional targets, and inhibiting the epigenetic enzymes that it recruits.

2) Use CRISPR screens to identify therapeutic targets to trigger aRMS cell differentiation or death.

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 11/2021-11/2024

* Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
1. 2022	1.20 Calendar
2. 2023	1.20 Calendar
3. 2024	1.20 Calendar

*Title: Molecular and Cellular Basis for Digestive Disease

Major Goals: The VDDRC, through technologies provided by innovative Research Cores and financial and professional support of junior and seasoned investigators, seeks to improve the prevention, treatment and outcomes of human digestive diseases by serving as a catalyst for collaborative and impactful investigative research.

*Status of Support: Active - **New**

Project Number: **2P30 DK058404-21**

Name of PD/PI: Peek, Richard M.

*Source of Support: NIH/NIDDK (VUMC59165 / ISA00000179)

*Primary Place of Performance: Vanderbilt University

Role: Co-Scientific Director

Procuring Contracting/Grants Officer and Address: Peter J Perrin peter.perrin@nih.hhs.gov

Specific Aims: 1) Promote digestive diseases research in an integrative, collaborative, and multidisciplinary manner

2) Attract new investigators to the study of these disorders

3) Enhance the innovative research capabilities of members

4) Promote the career development of junior investigators

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 07/2022 – 05/2027

* Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
21. 2023	0.30 Calendar
22. 2024	0.30 Calendar
23. 2025	0.30 Calendar
24. 2026	0.30 Calendar
25. 2027	0.30 Calendar

Name of Individual: Hiebert, Scott W.

PENDING:

*Title: Unveiling the cell context dependency of oncogenic pathway activation by p53 DNA contact mutants
Major Goals: Dr. Hiebert will provide expertise in genomic analysis of cells containing TP53 mutations. This analysis will be focused on PRO-seq using short time scales in the presence or absence of TP53 degradation, but may include other approaches such as ATAC-seq. PRO-seq will identify changes in nascent transcription, which represent the direct transcriptional targets of TP53. He will also oversee the informatic analyses of these datasets and work with Dr. Aaronson on the interpretation of these data and the communication and sharing of these data.

*Status of Support: Pending

Project Number: **1R01 CA269927-01A1**

Name of PD/PI: Aaronson, Stuart

*Source of Support: NIH/NCI (Icahn School of Medicine at Mount Sinai)

*Primary Place of Performance: Vanderbilt University

Role: Subcontract PI

Procuring Contracting/Grants Officer and Address: Unknown

Specific Aims: 1) Further elucidate the mechanism of oncogenic TEAD/YAP activation by p53 DNA contact mutants; 2) Compare and contrast mechanisms by which p53 conformational mutations exert a different GOF from DNA contact mutants in the same cellular context; 3) Investigate a novel p53 DNA contact mutant GOF in colorectal carcinomas (CRCs).

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 04/2023 – 03/2028

*Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
1. 2024	0.60 Calendar
2. 2025	0.60 Calendar
3. 2026	0.60 Calendar
4. 2027	0.60 Calendar
5. 2028	0.60 Calendar

*Title: Biochemical and Chemical Training for Cancer Research

Major Goals: This training grant provides support for 4 pre-doctoral and 3 post-doctoral candidates.

*Status of Support: Pending

Project Number: **2T32 CA009582-36**

Name of PD/PI: Hiebert, Scott

*Source of Support: NIH/NCI

*Primary Place of Performance: Vanderbilt University

Role: PI

Procuring Contracting/Grants Officer and Address: Susan E. Lim, lims@mail.nih.gov

Specific Aims: 1) To provide integrated training in biochemistry, chemistry, structural biology, and cancer biology which is critical to the molecular understanding of cancer and the development of new approaches to molecularly targeted therapeutics.

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 06/2023-05/2028

Person Months (Calendar/Academic/Summer) per budget period.

Name of Individual: Hiebert, Scott W.

Year Ending (YYYY)	Person Months (##.##)
36. 2024	0.00 Calendar
37. 2025	0.00 Calendar
38. 2026	0.00 Calendar
39. 2027	0.00 Calendar
40. 2028	0.00 Calendar

COMPLETED (last 5 years):

Title: SPORE in GI Cancer

Goals: Focused on colorectal cancer (CRC), a disease that represents the second leading cause of cancer deaths in the US where it affects more men and women than all other gastrointestinal malignancies combined.

Agency Number: 6P50 CA095103-15, UNIV58948

Principal Investigator: Scott Hiebert Ph.D. from Robert Coffey.

Time Commitments/Role: 5.0%, 0.60 Calendar Months, Consortium P.I.

Supporting Agency: NIH/NCI via consortium from VUMC

Procuring Contracting/Grants Officer: Peter Ujhazy, ujhazyp@mail.nih.gov

Performance Period: 04/30/2016-04/30/2018

Specific Aims:

- 1) Project 1 Multimodal imaging & targeted therapeutics of stem cell-derived colon cancer
- 2) Project 2 Targeting K-RAS in CRC
- 3) Project 3 Molecular markers of CRC recurrence
- 4) Project 4 Genetic & epigenetic markers of colorectal adenoma recurrence

*Title: Cancer Center Support Grant – Pilot: Mitochondrial Regulation of Acute Myeloid Leukemia VICC
Multi-PI Pilot Grant Proposal

Major Goals: Test the role of epigenetic modifiers on histone methylation and gene expression in AML. Test epigenetic modifiers on AML in vivo. Determine if epigenetic modifiers frequently mutated in AML mediate metabolic regulation of AML epigenetic marks.

*Status of Support: Completed – **New / Ended**

Project Number: 5P30 CA064140-22 UNIV60527

Name of PD/PI: Scott Hiebert Ph.D. from Jennifer Pietenpol

Time Commitments/Role: 0.00%, 0.00 Calendar Months, P.I.

*Source of Support: NIH/NCI via consortium from VUMC

Procuring Contracting/Grants Officer: Sonya Roberson robersos@mail.nih.gov

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date: (MM/YYYY) (if available): 05/2018-04/2019

Specific Aims:

- 1) Not applicable to this Pilot Project.

*Title: Role of AML1-ETO in Acute Leukemia

Major Goals: The t(8;21) occurs in roughly 10% of the acute myeloid leukemia cases and about half of these patients will suffer relapse with a poor prognosis. This work will define how the fusion protein made by this chromosomal translocation causes leukemia. We will use this information to test new therapeutic strategies for

Name of Individual: Hiebert, Scott W.

the treatment of this deadly disease, and at the same time determine how these new drugs work, so that we can better use these compounds in the future.

*Status of Support: Completed

Project Number: 5R01 CA064140-25

Name of PD/PI: Hiebert, Scott

Time Commitments/Role: 16%, 1.92 Calendar Months, P.I.

*Source of Support: NIH/NCI

Procuring Contracting/Grants Officer: Malgorzata Klauzinska klauzing@mail.nih.gov

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date: (MM/YYYY) (if available): 07/2014-10/2019

Specific Aims:

- 1) Genetically dissect how elongation factor interactions and endogenous MTGs affect AML1-ETO-dependent leukemia
- 2) Determine whether AML1-ETO regulates transcription initiation, pausing, and elongation

*Title: Incyte Research Support

Major Goals: Provide oversight for the Vanderbilt-Incyte Alliance research portfolio contributed by VU faculty who are VICC members. Provide oversight for the research performed in Pillar Project II entitled "Drugging Transcription in Myeloid Malignancies."

*Status of Support: Completed – **New / Ended**

Project Number: UNIV58882

Principal Investigator: Scott W. Hiebert, Ph.D. from Michael Savona

Time Commitments/Role: 5%, 0.60 Calendar Months, Co-Investigator

*Source of Support: Incyte Corporation via consortium from VUMC

Procuring Contracting/Grants Officer: Reid Huber, Executive V.P. C.S.O. 1801 Augustine Cut-off, Wilmington, DE, 19803 Attn: General Counsel

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date: (MM/YYYY) (if available): 04/2016-12/2019

Specific Aims:

- 1) Validation of multi-atlas content labeling algorithm
- 2) Demonstration of clinical utility of multi-atlas content labeling

*Title: Regulation of transcription and tumor suppression by MTG family members

Major Goals: This work will examine the role of MTG proteins in the control of gene expression at the level of transcriptional pausing and elongation. In addition, this proposal will determine the physiological function of cancer-associated mutations in transcriptional control and tumor development in vivo.

*Status of Support: Completed - **Ended**

Project Number: 5R01 CA178030-05

Name of PD/PI: Hiebert, Scott

Time Commitments/Role: 16%, 1.92 Calendar Months, P.I.

*Source of Support: NIH/NCI

Procuring Contracting/Grants Officer: Malgorzata Klauzinska klauzing@mail.nih.gov

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date: (MM/YYYY) (if available): 05/2014-04/2020 (NCE)

Specific Aims:

Name of Individual: Hiebert, Scott W.

- 1) Determine the mechanism by which MTGs regulate MLL, Aff4, and Dot1L function
- 2) Determine the effects of Mtg8 inactivation on transcriptional pausing and elongation
- 3) Use mouse models to test the effect of cancer-related point mutations that affect the associations with elongation factors in tumor suppression

*Title: What would happen if we had a drug for PAX3-FOXO1 in Rhabdomyosarcoma?

Major Goals: Alveolar rhabdomyosarcoma is one of the most aggressive and difficult to treat tumors in children. The cause in 60% of aggressive cases is a specific genetic abnormality. We will determine how this new gene causes cancer and determine what would happen to these sarcoma cells if we had a drug specific for this new gene.

*Status of Support: Completed – **New / Ended**

Project Number: 712154

Name of PD/PI: Hiebert, Scott

Time Commitments/Role: 10.00%, 1.20 Calendar Months

*Source of Support: St. Baldrick's Foundation

Procuring Contracting/Grants Officer: Laura Chung Laura@stbaldricks.org

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date: (MM/YYYY) (if available): 07/2020-12/2021 (NCE)

Specific Aims:

- 1) Define the therapeutic value of degrading PAX3-FOXO1 and/or inhibiting its targets
- 2) Define the mechanism of PAX3-FOXO1 transcriptional control

Overlap: None

*Title: Cancer Center Support Grant - Pilot: Using mutant Cas9 linked to Apex2 in an unbiased method to identify regulators of transcription in the non-coding genome

Major Goals: The goal of the Hiebert lab for this pilot funding is to establish an APEX2 system here at Vanderbilt using our AML1-ETO-FKBP12 system in which we can define the factors recruited by endogenous AML1-ETO-FKBP12 to repress individual enhancers and then rapidly degrade AML1-ETOFKBP12 to define the factors that repopulate the enhancer to reactivate these genes. We propose to use this well-validated system coupled with SILAC LC-MS/MS analysis to establish this system using the robust CD82 enhancer (1st intron), but feel that if we are successful the potential impact of this system will be VICC-wide.

*Status of Support: Completed – **New / Ended**

Project Number: 5P30 CA068485-24-Pilot, UNIV61682

Name of PD/PI: Scott Hiebert, Ph.D. from Jennifer Pietenpol, Ph.D.

Time Commitments/Role: 0.00%, 0.00 Calendar Months

*Source of Support: NIH/NCI via consortium from VUMC

Procuring Contracting/Grants Officer: Sonya Roberson robersons@mail.nih.gov

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date: (MM/YYYY) (if available): 07/2020-08/2021 (NCE)

Specific Aims:

- 1) Not applicable to this Pilot Project.

Overlap: None

*Title: Bridging Epigenomics and Patient Data to Detect Enhanceropathies in Cancer

Major Goals: The objective of this proposal is to identify the genetic determinants of disease susceptibility and treatment response that are frequently found in the non-coding regulatory regions of the genome. We will integrate

Name of Individual: Hiebert, Scott W.

gene regulation data with large-scale patient genome data and Electronic Medical Record (EMR) data to identify and characterize novel genetic markers of lymphoma and blood cancers.

*Status of Support: Completed – **This Grant Report / Ended**

Project Number: **W81XWH-20-1-0522**

Name of PD/PI: Hodges, Emily

*Source of Support: Department of Defense

*Primary Place of Performance: Vanderbilt University

Role: Collaborator

Procuring Contracting/Grants Officer and Address: Jamie A. Shortall Jamie.a.shortall.civ@mail.mil

Specific Aims: 1) Identify lymphoid “onco-enhanceropathies” in iHMR-defined B cell enhancers using genotyping and patient data.

2) Mechanistically interrogate enhancer regions nominated by ncSNPs associated with lymphoid cancers.

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 07/2020-10/2022 (NCE)

* Person Months (Calendar/Academic/Summer) per budget period.

Year Ending (YYYY)	Person Months (##.##)
1. 2021	0.40 Calendar
2. 2022	0.40 Calendar
2. 2022	0.13 Calendar

*Title: Incyte Research Support

Major Goals: To mobilize the skills, knowledge and experience developed at Vanderbilt-Ingram Cancer Center for a focused and coordinated attack on cancer with the goal of expanding understanding of cancer biology, the mechanisms of action of Incyte compounds on both solid and liquid tumors, approaches to patient selection and stratification, and new therapeutic targets in oncology.

*Status of Support: Completed - **Ended**

Project Number: **Incyte-2020**

Name of PD/PI: Savona, Michael

*Source of Support: Incyte Corporation (VUMC82169 / 61758)

*Primary Place of Performance: Vanderbilt University

Role: Co-Investigator

Procuring Contracting/Grants Officer and Address: Reid Huber, Executive VP, Chief SO,

Med_Affairs_Grants@incyte.com

Specific Aims: 1) Provide oversight for the Vanderbilt-Incyte Alliance research portfolio contributed by VU faculty who are VICC members

2) Provide oversight for the research performed in Pillar Project II entitled “Drugging Transcription in Myeloid Malignancies”

3) The Landman team will support the aims of the project through algorithm optimization and development, integration with high performance computing, maintaining data providence, and ensuring that backup / retention policies are followed.

Statement of Potential Overlap: None

Project/Proposal Start and End Date: (MM/YYYY) (if available): 01/2020 – 12/2022

Name of Individual: Hiebert, Scott W.

IN-KIND:

*Summary of In-Kind Contribution: VICTR Scholarships - Dr. Hiebert also receives a variable amount in vouchers for shared resources from the Vanderbilt-Ingram Cancer Center . These vouchers provide access to expensive, specialized equipment that is open to all at Vanderbilt and that support his grant funded research.

*Source of Support: Vanderbilt Institute for Clinical and Translational Research

*Status of Support: Pending

Overall Objectives: These are vouchers that support the overall research goals of the laboratory by making cancer center supported shared resources more affordable. Thus, we are able to access expensive, specialized equipment that is open to all at Vanderbilt and that support his grant funded research.

Statement of Potential Overlap: No

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date : (MM/YYYY) (if available): Pending-08/2023

*Summary of In-Kind Contribution: CCSG Scholarships - Dr. Hiebert also receives a variable amount in vouchers for shared resources from the Vanderbilt-Ingram Cancer Center . These vouchers provide access to expensive, specialized equipment that is open to all at Vanderbilt and that support his grant funded research.

*Source of Support: Vanderbilt-Ingram Cancer Center / Vanderbilt University Medical Center

*Status of Support: Pending

Overall Objectives: These are vouchers that support the overall research goals of the laboratory by making cancer center supported shared resources more affordable. Thus, we are able to access expensive, specialized equipment that is open to all at Vanderbilt and that support his grant funded research.

Statement of Potential Overlap:

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date : (MM/YYYY) (if available): Pending-08/2023

*Summary of In-Kind Contribution: DDRC Scholarships - Dr. Hiebert also receives a variable amount in vouchers for shared resources from the Vanderbilt Digestive Diseases Research Center. These vouchers provide access to expensive, specialized equipment that is open to all at Vanderbilt and that support his grant funded research.

*Source of Support: Vanderbilt Digestive Diseases Research Center / Vanderbilt University Medical Center

*Status of Support: Pending

Overall Objectives: These are vouchers that support the overall research goals of the laboratory by making cancer center supported shared resources more affordable. Thus, we are able to access expensive, specialized equipment that is open to all at Vanderbilt and that support his grant funded research.

Statement of Potential Overlap:

*Primary Place of Performance: Vanderbilt University

Project/Proposal Start and End Date : (MM/YYYY) (if available): Pending-08/2023

*

Current and Pending Support

John Anthony Capra, Ph.D.

CURRENT:

Title: The Evolution of Gene Regulation and Human Disease

Agency Number: 5 R35 GM127087-02

Principal Investigator: John Anthony Capra

Time Commitments/Role: 51%, 6.12 calendar months / PI

Supporting Agency: NIH/NIGMS

Procuring Contracting/Grants Officer: Veerasamy Ravichandran
veerasamy.ravichandra@nih.gov

Performance Period: 04/01/2018 – 03/31/2023

Goals: The major goal is develop computational models that leverage machine learning, evolutionary patterns, and large-scale clinical biobanks to identify mutations that disrupt proper control of genes and lead to disease.

Specific Aims:

1) Not applicable to this type of grant.

Overlap: None

Title: Integrating genomic studies of gestational duration and birth weight to understand maternal and fetal causes of adverse pregnancy outcomes and links with later diseases

Agency Number: N/A - **New**

Principal Investigator: Ge Zhang

Time Commitments/Role: 5%, 0.6 calendar months / PI

Supporting Agency: NIH (through Cincinnati Children's Hospital Medical Center)

Procuring Contracting/Grants Officer:

Performance Period: 12/01/2020-11/30/2024

Goals: The major goals of this subcontract are bioinformatic and evolutionary genomic analysis of loci identified through genome-wide association analyses of gestational duration and fetal growth

Specific Aims: N/A

Overlap: None

Title: Personalized Structural Biology: Enabling Exome Interpretation in Undiagnosed Diseases

Agency Number: 1R01 LM013434-01 - **New**

Principal Investigator: John Anthony Capra

Time Commitments/Role: 5%, 0.6 calendar months / PI

Supporting Agency: NIH/NLM

Procuring Contracting/Grants Officer: Jane Ye yej@mail.nih.gov

Performance Period: 07/01/2020-06/30/2024

Goals: Identifying the genetic variants responsible for patient phenotypes is challenging, especially for rare diseases. We will develop algorithms for interpreting variants of unknown clinical significance by modeling the effects of all variants in a patient on protein 3D structures, integrating structural features with machine learning, and using systems biology to identify functional interactions between variants. In collaboration with the Undiagnosed Diseases Network, we will develop and validate mechanistic hypotheses about the genetic basis for disease and establish a foundation for personalized treatments that reflect patients' unique genetic backgrounds.

Specific Aims:

- 1) Rapidly and efficiently evaluate the structural effects of a patient's variants
- 2) Evaluate the synergistic potential of variants in the context of protein interaction networks and the genomes of unaffected relatives

Overlap: None

Title: Integrative Analysis of Multiomic Datasets for Discovery of Molecular Underpinnings of Large-scale Human Grain Networks

Agency Number: R01 MH125933-A1 - **New**

Principal Investigator: Mikail Rubinov

Time Commitments/Role: 20%, 2.4 calendar months / co-I

Supporting Agency: NIH/NIMH

Procuring Contracting/Grants Officer: Ming Zhan ming.zhan@nih.gov

Performance Period: 07/01/2021-06/30/2024

Goals: The primary aim of this proposal is to advance our understanding of causal molecular pathways that underpin the organization of large-scale brain networks.

Specific Aims:

Aim 1. Discover molecular pathways underlying brain network variation in healthy individuals.

Aim 2. Distinguish between adaptive and non-adaptive network-neuroscience phenotypes.

Aim 3. Delineate coupled gene expression and brain-network phenotypes in development and aging.

Overlap: None

Title: Towards an Integrated Data Repository for Preterm Birth

Agency Number: 22-FY23-16 - **New**

Principal Investigator: Marina Sirota

Time Commitments/Role: 5%/ 0.6 calendar months/co-I

Supporting Agency: March of Dimes Birth Defects Foundation

Procuring Contracting/Grants Officer: Jon Cherry Sr.

Performance Period: 1/1/2020-12/31/2023

Goals: The goal of this effort is to create an integrated database to organize scientific data and research across the March of Dimes Transdisciplinary Centers (TDC) on Preterm Birth. We aim to aggregate and integrate diverse datasets that are generated by the individual MOD centers enabling data integration and collaboration leading to novel discoveries advancing preterm birth research.

Specific Aims:

Overlap: None

PENDING:

N/A

COMPLETED (last 5 years):

Title: March of Dimes Prematurity Research Center Ohio Collaborative

Agency Number: N/A – **New / Ended**

Principal Investigator: Sing Sing Way

Time Commitments/Role: ~5% effort, 0.6 calendar/co-I

Supporting Agency: March of Dimes (through CCHMC)

Procuring Contracting/Grants Officer: Unknown

Performance Period: 01/01/2022-12/31/2022

Specific Aims:

- 1) Develop machine-learning-based risk prediction algorithms for preterm birth in diverse populations
- 2) Investigate the genetic architecture of reproductive traits and diseases
- 3) Integrate diverse 'omics data to identify candidate sPTB genes

Overlap: None

Title: Bridging Epigenomics and Patient Data to Detect Enhanceropathies in Cancer

Agency Number: W81XWH-19-PRCRP-IASF CA190332 – **This Grant Report / Ended**

Principal Investigator: Emily Hodges

Time Commitments/Role: 5%, 0.6 calendar months/Collaborator

Supporting Agency: Department of Defense (via Vanderbilt University)

Procuring Contracting/Grants Officer: Unknown

Performance Period: 09/01/2020 – 10/31/2022

Specific Aims:

Overlap: None

Title: Vanderbilt Center for Undiagnosed Diseases (VCUD)

Agency Number: 5 U01 HG007674-07 - **Ended**

Principal Investigator: John Atlas Phillips III

Time Commitments/Role: 0%: 0.00 Academic; 0.00 Summer Months / PI Subcontract

Supporting Agency: NIH/NHGRI

Procuring Contracting/Grants Officer: Argenia Lanisha Necole Dossargenia.doss.ctr@usuhs.edu

Performance Period: 09/19/2018 – 06/30/2022

Goals: The major goal of this project is to coordinate the development and implementation of the 3D protein structural analysis and variant clustering pipeline and to apply it to candidate variants identified in UDN patients.

Specific Aims:

- 1) Improve the diagnosis and care of UD patients of all ages
- 2) Determine causes of and improve treatment options for UD
- 3) Test the hypothesis that BioVU and Structural Biology can be used as an integrated approach to identify causative variants in UDN patients.

Overlap: None

Title: An Atlas of Clinical Associations mapping to Variants within Protein Structures

Agency Number: 5R01GM126249-02 (UNIV60663) - **Ended**

Principal Investigator: William S. Bush

Time Commitments/Role: 5%, 0.60 Academic; 0.00 Summer Months

Supporting Agency: NIH/NIGMS consortium from Case Western Reserve University

Procuring Contracting/Grants Officer: Veerasamy Ravichandran
veerasamy.ravichandra@nih.gov

Performance Period: 09/30/2018-06/30/2021

Goals: The goal of this project is develop to tools and software for projecting low-frequency genetic variation (captured by genomic sequencing) into protein structural space, and to perform grouped-variant association tests to establish an online atlas, mapping functional protein regions to the thousands of clinical traits they potentially influence.

Specific Aims:

1) None Provided

Overlap: None

Title: Integrating 'Omics and Electronic Health Records to Elucidate the Genetic Architecture of Preterm Birth

Agency Number: Preterm Birth Initiative

Principal Investigator: John Anthony Capra

Time Commitments/Role: 5%, 0.45 Academic; 0.69 Summer Months, PI

Supporting Agency: Burroughs Wellcome Fund

Procuring Contracting/Grants Officer: Unknown

Performance Period: 06/01/2017-03/31/2021

Goals: The goal of this project is to integrate genomic-wide 'omics data with large electronic health record databanks to model how genetic variation influences risk of preterm birth.

Specific Aims:

1) Not applicable to this type of grant.

Title: Integrating 'omics and electronic health records to elucidate the genetic architecture of preterm birth

Agency Number: BWF_1017283 - **Ended**

Principal Investigator: John Anthony Capra, Ph.D.

Time Commitments/Role: 5%: 0.6 calendar months / PI

Supporting Agency: Burroughs Wellcome Fund

Procuring Contracting/Grants Officer: Unknown

Performance Period: 06/01/2017-03/31/2021

Specific Aims:

1) Integrate diverse 'omics data to identify candidate sPTB genes

2) Integrate diverse 'omics data to identify candidate sPTB gene regulatory loci

3) Conduct a phenome-wide scan of electronic health records to discover candidate gene associations with sPTB and correlated adverse pregnancy outcomes

Overlap: None

Title: Analyses, Validation, and Resource Creation for Genome Sequencing of Complex Diseases

Agency Number: 5 U01 HG009086-04 - **Ended**

Principal Investigator: Bingshan Li

Time Commitments/Role: 5%: 0.60 Academic; 0.00 Summer Months/Co-Investigator

Supporting Agency: NIH/NHGRI

Procuring Contracting/Grants Officer: Adam Felsenfeld felsenfa@mail.nih.gov

Performance Period: 05/18/2016 – 04/30/2020

Goals: The major goal of this project is to a suite of statistical and computational methods to identify genes and variants associated with complex disease and use BioVU, a DNA BioBank linked to electronic health records, to validate and investigate genetic architecture of multiple complex diseases.

Specific Aims:

1) To develop novel methods, software and analysis pipelines for integrated analysis of common and rare variants for the analysis of whole genome sequence on 50,000- 100,000 individuals for any given common disease

- 2) To develop and apply novel approaches for prioritizing results from these large-scale sequencing studies using BioVU, the 200,000 member biobank at Vanderbilt that is associated with 30 years of high quality electronic health records
- 3) To develop queryable results databases and a comprehensive web portal to serve results of our studies on the sequence data for both the internal sequencing community and for the broader scientific community.

Title: Analyses, Validation, and Resource Creation for Genome Sequencing of Complex Diseases

Agency Number: 5 U01 HG009086-04 - **New / Ended**

Principal Investigator: Bingshan Li

Time Commitments/Role: 5%: 0.60 Academic; 0.00 Summer Months/Co-Investigator

Supporting Agency: NIH/NHGRI

Procuring Contracting/Grants Officer: Adam Felsenfeld felsenfa@mail.nih.gov

Performance Period: 05/18/2016 – 04/30/2020

Goals: The major goal of this project is to a suite of statistical and computational methods to identify genes and variants associated with complex disease and use BioVU, a DNA BioBank linked to electronic health records, to validate and investigate genetic architecture of multiple complex diseases.

Specific Aims:

- 1) To develop novel methods, software and analysis pipelines for integrated analysis of common and rare variants for the analysis of whole genome sequence on 50,000- 100,000 individuals for any given common disease
- 2) To develop and apply novel approaches for prioritizing results from these large-scale sequencing studies using BioVU, the 200,000 member biobank at Vanderbilt that is associated with 30 years of high quality electronic health records
- 3) To develop queryable results databases and a comprehensive web portal to serve results of our studies on the sequence data for both the internal sequencing community and for the broader scientific community.

Title: From Gene Discovery in Human Pregnancy to Mechanisms of Birth Timing

Agency Number: 22-FY18-8924 (UNIV60581)

Principal Investigator: Antonis Rokas

Time Commitments/Role: 3.33%, 0.30 Academic; 0.00 Summer Months, Section Lead

Supporting Agency: Cincinnati Children's Hospital/March of Dimes

Procuring Contracting/Grants Officer: Unknown

Performance Period: 10/01/2018-09/30/2019

Goals: The major goal of this project is to develop a web-based encyclopedia for discovery and exploration of preterm birth genes and genetic elements.

Specific Aims:

- 1) Not applicable to this type of grant.

Title: Modeling the dynamics of genome-scale data across trees

Agency Number: 5R01 GM115836-03

Principal Investigator: Dennis Kostka

Time Commitments/Role: 27%, 2.25 Academic; 1.00 Summer Months, Multi-PI

Supporting Agency: NIH/NIGMS via University of Pittsburgh

Procuring Contracting/Grants Officer: Veerasamy Ravichandran
veerasamy.ravichandra@nih.gov

Performance Period: 08/01/2015 – 06/30/2019

Goals: The major goal of this project is to develop statistical methods and software that enable the integration and interpretation of genome-wide epigenetic and expression data from related contexts, such as stages across cellular differentiation.

Title: Elucidating the Gene Regulatory Landscape of Human Preterm Birth

Agency Number: N/A

Principal Investigator: John Anthony Capra

Role: PI

Supporting Agency: March of Dimes Innovation Catalyst Award

Procuring Contracting/Grants Officer: Unknown

Performance Period: 01/01/2015 – 01/01/2017

Goals: The goal of this project is to identify non-protein-coding gene regulatory regions relevant to pregnancy and parturition and to study their function and evolution.

Cross-tissue patterns of DNA hypomethylation reveal genetically distinct histories of cell development

Timothy Scott¹, Tyler Hansen², Evonne McArthur¹, Elizabeth Dorans¹, Emily Hodges^{1, 2, *}

¹Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, TN 37212;

²Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN 37212

DNA methylation (DNAm) is essential for balanced multi-lineage cellular differentiation, but exactly how DNAm drives cellular phenotypes is unclear. While >80% of CpG sites are stably methylated, tens of thousands of discrete CpG loci form hypomethylated regions (HMRs). Because they lack DNAm, HMRs are considered transcriptionally permissive, but not all HMRs participate in genome regulation. Contrary to long-held views about the role of DNAm in controlling promoter function, most promoter HMRs are invariable across cell-types regardless of transcriptional status. In contrast, a subset of HMRs is cell-type specific and enriched for tissue specific enhancers. Recently, we showed that, while enhancer HMRs correlate with chromatin accessibility and other indicators of permissive chromatin, their temporal dynamics are distinct from chromatin—HMRs can persist long after accessibility is lost. These studies suggest that establishment of HMRs is an important step in enforcing cell identity, and that patterns of HMR establishment are more complex than previously appreciated.

To understand their functional significance, we systematically dissected HMR patterns across diverse cell types and developmental timepoints, including human embryonic stem cells, fetal spine and heart, adult hematopoietic lineage, liver, and adrenal tissues. Unsupervised clustering of 102,390 distinct HMRs revealed that levels of HMR “sharedness” across cell-types reflects a developmental hierarchy supported by enrichment of stage-specific transcription factors and gene ontologies. Using a pseudo-time course of embryonic to adult hematopoietic development, we further show that a majority of HMRs observed in differentiated cells (~70-75%) are established at early developmental stages and accumulate as development progresses. HMRs that arise during differentiation frequently (~35%) establish near existing HMRs (≤ 6 kb away), leading to the formation of HMR clusters associated with stronger enhancer activity. Using SNP-based partitioned heritability from GWAS summary statistics of diverse traits and clinical lab values, we discovered that genetic contribution to trait heritability is enriched within HMRs. Moreover, heritability of cell-relevant traits increases with both increasing developmental specificity and HMR clustering, supporting the role for distinct HMR subsets in regulating normal cell function.

Altogether, our findings reveal that HMRs can predict cellular phenotype by providing genetically distinct historical records of a cell’s journey through development, and this attribute distinguishes DNAm from other epigenomic features.

Keywords:

(57) Epigenetics, (119) Methylation, (76) Gene regulation, (14) Bioinformatics

Primary category:

- I. **Epigenetics (I-III):** III Genome-scale, integrative, and/or comparative epigenomic analyses

1

2

3

4 **Cross-tissue patterns of DNA hypomethylation reveal genetically distinct histories**
5 **of cell development**

6

7 Timothy J. Scott¹, Tyler J. Hansen², Evonne McArthur¹, Emily Hodges^{1, 2, *}

8

9 ¹Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, TN, 37232,
10 USA.

11 ²Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN, 37232,
12 USA.

13

14 Short title: DNA hypomethylation and cell developmental histories

15

16 *Corresponding author

ABSTRACT (300 words)

Establishment of DNA methylation (DNAm) patterns is essential for balanced multi-lineage cellular differentiation, but exactly how these patterns drive cellular phenotypes is unclear. While >80% of CpG sites are stably methylated, tens of thousands of discrete CpG loci form hypomethylated regions (HMRs). Because they lack DNAm, HMRs are considered transcriptionally permissive, but not all HMRs actively regulate genes. Unlike promoter HMRs, a subset of non-coding HMRs is cell-type specific and enriched for tissue specific gene regulatory functions. Our data further argues not only that HMR establishment is an important step in enforcing cell identity, but also that complex HMR patterns are functionally instructive to gene regulation. To understand the significance of non-coding HMRs, we systematically dissected HMR patterns across diverse human cell types and developmental timepoints, including embryonic, fetal, and adult tissues. Unsupervised clustering of 102,390 distinct HMRs revealed that levels of HMR specificity reflects a developmental hierarchy supported by enrichment of stage-specific transcription factors and gene ontologies. Using a pseudo-time course of development from embryonic stem cells to adult stem and mature hematopoietic cells, we find that most HMRs observed in differentiated cells (~70-75%) are established at early developmental stages and accumulate as development progresses. HMRs that arise during differentiation frequently (~35%) establish near existing HMRs ($\leq 6\text{kb}$ away), leading to the formation of HMR clusters associated with stronger enhancer activity. Using SNP-based partitioned heritability from GWAS summary statistics across diverse traits and clinical lab values, we discovered that genetic contribution to trait heritability is enriched within HMRs. Moreover, the contribution of heritability to cell-relevant traits increases with both increasing developmental specificity and HMR clustering, supporting the role of distinct HMR subsets in regulating normal cell function. Altogether, our findings reveal that HMRs can predict cellular phenotypes by providing genetically distinct historical records of a cell's journey through development.

AUTHOR SUMMARY

Studies aiming to understand the relationship between DNA methylation patterns and phenotypic outcomes have focused largely on individual differentially methylated regions without consideration of combinatorial changes that drive phenotypes. In non-disease contexts, most of the human genome is stably methylated, except for thousands of discrete DNA hypomethylated regions (HMRs) coinciding with gene regulatory elements. Here, we comprehensively characterize HMR relationships both within and between developmentally diverse cell types to

understand the functional significance of complex HMR patterns. We show that levels of HMR specificity across cell-types captures time-point specific branchpoints of development. Our analysis further reveals that HMRs form clusters in proximity to cell identity genes and are associated with stronger gene enhancer activity. This is a wide-spread phenomenon and only a very small subset of HMR clusters is explained by overlapping super-enhancer annotations. Partitioned heritability revealed the functional significance of different HMR patterns linked to specific phenotypic outcomes and indicates a quantitative relationship between HMR patterns and complex trait heritability. Altogether, our findings reveal that HMRs can predict cellular phenotypes by providing genetically distinct historical records of a cell's journey through development, ultimately providing novel insights into how DNA hypo-methylation mediates genome function.

INTRODUCTION

Of twenty-eight million CpG sites in the human genome, the majority (80-85%) undergo constant DNA methylation (DNAm) (1-5). However, a subset of sites forms regions that are either broadly hypomethylated across cell-types or hypomethylated in a cell-type or lineage-dependent manner (hypomethylated regions, HMRs) (6-14). Contrary to long-held views of the role of DNAm in controlling promoter function, promoters are stably hypomethylated and largely invariant across cell types (15, 16). Thus, promoter HMRs poorly predict gene transcriptional status and ultimately cellular phenotypes. By contrast, the most dynamically hypomethylated HMRs are distal to promoters and often coincide with signature chromatin states typical of enhancers (5, 17-20). Yet, the functional significance of cell-type and lineage specific HMRs and their causal relationship with genes and cellular phenotypes is not well-defined.

Differential DNAm, like differential gene expression, is sufficient to distinguish diverse cell types and recapitulate their functional and developmental relationships (4, 9, 10, 21). Furthermore, differentially methylated regions are proximal to genes with tissue-specific functions and are enriched for GWAS (genome-wide association study) variants linked to tissue-relevant diseases (11, 22-25). While individual differentially methylated regions frequently overlap enhancers, few studies have considered the combinatorial significance of enhancer methylation patterns in a genome-wide manner across developmentally diverse datasets. For example, selective and continuous hypomethylation of individual enhancer units within super-enhancers (SEs) has been demonstrated in mouse embryonic stem cells (ESCs) during exit from naïve pluripotency, suggesting that coordinated hypomethylation of enhancers through cell fate transitions serves to

uphold specific cellular states (26-28). Furthermore, our recent work shows that, while HMRs correlate with chromatin accessibility and other indicators of permissive chromatin, the temporal dynamics of HMR establishment are distinct from chromatin remodeling—HMRs persist long after chromatin remodeling changes (12). These studies indicate that HMRs uniquely capture both active and previously active gene regulatory elements; however, this hypothesis and its link to phenotypic outcomes remains to be tested across diverse tissues and developmental timepoints in a genome-wide manner.

Here, we performed a comparative analysis of whole genome methylation data from diverse tissues representing distinct organ systems and developmental timepoints. Unlike previous studies that emphasize pairwise differential methylation or locus-specific changes during limited differentiation time courses, we comprehensively characterize both HMR relationships within and between cell types to understand the functional significance of complex HMR patterns. We show that hierarchical conservation of HMRs across tissues can distinguish different types of enhancers and provides a map for broad developmental timelines underlying cellular states. We further demonstrate that this hierarchical relationship highlights meaningful HMR sequence contribution to complex trait heritability, ultimately providing novel insights into how DNAm mediates genome function.

RESULTS

Shared HMR patterns among diverse cell types reveal common functional and developmental histories.

Non-coding HMRs can be functionally categorized by their cell-type specificity (9, 29). Previous work comparing distantly related cell types (embryonic stem cells versus terminally differentiated primary B cells) revealed two broad classes of HMRs: 1) highly shared, developmentally constitutive HMRs and 2) cell type specific HMRs that coincide with enhancer chromatin marks. Motivated by these studies, we hypothesized that differential comparisons of non-coding HMR profiles across multiple diverse cell types could provide insight into the presence and status of regulatory elements in individual cell types. To illustrate this idea, genome browser tracks of methylation data are displayed for datasets representing diverse lineages and developmental timepoints at a B cell enhancer cluster upstream of the CD27 gene (**Fig 1A**). A comparison of HMRs reveals different levels of cell-type and lineage specificity including HMRs conserved in all samples (developmentally constitutive), HMRs shared exclusively among lineage-related samples (e.g., hematopoietic cells), and HMRs present only in B cells.

To investigate the intersection of HMR patterns globally, we determined a set of high-confidence HMRs using whole genome bisulfite sequencing (WGBS) data from 7 different cell types and tissues, including embryonic stem cells (H1ESCs), fetal heart, fetal spinal cord, liver, adrenal gland, T cells, and B cells (see Methods). As the resolution of HMR specificity is contingent on the quantity and interrelatedness of cell types included in the analysis, we maximized comparative potential by including datasets representing a diversity of organ systems and developmental stages; additionally, we used high quality methylation data with a minimum base pair coverage of >10x sequence reads (30). This resulted in a total set of 103,017 unique HMRs with an average length of ~784bp, excluding TSS and exon spanning HMRs.

Hierarchical clustering applied to these datasets was sufficient to recapitulate both related and distant cell type relationships, demonstrating the quality and specificity of HMR calls (**Fig 1B**). Next, we utilized *k*-means clustering to group non-coding HMR methylation levels across the 7 different cell types and tissues in an unsupervised manner. A gap statistic determined an optimal number of *k*-means clusters ($n=7$) and the resultant heatmap revealed groups of HMRs highly stratified by both cell type and developmental stage (**Fig 1C, S1 Fig**). We manually classified each *k*-means group according to cell types displaying average HMR methylation $\leq 50\%$ for each group. For example, in the “Differentiated” HMR group all cell types except H1ESC display low methylation levels, whereas the “Early Developmental” HMR group is dominated by H1ESCs. Likewise, a group of HMRs is specific to the “Fetal” developmental state compared to stem and adult cells. Using this analysis, we achieve remarkable resolution to distinguish highly related cell types such as B and T cells; further, we identify a more specific group of exclusive B cell HMRs.

The specificity of the unsupervised groupings points to gene regulatory elements with shared functional and developmental histories. Since transcription factors (TFs) govern the functional progression and specialization of cell types, we performed TF motif enrichment analysis to understand the gene regulatory significance of each *k*-means group. Top motifs stratify strongly by *k*-means cluster (**Fig 1D**). Furthermore, representative TFs from top results show *k*-means cluster-specific enrichment of canonical TFs indicative of their respective cell types. For example, the ubiquitous CTCF is enriched in the *All* cluster (31); pluripotency factors OCT4-Sox2-Nanog are primarily in the *Early Developmental* *k*-means cluster (32); Lymphoid factor ETS1 is in the *Lymphoid specific* group (33, 34); and early B cell factor EBF is in the highly specific *B cell* group (35). Similarly, hepatocyte nuclear factor-4 alpha (HNF4 α) and retinoic acid receptor alpha (RAR α) motifs are present exclusively in the liver/adrenal-specific *Endocrine axis* group.

Given the specificity of the TF enrichment analysis supporting cell- and lineage-specific functions, we considered whether associated genes displayed similar biological specificity. Using GREAT ontology enrichment analysis (36), we show enrichment of distinct biological processes representative of the cell type and developmental stage associated with HMR groups (**Fig 1E**). Interestingly, T cell-related ontologies are strongly and uniquely enriched in the *Lymphoid specific* *k*-means cluster, whereas B cell-specific ontologies are enriched in the *B cell* group; this highlights the ability of HMRS to distinguish not only disparate lineages and developmental stages, but also highly related cell types. Together these data show that HMRS alone can recapitulate functional relationships between cell types. Furthermore, by comparing HMRS within and across lineages, we discovered that levels of HMR specificity can reflect deep developmental roots of gene regulation, capturing time-point specific branchpoints of development.

HMRS accumulate and persist through subsequent developmental transitions.

Terminally differentiated cells exhibit four times the number of non-coding HMRS compared to embryonic stem cells (**Fig 1C**). A minor subset of H1ESC HMRS are cell-type specific, while most H1ESC HMRS are highly shared across the cell types analyzed (2,822 vs. 15,413 non-coding HMRS). Our comparative analysis further reveals specific HMR groups defined by developmental stage (fetal vs. adult, differentiated vs. undifferentiated), lineage, and cell type (**Fig 1C**). These data suggest a model whereby H1ESCs supply a base HMR set to which additional HMRS are added at distinct lineage commitments through cell development. This is important because it suggests that a developmental hierarchy exists among HMRS and that HMRS accumulate as cells differentiate.

To determine whether progressive HMR establishment can be traced in developmentally derived cell types, we used publicly available WGBS data to construct a pseudo-time course of pluripotent (H1ESCs), multipotent (hematopoietic stem and progenitor cells, HSPCs), and terminally differentiated myeloid (macrophages) and lymphoid (B cells) lineage cells (**Fig 2A**). In general, we observe that HMRS increase in number with increasing cell maturity. An increase of total HMRS could be explained by (1) a simple accumulation of additional HMRS, or (2) a net increase with high turnover of HMRS. To differentiate between these two modes of HMR expansion, we measured HMR overlap between either embryonic stem cells or hematopoietic stem cells and mature hematopoietic cell types (**Fig 2B**). Of 18,235 HMRS observed in H1ESCs, 11,451 (62.79%) were represented in the multipotent HSPC dataset. In the fully differentiated cell types, 11,036 (60.52%) and 10,393 (56.99%) were maintained in the macrophage and B cell datasets, respectively. Next, of 34,605 HMRS established in HSPCs but absent in H1ESCs, 27,307

(78.91%) and 15,305 (44.23%) were represented in the macrophage and B cell datasets, respectively.

These data show that a majority of the HMRs observed in differentiated cells (~60%) are established at early developmental stages and suggest a pattern of HMR accumulation in relation to developmental progression. In addition to acquiring new HMRs, macrophages retain a majority of HMRs established in HSPCs, whereas B cells retain half as many HSPC-derived HMRs and fewer total HMRs compared to macrophages. This observation is consistent with previous studies demonstrating a lineage priming imbalance that favors myeloid differentiation in the absence of DNMT3A (37, 38); thus, lymphoid commitment and myeloid restriction requires re-methylation of specific early hematopoietic regulatory elements in parallel to demethylation of lymphoid-specific elements. Overall, our analysis supports a model where new HMRs are progressively established through successive developmental stages and persist through later stages of cell differentiation.

HMRs are non-randomly established into spatially organized clusters.

Locus-specific analysis of individual WGBS datasets indicates that multiple distinct HMRs are frequently located near one another, rather than being randomly distributed across linear genomic space. Moreover, these HMR groups appear to be spatially organized with HMRs displaying varying degrees of cell type specificity, from developmentally constitutive to B cell specific. An example locus shown in **Fig 1A** depicts a group of adjacent HMRs near the *CD27* gene, containing a cluster of genes which play a key role in B cell function (39-41). To quantify this genome-wide, we calculated observed and expected distributions of inter-HMR distances utilizing the broad set of diverse cell types represented in **Fig 1**. Expected distributions were simulated by random shuffling (n=10,000) HMRs across the genome for each dataset, excluding a blacklist of protein coding RefSeq promoters and exons. HMR distances are significantly closer to each other than expected by random chance (**Fig 3A**, Wilcoxon rank sum, p-value < $2.2e^{-16}$). Interestingly, differentiated cell types consistently show lower expected and observed distances (~40-50kb and ~12kb, respectively) compared to those of H1ESCs, which feature the largest inter-HMR distances; this is consistent with having fewer HMRs overall, supporting its role as a basal HMR set.

These data suggest that clustered HMRs play a distinct regulatory role compared to their unclustered counterparts. To characterize the features that distinguish “clustered” and “unclustered” HMRs we first determined a set of heuristic criteria to define clusters (**Fig 3B**). We plotted per-cell type distributions for non-coding inter-HMR distances and measured distance

quantiles. From this, we analyzed "end-to-end" cluster lengths based on three maximum linking values: the ≤ 12.5 kb stitching distance commonly used in ChIP-seq-based super-enhancer studies (42-46), the approximate median HMR distance of ≤ 10.5 kb, and ≤ 6 kb which represents the median after filtering for end-to-end cluster lengths of 50kb (**S1 Table**). Previous studies that have characterized clustered super enhancers have used a common linking distance threshold of 12.5 kb. However, this distance was empirically identified in ChIP-seq datasets, and such a distance applied to methylation data results in extraneously long stitched regions, the biological function of which is difficult to assign; some exceed 1Mb, which can result from HMRs spread across gene deserts, or large topological domains with low methylation levels or CpG frequency. By comparison, a linking distance of 6kb results in stitched regions with an overall mean value of ~ 10.5 kb, which is consistent with other clustered enhancer annotations such as stretch and super enhancers ((43, 47), **S2 Table, S2 Fig**). Using a linking distance of 6kb, we determined the fraction of HMRs that are clustered or unclustered for a subset of cell types, including H1ESC, HSPC, B cell and Liver (**Fig 3C**). To avoid confounding contributions of promoter characteristics to our analysis, clusters were not allowed to cross promoters or exons. At a 6kb threshold, non-promoter HMR groupings that exist as pairs or as clusters of 3 or more constitute $\sim 35\%$ of all HMRs.

As demonstrated in **Fig 1A**, a typical cluster consists of multiple HMRs with different levels of cell type specificity between them—broadly shared (developmentally constitutive), lineage-shared or cell-specific. This means that a cluster identified in one cell type, may not exist across all cell types, particularly in the case of H1ESCs. Moreover, many HMR clusters ($\sim 35\text{-}40\%$) contain at least one lineage- and/or cell type specific HMR. Given that HMRs accumulate over developmental timelines, this observation raises the possibility that HMRs are preferentially added to clusters in a lineage specific manner as cells differentiate. Indeed, we observe a positive correlation between clustering and developmental state. Clustering percentage increases as development progresses (**Fig 3D**, *H1ESC to HSPC & HSPC to Macrophage*: $p < 2.2 \times 10^{-16}$), and this is accompanied by a relative decrease in unclustered HMRs. Tracking these HMRs temporally for each pseudo-timepoint reveals that a substantial fraction of early-established HMRs is joined by additional HMRs in subsequent developmental states. The establishment of new HMRs near existing HMRs can lead to clustering, where an HMR may be classified as unclustered at an early timepoint but become clustered in a differentiated cell type. These growing clusters of HMRs are often in proximity to lineage-specific genes, as suggested by gene ontology analysis (**Fig 3E-F**). Altogether, these data show that HMRs can be broadly distinguished by 1)

the number of cell types that share them – a corollary of temporal establishment or developmental time, and 2) their clustering behavior, which may reflect a collective and unique developmental function that distinguishes clustered HMRs from other types of genomic regions.

Clustered HMRs are functionally distinct from unclustered HMRs.

Sequencing approaches have enabled the discovery of many spatially clustered regulatory elements genome-wide using chromatin accessibility (47), histone modifications (48-53), and transcription factor binding (42, 43). More recently, clustering of enhancers has been commonly associated with concepts such as super enhancers (SEs) (43), stretch enhancers (47), shadow enhancers and locus control regions (LCR) (54, 55), which are thought to provide regulatory additivity, synergy, and redundancy to their target genes in a tissue specific manner.

Comparison of clustered B cell HMRs with histone H3K27ac-defined B cell super enhancers shows that, while HMRs capture a majority proportion of SEs, SEs explain only a minor proportion (~1.5%) of all clustered HMRs (**Fig 4A**). This discrepancy may be explained by the fact that SEs are often defined by strong H3K27 acetylation or strong TF binding, and only a fraction of SEs exist as clusters in linear genomic space. On the other hand, HMRs are more pervasive, and their clustering frequency may reflect a broader gene regulatory function or phenomenon. To address this question, we used ChromHMM annotations to functionally categorize HMRs based on clustering behavior in B cells (**Fig 4B**) (48, 49). Notably, clustered HMRs are enriched for “strong enhancers” ($X^2=46.358$, $p=9.848 \times 10^{-12}$) while unclustered HMRs show higher enrichment of “heterochromatin” ($X^2=1434.3$, $p<2.2 \times 10^{-16}$) and “insulators” ($X^2=731.47$, $p<2.2 \times 10^{-16}$). This suggests clustered HMRs are enriched for active regulatory regions while unclustered HMRs tag elements involved in three-dimensional chromatin structure. This result is corroborated by the strong enrichment of the CTCF motif in unclustered HMRs, while both clustered and unclustered B cell HMRs show comparable enrichment of lymphoid-relevant transcription factors, including PU.1, SpiB, and ETS family members (**Fig 4C**).

Given the enrichment of strong enhancer annotations in clustered HMRs, we investigated their transcriptional regulatory activity by comparing with our recently published ATAC-STARR-seq data for immortalized B cells (**Fig 4D**) (56). ATAC-STARR-seq is a massively parallel reporter assay that uses Tn5 transposase to selectively clone accessible DNA from native chromatin into a plasmid-based reporter to test accessible chromatin regions for active and silent regulatory activity (56, 57). Since a majority of B cell HMRs overlap accessible chromatin regions in lymphoblastoid cells (**S3A Fig**), we measured the proportion of HMRs that contain an activator or

silencer (**Fig 4D**). Despite being fewer in number, clustered HMRs contain a significantly higher proportion of transcriptional regulators, including both activators and silencers ($p = 2.39^{-13}$ and 0.0106, respectively), than unclustered HMRs.

Based on these results, we reasoned that target genes of clustered HMRs display increased transcriptional output compared to those of unclustered HMRs. To define putative HMR target genes, we used a gene assignment strategy that identifies the nearest expressed neighboring gene within a topologically associated domain (TAD) containing both the gene and the HMR(s). A recent study showed that a combination of nearest neighbor assignment in conjunction with a minimum expression threshold increased associated-gene prediction accuracy above several gene assignment methods, including the commonly utilized simple nearest neighbor (58). Using this assignment strategy and lymphoblastoid RNA-seq data from ENCODE, we observe that clustered B cell HMRs are associated with significantly increased gene expression ($p < 2.2e-16$; **Fig 4E**) (59). When these comparisons are binned by HMR-gene distance, significant differences in expression are observed at both proximal (<10kb) and distal (50-100kb) HMR-gene distances (**Fig 4F**). We performed the same analysis for liver clustered HMR target genes, finding the pattern is consistent across cell types (**S3B-C Fig**).

Non-coding HMR patterns are highly enriched for genetic variants linked to specific clinical phenotypes.

Genome-wide associations studies (GWAS) have demonstrated that a substantial portion of human phenotype-associated single nucleotide polymorphisms (SNPs) are located in functional regulatory elements (60-64). Integration of GWAS with functional genomic data reveals that disease risk variants also localize primarily within cell-type specific enhancers of disease-relevant tissues (65). Studies examining the relationships between disease loci and molecular phenotypes such as gene expression, chromatin accessibility or the DNase status of *cis*-acting enhancers have identified a strong connection between non-coding genetic variants and epigenetic regulation (66-68). Based on these previous studies, we asked whether specific HMR patterns harbor genetic variants linked to cell-type-related clinical phenotypes, and, in turn, whether these relationships can inform the functional significance of different HMR patterns.

We reasoned that GWAS SNPs could be leveraged to reveal genetic variants in HMRs of critical importance to normal cell development and function. As in **Fig 2**, we defined B cell HMRs that are H1ESC-derived (developmentally constitutive), HSPC-derived (lineage-shared) or B cell-specific. We used stratified linkage disequilibrium score regression (S-LDSC) to perform partitioned

heritability (69) analysis from GWAS summary statistics of 79 traits and clinical lab values representing a range of organ systems (**S3 Table**). We found significant enrichment of trait heritability within lineage- and cell-specific HMRs (**Fig 5A, S4 Fig**). Moreover, trait specificity enriches with increased HMR specificity. For example, H1ESC-derived B cell HMRs are nominally enriched for traits not immediately attributable to B cell function, such as *cardiomyopathy* and *morning person*. This is unsurprising due to the pleiotropy of gene regulation and the shared genetic architecture between many complex traits. However, HSPC-derived HMRs are enriched for genetic heritability of general hematopoietic traits including *white blood cell*, *platelet*, and *neutrophil counts*. In highly B cell specific HMRs, we identify a notable enrichment of specific immune-related clinical traits and lab values, several of which achieve significance after multiple testing correction ($p < \text{Bonferroni}, n=79$). These observations hold true for S-LDSC analysis in H1ESC-derived, and cell-specific liver HMRs (**S5 Fig**).

HMRs stratified solely by clustering behavior also demonstrate heritability enrichment patterns associated with specific lymphoid traits (**Fig 5B**). In fact, compared to clustered HMRs, unclustered HMRs show no statistically significant trait enrichment above significance thresholds, suggesting that results in **Fig 5A** are powered predominantly by clustered HMRs. Accordingly, this trend was observed in gene-based disease enrichment analyses (disease ontology) applied to the same HMR groups analyzed by S-LDSC (**Fig S5**) (70, 71). For example, the top disease ontology enrichments for H1 ESC-derived HMRs uniquely include morphogenic ontologies such as *craniofacial abnormalities*, and the top ontologies for B cell specific HMRs include multiple lymphoid- and leukemia-related ontologies, reflecting the biological state associated with each HMR group. Together with the partitioned heritability results, these data suggest clustered cell-specific HMRs are both near lineage-specific genes and enrich for SNP variation associated with cell-specific traits.

To better contextualize the partitioned heritability enrichment results from B cell data, we compared results across cell types and against other known functional genomic feature annotations. We compared S-LDSC enrichment levels on a per-trait basis for B cell and liver HMR annotations (those from **Fig 5A and S4 Fig**) and other functional genomic annotations (**Fig 5C-D, S6A-B Fig**). For both immune-related clinical lab values and disease traits, we observe increasingly stronger enrichment from H1 ESC-derived to HSPC-derived to B cell-specific HMRs. In contrast, both H1 ESC-derived and liver-specific HMRs show positive enrichment for ALT (alanine transaminase) compared to B cell HMRs, as expected. This shows that SNP-based trait enrichment is capable of distinguishing HMR patterns from both distant and highly related cell

types. Across cell relevant traits, we observe SNP-based heritability enrichment values that surpass those of promoters, expression quantitative loci (eQTLs), and histone marks of open chromatin (H3K27ac) often used to approximate active regulatory regions. Enrichment values associated with cell specific HMRs are comparable to those of FANTOM5 enhancers, supporting the notion that developmentally specific HMRs mark enhancers important for cell identity. Altogether, this analysis highlights the functional significance of different HMR patterns and indicates a quantitative relationship with complex trait heritability.

DISCUSSION

While previous studies have characterized enhancer-associated methylation in a locus- or tissue-specific manner (72-75), we use comparative hypomethylation profiling to assess global hypomethylation patterns across cell types. This broader analysis reveals complex patterns of HMR establishment across a developmentally diverse dataset. By examining HMRs in a hematopoietic developmental context, we show that HMRs accumulate at distinct developmental stages and commonly persist through sequential lineage commitments. These developmentally hypomethylated regions are associated with distinct, stage-appropriate transcription factors, gene pathways, and SNP-based genetic heritability. These data allude to a model where H1 ESCs, with the fewest HMRs, present a basal set of HMRs to which additional regions are hypomethylated through development. In fact, about two-thirds of HMRs established in H1 ESCs remain in HMR datasets of differentiated cell types, highlighting their early establishment and continuous hypomethylation across time. Consequently, most (~3/4) HMRs in B cells were traced back to either H1 ESCs or HSPCs, indicating that the majority of HMRs are established at early cellular states. This further implies that biological differences between these cell types are driven by the minority population of differentially methylated HMRs. DMRs have been commonly used as a unit for studying DNA methylation (76-80); however, this annotation can lead to the loss of information of HMRs that are shared among various degrees of developmentally related cell types. Our results demonstrate that the entire HMR repertoire within a cell-type, rather than just the cell-type specific HMRs, stores information that is key to understanding and potentially predicting a cells phenotypic trajectory. In addition, we uncover a quantitative relationship between complex trait heritability and levels of HMR specificity, a feature that relates to HMR history. In general, we observe a strong pattern of hierarchical HMR establishment associated with stage-specific regulatory needs.

We further show that HMRs are preferentially established near existing HMRs, leading to the progressive enrichment of HMR clusters in differentiated cell types. Notably, clustered HMRs compose about 1/3 of all HMRs in differentiated cells, compared to less than 1/6 in H1 ESCs, indicating clustering increases proportionally to developmental progression. These spatially correlated HMRs are enriched for unique stage-relevant gene ontologies, trait-associated genetic heritability, and ChromHMM annotations, implying distinct regulatory roles compared to their unclustered counterparts. Previous investigations into enhancers describe subsets of clustering enhancers, including super-enhancers and hub enhancers (43, 54). While showing modest correlation with these annotations, HMRs capture additional non-coding space, suggesting clustered HMRs represent a broader epigenetic phenomenon than previously described.

We note that HMR clusters show patterns of hierarchical establishment that logically follow developmental paths. However, it is unclear if HMRs that persist through cell states remain epigenetically active at later stages. Clusters may include a combination of active and decommissioned, inactive enhancers recorded in HMR patterns. Murine models of early development have highlighted combinatorial and contrasting dynamics between spatially and functionally related enhancers during exit from pluripotency, where some require remethylation while others retain hypomethylation (27, 81). These enhancers may serve to uphold cellular states during cell fate transitions. Our lab has also observed ‘vestigial’ enhancers on shorter timescales by applying ATAC-Me-seq to a differentiation time course, simultaneously measuring chromatin accessibility and DNA methylation (82); these analyses reveal a subset of regions that undergo chromatin closing while simultaneously maintaining hypomethylation levels.

Partitioned heritability analysis of B cell HMRs established at three distinct developmental stages revealed enrichment of traits that reflect stage-relevant biology; in general, broadly shared HMRs were enriched for heritability of broader phenotypes while B cell specific HMRs were enriched for lymphocyte-relevant traits. Each B cell HMR subset likely suffers from power limitations, representing between 3,187,775 and 9,228,469 bp, or as low as ~0.1% of the genome. Despite this limitation, we observe a remarkable correspondence between heritability enrichment and stage-specific HMRs.

Our findings highlight DNA methylation as a unique epigenetic mark compared to common enhancer-associated histone marks. We highlight the unique accumulation of HMRs through developmental progression into clusters, enriched for stage-relevant SNP-based heritability. Through this process, epigenetic information can be maintained state to state. Thus, our results

support that the methylome presents a historical documentation of developmental choices which could assist in the prioritization and interpretation of SNP data associated with clinical traits and diseases. These conclusions may further assist in understanding the complex role of the methylome in development and epigenetic gene regulation.

METHODS

HMR selection/exclusion dataset

DNA HMRs were obtained through the *MethBase* DNA Methylation trackhub from the UCSC Genome Browser, which references data processed through the *MethPipe* software for processing bisulfite sequencing data (30). To achieve a high-confidence genome-wide methylation dataset, cell types were included based on a minimum coverage of 30x (83). This resulted in the selection of: *adrenal*, *fetal heart*, *fetal spinal cord*, *liver*, and *T cell* from the NIH Roadmap Epigenomics Consortium (18); *H1 ESC* from Lister, et al. (4); and *B cell* from Hodges, et al. (9). As a primary cleaning step, with interest in only non-coding HMRs, we removed promoter- and exon-overlapping HMRs. To do this, we combined RefSeq exon and protein-coding gene TSSs (-1000, +2000 bp) annotations to form an exclusion BED file. Next, we referenced this file to eliminate promoter- and exon-overlapping HMRs using the *intersect* function from the *Bedtools* package with option '-v' (84). Exclusion was defined by any basepair overlap. To increase the quality of our non-coding HMR dataset, we excluded HMRs based on a size minimum of 50bp. A subset of HMR datasets feature regions <8bp, suggesting artifacts from algorithmic noise, most likely incapable of featuring a functional TF binding site.

HMR dendrogram

This analysis was performed using the per-HMR average CpG numerical matrix as composed in the methylation heatmap analysis. We used the *hclust()* function and *ggdendro* package to perform hierarchical clustering and dendrogram composition in R.

Methylation heatmap

Heatmaps were generated in R with the package *pheatmap* (85). Numerical matrices representing per-HMR DNA methylation per cell type were used as input. These were generated in bash using methylation bigWig files from the *MethBase* DNA Methylation trackhub hosted on the UCSC Genome Browser. We used the *KentUtils* binary package to convert bigWig files to bedGraph files. CpGs that likely mark CTCF binding sites were excluded, as measured by genomic overlap

with a CTCF binding site as measured by Hoffman, et al. bedGraph score columns were used to populate a numerical matrix representing the sample-population methylation proportion at individual HMRs in rows. Heatmaps were generated using R package: *pheatmap* using options: *kmeans_k* = 7, *cluster_cols* = FALSE, *cuttree_rows* = 7.

Transcription factor motif enrichment analysis

The *HOMER* perl package was used to calculate transcription factor motif enrichment(86). Background regions were defined per specific cell type dataset as cell type HMR-subtracted whole-genome hypomethylated space, represented by the merged HMR datasets of other cell types. Raw hypergeometric values were used to rank motif enrichment output. *Scale fold enrichment* was calculated by the quotient of two *HOMER* output values: [%target/ %background]. Top representative TF from the top 15 results are displayed in **Fig 1**. All data is represented in **Fig 4** to visualize TF enrichment differences attributable to clustering. Data visualization is scaled by TF to show relative cell specific enrichment. Graphing was performed in *R* using the ggplot2 package (87).

k-means clustering gene ontology

Gene ontology was conducted using the web-based tool: *GREAT*(36). Specifically, *GREAT* assigns regulatory domains around gene TSSs, extending to the nearest gene's central domain up to a maximum extension distance. Here, we used the default gene annotation protocol from *GREAT* with a maximum extension of 1Mb. For a background file, we used a cumulative BED file representing all HMRs filtered for size and RefSeq promoters/exons from all cell types. Standard settings for maximum region-gene distance and gene assignment were used. Top results were downloaded from the web app using the "Shown ontology data as .tsv" selection. Output files were filtered to exclude the top row before import to R. Top ranked results are displayed as bar plots using ggplot2 and *geom_bar()*, ranked by binomial q-value.

Inter-HMR distances

To measure inter-HMR distances, we employed the *Bedtools closest* function with the '-io -d' options to calculate the distance from each HMR to the nearest HMR per cell type. Next, we extracted the output distance column to represent our observed distribution for graphing in R. To compare this distribution to random expectation, we used a script by Benton ML., 2018 which uses *Bedtools closest* to calculate distances between shuffled non-coding HMRs per cell type.

The *Bedtools closest* function takes two input files. For this analysis, the input dataset is submitted twice, though one input is shuffled. A region blacklist was used to exclude placement of HMRs during the shuffle into coding space, defined by RefSeq promoters and exons; this file was obtained from the HMR annotation step. We iterated the random shuffle-closest procedure 100,000 times to create a null expectation of genomic positioning given random placement. Distances per shuffle-closest were summarized as means, yielding a distribution of average distances per shuffle. Distributions were plotted using the ggplot2 R package.

Clustering percentage

To assess the prevalence of clustering per cell type, we utilized the same procedures for annotating HMR clustering as were used in the primary HMR annotation script. Here, we specifically defined clusters by their amount of constituent HMRs; again, clusters were permitted a maximum of 6kb between HMRs. We used *Bedtools merge* with the '-d' distance and '-c' count flags to generate a dataset of merge regions. Regions that composed merged regions were then identified using *Bedtools intersect* with the original file as the -a file and merged datasets as -b files. By isolating various merged regions by number of constitutive HMRs, we were able to identify the quantity of HMRs involved in varying degrees of clustering. Data was compiled and binned into 1, 2, or 3+ HMRs. Graphing was performed with the ggplot2 R package.

Sankey Diagram

HMR counts for each Sankey node and flow were determined using bash and the *Bedtools* suite. Nodes represent the total quantity of clustered and unclustered HMRs per cell type. Plots were generated in R using the package *networkD3*. To accurately represent the total quantity of HMRs per cell type, additional nodes were input and later processed with *Adobe Illustrator*.

Sankey gene ontology

Gene ontology was conducted using the web-based tool: *GREAT*(36), as with the *k*-means clustering gene ontology analysis. Here, we again used the default gene annotation protocol from *GREAT*. For a background file, we used the default "Whole genome" option. Standard settings for maximum region-gene distance and gene assignment were used. Top results were downloaded from the web app using the "Shown ontology data as .tsv" selection. Output files were filtered to exclude the top row before import to R, and the preceding "#" is removed from the

492 second row. Top results as ranked by binomial q-value are displayed as a bar plot using ggplot2
493 and *geom_bar*.

494 **Super enhancer annotation**

495 A super-enhancer catalog was downloaded from Whyte, et al. GM12878 SEs were isolated as a
496 text file before comparison with clustered and all B cell HMR datasets using *Bedtools intersect*.
497 Plotting was performed using the *R* package, *eulerr*, to maintain proportionally sized ellipses.

498 **ChromHMM annotation**

499 A ChromHMM 15-state annotation file was downloaded from the UCSC Genome Browser in hg19
500 as assayed in the GM12878 cell line. Intersections were assessed using *Bedtools intersect* and
501 B cell HMRs. Graphing was performed in *R* with the package, ggplot2.

502 **ATAC-STARR-seq comparison**

503 BED files for GM12878 ATAC-STARR-seq regulatory elements were obtained from Hansen &
504 Hodges (56) (GSE181317). HMRs were converted to GRCh38 for comparison using *liftOver*
505 (parameters: -bedPlus=3). We determined the number of overlaps between the datasets with
506 *Bedtools intersect* (default parameters) piped to a line count command (*wc -l*). The proportion of
507 overlapping HMRs was calculated as $[\#overlapping/\#total]$ and then plotted with *ggplot* in *R*. We
508 performed a two-tailed, two-sample Z test of proportions with the *prop.test* function in *R* to obtain
509 p-values.

510 **Nearest-neighbor transcriptional analysis**

511 To measure the transcriptional differences associated with clustered or unclustered HMRs, we
512 first identified a gene list with transcriptional values. To do this, we downloaded two replicate
513 RNA-seq datasets acquired from the ENCODEv3 release. Here, we elected to use data for
514 GM12878s, a lymphoblastoid cell line, to match B cells as closely as possible. Using the
515 *DESeqDataSetFromHTSeqData*, *DESeq*, *collapseReplicates*, and *counts* functions with default
516 options, we normalized replicate counts. Next, we combined *DESeq2* normalized count data with
517 genomic coordinates, for comparison with HMRs in BED format, using the gene set defined by
518 RefSeq *NM* genes. This file was downloaded from UCSC Genome Browser in *hg19* genome
519 version and referenced earlier in the HMR annotation script. Gene coordinates and normalized
520 counts were matched in *R* using the *merge* function. We also took this opportunity to add gene

symbol identifiers with the *bitr* function from the *ClusterProfiler* package (version 3.0.4) for easier interpretability in downstream analyses. Next, to annotate nearest neighbor, we used *Bedtools closest* with the ‘-d’ distance option to identify associated genes with clustered and unclustered HMR separately. Using the previously described RefSeq promoter/exon file, *awk*, and *Bedtools intersect* (‘-v’ option), we filtered HMR-TSS pairs to keep those that do not cross a promoter or exon boundary. These were then additionally filtered against a GM12878 TAD file, downloaded from the 3D genome browser provided by the Yue lab (<http://3dgenome.fsm.northwestern.edu/publications.html>), to eliminate pairs that cross a TAD boundary. An expression filter (normalized count >0) was applied in R to eliminate HMR-TSS pairs that are likely inactive and a false positive pair assignment. The *tidyverse* package was used to bin HMR-TSS distances. *ggplot2* and *ggpubr* were used for visualization in R.

S-LDSC

Stratified LD-score regression was performed with *LDSC* using the appropriate python scripts distributed by the Price lab (<https://github.com/bulik/ldsc>). Reference base annotation files were downloaded from the Price repository (Phase 3, version 2.2 annotations). We used the appropriate reference files coordinating with the 1000 Genomes baseline v2.2 scores and HapMap 3 SNPs (<https://alkesgroup.broadinstitute.org/LDSCORE/>). Summary statistics were collected from both the Price lab (https://alkesgroup.broadinstitute.org/LDSCORE/independent_sumstats/) as well as the Neale lab heritability repository (https://nealelab.github.io/UKBB_ldsc/index.html). Traits were obtained based on their determined relevance to either broad cell-agnostic etiology or to biology specifically relating to B cells or Liver. This provided us the ability to determine specificity of results associated with varying cell specificity of HMRs. In total, we assessed 79 traits as described in **S3 Table**. The primary *LDSC* program was run per HMR annotation per trait. Results for individual traits were tabularized per HMR annotation. Results were visualized using *ggplot* in R with the functions *geom_point* and *case_when* for conditional coloring. To determine B cell developmentally derived HMRs, we used *Bedtools intersect* to compare HMR files. H1 ESC-derived B cell HMRs were defined by B cell HMR coordinates and had to have overlap with HMRs from HSPC as well as H1 ESC, together. HSPC-derived B cell HMRs had to have overlap with HSPC HMRs but not H1 ESC HMRs. B cell specific HMRs had to have no overlap with any HMRs from the collection of adrenal gland, liver, fetal heart, fetal spinal cord, H1 ESC, HSPC, Macrophage, and T cell. In the clustering analysis, all clustered or all unclustered HMRs were used. Liver HMRs were defined as H1 ESC-derived based on any overlap with H1 ESC HMRs. Cell-specific Liver HMRs were also defined against the same

comparative cell type collection used with B cell for this specific analysis. Annotations used to compare against HMR groups were selected from those included in the “baselineLF_v2.2.UKB.tar.gz” from the Price lab LD-score website. Annotations were selected for their relevance to enhancers; CTCF, a ubiquitous transcription factor, was included as a negative control for cell-specific disease enrichment.

WebGestalt Gene Ontology Analysis

Developmentally grouped B cell HMR BED files, as used in the S-LDSC analysis, were used as input in addition to BED files for all B cell clustered or all B cell unclustered HMRs. GREAT Input was used to identify nearest neighbor genes in hg19 (36). We used the default gene assignment parameters under “Association rule settings” called “Basal plus extension,” which in most cases replicates a two-nearest neighbor gene association strategy. In the web app, we downloaded the “Gene -> genomic region association table” file from the “genomic region-gene associations” page. Gene symbols were extracted from the GREAT input downloaded files using the first column. These were input into the WebGestalt web app to perform an over-representation analysis on the disease functional database, GLAD4U (70, 71, 88). For a reference gene set, we selected *genome protein-coding*. Results were downloaded, and the enrichment values file was used to plot enrichment ratio values for top diseases in R using ggplot2.

NHGRI trait enrichment analysis

For SNP-associated trait enrichment analysis, we first obtained SNP-trait associations with the NHGRI SNP-trait catalog in *hg38* for information completeness (<https://www.ebi.ac.uk/gwas/docs/file-downloads>). HMR datasets were converted from *hg19* to *hg38* genome build versions using the *liftOver* tool from UCSC. *Bedtools intersect* was used to identify HMR-overlapping SNPs and associated traits. Resulting SNP-trait pairs were tabulated per HMR annotation group. The expected SNP-trait distribution was informed by the entire NHGRI catalog. Fold change was calculated against this null distribution. P-values were calculated using a binomial test to account for replacement and multiple SNPs associating with a single trait. Top results were ranked by p-value as displayed in the manuscript. Code was written in *Python* and graphing was performed in *R* using *ggplot*.

COMPETING INTEREST STATEMENT

None declared.

584 **ACKNOWLEDGEMENTS**

585 We thank Tony Capra, Lea Davis, Melinda Aldrich, Eric Gamazon, Mary Philip, and members of
586 the Hodges Lab for helpful feedback and discussions. We are grateful for support of the project
587 and the time invested in producing this manuscript by NIH awards [1R01GM147078-01 to E.H,
588 T32-GM080178-14 to T.S.], Department of Defense Idea Award [W81XWH-20-1-0522 to E.H],
589 and American Cancer Society (ACS) Institutional Research Grant (#IRG-15-169-56).

590 REFERENCES

- 591 1. Van der Ploeg L, Flavell R. DNA methylation in the human $\gamma\delta\beta$ -globin locus in erythroid
592 and nonerythroid tissues. *Cell*. 1980;19(4):947-58.
- 593 2. Gruenbaum Y, Stein R, Cedar H, Razin A. Methylation of CpG sequences in eukaryotic
594 DNA. *FEBS letters*. 1981;124(1):67-71.
- 595 3. Kunnath L, Locker J. Characterization of DNA methylation in the rat. *Biochimica et*
596 *Biophysica Acta (BBA)-Gene Structure and Expression*. 1982;699(3):264-71.
- 597 4. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA
598 methylomes at base resolution show widespread epigenomic differences. *Nature*.
599 2009;462(7271):315-22.
- 600 5. Gu J, Stevens M, Xing X, Li D, Zhang B, Payton JE, et al. Mapping of variable DNA
601 methylation across multiple cell types defines a dynamic regulatory landscape of the human
602 genome. *G3: Genes, Genomes, Genetics*. 2016;6(4):973-86.
- 603 6. Gama-Sosa MA, Midgett RM, Slagel VA, Githens S, Kuo KC, Gehrke CW, et al. Tissue-
604 specific differences in DNA methylation in various mammals. *Biochimica et Biophysica Acta*
605 *(BBA)-Gene Structure and Expression*. 1983;740(2):212-9.
- 606 7. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale
607 DNA methylation maps of pluripotent and differentiated cells. *Nature*. 2008;454(7205):766-70.
- 608 8. Portela A, Esteller M. Epigenetic modifications and human disease. *Nature biotechnology*.
609 2010;28(10):1057-68.
- 610 9. Hodges E, Molaro A, Dos Santos CO, Thekkat P, Song Q, Uren PJ, et al. Directional DNA
611 methylation changes and complex intermediate states accompany lineage specificity in the adult
612 hematopoietic compartment. *Molecular cell*. 2011;44(1):17-28.
- 613 10. Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, et al. Sperm
614 methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*.
615 2011;146(6):1029-41.
- 616 11. Bock C, Beerman I, Lien W-H, Smith ZD, Gu H, Boyle P, et al. DNA methylation dynamics
617 during in vivo differentiation of blood and skin stem cells. *Molecular cell*. 2012;47(4):633-47.
- 618 12. Barnett KR, Decato BE, Scott TJ, Hansen TJ, Chen B, Attalla J, et al. ATAC-Me captures
619 prolonged DNA methylation of dynamic chromatin accessibility loci during cell fate transitions.
620 *Molecular cell*. 2020;77(6):1350-64. e6.
- 621 13. He Y, Hariharan M, Gorkin DU, Dickel DE, Luo C, Castanon RG, et al. Spatiotemporal
622 DNA methylome dynamics of the developing mouse fetus. *Nature*. 2020;583(7818):752-9.
- 623 14. Razin A, Szyf M. DNA methylation patterns formation and function. *Biochimica et*
624 *Biophysica Acta (BBA)-Gene Structure and Expression*. 1984;782(4):331-42.
- 625 15. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, et al. Distribution,
626 silencing potential and evolutionary impact of promoter DNA methylation in the human genome.
627 *Nature genetics*. 2007;39(4):457-66.
- 628 16. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation
629 profiling of human chromosomes 6, 20 and 22. *Nature genetics*. 2006;38(12):1378-85.
- 630 17. Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, et al. Intragenic DNA
631 methylation prevents spurious transcription initiation. *Nature*. 2017;543(7643):72-7.
- 632 18. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative
633 analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-30.
- 634 19. Schlesinger F, Smith AD, Gingeras TR, Hannon GJ, Hodges E. De novo DNA
635 demethylation and noncoding transcription define active intergenic regulatory elements. *Genome*
636 *research*. 2013;23(10):1601-14.

20. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*. 2015;523(7559):212-6.
21. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*. 2011;480(7378):490-5.
22. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*. 2013;500(7463):477-81.
23. Mendizabal I, Shi L, Keller TE, Konopka G, Preuss TM, Hsieh T-F, et al. Comparative methylome analyses identify epigenetic regulatory loci of human brain evolution. *Molecular biology and evolution*. 2016;33(11):2947-59.
24. Yuan W, Xia Y, Bell CG, Yet I, Ferreira T, Ward KJ, et al. An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. *Nature communications*. 2014;5(1):1-7.
25. Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*. 2009;106(23):9362-7.
26. Li Y, Rivera CM, Ishii H, Jin F, Selvaraj S, Lee AY, et al. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PloS one*. 2014;9(12):e114485.
27. Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, Dang MD, et al. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature genetics*. 2013;45(10):1198-206.
28. Bell E, Curry EW, Megchelenbrink W, Jouneau L, Brochard V, Tomaz RA, et al. Dynamic CpG methylation delineates subregions within super-enhancers selectively decommissioned at the exit from naive pluripotency. *Nature communications*. 2020;11(1):1-16.
29. Molaro A, Falcioni I, Hodges E, Aravin AA, Marran K, Rafii S, et al. Two waves of de novo methylation during mouse germ cell development. *Genes & development*. 2014;28(14):1544-9.
30. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PloS one*. 2013;8(12):e81148.
31. Phillips JE, Corces VG. CTCF: master weaver of the genome. *Cell*. 2009;137(7):1194-211.
32. Rodda DJ, Chew J-L, Lim L-H, Loh Y-H, Wang B, Ng H-H, et al. Transcriptional regulation of nanog by OCT4 and SOX2. *Journal of Biological Chemistry*. 2005;280(26):24731-7.
33. Kistler B, Pfisterer P, Wirth T. Lymphoid-and myeloid-specific activity of the PU. 1 promoter is determined by the combinatorial action of octamer and ets transcription factors. *Oncogene*. 1995;11(6):1095-106.
34. Chen J. The proto-oncogene c-ets is preferentially expressed in lymphoid cells. *Molecular and cellular biology*. 1985;5(11):2993-3000.
35. Åkerblad P, Sigvardsson M. Early B cell factor is an activator of the B lymphoid kinase promoter in early B cell development. *The Journal of Immunology*. 1999;163(10):5453-61.
36. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*. 2010;28(5):495-501.
37. Ji H, Ehrlich LI, Seita J, Murakami P, Doi A, Lindau P, et al. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*. 2010;467(7313):338-42.
38. Bröske A-M, Vockentanz L, Kharazi S, Huska MR, Mancini E, Scheller M, et al. DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. *Nature genetics*. 2009;41(11):1207-15.

39. Hendriks J, Gravestien LA, Tesselaar K, van Lier RA, Schumacher TN, Borst J. CD27 is required for generation and long-term maintenance of T cell immunity. *Nature immunology*. 2000;1(5):433-40.
40. Agematsu K, Hokibara S, Nagumo H, Komiyama A. CD27: a memory B-cell marker. *Immunology today*. 2000;21(5):204-6.
41. Lens SM, Tesselaar K, van Oers MH, van Lier RA, editors. Control of lymphocyte function through CD27–CD70 interactions. *Seminars in immunology*; 1998: Elsevier.
42. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*. 2013;153(2):307-19.
43. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. *Cell*. 2013;155(4):934-47.
44. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*. 2013;153(2):320-34.
45. Sabari BR, Dall'Agnese A, Boija A, Klein IA, Coffey EL, Shrinivas K, et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science*. 2018;361(6400):eaar3958.
46. Vahedi G, Kanno Y, Furumoto Y, Jiang K, Parker SC, Erdos MR, et al. Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature*. 2015;520(7548):558-62.
47. Parker SC, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences*. 2013;110(44):17921-6.
48. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*. 2012;9(3):215-6.
49. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nature protocols*. 2017;12(12):2478-92.
50. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*. 2007;39(3):311-8.
51. Creighton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*. 2010;107(50):21931-6.
52. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*. 2009;457(7231):854-8.
53. Capra JA. Extrapolating histone marks across developmental stages, tissues, and species: an enhancer prediction case study. *BMC genomics*. 2015;16(1):1-9.
54. Huang J, Li K, Cai W, Liu X, Zhang Y, Orkin SH, et al. Dissecting super-enhancer hierarchy based on chromatin interactions. *Nature communications*. 2018;9(1):1-12.
55. Fraser P, Pruzina S, Antoniou M, Grosveld F. Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. *Genes & development*. 1993;7(1):106-13.
56. Hansen TJ, Hodges E. ATAC-STARR-seq reveals transcription factor-bound activators and silencers within chromatin-accessible regions of the human genome. *Genome Research*. 2022;32(8):1529-41.
57. Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nature communications*. 2018;9(1):1-15.
58. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature genetics*. 2019;51(12):1664-9.

59. Consortium EP. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology*. 2011;9(4):e1001046.
60. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
61. Roederer M, Quaye L, Mangino M, Beddall MH, Mahnke Y, Chattopadhyay P, et al. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. *Cell*. 2015;161(2):387-403.
62. Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat Genet*. 2017.
63. Guo MH, Nandakumar SK, Ulirsch JC, Zekavat SM, Buenrostro JD, Natarajan P, et al. Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. *Proc Natl Acad Sci U S A*. 2017;114(3):E327-E36.
64. Vockley CM, Barrera A, Reddy TE. Decoding the role of regulatory element polymorphisms in complex disease. *Curr Opin Genet Dev*. 2017;43:38-45.
65. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-5.
66. Koues OI, Kowalewski RA, Chang LW, Pyfrom SC, Schmidt JA, Luo H, et al. Enhancer sequence variants and transcription-factor deregulation synergize to construct pathogenic regulatory circuits in B-cell lymphoma. *Immunity*. 2015;42(1):186-98.
67. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*. 2016;167(5):1398-414.e24.
68. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2015;518(7539):337-43.
69. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*. 2015;47(11):1228-35.
70. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research*. 2005;33(suppl_2):W741-W8.
71. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic acids research*. 2019;47(W1):W199-W205.
72. Kuroda A, Rauch TA, Todorov I, Ku HT, Al-Abdullah IH, Kandeel F, et al. Insulin gene expression is regulated by DNA methylation. *PloS one*. 2009;4(9):e6953.
73. Schmidl C, Klug M, Boeld TJ, Andreesen R, Hoffmann P, Edinger M, et al. Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome research*. 2009;19(7):1165-74.
74. Bloushtain-Qimron N, Yao J, Snyder EL, Shipitsin M, Campbell LL, Mani SA, et al. Cell type-specific DNA methylation patterns in the human breast. *Proceedings of the National Academy of Sciences*. 2008;105(37):14076-81.
75. Deaton AM, Webb S, Kerr AR, Illingworth RS, Guy J, Andrews R, et al. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome research*. 2011;21(7):1074-86.
76. Park I-H, Wen B, Murakami P, Aryee MJ, Irizarry R, Herb B, et al. Differential methylation of tissue-and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nature genetics*. 2009;41(12):1350-3.
77. Frigola J, Song J, Stirzaker C, Hinshelwood RA, Peinado MA, Clark SJ. Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nature genetics*. 2006;38(5):540-9.

78. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*. 2009;41(2):178-86.
79. Bell JT, Tsai P-C, Yang T-P, Pidsley R, Nisbet J, Glass D, et al. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS genetics*. 2012;8(4):e1002629.
80. Hotta K, Kitamoto A, Kitamoto T, Ogawa Y, Honda Y, Kessoku T, et al. Identification of differentially methylated region (DMR) networks associated with progression of nonalcoholic fatty liver disease. *Scientific reports*. 2018;8(1):1-11.
81. Schulz M, Teissandier A, de la Mata E, Armand M, Iranzo J, El Marjou F, et al. DNA methylation restricts coordinated germline and neural fates in embryonic stem cell differentiation. *bioRxiv*. 2022:2022.10.22.513040.
82. Guerin LN, Barnett KR, Hodges E. Dual detection of chromatin accessibility and DNA methylation using ATAC-Me. *Nature Protocols*. 2021;16(12):5377-97.
83. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu Y, et al. The UCSC genome browser database. *Nucleic acids research*. 2003;31(1):51-4.
84. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-2.
85. Kolde R. Pheatmap: pretty heatmaps. R package version. 2012;1(2):726.
86. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*. 2010;38(4):576-89.
87. Wickham H. Package 'ggplot2': elegant graphics for data analysis. Springer-Verlag New York doi. 2016;10:978-0.
88. Jourquin J, Duncan D, Shi Z, Zhang B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC genomics*. 2012;13(8):1-12.
89. Phanstiel DH, Boyle AP, Araya CL, Snyder MP. Sushi. R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics*. 2014;30(19):2808-10.

FIGURE LEGENDS

Figure 1: Levels of HMR specificity recapitulate developmental relationships.

(A) Multiple alignment around the CD27 locus of methylation and HMR tracks across 7 cell types: H1 ESC, fetal spinal cord, fetal heart, adrenal gland, liver, B cell, and T cell. Methylation tracks are represented by orange vertical bars showing methylation value per CpG site. HMRs are shown by dark blue horizontal bars. Developmentally constitutive, lineage-specific, and cell-specific HMRs are highlighted. Sushi R package was used to generate genome browser snapshot (89). (B) Dendrogram of HMRs by average methylation per cell type through hierarchical clustering. (C) Heatmap of average methylation per HMR across cell types. Non-coding HMRs were k-means clustered based on their average CpG methylation values across 7 cell types represented in (A). A k-means of 7, assessed by the elbow method, was used to cluster HMRs into groups that are consistent with the biological relationships of their cell types. Groups are manually labeled to reflect their biological relationships. (D) The transcription factor (TF) motif enrichment of each k-means group reflects biological relationships captured in (A). Representative TFs were selected from top significant hits filtered by Benjamini-Hochberg adjusted p-value. Fold enrichment values are normalized from 0 to 1 across TFs. The background comparison file comprises HMRs across all represented cell types. (E) GREAT gene ontology enrichments are shown for selected groups from the heatmap in (C). Top 7 results by hypergeometric q-value are displayed. Bar plots show ontologies and q-values.

Figure 2: HMRs accumulate and persist through differentiation.

(A) Bar graph of the total number HMRs for each cell type, arranged by developmental progression. (B) Bar graph measuring the presence of HMRs established in either H1 ESCs (*top; green*) or HSPCs (*bottom; grey*) in developmentally progressive cell types. The software Bedtools *intersect* was used to determine overlap between cell type HMR datasets using default settings. Overlap was defined as a 1bp minimum.

Figure 3: HMRs cluster more than expected.

(A) Distribution plots of inter-HMR distances by cell type. The green distributions represent observed values from HMR datasets per cell type. Vertical black bars show median values. Gray distributions show expected values by random shuffling across the non-coding genome. (B) Diagram of HMR clustering and cell specificity workflow. HMR datasets are cleaned, removing regions that overlap RefSeq protein-coding promoters (TSS +2000/-1000) and exons. HMRs can be annotated for cell specificity and/or clustering behavior. Clustering refers to groups of HMRs in a cell type that are located a maximum of 6kb end-to-end from the next HMR, linking 3 or more HMRs. Cell specificity can also be annotated, with any base pair overlap between HMRs constituting overlap. (C) Bar graph of HMR clustering annotations discussed in (B) as percentages of total HMRs by cell type. Selected cell types represent members of the hematopoietic and hepatic lineages. Colors reflect cell types representing different developmental stages and lineages. (D) Bar graph of proportion of cell type HMRs that are clustered HMR. (E) Sankey diagram showing the flow of B cell HMRs. B cell HMRs are divided on the right of the panel into clustering groups. The left shows HSPC HMRs that overlap B cell HMRs, and are hierarchically categorized as *clustered HSPC HMR*, *unclustered HSPC HMR*, *shared*, or *cell specific*. To define cell specificity, B cell HMRs were compared to datasets from *adrenal gland*, *H1 ESC*, *HSPC*, *fetal spinal*, *fetal heart*, *liver*, *macrophage*, and *T cell*. (F) The bar graph shows the top biological process gene ontology results for the Sankey group of HMRs that progress from *HSPC unclustered* to *B cell clustered* (indicated in red). Results from GREAT Gene Ontology using

default background and gene assignment settings are represented by bars showing binomial q-value.

Figure 4: Clustered HMRs show distinct enhancer-associated characteristics compared to unclustered HMRs.

(A) Venn diagram showing partially overlapping sets between three region datasets: All B cell HMRs (blue line); clustered B cell HMRs (green line; subset of All); and GM12878 super-enhancers (solid red circle) (42). (B) Bar graph of HMR overlap with selected ChromHMM annotations: *strong enhancer*, *weak enhancer*, *heterochromatin*, *repressed*, *insulator*, *weak transcription*. The height of the bars represents the fraction of clustered and unclustered HMRs that overlap each annotation. (C) TF motif enrichment in clustered (left; green) and unclustered (right; blue) HMRs. Results are plotted as $-\log_{10}p$ -value by fold enrichment, measured as percentage of target regions containing motif divided by the percentage of background regions. Background represents all clustered and unclustered HMRs. (D) Boxplot of ATAC-STARR-seq regulatory element overlap by clustered and unclustered HMRs. Overlap is measured at the unit of HMRs, and values depict fraction of total HMRs that contain a regulatory element. A Wilcoxon rank sum test was used to determine statistical significance. (E) Boxplot of normalized read counts of nearest neighbor RefSeq protein-coding genes to clustered and unclustered HMRs. Nearest neighbor genes were filtered for TAD boundary crossing. Results are displayed for (E) all genes, or (F) genes binned by distance between the HMR and nearest gene. Statistical significance was measured by a Wilcoxon rank sum test (* = <0.05 ; ** = $<5 \times 10^{-5}$; *** = $<5 \times 10^{-10}$).

Figure 5: S-LDSC identifies HMR annotation-specific trait enrichments.

(A) Volcano-style plots of S-LDSC partitioned heritability results across 79 traits are shown for three B cell HMR groups: *H1 ESC-derived*, *HSPC-derived*, and *cell-specific*. HMRs are ordered by the developmentally distinct cell type in which they were established. Each HMR group was tested for enrichment of genetic heritability with a standard set of 98 base annotations against traits that include both clinical diseases as well as clinical lab values. Negative enrichment values were clipped to the lowest positive enrichment value for each row of plots (A: 0.1174537; B: 0.25754925). The size of each point represents the $-\log_{10}p$ -value of the enrichment, and the color shows the $\log_{10}$ enrichment value. Points with a p -value ≤ 0.05 or an enrichment > 10 are labeled by their trait name where available. (B) Further partitioned heritability analysis applied to B cell HMRs grouped only by clustering behavior is also represented. (C) Point and line plots of S-LDSC enrichment values by annotation group per trait. These graphs include data from developmentally-derived B cell HMRs compared against other enhancer-associated groups, including ancient human enhancer sequence age, FANTOM 5 enhancers, eQTLs, super enhancers, and the H3K27ac histone mark. Genomic controls were also included, such as phastCons 46-way annotations as well as promoters and CTCF. The x-axis represents enrichment values, and the y-axis displays genomic annotations. Points show enrichment point estimates and lines display 95% confidence intervals. The red dotted line marks an enrichment score of 0.

SUPPLEMENTARY FIGURE LEGENDS

S1 Fig. Dotplot of elbow method to determine appropriate number of k-means for methylation heatmap.

Figure displays within sum of squares estimates for clusters at each value of k-means group amount from 1 to 12. Estimates are derived from the *kmeans()* function in R.

S2 Fig. HMR cluster lengths are consistent across cell types.

The graph shows the lengths of HMR clusters, end-to-end, per cell type. Data is represented by both a violin plot and boxplot. The boxplot shows the interquartile range, and the bold black line shows the median value per cell type. The red dotted line shows the value 10,500 bp, which approximates the mean cluster length of 10,551.2 bp, measured across the cell types: *adrenal gland*, *fetal heart*, *fetal spinal cord*, *H1 ESC*, and *Liver*.

S3 Fig. Euler plot comparing B cell HMRs and open chromatin.

(A) Euler plot of all B cell HMRs and open chromatin defined by DNase I hypersensitivity sites in GM12878 cells. The DNase file was downloaded from the UCSC Genome Browser Table Browser using the following main settings: *clade*: "mammal"; *genome*: "human"; *assembly*: "Feb 2009 (GRCh37/hg19)"; *group*: "Regulation"; *track*: "Duke DNaseI HS"; *table*: "GM12878 Pk (wgEncodeOpenChromDnaseGm12878Pk)" [ENCODE file ID: ENCFF001UVC]. (B-C) Boxplot of normalized read counts of nearest neighbor RefSeq protein-coding genes to clustered and unclustered Liver HMRs. Nearest neighbor genes were filtered for TAD boundary crossing. Results are displayed for (B) all genes, or (C) genes binned by distance between the HMR and nearest gene. Statistical significance was measured by a Wilcoxon rank sum test (* = <0.05; ** = <5x10⁻⁵; *** = <5x10⁻¹⁰).

S4 Fig. S-LDSC identifies HMR annotation-specific trait enrichments in Liver.

(A) Volcano-style plots of S-LDSC partitioned heritability results across 79 traits are shown for two *Liver* HMR groups: *H1 ESC-derived* and *cell-specific*. HMRs are ordered by the developmentally distinct cell type in which they were established. Each HMR group was tested for enrichment of genetic heritability with a standard set of 98 base annotations against traits that include both clinical diseases as well as clinical lab values. Negative enrichment values were clipped to the lowest positive enrichment value for each row of plots (A: 0.02781896; B: 0.03787533). The size of each point represents the -log₁₀*p*-value of the enrichment, and the color shows the log₁₀enrichment value. Points with a *p*-value ≤ 0.05 or an enrichment > 10 are labeled by their trait name where available.

S5 Fig. Disease ontology for developmentally specific and clustered B cell HMRs.

Lollipop plots show top ten disease ontology enrichments as analyzed through WebGestalt with default parameters. The x-axis shows enrichment ratios, and the y-axis displays disease ontologies sourced from the GLAD4U disease database. The y-axis is sorted by enrichment value. The color for each bar represents the *p*-value for that trait. Individual graphs show results from B cell HMR developmental and clustering groups: (A) *H1 ESC-derived*, (B) *HSPC-derived*, (C) cell specific, (D) clustered and (E) unclustered.

S6 Fig. S-LDSC B cell by trait across genomic annotations.

Point and line plots of S-LDSC enrichment values by annotation group per trait. The x-axis represents enrichment values, and the y-axis displays genomic annotations. Points show enrichment point estimates and lines display 95% confidence intervals. The red dotted line marks an enrichment score of 0. Annotation groups include popular enhancer-associated groups such as ancient human enhancer sequence age, FANTOM 5 enhancers, eQTLs, super enhancers, and the H3K27ac histone mark. Genomic controls were also included, such as phastCons 46-way annotations as well as promoters and CTCF. These graphs include data from (A)

947 developmentally derived B cell HMRs. (B) This graph shows S-LDSC results for alanine
948 transaminase. The data includes the annotations from (A) in addition to developmentally derived
949 Liver HMRs.

950 **S1 Table. Inter-HMR lengths by cell type.**

951 The table shows the distance between HMRs per cell type. Values represent the inter-HMR
952 distances at different quantiles. The values have also been calculated at the .45 and .55 quantile
953 values for more specificity around the median. Max values were limited to a maximum of 50kbp.

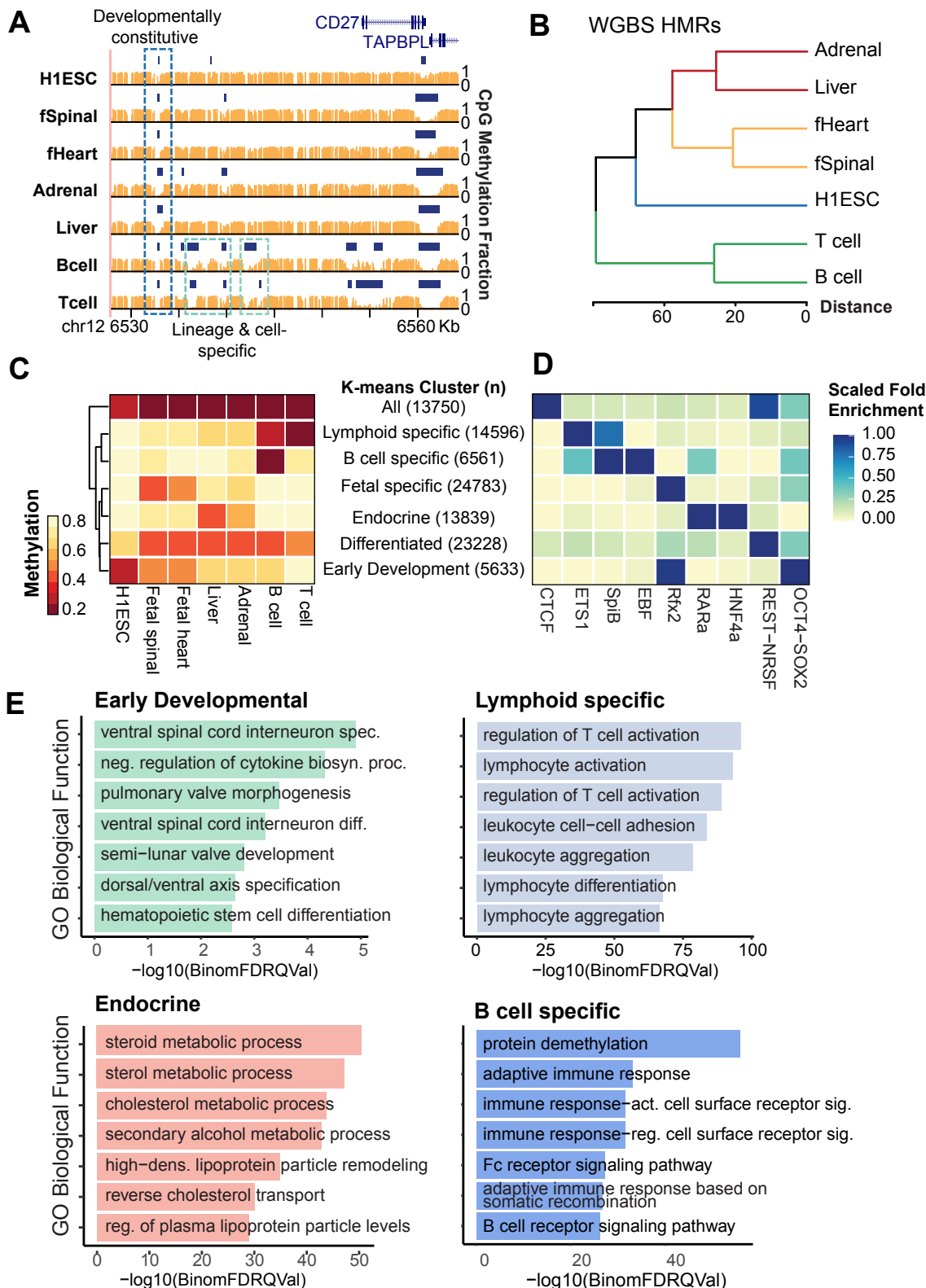
954 **S2 Table. Clustering group counts by clustering distance.**

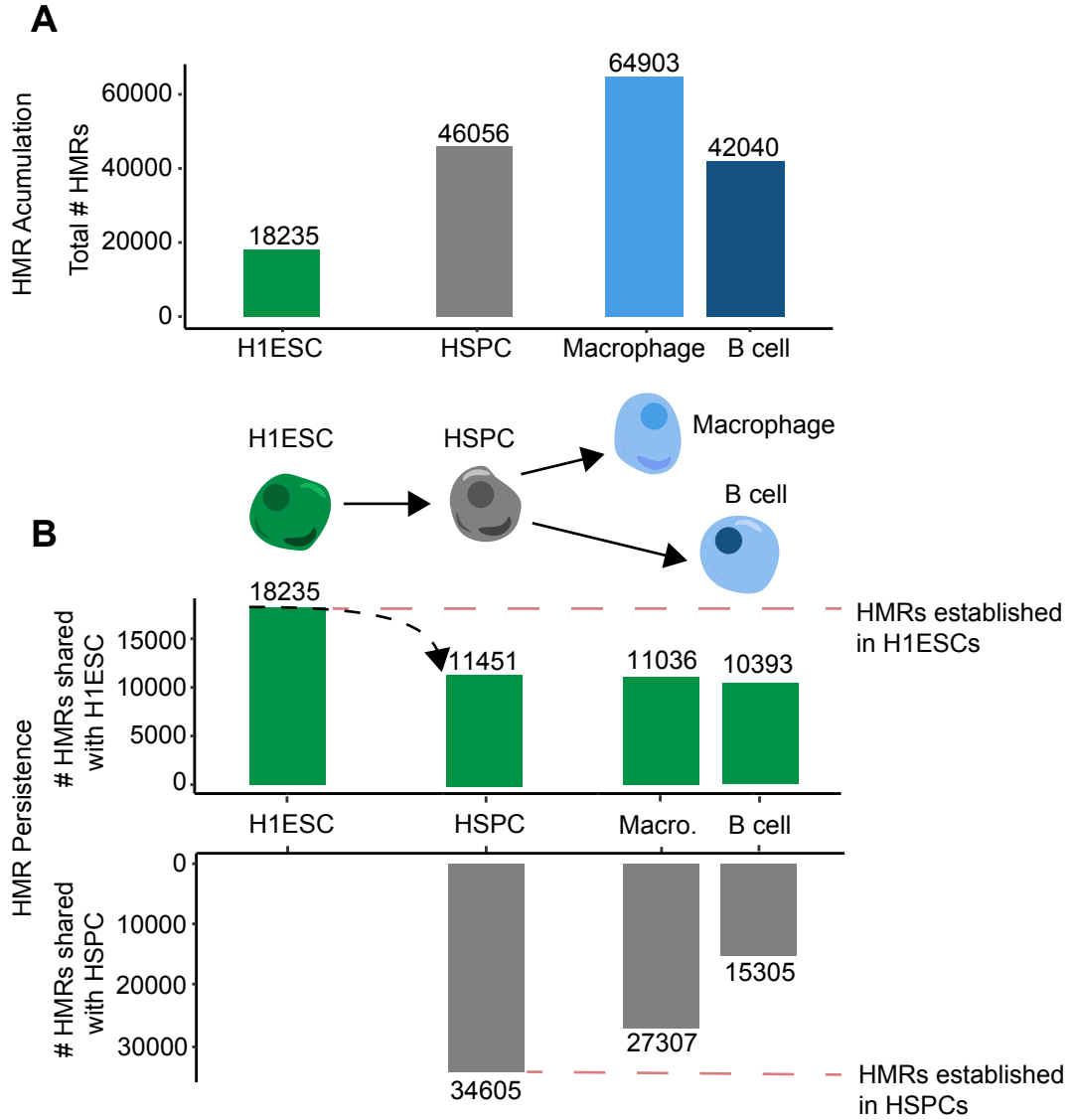
955 Table of HMR counts for clustering groups: unclustered; clusters of 2 HMRs where one is a
956 RefSeq promoter; clusters of 2 non-coding HMRs; and clusters of 3+ HMRs. Linking distances
957 include 6kb (used for our analyses); 10.5kb (approximate mean inter-HMR length); and 12.5kb
958 (distance commonly used in super-enhancer definitions applied to histone modification ChIP-seq
959 datasets).

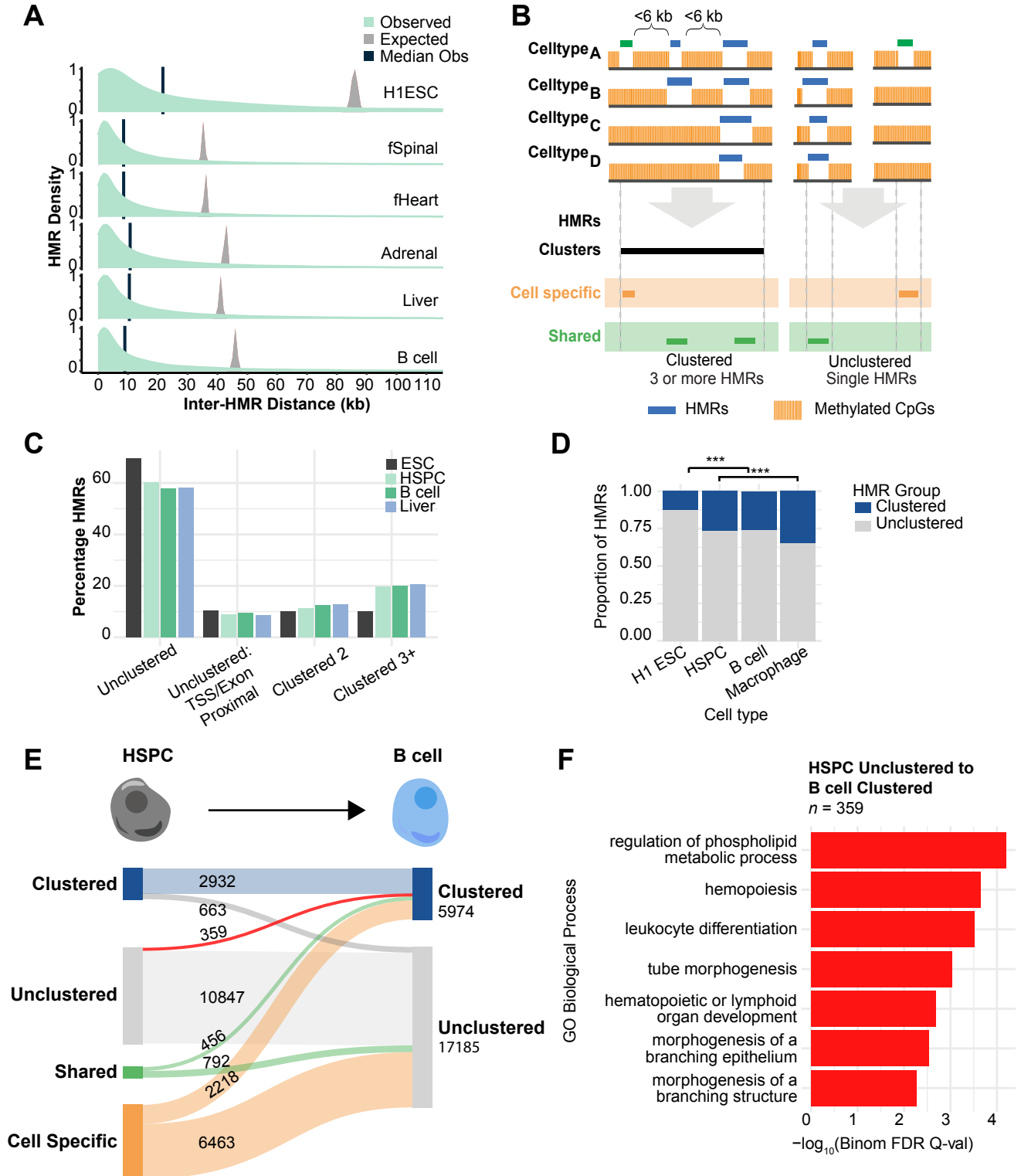
960 **S3 Table. List of 79 summary statistics files used for S-LDSC analyses.**

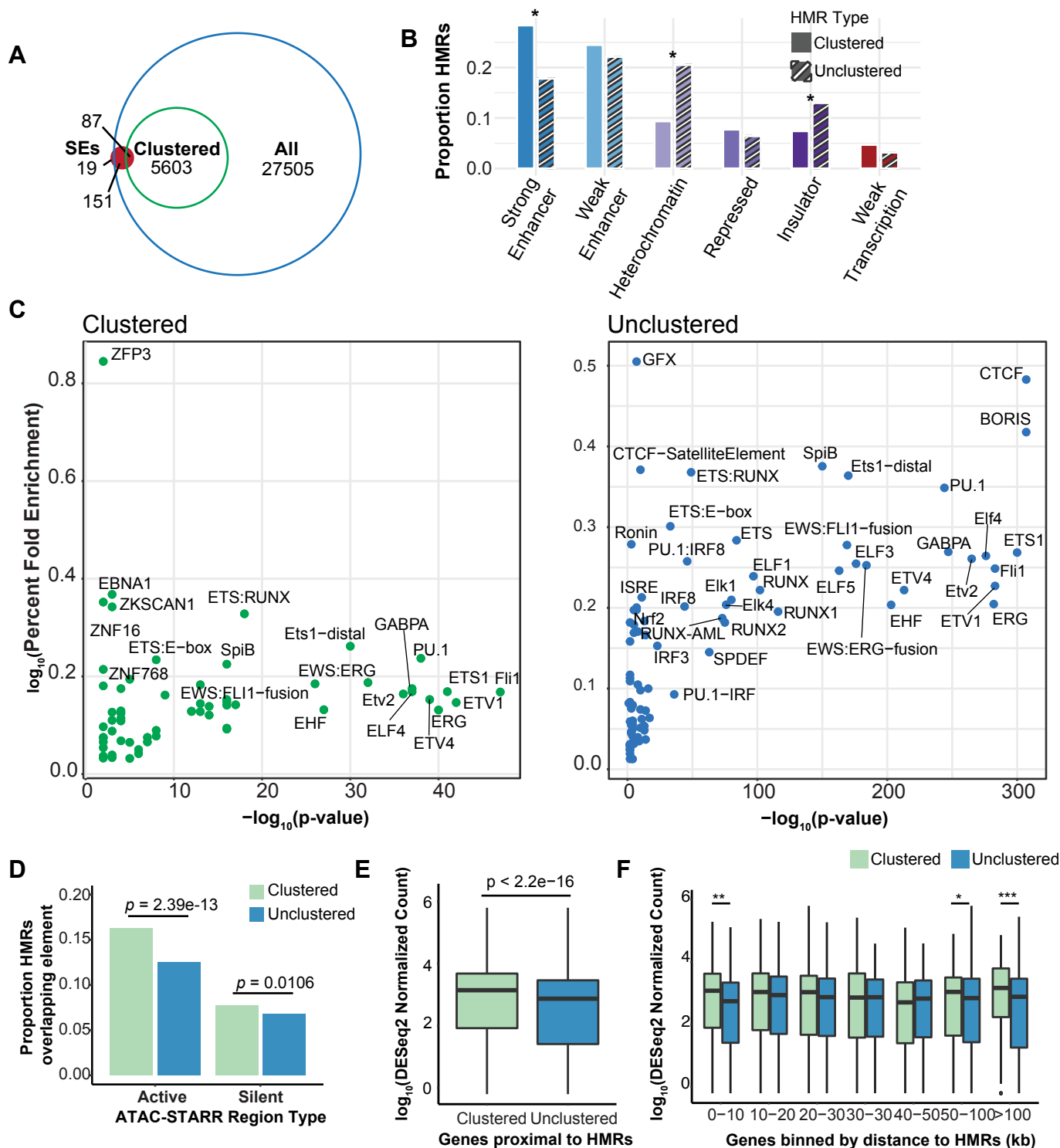
961 Table of trait names used in S-LDSC. Files were sourced from the Alkes Price lab website or
962 Neale lab LDSC summary statistics browser.

963









A

H1 ESC-derived
n = 10,214

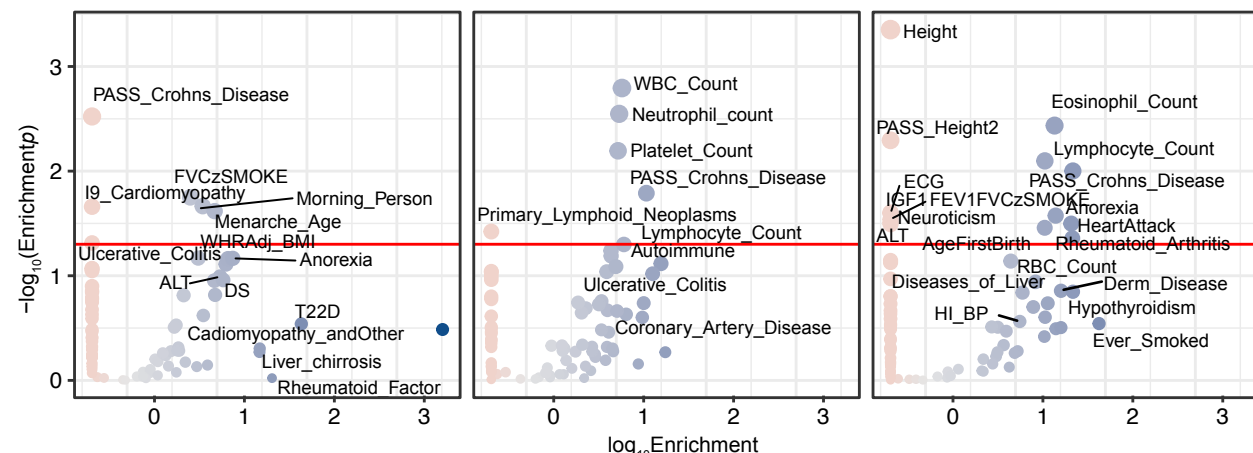
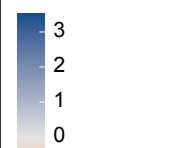
HSPC-derived
n = 15,414

B cell Specific
n = 4,442

$-\log_{10}(\text{Enrichment}p)$

● 1
● 2
● 3

$\log_{10}\text{Enrichment}$



B

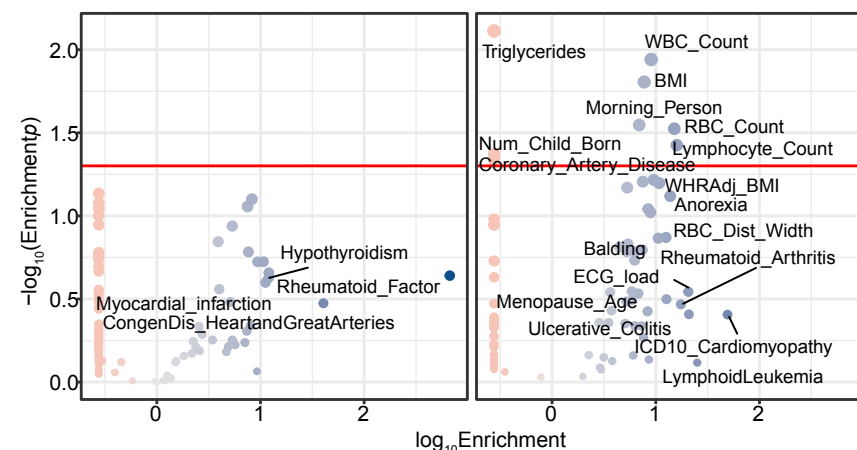
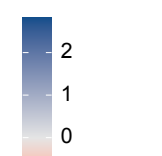
B cell Unclustered
n = 17,185

B cell Clustered
n = 5,974

$-\log_{10}(\text{Enrichment}p)$

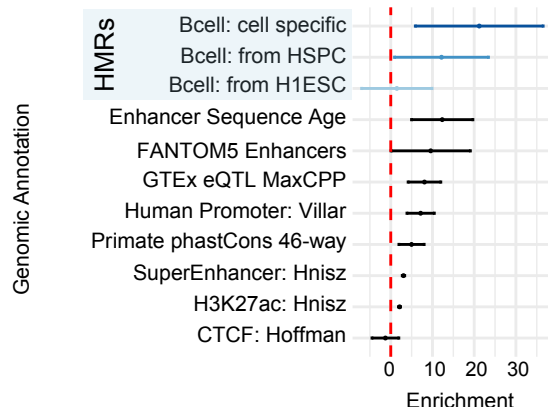
● 0.5
● 1.0
● 1.5
● 2.0

$\log_{10}\text{Enrichment}$



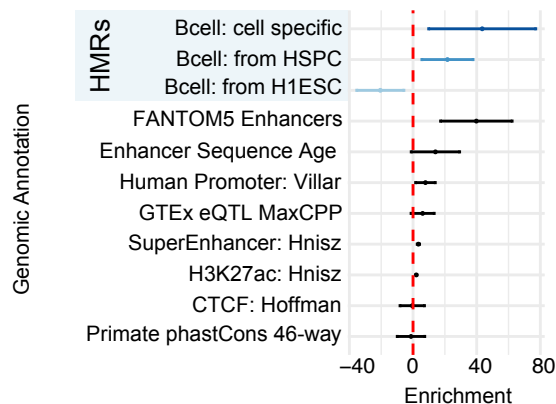
C

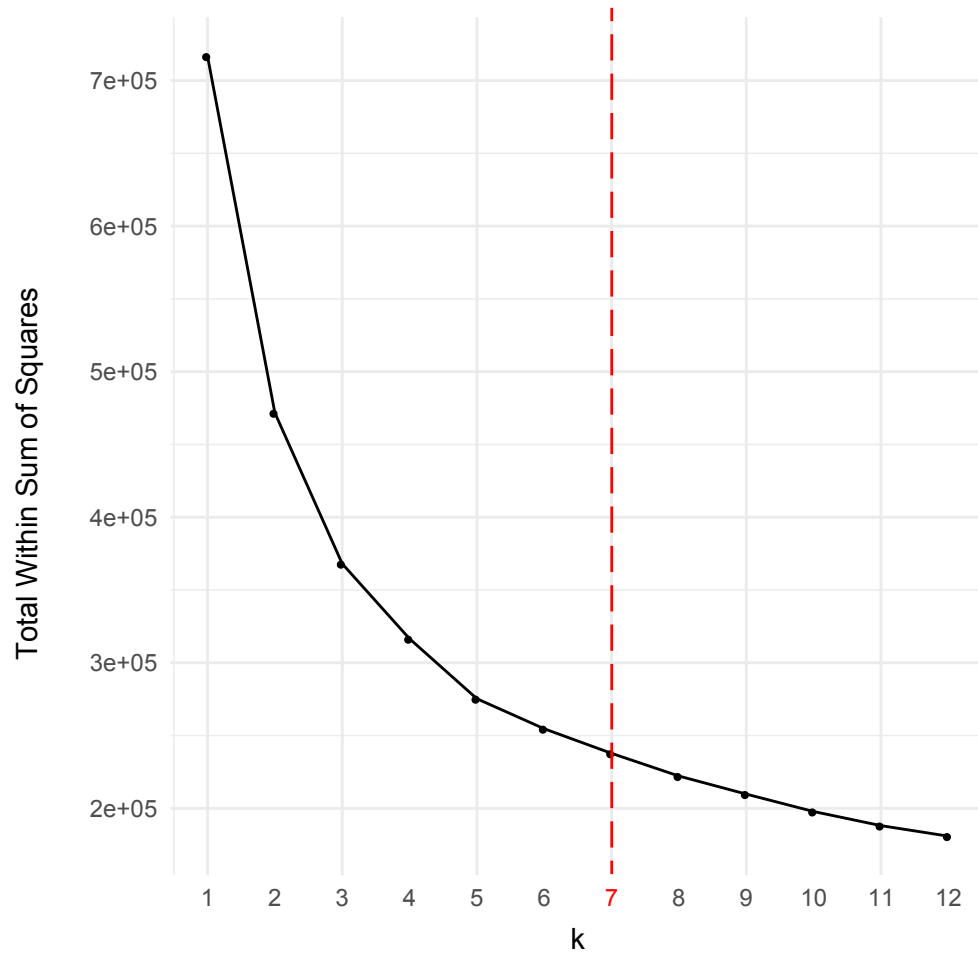
Lymphocyte Count

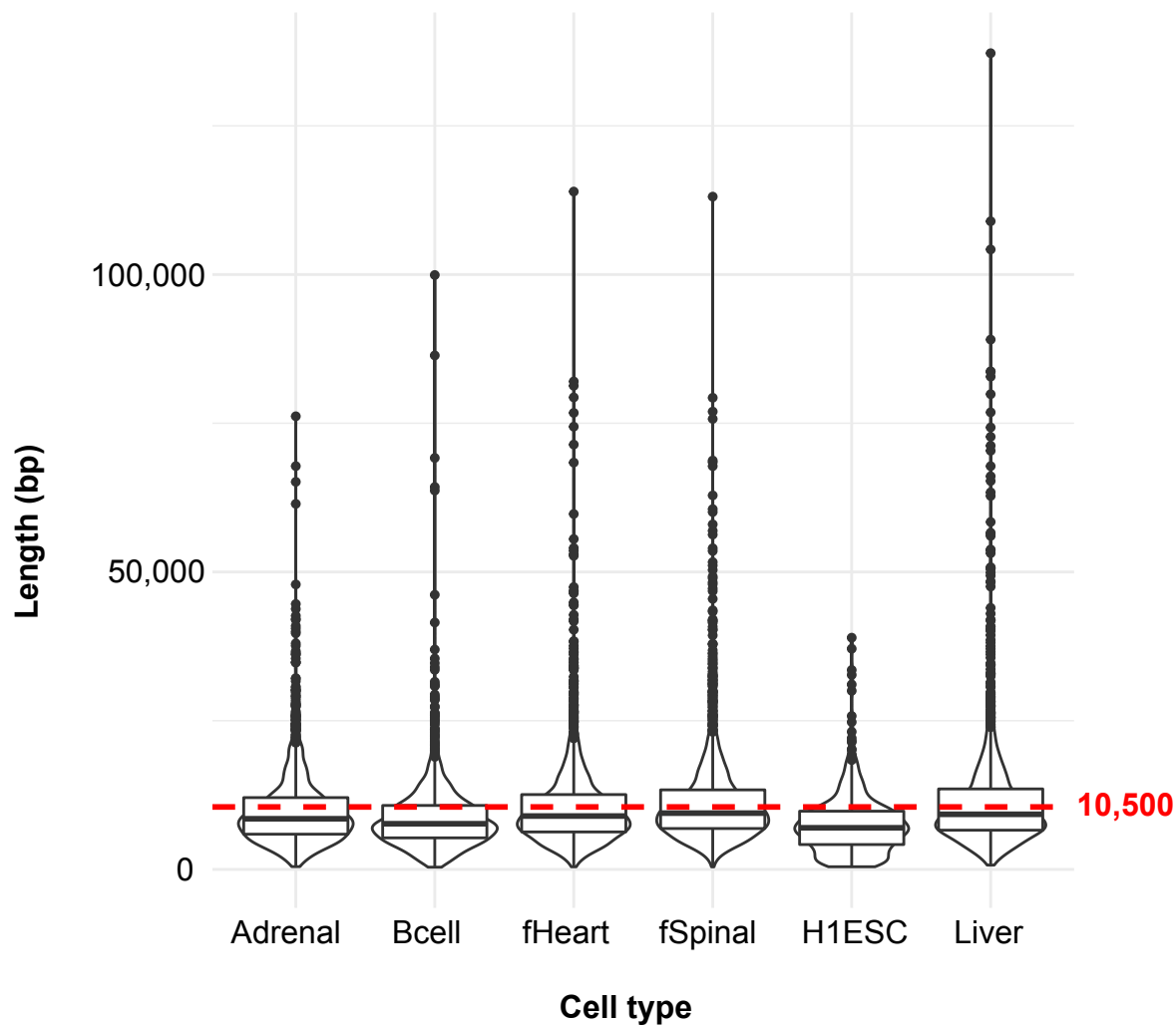


D

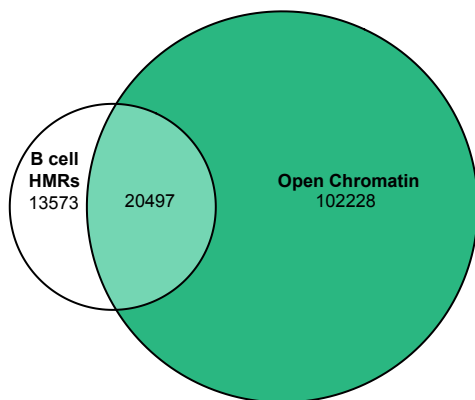
Crohn's Disease



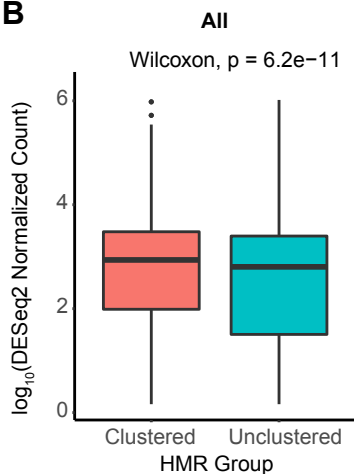




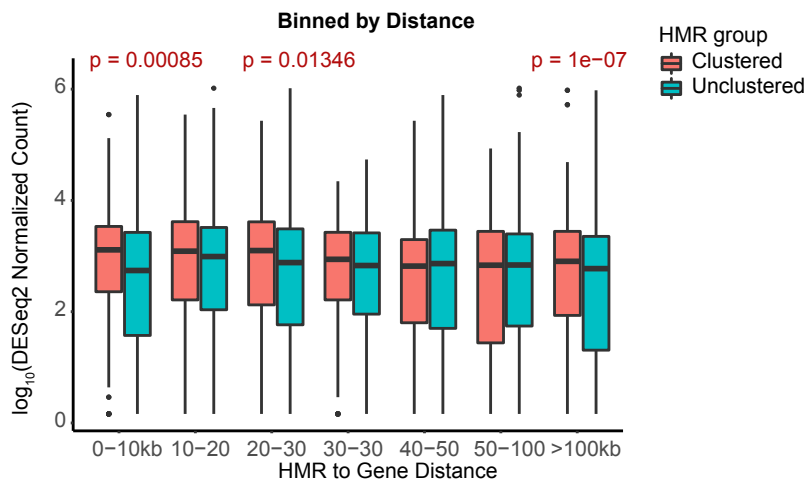
A



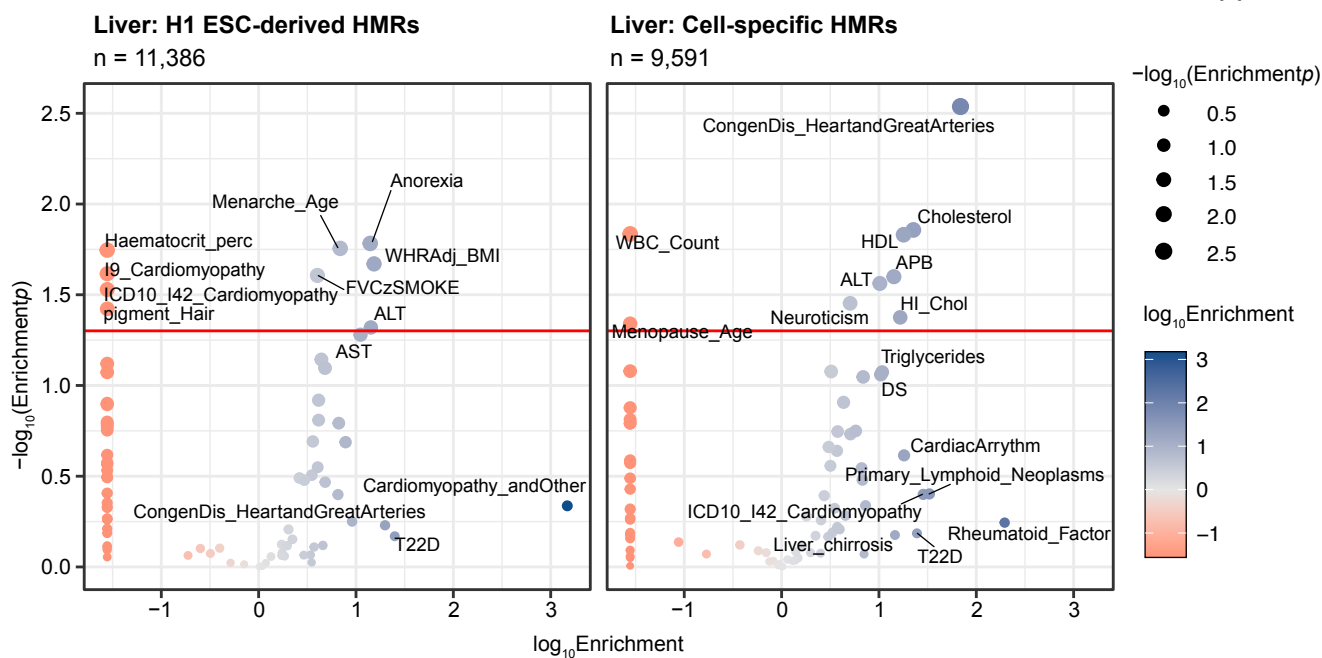
B



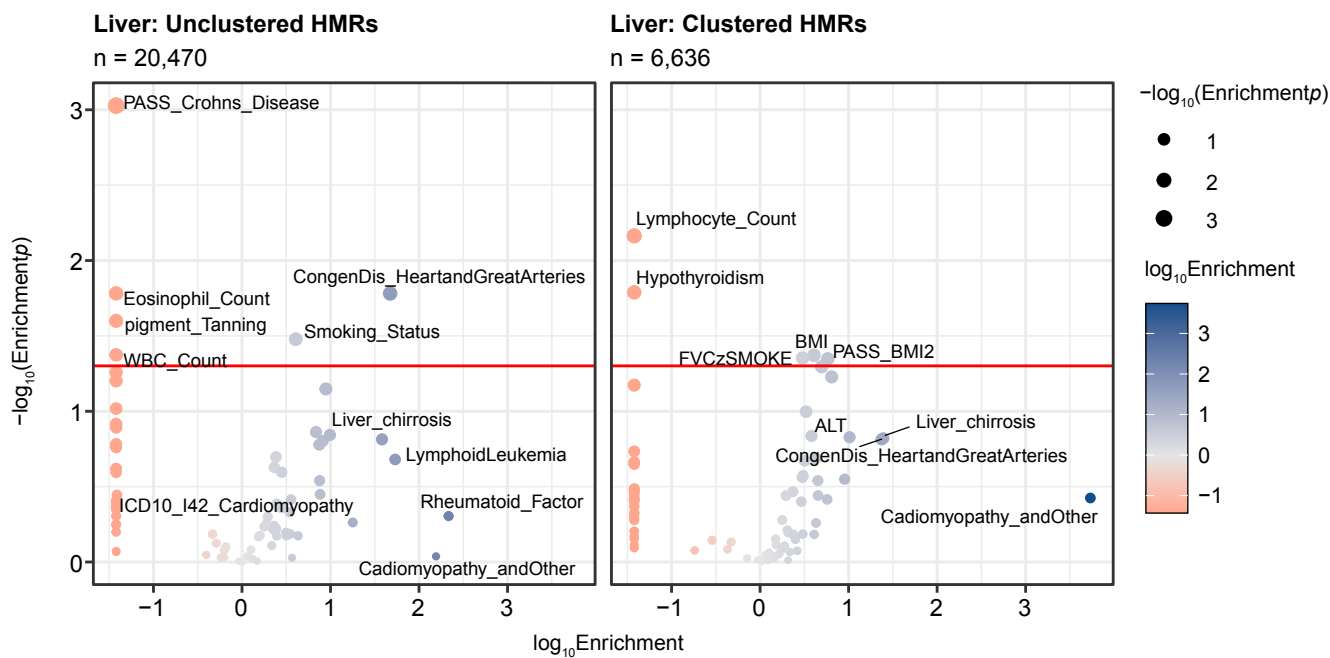
C



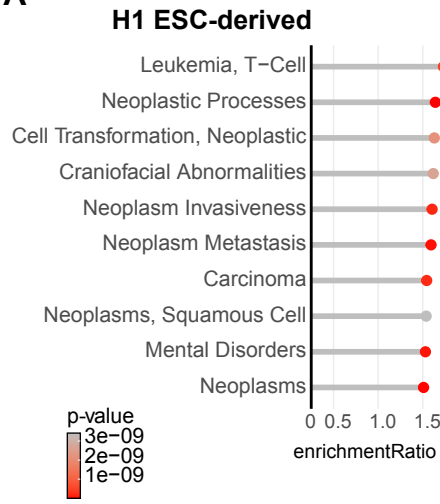
A



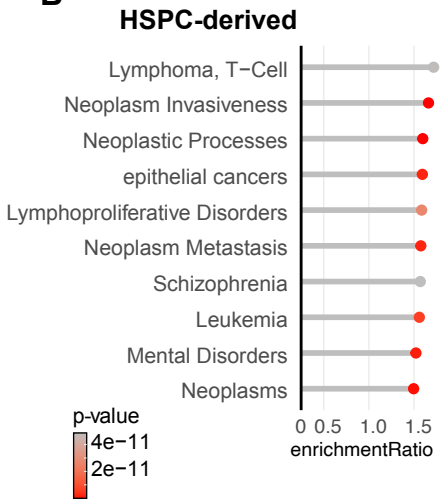
B



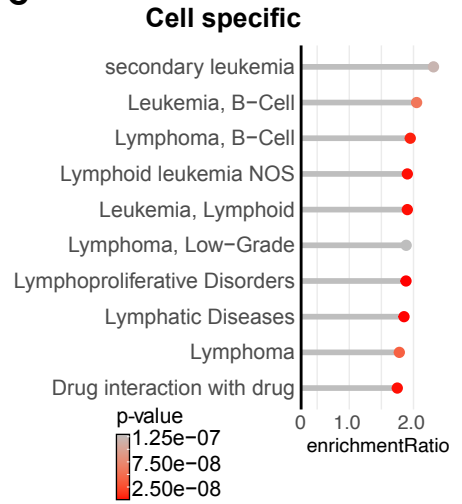
A



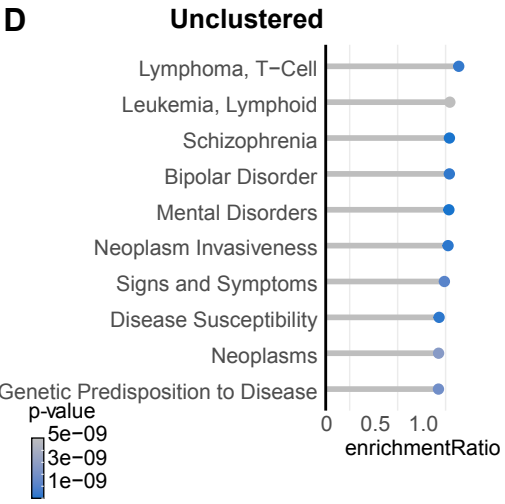
B



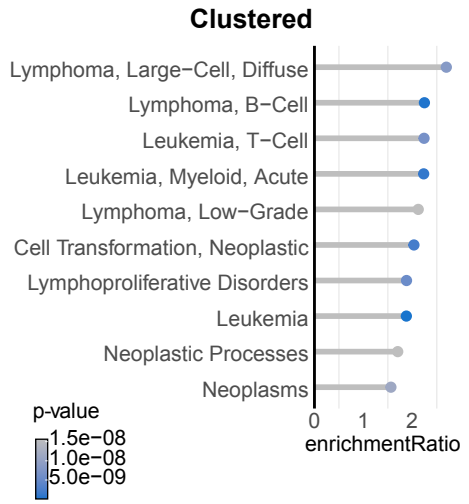
C

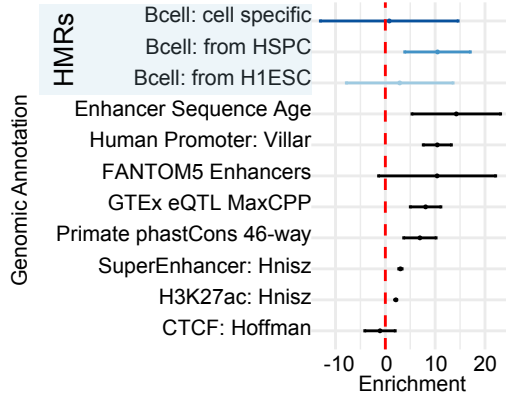
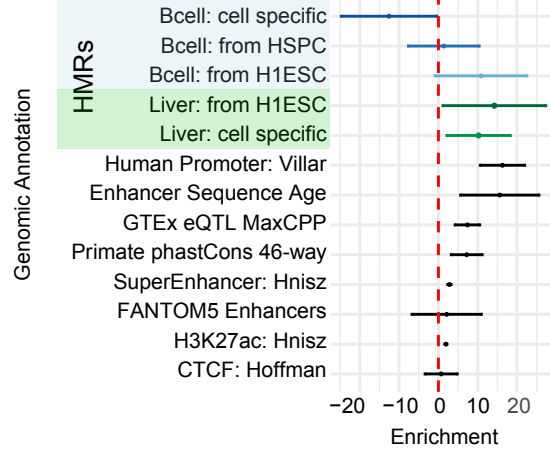


D



E



A**UK Biobank: Platelet Count****B****Alanine Transaminase (ALT)**

S1 Table. Inter-HMR lengths by cell type.

The table shows the distance between HMRs per cell type. Values represent the inter-HMR distances at different quantiles. The values have also been calculated at the .45 and .55 quantile values for more specificity around the median. Max values were limited to a maximum of 50kbp.

Quantile	H1 ESC	Fetal Spinal	Fetal Heart	Adrenal	Liver	B cell
0.1	227	881	761	673	633	436.3
0.2	1283	1770	1626	1678	1508	1259
0.3	2953.8	2960	2787	2969	2716.8	2410
0.4	5626.6	4574	4329	4777	441.4	4030
0.45	7102.65	5484	5309	5990	5540	4998
0.5	8921	6555	6361	7311	6846	6142
0.55	11013.15	7805	7689	8799.3	8461	7474
0.6	13299.8	9237	9183	10524	10148.2	9112.8
0.7	18970	13267	13231	15015.6	14803	13573.1
0.8	26779	19213	19414	21739.2	21582	19841.2
0.9	37041.2	29618	29903	32016.2	31880.6	30592.7

S2 Table. Clustering group counts by clustering distance.

Table of HMR counts for clustering groups: unclustered; clusters of 2 HMRs where one is a RefSeq promoter; clusters of 2 non-coding HMRs; and clusters of 3+ HMRs. Linking distances include 6kb (used for our analyses); 10.5kb (approximate mean inter-HMR length); and 12.5kb (distance commonly used in super-enhancer definitions applied to histone modification ChIP-seq datasets).

Linking Distance	Unclustered	Clusters 2	Clusters 2	Clusters
		Promoter Proximal		3+
6,000	17,185	2,836	3,475	5,974
10,500	12,975	3,398	3,987	8,805
12,500	11,668	3,395	3,910	9,717

S3 Table. List of 78 summary statistic files used for S-LDSC analyses.

Table of trait names used in S-LDSC. Files were sourced from the Alkes Price lab website or Neale lab LDSC summary statistics browser.

Traits:

Albumin	I9_IHD_wideDefinition
ALP	ICD10_I42_Cardiomyopathy
ALT	ICD10_I48_atrialFibrillationAndFlutter
Angina_byDoctor	IGF1
Apolipoprotein_B	Liver_chirrosis
AST	lung_FEV1FVCzSMOKE
blood_EOSINOPHIL_COUNT	lung_FVCzSMOKE
blood_PLATELET_COUNT	Lymphocyte_count
blood_RBC_DISTRIB_WIDTH	LymphoidLeukemia
blood_RED_COUNT	mental_NEUROTICISM
blood_WHITE_COUNT	Myocardial_infarction
bmd_HEEL_TSCOREz	Neutrophil_count
body_BALDING1	other_MORNINGPERSON
body_BMIz	PASS_AgeFirstBirth
body_HEIGHTz	PASS_Anorexia
body_WHRadjBMIz	PASS_Autism
bp_SYSTOLICadjMEDz	PASS_BMI1
Cadiomyopathy_andOther	PASS_Coronary_Artery_Disease
CardiacArrythm	PASS_Crohns_Disease
Cholesterol	PASS_DS
Coffee_type	PASS_Ever_Smoked
Congen_Heart_andGreatArteries	PASS_HDL
cov_EDU_YEARS	PASS_Height1
cov_SMOKING_STATUS	PASS_LDL
disease_AID_SURE	PASS_NumberChildrenEverBorn
disease_ALLERGY_ECZEMA_DIAGNOSED	PASS_Rheumatoid_Arthritis
disease_DERMATOLOGY	PASS_Schizophrenia
disease_HI_CHOL_SELF_REP	PASS_Type_2_Diabetes
disease_HYPOTHYROIDISM_SELF_REP	PASS_Ulcerative_Colitis
disease_RESPIRATORY_ENT	PASS_Years_of_Education2
Diseases_of_liver	pigment_HAIR
disease_T2D	pigment_SKIN
ECG_load	pigment_SUNBURN
ECG_phaseTime	pigment_TANNING
ECG	Primary_lymphoid_neoplasms
Haematocrit_percentage	repro_MENARCHE_AGE

HeartAttack_byDoctor	repro_MENOPAUSE_AGE
HighBloodPressure_byDoctor	RheumatoidFactor
I25_chronicIHD	Triglycerides
I9_Cardiomyopathy	