REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
01-09-2022	Final Report	1-Jun-2019 - 31-May-2022

4. TITLE AND SUBTITLE
Final Report: Single cell Analysis for Forensic Epigenetics (SAFE)

5a. CONTRACT NUMBER
W911NF-19-2-0185

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Salk Institute for Biological Studies
P. O. Box 85800

San Diego, CA 92186 -5800

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
75152-BB-DRP.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited

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

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Joseph Ecker
UU	UU	UU	UU		19b. TELEPHONE NUMBER 858-453-4100

RPPR Final Report
as of 07-Dec-2022

Agency Code: 21XD

Proposal Number: 75152BBDRP **Agreement Number: W911NF-19-2-0185**
INVESTIGATOR(S):

Name: Joseph R. Ecker
Email: ecker@salk.edu
Phone Number: 85845341001106
Principal: Y

Organization: **Salk Institute for Biological Studies**
Address: P. O. Box 85800, San Diego, CA 921865800
Country: USA

DUNS Number: 078731668 EIN: 952160097
Report Date: 31-Aug-2022 Date Received: 01-Sep-2022
Final Report for Period Beginning 01-Jun-2019 and Ending 31-May-2022
Title: Single cell Analysis for Forensic Epigenetics (SAFE)
Begin Performance Period: 01-Jun-2019 **End Performance Period:** 31-May-2022
Report Term: 0-Other

Submitted By: Todd Norell Email: tnorell@ihmc.org
Phone: (571) 236-2931

Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors

STEM Degrees: **STEM Participants:**

Major Goals: The Single-cell Analysis for Forensic Epigenetics (SAFE) project will identify molecular changes in individuals exposed to biological and/or CBRNE threats and could mitigate risks posed by WMD. Specifically, the SAFE project will identify single-cell DNA methylation patterns and chromatin accessibility changes in individuals who have been exposed to high priority threats. The SAFE project also includes acquisition of stabilized RNA and peripheral blood mononuclear cells (PBMCs) from patients exposed to CBRNE threats, and distribution of these biosamples to ACA team members and ECHO performers.

Additionally, SAFE-Rudimentary (SAFER) signature development is aimed at identifying epigenetic dynamics within samples (across PBMC cell types) and across samples (temporal and/or exposure - specific), with a focus on single cell epigenetic assays (SAFE) – snmC-seq and snATAC-seq, and correlating to single-cell RNA-seq (sc-RNA seq) datasets.

Accomplishments: • Variation in DNA methylation patterns are a reliable marker to accurately distinguish Bacterial vs Viral vs Chemical exposures

- Variation in DNA methylation patterns can be used to accurately predict time since exposure

- A minimal set of genomic loci (even a few hundred sites) can be used as bio-markers with a methylation signature predictive for exposure type and time since exposure

- Each of the seven immune cell types of PBMC used had a unique set of bio-markers allowing flexibility in the choice of cell type to use for field-assays / future studies.

Training Opportunities: Nothing to Report

Results Dissemination: Nothing to Report

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

RPPR Final Report
as of 07-Dec-2022

PARTICIPANTS:

Participant Type: Staff Scientist (doctoral level)
Participant: Manoj Hariharan
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: PD/PI
Participant: Joseph Ecker
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: Y

Participant Type: Non-Student Research Assistant
Participant: Cesar Barragan
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Non-Student Research Assistant
Participant: Rosa Castanon
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Non-Student Research Assistant
Participant: Jordan Altshul
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
National Academy Member: N

RPPR Final Report
as of 07-Dec-2022

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Todd Norell
Signature Date: 9/1/22 12:25AM

The Salk Institute for Biological Studies Epigenetic CHaracterization and Observation (ECHO) Final Technical Report

Period Covered by the Report:01 June 2019 through 31 May 2022

Date of Report: 26 August 2022

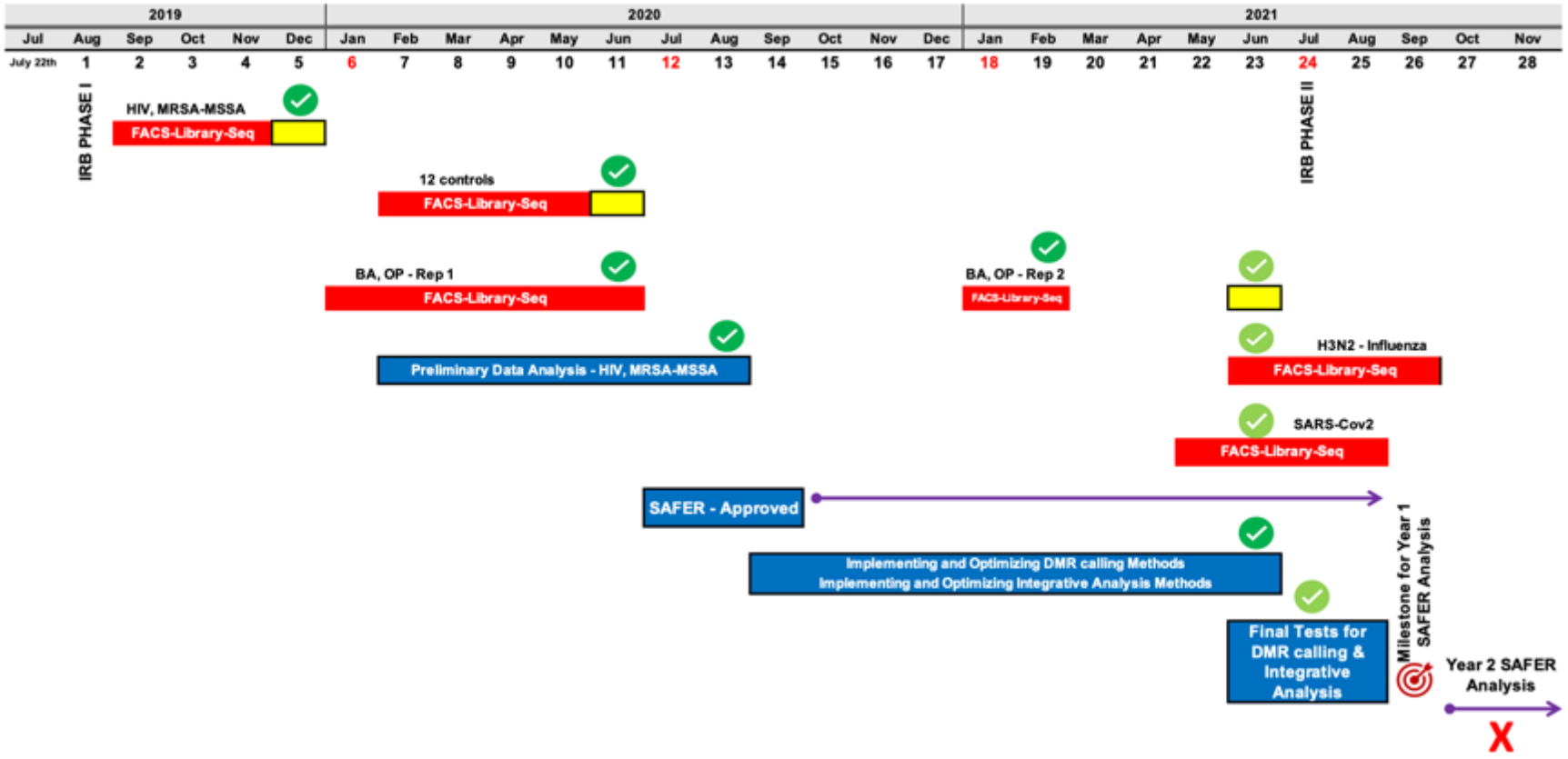
Project Title: Epigenetic CHaracterization and Observation (ECHO)
Agreement Number: W911NF1920185
Total Dollar Value: $4,778,942.00
Subcontractors: Stanford University, Institute for Human & Machine Cognition (IHMC)
Program Manager: Dr. Eric Van Gieson, DARPA/BTO

Submitted by:
Prof. Joseph Ecker
The Salk Institute for Biological Studies, 10010 N Torrey Pines Road, La Jolla, CA – 92037
858 453 4100
ecker@salk.edu

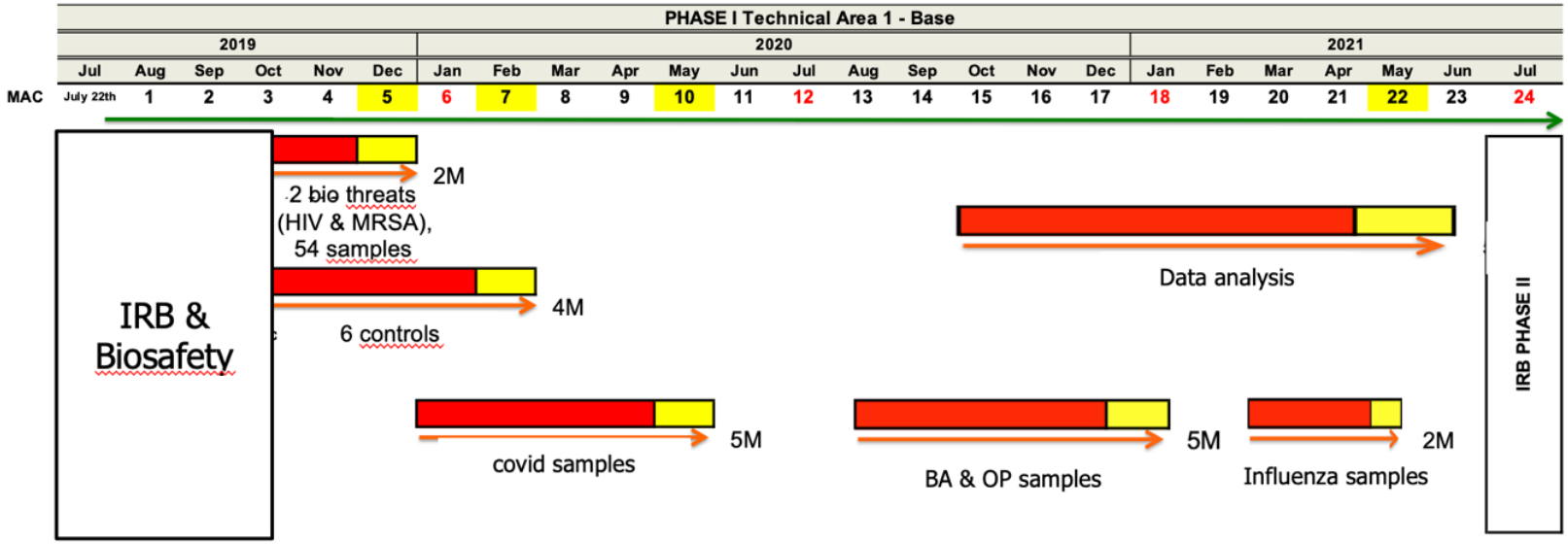
1 High-Level Project Progress, including Big Wins

- Variation in DNA methylation patterns are a reliable marker to accurately distinguish Bacterial vs Viral vs Chemical exposures
- Variation in DNA methylation patterns can be used to accurately predict time since exposure
- A minimal set of genomic loci (even a few hundred sites) can be used as bio-markers with a methylation signature predictive for exposure type and time since exposure
- Each of the seven immune cell types of PBMC used had a unique set of bio-markers allowing flexibility in the choice of cell type to use for field-assays / future studies.

2 Schedule: Milestones and Deliverables



High-level Gantt chart with major milestones and deliverables for snmC-seq



Red Bar Tasks: Sample preparation, Library Preparation, Sequencing, Data Processing
Yellow Bar Task: Data Upload/integrative analysis

High-level Gantt chart with major milestones and deliverables for snATAC-seq

3 Tasks: Progress, Accomplishments, and Plans

The goal of the DARPA Epigenetic CHaracterization and Observation (ECHO) program is to define changes in the features of an individual's epigenome to determine their history of exposure to Chemical, Biological, Radiological, Nuclear, and Explosive (CBRNE) weapons of mass destruction (WMD) and their precursors. The program will build a field-deployable platform capable of using the epigenome to assess CBRNE threat exposures, and support military forensics operations to counter-WMD proliferation.

The Salk Team provided the personnel, facilities and equipment necessary to complete Phase I of the program and an additional funded option focused on signature development.

Phase II of the program was not funded and additional changes were made throughout Phase I to remove specific tasks from the SOW. These changes will be highlighted below.

The program started with an ACA team to coordinate the research of independent contracts and cooperative agreements. This structure initially created several delays and by the end of Phase I, the lead ACA team was removed from the program.

Phase I focused on two molecular assays – (1) single nucleus methyl-C sequencing (snmC-seq) to generate detailed molecular profiles of DNA methylation and (2) single nucleus Assay for Transposase-Accessible Chromatin sequencing (snATAC-seq) to identify areas of open chromatin structure.

To complete Phase I, the following Tasks were completed:

Basic Task Technical Requirements (Base Phase, Phase I), Technical Area 1

3.1 Program Management

The Salk Team attended a kickoff meeting in Arlington VA on March 27 and 28 2019. While the ACA team was in place, bi-weekly team updates were held. In addition, regular meetings with DARPA with just the Salk Team were held to discuss progress and future direction. Monthly technical and financial status reports were provided.

3.2 Sample Descriptions

The Salk Team received the following sample types during Phase:

Exp #	Exposure	Source	
	HIV	GOV'T (MHRP)	
1	HIV-controls	GOV'T (MHRP)	
	MRSA	DUKE	
2	MRSA-controls	DUKE	First 6 months
3	B. anthracis	BATTELLE	
4	Organophosphates	DUKE	
5	COVID-19	ISMMS-DUKE-GOV'T	Months 6-12
6	Influneza H3N2	GOV'T (BARDA)	

Task 1: Human Blood Sample Accrual from 3 CBRNE Threat Exposures

At the start of the program, the Salk Team was tasked with collected PBMC's from two separate exposures, fentanyl and Co-60 radiation. The Salk Team coordinated with the Dayton VAMC Foundation to collect, process and store stabilized RNA and PBMCs samples from the two separate exposures. The Salk Team met with the physicians at the Kettering Health Network (KHN) leading to design the research protocol. For each of the two trials, samples were to be prospectively collected, processed and stored. At each time point, 20mL of venous blood was to be collected and processed. One 2.5mL PAXgene RNA tube were to be processed following the manufacturer's protocol for RNA stabilization and cryopreservation. Two 8mL BD Cell Preparation Tubes were to be processed following the Duke University protocol for PBMC isolation.1 mL ~ 10M PBMC aliquots will be prepared and cryopreserved following the Duke University protocol.

The protocol was completed, a reliance agreement between KHN and WSU IRB was signed and IRB/HRPO approval was underway. In October 2019, the contract agent and DARPA determined the Salk cooperative agreement was not the best mechanism to fund the data collection. No additional work was performed on this task.

Task 2: snmC-seq Analysis of Samples from 7 Threat Exposures (2 Biological and 5 Non-Biological)

Task 2.1: Obtain IRB approvals

- IRB approval for performing snmC-seq on CBRN exposures were obtained
- 100% complete
- Total funds for task
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

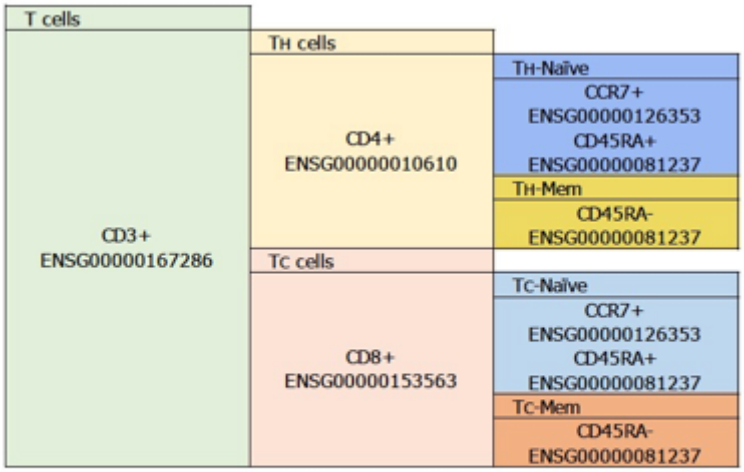
Task 2.2: Obtain cryopreserved PBMC samples from ACA collaborating teams

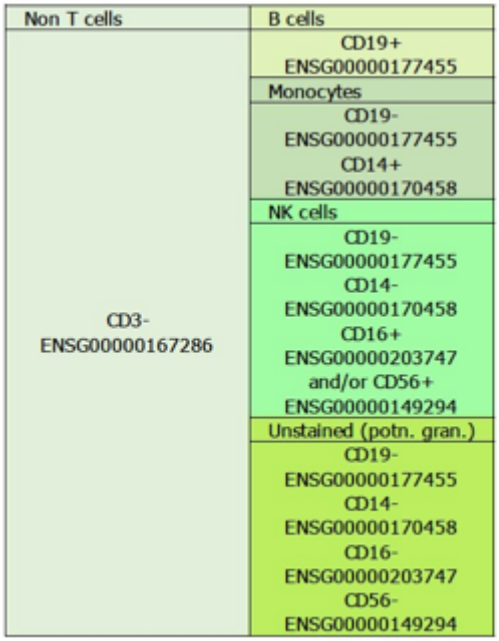
- Cryopreserved PBMC samples were received from collaborators
- 100% complete
- Total funds for task
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 2.3: FACS sort PBMCs cells into 384 well plates

- Seven PBMC cell types that are most long-lived and responsive to CBRNE exposures were identified (shown below)
- Antibodies were procured to label cell-surface proteins unique to each of the seven cell types

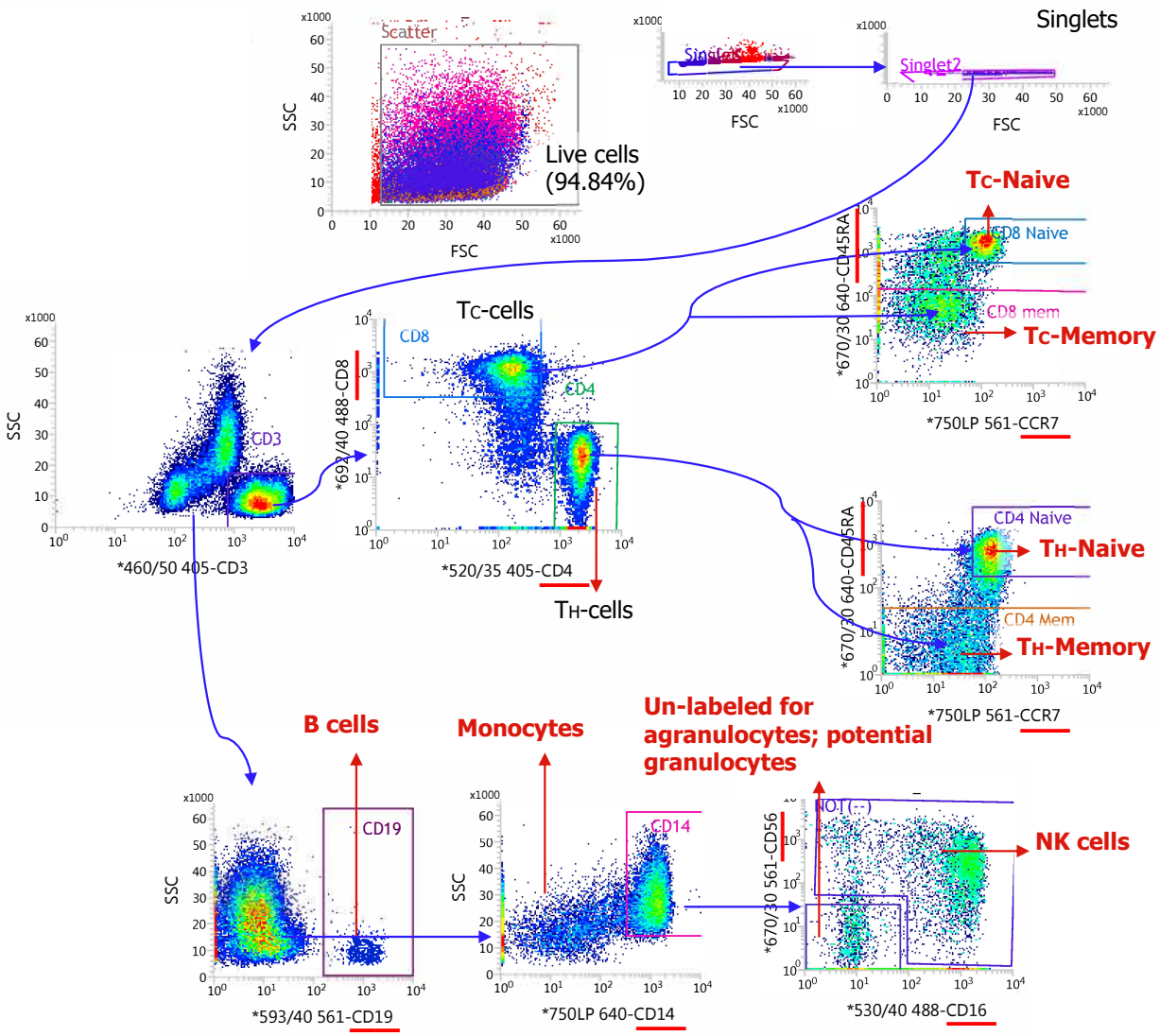
- An antibody cocktail was optimized for fluorescence-activated cell sorting (FACS)





- BV421-CD3
- PCP5.5-CD8
- BV510-CD4
- APC-CD45RA
- PE/Cy7-CCR7
- PE-CD19
- APCFire750-CD14
- PE/Cy5-CD56
- 488-CD16

- A gating strategy was developed to obtain equal numbers of all seven cell types in 384-well plates (48 cells of each cell type in each plate).



- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 2.4: Perform bisulfite reaction in each well

- Sodium bisulfite treatment was performed in each of the 384 wells of a plate based on the protocol developed in the lab.
-

References:

Luo, C., Keown, C.L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J.R., Sandoval, J.P., et al. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science *357*, 600–604.

Luo, C., Rivkin, A., Zhou, J., Sandoval, J.P., Kurihara, L., Lucero, J., Castanon, R., Nery, J.R., Pinto-Duarte, A., Bui, B., et al. (2018). Robust single-cell DNA methylome profiling with snmC-seq2. Nat. Commun. *9*, 3824.

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 2.5: Incorporate a 5'-sequencing adapter, amplify using PCR

- Library preparation, which includes incorporating a 5'-sequencing adapter to DNA and its amplification using PCR was carried out using a semi-automated process.
- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 2.6: Perform sequencing on the Illumina platform, raw data demultiplexing and quality assessment

- The libraries were sequenced first on an Illumina Miseq ® sequencer as a pilot run to assess the quality of the libraries
- Once all wells were shown to have consistently high-quality libraries, they were sequenced on a high-throughput Illumina Novaseq ® sequencer.
- Raw data demultiplexing and mapping were performed on the sequenced data
- All sequenced data passed stringent QC thresholds as defined in the detailed SOW (shown below).

snmC-seq QC Rules	Not Pass	Pass	Good
Percent genome	< 1%	[1%, 5%)	≥ 5%
CCC methylation rate	> 3%	(2%, 3%]	≤ 2%
Reads level %uniquely_mapped	< 40%	[40%, 70%)	≥ 70%
Reads per cell	< 10K	[10K, 1M)	≥ 1M
Cell level overall mCG rate (after mapping)	< 0.5	[0.5, 0.7)	≥ 0.7
Cell level overall mCH rate (after mapping)	≥ 0.08	< 0.08	
Dataset gender matches sample gender	No	Yes	
Average read length after trim	< 30 bp	≥ 30 bp	
Experiment passes metadata audits	No	Yes	
Dataset correlation	< 0.8	≥ 0.8	

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 2.7: Upload sequence data and sequencing metadata to IV&V

All raw data (fastq files) and processed files (BAM files and allC files) were uploaded to MIT-LL servers.

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 3: snATAC-seq Analysis of Samples from 7 Threat Exposures (2 Biological and 5 Non-Biological)

We performed the snATAC-seq analysis on the samples from 6 threat exposures (instead of 7), including 5 biological and 1 non-biological exposures.

Instead of isolating the single cells into 384-well PCR plates, we utilized the 10x genomics droplet-based platform to get single cells, which significantly increases the capacity (number of single cells in one assay) and for each sample, we can capture several thousand cells. We also prepared the libraries using the "10x Single Cell ATAC kits". The quality of the snATAC-seq data is based on the number of fragments & TSS enrichment in every single cell. Usually, cells with a number of fragment > 1000 & TSS enrichment > 5 is regarded as "good" cells; samples with >= 500 "good" cells are qualified for downstream analysis. In total, we generated qualified snATAC data for most exposures, except for MRSA/MSSA, as the MRSA/MSSA materials we received had limited PBMCs per sample (< 500,000 cells for each sample) with very poor cell state (partially due to highly diluted cells).

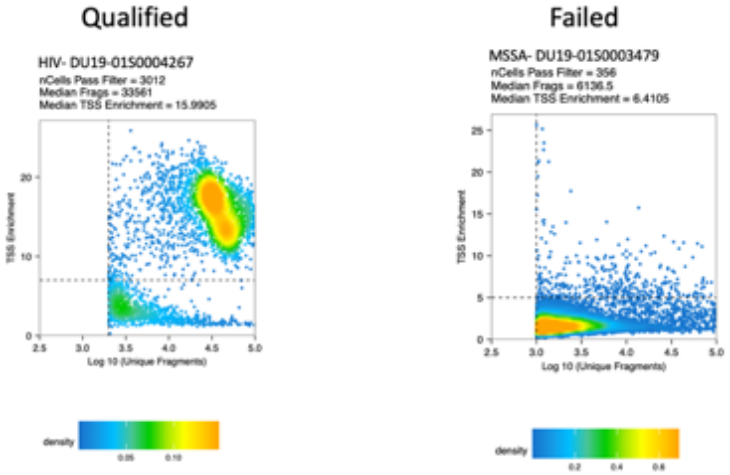


Figure: Examples of qualified and failed snATAC data. Note that we usually use higher cutoffs to capture better quality single cells for analysis if the sample quality is good enough.

Table 1 Numbers of samples with qualified/failed snATAC data

	HIV	MRSA/MSSA	BA	COVID19	Influenza	OP
Qualified	25	15	34	25	32	35
failed	2	13	0	0	2	0

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 4: snmC-seq Analysis of Control Samples from Subjects without CBRNE Threat Exposure

Task 4.1: Obtained IRB approval for snmC-seq of PBMCs from subjects without CBRNE threat exposure.

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 4.2: Cryopreserved PBMC control samples for snmC-seq were not provided by any of the ACA collaborators or DARPA performers. But since control samples are essential for any meaningful data analysis, samples were procured by the Salk team from commercial vendor Dx Biosamples after getting all required approvals from the DARPA team. These samples were selected to have varied ethnic backgrounds, mixed age groups, and sex. Instead of getting two-time points from 6 donors, a single time point from 12 donors was used as the 12 control samples. This was done to maximize the number of biological replicates. The table below shows the metadata of samples.

Subject ID	Blood Draw Date	Race	Gender	Age
R537415	12/5/19	Caucasian	Female	45
R537308	12/3/19	African American	Female	43
R537264	12/2/19	African American	Female	37
R537269	12/2/19	African American	Male	44
R537305	12/3/19	African American	Male	51
R537779	12/12/19	African American	Male	65
R537769	12/12/19	White/Caucasian	Male	39
R537560	12/9/19	White/Caucasian	Male	57
R537579	12/9/19	White/Caucasian	Female	32
N403841	12/9/19	White/Caucasian	Female	33
N100301	2/17/20	African American	Female	68
N100302	2/17/20	Caucasian	Female	66

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 4.3: Isolated single cells using fluorescence-activated cell sorter (FACS) into 384-well PCR plates.

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 4.4: Performed bisulfite reaction in each well. Unmethylated lambda will be spiked in to assess the conversion rate.

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 4.5: Incorporate a 5'-sequencing adapter, amplify using PCR

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 4.6: Performed sequencing on the Illumina platform, raw data demultiplexing and quality assessment.

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 4.7: Uploaded sequence data and metadata to the ACA partner site for access by the Government IV&V team and ACA collaborators.

- 100% complete
- Total funds spent – 100%

- Amount remaining – $0.00
- No deviations from actual

Task 4.8: Performed statistical analysis to identify potential snmC-seq biomarkers in collaboration with ACA team members. Refer to Analysis Tasks Report Section for more details.

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Task 5: snATAC-seq of Control Samples from Subjects without CBRNE Threat Exposure

We also received commercially available PBMC samples from healthy donors from collaborators, and performed the snATAC-seq analysis on them using the same commercial "10x Single Cell ATAC kits". Only 4 out of 6 control samples were processed as we found out these control samples were highly batch confounded, it's impossible to compare them with other exposures in our dataset.

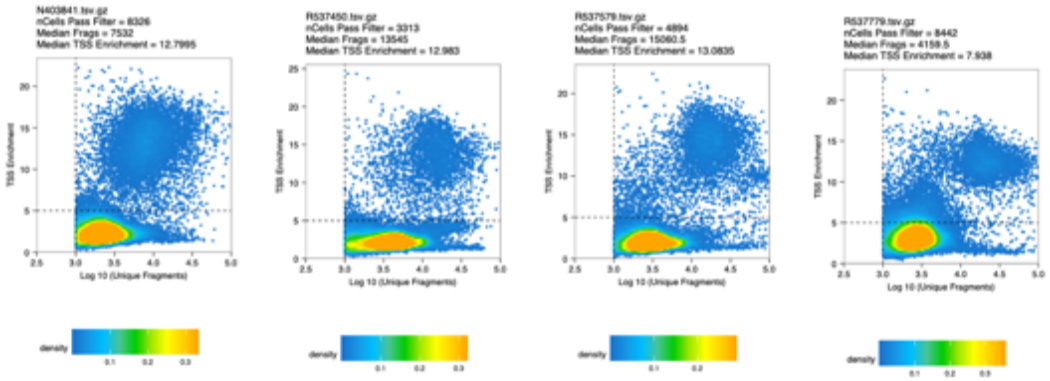


Figure: QC of the 4 control samples processed. All 4 samples are qualified.

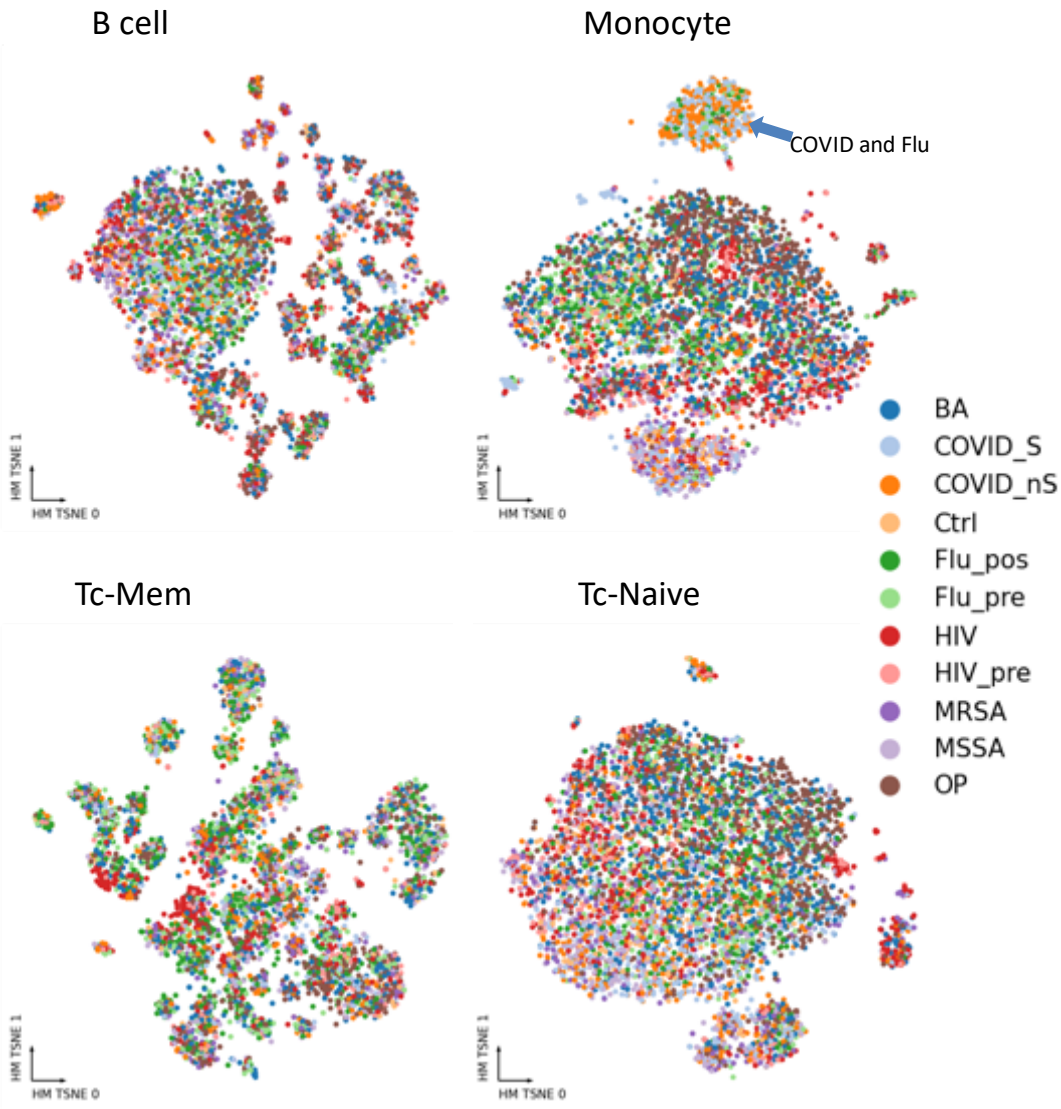
- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- The control samples were highly batch confounded, it's impossible to compare them with other exposures in our dataset.

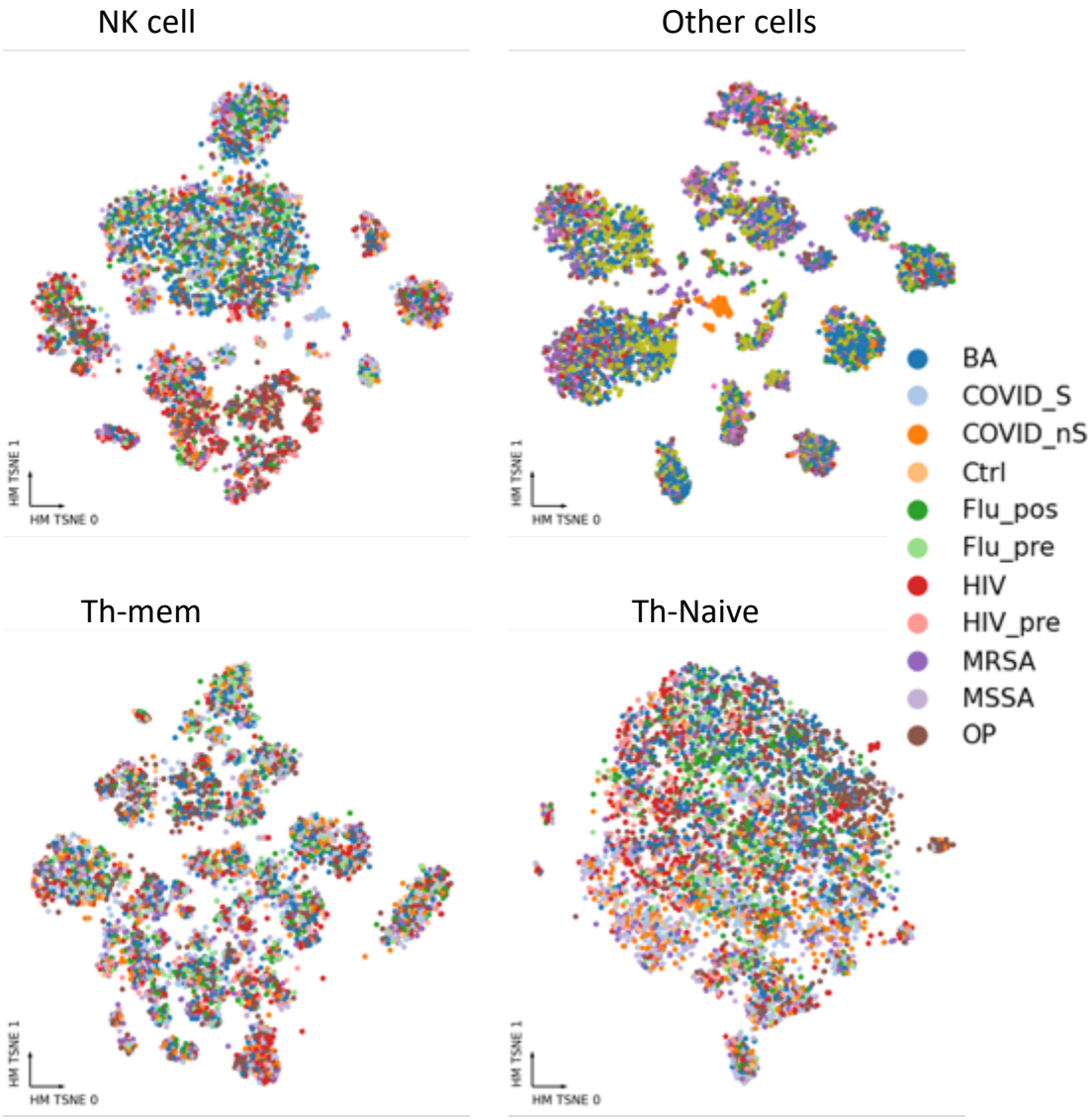
Supplementary Funding for Data Analysis – SAFER (SAFE Rudimentary Signatures)

Initially, we were not funded to do a detailed analysis of the DNA methylation datasets. But taking into consideration our long-standing experience in analyzing such data, we were provided supplementary funds for one year (MAC1 through MAC12 of SAFER) to identify differentially

methylated regions and develop machine learning approaches to identify signatures of exposure-response as biomarkers.

Analysis Tasks 1.1.1 and 1.1.2: Visual representation of global DNA methylation levels and de-novo Clustering of cells based on mCG levels across various exposures in different cell types were completed. Example results are shown below:

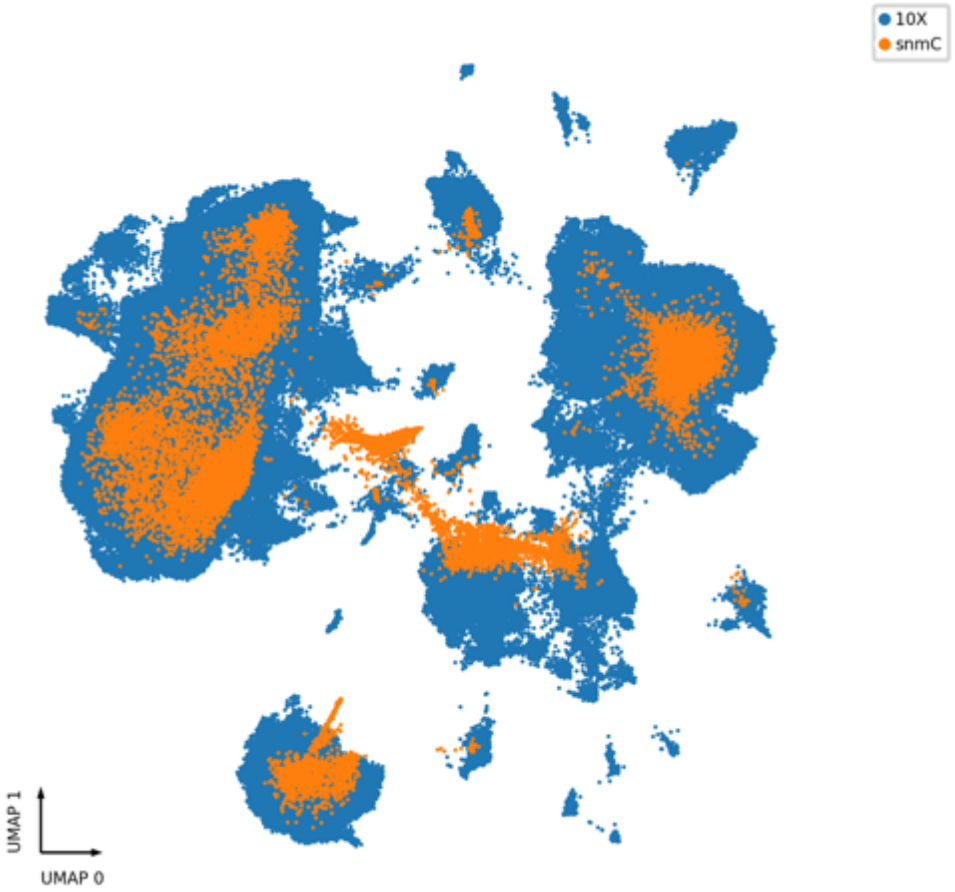




- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Analysis Tasks 1.1.3, 1.1.4, 1.1.9: De-novo Identification of Marker Genes, integration with sc-RNA-seq and Identification of Marker Genes based on sample information, and integration with sc-RNA-seq, Integration with other single-cell assays.

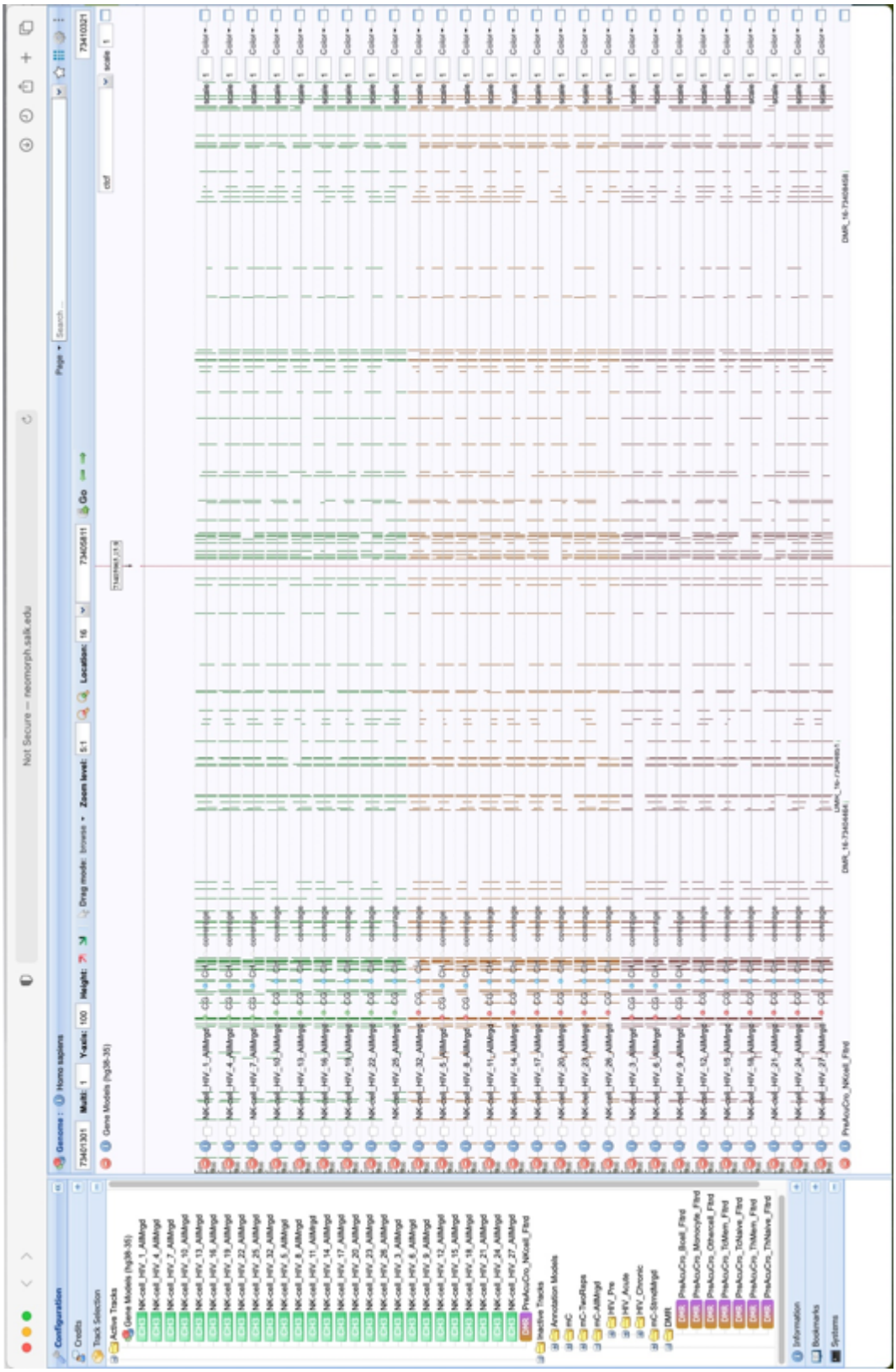
DNA methylation data was integrated with single-cell RNA-seq data generated by ISMMS. Example shown below with HIV cohort:



The blue data points correspond to the gene expression profile of cells derived from single-cell RNA-seq data while the orange data points correspond to the global methylation levels obtained from snmC-seq.

- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

- Analysis Tasks 1.1.5: Develop a single-cell browser. Example results are shown below:

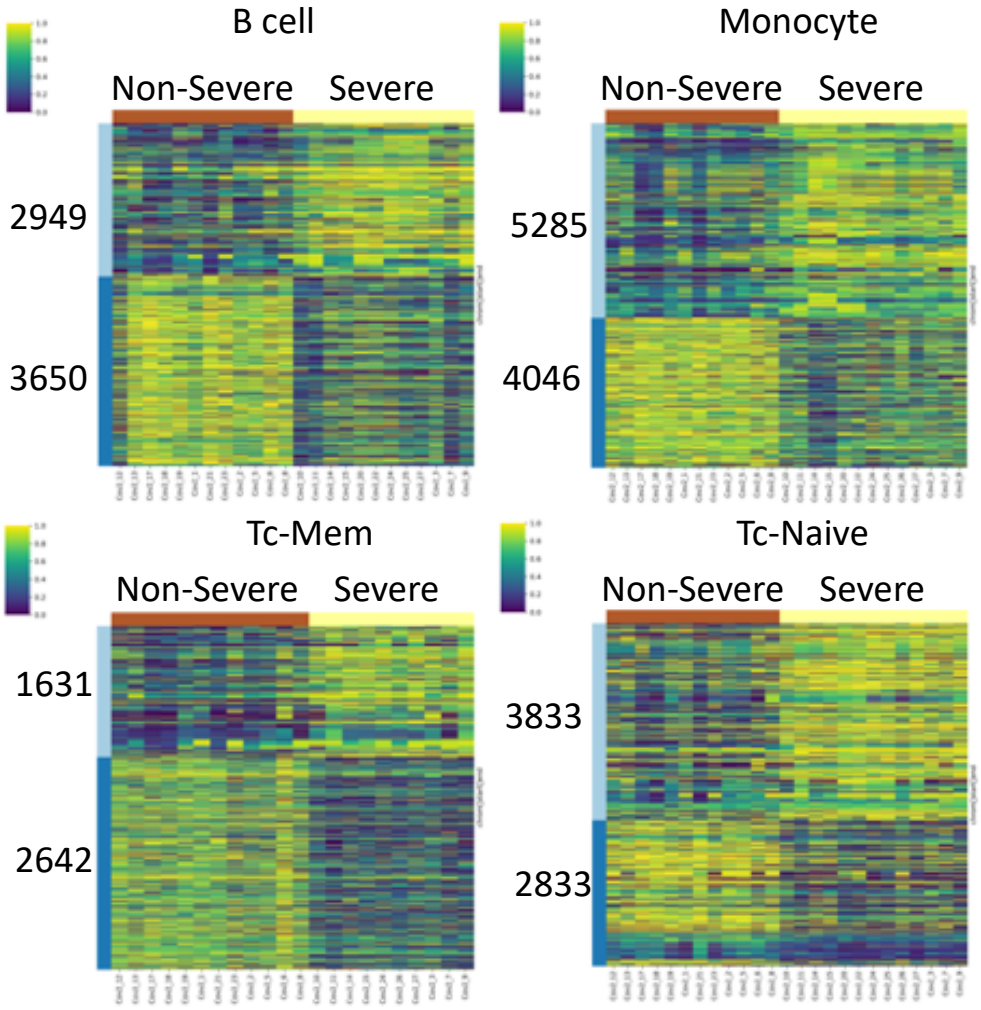


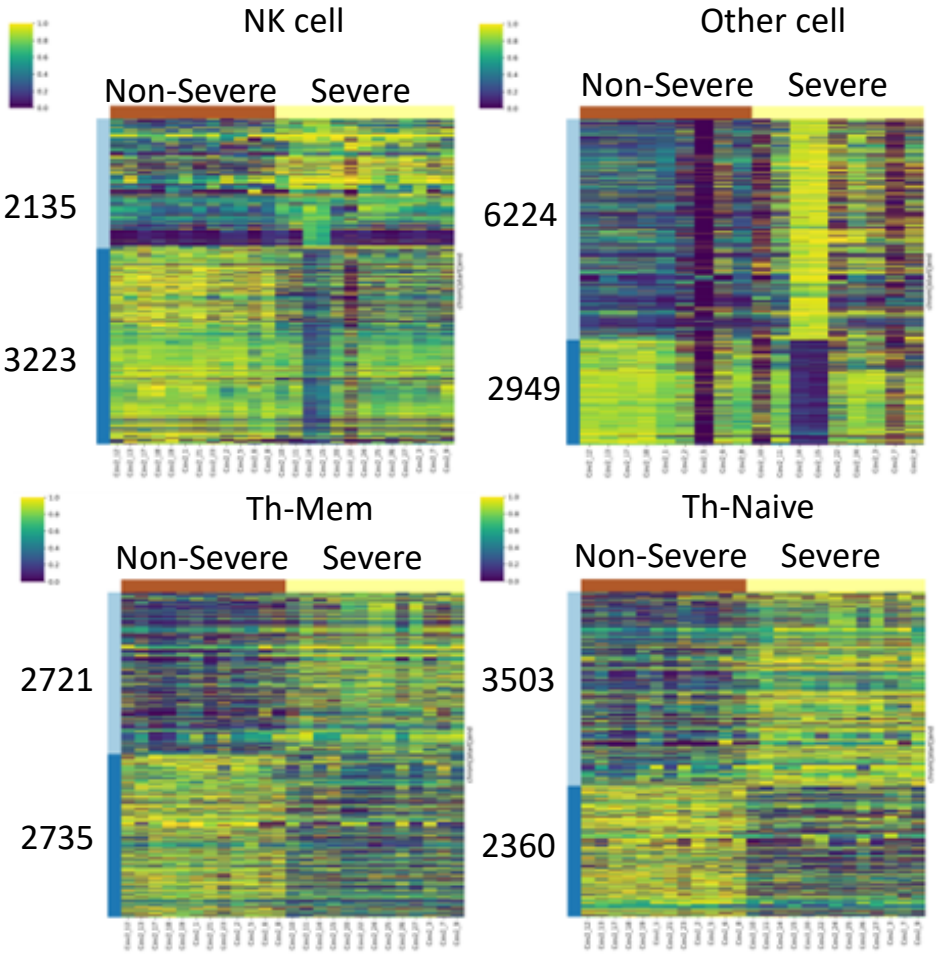
- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Analysis Tasks 1.1.6: Identifying Differentially Methylated Regions (DMR).

Example results are shown below for the COVID cohort:

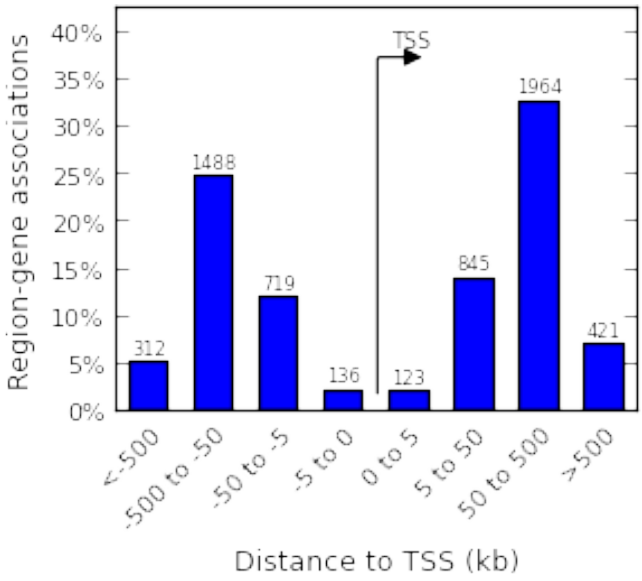
Numbers on the Y-axis denote the number of DMRs. Color scales represent the percentage of DNA methylation at each differentially methylated site averaged over the cytosines in the region.





- 100% complete
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Analysis Tasks 1.1.7: Identification of regulatory features within DMRs.
Example results shown below for the HIV cohort: A total of 386,945 DMRs were obtained in the HIV cohort. Further filtering by applying a criterion that at least three biological replicates out of nine were present in either the hypomethylated or hypermethylated category gave a set of 3,410 DMRs. Below is the summary of regulatory regions associated with each of these DMRs.



- 100% complete
- Total funds for task
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Analysis Tasks 1.1.8: Functional annotation of Marker Genes and DMRs.
Example results shown below for the HIV cohort. Applying a criterion that at least three biological replicates out of nine were present in either the hypomethylated or hypermethylated category gave a set of 3,410 DMRs. Below is an example of function annotation of these DMRs (biological processes enriched in the set of DMRs):

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*
regulation of interleukin-10 production	10	4.0090e-7	5.2755e-4
embryonic foregut morphogenesis	50	1.4454e-5	3.8040e-3
foregut morphogenesis	60	2.1595e-5	4.7361e-3
positive regulation of meiotic nuclear division	83	5.6389e-5	8.9400e-3
regulation of interleukin-10 secretion	87	6.1713e-5	9.3342e-3
loop of Henle development	97	7.3274e-5	9.9403e-3
regulation of hair follicle maturation	107	1.0541e-4	1.2964e-2
nephron tubule formation	110	1.1772e-4	1.4082e-2
Wnt signaling pathway involved in midbrain dopaminergic neuron differentiation	137	2.3390e-4	2.2466e-2
cellular response to laminar fluid shear stress	190	5.9690e-4	4.1340e-2
positive regulation of interleukin-10 secretion	207	7.1449e-4	4.5420e-2

- 100% complete
- Total funds for task
- Total funds spent – 100%
- Amount remaining – $0.00
- No deviations from actual

Analysis Task I.2: SAFER signatures based on single nucleus ATAC-seq (snATAC-seq)

We clustered the single cells based on the “TileMatrix” generated from the snATAC-seq data, leading to the clustering of different cell identities, including CD4 T cells, CD8 T cells, B cells, monocytes, and NK cells. We also identified specific marker genes, as well as specific accessible regions for different cell types. Based on the motif enrichment analysis on differential chromatin accessible regions, we also identified specific transcription factors activated in different cell types, indicating the important roles of these transcription factors in those cell types.

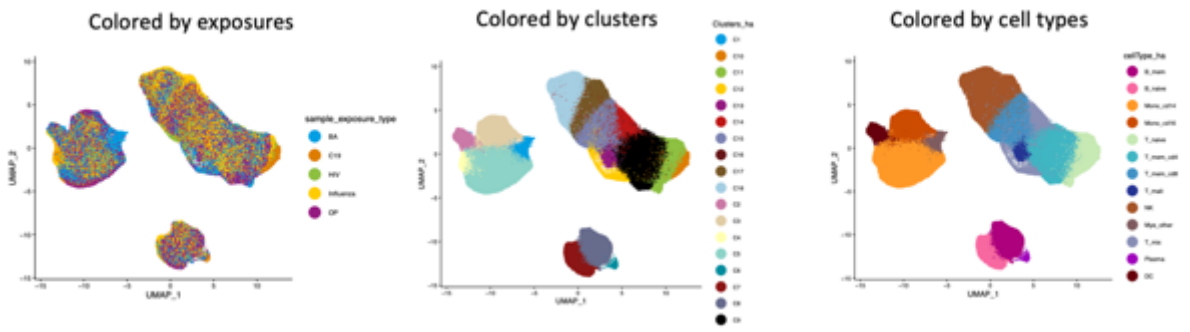


Figure: Integrative analysis of all PBMC samples from all exposures (except MRSA/MSSA) after batch correction. Different cell types from different exposures are clustered together after the batch correction.

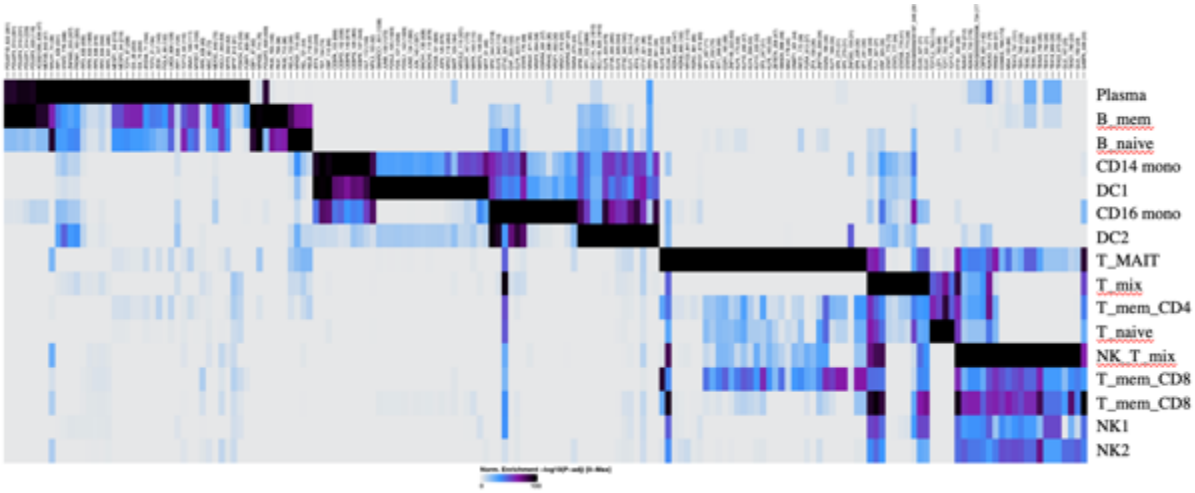


Figure: Heatmap on TF activity in different cell types based on motif enrichment analysis in differentially accessible regions in those cell types.

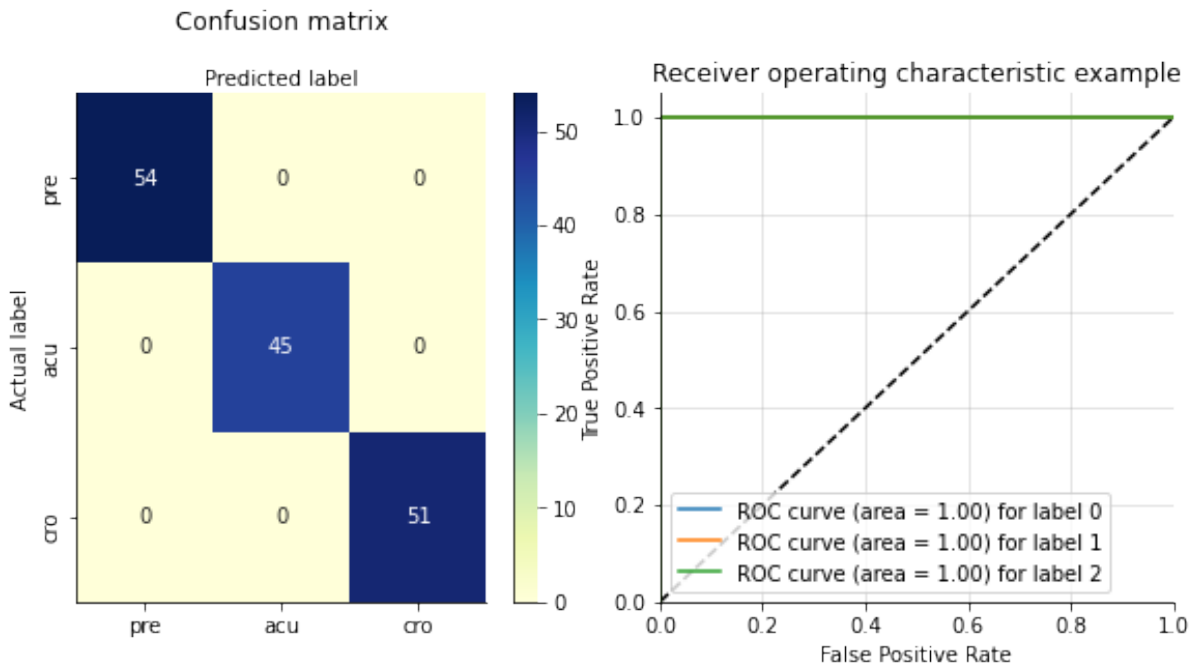
Significant Results:

For each exposure, we were able to identify a set of genomic loci that contain unique DNA methylation patterns specific to each cell type that can serve as biomarkers to identify the type of exposure and the time since exposure. This is the “signature” of each exposure on the various cell types.

Major result from HIV cohort:

Machine Learning Approach to Identify HIV infection

- Permutated 750 samples in ‘pre’, ‘acute’ and ‘chronic’ states.
- 600 as training, 150 as test.
- Select DMLs that can discriminate HIV stages



DMLs from all 8 cell types can perfectly predict the ‘pre’, ‘acute’ and ‘chronic’ states (100% accuracy)

Minimal set of DMLs that can perfectly predict HIV infection and Time since Exposure:

Cell type	B cell	Monocyte	NK cell	Other cell	Tc-Mem	Tc-Naïve	Th-Mem	Th-Naïve
# features	412	318	809	517	21,467	7,574	3,537	2,801

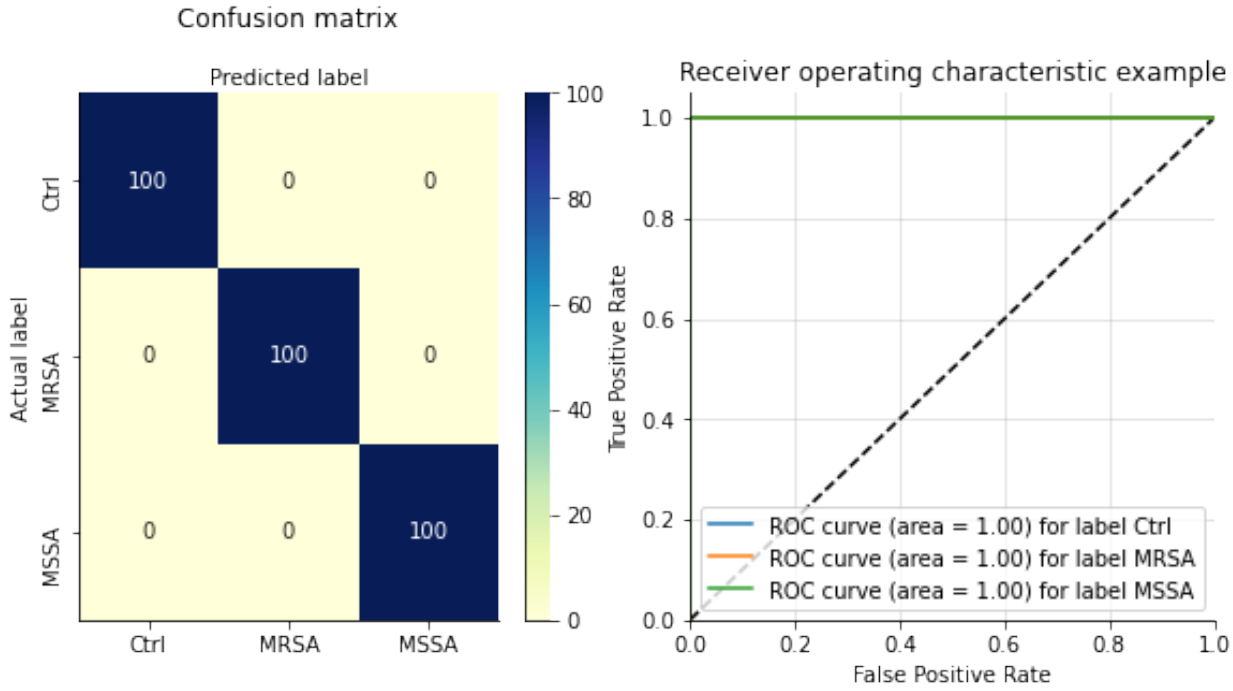
(Preliminary Finding)

- Major result from MRSA-MSSA cohort:

Machine Learning Approach to Identify MRSA-MSSA infection

- Permutate 400 samples across ‘Ctrl’, ‘MRSA’ and ‘MSSA’
- 300 as training set, 100 as test set
- Select features that can discriminate the exposures

Prediction based on DMLs in 8 cell types are accurate in classifying ‘MRSA’ from ‘MSSA’



Minimal set of DMLs that can perfectly predict MRSA-MSSA infection:

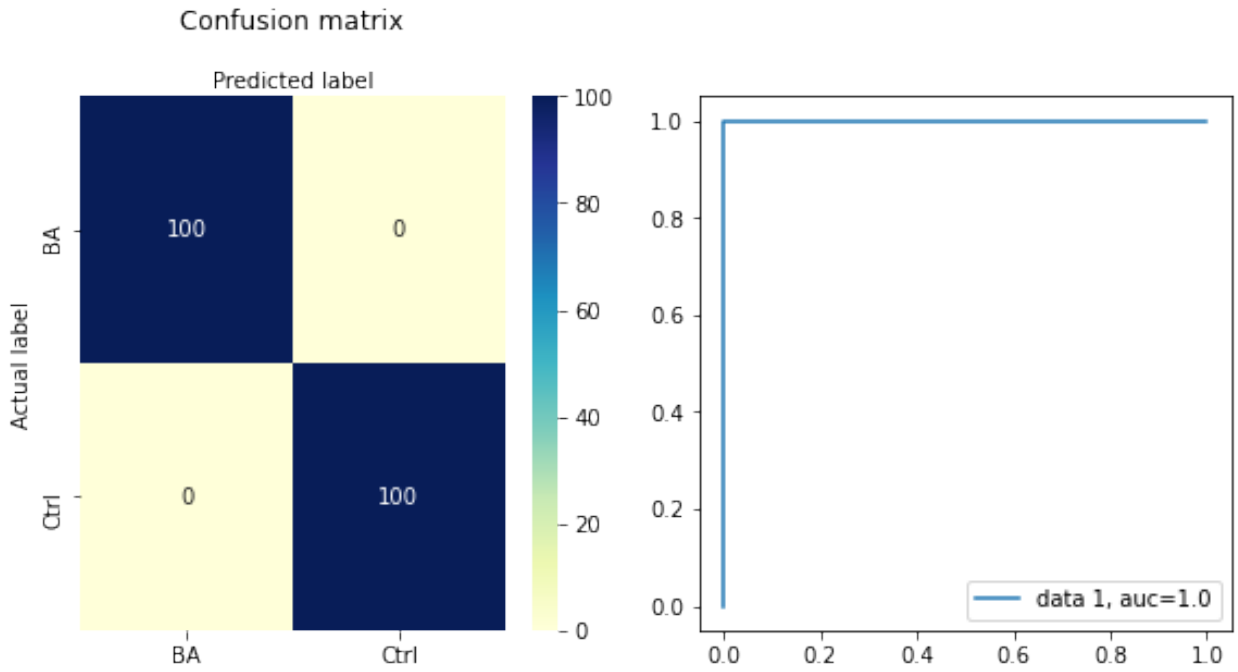
Cell type	B cell	Monocyte	NK cell	Other cell	Tc-Mem	Tc-Naïve	Th-Mem	Th-Naïve
# features	670	725	1,004	2,500	1,175	730	1,029	814

(Preliminary Finding)

Major result from Anthrax handling cohort:

Machine Learning Approach to Identify persons handling *B. anthracis*

- Permutate 400 samples, 300 as training, 100 as test.

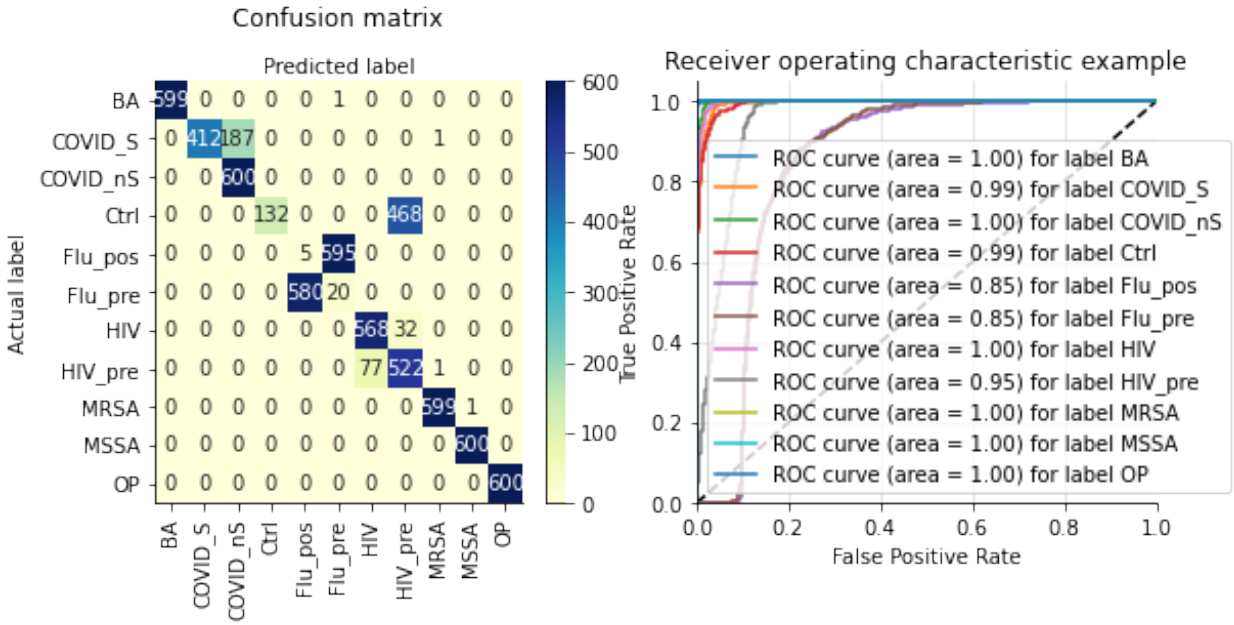


Minimal set of DMLs that can perfectly predict Anthrax handling:

Cell type	B cell	Monocyte	NK cell	Other cell	Tc-Mem	Tc-Naïve	Th-Mem	Th-Naïve
# features	392	221	2,596	819	434	205	227	269

(Preliminary Finding)

Consolidated Summary of Signatures of each Exposure:



Accomplishments

All goals, milestones, and deliverables of each of the tasks were met.

We developed a unique antibody cocktail capable of consistently obtaining seven major immune cell types from historically cryopreserved samples of PBMC.

We developed a gating strategy capable of consistently obtaining seven major immune cell types from historically cryopreserved samples of PBMC.

We developed and utilized a streamlined semi-automated pipeline for bisulfite conversion of DNA within single nuclei to efficiently converted all unmethylated cytosines in the genome with high coverage (greater than 99.9% efficiency) for every cell in all 384 cells consistently.

We implemented a custom computational pipeline that efficiently utilizes available compute resources to perform QC, mapping, and quantification of DNA methylation levels at a whole-genome level.

We collaborated with the Blish group from Stanford and integrated our snATAC data on COVID-19 samples with their scRNA-seq (transcriptome) and CyTOF data (proteome), covering in total 64 COVID-19 patients across the full range of disease severity, from outpatients with mild disease to fatal cases. The data revealed widespread dysfunction of peripheral innate immunity in severe and fatal COVID-19. We also identified chromatin accessibility changes at NF-κB binding sites within cytokine gene loci as a potential mechanism for the striking lack of pro-

inflammatory cytokine production observed in monocytes in severe and fatal COVID-19. This work was published in the Journal of Experimental Medicine in August 2021.

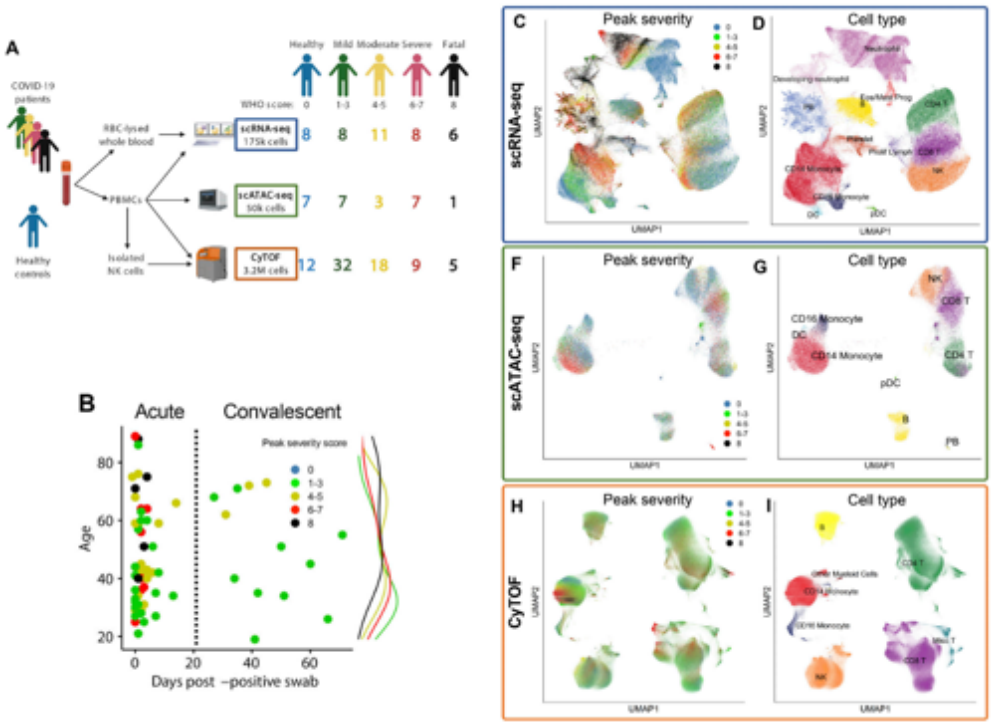


Figure: A trimodal single-cell atlas of the peripheral immune response to COVID-19 across a range of disease severities. (A) Pipeline for sample processing and a number of patients analyzed, summarized by modality and peak disease severity score. (B) Summary of key patient metadata, including age, peak disease severity score, and days after first positive nasopharyngeal PCR test. The vertical dotted line placed at 21 d after a positive test indicates the threshold after which patient samples are considered convalescent. (C, D, and F–I) UMAP projections of complete scRNA-seq (C and D), scATAC-seq (F and G), and CyTOF (H and I) datasets colored by peak disease severity score of the sample (C, F, and H) or cell type (D, G, and I). It shows there is the most significant difference among different severities in monocytes and neutrophils.

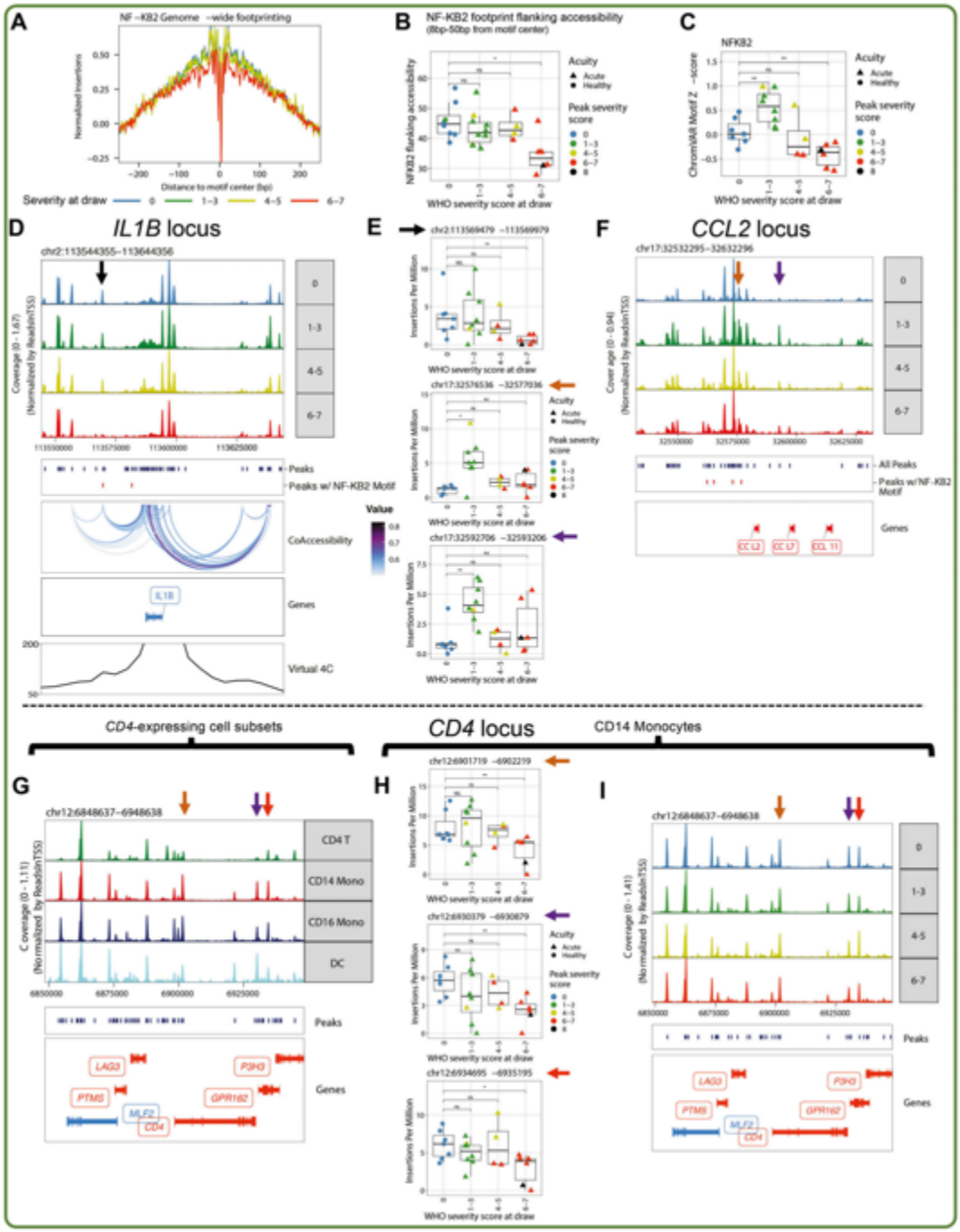


Figure: (A) Genome-wide footprinting of the NF-κB2 binding motif in CD14 monocytes from different severity groups shown in different colors. (B) Box plot depicting quantification of the "flanking accessibility" (Baek et al., 2017; Corces et al., 2018) for NF-κB2 motif footprints in CD14 monocytes from different samples. Each dot indicates the average "flanking accessibility" value for each sample. (C) Box plot depicting the average chromVAR z-scores of NF-κB2 binding motifs in CD14 monocytes from different samples. (D and F) The genome tracks show genomic regions near IL1B (D) and CCL2 (F) genes. The top panel indicates coverage at different peak regions for CD14 monocytes in different severity groups; the box below shows peaks called from all CD14 monocytes (dark blue) in the 100-kb region and peaks containing

putative strong NF-κB2 binding sites (red); the CoAccessibility box in D shows the accessibility correlated peak pairs across all CD14 monocytes near the IL1B locus; the Genes box shows the location of IL1B (D) or CCL2 (F) together with other adjacent genes; the bottom Virtual 4C track in D shows Knight-Ruiz–normalized contact frequencies to the IL1B promoter in THP-1 monocytic cells; blue color means the gene is located on the minus strand, and red color means the gene is located on the plus strand. The arrows indicate peaks of interest whose accessibility is quantified in the corresponding box plots (E). (G and I) The genome tracks show genomic regions near the CD4 gene. The top panel indicates coverage at different peak regions for different cell subsets (G) and CD14 monocytes in different severity groups (I); the box below shows peaks called from all PBMCs (G) or from the CD14 monocytes (I) in that region (dark blue); the bottom Genes box shows the location of CD4 and other adjacent genes; blue color means the gene is located on the minus strand, and red color means the gene is located on the plus strand. The arrows indicate monocyte-specific peaks with higher accessibility in monocytes and DCs than in CD4 T cells. (E and H) Box plots depicting the Tn5 insertions per million at the peaks marked with the corresponding arrows in CD14 monocytes. Exact P values for E: top, P = 0.0081 healthy versus severe; middle, P = 0.014 healthy versus mild; bottom, P = 0.0037. Exact P values for H: top, P = 0.0047 healthy versus severe; middle, P = 0.0047 healthy versus severe; bottom, P = 0.022 healthy versus mild. Points are colored by the peak disease severity score, shaped according to disease acuity, and grouped by the disease severity score at the time of sample collection. *, P < 0.05; **, P < 0.01; ns, not significant at P = 0.05 by two-sided Wilcoxon rank-sum test with Bonferroni correction for multiple hypothesis testing.

4 Project Coordination, Dissemination, and Translation Efforts

4.1 Project Coordination

The Salk Team held bi-weekly meetings with the DARPA team for the first 18 months of the program. At that time, the frequency of the meetings was reduced to monthly meetings at the request of DARPA. For each update, the Salk Team provided slides in advance summarizing recent progress and future plans. More formal quarterly reports, annual reports, and monthly financial reports were provided to DARPA to track program progress as well.

Other than the PI meeting, travel and in-person meeting attendance were cancelled due to COVID-19.

There were no changes in Key Personnel for the Salk Team. Joe Ecker at Salk, Will Greenleaf at Stanford, and Tim Broderick at IHMC supported the program through the effort.

5 Publications and Presentations

Title, Authors	Description/Type	Date Sent to DARPA/Agent	Status
Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in COVID-19. Aaron J Wilk et al. William J Greenleaf, Catherine A Blish.	Publication		Published