A multimodal single-cell reference dataset of the human bone marrow

I. Motivation

A. Why was the dataset created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

The last decade has witnessed a technological arms race to encode the molecular states of cells into DNA libraries, turning DNA sequencers into scalable single-cell microscopes. Single-cell measurement of chromatin accessibility (DNA), gene expression (RNA), and proteins has revealed rich cellular diversity across tissues, organisms, and disease states. Recent advances in multimodal single-cell technologies that can measure two or more of these layers, such as joint profiling of DNA and RNA, are a major step towards developing integrative models of the genetic regulatory programs that organize biology. However, single-cell data poses a new set of challenges for biomedical data science, and multimodal datasets compound those difficulties. We sought to create a high quality reference dataset that can be used to benchmark future developments in multimodal single-cell algorithms. The dataset was created with three benchmarks in mind involving multimodal single-cell data integration. The tasks involve predicting one modality from the other, matching profiles from each modality, and learning meaningful embeddings of jointly profiled data. We hope in the future there are new tasks that will be formulated off this dataset involving denoising, visualization, and others we can't anticipate.

B. Has the dataset been used already? If so, where are the results so others can compare?

This dataset is being used in the NeurIPS 2021 Multimodal Single-cell Data Integration Competition. The competition rules and results are accessible at `https://openproblems.bio/neurips_2021`.

C. What (other) tasks could the dataset be used for?

In addition to benchmarking, this dataset represents one of the largest multimodal atlases of the human bone marrow. We expect this resource will be useful to researchers in hematology and immunology who seek to understand the diversity of cell states in this highly proliferative niche responsible for all immune cells.

D. Who funded the dataset creation?

Major funding for this dataset was provided by the Chan Zuckerberg Initiative under grants DAF2021-235155 and DAF2021-235076. Support for incidental reagents and technician time were provided by Cellarity, Chan Zuckerberg Biohub, and Helmholtz Munich, and the Yale University Center for Genome Analysis.

II. Dataset Composition

A. What are the instances?

This dataset comprises single-cell profiles of human bone marrow mononuclear cells from human donors. Samples from each donor were measured using two multimodal technologies. The first technology measures cell surface protein markers using antibody-derived tags (ADT) and RNA gene expression (GEX). The second technology jointly measures DNA accessibility using to an technique called the assay for transposase-accessible chromatin (ATAC) and RNA gene expression (GEX) in single cells.

The dataset generation was designed to generate a nested batch structure across multiple sites as shown in **Figure 1**. We picked a single donor to be used as a reference across all four sites. The remaining 8 donors were distributed randomly 2 per site. The data was then split with samples from 3 sites used for training and one use for test. This enables methods to learn to generalize across donors and sites. The shared reference donor sample in the test set provides an anchor between the test and training sets.

The data is also annotated using expert curated cell type markers. A description of the annotation process is available in the Appendix of the accompanying manuscript.

B. How many instances are there in total (of each type, if appropriate)?

There are roughly 150,000 profiles in the raw data evenly split between the ATAC+GEX and ADT+GEX modalities. Filtering and preprocessing removes roughly 20% of samples, yielding an expected total of 120,000 cells. This number will be finalized after processing of each sample is finished.

C. What data does each instance consist of?

We measured the accessibility of 119,254 genomic regions, the expression of 15,189 genes, and the abundance of 134 surface proteins with ATAC+GEX and ADT+GEX in a multi-site, multi-donor dataset of a complex biological system.

Donor ID	Donor Age	Donor BMI	Donor Blood Type	Donor Race	Ethnicity	Donor Gender
14866	26	29.2	O+	Other Race	HISPANIC OR LATINO	Male
12710	27	32.1	O+	White	NOT HISPANIC OR LATINO	Male
10886	35	28.6	B+	Asian	NOT HISPANIC OR LATINO	Female
18303	33	24	O+	Asian	NOT HISPANIC OR LATINO	Male
28045	36	23.8	A+	Other Race	HISPANIC OR LATINO	Female
23356	34	24.7	O+	White	HISPANIC OR LATINO	Male
16710	40	27.8	O+	White	HISPANIC OR LATINO	Female
11466	22	31.5	A+	Asian	NOT HISPANIC OR LATINO	Female
15078	34	24.8	B-	White	HISPANIC OR LATINO	Male

TABLE I

DEMOGRAPHIC INFORMATION OF THE 9 HUMAN DONORS FROM WHOM BONE MARROW MONONUCLEAR CELLS WERE SOURCED.

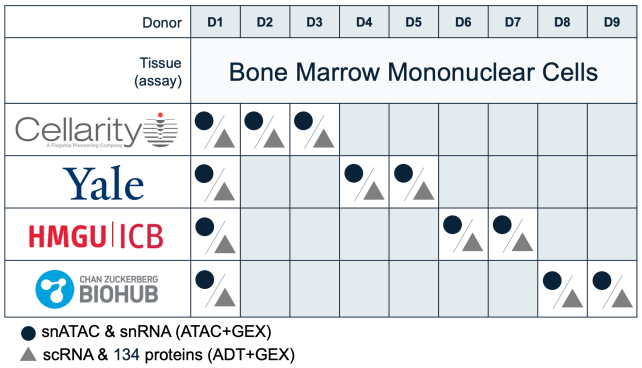

Fig. 1. A schematic showing the design of the dataset generation. The assignment of donors to sites was randomized, along with the train / test split. Nine human donors were measured using ATAC+GEX and ADT+GEX across 4 sites. One donor was measured at all sites with 2 donors measured uniquely at each site.

Each instance in the dataset is a single-cell measured using either ATAC+GEX joint profiling or ADT+GEX joint profiling. Each observation is indexed by a unique cellular barcode [1] that associated the profiles in either modality. For the GEX profiles, the features of the dataset are raw counts of unique molecular identifiers (UMI) that represent the absolute number of observed RNA molecules in each cell. For the ATAC profiles, the features are the number of reads falling in ATAC peaks as described in the documentation for CellRanger Arc v2.0 https://support.10xgenomics.com/single-cell-multiome-atac-gex/software/pipelines/latest/algorithms/overview. For the ADT profiles, the features are the UMI counts associated with each of the ADTs detected in each sample.

Each cell in each dataset is also associated with the donor ID and corresponding metadata associated with the sample, the site ID at which the sample was processed, the cell type annotation, the percentage mitochondrial content (a measure of cell health), pseudotemporal ordering for a subset of the cells [2], and the cell cycle phase.

D. Is there a label or target associated with each instance?

In our competition, we use the joint profiles as the target for each instance, akin to the multiple languages a sentiment is in for machine translation tasks. We also include a data integration task in which the cell type labels, cell cycle, pseudotime, and batch labels are used to measure how effectively batch effects are removed while preserving biology.

E. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing.

Only complete instances were included in this dataset.

F. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Profiles of the same cell are made explicit through the cell barcodes. Samples from the same donor or site are linked by the donor and site IDs, respectively. Cells of similar type are linked by the cell type identifiers, but these labels were associated with each dataset independently.

G. Does the dataset contain all possible instances or is it a sample of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not?

These samples were generated from bone marrow mononuclear cells of human donors. We filtered out granulocytes because presence of these cells can lead to sample quality deterioration [3]. We also removed doublets from the dataset using a procedure described in the Appendix of the associated submission.

H. Are there recommended data splits? If so, please provide a description of these splits, explaining the rationale behind them.

How the benchmark data is split into train, validation and test data is determined by its batch structure. As shown in **Figure 1**, our dataset nests donor batches within data generation sites. We refer to a unit in this nested structure as a sample. While cellular profiles can strongly differ depending on their identity, each sample contains broadly the same cellular identities in varying proportions (**Figure 1f,g**). Differences between cellular profiles of the same identity across samples are dominated by batch effects.

As we are challenging algorithms to overcome these batch effects, the data must be split by samples or sites. For the

final round of the NeurIPS competition we are holding out all three samples from one site and provide all samples from the three other sites as training data. To simulate a train-test split on the currently available samples, we recommend using one sample as test data. Here, it is advisable to avoid using donor 1 as test sample as data from this donor is included multiple times from different sites. For validation purpose, we suggest using an N-fold cross-validation approach where N refers to the number of samples in the training data.

I. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Single-cell data is subject to under-counting due to the small amount of starting material in each cell for biochemical reactions. The nature of the exact noise pattern has been long contested [4]. For droplet-based single-cell GEX profiles, such as those in this dataset, the consensus is emerging that the undercounting follows a negative binomial distribution [5]. As a result, many methods for denoising scRNA-seq have been proposed, with several noteable benchmarks [6], [7]. However, even the nature of benchmarking denoising methods is under dispute [8].

Research into noise models for single-cell ATAC profiles is newer, but some methods for denoising have been described [9], [10], [11]. There is also some literature on denoising ADT data [12].

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained.

III. Collection Process

A. What mechanisms or procedures were used to collect the data? How were these mechanisms or procedures validated?

Protocols for sample preparation and collection were based on validated protocols associated with commercially available products validated by the vendors of each product. Detailed experimental protocols will be deposited at the public protocol sharing platform protocols.io shortly after submission. Copies of the protocols can be found attached to this datasheet.

B. How was the data associated with each instance acquired? Was the data directly observable, reported by subjects, or indirectly inferred/derived from other data?

Human bone marrow mononuclear cells were sourced from AllCells (Alameda, CA). Access to donor samples was limited due to supply chain issues associated with COVID-19, but we were able to achieve moderate diversity of samples with equal representation of male and female, Hispanic and non-Hispanic, and white and non-white donors.

To generate ADT+GEX data, we used the 10X Genomics (Pleasanton, CA) Single Cell Gene Expression 3' v3.1 with Feature Barcoding using the Biolegend (San Diego, CA) TotalSeq-B Human Universal Cocktail v1.0. Data was generating following the ATAC protocol attached to this datasheet.

To generate ATAC+GEX data, we used the 10X Genomics Chromium Next GEM Single Cell Multiome ATAC + Gene Expression kit. Data was generated following the CITE protocol attached to this datasheet.

This data is directly observable using DNA sequencing technology.

C. Who was involved in the data collection process and how were they compensated?

Data was generated by a consortium of graduate students, postdoctoral fellows, and laboratory technicians at the sites responsible for generating the data. This project falls within the regular work duties of each contributor, who are are paid employees of their respective institutions. As such, they did not receive additional compensated for this work.

D. Over what timeframe was the data collected?

The data was generated in July and August 2021.

IV. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

B. Gene expression data

The data was prepossessed according to current best practices in single-cell analysis [13]. We used the Scanpy platform [14] as a basis for quality control, normalization, dimensionality reduction, clustering, feature selection, and trajectory inference.

C. Open chromatin data

The chromatin accessibility data acquired by ATAC-seq as part of the 10X Multiome protocol was processed using Signac version 1.3.0 [15], an extension of the Seurat toolkit version 4.0.3 [16], and the Scanpy platform version 1.7.2 [14]. To ensure the same set of features across samples, accessible regions (also referred to as peaks) were called jointly using Cell Ranger arc version 2.0.0. Quality control, dimensionality reduction and translating peaks to gene activity scores was performed using Signac, following the authors' instructions. Downstream analysis steps including cell type annotation and trajectory inference were done in Scanpy.

D. Protein data

The workflow of analyzing cell surface protein levels captured as antibody-derived tags (ADT) in the CITE-seq protocol was adapted from our pipeline to process gene

expression data and mainly performed using the Scanpy platform version 1.7.2 [14]. The TotalSeq-B antibody panel from BioLegend Inc. used in this study comprises 134 primary antibodies capturing human cell surface proteins and 6 isotype controls without any human target protein that can be used to assess the level of unspecific binding in each cell. A full description of the analysis can be found in the accompanying manucript and notebooks to reproduce the analysis will be made available on GitHub at https://github.com/openproblems-bio/neurips2021-notebooks.

E. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Raw sequencing data will be uploaded to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository after publication.

F. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

The computational pipelines for generating counts matrices and ATAC profiles are available on GitHub under an open-source license at `https://github.com/czbiohub/utilities`. Notebooks for dataset processing are available on GitHub under an open-source license at `https://github.com/openproblems-bio/neurips2021-notebooks`.

G. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

A major limitation of this dataset is that each donor sample has no technical replicates per site. This limitation arose from a lack of access to additional donor samples and funding limitations. This dataset also only measures a single tissue in a single species. Future work is likely to increase the diversity of tissues measures in the dataset.

V. Dataset Distribution

A. How will the dataset be distributed?

During the competition, the datasets will be made available via a public Amazon Simple Storage Service (S3) bucket at s3://openproblems-bio/public. Each dataset is stored in two AnnData objects [14], one for each modality. After the competition, the datasets will be made available at the CZI cellxgene portal at `https://cellxgene.cziscience.com/`.

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

The dataset will be released in September 2021 under a CC-BY License.

C. Are there any copyrights on the data?

No.

D. Are there any fees or access/export restrictions?

No.

E. Any other comments?

No.

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

During the competition, hosting of the dataset is provided by Saturn Cloud (New York, NY). After the competition, the dataset will be hosted by the Chan Zuckerberg Initiative.

B. Will the dataset be updated? If so, how often and by whom?

We do not anticipate the dataset will be updated.

C. How will updates be communicated?

In the event that we need to communicate updates, we will log them in the CZI cellxgene portal version notes.

D. If the dataset becomes obsolete how will this be communicated?

The dataset will be removed from the CZI cellxgene portal.

E. Is there a repository to link to any/all papers/systems that use this dataset?

No.

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

We designed this dataset to facilitate extension, augmentation, and validation. All protocols were performed using commercially available reagents. The source for cells, sample preparation, and the sequencing procedures are freely available on protocols.io. Interested parties may contact the organizers listed on `https://openproblems.bio` to indicate their intent to augment the dataset. Individuals interested in expanding on the data annotations may similarly contact the organizers at the above address.

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

There was no ethical review process conducted as all samples were obtained under a permissive universal consent form that explicitly provides consent for public distribution of sequencing data.

B. Does the dataset contain data that might be considered confidential? If so, please provide a description.

No.

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

No.

D. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

E. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

See **Table I**.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

It has been shown that it is possible to identify individuals uniquely using genomic information available in public repositories [17]. However, we note that the donors whose tissue is used in this study explicitly consented to distribution of their genomic samples in an unrestricted scientific database like GEO.

The following is an excerpt from the Informed Consent (IC) form used by AllCells:

> It is possible that genomic information (data) will be generated during research using your sample. This data will be freely available in a public, unrestricted scientific database that anyone can use (e.g. GEO, ENCODE portal). The public database will include information on hundreds of thousands of genetic variations in your DNA code, as well as your age, ethnic group and sex. The only health information included will be that you are a healthy volunteer. This public information will not be labeled with your name or other information that could be used to easily identify you. However, it is possible that the information from your genome, when combined with information from other public sources could be used to identify you, though we believe it is unlikely that this will happen.

A full copy of the IC form can be found at the end of the datasheet.

G. Does the dataset contain data that might be considered sensitive in any way (i.e., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

This dataset contains demographic information as described in **Table I**. The data also includes genomic information as discussed above and below.

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Data was obtained using tissue samples sources from a third party, AllCells.

I. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Donors were notified about data collection when they volunteered to donate samples. The notification was performed by AllCells using the IC form appended to this datasheet.

J. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

All donors freely signed the IC form appended to this datasheet. Signed consent forms are maintained by AllCells.

K. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

As described in the IC form:

> Your participation as a research subject is strictly voluntary. You have the right to discontinue your participation at any time before or after commencement of the procedure without penalty or loss of benefits. You may refuse to donate samples without penalty or consequence to the care provided to

you by the study doctor, associated physicians, the nurses, and the research staff.

If you decide to withdraw your consent after the donation, you may request that your samples be removed and destroyed if they have not yet left LeukoLab. However, once your samples depart LeukoLab, they cannot be retrieved.

(*emphasis theirs*).

Contact information to revoke consent was provided on the first page of the IC form.

L. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

We did not conduct a DIPA as this research does not pose a "high risk" to the rights and freedoms of the donors as defined by the GDPR Guidelines `https://gdpr.eu/data-protection-impact-assessment-template/`. We note that although GDPR makes note of new technologies, the sequencing of DNA is the relevant technique that relates to subject's personal rights and freedoms. This technology has been widely used over the past 20 years, and the donors explicitly consented to unrestricted distribution of their genomic data.

References

[1] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells - Nature Communications. *Nat. Commun.*, 8(14049):1–12, Jan 2017.

[2] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods - Nature Biotechnology. *Nat. Biotechnol.*, 37(5):547–554, May 2019.

[3] Elliott Swanson, Cara Lord, Julian Reading, Alexander T. Heubeck, Palak C. Genge, Zachary Thomson, Morgan D. A. Weiss, Xiao-jun Li, Adam K. Savage, Richard R. Green, Troy R. Torgerson, Thomas F. Bumol, Lucas T. Graybuck, and Peter J. Skene. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife*, Apr 2021.

[4] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics - Nature Methods. *Nat. Methods*, 11(6):637–640, Jun 2014.

[5] Valentine Svensson. Droplet scRNA-seq is not zero-inflated - Nature Biotechnology. *Nat. Biotechnol.*, 38(2):147–150, Feb 2020.

[6] Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C. Hicks. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.*, 21(1):1–30, Dec 2020.

[7] Tallulah S. Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7, 2018.

[8] Wei Vivian Li, Jingyi Jessica Li, Wei Vivian Li, and Jingyi Jessica Li. Issues arising from benchmarking single-cell RNA sequencing imputation methods. Aug 2019. [Online; accessed 26. Aug. 2021].

[9] Rongxin Fang, Sebastian Preissl, Yang Li, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau, Xinzhu Zhou, Fangming Xie, Eran A. Mukamel, Kai Zhang, Yanxiao Zhang, M. Margarita Behrens, Joseph R. Ecker, and Bing Ren. Comprehensive analysis of single cell ATAC-seq data with SnapATAC - Nature Communications. *Nat. Commun.*, 12(1337):1–15, Feb 2021.

[10] Zhicheng Ji, Weiqiang Zhou, Wenpin Hou, and Hongkai Ji. Single-cell ATAC-seq signal extraction and enhancement with SCATE. *Genome Biol.*, 21(1):1–36, Dec 2020.

[11] Avantika Lal, Zachary D. Chiang, Nikolai Yakovenko, Fabiana M. Duarte, Johnny Israeli, and Jason D. Buenrostro. Deep learning-based enhancement of epigenomics data with AtacWorks - Nature Communications. *Nat. Commun.*, 12(1507):1–11, Mar 2021.

[12] Matthew P. Mulè, Andrew J. Martins, and John S. Tsang. Normalizing and denoising protein expression data from droplet-based single cell profiling. *bioRxiv*, page 2020.02.24.963603, Feb 2021.

[13] Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, 15(6):e8746, Jun 2019.

[14] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):1–5, Dec 2018.

[15] Tim Stuart, Avi Srivastava, Caleb Lareau, and Rahul Satija. Multimodal single-cell chromatin analysis with signac. *bioRxiv*, 2020.

[16] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, may 2021.

[17] Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. Identifying Personal Genomes by Surname Inference. *Science*, 339(6117):321–324, Jan 2013.

Study Title: Bone Marrow Collection from Healthy Donors for the Research Market

Sponsor: LeukoLab, A Clinical Division of AllCells, LLC

Protocol Number: 7000-SOP-046
Protocol Version: 9.0 (Alpha IRB)

Principal Investigator: <<PI Name>>

<<Site Name>>
<<Site address>>

Telephone: <<phone number>>
After Hours: <<24 hr phone number>>

TEMPLATE - NOT FOR SUBJECT USE

Background

Over the past several decades tremendous progress has been made in treating cancer and blood diseases both through the research and development (R&D) of new drugs and therapies, and more recently from tests being developed to target these therapies. Academic institutions, pharmaceutical, biotechnology, and other organizations involved in conducting or supporting biomedical research require donation of various types of blood or bone marrow cells. Additionally, R&D may involve analysis of different cell types and other substances contained in these tissues including, but not limited to proteins, enzymes, hormones, DNA and RNA. Your tissue donation will remain anonymous (de-identified) to scientific researchers. However information and research data pertaining to such de-identified tissue may possibly be published for review by other members of the biomedical research community.

The sponsor AllCells, Limited Liability Company of Alameda, California, (along with its blood collection division known as LeukoLab) is a private company that collects, processes and sells blood cells to scientists who require bone marrow cells on which to conduct this type of research.

This informed consent form has information to help you decide if you want to participate in a research study. Before you agree to take part in this study, it is important for you to read all of the information that follows. If there are any words or information that you do not understand, your study doctor, or the research staff, will explain them to you and answer any other questions you may have regarding this study. Reading this form and discussing it with your study doctor or the research staff can help you decide whether or not to participate. If you decide to become a donor, you must sign, initial and date this form before you participate.

Purpose of the Study

AllCells, LLC/LeukoLab wishes to obtain sample(s) of bone marrow aspirate (the liquid part of the bone marrow) from health donors for the research market. It is well recognized that up to 20mL of bone marrow per kilogram of body weight can be safely removed from donors for bone marrow transplantation. For an adult weighing 140

pounds, this would be 1,272 mL or approximately 5 cups. The samples being collected today at LeukoLab are never intended for transfusion or other use in humans or animals.

Procedure

Bone marrow is the blood-forming organ of the body and is found in bones throughout the human body. A physician or nurse practitioner will perform a procedure called a bone marrow aspiration to obtain a sample of bone marrow.

Your samples, genomic data (information about your genetic makeup), health and demographic information will be stored and shared with other researchers. The samples and information will be available for any research question, such as research to understand what causes certain diseases (for example heart disease, cancer, or psychiatric disorders), development of new scientific methods, or the study of where different groups of people may have come from.

Although we will not give researchers your name, we will give them basic information such as your race, ethnic group, geographic region, age range, and sex. This information may help researchers' study whether the factors that lead to health problems are the same in different groups of people. It is possible that such findings could one day help people in the same groups as you.

Description of Bone Marrow Collection:

You will lie down on an exam table, on your stomach with your head resting on your arms. Bone Marrow will be collected either using a manual technique or with a device called Marrow Miner (an FDA approved medical device). The skin over the upper back part of your hip bone(s), known as your pelvic bone will be thoroughly cleansed using an antiseptic solution. A local anesthetic (numbing medication) called lidocaine (or Marcaine) will be slowly injected through a small needle into your skin at the right and left hip bone site, called the iliac crest. Additional lidocaine (numbing medication) will be injected into the surface of the bone at each of these sites. When the areas are numb, a needle will be passed through the bone then into the bone marrow space itself either on the right or left hip bone site first, followed by the other hip bone side. For each hip bone side, between 50 - 200 mL (4 tablespoons to less than 1 cup) of the liquid marrow will be gently pulled into a syringe(s). At times, the needle may need to be removed and reinserted into the same numbed area. After obtaining the targeted volume of bone marrow, the needle will be removed and a pressure bandage will be placed on the area(s) of skin where the needles were inserted. You will be instructed to leave your bandage in place until the following morning.

The removal of 50-300 mL of bone marrow during this procedure will in no way affect the normal function of the rest of your bone marrow. Your body will replace the fraction of bone marrow removed within a matter of hours.

Duration

For each needle site, the needle will be in your hip bone(s) for approximately 1 – 4 minutes each site. The entire procedure should take no longer than 60 minutes.

TEMPLATE - NOT FOR SUBJECT USE

POTENTIAL SIDE EFFECTS, RISKS AND DISCOMFORTS

If your Bone Marrow donation is used by researchers to study your genome (genetic makeup), and if this information is accidently released by the researchers, it might be possible that the information gathered about you as part of this study could become available to someone else outside the study.

It is possible that genomic information (data) will be generated during research using your sample. This data will be freely available in a public, unrestricted scientific database that anyone can use (e.g. GEO, ENCODE portal). The public database will include information on hundreds of thousands of genetic variations in your DNA code, as well as your age, ethnic group and sex. The only health information included will be that you are a healthy volunteer. This public information will not be labeled with your name or other information that could be used to easily identify you. However, it is possible that the information from your genome, when combined with information from other public sources could be used to identify you, though we believe it is unlikely that this will happen.

Your samples, genomic data and health information as described above will be stored by the recipient researcher and shared with other researchers working on the same project (consortium) or for other unrelated research projects. The samples and genomic data may be used for future research for an unlimited period of time. The samples and information will be available for any research question, such as research to understand what causes certain diseases (for example heart disease, cancer, or psychiatric disorders), development of new scientific methods, to deeply characterize (define) cells using a variety of genomic tests or the study of where different groups of people may have come from. It is also possible that research findings could be used inappropriately to support negative stereotypes, stigmatize, or discriminate against members of the same groups as you.

A Federal law, called the Genetic Information Non-discrimination Act (GINA), generally makes it illegal for health insurance companies, group health plans, and most employers to discriminate against you based on your genetic information. However, this Federal law does not protect you against genetic discrimination by companies that sell life insurance, disability insurance, or long-term care insurance.

Possible Risks with Anesthetics

Risks associated with the use of the local anesthetic Lidocaine (or long acting local anesthetic Marcaine with or without Epinephrine) include burning or pain at the site of injection and or numbness or tingling in the area. The study team will look for any allergic reactions you may experience that are caused by the anesthetic. If you receive local anesthetic, you may experience headaches and an inability to pass urine for a short period of time.

Needle Insertion for Bone Marrow Collection

The amount of bone marrow removed from your body is a small fraction of your total bone marrow and its removal will in no way affect the normal function of the rest of your bone marrow. Serious side effects are rarely seen with this common procedure. However, minor bleeding at the skin puncture site, bruising, infection, lightheadedness, fainting, or temporary numbness and/or tingling to the face and/or extremities may be experienced. You will be monitored closely by our clinical staff during the procedure for

DRAFT AFF - NOT IRB SUBJECT TO

any adverse reactions. Emergency medications and equipment are available if needed as well as access to 911.

Rarely, the needle may go into a "dry" portion of the bone (no liquid bone marrow present) and an additional needle insertion would be necessary.

Most donors do not need to take any pain medication after donation. If you do, we suggest you use Acetaminophen (Tylenol). You should avoid taking Ibuprofen (Motrin, Advil), aspirin or Naproxen for at least 24 hours after your donation, as these drugs may increase risk of bleeding at the site.

Activity post-donation will be as you tolerate it; though you may experience less soreness in your hip area if you minimize strenuous activity the day of your donation.

Pregnancy
You may not enter the study if you are pregnant or trying to become pregnant. There are no known risks of the study procedures which would affect pregnant women or unborn children. The study doctors have decided, however, not to include this population in order to avoid any potential unknown risks. You must use a reliable method of birth control to prevent pregnancy during your participation in this program.

Frequency of Donations
Donations up to 50 mL may be repeated once, after a waiting period of 6 weeks (42 days). Donations of 51 - 300 mL can be repeated every 12 weeks (84 days). Maximum donation volume in any 12-month period (365 days) is 1,000 mL. There is no life-time donation volume limit.

Your Responsibilities
You will be expected to arrive on time and well nourished for your appointment. Compensation for participation is dependent upon completion of the study procedure.

Benefits
There will be no direct scientific or clinical benefit to you, nor will you gain any commercial or financial rights from participating in this donor program. These samples will be used only for scientific research. You will not be given any results of the research or stored samples.

No Ownership of Donation
You will retain no ownership interest whatsoever in your donation. You will have no ownership interest whatsoever in anything that is produced with the donation or in any invention that is made with the use of the donation. After you sign this informed consent form, your marrow donation and any derivative may be used, modified, altered, transferred and disposed of without your notice or further consent.

Infectious Agent Testing
In order to qualify as a study participant, at least 3 business days prior to the bone marrow donation, the LeukoLab staff will test your blood for human immunodeficiency virus (HIV), hepatitis B virus (HBV) and hepatitis C virus (HCV). If requested by the researchers who will be studying your blood, additional testing may be performed. These tests may include, but are not limited to HTLV (human T-lymphocyte virus), CMV

SAMPLE - NOT FOR SUBJECT USE

(cytomegalovirus), TSE (human transmissible spongiform encephalitis), CJD (Creutzfeldt-Jakob disease), West Nile virus, T.cruzi (Chagas' Disease), Zika Virus and syphilis. Any positive results will be reported to you, and you will be referred to your private physician for follow up. In some circumstances, there may be false positive results and LeukoLab will recommend that the tests should be repeated by your private physician before determining the significance of a positive result.

Positive results may be reported to the State Department of Public Health by the lab performing the tests, as required by law and your name may be required for identification purposes. Otherwise, all samples will be identified with a code number only – no names will be used. If you test positive for any of these infectious agents, you will not be able to continue participation in this study.

Unforeseen Risks

There may be risks to the blood marrow aspirate collection procedure that are not yet known. You will be informed in a timely manner both verbally and in writing of any new information or findings that might affect your willingness to continue participation in the study.

Alternatives to Participation

This is a study for healthy bone marrow donors. Your alternative to participation in the study is to not be in this study.

Confidentiality

You are participating in this study as an anonymous study donor. A unique number will be assigned to you and to the sample you donate, to protect your identity. The study doctor, contracted laboratory for infectious disease testing and the research staff at this facility will have access to records that link the sample number(s) with your identifying information. Your records may also be reviewed for audit purposes by authorized customer or their collaborators. They will be bound by the same provisions of confidentiality. ***Neither your name nor any other identifying information will be revealed to the auditors or the research team that will eventually receive your donation***.

The research staff at this facility will retain records linking your identity to your sample for an indefinite period of time. Your study related medical information which may contain your name and other direct identifiers may be submitted to AllCells, LLC, LeukoLab, A Clinical Division of AllCells, collaborators or sponsor representatives, Alpha Independent Review Board (Alpha IRB), a group that reviews research to protect research subject's rights and welfare, the Department of Health and Human Services (DHHS), and/or other regulatory authorities as required by law. Confidentiality will be protected to the extent allowed by the law; however, absolute confidentiality cannot be guaranteed.

Voluntary Participation

Your participation as a research subject is strictly voluntary. You have the right to discontinue your participation at any time before or after commencement of the procedure without penalty or loss of benefits. You may refuse to donate samples without penalty or consequence to the care provided to you by the study doctor, associated physicians, the nurses, and the research staff.

SAMPLE DONOR SUBJECT USE

If you decide to withdraw your consent after the donation, you may request that your samples be removed and destroyed if they have not yet left LeukoLab. However, once your samples depart LeukoLab, they cannot be retrieved.

The study doctor or LeukoLab staff can stop your participation at any time without your consent for the following reasons:

- If it appears to be medically, mentally or emotionally harmful to you;
- If you fail to follow directions for participating in the study;
- If it is discovered that you do not meet the study requirements;
- If the study is canceled

Costs/compensation

There will be no cost to you for agreeing to donate your bone marrow. There is no charge to you or your insurance company for obtaining the samples for research or for the research performed using your samples.

For your time and effort related to your participation, you will be compensated as follows:

Volume Collected:	Up to 50 mL =	XXXXXX
	Up to 100 mL =	XXXXXX
	Up to 200 mL =	XXXXXX
	Up to 300 mL =	XXXXXX

At the completion of each visit a cash card will be issued to you. For first time donors, the card will be mailed to the address you provided to LeukoLab. For all other visits, the cash card will be given to you at the end of your visit. Funds will be available immediately upon receipt of your card.

Research-related injury

In the event of a research-related injury, medical care will be available. However, there will be no compensation for treatment of a research-related injury, and payment for such care will be the responsibility of you and/or your insurance company. Your insurance, including Medicare, may or may not cover these charges. There will be no payment for other things, such as disability, transportation, or loss of wages. The doctor, this facility, and AllCells, LLC/LeukoLab will not pay for treatment of research-related injuries, although you are not precluded from seeking to collect compensation for injury related to malpractice, fault, or blame on the part of those involved in the research.

Questions

If you have any questions about the study, the sample donation process or the uses of the bone marrow samples, or if you have a research related injury, please contact the study doctor or the research staff at the phone number(s) listed on page 1 of the consent form.

If you have questions, concerns or complaints about your rights as a research volunteer or about taking part in this study, or to obtain information or offer input, you may

TEMPLATE - NOT FOR SUBJECT USE

contact Alpha IRB, Attn: Marianne Thornton, toll free at (888) 265-5766 between the hours of 8:00am-5:00pm Pacific Time.

Alpha Independent Review Board
1001 Avenida Pico, Suite C #497
San Clemente, CA 92673
(888) 265-5766 (toll free)

Alpha IRB is a group of people who perform independent review of research studies to protect the rights and welfare of study participants. Although Alpha IRB has approved the information provided in this informed consent form and has approved for the study doctor to do the study, this does not mean Alpha IRB has approved you being in the study, or that the study is without risks. You must consider the information in this consent form for yourself and decide if you want to be in this study.

TEMPLATE - NOT FOR SUBJECT USE

Consent to Participate

Your signature below indicates that you have read the above information and have had adequate time to ask questions regarding your participation in this study. You have completed the history forms with required screening and medical information accurately and truthfully. Questions have been answered to your satisfaction. You agree to participate as a donor in the study as it is described above. You are aware that any specimens you donate will be used for scientific research only, and you will not be paid as a result of the commercialization of any product, process, or service developed from your cells, blood, blood marrow or other specimens. You will receive copies of this signed and dated consent form and the Experimental Subjects Bill of Rights form.

You do not give up any legal rights by signing this informed consent form.

_____ I agree to donate up to 50 mL bone marrow and understand that I will be compensated XXXXXX.

_____ I agree to donate up to 100 mL bone marrow and understand that I will be compensated XXXXXX.

_____ I agree to donate up to 200 mL bone marrow and understand that I will be compensated XXXXXX.

_____ I agree to donate up to 300 mL bone marrow and understand that I will be compensated XXXXXX.

_____	_____	_____
Donor/Subject's Printed Name	Donor/Subject's Signature	Date
_____	_____	_____
Person Conducting Consent Printed Name	Person Conducting Consent Signature	Date

TEMPLATE - NOT FOR SUBJECT USE

Experimental Research Subject's Bill of Rights

Any person asked to take part as a subject in research involving a medical experiment, or any person asked to consent to such participation on behalf of another, is entitled to receive the following list of rights written in a language in which the person is fluent. This list includes the right to:

1. Be informed of the nature and purposes of the experiment.
2. Be given an explanation of the procedures to be followed in the medical experiment and any drug or device to be utilized.
3. Be given a description of any attendant discomforts and risks reasonably to be expected from the experiment.
4. Be given an explanation of any benefits to the subject reasonably to be expected from the experiment, if applicable.
5. Be given a disclosure of any appropriate alternative procedures, drugs, or devices that may be advantageous to the subject, and their relative risks and benefits.
6. Be informed of the avenues of medical treatment, if any, available to the subject after the experiment if complications should arise.
7. Be given an opportunity to ask any questions concerning the experiment or the procedures involved.
8. Be instructed that consent to participate in the medical experiment may be withdrawn at any time and the subject may discontinue participation in the medical experiment without prejudice.
9. Be given a copy of the signed and dated written consent form.
10. Be given the opportunity to decide to consent or not to consent to a medical experiment without the intervention of any element of force, fraud, deceit, duress, coercion, or undue influence on the subject's decision.

TEMPLATE, NOT FOR SUBJECT USE

_____	_____	_____
Donor/Subject's Printed Name	Donor/Subject's Signature	Date

ATAC-Seq protocol Chan Zuckerberg Biohub

1 *Works for me*

Maurizio Morri
Chan Zuckerberg Biohub

PROTOCOL INFO

Maurizio Morri: ATAC-Seq protocol Chan Zuckerberg Biohub. **protocols.io**
https://protocols.io/view/atac-seq-protocol-chan-zuckerberg-biohub-bxpupmnw

CREATED

Aug 25, 2021

LAST MODIFIED

Aug 27, 2021

PROTOCOL INTEGER ID

52692

Cell Preparation

1 Cells were removed from liquid nitrogen storage.

2 Frozen cells should be kept on dry ice until ready for thawing.

3 Recommended media for diluting and washing cells are RPMI 1640, IMDM, or DMEM supplemented with 10% heat inactivated FBS warmed to room temperature (20-25°C).

4 Proper aseptic technique should be used when handling and manipulating cells.

5 Check that the vial is tightly capped and place in a 37°C water bath.

6 After 90 seconds (1.2mL vial fill) or 2 minutes (1.8mL vial fill) at 37°C, remove the vial from the water bath every 15-20 seconds and gently invert 3-4 times to check the level of thawing.

7 When the vial contents are 50-75% liquid, remove from the water bath and continue to gently invert the vial until the entire contents are liquid.

8 Place the vial on wet ice and transport it to a biosafety cabinet.

9 Wipe the outside of the vial with an alcohol wipe and gently invert 5 times to mix the cells

10 Check to see if there is liquid caught up in the cap. If yes, gently tap the upright vial on a benchtop to force the liquid into the vial.

11 Immediately remove 20µL of the contents for cell concentration and viability determination. Keep the sample on ice until assayed.

12 Transfer the remaining vial contents to a 15mL tube at RT (20-25°C).

13 Add 1mL of media to rinse the vial and transfer any remaining cells to the 15mL tube.

14 Slowly add 10mL media (about 1mL per second) to the cells. Securely cap the tube and gently invert 3-5 times after adding all the media.

15 Centrifuge the tube at 200g x 8 minutes at RT (20-25°C).

16 Aspirate the supernatant without disturbing the pellet and resuspend the cells in 12mL media.

17 Centrifuge the tube at 200g x 8 minutes at the RT and resuspend in 1ml of cold PBS+0.04% BSA.

18 Transfer the resuspended cells to a 2-ml microcentrifuge tube. Rinse the 15ml tube with 0.5ml PBS+0.04% BSA and transfer the rinse to the 2ml tube containing the cells. Mix by gently inverting the tube.

19 Centrifuge cells at 300rcf for 5 min.

20 Remove the supernatant using a 1ml pipette without disturbing the cell pellet by leaving behind a little supernatant and resuspend the pellet gently in 1 ml PBS + 0.04% BSA.

21 Pass cell suspension through 40uM Flowmi Cell Strainer.

22 Determine the cell concentration using an Automated cell counter.

23 Proceed to cell sorting. Do not sort cells if the cell count is <100,000.

Cell sorting pre atac-seq

24 Granulocytes were removed using a size-based approach on a SONY-SH800 sorter.

25 Collect the sorted cells in a 5-ml FACS tube containing 500 µl PBS + 0.04% BSA.

26 Centrifuge the collected cells at 3**00 rcf for 7 min (instead of 5 mins)** at 4°C.

27 Remove the supernatant without disrupting the cell pellet and resuspend in 500 µl PBS + 0.04% BSA.

28 Determine the cell concentration using an Automated Cell Counter. Place the cells on ice until ready to proceed.

29 Proceed to nuclei isolation from 100,000-1,000,000 cells using the protocol below.

30 Nuclei isolation was performed using the protocol suggested from 10x (page 5)
Nuclei Isolation for Single Cell Multiome ATAC + Gene Expression Sequencing
CG000365_DemonstratedProtocol_NucleiIsolation_ATAC_GEX_Sequencing_RevB.pdf

31 The following 10X protocol was used for the preparation of ATAC-seq libraries
CG000338_ChromiumNextGEM_Multiome_ATAC_GEX_User_Guide_RevD.pdf

Sequencing

32 Pooled libraries were run on an Agilent 4150 TapeStation for visual inspection, and quantified by qPCR, performed on a BioRad CFX96 RT PCR thermal cycler using the KAPA library quantification kit (#KK4923) for quality control assurance prior to sequencing.

33 ATAC multiome libraries were sequenced on an Illumina NextSeq 500 yielding paired-end reads with the following sequencing parameters: Read 1= 50 cycles, Index 1=8 cycles, Index 2= 24 cycles, Read 2= 49 cycles. A custom recipe was necessary to include 8 dark cycles and 16 nt cycles for Index 2 for this specific sequencer. The targeted sequencing depth was 25,000 reads per nucleus. BCLs were converted to FASTQ files and demultiplexed using cellranger-atac-1.2.0.

ATAC 3' gene expression libraries were sequenced on an Illumina NextSeq 2000 yielding paired-end reads with the following sequencing parameters: Read 1= 28 cycles, Index 1=10 cycles, Index 2= 10 cycles, Read 2= 90 cycles. The targeted sequencing depth was 20,000 reads per nucleus. BCLs were converted to FASTQ files and demultiplexed using cellranger-3.0.1.

CITE-seq protocol Chan Zuckerberg Biohub

1 *Works for me*

Maurizio Morri
Chan Zuckerberg Biohub

PROTOCOL INFO

Maurizio Morri: CITE-seq protocol Chan Zuckerberg Biohub. **protocols.io**
https://protocols.io/view/cite-seq-protocol-chan-zuckerberg-biohub-bxpxpmpn

CREATED

Aug 25, 2021

LAST MODIFIED

Aug 27, 2021

PROTOCOL INTEGER ID

52695

Cell Thawing

1 Cells were removed from liquid nitrogen storage.

2 Frozen cells should be kept on dry ice until ready for thawing.

3 Recommended media for diluting and washing cells are RPMI 1640, IMDM, or DMEM supplemented with 10% heat inactivated FBS warmed to room temperature (20-25°C).

4 Proper aseptic technique should be used when handling and manipulating cells.

5 Check that the vial is tightly capped and place in a 37°C water bath.

6 After 90 seconds (1.2mL vial fill) or 2 minutes (1.8mL vial fill) at 37°C, remove the vial from the water bath every 15-20 seconds and gently invert 3-4 times to check the level of thawing.

7 When the vial contents are 50-75% liquid, remove from the water bath and continue to gently invert the vial until the entire contents are liquid.

8 Place the vial on wet ice and transport it to a biosafety cabinet.

9 Wipe the outside of the vial with an alcohol wipe and gently invert 5 times to mix the cells

10 Check to see if there is liquid caught up in the cap. If yes, gently tap the upright vial on a benchtop to force the liquid into the vial.

11 Immediately remove 20µL of the contents for cell concentration and viability determination. Keep the sample on ice until assayed.

12 Transfer the remaining vial contents to a 15mL tube at RT (20-25°C).

13 Add 1mL of media to rinse the vial and transfer any remaining cells to the 15mL tube.

14 Slowly add 10mL media (about 1mL per second) to the cells. Securely cap the tube and gently invert 3-5 times after adding all the media.

15 Centrifuge the tube at 200g x 8 minutes at RT (20-25°C).

16 Aspirate the supernatant without disturbing the pellet and resuspend the cells in 12mL media.

17 Centrifuge the tube at 200g x 8 minutes at the RT and resuspend in 1ml of cold PBS+0.04% BSA.

18 Transfer the resuspended cells to a 2-ml microcentrifuge tube. Rinse the 15ml tube with 0.5ml PBS+0.04% BSA and transfer the rinse to the 2ml tube containing the cells. Mix by gently inverting the tube.

19 Centrifuge cells at 300rcf for 5 min.

20 Remove the supernatant using a 1ml pipette without disturbing the cell pellet by leaving behind a little supernatant and resuspend the pellet gently in 1 ml PBS + 0.04% BSA.

21 Pass cell suspension through 40uM Flowmi Cell Strainer.

22 Determine the cell concentration using an Automated cell counter.

23 Proceed to cell sorting

Cell sorting Pre CITE-seq

24 Granulocytes were removed using a size-based approach on a SONY-SH800 sorter.

25 Collect the sorted cells in a 5-ml FACS tube containing 500 µl PBS + 0.04% BSA.

26 Centrifuge the collected cells at 300 rcf for 7 min (instead of 5 mins) at 4°C.

27 Remove the supernatant without disrupting the cell pellet and resuspend in 500 µl PBS + 0.04% BSA.

28 Determine the cell concentration using an Automated Cell Counter. Place the cells on ice until ready to proceed.

Biolegend Antibody Lyophilized Panel Reconstitution

29 Reconstitute the TotalSeq-B lyophilized panel from Biolegend (cat# 399904) by following steps 1-9 as per the protocol attached

20-0014-00_TotalSeq-UC_8x11.pdf

Cell staining

30 Follow steps 8-14 on page 3 of the following protocol for cell staining.

Protocol - TotalSeq™-B or -C with 10x Feature Barcoding Technology.pdf

CITE-seq 10x

31 The following 10X protocol was used for the preparation of CITE-seq libraries.

CG000317_ChromiumNextGEMSingleCell3_v3.1_CellSurfaceProtein_Dual_Index_RevA.pdf

Sequencing

32 Pooled libraries were run on an Agilent 4150 TapeStation for visual inspection, and quantified by qPCR, performed on a BioRad CFX96 RT PCR thermal cycler using the KAPA library quantification kit (#KK4923) for quality control assurance prior to sequencing.

33 CITE-seq 3' gene expression libraries and cell surface protein libraries were pooled at 4:1 ratio and sequenced on an Illumina NextSeq 2000 yielding paired-end reads with the following sequencing parameters: Read 1= 28 cycles, Index 1=10 cycles, Index 2= 10 cycles, Read 2= 90 cycles. The targeted sequencing depth for 3' gene expression libraries was 20,000 reads per cell and 5,000 reads per cell for cell surface protein libraries. BCLs were converted to FASTQ files and demultiplexed using cellranger-3.0.1.