MOTIVATION FOR DATASHEET CREATION

[question1] Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

[answer1] 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 micro-scopes. 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 technolo-gies that can measure two or more of these layers, such as joint profiling of DNA and RNA, are a major step to-wards 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.

[question2] Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

[answer2] This dataset is being used in the NeurIPS 2021 Multi-modal Single-cell Data Integration Competition. The com-petition rules and results are accessible at https:// openproblems.bio/neurips_2021.

[question3] What (other) tasks could the dataset be used for?

[answer3] 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.

[question4] Who funded the creation dataset?

[answer4] Major funding for this dataset was provided by the Chan Zuckerberg Initiative under grants DAF2021-235155 and DAF2021-235076. Support for incidental reagents and tech-nician time were provided by Cellarity, Chan Zuckerberg Biohub, and Helmholtz Munich, and the Yale University Center for Genome Analysis.

[question5] Any other comment?

[answer5] No.

DATASHEET COMPOSITION

[question6] What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

[answer6] his dataset comprises single-cell profiles of human bone marrow mononuclear cells from human donors. Samples from each donor were measured using two multimodal tech-nologies. 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.

[question7] How many instances are there in total (of each type, if appropriate)?

[answer7] 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 sam-ples, yielding an expected total of 120,000 cells. This number will be finalized after processing of each sample is finished.

[question8] What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the in-stances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

[answer8] We measured the accessibility of 119,254 genomic re-gions, 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. 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.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] 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.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

[answer10] Only complete instances were included in this dataset.

[question11] 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.

[answer11] 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.

[question12] Does the dataset contain all possible instances or is it a sample (not necessarily random) 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 (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

[answer12] These samples were generated from bone marrow mononuclear cells of human donors. We filtered out gran-ulocytes 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.

[question13] Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer13] 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.

[question14] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer14] 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 distribu-tion [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 pro-files is newer, but some methods for denoising have been described [9], [10], [11]. There is also some literature on denoising ADT data [12].

[question15] Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer15] The dataset is self-contained.

[question16] Any other comments?

[answer16] No

COLLECTION PROCESS

[question17] What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

[answer17] Protocols for sample preparation and collection were based on validated protocols associated with commercially avail-able 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.

[question18] How was the data associated with each instance ac-quired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer18] 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.

[question19] If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

[answer19]

[question20] Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

[answer20] 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.

[question21] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer21] The data was generated in July and August 2021.

DATA PREPROCESSING

[question22] Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

[answer22] 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.

Open chromatin data

The chromatin accessibility data acquired by ATAC-seq as part of the 10X Multiome protocol was processed us-ing 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.

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 pro-tein 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 note-books to reproduce the analysis will be made available on GitHub at https://github.com/openproblems-bio/neurips2021-notebooks.

[question23] 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.

[answer23] Raw sequencing data will be uploaded to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository after publication.

[question24] Is the software used to preprocess/clean/label the in-stances available? If so, please provide a link or other access point.

[answer24] 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.

[question25] 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?

[answer25] 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.

[question26] Any other comments

[answer26] No

DATASET DISTRIBUTION

[question27] How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

[answer27] 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 com-petition, the datasets will be made available at the CZI cel-lxgene portal at https://cellxgene.cziscience.com/.

[question28] When will the dataset be released/first distributed? What license (if any) is it distributed under?

[answer28] The dataset will be released in September 2021 under a CC-BY License.

[question29] Are there any copyrights on the data?

[answer29] No.

[question30] Are there any fees or access/export restrictions?

[answer30] No.

[question31] Any other comments?

[answer31] No.

DATASET MAINTENANCE

[question32] Who is supporting/hosting/maintaining the dataset?

[answer32] 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.

[question33] Will the dataset be updated? If so, how often and by whom?

[answer33] We do not anticipate the dataset will be updated.

[question34] How will updates be communicated? (e.g., mailing list, GitHub)

[answer34] In the event that we need to communicate updates, we will log them in the CZI cellxgene portal version notes.

[question35] If the dataset becomes obsolete how will this be communicated?

[answer35] The dataset will be removed from the CZI cellxgene portal.

[question36] Is there a repository to link to any/all papers/systems that use this dataset?

[answer36] No.

[question37] 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?

[answer37] We designed this dataset to facilitate extension, augmen-tation, and validation. All protocols were performed using commercially available reagents. The source for cells, sam-ple 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.

LEGAL AND ETHICAL CONSIDERATIONS

[question38] 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.

[answer38] 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.

[question39] Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

[answer39] No.

[question40] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

[answer40] No.

[question41] Does the dataset relate to people? If not, you may skip the remaining questions in this section.

[answer41] Yes.

[question42] 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.

[answer42] See Table I

[question43] 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.

[answer43] It has been shown that it is possible to identify indi-viduals 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.

[question44] Does the dataset contain data that might be considered sensitive in any way (e.g., 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.

[answer44] This dataset contains demographic information as de-scribed in Table I. The data also includes genomic information as discussed above and below

[question45] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

[answer45] Data was obtained using tissue samples sources from a third party, AllCells.

[question46] 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.

[answer46] Donors were notified about data collection when they vol-unteered to donate samples. The notification was performed by AllCells using the IC form appended to this datasheet.

[question47] 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.

[answer47] All donors freely signed the IC form appended to this datasheet. Signed consent forms are maintained by AllCells.

[question48] 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).

[answer48] 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 commence-ment 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.

[question49] 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.

[answer49] 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.

[question50] Any other comments?

[answer50] No