Association of
Biomolecular Resource Facilities

Research • Technology • Communication • Education

RG Study Proposal

Research Group: GBIRG

Study Proposal

Submitted by: Charlie Whittaker and Shawn Polson

Reviewed and Approved: _____

Please Note: The overarching goal of RG studies is to create a publishable body of work.

The study could focus on establishing SOPs; creating standards; comparing protocols, instruments, or workflows; developing software; etc.

Study Title: Bioinformatics Assessment of Fixed/Frozen Single-Cell RNA-Seq Data across Three Platforms [DSRG/GBIRG Collaboration]

Specific Aim(s):

A. Prepare count matrices from 10x, Honeycomb and Parse sequence data using vendor-supported software. Investigate the possibility of using open-source software (salmon/alevin) to prepare similar count matrices from FASTQ files. Attempt to establish equivalence between the methods through comparison of QC metrics, global expression profiles and comparisons with publicly available bulk RNA-Seq datasets. QC metrics. downsample fastq files and reprocess the data using either vendor or open-source software. Subset cell sets to assess performance of various methods with decreased coverage and population sizes.
B. Configure and deploy a shared R instance that can be accessed and utilized by members of both the DSRG and GBIRG so that all code development and analytical activities can occur in a consistent shared environment.
C. Perform standard quality control analyses and make comparisons of QC metrics such as number of features detected, count of UMIs/cell and percent mitochondrial reads. This

processing will include ambient RNA adjustment and doublet detection and doublet rates will be compared between platforms.

D. Execute data integration and batch correction to produce a single seurat object containing all cells from each site and platform.
E. Perform unsupervised clustering analyses to identify the total number of robust clusters and enumerate the site and platform-specific contribution to each cluster.
F. Perform cell type classification of the resulting clusters to test the hypothesis that some cell types (granulocytes) are captured more effectively using different technologies.
G. Perform differential gene expression analysis to identify cell-type markers and make cross-platform comparisons of these results at both the gene and gene set level.

Background:

Single cell RNA-Seq is a critical and common service provided by genomics and bioinformatics cores across the country. The most common technology in use is the whole-transcriptome 10x Genomics platform that relies on freshly isolated cell suspensions. Fresh cells can pose experimental challenges because gene expression may not be static following tissue dissociation. The ability to fix or freeze cells immediately after isolation may facilitate more accurate capture of the transcriptional state of cells under investigation. Fixation and freezing may also alleviate logistical challenges caused by the lack of proximity between collection of biological material and the equipment necessary for preparation of single cell libraries. These issues have led companies to develop methods that work with fixed or frozen cells. 10x Genomics now provides a fixed-cell and probe-based protocol that competes with the Parse Biosciences method and a frozen cell approach available from Honeycomb Biotechnologies. The probe-based strategy employed by the 10x fixed cell approach restricts sequencing to genes that have been targeted by gene-specific primers thereby adding an additional level of variability to the experiment. The DSRG has designed the study outlined in appendix 1 to compare fresh human leukocytes processed with 10x with those prepared using the fixed 10x protocol, fixed Parse, and frozen Honeycomb samples collected at Dartmouth and processed in four different cores. The GBIRG will provide bioinformatics support for the analysis and interpretation of the data produced by the DSRG

Study Plan:

Preprocessing pipelines from each vendor will be used to prepare count matrices. A single replicate of fresh leukocytes will be processed with 10x and sequenced at the Dartmouth facility. This facility will also prepare four batches of two replicates of fixed cells for 10x and four batches of two replicates of fixed cells for Parse and four batches of two replicates of frozen cells for Honeycomb. Fixed and frozen samples will be stored for approximately 4 weeks and, during this time, will be distributed to four different facilities for library prep and sequencing. Together, this design will allow consideration of fresh samples, two different fixation protocols, one frozen cell protocol and four sequencing sites.

Following sequencing, the fresh and fixed 10x samples will be processed using the cellranger mkfastq and count packages. Parse and Honeycomb data will be processed using vendor-supplied software. The ParseBiosciences pipeline will be run according to the manufacturer's suggestions. The Honeycomb HIVE pipeline Beenet will be used to create bam files and count matrices.

We will also investigate the possibility of preparing count matrices using the open-source software Salmon/Alevin (Srivastava et al. 2019). This is possible with 10x data, likely possible with Honeycomb data, but Parse data remains to be determined. This work may allow for sequence processing of primary fastq data from the different platforms and sites using a consistent software package possibly resulting in reduced variability. It may also be faster and could allow for custom alignment targets to be used. In addition, Aim A of our proposal includes investigating the effect of downsampling of raw data so that we can better understand the stability of biological conclusions in the context of reduced cell number and lower sequencing depth. These data will require multiple runs of reprocessing from raw sequence. Our proposed analyses would benefit from a consistent and shared pipeline if one can be established..

The various count matrices will then be used as input to Seurat (Hao et al. 2021; Butler et al. 2018; Satija et al. 2015). We will strive to maintain consistency between platforms in downstream analysis, however we will review manufacturer's best practices for downstream analysis and incorporate platform specific recommendations as appropriate (e.g. HIVE provides a tutorial for secondary analysis in Seurat for performing QC, basic clustering, and cell type annotation for a single sample). We will prepare and configure an R environment running on an Rstudio Server instance at the Koch Institute. All members of both the DSRG and GBIRG will have access to this resource thus facilitating a consistent computational environment for all participants. These data will then be processed with a shared R markdown file prepared in this collaborative environment and the project will leverage a GitHub repository to enable sharing of both data and analysis code .

Seurat objects will be prepared from the count matrices for each run. For data prepared with 10x genomics, ambient RNA adjustment will be done using soupx (Young and Behjati 2020). We will investigate the usage of this software with data prepared using Parse and Honeycomb. Doublets will be quantified using DoubleFinder (McGinnis, Murrow, and Gartner 2019). Quality control metrics such as number of genes detected per cell, UMI counts per cell and mitochondrial RNA percentage will be gathered and comparison of these data between platforms and sites will be performed. Appropriate QC filtering will be done, Seurat objects will be merged and batch effects will be assessed through cluster analysis and data visualization.

Given the diverse platforms and processing routines, it is likely that substantial batch effects will exist in the data so data integration in Seurat will be performed followed by cluster analysis. Cell type identity will then be assigned to clusters using singleR (Dvir Aran, Aaron Lun, Daniel Bunis, Jared Andrews, Friederike Dündar n.d.) using various available human reference data. Comparison of cell counts within each cell type/cluster from each platform and site will allow for

identification of cells/clusters that are efficiently detected by all platforms and those that may be preferentially captured by a specific platform. Examples of these may include neutrophils and granulocytes which are reportedly difficult to capture with fresh 10x processing. It is possible that clusters could result from location or technical variables. Human leukocytes are a well-studied biological material, FACS experiments have documented different populations in the material and bulk public RNA-Seq data are available. Our datasets will first be subjected to cell type classification to identify the leukocyte populations present, and used to generate pseudobulk estimates that represent the average expression profile cell type. We will then determine the global gene expression correlation between single cell and publicly available bulk RNA-seq profiles in a cell type specific manner. This analysis will allow performance of each pipeline to be compared, as the best performing pipelines should share high correlation with bulk RNA-seq measurements. We hope to combine this analysis with the 10x fresh cell data and FACS-based information to validate biologically meaningful clusters and characterize technical results in the data from different platforms.

Following the cluster and cell type analyses, the FindAllMarkers function of Seurat will be used to identify gene markers specifically expressed in the various clusters or cell-types derived from different platforms and/or cell-preparations. The magnitude and consistency of these markers will be compared to assess the performance of the platforms. These analyses will be highly dependent on platform-dependent cluster representation but certain abundant cell types are likely to be amenable to analysis. We will quantify the expression of gene sets known to be relevant in different populations using VAM (Frost 2020) and the magnitude of gene set expression will be compared between platforms and sites.

The cells sequenced in this study are likely to be terminally differentiated and not undergoing cell state changes. This may complicate trajectory analyses but these types of analyses will be considered depending on the results. If possible, we will compare any developmental trajectories that are detected in data produced by different platforms.

References:

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. 2018. "Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species." *Nature Biotechnology* 36 (5): 411–20. https://doi.org/10.1038/nbt.4096.

Dvir Aran, Aaron Lun, Daniel Bunis, Jared Andrews, Friederike Dündar. n.d. *SingleR*. Bioconductor. Accessed July 6, 2022. https://doi.org/10.18129/B9.BIOC.SINGLER.

Frost, Hildreth Robert. 2020. "Variance-Adjusted Mahalanobis (VAM): A Fast and Accurate Method for Cell-Specific Gene Set Scoring." *Nucleic Acids Research* 48 (16): e94–e94. https://doi.org/10.1093/nar/gkaa582.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184 (13): 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

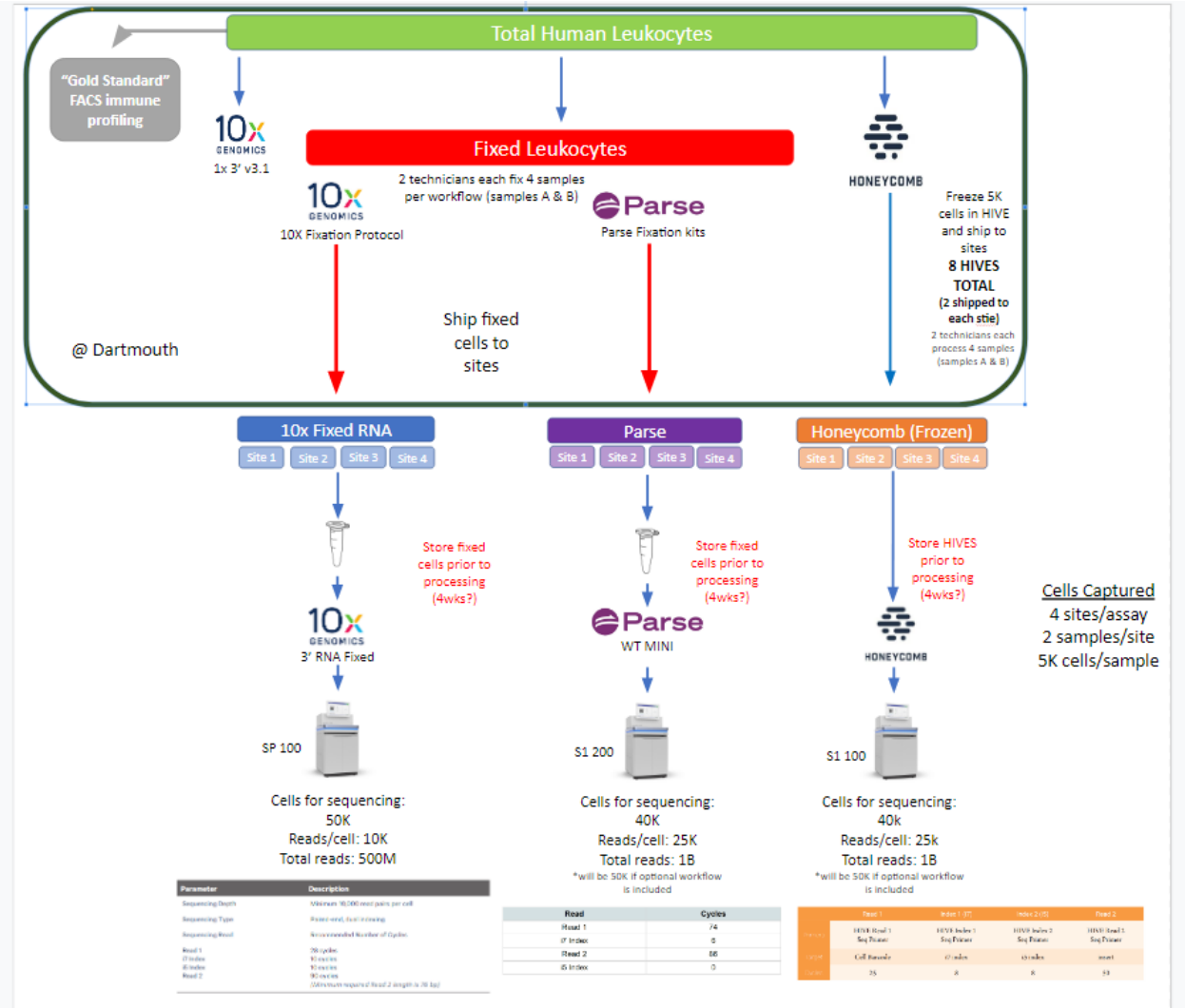
McGinnis, Christopher S., Lyndsay M. Murrow, and Zev J. Gartner. 2019. "DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors." *Cell Systems* 8 (4): 329-337.e4. https://doi.org/10.1016/j.cels.2019.03.003.

Satija, Rahul, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. 2015. "Spatial Reconstruction of Single-Cell Gene Expression Data." *Nature Biotechnology* 33 (5): 495–502. https://doi.org/10.1038/nbt.3192.

Srivastava, Avi, Laraib Malik, Tom Smith, Ian Sudbery, and Rob Patro. 2019. "Alevin Efficiently Estimates Accurate Gene Abundances from DscRNA-Seq Data." *Genome Biology* 20 (1): 65. https://doi.org/10.1186/s13059-019-1670-y.

Young, Matthew D, and Sam Behjati. 2020. "SoupX Removes Ambient RNA Contamination from Droplet-Based Single-Cell RNA Sequencing Data." *GigaScience* 9 (12): giaa151. https://doi.org/10.1093/gigascience/giaa151.

Appendix 1 - Study Design



Participants:

GBIRG Members: Alper Kucukural, Chris Gates, Shannan Ho Sui, Krishna Karuturi, George Bell, Alex Lemenze, Julie Dragon, Maddy MacDonald, Jessie Li, Huiming Ding, Duan Ma, Owen Wilkins, Jyothi Thimmapuram, Nadia Lanman, Charlie Whittaker, Shawn Polson

Work will be done in collaboration with the DSRG who designed the study and will be generating the data.

Study Budget / Cost:

No ABRF funds are required for this study.