SUPPLEMENTAL MATERIAL

Leveraging single cell RNA sequencing experiments to model intratumor heterogeneity

Meghan C. Ferrall-Fairbanks (1), Markus Ball (2,3),
Eric Padron (2), and Philipp M. Altrock* (1)

1: Department of Integrated Mathematical Oncology, Moffitt Cancer Center and Research Institute

2: Department of Malignant Hematology, Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

3: present address: Institute of Pathologiy, University Hospital of Cologne, Kerpener Str.62, 50937 Cologne, Germany

*Corresponding author: Philipp M. Altrock, philipp.altrock@moffitt.org
Moffitt Cancer Center, 12902 USF Magnolia Drive, SRB 24007, Tampa, FL 33612, USA
Phone: +1 813 745-5897 Fax: +1 813 745- 8357

SUPPLEMENTAL MATERIALS AND METHODS

DROPLET-BASED scRNA$_{SEQ}$ SAMPLES

The 10X Genomics platform offered a variety of data sets[1] that were used to create our pipeline for quantifying intraleukemic heterogeneity. The major focus of our analysis and pipeline development was of Healthy/Control and AML patient bone marrow mononulear cells (BMMCs). The Healthy Controls 1 and 2 BMMCs had been sequenced on Illumina Hiseq 2500 Rapid Run V2 with 90-135 thousand reads per cell and 2000-2400 cells detected. The AML027 and AML035 pre-transplant BMMCs had been sequenced on Illumnia NextSeq 500 High Output with 16.6-58 thousand reads per cell and 3500-3900 cells detected. The AML027 and AML035 post-transplant BMMCs had also been sequenced on Illumina NextSeq 500 High Output with 41-51 thousand reads per cell and 900-3900 cells detected. Additional data sets[1] for CD34+, CD14+, CD19+, and CD4+ (hematopoietic cell type set) cells as well as patient normal-tumor matched lung cancer samples[2] available in ArrayExpress under accessions E-MTAB-6653 and E-MTAB-6149 were used to test the robustness of the pipeline. The CD34+ data set was CD34+ cells enriched from peripheral blood mononuclear cells (PBMCs), sequenced on an Illumnia NextSeq 500 High Output with 24.7 thousand reads per cell with 9000 cells detected. The CD14+ data set was enriched from PBMCs, sequenced on Illumnia NextSeq 500 Output with 100 thousand reads per cell with 2600 cells detected. The CD19+ data set was enriched from PBMCs, sequenced on Illumnia NextSeq 500 Output with 25 thousand reads per cell with 10000 cells detected. The CD4+ data set was enriched from PBMCs, sequenced on Illumnia NextSeq 500 High Output with 21 thousand reads per cell with 11000 cells detected. All lung tissue samples were prepared as single-cell suspensions, sequenced on Illumina HiSeq4000, and aimed for an estimated 4000 cells per library[2]. There were six patients analyzed, each patient had 4 distinct

tissues locations sampled: an adjuacent normal lung sample (normal), a sample from the tumor core (core), a sample from the tumor margin (edge), and a sample between the tumor core and margin/edge (middle).

INITIAL DATA PROCESSING

First, the all samples were run through the 10x Genomics Cell Ranger count pipeline for transcriptome alignment, calling individual cell barcodes, estimating multiplet rates, and finally clustering of normalization corrected FPKM (fragments per kilobase million; expression level) values (**Fig. S1**, pipeline X). In our analysis of the Healthy versus AML high-throughput scRNA-seq, we used Cell Ranger aggr to pool together the individual data set results, so we could compare intratumor heterogeneity (ITH) between samples with the same clustering. Using a graph-based clustering algorithm, 23 different clusters were identified based on clustering cells by expression similarity (shown in **Fig. 2B-C**). The clustering was performed by forming a graph with cells as vertices and edges indicating pairs of cells that are sufficiently similar (implemented by Cell Ranger). The Cell Ranger graph-based clustering with Louvain Modularity Optimization[3] was used to partition this graph into clusters of similar cells. For analysis with the hematopoietic cell set and the lung cancer samples the cell/gene matrices produced by Cell Ranger count pipeline were then loaded into R using the Seurat package[4] (designed for QC, analysis, and exploration of single cell RNA-seq data; **Fig. S1**, pipelines M and S; pipelines implemented using version 2.3). With Seurat, cells were removed that had either fewer than 200 UMIs and greater than 6000 UMIs and more than 10% mitochondrial DNA[2], then scaled and normalized before clustering using graph-based clustering.

QUANTIFYING SUMMARY DIVERSITY METRICS

After the data was clustered, we then sought to quantify whether the clustering could disentangle healthy and AML samples by average gene expression. To approach within-sample differences in overall gene expression, we computed Euclidean distance matrices of the mean expression values in each cluster, to establish a distance metric of cluster differences and a subsequent diversity of distances across samples (standard deviation, ANOVA) (**Fig. 2A**). For each cluster, the geometric mean of each unique molecular identifier (UMI) across all genes were computed, then the Euclidean distance was computed between clusters. This was plotted as a graph where each node represents each clusters identified in the leukemic data set (**Fig. 2D**). The size of the node indicates the total number of cells in that cluster and the color identifies the major species (AML, Healthy, or postBMT) present in that cluster. The distribution of cells per cluster (**Fig. 2E**) was also included to show which other conditions were also present in each cluster and how the number of cells compared to the major species used to dictate the color of the nodes in graph describing the connectivity of the clusters in **Fig. 2D**.

Next, we sought to characterize across-sample differences by calculating the Kolmogorov-Smirnov (KS) distance[5] of the cell count distributions in each cluster (**Fig. 3A**). The KS distance is a non-parametric measure that is calculated as the supremum of paired differences of two empirical probability mass functions[5]. Last, we calculated a continuum of the ecological diversity index[6] based on the individual cell frequencies in each of the clusters identified by the graph-based clustering (**Fig. 3H**), across all clusters, which can be written as:

$$ {}^{q}D = \left(\sum_{I=1}^{n} p_i^q \right)^{\frac{1}{1-q}} $$

where n is the number of clusters in the data set, p_n is the frequency of each cluster, and q is the order of diversity. q is a hyperparameter that would be optimized in a clinical setting. The two distributions for AML were joined together and the two distributions of the healthy indiviuals were joined together and plotted across orders of diversity (**Fig. 4A-C**). The same technique was used to group the lung cancer samples based on location as reported in (**Fig. 5B**).

DOWNSAMPLING

We sought to test the robustness of our metrics by downsampling the number of cells and re-calculating the number of clusters identified as well as changes in key diversity scores. This downsampling was preformed removing increments of 10% of the data at a time and the results were reported as the summary statistics of 1000 runs at each downsampling percentile (**Fig. S2**). The downsampling results were also used to quantify confidence in the diversity index spectrum and was calculated for the following orders of diversity: 10^{-2}, 10^{-1}, 10^{0}, 2, 10^{1}, 10^{2}. The relative change in diversity score was reported for the downsampling results with 50% the initial amount of cells in **Fig. 4D-I**. Relative change in the diversity score was calculated by subtracting the mean of the diversity score from all 1000 and reported as the distributions of scores around that mean score.

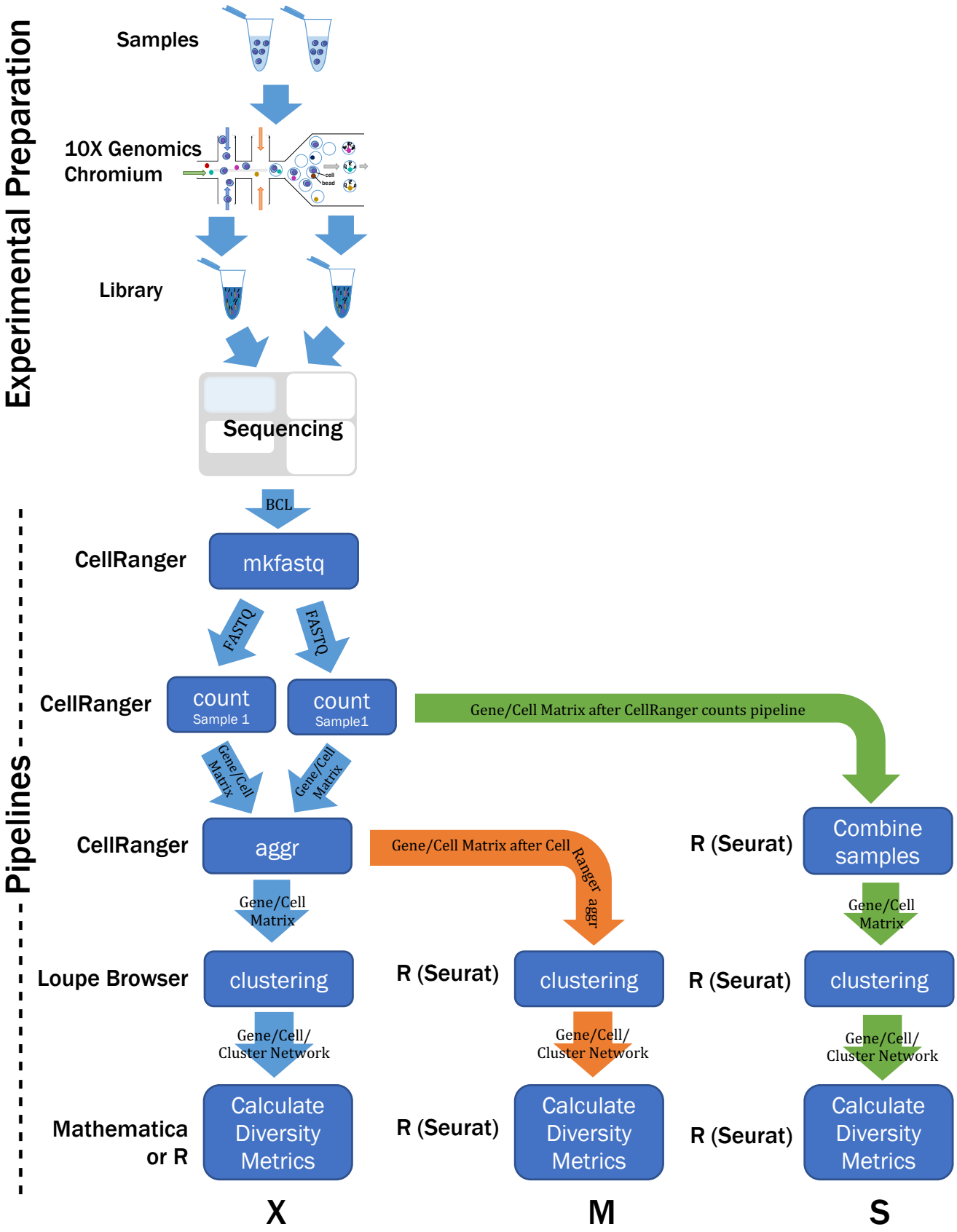
CODE AVAILABILITY

The code used in the pipelines described in **Fig. S1** were uploaded to a GitHub repository (https://github.com/MathOnco/scRNAseqITH/releases/tag/1.0). These pipelines were implemented using data run through at least Cell Ranger counts, and then post-processed with Cell Ranger Loupe Browser, Mathematica, and R. Code available includes:

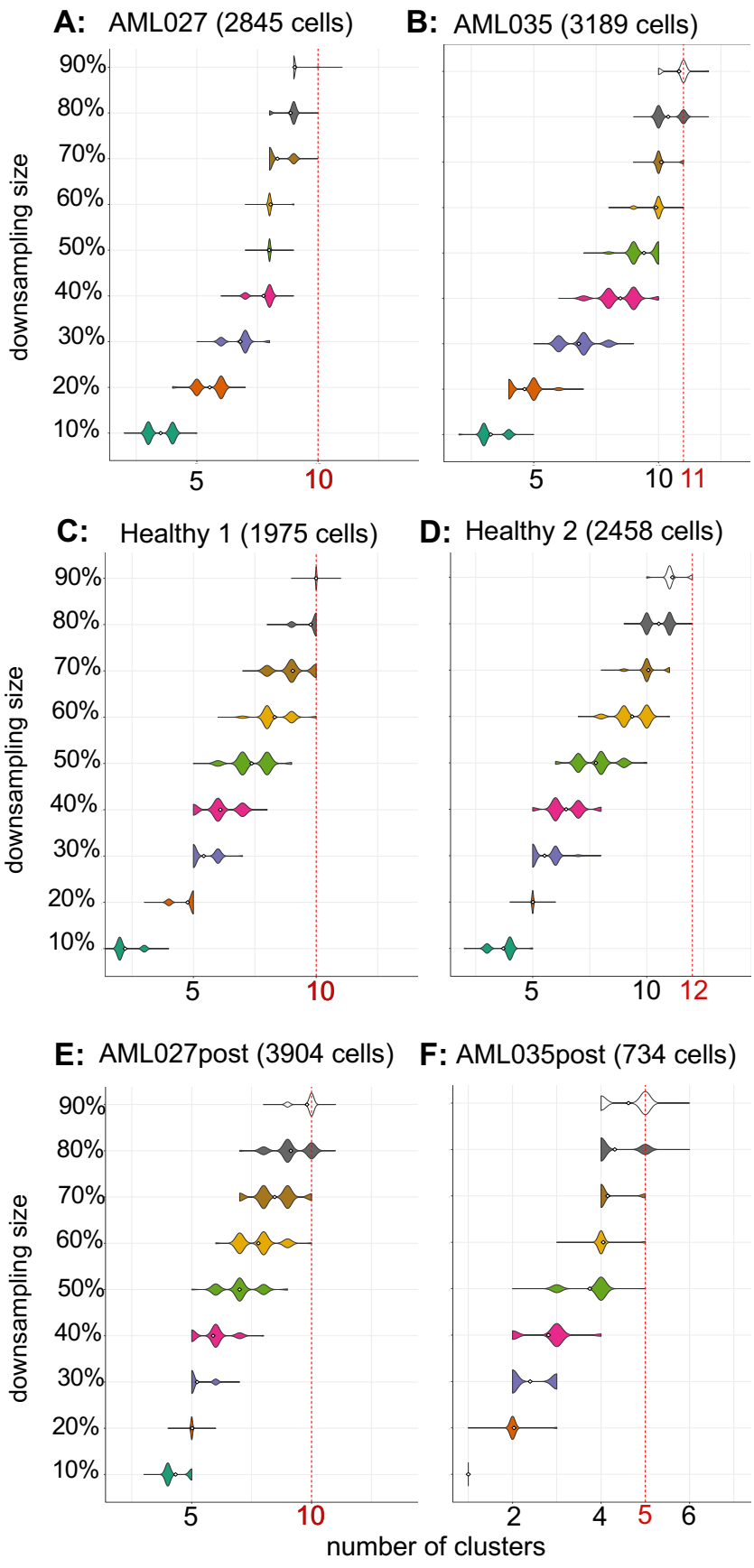
- Mathematica code and R scripts and data described to implement all pipelines (described in **Fig. S1**)

- Mathematica code for calculating the geometric mean of UMIs and Euclidean distances between clusters (for **Fig. 2**)
- R code to draw the graph of the clusters (**Fig. 2D**)
- Mathematica code for calculating the KS distance and diversity spectrum across all clusters for the leukemic data set (**Fig. 3**)
- R code used to downsample data sets, then cluster and calculate diversity indices (**Fig. 4** and **S2**)
- R code for calculating the diversity spectrum for the hematopioteic subtype and matched lung carcinoma data sets (**Fig. 5**)

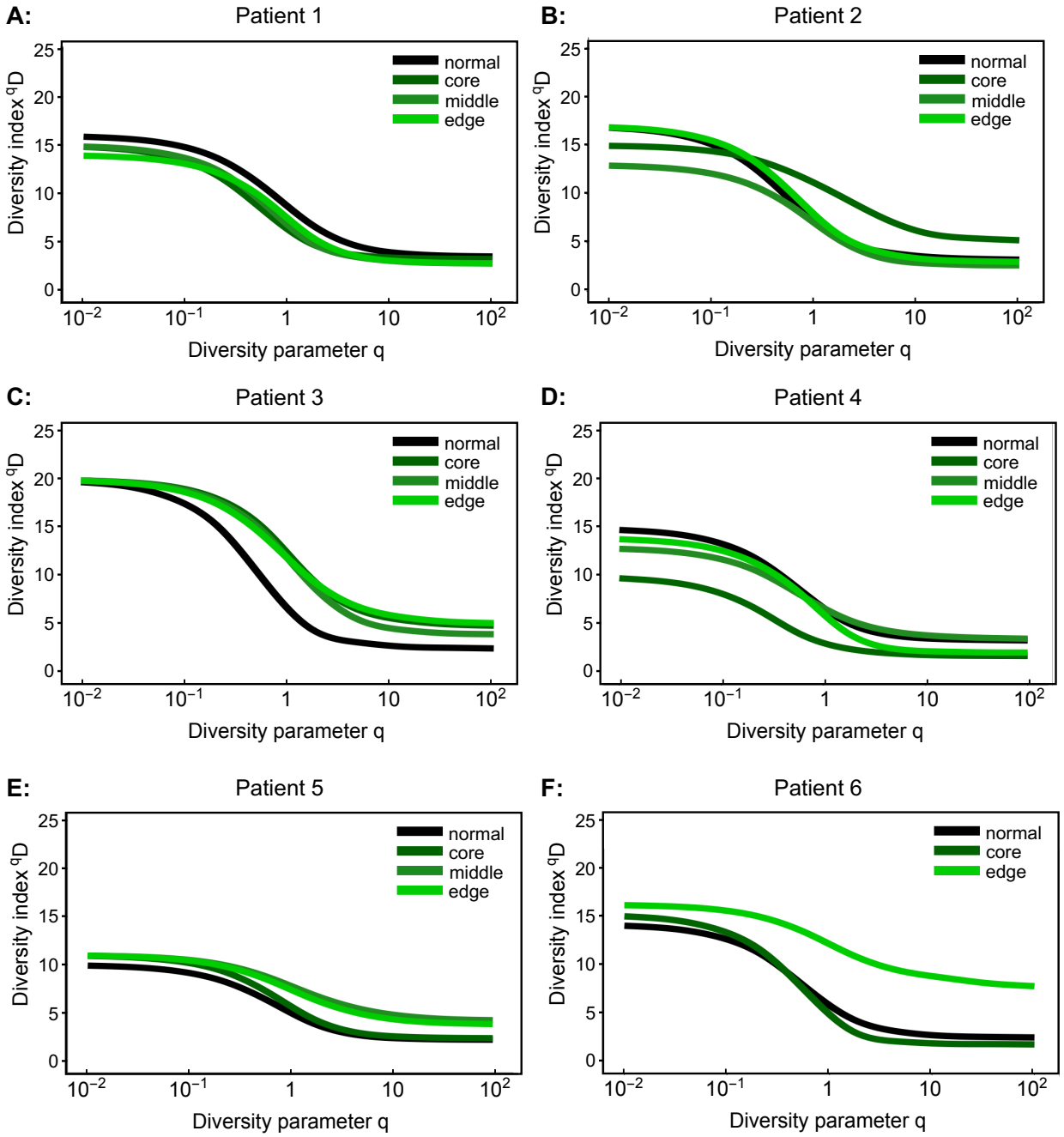
Supplemental Figure 1. Workflow diagram for the scRNAseq quantifying intraleukemia heterogeneity for an example comparing two samples. Cells were processed through the 10X Genomics Chromium and ultimately sequenced, producing a raw base call (BCL) file, which was demultiplexed for each flowcell directory and converted to a FASTQ file using Cell Ranger mkfastq. Then Cell Ranger count was run separately for each library to align, filter, and count barcodes and UMIs. These instances were then aggregated into a single instance using Cell Ranger aggr to normalize runs to the same sequencing depth and recompute analysis on combined data set. Following along pipeline X, from the aggregated analysis, the Loupe Browser contains the clustered data, which were exported and run through a Mathematic script to calculate the diversity metrics. This pipeline was used to generate Figs. 3 and 4. This pipeline was expanded to give the user increased flexibility by using the gene-barcode matrices output by Cell Ranger and then processing the data in R using the Seurat package (version 2.3), which allows users to set the filtering criteria (M) and allows data to be combined without running through the Cell Ranger aggr pipeline (S) as well as both new pipelines allow users to implement additional clustering algorithms. Pipeline S was used to generate Fig. 5 from gene-barcode matrices available from additional data from Zheng, et. al.[18].


Experimental Preparation
Samples
10X Genomics Chromium
Library
Sequencing
Pipelines
BCL
CellRanger
mkfastq
FASTQ
FASTQ
CellRanger
count Sample 1
count Sample1
Gene/Cell Matrix after CellRanger counts pipeline
Gene/Cell Matrix
Gene/Cell Matrix
CellRanger
aggr
Gene/Cell Matrix after Cell Ranger aggr
R (Seurat)
Combine samples
Gene/Cell Matrix
Loupe Browser
clustering
R (Seurat)
clustering
R (Seurat)
clustering
Gene/Cell/ Cluster Network
Gene/Cell/ Cluster Network
Gene/Cell/ Cluster Network
Mathematica or R
Calculate Diversity Metrics
R (Seurat)
Calculate Diversity Metrics
R (Seurat)
Calculate Diversity Metrics
X
M
S

Supplemental Figure 2. AML populations converge to consistent number of clusters that describe the population sooner than healthy populations. graph-based clustering. Each original population's gene-barcode matrix obtain from publicly available data from Zheng, et. al.[18] was then downsampled to contain 10% to 90% of the cells originally in the matrix. This data was loaded into R using the Seurat package, downsampled, and then clustered to determine how sensitive the cluster metrics was to the starting number of cells. Downsampling was performed 1000 times per cutoff and the violin plots show the distribution of clusters identified. The red-dashed lined indicates the number of clusters identified with no cells removed from the data set. Gray numbers indicate the number of cells present in each downsampling condition. AML populations (**A**, **B**) showed quickest convergence to number of clusters after about 1500 cells were present in clustering. Healthy populations (**C**, **D**) showed that as more cells were added, an additional cluster could be found. AML post-bone marrow transplant populations (**E**, **F**) behaved more similar to healthy than AML populations.


A: AML027 (2845 cells)
B: AML035 (3189 cells)
C: Healthy 1 (1975 cells)
D: Healthy 2 (2458 cells)
E: AML027post (3904 cells)
F: AML035post (734 cells)
downsampling size
90%
80%
70%
60%
50%
40%
30%
20%
10%
5
10
11
12
2
4
6
number of clusters

Supplemental Figure 3: Individual lung cancer patient diversity scores show same trends as aggregated clustering results with tumor samples have greater diversity scores than their matched normal counterpart. Individual patients were run through pipeline S to aggregate normal, core, middle, and edge tissue samples.



REFERENCES

1. Zheng GX, Terry JM, Belgrader P, et al: Massively parallel digital transcriptional profiling of single cells. Nat Commun 8:14049, 2017

2. Lambrechts D, Wauters E, Boeckx B, et al: Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med 24:1277-1289, 2018

3. 10xGenomics: https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/algorithms/overview, 2018

4. Butler A, Hoffman P, Smibert P, et al: Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36:411-420, 2018

5. Carruth J, Tygert M, Ward R: A comparison of the discrete Kolmogorov-Smirnov statistic and the Euclidean distance. arxiv.org:arXiv:1206.6367, 2012

6. Lou J: Entropy and diversity. Oikos 113:363-375, 2006