Comparative single-cell lineage bias in human and murine hematopoietic stem cells

Isaac Shamie,[1*] Meghan Bliss-Moreau,[2,3*] Jamie Casey Lee,[4*] Ronald Mathieu,[2] Harold M. Hoffman,[4] Bob Geng,[5] Nathan E. Lewis,[1,6,7#] Yanfang Peipei Zhu[4,8#] and Ben A. Croker[2-4#]

[1]Department of Bioengineering, UC San Diego, La Jolla, CA; [2]Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA; [3]Department of Pediatrics, Harvard Medical School, Boston, MA; [4]Division of Rheumatology, Allergy & Immunology, Department of Pediatrics, School of Medicine, UC San Diego, La Jolla, CA; [5]Rady Children's Hospital San Diego, San Diego, CA; [6]Division of Host-Microbe Systems and Therapeutics, Department of Pediatrics, UC San Diego, La Jolla, CA; [7]Center for Molecular Medicine, Complex Carbohydrate Research Center, Department of Biochemistry and Molecular Biology, University of Georgia, GA and [8]Immunology Center of Georgia, Medical College of Georgia, Department of Biochemistry and Molecular Biology, Augusta University, GA, USA

IS, MB-M and JCL contributed equally as first authors.

#NEL, YPZ and BAC contributed equally as senior authors.

Correspondence: B.A. Croker
bcroker@health.ucsd.edu

Y.P. Zhu
pzhu@augusta.edu

N.E. Lewis
natelewis@uga.edu

Received: March 25, 2025.
Accepted: July 11, 2025.
Early view: July 17, 2025.

https://doi.org/10.3324/haematol.2025.287897

Supplemental Methods

Human primary CD34+ cells

Cryopreserved $CD34^+$ hematopoietic stem and progenitor cells were obtained from Gilead (Donors 1-4, Donors A-D for FACS sorting) or StemCell Technologies (Donors 5-8). Adult donors gave informed consent for the collection of CD34+ cells. Samples, where applicable, were cultured for 72h in DMEM/10%FBS/10%CO_2 and a cytokine culture consisting of 100 ng/mL recombinant human SCF/IL-3/IL-6/Flt3L/G-CSF/GM-CSF. The $CD34^+$ samples were de-identified and processed in both the mt-scATAC-seq library preparation and FACS sorting.

Processing of mt-scATAC-seq sequencing fragments

Processing of mt-scATAC-seq reads was performed as previously reported[29]. The cellranger-atac count command from cellranger v6.1.1 was used to generate bam, peak genomic regions and peak-fragment count files. The hg38 reference genome was modified by hard-masking nuclear regions that align to the MT genome with single bp errors[29] (regions taken from https://github.com/caleblareau/mitoblacklist/tree/master/combinedBlacklist). Reads were trimmed to remove the adapter and primer sequences, and then aligned using BWA-MEM[30]. Open-chromatin peaks were detected, and cell barcodes were filtered. For peak-calling, reads were aggregated across all cells to boost signal, and a global threshold was applied to select candidate regions above background genomic noise. This was done by fitting negative-binomial distributions to estimate background and peak likelihood in the candidate regions. Local-maxima peaks within this region were then found and a local threshold was applied, generating peaks of various sizes. For cell-calling, potential barcode multiplets were collapsed by masking the minor barcode, and barcodes were removed using a threshold for fraction of fragments in the peak using a mixture model of two negative binomial distributions to capture the signal and noise, with an odds ratio threshold of 100000.

Variant calling in the MT genome

Cells were filtered with less than 200 bp in the MT genome and fragment duplicates removed. Positions were removed with less than ten cells with at least 50x coverage, and with less than 10 cells having 5x coverage of a putative variant at that position. Additionally, cells required an average Phred base quality score (BQ) of over 20 at the putative variant. MGATK filters removed variants with low strand concordance and low variance-mean ratio for each variant across all cells in a sample. The thresholds used were the same as previously reported[29], with concordance of 0.65 and log 10 variance-mean ratio of -2.

Separating multiplexed donor cells

To separate donors from the same sequencing run, the algorithm Vireo[31] was used, which is a variational Bayesian inference algorithm that reconstructs each donor's allele frequency

profile (the donor's mean allele frequency is the latent variable) and assigns a probability of each cell to that donor. Any cell with less than 0.9 probability to be assigned to a clone was removed. The algorithm also assigns a 'doublet' probability for each cell, which is the likelihood of the cell being part of multiple donors versus one. Cells with more than 0.1 probability of a doublet were also removed. To ensure the donors called were correct, the number of donors in Vireo +/- 2 from the true number of donors was examined. The model's reconstruction likelihood score, the evidence lower bound (ELBO), used in variational autoencoders, is saved for each donor parameter, and the 'elbow rule' is then used, which finds the error's inflection point upon increasing the number of donors. Donor specific homozygous variants were calculated as having a mean allele frequency greater than 0.9. In all our cases, the true number of donors is where the elbow occurs.

Clonal detection using MT barcodes

After computationally separating the donors, the single-cell variant allele frequency was calculated for high-coverage positions to reduce spurious clone-calling, and then MGATK was performed providing a new set of called variants for each donor. To detect cells of the same clone, the k-nearest-neighbors Leiden-based community detection algorithm was used[32]. The resolution parameter was set to 30, after assessing values of 30-50, and the cosine distance cutoff of the algorithm was set to 3.5. To measure consistency across workflows, cell pairs were examined to determine if they were either assigned to the same clone in both methods, assigned different clones in both, or assigned the same clone in one method but not the other (negative samples). We compared the fraction of the cell pairs that overlapped with each other (Figure S28). In Figure S2B, cell population was subsampled, and an adjusted normalized mutual information score was calculated between the cell-clone assignment in the sampled clone composition and the full sample detected clones.

To calculate the percentage of cells with the barcode in a clone and outside a clone in Figure 2D, variants were binarized with a minimum of 2 reads and an allele frequency of 0.001. The top 3 variants with the highest positive difference in percentage between clones and non-clones was chosen. For Figure 2E, complete-linkage using cosine similarity was used, setting allele frequency of >0.2 to 0.2 to improve visibility. Barcodes with an average of less than 0.01 in each clone were removed. In Figure 2A, the distribution of each barcode was plotted across cells in each clone using a boxenplot with default parameters in seaborn v0.11.2, which is a modified form of a boxplot that better represents the distribution for large data (https://github.com/heike/stat590f).

Processing single-cell nuclear open-chromatin regions

To examine the peaks detected using the nuclear open-chromatin reads in each cell, the Signac (V1.4) protocol was used to integrate conditions, preprocess, and binarize the cells, run

latent-semantic indexing (LSI), followed by UMAP dimensionality reduction, and KNN Louvain clustering to assign cluster labels[33]. Integration was done by comparing input and cultured cells or by integrating all sequencing runs.

To examine open-chromatin regions and aggregate data across experimental runs, the detected peaks were merged by expanding the peaks with overlap across runs. Peaks < 20 bp and >10,000 bp were removed and fragment counts were re-computed. A Signac model was used to remove regions with < 10 cells, and cells with < 200 features. Additionally, data were filtered by keeping peaks with: a) ≥ 10 and < 15,000 fragments; and b) ≥15% of the nucleotides in reads found in the peak was also covered in the peak (since a read can span the peak region and outside the region). Cells were also retained: a) with a nucleosome signal of ≥4 (i.e. the ratio of mononucleosomal to nucleosome-free fragments per cell); b) with a TSS enrichment of ≥ 0.2 (as defined previously); and c) with a ratio of reads aligned to blacklist regions over reads aligned to peaks < 0.05.

Peaks were binarized and a term frequency–inverse document frequency (TF-IDF) was assessed followed by SVD, which combined is the latent-semantic indexing method. UMAP was then run on dimensions 2-50, as the first factor correlates with depth. After this, runs were integrated using *FindIntegrationAnchors* of the Seurat package using the lsi transformed data[33]. After integration, UMAP on dimensions 2 to 30 of the integrated lsi components was utilized, then clustered using *FindNeighbors* and *FindClusters* with the SLM algorithm[34].

Annotating cell clusters using lineage markers

Cells were annotated by taking known lineage markers of both gene activity and TF activity and overlaying the density of the feature across the UMAP embedding. Gene activity scores for each gene was calculated by summing the number of peaks found in a gene and 2 kb upstream. Feature counts for each cell are divided by the total counts for that cell, multiplied by the median gene activity in that cell, and then natural log transformed to obtain an activity score. TF activity was calculated using the chromVAR extension in Signac, which estimates activity based on the number of TF motifs detected in a cell's open-chromatin peaks[35]. Manual annotation was performed on the clusters using both the gene and TF activity in known markers.

Hypergeometric test to measure lineage bias in clones

To detect clonal bias towards a specific lineage, a hypergeometric cumulative distribution test was used for each clone-cluster pair, and p-values were adjusted using the Benjamini-Hochberg method to control the false discovery rate. A significance threshold of 0.1 was used, but to account for clone and cluster sizes affecting the test, a non-parametric null distribution was created in which the cluster labels for each cell were shuffled 1000 times and the p-values for each clone-cluster pair computed. The p-values in each simulation were used as a background distribution, and empirical p-values were calculated for each clone-cluster pair, a significance of p=0.1 was used in reporting significance values.

Clone and lineage entropy measures

To measure the lineage-bias across clones in Figure S3F, a normalized entropy metric was used. The 'HSPC' lineage clusters were removed, and the frequency of each cell type was assessed in each clone, and then used as the probability distribution. The standard entropy measure was calculated using entropy from the SciPy v1.7.3 stats package[36], and was normalized to a value between 0 and 1 by dividing by the natural log of the number of clones.

Flow-cytometry for human CD34^{+} cell cultures

Human: flow-cytometry was done for four healthy CD34+ donors, and culturing was done as mentioned above. Staining was performed in FACS buffer (D-PBS + 1% human serum + 0.1% sodium azide + 2mM EDTA) on ice. Cells were filtered through sterile 70 µm cell strainers to obtain a single cell suspension. Prior to staining, human Fc receptors blocking reagent (Biolegend) was added for 15 min. Staining was performed for 30 minutes in a final volume of 100ul. Cells acquired using a LSR Fortessa (BD Biosciences). All flow cytometry analysis performed on live cells. The markers used for dimensionality reduction in were HLA-DR, CD117, CD11c, CD11b, CD34, CD10, CD45, CD86, FcεRIα, CD16, CD14, CD66b, CD101, Siglec8, CD3, CD19, CD56.

Code availability

All code used for data processing and analysis for this study has been deposited here, where it will be made publicly available upon acceptance of this work:

https://github.com/LewisLabUCSD/Mito_Trace

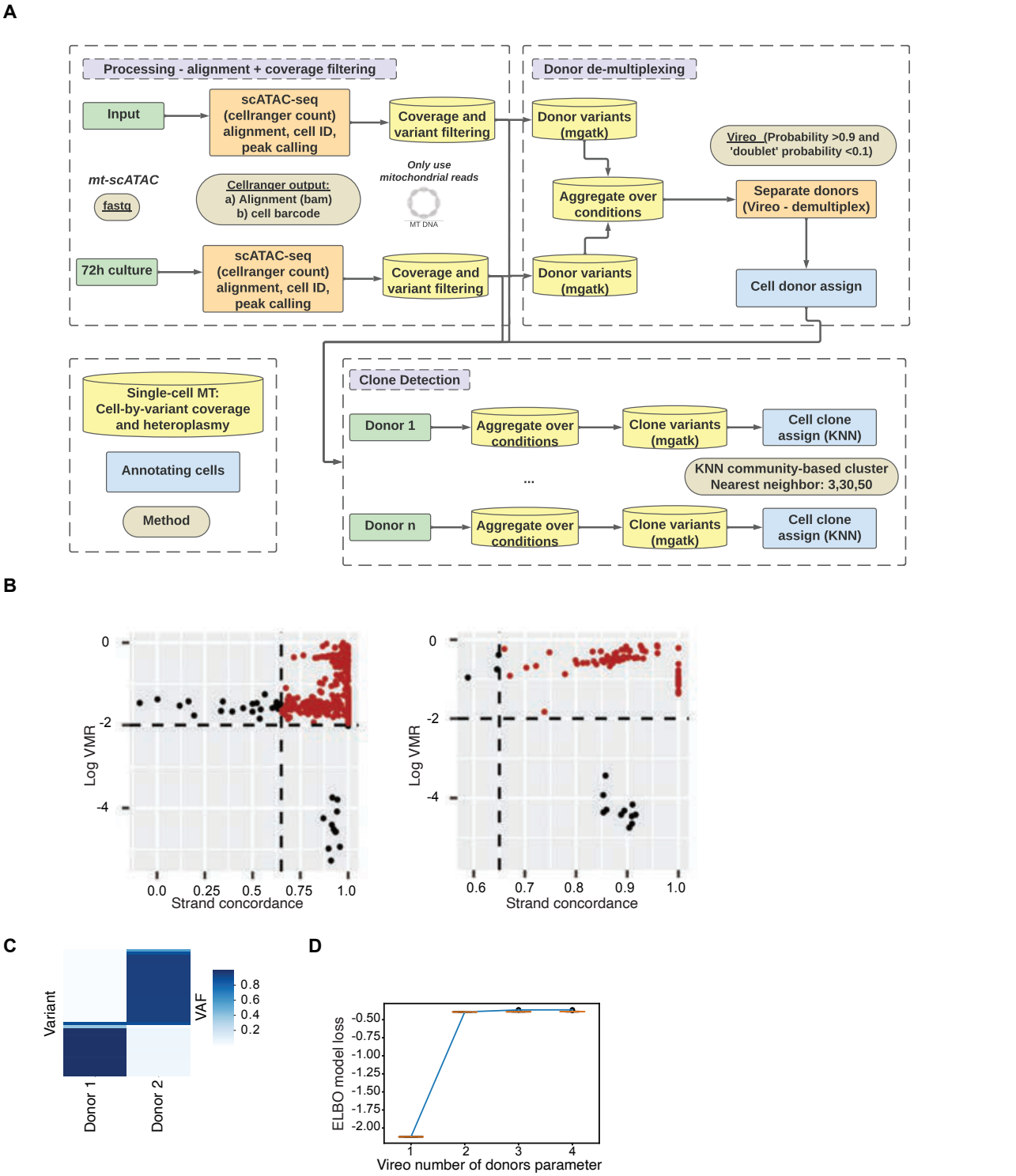

Figure S1. Pipeline and de-multiplexing in mt-scATAC-Seq experiments. (A) NGS processing, donor de-multiplexing and clone detection workflow. (B) MGATK algorithm used to call variants in the MT genome. Each point is a variant, and variants colored red pass the variance-mean ratio (VMR) and strand concordance thresholds. Left panel: input cells; right panel: 72 h culture. (C) Donor mean allele frequency. (D) The number of clusters (i.e. the number of donors) was varied and the Vireo likelihood score, the evidence lower bound (ELBO), was calculated. The “elbow rule” was then used to confirm that the true number of donors (n=2) was the inflection point in which performance gain was reduced when additional possible donors were added to the model.

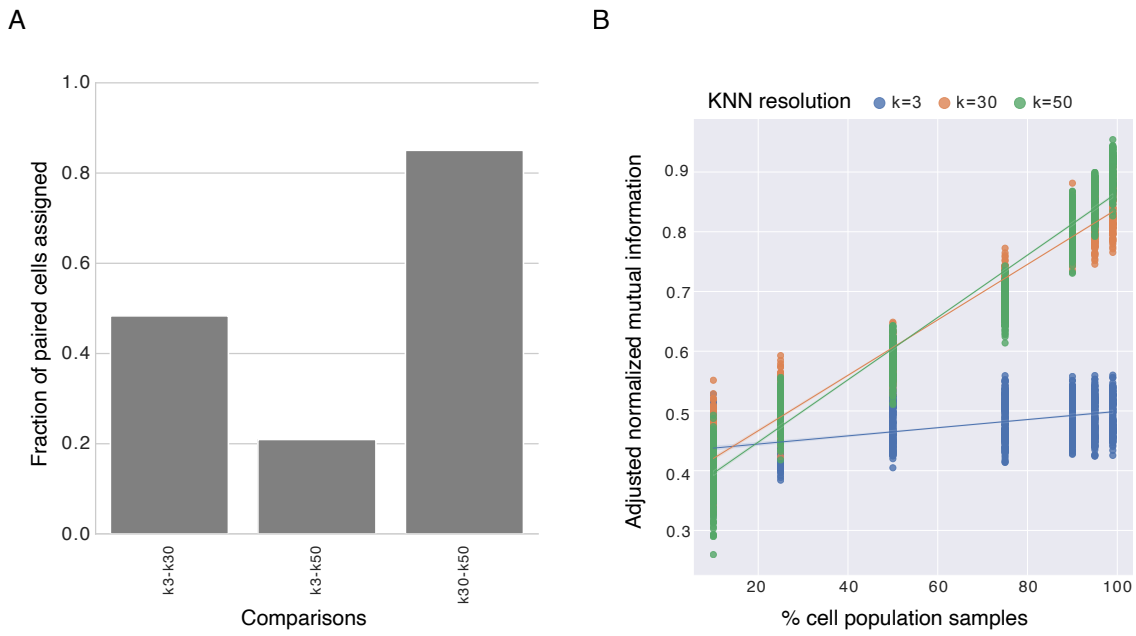

Figure S2. Clone assignment in mt-scATAC-Seq experiments. (A) Clone assignment comparison. K nearest neighbor of 3, 30, and 50 were compared by finding the number of cell pairs that are either in the same clone or in different clones in both methods. The score is then normalized to the fraction of paired cells assigned to the same clone across KNN resolution. (B) Subsampling cells from 10-99% 1000 times, calculating adjusted normalized mutual information between clones in sub-sampled run and the full population.

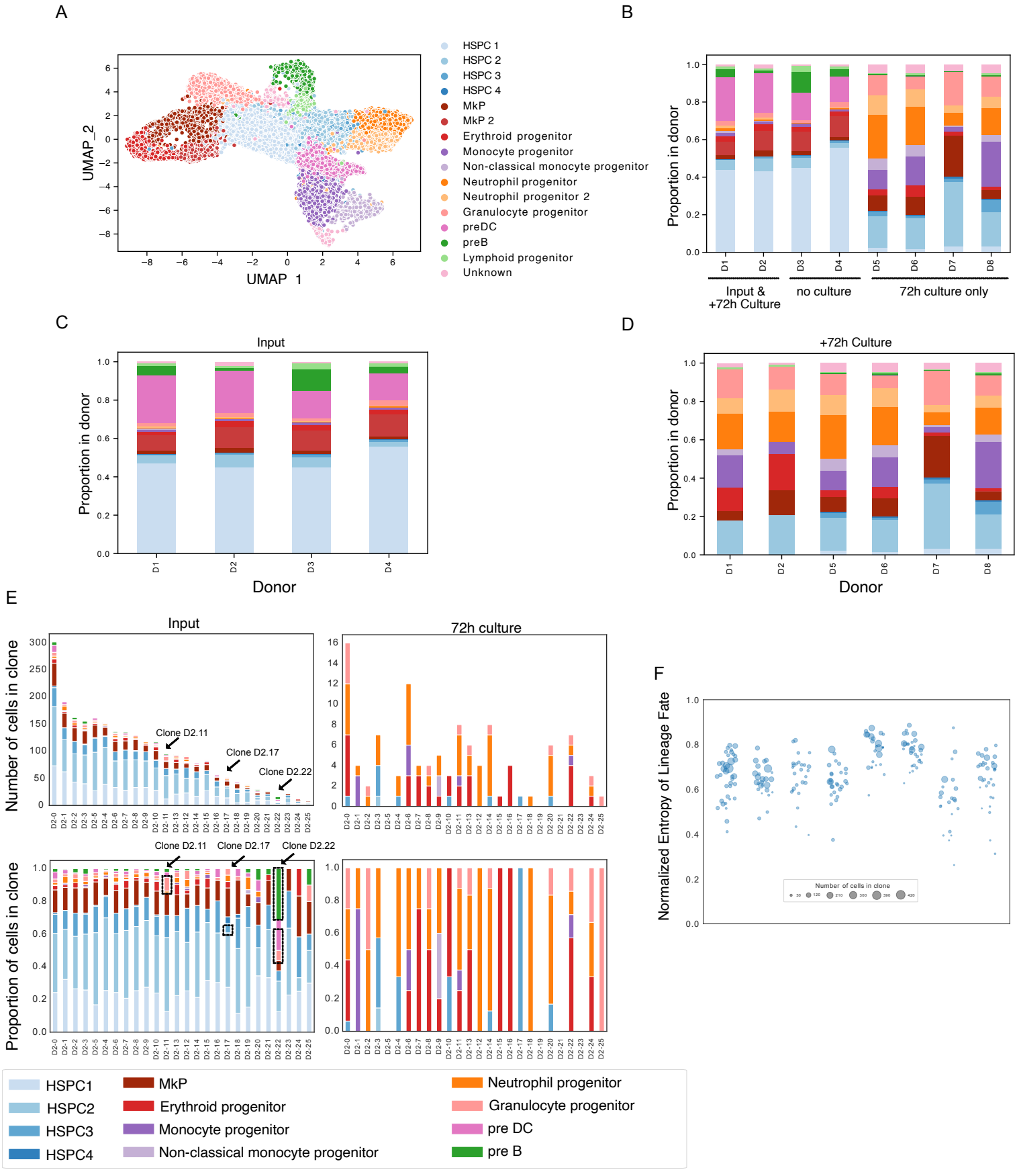

Figure S3. Minimum lineage-bias in human CD34+ HSPC clones across all donors. (A) Distribution of cells across all donors (n=8) and conditions on UMAP, colored by annotated cluster labels (B-D) Proportion of cells across HSPC clusters in each cell population studied, both input CD34+ cells and in cells cultured for 72h. (E) Raw cell counts (upper) and percent (lower) of immune lineage clusters in each clone for donor 2 before (input) and after 72 h culture. (F) Normalized entropy of lineage fate in each clone after 72h culture, sorted by rank within each donor.

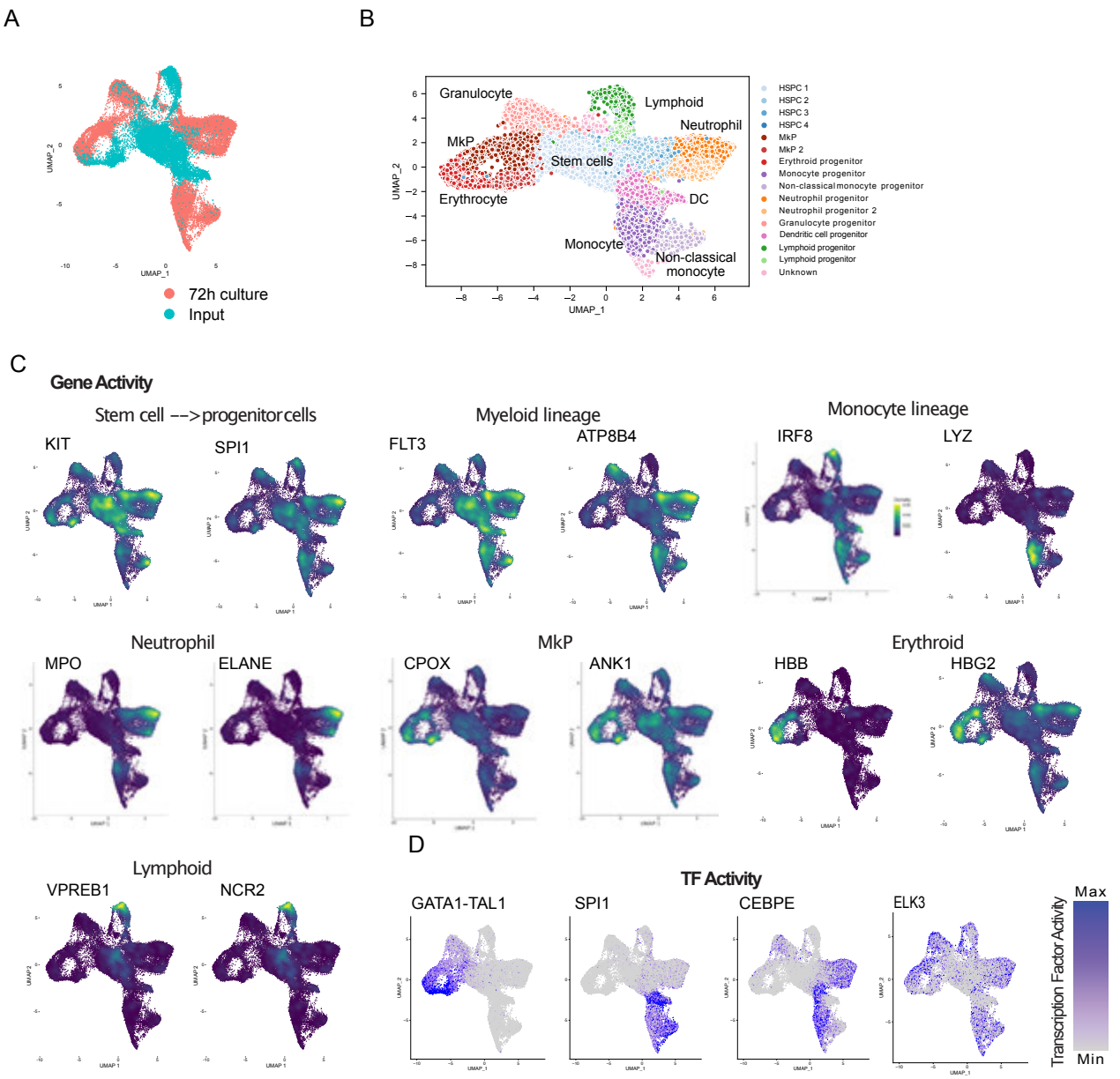

Figure S4. Characterizing cells by lineage markers in nuclear open-chromatin peaks across all donors. (A)All eight donor samples from input and/or cultured CD34+ cells overlaid on UMAP. (B) UMAP colored by cell type using Seurat's SNN method. (C) Gene activity scores for select markers overlaid on UMAP. (D) Transcription factor activity scores for select markers.

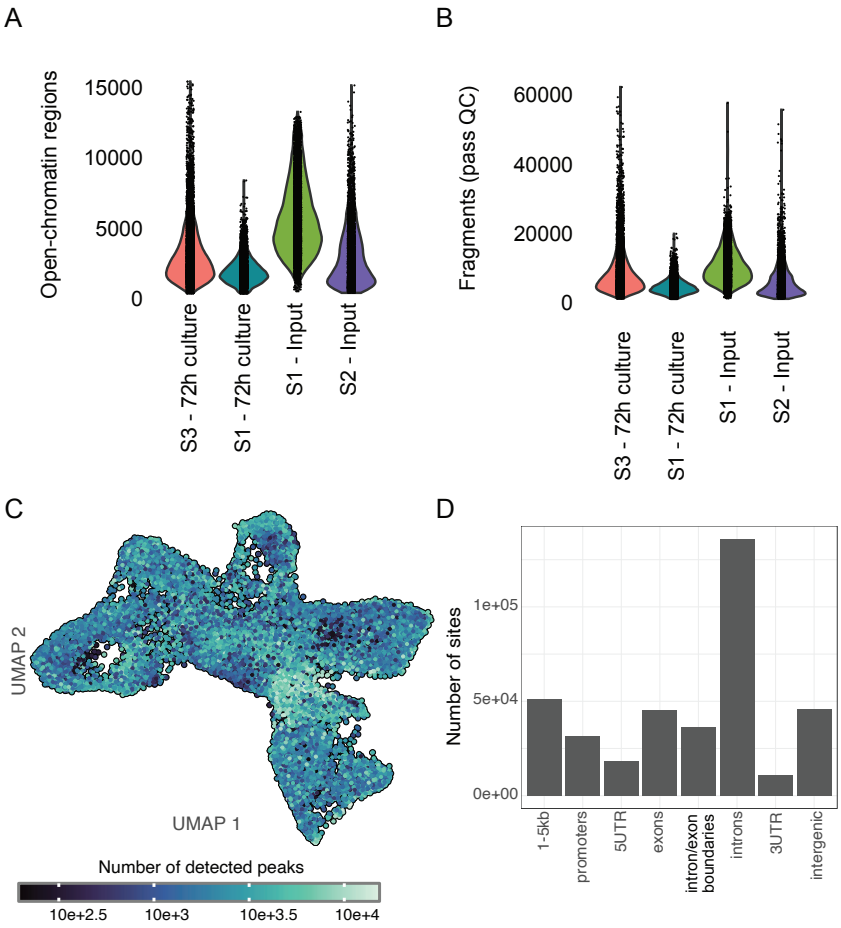

Figure S5. Detecting nuclear open-chromatin peaks. (A) Number of detected open-chromatin peak regions per cell. (B) Number of fragments that are non-duplicated and pass QC filters (see Methods). (C) UMAP illustrating the number of peaks detected for each cell. (D) The genomic location of each peak for each open-chromatin site detected.

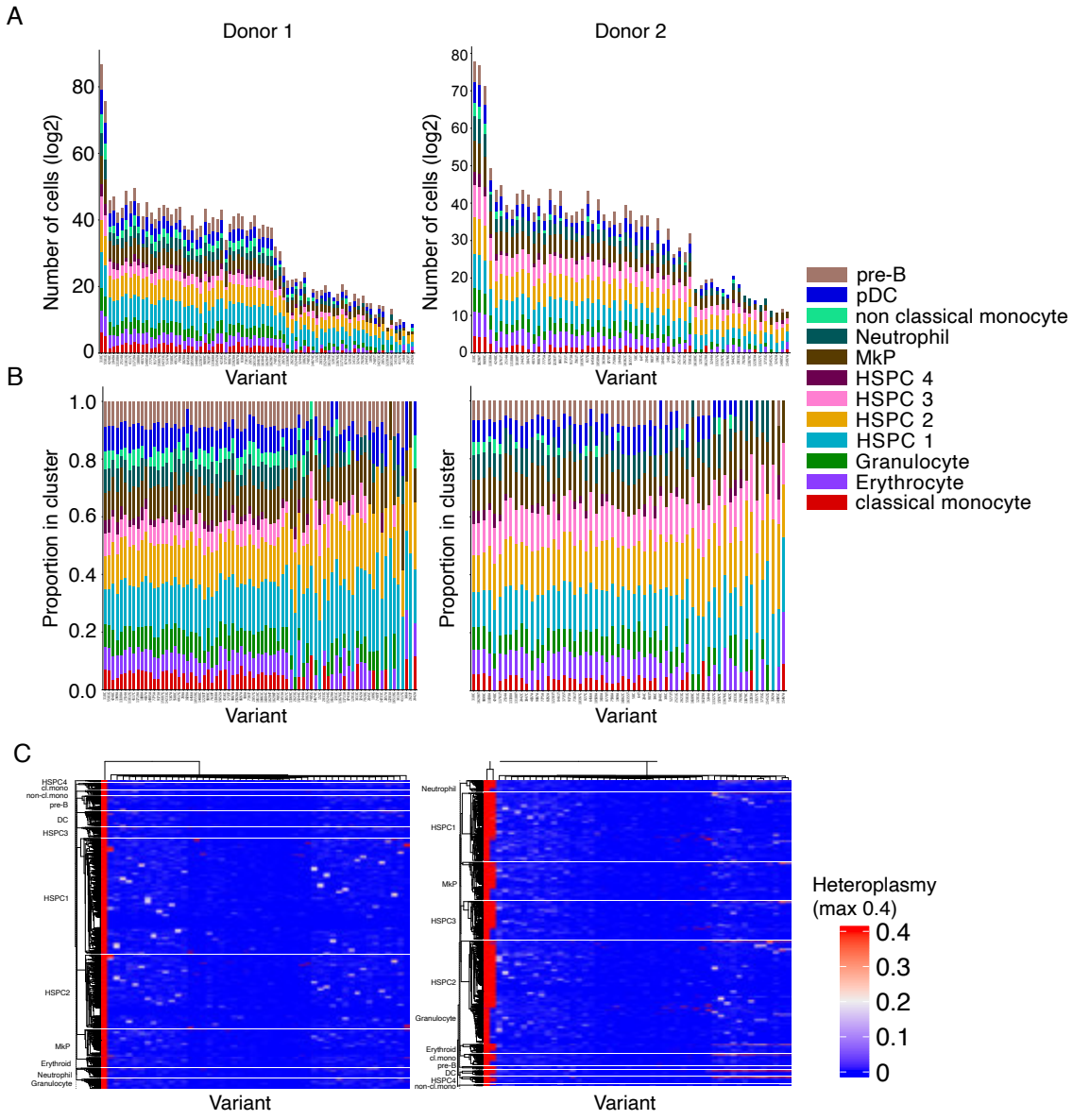

Figure S6. MT barcodes across lineage clusters. (A) Total cell counts (log2) for barcodes (allele frequency>0.01, coverage>10) across hematopoietic clusters. (B) Barcodes (allele frequency>0.01, coverage>10) across hematopoietic clusters normalized within each variant. (C) Cell-by-variant heteroplasmy heatmap for top differentiating variants in Donor 1 and Donor 2, ordered by single-linkage hierarchical clustering within each cell type.

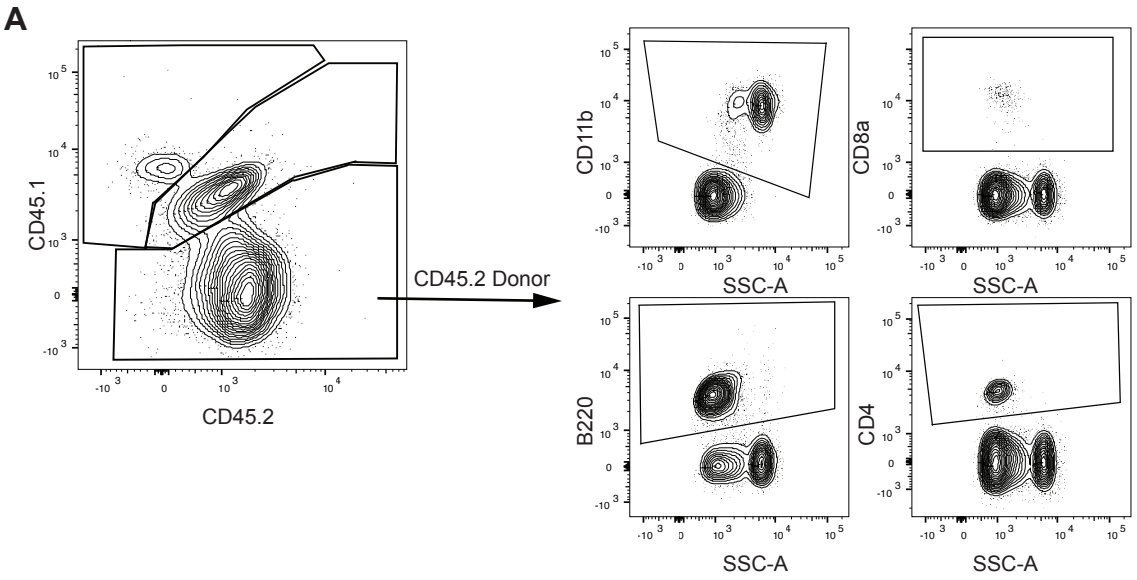

Figure S7. Differentiation of donor, support, and recipient cells in murine transplant experiments. Identification of donor hematopoietic cells (CD45.2) from recipient (CD45.1) and support (CD45.1/CD45.2) cells by flow cytometry to allow for appropriate quantification of donor LT-HSC transplant reconstitution lineage contribution, myeloid (e.g. CD11b), B cell (e.g. B220), T cell (e.g. CD4 and CD8).

Table S1. Clone characteristics in CD34+ cells from human donors

		Number of cells in clone			Number of nuclear peaks			Number of cells in clone (fraction of donor)		
Donor	Number of clones	mean +/- std	median	max	mean +/- std	median	max	mean +/- std	median	max
Donor 1	36	101.61 +/- 79.23	73	429	5612.50 +/- 355.74	5611	6406	0.03 +/- 0.02	0.02	0.12
Donor 2	26	100.85 +/- 70.69	97.5	317	5405.59 +/- 531.32	5389	7457	0.04 +/- 0.03	0.04	0.12
Donor 3	34	38.62 +/- 24.36	41	91	3095.06 +/- 644.63	3020	5184	0.03 +/- 0.02	0.03	0.07
Donor 4	27	76.11 +/- 48.80	62	247	3216.47 +/- 366.80	3154	3892	0.04 +/- 0.02	0.03	0.12
Donor 5	33	63.39 +/- 74.85	26	264	3016.17 +/- 517.84	3042	3895	0.03 +/- 0.04	0.01	0.13
Donor 6	35	56.63 +/- 52.50	40	193	3324.59 +/- 492.69	3352	4498	0.03 +/- 0.03	0.02	0.1
Donor 7	41	30.68 +/- 40.60	10	213	3029.99 +/- 567.27	2948	4437	0.02 +/- 0.03	0.01	0.17
Donor 8	50	35.58 +/- 39.73	23	180	2401.89 +/- 573.19	2281	3902	0.02 +/- 0.02	0.01	0.1