OMTM, Volume *18*

# Supplemental Information

# Purification of Human CD34$^{+}$CD90$^{+}$ HSCs Reduces Target Cell Population and Improves Lentiviral Transduction for Gene Therapy

**Stefan Radtke, Dnyanada Pande, Margaret Cui, Anai M. Perez, Yan-Yi Chan, Mark Enstrom, Stefanie Schmuck, Andrew Berger, Tom Eunson, Jennifer E. Adair, and Hans-Peter Kiem**

**SUPPLEMENTAL INFORMATION for**

# Purification of human CD34$^+$CD90$^+$ HSC reduces target cell population and improves lentiviral transduction for gene therapy

**Authors:** Stefan Radtke[1,2,#], Dnyanada Pande[1,2], Margaret Cui[1,2], Anai M. Perez[1,2], Yan-Yi Chan[1,2], Mark Enstrom[1,2], Stefanie Schmuck[1,2], Andrew Berger[2], Tom Eunson[2], Jennifer E. Adair[1,2], Hans-Peter Kiem[1,2,3,4,#]

## LIST OF SUPPLEMENTAL ITEMS:

## SUPPLEMENTAL METHODS:

### Expression Analysis for Bulk RNAseq

RNAseq expression analysis was performed in shared resources at the Fred Hutchinson Cancer Research Center. RNAseq libraries of GCSF-mobilized CD34 subsets were prepared using the NuGEN Ovation SoLo RNAseq System (Tecan Genomics, Redwood City, CA, USA). RNAseq libraries of steady-state BM CD34 subsets were prepared using the SMART-Seq v4 Ultra Low Input RNA Kit (Takara Bio Inc., Kusatsu, Japan) and Nextera XT Index Kit v2 (Illumina, Inc., San Diego, CA, USA). Work was performed on a Sciclone NGSx Workstation (PerkinElmer, Waltham, MA, USA). Library size distribution was validated using an Agilent 2200 TapeStation (Agilent Technologies, Santa Clara, CA, USA). Additional library QC, blending of pooled indexed libraries, and cluster optimization was performed using Life Technologies Invitrogen Qubit® 2.0 Fluorometer (Life Technologies-Invitrogen, Carlsbad, CA, USA). RNAseq libraries were pooled and clustered onto a flow cell lane.

### Quantification of Transcripts

The quantification was performed using kallisto (v0.43.1).[51] Human genome assembly (GRCh38) from National Center for Biotechnology Information (NCBI) was used as the reference. The compressed fastq files (.fastq.gz) were input to kallisto. The human reference transcriptome was processed to create a transcriptome index using “kallisto index ” option with the default k-mer length. The abundances of the transcripts were quantified by aligning the raw reads to the reference with bootstrapping, using the option “kallisto quant -b 100”. The bootstrapping was performed to obtain confidence intervals on transcript quantification. Kallisto generated two output files with the alignment information. The abundances.tsv reported the abundances as estimated counts (est_counts) and transcripts per million (tpm), while the abundances.h5 file had the abundance estimates, bootstrap estimates, transcript length information, and the run information.

***Data analysis.*** The counts (abundances.tsv) from kallisto were imported into R in the form of a matrix with the tximport package (v.1.10.1). The Human RefSeq Reference Genome Annotation file (v.38_p12), was downloaded from the Human Genome Resources at NCBI to obtain the gene IDs. Each transcript ID and its count was then associated with the corresponding gene ID for summarization of gene-level counts. The count matrix was analyzed for differential gene expression using the DESeq2 package from Bioconductor in R (v.1.22.2).[52] The count matrix was pre-filtered by keeping the rows that have a minimum of one transcript before analysis with DESeq2. The result obtained was a list of differentially expressed genes with significant p-values and log-fold changes. Clustering and principal component analysis (PCA) was performed on the normalized data, which identified the genes that were contributing to the variance in the samples.

### Alignment and Counting

The 10X Genomics Cell Ranger software suite (v2.0.0) was used to covert the raw sequence reads into single-cell gene expression counts. The “cellranger count” command with default option was run for alignment, filtering, cell barcode counting, and UMI counting. Cell barcode is a known nucleotide sequence that acts as a unique identifier for a single GEM (Gelbead-in-Emulsion) droplet. Each barcode contains reads from a single cell. UMIs are random 10bp nucleotide sequences that help determine which reads came from the same transcript. The cDNA was aligned to human reference genome (hg38) using the STAR aligner (v.2.6.1). UMIs were also filtered for a minimum of Qual = 10. Reads were marked as PCR duplicates if two or more read pairs shared the same cell barcode, UMI, and gene ID. Valid cell barcodes were determined based on the final UMI distributions. Valid cell barcodes with a valid UMI mapped to exons (Ensembl GTF GRCh38) were used to generate the final cell barcode matrix (.mtx).

### Dimensional Reduction and Clustering

The single cell data analysis was performed using Seurat (v2.3.4),[26] an R toolkit for single cell genomic data. The 10X runs for the CD34$^+$ cells and the CD34-subsets were merged by combining the cell barcode matrices into a single Seurat object. The gene expression data for each cell was log normalized. The genes were regressed based on the number of UMIs (nUMI), then scaled and centered to improve downstream analysis. PCA was run on the highly variable genes to compute linear dimensional reduction. The cells were clustered based on similar gene expression patterns using the first 10 principal components (PC) with a resolution of 0.4. t-distributed Stochastic Neighbor Embedding (tSNE) was used to visualize the gene clusters and the CD34$^+$ cells and CD34-subsets. The positively differentially expressed genes were found for all the clusters based on the Wilcoxon rank sum test with a log-fold change threshold of 0.25. The gene expression patterns of marker genes were visualized on a tSNE dimensional reduction plot and a PCA dimensional reduction plot.

### Single Cell and Bulk RNAseq Combined Analysis

DESeq2 (v.1.22.2) was run on the bulk RNA data as described above to create an un-normalized count matrix. The raw counts were transformed into a Single Cell Experiment (SCE) object along with the corresponding donor and gene information. The SCE is an R package that includes methods to store single cell data information. The raw counts were used to compute the normalized counts and log counts, which are necessary to convert the SCE data object into a Seurat data object. Using the Seurat package, the bulk RNAseq data was converted from an SCE object to a Seurat object. UMI counts were generated for the bulk RNA data and added as metadata to the object. Next, the bulk RNAseq data was merged with the single cell RNAseq data to create a combined Seurat dataset. The combined dataset was then log normalized and scaled as described above. This maintained uniformity in the scaling and normalization of both the single cell RNA and bulk RNA data together.

### Transforming the Data with Significant Principal Components

The 10X run for the CD34$^+$ cells was also analyzed by Seurat (v.2.3.4). The data was normalized and scaled. Variable genes were identified for the data and PCA was run on the variable genes. The genes were clustered using the first 10 PCs with a resolution of 0.4. The genes that defined PC1 and PC2 were extracted from the Seurat object.

A matrix was created by sub-setting the scaled count data matrix of the combined dataset using the PC1 and PC2 genes from the CD34$^+$ data. This matrix was then multiplied with PC1 and PC2 values. The combined single cell and bulk RNA data was thus linearly transformed with the CD34$^+$ cells as the reference and was used for further downstream analysis.

### Overlaying the Cell Populations on the Reference CD34$^+$ Cell Population

Points specific for each of the different cell types from the bulk RNA data, the CD34$^+$ cell population and the CD34 subset cell population were extracted from the combined dataset. PCA was used as the linear dimensional transform. The CD34$^+$ population was plotted as the reference, and the cell types from the bulk RNA data were overlaid on the reference to see where they map. Similarly, the CD34 subsets were visualized against the CD34$^+$ reference map.

### Software and Packages

FlowJo v.10.2 and higher https://www.flowjo.com

Kallisto v.0.43.1 - https://pachterlab.github.io/kallisto.2 -
http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html
Tximport v.1.10.1 - http://bioconductor.org/packages/release/bioc/html/tximport.html
10X Genomics Chromium- https://www.10xgenomics.com/product-list/#single-cell
10X Genomics Cell Ranger v.2.0.0 - https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome
STAR aligner v.2.6.1 - https://github.com/alexdobin/STAR
Seurat v.2.3.4 - https://satijalab.org/seurat/
SCE v.1.4.1 -
https://www.bioconductor.org/packages/release/bioc/html/SingleCellExperiment.html

**Table S1.** Experimental parameters for scRNAseq

| **Donor** | **1** | | | | **2** | |
|---|---|---|---|---|---|---|
| **Population** | **CD34$^+$** | **CD133$^+$** | **CD38$^{low/-}$** | **CD90$^+$** | **CD34$^+$** | **CD90$^+$** |
| # of cells | 2,162 | 2,019 | 2,472 | 1,523 | 1,449 | 1,189 |
| Mean reads/cell | 75,692 | 71,796 | 64,918 | 75,971 | 62,234 | 64,282 |
| Sequencing saturation | 76.6% | 76.9% | 81.2% | 81.7% | 75.2% | 79.1% |
| Fraction reads in cells | 94.7% | 94.6% | 96.7% | 96.4% | 90.5% | 94.5% |
| Valid barcodes | 97.7% | 98.0% | 97.6% | 97.9% | 98.4% | 98.4% |
| Total genes detected | 18,132 | 18,036 | 18,150 | 17,430 | 17,183 | 16,321 |
| Q30 bases in barcodes | 97.4% | 96.8% | 97.4% | 97.0% | 96.2% | 96.2% |
| Q30 bases in RNA reads | 91.3% | 87.2% | 87.7% | 86.0% | 73.5% | 73.6% |
| Q30 bases in sample index | 96.3% | 96.3% | 96.5% | 96.2% | 96.2% | 95.0% |
| Q30 bases in UMI | 97.5% | 96.8% | 97.4% | 97.0% | 96.3% | 96.3% |

**Table S9.** Summary of mobilization, leukapheresis and CD34 enrichment parameters

| Donor | GCSF dose | # collections | WBC count | CD34 count | CD34 purity [%] | Comment |
|---|---|---|---|---|---|---|
| 1 | 5mg/kg | 2 | 4.74e10 | 2.68e8 | 99.00 | Cryopreserved |
| 2 | 5mg/kg | 2 | 9.70e10 | 2.88e8 | 92.00 | Cryopreserved |
| 3 | 5mg/kg | 2 | 7.72e10 | 2.87e8 | 91.80 | Cryopreserved |
| 4 | 5mg/kg | 2 | 2.07e10 | 5.92e6 | 35.10 | Discontinued |
| 5 | 5mg/kg | 1 | 3.00E+10 | 1.30e8 | 97.00 | Fresh processing |
| 6 | 7.5mg/kg | 1 | 5.98e10 | 3.70e8 | 87.90 | Fresh processing |

All donors were selected for adjusted body weight >120% of ideal body weight.

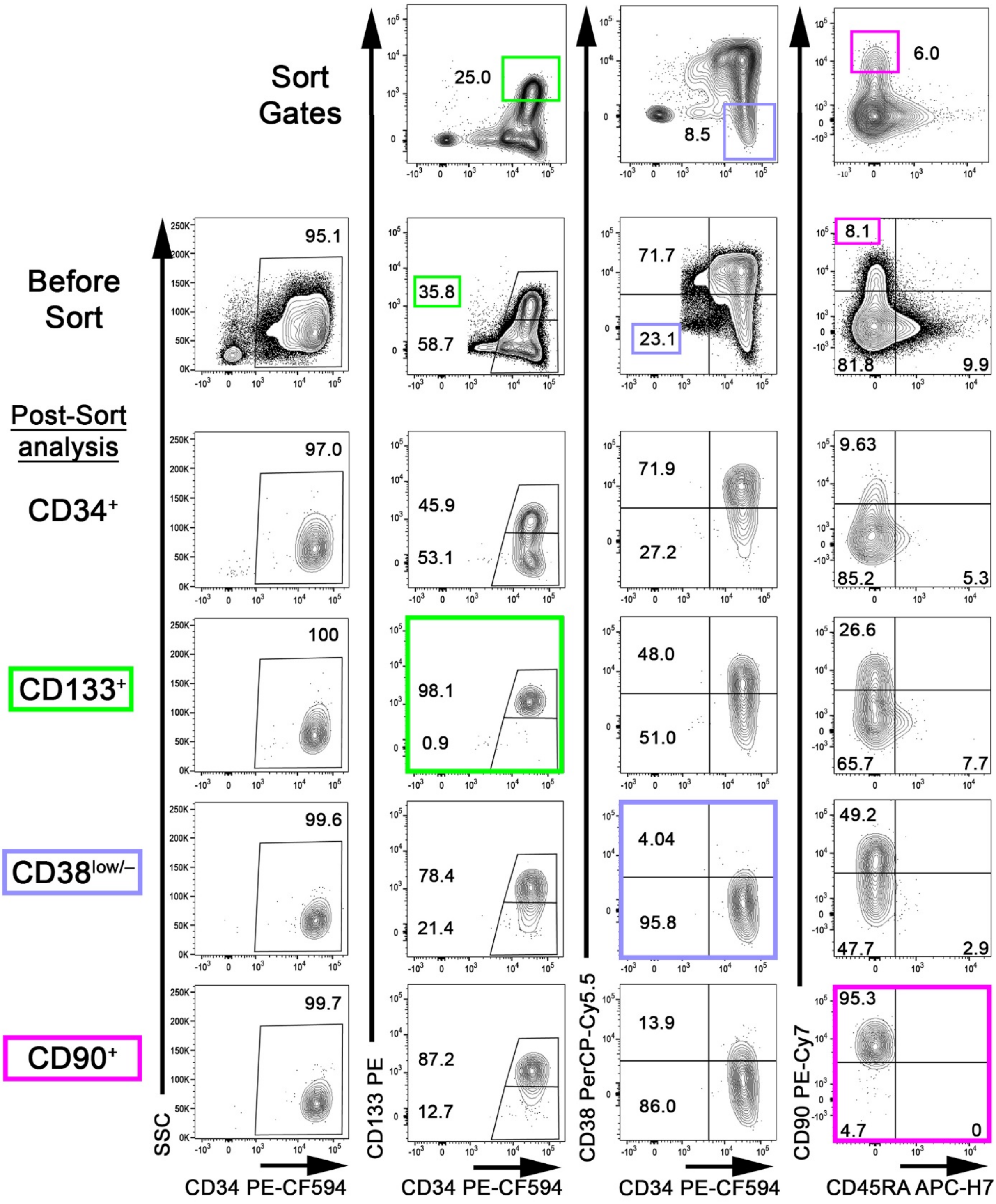


**Figure S1. Quality control of sort-purified CD34-subpopulations. Sort gates (top row, Sort Gates),** flow-cytometric assessment of bulk CD34+ cells (2nd row, Before Sort) and sort-purified CD34+ (3rd row), CD133+ (4th row), CD38low/- (5th row) and CD90+ (6th row) HSPCs (Post-Sort analysis). Sorted target cell fractions are framed and color-coded. Numbers indicate frequency of gated population.

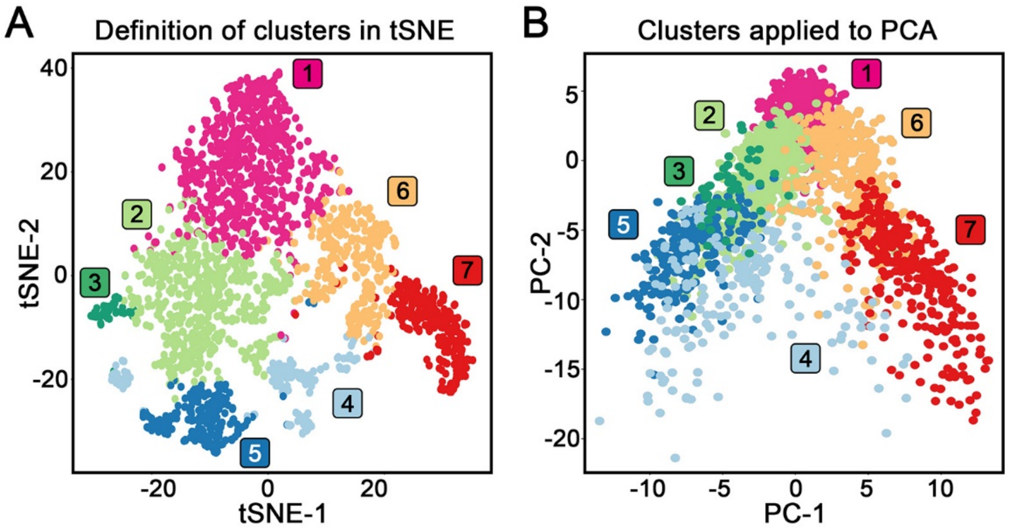


**Figure S2. Transcriptionally distinct ssBM CD34 clusters in a second donor.** (**A**) Graph-based clustering of ssBM-derived CD34$^+$ cells. Transcriptionally distinct CD34 clusters were color-coded and numbered. (**B**) Clusters defined in A projected onto the PCA analysis.

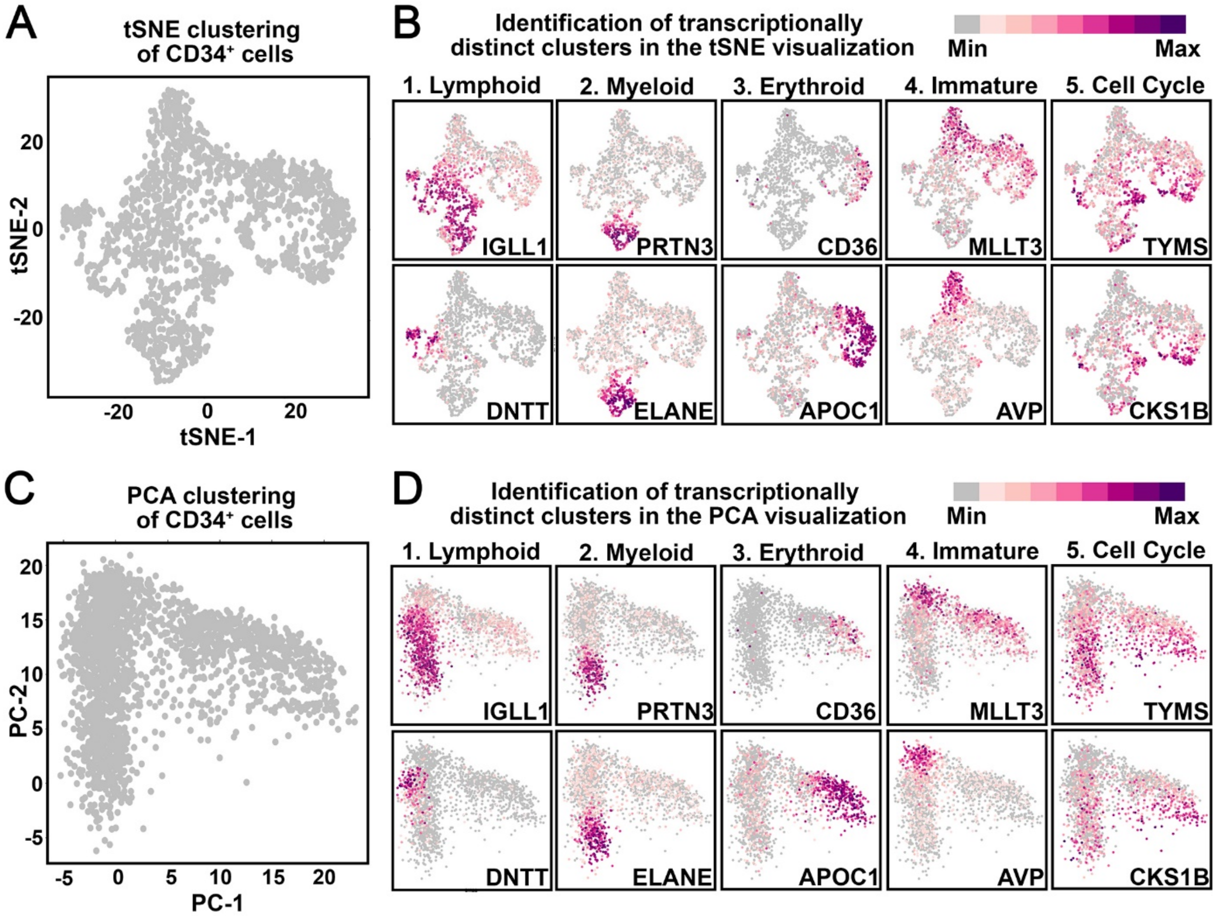


**Figure S3. ScRNAseq of ssBM-derived CD34$^+$ HSPCs and sort-purified CD34 subsets.** (**A**) Dimensional reduction (tSNE) of scRNAseq data from ssBM-derived CD34$^+$ cells. (**B**) Feature plots showing the expression of representative genes associated with lymphoid-, myeloid-, erythroid-primed, immature, and proliferating HSPCs. Level of expression is color coded as shown in the legend. (**C**) PCA based transformation and (**D**) expression of representative genes for the same dataset shown in panel A.

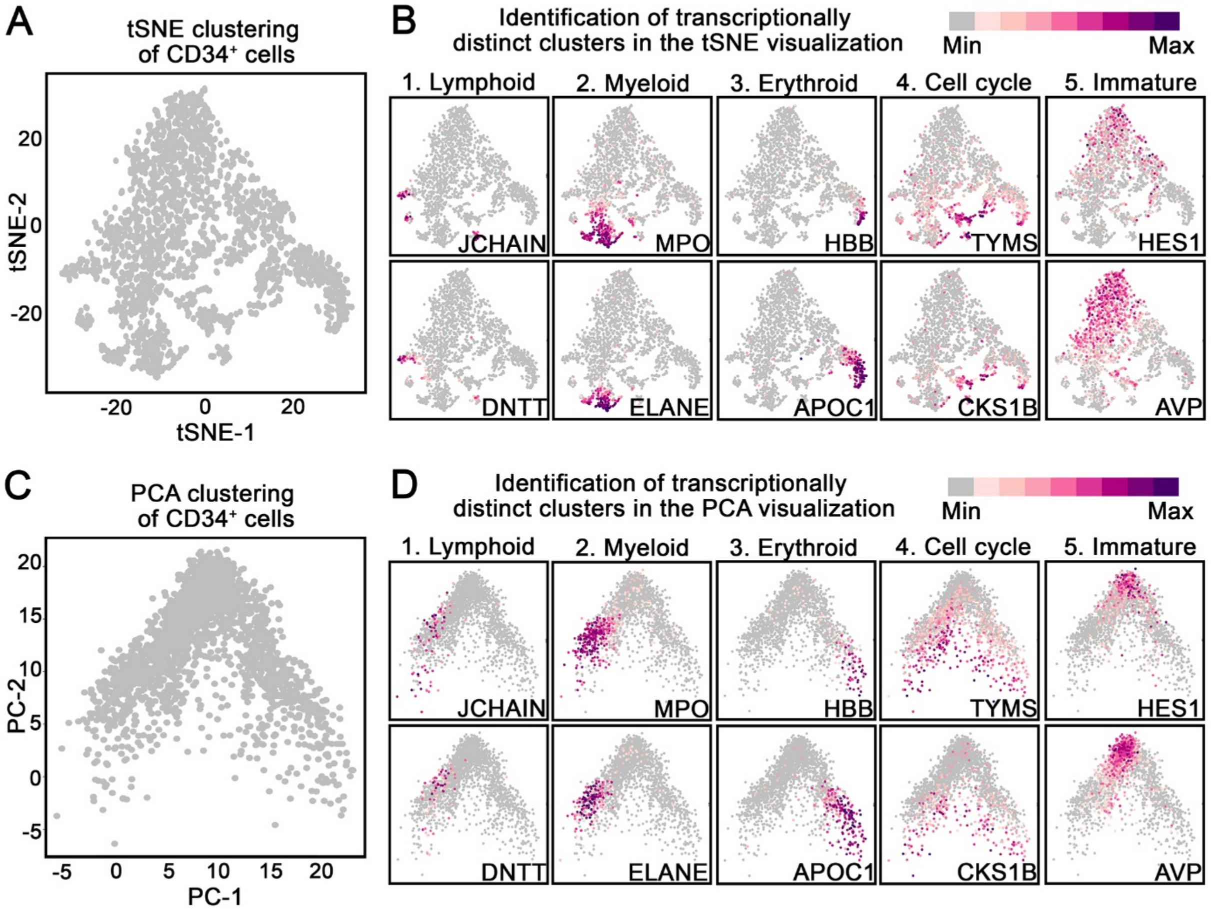


**Figure S4. Donor-independent reproducibility of the scRNAseq ssBM reference map.** (**A**) tSNE and (**C**) PCA clustering of scRNAseq data from ssBM-derived CD34⁺ cells from a second donor. (**B** and **D**) Feature plots showing the expression of representative genes associated with lymphoid-, myeloid-, erythroid-primed, proliferating, and immature HSPCs. Level of expression is color coded as shown in the legend.

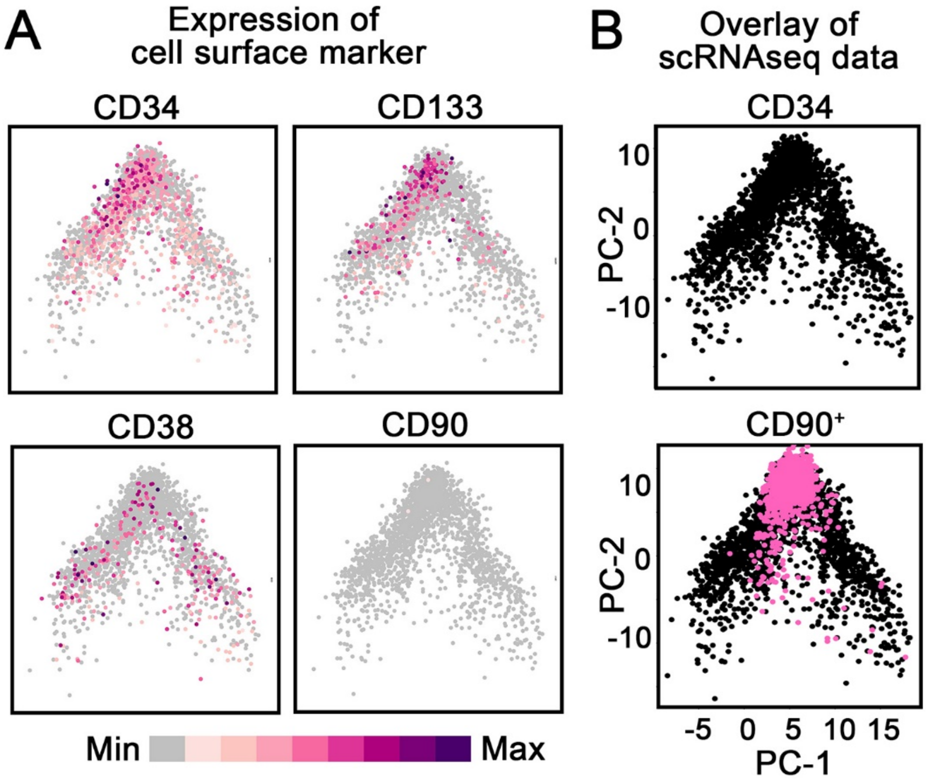


**Figure S5. Transcriptional mapping of sort-purified CD34 subsets from a second donor.** (**A**) Expression of CD34, CD133, CD38 and CD90 in ssBM-derived CD34$^+$ cells. Level of expression is color coded as shown in the legend. (**B**) Overlay of scRNAseq data from CD34$^+$ cells (black, top plot) with sort-purified CD90$^+$ (pink, lower plot) HSPCs.

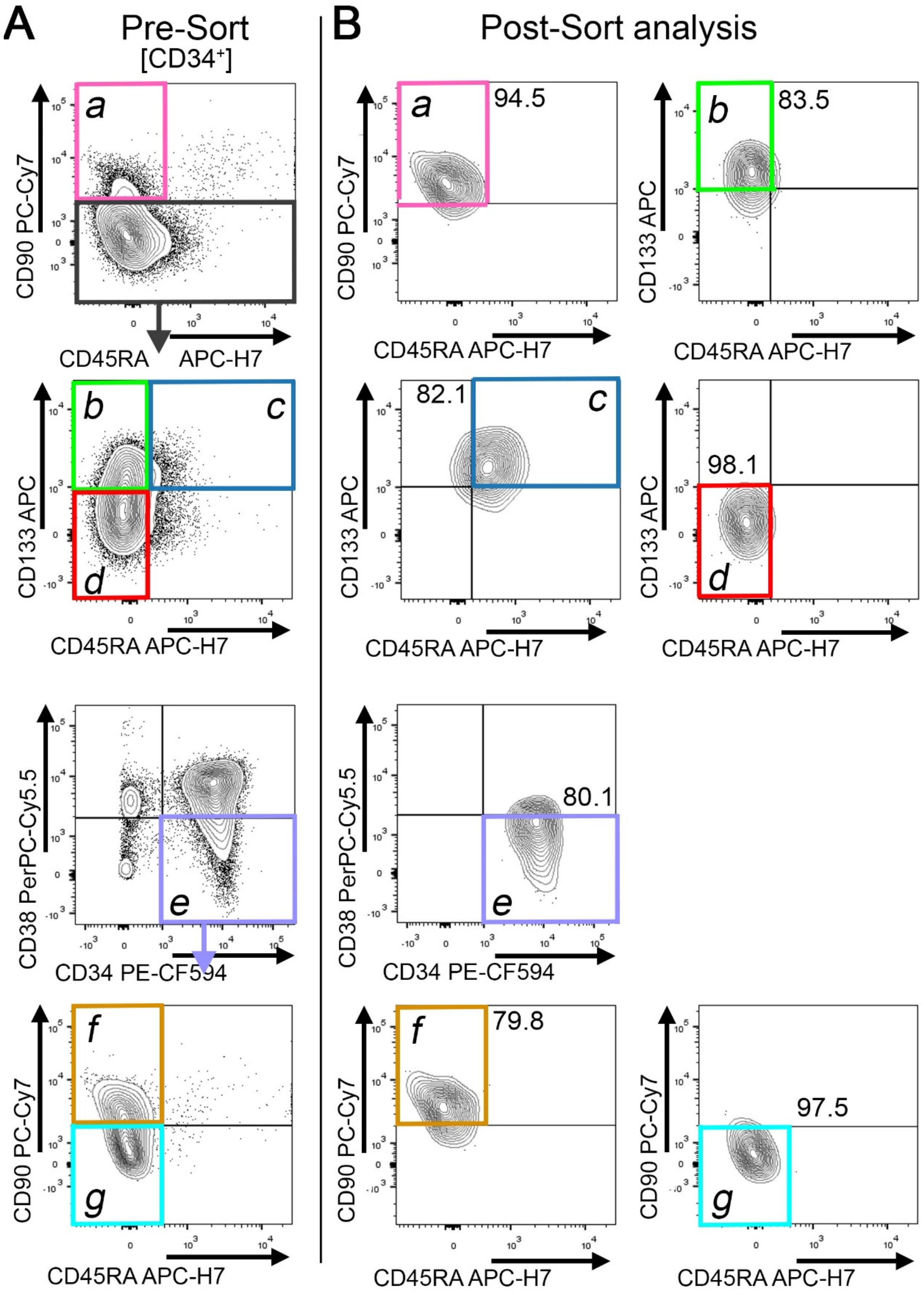
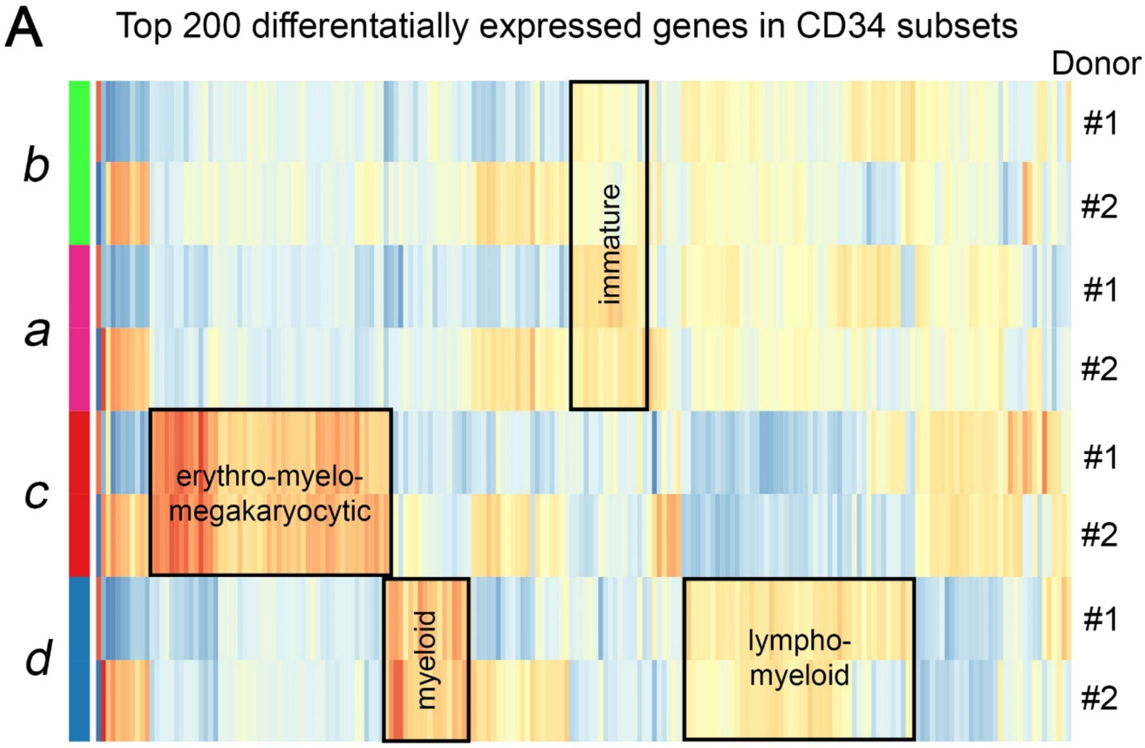


**Figure S6. Quality control of sort-purified CD34-subpopulations for bulk RNAseq.** (**A**) Gating of ssBM-derived CD34 subpopulation defined in Figure 3A. (**B**) Flow-cytometric quality control of sort-purified CD34$^+$ subsets for bulk RNAseq. Sorted cell fractions are framed and color-coded. Numbers indicate frequency of gated population.



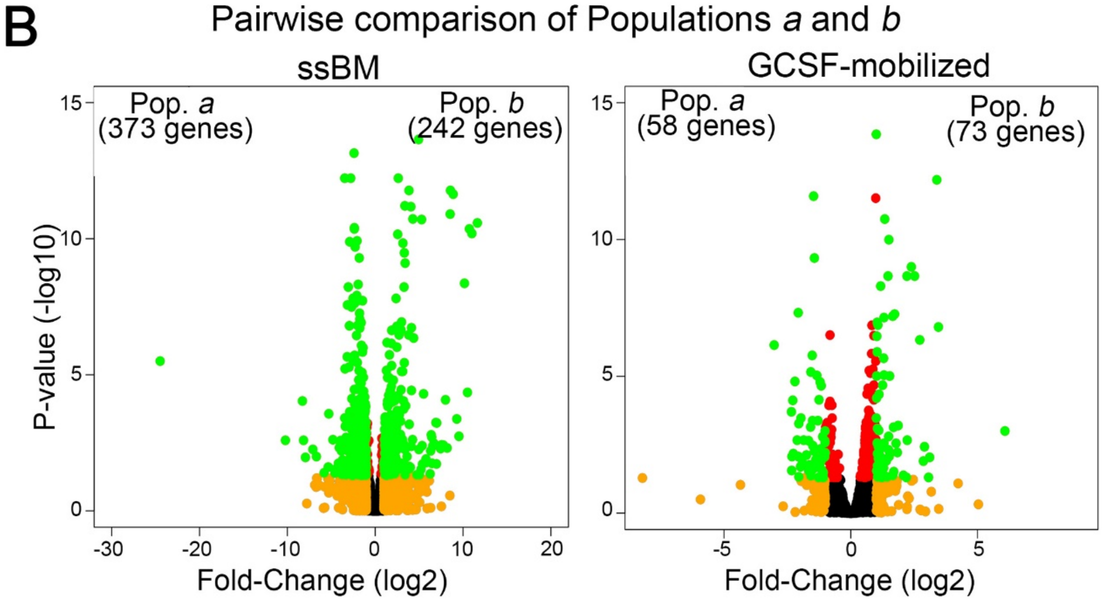


**Figure S7. Differentially expressed genes in GCSF-mobilized bulk CD34 subsets.** (**A**) Heat map of the Top 200 differentially expressed genes in phenotypic GCSF-mobilized CD34 subpopulations *a*–*d* from two independent human donors. (**B**) Pair-wise comparison of the gene-expression in the ssBM and GCSF-mobilized subpopulations *a* and *b*. Differentially expressed genes are color coded according to the figure legend in the top left. Color-code: green = p-value < 0.05 and fold-change (FC) >1; red = p-value > 0.05 and FC >1; yellow = p-value > 0.05 and FC >1; black = p-value > 0.05 and FC < 1.

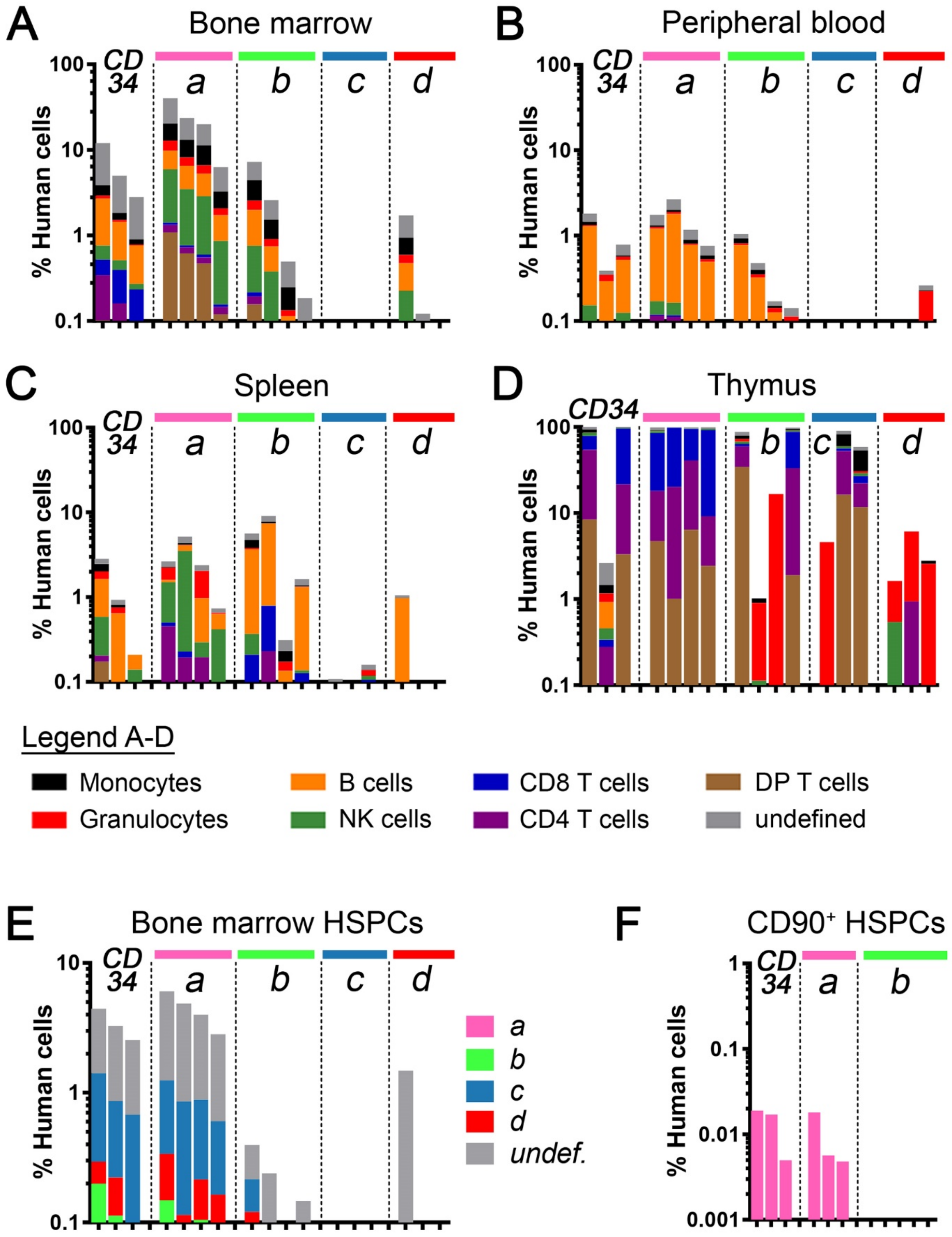


**Figure S8. Multilineage engraftment potential of human CD34 subpopulations.** Human multilineage engraftment in the (**A**) BM, (**B**) PB, (**C**) spleen and (**D**) thymus after transplantation of bulk CD34$^+$ HSPCs as well as sort-purified CD34 subpopulations (1e5 cells per mouse). Mice in all graphs and within each group are organized from the highest to the lowest engraftment level in the BM (A). (**E**) Frequency of human CD34$^+$ cells (total height of bars) and CD34$^+$ subpopulations (color-coded, as defined in Figure 3A). (**F**) Frequency of human CD90$^+$ HSPCs in the BM of mice transplanted with populations *a* and *b* only.

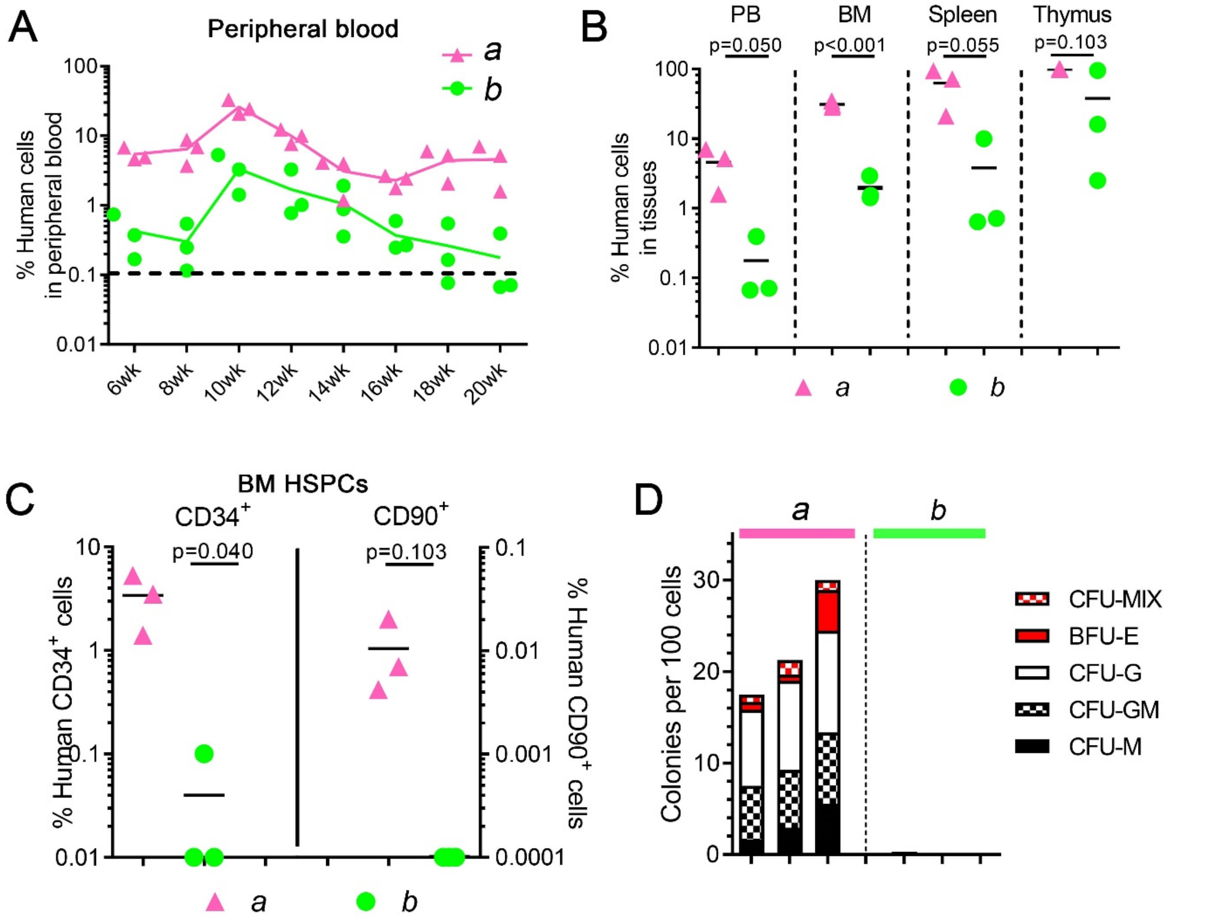


**Figure S9. Engraftment potential of human CD34 subsets.** (**A**) Longitudinal tracking of human CD45$^+$ engraftment in the PB of mice transplanted with $2.5\times10^5$ HSPCs cells per mouse from Population *a* or Population *b*. (**B**) Side-by-side comparison of human CD45$^+$ engraftment in the PB, BM, spleen and thymus. PB and BM use left y-axis, spleen and thymus right y-axis. (**C**) Frequency of human CD34$^+$ cells (left y-axis) and CD90$^+$ HSPCs (right y-axis) in the BM of engrafted mice. (**D**) Erythroid, myeloid and erythro-myeloid colony-forming potential of engrafted human HSPCs.

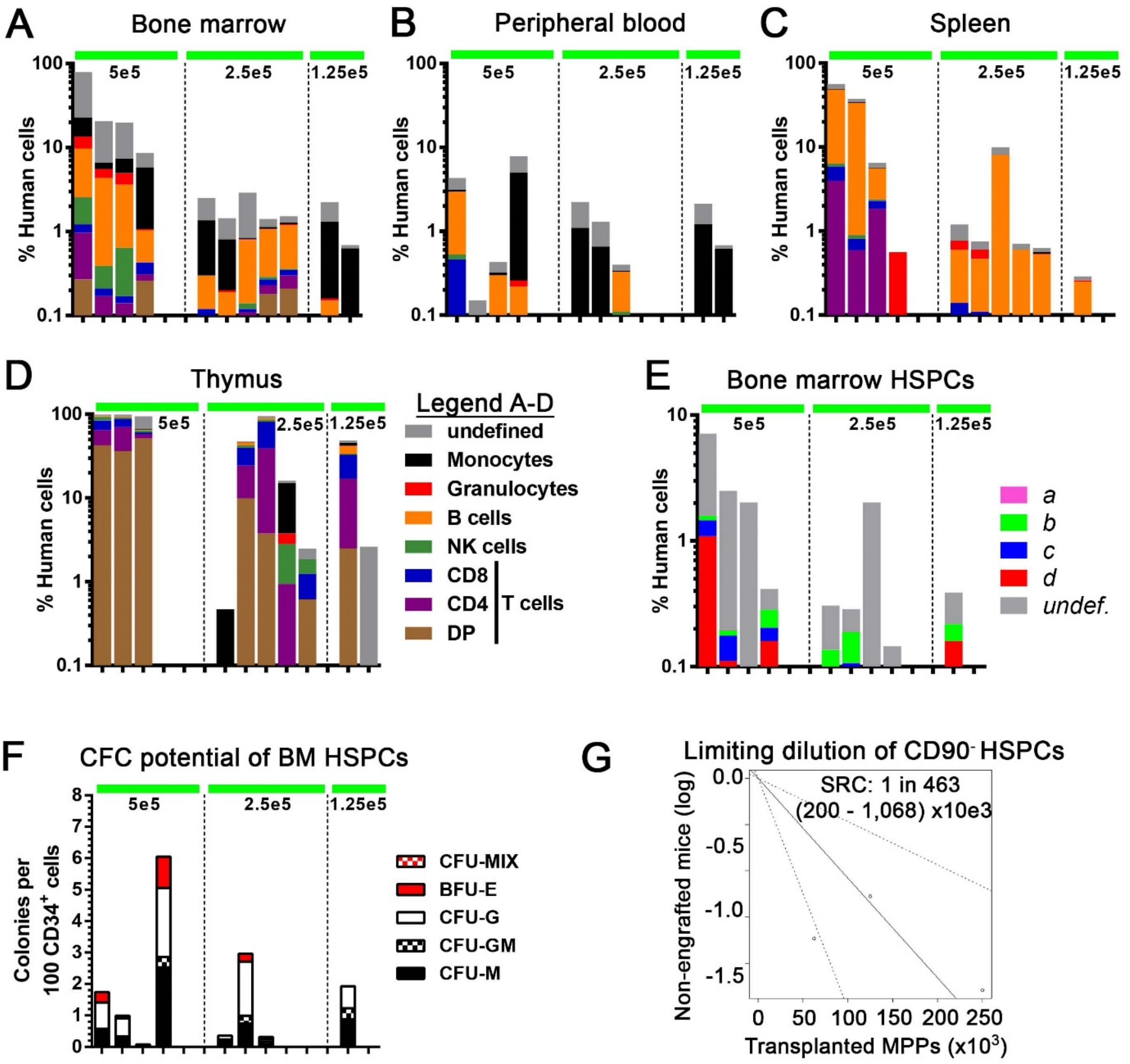


**Figure S10. Engraftment potential of human HSPCs from Population *b*.** Human multilineage engraftment in the (**A**) BM, (**B**) PB, (**C**) spleen and (**D**) thymus. Mice in all graphs and within each group are organized from the highest to the lowest engraftment level in the BM (A). (**E**) Frequency of human CD34$^+$ cells (total height of bars) and CD34$^+$ subpopulations in the BM of transplanted mice. (**F**) Erythroid, myeloid and erythro-myeloid colony-forming potential of engrafted human HSPCs. (**G**) Calculation of human HSPCs from population *b* with SRC potential using a limiting dilution approach as previously described.[50]

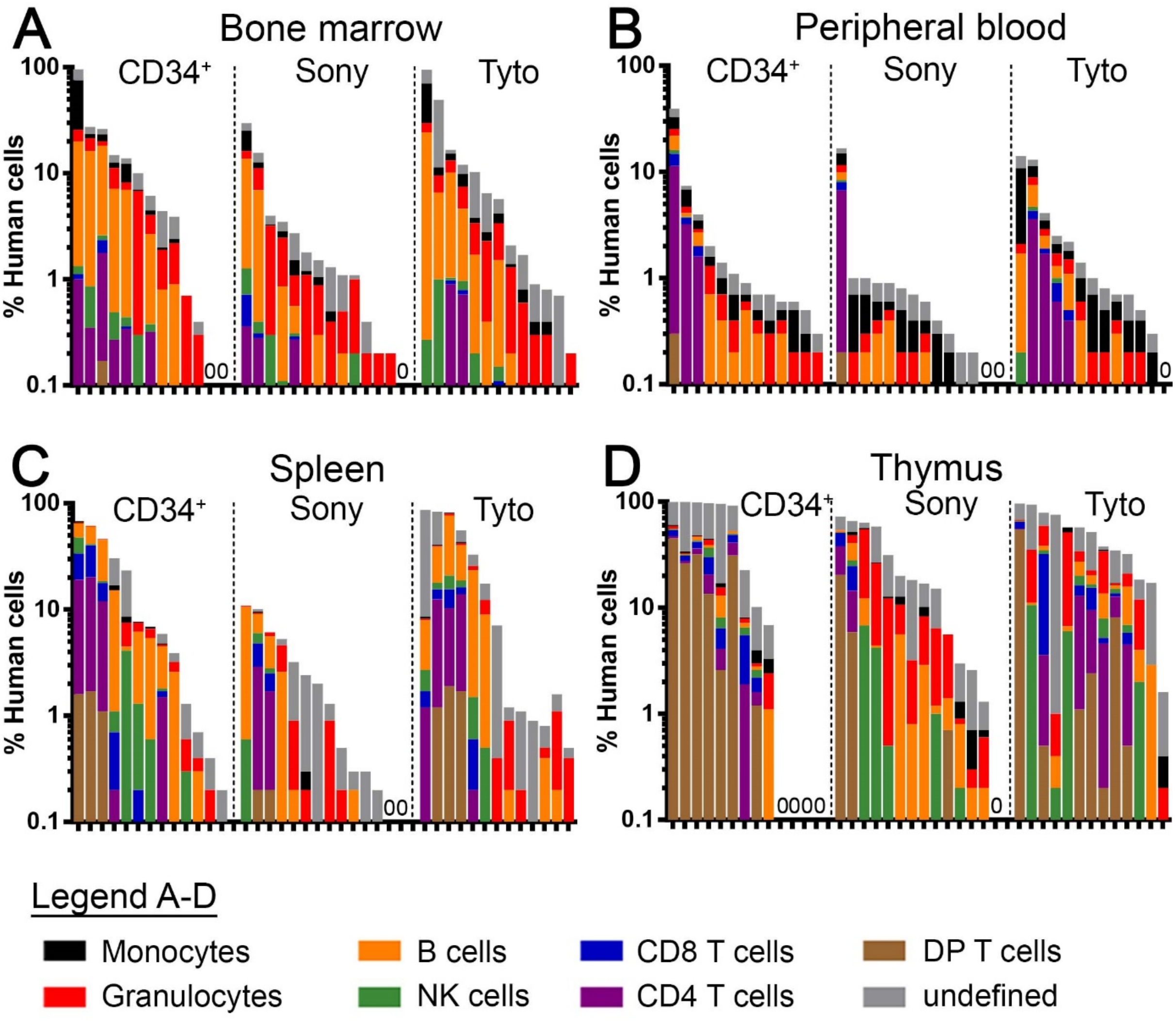


**Figure S11. Engraftment potential of gene-modified human bulk CD34$^+$ and sort-purified CD34$^+$CD90$^+$ cells.** Human multilineage engraftment in the (**A**) BM, (**B**) PB, (**C**) spleen and (**D**) thymus of transplanted mice at 20 weeks post-transplant.