Transcriptional landscape of direct reprogramming toward the hematopoietic lineage

Jillian Cwycyshyn[1,2, †], Cooper Stansbury[1,†], Sarah Golts[1], Hyunsu Lee[3], Joshua Pickard[1], Walter Meixner[1], Indika Rajapakse[1, 2, 4], and Lindsey A. Muir[1*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109 USA
[2]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48109 USA
[3]iReprogram Inc., Ann Arbor, MI
[4]Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 USA
[†]These authors contributed equally.
[*]To whom correspondence should be addressed (+1-734-904-3013; lindmuir@umich.edu; 1600 Huron Parkway, NCRC B520-3368, Ann Arbor, MI 48109).

2026-01-16

Abstract

Direct reprogramming of human fibroblasts into hematopoietic stem cells (HSCs) offers a promising strategy for generating autologous cells to treat blood and immune disorders. However, low conversion efficiency and incomplete characterization of reprogrammed states limit current protocols. We therefore assembled a single-cell transcriptomic reference atlas spanning hematopoietic, stromal, and engineered cell types for evaluating reprogramming outcomes, and tested an algorithmically-predicted transcription factor recipe—GATA2, GFI1B, FOS, REL, and STAT5A—for HSC induction. Long-read single-cell RNA sequencing of $CD34^+$ cells revealed heterogeneous reprogramming intermediates expressing fibroblast-, endothelial-, and hematopoietic-associated genes, consistent with partial reprogramming. We developed a quantitative benchmarking approach by mapping reprogrammed cells' transcriptomic positions to reference atlas cell types, revealing an endothelial-to-hematopoietic transition-like trajectory. Finally, we cataloged isoform diversity in reprogrammed cells, initial fibroblasts, and target HSCs, demonstrating transcriptional remodeling that is not apparent from gene-level analyses alone. This experimental-computational framework provides a generalizable strategy for characterizing and optimizing reprogramming protocols.

1 Introduction

The generation of hematopoietic stem cells (HSCs) through cell reprogramming would enable novel treatments for hematopoietic disorders. Current reprogramming strategies fall into two categories: direct reprogramming between somatic cell types, and differentiation of induced pluripotent stem cells (iPSCs) into cell types of interest. While the latter approach has achieved cells that have some functional properties of native HSCs[1], direct reprogramming offers notable advantages for clinical translation. An important consideration for clinical applications is the use of autologous initial cells, which can circumvent the immunological incompatibility and elevated risk of mortality associated with allogeneic transplantation[2]. Since direct reprogramming does not require a pluripotent intermediate, has fewer target cell states to achieve and validate, and has a lower risk of tumorigenesis, this approach may be more efficient for generating autologous cells from a patient sample[3–7]. However, direct fibroblast-to-HSC reprogramming is limited by low conversion rates and has not yet produced cells that have durable engraftment potential[8–15].

Assays that test cell function are the gold-standard for evaluating whether reprogramming cells have sufficiently adopted the target cell phenotype. While functional assays aid validation, they require cells to be fully reprogrammed, and therefore have limited utility until after an efficient direct reprogramming protocol is established. Better methods for evaluating reprogramming cells prior to this point would help identify successful and efficient paths of reprogramming. For example, cells that nearly achieve a target state might need only minimal intervention to fully reprogram, and methods that identify these cells would enable new protocols to be developed more quickly.

Single-cell transcriptomics has vastly improved the characterization of phenotype transitions such as differentiation in the hematopoietic lineage in mice[16–18] and humans[19–21]. Similar approaches have been used to examine heterogeneity[1] and identify intermediate or off-target states in reprogramming[22]. However, the extent to which reprogrammed cells and their gene expression dynamics compare to native cell types remains challenging to quantify. Comprehensive single-cell atlases can provide a powerful framework for placing incompletely understood or engineered cell types into the context of well-characterized phenotypes[23–25], but this approach has not yet been used for characterizing cell states in direct reprogramming to HSCs.

Here, we construct and evaluate the utility of a single-cell transcriptomic reference atlas spanning hematopoietic, endothelial, stromal, and engineered cell types for evaluating direct fibroblast-to-HSC reprogramming of human cells. Our approach includes testing a novel combination of transcription factors (TF) —GATA2, GFI1B, FOS, REL, and STAT5A—that were predicted by an algorithmic framework[26]. We profile the resulting $CD34^+$ cells using long-read single-cell RNA sequencing (scRNA-seq) and characterize several reprogramming intermediates within the context of our reference atlas. Lastly, we quantify gene- and isoform-level diversity in initial fibroblasts, reprogrammed cells, and native HSCs from bone marrow. Through this integrated experimental-computational approach, we provide a generalizable strategy that can be used to guide and optimize protocols for direct reprogramming.

2 Results

2.1 Construction of an atlas for hematopoietic cell reprogramming

Standard analysis of scRNA-seq data relies on cluster annotation, in which transcriptionally similar cells are grouped and assigned cell type identities based on marker gene expression for expected cell types[27,28]. This approach works well for native cell types with well-characterized transcriptional profiles, but is less effective in reprogramming contexts where induced cells may occupy poorly defined intermediate states that do not clearly match existing cell types[29]. Here, we sought to establish a reference framework that aids assessment of cell reprogramming outcomes, serving as a ground truth for assessing similarity, divergence, and partial identity across reprogrammed cell states.

We constructed a single-cell reference atlas drawn from 270 publicly available datasets covering over one million human hematopoietic, immune, and stromal cells (Supplementary Fig. 1). To mitigate technical variability, data were uniformly preprocessed and harmonized using single-cell variational inference (scVI) and conditioned on standardized cell type annotations provided by the source datasets (Methods 4.1)[30]. Source cell type annotations were consolidated into major cell compartments (e.g., fibroblast, endothelial, HSCs), which serve as the primary units of analysis throughout this study. To visualize gene expression patterns in the reference atlas, we generated a low-dimensional t-distributed stochastic neighbor embedding (t-SNE) from the harmonized latent space (Fig. 1A-B)[31].

To validate the integrated reference atlas, we evaluated lineage-specific gene expression and differentially expressed genes (DEGs) across the major cell compartments. Lineage-specific gene expression was consistent with annotations for fibroblasts and mesenchymal cells (*DCN*, *COL1A2*, *CALD1*), endothelial cells (*ACE*, *PECAM1*, *ENG*), hematopoietic stem and progenitor cells (*CD34*, *GATA2*, *CD164*), myeloid cells (*SPI1*, *CD14*, *ITGAM*), and lymphoid cells (*CD7*, *CD3*, *CD52*) (Fig. 1C-D and Supplementary Fig. 2), while lineage-specific marker transcripts were low or absent in HSCs (Fig. 1C, gray boxes). Likewise, DEGs between the major cell compartments were consistent with established marker sets and TFs (Fig. 1E-F, Supplementary Fig. 3, Supplementary Tables 1, 2)[28,32]. Hematopoietic markers (*MYB*, *LMO2*, *HOXA9*, *SPINK2*) and TFs (*HMGA1*, *MSI2*) were enriched in the hematopoietic compartment (Fig. 1C, Supplementary Figs. 3, 4, 5, 6), fibroblasts and mesenchymal cells expressed extracellular matrix genes (*LUM*, *FBLN1*, *COL6A1*) and mesenchymal regulators (*PRRX1*, *TWIST1*) (Fig. 1C, F and Supplementary Fig. 8), and the endothelial compartment expressed endothelial markers (*VWF*, *EMCN*, *ECSCR*) and TFs (*ETS1*, *EPAS1*, *SOX17*) and was enriched for transcripts related to angiogenesis, immune signaling, and vascular transport (Fig. 1C, E and Supplementary Figs. 3, 7).

Together, these data establish a comprehensive transcriptomic reference atlas for evaluating the outcomes of fibroblast-to-HSC direct reprogramming with respect to native cell types.

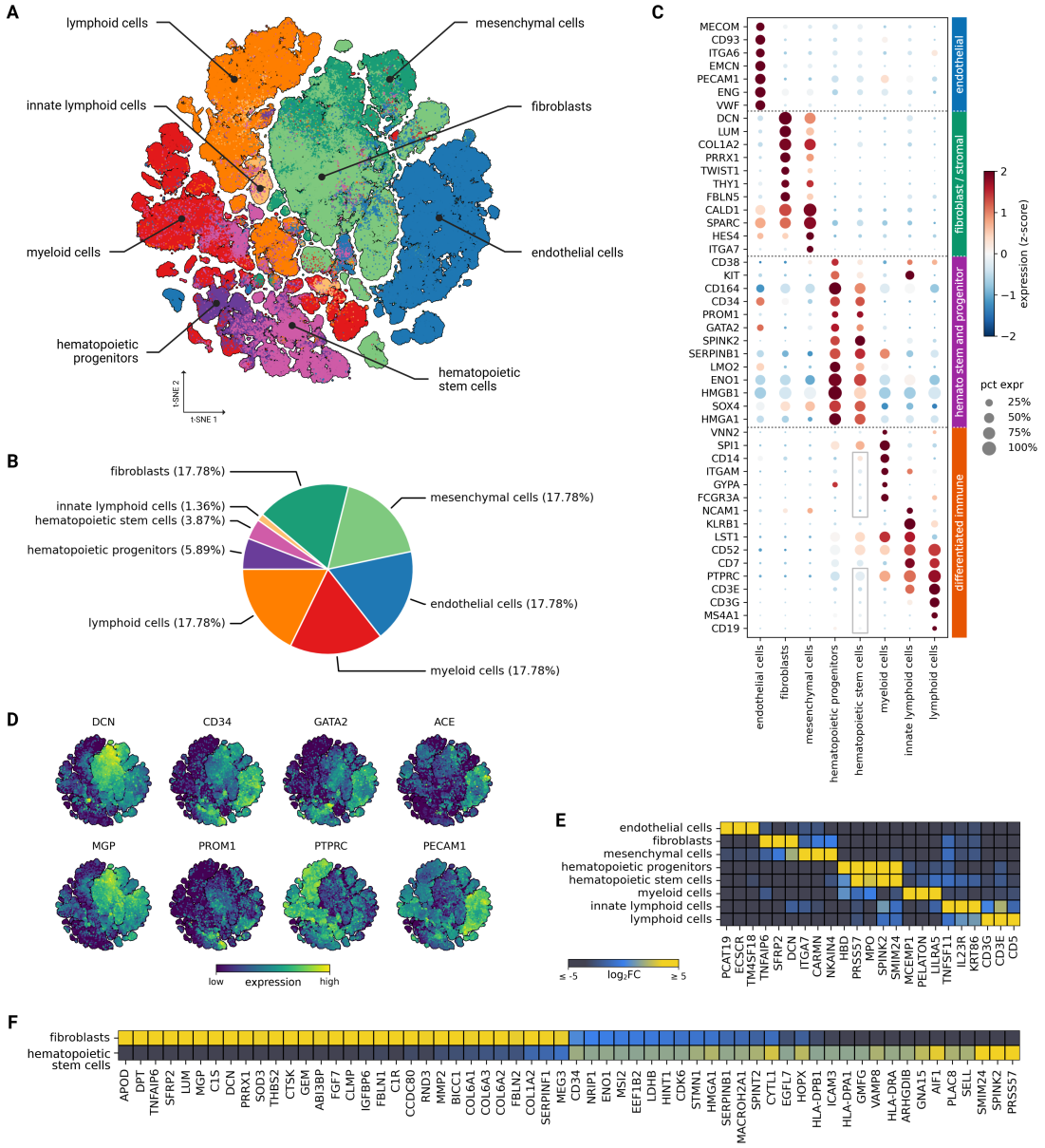

Figure 1: **A reference atlas for human hematopoietic cell reprogramming** **(A)** t-SNE embedding of scANVI-integrated cell types relevant to fibroblast-to-HSC reprogramming (1.12 million cells). **(B)** Proportions of major cell compartments in (A). **(C)** Average expression (z-score) of marker genes for endothelial, fibroblast/stromal, hematopoietic stem and progenitor, and differentiated immune cells in the major cell compartments in (A). Dot size represents the percentage of cells expressing a given gene. Gray boxes indicate genes encoding standard lineage-negative markers for HSCs. **(D)** Expression of fibroblast, hematopoietic, and endothelial marker genes projected onto the t-SNE coordinates from (A). **(E)** Top three DEGs per major cell compartment (Wilcoxon rank-sum test, Benjamini–Hochberg adjusted $p \leq 0.05$; detected in $\geq 15\%$ of cells per group). Color reflects log_2 fold changes relative to the other cell compartments. **(F)** Top 30 DEGs detected in $\geq 35\%$ of fibroblasts and hematopoietic stem cells.

2.2 A novel data-guided TF-recipe produces CD34$^+$ cells

Although a number of TF recipes have been used to reprogram human fibroblasts into HSC-like cells[13] (Supplementary Fig. 9), systematic approaches for identifying and evaluating candidate reprogramming factors remain limited. In prior work, we constructed the data-guided control (DGC) algorithm to predict candidate combinations of TFs for direct reprogramming[26]. Here, we leveraged the DGC algorithm to generate candidate TFs for fibroblast-to-HSC reprogramming (Methods 4.2.2), with the goal of generating a dataset to evaluate within our reference atlas. TFs previously reported to induce an HSC-like phenotype in human fibroblasts (3TF: GATA2, GFI1B, and FOS)[10] were among the top-ranked DGC candidates. Two additional top-ranked candidates, STAT5A and REL, were selected for inclusion in a 5TF protocol.

Human neonatal fibroblasts were transduced with inducible Tet-ON lentiviral vectors encoding the 5TF recipe (Fig. 2A, Methods 4.2.3, 4.2.4). On day 0, doxycycline (Dox) was added for 48 hours to induce TF expression (Methods 4.2.5). qRT-PCR analysis on day 3 confirmed upregulation of all five TFs in 5TF-induced cells compared to untransduced Dox-treated fibroblasts (unpaired t-test with Welch's correction, $p < 0.01$) (Fig. 2B, Methods 4.2.6, Supplementary Table 3). To quantify reprogramming efficiency, surface expression of CD34 was assessed by flow cytometry on day 7 using primary human bone marrow and uninduced 5TF cells as positive and negative controls, respectively (Methods 4.2.7, 4.2.9). Approximately 3% of 5TF-induced cells expressed CD34, a significant increase over negative controls (Mann-Whitney U test, $p = 0.0357$) (Fig. 2C). Immunofluorescence staining further validated CD34 protein expression in induced versus uninduced 5TF cells (Fig. 2D, Methods 4.2.8). Replicates of the 5TF treatment consistently produced CD34$^+$ cells, as did variations on these TF combinations (Supplementary Fig. 10). Expression of the hematopoietic marker CD143 was also detected, though at lower levels. Other common hematopoietic surface markers, including CD49f and CD9, were expressed by fibroblasts and thus were not used for further evaluation of the reprogramming population.

To further characterize 5TF-induced cells, we performed transcriptomic profiling using long-read scRNA-seq (Fig. 2E, Methods 4.3). We prepared two scRNA-seq libraries: Library 1 contained both CD34$^+$ 5TF-induced cells (reprogrammed; Supplementary Fig. 10D, Methods 4.2.7) and CD34$^-$ spike-in controls, and Library 2 contained untransduced fibroblasts. Single cells from both libraries were barcoded using the 10x Genomics Chromium platform and sequenced on Oxford Nanopore platforms. Raw reads were processed using the EPI2ME Labs scRNA-seq pipeline, and quality controlled data from both libraries was pooled, harmonized, and clustered, revealing 15,317 cells with three clusters of reprogrammed cells (R1-R3) and two clusters of fibroblast controls (C4, C5) (Fig. 2F-G, Methods 4.4.1). As expected, CD34$^-$ controls in Library 1 clustered together with fibroblasts from Library 2, indicating harmonization across libraries (Supplemental Fig. 11). Spike-in controls were excluded from downstream analyses after confirming their identity.

To assess transcriptional identity during reprogramming, we explored fibroblast and hematopoietic gene expression across both libraries (Fig. 2H-I). While fibroblast gene expression (e.g., *PRRX1*, *DCN*, *LUM*) was broadly reduced, residual fibroblast identity persisted in a subset of reprogrammed cells (Supplementary Figs. 11E, 12C). This was most evident in cluster R1, which retained elevated expression of multiple fibroblast-associated genes. In contrast, R2 and R3 exhibited a more pronounced loss of fibroblast markers, although expression of structural genes such as *COL1A1* and *S100A4* remained high, indicating incomplete silencing of the fibroblast program. Hematopoietic genes were partially induced in reprogrammed cells, including elevated expression of *CD34*, *ITGA6*, and *CD9*, and increased proportions of cells expressing genes such as *ETS2*, *MSI2*, and *G0S2* (Supplementary Fig. 12A). Interestingly, reprogrammed cells simultaneously showed increased expres-

sion of endothelial-associated genes, including *ACE*, *PECAM1*, and *PODXL*, consistent with previously reported intermediate cells reprogrammed with the 3TF recipe (Supplementary Fig. 12B)[10].

We next performed differential expression analysis to identify genes globally altered by reprogramming (Fig. 2J). Among the most upregulated genes in reprogrammed cells were *CD24* and *ZBTB16*, a TF implicated in balancing self-renewal and differentiation in HSCs[33], along with genes associated with membrane transport (*AQP3*, *SLC12A8*), cytoskeletal remodeling (*NEBL*), and metabolic regulation (*NSG1*, *CALB2*), consistent with a more plastic, reprogramming-permissive state. In contrast, interferon-stimulated genes (*MX1*, *MX2*, *IFIT1*) were reduced, potentially reflecting suppression of a steady-state innate immune program upon reprogramming[34]. At the cluster level, interferon-responsive genes and fibroblast markers were enriched in C4/C5 and broadly downregulated across R1-R3 (Fig. 2K). R1 and R2 upregulated remodeling-associated genes (*MGP*, *KRT18*, *CLU*), with R2 showing the strongest induction of progenitor-associated genes (*CD24*, *RBP1*). In contrast, R3 displayed a distinct proliferative and chromatin-remodeling signature characterized by upregulation of cell cycle genes and histone transcripts (*TOP2A*, *CIT*, *H1-3*).

Together, these results show that the 5TF recipe converts $CD34^-$ fibroblasts into transcriptionally distinct $CD34^+$ cells. The reprogrammed population co-expressed markers of fibroblast, endothelial, and hematopoietic programs, reflecting a heterogeneous, partially reprogrammed transcriptional state.

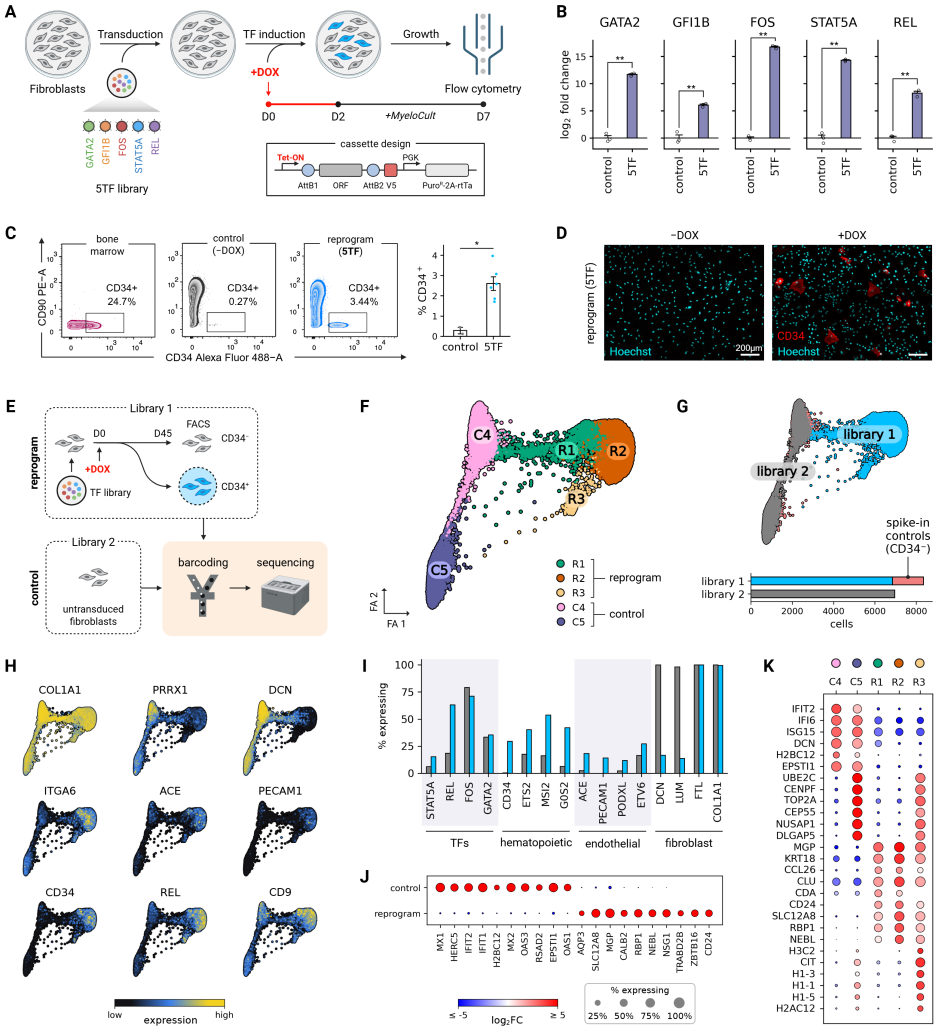

Figure 2: **Direct reprogramming of fibroblasts into CD34$^+$ cells. (A)** Overview of the reprogramming workflow and design of the inducible TF expression cassette. **(B)** 5TF expression as measured by qRT-PCR in control and 5TF-induced fibroblasts. Bars show mean ± SEM of $-\Delta\Delta$Ct, normalized to *LDLR* and expressed relative to control (unpaired *t*-test with Welch's correction; ** $p < 0.01$, $n = 3$ replicates). **(C)** Left: Flow cytometry plots for bone marrow, uninduced fibroblast controls, and 5TF-induced fibroblasts showing CD34. Right: CD34$^+$ cells as a percentage of live cells. Bars show mean ± SEM (Mann-Whitney U test; * $p < 0.05$, $n = 6$ replicates for 5TF, $n = 2$ replicates for control). **(D)** Immunofluorescent staining of CD34 in 5TF cells with or without Dox. Scale bar = 200μm. **(E)** Overview of scRNA-seq library preparation. Library 1 (reprogram): CD34$^+$ 5TF-induced cells with addition of CD34$^-$ spike-in control fibroblasts (not shown). Library 2 (control): untransduced fibroblasts. **(F)** Force-directed graph layout of single cells colored by Leiden cluster. **(G)** Graph from (F) colored by library. Cell counts shown below (Library 1: $n = 6,849$ 5TF cells, $n = 1,508$ spike-in controls; Library 2: $n = 6,956$ cells). **(H)** Normalized expression of select fibroblast, endothelial, and hematopoietic genes on the graph from (F). **(I)** Fraction of cells with nonzero expression for key TFs, hematopoietic, endothelial, and fibroblast genes in cells from both libraries. **(J)** Dot plot showing the top protein-coding DEGs between control (C4/C5) and reprogrammed (R1/R2/R3) cells. **(K)** Dot plot showing top protein-coding DEGs per cluster. In (J) and (K), dot color represents $\log_2$ fold change (Wilcoxon Rank-sum test, corrected $\alpha = 0.05$), and dot size reflects the percentage of cells expressing a gene in a given group.

2.3 Trajectory inference reveals emergence of early hematopoietic signature

The transcriptional heterogeneity observed in 5TF-induced cells suggests the presence of intermediate stages during fibroblast-to-HSC reprogramming, which remain poorly defined. These intermediate states can exist along novel reprogramming trajectories that differ substantially from known cell states[22,29,35]. Thus, to explore the dynamic progression from initial fibroblasts to 5TF-induced cells, we inferred a reprogramming trajectory using RNA velocity (Methods 4.4.4), which leverages the relative abundance of spliced and unspliced transcripts to estimate directional gene expression changes and position cells along a reprogramming continuum[36,37].

We established a velocity-derived pseudotemporal ordering for reprogrammed cells using control fibroblasts as a reference for the initial state (Fig. 3A-C, Methods 4.4.5). To mitigate cell cycle-driven artifacts, we selected control cells in both G1/S and G2/M phases and computed velocity pseudotime using the mean ordering derived from these cells (Fig. 3B-C, Supplementary Figs. 13, 14A-B, Supplementary Table 4). Diffusion-based pseudotime analysis[38] confirmed this ordering (Spearman's $\rho = 0.90$, Supplementary Fig. 14C-E). The resulting trajectory showed clear progression from control fibroblasts at early pseudotime (clusters C4/C5) to reprogrammed cells at late pseudotime (R2, R3), with cluster R1 spanning an intermediate range (Fig. 3C-D). Velocity vectors indicated flow from proliferating fibroblasts (C5) toward G1/G0-like states (C4), then outward into reprogrammed populations (Fig. 3E). Within reprogrammed cells, velocity vectors diverged into at least two distinct trajectories (R2 and R3), while R1 showed heterogeneous orientations, with some cells directed toward control states and others toward reprogrammed states.

Genes correlated with velocity pseudotime (Spearman's ρ, adjusted $p < 0.05$) revealed coordinated transcriptional programs (Fig. 3F). Early pseudotime genes included fibroblast markers (*DCN*, *LUM*, *TNC*) and interferon-stimulated genes (*STAT1*, *IFI6*, *ISG15*), consistent with diffusion pseudotime (Supplementary Fig. 14G) and findings from differential expression analysis. Late pseudotime genes included matrix remodeling factors (*MGP*, *CLU*, *FLNB*), cytoskeletal regulators (*KRT18*, *NEBL*, *S100A10*), and transport-related genes (*SLC12A8*, *RBP1*, *KCNMA1*). Positively correlated TFs (high in late pseudotime) included early hematopoietic regulators (*ZBTB16*, *ETS2*, *REL*) and developmental factors (*FOXN3*, *SOX13*, *SMAD6*), while negatively correlated TFs were enriched for innate immune responses (*IRF1*, *STAT2*, *SP100*) and fibroblast identity (*PRRX1*, *MKX*) (Supplemental Fig. 15), suggesting coordinated suppression of the fibroblast program and induction of early progenitor-like transcriptional regulators.

Cluster-specific velocity genes revealed distinct transcriptional drivers across reprogrammed populations (Fig. 3G). R1 was characterized by expression of matrix remodeling genes (*SPARC*, *CLU*, *MGP*), consistent with a transitional identity. R2 was enriched for endothelial (*HEG1*) and hematopoietic specification genes (*SMAD6*, *WNT9A*), indicating a potential endothelial-to-hematopoietic transition (EHT)-like state. R3 exhibited strong velocity signals for cell cycle genes (*CENPF*, *KIF4A*, *TPX2*), representing a distinct proliferative trajectory.

Expression dynamics along the reprogramming trajectory further supported EHT-like features (Fig. 3H). Fibroblast markers decreased over pseudotime while endothelial/EHT markers (*ACE*, *CD44*) and HSC markers (*CD34*, *ITGA6*) increased, peaking in R2. Genes involved in HSC maintenance and self-renewal (*MSI2*, *ZBTB16*) also increased along the trajectory. Notably, canonical EHT regulators (*RUNX1*, *GATA2*, *SMAD3*, *PECAM1*) and long-term HSC genes (*PROCR*, *MEIS1*, *IL3RA*, *EPAS1*) showed minimal change in steady-state expression but displayed positive velocity signals at late pseudotime, indicating transcriptional priming of the hematopoietic program (Sup-

plementary Fig. 16D-E).

These findings demonstrate that 5TF reprogramming drives directional exit from fibroblast identity through a heterogeneous intermediate state exhibiting early hematopoietic and endothelial features, with key HSC regulators transcriptionally primed but not yet fully expressed.

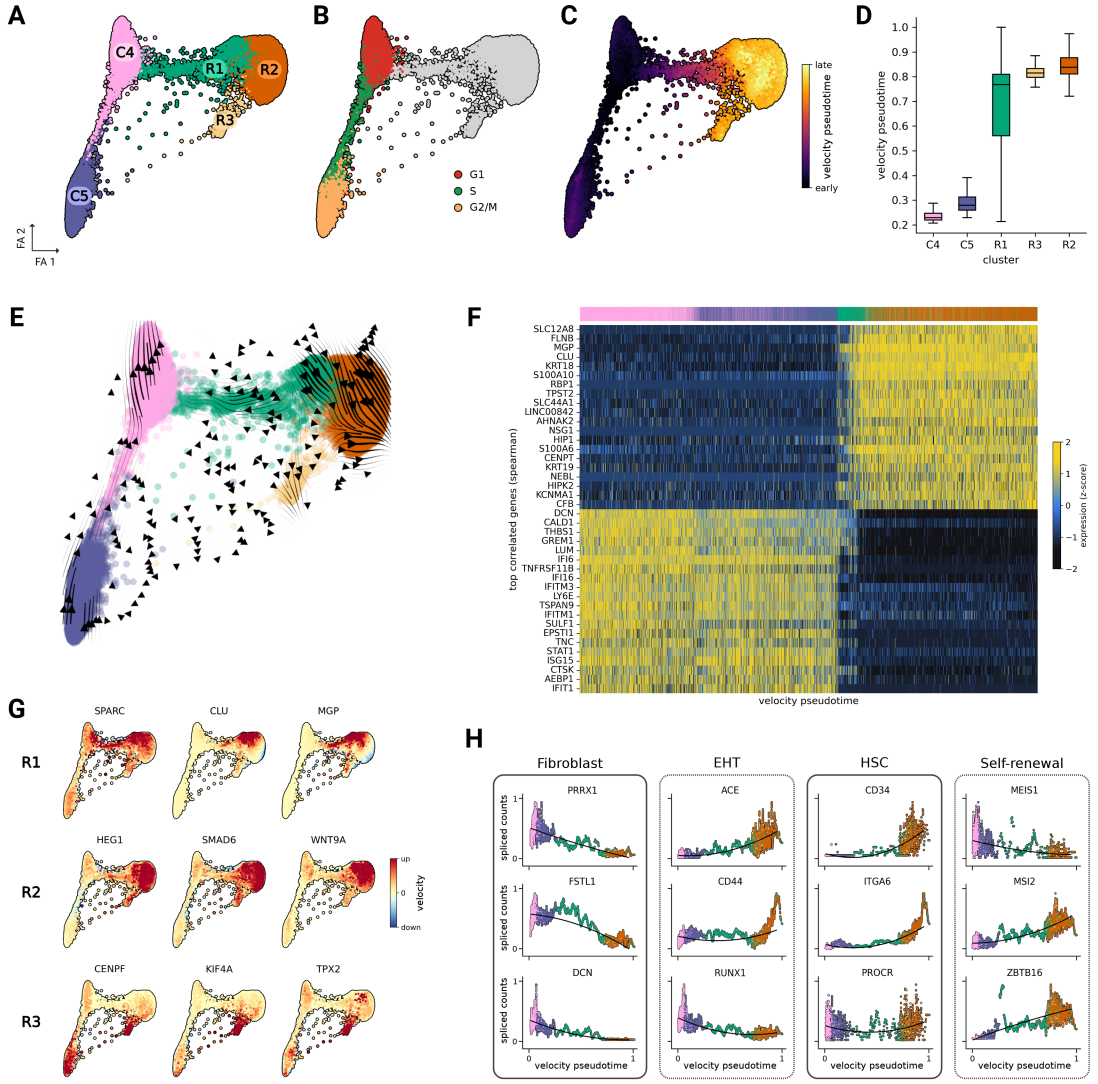

Figure 3: **Cell state transitions and gene expression dynamics from RNA velocity.** **(A)** Force-directed graph layout colored by Leiden cluster assignment. **(B)** Same layout with cells from Library 2 colored by cell cycle phase (G1, S, G2/M) used to select root cells for pseudotemporal ordering. **(C)** Cells colored by velocity pseudotime. **(D)** Per-cluster distributions of pseudotemporal ordering (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range). **(E)** Low-dimensional projection of RNA velocity estimates. **(F)** Z-scored expression of the top positively and top negatively pseudotime-correlated genes. Cells (columns) are ordered by velocity pseudotime and annotated by cluster (color bar, top). Correlations were computed using Spearman's ρ, and significance was assessed by Benjamini-Hochberg FDR correction ($\alpha = 0.05$). Only genes with adjusted $p < 0.05$ were considered. **(G)** Top velocity genes for reprogrammed cell clusters. **(H)** Spliced (steady-state) gene expression shown over velocity pseudotime for select fibroblast, EHT, and HSC genes colored by cluster.

2.4 Benchmarking against native and engineered cell types

Despite clear divergence from fibroblasts and acquisition of hematopoietic features, the degree to which reprogrammed cells resemble true HSCs remains incompletely characterized[10,39]. We therefore generated a third long-read scRNA-seq library (Library 3) of CD34$^+$ hematopoietic stem and progenitor cells (HSPCs) from human bone marrow (Fig. 4A; Methods 4.2.1, 4.2.10, 4.3). Cell types were annotated based on expression of canonical hematopoietic marker genes, identifying HSCs and lineage-restricted progenitors (Supplementary Fig. 17), consistent with published data[40].

We integrated our three long-read datasets using scVI to learn a harmonized latent space while controlling for technical bias and cell cycle effects (Fig. 4B, Methods 4.4.6)[30]. A uniform manifold approximation and projection (UMAP) embedding of this latent space showed graded expression of *CD34* and other hematopoietic genes across libraries (Supplementary Fig. 18A)[41], with the highest expression in bone marrow cells. Differential expression analysis revealed substantial downregulation of fibroblast programs in R1-R3, although some extracellular matrix and mesenchymal gene expression persisted in subsets of reprogrammed cells (Fig. 4C, Supplementary Fig. 18B). Several hematopoietic genes were upregulated in R1-R3 but remained lower than in native HSCs, supporting incomplete acquisition of the target state (Fig. 4C, Supplementary Fig. 18A).

To quantify transcriptional similarity, we computed per-cell Euclidean and cosine distances in the scVI latent space between R1-R3 and either initial fibroblasts or target HSCs (Fig. 4D). For each reprogrammed cell, distances were calculated to all cells in the reference population and averaged. We defined reference thresholds for "close" (mean HSC-to-progenitor distance) and "far" (mean HSC-to-fibroblast distance) relationships. R1 showed the greatest similarity to initial fibroblasts across both metrics, consistent with limited reprogramming progress. R2 and R3 showed reduced similarity to fibroblasts and increased alignment with HSCs. Notably, Euclidean distance—reflecting absolute separation in the latent space—suggested that R2/R3 remained closer to the initial state, whereas cosine distance—which captures similarity in the direction of latent representations—indicated greater similarity to HSCs. This suggests that R2/R3 have acquired hematopoietic-like transcriptional programs while retaining some aspects of their original fibroblast identity.

To contextualize reprogrammed cells within the broader hematopoietic landscape, we integrated our three libraries with cell types from the native reference atlas (fibroblasts, mesenchymal cells, endothelial cells, HSCs, hematopoietic progenitors), as well as engineered hematopoietic cell types from Ng *et al.*[1] and Gomes *et al.*[10] (Fig. 4E-G, Supplementary Fig. 19, Methods 4.4.7). Leiden clustering of scVI-harmonized data revealed 16 clusters (A1-A16) with substantial overlap between similar cell types across data sources (Fig. 4E, Supplementary Table 6). R2/R3 localized near engineered HSPCs (iHSPCs) while remaining partially proximal to both initial fibroblasts and subsets of native HSCs and hematopoietic progenitors (Fig. 4F-G, Supplementary Fig. 20).

We computed Euclidean distances between the R2/R3 centroid and individual cells within cluster-resolved cell types in the integrated atlas (Fig. 4H-I). R2/R3 were most similar to subsets of iHSPCs and native HSCs, but were also close to initial fibroblasts and native hematopoietic progenitors. A similar analysis of six independent human hematopoietic scRNA-seq datasets showed that reprogrammed cells were consistently closer to engineered hematopoietic and stromal populations than to native HSCs (Fig. 4I, Supplementary Fig. 21, Supplementary Table 5, Methods 4.4.8) and aligned most closely with day 25 3TF-reprogrammed cells[10]. Application of `Capybara`[22], a tool for aiding classification of hybrid cell identities, further supported a high degree of similarity between R1-R3 and engineered cell types (Fig. 4J, Supplementary Fig. 22, Methods 4.4.9).

Together, benchmarking analyses place reprogrammed cells in an intermediate state between

initial cells, endothelial cells, and HSCs, supporting an EHT-like trajectory that places them on a well-defined path to HSC reprogramming.

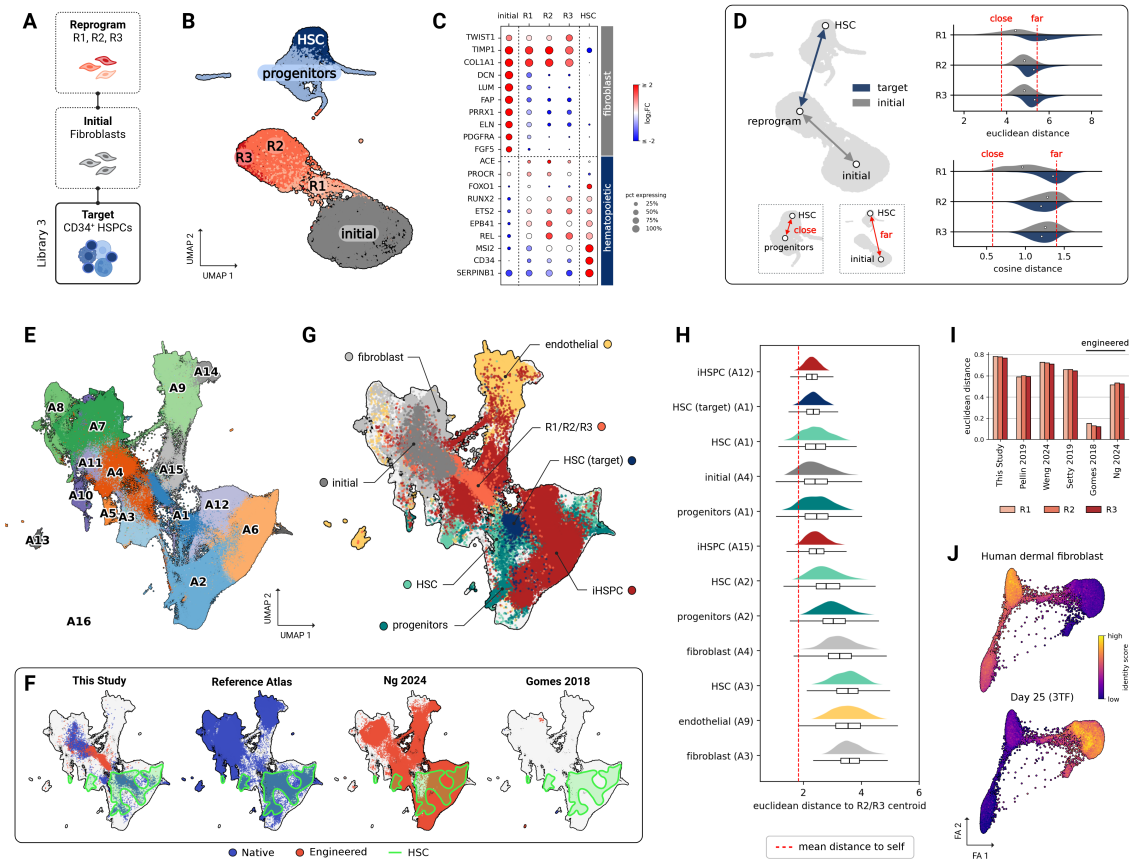

Figure 4: **Benchmarking 5TF-induced direct reprogramming.** **(A)** Long-read scRNA-seq datasets used for benchmarking, including data from human CD34$^+$ bone marrow cells (Library 3; $n = 5,458$). **(B)** scVI-corrected UMAP embedding of cells from (A). **(C)** Select marker genes. Dot color represents $\log_2$ fold change (Wilcoxon Rank-sum test, corrected $\alpha = 0.05$), and dot size reflects the percentage of cells expressing a gene in a given group. **(D)** Left: schematic of cell–cell distance comparisons; Right: distributions of mean pairwise Euclidean and cosine distances from R1-R3 to the target and initial cell types. White dots indicate the mean distance. Red dashed lines represent the mean distance between close (HSC to progenitors) and far (HSC to initial) cells. **(E)** UMAP of 420,108 cells from the reference atlas and selected sources based on scVI latent space, colored by Leiden cluster. **(F)** Source of cells in (E), colored by native or engineered cell types. The green outline represents HSCs from the native reference atlas. **(G)** Cell groupings of interest from (E). **(H)** Distributions of per-cell Euclidean distances to the centroid of R2/R3 per cluster-resolved cell annotation, sorted by mean distance (box plots: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range). Red dashed line represents the mean distance of R2/R3 cells to the R2/R3 centroid. **(I)** Normalized euclidean distances from pseudobulked profiles of R1-R3 to native and engineered HSCs from a variety of hematopoietic scRNA-seq datasets. **(J)** Capybara[22] identity scores for human dermal fibroblasts and Day 25 3TF-reprogrammed cells from Gomes *et al.*[10] projected onto the force-directed graph layout from Fig. 2F.

2.5 Transcript-level diversity in direct reprogramming

While our reference atlas enabled a quantitative comparison of reprogrammed cells to a variety of native and engineered cell types, these analyses relied primarily on gene-level expression, which incompletely reflects molecular and phenotypic changes[42,43]. Commonly used functional assays for evaluating reprogrammed HSCs, such as transplantation or colony-forming assays, lack the resolution needed to quantitatively define intermediate cell states or inform iterative optimization of reprogramming protocols[39]. Protein-level measurements can capture key aspects of cell identity but remain limited in throughput and scalability[44]. Thus, existing approaches provide only a partial view of the molecular changes accompanying reprogramming. Alternatively spliced (AS) transcript isoforms represent an additional regulatory layer that is more directly linked to protein structure, regulation, and post-transcriptional control, and are incompletely captured by short-read sequencing[45,46]. Leveraging our long-read scRNA-seq data, we sought to examine transcript-level changes during reprogramming (Methods 4.4.10), hypothesizing that isoform-level analyses could reveal features of cell state obscured at the gene level.

We first characterized global patterns of AS across all three libraries. As expected for mammalian cells, exon skipping was the most common AS event observed in each condition (Fig. 5A, Methods 4.4.11)[47,48]. To assess transcript-level diversity, we quantified the number of detected transcripts per gene and classified genes as not expressed (NE), expressed as a single isoform, or expressed as multiple isoforms. Between conditions, genes were scored as showing a gain in isoform diversity (NE to single or multiple, single to multiple) or a loss in diversity (multiple to single or NE, or single to NE). A visual representation of this analysis is shown in Figure 5B. Among reprogrammed populations, cluster R2 showed the largest gain in isoform diversity relative to initial fibroblasts, and the greatest overlap with hematopoietic progenitors and HSCs. In contrast, R3 exhibited the greatest loss of isoform diversity and the least overlap with both initial and target states, consistent with more divergent or heterogeneous transcript usage.

To further assess transcript-level differences, we identified protein-coding genes whose expression was dominated by a single transcript, designating this transcript as the 'major isoform'. HSCs exhibited the largest number of unique major isoforms, consistent with a distinct transcriptomic architecture (Fig. 5C). Among reprogrammed clusters, R2 shared the greatest number of major isoforms with HSCs, whereas R3 showed the least overlap with the initial fibroblast state. R2/R3 displayed the highest numbers of unique major isoforms among reprogrammed populations, reflecting progressive transcript-level remodeling associated with later reprogramming states. Similar patterns were observed when examining major isoform overlap across curated gene sets, including fibroblast-, endothelial-, HSC-, and stem cell-associated genes (Fig. 5D). For all gene sets, reprogrammed cells shared the greatest proportion of major isoforms with initial fibroblasts, while exhibiting increased overlap with HSCs relative to initial cells. Across gene sets, all cell populations retained moderate to high fractions of unique major isoforms, highlighting the prevalence of transcript-level differences across initial, reprogrammed, and target cell states.

We next identified genes undergoing isoform switches (IS) across cell types (Fig. 5E), defined as changes in the dominant protein-coding transcript. Several genes exhibited consistent IS between reprogrammed cells and the initial fibroblast state, with R1-R3 preferentially expressing transcript isoforms that encode distinct protein products. For example, initial fibroblasts predominantly expressed the Ensembl canonical isoform *TEAD4-202* (UniProt: Q53GI4), whereas R1-R3 primarily expressed *TEAD4-207* (UniProt: H0YGS2), which lacks annotated binding domains present in the canonical transcript. This example illustrates how IS can introduce qualitative changes in predicted protein structure that are not captured by gene-level expression analyses.

To quantify transcript-level regulation independent of gene-level differential expression, we identified differentially expressed isoforms (DEIs) in clusters R1-R3 relative to initial fibroblasts and assessed whether the corresponding genes were differentially expressed at the gene level (Fig. 5F). Across all reprogrammed clusters, approximately 30% of DEIs occurred in genes without detectable gene-level differential expression, suggesting that a substantial fraction of transcriptional remodeling would be missed by gene-centric analyses. Visualization of representative DEIs from genes lacking gene-level differential expression revealed pronounced changes in transcript usage across conditions (Fig. 5G). Notably, several DEIs exhibited changes in both coding sequences and untranslated regions, which regulate mRNA stability, localization, and translation efficiency[49].

Together, these findings demonstrate that isoform-level analyses reveal substantial transcriptional remodeling during reprogramming that is not apparent from gene-level expression alone, and underscore the value of long-read transcriptomics for resolving molecular features of reprogramming cell states.

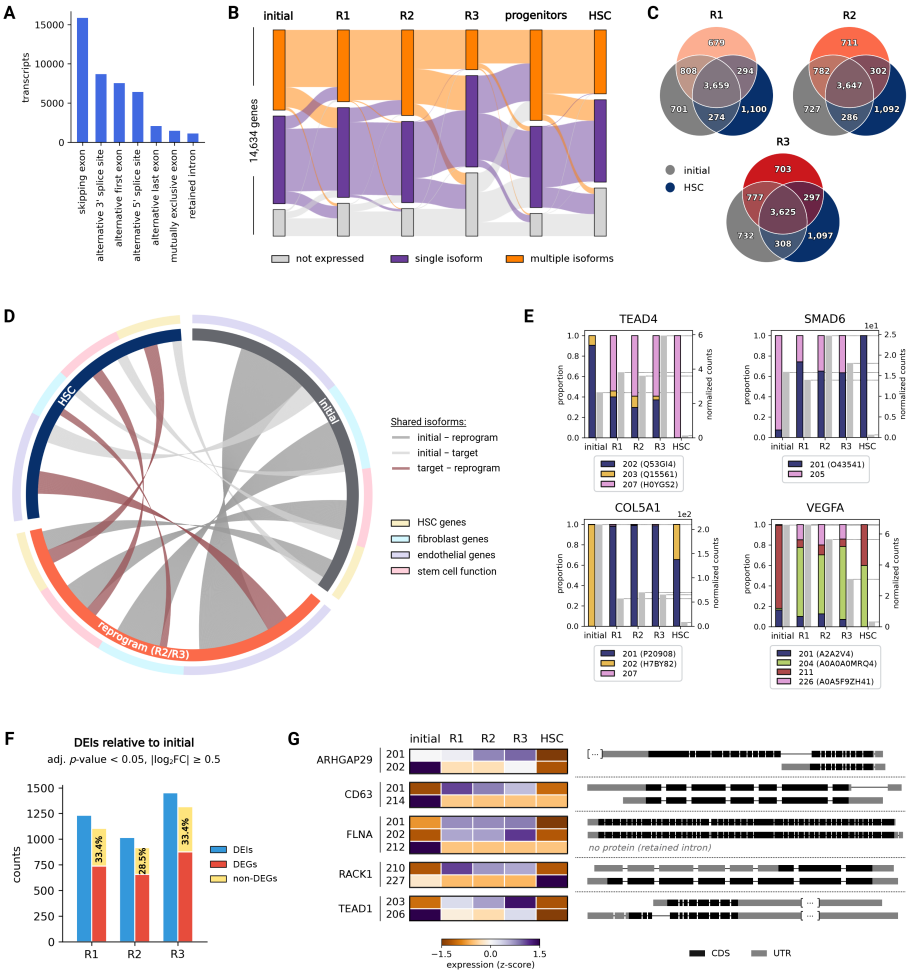

Figure 5: **Isoform diversity in fibroblast to HSC reprogramming. (A)** Count of pooled transcripts by splicing event. **(B)** Sankey diagram showing changes in isoform usage for protein-coding genes across reprogramming stages. Transcripts with less than 10 counts per cell grouping were excluded. **(C)** Overlap of shared major isoforms between reprogrammed clusters and initial or target cell types. **(D)** Shared major isoforms across cell groupings for curated gene sets from GO Biological Process[50,51] and marker gene databases[28,32,52]. Links between transcripts signify shared major isoforms between cell groups. Unlinked transcripts represent unique major isoforms. **(E)** Isoform switching events. Bar plots show the proportion of total gene expression for each detected transcript, with gray bars signifying total normalized counts per gene (counts per million, CPM). Transcripts are labeled with corresponding UniProt accessions (Methods 4.4.10). **(F)** Left bars: Number of DEIs for each reprogrammed cluster relative to initial cells. Right bars: Proportion of corresponding genes that are or are not differentially expressed at the gene-level. (Wilcoxon Rank-sum test, corrected $\alpha = 0.05$, $|\log_2 \text{ fold change}| > 0.5$, expressed in $> 30\%$ of group where $\log_2$ fold change > 0) **(G)** Left: Z-scored expression (CPM) for representative transcripts where the gene is not differentially expressed at the gene-level. Transcripts were filtered for those accounting for $> 10\%$ of total counts per gene. Right: Exon structures for corresponding transcripts with coding sequences (CDS) and untranslated regions (UTR) annotated according to the GRCh38 reference genome (10x Genomics, 2024-A).

3 Discussion

Our study clarifies both the capabilities and current limitations of TF-driven direct reprogramming protocols for generating autologous HSCs. Our prior DGC algorithm enabled prediction of a novel TF recipe for fibroblast-to-HSC reprogramming, while our reference atlas aided evaluation of reprogrammed cells, revealing acquisition of an early hematopoietic- or endothelial-like program and previously uncharacterized patterns in gene and isoform expression. Reprogrammed cells showed transcriptomic shifts that distinguished them from fibroblasts, but they remained distinct from native HSCs.

Notably, the TFs used for direct cell reprogramming in this study were predicted using a data-guided algorithm[26]. Among the predicted TFs, GATA2, GFI1B, and FOS are key hematopoietic regulators that have been used in previous studies for direct reprogramming to HSCs[10,11,53], supporting that algorithm-predicted TFs are a practical starting point for recipe discovery. Of the novel predicted factors, STAT5A is a hematopoietic regulator that supports creation of HSCs from murine embryonic stem (ES) cells and has a role in HSC self-renewal, proliferation, and survival[54–56]. REL encodes a subunit of the NF-κB complex, which has an important role in the emergence of HSCs during embryonic development[57]. Proinflammatory signaling via NF-κB is involved in HSC specification alongside Notch signaling in the hemogenic endothelium[58,59]. REL also regulates apoptosis by promoting expression of anti-apoptotic genes *BCL* and *BCL2L1*[60–62]. Together, STAT5A and REL are shown to regulate chromatin accessibility at shared shared target loci[63], supporting a model in which they act as complementary regulators in the hematopoietic gene expression program. However, the context-dependent role of these TFs in reprogramming requires further elucidation.

Prior studies support a model in which fibroblast-to-HSC reprogramming proceeds through an endothelial precursor, exhibiting core aspects of developmental EHT[12,64]. For example, murine fibroblast-to-HSC reprogramming studies report endothelial-like and hemogenic endothelial intermediates during early reprogramming stages, including expression of embryonic hemoglobin (*b-H1*), *Pecam1*, and *Vwf*[65,66]. Human fibroblast-to-HSC studies likewise show expression of endothelial genes such as *PPARG*, *VWF*, and *FOXC2*[10,11,67], and similarly, our reprogrammed cells expressed endothelial genes such as *ACE*, *PECAM1*, and *PODXL*. Consistent with this endothelial intermediate model, gene expression dynamics along the 5TF reprogramming trajectory revealed transcriptional priming of key EHT regulators, including *RUNX1* and *GATA2*.

Notably, within our integrated reference atlas, 5TF cells showed substantial overlap with subsets of iPSC-derived iHSPCs from Ng *et al.*, a system in which multilineage bone marrow engraftment has been demonstrated in immune-deficient mice[1]. The transcriptional proximity of our cells to an engineered hematopoietic system with demonstrated in vivo potential supports that 5TF cells are progressing along a trajectory compatible with HSC specification. While Ng *et al.* cells generate a broad spectrum of hematopoietic cell types, our 5TF cells appear to follow a more constrained, direct trajectory from fibroblasts toward HSCs, forming a condensed population positioned between endothelial cells, HSCs, and fibroblasts. Given that iPSC-based approaches recapitulate aspects of developmental EHT, our findings raise the possibility that direct reprogramming may similarly engage key features of the developmental program but through a more condensed or accelerated route. Future studies will be required to determine whether direct fibroblast-to-HSC reprogramming follows a path similar to iPSC-based hematopoiesis, and whether this potentially accelerated path can likewise produce functionally equivalent HSCs.

Long-read single-cell profiling revealed isoform-level changes during direct reprogramming that are missed at the gene level. Across our reprogrammed cells, hundreds of transcripts showed differ-

ential isoform usage without corresponding changes in total gene expression. These isoform-level differences may have important functional consequences: distinct transcripts from the same gene can encode proteins with different domain architectures or contain alternative untranslated regions that influence post-transcriptional regulation[48,49]. Such mechanisms are particularly relevant in reprogramming, where subtle changes in TF activity or signaling pathways can alter how cells interpret cues and respond to fate-determining signals. Indeed, prior work has shown that reprogramming to pluripotency involves distinct, sequential phases of splicing regulation[68–71] and that isoform switching in TFs can influence reprogramming outcomes[72,73]. These observations raise the possibility that TF recipes shape reprogramming not only by activating target genes, but also by influencing transcript processing and isoform usage. Reprogramming protocols may therefore benefit from considering how TF combinations influence transcript processing. Importantly, the functional impact of individual isoforms cannot be inferred from transcriptomic data alone, and targeted experimental validation will be required to determine how specific isoforms contribute to reprogramming efficiency or the stability of a specific cell state.

Despite acquisition of $CD34^+$ surface expression and marked transcriptional changes, the reprogrammed cells retained expression of key fibroblast markers such as *COL1A1* and *FTL*[32]. Complete reprogramming may require overcoming barriers in cell state transitions, including both local stability and maintenance of the new identity[74–76]. Quantifying intermediate or hybrid cell states remains a key challenge[22,35]. Computational tools such as `Capybara`[22] enable partial assignment of transcriptomic profiles to single cells, supporting estimation of hybrid identities; however, the accuracy of this approach is limited by the availability of suitable reference data. Our approach parallels other recent atlasing strategies for identifying rare, diseased, or engineered cell states and evaluating new engineering protocols[24,77,78]. Future studies may assess the applicability of emerging single-cell foundation models[23,25,79,80], though their capacity to represent engineered and intermediate cell types remains to be determined.

Contextualization of reprogrammed cell states provides critical feedback for refining reprogramming protocols. Systematic characterization of similarities and differences between reprogrammed and native cell types can reveal barriers to efficient conversion and identify opportunities to overcome them. Beyond the present system, our approach is broadly applicable to other contexts where engineered or perturbed cells occupy intermediate states, including regenerative medicine, disease modeling, and synthetic biology. In summary, our approach offers a generalizable framework for evaluating unknown cell identities and contributes to the effort to improve direct reprogramming methods for generating autologous cells.

4 Methods

4.1 Reference Atlas Construction

Single cell transcriptomic data were retrieved from the CellxGene Census (2024-07-01) using the `cellxgene-census` Python API. We selected 1,125,041 primary cells (52,164 genes) from non-diseased human samples representing major hematopoietic, myeloid, lymphoid, innate lymphoid, fibroblast, mesenchymal, and endothelial cell types. Standardized source data metadata, such as cell type and tissue annotations, were stored along with gene expression data for downstream processing and analysis using `Scanpy`[81] (version 1.11.4) and/or `rapids-singlecell`[82] (version 0.12.3). Consistent quality control was applied across all cells. First, data were filtered for cells with at least 200 UMI reads and genes with a minimum of 50 reads. Expression was normalized using `normalize_total` with `target_sum=`10^4, followed by log-transformation using `log1p`.

We performed batch harmonization and biologically meaningful latent space extraction in two steps. First, we used scVI[30] implemented in the `scvi-tools` (version 1.3.3) Python library[83]. We developed a `scVI` model with two hidden layers and a 24-dimensional latent space, a negative binomial likelihood based on individual dataset set identifiers and cell type annotations. Training was performed on GPU (NVIDIA Tesla V100 or A100) for up to 100 epochs with early stopping (patience = 5), using a batch size of 1000 and monitoring validation loss. Next, we initialized a semi-supervised `scANVI` from the pretrained `scVI` model to improve cell type influence on the learned latent space[84]. The `scANVI` model was trained for up to 20 epochs using GPU acceleration, with early stopping (patience = 5) based on validation loss. Training used a batch size of 1,000 and 100 samples per label. Training was performed on the same GPU hardware while monitoring validation loss. For visualization, we developed a t-SNE embedding based on the `scANVI` model latent space.

To identify DEGs, we used `scanpy.tl.rank_genes_groups` using the Wilcoxon rank-sum test (two-sided) over log-normalized expression. DEGs were identified for major cell compartments and cell types, excluding cell types with fewer than 500 cells.

4.2 5TF Reprogramming

4.2.1 Initial and Target Cell Sources

Human neonatal fibroblasts were used as the initial cell type for reprogramming experiments. Cryopreserved BJ fibroblasts were purchased from ATCC (CRL-2522, RRID: CVCL_3653), thawed, and expanded in standard growth medium (GM): Dulbecco's Modified Eagle Medium - high glucose (DMEM; Gibco, 11965092) with 1% L-glutamine and 10% fetal bovine serum (FBS; Corning 35-015-CV). Cells were cryopreserved at passage 2 or 4 and freshly thawed for each experiment. 1x TrypLE^{TM} Express Enzyme (Gibco #12604-021) was used for dissociation and cell passaging. To represent the target cell type in single-cell transcriptomics experiments, primary CD34$^+$ hematopoietic stem and progenitor cells (HSPCs) from adult human bone marrow were purchased from STEMCELL Technologies (Cat#70002), containing 1×10^6 cells. CD34$^+$ HSPCs were thawed in pre-warmed GM and prepared according to the manufacturer's protocol.

4.2.2 Algorithmically Guided TF Recipes

TFs for fibroblast-to-HSC reprogramming were selected using the previously constructed "data-guided control" (DGC) algorithm[26]. The DGC algorithm provided a ranked list of TFs that partially overlapped with a prior published TF recipe (GATA2, GFI1B, FOS), and STAT5A and REL were selected since they were among the top five TFs by rank. Briefly, in this framework, TFs are identified by minimizing the distance between the initial and target cell states according to the following linear difference equation:

$$\mathbf{x}_{k+1} = \mathbf{A}_k \mathbf{x}_k + \mathbf{B}\mathbf{u}_k$$

where:

- $\mathbf{x}_k \in \mathbb{R}^N$: gene expression vector at time k
- $\mathbf{A}_k \in \mathbb{R}^{N \times N}$: transition matrix representing time-varying changes in gene expression
- $\mathbf{B} \in \mathbb{R}^{N \times M}$: binary input matrix capturing interactions between genes (N) and TFs (M)
- $\mathbf{u}_k \in \mathbb{R}^M$: input vector indicating which TF(s) are introduced at time k

TFs are ranked through the following optimization:

$$\text{minimize}_{\mathbf{u}} \, \|\mathbf{x}_T - \mathbf{z}_F(\mathbf{u})\|$$

The objective is to minimize the Euclidean norm, $\|\cdot\|$, of the difference between the target state $\mathbf{x}_T$ and the final state $\mathbf{z}_F(\mathbf{u})$, where $\mathbf{z}_F$ depends on the initial condition $\mathbf{x_0}$ and the input signal $\mathbf{u}$.

4.2.3 Lentiviral Vector Design

Human GATA2, GFI1B, FOS, STAT5A, and REL plasmids were procured from the DNASU repository, subcloned into pLIX-403 Tet ON lentiviral backbones (RRID:Addgene_41395), and verified by Sanger sequencing. For TF combinations labeled FH3 (GATA2, GFI1B, FOS) and FH6 (GATA2, GFI1B, FOS, STAT5A, REL), lentiviral vectors were prepared individually for each plasmid, then pooled for transductions (Table 1). For recipe FH4a (GATA2, GFI1B, FOS, STAT5A, REL), plasmids were pooled in equimolar amounts followed by lentiviral packaging. All-Prep DNA/RNA/Protein Mini Kits (Qiagen 80004) were used to process cells for testing genome integration and protein and gene expression.

4.2.4 Lentiviral Transduction

BJ fibroblasts were plated at a density of 21,000 cells/cm^2 in either six-well plates or 10 *cm* dishes. The next day, cells were transduced with one of three TF recipes in Table 1. Transductions were performed in standard GM containing 8 μg/mL polybrene (Millipore Sigma TR-1003-G). GM supplemented with 1% penicillin/streptomycin (P/S; Gibco 15140-122) was exchanged 24 hours post-transduction. Recipes FH3 and FH6 were used for flow cytometry experiments, while FH4a was used for transductions in all other experiments.

FH3	FH4a	FH6
GATA2 GFI1B FOS	GATA2 GFI1B FOS REL STAT5A	GATA2 GFI1B FOS REL STAT5A
Vector preparation		
individual lenti per TF	single lenti	individual lenti per TF

Table 1: **TF recipes and lentiviral vector preparation methods.**

4.2.5 Reprogramming Induction

All cells were cultured at 37°C with 5% CO_2. 48 hours after transduction, TF expression was induced via addition of 1 μg/mL doxycycline hyclate (Dox; Millipore Sigma, D9891). We consider this Day 0 of reprogramming. On Day 2, cells were plated in six-well plates coated with 0.1% porcine gelatin (Millipore, #SF008) in complete MyeloCult H5100 (STEMCELL Technologies, 05150) + 10e-6 M hydrocortisone (Millipore Sigma, #H0888) + 1% P/S. Media was exchanged every other day for the duration of experiments. Reprogramming cultures were grown for 7-48 days, with cell passaging occurring every 6 days. Cells used for control and reprogramming cultures were used at initial passage number between 2 and 4, and were at 10 or fewer passages for endpoints.

4.2.6 qRT-PCR Validation of 5TF Expression

Total RNA was collected from independent replicates ($n = 3$) of fibroblasts transduced with FH4a (Table 1) using the RNeasy Kit (Qiagen 74104) after 3 days of 1 μg/mL Dox treament in complete Myelocult media. The yield and purity of RNA samples were assessed by Nanodrop and Qubit (Thermo Fisher Scientific). Using 1 μg of purified RNA per sample, first-strand cDNA synthesis was performed using SuperScript IV Reverse Transcriptase (Invitrogen 18090010) with random hexamers (Invitrogen N8080127). Quantitative real-time PCR (qRT-PCR) was conducted using SYBR Green qPCR Master Mix (Thermo Fisher Scientific A66732) for all samples in triplicate in a QuantStudio 5 Real-Time PCR System (Applied Biosystems). Relative quantification was determined using the $\Delta\Delta$Ct method and normalized to the endogenous control LDLR. Primer efficiency was verified for each exogenous TF prior to running qRT-PCR. See Table 2 for a list of primers. Statistical significance was determined in GraphPad Prism version 10.4.1, where quantile-quantile plots were used to assess data normality, and group differences were evaluated using an unpaired t-test with Welch's correction (two-sided). Statistical details are provided in Supplementary Table 3.

4.2.7 Flow Cytometry

Antibodies used for flow and/or cell sorting included anti-human CD90 PE (1:250, BioLegend Cat# 328109, RRID: AB_893442), CD34 Alexa Fluor 488 (1:250, BioLegend Cat# 343517, RRID: AB_1937204) or CD34 PE (1:250, BioLegend Cat# 343605, RRID:AB_1732033), and CD143/ACE (1:250, BioLegend Cat# 344207, RRID: AB_2783230). LIVE/DEAD® Fixable Yellow Dead Cell Stain Kit (0.8 μL per test, Invitrogen L34968) and DAPI were used as viability markers. Cells were incubated with primary antibodies or LIVE/DEAD stain for 30 minutes at 4°C in Hank's balanced

Primers	
GATA2-V5 Fwd	TTCGAGGAGCTGTCAAAGTG
GFI1B-V5 Fwd	CCAGAAGTCCGACATGAAGAAG
FOS-V5 Fwd	GGACCTATCTGGGTCCTTCTA
STAT5A-V5 Fwd	CAAGTGGTCCCTGAGTTTGT
REL-V5 Fwd	TCCATGCCATCAGCAGATTTA
Common-V5 Rev	CTAGGAGTGGGTTTGGGATTG
LDLR Fwd	ACCGACTCTGTCCTGGGCACTG
LDLR Rev	AGTTCCCCAGTCAGTCCAGTA

Table 2: **Primers used for qRT-PCR validation of 5TF expression.**

salt solution or phosphate buffered saline (PBS). UltraComp eBeads (Invitrogen Cat# 01-2222) and ArC Amine Reactive Beads (Invitrogen Cat# A10628) were used for single color compensation controls. After gating for singlets and live cells, fluorescence-minus-one (FMO) controls were used to aid gating for reprogrammed cells. As additional staining controls, human primary bone marrow $CD34^+$ cells were purchased from ATCC (PCS-800-012TM, Lot 80516705) and cultured in complete MyeloCult H5100 (STEMCELL Technologies, 05150) for 12-16 hours prior to analysis, and Veri-Cells CD34 PBMCs (Biolegend, Cat# 426901 Lot B375717) were reconstituted following the manufacturer's protocol. Flow cytometry data were collected on a Bio-Rad ZE5 Cell Analyzer and cells were sorted using a MoFlo Astrios or a Bigfoot Spectral Cell Sorter. FlowJo version 10.10.0 (Becton Dickinson) was used for flow cytometry analyses. For % $CD34^+$ comparisons between independent replicates of 5TF-induced cells ($n = 6$) and fibroblast controls ($n = 2$), statistical significance was determined using `SciPy` (version 1.16.1) `scipy.stats.mannwhitneyu` (one-sided).

4.2.8 Immunofluorescence

On Day 26 of reprogramming, cells were fixed with 4% paraformaldehyde (PFA) for 10 minutes, washed with cold PBS, and permeabilized for 10 minutes in PBS containing either 0.25% Triton X-100, 100 µM digitonin, or 0.5% saponin. After washing with PBS, blocking was performed using 1% bovine serum albumin (BSA) and 22.52 mg/mL glycine in PBS with 0.1% Tween 20 for 30 minutes. Cells were then incubated overnight at 4°C in primary antibody solution containing anti-human CD34 PE (1:200, Biolegend, Clone 581, Cat #343505, Lot B373336) and 1% BSA/PBST. Cells were washed with PBS and counterstained with 2 µM Hoechst 33342 (Cayman Chemical Cat# 15547) for 10 minutes. Imaging was performed using a Zeiss LSM 710 confocal microscope.

4.2.9 Immunolabeling and Cell Sorting of $CD34^+$ HSPCs

DAPI was used as a cell viability marker. Veri-Cells CD34 PBMCs (BioLegend) were reconstituted following the manufacturer's instructions and served as staining controls. Approximately 3×10^5 Veri-Cells were used for full staining and FMO controls. UltraComp eBeads (Invitrogen Cat#01-2222) were used for single-color compensation. Approximately 1×10^6 bone marrow $CD34^+$ HSPCs were prepared for sorting. Cells were first incubated with Fc receptor block (BioLegend, Cat#422301) for 10 minutes on ice to minimize nonspecific antibody binding. Subsequently, cells were incubated with conjugated antibodies for 30 minutes on ice in the dark, washed with 2 mL of cell staining buffer (BioLegend Cat#422301), pelleted, and resuspended in fresh buffer

(BioLegend Cat#420201). Antibodies included APC anti-human CD34 (Clone 581, BioLegend Cat#343509, RRID: AB_1877154), PE anti-human CD38 (Clone HIT2, BioLegend Cat#303505, RRID: AB_314357), and FITC-conjugated human Lineage Cocktail (Clone UCHT1; HCD14; 3G8; HIB19; 2H7; HCD56, BioLegend Cat#348801, RRID: AB_10612570), which labels T cells, monocytes/macrophages, NK cells, B cells, neutrophils, and eosinophils. Dead cells and FITC-positive cells were excluded during gating. $CD34^+/CD38^-$ and $CD34^+/CD38^+$ cell populations were sorted on a Cytek Aurora-1 instrument.

4.2.10 Cell Hashing of $CD34^+$ HSPCs

Sorted cells were re-blocked with Fc receptor block for 10 minutes on ice and incubated with mixtures of hashtag antibodies for 30 minutes at 4°C. Prior to use, antibody mixtures were centrifuged at 14,000 g for 10 minutes at 4°C to remove protein aggregates. $CD34^+/CD38^-$ cells were labeled with TotalSeq-B0251 while $CD34^+/CD38^+$ cells were labeled with TotalSeq-B0252 for hashing. After labeling, cells were washed three times with 2 mL of cell staining buffer (BioLegend), with centrifugation at 200 g for 5 minutes at 4°C after each wash.

4.3 Single-Cell Transcriptomics

Day 26 cryopreserved fibroblasts transduced with FH4a were thawed and expanded in complete Myelocult supplemented with Dox. On Day 48, full-length barcoded cDNA amplicons were obtained from FACS-sorted reprogrammed cells (Library 1) using the 10x Genomics Next GEM Single Cell 3' Kit V3.1. Briefly, 20,000 sorted $CD34^+$ reprogrammed cells and 2,000 control fibroblasts were loaded onto the 10x Genomics Chromium Controller (X) to partition single cells into Gel Beads-In Emulsion (100% viability). cDNA was generated from polyadenylated mRNA transcripts, incorporating a unique molecular identifier (UMI) and a cell-specific barcode for each transcript. Barcoded cDNA was pooled, purified with SILANE magnetic beads (10X Genomics, #PN-2000048), and amplified by PCR. Barcoded cDNA libraries were prepared for sequencing according to the Oxford Nanopore Technologies (ONT) 3' cDNA protocols (Table 3). Briefly, cDNA amplicons were biotin tagged by PCR with the primers shown in Table 3. Biotinylated cDNA was purified with Dynabeads™ M-280 Streptavidin beads (Invitrogen, #11205D), followed by secondary PCR amplification and Agencourt AMPure XP bead clean-up (Beckman Coulter, #A63881). Products were assessed for quality following ONT recommendations. Long-read cDNA libraries were prepared for sequencing as in Table 3 and sequenced on an ONT GridION or PromethION Solo 2 (P2) sequencer. Raw reads were base-called with `guppy` using the High Accuracy basecalling model and stored as fastq files.

Single cell data from untransduced BJ fibroblasts (passage 6) (Library 2) and from bone marrow $CD34^+$ HSPCs (Library 3) were obtained as for the reprogrammed cells except using the Next GEM Single Cell 3' Kit V4. For Library 2, fibroblasts were stained with 16 μM Hoechst 33342 for 50 minutes then sorted by cell cycle phase based on DNA content using a Bigfoot Spectral Cell Sorter. Isolated cells from G1, S, and G2/M phases were labeled individually with TotalSeq-B human hashing antibodies (BioLegend, Cat# 394631, #394603, and #394605, respectively) before pooling and single cell barcoding (100% viability). For Library 3, approximately 8,000 hashed $CD34^+$ HSPCs (6,400 $CD34^+/CD38^-$ and 1,600 $CD34^+/CD38^+$) were used for single cell barcoding (~72% viability). Transcript and hashing libraries were prepared for sequencing as described above and sequenced on an ONT P2 sequencer.

10x Genomics		
Next GEM Single Cell 3' Kit (V3.1,V4)		
ONT		
Sequencer	**3' Protocol Version**	**Adapter Kit**
GridION	SST_V9148_V111_REVB_12JAN2022	cDNA-PCR Sequencing Kit (SQK-PCS111)
PromethION P2 Solo	SST_V9198_V114_REVB_06DEC2023	Ligation Sequencing Kit V14 (SQK-LSK114)
Primers		
Transcript libraries		
Fwd: 5' - [Biotin] -	CAGCACTTGCCTGTCGCTCTATCTTCCTACACGACGCTCTTCCGATCT	- 3'
Rev: 5' -	CAGCTTTCTGTTGGTGCTGATATTGCAAGCAGTGGTATCAACGCAGAG	- 3'
Hashing library		
Fwd: 5' - [Biotin] -	GCAGCGTCAGATGTGTATAAGAGACAG	- 3'
Fwd: 5' -	GTGACTGGAGTTCAGACGT	- 3'

Table 3: **Protocols and primers used for single cell barcoding and ONT sequencing.**

4.4 scRNA-seq Data Analysis

4.4.1 Data Processing

All raw sequencing reads from each library were processed using the EPI2ME Labs `wf-single-cell` pipeline (github.com/epi2me-labs/wf-single-cell). Reads were first aligned to the GRCh38 reference genome (10x Genomics, 2024-A), and gene-level count matrices were generated using either 3' V3 or 3' V4 barcode sets. The following parameters were used: `full_length = True`, `expected_cells = 10000`, `barcode_min_quality = 15`, `barcode_max_ed = 2`, and `gene_assigns_minqv = 30`.

Reads from each library were subjected to quality control using `Scanpy`. To integrate Library 1 (reprogram) and Library 2 (control), low quality cells were filtered in a library-aware manner using median absolute deviation (MAD) thresholds. Cells exceeding five MADs from the library-specific median for log-transformed total counts or detected genes, or showing high mitochondrial content (> 5 MADs or $> 20\%$), were filtered out. Across both libraries, genes with fewer than 10 total UMI counts were removed, and raw counts were normalized to 10,000 counts per cell and log-transformed using `scanpy.pp.log1p`. Highly variable genes were selected using the `seurat_v3` method with library as the batch key, and principal component analysis (PCA) was performed using `scanpy.pp.pca`. Data was integrated using `Harmony`[85] (version 0.0.10) with library as the batch key. A neighborhood graph was computed using the Harmony-corrected PCA space (`n_neighbors=100`, `n_pcs=25`). A force-directed graph was constructed using `scanpy.tl.draw_graph` with `layout='fa'`, and Leiden clustering was performed with `resolution=0.3`.

Cells from Library 3 (bone marrow) were processed separately. Cells with fewer than 1,000 detected genes or fewer than 500 total UMI counts were filtered out, and genes with fewer than 10 total UMI counts were removed. Unique UMI counts were normalized to 10,000 counts per cell and log-transformed using `log1p`. PCA was performed on highly variable genes, and a neighborhood graph was computed with 45 nearest neighbors. A UMAP embedding was constructed (`min_dist=0.2`), and Leiden clustering was performed with `resolution=0.38`. Clusters were annotated based on marker gene expression for expected cell types.

4.4.2 BioLegend TotalSeq-B Data Processing

We developed a pipeline using `Snakemake` to process BioLegend TotalSeq-B libraries sequenced on ONT platforms. Briefly, reads were demultiplexed based on hash sequences using `nanoplexer` (https://github.com/hanyue36/nanoplexer), with parameters `-L 150 -m 3 -x 2 -o 3 -e 1`. Following demultiplexing, barcode correction was performed using `Flexiplex`[86], with parameters `-e 1 -x '' -b ??????????????? -u '' -x ''`. Corrected barcode-hash sequence pairs were quantified and aggregated per cell. For Library 2, the hash-sequence with the most read-pairs per cell barcode was taken as the barcoded phase for each cell.

4.4.3 Differential Expression Analysis

Differential expression analysis was performed with the `scanpy.tl.rank_genes_groups` over log-normalized expression using the Wilcoxon rank-sum test (two-sided) with Benjamini-Hochberg correction. Unless stated otherwise, differential genes were identified in each cell type or cluster with the following criteria: expressed in $> 30\%$ of group; $|\log_2\text{FoldChange}| > 0.5$; adjusted p-value < 0.05.

4.4.4 RNA Velocity Inference

We developed a `Snakemake` workflow to preprocess BAM files for RNA velocity analysis with `velocyto`[36]. Alignment files generated by the EPI2ME Labs `wf-single-cell` pipeline were used as input for `velocyto`. Pre-processed gene expression data was merged with spliced/unspliced counts from sample-specific `velocyto` output files. Cell barcodes were aligned, and barcodes that did not pass quality control in the gene expression data were excluded from downstream analysis. All RNA velocity analysis was carried out in `scVelo`[37] (version 0.3.3). First, we computed a nearest-neighbor graph (`n_neighbors`=150) on the Harmony-corrected PCA space, after which moments were estimated with `scv.pp.moments`. Gene-specific kinetics were inferred using the dynamical model (`scv.tl.recover_dynamics`). Per-cell velocities were calculated with `scv.tl.velocity`, and confidence of velocities were calculated with `scv.tl.velocity_confidence`. A velocity graph was then constructed (`scv.tl.velocity_graph`), terminal states and latent times were inferred (`scv.tl.terminal_states`, `scv.tl.latent_time`), and velocity-associated genes were ranked (`scv.tl.rank_velocity_genes`). Stream and grid embeddings were visualized on a force-directed layout using `scv.pl.velocity_embedding_stream`.

4.4.5 Pseudotemporal Inference

To identify biologically-informed root cells for pseudotime inference, cell cycle phases were assigned to individual cells with `scanpy.tl.score_genes_cell_cycle()` using established S and G2/M gene signatures[87]. Predicted phases were compared to DNA content-based phase assignments obtained from the hashing barcodes that were added after cell sorting (Supplementary Table 4). Representative root cells were selected from the control cell population. We chose root cells from both G1 and G2/M cell cycle phases based on minimum Euclidean distance from the centroid of the respective phase group in Harmony-corrected PCA space[85]. For each root, velocity psueodotime was computed using the `scVelo` function `scv.pl.velocity_pseudotime`, producing pseudotemporal orderings for G1 and G2/M, which were averaged for each cell. Aggregate pseudotemporal orderings serve as root-agnostic measure of transcriptional distance from the initial state. To analyze transcriptional

dynamics along the inferred reprogramming trajectory, log-normalized gene expression was correlated with aggregated pseudotime. We computed Spearman correlation coefficients and associated p-values for each gene relative to pseudotime, followed by multiple testing correction using the Benjamaini-Hochberg method to control the false discovery rate[88]. Only genes with adjusted $p < 0.05$ were included for downstream analyses.

For diffusion-based pseudotime[38], a diffusion map was constructed on the Harmony-corrected PCA space (`rsc.tl.diffmap`), and a neighborhood graph was constructed from the resulting components. For G1 and G2/M root cells, diffusion pseudotime orderings were inferred using `scanpy.tl.dpt`, which were averaged for each cell. Gene expression correlations with diffusion pseudotime were computed as described above.

4.4.6 Harmonization and Visualization of Long-Read scRNA-seq Datasets

To harmonize data across all three long-read scRNA-seq libraries, we used `scVI`[30] from `scvi-tools`[83]. Processed UMI count matrices were concatenated and cell cycle scores we computed using `scanpy.tl.score_genes_cell_cycle()` with established gene sets[87]. Next, we defined a variational autoencoder with the batch identifier as a categorical covariate and S and G2/M scores as continuous covariates. We used a zero-inflated negative binomial (ZINB) likelihood, with per-gene–per-cell dispersion and two hidden layers of 128 units each with a dropout rate of 0.25, and a 24-dimensional latent representation. Training was performed on an NVIDIA Tesla V100 or A100 for up to 1,000 epochs with early stopping (`patience=10`). Next, we performed PCA on the scVI latent space, and computed a neighborhood graph using 35 neighbors. Cluster connectivity was inferred using `scanpy.tl.paga`[89], grouping cells by their annotated labels. The resulting Partition-based Graph Abstraction (PAGA) coordinates were used to initialize a 2-dimensional UMAP embedding with parameters `spread=0.55` and `min_dist=0.2`.

Pairwise Euclidean and cosine distances were used to quantify similarity between reprogrammed cells and initial or target populations in the 24-dimensional latent representation. For each reprogrammed cell, distances were calculated to all cells annotated as initial or HSC (`cuml.metrics.pairwise_distances`) and averaged. Per-cell average distances were grouped by cluster (R1-R3) and visualized. For visual interpretation of distances, "close" and "far" reference thresholds were calculated as the mean pairwise distance between HSCs and hematopoietic progenitors ("close") and between HSCs and initial fibroblasts ("far").

4.4.7 Integrated Atlas Benchmarking

We compiled scRNA-seq data from four sources: the native reference atlas (Methods 4.1), Ng 2024[1], Gomes 2018[10], and this study (Methods 4.3), yielding 983,455 cells and 105 annotated cell types. After excluding immune populations, the reference atlas dataset contributed 709,808 cells across 72 cell types. The Ng 2024 dataset provided 252,607 cells spanning 20 types, Gomes 2018 contributed 286 cells from five populations, and this study added 20,754 cells across 12 clusters. All datasets were filtered to remove unwanted cell types and non-coding genes, and genes with fewer than 10 total counts were excluded. Up to 250,000 cells were randomly sampled per dataset, retaining relevant metadata fields where available. After sampling, 521,040 cells and 17,607 genes remained. To ensure robust downstream integration, batch labels with fewer than 100 cells were excluded. Additional filtering removed cells with fewer than 500 total counts (17,783 cells) and genes with low or zero counts (234 genes). Cells with fewer than 1,000 detected genes were also removed. The final dataset contained 420,108 cells and 17,373 genes, which were normalized to 10,000 total counts per

cell and log-transformed. Cell-cycle phase scores were predicted for each cell as in Methods 4.4.6. Highly variable genes were selected using the `seurat_v3` method, retaining the top 3,000 features. Genes annotated as mitochondrial or ribosomal were excluded from highly variable gene selection.

To harmonize single-cell profiles, a `scVI` model was trained using the data source as the batch key and including continuous covariates for cell cycle phase (S and G2M scores). The model architecture included two encoder/decoder layers and 128 hidden units each with a dropout rate of 0.15, and a 24-dimensional latent representation. Gene expression counts were modeled with a ZINB likelihood and per-gene-per-cell dispersion parameters (`dispersion=`gene-cell'`). Training was performed on either an NVIDIA Tesla V100 or A100 for up to 500 epochs with early stopping (patience=30), a learning rate of 1×10^{-3}, and a batch size of 500. Model performance was monitored by tracking training and validation loss, evidence lower bound (ELBO), reconstruction error, and Kullback–Leibler (KL) divergence. We performed PCA on the scVI latent space, computed a neighborhood graph (`n_neighbors=15`), and generated a UMAP embedding (`min_dist=0.2`). Leiden clustering was performed with `resolution=0.45`.

To compute the abstract connectivity graph of cell clusters, we applied PAGA[89] using `scanpy.tl.paga`, and visualization was generated with `scanpy.pl.paga`. Pairwise Euclidean distances were used to compare pseudobulk profiles of cell clusters in the latent space using `cuml.metrics.pairwise_distances`, where pseudobulk profiles were obtained by taking the mean of the latent vector per cluster. To compare single-cell profiles, Euclidean distances were computed between each cell and the mean vector of R2/R3 cells in the latent space using `CuPy.linalg.norm`. Per-cell distances were grouped by cluster-resolved cell type annotations.

4.4.8 Curated Dataset Benchmarking

We benchmarked our scRNA-seq data on six publicly available scRNA-seq datasets (Supplementary Table 5). Briefly, we obtained 359,848 single cells, and a broad spectrum of 71 annotated cell types from[1,10,19,21,40,90]. For each dataset, pseudobulk profiles were obtained by summing over raw UMI counts for each source-defined cell type label. Aggregated counts were normalized using `scanpy.tl.normalize_total` with `target_sum=`10^4 and log-transformed using `scanpy.tl.log1p`. Cell-cycle phase scores were predicted for each pseudobulk profile as in Methods 4.4.6. We filtered genes annotated as mitochondrial, ribosomal, non-protein-coding, or cell-cycle related[87], and only genes whose mean expression lay in the upper quartile in both the test dataset and the curated datasets were retained, yielding 15,216 shared genes across 82 pseudobulk profiles. Pseudobulk profiles were harmonized as in Methods 4.4.7 with a dropout rate of 0.25, batch size of 25, and a 90/10 train/validation split. PCA was then computed on the harmonized latent space. Pairwise Euclidean distances were used to compare pseudobulk profiles in the PCA representation of the latent space using (`cuml.metrics.pairwise_distances`).

4.4.9 Partial Identity Scoring in Single Cells

We used `Capybara`[22] to estimate the fractional identity scores of single-cells based on aggregated pseudobulk profiles as described in Methods 4.4.8. Briefly, we used the intersection of gene sets between the input data and the reference pseudobulk profiles. Identity scores were computed using `single.round.QP.analysis` with 24 CPU cores, log-transformed and scaled both bulk and single-cell counts, and forced no equality constraints (`force.eq=0`). Resulting scores were stored for downstream analysis.

4.4.10 Isoform Identification

Transcript-level UMI count matrices were generated with the EPI2ME Labs `wf-single-cell` pipeline (github.com/epi2me-labs/wf-single-cell) using default parameters. Sequencing reads were aligned to the GRCh38 reference genome (10x Genomics, 2024-A) prior to quantification.

Reads were assigned to transcripts following a strategy adapted from `FLAMES`[91]. Alignments mapping to intronic query transcripts were labeled as unknown ("`--`"). Reads with a single, unique alignment to the annotated transcriptome were directly attributed to that transcript. For ambiguously mapped reads, alignments were ordered by alignment score (AS); if the top hit exhibited a higher AS or greater query coverage than the second hit and covered at least 40% of the transcript, the read was assigned accordingly, otherwise it was marked unknown. When the top two alignments shared identical AS values, the alignment with the greater transcript coverage was selected provided its coverage exceeded 80%. After filtering out cell barcodes that failed gene-expression quality control, the final dataset is comprised of 20,754 cells across three libraries and 113,335 detected transcripts. For downstream analyses, transcripts with fewer than 30 total UMI counts were filtered out in `Scanpy`. Raw counts were normalized to 10,000 counts per cell and log-transformed. Protein identifiers were assigned to transcripts by mapping Ensembl transcript IDs to UniProt accessions from `HUMAN_9606_idmapping_selected.tab.gz`.

4.4.11 Splicing Event Annotation

AS events were annotated with `SUPPA`[47] using the GRCh38_2024-A gene annotation. Analyses were run in intron-over-intron (`ioi`) and intron-over-exon (`ioe`) modes with gene pooling enabled; the `ioe` run was restricted to the SE, SS, MX, RI, and FL event classes (Table 4).

Code	Description
SE	Skipping exon
A5	Alternative 5′ splice site
A3	Alternative 3′ splice site
MX	Mutually exclusive exon
RI	Retained intron
AF	Alternative first exon
AL	Alternative last exon

Table 4: **AS event types.**

4.5 Statistics and Reproducibility

Statistical analyses were performed using GraphPad Prism (version 10.4.1) or SciPy (`scipy.stats`, version 1.16.1). Unless stated otherwise, p or FDR values < 0.05 were considered as statistically significant. All sample sizes and replicate definitions are described in corresponding Methods sections or in the main text.

5 Author contributions

LAM and IR initiated and supervised the project, including setting overall goals and designing the experiments. WM, HL, LAM, and JC processed cells and tissues and collected sequencing data. JC, CMS, SG, LAM, and IR conducted data analysis and interpretation and created data visualizations. LAM and IR supervised the use of computational and bioinformatics methods, which were implemented by CMS and JC. IR and LAM provided project resources. All authors contributed to the development of methodologies and to the writing and editing of the manuscript.

6 Acknowledgments

We thank Max Wicha at the University of Michigan for valuable discussions conversations and ideas on how to present results. We thank the University of Michigan Advanced Genomics Core for assistance with 10x Genomics single cell barcoding and the Flow Cytometry Core for assistance with cell sorting. We acknowledge BioRender, which was used to assemble figures and for generating the graphical abstract. This work was supported by the National Science Foundation (NSF) award number 2035827 (IR) subcontracted from iReprogram, Inc., the Air Force Office of Scientific Research (AFOSR) award number FA9550-22-1-0215 (IR), the Defense Advanced Research Projects Agency (DARPA) award number HR00112490472 (IR and LM), NIGMS GM150581 (JC), iReprogram, Inc, and ONT. We thank Lakmal Jayasinghe and the research and support teams at ONT for helpful discussions and technical assistance.

7 Data Availability

Raw data generated in this study is available in the Gene Expression Omnibus with accession number GSE283658.

8 Code Availability

All source code for data processing and analysis for this study is available at:
https://github.com/jrcwycy/Hematokytos

9 Competing interests

IR and LM are founders of and shareholders in iReprogram, Inc., which may benefit financially from the subject matter or materials discussed in this manuscript.

References

1. Ng, E. S. *et al.* Long-Term Engrafting Multilineage Hematopoietic Cells Differentiated from Human Induced Pluripotent Stem Cells. *Nat Biotechnol,* 1–14 (2024).
2. Holtan, S. G. *et al.* Disease Progression, Treatments, Hospitalization, and Clinical Outcomes in Acute GVHD: A Multicenter Chart Review. *Bone Marrow Transplant* **57,** 1581–1585 (2022).
3. Schlaeger, T. M. *et al.* A Comparison of Non-Integrating Reprogramming Methods. *Nat Biotechnol* **33,** 58–63 (2015).
4. Volpato, V. & Webber, C. Addressing variability in iPSC-derived models of human disease: guidelines to promote reproducibility. *Disease Models & Mechanisms* **13,** dmm042317 (2020).
5. Wang, H., Yang, Y., Liu, J. & Qian, L. Direct Cell Reprogramming: Approaches, Mechanisms and Progress. *Nat Rev Mol Cell Biol* **22,** 410–424 (2021).
6. Han, L. *et al.* Distinctive Clinical and Pathologic Features of Immature Teratomas Arising from Induced Pluripotent Stem Cell-Derived Beta Cell Injection in a Diabetes Patient. *Stem Cells and Development* **31,** 97–101 (2022).
7. Movahed, A. Y., Bagheri, R., Savatier, P., Šarić, T. & Moradi, S. Elimination of Tumorigenic Pluripotent Stem Cells from Their Differentiated Cell Therapy Products: An Important Step toward Ensuring Safe Cell Therapy. *Stem Cell Reports* **20** (2025).
8. Szabo, E. *et al.* Direct Conversion of Human Fibroblasts to Multilineage Blood Progenitors. *Nature* **468,** 521–526 (2010).
9. Daniel, M. G., Pereira, C.-F., Lemischka, I. R. & Moore, K. A. Making a Hematopoietic Stem Cell. *Trends in Cell Biology* **26,** 202–214 (2016).
10. Gomes, A. M. *et al.* Cooperative Transcription Factor Induction Mediates Hemogenic Reprogramming. *Cell Reports* **25,** 2821–2835.e7 (2018).
11. Silvério-Alves, R., Gomes, A. M., Kurochkin, I., Moore, K. A. & Pereira, C.-F. Hemogenic Reprogramming of Human Fibroblasts by Enforced Expression of Transcription Factors. *JoVE (Journal of Visualized Experiments),* e60112 (2019).
12. Daniel, M. G. *et al.* Induction of Human Hemogenesis in Adult Fibroblasts by Defined Factors and Hematopoietic Coculture. *FEBS Letters* **593,** 3266–3287 (2019).
13. Wei, C., Yu, P. & Cheng, L. Hematopoietic Reprogramming Entangles with Hematopoiesis. *Trends in Cell Biology* **30,** 752–763 (2020).
14. Ding, J., Alavi, A., Ebrahimkhani, M. R. & Bar-Joseph, Z. Computational Tools for Analyzing Single-Cell Data in Pluripotent Cell Differentiation Studies. *Cell Reports Methods* **1** (2021).
15. Liu, G., David, B. T., Trawczynski, M. & Fessler, R. G. Advances in Pluripotent Stem Cells: History, Mechanisms, Technologies, and Applications. *Stem Cell Rev and Rep* **16,** 3–32 (2020).
16. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163,** 1663–1677 (2015).
17. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage Tracing on Transcriptional Landscapes Links State to Fate during Differentiation. *Science* **367,** eaaw3381 (2020).

18. Watcham, S., Kucinski, I. & Gottgens, B. New Insights into Hematopoietic Differentiation Landscapes from Single-Cell RNA Sequencing. *Blood* **133,** 1415–1426 (2019).

19. Pellin, D. *et al.* A Comprehensive Single Cell Transcriptional Landscape of Human Hematopoietic Progenitors. *Nat Commun* **10,** 2395 (2019).

20. Ranzoni, A. M. *et al.* Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell* **28,** 472–487.e7 (2021).

21. Weng, C. *et al.* Deciphering Cell States and Genealogies of Human Haematopoiesis. *Nature* **627,** 389–398 (2024).

22. Kong, W. *et al.* Capybara: A Computational Tool to Measure Cell Identity and Fate Transitions. *Cell Stem Cell* **29,** 635–649.e11 (2022).

23. Theodoris, C. V. *et al.* Transfer Learning Enables Predictions in Network Biology. *Nature* **618,** 616–624 (2023).

24. He, Z. *et al.* An Integrated Transcriptomic Cell Atlas of Human Neural Organoids. *Nature* **635,** 690–698 (2024).

25. Heimberg, G. *et al.* A Cell Atlas Foundation Model for Scalable Search of Similar Human Cells. *Nature* **638,** 1085–1094 (2025).

26. Ronquist, S. *et al.* Algorithm for Cellular Reprogramming. *PNAS* **114,** 11832–11837 (2017).

27. Heumos, L. *et al.* Best Practices for Single-Cell Analysis across Modalities. *Nat Rev Genet* **24,** 550–572 (2023).

28. THE TABULA SAPIENS CONSORTIUM. The Tabula Sapiens: A Multiple-Organ, Single-Cell Transcriptomic Atlas of Humans. *Science* **376,** eabl4896 (2022).

29. Morris, S. A. Redefining cellular reprogramming with advanced genomic technologies. *Nature Reviews Genetics,* 1–19 (2025).

30. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep Generative Modeling for Single-Cell Transcriptomics. *Nat Methods* **15,** 1053–1058 (2018).

31. van der Maaten, L. & Hinton, G. Visualizing Data Using T-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

32. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: A Web Server for Exploration of Mouse and Human Single-Cell RNA Sequencing Data. *Database* **2019,** baz046 (2019).

33. Maeda, T. Regulation of hematopoietic development by ZBTB transcription factors. *International journal of hematology* **104,** 310–323 (2016).

34. Krausgruber, T. *et al.* Structural cells are key regulators of organ-specific immune responses. *Nature* **583,** 296–302 (2020).

35. MacLean, A. L., Hong, T. & Nie, Q. Exploring Intermediate Cell States through the Lens of Single Cells. *Current Opinion in Systems Biology. Mathematic Modelling* **9,** 32–41 (2018).

36. La Manno, G. *et al.* RNA Velocity of Single Cells. *Nature* **560,** 494–498 (2018).

37. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling. *Nat Biotechnol* **38,** 1408–1414 (2020).

38. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion Pseudotime Robustly Reconstructs Lineage Branching. *Nat Methods* **13,** 845–848 (2016).

39. Sugimura, R. *et al.* Haematopoietic Stem and Progenitor Cells from Human Pluripotent Stem Cells. *Nature* **545,** 432–438 (2017).

40. Setty, M. *et al.* Characterization of Cell Fate Probabilities in Single-Cell Data with Palantir. *Nat Biotechnol* **37,** 451–460 (2019).

41. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat].* arXiv: 1802.03426 [cs, stat] (2020).

42. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology* **34,** 1145–1160 (2016).

43. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165,** 535–550 (2016).

44. Labib, M. & Kelley, S. O. Single-cell analysis targeting the proteome. *Nature Reviews Chemistry* **4,** 143–158 (2020).

45. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nature reviews Molecular cell biology* **18,** 437–451 (2017).

46. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nature methods* **10,** 1177–1184 (2013).

47. Trincado, J. L. *et al.* SUPPA2: Fast, Accurate, and Uncertainty-Aware Differential Splicing Analysis across Multiple Conditions. *Genome Biology* **19,** 40 (2018).

48. Wright, C. J., Smith, C. W. J. & Jiggins, C. D. Alternative Splicing as a Source of Phenotypic Diversity. *Nat Rev Genet* **23,** 697–710 (2022).

49. Wieder, N. *et al.* The role of untranslated region variants in Mendelian disease: a review. *European Journal of Human Genetics,* 1–10 (2025).

50. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25,** 25–29 (2000).

51. Aleksander, S. A. *et al.* The gene ontology knowledgebase in 2023. *Genetics* **224,** iyad031 (2023).

52. Hu, C. *et al.* CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic acids research* **51,** D870–D876 (2023).

53. Silvério-Alves, R. *et al.* GATA2 Mitotic Bookmarking Is Required for Definitive Haematopoiesis. *Nat Commun* **14,** 4645 (2023).

54. Schuringa, J. J., Wu, K., Morrone, G. & Moore, M. A. Enforced Activation of STAT5A Facilitates the Generation of Embryonic Stem–Derived Hematopoietic Stem Cells That Contribute to Hematopoiesis In Vivo. *Stem Cells* **22,** 1191–1204 (2004).

55. Schuringa, J. J., Chung, K. Y., Morrone, G. & Moore, M. A. Constitutive Activation of STAT5A Promotes Human Hematopoietic Stem Cell Self-Renewal and Erythroid Differentiation. *Journal of Experimental Medicine* **200,** 623–635 (2004).

56. Kollmann, S. *et al.* A STAT5B–CD9 axis determines self-renewal in hematopoietic and leukemic stem cells. *Blood* **138,** 2347–2359 (2021).

57. Espín-Palazón, R. & Traver, D. The NF- B family: Key players during embryonic development and HSC emergence. *Experimental Hematology* **44,** 519–527 (2016).

58. Espín-Palazón, R. *et al.* Proinflammatory signaling regulates hematopoietic stem cell emergence. *Cell* **159,** 1070–1085 (2014).

59. He, Q. *et al.* Inflammatory signaling regulates hematopoietic stem and progenitor cell emergence in vertebrates. *Blood, The Journal of the American Society of Hematology* **125,** 1098–1106 (2015).

60. Turco, M. C. *et al.* NF-κB/Rel-mediated Regulation of Apoptosis in Hematologic Malignancies and Normal Hematopoietic Progenitors. *Leukemia* **18,** 11–17 (2004).

61. Gilmore, T. D. & Gerondakis, S. The C-Rel Transcription Factor in Development and Disease. *Genes & Cancer* **2,** 695–711 (2011).

62. Fallahi, S. *et al.* Impact of C-rel Inhibition of Cord Blood-Derived B-, T-, and NK Cells. *Journal of Immunotoxicology* **14,** 15–22 (2017).

63. Gerloff, D. *et al.* NF-κB/STAT5/miR-155 network targets PU. 1 in FLT3-ITD-driven acute myeloid leukemia. *Leukemia* **29,** 535–547 (2015).

64. Calvanese, V. *et al.* Mapping Human Haematopoietic Stem Cells from Haemogenic Endothelium to Birth. *Nature* **604,** 534–540 (2022).

65. Pereira, C.-F. *et al.* Induction of a Hemogenic Program in Mouse Fibroblasts. *Cell Stem Cell* **13,** 205–218 (2013).

66. Batta, K., Florkowska, M., Kouskoff, V. & Lacaud, G. Direct Reprogramming of Murine Fibroblasts to Hematopoietic Progenitor Cells. *Cell Reports* **9,** 1871–1884 (2014).

67. Daniel, M. G., Lemischka, I. R. & Moore, K. Converting Cell Fates: Generating Hematopoietic Stem Cells de Novo via Transcription Factor Reprogramming. *Annals of the New York Academy of Sciences* **1370,** 24–35 (2016).

68. Ohta, S., Nishida, E., Yamanaka, S. & Yamamoto, T. Global Splicing Pattern Reversion during Somatic Cell Reprogramming. *Cell Reports* **5,** 357–366 (2013).

69. Kanitz, A., Syed, A. P., Kaji, K. & Zavolan, M. Conserved Regulation of RNA Processing in Somatic Cell Reprogramming. *BMC Genomics* **20,** 100 (2019).

70. Vivori, C. *et al.* Dynamics of Alternative Splicing during Somatic Cell Reprogramming Reveals Functions for RNA-binding Proteins CPSF3, hnRNP UL1, and TIA1. *Genome Biol* **22,** 171 (2021).

71. Zavolan, M. & Kanitz, A. RNA Splicing and Its Connection with Other Regulatory Layers in Somatic Cell Reprogramming. *Current Opinion in Cell Biology. Cell Nucleus* **52,** 8–13 (2018).

72. Han, H. *et al.* MBNL Proteins Repress ES-cell-specific Alternative Splicing and Reprogramming. *Nature* **498,** 241–245 (2013).

73. Toh, C.-X. D. *et al.* RNAi Reveals Phase-Specific Global Regulators of Human Somatic Cell Reprogramming. *Cell Reports* **15,** 2597–2607 (2016).

74. Waddington, C. H. *The Strategy of the Genes* ISBN: 978-1-315-76547-1 (Routledge, London, 2014).

75. Kitano, H. Biological Robustness. *Nat Rev Genet* **5,** 826–837 (2004).

76. Ferrell, J. E. Bistability, Bifurcations, and Waddington's Epigenetic Landscape. *Current Biology* **22,** R458–R466 (2012).

77. Lee, W. *et al.* A single-cell atlas of in vitro multiculture systems uncovers the in vivo lineage trajectory and cell state in the human lung. *Experimental & Molecular Medicine* **55,** 1831–1842 (2023).

78. Huang, C. *et al.* Single-cell transcriptomic analysis of chondrocytes in cartilage and pathogenesis of osteoarthritis. *Genes & Diseases* **12,** 101241 (2025).

79. Rood, J. E., Hupalowska, A. & Regev, A. Toward a Foundation Model of Causal Cell and Tissue Biology with a Perturbation Cell and Tissue Atlas. *Cell* **187,** 4520–4545 (2024).

80. Szałata, A. *et al.* Transformers in Single-Cell Omics: A Review and New Perspectives. *Nat Methods* **21,** 1430–1443 (2024).

81. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biol* **19,** 1–5 (2018).

82. Dicks, S. *et al. Scverse/Rapids_singlecell: V0.10.6* Zenodo. 2024.

83. Gayoso, A. *et al.* A Python Library for Probabilistic Analysis of Single-Cell Omics Data. *Nat Biotechnol* **40,** 163–166 (2022).

84. Xu, C. *et al.* Probabilistic Harmonization and Annotation of Single-cell Transcriptomics Data with Deep Generative Models. *Molecular Systems Biology* **17,** e9620 (2021).

85. Korsunsky, I. *et al.* Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony. *Nat Methods* **16,** 1289–1296 (2019).

86. Cheng, O. *et al.* Flexiplex: A Versatile Demultiplexer and Search Tool for Omics Data. *Bioinformatics* **40,** btae102 (2024).

87. Tirosh, I. *et al.* Single-Cell RNA-seq Supports a Developmental Hierarchy in Human Oligodendroglioma. *Nature* **539,** 309–313 (2016).

88. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57,** 289–300 (1995).

89. Wolf, F. A. *et al.* PAGA: Graph Abstraction Reconciles Clustering with Trajectory Inference through a Topology Preserving Map of Single Cells. *Genome Biology* **20,** 59 (2019).

90. Georgolopoulos, G. *et al.* Discrete Regulatory Modules Instruct Hematopoietic Lineage Commitment and Differentiation. *Nat Commun* **12,** 6790 (2021).

91. Dong, X. *et al.* The Long and the Short of It: Unlocking Nanopore Long-Read RNA Sequencing Data with Short-Read Differential Expression Analysis Tools. *NAR Genomics and Bioinformatics* **3,** lqab028 (2021).