Transcriptional landscape of direct reprogramming toward hematopoietic stem cells

Cooper Stansbury[1,†], Jillian Cwycyshyn[1,2, †], Sarah Golts[1], Hyunsu Lee[1], 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).

2025-08-05

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. To address this, we assembled a single-cell transcriptomic reference atlas spanning hematopoietic, endothelial, and fibroblast cell types, and evaluated a novel algorithmically predicted transcription factor (TF) recipe—GATA2, GFI1B, FOS, REL, and STAT5A—for induction of HSC-like states. The 5TF recipe increased conversion efficiency of fibroblasts to $CD34^+$ cells by approximately three-fold compared to previous protocols. Long-read single-cell RNA sequencing revealed heterogeneous expression of hematopoietic- and endothelial-associated genes, consistent with partial reprogramming. To benchmark reprogrammed cells, we developed an approach to map their transcriptomic positions relative to native and initial cell states within the reference atlas. Finally, we cataloged isoform diversity in reprogrammed cells, suggesting a role for alternative splicing in reprogramming dynamics. These findings highlight both the promise and complexity of direct HSC reprogramming and provide a framework for optimizing future protocols.

1 Introduction

The generation of hematopoietic stem cells (HSCs) through cell reprogramming would enable development of novel treatments for hematopoietic disorders. Autologous cell reprogramming would further avoid major causes of morbidity and mortality such as graft-versus-host-disease [1].

Reprogramming approaches for generating HSCs fall into two general categories: direct reprogramming between two somatic cell types and protocols for differentiating HSCs from somatic cells

that have been reprogrammed into induced pluripotent stem cells (iPSC). In the latter approach, a study recently demonstrated the production of cells that have all functional attributes of native HSCs [2]. While iPSC-based approaches have the potential to generate a large number of therapeutically-relevant cells, they still face significant barriers to use in autologous cell therapies, including long protocol timelines and low conversion efficiency, the presence of genomic instability, and variation in outcomes between individuals [3, 4]. Direct reprogramming between somatic cells, without using a pluripotent intermediate, has the potential to more rapidly produce clinically relevant cells that have a lower risk of tumorigenesis [5–7]. However, direct human somatic cell fibroblast-to-HSC reprogramming is limited by low conversion efficiency [8–13] and insufficient HSC function [5, 13–15] as measured by gold standard assays such as long-term reconstitution and multilineage engraftment potential after transplantation [13].

Gold standard assays for function require that cells reach a fully reprogrammed state, but evaluating partially reprogrammed cells is important for understanding the mechanisms governing reprogramming. A cell that fails to fully reprogram but achieves a state close to the target cell type might have major value, requiring only minor perturbation or modification of the reprogramming protocol to achieve full reprogramming. Additionally, a better understanding of intermediate states during reprogramming can aid prediction of whether cells are likely to achieve full reprogramming. However, rigorous methods for measuring how closely partially reprogrammed cells align to the target state are remain underdeveloped.

Single-cell transcriptomics has helped characterize differentiation in complex lineages, including the hematopoietic lineage in mice [16–18] and humans [19–21]. Similar approaches have been used to characterize phenotypic heterogeneity [2] and intermediate and off-target cell states in reprogramming [22]. The extent to which reprogrammed cells and their trajectories compare to native cell types remains incompletely characterized. Comprehensive single cell atlases have utility in placing lesser understood cell types within the context of known cell phenotypes [23–25], but this approach has not been used for understanding intermediate cell states in reprogramming.

In this study, we systematically compile and analyze a single-cell transcriptomic reference atlas of hematopoietic, endothelial, and fibroblast cell types. We test direct reprogramming of human fibroblasts toward the hematopoietic lineage with a novel algorithmically-predicted TF recipe—GATA2, GFI1B, FOS, REL, and STAT5A [26]—and evaluate the transcriptome profiles of the resulting $CD34^+$ cells using long-read single-cell RNA sequencing (scRNA-seq). We mapped our scRNA-seq data onto the reference atlas and characterized transcriptomic diversity among reprogrammed cells at both gene and isoform levels. Our approach establishes a framework for integration of algorithmic prediction with experimental validation that can guide optimization of reprogramming protocols.

2 Results

2.1 Construction of an atlas for hematopoietic cell reprogramming

A significant challenge in the optimization of direct reprogramming protocols is the evaluation of engineered cells, which exist in intermediate states not observed in typical biological settings. The increased availability of single-cell data presents an opportunity to benchmark intermediate cell states against diverse transcriptomes from large populations of target cells, allowing for unbiased characterization that accounts for both technical and biological heterogeneity [22]. To our knowledge, no such atlas exists for fibroblast to HSC direct reprogramming.

To this end, we constructed a reference atlas by integrating 1.12 million single-cell transcriptomes drawn from 270 publicly available datasets spanning 188 cell types from diverse tissues (Supplementary Figure 1), donors, and experimental contexts (Supplementary Figure 2A-C). To mitigate technical variability, data were uniformly preprocessed and harmonized using single-cell variational inference (scVI) and conditioned on standardized cell type annotations (Methods 4.1) [27]. A low-dimensional *t*-distributed stochastic neighbor embedding (t-SNE) was generated from the harmonized latent space to visualize gene expression patterns in the reference atlas (Figure 1A–B) [28]. Analysis of lineage-specific marker genes revealed strong enrichment of stromal-associated genes (*DCN/LUM*), canonical stem and progenitor markers (*CD34/GATA2*), pan-immune markers (*PTPRC/CD45*) and endothelial genes (*ACE/PECAM1*) within their respective compartments (Fig. 1A, C, Supplementary Figure 3).

Differential expression analysis revealed distinct transcriptional programs for the major cellular compartments, including endothelial, stromal, hematopoietic progenitor, and immune populations (Fig. 1D-F; Supplementary Figure 6A-C). Signatures of differentially expressed genes and TFs were consistent with established marker sets (Supplementary Tables 4, 5) [29, 30]. The hematopoietic compartment was enriched for key markers (*MYB*, *LMO2*, *HOXA9*) and TFs (*HMGA1*, *MSI2*) in progenitors and HSCs (Supplementary Figures 6A-D, 4A-B, 7, 10, 5). Similarly, the endothelial compartment broadly expressed markers like *ACE*, *PECAM1*, and *VWF*, alongside ETS-family and hypoxia-related TFs such as *ETS1*, *EPAS1*, and *SOX17*, and was enriched for genes implicated in angiogenesis, immune signaling, and vascular transport (Fig. 1C; Supplementary Figures 6A-D, 8, 11). Finally, fibroblasts were characterized by the expression of known mesenchymal regulators, including *PRRX1* and *TWIST1* (Supplementary Figures 9, 12).

Together, these data establish a comprehensive transcriptomic reference atlas for evaluating intermediate states in fibroblast to HSC direct reprogramming with respect to native cell types.

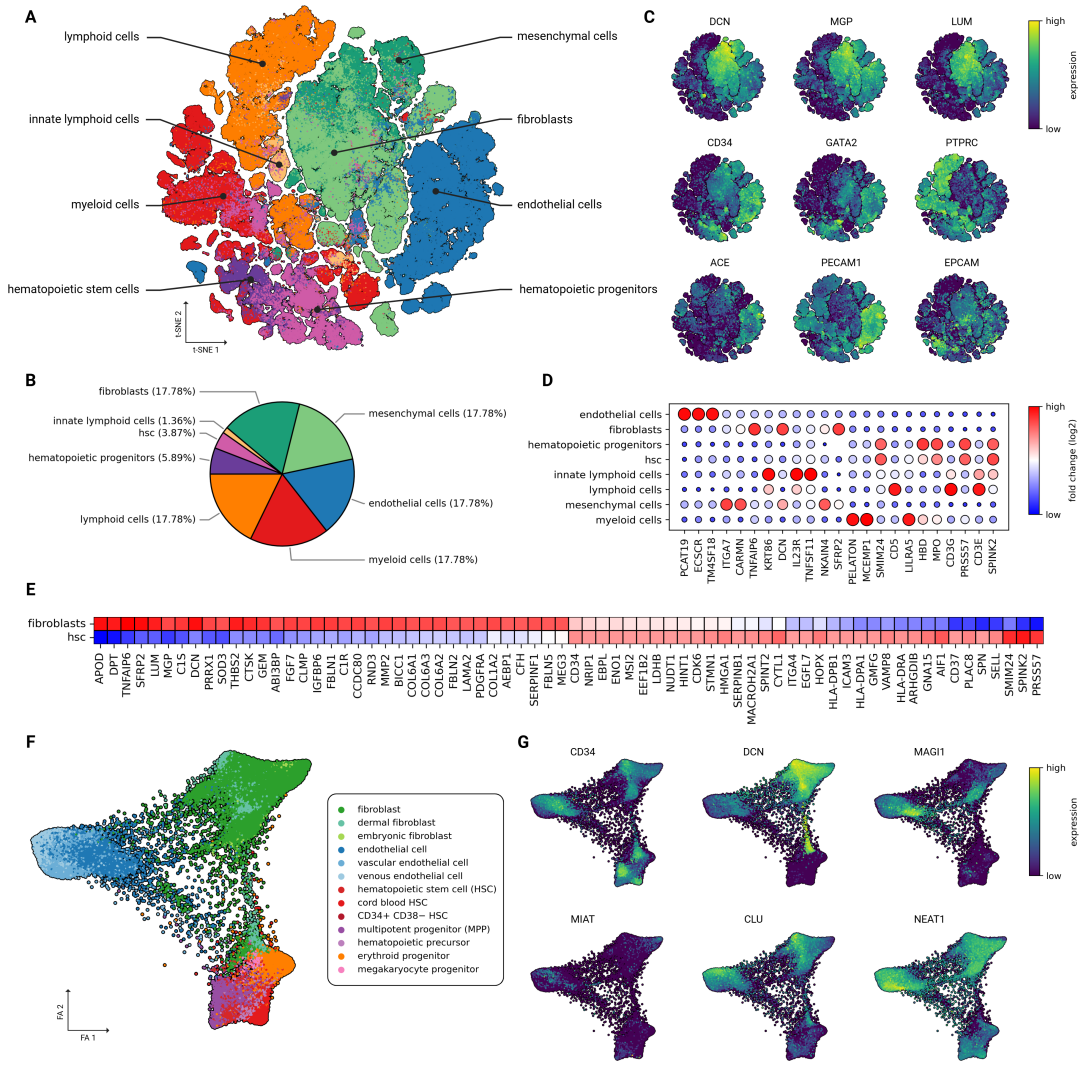

Figure 1: **A reference atlas of human hematopoietic cell types.** **(A)** t-SNE embedding of scANVI-integrated cell types relevant to FB-to-HSC reprogramming (1.12 million cells). **(B)** Proportions of major cell categories in (A). **(C)** Expression of fibroblast, hematopoietic, and endothelial marker genes projected onto the t-SNE coordinates from (A). **(D)** Top three differentially expressed genes per group (Wilcoxon rank-sum test, Benjamini–Hochberg adjusted $p \leq 0.05$; detected in 15% of cells per group). **(E)** Top 35 differentially expressed genes detected in 35% of fibroblasts and HSCs. **(F)** Force-directed graph layout of a sample of 74,539 fibroblasts, endothelial cells, and hematopoietic cells shown with their cell type annotations. **(G)** Expression of selected genes on the coordinates in (F).

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

Although a number of TF recipes have been used to reprogram human fibroblast cells into HSC-like cells [13] (Supplementary Figure 13), relatively little progress has been made toward improving conversion efficiency. We explored whether computationally predicted TFs from our data-guided control (DGC) algorithm could improve direct reprogramming of fibroblasts into hematopoietic-like cells (Methods 4.2.1) [26]. TFs which were previously demonstrated by Gomes *et. al.* to reprogram human fibroblasts into an HSC-like phenotype (3TF: GATA2, GFI1B, and FOS) [10] were among the top DGC-predicted candidates. Additionally, top DGC-predicted candidates STAT5A and REL were added to the 3TF recipe to test whether the novel 5TF recipe could increase the yield of CD34$^+$ cells from human fibroblasts.

Human dermal fibroblasts were transduced with pooled TFs delivered via lentiviral vectors under doxycycline (Dox)-inducible TET-ON expression cassettes (Figure 2A-B). Two days post-transduction, TF expression was induced, and surface expression of CD34 was assessed by flow cytometry using primary human bone marrow as a reference (Figure 2C). The 5TF cocktail enhanced reprogramming efficiency approximately three-fold compared to the previously published 3TF recipe, suggesting that the addition of *STAT5A* and *REL* promotes conversion of fibroblasts into CD34$^+$ cells (Figure 2C). Immunostaining and fluorescence microscopy of reprogrammed cells revealed CD34 surface expression and morphological differences between induced and non-induced cells, further supporting phenotypic shifts from the initial cell type (Supplementary Figure 15, Methods 4.2.5).

To further characterize reprogrammed cells, we performed transcriptomic profiling using Oxford Nanopore Technologies (ONT) long-read scRNA-seq. We prepared two scRNA-seq libraries (Figure 2D, E): Library 1 contained both CD34$^+$ fibroblasts reprogrammed with the 5TF recipe and CD34$^-$ controls. Library 2 contained fibroblasts without the addition of the 5TF recipe (Figure 2E). Both libraries were barcoded using the 10x Genomics Chromium platform and sequenced using ONT platforms (Methods 4.4). Raw reads were processed using the EPI2ME Labs scRNA-seq pipeline and quality controlled data from both libraries was pooled, harmonized, and clustered, revealing 5 clusters (C1-C5) from 15,808 cells (Figure 2F, Methods 4.5.1).

As expected, CD34$^-$ controls in Library 1 clustered together with cells from Library 2, indicating harmonization across libraries (Figure 2G). We evaluated the proportion of non-zero expression for selected genes across both libraries, finding higher rates of hematopoietic and endothelial-associated genes and lower rates of fibroblast markers in Library 1 compared to Library 2 (Figure 2H-I). Two clusters of cells from Library 1 (C2/C3) showed elevated expression of *CD34* and endothelial markers (*ACE* and *ITGA6*), consistent with previously reported intermediate cells reprogramed using the 3TF recipe (Figure 2I, Supplementary Figure 16B) [10]. Despite moderate enrichment of endothelial genes, Library 1 cells also displayed strong residual expression of fibroblast-associated genes *COL1A1* and *LUM* (Figure 2I, Supplementary Figure 16A-C). Interestingly, several other fibroblast-associated transcripts such as *PRRX1* and *DCN* had lower expression in clusters C2/C3 (Figure 2I).

Taken together, these results demonstrate conversion of transcriptionally distinct CD34$^+$ cells from an initial CD34$^-$ fibroblast population via the induced overexpression of the 5TF recipe.

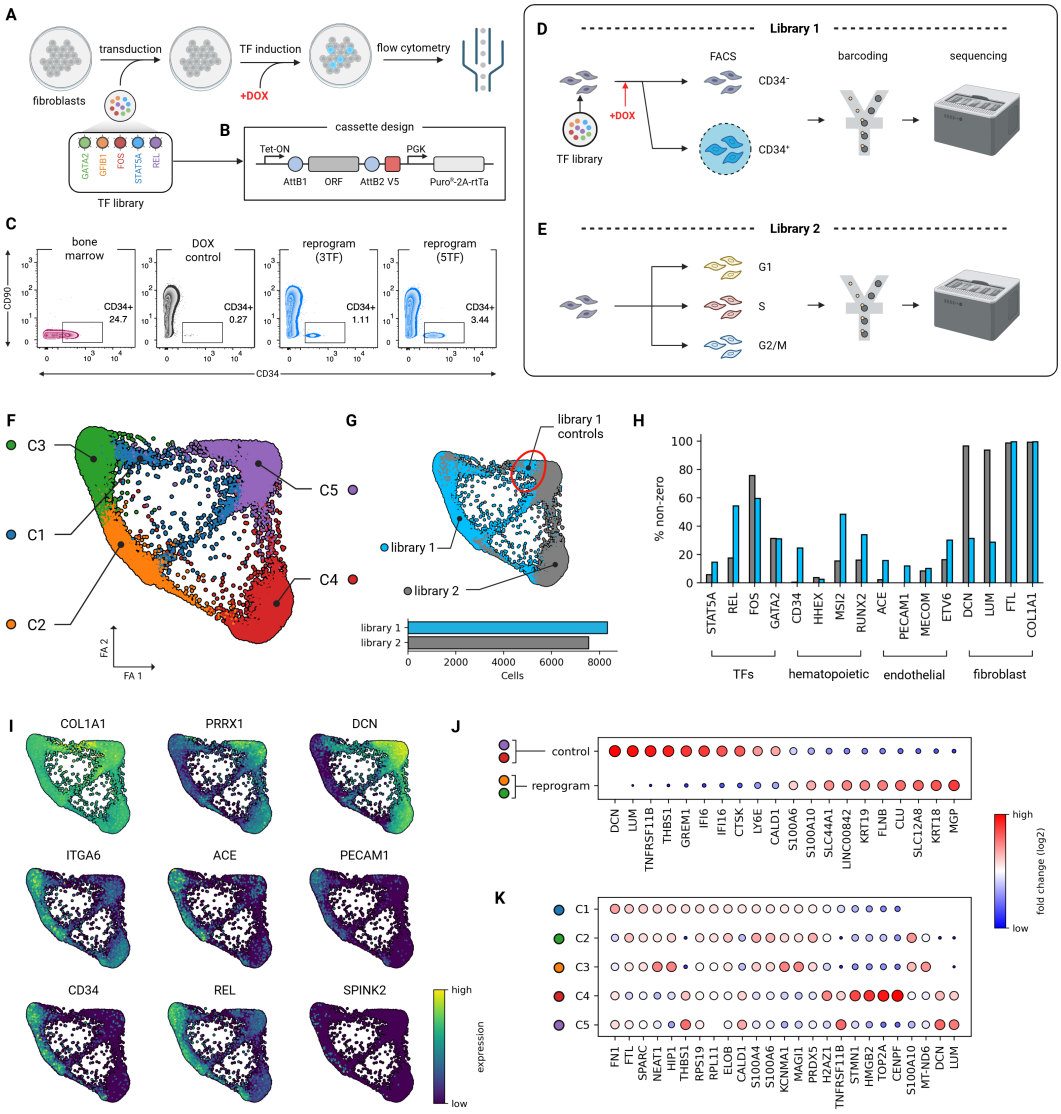

Figure 2: **Data-guided direct reprogramming of fibroblasts into CD34$^+$ cells. (A)** Lentiviral transduction of 5TF recipe, induction with doxycycline (+Dox), and quantification of CD34$^+$ cells using flow cytometry. **(B)** Design of inducible TF expression cassette. **(C)** CD34/CD90 flow cytometry plots for bone marrow, uninduced fibroblast controls, and reprogrammed fibroblasts with 3TF or 5TF recipes. **(D)** Library 1: CD34$^+$ cells are isolated post-reprogramming with 5TF recipe and profiled using 10x barcoding and nanopore sequencing. **(E)** Library 2: uninduced fibroblasts are sorted by cell cycle phase (G1, S, G2/M) and sequenced. **(F)** Force-directed graph layout of single cells colored by Leiden cluster (C1–C5). **(G)** Graph from (F) colored by library. Cell counts shown below. **(H)** Fraction of cells with nonzero expression for key TFs, hematopoietic, endothelial, and fibroblast genes in cells from both libraries. **(I)** Force-directed graphs colored by normalized expression of selected fibroblast, endothelial, and hematopoietic genes. **(J)** Dot plot showing differentially expressed genes (Wilcoxon Rank-sum test, corrected $\alpha = 0.05$) between control (C4/C5) and reprogrammed (C2/C3) cells. **(K)** Cluster-resolved dot plot highlighting expression of differentially expressed genes (Wilcoxon Rank-sum test, corrected $\alpha = 0.05$) across clusters.

2.3 Pseudotime reveals emergence of endothelial signature

Intermediate states in direct reprogramming of fibroblast to HSC remain poorly defined. Each TF recipe configuration may produce intermediate states along a novel reprogramming trajectory, whose phenotypes differ substantially from previously described cell states [22, 31]. To characterize the degree to which our induced cells displayed divergence from their initial state, we analyzed gene expression changes along the 5TF-induced reprogramming trajectory using two trajectory inference methods: diffusion-based pseudotemporal ordering [32] and RNA velocity estimation [33, 34].

First, we sought to establish a diffusion-based pseudotemporal ordering for cells from Library 1 using unperturbed cells from Library 2 as a reference for the initial state (Figure 3A-C, Methods 4.5.2). We used the mean pseudotime from cell-cycle phase assignments (G1, S/G2M) from Library 2 to measure reprogramming divergence from the initial state (Figure 3B-C, Supplementary Figures 21, 17, 18, 19, Supplementary Table 6, Methods 4.5.3). Interestingly, we observed that markers of cell-cycle arrest were elevated in clusters C1-C3 compared to C4 and C5—potentially indicative of cell-cycle exit (Supplementary Figure 20).

Transcriptomic patterns along the reprogramming trajectory were assessed using Pearson correlation between pseudotime and gene expression. Genes with positive correlation (upwards trend in reprogramming) included endothelial-associated transcripts such as *KRT18*, *CLU*, and *FLNB*, as well as genes involved in cell adhesion and matrix organization (*MGP*, *SLC44A1*, and *MAGI1*, Figure 3E). In contrast, negatively correlated genes were enriched for cell-cycle associated function, including *STMN1*, *MKI67*, *TOP2A*, and *DUT*, potentially suggesting a downregulation of fibroblast proliferative programs during reprogramming (Figure 3E-G). Further, expression of chromatin remodelers and proliferation-related TFs such as *STAT1*, *HMGB2*, and *EZH2* also decreased with pseudotime (Supplementary Figure 22). Interestingly, several TFs with known roles in stem cell maintenance and pluripotency (*SMAD3*, *RBFOX2*, and *BCL3*) showed increased expression over pseudotime, consistent with direct reprogramming towards a hemogenic-like cell state (Supplementary Figure 22).

Next, we inferred unspliced and spliced mRNA counts from our long-read scRNA-seq data using `velocyto` and used the their ratio to estimate RNA velocity using `scVelo` (Supplementary Figure 23A, Methods 4.5.4) [33, 34]. As expected, velocities showed reprogrammed cell clusters C2/C3 transitioning away from the initial cell states (Figure 3H). Next, we identified genes underlying cell state transitions based on their velocity estimates (Supplementary Figure 23 23D-E), finding that RNA velocity analysis were largely consistent with our findings from pseudotime (Supplementary Figure 23 23F-H). For example, genes with strong relationships with the inferred reprogramming trajectory included endothelial transcripts such as *CD34*, *ACE*, and *ITGA6* (Figure 3I).

Together, these results demonstrate divergence of the reprogrammed cells from the initial fibroblast cell type and reveal a partial acquisition of an endothelial-like intermediate state.

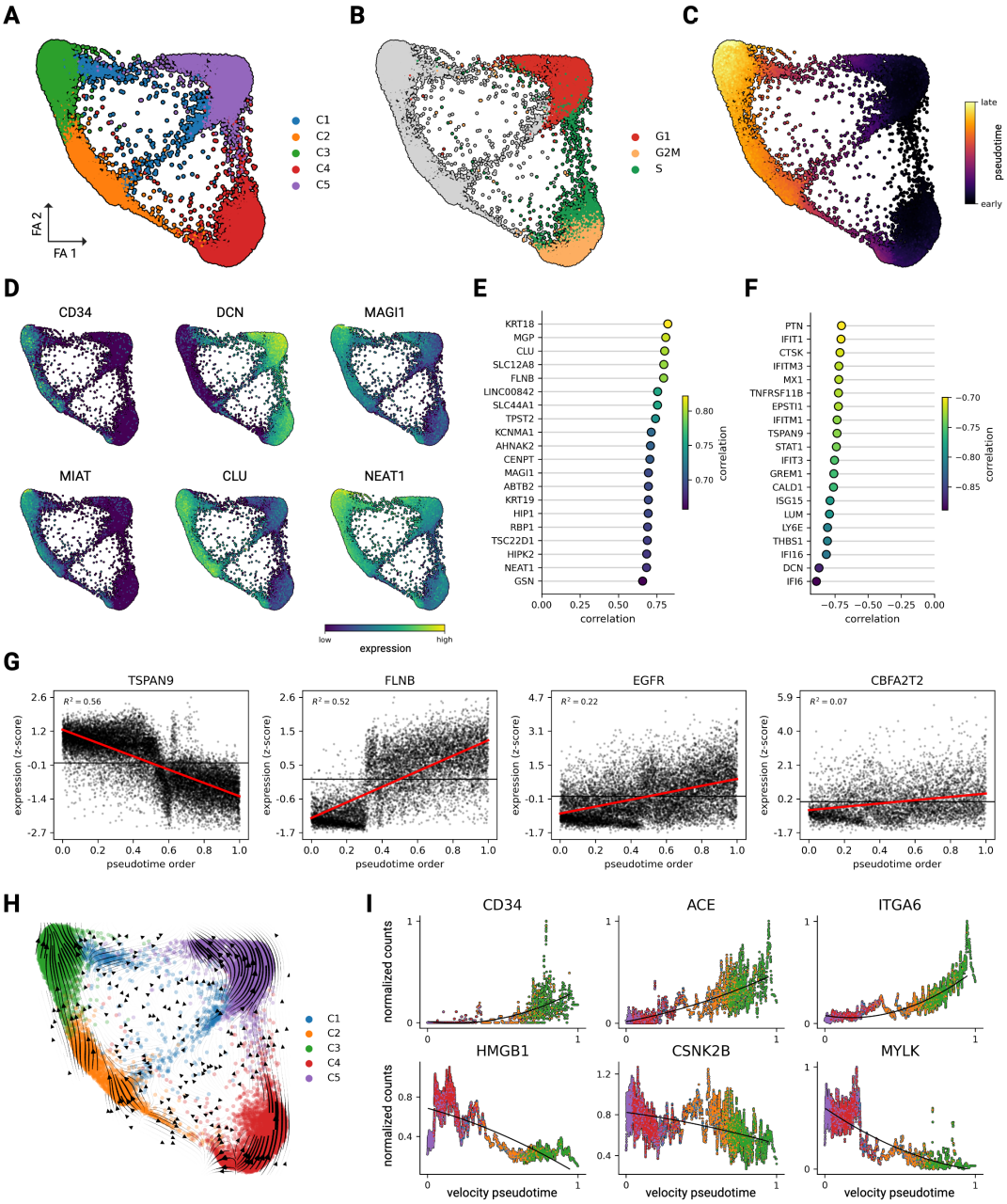

Figure 3: **Cell state transitions and gene expression dynamics along pseudotime. (A)** Force-directed layout of the k-nearest neighbor (kNN) graph colored by Leiden cluster assignment. **(B)** Same layout colored by cell cycle phase (G1, S, G2/M). **(C)** Cells colored by pseudotime. **(D)** Expression of selected genes. **(E)** Top genes positively correlated (Pearson) with pseudotime. **(F)** Top genes negatively correlated (Pearson) with pseudotime. **(G)** Correlation of selected gene expression (Z-score) with pseudotemporal ordering, Pearson correlation coefficient shown. **(H)** Low-dimensional projection of RNA velocity estimates. **(I)** Selected gene expression shown over velocity pseudotime.

2.4 Benchmarking against native and engineered cell types

Despite clear divergence from the initial cell state, the degree to which reprogrammed cells resemble the target HSC cell type remains incompletely characterized [10, 35]. To address this, we generated a third long-read scRNA-seq library (Library 3) from CD34$^+$ human bone marrow cells (Figure 4A, Methods 4.3, 4.4). We hypothesized that transcriptionally distinct populations previously identified in our reprogramming data (C2/C3) would display variable similarity to native cell types in bone marrow.

Clustering of cells from Library 3 revealed expected cell types such as HSCs (CD34$^+$/CD38$^-$) and several lineage-restricted progenitor populations, consistent with previously published bone marrow scRNA-seq data (Supplementary Figure 24) [36]. To compare cells across all three libraries, we constructed a harmonized latent space using scVI, controlling for the effects of the cell cycle (Figure 4B, Methods 4.5.5) [27]. We constructed a uniform manifold approximation and projection (UMAP) embedding from the learned latent space, which showed graded CD34 expression across the three libraries, with the highest expression in bone marrow populations (Figure 4C) [37]. We observed that endothelial genes such as *ACE* were mostly highly enriched in reprogrammed cells and largely absent from bone marrow (Figure 4C).

To characterize cell-cell differences, we computed mean pairwise Euclidean distances between cells in clusters C2/C3, and both the initial cells (C4/C5) and target cells (CD34$^+$/CD38$^-$ HSCs) using the learned latent space (Figure 4D). C3 cells showed larger average distances to the initial cell type and lower distances to HSCs compared to C2 cells, yet remained closer to the initial state than to the target state, consistent with partial reprogramming. Despite being more similar to initial cell populations, reprogrammed cells did not clearly align with fibroblasts or cell types in bone marrow (Figure 4B).

To investigate this further, we placed reprogrammed cells in a broader context by integrating them with native hematopoietic cell types, fibroblast, and endothelial cell types from the reference atlas. In addition, we included engineered scRNA-seq profiles from Ng et al. [2] and from Gomes et al. [10] (Figure 4E-G, Methods 4.5.7). Leiden clustering of harmonized data revealed substantial overlap between similar cell types from different data sources (Figure 4F, Supplementary Table 8). As expected, bone marrow cells and HSCs from Library 3 aligned most closely with native hematopoietic cell types in the reference (Figure 4F-G). Consistent with previous findings, reprogrammed cells aligned similarly with fibroblast and endothelial cell types (Figure 4E-G, Supplementary Figure 25A-B). However, reprogrammed cells were also distinct from native cell types in terms of their cell-cell distance profiles (Supplementary Figure 25A-B).

Finally, we performed a similar analysis to published human hematopoietic scRNA-Seq datasets (Figure 4G–H, Supplementary Figure 26, Supplementary Table 7, Methods 4.5.8), revealing diverse similarity profiles across a range of protocols and hematopoietic cell types. Reprogrammed cells most closely resembled multipotent progenitors and endothelial/stromal cells, and differed most from more differentiated immune cell types. Across all comparisons, our cells aligned most closely with day 25 3TF-reprogrammed cells (Figure 4H–I) [10]. Application of `Capybara` [22] to compare our single-cell data with bulk RNA-seq from Gomes *et al.* further supported that clusters C2/C3 had high similarity to engineered cell types (Figure 4I, Methods 4.5.9).

These results indicate reprogrammed cells occupy a transcriptomic space that is distinct from the initial and target cell types, and also from native cell types along the EHT spectrum.

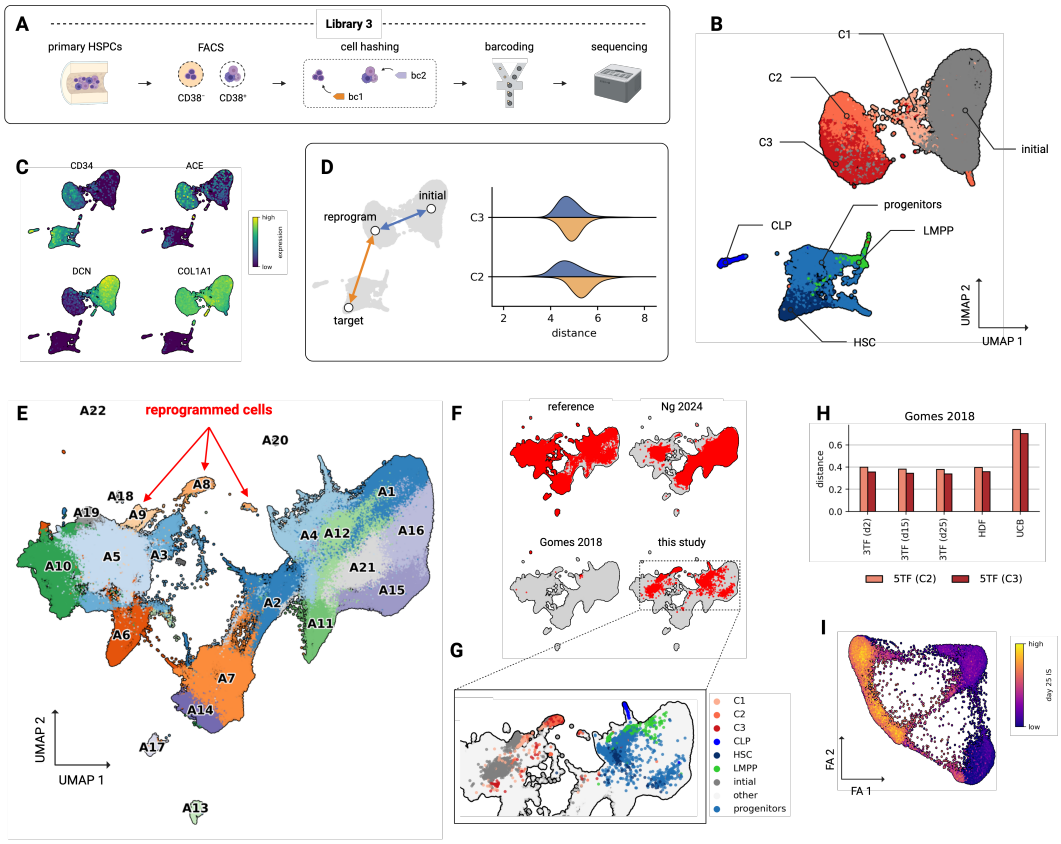

Figure 4: **Benchmarking data-guided direct reprogramming.** **(A)** Experimental workflow for single-cell RNA-seq of human bone marrow cells (Library 3). **(B)** UMAP embedding of bone marrow cells, reprogrammed cells and fibroblasts. **(C)** Expression of key marker genes in (B). **(D)** Top: schematic of cell–cell distances; bottom: distributions of mean Euclidean distances from reprogrammed cells to the target and initial cell types. **(E)** UMAP of 420,366 cells from reference atlas and selected sources based on scVI latent space, shown with Leiden clusters (resolution = 0.55). **(F)** Source of cells in (E). **(G)** Zoomed in view of (E) with cell an notions from part (B). **(H)** Mean Euclidean distance to pseudobulk signatures from [10]. **(I)** D25 identity score projected over UMAP coordinates (Fig. 2F).

2.5 Diversity of alternative splicing in direct reprogramming

Long-read sequencing can be used to quantify transcript isoforms produced by alternative splicing (AS) of a gene [38–40]. AS is observed during reprogramming to pluripotency [41–43] and plays an important role in fate specification during hematopoiesis and in the function of mature hematopoietic cells [44–48]. However, isoform expression remains largely uncharacterized in direct reprogramming.

Given the link between AS and cell identity, we hypothesized that changes in isoform diversity would accompany fibroblast to HSC reprogramming. For each of the three libraries, counts were obtained for the transcripts whose cell barcodes passed QC at the gene expression level (Methods 4.5.10). As expected for mammalian cells, analysis of these transcripts showed that exon skipping was the most common AS event across all three libraries (Figure 5A, Methods 4.5.11) [49, 50].

Next, we quantified the number of detected transcripts per gene as a measure of transcript-level diversity in each cell type (Figure 5B). Fibroblasts and lineage-restricted hematopoietic progenitors displayed the highest fraction of multi-isoform genes (Figure 5A). Reprogrammed cells (C2/C3) showed differences in isoform diversity relative to the initial fibroblast state, with C3 showing greater, though limited, overlap with HSCs (Figure 5B). The greatest overall difference in transcript diversity was observed between initial and target cells (6,753 genes).

To uncover differences in splicing patterns, we identified genes whose expression was dominated by a specific transcript and annotated the associated transcript as the 'major isoform.' Consistent with similarity of gene expression, reprogrammed cells and initial cells shared many major isoforms, suggesting similar regulation of AS (Figure 5C). However, we also identified genes whose major isoform differed by cell type (Figure 5D). Notably, we identified endothelial-associated genes *FTL1*, *MECOM*, *SOX18*, and *EMCN*, and genes implicated in the Hippo-signaling pathway (TEAD-family TFs, *YAP1*) (Figure 5D).

Next, we computed the proportion of expression attributed to each isoform by cell type (Figure 5E, Supplementary Figure 28). We observed differences in isoform expression for genes such as *TEAD4*, *SOX6*, and *ETV4* (Figure 5E). For example, initial cells predominantly express *SOX6-218* (UniProt: P35712-2) while reprogrammed and target cells show elevated expression of *SOX6-213* (UniProt: E9PQ78), which is predicted to undergo nonsense-mediated decay.

At the gene-level, *TEAD4* was not differentially expressed between initial and reprogrammed cells (Figure 5E). However, reprogrammed cells showed elevated expression of *TEAD4-207*, while initial cells predominantly expressed the Ensembl canonical isoform *TEAD4-202* (Figure 5F). *TEAD4-202* and *TEAD4-207* encode for different proteins with significant structural differences (Figure 5G, Methods 4.5.12). Notably, the last 4 exons of *TEAD4-202* encode the YAP binding domain, which are truncated in *TEAD4-207* (Figure 5G-H).

These isoform results highlight the additional complexity in transcriptional regulation beyond gene expression and support that alternative splicing could contribute to heterogeneity in cell reprogramming outcomes.

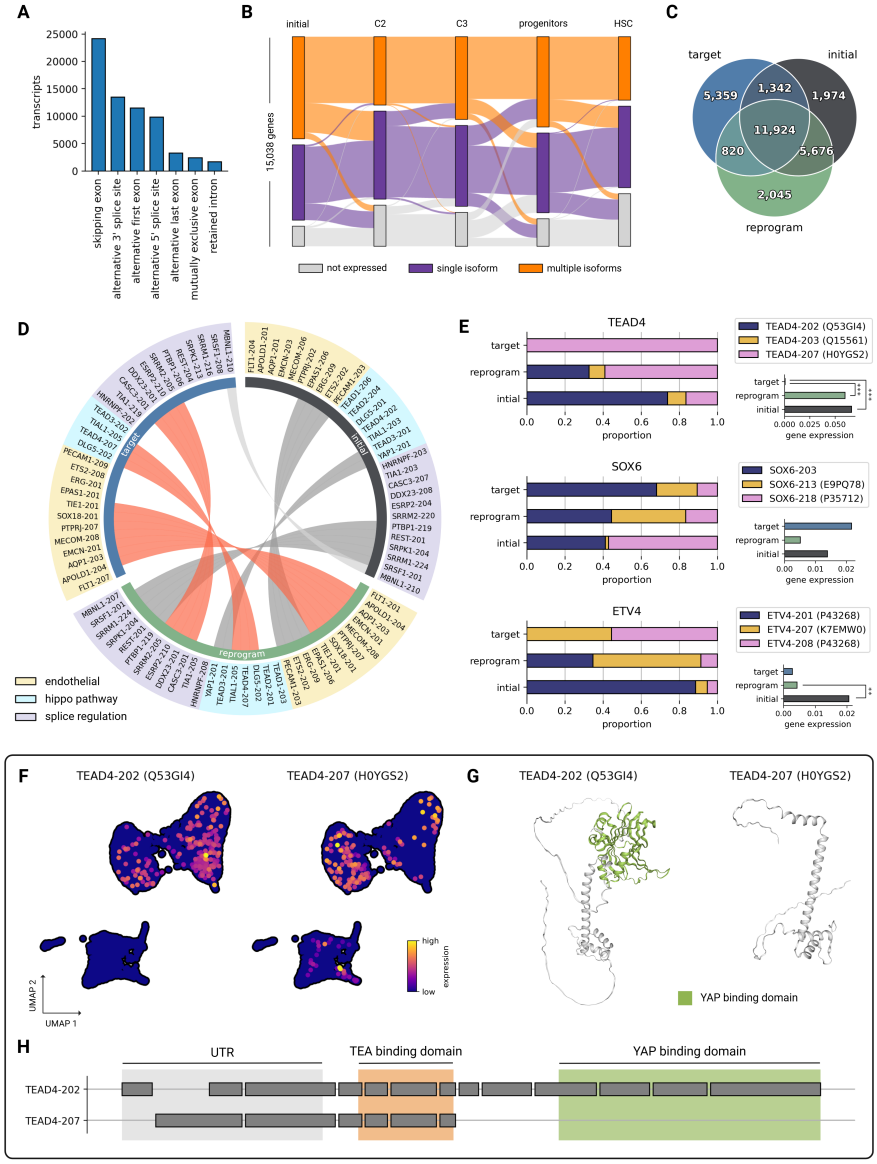

Figure 5: **Alternative splicing across fibroblast to HSC reprogramming.** **(A)** Count of pooled transcripts by splicing event. **(B)** Sankey diagram showing changes in isoform usage across reprogramming stages. **(C)** Overlap of shared dominant isoforms between cell types. **(D)** Isoform switching events. Links between transcripts signify shared dominant isoforms between cell groups. **(E)** Proportion of total expression for each detected isoform per group. Bottom right bar plots are mean log-normalized expression values at the gene level. Asterisks represent significance of differential expression (Wilcoxon Rank-sum test, corrected $\alpha = 0.05$; ** adjusted p-value < 0.01, *** adjusted p-value < 0.001). **(F)** UMAP of TEAD4 isoform expression. **(G)** AlphaFold structural predictions of TEAD4 isoforms. **(H)** Exon structure of TEAD4-202 and TEAD4-207.

3 Discussion

We tested a novel algorithmically-predicted TF recipe for fibroblast-to-HSC conversion. We systematically evaluated single-cell transcriptomic profiles of intermediate cell states, revealing partial acquisition of an endothelial-like phenotype and previously uncharacterized patterns in gene and isoform expression. Reprogrammed cells showed transcriptomic shifts that distinguish them from the initial fibroblast cell type, but remained distinct from native HSCs and related hematopoietic cell types. Our study clarifies both the capabilities and current limitations of TF-driven direct reprogramming protocols for generating autologous HSCs.

To our knowledge, the 5TF recipe is novel for fibroblast-to-HSC direct reprogramming. Among these, GATA2, GFI1, and FOS are established factors [10]. GATA2 is required for definitive hematopoiesis, and both GATA2 and GFI1B function as chromatin bookmarks for key hematopoietic regulators [11, 51]. Of the novel factors, STAT5A is a critical regulator of hematopoiesis, with elevated expression in HSCs and a well-established role in promoting self-renewal and expansion of hematopoietic progenitors, both in vitro and in human $CD34^+$ cord blood cells [52, 53]. The potential function of REL in reprogramming is less understood. Members of the NF-κB/Rel family, including REL, are key regulators of anti-apoptotic gene expression in hematopoietic cells [54]. REL is highly expressed in lymphocytes [55] and promotes anti-apoptotic genes such as *Bcl-2* and *Bcl-xL*. Inhibition of REL increases apoptosis and impairs differentiation and expansion of lymphoid lineages derived from cord blood, underscoring its role in cell survival during hematopoietic differentiation [56]. Together, STAT5A and REL may act as complementary regulators supporting long-term survival of reprogrammed cells in vitro. However, the context-dependent role of these TFs in reprogramming require further elucidation.

Studies in mouse [57, 58] and human cells [10, 11, 59] support that reprogramming to HSC-like cells proceeds through an endothelial-like intermediate. Gata2, cFos, Etv6, and Gfi1b induction in mouse fibroblasts generates an endothelial precursor marked by $Prom1^+Sca\text{-}1^+CD34^+CD45^-$ [60]. Another study found that reprogrammed murine cells express embryonic hemoglobin (*b-H1*) and show a peak in endothelial gene expression, including *Pecam1* and *Vwf* early in reprogramming [57]. Human cells transduced with the 3TF recipe also upregulate endothelial genes, including *PPARG*, *VWF*, and *FOXC2*, leading to the generation of a population of $CD49f^+ACE^+$ cells [10]. In line with these observations, our reprogrammed cells show elevated expression of *ACE*, *ITGA6*, and *PECAM1*, supporting the existence of an endothelial-like intermediate during reprogramming.

In development, HSCs arise via the endothelial-to-hematopoietic transition (EHT) [12, 61]. Previous studies have connected Hippo signaling to early hematopoietic specification. Goode et al. [62] implicated the Hippo effectors YAP1 and TEAD family TFs in murine EHT, consistent with Hippo pathway activation. Inhibiting the YAP–TEAD interaction impaired the generation of $CD41^+$ precursors before, but not after, specification, suggesting that TEAD/YAP activity is required to initiate hematopoietic fate. Further, Nakajima et al. demonstrated YAP1 activity in endothelial cells under EHT-like mechanical stimulation [63]. These findings suggest that YAP integrates mechanical and developmental signals to modulate hematopoietic output.

Both gene and isoform-level analyses support a role for Hippo signaling in HSC reprogramming. In our reference atlas, Hippo pathway genes are most highly expressed in mesenchymal and endothelial populations (Supplementary Tables 9, 10). In our data, reprogrammed cells exhibit intermediate levels of Hippo gene expression, including reduced expression of *TEAD4* and *YAP1* relative to the initial cells (Supplementary Table 11). We also observed shifts in dominant isoform usage for TEAD-family TFs and related genes such as *FTL1* and *ESRP2* in reprogrammed cells

(Figure 5D). *ESRP2*, a splicing regulator of *TEAD1* and other Hippo-related genes (e.g., *NF2*, *YAP1*, *CSNK1D*), have been shown to coordinate a neonatal-to-adult splicing shift that rewires Hippo signaling and restricts YAP/TAZ activation [64, 65]. Together, these results support a relationship between AS and Hippo pathway activation during reprogramming.

Reprogramming to pluripotency involves distinct, sequential phases of splicing regulation [41–43, 66], and isoform switching of TFs has been shown to influence reprogramming outcomes in mice [67, 68]. To our knowledge, this study is the first to characterize direct reprogramming in human systems using long-read scRNA-seq. We observed cell-type-specific variability in isoform expression suggesting that isoform-level analysis can provide finer resolution for distinguishing cell states.

Despite acquisition of $CD34^+$ surface expression and marked transcriptional changes, reprogrammed cells retained expression of key fibroblast markers, such as *COL1A1* and *FTL* [30]. Complete reprogramming may require overcoming barriers in cell state transitions, including both local stability and maintenance of the new identity [69–71]. Quantifying intermediate or hybrid cell states remains a key challenge [22, 31]. Computational tools such as `Capybara` [22] enable partial assignment of transcriptomic profiles to single cells, supporting estimation of hybrid identities; however, their accuracy is limited by the availability of suitable reference data. Our approach parallels recent strategies developed for engineered neural organoids [24]. Future studies may assess the applicability of emerging biological foundation models [23, 25, 72, 73], though their capacity to represent engineered 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 reveal opportunities to overcome them. Overall, our findings contribute to the broader 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` [74] and/or `rapids-singlecell` [75]. 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 single-cell variation inference (scVI) [27] implemented in the `scvi-tools` Python library [76]. 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 [77]. 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-distributed stochastic neighbor embedding (t-SNE) embedding based on the `scANVI` model latent space.

To identify differentially expressed genes (DEGs), we used `Scanpy` `sc.tl.rank_genes_groups` using the Wilcoxon rank-sum test 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 Algorithmically Guided Recipe subsection

TFs for fibroblast-to-HSC reprogramming were selected as described in [26]. TFs were identified by minimizing the distance between the initial and target cell states using the Data-Guided Control (DGC) algorithm [26]. The DGC model is based on 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 $\mathbf{u}$.

4.2.2 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. FH3 and FH6 were used for flow cytometry experiments shown in Figure 2C, while FH4a was used for transductions in all other experiments. AllPrep DNA/RNA/Protein Mini Kits (Qiagen 80004) were used to process cells for testing genome integration and protein and gene expression.

FH3	FH4a	FH6
GATA2	GATA2	GATA2
GFI1B	GFI1B	GFI1B
FOS	FOS	FOS
	REL	REL
	STAT5A	STAT5A

Table 1: **TF Recipes.**

4.2.3 Cell Culture

All cells were cultured at 37 °C with 5% CO_2. Neonatal foreskin fibroblasts (BJ, ATCC CRL-2522, RRID: CVCL _3653) were used as the initial cell type for reprogramming. 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. For reprogramming cultures, plates were coated with 0.1% porcine gelatin (Millipore #SF008). Fibroblasts were cultured in high glucose Dulbecco's Modified Eagle Medium (DMEM; Gibco 11965092) + 10% fetal bovine serum (FBS; Corning 35-015-CV) + 1% penicillin/streptomycin (P/S; Gibco 15140-122) + 1% L-glutamine. Lentiviral transductions were carried out in complete medium without P/S in the presence of 8 μg/mL polybrene (Millipore Sigma TR-1003-G). Integrated TF expression cassettes were induced with 1 μg/mL doxycycline hyclate (Dox, Millipore Sigma D9891) supplemented in the culture medium for 48 hours. Control fibroblasts were untransduced for single cell transcriptomics experiments. Reprogramming cells were switched to complete MyeloCult H5100 (STEMCELL Technologies 05150) with 10e-6 M hydrocortisone (Millipore Sigma #H0888) and 1% P/S for 48 hours after the initial Dox treatment. Normal human primary human bone marrow $CD34^+$ cells were purchased from ATCC (PCS-800-012TM Lot 80516705) and cultured in complete Myelocult H5100 for 12 - 16 hours prior to analysis. Adherent cells were dissociated using 1x TrypLE^{TM} Express Enzyme (Gibco #12604-021).

4.2.4 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 incubated with primary antibodies or LIVE/DEAD stain for 30 minutes at 4°C in Hank's balanced 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 and human primary $CD34^+$ bone marrow cells were used to aid gating for iHSCs.

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. Veri-Cells CD34 PBMCs (Biolegend Cat# 426901 Lot B375717) were reconstituted following the manufacturer's protocol and used as an additional staining control during cell sorting. FlowJo v10.10.0 (Becton Dickinson) was used for flow cytometry analyses.

4.2.5 Immunofluorescence

For immunofluorescent staining, 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.

4.3 Bone Marrow

Human bone marrow $CD34^+$ cells were purchased from STEMCELL Technologies (Catalog #70002), containing 1×10^6 cells. Cells were thawed by adding pre-warmed DMEM supplemented with 10% FBS and prepared according to the manufacturer's protocol.

4.3.1 Immunolabeling and Cell Sorting

DAPI was used as a viability marker. Veri-Cells CD34 PBMCs (BioLegend Cat#426901, Lot B375717) were reconstituted following the manufacturer's instructions and served as staining controls. Approximately 3×10^5 Veri-Cells were used for full staining and fluorescence-minus-one (FMO) controls. UltraComp eBeads (Invitrogen Cat#01-2222) were used for single-color compensation. Approximately 1 million bone marrow $CD34^+$ cells 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.

4.3.2 Feature Barcoding and Cell Hashing

Sorted cells were re-blocked with Fc receptor block for 10 minutes on ice and incubated with mixtures of hashtag and feature-barcoded 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. All bone sorted marrow cells were labeled with TotalSeq-B0054, -B0060, -B0594, -B0063, and -B0070 for surface marker labeling. 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.

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

4.4 Single-Cell Transcriptomics

Full-length barcoded cDNA amplicons were obtained from 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 (GEMs). 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 2). Briefly, cDNA amplicons were biotin tagged by PCR with the primers shown in Table 2. Biotinylated cDNA was purified with Dynabeads^{TM} 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 2 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 fibroblasts alone (Library 2) and from $CD34^+$ human bone marrow cells (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 BioLegend TotalSeq-B human hashing antibodies (Cat# 394631, #394603, and #394605, respectively) before pooling and single cell barcoding (viability >97%). For Library 3, approximately 8,000 hashed and feature barcoded $CD34^+$ cells (6,400 $CD34^+/CD38^-$ and 1,600 $CD34^+/CD38^+$) were used for single cell barcoding (viability 72%). Transcript and hashing libraries were prepared for sequencing as described above and sequenced on an ONT P2 sequencer.

4.5 scRNA-seq Data Analysis

4.5.1 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 consistent quality control using `Scanpy` [74]. Cells with fewer than 1,000 detected genes or fewer than 500 total UMI counts were filtered out. Mitochondrial and ribosomal genes were annotated for downstream analysis. Unique UMI counts were normalized to 10,000 counts per cell and log-transformed using `log1p`.

4.5.2 Cell Cycle Phase Assignment

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` [78], with parameters `-e 1 -x '' -b ???????????????? -u '' -x ''`. Corrected barcode-hash sequence pairs were quantified and aggregated per cell. The hash-sequence with the most read-pairs per cell barcode was taken as the barcoded phase for each cell.

To predict cell cycle phases, we used established gene signatures from [79]. Cells were scored for S and G2/M programs using the `score_genes_cell_cycle()` function in `Scanpy`. Predicted phases were compared to DNA content-based phase assignments obtained from the hashing barcodes that were added after cell sorting (Supplementary Table 6).

4.5.3 Pseudotemporal Inference

To quantify transcriptional divergence from the initial state after induction, we used diffusion-based pseudotime [32]. Representative root cells were selected from the control cell population. We chose root cells from both G1 and G2/M based on minimum Euclidean distance from the centroid of the respective phase group in Harmony-corrected PCA space [80]. A diffusion map was constructed on the Harmony-corrected PCA space (`rsc.tl.diffmap`), and a k-nearest neighbor graph was constructed from the resulting components. For each root, pseudotime trajectories were inferred using diffusion pseudotime (`sc.tl.dpt`), 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 cell cycle, gene expression was correlated with aggregated pseudotime. First, we computed the z-score of log-normalized expression across cells. For each gene, correlation coefficients and associated p-values were calculated relative to pseudotime. Multiple testing correction was performed using the Benjamini-Hochberg method to control the false discovery rate [81].

4.5.4 RNA Velocity Inference

We developed a `Snakemake` workflow to preprocess BAM files for RNA velocity analysis with `velocyto` [33]. 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 and 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` [34]. First, we computed a a nearest-neighbour 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`), and per-cell velocities were calculated with `scv.tl.velocity`. 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`.

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

To harmonize data across al three long-read scRNA-seq libraries, we used `scVI` [27] from `scvi-tools` [76]. Processed UMI count matrices were concatenated and cell cycle scores we computed using the `` `score_genes_cell_cycle()` `` function in `Scanpy` with the gene set from [79]. Next, 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. Training was performed on an an NVIDIA Tesla V100 or A100 with early stopping for up to 500 epochs. Next, we performed principal component analysis (PCA) on the scVI latent space, and computed a neighborhood graph using 35 neighbors. Cluster connectivity was inferred using `scanpy.tl.paga` [82], grouping cells by their annotated labels. The resulting Partition-based Graph Abstraction (PAGA) coordinates were used to initialize a 2-dimensional uniform manifold approximation and projection (UMAP) embedding with parameters `spread=0.55` and `min_dist=0.2`.

4.5.6 Reference Atlas Benchmarking

Pre-processed data from the reference atlas (Methods 4.1) and experimental Libraries 1–3 (Methods 4.5.1) were combined. Pseudobulk profiles were obtained by summing over raw UMI counts for each cell-type. Aggregated counts were normalized using `normalize_total` with `target_sum`=10^4 and log-transformed using `log1p`. Cell-cycle phase scores were predicted for each pseudobulk profile as in Methods 4.5.2. We filtered genes expressed exclusively in the experimental data, as well as those annotated as mitochondrial, ribosomal, non-protein-coding, or cell-cycle related [79], yielding 16,131 genes over which we conducted downstream analysis.

To harmonize pseudobulk profiles, we used a single-cell variational inference model (scVI) [27] from `scvi-tools` [76] with S-phase and G2/M-phase scores as continuous covariates. The scVI model architecture was two fully connected hidden layers of 128 units each with a dropout rate of 0.25, and a 24-dimensional latent representation. Gene expression counts were modeled with a zero-inflated negative binomial (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=10), a learning rate of 1×10^{-3}, a batch size of 25, and a 90/10 train/validation split. The resulting latent space provided a harmonized, cell-cycle-aware embedding used in downstream analyses. Pairwise Euclidean distances were used to compare pseudobulk profiles in the latent space using `scipy.spatial.distance.pdist`.

4.5.7 Integrated Dataset Benchmarking

We compiled single-cell RNA-seq data from four sources: the reference atlas, Ng 2024 [2], Gomes 2018 [10], and this study, yielding 983,763 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 21,062 cells across eight 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, 521,348 cells and 17,601 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 (17,898 cells) and genes with low or zero counts (227 genes). Cells with fewer than 1,000 detected genes were also removed. The final dataset contained 420,366 cells and 17,374 genes, which were normalized to 10,000 total counts per cell and log-transformed.

Single-cell profiles were harmonized following the approach in Methods 4.5.6. Highly variable genes were selected using the `seurat_v3` method, retaining the top 3,500 features. A `scVI` model was trained to correct for batch effects using the data source as the batch key and including continuous covariates for cell cycle phase (S and G2M scores). The model architecture included 24 latent dimensions, 128 hidden units, and two encoder/decoder layers. Training was performed on GPU with early stopping (patience = 30, max epochs = 500). Model performance was monitored by tracking training and validation loss, evidence lower bound (ELBO), reconstruction error, and Kullback–Leibler (KL) divergence.

4.5.8 Curated Dataset Benchmarking

We benchmarked our scRNA-seq data on six publicly available single-cell RNA-seq datasets (Supplementary Table 7). Briefly, we obtained 359,848 single cells, and a broad spectrum of 71 annotated cell types from [2, 10, 19, 21, 36, 83]. For each dataset we aggregated raw UMI counts by the source-defined cell-type labels to create pseudobulk profiles as in (Methods 4.5.6). Only genes whose mean expression lay in the upper quartile in both the test dataset and the curated datasets were retained, yielding 15,022 shared genes across 74 pseudobulk profiles.

Pseudobulk profiles were harmonized as in Methods 4.5.6. To compute the abstract connectivity graph of cell clusters, we applied PAGA [82] using `rapids_singlecell` and `scanpy`. We computed the neighborhood graph with `rsc.pp.neighbors`, using 15 neighbors and the top 50 principal components from the scVI latent embedding. PAGA analysis was performed using `sc.tl.paga`, grouping cells by their cell type annotations. Visualization was generated using `sc.pl.paga`. Pairwise Euclidean distances were used to compare pseudobulk profiles in the latent space using `scipy.spatial.distance.pdist`.

4.5.9 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 4.5.8. 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.

4.5.10 Isoform Identification

Transcript-level unique molecular identifier (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` [84]. 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 21,062 cells across three libraries and 69,509 detected transcripts.

4.5.11 Splicing Event Annotation

Alternative splicing events were annotated with `SUPPA` using the GRCh38_2024-A gene annotation [49]. 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 3).

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 3: **Alternative splicing event types.**

4.5.12 Protein Structure Predictions

Protein structure predictions predictions for TEAD4-202 and TEAD4-207 were obtained from the Human Protein Atlas version 24.0 (Alphafold v2.3.2). The YAP binding domain falls within a high-confidence region of TEAD4-202 (pLDDT > 90).

5 Author contributions

LAM and IR initiated and supervised the project, including setting overall goals and designing the experiments. WM, LAM, and JC processed cells and tissues and collected sequencing data. CMS, JC, 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), iReprogram, Inc, and Oxford Nanopore Technologies (ONT). We thank Lakmal Jayasinghe and the research and support teams at ONT for helpful discussions and technical assistance.

7 Code Availability

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

8 Data Availability

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

9 Conflicts of Interest Statement

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] Shernan G. Holtan et al. "Disease progression, treatments, hospitalization, and clinical outcomes in acute GVHD: a multicenter chart review". In: *Bone Marrow Transplantation* 57.10 (2022).

[2] Elizabeth S. Ng et al. "Long-Term Engrafting Multilineage Hematopoietic Cells Differentiated from Human Induced Pluripotent Stem Cells". In: *Nature Biotechnology* (2024).

[3] Thorsten M. Schlaeger et al. "A comparison of non-integrating reprogramming methods". In: *Nature Biotechnology* 33.1 (2015).

[4] Viola Volpato and Caleb Webber. "Addressing variability in iPSC-derived models of human disease: guidelines to promote reproducibility". In: *Disease Models & Mechanisms* 13.1 (2020).

[5] Haofei Wang et al. "Direct Cell Reprogramming: Approaches, Mechanisms and Progress". In: *Nature Reviews Molecular Cell Biology* 22.6 (2021).

[6] Lei Han et al. "Distinctive Clinical and Pathologic Features of Immature Teratomas Arising from Induced Pluripotent Stem Cell-Derived Beta Cell Injection in a Diabetes Patient". In: *Stem Cells and Development* 31.5-6 (2022).

[7] Afsaneh Yazdani Movahed et al. "Elimination of tumorigenic pluripotent stem cells from their differentiated cell therapy products: An important step toward ensuring safe cell therapy". In: *Stem Cell Reports* 20.7 (2025).

[8] Eva Szabo et al. "Direct Conversion of Human Fibroblasts to Multilineage Blood Progenitors". In: *Nature* 468.7323 (2010).

[9] Michael G. Daniel et al. "Making a Hematopoietic Stem Cell". In: *Trends in Cell Biology* 26.3 (2016).

[10] Andreia M. Gomes et al. "Cooperative Transcription Factor Induction Mediates Hemogenic Reprogramming". In: *Cell Reports* 25.10 (2018).

[11] Rita Silvério-Alves et al. "Hemogenic Reprogramming of Human Fibroblasts by Enforced Expression of Transcription Factors". In: *JoVE (Journal of Visualized Experiments)* 153 (2019).

[12] Michael G. Daniel et al. "Induction of Human Hemogenesis in Adult Fibroblasts by Defined Factors and Hematopoietic Coculture". In: *FEBS Letters* 593.23 (2019).

[13] Chuijin Wei, Pei Yu, and Lin Cheng. "Hematopoietic Reprogramming Entangles with Hematopoiesis". In: *Trends in Cell Biology* 30.10 (2020).

[14] Jun Ding et al. "Computational Tools for Analyzing Single-Cell Data in Pluripotent Cell Differentiation Studies". In: *Cell Reports Methods* 1.6 (2021).

[15] Gele Liu et al. "Advances in Pluripotent Stem Cells: History, Mechanisms, Technologies, and Applications". In: *Stem Cell Reviews and Reports* 16.1 (2020).

[16] Franziska Paul et al. "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors". In: *Cell* 163.7 (2015).

[17] Caleb Weinreb et al. "Lineage Tracing on Transcriptional Landscapes Links State to Fate during Differentiation". In: *Science* 367.6479 (2020).

[18] Sam Watcham, Iwo Kucinski, and Berthold Gottgens. "New Insights into Hematopoietic Differentiation Landscapes from Single-Cell RNA Sequencing". In: *Blood* 133.13 (2019).

[19] Danilo Pellin et al. "A Comprehensive Single Cell Transcriptional Landscape of Human Hematopoietic Progenitors". In: *Nature Communications* 10.1 (2019).

[20] Anna Maria Ranzoni et al. "Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis". In: *Cell Stem Cell* 28.3 (2021).

[21] Chen Weng et al. "Deciphering Cell States and Genealogies of Human Haematopoiesis". In: *Nature* 627.8003 (2024).

[22] Wenjun Kong et al. "Capybara: A Computational Tool to Measure Cell Identity and Fate Transitions". In: *Cell Stem Cell* 29.4 (2022).

[23] Christina V. Theodoris et al. "Transfer Learning Enables Predictions in Network Biology". In: *Nature* 618.7965 (2023).

[24] Zhisong He et al. "An Integrated Transcriptomic Cell Atlas of Human Neural Organoids". In: *Nature* 635.8039 (2024).

[25] Graham Heimberg et al. "A Cell Atlas Foundation Model for Scalable Search of Similar Human Cells". In: *Nature* 638.8052 (2025).

[26] Scott Ronquist et al. "Algorithm for Cellular Reprogramming". In: *Proceedings of the National Academy of Sciences* 114.45 (2017).

[27] Romain Lopez et al. "Deep Generative Modeling for Single-Cell Transcriptomics". In: *Nature Methods* 15.12 (2018).

[28] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data Using T-SNE". In: *Journal of Machine Learning Research* 9.86 (2008).

[29] THE TABULA SAPIENS CONSORTIUM. "The Tabula Sapiens: A Multiple-Organ, Single-Cell Transcriptomic Atlas of Humans". In: *Science* 376.6594 (2022).

[30] Oscar Franzén, Li-Ming Gan, and Johan L M Björkegren. "PanglaoDB: A Web Server for Exploration of Mouse and Human Single-Cell RNA Sequencing Data". In: *Database* 2019 (2019).

[31] Adam L. MacLean, Tian Hong, and Qing Nie. "Exploring Intermediate Cell States through the Lens of Single Cells". In: *Current Opinion in Systems Biology.* Mathematic Modelling 9 (2018).

[32] Laleh Haghverdi et al. "Diffusion Pseudotime Robustly Reconstructs Lineage Branching". In: *Nature Methods* 13.10 (2016).

[33] Gioele La Manno et al. "RNA Velocity of Single Cells". In: *Nature* 560.7719 (2018).

[34] Volker Bergen et al. "Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling". In: *Nature Biotechnology* 38.12 (2020).

[35] Ryohichi Sugimura et al. "Haematopoietic Stem and Progenitor Cells from Human Pluripotent Stem Cells". In: *Nature* 545.7655 (2017).

[36] Manu Setty et al. "Characterization of Cell Fate Probabilities in Single-Cell Data with Palantir". In: *Nature Biotechnology* 37.4 (2019).

[37] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv:1802.03426 [cs, stat]* (2020). arXiv: 1802.03426 [cs, stat].

[38] Vivien Marx. “Method of the Year: Long-Read Sequencing”. In: *Nature Methods* 20.1 (2023).

[39] Siyuan Wu and Ulf Schmitz. “Single-Cell and Long-Read Sequencing to Enhance Modelling of Splicing and Cell-Fate Determination”. In: *Computational and Structural Biotechnology Journal* 21 (2023).

[40] Andrey D. Prjibelski et al. “Accurate Isoform Discovery with IsoQuant Using Long Reads”. In: *Nature Biotechnology* 41.7 (2023).

[41] Sho Ohta et al. “Global Splicing Pattern Reversion during Somatic Cell Reprogramming”. In: *Cell Reports* 5.2 (2013).

[42] Alexander Kanitz et al. “Conserved Regulation of RNA Processing in Somatic Cell Reprogramming”. In: *BMC Genomics* 20.1 (2019).

[43] Claudia Vivori et al. “Dynamics of Alternative Splicing during Somatic Cell Reprogramming Reveals Functions for RNA-binding Proteins CPSF3, hnRNP UL1, and TIA1”. In: *Genome Biology* 22.1 (2021).

[44] Fang Wang et al. “Single-Cell Architecture and Functional Requirement of Alternative Splicing during Hematopoietic Stem Cell Formation”. In: *Science Advances* 8.1 (2022).

[45] Laura Mincarelli et al. “Single-Cell Gene and Isoform Expression Analysis Reveals Signatures of Ageing in Haematopoietic Stem and Progenitor Cells”. In: *Communications Biology* 6.1 (2023).

[46] Lu Chen et al. “Transcriptional Diversity during Lineage Commitment of Human Blood Progenitors”. In: *Science* 345.6204 (2014).

[47] Yapu Li et al. “A Splicing Factor Switch Controls Hematopoietic Lineage Specification of Pluripotent Stem Cells”. In: *EMBO reports* 22.1 (2021).

[48] Yimeng Gao, Radovan Vasic, and Stephanie Halene. “Role of Alternative Splicing in Hematopoietic Stem Cells during Development”. In: *Stem Cell Investigation* 5.0 (2018).

[49] Juan L. Trincado et al. “SUPPA2: Fast, Accurate, and Uncertainty-Aware Differential Splicing Analysis across Multiple Conditions”. In: *Genome Biology* 19.1 (2018).

[50] Charlotte J. Wright, Christopher W. J. Smith, and Chris D. Jiggins. “Alternative Splicing as a Source of Phenotypic Diversity”. In: *Nature Reviews Genetics* 23.11 (2022).

[51] Rita Silvério-Alves et al. “GATA2 Mitotic Bookmarking Is Required for Definitive Haematopoiesis”. In: *Nature Communications* 14.1 (2023).

[52] Jan Jacob Schuringa et al. “Enforced Activation of STAT5A Facilitates the Generation of Embryonic Stem–Derived Hematopoietic Stem Cells That Contribute to Hematopoiesis In Vivo”. In: *Stem Cells* 22.7 (2004).

[53] Jan Jacob Schuringa et al. “Constitutive Activation of STAT5A Promotes Human Hematopoietic Stem Cell Self-Renewal and Erythroid Differentiation”. In: *Journal of Experimental Medicine* 200.5 (2004).

[54] M. C. Turco et al. “NF-κB/Rel-mediated Regulation of Apoptosis in Hematologic Malignancies and Normal Hematopoietic Progenitors”. In: *Leukemia* 18.1 (2004).

[55] Thomas D. Gilmore and Steve Gerondakis. “The C-Rel Transcription Factor in Development and Disease”. In: *Genes & Cancer* 2.7 (2011).

[56] Shirin Fallahi et al. “Impact of C-rel Inhibition of Cord Blood-Derived B-, T-, and NK Cells”. In: *Journal of Immunotoxicology* 14.1 (2017).

[57] Kiran Batta et al. “Direct Reprogramming of Murine Fibroblasts to Hematopoietic Progenitor Cells”. In: *Cell Reports* 9.5 (2014).

[58] Raphael Lis et al. “Conversion of Adult Endothelium to Immunocompetent Haematopoietic Stem Cells”. In: *Nature* 545.7655 (2017).

[59] Michael G. Daniel, Ihor R. Lemischka, and Kateri Moore. “Converting Cell Fates: Generating Hematopoietic Stem Cells de Novo via Transcription Factor Reprogramming”. In: *Annals of the New York Academy of Sciences* 1370.1 (2016).

[60] Carlos-Filipe Pereira et al. “Induction of a Hemogenic Program in Mouse Fibroblasts”. In: *Cell Stem Cell* 13.2 (2013).

[61] Vincenzo Calvanese et al. “Mapping Human Haematopoietic Stem Cells from Haemogenic Endothelium to Birth”. In: *Nature* 604.7906 (2022).

[62] Debbie K. Goode et al. “Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation”. In: *Developmental Cell* 36.5 (2016).

[63] Hiroyuki Nakajima et al. “Flow-Dependent Endothelial YAP Regulation Contributes to Vessel Maintenance”. In: *Developmental Cell* 40.6 (2017).

[64] Sushant Bangru et al. “Alternative Splicing Rewires Hippo Signaling Pathway in Hepatocytes to Promote Liver Regeneration”. In: *Nature Structural & Molecular Biology* 25.10 (2018).

[65] Sunkyung Choi, Namjoon Cho, and Kee K. Kim. “The Implications of Alternative Pre-mRNA Splicing in Cell Signal Transduction”. In: *Experimental & Molecular Medicine* 55.4 (2023).

[66] Mihaela Zavolan and Alexander Kanitz. “RNA Splicing and Its Connection with Other Regulatory Layers in Somatic Cell Reprogramming”. In: *Current Opinion in Cell Biology*. Cell Nucleus 52 (2018).

[67] Hong Han et al. “MBNL Proteins Repress ES-cell-specific Alternative Splicing and Reprogramming”. In: *Nature* 498.7453 (2013).

[68] Cheng-Xu Delon Toh et al. “RNAi Reveals Phase-Specific Global Regulators of Human Somatic Cell Reprogramming”. In: *Cell Reports* 15.12 (2016).

[69] C. H. Waddington. *The Strategy of the Genes*. London: Routledge, 2014. ISBN: 978-1-315-76547-1.

[70] Hiroaki Kitano. “Biological Robustness”. In: *Nature Reviews Genetics* 5.11 (2004).

[71] James E. Ferrell. “Bistability, Bifurcations, and Waddington’s Epigenetic Landscape”. In: *Current Biology* 22.11 (2012).

[72] Jennifer E. Rood, Anna Hupalowska, and Aviv Regev. “Toward a Foundation Model of Causal Cell and Tissue Biology with a Perturbation Cell and Tissue Atlas”. In: *Cell* 187.17 (2024).

[73] Artur Szałata et al. “Transformers in Single-Cell Omics: A Review and New Perspectives”. In: *Nature Methods* 21.8 (2024).

[74] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. “SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis”. In: *Genome Biology* 19.1 (2018).

[75] Severin Dicks et al. *Scverse/Rapids_singlecell: V0.10.6*. Zenodo. 2024.

[76] Adam Gayoso et al. "A Python Library for Probabilistic Analysis of Single-Cell Omics Data". In: *Nature Biotechnology* 40.2 (2022).

[77] Chenling Xu et al. "Probabilistic Harmonization and Annotation of Single-cell Transcriptomics Data with Deep Generative Models". In: *Molecular Systems Biology* 17.1 (2021).

[78] Oliver Cheng et al. "Flexiplex: A Versatile Demultiplexer and Search Tool for Omics Data". In: *Bioinformatics* 40.3 (2024).

[79] Itay Tirosh et al. "Single-Cell RNA-seq Supports a Developmental Hierarchy in Human Oligodendroglioma". In: *Nature* 539.7628 (2016).

[80] Ilya Korsunsky et al. "Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony". In: *Nature Methods* 16.12 (2019).

[81] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995).

[82] F. Alexander Wolf et al. "PAGA: Graph Abstraction Reconciles Clustering with Trajectory Inference through a Topology Preserving Map of Single Cells". In: *Genome Biology* 20.1 (2019).

[83] Grigorios Georgolopoulos et al. "Discrete Regulatory Modules Instruct Hematopoietic Lineage Commitment and Differentiation". In: *Nature Communications* 12.1 (2021).

[84] Xueyi Dong et al. "The Long and the Short of It: Unlocking Nanopore Long-Read RNA Sequencing Data with Short-Read Differential Expression Analysis Tools". In: *NAR Genomics and Bioinformatics* 3.2 (2021).

[85] Alex H. M. Ng et al. "A Comprehensive Library of Human Transcription Factors for Cell Fate Engineering". In: *Nature Biotechnology* 39.4 (2021).