Data-guided direct reprogramming of human fibroblasts into the hematopoietic lineage

Cooper Stansbury[1, 2], Jillian Cwycyshyn[1, 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]The Michigan Institute for Computational Discovery and Engineering, University of Michigan, Ann Arbor, MI 48109 USA
[3]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48109 USA
[4]Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 USA
[*]To whom correspondence should be addressed (+1-734-904-3013; lindmuir@umich.edu; 1600 Huron Parkway, NCRC B520-3368, Ann Arbor, MI 48109).

November 14, 2024

Abstract

Direct reprogramming of human fibroblasts into hematopoietic stem cells (HSCs) shows promise for generating autologous cells for treatment of blood and immune-related diseases. However, low conversion efficiency of existing protocols points to gaps in our understanding and opportunities for identification of optimal transcription factor (TF) combinations, which is a major bottleneck for HSC generation. In this study, we tested a novel algorithmically-predicted TF recipe (GATA2, GFIB1, FOS, REL, and STAT5A) for inducing HSC-like states. Our recipe induced $CD34^{+}$surface expression and single-cell transcriptomic signatures similar to those of native HSCs. Transcriptional heterogeneity within reprogrammed cells included differences in expression of HSC and endothelial-associated genes and in alternatively spliced transcripts as measured by single-cell long-read RNA-sequencing. Further, we proposed an approach to quantify the relative position of reprogrammed cells within the spectrum of initial and target cell states. This study lays the foundation for efficient optimization of direct reprogramming protocols.

1 Introduction

Hematopoietic stem cells (HSCs) are responsible for the continuous regeneration of blood and immune cells, making them a critical resource for treating blood disorders and immune deficiencies. While allogeneic HSC transplantation remains a viable option for many patients, it presents significant challenges such as the risk of immune incompatibility and graft-versus-host disease (GVHD) [1, 2]. Further, the need for human leukocyte antigen (HLA) matching is exacerbated by limited availability in existing donor registries, especially for underrepresented groups, who experience fewer matches and worse clinical outcomes (CIBMTR data 2016-2020) [3]. These limitations underscore the need for alternative approaches to generate HSCs.

Direct cell reprogramming via overexpression of transcription factors (TFs) enables generation of autologous cells for use in cell-based therapeutics [4]. Although strategies for cell conversion have been established [5], identifying optimal TF recipes for efficient reprogramming remains challenging. Key barriers to the generation of autologous HSCs include the labor-intensive process of identifying optimal TF recipes and the limited conversion efficiency of current protocols. These challenges are compounded by the difficulty of assessing experimental validation and the incomplete understanding of the molecular mechanisms underlying reprogramming dynamics.

Direct reprogramming of HSC-like cells from fibroblasts has been demonstrated, but conversion efficiency remains low, with small proportions of the initial cell population transitioning into HSC-like states [6–10]. Despite in-depth characterization of the hematopoietic lineage in mice [11, 12] and in humans [13–15], much is unknown about whether reprogrammed cells reach their target state, and to what extent they resemble native cell types. Quantitative comparisons are not well developed and the position of intermediate cell states along reprogramming trajectories remains incompletely understood. To optimize the search for novel reprogramming recipes, we require improved evaluation of reprogrammed cells that more closely predict their position and functional state [16].

In this study, we tested direct reprogramming of human dermal fibroblasts into the hematopoietic lineage using a novel algorithm-predicted [17] TF recipe: GATA2, GFIB1, FOS, REL, and STAT5A. We evaluated reprogrammed cells using flow cytometry and long-read single-cell RNA-sequencing, finding an increased proportion of $CD34^+$ cells using the algorithm-predicted recipe and transcriptomic overlap between reprogrammed cells and native HSCs. Finally, we identified novel heterogeneity in transcriptional programs among iHSCs at both gene and isoform levels which advance our understanding of the reprogramming cell states. We propose a framework which closes the loop between algorithmic prediction and experimental validation, thus enabling optimization of reprogramming protocols.

2 Results

2.1 A Novel Data-Driven TF-Recipe Produces CD34$^+$ Cells

Although a limited number of TF recipes have been used to reprogram human fibroblast cells into HSCs (Supplementary Figure 1), relatively little progress has been made toward improving reprogramming efficiency. We explored whether computationally predicted TFs from our previously published data-guided control (DGC) algorithm could improve direct reprogramming of fibroblasts into HSCs (Methods 4.2) [17]. Notably, GATA2, GFI1B, and FOS, which were previously demonstrated by Gomes and colleagues to reprogram human fibroblasts into an HSC-like phenotype [8], were among the top DGC-predicted TFs.

The DGC-predicted candidates STAT5A and REL were added to the 3TF recipe GATA2, GFI1B, and FOS to test whether the novel 5TF recipe (Figure 1A) could increase the yield of reprogrammed cells. Human dermal BJ fibroblasts were transduced with lentiviral vectors carrying individual TFs within TET-ON cassettes (Figure 1B). Two days after transduction, cells were treated with doxycycline (Dox) to induce TF expression. After 8 days in reprogramming conditions, flow cytometry was used to evaluate surface expression of CD90 and the HSC marker CD34 (Figure 1C). Using CD34$^+$ human bone marrow cells as a positive control for gating, lentivirally-transduced, non-Dox-treated controls had very few CD34$^+$ cells. Approximately 1% of the 3TF-treated cultures became CD34$^+$ while the 5TF recipe produced about 3-fold higher frequency of CD34$^+$ cells (Figure 1C-D). Consistent with the flow results, immunostaining showed surface expression of CD34 in approximately 3% of the cells in the reprogrammed cultures (Figure 1E).

Taken together, our results suggest that addition of predicted TFs increased the efficiency of reprogramming toward an HSC phenotype.

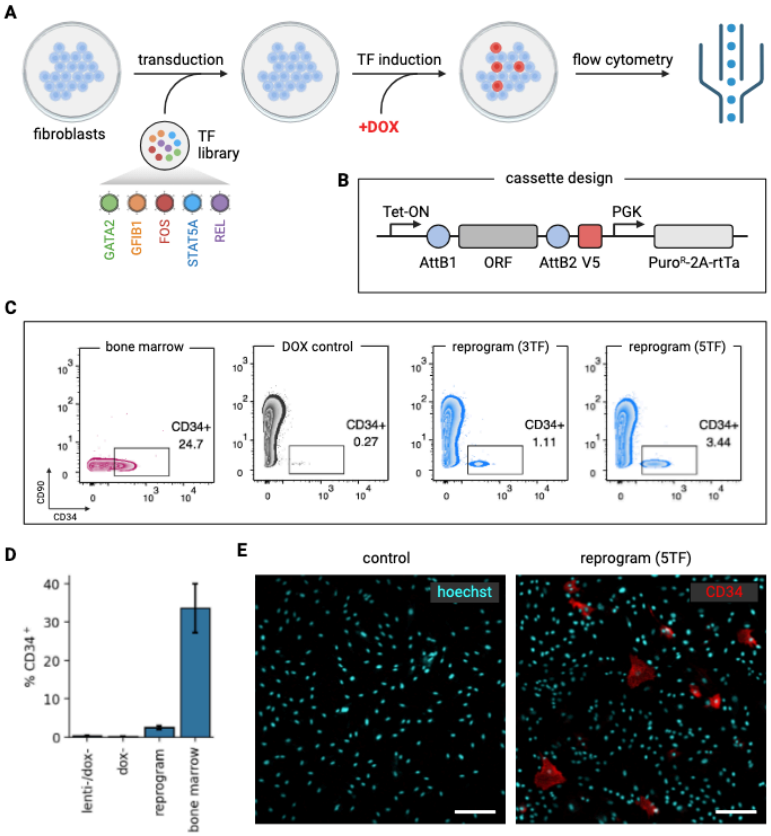

Figure 1: **A 5TF Recipe Improves Reprogramming of Human Fibroblasts Into CD34$^+$ Cells.** **(A)** Schematic representing the process for reprogramming human fibroblasts to induced hematopoietic stem cells (iHSCs), including lentiviral transduction of a transcription factor recipe, induction with doxycycline (+Dox), and quantification of CD34$^+$ cells using flow cytometry. **(B)** Cassette design. **(C)** Flow cytometry data showing surface expression of CD34 and CD90 in primary human CD34$^+$ bone marrow cells, transduced fibroblasts without Dox, and fibroblast-to-HSC reprogramming using the 3TF recipe (GATA2, GFI1B, FOS) [8] and the 5TF recipe (GATA2, GFI1B, FOS, STAT5A, and REL). Cells were gated on live singlets. **(D)** Flow cytometry data on the percentage of CD34$^+$ cells in live singlets. **(E)** Fluorescence microscopy of reprogrammed cultures with CD34 immunostaining and Hoechst staining for nuclei. Scale bar is 200μm for both panels.

2.2 Single-Cell Sequencing Reveals Transcriptomic Heterogeneity in Reprogrammed Cells

Reprogrammed cells persisted in long-term cultures, with CD34$^+$ cells maintaining similar populations for up to 50 days (Supplementary Figure 2). Reprogrammed cell cultures expanded at a similar rate as source BJ fibroblasts and were passaged several times, suggesting that the reprogrammed cells were proliferating at a similar rate as non-reprogrammed cells. Importantly, maintenance of a CD34$^+$ population suggests that the transcriptional programs active in early CD34-expressing cells were maintained in their progeny.

To gain deeper understanding of the transcriptomic state of the reprogrammed cells, we performed long-read single-cell RNA-sequencing using the Oxford Nanopore Technologies (ONT) platform (Figure 2A). We obtained 8,563 single-cell transcriptomic profiles, including a population of CD34$^-$ fibroblasts as a reference for the initial cell state (Supplementary Table 2, Methods 4.6). Clustering of the single-cell data resulted in six major groups (Figure 2B-C, Supplementary Table 3, Methods 4.7.1) [18, 19]. Cells in clusters C1-C3 and C5 showed variable expression of the reprogramming TFs and CD34, with the highest CD34 expression in C2 (Figure 2D, Supplementary Table 4). We identified cluster C4 as the CD34$^-$controls (Supplementary Figure 6) and C6 was excluded from downstream analysis due to low cell count and read-depth.

iHSCs exhibited considerable transcriptomic heterogeneity, with elevated expression of fibroblast-associated transcripts such as *FSTL1* and *KLF2* in clusters C1 and C3, elevated HSC marker expression of *MGST1* and *MPL* in C2 and elevated endothelial-associated expression of *THY1/CD90* and *PECAM1* in C5 (Supplementary Figure 3, Supplementary Table 7). Hematopoietic lineage genes such as *CD34*, *SOX4*, and *PTPRC*, were detected in 20-30% of iHSCs (Figure 2E, Supplementary Table 9). Compared to population-level BJ fibroblast data from [20], iHSCs expressed HSC-associated transcripts at higher levels and fibroblast-associated transcripts at lower levels, suggesting a phenotypic shift from the initial cell state (Figure 2F, Supplementary Figure 4A-C) [21].

Comparison of iHSC with single-cell data from BJ fibroblasts yielded similar results (Methods 4.1, 4.6, 4.7, 4.7.2). iHSCs showed elevated expression of reprogramming TFs, CD34, and other transcripts associated with the hematopoietic lineage (Supplementary Figures 5A–F, 9A–B). Joint clustering of iHSCs and controls revealed distinct expression patterns in cluster C2, including elevated expression of immune genes such as *KAT7*, *IFNA17*, *ADGRG3*, and *BLNK* (Supplementary Figure 7, Methods 4.7.2). Pairwise distances between transcriptomic profiles supported that clusters C2 and C5 are the most transcriptomically distinct subpopulations, which were distinguishable by higher expression of HSC-associated genes compared to other iHSCs and controls (Figure 2G-H, Supplementary Figure 5G–H). That clusters C2 and C5 were undergoing active reprogramming was further supported by RNA velocity analysis, which revealed up-regulation of endothelial pathways in cluster C2 and C5 (Supplementary Figure 8, Methods 8).

Overall, these results underscore the transcriptomic heterogeneity within

reprogrammed cultures, and identify clusters C2 and C5 as displaying expression patterns suggestive of phenotypic shift towards an HSC-like state.

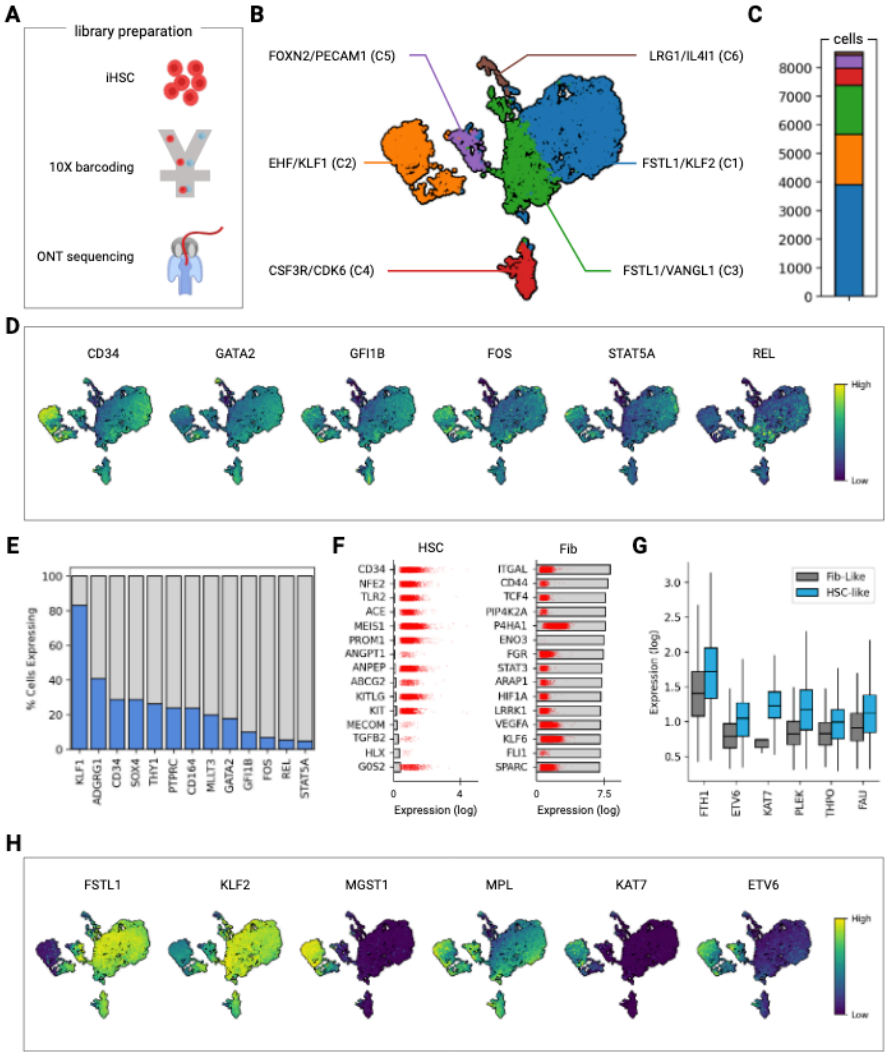

Figure 2: **Single-cell Characterization of iHSCs.** **(A)** Overview of iHSC single-cell long-read sequencing. **(B)** UMAP plot showing clustering results for single-cell profiles. **(C)** The number of cells in each iHSC cluster. **(D)** MAGIC-imputed expression patterns of CD34 and TFs used for reprogramming on the UMAP coordinates from part B. **(E)** The percentage of single-cell profiles that express key HSC transcripts. **(F)** Single-cell expression of HSC (left) and fibroblast genes (right) compared to BJ fibroblast at the population level. The grey bars show fibroblast gene expression levels and the red points show expression levels of randomly sampled iHSC cells. **(G)** Expression of HSC transcripts differentially expressed between clusters C2/C5 vs. C1/C3/C4 (Wilcoxon-rank sum test, $\alpha \leq 0.05$, log2 fold-change > 1). **(H)** MAGIC-imputed expression patterns of fibroblast and HSC transcripts on the UMAP coordinates from part B.

2.3 Clustering of Reprogrammed Cells and Native Cell Types Identifies Common Expression Patterns

The degree to which reprogrammed cells resemble native cell types remains poorly understood. To evaluate transcriptional patterns in reprogrammed cells relative to native cell types, we integrated single-cell transcriptomic data from our iHSCs with single-cell data from publicly available human hematopoietic cells [13, 15] and fibroblasts [22] (Methods 4.8, Supplementary Table 11). Clustering of single-cell data based on batch-corrected marker gene expression resulted in 21 groups (B1-B21), which we visualized using Uniform Manifold Approximation and Projection (UMAP) (Figure 3A) and annotated according to the dominant cell type in each cluster (Figure 3B, Methods 4.8.1).

To investigate the heterogeneity of native cell types and identify those most similar to iHSCs, we analyzed the distribution of cell types in each cluster (Figure 3C). As expected, clusters containing fibroblasts (B4-6, B12-13, B15, B18) were separated based on tissue origin and were distinct from clusters containing hematopoietic cell types (Figure 3C). In contrast, clusters containing hematopoietic cell types were more heterogeneous due to similar expression patterns of HSC markers. For example, several clusters contained native HSCs along with lineage negative cells (B3, B9, B19, B20) or native HSCs with multipotent progenitors (B1, B3, B9, B17) (Figure 3C). Interestingly, iHSCs clustered with both fibroblast (B12) and hematopoietic cell types (B2, B11, B14, B16, B19) separately (Figure 3C). The hematopoietic markers *GATA2*, *GFI1B* and *PTPRC* were expressed similarly in iHSCs and hematopoietic cell types, while *CD34* and *THY1* distinguished iHSCs and fibroblasts less clearly (Figure 3D, Supplementary Figures 10, Supplementary Table 12).

We then evaluated the distribution of the original iHSC clusters (C1-C6) within the new clusters (B1-B21). The distribution of new clusters among the original iHSC clusters showed B2 as the dominant iHSC grouping (Figure 3E). Consistent with previous findings, C2 and C5 cells were partitioned into B11 (Figure 3E), which was characterized by high differential expression of several HSC marker genes (Supplementary Figure 10, Supplementary Figure 11).

Integrated analysis highlighted the diverse transcriptomic landscape of native cell types and identified iHSC subpopulations with transcriptomic similarity to native HSCs.

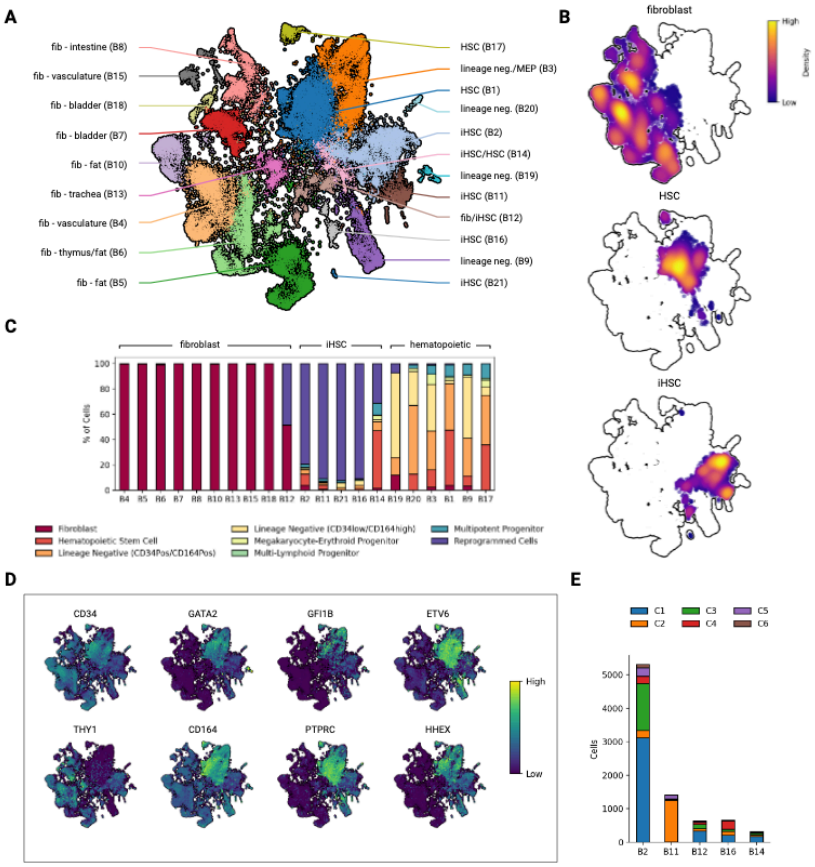

Figure 3: **Reprogrammed Cells and the Hematopoietic Landscape**. **(A)** UMAP plot showing clustering of fibroblast, hematopoietic cell types, and iHSCs **(B)** Density of fibroblast, HSC, and iHSC on UMAP coordinates from part (A) **(C)** Distribution of cell types by cluster **(D)** Log-normalized expression patterns on UMAP coordinates from part (A) **(E)** Count of iHSCs, grouped by original cluster, in the new clusters that contained at least 100 iHSCs.

2.4 Quantifying Transcriptomic Similarity to Native HSCs

We sought to address whether reprogrammed cells are transcriptomically similar to their reprorgamming target. To quantify the extent to which the expression patterns in iHSCs aligned with those in native HSCs, we used distance metrics to represent transcriptomic similarity between single cells [23]. Cosine distance was computed on the expression of marker genes to analyze differences between the core transcriptional programs of different cell types. First, we measured the distribution of distances between iHSCs and canonical HSC and fibroblast populations represented by the centroids of principal component embeddings of HSC and fibroblast marker genes (Methods 4.8.1). Overall, iHSCs were closer to the HSC centroid than to the fibroblast centroid (Figure 4A).

Next, we compared our iHSC data to 3,243 single-cell native HSC profiles ($CD34^+/CD45RA^-/CD90^+$) from Weng *et al.* [15]. Using 25 HSC marker genes (Supplementary Table 8), we computed transcriptomic similarity between each iHSC-HSC cell pair. Figure 4B shows the minimum distances between each iHSC and their closest HSC, with more similar cells (distances <1) highlighted in red. As expected, differential expression analysis revealed transcripts involved in hematopoiesis such as *ETV6* were upregulated in the closest iHSCs when compared to those with the largest distances (Supplementary Figures 13, 14, 15A-B). The closest iHSCs expressed HSC-associated transcripts that were not used in distance calculations, such as *TYROBP* and *MPL* (Supplementary Figures 15A) and had similar expression patterns of marker genes to native HSCs (Supplementary Figures 13). In contrast, iHSCs furthest from HSCs had higher expression of fibroblast-associated genes *KLF2* and *FSTL1* (Supplementary Figures 15C-D).

Among iHSC clusters, C2, C4, and C5 had the highest similarity to native HSCs (Figure 4C). Importantly, cluster C4 was the most similar to native HSCs due to transcripts implicated in cell cycle regulation, highlighting a central challenge in quantification of high-dimensional transcriptomic profiles. Consistent with their higher expression of HSC-associated genes, cluster C2 and C5 were also close to native HSCs (Figure 2). Additionally, the overall distance between all iHSCs and HSCs was lower than that between iHSCs and adult dermal fibroblasts (Supplementary Figure 16). These results demonstrate that iHSCs in clusters C2 and C5 have similar expression of HSC-associated transcripts to native HSCs.

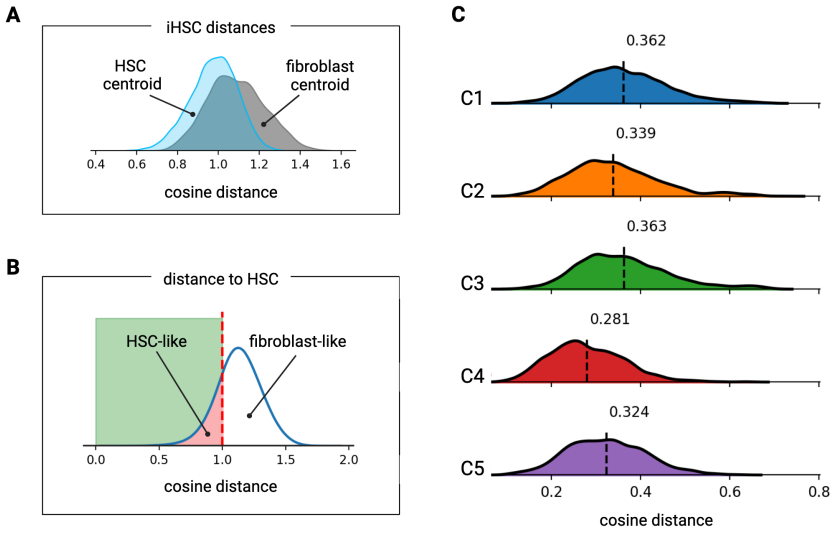

Figure 4: **Marker Gene Expression Distance between iHSC and HSC.** **(A)** Cosine distance distribution for all iHSC cells to fibroblast and HSC centroids. **(B)** Distribution of median cosine distances between iHSCs and gold-standard HSC profiles, with the closest iHSCs highlighted in red. **(C)** Minimum cosine distance distributions for iHSC clusters C1 to C5.

2.5 Positioning of iHSCs in the Hematopoietic Landscape

To address the gap in methods for quantitatively assessing reprogrammed cell states, we developed an approach to characterize the relative position of our reprogrammed cells within the spectrum of initial and target cell states. Our method is composed of three steps: (1) identification of a suitable representation of cell state, (2) calculation of pairwise distances between the groups of interest, and (3) analysis of inter-group distances using manifold learning techniques. In the absence of datasets spanning full reprogramming trajectories, our approach provides a quantitative framework for evaluating the relative position of reprogrammed cells in the context of native cell types.

We applied our method to contextualize our iHSC signatures with respect to single-cell profiles of hematopoietic stem cells (target) [15], adult fibroblasts [22], and BJ fibroblasts (initial). To harmonize single-cell data from different sequencing platforms and develop a representation of initial and target cell states, we trained a semi-supervised variational inference model (scANVI) on initial and target cell states [24–27]. Using the trained scANVI model as the reference, we mapped our iHSC signatures onto the latent space using the scArches method [25, 27, 28] (Methods 4.9).

Based on expression of HSC and fibroblast marker genes, we labeled iHSC clusters C2 and C5 as $\text{iHSC}_{\text{high}}$ and clusters C1 and C3 as iHSC_{low} (Figure 5A). To calculate pairwise distances between cell groups, we averaged batch-corrected gene expression values within each group and then computed the cosine distance between the resulting aggregated expression signatures. The principal multidimensional scaling (MDS) component was used to represent distance on a 1-dimensional spectrum from initial to target state (Figure 5B).

The relative positioning of iHSC signatures indicated that iHSC cells are closer to the initial state than the target. Among iHSCs, $\text{iHSC}_{\text{high}}$ cells were positioned closer to the target state than iHSC_{low} cells (Figures 5B, C). In agreement with previous results, iHSC cluster C2 was nearest to the target state, while cluster C4 resembled the initial cell type (Figure 5D).

We then examined gene expression patterns in iHSCs as function of their distance from the target state. To investigate the genetic programs distinguishing $\text{iHSC}_{\text{high}}$ and iHSC_{low} from the initial cell state, we calculated the Pearson correlation between each gene's batch-corrected expression and the principal MDS component. Genes that were upregulated in HSCs relative to fibroblasts (Figure 5E) showed high positive correlations, such as *RUNX1* ($r = 0.997$) and *HLF* ($r = 0.991$). Similarly, down-regulated genes (Figure 5F) showed negative correlations; *FGF2* ($r = -0.962$) and *SPARC* ($r = -0.967$). We then extracted genes expressed more highly in $\text{iHSC}_{\text{high}}$ than in the initial cell state and visualized their normalized expression on the scANVI-based UMAP coordinates (Figure 5G). Among genes correlated with fibroblast-to-HSC position and upregulated in $\text{iHSC}_{\text{high}}$ were *CMPK2* ($r = -0.369$) and *PTPRK* ($r = -0.377$) and HSC marker gene *MS4A6A* ($r = -0.255$).

These results quantify the relative positioning of iHSCs within the continuum of initial and target cell states and suggest that distinct subsets of *CD34*$^+$iHSCs

advance further along the reprogramming trajectory with respect to the target state.

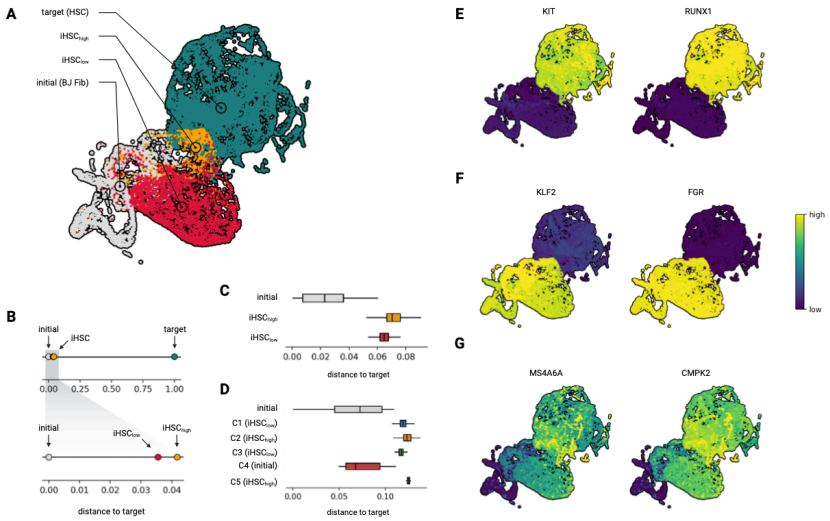

Figure 5: **Positioning of iHSCs in the Hematopoietic Landscape. (A)** UMAP embedding of iHSC$_{\text{high}}$, iHSC$_{\text{low}}$, HSC and BJ fibroblast based on scANVI embeddings. **(B)** Cosine distance between the batch-corrected normalized expression of aggregated single-cell profiles projected onto a line using MDS. Top shows initial (BJ fibroblast) relative to the target (native HSC). Bottom shows iHSC$_{\text{high}}$ and iHSC$_{\text{low}}$ relative to the initial condition. **(C)** Distribution of MDS component of 1,000 randomly sampled single-cells by group. **(D)** Distribution of MDS component of 1,000 randomly sampled single-cells by initial iHSC cluster. **(E)** Normalized batch-corrected expression of HSC-marker genes upregulated in the target cell type. **(F)** Normalized batch-corrected expression of genes upregulated in iHSC$_{\text{high}}$ cells. **(G)** Normalized batch-corrected expression of genes upregulated in the initial cell type.

2.6 Alternative Splicing in iHSCs

A major advantage of long-read sequencing technology is the ability to quantify full length transcripts and resolve specific isoforms that arise as a result of alternative splicing (AS) [29–31]. AS is important during fate-specification in hematopoiesis and during cellular responses in mature hematopoietic lineage cells [32–35]. Major changes in splicing are also observed during cell reprogramming to pluripotent cells [36–38]. We reasoned that if AS is cell type specific and has a role in shaping cell identity, changes in isoform balance should occur during fibroblast-to-HSC reprogramming. We therefore further queried the 138 million pooled long-reads from the iHSC and BJ fibroblast single-cell datasets that mapped to known isoforms [31] (Methods 4.10).

Evaluation of global patterns in the isoform data showed that iHSCs expressed a higher proportion of noncoding transcripts compared to fibroblasts (Figure 6A). We then investigated AS in the core transcriptomic program of the initial and target cell types. For fibroblast marker genes, the number of isoforms per gene was similar in iHSCs and fibroblasts. However, for HSC markers, there were significantly fewer isoforms per gene in iHSCs compared to fibroblasts (Figure 6B), suggesting an AS-related shift in isoform usage for these genes.

Next, we compared isoform usage as a proportion of the total reads per gene. We observed differences in total expression levels and isoform usage for several transcripts. Genes *CD34* and *GATA2* had higher expression in iHSCs than in controls and *PRRX1* and *COL15A1* had higher expression in controls compared to iHSCs (Figure 6C). Isoform usage was also different for these genes (Figure 6C).

We then compared isoform usage for genes whose transcripts mapped to multiple UniProt protein IDs in both libraries (Figure 6D). We identified 166 genes with a dominant isoform in iHSCs that differed from the dominant isoform detected in fibroblasts. In some cases, an isoform contributed more than 50% of the read counts for the gene in one library but not the other (Figure 6E). Among genes with differing dominant isoforms were HSC genes such as *ELF1*, *FOS*, *FTH1*, *EEF1A1* and several ribosomal genes such as *RPS19*, *RPL14*, *RPL38*, and *RPL35A* which are upregulated in HSCs [22]. Similarly, the list included fibroblast-associated genes such as *LOX*, *EFEMP1*, *SERPINH1*, and *C1R*. These shifts in isoform dominance may suggest a role for AS during direct reprogramming to HSCs through the modulation of both fibroblast and HSC identity markers.

Differences in isoform usage at the library-level led us to consider the potential to analyze isoform expression at the single-cell level. We extracted unambiguously identified known isoforms and assigned them to error-corrected barcodes. We visualized single-cell isoform expression patterns by mapping detected isoforms onto a UMAP embedding based on single-cell gene expression (Figure 7A). Tests for differential expression were used to determine isoforms that were up-regulated in iHSCs compared to fibroblasts. Isoforms of HSC genes *ALDH1A1* and *PTPRS* showed differences in the MAGIC-imputed expression patterns by cluster (Figure 7 B-C). Further, specific isoforms of HSC-related

genes were upregulated in iHSCs when compared to fibroblasts (Figure 7D).

These results highlight alternative splicing diversity in reprogrammed cells and may suggest a role for AS in direct reprogramming.

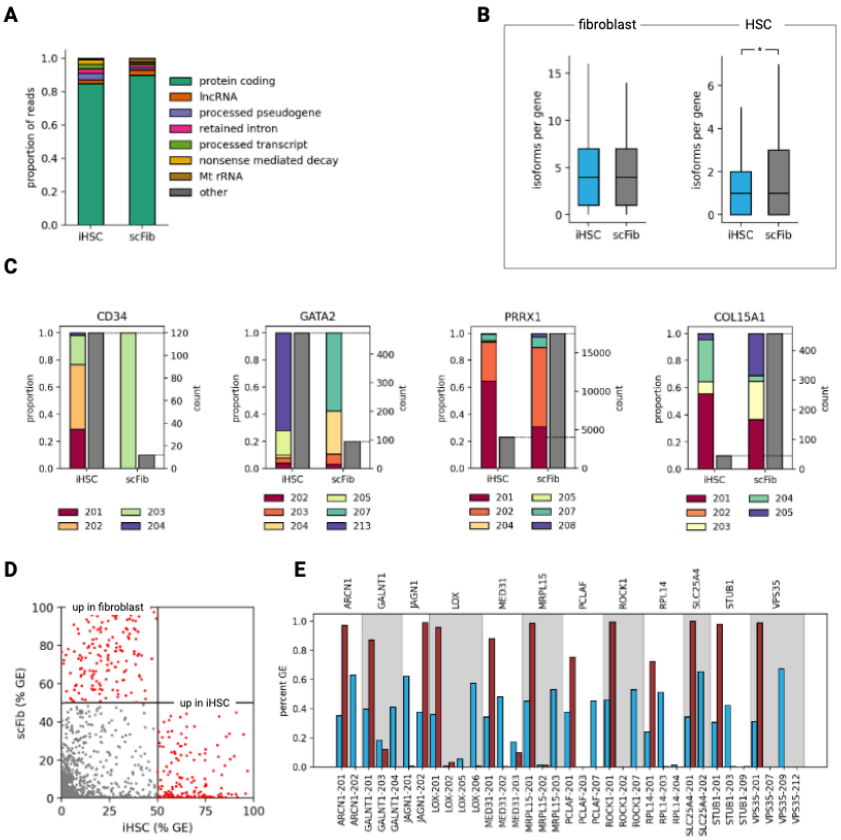

Figure 6: **Alternative Splicing in iHSCs. (A)** The proportion of detected known isoforms for iHSC and BJ fibroblast datasets by transcript biotype. **(B)** The number of detected isoforms per gene for fibroblast (left) and HSC marker genes (right) by dataset. **(C)** Isoform usage ratio (left bars) and total reads (right, grey bars) of select genes by dataset. **(D)** Scatterplot showing isoforms which make up large proportions of the total gene expression (GE) in BJ fibroblast vs. iHSC data sets. **(E)** Examples of isoforms with differing percentage of total gene expression by dataset.

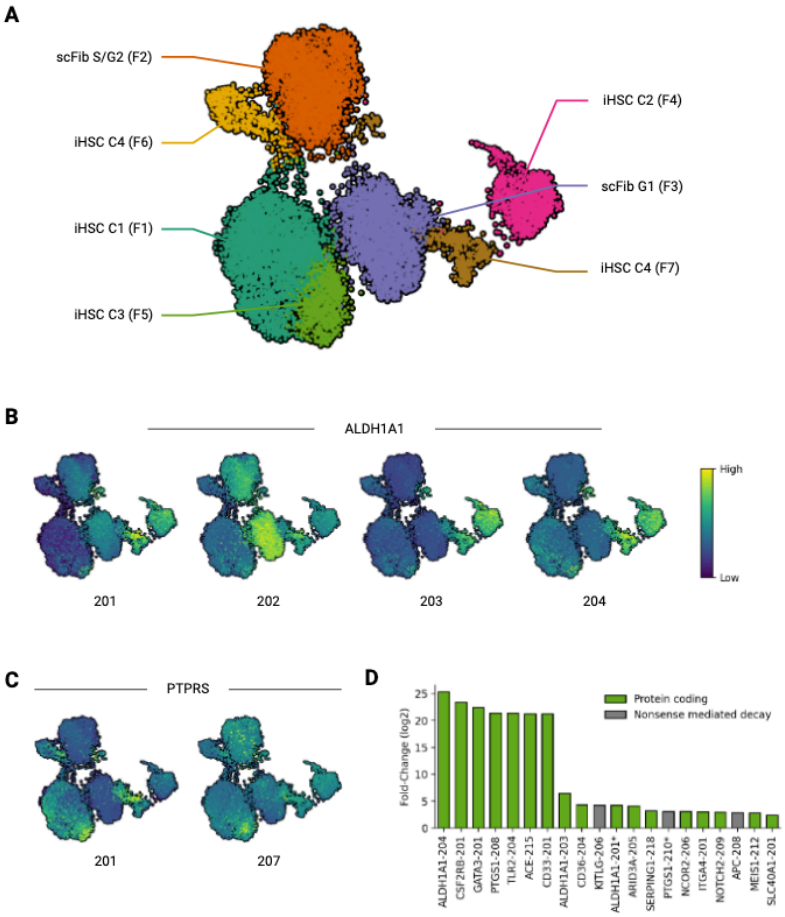

Figure 7: **Isoform Expression in iHSCs. (A)** UMAP of BJ fibroblast iHSC clusters generated on gene expression values. **(B)** MAGIC-imputed single-cell expression patterns of ALDH1A1 isoforms **(C)** MAGIC-imputed single-cell expression patterns of PTPRS isoforms **(D)** Fold-change (log2) of expression for HSC marker genes for HSC-like cells (F4/F7) vs. fibroblast-like cells. Asterisk denotes that the transcript is statistically differentially expression in clusters (F4/F7) vs. fibroblast-like cells ($\alpha < 0.05$)

3 Discussion

Here we demonstrate that the addition of algorithm-predicted TFs improved the conversion efficiency of reprogramming human fibroblasts into the hematopoietic lineage and produced a subset of cells that bear transcriptomic similarity to native HSCs.

Prior work proposed an endothelial-like precursor during HSC reprogramming in mice [39, 40] and humans [8, 9, 41]. Consistent with this hypothesis, our iHSC clusters that were the most similar to native HSCs (C2/C5) also expressed endothelial markers such as *ADORA2A* and *CASZ1* (Supplementary Figures 8, 18, 19, 17).

While our analyses support that iHSCs shifted toward the target state, significant transcriptomic differences remained between iHSCs and native HSCs (Figure 5B). These results suggest a partially reprogrammed state, whose position along a reprogramming trajectory will require time-series data to fully resolve. Despite these differences, our reprogrammed cells show clear signs of a down-regulated fibroblast program, increased hematopoietic gene expression (Figure 2F, H), and similarity to previously published iHSC signatures (Supplementary Figure 20A-D) [8].

Precise contextualization of reprogrammed cell states allows for refined feedback during iterative development of reprogramming protocols. Changes to TF recipe constitution, delivery, and/or timing may be closely monitored with respect to their impact on outcome metrics such as distance from the target state. For example, identification of the transcriptional program that distinguishes iHSCs by their distance from native HSCs has the potential to reveal both barriers and drivers of cell type conversion. Overall, our results underscore the utility of characterizing reprogramming cell states and provide a benchmark for assessing the transcriptomic proximity of partially reprogrammed cells to the target phenotype.

To our knowledge, this study is the first to characterize direct HSC reprogramming using single-cell long-read sequencing, which enabled high-accuracy detection of isoforms [31]. Variance in isoform expression by iHSC cluster suggests a potential role for AS in cell type conversion. However, further studies will be required to determine the overlap between iHSCs and native HSCs at the isoform level and the impact of transient AS mechanisms on reprogramming outcomes. Relatively few studies have investigated the role of AS in hematopoiesis, and much of our understanding to date comes from mouse models [32, 33]. Changes in AS are also observed in other reprogramming settings. Studies on reprogramming to pluripotency show sequential phases of AS [36–38, 42]. Other examples in mice show that isoform switching in TFs impacts reprogramming [43, 44]. These studies support the notion that isoform usage can impact reprogramming and may also provide higher resolution for discriminating cell states.

Overall, our findings contribute to the broader effort to advance cellular reprogramming techniques used to generate autologous HSCs. We further provide a benchmark for coupling algorithmic prediction with experimental validation

to accelerate development of direct reprogramming protocols.

4 Methods

4.1 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 DMEM containing sodium pyruvate + 10% fetal bovine serum + 1% penicillin/streptomycin + 1% L-glutamine. Lentiviral transductions were carried out in complete medium without penicillin/streptomycin 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% penicillin/streptomycin 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 Express Enzyme (Gibco #12604-021).

4.2 Algorithmically Guided Recipe Generation

Transcription factors (TFs) for fibroblast-to-HSC reprogramming were selected as described in [17]. TFs were identified by minimizing the distance between the initial and target cell states using the Data-Guided Control (DGC) algorithm [17]. 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 $\mathbf{u}$.

4.3 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. 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 experiments shown in Figure 1, 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

4.4 Immunofluorescence

For immunofluorescent staining, cells were fixed with 4% paraformaldehyde (PFA) for 10 minutes, washed with cold phosphate-buffered saline (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.

4.5 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. 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 FACS. FlowJo v10.10.0 (Becton Dickinson) was used for flow cytometry analyses.

4.6 Single-Cell Transcriptomics

Full-length barcoded cDNA amplicons were obtained from single cells using the 10X Genomics Next GEM Single Cell 3' Kit V3.1. Briefly, 20,000 sorted $CD34^+$ 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 1). Briefly, cDNA amplicons were biotin tagged by PCR with the primers shown in Table 1. 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 1 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 were obtained as for the reprogrammed cells except using the Next GEM Single Cell 3' Kit V4. Additionally, fibroblasts were sorted by cell cycle phase after staining with 16 μM Hoechst 33342 for 50 minutes. Samples of cells in G1, S, and G2/M were labeled individually with BioLegend TotalSeq-B human hashing antibodies (Cat# 394631, #394603, and #394605, respectively), mixed, and single cell barcoded (viability >97%). Transcript and hashing libraries from this sample were processed barcoded libraries returned (transcript and hashing).

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

4.7 scRNA-seq Data Analysis

An analysis pipeline for quality control and analysis of long-read single-cell data was developed using `Snakemake` [45]. Raw sequencing files were demultiplexed with `blaze2` [46], with the parameter `--expect-cells` set to 10,000. For iHSCs, barcodes from `3M-february-2018.txt` were used, while control fibroblasts were analyzed using `3M-3pgex-may-2023.txt`. Demultiplexed reads were aligned to the GRCh38 genome [47] using `minimap2` [48] with parameters `-ax splice -uf --secondary=no --MD`.

Count matrices were generated with HTSeq `htseq-count` [49], and these matrices were stored as `anndata` [50, 51] objects. Analysis was conducted in `Python` using the `scanpy` library [19]. Single-cell expression data were filtered to retain cells expressing more than 200 genes and genes expressed in fewer than 3 cells were filtered out. Read counts were normalized to 10^4 and logarithmized. Where indicated, expression values were imputed using MAGIC with default parameters [19, 52].

4.7.1 Clustering of iHSCs

We analyzed the single-cell data using a curated set of genes including marker genes from PanglaoDB [21] and highly variable genes identified from the iHSC data. Principal component analysis (PCA) was performed on normalized counts and the top 10 principal components were selected based on the variance ratio. A nearest neighbor graph was constructed with $k = 100$ and clustering was performed using the Leiden algorithm with a resolution parameter of 0.2 [18, 19].

4.7.2 Integration of iHSC and Fibroblast scRNA-seq Data

For iHSC and BJ fibroblast scRNA-seq data, genes expressed in fewer than 20 cells were filtered. Sequencing libraries were concatenated, and further filtering

removed cells with fewer than 500 genes and genes detected in fewer than 10 cells. Gene counts were normalized to a target of 10^4 and log-transformed.

Batch effects between datasets were corrected using ComBat [53], with total counts and the number of detected genes as covariates. Dropout events were corrected using MAGIC [19, 52]. Highly variable genes were selected as features from ComBat-corrected read counts and embedded using PCA with 50 components. A neighborhood graph of 25 nearest neighbors was constructed, and clustering was performed using the Leiden algorithm with a resolution of 0.5.

4.7.3 RNA Velocity Estimation

BAM files were prepared for RNA velocity analysis using `velocyto` [54]. RNA velocity analysis was performed using `scVelo` [55, 56]. Cells with fewer than 500 genes detected and genes expressed in fewer than 5 cells were filtered. The top 6,000 highly variable genes were selected and PCA was performed using the first 20 components. A neighborhood graph was constructed using 15 nearest neighbors. Moment computation was performed with 5 nearest neighbors. RNA velocity was estimated in 'dynamical' mode, and velocity vectors were computed using the velocity graph with 15 neighbors.

4.8 Analysis of Publicly Available Data

Single-cell data from hematopoietic stem cells (HSCs) and related cell types were sourced from [13] and [15]. Fibroblast data were obtained from [22]. Expression data were structured as count matrices using the `scanpy` package [19] and the `anndata` framework [50, 51]. Cell-type annotations from the data sources were retained. Counts were concatenated into a single matrix with iHSC cell profiles. Annotations for protein-coding genes were standardized using the GTF file for GRCh38 [47].

Cells with fewer than 100 genes detected and genes expressed in fewer than 3 cells were filtered. Expression values were normalized to a target of 10^4 and logarithmically transformed. Batch correction was performed using ComBat [53].

4.8.1 Clustering of Hematopoietic Cell Types

Feature selection was based on marker genes for hematopoietic, immune, and fibroblast cell types from PanglaoDB [21]. PCA was applied to the batch-corrected feature counts, using the top 50 principal components. A nearest neighbor graph with $k = 100$ neighbors was constructed and clustering was performed using the Leiden algorithm with a resolution of 0.65.

4.9 Reference Mapping

Single cell RNA sequencing data from our study, as well as data from [22] and [15], were preprocessed to improve the detection of biological signals and

mitigate technical variation. Data were normalized to 10^4 total counts per cell followed by log transformation. Cells expressing fewer than 1,000 genes were excluded, as were genes expressed in fewer than 500 cells. The top 13,000 highly variable genes were selected as features based on log-normalized counts across batch identifiers.

The initial `scVI` model utilized a three-layer variational autoencoder architecture, configured with 128 nodes per hidden layer, a dropout rate of 20%, and layer normalization in both the encoder and the decoder. The model was trained for a maximum of 500 epochs, with early stopping applied based on the evidence lower bound (ELBO) on validation data, and a patience of 5 epochs. An initial learning rate of 0.001 was set, with a KL divergence warm-up over the first 10 epochs. Learning rate was reduced on plateau with a patience of 8 epochs.

Following this training phase, we initialized a semi-supervised `scANVI` model using the pre-trained `scVI` architecture, training it for an additional 500 epochs under with the same training parameters. This model facilitated comparative analysis of iHSC cell distributions compared to those of the reference data sets. For query mapping, both `scVI` and `scANVI` models were adapted to the query data, utilizing the pretrained representations for embedding and count normalization across data sets. This enabled accurate integration with the reference data.

4.10 Isoform Identification

We performed isoform identification and quantification of known isoforms using `IsoQuant`, a tool designed for genome-based analysis of long RNA reads [31]. Briefly, the isoform quantification workflow contains the following steps: (1) construction of an annotation database from an input GTF file, (2) read-assignment to annotated isoforms based on their intron and exon structure.

The following parameters were used with `IsoQuant`: `--reference=GRCh38 --count_exons --data_type` `nanopore' `--complete_genedb --bam_tags` `CB,UB,RD'`. BAM files were generated using `minimap2` [48] in splice mode, as previously described.

Isoforms were assigned to error-corrected barcodes using parsed `IsoQuant` outputs filtered for read IDs with assignment types of either 'unique' or 'unique minor difference'. Each matching read ID was associated with a unique cell barcode-UMI combination in alignment files (BAM) and assigned a count of 1. The data were then structured into a cell-by-isoform counts matrix, with gene-level annotations merged according to the call barcode.

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, 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. 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/ONT-single-cell and
https://github.com/CooperStansbury/Nanopore10x_BJsc.

8 Data Availability

Raw sequencing data generated in this study will be deposited in the Gene Expression Omnibus upon acceptance and is available upon request.

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] H. Joachim Deeg. "Reward from half a match". In: *Blood* 141.25 (2023).

[2] Amy E. DeZern et al. "Alternative donor BMT with posttransplant cyclophosphamide as initial therapy for acquired severe aplastic anemia". In: *Blood* 141.25 (2023).

[3] Karen Ballen et al. "Impact of Race and Ethnicity on Outcomes After Umbilical Cord Blood Transplantation". In: *Transplantation and Cellular Therapy* 30.10 (2024).

[4] Haofei Wang et al. "Direct cell reprogramming: approaches, mechanisms and progress". In: *Nature Reviews Molecular Cell Biology* 22.6 (2021).

[5] Kenichi Horisawa and Atsushi Suzuki. "Direct cell-fate conversion of somatic cells: Toward regenerative medicine and industries". In: *Proceedings of the Japan Academy, Series B* 96.4 (2020).

[6] Eva Szabo et al. "Direct conversion of human fibroblasts to multilineage blood progenitors". In: *Nature* 468.7323 (2010).

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

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

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

[10] Michael G. Daniel et al. "Induction of human hemogenesis in adult fibroblasts by defined factors and hematopoietic coculture". In: *FEBS Letters* 593.23 (2019).

[11] Caleb Weinreb et al. "Lineage tracing on transcriptional landscapes links state to fate during differentiation". In: *Science* 367.6479 (2020).

[12] Sam Watcham, Iwo Kucinski, and Berthold Gottgens. "New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing". In: *Blood* 133.13 (2019).

[13] Danilo Pellin et al. "A comprehensive single cell transcriptional landscape of human hematopoietic progenitors". In: *Nature Communications* 10.1 (2019).

[14] 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).

[15] Chen Weng et al. "Deciphering cell states and genealogies of human haematopoiesis". In: *Nature* 627.8003 (2024).

[16] 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).

[17] Scott Ronquist et al. "Algorithm for cellular reprogramming". In: *Proceedings of the National Academy of Sciences* 114.45 (2017).

[18] V. A. Traag, L. Waltman, and N. J. van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific Reports* 9.1 (2019).

[19] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome Biology* 19.1 (2018).

[20] Haiming Chen et al. "Functional organization of the human 4D Nucleome". In: *Proceedings of the National Academy of Sciences* 112.26 (2015).

[21] 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).

[22] THE TABULA SAPIENS CONSORTIUM. "The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans". In: *Science* 376.6594 (2022).

[23] Yuge Ji et al. *Optimal distance metrics for single-cell RNA-seq populations.* 2023.

[24] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature Methods* 15.12 (2018).

[25] Mohammad Lotfollahi et al. "Mapping single-cell data to reference atlases by transfer learning". In: *Nature Biotechnology* 40.1 (2022).

[26] Malte D Luecken and Fabian J Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular Systems Biology* 15.6 (2019).

[27] Lukas Heumos et al. "Best practices for single-cell analysis across modalities". In: *Nature Reviews Genetics* 24.8 (2023).

[28] Malte D. Luecken et al. "Benchmarking atlas-level data integration in single-cell genomics". In: *Nature Methods* 19.1 (2022).

[29] Vivien Marx. "Method of the year: long-read sequencing". In: *Nature Methods* 20.1 (2023).

[30] 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).

[31] Andrey D. Prjibelski et al. "Accurate isoform discovery with IsoQuant using long reads". In: *Nature Biotechnology* 41.7 (2023).

[32] Fang Wang et al. "Single-cell architecture and functional requirement of alternative splicing during hematopoietic stem cell formation". In: *Science Advances* 8.1 (2022).

[33] 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).

[34] Lu Chen et al. "Transcriptional diversity during lineage commitment of human blood progenitors". In: *Science* 345.6204 (2014).

[35] Yapu Li et al. "A splicing factor switch controls hematopoietic lineage specification of pluripotent stem cells". In: *EMBO reports* 22.1 (2021).

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

[37] Alexander Kanitz et al. "Conserved regulation of RNA processing in somatic cell reprogramming". In: *BMC Genomics* 20.1 (2019).

[38] 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).

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

[40] Raphael Lis et al. "Conversion of adult endothelium to immunocompetent haematopoietic stem cells". In: *Nature* 545.7655 (2017).

[41] 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).

[42] 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).

[43] Hong Han et al. "MBNL proteins repress ES-cell-specific alternative splicing and reprogramming". In: *Nature* 498.7453 (2013).

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

[45] Felix Mölder et al. "Sustainable data analysis with Snakemake". In: *F1000Research* 10 (2021).

[46] Yupei You et al. "Identification of cell barcodes from long-read single-cell RNA-seq with BLAZE". In: *Genome Biology* 24.1 (2023).

[47] Fergal J Martin et al. "Ensembl 2023". In: *Nucleic Acids Research* 51.D1 (2023).

[48] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18 (2018).

[49] Givanna H Putri et al. "Analysing high-throughput sequencing data in Python with HTSeq 2.0". In: *Bioinformatics* 38.10 (2022).

[50] Isaac Virshup et al. *anndata: Annotated data*. 2021.

[51] Isaac Virshup et al. "The scverse project provides a computational ecosystem for single-cell omics data analysis". In: *Nature Biotechnology* 41.5 (2023).

[52] David van Dijk et al. "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion". In: *Cell* 174.3 (2018).

[53] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007).

[54] Gioele La Manno et al. "RNA velocity of single cells". In: *Nature* 560.7719 (2018).

[55] Volker Bergen et al. "Generalizing RNA velocity to transient cell states through dynamical modeling". In: *Nature Biotechnology* 38.12 (2020).

[56] Volker Bergen et al. "RNA velocity—current challenges and future perspectives". In: *Molecular Systems Biology* 17.8 (2021).

[57] Jiaqi Li et al. "Comparison of Scanpy-based algorithms to remove the batch effect from single-cell RNA-seq data". In: *Cell Regeneration* 9.1 (2020).