Nucleic Acids Research, 2024, **52**, D1042–D1052
https://doi.org/10.1093/nar/gkad1063
Advance access publication date: 11 November 2023
Database issue

StemDriver: a knowledgebase of gene functions for hematopoietic stem cell fate determination

Yangyang Luo[1,†], Jingjing Guo[1,†], Jianguo Wen[2], Weiling Zhao[2], Kexin Huang[1], Yang Liu[1], Grant Wang[2], Ruihan Luo[1], Ting Niu[1], Yuzhou Feng[1], Haixia Xu[1], Pora Kim[2,*] and Xiaobo Zhou[2,3,4,*]

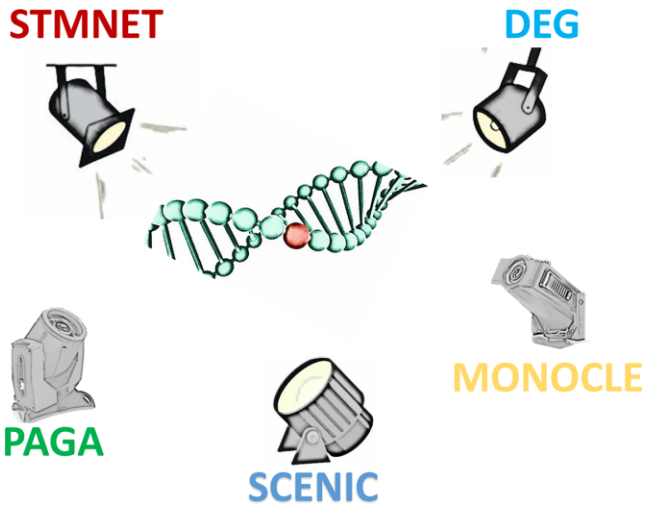
[1]Department of Hematology and West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, P.R. China
[2]Center for Computational Systems Medicine, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[3]McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[4]School of Dentistry, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[*]To whom correspondence should be addressed. Tel: +1 713 500 3923 and 3636; Email: Xiaobo.zhou@uth.tmc.edu
Correspondence may also be addressed to Pora Kim. Email: pora.kim@uth.tmc.edu
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

StemDriver is a comprehensive knowledgebase dedicated to the functional annotation of genes participating in the determination of hematopoietic stem cell fate, available at http://biomedbdc.wchscu.cn/StemDriver/. By utilizing single-cell RNA sequencing data, StemDriver has successfully assembled a comprehensive lineage map of hematopoiesis, capturing the entire continuum from the initial formation of hematopoietic stem cells to the fully developed mature cells. Extensive exploration and characterization were conducted on gene expression features corresponding to each lineage commitment. At the current version, StemDriver integrates data from 42 studies, encompassing a diverse range of 14 tissue types spanning from the embryonic phase to adulthood. In order to ensure uniformity and reliability, all data undergo a standardized pipeline, which includes quality data pre-processing, cell type annotation, differential gene expression analysis, identification of gene categories correlated with differentiation, analysis of highly variable genes along pseudo-time, and exploration of gene expression regulatory networks. In total, StemDriver assessed the function of 23 839 genes for human samples and 29 533 genes for mouse samples. Simultaneously, StemDriver also provided users with reference datasets and models for cell annotation. We believe that StemDriver will offer valuable assistance to research focused on cellular development and hematopoiesis.

Graphical abstract



Received: August 18, 2023. **Revised:** October 24, 2023. **Editorial Decision:** October 24, 2023. **Accepted:** November 1, 2023

Introduction

Hematopoiesis plays an important role in the intricate workings of the human body. Vital functions such as oxygen transport, immune system, blood clotting, and blood regeneration all rely on the proper orchestration of hematopoiesis (1). The normal self-renewal and differentiation processes of hematopoietic stem cells (HSCs) are closely connected with human well-being and health. Although HSCs are the first tissue-specific adult stem cells to be successfully isolated and used for clinical treatment (2), the underlying mechanism that governs the fate determination of hematopoietic stem cells (HSCs) still remains unknown. By employing single-cell RNA sequencing data, several recent studies have attempted to chart the dynamic molecular changes that underlie the differentiation of HSCs (3–5). However, individual studies concentrate on one or a few specific cell types derived from HSC are limited in their ability to support broad biological research. Current hematopoiesis-related databases such as BloodSpot (6), Haemopedia RNA-seq (7) and CODEX (8) provided gene-expression profiles without systematic functional annotations to elucidate the molecular functions behind cell fate determination. The information provided by these databases have become insufficient to address the current research demands. There is an urgent need for a database that analyzes genes from various perspectives, providing users with a platform for comprehensive and multifaceted gene exploration.

To address this gap, we have developed StemDriver, a comprehensive knowledge database focused on gene functions related to the determination of hematopoietic stem cell fate. Recent studies have shown that hematopoietic stem cells (HSCs) originate in the yolk sac and later appear in the aorta/gonad/mesonephros (AGM) region. Afterward, they migrate to the fetal liver and fetal bone marrow, where they undergo a phase of expansion (9–11). In order to encompass the entirety of hematopoiesis, we curated datasets that span various developmental stages, ranging from embryonic to adult phases. In the end, StemDriver includes 42 datasets from both human and mouse samples, covering 14 distinct tissue types. By utilizing the collected data, we have generated a lineage map that encompasses 22 major cell types, starting from the early formation of HSCs during embryogenesis and extending to the emergence of terminal unipotent cells in adults.

StemDriver offers comprehensive gene annotation, delving into gene roles in hematopoietic stem cell differentiation from multiple perspectives. Initially, we explored the correlation between gene expression and differentiation direction and extent. Subsequently, we analyzed gene expression differences across cell types and within stem cell subsets with varying differentiation directions. Lastly, we monitored dynamic gene expression changes and identified highly variable genes along developmental trajectories. The synthesis of these analyses provides insights into the gene's effects on differentiation direction, degree, and specificity, its potential to initiate specific differentiation pathways, and its significant influence on cell transitions during distinct stages. Moreover, we probed the enrichment of gene expression regulatory networks within each lineage commitment. Finally, we carried out a detailed characterization of the roles played by 23 839 genes in humans and 29 533 genes in mice during the differentiation processes of hematopoietic stem cells and their progenitors. The comprehensive annotation results provided by StemDriver enable us to gain profound insights into the molecular characteristics associated with the differentiation of pluripotent stem cells into various cell lineages. All these reuslts will contribute to the identification of novel targets suitable for cellular engineering or disease treatment.

Materials and methods

Data collection

We conducted a search of previous studies centered on HSCs using scRNA-seq data through the PubMed website, employing keywords such as 'hematopoietic stem cell', 'single cell', and 'RNA-seq'. Subsequently, a total of 245 relevant papers were identified. From these datasets, we applied the following criteria for dataset selection: (i) The datasets had to be publicly accessible. (ii) Only datasets generated using single-cell RNA sequencing techniques were considered. (iii) The study focused on organisms within the *Homo sapiens* and *Mus musculus* species. (iv) Included cell types were limited to early formed HSCs, HSCs, and their resultant cell types. (v) The datasets had to meet the quality control standards outlined in the data preprocessing section (Figure 1). Following the application of these filters, a total of 42 scRNA-seq datasets remained. These datasets were originated from 14 different types of tissues, including adult bone marrow, adult peripheral blood, adult spleen, Aorta-Gonad-Mesonephros (AGM), cord blood (CB), fetal bone marrow (FBM), fetal liver (FL), fetal kidney (FK), fetal genitourinary system (FG), fetal thymus (FT), fetal skin (FS), fetal artery (FA), yolk sac (YS), and placenta (PL). The scRNA-seq datasets were generated using a variety of sequencing platforms, including SMART-seq2, START-seq, CITE-seq, inDrop, 10× Genomics, Fluidigm C1 and Microarray.

Data pre-processing

The raw data collected in StemDriver underwent preprocessing using a standardized pipeline. Here is an overview of the steps involved: (i) Gene symbol mapping: The original datasets, which used Entrez ID for gene information, were mapped to gene symbols using the org.Hs.eg.db (12) in R for human data and org.Mm.eg.db (13) for mouse data. To standardize gene symbols, we retained human genes recorded as approved genes in the HGNC database (14) and mouse genes, excluding withdrawn marker symbols, recorded by MGI (15). (ii) Gene expression matrix construction: A gene expression matrix with raw data was constructed for each data set. Datasets that only provided normalized data were excluded, as SCENIC (16) analysis requires the raw count data. (iii) Quality control: We first excluded genes expressed in less than 5 single cells, cells with >200 genes, and <20% mitochondrial genes retained for downstream analyses.

Dataset Integration

In this study, data integration was selectively applied to datasets exclusively composed of stem cells or progenitors. These datasets primarily consisted of cells with minimal differentiation variation, which posed limitations for cell differentiation analysis. To overcome this, these cells were integrated with more mature cell types to facilitate trajectory analysis. Briefly, pre- or HSC cells were integrated with progenitors, including common lymphoid progenitors (CLP), granulocyte-monocyte progenitors (GMP) and megakaryocyte-erythroid

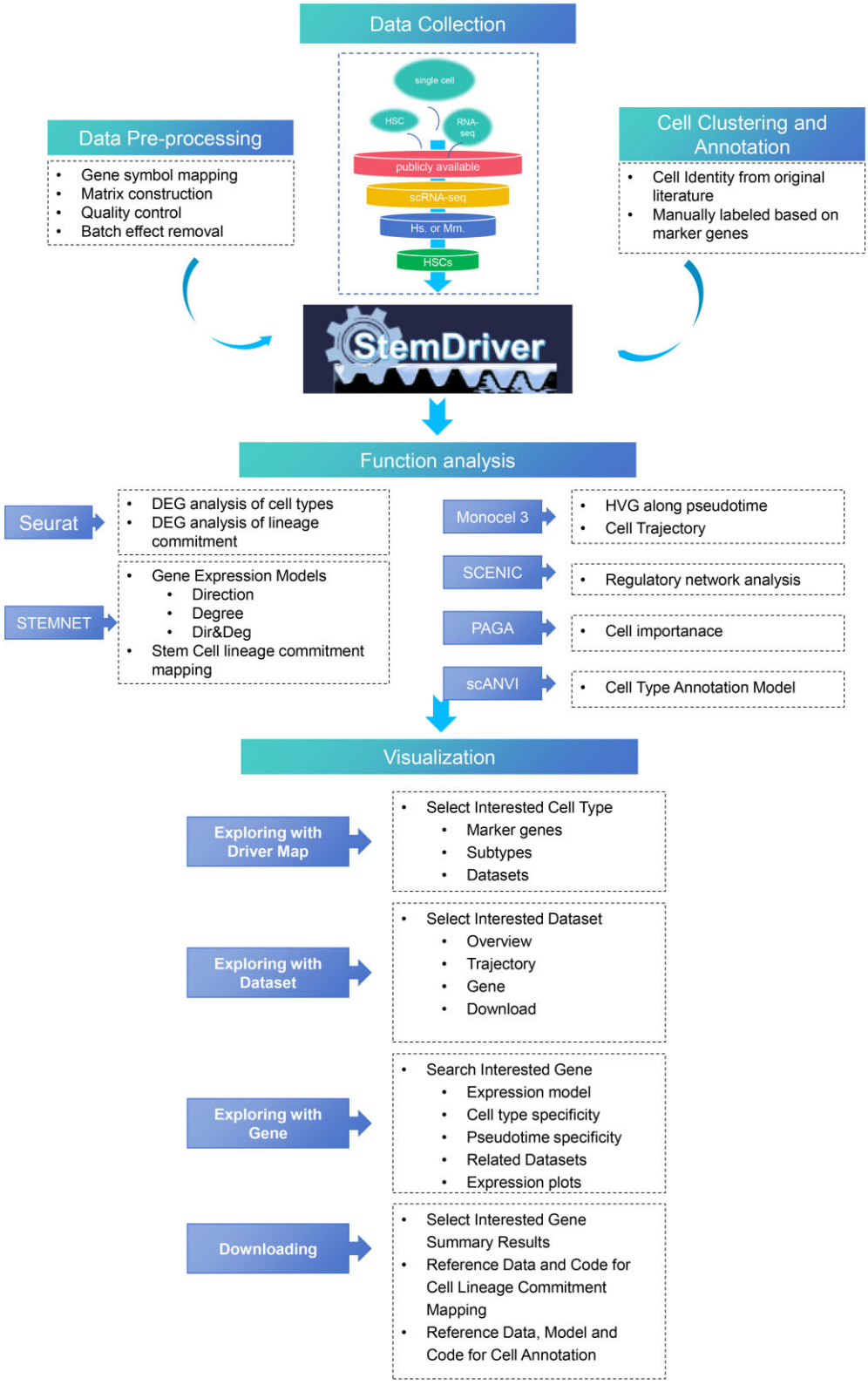


Figure 1. StemDriver workflow and overview of functional analysis. In total, there are 42 datasets covering 22 major cell types sourced from 14 different tissues. DEG, differentially expressed genes. HVG, highly variable genes. Datasets include in StemDriver screened with a uniform criterion and pre-processed with a standard workflow. StemDriver provides functional analysis of genes including Differential gene expression analysis.

progenitors (MEP). Subsequently, these progenitors were merged with fully developed mature cells.

Batch effects removal

Datasets including different sources underwent batch effects removal utilizing harmony (17) within the Seurat (Version 4.3.0) (18). In detail, Seurat objects from various samples were consolidated into a unified global object using the merge function. Following this, log-normalization and feature selection procedures were carried out. Principal components were computed using the RunPCA function with the parameter set to npcs = 50. The RunHarmony function was employed to integrate the data by specifying the reduction parameter as 'pca' The reduction method 'harmony' was applied to the integrated data, and the cells were clustered using a resolution of 0.5. The subsequent downstream analysis of the integrated data adhered to the same procedures as applied to the other datasets, ensuring consistency in our analytical approach.

Cell clustering and annotation

Datasets without integration processed with standard Seurat workflow. Briefly, the raw count matrix underwent a logarithmic transformation using a scale factor of 10 000. Then, the top 2000 highly variable genes were obtained using the FindVariableGenes function in Seurat with default parameters. Principal component analysis (PCA) was performed using the top 2000 highly variable genes, and the top 15 resulting principal components were used for subsequent UMAP analysis. The cell identities for each cluster were determined based on the original study results. In most cases, the original literature provided cell type information for each cell. However, for some studies that only provided marker genes for each cell type, the expression pattern of the marker gene list was projected onto a DotPlot. Cell identities were then manually annotated based on the expression patterns observed in the DotPlot.

Differential gene expression analysis

To identify genes that are highly expressed in specific cell types, the FindAllMarkers function in Seurat was employed. This analysis aimed to identify genes that exhibit significant differential expression in a particular cell type compared to other cell types within the dataset. For this analysis, genes with a $\log_2$ fold-change in average expression bigger than 1 or lower than −1 and an adjusted *P*-value less than 0.05 were retained as cell-type-specific highly expressed genes.

Cell Trajectory analysis with STEMNET, identification of genes with effective roles in cell differentiation

StemDriver classified genes into four categories based on gene expression patterns correlated with cell differentiation as below. (i) Direction (Dir) gene: these genes exhibit consistent up- or down-regulation from early lineage priming throughout the entire differentiation direction that the stem cells follow. (ii) Degree (Deg) gene: these genes show up or down-regulation at a specific degree of cell differentiation, independent of the differentiation direction. (iii) Dir&Deg gene: these genes are up- or down-regulated at a specific degree of differentiation in a specific direction, combining features of both the direction and degree gene models. (iv) Neither: these genes do not exhibit consistent and systematic changes during cell development, which may not play an important role during stem cell differentiation.

To identify the gene expression patterns, we employed the STEMNET packages (19). The procedure involved selecting target cell types as differentiation endpoints, using cells labeled with these target cell types as anchors, and mapping stem cells and progenitors to different directions using the runSTEMNET function. The gene expression features were then fitted into the four categories mentioned above using the mclapply function. To speed up computation, mc.cores = 40 was set. The optimal categories for each gene were determined by comparing the models' Bayesian Information Criteria.

Cell trajectory analysis with PAGA, calculating gene importance across trajectories

In this study, we utilized the Partition-based Graph Abstraction (PAGA) (20) method within the dynverse tool (21) to track gene expression changes along intricate developmental trajectories. In brief, raw and normalized counts of the top 2000 highly variable genes were employed to create a dynverse object. The selection of the root cell, possessing the lowest differentiation degree, was based on STEMNET results. Cell trajectory analysis was conducted using the 'paga-tree' method within the infer_trajectory function. Furthermore, the influence of gene expression on trajectory branch points was assessed using the calculate_milestone_feature_importance function, resulting in importance scores for each gene at each branch point.

Cell trajectory analysis with Monocle 3, identify highly variable genes along the pseudo-time

To capture the dynamic changes in gene expression from stem cells to mature cell types, we used Monocle 3 (version 1.3.1) (22), which is able to identify correlated genes on the complex trajectory. The genes that exhibit high variability in expression between cell types along the trajectory were identified using graph-autocorrelation analysis through the graph_test function of Monocle 3. Co-expression gene modules were calculated based on the identified variable genes using specific criteria, including p-value < 0.05, q-value < 0.05, and Morans_I > 0.1. Morans_I is a measurement of spatial autocorrelation, ranging from -1 to + 1. A higher value of Morans_I indicate stronger positive autocorrelation. The enrichment of co-expressed gene modules can be visualized in a heatmap. For individual genes, only those with a Morans_I > 0.5 were included in the dynamic expression profile on pseudo-time.

Gene expression regulatory network analysis

Gene regulatory network analysis was conducted by using pySCENIC (version 0.11.2) (23). First, the co-expression gene modules were calculated based on the raw count matrix. Then we built the regulons, which consist of transcriptional factors and candidate target genes, using enriched DNA motifs from gene modules with a normalized enrichment score (NES) of 3.0 or higher. The regulons enriched in lineage branches were then calculated based on area under the curve (AUC) values. The heatmap of regulons enriched in each lineage branch was included in the visualization.

Table 1. StemDriver statistics on gene analysis

Methods	Major group	Sub group	Gene number in human	Gene number in mouse
STEMNET	Gene expression associated with stem cell differentiation in terms of	Direction	14 873	13 627
		Degree	9733	12 954
		Dir&Deg	17 658	25 045
PAGA	Gene importance at branch point		2000	2000
Monocle 3	Highly variable genes in trajectory along pseudo-time (trajectory_hvg)	33	13 502	11 922
Seurat	Differential expressed genes in cell types (celltype_DEG)	70	6733	5558
Seurat	Differential expressed genes in trajectory (direction_DEG)		3009	2254

Cell annotation with scANVI approach

Cells from different datasets were integrated into the scANVI model (24). Initially, the cells underwent normalization and pre-processing using the Scanpy (version 1.9.1) workflow. Subsequently, the top 2000 highly variable genes were selected for joint embedding. Dimension reduction and clustering were carried out using the scANVI model in scvi-tools (version 0.20.3) (25). The model was trained for 500 epochs and a weight_decay set to 0. The trained scANVI model and its latent embedding were exported for transfer learning with scArches (26). When users employ our trained scANVI model for cell annotation, it is essential to configure the batch_key as 'Dataset', the labels_key as 'celltype', and the unlabeled_category as 'Unknown.

Results

Overview of stemdriver

To evaluate the role of genes in the differentiation process of hematopoietic stem cells and their progenitors, we conducted an analysis focusing on several aspects. Differentially expressed genes (DEGs) analysis evaluated gene expression on cell-type-specificity, which is conducted with Seurat software (18). Gene expression associate with stem cell differentiation in terms of direction, degree, both direction and degree, or neither were identified with STEMNET analysis (19). The gene importance of the top 2000 highly valuable genes across developmental trajectories were assessed by PAGA (20) analysis. Furthermore, the highly variable genes alone pseudo-time of each cell trajectory were identified with Monocle 3 (22) analysis. Ultimately, we comprehensively assessed the functions of 23 839 human genes and 29 533 mouse genes. Table 1 shows the overall statistics for our findings. In human, 14 873 genes influence cell differentiation direction (Table 1, Direction), 9 733 genes impact the extent of cell differentiation (Table 1, Degree), and 17 658 genes simultaneously affect both direction and extent of cell differentiation (Table 1, Dir&Deg). In mice, 13 627 genes influence cell differentiation direction, 12 954 genes affect cell differentiation degree, and 25 045 genes similarly influence both direction and degree of cell differentiation. Furthermore, across 33 differentiation trajectories, 13 502 human genes and 11 922 mouse genes were identified to influence the transition of cells from their initial state to the final state (Table 1, trajectory_hvg). Lastly, in humans, 3009 genes and in mice, 2254 genes, may potentially initiate the differentiation of stem cells into specific downstream cell types (direction_DEG). Additionally, 6 733 human genes and 5 558 mouse genes exhibit differential expression across various cell types (celltype_DEG). Users can access the detailed outcomes of the aforementioned analyses by visiting the homepage of the StemDriver website and selecting the specific functional analysis results of interest (Figure 2).

StemDriver additionally provides three modules designed to facilitate easy navigation through our analyses. The first module, named 'Driver Map,' offers comprehensive cell type information, including marker genes, subtypes, and associated datasets. The second module, 'Dataset,' offers essential contextual information and organizes analysis outcomes into five distinct sections. Lastly, the 'Gene' module offers detailed gene annotations spanning across datasets. Further elaboration on these modules is available below.

Driver map module

By utilizing the collected data, we have constructed a comprehensive lineage map that spans 22 major cell types. This map commences with the early formation of hematopoietic stem cells (HSCs) during embryogenesis and extends to the emergence of terminal unipotent cells in adults (Figure 3). Early-forming HSCs were predominantly identified in the yolk sac, AGM region, fetal liver, and fetal bone. Subsequently, hematopoietic stem cells undergo a series of sequential differentiation stages, progressively transforming into various types of blood cells to support normal hematopoiesis and immune functions. In the initial stages, hematopoietic stem cells primarily differentiate into multipotent progenitors (MPPs), which exhibit a high degree of similarity to HSCs. Following this, MPPs further differentiate into lineage-committed progenitors, including common myeloid progenitors (CMPs), granulocyte-monocyte progenitors (GMPs), lymphoid-primed multipotent progenitors (LMPs), and others. These lineage-committed progenitors subsequently undergo further differentiation into functionalized unipotent cells. For instance, CMPs undergo a stepwise transformation into megakaryocyte-erythroid progenitors (MEPs) and erythrocytes with oxygen transport functions. Similarly, LMPs experience successive differentiation stages leading to the formation of common lymphoid progenitors (CLPs), which further give rise to T cells, B cells, and NK cells involved in immune responses. Interestingly, both CMPs and LMPs can also give rise to GMPs, which subsequently mature into neutrophils, monocytes, eosinophils, basophils, dendritic cells, and macrophages. These cell types play crucial roles in maintaining normal immune functions.

While differentiation of hematopoietic stem cells within the bone marrow has been extensively studied, the understanding of HSC formation during the embryonic period is still in its early stages. StemDriver incorporates recently published

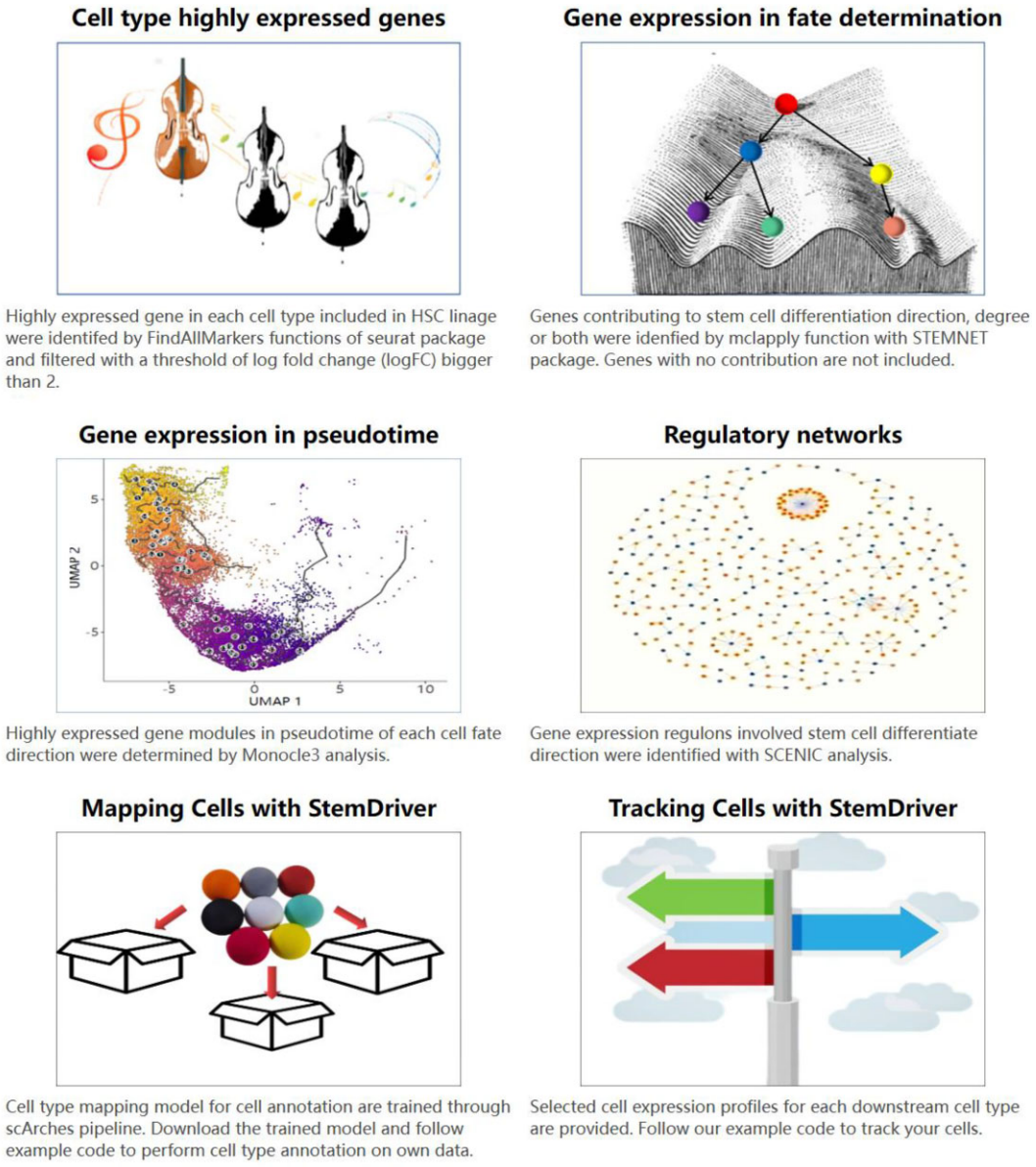


Figure 2. Summary information for analysis results. Users can download the results by choosing an interesting analysis. In addition, we provided reference cells used to analyze cell fate choice and trained the scANVI model for cell type annotation. The reference data and code can be downloaded from the StemDriver website.

sequencing data that shed light on HSC formation across different tissues during the embryonic period. Users can explore the DriverMap to select specific cell types of interest, gaining access to more detailed information such as cell subtypes recorded in StemDriver, highly expressed genes, and associated datasets.

Dataset module

In the Dataset module, users can access literature information for each dataset (Figure 4A). The analysis results for each dataset are presented in four categories: 'Overview', 'Trajectory', 'Gene', 'Download'. The 'Overview' section includes UMAP plots for visualizing cell composition, donut plots depicting cell numbers per cell type, and histograms displaying cell type proportions at different stages (Figure 4B). In the 'Trajectory' section, In the 'Trajectory' section, we presented a broad overview of four analysis methods. STEMNET analysis was employed to predict potential lineage commitments of stem cells and progenitor cells and visualize the results in a star plot. Cells positioned at the center of the plot represent the lowest degree of differentiation, while cells situated at the vertices of the plot correspond to the highest degree of differentiation. This indicates a more mature status within the corresponding lineages. In contrast to the STEMNET analysis, the PAGA analysis also predicted cell trajectories. However, PAGA analysis diverged by evaluating gene importance at each trajectory branch point. Subsequently, we applied Monocle 3 analysis to study the development of each cell lineage. During this analysis, genes that potentially play a significant role along pseudo-time were assessed. Readers have the option to select a trajectory of interest to gain insight into the overall status of cells and evaluate the importance of specific genes in that context. In addition to the three trajectory analysis methods mentioned above, we also conducted an analysis of the gene expression regulatory network using SCENIC.

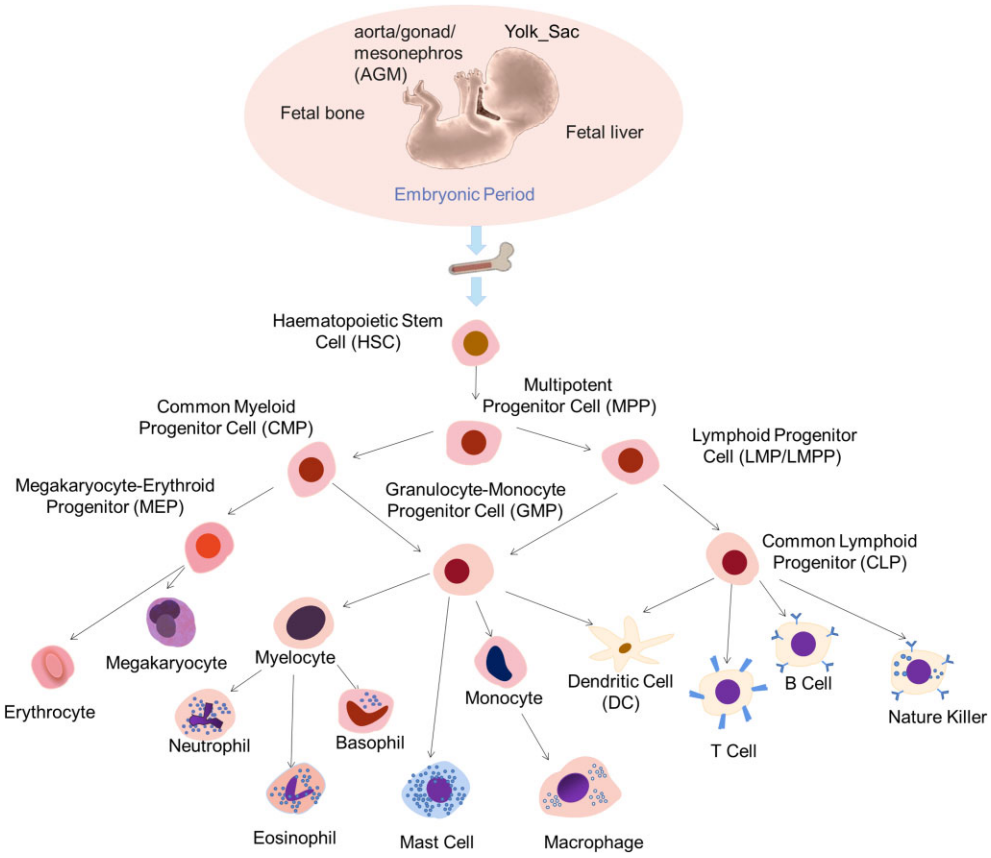


Figure 3. DriverMap, a complete lineage map of hematopoiesis based on collected data.

The enrichment of gene sets for each developmental trajectory can be visualized through the heatmap. The results of gene evaluation from each of these analysis approaches were summarized and can be referred to in the table located at the bottom (Figure 4C, Supplementary Figure 1B). In the 'Gene' section (Figure 4D, we offer detailed visualizations of genes that stand out from the analyses mentioned earlier. It's important to note that since the above methods characterize genes from different perspectives, not all genes may have consistent representations across all analysis results. Readers have the flexibility to search for their specific genes of interest to gain a comprehensive understanding of their characteristics. Furthermore, the results of these analyses from the four methods can be downloaded from the 'Download' section.

Taking the E-MTAB-11343 dataset as an example (Supplementary Figure 1) (3), this dataset comprises sixteen distinct cell types, totaling 22 098 cells. These cells originate from human embryos, range from the 4th post-conception week (PCW) to the 17th PCW (Supplementary Figure 1A). This dataset encompasses various matured cell types. Cell development lineages can be visualized through the star plot generated using STEMNET analysis. As depicted in the image (Supplementary Figure 1B), the central portion is composed of stem cells and progenitors, while B cells, T cells, and other matured cells are situated at the vertices. Cells that have initiated differentiation are progressing toward their respective directions (Supplementary Figure 1B.i). The PAGA analysis segmented the cell trajectories into 27 branch points, and it assessed the importance of the top 2000 highly valuable genes at these branch points. The results of this evaluation are available in the table (Supplementary Figure 1B.ii, B.iii). Information about the enrichment of gene regulons for cell lineages can be found in the PySCENIC method section (Supplementary Figure 1C).

Gene module

Users can search their interested genes either through 'Dataset' page or 'Gene' page online. The 'Dataset' page, included the discovered genes obtained from above four analysis methods. Taking MZB-associated genes (MZB1) as an example (Figure 5A), the star plot showcases the expression levels of MZB1 across different lineage commitments. Based on the results of differential expression analysis, MZB1 exhibits higher expression in B cells and T cells, as well as within subgroups of MLPs that undergo differentiation into B cells and T cells (Supplementary Figure 1C). In order to catch the gene expression characteristics alone pseudo-time, we checked the results obtained from Monocle 3 (Supplementary table 1). Monocle 3 evaluated the significance of genes along pseudo-time by assessing their autocorrelation, which represented as the Morans_I value. The Morans_I value ranges from -1 to + 1, with a higher value indicating a stronger positive autocorrelation. This suggests a higher possibility of gene impact on cell development. According to the results from Monocle 3, MZB1 has a relatively high Morans_I value in the T cell (0.490295551) and B cell (0.260621267) development lineage, which is consistent with the results of DEG analysis. In addition, the results of PAGA analysis showed that MZB1 had a significant impact on cell cluster of 8, 13, 11, 1, 7, 14, 2, which is a mix of B cell and progenitors, or a mix of T cell and progenitors.

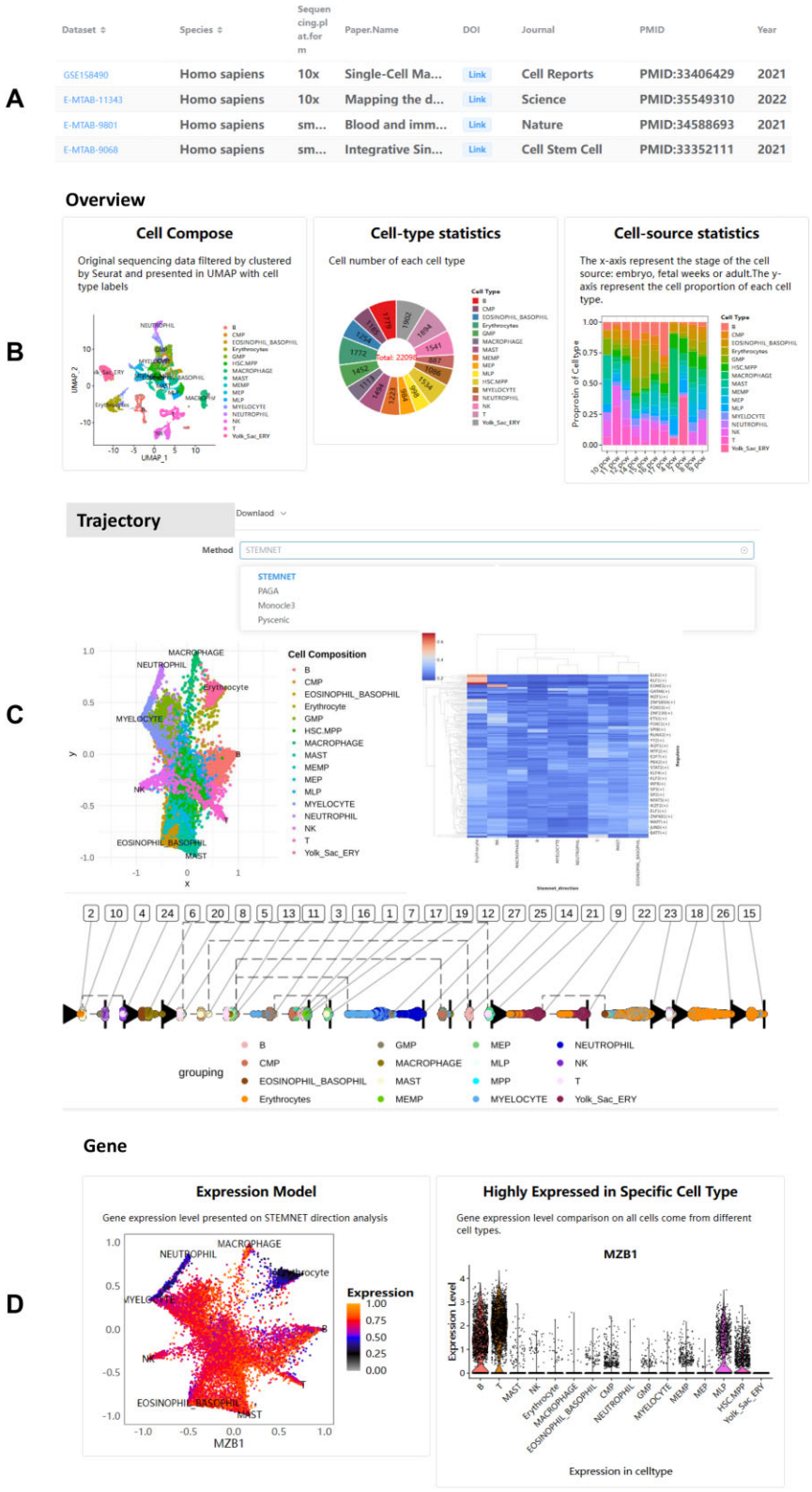


Figure 4. Exploring dataset module. (**A**) The literature information on datasets can be found in the table. (**B**) The 'Overview' section. General cell type information is included in the 'Overview' section. (**C**) The 'Trajectory' section offers a summarized overview of the results from trajectory analysis conducted using STEMNET, PAGA, and MONOCEL3. In the STEMNET analysis, potential lineage commitments of stem cells and progenitor cells were predicted. PAGA analysis assessed the importance of genes at each trajectory branch point, while Monocle 3 analysis identified significant genes along the pseudo-time. Furthermore, the results of the gene regulatory network analysis conducted with SCENIC are also included in this section. (**D**) The 'Gene' section contains the expression profiles of each gene. Due to the extensive number of profiles, expression data is only available for genes included in the analysis results. As different analysis methods are applied, not every gene may be included in every set of results, and consequently, not all genes will have displays for every analysis method.

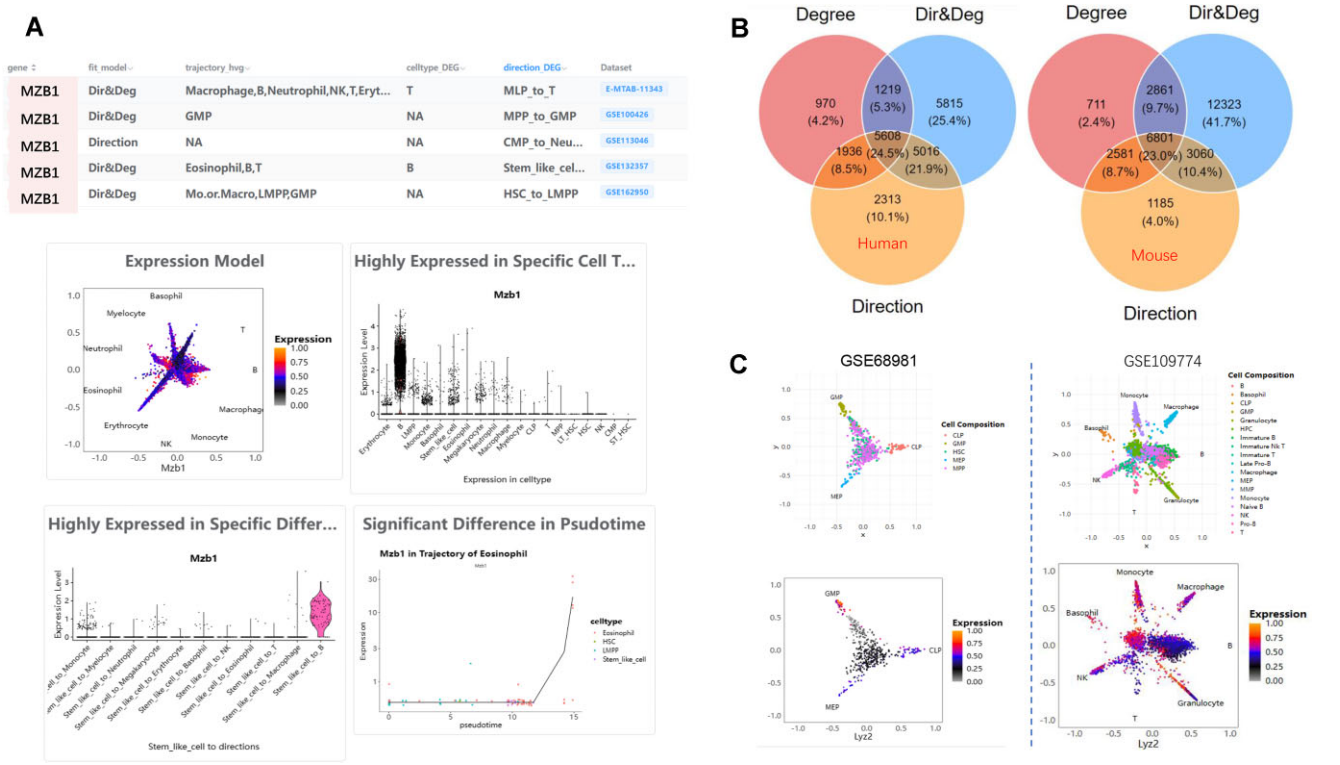


Figure 5. Results exploring with gene. (**A**) Example of summarized gene information obtained by searching. (**B**) Statistical analysis of gene expression patterns correlates with differentiation. (**C**) Expression of gene Lyz2 in datasets of GSE68981 and Gse109774.

Above analysis results indicate that MZB1 is likely involved in the development of both B cells and T cells. MZB1 is established as a marker gene for marginal zone B cells (MZB) and is well-documented for its substantial role in antibody secretion (27,28). Additionally, a recent study reported an upregulation of MZB1 in transitional B cells with high IgM expression, and these cells were identified as following a developmental trajectory towards MZB cells (29). The involvement of MZB1 in T cell development has also been documented, with evidence demonstrating that targeting the MZB1 gene with microRNA-185 leads to the arrest of T cell development (30).

To gain a more comprehensive understanding of the role of MZB1, we further examined its expression in other datasets. As indicated in supplementary table 2, the MZB1 gene exhibited significantly higher expression across multiple cell types, spanning various stages of B cell development, including pre-pro-B, pro-B, late pro-B and B cells. This strongly suggests that MZB1 plays a crucial role in B cell development. A previous study has reported that the expression of Mzb1 serves as a hallmark for identifying pro- and pre-B cells, with these genes playing essential roles in governing the proliferation and maturation of B cells (31). Additionally, MZB1 has been observed to exhibit high expression in plasmacytoid dendritic cells (pDC) in multiple datasets, suggesting a significant connection between MZB1 and pDC. While there are limited reports on the specific role of MZB1 in pDC development, it has been documented to enhance the immune functions of pDC (27)'.

Conclusion and discussion

StemDriver aims to provide valuable data and gene information to assist researchers in identifying potential candidates for future research. StemDriver have curated a comprehensive scRNA-seq dataset that spans from the embryonic phase to adulthood, capturing the entire journey from the initial formation of hematopoietic stem cells to the maturation of fully functional terminal cells. StemDriver provided a comprehensive evaluation of gene characteristics related to cell differentiation. STEMNET categorized genes based on their expression patterns concerning stem cell differentiation. PAGA analysis assessed gene importance at trajectory branch points. Monocle 3 identified highly variable genes along pseudotime, and DEG analysis revealed cell-type-specific gene expression. StemDriver offers a platform for cross-dataset comparisons. We have collected data from 42 different studies and performed standardized analysis. By leveraging resources from various studies, users can gain a more comprehensive understanding of gene expression features related to cell differentiation.

After being processed with a standardized workflow, a total of 23839 human genes and 29533 mouse genes were identified. These genes were annotated based on their association with stem cell differentiation, their significance at trajectory branch points, their expression along pseudo-time on specific cell trajectory development, and their expression specificity across cell types. The statistical analysis (Figure 5B) clearly demonstrates that over 50% of the identified genes with impactful roles fall under the 'Dir&Deg' category for both humans and mice. This underscores their crucial involvement in driving the differentiation of stem cells and progenitors towards distinct lineage commitments. However, a notable portion of genes also align with multiple expression patterns. This variability might stem from their distinct functions across diverse lineage branches or involvement in different cells associated with lineage commitment. For instance, the gene Lyz2

(Lysozyme 2) was categorized as a 'Degree gene' in the dataset GSE6898 (32), indicating that Lyz2 doesn't show specific trajectory preferences while promoting HSC differentiation into GMP, MEP, and CLP. However, in dataset GSE10977 (33), it was classified as a 'Dir&Deg' gene, indicating that its expression promotes stem cells to differentiate into specific unipotent cells. This observation concurs with the outcomes of gene differential expression analysis in the directional context. Importantly, Lyz2 demonstrates increased expression in progenitors undergoing differentiation into granulocytes and monocytes (Figure 5C). This discovery enhances our comprehension of the roles played by individual genes during varied differentiation processes.

StemDriver offers comprehensive gene annotations from multiple perspectives, rendering it highly valuable in the field of cell engineering. Although there has been well-documented progress in cell-based immunotherapy against blood cancers, the limited availability of specific cell subsets has presented significant challenges in the development and implementation of these therapies. Consequently, extensive studies have been directed towards modifying induced pluripotent stem cells (iPSCs) to generate desired cell subsets, including iPSC-derived natural killer (NK) cells (34,35). To identify genes that play a crucial role in initiating the differentiation of stem cells and their progenitors into NK cells, we conducted a screening using the StemDriver database, focusing on the differentiation stage of NK cells. As a result, we identified a total of 321 genes. The functions of these genes are pivotal in governing various aspects of NK cell behavior. For instance, IL32 (36) plays a role in NK cell activity, (HSPA6) (37) is associated with cytotoxicity, and HLA-A, HLA-B and HLA-C (38) are involved in recognition processes. StemDriver offers thorough gene expression annotations across the trajectory of target cells derived from diverse progenitors. This resource empowers researchers with a comprehensive grasp of gene functions, aiding in the identification of efficient target genes for iPSC engineering. Moreover, it can be instrumental in designing a cell culture environment that facilitates the stimulation of target gene expression.

The primary objective of StemDriver is to offer an intricate molecular overview of hematopoiesis, wherein the intricate molecular changes within specific cell subgroups are best comprehended through single-cell resolution sequencing. Therefore, the bulk cell data is not included. While single-cell assay for transposase-accessible chromatin with high-throughput sequencing (scATAC-seq) also provides information at single-cell resolution, there is currently insufficient data to cover all the relevant cell types involved in hematopoiesis. Therefore, we are planning to include the scATAC-seq data in the near future. By leveraging a wide range of omics data, we aim to construct a more comprehensive regulatory landscape of hematopoietic differentiation, spanning from the initial formation of hematopoietic stem cells to the development of each distinct functional terminal cell type. In summary, StemDriver serves as a pioneering and unique database that offers a systematic characterization of molecular functions in HSC formation and differentiation at the single-cell level. The wealth of information contained in StemDriver presents numerous intriguing stories and insights waiting to be explored by users.

Data availability

All data and results can be downloaded on the StemDriver website (http://biomedbdc.wchscu.cn/StemDriver/). The trained scANVI model and code used to perform cell annotation on new data are also included on the download page. For more details and discussions, contact Dr Xiaobo Zhou.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We would like to thank Mrs. Rong Yu and Mr. Shengqiang Zhou from West China Biomedical Big Data Center for their support on computer server maintenance. Luo YY would like to acknowledge Dr Ailin Zhang for certain discussions. We also would like to thank the researchers who share the single-cell sequencing data for public study in the field of hematopoiesis.

Funding

Luo etc. was supported by Center of Excellence-International Collaboration Initiative Grant, West China Hospital, Sichuan University [139170052]; Sichuan Science and Technology Program [2023YFS0200, 2022YFS0228]; Drs Zhou, Zhao and Wen were partially supported by NIH [R01GM123037, U01AR069395, R01CA241930]; NSF [2217515, 2326879]; Dr. Kim is partially supported by NIH [R35GM138184]. Funding for open access charge: West China Hospital, Sichuan University [139170052].

Conflict of interest statement

None declared.

References

1. Skulimowska,I., Sosniak,J., Gonka,M., *et al.* (2022) The biology of hematopoietic stem cells and its clinical implications. *FEBS J.*, **289**, 7740–7759.
2. Gunsilius,E., Gastl,G. and Petzer,A.L. (2001) Hematopoietic stem cells. *Biomed. Pharmacother.*, **55**, 186–194.
3. Suo,C., Dann,E., Goh,I., *et al.* (2022) Mapping the developing human immune system across organs. *Science*, **376**, eabo0510.
4. Bunis,D.G., Bronevetsky,Y., Krow-Lucal,E., *et al.* (2021) Single-cell mapping of progressive fetal-to-adult transition in human naive T cells. *Cell.Com.*, **34**, 1.
5. Zheng,Z., He,H., Tang,X.T., *et al.* (2022) Uncovering the emergence of HSCs in the human fetal bone marrow by single-cell RNA-seq analysis. *Cell Stem Cell*, **29**, 1562–1579.
6. Bagger,F.O., Sasivarevic,D., Sohi,S.H., *et al.* (2016) BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. *Nucleic Acids Res.*, **44**, D917–D924.
7. Choi,J., Baldwin,T.M., Wong,M., *et al.* (2019) Haemopedia RNA-seq: a database of gene expression during haematopoiesis in mice and humans. *Nucleic Acids Res.*, **47**, D780–D785.
8. Sánchez-Castillo,M., Ruau,D., Wilkinson,A.C., Ng,F.S.L., Hannah,R., Diamanti,E., Lombard,P., Wilson,N.K. and Gottgens,B. (2015) CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.*, **43**, D1117–D1123.
9. Ueno,H. and Weissman,I.L. (2010) The origin and fate of yolk sac hematopoiesis: application of chimera analyses to developmental studies. *Int. J. Dev. Biol.*, **54**, 1019–1031.
10. Lewis,K., Yoshimoto,M. and Takebe,T. (2021) Fetal liver hematopoiesis: from development to delivery. *Stem. Cell Res. Ther.*, **12**, 139.

11. Drissen,R., Thongjuea,S., Theilgaard-Mönch,K., *et al.* (2019) Identification of two distinct pathways of human myelopoiesis. *Sci. Immunol.*, **4**, eaau7148.
12. Carlson,M. (2021) org.Hs.eg.db: Genome wide annotation for Human. R package version 3.14.0. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.14.0. (**2021**).
13. Marc,C. (2021) org.Mm.eg.db: Genome wide annotation for Mouse. R package version 3.14.0. org.Mm.eg.db: Genome wide annotation for Mouse. R package version 3.14.0. (**2021**).
14. Seal,R.L., Braschi,B., Gray,K., *et al.* (2023) Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.*, **51**, D1003–D1009.
15. Blake,J.A., Baldarelli,R., Kadin,J.A., *et al.* (2021) Mouse Genome Database (MGD): knowledgebase for mouse-human comparative biology. *Nucleic Acids Res.*, **49**, D981–D987.
16. Aibar,S., González-Blas,C.B., Moerman,T., *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**, 1083–1086.
17. Korsunsky,I., Millard,N., Fan,J., *et al.* (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods*, **16**, 1289–1296.
18. Stuart,T., Butler,A., Hoffman,P., *et al.* (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
19. Velten,L., Haas,S.F., Raffel,S., *et al.* (2017) Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.*, **19**, 271–281.
20. Wolf,F.A., Hamey,F.K., Plass,M., *et al.* (2019) PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.*, **20**, 59.
21. Saelens,W., Cannoodt,R., Todorov,H., *et al.* (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, **37**, 547–554.
22. Cao,J., Spielmann,M., Qiu,X., *et al.* (2019) The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, **566**, 496–502.
23. Van de Sande,B., Flerin,C., Davie,K., *et al.* (2020) A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.*, **15**, 2247–2276.
24. Xu,C., Lopez,R., Mehlman,E., *et al.* (2021) Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.*, **17**, e9620.
25. Lopez,R., Regier,J., Cole,M.B., *et al.* (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, **15**, 1053–1058.
26. Lotfollahi,M., Naghipourfar,M., Luecken,M.D., *et al.* (2022) Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.*, **40**, 121–130.
27. Kapoor,T., Corrado,M., Pearce,E.L., *et al.* (2020) MZB1 enables efficient interferon α secretion in stimulated plasmacytoid dendritic cells. *Sci. Rep.*, **10**, 21626.
28. Flach,H., Rosenbaum,M., Duchniewicz,M., *et al.* (2010) Mzb1 protein regulates calcium homeostasis, antibody secretion, and integrin activation in innate-like B cells. *Immunity*, **33**, 723–735.
29. Tull,T.J., Pitcher,M.J., Guesdon,W., *et al.* (2021) Human marginal zone B cell development from early T2 progenitors. *J. Exp. Med.*, **218**, e20202001.
30. Belkaya,S., Murray,S.E., Eitson,J.L., *et al.* (2013) Transgenic expression of microRNA-185 causes a developmental arrest of T cells by targeting multiple genes including Mzb1. *J. Biol. Chem.*, **288**, 30752–30762.
31. Anderson,D., Skut,P., Hughes,A.M., *et al.* (2020) The bone marrow microenvironment of pre-B acute lymphoblastic leukemia at single-cell resolution. *Sci. Rep.*, **10**, 19173.
32. Yang,J., Tanaka,Y., Seay,M., *et al.* (2017) Single cell transcriptomics reveals unanticipated features of early hematopoietic precursors. *Nucleic Acids Res.*, **45**, 1281–1296.
33. Schaum,N., Karkanias,J., Neff,N.F., *et al.* (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**, 367–372.
34. Lupo,K.B., Moon,J.I., Chambers,A.M., *et al.* (2021) Differentiation of natural killer cells from induced pluripotent stem cells under defined, serum- and feeder-free conditions. *Cytotherapy*, **23**, 939–952.
35. Woan,K.V., Kim,H., Bjordahl,R., *et al.* (2021) Harnessing features of adaptive NK cells to generate iPSC-derived NK cells for enhanced immunotherapy. *Cell Stem Cell*, **28**, 2062–2075.
36. Gorvel,L., Korenfeld,D., Tung,T., *et al.* (2017) Dendritic cell-derived IL-32α: a novel inhibitory cytokine of NK cell function. *J. Immunol.*, **199**, 1290–1300.
37. Yang,D., Zhao,F., Su,Y., *et al.* (2023) Integrated analysis of single-cell and bulk RNA-sequencing identifies a signature based on NK cell marker genes to predict prognosis and immunotherapy response in hepatocellular carcinoma. *J. Cancer Res. Clin. Oncol.*, **149**, 10609–10621.
38. Wagtmann,N., Rajagopalan,S., Winter,C.C., *et al.* (1995) Killer cell inhibitory receptors specific for HLA-C and HLA-B identified by direct binding and by functional transfer. *Immunity*, **3**, 801–809.

Received: August 18, 2023. **Revised:** October 24, 2023. **Editorial Decision:** October 24, 2023. **Accepted:** November 1, 2023