Data in Brief 36 (2021) 107080

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Genetic pathways regulating hematopoietic lineage speciation: Factorial latent variable model analysis of single cell transcriptome

Zhaoyan Liu [a,*], Wei Zhu [a,*], Dmitri V. Gnatenko [b], Natasha M. Nesbitt [b], Wadie F. Bahou [b,*]

[a] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794 (USA)
[b] Department of Medicine, Stony Brook University, Stony Brook, NY 11794 (USA)

ARTICLE INFO

Article history:
Received 4 January 2021
Revised 25 March 2021
Accepted 15 April 2021
Available online 22 April 2021

Keywords:
Single-cell RNA sequencing analysis
Pathway annotation
Factor analysis
Spatial reconstruction

ABSTRACT

Genetic pathways regulating hematopoietic lineage commitment at critical stages of development remain incompletely characterized. To better delineate genetic sources of variability regulating cellular speciation during steady-state hematopoiesis, we applied a factorial single-cell latent variable model (f-scLVM) to decompose single-cell transcriptome heterogeneity into interpretable biological factors (refined pathway annotations or gene sets without annotation) dynamically regulating cell fate. Hematopoietic single cell transcriptomic raw sequencing data extracted from 1,920 hematopoietic stem and progenitor cells (HSPCs) derived from 12-week-old female mice were used for data analysis and model development. These single cell RNA sequencing data were subsequently analyzed using the factorial single-cell latent variable model (f-scLVM), with their heterogeneity decomposed into interpretable biological factors. The top biological factors underlying the basal hematopoiesis were subsequently identified for the aggregate, and lineage-restricted (myeloid, megakaryocyte, erythroid) progenitor cells. For a subset of factors, data were independently verified experimentally

DOI of original article: 10.1016/j.freeradbiomed.2020.12.015
* Corresponding author.
E-mail addresses: zhaoyan.liu@stonybrook.edu (Z. Liu), wei.zhu@stonybrook.edu (W. Zhu), wadie.bahou@stonybrookmedicine.edu (W.F. Bahou).

https://doi.org/10.1016/j.dib.2021.107080

in a companion research paper [1]. These data facilitate the identification of novel subpopulations and adjust gene sets to discover new marker genes and hidden confounding factors driving basal hematopoiesis.

Specifications Table

Subject	Biological sciences
Specific subject area	Single cell transcriptomics, Hematopoiesis
Type of data	Tables Figures
How data were acquired	Raw data was acquired from GEO (accession number GSE81682, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81682). Data analysis was completed using the R software
Data format	Analyzed Secondary Data
Parameters for data collection	Quality control: removed cells expressing (i) less than 4,000 detected genes, (ii) less than 200,000 reads mapped to nuclear genes, or (iii) more than 10% of mapped reads mapping to the mitochondrial genome. Normalization: used function Seurat::NormalizeData() with normalization.method = "LogNormalize", scale.factor = 10000, margin = 1. Select highly variable genes: used function Seurat::FindVariableFeatures() with selection.method = "vst", nfeatures = 2000. Scaling: used function Seurat::ScaleData() for all genes. PCA: used function Seurat::RunPCA() with npcs=30. Clustering: used 30 PCs as reduction, set k=25 for the kNN algorithm, and selected 0.7 as the resolution parameter. Fsclvm: used hallmark gene sets derived from MSigDB version 7.0 as annotated gene sets, set the number of hidden factors to fit in the model to be 5, the minimum number of genes required in order to retain a gene set for analysis was set to 15, and the maximum number of iterations to use in training the model was set to 8000.
Description of data collection	The secondary data was obtained by performing quality control, normalization, highly variable genes (HVGs) selection and scaling on the raw count data. Based on HVGs, PCA was conducted, and then clustering was done on by using top 30 PCs. The fsclvm was trained by using all genes in the secondary data and the hallmark gene sets derived from MSigDB version 7.0. The loadings of top 30 genes in top 2 annotated factors of clusters were printed out.
Data source location	Institution: Department of Applied Mathematics and Statistics, Stony Brook University City/Town/Region: Stony Brook Country: United States Primary data source: Institution: Lab Gottgens, Department of Haematology, University of Cambridge City/Town/Region: Cambridge Country: United Kingdom Data repository: GEO (GSE81682,https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81682)
Data accessibility	Liu, Zhaoyan (2021), "HSPC_FSCLVM_DATA&CODE", Mendeley Data, V1, doi: 10.17632/3cxw2s7jw5.1 The link is: http://dx.doi.org/10.17632/3cxw2s7jw5.1
Related research article	Natasha M. Nesbitt, Lisa E. Malone, Zhaoyan Liu, Alexander Jares, Dmitri V. Gnatenko, Yupo Ma, Wei Zhu, and Wadie F. Bahou. Divergent erythroid megakaryocyte fates in Blvrb-deficient mice establish non-overlapping cytoprotective functions during stress hematopoiesis. Free Radic Biol Med, Volume 164, 2021, Pages 164-174, ISSN 0891-5849, https://doi.org/10.1016/j.freeradbiomed.2020.12.015.

Value of the Data

- The data shows the biological factors (gene sets) that contribute the most to the heterogeneity in gene expression in hematopoietic stem cell populations during basal haematopoiesis. The contributions (relevance) of individual factors (genes) are estimated, pathway annotations are refined, and factors without annotation are inferred.
- Researchers who study hematopoietic lineage development would find this data helpful.
- The latent factors identified can help elucidate the underlying biological mechanism and be used in the clustering analysis of HSPCs to identify sub genetic/phenotypic populations.

1. Data Description

1.1. Raw data

Hematopoietic single cell transcriptomic raw sequencing data were obtained from GEO (accession number GSE81682) [2] and included 1,920 hematopoietic stem and progenitor cells (HSPC) derived from 12-week-old female mice.

1.2. Secondary data generated through statistical analyses

1.2.1. Part I: Estimated latent factors obtained via factorial single-cell latent variable model (f-scLVM)

As described in the method section below, each latent factor generated is simply a linear combination of known gene sets and unknown major factors corresponding to the diversity of the single cell transcriptome profiles. They are ranked based on the percentage of variation explained, with that accounting for the most ranked the highest and so on, so forth. The biological identity of each estimated factor can be inferred based on the known gene set and pathway information.

Fig. 1 shows the weights of the most important genes in the top 2 pathways (G2M CHECKPOINT and E2F TARGETS) identified by f-scLVM for cells in Cluster 2, which is annotated as

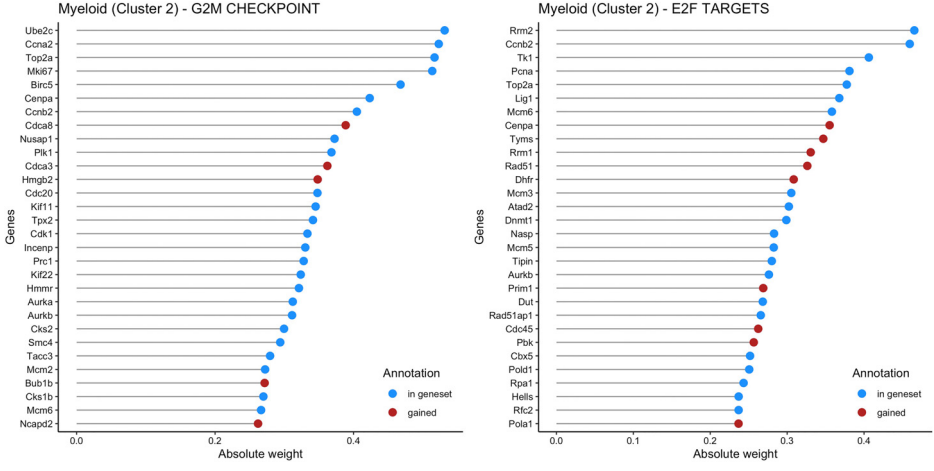

Fig. 1. Weights for the most important genes in the top 2 factors of Myeloid.

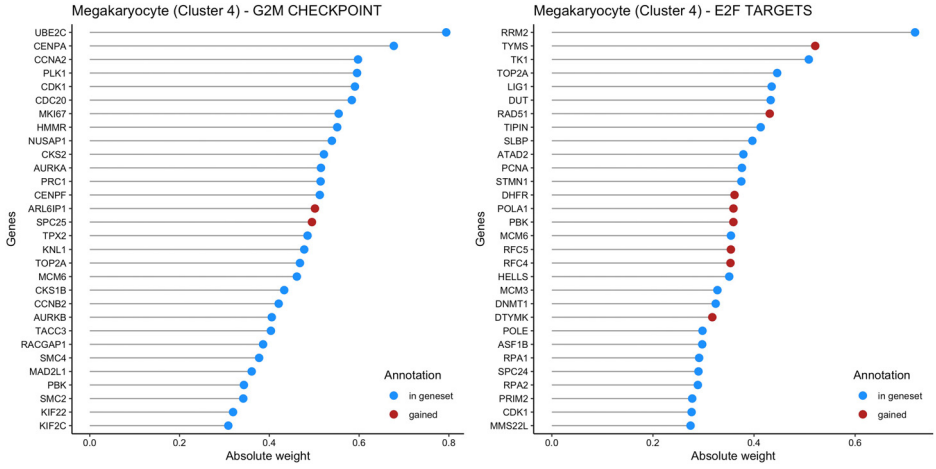

Fig. 2. Weights for the most important genes in the top 2 factors of Megakaryocyte.

Myeloid. Genes that were pre-annotated by MSIGDB are in *blue* and genes added by the model are in *red*.

Fig. 2 shows the weights of the most important genes in the top 2 pathways (G2M CHECKPOINT and E2F TARGETS) identified by f-scLVM for cells in Cluster 4, which is annotated as Megakaryocyte. Genes that were pre-annotated by MSIGDB are in *blue* and genes added by the model are in *red*.

Fig. 3 shows the weights of the most important genes in the top 2 pathways identified by f-scLVM for cells in Cluster 7, 8 and 9. Three clusters represent three successive stages of Erythroid development. Genes that were pre-annotated by MSIGDB are in *blue* and genes added by the model are in red.

Fig. 4 shows two t-distributed stochastic neighbor embedding (t-SNE) plots of the single-cell variation captured by annotated factors (left) and unannotated factors (right), where each point is a cell and it is positioned based on the cell embeddings determined by the reduction technique. Colors of points correspond to cell types identified in [2]. (LT-HSC represents long-term hematopoietic stem cells; HSPC represents hematopoietic stem progenitor cells; Prog represents progenitor cells.) While annotated factors tend to resolve intra-cell type variation, unannotated factors tend to capture inter-cell type differences that cannot be readily captured by the annotated factors.

1.2.2. Part II: Principal components obtained via the principal component analysis (PCA)

As described in the method section below, each principal component generated is also a linear combination of the genes. They are ranked based on the percentage of variation explained, with that accounting for the most ranked the highest and so on, as well. However, no prior biological knowledge is utilized in the entire analysis, and therefore the resulting principal components are hard to relate to, biologically as this method can hardly separate underlying biological factors responsible for the bio-diversity, from random noises or other biases. This is what we have found in our studies too. In the accompanying research paper [1], we were able to experimentally validate the top latent factors – however, we could not deduce the biological identities of the principal components. This contrast confirmed the superiority of the latent factor approach.

Fig. 5 shows the weights (loadings) of the most important genes in the top 2 PCs.

Fig. 6 visualizes cell embeddings of top 2 PCs, where each point is a cell and it is positioned based on the cell embeddings determined by the first 2 PCs. Colors of points correspond to cell

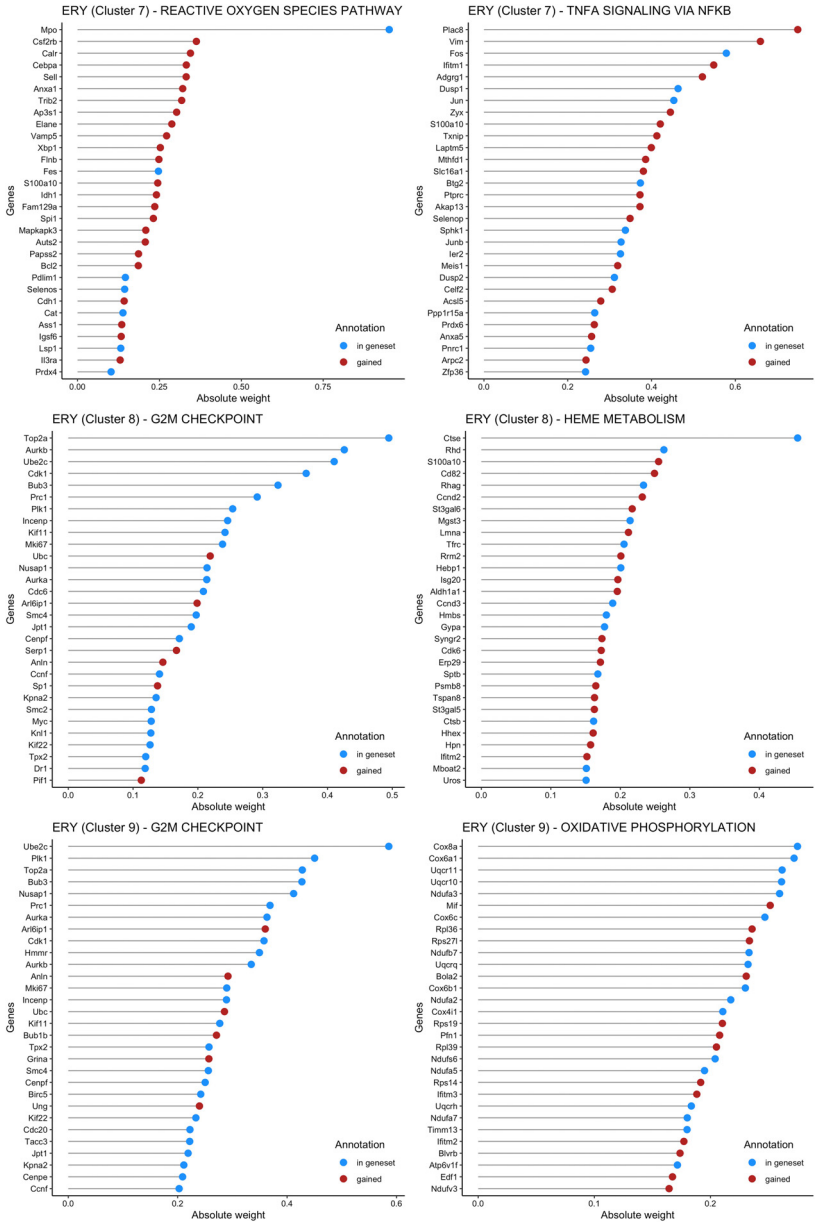

Fig. 3. Weights for the most important genes in the top 2 factors of Erythroid.

types identified in [2]. (LT-HSC represents long-term hematopoietic stem cells; HSPC represents hematopoietic stem progenitor cells; Prog represents progenitor cells.)

Fig. 7 shows a t-distributed stochastic neighbor embedding (t-SNE) plot of the single-cell variation captured by first 30 Principal Components (PCs), where each point is a cell and it is positioned based on the cell embeddings determined by the reduction technique. Colors of points correspond to cell types identified in [2]. (LT-HSC represents long-term hematopoietic stem cells; HSPC represents hematopoietic stem progenitor cells; Prog represents progenitor cells.)

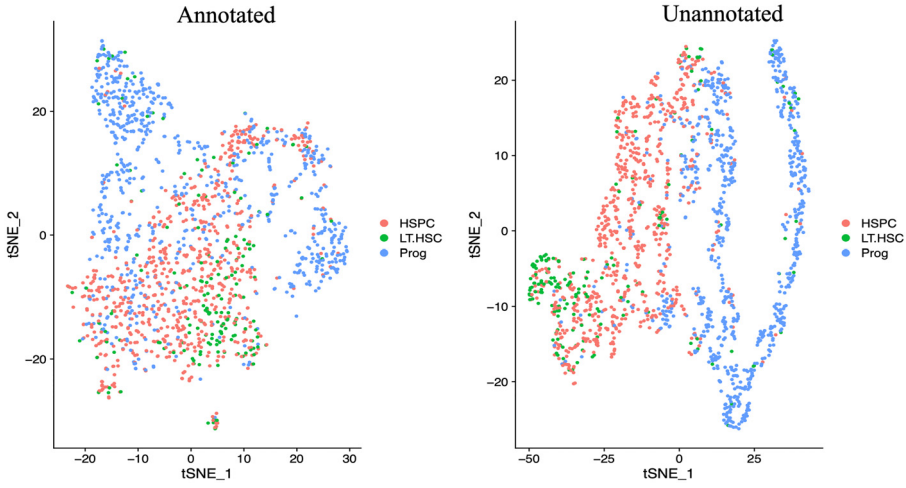

Fig. 4. t-SNE visualization of the single-cell variation captured by annotated factors (left) and unannotated factors (right).

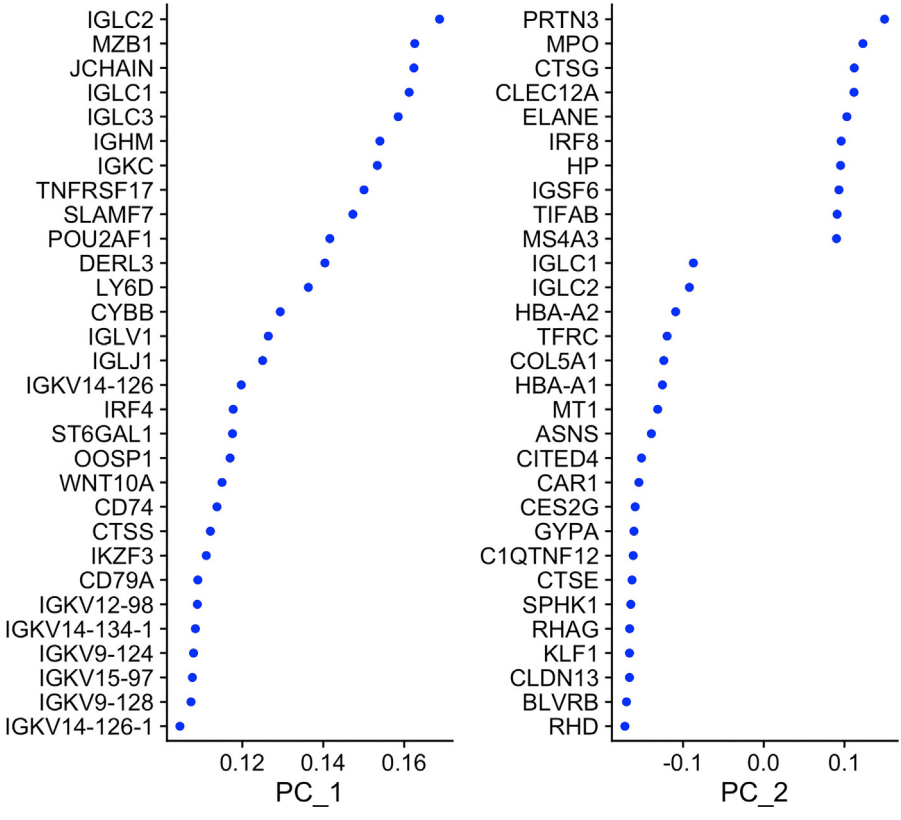

Fig. 5. Weights for the most important genes in the top 2 PCs.

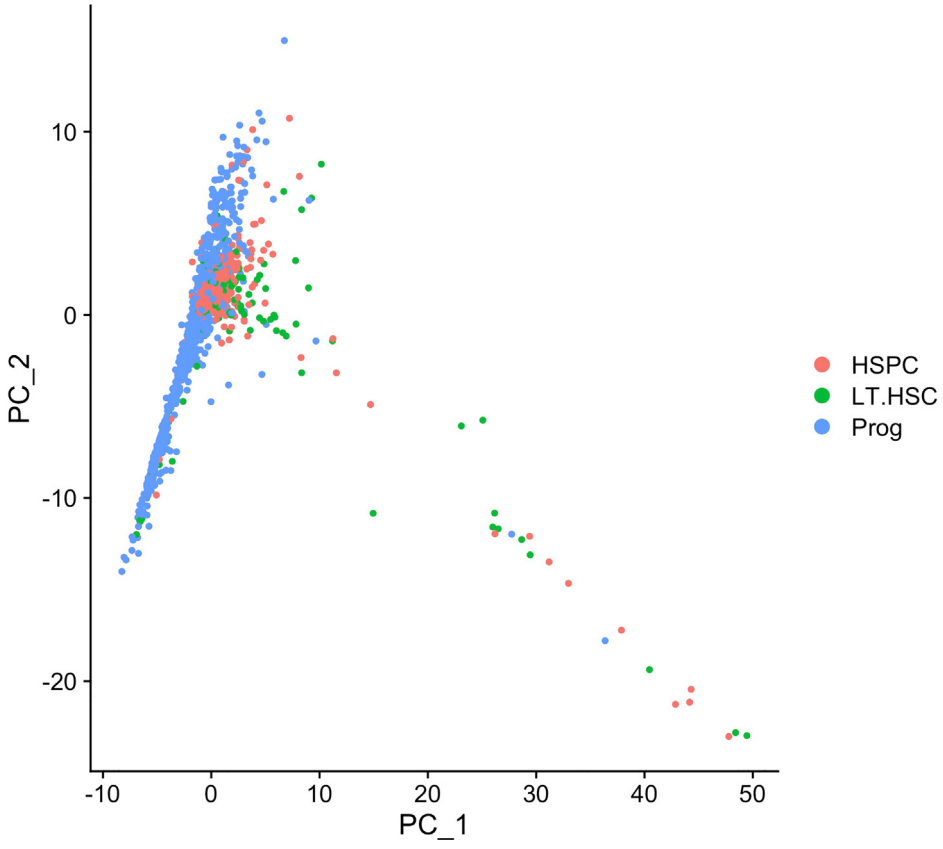

Fig. 6. PCA plot for cells.

2. Experimental Design, Materials and Methods

2.1. Preprocessing

Pre-processing step excluded 264 cells expressing (i) less than 4,000 detected genes, (ii) less than 200,000 reads mapped to nuclear genes, or (iii) more than 10% of mapped reads mapping to the mitochondrial genome. We normalized the data for each cell by the total expression, multiplied this by a scale factor 10,000, and log-transformed the results. Genes exhibiting high variability across cells were identified using the variance-stabilizing method introduced in [6], and 2,000 genes with the highest standardized variance were selected.

2.2. Factor analysis

In order to decompose single-cell transcriptome heterogeneity into interpretable biological drivers, the factorial single-cell latent variable model (f-scLVM) [3] was applied. F-scLVM is based on sparse factor analysis, a linear latent variable model for dimensionality reduction. It factorizes the gene expression matrix into the sum of annotated factors, unannotated factors and residual noise. For annotated factors, prior annotations can be derived from public pathway databases

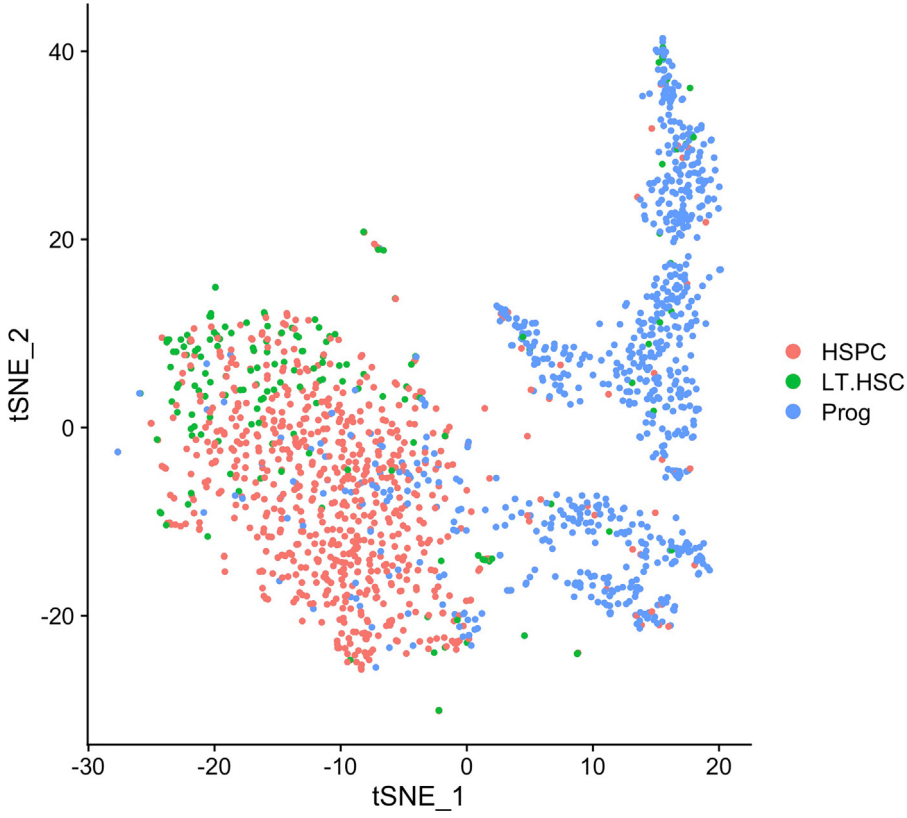

Fig. 7. t-SNE visualization of the single-cell variation captured by first 30 principal components.

such as MSigDB [4] or REACTOME [5]. The number of unannotated factors is fixed and defined by user. Gene assignments, factor weights and factor states are inferred by using computationally efficient variational Bayesian inference. The relevance of a factor, representing how much the variability in the given dataset the factor explains, is inferred by calculating the expected variance in expression levels across cells using genes assigned to the factor.

In this study, the normalized expression matrix of 42,512 genes and 1,656 cells was used as input of f-scLVM. Hallmark gene sets derived from MSigDB version 7.0 was used as the reference pathway database. The false positive rate (FPR) and false negative rate (FNR) for assigning genes to factors are 0.01 and 0.001, respectively; factors with mean absolute deviation lower than 0.2 were filtered out. For fitting, five hidden factors were expected to be included, and gene sets with at least 10 genes presented in the single-cell transcriptomic dataset were retained.

2.3. Principal Component Analysis (PCA)

PCA was performed for the normalized expression matrix of 2,000 highly variable genes and 1,656 cells. Top 30 PCs were selected for the downstream analysis. We see that the major advantage of the factor analysis is that it not only utilizes existing biological knowledge but also generates new biological knowledge through the sparse factor analysis — and subsequently combines the two into a set of unified biological factors with clear biological interpretations and

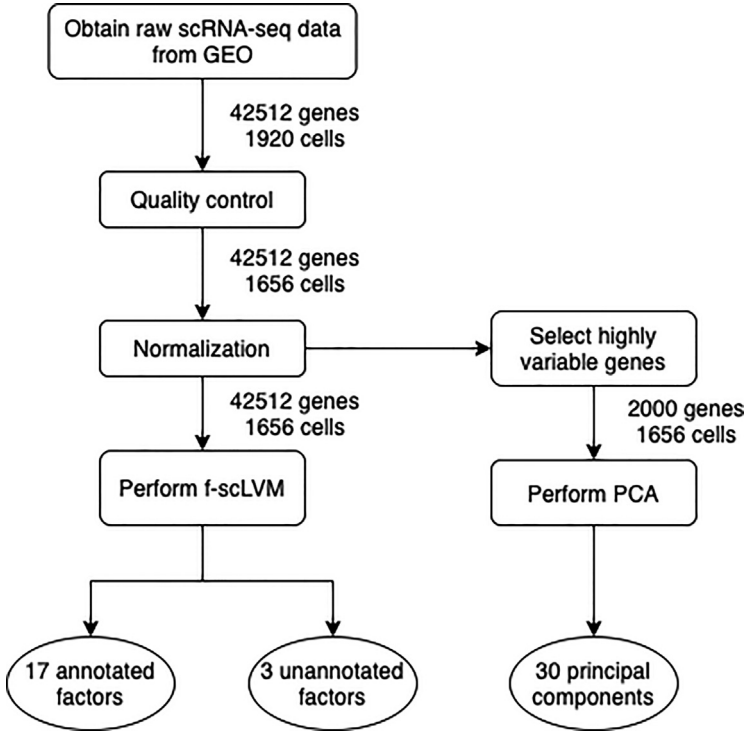

Fig. 8. Workflow.

meanings. In our original research paper [1], we were able to validate those top latent factors experimentally. In contrast, the PCA based method is entirely math driven and devoid of any existing biological knowledge – and thus it would be ineffective in separating biological factors from other random noises and provide meaningful biological interpretations. For comparison reasons, we have also provided the top principal components, which are also linear combinations of the underlying genes in the same way as the estimated latent factors.

2.4. Visualization

The t-distributed stochastic neighbor embedding (t-SNE) was applied to reduce the high dimensional data into 2-dimensional, so that we visualized cells on a 2-D scatter plot basing on the dimensionality reduction.

2.5. Software

All analyses were performed in the R programming environment (https://www.r-project.org). Preprocessing and PCA were performed by using the R package Seurat [6]. The f-scLVM was implemented by the R package slalom [3]. The t-SNE plots were generated by using the package Seurat [6].

Ethics Statement

Human subjects research: Not applicable

Animal experiments: All experiments comply with the Animal Research: Reporting of *In Vivo* Experiments (ARRIVE) guidelines and were carried out in accordance with the National Institutes of Health guide for the care and use of Laboratory animals (NIH Publications No. 8023, revised 1978).

Social media platforms: Not applicable.

CRediT Author Statement

Zhaoyan Liu: Conceptualization, Methodology, Software, Data Analysis and Writing; **Wei Zhu**: Supervision, Conceptualization, Methodology, Software and Data analysis; **Natasha M. Nesbitt**: Writing and Conceptualization; **Dmitri V. Gnatenko:** Conceptualization and Data Analysis; **Wadie F. Bahou**: Conceptualization, Writing-Reviewing and Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

This work was supported by NIH grants HL091939 and HL12945, an American Society of Hematology Bridge grant, and a Burroughs Wellcome Fund Collaborative Research Travel Grant (#1019947).

References

[1] Natasha M. Nesbitt, Lisa E. Malone, Zhaoyan Liu, Alexander Jares, Dmitri V. Gnatenko, Yupo Ma, Wei Zhu, Wadie F. Bahou, Divergent erythroid megakaryocyte fates in Blvrb-deficient mice establish non-overlapping cytoprotective functions during stress hematopoiesis, Free Radic. Biol. Med. 164 (2021) 164–174 ISSN 0891-5849, doi:10.1016/j.freeradbiomed.2020.12.015.

[2] S Nestorowa, FK Hamey, B Pijuan Sala, et al., A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation, Blood 128 (8) (2016) e20–e31, doi:10.1182/blood-2016-05-716480.

[3] F Buettner, N Pratanwanich, DJ McCarthy, JC Marioni, O Stegle, f-scLVM: Scalable and versatile factor analysis for single-cell RNA-seq, Genome Biol 18 (1) (2017) 1–13, doi:10.1186/s13059-017-1334-8.

[4] A Liberzon, A Subramanian, R Pinchback, H Thorvaldsdottir, P Tamayo, JP Mesirov, Molecular signatures database (MSigDB) 3.0, Bioinformatics 27 (12) (2011) 1739–1740.

[5] D Croft, AF Mundo, R Haw, M Milacic, J Weiser, G Wu, M Caudy, P Garapati, M Gillespie, MR Kamdar, et al., The Reactome pathway knowledgebase, Nucl. Acids Res 42 (2014) D472–D477.

[6] T Stuart, A Butler, P Hoffman, C Hafemeister, E Papalexi, WM Mauck 3rd, et al., Comprehensive Integration of Single-Cell Data, Cell 177 (7) (2019) 1888-902 e21.