

EryDB: A Transcriptomic Profile Database for Erythropoiesis and Erythroid-related Diseases

Guangmin Zheng (郑光敏) ^{1,2,3,#}, Song Wu (吴松) ^{1,2,3,4,#}, Zhaojun Zhang ^{1,2,3,#}, Zijuan Xin (辛子娟) ^{1,2,3}, Lijuan Zhang (张立娟) ^{1,2,3}, Siqi Zhao ^{1,2,3}, Jing Wu (吴静) ^{1,2,3}, Yanxia Liu (刘彦霞) ^{1,2}, Meng Li (李蒙) ^{1,2}, Xiuyan Ruan ^{1,2}, Nan Qiao ⁵, Yiming Bao (鲍一明) ^{1,2,3,4,*}, Hongzhu Qu (渠鸿竹) ^{1,2,3,*}, Xiangdong Fang (方向东) ^{1,2,3,*}

¹China National Center for Bioinformation, Beijing 100101, China

²Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

³University of Chinese Academy of Sciences, Beijing 100049, China

⁴National Genomics Data Center, China National Center for Bioinformation, Beijing 100101, China

⁵Huawei Cloud Computing Technologies Co., Ltd., Guiyang 550029, China

*Corresponding authors: baoym@big.ac.cn (Bao Y), quhongzhu@big.ac.cn (Qu H), fangxd@big.ac.cn (Fang X).

#Equal contribution.

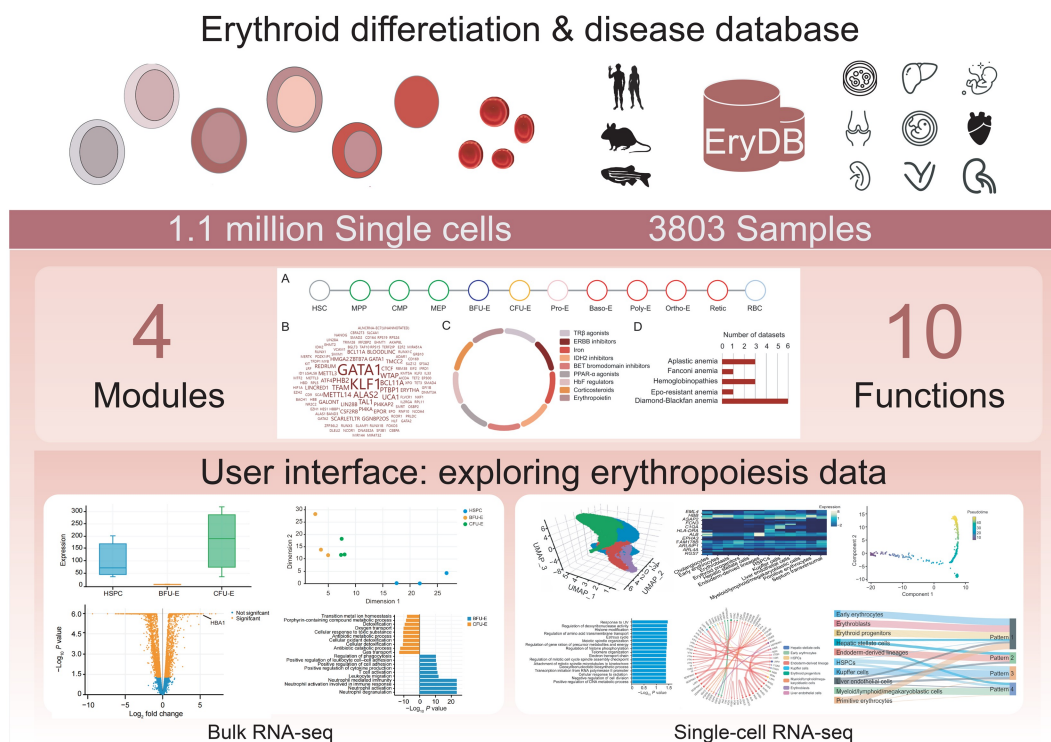
Handling Editor: An-Yuan Guo

Abstract

Erythropoiesis is a finely regulated and complex process that involves multiple transformations from hematopoietic stem cells to mature red blood cells at hematopoietic sites from the embryonic to adult stages. Investigations into its molecular mechanisms have generated a wealth of expression data, including bulk and single-cell RNA sequencing data. A comprehensively integrated and well-curated erythropoiesis-specific database will greatly facilitate the mining of gene expression data and enable large-scale research of erythropoiesis and erythroid-related diseases. Here, we present EryDB, an open-access and comprehensive database dedicated to the collection, integration, analysis, and visualization of transcriptomic data for erythropoiesis and erythroid-related diseases. Currently, the database includes expertly curated quality-assured data of 3803 samples and 1,187,119 single cells derived from 107 public studies of three species (*Homo sapiens*, *Mus musculus*, and *Danio rerio*), nine tissue types, and five diseases. EryDB provides users with the ability to not only browse the molecular features of erythropoiesis between tissues and species, but also perform computational analyses of single-cell and bulk RNA sequencing data, thus serving as a convenient platform for customized queries and analyses. EryDB v1.0 is freely accessible at <https://ngdc.cnbc.ac.cn/EryDB/home>.

Key words: Erythropoiesis; Database; Transcriptome; scRNA-seq; Data integration.

Graphical abstract



Received: 6 March 2023; Revised: 4 January 2024; Accepted: 29 January 2024.

© The Author(s) 2024. Published by Oxford University Press and Science Press on behalf of the Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Erythrocytes or red blood cells (RBCs), the most abundant type of blood cells, are responsible for transporting oxygen from lungs to other tissues [1]. Erythrocyte impairment is attributed to acquired/hereditary erythropoietic diseases in which abnormalities occur in causative genes, including genes that encode erythrocyte membrane proteins, globins, and enzymes, as well as genes that are required for erythropoiesis [2]. Erythropoiesis is a finely regulated and complicated process that involves multiple transformations from hematopoietic stem cells to mature RBCs at hematopoietic sites from the embryonic to adult stages [3]. Moreover, the current understanding of erythropoiesis and the pathophysiology of erythropoietic defects is limited. The underlying mechanisms need to be better defined to improve *in vitro* regeneration of RBCs [4,5] for the development of clinical treatments [3,6].

Large amounts of multi-omics data are now available [7–13], providing unprecedented insights into erythroid biology and generating many novel hypotheses. However, most of the omics data are available as raw data deposited in public repositories, such as the Genome Sequence Archive (GSA) [14,15] and Sequence Read Archive (SRA) [16], making it challenging for clinicians and researchers to reanalyze the data and gain biological insights [17]. Although some research groups have made their erythroid profile data web-accessible [9,11,18], no erythropoiesis-specific database that can facilitate the reuse of omics data is available to the erythroid biology community. An authoritative and integrated omics database that contains the molecules, pathways, and events involved in erythropoiesis will help to mitigate these shortcomings and accelerate research in this area.

Here, we present EryDB (<https://ngdc.cncb.ac.cn/EryDB/home>), a database that contains curated, quality-assured, and

pre-analyzed transcriptomic data of erythropoiesis and erythroid-related diseases. The EryDB platform allows users to search, browse, and analyze gene expression profiles or related data among various tissues or experimental types. Furthermore, gene expression changes under various pathological conditions can be analyzed toward developing diagnostic and prognostic tools for various erythropoietic diseases.

In this study, we describe the breadth of the transcriptomic data available in EryDB, which includes bulk and single-cell RNA sequencing (RNA-seq) data. The datasets are classified into four major erythropoiesis-related categories, namely Erythroid Differentiation, Genes, Compounds, and Diseases that users can mine to obtain erythroid-specific information. Customized filtering of the data can be performed under any of the four categories. These tools allow users to easily retrieve erythroid-specific data of interest to them. Moreover, EryDB provides a functional module named the Erythroid Atlas that can be used to integrate and compare erythropoiesis-related datasets from different tissues, namely bone marrow, peripheral blood, cord blood, and embryos. The rapid data query and analysis capabilities provided by EryDB enable efficient data mining for basic and clinical research.

Data collection and processing

Data collection and classification

The process that was used to construct EryDB is outlined in Figure 1. Transcriptomic data of erythroid cells were obtained from the Gene Expression Omnibus (GEO) [19], SRA [16], and GSA [14,15] databases using the keywords “erythroid” or “erythrocyte”. We also conducted a PubMed search to identify recent single-cell transcriptome studies using the keywords “peripheral blood mononuclear cell” OR “hematopoiesis” AND “erythroid” as well as “single cell”.

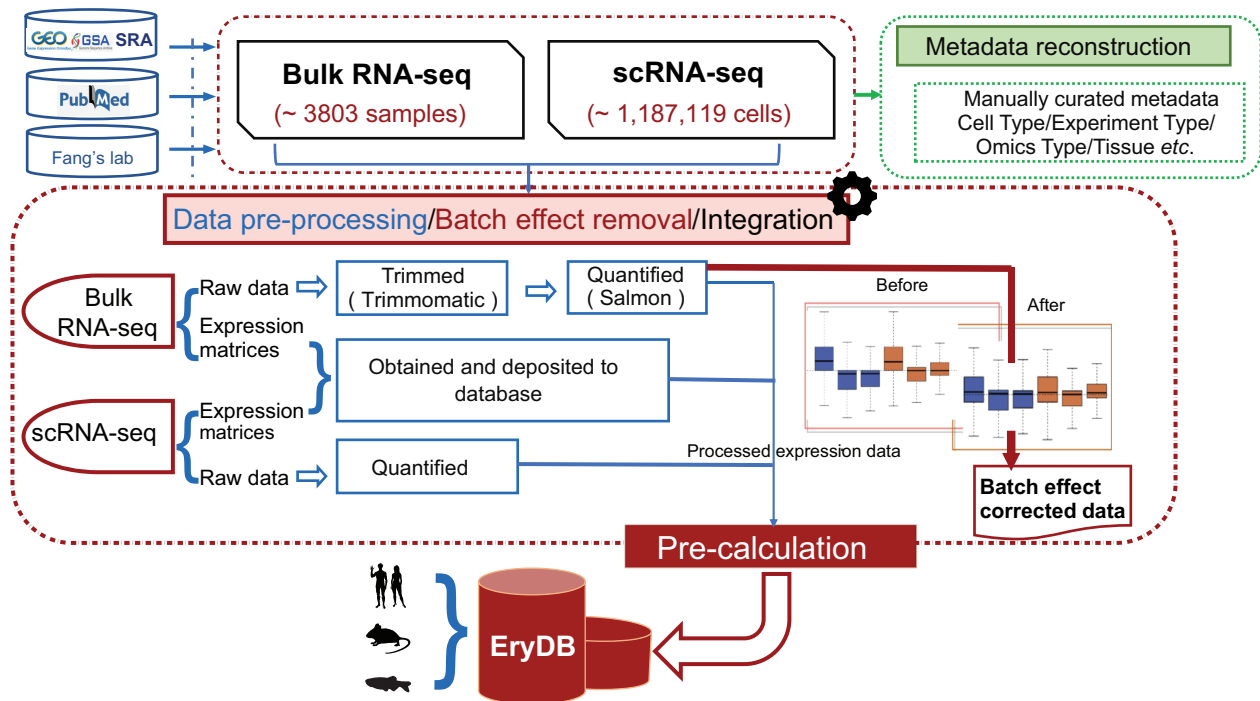


Figure 1 Construction process of EryDB

Transcriptomic data for erythroid cells were obtained from GEO, SRA, and GSA. Metadata were downloaded from GEO, SRA, and GSA, and manually curated by experts in erythropoiesis. GEO, Gene Expression Omnibus; SRA, Sequence Read Archive; GSA, Genome Sequence Archive; RNA-seq, RNA sequencing; scRNA-seq, single-cell RNA sequencing.

Furthermore, we screened the titles, abstracts, and sample information to identify studies that specifically performed gene expression profiling of erythroid cells. For the single-cell RNA-seq (scRNA-seq) data, only erythroid-related studies with > 300 cells were included. We also included published [20–23] and unpublished data from our laboratory. Finally, we assigned a unique custom identity document (ID) to each dataset in the EryDB to facilitate user-friendly searches. Additionally, we retained the data source as part of the dataset metadata.

Metadata curation

Metadata were downloaded from GEO, SRA, and GSA, and manually curated by experts in erythropoiesis (Figure 1). Repetitive and redundant information was removed. Each dataset was manually annotated with biologically relevant information, including model organism, sample tissue, experiment type, and omics type. A description of each study, GEO or GSA identifiers, and references to the original publications were also included in the metadata. All the datasets are labeled with species, tissue, and experiment type (*in vivo/in vitro*).

Pre-processing and quality control of bulk transcriptomic data

For the RNA-seq datasets, we obtained raw sequencing data whenever possible; in cases where raw sequencing data were unavailable, we included pre-processed data or raw count files. We employed an identical processing pipeline for datasets with access to raw sequencing data. The adaptor and low-quality sequences were removed using Trimmomatic (v.0.36) [24]. Gene expression was quantified using Salmon (v.1.4.0) [25] (Figure 1). Trimmed mean of M-values and voom normalization were performed using the edgeR (v.3.42.4) [26] and limma (v.3.48.3) [27] packages in R (v.4.3.0), respectively. For the integrated datasets within Erythroid Atlas module in EryDB, all samples have been processed using the same analysis pipeline followed by batch effect removal before proceeding to downstream analysis. Gene annotations for *Homo sapiens* (GRCh38), *Mus musculus* (GRCm38), and *Danio rerio* (Zv9) were sourced from Ensemble (release 91).

Pre-processing and quality control of single-cell transcriptomic data

For the scRNA-seq datasets, we downloaded the processed expression matrices when raw sequencing data were unavailable. The raw sequencing data were pre-processed using the methods described in the original publication, including quality control and normalization. Low-quality cells (expressing < 300 genes) and mitochondrial genes (> 20% of the reads mapped to mitochondrial genes) were removed. Library size normalization was performed using the Seurat (v.4.3.0) [28] R package. The batch effects of combined expression data were removed using Seurat canonical correlation analysis (CCA) function or Harmony (v.0.1.1) [29]. For the integrated single-cell datasets within Erythroid Atlas module in EryDB, all raw sequencing data have been pre-processed by an identical processing pipeline, and the comparative analyses were performed after integration and batch correction.

Implementation

EryDB is an interactive web application built using the Python/Flask (v.1.0.2) web framework. We used Vue, Plotly,

ECharts, and Highcharts with Python for data visualization, and Table module in iView to construct searchable tables.

Database content and usage

EryDB is a free public biological resource for erythropoiesis-related research. Users can browse datasets, examine the expression of erythropoiesis-related genes, and compare transcriptome differences and cell variations in the context of erythropoiesis and erythroid-related diseases.

Data records

In the current release, EryDB contains high-quality analytical results of 118 datasets and 3803 samples from 107 studies of 3 species, *H. sapiens*, *M. musculus*, and *D. rerio*. Fifteen of these studies were scRNA-seq studies involving 1,187,119 single cells; the others were bulk RNA-seq studies. Among the datasets, 60 (~ 50%), 49 (~ 42%), and 9 (~ 8%) were for *H. sapiens*, *M. musculus*, and *D. rerio*, respectively. The 118 datasets were associated with 11 tissue types: bone marrow, cord blood, peripheral blood, spleen, embryo, fetal tissue, heart, kidney, induced pluripotent stem cells (iPSCs), cell line, and other (unknown). Data statistics for EryDB are shown in Table 1. The accession numbers for all the erythroid-related datasets are listed in Table S1.

User-friendly search modules for data queries

EryDB provides three methods for dataset searches. First, a quick-search box on the home page allows real-time queries by specifying genes, tissues, dataset ID, organism name, or PubMed Unique Identifier (PMID). Second, category-specific searches by Erythroid Differentiation, Genes, Compounds, or Diseases are also available on the home page: Erythroid Differentiation — the process of erythroid differentiation is subdivided into 12 cell types (Figure 2A), and datasets that contain a cell type of interest are listed by clicking on the cell type; Genes — key genes related to erythroid differentiation and development are summarized as a keyword cloud and a sorted table, and datasets associated with a gene of interest can be accessed by clicking on the gene name (Figure 2B); Compounds — molecular compounds related to erythroid differentiation and development are listed, and datasets associated with a compound of interest can be accessed by clicking on the compound name (Figure 2C); and Diseases — the five main diseases related to erythropoiesis are available in 14 datasets (Figure 2D), and datasets associated with a disease of interest can be accessed by clicking on the disease name. The query results from these four categories are presented as tables that contain the dataset ID, species, tissue, and experiment type (Figure 2E). The table heading in each column is a drop-down box that enables further filtering of the datasets according to specific conditions. Third, target datasets can be selected by clicking on “Search” at the top of the home page and choosing one or more of the options on the navigation bar on the left-hand side of the page, namely cell type, reported gene, compound type, disease type, species, tissue, experiment type, and omics type (Figure 2F). Detailed information about the content of the datasets that can be searched under these conditions is listed in Table S2. The results of such queries are listed in tables that are similar to those obtained by category-specific searches.

Clicking on any word in the results tables from all three types of searches expands the overview and the gene

Table 1 Statistics of the contents of EryDB

Species	Tissue	Omics type	No. of studies	No. of datasets	No. of samples	No. of single cells
<i>Homo sapiens</i>	Fetal tissue (e.g., liver, spleen, and kidney)	scRNA-seq	3	3	98	434,156
		Bulk RNA-seq	5	5	80	/
	Bone marrow	scRNA-seq	6	7	130	267,566
		Bulk RNA-seq	3	6	88	/
	Peripheral blood	scRNA-seq	1	1	2	26,740
		Bulk RNA-seq	10	7	474	/
	Induced pluripotent stem cells	scRNA-seq	2	2	731	27,666
		Bulk RNA-seq	1	1	5	/
	Cord blood	scRNA-seq	2	2	7	29,920
		Bulk RNA-seq	13	13	142	/
Cell line	Bulk RNA-seq	6	6	82	/	
	Other (unknown)	Bulk RNA-seq	1	1	6	/
<i>Mus musculus</i>	Fetal tissue (e.g., liver, yolk sac, and caudal half)	scRNA-seq	1	1	6	353,528
		Bulk RNA-seq	18	22	542	/
	Bone marrow	scRNA-seq	2	1	9	28,224
		Bulk RNA-seq	10	15	1162	/
	Cell line	Bulk RNA-seq	4	6	46	/
		Bulk RNA-seq	4	6	43	/
	Spleen	Bulk RNA-seq	4	6	43	/
		Bulk RNA-seq	3	3	24	/
	Embryo	Bulk RNA-seq	2	1	6	/
		Bulk RNA-seq	2	1	6	/
<i>Danio rerio</i>	Fetal tissue	scRNA-seq	1	1	1	19,319
		Bulk RNA-seq	7	6	93	/
	Kidney	Bulk RNA-seq	2	2	24	/
Sum			107	118	3803	1,187,119

Note: RNA-seq, RNA sequencing; scRNA-seq, single-cell RNA sequencing.

expression profiles of the dataset. The overview includes a research summary, research organization, sequencing platform information, specific sample information, data processing method, overall design introduction, data contributors, and contact information for the dataset (Figure 2G). The sample accession numbers are linked to the database source of the original data. Because of the different technologies used for bulk RNA-seq and scRNA-seq, the gene expression profile datasets from the two technologies are presented separately.

Comprehensive multidimensional online data exploration

Bulk RNA-seq data exploration

EryDB provides four analysis modules for the bulk transcriptome datasets on the overview page: Expression Profile — users can perform the analysis of expression changes for a gene, sample group, and dataset specified by users, and the detailed results are shown when the sample bar is selected (Figure 3A); Principal Component Analysis — a two-dimensional (2D) plot of the principal component analysis results is obtained, where individual sample's group is color-coded (Figure 3B); Differential Analysis — a volcano plot of differentially expressed genes (DEGs) between two user-specified groups and a browsable table below the plot are obtained (Figure 3C); and Enrichment Analysis — enrichment analysis results of DEGs between user-specified groups, including Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) functional and pathway annotations (Figure 3D).

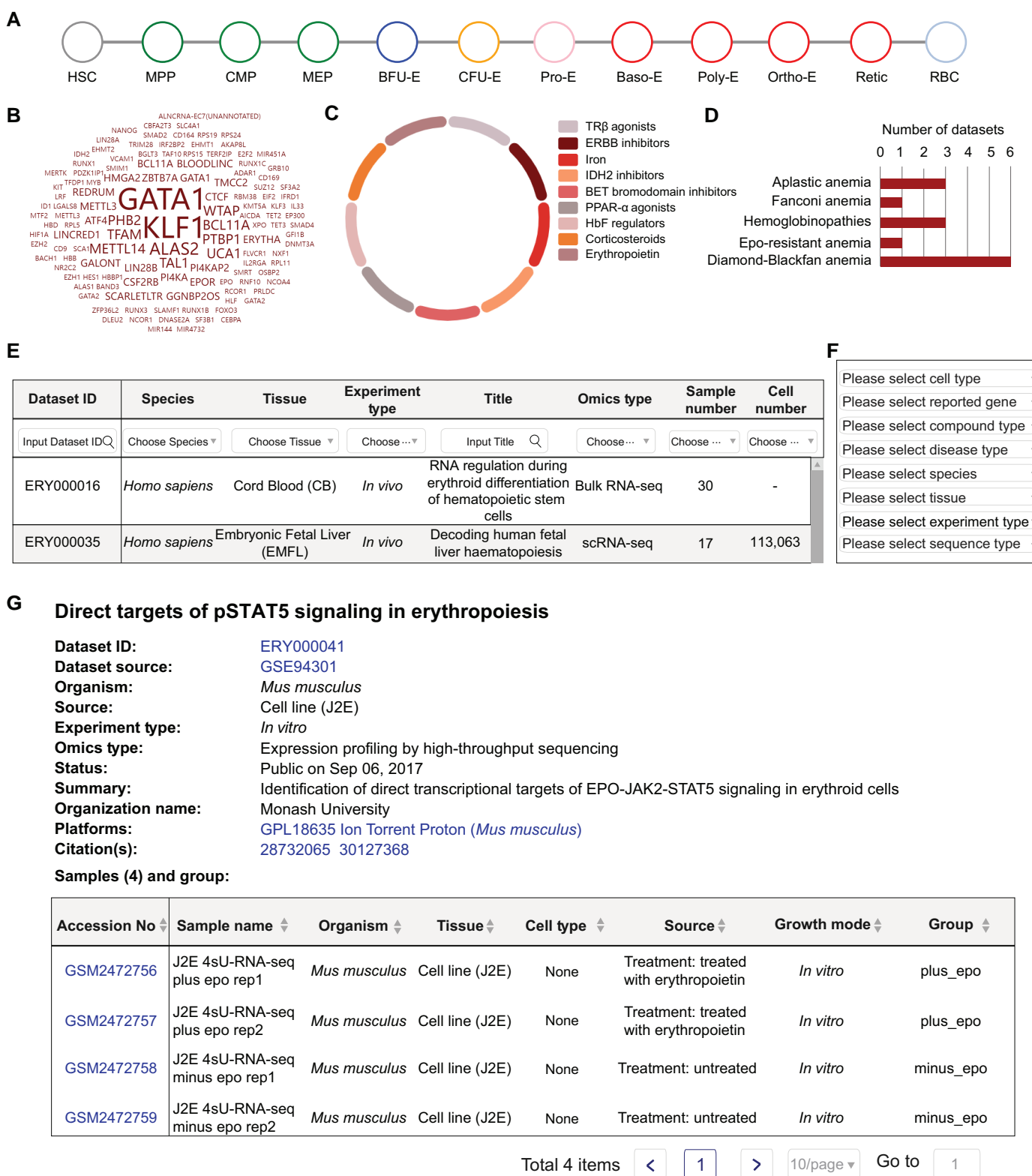
To demonstrate the database functions, we applied EryDB to the GEO dataset GSE128268, which is a bulk RNA-seq dataset of an erythroid cell culture of human umbilical cord-derived CD34⁺ hematopoietic stem and progenitor cells [30]. We found that the hemoglobin subunit gene *HBA1* was highly expressed at the colony-forming unit-erythroid (CFU-E) stage

(Figure 3A). Furthermore, the burst-forming unit-erythroid (BFU-E) and CFU-E cells had transcriptomic profiles that were clearly distinct, and both profiles were distinct from those of umbilical cord-derived CD34⁺ hematopoietic stem and progenitor cells (Figure 3B). Analysis of the DEGs showed that hemoglobin markers (e.g., *HBA1*, *HBA2*, and *HBB*) were among the top 25 most upregulated genes at the CFU-E stage (Figure 3C). To determine which pathways were most upregulated in CFU-E cells, we compared the gene expression signatures between CFU-E and BFU-E cells and found that gas and oxygen transport were upregulated in CFU-E cells (Figure 3D). Together, these results suggest that CFU-E cells undergo a switch that results in a hemoglobin-based oxygen carrier phenotype.

The Erythroid Atlas page in EryDB allows comparisons to be made between transcriptome datasets associated with erythroid development. Users can explore the gene expression profiles of different differentiation systems in humans, mice, and zebrafish (Figure 3E). For example, changes in the expression of the transcription factor gene *RUNX2* at different erythroid differentiation stages in humans (*in vivo*) or changes induced *in vitro* can be browsed (Figure 3F). The results were obtained by the integration of 11 datasets that contain a total of 207 samples. All the data were analyzed and normalized in a standard pipeline to eliminate the batch effect of the data sources. Moreover, DEGs can be identified from volcano plots and tables between any two stages of the erythroid differentiation process by selecting corresponding tags in the drop-down box, which enables users to identify genes of interest for further study.

scRNA-seq data exploration

EryDB provides six analysis modules for scRNA-seq data. Visualization & Feature — a module enables dimension reduction analysis and signature expression queries among

**Figure 2 Dataset query strategies in EryDB**

A. The Erythroid Differentiation category allows datasets to be filtered by cell type. **B.** The Genes category allows datasets to be filtered by selecting a gene of interest. **C.** The Compounds category allows datasets under the action of a certain molecular compound to be obtained. **D.** The Diseases category allows datasets to be obtained by selecting a specific disease. **E.** Query results are presented as a table. **F.** Navigation bar on the left-hand side of the "Search" page lists the conditions for selecting target datasets. **G.** Detailed meta information included in the dataset overview page. HSC, hematopoietic stem cell; MPP, multipotent progenitor cell; CMP, common myeloid progenitor; MEP, megakaryocyte-erythroid progenitor; BFU-E, burst-forming unit-erythroid; CFU-E, colony-forming unit-erythroid; Pro-E, proerythroblast; Baso-E, basophilic erythroblast; Poly-E, polychromatophilic erythroblast; Ortho-E, orthochromatic erythroblast; Retic, reticulocyte; RBC, red blood cell; BET, bromodomain and extra-terminal domain; TR β , thyroid hormone receptor β ; ERBB, epidermal growth factor receptor; IDH2, isocitrate dehydrogenase 2; PPAR- α , peroxisome proliferator-activated receptor α ; HbF, fetal hemoglobin.

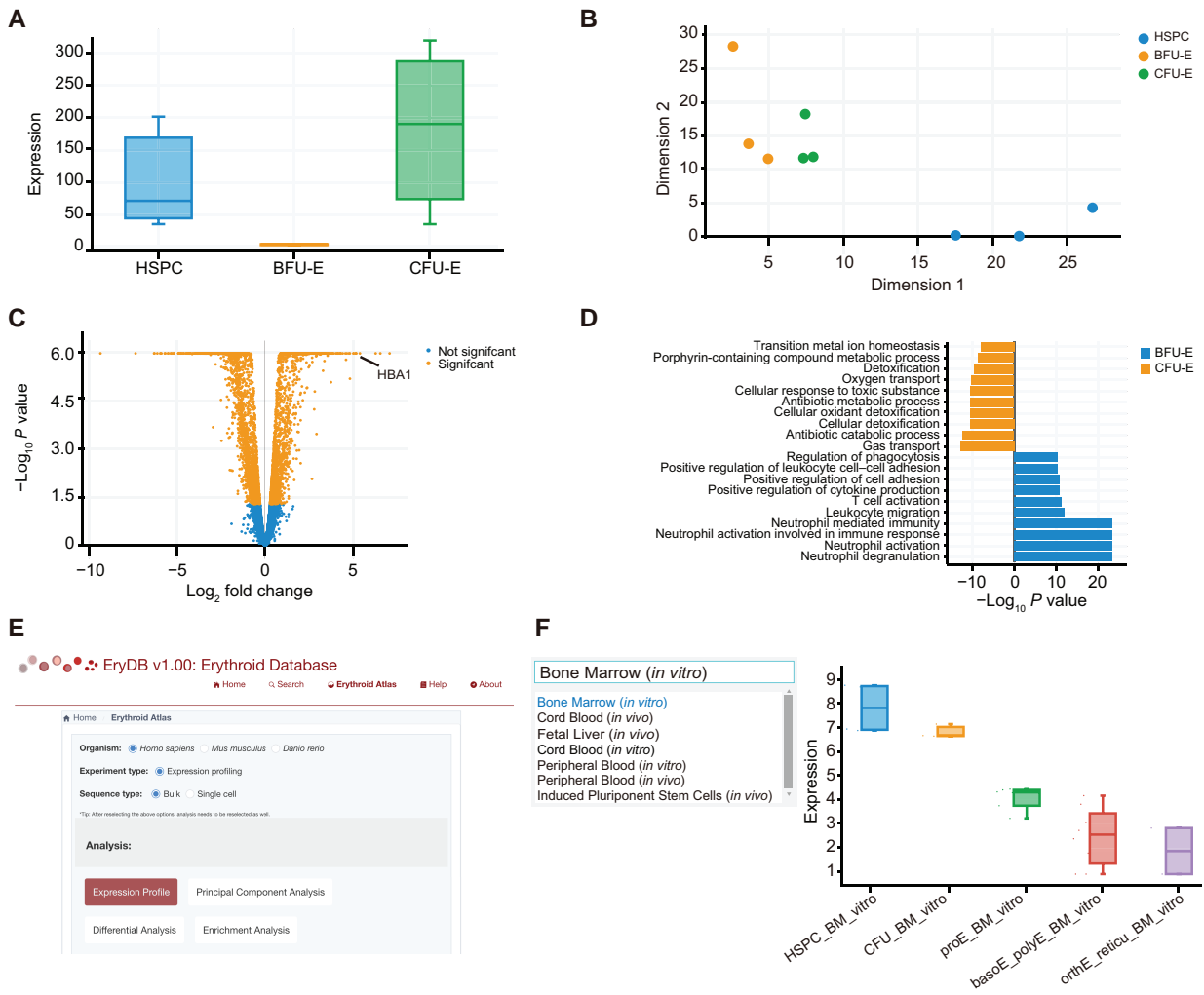


Figure 3 Statistical analysis function for bulk RNA-seq data

A. Expression profile of the hemoglobin subunit gene *HBA1* in the GSE128268 dataset as an example to demonstrate the functions of EryDB. **B.** Principal component analysis of samples in the GSE128268 dataset. **C.** Volcano plot of DEGs between samples from the groups of CFU-E and BFU-E in GSE128268. **D.** Enrichment analysis (biological process terms of GO) of DEGs between samples from the CFU-E and BFU-E groups in GSE128268. **E.** Screenshot of the Erythroid Atlas page. Datasets for different differentiation stages can be explored on this page. **F.** Expression of the transcription factor gene *RUNX2* among different developmental stages across multiple datasets of bone marrow tissue as an example. The tissue of interest can be selected from the drop-down box on the left-hand side. DEG, differentially expressed gene; GO, Gene Ontology; HSPC, hematopoietic stem and progenitor cell.

cells. Gene expression among single cells is displayed as interactive *t*-distributed stochastic neighbor embedding (*t*-SNE) or uniform maximal approximation projection (UMAP) plots (Figure 4A and B), and the expression of specific genes among cells can be queried in an extra *t*-SNE plot (Figure 4C). Marker & Enrichment — a module provides marker gene expression of cell types as a heatmap (Figure 4D) and enables functional enrichment analysis of genes in cell populations. The KEGG pathway or GO enrichment results for each stage are presented as bar graphs (Figure 4E). Difference & Enrichment — the expression of DEGs in the same cell type from different groups is presented in a volcano plot and the functional enrichment results of these DEGs are presented in a bar plot. Differentiation Trajectory — pseudotime analysis of cells can be shown, where pseudotime trajectories of cell differentiation are visualized in 2D space (Figure 4F), and cell differentiation trajectories can be displayed by cell type (Figure 4G). Cell–Cell Interaction — the interaction of ligands and receptors

between different cell types can be viewed as a circos plot, where the thickness of the lines indicates the interaction strength. The communication types include growth factors, cytokines, checkpoints, and others, which can be queried by selection in the box provided (Figure 4H). Cell–Cell Communication — cellular communication is visualized as a river diagram in which each pattern represents different levels of communication (Figure 4I). User can view the interaction of a signaling pathway of interest (Figure 4J) and the contribution of specific ligand–receptor pairs in this pathway (Figure 4K). Together, these analyses cover most of the statistical bases of single-cell transcriptomes.

To demonstrate the database functions, we applied EryDB to the GSA dataset CRA002443, which is a single-cell transcriptome dataset of human 5–19 week post-coitum fetal tissue (W5–W19) [7], and explored the erythroid lineage properties. A total of 12 cell clusters were identified in the W9 sample, and the number of erythroid cell lineages in this sample was higher than the numbers in the other samples

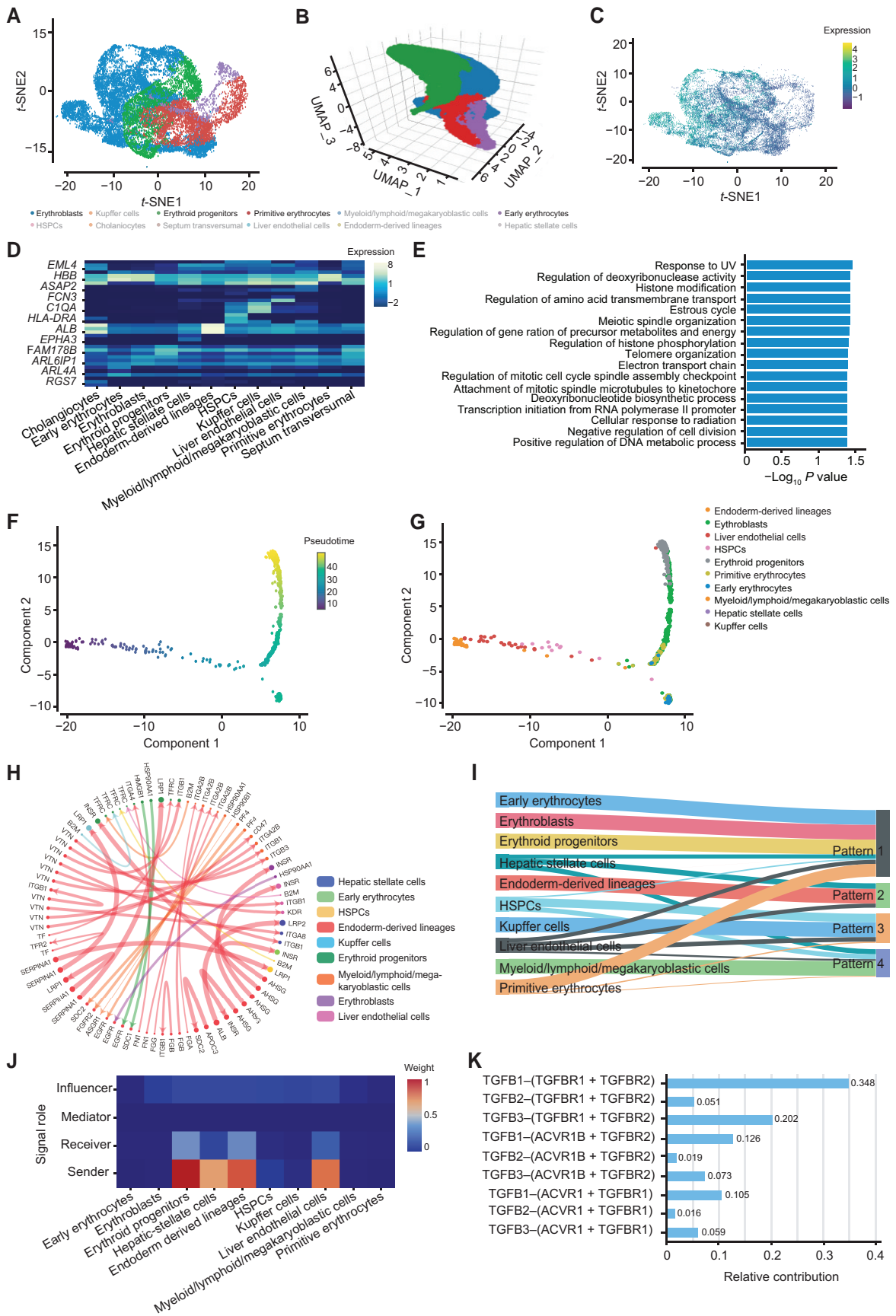


Figure 4 Statistical analysis function for scRNA-seq data

A. Interactive 2D t-SNE plot of the single cells in the W9 group in the CRA002443 dataset as an example to demonstrate the functions of EryDB. Cell clusters are color-coded. **B.** Interactive 3D UMAP plot of the single cells in the W9 group. Cell clusters are color-coded. **C.** Expression of the cyclin gene *CCNB1* among erythroblasts as an example. Panels A, B, and C illustrate the usage of the Visualization & Feature module. **D.** Expression of marker genes at different cell stages. **E.**

(Figure 4A and B). Expression of the cyclin gene *CCNB1* was upregulated in erythroblasts (Figure 4C). Markers for each cell population are plotted as a heatmap (Figure 4D) that can be enlarged for convenient browsing of gene expression. The hemoglobin gene *HBB* was highly expressed in erythroid cell lineages in the W9 sample, and genes expressed in erythroblasts were enriched in several biological processes, including response to UV and regulation of deoxyribonuclease activity (Figure 4E). In the W9 group, endoderm-derived cell lineages had the strongest cell–cell interactions among the top 40 pairs of ligands and receptors, but they showed much less interactions in early erythroblasts (Figure 4H). Four communication patterns were identified among different cell types (Figure 4I). The transforming growth factor TGF β signaling pathway in two of the patterns (patterns 1 and 2) was communicated between most of the cells by multiple ligands and receptors (Figure 4J and K). Similarly, the Erythroid Atlas page provides integrative analysis via six modules across differentiated systems from the scRNA-seq datasets.

In summary, these examples demonstrate the power of EryDB to efficiently test novel hypotheses using publicly available data of genes that regulate erythroid differentiation and development. All visualizations can be downloaded as scalable vector graphics (SVG) or portable network graphics (PNG) files, and related result data tables can be downloaded in comma-separated value (CSV) format or Excel sheets. For advanced users, the integrated expression matrix and metadata can be downloaded for personalized analysis. Besides, the analysis results can be filtered with adjustable parameters. The fully functional design of EryDB is intended to facilitate the widespread reuse of publicly available datasets to address novel and unresolved questions and allow comparative analyses of similar studies to resolve study-specific biases.

Discussion

A survey of erythroid-related omics databases confirmed that the currently available databases are important resources that can be used to interpret gene regulation or mutations related to erythropoiesis and erythroid-related diseases. Some erythroid-related databases are no longer accessible (e.g., EpoDB [31], Hembase [32], and ErythronDB [33]). Accessible databases (e.g., dbRBC [34], RBCmembrane [35], and BloodSpot [36,37]) that contain specific omics data associated with RBCs (Table S3) tend to lack diversity in their data sources, species covered, or functional analysis tools, and do not contain single-cell sequencing data.

EryDB was designed to integrate most of the accessible public transcriptome datasets related to erythropoiesis and enable the investigation of gene expression dynamics in erythroid differentiation and erythropoiesis-related diseases. Compared with other RBC-related databases, EryDB has

more user-friendly data search options, more extensive data sources, and more comprehensive analysis capabilities. In the sub-module search of the datasets, users can select the search that meets their research interests by selecting the relevant module. For example, in the Disease module, researchers can easily examine the gene expression changes under certain pathological conditions for blood disease diagnosis [38–41], and in the Compounds module, users can obtain information about the small molecules that promote or inhibit erythroid differentiation. The *in vitro* RBC regeneration data from our laboratory are unique in EryDB and can be integrated with *in vivo*-derived erythroid cell data. This allows users to compare the differences and mechanisms underlying erythropoiesis from different global sources. EryDB also provides comprehensive functional analysis tools. For example, cell–cell interaction analysis based on scRNA-seq datasets generated a broad spectrum of information on the interactions between erythroid cells and other cell types. We found that *in vivo* erythroid cells had stronger cellular interactions with other cell types than *in vitro* erythroid cells had.

EryDB will be useful not only for computational biologists but also for bench clinicians interested in erythropoietic disorders and for researchers involved in basic erythropoiesis-related research. In future releases of EryDB, the transcriptomic data will be continuously updated, and other types of omics data will be added. We will also develop new analytical features for further exploration of the multi-omics data.

Data availability

EryDB v1.0 is freely accessible at <https://ngdc.cncb.ac.cn/EryDB/home>. It has also been submitted to Database Commons [42] at the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB), which is publicly accessible at <https://ngdc.cncb.ac.cn/databasecommons/database/id/10216>.

CRedit author statement

Guangmin Zheng: Software, Visualization, Writing – original draft, Writing – review & editing. **Song Wu:** Software, Visualization, Validation, Writing – review & editing. **Zhaojun Zhang:** Data curation, Resources, Investigation, Validation, Writing – review & editing. **Zijuan Xin:** Data curation, Resources, Investigation. **Lijuan Zhang:** Data curation, Resources, Investigation. **Siqi Zhao:** Validation. **Jing Wu:** Validation. **Yanxia Liu:** Validation. **Meng Li:** Validation. **Xiuyan Ruan:** Validation. **Nan Oiao:** Software, Resources, Validation. **Yiming Bao:** Supervision, Project administration, Funding acquisition, Resources, Conceptualization. **Hongzhu Qu:** Validation, Writing – review & editing. **Xiangdong Fang:** Supervision, Project administration, Funding acquisition, Resources, Conceptualization. All authors have read and approved the final manuscript.

Figure 4 Continued

GO and KEGG functional and pathway enrichment for genes in the erythroblasts cluster. The cluster and enriched type can be selected in the drop-down boxes in the Marker & Enrichment module. **F.** Pseudotime trajectories of cell differentiation visualized in 2D. **G.** Cell differentiation trajectories displayed by cell type. Panels F and G show the usage of the Differentiation Trajectory module. **H.** Interaction of ligands and receptors between different cell types. Interaction of the top 40 ligand–receptor pairs is shown. The communication type can be defined in the drop-down boxes in the Cell–Cell Interaction module. **I.** Cell–cell communication is visualized as a river diagram. **J.** Interaction of the TGF β signaling pathway between cells as an example. **K.** Contribution of specific ligand–receptor pairs in the TGF β signaling pathway. Panels I, J, and K illustrate the usage of the Cell–Cell Communication module. *t*-SNE, *t*-distributed stochastic neighbor embedding; UMAP, uniform maximal approximation projection; W9, 9 weeks; 2D, two-dimensional; 3D, three-dimensional; TGF, transforming growth factor; KEGG, Kyoto Encyclopedia of Genes and Genomes.

Competing interests

Nan Qiao is the current employee of Huawei Cloud Computing Technologies Co., Ltd. All the other authors have declared no competing interests.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (<https://doi.org/10.1093/gpbjnl/qzae029>).

Acknowledgments

We would like to thank Yunxiao Ren, Rudan Xiao, Xinyue Wang, and Yongsen Chen for their assistance in RNA data curation. This research was supported by the National Key R&D Program of China (Grant No. 2022YFC2406803), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA16010602), and the National Natural Science Foundation of China (Grant Nos. 81870097, 82070114, and 82270126). We would like to thank Liwen Bianji (Edanz) (www.liwenbianji.cn) and Editage (www.editage.cn) for the English language review.

ORCID

0000-0002-3781-170X (Guangmin Zheng)
 0000-0002-0923-639X (Song Wu)
 0000-0003-0490-6507 (Zhaojun Zhang)
 0000-0001-9418-4848 (Zijuan Xin)
 0009-0009-4246-929X (Lijuan Zhang)
 0000-0003-2341-4989 (Siqi Zhao)
 0000-0001-6032-0979 (Jing Wu)
 0000-0002-0608-2217 (Yanxia Liu)
 0000-0003-0369-5867 (Meng Li)
 0000-0002-2181-1529 (Xiuyan Ruan)
 0000-0002-2740-262X (Nan Qiao)
 0000-0002-9922-9723 (Yiming Bao)
 0000-0001-7013-8409 (Hongzhu Qu)
 0000-0002-6628-8620 (Xiangdong Fang)

References

- Palis J. Primitive and definitive erythropoiesis in mammals. *Front Physiol* 2014;5:3.
- Di Pierro E, Brancaloni V, Granata F. Advances in understanding the pathogenesis of congenital erythropoietic porphyria. *Br J Haematol* 2016;173:365–79.
- Caulier AL, Sankaran VG. Molecular and cellular mechanisms that regulate human erythropoiesis. *Blood* 2022;139:2450–9.
- Batta K, Menegatti S, Garcia-Alegria E, Florkowska M, Lacaud G, Kouskoff V. Concise review: recent advances in the *in vitro* derivation of blood cell populations. *Stem Cells Transl Med* 2016;5:1330–7.
- Christaki EE, Politou M, Antonelou M, Athanasopoulos A, Simantirakis E, Seghatchian J, et al. *Ex vivo* generation of transfusable red blood cells from various stem cell sources: a concise revisit of where we are now. *Transfus Apher Sci* 2019;58:108–12.
- Weiss MJ, dos Santos CO. Chaperoning erythropoiesis. *Blood* 2009;113:2136–44.
- Wang X, Yang L, Wang YC, Xu ZR, Feng Y, Zhang J, et al. Comparative analysis of cell lineage differentiation during hepatogenesis in humans and mice at the single-cell transcriptome level. *Cell Res* 2020;30:1109–26.
- Yang D, Jang I, Choi J, Kim MS, Lee AJ, Kim H, et al. 3DIV: a 3D-genome Interaction Viewer and database. *Nucleic Acids Res* 2018;46:D52–7.
- Tusi BK, Wolock SL, Weinreb C, Hwang Y, Hidalgo D, Zilionis R, et al. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* 2018;555:54–60.
- Weinreb C, Wolock S, Tusi BK, Socolovsky M, Klein AM. Fundamental limits on dynamic inference from single-cell snapshots. *Proc Natl Acad Sci U S A* 2018;115:E2467–76.
- Popescu DM, Botting RA, Stephenson E, Green K, Webb S, Jardine L, et al. Decoding human fetal liver haematopoiesis. *Nature* 2019;574:365–71.
- Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. *Science* 2020;370:eaba7721.
- Vanuytsel K, Matte T, Leung A, Naing ZH, Morrison T, Chui DHK, et al. Induced pluripotent stem cell-based mapping of beta-globin expression throughout human erythropoietic development. *Blood Adv* 2018;2:1998–2011.
- CNCB-NGDC Members and Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformatics in 2022. *Nucleic Acids Res* 2022;50:D27–38.
- Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. *Genomics Proteomics Bioinformatics* 2021;19:578–83.
- Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res* 2011;39:D19–21.
- Rung J, Brazma A. Reuse of public genome-wide gene expression data. *Nat Rev Genet* 2013;14:89–99.
- Reyes M, Filbin MR, Bhattacharyya RP, Billman K, Eisenhaure T, Hung DT, et al. An immune-cell signature of bacterial sepsis. *Nat Med* 2020;26:333–40.
- Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res* 2007;35:D760–5.
- Xin Z, Zhang W, Gong S, Zhu J, Li Y, Zhang Z, et al. Mapping human pluripotent stem cell-derived erythroid differentiation by single-cell transcriptome analysis. *Genomics Proteomics Bioinformatics* 2021;19:358–76.
- Ren Y, Zhu J, Han Y, Li P, Wu J, Qu H, et al. Regulatory association of long noncoding RNAs and chromatin accessibility facilitates erythroid differentiation. *Blood Adv* 2021;5:5396–409.
- Ding N, Xi J, Li Y, Xie X, Shi J, Zhang Z, et al. Global transcriptome analysis for identification of interactions between coding and noncoding RNAs during human erythroid differentiation. *Front Med* 2016;10:297–310.
- Yang Y, Wang H, Chang KH, Qu H, Zhang Z, Xiong Q, et al. Transcriptome dynamics during human erythroid differentiation and development. *Genomics* 2013;102:431–41.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–20.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;14:417–9.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–40.
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43:e47.
- Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. *Cell* 2019;177:1888–902.e21.
- Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019;16:1289–96.

- [30] Schulz VP, Yan H, Lezon-Geyda K, An X, Hale J, Hillyer CD, et al. A unique epigenomic landscape defines human erythropoiesis. *Cell Rep* 2019;28:2996–3009.e7.
- [31] Stoeckert CJ Jr, Salas F, Brunk B, Overton GC. EpoDB: a prototype database for the analysis of genes expressed during vertebrate erythropoiesis. *Nucleic Acids Res* 1999;27:200–3.
- [32] Goh SH, Lee YT, Bouffard GG, Miller JL. Hembase: browser and genome portal for hematology and erythroid biology. *Nucleic Acids Res* 2004;32:D572–4.
- [33] Kingsley PD, Greenfest-Allen E, Frame JM, Bushnell TP, Malik J, McGrath KE, et al. Ontogeny of erythroid gene expression. *Blood* 2013;121:e5–13.
- [34] Patnaik SK, Helmberg W, Blumenfeld OO. BGMUT: NCBI dbRBC database of allelic variations of genes encoding antigens of blood group systems. *Nucleic Acids Res* 2012;40:D1023–9.
- [35] An X, Mohandas N. Disorders of red cell membrane. *Br J Haematol* 2008;141:367–75.
- [36] Bagger FO, Sasivarevic D, Sohi SH, Laursen LG, Pundhir S, Sønderby CK, et al. BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. *Nucleic Acids Res* 2016;44:D917–24.
- [37] Bagger FO, Kinalis S, Rapin N. BloodSpot: a database of healthy and malignant haematopoiesis updated with purified and single cell mRNA sequencing profiles. *Nucleic Acids Res* 2019;47:D881–5.
- [38] Masuda T, Wang X, Maeda M, Canver MC, Sher F, Funnell AP, et al. Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. *Science* 2016;351:285–9.
- [39] Lee HY, Gao X, Barrasa MI, Li H, Elmes RR, Peters LL, et al. PPAR- α and glucocorticoid receptor synergize to promote erythroid progenitor self-renewal. *Nature* 2015;522:474–7.
- [40] Wan Y, Zhang Q, Zhang Z, Song B, Wang X, Zhang Y, et al. Transcriptome analysis reveals a ribosome constituents disorder involved in the RPL5 downregulated zebrafish model of Diamond-Blackfan anemia. *BMC Med Genomics* 2016;9:13.
- [41] Iskander D, Wang G, Heuston EF, Christodoulidou C, Psaila B, Ponnusamy K, et al. Single-cell profiling of human bone marrow progenitors reveals mechanisms of failing erythropoiesis in Diamond-Blackfan anemia. *Sci Transl Med* 2021;13:eabf0113.
- [42] Ma L, Zou D, Liu L, Shireen H, Abbasi AA, Bateman A, et al. Database Commons: a catalog of worldwide biological databases. *Genomics Proteomics Bioinformatics* 2023;21:1054–8.