frontiers in Genetics

ORIGINAL RESEARCH
published: 26 September 2019
doi: 10.3389/fgene.2019.00903

SCDevDB: A Database for Insights Into Single-Cell Gene Expression Profiles During Human Developmental Processes

*Zishuai Wang†, Xikang Feng† and Shuai Cheng Li**

Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

Single-cell RNA-seq studies profile thousands of cells in developmental processes. Current databases for human single-cell expression atlas only provide search and visualize functions for a selected gene in specific cell types or subpopulations. These databases are limited to technical properties or visualization of single-cell RNA-seq data without considering the biological relations of their collected cell groups. Here, we developed a database to investigate single-cell gene expression profiling during different developmental pathways (SCDevDB). In this database, we collected 10 human single-cell RNA-seq datasets, split these datasets into 176 developmental cell groups, and constructed 24 different developmental pathways. SCDevDB allows users to search the expression profiles of the interested genes across different developmental pathways. It also provides lists of differentially expressed genes during each developmental pathway, T-distributed stochastic neighbor embedding maps showing the relationships between developmental stages based on these differentially expressed genes, Gene Ontology, and Kyoto Encyclopedia of Genes and Genomes analysis results of these differentially expressed genes. This database is freely available at https://scdevdb.deepomics.org

Keywords: single cell, gene expression, development, database, cell type, differential expression

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis (Beijing) Co. Ltd, China

Reviewed by:
Suman Ghosal,
National Institutes of Health (NIH), United States
Lei Chen,
Shanghai Maritime University, China

****Correspondence:***
Shuai Cheng Li
shuaicli@cityu.edu.hk

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *12 May 2019*
Accepted: *26 August 2019*
Published: *26 September 2019*

Citation:
Wang Z, Feng X and Li SC (2019) SCDevDB: A Database for Insights Into Single-Cell Gene Expression Profiles During Human Developmental Processes. Front. Genet. 10:903. doi: 10.3389/fgene.2019.00903

INTRODUCTION

In developmental biology, gene expression changes during the developmental process is an important feature to understand developmental questions such as cell growth, cell differentiation, cell fate decisions, etc. (Ko, 2001; Merks and Glazier, 2005; Gittes, 2009). Recently, high-throughput RNA sequencing technique has been widely used to study gene expression in developmental processes (Spitz and Furlong, 2006). Bulk RNA sequencing typically uses hundreds to millions of cells and reveals only the average expression level for each gene across a large population of cell populations (Wang and Bodovitz, 2010; Sanchez and Golding, 2013). Single-cell RNA-seq measures the distribution of expression levels for each gene across a population of cells and provides a more accurate representation of cell-to-cell variations instead of the stochastic average (Saliba et al., 2014). Therefore, single-cell RNA-seq is particularly apposite for developmental biology (Liu et al., 2014; Griffiths et al., 2018).

High-resolution single-cell transcriptome analysis has been performed during many developmental processes including preimplantation development from oocyte to morula (Xue et al., 2013; Yan et al., 2013), early forebrain and mid/hindbrain cell differentiation from human embryonic stem

cells (hESCs) (Yao et al., 2017), and digestive tract development from human embryos between 6 and 25 weeks (Gao et al., 2018), etc. These studies not only revealed many biological features, including developmental processes, signaling pathways, cell cycle, and transcription factor networks but also provided resources to investigate the gene expression patterns during different developmental processes. Therefore, there is a strong need for a web resource that curates and provides single-cell gene expression profiles during different developmental processes.

So far, several web resources for human single-cell transcriptome data have been reported. scRNASeqDB contains 38 datasets covering 200 human cell lines or cell types and 13,440 samples (Cao et al., 2017). The single-cell expression atlas, launched by the European Bioinformatics Institute (https://www.ebi.ac.uk/gxa/sc/home), contains 52 single-cell RNA-Seq studies, consisting of 61,073 cells from 9 different species. The single-cell centric database "SCPortalen" covers 23 human single-cell transcriptomics datasets that are publicly available from the International Nucleotide Sequence Database Collaboration sites (Abugessaisa et al., 2017). PanglaoDB integrated 209 human single-cell datasets consisting of gene expression measurements from cells originating from a common biological source or experiment (Franzén et al., 2019). However, users of these databases can only query gene expression in specific cell types or population heterogeneity processed by the authors. Researchers who are interested in gene expression changes during a specific developmental process are not easily able to extract these dynamic features from these databases.

Here, we developed a database to investigate single-cell gene expression profiling during different developmental processes (SCDevDB). In this database, we collected 10 human single-cell RNA-seq datasets, split these datasets into 176 developmental cell groups, and constructed 24 different developmental pathways. Users of SCDevDB are easy to view the expression changes of their interested genes showed with a boxplot. In addition, users can also download differentially expressed (DE) genes during each developmental pathway, the T-distributed stochastic neighbor embedding (t-SNE) map constructed with these genes, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis results of these differentially expressed (DE) genes. This database is publicly available at https://scdevdb.deepomics.org. It helps researchers within the fields of developmental biology to facilitate gene expression studies in human single cells.

MATERIALS AND METHODS

Transcriptomic Data Collection

We searched the National Center for Biotechnology Information Gene Expression Omnibus database using the successfully utilized keywords, single-cell RNA-seq, single-cell RNA-seq, single-cell transcriptome, and selected the species to humans. In this study, we only focused on the normal human developmental processes; thus, we abnegated experiments using tumor and other samples treated with chemical reagents. After carefully reviewing the resultant papers and datasets, we obtained 10 datasets for human single-cell RNA-seq using normal cell type, tissue, or organs. These datasets including human cell groups related to the nervous system, digestive system, the heart, the brain, hESC, cell lines, and others. Single cells originating from the same cell lines, tissue regions, or organ regions at the same developmental time points are treated as a cell group. Based on this rule, we classified the 18,413 single cells into 176 cell groups (**Supplemental Table S1**). Cell groups originating from the same cell lines, tissue regions, or organ regions but at different developmental time points were regarded as one developmental stage. Therefore, the 176 cell groups were merged into 35 developmental stages (**Supplemental Table S1**).

Data Processing and Gene Expression Profiling Analysis

For the selected RNA-Seq experiments, the gene expression matrices were also retrieved from the Gene Expression Omnibus. For cells in datasets where the fragments per kilobase of exon per million reads mapped (FPKM) were available, we computed the TPM for gene i in cell j, according to:

$$TPM_i = \left(\frac{FPKM_i}{\sum_j FPKM_j} \right) \times 10^6$$

This conversion enables the units to be consistent for dataset-to-dataset comparison. Then, for each dataset, we merged cells originating from the same tissue or organ into one file and performed imputation using the R package single-cell analysis via expression recovery with default parameters. Single-cell analysis via expression recovery takes in a matrix and performs library size normalization during denoising step, which can reduce noise including sequencing depth, the number of cells, and cell composition (Huang et al., 2018). We eventually got 176 different files which are consistent with 176 different cell groups.

Differential Gene Expression and T-SNE Analysis

For each developmental pathway, we merged the expression data of all developmental stages in this pathway into one file. Then, we conducted DE gene analysis between cell groups in the same developmental pathway using Monocle, which will do all needed normalization steps internally, with default parameters (Qiu et al., 2017). We extracted expression data of the DE genes and performed t-SNE analysis with different perplexity for different process (Pedregosa et al., 2011).

GO and KEGG Enrichment Analysis

The symbol names of DE genes were used as the gene list input into R packages "GOstats" (Falcon and Gentleman, 2006) and "KEGG.db" (Carlson et al., 2016) for GO and KEGG analysis, respectively. We selected the "ontology" parameter as "BP," "MF," and "CC" for GO analysis and "pvalueCutoff" parameter as 0.5 for both GO and KEGG analysis. Top 20 significantly enriched GO terms and KEGG terms were selected to show potential functions of DE genes.

Database Construction

The SCDevDB website was built using the Django Python Web framework (https://www.djangoproject.com) coupled

with the MySQL database. The front-end interface was developed based on the Bootstrap open source toolkit (https://getbootstrap.com). The web interactive visualization graphs were developed using Plotly JavaScript Open Source Graphing Library (https://plot.ly/javascript/). SCDevDB was published using the Apache http server and is accessible at https://scdevdb.deepomics.org/.

RESULTS

Datasets Summary and the Developmental Tree Construct

At the time of this publication, the database contains 10 datasets covering 18,413 single cells and 176 cell groups (see Methods). According to the notation of the data resources, we classified these cell groups into 35 developmental stages. Every mammalian individual is developed from the totipotent zygote. Mammalian preimplantation development is a complex process including a series of cell divisions from 1 to 2 cells, 2 to 4 cells, 4 to 8 cells, 8 to 16 cells, and 16 cells to blastocyst (Niakan et al., 2012). After that, nearly all of the human tissues are original from embryoblast (hESC). Then, a developmental tree was constructed based on the development process of the multicellular organism (Hall, 2012) (**Figure 1**). Specifically, we first considered the developmental process from oocyte to hESC as the root process; then, the left 27 developmental stages were classified into 24 different developmental pathways by combining with the root process (**Supplemental Table S1**). The detailed cell number in each stage is shown in **Figure 2**, and the datasets summary is available at https://scdevdb.deepomics.org/data-summary/.

User Interface to the SCDevDB

In order to provide users easy access to the SCDevDB, we designed an interface to allow users to perform basic operations, such as searching, viewing, and downloading data. SCDevDB is composed of two functional pages: "Gene Expression Search" page and "Differential Gene List Collection" page. The web interface of SCDevDB is summarized in **Figures 3** and **4**.

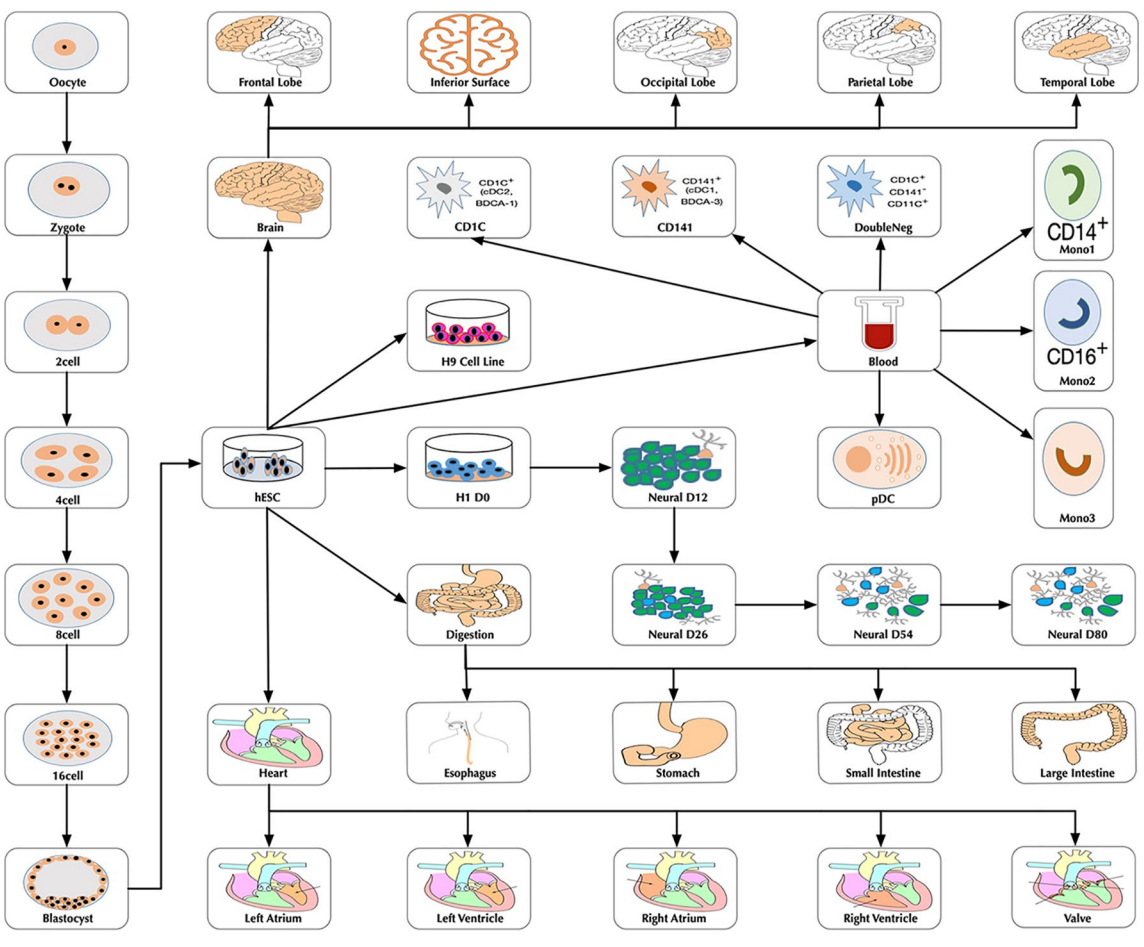

FIGURE 1 | The developmental tree. Figures of the brain, heart and digestion originate from Wikimedia Commons (https://commons.wikimedia.org/wiki/Main_Page).

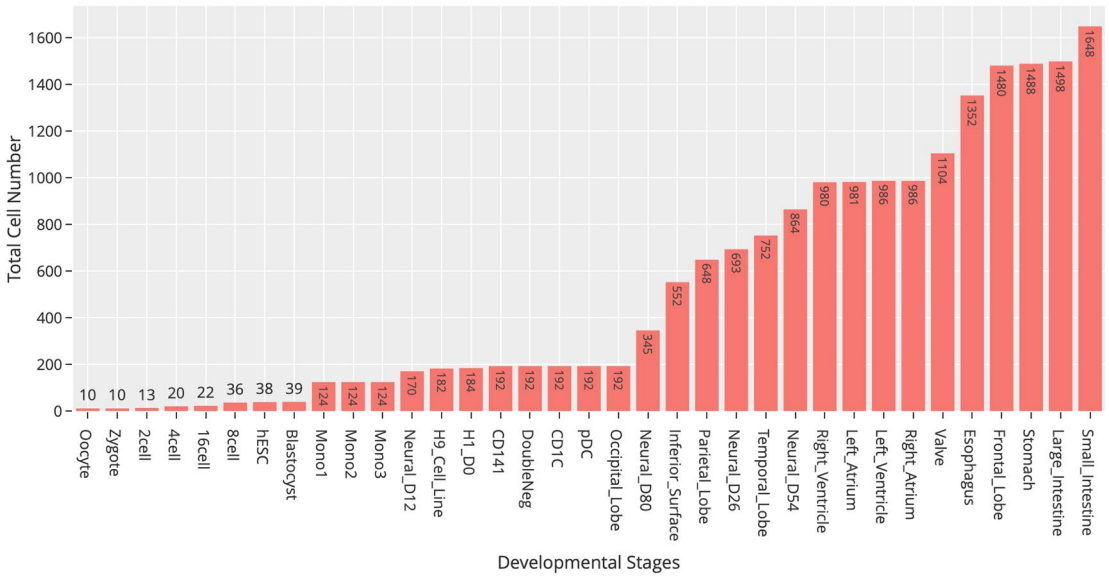

FIGURE 2 | Statistics of cell numbers of 35 developmental stages.

FIGURE 3 | Overview of the gene expression search page. **(A)** Searching result of the gene "MYL2". **(B)** Boxplot shows expression level distribution of MYL2 during developmental process by clicking the image. **(C)** The function of removing uninterested developmental stages by clicking the name of the stage listed in the figure legend. **(D)** An example of double clicking on a stage name.

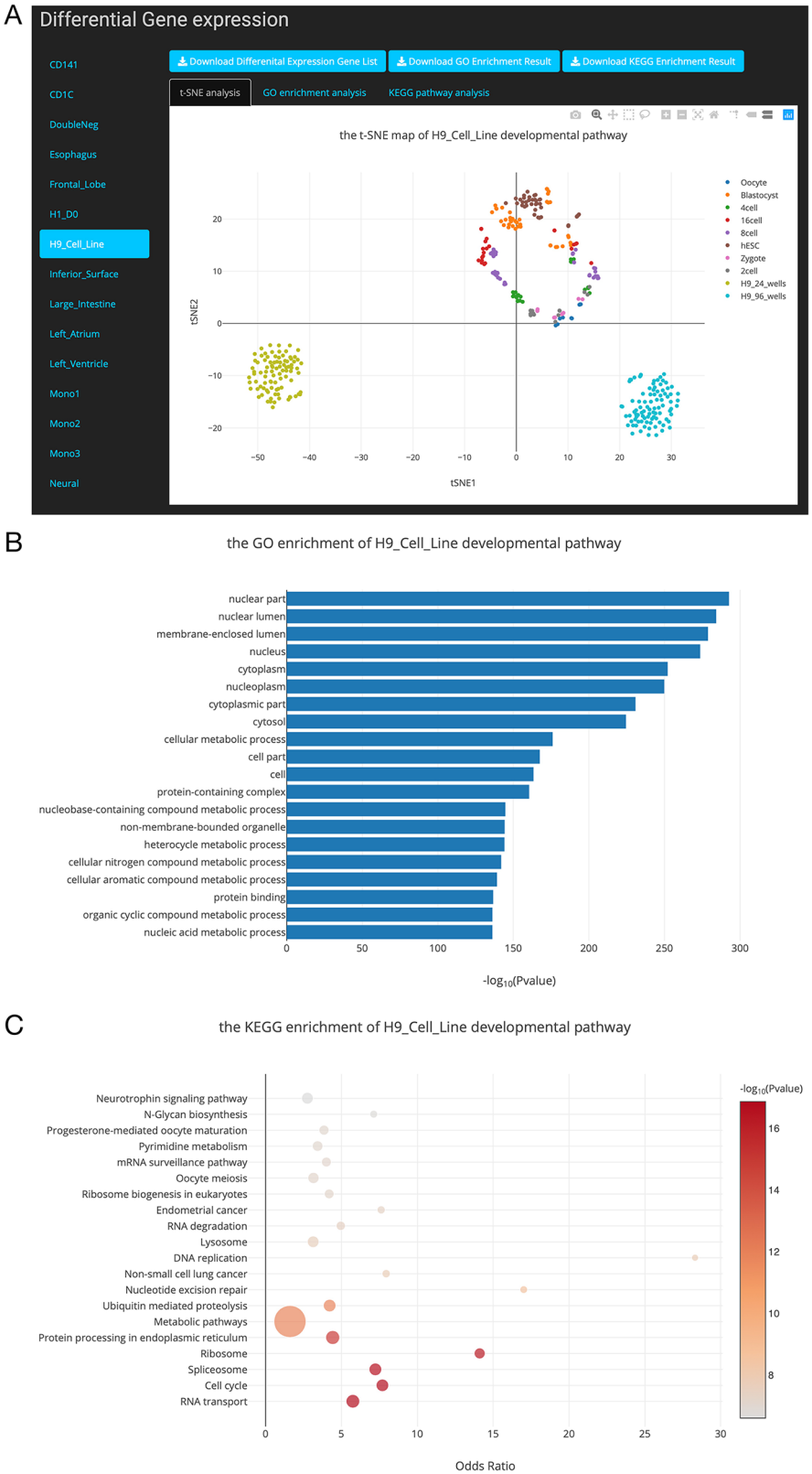

FIGURE 4 | Overview of the differential gene list collection page. **(A)**. T-distributed stochastic neighbor embedding (t-SNE) maps showing the relationships between developmental stages based on these differentially expressed genes. **(B)**. Top 20 Gene Ontology (GO) terms of differential expression genes. **(C)**. Top 20 Kyoto Encyclopedia of Genes and Genomes (KEGG) terms of differential expression genes.

Query Function to Search Gene Expression Across 35 Developmental Stages

In this page, users can view whether an interested gene is expressed in different developmental stages by giving a gene symbol (e.g., APMAP) or an Ensembl ID (e.g., ENSG00000101474) in the searching input box. The searching result will be displayed in the developmental tree. Specifically, if the searched gene ("MYL2" gene as an example) is not expressed at one stage, the stage image will be disabled and cannot be clicked (the light-colored images in **Figure 3A**). Furthermore, the interactive boxplot of gene expression level along with a selected developmental pathway is available by clicking the stage image (**Figure 3B**). To illustrate the interactive function of this boxplot, we took the distribution of the MYL2 expression during left ventricle process as an example. Clicking on the stage name "Left_Ventricle_E7w" listed in the graph legend can remove the boxplot data of this stage (**Figure 3C**). This function allows users to compare their interested stages. Moreover, double clicking on a stage name allows users to view detail gene expression value of this stage (**Figure 3D**). These boxplots can be download in PNG format for further usage.

Differential Gene List Collection for 24 Developmental Pathways

In this study, we performed DE gene analysis for 24 developmental pathways. Finally, 24 differential gene lists were collected into the SCDevDB. Users can download these gene lists by clicking the Download button in Differential Gene List Collection page. Moreover, we performed t-SNE analysis using these differential gene lists, and the result is displayed using an interactive scatterplot (**Figure 4A**). Subsequently, GO and KEGG enrichment analysis of the DE genes were performed using R packages, and top 20 significantly enriched GO or KEGG terms were selected to show potential functions of these DE genes (**Figures 4B**, **C**). In addition, tables showing all of the GO or KEGG terms are also available and free to download on the "Differential Gene List Collection" page. These scatterplots and bar chart can be downloaded in PNG format for further usage.

Case Study

Myosin light chain-2 (MYL2, also called MLC-2) is a protein that belongs to the EF-hand calcium binding protein superfamily and exists as three major isoforms encoded by three distinct genes in mammalian striated muscle (Sheikh et al., 2015). Diseases associated with MYL2 include cardiomyopathy, familial hypertrophic, and congenital fiber-type disproportion (Flavigny et al., 1998; Weterman et al., 2013). Here, we used this gene as an interested example to test the functions of SCDevDB. Previous studies using bulk-seq data have shown that MYL2 is highly expressed in tissue of muscles including skeletal muscle, myocardial, and smooth muscles (Hsu et al., 2012; Lindholm et al., 2014; Renaudin et al., 2018). Searching result of the SCDevDB is consistent with these studies as shown in **Figure 3A**. Moreover, comparing with the expression levels in cells of the atriums, MYL2 has higher levels in cells of the ventricles (**Figure 5**). This result indicated that MYL2 can be used as a marker gene to distinguish ventricle and atrium cells in subpopulation analysis.

hESC lines has been used as a source of cells for regenerative medicine, as well as valuable tools for drug discovery and for understanding human development and disease (Allegrucci and Young, 2006). Notably, H9 is one of the first five lines derived in the University of Wisconsin, i.e. H1, H7, H9, H13 and H14 (Denning et al., 2003), which has been used as an important material in many publications (Amit et al., 2000; Gafni et al., 2013; Kim et al.,

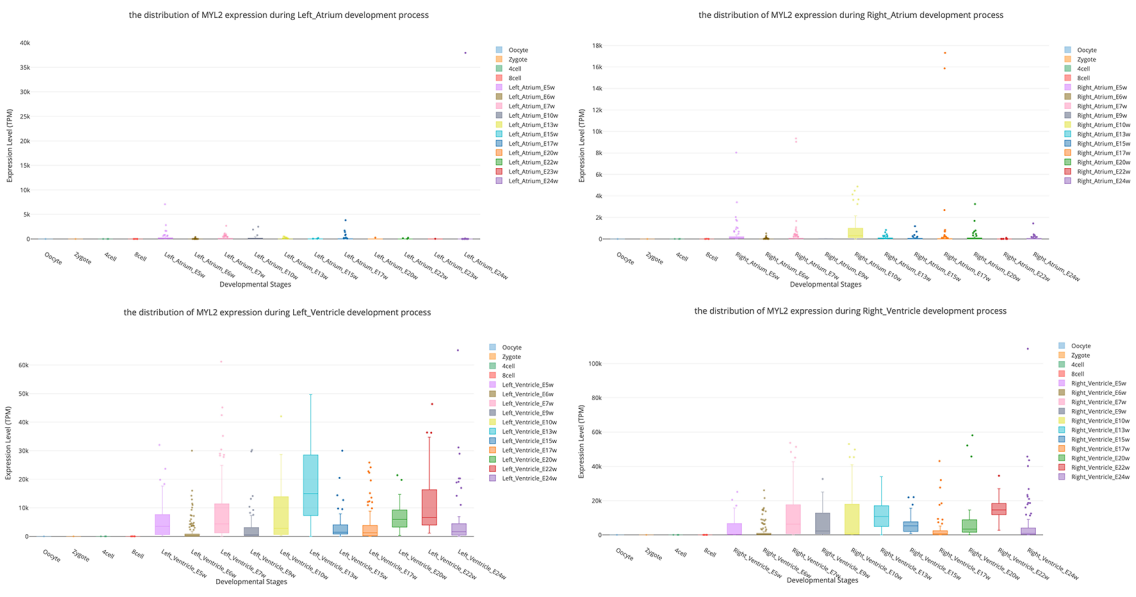

FIGURE 5 | Comparison of MYL2 expression distributions between atrium cell types and ventricle cell types.

2014). In our "Differential Gene List Collection" page, when we selected the "H9_Cell_Line" developmental pathway, the t-SNE map indicates that H9 cell lines are distinct from preimplantation cell types (**Figure 4A**). This result is reasonable as the H9 cell line are different from embryonic stem cells in expression levels of various genes (Telugu et al., 2013). Our GO and KEGG analysis results showed that the potential functions of the DE genes during the H9_Cell_Line developmental pathway were enriched in developmental-related biology processes including cellular metabolic process, nucleobase-containing compound metabolic process, RNA transport, and cell cycle pathways.

CONCLUSION

In summary, unlike previous databases, SCDevDB is an interactive database providing human single cell resources to profiling gene expression distributions in different developmental pathways. This database also provides DE gene lists in each developmental pathway, t-SNE map, and GO and KEGG enrichment analysis based on these differential genes. We believe that this database will facilitate researchers within the fields of developmental biology to investigate gene expression changes during human developmental pathways in the single-cell level.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://www.ncbi.nlm.nih.gov/geo/.

AUTHOR CONTRIBUTIONS

ZW performed the data collection and analysis, XF developed the database, SL designed and supervised the study, and ZW, XF, and SL wrote the manuscript.

FUNDING

This research was funded by a GRF Project grant from the RGC General Research Fund (9042181; CityU 11203115), the GRF Research Project (9042348; CityU 11257316).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00903/full#supplementary-material

REFERENCES

Abugessaisa, I., Noguchi, S., Böttcher, M., Hasegawa, A., Kouno, T., Kato, S., et al. (2017). SCPortalen: human and mouse single-cell centric database. *Nucleic Acids Res.* 46, D781–D787. doi: 10.1093/nar/gkx949

Allegrucci, C., and Young, L. (2006). Differences between human embryonic stem cell lines. *Hum. Reprod. Update* 13, 103–120. doi: 10.1093/humupd/dml041

Amit, M., Carpenter, M. K., Inokuma, M. S., Chiu, C.-P., Harris, C. P., Waknitz, M. A., et al. (2000). Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. *Dev. Biol.* 227, 271–278. doi: 10.1006/dbio.2000.9912

Cao, Y., Zhu, J., Jia, P., and Zhao, Z. (2017). scRNASeqDB: a database for RNA-Seq based gene expression profiles in human single cells. *Genes* 8, 368. doi: 10.3390/genes8120368

Carlson, M., Falcon, S., Pages, H., and Li, N. (2016). KEGG. db: A set of annotation maps for KEGG. *R Package Version* 3, 2016. doi: 10.18129/B9.bioc.KEGG.db

Denning, C., Allegrucci, C., Priddle, H., Barbadillo-Muñoz, M. D., Anderson, D., Self, T., et al. (2003). Common culture conditions for maintenance and cardiomyocyte differentiation of the human embryonic stem cell lines, BG01 and HUES-7. *Int. J. Dev. Biol.* 50, 27–37. doi: 10.1387/ijdb.052107cd

Falcon, S., and Gentleman, R. (2006). Using GOstats to test gene lists for GO term association. *Bioinformatics* 23, 257–258. doi: 10.1093/bioinformatics/btl567

Flavigny, J., Richard, P., Isnard, R., Carrier, L., Charron, P., Bonne, G., et al. (1998). Identification of two novel mutations in the ventricular regulatory myosin light chain gene (MYL2) associated with familial and classical forms of hypertrophic cardiomyopathy. *J. Mol. Med.* 76, 208–214. doi: 10.1007/s001090050210

Franzén, O., Gan, L.-M., and Björkegren, J. L. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* 2019 doi: 10.1093/database/baz046

Gafni, O., Weinberger, L., Mansour, A. A., Manor, Y. S., Chomsky, E., Ben-Yosef, D., et al. (2013). Derivation of novel human ground state naive pluripotent stem cells. *Nature* 504, 282. doi: 10.1038/nature12745

Gao, S., Yan, L., Wang, R., Li, J., Yong, J., Zhou, X., et al. (2018). Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat. Cell Biol.* 20, 721. doi: 10.1038/s41556-018-0105-4

Gittes, G. K. (2009). Developmental biology of the pancreas: a comprehensive review. *Dev. Biol.* 326, 4–35. doi: 10.1016/j.ydbio.2008.10.024

Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* 14, e8046. doi: 10.15252/msb.20178046

Hall, B. K. (2012). *Evolutionary developmental biology*. Springer Science & Business Media.

Hsu, J., Hanna, P., Van Wagoner, D. R., Barnard, J., Serre, D., Chung, M. K., et al. (2012). Whole genome expression differences in human left and right atria ascertained by RNA sequencing. *Circ. Cardiovasc. Genet.* 5, 327–335. doi: 10.1161/CIRCGENETICS.111.961631

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., et al. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539. doi: 10.1038/s41592-018-0033-z

Kim, J. J., Khalid, O., Namazi, A., Tu, T. G., Elie, O., Lee, C., et al. (2014). Discovery of consensus gene signature and intermodular connectivity defining self-renewal of human embryonic stem cells. *Stem Cells* 32, 1468–1479. doi: 10.1002/stem.1675

Ko, M. S. (2001). Embryogenomics: developmental biology meets genomics. *Trends Biotechnol.* 19, 511–518. doi: 10.1016/S0167-7799(01)01806-6

Lindholm, M. E., Huss, M., Solnestam, B. W., Kjellqvist, S., Lundeberg, J., and Sundberg, C. J. (2014). The human skeletal muscle transcriptome: sex differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. *FASEB J.* 28, 4571–4581. doi: 10.1096/fj.14-255000

Liu, N., Liu, L., and Pan, X. (2014). Single-cell analysis of the transcriptome and its application in the characterization of stem cells and early embryos. *Cell. Mol. Life Sci.* 71, 2707–2715. doi: 10.1007/s00018-014-1601-8

Merks, R. M., and Glazier, J. A. (2005). A cell-centered approach to developmental biology. *Physica A* 352, 113–130. doi: 10.1016/j.physa.2004.12.028

Niakan, K. K., Han, J., Pedersen, R. A., Simon, C., and Pera, R.A. (2012). Human pre-implantation embryo development. *Development* 139, 829–841. doi: 10.1242/dev.060426

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* 14, 309. doi: 10.1038/nmeth.4150

Renaudin, P., Janin, A., Millat, G., and Chevalier, P. (2018). A novel missense mutation p. Gly162Glu of the gene MYL2 involved in hypertrophic cardiomyopathy: a pedigree analysis of a proband. *Mol. Diagn. Ther.* 22, 219–223. doi: 10.1007/s40291-018-0324-1

Saliba, A.-E., Westermann, A. J., Gorski, S. A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* 42, (14) 8845–8860. doi: 10.1093/nar/gku555

Sanchez, A., and Golding, I. (2013). Genetic determinants and cellular constraints in noisy gene expression. *Science* 342, 1188–1193. doi: 10.1126/science.1242975

Sheikh, F., Lyon, R. C., and Chen, J. (2015). Functions of myosin light chain-2 (MYL2) in cardiac muscle and disease. *Gene* 569, 14–20. doi: 10.1016/j.gene.2015.06.027

Spitz, F., and Furlong, E. E. (2006). Genomics and development: taking developmental biology to new heights. *Dev. Cell* 11, 451–457. doi: 10.1016/j.devcel.2006.09.013

Telugu, B., Adachi, K., Schlitt, J., Ezashi, T., Schust, D., Roberts, R., et al. (2013). Comparison of extravillous trophoblast cells derived from human embryonic stem cells and from first trimester human placentas. *Placenta* 34, 536–543. doi: 10.1016/j.placenta.2013.03.016

Wang, D., and Bodovitz, S. (2010). Single cell analysis: the new frontier in 'omics'. *Trends Biotechnol.* 28, 281–290. doi: 10.1016/j.tibtech.2010.03.002

Weterman, M. A., Barth, P. G., Van Spaendonck-Zwarts, K. Y., Aronica, E., Poll-The, B.-T., Brouwer, O. F., et al. (2013). Recessive MYL2 mutations cause infantile type I muscle fibre disease and cardiomyopathy. *Brain* 136, 282–293. doi: 10.1093/brain/aws293

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-Y., Feng, Y., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500, 593. doi: 10.1038/nature12364

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1131. doi: 10.1038/nsmb.2660

Yao, Z., Mich, J. K., Ku, S., Menon, V., Krostag, A.-R., Martinez, R. A., et al. (2017). A single-cell roadmap of lineage bifurcation in human ESC models of embryonic brain development. *Cell Stem Cell* 20, 120–134. doi: 10.1016/j.stem.2016.09.011

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.