Unraveling the Genetic Drivers of Heart Failure from Cardiac Endothelial Cells via Single-Cell RNA-Sequencing Data and Machine Learning Model

Jisheng Zhong
Capital Medical University
Dongdong Wu
Fuwai Hospital
Junquan Chen
Tianjin Medical University
Aijun Liu
Capital Medical University
Gang Li
Capital Medical University
Junwu Su (sujunwu@ccmu.edu.cn)
Capital Medical University
Yu Liu
Capital Medical University

Research Article

Keywords:

Posted Date: May 8th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2875387/v1

Abstract

Background

Heart failure is a complex syndrome that hinders the heart's ability to provide oxygen to the tissues and is a significant cause of death globally. Given that left ventricular failure is more common than left atrial failure, this study utilizes single-cell RNA-sequencing data to detect differentially expressed genes (DEGs) between the endothelial cells of the two chambers and examines whether these DEGs are related to heart failure development.

Method

The analysis of a healthy human dataset was performed using Seurat, an R package, to identify differentially expressed genes (DEGs) between endothelial cells from the left ventricle and the left atrium of the heart. These findings were validated using two datasets that included both humans and mice with and without heart disease. The overlapped DEGs from the datasets were then utilized to develop a risk prediction model by using linear regression, which can determine an individual's risk for heart failure based on the expression levels of the overlapped DEGs.

Results

Seven genes, including MYL3, MYH6, TNNC1, FN1, B2M, MYL2, and SAT1, were identified with a significant p-value in all three datasets. Enrichment analysis has shown that these genes play a critical role in muscle contraction and heart regulation, and mutations in these genes have been linked to cardiomyopathy. The risk prediction model had a high accuracy rate of 85% in an independent validation dataset.

Conclusion

This study has successfully identified significant genes in endothelial cells that are associated with heart failure and may explain the difference in morbidity between the left ventricle and left atrium.

Introduction

Heart disease remains a leading cause of death worldwide, causing over 17.9 million deaths annually and accounting for an estimated 32% of all deaths (CDC). Among various heart diseases, heart failure, a condition in which the heart muscle is unable to pump sufficient blood to meet the body's demands, continues to increase in prevalence and severity. Heart failure affects over 26 million people globally, contributing to approximately 287,000 deaths in the United States alone each year (Ponikowski et al. 2014; Owen et al. 2009). Despite the varied environmental factors associated with heart failure, recent

studies have revealed that genetic factors may play a critical role in the formation of this disease (Czepluch et al. 2018). For instance, the genetic form of dilated cardiomyopathy accounts for nearly 40% of all cases (Mestroni et al. 2014). However, there is limited understanding of how genetic factors influence the prevalence of heart failure in different cardiac chambers. Previous research has suggested that heart failure occurs more frequently in the left ventricle than in the left atrium (Nagy et al. 2018), but the underlying genetic mechanisms remain unclear.

Previous studies have identified significant differences in gene expression levels between cardiomyocytes in the left ventricle and left atrium (Li Wang et al., 2020). Cardiac endothelial cells (ECs) play a crucial role in the contraction and regulation of the heart by lining the inside of heart muscles and blood vessels. ECs release specific molecules, such as nitric oxide, which directly affect the rhythmicity of the heart. The present study aims to identify differentially expressed genes (DEGs) in the endothelial cells of the left ventricle and left atrium associated with the formation of heart failure and to elucidate why heart failure is more prevalent in the left ventricle. This study adopts a similar approach to previous single-cell research in identifying DEGs in cardiomyocytes to examine the role of DEGs in left ventricular and left atrial endothelial cells (ECs) in the formation of heart failure. Through this approach, the study aims to uncover new insights into the genetic mechanisms underlying heart failure and help develop more effective treatment strategies.

Methods

Single-cell RNA Sequencing Data. In this study, five public single-cell RNA sequencing datasets (Table 1) were collected to investigate different aspects of heart disease. The first dataset included 5,055 cardiac cells from 10 healthy samples to identify endothelial cells and differentially expressed genes (DEGs) between the left ventricle and the left atrium. The second dataset included 274 cardiac cells from 2 healthy samples and 877 cardiac cells from 4 patients with dilated cardiomyopathy (DCM) to identify DEGs between healthy controls and DCM patients. The third dataset consisted of 2,460 healthy mouse cells and 1,957 cells from mice with transverse aortic transcription, a surgical procedure performed on mice to investigate pressure overload-induced cardiac hypertrophy and heart failure. Samples from this dataset were collected periodically over an 11-week period, and the cells from each sample were sequenced and analyzed periodically to identify significant DEGs between samples in week 1 and week 11. The fourth dataset included 1,440 cardiac cells from 2 coronary atherosclerotic disease samples and 274 cells from 2 healthy samples to build an external independent validation dataset for the disease prediction model. Lastly, the fifth dataset was composed of 145,028 fetal cells (13,315 fetal ECs) from day 90 to day 122 to examine significant genes' expression levels during fetal development over time. These datasets were used to gain insights into the genetic mechanisms underlying heart disease and identify potential biomarkers for early detection and treatment of heart failure.

Table 1
Summary of the five public datasets studied

Dataset	1	2	3	4	5
Species	Homo Sapiens	Homo Sapiens	Mus musculus (C57BL/6 mice)	Homo Sapiens	Homo Sapiens
Number of Cells	5,055 cardiac cells from 10 healthy samples	274 cardiac cells from 2 healthy samples, 877 cardiac cells from 4 patients with dilated cardiomyopathy (DCM)	2,468 healthy cells, 1,957 cells from samples with transverse aortic constriction (TAC). The mice developed TAC over an 11 week period	1,440 cardiac cells from 2 coronary atherosclerotic disease (CAD) samples, 274 cells from 2 healthy samples	145,028 human fetal cells
Source of Data	Wang Li, et al., 2020	Wang Li, et al., 2020	Ren Zongna, et al., 2020	Wang Li et al., 2020	Cao Junyue, et al., 2020
Collection Techniques	Single live cells, defined by Hoechst-positive and Propidium Iodide-negative staining, were selected and subjected to reverse transcription.				

Data Analysis of snRNA-seq. Only genes expressed in at least three cells and cells with a detected gene count higher than 200 were retained. Additionally, cells with a high percentage of mitochondrial genes (> 5%) were filtered out. The R package SCTransform (Christoph et al., 2019) was used to normalize gene expression for each cell by fitting the Gamma-Poisson generalized linear model. The resulting log-transformed, normalized single-cell expression values were used for visualizations and differential expression tests. Statistically significant principal components were determined by a resampling test and were retained for the Uniform Manifold Approximation and Projection (UMAP) analysis. Differential expression analysis among clusters was conducted using a likelihood-ratio test, comparing cells within each cluster against all other cells. Gene A was defined as a biomarker for cluster X if it was detected in at least 25% of cells, had an adjusted p-value less than 0.05, and had a log e fold change of at least 0.25 between cells of cluster X and all other cells. These analyses were performed using the Seurat package v4.0. DEGs were analyzed for GO terms and KEGG pathways enrichment by using KOBAS(Xie et al., 2011). A significance threshold of FDR < 0.05 was used during the enrichment analysis to identify significant results. The protein-protein interaction network was analyzed by using the STRING website (Damian et al., 2023)

Risk Prediction Model. Subsequent to the identification of commonly expressed differentially expressed genes (DEGs), a linear regression model was developed to predict disease status, utilizing the identified genes as the foundation. The pipeline and framework of the training and testing of the model are depicted in Fig. 1 below. In order to construct and evaluate the risk prediction model, the fourth dataset was utilized. The expression levels of the overlapping DEGs, along with the CAD disease status of each sample, were employed as input for a 10-fold cross-validation model. 90% of the data was reserved for training the model, while the remaining 10% was utilized to assess its accuracy. Subsequently, the trained

model was employed to validate the second dataset comprising of DCM patients. The Receiver Operating Characteristic (ROC) curve was used to determine the precision of the prediction model.

Results

Heterogeneity of ECs between LA and LV. By analyzing single-cell RNA sequencing data from healthy cardiac cells obtained from the left ventricle and left atrium in the first dataset, we identified 11 distinct clusters (Fig. 2a). Feature plots revealed that PECAM1 and VHF, well-known EC biomarkers, were highly expressed in clusters 0, 5, and 10 (Fig. 2b). We zoomed in all ECs only and distinguished two relatively distinct clusters of ECs originating from the left ventricle and atrium (Fig. 2c), and we found 621 DEGs between these two chambers of the heart. The top 10 most significant DEGs' expression levels are represented in the heatmap (Fig. 2d).

To identify potential disease-causing genes, we compared gene expression levels between healthy controls and patients with dilated cardiomyopathy (DCM) in the second dataset, which yielded 146 differentially expressed genes (DEGs) associated with the disease. Additionally, we analyzed the third dataset to compare gene expression levels between week 1 and week 11 in mice with transverse aortic constriction (TAC), identifying 145 DEGs. These three datasets had seven overlapped genes: B2M, FN1, SAT1, MYL2, MYH6, MYL3, and TNNC1 (Fig. 3a). To explore protein-protein interactions between these seven genes, we constructed a network using STRING and found that MYL2, MYH6, TNNC1, and MYL3 were highly associated with each other (Fig. 3b). SAT1, an enzyme that catalyzes the acetylation of polyamines, showed no clear association with the other six genes.

We then analyzed the expression levels of the seven overlapped genes in the left ventricle and left atrium (Fig. 3c). MYH6, SAT1, FN1, and B2M were significantly upregulated in the left atrium, while MYL2, MYL3, and TNNC1 showed significantly higher expression levels in the left ventricle. These genes are involved in cardiac functions such as muscle filament sliding (p-value = 7.78E-12), regulation of the force of heart contraction (p-value = 2.99E-07), ventricular cardiac muscle tissue morphogenesis (p-value = 1.71E-11), and cardiac muscle contraction (p-value = 4.13E-08, Fig. 3d).

Predicting heart disease by using overlapping genes. To investigate the impact of the 7 overlapped DEGs on disease development, a linear regression model was trained using the expression matrix of these 7 genes in the second dataset, which contained approximately 1,100 cells from 2 healthy samples and 4 DCM samples. To validate the model, an external independent dataset of 1,700 cells from 2 healthy samples and 2 CAD samples, a related disease that also leads to heart failure, was used. Each cell is treated as a sample in the prediction model. The prediction model achieved excellent results, with 83.08% accuracy in the DCM dataset (Fig. 4a) and 85.67% accuracy in the independent CAD dataset (Fig. 4b), using the 7 overlapped genes as the basis. These results provide further evidence of the significant contribution of these 7 genes' expression levels to heart failure development and suggest a potential avenue for early prediction of heart failure.

Characterizing potential disease-causing genes in the development of the fetal heart. To understand if 7 potential disease-causing genes contribute to the development of heart disease by affecting the development of human's heart during fetal stage, the expression levels of these genes were analyzed during the fetal heart development. The data was collected over 7 different time periods and compared between adjacent days to generate 6 lists of DEGs. Among these lists, 65, 21, 42, 42, 82, and 45 significant DEGs were found between days 90 and 94, 94 and 110, 110 and 113, 113 and 115, 115 and 120, and 120 and 122, respectively. Two genes, B2M and FN1, were found to be common in both the previous list of 7 DEGs and these DEGs. B2M plays a key role in antigen presentation, processing, inflammation, the complement cascade, and stress response, and previous research suggests positive associations of higher B2M levels with cardiovascular disease outcomes. FN1, on the other hand, is involved in various cell processes, such as embryogenesis, wound healing, and blood coagulation. Interestingly, these two genes exhibit a similar expression patterns in the different time stage, indicates a potential joint action. Figure 5 shows the average expression levels of these two genes over time.

Discussion

This is the first study to utilize single-cell RNA-sequencing data to identify potential heart disease-causing DEGs in cardiac endothelial cells by comparing expression levels between the left ventricle and the left atrium, integrating over 100,000 cells from 5 public datasets. Seven genes were identified: MYL2, MYL3, MYH6, TNNC1, B2M, FN1, and SAT1. Previous studies have shown that MYL2, MYL3, MYH6, and TNNC1 genes work together to contract cardiac striated muscle, specifically cardiac muscle, and their mutations can cause left ventricular noncompaction and heart failure (Tian et al. 2014). MYL2 downregulation is associated with chronic heart failure, which may explain why left ventricular failure is more common than left atrial failure (Li et al. 2020). Although SAT1 has no known interactions with the other six genes, it was found to be generally expressed more in the left ventricle of the heart than in the left atrium and associated with the formation of heart failure (Thakur et al. 2019). Future research is needed to determine the specific role of SAT1 in heart function and its potential interactions with the other six genes.

We have also proved the strong association between the overlapped DEGs and heart failure formation by the high accuracy of the disease prediction model in distinguishing healthy samples and samples with heart disease. The accuracy of the external independent validation datasets reached a really great AUC as 0.85. However, further research is required to determine the specific impact of these genes on the different types of heart disease in the left ventricle and the left atrium. To improve the model's efficacy, future studies should integrate additional sequencing data and lifestyle factors since heart disease is influenced by both genetic and lifestyle factors.

There are still several limitations in this study that need to be addressed. Firstly, the mechanism of how these 7 genes contribute to the development of heart failure disease in the LV has not been fully validated by wet lab experiments. Although these genes have previously been reported to have a cardiac function, it is unclear if there are any differences in their function between the LV and LA. Furthermore, since this study only focused on EC-related genes, sequencing data from other cell types, such as cardiomyocytes

and fibroblasts, should be incorporated. The formation of heart failure disease is a complex process involving multiple cell types, and further research is needed to fully understand the disease mechanism.

Declarations

Author Contributions:

Conceptualization, Jisheng Zhong; Data curation, Gang Li; Formal analysis, Jisheng Zhong; Funding acquisition, Junwu Su; Investigation, Junquan Chen; Methodology, Jisheng Zhong; Project administration, Junwu Su; Resources, Junquan Chen; Software, Dongdong Wu; Supervision, Junwu Su; Validation, Dongdong Wu; Visualization, Jisheng Zhong and Aijun Liu; Writing – original draft, Jisheng Zhong; Writing – review & editing, Yu Liu, Aijun Liu and Gang Li.

Ethics approval and consent to participate:

Not applicable

Consent for publication:

Not applicable

Availability of data and materials:

All data analyzed are from public dataset. The dataset 1,2 and 4 can be found in the Gene Expression Omnibus (GEO) under accession codes GSE109816 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109816) and GSE121893 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121893). The dataset 3 is in the Gene Expression Ominbus (GSE120064, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120064). The dataset 5 is in the Gene Expression Ominbus (GSE156793,https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156793).

Competing interests:

Not applicable

Acknowledgment:

Not applicable

Sources of Funding:

Not applicable

References

1. Berggård I, Bearn AG. Isolation and properties of a low molecular weight beta-2-globulin occurring in human biological fluids. J Biol Chem. 1968 Aug 10;243(15):4095-103. PMID: 4175239.
2. Cao, J., et al. (2020). A human cell atlas of fetal gene expression. *Science*, *370*(6518). https://doi.org/10.1126/science.aba7721.
3. Czepluch, Frauke S., et al. "Genetic Determinants of Heart Failure: Facts and Numbers." *ESC Heart Failure*, vol. 5, no. 3, 2018, pp. 211–217., https://doi.org/10.1002/ehf2.12267.
4. Horton R., Wilming L., Rand V. Gene map of the extended human MHC. *Nat. Rev. Genet.* 2004;**5**(12):889–899. doi: 10.1038/nrg1489.
5. Li, Y., Wu, G., Tang, Q., Huang, C., Jiang, H., Shi, L., Tu, X., Huang, J., Zhu, X., & Wang, H. (2010). Slow cardiac myosin regulatory light chain 2 (MYL2) was down-expressed in chronic heart failure patients. *Clinical Cardiology*, *34*(1), 30–34. https://doi.org/10.1002/clc.20832.
6. Mestroni, Luisa, et al. "Genetic Causes of Dilated Cardiomyopathy." *Progress in Pediatric Cardiology*, vol. 37, no. 1-2, 2014, pp. 13–18., https://doi.org/10.1016/j.ppedcard.2014.10.003.
7. Nagy, A. I., Hage, C., Merkely, B., Donal, E., Daubert, J.-C., Linde, C., Lund, L. H., & Manouras, A. (2018). Left atrial rather than left ventricular impaired mechanics are associated with the pro-fibrotic ST2 marker and outcomes in heart failure with preserved ejection fraction. *Journal of Internal Medicine*, *283*(4), 380–391. https://doi.org/10.1111/joim.12723.
8. Owen JS, Khatib S, Morin DP. Cardiac resynchronization therapy. Ochsner J. 2009 Winter;9(4):248-56. PMID: 21603451; PMCID: PMC3096278.
9. Ponikowski, Piotr, et al. "Heart Failure: Preventing Disease and Death Worldwide." *ESC Heart Failure*, vol. 1, no. 1, 2014, pp. 4–25., https://doi.org/10.1002/ehf2.12005.
10. Ren, Zongna, et al. "Single-Cell Reconstruction of Progression Trajectory Reveals Intervention Principles in Pathological Cardiac Hypertrophy." *Circulation*, 26 Feb. 2020, www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.119.043053.
11. Shi, Fanchao et al. "Association of beta-2-microglobulin and cardiovascular events and mortality: A systematic review and meta-analysis." *Atherosclerosis* vol. 320 (2021): 70-78. doi:10.1016/j.atherosclerosis.2021.01.018.
12. String-db.org. 2021. *STRING: functional protein association networks*. [online] Available at: <https://string-db.org/> [Accessed 24 October 2021].
13. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun 13;177(7):1888-1902.e21. doi: 10.1016/j.cell.2019.05.031. Epub 2019 Jun 6. PMID: 31178118; PMCID: PMC6687398.
14. Thakur, V.S., Aguila, B., Brett-Morris, A. *et al.* Spermidine/spermine N1-acetyltransferase 1 is a gene-specific transcriptional regulator that drives brain tumor aggressiveness. *Oncogene* 38, 6794–6800 (2019). https://doi.org/10.1038/s41388-019-0917-0.
15. Wang, L., Yu, P., Zhou, B. *et al.* Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function. *Nat Cell Biol* 22, 108–119 (2020). https://doi.org/10.1038/s41556-019-0446-7.

16. Hafemeister, C., Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 20, 296 (2019). https://doi.org/10.1186/s13059-019-1874-1
17. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W316-22. doi: 10.1093/nar/gkr483. PMID: 21715386; PMCID: PMC3125809.
18. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023 Jan 6;51(D1):D638-D646. doi: 10.1093/nar/gkac1000. PMID: 36370105; PMCID: PMC9825434.
19. Tian, T., Wang, J., Wang, H. *et al.* A low prevalence of sarcomeric gene variants in a Chinese cohort with left ventricular non-compaction. *Heart Vessels* 30, 258–264 (2015). https://doi.org/10.1007/s00380-014-0503-x.

Figures

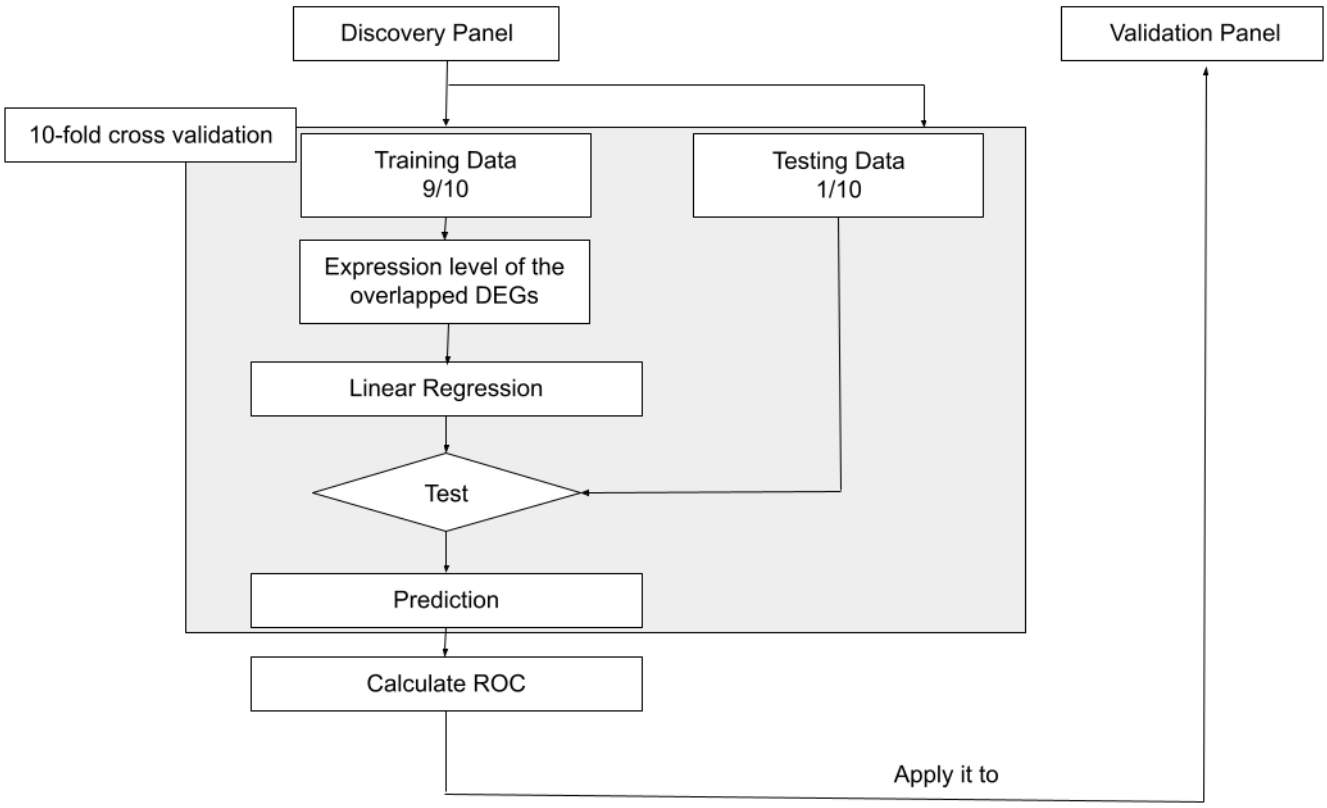

Figure 1

Disease prediction model framework.

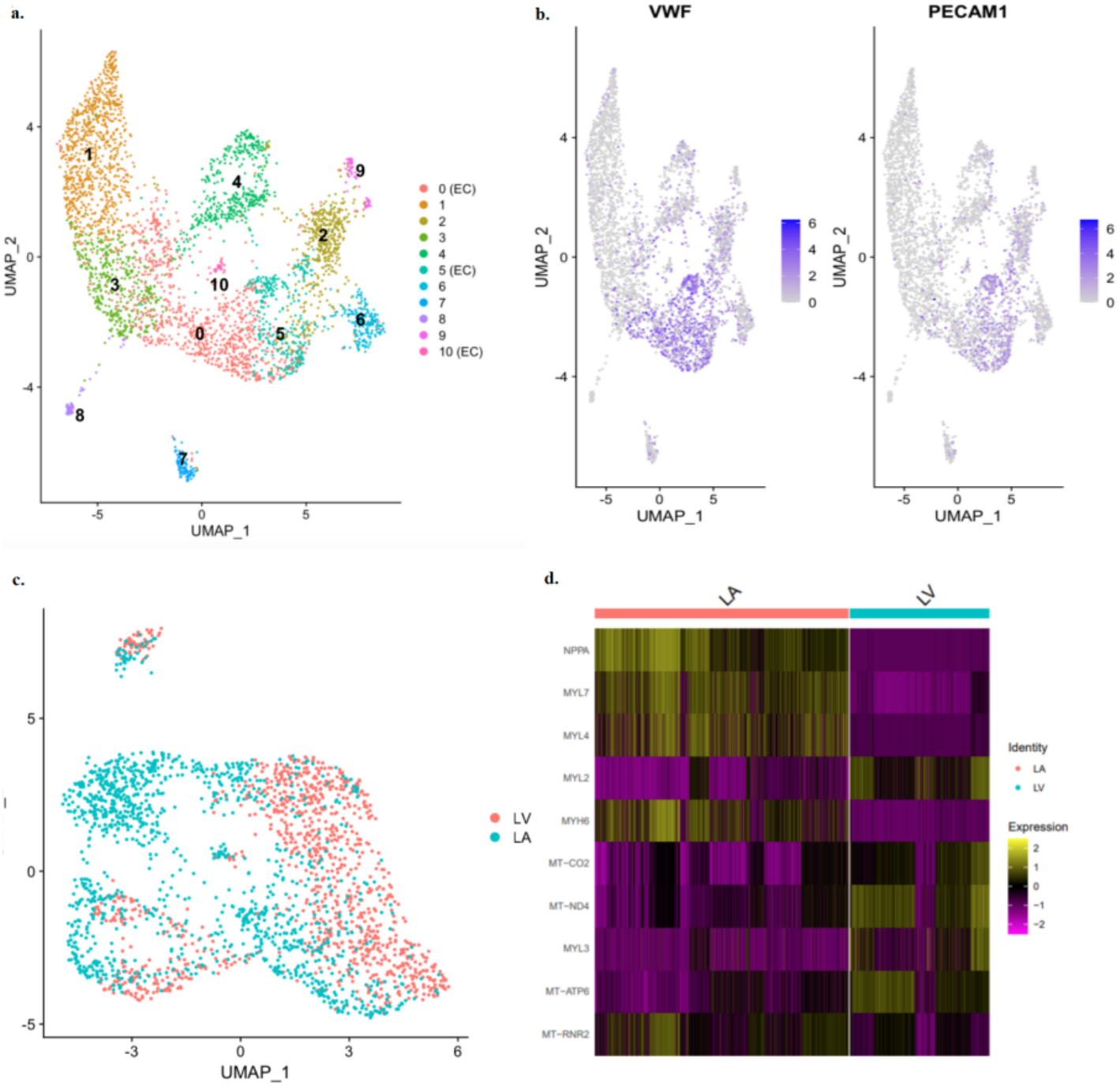

Figure 2

Identification and distribution of EC in different chambers. (a). UMAP of all 5,055 cells form 10 healthy samples. (b). Feature plots of EC marker genes VWF and PECAM1. (c). Distribution of EC in different chambers by zooming in ECs identified. (d). Heatmap of the top 10 DEGs between ECs of LA and LV.

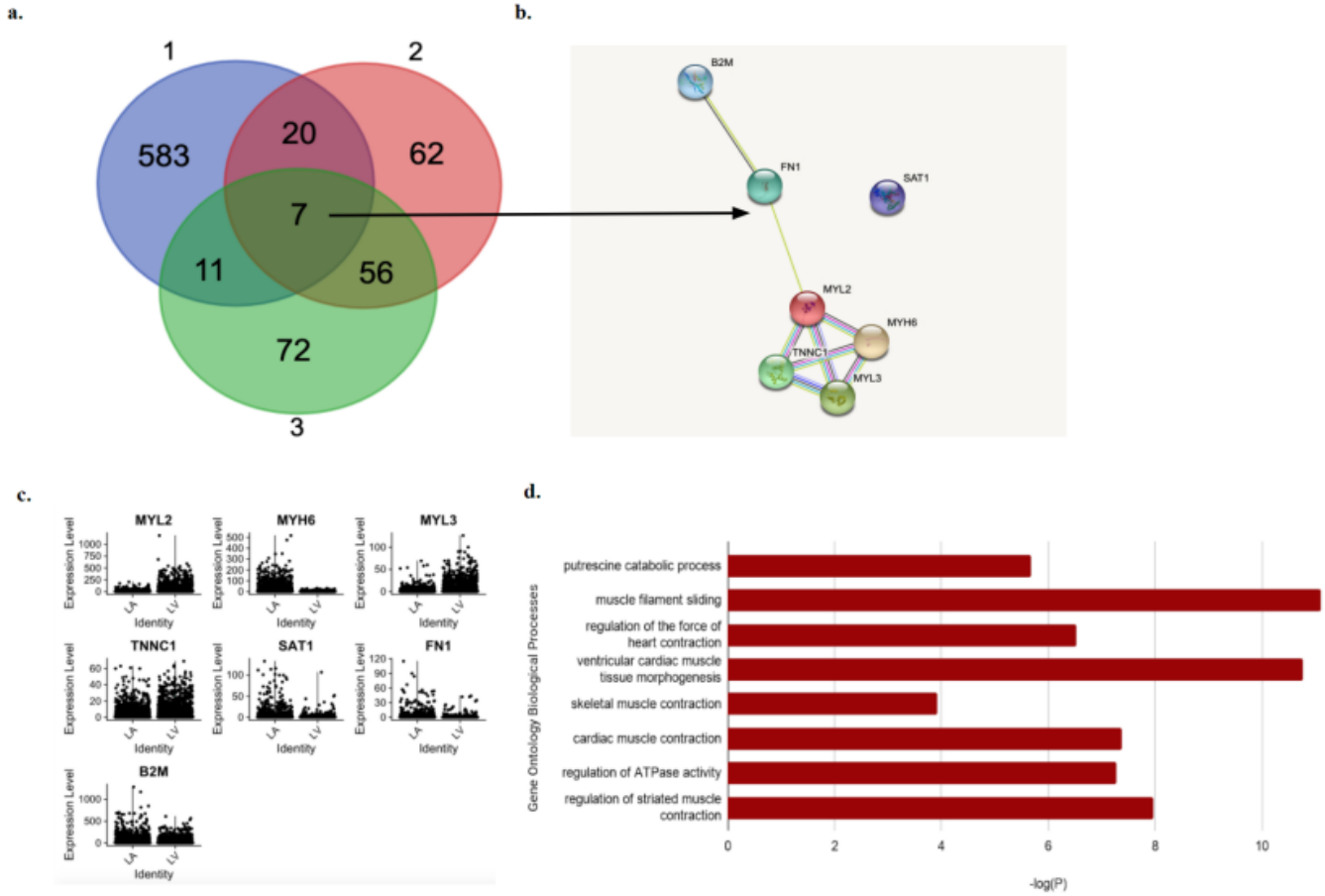

Figure 3

Identification of disease associated DEGs between ECs of LA and LV. (a). DEGs that were common between disease-associated DEGs and DEGs of ECs between LA and LV. (b). Protein-protein interaction network of overlapped 7 DEGs. (c). Expression levels of 7 DEGs in LA and LV. (d). Results of pathway enrichment analysis by using 7 DEGs.

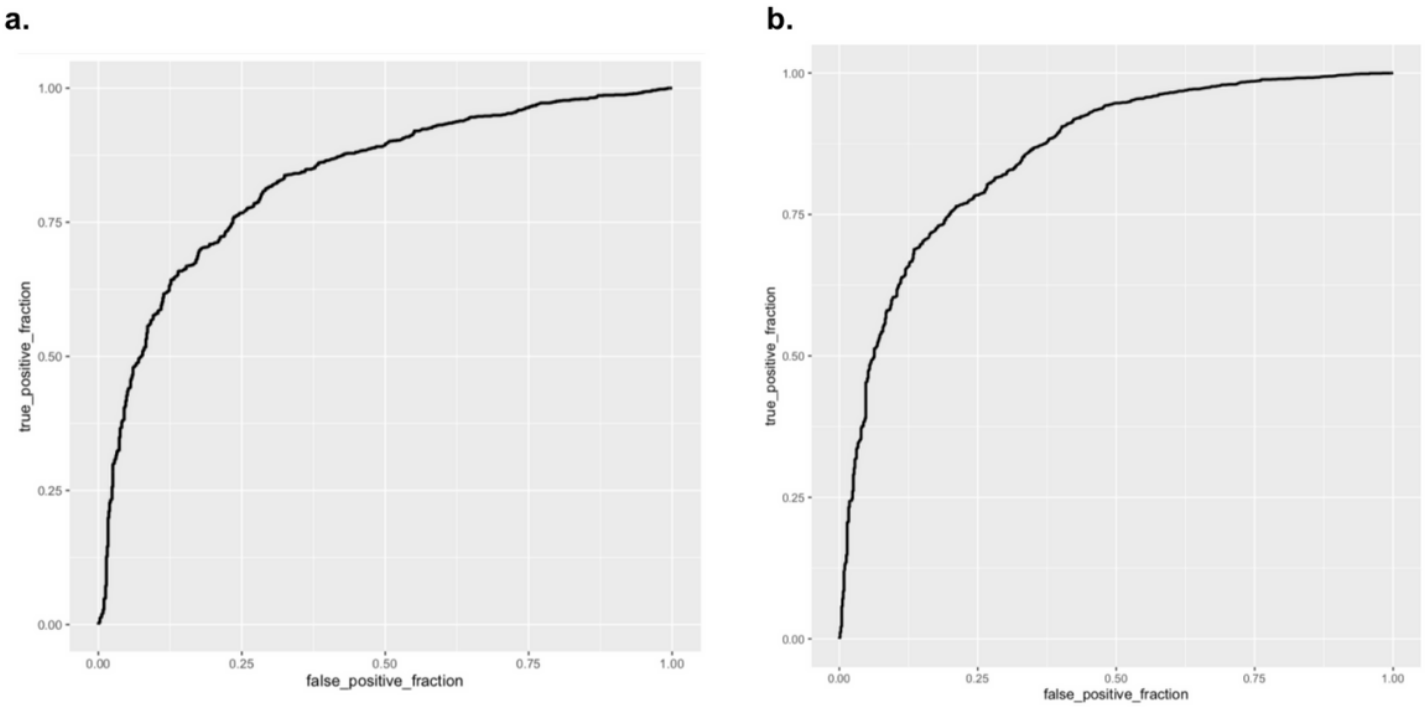

Figure 4

ROC of the prediction models by using the expression matrix of 7 overlapped DEGs. (a). The ROC of DCM prediction, AUC=0.8308. (b). The ROC of the independent CAD prediction, AUC=0.8567.

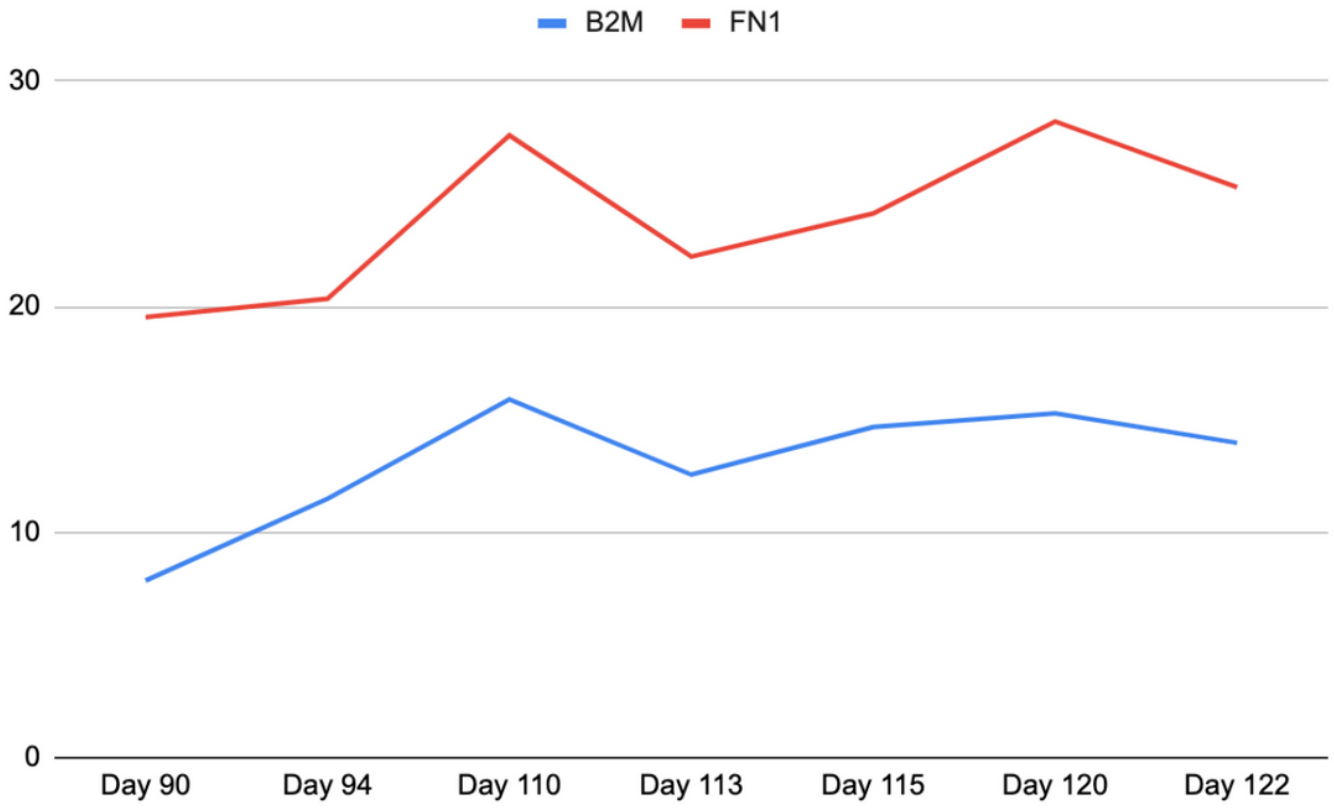

Figure 5

The average expression level of genes B2M and FN1 in different fetal time