Recognize Cell Types Underlying Coronary Artery Disease Across Humans and Mice

Haitian Hao[1], Jiyuan Yang[2], Siqi Zhang[3], Songwei Chen[4]

[1] *Department of Computer Science, University of Maryland, College Park, MD 20742, United States*

[2] *School of Information Engineering, Zhengzhou University, Zhengzhou, 450001, China*

[3] *Bryn Mawr School, Baltimore, MD 21212, United States*

[4] *Nanjing Foreign Language School Xianlin Campus, Nanjing, 210019, China*

Abstract. Heart disease, such as coronary heart disease, arrhythmia Dilated cardiomyopathy, and congenital heart disease, caused a large number of deaths in this world. We decided to analyze the possible causes of a specific type of heart disease - coronary artery disease (CAD). In this project, we managed to uncover relations between human cell types and CAD, based on relations across genes and cell types of humans and mice that we discovered by applying techniques including principal component analysis (PCA), K-nearest-neighbor (KNN) analysis and a very powerful R package tool - Seurat. We concluded that CAD has a significant connection with macrophages and Fibroblasts, which means that we can recognize the disease conditions of CAD patients by recognizing the status of macrophages and fibroblasts.

Keywords: Heart disease, CAD, Coronary artery disease, Human, Mouse, Gene, SNPs, Cell types, Single cell

1. Introduction

Heart disease is one of the most deadly diseases in the entire world. There are more than 26 million heart disease patients in the world, and the funds spent on heart disease are 108 billion US dollars each year. The risk of heart disease increases by age, in general, the risk of men is higher than that of women. Among people over 40, one in five (20%) is at risk of heart disease. Heart disease is the main reason for hospitalization for patients over 65 years of age. The 5-year survival rate is less than 50%, and the mortality rate is 2-3 times that of breast and bowel cancer.

Our research group chose to focus on coronary artery disease (CAD), one of the most representative types of heart disease. The complex phenotype of CAD is driven by genetic and environmental factors. Since the first CAD genome-wide association study (GWAS) was conducted in 2007, the sample size of many other studies has gradually increased, and 97 genome-wide important genetic loci related to CAD5-10 have been identified during analysis. In 2018, other researchers discovered 64 new loci related to CAD[3].

When we carried out our research, we were frequently restricted by the problem caused by insufficient samples[1]. In many studies, researchers use animal models and genomics to study related diseases[4,5]. Inspired by that, we used a combination of mice and humans models to study the expression of coronary artery disease (CAD) pathogenic genes in different cell types. We used single-cell analysis to find the expression of genes in mouse cells. Then, we built a link between humans and mice through the COE correlation values to study the expression of CAD-related pathogenic genes.

2. Data Access

2.1 Human pathogenic gene data are available here.
2.2 Raw data of mouse single cells are available at ArrayExpress E-MTAB-6173.

In this study, the dataset contains quantitative relations between mice genes and cells. The values in this matrix represent the number of molecules for each feature (i.e. gene; row) detected in each cell (column).

3. Methods and results:

3.1 Find disease-causing genes of human:

In the paper[3] by Pim van der Harst and Niek Verweij, A genome-wide association study was conducted on 34541 CAD cases and 261984 UK Biobank resources, and then 88192 cases and 162544 cases were reproduced from CARDIoGRAMplusC4D. Based on the research, we found disease-causing genes related to CAD based on single-nucleotide polymorphisms (SNPs) associated with CAD.[3]

3.2 PCA model

Principal Component Analysis (PCA) technique is a dimensionality-reduction method that is often used to reduce the dimensionality of a group of large data. It aims to transform a large set of variables into a smaller one that still contains most of the information in the large set. Since our purpose of this research is to discover the several most influencing CAD-related genes, we inevitably need to reduce the genes of interests from all to a few. Thus, PCA is the first tool we adopted to analyze the data. However, reducing the number of variables of a dataset naturally comes at the expense of accuracy, we managed to reduce the number of variables while preserving as much information as possible.

3.2.1 Pre-processing workflow

We filtered out cells with a unique feature count of more than 2500 or less than 200 and a mitochondrial count of more than 5%.

3.2.2 Normalize data

In this work, the unnecessary cells were deleted before we officially started operating on the data. "LogNormalize" is our global normalization method. Specifically, the measured value of the characteristic expression and the total expression of each cell are first normalized by us, and then multiplied by the scale factor (the default is 10,000), and the result is obtained after logarithmic transformation.

3.2.3 Feature selection

We identified highly variable work by calculating a subset of features that exhibited high intercellular differences in the data. Specifically, they were highly expressed in some cells and low in others. If we focused on these genes in downstream analysis, it would be easier for us to highlight the biological signals in the single-cell data set. Each data set returned 2000 features.

3.2.4 Scale data and perform linear dimensional reduction

First, the expressions of each gene were altered by us so that the average expressions of the entire cell became zero. In order to make genes with high expression levels not dominate, equal weights need to be obtained in downstream analysis. In order to achieve that, we scale the expression of each gene such that the variance of the whole data equal to 1. After that, we performed the PCA analysis on the data, in which the input was the previously determined variable features.

We used the powerful tool in R - Seurat to visualize data. Exploring the main source of heterogeneity in the data set was an important step. For this, we applied the DimHeatmap library, a helpful tool in determining which PCs to include for further downstream analysis. In Figure 1, genes were ranked based on their PCA scores, giving us sufficient insights of the inner relations of the data.

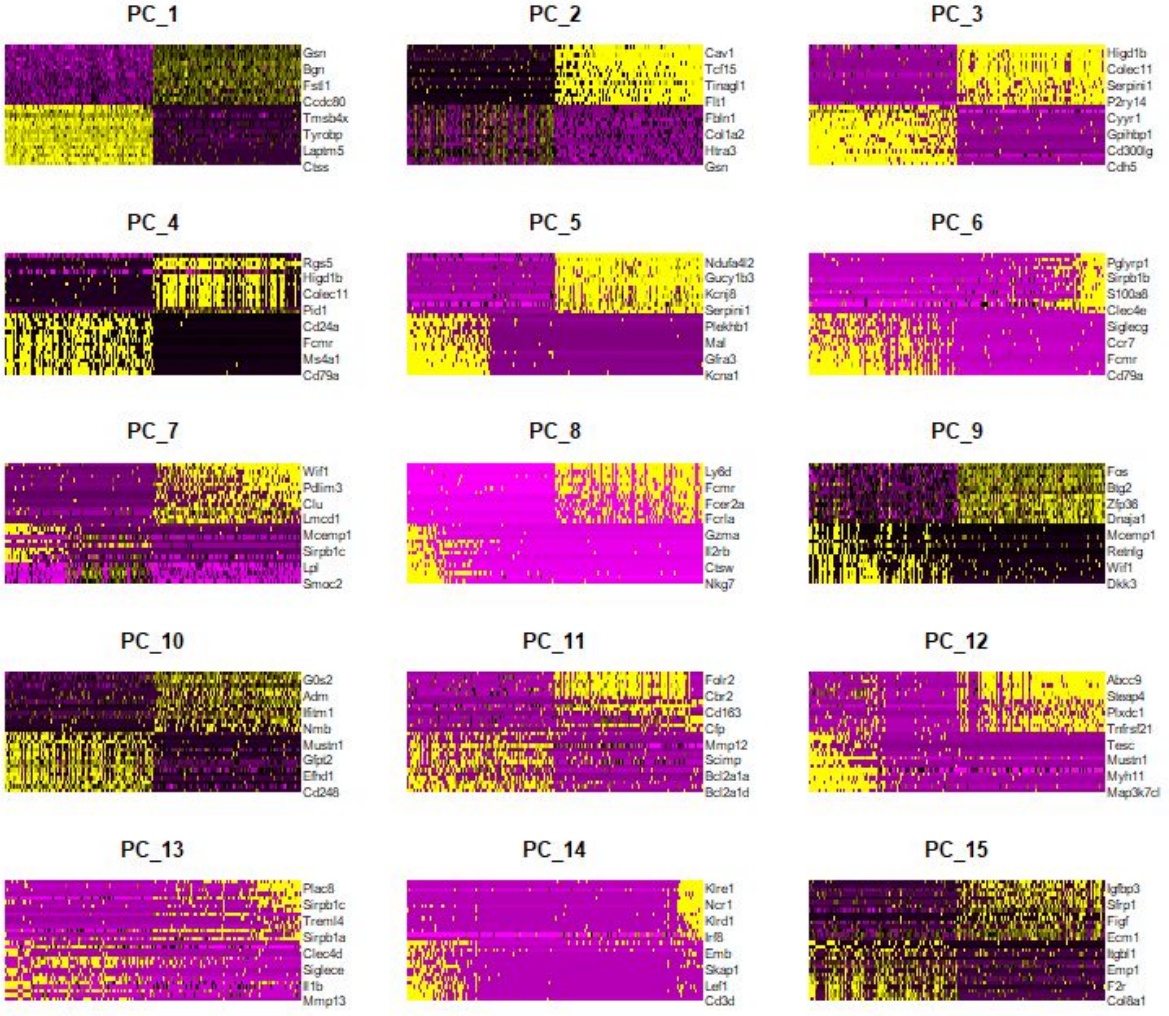

Figure 1. The main sources of heterogeneity in a dataset

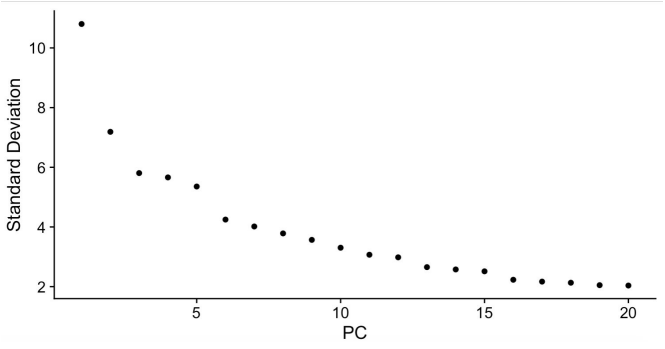

Figure 2. The Elbow plot

We drew the elbow diagram to determine the number of clusters we would use in our analysis. The plot conforms to the percentage of variance of each principal component, that is, the elbow plot shows the rankings of the principal components based on their impacts. In this example, we observed a change of trend at about 20

on x-axis, which indicates that most signals of interests were within the first 20 PCs. Thus, the dimensions were set to 20 in the following steps.

3.3 Expression of different genes in mouse cells

3.3.1 Cluster the cells and Run non-linear dimensional reduction

Using the Euclidean distance in the PCA space, we first constructed a KNN (K-nearest neighbors) graph and then adjusted the edge weight (Jaccard similarity) between any two cells based on the shared overlap in its local neighborhood. Use the FindNeighbors function to perform this step and take the previously defined dimensions of the data set (the first 20 PCs) as input.

We applied some Seurat library tools to perform nonlinear dimensionality reduction to visualize and explore the data. Eventually, genes were clustered into 13 clusters based on their composition.

3.3.2 Assigning cell type identity to clusters

We matched the unbiased clusterings to known cell types using cell markers. The only problem was that at some point we ran into troubles when matching the 11th cluster, since it was difficult to distinguish whether the 11th cluster is related to Granulocytes or Macrophages. After filtering, 11 clusters were related to Granulocytes by expressing differential genes Csf3r (myAUC = 0.642) and Slpi (myAUC = 0.586)(Table 1). In addition, this result conforms to the previous studies - it has been proved in previous studies that both Csf3r and Slpi are related to Granulocytes.[2]

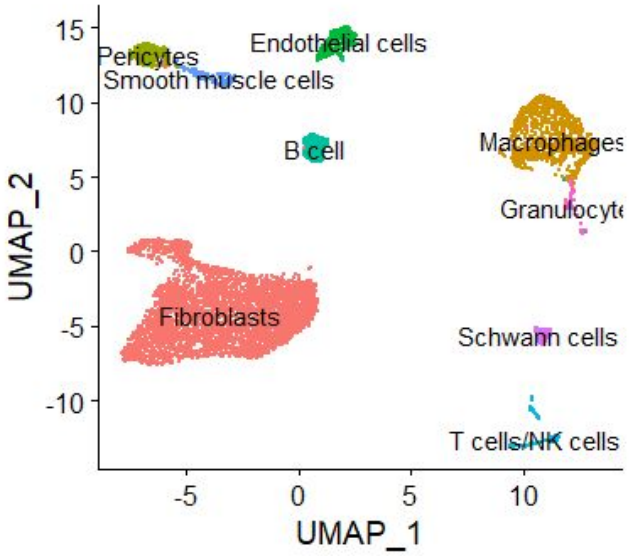

Figure 3. According to cell markers, we converted 13 clusters into 10 cell types: Fibroblasts, Macrophages, Pericytes, Endothelial cells, B cells, T cells/NK cells, Smooth muscle cells, Schwann cells, Granulocytes. In addition, T cells and NK cells are divided as a cluster.

3.4 Establishing links between human genes and mouse genes by COE value

We first applied the CAD-related human genes[3] and filtered the genes with FDR smaller than 0.05, which were considered significant. Then, we compared these candidate genes with the clustered mouse marker genes and thus got the corresponding genes. To further screen the genes, we filtered one-to-one orthologs that are expressed in both

species based on their conservation of expression (COE) values. We eventually obtained 69 ideal human-mouse genes.

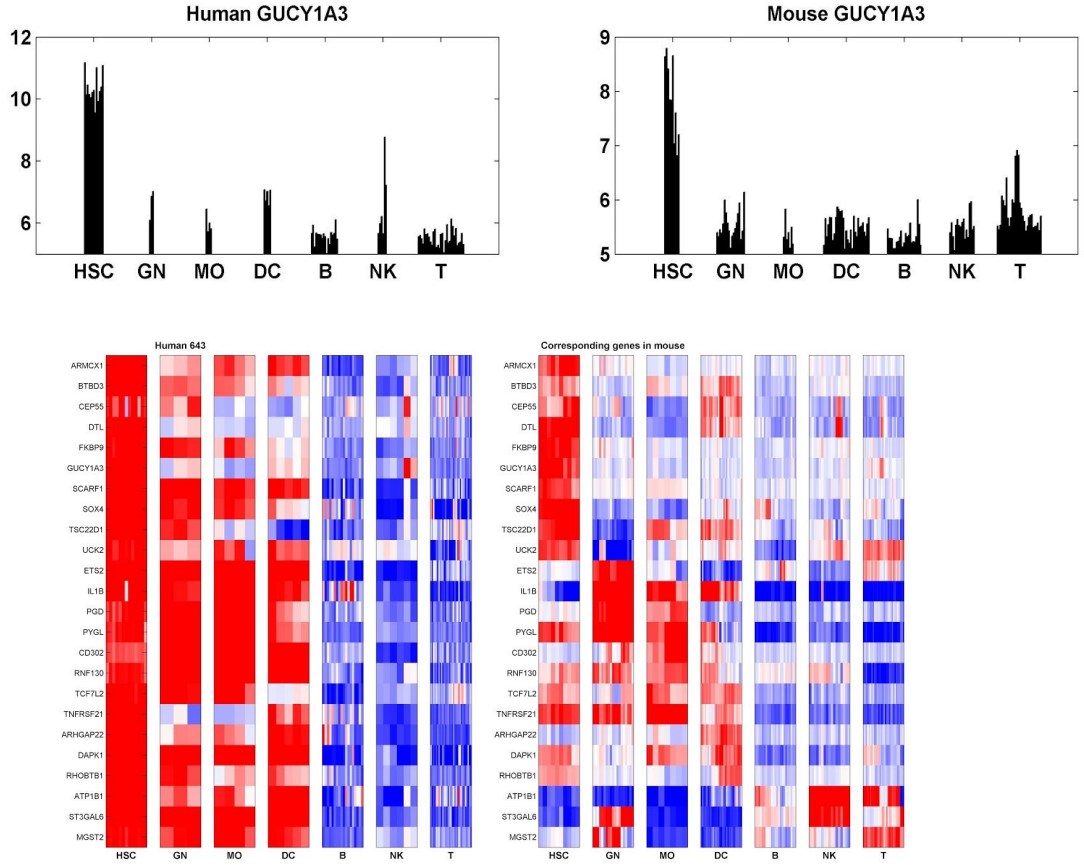

Figure 4. A heatmap of GUCY1A3 (COE=0.93) human gene module and its corresponding genes in mice. GUCY1A3 is one of the selected genes from all 69 human-mouse orthologs.

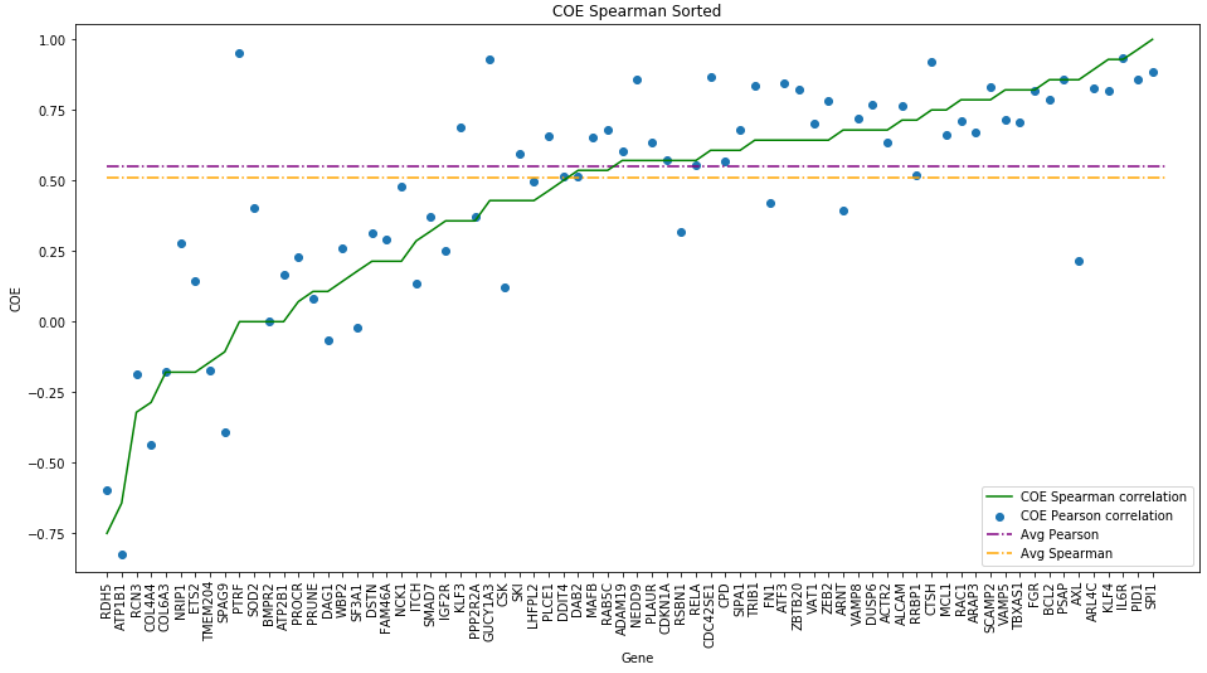

Figure 5. COE Pearson correlation and COE Spearman correlation between 69 respective human and mouse genes. Pearson correlation focuses on a linear relationship; Spearman correlation measures a monotonic relationship. These two methods show a similar pattern of the gene expressions.

3.5 Observe the expression of pathogenic genes in mouse cells

3.5.1 Expression probability distributions

After establishing the connection between mouse and human genes, we found 69 related genes, the following series of analysis. Among 69 genes found by COE value, we selected 42 genes with high correlation (|COE|>0.5)(Table 2). We selected a few of them to observe their expression. In this study, we found that most of them were related to macrophages. In order to prove our conclusion, we carried on our further analysis using clusterings.

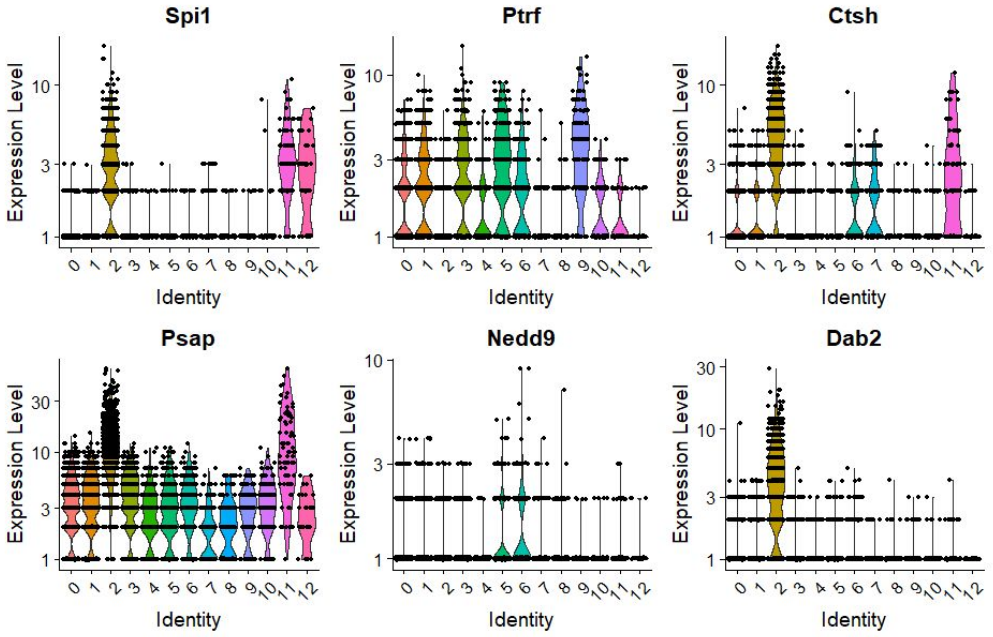

Figure 6. Expression probability distributions. The chart shows the expression of 6 gene components in clusters 0-12. clusters 0-12 respectively represent different cell types, and there may be two or more clusters representing the same cell type

3.5.2 Feature expression

We used VlnPlot to show expression probability distributions across clusters. We randomly selected 6 from 42 genes to observe. We discovered that Spi, Ptrf, Ctsh, Psap, Daba2 are all highly expressed in cluster 2.

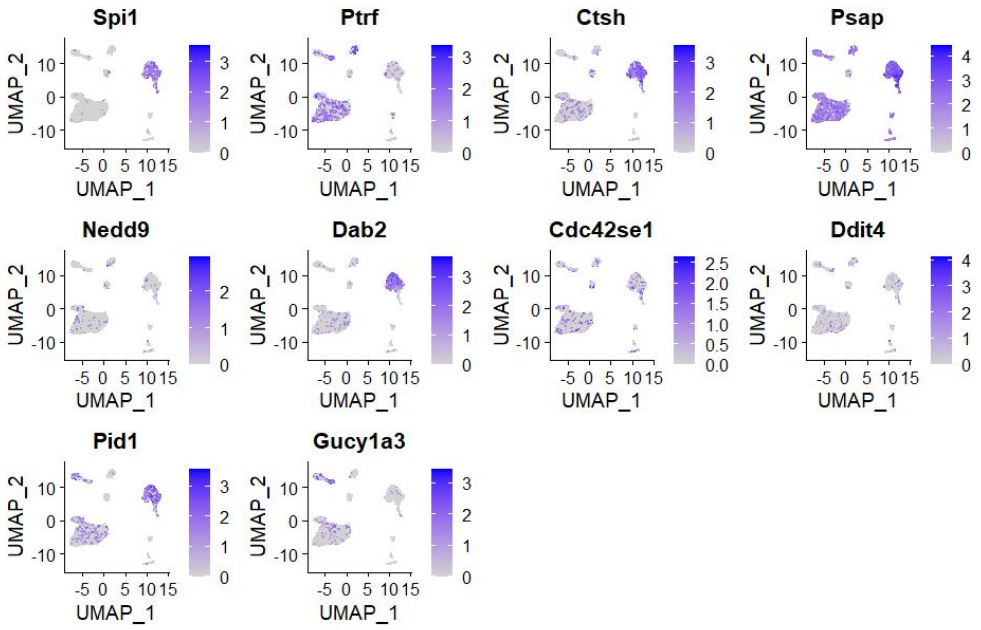

Figure 7. Feature expression. Similarly, the chart shows the PCA-based feature expression, indicating the genetic components in 13 clusters.

We also used the FeaturePlot tool to visualize feature expressions on a PCA plot. Similarly, we added 4 more genes to the already selected genes, making it a total of 10 genes for observation. In this study, most of them were expressed in macrophages, and a few were expressed highly in Fibroblasts, Schwann cells, Endothelial cells, and Pericytes.

4. Conclusion

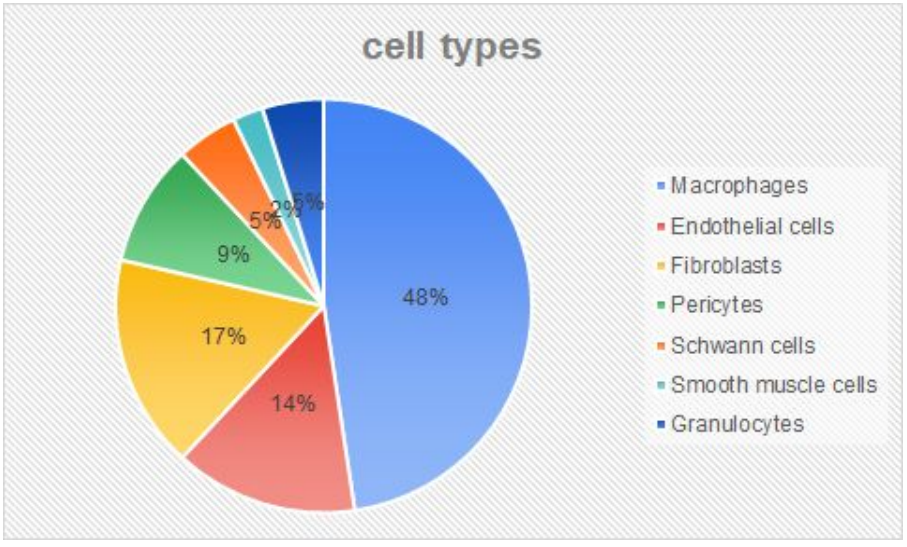

Figure 8. Cell types related to disease-causing genes.

In the present study, 42 genes were selected and 48% of them were highly expressed in macrophages and 17% are in Fibroblasts. Therefore, we concluded that CAD has a significant connection with macrophages and Fibroblasts, which means that we can recognize the disease conditions of CAD patients by cognizing the status of macrophages and fibroblasts.

5. Supplementary materials

5.1 Table 1

After clustering data with roc curve0, the origin data were divided into 12 clusters, and eventually assigned to 10 different cell types.

5.2 Table 2

After preliminary screening, 69 genes with correlation. After the second screening, it shows that 42 genes have high Pearson/Spearman correlation (|COE|>0.5) between humans and mice.

6. Reference

1. Liew CC, Dzau VJ. Molecular genetics and genomics of heart failure. Nature reviews Genetics. 2004;5(11):811–825. [PubMed] [Google Scholar]

2. Daniel A.Skelly, Galen T.Squiers, Micheal A.McLellan, Mohan T.Bolisetty, PaulRobson, Nadia A.Rosenthal, Alexander R.Pinto.Single-Cell Transcriptional Profiling Reveals Cellular Diversity and Intercommunication in the Mouse HeartCell Rep (IF: 8.109; Q1) . 2018 Jan 16;22(3):600-610. doi:10.1016/j.celrep.2017.12.072.[PubMed][Google Scholar]

3. Pim van der and Niek Verweij, Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ Res. 2018 Feb 2; 122(3): 433–443. doi: 10.1161/CIRCRESAHA.117.312086.[PMC free article][PubMed][Google Scholar]

4. Jianglin Fan, Yajie Chen, Haizhao Yan, Manabu Niimi, Yanli Wang, Jingyan Liang, Principles and Applications of Rabbit Models for Atherosclerosis Research, J Atheroscler Thromb. 2018 Mar 1; 25(3): 213–220. doi: 10.5551/jat.RV17018. [PubMed][Google Scholar]

5. Kelly S Swanson, Meredith J Mazur, Kapil Vashisht, Laurie A Rund, Jonathan E Beever, Christopher M Counter, Lawrence B Schook, Genomics and clinical medicine: rationale for creating and effectively evaluating animal models, Exp Biol Med (Maywood) (IF: 3.139; Q1). 2004 Oct;229(9):866-75. doi: 10.1177/153537020422900902. [PubMed][Google Scholar]