STAT605 Project Report: Find potential cholesterol genes by using CHTC

Shuyang Chen(schen662), Xiyue Wang(xwang2342), Hongwei Pan(hpan55), Anne Huen Wai Wong(awong43)

12/10/2019

Introduction

The purpose of our project is to find potential cholesterol genes using parallel computing through CHTC. Two different datasets, one is about human genes and one is about mouse genes are analyzed. We calculated biweight midcorrelation between gene "LDLR" and other genes, and defined highly correlated ones as potential cholesterol genes.

Data Description

- Source: ARCHS4: https://amp.pharm.mssm.edu/archs4/download.html
- Size: R human_matrix.rda v7 (4.9GB), R mouse_matrix.rda v7(4.8GB)

These two datasets are gene expressions for 35,000 genes over 200,000 mouse samples and 16,000 human samples in rda format. Expressions of each gene over all the samples are stored as a row vector in a big matrix. During data cleaning process, since liver is the main organ that process cholesterol, we only select those liver samples to analyze and filter out some meta-information variables.

Statistical Computation

Intuitively, we think that genes which have similar distributions among samples may have similar functions. We consider measuring the similarity by calculating the correlation between a commonly known cholesterol gene "LDLR" and other genes, and select the most similar ones as potential cholesterol genes.

Our statistical computation is based on biweight midcorrelation, which is also called bicor. It is a measure of similarities between samples based on median. Compare with other metrics, for example, Pearson correlation which is based on mean, bicor is more robust and less sensitive to outliers.

Formulas to calcultate biweight midcorrelation are list follows, as shown in wikipedia:

$$bicor(x,y) = \sum_{i=1}^{m} \tilde{x_i}\tilde{y_i}$$

where

$$\tilde{x_i} = \frac{(x_i - med(x))w_i^{(x)}}{\sqrt{\sum_{j=1}^{m}[(x_j - med(x))w_j^{(x)}]^2}}, \tilde{y_i} = \frac{(y_i - med(y))w_i^{(y)}}{\sqrt{\sum_{j=1}^{m}[(y_j - med(y))w_j^{(y)}]^2}}$$

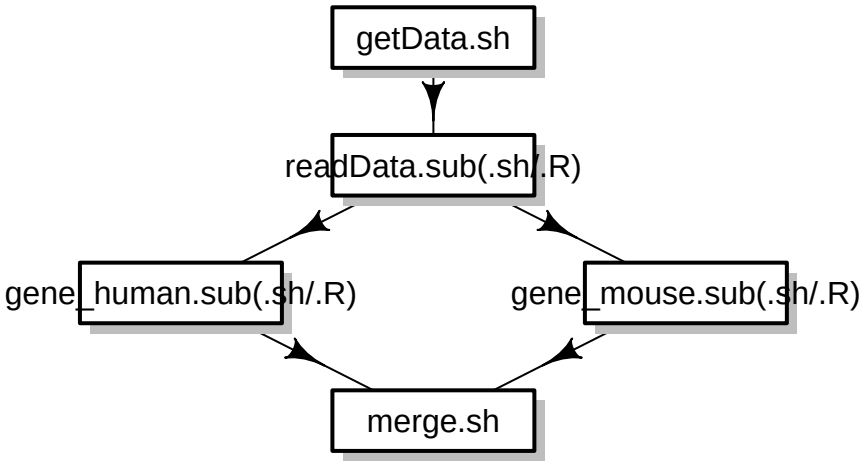
and

$$w_i^{(x)} = (1-u_i^2)^2 I(1-|u_i|), w_i^{(y)} = (1-v_i^2)^2 I(1-|v_i|)$$

$$u_i = \frac{x_i - med(x)}{9med(x)}, v_i = \frac{y_i - med(y)}{9med(y)}$$

We use function `bicorAndPvalue()` in R package `WGCNA` to calculate biweight midcorrelation.

This flowchart shows what we do through CHTC:



Since our original data format is Rda, R, instead of shell command, helps us split every file into ten smaller txt files to do parallel computing through CHTC. After we write out new files, bicorrelation coefficient, p-value between LDLR and other genes are calculated respectively. No matter the coefficient is positive or negative, only the absolute value influence the relevance we want to know.

After parallel computing, we merge and sort to get two final files that contain coefficient and p-value of human and mouse in ascending order. Aim at highly correlated genes, which are potential cholesterol genes, we analyze them out of CHTC.

Finally, we select 10 genes that are highly correlated with LDLR, and make comparison plots.

Results

The top 10 genes correlated with LDLR in human genes:

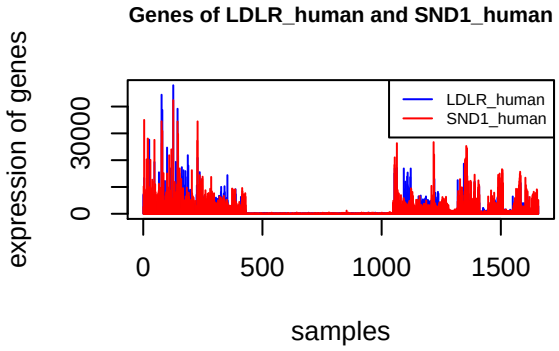
```
##       gene bicorrelation   p-value
## 1     SND1     0.6985117 8.43e-243
## 2      VCP     0.6932201 1.15e-237
```

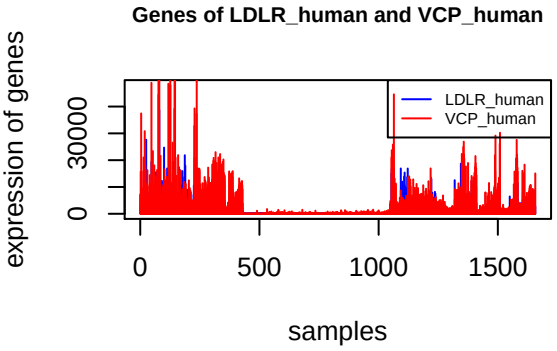
```
## 3     PHB2     0.6928290 2.73e-237
## 4    FKBP4     0.6907002 2.93e-235
## 5   PITRM1     0.6884732 3.74e-233
## 6   AFG3L2     0.6878611 1.41e-232
## 7     IMMT     0.6861711 5.37e-231
## 8     FAF2     0.6859340 8.94e-231
## 9    PSMD3     0.6846405 1.42e-229
## 10 MAPKAP1     0.6843025 2.93e-229
```

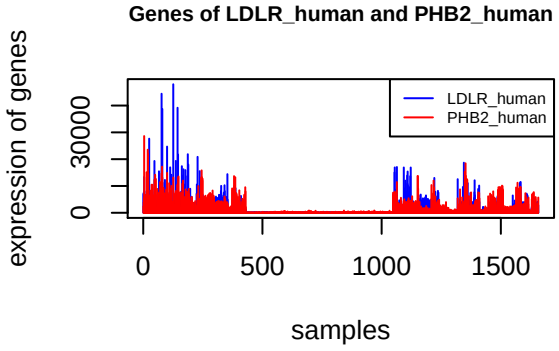
The top 10 genes correlated with LDLR in mouse genes:

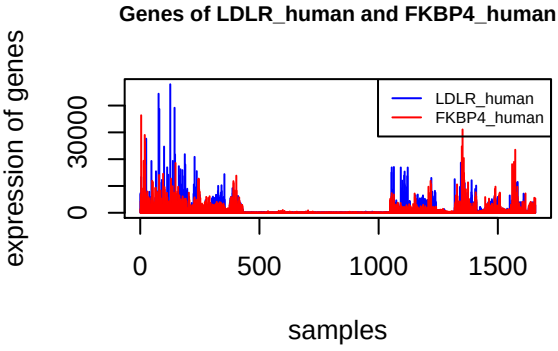
```
##         gene bicorrelation        p-value
## 1      AKAP1     0.8481386 <2.225074e-308
## 2       ACO1     0.8453033 <2.225074e-308
## 3     ZFYVE1     0.8439302 <2.225074e-308
## 4       PEX5     0.8427288 <2.225074e-308
## 5       CDS2     0.8336746 <2.225074e-308
## 6     HECTD3     0.8310280 <2.225074e-308
## 7       VAV2     0.8273318 <2.225074e-308
## 8  D17WSU92E     0.8265241 <2.225074e-308
## 9       RAB5B     0.8263013 <2.225074e-308
## 10     TRAK1     0.8260291 <2.225074e-308
```

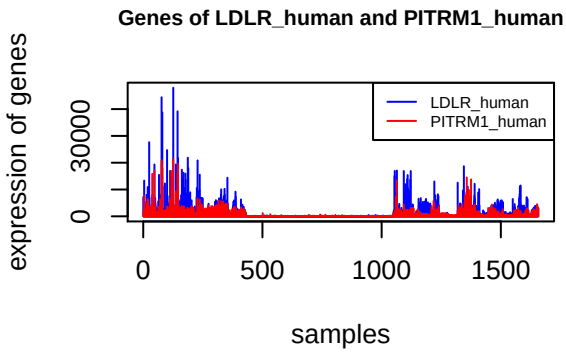
Plots for top 10 genes compared with LDLR in human genes:

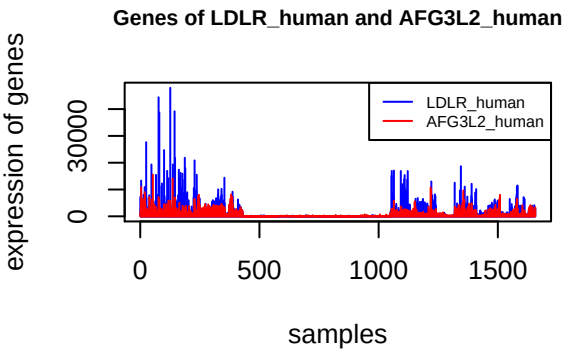


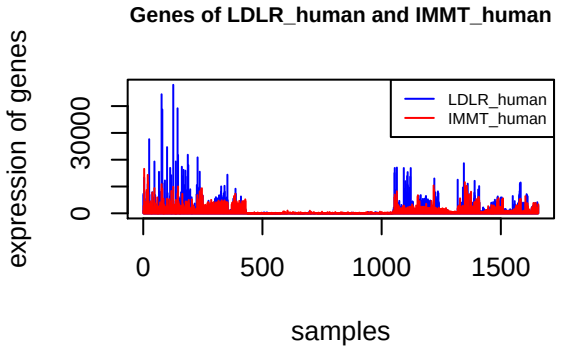


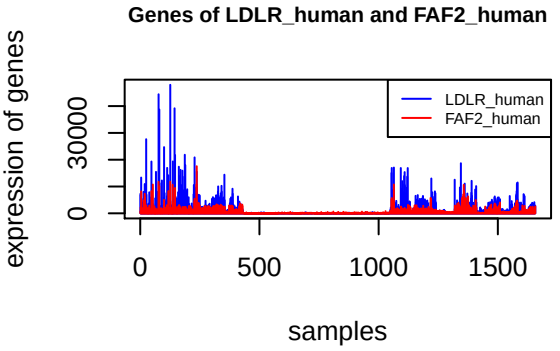


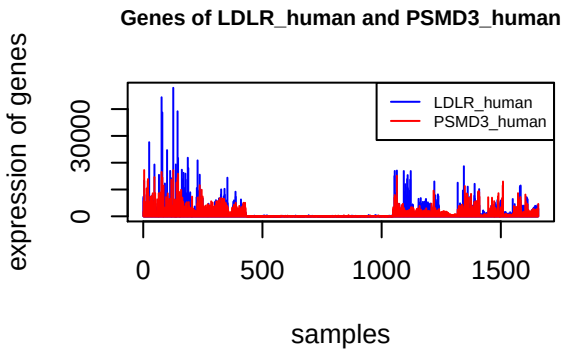


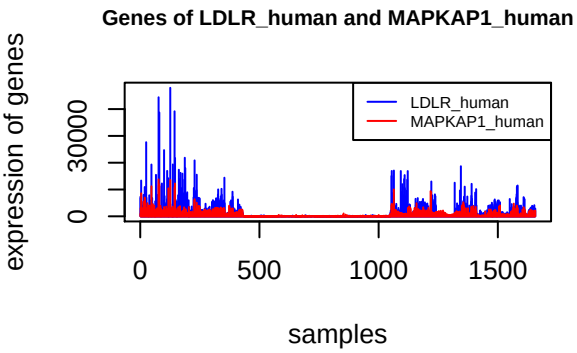

Genes of LDLR_human and PITRM1_human
expression of genes
0
30000
LDLR_human
PITRM1_human
0
500
1000
1500
samples


Genes of LDLR_human and AFG3L2_human
expression of genes
0
30000
LDLR_human
AFG3L2_human
0
500
1000
1500
samples

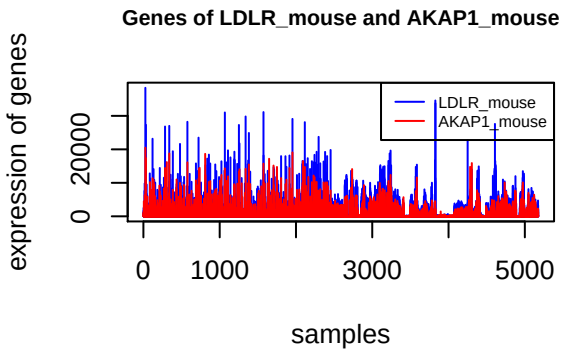

Genes of LDLR_human and IMMT_human
expression of genes
0
30000
LDLR_human
IMMT_human
0
500
1000
1500
samples

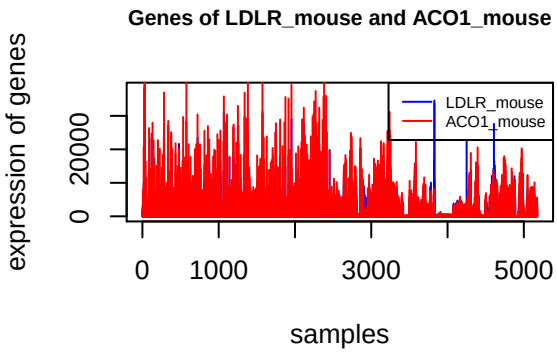

Genes of LDLR_human and FAF2_human
expression of genes
0
30000
LDLR_human
FAF2_human
0
500
1000
1500
samples

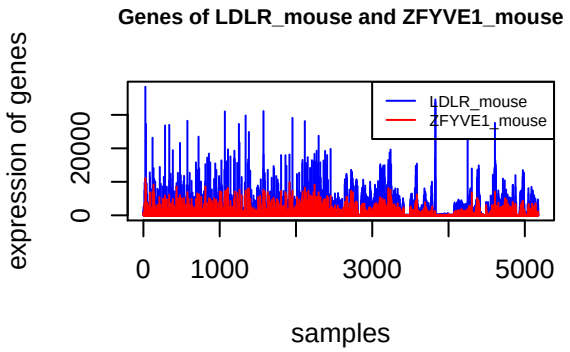


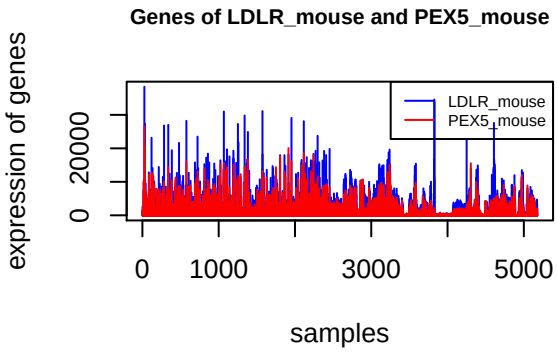


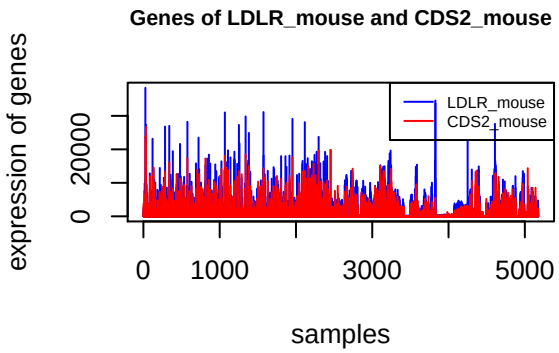
Plots for top 10 genes compared with LDLR in mouse genes:

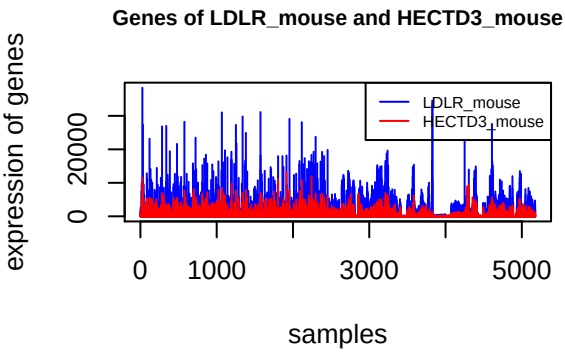

Genes of LDLR_mouse and AKAP1_mouse
expression of genes
0
20000
LDLR_mouse
AKAP1_mouse
0
1000
3000
5000
samples

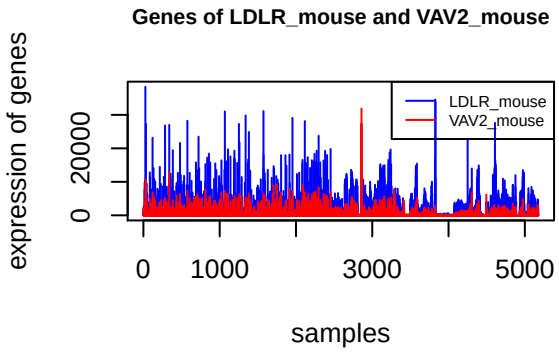

Genes of LDLR_mouse and ACO1_mouse
expression of genes
0
20000
LDLR_mouse
ACO1_mouse
0
1000
3000
5000
samples

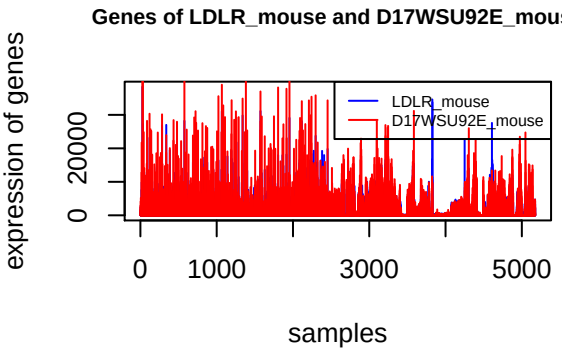

Genes of LDLR_mouse and ZFYVE1_mouse
expression of genes
0
20000
LDLR_mouse
ZFYVE1_mouse
0
1000
3000
5000
samples

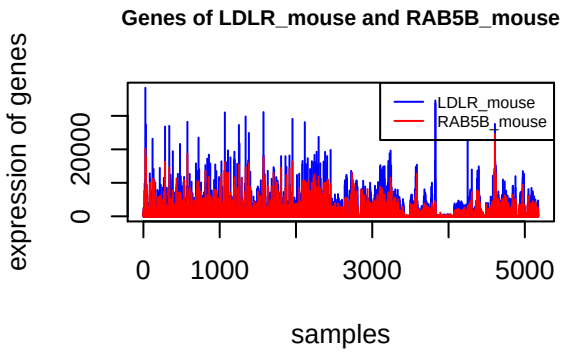

Genes of LDLR_mouse and PEX5_mouse
expression of genes
0
20000
LDLR_mouse
PEX5_mouse
0
1000
3000
5000
samples

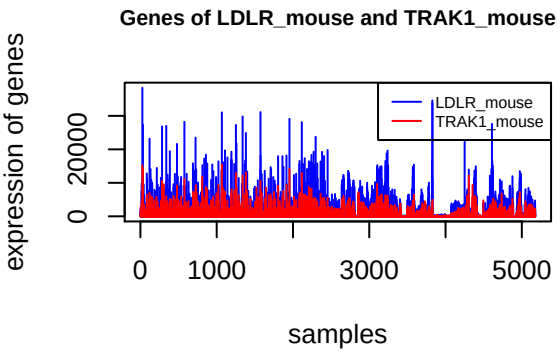

Genes of LDLR_mouse and CDS2_mouse
expression of genes
0
20000
LDLR_mouse
CDS2_mouse
0
1000
3000
5000
samples


Genes of LDLR_mouse and HECTD3_mouse
expression of genes
0
20000
LDLR_mouse
HECTD3_mouse
0
1000
3000
5000
samples


Genes of LDLR_mouse and VAV2_mouse
expression of genes
0
20000
LDLR_mouse
VAV2_mouse
0
1000
3000
5000
samples


Genes of LDLR_mouse and D17WSU92E_mouse
expression of genes
0
20000
LDLR_mouse
D17WSU92E_mouse
0
1000
3000
5000
samples





Descrptions of some genes correlated with LDLR:

VCP: Valosin-containing protein (VCP) is important for the cholesterol-accelerated degradation. [1]

IMMT: Encoded mitochondrial inner membrane protein. And oxidation helps the process of building up cholesterol. [2]

Although p-values and plots seem good, part of the reason is that huge amounts of samples are used in calculation. And due to the lack of biological knowledge, there might be other influence factors that are able to decide the expression of genes but we don't know. It's difficult for us to verify whether there is any actual meaning for our findings.

Conclusion

- There might be probability to narrow relevant genes down by finding highly correlated genes.
- There is obvious difference between the LDLR gene of human and mouse.
- Our results may be used to build gene modules and to make network inferences in the future.

[1] Ngee Kiat Chua, Nicola A. Scott, Andrew J. Brown; Valosin-containing protein mediates the ERAD of squalene monooxygenase and its cholesterol-responsive degron. Biochem J 30 September 2019; 476 (18): 2545–2560.

[2] Chinese Journal of Cell Biology:1-8[2019-12-10]