Omics Central

Amrit Singh

2020-02-25

Contents

Chapter 1

Rationale

This project was developed in order to create a resource warehouse for researchers analyzing omics datasets of various types such as transcriptomics, proteomcs, metabolomics. I expect this resource to grow as others contribute to it. Think of it as an awesome-resource github repo but in a bookdown format. However, since this book is meant as documentation to the omics central web application, adding new methods will require pull requests to the omics central web app repos (omics-central-frontend, omics-central-backend and omics-central-docker) and bookdown repos (omics-central-learn and omics-central-contribute omics-central-learn).

Site under development…

Chapter 2

Introduction

\TODO

Chapter 3

Data-types

3.1 Microarrays

3.2 RNA sequencing

3.3 Nanostring

3.4 Biocrates

3.5 Multiple Reaction Monitoring

Chapter 4

Exploratory Data Analysis

//TODO insert video of EDA using Omics Central here

4.1 Principal Component Analysis

4.1.1 Method

4.1.1.1 What is PCA?

- method to turn a dataset with correlated variables into another dataset with linearly uncorrelated variables called principal components (PCs).

4.1.1.2 Why is PCA useful?

- The first few PCs capture most of the variability in the data.
- PCA can be used to visualize clustering patterns (samples or variables) in the data, determine relationships between samples (see Principal Component plot), between variables (see Correlation circle), between samples and variables (see Biplot).
- PCA is also useful in determining the influece of covariates, both techincal (*e.g.* batch effects) or biological (*e.g.* sex).

4.1.1.3 What is a principal component (PC)?

- a PC is a weighted average of the original predictors, $\mathbf{PC}_i = \mathbf{X}\mathbf{v}_i$, where $\mathbf{X}$ is a centered matrix and $i=1,\ldots,n$.

4.1.1.4 What do the vector of weights $\mathbf{v}_i$ do?

- v_i maximizes the variance; $\mathbf{X^T X}$ and are called eigenvectors, weights or loadings.

4.1.1.5 How do I compute the vector of weights, v_i?

- apply a factorization method called singular value decomposition (SVD). SVD decomposes a matrix X into a product of 3 matrices, $\mathbf{UDV^T}$; $\mathbf{X}_{np}$ = $\mathbf{U}_{nxp}$ x $\mathbf{D}$~pxp~ x $\mathbf{V^T}_{pp}$ or $\mathbf{X^T X} = \mathbf{VD^2V^T}$.
- The columns of $\mathbf{V}$ are the weights/loadings for each principal component.
- $\mathbf{D}$ is a diagnoal matrix where entry $\mathbf{D}_{i,i}$ is the standard deviation of the *ith* principal component (PC).
- Only the first k PCs are needed to capture the majority of the variation in the high dimensional dataset ($n << p$ and $k << p$); $\mathbf{X}_{nk}$ = $\mathbf{U}_{nxk}$ x $\mathbf{D}_{pxk}$ x $\mathbf{V^T}_{nk}$ such that $\mathbf{X}_{nk} \approx \mathbf{X}_{np}$.

4.1.1.6 Why scale the data before applying PCA?

- The clinical variables are on different unit scales (*e.g.* Age (years) *vs.* Ejection fraction (%)). Scaling makes the mean of each variable zero and the standard deviation one.

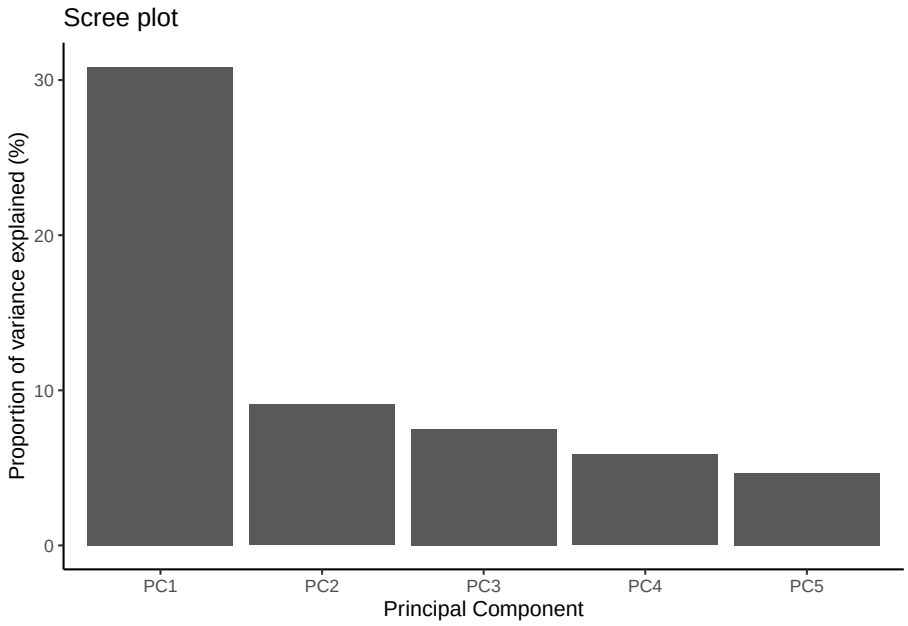
References

1. page 64-66 from ESL: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print10.pdf
2. Wikipedia: https://en.wikipedia.org/wiki/Principal_component_analysis

4.1.2 Visualizations

4.1.2.1 Scree plot

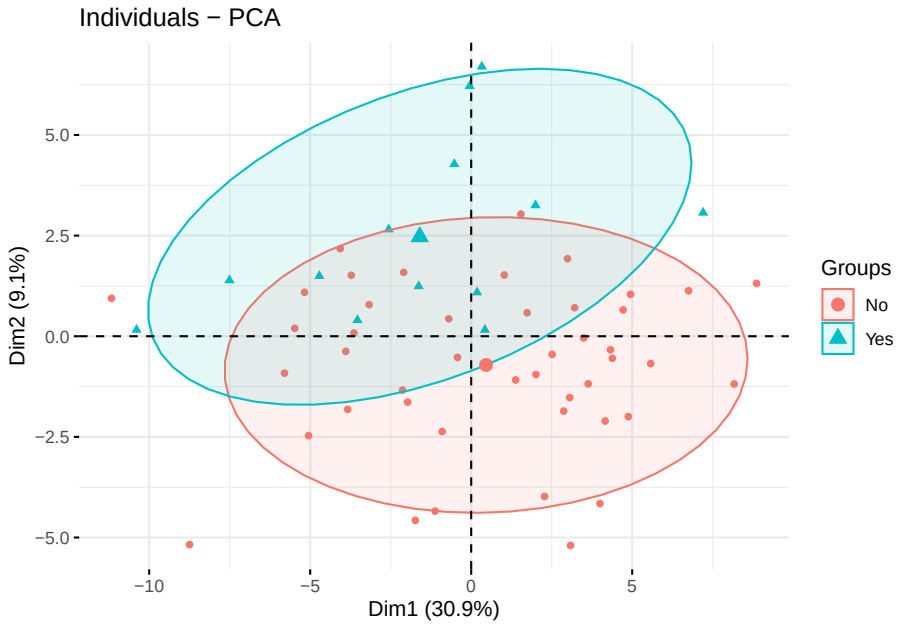
- determine the proportion of variation explained by each principal component.



The barplot depicts the proportion of variation that is captured by the first five PCs; the first PC captures ~30.9% of the variability in the dataset consisting of 65 variables.

4.1.2.2 Component plot

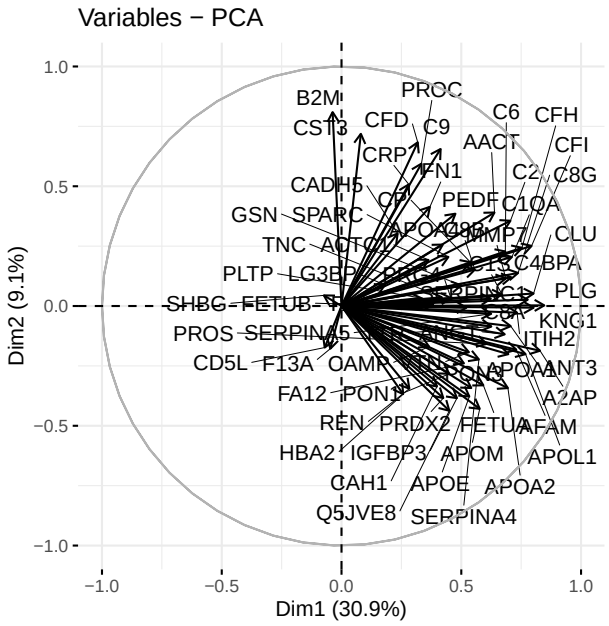
- visualize the clustering of the samples and identify any clustering with respect to covariates of interest.



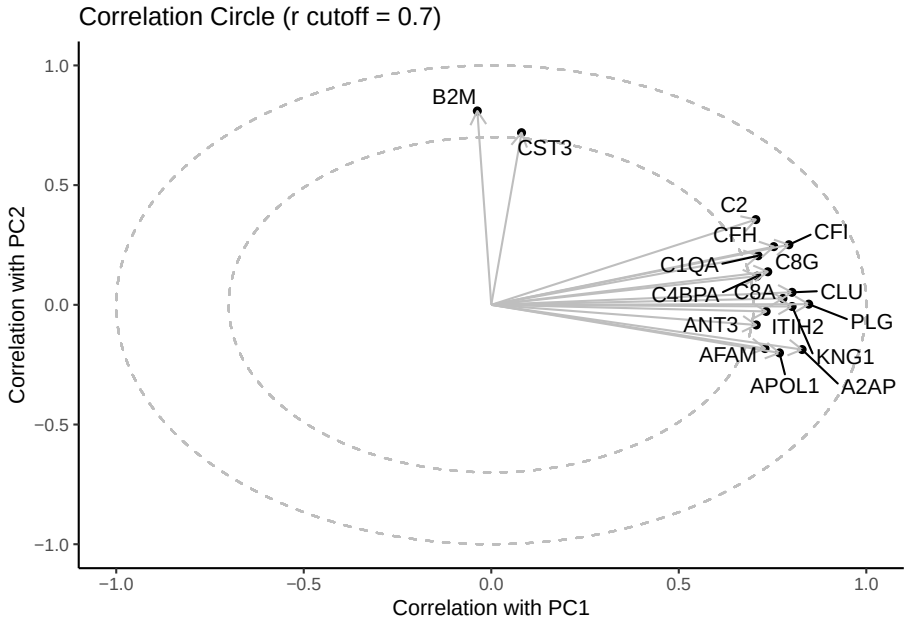
The scatter plot above is a 2D depiction of a 65 (# of clinical variables) dimensional dataset. PC1 and PC2 together capture 40% of the variability in the clinical dataset. Some separation between the groups of interest can be observed.

4.1.2.3 Correlation Circle

- determine relationship between variables (based on the correlation between each variable and PCs).
- the angle (θ) between two vectors determines the correlation between the two variables:
- θ=0: postive correlation (corr=1)
- 0<θ<90: postive correlation
- θ=90: zero correlation
- 90<θ<180: negative correlation
- θ=180: negative correlation (corr=-1)



4.1.2.4 Correlation Circle (with a cut-off)



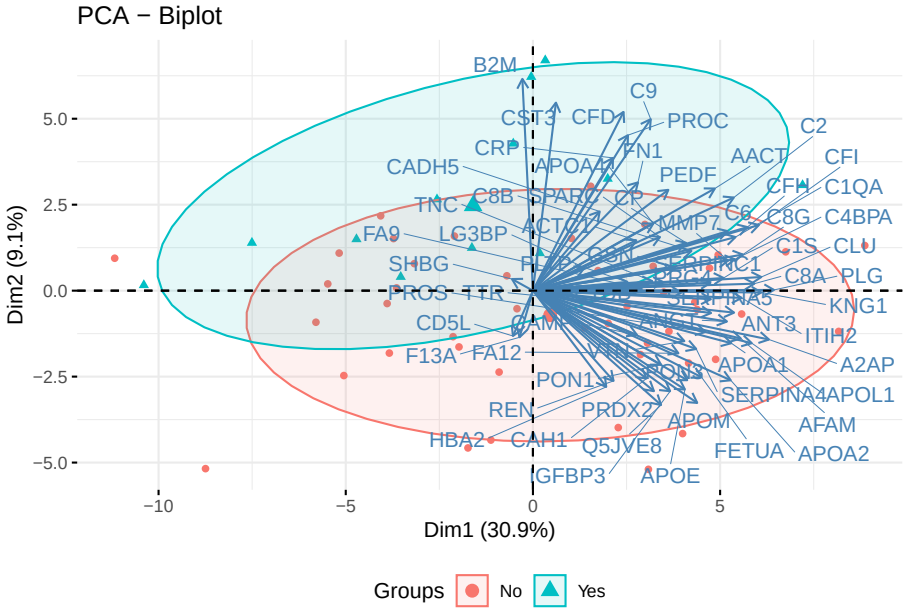
The above plot only displays the variables if they have a correlation

greater than 0.5 with either PC1 or PC2. Ischemia and Statins are positively correlated suggesting that patients with ischemia are likely to be on statins. BNP (Brain Natriuretic Peptide) is positively correlated with age and negatively correlated with Heart Rate.

References
1. Figure 1 from BioData Mining volume 5, Article number: 19 (2012)
2. plotVar(): mixOmics R-library 3. fviz_pca_var(): factoextra R-library

4.1.2.5 Biplot

- superimpose the principal components with loadings vectors.



Each arrow can be thought of as an axis. For example, BNP points to the left which means that patients on the left (PC1 < 0) have lower BNP levels than patients on the right (PC1 > 0). Patients at the center (PC=1) have an average BNP level. Note that this aligns well with the hospitalization status; *ie.* patients on the left are more likely to be hospitalized as compared to patients on the right.

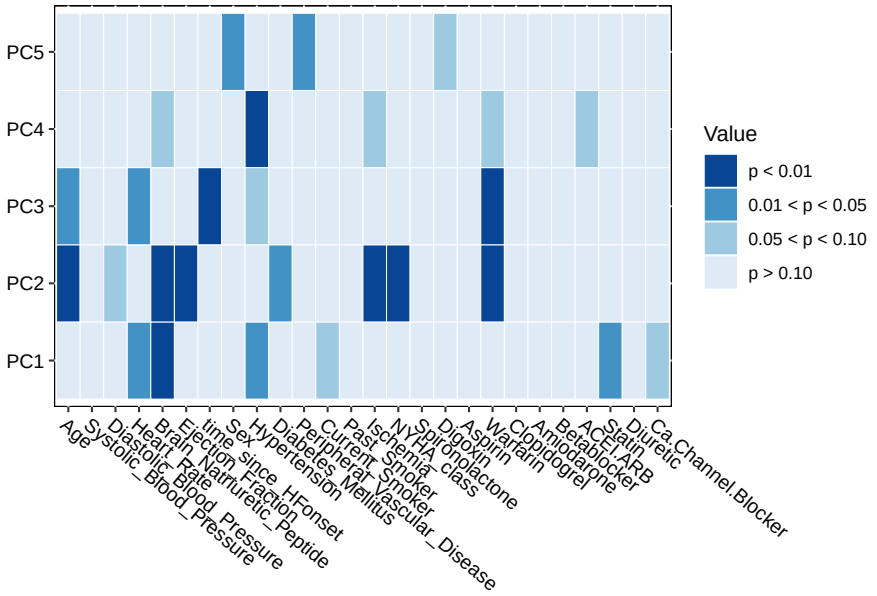
References
1. ggbiplot(): https://github.com/vqv/ggbiplot
2. Biplot: https://stackoverflow.com/questions/6578355/plotting-pca-biplot-with-ggplot2
3. biplot(): K. R. Gabriel (1971). The biplot graphical display of matrices with application to principal component analysis. Biometrika, 58, 453–467. doi:

10.2307/2334381.
4. fviz_pca_biplot(): http://www.sthda.com/english/wiki/fviz-pca-quick-principal-component-analysis-data-visualization-r-software-and-data-mining

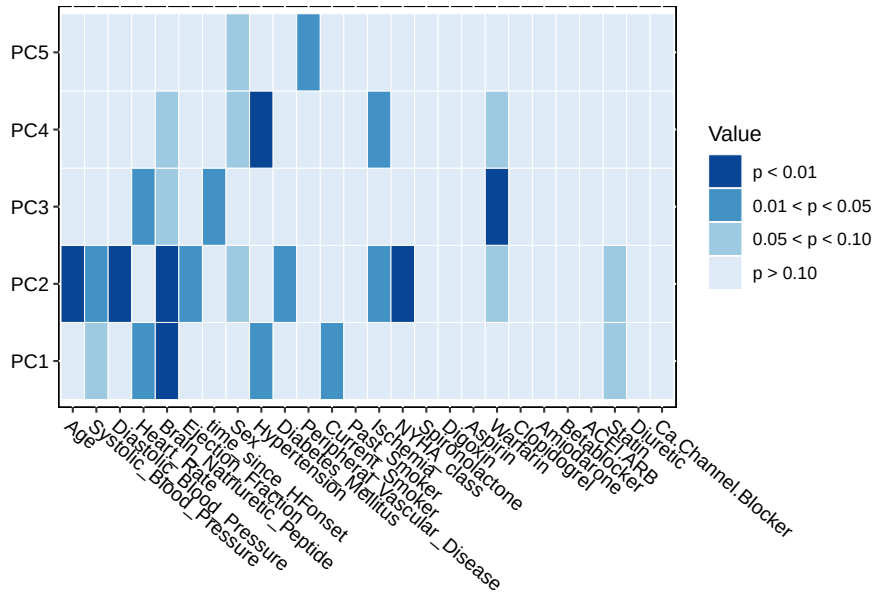
4.1.2.6 Are the major sources of variation in the proteomics dataset related to any demographics variables?

- this is often answers by correlating the PCs with demographics variables such as batch or disease of interest.

4.1.2.6.1 Test the Pearson correlation between PCs and demographic variables



4.1.2.6.2 Test the Spearman correlation between PCs and demographic variables



The association between PC1 and BNP has a p-value of < 0.01 which supports the Biplot in which BNP was parallel to PC1 (x-axis).

WARNING: This is only to be used for exploratory purposes and not for inference since spurious correlations may arise.

References 1. BioData Mining volume 5, Article number: 19 (2012)
2. PH525x series: http://genomicsclass.github.io/book/ 3. mixOmics: https://mixomicsteam.github.io/Bookdown 4. EDA in R: https://bookdown.org/rdpeng/exdata/

Chapter 5

Batch Correction

5.1 ComBat

5.2 Surrogate Variable Analysis

5.3 Model adjustment

5.4 References

1. Batch effect simluations: http://jtleek.com/svaseq/simulateData.html
2. Surrogate Variable Analysis: https://bioconductor.org/packages/release/bioc/vignettes/sva/inst/doc/sva.pdf

Chapter 6

Differential Expression Analysis

//TODO insert video of performing differential expression analysis using Omics Central here

6.1 Methods

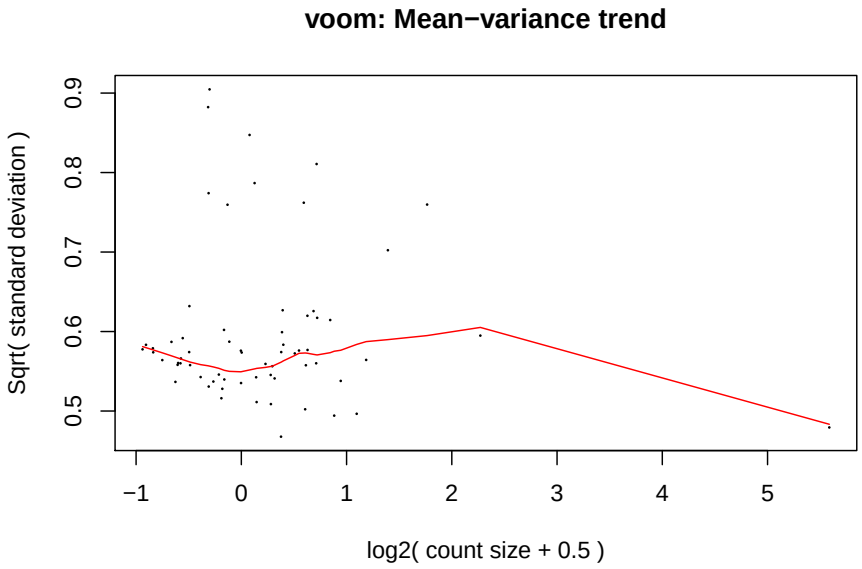
6.1.1 Ordinary Least Squares

6.1.2 LInear Models for MicroArrays and RNA-Seq

6.1.2.1 LIMMA

6.1.2.2 Robust LIMMA

6.1.2.3 LIMMA VOOM (adjusts for heteroscdasticity)



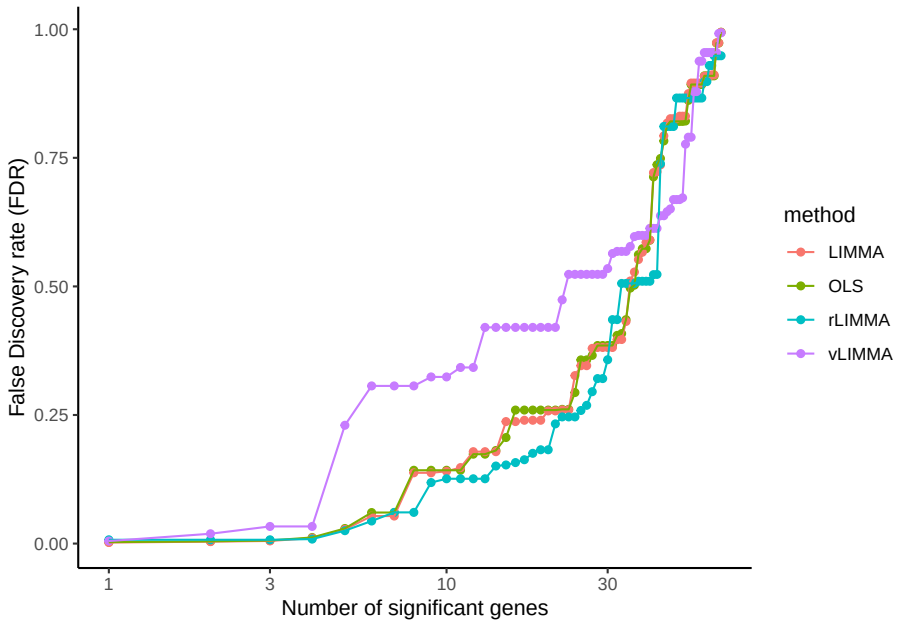
6.1.3 Significance Analysis for Microarrays (SAM)

6.1.4 cell-specific Analysis for Microarrays (csSAM)

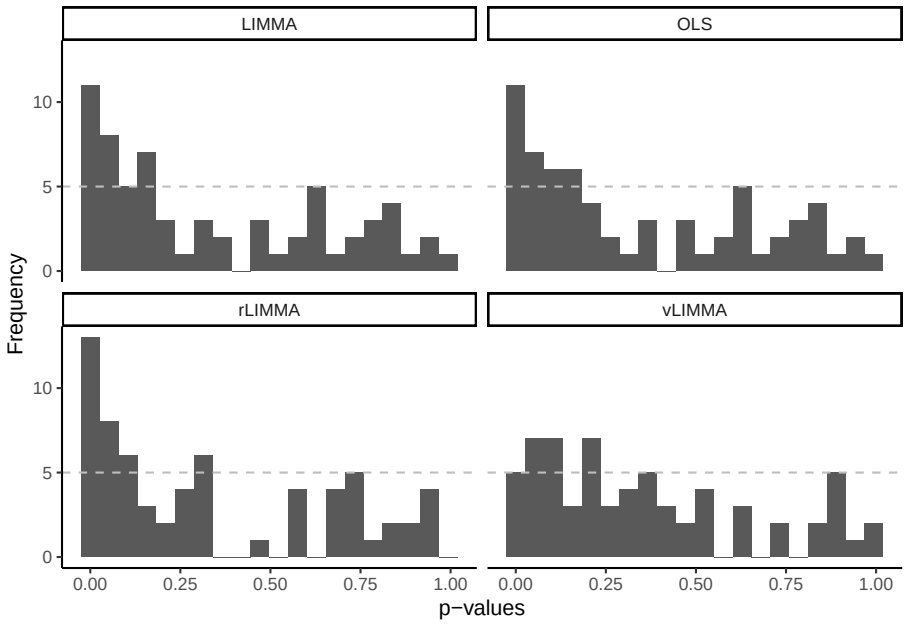
```
## [1] TRUE
```

6.2 Visualizations

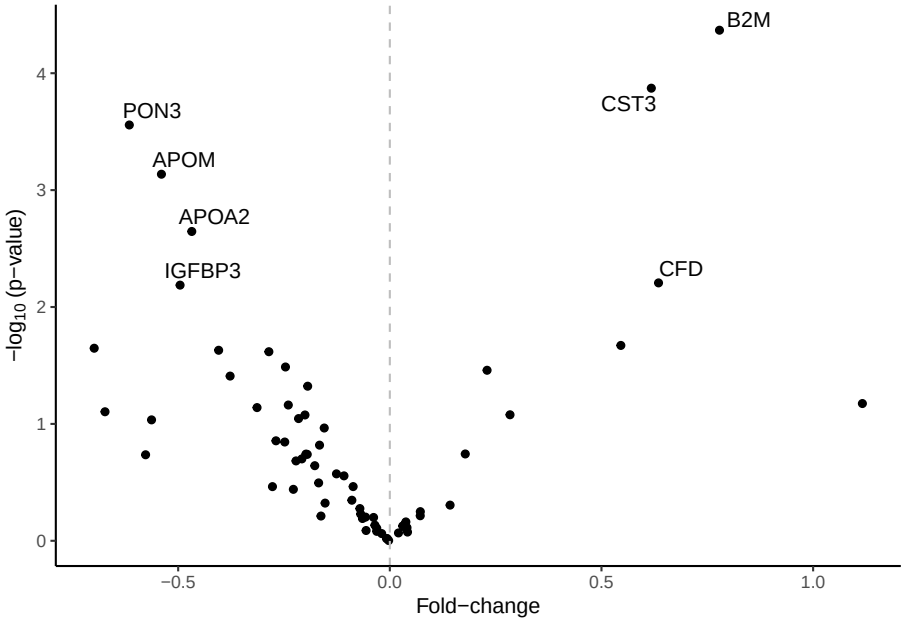
6.2.1 Number of differentially expressed genes



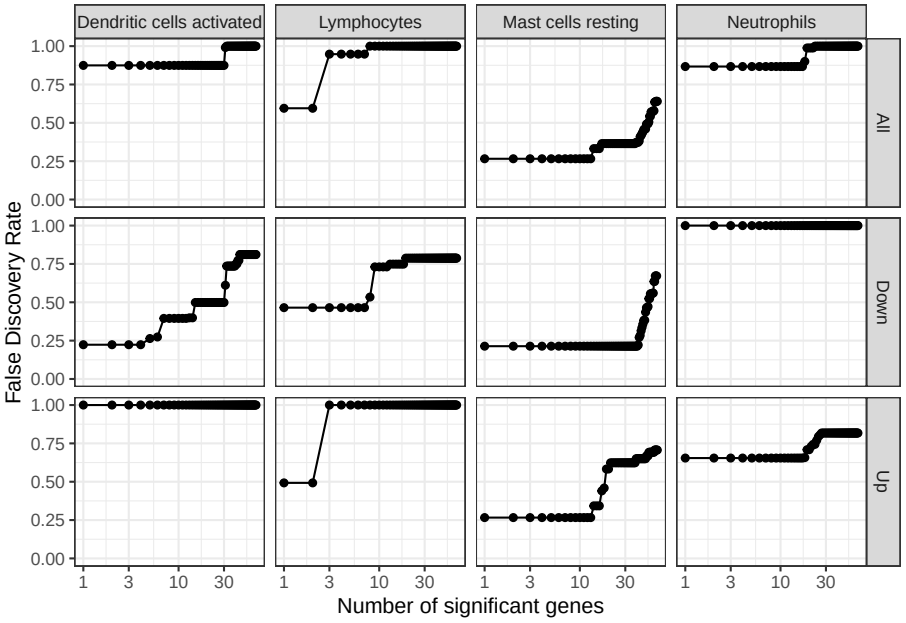
6.2.2 P-value histograms



6.2.3 MA plot


B2M
CST3
PON3
APOM
APOA2
IGFBP3
CFD
−log10 (p−value)
Fold−change
4
3
2
1
0
−0.5
0.0
0.5
1.0

6.2.4 csSAM


Dendritic cells activated
Lymphocytes
Mast cells resting
Neutrophils
All
Down
Up
False Discovery Rate
1.00
0.75
0.50
0.25
0.00
1
3
10
30
Number of significant genes

Chapter 7

Network Analysis

7.1 DINGO

7.2 WGCNA

7.3 PANDA

7.4 BioNetStat

Chapter 8

Data Integration

8.1 Supervised

8.1.1 DIABLO (SGCCDA)

8.1.2 Ensemble of glmnet classifiers

8.1.3 DIABLO2 (sMB-PLSDA)

8.2 References

1. caret: https://topepo.github.io/caret/index.html

8.3 Unsupervised

8.3.1 PANDA

8.3.2 MOFA

8.3.3 JIVE

8.3.4 SNF

Chapter 9

Biological Enrichment

9.1 Enrichr

9.2 SEAR

9.2.1 hypergeometric tests

9.2.1.1 hypergeometric probabilities

- The sample space consists of a total of n genes, out of which m genes belong to Pathway A. Select k genes at random (without replacement). What is the probability that i of the selected genes belong to Pathway A.
- parameters include:
- n: total number of genes observed
- m: number of genes in Pathway A
- k: genes selected at random
- i: number of selected genes that belong to Pathway A

Size of sample space (Ω) = $\binom{n}{k}$: all ways to draw k genes from n genes Event of interest: # of ways to get i genes from Pathway A after drawing k genes = (# of ways to select i genes from Pathway A from a total of m genes in Pathway A, $\binom{m}{i}$) x (# of ways to get k-i from the remaining n-m genes not in Pathway A, $\binom{n-m}{k-i}$)

9.2.1.2 References

1. Probability - The Science of Uncertainty and Data

2. Falcon S., Gentleman R. (2008) Hypergeometric Testing Used for Gene Set Enrichment Analysis. In: Bioconductor Case Studies. Use R!. Springer, New York, NY

9.3 CAMERA

9.4 Network-based Gene Set Analysis

Chapter 10

Literature Mining