Topics in Statistical Learning

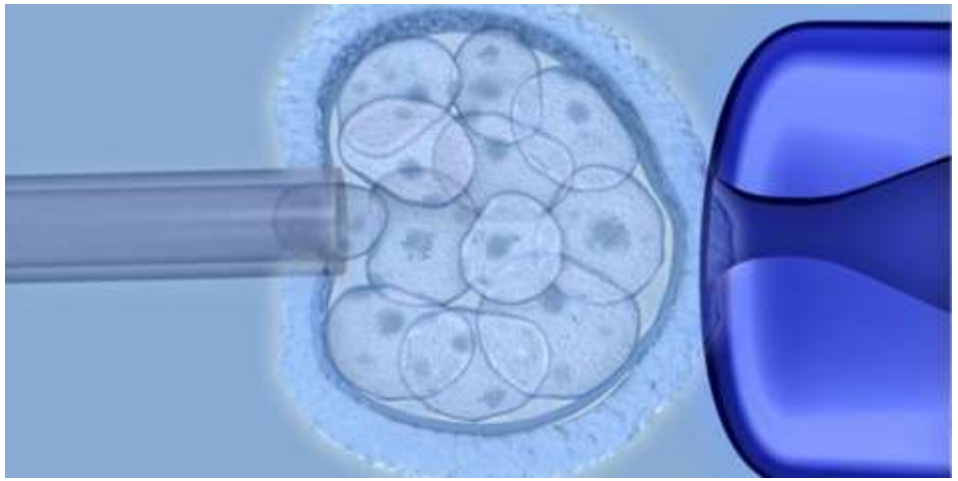
Analysis of transcriptomic differences between iPSC and ESC

Luca Fusar Bassini Alessandro Marincioni Andrea Ghigi

What are ESC and iPSC?

Embryonic stem cells (ESC)

- Derive from the first cell divisions of the zygote.
- Can differentiate into any type of cell of the adult body.

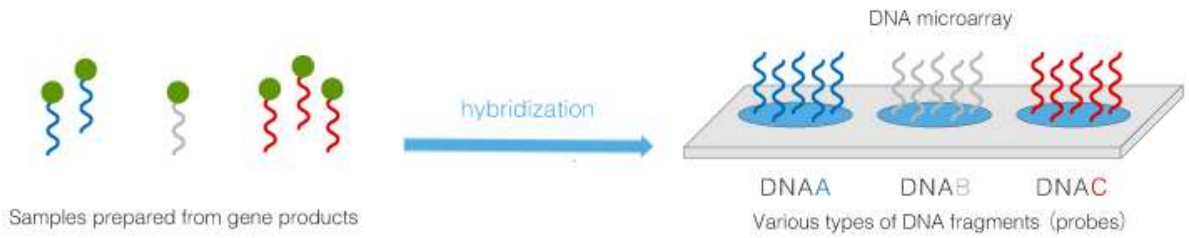


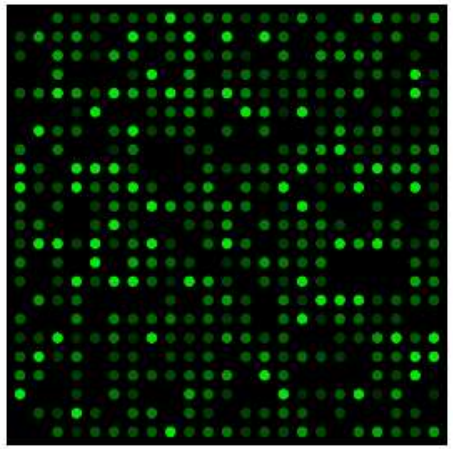
Induced pluripotent stem cells (iPSC)

- Are obtained by artificially reprogramming mature cells.
- Retain some differences in gene expression when compared to ESC.

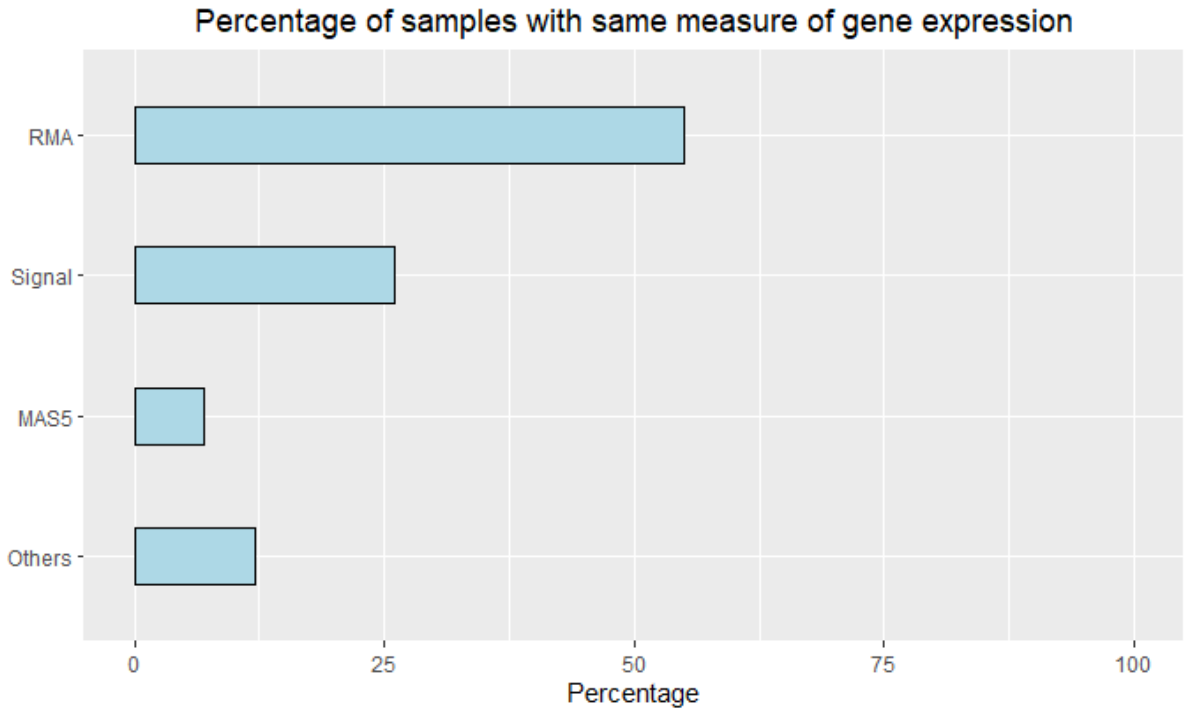
Data collection from Gene Expression Omnibus (GEO)

In every spot of a microarray the intensity of the signal is proportional to the gene expression of a gene

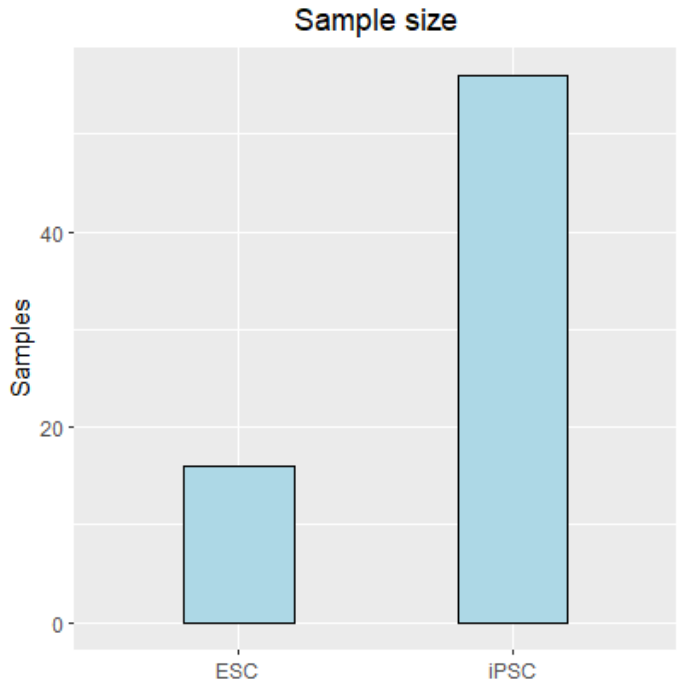


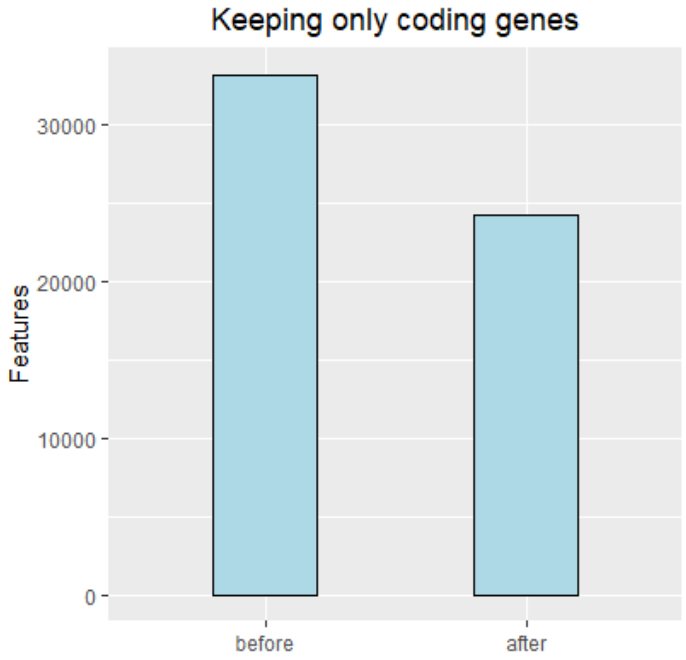


We identified 130 samples, but gene expression was measured with different algorithms.



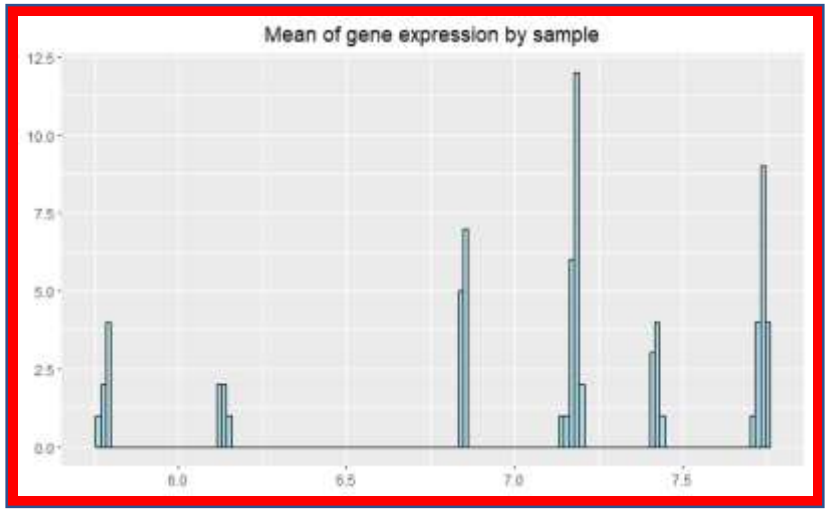
An overview on the dataset

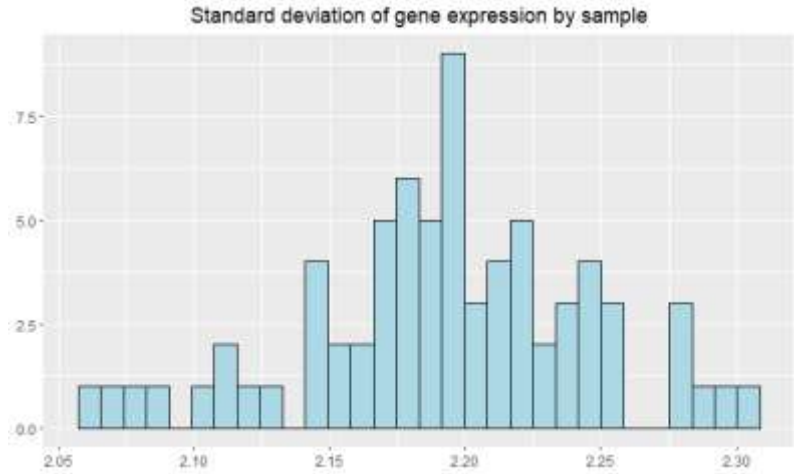


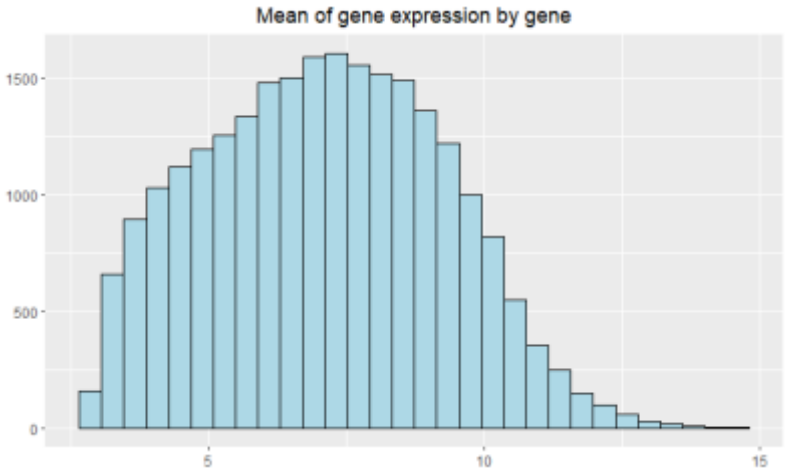


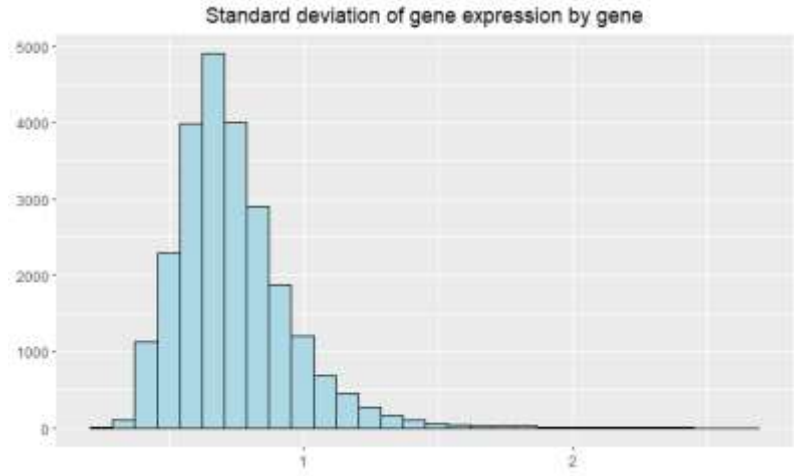
Preliminary analysis on the dataset

The distribution is not normal at all!



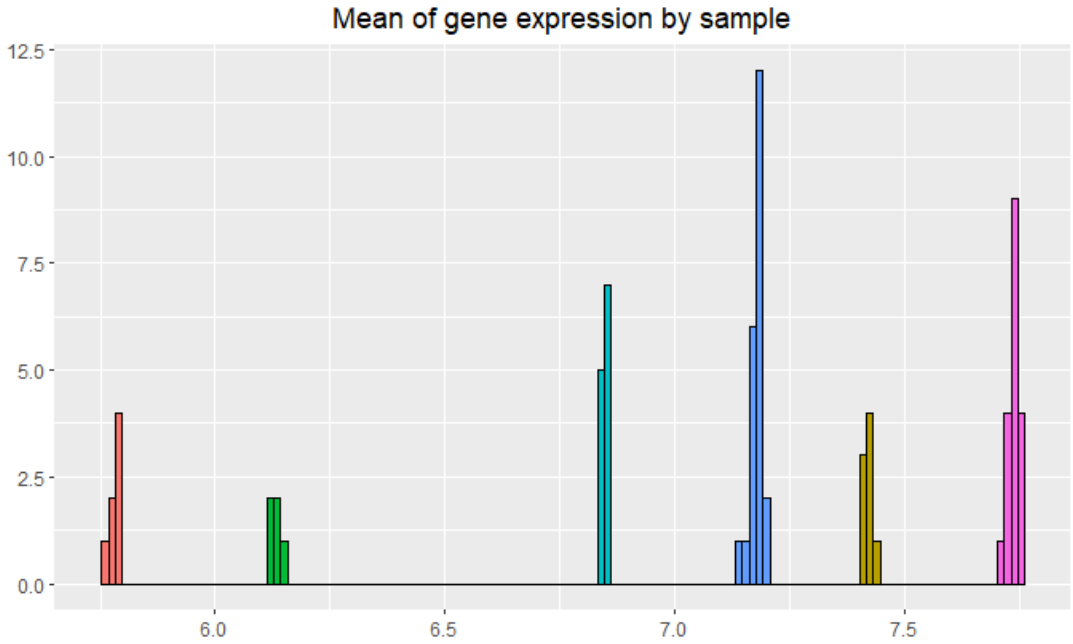




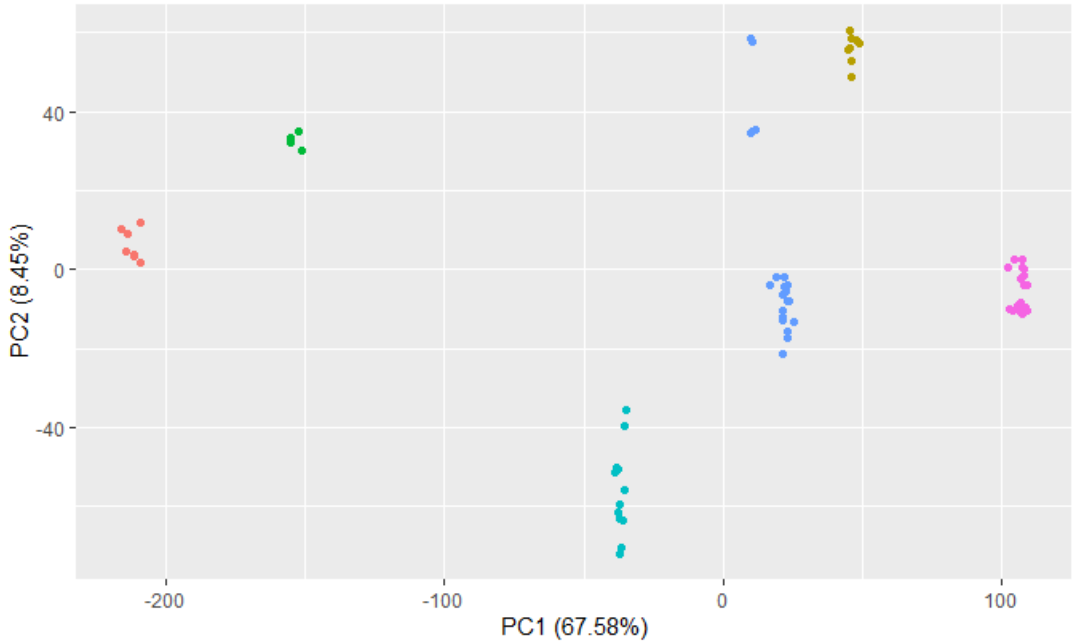


Preliminary analysis on the dataset

Histogram coloured by experiment confirms a strong experiment-effect

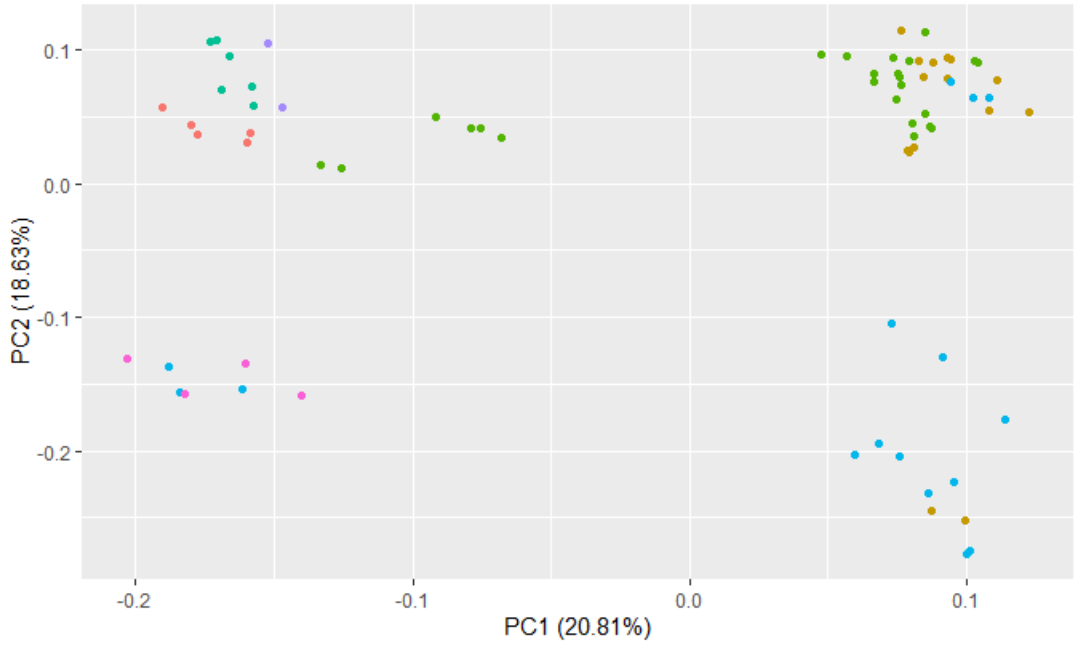


PCA also confirms that most of the variability is explained by experiment

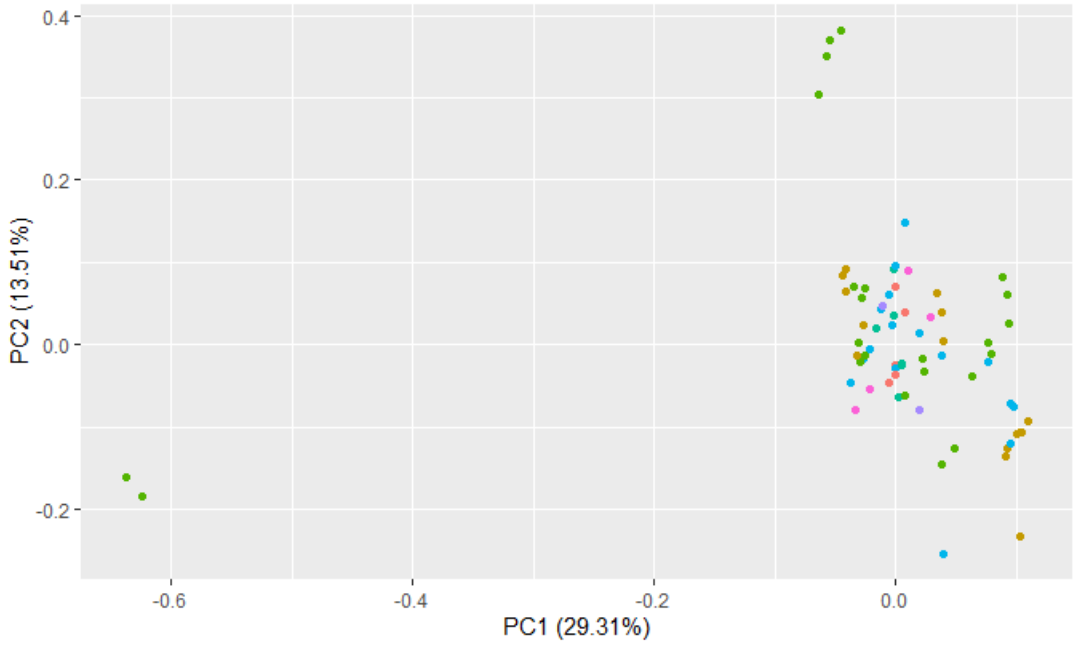


Scaling by sample or by experiment?

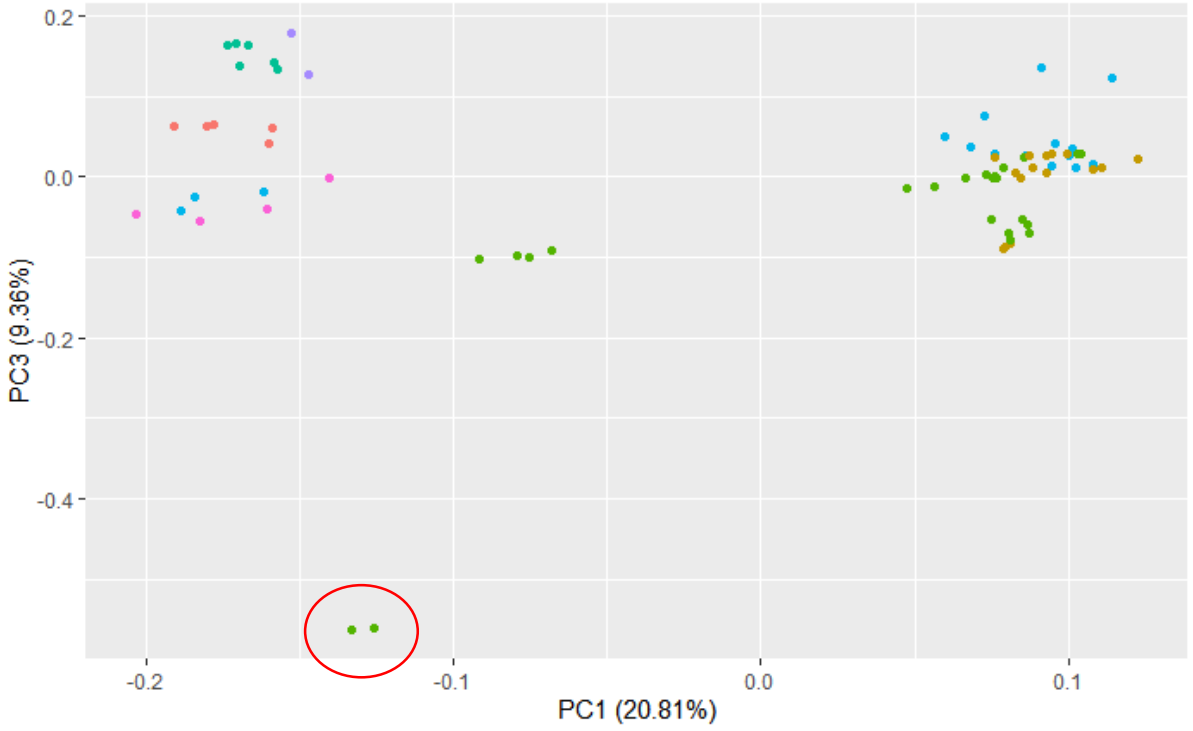
PCA after scaling with Z-score normalization



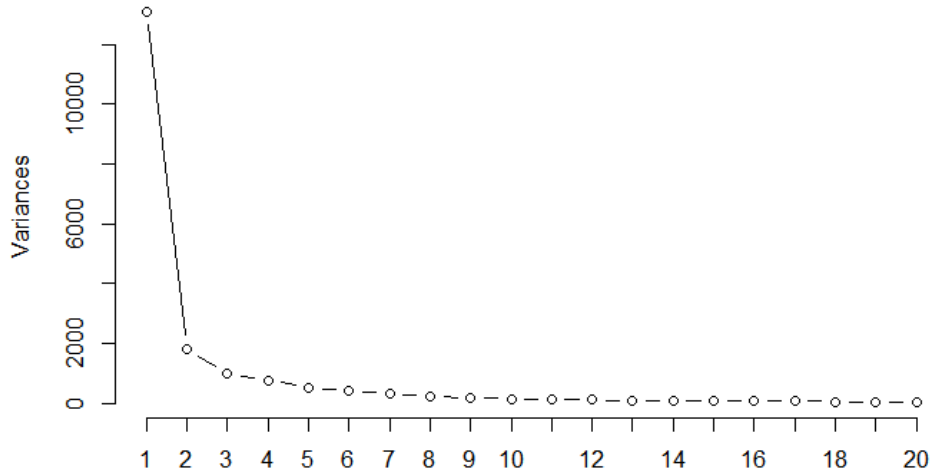
PCA after scaling by experiment using ANOVA



Managing outliers



PC1-PC3 plot of the dataset

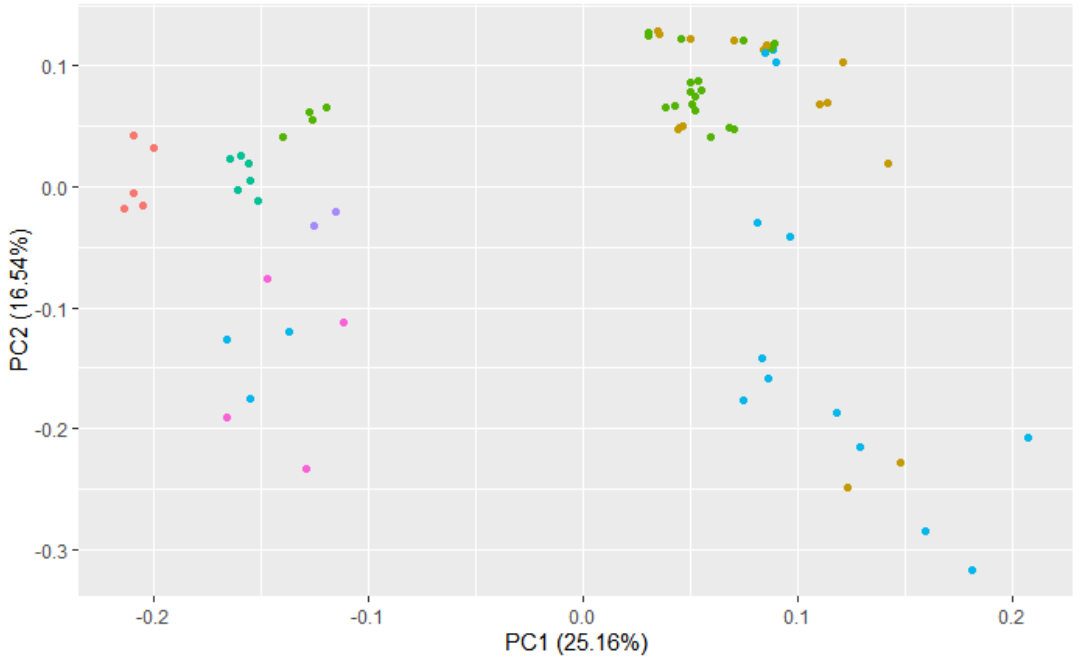


Scree plot

Reducing the number of features with GO

Gene/product	Gene/product name	Organism	PANTHER family	Type	Source	Synonyms
FBN2	Fibrillin-2	Homo sapiens	fibrillin-related pthr24039	protein	UniProtKB	
FBN1	Fibrillin-1	Homo sapiens	fibrillin-related pthr24039	protein	UniProtKB	FBN
MSX2	Homeobox protein MSX-2	Homo sapiens	homeobox protein msx pthr24338	protein	UniProtKB	HOX8
PTCH1	Protein patched homolog 1	Homo sapiens	protein patched pthr46022	protein	UniProtKB	PTCH
TCF15	Transcription factor 15	Homo sapiens	basic helix-loop-helix transcription factor, twist pthr23349	protein	UniProtKB	BHLHA40 BHLHEC2
WNT9A	Protein Wnt-9a	Homo sapiens	wnt related pthr12027	protein	UniProtKB	WNT14
GRSF1	G-rich sequence factor 1	Homo sapiens	heterogeneous nuclear ribonucleoprotein-related pthr13976	protein	UniProtKB	
LIF	Leukemia inhibitory factor	Homo sapiens	leukemia inhibitory factor pthr10633	protein	UniProtKB	HILDA
BPTF	Nucleosome-remodeling factor subunit BPTF	Homo sapiens	nucleosome-remodeling factor subunit bptf pthr45975	protein	UniProtKB	FAC1 FALZ
EFEMP1	EGF-containing fibulin-like extracellular matrix protein 1	Homo sapiens	fibrillin-related pthr24039	protein	UniProtKB	FBLN3 FBNL
ATP8A2	Phospholipid-transporting ATPase IB	Homo sapiens	probable phospholipid-transporting atpase pthr24092	protein	UniProtKB	ATPIB
PAX5	Paired box protein Pax-5	Homo sapiens	paired box protein pax-6-related-related pthr45636	protein	UniProtKB	
ID3	DNA-binding protein inhibitor ID-3	Homo sapiens	dna-binding protein inhibitor pthr11723	protein	UniProtKB	1R21 BHLHB25

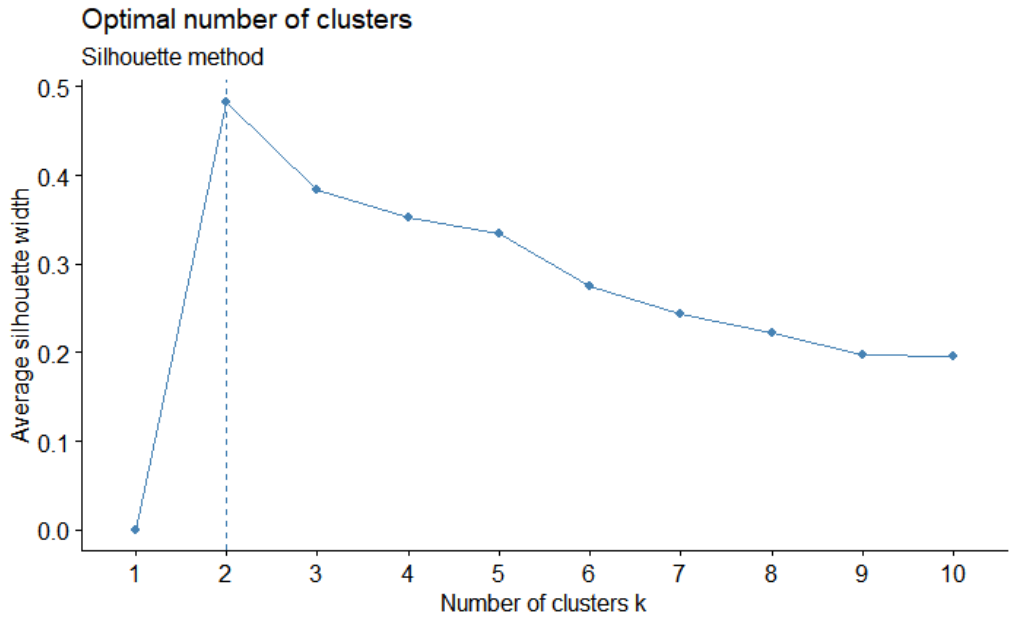
Example of Gene Ontology query result

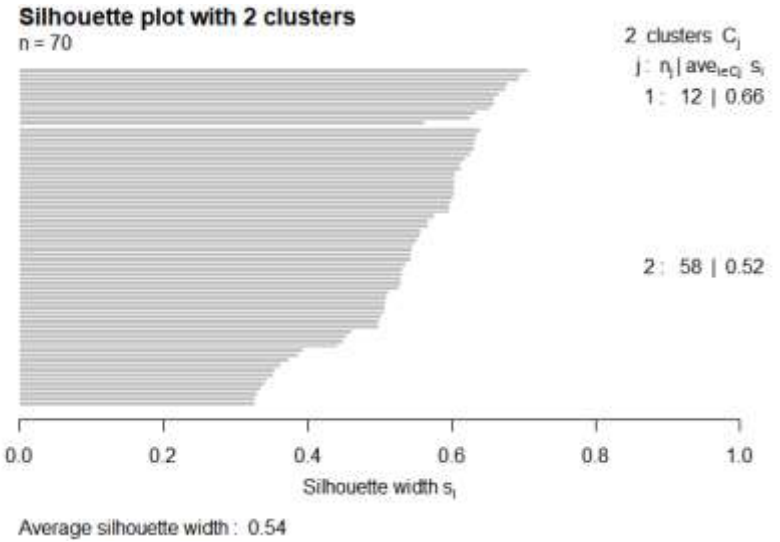


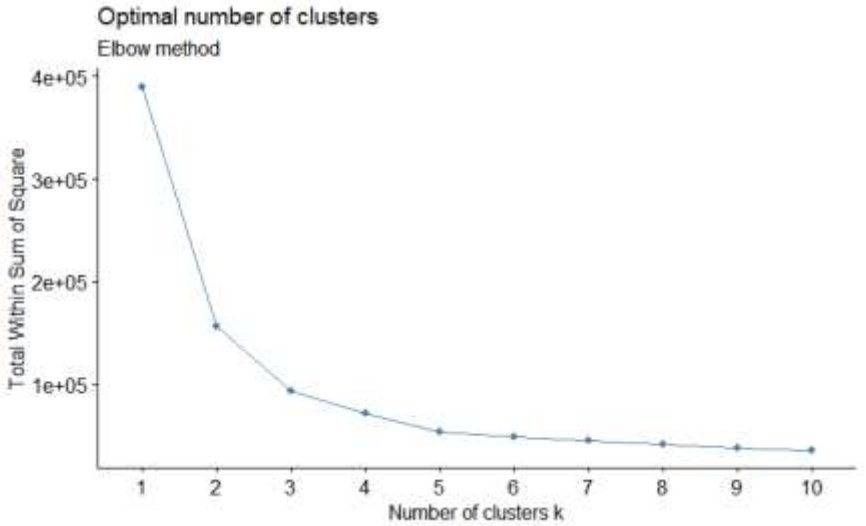
Reduced dataset

Choosing the optimal number of clusters

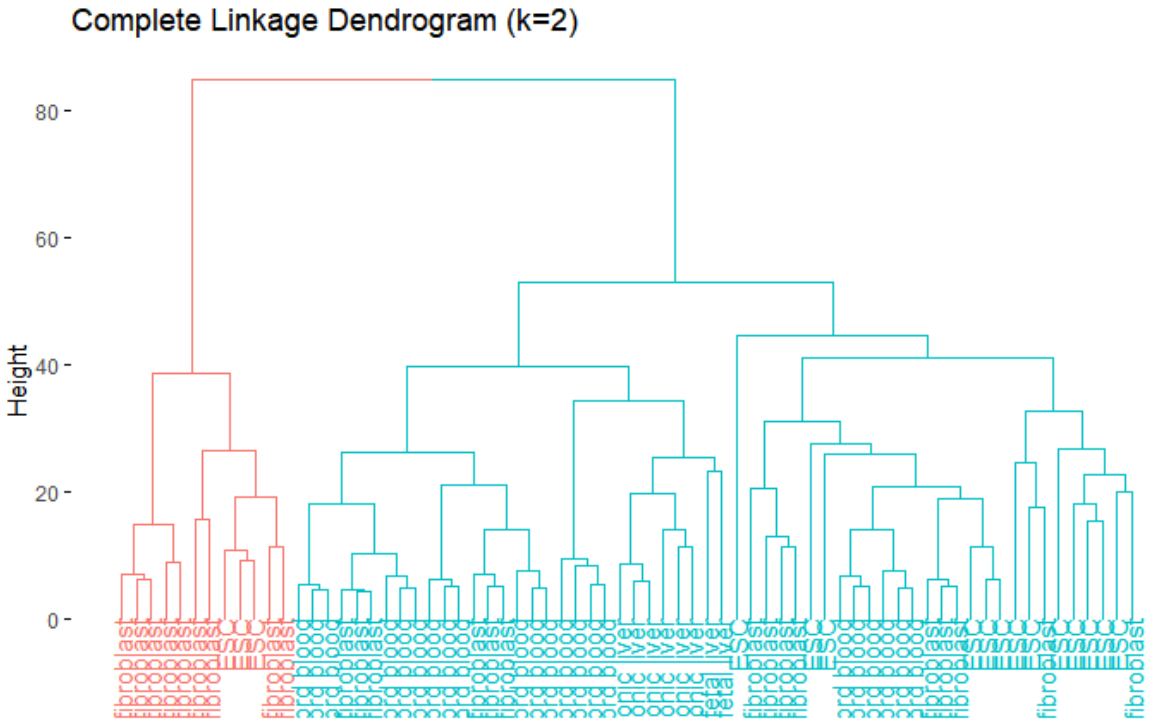
We use the Silhouette method to choose the number, and then compare the results with the corresponding silhouette and wss plot.

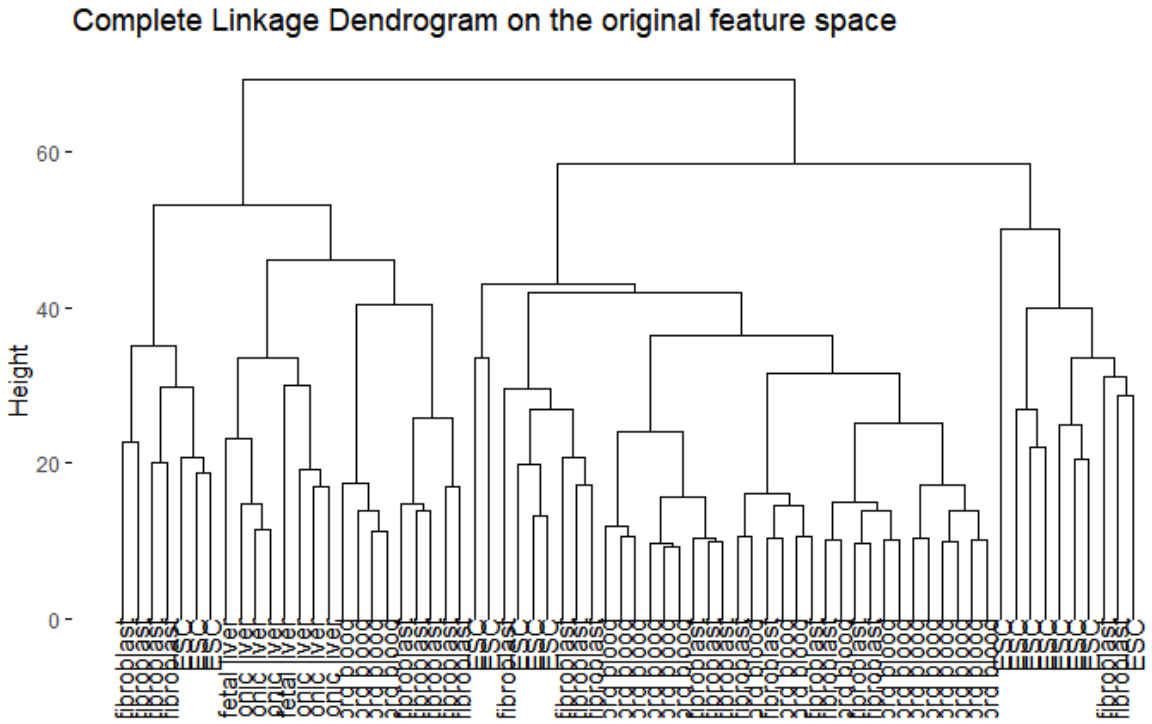






Dendrogram clusters

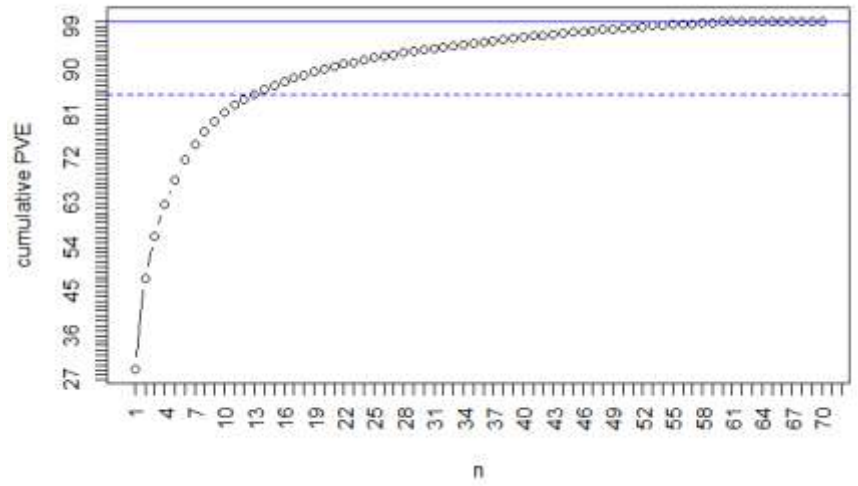




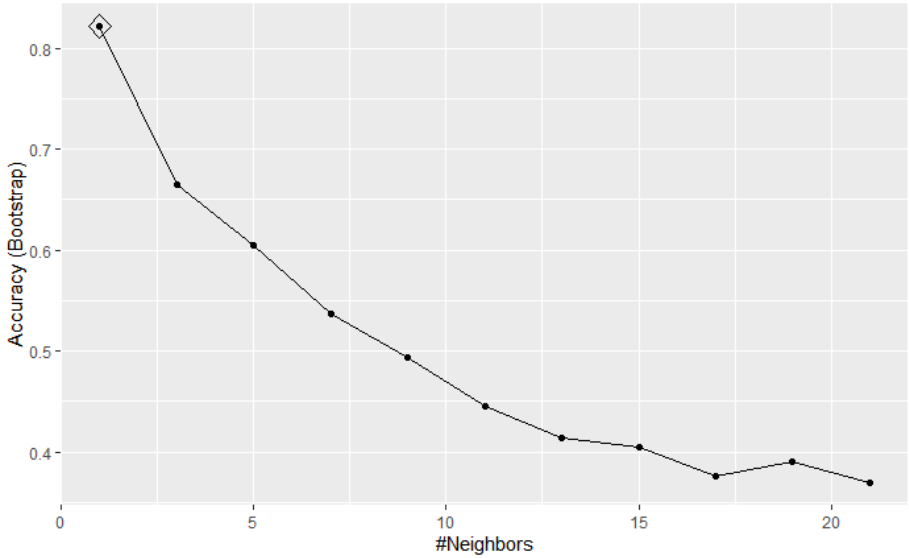
Origin tissue classification with knn

- Task: Predict origin tissue from gene expression.
- 6 classes: "ESC", "cord blood", "adult fibroblast", "embryonic liver" , "neonatal fibroblast", "AD specific fibroblast"

Using a proxy dataset obtained from PCA



The dashed line shows the 85% cutoff

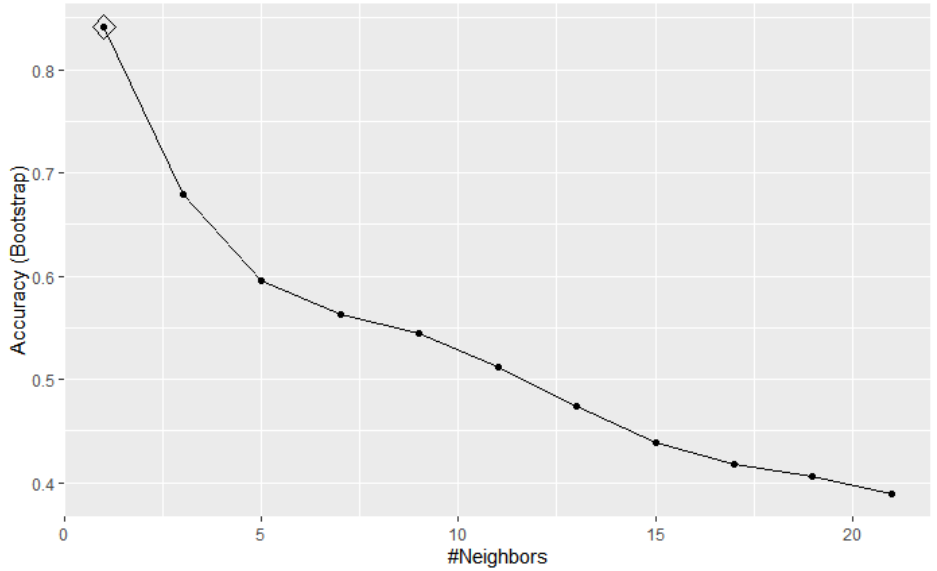


Model performance over different values of k

Origin tissue classification with knn

- Task: Predict origin tissue from gene expression.
- 6 classes: "ESC", "cord blood", "adult fibroblast", "embryonic liver" , "neonatal fibroblast", "AD specific fibroblast"

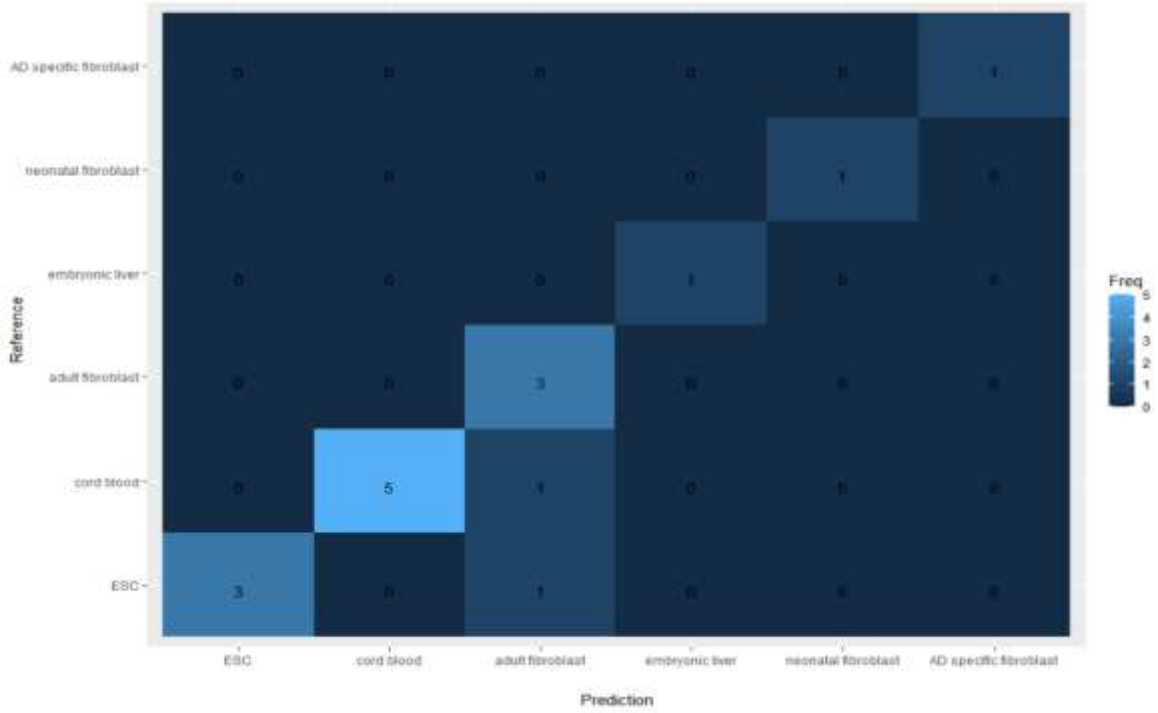
Using the dataset obtained from GO



Model performance over different values of k

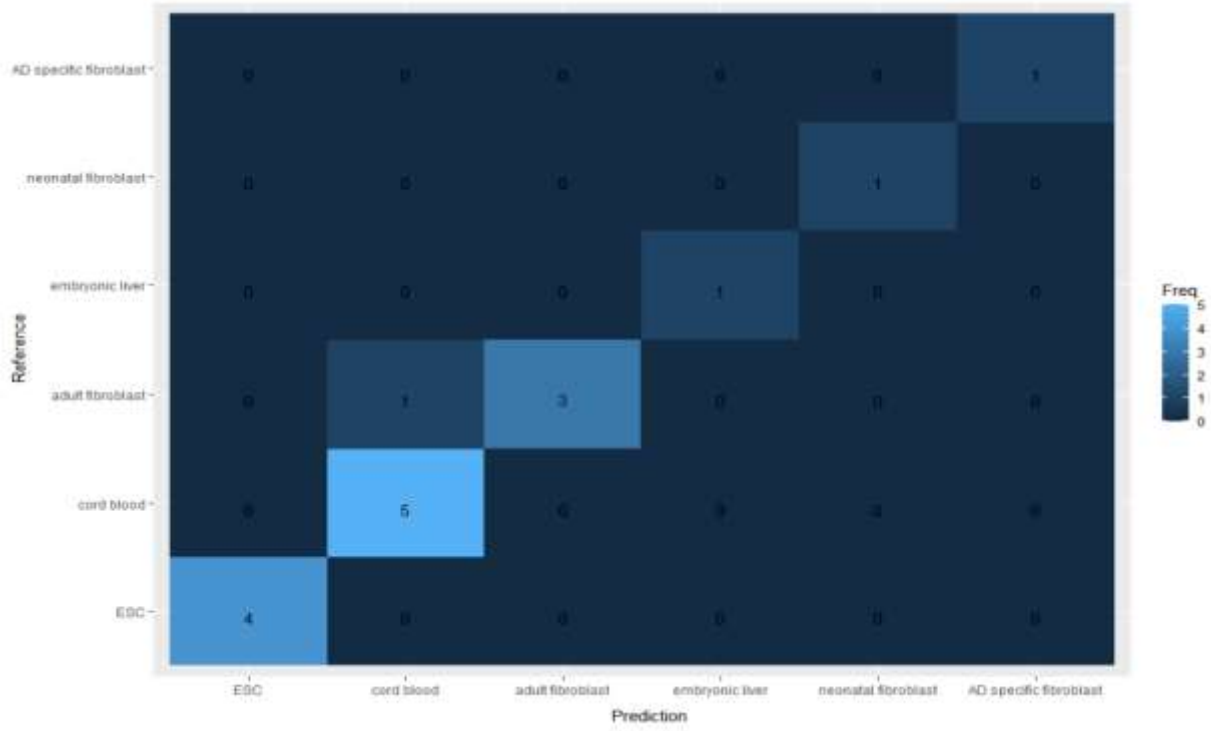
Confusion matrices comparison

Results using the 12 PC proxy



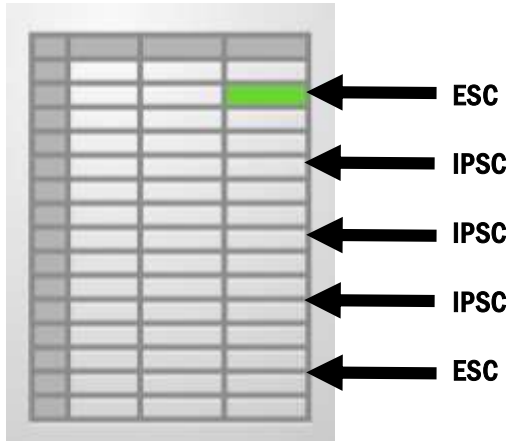
Accuracy: 87.5%

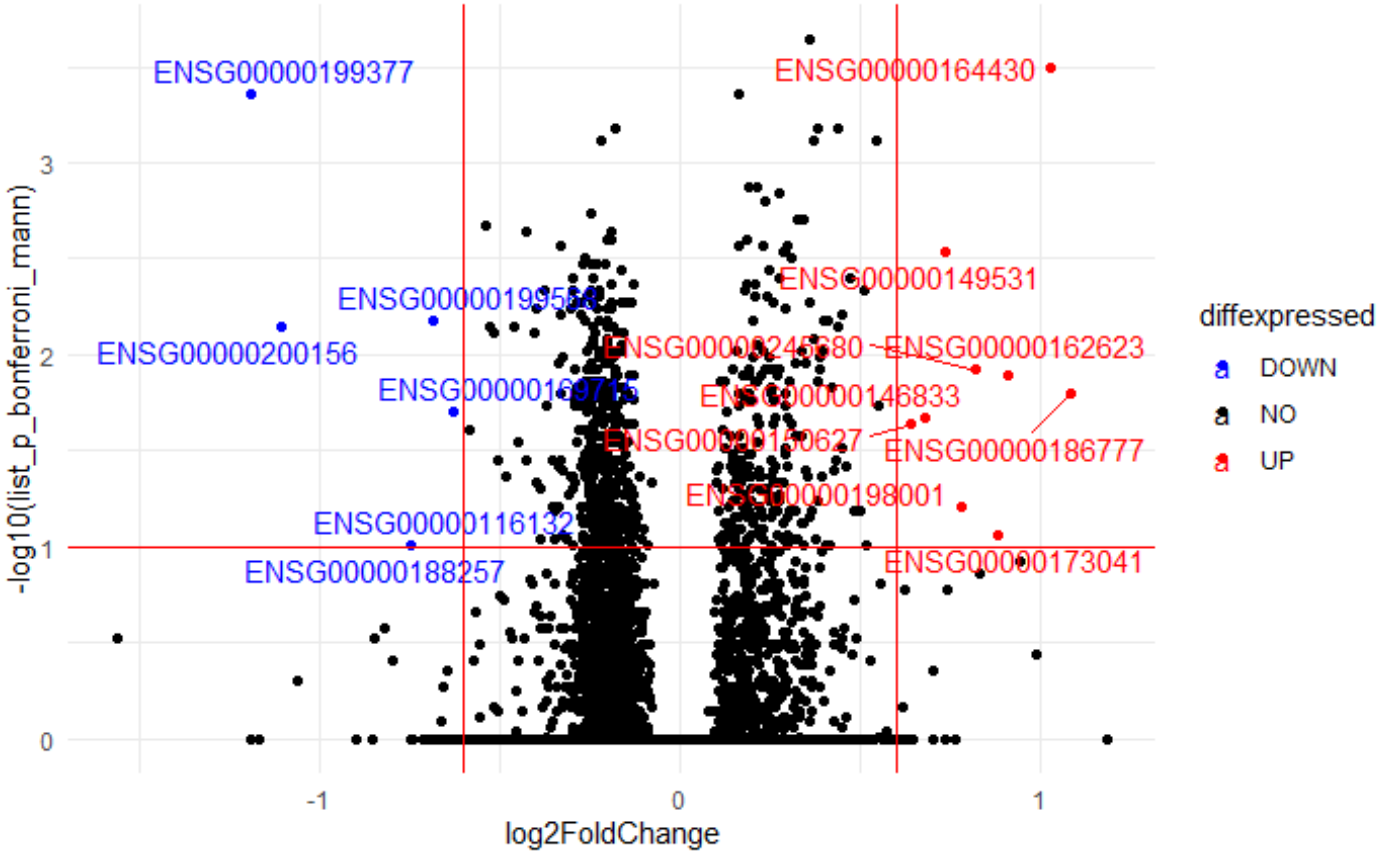
Results using the feature space from GO



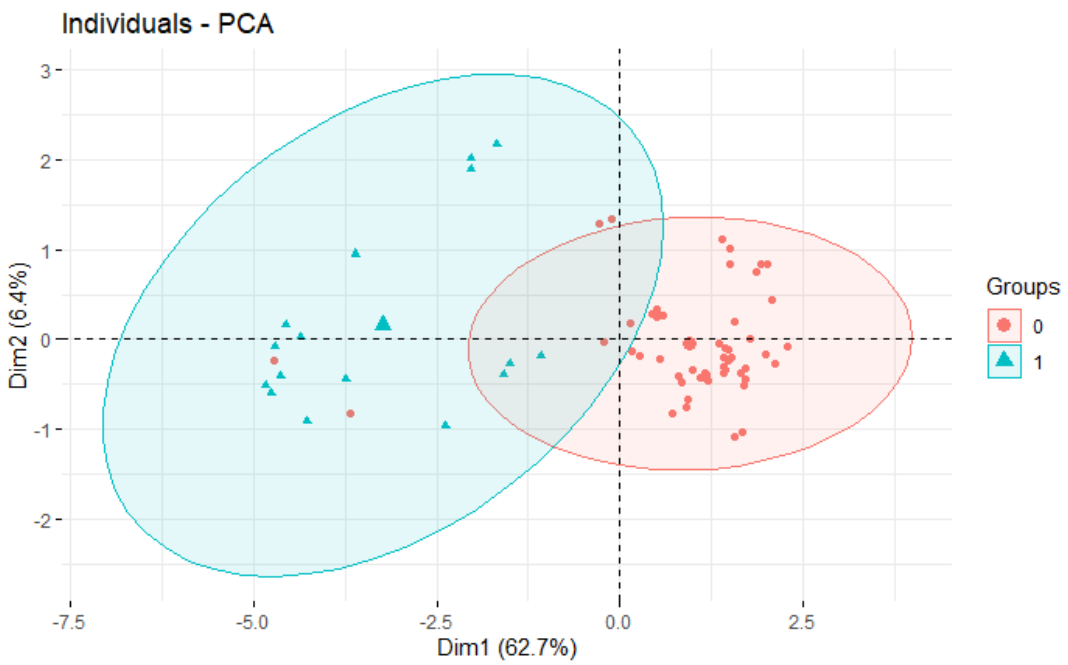
Accuracy: 93.75%

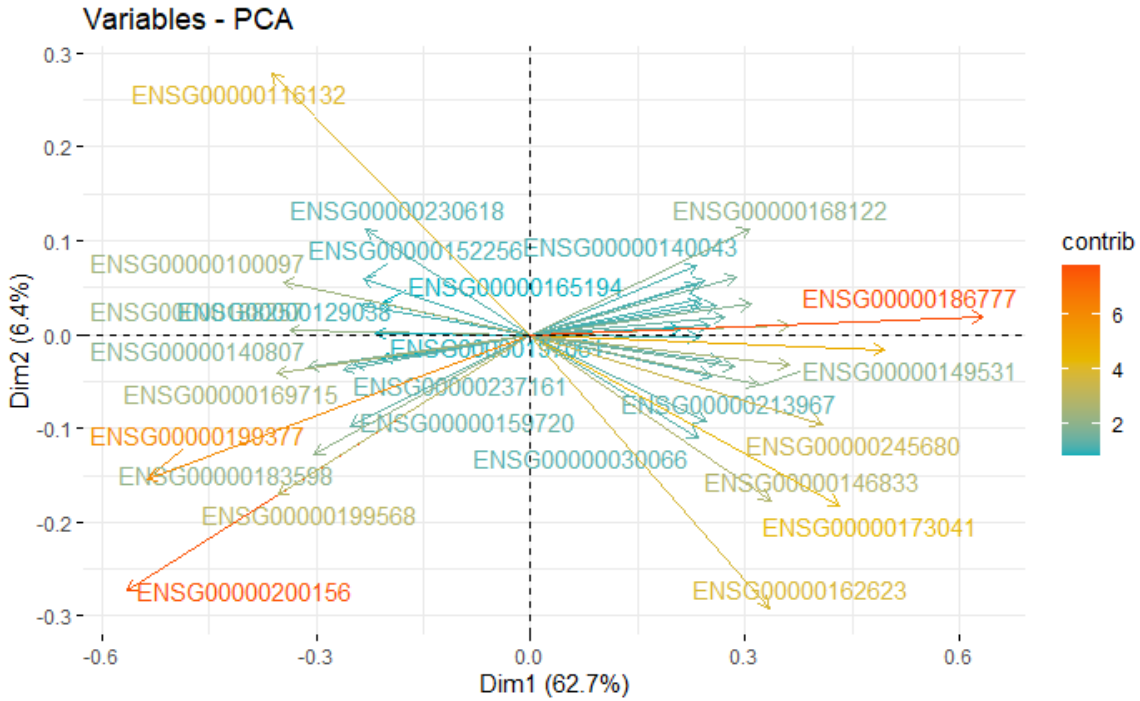
Which genes are markedly differentially expressed?



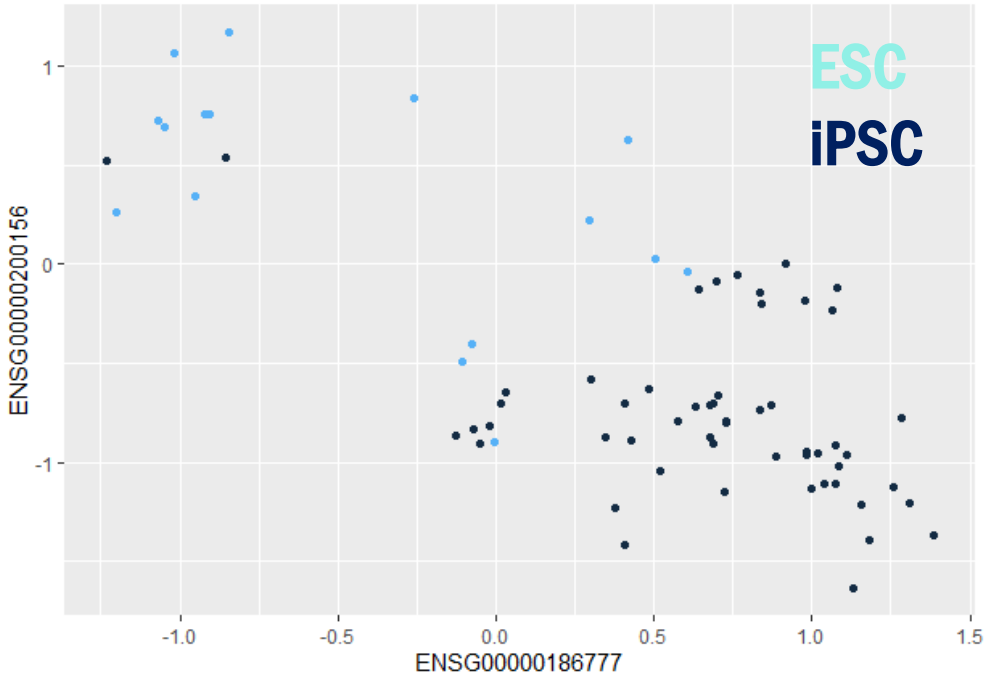


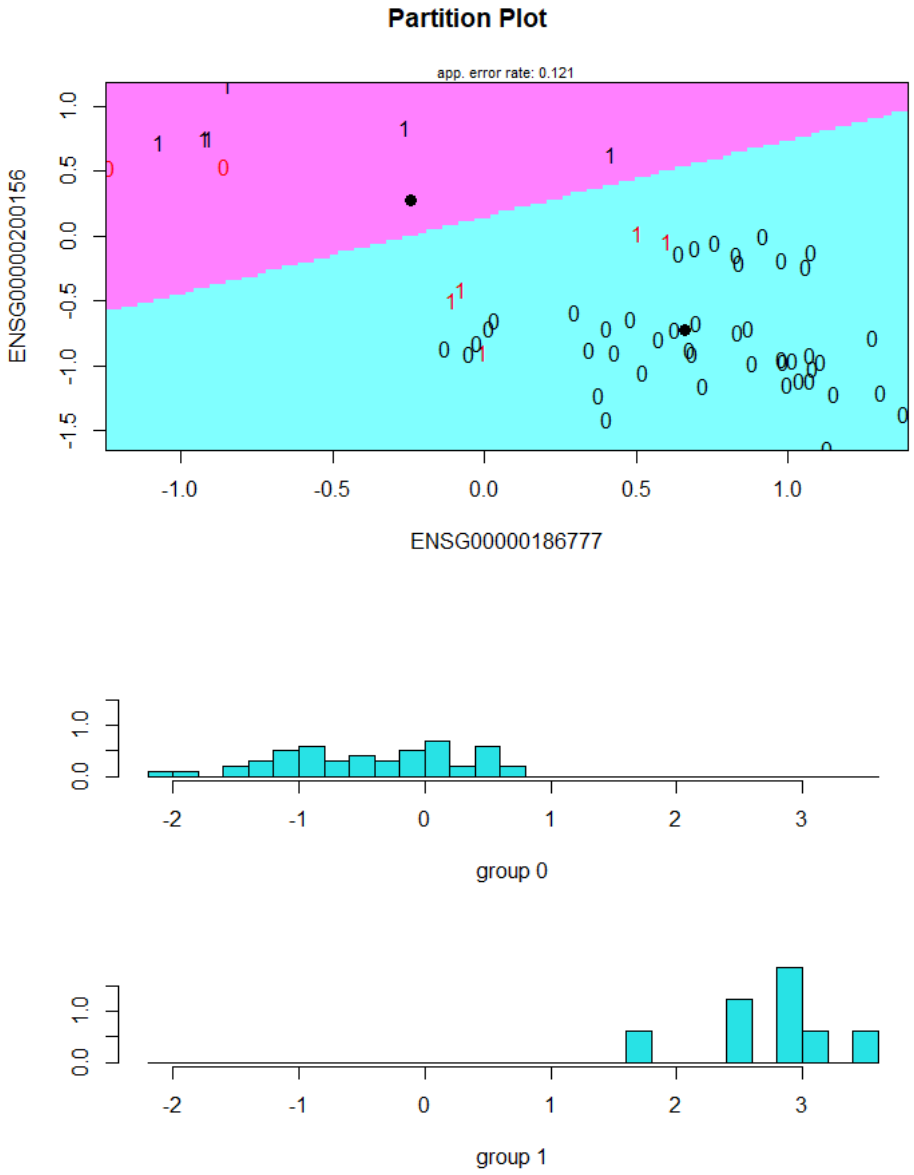
Which genes can discriminate ESCs versus IPSCs?



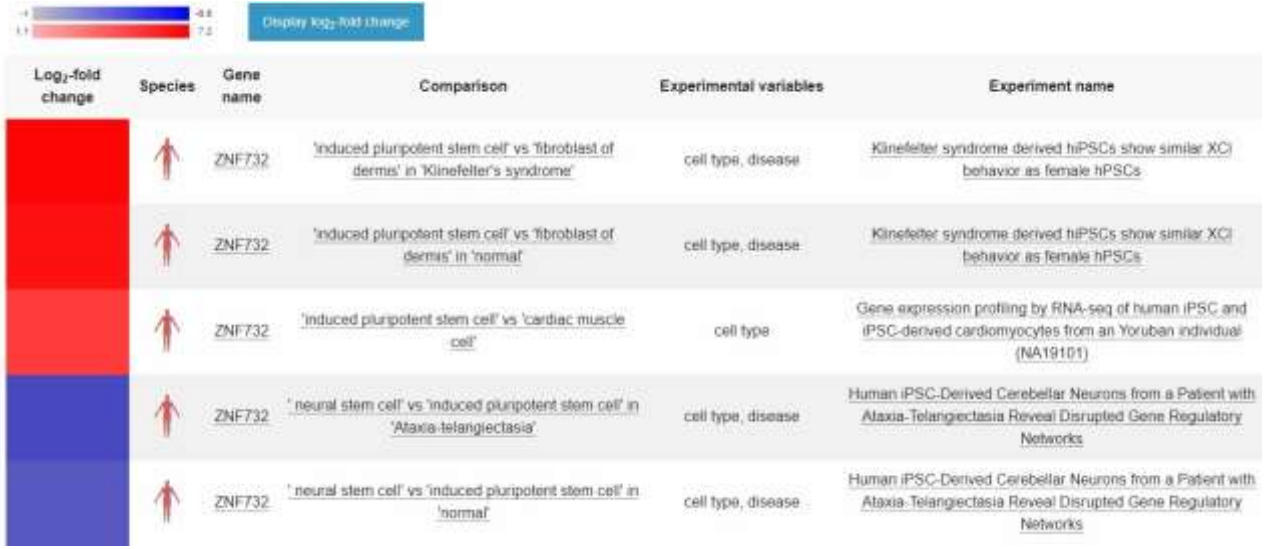


Two genes are enough!



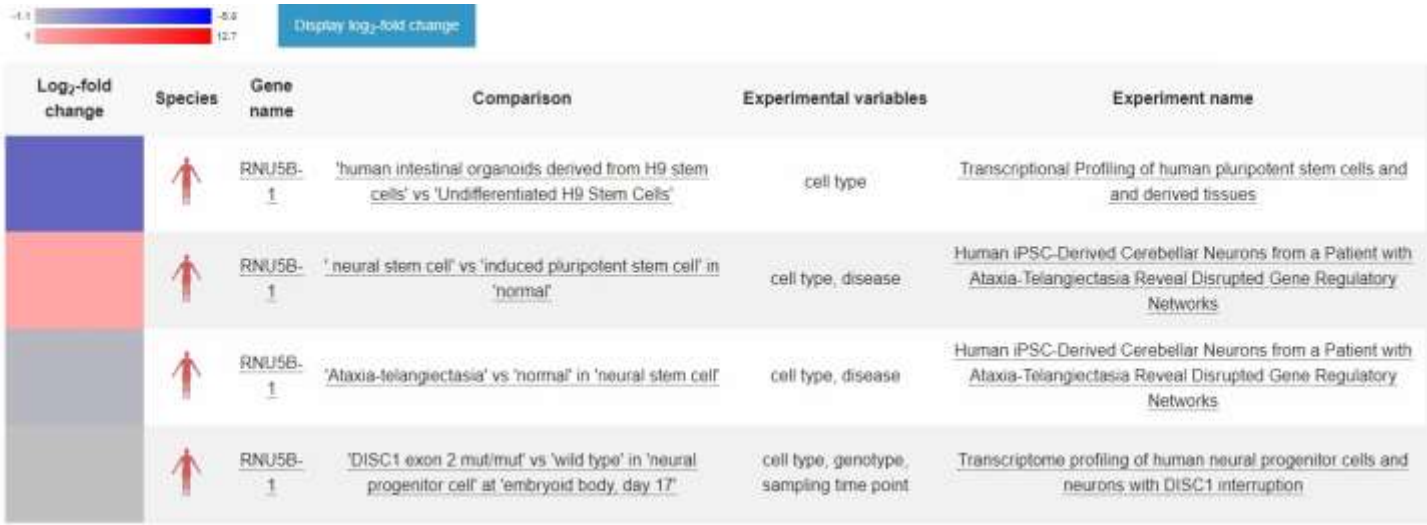


Who are these two genes?



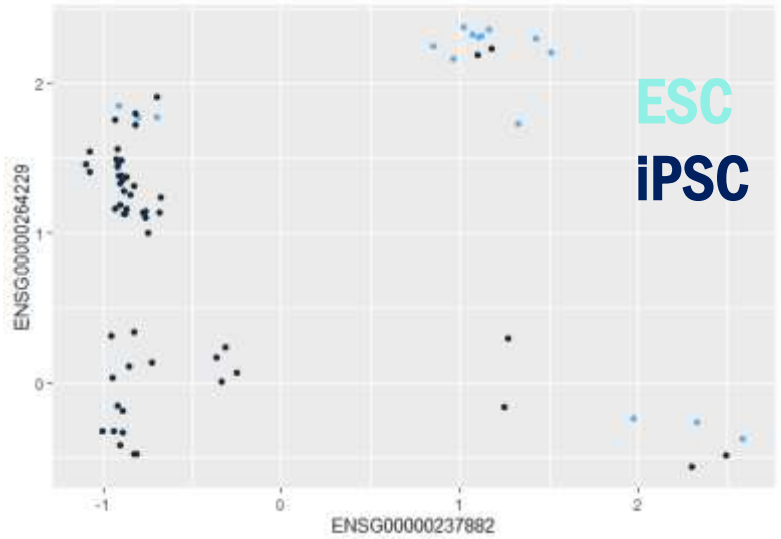
EBI Expression Atlas

ZFP-732

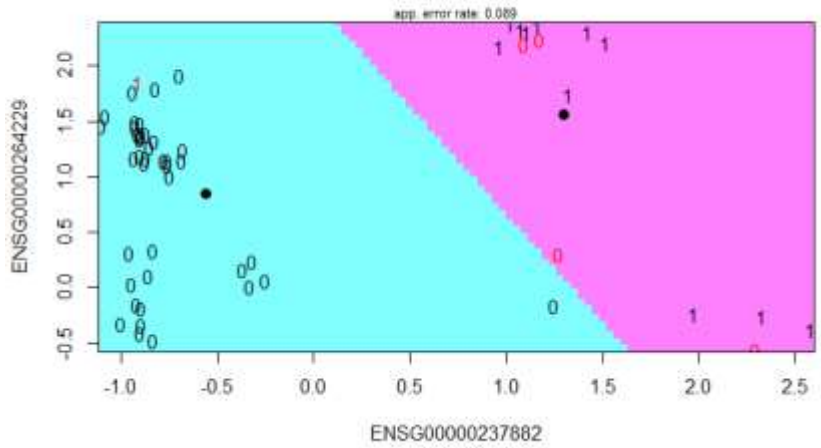


U5B sn1 RNA

Does accuracy improve if we add more genes?



Partition Plot



2-genes model # 1			
		LDA OUTPUT	
		IPSC	ESC
REALITY	IPSC	52	2
	ESC	6	10

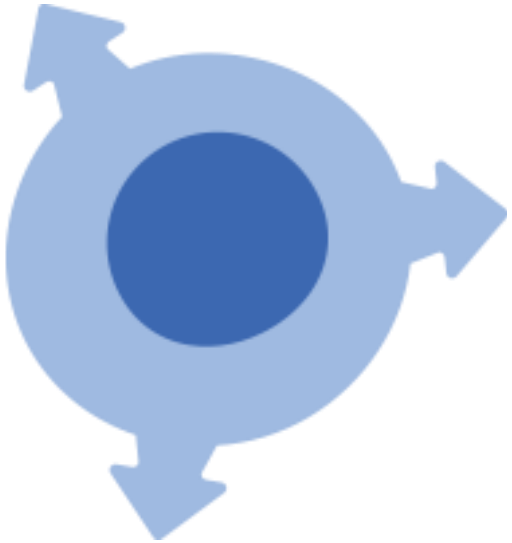
2-genes model # 2			
		LDA OUTPUT	
		IPSC	ESC
REALITY	IPSC	50	4
	ESC	4	12

6-genes model			
		LDA OUTPUT	
		IPSC	ESC
REALITY	IPSC	52	2
	ESC	4	12

Any biological meaning?

iPSC upregulated GO Terms

- mesenchyme migration and morphogenesis
- Wnt signaling pathway
- heart development
- negative regulation of morphogenesis of an epithelium
- outflow tract septum morphogenesis
- chromatin organization involved in negative regulation of transcription
- negative regulation of gene expression, epigenetic
- endoderm development



Different protocols and tissues of origin still affect some gene modules, making ***iPSCs imprecise replicas of ESCs***.

References

- Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics 17, 29–39 (2016).
- Leek, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11, 733–739 (2010).
- Fajarda, O., Duarte-Pereira, S., Silva, R. M. & Oliveira, J. L. Merging microarray studies to identify a common gene expression signature to several structural heart diseases. BioData Min. 13, 8 (2020).
- Kapushesky, M. et al. Gene expression atlas at the European bioinformatics institute. Nucleic Acids Res. 38, D690–D698 (2010).
- 5. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell 126, 663–676 (2006).