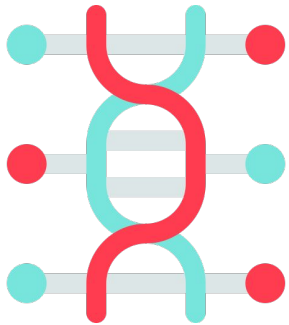
Small nucleolar RNAs in breast cancer

Laboratory of Biological Data Mining

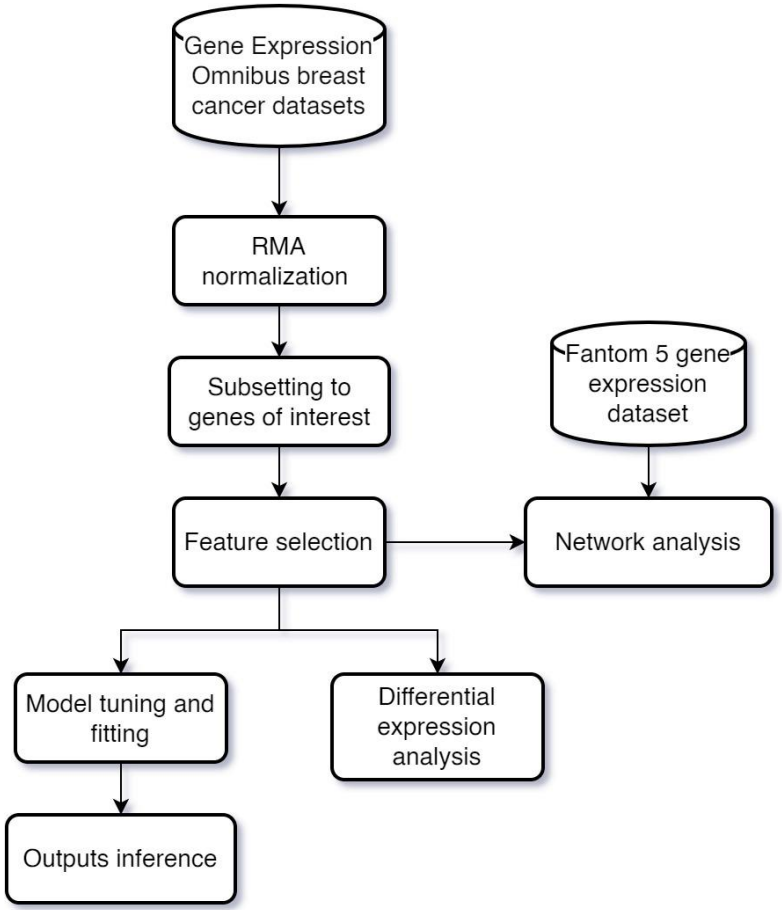
Group 1: Lorenzo Bocchi, Valentino Frasnelli, Erich Robbi, Annalisa Xamin

Introduction

Problem definition



Project Pipeline



Data collection

snoDB

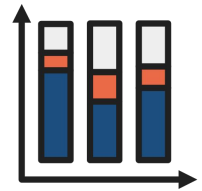
FANTOM

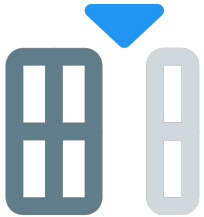
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

GEO

Gene Expression Omnibus

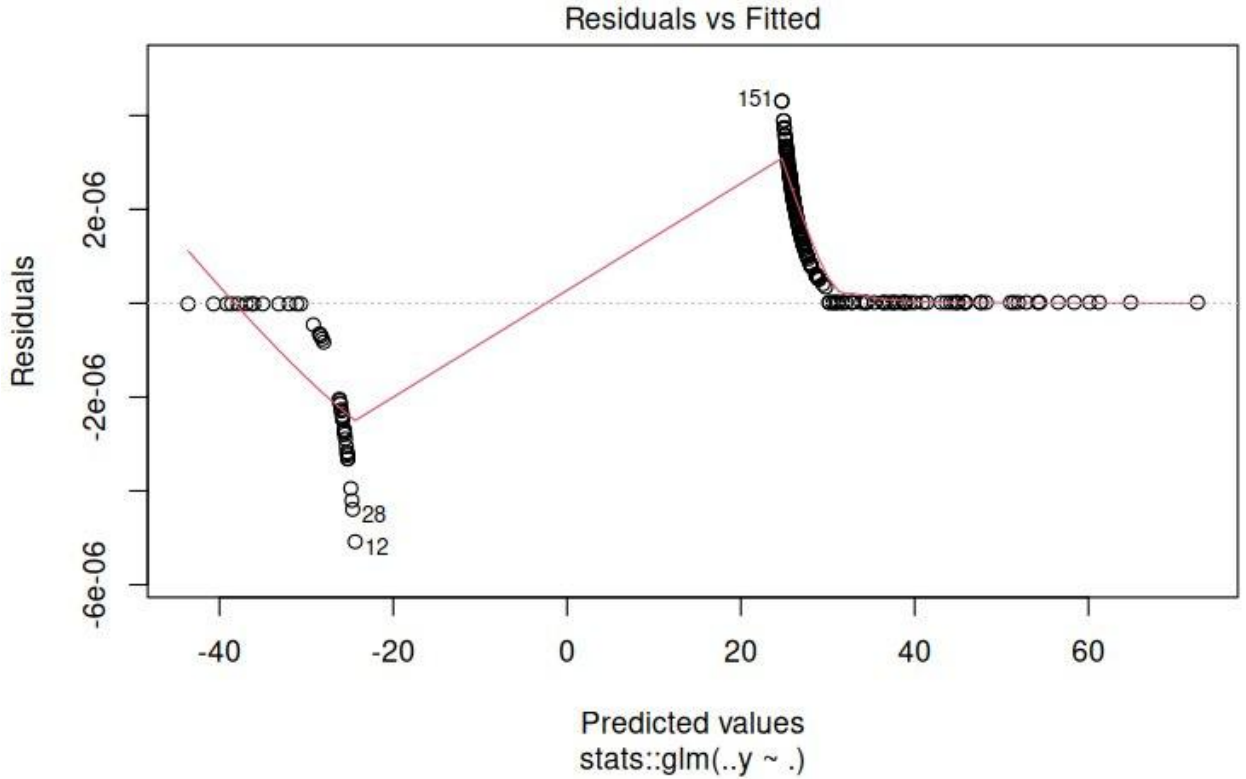
Data preparation





Model fitting

First dataset - suspicious results



Random forest performance measures on our genes of interest (old data)

	Accuracy	Precision	Recall	F1
No Cancer	0.9615	1	0.80	0.89
Cancer	0.9615	0.95	1	0.98

Random forest fit on housekeeping genes (old data)

	Accuracy	Precision	Recall	F1
No Cancer	0.9221	1	0.60	0.75
Cancer	0.9221	0.91	1	0.95

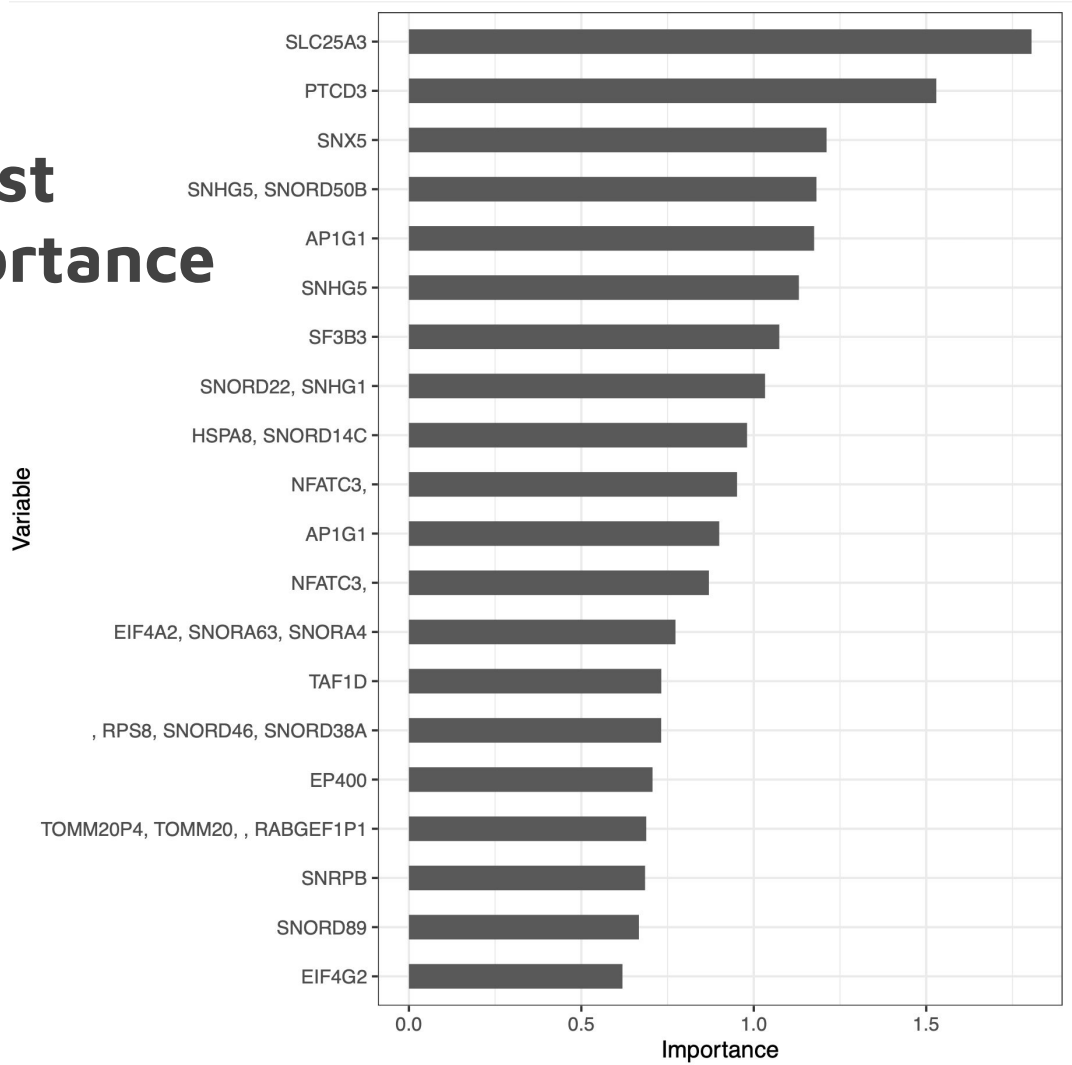
Random forest performance measures on our genes of interest (new data)

	Accuracy	Precision	Recall	F1
Basal	0.8571	0.75	0.82	0.78
HER	0.8286	0.63	0.63	0.63
L. A	0.8857	0.67	0.86	0.75
L. B	0.8833	0.60	0.75	0.67
Normal	1	1	1	1

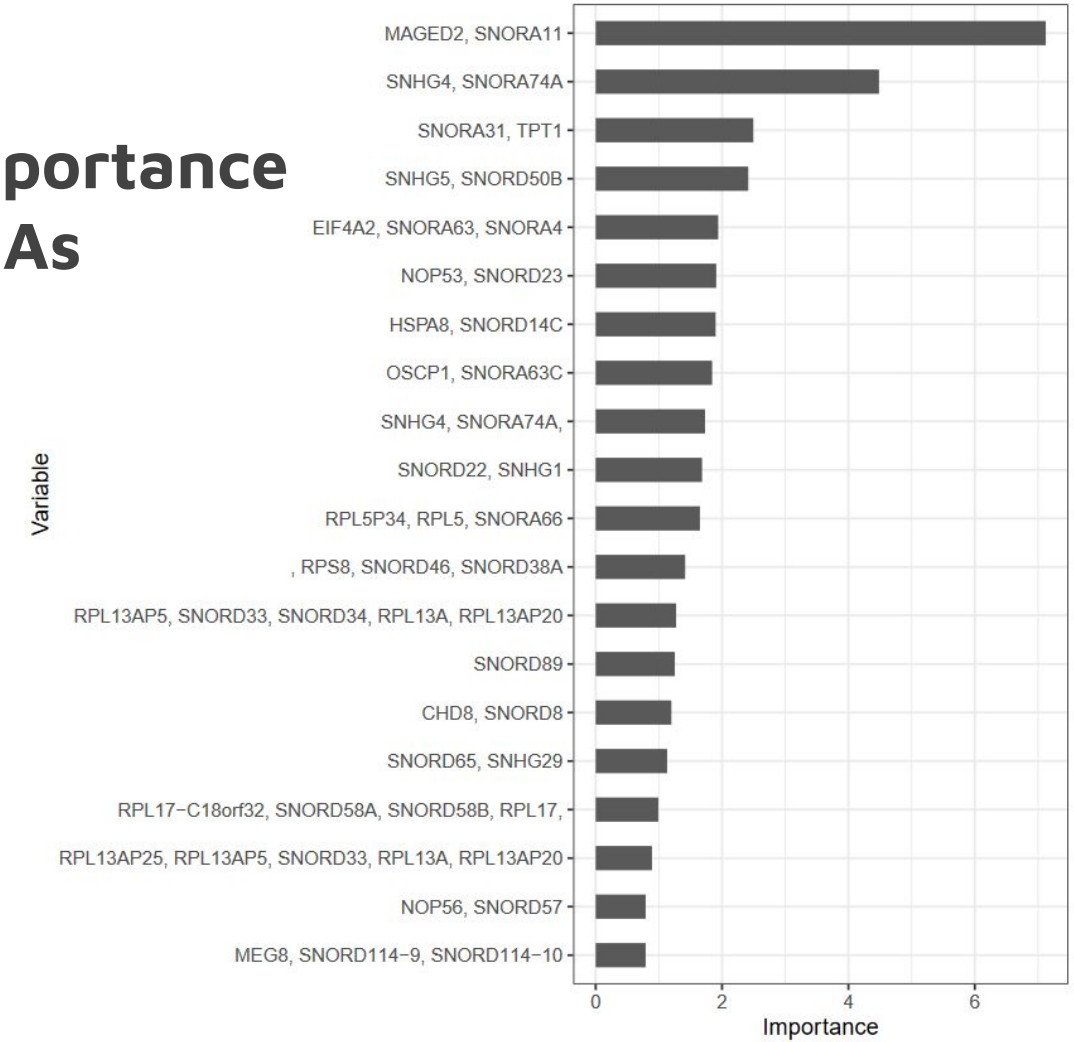
Random forest performance metrics on lesser variant genes (new data)

	Accuracy	Precision	Recall	F1
Basal	0.6	0.41	0.64	0.5
HER	0.65	0	0	0
L. A	0.74	0.24	0.14	0.18
L. B	0.71	0.40	0.5	0.44
Normal	0	0	0	0

Random Forest variable importance



Variable importance only snoRNAs



Final SVM model

Property	Value
Cost	0.005332322
Support Vectors	92/102
Training Error	0.019608

Final SVM model performance measures

	Accuracy	Precision	Recall	F1
Basal	0.8857	0.77	0.91	0.83
HER	0.8857	0.75	0.75	0.75
L. A	0.9429	0.78	1	0.88
L. B	0.8857	1	0.5	0.67
Normal	1	1	1	1

	Accuracy	Precision	Recall	F1
Basal	0.8571	0.75	0.82	0.78
HER	0.8286	0.63	0.63	0.63
L. A	0.8857	0.67	0.86	0.75
L. B	0.8833	0.60	0.75	0.67
Normal	1	1	1	1

Random forest

SVM

	Accuracy	Precision	Recall	F1
Basal	0.8857	0.77	0.91	0.83
HER	0.8857	0.75	0.75	0.75
L. A	0.9429	0.78	1	0.88
L. B	0.8857	1	0.5	0.67
Normal	1	1	1	1

Logistic regression with L1 penalty performance measures

	Accuracy	Precision	Recall	F1
Basal	0.8857	0.82	0.82	0.82
HER	0.8571	0.71	0.63	0.67
L. A	0.8571	0.60	0.86	0.71
L. B	0.8286	.67	0.5	0.57
Normal	1	1	1	1

Most important variables

Coefficient (Probe id)	Estimate	Gene Symbols
214744_s_at	-5.807828e-01	RPL23
228971_at	-5.005921e-01	SLC25A3
232355_at	-4.876120e-01	MEG8, SNORD114-3
235102_x_at	4.474731e-01	SNORD3B-1, SNORD3B-2, SNORD3D, SNORD3A, SNORD3C
223666_at	4.409927e-01	SNX5
231096_at	-4.260707e-01	PCAT4
242856_at	-4.194403e-01	MEG8, SNORD114-9, SNORD114-10
221621_at	3.693238e-01	SNHG20
228879_at	-3.637258e-01	SNORD104, SNORA50C
241448_at	2.737583e-01	TOMM20

Least important variables

Coefficient (Probe id)	Estimate	Gene Symbols
1555177_at	0	PRKAA1
215011_at	0	SNHG3
229038_at	0	CWF19L1
226428_at	0	TNPO2
240083_at	0	MEG8
200716_x_at	-8.590805e-05	RPL13AP5, RPL13A
200031_s_at	-1.930631e-04	RPS11
224741_x_at	-2.245289e-04	GAS5
200692_s_at	2.554860e-04	HSPA9
209476_at	-2.689458e-04	TMX1

Network Analysis

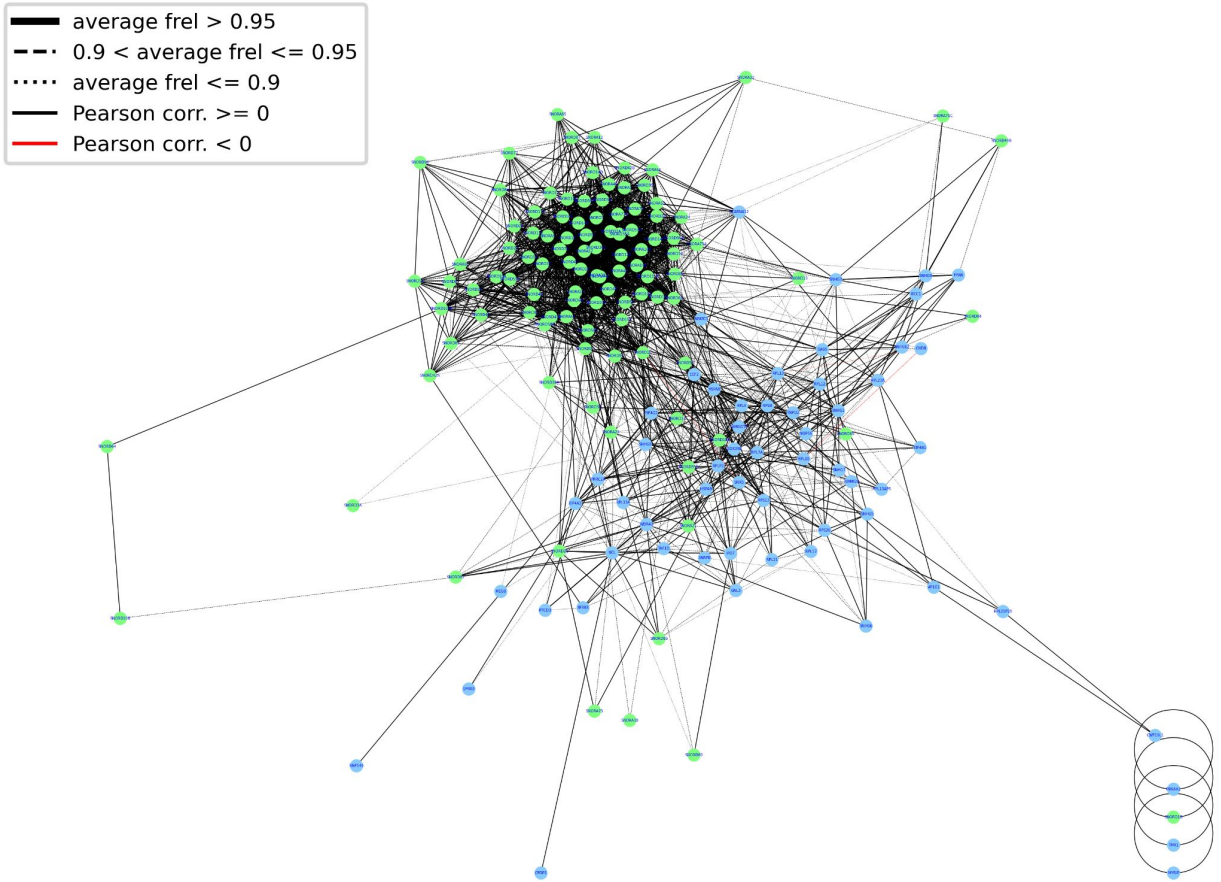
List of genes not yet expanded by gene@home

PCAT4	SNORD114-14	RPL23
SNORD114-1	RPS3	SNORD114-19
RACK1	SNORD113-4	NAN
SNORD114-13	RPS11	RPS2
SNORD114-20	SNHG20	SNORD115-23
SNORD116-13	CDKN2B-AS1	SNORD113-3
SNORD115-32	SLC25A3	ZFAS1
HIF1A-AS2	SNORD114-21	

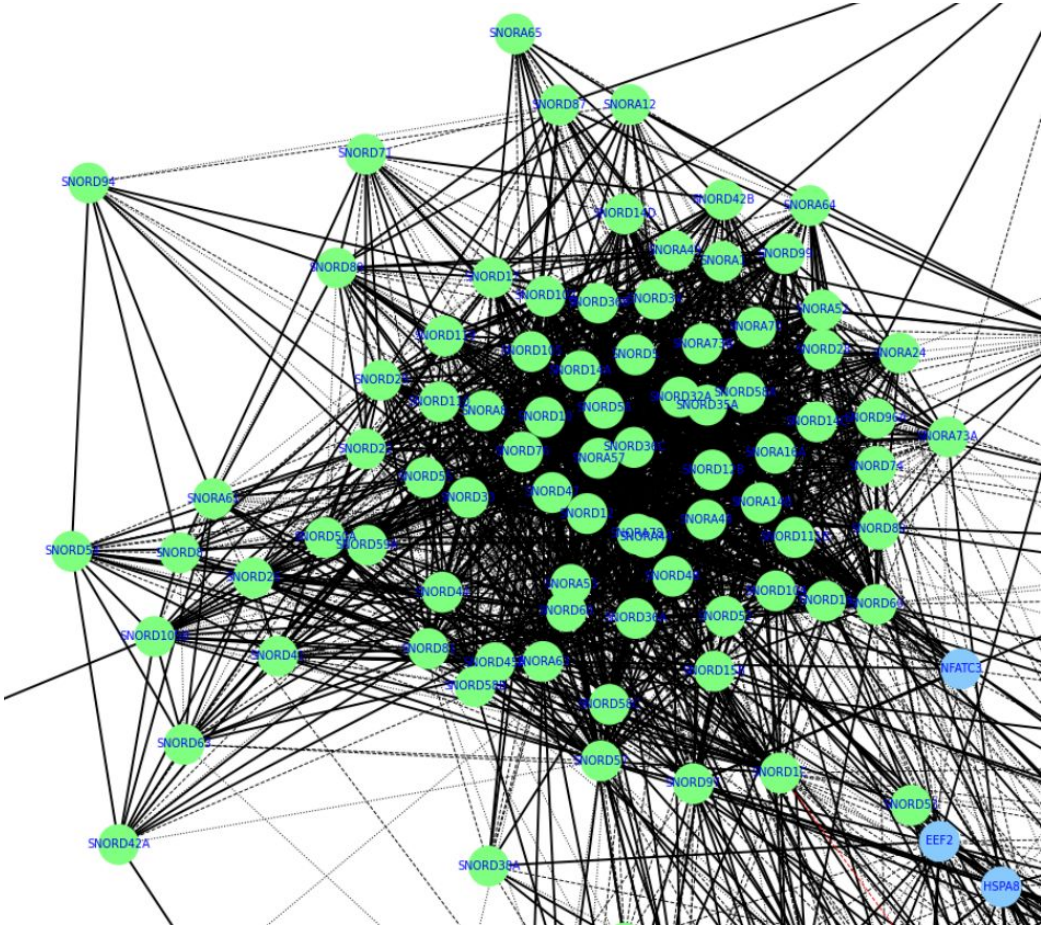
List of genes not yet expanded by gene@home

PCAT4	~~SNORD114-14~~	RPL23
~~SNORD114-1~~	RPS3	~~SNORD114-19~~
RACK1	~~SNORD113-4~~	NAN
~~SNORD114-13~~	RPS11	RPS2
~~SNORD114-20~~	SNHG20	~~SNORD115-23~~
~~SNORD116-13~~	CDKN2B-AS1	~~SNORD113-3~~
~~SNORD115-32~~	SLC25A3	ZFAS1
~~HIF1A-AS2~~	~~SNORD114-21~~	

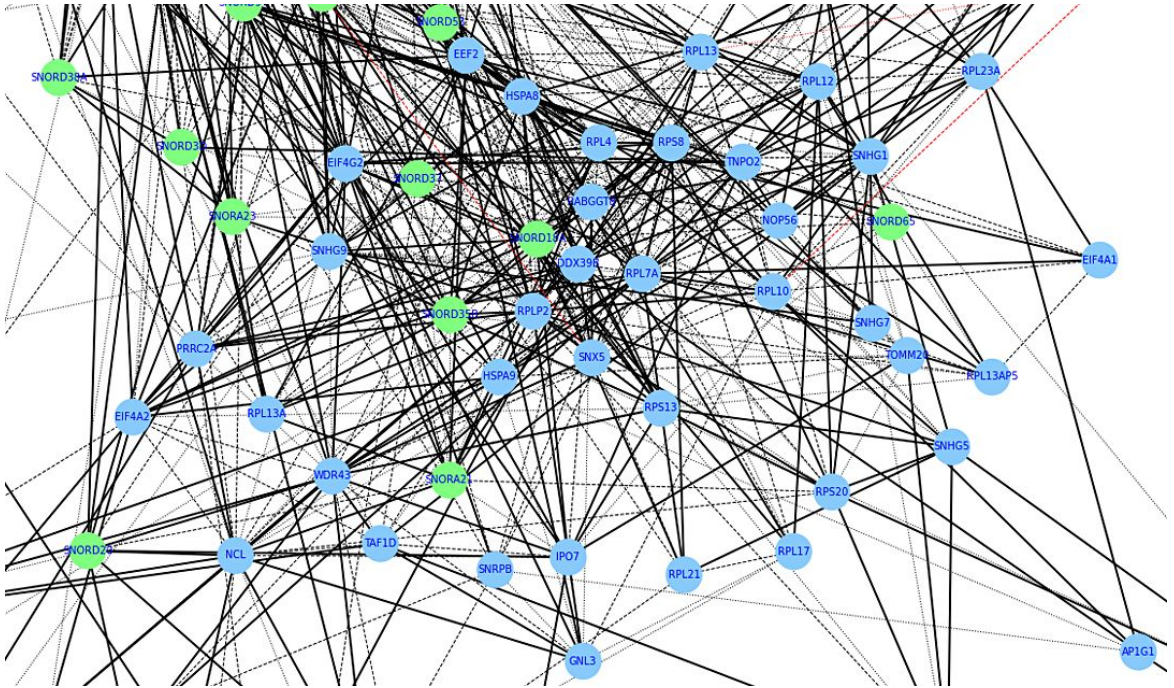
Complete graph



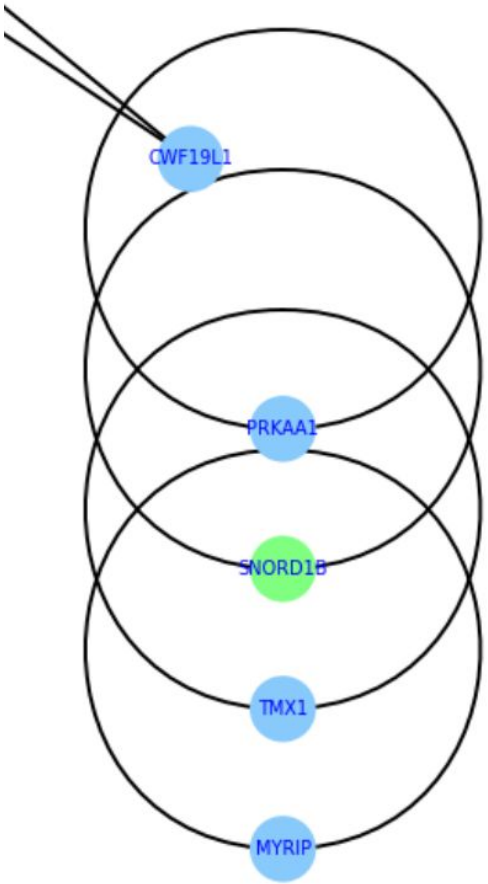
Complete graph



Complete graph



Complete graph



Validation @0.85

SNORD41 → SNORD60

SNORA73B → SNORA73A

SNORA78 → SNORD28

SNORD50A → SNORD59A

SNORD20 → SNORA75

SNORD108 → SNORD64

SNORD12 → SNORD12B

SNORD47 → SNORD59A

SNORD47 → SNORD60

SNORA70 → SNORD52

SNORD94 → SNORD22

Validation @0.95

SNORA73B → SNORA73A

SNORD108 → SNORD64

SNORD47 → SNORD59A

SNORD47 → SNORD60

SNORD20 → SNORA75

SNORD12 → SNORD12B

SNORD94 → SNORD22

Connections between snoRNAs and their respective host genes

snoRNA	Host Gene
SNORD18A	RPL4
SNORD20	NCL
SNORA73A	SNHG3
SNORA14B	TOMM20
SNORA38	PRRC2A
SNORD82	NCL
SNORA78	SNHG9
SNORA48	NFATC3
SNORD68	RPL13
SNORD97	EIF4G2
SNORA75	NCL
SNORD37	EEF2

JOINED RESULTS

TOMM20

- Important predictor for both the random forest models
- When compared to the other predictors, its coefficient has experienced the least shortening.

snoRNA	freq. rel.
SNORA14B	0.93
SNORD13	0.86

SNX5

- Important predictor for the lasso model
- Important predictor for the RF

snoRNA	freq. rel.
SNORD104	0.98
SNORD1C	0.92
SNORD53	0.94
SNORD58C	0.91
TOMM20	0.8584

SNHG1

- Important predictor for RF

snoRNA	freq. rel.
SNORD104	0.99
SNORD57	0.99
GAS5	0.97

HSPA8

- Important predictor for RF

- SNORA16A - 0.94
- SNORA44 - 0.97
- SNORA61 - 0.86
- SNORD119 - 0.98
- SNORD13 - 0.88
- SNORD15B - 0.95
- SNORD18A - 0.98
- SNORD1C - 0.87
- SNORD34 - 0.92
- SNORD35B - 0.95
- SNORD37 - 0.88
- SNORD57 - 1
- SNORD68 - 0.88
- SNORD84 - 0.95
- SNORD97 - 0.97

NFATC3

- Important predictor for RF

- SNORA14B - 0.95
- SNORA16A - 0.99
- SNORA44 - 0.86
- SNORA48 - 0.98
- SNORA53 - 0.98
- SNORA57 - 0.92
- SNORA63 - 0.98
- SNORA73B - 0.94
- SNORA78 - 0.93
- SNORD10 - 0.89
- SNORD104 - 0.90
- SNORD14A - 0.90
- SNORD22 - 0.90
- SNORD56 - 0.95
- SNORD57 - 0.89
- SNORD97 - 0.97
- **SNX5** - 0.89

Differential Gene Expression

Published online 20 January 2015

Nucleic Acids Research, 2015, Vol. 43, No. 7 ***e47***
doi: 10.1093/nar/gkv007

limma powers differential expression analyses for RNA-sequencing and microarray studies

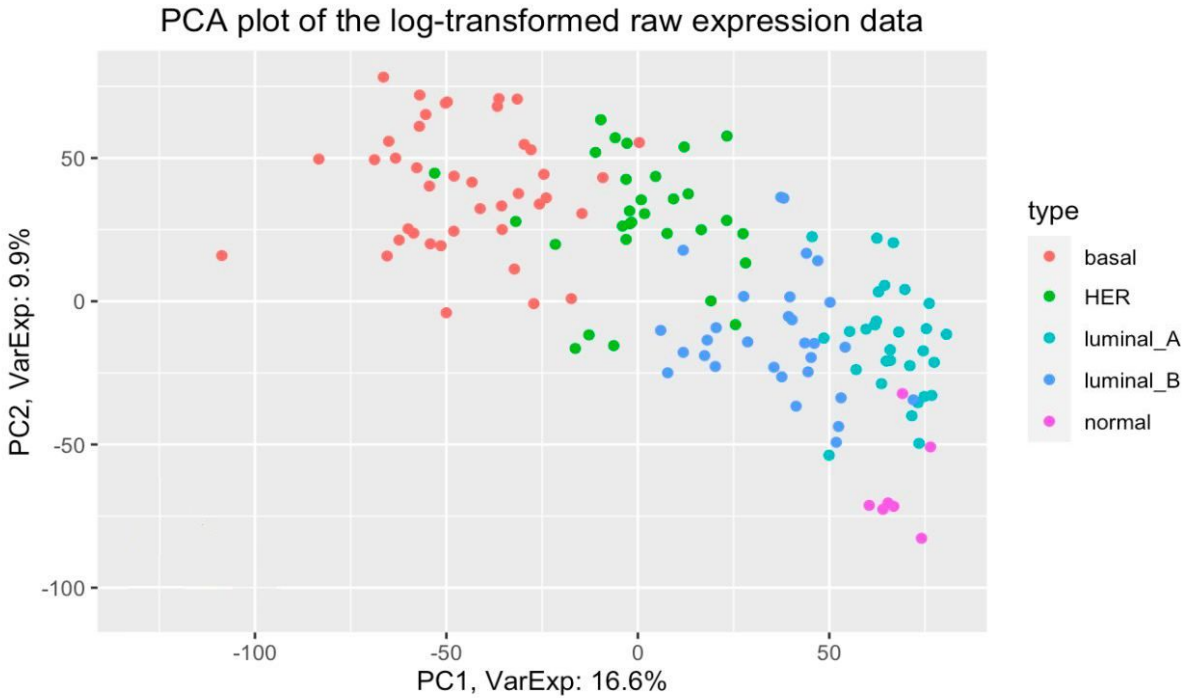
Matthew E. Ritchie[1,2], Belinda Phipson[3], Di Wu[4], Yifang Hu[5], Charity W. Law[6], Wei Shi[5,7] and Gordon K. Smyth[2,5,*]

[1]Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, [2]Department of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia, [3]Murdoch Childrens Research Institute, Royal Children's Hospital, 50 Flemington Road, Parkville, Victoria 3052, Australia, [4]Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138-2901, USA, [5]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, [6]Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland and [7]Department of Computing and Information Systems, The University of Melbourne, Parkville, Victoria 3010, Australia

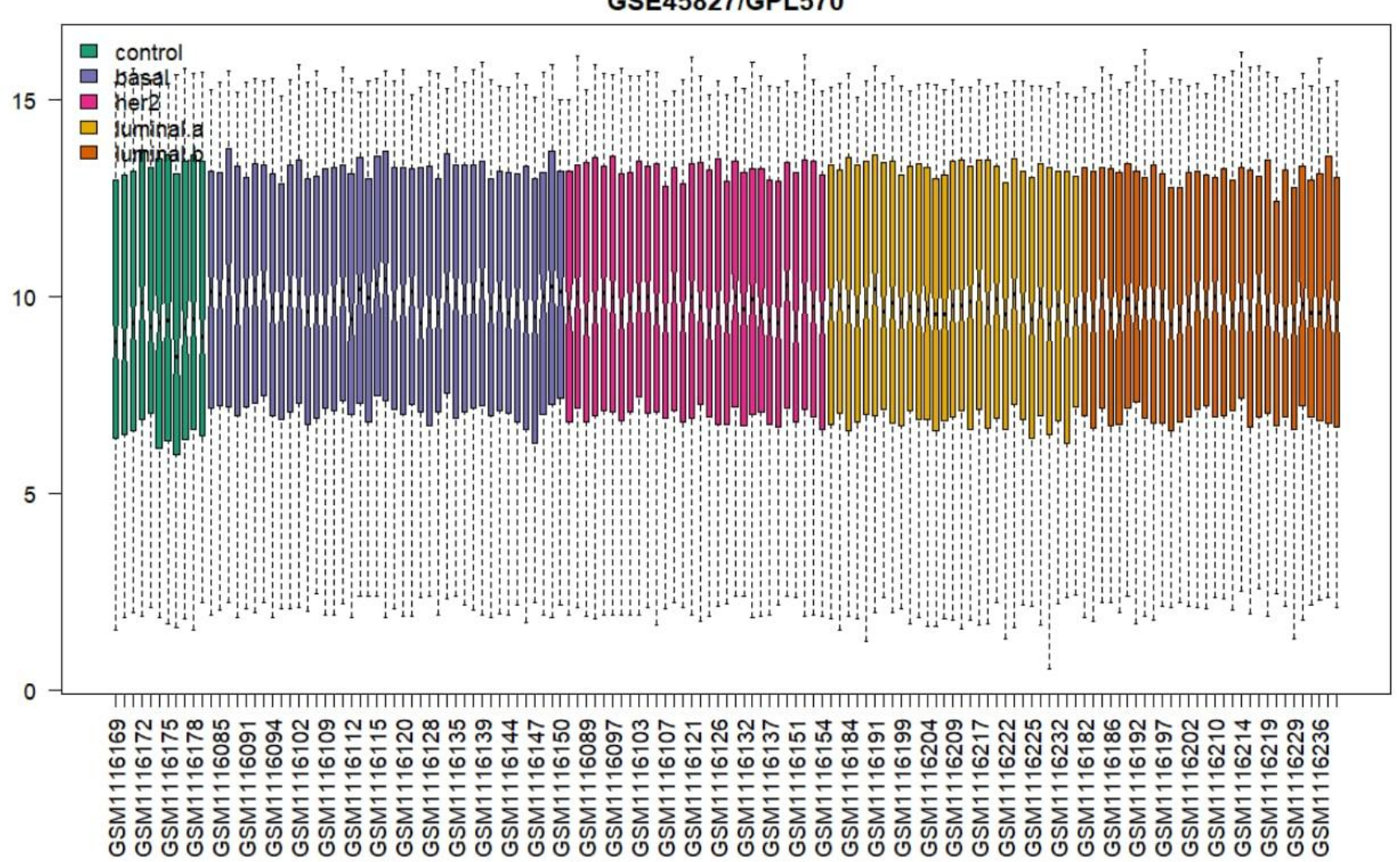
Received November 09, 2014; Revised January 04, 2015; Accepted January 06, 2015

https://doi.org/10.1093/nar/gkv007
https://bioconductor.org/packages/release/bioc/html/limma.html

Exploratory Data Analysis



PCA without cell lines

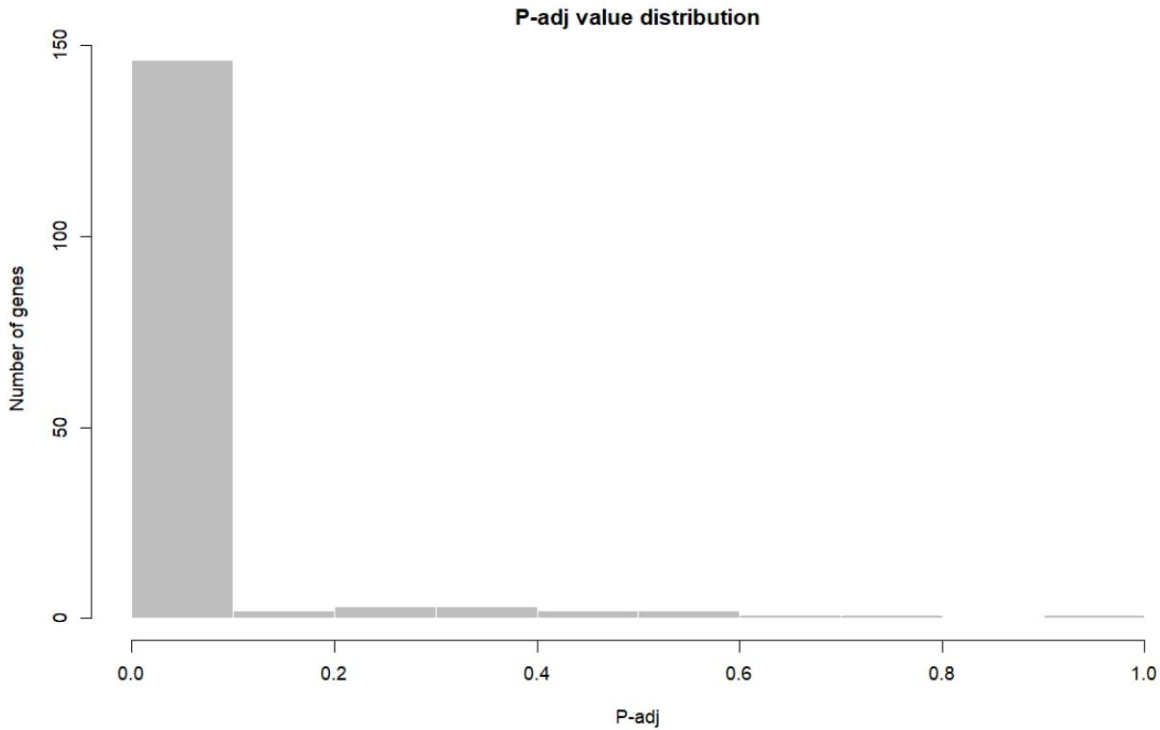

GSE45827/GPL570
control
basal
her2
luminal.a
luminal.b
15
10
5
0
GSM116169
GSM116172
GSM116175
GSM116178
GSM116085
GSM116091
GSM116094
GSM116102
GSM116109
GSM116112
GSM116115
GSM116120
GSM116128
GSM116135
GSM116139
GSM116144
GSM116147
GSM116150
GSM116089
GSM116097
GSM116103
GSM116107
GSM116121
GSM116126
GSM116132
GSM116137
GSM116151
GSM116154
GSM116184
GSM116191
GSM116199
GSM116204
GSM116209
GSM116217
GSM116222
GSM116225
GSM116232
GSM116182
GSM116186
GSM116192
GSM116197
GSM116202
GSM116210
GSM116214
GSM116219
GSM116229
GSM116236

Estimating the proportion of true null hypotheses

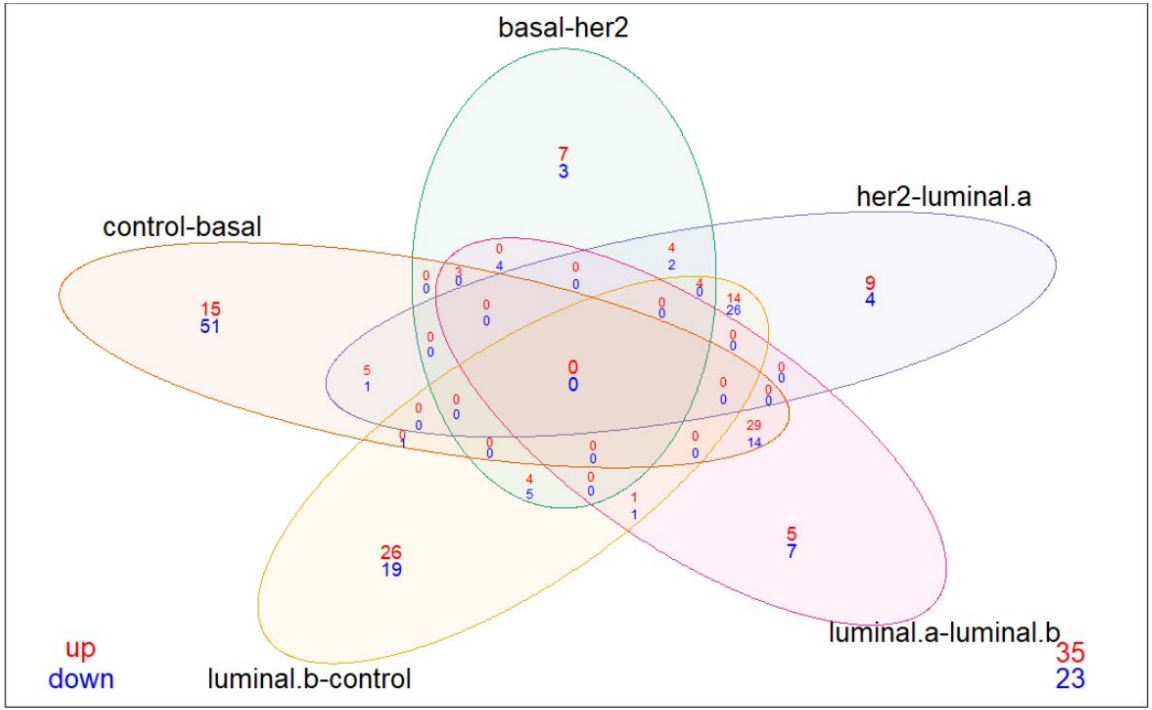
We build a histogram of P-values for all genes.

The null hypothesis is that most genes are not differentially expressed.

We reject the null hypothesis.

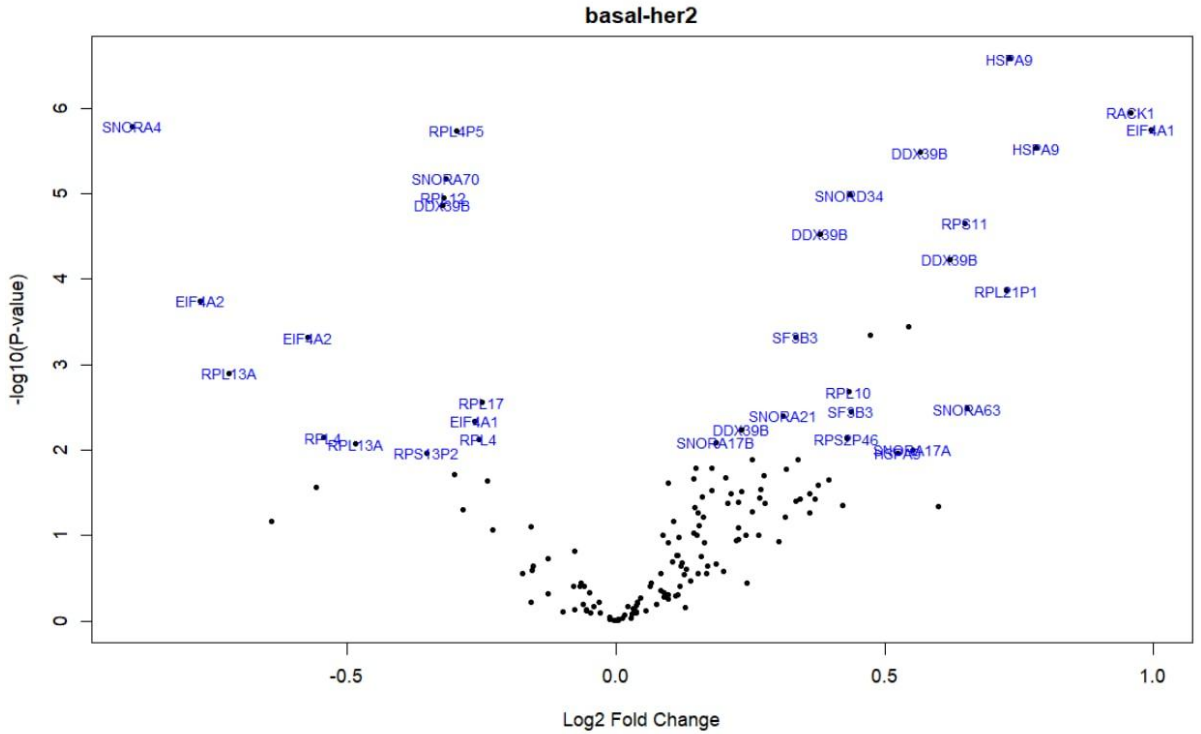


Venn diagram with both up and down-regulated genes

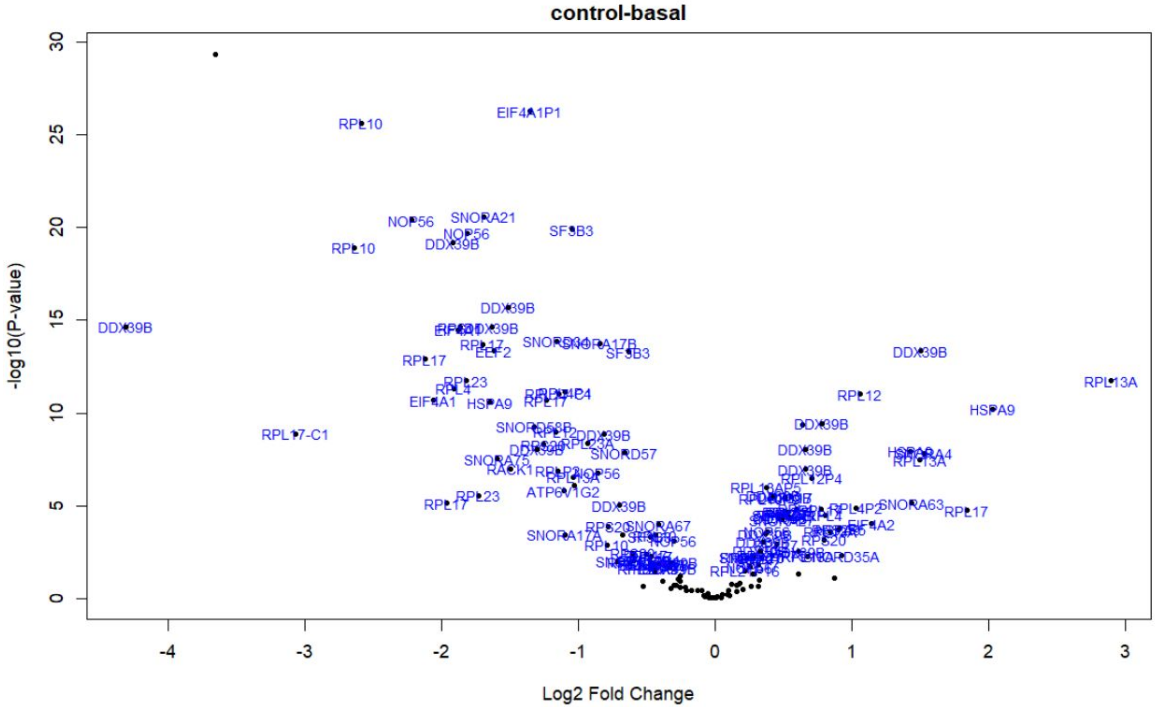


Venn diagram showing overlap in the number of DE genes for the 5 comparisons, generated by the *vennDiagram* function.

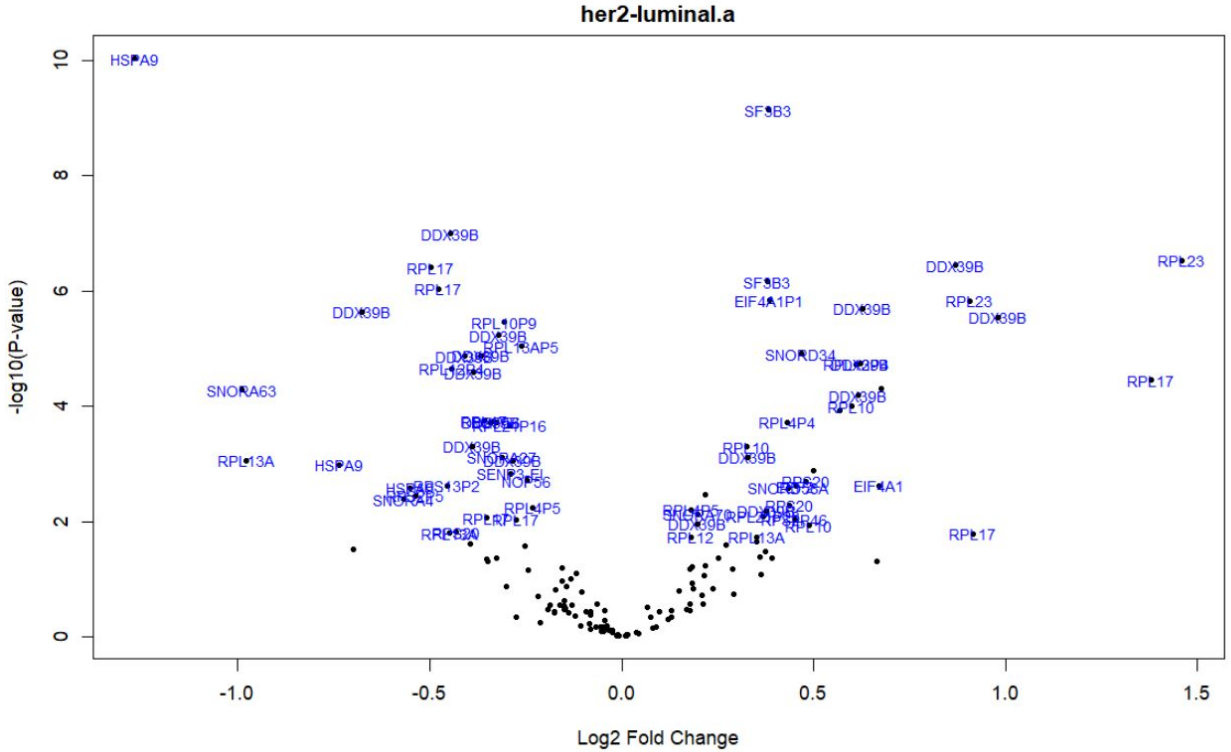
Volcano plot (1/5) - Basal VS HER2



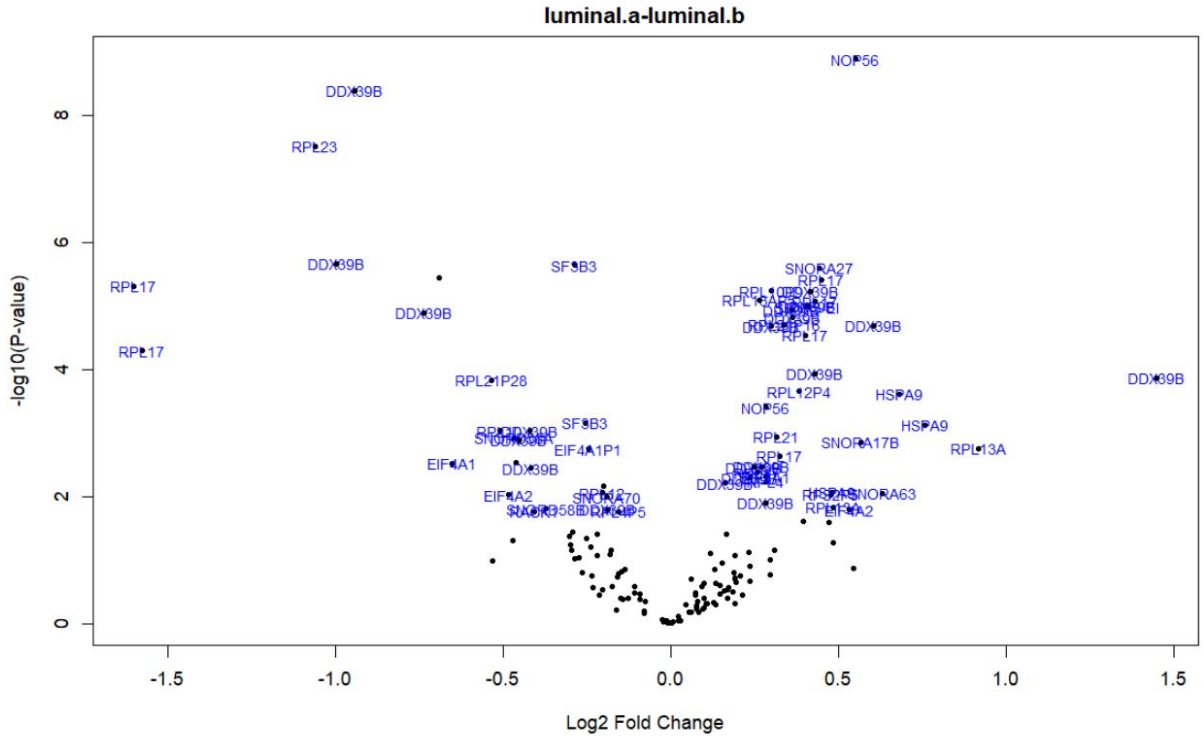
Volcano plot (2/5) - Control VS Basal



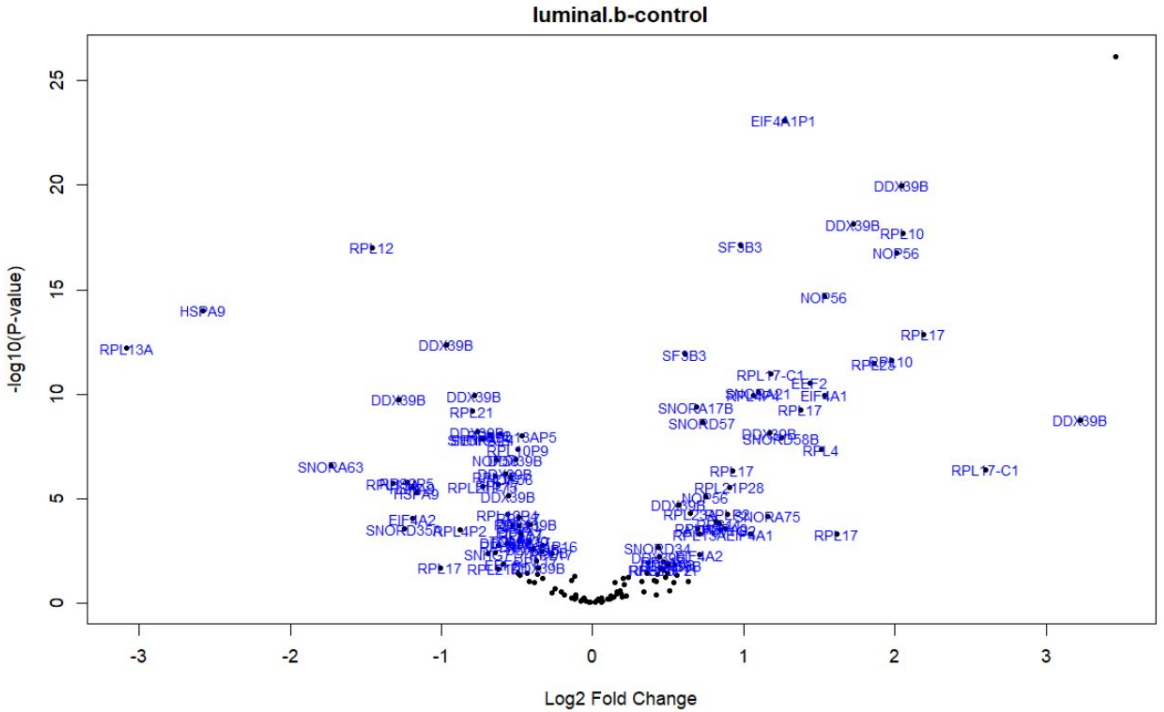
Volcano plot (3/5) - HER2 VS Luminal A



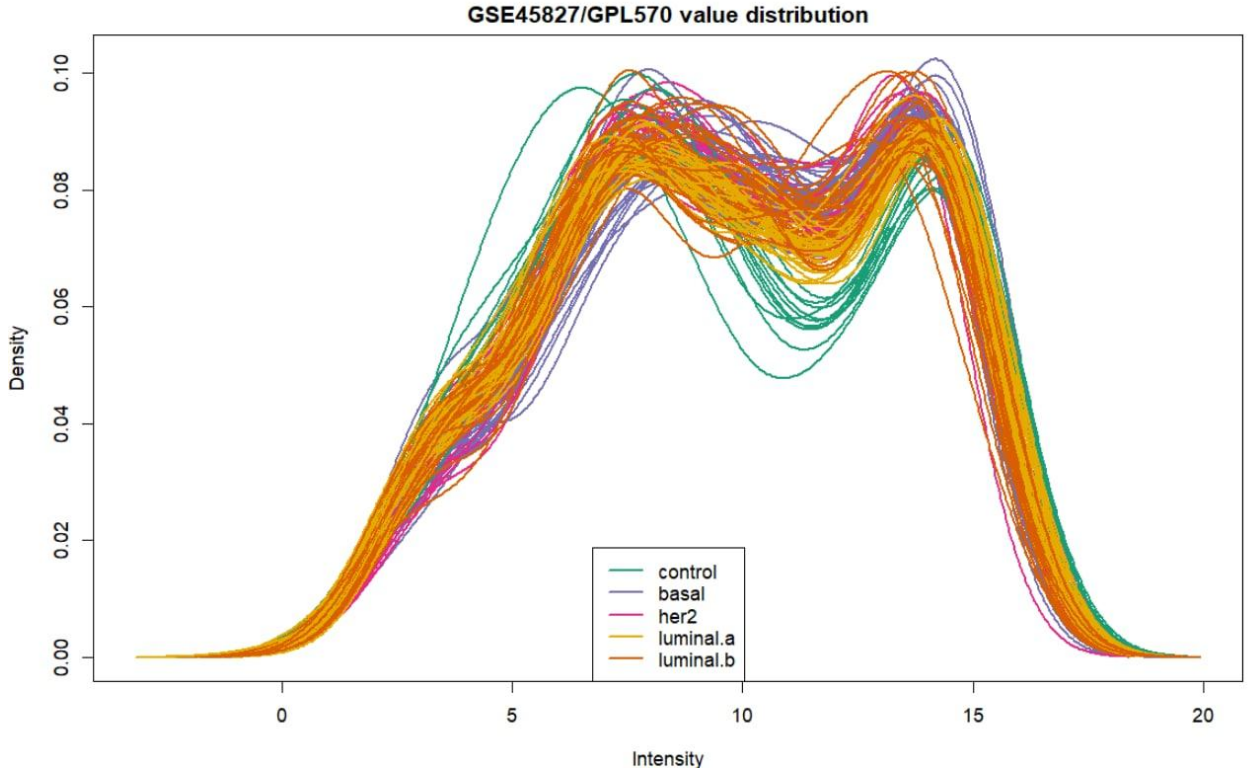
Volcano plot (4/5) - Luminal A VS Luminal B



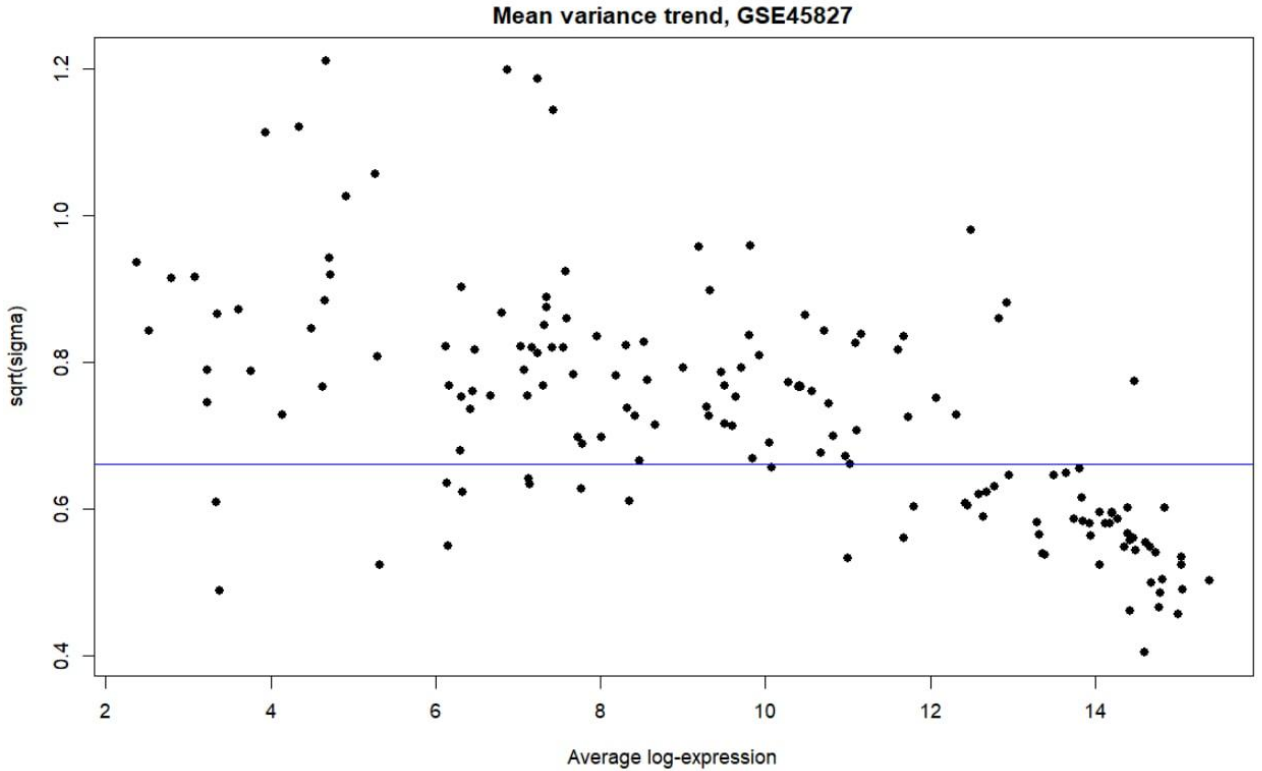
Volcano plot (5/5) - Luminal B VS Control



Expression value distribution



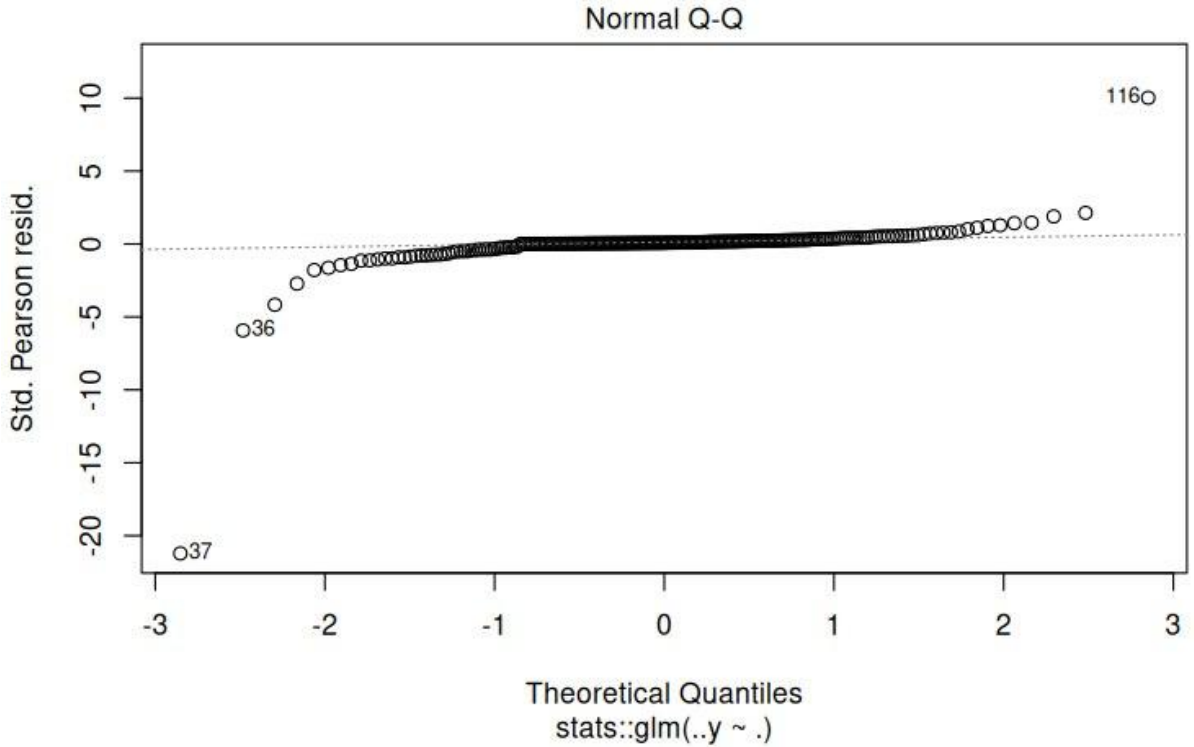
Mean-variance trend



Thanks for your attention

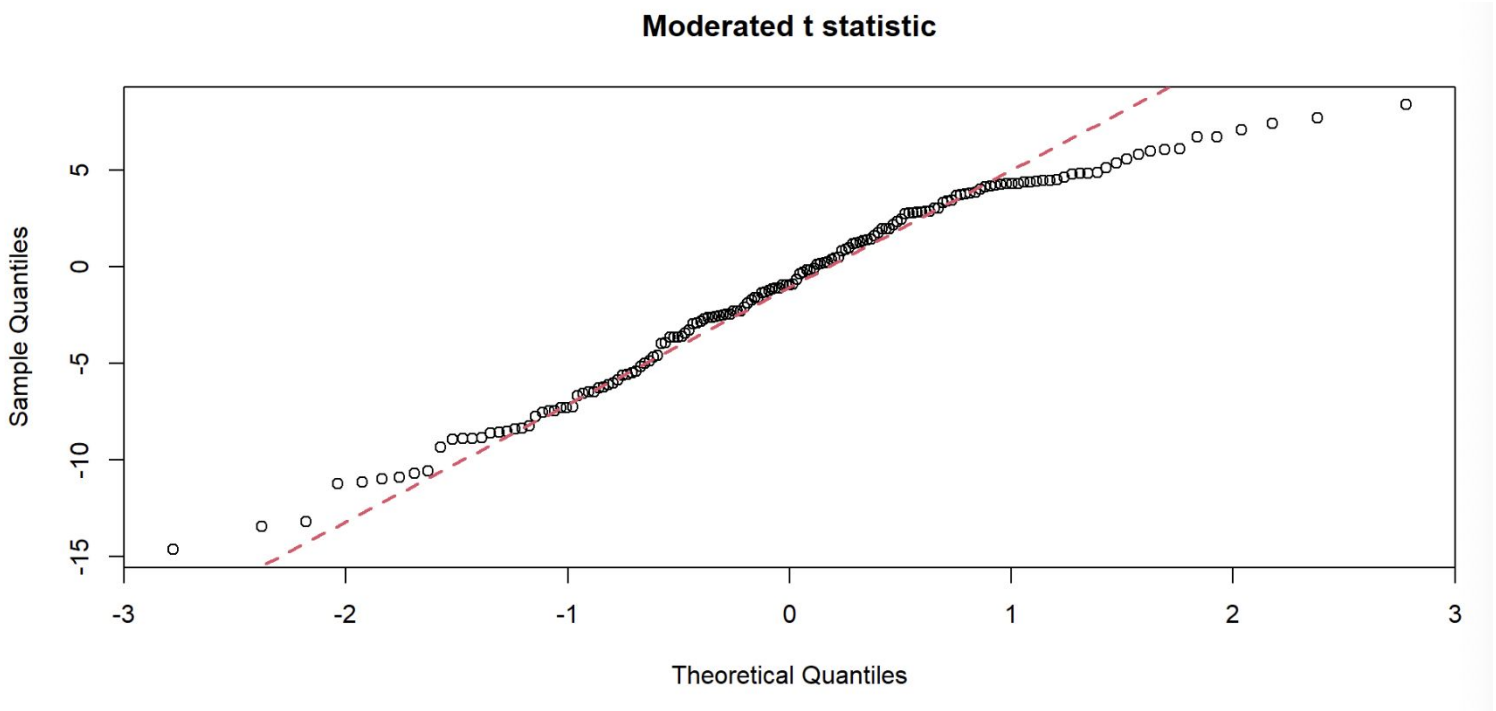
Appendix

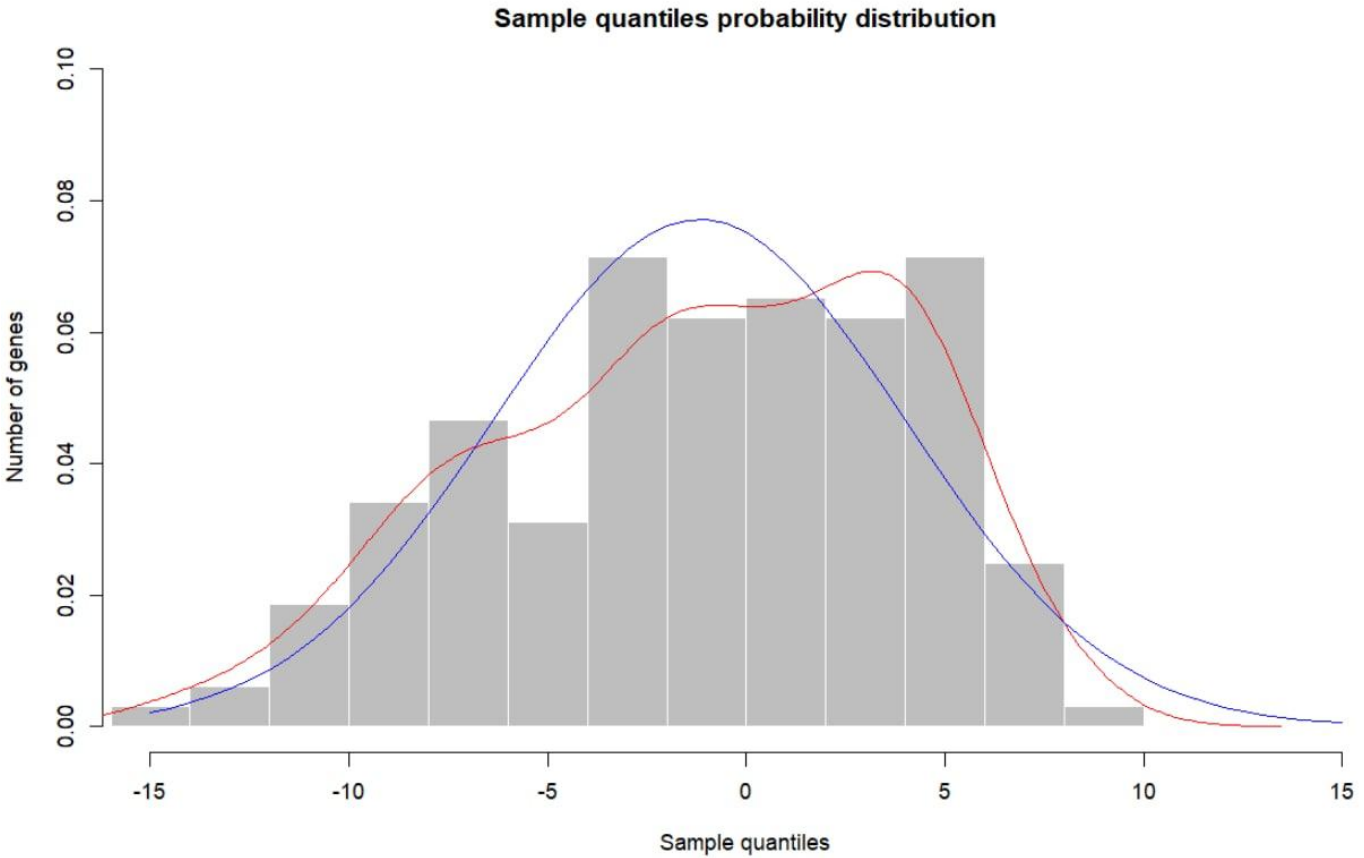
Normal QQ plot



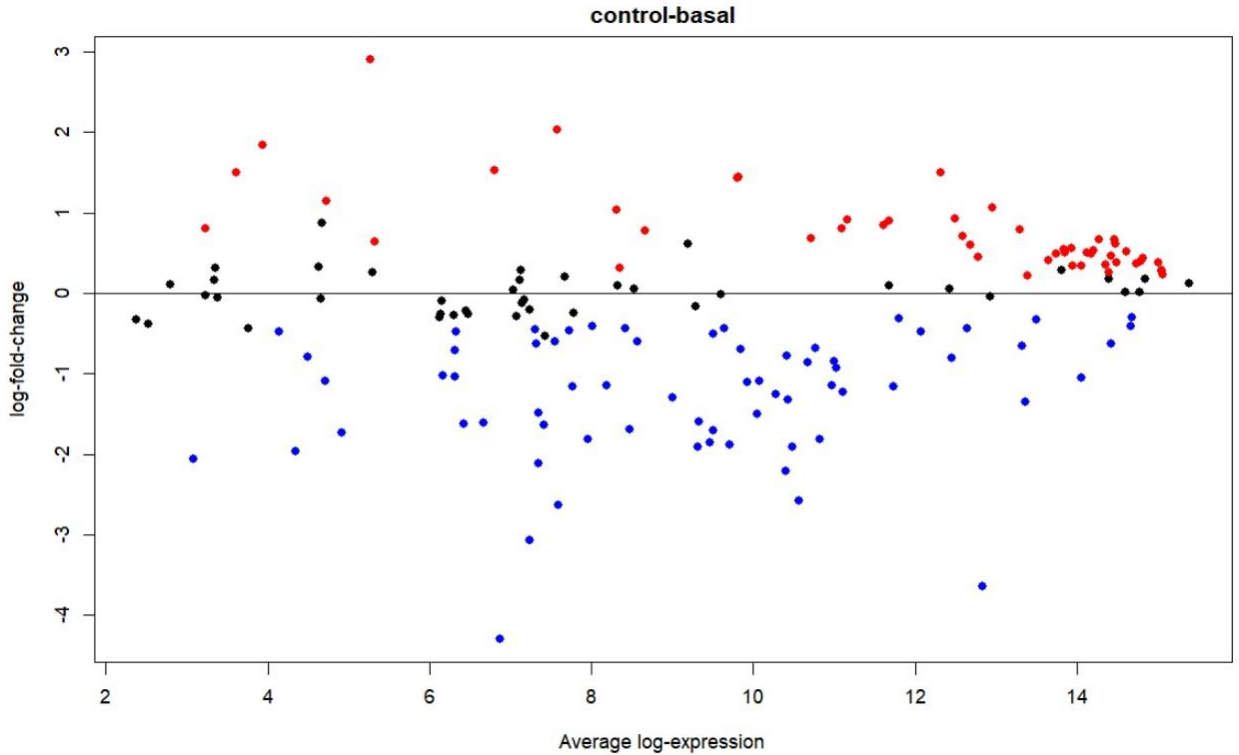
Differential Gene Expression Analysis

Q-Q plot for t-statistic

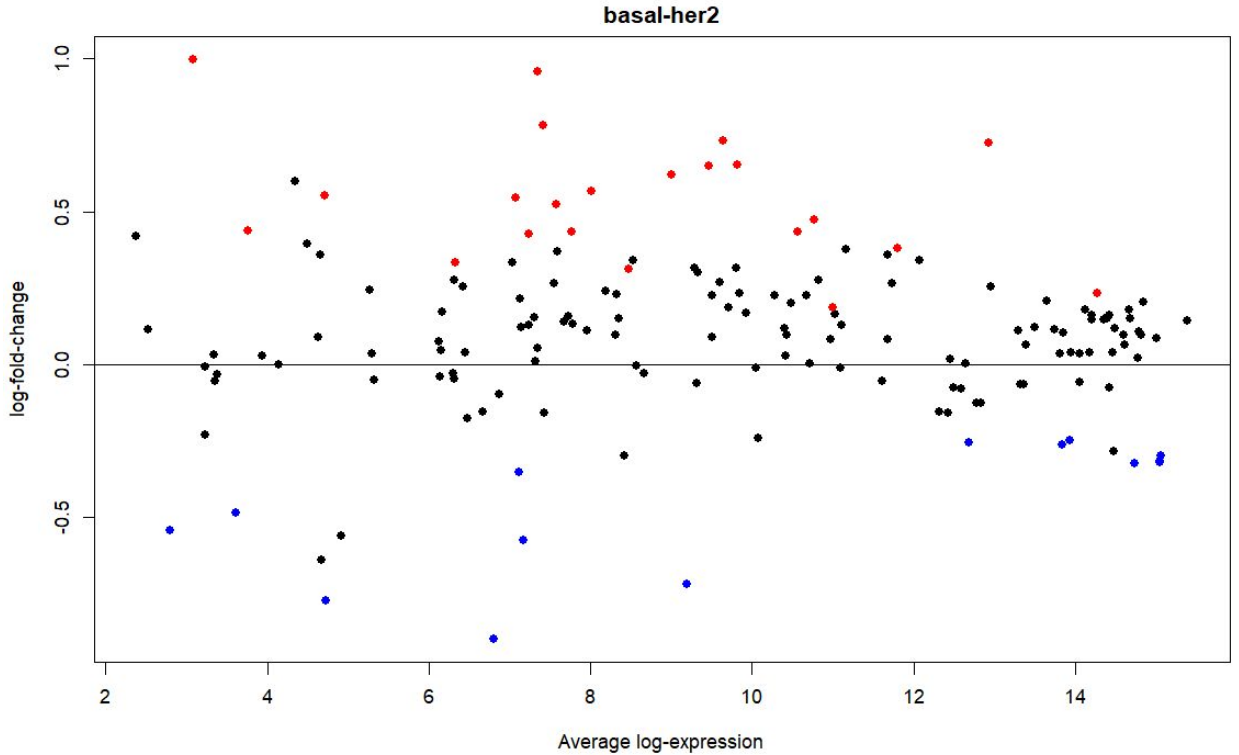



Sample quantiles probability distribution
Number of genes
0.00
0.02
0.04
0.06
0.08
0.10
-15
-10
-5
0
5
10
15
Sample quantiles

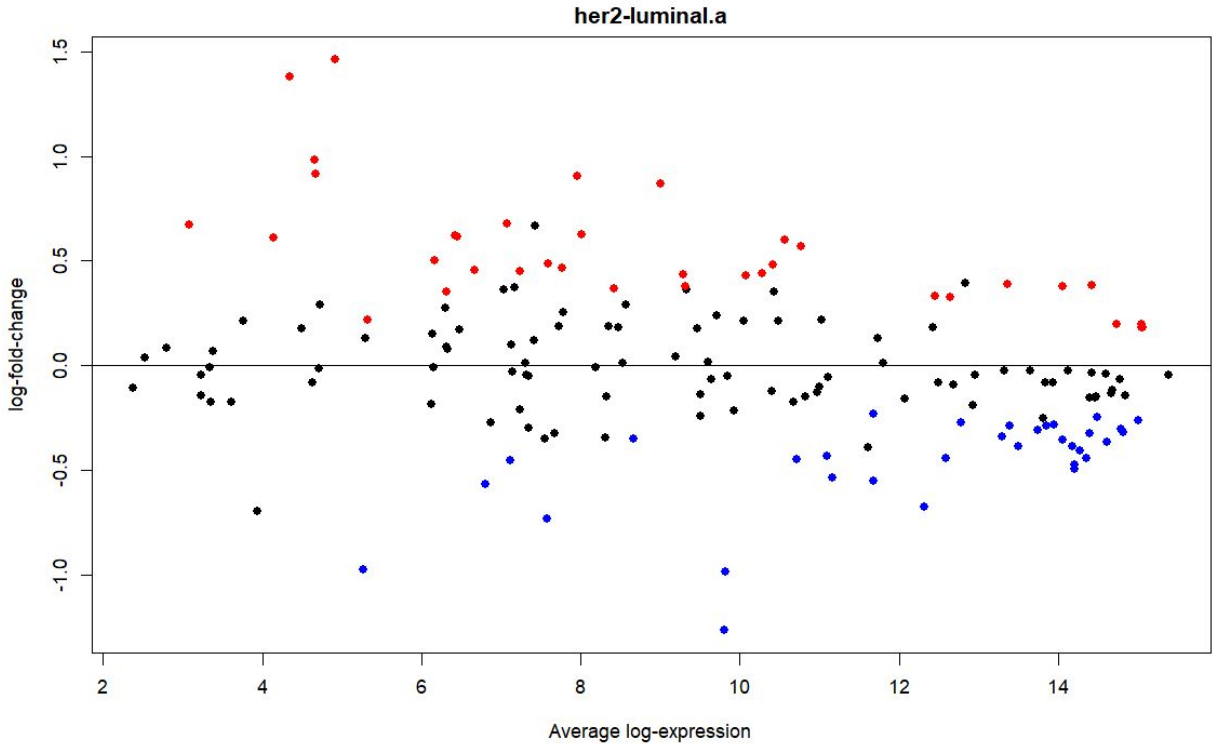
Mean-Difference plot (1/5)



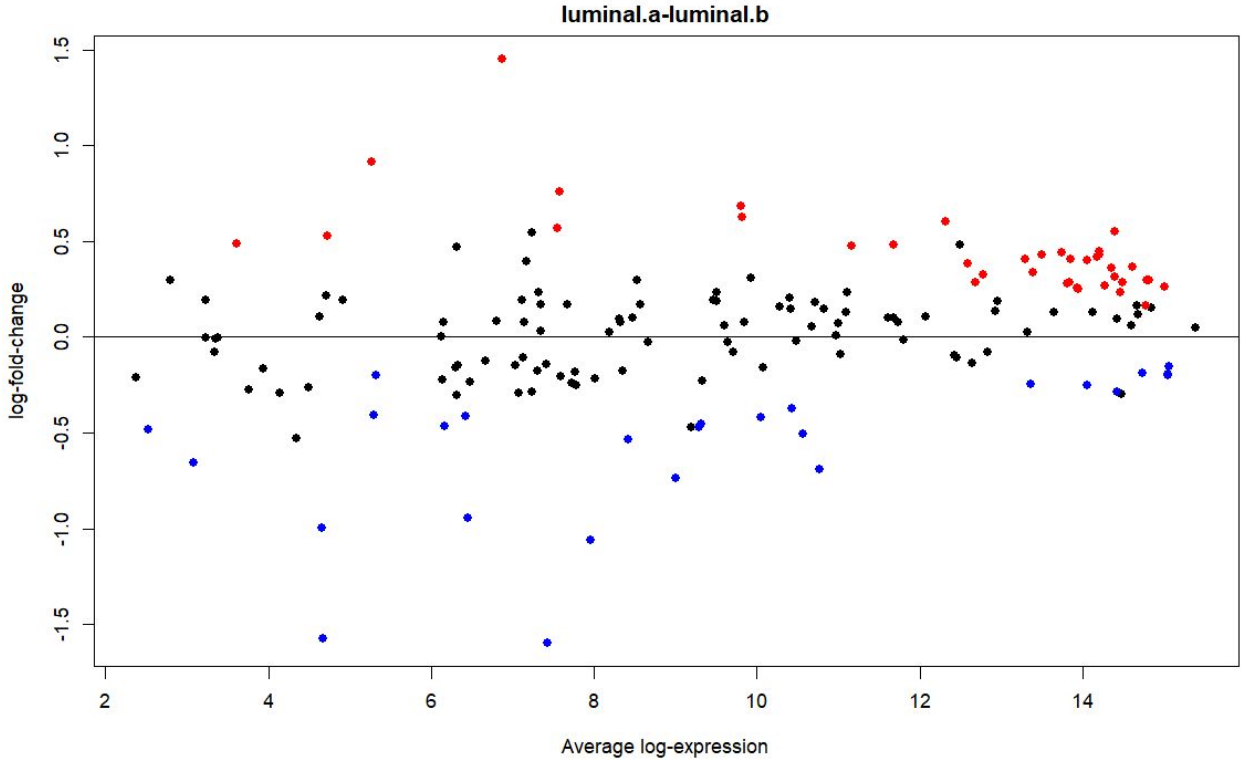
Mean-Difference plot (2/5)



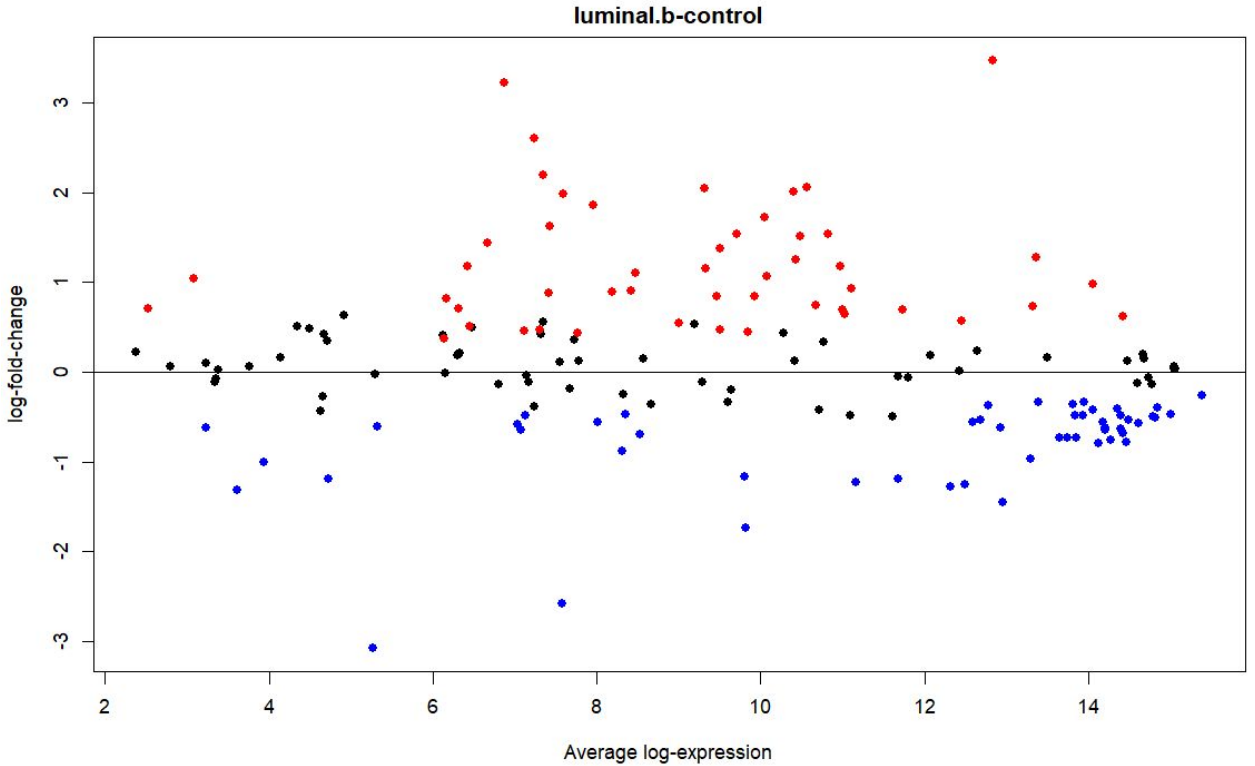
Mean-Difference plot (3/5)



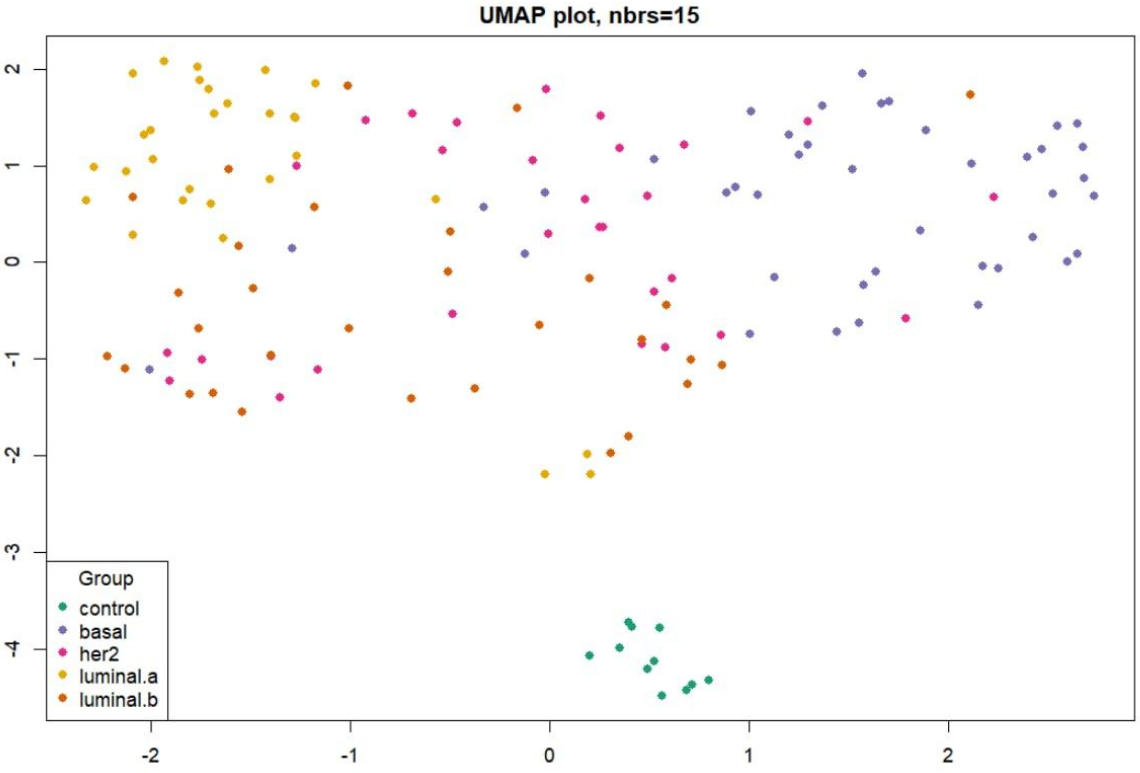
Mean-Difference plot (4/5)



Mean-Difference plot (5/5)

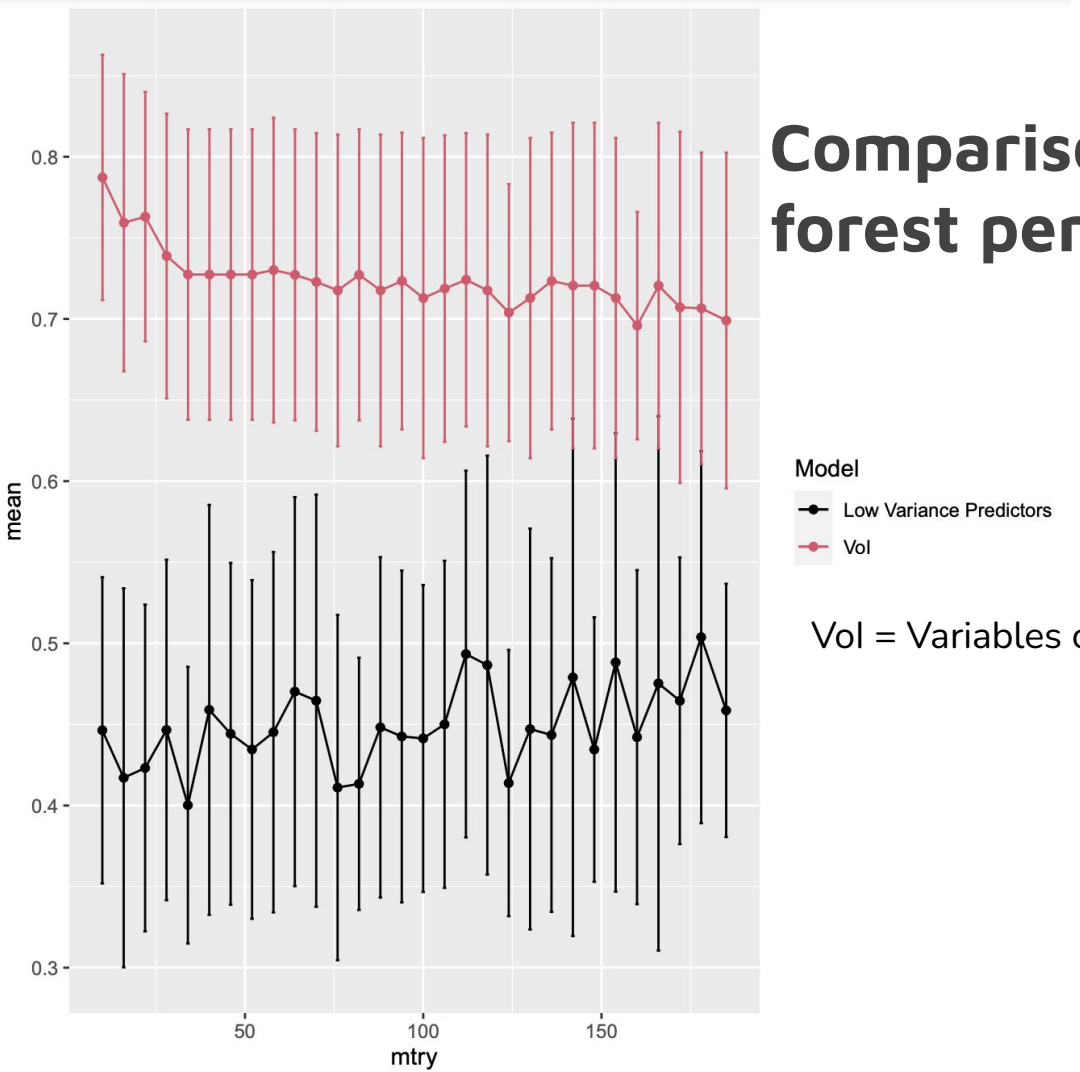


UMAP


UMAP plot, nbrs=15
Group
control
basal
her2
luminal.a
luminal.b
2
1
0
-1
-2
-3
-4
-2
-1
0
1
2

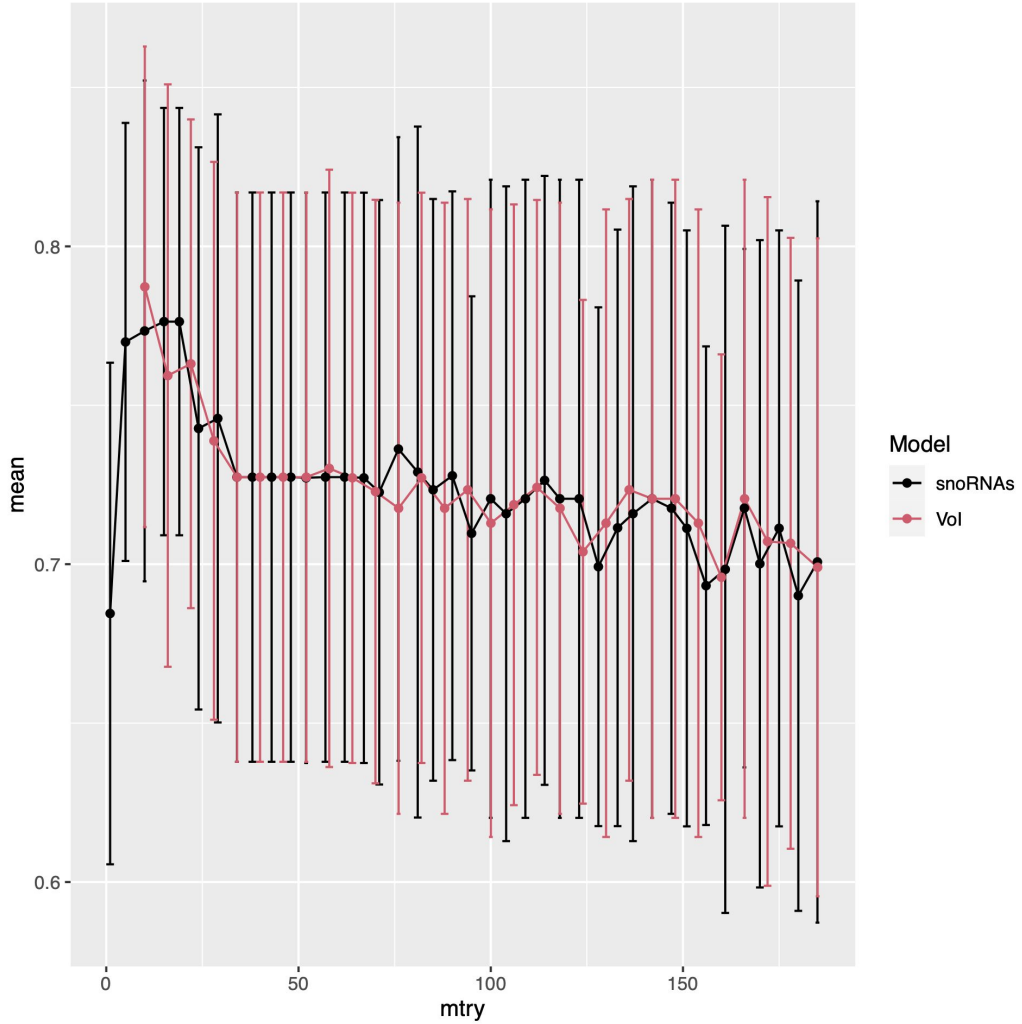
Random Forest

Comparison random forest performance



VoI = Variables of Interest

RF snoRNAs vs variables of interest



Random forest fit on our genes of interest (old data)

		Truth	
		No Cancer	Cancer
Predicted	No Cancer	12	0
	Cancer	3	62

mtry*	Measure	Mean	N. folds	Std. error
10	F1	0.7872726	5	0.03857438
22	F1	0.7630541	5	0.03922816
16	F1	0.7593441	5	0.04674310
28	F1	0.7388160	5	0.04477342
58	F1	0.7301482	5	0.04794219

*mtry is the number of predictors used for the splits

Predicted \ Truth	Basal	HER	L. A	L. B	Normal
Basal	9	2	0	1	0
HER	2	5	0	1	0
L. A	0	1	6	2	0
L. B	0	0	1	4	0
Normal	0	0	0	0	1

mtry*	Measure	Mean	N. folds	Std. error
178	F1	0.5038095	5	0.05853361
112	F1	0.4933707	5	0.05767972
154	F1	0.4882505	5	0.07213041
118	F1	0.4865629	5	0.06589169
142	F1	0.4790185	5	0.08137526

*mtry is the number of predictors used for the splits

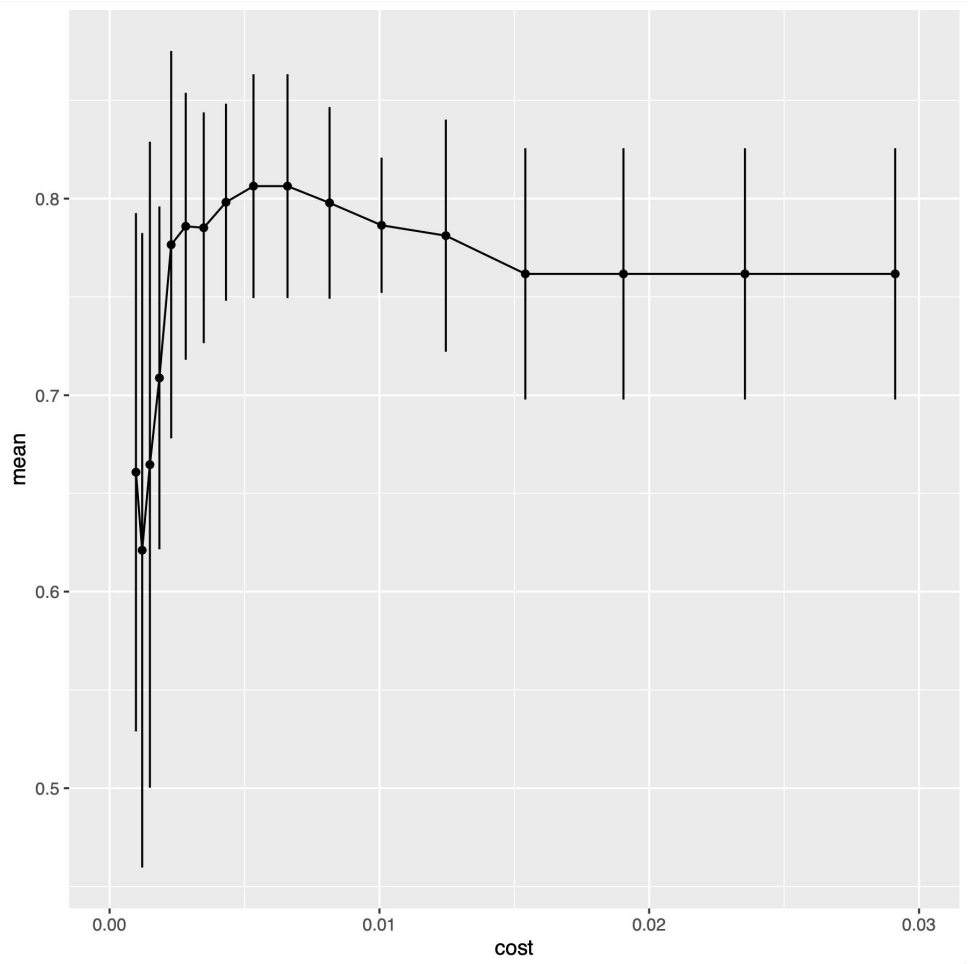
Random forest confusion matrix on lesser variant genes

Predicted	Truth: Basal	HER	L. A	L. B	Normal
Basal	7	6	2	2	0
HER	2	0	1	1	0
L. A	1	1	1	1	0
L. B	1	1	3	4	1
Normal	0	0	0	0	0

cost	Measure	Mean	N. folds	Std. error
0.005332322	F1	0.8063554	5	0.02904139
0.006592777	F1	0.8063554	5	0.02904139
0.004312850	F1	0.7981699	5	0.02556259
0.008151179	F1	0.7978359	5	0.02487277
0.010077956	F1	0.7864484	5	0.01755194

SVM

SVM cost

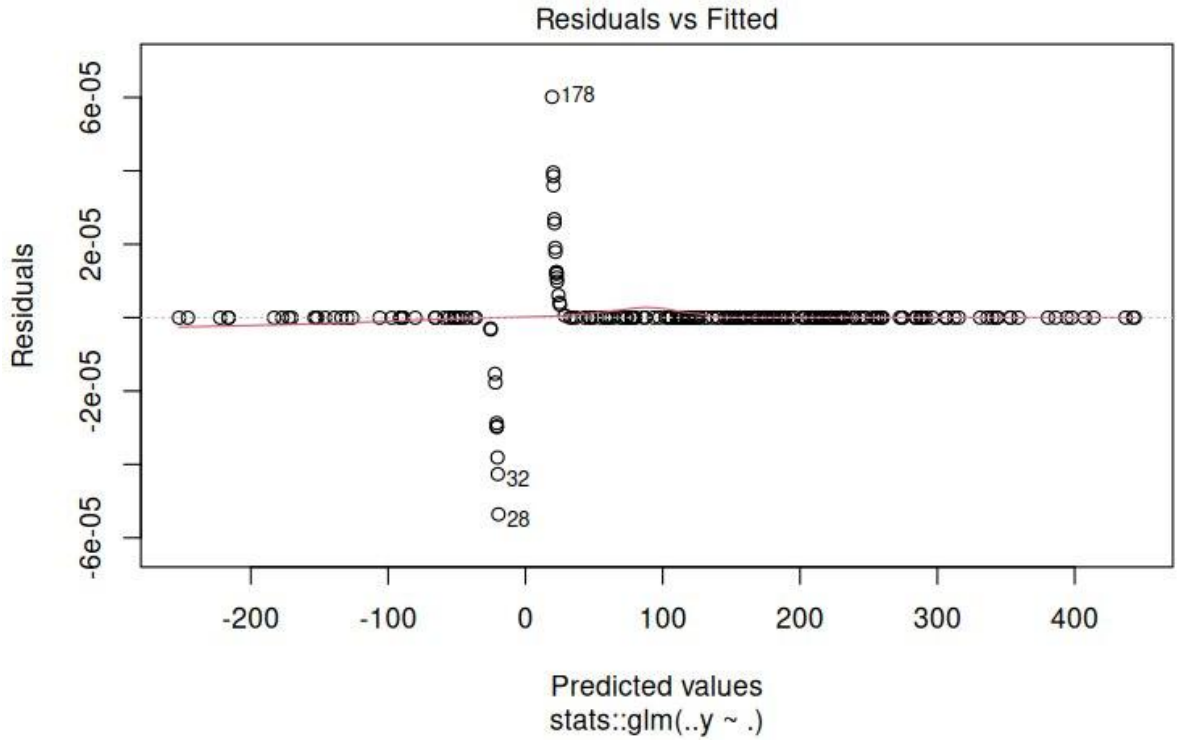


Final SVM model confusion matrix

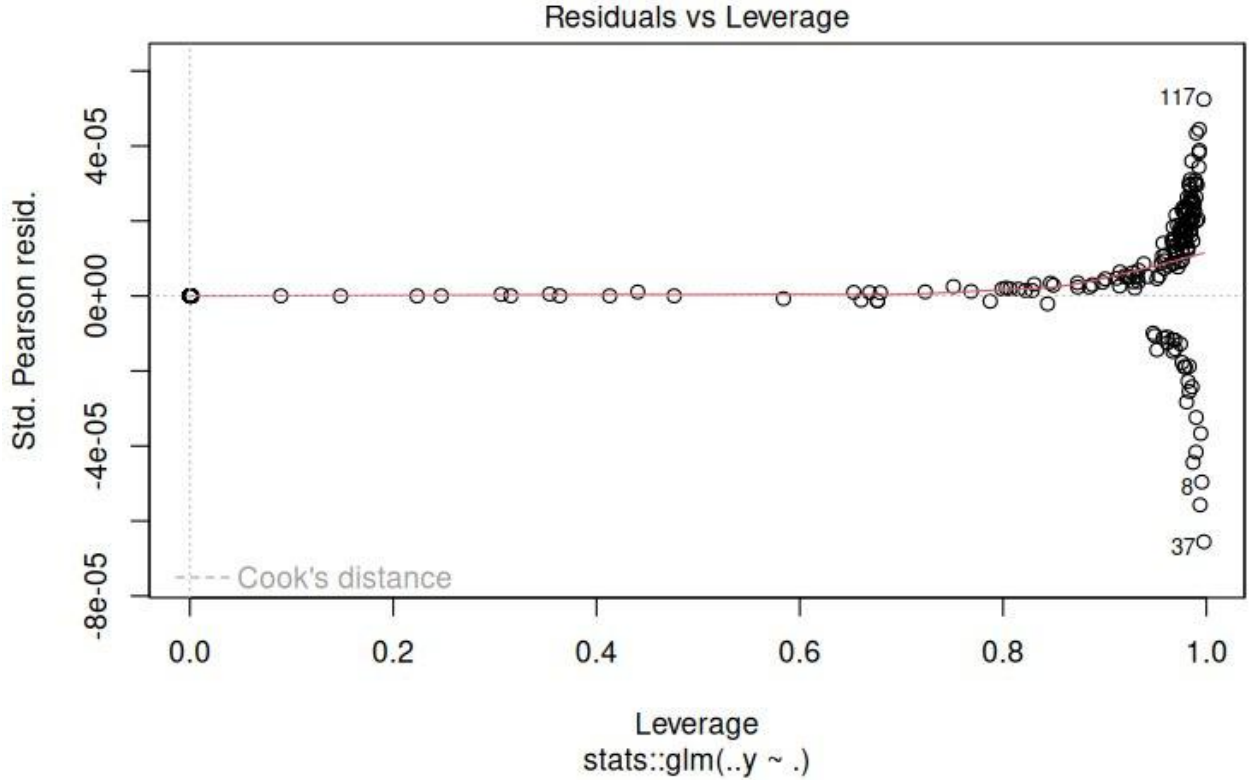
Predicted \ Truth	Basal	HER	L. A	L. B	Normal
Basal	10	2	0	1	0
HER	1	6	0	1	0
L. A	0	0	7	2	0
L. B	0	0	0	4	0
Normal	0	0	0	0	1

Logistic Regression

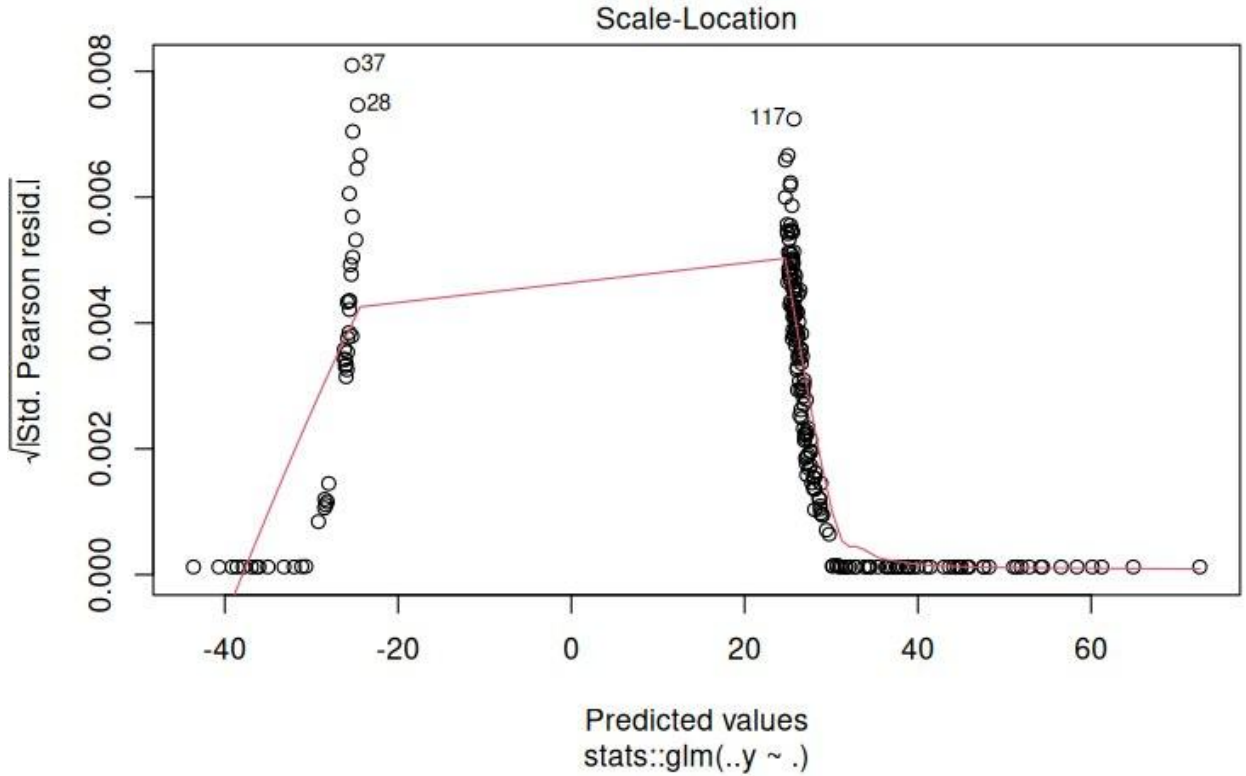
Residuals vs fitted old



Reslev old


Residuals vs Leverage
Std. Pearson resid.
4e-05
0e+00
-4e-05
-8e-05
117
8
37
Cook's distance
0.0
0.2
0.4
0.6
0.8
1.0
Leverage
stats::glm(..y ~ .)

Scaleloc old



Logistic regression with L1 penalty crossvalidation results

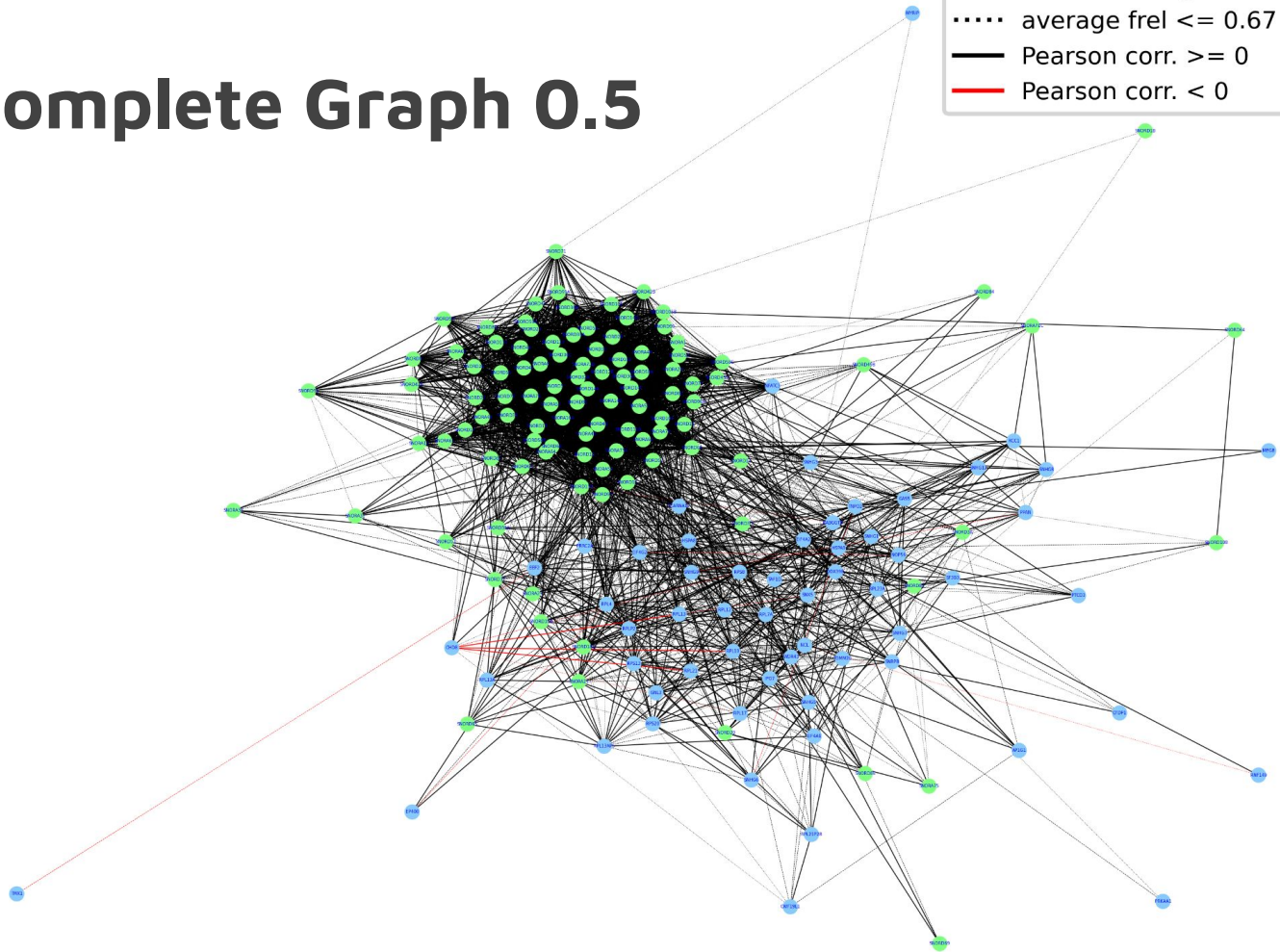
penalty	Measure	Mean	N. folds	Std. error
6.250552e-01	F1	0.8195279	5	0.04381708
1.264855e-05	F1	0.8137864	5	0.03345629
1.206793e-06	F1	0.8099060	5	0.04936689
1.757511e-08	F1	0.8039975	5	0.04661561
6.866488e-09	F1	0.8030488	5	0.04050916

Logistic regression with L1 penalty confusion matrix

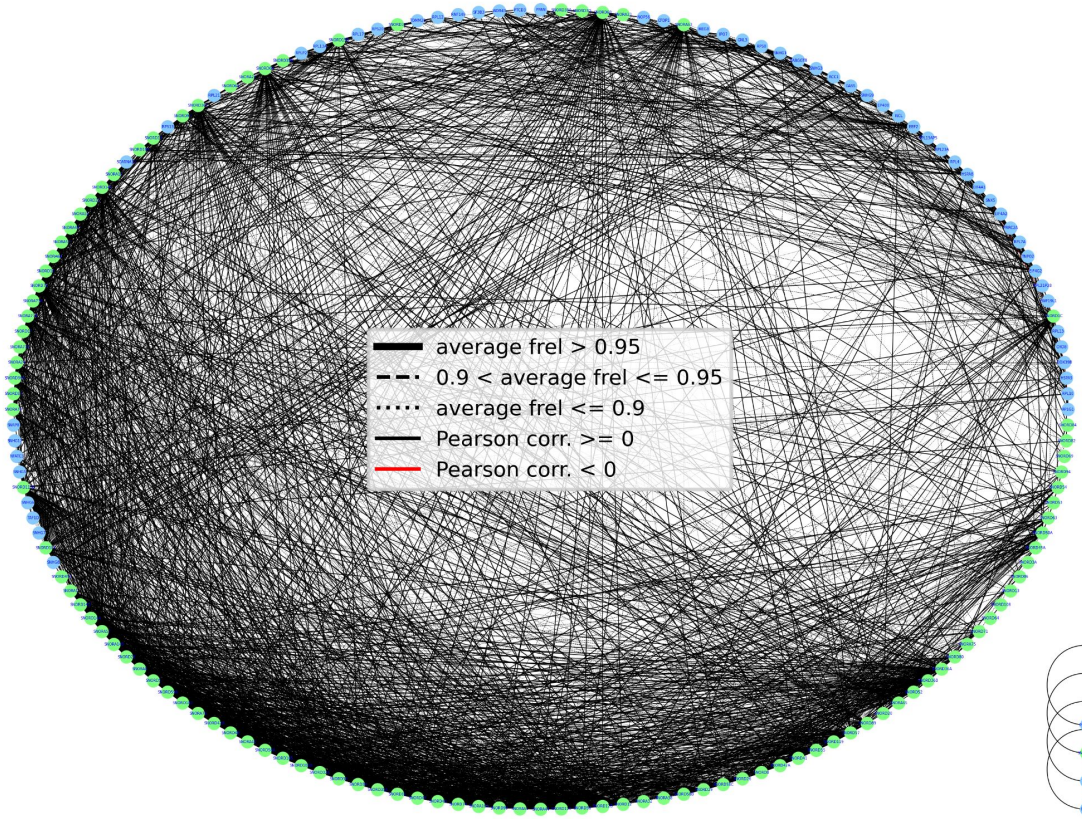
		Truth				
		Basal	HER	L. A	L. B	Normal
Predicted	Basal	9	1	0	1	0
	HER	2	5	0	0	0
	L. A	0	1	6	3	0
	L. B	0	1	1	4	0
	Normal	0	0	0	0	1

Network analysis

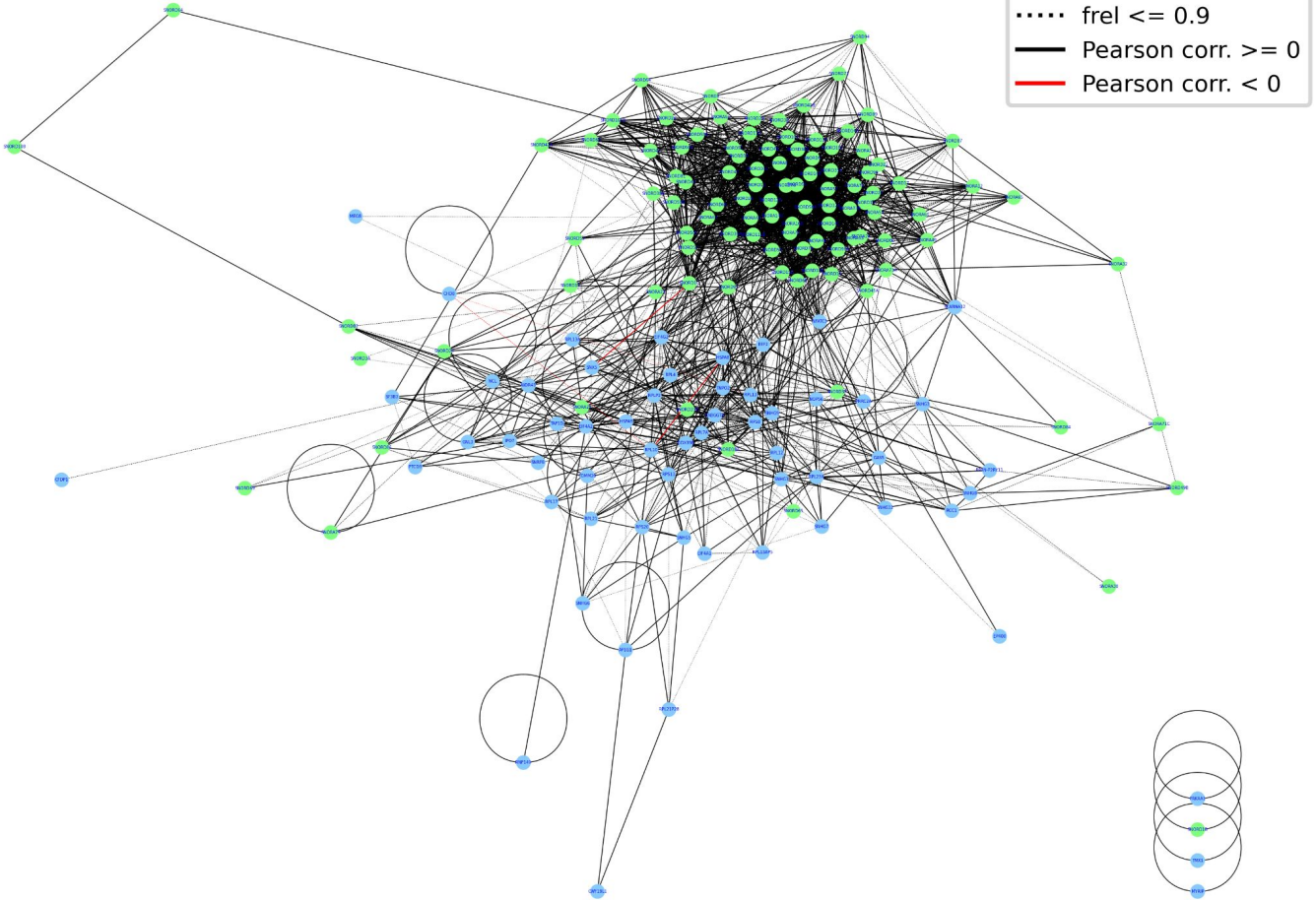
Complete Graph 0.5



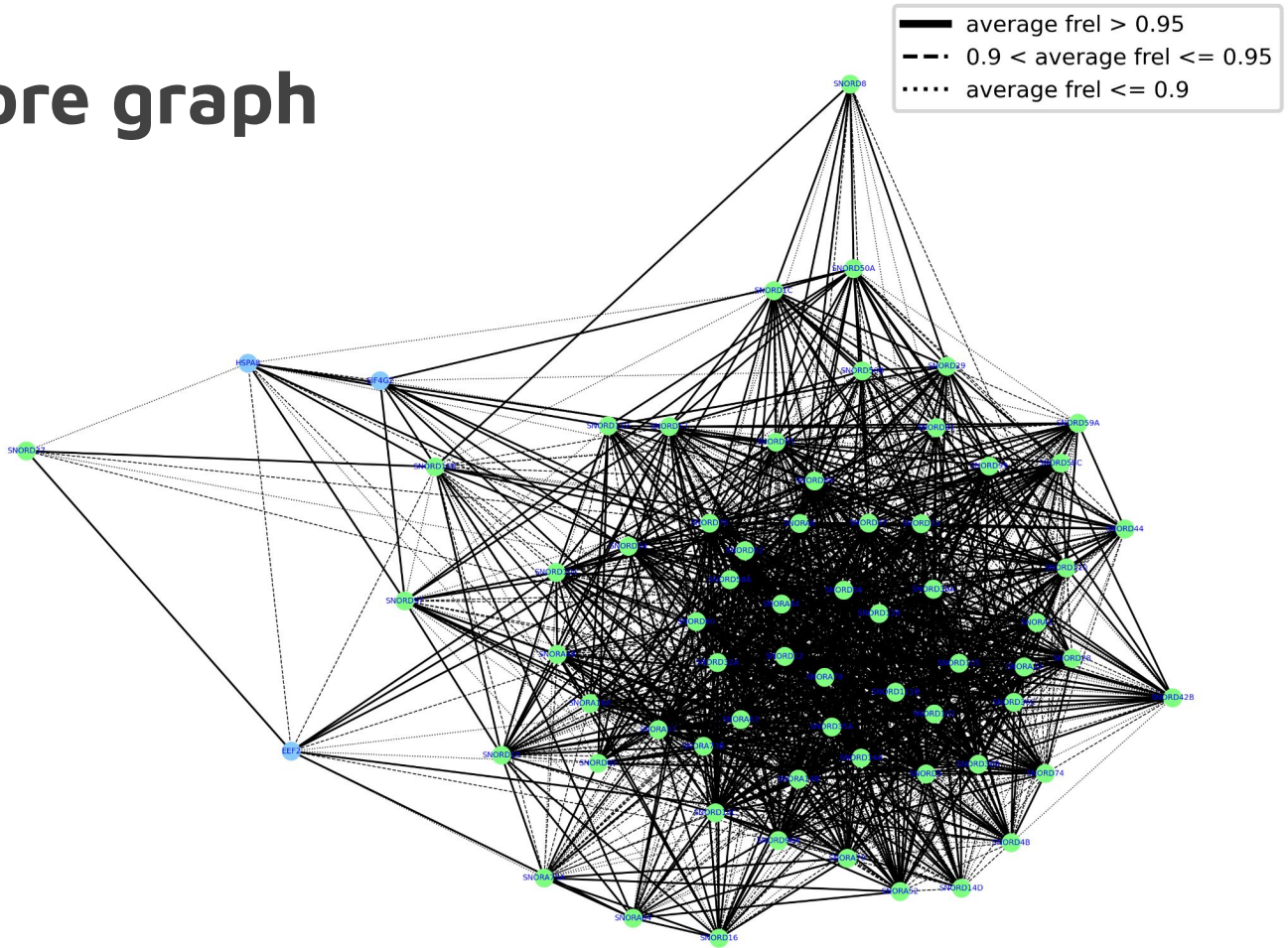
Complete Graph Circular



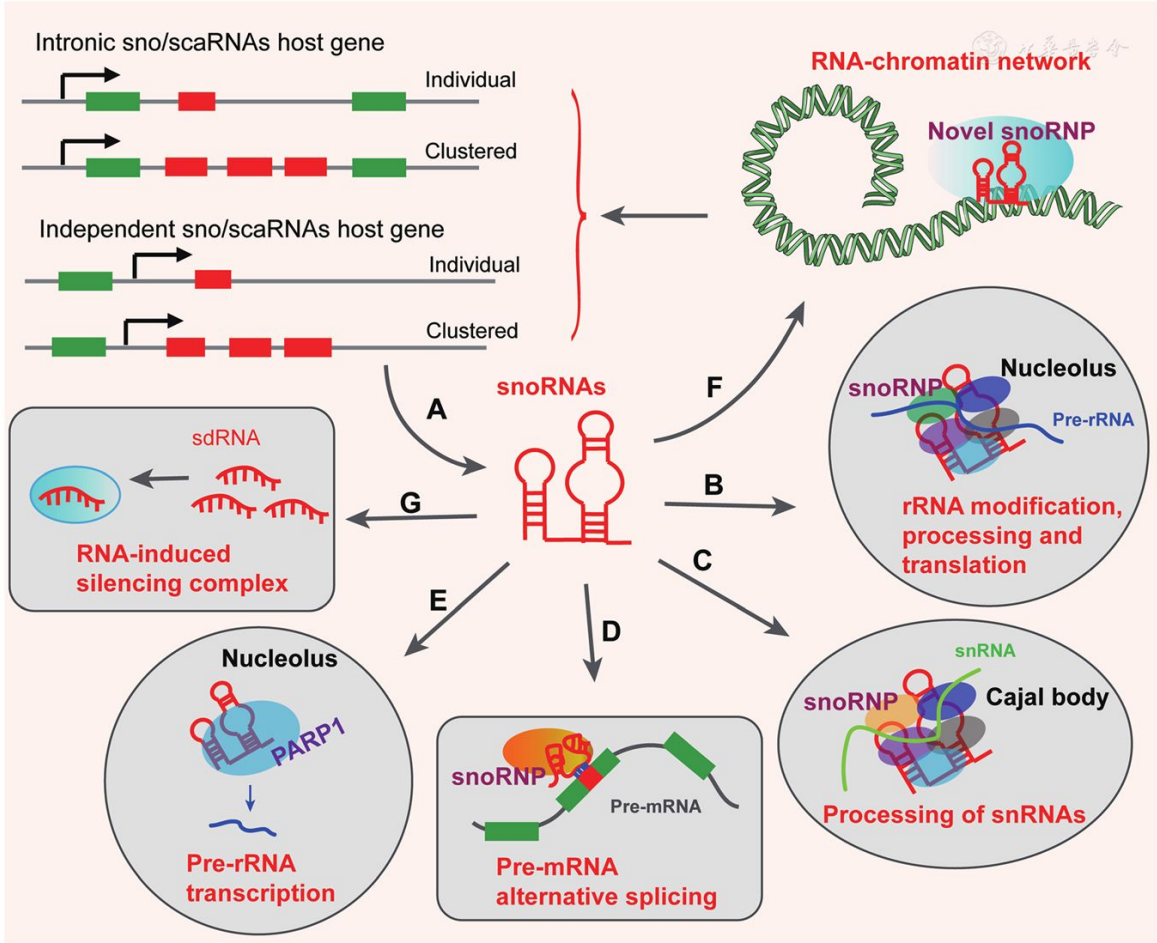
Complete Graph not ignored



Core graph



Theory part


Intronic sno/scaRNAs host gene
Individual
Clustered
Independent sno/scaRNAs host gene
Individual
Clustered
RNA-chromatin network
Novel snoRNP
snoRNAs
A
B
C
D
E
F
G
sdRNA
RNA-induced silencing complex
Nucleolus
snoRNP
Pre-rRNA
rRNA modification, processing and translation
Nucleolus
PARP1
Pre-rRNA transcription
snoRNP
Pre-mRNA
Pre-mRNA alternative splicing
snRNA
snoRNP
Cajal body
Processing of snRNAs

Breast cancer subtypes

Luminal A breast cancer is estrogen receptor-positive and progesterone receptor-positive, HER2-negative, and has low levels of the protein Ki-67, which helps control how fast cancer cells grow. Luminal A cancers tend grow more slowly than other cancers, be lower grade, and have a good prognosis.

Luminal B breast cancer is estrogen receptor-positive and HER2-negative, and also has either high levels of Ki-67 (which indicate faster growth of cancer cells) or is progesterone receptor-negative.

Breast cancer subtypes

HER2-enriched breast cancer is estrogen receptor-negative and progesterone receptor-negative and HER2-positive. HER2-enriched cancers tend to grow faster than luminal cancers and can have a worse prognosis, but are usually successfully treated with targeted therapy medicines aimed at the HER2 protein.

Triple-negative or basal-like breast cancer is estrogen receptor-negative, progesterone receptor-negative, and HER2-negative.

Triple-negative breast cancer is considered more aggressive than either luminal A or luminal B breast cancer.

Font: https://www.breastcancer.org/types/molecular-subtypes