Final Project

Xulin Ge

AS.410.671.82.SU22 Gene Expression Data Analysis and Visualization

Data set: GEO accession GSE56323

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56323

Introduction & Background

- Wilms Tumor is the most common pediatric kidney cancer that evolves from the failure of terminal differentiation of the embryonic kidney
- The authors used mice model and hypothesized that the regulator Lin28 during kidney development plays a role in cancer development
- The researchers performed gene expression analysis on a total of 8 kidneys samples, including 4 tumors samples from Lin28 transgenic mice and 4 control kidneys
- Platform: GPL6885 Illumina MouseRef-8 v2.0 expression beadchip
- Dataset: GSE56323 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56323
- GEO2R analysis: GSE56323 https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE56323

Question in this project: Find out the genes that show expression difference between tumor and control. Is Lin 28 one of them?

Ref:Urbach A, Yermalovich A, Zhang J, Spina CS, Zhu H, Perez-Atayde AR, Shukrun R, Charlton J, Sebire N, Mifsud W, Dekel B, Pritchard-Jones K, Daley GQ. Lin28 sustains early renal progenitors and induces Wilms tumor. Genes Dev. 2014 May 1;28(9):971-82. doi: 10.1101/gad.237149.113. Epub 2014 Apr 14. PMID: 24732380; PMCID: PMC4018495.

Methods

- Retrieve dataset from GEO
  Take a brief look at GEO2R analysis

- Use 3 plots to test for outliers:
  Correlation plot (heat map)
  Hierarchical clustering dendrogram
  Average correlation plot

- Plot the gene expression Densities. Filter out low expression genes that <=0

- Perform Student's two-sample t-test on all genes.
  Because there're two groups (tumor vs. control).

- Calculate the fold change between the groups.

Methods

- Select genes that match two criterions:
  1. p-value < Bonferroni threshold
  2. linear |fold change|>2

  And plot the scores of retained genes

- PCA analysis on subset data and

  Plot the first two components in PCA plot.

  Plot a two-dimensional embedding of the weighted graph Laplacian

- k-means clustering of the subset data and the original data (k=2)

  Plot samples in text

  Colored from k-means cluster membership

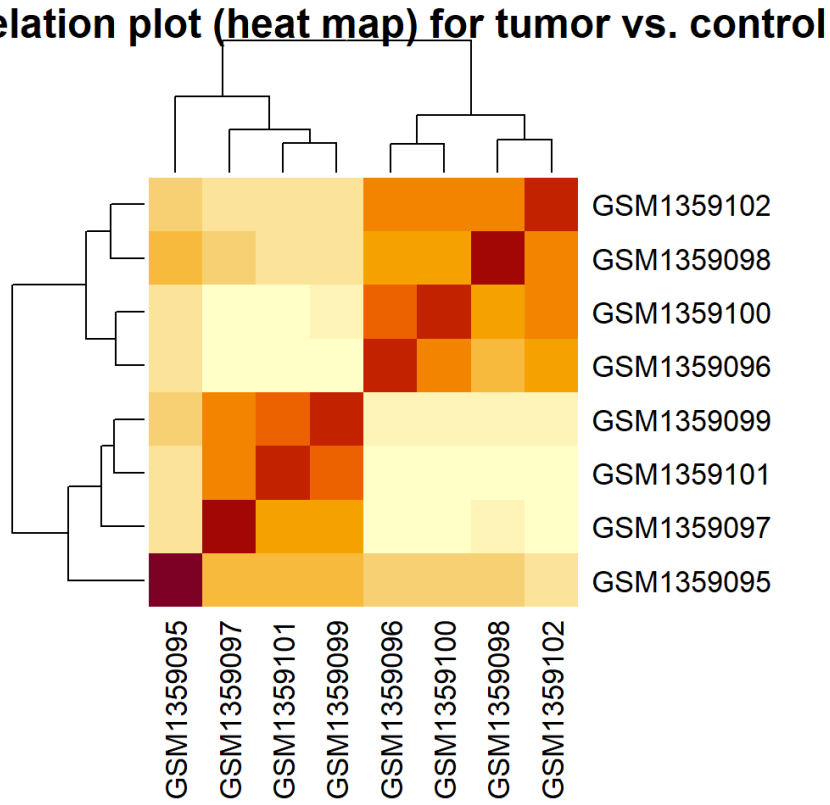
- Report gene functional information on NCBI's DAVID

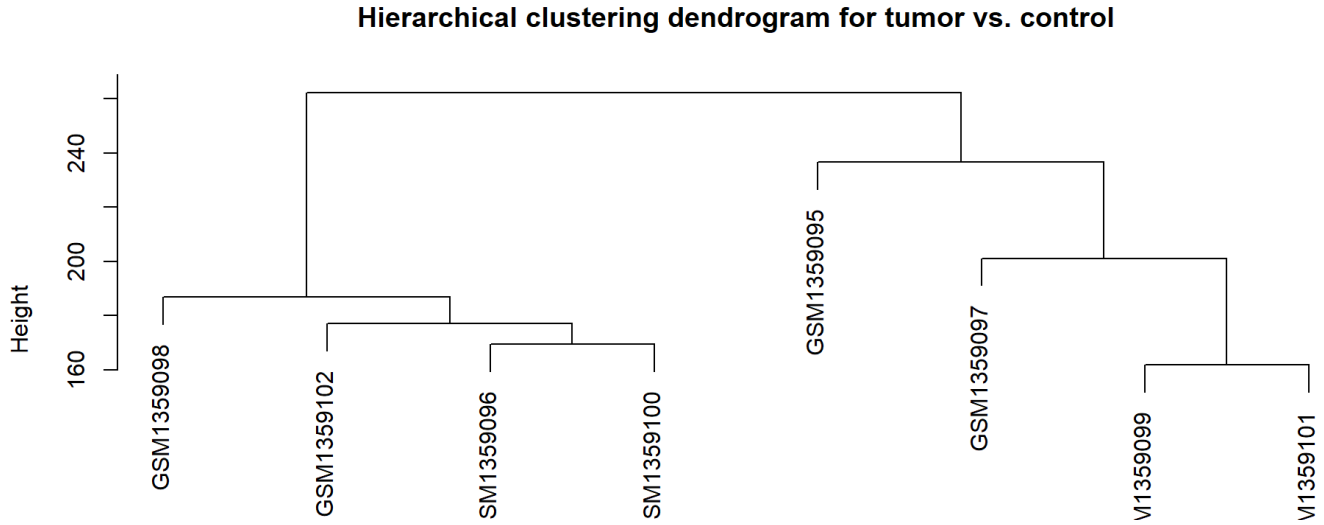
Results - GEO2R

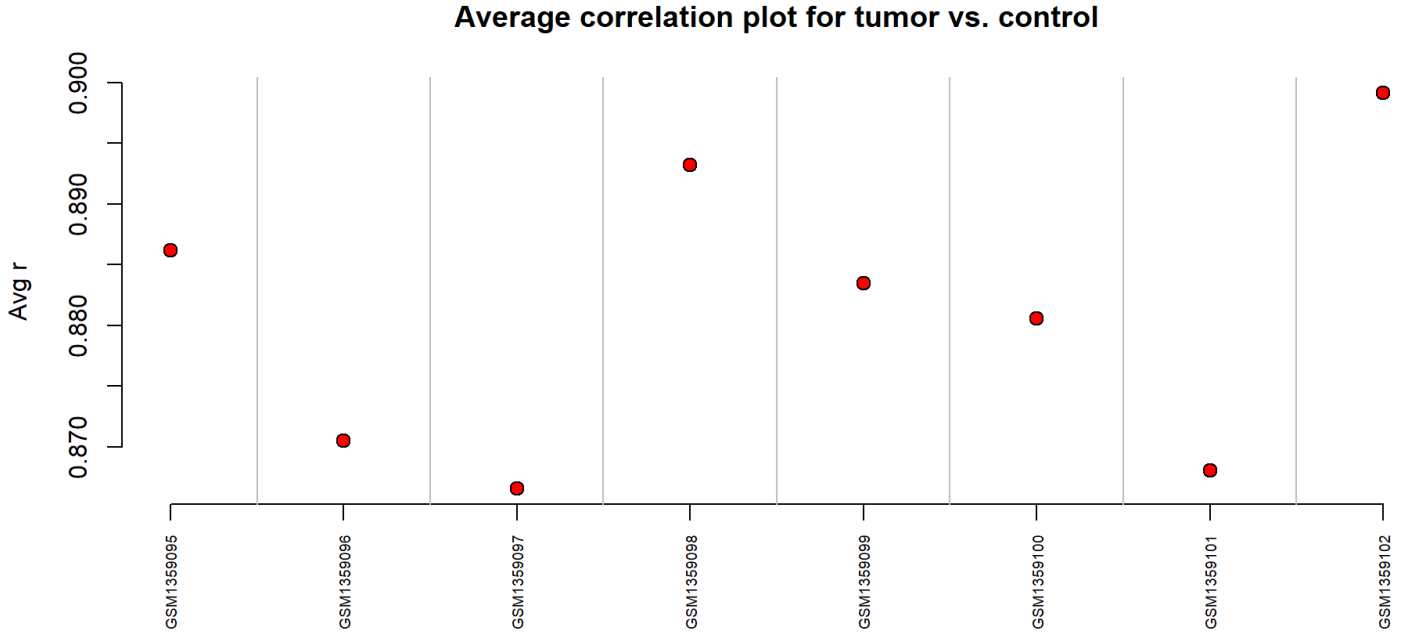
Group	Accession	Title	Source name	Tissue	Tumor status	Age
1	GSM1359095	kidney tumor #1	kidney tumor #1	kidney	tumorous	5W
2	GSM1359096	control kidney #1	control kidney #1	kidney	non-tumorous	5W
1	GSM1359097	kidney tumor #2	kidney tumor #2	kidney	tumorous	4M
2	GSM1359098	control kidney #2	control kidney #2	kidney	non-tumorous	4M
1	GSM1359099	kidney tumor #3	kidney tumor #3	kidney	tumorous	4M
2	GSM1359100	control kidney #3	control kidney #3	kidney	non-tumorous	4M
1	GSM1359101	kidney tumor #4	kidney tumor #4	kidney	tumorous	4M
2	GSM1359102	control kidney #4	control kidney #4	kidney	non-tumorous	4M

The dataset includes the class structure with different levels (tumor vs. control)

Results - Test for outlier samples



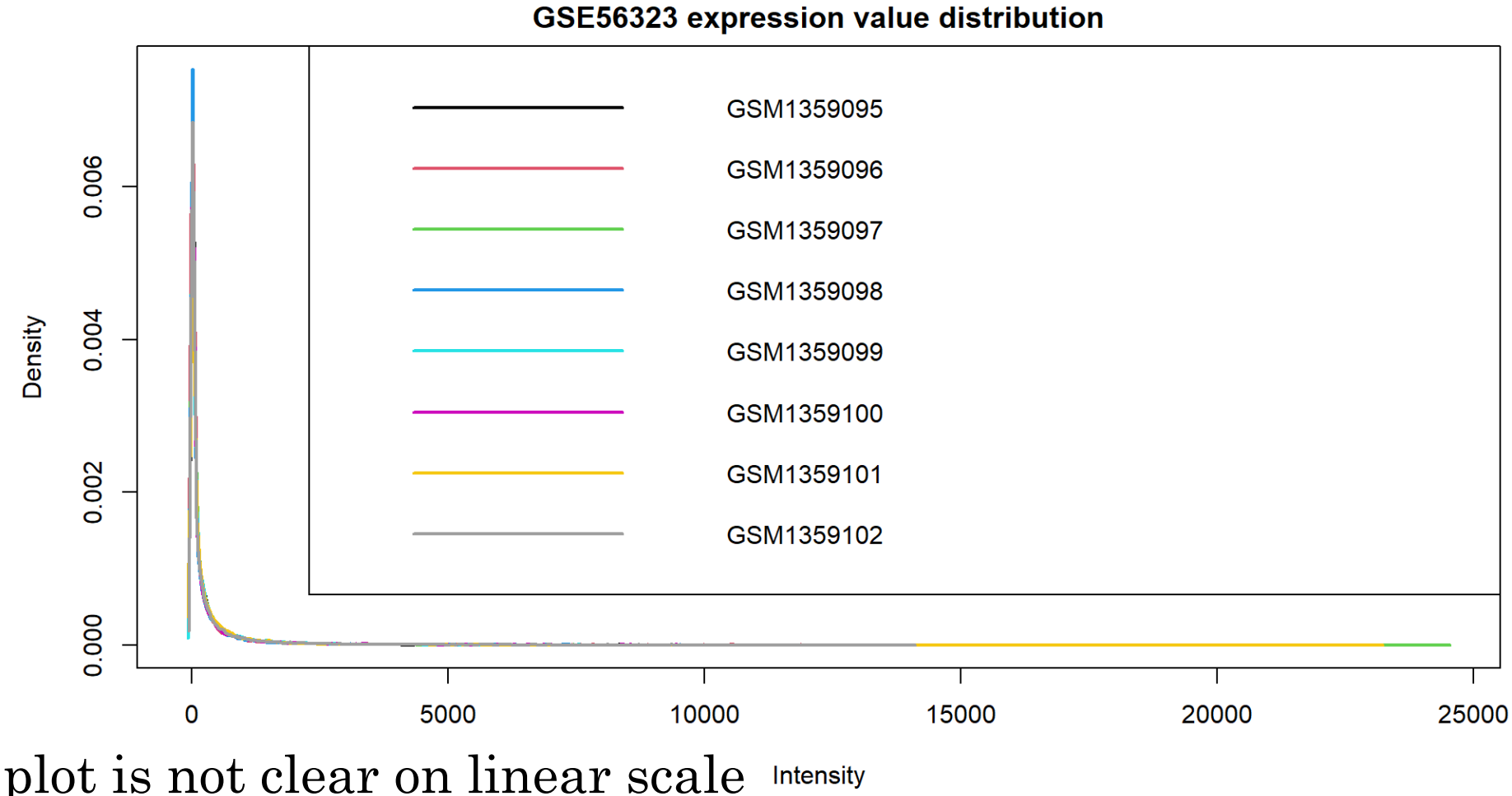




- There's no outlier in the dataset, based on three plots
- No need to remove outlier(s)

Results - Filter out low expression genes

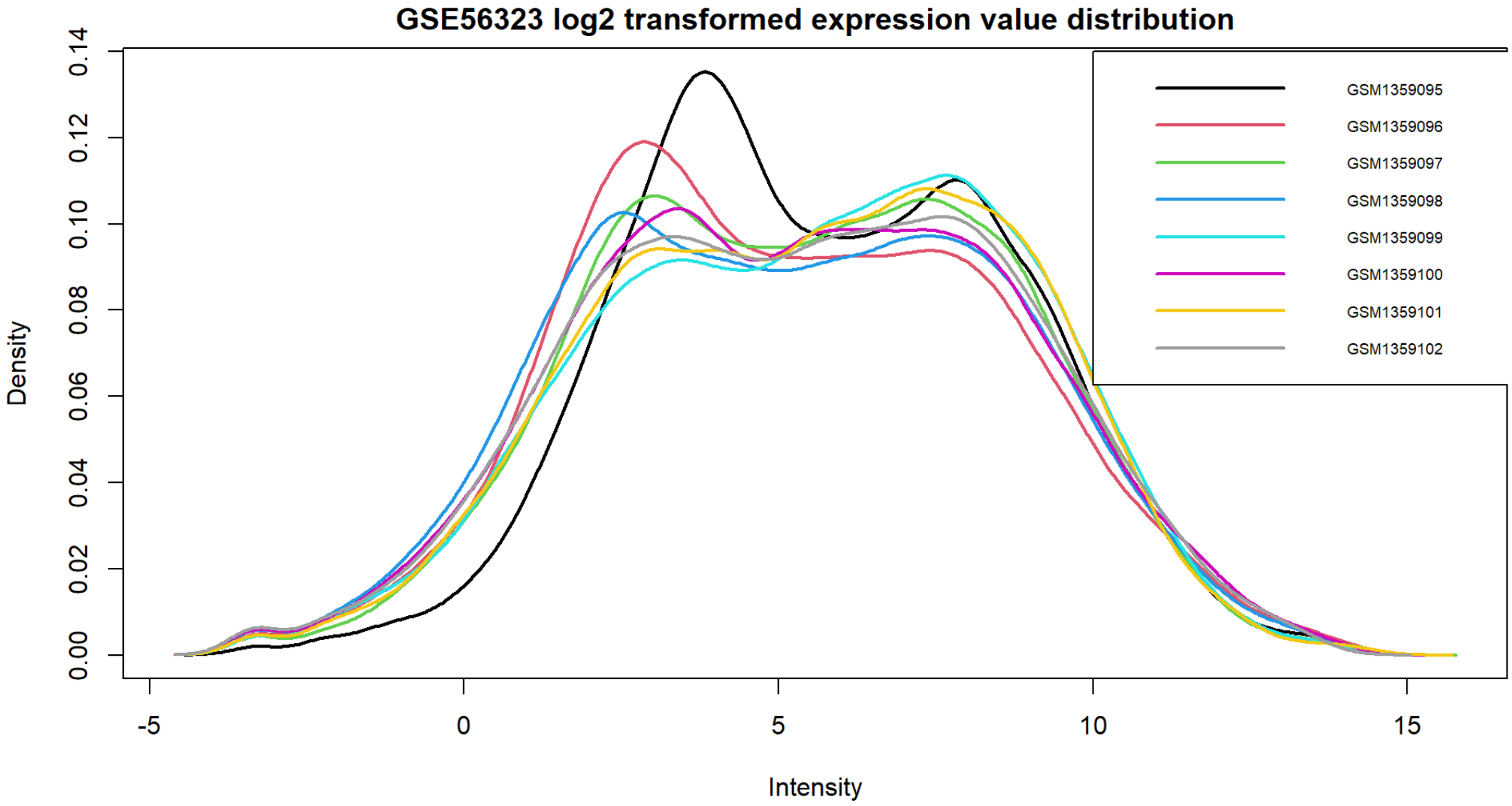
First, take a look at the gene expression distribution.



The gene expression plot is not clear on linear scale

Transform to log2 scale

Results - Filter out low expression genes



The expression plot in log2 scale

Results - Filter out low expression genes

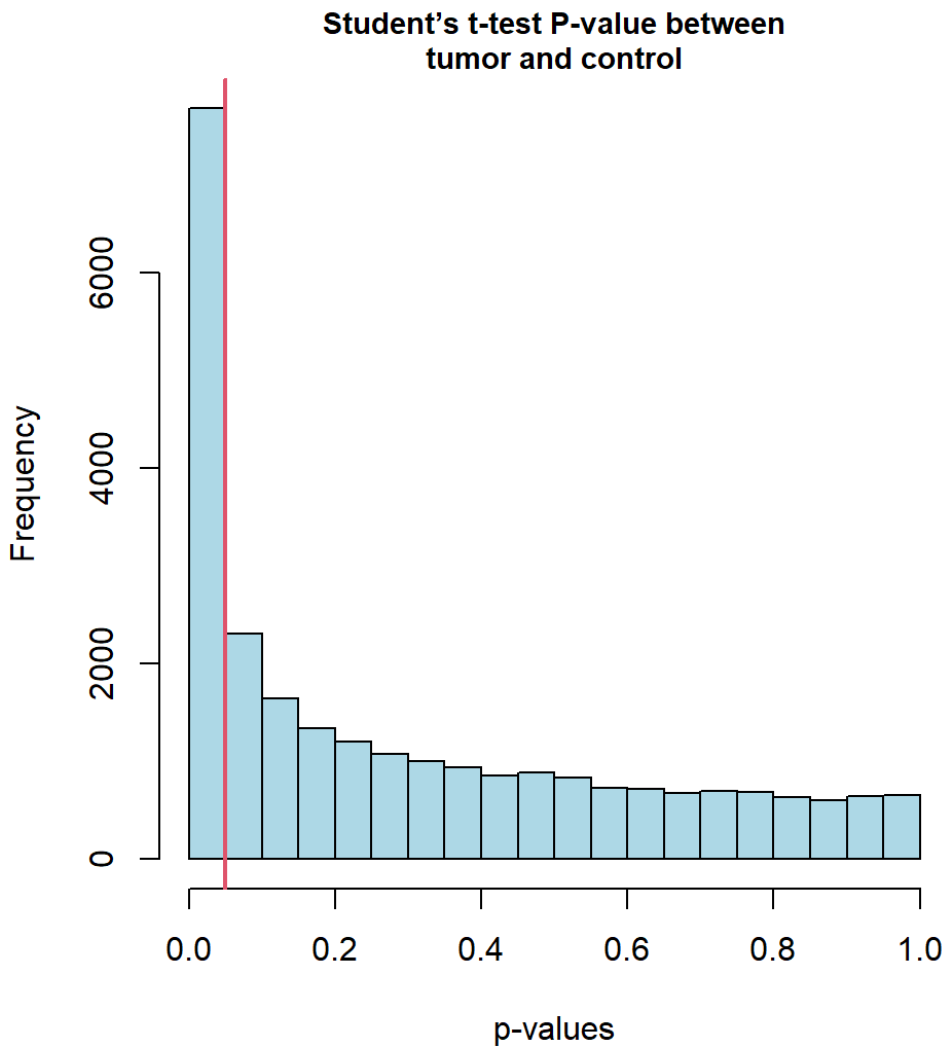
```r
ex <- exprs(gset)
quantile(rowMeans(ex), na.rm=T)
#       0%        25%        50%        75%       100%
# -10.4375     0.8250    17.8000   171.6625 17239.4500
```

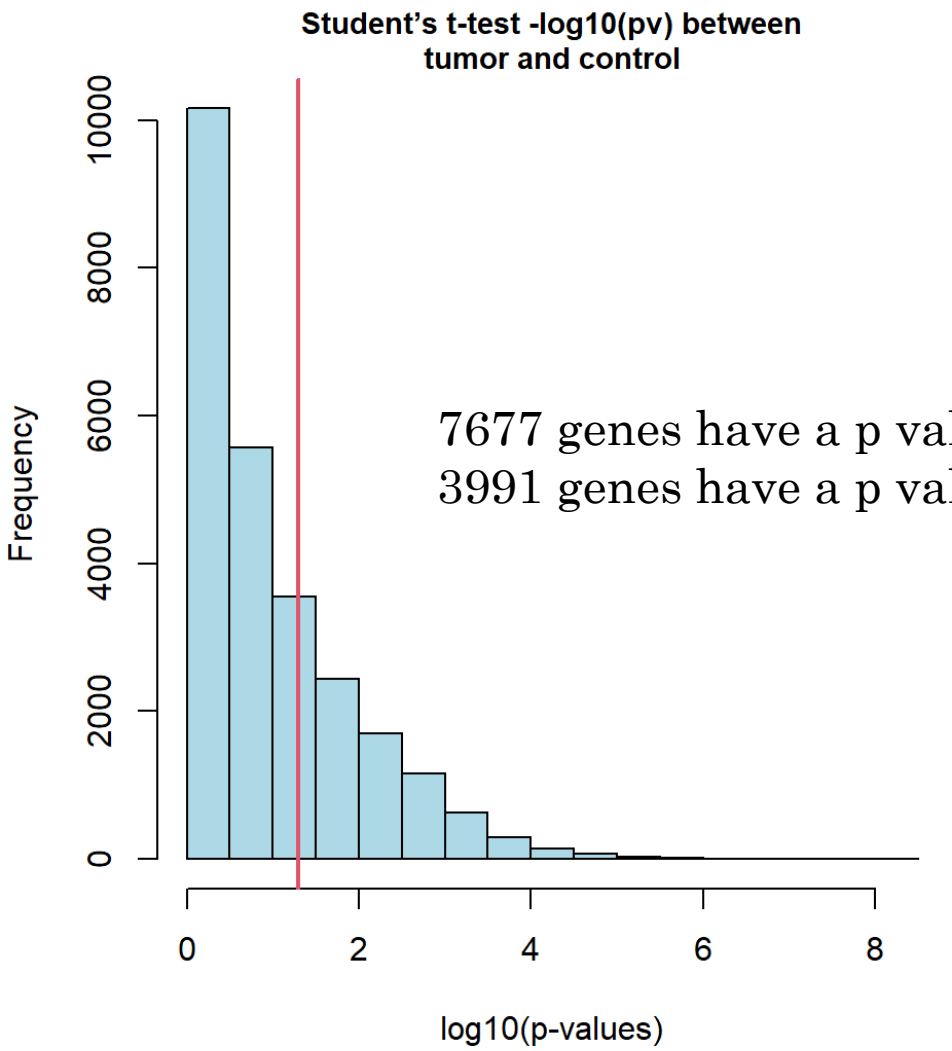
So, the data does contain low expression genes

Filter out these genes (ex <= 0)

Results - Feature selection with a statistical test

Student's two-sample t-test on all genes



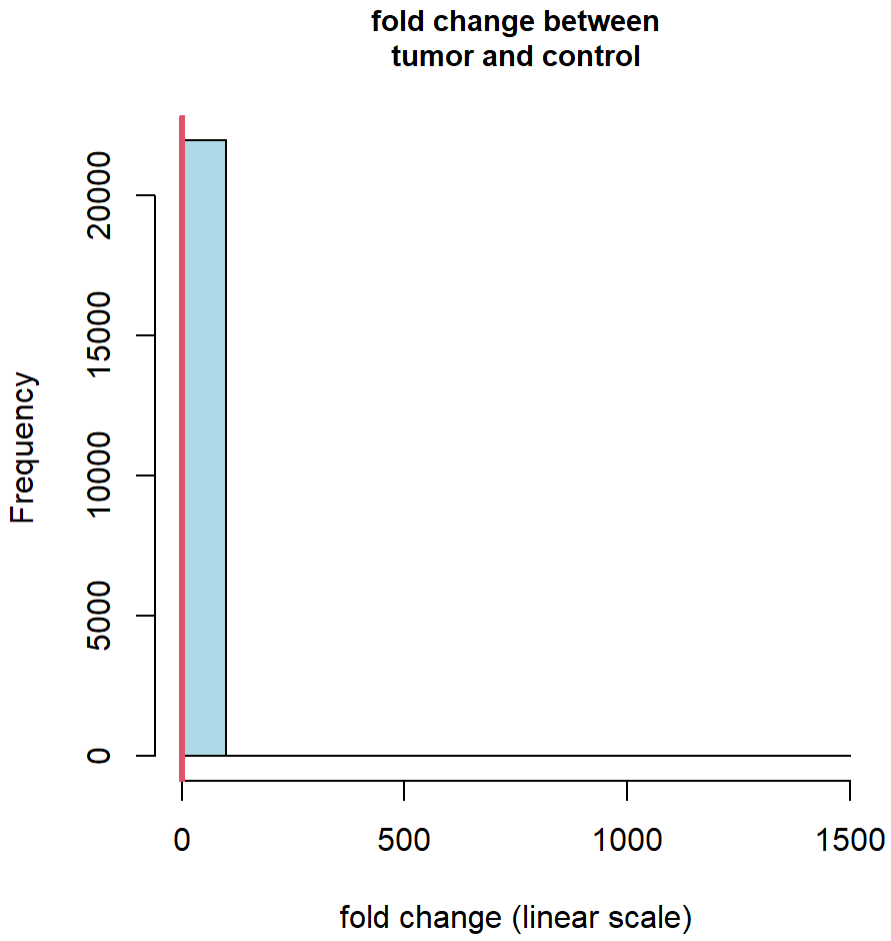


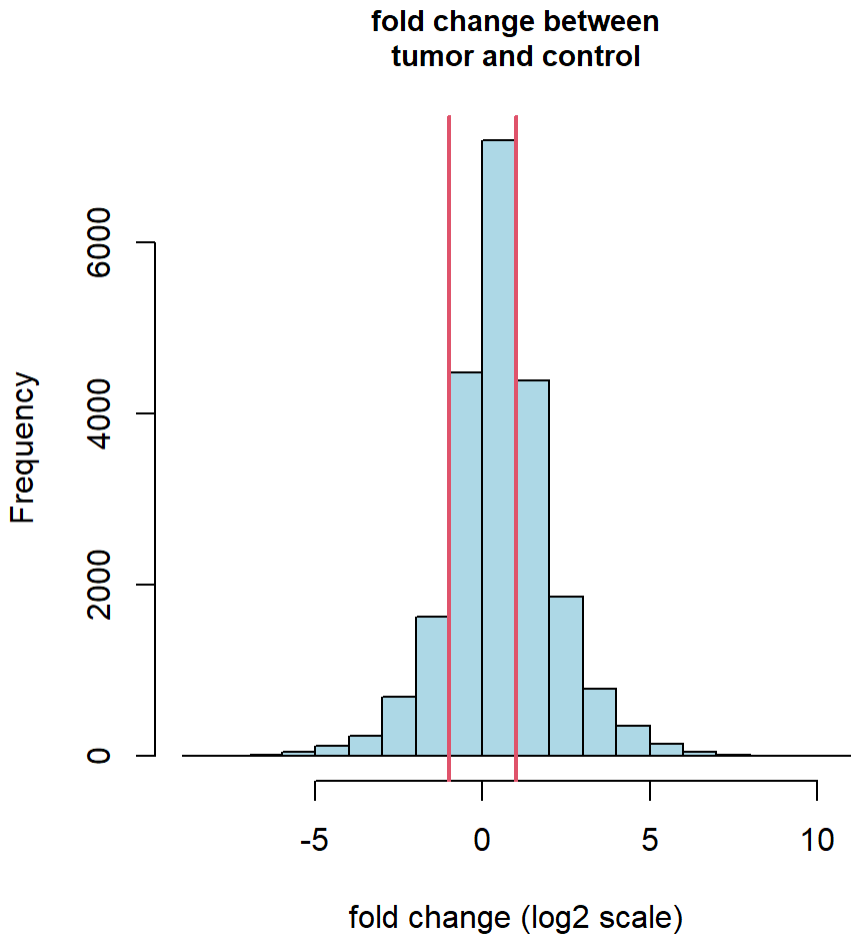
7677 genes have a p value <.05 and
3991 genes have a p value <.01

Histograms of p-values on linear and -log10 scale. Vertical line: p = 0.05

Results - Feature selection with a statistical test

Calculate the fold change between the groups





Histograms of fold change on linear and log2 scale. Vertical line: |fold change|=2

Results - Genes retained

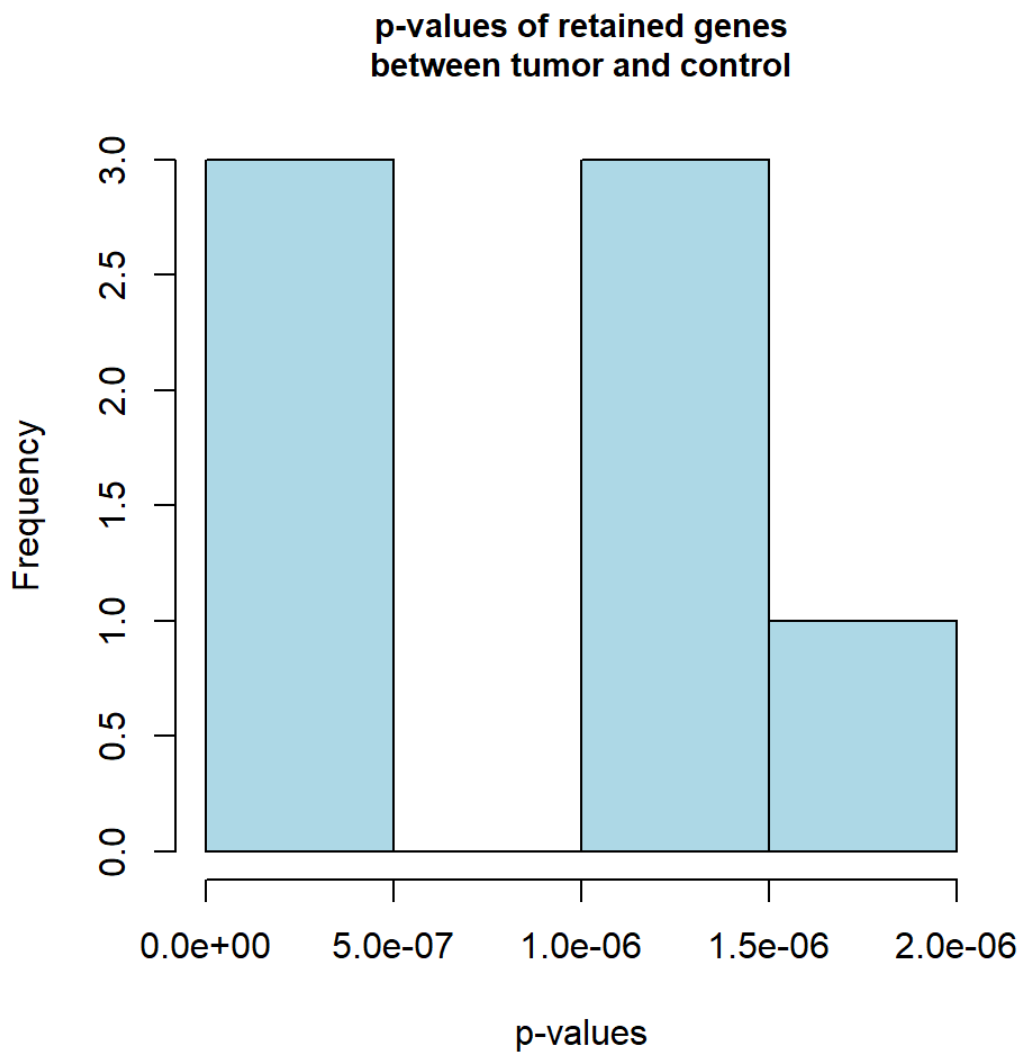
Retain the genes with the two criterions:

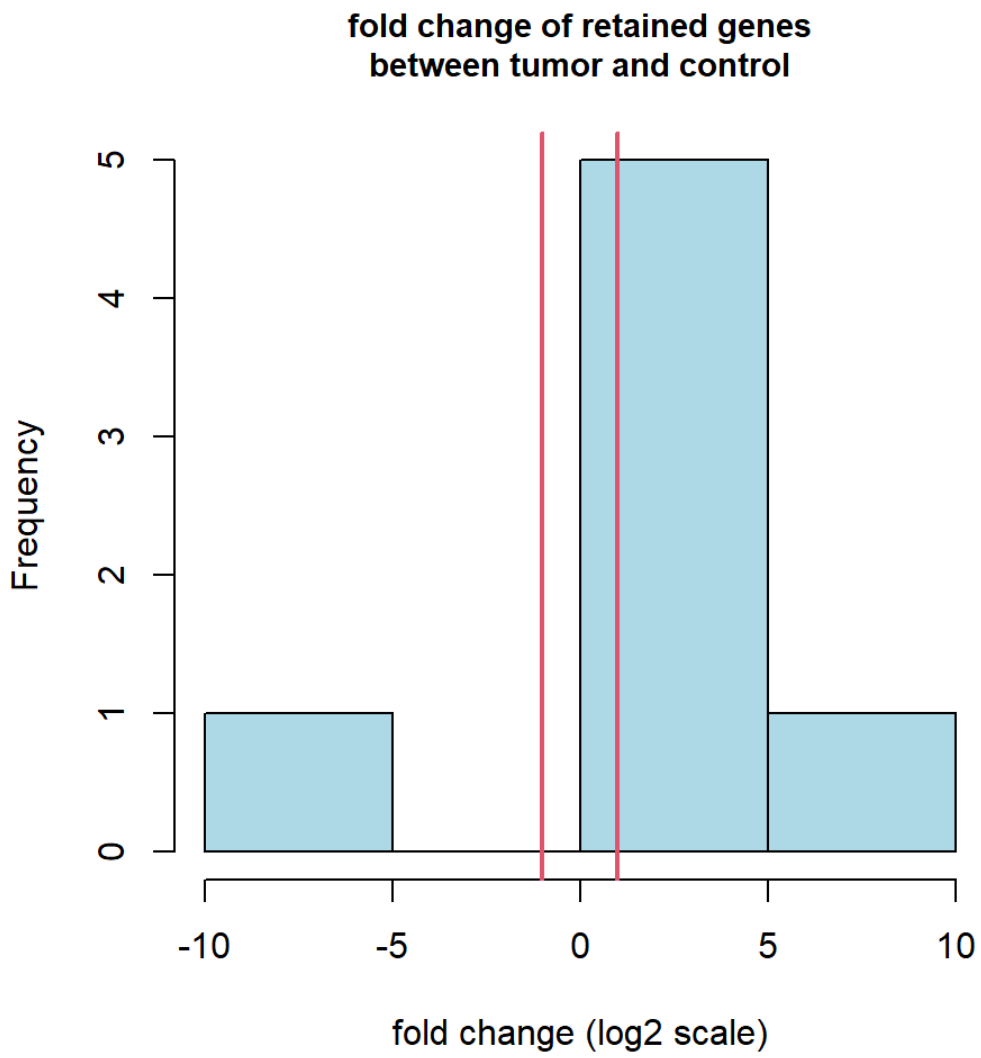
1. p-value < Bonferroni threshold = 0.05/length(dat[,1]) = 1.945752e-06
2. linear |fold change|>2

```
pv.filter <- pv[pv<alpha & abs(fold)>log2(2)]
length(pv.filter)
# [1] 7
Thus, there are 7 genes meet the requirements.
```

There are 7 genes meet the requirements. Plot the scores of those genes in a histogram

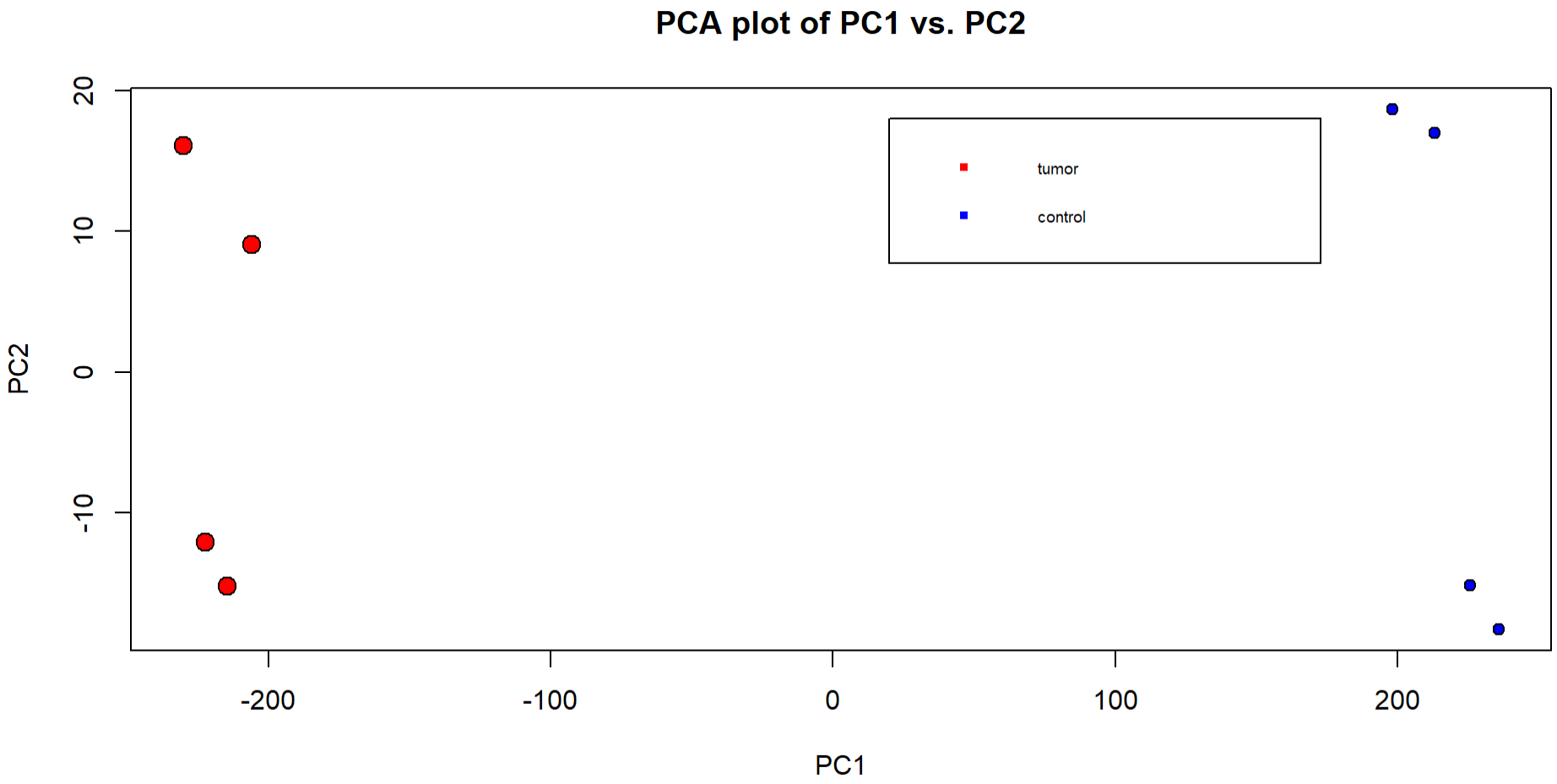
Results - Genes retained





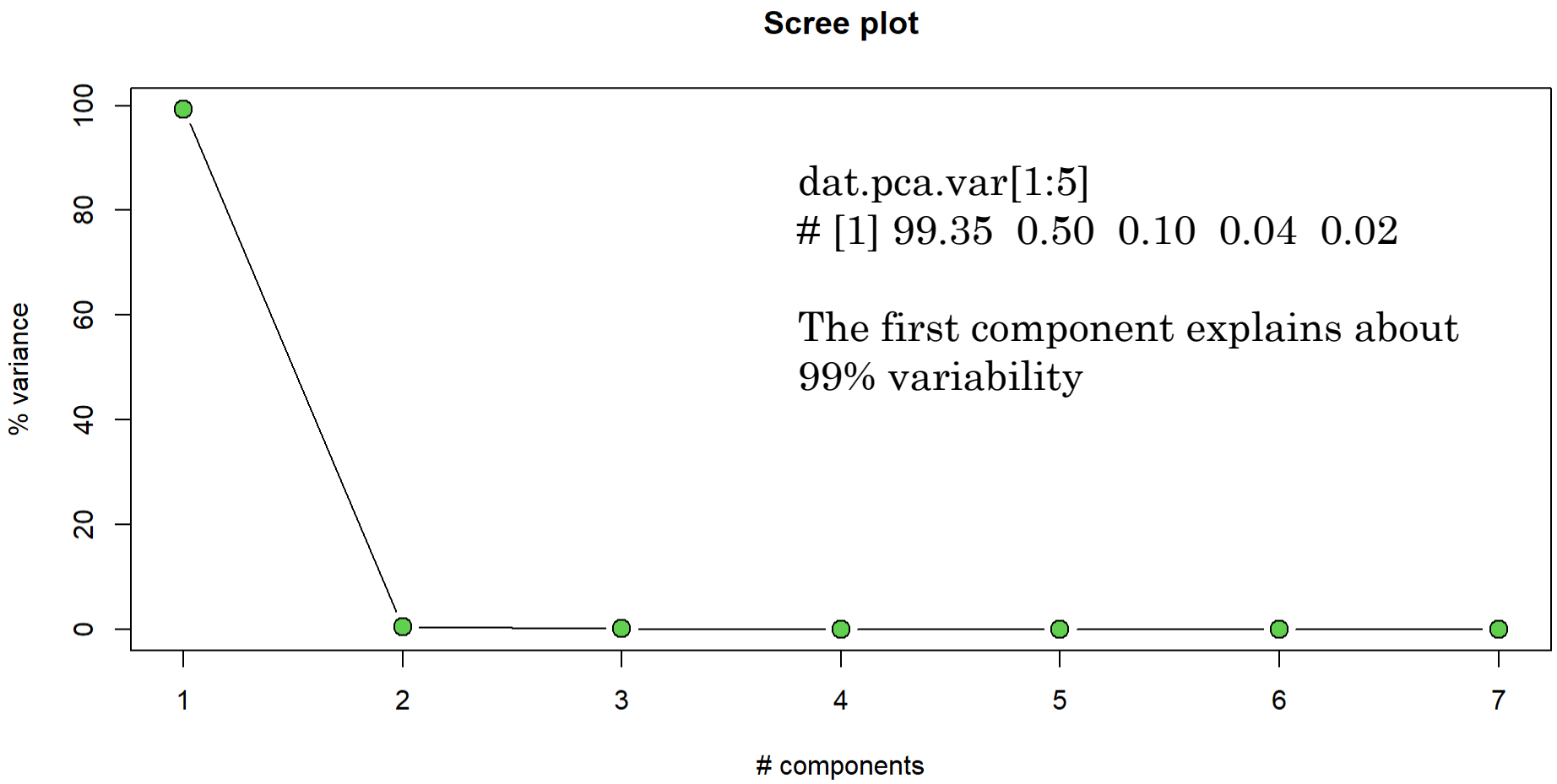
Histograms of p value and fold change of those genes retained

Results – PCA analysis on subset data



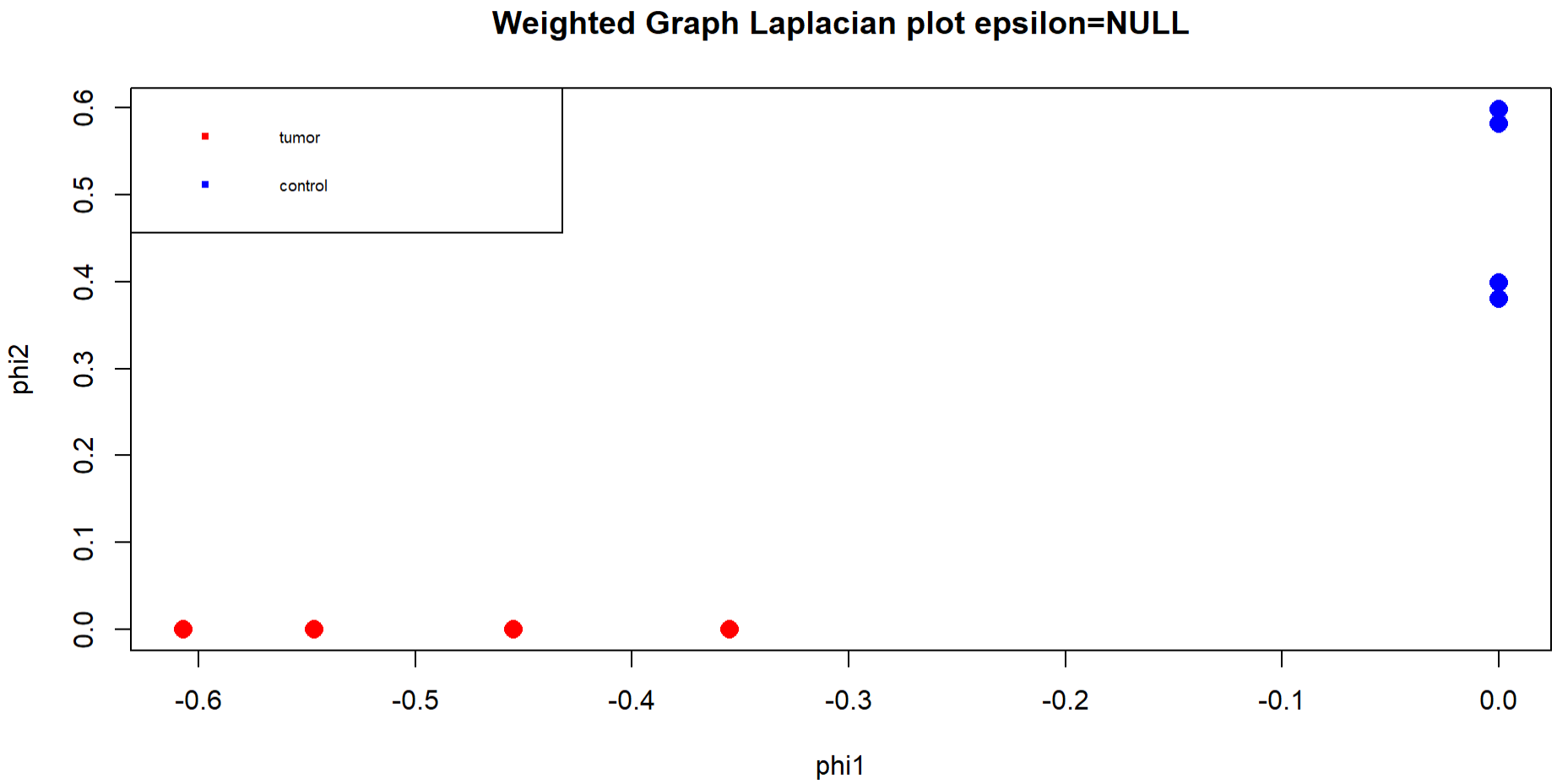
First 2 components of subset data in PCA plot

Results – PCA analysis on subset data



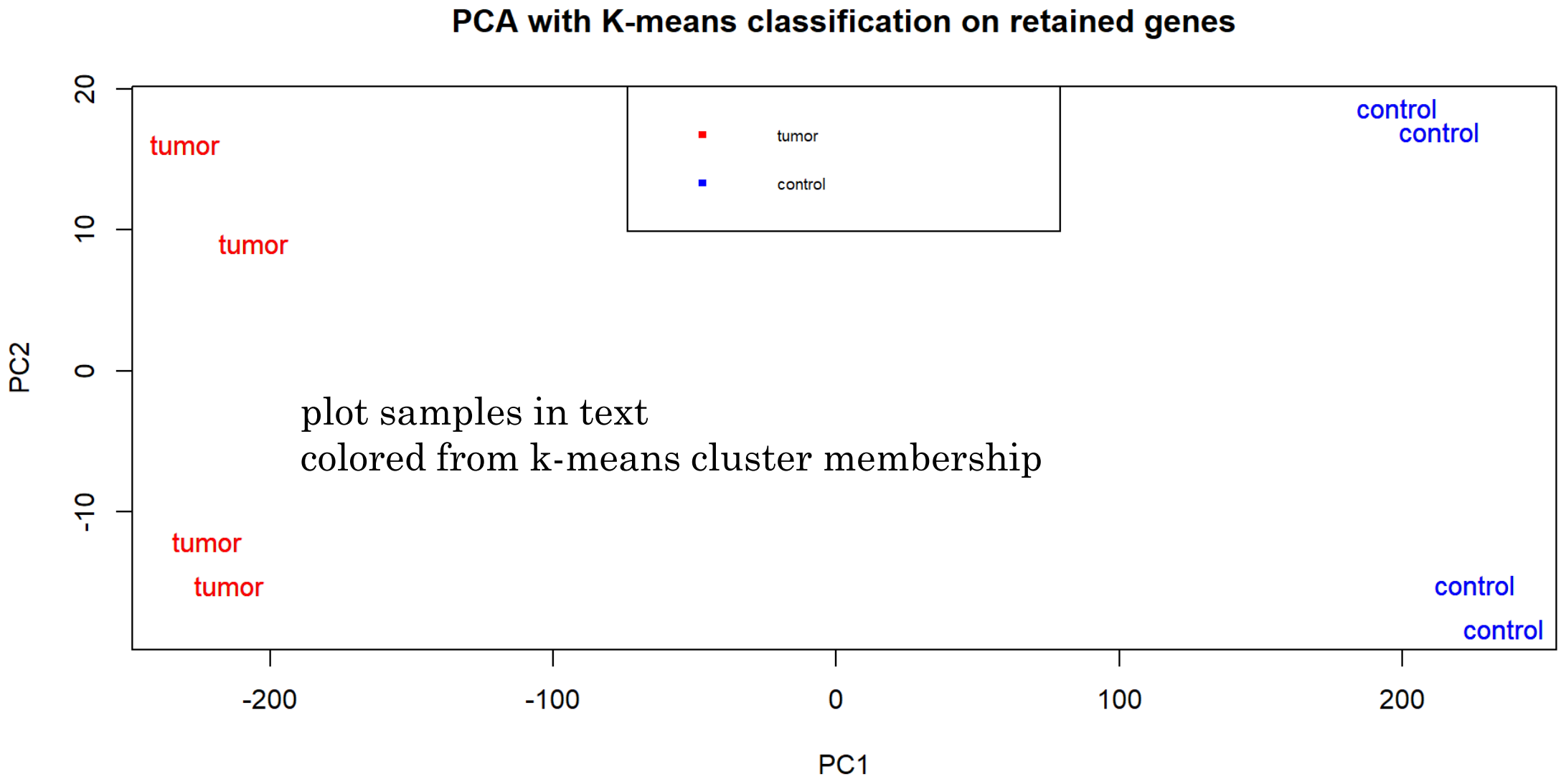
Scree plot that corresponds to the PCA

Results – PCA analysis on subset data



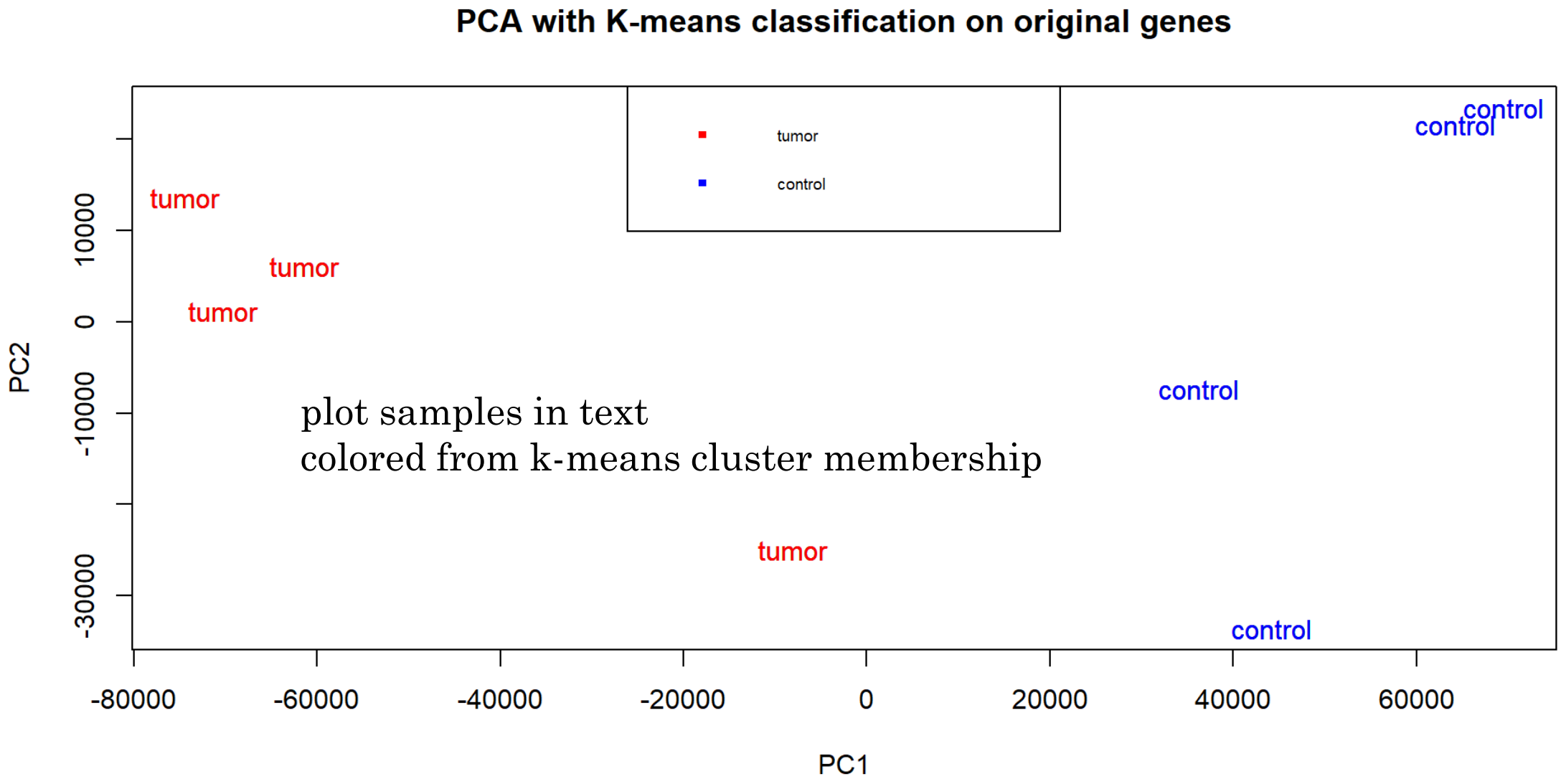
Two-dimensional embedding of the weighted graph Laplacian

Results – Classify the samples into classes



k-means clustering of the subset data based on first 2 PCA components (k=2)

Results – Classify the samples into classes



k-means clustering of the original data based on first 2 PCA components (k=2)

Results - Top 5 discriminant genes

Looking at the annotation file GPL6885.annot, these genes are from Mus musculus :

positive direction
ILMN_1212607 CASP2 and RIPK1 domain containing adaptor with death domain Cradd
ILMN_1212612 regulator of calcineurin 2 Rcan2
ILMN_1212619 microfibrillar-associated protein 3-like Mfap3l
ILMN_1212628 transcription elongation regulator 1-like Tcerg1l
ILMN_1212632 Mab-21 domain containing 2 Mb21d2

negative direction
ILMN_3163569 solute carrier family 16 (monocarboxylic acid transporters), member 3 Slc16a3
ILMN_3163572 adenylate cyclase 7 Adcy7
ILMN_3163577 ACACCGGCCCTTTGTGAAGACCACAAGACTAATACCCCTGCGAGTCACTG BLAST search returns 100% match of NM_001359700.1, which is from gene Scn3b
ILMN_3163581 engrailed 1 En1
ILMN_3163582 L1 cell adhesion molecule L1cam

Results - Functional information on NCBI's DAVID

ENSEMBL_GENE_ID	Species	David Gene Name	BIOLOGICAL PROCESS	GO TERM
71306	Mus musculus	microfibrillar-associated protein 3-like(Mfap3l)	Immunoglobulin domain, Signal, Transmembrane,	nucleus, nucleoplasm, cytoplasm, plasma membrane, membrane, integral component of membrane, cell junction,
16728	Mus musculus	L1 cell adhesion molecule(L1cam)	Cell adhesion, Differentiation, Neurogenesis,	endosome, plasma membrane, external side of plasma membrane, cell surface, membrane, integral component of membrane, axon, dendrite, presynaptic membrane, cell projection, neuronal cell body, terminal bouton, dendritic growth cone, axonal growth cone, membrane raft, Schaffer collateral - CA1 synapse
12905	Mus musculus	CASP2 and RIPK1 domain containing adaptor with death domain(Cradd)	Apoptosis,	nucleus, nucleolus, cytoplasm, endopeptidase complex,
239796	Mus musculus	Mab-21 domain containing 2(Mb21d2)	Nucleotidyltransferase, Transferase,	transferase activity, nucleotidyltransferase activity, macromolecular complex binding,
80879	Mus musculus	solute carrier family 16 (monocarboxylic acid transporters), member 3(Slc16a3)	Symport, Transport,	monocarboxylic acid transport, plasma membrane lactate transport, transmembrane transport,
70571	Mus musculus	transcription elongation regulator 1-like(Tcerg1l)	transcription cofactor activity, RNA polymerase binding,	nucleus,
11513	Mus musculus	adenylate cyclase 7(Adcy7)	cAMP biosynthesis,	plasma membrane, integral component of plasma membrane, membrane, integral component of membrane,
13798	Mus musculus	engrailed 1(En1)	Developmental protein, DNA-binding, Developmental protein,	nucleus, membrane,
235281	Mus musculus	sodium channel, voltage-gated, type III, beta(Scn3b)	Ion transport, Sodium transport, Transport,	voltage-gated sodium channel complex, plasma membrane, membrane, integral component of membrane, Z disc,
53901	Mus musculus	regulator of calcineurin 2(Rcan2)	Thyroid hormone signaling pathway,	nucleus, cytoplasm,

Table of gene name and functional information for 10 discriminant genes

Results - Functional information on NCBI's DAVID

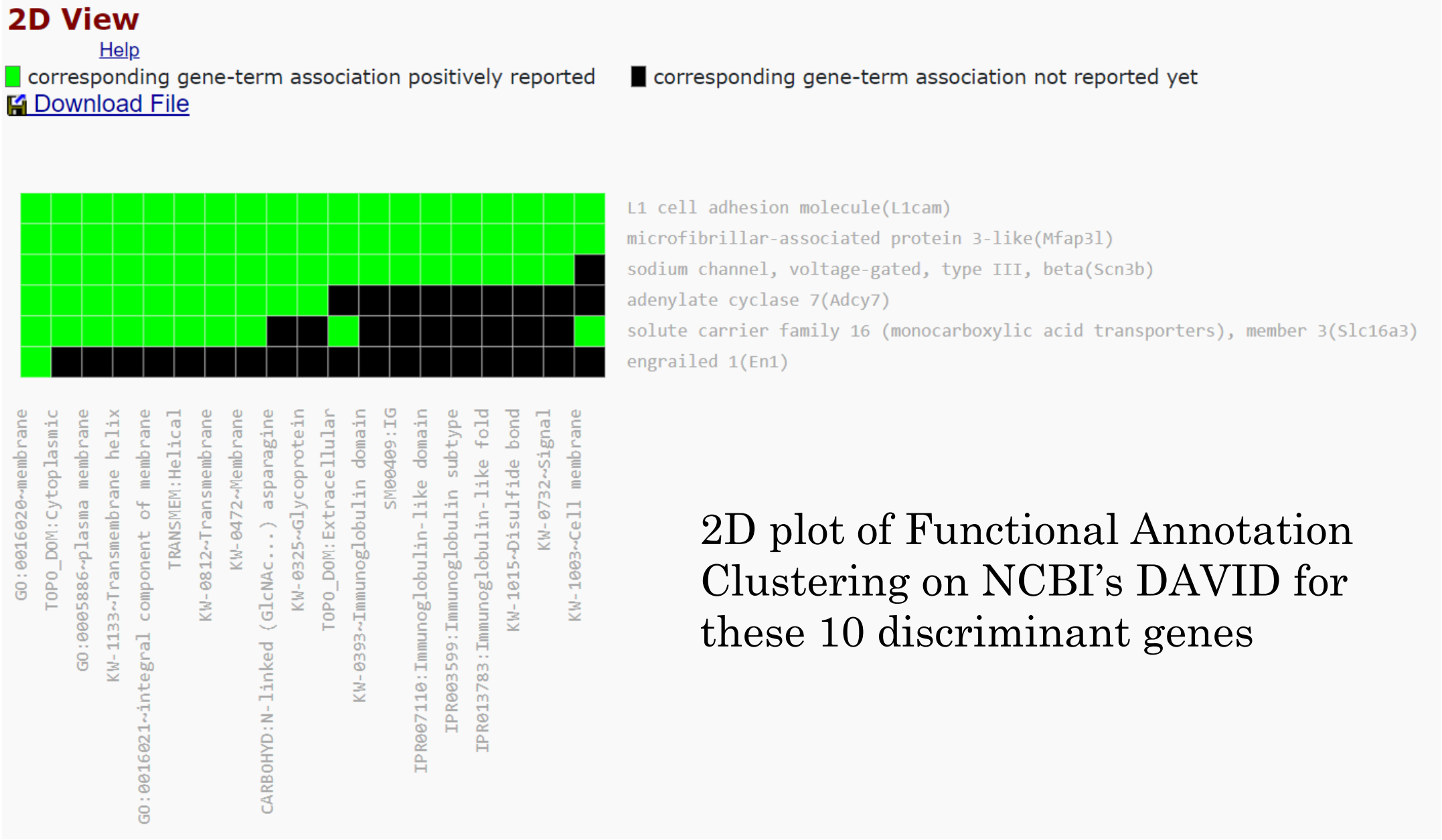
12 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	UP_KW_DOMAIN	Immunoglobulin domain	RT		3	30.0	1.5E-2	1.1E-1
☐	GOTERM_BP_DIRECT	calcium-mediated signaling	RT		2	20.0	2.7E-2	1.0E0
☐	INTERPRO	Immunoglobulin subtype	RT		3	30.0	2.7E-2	6.5E-1
☐	UP_SEQ_FEATURE	TOPO_DOM:Cytoplasmic	RT		5	50.0	3.0E-2	6.9E-1
☐	UP_SEQ_FEATURE	DOMAIN:Ig-like C2-type	RT		2	20.0	4.2E-2	6.9E-1
☐	UP_SEQ_FEATURE	REGION:Disordered	RT		9	90.0	5.7E-2	6.9E-1
☐	SMART	IG	RT		3	30.0	5.7E-2	6.3E-1
☐	UP_SEQ_FEATURE	TOPO_DOM:Extracellular	RT		4	40.0	6.3E-2	6.9E-1
☐	INTERPRO	Immunoglobulin I-set	RT		2	20.0	6.3E-2	6.5E-1
☐	INTERPRO	Immunoglobulin-like domain	RT		3	30.0	6.6E-2	6.5E-1
☐	GOTERM_CC_DIRECT	membrane	RT		6	60.0	7.4E-2	1.0E0
☐	INTERPRO	Immunoglobulin subtype 2	RT		2	20.0	9.9E-2	6.5E-1

Functional Annotation Chart on NCBI's DAVID for these 10 discriminant genes

Results - Functional information on NCBI's DAVID



2D plot of Functional Annotation Clustering on NCBI's DAVID for these 10 discriminant genes

Conclusions

- The analysis pipeline performed Student's two-sample t-test, PCA analysis, and k-means clustering on tumor and control kidney samples.

- The analysis found a few discriminant genes that show significant expression difference between tumor and control kidney groups.

- However, Lin28 was not one of them

- Besides, those discriminant genes are involved in the processes like transport and biosynthesis, etc.
- None of them are related to tumorigenesis or Lin28/Let-7 pathway as suggested by the authors.