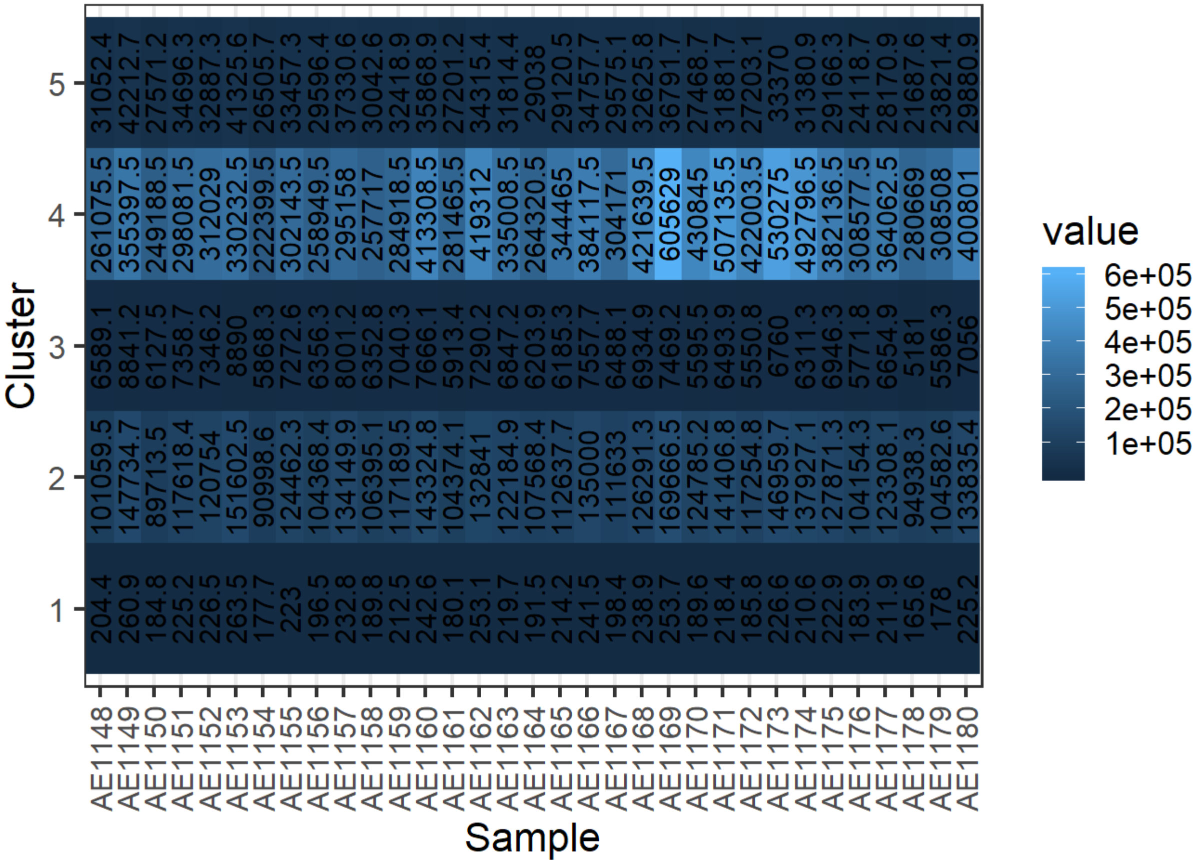
# HW4 – Bioinformatics – 236523

Anna Romanov 321340580 annarom@campus.technion.ac.il
Maxim Kolchinsky 320983216 kolchinsky@campus.technion.ac.il

## Question 1

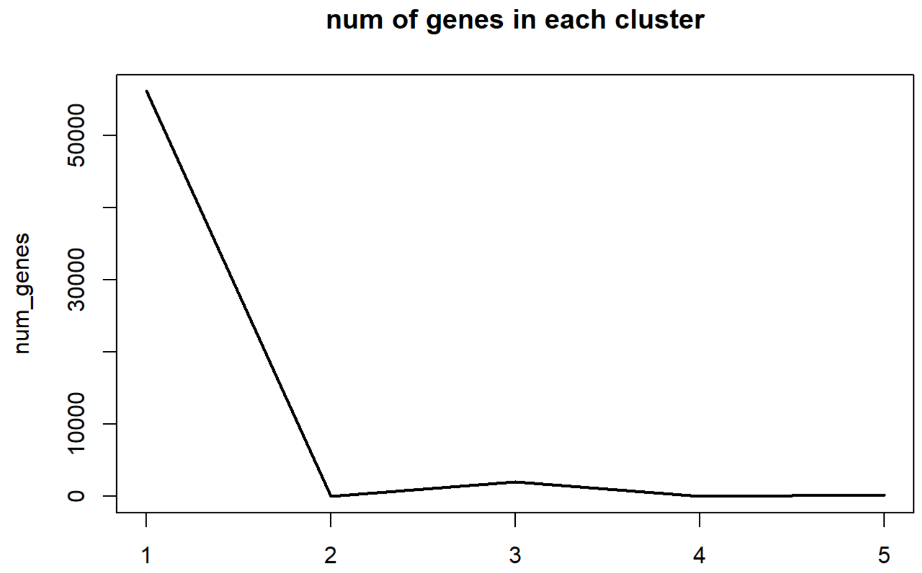
2. c. ii.



iii.

```
##     1     2     3     4     5
## 56101    21  2054     2   169
```

Plot:



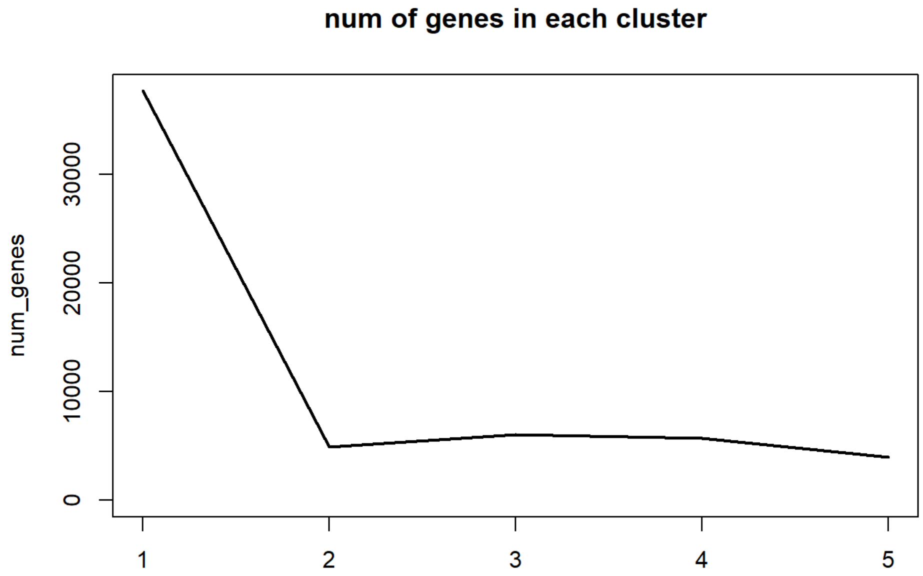
iv. We see that the number of genes in each cluster varies significantly – from 2 genes in cluster number 1, to 56,101 genes in cluster number 5. This is a shortcoming since most of the genes appear in a single cluster which is not very informative for expression analysis.

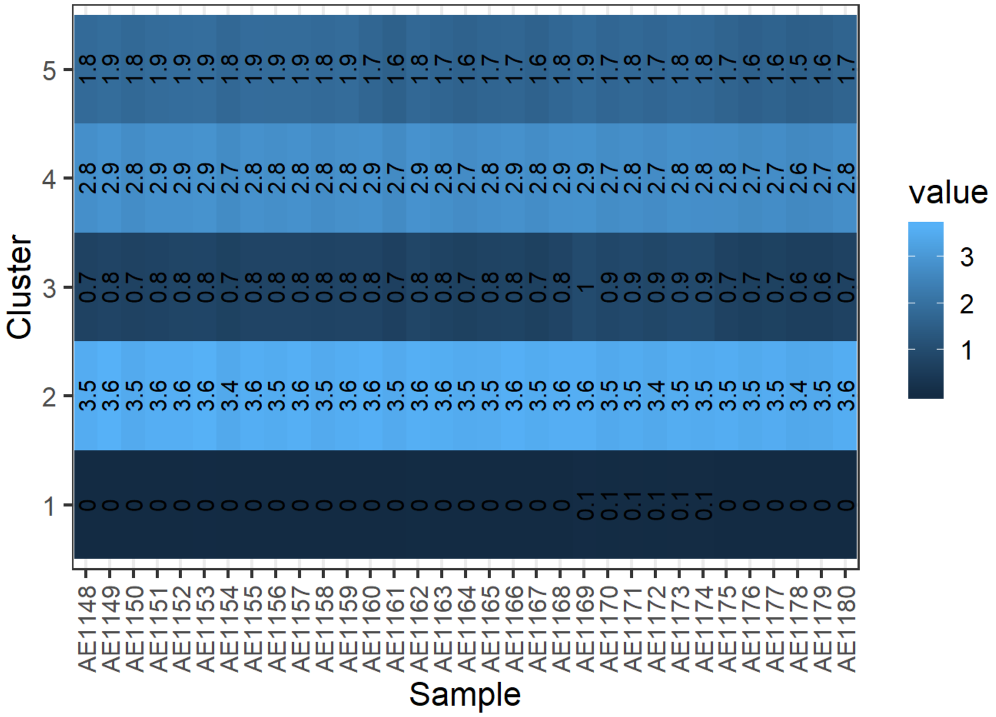
Another problem is that for clusters 1 and 2, the range of values that belong to the same cluster is big, therefore in the heatmap it looks like cluster 1 is not homogeneous. There are two possibilities: either the big difference in values (for example there is a value of 605,629 together with 222,399 in cluster 1) has biological meaning and those samples being in the same cluster is an error; or in the range of such high values, big differences are less significant biologically (a difference of ~200,000 is still in the same order of magnitude), so the clustering itself is correct but the heatmap suggests that there might be significant differences in examples in the same cluster. In the second case, a better way of representing the results could be used (as in the log transformation below).

A possible reason for these problems could be that the expression values are within a very large range, being more dense in the smaller values (we see that in cluster 5, the maximum difference is around ~100, while in cluster 1 it is around ~300,000). This in turn is caused by large differences in counts of different genes as appear in ‘rawcounts.csv’.
This way, to make distinction between clusters the algorithm divides intro groups roughly by order of magnitude, but it might be that after a certain threshold big differences are not significant, while in the small values the algorithm should be more sensitive to small fluctuations.

v. After applying log transformation:

```
##     1     2     3     4     5
## 37664  4918  6076  5699  3990
```





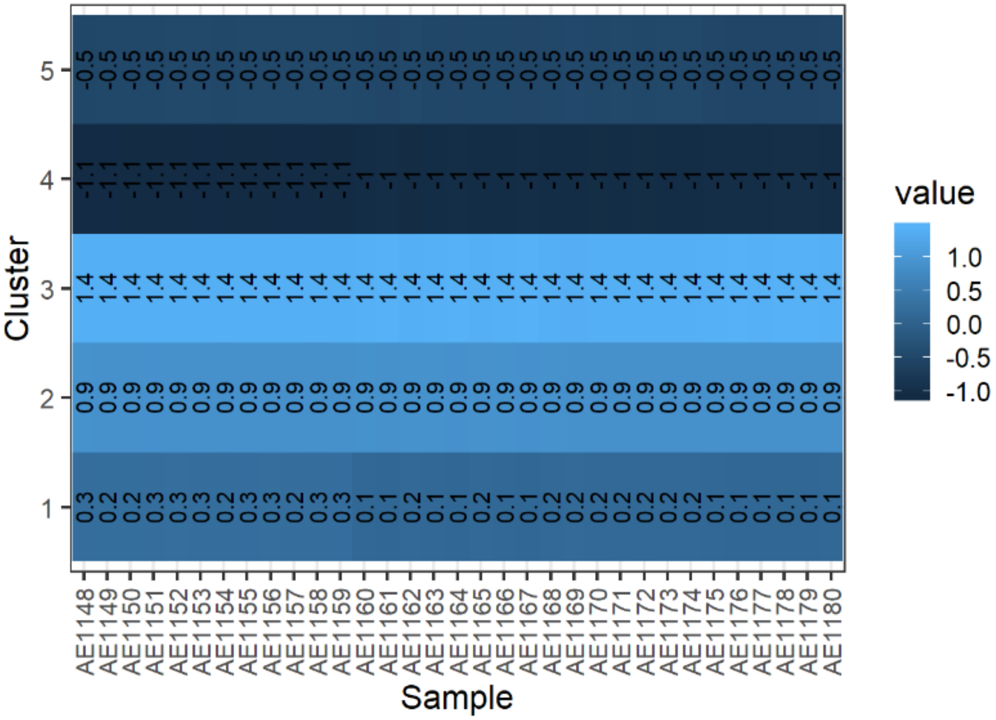
We see that cluster sizes are now more balanced, and also the expression values in each cluster are closer to each other (in contrast to the previous configuration), making clusters more homogenic. This can be explained by the fact that expression values are now in smaller range – from 0 up to 4, so more of the larger values are found in the same cluster (since big differences after log transformation become much smaller and examples which previously were considered ‘too different’ are now similar).

vi. We noticed that many of the rows in the dataset have values of zero in all or nearly all columns, and therefore such rows are not valuable for expression analysis of genes – our goal is to find genes which have different expression levels for different treatments, however if a gene has zero expression for all samples, we can’t conclude useful information based on this gene. Therefore, an improvement we suggest is to filter out rows with zero or very small expression. Specifically, we sum all rows and filter out rows with a sum less than 10.
In addition, we performed standardization on the data to prepare it for clustering.

vii. The results we got (using log transform too):

Cluster sizes:

```
   1    2    3    4    5
3580 5810 4362 9297 4650
```

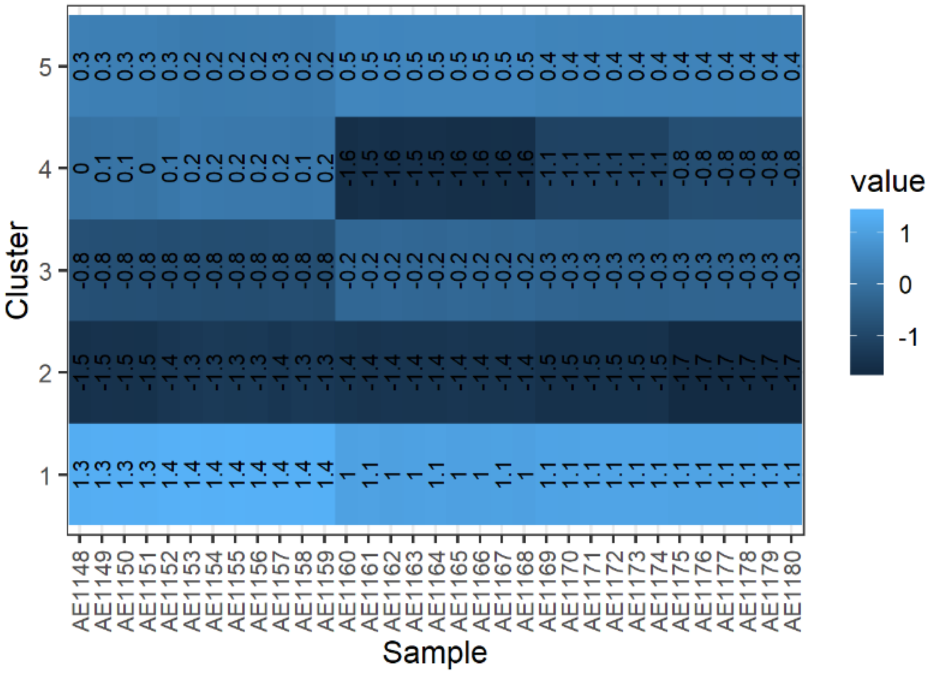


As we can see, the clusters are more homogenous now and also different from each other.

x. We used the file “sigresults.csv” given in the course website.
The results after filtering the data (and applying all the steps as described in the previous section – ignoring rows of zeroes, log-transformation and scaling):

```
  1    2    3    4    5
```

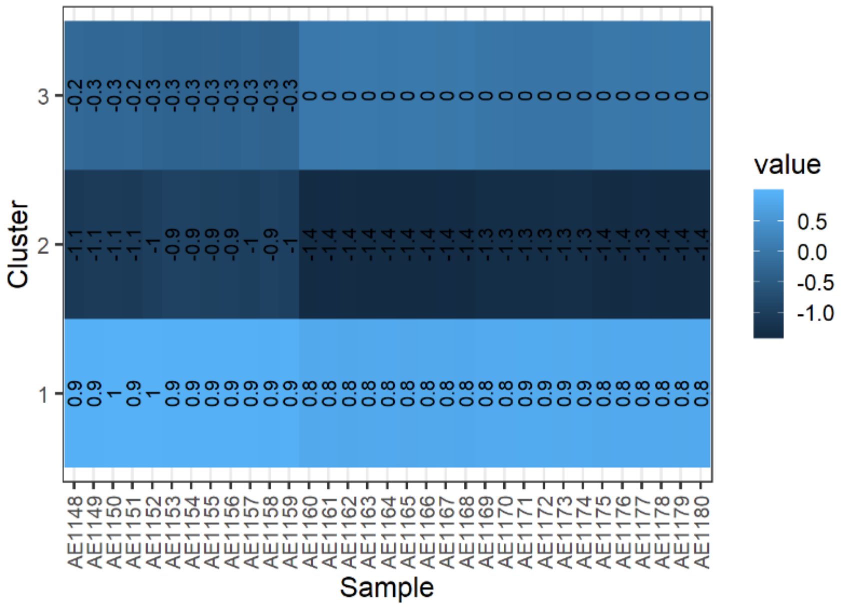
```
 171 121 202  79 348
```



We see that the clusters are less homogenous now and are also less distinct from each other. For example, cluster 4 looks like a combination of clusters 2 and 3. The conclusion is that filtering the data to contain only the differentially expressed genes didn't help the clustering algorithm. This is against our expectation that leaving only DE genes would show better results, since clustering is done only on the most 'relevant' genes. A possible explanation could be that DE genes don't hold enough information to compensate the loss of data (since we filtered out thousands of genes). A solution could be to try a smaller number of clusters, since we assume that there is less variability in the data after the filtering. And indeed, with 3 clusters we got the following results, which appear to be better:

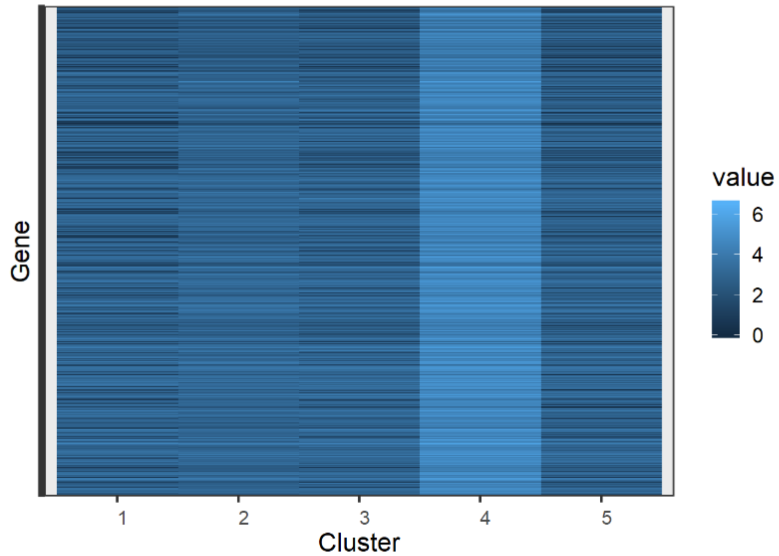
Cluster sizes -

```
  1   2   3 
341 216 364 
```



d. ii. Using log-transform, removing low-expression rows and filtering the dataset to contain DE genes, we got the following results:
cluster sizes (with 5 clusters) –

```
1  2  3  4  5
9 12  6  1  6
```



iii. Clustering the genes, we got good results with the 5 clusters being homogenous and well-separated from each other (meaning that in the same cluster, genes have similar expressions levels, and different from those of genes in other clusters). From the clustering results, we can learn which genes have similar behavior (having similar expression levels on different samples). For example, if we know that a certain gene's expression is  affected by some treatments, using clustering we can find out which other genes are affected by the same treatments (or, if we know that a gene is related to a disease, we can look for more genes related to the same disease using clustering of genes).

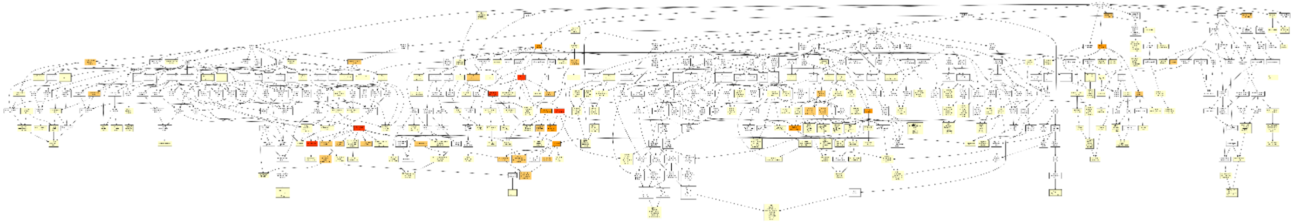
The results of clustering samples are harder to work with, as many samples in different clusters have similar expression levels in many genes. However, there are some differences between clusters, and some genes do show distinct values. Clustering samples can, for example, help us understand which treatments have similar effects, or in which cell types there are similar expression levels of some genes (possibly showing common functionalities affected by cancer).

## Question 2

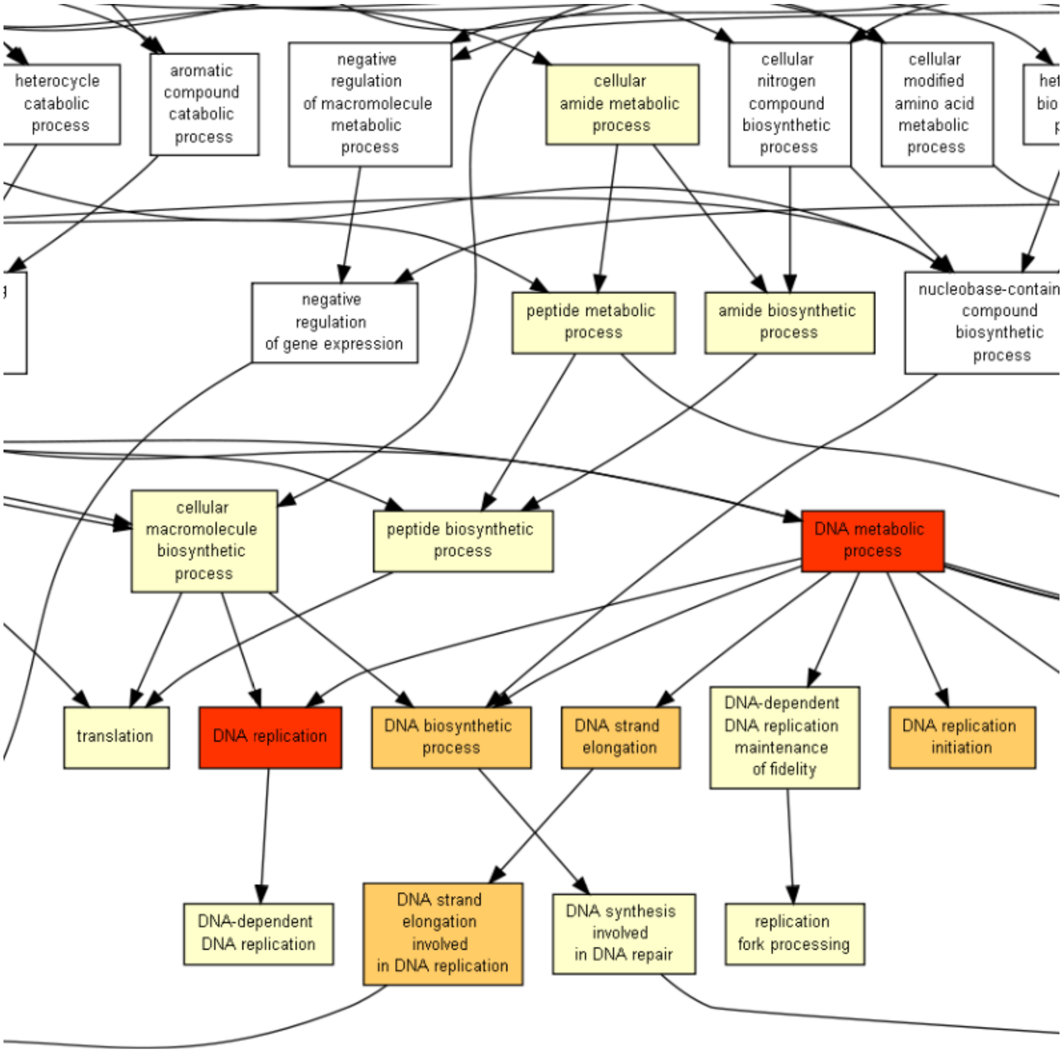
1. C. We got a long list of GO terms that were found by Gorilla, the most significant ones are:

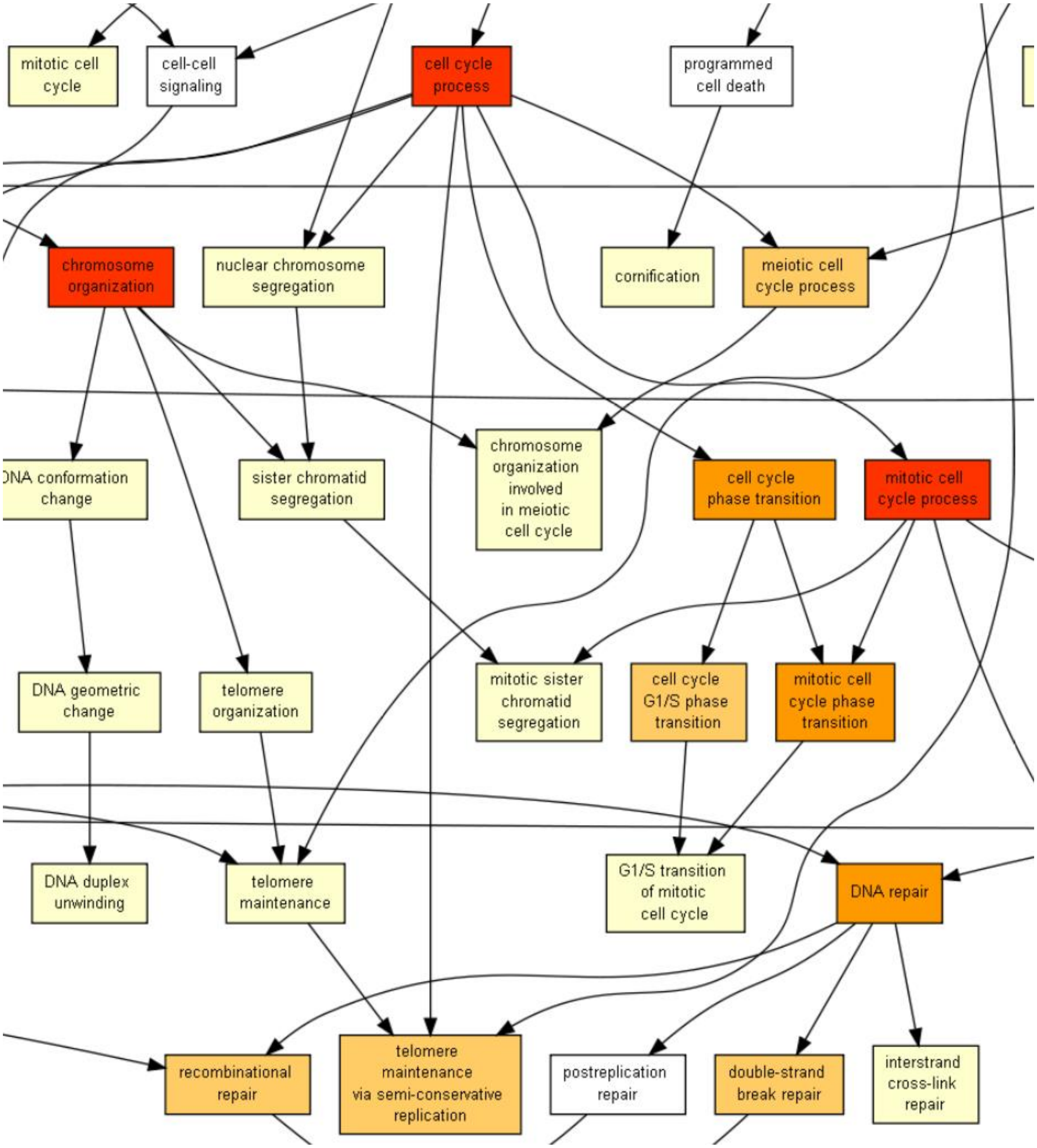
| GO term | Description | P-value | FDR q-value | Enrichment (N, B, n, b) |
|---|---|---|---|---|
| GO:0022402 | cell cycle process | 8.7E-17 | 1.22E-12 | 2.01 (11732,764,1261,165) |
| GO:0006260 | DNA replication | 2.6E-14 | 1.82E-10 | 4.04 (11732,136,940,44) |
| GO:1903047 | mitotic cell cycle process | 2.91E-14 | 1.36E-10 | 2.18 (11732,504,1261,118) |
| GO:0051276 | chromosome organization | 6.96E-13 | 2.44E-9 | 2.50 (11732,294,1163,73) |
| GO:0006259 | DNA metabolic process | 5.62E-12 | 1.57E-8 | 2.15 (11732,641,819,96) |
| GO:0009987 | cellular process | 1.59E-9 | 3.72E-6 | 1.09 (11732,9073,1237,1044) |
| GO:0007093 | mitotic cell cycle checkpoint | 6.46E-9 | 1.29E-5 | 3.38 (11732,89,1208,31) |
| GO:0000075 | cell cycle checkpoint | 1.49E-8 | 2.61E-5 | 2.99 (11732,117,1208,36) |
| GO:0050896 | response to stimulus | 2.04E-8 | 3.18E-5 | 1.35 (11732,3063,833,294) |
| GO:0006281 | DNA repair | 2.56E-8 | 3.58E-5 | 2.19 (11732,403,837,63) |
| GO:0044772 | mitotic cell cycle phase transition | 2.95E-8 | 3.75E-5 | 2.37 (11732,203,1221,50) |
| GO:0044770 | cell cycle phase transition | 3.69E-8 | 4.31E-5 | 2.33 (11732,210,1221,51) |
| GO:1903046 | meiotic cell cycle process | 7.38E-7 | 7.95E-4 | 2.90 (11732,104,1167,30) |

| | | | | |
|---|---|---|---|---|
| GO:0031023 | microtubule organizing center organization | 8.31E-7 | 8.31E-4 | 3.63 (11732,57,1192,21) |
| GO:0060249 | anatomical structure homeostasis | 8.74E-7 | 8.15E-4 | 2.27 (11732,191,1188,44) |
| GO:0006271 | DNA strand elongation involved in DNA replication | 9.31E-7 | 8.15E-4 | 9.96 (11732,12,785,8) |



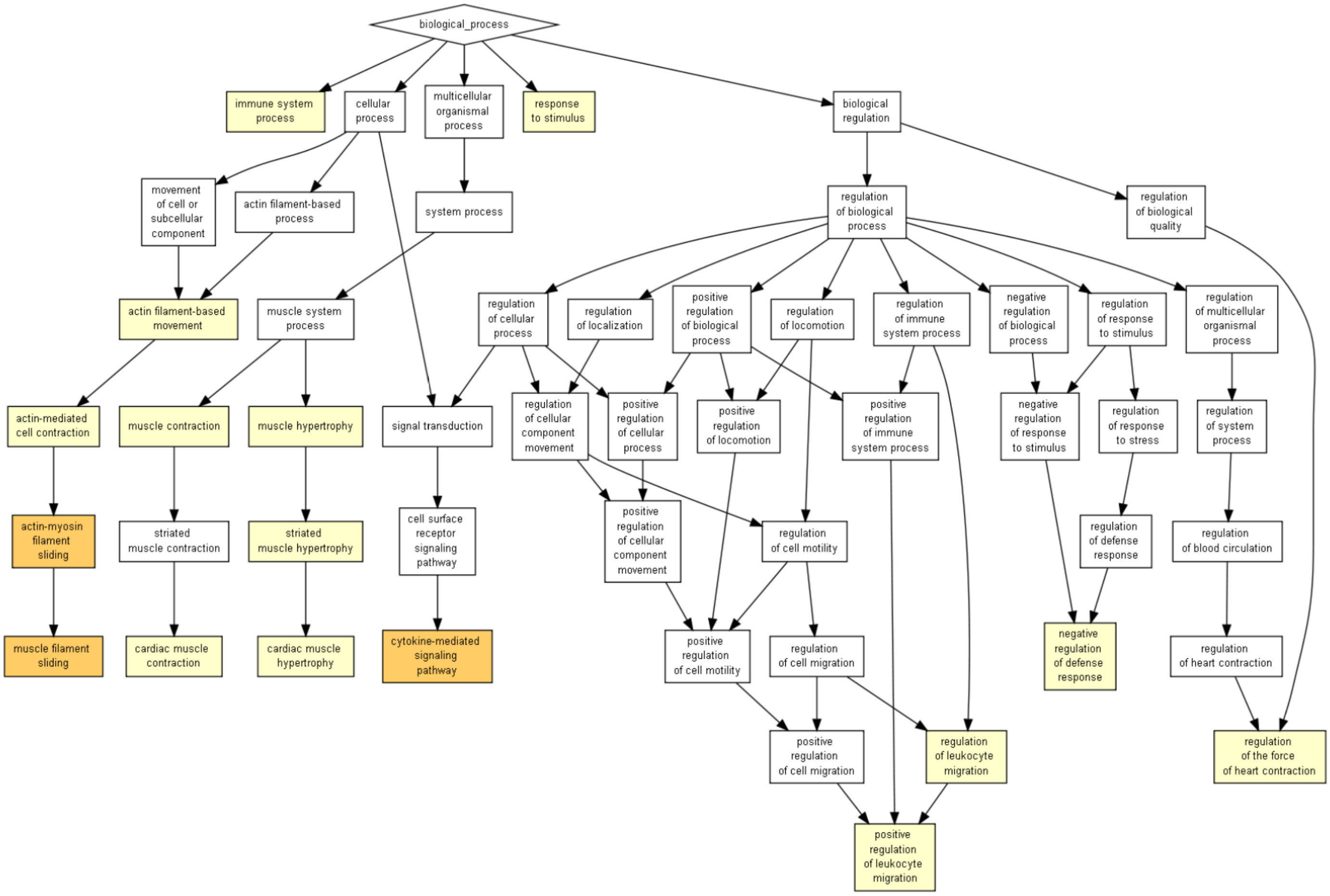
Relevant parts of the diagram are:


heterocycle catabolic process
aromatic compound catabolic process
negative regulation of macromolecule metabolic process
cellular amide metabolic process
cellular nitrogen compound biosynthetic process
cellular modified amino acid metabolic process
het bio
negative regulation of gene expression
peptide metabolic process
amide biosynthetic process
nucleobase-contain compound biosynthetic process
cellular macromolecule biosynthetic process
peptide biosynthetic process
DNA metabolic process
translation
DNA replication
DNA biosynthetic process
DNA strand elongation
DNA-dependent DNA replication maintenance of fidelity
DNA replication initiation
DNA-dependent DNA replication
DNA strand elongation involved in DNA replication
DNA synthesis involved in DNA repair
replication fork processing


mitotic cell cycle
cell-cell signaling
cell cycle process
programmed cell death
chromosome organization
nuclear chromosome segregation
cornification
meiotic cell cycle process
DNA conformation change
sister chromatid segregation
chromosome organization involved in meiotic cell cycle
cell cycle phase transition
mitotic cell cycle process
DNA geometric change
telomere organization
mitotic sister chromatid segregation
cell cycle G1/S phase transition
mitotic cell cycle phase transition
DNA duplex unwinding
telomere maintenance
G1/S transition of mitotic cell cycle
DNA repair
recombinational repair
telomere maintenance via semi-conservative replication
postreplication repair
double-strand break repair
interstrand cross-link repair

D. The biological insight we gain is that differentially the expressed genes found in HW3 are part of certain biological processes as listed above. In particular, the lowest P-value (most significant) belongs to the GO term GO:0022402: cell cycle process. Other significant processes are DNA replication, mitotic cell cycle process, chromosome organization and more. We can conclude from the results which are the processes that are influenced most by the treatment. If the treatment improves the condition, the results could be used to generalize and see which processes are most affected by the disease and should be targeted by other treatments as well.

2. a. To obtain the target list, we filtered the list of all genes (that appear in DE_results_corrected.csv) to contain only genes with log2FoldChange values above or below 2 and -2 respectively, and also padj less than 0.05.
   c. The results of running Gorilla:



biological_process
immune system process
cellular process
multicellular organismal process
response to stimulus
biological regulation
movement of cell or subcellular component
actin filament-based process
system process
regulation of biological process
regulation of biological quality
actin filament-based movement
muscle system process
regulation of cellular process
regulation of localization
positive regulation of biological process
regulation of locomotion
regulation of immune system process
negative regulation of biological process
regulation of response to stimulus
regulation of multicellular organismal process
actin-mediated cell contraction
muscle contraction
muscle hypertrophy
signal transduction
regulation of cellular component movement
positive regulation of cellular process
positive regulation of locomotion
positive regulation of immune system process
negative regulation of response to stimulus
regulation of response to stress
regulation of system process
actin-myosin filament sliding
striated muscle contraction
striated muscle hypertrophy
cell surface receptor signaling pathway
positive regulation of cellular component movement
regulation of cell motility
regulation of defense response
regulation of blood circulation
muscle filament sliding
cardiac muscle contraction
cardiac muscle hypertrophy
cytokine-mediated signaling pathway
positive regulation of cell motility
regulation of cell migration
negative regulation of defense response
regulation of heart contraction
positive regulation of cell migration
regulation of leukocyte migration
regulation of the force of heart contraction
positive regulation of leukocyte migration

| GO term | Description | P-value | FDR q-value | Enrichment (N, B, n, b) |
|---|---|---|---|---|
| GO:0019221 | cytokine-mediated signaling pathway | 4.56E-7 | 6.95E-3 | 13.15 (17785,526,18,7) |
| GO:0030049 | muscle filament sliding | 5.11E-6 | 3.89E-2 | 87.18 (17785,34,18,3) |
| GO:0033275 | actin-myosin filament sliding | 5.11E-6 | 2.6E-2 | 87.18 (17785,34,18,3) |
| GO:0070252 | actin-mediated cell contraction | 1.12E-5 | 4.28E-2 | 67.37 (17785,44,18,3) |
| GO:0030048 | actin filament-based movement | 3.02E-5 | 9.21E-2 | 48.59 (17785,61,18,3) |

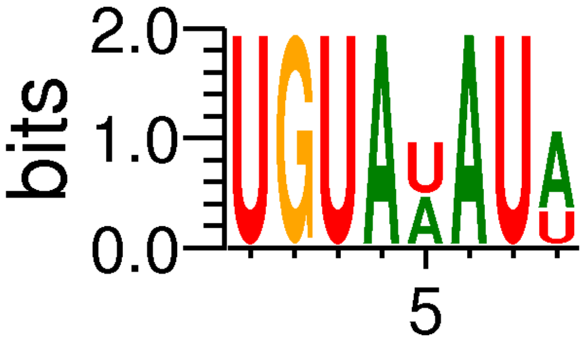
| | | | | |
|---|---|---|---|---|
| GO:0002687 | positive regulation of leukocyte migration | 1.62E-4 | 4.11E-1 | 27.70 (17785,107,18,3) |
| GO:0002376 | immune system process | 1.67E-4 | 3.64E-1 | 4.46 (17785,1772,18,8) |
| GO:0003300 | cardiac muscle hypertrophy | 1.82E-4 | 3.47E-1 | 98.81 (17785,20,18,2) |
| GO:0014897 | striated muscle hypertrophy | 2.01E-4 | 3.4E-1 | 94.10 (17785,21,18,2) |
| GO:0014896 | muscle hypertrophy | 2.21E-4 | 3.37E-1 | 89.82 (17785,22,18,2) |
| GO:0002026 | regulation of the force of heart contraction | 2.64E-4 | 3.65E-1 | 82.34 (17785,24,18,2) |
| GO:0060048 | cardiac muscle contraction | 4.14E-4 | 5.26E-1 | 65.87 (17785,30,18,2) |
| GO:0050896 | response to stimulus | 4.31E-4 | 5.06E-1 | 2.51 (17785,4731,18,12) |
| GO:0002685 | regulation of leukocyte migration | 5.38E-4 | 5.86E-1 | 18.41 (17785,161,18,3) |
| GO:0006936 | muscle contraction | 8.18E-4 | 8.32E-1 | 15.94 (17785,186,18,3) |
| GO:0031348 | negative regulation of defense response | 8.18E-4 | 7.8E-1 | 15.94 (17785,186,18,3) |

d. This time Gorilla found less values and with higher P-values than in the previous run. We see that the significant results are mainly in the subtrees of ‘cellular-process’ and ‘multicellular organismal process’, most significant ones being cytokine-mediated signaling pathway, muscle filament sliding, actin-myosin filament sliding, actin-mediated cell contraction.

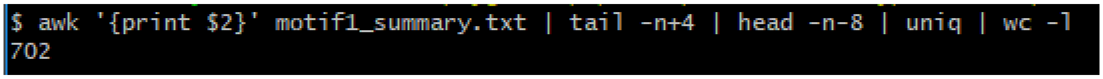
3. In the first run we got a larger amount of enriched terms, the highest ranked ones were more significant than the highest ranked terms in the second run. Moreover, some of the terms that we got in the second run were not recognized as enriched terms in the first run at all.
The difference in results might be due to the size of the target list, obtained as described in section 2a. After filtering out the complete gene list, only 20 genes remained with the adjusted p-value and log-fold values that match the conditions, while the background list contains over 50,000 genes.

## Question 3

2. The resulting motif is:



3. Since a sequence may contain more than one occurrence we will have to uniq filter our results. I took the motif occurrences file motif1_summary.txt and ran the following linux command to get the amount of different sequences:



```
$ awk '{print $2}' motif1_summary.txt | tail -n+4 | head -n-8 | uniq | wc -l
702
```

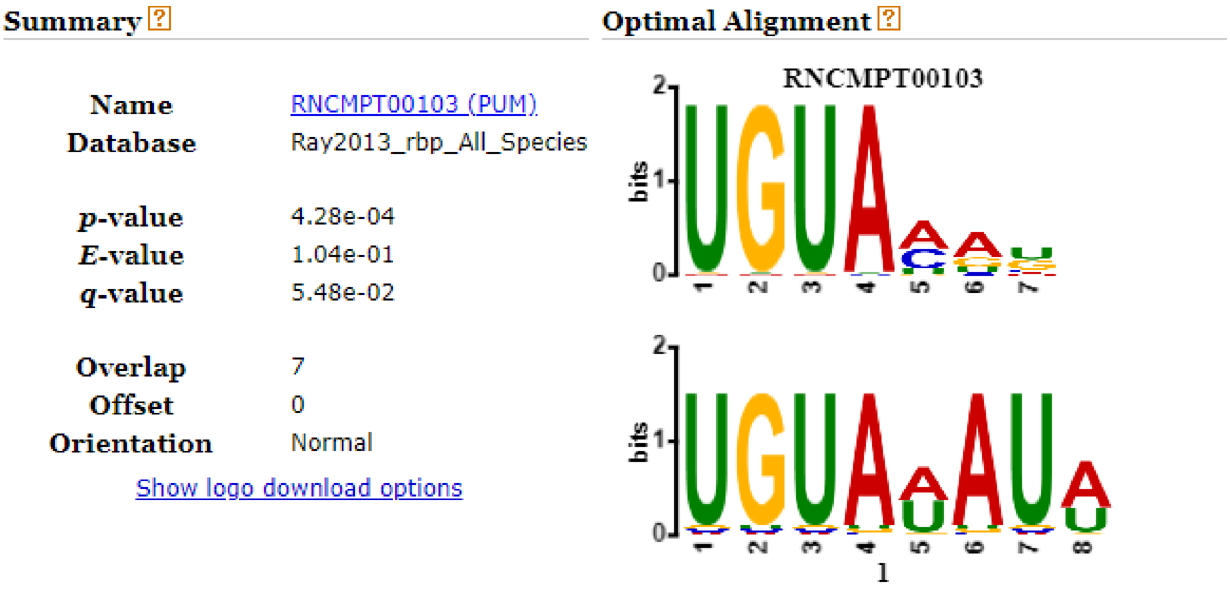
So we have 702 sequences that contain a k-mer associated with the motif.

4. PSSM:

| PSSM: | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ----- | | | | | | | | |
| A | 0 | 0 | 0 | 1 | 0.52 | 1 | 0 | 0.66 |
| T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| U | 1 | 0 | 1 | 0 | 0.48 | 0 | 1 | 0.34 |

What can you say about the motif?

This is the motif we found in the first result of TOMTOM search



Summary ?

| | |
|---|---|
| Name | RNCMPT00103 (PUM) |
| Database | Ray2013_rbp_All_Species |
| p-value | 4.28e-04 |
| E-value | 1.04e-01 |
| q-value | 5.48e-02 |
| Overlap | 7 |
| Offset | 0 |
| Orientation | Normal |

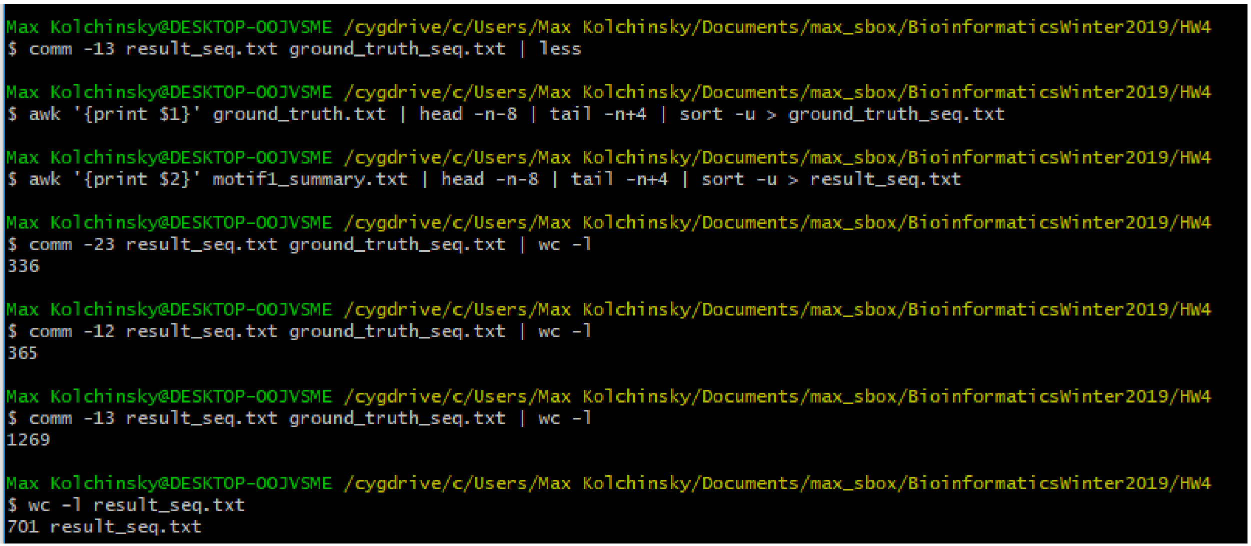
Show logo download options

Optimal Alignment ?

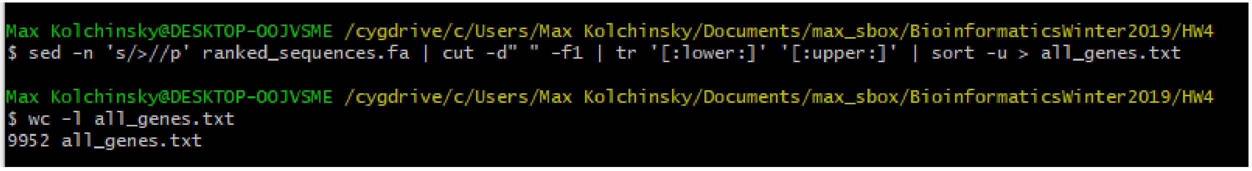
What possible experiment generated the ranked list of sequences?

The list of sequences is a fasta file. An example for a possible experiment that made this file could be PCA on a patient being treated for cancer.

6. The following commands gave me the results:



```
Max Kolchinsky@DESKTOP-OOJVSME /cygdrive/c/Users/Max Kolchinsky/Documents/max_sbox/BioinformaticsWinter2019/HW4
$ comm -13 result_seq.txt ground_truth_seq.txt | less

Max Kolchinsky@DESKTOP-OOJVSME /cygdrive/c/Users/Max Kolchinsky/Documents/max_sbox/BioinformaticsWinter2019/HW4
$ awk '{print $1}' ground_truth.txt | head -n-8 | tail -n+4 | sort -u > ground_truth_seq.txt

Max Kolchinsky@DESKTOP-OOJVSME /cygdrive/c/Users/Max Kolchinsky/Documents/max_sbox/BioinformaticsWinter2019/HW4
$ awk '{print $2}' motif1_summary.txt | head -n-8 | tail -n+4 | sort -u > result_seq.txt

Max Kolchinsky@DESKTOP-OOJVSME /cygdrive/c/Users/Max Kolchinsky/Documents/max_sbox/BioinformaticsWinter2019/HW4
$ comm -23 result_seq.txt ground_truth_seq.txt | wc -l
336

Max Kolchinsky@DESKTOP-OOJVSME /cygdrive/c/Users/Max Kolchinsky/Documents/max_sbox/BioinformaticsWinter2019/HW4
$ comm -12 result_seq.txt ground_truth_seq.txt | wc -l
365

Max Kolchinsky@DESKTOP-OOJVSME /cygdrive/c/Users/Max Kolchinsky/Documents/max_sbox/BioinformaticsWinter2019/HW4
$ comm -13 result_seq.txt ground_truth_seq.txt | wc -l
1269

Max Kolchinsky@DESKTOP-OOJVSME /cygdrive/c/Users/Max Kolchinsky/Documents/max_sbox/BioinformaticsWinter2019/HW4
$ wc -l result_seq.txt
701 result_seq.txt
```



```
Max Kolchinsky@DESKTOP-OOJVSME /cygdrive/c/Users/Max Kolchinsky/Documents/max_sbox/BioinformaticsWinter2019/HW4
$ sed -n 's/>//p' ranked_sequences.fa | cut -d" " -f1 | tr '[:lower:]' '[:upper:]' | sort -u > all_genes.txt

Max Kolchinsky@DESKTOP-OOJVSME /cygdrive/c/Users/Max Kolchinsky/Documents/max_sbox/BioinformaticsWinter2019/HW4
$ wc -l all_genes.txt
9952 all_genes.txt
```

So:

Left_only = 336

Right_only = 1269

Both = 365

TotalGenesLeft = 701

TP = Both = 365

FP = TotalGenesLeft – Both = 701 – 365 = 336

FN = Right_only = 1269

TN = TotalGenes – FP = 9952 – 1269 = 8636

From which I can conclude that my confusion matrix will be like so:

| | Predicted:<br>No | Predicted:<br>Yes |
|---|---|---|
| Actual:<br>No | 8636 | 336 |

| | | |
|---|---|---|
| Actual:<br>Yes | 1269 | 365 |

7. Calculation:

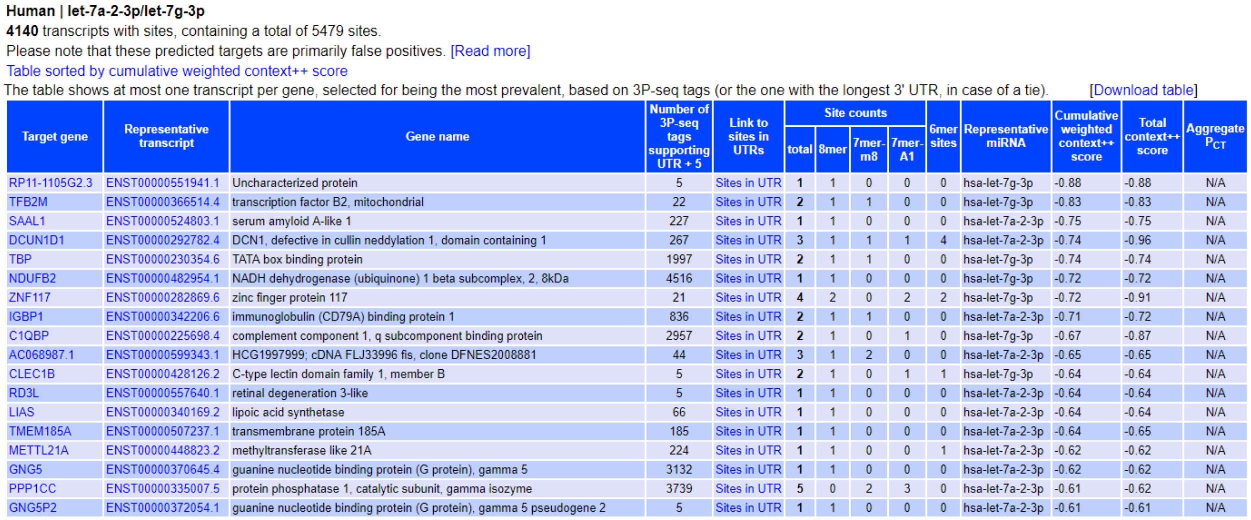
$$Sensitivity = \frac{TP}{TP + FN} = \frac{365}{365 + 1269} = 0.223 = 22.3\%$$

$$Specifity = \frac{TN}{TN + FP} = \frac{8636}{8363 + 336} = 0.962 = 96.2\%$$

8. If the tool is very complex it is also very slow. Since we already have 96.2% specifity if we value specifity a lot more than sensitivity and we also value fast runtime we will prefer to run the motif search to get almost 100% results a lot faster.

## Question 4

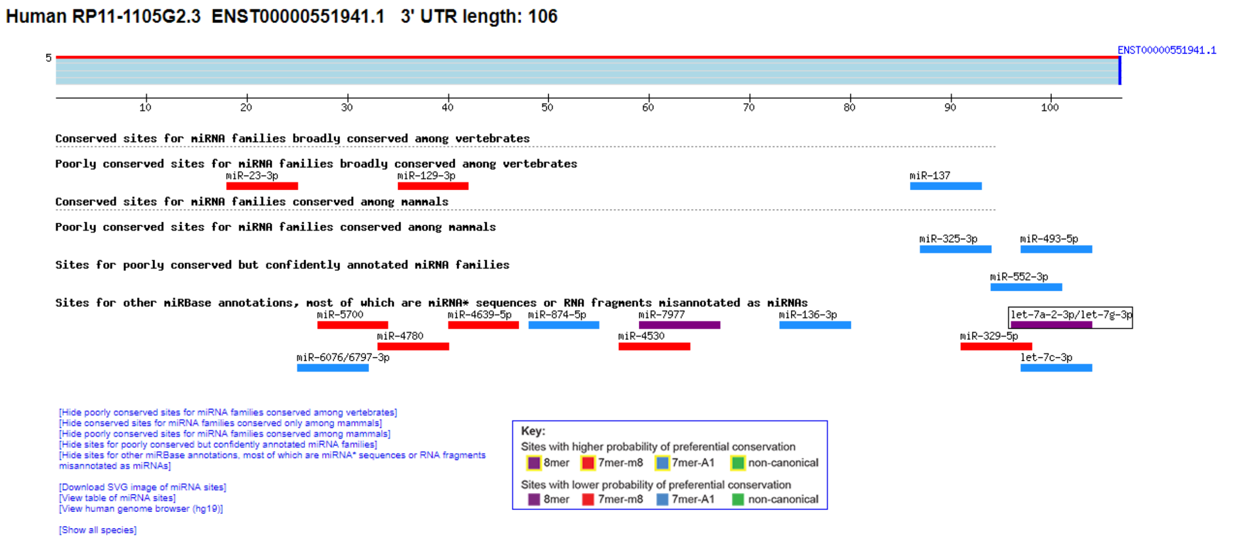
1. Screenshot of the results:



**Human | let-7a-2-3p/let-7g-3p**
**4140** transcripts with sites, containing a total of 5479 sites.
Please note that these predicted targets are primarily false positives. [Read more]
Table sorted by cumulative weighted context++ score
The table shows at most one transcript per gene, selected for being the most prevalent, based on 3P-seq tags (or the one with the longest 3' UTR, in case of a tie). [Download table]

| Target gene | Representative transcript | Gene name | Number of 3P-seq tags supporting UTR + 5 | Link to sites in UTRs | Site counts | | | | 6mer sites | Representative miRNA | Cumulative weighted context++ score | Total context++ score | Aggregate $P_{CT}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | total | 8mer | 7mer-m8 | 7mer-A1 | | | | | |
| RP11-1105G2.3 | ENST00000551941.1 | Uncharacterized protein | 5 | Sites in UTR | **1** | 1 | 0 | 0 | 0 | hsa-let-7g-3p | -0.88 | -0.88 | N/A |
| TFB2M | ENST00000366514.4 | transcription factor B2, mitochondrial | 22 | Sites in UTR | **2** | 1 | 1 | 0 | 0 | hsa-let-7g-3p | -0.83 | -0.83 | N/A |
| SAAL1 | ENST00000524803.1 | serum amyloid A-like 1 | 227 | Sites in UTR | **1** | 1 | 0 | 0 | 0 | hsa-let-7a-2-3p | -0.75 | -0.75 | N/A |
| DCUN1D1 | ENST00000292782.4 | DCN1, defective in cullin neddylation 1, domain containing 1 | 267 | Sites in UTR | **3** | 1 | 1 | 1 | 4 | hsa-let-7a-2-3p | -0.74 | -0.96 | N/A |
| TBP | ENST00000230354.6 | TATA box binding protein | 1997 | Sites in UTR | **2** | 1 | 1 | 0 | 0 | hsa-let-7g-3p | -0.74 | -0.74 | N/A |
| NDUFB2 | ENST00000482954.1 | NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 2, 8kDa | 4516 | Sites in UTR | **1** | 1 | 0 | 0 | 0 | hsa-let-7g-3p | -0.72 | -0.72 | N/A |
| ZNF117 | ENST00000282869.6 | zinc finger protein 117 | 21 | Sites in UTR | **4** | 2 | 0 | 2 | 2 | hsa-let-7g-3p | -0.72 | -0.91 | N/A |
| IGBP1 | ENST00000342206.6 | immunoglobulin (CD79A) binding protein 1 | 836 | Sites in UTR | **2** | 1 | 1 | 0 | 0 | hsa-let-7a-2-3p | -0.71 | -0.72 | N/A |
| C1QBP | ENST00000225698.4 | complement component 1, q subcomponent binding protein | 2957 | Sites in UTR | **2** | 1 | 0 | 1 | 0 | hsa-let-7g-3p | -0.67 | -0.87 | N/A |
| AC068987.1 | ENST00000599343.1 | HCG1997999; cDNA FLJ33996 fis, clone DFNES2008881 | 44 | Sites in UTR | **3** | 1 | 2 | 0 | 0 | hsa-let-7a-2-3p | -0.65 | -0.65 | N/A |
| CLEC1B | ENST00000428126.2 | C-type lectin domain family 1, member B | 5 | Sites in UTR | **2** | 1 | 0 | 1 | 1 | hsa-let-7g-3p | -0.64 | -0.64 | N/A |
| RD3L | ENST00000557640.1 | retinal degeneration 3-like | 5 | Sites in UTR | **1** | 1 | 0 | 0 | 0 | hsa-let-7a-2-3p | -0.64 | -0.64 | N/A |
| LIAS | ENST00000340169.2 | lipoic acid synthetase | 66 | Sites in UTR | **1** | 1 | 0 | 0 | 0 | hsa-let-7a-2-3p | -0.64 | -0.64 | N/A |
| TMEM185A | ENST00000507237.1 | transmembrane protein 185A | 185 | Sites in UTR | **1** | 1 | 0 | 0 | 0 | hsa-let-7a-2-3p | -0.64 | -0.65 | N/A |
| METTL21A | ENST00000448823.2 | methyltransferase like 21A | 224 | Sites in UTR | **1** | 1 | 0 | 0 | 1 | hsa-let-7a-2-3p | -0.62 | -0.62 | N/A |
| GNG5 | ENST00000370645.4 | guanine nucleotide binding protein (G protein), gamma 5 | 3132 | Sites in UTR | **1** | 1 | 0 | 0 | 0 | hsa-let-7a-2-3p | -0.62 | -0.62 | N/A |
| PPP1CC | ENST00000335007.5 | protein phosphatase 1, catalytic subunit, gamma isozyme | 3739 | Sites in UTR | **5** | 0 | 2 | 3 | 0 | hsa-let-7a-2-3p | -0.61 | -0.62 | N/A |
| GNG5P2 | ENST00000372054.1 | guanine nucleotide binding protein (G protein), gamma 5 pseudogene 2 | 5 | Sites in UTR | **1** | 1 | 0 | 0 | 0 | hsa-let-7a-2-3p | -0.61 | -0.61 | N/A |

2. Predicted target sites on the UTR of target gene RP11-1105G2.3



**Human RP11-1105G2.3 ENST00000551941.1 3' UTR length: 106**

5
ENST00000551941.1
10
20
30
40
50
60
70
80
90
100
Conserved sites for miRNA families broadly conserved among vertebrates
Poorly conserved sites for miRNA families broadly conserved among vertebrates
miR-23-3p
miR-129-3p
miR-137
Conserved sites for miRNA families conserved among mammals
Poorly conserved sites for miRNA families conserved among mammals
miR-325-3p
miR-493-5p
Sites for poorly conserved but confidently annotated miRNA families
miR-552-3p
Sites for other miRBase annotations, most of which are miRNA* sequences or RNA fragments misannotated as miRNAs
miR-5700
miR-4639-5p
miR-874-5p
miR-7977
miR-136-3p
let-7a-2-3p/let-7g-3p
miR-4780
miR-4530
miR-329-5p
miR-6076/6797-3p
let-7c-3p

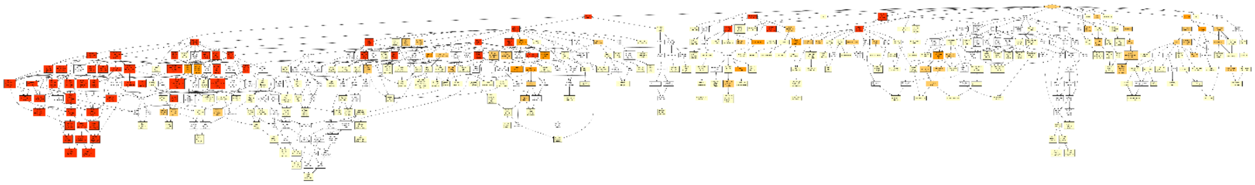
[Hide poorly conserved sites for miRNA families conserved among vertebrates]
[Hide conserved sites for miRNA families conserved only among mammals]
[Hide poorly conserved sites for miRNA families conserved among mammals]
[Hide sites for poorly conserved but confidently annotated miRNA families]
[Hide sites for other miRBase annotations, most of which are miRNA* sequences or RNA fragments misannotated as miRNAs]

[Download SVG image of miRNA sites]
[View table of miRNA sites]
[View human genome browser (hg19)]

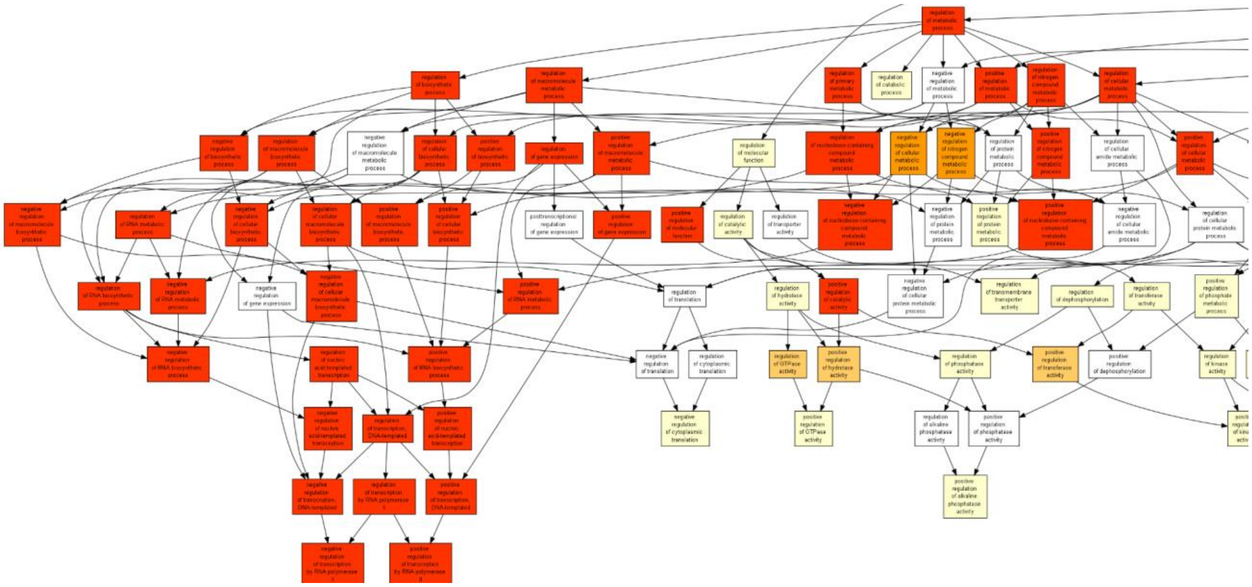
[Show all species]

Key:
Sites with higher probability of preferential conservation
🟪 8mer 🟥 7mer-m8 🟦 7mer-A1 🟩 non-canonical
Sites with lower probability of preferential conservation
🟪 8mer 🟥 7mer-m8 🟦 7mer-A1 🟩 non-canonical

3. I used background_list.txt from Q2 as the background of the search and the table from section 1 as the target list.
   Results of Gorilla (for biological process):



From what we can see in the Gorilla result (after zooming in) is that there is considerable differentiation in many categories which start at the “biological regulation” sub tree while most of the results are under “regulation of metabolic process” as can be seen in the following zoomed in part:


regulation of metabolic process
regulation of biosynthetic process
regulation of macromolecule metabolic process
regulation of primary metabolic process
regulation of catabolic process
negative regulation of metabolic process
positive regulation of metabolic process
regulation of nitrogen compound metabolic process
regulation of cellular metabolic process
negative regulation of biosynthetic process
regulation of macromolecule biosynthetic process
negative regulation of macromolecule metabolic process
regulation of cellular biosynthetic process
positive regulation of biosynthetic process
regulation of gene expression
positive regulation of macromolecule metabolic process
regulation of molecular function
regulation of nucleobase-containing compound metabolic process
negative regulation of cellular metabolic process
negative regulation of nitrogen compound metabolic process
regulation of protein metabolic process
positive regulation of nitrogen compound metabolic process
regulation of cellular amide metabolic process
positive regulation of cellular metabolic process
negative regulation of macromolecule biosynthetic process
regulation of RNA metabolic process
negative regulation of cellular biosynthetic process
regulation of cellular macromolecule biosynthetic process
positive regulation of macromolecule biosynthetic process
positive regulation of cellular biosynthetic process
posttranscriptional regulation of gene expression
positive regulation of gene expression
positive regulation of molecular function
regulation of catalytic activity
regulation of transporter activity
negative regulation of nucleobase-containing compound metabolic process
negative regulation of protein metabolic process
positive regulation of protein metabolic process
positive regulation of nucleobase-containing compound metabolic process
negative regulation of cellular amide metabolic process
regulation of cellular protein metabolic process
regulation of RNA biosynthetic process
negative regulation of RNA metabolic process
negative regulation of gene expression
negative regulation of cellular macromolecule biosynthetic process
positive regulation of RNA metabolic process
regulation of translation
regulation of hydrolase activity
positive regulation of catalytic activity
negative regulation of cellular protein metabolic process
regulation of transmembrane transporter activity
regulation of dephosphorylation
regulation of transferase activity
positive regulation of phosphate metabolic process
negative regulation of RNA biosynthetic process
regulation of nucleic acid-templated transcription
positive regulation of RNA biosynthetic process
negative regulation of translation
regulation of cytoplasmic translation
regulation of GTPase activity
positive regulation of hydrolase activity
regulation of phosphatase activity
positive regulation of transferase activity
positive regulation of dephosphorylation
regulation of kinase activity
negative regulation of nucleic acid-templated transcription
regulation of transcription, DNA-templated
positive regulation of nucleic acid-templated transcription
negative regulation of cytoplasmic translation
positive regulation of GTPase activity
regulation of alkaline phosphatase activity
positive regulation of phosphatase activity
negative regulation of transcription, DNA-templated
regulation of transcription by RNA polymerase II
positive regulation of transcription, DNA-templated
positive regulation of alkaline phosphatase activity
negative regulation of transcription by RNA polymerase II
positive regulation of transcription by RNA polymerase II

As we can see in the table below there are other differentially expressed genes with higher p-value as can be seen here:

| GO term | Description | P-value | FDR q-value | Enrichment (N, B, n, b) | Genes |
|---|---|---|---|---|---|
| GO:0051252 | regulation of RNA metabolic process | 2.42E-27 | 3.72E-23 | 1.32 (18016,3456,3966,1003) | [+] Show genes |

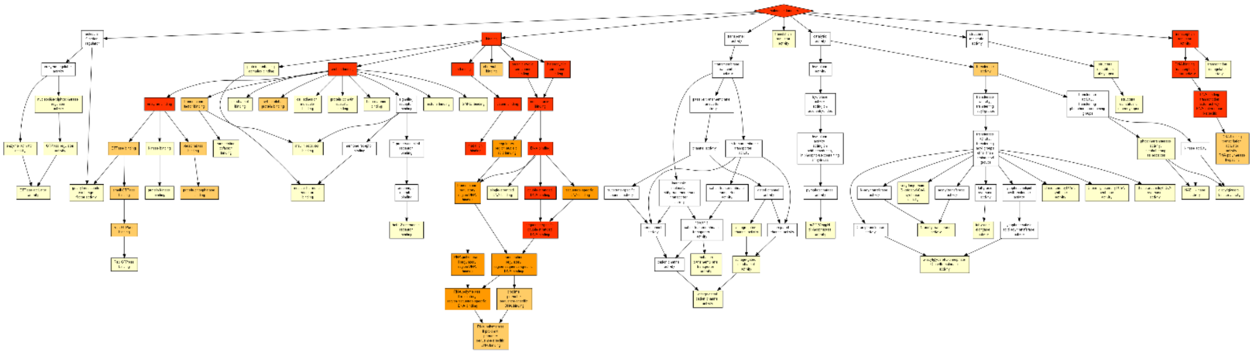
| | | | | | |
|---|---|---|---|---|---|
| GO:0019219 | regulation of nucleobase-containing compound metabolic process | 5.48E-27 | 4.21E-23 | 1.30 (18016,3709,3966,1063) | [+] Show genes |
| GO:2000112 | regulation of cellular macromolecule biosynthetic process | 2.51E-25 | 1.29E-21 | 1.30 (18016,3612,3966,1031) | [+] Show genes |
| GO:1903506 | regulation of nucleic acid-templated transcription | 1.23E-24 | 4.72E-21 | 1.31 (18016,3221,3966,932) | [+] Show genes |
| GO:0010556 | regulation of macromolecule biosynthetic process | 1.28E-24 | 3.93E-21 | 1.29 (18016,3722,3966,1054) | [+] Show genes |
| GO:2001141 | regulation of RNA biosynthetic process | 1.36E-24 | 3.49E-21 | 1.31 (18016,3226,3966,933) | [+] Show genes |
| GO:0006355 | regulation of transcription, DNA-templated | 1.52E-24 | 3.34E-21 | 1.32 (18016,3170,3966,919) | [+] Show genes |
| GO:0031326 | regulation of cellular biosynthetic process | 1.25E-23 | 2.4E-20 | 1.27 (18016,3860,3966,1082) | [+] Show genes |
| GO:0009889 | regulation of biosynthetic process | 2.43E-23 | 4.15E-20 | 1.27 (18016,3929,3966,1097) | [+] Show genes |
| GO:0006357 | regulation of transcription | 8.87E-22 | 1.36E-18 | 1.35 (18016,2475,3966,733) | [+] Show genes |

| | | | | | |
|---|---|---|---|---|---|
| | by RNA polymerase II | | | | |
| GO:0031323 | regulation of cellular metabolic process | 7.93E-21 | 1.11E-17 | 1.19 (18016,5665,3966,1490) | [+] Show genes |
| GO:0080090 | regulation of primary metabolic process | 7.12E-20 | 9.11E-17 | 1.19 (18016,5536,3966,1454) | [+] Show genes |
| GO:0051171 | regulation of nitrogen compound metabolic process | 8.82E-20 | 1.04E-16 | 1.20 (18016,5383,3966,1418) | [+] Show genes |
| GO:0016043 | cellular component organization | 9.36E-18 | 1.03E-14 | 1.20 (18016,4740,3966,1255) | [+] Show genes |
| GO:0071840 | cellular component organization or biogenesis | 1.73E-17 | 1.77E-14 | 1.20 (18016,4782,3966,1263) | [+] Show genes |
| GO:0048522 | positive regulation of cellular process | 3.78E-17 | 3.63E-14 | 1.19 (18016,4895,3966,1287) | [+] Show genes |
| GO:0048518 | positive regulation of biological process | 7.16E-17 | 6.47E-14 | 1.18 (18016,5478,3966,1420) | [+] Show genes |
| GO:0032502 | developmental process | 8.69E-17 | 7.42E-14 | 1.20 (18016,4567,3966,1208) | [+] Show genes |
| GO:0050794 | regulation of cellular process | 1.49E-16 | 1.2E-13 | 1.10 (18016,9862,3966,2397) | [+] Show genes |

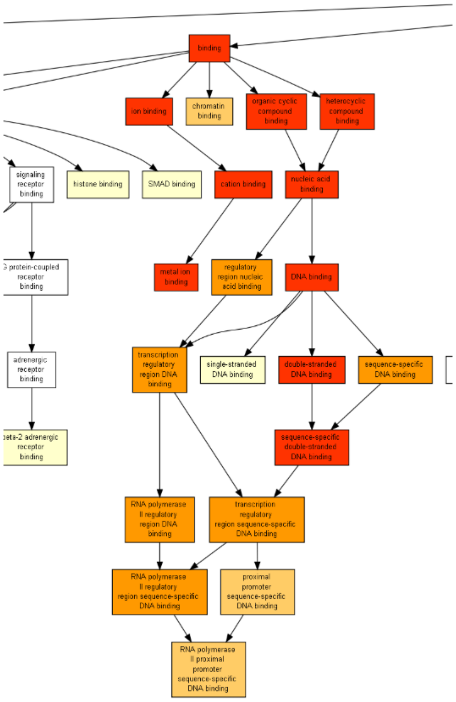
| | | | | | |
|---|---|---|---|---|---|
| GO:0051254 | positive regulation of RNA metabolic process | 1.91E-15 | 1.46E-12 | 1.37 (18016,1591,3966,479) | [+] Show genes |

These are the top go terms from the table of results sorted by p-value. The table of results sorts the results in the go tree by p-value to show the order of differentiation of each term by p-value, fdr q-value and Enrichment.

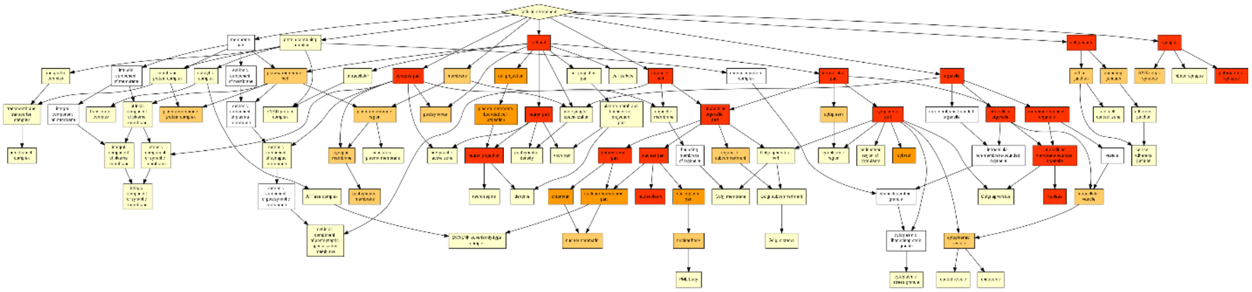
Here are the results for differentiation in molecular function:



Here we can see considerable differentiation mostly in the "binding" sub tree as can be see zoomed in here:


binding
ion binding
chromatin binding
organic cyclic compound binding
heterocyclic compound binding
signaling receptor binding
histone binding
SMAD binding
cation binding
nucleic acid binding
protein-coupled receptor binding
metal ion binding
regulatory region nucleic acid binding
DNA binding
adrenergic receptor binding
transcription regulatory region DNA binding
single-stranded DNA binding
double-stranded DNA binding
sequence-specific DNA binding
eta-2 adrenergic receptor binding
sequence-specific double-stranded DNA binding
RNA polymerase II regulatory region DNA binding
transcription regulatory region sequence-specific DNA binding
RNA polymerase II regulatory region sequence-specific DNA binding
proximal promoter sequence-specific DNA binding
RNA polymerase II proximal promoter sequence-specific DNA binding

And the results for cellular component:



Here differentiation is not mostly under a single sub tree but is mostly spread around all sub trees.

4. A phenotype of this miRNA will most likely be associated with a gene which is differentially expressed according to Gorilla. We will go to the biological process go tree and check the genes associated with the first go term (the go term with the lowest p-value which is the first one in the table).
   We will pick a gene from the list, for example, GFI1 (growth factor independent 1 transcription repressor).
   According to NCBI, potential phenotypes of this gene are:
   Neutropenia, nonimmune chronic idiopathic, of adults
   Severe congenital neutropenia 2, autosomal dominant

# HW4 - Gene expression - Clustering and Enrichment

Submitted by:
ID1: 320983216 Name1: Maxim Kolchinsky

ID2: 321340580 Name2: Anna Romanov

## The Biology

The data for this lesson comes from:
> Saigi *et al.* “MET-Oncogenic and JAK2-Inactivating Alterations Are Independent Factors That Affect Regulation of PD-L1 Expression in Lung Cancer” *PLoS ONE*. 2018 Jun 13;9(6):e99625. PMID: 24926665.

Purpose: The blockade of immune checkpoints such as PDL1 and PD-1 is being exploited therapeutically in several types of malignancies. Here, we aimed to understand the contribution of the genetics of lung cancer to the ability of tumor cells to escape immunosurveillance checkpoints. Experimental Design: More than 150 primary non-small cell lung cancers, including pulmonary sarcomatoid carcinomas, were tested for levels of the HLA-I complex, PD-L1, tumor-infiltrating CD8þ lymphocytes, and alterations in main lung cancer genes. Correlations were validated in cancer cell lines using appropriate treatments to activate or inhibit selected pathways. We also performed RNA sequencing to assess changes in gene expression after these treatments. Results: MET-oncogenic activation tended to associate with positive PD-L1 immunostaining, whereas STK11 mutations were correlated with negative immunostaining. In MET-altered cancer cells, MET triggered a transcriptional increase of PD-L1 that was independent of the IFNgmediated JAK/STAT pathway. The activation of MET also upregulated other immunosuppressive genes (PDCD1LG2 and SOCS1) and transcripts involved in angiogenesis (VEGFA and NRP1) and in cell proliferation. We also report recurrent inactivating mutations in JAK2 that co-occur with alterations in MET and STK11, which prevented the induction of immunoresponse-related genes following treatment with IFNg. Conclusions: We show that MET activation promotes the expression of several negative checkpoint regulators of the immunoresponse, including PD-L1. In addition, we report inactivation of JAK2 in lung cancer cells that prevented the response to IFNg. These alterations are likely to facilitate tumor growth by enabling immune tolerance and may affect the response to immune checkpoint inhibitors

## Data

This data was downloaded from GEO (GSE:GSE109720)

### Import count data and metadata

```r
library(readr)
library(dplyr)
library(ggplot2)

rawcounts <- read_csv("data/lung_counts.csv")
metadata <-  read_csv("data/lung_metadata.csv")

rawcounts
```

```
## # A tibble: 58,347 x 34
##    ensgene AE1148 AE1149 AE1150 AE1151 AE1152 AE1153 AE1154 AE1155 AE1156
##    <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 ENSG00~      0      0      5      0      7      0      0      0      5
##  2 ENSG00~    374    529    304    366    466    519    319    379    326
##  3 ENSG00~      0      0      0      0      0      0      0      0      0
##  4 ENSG00~      0      0      0      0      0      0      0      0      0
##  5 ENSG00~      0      0      0      0      0      0      0      0      0
##  6 ENSG00~      0      0      0      0      0      0      0      0      0
##  7 ENSG00~      0      0      0      0      0      0      0      0      0
##  8 ENSG00~      0      0      0      0      0      0      0      0      0
##  9 ENSG00~      0      0      0      0      0      0      0      0      0
## 10 ENSG00~     11      1      8     24     18     21     25     17     17
## # ... with 58,337 more rows, and 24 more variables: AE1157 <dbl>,
## #   AE1158 <dbl>, AE1159 <dbl>, AE1160 <dbl>, AE1161 <dbl>, AE1162 <dbl>,
## #   AE1163 <dbl>, AE1164 <dbl>, AE1165 <dbl>, AE1166 <dbl>, AE1167 <dbl>,
## #   AE1168 <dbl>, AE1169 <dbl>, AE1170 <dbl>, AE1171 <dbl>, AE1172 <dbl>,
## #   AE1173 <dbl>, AE1174 <dbl>, AE1175 <dbl>, AE1176 <dbl>, AE1177 <dbl>,
## #   AE1178 <dbl>, AE1179 <dbl>, AE1180 <dbl>
```

```
metadata
```

```
## # A tibble: 33 x 4
##    id     celltype dex          geo_id
##    <chr>  <chr>    <chr>        <chr>
##  1 AE1160 EBC1     No treatment GSM2949380
##  2 AE1163 EBC1     No treatment GSM2949381
##  3 AE1166 EBC1     No treatment GSM2949382
##  4 AE1162 EBC1     Crizotinib   GSM2949383
##  5 AE1165 EBC1     Crizotinib   GSM2949384
##  6 AE1168 EBC1     Crizotinib   GSM2949385
##  7 AE1161 EBC1     Interferon <U+03B3> GSM2949386
##  8 AE1164 EBC1     Interferon <U+03B3> GSM2949387
##  9 AE1167 EBC1     Interferon <U+03B3> GSM2949388
## 10 AE1169 H1573    No treatment GSM2949389
## # ... with 23 more rows
```

### Clustering - k-means

```
library(tibble)
library(tidyr)
# a plotting function
plot_heatmap <- function(km) {
  centers <- km$centers %>%
      tbl_df() %>%
      rownames_to_column('Cluster') %>%
      gather(Sample, value, -Cluster) %>%
      mutate(
          Cluster = factor(Cluster),
          Sample = factor(Sample)
      )
   ggplot(centers,aes(Sample,Cluster)) + geom_tile(aes(fill=value)) + geom_text(aes(label = round(value
}
```

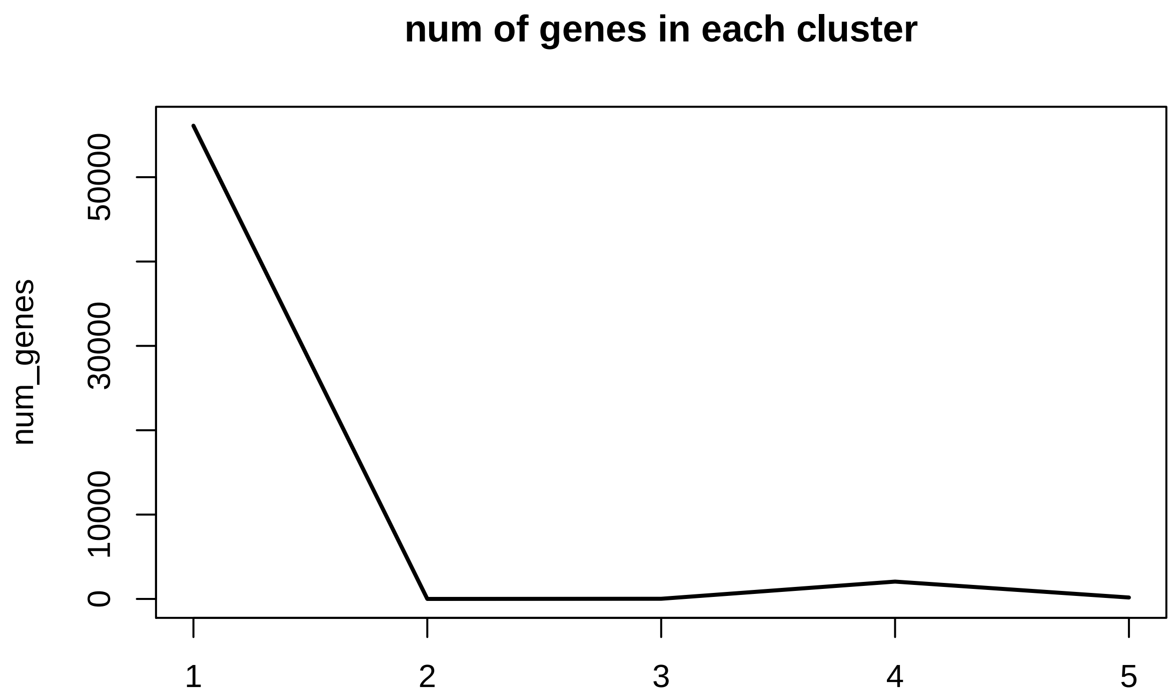
```r
# cluster the genes

rawcounts_RN<- data.frame(rawcounts, row.names=1)
km <- kmeans(rawcounts_RN,5)
km$centers
```

```
##        AE1148      AE1149      AE1150     AE1151      AE1152      AE1153
## 1    204.3794    260.8659    184.8346    225.175    226.5053    263.4797
## 2 261075.5000 355397.5000 249188.5000 298081.500 312029.0000 330232.5000
## 3 101059.4762 147734.7143  89713.5238 117618.429 120754.0000 151602.5238
## 4   6589.0833   8841.2332   6127.4839   7358.684   7346.2171   8890.0477
## 5  31052.4260  42212.6686  27571.1893  34696.260  32887.3373  41325.6095
##        AE1154      AE1155      AE1156      AE1157      AE1158     AE1159
## 1    177.6818    222.9686    196.5011    232.7849    189.8281    212.521
## 2 222399.5000 302143.5000 258949.5000 295158.0000 257717.0000 284918.500
## 3  90998.6190 124462.2857 104368.4286 134149.9048 106395.0952 117189.476
## 4   5868.3233   7272.5940   6356.2941   8001.1913   6352.8116   7040.299
## 5  26505.6568  33457.2663  29596.4379  37330.6331  30042.5799  32418.935
##       AE1160      AE1161      AE1162      AE1163      AE1164      AE1165
## 1    242.584    180.0566    253.0504    219.7236    191.4948    214.1561
## 2 413308.500 281465.5000 419312.0000 335008.5000 264320.5000 344465.0000
## 3 143324.762 104374.0952 132841.0000 122184.8571 107568.4286 112637.7143
## 4   7666.075   5913.3574   7290.1723   6847.2006   6203.9289   6185.2936
## 5  35868.917  27201.2367  34315.4142  31814.4260  29037.9941  29120.4734
##        AE1166      AE1167      AE1168      AE1169      AE1170      AE1171
## 1    241.4646    198.3505    238.9257    253.7209    189.6234    218.4082
## 2 384117.5000 304171.0000 421639.5000 605629.0000 430845.0000 507135.5000
## 3 135000.0476 111633.0476 126291.2857 169666.4762 124785.2381 141406.8095
## 4   7557.6928   6488.0501   6934.8676   7469.2434   5595.4722   6493.8734
## 5  34757.7219  29575.0533  32625.8402  36791.6982  27468.7456  31881.7219
##        AE1172      AE1173      AE1174     AE1175      AE1176      AE1177
## 1    185.7504    226.5967    210.5955    222.858    183.9239    211.8739
## 2 422003.5000 530275.0000 492796.5000 382136.500 308577.5000 364062.5000
## 3 117254.7619 146959.6667 137927.1429 127871.333 104154.3333 123308.0952
## 4   5550.7936   6760.0141   6311.3476   6946.255   5771.8062   6654.9367
## 5  27203.1006  33370.0118  31380.9112  29166.260  24118.6923  28170.8698
##        AE1178      AE1179      AE1180
## 1    165.6252    178.0214    225.1644
## 2 280669.0000 308508.0000 400801.0000
## 3  94938.3333 104582.6190 133835.4286
## 4   5181.0136   5586.3374   7055.9640
## 5  21687.6036  23821.4024  29880.8935
```

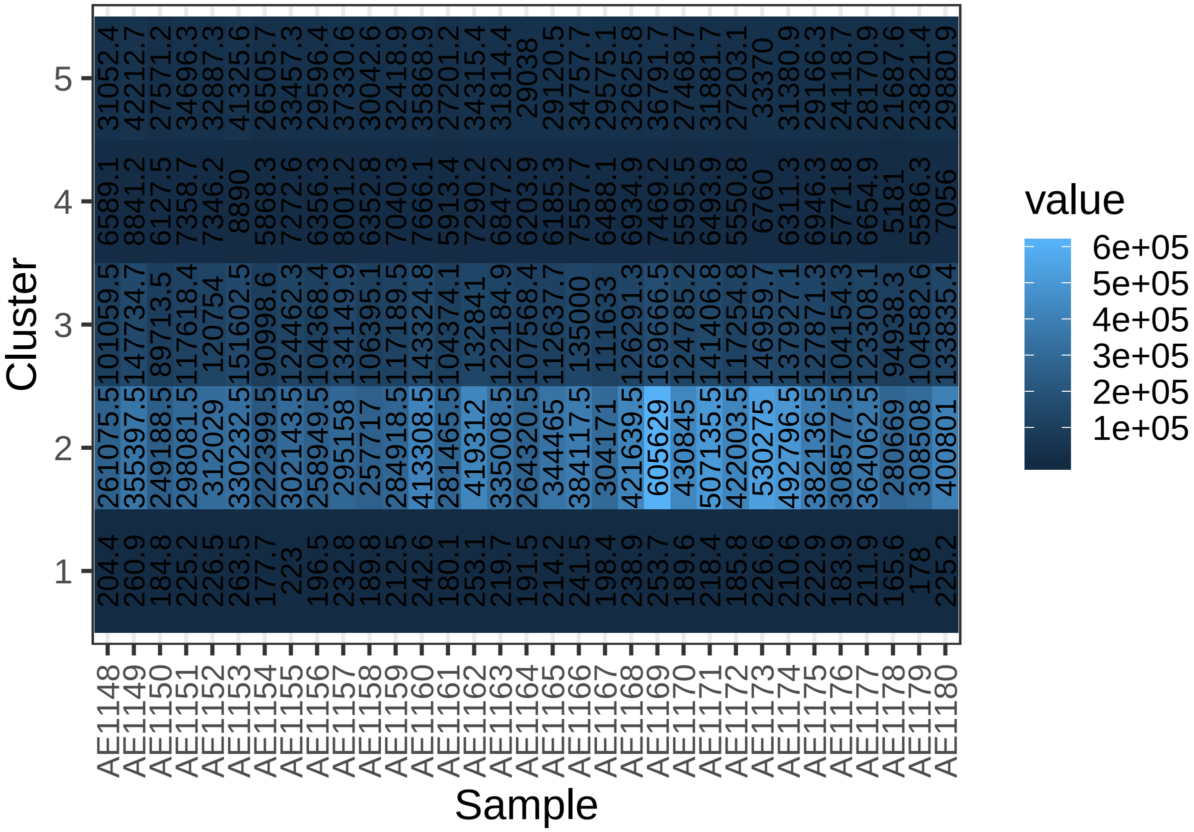
```r
num_genes <- table(km$cluster)

num_genes
```

```
## 
##     1     2     3     4     5 
## 56101     2    21  2054   169
```

```
plot(num_genes, type="l", main='num of genes in each cluster')
```



```
plot_heatmap(km)
```



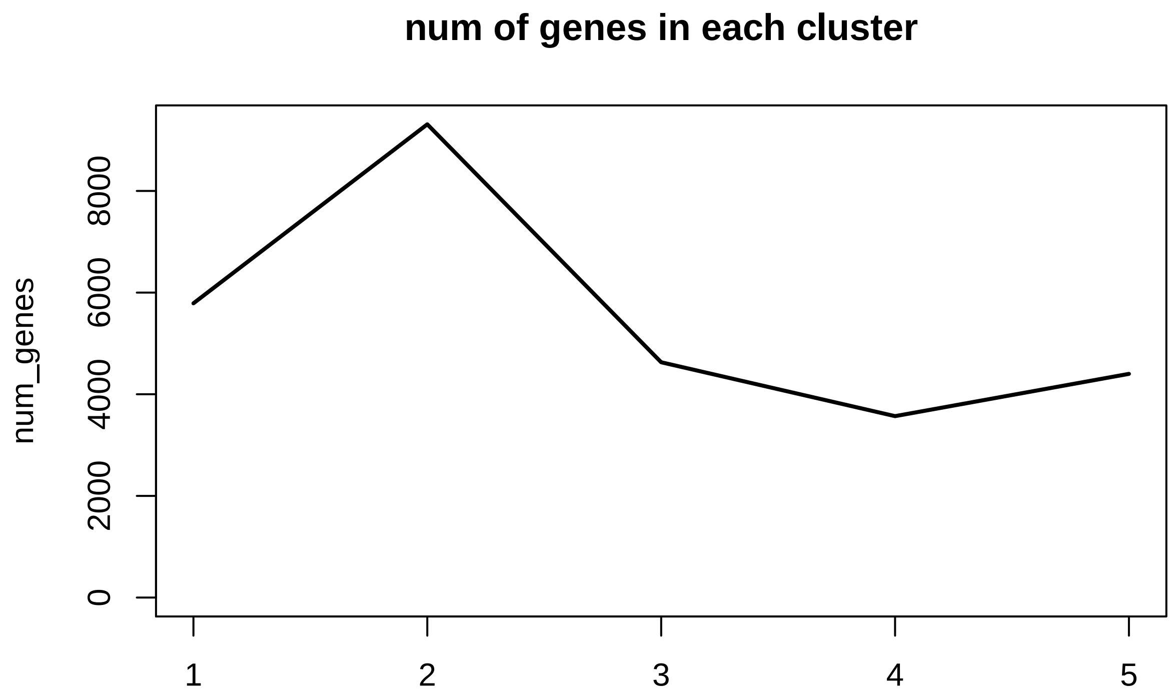
```
# remove rows of zeroes
rawcounts_RN$sum <- rowSums(rawcounts_RN)
rawcounts_clean <- rawcounts_RN[ rawcounts_RN$sum >= 10, ]
rawcounts_clean$sum <- NULL

# apply log scaling
rawcounts_log = log10(1+rawcounts_clean)
km <- kmeans(rawcounts_log,5)
km$centers
```

```
##      AE1148    AE1149    AE1150    AE1151    AE1152    AE1153    AE1154
## 1 2.8768341 2.9629047 2.8159122 2.9162445 2.9151470 2.9548990 2.7748572
## 2 0.2224750 0.2595178 0.2365299 0.2465905 0.2664494 0.2856338 0.2270980
## 3 0.9926849 1.0658002 0.9937652 1.0406866 1.0859241 1.1021633 0.9782394
## 4 2.0053516 2.0718609 1.9683162 2.0457680 2.0753388 2.0995963 1.9408546
## 5 3.5738212 3.6867595 3.5233650 3.6183988 3.6159370 3.6748730 3.4829931
##      AE1155    AE1156    AE1157    AE1158    AE1159    AE1160    AE1161
## 1 2.9035514 2.8555548 2.9088671 2.8171388 2.8838187 2.9247148 2.7762690
## 2 0.2611225 0.2488747 0.2554619 0.2473489 0.2562278 0.2735666 0.2295744
## 3 1.0667283 1.0603057 1.0567061 1.0293423 1.0579908 1.0154241 0.8954655
## 4 2.0608478 2.0370740 2.0484974 1.9979362 2.0467039 1.8938259 1.7459993
## 5 3.6039529 3.5517147 3.6395846 3.5207057 3.5946933 3.6473609 3.5152520
##      AE1162    AE1163    AE1164    AE1165    AE1166    AE1167    AE1168
## 1 2.9481320 2.8845318 2.8080246 2.8748433 2.9221639 2.8221034 2.9211485
## 2 0.3069437 0.2660096 0.2308454 0.2634969 0.2683428 0.2351569 0.2865346
## 3 1.0746557 0.9825569 0.9117918 0.9844687 1.0037558 0.9144569 1.0278350
## 4 1.9523912 1.8613406 1.7745325 1.8706251 1.8907524 1.7858088 1.9209397
## 5 3.6335486 3.5997041 3.5385613 3.5622239 3.6428058 3.5588146 3.6114502
##      AE1169   AE1170    AE1171    AE1172   AE1173    AE1174    AE1175
## 1 2.9333166 2.806446 2.8682297 2.7963220 2.882997 2.8517416 2.8476059
## 2 0.4410393 0.373576 0.4068394 0.3665689 0.409207 0.3943318 0.2684219
## 3 1.1728400 1.052466 1.1166752 1.0400485 1.123027 1.1024374 0.9364739
## 4 2.0208283 1.892123 1.9587703 1.8819145 1.968288 1.9379473 1.8364030
## 5 3.6191436 3.491511 3.5584802 3.4891624 3.575307 3.5434462 3.5860752
##      AE1176    AE1177    AE1178    AE1179    AE1180
## 1 2.7616280 2.8245047 2.7176708 2.7502298 2.8514261
## 2 0.2289570 0.2646176 0.2237919 0.2291777 0.2697421
## 3 0.8611594 0.9207595 0.8351714 0.8435144 0.9420726
## 4 1.7454409 1.8129884 1.7102739 1.7326465 1.8332838
## 5 3.5038501 3.5662006 3.4583779 3.4908875 3.5926027
```

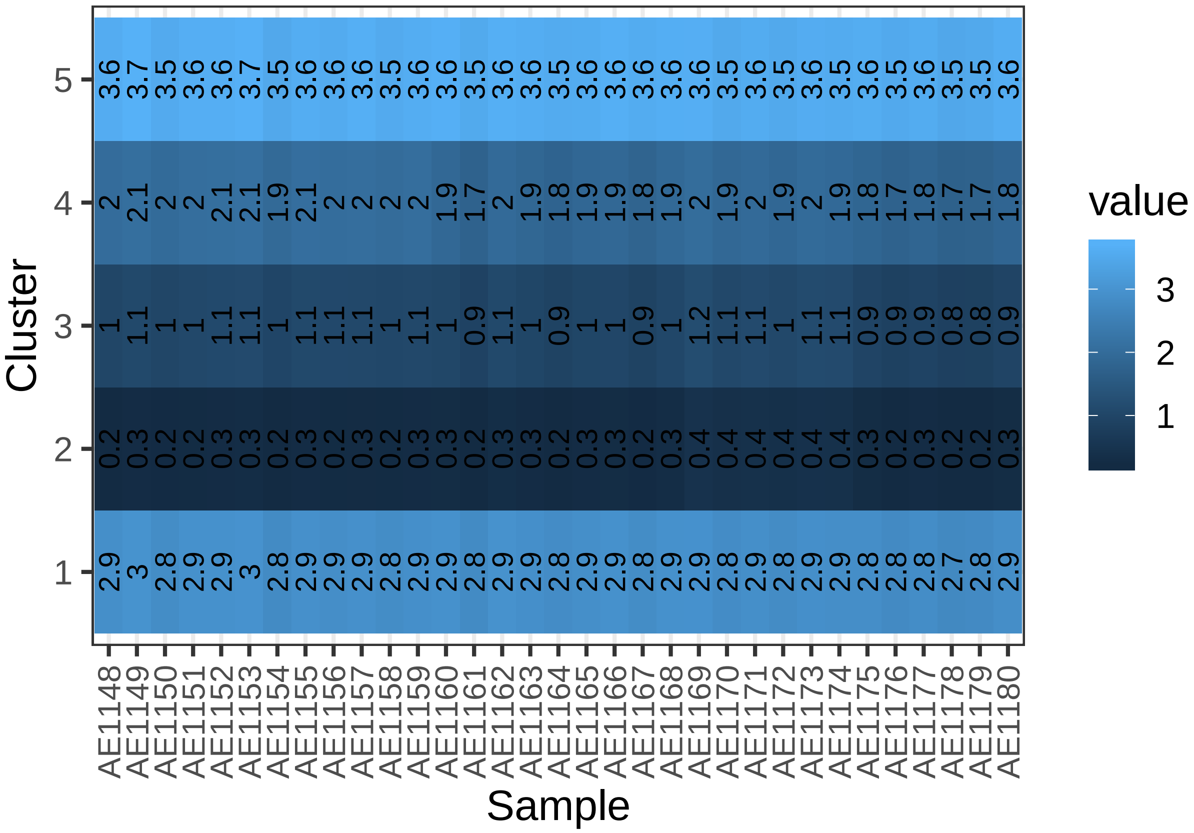
```
num_genes <- table(km$cluster)

num_genes
```

```
##
##    1    2    3    4    5
## 5789 9311 4629 3569 4401
```

```
plot(num_genes, type="l", main='num of genes in each cluster')
```



```r
plot_heatmap(km)
```



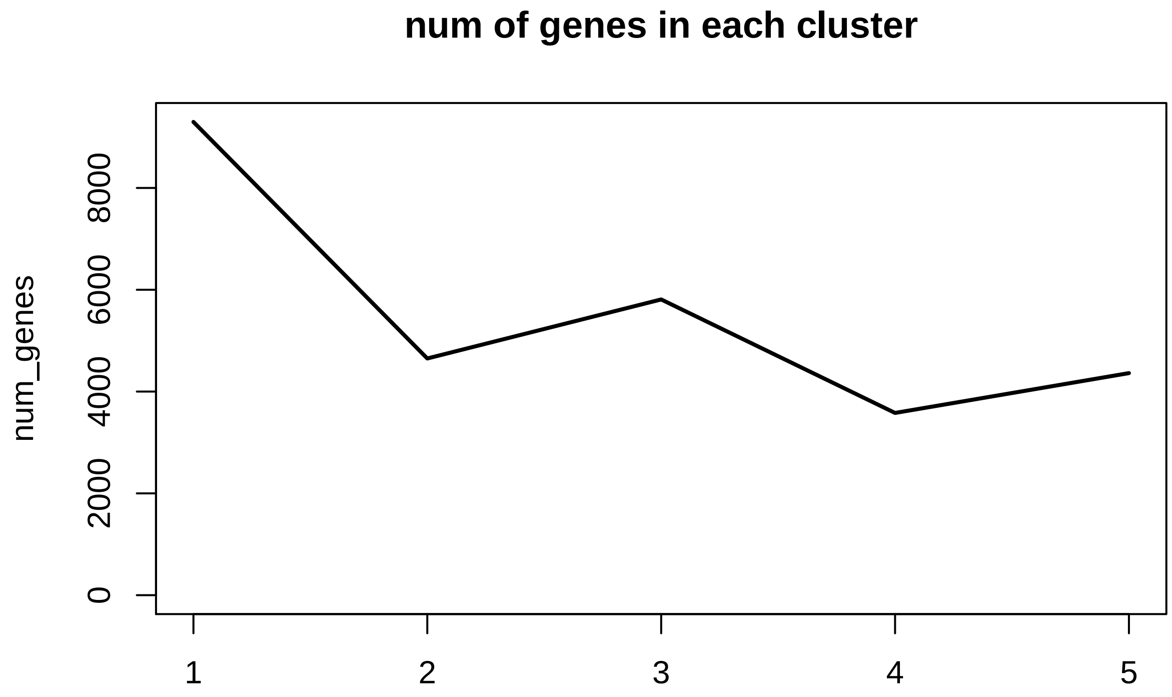
```r
# Apply standardization
rawcounts_scaled <- scale(rawcounts_log)
km <- kmeans(rawcounts_scaled,5)
km$centers
```

```
##       AE1148     AE1149     AE1150     AE1151     AE1152     AE1153
## 1 -1.0693269 -1.0688768 -1.0633402 -1.0722110 -1.0782449 -1.0693484
## 2 -0.5008578 -0.4873795 -0.4943596 -0.4901676 -0.4734303 -0.4761762
## 3  0.8964867  0.8893151  0.8828349  0.8935212  0.8863283  0.8799729
## 4  0.2528079  0.2451968  0.2458377  0.2538863  0.2649167  0.2567951
## 5  1.4114811  1.4119526  1.4156985  1.4092971  1.4048429  1.4039390
##       AE1154     AE1155     AE1156     AE1157     AE1158     AE1159
## 1 -1.0586070 -1.0739779 -1.0785432 -1.0690393 -1.0662110 -1.0741021
## 2 -0.4916615 -0.4791412 -0.4702332 -0.4830918 -0.4779076 -0.4795858
## 3  0.8760129  0.8860388  0.8821716  0.8804559  0.8744648  0.8826296
## 4  0.2447623  0.2625378  0.2683054  0.2498533  0.2583316  0.2608463
## 5  1.4127032  1.4041744  1.4048258  1.4157064  1.4051733  1.4108421
##       AE1160     AE1161     AE1162     AE1163     AE1164     AE1165
## 1 -1.0365985 -1.0180156 -1.0427133 -1.0321912 -1.0220277 -1.0340571
## 2 -0.4976051 -0.5213826 -0.4809970 -0.5061895 -0.5183713 -0.5024793
## 3  0.8973837  0.8894981  0.8979515  0.8978315  0.8936499  0.9007449
## 4  0.1468150  0.1198996  0.1673943  0.1452124  0.1274776  0.1580055
## 5  1.4240540  1.4423877  1.4017358  1.4245303  1.4359796  1.4101721
##       AE1166     AE1167     AE1168     AE1169     AE1170     AE1171
## 1 -1.0346146 -1.0215992 -1.0380963 -1.0333268 -1.0232746 -1.0306172
## 2 -0.5015744 -0.5215339 -0.4957272 -0.4671921 -0.4872764 -0.4752458
## 3  0.8979212  0.8933030  0.8994738  0.8715029  0.8722969  0.8709739
## 4  0.1481989  0.1284032  0.1652151  0.1754403  0.1611897  0.1696153
## 5  1.4222051  1.4381401  1.4073590  1.3956382  1.4062621  1.4039338
##       AE1172     AE1173     AE1174     AE1175     AE1176     AE1177
## 1 -1.0237152 -1.0311316 -1.0302534 -1.0126447 -1.0052797 -1.0095451
## 2 -0.4907044 -0.4749960 -0.4750045 -0.5235211 -0.5360008 -0.5268240
## 3  0.8720849  0.8718143  0.8711739  0.8850181  0.8846325  0.8835753
## 4  0.1597162  0.1691050  0.1655199  0.1384030  0.1238105  0.1326525
## 5  1.4123471  1.4040635  1.4059962  1.4240011  1.4340972  1.4275570
##       AE1178     AE1179     AE1180
## 1 -1.0017273 -1.0019488 -1.0131401
## 2 -0.5422155 -0.5441582 -0.5208928
## 3  0.8831780  0.8858309  0.8852047
## 4  0.1194956  0.1214050  0.1338500
## 5  1.4386298  1.4360721  1.4257434
```

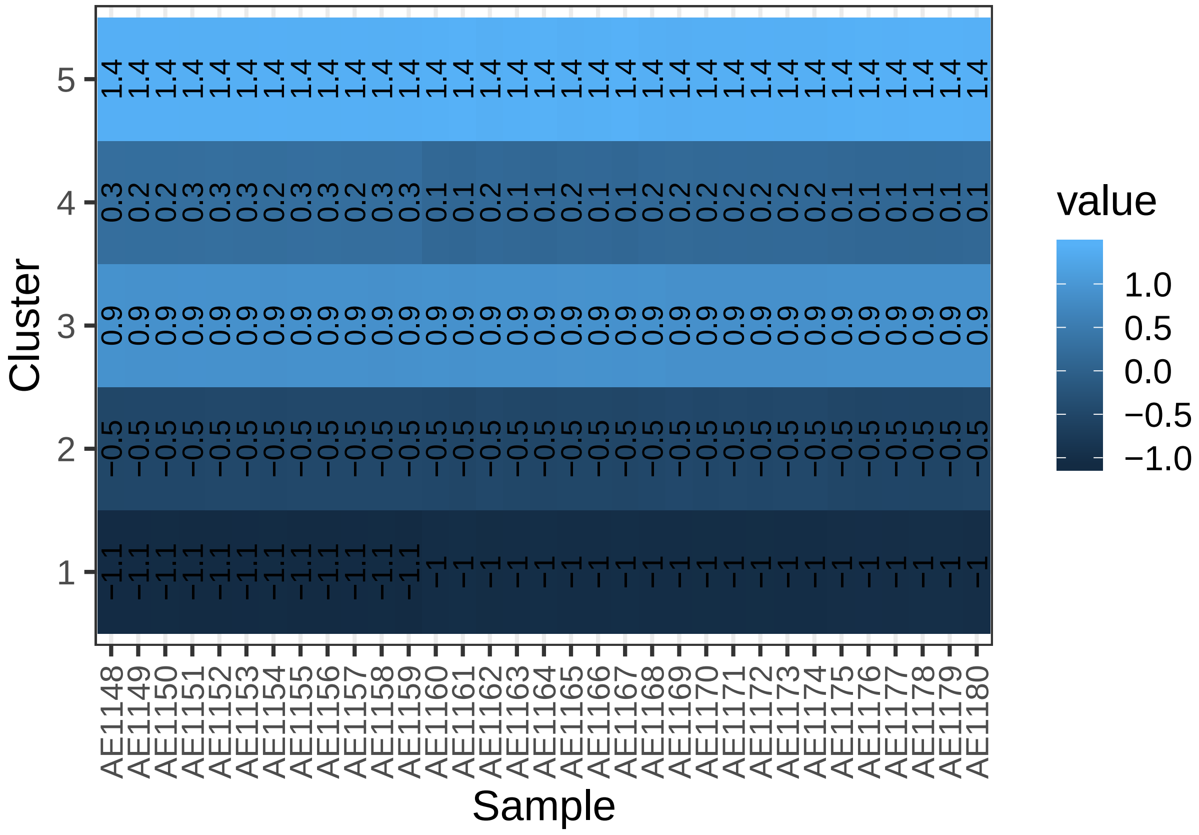
```
num_genes <- table(km$cluster)

num_genes
```

```
##
##    1    2    3    4    5
## 9297 4650 5810 3580 4362
```

```
plot(num_genes, type="l", main='num of genes in each cluster')
```



```
plot_heatmap(km)
```



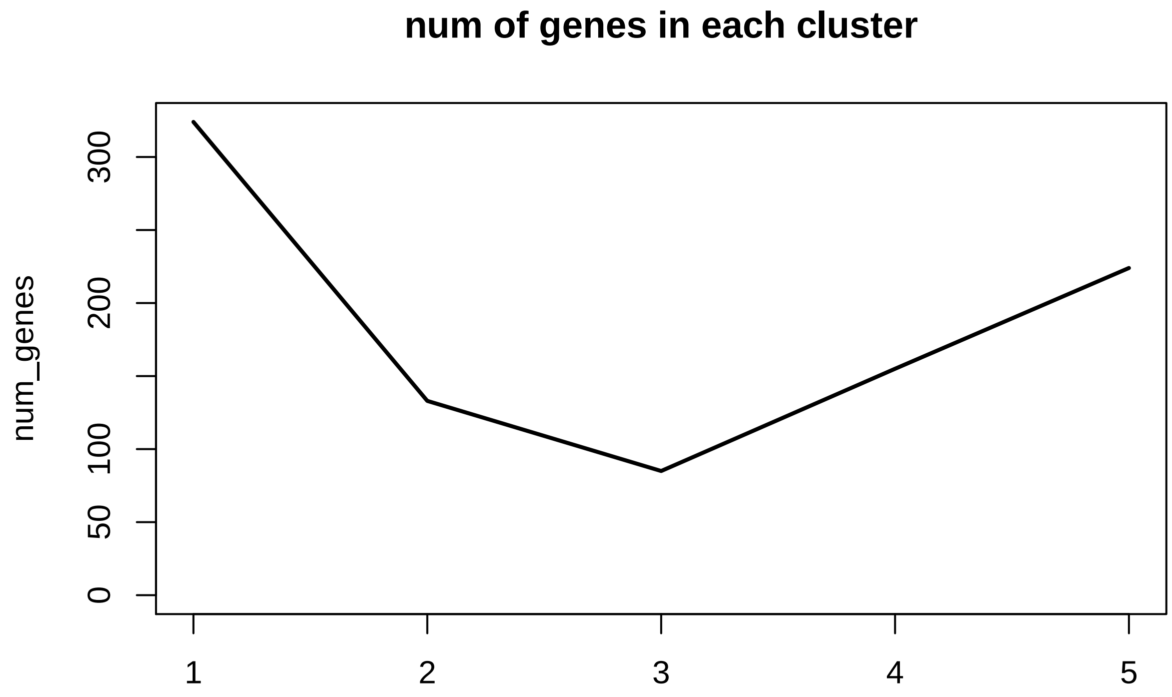
```
# load DE gene list
DE_res <- read_csv("data/sigresults.csv")
DE_genes <- DE_res$row

# example of filtering a data frame
dat_filtered = rawcounts_clean[DE_genes,]
rawcounts_log = log10(1+dat_filtered)
rawcounts_scaled <- scale(rawcounts_log)
km <- kmeans(rawcounts_scaled,5)
km$centers
```

```
##        AE1148      AE1149      AE1150      AE1151       AE1152      AE1153
## 1  0.42403366  0.40593559  0.41257406  0.41343195  0.378408740  0.32836900
## 2 -1.57036467 -1.53889857 -1.54205665 -1.55319030 -1.450632219 -1.39290277
## 3 -0.09426217 -0.05868337 -0.05744064 -0.07496552  0.005271765  0.08664749
## 4  1.32824453  1.34815737  1.35139359  1.34733238  1.382576790  1.40244856
## 5 -0.56425902 -0.58404432 -0.59448221 -0.57965220 -0.644722541 -0.65125022
##        AE1154      AE1155      AE1156      AE1157      AE1158      AE1159
## 1  0.33578948  0.34837334  0.35360157  0.36186704  0.34171469  0.36054754
## 2 -1.36066628 -1.39234203 -1.38216899 -1.45594594 -1.36793485 -1.41289991
## 3  0.06416777  0.04142446  0.03334408  0.03713265  0.01772979  0.04332771
## 4  1.38982555  1.39478217  1.39122149  1.38779412  1.38078743  1.38229875
## 5 -0.66385891 -0.65805263 -0.66612505 -0.63334140 -0.64423798 -0.65553963
##       AE1160     AE1161      AE1162     AE1163     AE1164     AE1165
## 1  0.4773720  0.4799396  0.47498824  0.4770977  0.4788714  0.4817270
## 2 -1.2482030 -1.2939910 -1.23358972 -1.2646579 -1.2663800 -1.2364346
## 3 -1.6214115 -1.5660180 -1.63189301 -1.6018460 -1.5875136 -1.6368159
## 4  1.1223981  1.1281802  1.09330082  1.1218241  1.1347879  1.0963944
## 5 -0.1107556 -0.1123038 -0.09187267 -0.1076159 -0.1235682 -0.1002032
##       AE1166     AE1167     AE1168     AE1169     AE1170     AE1171
## 1  0.4877514  0.4893930  0.4883743  0.4742062  0.4760763  0.4704565
## 2 -1.2599205 -1.2718339 -1.2509781 -1.3490956 -1.3586094 -1.3423383
## 3 -1.6219404 -1.6018767 -1.6234543 -1.3027711 -1.2760303 -1.2699420
## 4  1.1093457  1.1251012  1.0964178  1.1211957  1.1165703  1.1442426
## 5 -0.1095790 -0.1233956 -0.1062694 -0.1663521 -0.1703549 -0.1933458
##       AE1172     AE1173     AE1174     AE1175     AE1176     AE1177
## 1  0.4729096  0.4680305  0.4681295  0.4510429  0.4564274  0.4520655
## 2 -1.3416282 -1.3337278 -1.3261636 -1.5324000 -1.5477186 -1.5325070
## 3 -1.2875782 -1.2733156 -1.2828794 -0.9697346 -0.9441060 -0.9532551
## 4  1.1438410  1.1372523  1.1470611  1.1190033  1.1226214  1.1195577
## 5 -0.1903454 -0.1888319 -0.1966247 -0.1488695 -0.1597912 -0.1569221
##       AE1178     AE1179     AE1180
## 1  0.4545237  0.4518388  0.4539705
## 2 -1.5486074 -1.5390727 -1.5253211
## 3 -0.9526021 -0.9620322 -0.9742974
## 4  1.1219108  1.1278541  1.1172293
## 5 -0.1527941 -0.1551061 -0.1543482
```

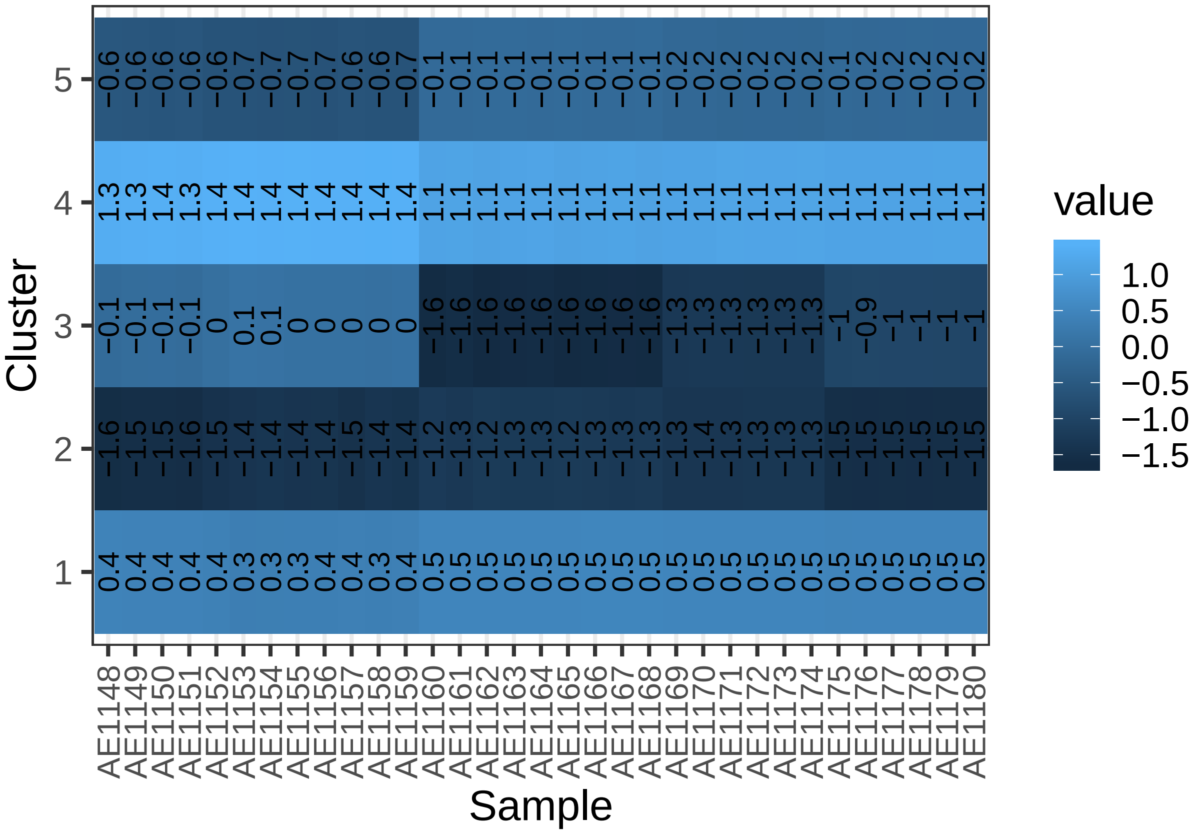
```r
num_genes <- table(km$cluster)

num_genes
```

```
##
##   1   2   3   4   5
## 324 133  85 155 224
```

```r
plot(num_genes, type="l", main='num of genes in each cluster')
```



```
plot_heatmap(km)
```



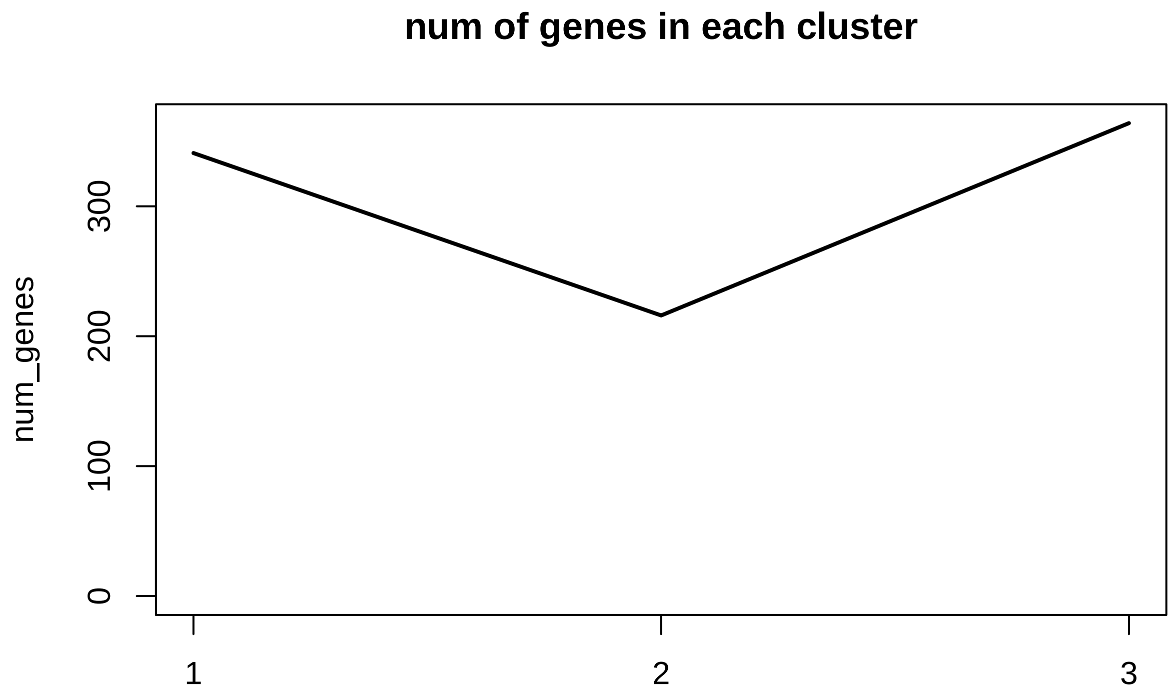
```r
# try again with less clusters
km <- kmeans(rawcounts_scaled,3)
km$centers
```

```
##       AE1148     AE1149     AE1150     AE1151     AE1152     AE1153
## 1  0.9421614  0.9443945  0.9513452  0.9479200  0.9544944  0.9367624
## 2 -1.0897526 -1.0578491 -1.0588992 -1.0754732 -0.9956799 -0.9289602
## 3 -0.2359628 -0.2569866 -0.2628749 -0.2498311 -0.3033400 -0.3263202
##       AE1154     AE1155     AE1156     AE1157     AE1158     AE1159
## 1  0.9349843  0.9437989  0.9441724  0.9456099  0.9302650  0.9452465
## 2 -0.9224395 -0.9499582 -0.9432554 -0.9797709 -0.9325391 -0.9582931
## 3 -0.3285240 -0.3204519 -0.3247792 -0.3044573 -0.3181097 -0.3168620
##        AE1160     AE1161      AE1162      AE1163      AE1164      AE1165
## 1  0.82571167  0.8338699  0.80900981  0.82507007  0.83521581  0.81472352
## 2 -1.36041448 -1.3633053 -1.35171193 -1.36449421 -1.35355507 -1.35932245
## 3  0.03374135  0.0278140  0.04422371  0.03676334  0.02076732  0.04338717
##        AE1166      AE1167      AE1168      AE1169      AE1170      AE1171
## 1  0.82544492  0.83656652  0.82002231  0.84714767  0.84572955  0.85731922
## 2 -1.37172915 -1.36239212 -1.36760089 -1.31968529 -1.31536911 -1.30887236
## 3  0.04070543  0.02474592  0.04333567 -0.01050916 -0.01174189 -0.02645446
##        AE1172      AE1173      AE1174      AE1175      AE1176      AE1177
## 1  0.85917638  0.85284727  0.85763275  0.83012936  0.83458808  0.83105863
## 2 -1.31136582 -1.30026943 -1.29824918 -1.35161032 -1.35404301 -1.34482053
## 3 -0.02671463 -0.02737011 -0.03305205  0.02437834  0.02164494  0.01947869
##        AE1178      AE1179     AE1180
## 1  0.83425955  0.83615931  0.8312262
## 2 -1.35443662 -1.35491449 -1.3511697
## 3  0.02218627  0.02069012  0.0230893
```

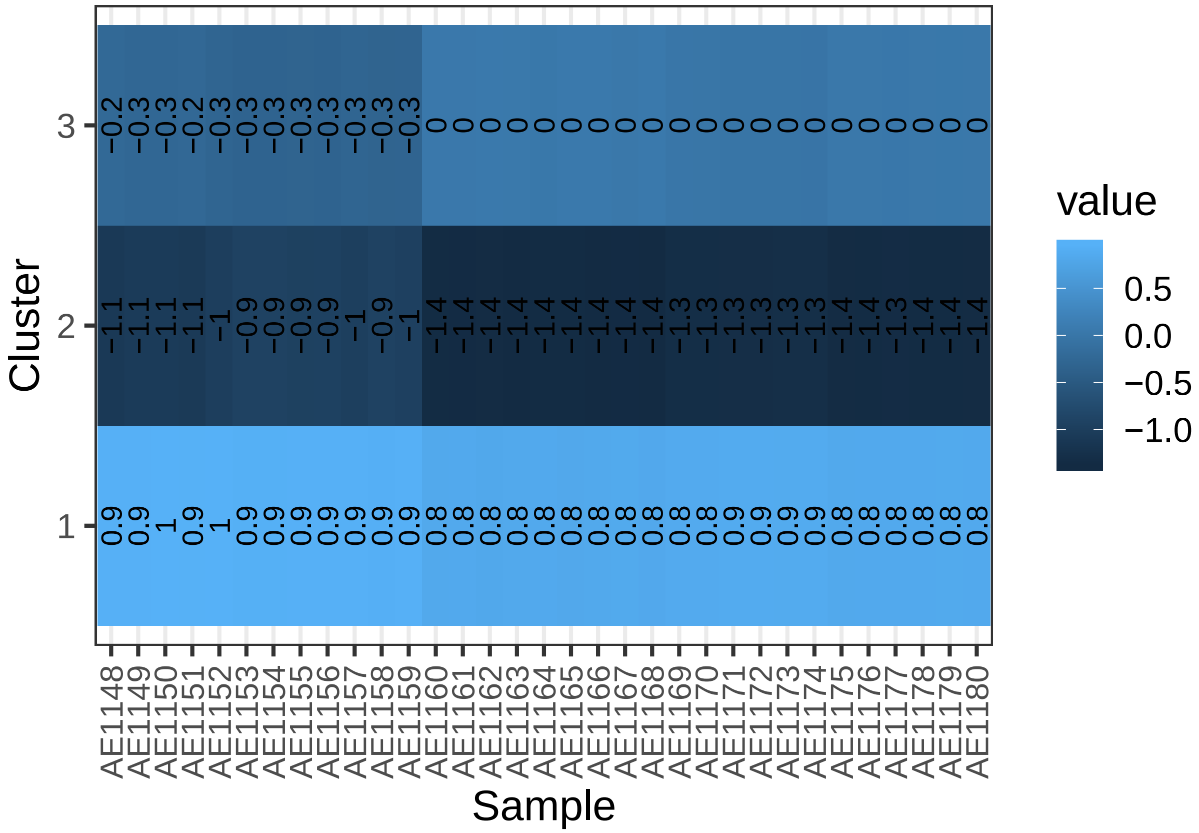
```
num_genes <- table(km$cluster)

num_genes
```

```
##
##   1   2   3
## 341 216 364
```

```
plot(num_genes, type="l", main='num of genes in each cluster')
```



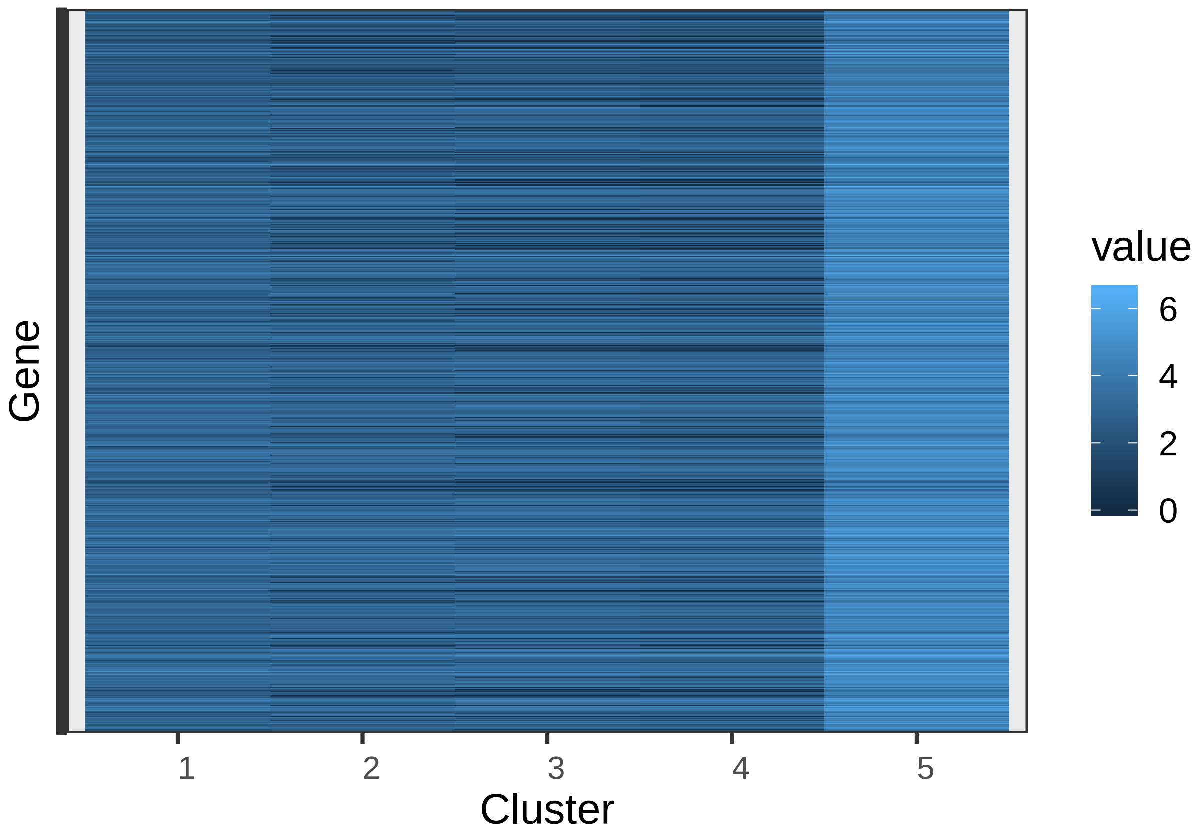
```
plot_heatmap(km)
```



```r
# a plotting function
plot_heatmap <- function(km) {
  centers <- km$centers %>%
      tbl_df() %>%
      rownames_to_column('Cluster') %>%
      gather(Gene, value, -Cluster) %>%
      mutate(
          Cluster = factor(Cluster),
          Gene = factor(Gene)
      )
   ggplot(centers,aes(Cluster,Gene)) + geom_tile(aes(fill=value)) + theme(axis.text.x=element_text(angl
}

# cluster the samples
rawcounts_clean <- rawcounts_RN[ rawcounts_RN$sum >= 10, ]
rawcounts_log = log10(1+rawcounts_clean)
DE_res <- read_csv("data/sigresults.csv")
DE_genes <- DE_res$row
dat_filtered = rawcounts_log[DE_genes,]
km <- kmeans(t(dat_filtered),5)
#km$centers
table(km$cluster)
```

```
##
##  1  2  3  4  5
## 12 12  6  3  1
```

```
plot_heatmap(km)
```



## Record sessionInfo()

The sessionInfo() prints version information about R and any attached packages. It's a good practice to always run this command at the end of your R session and record it for the sake of reproducibility in the future.

```
sessionInfo()
```

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_Israel.1252  LC_CTYPE=English_Israel.1252
## [3] LC_MONETARY=English_Israel.1252 LC_NUMERIC=C
## [5] LC_TIME=English_Israel.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
## [1] bindrcpp_0.2.2 tidyr_0.8.2    tibble_1.4.2   dplyr_0.7.8
## [5] readr_1.2.1    ggplot2_3.1.0  knitr_1.20
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.0       pillar_1.3.0     compiler_3.5.1   plyr_1.8.4
##  [5] bindr_0.1.1      tools_3.5.1      digest_0.6.18    evaluate_0.12
##  [9] gtable_0.2.0     pkgconfig_2.0.2  rlang_0.3.0.1    cli_1.0.1
## [13] yaml_2.2.0       withr_2.1.2      stringr_1.3.1    hms_0.4.2
## [17] grid_3.5.1       tidyselect_0.2.5 glue_1.3.0       R6_2.3.0
## [21] fansi_0.4.0      rmarkdown_1.11   purrr_0.2.5      magrittr_1.5
## [25] scales_1.0.0     codetools_0.2-15 htmltools_0.3.6  assertthat_0.2.0
## [29] colorspace_1.3-2 labeling_0.3     utf8_1.1.4       stringi_1.2.4
## [33] lazyeval_0.2.1   munsell_0.5.0    crayon_1.3.4
```