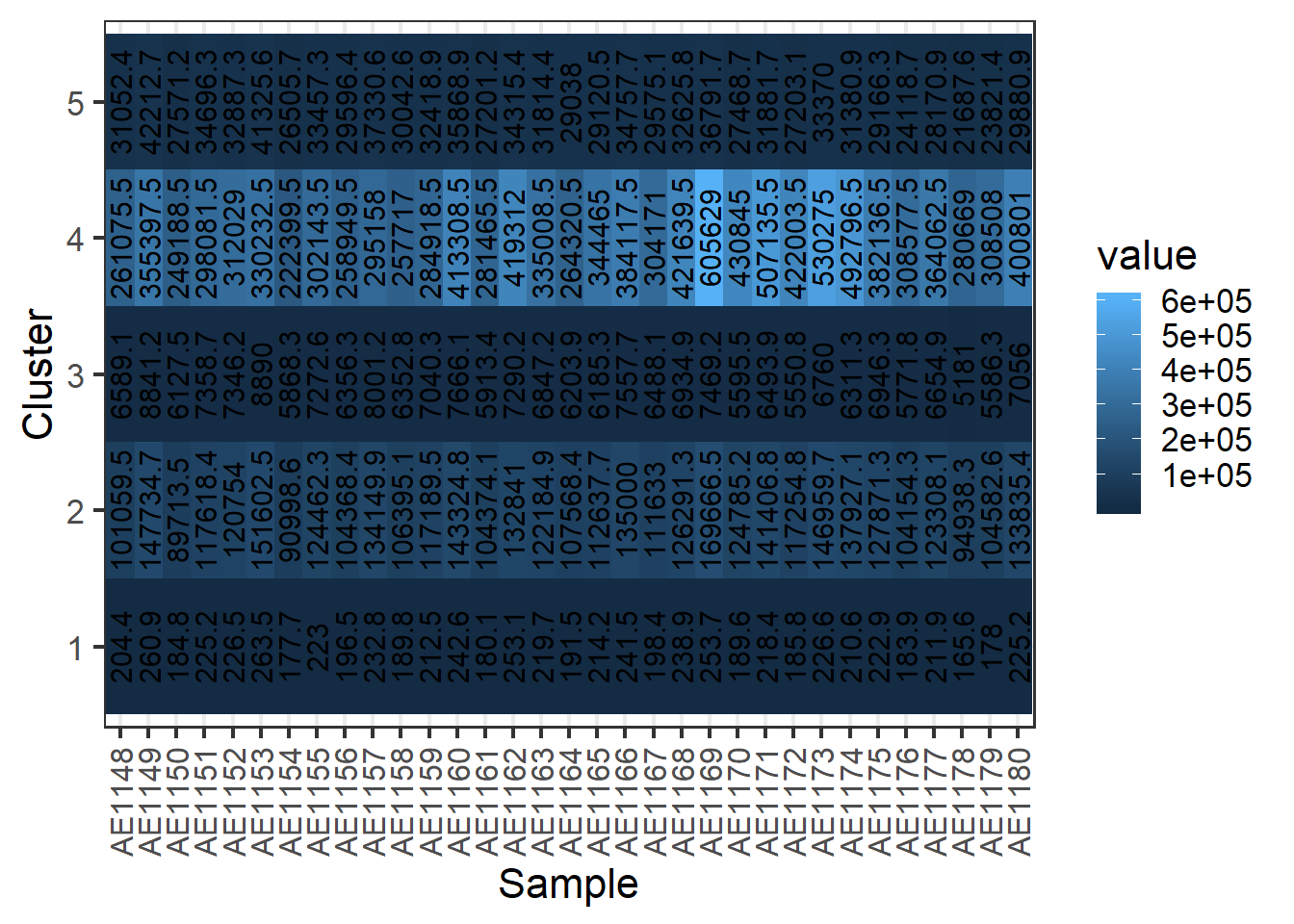
HW3 – Bioinformatics – 236523

Anna Romanov 321340580 [annarom@campus.technion.ac.il](mailto:annarom@campus.technion.ac.il)

Maxim Kolchinsky 320983216 [kolchinsky@campus.technion.ac.il](mailto:kolchinsky@campus.technion.ac.il)

# Question 1

2. c. ii.

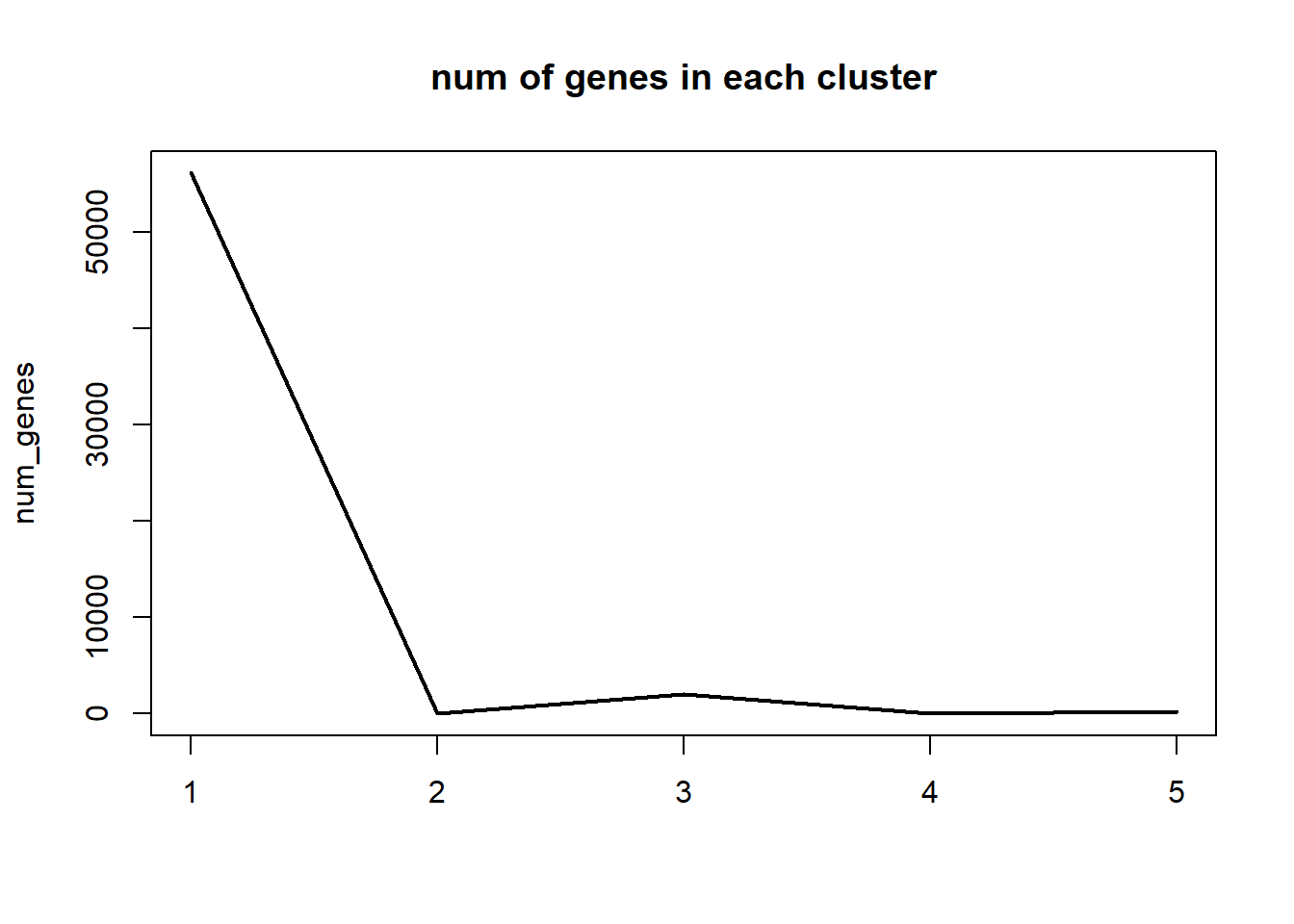


iii.

## 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.

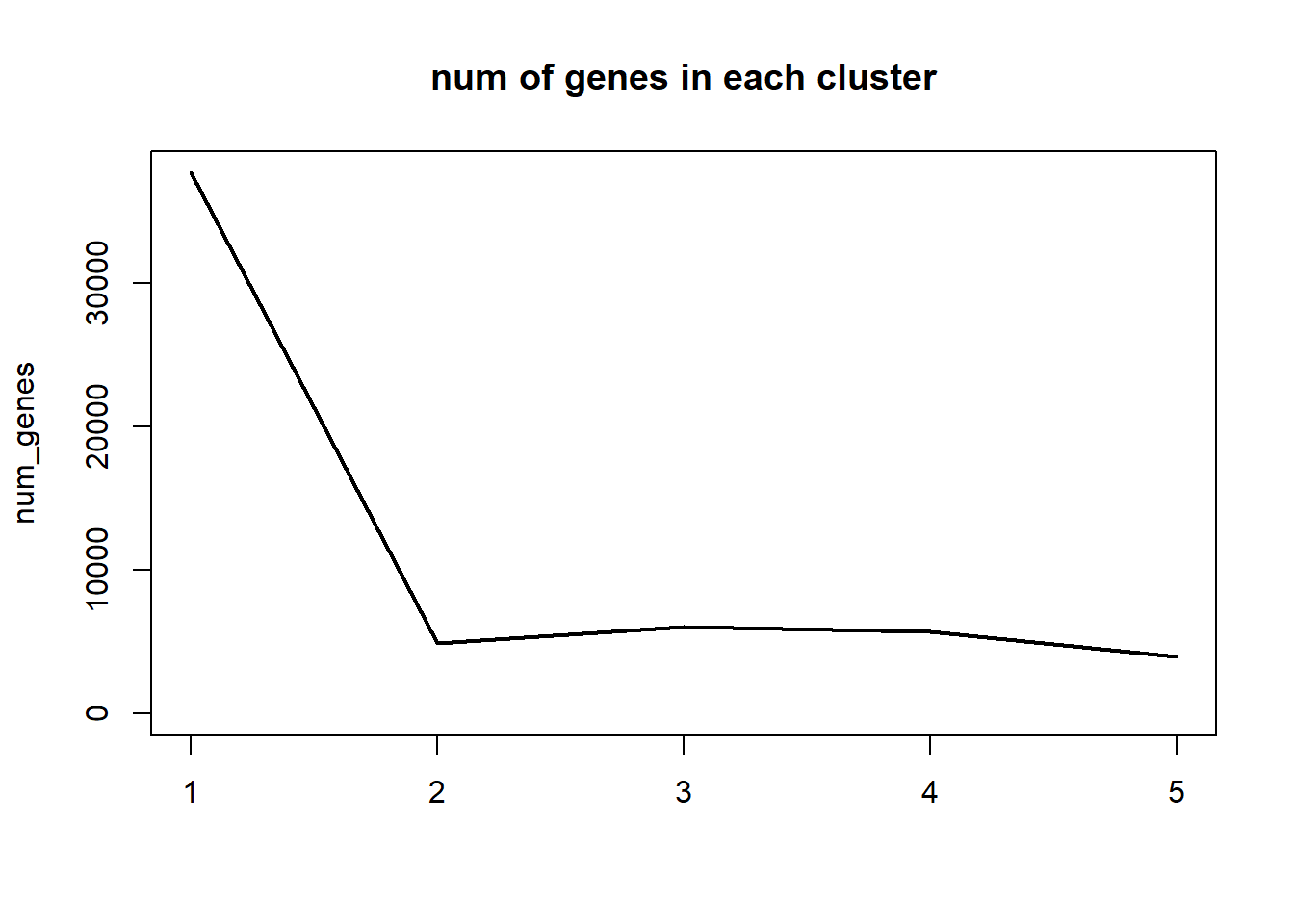
Another problem is that for clusters 1 and 2, the range of values that belong to the same cluster is big, therefore in the heatmap it looks like cluster 1 is not homogeneous. There are two possibilities: either the big difference in values (for example there is a value of 605,629 together with 222,399 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).

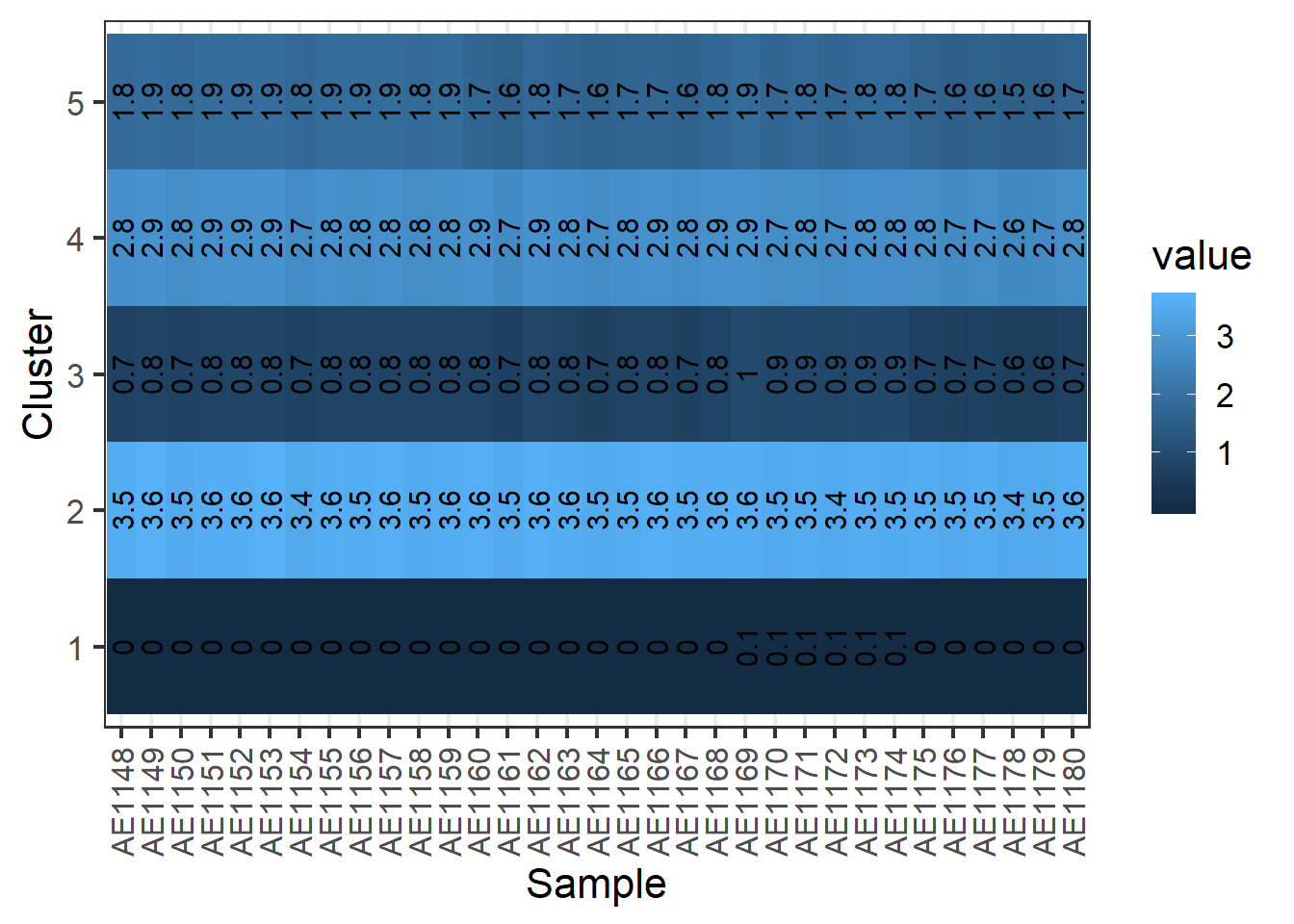
A possible reason for these problems could be that the expression values are within a very large range, being more dense in the smaller values (we see that in cluster 5, the maximum difference is around ~100, while in cluster 1 it is around ~300,000). This in turn is caused by large differences in counts 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).

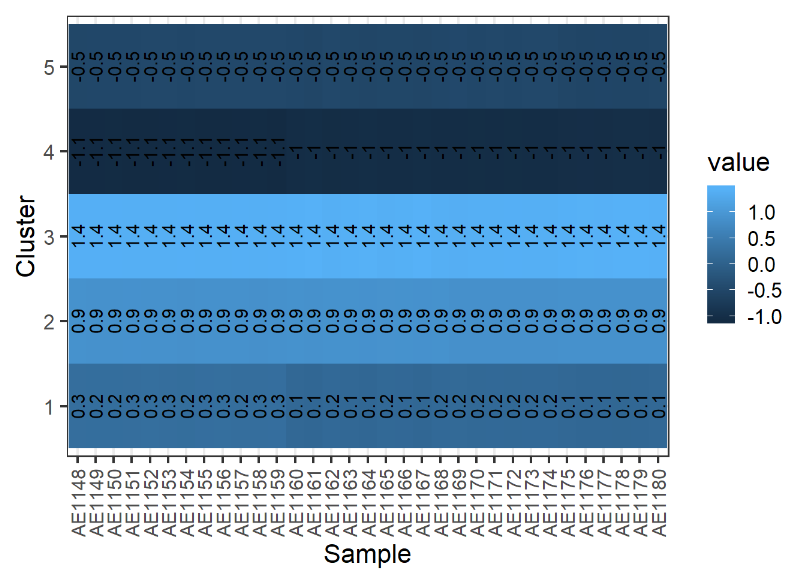
vi. We noticed that many of the rows in the dataset have values of zero in all or nearly all columns, and therefore such rows are not valuable for expression analysis of genes – our goal is to find genes which have different expression levels for different treatments, however if a gene has zero expression 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

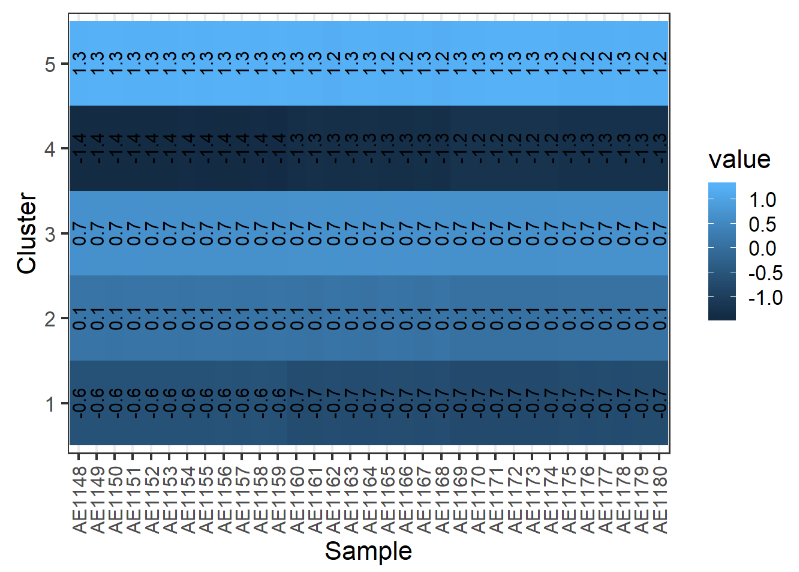


As we can see, the clusters are more homogenous now and also different from each other.

x. The results after filtering the data (and applying all the steps as:

1 2 3 4 5

3055 3267 5158 3895 2582

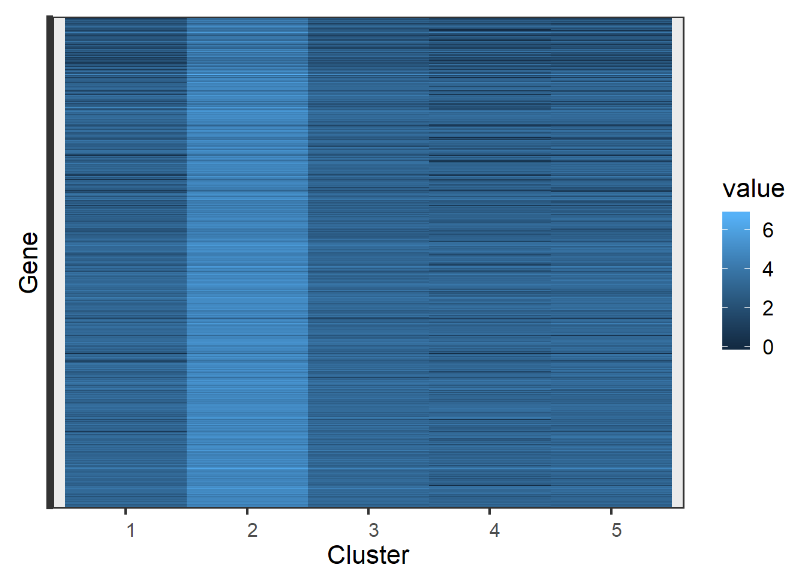


The results are slightly different, we can see that filtering out non-differentially expressed genes evened out the clusters’ sizes.

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

6 1 9 6 12



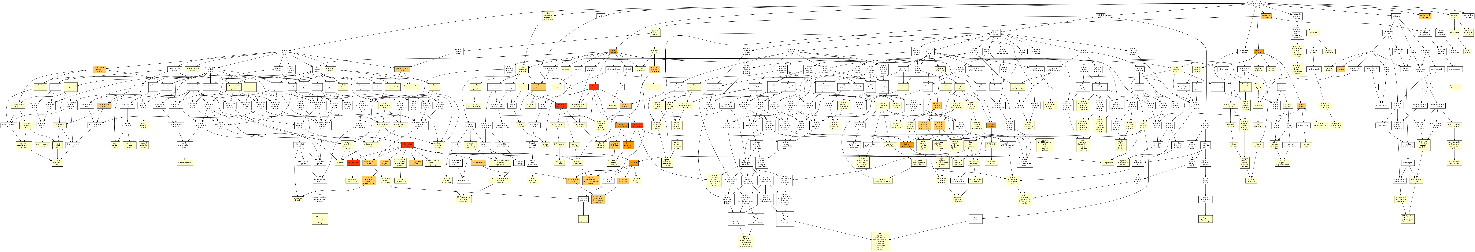
iii. Clustering the genes, we got good results with the 5 clusters being homogenous and well-separated from each other (meaning that in the same cluster, genes have similar expressions levels, and different from those of genes in other clusters). From the clustering results, we can learn which genes 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).

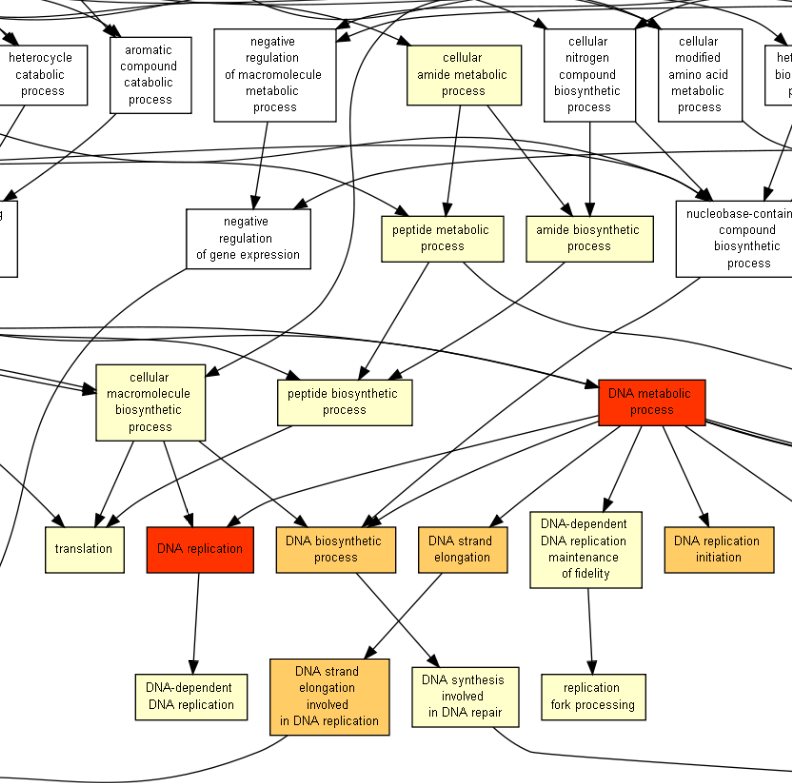
# Question 2

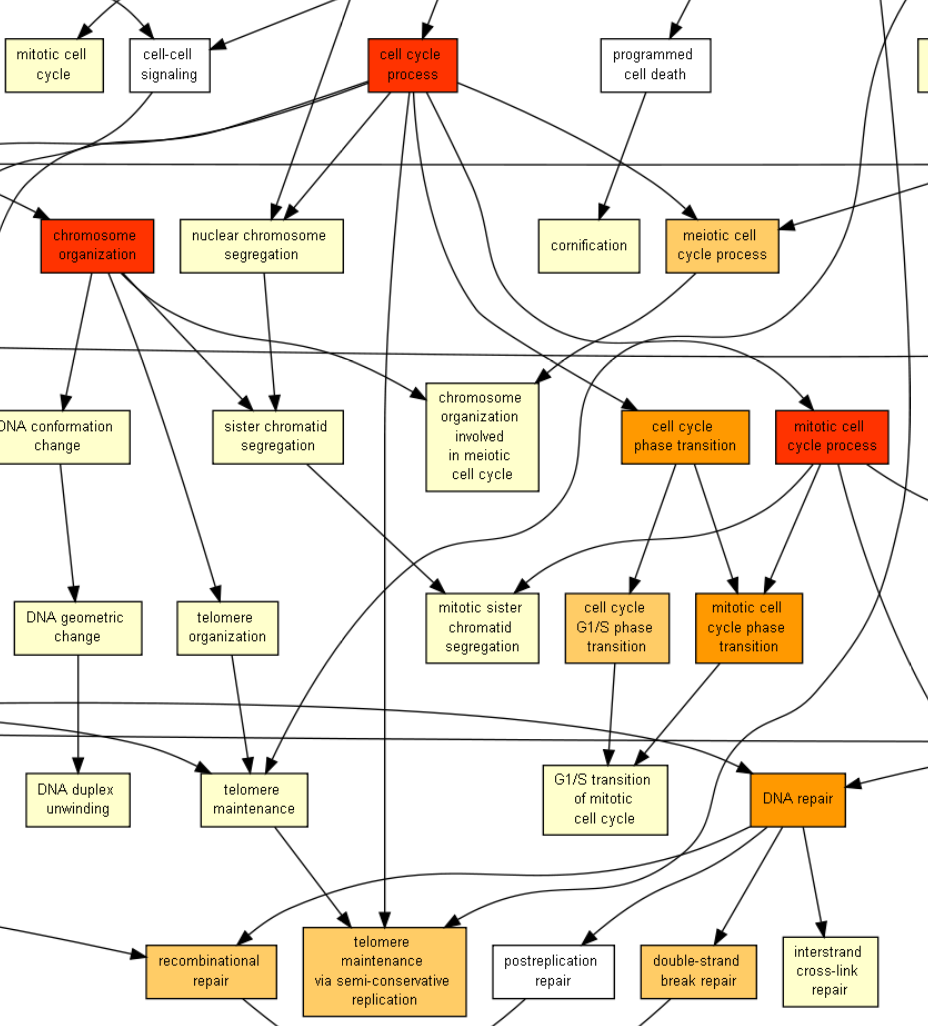
1. C. We got a long list of GO terms that were found by Gorilla, the most significant ones are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **GO term** | **Description** | [**P-value**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/gthwrzkj/GOResultsPROCESS.html#p_value_info) | [**FDR q-value**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/gthwrzkj/GOResultsPROCESS.html#fdr_info) | [**Enrichment (N, B, n, b)**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/gthwrzkj/GOResultsPROCESS.html#enrich_info) |
| [GO:0022402](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0022402&view=details) | cell cycle process | 8.7E-17 | 1.22E-12 | 2.01 (11732,764,1261,165) |
| [GO:0006260](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0006260&view=details) | DNA replication | 2.6E-14 | 1.82E-10 | 4.04 (11732,136,940,44) |
| [GO:1903047](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:1903047&view=details) | mitotic cell cycle process | 2.91E-14 | 1.36E-10 | 2.18 (11732,504,1261,118) |
| [GO:0051276](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0051276&view=details) | chromosome organization | 6.96E-13 | 2.44E-9 | 2.50 (11732,294,1163,73) |
| [GO:0006259](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0006259&view=details) | DNA metabolic process | 5.62E-12 | 1.57E-8 | 2.15 (11732,641,819,96) |
| [GO:0009987](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0009987&view=details) | cellular process | 1.59E-9 | 3.72E-6 | 1.09 (11732,9073,1237,1044) |
| [GO:0007093](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0007093&view=details) | mitotic cell cycle checkpoint | 6.46E-9 | 1.29E-5 | 3.38 (11732,89,1208,31) |
| [GO:0000075](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0000075&view=details) | cell cycle checkpoint | 1.49E-8 | 2.61E-5 | 2.99 (11732,117,1208,36) |
| [GO:0050896](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0050896&view=details) | response to stimulus | 2.04E-8 | 3.18E-5 | 1.35 (11732,3063,833,294) |
| [GO:0006281](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0006281&view=details) | DNA repair | 2.56E-8 | 3.58E-5 | 2.19 (11732,403,837,63) |
| [GO:0044772](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0044772&view=details) | mitotic cell cycle phase transition | 2.95E-8 | 3.75E-5 | 2.37 (11732,203,1221,50) |
| [GO:0044770](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0044770&view=details) | cell cycle phase transition | 3.69E-8 | 4.31E-5 | 2.33 (11732,210,1221,51) |
| [GO:1903046](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:1903046&view=details) | meiotic cell cycle process | 7.38E-7 | 7.95E-4 | 2.90 (11732,104,1167,30) |
| [GO:0031023](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0031023&view=details) | microtubule organizing center organization | 8.31E-7 | 8.31E-4 | 3.63 (11732,57,1192,21) |
| [GO:0060249](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0060249&view=details) | anatomical structure homeostasis | 8.74E-7 | 8.15E-4 | 2.27 (11732,191,1188,44) |
| [GO:0006271](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0006271&view=details) | DNA strand elongation involved in DNA replication | 9.31E-7 | 8.15E-4 | 9.96 (11732,12,785,8) |



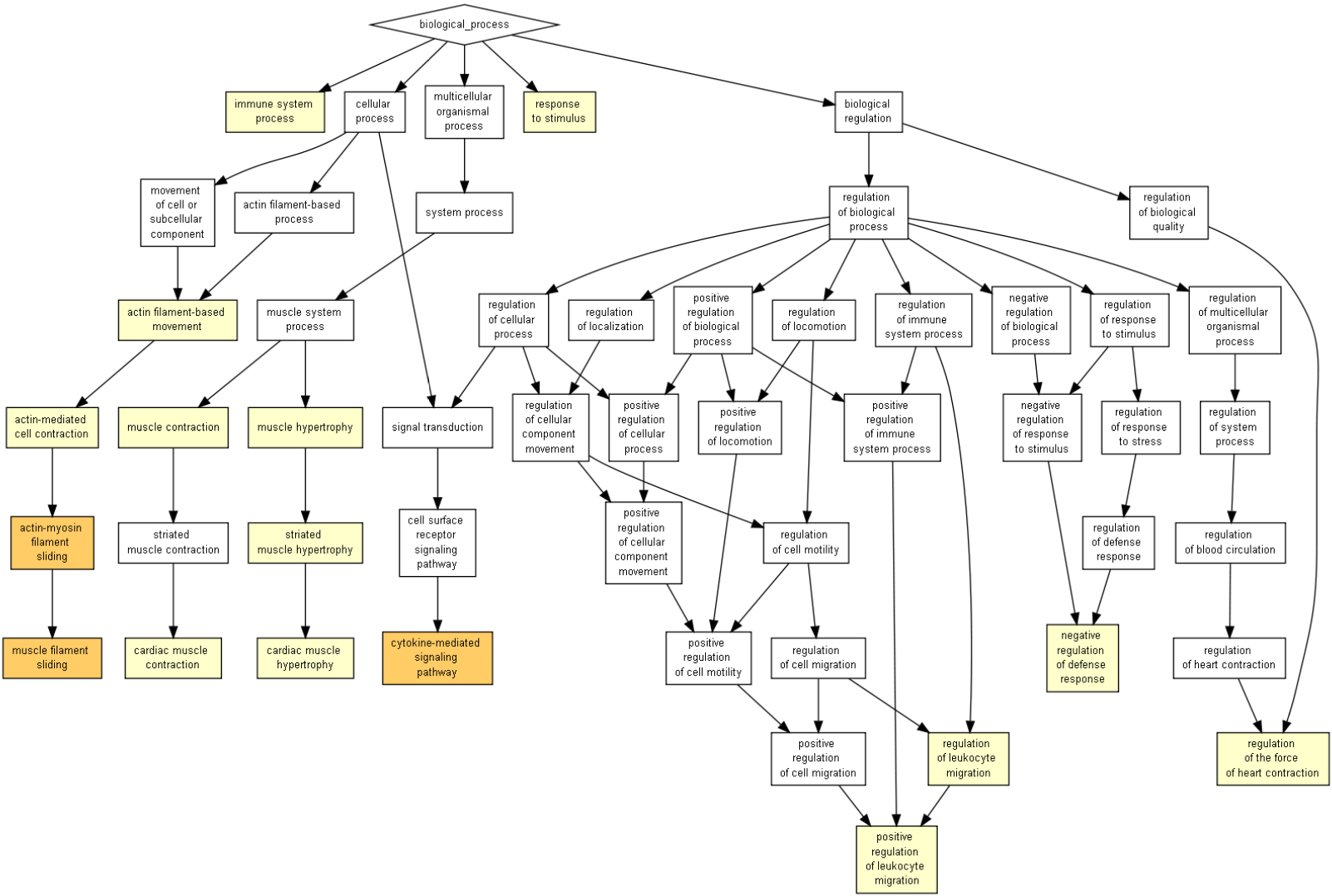
Relevant parts of the diagram are:





 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.

1. 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:



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **GO term** | **Description** | [**P-value**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/8b037647/GOResultsPROCESS.html#p_value_info) | [**FDR q-value**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/8b037647/GOResultsPROCESS.html#fdr_info) | [**Enrichment (N, B, n, b)**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/8b037647/GOResultsPROCESS.html#enrich_info) |
| [GO:0019221](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0019221&view=details) | cytokine-mediated signaling pathway | 4.56E-7 | 6.95E-3 | 13.15 (17785,526,18,7) |
| [GO:0030049](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0030049&view=details) | muscle filament sliding | 5.11E-6 | 3.89E-2 | 87.18 (17785,34,18,3) |
| [GO:0033275](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0033275&view=details) | actin-myosin filament sliding | 5.11E-6 | 2.6E-2 | 87.18 (17785,34,18,3) |
| [GO:0070252](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0070252&view=details) | actin-mediated cell contraction | 1.12E-5 | 4.28E-2 | 67.37 (17785,44,18,3) |
| [GO:0030048](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0030048&view=details) | actin filament-based movement | 3.02E-5 | 9.21E-2 | 48.59 (17785,61,18,3) |
| [GO:0002687](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0002687&view=details) | positive regulation of leukocyte migration | 1.62E-4 | 4.11E-1 | 27.70 (17785,107,18,3) |
| [GO:0002376](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0002376&view=details) | immune system process | 1.67E-4 | 3.64E-1 | 4.46 (17785,1772,18,8) |
| [GO:0003300](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0003300&view=details) | cardiac muscle hypertrophy | 1.82E-4 | 3.47E-1 | 98.81 (17785,20,18,2) |
| [GO:0014897](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0014897&view=details) | striated muscle hypertrophy | 2.01E-4 | 3.4E-1 | 94.10 (17785,21,18,2) |
| [GO:0014896](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0014896&view=details) | muscle hypertrophy | 2.21E-4 | 3.37E-1 | 89.82 (17785,22,18,2) |
| [GO:0002026](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0002026&view=details) | regulation of the force of heart contraction | 2.64E-4 | 3.65E-1 | 82.34 (17785,24,18,2) |
| [GO:0060048](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0060048&view=details) | cardiac muscle contraction | 4.14E-4 | 5.26E-1 | 65.87 (17785,30,18,2) |
| [GO:0050896](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0050896&view=details) | response to stimulus | 4.31E-4 | 5.06E-1 | 2.51 (17785,4731,18,12) |
| [GO:0002685](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0002685&view=details) | regulation of leukocyte migration | 5.38E-4 | 5.86E-1 | 18.41 (17785,161,18,3) |
| [GO:0006936](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0006936&view=details) | muscle contraction | 8.18E-4 | 8.32E-1 | 15.94 (17785,186,18,3) |
| [GO:0031348](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0031348&view=details) | 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

1. 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:



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

1. 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 |

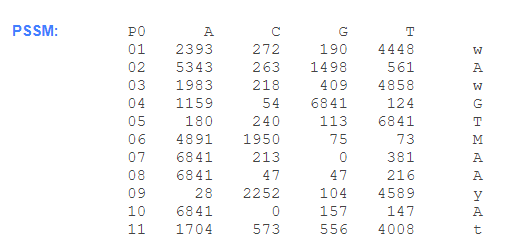
What can you say about the motif?

The motif is FOXBD1 taken from the Homo Sapiens Database.

The length of the motif is 11.



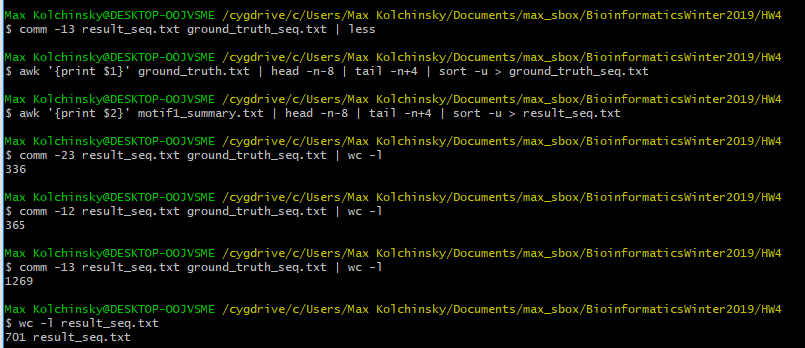
The PSSM of the motif is:

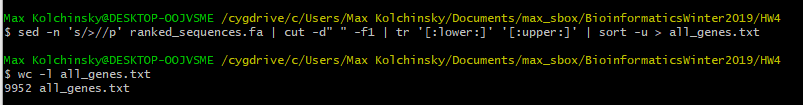


What possible experiment generated the ranked list of sequences?

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.

1. The following commands gave me the results:





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

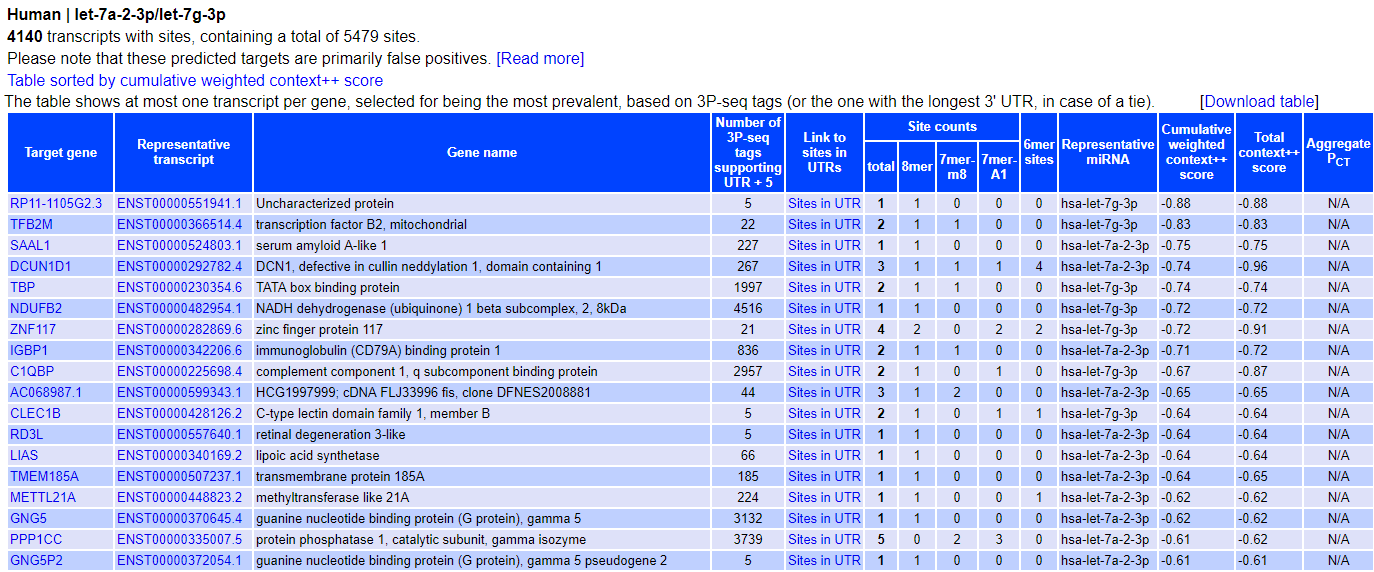
From which I can conclude that my confusion matrix will be like so:

|  |  |  |
| --- | --- | --- |
|  | Predicted:  No | Predicted:  Yes |
| Actual:  No | 8636 | 336 |
| Actual:  Yes | 1269 | 365 |

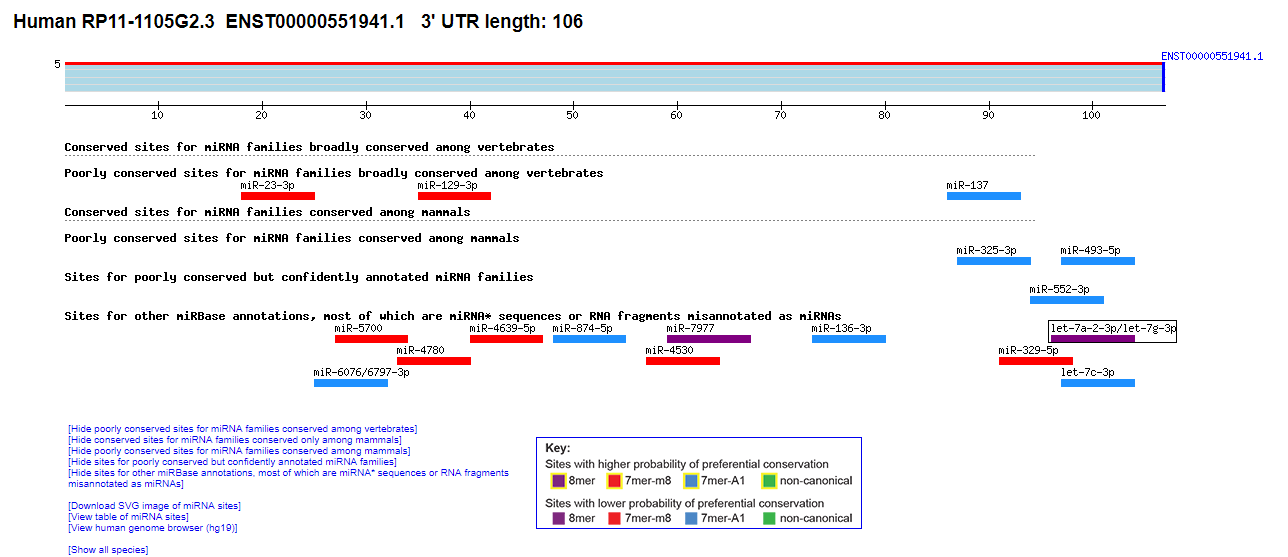
1. Calculation:
2. If the tool is very complex it is also very slow. Since we already have 96.2% specifity if we value specifity a lot more than sensitivity and we also value fast runtime we will prefer to run the motif search to get almost 100% results a lot faster.

# Question 4

1. Screenshot of the results:

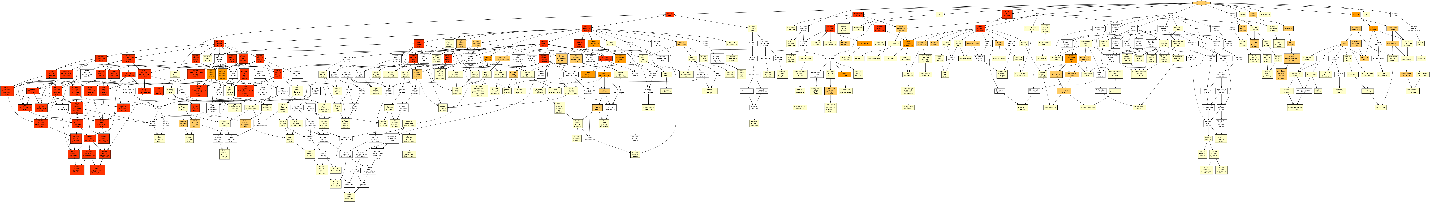


1. Predicted target sites on the UTR of target gene [RP11-1105G2.3](http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000258365.1)

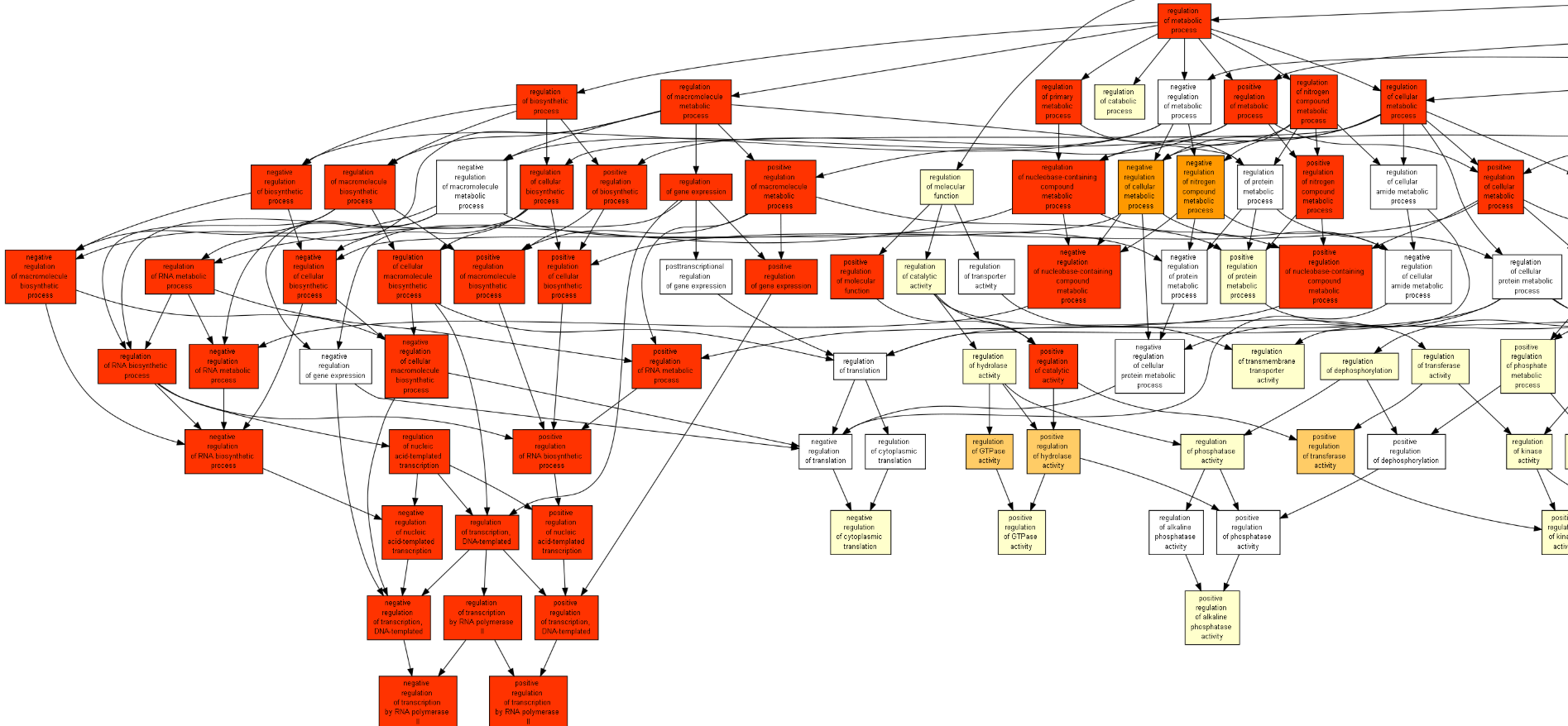


1. I used background\_list.txt from Q2 as the background of the search and the table from section 1 as the target list.

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:

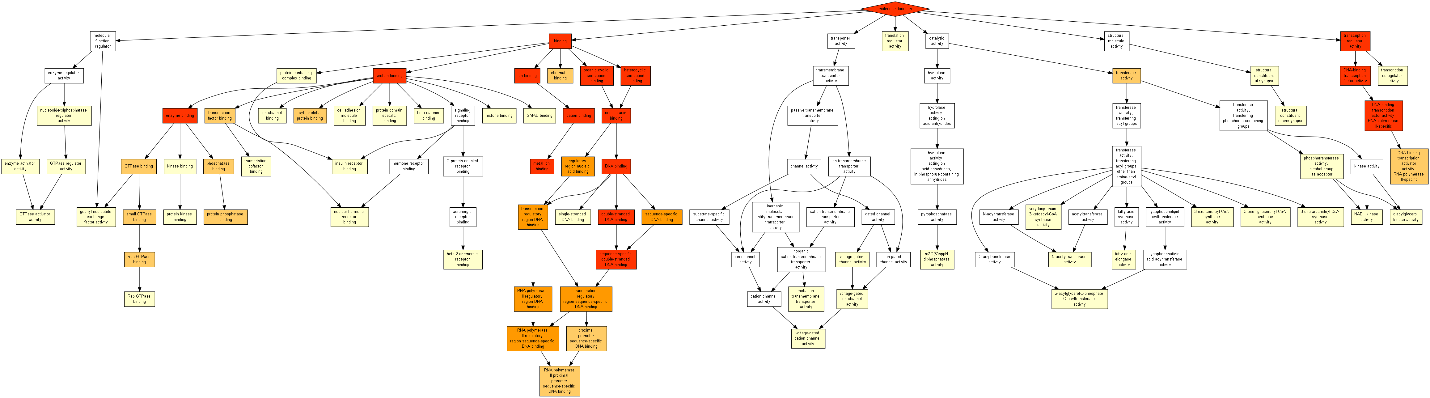


As we can see in the table below there are other differentially expressed genes with higher p-value as can be seen here:

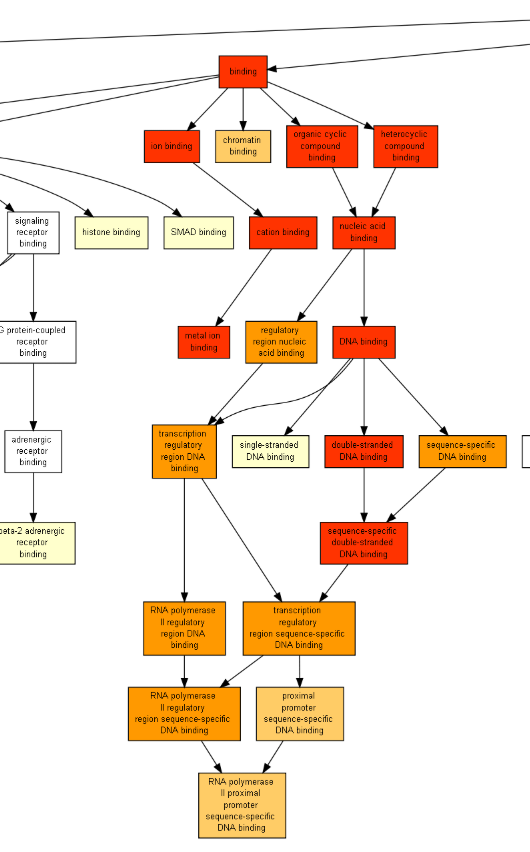
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **GO term** | **Description** | [**P-value**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/h7xvnja1/GOResultsPROCESS.html#p_value_info) | [**FDR q-value**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/h7xvnja1/GOResultsPROCESS.html#fdr_info) | [**Enrichment (N, B, n, b)**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/h7xvnja1/GOResultsPROCESS.html#enrich_info) | [**Genes**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/h7xvnja1/GOResultsPROCESS.html#genes_info) |
| [GO:0051252](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0051252&view=details) | regulation of RNA metabolic process | 2.42E-27 | 3.72E-23 | 1.32 (18016,3456,3966,1003) | [[+] Show genes](javascript:toggle('elements_GO:0051252')) |
| [GO:0019219](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0019219&view=details) | regulation of nucleobase-containing compound metabolic process | 5.48E-27 | 4.21E-23 | 1.30 (18016,3709,3966,1063) | [[+] Show genes](javascript:toggle('elements_GO:0019219')) |
| [GO:2000112](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:2000112&view=details) | regulation of cellular macromolecule biosynthetic process | 2.51E-25 | 1.29E-21 | 1.30 (18016,3612,3966,1031) | [[+] Show genes](javascript:toggle('elements_GO:2000112')) |
| [GO:1903506](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:1903506&view=details) | regulation of nucleic acid-templated transcription | 1.23E-24 | 4.72E-21 | 1.31 (18016,3221,3966,932) | [[+] Show genes](javascript:toggle('elements_GO:1903506')) |
| [GO:0010556](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0010556&view=details) | regulation of macromolecule biosynthetic process | 1.28E-24 | 3.93E-21 | 1.29 (18016,3722,3966,1054) | [[+] Show genes](javascript:toggle('elements_GO:0010556')) |
| [GO:2001141](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:2001141&view=details) | regulation of RNA biosynthetic process | 1.36E-24 | 3.49E-21 | 1.31 (18016,3226,3966,933) | [[+] Show genes](javascript:toggle('elements_GO:2001141')) |
| [GO:0006355](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0006355&view=details) | regulation of transcription, DNA-templated | 1.52E-24 | 3.34E-21 | 1.32 (18016,3170,3966,919) | [[+] Show genes](javascript:toggle('elements_GO:0006355')) |
| [GO:0031326](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0031326&view=details) | regulation of cellular biosynthetic process | 1.25E-23 | 2.4E-20 | 1.27 (18016,3860,3966,1082) | [[+] Show genes](javascript:toggle('elements_GO:0031326')) |
| [GO:0009889](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0009889&view=details) | regulation of biosynthetic process | 2.43E-23 | 4.15E-20 | 1.27 (18016,3929,3966,1097) | [[+] Show genes](javascript:toggle('elements_GO:0009889')) |
| [GO:0006357](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0006357&view=details) | regulation of transcription by RNA polymerase II | 8.87E-22 | 1.36E-18 | 1.35 (18016,2475,3966,733) | [[+] Show genes](javascript:toggle('elements_GO:0006357')) |
| [GO:0031323](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0031323&view=details) | regulation of cellular metabolic process | 7.93E-21 | 1.11E-17 | 1.19 (18016,5665,3966,1490) | [[+] Show genes](javascript:toggle('elements_GO:0031323')) |
| [GO:0080090](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0080090&view=details) | regulation of primary metabolic process | 7.12E-20 | 9.11E-17 | 1.19 (18016,5536,3966,1454) | [[+] Show genes](javascript:toggle('elements_GO:0080090')) |
| [GO:0051171](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0051171&view=details) | regulation of nitrogen compound metabolic process | 8.82E-20 | 1.04E-16 | 1.20 (18016,5383,3966,1418) | [[+] Show genes](javascript:toggle('elements_GO:0051171')) |
| [GO:0016043](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0016043&view=details) | cellular component organization | 9.36E-18 | 1.03E-14 | 1.20 (18016,4740,3966,1255) | [[+] Show genes](javascript:toggle('elements_GO:0016043')) |
| [GO:0071840](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0071840&view=details) | cellular component organization or biogenesis | 1.73E-17 | 1.77E-14 | 1.20 (18016,4782,3966,1263) | [[+] Show genes](javascript:toggle('elements_GO:0071840')) |
| [GO:0048522](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0048522&view=details) | positive regulation of cellular process | 3.78E-17 | 3.63E-14 | 1.19 (18016,4895,3966,1287) | [[+] Show genes](javascript:toggle('elements_GO:0048522')) |
| [GO:0048518](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0048518&view=details) | positive regulation of biological process | 7.16E-17 | 6.47E-14 | 1.18 (18016,5478,3966,1420) | [[+] Show genes](javascript:toggle('elements_GO:0048518')) |
| [GO:0032502](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0032502&view=details) | developmental process | 8.69E-17 | 7.42E-14 | 1.20 (18016,4567,3966,1208) | [[+] Show genes](javascript:toggle('elements_GO:0032502')) |
| [GO:0050794](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0050794&view=details) | regulation of cellular process | 1.49E-16 | 1.2E-13 | 1.10 (18016,9862,3966,2397) | [[+] Show genes](javascript:toggle('elements_GO:0050794')) |
| [GO:0051254](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0051254&view=details) | positive regulation of RNA metabolic process | 1.91E-15 | 1.46E-12 | 1.37 (18016,1591,3966,479) | [[+] Show genes](javascript:toggle('elements_GO:0051254')) |

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.

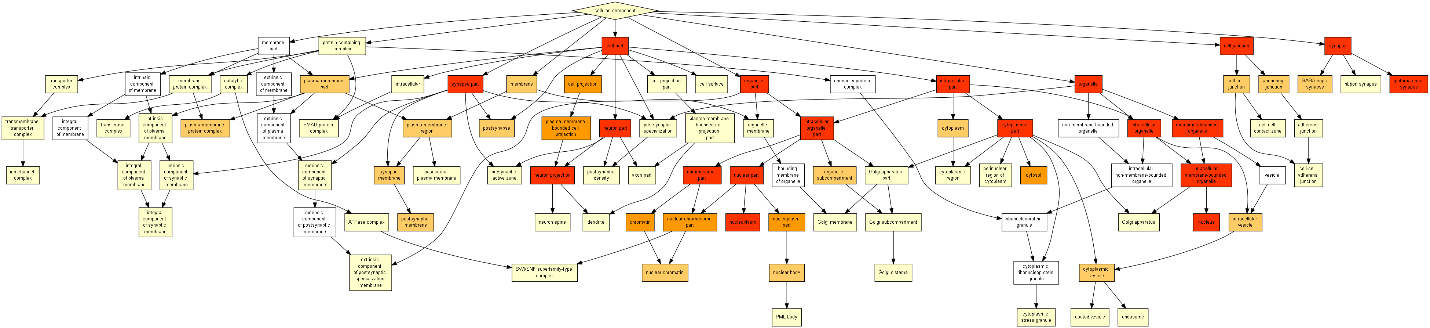
Here are the results for differentiation in molecular function:



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



And the results for cellular component:



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

1. 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](https://www.ncbi.nlm.nih.gov/gtr/conditions/C1842930)

[Severe congenital neutropenia 2, autosomal dominant](https://www.ncbi.nlm.nih.gov/gtr/conditions/C2751288)