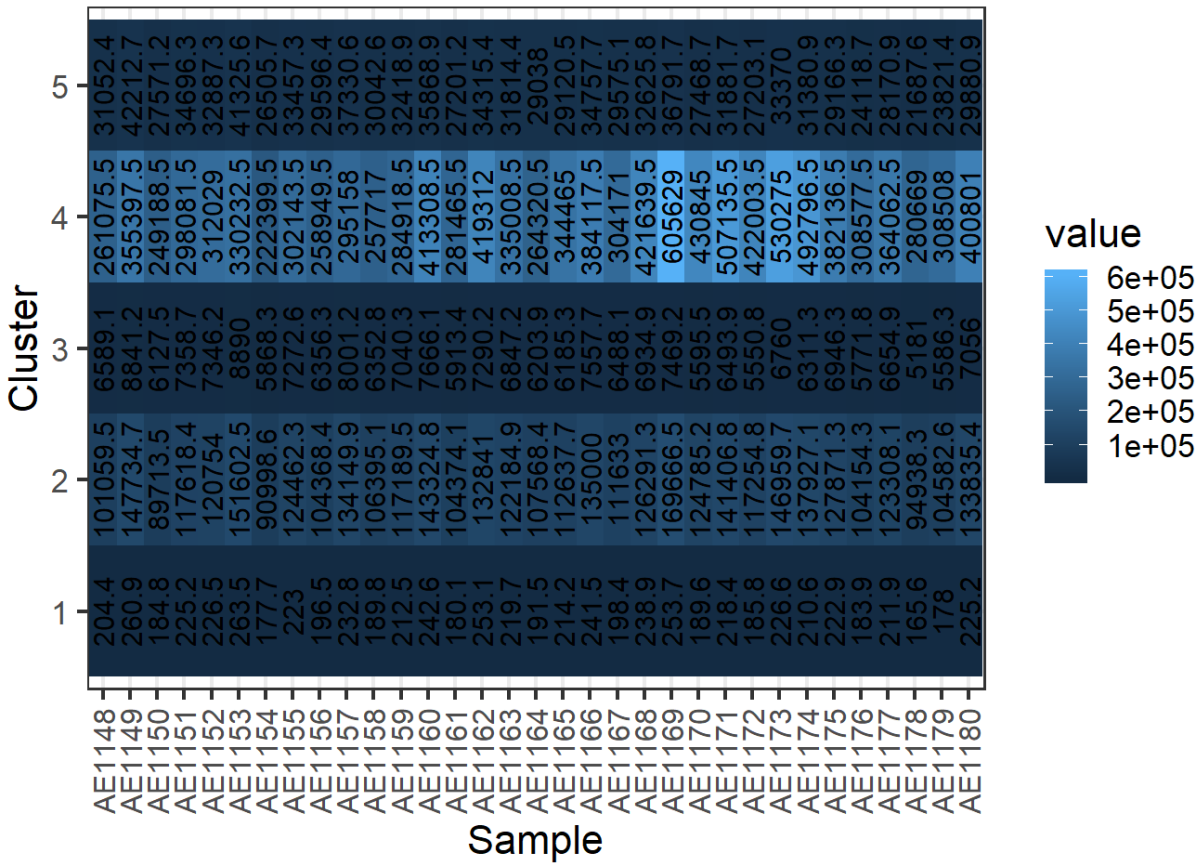
HW4 – Bioinformatics – 236523

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

Question 1

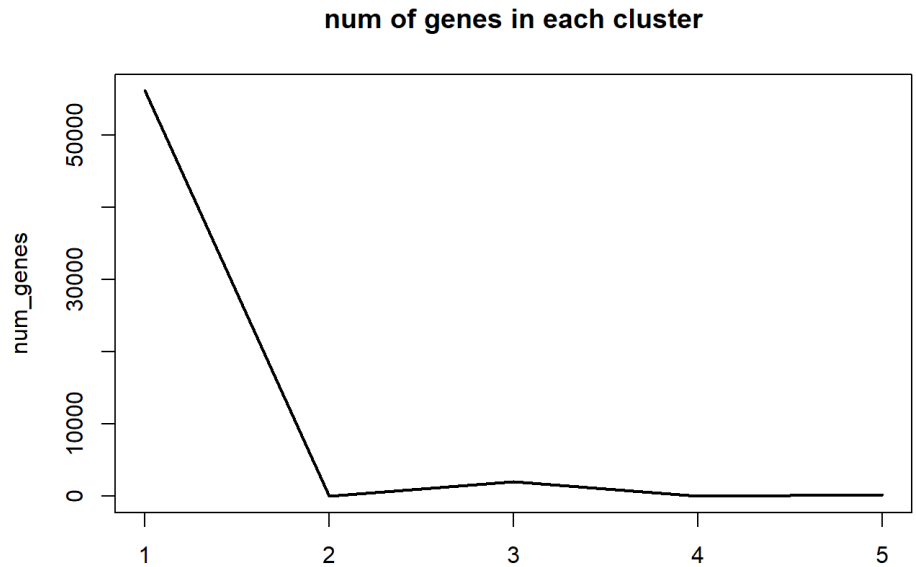
2. c. ii.



iii.

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

Plot:



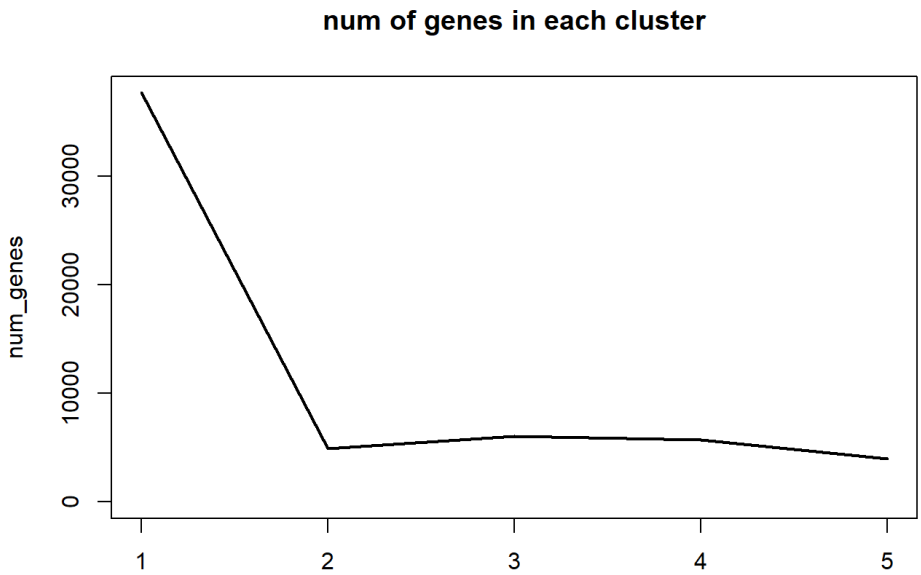
iv. We see that the number of genes in each cluster varies significantly – from 2 genes in cluster number 1, to 56,101 genes in cluster number 5. This is a shortcoming since most of the genes appear in a single cluster which is not very informative for expression analysis.

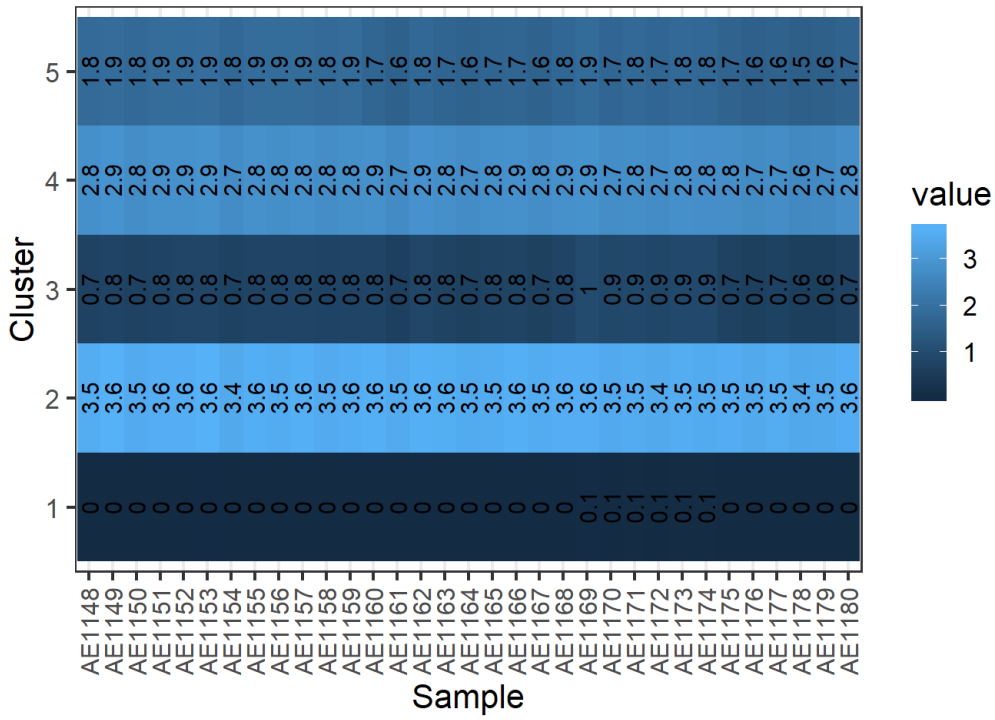
Another problem is that for clusters 1 and 2, the range of values that belong to the same cluster is big, therefore in the heatmap it looks like cluster 1 is not homogeneous. There are two possibilities: either the big difference in values (for example there is a value of 605,629 together with 222,399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

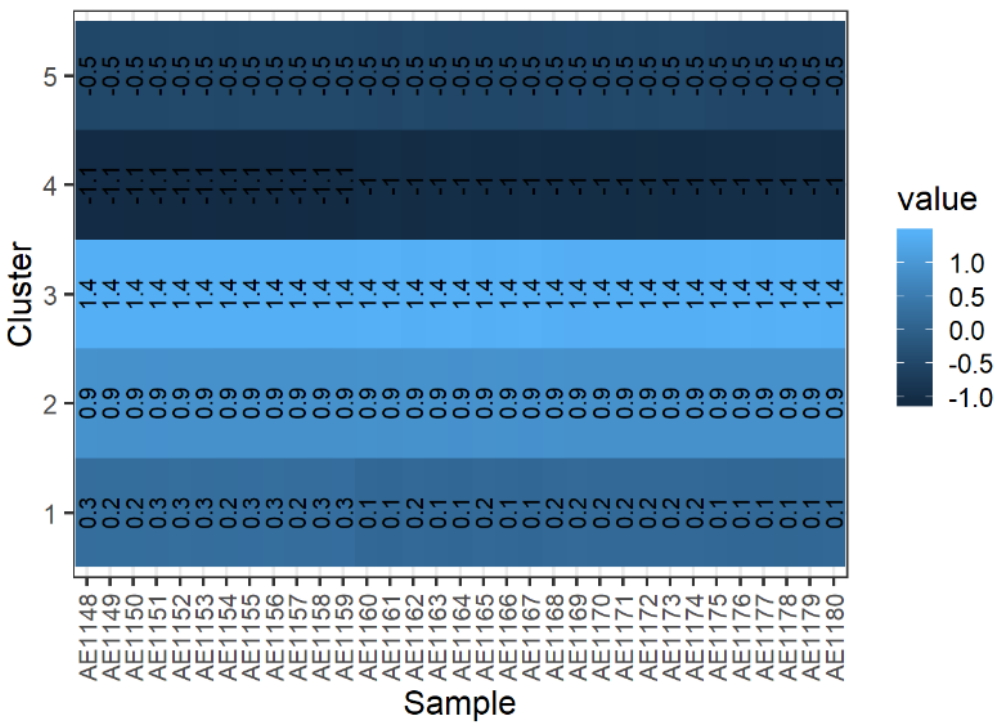
of the larger values are found in the same cluster (since big differences after log transformation become much smaller and examples which previously were considered ‘too different’ are now similar).

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
```

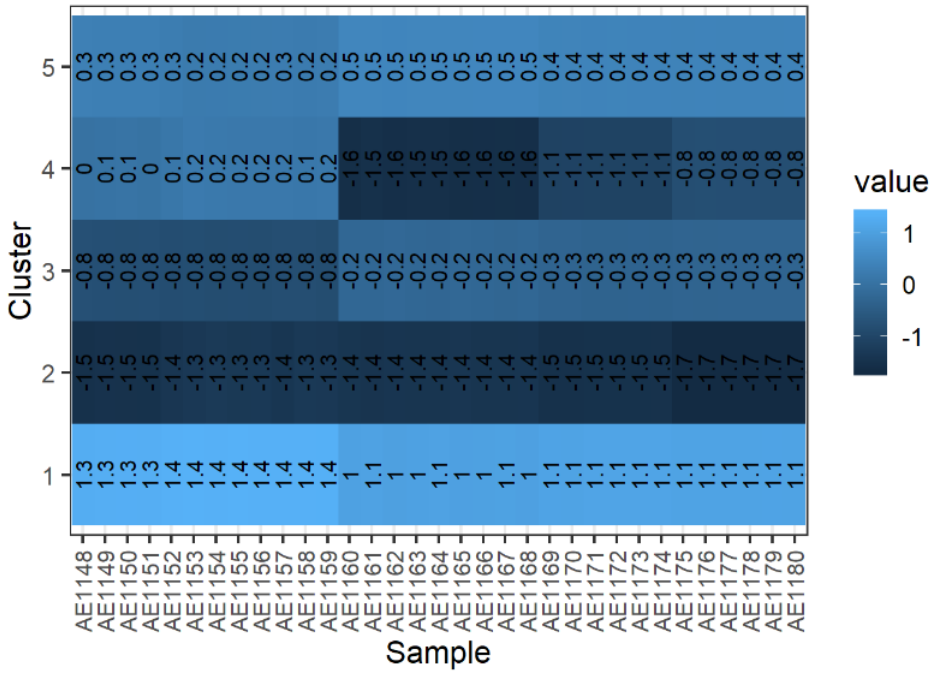


As we can see, the clusters are more homogenous now and also different from each other.

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
```

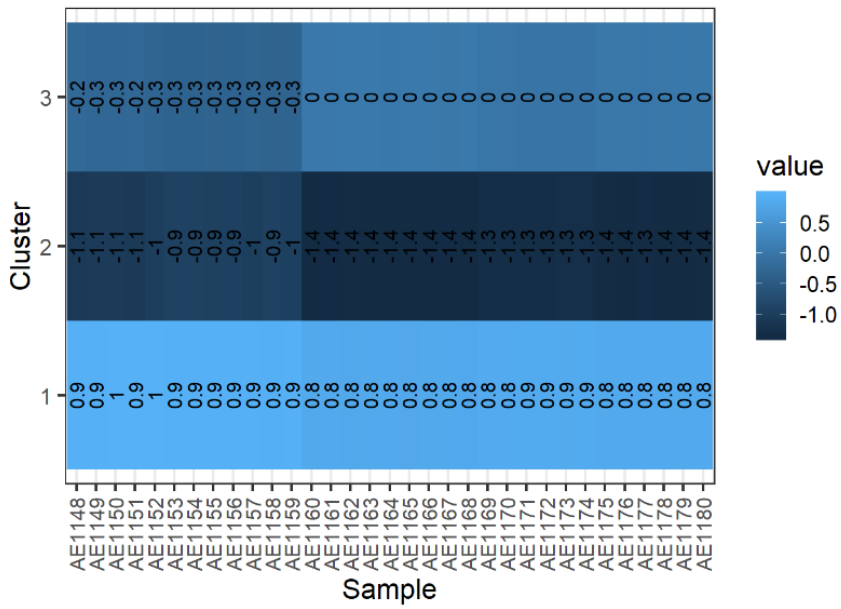
```
171 121 202  79 348
```



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:

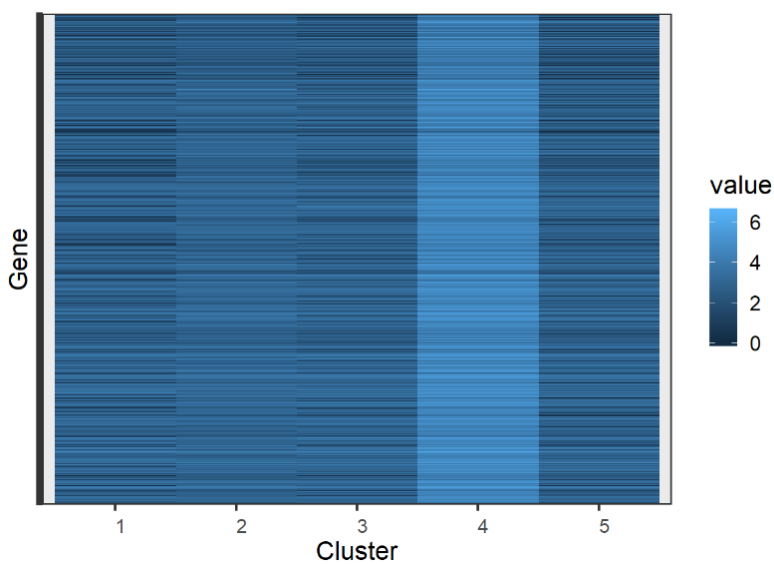
Cluster sizes -

```
  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).

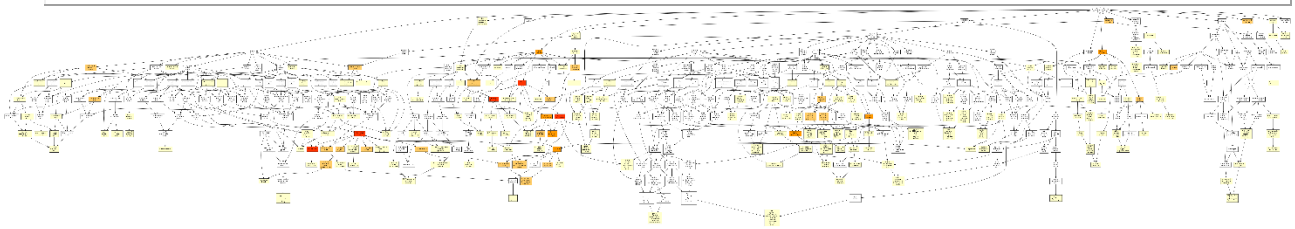
The results of clustering samples are harder to work with, as many samples in different clusters have similar expression levels in many genes. However, there are some differences between clusters, and some genes do show distinct values. Clustering samples can, for example, help us understand which treatments 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

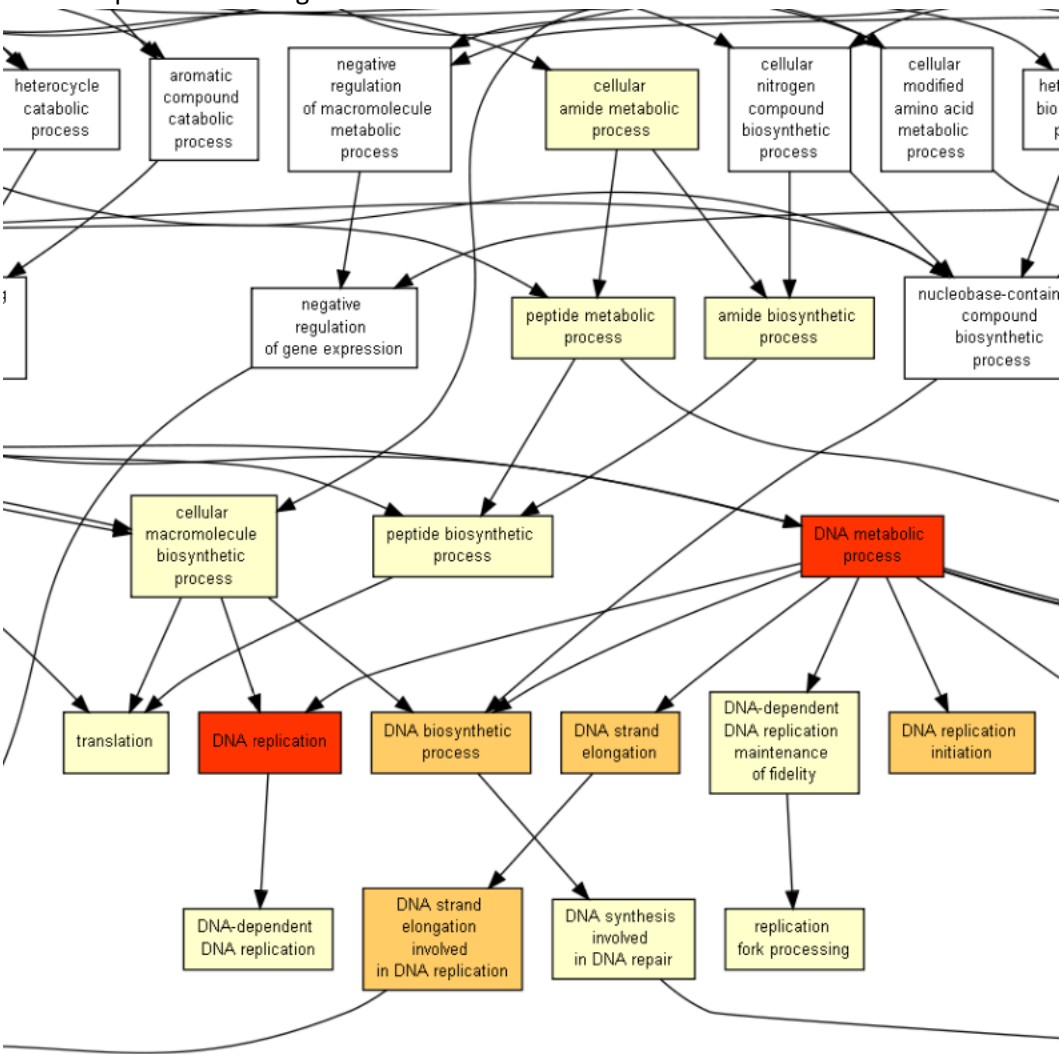
1. C. We got a long list of GO terms that were found by Gorilla, the most significant ones are:

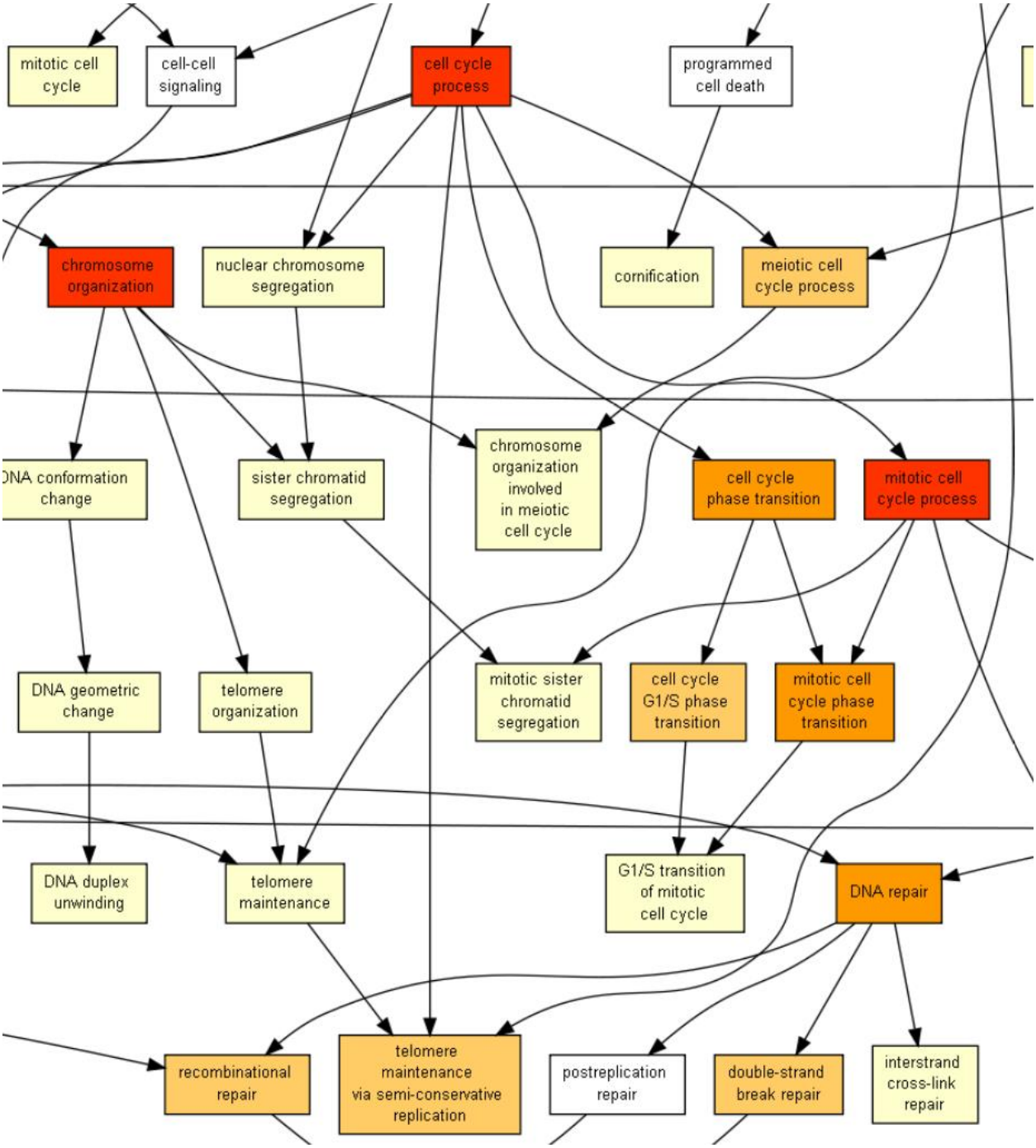
GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
GO:0022402	cell cycle process	8.7E-17	1.22E-12	2.01 (11732,764,1261,165)
GO:0006260	DNA replication	2.6E-14	1.82E-10	4.04 (11732,136,940,44)
GO:1903047	mitotic cell cycle process	2.91E-14	1.36E-10	2.18 (11732,504,1261,118)
GO:0051276	chromosome 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)



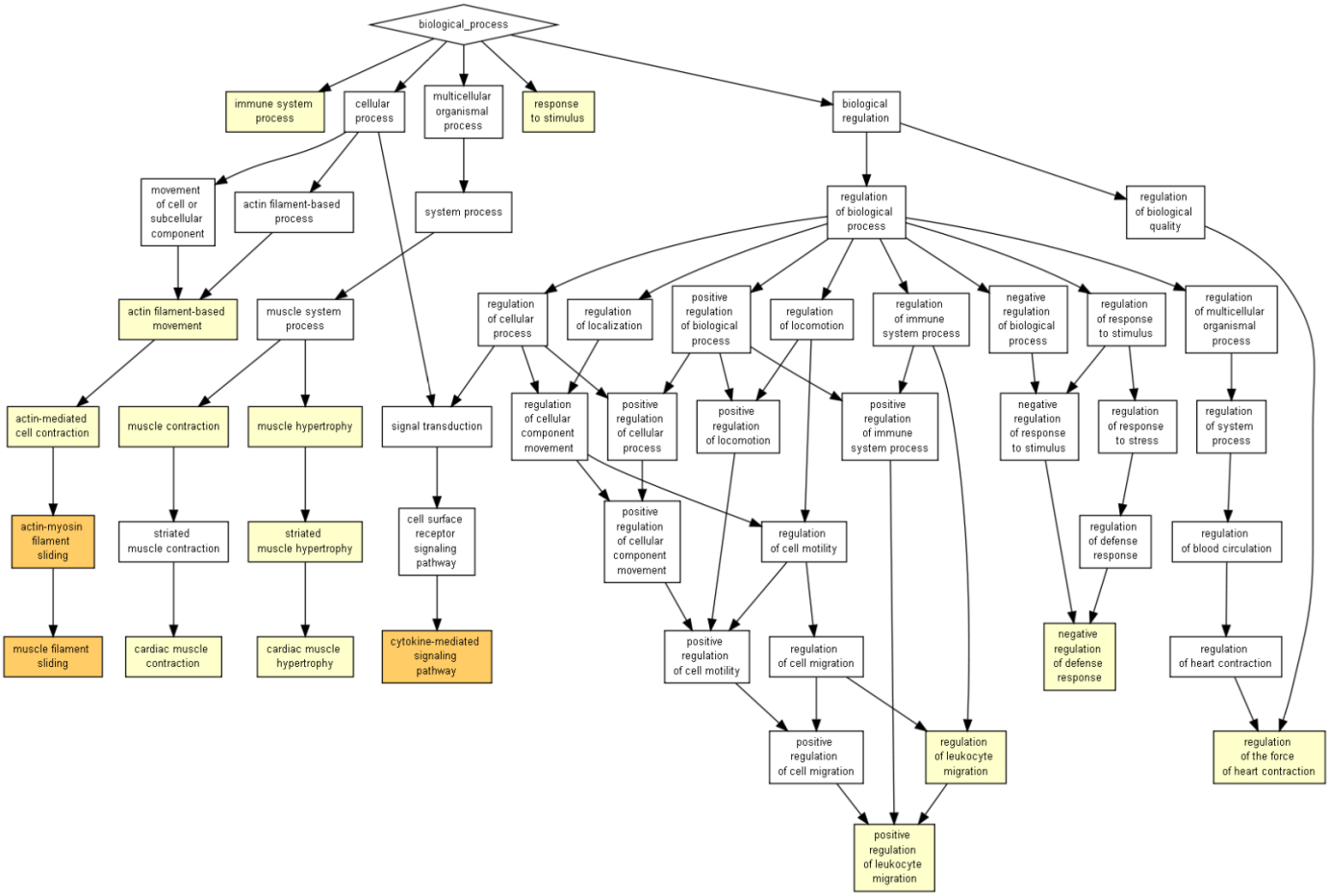
Relevant parts of the diagram are:





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:



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)

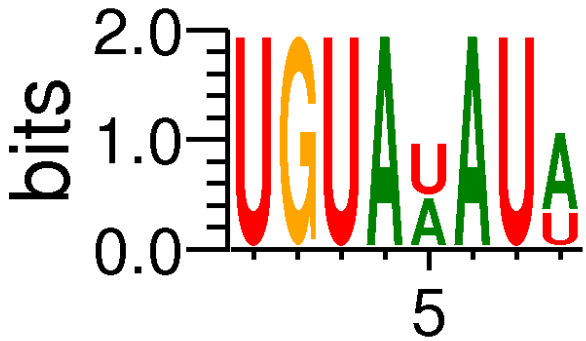
GO:0002687	positive regulation of leukocyte migration	1.62E-4	4.11E-1	27.70 (17785,107,18,3)
GO:0002376	immune system process	1.67E-4	3.64E-1	4.46 (17785,1772,18,8)
GO:0003300	cardiac muscle hypertrophy	1.82E-4	3.47E-1	98.81 (17785,20,18,2)
GO:0014897	striated muscle hypertrophy	2.01E-4	3.4E-1	94.10 (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

<u>What can you say about the motif?</u>

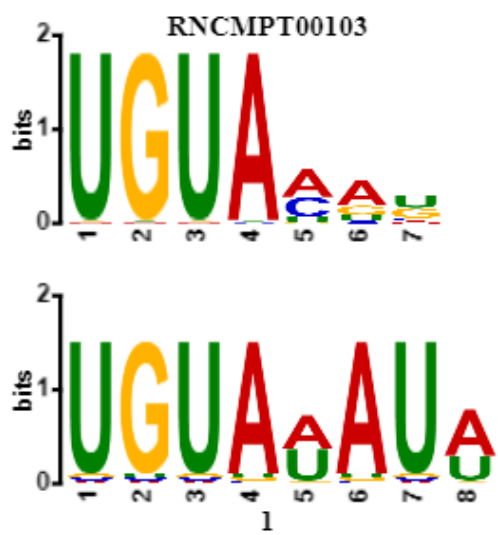
This is the motif we found in the first result of TOMTOM search

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: No	Predicted: Yes
Actual: No	8636	336

Actual: 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

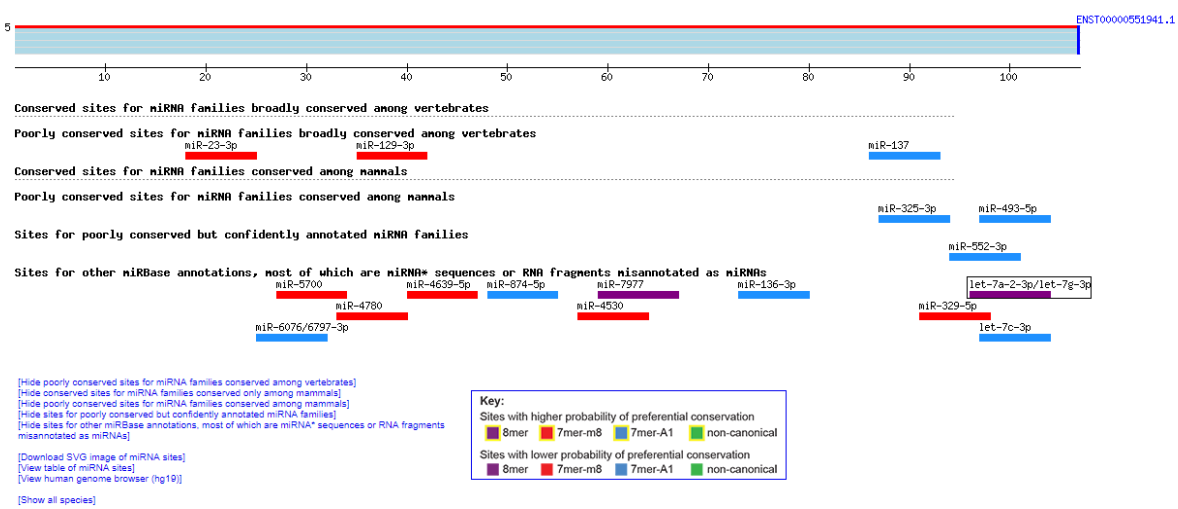
1. Screenshot of the results:

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]

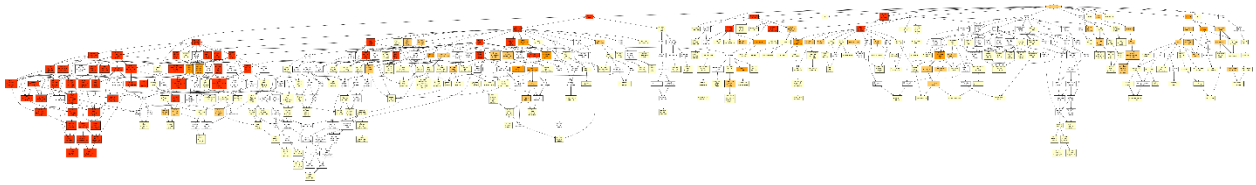
Target gene	Representative transcript	Gene name	Number of 3P-seq tags supporting UTR + 5	Link to sites in UTRs	Site counts				6mer sites	Representative miRNA	Cumulative weighted context++ score	Total context++ score	Aggregate P_{CT}
					total	8mer	7mer-m8	7mer-A1					
RP11-1105G2.3	ENST00000551941.1	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

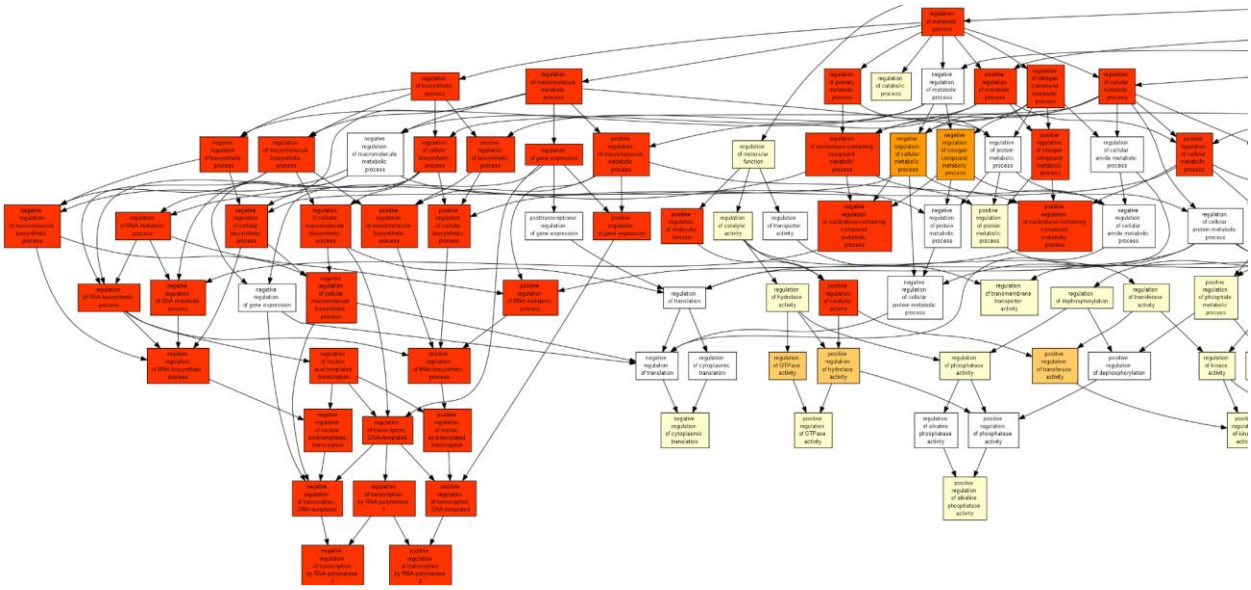
Human RP11-1105G2.3 ENST00000551941.1 3' UTR length: 106



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:



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

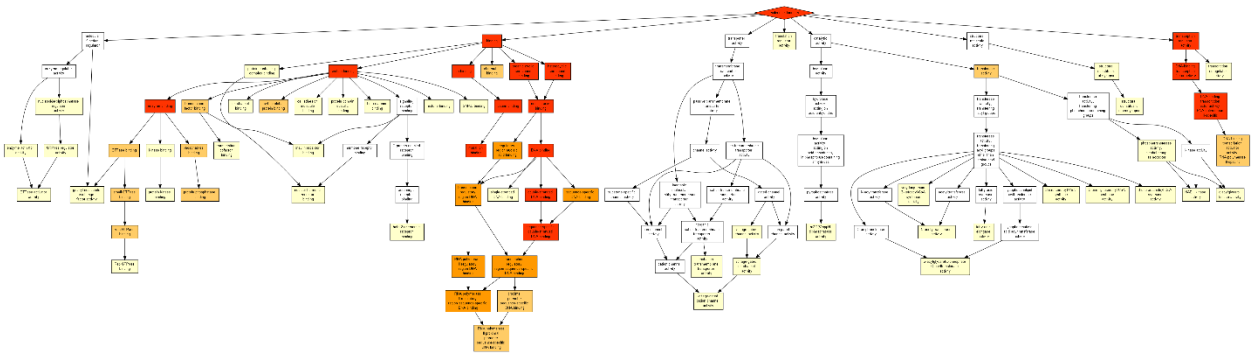
GO:0019219	regulation of nucleobase-containing compound metabolic process	5.48E-27	4.21E-23	1.30 (18016,3709,3966,1063)	[+] Show genes
GO:2000112	regulation of cellular macromolecule biosynthetic process	2.51E-25	1.29E-21	1.30 (18016,3612,3966,1031)	[+] Show genes
GO:1903506	regulation of nucleic acid-templated 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

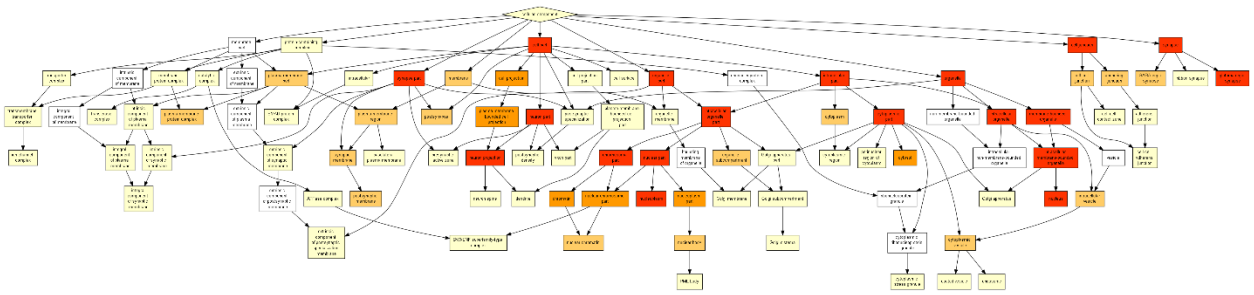
These are the top go terms from the table of results sorted by p-value. The table of results sorts the results in the go tree by p-value to show the order of differentiation of each term by p-value, fdr q-value and Enrichment.

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:

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