Assignment Molecular Systems Biology (SSB-30306). Dataset 5. Fernando Bueno Gutierrez.

PART I. Preliminary steps

There are 177 genes in 287 conditions.

There are 8 missing values. They correspond to 4 conditions and 2 genes, as follows:

experiment/gene	Rv3563	Rv3199c
SV.D09	NA	NA
X0.12mg.mL.PZA.pH5.6..5.5h.	NA	NA
X1ug.mL.ARP2..DMSO..6h	NA	NA
X25ug.mL.CPZ...0.1mM.GSNO..DMSO.ctrl..2.5h	NA	NA

Illustration 1: Mising data

As preliminary steps we check:

- Whether the data has been normalized.
- Overview of the difference between the expression pattern of the genes generating a dendrogram, a heatmap and with PCA analysis.
- Correlation between the expression patterns of the genes.

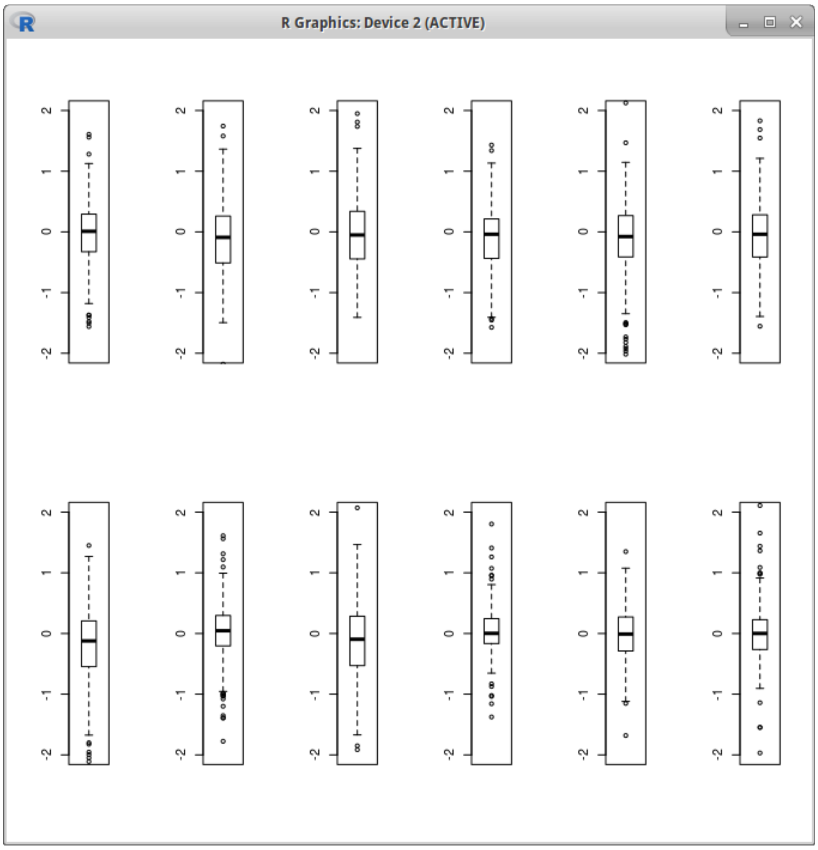


Illustration 2: Check normalization of experimental conditions

Illustration 2 corresponds to the box-plot of 12 randomly chosen experimental conditions. The data seems to be normalized. To be safe, it is require to check the box-plot of each experimental condition (not only 12).

Comparing euclidean vs correlation distances of genes will give us some insights on whether the data has been standardized (if the data has been standarized, euclidean and correlation distances are equivalent). First we will have a look at a dendrogram, of the genes based on euclidean distance.

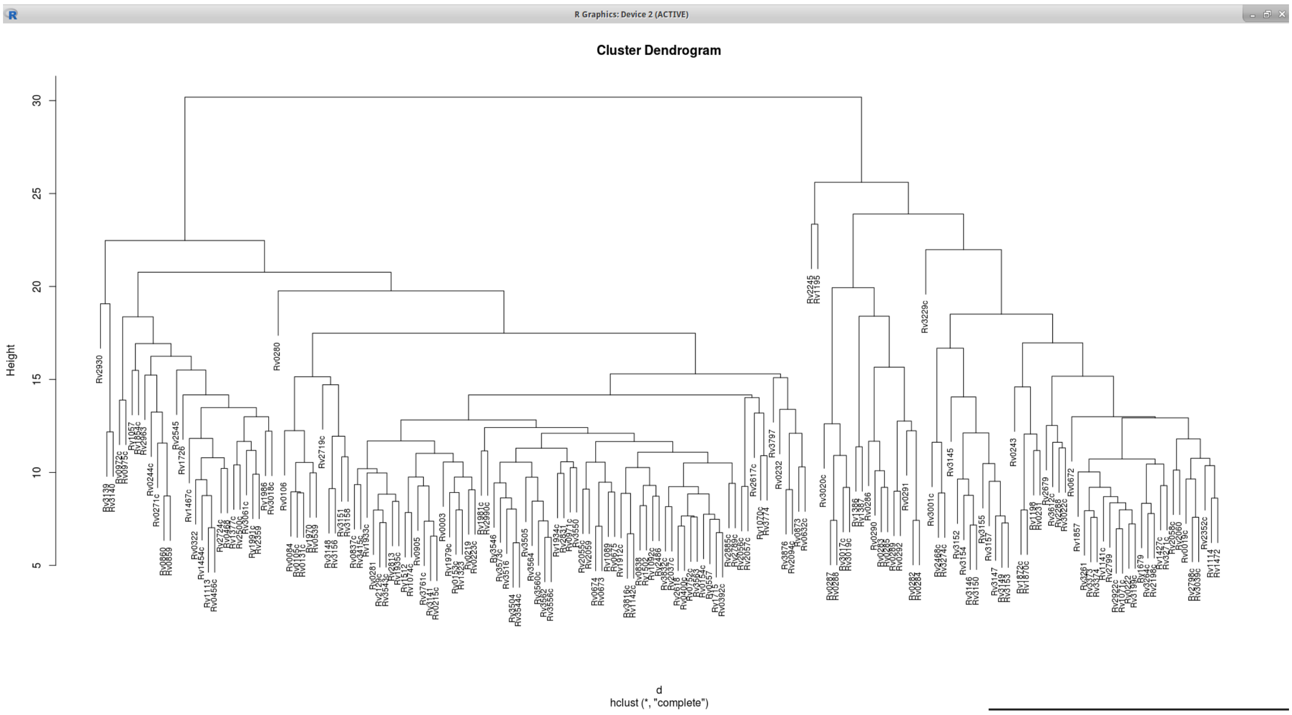


Illustration 3: Overview dendrogram

Genes RV1114 and RV1472, for instance, (right corner) seem to have similar expression pattern and very different from genes such as RV2930 and RV3139, for instance. The dendrogram allows for a first overview of the similarities between genes. There seem to be a large cluster with about 2/3 of the genes and another cluster with about 1/3. These are, in principle good proportions. As the clusters are made it is always desirable to increase the variability between clusters and reduce the variability within clusters.

Now we compare correlation vs euclidean distance by comparing the dendrograms:

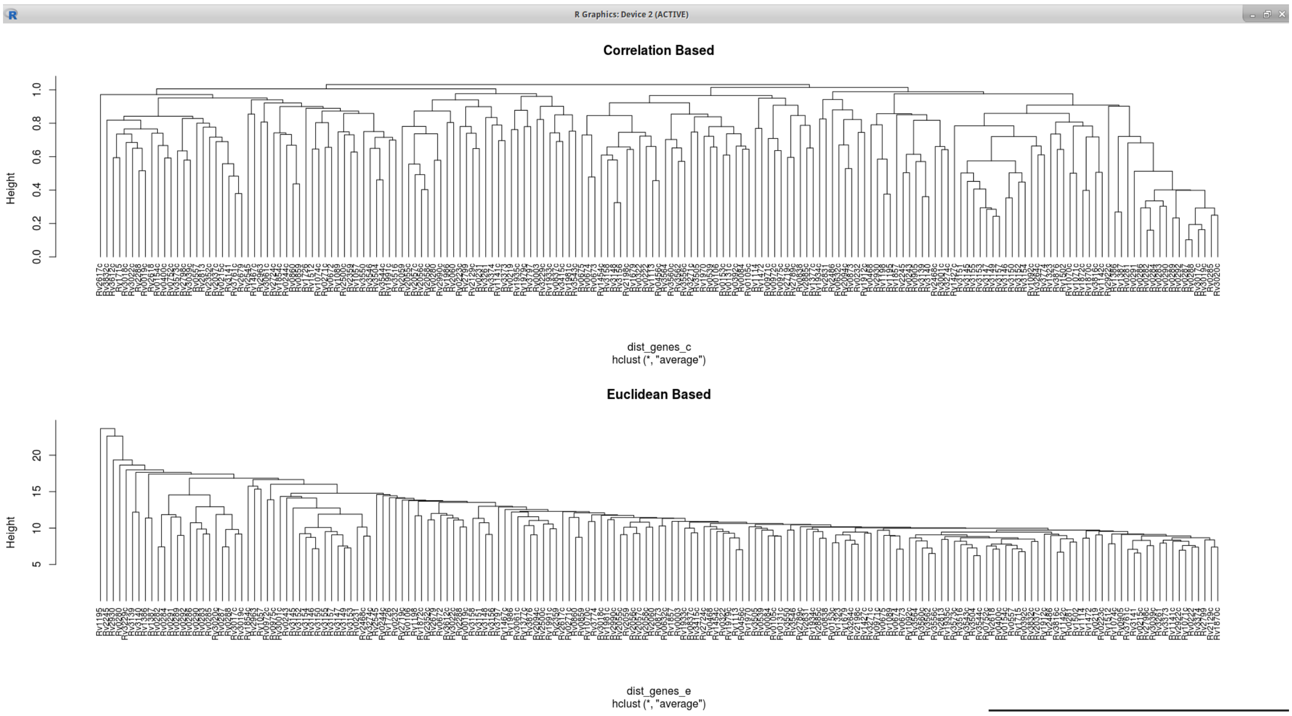


Illustration 4: Euclidean vs correlation dendrogram

The euclidean distance is based on the overall expression values of the genes, whereas the correlation distance is based on how these values change from one condition to another. To identify clusters of genes with similar functions given expression data, correlation distance is preferred. Since both dendrograms differ substantially, we may have to check at this point whether the data has been standardized.

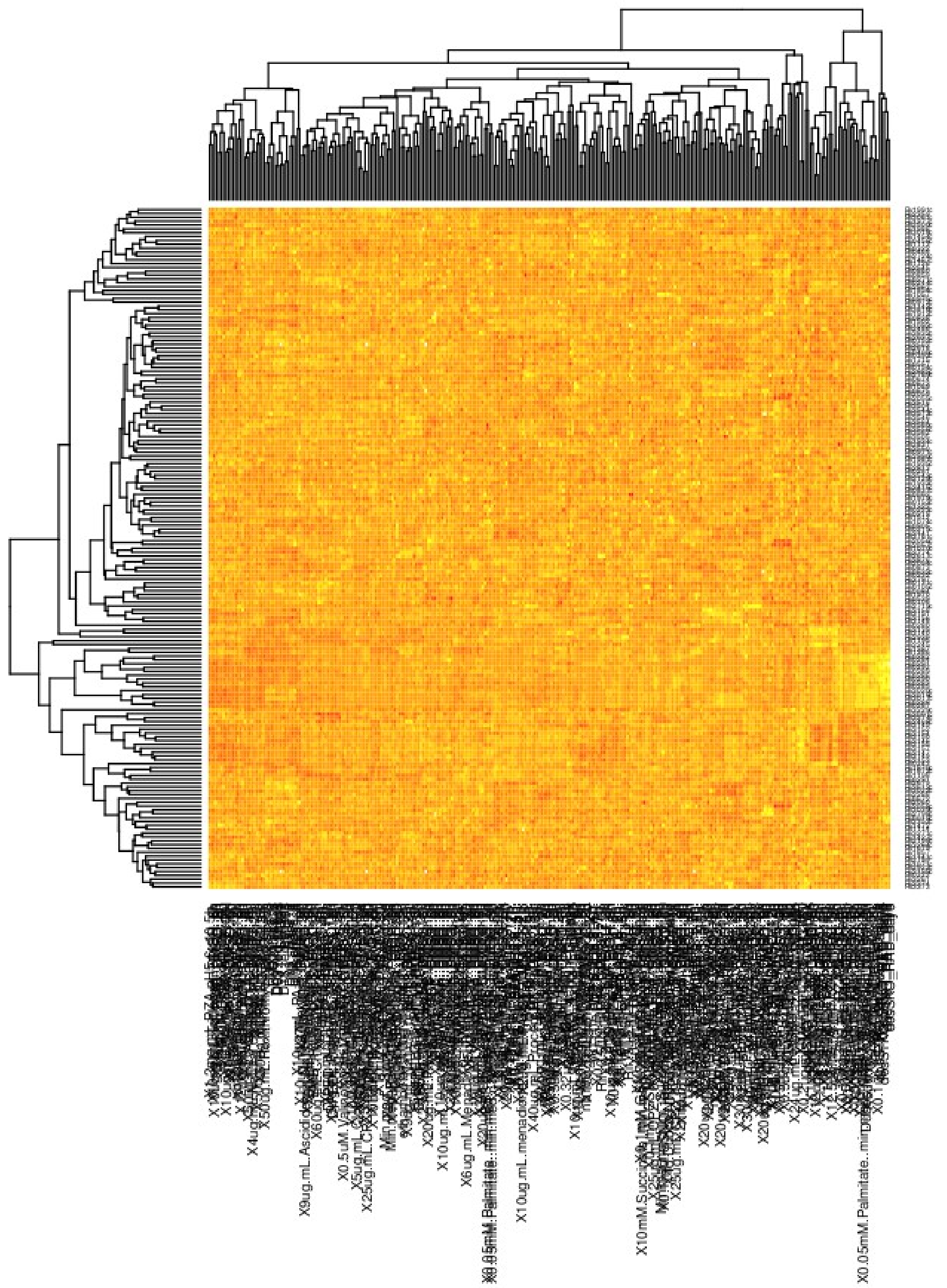
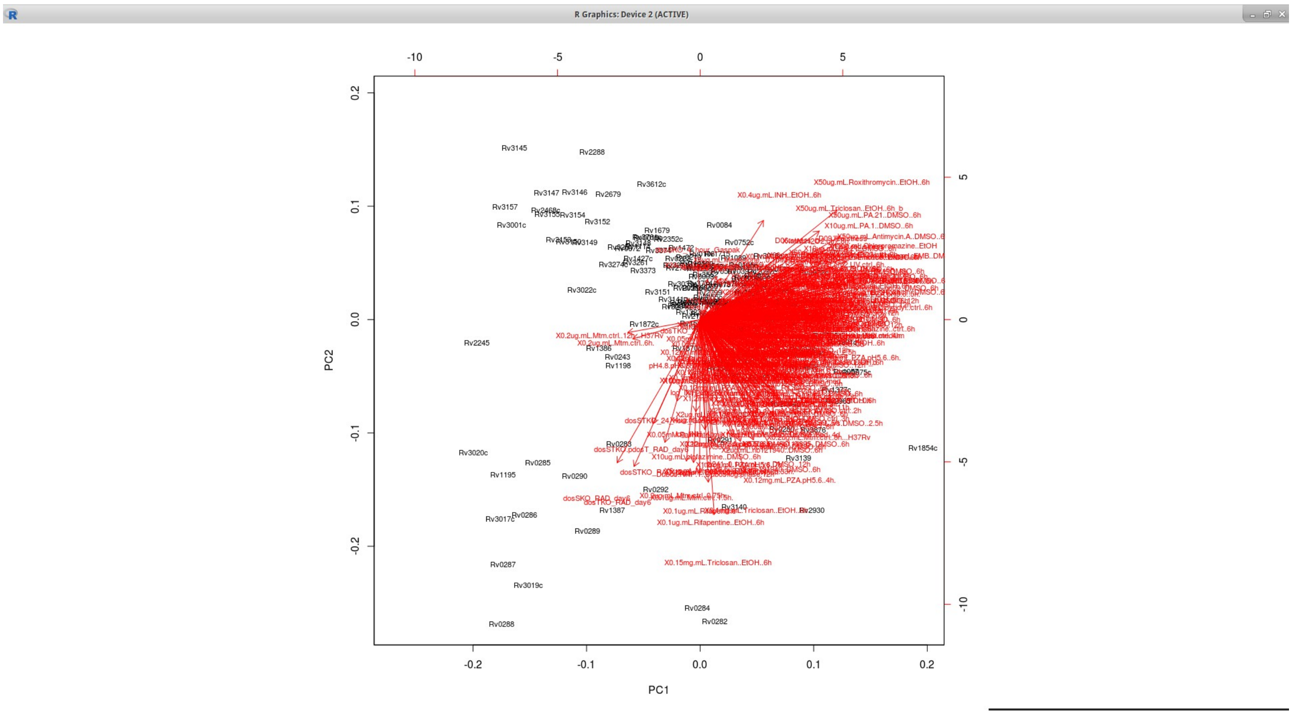


Illustration 5: Overview heatmap of gnes and experimental conditions.

The heatmap includes a dendrogram for the genes and another for the experimental conditions. When zooming in we can see that genes like RV1114 and RV1472 have similar expression patter, as in Illustration 3. The heatmap does not reveal much difference between genes or experimental conditions. Moreover, the variation across genes and across experimental conditions seems to be proportioned.

The PCA analysis allows for a simple interpretation of which genes and experimental conditions are more alike. In this case, however, the interpretation may be too simplistic because the first 2 PCA explain less than 30% of the variance (29,67%).



Also here we can compare results when we use euclidean vs correlation distance.

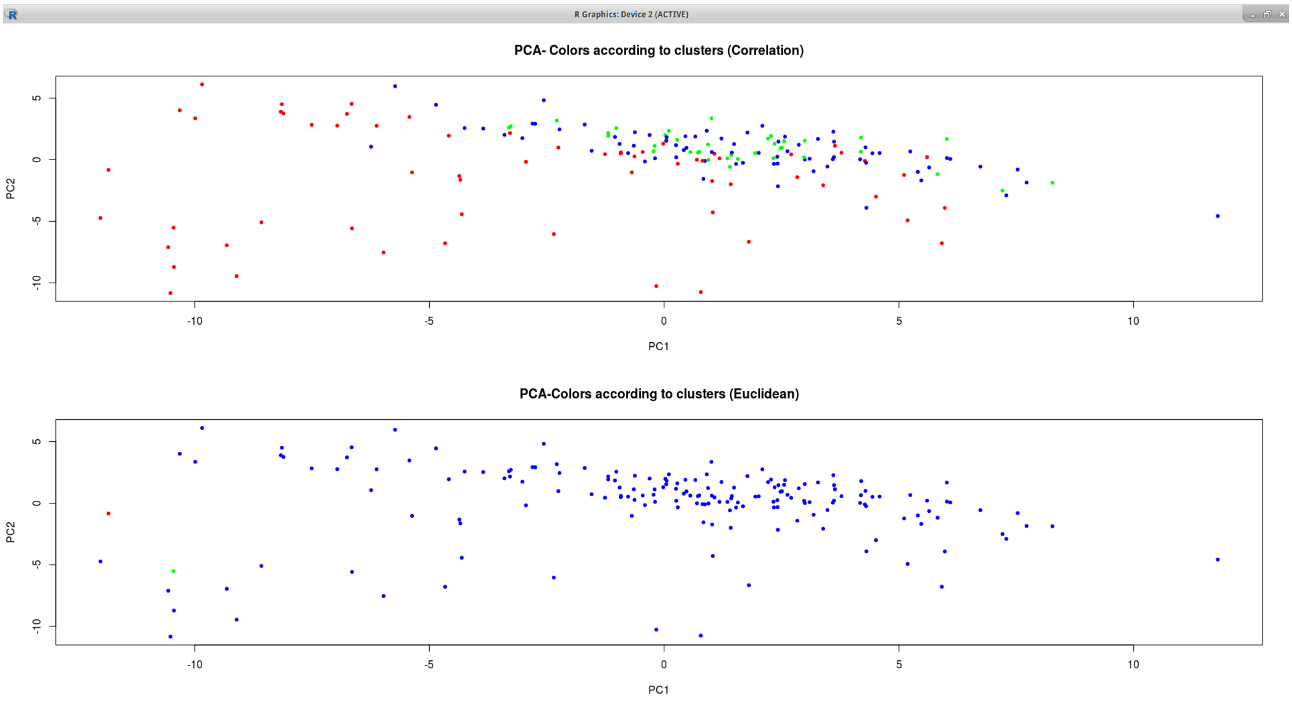

R Graphics: Device 2 (ACTIVE)
PCA- Colors according to clusters (Correlation)
PCA-Colors according to clusters (Euclidean)
PC1
PC2
-10
-5
0
5
10

Illustration 6: PCA euclidean vs correlation

We observe that the correlation method in this cases allows for a more balanced clusters size. If the principal components are a good representation of the futures (experimental conditions), the genes that appear together in the deprogram or that show intense colors in the heatmap, will also be grouped together in the PCA.

Now we calculate the correlations between genes.
The top 10 correlations are:

corr	gene1	gene2
0.897	Rv0287	Rv0288
0.847	Rv0288	Rv3019c
0.84	Rv0282	Rv0284
0.837	Rv3017c	Rv3019c
0.81	Rv0287	Rv3019c
0.79	Rv0287	Rv3017c
0.77	Rv0288	Rv3017c
0.77	Rv0285	Rv0287
0.767	Rv0289	Rv0292
0.76	Rv0286	Rv0288

Table 1: Top 10 strongest positive correlations

Whereas the top 10 negative correlations are:

corr	gene1	gene2
-0.405	Rv3543c	Rv0456c
-0.410	Rv2930	Rv3157
-0.415	Rv1195	Rv0752c
-0.420	Rv1195	Rv0400c
-0.425	Rv1198	Rv0271c
-0.426	Rv1387	Rv2500c
-0.427	Rv0557	Rv1198
-0.438	Rv1386	Rv2500c
-0.442	Rv1857	Rv0400c
-0.449	Rv2468c	Rv2359
-0.519	Rv2930	Rv0271c

Table 2: Top 10 strongest negative correlations

The top 10 weakest correlations are:

corr	gene1	gene2
9.47E-005	Rv0105c	Rv0281
7.47E-005	Rv3140	Rv3154
7.11E-005	Rv3573c	Rv0673
1.26E-005	Rv3761c	Rv0632c
6.17E-006	Rv3816c	Rv0392c
-4.58E-006	Rv0859	Rv3157
-7.92E-005	Rv2963	Rv1427c
-1.10E-004	Rv1377c	Rv2060
-1.20E-004	Rv1195	Rv3150
-1.27E-004	Rv1679	Rv1070c

Table 3: Top 10 weakest correlations

Pairs of genes that are highly correlated with positive value may be involved in related functions and respond either, positively or negatively to the treatment. Whereas pairs of genes that show a strong negative correlation correspond to pairs in which when one gene responds positively to the treatment the other responds negatively (less expression). These genes may be involved in competing biological functions. Genes with week correlation are expected to be involved in unrelated functions. The following image allows for a simple interpretation, where the X-axis are the 287 experimental conditions and the Y-axes are the Fold Changes between the studied condition and a reference condition.

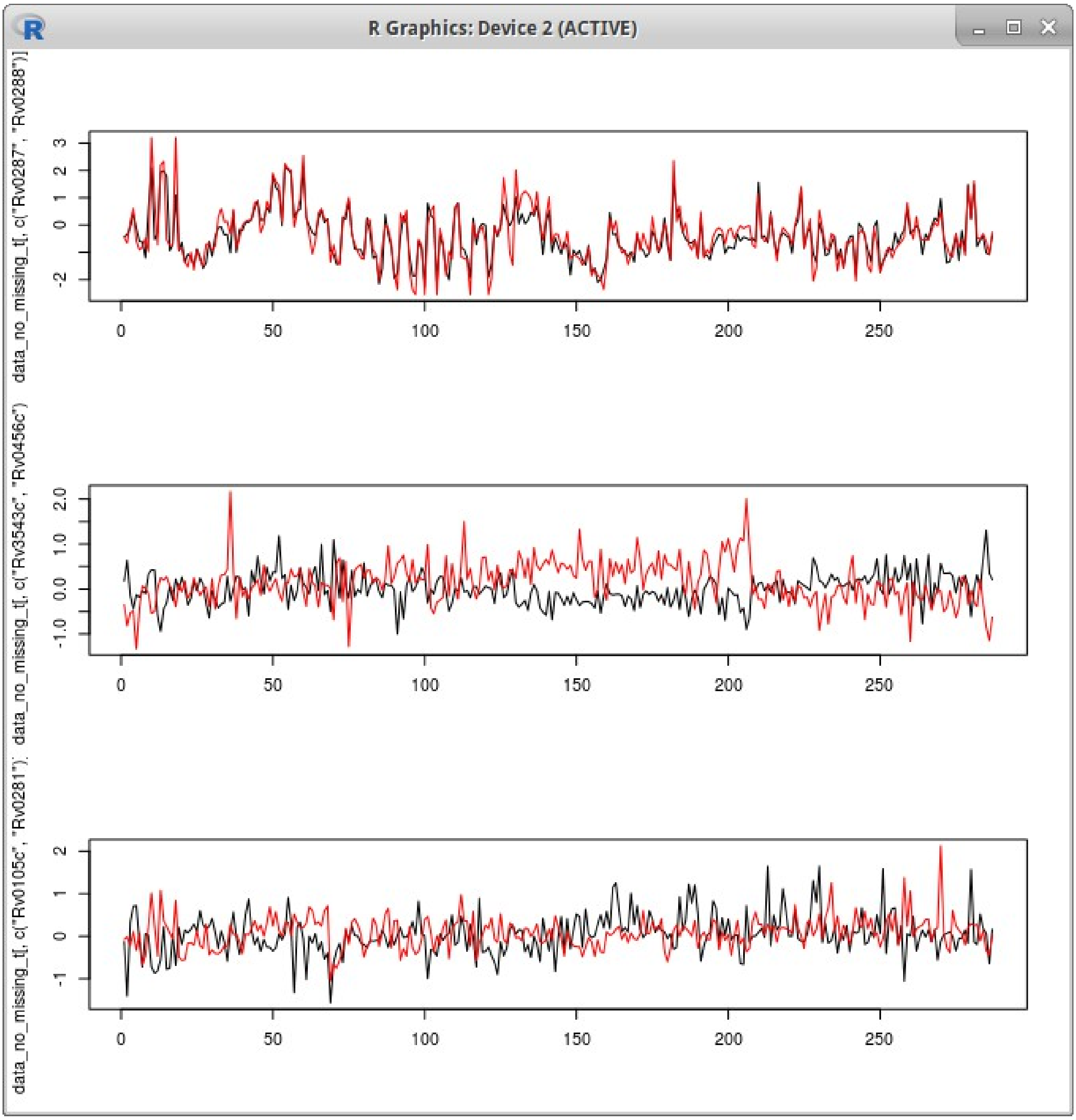


Illustration 7: Expression pattern of the pair of genes with strongest positive correlation, weakest correlation and strong negative correlation, receptively.

The expression pattern is similar for genes that are strongly correlated, it is opposite for genes that are negatively correlated and it is is arbitrary for pairs of genes that are weakly correlated.

PART II. Gene networks

We study three of the main methods of network inference: Context Likelihood of Relatedness (CLR), ARACNE and minimum-redundancy network (MRNET).
With CLR it is easy to introduce false edges caused by indirect interactions. Whereas ARACNE is based on Data Processing Inequality, wherein indirect interactions in interaction triangles are considered. MRNET is useful for feature selection strategy, such as finding the genes that are more associated with a given function.
Since the genes in the dataset are overall highly co-express (high number of edges), CLR is not the most appropriate method, as the number of edges would be even higher. We will start the analysis with the three methods to get an overview of the differences and then we will proceed with ARACNE and MRNET.

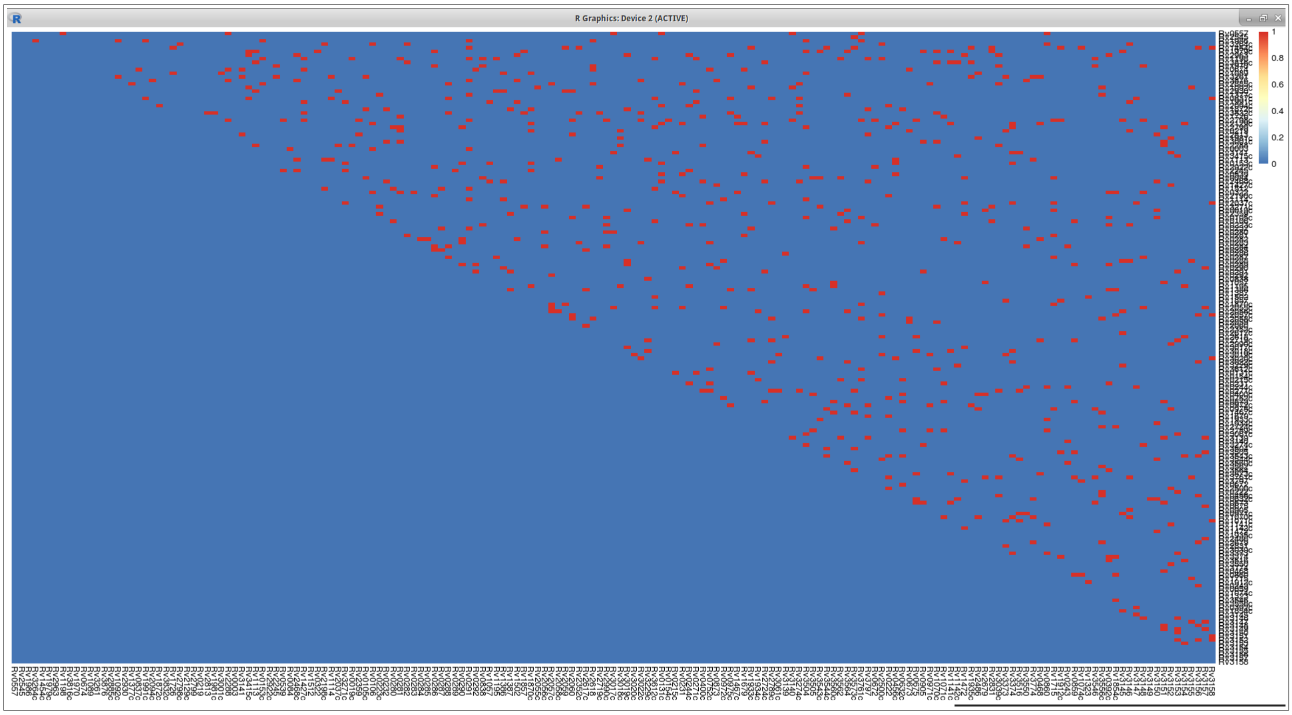


Illustration 8: Heatmap_method_ ARACNE

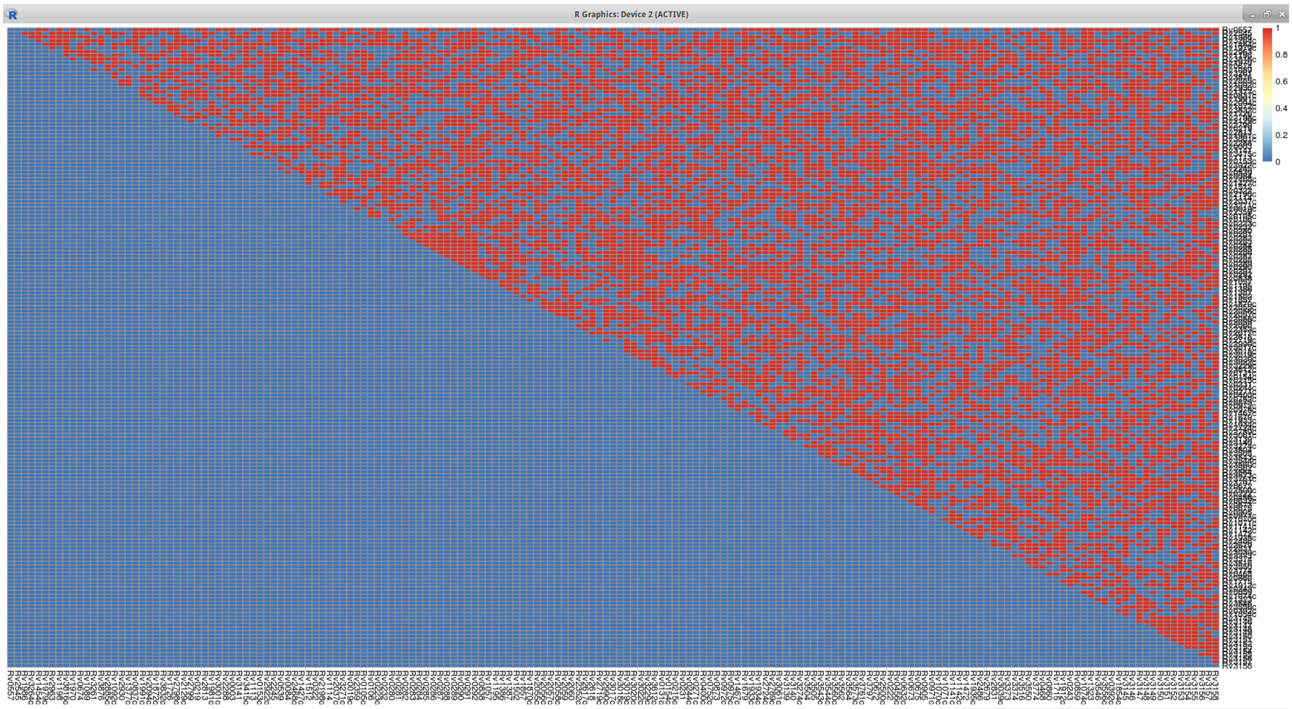


Illustration 9: heatmap_method_CLR

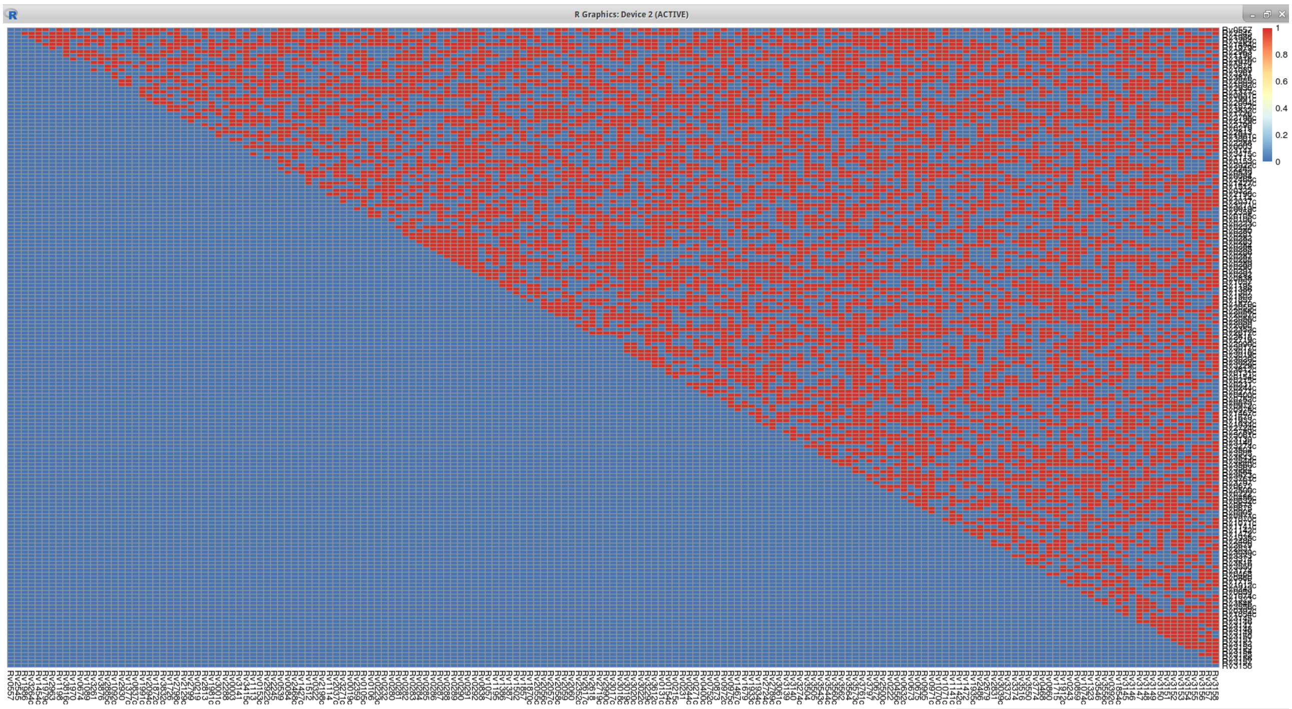


Illustration 10: heatmap_method_ MRNET

In illustrations 7 to 10 it can be observed that the method ARACNE is more strict and the number of correlations is lower. This is because there are few interaction triangles. With the method MRNET the number of direct interactions have increased through an iterative search for interactions. The magnitude of these interactions, however, may be lower as we see in the next plots.

An important decision is what we consider “coexpression”. In other words, how strict we should be when we ask whether a pair of genes have a “similar” co-expression pattern. We test two thresholds: 0.1 and 0.5. 0.5 is more strict, so less number of edges.

The networks can be visualized with the R package “ggnet”. The following plots correspond to the methods ARACNE with a threshold of 0.1 (697 egdes), ARACNE with threshold of 0.5 (332 edges) and MRNET with threshold 0.1 (319 edges). The number of nodes is in all cases 175 genes. In the case of ARACNE 0.5, the network is disconnected. The higher the threshold, the more disconnected.

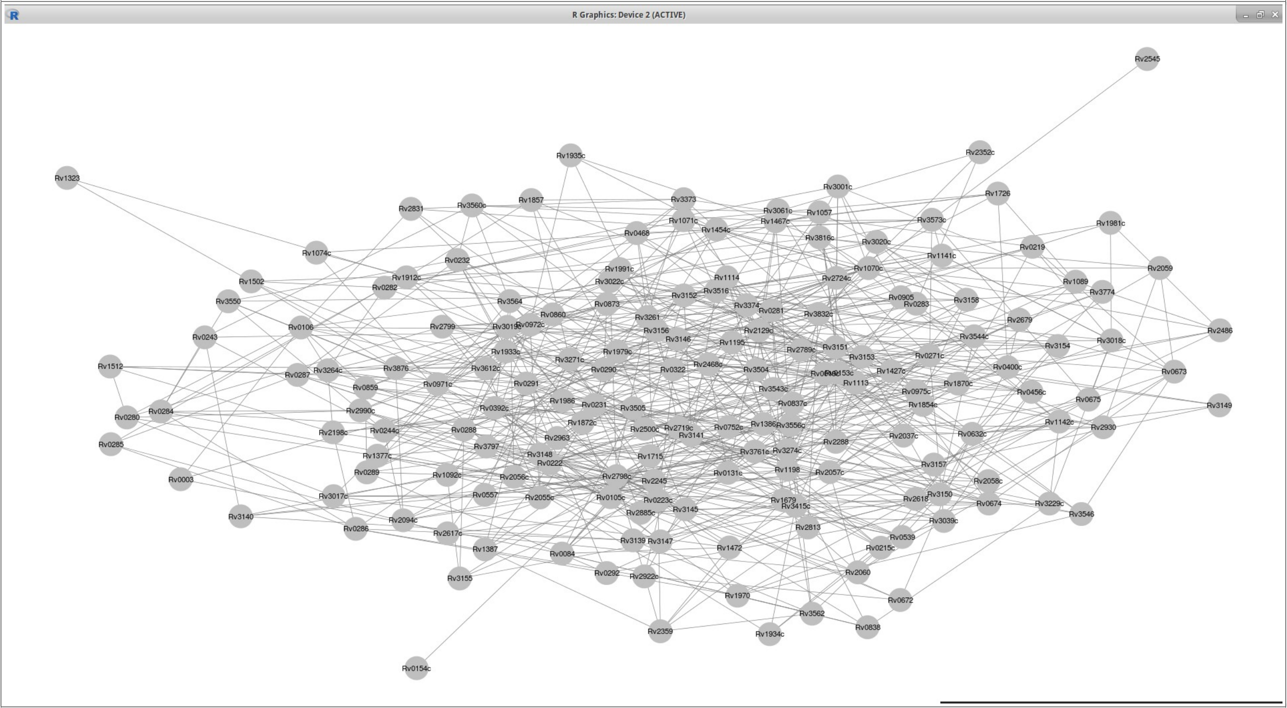


Illustration 11: Network ARACNE 0.1, visualized with R ggnet

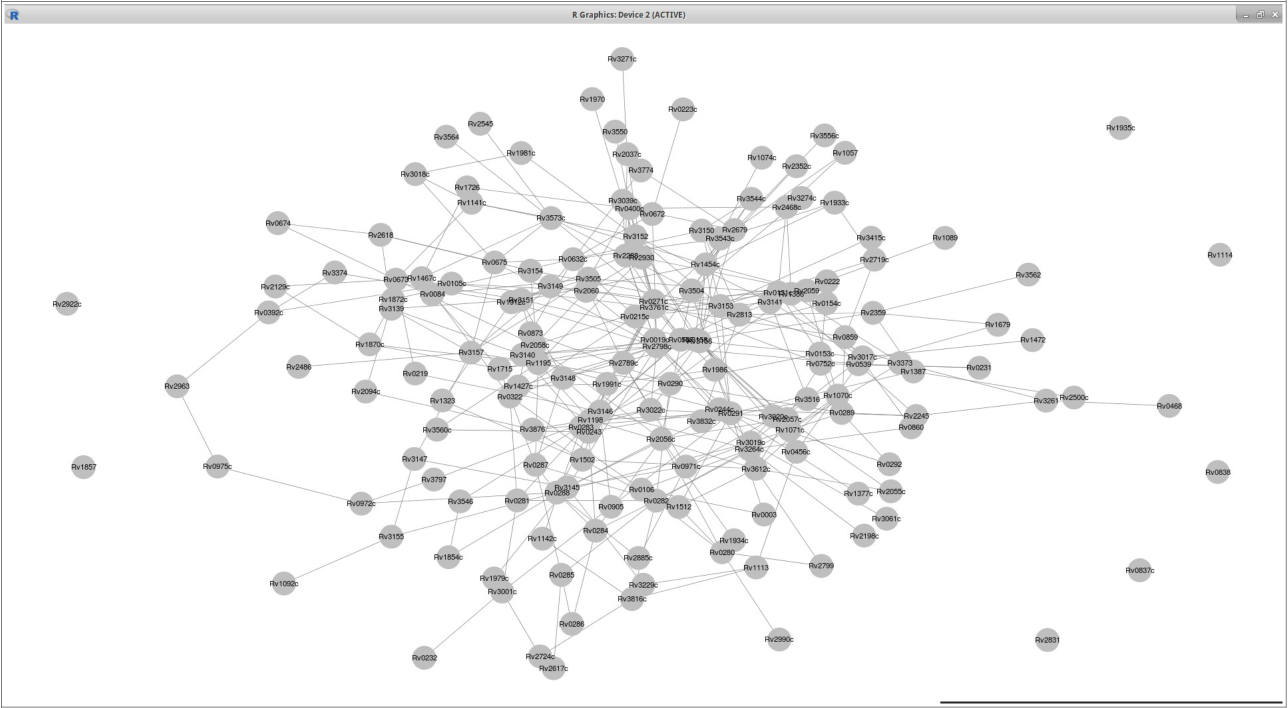


Illustration 12: Network ARACNE 0.5, visualized with R ggnet

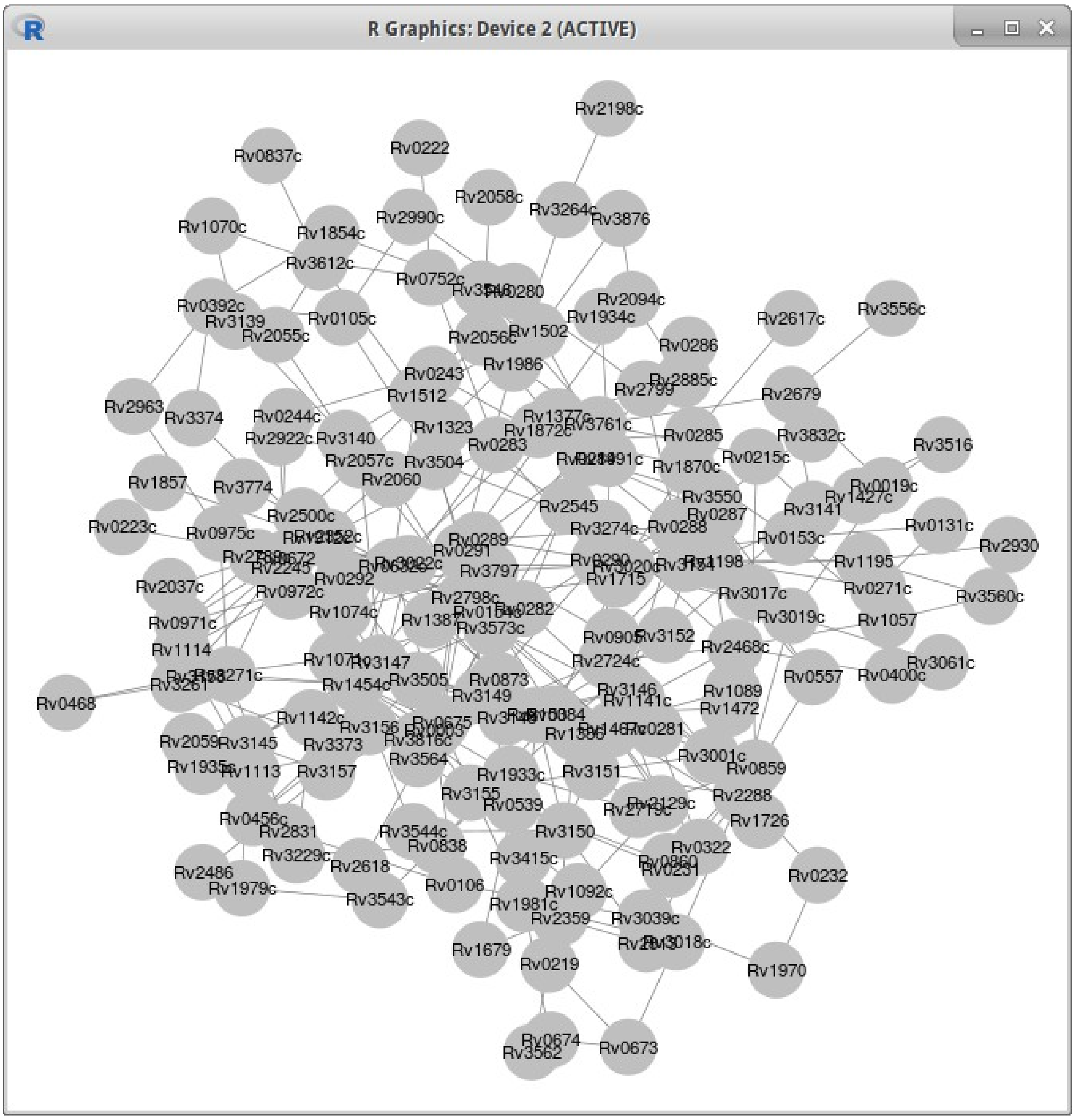
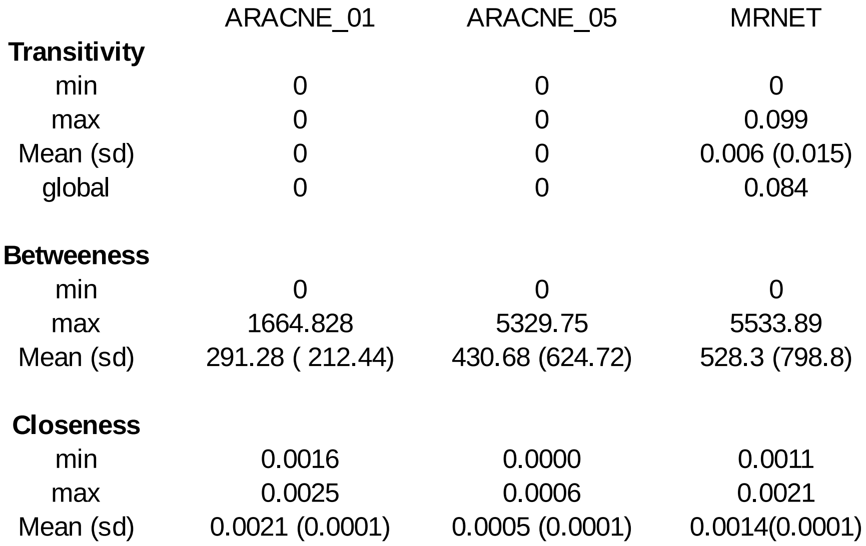


Illustration 13: Network MRNET 0.1, visuealized with R ggnet

For MRNET the transitivity is 0.084, whereas for ARACNE 0.1 and 0.5, the transitivity was 0, meaning that there are not triangular connections in the dataset. The low values of transitivity imply that genes within a cluster may not be well connected between them, suggesting that there is variation within clusters.

The transitivity, betweeness and closeness of each node are provided in the .R script.



	ARACNE_01	ARACNE_05	MRNET
Transitivity			
min	0	0	0
max	0	0	0.099
Mean (sd)	0	0	0.006 (0.015)
global	0	0	0.084
Betweeness			
min	0	0	0
max	1664.828	5329.75	5533.89
Mean (sd)	291.28 (212.44)	430.68 (624.72)	528.3 (798.8)
Closeness			
min	0.0016	0.0000	0.0011
max	0.0025	0.0006	0.0021
Mean (sd)	0.0021 (0.0001)	0.0005 (0.0001)	0.0014(0.0001)

Illustration 14: Transitiveness, betweeness and closeness of the network

As expected, for a fixed number of nodes (175), the closeness is larger in networks with more edges, because it is more easy to go from each one of the nodes to another. Similarly, it was also expected that the betweeness would be larger for networks with less number of edges because some of the shortest paths are no longer available. We did not observe any difference between MRNET and ARACNE (i.e ARACNE with 332 edges had a betweness of ~5330, whereas MRNET with 319 had betweeness equal to ~5534).

PART III. Gene communities and gene-ontogy enrichment

With the package Lyncomm it is possible to extract link communities using the “single” hierarchical clustering method. Communities are clusters in which each node may belong to more than one group. If the features used to study the genes are adequate, genes within the same communities are expected to be involved in related functions.

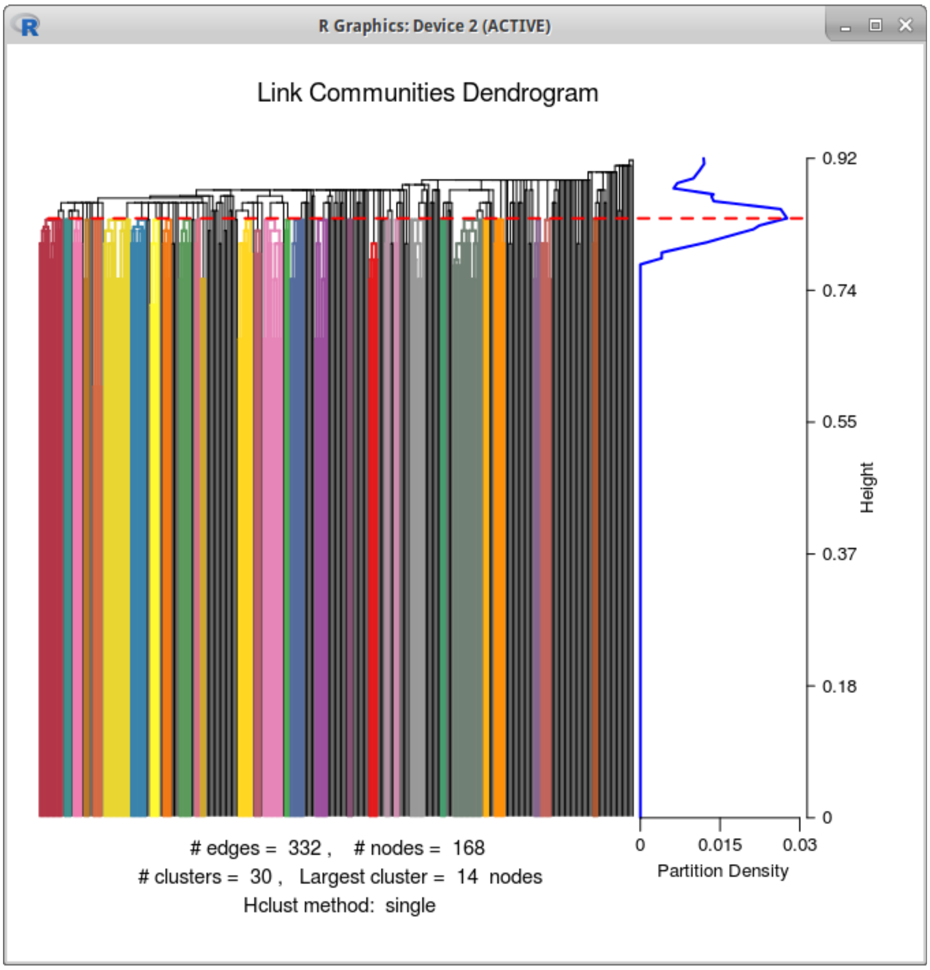

R Graphics: Device 2 (ACTIVE)
Link Communities Dendrogram
0.92
0.74
0.55
0.37
0.18
0
Height
0
0.015
0.03
Partition Density
# edges = 332 , # nodes = 168
# clusters = 30 , Largest cluster = 14 nodes
Hclust method: single

Illustration 15: Gene communities ARACNE_05

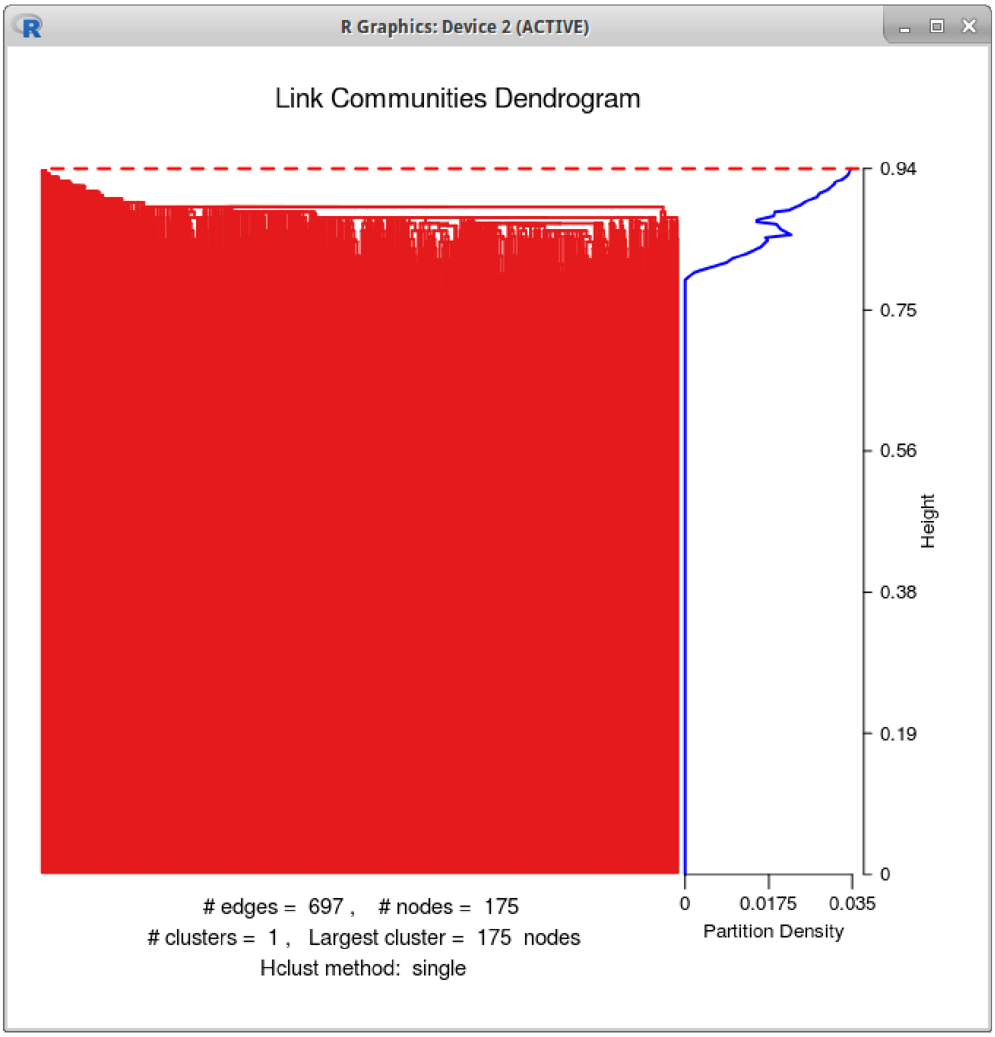

R Graphics: Device 2 (ACTIVE)
Link Communities Dendrogram
0.94
0.75
0.56
0.38
0.19
0
Height
0
0.0175
0.035
Partition Density
# edges = 697 , # nodes = 175
# clusters = 1 , Largest cluster = 175 nodes
Hclust method: single

Illustration 16: Gene communities ARACNE_01

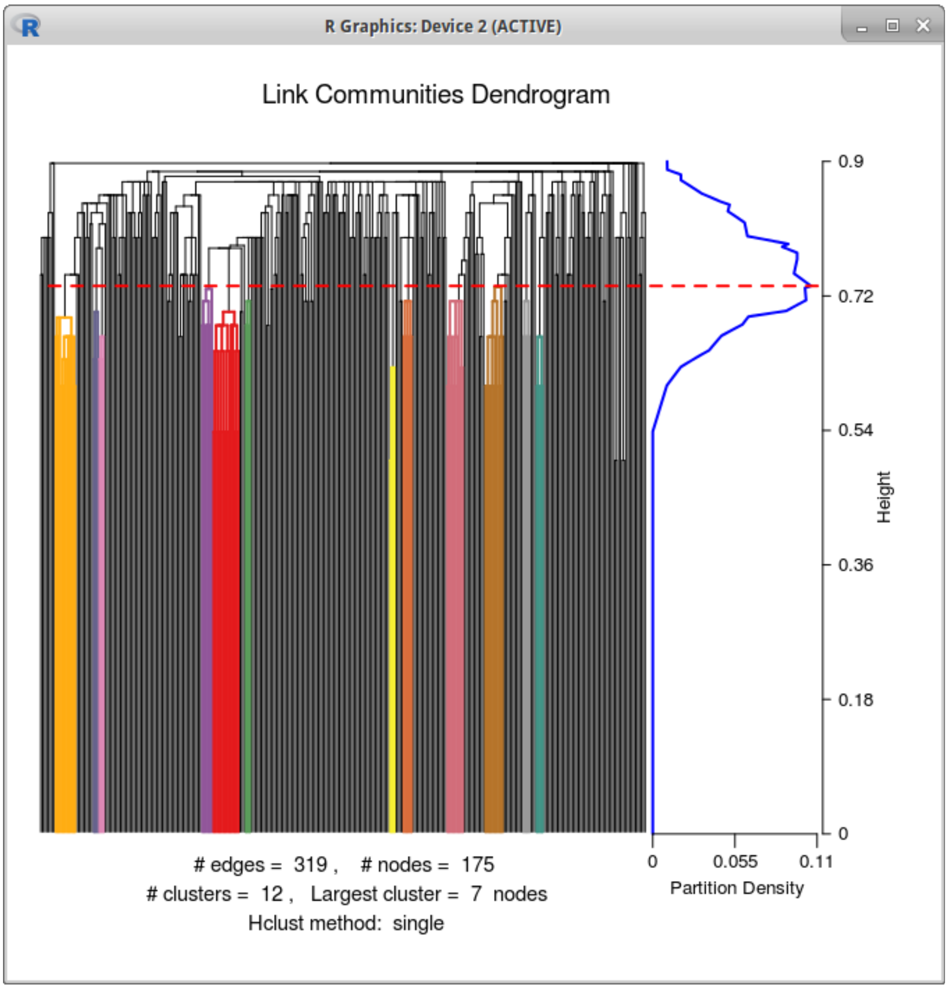


Illustration 17: Gene communities MRNET

When the threshod of expression was high, more communities were defined (ARACNE_05: 30 communities vs ARACNE_01:1 community). Also, for the same threshold, the method MRNET allowed to define more communities. Because we expect that there should be several communities, we will use the method MRNET from now on. We will use MRNET instead of ARCANCE_05 because a threshold of 0.5 is too strict and, as we saw in Ilustration 12, the network is disconnected.

The partition density plots on the right indicate at which point in the assignation of clusters to the communities, the number of links was highest for the overall communities. In other words, the partition density is 0 when none of the genes corresponds to more than one community and is maximum when the proportion of genes that belong to several classes is highest.

The object “lc$nodeclusters” in the .R script indicates to which cluster the genes have been assigned. We can create substes of genes based on this and see whether this subsets are enriched in any particular GO term. The analysis could be improved further by measuring the expression of other genes that are known to be associated with such GO terms. Then we could calculate the correlation of expression pattern between those genes and the genes of the gene community of interest.

The following illustrations show how the genes are grouped in communities for the MRNET method.

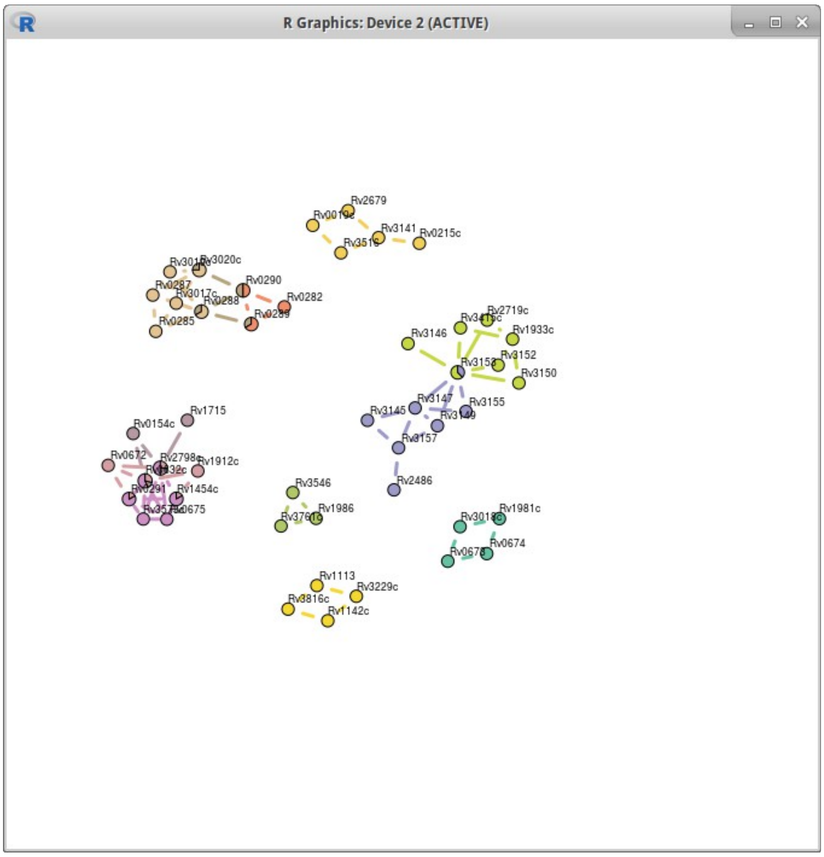


Illustration 18: gene communities. MRNET. Fruchterman.reingold

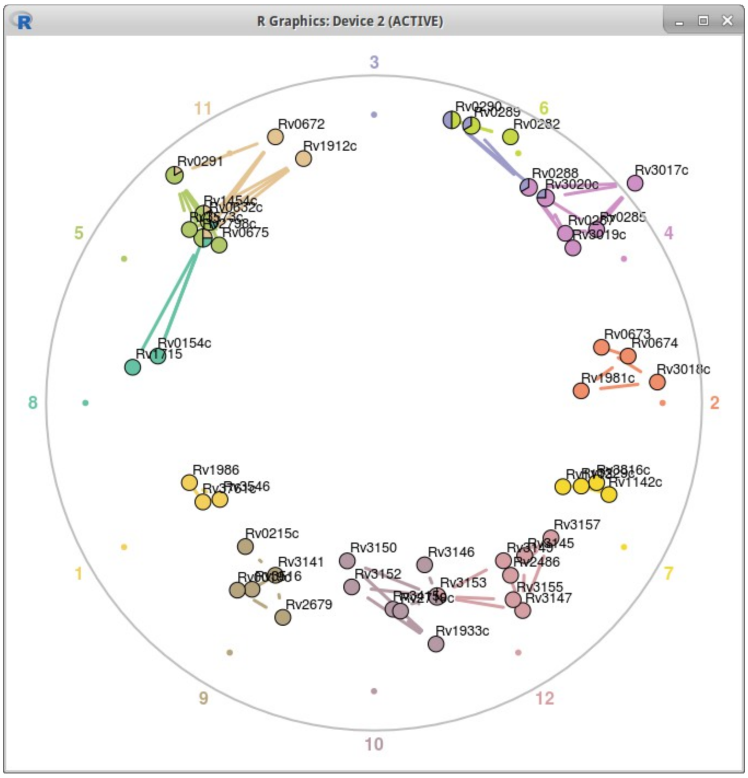


Illustration 19: MRNET_spencer_circle

We can also see how many of the genes correspond to at least, for instance, 3 communities. Only RV0632c and RV2798c (whose name appear in Illustration 20).

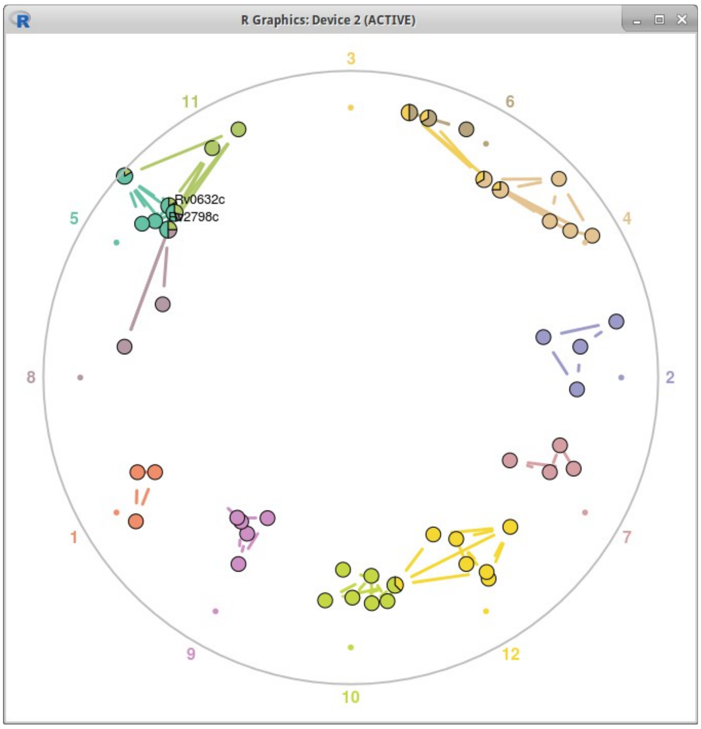


Illustration 20: MRNET_spencer_minimum_3_communities

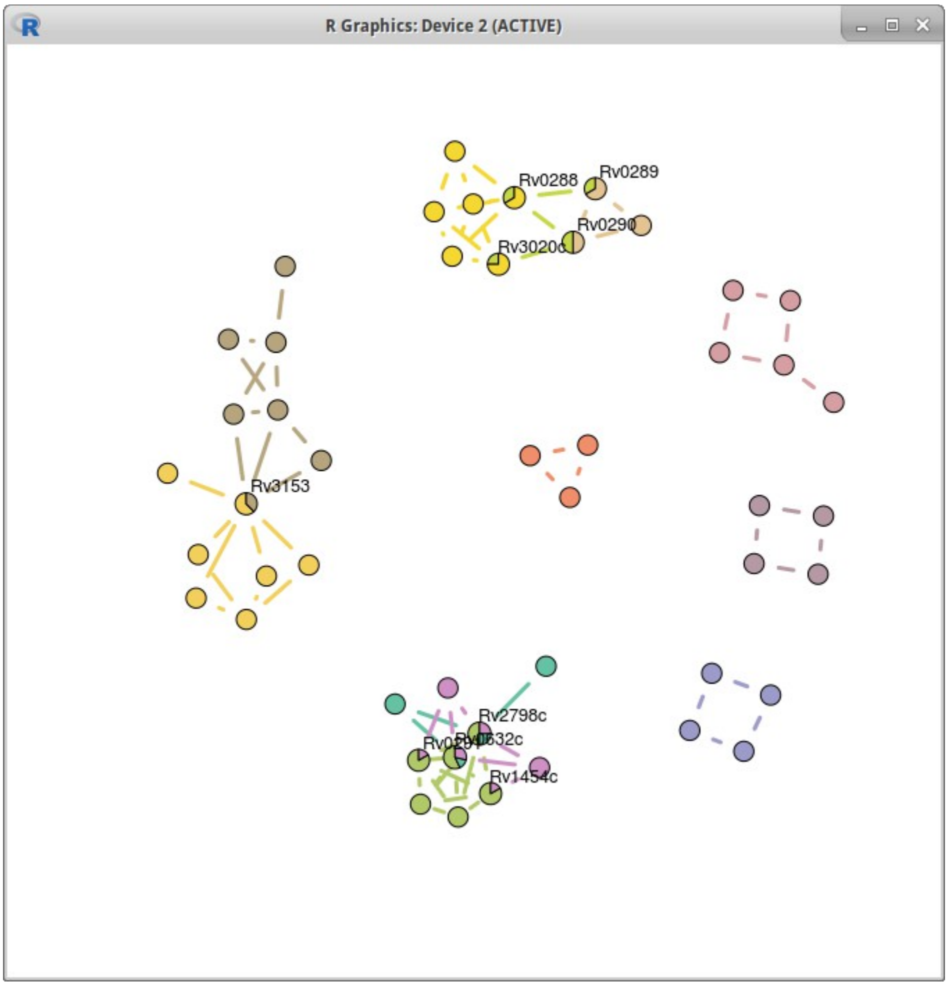
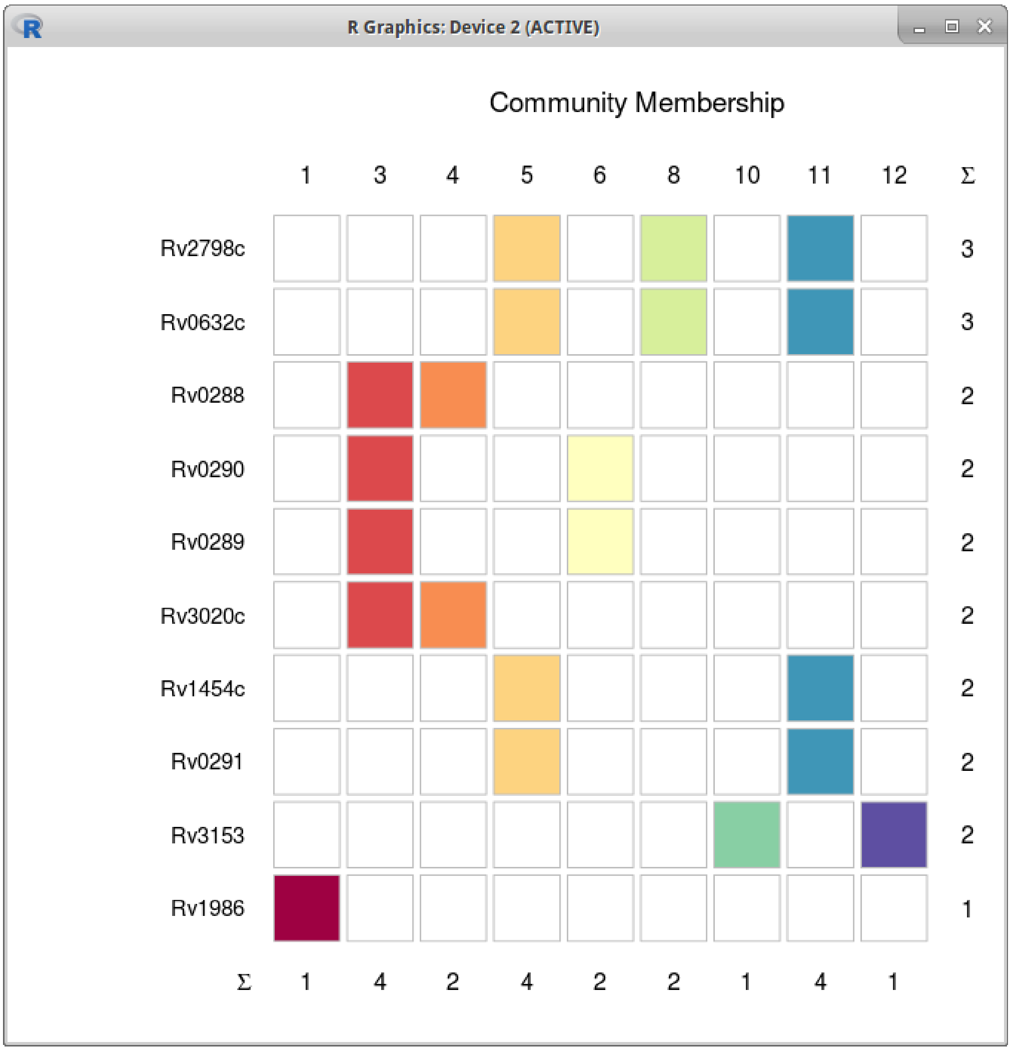


Illustration 21: MRNET_interactions_between_gene_communities



Community Membership

	1	3	4	5	6	8	10	11	12	Σ
Rv2798c				■		■		■		3
Rv0632c				■		■		■		3
Rv0288		■	■							2
Rv0290		■			■					2
Rv0289		■			■					2
Rv3020c		■	■							2
Rv1454c				■				■		2
Rv0291				■				■		2
Rv3153							■		■	2
Rv1986	■									1
Σ	1	4	2	4	2	2	1	4	1	

Illustration 22: Membership of the 10 genes that belong to a larger number of communities

Ilustration 21 allows to see the realtionship between the communities and Ilustration 22 the membership, from the genes that were assigned to a larger number of genes (above) to genes that were assigned only to one (below appreas one of them).

We check GO associations for the genes of the same community and store those assoications that appear at least for two genes within the same community. We get:

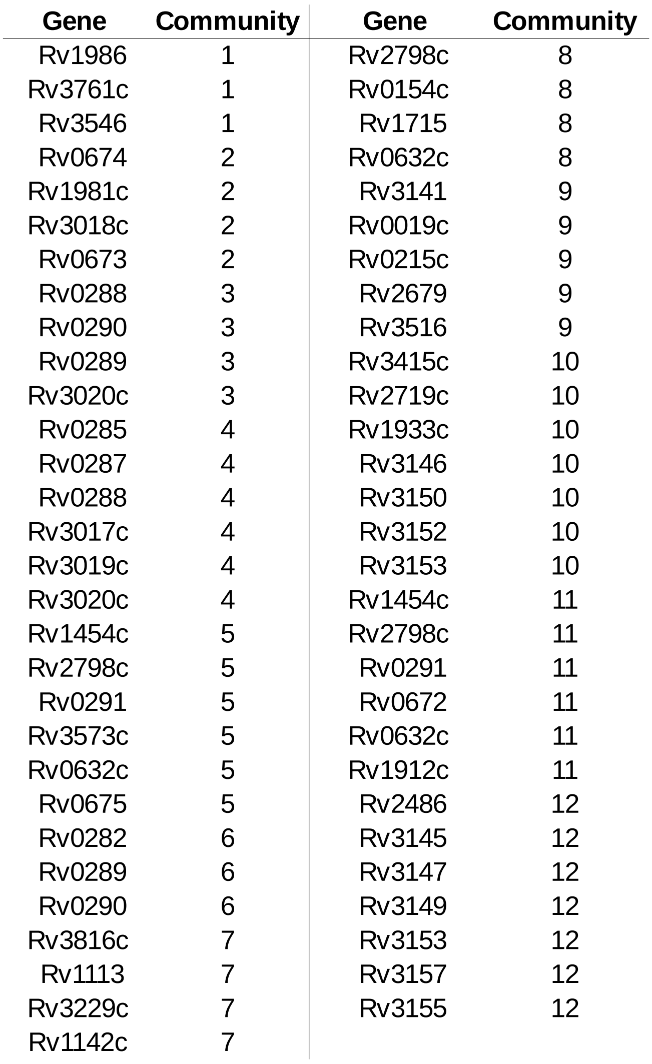
community 2:0055114
community 5:0055114
community 8:0008152
community 9:0055114
community 10:0016998
community 11:0055114
community 12:0055114
community 5:0008152
community 6:0040007
community 8:0055114
community 9:0052572
community 10:0055114
community 12:0006810
community 5:0040007
community 8:0006631
community 9:0008152
community 10:0006974
community 11:0008152
community 12:0042773

It seems that the genes in the dataset are involved in "oxidation-reduction process" (0055114)

(Communities 2, 5, 10-12). Another function in which the genes are involved is “metabolic process”, although this is a general GO term, and therefore the information is less useful.
Other Go terms of interest are for instance, “response to host immune response” in the genes of community 9 or “ATP synthesis coupled electron transport” in community 12. The links between communities in Inlustration 21, as well as the interactions between genes (in the last part of pre-processing) may allow us to know whether, for instance, “ATP synthesis coupled electron transport” could be relevant for the response to host immune response. Although this would require a more rigorous study”. Other aspects to be considered, are the number of genes that have a function within the cluster with respect to the total number of genes that have such function outside the community. Also, the number of edges between the genes of a community is relevant to associate groups of genes to GO terms.

Another analysis that could be done is to check in which experimental conditions the genes of a community that is associated with a GO term, increased (or decreased) more, and whether there is a biological explanation for such increase (or decrease).

The assignation of genes into communities is as follows:



Gene	Community	Gene	Community
Rv1986	1	Rv2798c	8
Rv3761c	1	Rv0154c	8
Rv3546	1	Rv1715	8
Rv0674	2	Rv0632c	8
Rv1981c	2	Rv3141	9
Rv3018c	2	Rv0019c	9
Rv0673	2	Rv0215c	9
Rv0288	3	Rv2679	9
Rv0290	3	Rv3516	9
Rv0289	3	Rv3415c	10
Rv3020c	3	Rv2719c	10
Rv0285	4	Rv1933c	10
Rv0287	4	Rv3146	10
Rv0288	4	Rv3150	10
Rv3017c	4	Rv3152	10
Rv3019c	4	Rv3153	10
Rv3020c	4	Rv1454c	11
Rv1454c	5	Rv2798c	11
Rv2798c	5	Rv0291	11
Rv0291	5	Rv0672	11
Rv3573c	5	Rv0632c	11
Rv0632c	5	Rv1912c	11
Rv0675	5	Rv2486	12
Rv0282	6	Rv3145	12
Rv0289	6	Rv3147	12
Rv0290	6	Rv3149	12
Rv3816c	7	Rv3153	12
Rv1113	7	Rv3157	12
Rv3229c	7	Rv3155	12
Rv1142c	7		

Table 4: Assignation of genes to communities