Babraham Bioinformatics

Exercises: Quantitative Gene Set Analysis

Version 2019-02

Licence

This manual is © 2018-19, Simon Andrews.

This manual is distributed under the creative commons Attribution-Non-Commercial-Share Alike 2.0 licence. This means that you are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

- Attribution. You must give the original author credit.
- Non-Commercial. You may not use this work for commercial purposes.
- Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under a licence identical to this one.

Please note that:

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.
- Nothing in this license impairs or restricts the author's moral rights.

Full details of this licence can be found at
http://creativecommons.org/licenses/by-nc-sa/2.0/uk/legalcode

Introduction

In this exercise we will look at how we can perform a quantitative gene set analysis where we don't need to define which genes we consider to be interesting or significant.

Software

The software which will be used in this session is listed below. In this case we are starting with data which has already been mapped and loaded into a SeqMonk project so we're only looking at the software we're using for visualisation and statistical analysis:

- SeqMonk (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/)

Data

The mouse RNA-Seq data used here are selected RNA-Seq samples from GEO accession GSE48364. The gene set data file is taken from Pathway Commons (Dec 2014 release).

The data file is "quantitative_gene_set_data.smk". The gene set file is "Mouse_GO_AllPathways.gmt.txt".

Exercise 1: Loading the expression data

The data for this exercise has already been loaded into a SeqMonk project, and has been quantitated as log2RPM values. You can start by loading the "`quantitative_gene_set_data.smk`" file into SeqMonk.

To get an idea of what sort of differences we're looking at you can do a scatterplot between the two replicate sets in the project (ESC and iPS).

`Plots > Scatter Plot`

Have a look at the general level of difference between the samples, and check that the normalisation looks OK.

Exercise 2: Identifying interesting Gene Sets

We are now going to try to find gene sets from a Pathway Commons data source which look like they are changing in an interesting way between the two replicate sets. Rather than an analysis at a single gene level we will be analysing whole sets of genes, which should increase the amount of statistical power we have, and will allow us to see things we might miss when looking at individual genes.

We will do this analysis twice using different statistical tests, the Kolmogorov-Smirnov test, and the Student's T-Test. Make sure you understand the conceptual difference between the types of changes we are hoping to identify with these two tests.

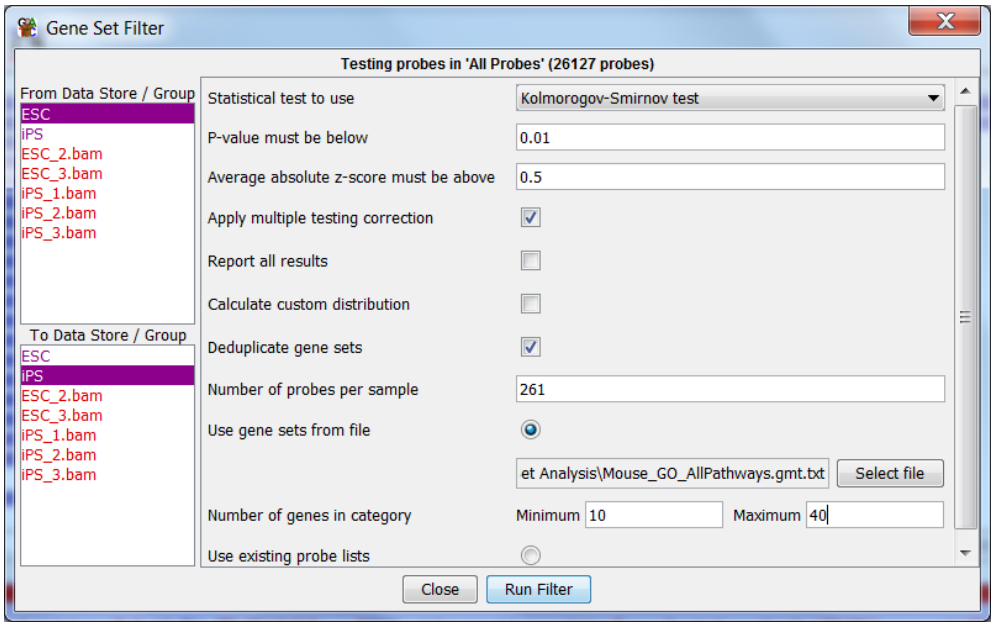
Step 2.1 – Analysis using a KS test

You can launch the options for the Gene Set test using:

`Filtering > Filter by Statistical Test > Subgroup statistics > Gene Set Enrichment`

There are a number of things you are going to need to select to perform the test.

1. Select your ESC replicate set in the "From Stores" and your iPS replicate set in the "To Stores"
2. Change the p-value we're looking for down to 0.01 to restrict ourselves to more significant events.
3. Set the z-score (the minimum amount of absolute change) to 0.5
4. Change the number of genes in a category to be between 10 and 40. This will exclude small categories where you are unlikely to hit statistical significance, and large categories where you are unlikely to get a concerted change (and where you may be over-powered)
5. Finally, press the "Select File" button to choose the Pathway Commons file which has the definitions of the gene sets in it.



Once you are happy that all of the correct options have been set, press “Run Filter”.

You should see a summary of the gene sets which were taken forward for analysis. What proportion of the sets were rejected? If you have time you could plot out the distribution of gene set sizes in the file you can see where what number of genes the majority of sets have.

Once the test has run you should see the table and graph of the results. Use the button below the plot to switch between a standard scatterplot, and the z-score plot used for the test. Can you see the difference between them and understand how this transformation is achieved?

Select some of the hits in the hit table and see if you can see why they have been selected. Try sorting the table by p-value and mean z-score and look at the top (and bottom for the z-score) hits for each metric. Can you see the relationship between the size of the difference and the number of points for each gene set?

Look at how far the points in each set fall from the centre of the z-score distribution. Are they consistent or not? How does this fit with what the KS test is testing?

Save the hits as individual probe lists. Press the “`Select All`” button, followed by “`Save Selected Probe Lists`”. Add ‘`KS`’ to the name of the saved lists so you can tell where it came from easily.

Step 2.2 – Analysis using a T-test

Repeat the gene set analysis, but change the test type to be a t-test. Look at the results and see if you can see any differences in the nature of the changes found in the most significant hits.

When you’re looking through the hits see if you can get an impression whether some of the lists contain largely the same sets of genes (they won’t be exactly the same as removing duplicates was one of the pre-filtration steps).

After reviewing the hits, again save them as separate probe lists.

Exercise 3: Clustering the hit sets

Since we have a large number of different categories which were hit, and since they appear to have substantial overlaps between them we're going to see if we can rationalise the hits to get a clearer picture of what's really been found in this comparison.

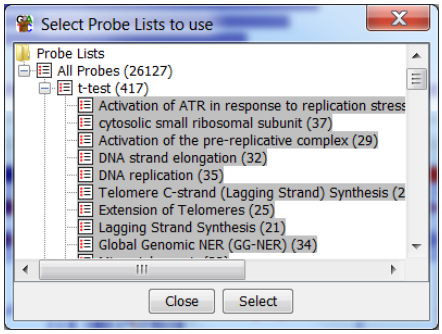
For this analysis we'll just take the hits from the T-Test (we could have used either). We will do a graphical clustering of these, using the degree of overlap in the genes in common between lists to determine how closely positioned different groups should be.

The plot we are going to draw is currently limited to 50 groups. With the filters we applied before you should have fewer groups than this in your list of hits, but if you ended up with more than that you'll need to randomly a pick a subset of them to be able to draw the plot.

Start by opening the plot options.

`Plots > Probe List Overlap > Probe List Overlap Plot`

You will see an expanded list of all of the probe lists in your project. You need to select all of the lists under your T-Test results group. You can do this by selecting the top group in the list, scrolling down to the bottom then pressing shift and clicking on the last group:



Press `Select` to draw the plot.

What you will see is a 2D plot of the probe lists which are positioned based on their similarity. The plot is started at a random position and is run through an interactive improvement to get to the final position. You can press the "`Go`" button a few times to re-run the simulation to see how robust the clustering is.

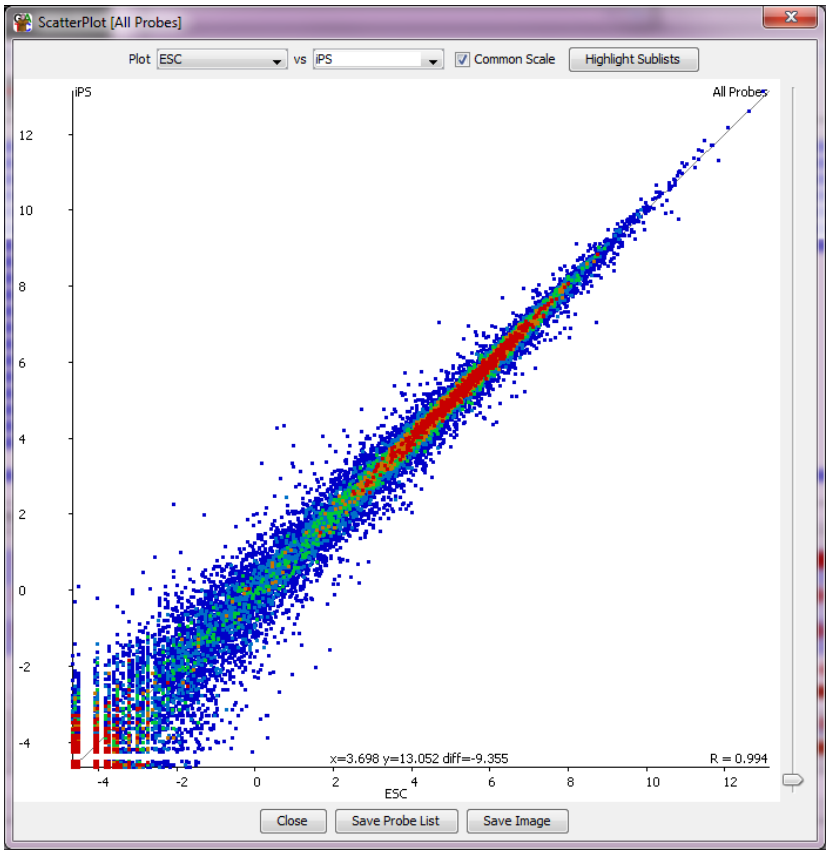
Have a look at the plot and try to answer the questions:

1. What proportion of the lists are singletons (don't relate to other groups)?
2. How many major connected groups are there?
3. What are the main biological themes of the groups?
4. Do you think these themes point to interesting underlying biology, or a potential artefact in the experiment?

Example Plots

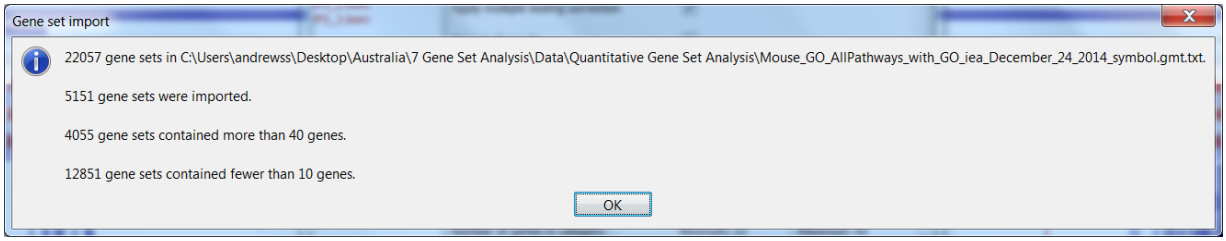
So you know what you should be seeing here are copies of the plots you should generate in this practical:

Exercise 1

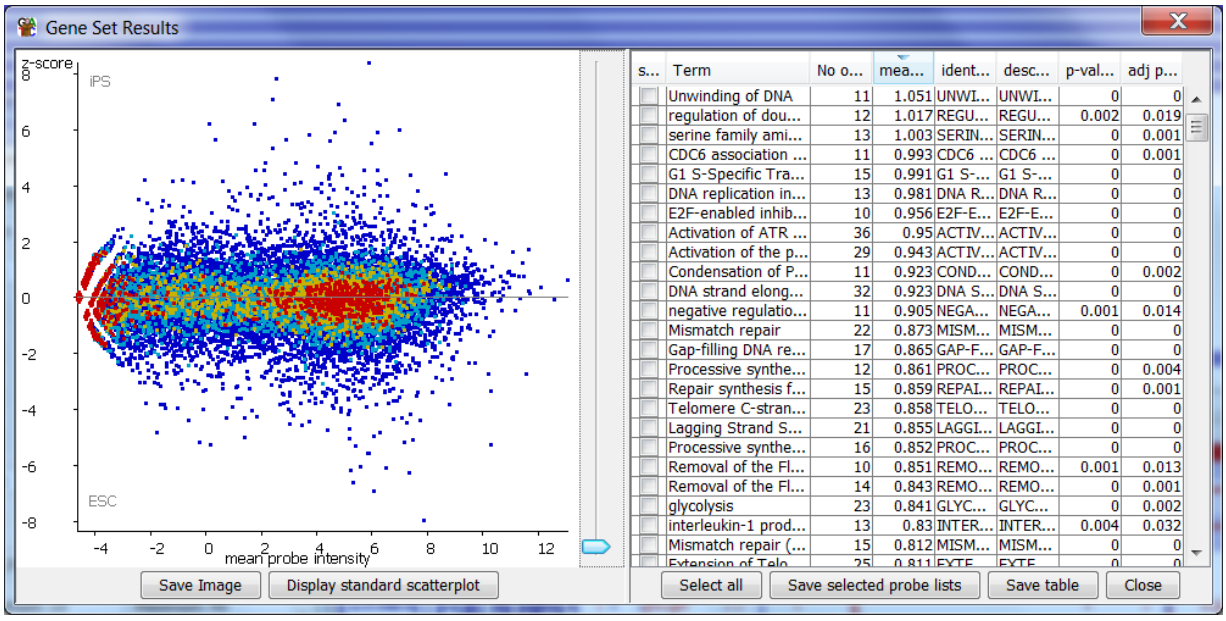


The two samples look pretty similar to each other overall, with some outliers. The data look well normalised and sit nicely on the diagonal.

Exercise 2



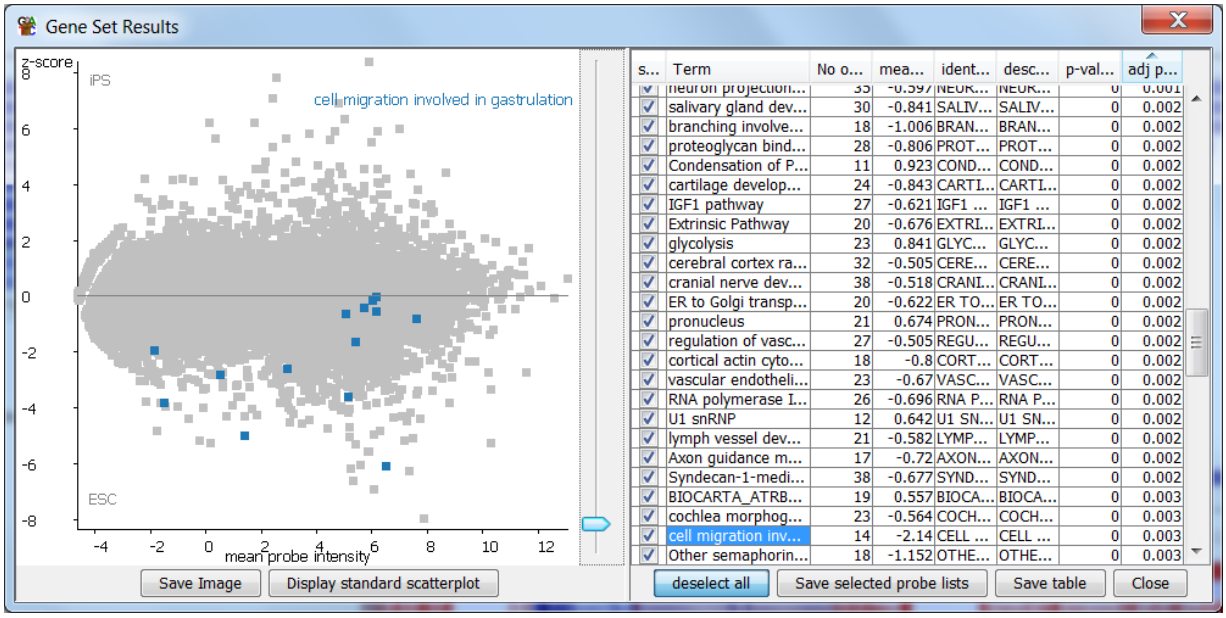
We actually opt not to analyse most of the lists. There are a very large number of categories with a very small number of genes in them. This is due to the fragmentary nature of the Gene Ontology, but the hierarchical nature of the structure means that the same genes will still be captured at one of the higher levels.



s...	Term	No o...	mea...	ident...	desc...	p-val...	adj p...
	Unwinding of DNA	11	1.051	UNWI...	UNWI...	0	0
	regulation of dou...	12	1.017	REGU...	REGU...	0.002	0.019
	serine family ami...	13	1.003	SERIN...	SERIN...	0	0.001
	CDC6 association ...	11	0.993	CDC6 ...	CDC6 ...	0	0.001
	G1 S-Specific Tra...	15	0.991	G1 S-...	G1 S-...	0	0
	DNA replication in...	13	0.981	DNA R...	DNA R...	0	0
	E2F-enabled inhib...	10	0.956	E2F-E...	E2F-E...	0	0
	Activation of ATR ...	36	0.95	ACTIV...	ACTIV...	0	0
	Activation of the p...	29	0.943	ACTIV...	ACTIV...	0	0
	Condensation of P...	11	0.923	COND...	COND...	0	0.002
	DNA strand elong...	32	0.923	DNA S...	DNA S...	0	0
	negative regulatio...	11	0.905	NEGA...	NEGA...	0.001	0.014
	Mismatch repair	22	0.873	MISM...	MISM...	0	0
	Gap-filling DNA re...	17	0.865	GAP-F...	GAP-F...	0	0
	Processive synthe...	12	0.861	PROC...	PROC...	0	0.004
	Repair synthesis f...	15	0.859	REPAI...	REPAI...	0	0.001
	Telomere C-stran...	23	0.858	TELO...	TELO...	0	0
	Lagging Strand S...	21	0.855	LAGGI...	LAGGI...	0	0
	Processive synthe...	16	0.852	PROC...	PROC...	0	0
	Removal of the Fl...	10	0.851	REMO...	REMO...	0.001	0.013
	Removal of the Fl...	14	0.843	REMO...	REMO...	0	0.001
	glycolysis	23	0.841	GLYC...	GLYC...	0	0.002
	interleukin-1 prod...	13	0.83	INTER...	INTER...	0.004	0.032
	Mismatch repair (...	15	0.812	MISM...	MISM...	0	0

You should see that the z-score normalised scatterplot looks much more evenly spread from the central line. The normalisation tries to equalise the magnitude of change at all points in the expression range, so that we don't end up biasing towards highly or lowly expressed genes.

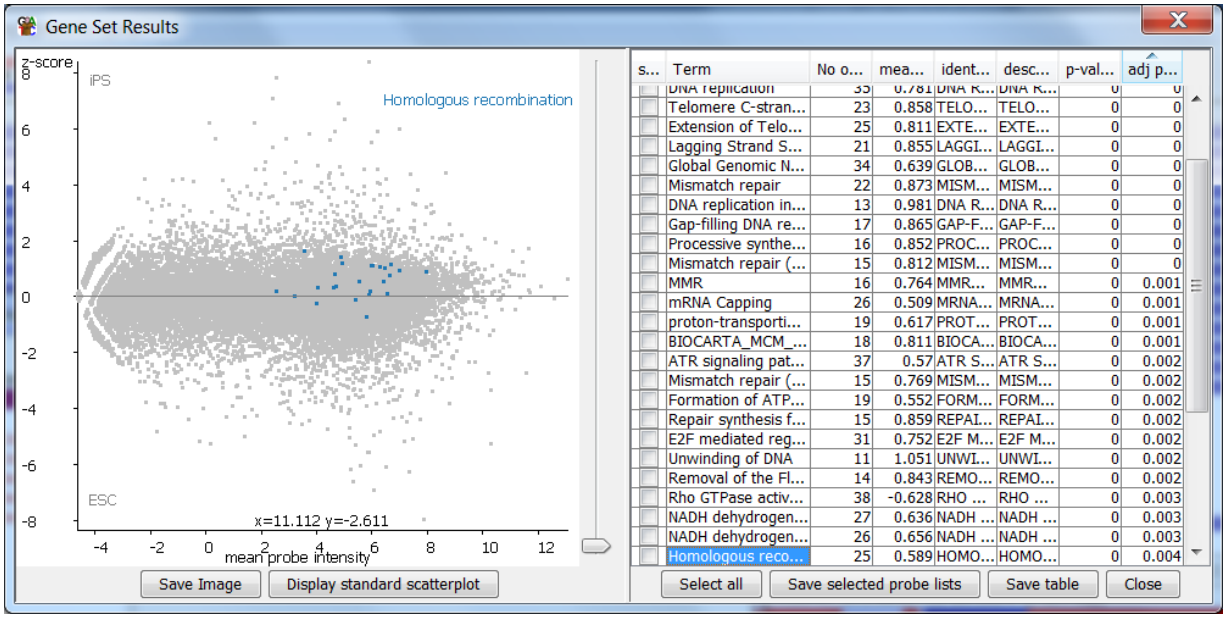
Step 2.1



s...	Term	No o...	mea...	ident...	desc...	p-val...	adj p...
✓	salivary gland dev...	30	-0.841	SALIV...	SALIV...	0	0.002
✓	branching involve...	18	-1.006	BRAN...	BRAN...	0	0.002
✓	proteoglycan bind...	28	-0.806	PROT...	PROT...	0	0.002
✓	Condensation of P...	11	0.923	COND...	COND...	0	0.002
✓	cartilage develop...	24	-0.843	CARTI...	CARTI...	0	0.002
✓	IGF1 pathway	27	-0.621	IGF1 ...	IGF1 ...	0	0.002
✓	Extrinsic Pathway	20	-0.676	EXTRI...	EXTRI...	0	0.002
✓	glycolysis	23	0.841	GLYC...	GLYC...	0	0.002
✓	cerebral cortex ra...	32	-0.505	CERE...	CERE...	0	0.002
✓	cranial nerve dev...	38	-0.518	CRANI...	CRANI...	0	0.002
✓	ER to Golgi transp...	20	-0.622	ER TO...	ER TO...	0	0.002
✓	pronucleus	21	0.674	PRON...	PRON...	0	0.002
✓	regulation of vasc...	27	-0.505	REGU...	REGU...	0	0.002
✓	cortical actin cyto...	18	-0.8	CORT...	CORT...	0	0.002
✓	vascular endotheli...	23	-0.67	VASC...	VASC...	0	0.002
✓	RNA polymerase I...	26	-0.696	RNA P...	RNA P...	0	0.002
✓	U1 snRNP	12	0.642	U1 SN...	U1 SN...	0	0.002
✓	lymph vessel dev...	21	-0.582	LYMP...	LYMP...	0	0.002
✓	Axon guidance m...	17	-0.72	AXON...	AXON...	0	0.002
✓	Syndecan-1-medi...	38	-0.677	SYND...	SYND...	0	0.002
✓	BIOCARTA_ATRB...	19	0.557	BIOCA...	BIOCA...	0	0.003
✓	cochlea morphog...	23	-0.564	COCH...	COCH...	0	0.003
✓	cell migration inv...	14	-2.14	CELL ...	CELL ...	0	0.003
✓	Other semaphorin...	18	-1.152	OTHE...	OTHE...	0	0.003

From the KS test it doesn't matter if we have similar deviations in expression – just that the overall set doesn't sit on the null centre line. Groups like the one above are highly significant even though some of the points are very different to others.

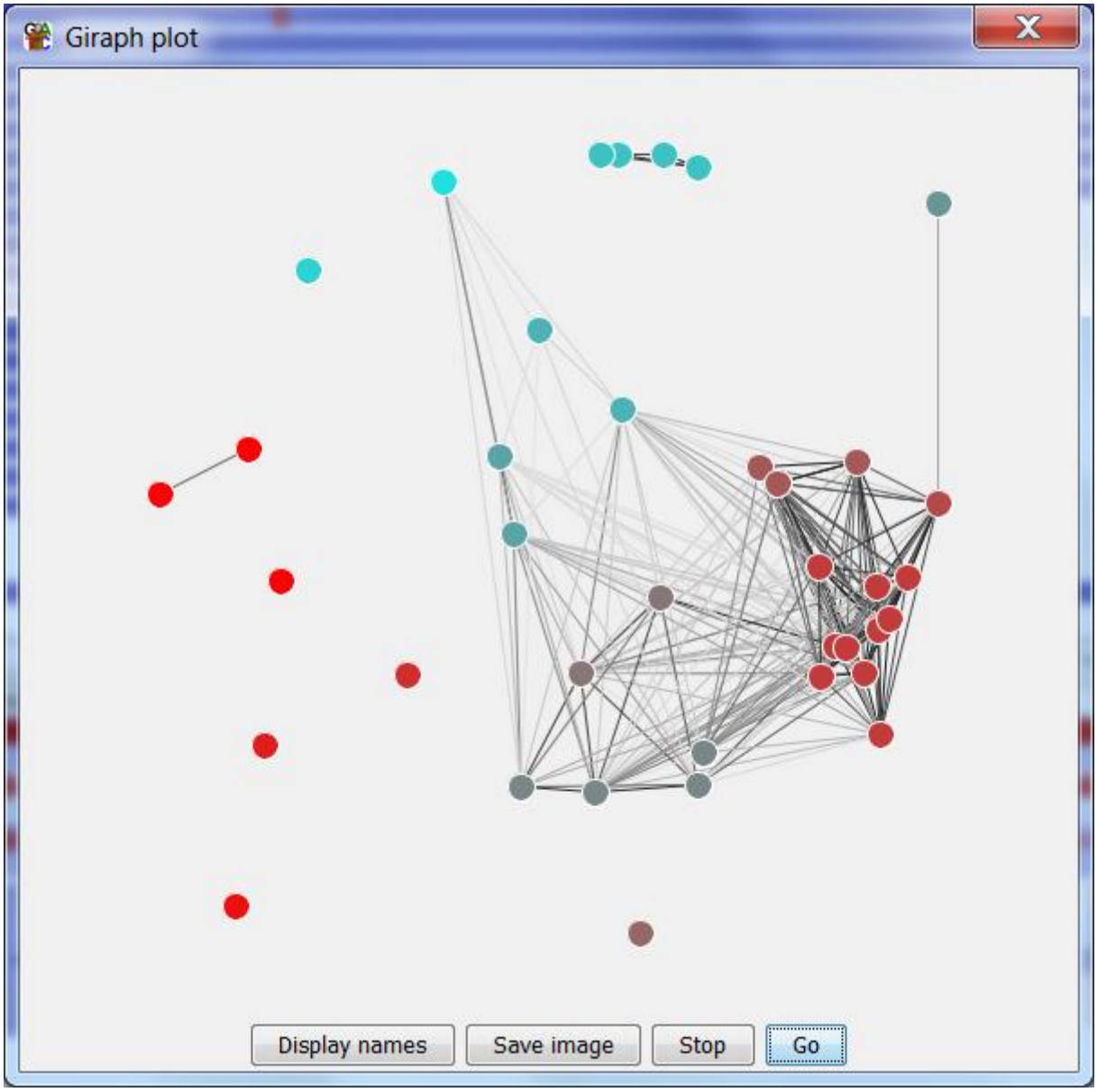
Step 2.2



s...	Term	No o...	mea...	ident...	desc...	p-val...	adj p...
	DNA replication	33	0.781	DNA R...	DNA R...	0	0
	Telomere C-stran...	23	0.858	TELO...	TELO...	0	0
	Extension of Telo...	25	0.811	EXTE...	EXTE...	0	0
	Lagging Strand S...	21	0.855	LAGGI...	LAGGI...	0	0
	Global Genomic N...	34	0.639	GLOB...	GLOB...	0	0
	Mismatch repair	22	0.873	MISM...	MISM...	0	0
	DNA replication in...	13	0.981	DNA R...	DNA R...	0	0
	Gap-filling DNA re...	17	0.865	GAP-F...	GAP-F...	0	0
	Processive synthe...	16	0.852	PROC...	PROC...	0	0
	Mismatch repair (...	15	0.812	MISM...	MISM...	0	0
	MMR	16	0.764	MMR...	MMR...	0	0.001
	mRNA Capping	26	0.509	MRNA...	MRNA...	0	0.001
	proton-transporti...	19	0.617	PROT...	PROT...	0	0.001
	BIOCARTA_MCM_...	18	0.811	BIOCA...	BIOCA...	0	0.001
	ATR signaling pat...	37	0.57	ATR S...	ATR S...	0	0.002
	Mismatch repair (...	15	0.769	MISM...	MISM...	0	0.002
	Formation of ATP...	19	0.552	FORM...	FORM...	0	0.002
	Repair synthesis f...	15	0.859	REPAI...	REPAI...	0	0.002
	E2F mediated reg...	31	0.752	E2F M...	E2F M...	0	0.002
	Unwinding of DNA	11	1.051	UNWI...	UNWI...	0	0.002
	Removal of the Fl...	14	0.843	REMO...	REMO...	0	0.002
	Rho GTPase activ...	38	-0.628	RHO ...	RHO ...	0	0.003
	NADH dehydrogen...	27	0.636	NADH ...	NADH ...	0	0.003
	NADH dehydrogen...	26	0.656	NADH ...	NADH ...	0	0.003
	Homologous reco...	25	0.589	HOMO...	HOMO...	0	0.004

In the T-Test there is expected to be a general level of homogeneity between the points in a group. Even where (as in the case above) the points are spread around the null line we can see that the overall noise is reasonably consistent, and that the mean would be above the line.

Exercise 3



This is the view of the gene set clustering. You can see that there is one very major connected group, with 3 separate subgroups within it. The major biological theme of the group would appear to be that there is a difference in the rate of cell division between the samples, since all of the groups relate to this. The 3 subgroups are the genes relating to DNA replication, regulatory genes which initiate replication or DNA unwinding and more general genes relating to the G1 to S transition.

Separately there is a smaller group which seems to relate to mitochondrial activity, and a much smaller group related to ribosomal sequences.