

Exercises:

Advanced Analysis with SeqMonk

Version 2018-12

# Licence

This manual is © 2013-2018, 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>

# DataSets

The example datasets used as examples in this course are taken from the public sequence repositories. The data used were:

1. All samples from ArrayExpress E-MTAB-822 Transcription profiling by high thoughput sequencing of human cell lines Ishikawa, MCF7 and T47D treated with estrogen, progesterone and their antagonists

### Exercise 1: Custom Tracks and Grouping

* Open the ‘Large\_RNA\_Seq.smk’ project file containing 18 RNA-Seq samples. Create a custom mRNA track containing only protein coding genes on autosomal chromosomes (exclude X, Y and MT)
* Do a standard RNA-Seq quantitation using this custom track and merging transcript isoforms. Use the distribution plots to check the normalisation.
* Use a condition tree to see how your samples cluster and check that they match the names of the groups you can see.
* Create replicate sets from the Ish, T47D, MCF7-Tam and MCF7 sample groups using the automatic group creation tool.

### Exercise 2: Pairwise comparison

* Re-run the RNA-Seq quantitation to get raw counts for your data. To do this you should just need to tick the box which says “Generate raw counts” in the pipeline options.
* Use the DESeq2 filter to find genes changing between Ish and MCF7-Tam
* Requantitate your data with normalised log2 transformed counts (the default)
* Draw a scatterplot of Ish vs MCF7-Tam and highlight the DESeq hits to check they look OK.
* Use the Intensity Difference filter to find transcripts which are changing between any of your replicate sets. Select all combinations and it will run every possible pairwise comparison.

### Exercise 3: Multi-comparison and clustering

* Cluster your hits using hierarchical clustering and view the results
* Try viewing the clustered results as replicate sets and individual replicates
* Generate lists from clusters connected R>0.7 and draw a summary line graph for these groups

### Exercise 4: More clustering

* Find genes whose expression increases steadily from Ish untreated – E2-3h – E2-12h.
* Find genes whose expression decreases steadily from Ish untreated – E2-3h – E2-12h
* View the two sets of results.

### Exercise 5: Single Cell QC

* Load the single\_cell\_rna\_seq.smk file. This contains data from XXX individual cells
	+ Note that this is a multi-genome project with ERCC control sequences
* Look at the data. Since the tracks are small you can select an individual sample from the data view and the track with expand so you can see it more clearly.
	+ Are there samples which look like they might be problematic?
	+ Run a Data Store Summary report and sort by Total Read Count. How many bad cells are there likely to be?
* Draw an RNA-Seq QC report. This plots out several metrics for your cells. Look for outliers to try to indentify problematic cells. You can click and drag in the report to select cells and can then opt to flag them by adding “BADQC” to the start of their names.
	+ Select the cells with high ERCC content, or low data size and flag them as BADQC
* Go to Edit Data Sets and delete the data for the cells which you flagged as having bad QC.
* Run the RNA-Seq quantitation pipeline using default parameters (log2 transformation of normalised merged data, assuming a non-strand-specific library)
* Since we have so many datasets try changing the display preferences to showing only probes and turn the representation into blocks.
* Create a replicate set called “All Cells” which contains….. all cells.
* Draw a variance plot using the All Cells replicate set, based on STDEV
	+ Note how some genes are unusually highly variable. Take a look at some of the highest / lowest variability genes and check you can see what aspects of the data would make them appear like this.
* Run a variance filter (Filters > Filter on variance) and select probes with variance 0.5 STDEV higher than the local average. Highlight the selected list on the variance plot to check that you’ve done the right thing.
* Select the variable probe list and draw a Data Store Tree (under Plots > Data Store Similarity)
	+ Do you see two obvious large groups of cells?
	+ Can you use the slider to separate them by cutting the tree and using the “Split Data Stores” option?
* Change the view to show your two new replicate sets and change the display preferences so that they are expanded (so you can see the individual data sets again).
	+ Have a look at some of the hyper-variable genes. Do they seem to segregate between the two groups?
* Generate a TSNE plot from your hyper-variable genes across all cells. Do you see obvious separation? Highlight on the two groups you separated from the Data Store Tree. Do they separate by TSNE as well?
* You can now try to do a statistical test to find genes which differ significantly between the two groups you found. To visualise this you can first do a scatterplot to plot all genes (not just your hyper variable ones) between your two groups.
	+ Can you see outliers?
	+ Do they look consistent when you look at the underlying quantitation?
* Since we have so many samples we can use a simple t-test to compare the two groups
	+ Filtering > Filter by Statistical test > Continuous values > Replicated Data > T-test
	+ Use a corrected p-value cutoff of 0.05
	+ Highlight the hits on your scatterplot.
* From your selected hits generate
	+ A hierarchical cluster plot
	+ An annotated probe report
	+ A probe list description report.