

Advanced Analysis with SeqMonk Exercises

Licence

This manual is © 2013-2014, 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. The UHR_directional_Tn-RNA-seq sample (GSM800443) from GEO GSE32307. Taken from Gertz J, Varley KE, Davis NS, Baas BJ et al. Transposase mediated construction of RNA-seq libraries. Genome Res 2012 Jan;22(1):134-41. PMID: 22128135
2. All samples from ArrayExpress E-MTAB-822 Transcription profiling by high throughput sequencing of human cell lines Ishikawa, MCF7 and T47D treated with estrogen, progesterone and their antagonists

Exercise 1: Reimporting and Wiggling

- Open the directional RNA-Seq project file
- Select a small region with obvious variation in coverage and construct a wiggle plot over the region
- Use the smoothing quantitation to smooth out the wiggle plot
- Use the antisense transcription pipeline to identify novel antisense transcription. Review the results and see if you agree with its predictions.

Exercise 2: 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
- Normalise your data as you see fit
- Do a condition tree to see how to group your samples and create replicate sets
- Create replicate sets from the Ish, T47D, MCF7-Tam and MCF7 sample groups. You can use a mixture of automatic and manual group creation.

Exercise 3: Simulation

- Select the subset of transcripts which are annotated as being Calmodulin binding proteins (GO:0005546)
- Use a Monte-Carlo simulation to test whether these show higher than average expression in the T47D sample

Exercise 4: Pairwise comparison

- Use the Intensity Difference filter to find transcripts which are changing between the Ish and T47D groups
- Create a report sorted by significance and look at the top hits
- Re-run the intensity difference filter to find changes between any of your replicate sets.

Exercise 5: 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 6: 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.