Babraham Bioinformatics

Exercises: Advanced Analysis with SeqMonk

Version 1.4

Licence

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 thoughput 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: 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 the Ish and MCF7-Tam groups. Combine this with the DESeq hits to see which agree between the two methods.
- Finally, run the intensity difference filter to find changes between any of your replicate sets.

Exercise 4: 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 5: 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.