EMBnet tutorial on RNA-seq, Valencia 14.5.2013

Eija Korpelainen
chipster@csc.fi

1. Start Chipster and import fastq file (raw reads)

Go to http://chipster.csc.fi/, click on the link Launch Chipster v2.5. Select File / Import files, navigate to the Chipster folder and select the file h1-hESC_RNAseq.fastq. When the file appears in Chipster, select it, go the visualization panel and select method View text.

2. Quality control with FastQC and PRINSEQ tools

Select the file, the tool Quality control / Read quality with FastQC and click Run. Inspect the results.

-how many reads are there and how long are they?
-what quality encoding is used?
-is the base quality good all along the reads?

3. Trim sequences based on quality

Select h1-hESC_RNAseq.fastq and the tool Utilities / Trim reads for several criteria and set parameters to trim from 3' ends so that the mean quality in a 3 base window is at least 20. Inspect the log file and figure out how many reads were discarded. Select the file trimmed.fastq and run the tool Quality control / Read quality with PRINSEQ. Select the result file reads-stats.html and visualization method Open in external web browser.

-does the quality look better?
-how is the read length distribution now?

4. Filter out reads that are shorter than 50 bases.

Select trimmed.fastq and the tool Filtering / Filter reads for length. How many reads get discarded?

5. Align reads to reference genome using Bowtie2 (for the interest of time, normally you would use TopHat)

Select accepted.fastq and run the tool Alignment / Bowtie2 for single end reads. Right-click on the BAM file and rename it to hESC.bam. Rename the index file to hESC.bam.bai.

-what was the overall alignment rate?
-how many alignments are there and how many of them have mapping quality higher than 1 (use the tool Utilities / Count alignments in BAM)

6. Count reads per genes using HTSeq

Select the BAM file and run RNA-seq / Map aligned reads to genes with HTSeq so that the minimum alignment quality is 1. Select htseq-counts.tsv and run Filtering / Filter table by column value

-how many genes have read counts?

7. Save session, get analysis history file, save and run workflow

Select File / Save session. Get a textual report on what you have done: Select filtered-NGS-result.tsv and click on the small paper icon in the workflow panel. Save also an automatic workflow: Select file h1-hESC_RNAseq.fastq and Workflow / Save starting from selected. Import file GM12878_RNAseq.fastq, drag it under the other files, and select Workflow / Run recent / your workflow.

8. Create count table and description file for the experiment

Select both htseq-counts.tsv files and run the tool Utilities / Define NGS experiment so that you set the count column and genomic coordinates parameters correctly. Fill in the group column in the phenodata file: enter 1 for h1-hESC and 2 for GM12878. Save the session.

We will return to this dataset, but for a while we move to a Drosophila dataset which is more interesting for differential expression analysis.

9. Open new session

Select File / Open session and the file pasilla.zip. Inspect the phenodata. This is a two-group comparison with 7 samples. Some samples were sequenced single end, some paired end.

10. Analyze differential expression with edgeR classic

Select the file counts.tsv and run the tool RNA-seq / Differential expression using edgeR.

-do the samples separate according to group or something else in the MDS plot?
-how many differentially expressed genes do you get?
-is common dispersion a good approximation of genewise dispersions, judged by the dispersion plot?
-how big fold change is required for DE genes, judged by the MA plot?
-how does the p-value distribution look like in the p-value plot?

Try filtering: Repeat the run so that you analyze only genes which are expressed in at least 3 samples.

-how does the number of DE genes change and why?
-did the p-value distribution change?
-did the MDS plot change?

11. Analyze differential expression with DESeq

Delete the results from the filtered run. Select the file counts.tsv and run RNA-seq / Differential expression using DESeq. Repeat the run so that you use fitted dispersion values always.

-which parameter setting finds more DE genes? Why?

12. Compare the result lists from DESeq and classic edger using Venn diagram

Select the files de-list-deseq.tsv and de-list-edger.tsv by keeping the ctrl key down. In the visualization panel select method Venn-diagram. Select genes found by both methods and create a new dataset out of them.

13. Analyze differential expression with edger glm

Delete the previous results. Select the file counts.tsv and run the tool RNA-seq / Differential expression using edgeR for multivariate experiments so that you analyze only genes which are expressed in at least 3 samples. Filter the file edger-glm.tsv using Filtering / Filter table by column value (column = PValue.group, cutoff = 0.05)

-how many genes were analyzed and how many are DE?

Repeat the run but incorporate the readtype info to analysis: Add readtype as main effect 2. Save the session.

-how many DE genes do you get now?

14. Visualize differentially expressed genes in genome browser

We return now to the original human dataset. Select File / Open session and the session that you saved in exercise 8. Select file ng-data-table.tsv and run edgeR classic so that you analyze only genes which are expressed in at least 1 sample. How do the plots look like when you don’t have replicates?

Open de-list-edger.bed, click Detach and put the new window aside. Select the BED file, both BAM files and visualization method Genome browser. Maximize the visualization panel, select genome Human hg19, and click Go. Click on the rows of the detached BED file to navigate from one differentially expressed gene to another. Zoom in and out with a mouse wheel.