RNA-Seq expression analysis in Galaxy *From A to Z*

Practicals

Youri Hoogstrate[1,2] & Saskia Hiltemann[1,2]
[1]ErasmusMC, [2]Translational Research IT (TraIT), CTMM

Table of Contents

Introduction

In this workshop we demonstrate a typical analysis done by a bioinformatician working with RNA-seq data using Galaxy. This involves quality control, aligning sequencing data to a known reference genome, doing differential gene expression analysis and visualization.

Action:

- Open a browser and go to: *<< url will be given during the practical >>*
- Register to galaxy

If you have questions, please ask.

Part 1: Raw data to alignment

Action:

- All the data that belongs to pactical 1 – *including results and answers* – can be found in:
  - Shared Data / Part 1: raw data to expression levels

To prepare yourself on the RNA-Seq topic, obtain the presentation from the *Shared Data* and keep the document open to fall back. We will start with a raw paired-end sequencing dataset and process it to gene expression levels and visualize it with an interactive genome browser. The (truncated) dataset is a small set of reads that will align to just a small region on the human genome. We have chosen for this small set to save resources.

Action:

- Import the following files from Shared Data:
  - Part 1: raw data to expression levels
    - `ucsc refseq.gtf`
    - `miR-23b_1.fq`
    - `miR-23b_2.fq`

Original data was treated with a miRNA

Take a look at one of the FASTQ files by pressing the eye-icon. The first part of the analysis is *Quality Assurance / Quality Control*. Therefore we want to get an impression of the quality of the reads.

Action:

- Run FastQC on both FASTQ files.
  - *Hint:* When selecting input files, you can choose multiple datasets and run in parallel.

The tool produces two output files per input file. Only keep the 'website' because the data is presented in a understandable format. We can remove the history items named "*FastQC on ...: RawData*". When we look at the output of FastQC, you will notice a lot of warnings and errors, which arise partially from the fact that we work with a (artificial) very small dataset. Take a look at on of the two websites and answer the questions.

***Questions*:** *you won't get an exam, but they help you understanding RNA-Seq*

- What is the total number of sequences?
  - You can find the number of sequences on top of the FastQ report: 9816
- What is the quality encoding? (**Write this down**, you will need it later on)
  - You can find the quality encoding on top of the FastQ report: *Sanger / Illumina 1.9*
- Take a look at the figures "*Per base sequence quality*", "*Per sequence quality scores*" and "*Sequence Length Distribution*". What is the reason for the warning at the "*Per base sequence quality*"?
  - These figures represent the average of all reads of this dataset. There is a part of the reads that have a low quality (<30) on the beginning , but not too low. Also a part of the reads have a too low quality (<20) bases on the end.

There are, unfortunately, several derivatives of the FASTQ file format. Differences are in particular the quality encoding. The current 'standard' is the FASTQ-Sanger format. To convert FASTQ files into FASTQ-Sanger in Galaxy, you should run the tool *FASTQ Groomer* on both files. Therefore you need to know the quality encoding of our two files (hint: you wrote this down as result of FastQC). If you need a more detailed

explaination of the FASTQ format and the coresponding quality encodings, go to: http://en.wikipedia.org/wiki/FASTQ_format#Encoding or ask for help.

Action:

- Run *FASTQ Groomer* on both `miR-23b_2.fq` and `miR-23b_2.fq`.
- After grooming, rename the files to something meaningful (e.g. "`miR-23b_1_groomed`" and "`miR-23b_2_groomed`"). Otherwise it will become complicated to track the files back.
- Check if the files are converted correctly.
  - Confirm whether the ***format*** is equal to *fastqsanger* (not *fastq*, not *fastqcssanger*!)

In the FASTQ report we saw that the per base quality was insufficient. We can use *Sickle* for trimming low quality bases from the ends of the reads.

Action:

- Select the tool *Sickle* and make sure:
  - Select the appropriate type of reads (paired-end or single-end)
  - Select the correct FASTQ: _1 is forward, _2 is reverse and use those that are groomed.
  - Make sure you use the right quality type (Remember we have fastsqsanger)

Question:

- After running the tool, you will obtain 3 files. Browse through one of them and scroll down.
  - What difference do you see when you look at the newly generated FASTQ files?
    - You will see that some reads are trimmed such that only the high quality bases remain. Also, you will see that you have now three instead of two files.
  - What are the "singletons"?
    - Singletons are the single pairs that remain when the mate is entirely trimmed.
- Run FastQC again, but now on either the forward or reverse files from Sickle.
  - If you run FastQC again, which metrics have improved / changed?
    - Base quality has improved and read length is shorter

To put our high quality reads into a biological context, we align the sequencing data to the human reference genome. The most widely used human reference genome at the moment is *hg19*. Please read the first paragraph (3 scentences) of the following url:
http://en.wikipedia.org/wiki/Reference_genome

Question:

- On how many individuals is hg19 based?
- "Human genome (build 37) is derived from thirteen anonymous volunteers from Buffalo, New York". This amount of genomes is not representative for the entire human population. This is important to understand because this makes you realize that mismatches and detected variants are not neccesairily disease causing.

For RNA-Seq we need specialized RNA aligners, able to cope with gaps that originate from splicing. For this exercise we will make use of a tool called RNA-STAR.

Action:

- Load STAR in galaxy (under the name '*rnastar*') and make sure:

    - single ended or mate-pair ended reads in this library = paired-end
    - forward reads = forward reads, from sickle
    - reverse reads = reverse reads, from sickle
    - built-in index = hg19
- In case RNA-star fails and only if the instructor says so, we will fall back to tophat2 (because of the resources)

Note 1:** during installation of the server it happened that the right **dbkey** was not assigned to the history item. If this is the case, please set it to ***hg19 manually.

***Note 2**: during installation it also occurred that visualization the bam file turned into an error. Rerunning usually worked. If this still doesn't work, raise your hand so the instructors can fix it.

From the results, the "*rnastarrun_starmapped.bam*" is the alignment. This file is in a binary compressed format and visualizing its file contents is not very useful. Instead, we can visualise it interactively in Galaxy.

Action:

- Visualise the BAM file by going to *Trackster*, and save it immediately. When Trackster is loaded, press the **[+]** in the right corner and add *refseq_genes.gtf* and save it again.

Questions:

- Can you find evidence for alternative splicing in region *chr16:15687753-15736550*
    - Hint: Change the visualisation mode of the alignment to "pack".
    - If you look carefully you can see that there are many 'gaps' in the aligned reads that align exactly to the junctions of the exons. These gaps indicate alternative splicing.
- Can you find mismatches in the alignment (or possibly even variants)?
    - Variants in Trackster are indicated with bright colors. There are not too many of them.

To measure the expression levels per gene, we lookup the coordinates of all known genes in a database (e.g. refseq). For each gene, we count the reads that fall within the genes exons.

Action:

- Load the tool *featureCounts* and select:
- alignment file: *rnastarrun_starmapped.bam*
- GFF/GTF source: *Use reference from history*
    - Gene annotation file: *ucsc_refseq.gtf*
- *Leave the other settings default and execute.*

This will result into two files; a summary that contains statistics on the counting and the actual counts. Take a look at the non-summary file; the actual counts. The second gene, WASH7P, which has a read-count of 0 (2nd column) has a *length* of 1769bp. Yet, when we look at the UCSC genome browser its genomic location is given as chr1:14407-29370 and similarly gene-cards says the gene's length is "Size: 15,445 bases".

- Why is there a difference in length?
    - Hint: in what region(s) did we count reads?
    - Since we're interested in statistically independent regions, we're only looking

in exons. The size given in the file is the region in which featureCounts has counted rather than the entire gene.

The featureCounts output has many 0-values. This is because we make use of an artificial small dataset. To show only relevant content, we can filter the 0 value lines out.

Action:

- Proceed with the tool "`Filter data on any column using simple expression`" and set:
  - File: the featureCounts non-summary output file
  - Condition: c2!=0
  - Number of lines to skip: 1

Take a look at the result.

Question:

- Which gene has the highest readcount?
  - ANXA2

Part 2: Differential expression analysis

In the previous exercise we found that gene ANXA2 had the highest readcount, but what does it mean if this gene always has a high readcount in every sample? To say something about expression levels, we should say it in a context relative to other expression levels. Therefore, we need normalization and apply statistical testing. A popular package that allows to do this is EdgeR.
In the following analyses you will determine the differentially expressed genes in the MCF7-cell line between samples that have been treated with the hormone β-estradiol (E2) and those that were left as control. For this analysis we will make use of samples all taken from this cell line. In the corresponding article, the statistical power has been estimated at different sequencing depths (subsampling) and with a different number of replicates.

Reference: http://www.ncbi.nlm.nih.gov/pubmed/24319002.

Tip: Look up where MCF7 cell lines originate from if you don't know this.

For this practical we have the readcounts for a sequencing depth of ~30M (full), 10M and 5M, for each replicate, already made available. Each analysis in this assignment will be to determine the number of differentially expressed (DE) genes and add this number to Table 01:

Table 01: results

Replicates	Seq. Depth	Significant DE genes
0	0	0
7	5,000,000	3287
7	10,000,000	3977
7	30,000,000	4717
5	30,000,000	3078

Important: keep track of this table. You can write these numbers down on paper or in a spreadsheet on the computer. You will need this data for the final assignment.

Subsampled datasets: 5M, 7+7

Create a new histroy called "*DGE MCF7*". Import from the Shared Data library the following items into your history:

- `GSE51403_design_matrix.txt`
- `GSE51403_expression_matrix_10M_coverage.txt`
- `GSE51403_expression_matrix_5M_coverage.txt`
- `GSE51403_expression_matrix_full.txt`

We have a two classes problem: the class itreated with estradiol is called "*E2*", and the other called "*Control*". The design matrix provides the mapping from the RNAseq counts per sample to the phenotype class each is associated with. Take a look at the design matrix and see if you can find samples that belong to those classes.

Action:

- Run the DGE analysis, load galaxy tool:
- `edgeR: Differential Gene(Expression) Analysis RNA-Seq gene expression analysis using edgeR (R package).`
- Analyse the dataset that has the lowest sequencing depth (5 million reads):
  - Select `GSE51403_expression_matrix_5M_coverage.txt` as expression matrix.
  - Select the `GSE51403_design_matrix.txt` as the design matrix.
  - **Define the contrast:** The more complicated part is the biologcial question we want to ask. This question is asked as a mathematical formulation in a format described by a well known R package *limma*. For two class problems it's very simple: *classTreated-classNormal*, which in our case is:
    - *Control-E2* (case sensitive!)
  - Don't select addition output files.

It is always important to check whether what we did was correct. Take a look at the file "*edgeR DGE on 1: GSE51403_design_matrix.txt - differentially expressed genes*". If everything is correct, the gene GREB1 is located in the top of the file. Please check its corresponding gene cards page:

http://www.genecards.org/cgi-bin/carddisp.pl?gene=GREB1

Questions:

- Can you find on the gene cards page a regulatory factor of the gene that relates to the E2 treatment?
  - Hint: what was E2 again?
  - "GREB1 (Growth Regulation By Estrogen In Breast Cancer 1) is a Protein Coding gene."
- Can you find on the gene cards page an association with MCF-7 cells?
  - Hint: what was MCF-7 for type of cell line again?
  - "Diseases associated with GREB1 include breast cancer."

The answers to the questions confirm that what we find with the DGE analysis is in agreement with the setup of the analysis. In the output file, each line represents one gene indicated by the gene symbol in the 2nd column. The *Pvalue* is a probability that represents the chance to find the expression values that belong to the gene given that they are from the same condition. The *FDR* is a multiple testing correction of the *Pvalue* and is usually used instead of the *Pvalue*. The lower this value, the less likely it is that the observed values are derived from the same condition. Thus, differentially

expressed genes will have a low *FDR* and *Pvalue*. To distinguish between differences considered to be caused by chance and differences that are that significantly large that they are considered to be from different conditions, we make use of a cut-off, commonly set to < 0.01 or <0.05.

To find how many genes are significantly differentially expressed, go to "`Filter data on any column using simple expressions`":

- File: "`edgeR DGE on 1: GSE51403_design_matrix.txt - differentially expressed genes`"
- Condition: c7<0.01
- Header lines to skip: 1
- Rename the file to "`DE Genes 5M, 7+7`"

Question:
How many genes are significant differentially expressed between *Control* and *E2*?

- Write this number down in *Table 01* (top of the practical)

Subsampled datasets: 10M, 7+7

In the previous analysis, the FASTQ files contained a total of 5.000.000 reads per sample. For the next analysis we will make use of twice the amount of raw data: 10M reads per sample.

Action:

- Run *`edgeR: Differential Gene(Expression) Analysis`* again
- Use `GSE51403_expression_matrix_10M_coverage.txt` as expression matrix and leave everything else the same (contrast: *Control-E2*, case sensitive!).
- Run the filter to find the number of significant genes
  - Condition: c7<0.01
  - Header lines to skip: 1
  - Rename the file to "`DE Genes 10M, 7+7`"
  - Write this number down in *Table 01* (top of the practical)

Question:

- Do we find more or less differentially expressed genes between the *Control* and *E2* using 10M instead of 5M raw reads per sample?
  - Hint: we have doubled our raw data
  - We doubled the amount of raw data and we have got more but not the doube amount of DE genes.

Subsampled datasets: 30M, 7+7

In the previous analyses, the FASTQ files contained a total of 5.000.000 or 10.000.000 reads per sample. The full data set contains more or less 30.000.000 raw reads per sample.

Action:

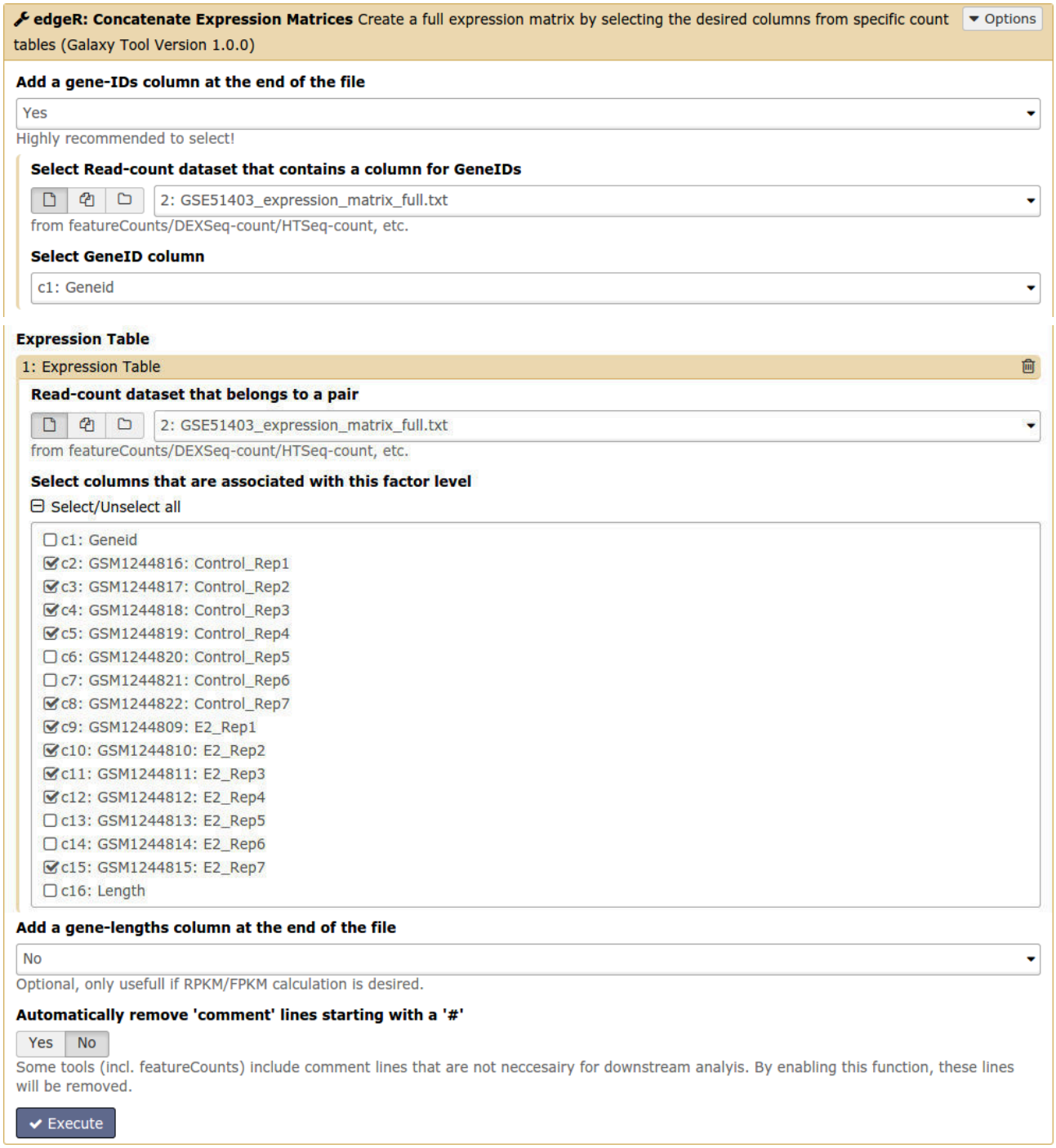
- Run *`edgeR: Differential Gene(Expression) Analysis`* again
- Use "`GSE51403_expression_matrix_full.txt`" as expression matrix and leave everything else the same (contrast: *Control-E2*, case sensitive!).
- Run the filter on the result, to find the number of significant DE genes
  - Condition: c7<0.01
  - Header lines to skip: 1
  - Rename the file to "`DE Genes ~30M, 7+7`"
  - Write this number down in *Table 01* (top of the practical)

Subsampled datasets: 30M, 5+5

We have now ran three tests with 7+7 replicates with different sequencing depths. To see what the effects are of sample replication, we should run the same analysis with a different number of replicates. To modify expression matrices within Galaxy (both concatenating and removal) we can make use of the tool "`edgeR: Concatenate Expression Matrices Create a full expression matrix`". We have used all our replicates in the previous analyses and so we can reduce the number of replicates to 5+5 by simply picking a subset of the samples.

Action:

- Load galaxy tool: `edgeR: Concatenate Expression Matrices Create a full expression matrix`
- Use expression matrix `GSE51403_expression_matrix_full.txt` and select the following 5 replicates per sample:



This will create a truncated version of the expression matrix, only including the desired 5+5 replicates.

- For convenience, rename the new expression matrix to: "`GSE51403_expression_matrix_full__5+5_replicates.txt`".

Action:

- Rerun the tool `edgeR: Differential Gene(Expression) Analysis`
- Use `GSE51403_expression_matrix_full__5+5_replicates.txt` as expression matrix and leave everything else the same (contrast: *Control-E2*, case sensitive!)
- **Enable** the optional output: "**MDS-plot (logFC-method)**".
- Leave everything else the same
- Run "`Filter data on any column using simple expressions`" on the new ...`differentially expressed genes` file:
  - Condition: c7<0.01
  - Header lines to skip: 1
- Count the lines and write it down in Table 01.

Take a look at the MDS plot. If you want to understand all details about MDS plots you should do some research online. For now, what matters, is that the distances between the samples in the plot should correspond more or less to the distances between the samples based on the expression of all (22.000) genes.

Question:

- Do you see separation between the samples from E2 and Control?
  - The classes separate very well.
- Could you think of an application where it would be desired to see high separation between classes?
  - This is very convenient in classification analysis. You can predict what type of sample you have (E2 or Control) based on the results of featureCounts / edgeR.

Question:

Would you expect more or less differentially expressed genes if the experiment was done on individual patient samples instead of cell-line replicates?

You would expect less differentially expressed genes because cell lines are technical replicates that have the same genome and have grown under the same conditions. Instead, patients all have their individual genotype and conditions. This will introduce more variation in biological processed and corresponding expression levels, which will reduce the number of DE genes.

Final question

Can you create a tab delimited file of Table 01 and upload it as a 'tabular' file within Galaxy?

Try to open it in Galaxy as Scatterplot (visualization on the history item) and discuss with other people about the impact of removing these 2 replicates on the statistical power.

You are finished!

In case you can't get enough of it, go to the Shared Data bonus section and answer the following question:

Question:
To which classes do the Unknown samples belong? (Tip: MDS / heatmap)