RNA-seq Analysis

Welcome to the World of RNA-seq. In this practical we will explore how to measure gene expression using sequencing data and bioinformatics. This is a critical skill in functional biology. However, most functional biologists rely on collaborators or technicians to analyse their expression data, losing control on how the data is being treated. Fortunately, you are about to learn how to deal with RNA-seq data yourself.

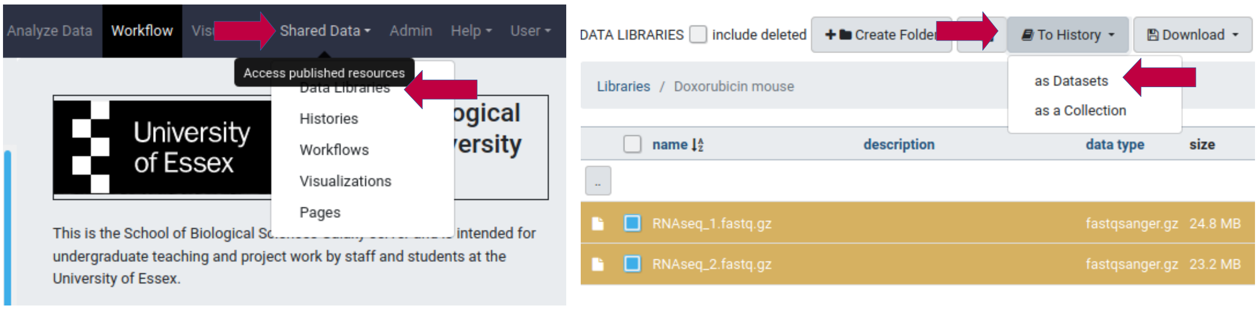
Throughout this practical, you will be using a platform called Galaxy, that integrates multiple state-of-the-art bioinformatics programs. Galaxy is intuitive and easy to use, and all programs you run will provide you with the actual code that Galaxy is running in the background, so you can run the process in a command line environment. But for this practical, we will only focus on the point-and-click use of Galaxy. Off we go!

The biology bit

In the previous part you detected which areas of the genome are targeted by p53. The question is: is this having an effect on gene regulation. Well, let's measure gene expression and see if p53 has an impact on it. We have six mouse cells samples. Three of them were treated with doxorubicin, a compound that activates p53, triggering an apoptotic signal. The other three are control (untreated) cells. By comparing the expression profile of these two sample groups, using a technique called differential gene expression, we will find out which genes are activated (or repressed) by p53 at the transcriptional level.

Mapping reads to a reference genome

You will start mapping sequence reads from a paired-end sequencing reaction against the mouse genome (assembly mm9[1]). First we need to have a look at the *fastq* files provided. Using your account Galaxy (http://galaxy.essex.ac.uk), go to 'Shared Data' → 'Data Libraries' at the top menu. Click on the 'Proficio 2020 - RNAseq' dataset and then select the two RNAseq fastq files and import the datasets by clicking in 'To History' → 'As Datasets' from the 'DATA LIBRARIES' menu (see below).

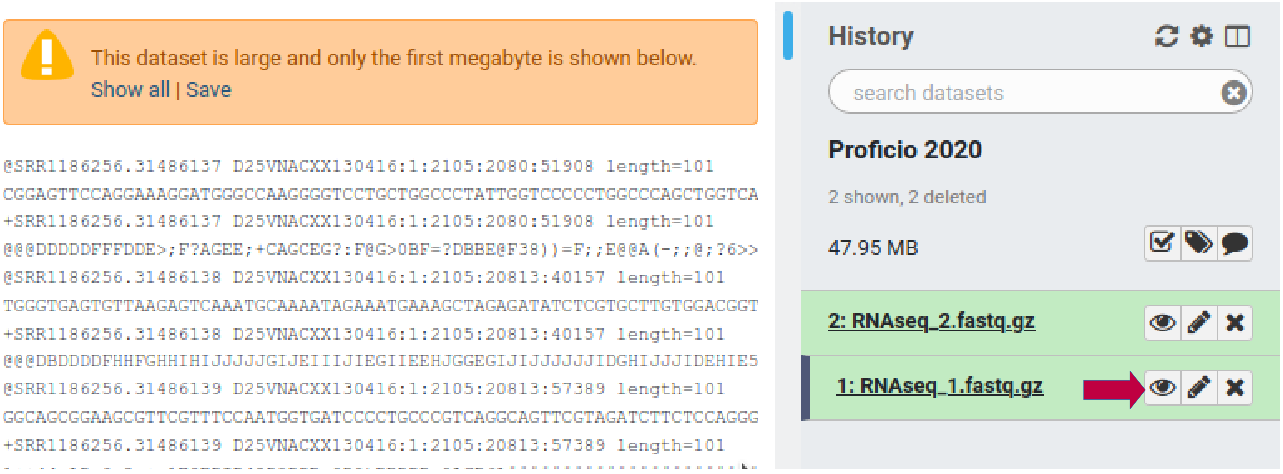


name	description	data type	size
..			
RNAseq_1.fastq.gz		fastqsanger.gz	24.8 MB
RNAseq_2.fastq.gz		fastqsanger.gz	23.2 MB

The RNAseq files will be shortly in your current history. Click on the 'Galaxy' (top-left) icon to go to your home page. You can see two *fastq* files, one ending in '_1.fastq' and another ending in '_2fastq'. These two files correspond to the same paired-ends sequencing run. Each sequence read

[1] Each genome assembly has a name. *mm9* is the 9th assembly of the *Mus musculus* (house mouse) genome. Is still used as it is very well annotate, although the most up-to-date version version is *mm10*. Likewise, the most recent human genome assembly is called *hg38*.

in the first file has a corresponding pair in the second file, and we need to map the two files together. Have a look at the first *fastq* file clicking in the 'eye' icon next to the dataset:

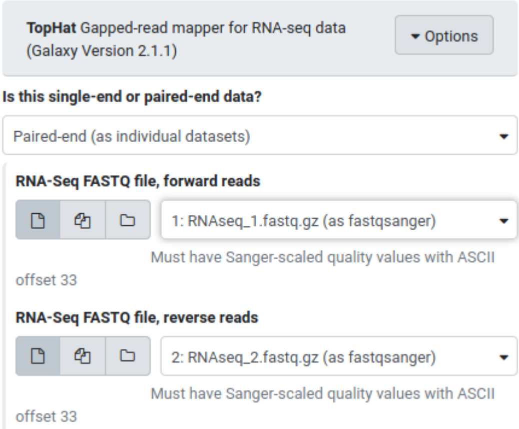


This dataset is large and only the first megabyte is shown below.
Show all | Save

```
@SRR1186256.31486137 D25VNACXX130416:1:2105:2080:51908 length=101
CGGAGTTCCAGGAAAGGATGGGCCAAGGGGTCCTGCTGGCCCTATTGGTCCCCCTGGCCCAGCTGGTCA
+SRR1186256.31486137 D25VNACXX130416:1:2105:2080:51908 length=101
@@@DDDDDFFFDDE>;F?AGEE;+CAGCEG?:F@G>0BF=?DBBE@F38))=F;;E@@A(-;;@;?6>>
@SRR1186256.31486138 D25VNACXX130416:1:2105:20813:40157 length=101
TGGGTGAGTGTTAAGAGTCAAATGCAAAATAGAAATGAAAGCTAGAGATATCTCGTGCTTGTGGACGGT
+SRR1186256.31486138 D25VNACXX130416:1:2105:20813:40157 length=101
@@@DBDDDDFHHFGHHIHIJJJJJGIJEIIIJIEGIIEEHJGGEGIJIJJJJJJIDGHIJJJIDEHIE5
@SRR1186256.31486139 D25VNACXX130416:1:2105:20813:57389 length=101
GGCAGCGGAAGCGTTCGTTTCCAATGGTGATCCCCTGCCCGTCAGGCAGTTCGTAGATCTTCTCCAGGG
+SRR1186256.31486139 D25VNACXX130416:1:2105:20813:57389 length=101
```

Does this format look familiar to you? This is the FASTQ format as described in the slides. Time to map these reads into the reference genome.

In the previous part you mapped sequence reads from a ChIP-seq reaction with `Bowtie`. As we discussed already, `Bowtie` can't map reads across introns. Fortunately, the program `TopHat` does precisely that. To make `TopHat` map your reads you can run it from Galaxy. In the 'search tools' dialog look for the TopHat program (by typing 'tophat'), and click on the program 'TopHat'. To the question 'Is this single-end or paired-end data?' select 'Paired end (as individual datasets)' from the menu, and then select RNAseq_1 as forward reads, and RNAseq_2 as reverse reads:



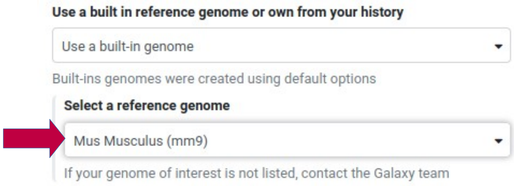
TopHat Gapped-read mapper for RNA-seq data (Galaxy Version 2.1.1)

Is this single-end or paired-end data?
Paired-end (as individual datasets)

RNA-Seq FASTQ file, forward reads
1: RNAseq_1.fastq.gz (as fastqsanger)
Must have Sanger-scaled quality values with ASCII offset 33

RNA-Seq FASTQ file, reverse reads
2: RNAseq_2.fastq.gz (as fastqsanger)
Must have Sanger-scaled quality values with ASCII offset 33

Select in the 'Select a reference genome' menu the 'Mus Musculus (mm9)' option:



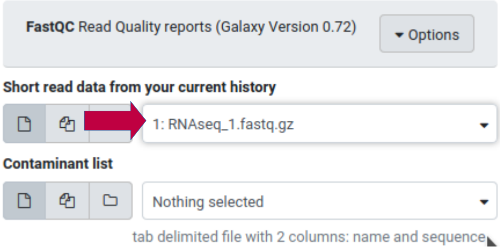
Use a built in reference genome or own from your history
Use a built-in genome
Built-ins genomes were created using default options

Select a reference genome
Mus Musculus (mm9)
If your genome of interest is not listed, contact the Galaxy team

Leave the other options with their default values and click on the 'Execute' blue button below. The program will run for approximately 15 minutes and after the mapped reads will be available in your current history.

Evaluating the quality of a reads file

While the reads are being mapped by TopHat, you can evaluate the quality of the reads, as we discussed during the lecture. To do so, we will be using FastQC in Galaxy. In the ‘search tools’ panel write ‘fastqc’ and then select the ‘FastQC’ tool. Select the RNAseq_1 dataset in the ‘Short read data from your current history’ menu, and ‘Execute’ the program:



FastQC Read Quality reports (Galaxy Version 0.72)

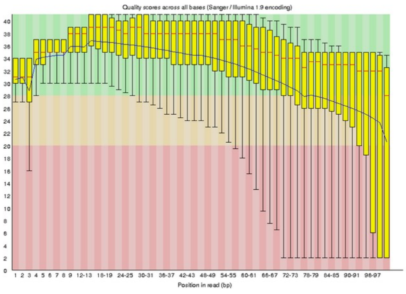
Short read data from your current history: 1: RNAseq_1.fastq.gz

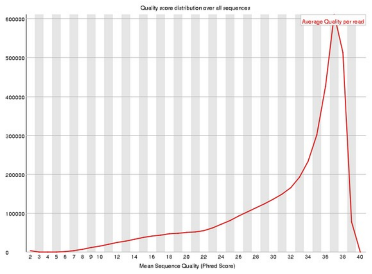
Contaminant list: Nothing selected

tab delimited file with 2 columns: name and sequence

Two new files will appear in your history window, and they will be in yellow while the process is running. In about 5 minutes the files will turn green. They’re ready!

Click on the ‘eye’ icon of the file `FastQC` on data 1: Webpage’. You will see a number of diagnosis plots, the most informative being the ‘Per base sequence quality’ and ‘Per sequence quality scores’. Have a look at them and discuss with colleagues. Your instructor will guide you through them. (If the images do not open in the Galaxy server, you can download the files using the ‘disk’ icon and open them locally in your computer with a web browser.)





How is the quality of the reads overall?

Count reads for each gene

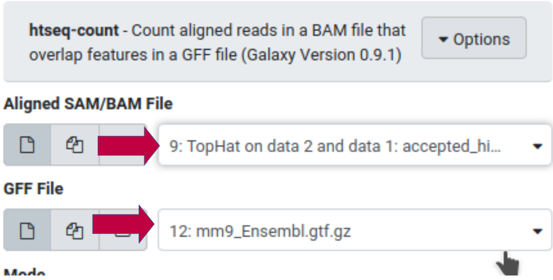
If everything went well, your reads are now mapped to the genome. That’s good news. The bad news is that the mapped reads are in a very tricky format called BAM, which is a binary compressed version of another format called SAM. But you don’t really need to understand these formats as we can convert them to something more intuitive: read counts. That is, we are going to read the output file of `TopHat` and convert it in a table that list all genes in the mouse genome and the number of reads mapped to this gene[2].

`TopHat` produces a few output files, as you can see in your history panel. For our purposes, most of these files are irrelevant. The reads mapped to the genome are in the output ending with ‘accepted_hits’. To find out how many reads are associated to each gene we need this file, but also a

[2] The number of reads mapped to a gene is a reflection of the amount of transcript produced, that is, the gene expression level.

file of genome coordinates that tells us where in the genome is each gene. This file is a GTF[3] file, and you can find it in the Shared Data folder ‘Proficio 2020 – RNAseq’. Go there and import it into your Galaxy history.

To get the read count we can (and we will) use the program `HTSeq`, again, in Galaxy. Find the program ‘htseq-count’ in the Tools menu and open it. Select the ‘accepted_hits’ output form TopHat in the ‘Aligned SAM/BAM File’ menu, and the *gtf* file in the ‘GFF File’ menu, and ‘Execute’ the program:

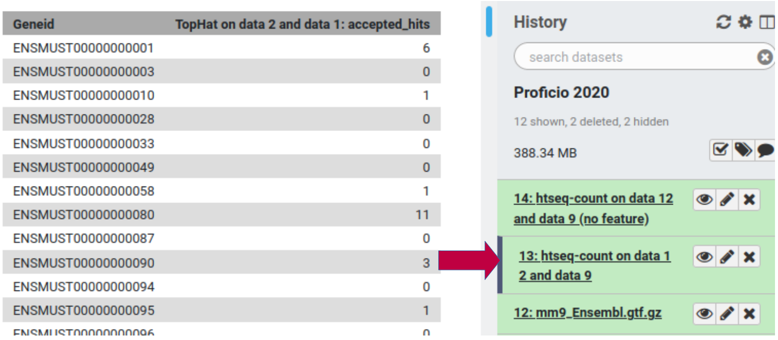


htseq-count - Count aligned reads in a BAM file that overlap features in a GFF file (Galaxy Version 0.9.1)

Aligned SAM/BAM File: 9: TopHat on data 2 and data 1: accepted_hi...

GFF File: 12: mm9_Ensembl.gtf.gz

That will take a few minutes. At the end, the output file will be in your history panel. If you have a quick look (‘eye’ icon) at the file you will see that it looks exactly as we expected, a list of genes (or Gene IDs) and the read counts associate to them:



Geneid	TopHat on data 2 and data 1: accepted_hits
ENSMUST00000000001	6
ENSMUST00000000003	0
ENSMUST00000000010	1
ENSMUST00000000028	0
ENSMUST00000000033	0
ENSMUST00000000049	0
ENSMUST00000000058	1
ENSMUST00000000080	11
ENSMUST00000000087	0
ENSMUST00000000090	3
ENSMUST00000000094	0
ENSMUST00000000095	1
ENSMUST00000000096	0

History

Proficio 2020

12 shown, 2 deleted, 2 hidden

388.34 MB

14: htseq-count on data 12 and data 9 (no feature)

13: htseq-count on data 1 2 and data 9

12: mm9_Ensembl.gtf.gz

Why the counts are very small (or 0) for most of the genes? Well, you have been mapping a subset of the actual reads produced during sequencing, but if you were to map the whole set, the mapping may take days to run! And we don’t have much time in this course. But at least you know how to do it, just keep in mind that when using your own data the whole process will take a while. In the next section we will be using the mapping reads from the actual RNAseq experiments.

Differential Gene Expression

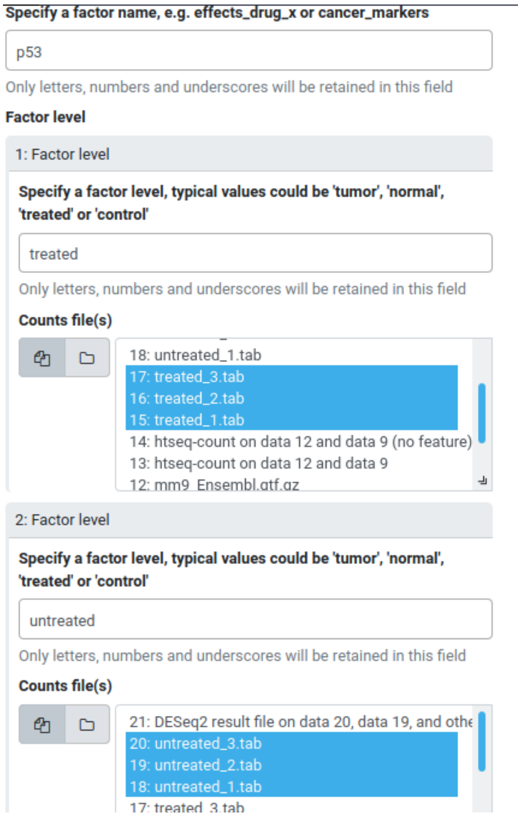
A convenient way to perform a differential gene expression analysis is with the `DESeq2` program installed in our Galaxy server.

Go again to ‘Shared Data’ → ‘Data libraries’ → ‘Proficio 2020 - RNAseq’. Select the six files with read counts, three ‘treated’ and three ‘untreated’, and click on ‘to History’ → Import. Go back to the main Galaxy page.

Find the `DESeq2` program in the tools window. In the ‘Factor name’ field write ‘p53’ (or any other meaningful title you may think of). On the box under ‘1:Factor Level’ type ‘treated’ and then select,

[3] General Transfer Format

using the CTRL key, the three ‘treated’ datasets. On the box ‘2:Factor Level’ type ‘untreated’ and then select the three ‘untreated’ datasets:



Specify a factor name, e.g. effects_drug_x or cancer_markers

p53

Only letters, numbers and underscores will be retained in this field

Factor level

1: Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

treated

Only letters, numbers and underscores will be retained in this field

Counts file(s)

18: untreated_1.tab
17: treated_3.tab
16: treated_2.tab
15: treated_1.tab
14: htseq-count on data 12 and data 9 (no feature)
13: htseq-count on data 12 and data 9
12: mm9_Ensembl.gtf.gz

2: Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

untreated

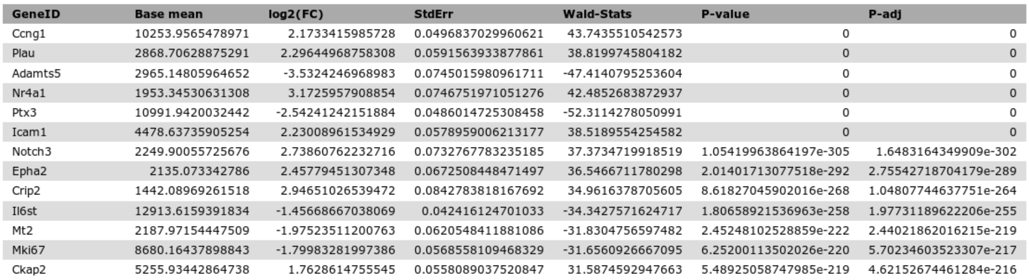
Only letters, numbers and underscores will be retained in this field

Counts file(s)

21: DESeq2 result file on data 20, data 19, and othe
20: untreated_3.tab
19: untreated_2.tab
18: untreated_1.tab
17: treated_3.tab

Leave the other options by default and click on ‘Execute’. Wait a couple of minutes.

One of the outputs, ‘DESeq2 plots’, shows a number of plots. Unfortunately, we don’t have time in a short course to discuss them, but it’s worth that you open them and try to understand. Another output is ‘DESeq2 result file’, which contains the main output we are going to analyse. Click on the ‘eye’ icon and you’ll see something like that:



GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
Ccng1	10253.9565478971	2.1733415985728	0.0496837029960621	43.7435510542573	0	0
Plau	2868.70628875291	2.29644968758308	0.0591563933877861	38.8199745804182	0	0
Adamts5	2965.14805964652	-3.5324246968983	0.0745015980961711	-47.4140795253604	0	0
Nr4a1	1953.34530631308	3.1725957908854	0.0746751971051276	42.4852683872937	0	0
Ptx3	10991.9420032442	-2.54241242151884	0.0486014725308458	-52.3114278050991	0	0
Icam1	4478.63735905254	2.23008961534929	0.0578959006213177	38.5189554254582	0	0
Notch3	2249.90055725676	2.73860762232716	0.0732767783235185	37.3734719918519	1.05419963864197e-305	1.6483164349909e-302
Epha2	2135.073342786	2.45779451307348	0.0672508448471497	36.5466711780298	2.01401713077518e-292	2.75542718704179e-289
Crip2	1442.08969261518	2.94651026539472	0.0842783818167692	34.9616378705605	8.61827045902016e-268	1.04807744637751e-264
Il6st	12913.6159391834	-1.45668667038069	0.042416124701033	-34.3427571624717	1.80658921536963e-258	1.97731189622206e-255
Mt2	2187.97154447509	-1.97523511200763	0.0620548411881086	-31.8304756597482	2.45248102528859e-222	2.44021862016215e-219
Mki67	8680.16437898843	-1.79983281997386	0.0568558109468329	-31.6560926667095	6.25200113502026e-220	5.70234603523307e-217
Ckap2	5255.93442864738	1.7628614755545	0.0558089037520847	31.5874592947663	5.48925058747985e-219	4.62152674461284e-216

The columns of interest in the output file are: GeneID, the gene names; log2(FC), the fold change of one sample with respect to other (I explain this below); and the P-adj, the p-value corrected for False Discovery Rate.

How to interpret the log2(FC)? First look at the symbol. If it’s positive, the gene is higher expressed in the treated samples than in the untreated one. If it’s negative is it’s the other way around. Now, look at the magnitude. A value of 1 indicates that expression in one sample is double than expression in the other. If it’s 2 the expression is 4 times higher in one sample, if its 3, the expression is 8 times higher, and so forth. For instance, a log2(FC) of -2 indicated that the specific genes is expressed 4 times less in the treated sample than in the untreated one.

Making sense of the results

From here, one can go in many directions, depending on the specific interest. In the last practical you will integrate RNA-seq results with ChIP-seq results. But for completeness, let's do an additional analysis that you may find useful in the future.

Let's find which genes are overexpressed in the 'treated' samples, that is, genes that are activated by p53. In Galaxy find the tool 'Filter' and click on it. Select the 'DESEq2 results file'. We are going to be very strict and select genes with a log2(FC) over 3, and a adjusted p-value of 0.001 or below. Thus, in the 'With following condition' box we type:

```
c3>3 and c7<=0.001
```

and we click on 'Execute'.

The output shows 22 genes that are heavily influenced (up-regulated) by p53 activity. Next step (which is not to be covered here) will be to go to the lab, and perform a few RT-qPCRs to validate these results.

Recommended readings

It's virtually impossible to cover RNA-seq in just a couple of hours (I run a whole semester on that), but I hope you get a fair overview. If you want to learn more about it, there are a few references you may consider. A general textbook is that one:

Pevsner (2015) *Bioinformatics and Functional Genomics*

A more specialised book (yet easy to read) is this:

Korpelainen et al. (2015) *RNA-seq Data Analysis: A practical approach*

But for almost anything you can do in Bioinformatics, ask Google first.