# Introduction to RNA-Sequencing Analysis for Differential Expression on Galaxy

## Part 2: DNA/RNA Sequencing Course, Westmead

**Authors**: Tracy Chew, Nicholas Ho, Rosemarie Sadsad

**Sydney Informatics Hub**
https://informatics.sydney.edu.au
sih.info@sydney.edu.au

## Course Overview

**Part A**: Introduction
- A1. Why sequence RNA?
- A2. How does RNA sequencing work?
- A3. Experimental design
- A4. Analysis workflow overview

**Part B**: Alignment and Visualisation
- B1. Uploading data on Galaxy
- B2. Alignment with HISAT2
- B3. Visualisation with IGV

**Part C**: Differential expression analysis
- C1. Generating raw count data with featureCounts
- C2. Differential expression with DESeq2
- C3. Functional annotation

**Part D**: Useful resources

**Before we begin**:
Please have the latest version of Java installed.
https://www.wikihow.com/Update-Java

To view my results on Galaxy, click the link below:
http://galaxy-mel.genome.edu.au/galaxy/u/tracyc/h/rnaseq2018

THE UNIVERSITY OF SYDNEY
Sydney Informatics Hub

## Part A: Introduction

### *Part A1: Why sequence RNA?*

**RNA sequencing** provides insights into the transcriptome of a cell, enabling scientists to understand how DNA is being expressed at a given point in time under certain conditions.

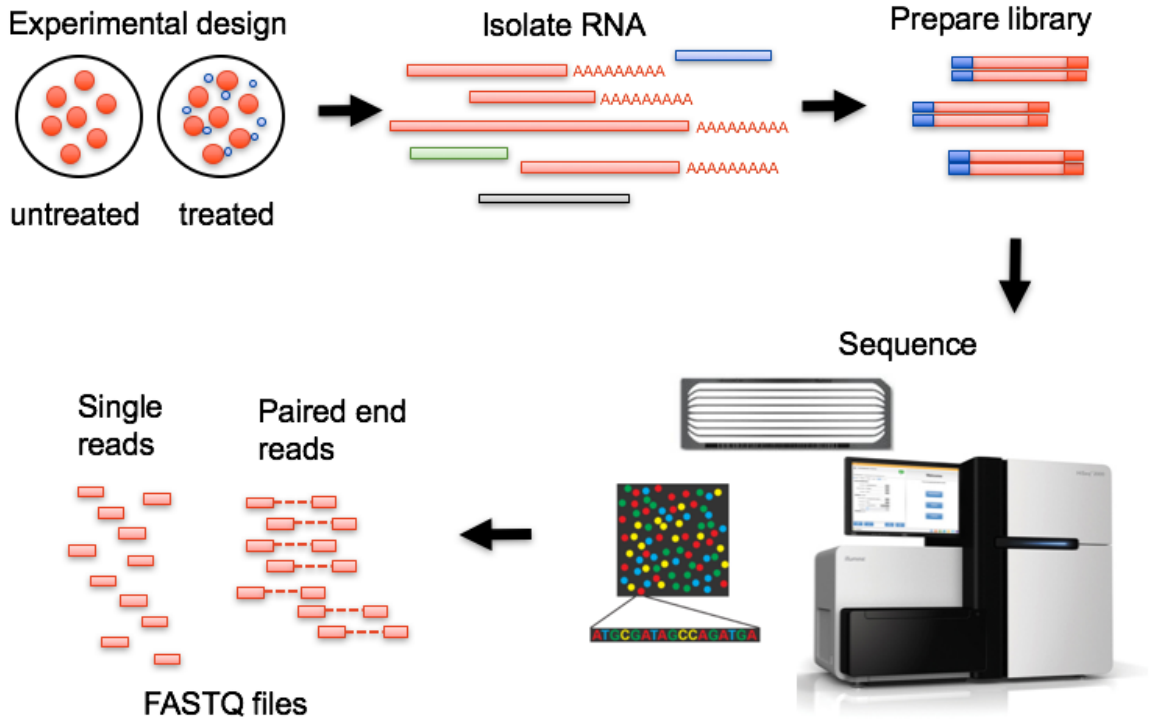
There are a number of research questions that we can answer using RNA sequencing, such as:

- Which genes are differentially expressed under certain conditions? (e.g. treatment vs. control, disease vs. healthy)
- Are there novel transcripts present in my sample?
- Is my gene of interest alternatively spliced?

In this course, we will learn how to conduct a full RNA-Sequencing analysis pipeline for **differential expression** on **Galaxy**.

### *Part A2: How does RNA sequencing work?*

The image below provides a summary of how RNA is sequenced.



**The basic steps of RNA sequencing:**

- Conducting your experiment
- Isolating high quality RNA (and usually selecting for mRNA to remove other RNA that are not of interest such as ribosomal RNA)
- Preparing libraries for sequencing (for Illumina sequencing platforms, it includes converting RNA to cDNA, fragmentation, size selection, attaching adapters)
- Sequencing. On an Illumina HiSeq platform, cDNA is sequenced on a flowcell. There are 8 lanes per flowcell, 200 million reads per lane with the option to have up to 24 samples per lane)
- Depending on your choice, you can have single or paired end sequencing reads which will be sent to you in FASTQ format.

***A3. Experimental design***

Before you start your experiment, we want to emphasize that you should carefully consider your project's **experimental design** to ensure that you have sufficient data to provide you with statistically sound results and that these can provide you with answers to the questions being asked. As with any other experiment, you should take measures to avoid **bias.**

For RNA sequencing experiments, it may be tempting to sequence fewer samples and to reduce the cost of the experiment. However, without enough **replicates,** you may not be able to capture enough variability for statistical testing to account for noise that may be present, particularly from **biological or technical** variation.

**Biological replicates:** these should include biologically distinct samples that capture random biological variation that may exist (sample to sample differences). Gene expression is likely to vary across different individuals and enough replication is required to determine what is natural individual variation and what is caused by the experimental condition of interest.
**Technical replicates:** these include repeated measurements of the same sample, to account for random noise associated with a laboratory protocol or equipment.

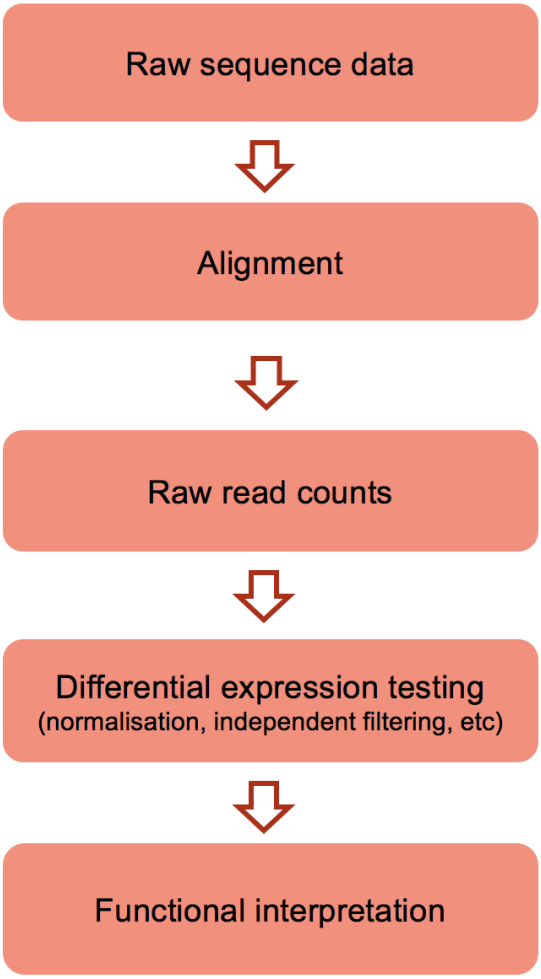
To ensure you have enough and the correct type of data to appropriately answer your question, you also need to consider stranded vs unstranded protocols, depth of coverage, read length, single read vs. paired end read sequencing, etc.

We recommend that you speak to your local statistician and sequencing company to determine the best balance between cost, amount of data and number of samples required. Also read some published articles, such as by:

- Schurch et al., 2016 who looked at the recommended number of biological replicates
- Corley et al., 2017 who compared single-end vs paired-end reads and stranded vs unstranded protocols
- Auer and Doerge., 2010 who provide advice on statistical design and analysis of RNA sequencing data, in particular using "balanced block designs" across lanes of a flowcell

***Part A4. A typical differential expression analysis workflow***

In the next couple of sections, we will learn how to perform differential expression analysis using RNA sequence data. A typical workflow is represented below:




Before performing your analysis, we recommend that you **store your raw data on a secure system** (such as the University's Research Data Store) and perform **quality control** on the raw read data.

FastQC is a popular program that generates quality score reports on raw sequence data (FASTQ files) and is available on Galaxy. We will not be going through this in the course today but you can find and explanation of the report here.

To learn how to properly manage and store your data using the Research Data Store, consider attending our "Data transfer and RDS for HPC course".

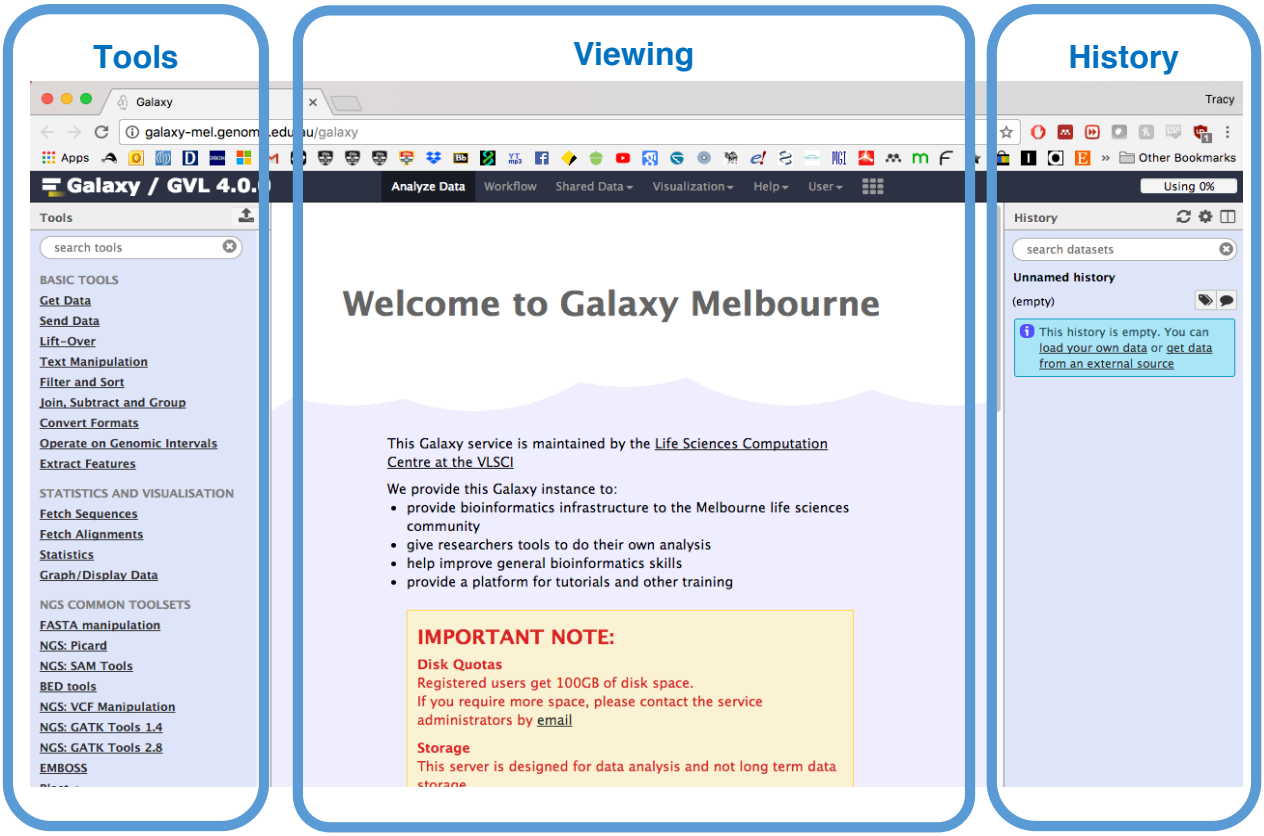
**Part B: Alignment and visualisation**

***Galaxy***

This course will be based on a web-based platform called **Galaxy**. It is a useful tool for biologists to perform analysis workflows and contains a lot of common bioinformatics tools. Registered users can use 100Gb of space for free.

We will use Galaxy (Melbourne instance) for aligning, counting and differential expression analysis.

- Go to the Melbourne Galaxy Server (in Firefox, Chrome or Safari): http://galaxy-mel.genome.edu.au/galaxy
- If you haven't already, register (click User > Register), then log in



***The study***

In this course we will work with simulated data obtained from a study by Corley et al., 2016. This study uses a mouse model to understand Williams-Beuren Syndrome (WBS). This is a rare developmental disorder that affects multiple body systems. It is commonly characteristed by:

- distinctive facial features
- mild to moderate intellectual disability
- cardiovascular abnormalities

The *Gtf2ird1* transcription factor gene is commonly disrupted in patients with the disease and are a prime candidate gene linked to the disease phenotype. To better improve our understanding of this gene, Corley et al created a ***Gtf2ird1*** **knockout mouse model** (see Palmer et al. 2007).

There were 3 replicates in each group:

Wildtype

| | | |
|---|---|---|
|  WT_1 |  WT_2 |  WT_3 |

Knockout

| | | |
|---|---|---|
|  KO_1 |  KO_2 |  KO_3 |

***Part B1. Uploading data***

Sequencing companies will often send you a direct link to download your data in the form of a FASTQ file (sometimes gzipped, with the extension .fq.gz). You can upload files onto Galaxy from your local computer or from the download URL.

Here we will upload the data using a download link but in practice, **I highly recommend** that you store your **raw data** in a secure place, such as the **Research Data Store (RDS)** systems provided by the University.

- Copy the links below

**FASTQ files**

https://informatics.sydney.edu.au/services/coursedocs/SRR3473984.fastq
https://informatics.sydney.edu.au/services/coursedocs/SRR3473985.fastq
https://informatics.sydney.edu.au/services/coursedocs/SRR3473986.fastq
https://informatics.sydney.edu.au/services/coursedocs/SRR3473987.fastq
https://informatics.sydney.edu.au/services/coursedocs/SRR3473988.fastq
https://informatics.sydney.edu.au/services/coursedocs/SRR3473989.fastq

- Go back to Galaxy
- Click the upload icon on the top left corner
- Click Paste/Fetch data
- Paste in the links
- Change file type to “fastqsanger”

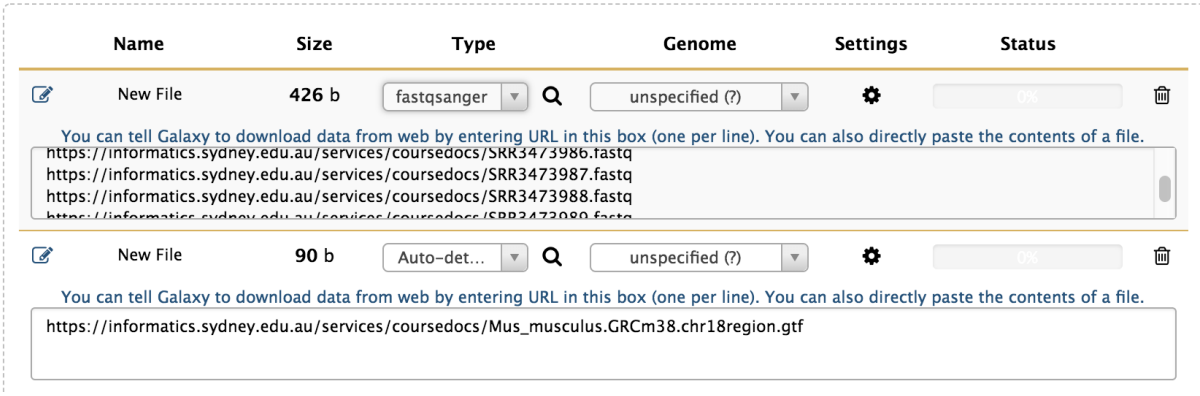
While we are here, we will upload the annotation file that is required later on in the course.

- To upload additional files of a different type, Paste/Fetch data
- Copy and paste the below link in the new box that appears
- Leave type as “Auto-detect” (Galaxy can recognize this file type)

**Annotation file (GTF)**

https://informatics.sydney.edu.au/services/coursedocs/Mus_musculus.GRCm38.chr18region.gtf

This step should appear as below:




- Click Start to begin uploading the six FASTQ and one GTF file
- Close the white box

The files will now appear in your history pane.

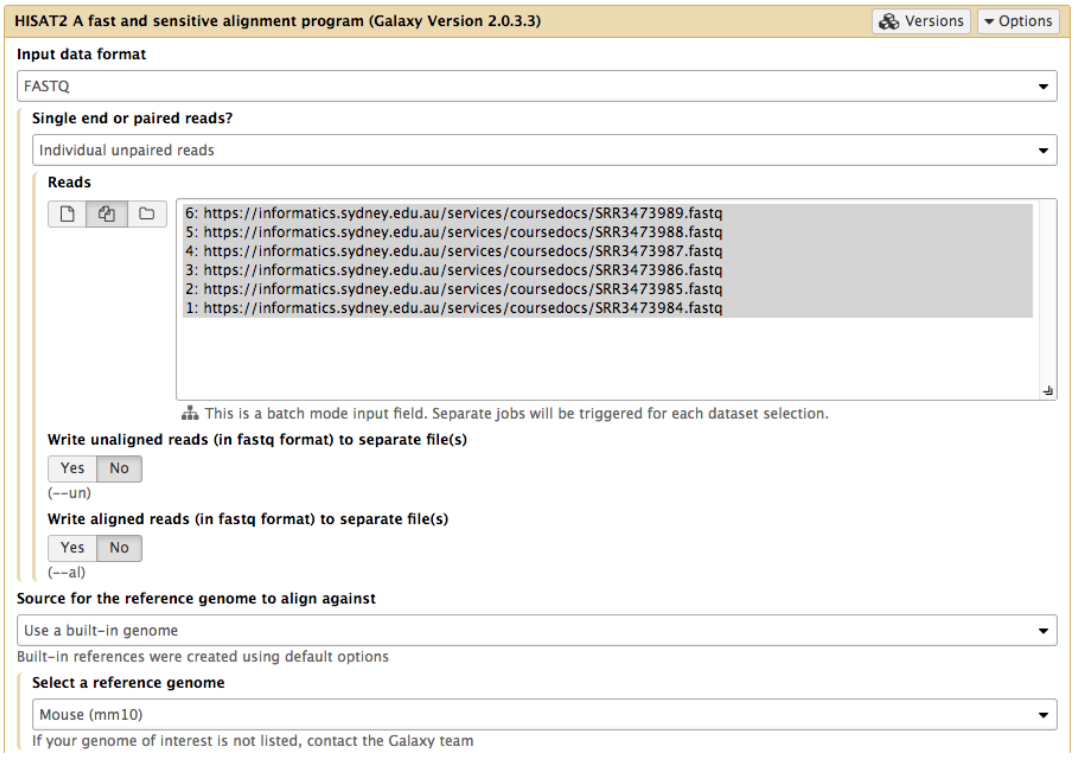
*[Optional]*

- Click on the eye icon to view what a FASTQ file looks like


***Part B2. Alignment with HISAT2***

Now that we have uploaded our raw FASTQ files, we are ready to begin aligning them to the mouse reference genome. HISAT2 is a fast and sensitive spliced alignment program.

- In the tools panel, click “NGS: RNA Analysis”
- Click “HISAT2”
- Under “Single end or paired reads?”, select “individual unpaired reads”
- Click the multiple datasets icon
- Highlight all six FASTQ files that we just uploaded
- Select “Mouse (mm10)” under “Select a reference genome”
- Leave the other values as default
- Click ✔ Execute




The alignments may take a couple minutes to complete.

***Part B3. Visualisation with the Integrated Genomics Viewer (IGV)***
IGV is a popular java application that allows the visualisation of alignments (.bam files). IGV can be launched directly from Galaxy (but requires the latest version of Java to be installed).

**Unfortunately, we have forgotten to label our samples and don't know which samples belong to the wildtype or knockout group!**

| File name | New sample ID<br>(e.g. WT_1, WT_2, WT_3, KO_1, KO_2, KO_3) |
|---|---|
| SRR3473984.fastq | |
| SRR3473985.fastq | |
| SRR3473986.fastq | |
| SRR3473987.fastq | |
| SRR3473988.fastq | |
| SRR3473989.fastq | |

Here, we will use IGV to visualise our alignments to assign each sample to their correct treatment group.

The key to this is understanding how the knockout mouse model was generated. In the paper, we can find that:

> "A loss of function mutation of *Gtf2ird1* was generated by a random insertion of a *Myc* transgene into the region, **resulting in a 40 kb deletion surrounding exon 1**"

**Exon 1 is located**:

chr5:134,332,897-134,481,480

Let's go back to Galaxy and check to see if our alignments are complete.

- In the history panel, click on "HISAT2 on data 1" (note that your data number may be different)

The pane will open to reveal alignment statistics and other options. Notice how the aligned files are in BAM format (binary SAM format) and are not human readable.

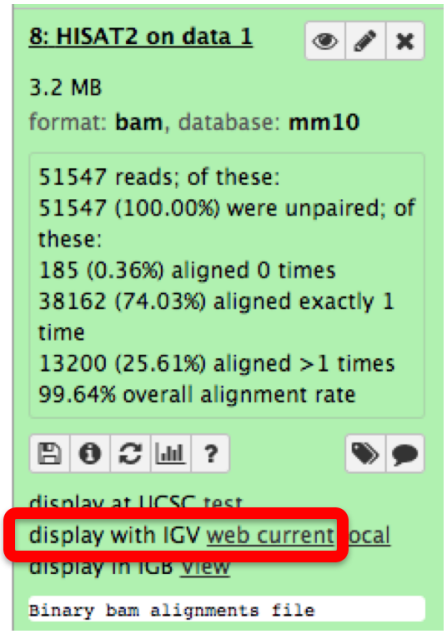
- Click on "web_current" next to "display with IGV"
- Save the "igv.jnlp" file
- Open the file

**Note for Mac users**: if you are prompted with a security error message, right click on the file, show in Finder, and select Open With > Java Web Start.

**Using IGV**

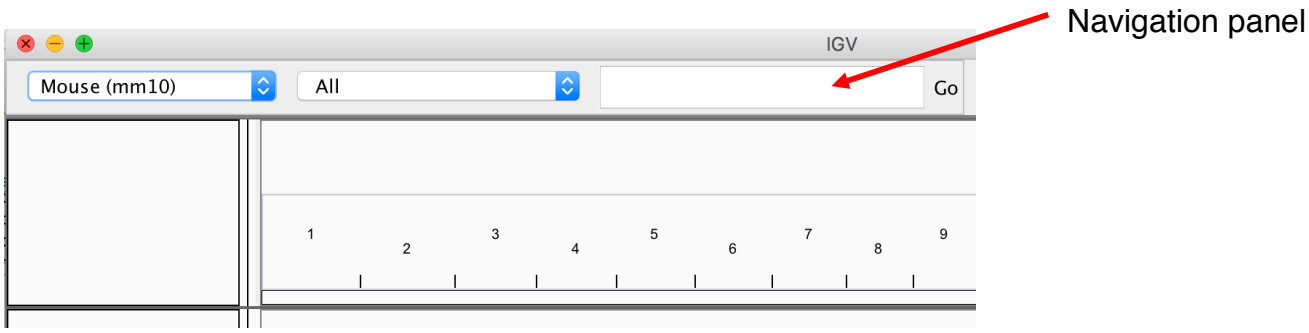
IGV should now be open. At the top you will see some navigation tools. Underneath are the viewing panels for chromosomes, your sample data, and Refseq gene annotations. These are customisable.

.




- At the top left corner, change the reference genome from the preloaded “Human (hg19)” genome to “Mouse (mm10)”

It will take a couple of seconds for IGV to load the genome. Notice that the number of chromosomes on the first viewing panel changes. On the left-hand panel, you will see that your alignment file has been loaded. In this case, “HISAT2 on data 1” has been loaded.

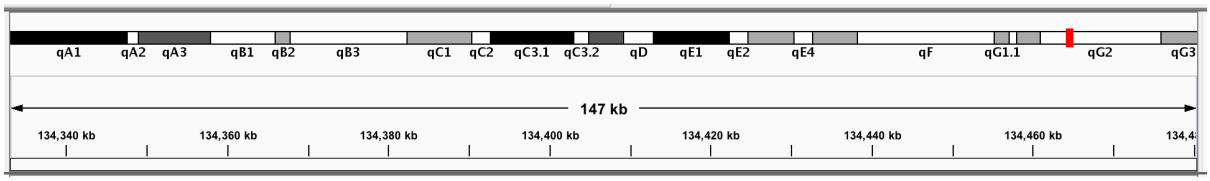



- Let’s open another 2 alignments (no more than this as viewing can become too small). Go back to Galaxy and click on a different alignment. Since IGV is already open, you can click “local” instead of “web_current”
- Navigate to the *Gtf2ird1* gene on chromosome 5 by copying this region into the navigation panel:

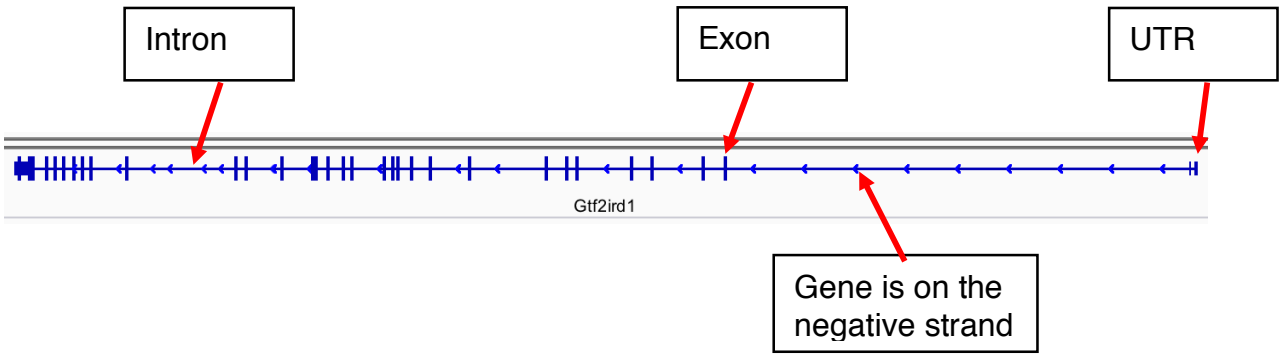
  chr5:134,332,897-134,481,480

- Click “Go”

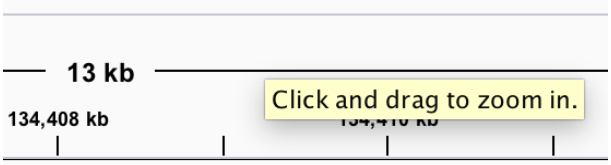
You will now see a chromosome map at the top. The region in red is the region currently being viewed with a scale and positions provided underneath. You should be on at qG2 of chromosome 5.




In the Refseq genes track, you will see Gtf2ird1. The positions above directly correspond to the features in the Refseq genes track.



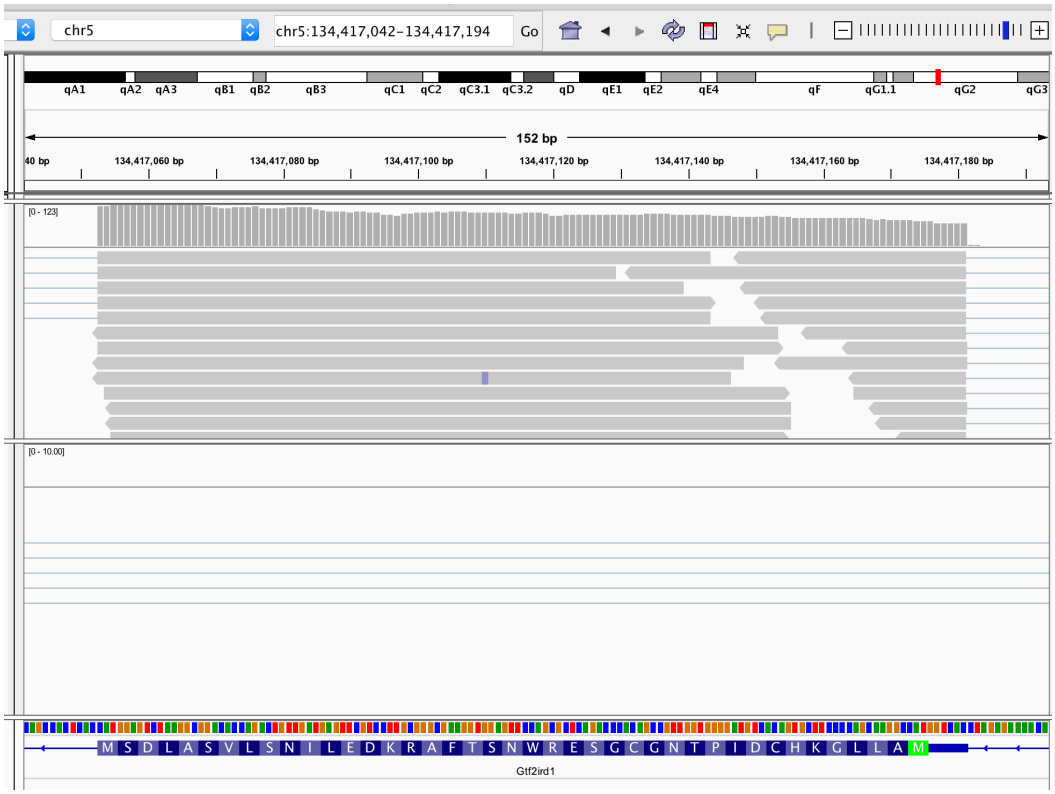

- Practice other forms of navigation, e.g.:
  - Click and dragging a region




  - Using the zoom in and zoom out tool

You will need to zoom in close enough to see your aligned reads. The grey bars represent your reads.

At the first exon, you should see either of these:

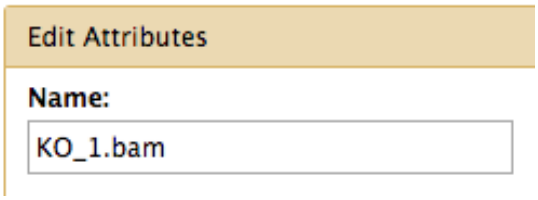



- To view this exact region, navigate to: chr5:134,417,042-134,417,194

**Renaming bam files on Galaxy**

Once you have identified which samples are part of the wildtype or treatment groups, rename them to something more meaningful.

- Navigate to the history panel
- On one of your bam files ("HISAT2 on data …") click on the edit attributes icon
- Change the name to something more meaningful, e.g. "WT_1.bam", "WT_2.bam", "WT_3.bam", "KO_1.bam", "KO_2.bam", "KO_3.bam".
- Click save or hit enter



Your files should look like this:


## Part C. Differential expression

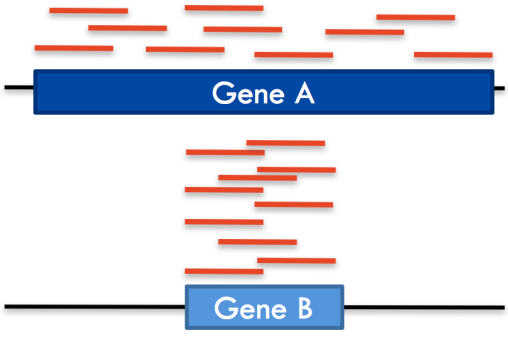
Determining which genes are differentially expressed requires counting how many reads aligned to each gene and determining if the number of reads aligned is significant between the treatment vs. control groups.

This relies on using a gene annotation file (gff3 or gtf). We uploaded one earlier (Mus_musculus.GRCm38.chr18region.gtf ).

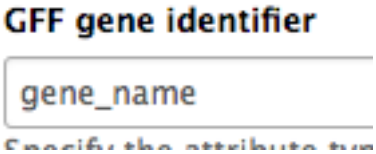
- Click on the eye icon to see what the file looks like.


### *Part C1. Generating count data with featureCounts*

We will use the program featureCounts to obtain raw counts. Here, for each sample and gene (from the annotation - GTF – file), the number of reads that aligned are counted. If expression in a particular gene is higher in one sample than the another, we would expect more reads to be aligned in the sample where expression is upregulated. We will use these raw counts to statistically determine if there is differential expression between treatment groups.



- In the tools panel under “NGS: RNA Analysis”, click on “featureCounts”
- Under Alignment file at the top, click on the multiple datasets icon 
- Highlight all six bam files
- Under “Gene annotation file”, select “in your history”. Select the “Mus_musculus.GRCm38.chr18region.gtf” file that was uploaded earlier.
- In “Advanced options”, change the GFF gene identifier field to “gene_name”




- Leave the other values as default and click 


**featureCounts output**

The program will output two files per bam file, a summary file containing the total number of reads that were “Assigned” to a gene, or “Unassigned” and why. Note, this file may be “hidden” in Galaxy. The second file contains your count information for each of the genes.

- Click on the eye icon to view the featureCounts output and observe the results. It should look like the image to the right.
- Rename the output files to something more meaningful (WT_1_counts, KO_1_counts, etc). We don’t need the summary files so you can ignore them (or if you wish to, remove them from the history pane by clicking  ).

| Geneid | KO_3.bam |
|---|---|
| Ccdc68 | 55 |
| 1700061H18Rik | 1 |
| 4930448D08Rik | 2 |
| Rab27b | 1186 |
| Dynap | 19 |
| Gm45879 | 0 |
| 4930503L19Rik | 112 |
| Stard6 | 2 |
| Poli | 187 |
| Mir6357 | 0 |
| Mbd2 | 4308 |
| Dcc | 235 |
| Gm25509 | 0 |

*[Optional]*

We can concatenate all of the count data for the six samples into a single file. This will allow us to compare counts across each sample and each gene easily. This can then later be compared with the statistical results provided by DESeq2 .

- In the tools panel under “NGS: RNA Analysis”, click “Generate count matrix”.
- Select all of count data.
- Keep all options set to default values and click 


- View the newly generated file by clicking on the icon

***Part C2. Differential expression analysis with DESeq2***

DESeq2 performs differential expression analysis based on the negative binomial distribution. It takes **raw** counts as input and automatically performs normalisation and independent filtering.

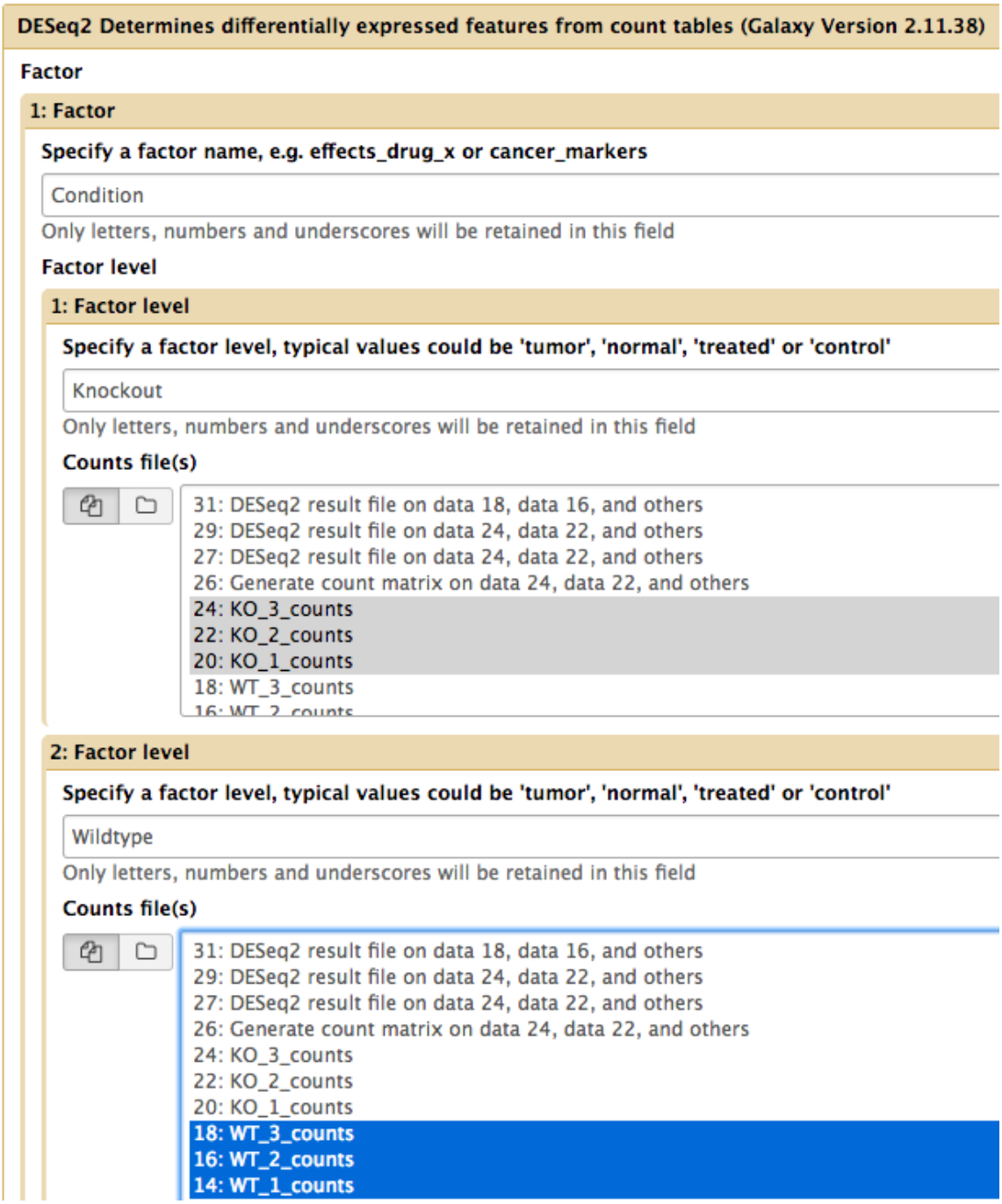
**Normalisation**: this process involves adjusting raw count data in order to make each gene and each sample comparable (for instance, making sample A comparable to sample B, even though it was sequenced to double the amount of coverage).
**Independent filtering:** this is performed to remove genes which have little chance at producing a significant result (for instance, genes will zero counts across all samples), in order to reduce the negative effect of multiple testing.

For further information, please see the DESeq2 paper and vignette.

**Using DESeq2 on Galaxy**

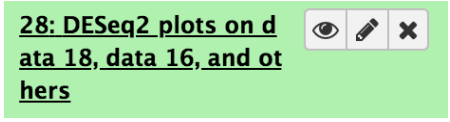
- In the tools panel, under “NGS: RNA Analysis”, click “DESeq2”
- Under “1. Factor”, “Specify a factor name, …” type “Condition”
- Under “1. Factor level” type “Knockout” and highlight all knockout count files
- Under “2. Factor level” type “Wildtype” and highlight all wildtype count files
- Leave all other settings as default and click Execute

**Note:** It is important to specify wildtype **last** so that it is used as a base level for our gene expression comparison between the two treatment groups.




DESeq2 will output two files: a “results” file and a “plots” file.

- Click the eye icon to view “DESeq2 plots…”




***DESeq2 plots***

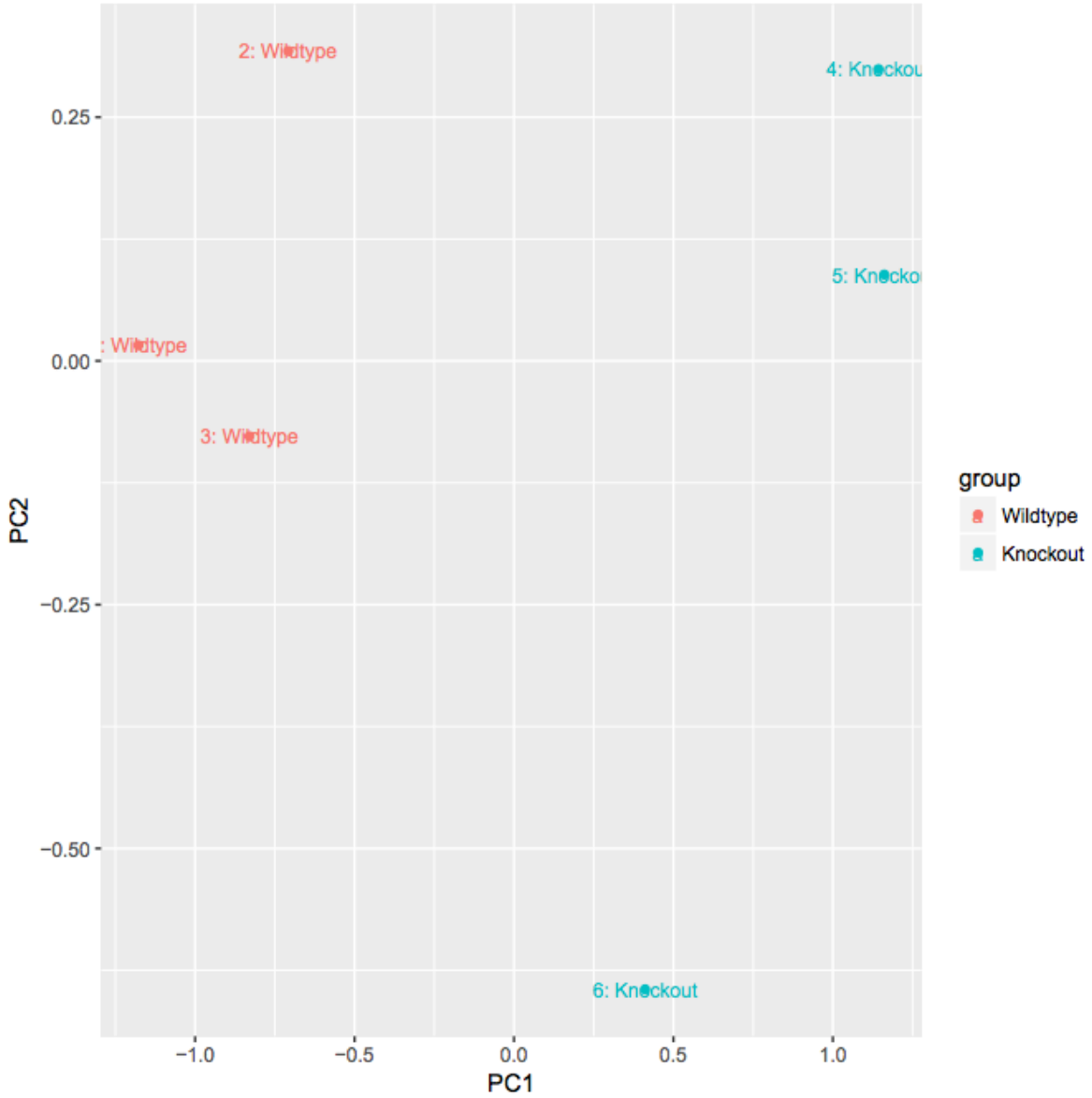
DESeq2 outputs five plots in Galaxy:

1. PCA plot
2. Sample-sample distances heatmap
3. Dispersion estimates
4. Histogram of p-values
5. MA-plot

We will briefly go through each of these here. Keep in mind that the data we have used here has been heavily reduced and that results should be interpreted accordingly.

***Principal components analysis (PCA) plot***

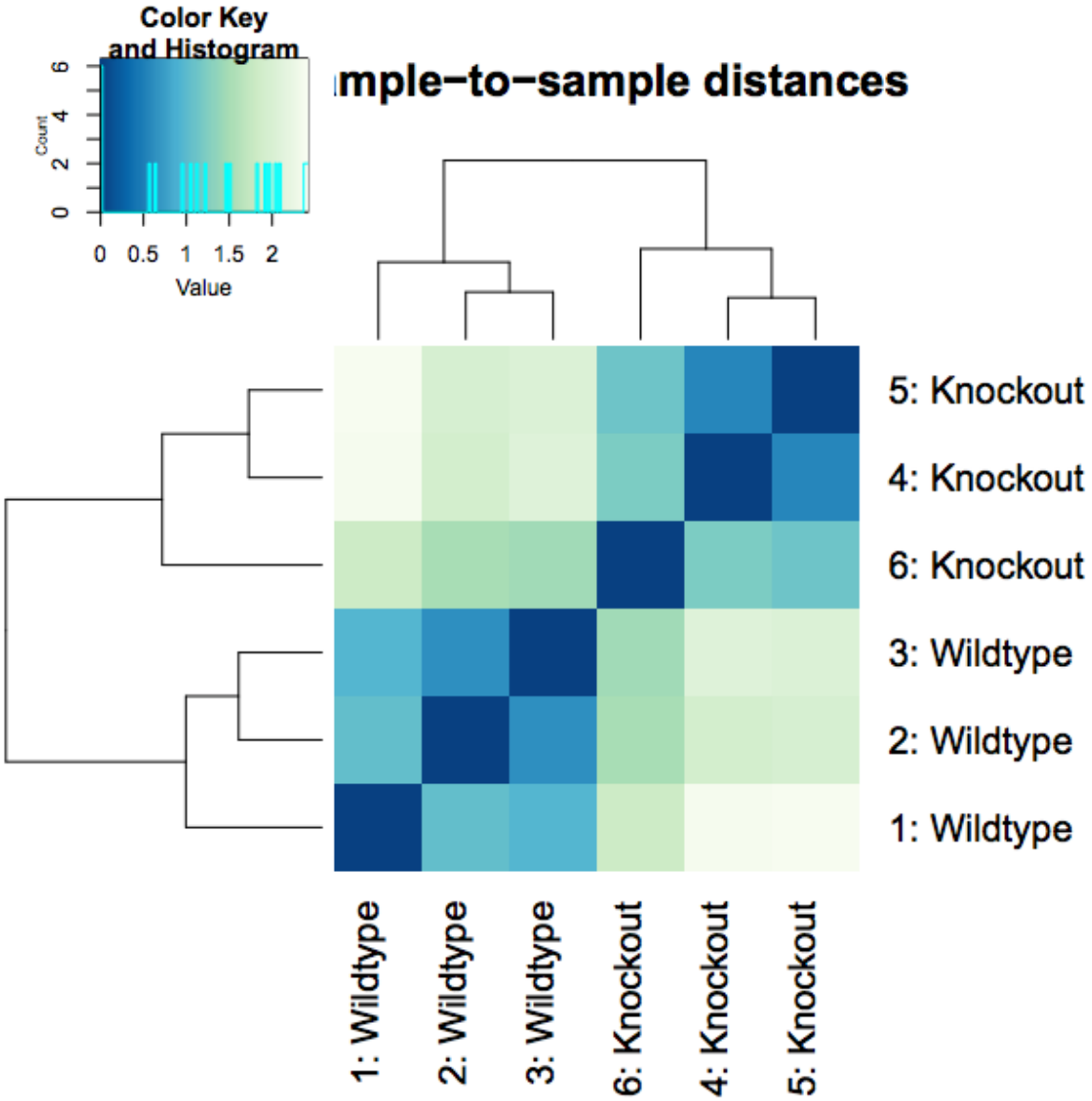
Using information from the count data, a principal component analysis can be used to assess the similarity between the samples. This information can be translated to a two-dimensional plot so that more similar samples are plotted closer together. The PCA plot is thus a useful plot in assessing whether your experiments ran as expected, or if there are potential issues with the sequencing data.




In general, our wildtype samples appear to cluster together well. In practice, knockout 6 may be a sample of concern as it appears to be an outlier and is clustering closer to wildtype 3.

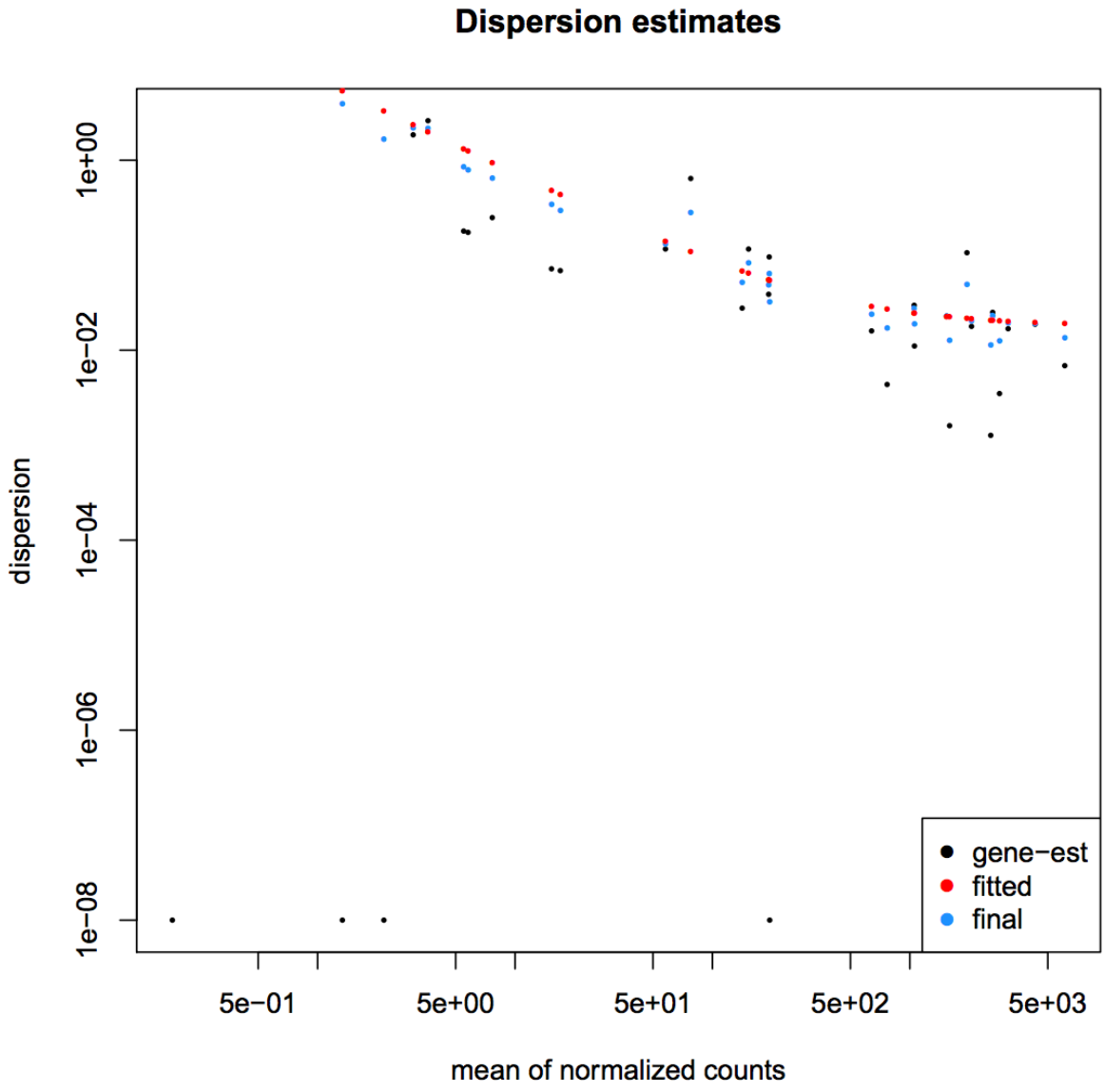
***Sample-sample distances heatmap***

The sample-sample distances heatmap provides a visualisation of similarities and differences between samples, similar to the PCA plot. From this plot, we can confirm that the samples cluster within treatment groups as expected, i.e. knockout samples cluster closer with each other than wildtype samples, and vice versa.



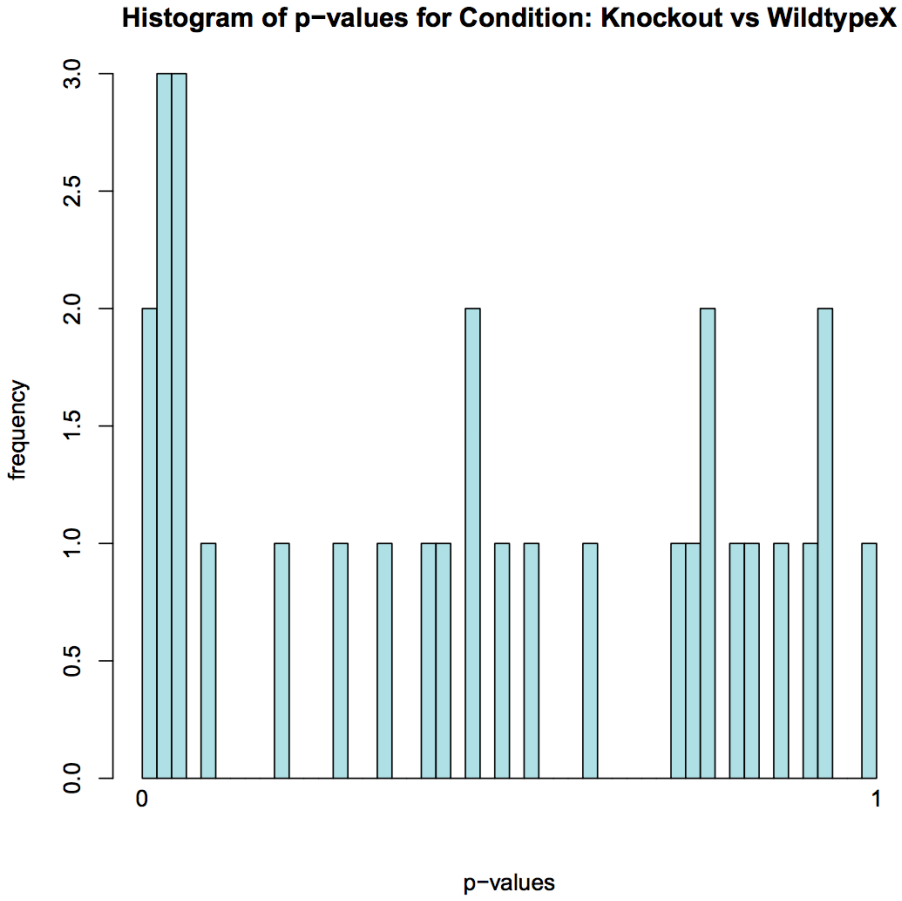
***Dispersion estimates***

RNA sequencing data is unique in that it contains extreme values and is heteroskedastic – there is a mean-variance dependency. In other words, genes with a lower mean average expression tend to have smaller variances than genes with a larger mean average expression which have higher variances. DESeq2 models this using the negative binomial distribution. The dispersion estimate plot represents how well this model fits the data where black dots represent the gene-wise estimates, the red line represents the fitted curve and the blue dots represent the final estimates (each gene's shrunken estimate). Heteroskedasticity is fitted via a 'parametric' fitting procedure. If this fitting is not appropriate, one can use a 'local' (local regression) or 'mean' (no apparent mean-variance dependence). See the DESeq2 vignette for more information.



### *Histogram of p-values*

Viewing the histogram of p-values can indicate issues with the design of the experiment or analysis. Distributions that don't fit the expected profile can indicate biases that have not been accounted for in the statistical model. As the data in this course has been heavily reduced, we will not be able to view a good representation of the histogram of p-values here, but for a thorough explanation on this please see the DESeq2 vignette.




Ideal plots from [Gonzalez, 2014] are below:

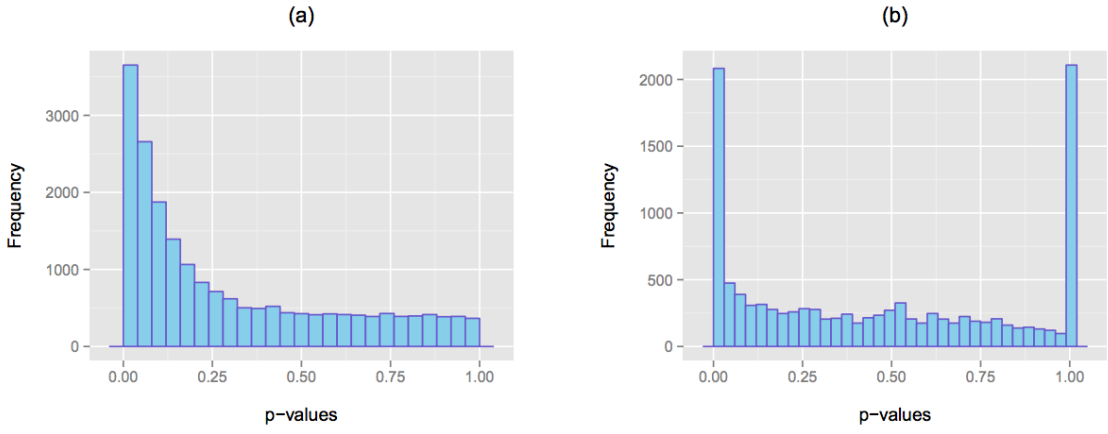
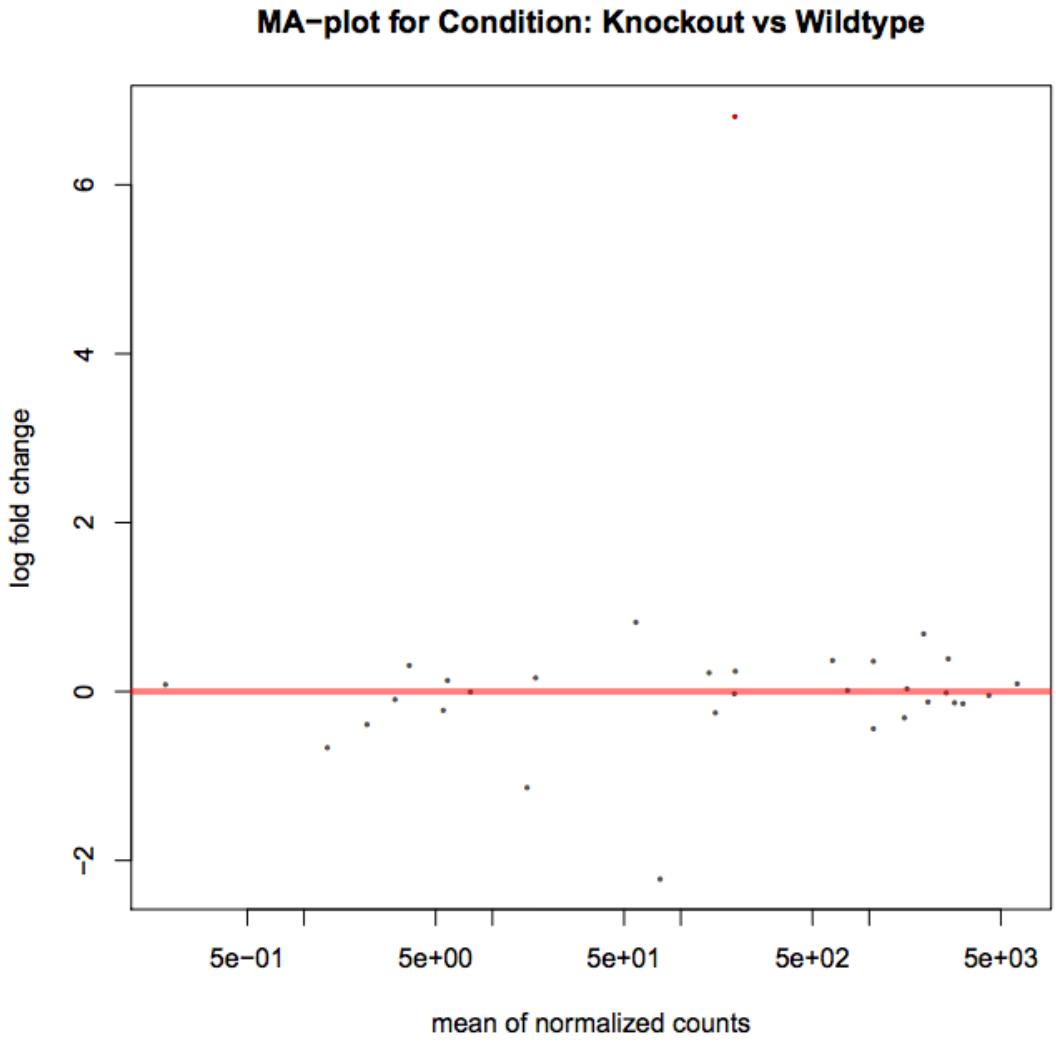



**Figure 19:** Desirable shape histograms of the p-values returned by the test for differential expression.

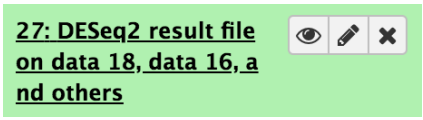
***MA plot***

MA plots represent log fold changes against the mean of normalised counts for each gene, where a gene is represented by a black dot (not significantly differentially expressed), or a red dot (significant). These can be used to visualise the minimum mean count required to obtain a significant P value. If the genes are not symmetrical around the x-axis, it can also potentially indicate issues with normalisation.



***DESeq2 results***

- Click the eye icon to view “DESeq2 result file…”




Amongst the plots, DESeq2 reports statistical results in the form of a table, where each row represents one gene and its corresponding statistical results for differential expression between the treatment and control groups.

The key statistics to look at here are the log2(FC) (i.e. the log fold change in log base 2 scale) and the P-adj value (adjusted for multiple testing using the Benjamini-Hochberg method). Please see the DESeq2 manual for descriptions on the other columns.

- The **P-adj value** indicates if expression is significantly different between treatment and control groups. A P-adj < 0.1 is considered as the significance threshold
- The **log2(FC)** indicates the magnitude of change between treatment and control groups. A positive value indicates upregulation and negative downregulation. If you have defined your groups in the order specified in the DESeq2 section, the wildtype group is your base level.

- From our results, how many/which genes are significantly differentially expressed?

| GeneID | Base mean | log2(FC) | StdErr | Wald-Stats | P-value | P-adj |
|---|---|---|---|---|---|---|
| Dcc | 193.907028695522 | 6.8136161026127 | 0.470827382593418 | 14.4715799346288 | 1.83214819175593e-47 | 5.49644457526779e-46 |
| Rab27b | 1946.79834588614 | 0.687071575831612 | 0.253388297309799 | 2.71153633820578 | 0.00669721998934165 | 0.100458299840125 |
| Ccdc68 | 57.8018403091394 | 0.828468116734194 | 0.405462048685272 | 2.04326920218683 | 0.041025796573053 | 0.163186623128358 |
| Me2 | 639.951862288326 | 0.373577095575733 | 0.184319201405252 | 2.0267942391654 | 0.0426834598679198 | 0.163186623128358 |
| Mapk4 | 15.2988070718808 | -1.13647835240699 | 0.562970054699193 | -2.01871901164306 | 0.0435164328342288 | 0.163186623128358 |
| Myo5b | 2627.5026675989 | 0.387973825052704 | 0.176792078134219 | 2.1945204171318 | 0.0281980177899421 | 0.163186623128358 |
| Lipg | 1055.21330831661 | 0.364361968542322 | 0.164079346162501 | 2.22064493224799 | 0.0263750199574691 | 0.163186623128358 |
| Smad7 | 1051.74424258594 | -0.437863390380022 | 0.194168457594231 | -2.25506962255971 | 0.0241289590300518 | 0.163186623128358 |
| Dym | 1536.17453328818 | -0.309980800215367 | 0.177122687177108 | -1.75009088420961 | 0.0801026324828614 | 0.267008774942871 |
| Gm23119 | 1.3348571411342 | -0.664959424137236 | 0.503950677698047 | -1.31949306462816 | 0.187004327203062 | 0.561012981609185 |
| Poli | 195.056496644244 | 0.239352186503826 | 0.221523292133726 | 1.08048315912232 | 0.279927082173054 | 0.763437496835603 |
| Mex3c | 2846.07385572655 | -0.130630808845863 | 0.13297236184622 | -0.982390679026486 | 0.325907416542936 | 0.81476854135734 |
| 4930503L19Rik | 152.417107402973 | -0.244105598914916 | 0.328084773406528 | -0.74403208774473 | 0.456857062882665 | 0.859354697491395 |

If you performed the optional task and have generated a count matrix, you can compare these statistical results to the raw count data, observing and comparing the counts for samples within each treatment group.

***Part C3. Functional annotation***

In this course, we only have one significantly differentially expressed gene. Uniprot Knowledgebase (UniprotKB) contains a database of functional information on proteins.

- Navigate to UniprotKB (https://www.uniprot.org/help/uniprotkb) or Google if you wish
- Search for the function of the significantly differentially expressed gene
- Does it match the disease of interest?

Typically, you will find many differentially expressed genes and may want to determine enriched biological pathways that are involved. By doing this, you can find out if your genes are associated with a particular biological process or molecular function.

The most popular methods of functionally annotating your differentially expressed genes are to classify them using Gene Ontology (GO) terms (http://geneontology.org/). The GO Consortium aims to provide an up to date, comprehensive database of knowledge on functions of genes and gene products in standardised nomenclature. They achieve this by classifying gene functions in a hierarchical manner – starting from board ontologies (molecular function, cellular component and biological process) to more specific terms.

Interactions between genes are commonly identified using the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database (http://www.genome.jp/kegg/pathway.html).

There are many software which can perform these analyses, including:

- GSEA (http://software.broadinstitute.org/gsea/index.jsp)
- DAVID (https://david.ncifcrf.gov/)
- PANTHER (http://www.pantherdb.org/)
- Ingenuity Pathway Analysis (IPA, QIAGEN proprietary software). Sydney University has paid for one license to be shared for **free** amongst USyd researchers. Please contact SIH if you would like access to IPA.

***Part D: Useful resources***

***DAVID***
https://david.ncifcrf.gov/

***DESeq2***
http://www.bioconductor.org/packages//2.13/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

***DESeq2 (Beginner's guide)***
https://bioc.ism.ac.jp/packages/2.14/bioc/vignettes/DESeq2/inst/doc/beginner.pdf

***Galaxy (Melbourne instance)***
https://galaxy-mel.genome.edu.au/galaxy

***Galaxy (University of Queensland instance)***
https://galaxy-qld.genome.edu.au/galaxy

***Gene Ontologies***
http://geneontology.org/

***GSEA***
http://software.broadinstitute.org/gsea/index.jsp

***HISAT2***
https://ccb.jhu.edu/software/hisat2/manual.shtml

***Ingenuity Pathway Analysis (also contact SIH for free access)***
https://www.qiagenbioinformatics.com/products/features/

***KEGG PATHWAY Database***
http://www.genome.jp/kegg/pathway.html

***PANTHER***
http://www.pantherdb.org/

***Statistical analysis of RNA-Seq data (a really great, comprehensive tutorial by Gonzalez, 2014)***
http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

***Sydney Informatics Hub – training courses***
https://informatics.sydney.edu.au/services/training/

***UniprotKB***
https://www.uniprot.org/help/uniprotkb/

**Articles referred to in the course:**

**The case study**

Corley SM, Canales CP, Carmona-Mora P, Mendoza-Reinosa V, Beverdam A, Hardeman EC, et al. RNA-Seq analysis of Gtf2ird1 knockout epidermal tissue provides potential insights into molecular mechanisms underpinning Williams-Beuren syndrome. BMC Genomics. 2016;17:450.
https://www.ncbi.nlm.nih.gov/pubmed/27295951

**Replicates in RNA sequencing studies**

Schurch NJ, Schofield P, Gierlinski M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22:839-51.
https://www.ncbi.nlm.nih.gov/pubmed/27022035

**Single-end versus paired-end reads, stranded versus non-stranded protocols**

Corley SM, MacKenzie KL, Beverdam A, Roddam LF & Wilkins MR. Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. BMC Genomics. 2017;18:399.
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3797-0

**Statistical design and Analysis of RNA Sequencing Data**

Auer PL & Doerge RW. Statistical Design and Analysis of RNA Sequencing Data. Genetics. 2010;2:405-416.
http://www.genetics.org/content/185/2/405#sec-6