

# Introduction to RNA-Seq on Galaxy

## Analysis for differential expression

**Tracy Chew & Cali Willet**

Senior Research Bioinformatics Technical Officer

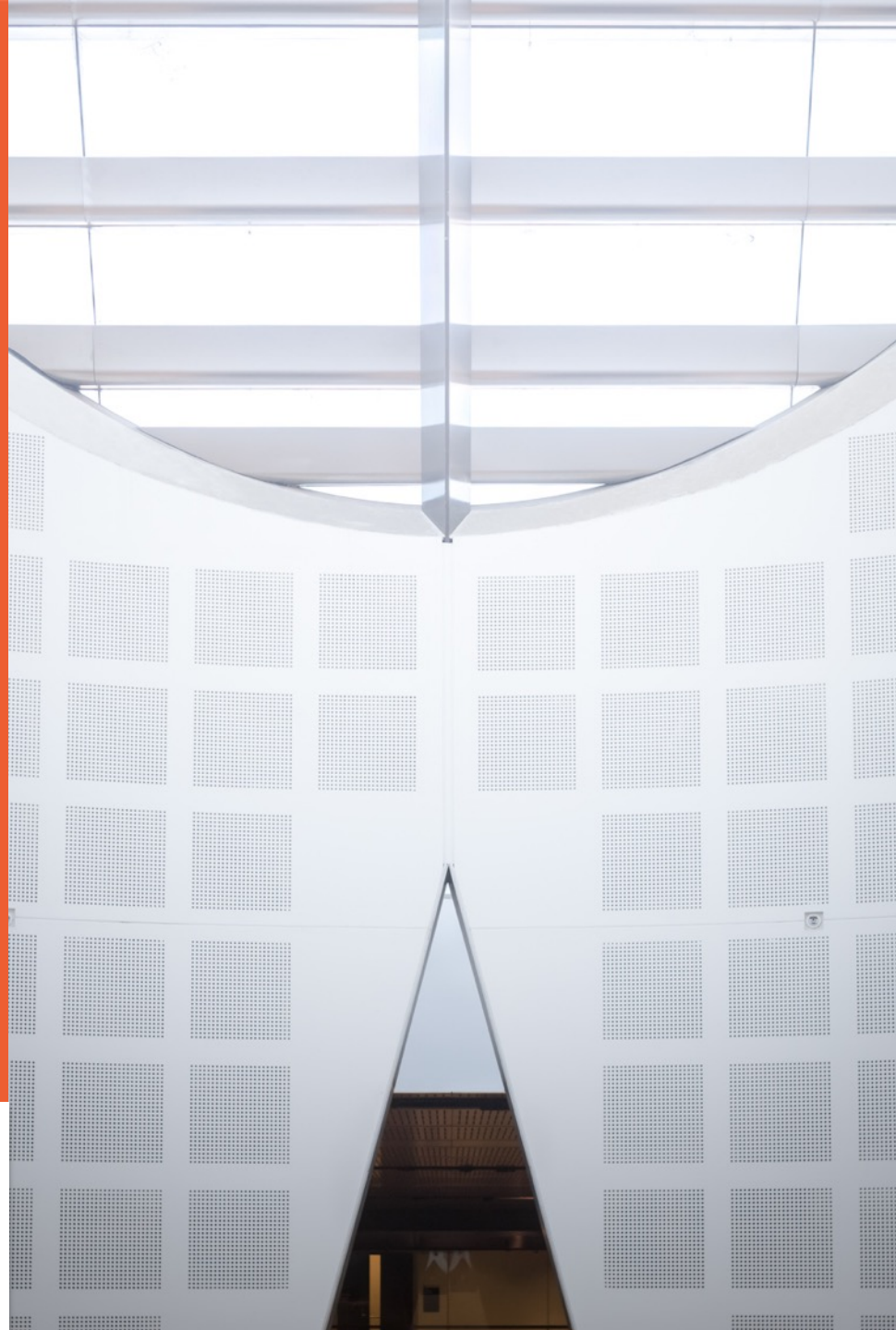
**Sydney Informatics Hub**

sih.info@sydney.edu.au

[https://usegalaxy.org.au:/u/tracy\\_chew/h/rna-seq-2019-1](https://usegalaxy.org.au:/u/tracy_chew/h/rna-seq-2019-1)



THE UNIVERSITY OF  
**SYDNEY**  
—  
Sydney  
Informatics Hub



# Course outline

## Part A: Introduction

- Why sequence RNA?
- How is the transcriptome sequenced?
- Experimental design considerations
- Analysis workflow overview

## Part B: Raw data and quality checking

- Uploading data on Galaxy
- Quality checking with fastQC and multiQC
- Trimming

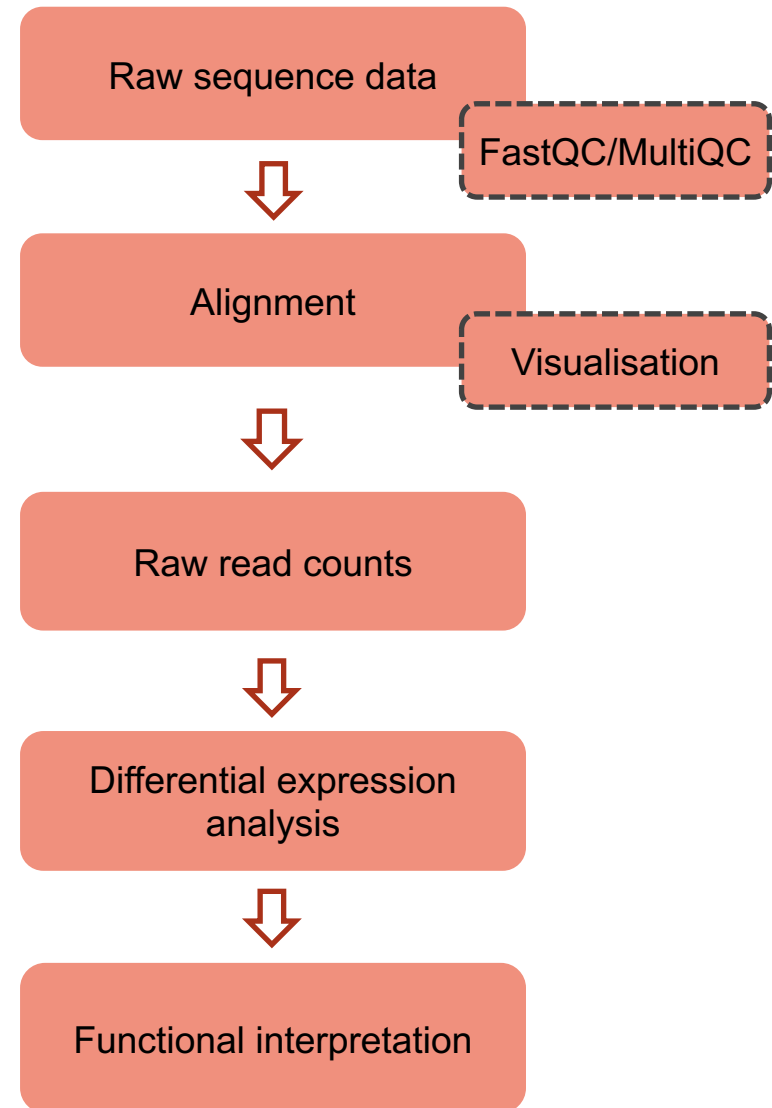
## Part C: Alignment and Visualisation

- Alignment with HISAT2
- Visualisation in IGV

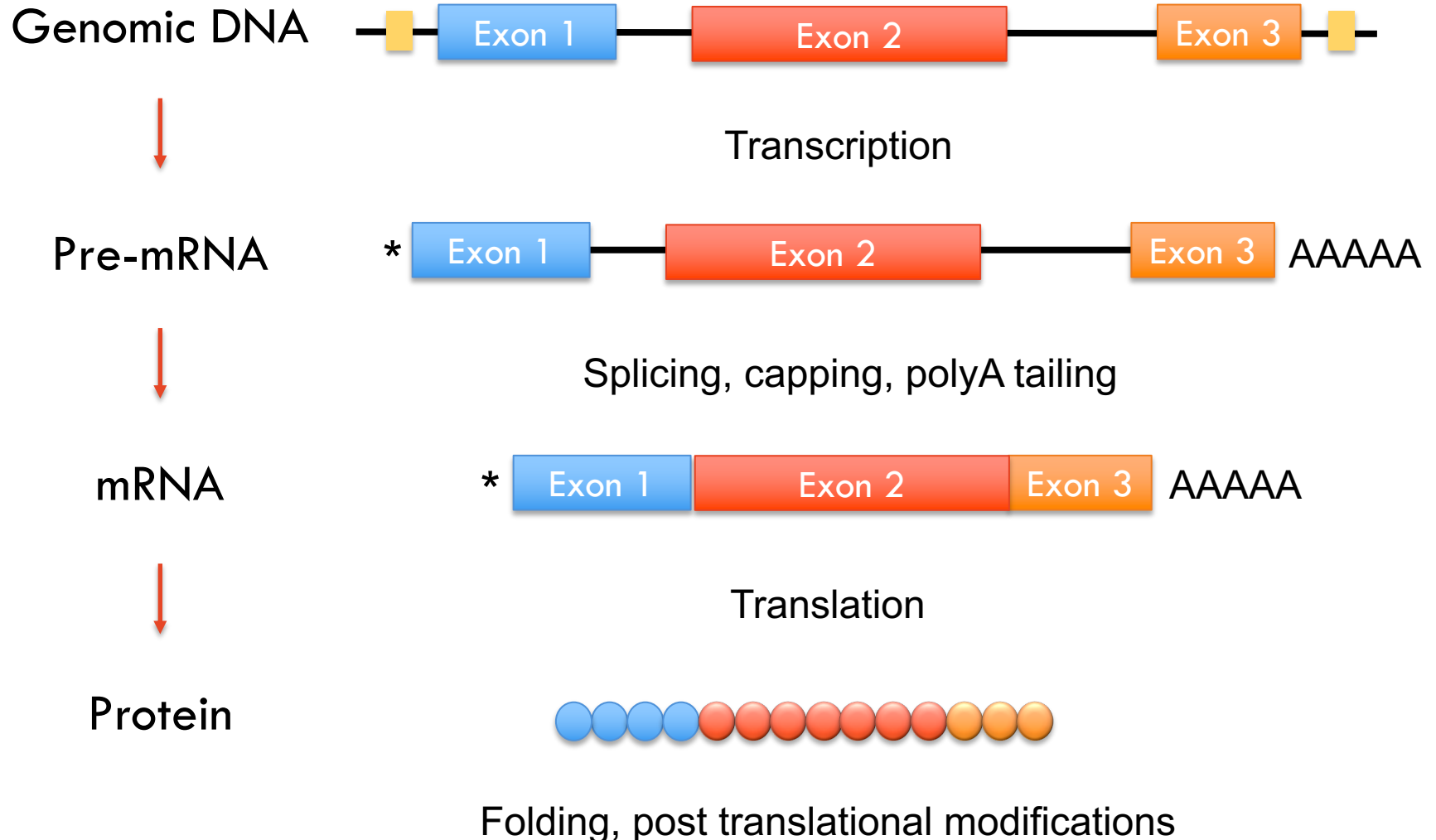
## Part D: Differential Expression Analysis

- Obtaining count data with featureCounts
- DESeq2
- Functional annotation

## Part E: Useful resources

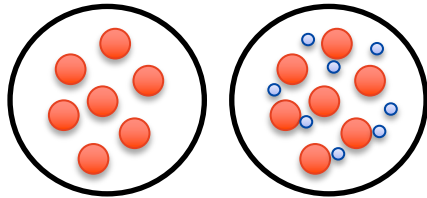


## Part A: Why sequence RNA?



# Part A: How does RNA sequencing work?

Experimental design



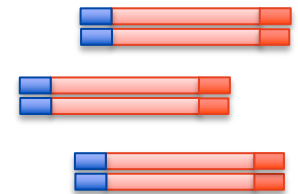
untreated

treated

Isolate RNA

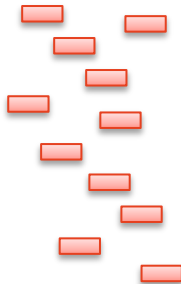


Prepare library

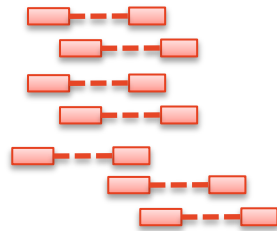


Sequence

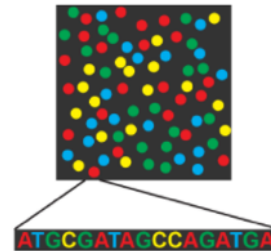
Single reads



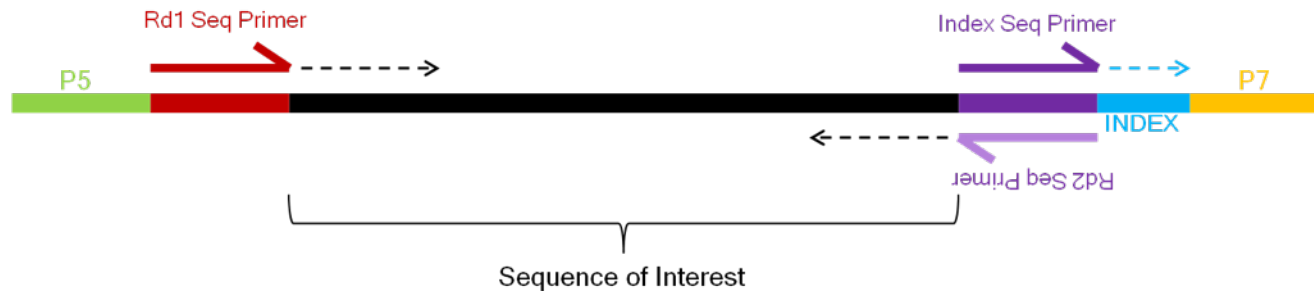
Paired end reads



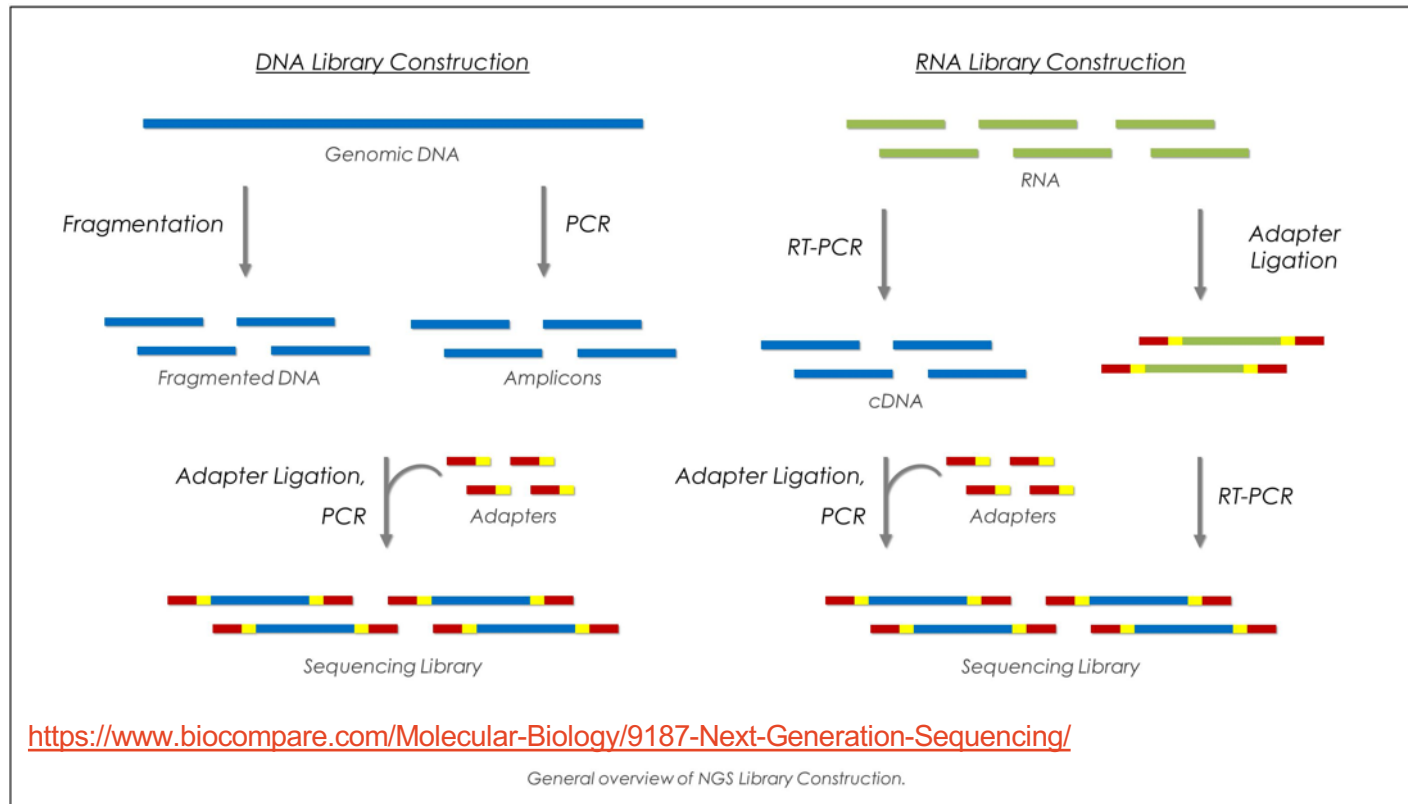
FASTQ files



# Part A: How does RNA sequencing work?



<http://nextgen.mgh.harvard.edu/IlluminaChemistry.html>



## Part A: Experimental Design

Want design to be able to give you results that are statistically sound and provide you with answers to your experimental questions.

**Replicates:** Technical vs Biological

**Data amount/type:** Read length, single vs paired end, stranded vs unstranded, desired depth of coverage

# Part A: Replicates and protocols

[RNA](#). 2016 Jun;22(6):839-51. doi: 10.1261/ma.053959.115. Epub 2016 Mar 28.

## How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

[Schurch NJ](#)<sup>1</sup>, [Schofield P](#)<sup>2</sup>, [Gierliński M](#)<sup>2</sup>, [Cole C](#)<sup>1</sup>, [Sherstnev A](#)<sup>1</sup>, [Singh V](#)<sup>3</sup>, [Wrobel N](#)<sup>4</sup>, [Gharbi K](#)<sup>4</sup>, [Simpson GG](#)<sup>5</sup>, [Owen-Hughes T](#)<sup>3</sup>, [Blaxter M](#)<sup>4</sup>, [Barton GJ](#)<sup>6</sup>.

⊕ Author information

### Erratum in

Erratum: How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? [[RNA](#). 2016]

## Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge

*GENETICS* June 1, 2010 vol. 185 no. 2 405-416; <https://doi.org/10.1534/genetics.110.114983>

## 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

[Susan M. Corley](#) ✉, [Karen L. MacKenzie](#), [Annemiek Beverdam](#), [Louise F. Roddam](#) and [Marc R. Wilkins](#)

*BMC Genomics* 2017 18:399

<https://doi.org/10.1186/s12864-017-3797-0> | © The Author(s). 2017

Received: 16 December 2016 | Accepted: 16 May 2017 | Published: 23 May 2017

# Part B: Analysis in Galaxy

<https://usegalaxy.org.au/>

The screenshot displays the Galaxy Australia web interface. The top navigation bar includes links for Analyze Data, Workflow, Visualize, Shared Data, Help, and User. A red arrow points to the 'User' dropdown menu. The interface is divided into three main panels:

- Tools Panel (Left):** Titled 'Tools', it features a search bar and a list of tool categories: FILE AND META TOOLS (Get Data, Send Data, Collection Operations), GENERAL TEXT TOOLS (Text Manipulation, Filter and Sort, Join, Subtract and Group), GENOMIC FILE MANIPULATION (FASTA/FASTQ, FASTQ Quality Control, SAM/BAM, BED, VCF/BCF, Convert Formats), COMMON GENOMICS TOOLS (Operate on Genomic Intervals, Extract Features, Fetch Sequences/Alignments), and GENOMICS ANALYSIS (Assembly).
- Viewing Panel (Center):** Titled 'Viewing Panel', it displays the Galaxy AUSTRALIA logo and two columns of news and events. The 'News' column lists updates such as 'Galaxy Australia upgraded to Galaxy version 19.05' and 'Text processing tools disabled'. The 'Events and Workshops' column lists upcoming events like the '2019 Galaxy Community Conference (GCC2019)' and 'Galaxy training workshops Brisbane - April 2019'. At the bottom, a 'Galaxy Australia Jobs (Last 12 hours)' section is partially visible.
- History Panel (Right):** Titled 'History', it shows a search bar for datasets and indicates that the history is currently empty. A message states: 'This history is empty. You can load your own data or get data from an external source'.



## Part B: The study

Knockout mouse model to study **Williams-Beuren Syndrome (WBS)**, a rare disease found in people

- distinctive facial features
- intellectual disability
- cardiovascular abnormalities

It is caused by a disruption in the **Gtf2ird1** gene

**RNA-Seq analysis of *Gtf2ird1* knockout epidermal tissue provides potential insights into molecular mechanisms underpinning Williams-Beuren syndrome**

Susan M. Corley , Cesar P. Canales, Paulina Carmona-Mora, Veronica Mendoza-Reinosa, Annemiek Beverdam, Edna C. Hardeman, Marc R. Wilkins and Stephen J. Palmer

*BMC Genomics* 2016 17:450

<https://doi.org/10.1186/s12864-016-2801-4> | © The Author(s). 2016

Received: 2 November 2015 | Accepted: 26 May 2016 | Published: 13 June 2016

## Part B: The study

To improve our understanding of this disease, Corley et al. 2016 created a knockout mouse model of this disease.

Wildtype



Knockout



**Which genes (if any) are upregulated or downregulated in our knockout mice and how do these relate to the disease phenotype?**

## Part B: Uploading data

Raw sequence files are sent in **FASTQ** format. In practice, download and store these in a safe place such as the **Research Data Store** systems provided by the University

- Copy the links to the FASTQ files

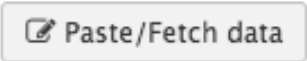
<https://informatics.sydney.edu.au/training/coursedocs/SRR3473984.fastq>  
<https://informatics.sydney.edu.au/training/coursedocs/SRR3473985.fastq>  
<https://informatics.sydney.edu.au/training/coursedocs/SRR3473986.fastq>  
<https://informatics.sydney.edu.au/training/coursedocs/SRR3473987.fastq>  
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

- Go back to Galaxy
- Click the upload icon





## Part B: Uploading data

A white box should appear.

- Click 
- Paste the links in the box that appears
- Change “Type” to “fastqsanger”
- Do the same for the annotation file, except leave “Type” as “Auto-detect”

[https://informatics.sydney.edu.au/training/coursedocs/Mus\\_musculus.GRCm38.chr18region.gtf](https://informatics.sydney.edu.au/training/coursedocs/Mus_musculus.GRCm38.chr18region.gtf)

## Part B. Uploading data

Name	Size	Type	Genome	Settings	Status
 New File	426 b	fastqsanger ▼	Q	unspecified (?) ▼	⚙️ 0% 🗑️
<p>You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.</p> <p><a href="https://informatics.sydney.edu.au/services/coursedocs/SRR3473988.fastq">https://informatics.sydney.edu.au/services/coursedocs/SRR3473988.fastq</a> <a href="https://informatics.sydney.edu.au/services/coursedocs/SRR3473989.fastq">https://informatics.sydney.edu.au/services/coursedocs/SRR3473989.fastq</a></p>					
 New File	90 b	Auto-det... ▼	Q	unspecified (?) ▼	⚙️ 0% 🗑️
<p>You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.</p> <p><a href="https://informatics.sydney.edu.au/services/coursedocs/Mus_musculus.GRCm38.chr18region.gtf">https://informatics.sydney.edu.au/services/coursedocs/Mus_musculus.GRCm38.chr18region.gtf</a></p>					

- Click 
- You may now close the upload box

### In the history panel

Grey panels – in queue

Yellow – running

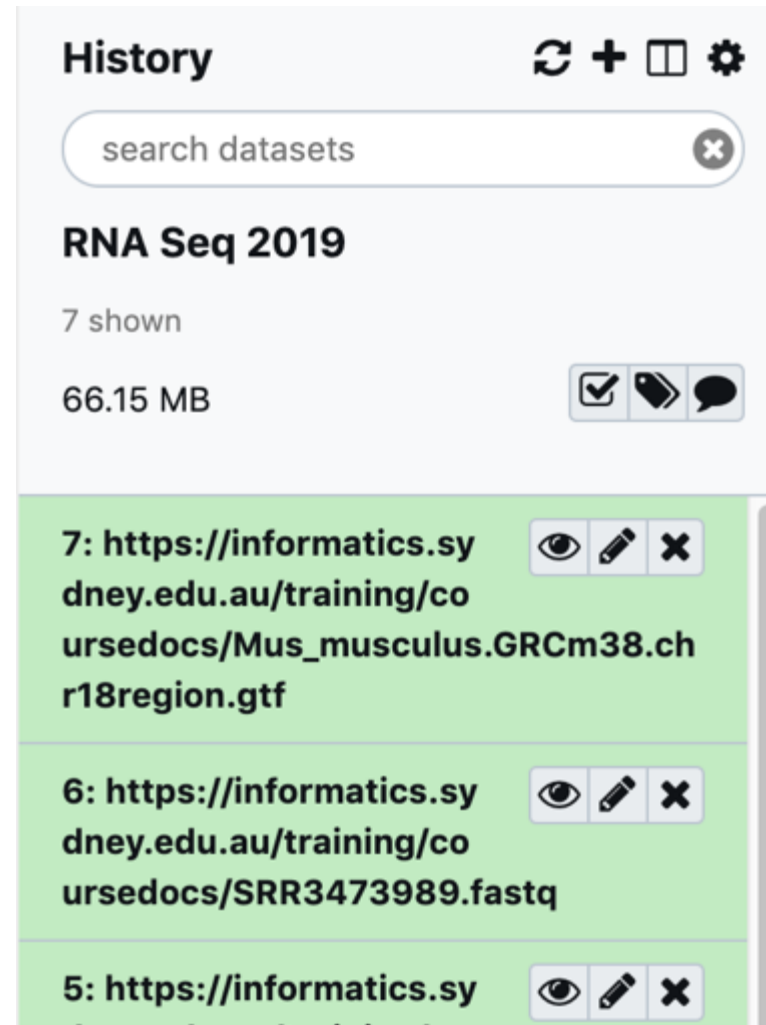
Green – job has completed

Red – job has failed

## Part B. Uploading data

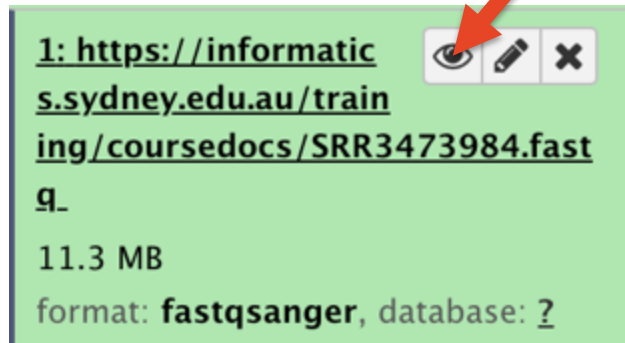
Your "upload job" will be submitted to the Galaxy server.

When it is complete, it will appear in green in your history pane.



## Part B: FASTQ files

Inspect your FASTQ files by clicking on the eye icon



@SRR3473984.R.33538298

GGAATTCTGTGGTCACTGTTACAGTTGTCATGGTGACTTCTGGCTTGGAGGGCGCTCAGAGGAGGCCTCCTCCGCCTGCTCCTGCTCGGGCTCCGGCGAT

+

CCCCFFFFHHFHHGIIJJJHIEIJJGJHIGIIJJHFGIJJJIIJJIIJJDFGGGGHGGGHIHFCHF?DCEBACCB>??A?CCDC@>3:29BB<?@99<555

Line 1	@ followed by sequence identifier. Usually contains some sequencing and pair membership information (e.g. @HWUSI-EAS100R:6:73:941:1973#0/1)
Line 2	Raw sequence
Line 3	+ optionally followed by sequence identifier/description
Line 4	Quality values for line 2 encoded in ASCII (usually Phred+33)

## Part B: Phred Quality Scores

Using corresponding 'Dec' values in this ASCII table (<http://www.asciitable.com/>), what is the Phred quality score of the first base in read provided in SRR3473984.fastq?

Phred quality scores  $Q$  are defined as a property which is logarithmically related to the base-calling error probability  $P$ .

$$Q = -10 \log_{10} P$$

or

$$P = 10^{\frac{-Q}{10}}$$

Now that you have determined the Phred score for the first base in the previous example, what is the probability ( $P$ ) that this base was incorrectly called by the sequencing machine?

Why do we use ASCII to encode quality scores?



## Part B: FASTQ files






```
S - Sanger          Phred+33,  raw reads typically (0, 40)
X - Solexa          Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+   Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+   Phred+64,  raw reads typically (3, 41)
      with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
      (Note: See discussion above).
L - Illumina 1.8+   Phred+33,  raw reads typically (0, 41)
```

## Part B: FastQC


- In the Tools panel (left), click FastQC under FASTQ Quality Control (or search for fastqc)
- Click on the multiple datasets icon and select all fastq files
- Leave other options as default and click execute


FastQC Read Quality reports (Galaxy Version 0.72) ☆ Favorite 🔄 Versions ▼ Options

Short read data from your current history

6: <https://informatics.sydney.edu.au/training/coursedocs/SRR3473989.fastq>  
5: <https://informatics.sydney.edu.au/training/coursedocs/SRR3473988.fastq>  
4: <https://informatics.sydney.edu.au/training/coursedocs/SRR3473987.fastq>  
3: <https://informatics.sydney.edu.au/training/coursedocs/SRR3473986.fastq>  
2: <https://informatics.sydney.edu.au/training/coursedocs/SRR3473985.fastq>  
1: <https://informatics.sydney.edu.au/training/coursedocs/SRR3473984.fastq>

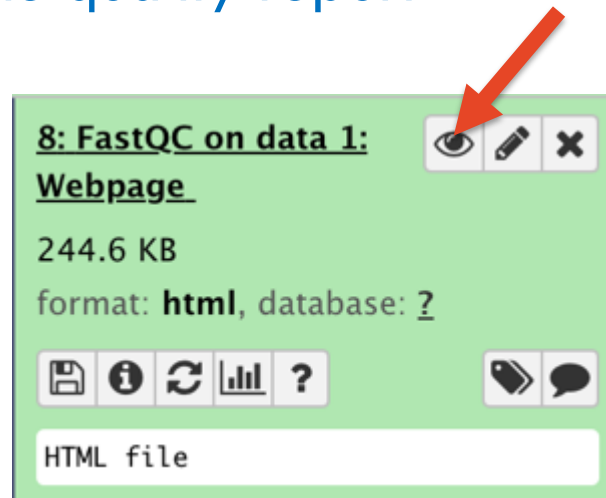


 This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

More info about the tool can be found at the bottom of the screen.

## Part B: FastQC

- For each fastq file, FastQC will create a “Webpage” and “RawData” output
- Click on the eye icon for the SRR3473984.fastq webpage output to view the quality report



The authors of FASTQC have provided a description of each category.

## Part B: FastQC

 passed QC

 failed QC

 warning

How many sequences were in SRR3473984.fastq?

What are the lengths of the reads in SRR3473984.fastq?

Which part of the reads tend to have worse per base sequence quality?

When inspecting FastQC reports for RNA seq data, you may notice the categories “Per base sequence content” and “Sequence Duplication Levels” marked as failing QC. Why is this?

# Part B: MultiQC

MultiQC  
v1.5

General Stats

FastQC

Sequence Quality Histograms

Per Sequence Quality Scores

Per Base Sequence Content

Per Sequence GC Content

Per Base N Content

Sequence Length Distribution

Sequence Duplication Levels

Overrepresented sequences

Adapter Content

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2018-10-22, 14:55 based on data in: /project/RDS-FMH-scRNAseqSepsis-RW/RNASEQ\_050918\_1/fastQC\_trimmed

Welcome! Not sure where to start?

Watch a tutorial video (6:06)

don't show again ✕

General Statistics

Copy table

Configure Columns

Plot

Showing 112/112 rows and 3/5 columns.

Sample Name	% Dups	% GC	M Seqs
RNASEQ_050918_1_HC-025_1	73.1%	42%	16.6
RNASEQ_050918_1_HC-027_2	51.3%	43%	4.8
RNASEQ_050918_1_HC-036_4	62.8%	47%	52.8
RNASEQ_050918_1_HC-037_1	62.2%	43%	19.6
RNASEQ_050918_1_HC-038_7	58.4%	43%	19.9
RNASEQ_050918_1_HC-039_8	47.3%	44%	14.5

The University of Sydney

Page 21

## Part B: Trimming

Trimming is sometimes performed to improve the quality of the raw data and potentially improve its mappability. There are several ways to perform trimming:

- Removal of poor quality reads or bases (e.g. ends of reads)
- Removal of adapter sequences
- Removal of polyA tails

**Be very wary about trimming RNA sequencing data**

Trimming of poor quality reads can affect gene expression estimates ([Williams et al., 2016](#))

Trimming of high quality adapter sequences was shown to increase quality and reliability of biological signals in RNA-seq data ([Dozmorov et al., 2015](#))

## Part C: Alignment with HISAT2

- In the tools panel, click “RNA-seq”
- Click HISAT2
- Input a reference genome
  - Source for the reference genome: Use a built-in genome
  - Select a reference genome: Mouse (*Mus Musculus*) mm10
- Input your reads
  - Single-end or paired-end reads?: Single-end
  - FASTA/Q: Click on the multiple datasets icon, highlight all six fastq files
- Specify strand information
  - Leave as unstranded
- Click execute

## Part C: Stranded vs Unstranded

There are two types of RNA sequencing sample preparation protocols: stranded and unstranded. It is important to know which you have in the downstream analysis.

Stranded protocols retain strandedness information (whether your RNA was transcribed from the forward or reverse strand). Unstranded protocols do not retain this information.

Mammalian genomes have many overlapping genes...e.g. BDNF locus in humans (UCSC hg19 genome browser coordinates chr11:27,671,365-27,684,616)



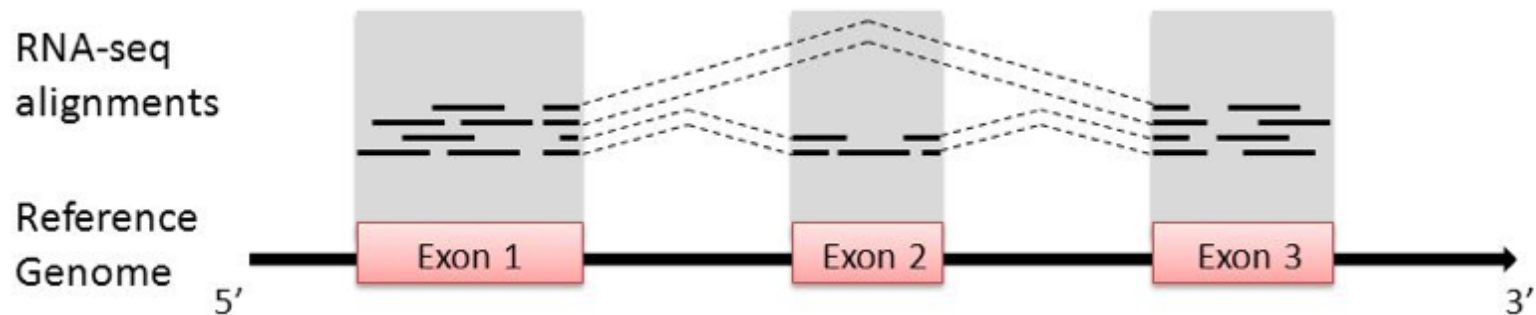
**Tip:** In the tools panel under RNA Analysis, you can use Infer Experiment to determine whether RNA sequencing was strand specific (forward/reverse) or unstranded.



# Part C: Alignment with HISAT2

## Mapping to a reference genome

- Allows transcript discovery (better with paired end data)
- Variant calling



## Part C: Visualisation with IGV

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

In the next task, we will use the Integrated Genomics Viewer (IGV) to visualise our alignments and assign samples to their correct treatment group (wildtype or knockout)

## Part C: The study

The key to this is:

**“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”**

SRR3473984.fastq  
SRR3473985.fastq  
SRR3473986.fastq  
SRR3473987.fastq  
SRR3473988.fastq  
SRR3473989.fastq

?

Wildtype



Knockout



## Part C: IGV

Go to: <http://software.broadinstitute.org/software/igv/download>

Click on the Launch button that is relevant to your machine. You may be asked to install the most recent version of Java.

Did you know that there is also an **IGV web application** that runs only in a web browser, does not use Java, and requires no downloads? See <https://igv.org/app>. Click on the [Help](#) link in the app for more information about using IGV-Web.

### Install IGV 2.5.x



#### IGV Mac App

Download and unzip the Mac App Archive, then double-click the IGV application to run it. You can move the app to the *Applications* folder, or anywhere else.



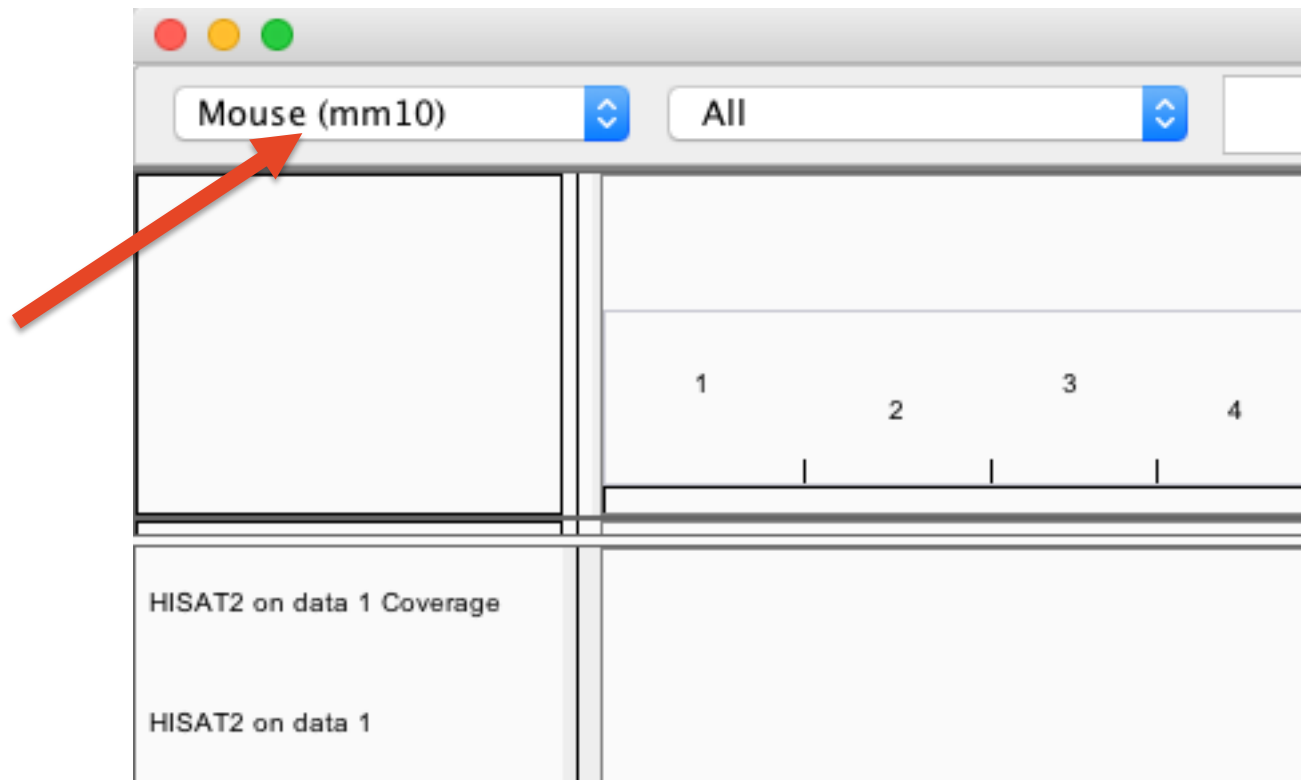
#### IGV for Windows

Download and run the installer. An IGV shortcut will be created on the Desktop; double-click it to run the application.

You can also download the bams/index files and upload these to the IGV web application: <https://igv.org/app>

## Part C: Visualisation with IGV

- Change the reference genome to “Mouse (mm10)”






## Part C: Visualisation with IGV

- Go back to Galaxy
- Click on “HISAT2 on data 1: aligned reads (BAM)”








Notice that our aligned files are in “BAM” format. This is in binary SAM format (if you click on the eye icon – you are actually viewing the SAM format). Also notice alignment stats provided.

- Click on “local”
- Your bam file should load in IGV

**20: HISAT2 on data 1: aligned reads (BAM)**   

3.3 MB  
format: **bam**, database: ?

51547 reads; of these:  
51547 (100.00%) were unpaired;  
of these:  
29 (0.06%) aligned 0 times  
47719 (92.57%) aligned exactly 1 time  
3799 (7.37%) aligned >1 times  
99.94% overall alignment rate

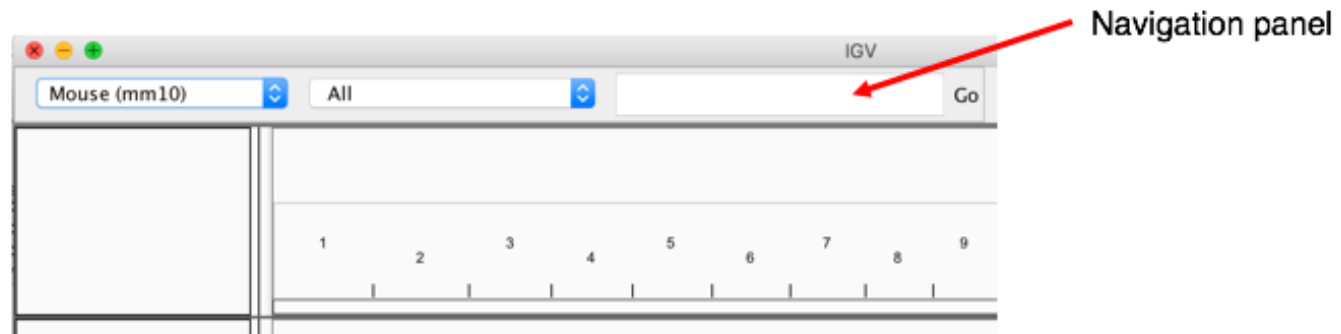
      

**display with IGV local**  
display in IGB View

Binary bam alignments file

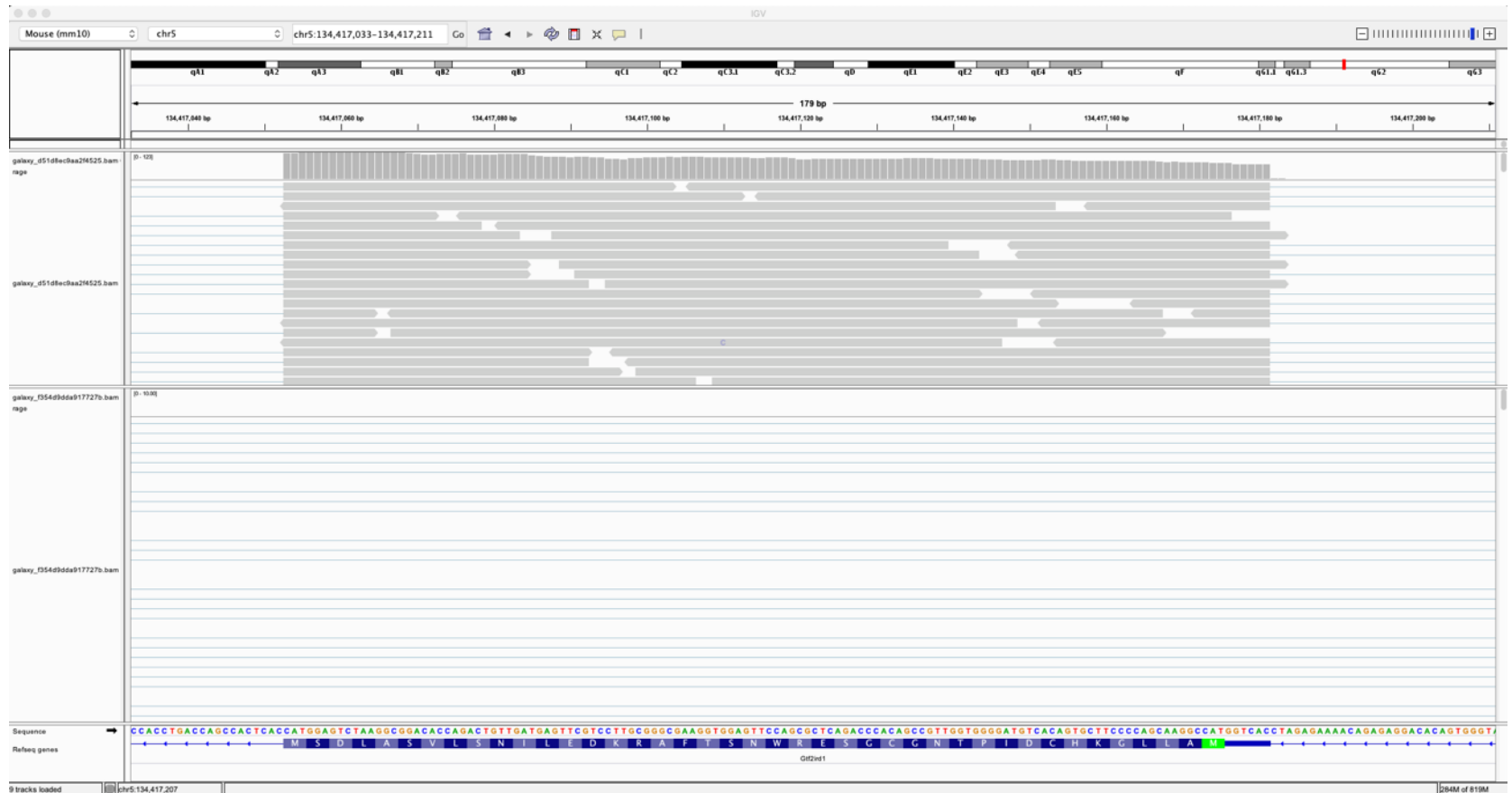
## Part C: Visualisation with IGV

- Let's open another 2 alignments
- Go back to Galaxy
- Navigate to another BAM file (e.g. “HISAT2 on data 2” and “HISAT2 on data 6”)
- Click “local”
- Practice navigating
- Navigate to **Gtf2ird1**: **chr5:134,332,897-134,481,480**




# Part C: Visualisation with IGV

Key is to look at exon 1...





## Part C: Renaming .bam files on Galaxy

- Once you have identified your samples, rename them to something more meaningful
- Click on the edit attributes button  next to your sample bam file (“HISAT2 on data ...”)
- Type in the new file name under “Name:”
- Click save

≡ Attributes ⚙️ Convert 📄 Datatypes 👤 Permissions

Edit attributes

↺ Auto-detect 💾 Save

**Name**

## Part D: Differential expression – count data

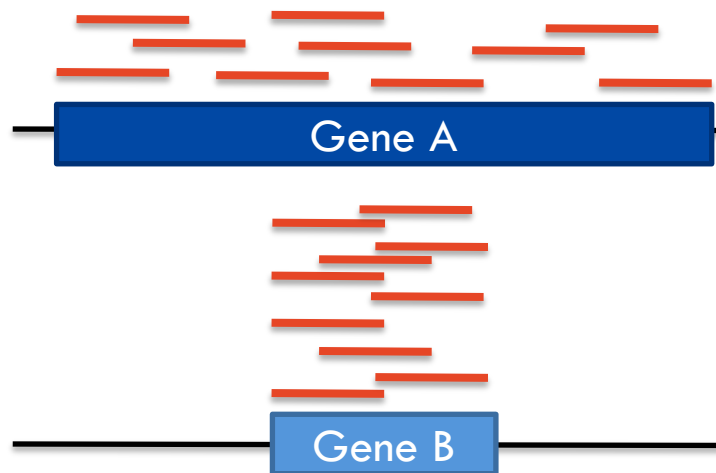
We are now ready to obtain raw count data. We want to count the number of reads that fall within each gene.

We will need an annotation file (GTF/GFF3) that tells us where the genes are located in the genome. We uploaded this file earlier (`Mus_musculus.GRCm38.chr18region.gtf`).

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Group
18	ensembl_havana	gene	69925426	69969484	.	+	.	gene_id "ENSMUSG00000"
18	havana	transcript	69925426	69944044	.	+	.	gene_id "ENSMUSG00000"
18	havana	exon	69925426	69925533	.	+	.	gene_id "ENSMUSG00000"
18	havana	exon	69938842	69938970	.	+	.	gene_id "ENSMUSG00000"
18	havana	exon	69940100	69940186	.	+	.	gene_id "ENSMUSG00000"

## Part D: Obtaining raw counts with featureCounts

- In the tools panel, under RNA-seq, click on featureCounts
- Alignment file
  - Click the multiple datasets icon and highlight all six bam files
- Gene annotation file
  - In your history
  - Gene annotation file: select the GTF file we uploaded earlier
- In “Advanced options” change “GFF gene identifier” to “gene\_name”
- Click execute



## Something else to be wary of...

Sample 1 has twice as many reads  
at gene A than sample 2.

The average coverage in sample 1  
is twice the amount as it is for sample 2.

**Is the expression for gene A higher for sample  
1 than sample 2?**

## Part D: Count data

- featureCounts outputs two files per sample: “counts” and “summary”. Carefully delete “summary” files to keep things tidy
- Observe the count data
- Rename the data to something more meaningful (e.g. “WT\_1\_counts”)

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

## Part D: Differential expression analysis with DESeq2

We are now ready to perform statistical testing to see which genes have significant differential expression between treatment groups.

- In the tools panel under RNA-seq, click on DESeq2
- Name “Condition” as your Factor
- Input wildtype and knockout count data as separate factors
- Specify wildtype **last** so that it is used as the base level
- Leave everything else as default, click Execute

The screenshot shows the DESeq2 web interface with the following configuration:

- 1: Factor**
  - Specify a factor name, e.g. effects\_drug\_x or cancer\_markers: Condition
  - 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': Knockout
    - Only letters, numbers and underscores will be retained in this field
    - Counts file(s)**
      - 36: KO\_3\_counts
      - 34: KO\_2\_counts
      - 32: KO\_1\_counts
      - 30: WT\_3\_counts
      - 28: WT\_2\_counts
      - 26: WT\_1\_counts
      - 7: [https://informatics.sydney.edu.au/training/coursedocs/Mus\\_musculus.GRCm38.chr18region.gtf](https://informatics.sydney.edu.au/training/coursedocs/Mus_musculus.GRCm38.chr18region.gtf)
- 2: Factor level**
  - Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control': Wildtype

## Part D: DESeq2 output files

DESeq2 produces two output files:

1. A "DESeq2 plots ..." pdf file containing 5 plots
  - Principal components analysis plot (PCA plot)
  - Sample-sample distances heatmap
  - Dispersion estimates
  - Histogram of p-values
  - MA plot
2. A "DESeq2 results..." file containing statistical results

Let's observe the plots first.

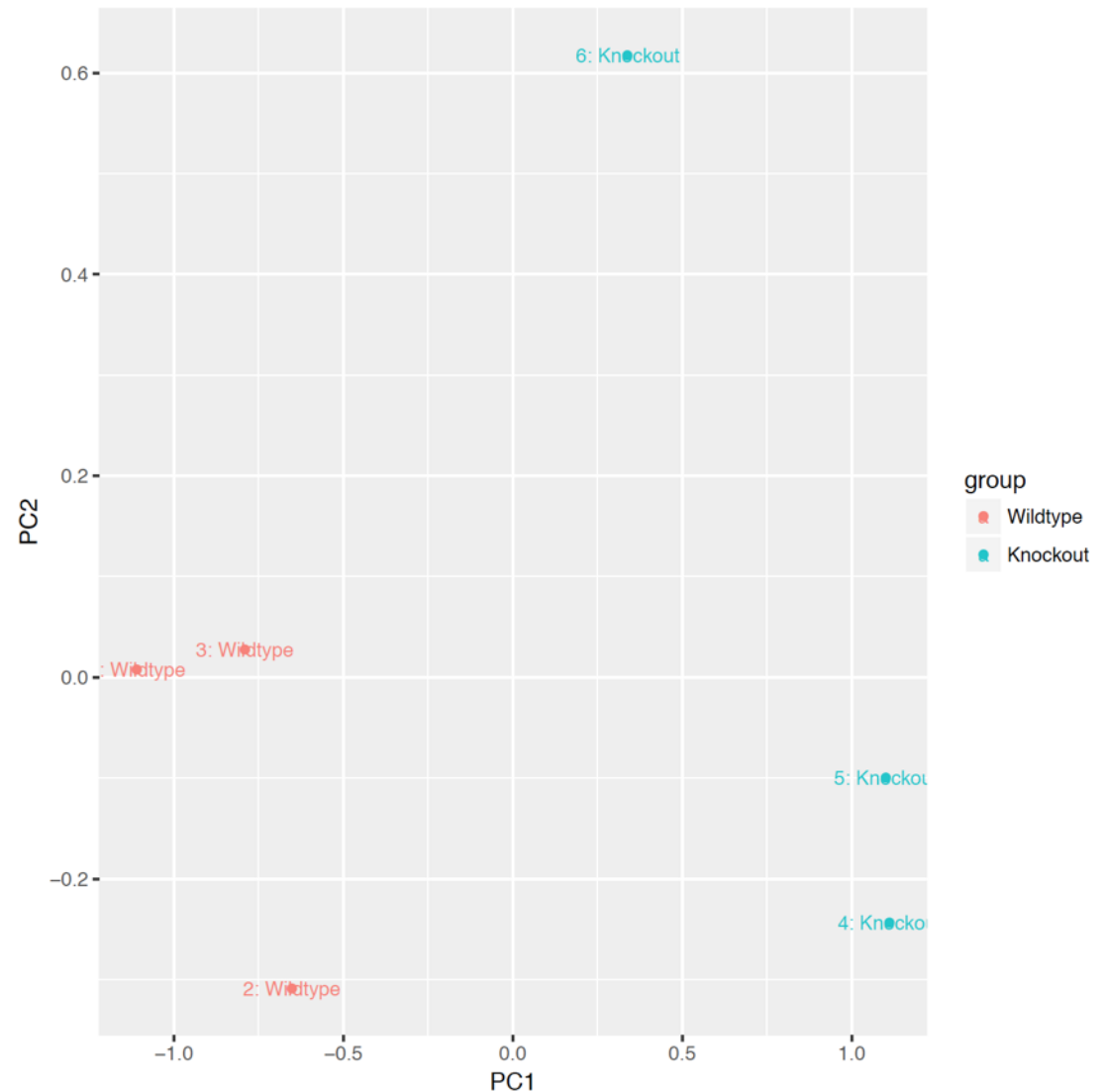
- Click the eye icon to view the plots

28: DESeq2 plots on data 18, data 16, and others



## Part D: DESeq2 plots – PCA

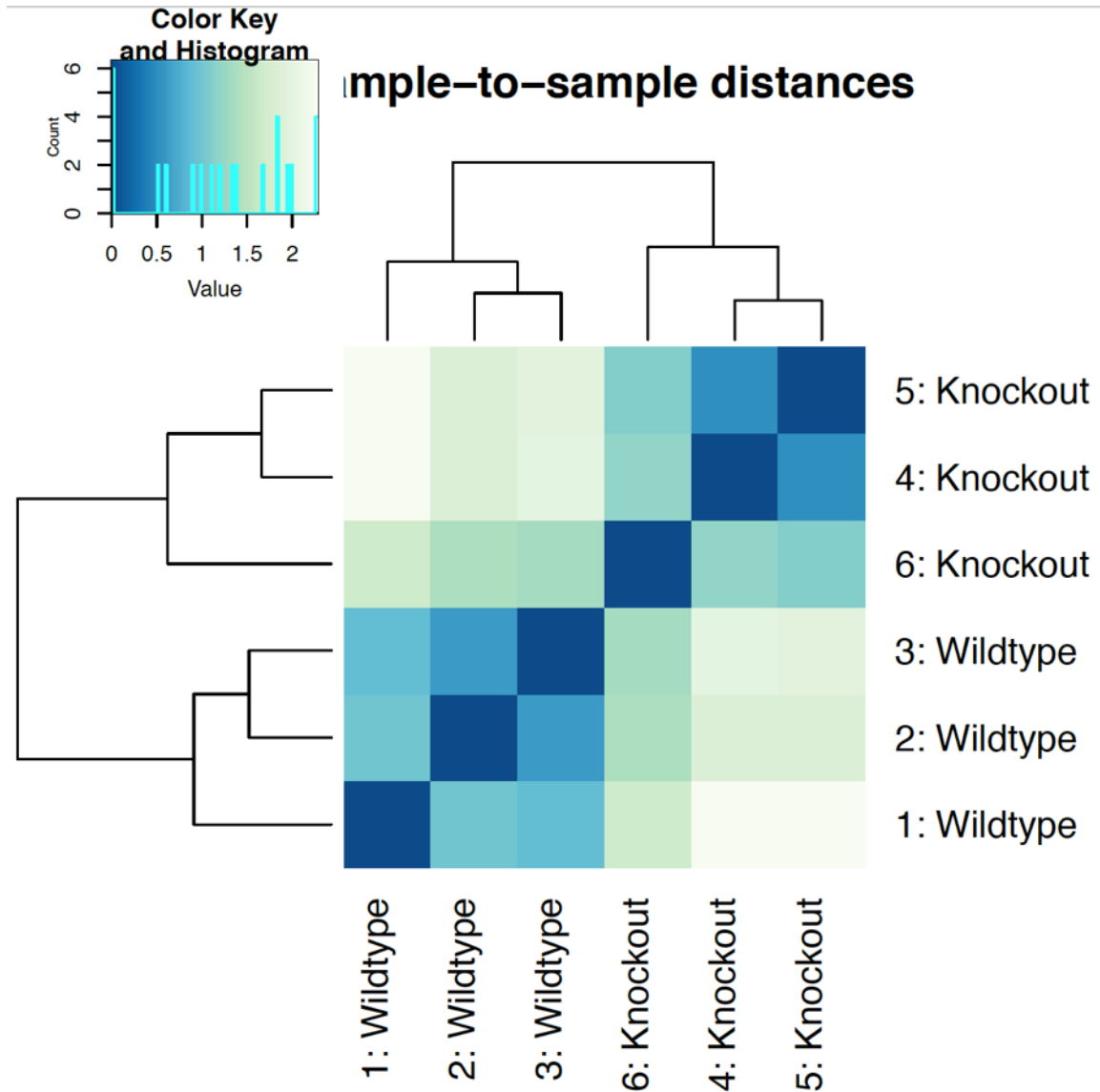
- Principal components analysis plot
- Sample clustering
- Indicates possible contamination, other issues





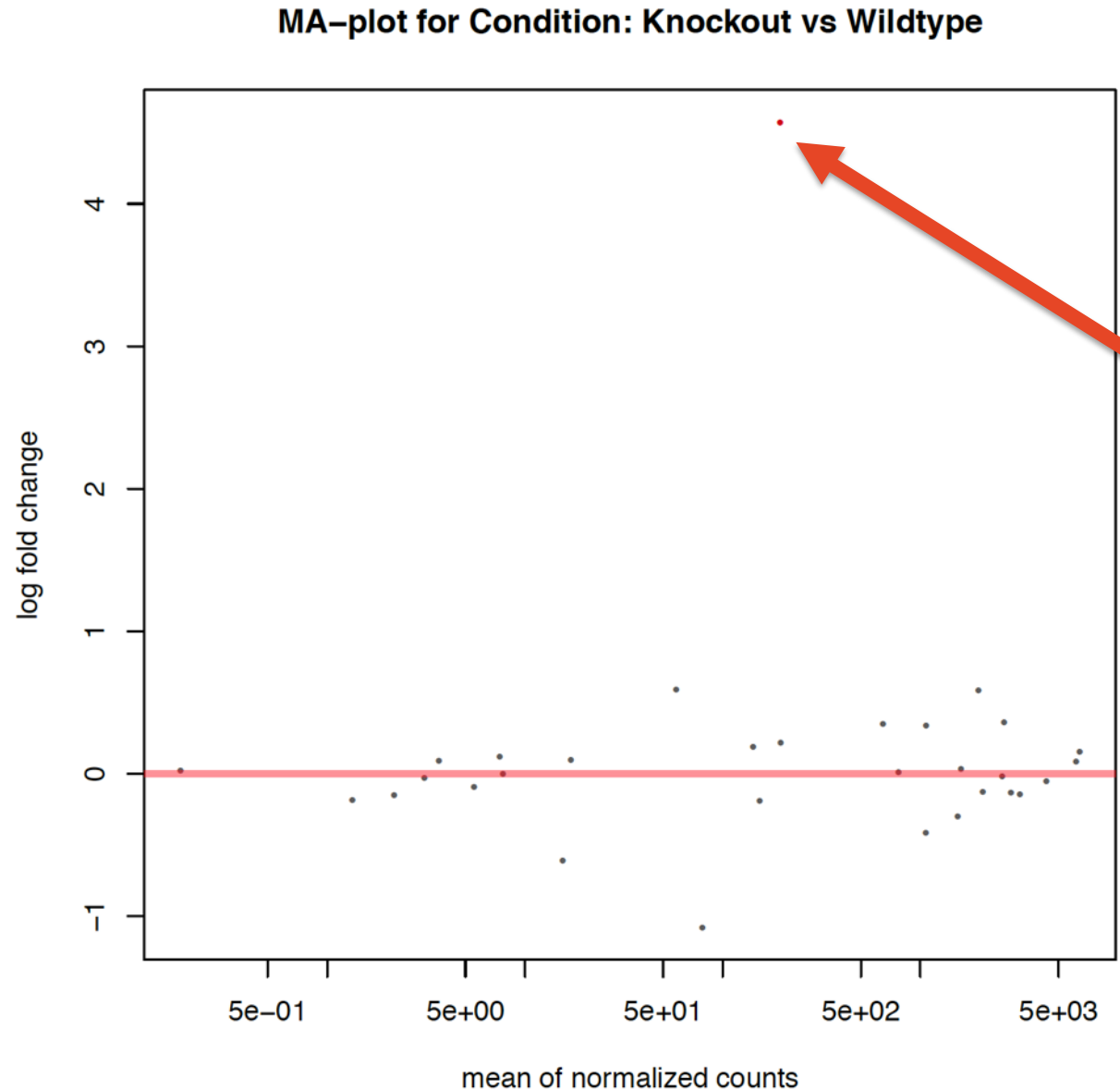
## Part D: DESeq2 plots – Sample to sample distances

- Sample clustering



## Part D: DESeq2 plots – MA plot

- Logfold changes for each gene vs mean of normalised counts
- Red dots: significantly differentially expressed genes



## Part D: DESeq2 – result file

- Click the eye icon to view the DESeq2 results file
- Two significant DE genes ( $P_{adj} < 0.1$ )
- $\log_2(FC)$  of 4.5 indicates that this gene is upregulated in the knockout group (wildtype was set as base level)

GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
Dcc	195.415277304316	4.57268885427617	0.281248624502068	16.2585287745738	1.94348873663644e-59	6.02481508357296e-58
Dynap	79.0541023643527	-1.08088086878273	0.300640858129957	-3.59525606567922	0.000324072661696416	0.00502312625629445
Rab27b	1967.66146410935	0.582984895350518	0.238394493181658	2.4454629281486	0.0144666320328227	0.125139444131148
Smad7	1064.29518567968	-0.4163308248882	0.17306913207441	-2.40557527444699	0.0161470250491804	0.125139444131148
Me2	646.261985114386	0.348500769717429	0.159674568282196	2.18256904318988	0.029067561591892	0.145864810341893
Myo5b	2653.49787779652	0.359608930891107	0.168605061369932	2.13284778030535	0.032937215238492	0.145864810341893

## Part D: DESeq2 – result file

- We can see the difference in expression of DCC between wildtype and knockout mice in the raw count data
- In tools under RNA-seq, click Generate count matrix
- Count files from your history
  - Highlight all 6 counts files
- Click execute
- This puts all count data into a single matrix (note these are raw counts)

1	2	3	4	5	6	7
gene_id	WT_3_counts	WT_2_counts	WT_1_counts	KO_3_counts	KO_2_counts	KO_1_counts
Ccdc68	26	56	39	62	129	52
Rab27b	1447	1707	1331	2675	3841	1193
Dynap	331	67	48	28	8	19
4930503L19Rik	151	221	135	126	253	72
Stard6	5	10	4	14	10	2
Poli	209	193	147	172	305	188
Mbd2	5514	3838	4083	3275	5600	4314
Dcc	0	2	0	399	589	236
Meis3	3598	2851	2674	2289	3660	2527

## Part D: Functional analysis

- Use your favourite database to search for associated phenotypes for this DE gene
- Does it relate to the disease of interest?

(A reminder...)

Knockout mouse model to study **Williams-Beuren Syndrome (WBS)**, a rare disease found in people

- distinctive facial features
- intellectual disability
- cardiovascular abnormalities

## Part D: Functional analysis

In a more typical RNA sequencing analysis, you will normally end up with hundreds to thousands of significantly differentially expressed genes.

*What is considered significant?*

- $P\text{-adj} < 0.05$
- $\text{Log}_2(\text{FC})$  over  $\pm 2$
- Somewhat arbitrary, but recommended to have 100-3,000 genes for pathway analysis

Tools to find enriched biological pathways for significantly differentially expressed

- DAVID
- PANTHER
- Ingenuity Pathway Analysis (Usyd has one shared license, contact SIH if you would like access)

# Acknowledgements

The Galaxy community

**Sydney Informatics Hub**

Rosemarie Sadsad

Nicholas Ho

Anushi Shah



# Part E: 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 Australia:** <https://usegalaxy.org.au/>

**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/>

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

You can also come to our monthly **Hacky Hour** event or contact the **Sydney Informatics Hub** if you need assistance with your projects.





# Part E: Useful resources

## 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://bmcbgenomics.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>

## Research Computing Services

Provides research computing expertise, training, and support

- Data analyses and support (bioinformatics, modelling and simulation, visualisation)
- Training and workshops
  - High Performance Computing (HPC)
  - Programming (R, Python, Matlab, Scripting, GPU)
  - Code management (Git)
  - Bioinformatics (RNA-Seq, Genomics)
- Research Computing Support
  - Artemis HPC
  - Argus Virtual Research Desktop
  - Bioinformatics software support (CLC Genomics Workbench, Ingenuity Pathways Analysis)
- Events and Competitions
  - HPC Publication Incentive – High quality papers that acknowledge SIH and/or HPC/VRD
  - Artemis HPC Symposium

## **Data Science Expertise**

Provides data science (e.g. machine learning, deep learning, AI, NLP) expertise, training, and support

## **Research Data Management and Digital Tools Support**

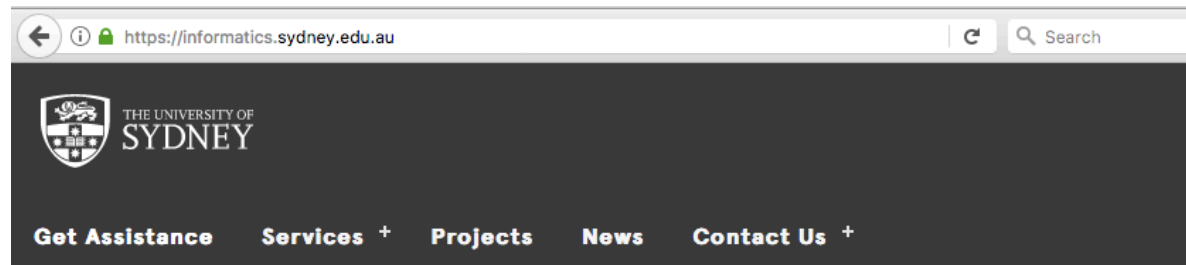
Provide expertise, training, and support on management of research data and use of digital tools.

- Digital research platforms supported
  - eNotebook - collaborative electronic notebook
  - REDCap - surveys and databases
  - GitHub - software repository management
  - Research Data Store
  - Dropbox
  - CloudStor
  - Office365/OneDrive

# Sydney Informatics Hub

W: <https://informatics.sydney.edu.au>

E: [sih.info@sydney.edu.au](mailto:sih.info@sydney.edu.au)



## Sydney Informatics Hub

The Sydney Informatics Hub (SIH) is a core research facility of the University of Sydney, providing **services** surrounding data and computation within the University. It delivers policies, systems, advice, engineering and training to our researchers and their external collaborators.

Researchers can access SIH's services primarily through attending **training workshops** or seeking assistance below.