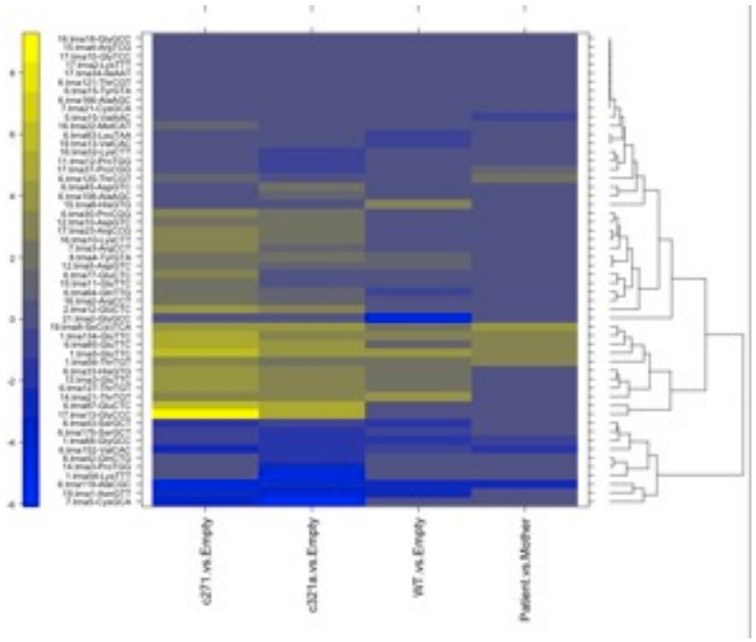
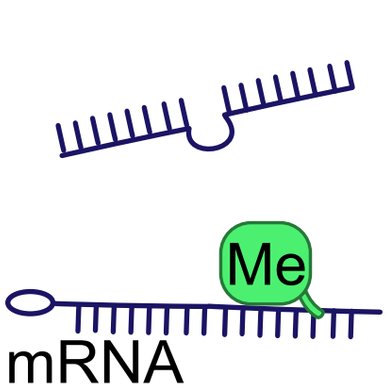
Approaches to transcriptome analysis



Dr Jelena Aleksic
University of Cambridge
ja313@cam.ac.uk

What I do all day

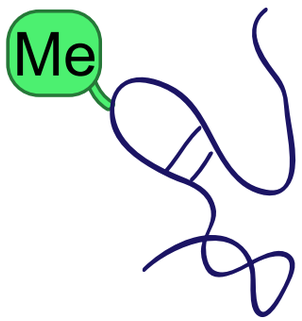
The types of data I work on



sRNA-seq

mRNA-seq

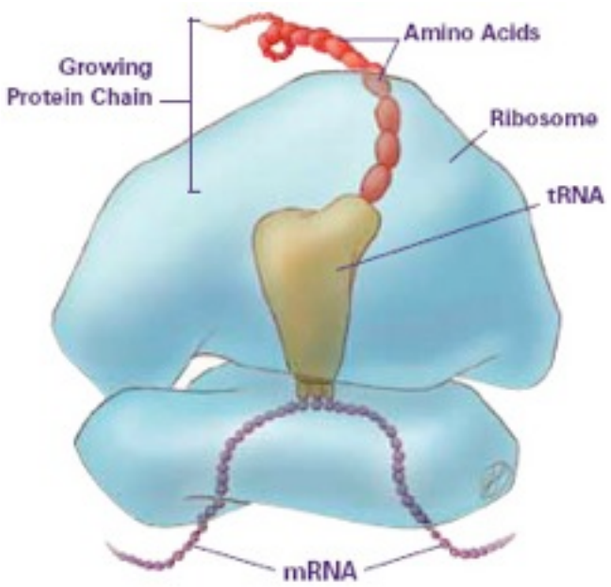
What RNA transcripts are present?



miCLIP

Bisulfite seq

Where is RNA methylation taking place?

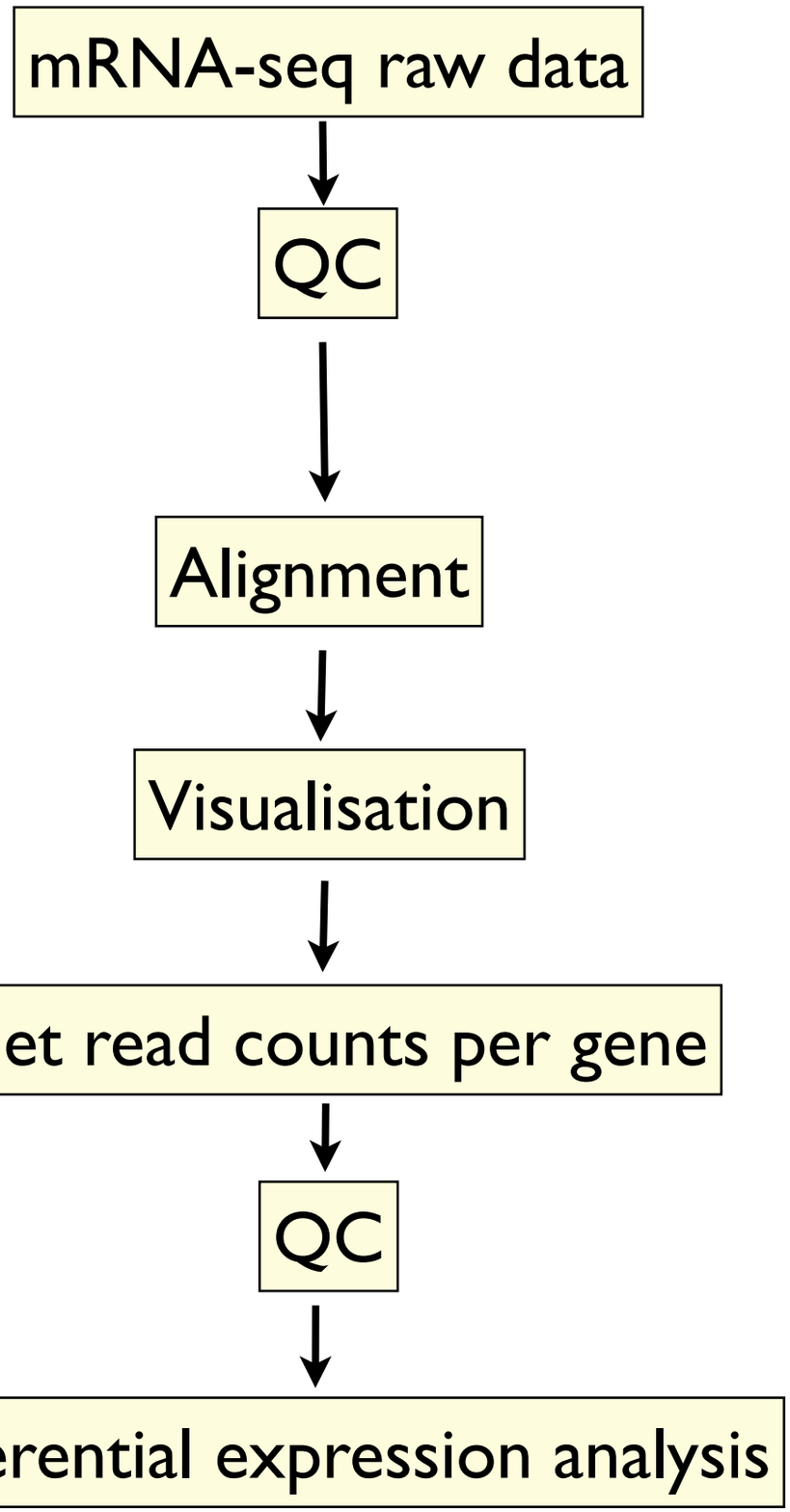


(Credit: http://www.sciencegeek.net)

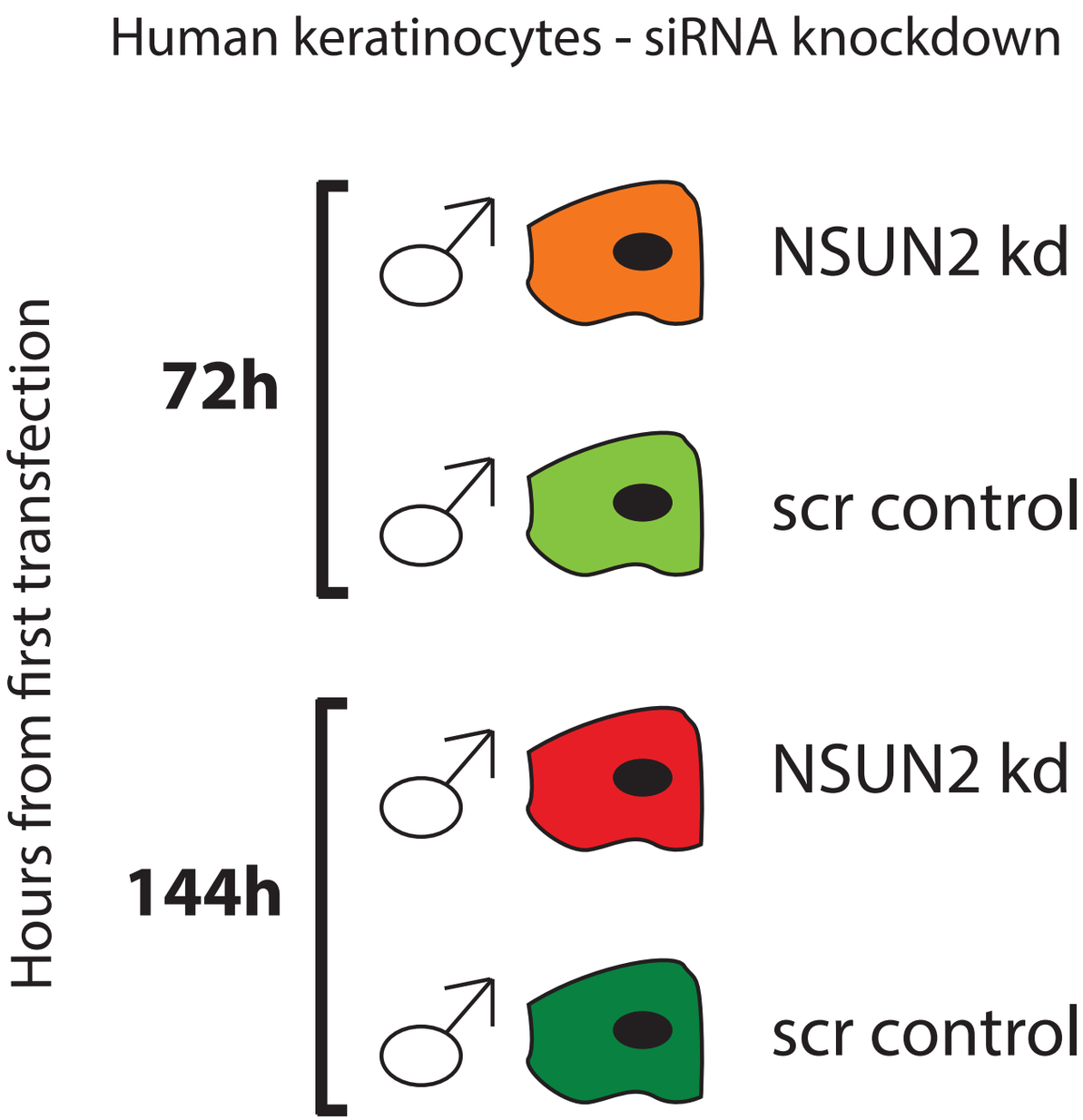
Ribosomal profiling

What is the rate of translation / pattern of ribosome occupancy?

A standard mRNAseq workflow

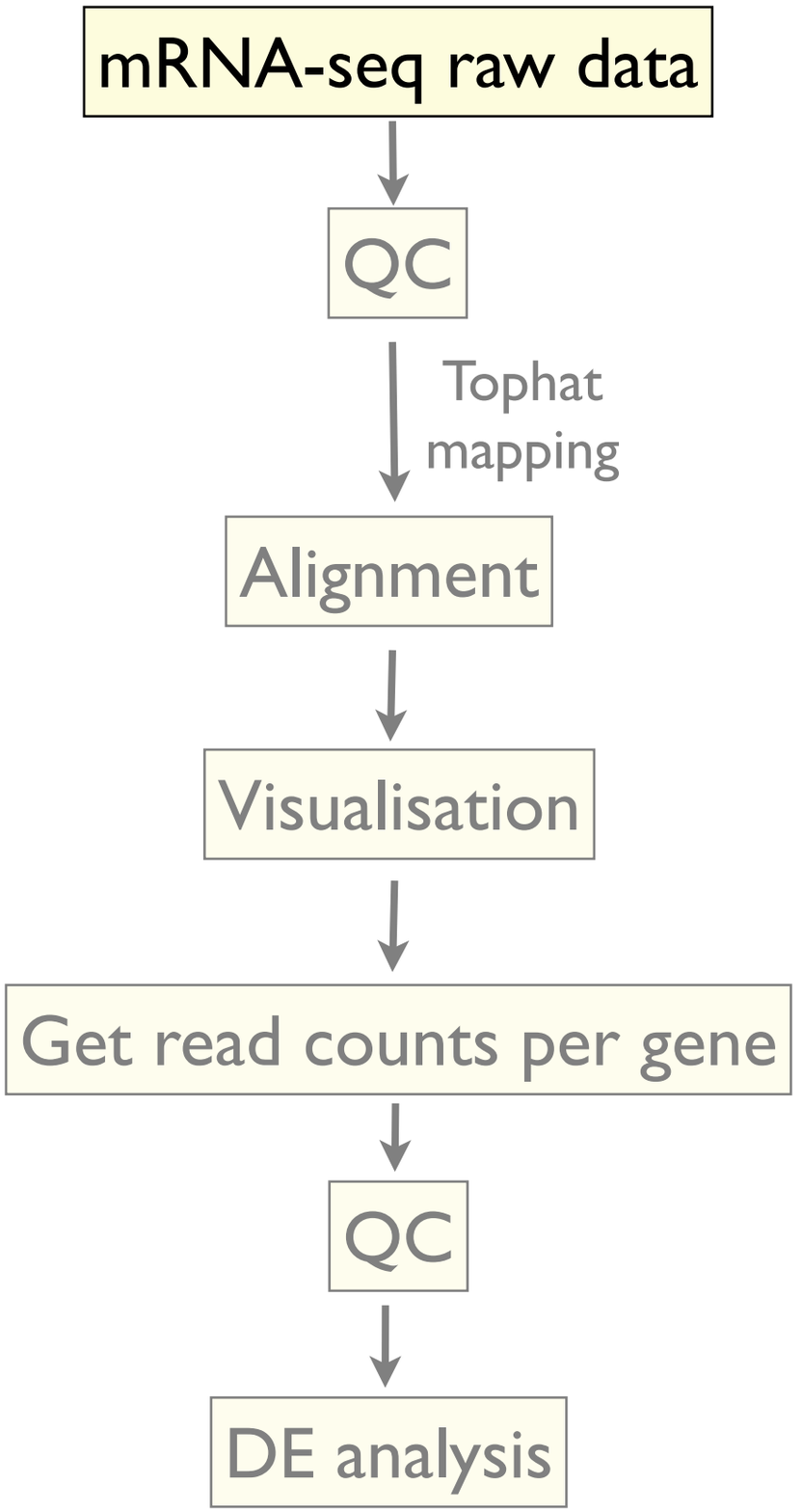


Time course example - verifying Nsun2 knockdown effects



Different methylation stages.
Most tRNAs are still
methylated at 72h,
and unmethylated at 144h

FastQ files



- Large files
- 4 lines per sequence
- For 30M sequences, that's 120M lines!

(something like 1-2GB per file)

FastQC report

mRNA-seq raw data

↓

QC

↓ Tophat mapping

Alignment

↓

Visualisation

↓

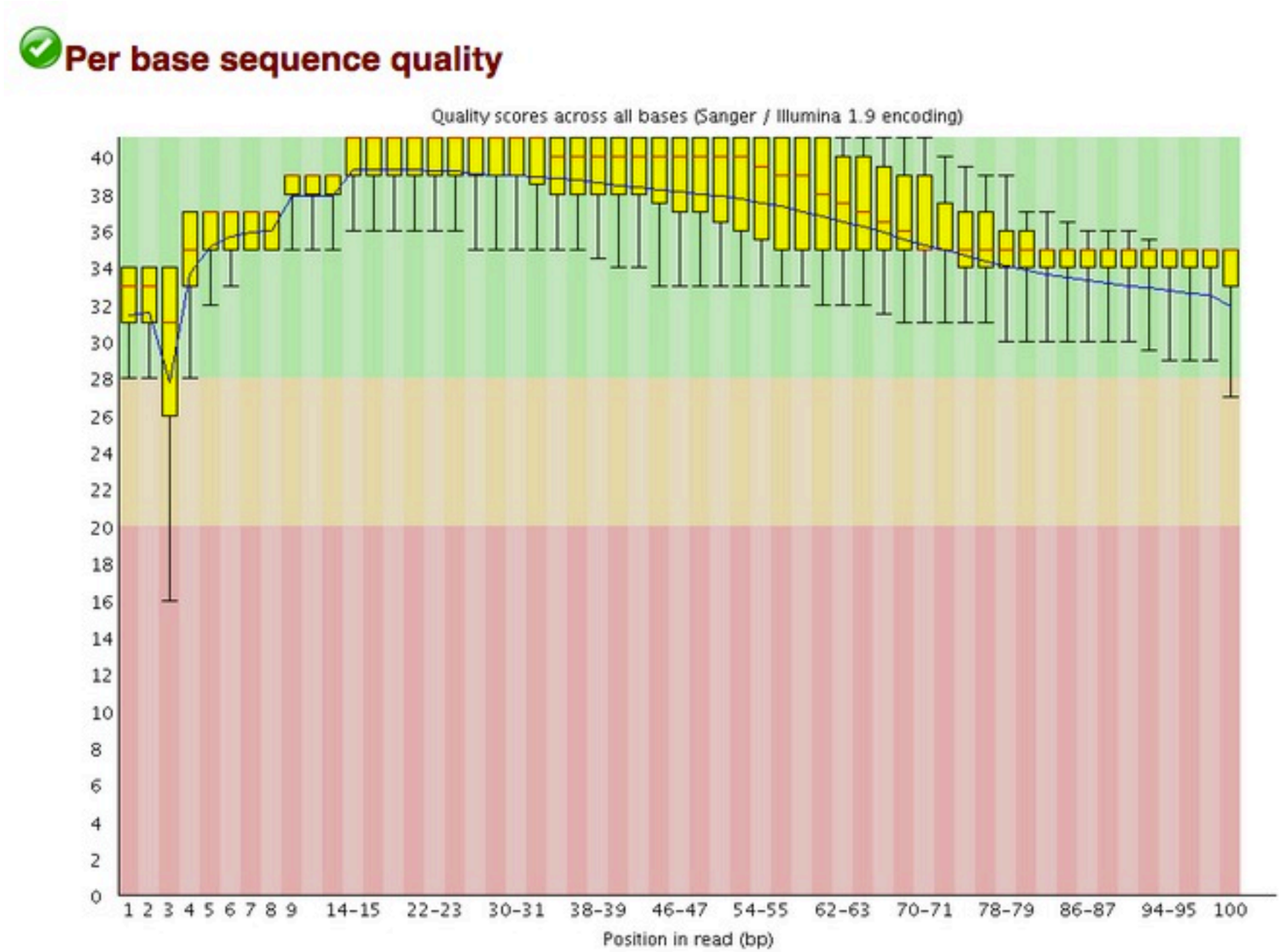
Get read counts per gene

↓

QC

↓

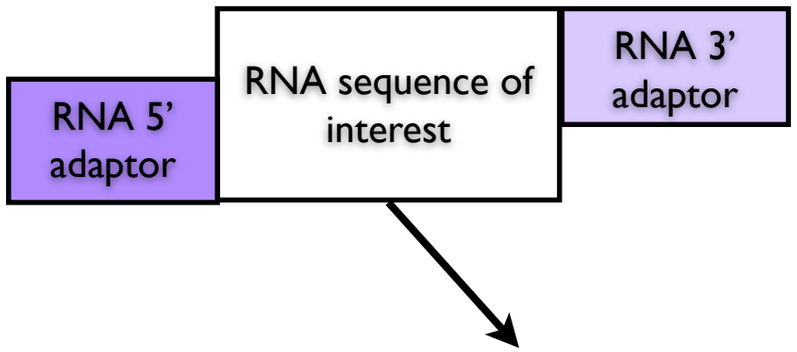
DE analysis



To trim or not to trim adapters

For mRNAseq:

Relatively few 3' adapter traces, and the aligner actually deals with them ok. Data would in theory be cleaner without them, but it's one extra step, and I found it only improved my mapping efficiency by 0.2%



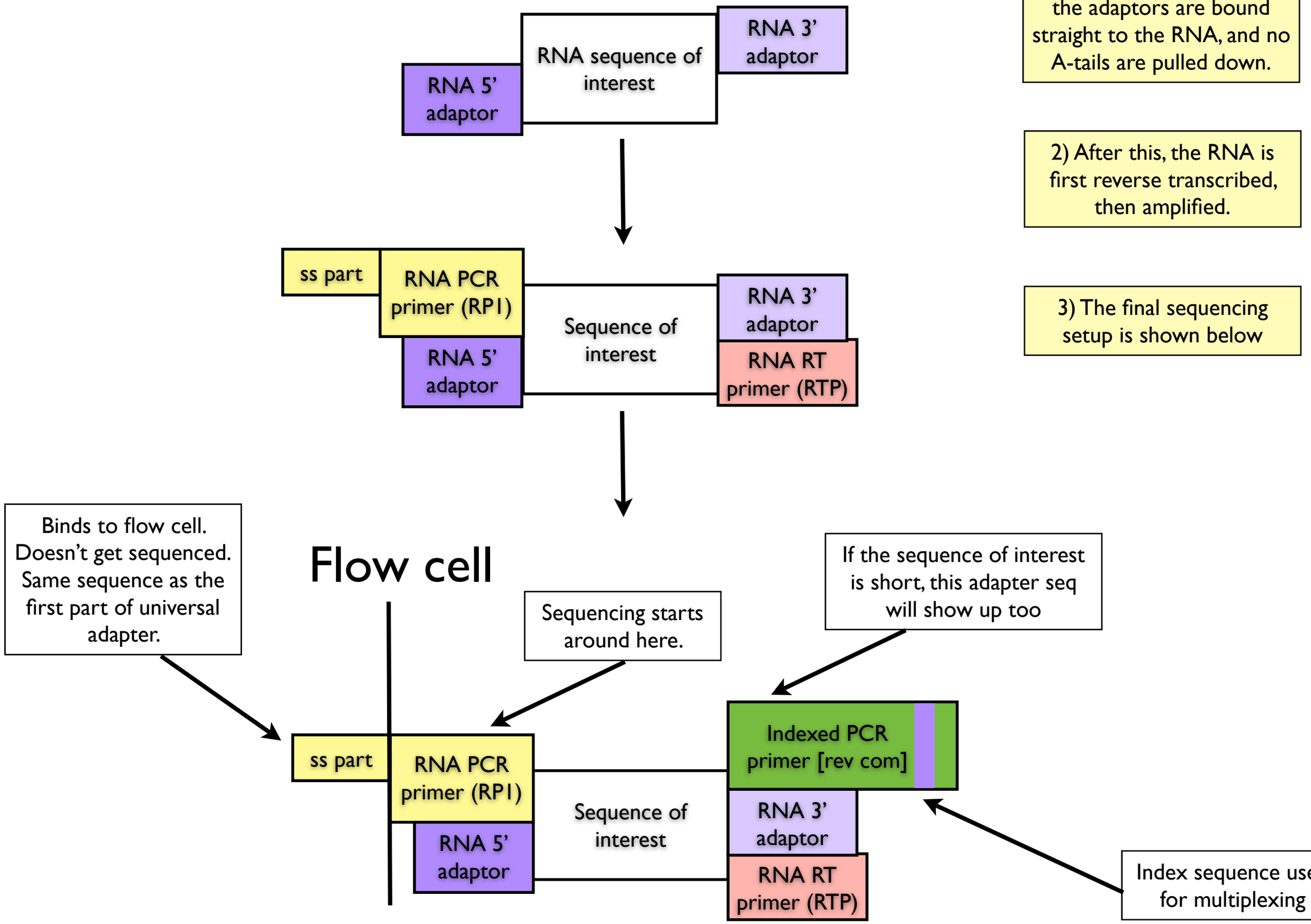
mRNA fragments =~ 80-200 bp

Sequencing to 100bp doesn't run over into the 3' adapter most of the time.

For sRNAseq / other applications:

Removing adapters is often absolutely critical.

sRNAseq / Riboseq adapter setup

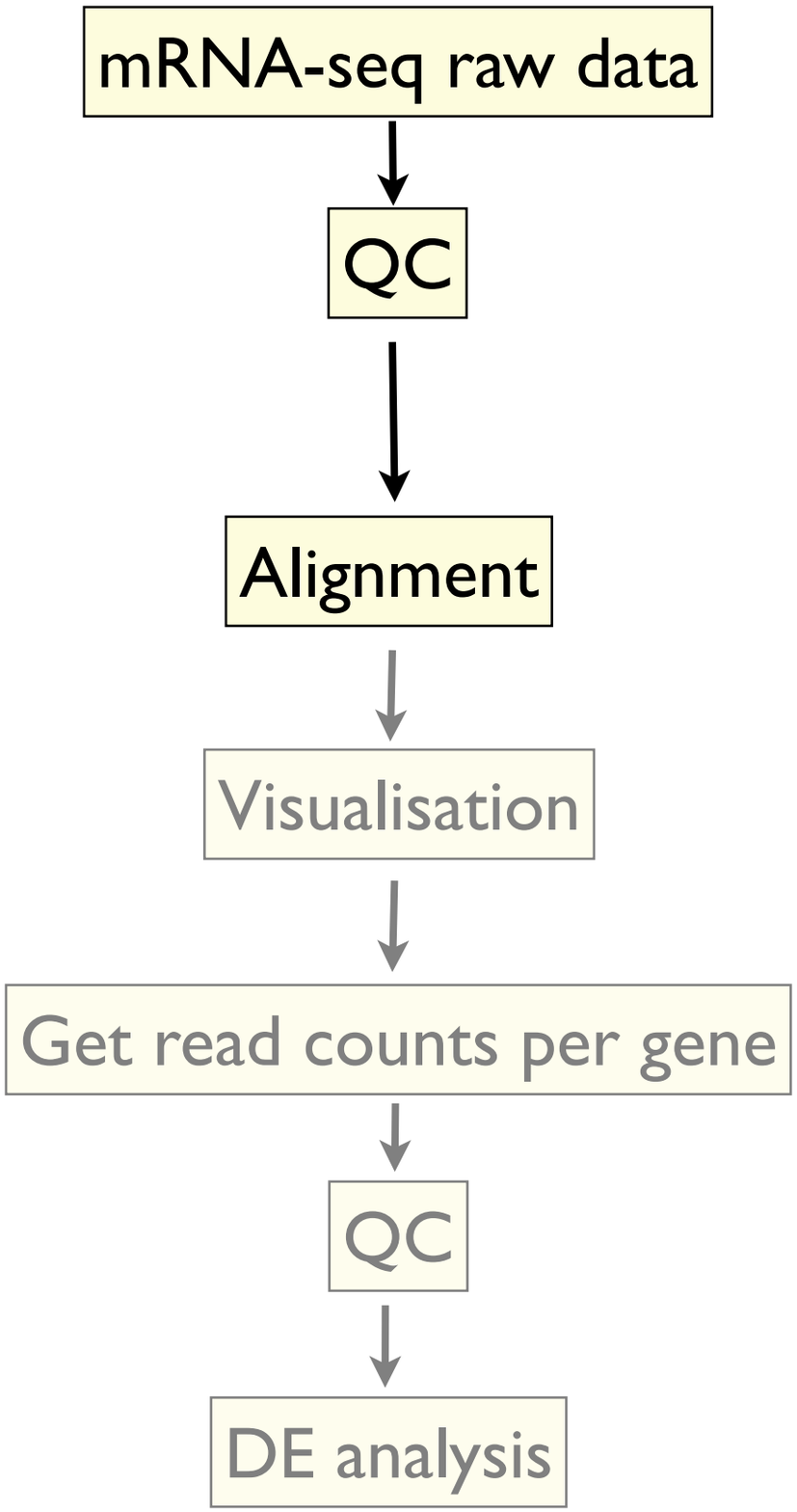


1) In the case of smallRNA, the adaptors are bound straight to the RNA, and no A-tails are pulled down.

2) After this, the RNA is first reverse transcribed, then amplified.

3) The final sequencing setup is shown below

Alignment



You have the option to align to:

- Genome
- Transcriptome

Both are usable options.

Genome:

You create a genome index file, and give it to your aligner alongside a gtf file of gene annotations.

Transcriptome:

Align to a fasta file of transcript sequences.

Sequence and annotation files

Genome and transcriptome: sequence files.

Fasta format:
>sequence1
ATGCGTGCACAGTT

Gene annotations

GTF format:

chr1 processed_transcript exon 11869 12227 . + .
gene_id "ENSG00000223972"; transcript_id "ENST00000456328";
exon_number "1"; gene_name "DDX11L1"; gene_source "havana";
gene_biotype "transcribed_unprocessed_pseudogene";
transcript_name "DDX11L1-002"; transcript_source "havana";
exon_id "ENSE00002234944";

Where you find genome sequences and alignments

www.ensembl.org ⟶ FTP download

Show 10 entries | Show/hide columns | Filter

★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)
Y	**Human** *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	Regulation (GFF)
Y	**Mouse** *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	Regulation (GFF)
Y	**Zebrafish** *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	-
	Alpaca *Vicugna pacos*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-
	Amazon molly *Poecilia formosa*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-
	Anole lizard *Anolis carolinensis*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-

Genome

Transcriptome

Gene annotations

Keeping track of versions

Both genome files and annotation files have versions.

e.g.
Human genome version GRCh38
Gene annotation version - Ensembl release 77

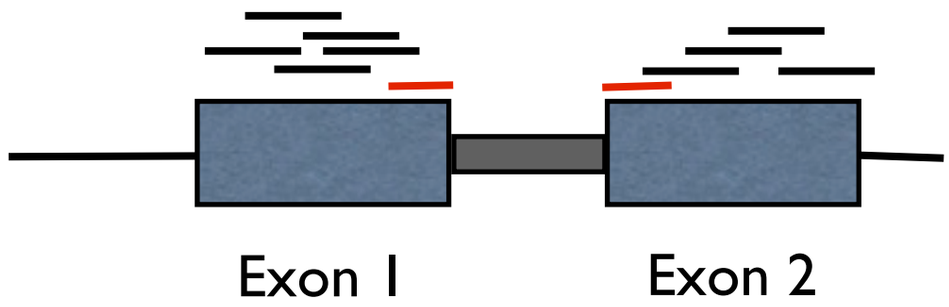
Important:
It is essential to note what genome and annotation versions you're working with.

Genome = changes every few years.
Annotations = change every few months.

Using Tophat for genome alignment

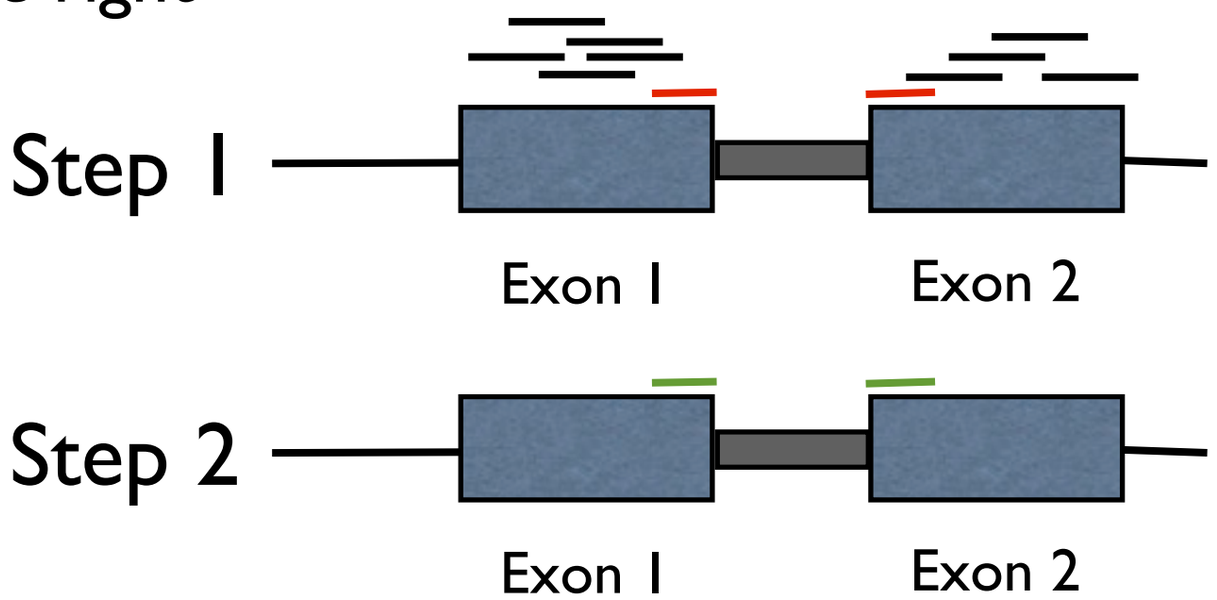
bowtie

- Super fast aligner
- Aligns sequences to the genome
- Forms the core of Tophat
- Not aware of splice junctions



TopHat

- Starts off by running bowtie
- For unmapped reads, it uses splice junction information to map them to the right exons/ transcripts.

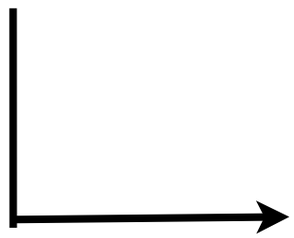


Using Tophat for genome alignment

TopHat

1. Download genome sequence + annotation files
2. Prepare index files
3. Run Tophat:

```
tophat -p 12 -o patient1_thout -G Homo_sapiens.GRCh38.76.gtf -m 2
~/Bowtie2Index/hg38_genome patient1_RNAseq.fastq
```



Output: accepted_hits.bam

(binary file of genome alignments)

Unique vs multiple mapping

TopHat

The .bam file from Tophat contains both **uniquely** and **multiply** mapped reads.

For mRNAseq, people often work with uniquely mapped reads only, so it is important to filter for these.

Converting from .bam to .sam using **samtools** will give you a text file you can read/process.

All uniquely mapped reads have a flag NH:i:1

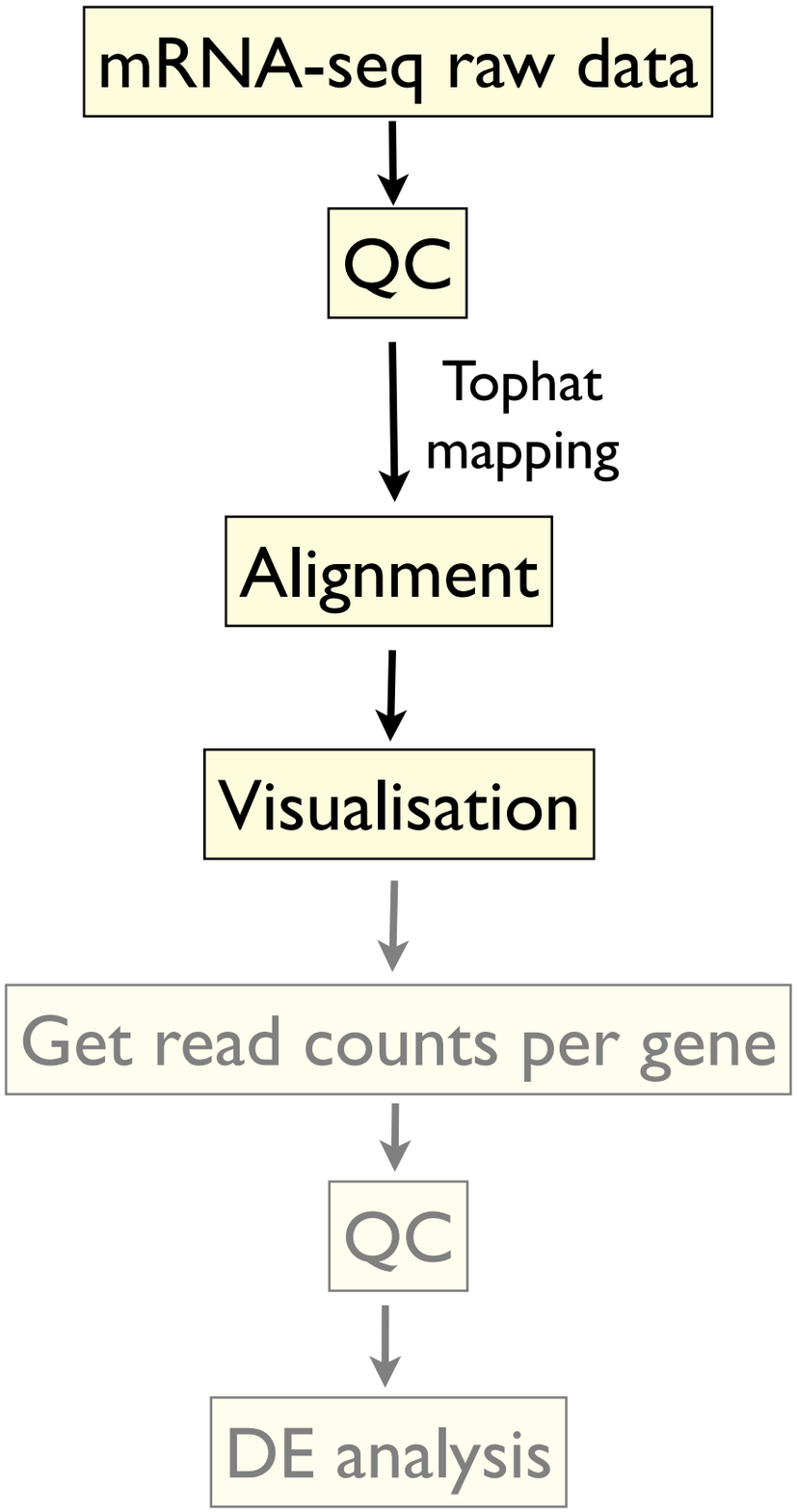
-> the number at the end signifies to how many locations the read has been mapped

Converting from .bam to .sam using **samtools** will give you a text file you can read/process.

Collecting alignment statistics

Time point	Sample	Replicate	Readsx10^6	% Reads mapped	% non-unique mapping	Uniq. mapped reads x10^6
72 h	NSUN2 KD	A	41.8	97.0	6.3	38.0
		B	41.0	97.2	6.4	37.3
		C	38.6	97.1	6.1	35.1
		D	41.7	97.1	7.0	37.7
	Scr Control	A	45.7	97.1	6.7	41.4
		B	40.2	97.1	7.2	36.3
		C	40.3	97.2	6.7	36.6
		D	40.7	97.1	7.0	36.7
144 h	NSUN2 KD	A	39.7	96.9	9.3	34.8
		B	38.8	96.8	9.0	34.1
		C	38.1	96.1	9.5	33.1
		D	39.4	96.4	10.3	34.1
	Scr Control	A	38.7	96.9	7.1	34.8
		B	39.9	96.5	6.9	35.8
		C	37.2	96.9	7.4	33.4
		D	36.9	96.6	7.1	33.2

Visualisation



Once the reads have been aligned, it is important to visualise the data.

- Sanity checks what the data looks like
- Can act as a quality control

Aligned sequence data is visualised using a **genome browser**.

Popular ones:

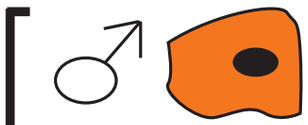
- UCSC (online)
- IGV

Nsun2 siRNA knockout RNAseq data

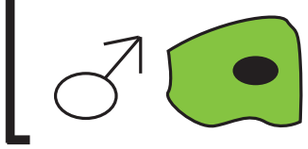
Human keratinocytes - siRNA knockdown

Hours from first transfection

72h

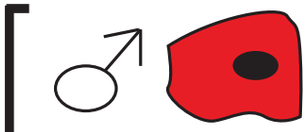


NSUN2 kd

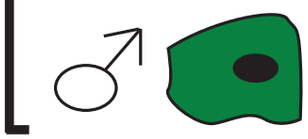


scr control

144h

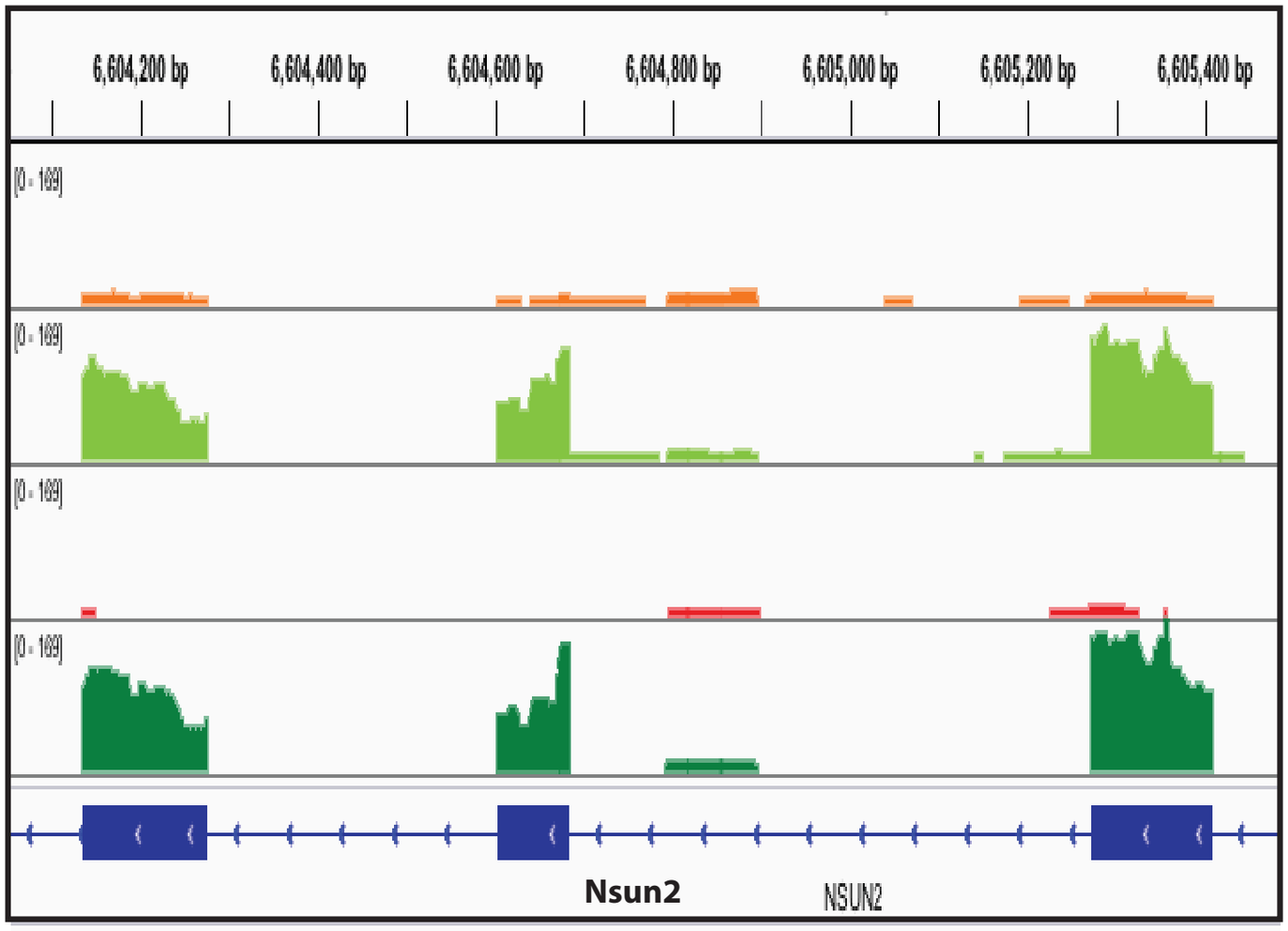


NSUN2 kd

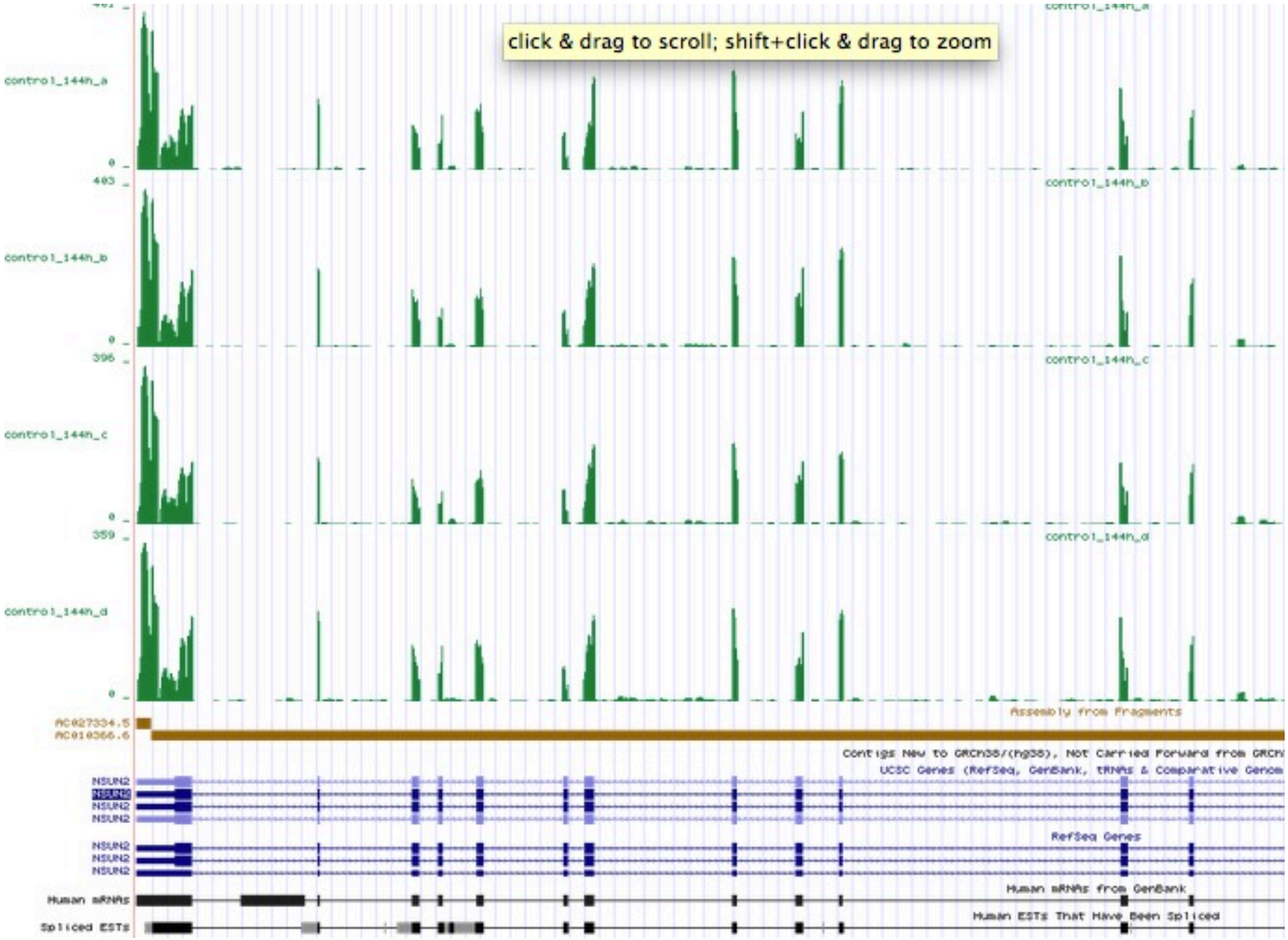


scr control

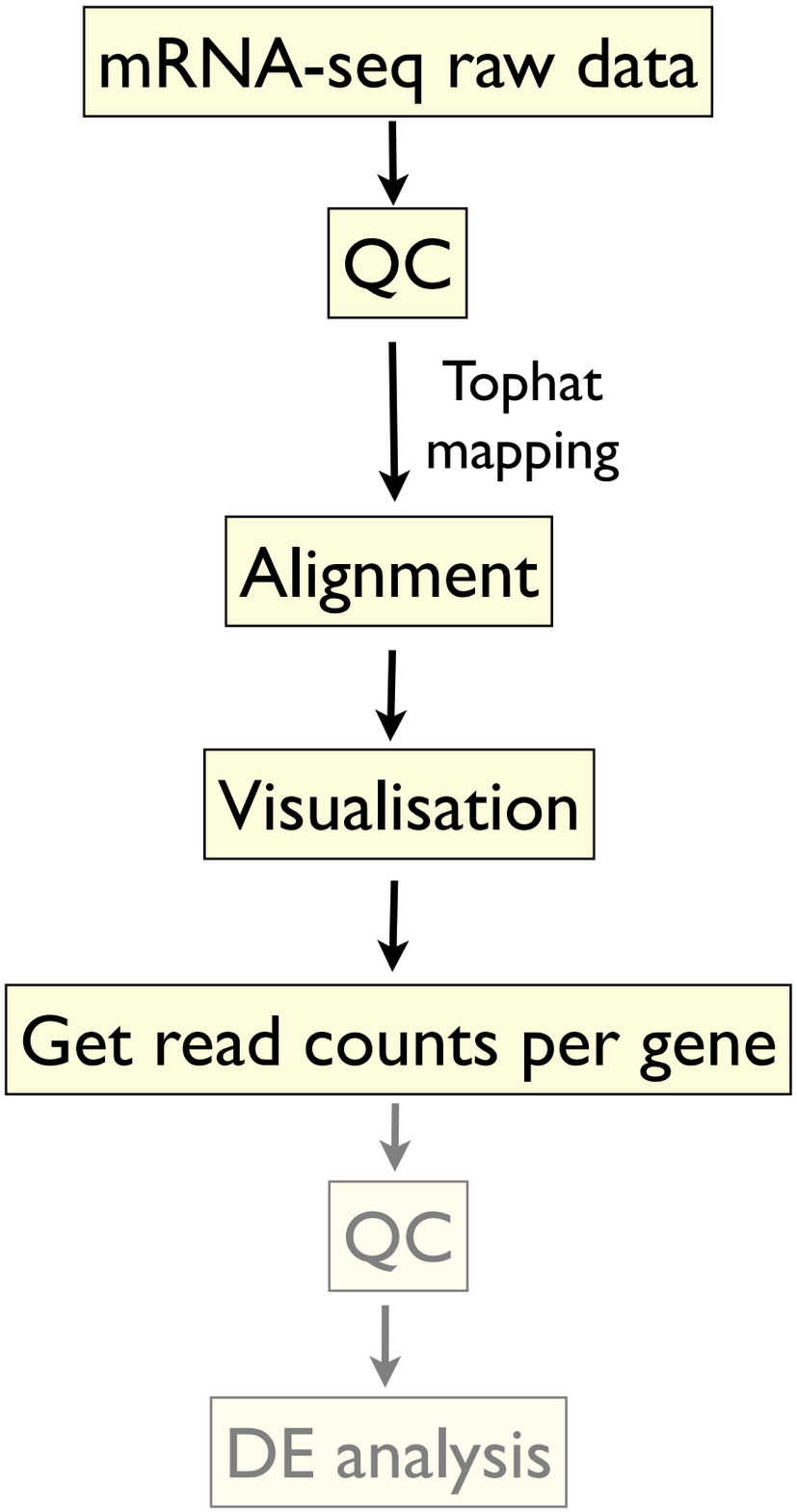
IGV snapshot



UCSC screenshot



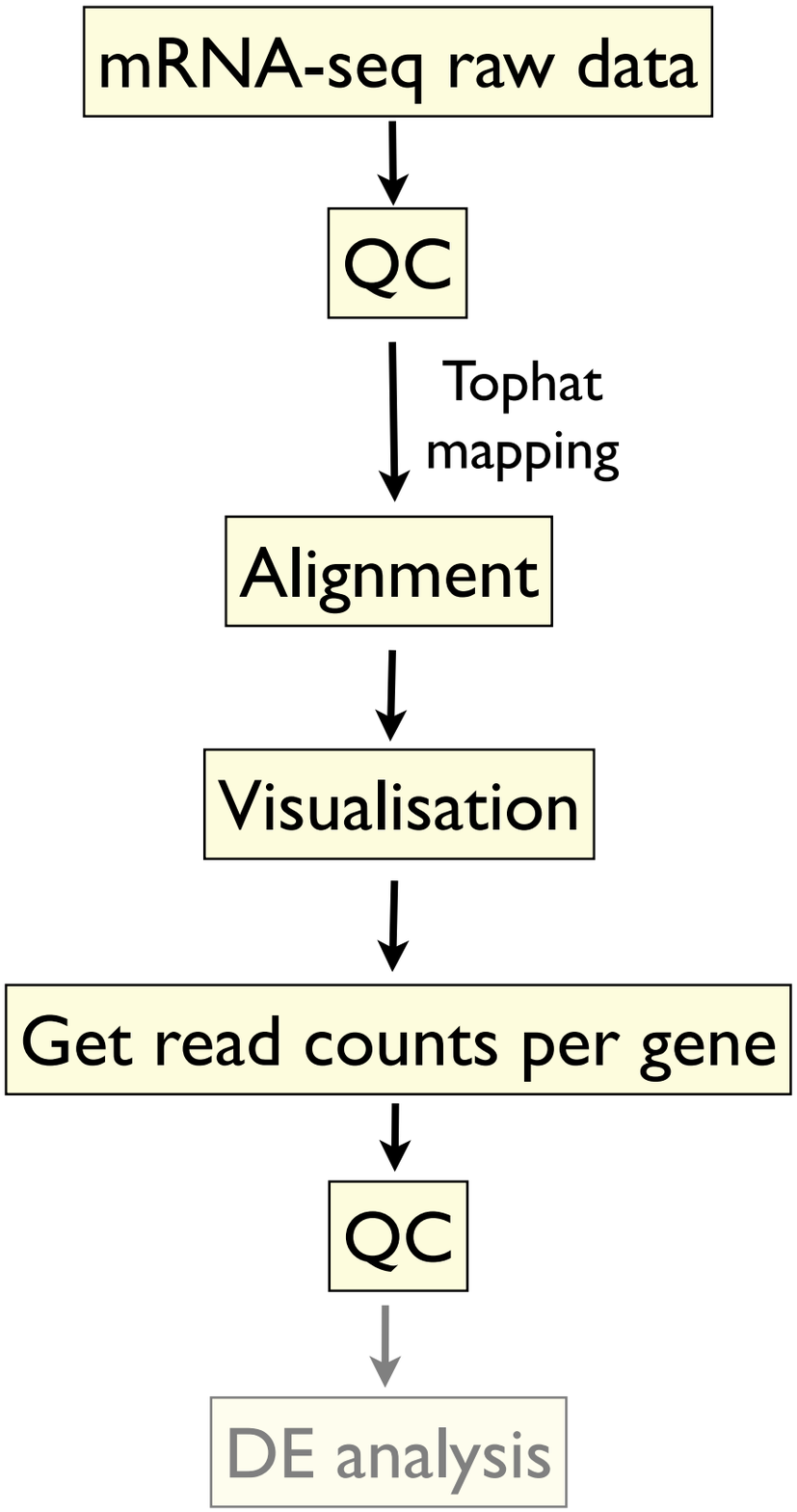
Reads per gene



Want to have the sum of reads per gene,
in order to determine the average expression
of each gene.

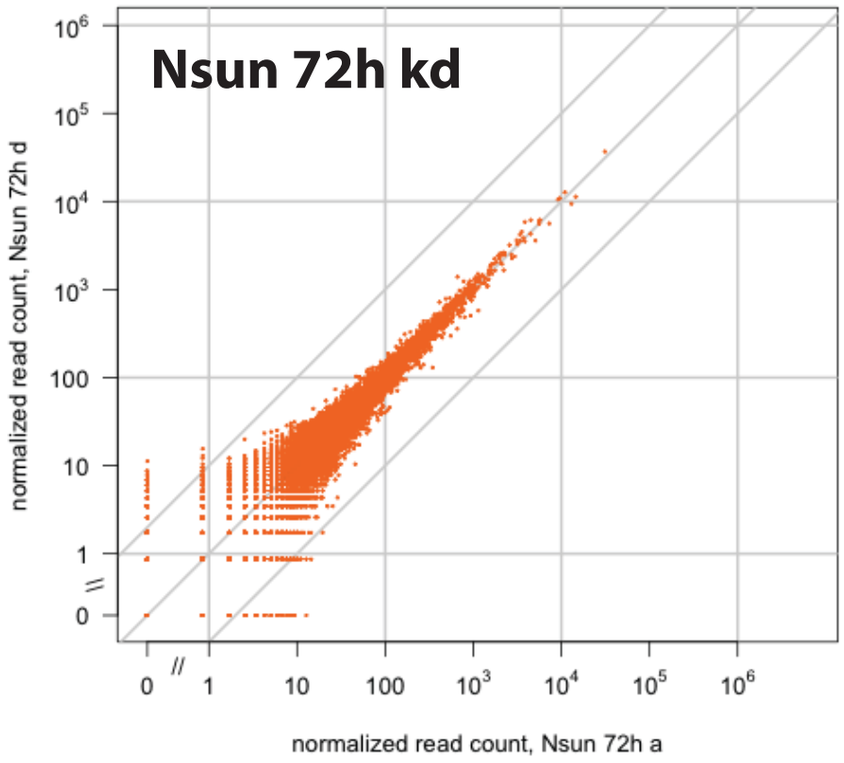
htseq-count -i gene_id -m union
patient1_unique.sam
Homo_sapiens.GRCh38.76.gtf > gene_counts.txt

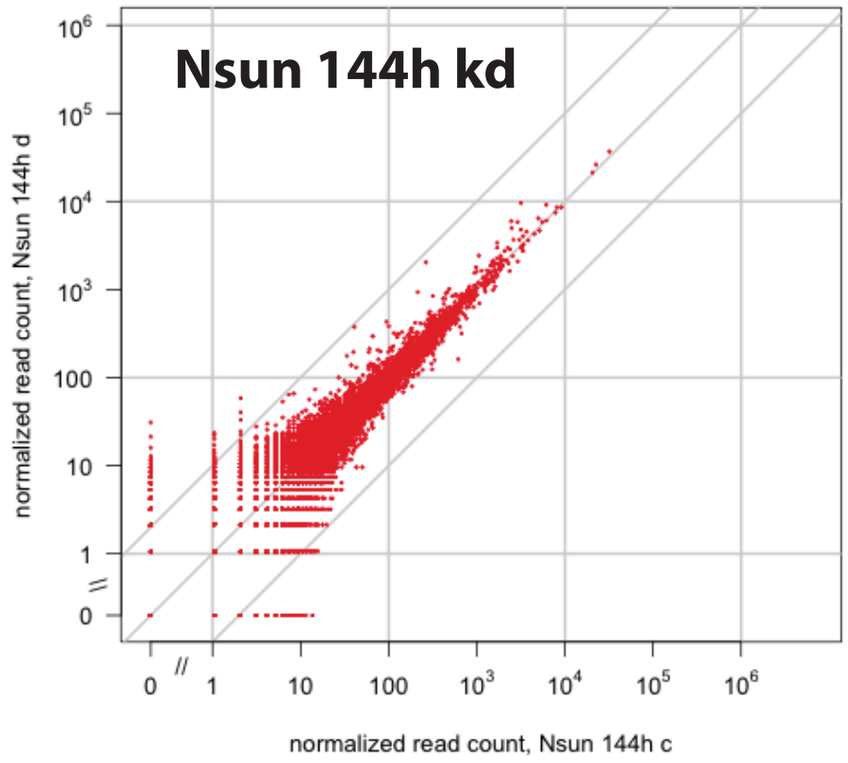
Will get a file of counts per gene for each gene.
Can then assemble a table for all replicates and
conditions.

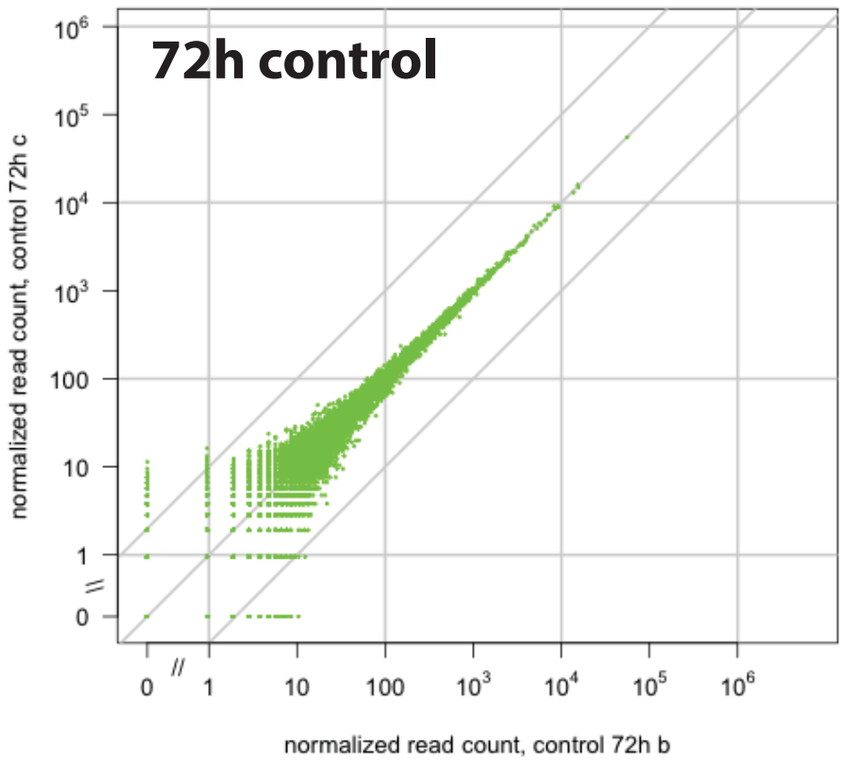


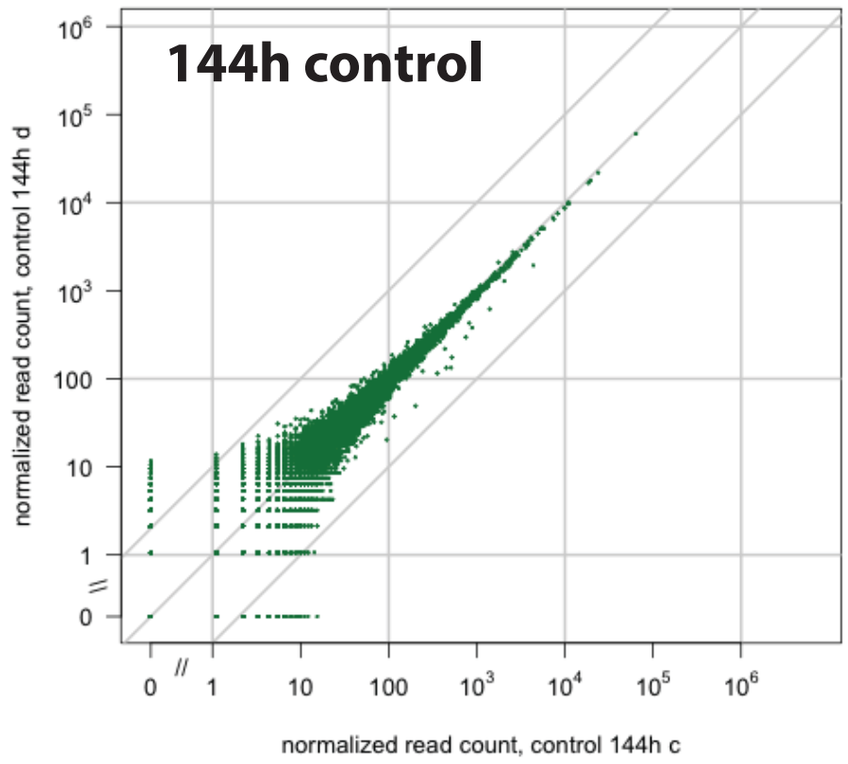
At this stage, QC mainly consists of verifying the data distribution, and the similarity between replicates.

Comparison of replicates within a condition

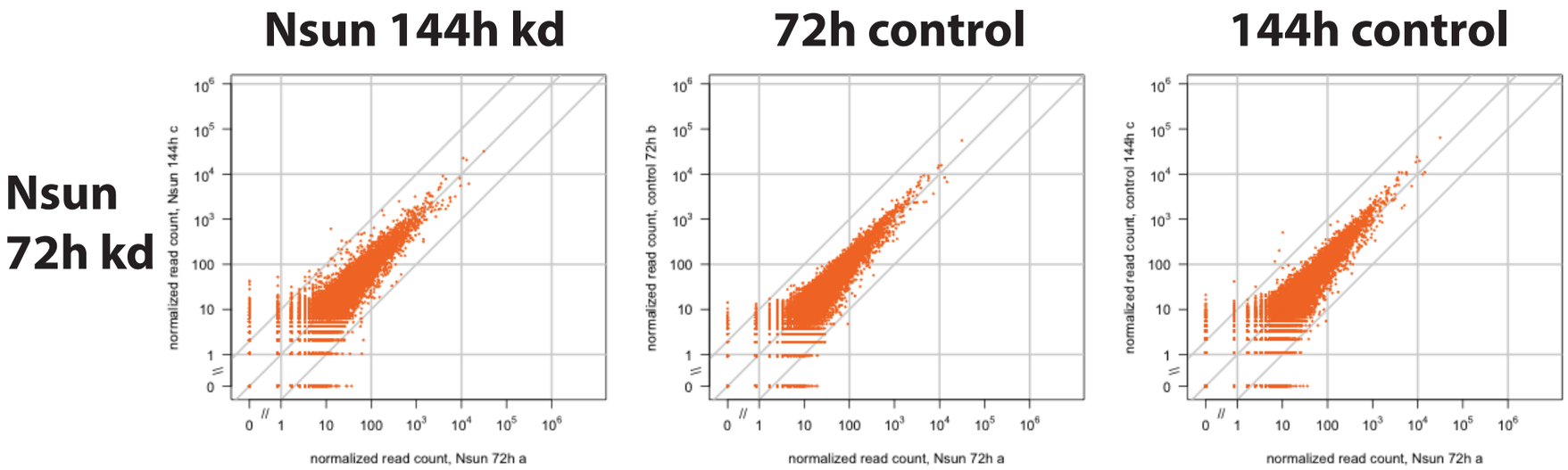


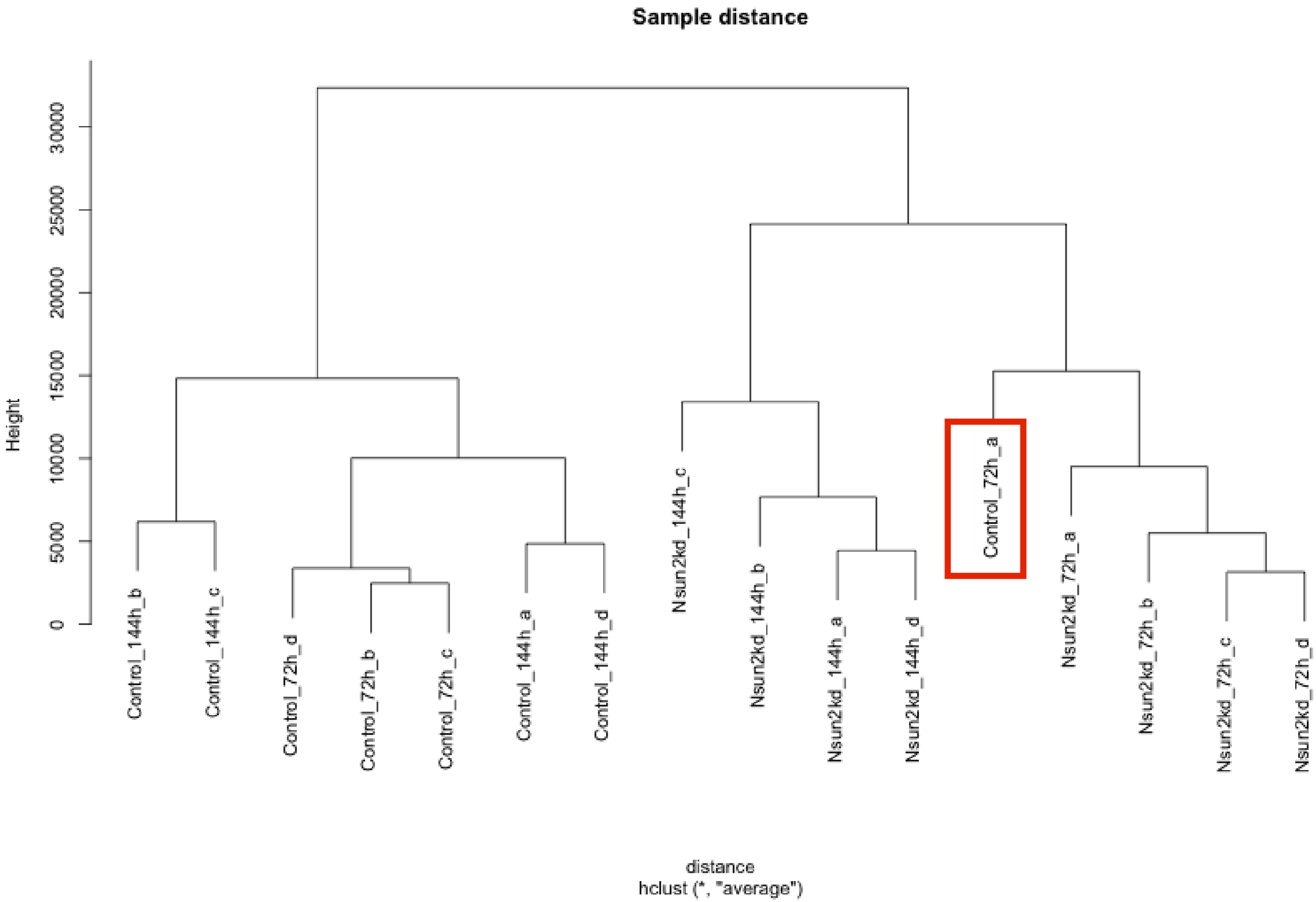


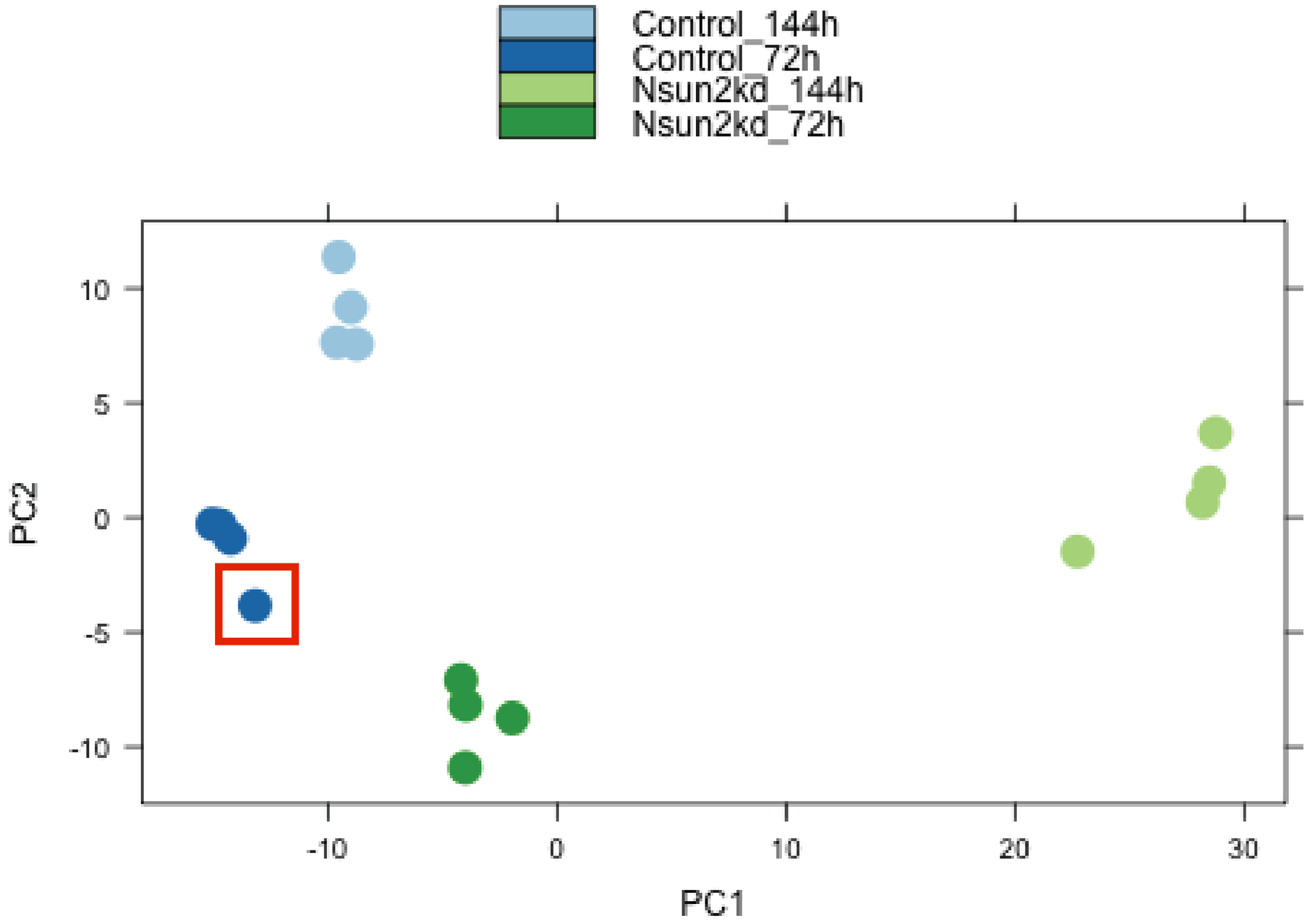


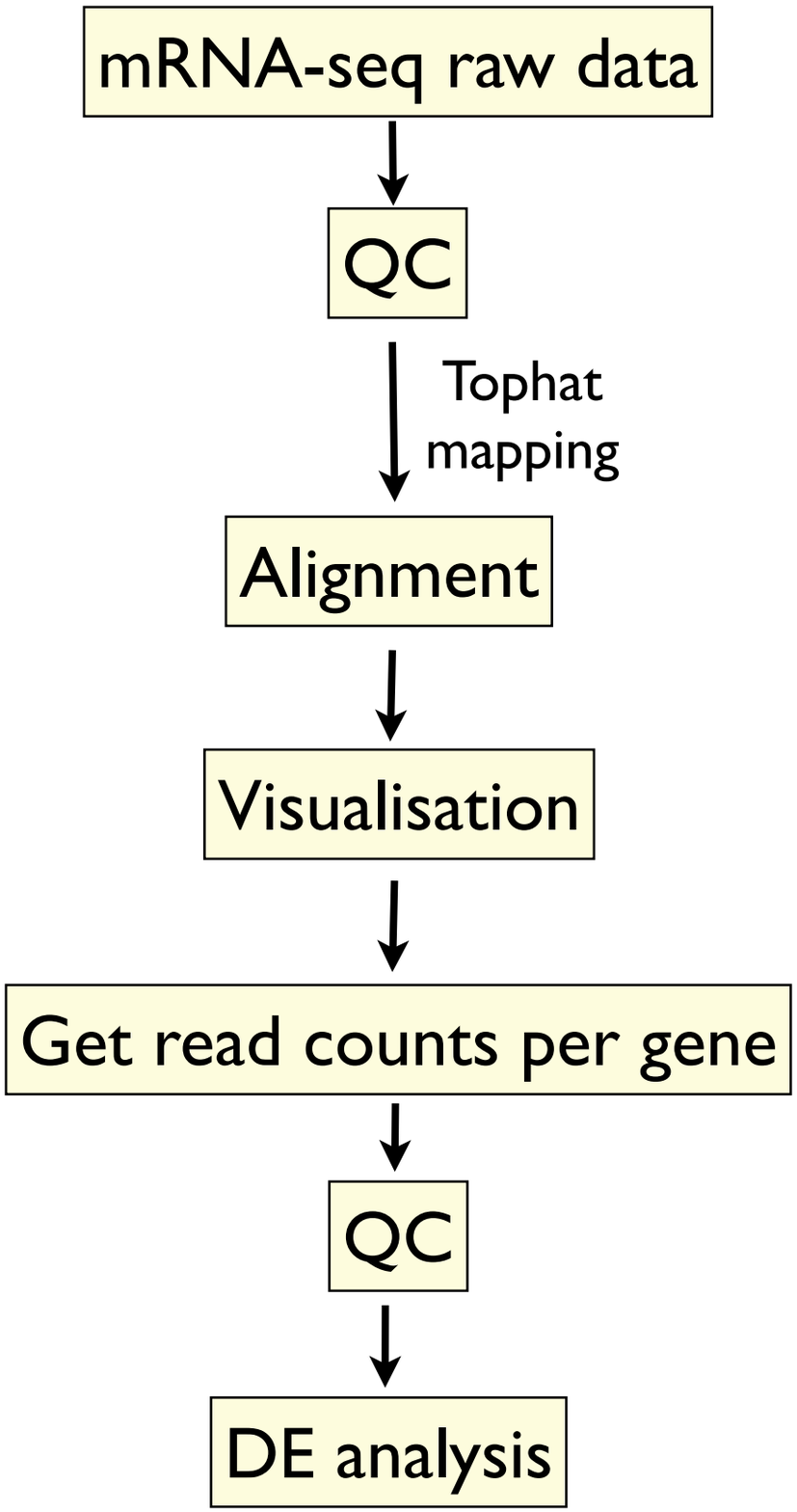


Comparison of replicates between conditions



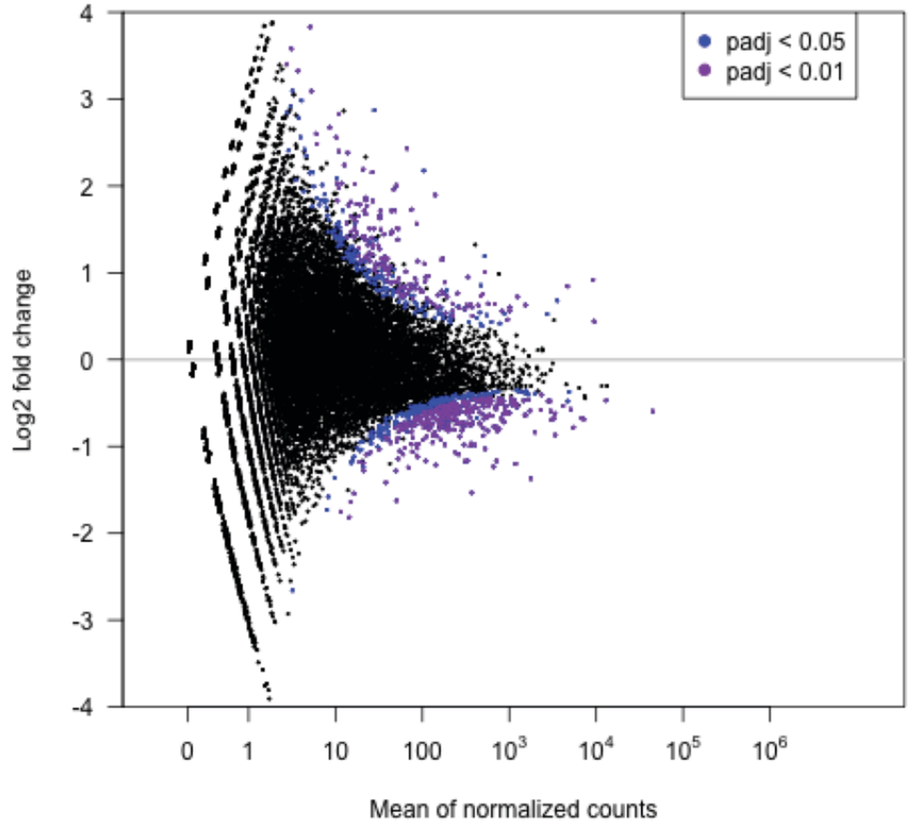

Sample distance
Height
0
5000
10000
15000
20000
25000
30000
Control_144h_b
Control_144h_c
Control_72h_d
Control_72h_b
Control_72h_c
Control_144h_a
Control_144h_d
Nsun2kd_144h_c
Nsun2kd_144h_b
Nsun2kd_144h_a
Nsun2kd_144h_d
Control_72h_a
Nsun2kd_72h_a
Nsun2kd_72h_b
Nsun2kd_72h_c
Nsun2kd_72h_d
distance
hclust (*, "average")


Control_144h
Control_72h
Nsun2kd_144h
Nsun2kd_72h
10
5
0
-5
-10
PC2
-10
0
10
20
30
PC1

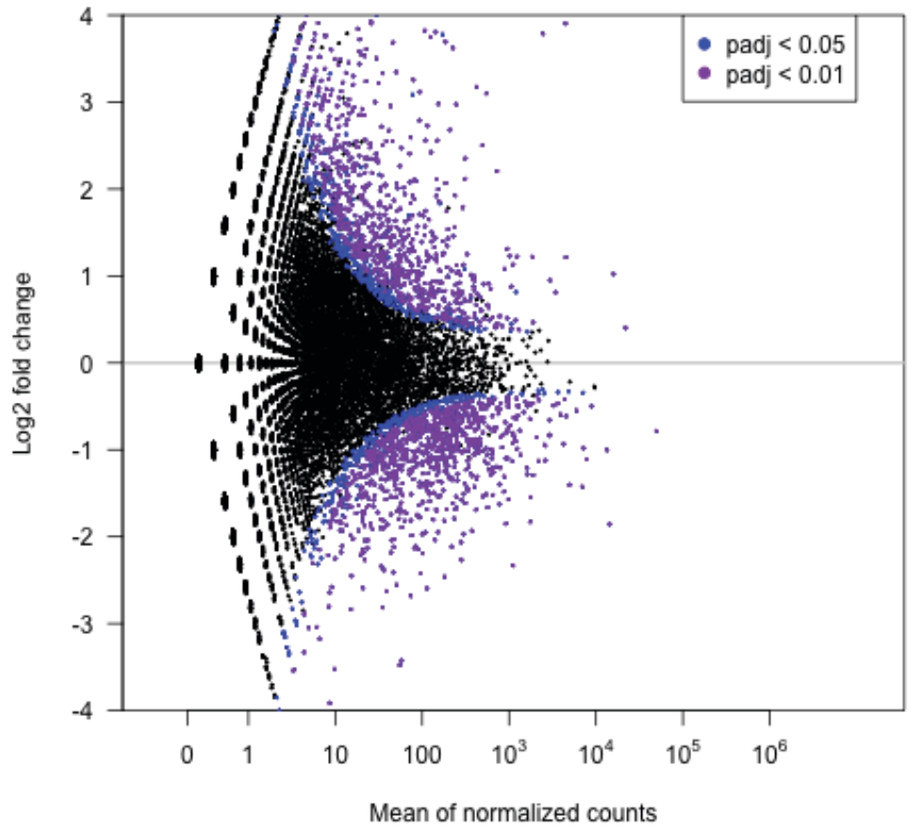


There are a number of packages for performing differential expression analysis. A popular one we will be covering this afternoon is an R package called DESeq.

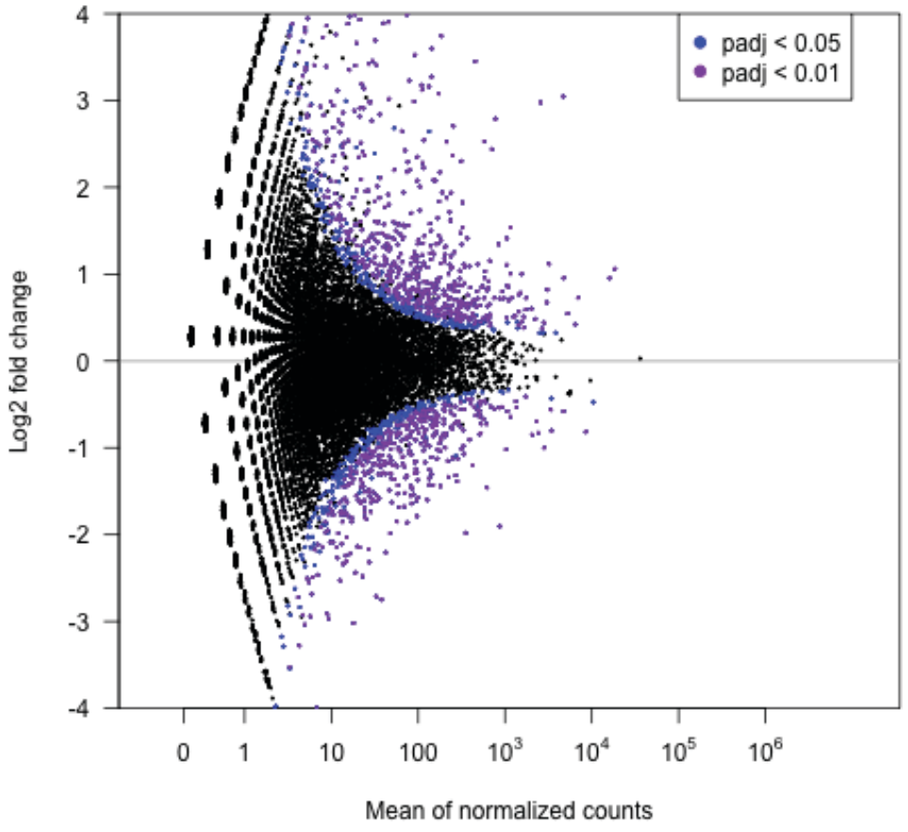
Nsun 72h vs control



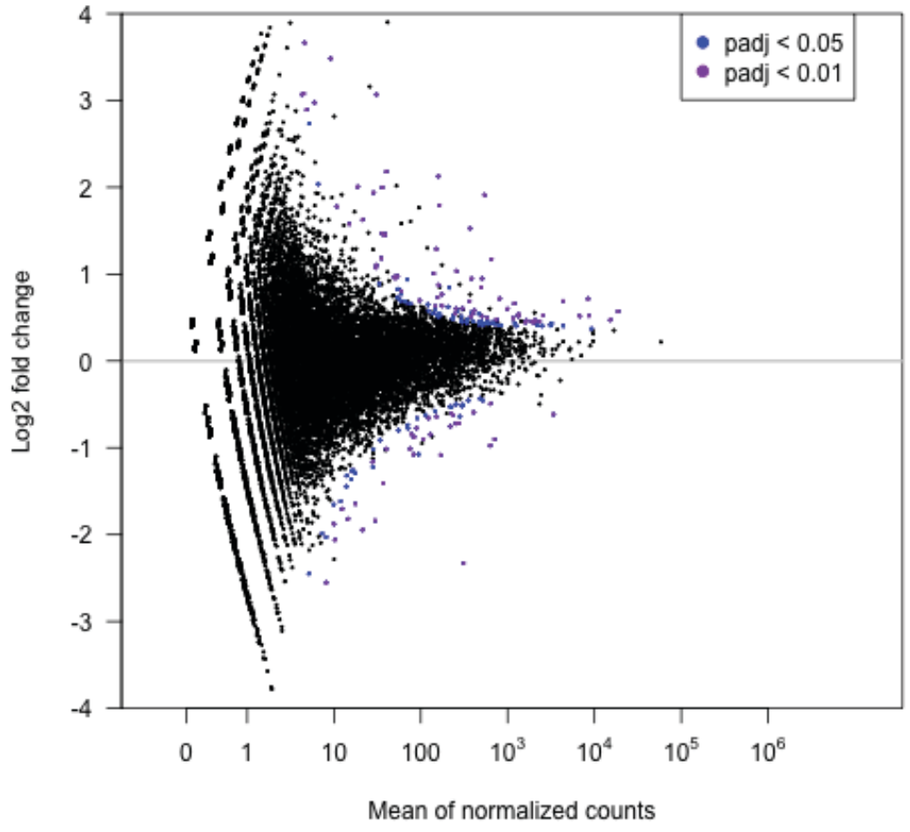
Nsun 144h vs control



Nsun 144h vs Nsun 72h



144h control vs 72h control



Differentially expressed genes at p-value < 0.01

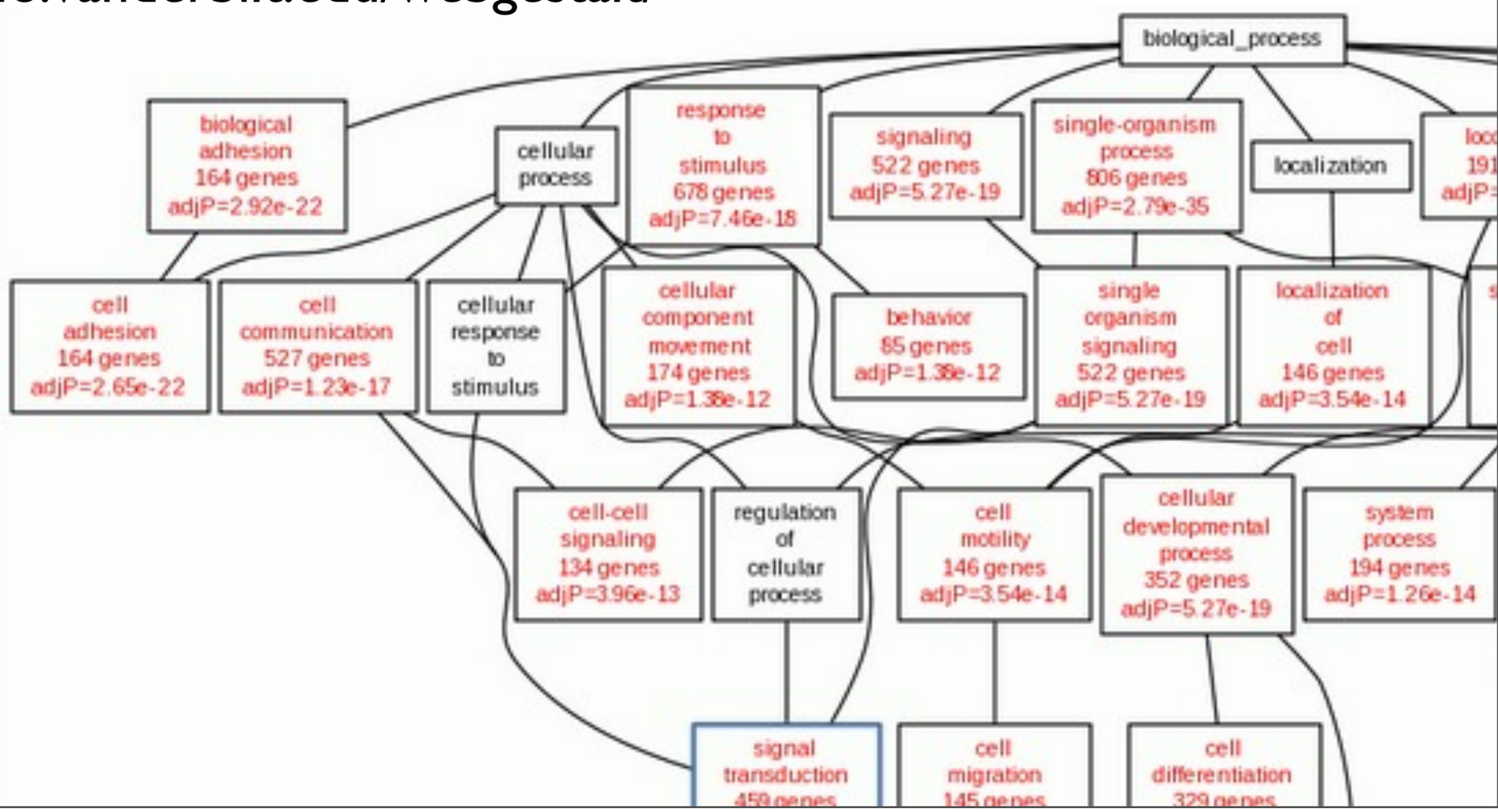
	Nsun 72h	Nsun 144h	Nsun 144 vs 72h	Control 144 vs 72h
Upregulated	174	449	409	26
Downregulated	377	744	325	24

Downstream analysis:

GO enrichment

- Extremely common way of checking the contents of your list of differentially expressed genes.

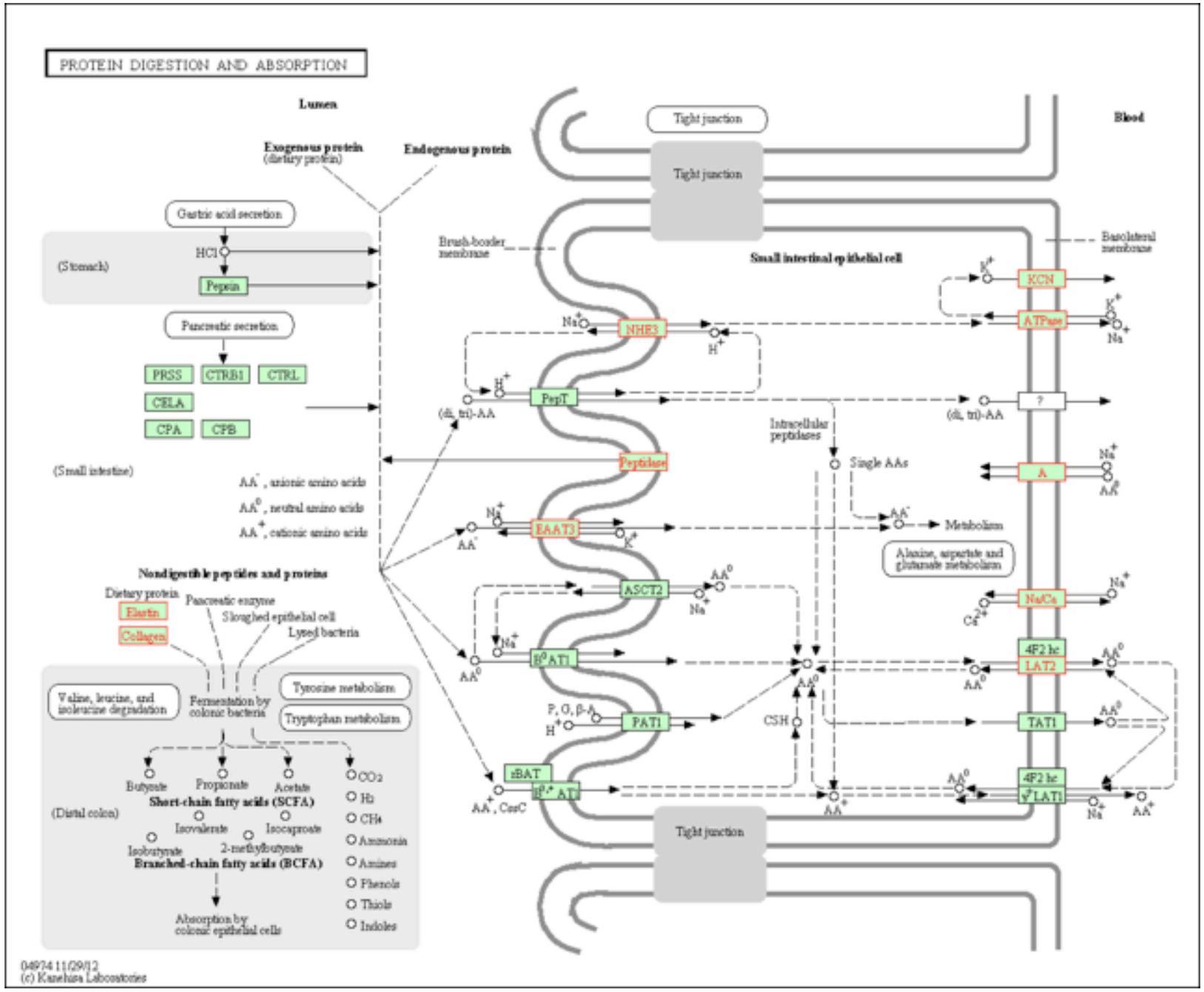
- Can be performed using R packages. There are also a large number of websites that will perform this for you. Here is one of them: http://bioinfo.vanderbilt.edu/webgestalt/



Downstream analysis

Pathway enrichment

- Another way of getting an idea of what your collection of differentially expressed genes might be doing.



Concluding remarks

- This lecture gives you an overview of how to pre-process RNAseq data, along with code examples. For a more detailed example, the code I use for analysis is at https://github.com/jelena121/RNAseq_scripts

- Differential expression is a very common application. However, it is also just the beginning - there is a lot of other downstream analysis you can perform with RNAseq data.

- For more ideas - flick through some RNAseq papers. If you find a paper particularly exciting, see what they've done and whether you might be able to apply it to your data too.