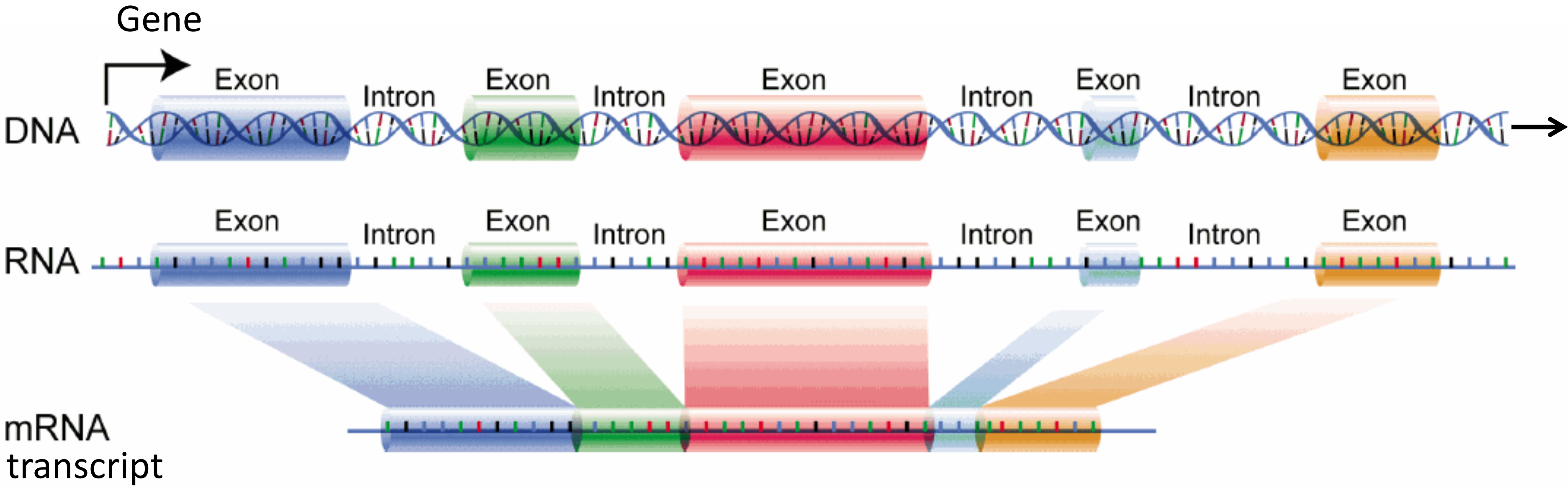
RNA-seq basics: From reads to differential expression

COMBINE RNA-seq Workshop

RNA sequencing (RNA-seq)

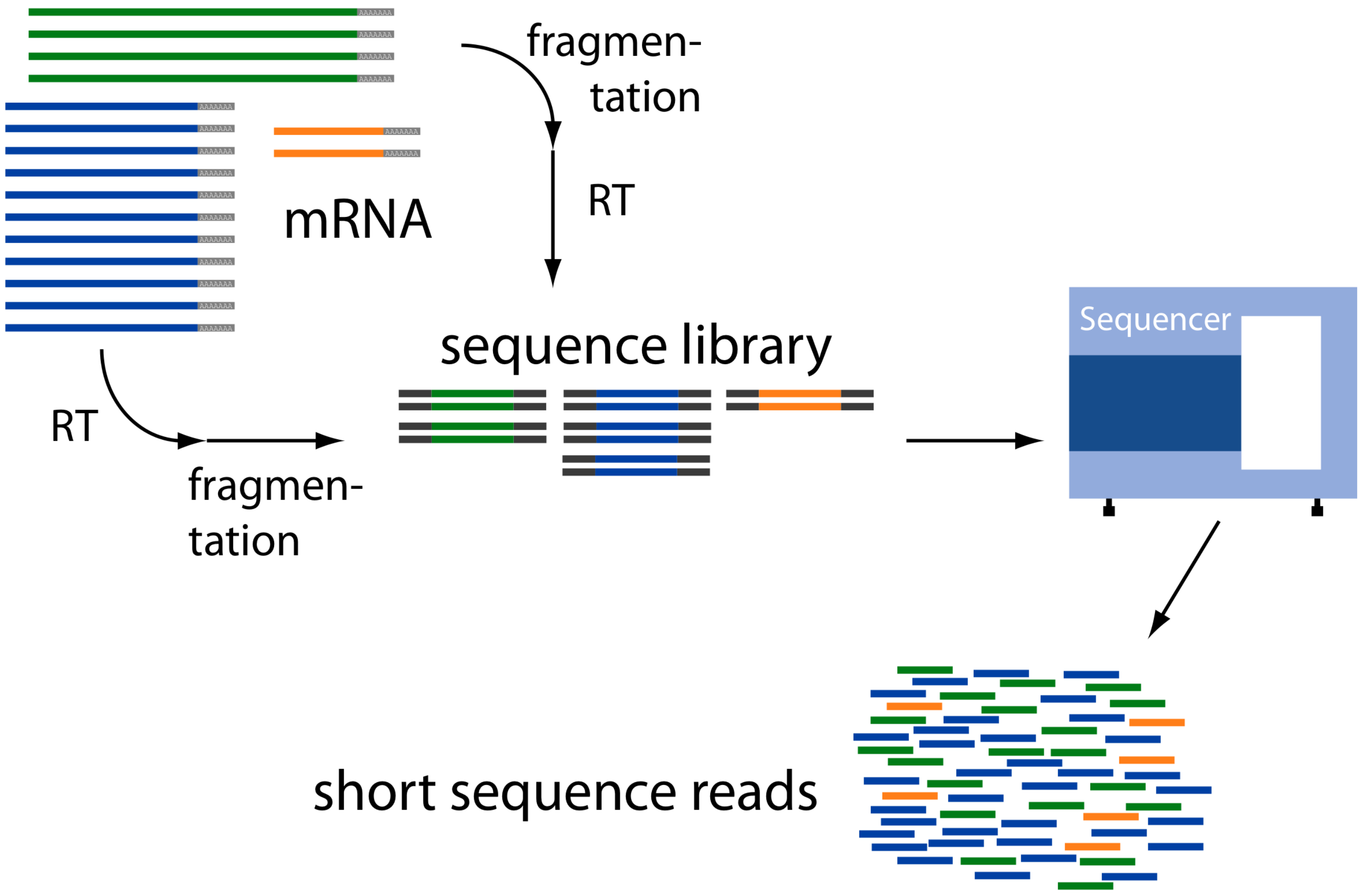
- Use of ultra high-throughput sequencing (‘next-’ or ‘second’-generation) technologies to study gene expression
- Many applications: differential expression, transcript discovery, splice variants, allele-specific expression
- In this hands-on course, you will learn how to use statistical methods to assess differential expression in RNA-seq data using popular tools in R/Bioconductor

Genes and transcripts



Slide from Alicia Oshlack

From transcripts to short reads



mRNA

fragmentation

RT

RT

fragmentation

sequence library

Sequencer

short sequence reads

Pepke et al, Nature Methods, 2009

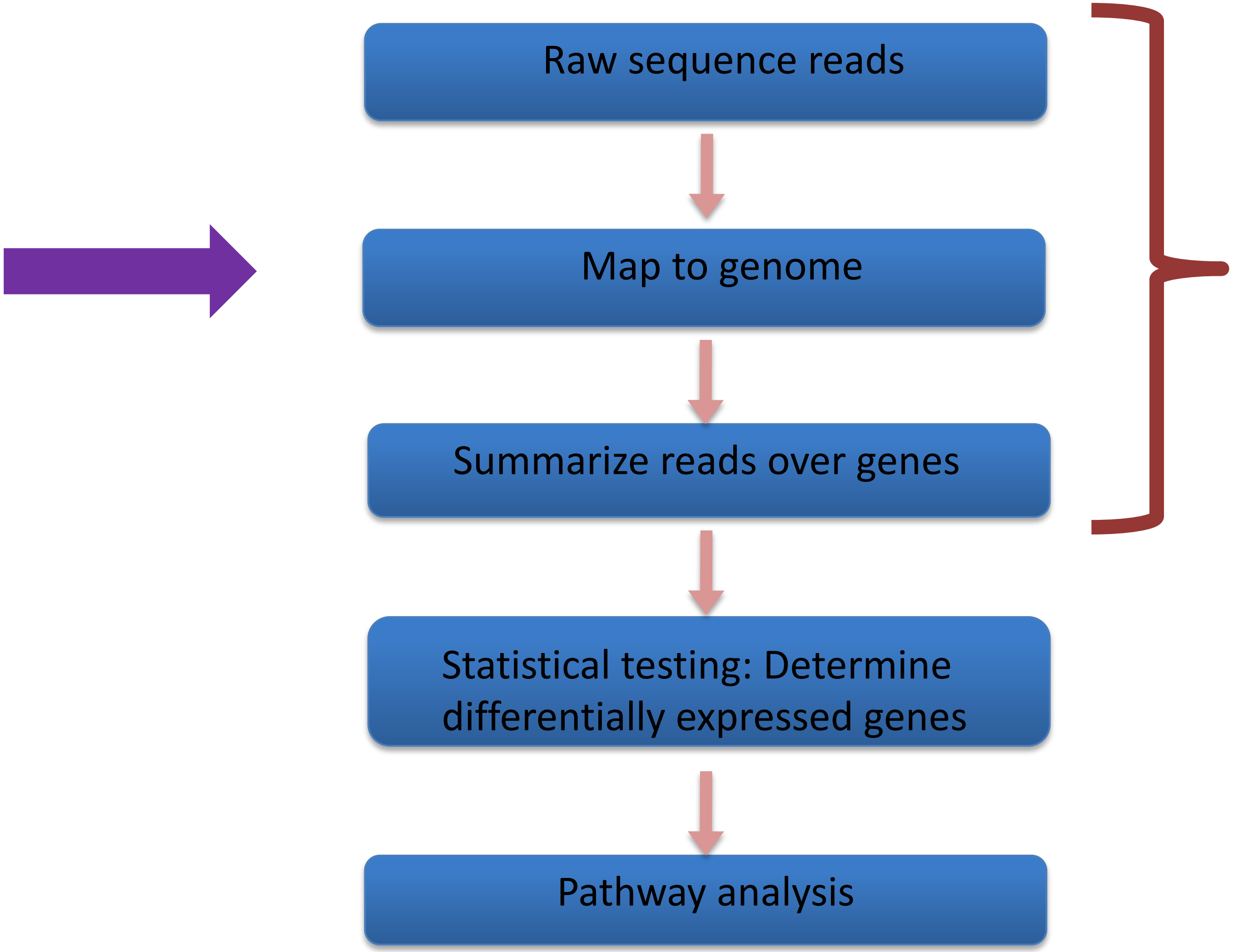
Raw data comes in fastq files

- Short sequence reads
- Quality scores

50 bp sequence

```
@HWI-ST1148:308:C694RACXX:5:1101:1768:1990 1:N:0:CGTACG
NTAGGCCTTGGCAGTTTTGGAGAATCACTGCTGCCAAAGAGTCTACTTGG
+
#0<FFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFIIIIIII
@HWI-ST1148:308:C694RACXX:5:1101:3409:1990 1:N:0:CGTACG
NAGTTACCCTAGGGATAACAGCGCAATCCTATTCTAGAGTCCATATCAAC
+
#000BFBFFFFFFF<BFFFFBBBBBFBBFF<<FBFFIBFFFBFFFIIBFF
```

RNA-seq analysis steps



Raw sequence reads

Map to genome

Summarize reads over genes

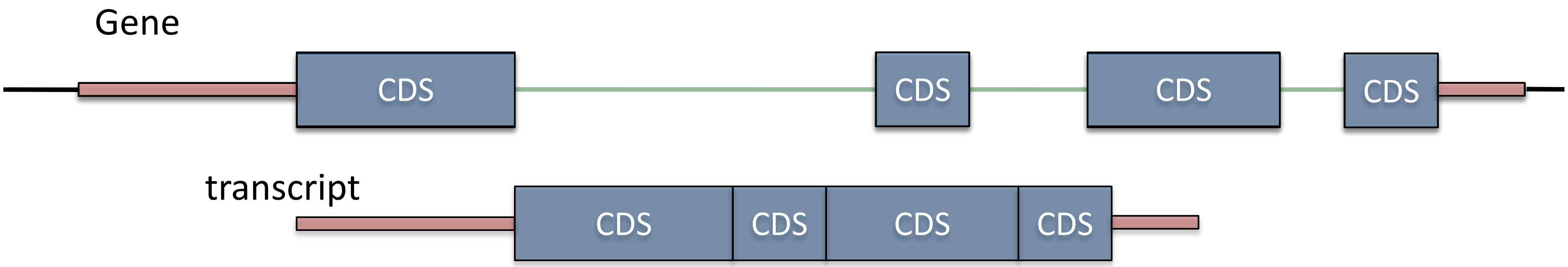
Statistical testing: Determine differentially expressed genes

Pathway analysis

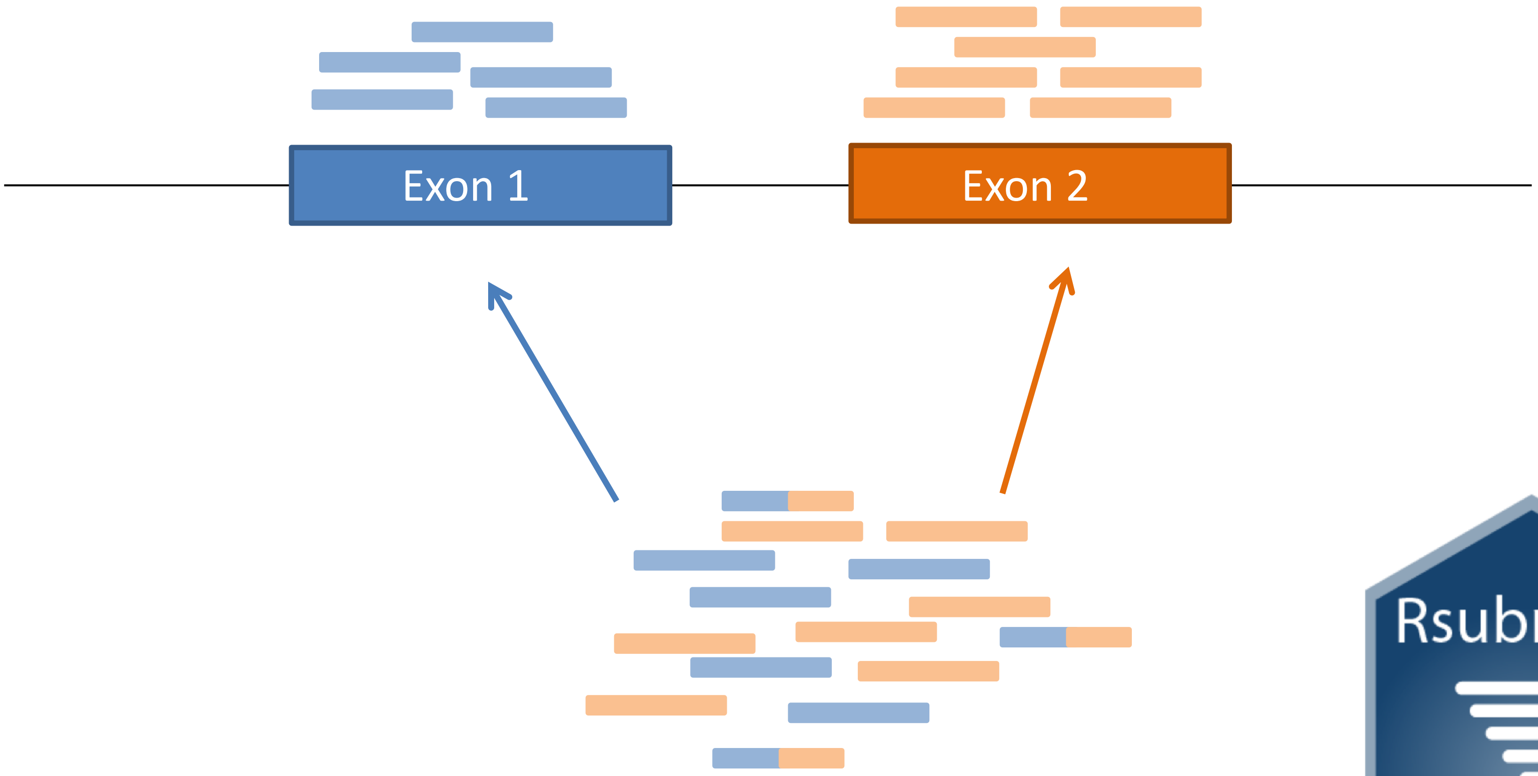
Slide from Alicia Oshlack

Mapping reads to the genome

- Where do the millions of short sequences come from in the genome?
- Sequencing transcripts, not the genome



Lots of good aligners handle splice junctions well

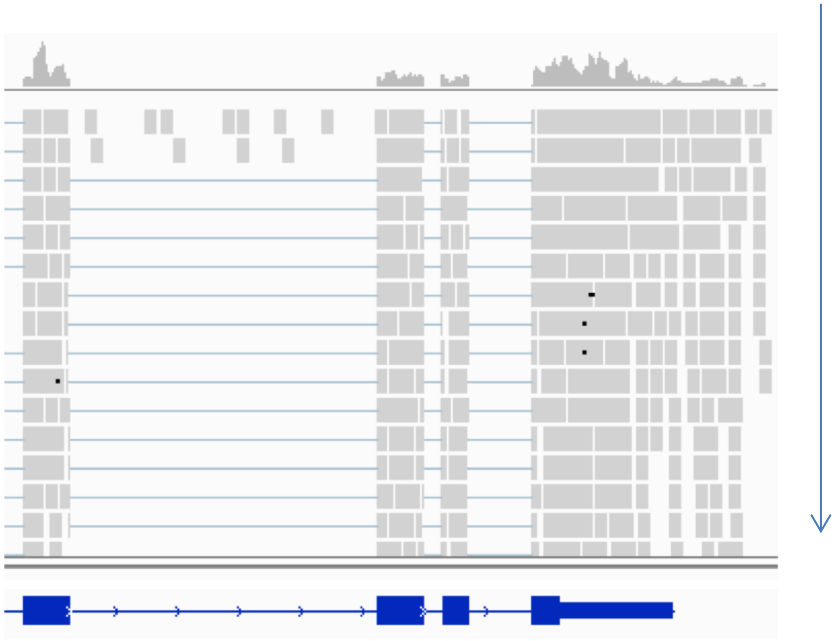


Rsubread

Aligned reads (bam files)

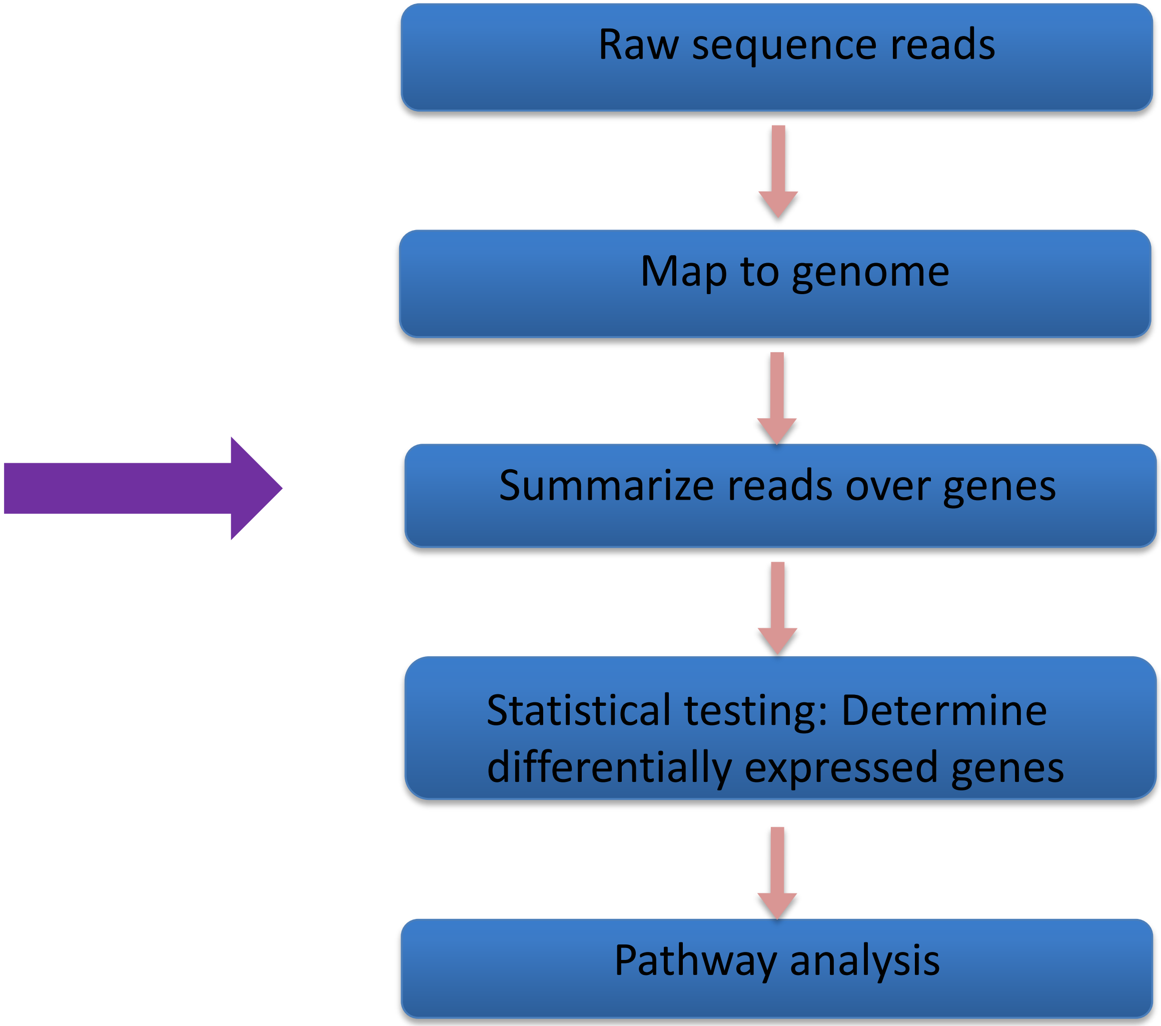
```
HWI-ST1148:308:C694RACXX:6:2209:15171:26188     272     chr10   76314   0       50M     *
0       0       CATCTGATCTTTGACAAACCTGACAAACACAAGCAATGGGGAAAGGATTC
IIIIIIIIIIIIIIIIIFIFFFIIIIIIIIIIIIIIIFFFFFFFFFFBBB      NH:i:10 HI:i:6  AS:i:49 nM:i:0

HWI-ST1148:308:C694RACXX:6:2306:17518:85846     272     chr10   76315   0       50M     *
0       0       ATCTGATCTTTGACAAACCTGACAAACACAAGCAATGGGGAAAGGATTCC
BFIIIIIIFFIFFBFFFFFFFFFFFFFIIIFFFFFFIFFFFFFFFFFBBB      NH:i:10 HI:i:7  AS:i:49 nM:i:0
```



A row for each sequence
Millions of rows...

RNA-seq analysis steps



Raw sequence reads

Map to genome

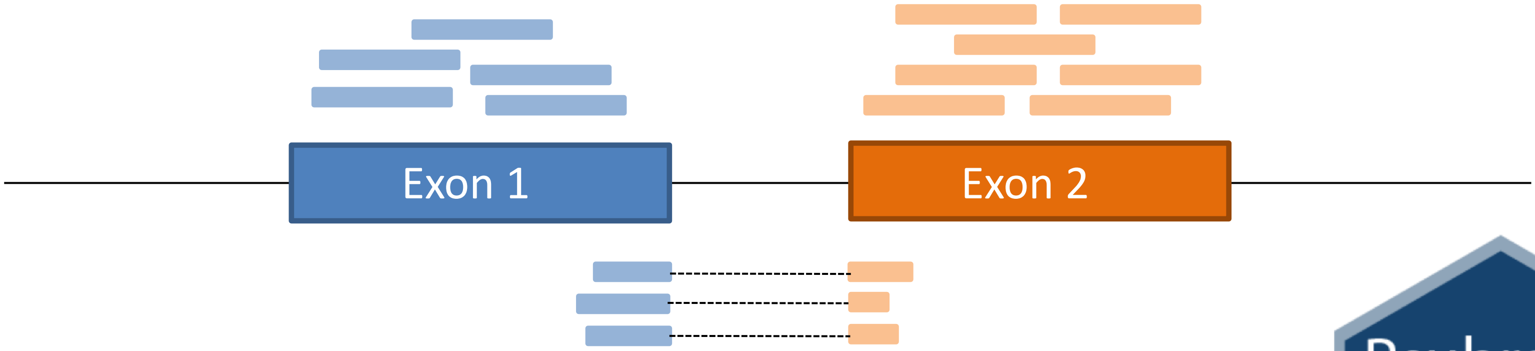
Summarize reads over genes

Statistical testing: Determine differentially expressed genes

Pathway analysis

Slide from Alicia Oshlack

Counting over exons vs counting over genes



Exon 1 = 8 reads
Exon 2 = 10 reads

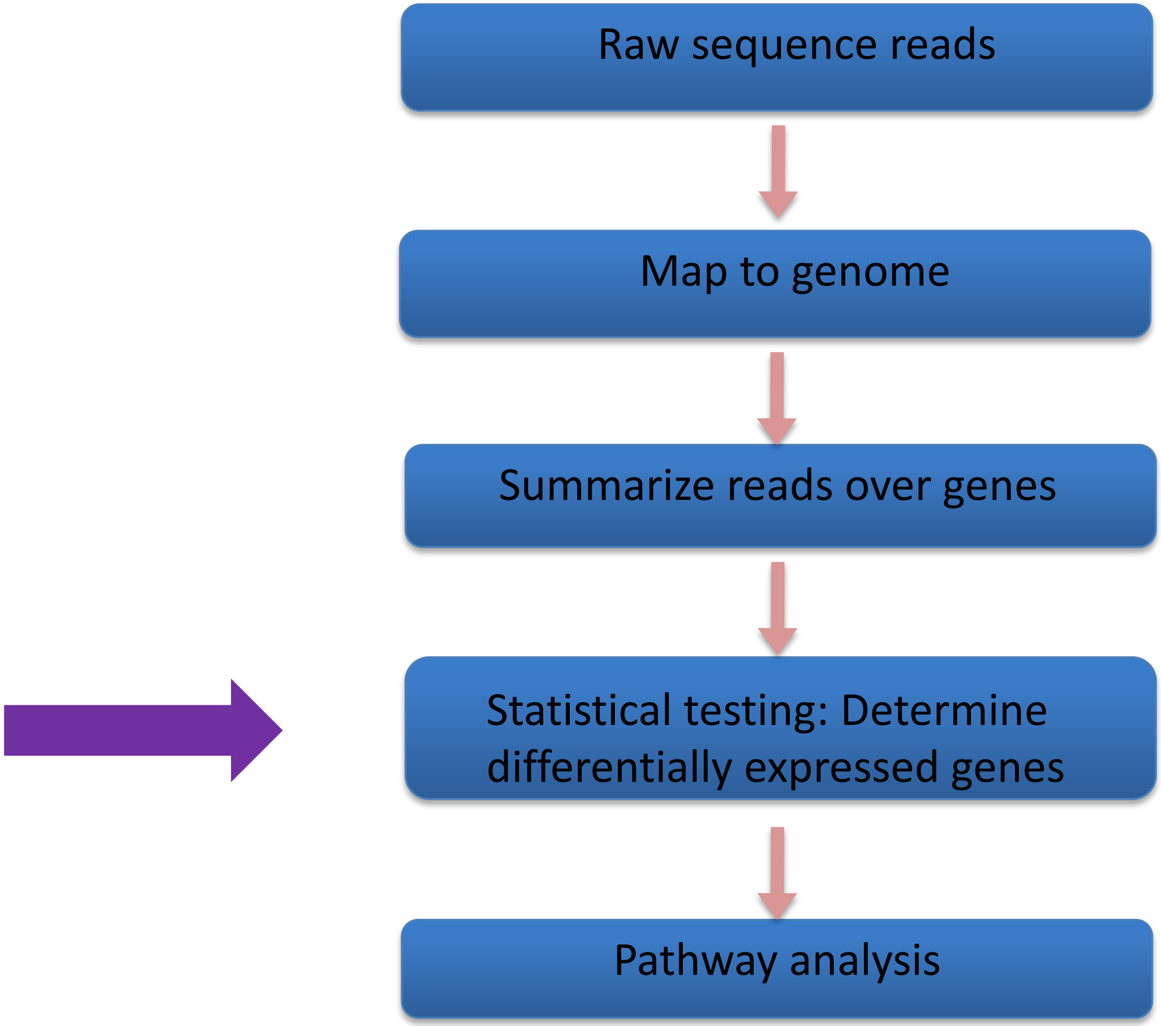
Counting over whole gene (Exon1 + Exon2) = 15

Summarization turns mapped reads into a table of counts

Tag ID	A1	A2	B1	B2
ENSG00000124208	478	619	4830	7165
ENSG00000182463	27	20	48	55
ENSG00000125835	132	200	560	408
ENSG00000125834	42	60	131	99
ENSG00000197818	21	29	52	44
ENSG00000125831	0	0	0	0
ENSG00000215443	4	4	9	7
ENSG00000222008	30	23	0	0
ENSG00000101444	46	63	54	53
ENSG00000101333	2256	2793	2702	2976
…	… tens of thousands more tags …			

** very high dimensional data **

RNA-seq analysis steps



Raw sequence reads

Map to genome

Summarize reads over genes

Statistical testing: Determine differentially expressed genes

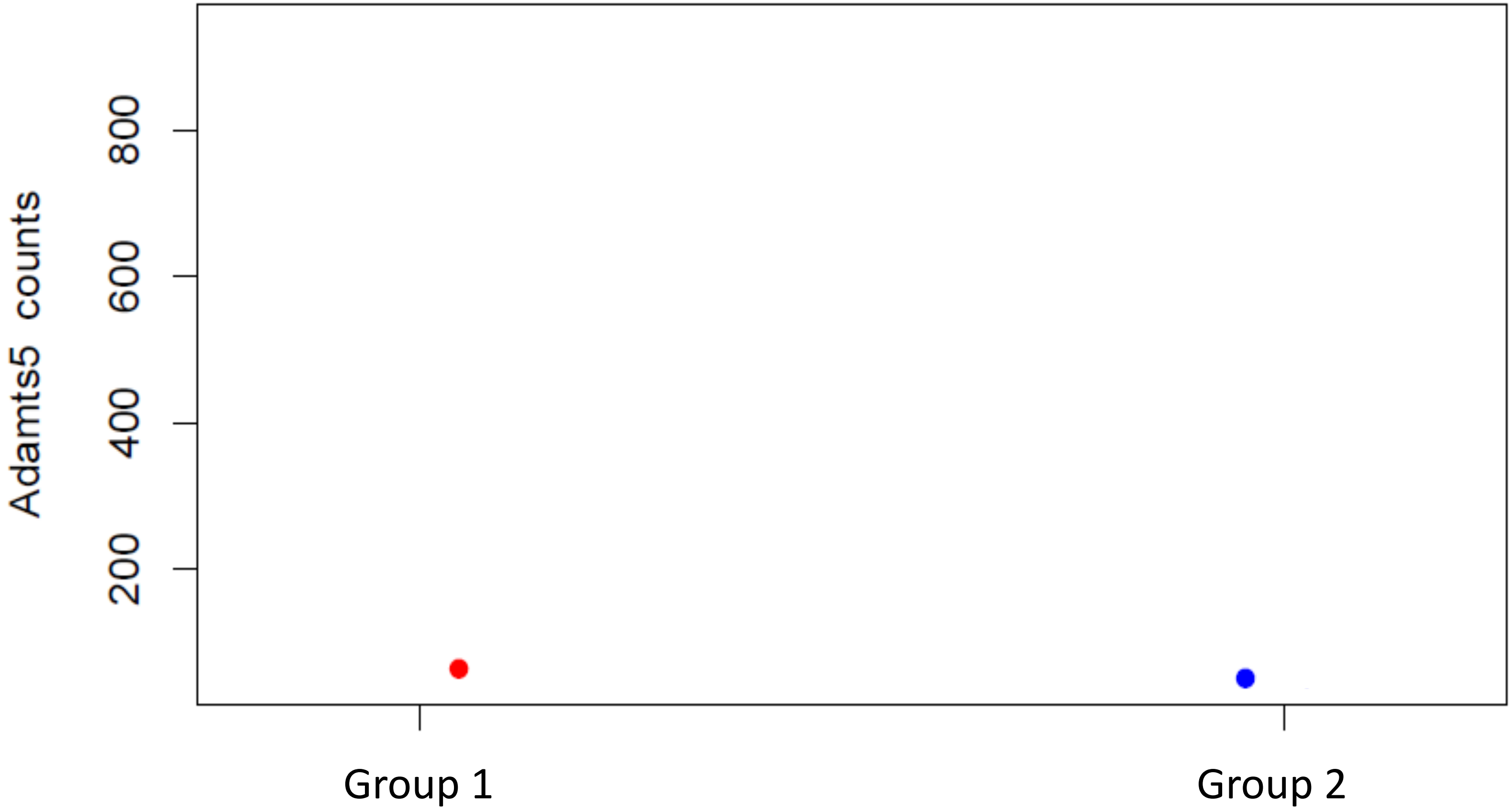
Pathway analysis

Slide from Alicia Oshlack

Assessing differential expression

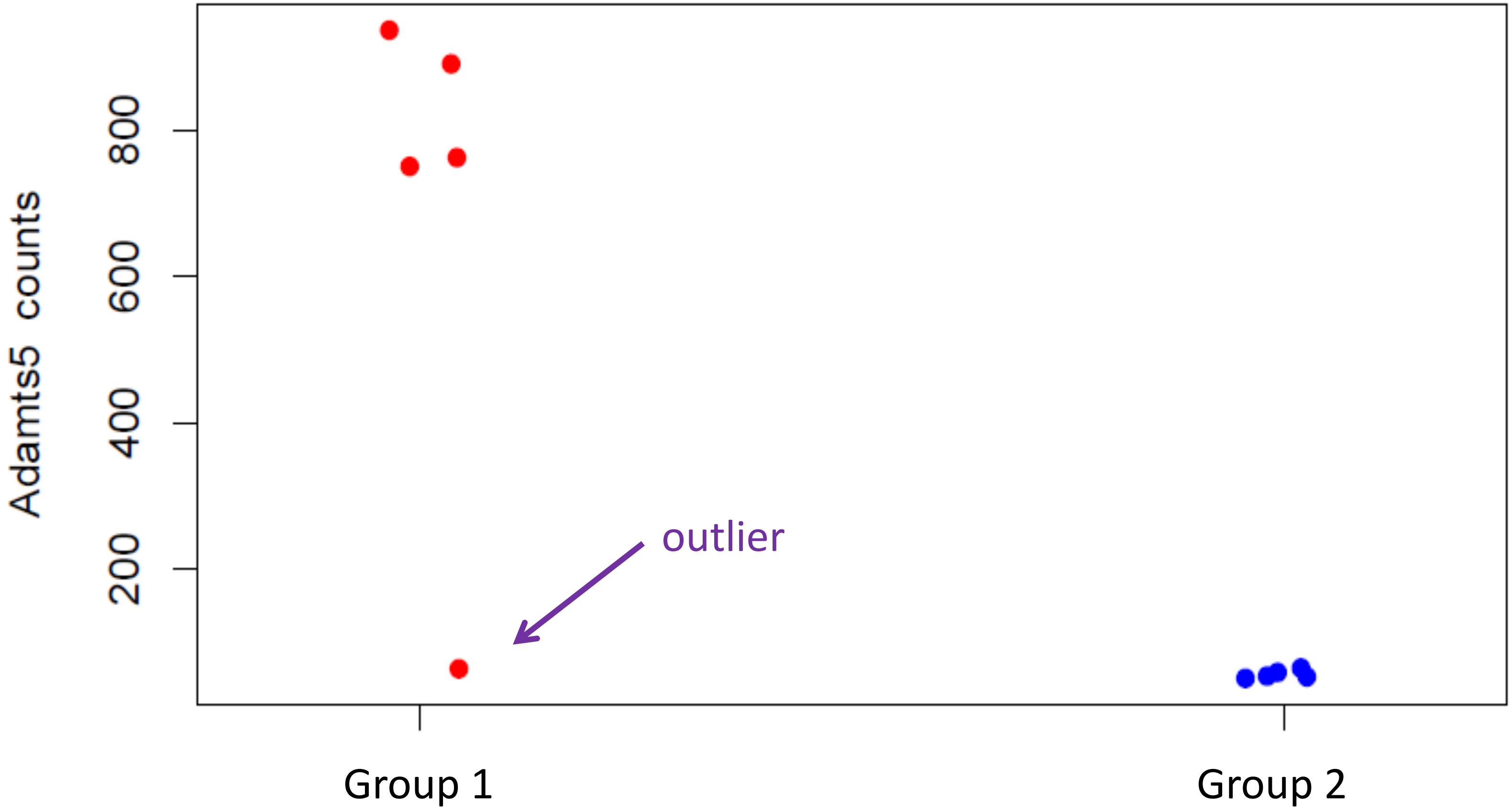
- For each gene in each sample we have a measure of abundance
  - Number of reads mapping across gene
- We want to know whether there is a statistically significant difference in abundance between treatments/groups/genotypes

Is this gene differentially expressed?



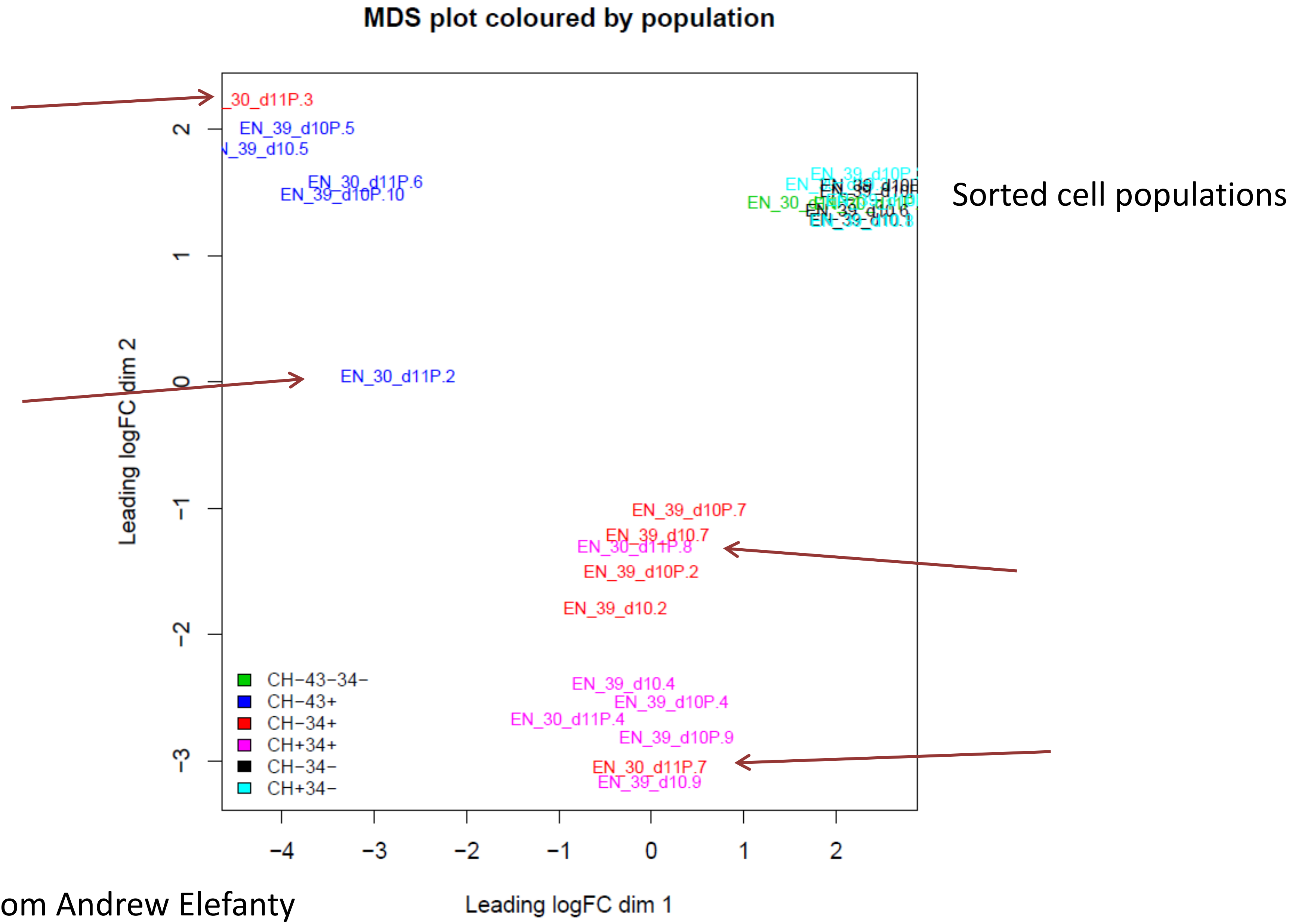
Data from Shireen Lamande

Is this gene differentially expressed?



Replication is really important!

Quality control – check your data!

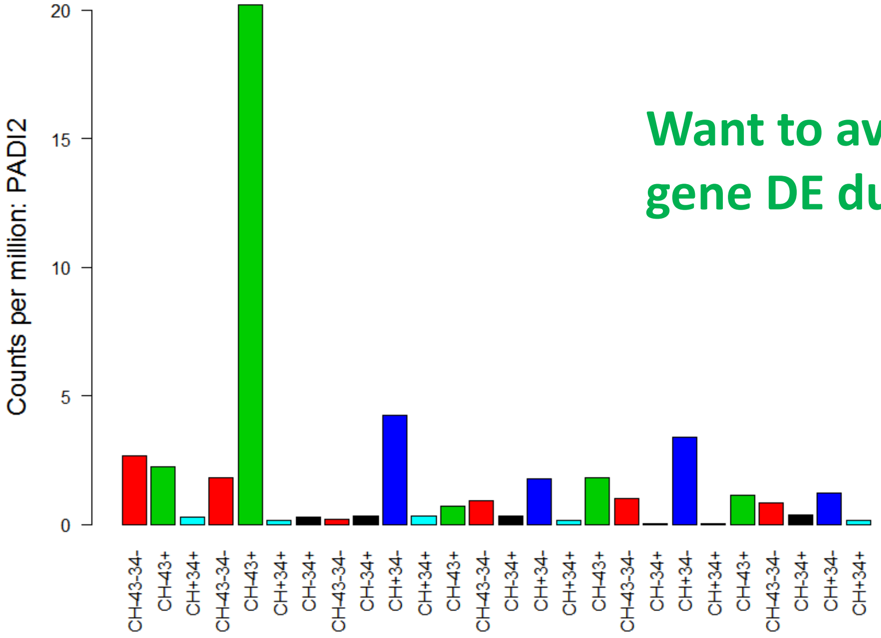


Sorted cell populations

Data from Andrew Elefanty

Things to think about before statistical testing

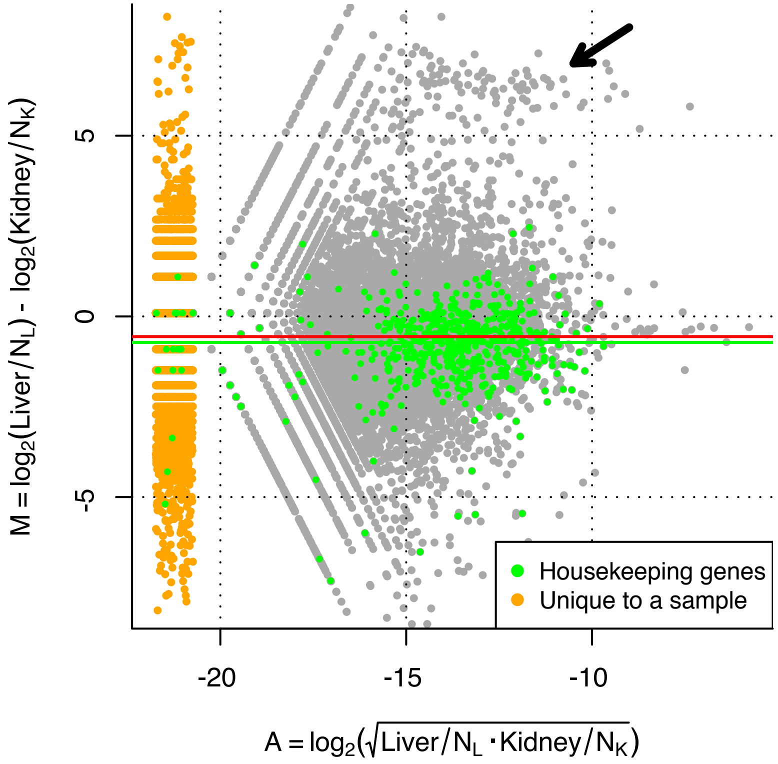
- Filtering out lowly expressed genes
  - Need to make decisions about cut-offs
  - Can be an iterative process



Want to avoid calling this gene DE due to one sample

Things to think about before statistical testing

- Normalisation
  - Library size (sequencing depth)
  - Composition bias (TMM)
  - Batch effects (RUVSeq)



Statistical testing for DE

- For each gene, is the mean expression level under one condition significantly different from the mean expression level under a different condition?

Tag ID	A1	A2	B1	B2
ENSG00000124208	478	619	4830	7165
ENSG00000182463	27	20	48	55
ENSG00000125835	132	200	560	408
ENSG00000125834	42	60	131	99
…	… tens of thousands more tags …			

Many different statistical methods

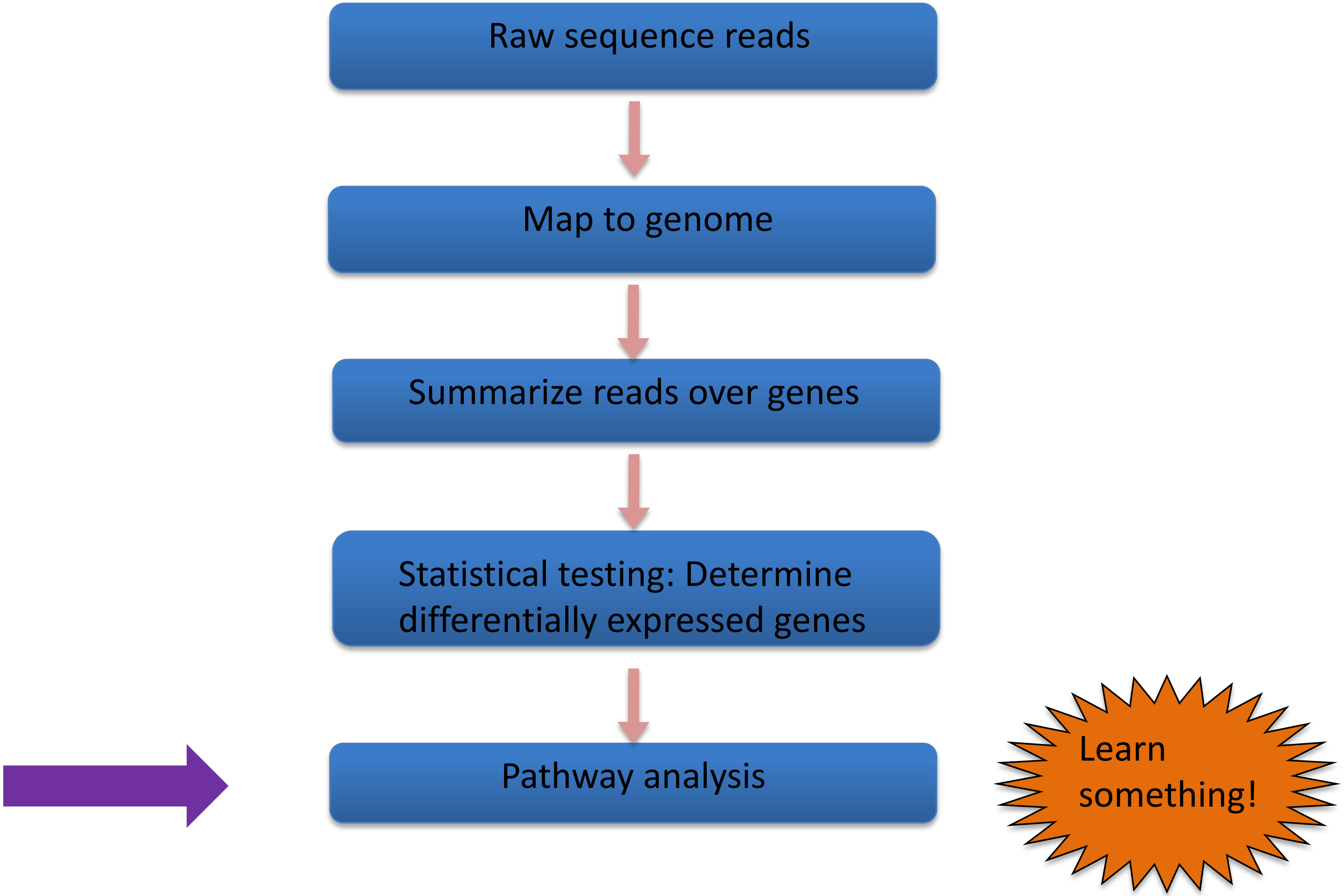
- Model the counts directly
  - Negative binomial modelling is best because it captures biological as well as technical variability
  - Most popular packages in R
    - *edgeR*
    - *DESeq/DESeq2*
    - Lots of others exist (*baySeq*, *NBPSeq*, …)
- Transform the counts and used normal based methods
  - voom in the *limma* package

Statistical testing gives each gene a p-value for evidence of DE

Tag ID	A1	A2	B1	B2
ENSG00000124208	478	619	4830	7165
ENSG00000182463	27	20	48	55
ENSG00000125835	132	200	560	408
ENSG00000125834	42	60	131	99
ENSG00000197818	21	29	52	44
ENSG00000125831	0	0	0	0
ENSG00000215443	4	4	9	7
ENSG00000222008	30	23	0	0
ENSG00000101444	46	63	54	53
ENSG00000101333	2256	2793	2702	2976
…	… tens of thousands more tags …			

Tag ID	P-value
ENSG00000124208	0.0002
ENSG00000182463	0.12
ENSG00000125835	0.034
ENSG00000125834	0.08
ENSG00000197818	0.64
ENSG00000125831	1
ENSG00000215443	1
ENSG00000222008	0.06
ENSG00000101444	0.73
ENSG00000101333	0.22
…	

RNA-seq analysis steps



Raw sequence reads

Map to genome

Summarize reads over genes

Statistical testing: Determine differentially expressed genes

Pathway analysis

Learn something!

Slide from Alicia Oshlack

Summary

- Lots of choices in analysis methodology
- Quality control is essential! Sometimes detective work is necessary.
- Each step of the analysis requires decisions that impact down-stream analysis
- Life gets harder when there’s no genome or poor quality genomes

RNA-seq analysis in R / Bioconductor

Acknowledgements

Slides:

- Alicia Oshlack
- Belinda Phipson
- Anthony Hawkins
- Gordon Smyth
- Davis McCarthy

Data:

- Andrew Elefanty and Elizabeth Ng
- Shireen Lamande