Transcriptome analysis using RNA-seq

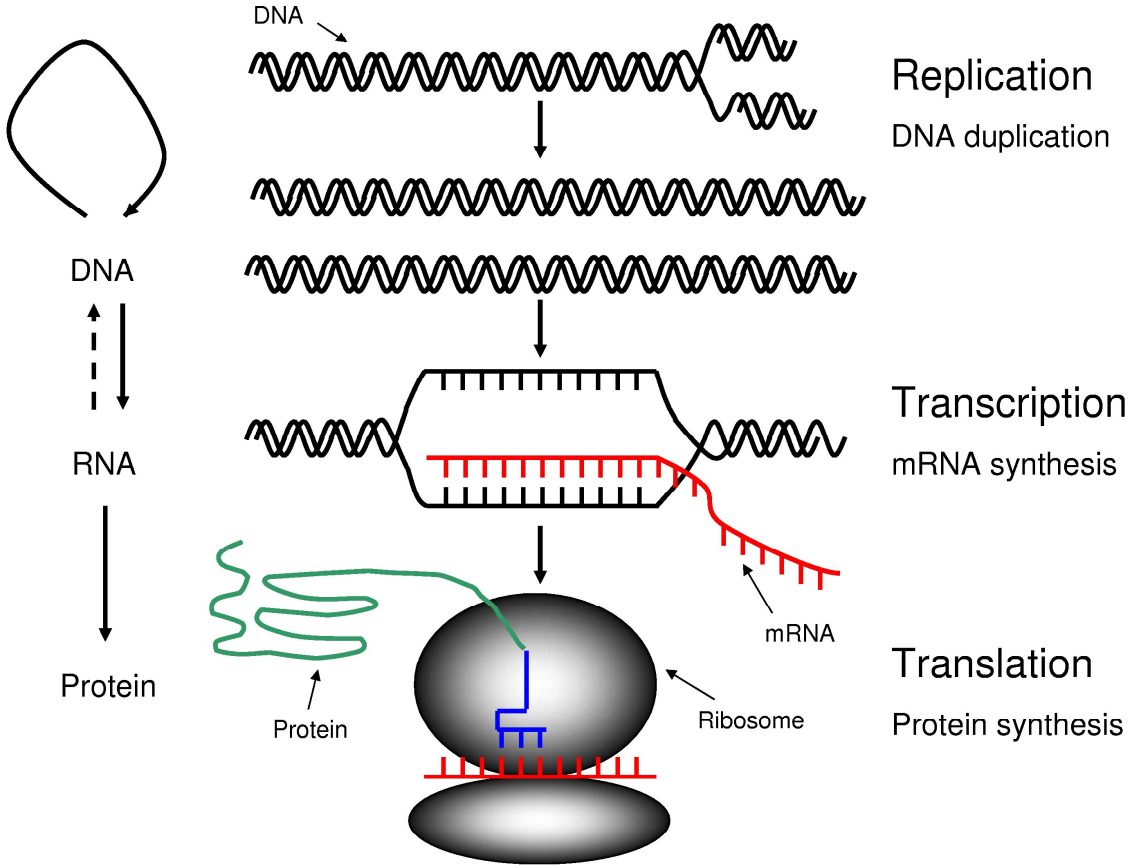
Adam Reid & Steve Doyle
Wellcome Sanger Institute/LSHTM

LSHTM Pathogen Genomics 2021

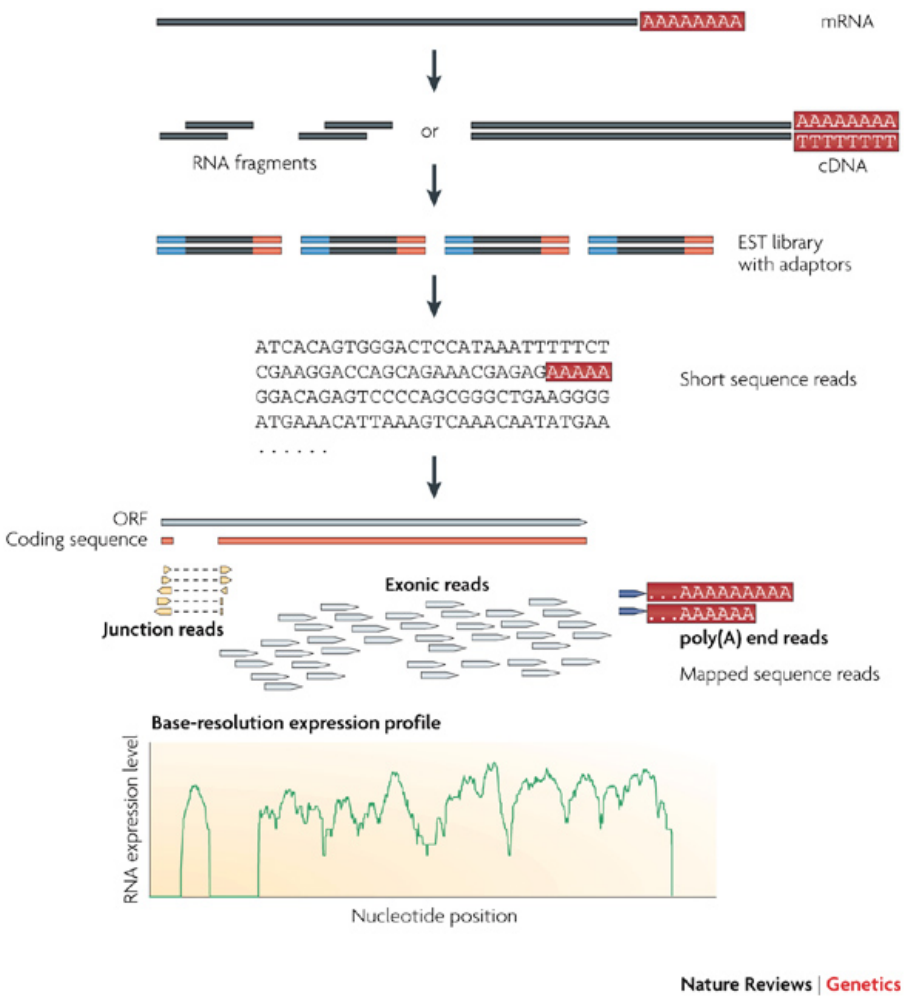
Summary

- RNA-seq background
- Mapping to the genome (*HISAT2* and *Artemis*)
- Mapping to the transcriptome and counting reads (*Kallisto*)
- Read count normalisation
- Differential expression (*Sleuth*)
- What to do with a gene list
- The exercise

Gene expression



RNA sequencing



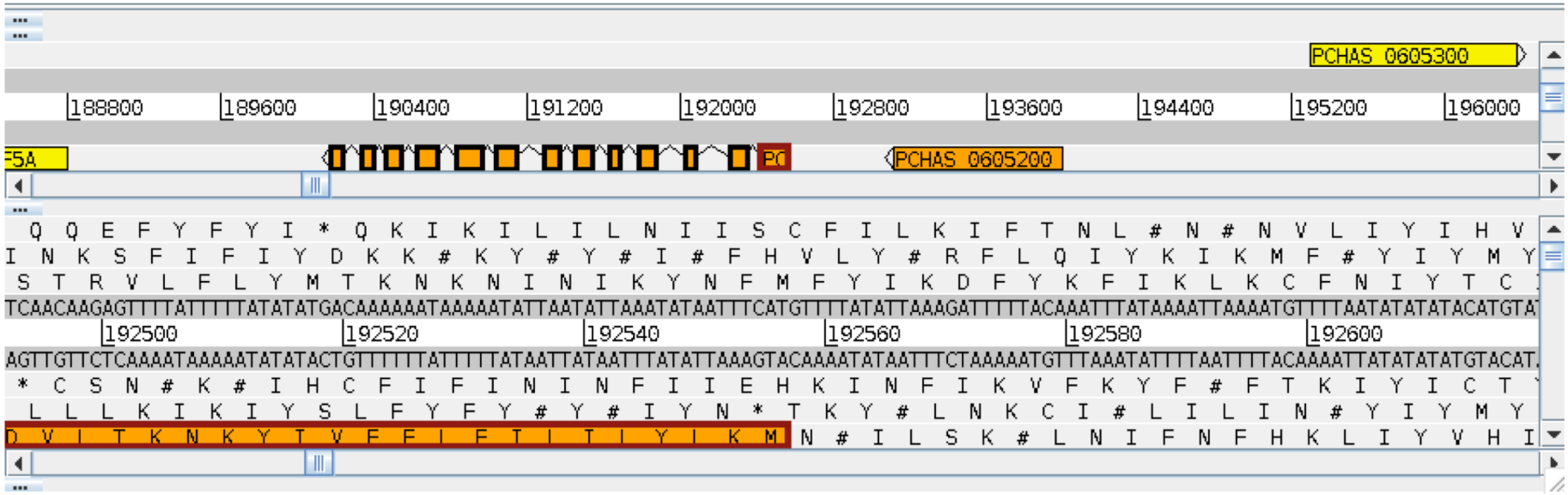
Experimental design

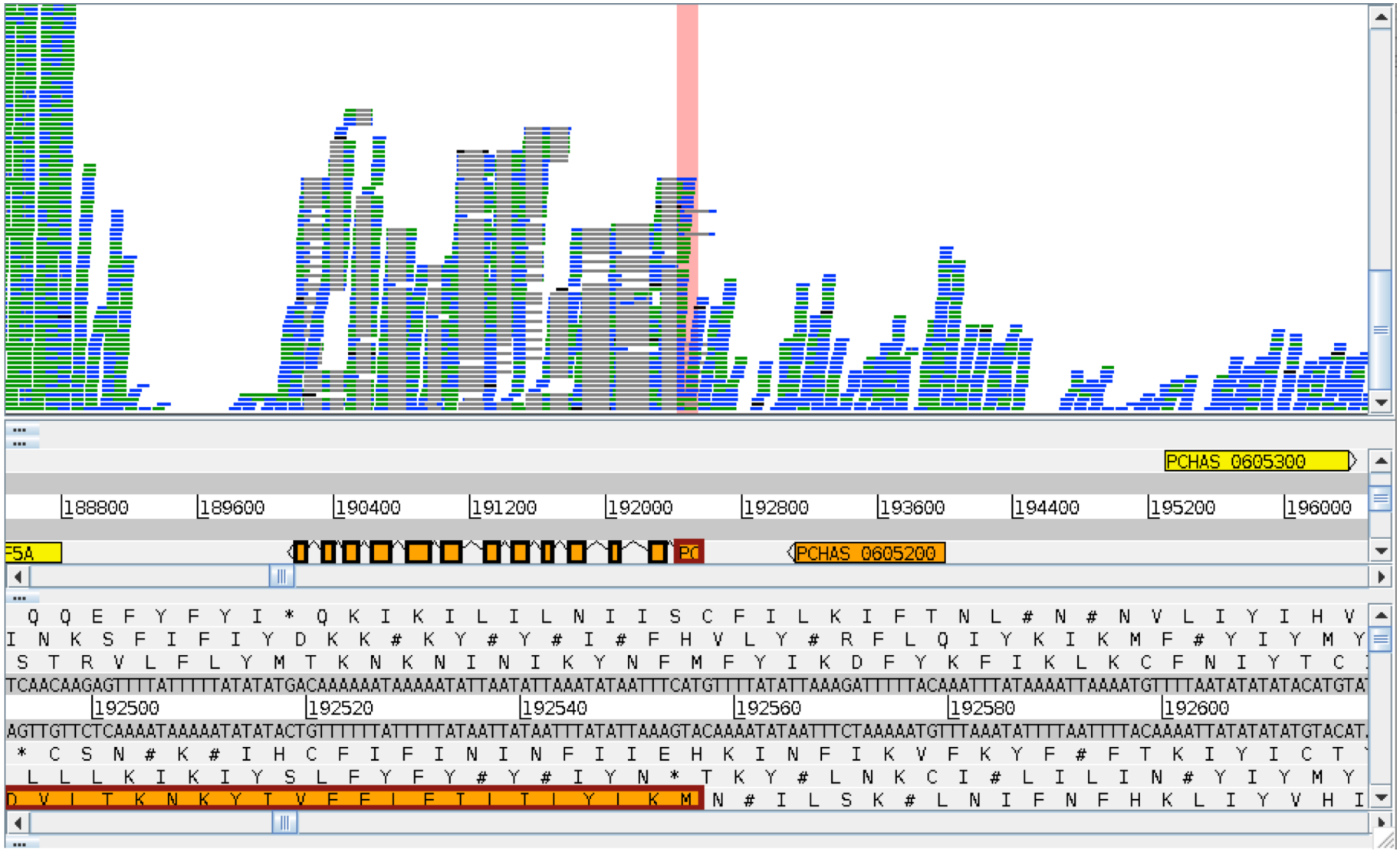
- Replicates
  - Relevant biological replicates are required
  - Technical replicates are not generally required, but try to arrange samples on plates to minimise potential problems
- Sequencing depth
  - Practical considerations e.g. amount of data on one lane, number of barcodes/tags
  - Suggested 2-5Gb for human, 0.5-1Gb for *Plasmodium*, but depends greatly on complexity of samples and how obvious the interesting biology is

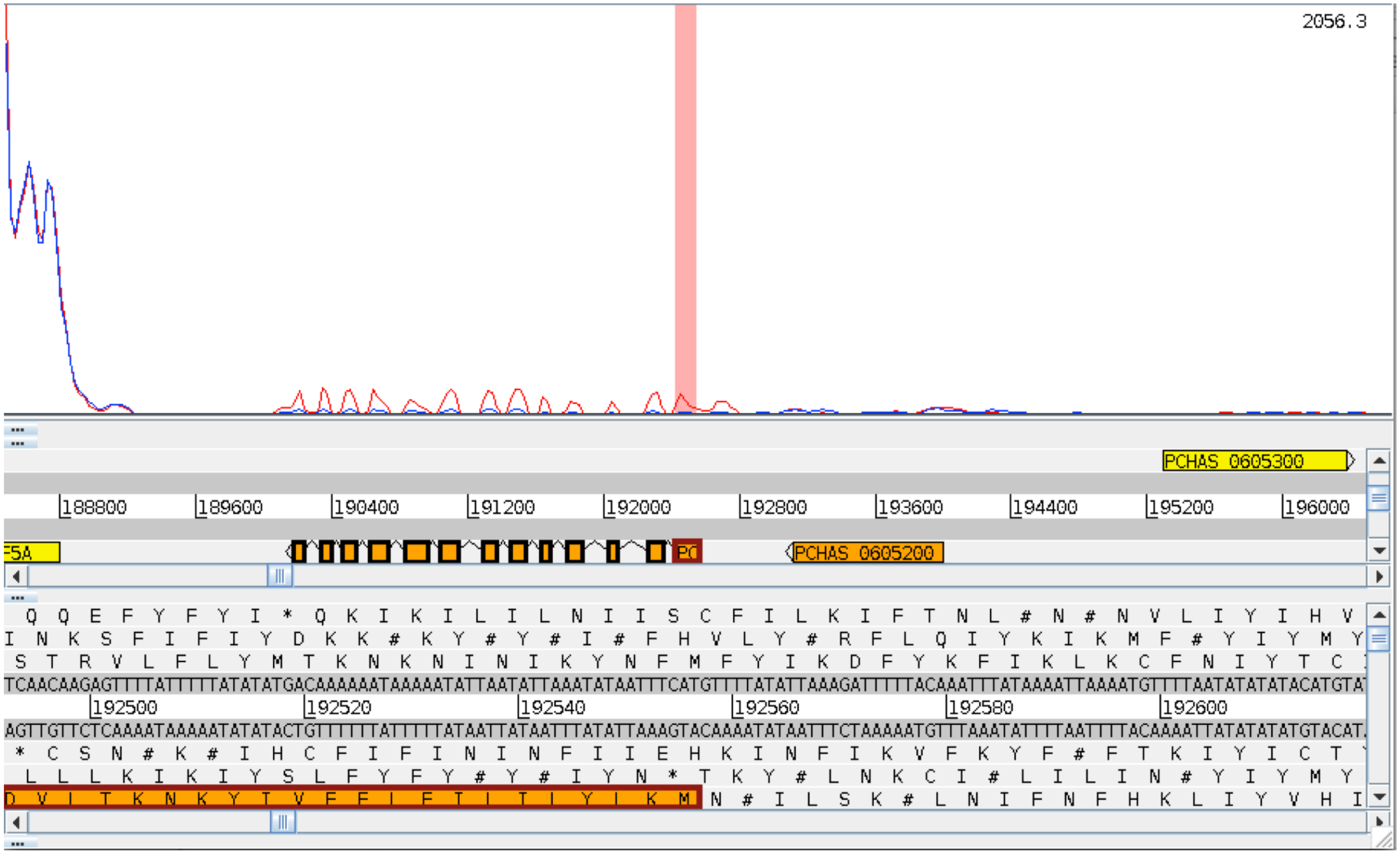
Mapping RNA-seq reads to the genome (HISAT2)

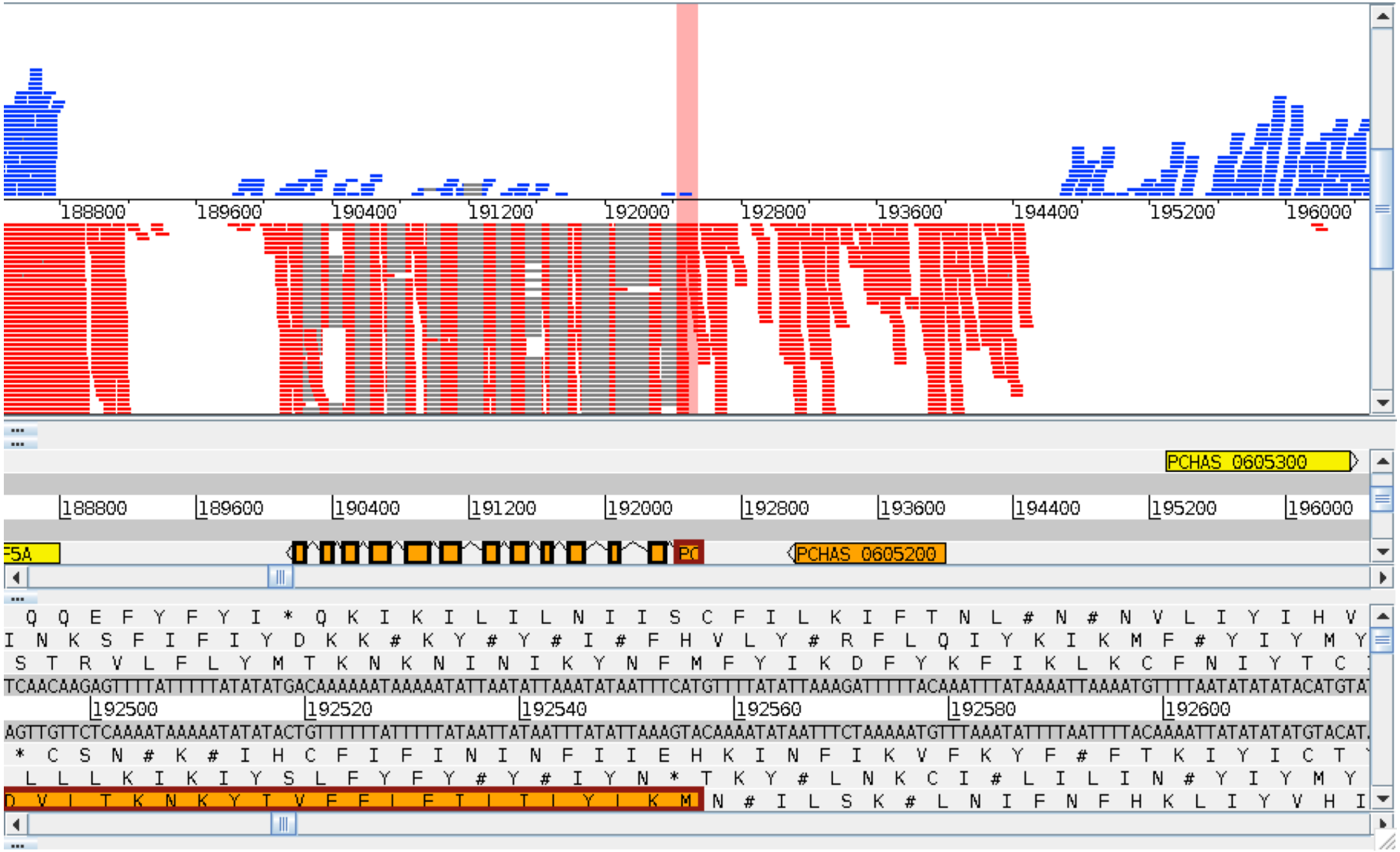
- Mapping to the genome is great for determining whether your RNA-seq data is of high quality and exploring the structure of genes of interest
- Eukaryotic genes have introns, which are not present in mature mRNA so special mapping algorithms are required
- HISAT2 is only one such algorithm, but is accurate, fast and easy to use

Artemis Genome Browser



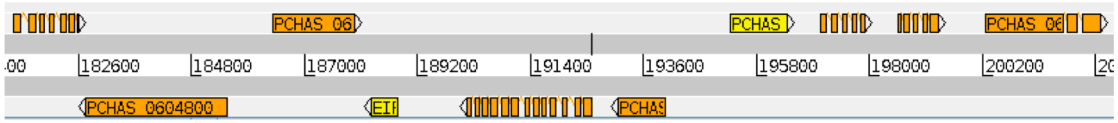

PCHAS 0605300
188800
189600
190400
191200
192000
192800
193600
194400
195200
196000
F5A
PC
PCHAS 0605200
192500
192520
192540
192560
192580
192600


2056.3
PCHAS_0605300
188800
189600
190400
191200
192000
192800
193600
194400
195200
196000
PCHAS_0605200
192500
192520
192540
192560
192580
192600

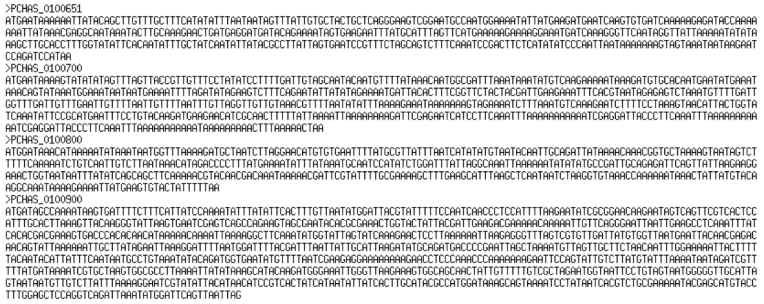

188800
189600
190400
191200
192000
192800
193600
194400
195200
196000
PCHAS_0605300
PCHAS_0605200
192500
192520
192540
192560
192580
192600

Mapping to the transcriptome and counting reads (*Kallisto*)

Genome sequence FASTA file



Transcript sequence FASTA file



- Multiple splice forms per gene introduce ambiguity into the mapping
- Mapping to the spliced transcript sequences allows this ambiguity to be taken into account and allows transcript-specific read counts
- It is also faster because there is less target sequence
- Recent improvements in algorithms (pseudoalignment) make this even faster
- Counting comes for free

Normalisation

- Read counts are biased because each sample will have a different total number of reads (solved by CPM)
- Different transcripts have different lengths, so we expect more reads from a longer transcript than a shorter one, even if the expression levels are the same (solved by RPKM/FPKM)

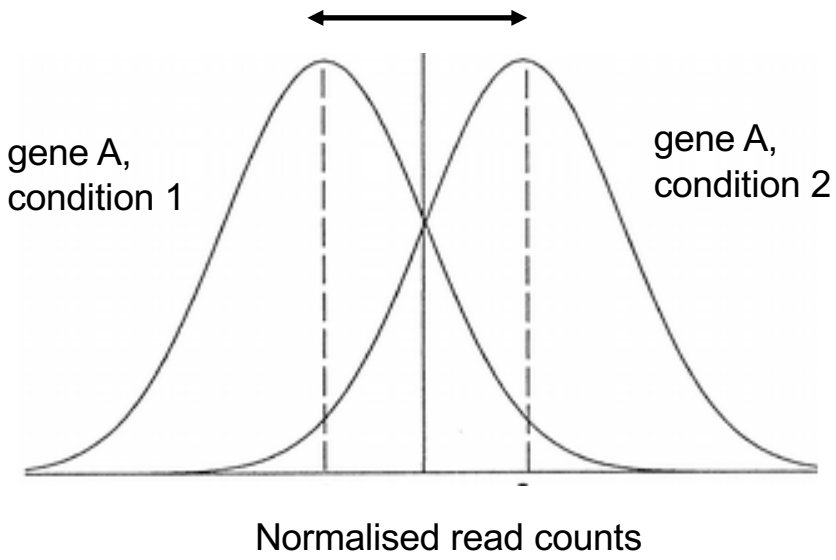
$$RPKM_i = \frac{10^9 m_i}{l_i M}$$

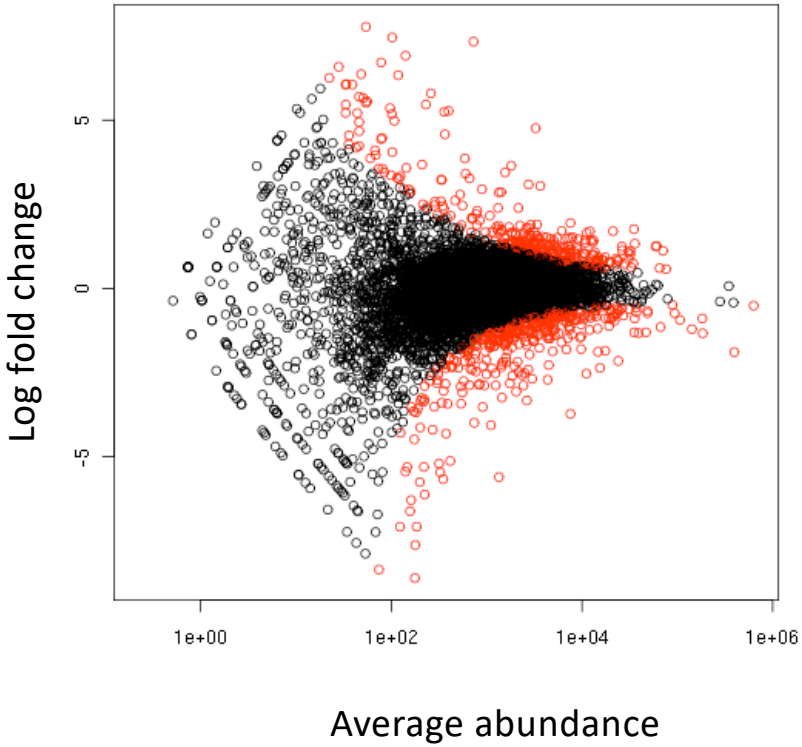
FPKM = RPKM for paired end reads

- However, RPKM has problems with highly expressed genes, so most methods use more complicated normalisation procedures (DESeq2 rlog, Sleuth)

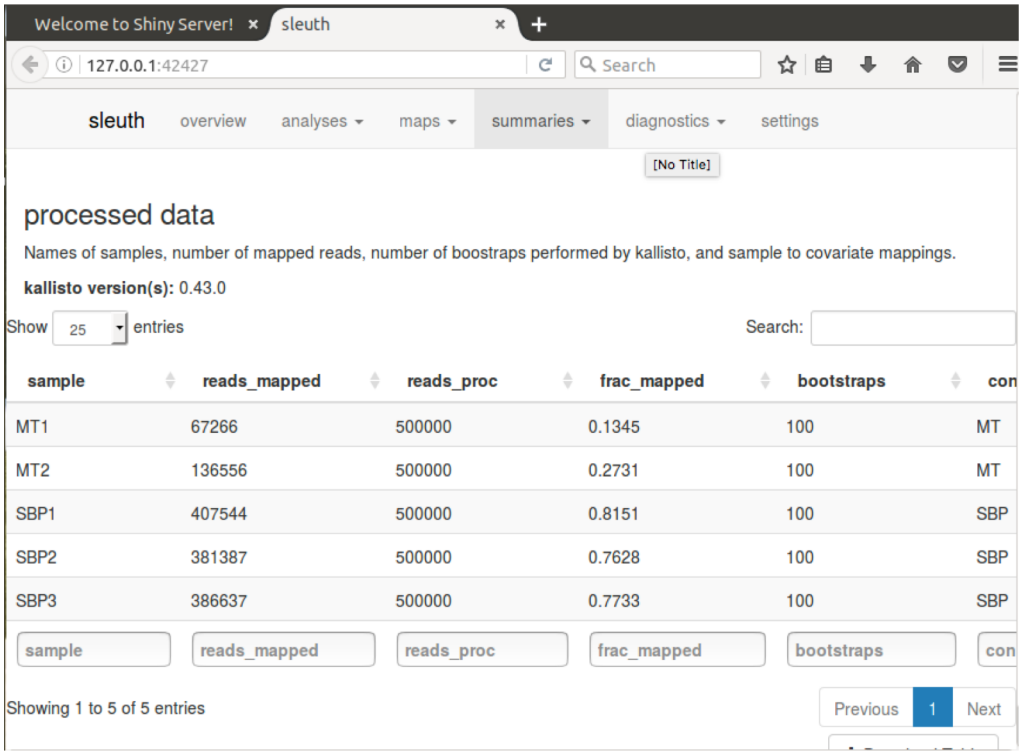
Determining differential expression (*Sleuth*)

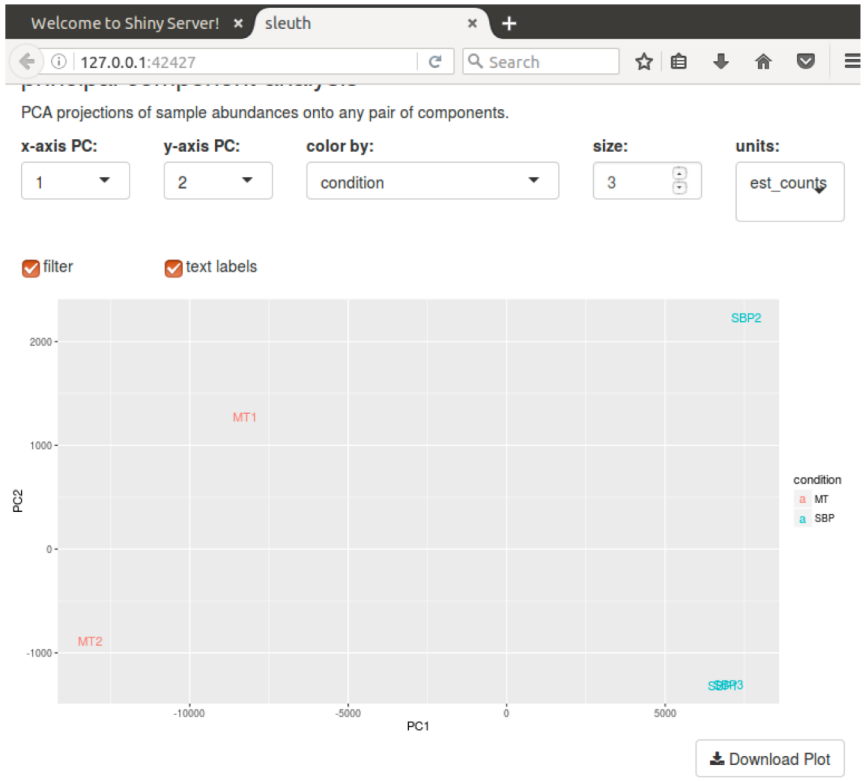
- We normally don't have enough replicates to do traditional tests of significance for RNA-seq data
- Instead most methods look for outliers in the relationship between average abundance and fold change, assuming most genes are not differentially expressed





QC with Sleuth





What to do with a gene list

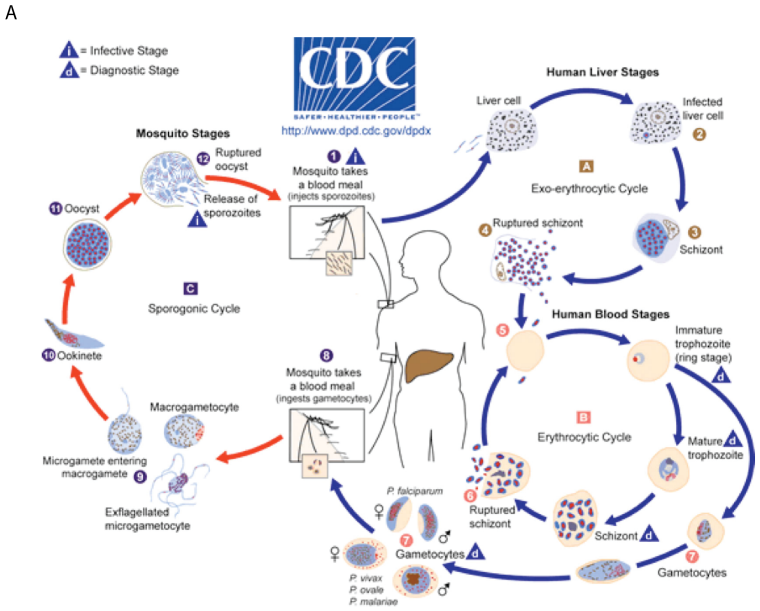
- What we have covered so far is well established methodology, which is generally applicable to most experiments
- When you have a list of differentially expressed genes, things start to get difficult. What to do:

1. Have a hypothesis already? Test it.
2. GO term/pathway analysis (GSEA, TopGO, InnateDB, Ingenuity Pathway Analysis etc.)
3. Work through list, Google, read papers
4. Overlay datasets on essentiality, populations, mutations, Pfam domains, chromosomal location, expression, proteome…

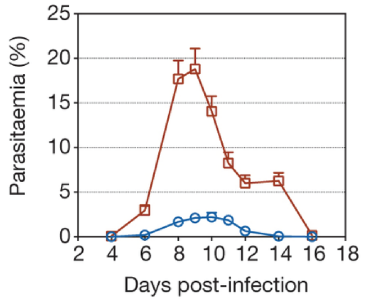
Then make a hypothesis about what genes are interesting and why. Can you test/explore this further bioinformatically? Design the next wet lab experiment.

This morning's exercise

A



B



C