

Parasite Genomics

Transcriptome analysis using RNA-seq

Adam Reid & Steve Doyle

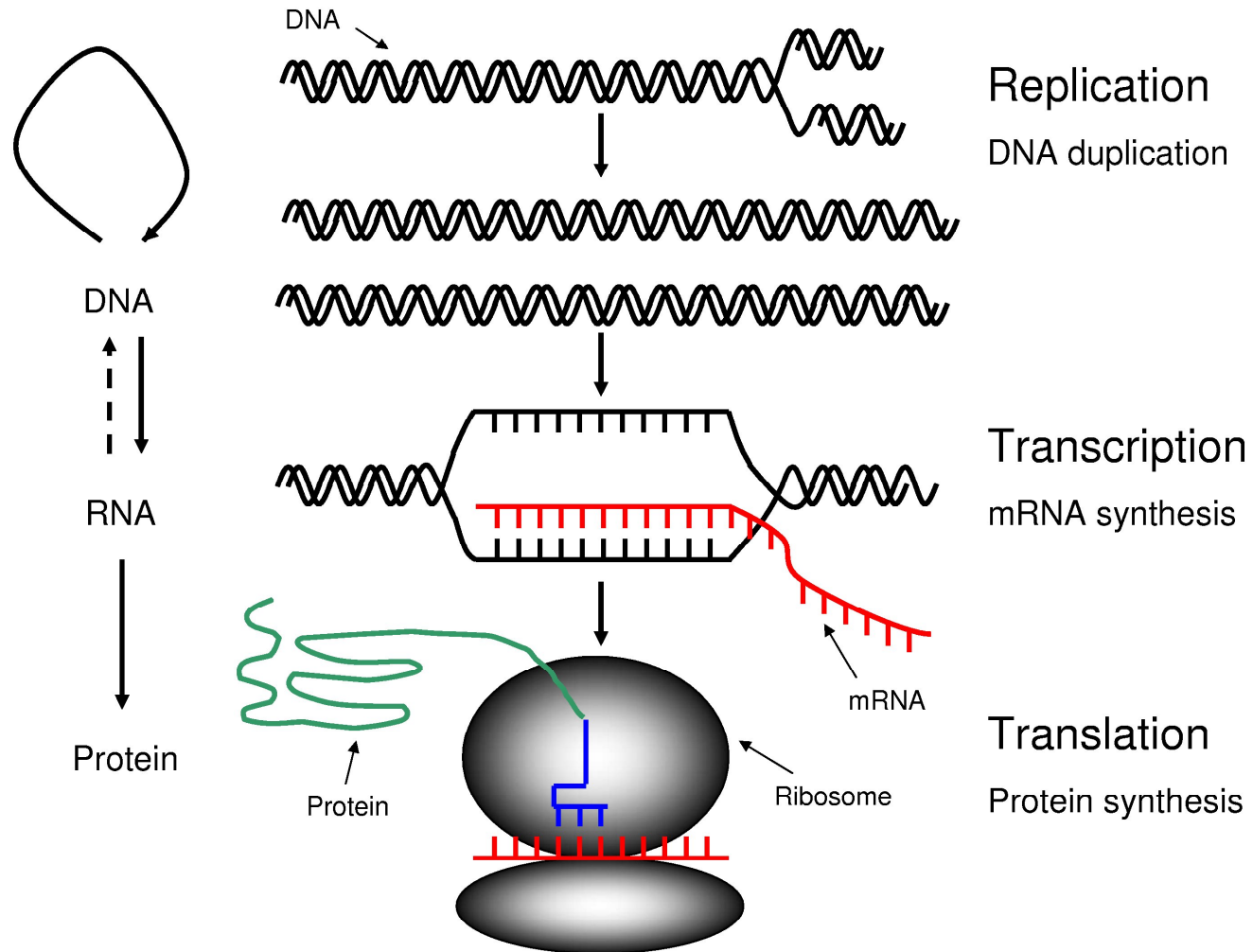
Wellcome Sanger Institute/LSHTM

LSHTM Pathogen Genomics

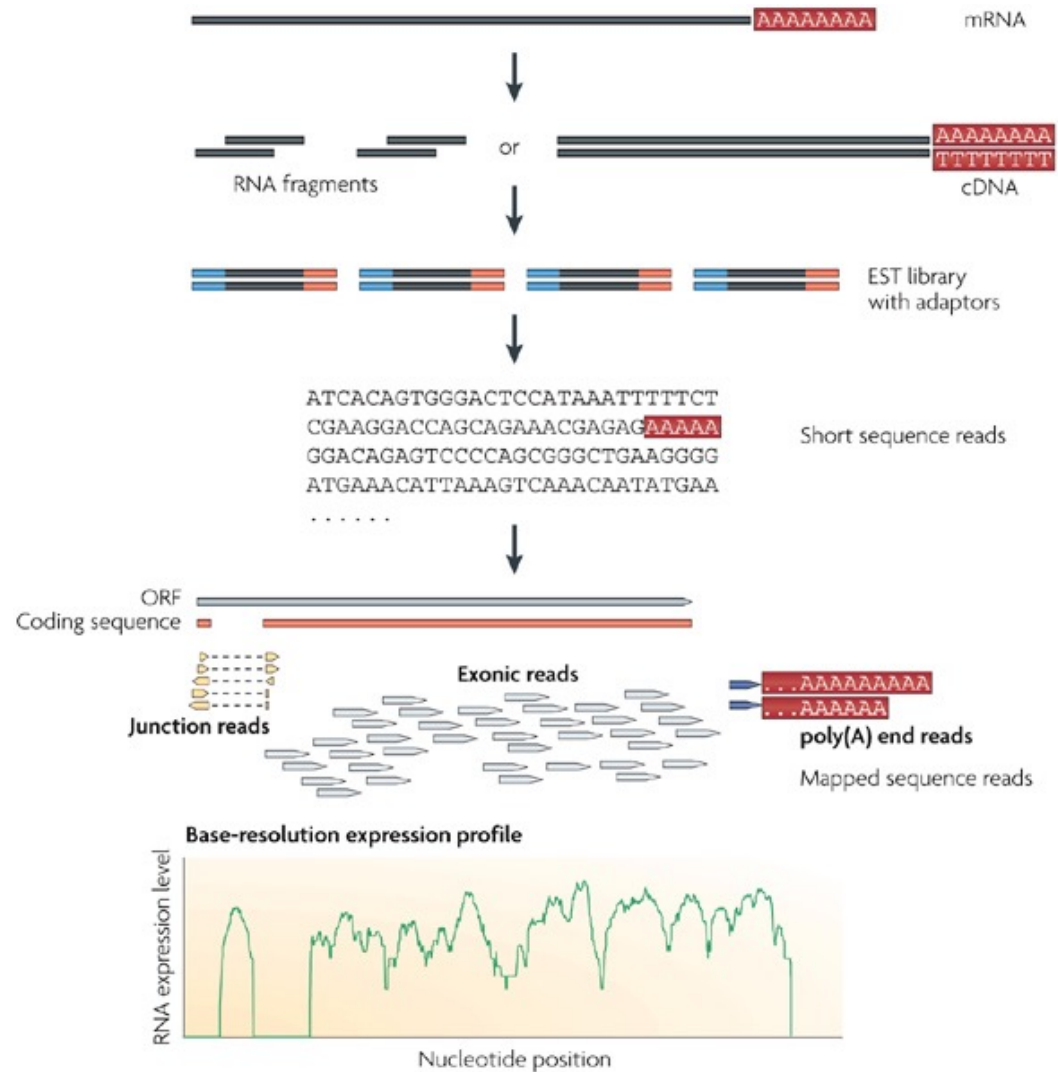
Summary

- 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



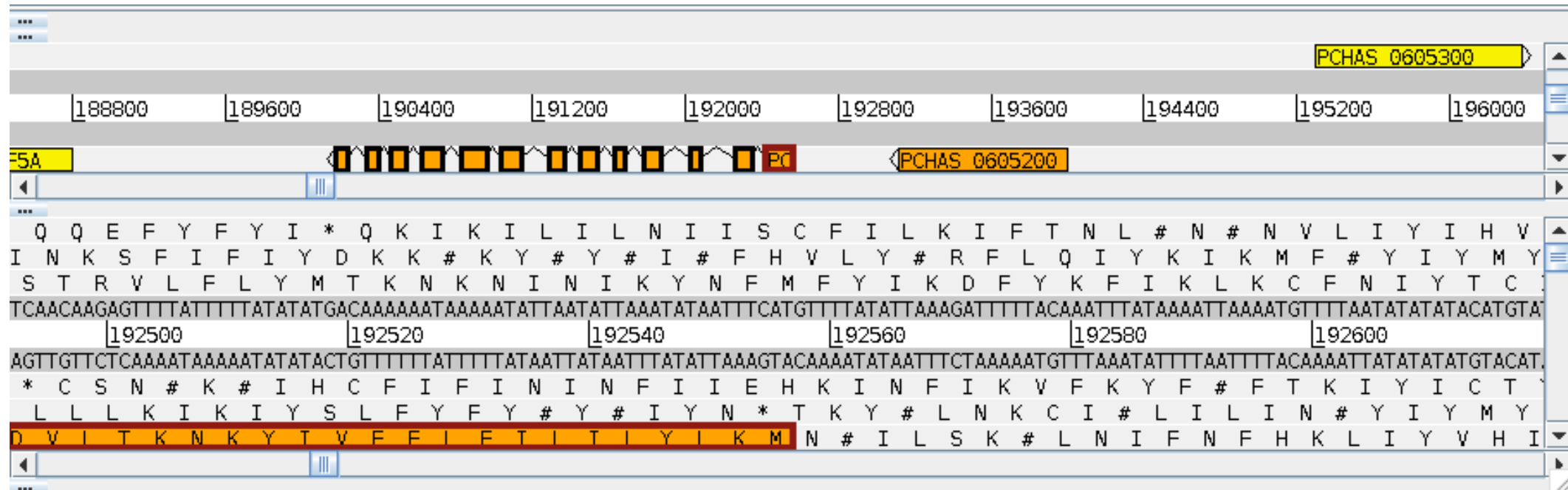
Experimental design

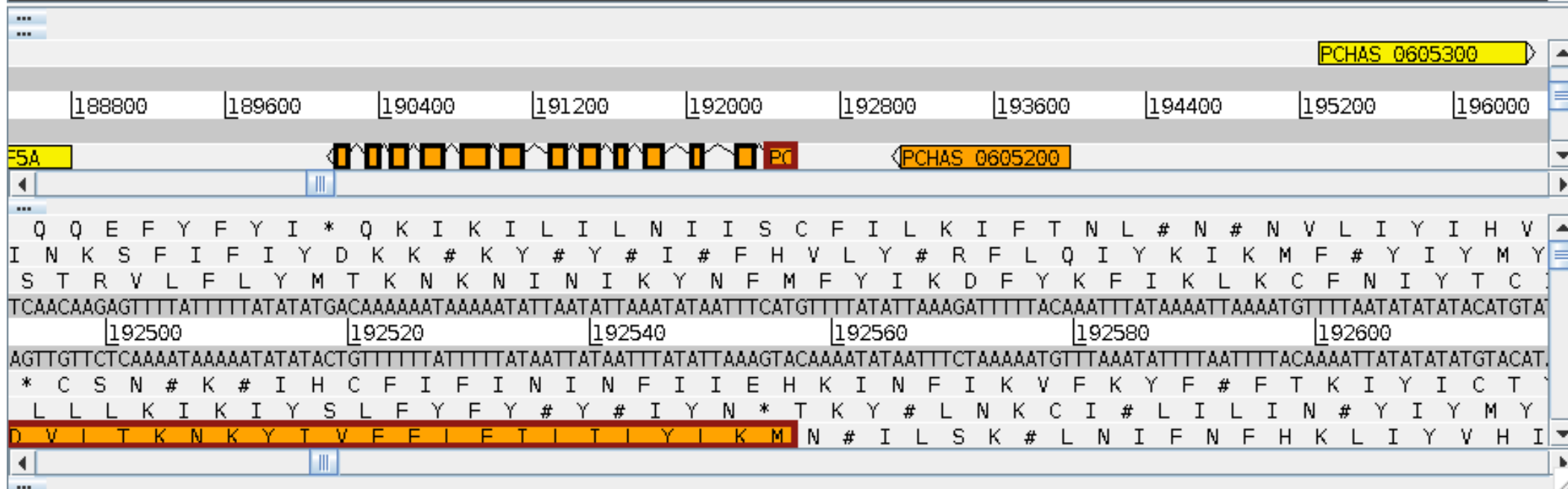
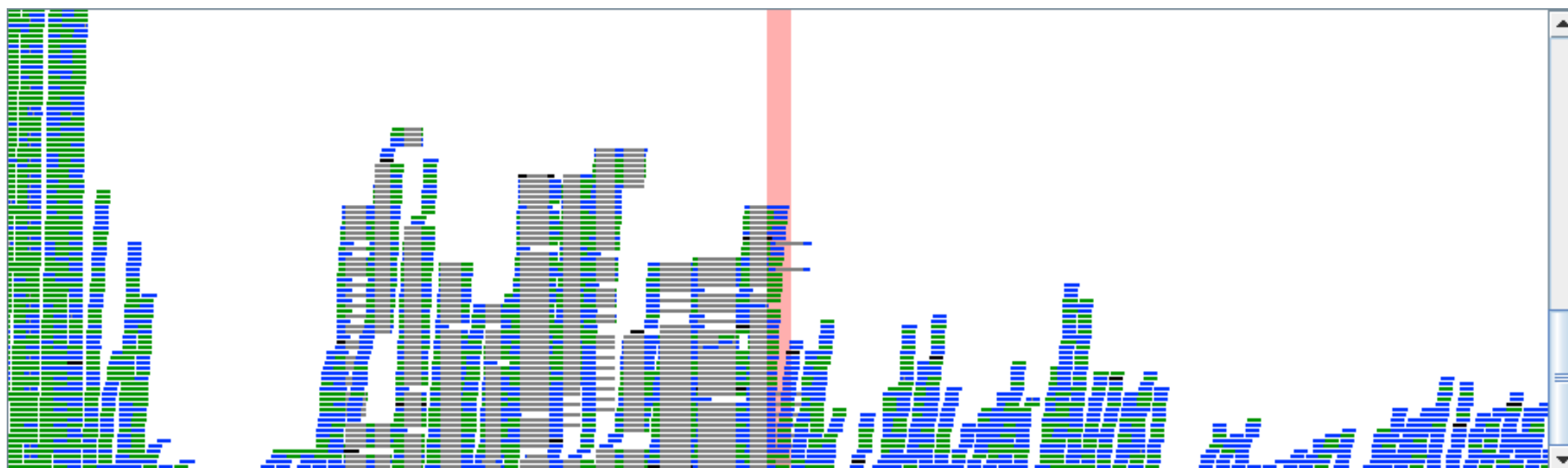
- Replicates
 - Relevant biological replicates are required
 - Technical replicates are not generally required, but try to arrange samples on plates to minimise potential problems
- Sequencing depth
 - Practical considerations e.g. amount of data on one lane, number of barcodes/tags
 - Suggested 2-5Gb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

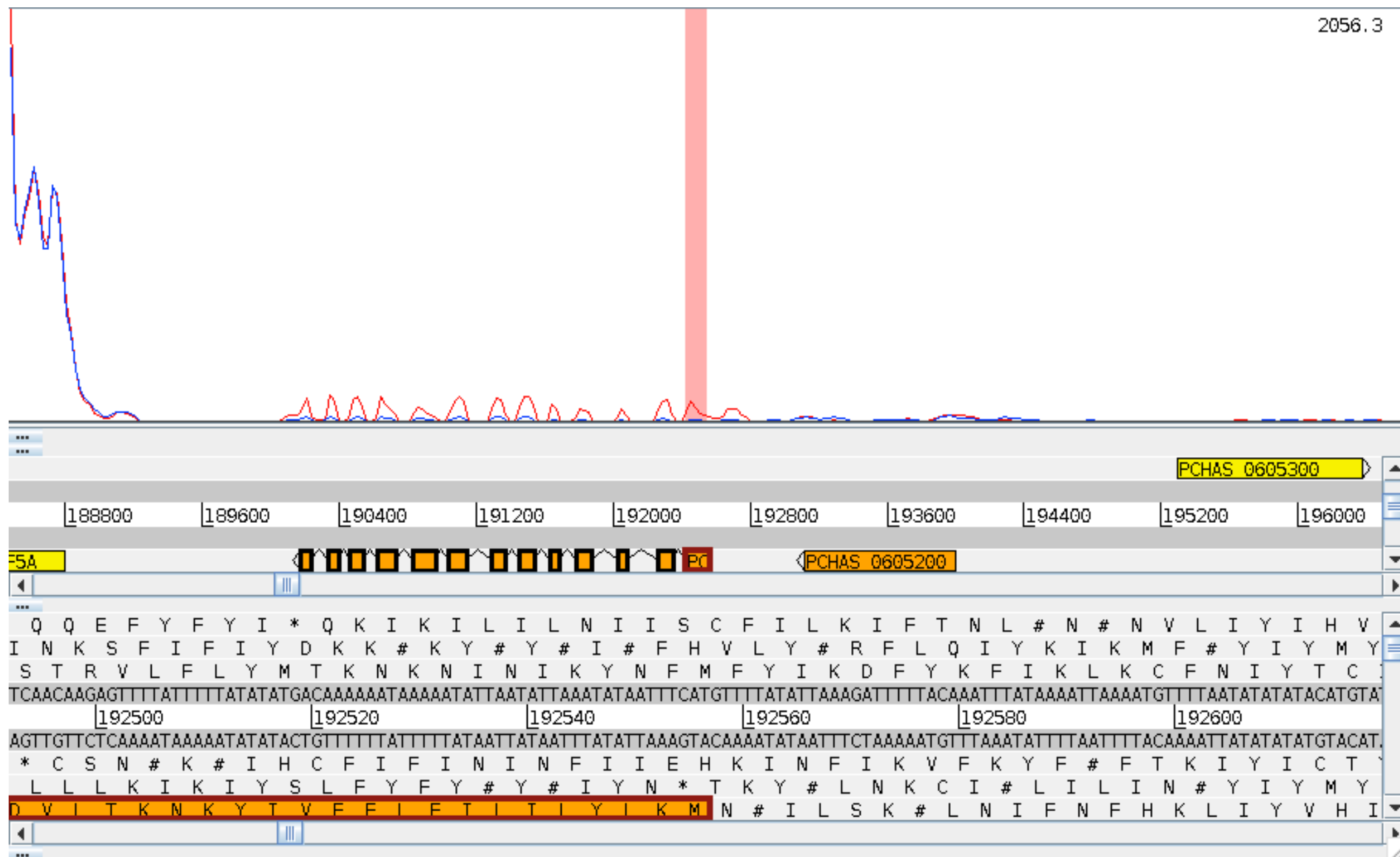
Mapping RNA-seq reads to the genome (HISAT2)

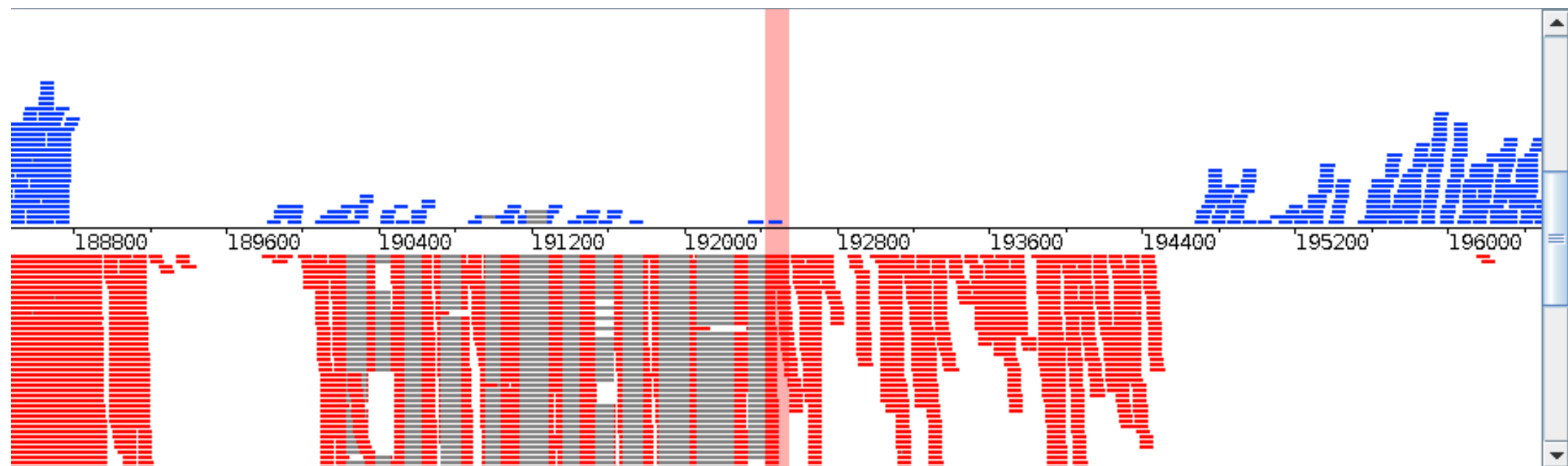
- Mapping to the genome is great for determining whether your RNA-seq data is of high quality and exploring the structure of genes of interest
- Eukaryotic genes have introns, which are not present in mature mRNA so special mapping algorithms are required
- HISAT2 is only one such algorithm, but is accurate,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

Artemis Genome Browser









...
...
PCHAS 0605300

188800 189600 190400 191200 192000 192800 193600 194400 195200 196000

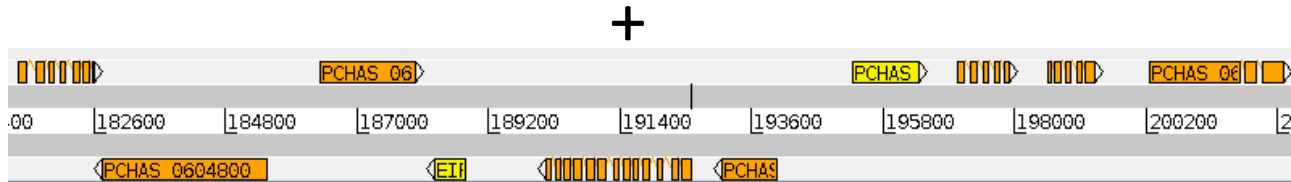
5A

PC PCHAS 0605200

Q Q E F Y F Y I * Q K I K I L I L N I I S C F I L K I F T N L # N # N V L I Y I H V
I N K S F I F I Y D K K # K Y # Y # I # F H V L Y # R F L Q I Y K I K M F # Y I Y M Y
S T R V L F L Y M T K N K N I N I K Y N F M F Y I K D F Y K F I K L K C F N I Y T C
TCAACAAGAGTTTTATTTTATATATGACAAAAAATAAAAAATATTAATATTTAAATATAATTTTCATGTTTTATATTAAGATTTTTACAAATTTATAAAATTTAAATGTTTTAATATATATACATGTA
192500 192520 192540 192560 192580 192600
AGTTGTTCTCAAAATAAAAAATATATACTGTTTTTATTTTATAATTATAATTTATATTTAAAGTACAAAATATAATTTCTAAAAATGTTTAAATATTTTAAATTTACAAAATTATATATATGTACAT
* C S N # K # I H C F I F I N I N F I I E H K I N F I K V F K Y F # F T K I Y I C T
L L L K I K I Y S L F Y F Y # Y # I Y N * T K Y # L N K C I # L I L I N # Y I Y M Y
D V I T K N K Y T V F F I F T I T I Y I K M N # I L S K # L N I F N F H K L I Y V H I

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

Genome sequence FASTA file



Transcript sequence FASTA file

```
>PCHAS_0100651
ATGATTAAGAAATATATACAGCTTGTGCTTCATATATTTAATATAGTTTATGTGCTACTGCTCAGGAGCTCGAATGCCAATGGAATATTTATGAGATGATCAGTGTGATCAAAAGAGATACCAAAA
AAATTTAAACAGCGCGCATTAATCTTCAGAGAGCTGATGAGATGATACAGAAATAGTGAAGATTTATGCTATTAGTTCATGAAAAAGAAAGAAATGATCAAGGCTTCATAGGTTTAAATATATAT
AAAGCTTGGCCCTTTGGTATATTCACATATTTGCTATCATATTTATACGCTTATAGTGAATCCGTTTCTAGCAGCTTTCATATCCGATTTAATAAAAAGTAAATATATAGAT
CCAGATCCATAA
>PCHAS_0100700
ATGAAATAAAGTATATATAGCTTTAGTTACCGTTGTTTCTATATCTTTTGTATGCTAGCAATCAATGTTTATAAACAATGCCGATTTAATAAATATGTCAGAAAAATAAAGATGTCACATCATATGAAT
AAACGATATAAATGAATATATGAATTTTAGATATAGAGCTCTTCAGATATTATATAGAAATGATTACACTTTCGCTTCTACTACATTGAAGAAATTTACGCTATAGAGAGCTTAAGCTTTTGTATT
GGTTTGATTGTTGATGTTGTTTAAATGTTTAAATTTGTTAGGTTGTTGTAACGCTTTTAAATATTTAAAGAAATAAAAAGTAGAAATCTTAAATGTCAGAAATCTTTCCTAAGTACATCTACTGTA
TCAATATTCGCGTGAATTTCTGCTACAGATGAAGAGCTCCGACTTTTATTAATTAATAAAAAAGATTGAGATCATCTCTCAATTTAAAAAAAATCGAGGATTACCTTCATTTAAAAAAA
AATCGAGATTTACCTTCATTTAAAAAAAATAAAAAAACTTTAAAACTAA
>PCHAS_0100800
ATGATAAACAATAAATATAAATAGGTTTAAAGATGCTAATCTTAGCAATGCTGATTTTATGCGTTTATTAATCATATATGTAATACATTTGCAATTAATAAACAACCGCTGCTAAAGTATAGTCT
TTTTCAAAATCTGCTATGCTTAAATACATAGACCCCTTTATGAATATTTAATATCAATCCATATCTGGATTTATAGGCAATTAATAAATATATATGCGGATTGCAAGATTCAGTTATTAGAGG
AACTGGTAATATTTATATCAGCAGCTTCAGAAAGCTACACCAATTAACAGCATTCGATTTTGGCAAGGCTTTGAGCATTATAGCTCAATATCTAAGGCTAAACCAAAATTAATATTATGTA
AGCAATTAAGAAATTTATGAAGTACTATTTTAA
>PCHAS_0100900
ATGATAGCAAAATAGTATTTCTTATATCAAAATATTTATATCACTTTTATATGAGATTAGTATTTTCCATCAACCCCTCAATTTAAGATATCCGGAACAGAACTAGTCTGCTACTCC
ATTTGCTATTAAGTTTACAGCTATTAGTGAATGACTGACGAGAGTACCAATACGCAAGCTGCTACTATTCCGATTAAGCAAGCAAAACAAATTTGTCAGGATTAATTAAGCTCAATTTAT
CAACGCAAGAGTACCCCAACACATAAACAAATTAAGAGGCTTCAATATGCTATTAGTATAGAGACTCTTAATAAATTAAGAGGCTTAGTCTGCTGATTTATGCTGATTAATGCAATTAACACAGAC
AACGATTAATAAATGCTTATAGATTAAGGATTTTATGATTTTACGATTTAATGCTATTAGATATGAGATGCAAGTACCCCAATTAAGTATGCTTAGTCTTACATTTGCAAAATTAATCTTT
TACATATATTTCAATATGCTGTAATATACAGATGCTGATGTTTTATGAGAGCAAAAGAGCACTCCCAACCCCAAAAGAGATCCGATTTGCTTTATGATTTAATATAGATGCTTT
TTATGATAAATCTGCTAGTGGCCCTTAAATTTATTAAGCATACAGATGGGAATGGTTAAGAGAGTGGCAGCACTATGTTTTGCTGCTAGATGTTATCTCTAGTATGAGGAGGCTGATTA
GTATATGTTGCTTTTAAAGGATCGTATATACATACATCCGCTATCATATATATCACTTCTACATCCGATGATTAAGCACTAATCTTATACCTGCTGCAAAATACGAGATGATCC
TTGAGGCTCCGCTGATTAATATGATTTAGT
```

- 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)

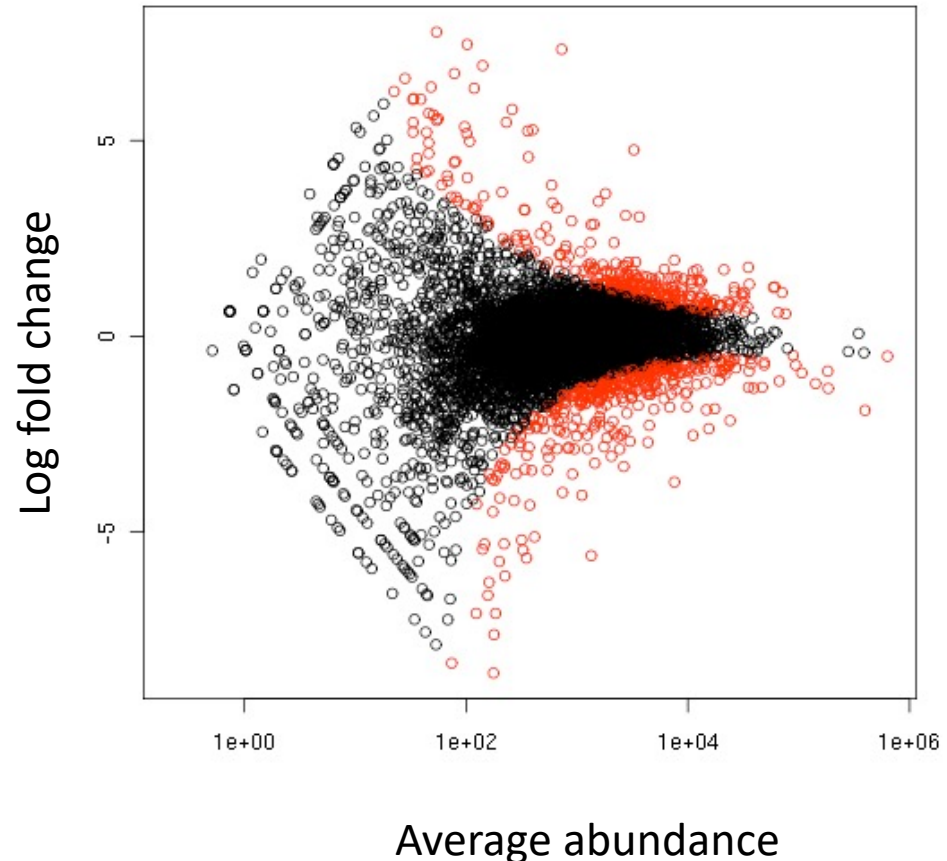
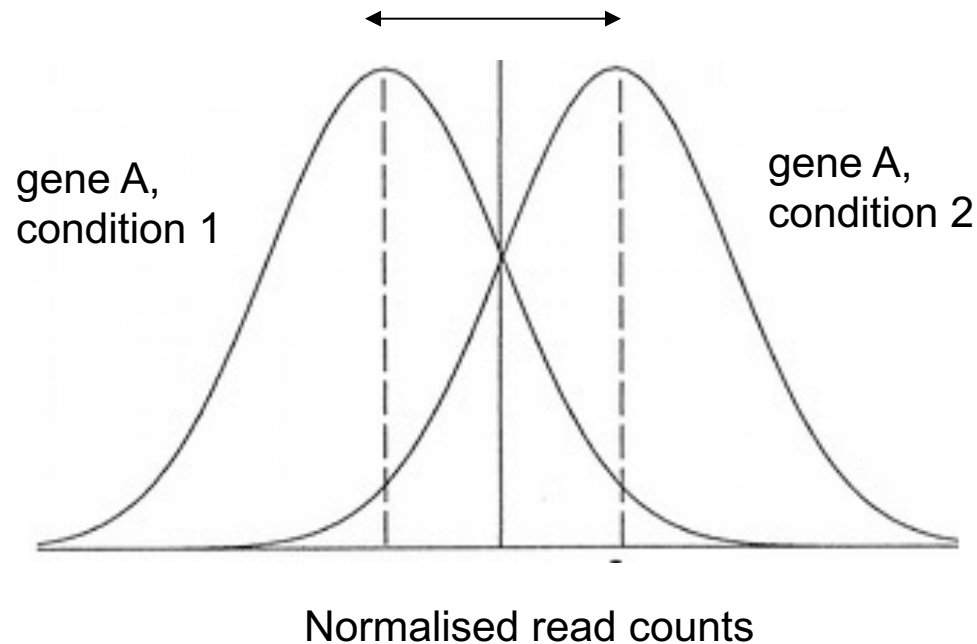
$$RPKM_i = \frac{10^9 m_i}{l_i M}$$

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

Welcome to Shiny Server! x sleuth x +

127.0.0.1:42427 Search

sleuth overview analyses maps summaries diagnostics settings

[No Title]

processed data

Names of samples, number of mapped reads, number of bootstraps performed by kallisto, and sample to covariate mappings.

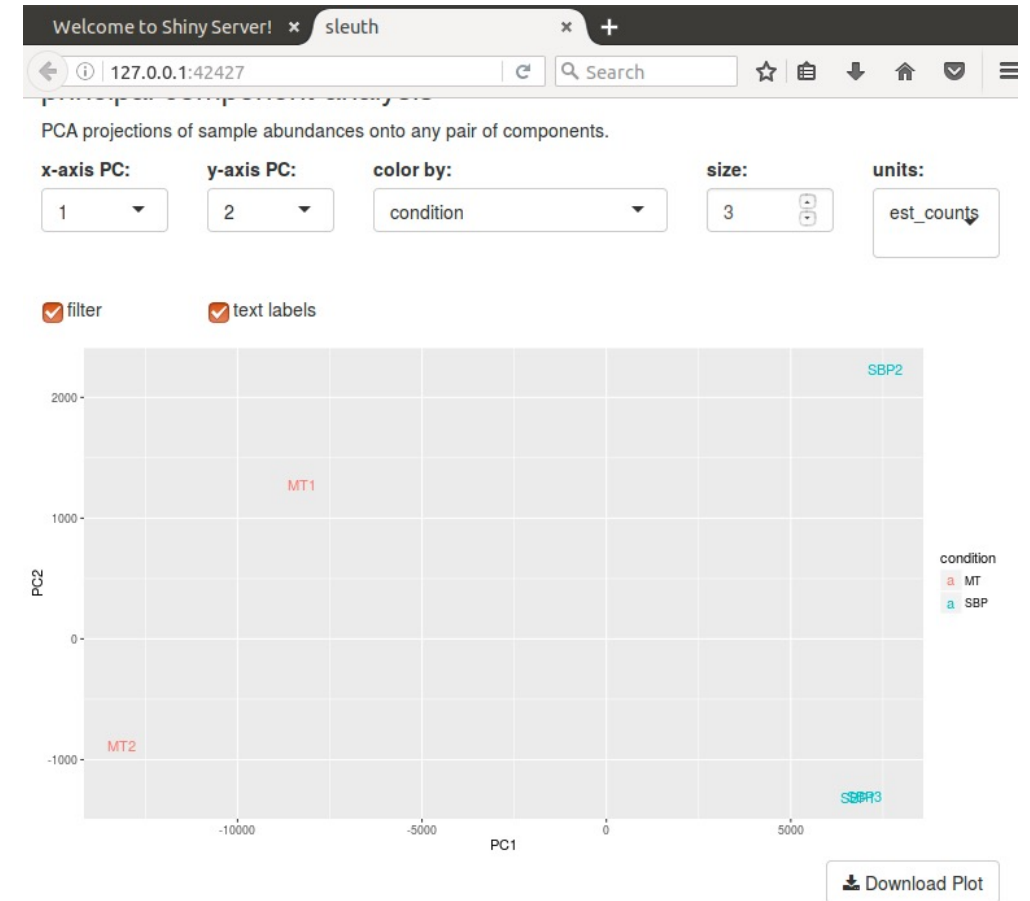
kallisto version(s): 0.43.0

Show 25 entries Search:

sample	reads_mapped	reads_proc	frac_mapped	bootstraps	condition
MT1	67266	500000	0.1345	100	MT
MT2	136556	500000	0.2731	100	MT
SBP1	407544	500000	0.8151	100	SBP
SBP2	381387	500000	0.7628	100	SBP
SBP3	386637	500000	0.7733	100	SBP

sample reads_mapped reads_proc frac_mapped bootstraps condition

Showing 1 to 5 of 5 entries Previous 1 Next



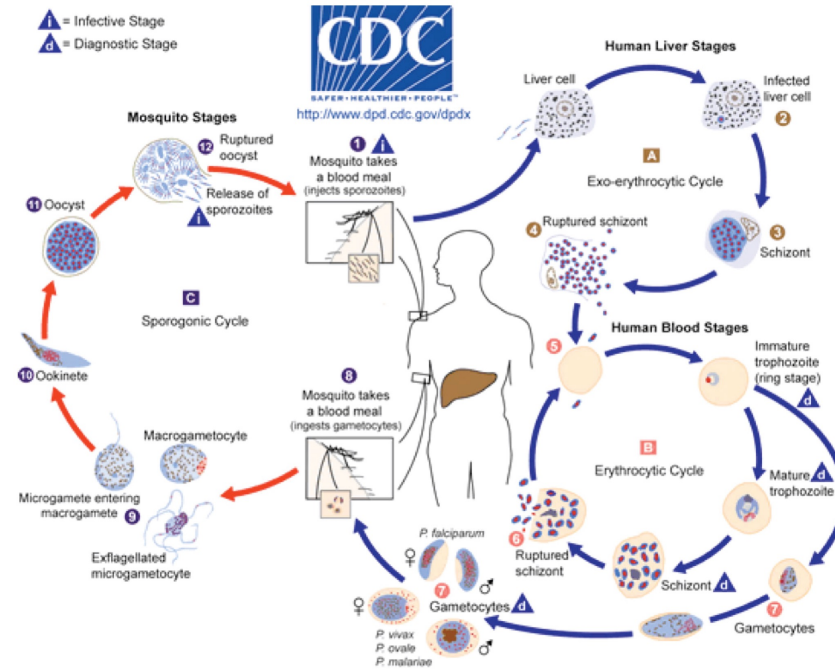
What to do with a gene list

- 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...

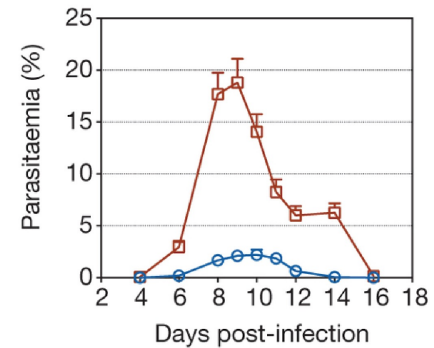
Then make a hypothesis about what genes are interesting and why. Can you test/explore this further bioinformatically? Design the next wet lab experiment.

Today's exercise

A



B



C

