Differential Expression using RNA-Seq

Victoria Offord

WTC NGS Bioinformatics

Hinxton 2019

Overview

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

What is the transcriptome?

"The complete set of transcripts in a cell

and their quantity

for a specific developmental stage or condition"

Wang *et al*. (2009)
Nature Reviews Genetics
(PubMed: 19015660)

Central dogma

DNA

transcription

Pre-mRNA

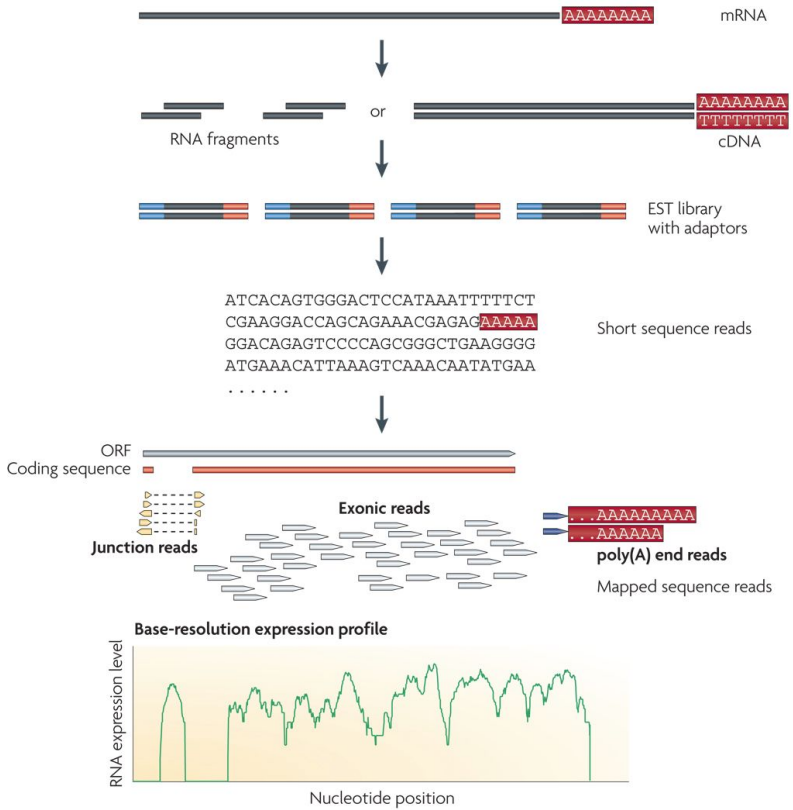
splicing
capping
tailing

mRNA

translation

Protein

RNA Sequencing



Wang *et al*. (2009)
Nature Reviews Genetics
(PubMed: 19015660)

Experimental design

- Successful RNA-Seq studies start with a good study design
- Considerations for generating data to answer your biological question include:
  - library type
  - sequencing depth
  - number of replicates
  - avoiding biases

Experimental design - library preparation

- Total RNA = mRNA + rRNA + tRNA + regulatory RNAs...
- Ribosomal RNA can represent > 90% total RNA
- Can enrich for the 1-2% mRNA or deplete rRNA
    - enrichment typically needs good RIN and high RNA proportion
    - some samples (e.g. tissue biopsies) may not be suitable
    - bacterial mRNA not polyadenylated -> ribosomal depletion
- Be aware of protocol being used (e.g. some will remove small RNAs)

Experimental design - library type

- **Stranded vs unstranded**
  - strand-specific protocols better for detangling antisense or overlapping transcripts
- **Single or paired end**
  - paired end better for *de novo* transcript discovery or isoform expression analysis
  - < 55% reads will span 2 or more exons

Experimental design - replicates

Biological replicates

- biologically distinct samples
- same type of organism treated or grown in the same condition
- understand biological variation (e.g. variation between individuals)
- relevant biological replicates are required

Technical replicates

- repeated measurements of the same sample
- understand the variation in equipment or protocols
- technical replicates are not generally required, but try to arrange samples on plates to minimise potential problems

Experimental design - sequencing depth / replicates

BIOINFORMATICS **DISCOVERY NOTE** Vol. 30 no. 3 2014, pages 301–304
doi:10.1093/bioinformatics/btt688

Gene expression

Advance Access publication December 6, 2013

RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

[1]Institute of Genomics and Systems Biology, [2]Committee on Development, Regeneration, and Stem Cell Biology and [3]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

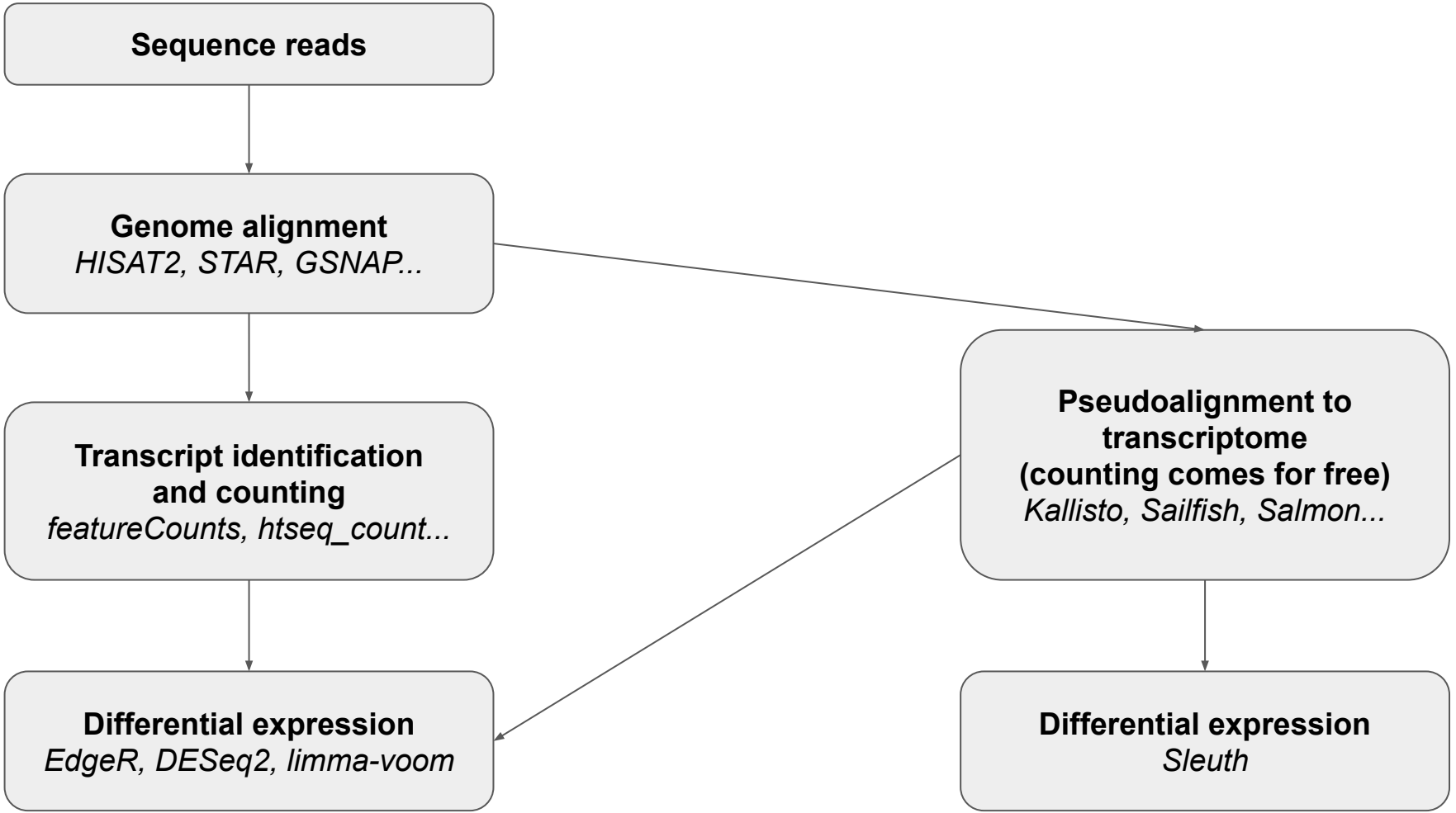
Associate Editor: Janet Kelso

- reduces the coefficient of variation

What do we need to know?

1. Which genes/transcripts do our reads belong to? **mapping / assembly**
2. How many reads align to a specific gene/transcript? **quantification**
3. Do different sample groups express genes/transcripts differently? **DGE analysis**

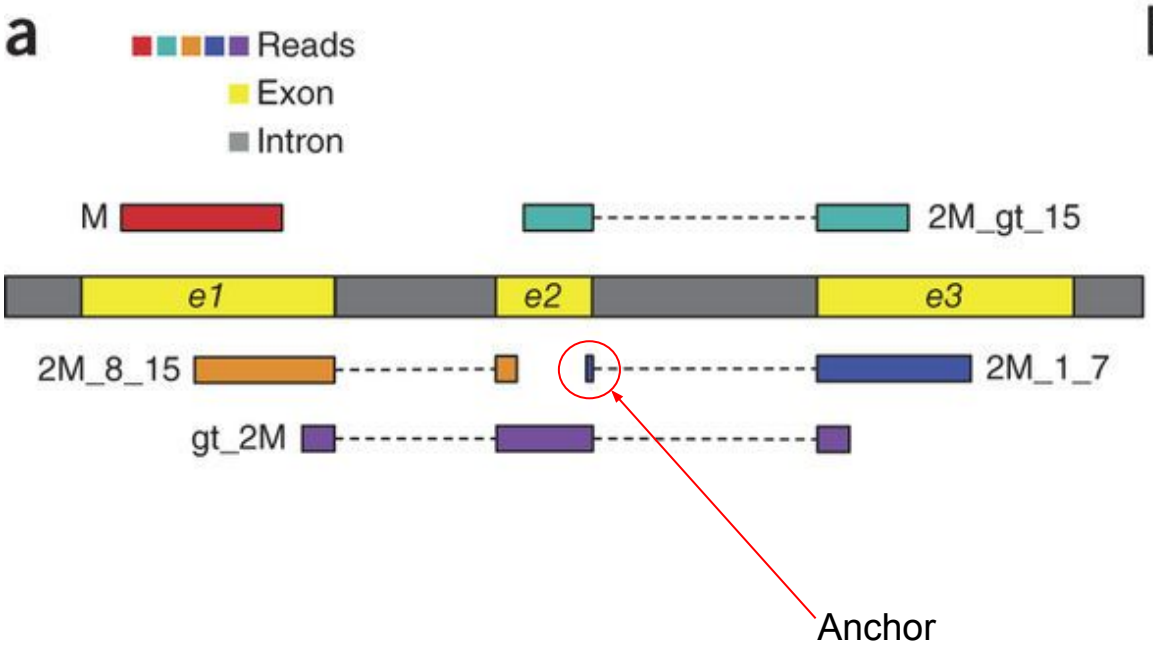
No universal pipeline to cover every analysis!!!

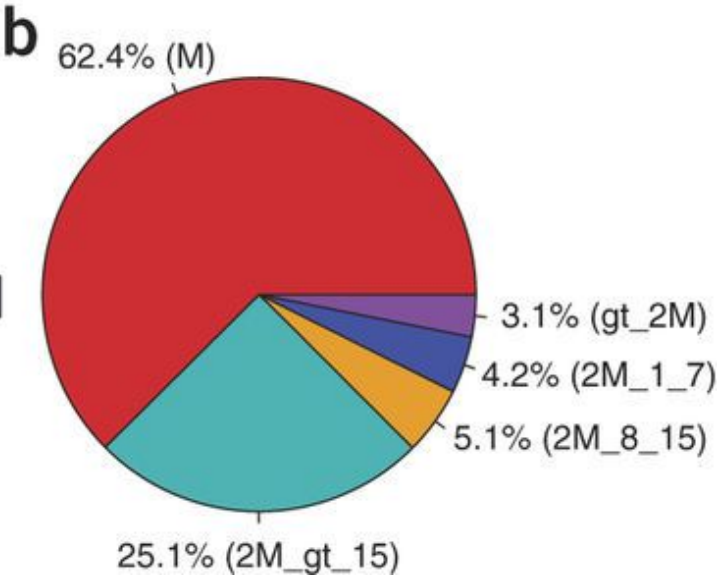

Sequence reads
Genome alignment
HISAT2, STAR, GSNAP...
Transcript identification and counting
featureCounts, htseq_count...
Differential expression
EdgeR, DESeq2, limma-voom
Pseudoalignment to transcriptome (counting comes for free)
Kallisto, Sailfish, Salmon...
Differential expression
Sleuth

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 (splice-aware)
- **HISAT2** is only one such algorithm, but is accurate, fast and easy to use

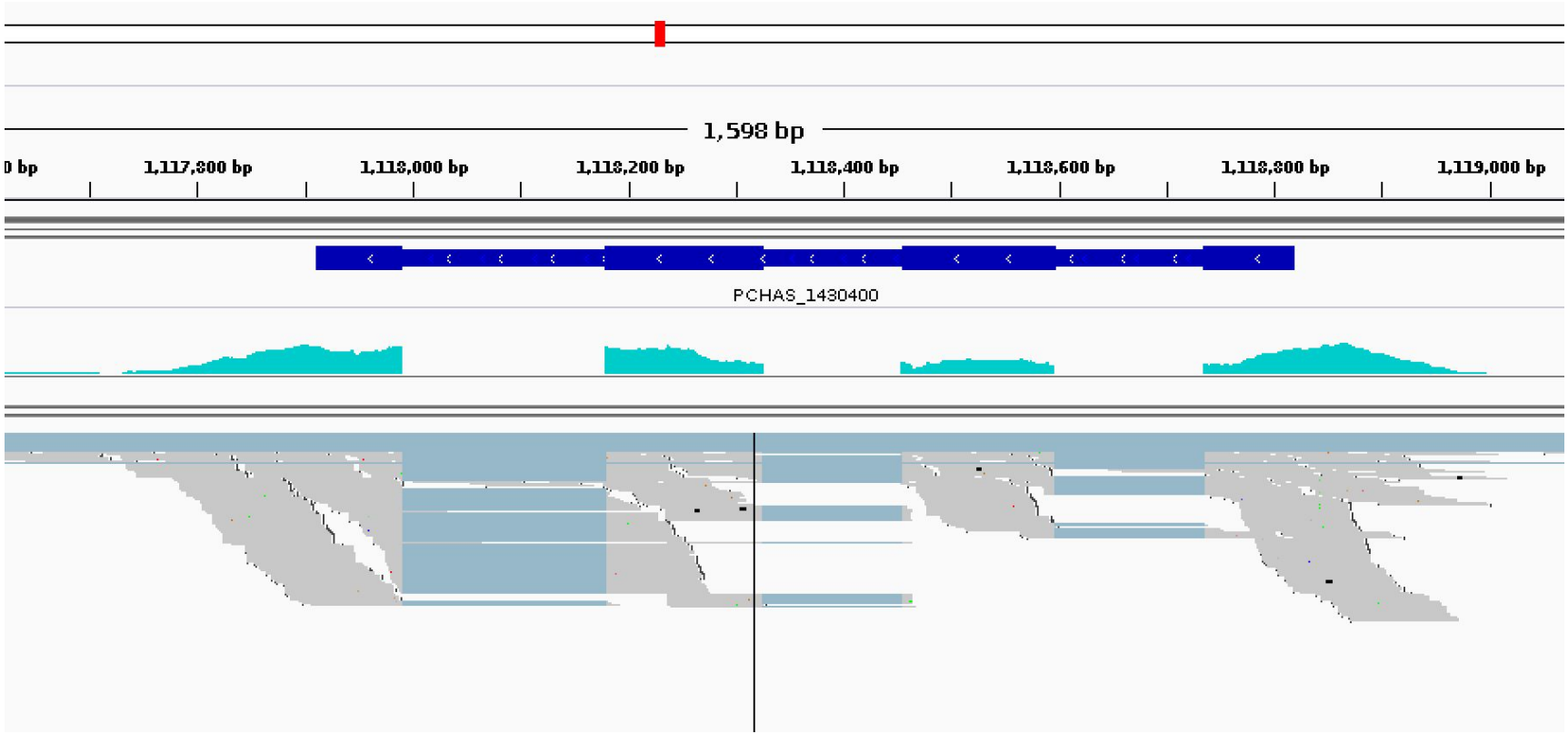
Splice aware alignment





Kim *et al*. (2015)
Nature Methods
(PubMed: 25751142)

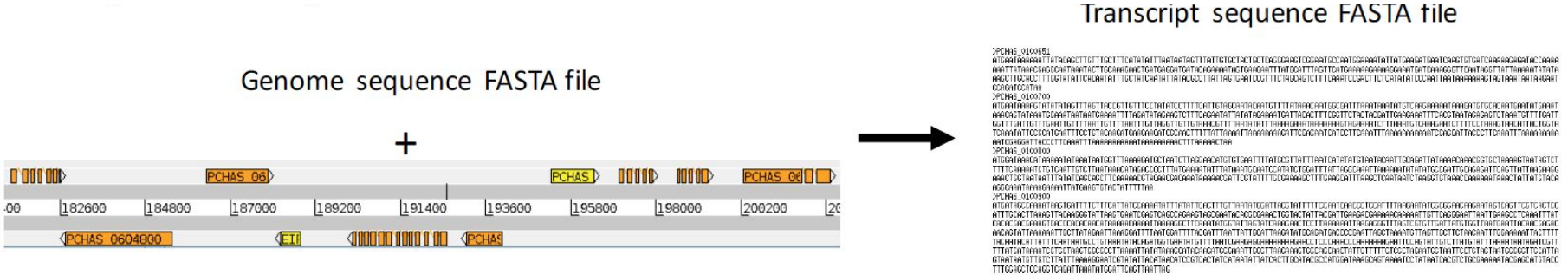
Visualisation: Integrative Genomics Viewer (IGV)



Mapping to the transcriptome and counting reads (Kallisto)

- 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
    - doesn’t care where in each transcript reads map to, just which of the transcripts they map to
- Counting comes for free

Mapping to the transcriptome and counting reads (Kallisto)

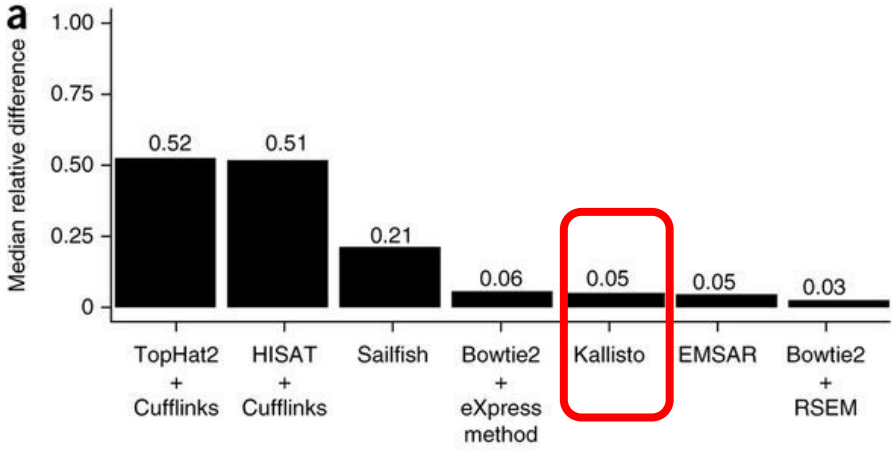


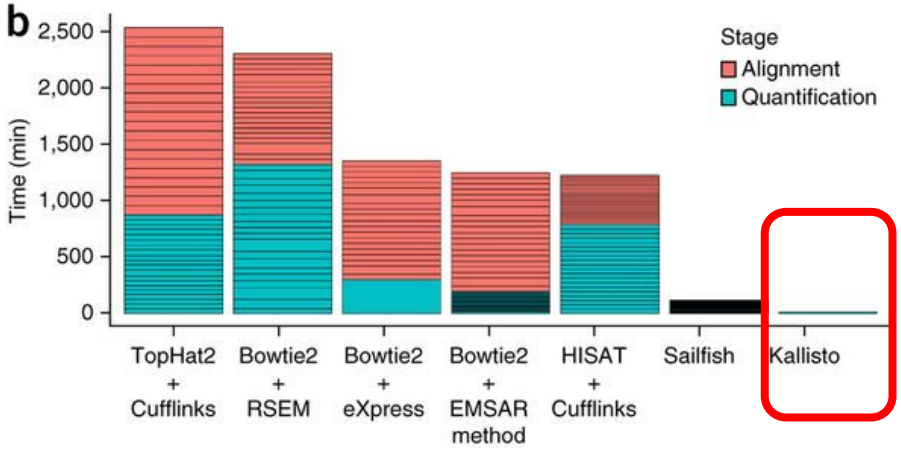
Kallisto has two steps:

1. Building an index from the spliced transcript sequences
2. Quantify reads against the index

Kallisto cannot be used to identify novel transcripts

Mapping to the transcriptome and counting reads (Kallisto)





Bray *et al*. (2016)
Nature Biotechnology
(PubMed: 27043002)

Normalisation

- Runs with more depth will have more reads mapping to each gene (**sequencing depth bias**)
- Longer genes will have more reads mapping to them (**gene length bias**)
- Most methods will normalise for sequencing depth and gene length

Normalisation methods

- **RPKM** - reads per kilobase per million
- **FPKM** - fragments per kilobase per million
- **TPM** - transcripts per million

Some of these methods have problems with highly expressed genes, so it's better to use more complicated normalisation procedures (**DESeq2 rlog**, **Sleuth**)

RPKM

BEFORE

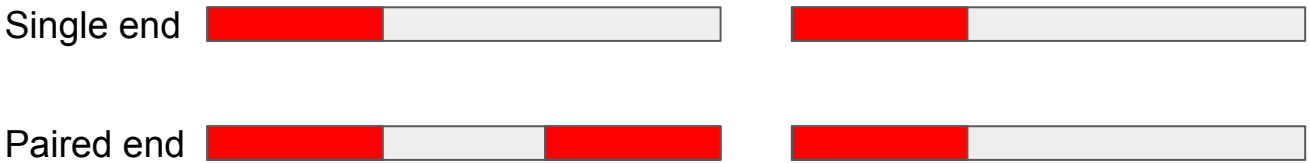
Gene	Replicate 1 Counts	Replicate 2 Counts	Replicate 3 Counts
A (2,000 bases)	10	12	30
B (4,000 bases)	20	25	60
C (1,000 bases)	5	8	15
D (10,000 bases)	0	0	1

AFTER

Gene (bases)	Replicate 1 RPKM	Replicate 2 RPKM	Replicate 3 RPKM
A (2,000 bases)	1.43	1.33	1.42
B (4,000 bases)	1.43	1.39	1.42
C (1,000 bases)	1.43	1.78	1.42
D (10,000 bases)	0	0	0.009

FPKM (fragments per kilobase million)

- RPKM for paired reads
- takes into account that two reads can map to one fragment (and so it doesn’t count this fragment twice)



RPKM vs TPM

RPKM

Gene	R1	R2	R3
A	1.43	1.33	1.42
B	1.43	1.39	1.42
C	1.43	1.78	1.42
D	0	0	0.009
Total	**4.29**	**4.5**	**4.25**

TPM

Gene	R1	R2	R3
A	3.33	2.96	3.326
B	3.33	3.09	3.326
C	3.33	3.95	3.326
D	0	0	0.02
Total	**10**	**10**	**10**

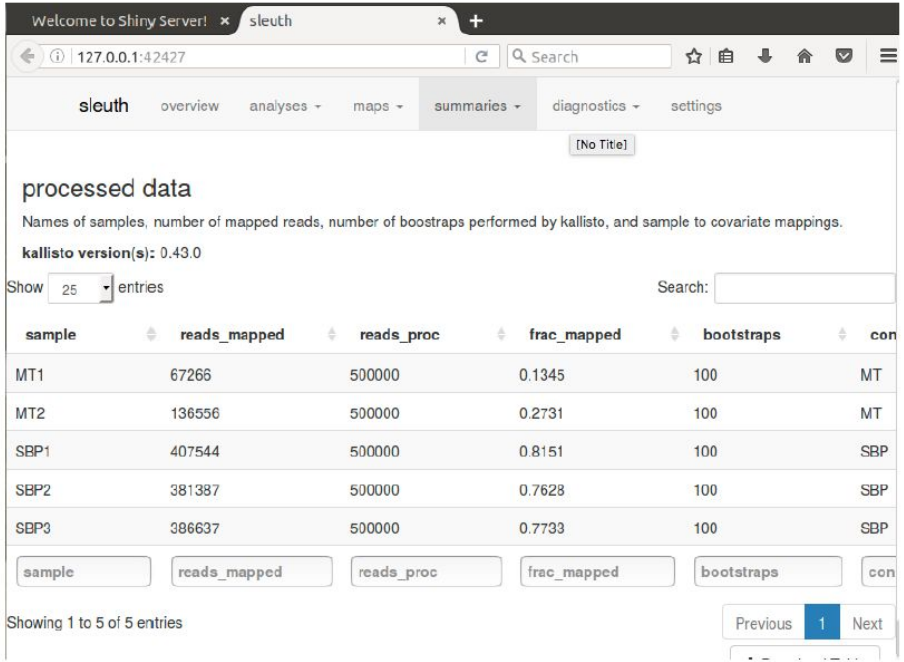
Easier to see the proportion of each gene within a sample as sum of TPMs same across samples

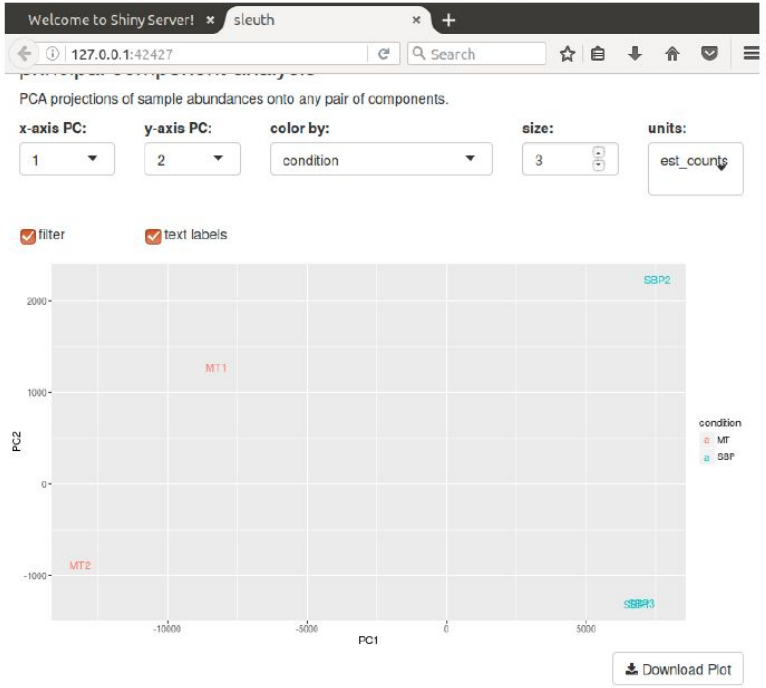
Adapted from StatQuest (http://statquest.org)

Determining differential expression (Sleuth)

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

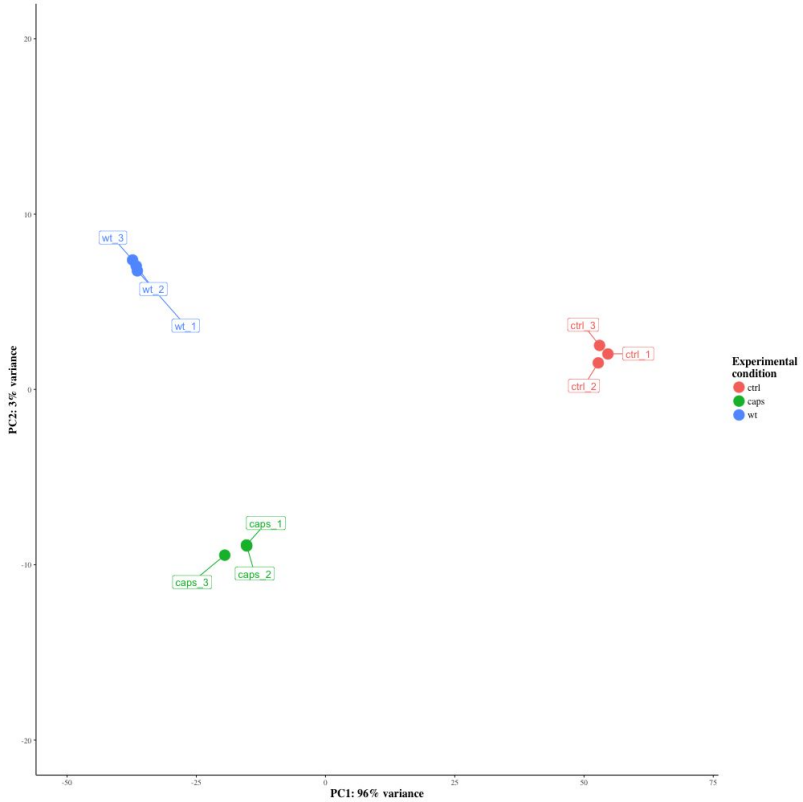
QC with Sleuth



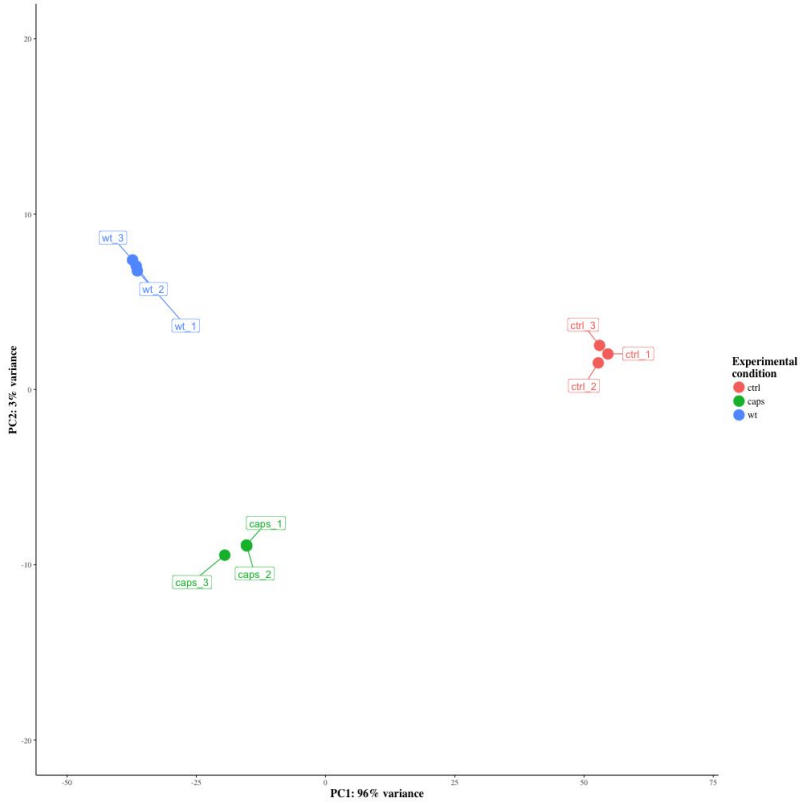


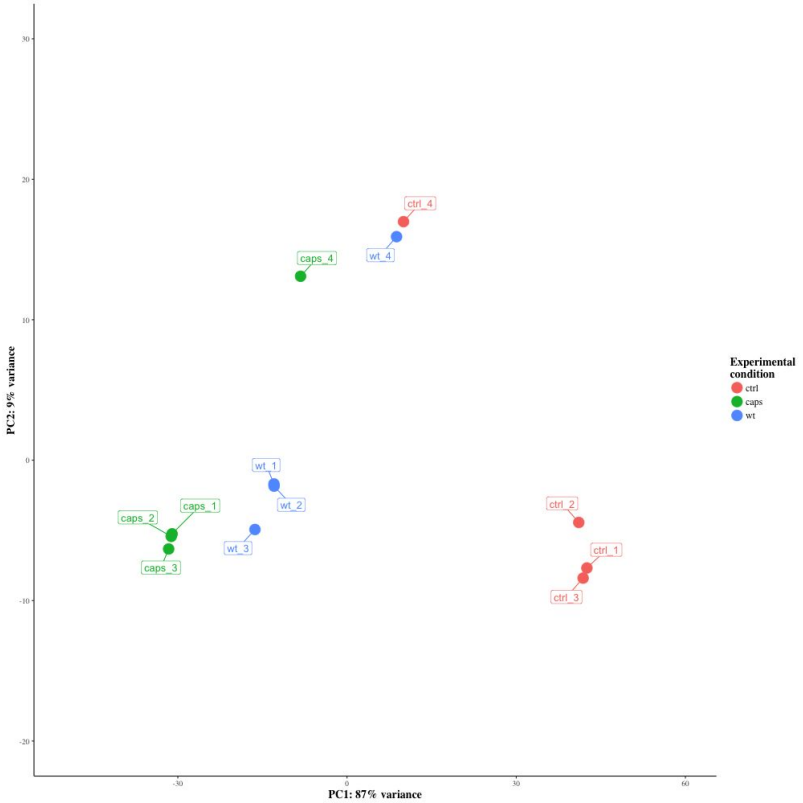
Principal component analysis (PCA)

- Use to look at variation and strong patterns within data
- Identifies uncorrelated variables or principal components (PC)
- Tries to explain the maximum amount of variance with the smallest number of principal components



Why QC our data?





What to do next with your gene list

When you have a list of differentially expressed genes, things start to get difficult.

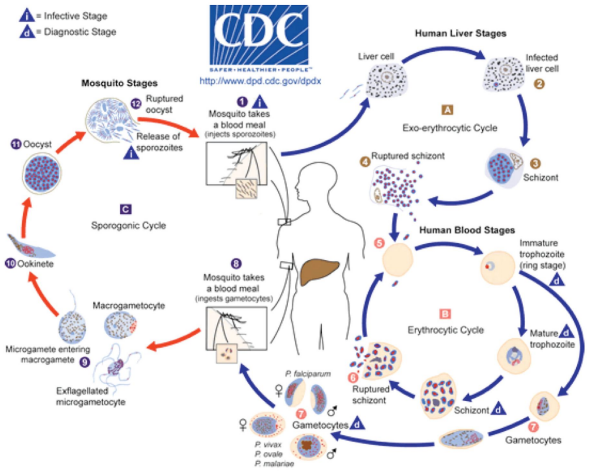
What to do:

1. Have a hypothesis already? Test it.
2. GO term/pathway/gene-set enrichment 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…

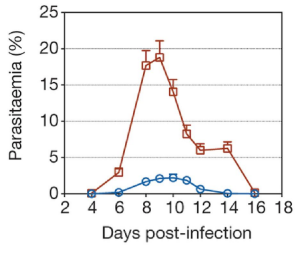
Then make a hypothesis about what genes are interesting and why. Can you test/explore this further bioinformatically? Design the next wet lab experiment

The exercise

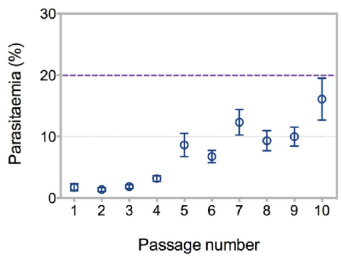
A



B



C



- *Plasmodium chabaudi*
- rodent malaria parasite
  - exhibits many characteristics associated with the pathogenesis of human infection
- serial blood passage (SBP)
  - direct injection from mouse to mouse
  - results in severe disease
- infection with parasites via mosquitoes (MT)
  - develop lower parasitaemia (presence of parasites in the blood)
  - mild, chronic disease

IS THE TRANSCRIPTOME OF MOSQUITO TRANSMITTED PARASITE DIFFERENT FROM ONE WHICH HAS NOT PASSED THROUGH A MOSQUITO?