wellcome
connecting
science

Differential Expression using RNA-Seq

Vivek Iyer
Based extensively on slides from Victoria Offord

Wellcome Sanger Institute
7 June 2024

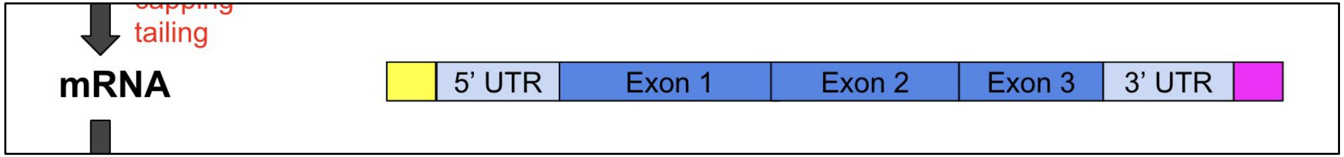
Learning outcomes

By the end of this module and tutorial you can expect to be able to:

- Appreciate the important **aspects of RNASeq experiment design**
- Understand the various technical **steps in RNASeq pipelines**
- **Align RNA-Seq reads** to a reference genome and a transcriptome
- **Visualise transcription data** using standard tools
- **Quantify the expression values** of your transcripts using standard tools
- **Perform QC** of NGS transcriptomic data
- Interpret **differential gene expression** data

This module in context

- A change of modality - sequence mRNA ! (indirectly)



- QC - basic thresholds and PCA
- See some actual stats -
  - Experiments are looking for associations between
    - gene expression (quantitative readout)
    - experimental conditions
  - Some sort of modelling needed
    - We'll see p-values and q-values flying around.
- Once you see these techniques, CRISPR screens make sense too

About me

Background

- PhD Theoretical physics 1996
- Software engineering (consultancy) 1996-2002
- Sanger 2002 - now
  - Java developer (Apollo genome browser)
  - Ensembl gene builder
  - High-throughput Mouse ESCell KO's (EuCOMM/KOMP)
  - Cancer bioinformatics / analysis (WGS, WES, CNV, CRISPR)
  - Human genetics programme informatics team (services to humgen)
    - WES/WGS variant calling and QC
    - RNA and scRNA calling and QC
    - TRE / "Data safe havens"
    - Software and disk space management

Scale + engineering

Lecture outline

- RNA-seq background
- Pipelines
  - 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

Lecture outline

- **RNA-seq background**
- Pipelines
    - 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)

What is the transcriptome?

“The complete set of transcripts in a cell

and their quantity

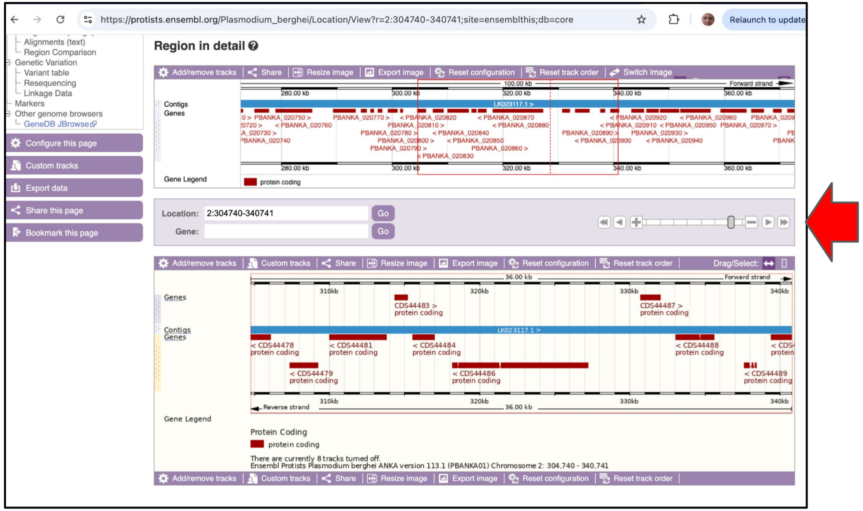
for a specific developmental stage or condition”

It’s a snapshot

- Fixed point in time
- Fixed set of conditions

RNASeq - uses NGS technology to measure this

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



https://protists.ensembl.org/Plasmodium_berghei/Location/View?r=2:304740-340741;site=ensemblthis;db=core

Region in detail

Location: 2:304740-340741

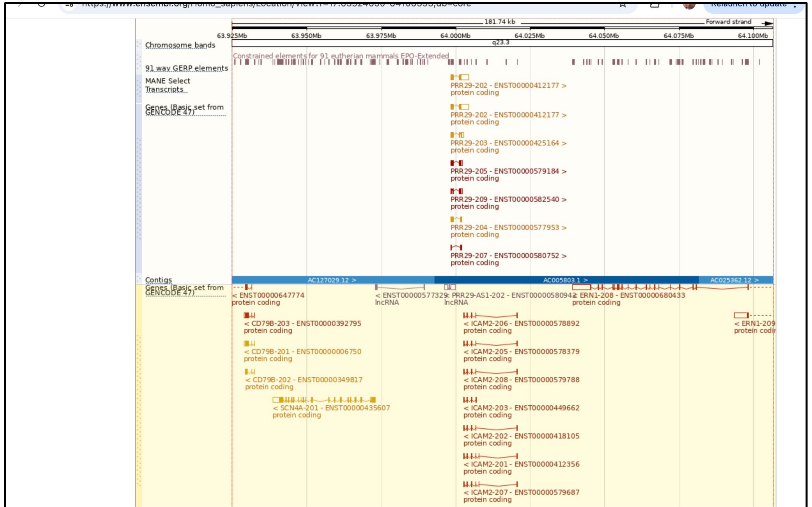
Gene:

There are currently 8 tracks turned off.
Ensembl Protists Plasmodium berghei ANKA version 113.1 (PBANKA01) Chromosome 2: 304,740 - 340,741

Plasmodium berghei

```
vvi   19254170 20 Oct 19:31 PccAS_v3_genome.fa
vvi   10554131 20 Oct 19:31 PccAS_v3_transcripts.fa
```

Transcriptome (10M) ~ 50% of genome (20M)
Exons are big cf introns / intergenic



Forward strand
Chromosome bands
91 way GERP elements
Constrained elements for 91 eutherian mammals EPO-Extended
MANE Select Transcripts
Genes (Basic set from GENCODE 47)
Contigs
PRR29-202 - ENST00000412177 >
PRR29-203 - ENST00000425164 >
PRR29-205 - ENST00000579184 >
PRR29-209 - ENST00000582540 >
PRR29-204 - ENST00000577953 >
PRR29-207 - ENST00000580752 >
protein coding
lncRNA
< CD79B-203 - ENST00000392795
< CD79B-201 - ENST00000006750
< CD79B-202 - ENST00000349817
< SCN4A-201 - ENST00000435607
< ICAM2-206 - ENST00000578892
< ICAM2-205 - ENST00000578379
< ICAM2-208 - ENST00000579788
< ICAM2-203 - ENST00000449662
< ICAM2-202 - ENST00000418105
< ICAM2-201 - ENST00000412356
< ICAM2-207 - ENST00000579687
< ERN1-209

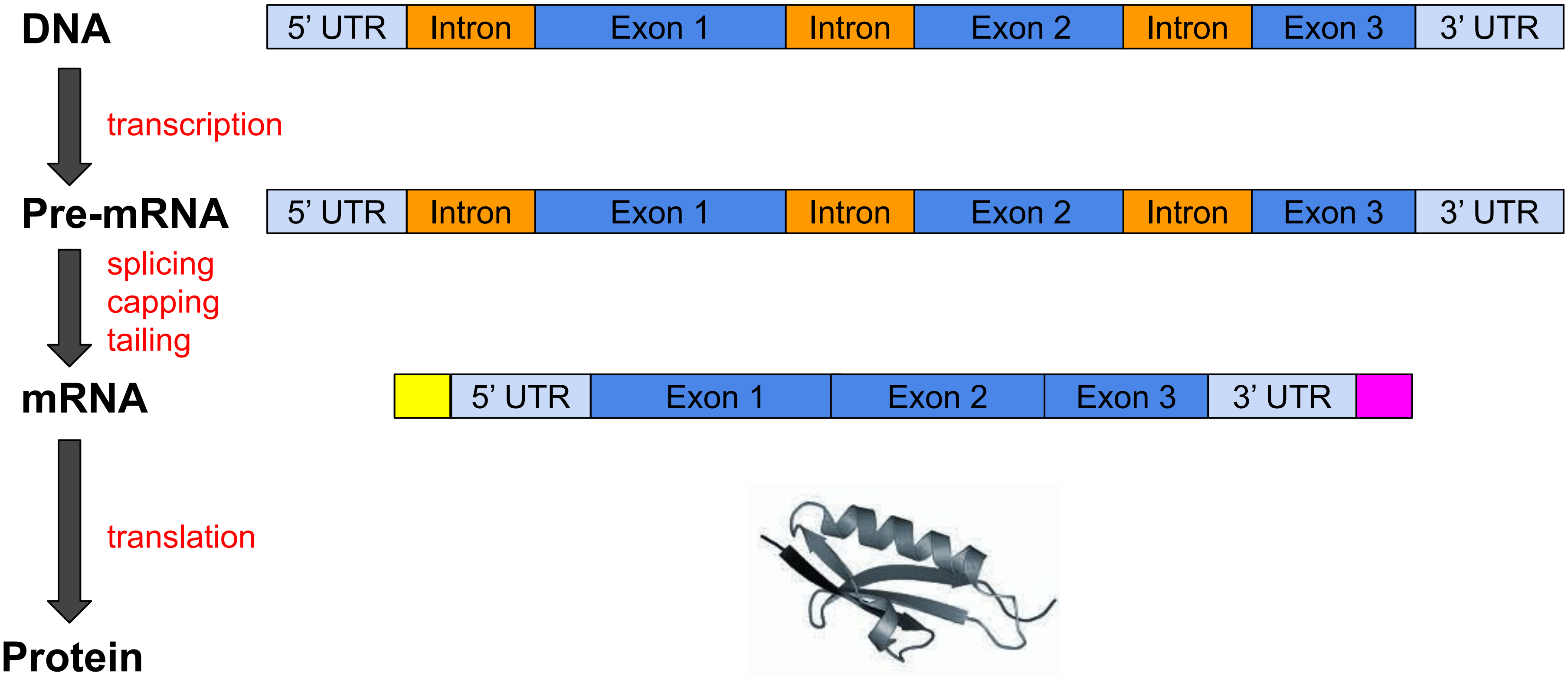
Homo Sapiens

3.0 G GRCh38_15/Homo_sapiens.GRCh38_15.fa

434 M Homo_sapiens.GRCh38.cdna.all.fa

Transcriptome ~ 10% of genome
Exons are *small* compared to introns / intergenic

Central dogma



DNA

5’ UTR | Intron | Exon 1 | Intron | Exon 2 | Intron | Exon 3 | 3’ UTR

transcription

Pre-mRNA

5’ UTR | Intron | Exon 1 | Intron | Exon 2 | Intron | Exon 3 | 3’ UTR

splicing
capping
tailing

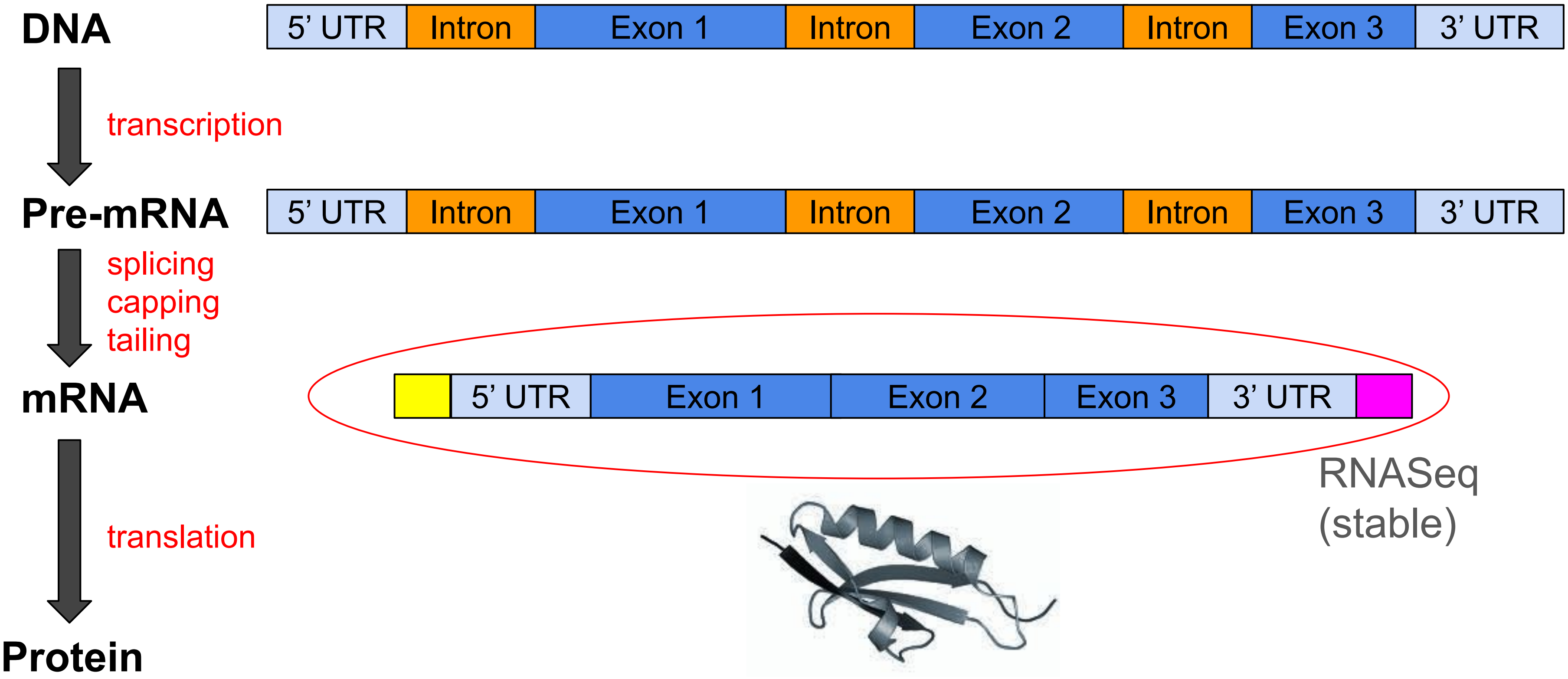
mRNA

🟨 | 5’ UTR | Exon 1 | Exon 2 | Exon 3 | 3’ UTR | 🟪

translation

Protein

Central dogma



DNA

5’ UTR | Intron | Exon 1 | Intron | Exon 2 | Intron | Exon 3 | 3’ UTR

transcription

Pre-mRNA

5’ UTR | Intron | Exon 1 | Intron | Exon 2 | Intron | Exon 3 | 3’ UTR

splicing
capping
tailing

mRNA

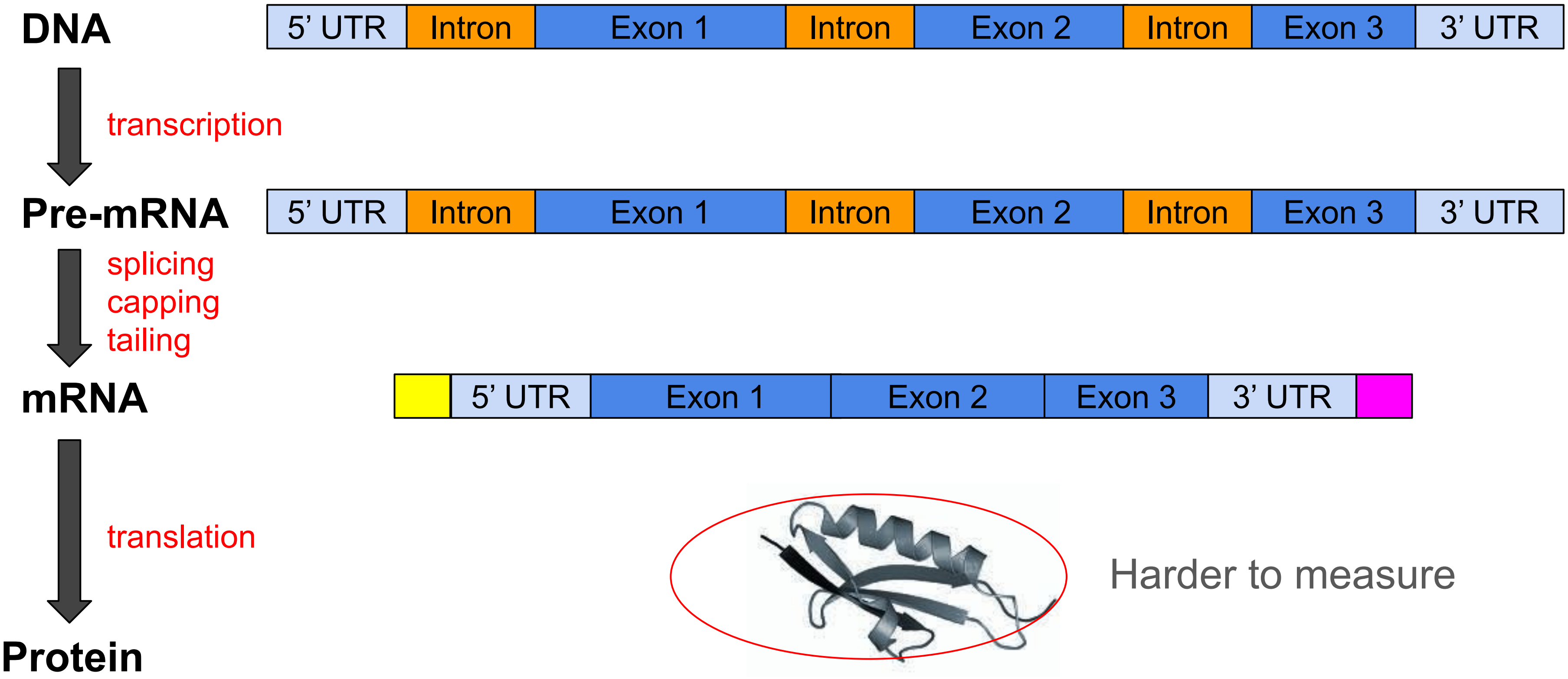
5’ UTR | Exon 1 | Exon 2 | Exon 3 | 3’ UTR

RNASeq
(stable)

translation

Protein

Central dogma



DNA

5’ UTR | Intron | Exon 1 | Intron | Exon 2 | Intron | Exon 3 | 3’ UTR

transcription

Pre-mRNA

5’ UTR | Intron | Exon 1 | Intron | Exon 2 | Intron | Exon 3 | 3’ UTR

splicing
capping
tailing

mRNA

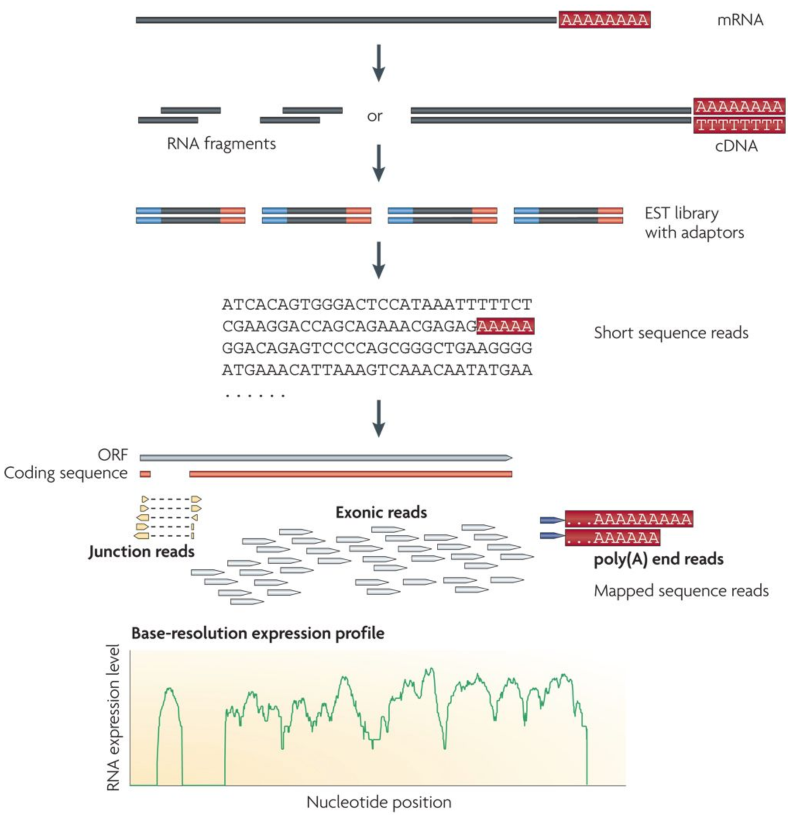
5’ UTR | Exon 1 | Exon 2 | Exon 3 | 3’ UTR

translation

Protein

Harder to measure

RNA Sequencing



- Convert to cDNA
- Fragmented
- Adapters
- sequenced

- Direct sequencing of mRNA via long-read

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 prep and type
  - sequencing depth
  - number of replicates

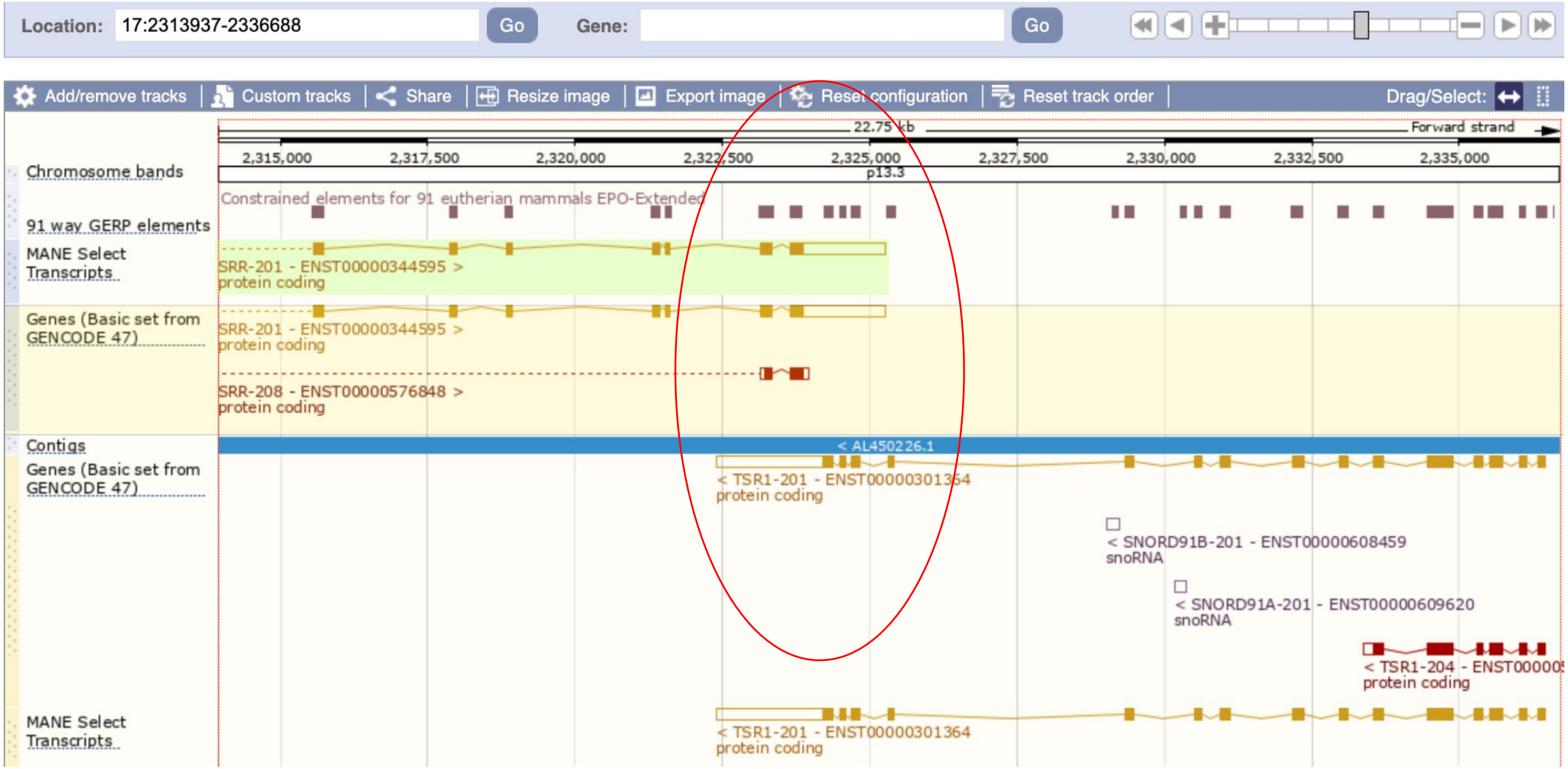
Experimental design - library preparation

- Total RNA = mRNA + **rRNA** + tRNA + regulatory RNAs...
- Ribosomal RNA can represent > 90% total RNA
- We need to 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

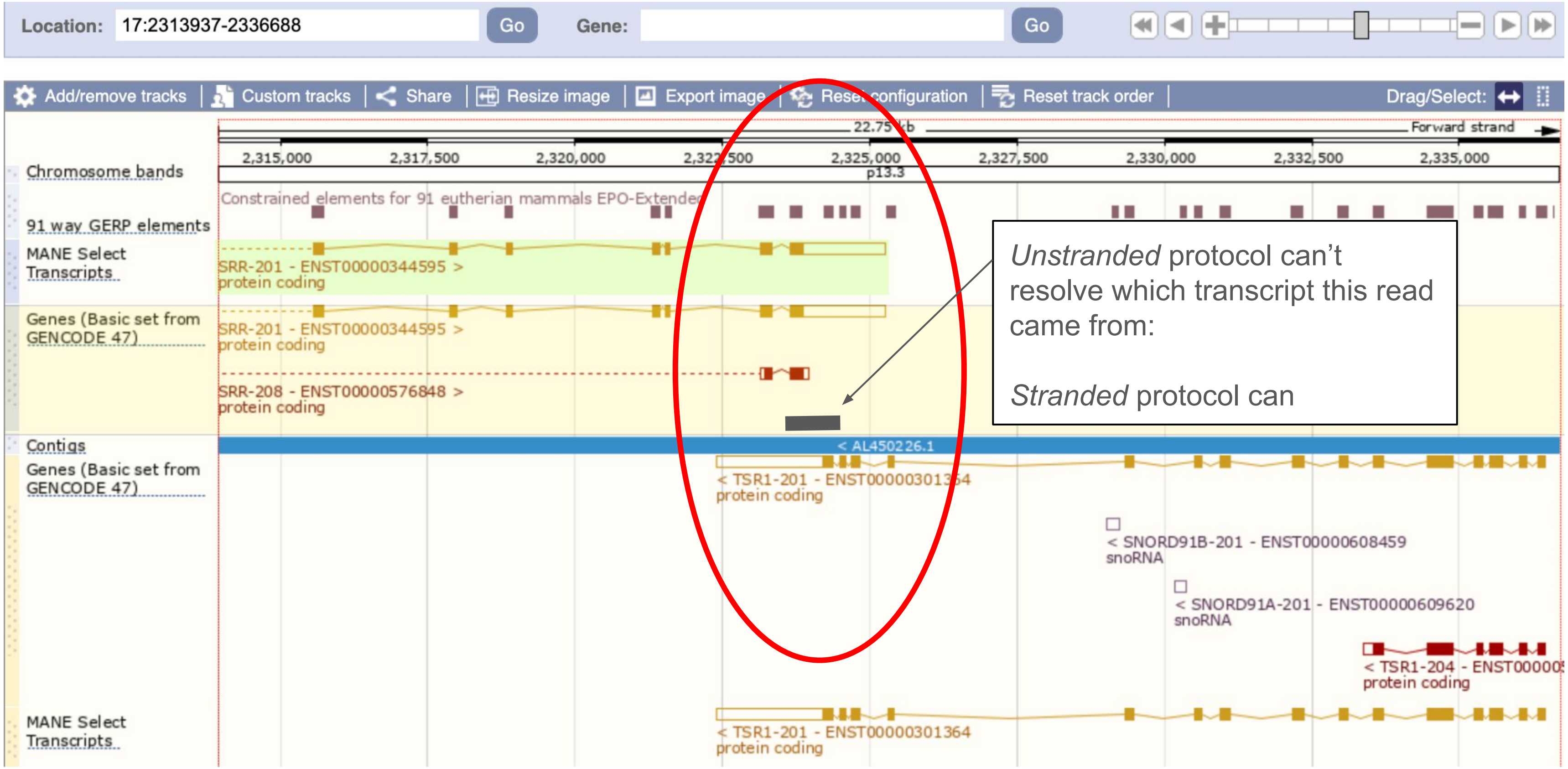
SRR 3’UTR overlaps TSR1 3’UTR + coding



Location: 17:2313937-2336688

Gene:

SRR 3’UTR overlaps TSR1 3’UTR + coding



Location: 17:2313937-2336688

Unstranded protocol can’t resolve which transcript this read came from:

Stranded protocol can

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 (***some packages adjust based on these "spike-ins"***)

Experimental design - sequencing depth

Sequencing depth: encodeproject.org

- 100bp Paired End, Human transcriptome: 30 million reads
- Novel transcripts, rare isoforms: 50-100 million reads

Next: this is how I think sequencing depth …

Experimental design - sequencing depth

This is how I think about it

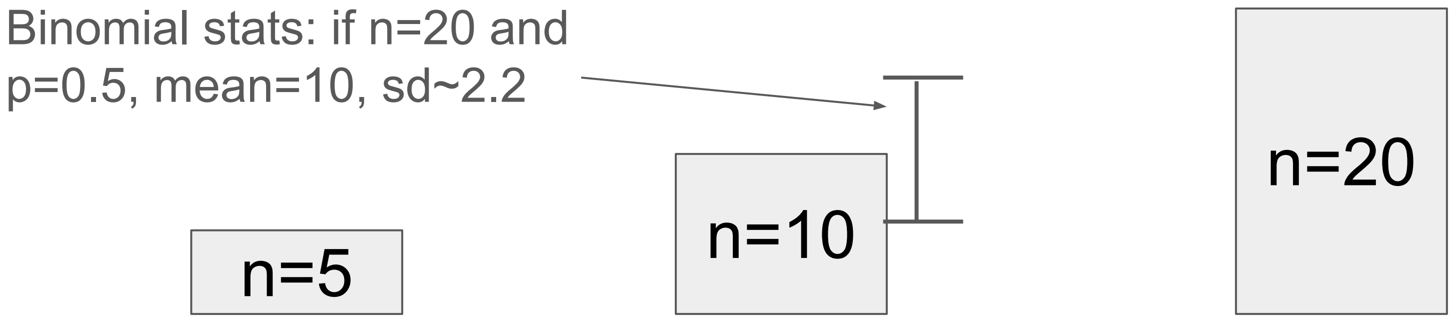
- 30 million x 100bp PE ~ 3x10^7 x 100 ~ 3 x 10^9 bases

- Transcriptome ~ 150 Mbp ~ 1.5 x 10^8 bases => **“Coverage” ~ 20x** (=30/1.5)

Experimental design - sequencing depth

This is how I think about it

- 30 million x 100bp PE ~ 3x10^7 x 100 ~ 3 x 10^9 bases
- Transcriptome ~ 150 Mbp ~ 1.5 x 10^8 bases => **“Coverage” ~ 20x** (=30/1.5)
- **This means you can “reasonably” tell apart : full / half / quarter**



Binomial stats: if n=20 and p=0.5, mean=10, sd~2.2

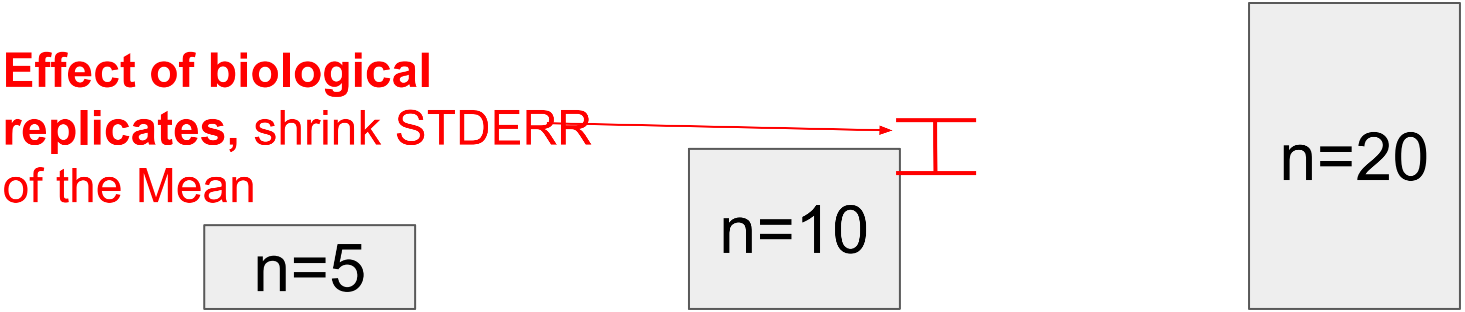
Experimental design - replicates

This is how I think about it

- 30 million x 100bp PE ~ 3x10^7 x 100 ~ 3 x 10^9 bases

- Transcriptome ~ 150 Mbp ~ 1.5 x 10^8 bases => **Coverage ~ 20x**

- **This means you can "reasonably" tell apart : full / half / quarter**



Effect of biological replicates, shrink STDERR of the Mean

Experimental design - sequencing depth *vs* replicates

- Increasing sequencing depth can increase the ability to detect low expression transcripts (i.e. increases ability to detect DE genes)
  - Returns diminish beyond a certain sequencing depth
- Increasing biological replicates increases the **accuracy of logFC** and absolute expression levels (particularly in low expression transcripts)

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

Experimental design - sequencing depth / replicates

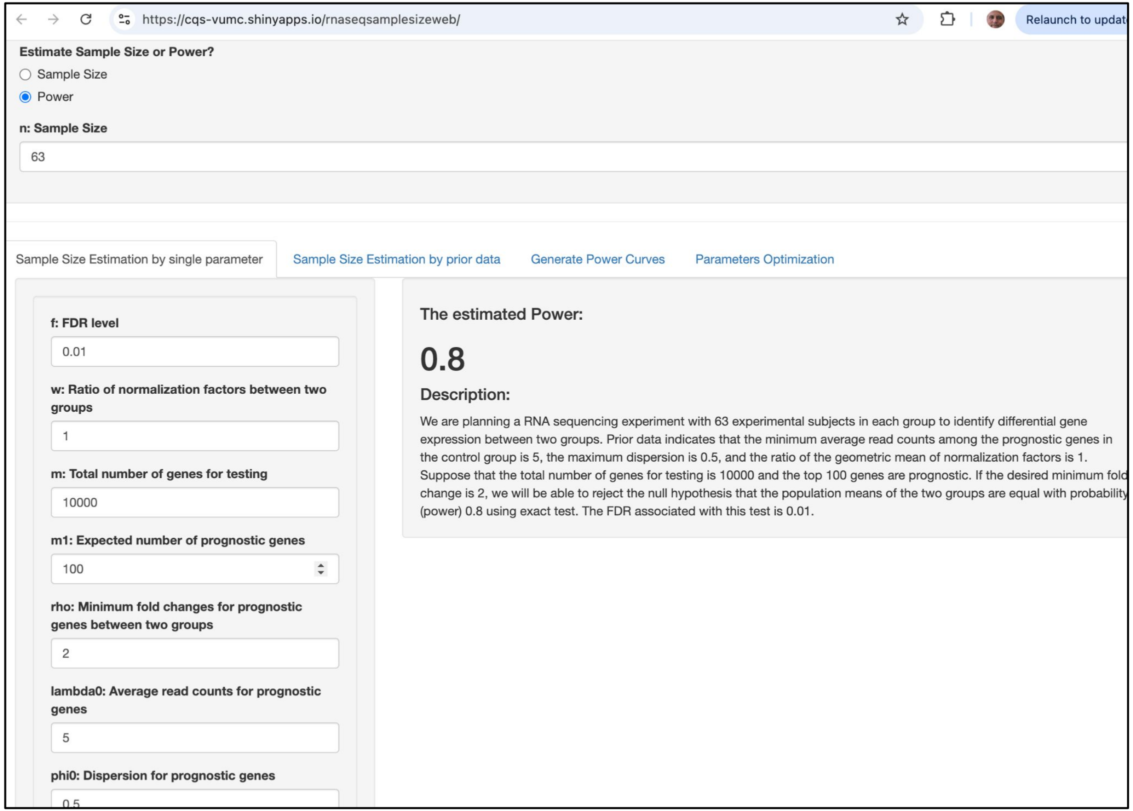
Use a power calculator to estimate sample size!

https://cqs-vumc.shinyapps.io/rnaseqsamplesizeweb/

Experimental design - sequencing depth / replicates

Plug in

- **Sample Size**
- FDR level
- Total number of genes
- *Expected variable genes*
- *Minimum FC between groups*
- **Average read counts**
- Dispersion for prognostic genes


https://cqs-vumc.shinyapps.io/rnaseqsamplesizeweb/
Relaunch to updat
Sample Size Estimation by single parameter
Sample Size Estimation by prior data
Generate Power Curves
Parameters Optimization

Estimate Sample Size or Power?

○ Sample Size

● Power

n: Sample Size

63

f: FDR level

0.01

w: Ratio of normalization factors between two groups

1

m: Total number of genes for testing

10000

m1: Expected number of prognostic genes

100

rho: Minimum fold changes for prognostic genes between two groups

2

lambda0: Average read counts for prognostic genes

5

phi0: Dispersion for prognostic genes

0.5

The estimated Power:

0.8

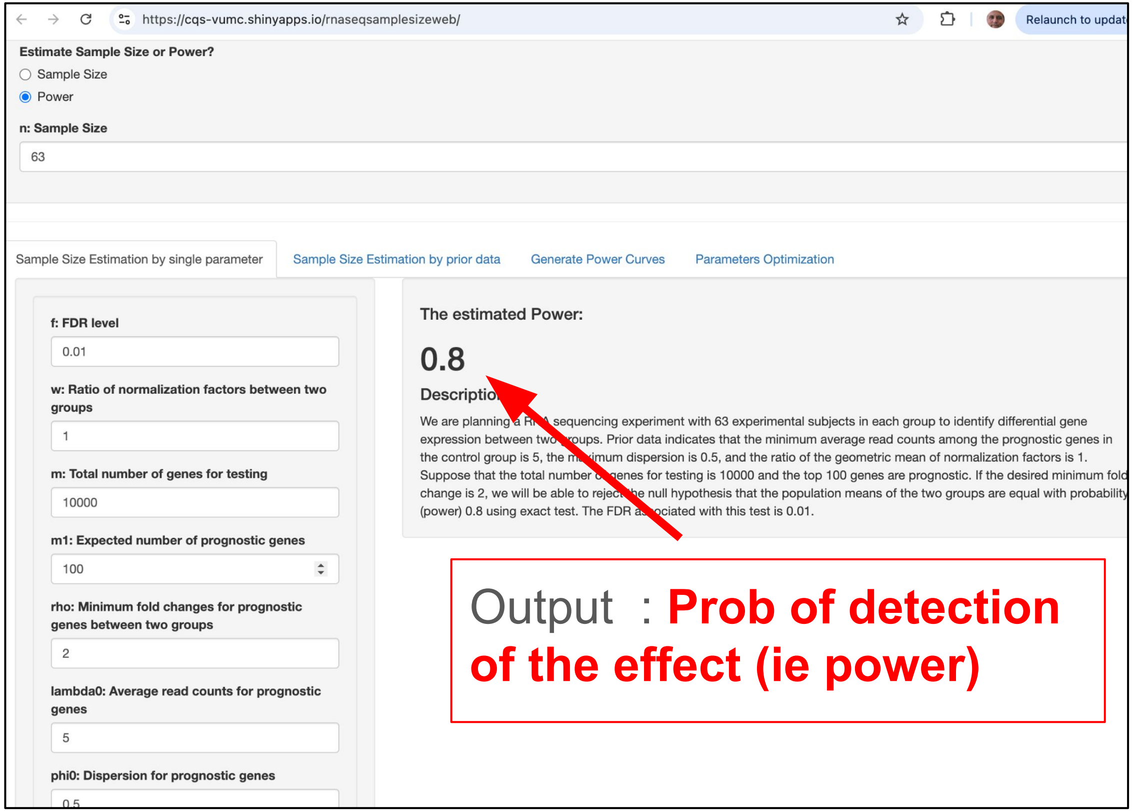
Description:

We are planning a RNA sequencing experiment with 63 experimental subjects in each group to identify differential gene expression between two groups. Prior data indicates that the minimum average read counts among the prognostic genes in the control group is 5, the maximum dispersion is 0.5, and the ratio of the geometric mean of normalization factors is 1. Suppose that the total number of genes for testing is 10000 and the top 100 genes are prognostic. If the desired minimum fold change is 2, we will be able to reject the null hypothesis that the population means of the two groups are equal with probability (power) 0.8 using exact test. The FDR associated with this test is 0.01.

Experimental design - sequencing depth / replicates

Plug in

- Sample Size
- FDR level
- Total number of genes
- *Expected variable genes*
- *Minimum FC between groups*
- Average read counts
- Dispersion for prognostic genes



https://cqs-vumc.shinyapps.io/rnaseqsamplesizeweb/
Relaunch to updat

Estimate Sample Size or Power?

○ Sample Size
● Power

n: Sample Size

63

Sample Size Estimation by single parameter
Sample Size Estimation by prior data
Generate Power Curves
Parameters Optimization

f: FDR level

0.01

w: Ratio of normalization factors between two groups

1

m: Total number of genes for testing

10000

m1: Expected number of prognostic genes

100

rho: Minimum fold changes for prognostic genes between two groups

2

lambda0: Average read counts for prognostic genes

5

phi0: Dispersion for prognostic genes

0.5

The estimated Power:

0.8

Description

We are planning a RNA sequencing experiment with 63 experimental subjects in each group to identify differential gene expression between two groups. Prior data indicates that the minimum average read counts among the prognostic genes in the control group is 5, the maximum dispersion is 0.5, and the ratio of the geometric mean of normalization factors is 1. Suppose that the total number of genes for testing is 10000 and the top 100 genes are prognostic. If the desired minimum fold change is 2, we will be able to reject the null hypothesis that the population means of the two groups are equal with probability (power) 0.8 using exact test. The FDR associated with this test is 0.01.

Output : **Prob of detection of the effect (ie power)**

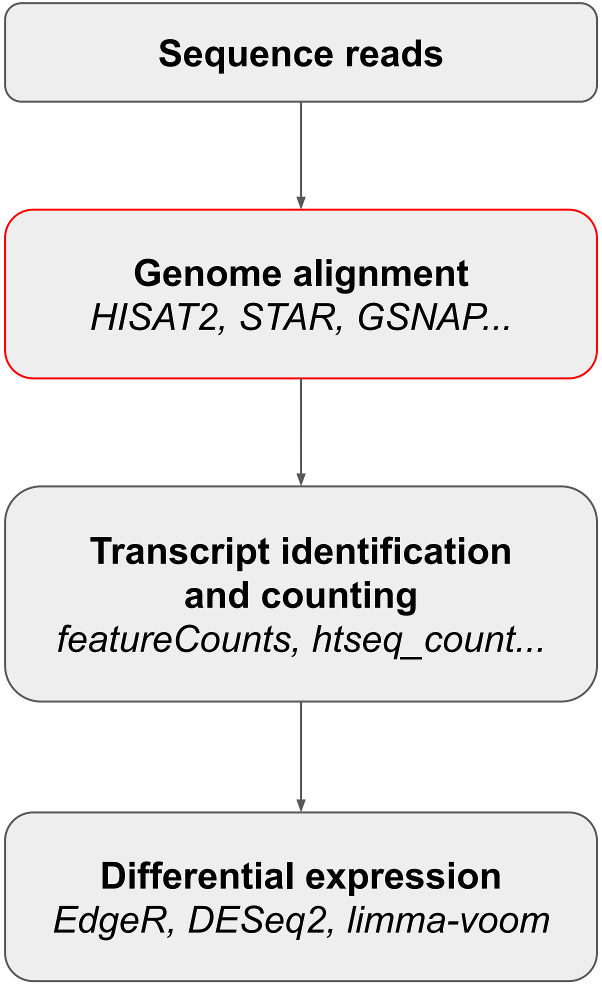
Lecture outline

- RNA-seq background
- **Pipelines**
    - 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

Key steps in any pipeline

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!!!

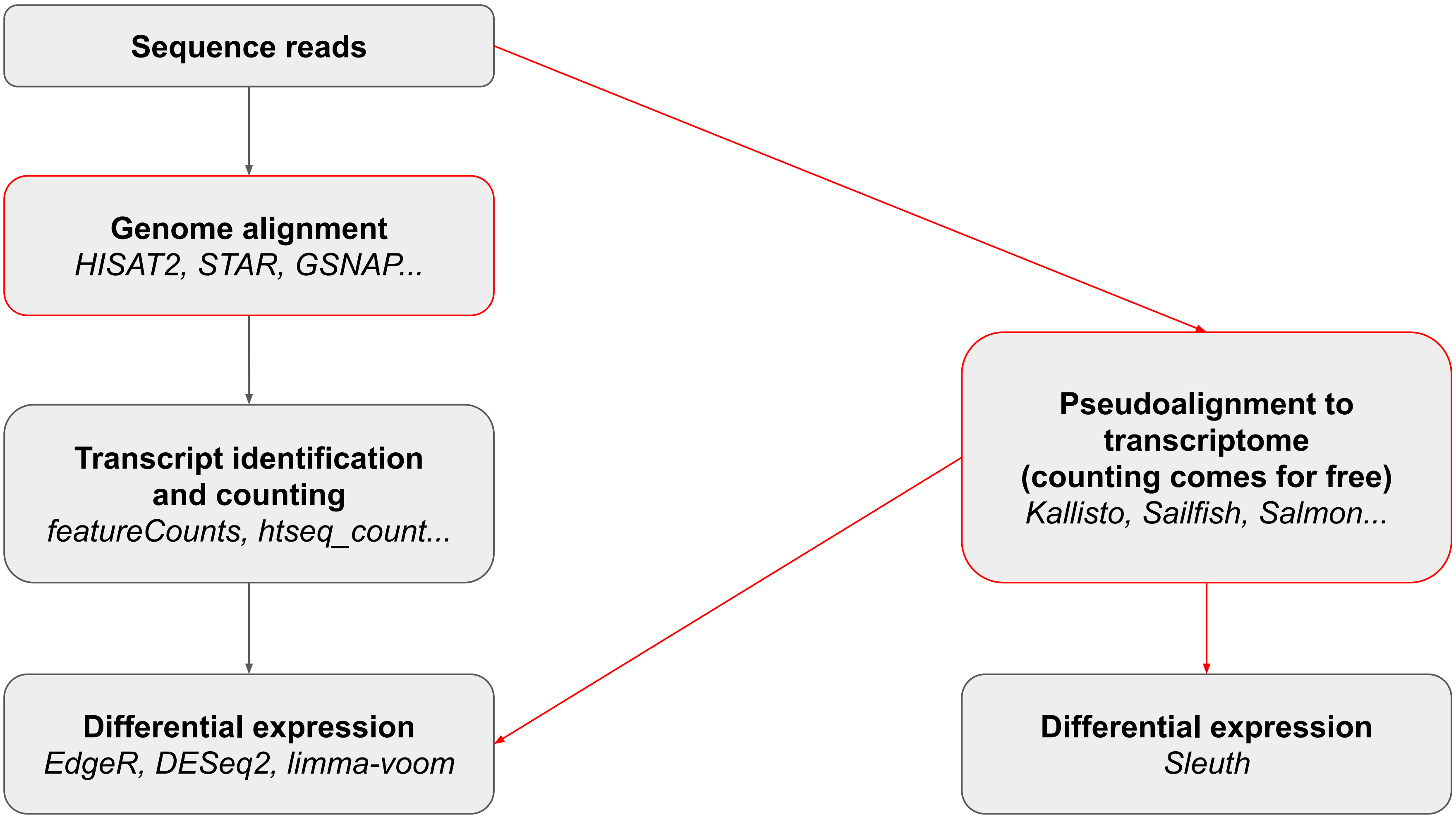


Sequence reads

Genome alignment
HISAT2, STAR, GSNAP...

Transcript identification and counting
featureCounts, htseq_count...

Differential expression
EdgeR, DESeq2, limma-voom



Sequence reads

Genome alignment
HISAT2, STAR, GSNAP...

Transcript identification and counting
featureCounts, htseq_count...

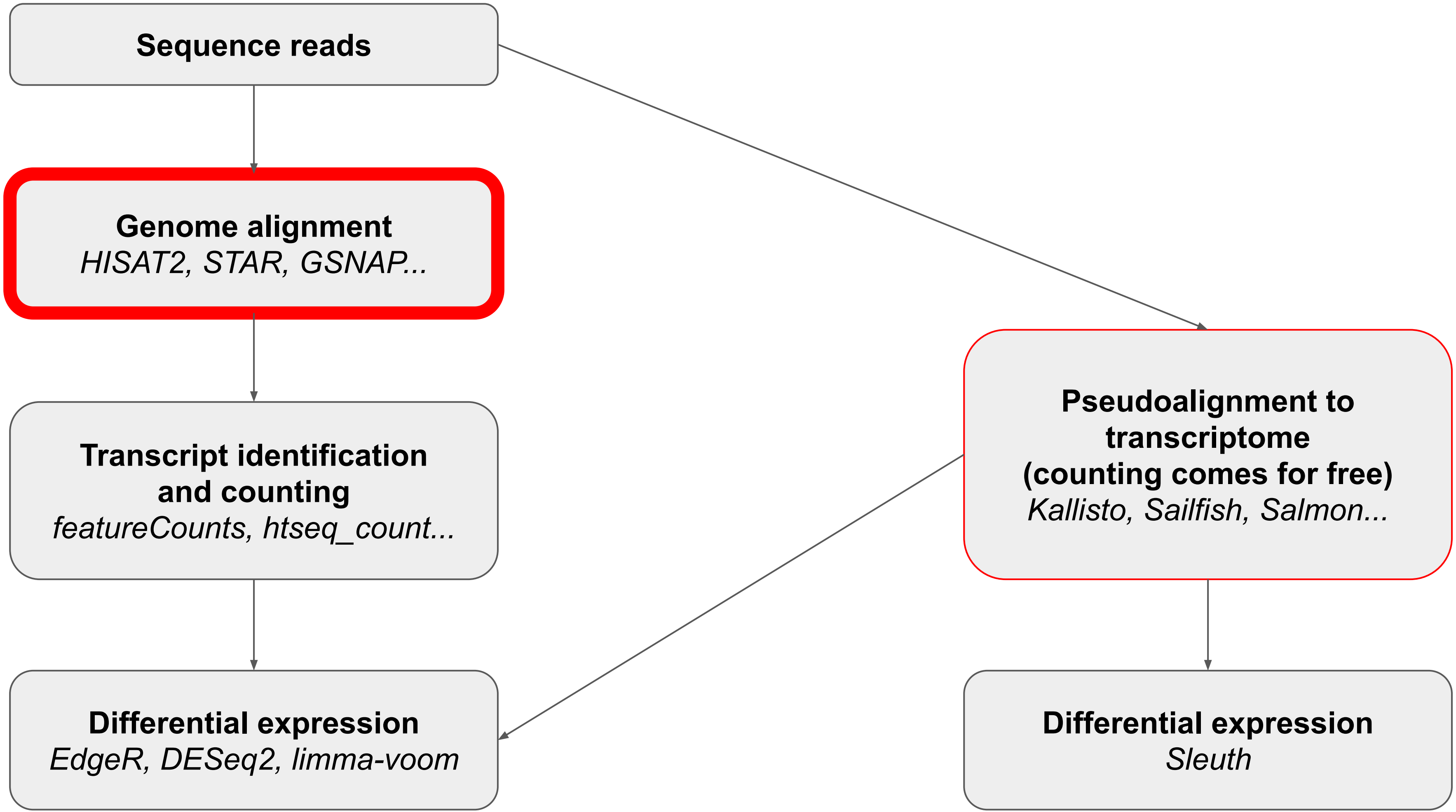
Differential expression
EdgeR, DESeq2, limma-voom

Pseudoalignment to transcriptome (counting comes for free)
Kallisto, Sailfish, Salmon...

Differential expression
Sleuth

Lecture outline

- RNA-seq background
- **Pipelines**
  - **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



Sequence reads

Genome alignment
HISAT2, STAR, GSNAP...

Transcript identification and counting
featureCounts, htseq_count...

Differential expression
EdgeR, DESeq2, limma-voom

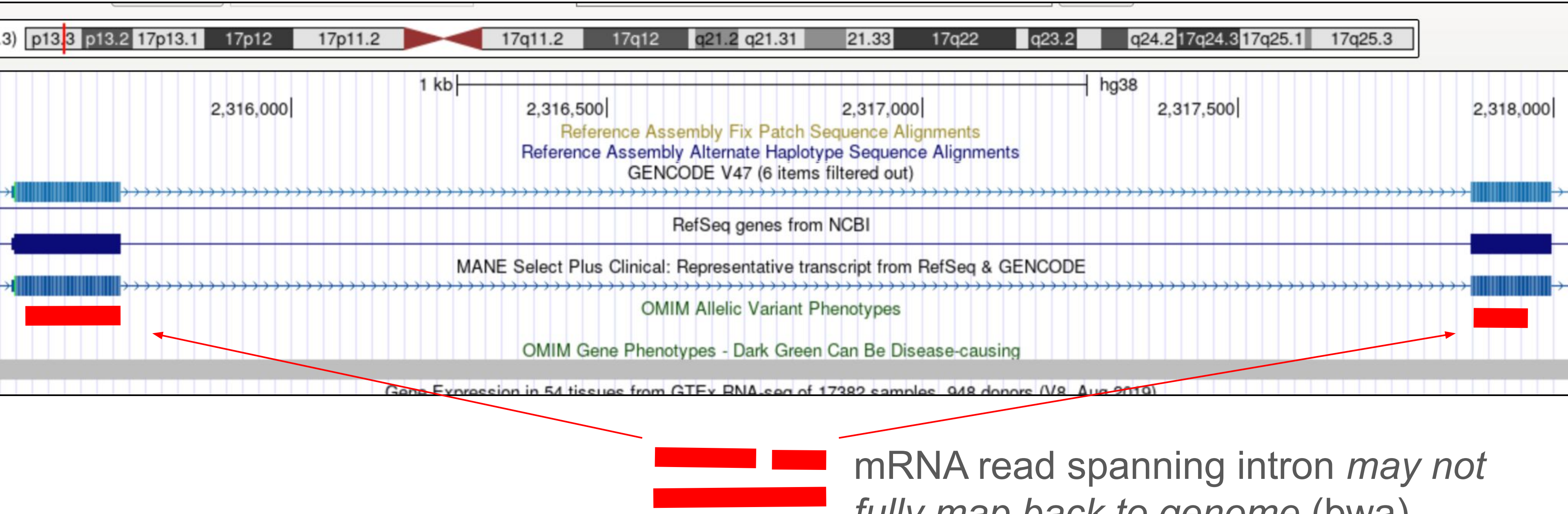
Pseudoalignment to transcriptome (counting comes for free)
Kallisto, Sailfish, Salmon...

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)


3) p13.3 p13.2 17p13.1 17p12 17p11.2 17q11.2 17q12 q21.2 q21.31 21.33 17q22 q23.2 q24.2 17q24.3 17q25.1 17q25.3
1 kb
hg38
2,316,000
2,316,500
2,317,000
2,317,500
2,318,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V47 (6 items filtered out)
RefSeq genes from NCBI
MANE Select Plus Clinical: Representative transcript from RefSeq & GENCODE
OMIM Allelic Variant Phenotypes
OMIM Gene Phenotypes - Dark Green Can Be Disease-causing
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)

mRNA read spanning intron *may not fully map back to genome* (bwa)

Need a splice-aware aligner

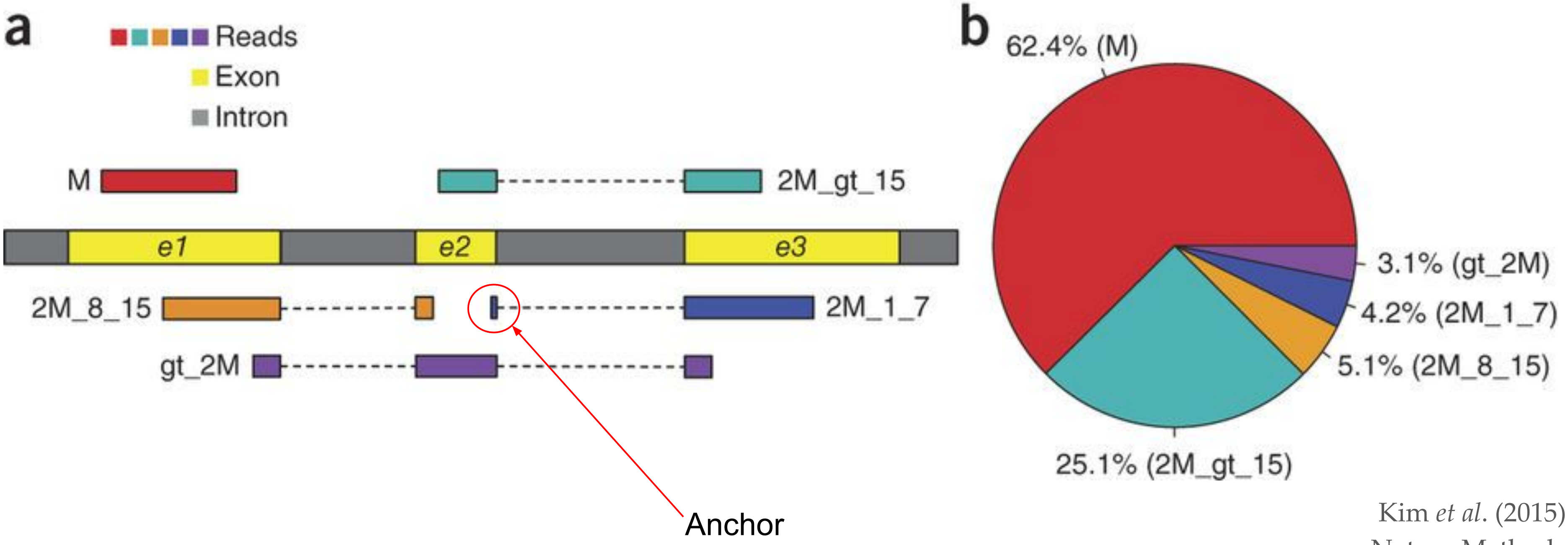
Mapping RNA-seq reads to the genome (**HISAT2**)

- 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. (others include STAR, and bowtie2 (v old))

Splice aware alignment



Anchor

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

Mapping RNA-seq reads to the genome (**HISAT2**)

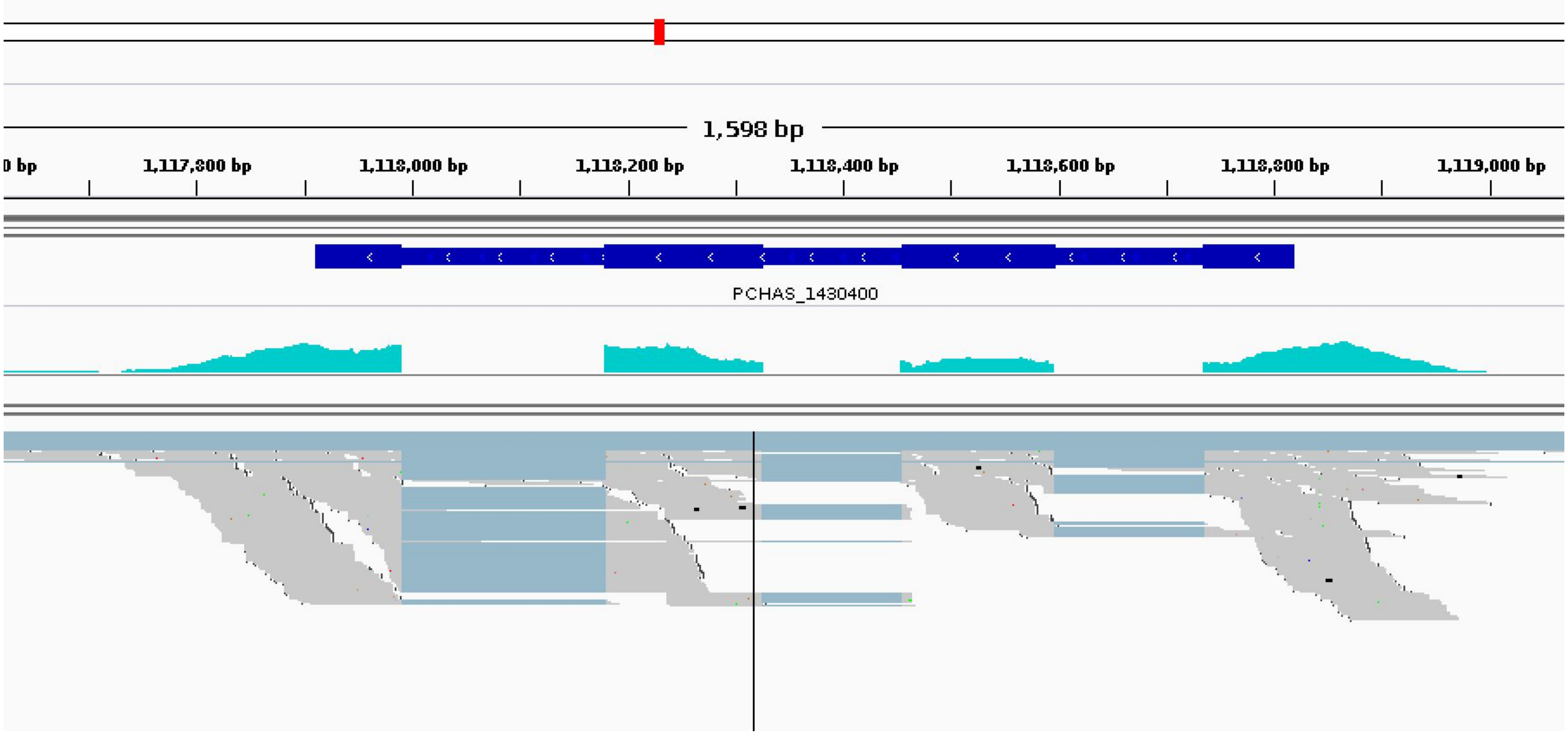
- Mapping to the genome is great for determining whether your RNA-seq data is of high quality and exploring the structure of genes of interest

- Eukaryotic genes have introns, which are not present in mature mRNA so special mapping algorithms are required (splice-aware)

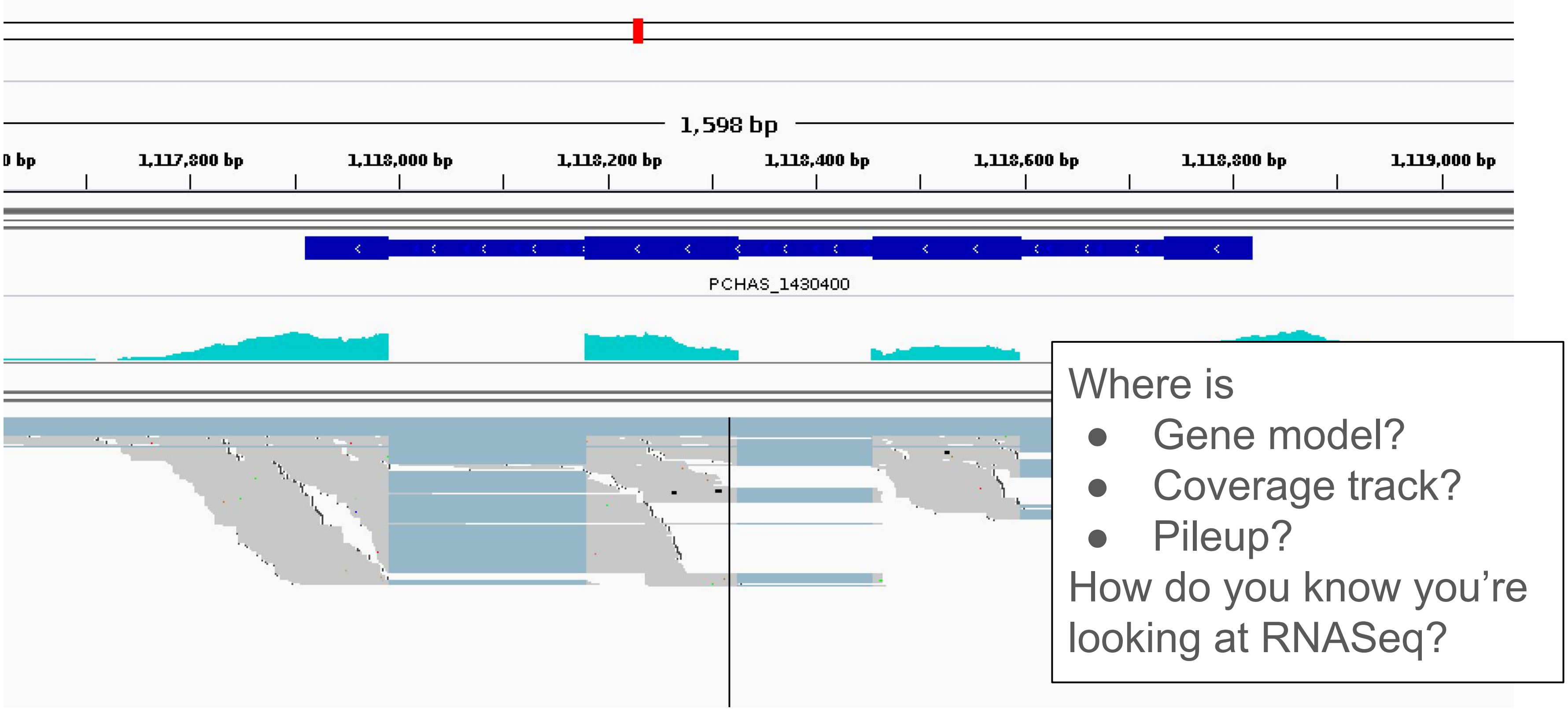
- **HISAT2** is only one such algorithm, but is accurate, fast and easy to use. (others include STAR, and bowtie2 (v old)

- HISAT2: memory footprint smaller, novel splice discovery faster

Visualisation: Integrative Genomics Viewer (IGV)



Visualisation: Integrative Genomics Viewer (IGV)



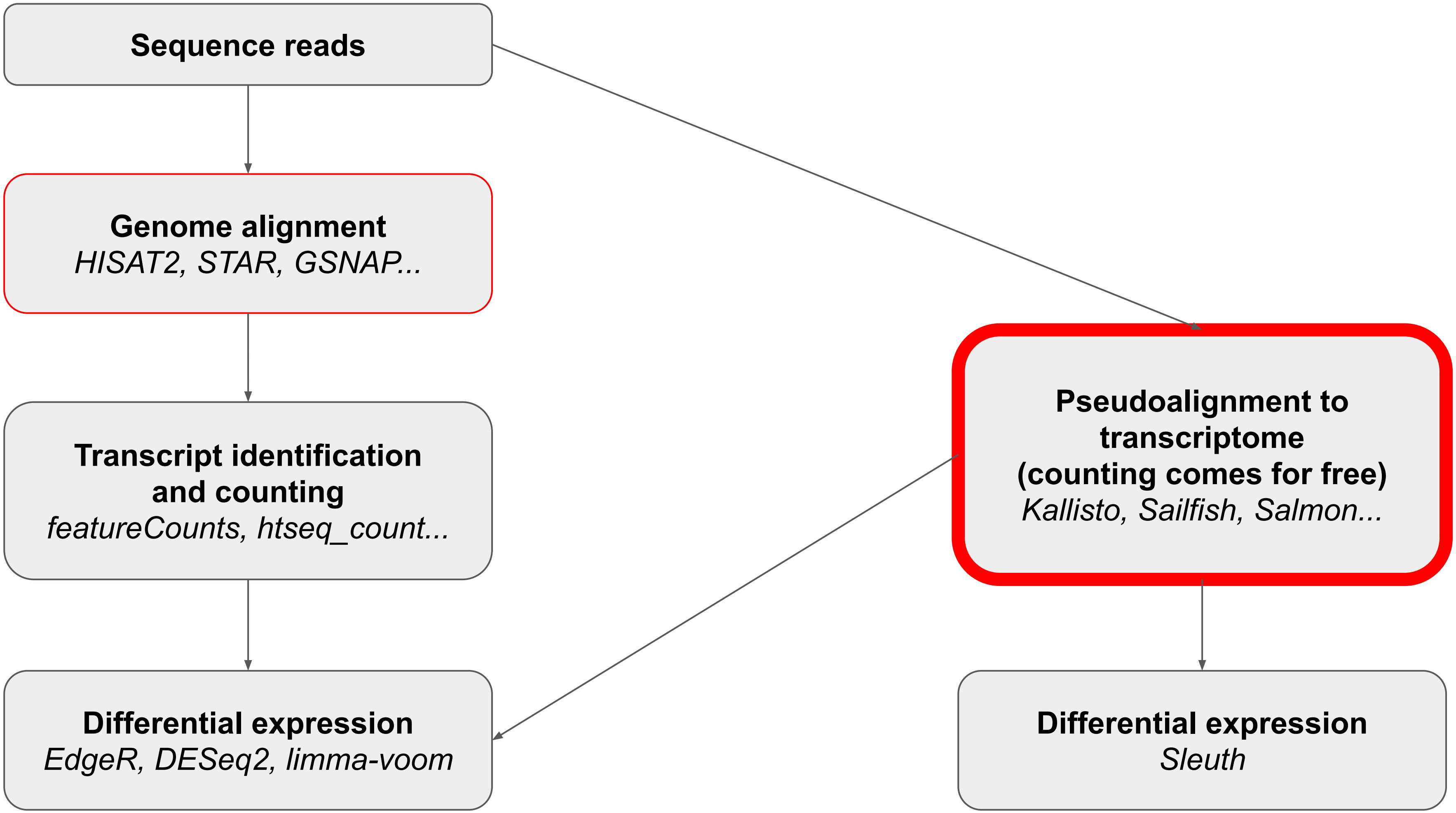
Where is

- Gene model?
- Coverage track?
- Pileup?

How do you know you're looking at RNASeq?

Lecture outline

- RNA-seq background
- **Pipelines**
  - 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



Sequence reads

Genome alignment
HISAT2, STAR, GSNAP...

Transcript identification and counting
featureCounts, htseq_count...

Differential expression
EdgeR, DESeq2, limma-voom

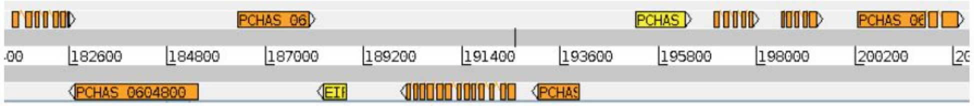
Pseudoalignment to transcriptome (counting comes for free)
Kallisto, Sailfish, Salmon...

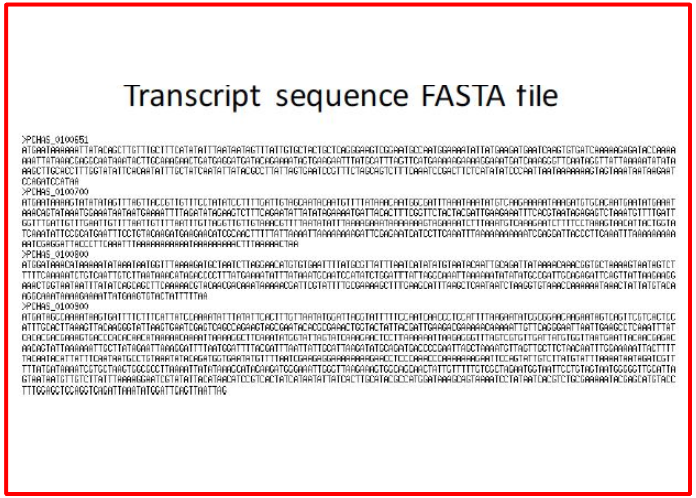
Differential expression
Sleuth

Mapping to the transcriptome and counting reads (Kallisto)

Genome sequence FASTA file

+





Transcript sequence FASTA file

```
>PCHAS_0100651
ATGAATAAAAAATTATACAGCTTGTTTGCTTTCATATATTTAATAATAGTTTATTGTGCTACTGCTCAGGGAAGTCGGAATGCCAATGGAAAATATTATGAAGATGAATCAAGTGTGATCAAAAGAGATACCAAAA
AAATTATAAACGAGGCAATAAATACTTGCAAAGAACTGATGAGGATGATACAGAAAATACTGAAGAATTTATGCATTTAGTTCATGAAAAAGAAAAGGAAATGATCAAAGGGTTCAATAGGTTATTAAAAATATATA
AAGCTTGCACCTTTGGTATATTCACAATATTTGCTATCAATATTATACGCCTTATTAGTGAATCCGTTTCTAGCAGTCTTTCAAATCCGACTTCTCATATATCCCAATTAATAAAAAAGTAGTAAATAATAAGAAT
CCAGATCCATAA
>PCHAS_0100700
ATGAATAAAAGTATATATAGTTTAGTTACCGTTGTTTCCTATATCCTTTTGATTGTAGCAATACAATGTTTTATAAACAATGGCGATTTAAATAAATATGTCAAGAAAAATAAAGATGTGCACAATGAATATGAAAT
AAACAGTATAAATGGAAATAATAATGAAAATTTTAGATATAGAAGTCTTTCAGAATATTATATAGAAAATGATTACACTTTCGGTTCTACTACGATTGAAGAAATTTCACGTAATAGAGAGTCTAAATGTTTTGATT
GGTTTGATTGTTTGAATTGTTTTAATTGTTTTAATTTGTTAGGTTGTTGTAAACGTTTTAATATATTTAAAAGAAATAAAAAAAGTAGAAAATCTTTAAATGTCAAAGAATCTTTTCCTAAAGTAACATTACTGGTA
TCAAATATTCCGCATGAATTTCCTGTACAAGATGAAGAACATCGCAACTTTTTATTAAAATTAAAAAAAGATTCGAGAATCATCCTTCAAATTTAAAAAAAAAATCGAGGATTACCCTTCAAATTTAAAAAAAAA
AATCGAGGATTACCCTTCAAATTTAAAAAAAAAAATAAAAAAAAACTTTAAAAACTAA
>PCHAS_0100800
ATGGATAAACATAAAAATATAAATAATGGTTTAAAAGATGCTAATCTTAGGAACATGTGTGAATTTTATGCGTTATTTAATCATATATGTAATACAATTGCAGATTATAAAACAAACGGTGCTAAAAGTAATAGTCT
TTTTCAAAAATCTGTCAATTGTCTTAATAAACATAGACCCCTTTATGAAAATATTTATAAATGCAATCCATATCTGGATTTTATTAGGCAAATTAAAAAATATATATGCCGATTGCAGAGATTCAGTTATTAAGAAGG
AAACTGGTAATAATTTATATCAGCAGCTTCAAAAACGTACAACGACAAATAAAAACGATTCGTATTTTGCGAAAAGCTTTGAAGCATTTAAGCTCAATAATCTAAGGTGTAAACCAAAAAATAAACTATTATGTACA
AGGCAAATAAAAGAAAATTATGAAGTGTACTATTTTTAA
>PCHAS_0100900
ATGATAGCCAAAATAAGTGATTTTCTTTCATTATCCAAAATATTTATATTCACTTTGTTAATATGGATTACGTATTTTTCCAATCAACCCTCCATTTTAAGAATATCGCGGAACAAGAATAGTCAGTTCGTCACTCC
ATTTGCACTTAAAGTTACAAGGGTATTAAGTGAATCGAGTCAGCCAGAAGTAGCGAATACACGCGAAACTGCTACTATTACGATTGAAGACGAAAAACAAAAATTGTTCAGGGAATTAATTGAAGCCTCAAATTTAT
CACACGACGAAAGTGACCCACACAACATAAAAACAAAATTAAAAGGCTTCAAATATGGTATTAGTATCAAAGAACTCCTTAAAAAATTAAGAGGGTTTAGTCGTGTTGATTATGTGGTTAATGAATTACAACGAGAC
AACAGTATTAAAAAATTGCTTATAGAATTAAAGGATTTTAATGGATTTTACGATTTAATTATTGCATTAAGATATGCAGATGACCCCGAATTAGCTAAAATGTTAGTTGCTTCTAACAATTTGGAAAAATTACTTTT
TACAATACATTATTTCAATAATGCCTGTAAATATACAGATGGTGAATATGTTTTAATCGAAGAGGAAAAAAAAGAACCTCCCAAACCCAAAAAAAGAATTCCAGTATTGTCTTATGTATTTAAAATAATAGATCGTT
TTTATGATAAAATCGTGCTAAGTGGCGCCTTAAAATTATATAAAGCATACAAGATGGAAATTGGCTTAAGAAAGTGGCAGCAACTATTGTTTTTGTCGCTAGAATGGTAATTCCTGTAGTAATGGGGTTGCATTA
GTAATAATGTTGTCTTATTTAAAAGGAATCGTATATTACATAACATCCGTCACTATCATAATATTATCACTTGCATACGCCATGGATAAAGCAGTAAAATCCTATAATCACGTCTGCGAAAAATACGAGCATGTACC
TTTGGAGCTCCAGGTCAGATTAAATATGGATTCAGTTAATTAG
```

Kallisto has two steps:

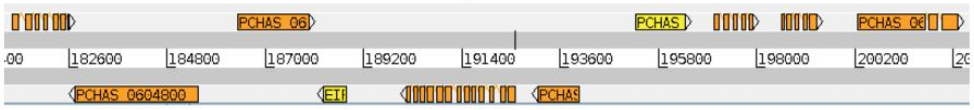
1. Building an index from the spliced transcript sequences
2. Quantify reads against the index

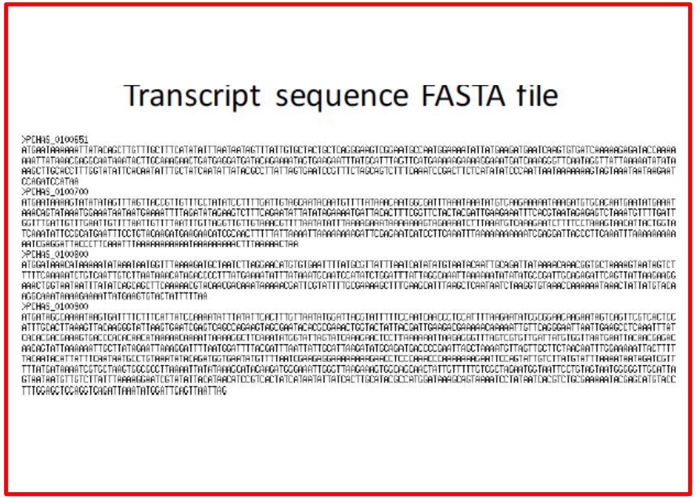
WHAT IS THE DISADVANTAGE ??

Mapping to the transcriptome and counting reads (Kallisto)

Genome sequence FASTA file

+




Transcript sequence FASTA file

```
>PCHAS_0100651
ATGAATAAAAAATTATACAGCTTGTTTGCTTTCATATATTTAATAATAGTTTATTGTGCTACTGCTCAGGGAAGTCGGAATGCCAATGGAAAATATTATGAAGATGAATCAAGTGTGATCAAAAAGAGATACCAAAA
AAATTATAAACGAGGCAATAAATACTTGCAAAGAACTGATGAGGATGATACAGAAAATAGTGAAGAATTTATGCATTTAGTTCATGAAAAAGAAAAGGAAATGATCAAAGGGTTCAATAGGTTATTAAAAAATATATA
AAGCTTGCACCTTTGGTATATTCACAATATTTGCTATCAATATTATACGCCTTATTAGTGAATCCGTTTCTAGCAGTCTTTCAAATCCGACTTCTCATATATCCCAATTAATAAAAAAGTAGTAAATAATAAGAAT
CCAGATCCATAA
>PCHAS_0100700
ATGAATAAAAGTATATATAGTTTAGTTACCGTTGTTTCCTATATCCTTTTGATTGTAGCAATACAATGTTTTATAAACAATGGCGATTTAAATAAATATGTCAAGAAAAATAAAGATGTGCACAATGAATATGAAAT
AAACAGTATAAATGGAAATAATAATGAAAATTTTAGATATAGAAGTCTTTCAGAATATTATATAGAAAATGATTACACTTTCGGTTCTACTACGATTGAAGAAATTTCACGTAATAGAGAGTCTAAATGTTTTGATT
GGTTTGATTGTTTGAATTGTTTTAATTGTTTTAATTTGTTAGGTTGTTGTAAACGTTTTAATATATTTAAAAGAAATAAAAAAAGTAGAAAATCTTTAAATGTCAAAGAATCTTTTCCTAAAGTAACATTACTGGTA
TCAAATATTCCGCATGAATTTCCTGTACAAGATGAAGAACATCGCAACTTTTTATTAAAATTAAAAAAAAGATTCGAGAATCATCCTTCAAATTTAAAAAAAAAATCGAGGATTACCCTTCAAATTTAAAAAAAAA
AATCGAGGATTACCCTTCAAATTTAAAAAAAAAAATAAAAAAAAACTTTAAAAACTAA
>PCHAS_0100800
ATGGATAAACATAAAAATATAAATAATGGTTTAAAAGATGCTAATCTTAGGAACATGTGTGAATTTTATGCGTTATTTAATCATATATGTAATACAATTGCAGATTATAAAACAAACGGTGCTAAAAGTAATAGTCT
TTTTCAAAAATCTGTCAATTGTCTTAATAAACATAGACCCCTTTATGAAAATATTTATAAATGCAATCCATATCTGGATTTATTAGGCAAATTAAAAAATATATATGCCGATTGCAGAGATTCAGTTATTAAGAAGG
AAACTGGTAATAATTTATATCAGCAGCTTCAAAAACGTACAACGACAAATAAAAACGATTCGTATTTTGCGAAAAGCTTTGAAGCATTTAAGCTCAATAATCTAAGGTGTAAACCAAAAAATAAACTATTATGTACA
AGGCAAATAAAAGAAAATTATGAAGTGTACTATTTTTAA
>PCHAS_0100900
ATGATAGCCAAAATAAGTGATTTTCTTTCATTATCCAAAATATTTATATTCACTTTGTTAATATGGATTACGTATTTTTCCAATCAACCCTCCATTTTAAGAATATCGCGGAACAAGAATAGTCAGTTCGTCACTCC
ATTTGCACTTAAAGTTACAAGGGTATTAAGTGAATCGAGTCAGCCAGAAGTAGCGAATACACGCGAAACTGGTACTATTACGATTGAAGACGAAAAACAAAAATTGTTCAGGGAATTAATTGAAGCCTCAAATTTAT
CACACGACGAAAGTGACCCACACAACATAAAAACAAAATTAAAAGGCTTCAAATATGGTATTAGTATCAAAGAACTCCTTTAAAAAATTAAGAGGGTTTAGTCGTGTTGATTATGTGGTTAATGAATTACAACGAGAC
AACAGTATTAAAAAATTGCTTATAGAATTAAAGGATTTTAATGGATTTTACGATTTAATTATTGCATTAAGATATGCAGATGACCCCGAATTAGCTAAAATGTTAGTTGCTTCTAACAATTTGGAAAAATTACTTTT
TACAATACATTATTTCAATAATGCCTGTAAATATACAGATGGTGAATATGTTTTAATCGAAGAGGAAAAAAAAGAACCTCCCAAACCCAAAAAAAGAATTCCAGTATTGTCTTATGTATTTAAAATAATAGATCGTT
TTTATGATAAAATCGTGCTAAGTGGCGCCTTAAAATTATATAAAGCATACAAGATGGAAATTGGCTTAAGAAAGTGGCAGCAACTATTGTTTTTGTCGCTAGAATGGTAATTCCTGTAGTAATGGGGTTGCATTA
GTAATAATGTTGTCTTATTTAAAAGGAATCGTATATTACATAACATCCGTCACTATCATAATATTATCACTTGCATACGCCATGGATAAAGCAGTAAAATCCTATAATCACGTCTGCGAAAAATACGAGCATGTACC
TTTGGAGCTCCAGGTCAGATTAAATATGGATTGAGTTAATTAG
```

Kallisto has two steps:

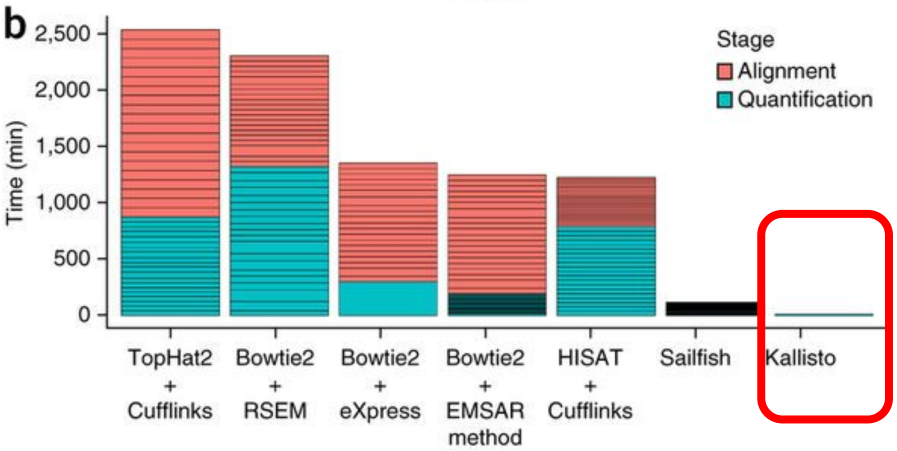
1. Building an index from the spliced transcript sequences
2. Quantify reads against the index

Kallisto cannot be used to identify novel transcripts

Mapping to the transcriptome and counting reads (Kallisto)

- It is faster because there is less target sequence
- *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
- 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

Mapping to the transcriptome and counting reads (Kallisto)



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

Lecture outline

- RNA-seq background
- **Pipelines**
  - ○ **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

- Good quality control method
- Novel splice junctions discovery
- Have to quantify afterwards
- If you have the compute, do it

- Very very fast
- Quantification comes for free
- Must have transcriptome defined
- Cant discover novel transcripts

Lecture outline

- RNA-seq background
- **Pipelines**
  - 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

The result of quantification will look like this

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

The result of quantification will look like this

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

Can you directly compare these two gene counts?
Why ? / Why not ?

The result of quantification will look like this

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

Can you directly compare these two gene counts?
Why ? / Why not ?
Overall sequencing depths differ

The result of quantification will look like this

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

Can you directly compare these two gene counts?
Why ? / Why not ?

The result of quantification will look like this

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

Can you directly compare these two gene counts?
Why ? / Why not ?
Gene lengths differ

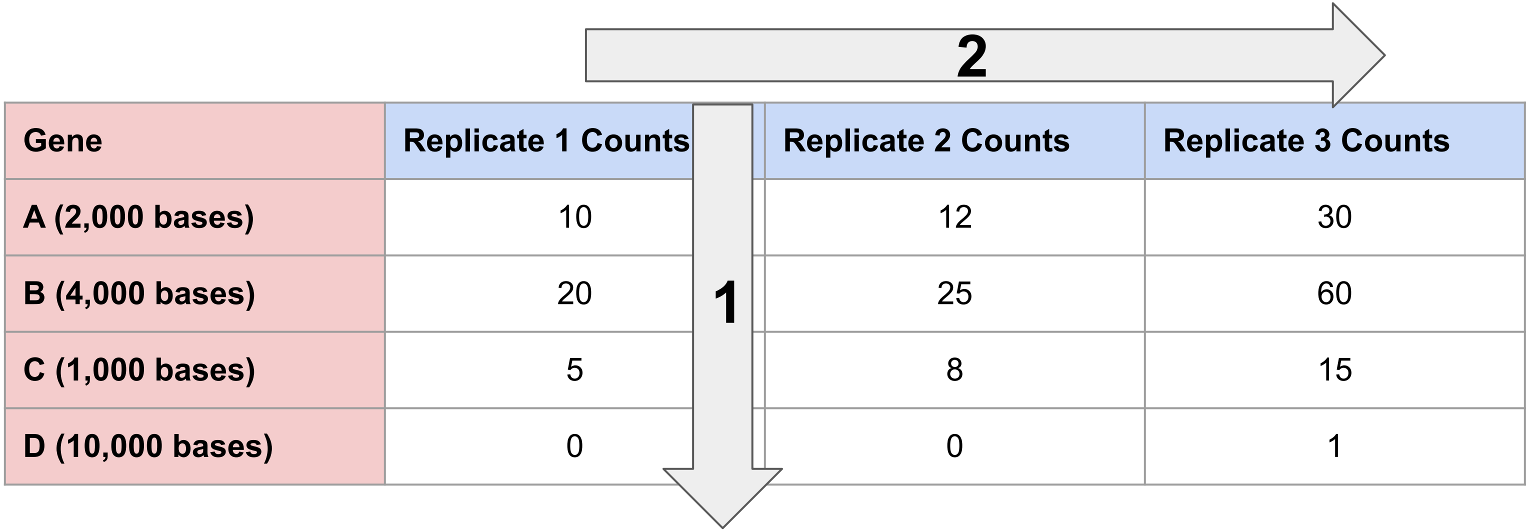
Normalisation

- 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

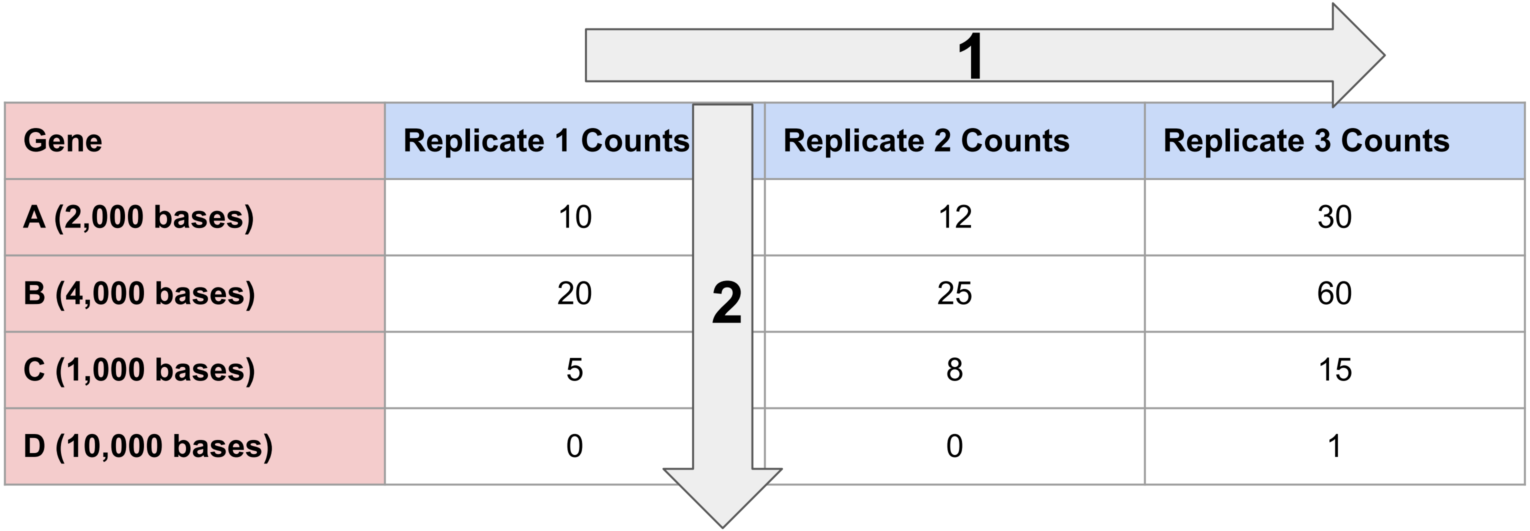
- **RPKM** - reads per kilobase per million
- **FPKM** - fragments per kilobase per million
- **TPM** - transcripts per million

RPKM - adjust for sequencing depth, *then* gene size



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

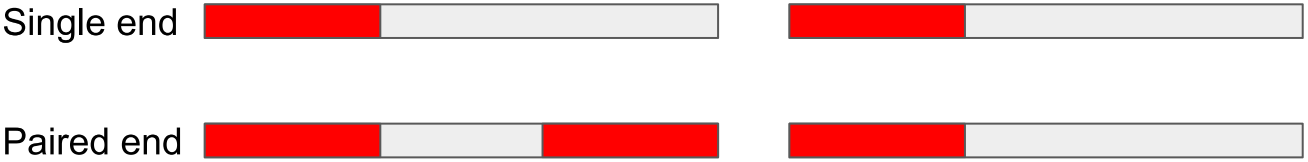
TPM - adjust for gene size, *then* sequencing depth



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

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)



Normalisation methods

- **RPKM** - reads per kilobase per million : TOTAL READ COUNT FIRST
- **FPKM** - fragments per kilobase per million : TOTAL READ COUNT FIRST
- **TPM** - transcripts per million : GENE LENGTH FIRST

Normalisation methods

- **RPKM** - reads per kilobase per million : TOTAL READ COUNT FIRST
- **FPKM** - fragments per kilobase per million : TOTAL READ COUNT FIRST
- **TPM** - transcripts per million : GENE LENGTH FIRST

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

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

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

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)

Lecture outline

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

Why QC our data?

- We gather samples and sequence
- We are trying to: detect associations between
  - genetics/expression/assay readout and
  - “condition”
- But things can go wrong
  - Samples can be tainted , e.g.
    - Mis-handled during collection (DNA or RNA degrades)
    - Sequencing of a lane can have problems (too little, too much, problem reagents)
  - Specific genomic loci can be problematic - e.g. sequencing “hotspots”, local capture problems
- Result
  - **Association is muddied** - false positive, false negative

Why QC our data?

	S1	S2	S3	S4	S5	…
L1						
L2						
L3						
L4						
L5						
…						
L						

Why QC our data?

	S1	S2	S3	S4	S5	…
L1	🟩	🟩	🟩	🟥	🟩	
L2	🟩	🟩	🟩	🟥	🟩	
L3	🟥	🟥	🟥	🟥	🟥	
L4	🟩	🟩	🟩	🟥	🟩	
L5	🟩	🟩	🟩	🟥	🟩	
…	🟩	🟩	🟩	🟥	🟩	
L	🟩	🟩	🟩	🟥	🟩	
	🟩	🟩	🟩	🟥	🟩	
	🟩	🟩	🟩	🟥	🟩	

Why QC our data?

	S1	S2	S3	S4	S5	…
L1	🟩	🟩	🟩	🟥	🟩	
L2	🟩	🟩	🟩	🟥	🟩	
L3	🟩	🟩	🟩	🟥	🟩	
L4	🟩	🟩	🟩	🟥	🟩	
L5	🟩	🟩	🟩	🟥	🟩	
…	🟩	🟩	🟩	🟥	🟩	
L	🟩	🟩	🟩	🟥	🟩	
	🟩	🟩	🟩	🟥	🟩	
	🟩	🟩	🟩	🟥	🟩	

Why QC our data?

	S1	S2	S3	S4	S5	…
L1	🟩	🟩	🟩	🟥	🟩	
L2	🟩	🟩	🟩	🟥	🟩	
L3	🟩	🟩	🟩	🟥	🟩	
L4	🟩	🟩	🟩	🟥	🟩	
L5	🟩	🟩	🟩	🟥	🟩	
…	🟩	🟩	🟩	🟥	🟩	
L	🟩	🟩	🟩	🟥	🟩	
	🟩	🟩	🟩	🟥	🟩	
	🟩	🟩	🟩	🟥	🟩	

Sample QC

- Measures properties of whole samples , e.g.
  - Alignment rate
  - Number of het sites
  - Value of PC1 etc
- Goal
  - Remove “outlying” samples

Why QC our data?

	S1	S2	S3	S4	S5	…
L1	🟩	🟩	🟩	🟩	🟩	
L2	🟩	🟩	🟩	🟩	🟩	
L3	🟥	🟥	🟥	🟥	🟥	
L4	🟩	🟩	🟩	🟩	🟩	
L5	🟩	🟩	🟩	🟩	🟩	
…	🟩	🟩	🟩	🟩	🟩	
L2000	🟩	🟩	🟩	🟩	🟩	
…	🟩	🟩	🟩	🟩	🟩	

Why QC our data?

	S1	S2	S3
L1	🟩	🟩	🟩
L2	🟩	🟩	🟩
L3	🟥	🟥	🟥
L4	🟩	🟩	🟩
L5	🟩	🟩	🟩
…	🟩	🟩	🟩
L2000	🟩	🟩	🟩
…	🟩	🟩	🟩

Locus QC

- DNASeq:
  - Measures properties of single locus (all samples) , e.g.
  - Allele Depth
  - Strand bias
  - …
  - Train model on known false and true loci
  - Goal
    - Remove problematic loci
- RNASeq: remove low-expressed / low-variability genes

Why QC our data?

	S1	S2	S3	S4	S5	…
L1	🟩	🟩	🟩	🟥	🟩	
L2	🟩	🟩	🟩	🟥	🟩	
L3				🟥	🟩	
L4				🟥	🟩	
L5				🟥	🟩	
…				🟥	🟩	
L				🟥	🟩	
				🟥	🟩	
	🟩	🟩	🟩	🟥	🟩	

Right now, for RNASeq:

Focus on Sample QC

- Number of sequencing reads
- Alignment rate
- PCA: Dimensionally reduced cross-transcriptome expression

Result - find outlying samples

Why QC our data?

Flagging outlying samples based on read count / align rate

Sample	Number of reads	Alignment Rate	QC Pass?
S1	60M	85%	
S2	10M	80%	
S3	50M	85%	
S4	65M	50%	
S5	55M	55%	
S6	70M	75%	

Why QC our data?

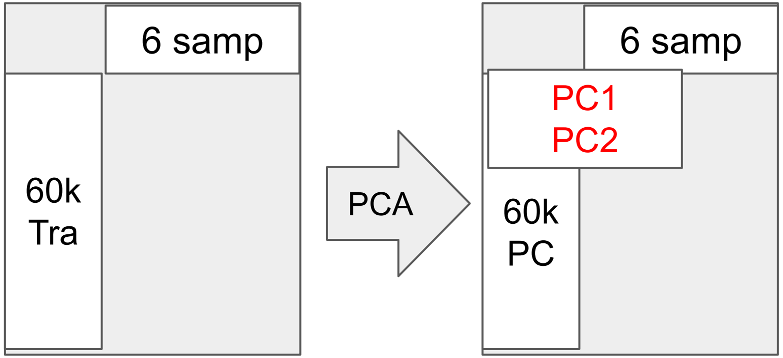
Flagging outlying samples based on read count / align rate

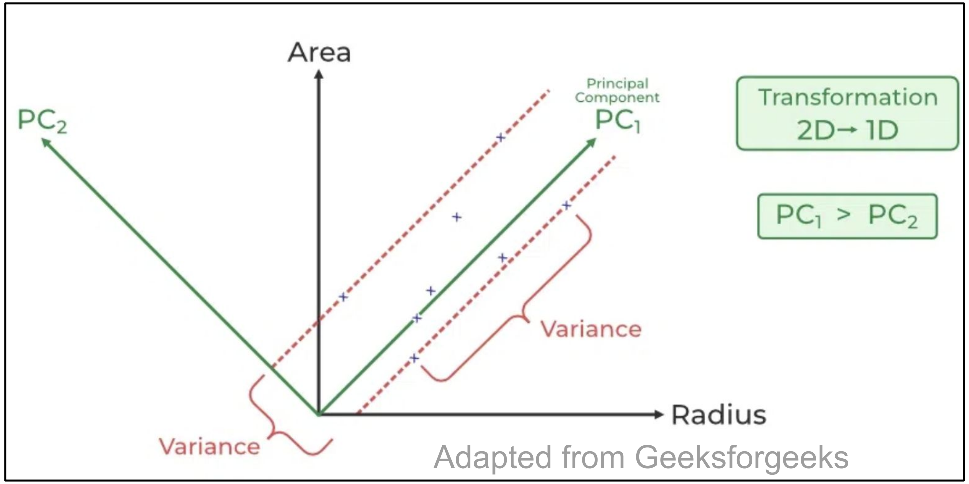
Sample	Number of reads	Alignment Rate	QC Pass?
S1	60M	85%	
S2	**10M**	80%	?
S3	50M	85%	
S4	65M	**50%**	?
S5	55M	**55%**	?
S6	70M	75%	

Why QC our data?

Flagging outlying samples based on PCA

- Start with N samples in D-dimensional space (RNASeq D~60k)
- Transform (rotate) into a *new* D-dim space (PC space)
- Dimensional reduction: Focus on first 2-4 PCs





Adapted from Geeksforgeeks

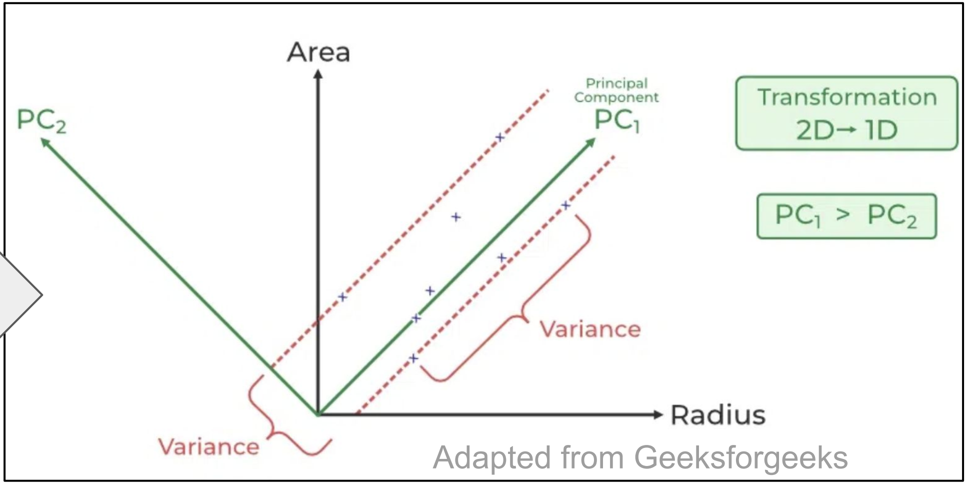
Why QC our data?

Flagging outlying samples based on PCA

- Start with N samples in D-dimensional space (RNASeq D~60k)
- Transform (rotate) into a new D-dim space (PC space)
- Dimensional reduction: Focus on first 2-4 PCs

Visualise N samples in new dimensions

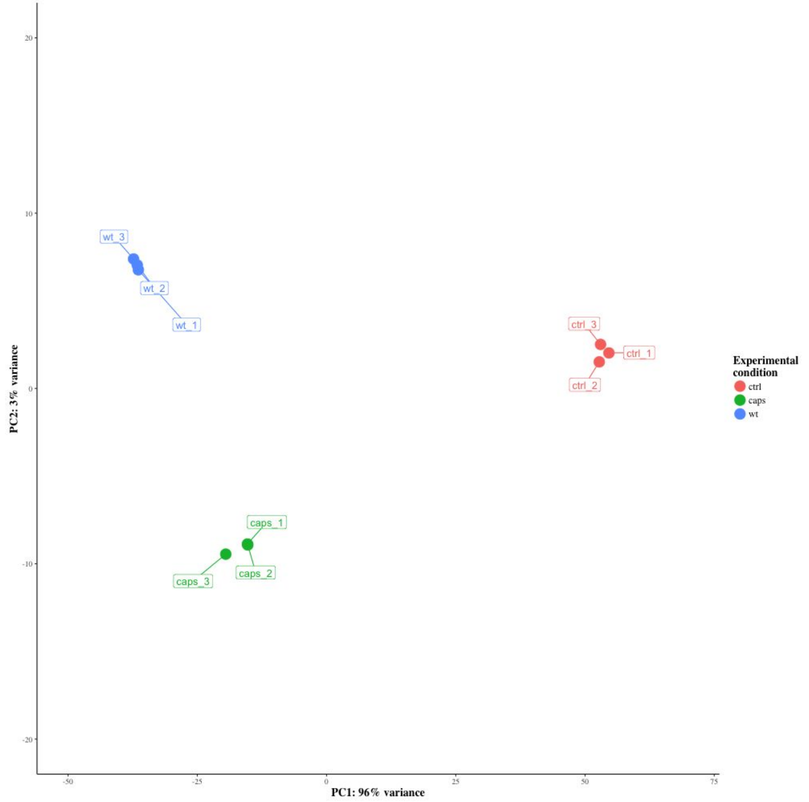
- Order new dimensions in order of sample variability
- The variability in each dimension is maximally independent (no covariance between dimensions)



Adapted from Geeksforgeeks

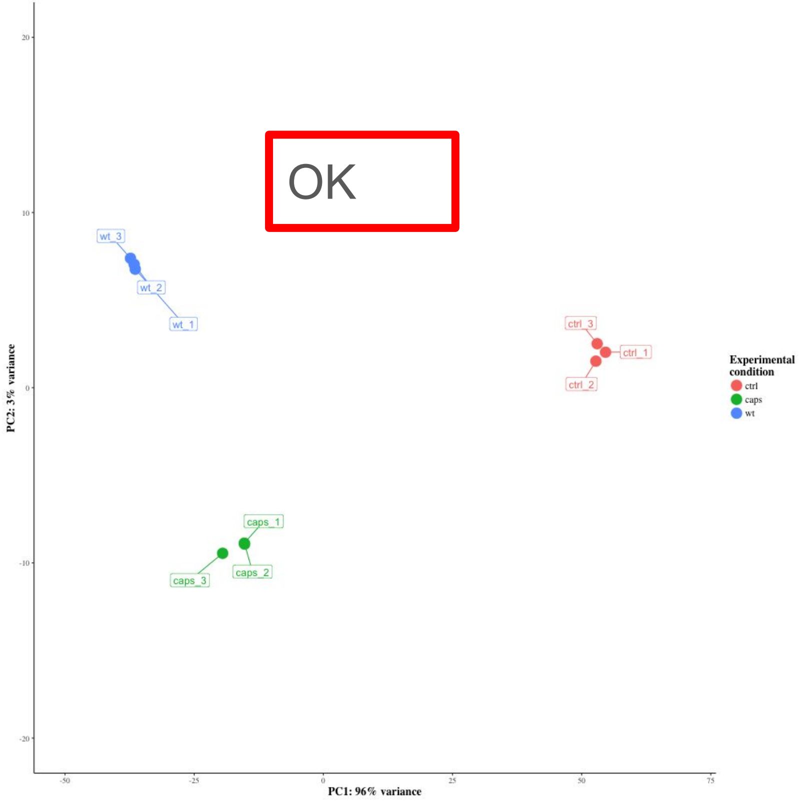
Principal component analysis (PCA)

- Use to look at variation and strong patterns within data

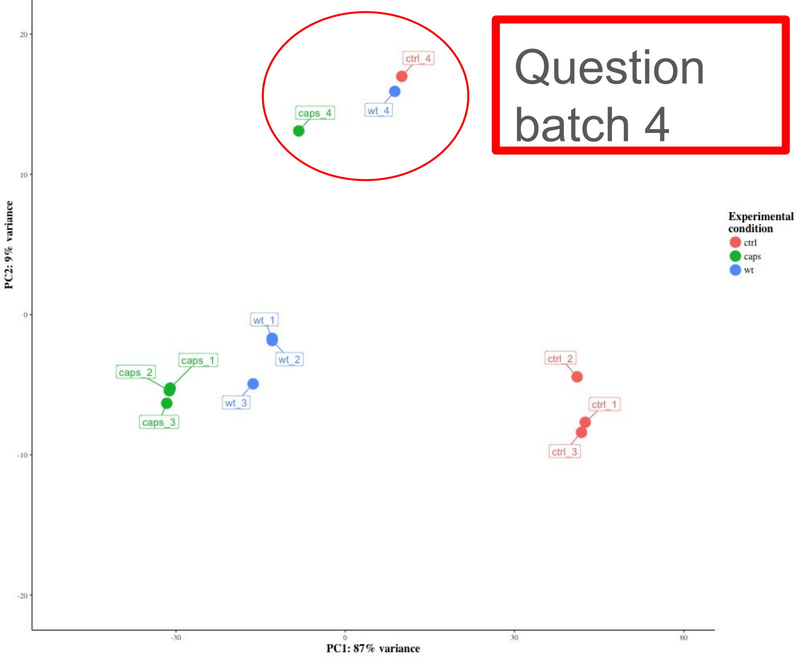


Why QC our data?

In practice, PCA picks up batch effects
Plot N samples in PC1 and PC2
You *don't want* samples clustering by an experimental artefact (eg processing batch)



OK



Question batch 4

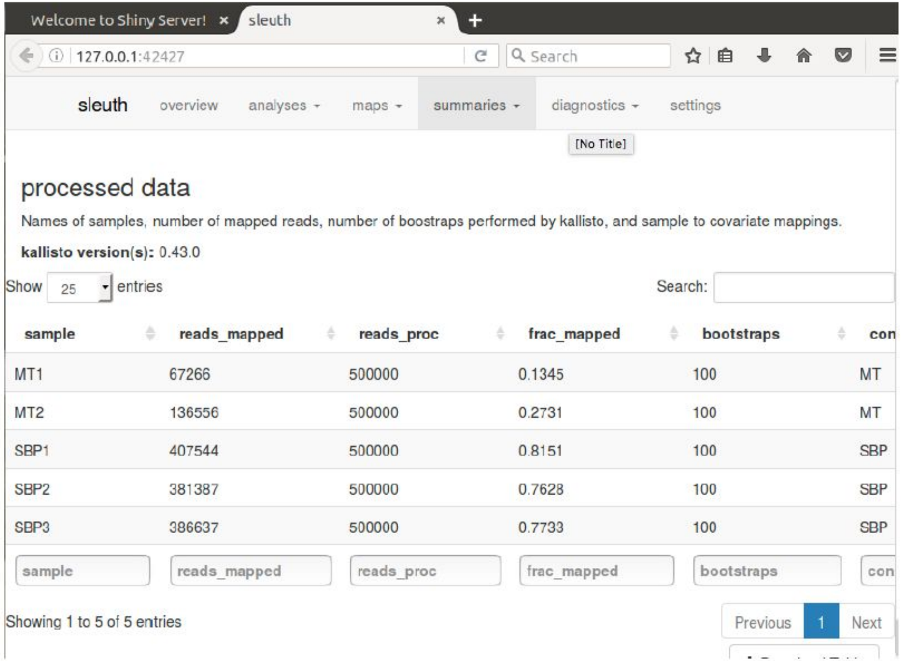
Lecture outline

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

Determining differential expression (Sleuth)

- Many packages to do Differential Gene Expression
    - DESeq2, EdgeR, Limma/Voom
    - Sleuth - companion to Kallisto
- Why you can’t (really) use a gene by gene t-test
    - “Size factors” - we can do better than the basic method in FPKM or TPM
    - **We don’t normally have enough replicates to do traditional tests of significance for RNA-seq data (methods do gene-variance modelling in some way)**
    - **You may want to account for many different input conditions (e.g. experimental + genetic)**
        - Methods compare two linear models - with and without experimental condition
    - Need to account for multiple-testing effect (q-value vs p-value)

QC with Sleuth



processed data

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

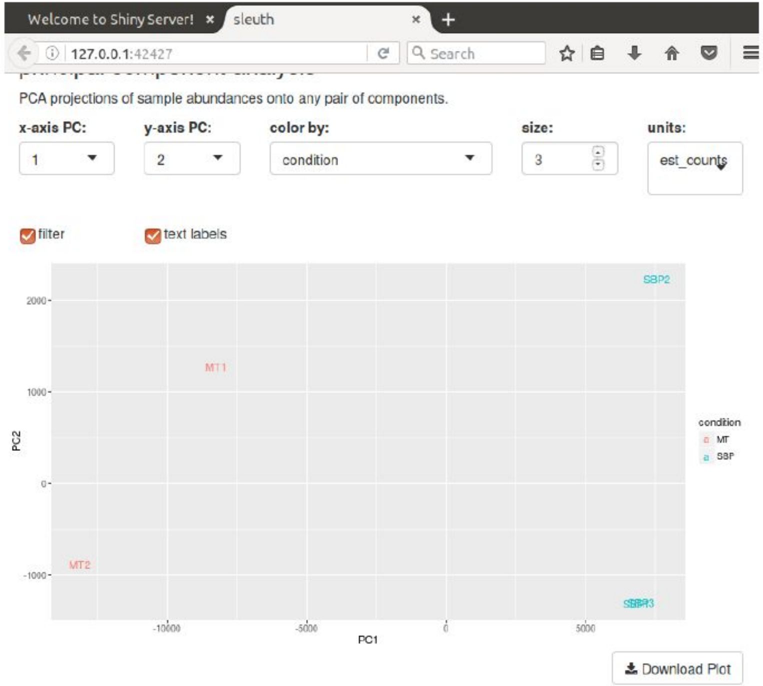
kallisto version(s): 0.43.0

Show 25 entries

Search:

sample	reads_mapped	reads_proc	frac_mapped	bootstraps	con
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

Showing 1 to 5 of 5 entries



PCA projections of sample abundances onto any pair of components.

x-axis PC: 1

y-axis PC: 2

color by: condition

size: 3

units: est_counts

☑ filter

☑ text labels

What to do next with your gene list

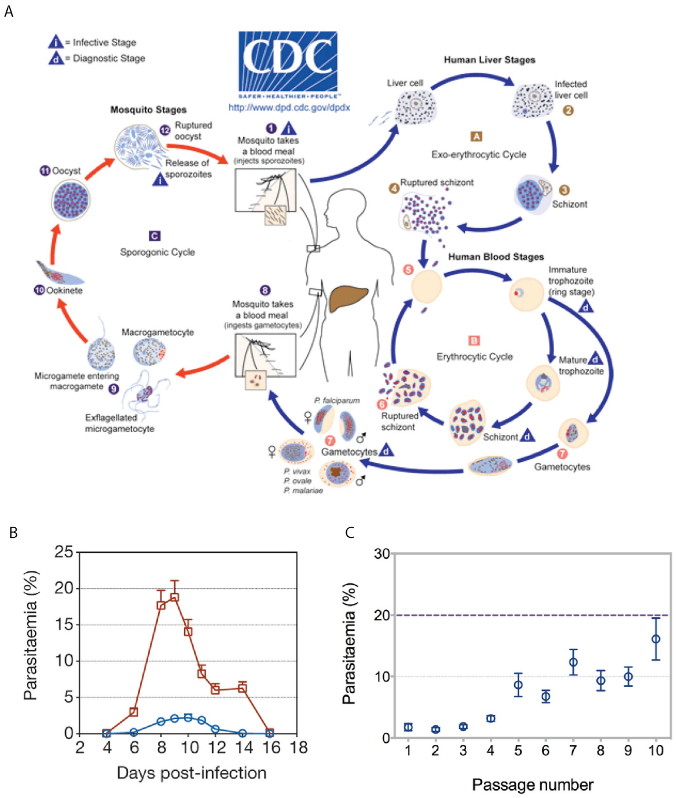
When you have a list of differentially expressed genes, things start to get difficult.

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. Stare at a volcano plot of effect size x p-value, draw cutoffs in effect size and cherry pick genes ;-)

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



- *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 (MT) *DIFFERENT FROM*
- ONE WHICH HAS NOT PASSED THROUGH A MOSQUITO (SBP)?