Placental Bioinformatics Course

Dr Russell S. Hamilton

Email: rsh46@cam.ac.uk
Twitter: @drrshamilton

Dr Xiaohui Zhao

Email: xz289@cam.ac.uk

Dr Malwina Prater

Email: mn367@cam.ac.uk

Course Materials:
https://github.com/CTR-BFX/2019-PlacentalBiologyCourse

Version 0.1: 20190619

Program

Lecture 1: What is RNA-Seq?

- Introduction to RNA-Seq and sequencer options
- From Sequencer to Quality Control and Aligning reads

Practical 1:

- From FASTQ to BAM

Lecture 2: Gene Counts to hypothesis testing

- Experimental Design
- Gene quantification
- Differential Gene Expression Analysis

Practical 2:

- BAM to DEGs

Program

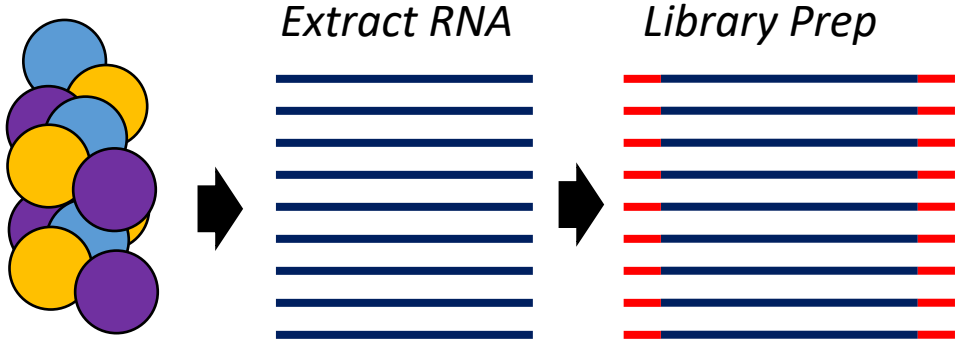
Lecture 1: What is RNA-Seq?

- Introduction to RNA-Seq and sequencer options
- From Sequencer to Quality Control and Aligning reads

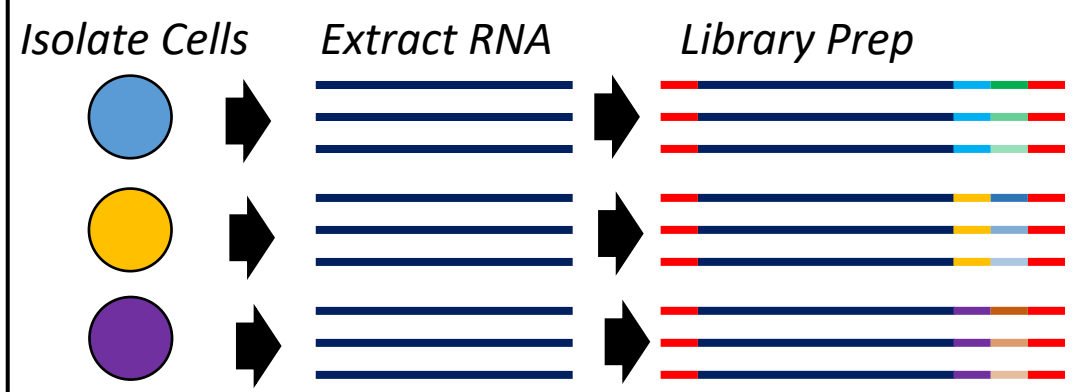
What is RNA-Seq?

Tissue / Cells of Interest

Bulk RNA-Seq (poly-A capture)



scRNA-Seq (e.g. DropSeq)

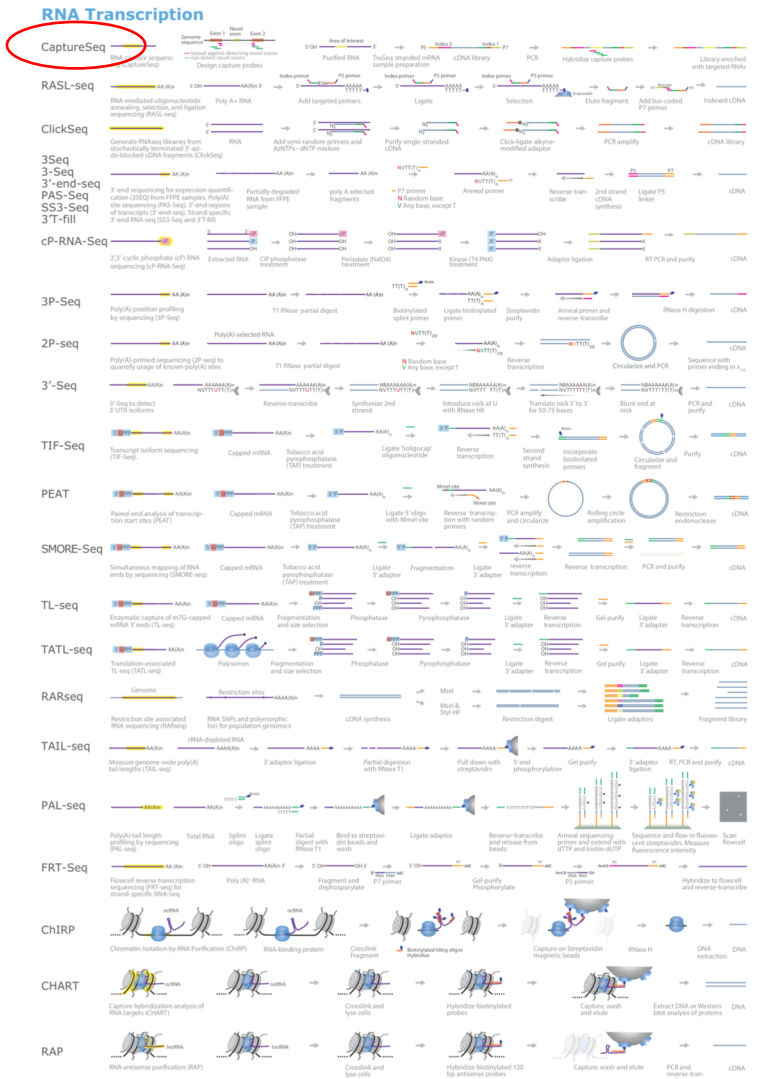


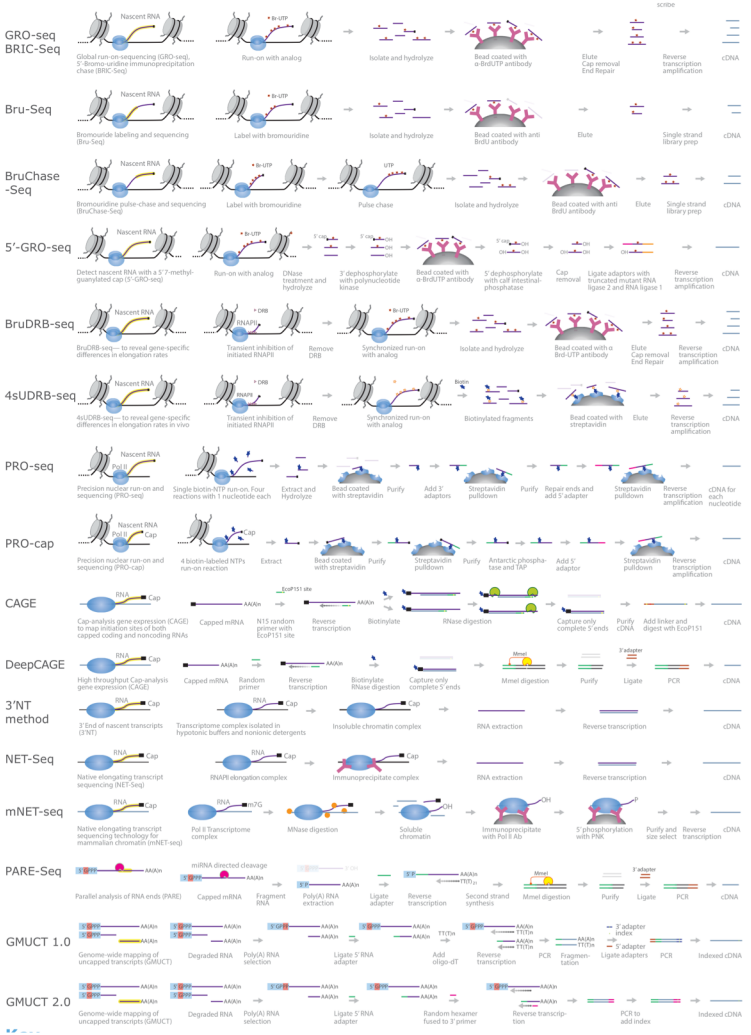
Sequencer

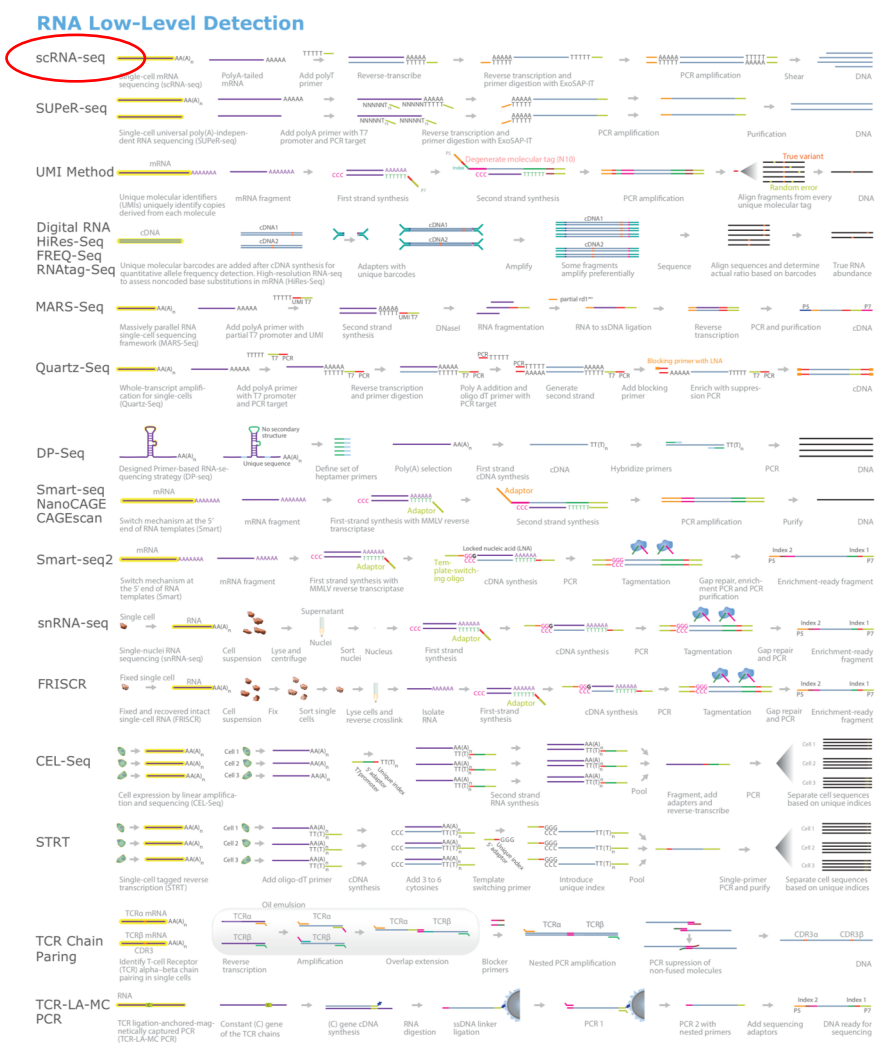
FASTQ

Russell S. Hamilton (rsh46@cam.ac.uk), Xiaohui Zhao (xz289@cam.ac.uk) & Malwina Prater (mn367@cam.ac.uk)

Different flavours of RNA-Seq?







https://emea.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/for-all-you-seq-rna.pdf

Sequencing Technologies

Sequencing By Synthesis

Illumina
Short Read Sequencers 50-300bp

Single Molecule Real Time Sequencing (SMRT)

PacBio
Long Read Sequencers 1-10Kb

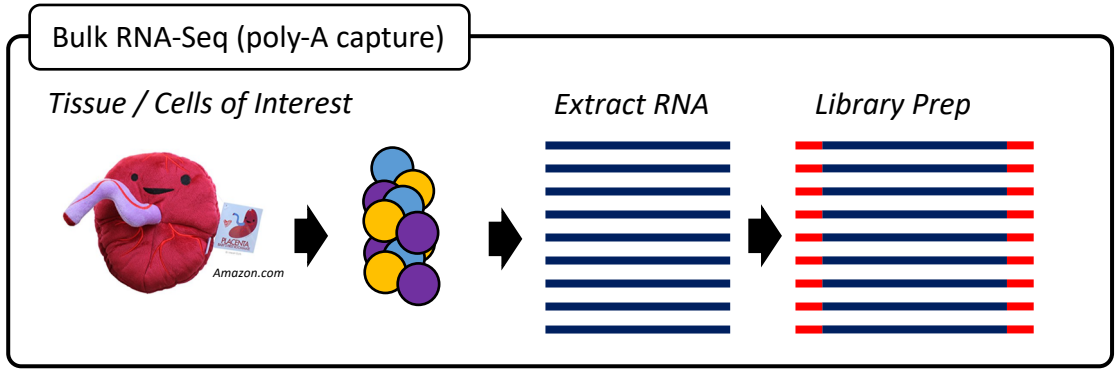
Sequencing through protein pores

Oxford Nanopore Technologies
Long Read Sequencers 1Kb – 1Mb

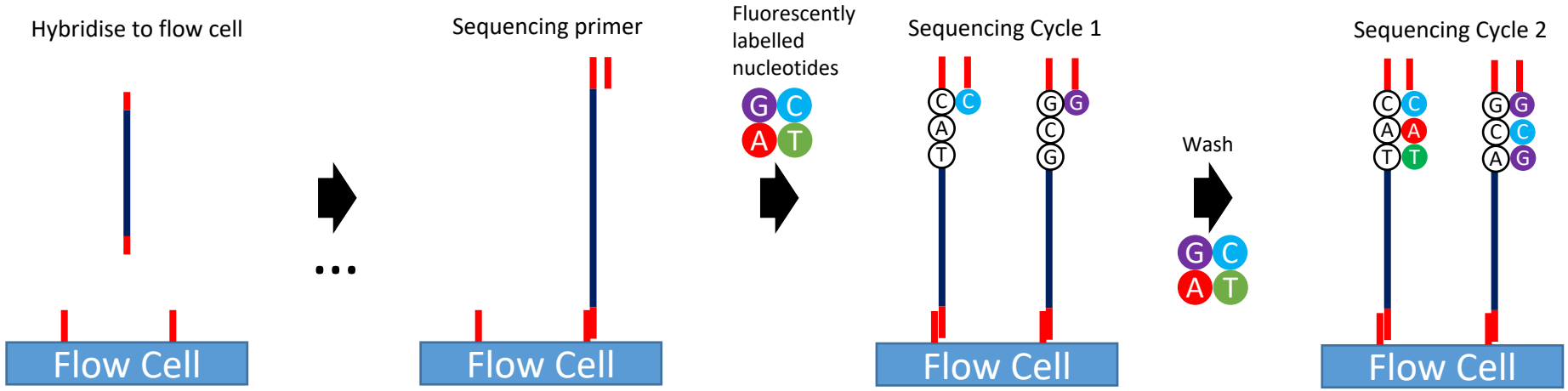
Cumulative Errors

Stochastic Errors

Sequencing By Synthesis

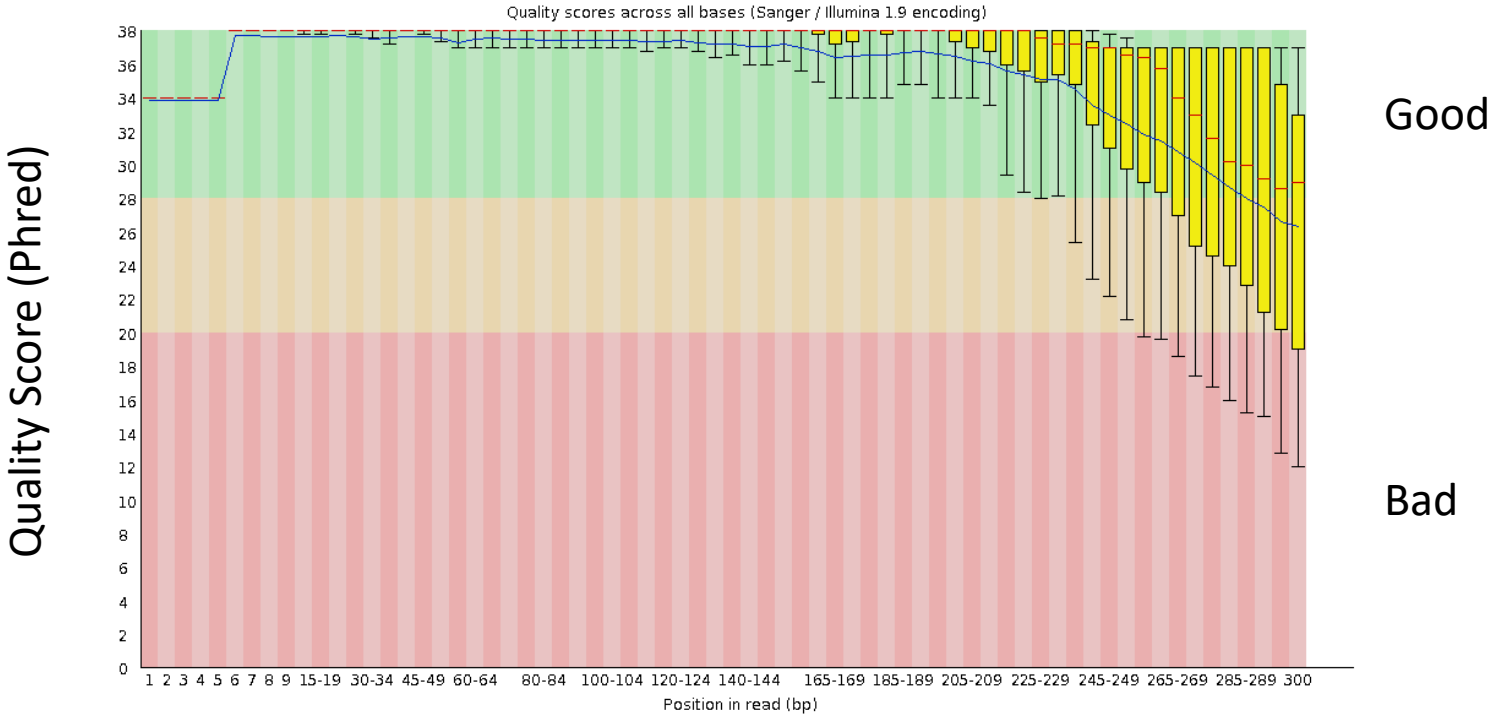


Sequencing By Synthesis (Illumina)



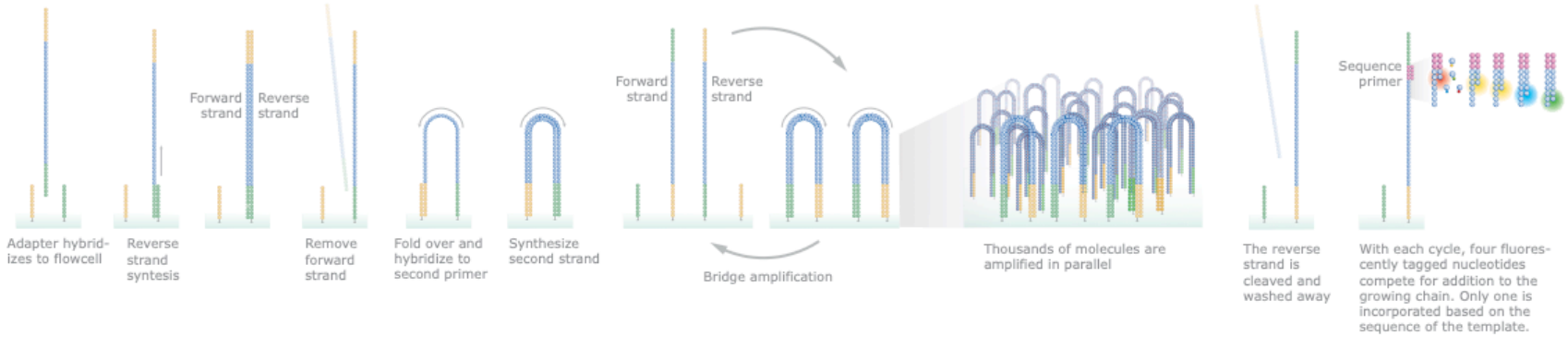
Sequencing Quality

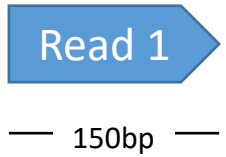
Sequencing by synthesis errors are cumulative due to phasing



Single End Vs Paired End

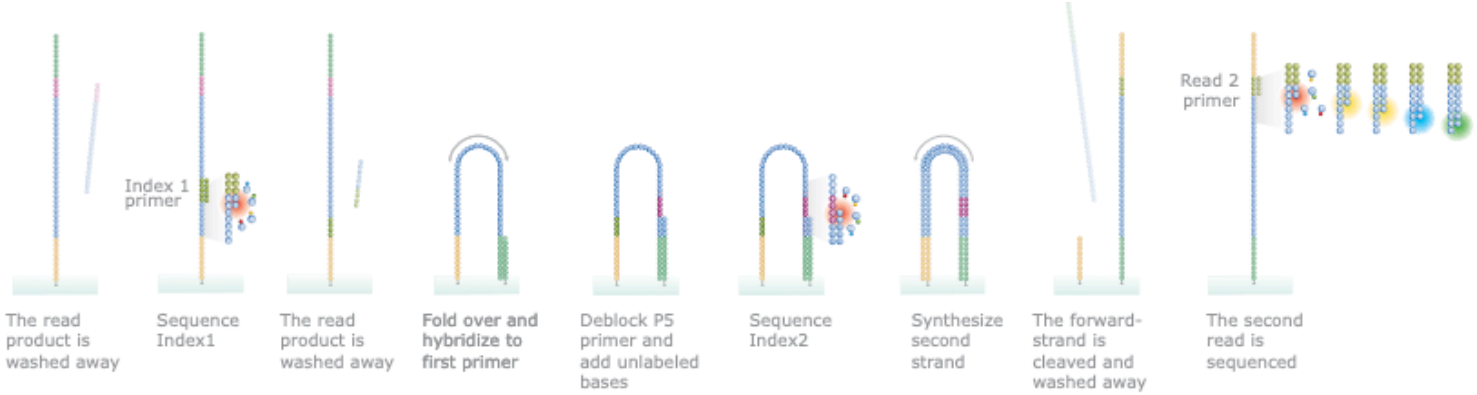
Single End Sequencing

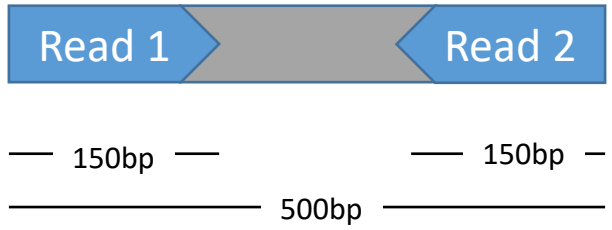




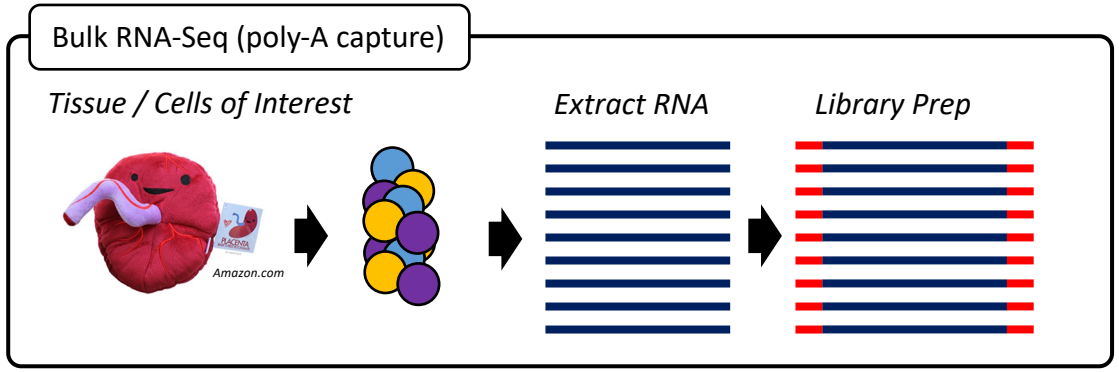
Paired End Sequencing

Starts from SE sequencing method





Sequencing Depth



Library prep includes a sample index, so multiple samples can be sequenced on the same run

Global transcriptome analysis **30-60M** reads
In depth transcriptome analysis **100-200** reads

Sequencer Choice

NextSeq 

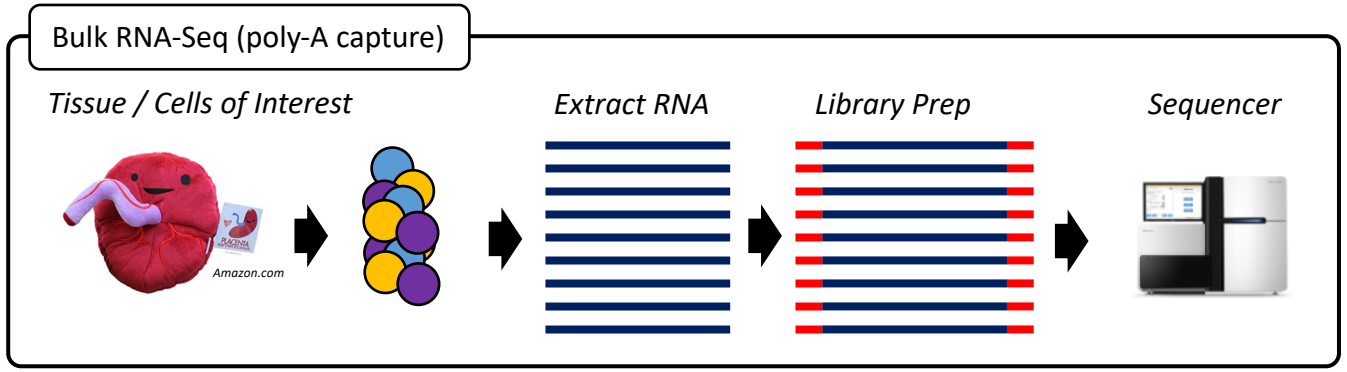
400M Reads Per Run

10 Samples = ~40M reads per sample

20 Samples = ~20M reads per sample

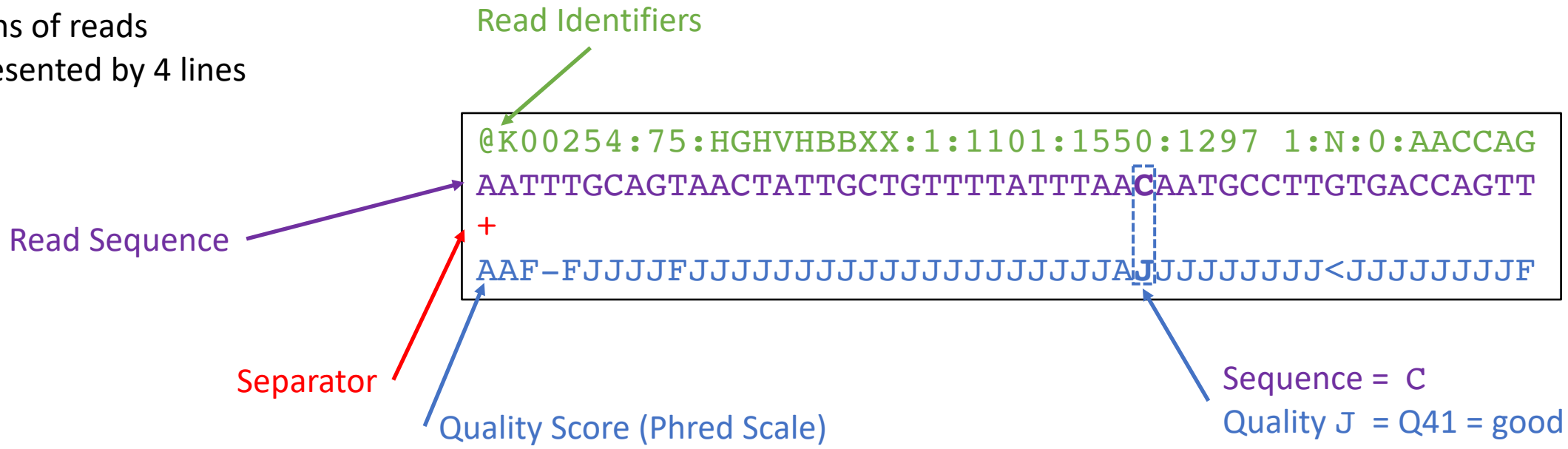
30 Samples = ~13M reads per sample

Sequencer Output



FASTQ File

- Contains millions of reads
- Each read represented by 4 lines



Post Sequencing

FASTQ

```
@K00254:75:HGHVHBBXX:1:1101:1550:1297 1:N:0:AACCAG
AATTTGCAGTAACTATTGCTGTTTTATTTAACAATGCCTTGTGACCAGTT
+
AAF-FJJJJFJJJJJJJJJJJJJJJJJJJJAJJJJJJJJJ<JJJJJJJJF
```

Did the sequencing perform well?

- Read qualities
- GC bias
- Adapter content

Practical Exercises
How to assess quality of sequencing output

No information on where the read came from

- Which gene is it from?
- How many reads for each gene?

How to align reads to a reference genome

Read Alignment to Reference Genome

FASTQ

```
@K00254:75:HGHVHBBXX:1:1101:1550:1297 1:N:0:AACCAG
AATTTGCAGTAACTATTGCTGTTTTATTTAACAATGCCTTGTGACCAGTT
+
AAF-FJJJJFJJJJJJJJJJJJJJJJJJJJAJJJJJJJJJ<JJJJJJJJF
```

Reference Genome

https://www.ensembl.org/Mus_musculus

Option 1: Genome Assembly

Chromosome Sequences

- Mus_musculus.GRCm38.dna.chromosome.10.fa.gz

Gene Annotations (location of genes and exons etc)

- Mus_musculus.GRCm38.84.gtf

Option 2: Transcriptome Assembly

Known Transcript Sequences

- Mus_musculus.GRCm38.cdna.all.fa.gz

Read Alignment to Reference Genome

Option 1: Genome Alignment using e.g. STAR

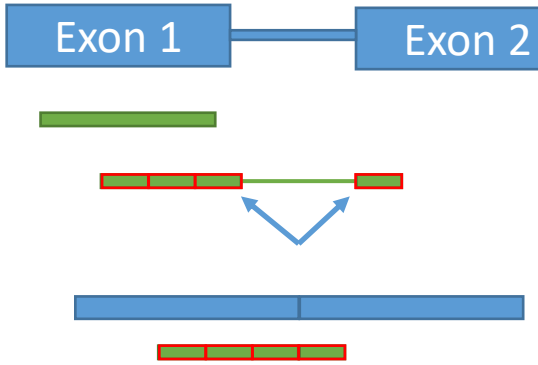
1. Align reads to reference genome
   - Get location of match (chr, start, end)
   - Check if alignment is unique

2. Does the read lie within a gene?
3. Does the read lie within or span an exon?

	STAR
Run time	hours
Hardware requirements	Multi-core
Novel Splice Sites	yes

Chromosome Sequence Files (FASTA file)

Gene and exon locations (GTF File)



Single exon mapping

Segments aligned and assembled

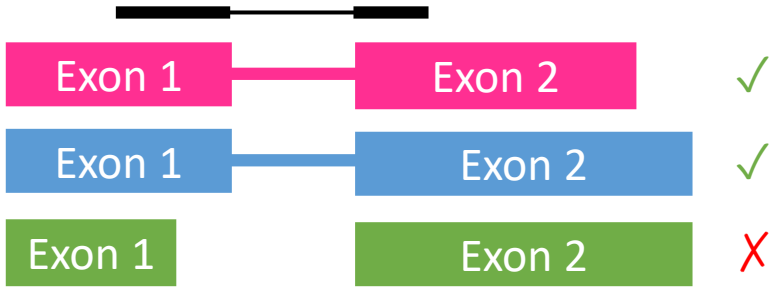
Read Alignment to Reference Genome

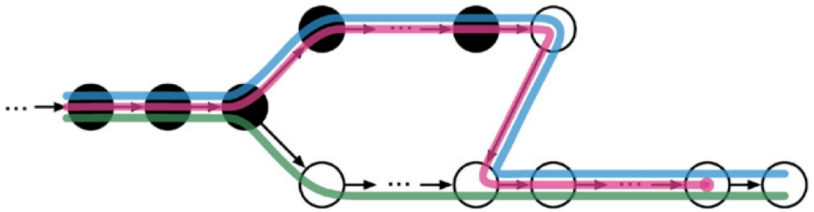
Option 2: Transcriptome Pseudo-Alignment using Kallisto

Does the reads pseudo-align to a transcript – and which isoform is most likely?

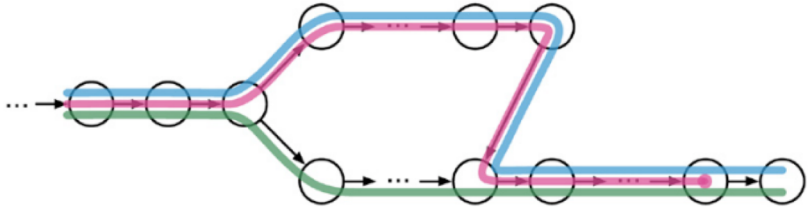
	Kallisto
Run time	minutes
Hardware requirements	Laptop
Novel Splice Sites	no

Reference Transcriptome Split into Kmers

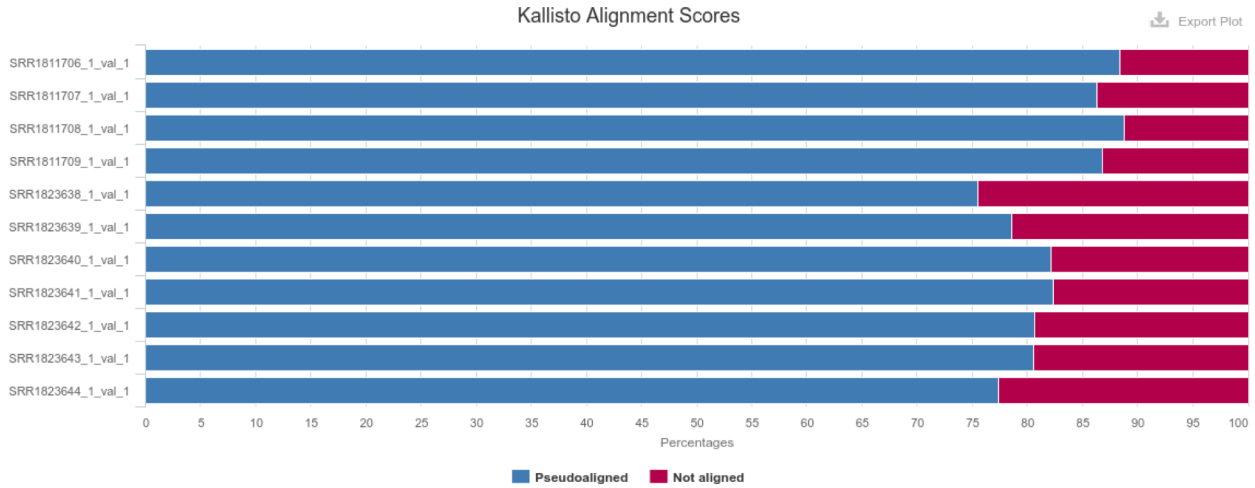




Encoded into a de Bruijn Graph



QC Alignments



Why do you never see 100% alignment?

- Incomplete reference genomes / transcriptomes
- Repetitive reads hard to map uniquely
- Sample: Structural Variants
  Copy Number Variants

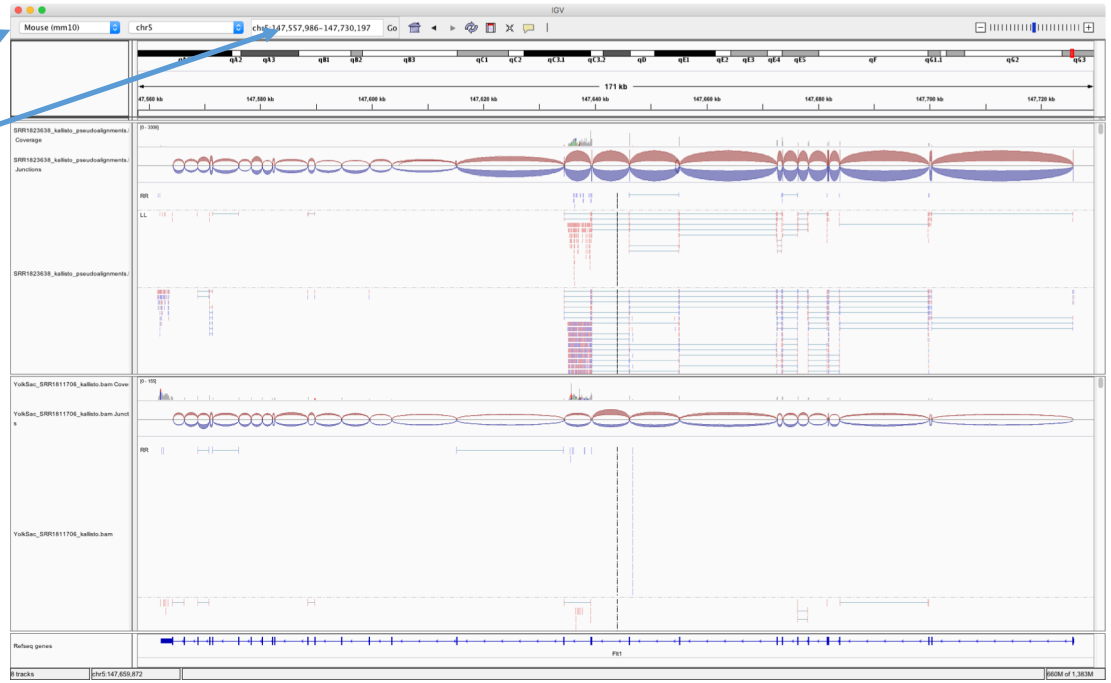
Visualising the Alignments

Have all the computational tools worked as expected?

Lots of genome viewers available
IGV (Integrated Genomics Viewer)
https://software.broadinstitute.org/software/igv/

Select Reference Genome: e.g. Mouse(mm10)

Search for gene of interest



Can you see the difference in coverage between samples?
Do the reads line up with exons?
(For mRNA reads should align to exons only)

(Not part of todays practical, but an essential part of any sequencing based analysis)

Dr Russell S. Hamilton Email: rsh46@cam.ac.uk

Dr Malwina Prater Email: mn367@cam.ac.uk

Dr Xiaohui Zhao Email: xz289@cam.ac.uk

Course Materials:
https://github.com/CTR-BFX/2019-PlacentalBiologyCourse