Oxford Genomics CENTRE

UNIVERSITY OF OXFORD

RNA-Seq Data Analysis
26-27th November, 2018

Organised and delivered by Bioinformatics Core at WHG:

Helen Lockstone M.Sc.

Ben Wright PhD

Eshita Sharma PhD

Santiago Revale M.Sc.

wellcome centre human genetics

Oxford Genomics CENTRE

Oxford Genomics Centre
The Wellcome Trust Centre for Human Genetics

Sequencing overview

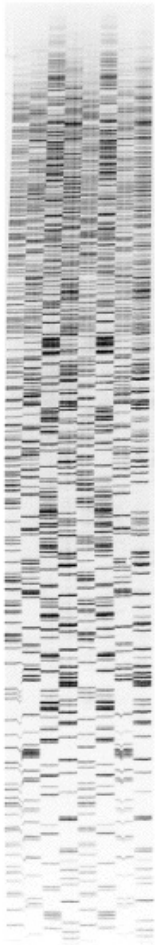
Santiago Revale, M.Sc.
Senior Bioinformatics Analyst, Bioinformatics Core Group

whg

UNIVERSITY OF OXFORD

Genomics Platforms

Oxford Genomics CENTRE



wellcome centre human genetics

UNIVERSITY OF OXFORD

Sequencers Comparison

Oxford Genomics CENTRE

	Illumina	Oxford Nanopore
Method	Sequencing by synthesis	Nanopore Sequencing
Read length	MiniSeq, NextSeq: 75-150 bp; MiSeq: 50-300 bp; HiSeq 2500: 50-250 bp; HiSeq 3/4000: 50-150 bp; HiSeq X: 150 bp	Dependent on library prep, not the device, so user chooses read length. (up to 500 kb reported)
Accuracy (single read not consensus)	99.9% (Phred30)	~92–97% single read
Reads per run	MiniSeq/MiSeq: 1-25 Million; NextSeq: 130-00 Million, HiSeq 2500: 300 million - 2 billion, HiSeq 3/4000 2.5 billion, HiSeq X: 3 billion	Dependent on read length selected by user
Time per run	1 to 11 days, depending upon sequencer and specified read length.	Data streamed in real time. Choose 1 min to 48 hrs
Cost per 1 million bases (in US$)	$0.05 to $0.15	$500–999 per Flow Cell, base cost dependent on experiment.
Advantages	Potential for high sequence yield, depending upon sequencer model and desired application.	Longest individual reads. Accessible user community. Portable (Palm sized).
Disadvantages	Equipment can be very expensive. Requires high concentrations of DNA.	Lower throughput than other machines, Single read accuracy in 90s.

Sequencing Outputs

Oxford Genomics CENTRE

Illumina HiSeq 4000

Output range	105 – 1500 Gb	
Reads per run	2.1 – 5 billion	
Max. read length	2 x 150 bp	
Run time	< 1 – 3.5 days	
Samples sequenced per:	Flowcell	Lane
polyA	80	10
Ribodepleted	40	5
3’ mRNA	384	48
CHIPseq	80	10

Illumina HiSeq 2500

Output range	9 – 1000 Gb	
Reads per run	0.3 – 4 billion	
Max. read length	2 x 250 bp	
Run time	< 1 – 6 days	
Samples sequenced per:	Run	Lane
Small RNA	168	21

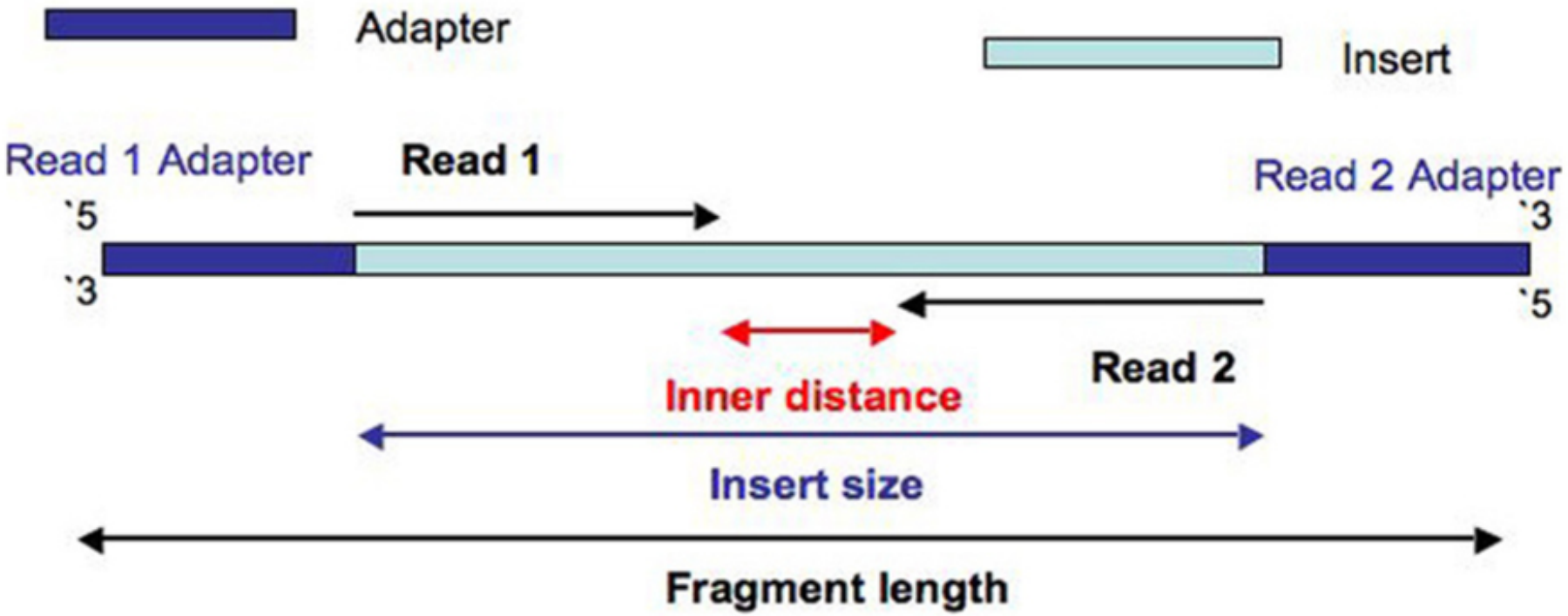
wellcome centre human genetics

UNIVERSITY OF OXFORD

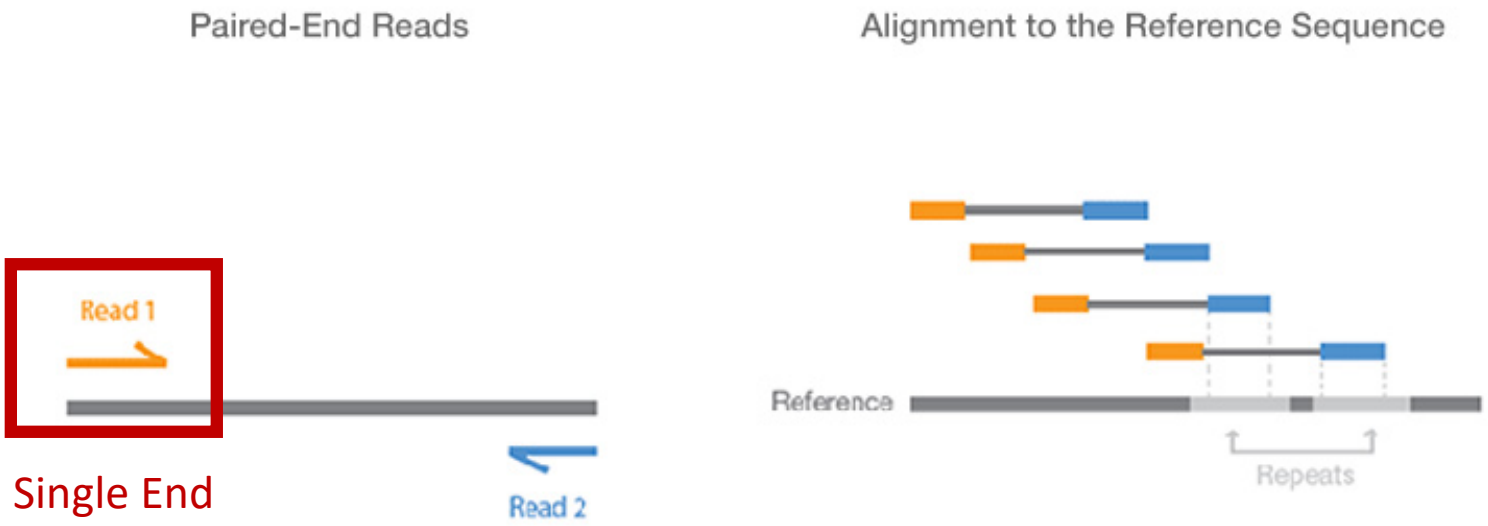
Sequencing By Synthesis (Illumina)

https://youtu.be/womKfikWlxM

Fragment/Insert Size



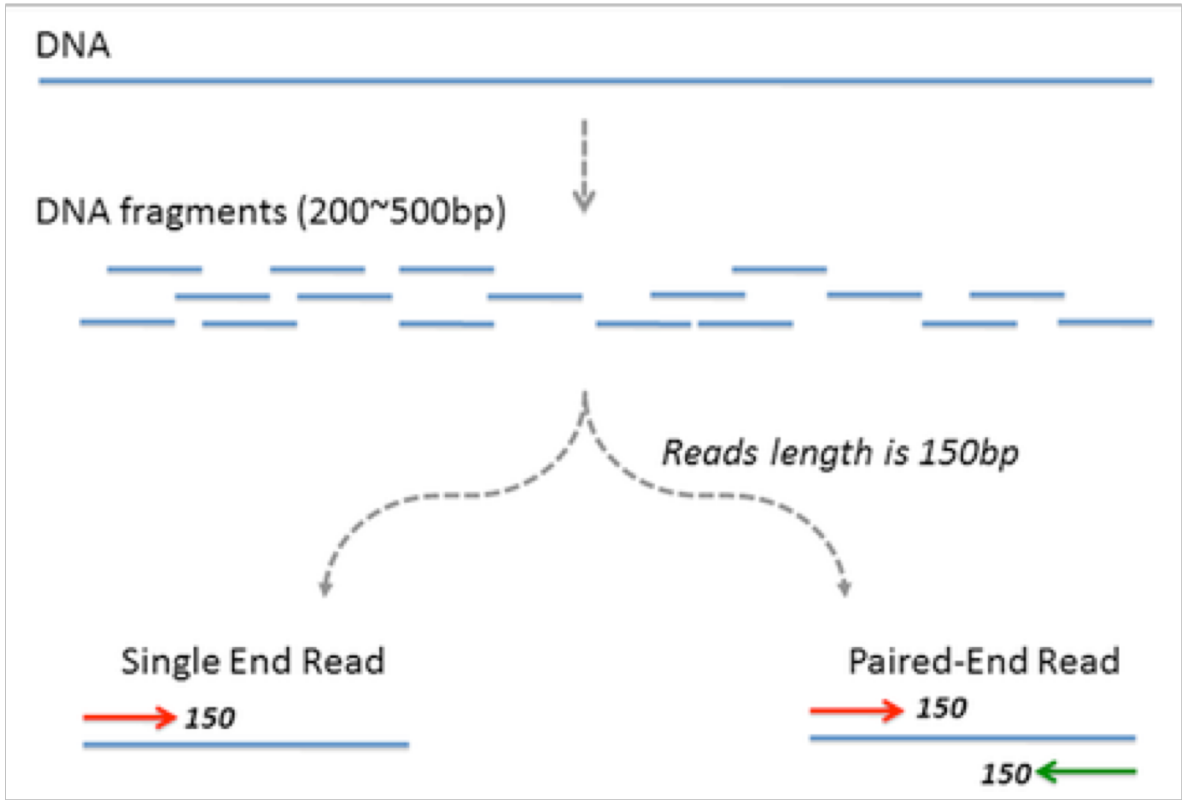
Paired End / Single End



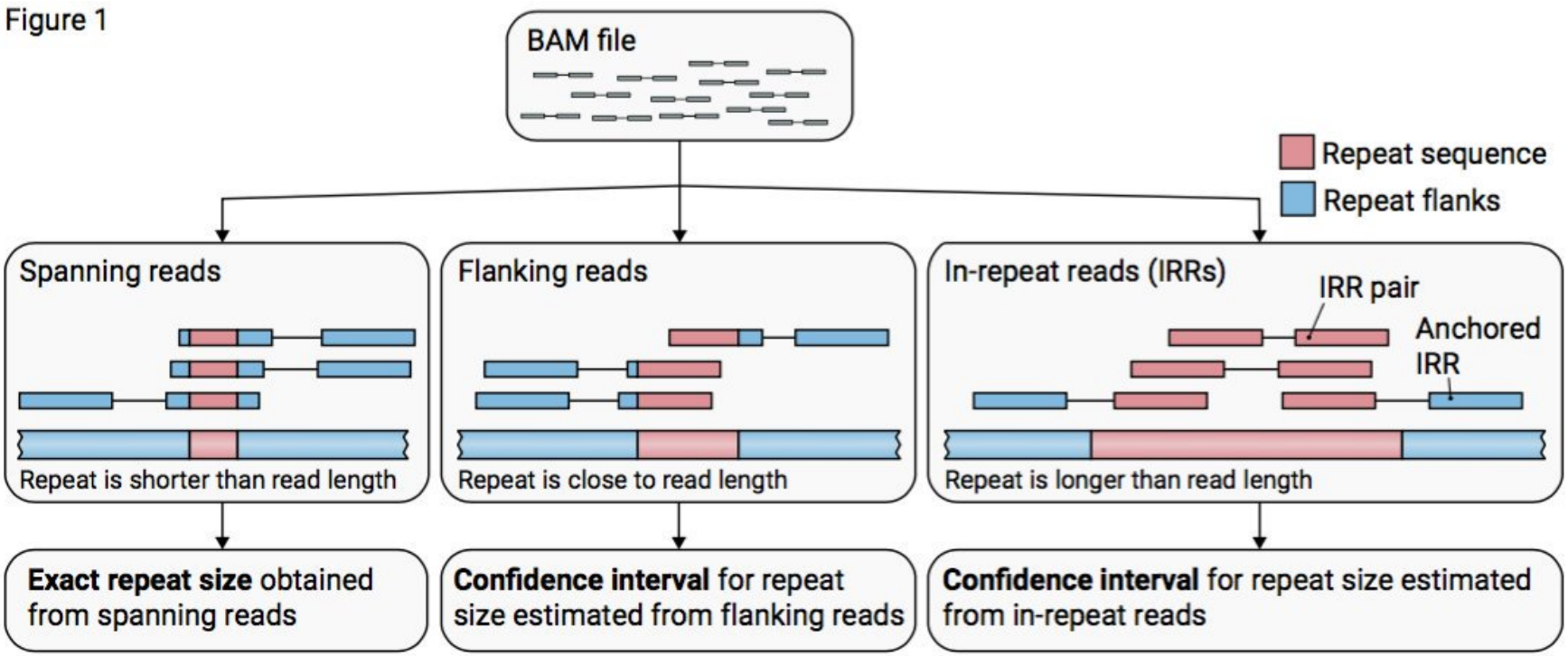
Single End

Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Read Length

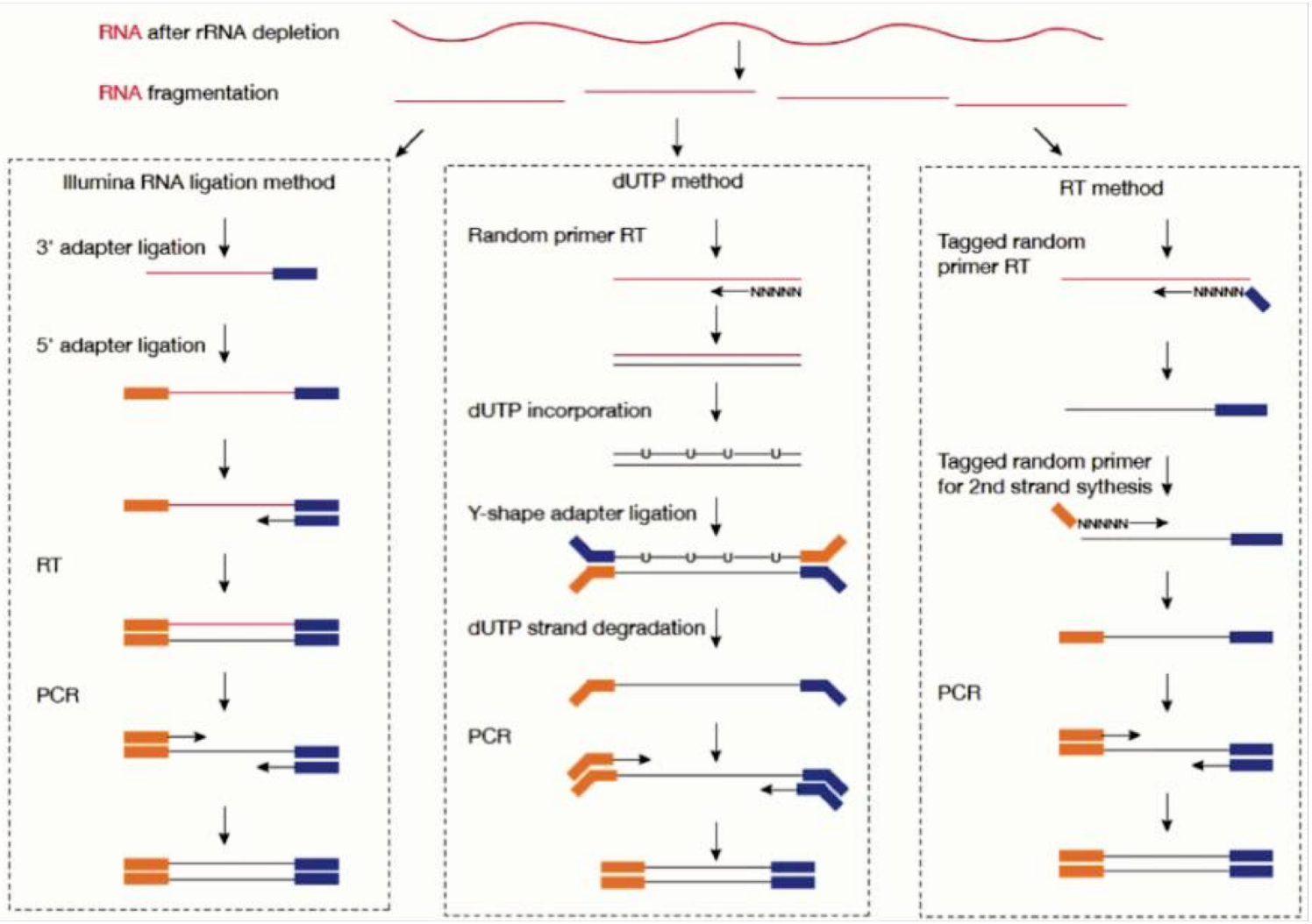

DNA
DNA fragments (200~500bp)
Reads length is 150bp
Single End Read
150
Paired-End Read
150
150

Read Length

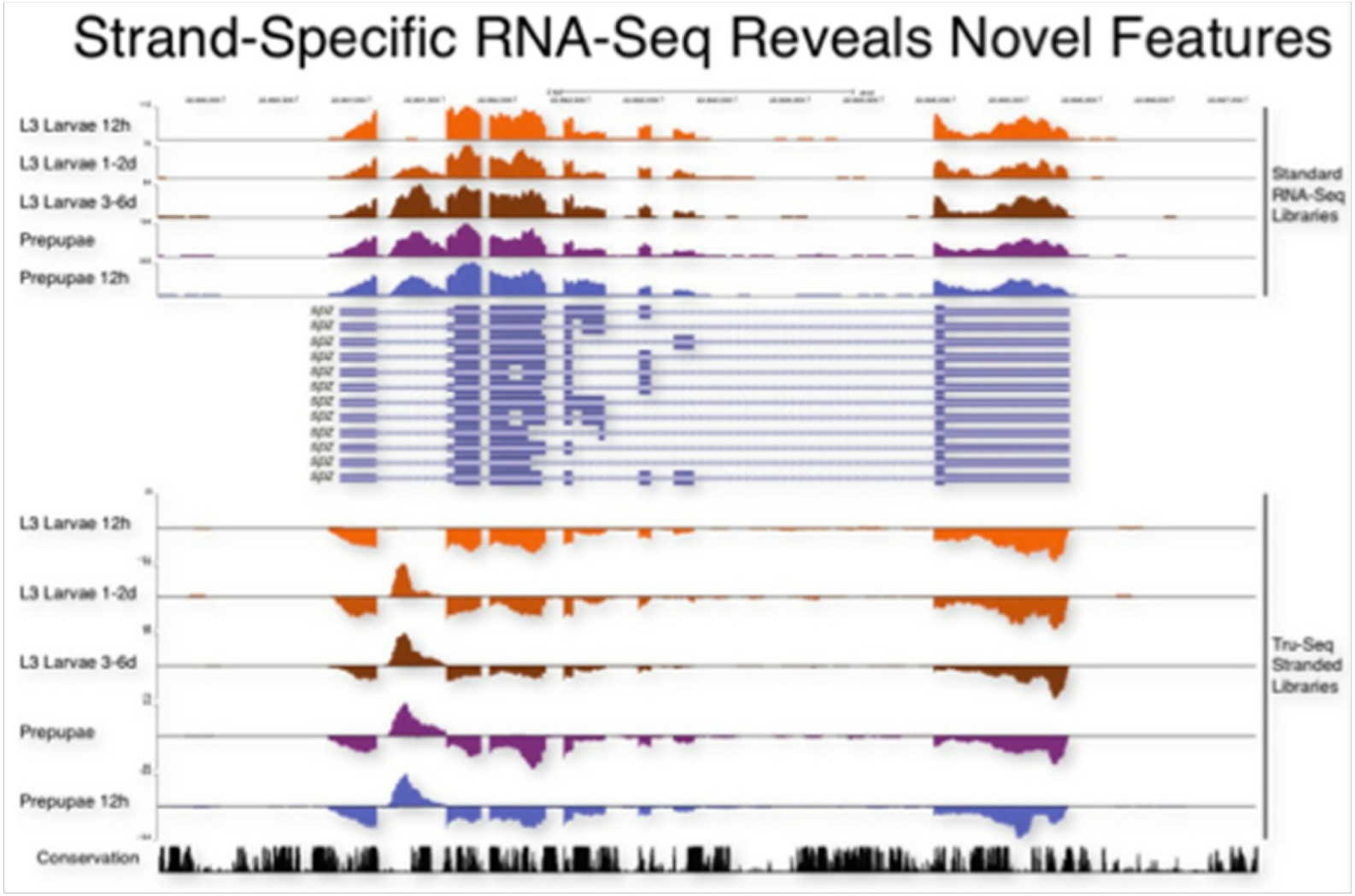


The better the read length, the higher the probability of resolving repeats or finding splice junctions.

Stranded / Unstranded

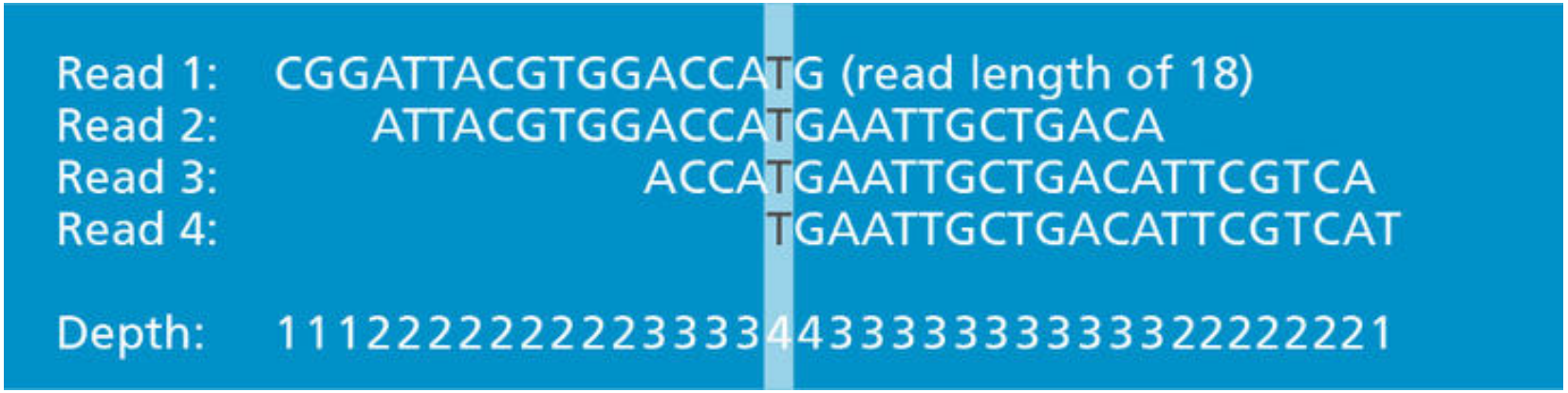


Stranded / Unstranded



Read Depth / Coverage

Number of unique reads that include a given nucleotide in the reconstructed sequence.



Deep sequencing refers to the general concept of aiming for high number of unique reads of each region of a sequence.

Oxford Genomics Centre

Oxford Genomics Centre
The Wellcome Trust Centre for Human Genetics

Data processing and evaluation

Santiago Revale, M.Sc.
Senior Bioinformatics Analyst, Bioinformatics Core Group

whg

UNIVERSITY OF OXFORD

Inspecting raw data

Biological Sequence Data Format

FastA

```
>SeqID HEADER
TAATTTGGTAACGGCTGATGGTGGACCGCA
AGAAGGTTATCCATATCGTG
```

It only contains sequence information.

Qual

```
>SeqID HEADER
33 33 33 37 37 37 37 37 37 37 37 37 37 40 40 40 40 37 37 40
40 33 37 37 40 40 40 37 40 40 40 40 40 40 37 40 37 37 37 40
40 37 40 37 33 06 15 27 15 22
```

It only contains quality information.
Heavy file: 3 bytes / base.

FastQ

```
@SeqID HEADER
TAATTTGGTAACGGCTGATGGTGGACCGCA
AGAAGGTTATCCATATCGTG
+
BBBFFFFFFFFFFIIIIFFIIBFFIIIFII
IIIIFIFFFIIFIFB'0<07
```

Contains both sequence *and* quality information.
Quality: 1 byte / base.

ASCII values: 33 to 126 → Quality values: 0 to 93.

Phred Quality

$$Q_{\text{phred}} = -10 \log_{10} e$$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy	ASCII	Character
20	1 in 100	99%	53	5
30	1 in 1,000	99.9%	63	?
40	1 in 10,000	99.99%	73	I

Oxford Genomics CENTRE

Quality Control

FastQC

Widely used for Illumina data because it's fast. It works on a subset of reads.

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

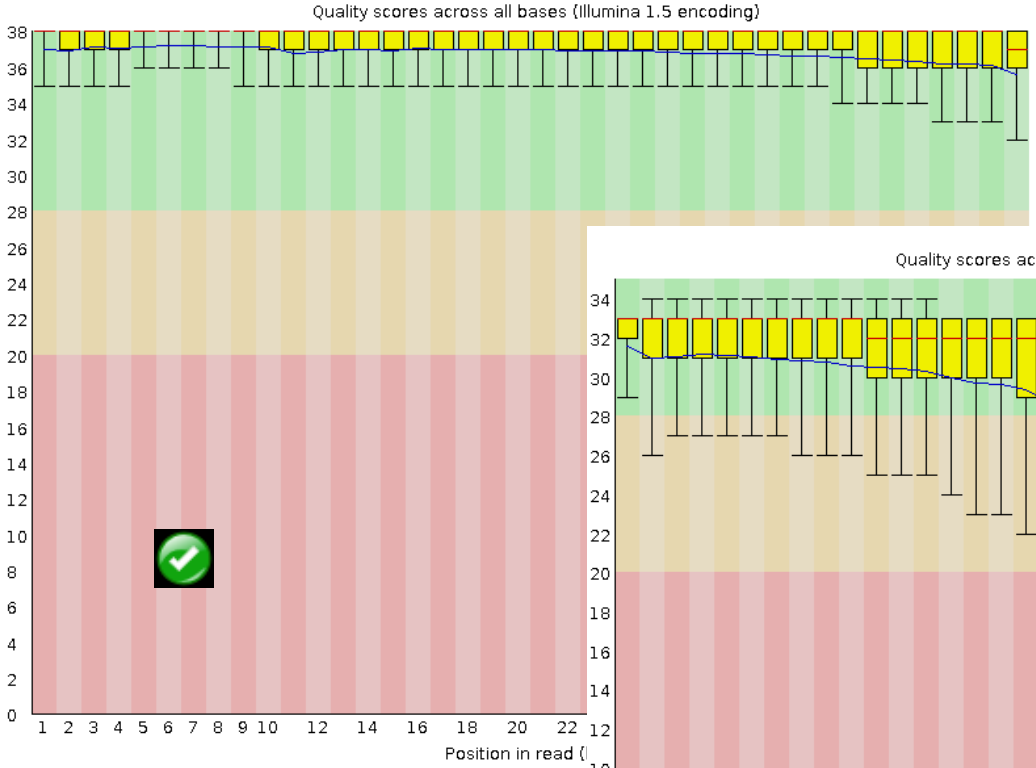
PrinseQ

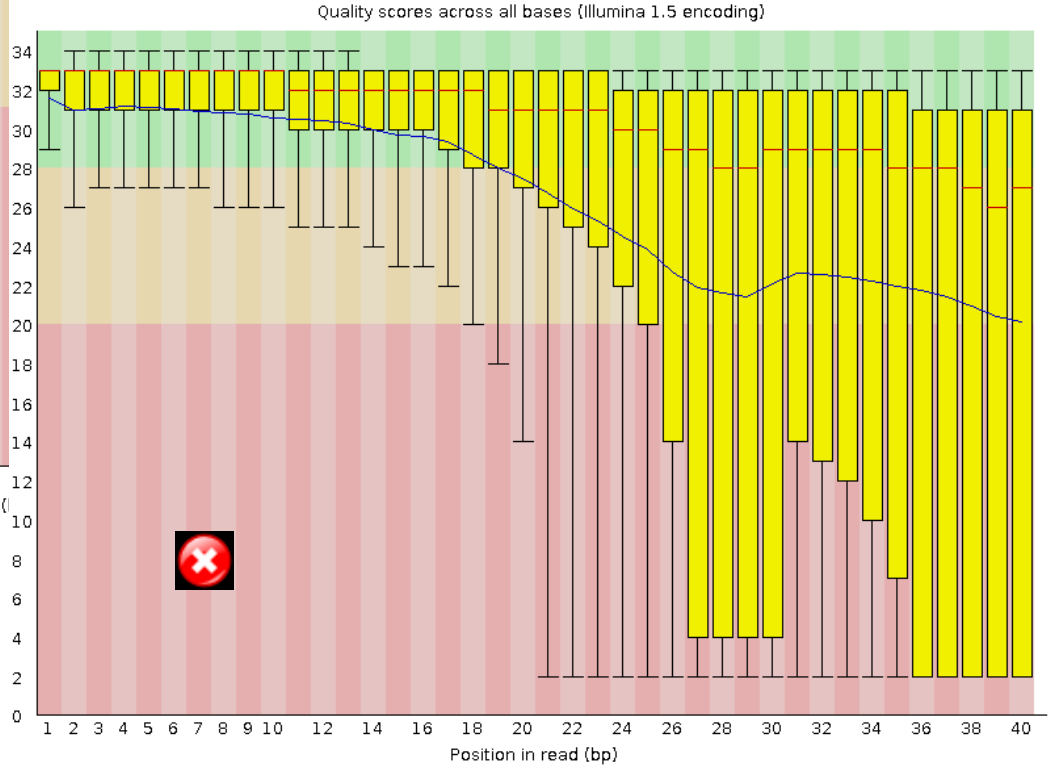
Used for smaller datasets because it computes every sequence.

http://prinseq.sourceforge.net/

Rawdata Quality Control

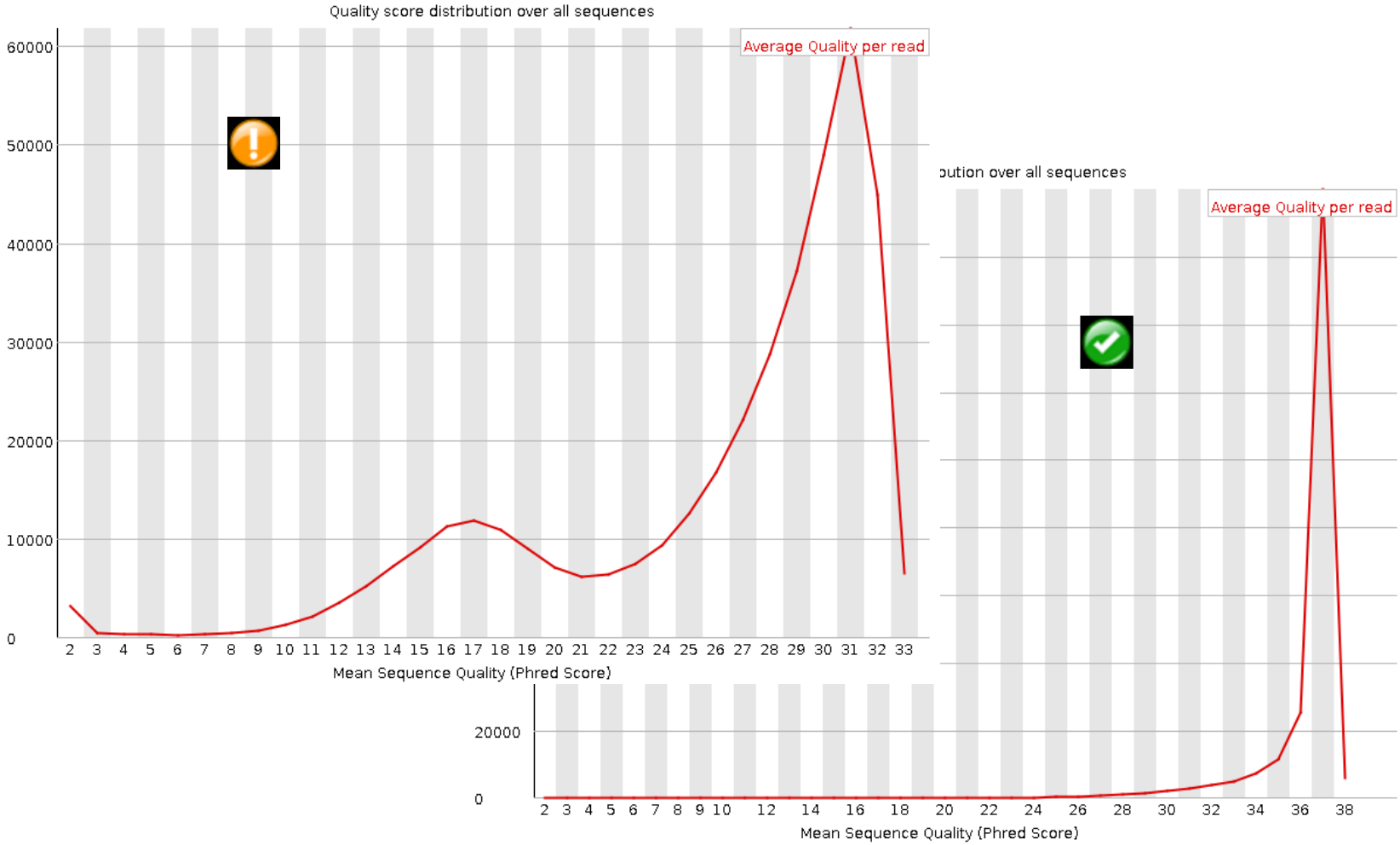
Per base sequence quality





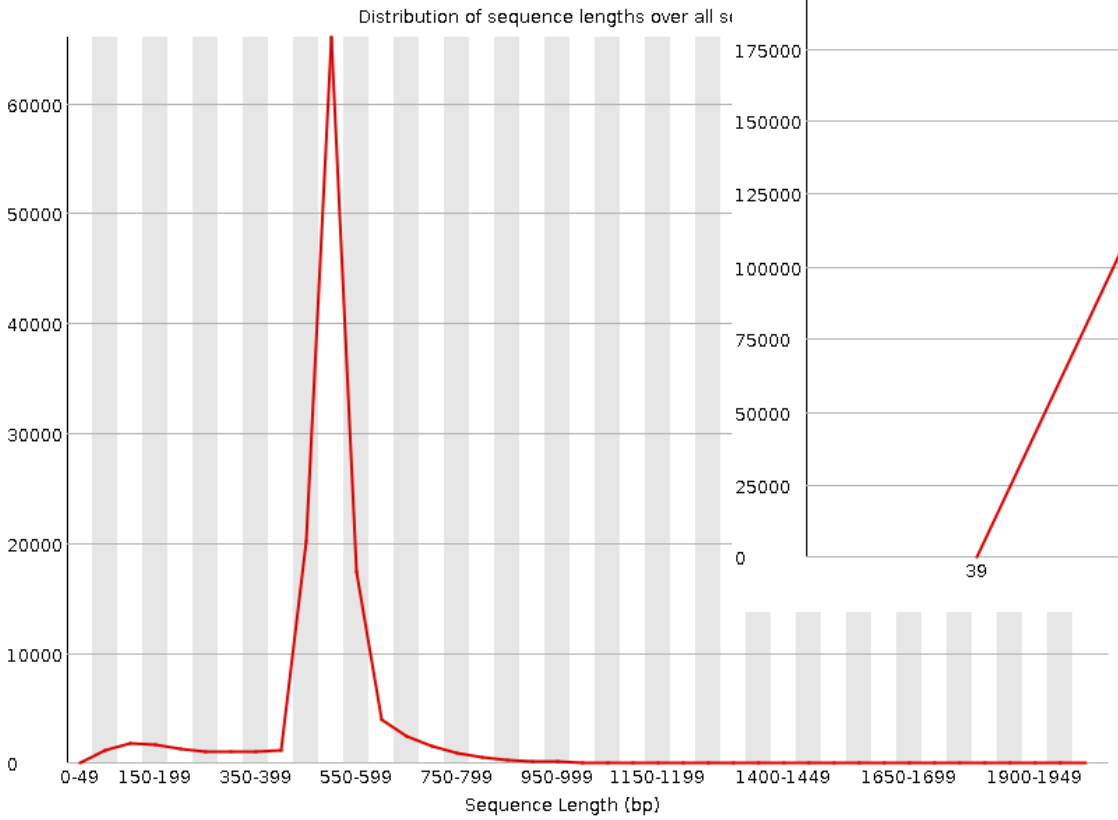
Rawdata Quality Control

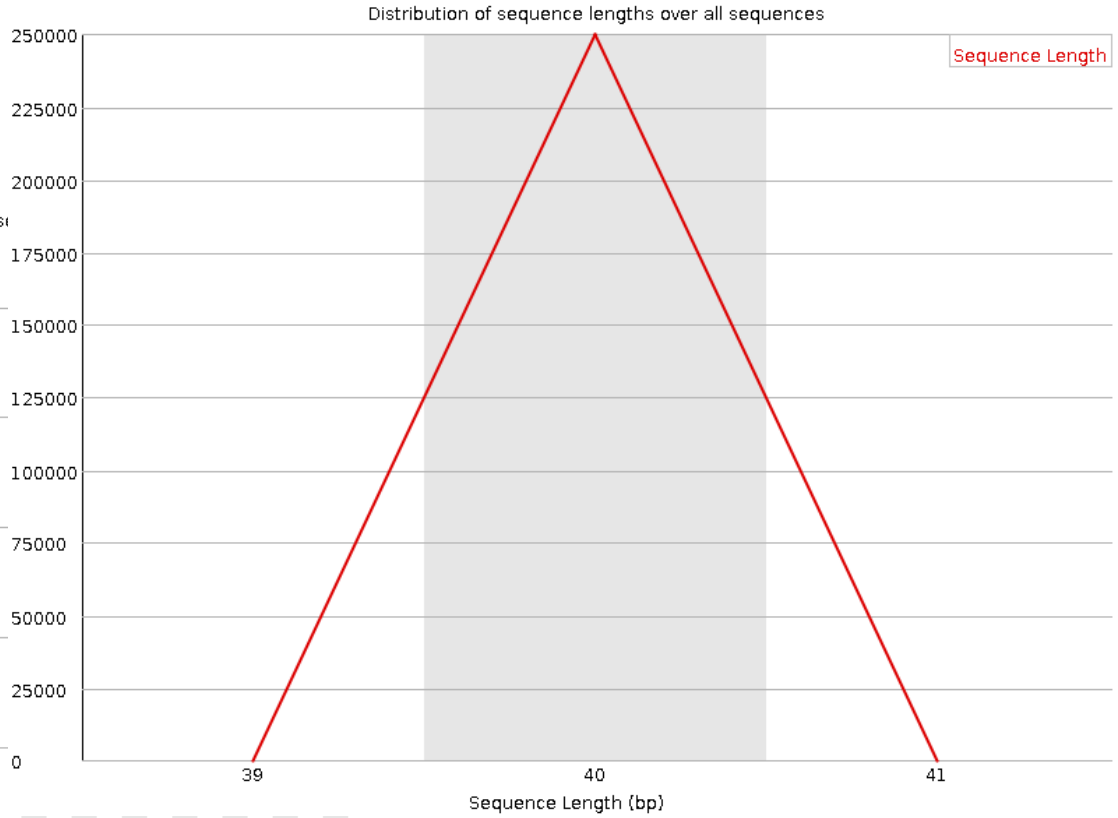
Per sequence quality scores



Rawdata Quality Control
Sequence length distribution

E.g. 454/Roche





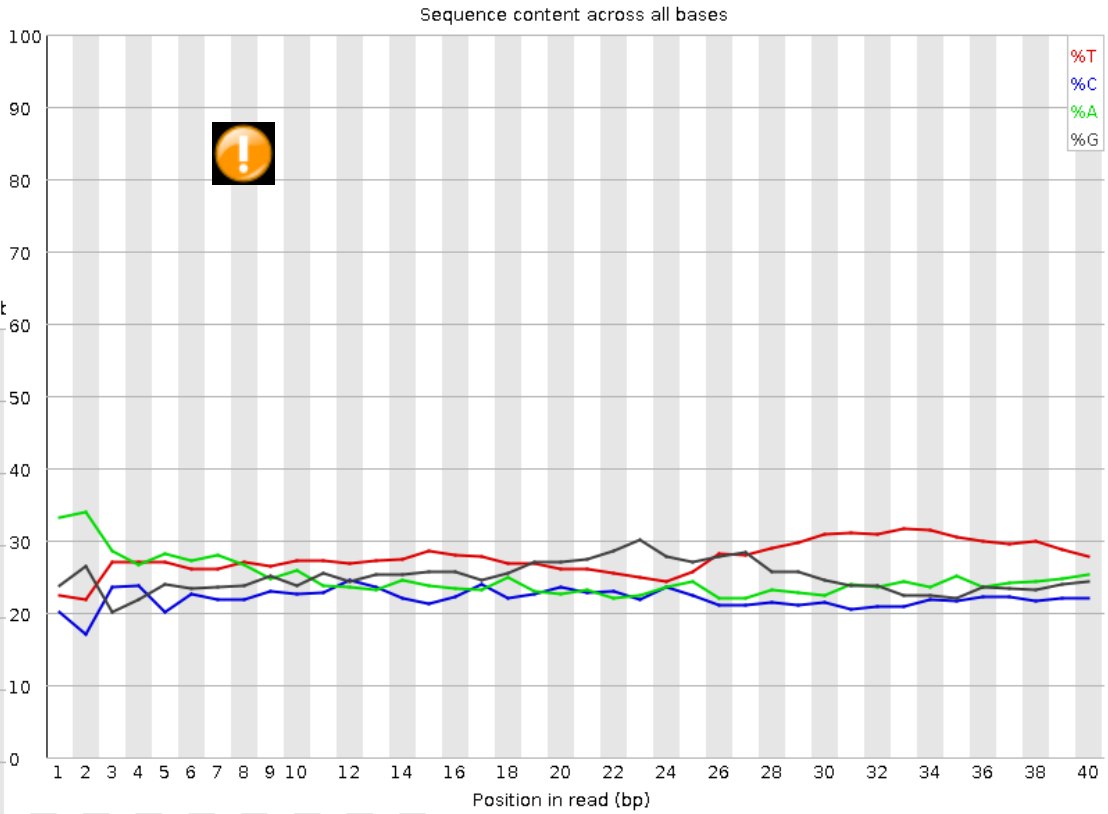
E.g. Illumina

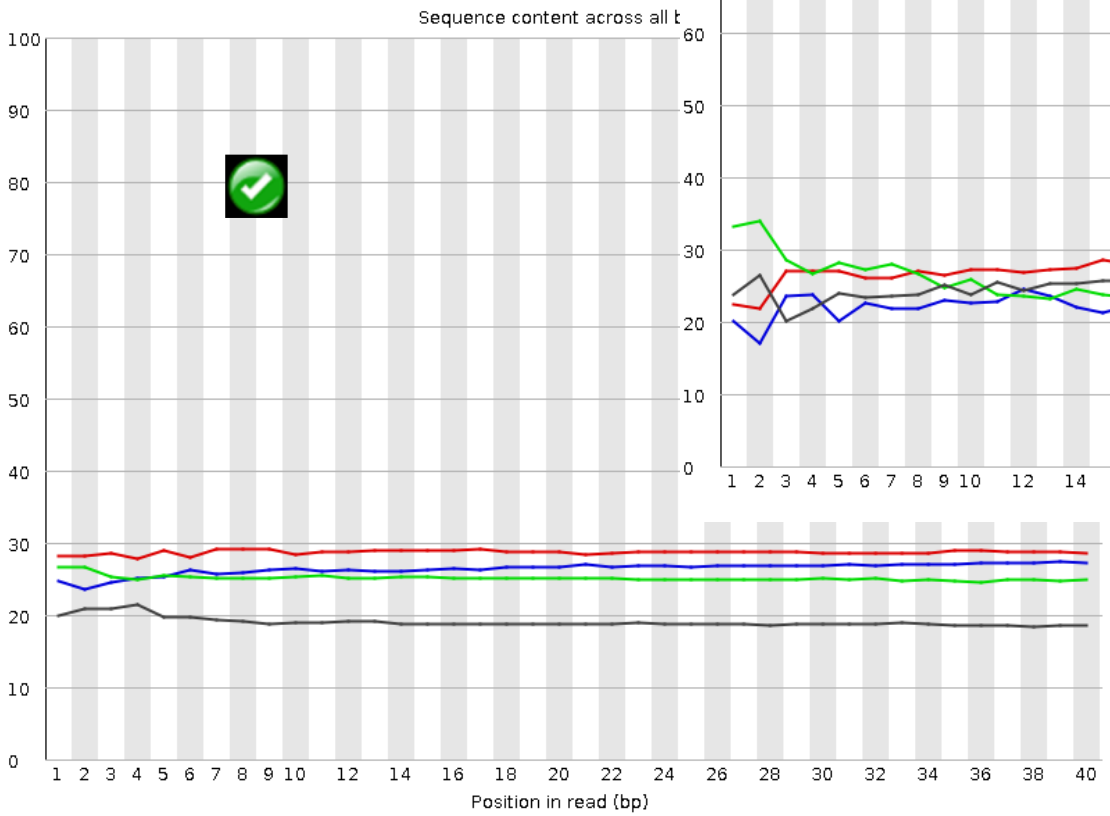
wellcome trust

UNIVERSITY OF OXFORD

Rawdata Quality Control

Per base sequence content

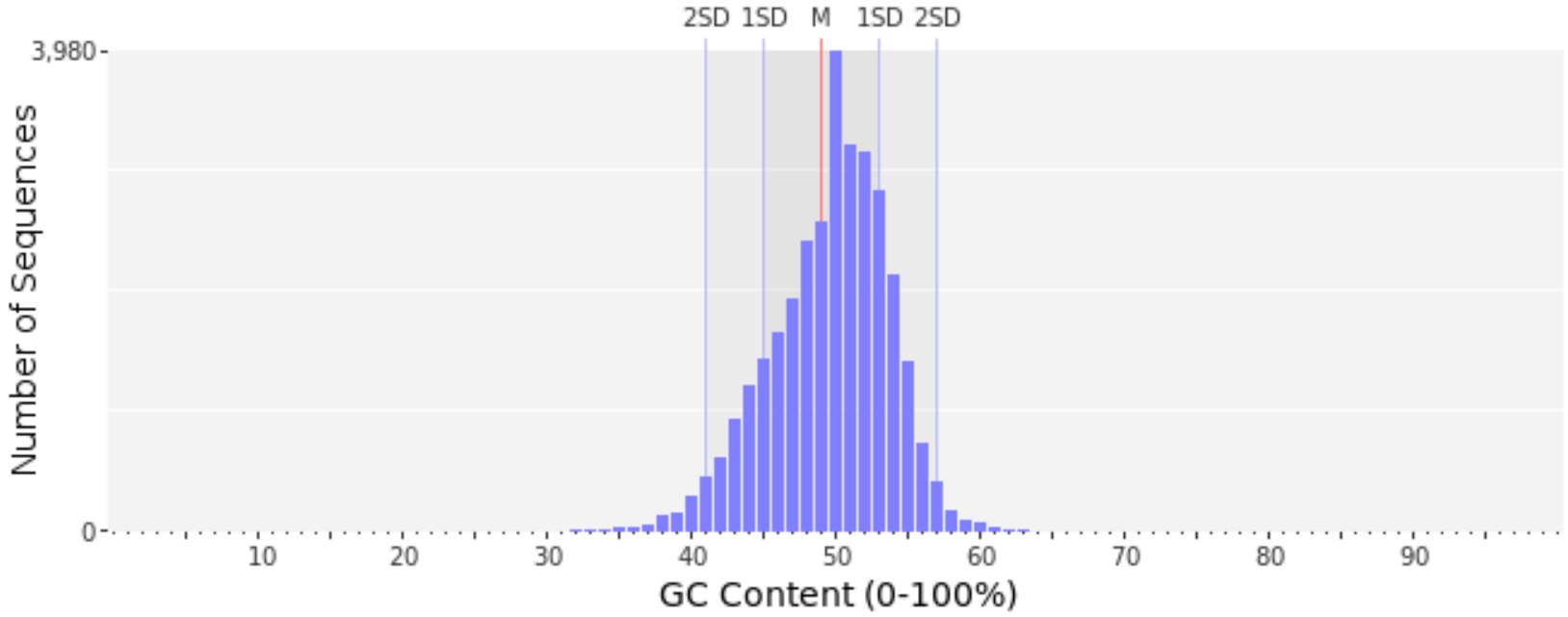




Rawdata Quality Control

GC content distribution

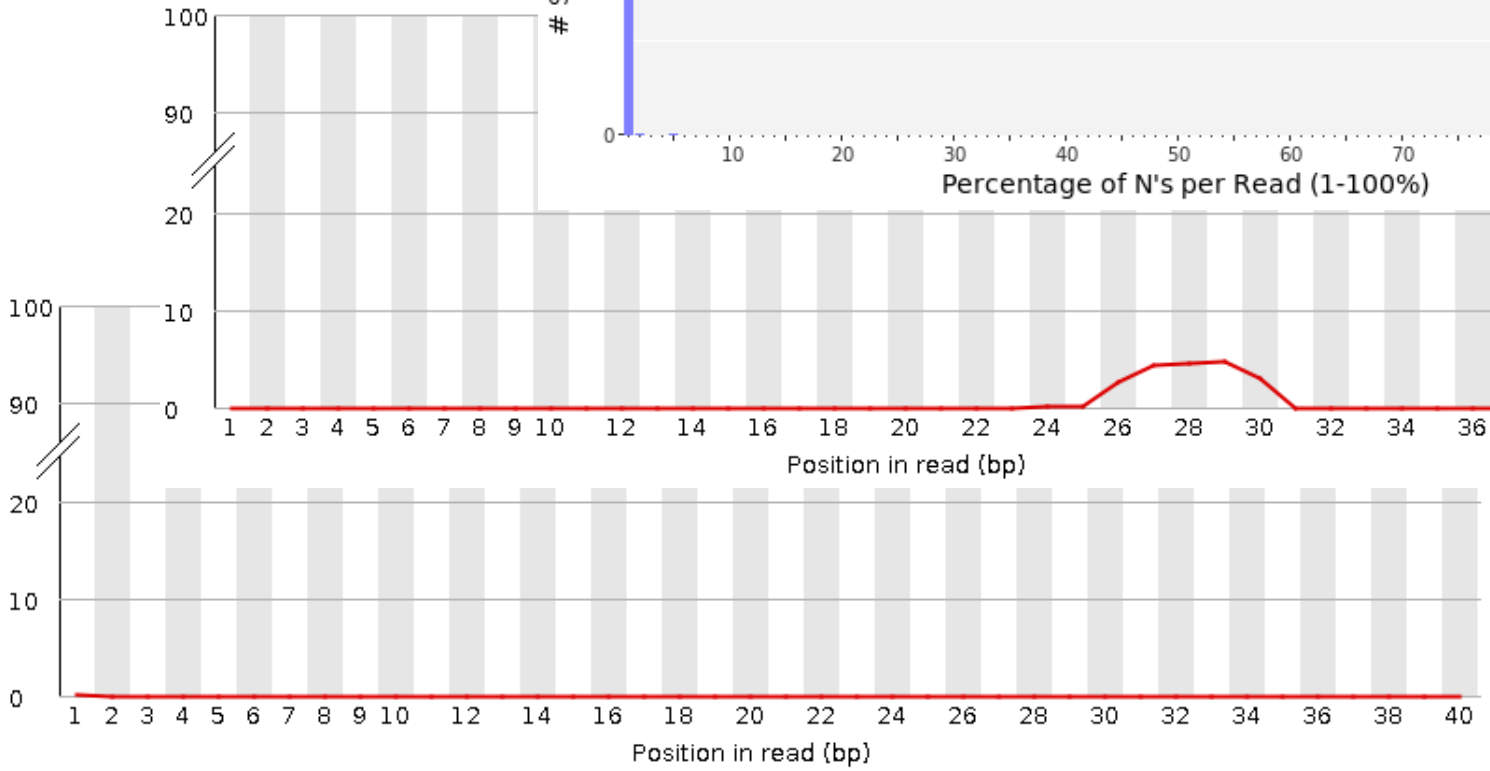
Mean GC content:	**49.55 ± 4.21 %**
Minimum GC content:	**20 %**
Maximum GC content:	**69 %**
GC content range:	**50 %**
Mode GC content:	**50 % with 3,977 sequences**



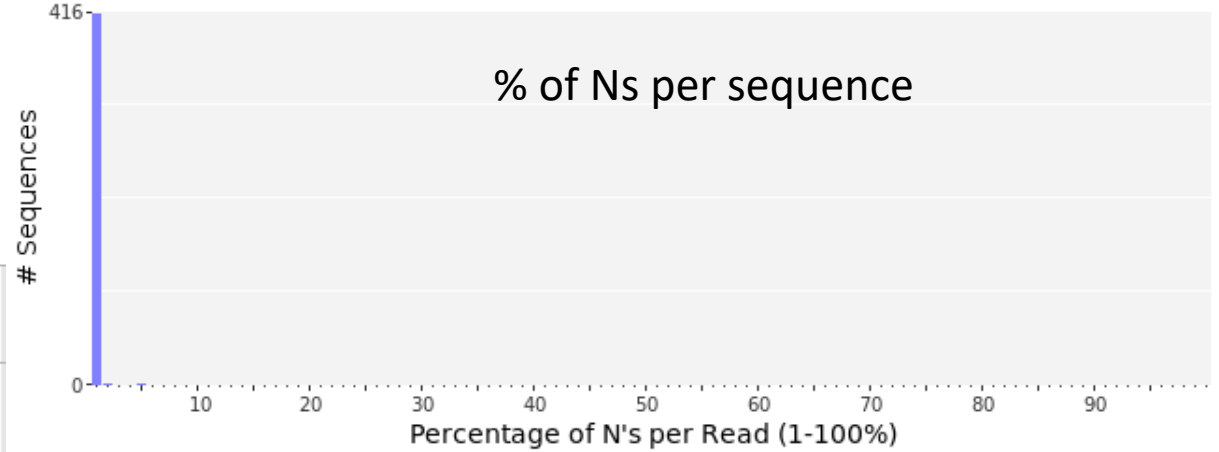
Rawdata Quality Control

N base content

Ns by sequence position

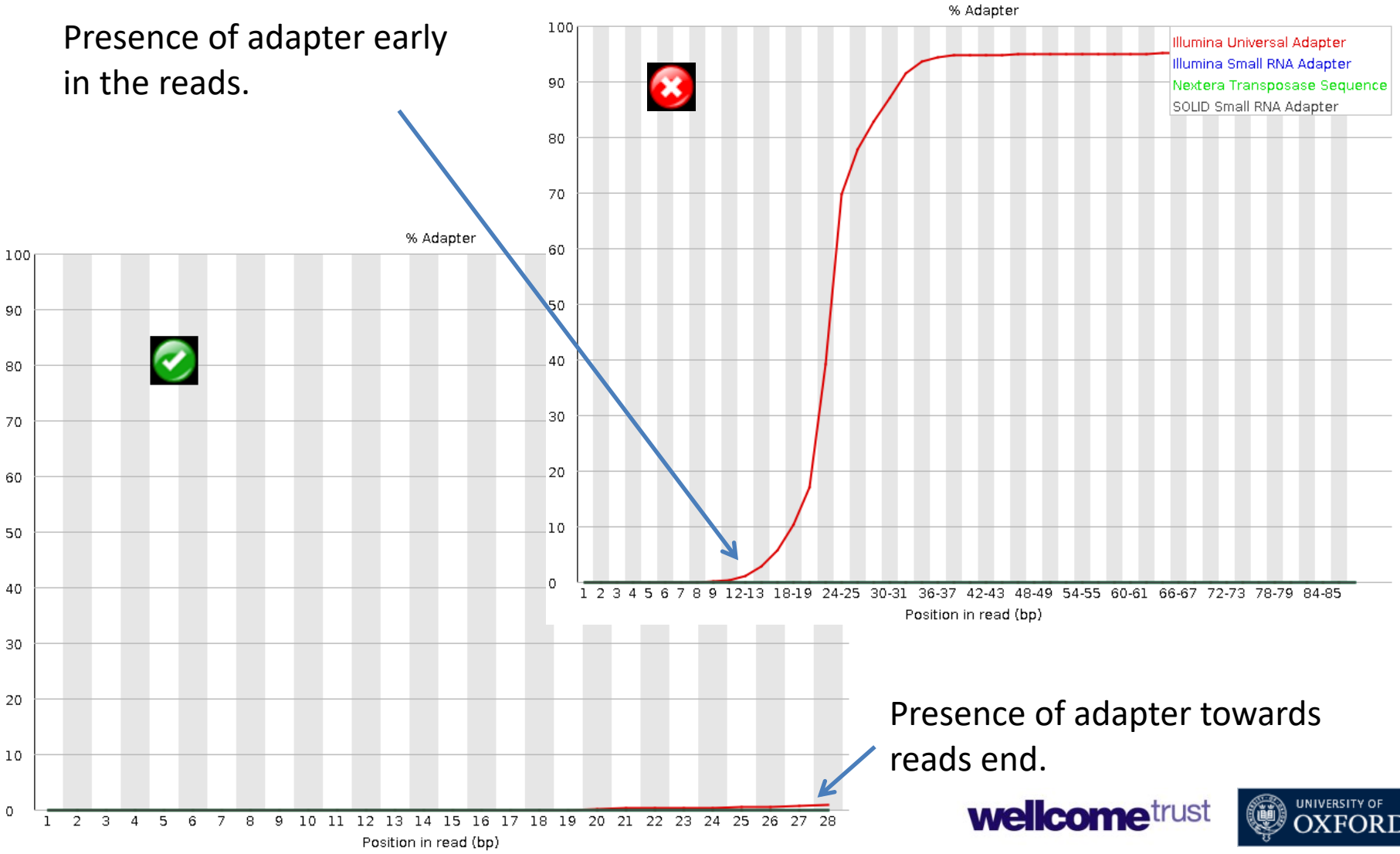


% of Ns per sequence



Rawdata Quality Control
Adapter content

Presence of adapter early in the reads.



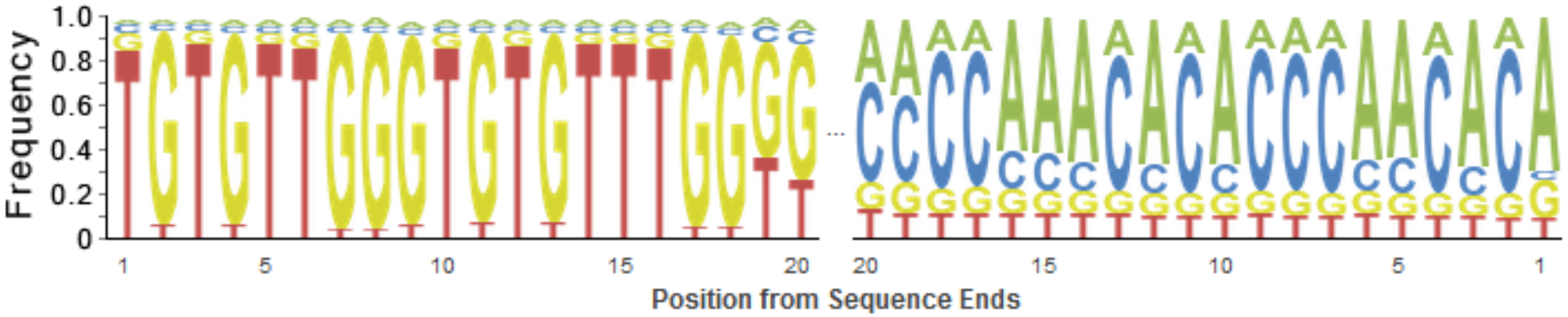
Presence of adapter towards reads end.

wellcome trust
UNIVERSITY OF OXFORD

Rawdata Quality Control

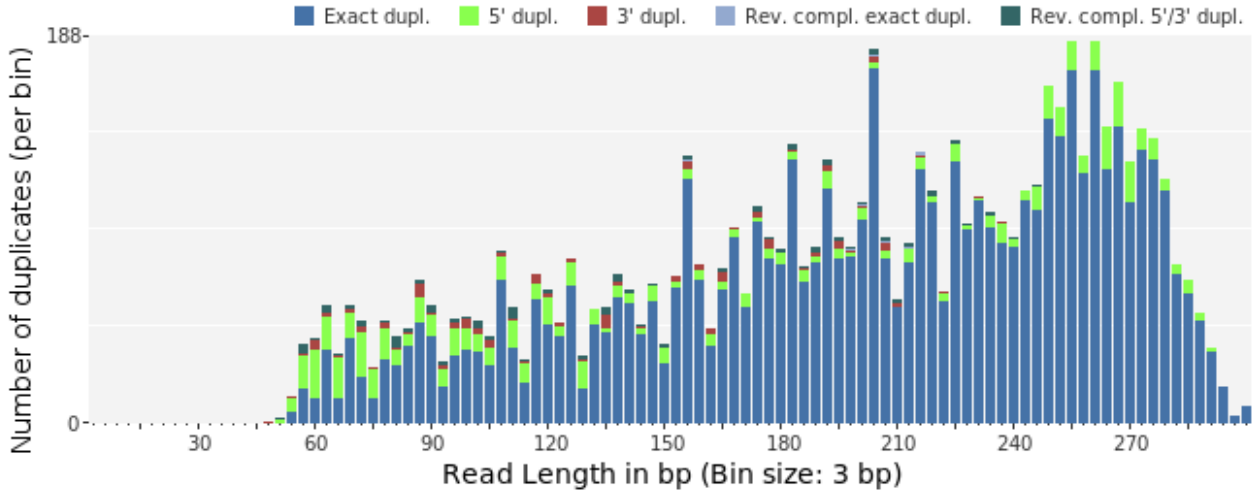
Tag sequence check

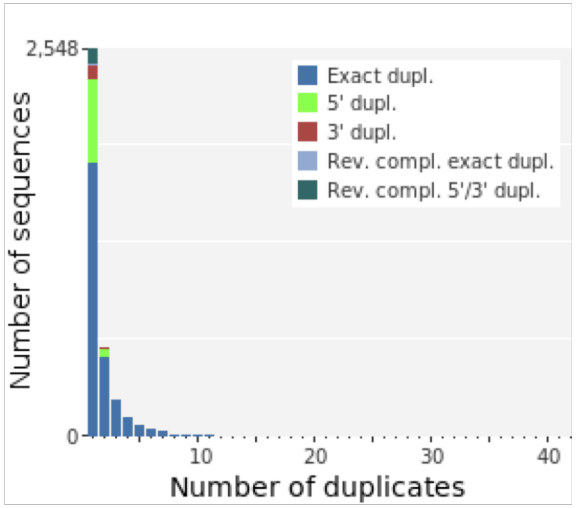
	5'-end	3'-end
Probability of tag sequence:	81 %	49 %
GSMIDs or RLMIDs:	none	

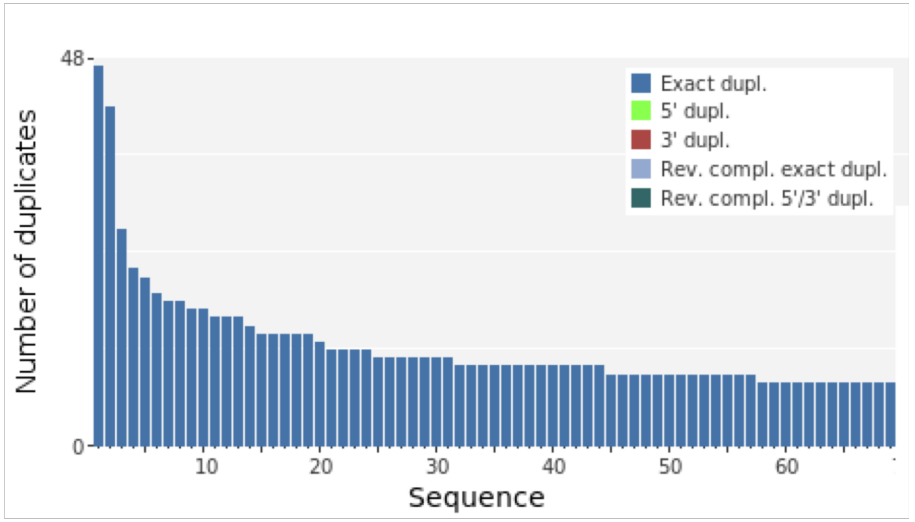


Rawdata Quality Control

Sequence duplication







Inspecting alignment data

Sequence Alignment Map (SAM) Format

```
K00198:242:HLGYVBBXX:8:1119:6137:36112          163     1     12636     1     75M     =     13406     845
            CCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCA
            AAFFFJJJJJJJJJJJJJJJJJJJJAJJJJJJJJJFJJJJ7FJJJ<FFJJFJFJJJJJJJJJJJJJJJJJFFJF
            NH:i:3      HI:i:1      AS:i:148    nM:i:0

K00198:242:HLGYVBBXX:8:1119:6137:36112    83     1     13406     1     75M     =     12636    -845
            CTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCCAGTTTCACCAGAAGTAGGCCTCTTCCTGAC
            JJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA
            NH:i:3      HI:i:1      AS:i:148    nM:i:0
```

Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

Flag 163

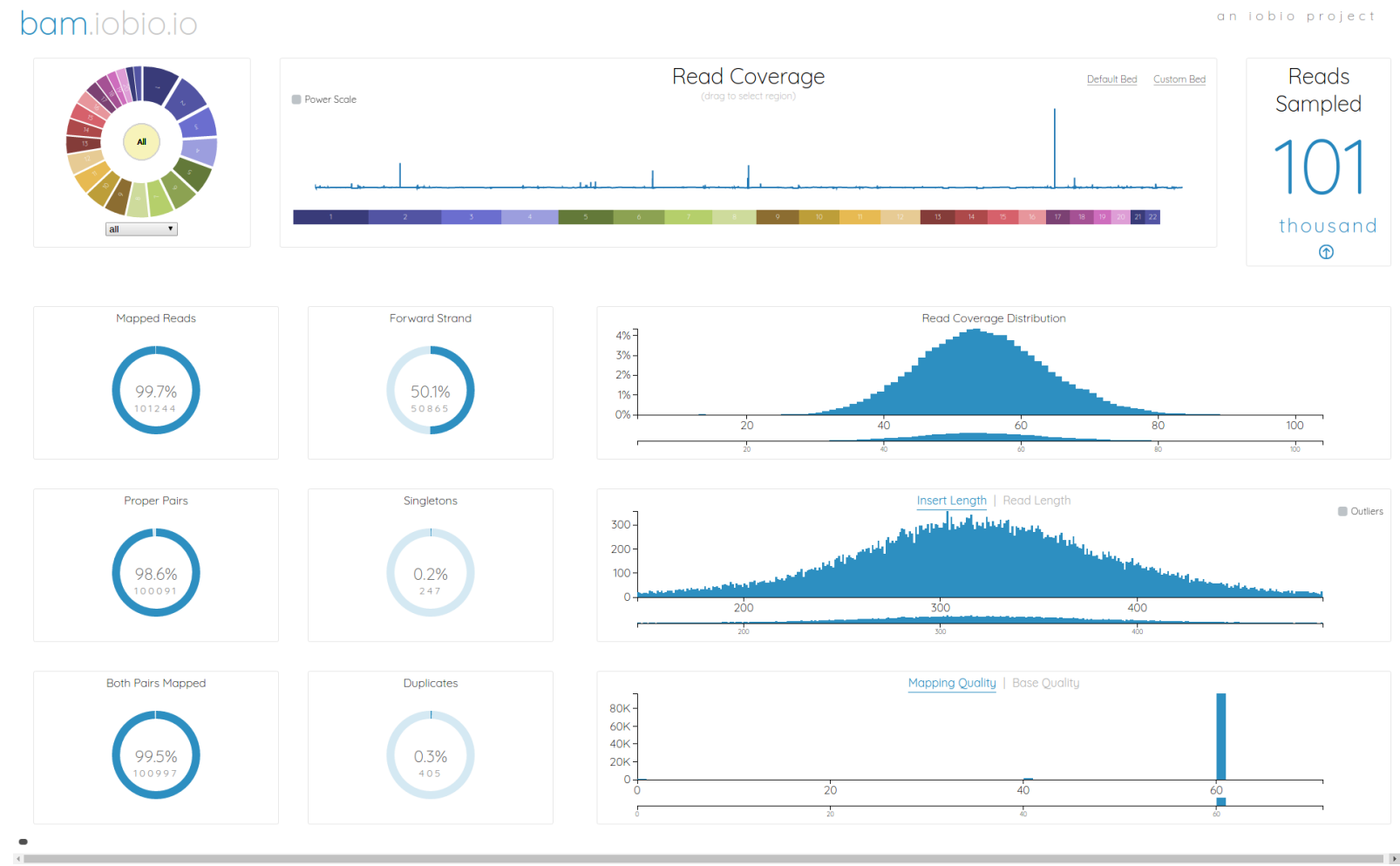
read paired (0x1)
read mapped in proper pair (0x2)
mate reverse strand (0x20)
second in pair (0x80)

Flag 83

read paired (0x1)
read mapped in proper pair (0x2)
read reverse strand (0x10)
first in pair (0x40)

Oxford Genomics CENTRE

Alignment Stats (using bam.iobio.io)

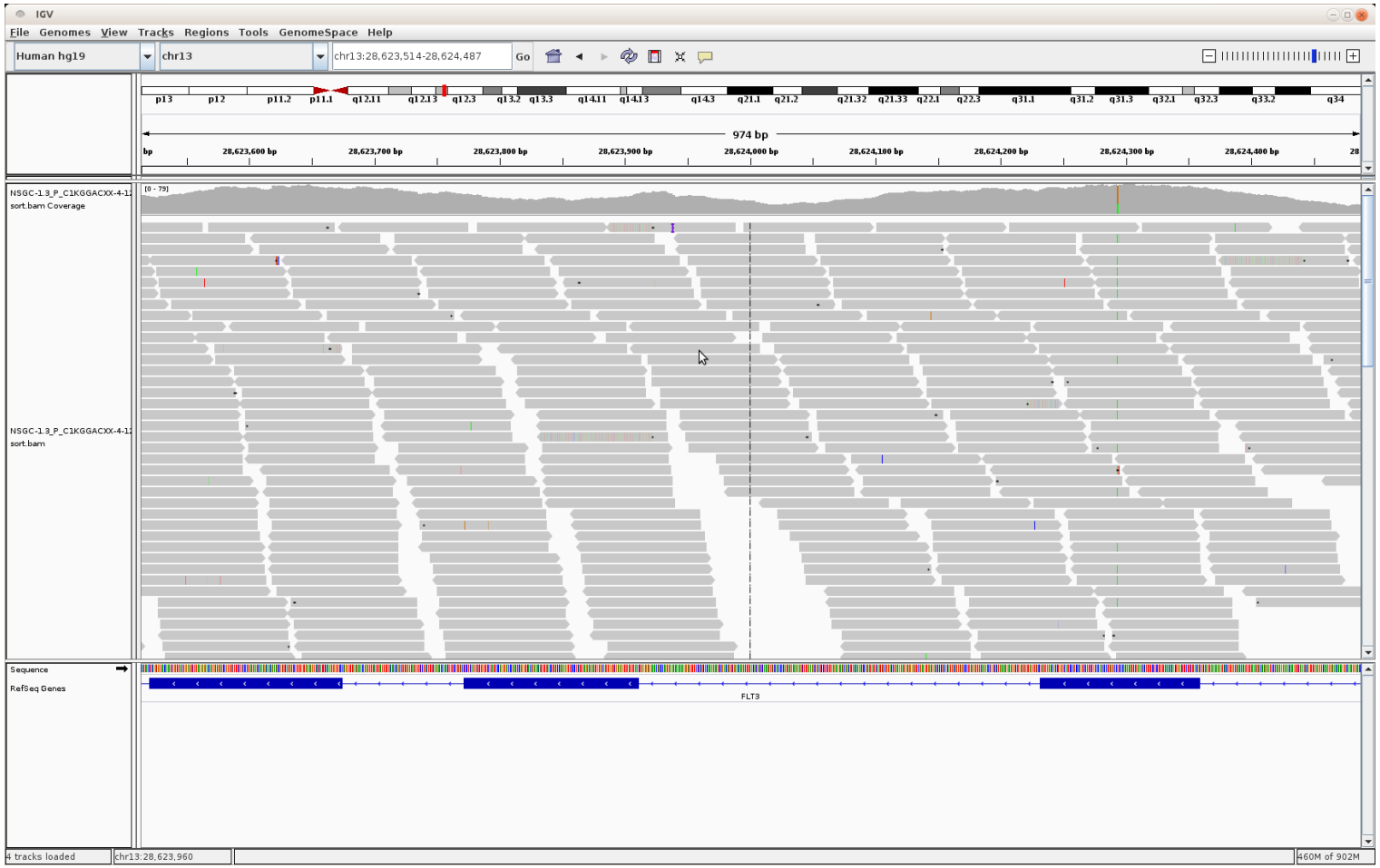


wellcome trust

UNIVERSITY OF OXFORD

Oxford Genomics CENTRE

Alignment Visualization (using IGV)



wellcome trust

UNIVERSITY OF OXFORD

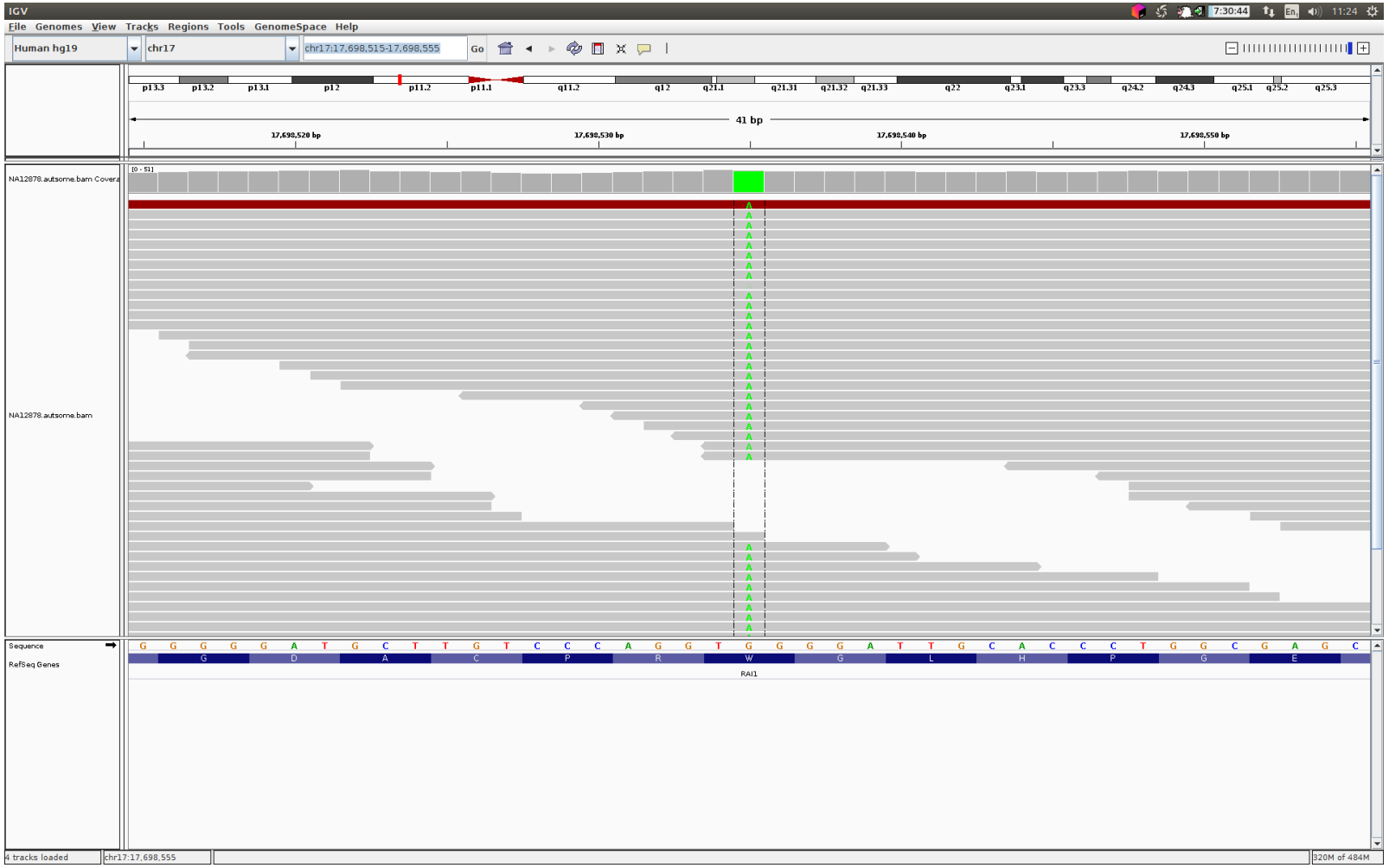
Oxford Genomics CENTRE

Alignment Visualization (using IGV)

wellcome trust

UNIVERSITY OF OXFORD

Alignment Visualization (using IGV)



Oxford Genomics CENTRE

Questions?

wellcome trust

UNIVERSITY OF OXFORD

Playing with some data

Oxford Genomics CENTRE

Playing with some data

Inspecting rawdata using PrinSeq

1) Go to http://prinseq.sourceforge.net/
2) Click on “Use PRINSEQ”
3) Click on “Access data”
4) Click on different example datasets and compare them.

1

Inspecting rawdata using FastQC

1) Download the ”rawdata” files from http://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Sept2018/Day1_260918/Session1/practical/rawdata/
2) Download FastQC from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
3) Run the software by clicking on the “run_fastqc.bat” or “fastqc” files.
4) Once loaded, from the ”File” menu, click ”Open” and select the downloaded files.
5) Once loaded, inspect both files and compare the results.

2

Oxford Genomics CENTRE

Playing with some data

Inspecting aligned data using bam.iobio.io

1) Go to http://bam.iobio.io/
2) Click on “choose bam url”.
3) Click on “Go”.
4) Play around with the data.

1

Inspecting aligned data using Qualimap2

1) Download the "alignment" files from http://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Sept2018/Day1_260918/Session1/practical/alignment/
2) Go to http://qualimap.bioinfo.cipf.es/
3) Download the version that is appropriate for your operating system.
4) Unzip the file and open the software through the ”qualimap” file.
5) Once opened, from the "File" menu, click ”New Analysis” -> ”BAM QC”, select one of the recently downloaded BAM files and click "Open".
6) Play around with the tool and check the different metrics that were produced.

2

Oxford Genomics CENTRE

Playing with some data

1

Inspecting aligned data using Integrative Genomics Viewer (IGV)

1) Download the "alignment" files from http://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Sept2018/Day1_260918/Session1/practical/alignment/
2) Go to https://software.broadinstitute.org/software/igv/download
3) Download the version that is appropriate for your operating system.
4) Open IGV by following the instructions provided on the download page.
5) Once opened, from the "File" menu, click "Load from file" and select both recently downloaded BAM files and click "Open".
6) Play around with the tool by choosing and zooming in different regions of the genome. In the blank box, you could even write the name of a gene to quickly go to it.