Oxford Genomics CENTRE

UNIVERSITY OF OXFORD

RNA-Seq Data Analysis
27-28 November, 2017

Taught module for DPhil programme in
Genomic Medicine and Statistics

Organised and delivered by Bioinformatics Core at WHG:

Helen Lockstone M.Sc.

Ben Wright PhD

Santiago Revale, M.Sc.

wellcome centre human genetics

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

Oxford Genomics CENTRE

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/

wellcome trust

UNIVERSITY OF OXFORD

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

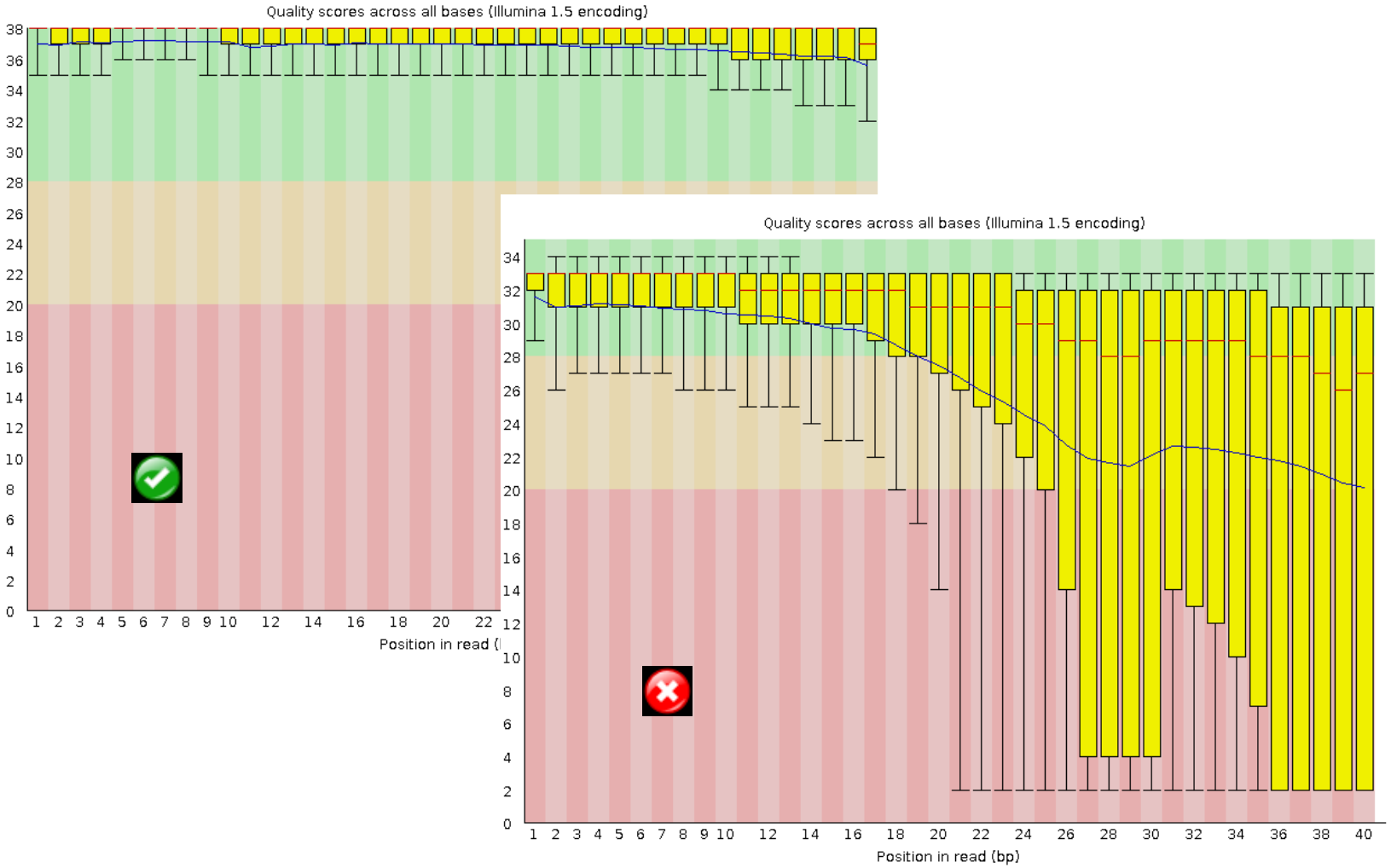
Inspecting rawdata using FastQC

1) Download the ”rawdata” files from http://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Data_Analysis/Monday_am/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

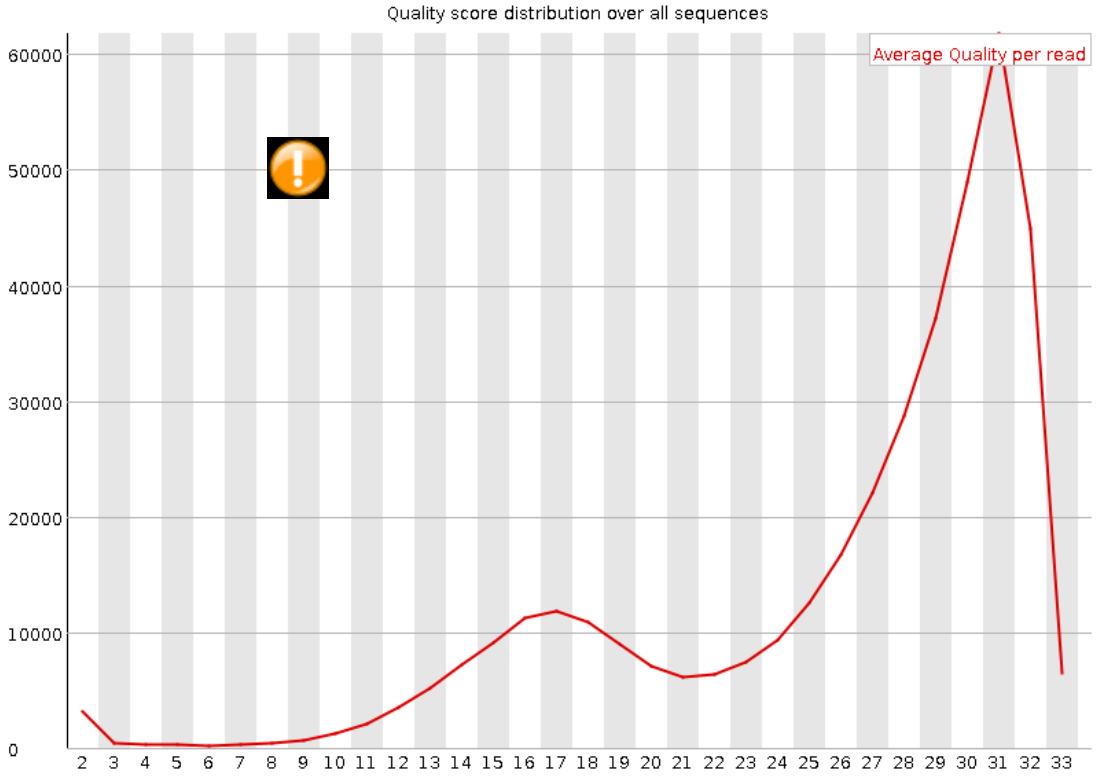
Rawdata Quality Control

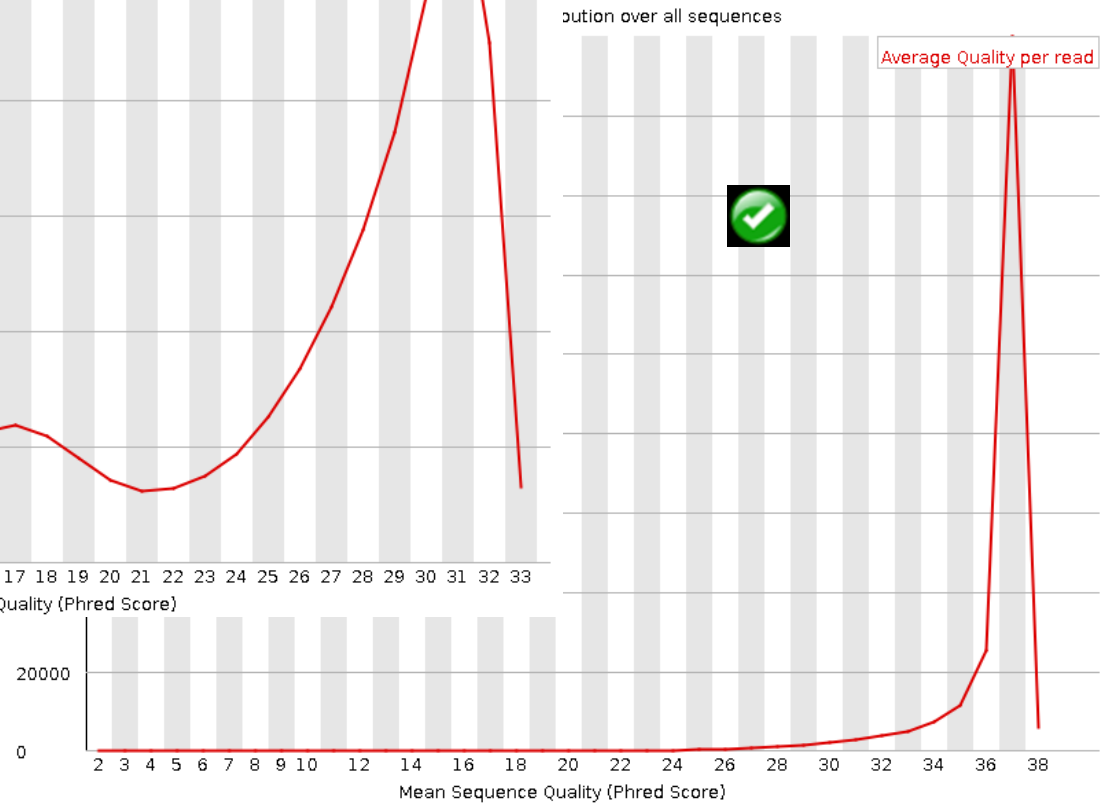
Per base sequence quality



Rawdata Quality Control

Per sequence quality scores

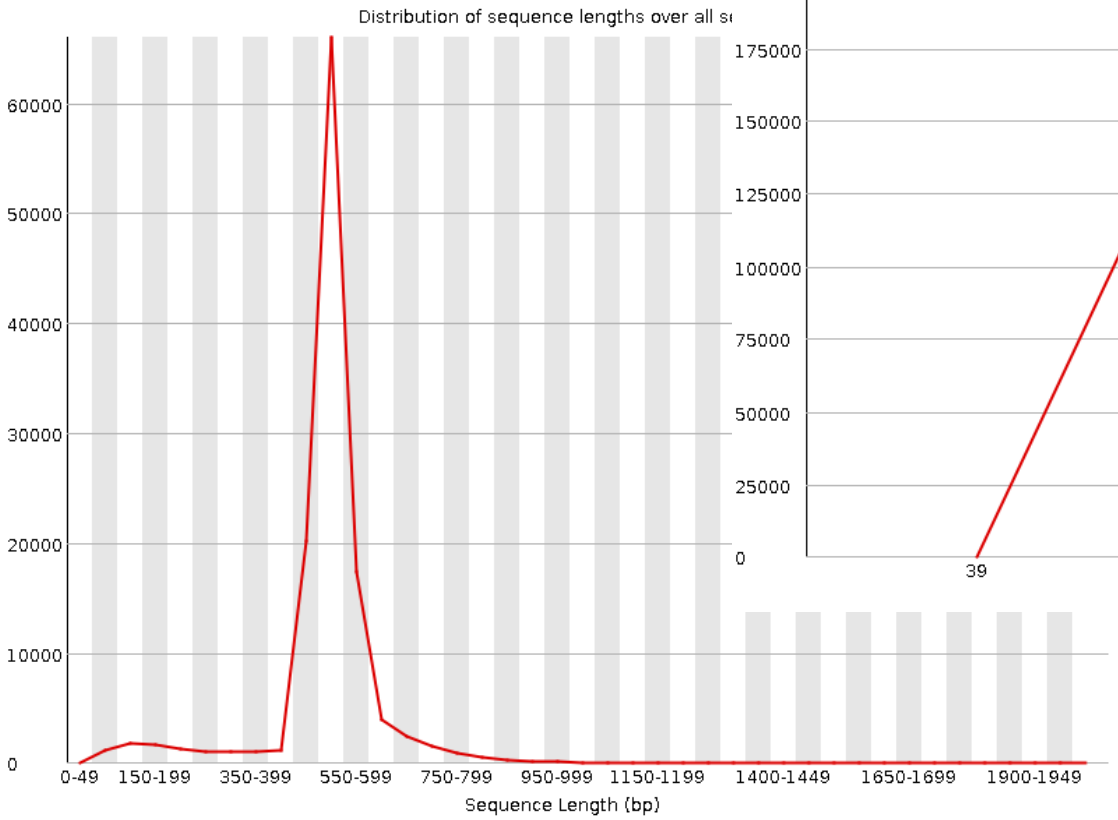


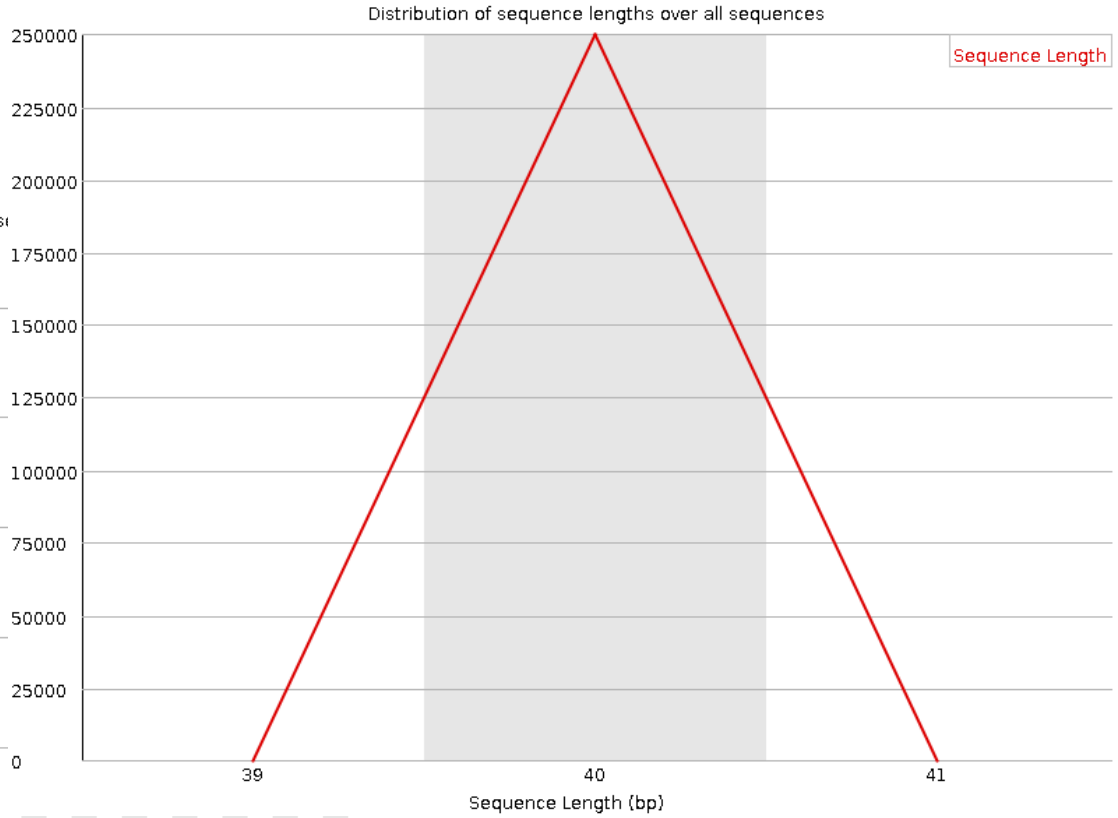


Rawdata Quality Control

Sequence length distribution

E.g. 454/Roche





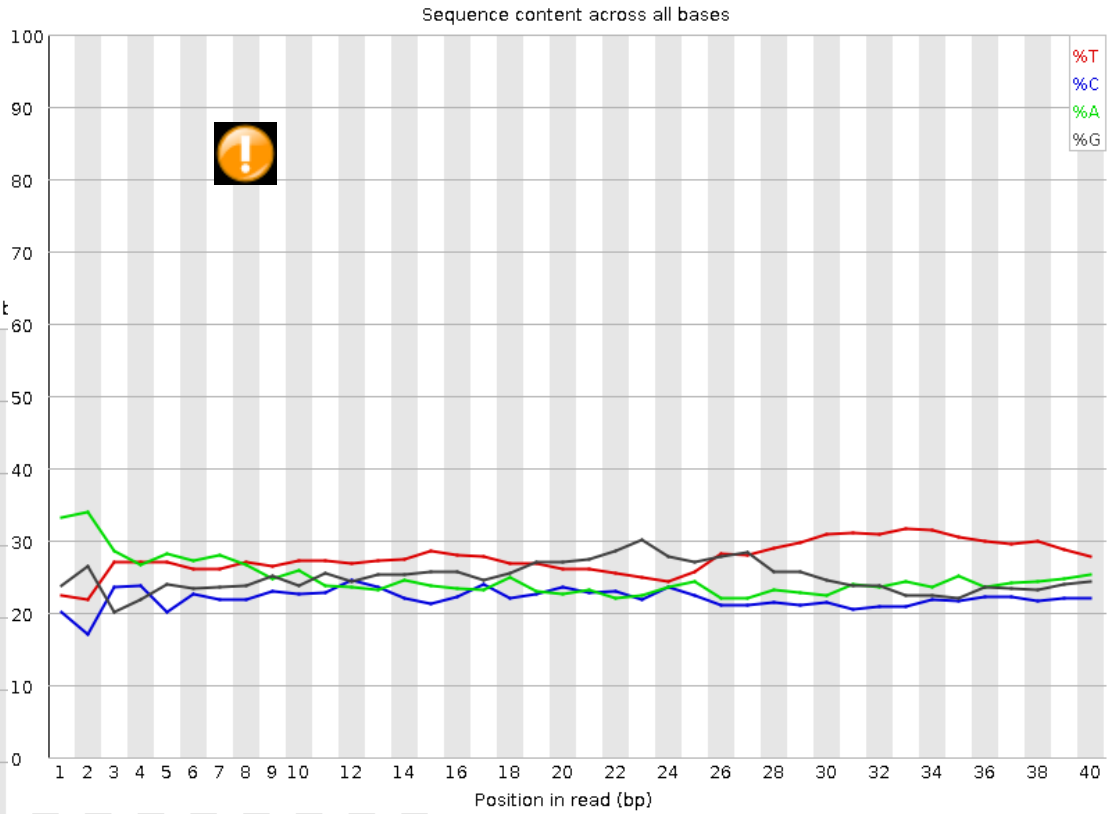
E.g. Illumina

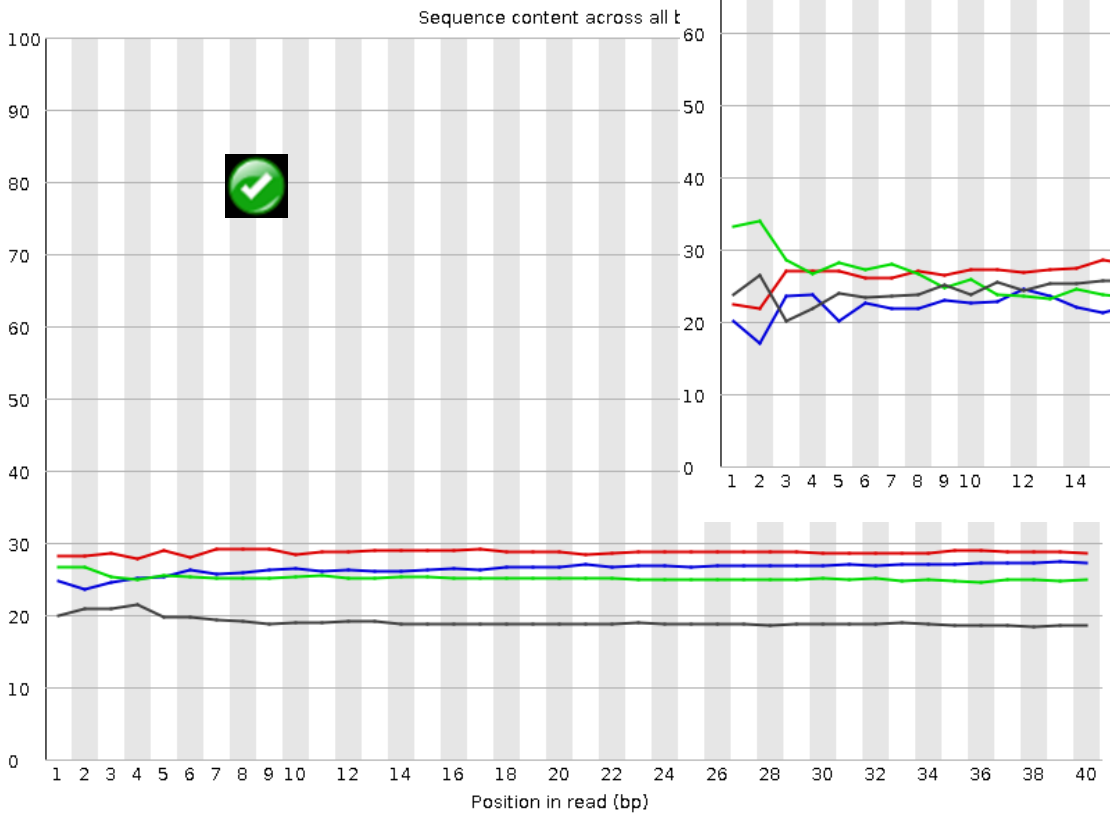
wellcome trust

UNIVERSITY OF OXFORD

Rawdata Quality Control

Per base sequence content

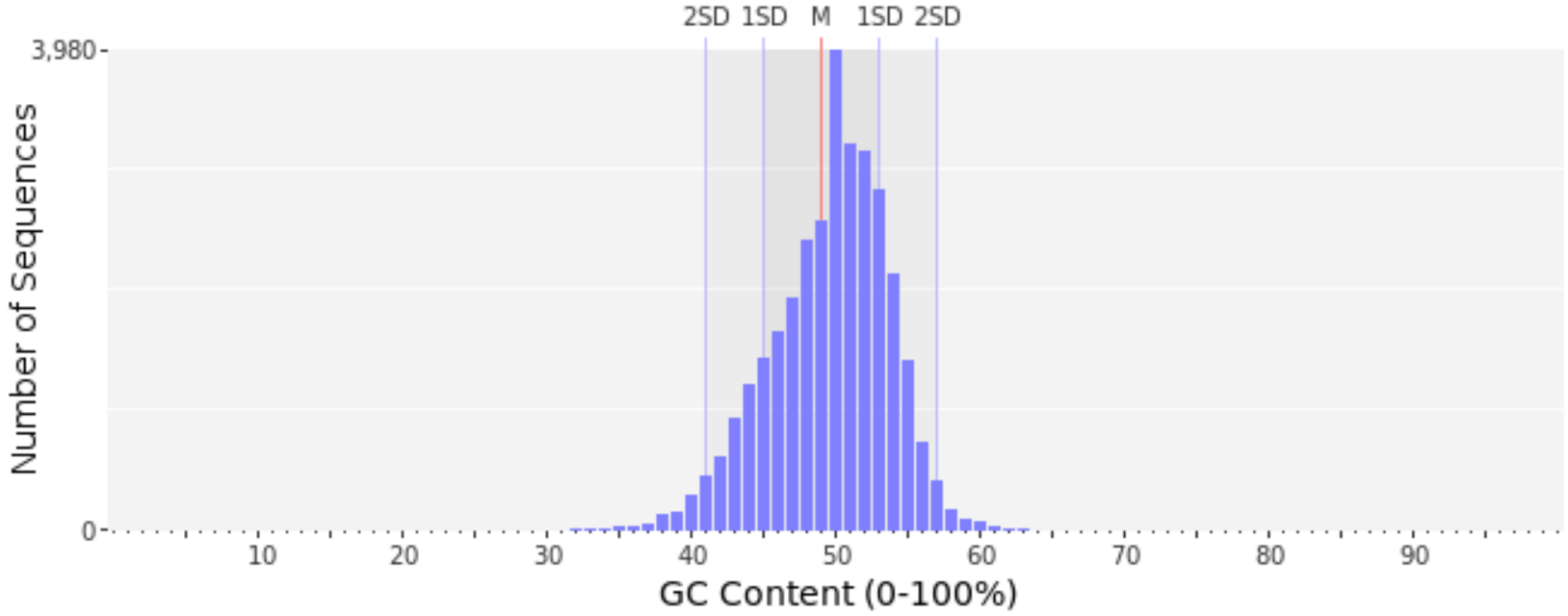




Rawdata Quality Control

GC content distribution

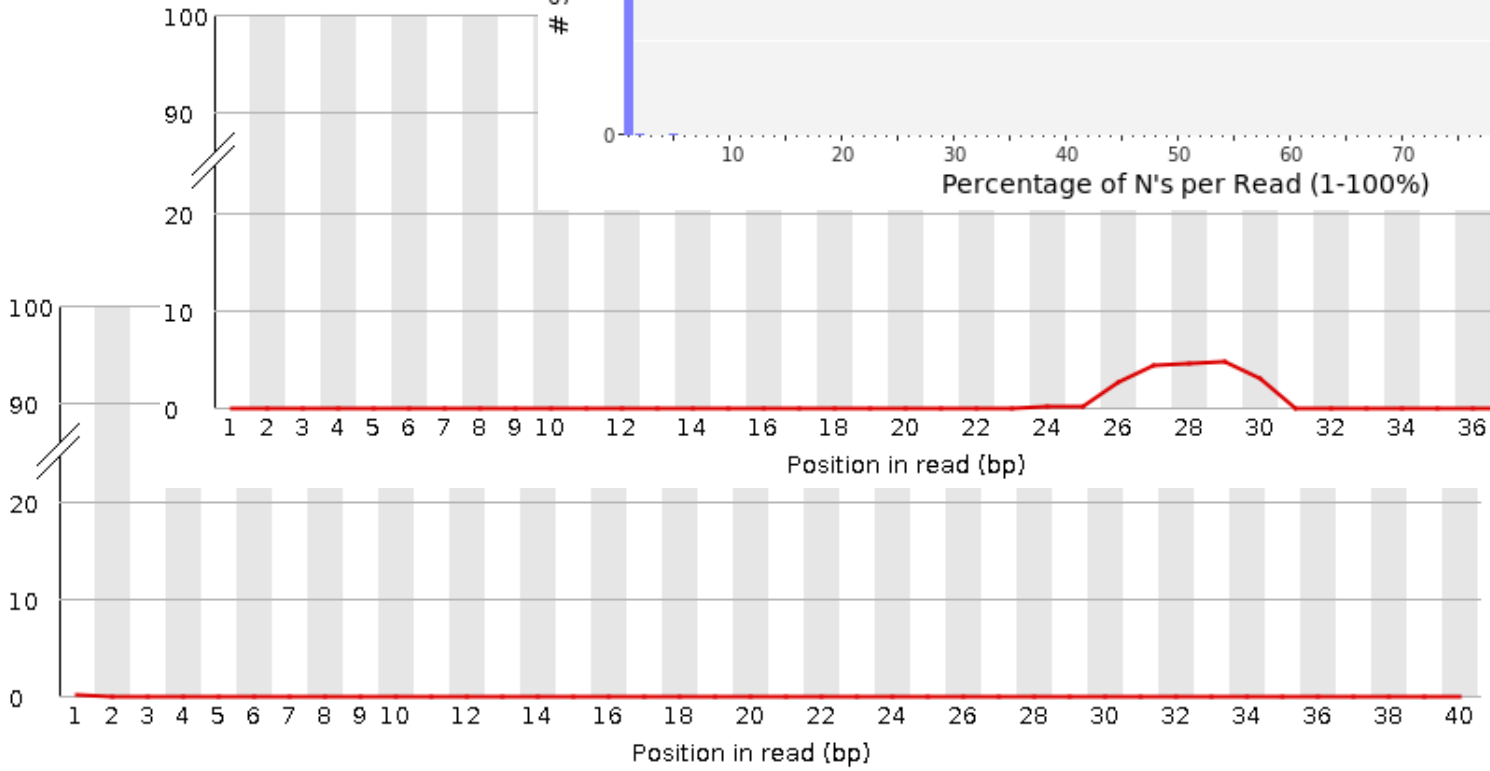
Mean GC content:	**49.55 ± 4.21 %**
Minimum GC content:	**20 %**
Maximum GC content:	**69 %**
GC content range:	**50 %**
Mode GC content:	**50 % with 3,977 sequences**



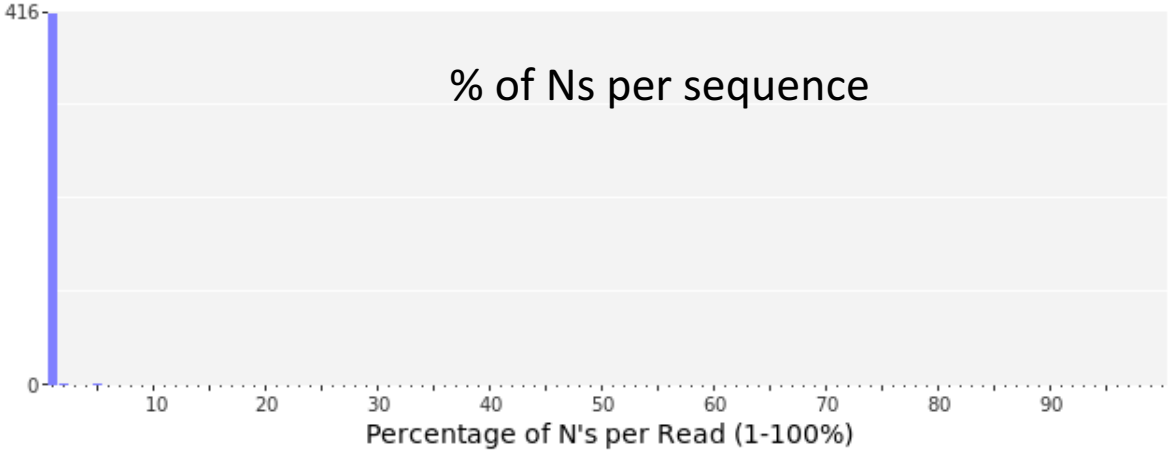
Rawdata Quality Control

N base content

Ns by sequence position



% of Ns per sequence



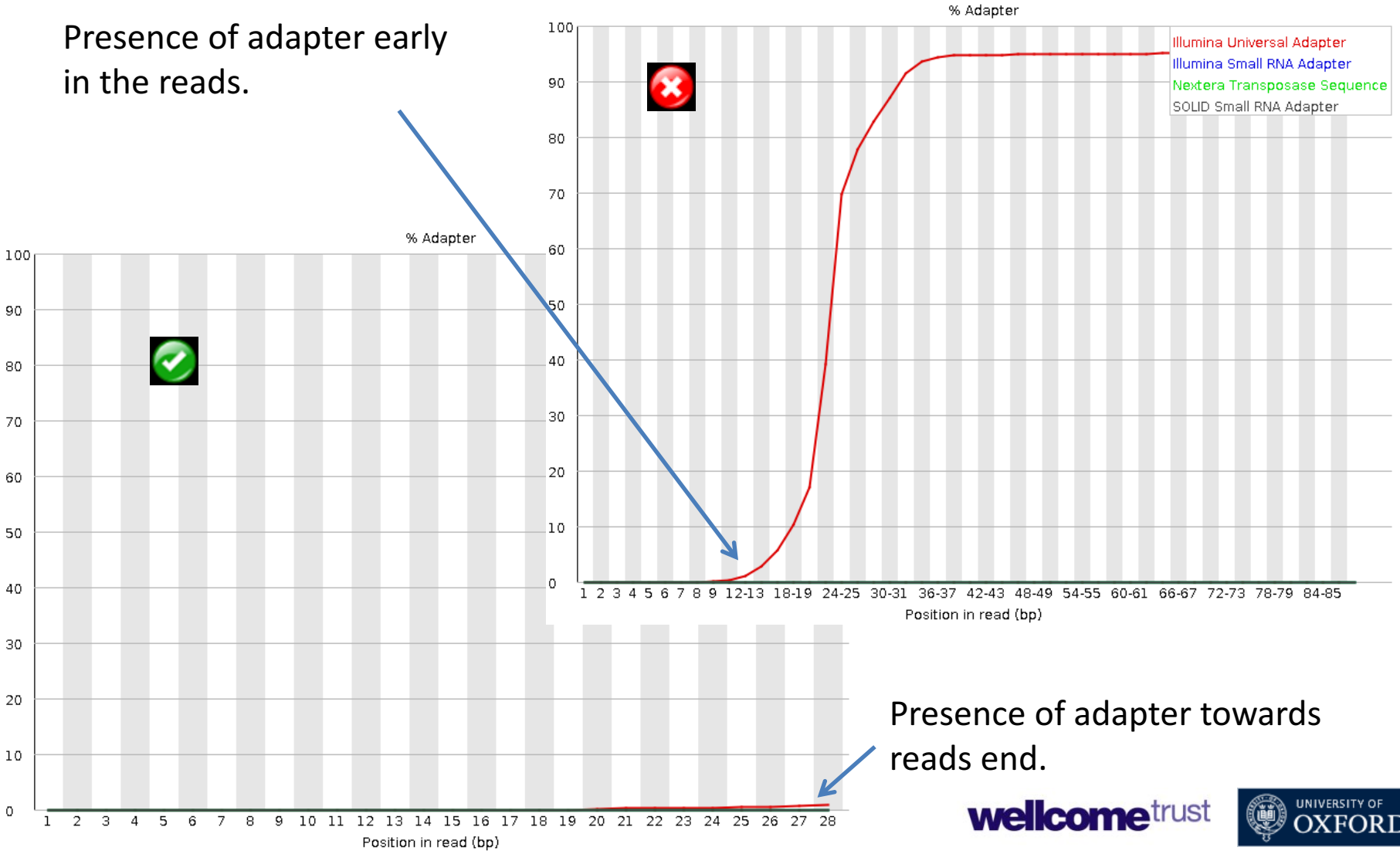
wellcome trust

UNIVERSITY OF OXFORD

Rawdata Quality Control

Adapter content

Presence of adapter early in the reads.



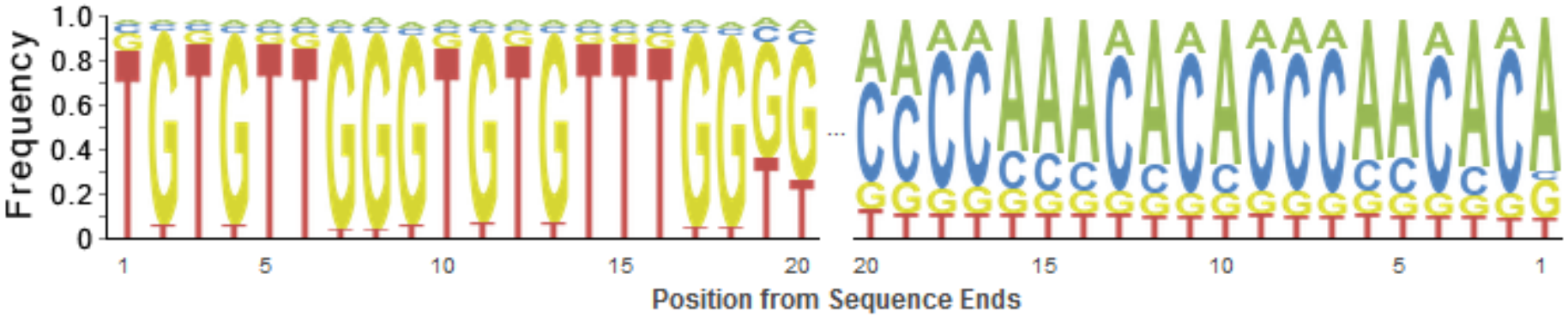
Presence of adapter towards reads end.

wellcome trust
UNIVERSITY OF OXFORD

Rawdata Quality Control

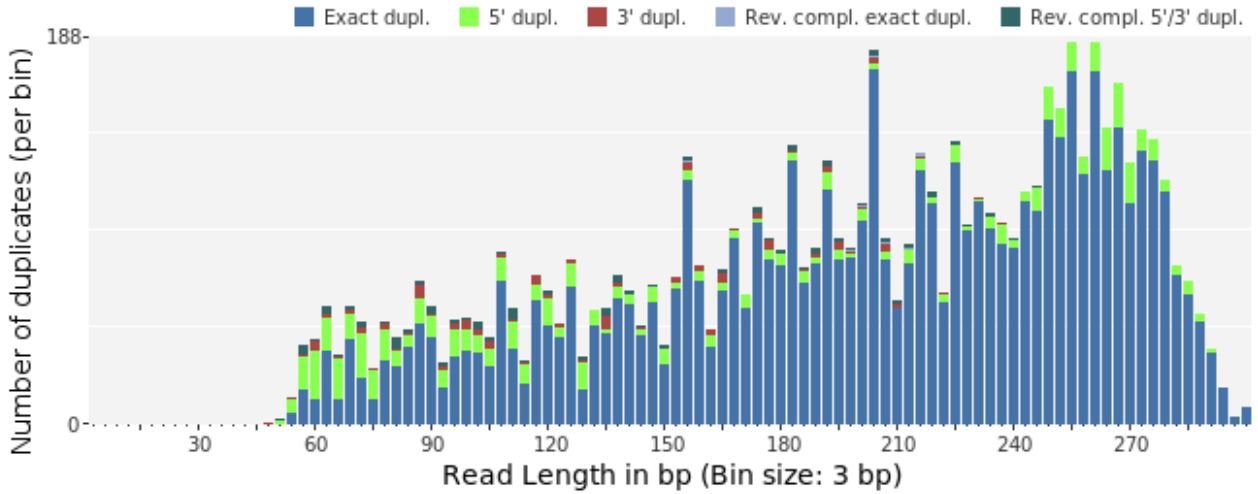
Tag sequence check

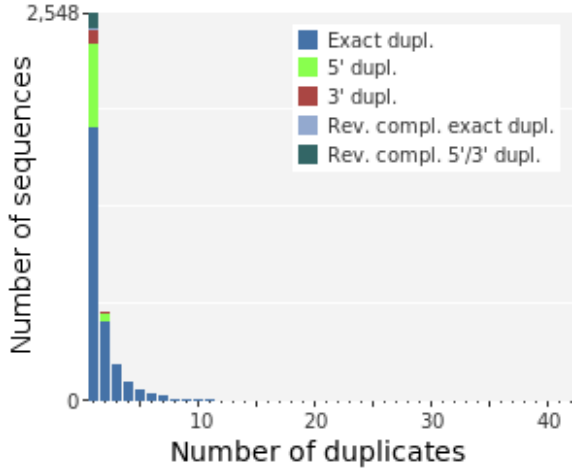
	5'-end	3'-end
Probability of tag sequence:	81 %	49 %
GSMIDs or RLMIDs:	none	

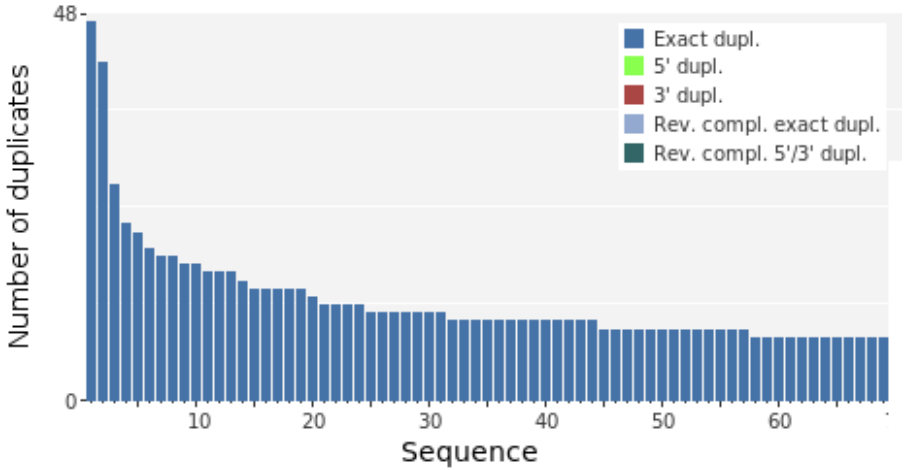


Rawdata Quality Control

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

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

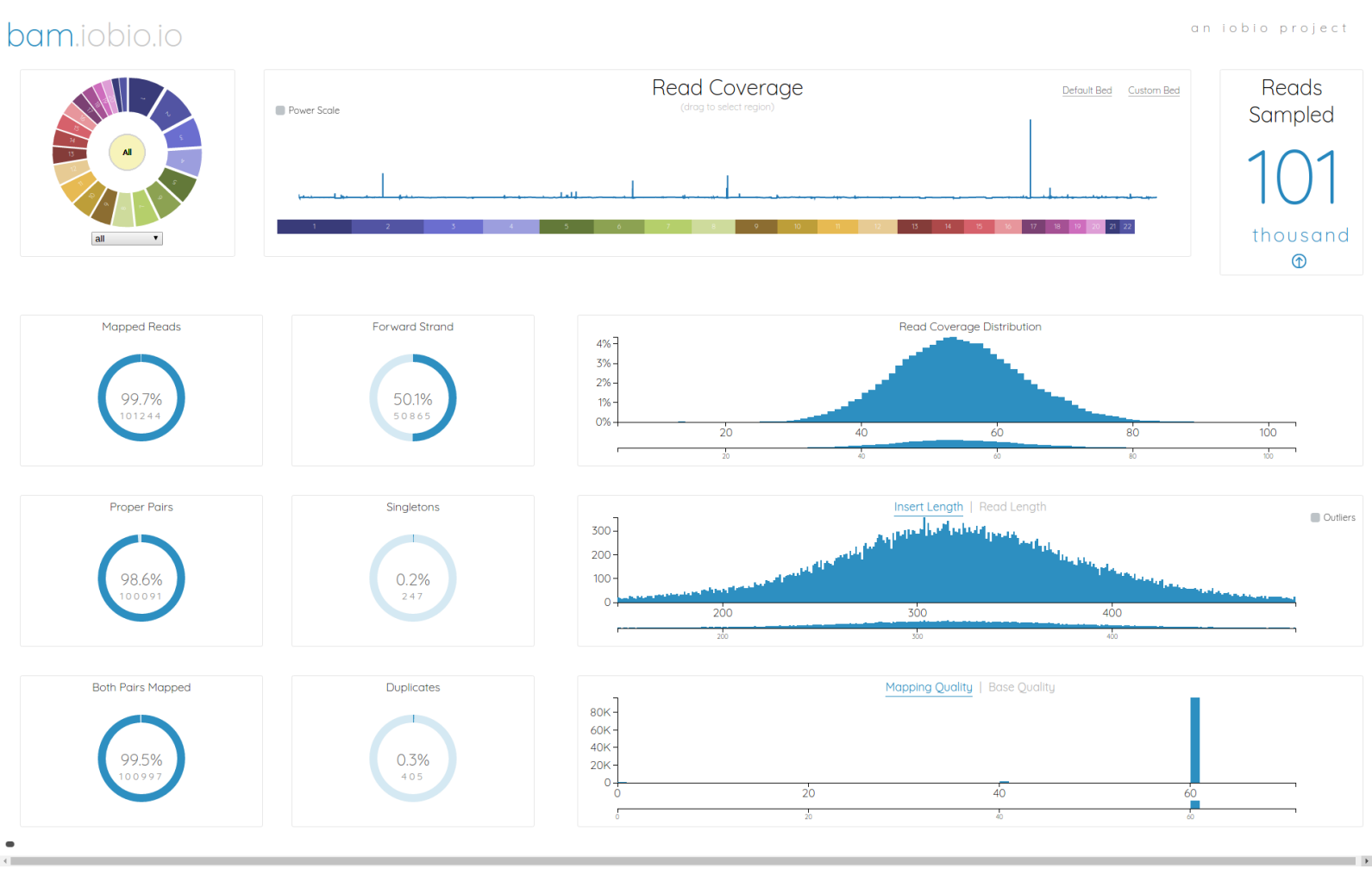
Inspecting aligned data using Qualimap2

1) Download the "alignment" files from http://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Data_Analysis/Monday_am/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

Alignment Stats (using bam.iobio.io)



wellcome trust

UNIVERSITY OF OXFORD

Oxford Genomics CENTRE

Playing with some data

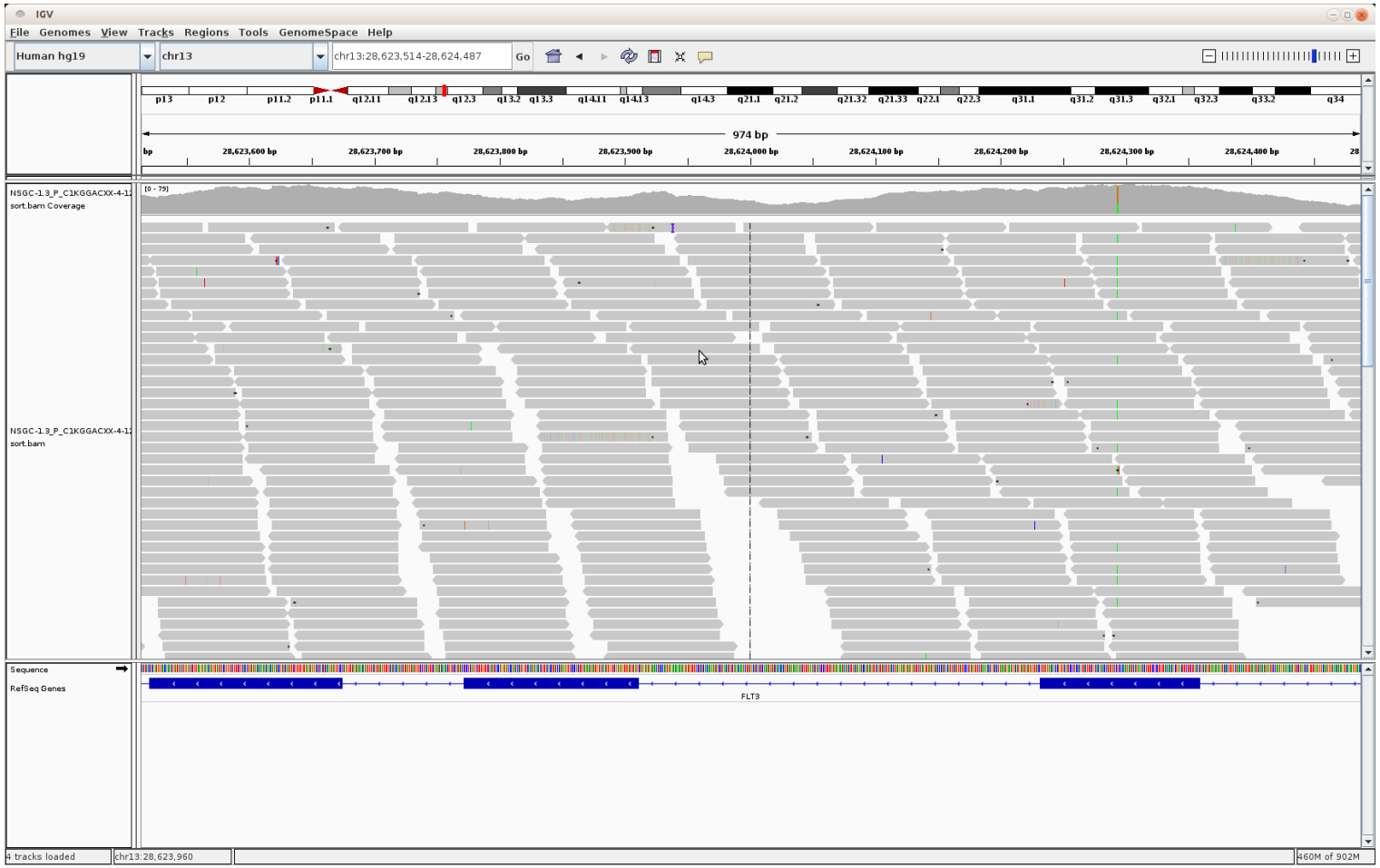
1

Inspecting aligned data using Integrative Genomics Viewer (IGV)

1) Download the "alignment" files from http://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Data_Analysis/Monday_am/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.

Oxford Genomics CENTRE

Alignment Visualization (using IGV)

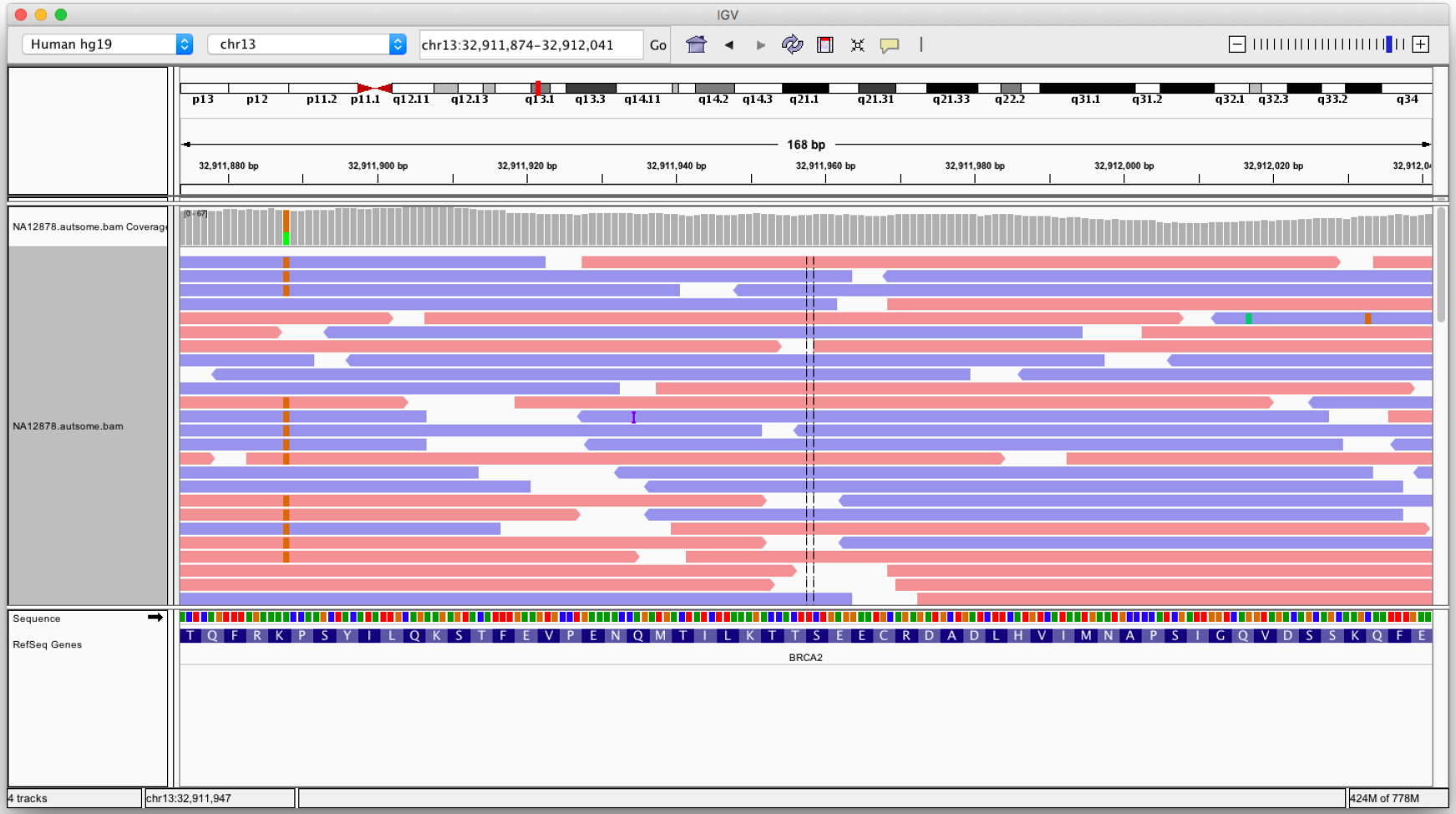


wellcome trust

UNIVERSITY OF OXFORD

Oxford Genomics CENTRE

Alignment Visualization (using IGV)

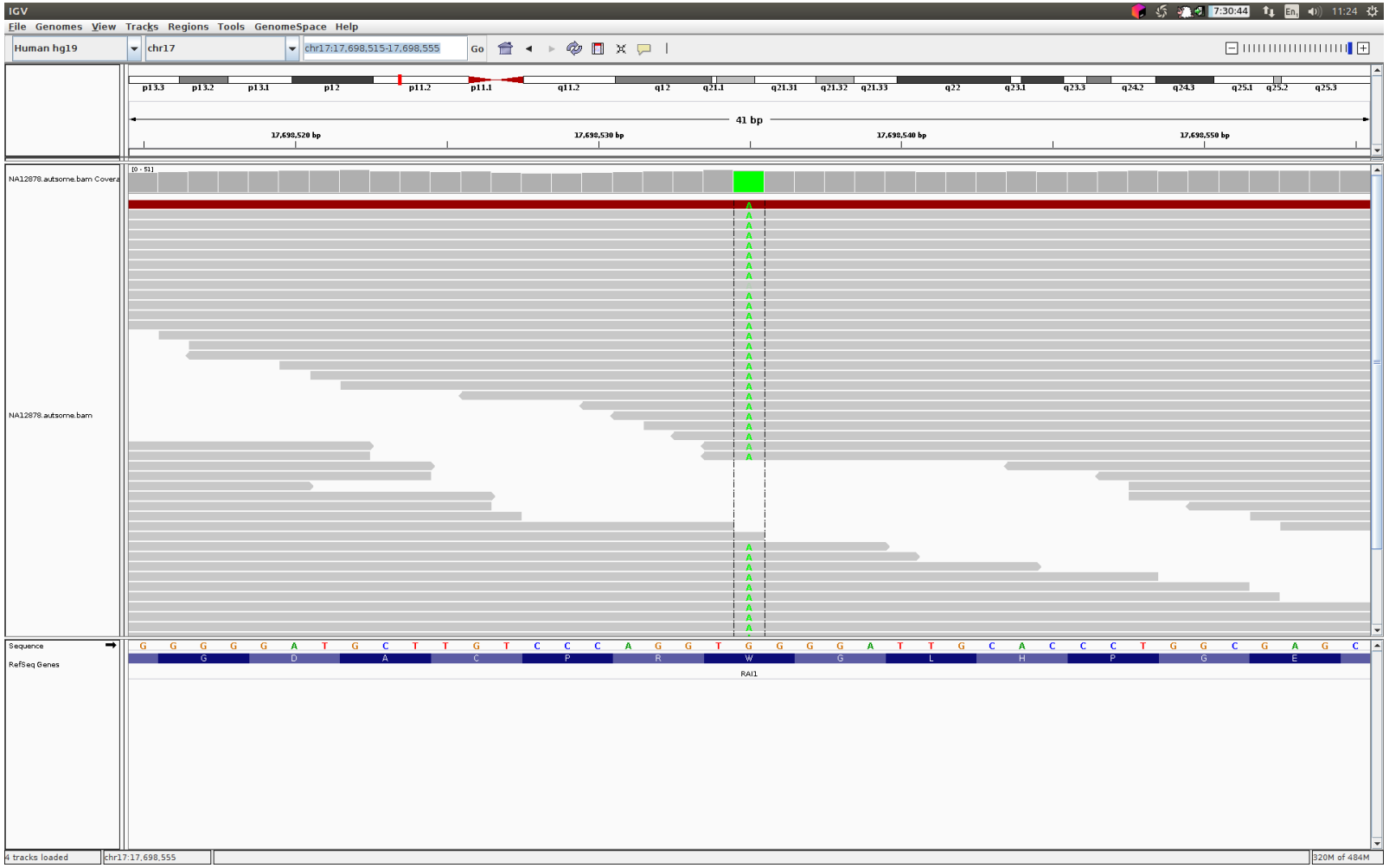


wellcome trust

UNIVERSITY OF OXFORD

Oxford Genomics CENTRE

Alignment Visualization (using IGV)



wellcome trust

UNIVERSITY OF OXFORD

Oxford Genomics CENTRE

Questions?

wellcome trust

UNIVERSITY OF OXFORD