Oxford Genomics CENTRE

UNIVERSITY OF OXFORD

Handling Sequencing Data 30th October, 2020

Taught module for DPhil programme in

Genomic Medicine and Statistics

Organised and delivered by Bioinformatics Core at the

wellcome centre human genetics

Oxford Genomics Centre

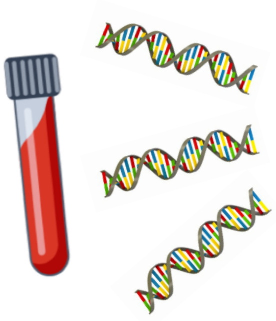
The Wellcome Trust Centre for Human Genetics

Genomic Sequencing Platforms

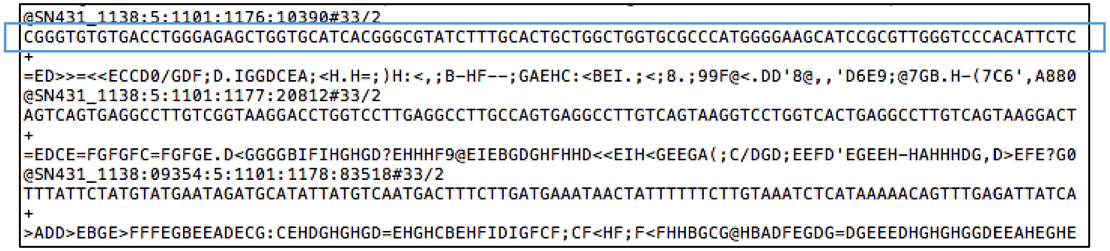
Matthieu Miossec, PhD
Bioinformatics Analyst, Bioinformatics Core Group

Sequencing At a Glance

- The process of determining a sequence of nucleotides and producing a digital representation (known as a **read** or **reads**) for further analysis.



A FASTQ file



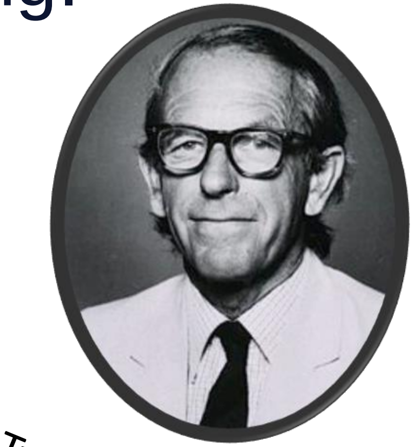
```
@SN431_1138:5:1101:1176:10390#33/2
CGGGTGTGTGACCTGGGAGAGCTGGTGCATCACGGGCGTATCTTTGCACTGCTGGCTGGTGCGCCCATGGGGAAGCATCCGCGTTGGGTCCCACATTCTC
+
=ED>>=<<ECCD0/GDF;D.IGGDCEA;<H.H=;)H:<,;B-HF--;GAEHC:<BEI.;<;8.;99F@<.DD'8@,,'D6E9;@7GB.H-(7C6',A880
@SN431_1138:5:1101:1177:20812#33/2
AGTCAGTGAGGCCTTGTCGGTAAGGACCTGGTCCTTGAGGCCTTGCCAGTGAGGCCTTGTCAGTAAGGTCCTGGTCACTGAGGCCTTGTCAGTAAGGACT
+
=EDCE=FGFGFC=FGFGE.D<GGGGBIFIHGHGD?EHHHF9@EIEBGDGHFHHD<<EIH<GEEGA(;C/DGD;EEFD'EGEEH-HAHHHDG,D>EFE?G0
@SN431_1138:09354:5:1101:1178:83518#33/2
TTTATTCTATGTATGAATAGATGCATATTATGTCAATGACTTTCTTGATGAAATAACTATTTTTTCTTGTAAATCTCATAAAAACAGTTTGAGATTATCA
+
>ADD>EBGE>FFFEGBEEADECG:CEHDGHGHGD=EHGHCBEHFIDIGFCF;CF<HF;F<FHHBGCG@HBADFEGDG=DGEEEDHGHGHGGDEEAHEGHE
```

In a perfect world, we could sequence entire chromosomes in real-time and get a single error-free read for each chromosome as our output.

(we don’t live in a perfect world)

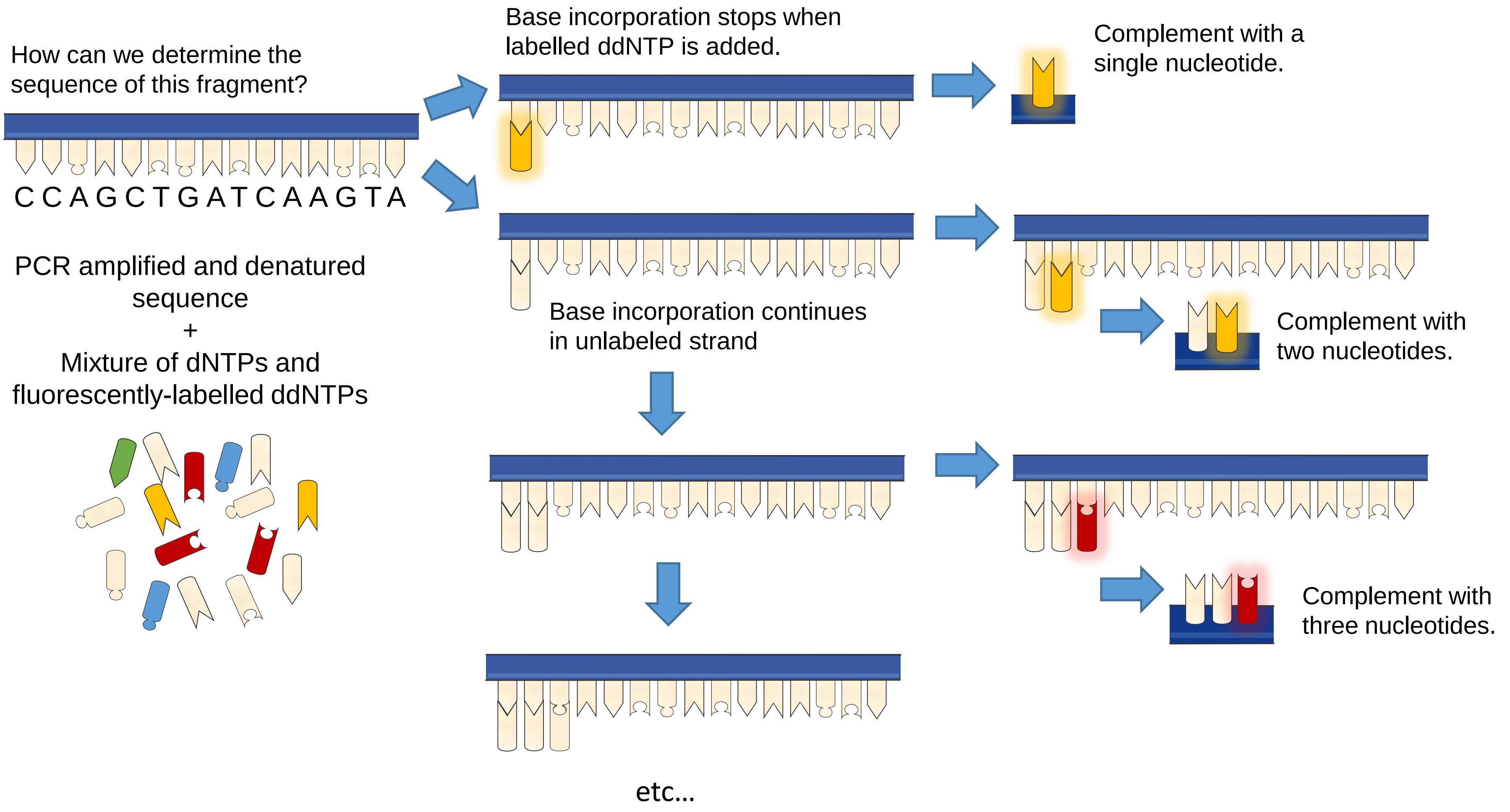
The Dawn of Sequencing

- For close to 40 years, the **Sanger method** was the dominant approach for sequencing.
  - Originally required a lot of lab work, but progressively automated.
  - Various organisms sequenced for the first time using the Sanger method, including *Homo sapiens*!
    - The famous Human Genome Project (HGP) took 13 years to ‘complete’ (earlier than expected!).



Twice Nobel Prize winner
Frederick Sanger

How does Sanger work? 1/2



How can we determine the sequence of this fragment?

CCAGCTGATCAAGTA

PCR amplified and denatured sequence
+
Mixture of dNTPs and fluorescently-labelled ddNTPs

Base incorporation stops when labelled ddNTP is added.

Complement with a single nucleotide.

Base incorporation continues in unlabeled strand

Complement with two nucleotides.

Complement with three nucleotides.

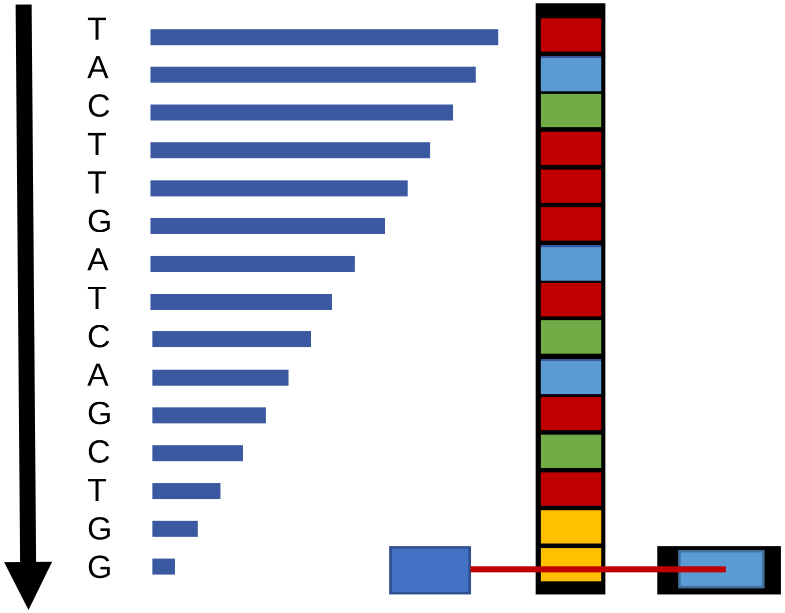
etc…

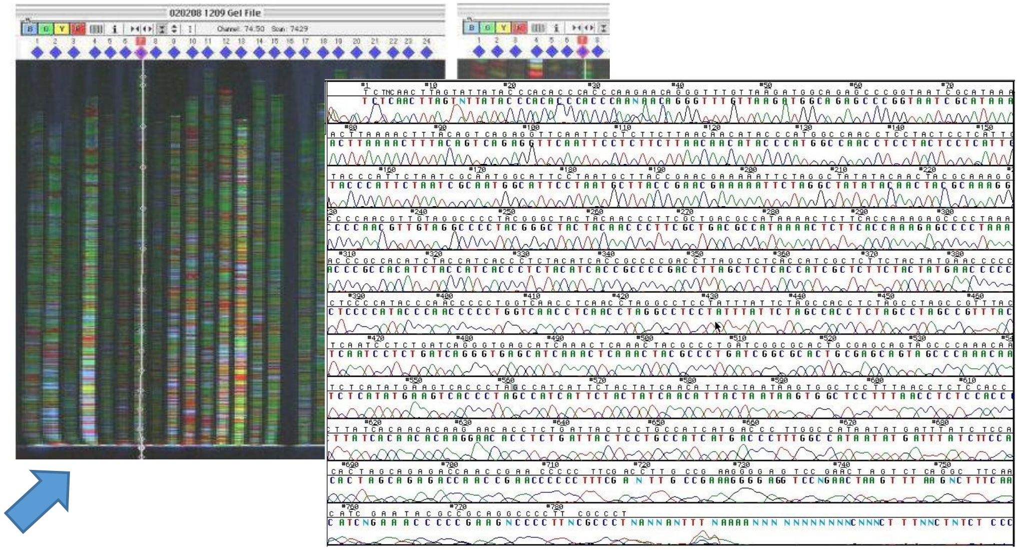
This process gives us multiple incomplete complement fragments of various lengths. The length is a proxy for the location at which the chain terminating ddNTP was incorporated.

dNTP = Deoxynucleoside triphosphate
ddNTP = Dideoxynucleoside triphosphate (chain-terminating)

How does Sanger work? 2/2

- Complement fragments are passed through a capillary tube using electrophoresis, their final ordering determined by size.
  - Smaller fragments pass through first.
  - Fluorescence is captured via laser excitation.





Sanger method’s attributes

- To this day, Sanger’s single read per base quality remains unmatched.
  - For a fragment 400-500 base pairs (bp), per base accuracy reaches **99.999%**.
  - Fragments can now reach up to 1000 bp in length.

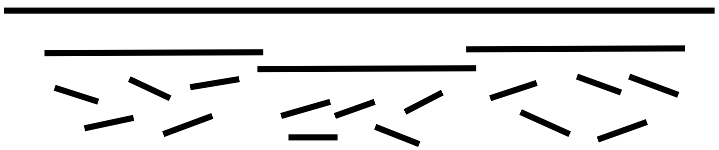
- How can you sequence the 3 billion base pairs of the human genome with sequencing technology that only takes fragments of up to 1000 bp at a time?

Shotgun Sequencing

- The HGP began with clone-by-clone sequencing…

laborious
costly
slow

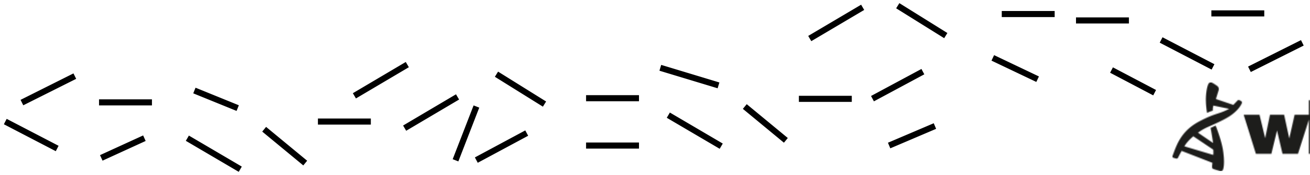
- Genome broken into large 150 kbp fragments, the position of each fragment carefully recorded.
    - Fragments were clonally amplified and then broken into smaller overlapping fragments themselves clonally amplified…



- ..then moved to shotgun.

cheaper
faster

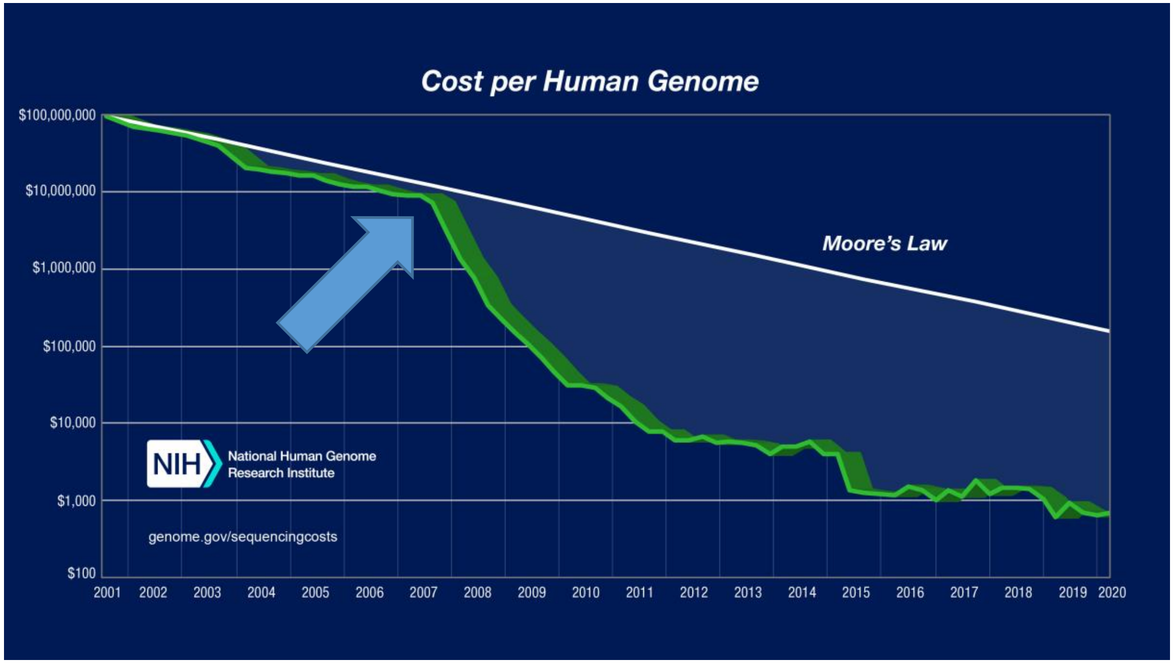
- Genome directly broken up into small overlapping fragments and clonally amplified. Faster but **no ordering is conserved in the process**.



Assembly/mapping after shotgun

Next-generation Sequencing

- Mid 2000s: several **high-throughput/massively parallel** sequencing (HTS/MPS) platforms released.
  - Eventually, what took 13 years to complete now takes hours and costs around 1000$.



The First Three

- Next-generation sequencing (NGS) originally referred two three sequencing platforms:

2005
Roche 454

2006
Solexa (now Illumina)
Genome Analysis
System

2007
ABI SOLiD

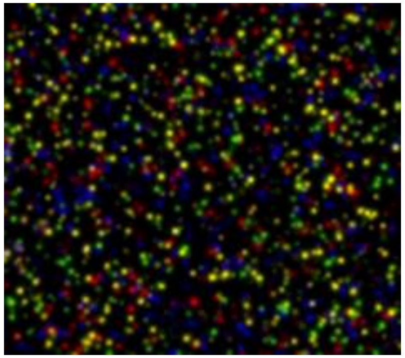
Innovations of the NGS Platforms

- NGS platforms (and subsequent iterations) use various distinct biochemical processes to produce reads from shotgun sequencing, but they do have some attributes in common:
  - PCR amplification is used to turn weak bioluminescent signals generated by a small fragment, into the strong signal of a cluster of identical fragments.
  - Sequencing-by-synthesis!

https://youtu.be/fCd6B5HRaZ8

Limitations of NGS

- While sequencing-by-synthesis makes high-throughput sequencing possible, it does place constraints on **read length** and **accuracy**.
  - Determining the bases of a sequence (aka **base calling**) via bioluminescence has some pitfalls:



❖ Signals from clusters in close proximity can interfere with each other.

❖ Synchronicity between strands in a cluster is gradually lost (one big reason for short reads).

❖ The intensity of a signal can vary.

❖ A signal can be ambiguous when bases repeat (e.g. Did the machine detect C-C or C-C-C?).

Sequencing Platforms 2/2

- Since the mid-2000s, NGS platforms have either been further refined or fallen out of favour.
  - E.g. Illumina now offers a wide range of platforms tailored to different throughput needs.

  - New platforms have also been developed to address some of the flaws of NGS technologies. (Problem = they have their own flaws…)

Sequencing Platform Comparison

	Illumina (NGS)	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 90% range.

Sequencing Outputs

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 Ribodepleted 3' mRNA CHIPseq	80 40 384 80	10 5 48 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

Oxford Genomics Centre

The Wellcome Trust Centre for Human Genetics

Sequencing Data Concepts

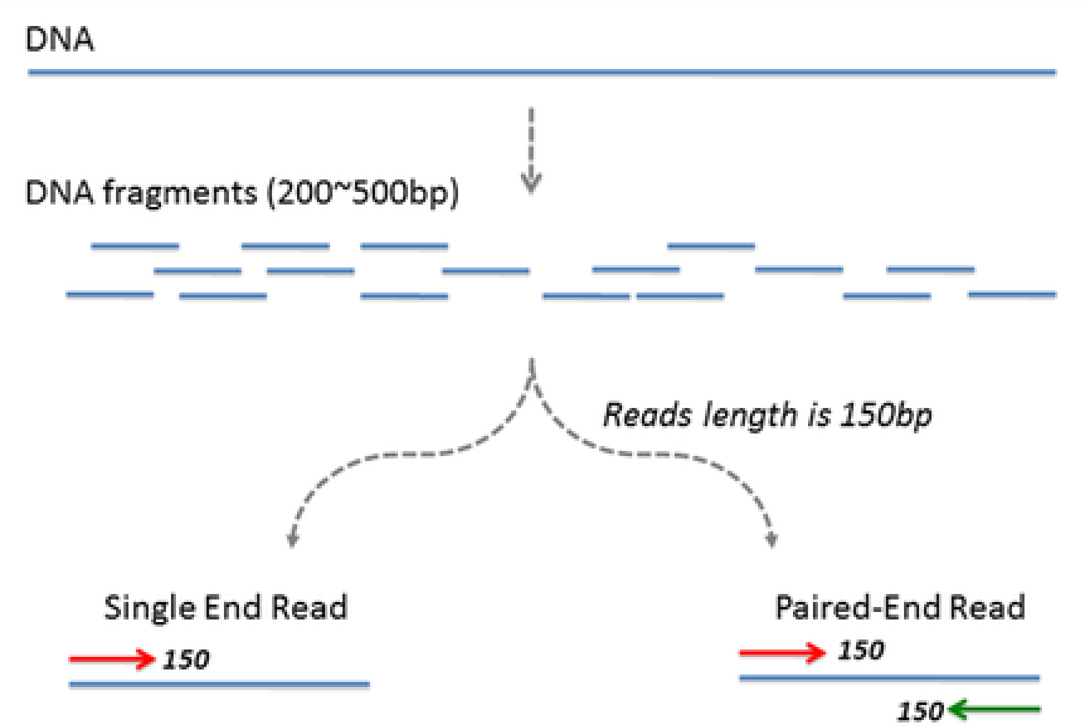
Matthieu Miossec, PhD
Bioinformatics Analyst, Bioinformatics Core Group

whg

UNIVERSITY OF OXFORD

Read Length

- Read length is constrained by the platform being used. It should not vary within a particular run.
  - A read is obtained from a fraction of a DNA/RNA fragment. Two reads can be obtained from a single fragment.

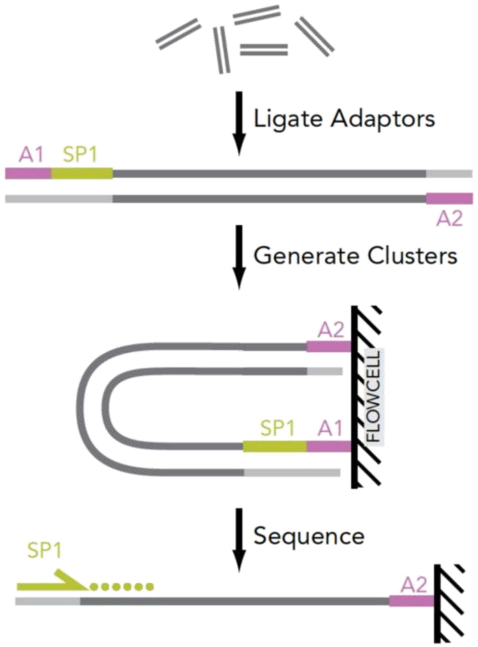


Single-End/Paired-End Reads

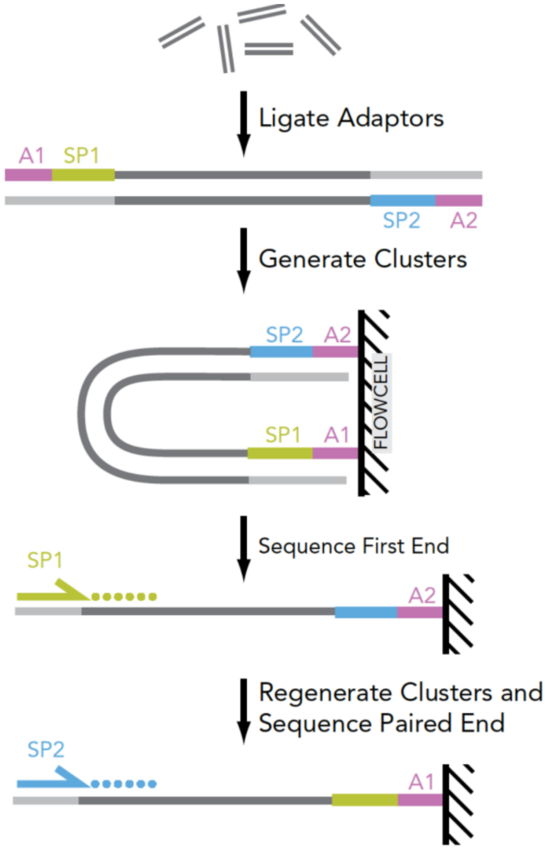
- From a DNA fragment, we can generate:

A single read from one end

Reads from both ends



Single-end read

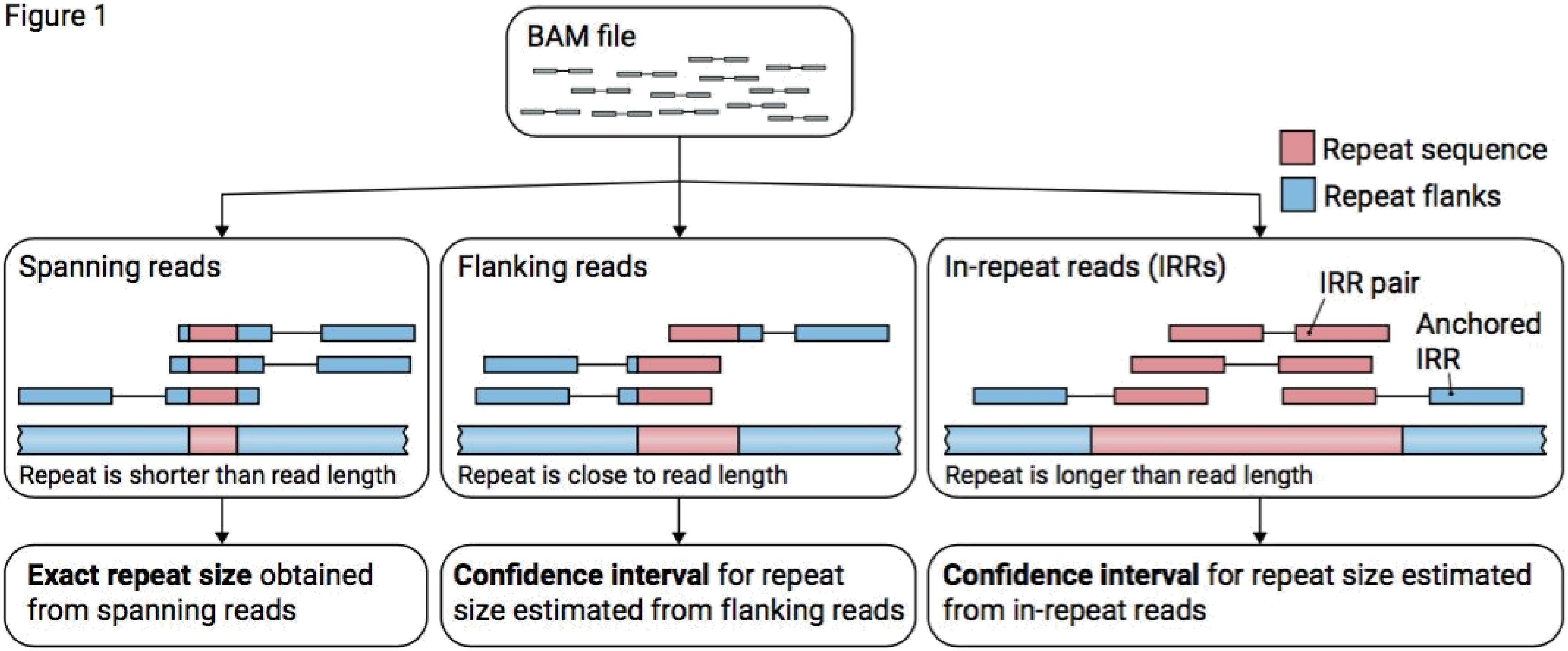


Paired-end reads

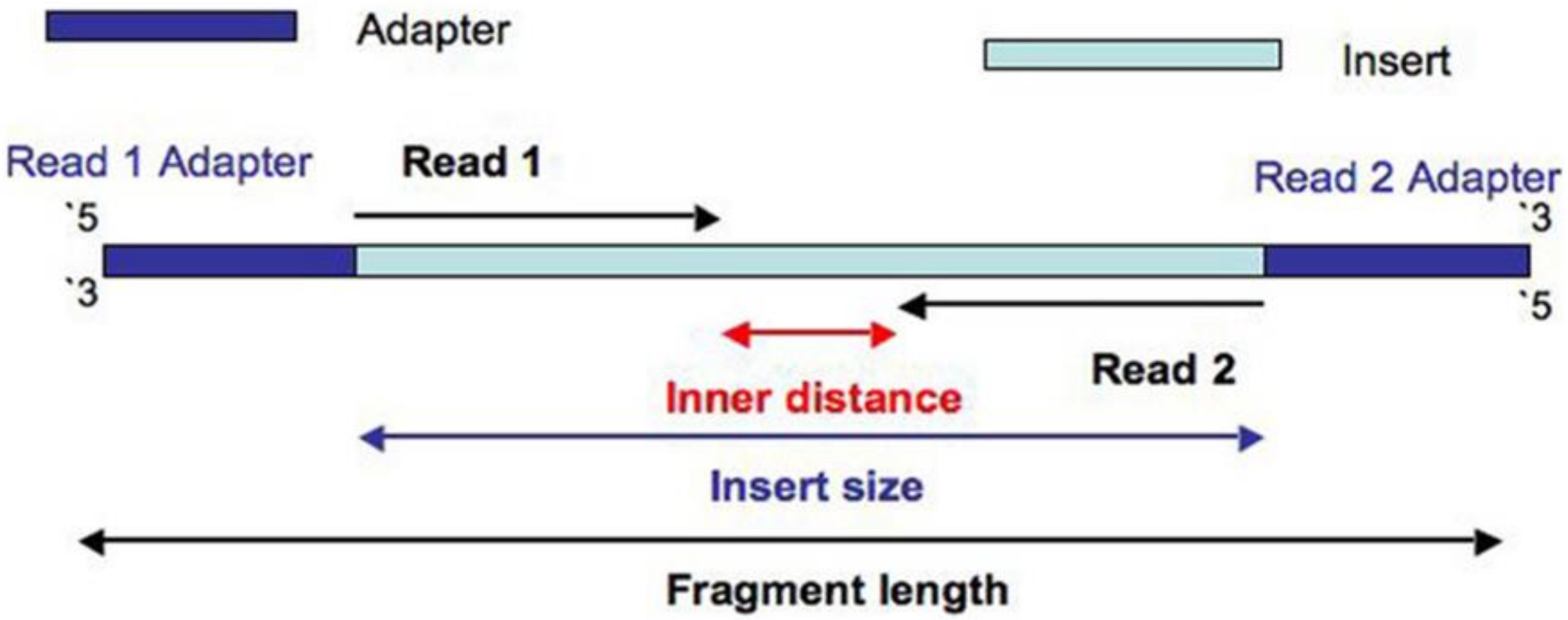
Read Length and Repeats

- Distance between paired-end reads is conserved and can be use to resolve difficult repeat/low complexity regions

Figure 1


BAM file
Repeat sequence
Repeat flanks
Spanning reads
Repeat is shorter than read length
Flanking reads
Repeat is close to read length
In-repeat reads (IRRs)
IRR pair
Anchored IRR
Repeat is longer than read length
Exact repeat size obtained from spanning reads
Confidence interval for repeat size estimated from flanking reads
Confidence interval for repeat size estimated from in-repeat reads

Fragment/Insert Size



Oxford Genomics Centre

The Wellcome Trust Centre for Human Genetics

Data processing and evaluation:

File Formats and Quality Control

Matthieu Miossec, PhD

Bioinformatics Analyst, Bioinformatics Core Group

FASTA File format

- One of the simplest sequence files for storing sequence data.
  - It contains at least one identifier line followed by a sequence (of any length).
    - It can contain several separate sequences stored one after the other.
      (e.g. human genome reference, sequences per chromosome)

Let’s take a look at an example on the cluster!

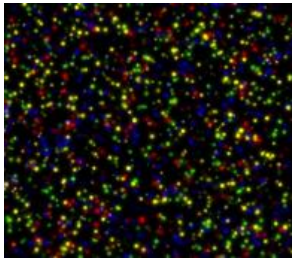
FASTQ File format

- Builds on FASTA, but crucially with an additional line for base calling quality.
  - When handling sequencing data, this is likely the first format you will encounter.
  - Base call quality for each nucleotide in the sequence is shown as ASCII values instead of a number.
    - ASCII character corresponds to the quality of the base right above it.

Let’s take a look at an example on the cluster!

Phred base Quality Score 1/2

- As mentioned earlier, the process of base calling is imperfect.



- Some of the resulting uncertainty captured by **Phred quality scores**.
  - Every base call has an estimated probability *P* of being incorrect (e.g. T called where there was a C).
  - This probability can be expressed in logarithmic form:

$$Q = -10 \log_{10} P$$

Giving us our **Phred** **<u>base</u>** **quality score**.

Phred base Quality Score 2/2

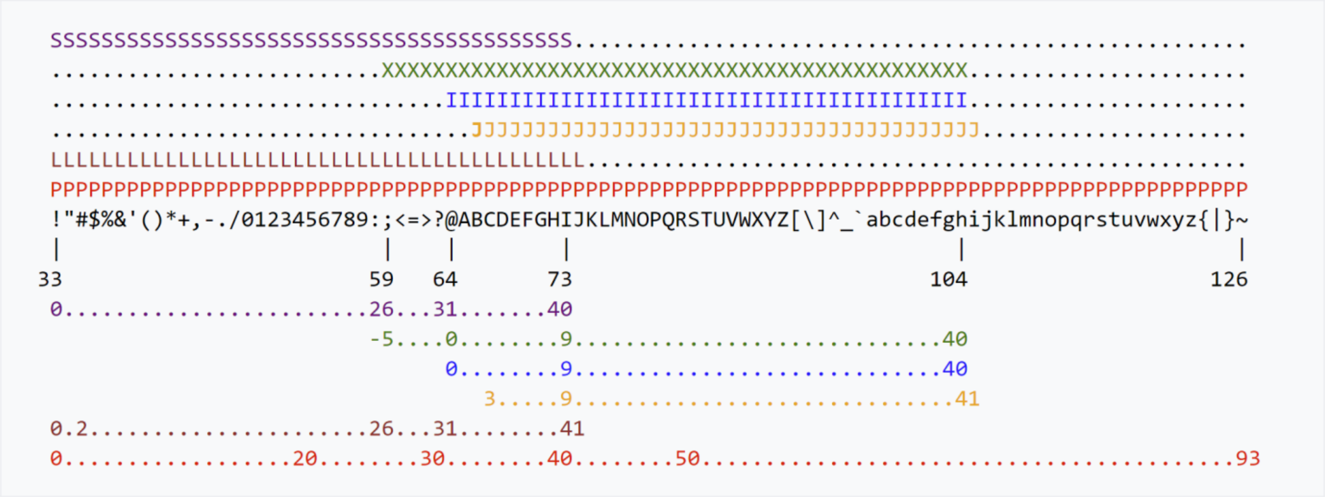
- The conversion between score and probability is fairly intuitive.

$$Q = -10 \, log_{10} P$$

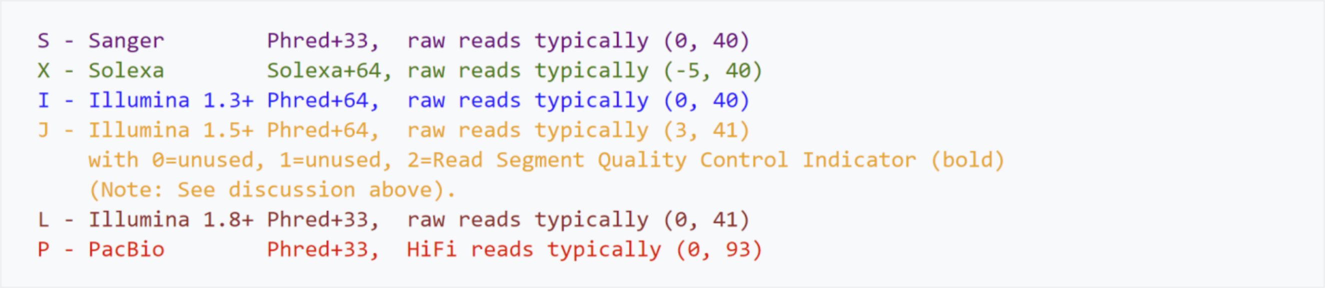
Phred quality score	Probability of incorrect base call	Base calling accuracy
10	1/10	90%
20	1/100	99%
30	1/1 000	99.9%
40	1/10 000	99.99%
50	1/100 000	99.999%

Phred+33 Encoding

- In FASTQ, sequence of Phred quality scores encoded in ASCII.
  - Phred+33 now the standard.
    (i.e. a quality of 0 encoded by ASCII symbol 33 or ‘!’)



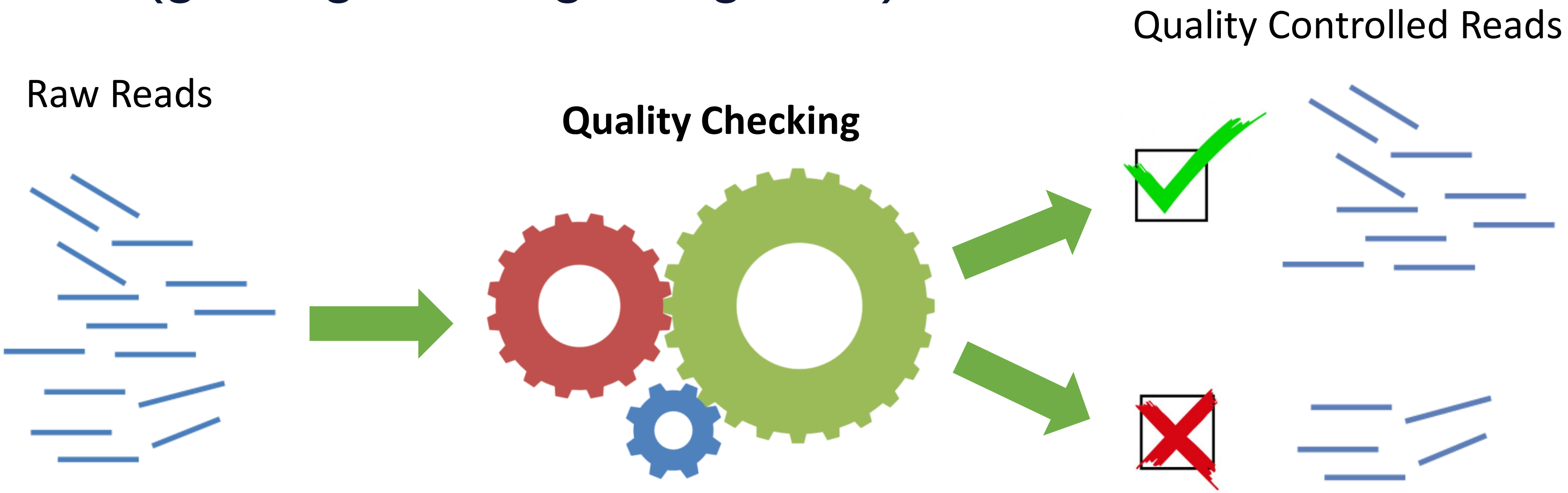
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                 0........9.............................40
                                    3.....9..............................41
  0.2......................26...31........41
  0..................20........30........40........50..........................................93
```



```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

The First Step: Quality Control

- Absolutely crucial step.
  - Bad quality data lead to dissapointing results (garbage in → garbage out).

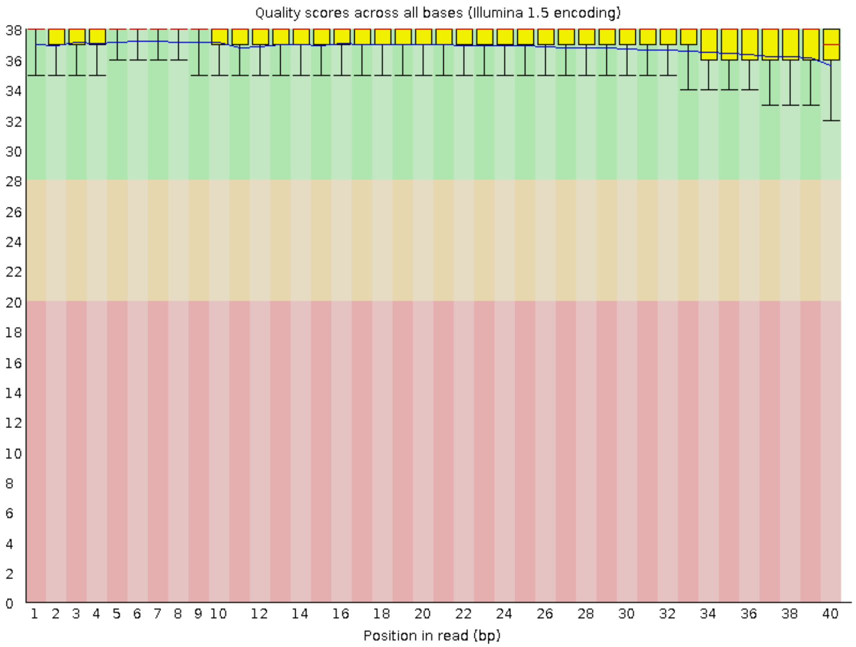


FastQC
Widely used for Illumina data because it’s fast. It works on a subset of reads.
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

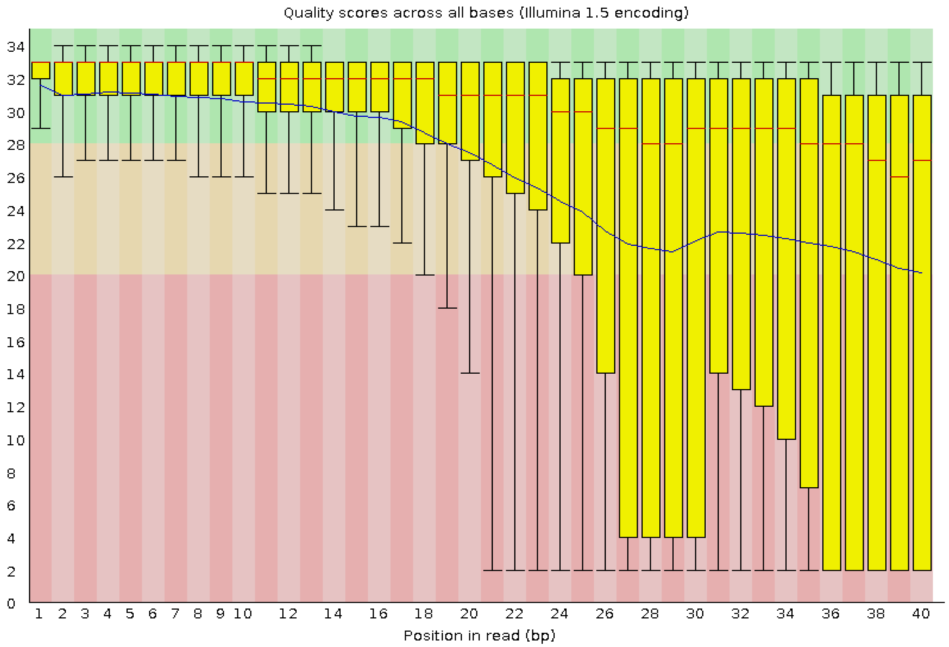
PRINSEQ
Used for smaller datasets because it computes every sequence.
http://prinseq.sourceforge.net/

Per Base Sequence Quality (FastQC)

✓



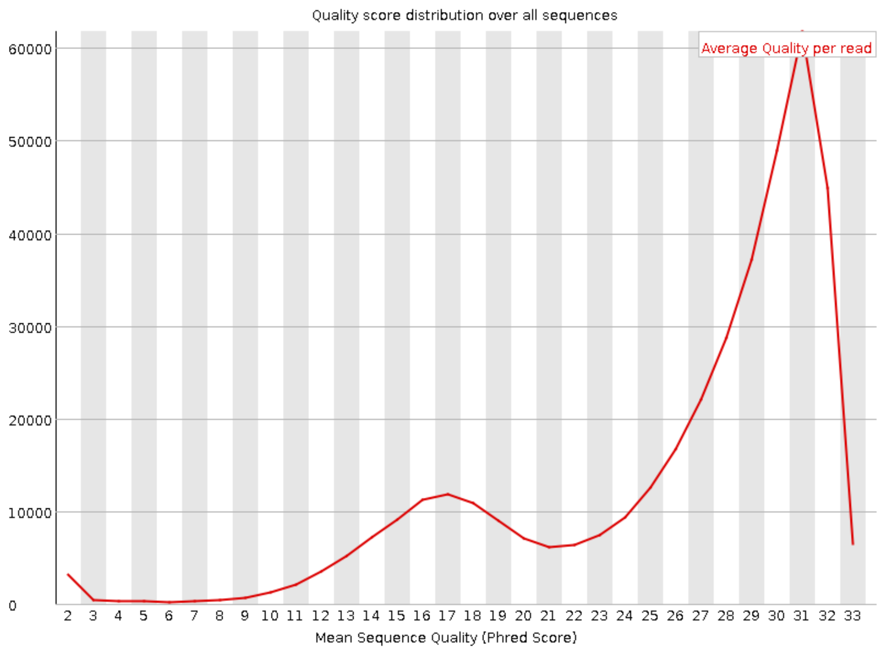
✗



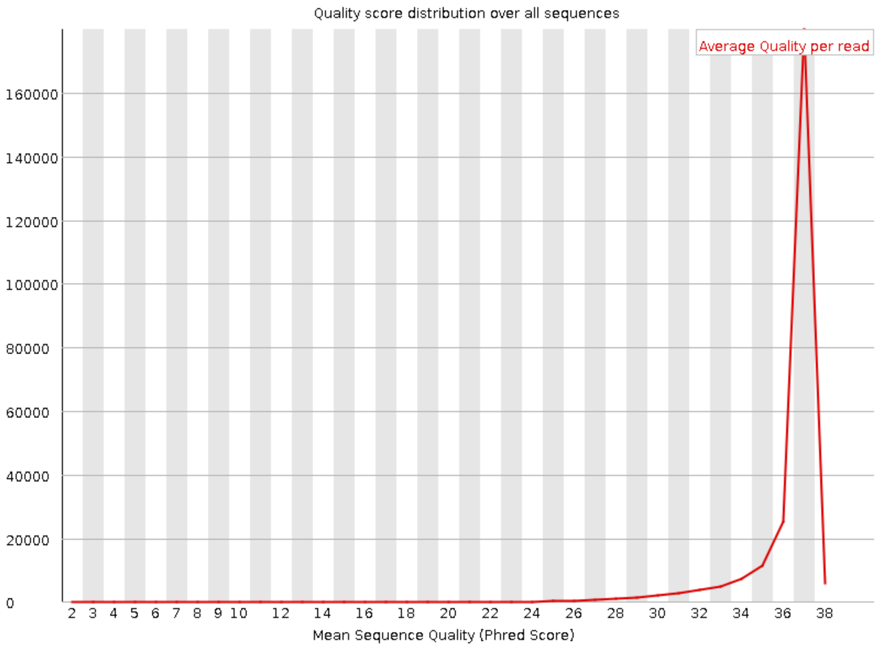
Per Sequence Quality (FastQC)

Oxford Genomics CENTRE

!



✓

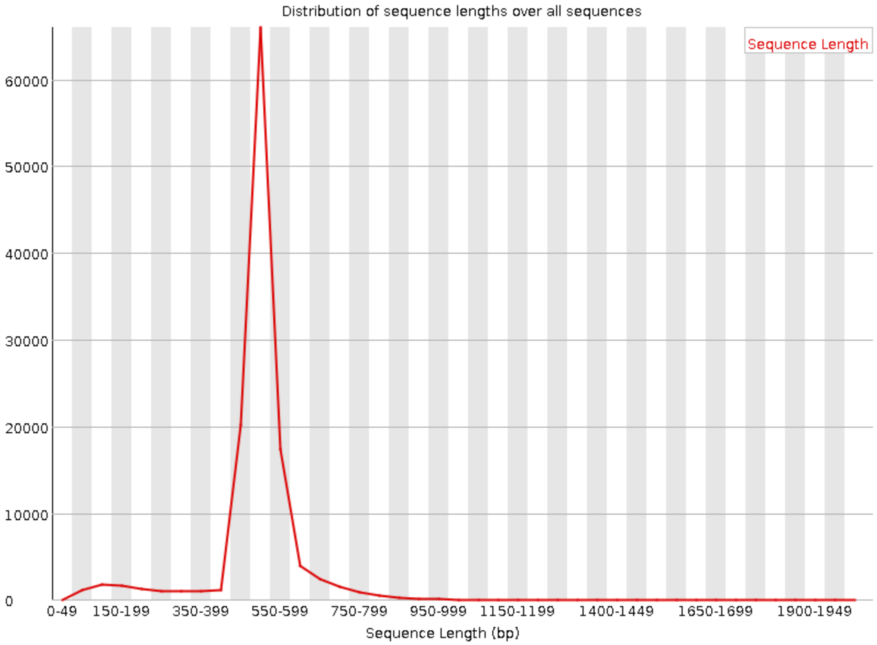


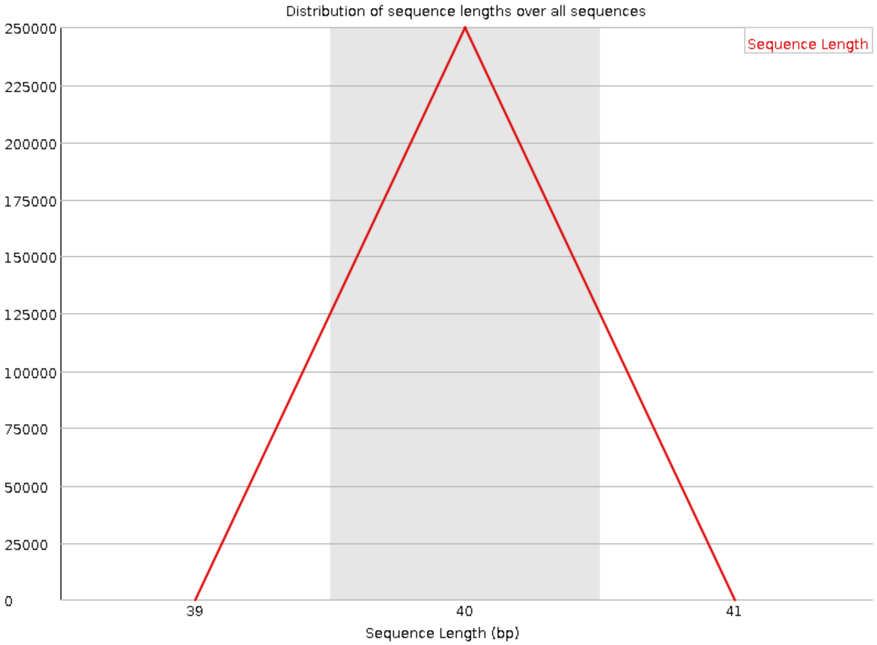
whg

UNIVERSITY OF OXFORD

Sequence Length Distribution (FastQC)

E.g. 454/Roche

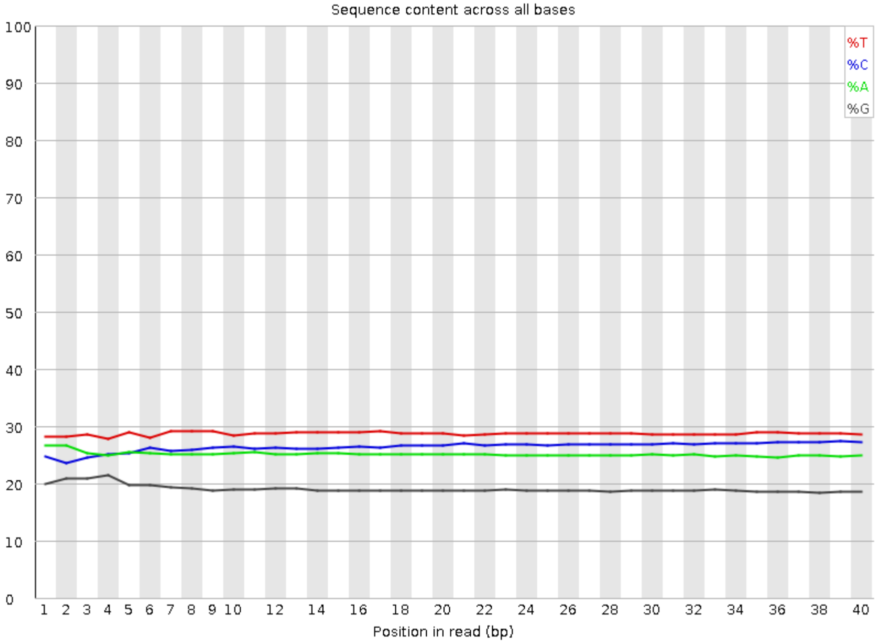




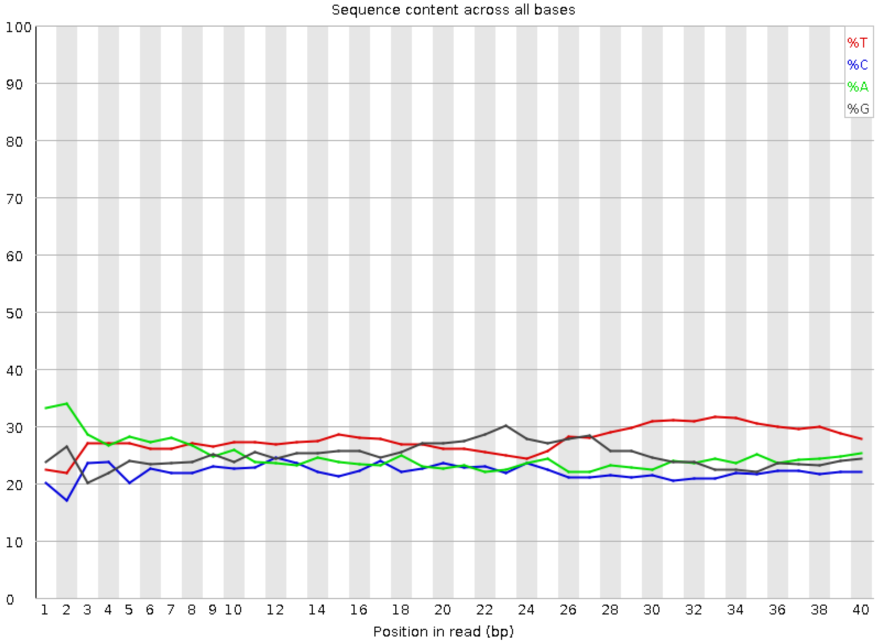
E.g. Illumina

Per Base Sequence Content (FastQC)

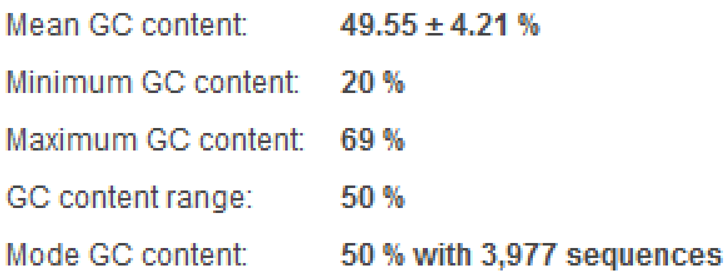
✓



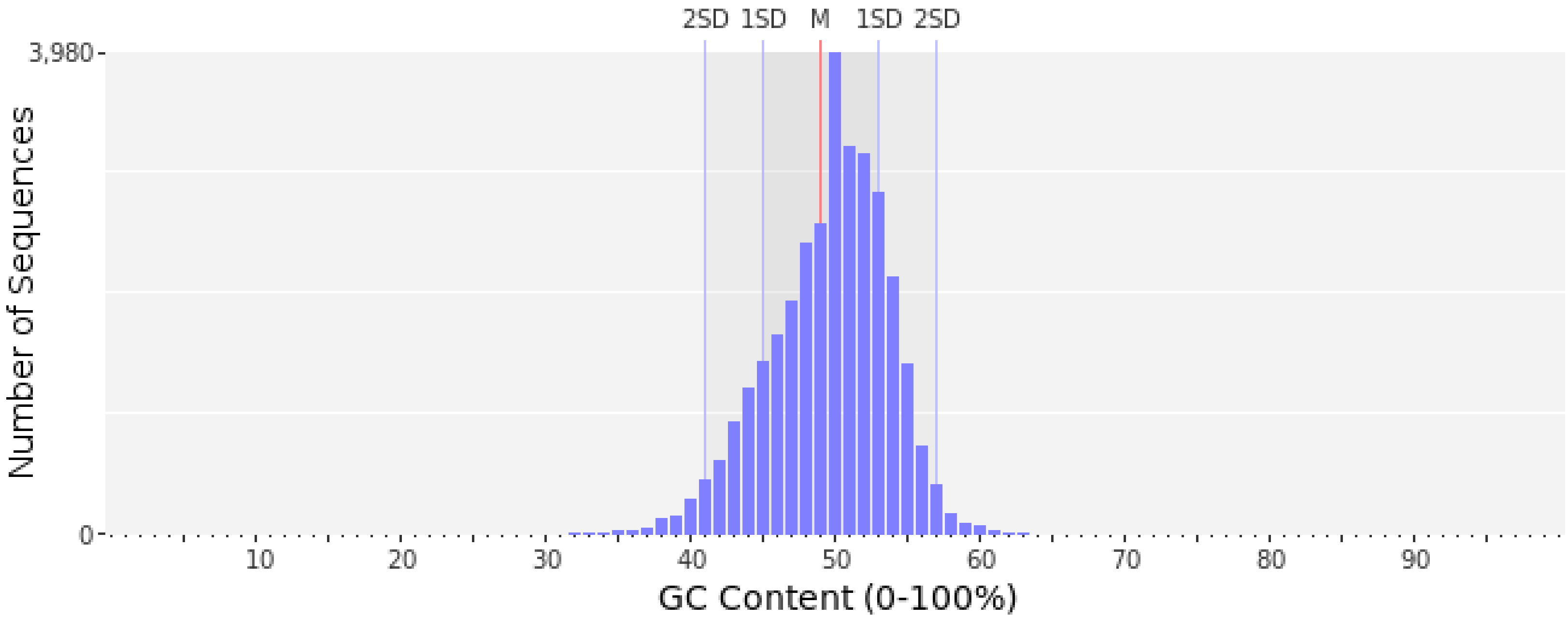
!



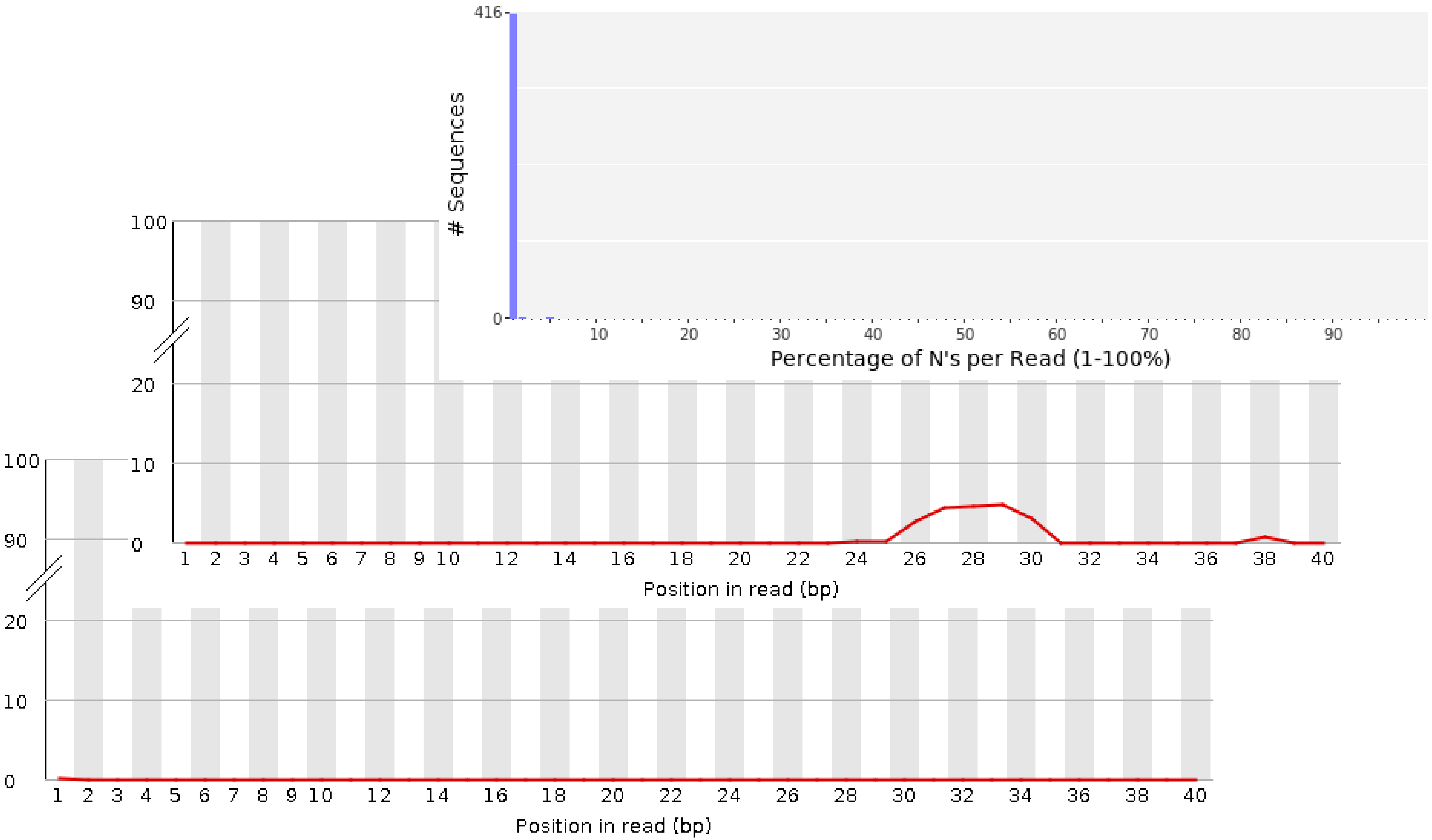
GC Content Distribution (PRINSEQ)



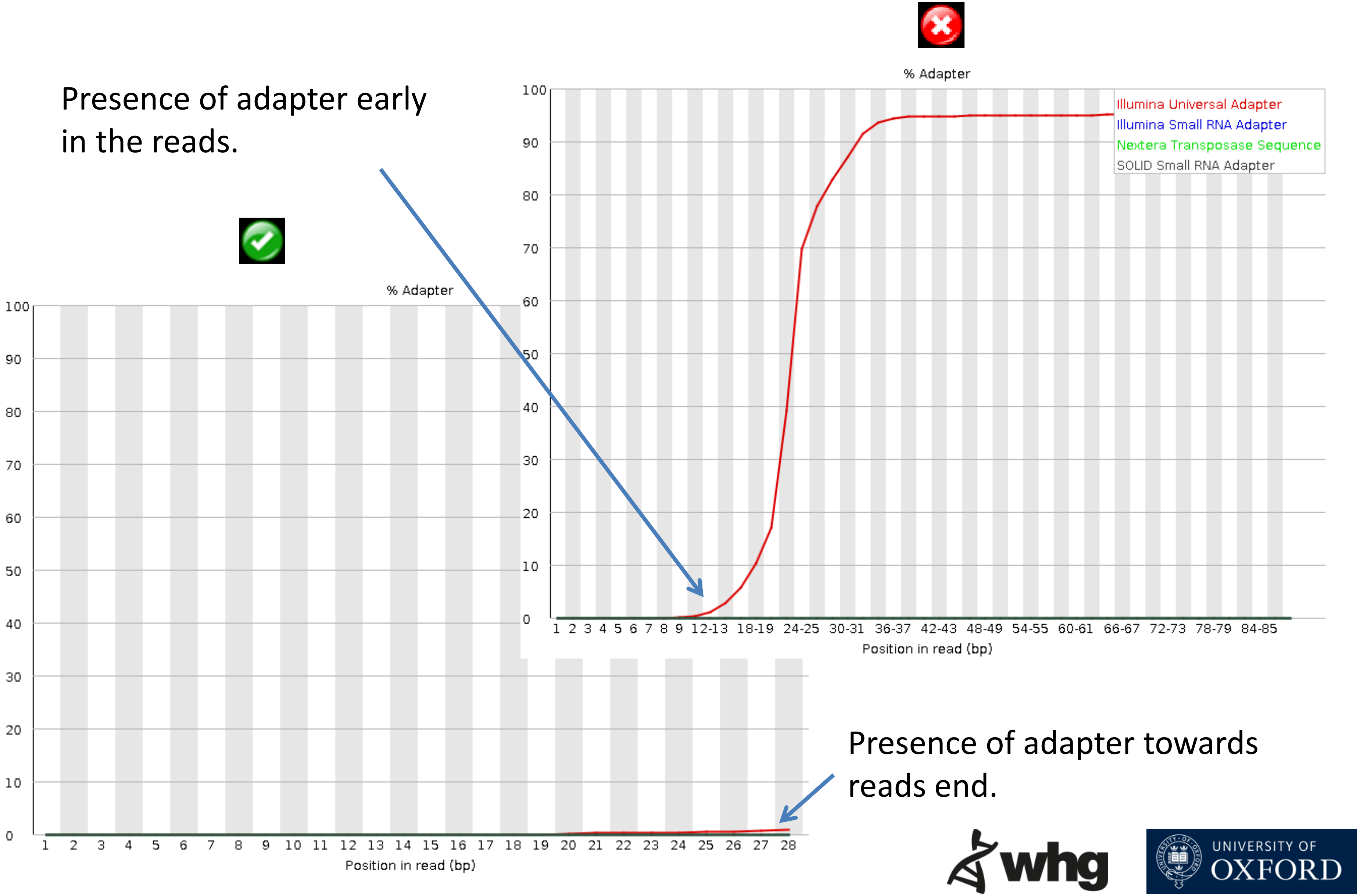
Mean GC content:	**49.55 ± 4.21 %**
Minimum GC content:	**20 %**
Maximum GC content:	**69 %**
GC content range:	**50 %**
Mode GC content:	**50 % with 3,977 sequences**



N Base Content (PRINSEQ)



Adapter Content (PRINSEQ)

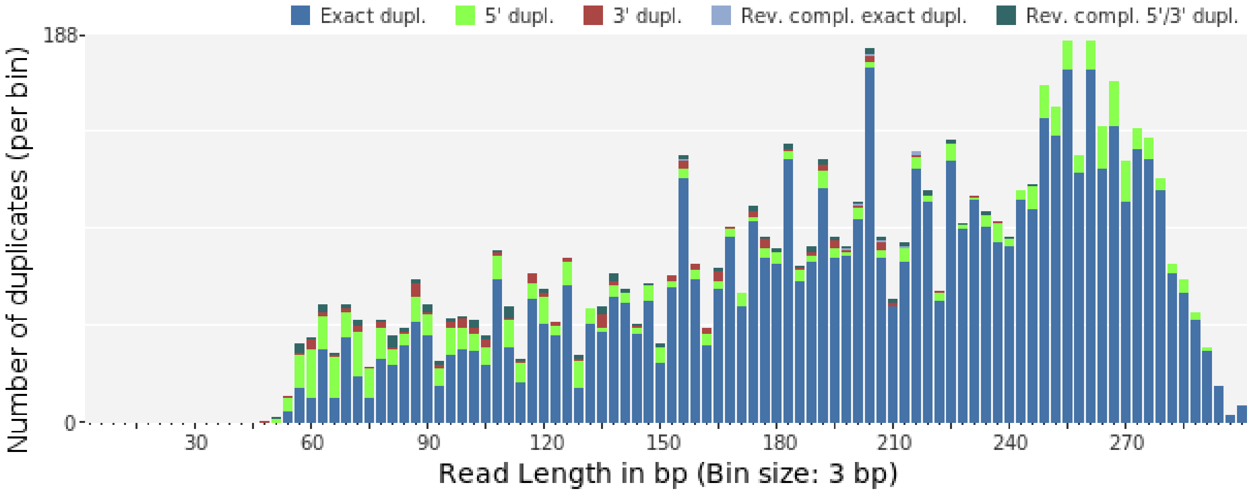


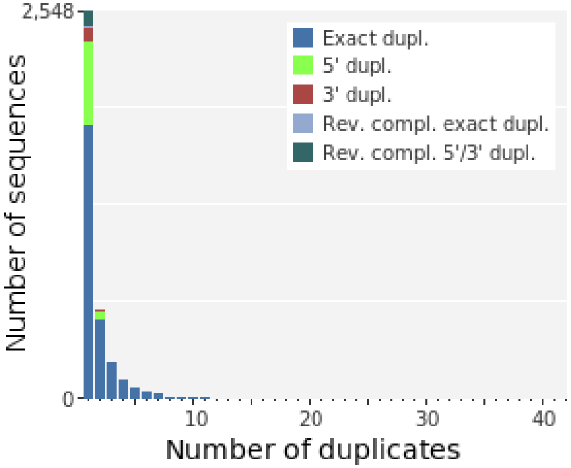
Presence of adapter early in the reads.

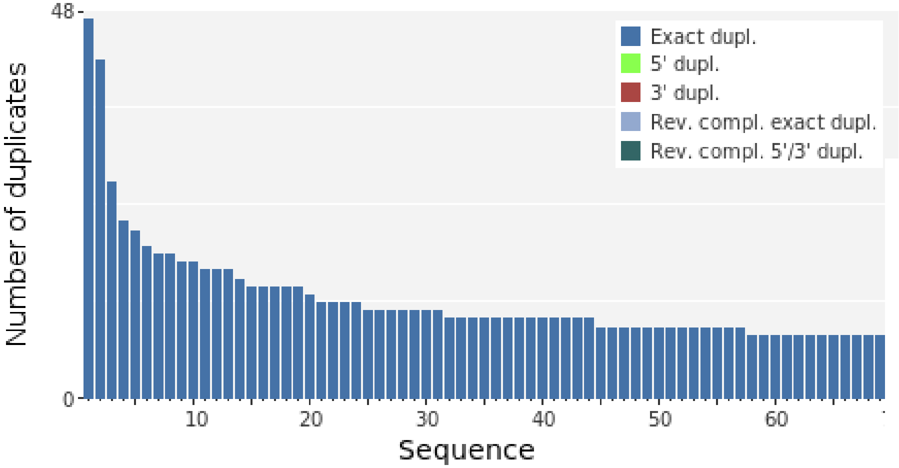
Presence of adapter towards reads end.

Sequence Duplication (PRINSEQ)

Oxford Genomics CENTRE







whg

UNIVERSITY OF OXFORD

Oxford Genomics Centre
The Wellcome Trust Centre for Human Genetics

Data processing and evaluation:

Sequence Alignment

Matthieu Miossec, PhD
Bioinformatics Analyst, Bioinformatics Core Group

Alignment to a Reference

- The first genome of a species has to be assembled from scratch (*de novo* assembly), a computationally intensive operation.
  - Luckily, once a (reference) genome exists, we can align/map any new individuals of the same species to that genome.
    - Many reads will align perfectly due to strong intra-species similarities.
    - Many other reads will only deviate from the reference by one or two bases.
- The human genome reference can be found here: http://hgdownload.soe.ucsc.edu/downloads.html#human

Challenges of Alignment

- Despite being less computationally demanding than *de novo* assembly, aligning to a reference does have its challenges:
  - A lot of short reads to align to an entire genome (a data structure like a hash table and some dynamic programming is needed to do this fast).
  - The presence of both errors and individual variation complicates alignment process.
  - Low complexity and repetitive regions are difficult to align to (paired-end reads help).

Types of deviation

Oxford Genomics CENTRE

- Errors and variation come in different forms. For alignment, we don’t need to distinguish the two (that’s what variant calling is for).
  - Single nucleotide alteration. source: variant (SNV) or error.

    Ref …ATGATGCCATGACTGACCCTGAT…

    Read …ATGATGCCATGACTGAC**A**CTGAT…

  - Insertion. source: real insertion or error.

    Ref …TCCATGTGTGACTA******CACC…

    Read …TCCATGTGTGACTATTTGTCACC…

  - Deletion. source: real insertion or error.

    Ref …AAACTTAGTGCAACAGTGCACGAG…

    Read …AAAC**AGTGCAACAGTGCACGAG…

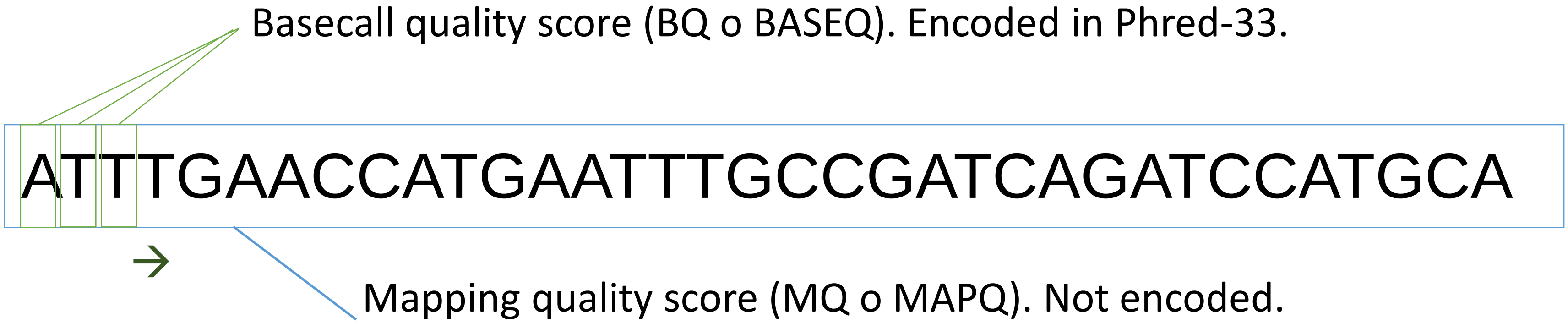
Referred together as indels

whg

UNIVERSITY OF OXFORD

Phred Quality Score Revisited

- Phred quality scores are also used to quantify mapping uncertainty.
  - Mapping quality is apply to a single read rather than individual bases.

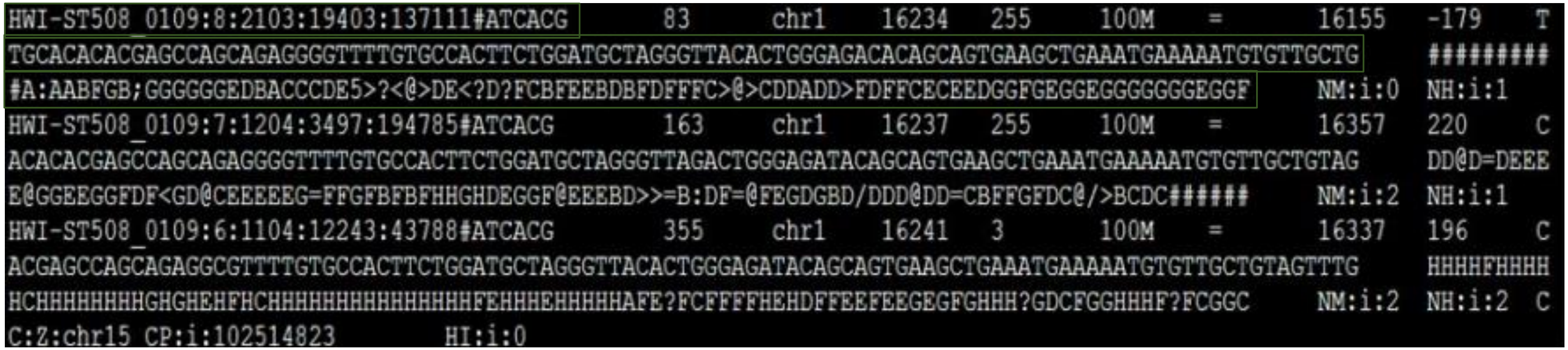


Basecall quality score (BQ o BASEQ). Encoded in Phred-33.
ATTTGAACCATGAATTTGCCGATCAGATCCATGCA
→
Mapping quality score (MQ o MAPQ). Not encoded.

- We also need to account for insertions and deletions in relation to the reference. This is done using a CIGAR.

Sequence Alignment/Mapping Format

- Most popular fast-aligners (e.g. BWA, Bowtie2) take FASTQ as input and produce SAM/BAM files.
    - The SAM format complements information from the FASTQ file with alignment information (i.e. position relative to reference, quality, presence of indels).



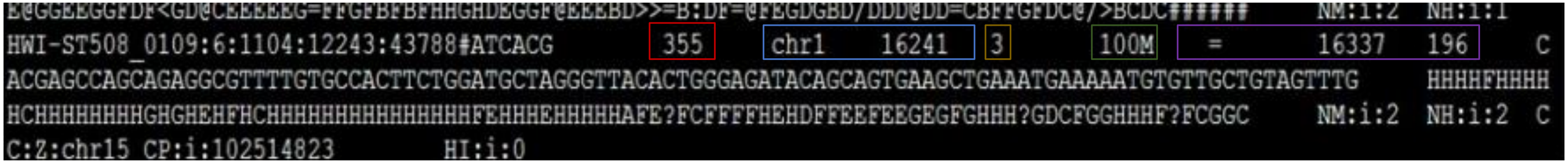
```
HWI-ST508_0109:8:2103:19403:137111#ATCACG        83      chr1    16234   255     100M    =       16155   -179    T
TGCACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGACACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTG      #########
#A:AABFGB;GGGGGGEDBACCCDE5>?<@>DE<?D?FCBFEEBDBFDFFFC>@>CDDADD>FDFFCECEEDGGFGEGGEGGGGGGGEGGF      NM:i:0  NH:i:1
HWI-ST508_0109:7:1204:3497:194785#ATCACG         163     chr1    16237   255     100M    =       16357   220     C
ACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTAGACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAG      DD@D=DEEE
E@GGEEGGFDF<GD@CEEEEEG=FFGFBFBFHHGHDEGGF@EEEBD>>=B:DF=@FEGDGBD/DDD@DD=CBFFGFDC@/>BCDC######      NM:i:2  NH:i:1
HWI-ST508_0109:6:1104:12243:43788#ATCACG         355     chr1    16241   3       100M    =       16337   196     C
ACGAGCCAGCAGAGGCGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAGTTTG      HHHHFHHHH
HCHHHHHHHHGHGHEHFHCHHHHHHHHHHHHHHHHHFEHHHEHHHHHAFE?FCFFFFHEHDFFEEFEEGEGFGHHH?GDCFGGHHHF?FCGGC      NM:i:2  NH:i:2  C
C:Z:chr15 CP:i:102514823        HI:i:0
```

Let’s take a look at an example on the cluster!

(Using SAMtools)

Alignment Information

- Here is what information a SAM/BAM file contains about the alignment.

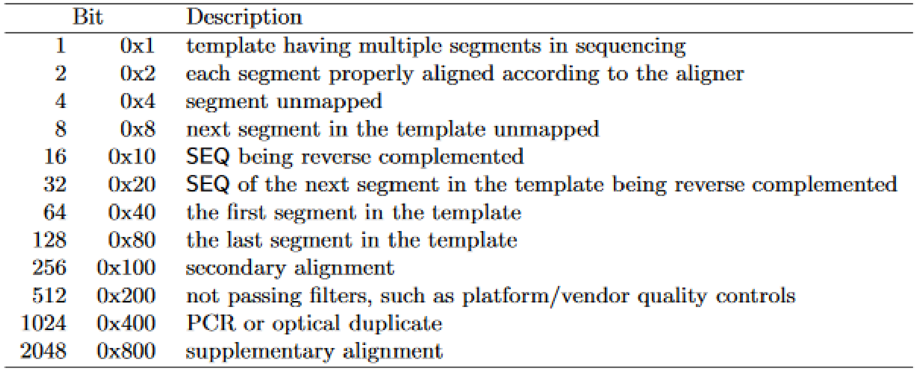


```
E@GGEEGGFDF<GD@CEEEEEG=FFGFBFBFHHGHDEGGF@EEEBD>>=B:DF=@FEGDGBD/DDD@DD=CBFFGFDC@/>BCDC######      NM:i:2  NH:i:1
HWI-ST508_0109:6:1104:12243:43788#ATCACG        355     chr1    16241   3       100M    =       16337   196     C
ACGAGCCAGCAGAGGCGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAGTTTG       HHHHFHHHH
HCHHHHHHHHGHGHEHFHCHHHHHHHHHHHHHHHFEHHHEHHHHHAFE?FCFFFFHEHDFFEEFEEGEGFGHHH?GDCFGGHHHF?FCGGC      NM:i:2  NH:i:2  C
C:Z:chr15 CP:i:102514823        HI:i:0
```

- Position in the reference where read is aligned (chromosome and locus) and the other half of the base pair (including relative to the first half).
- Mapping quality (MAPQ).
- The CIGAR (Concise Idiosyncratic Gapped Alignment Report)
- A Bitwise flag for additional information about the read.

Bitwise flag and CIGAR

- Crucial bits of information about a given read can be stored in a single bit.



Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

The bit is a sum of the statements that are true about a read (e.g. 1033 corresponds to 1, 8 and 1024).

- A CIGAR signals where a reads needs an insertion/deletion to match the reference.

40M5I30M2D25M (agnostic about single nucleotide alterations)

ATGATGCCATGACTGACCCTGATGGTCCATGTGTGACTA*****CACCACATGCTGGATAGGTGCCCGTGAAACTTAGTGCAACA
GTGCACGAGATGAGGAGTG

ATGATGCCATGACTGACCCTGATGGTCCATGTGTGACTATTTGTCACCACATGCTGTATAGGTGCCCGTGAAAC**AGTGCAACA
GTGCACGAGATGAGGAGTG

Oxford Genomics Centre

Oxford Genomics Centre
The Wellcome Trust Centre for Human Genetics

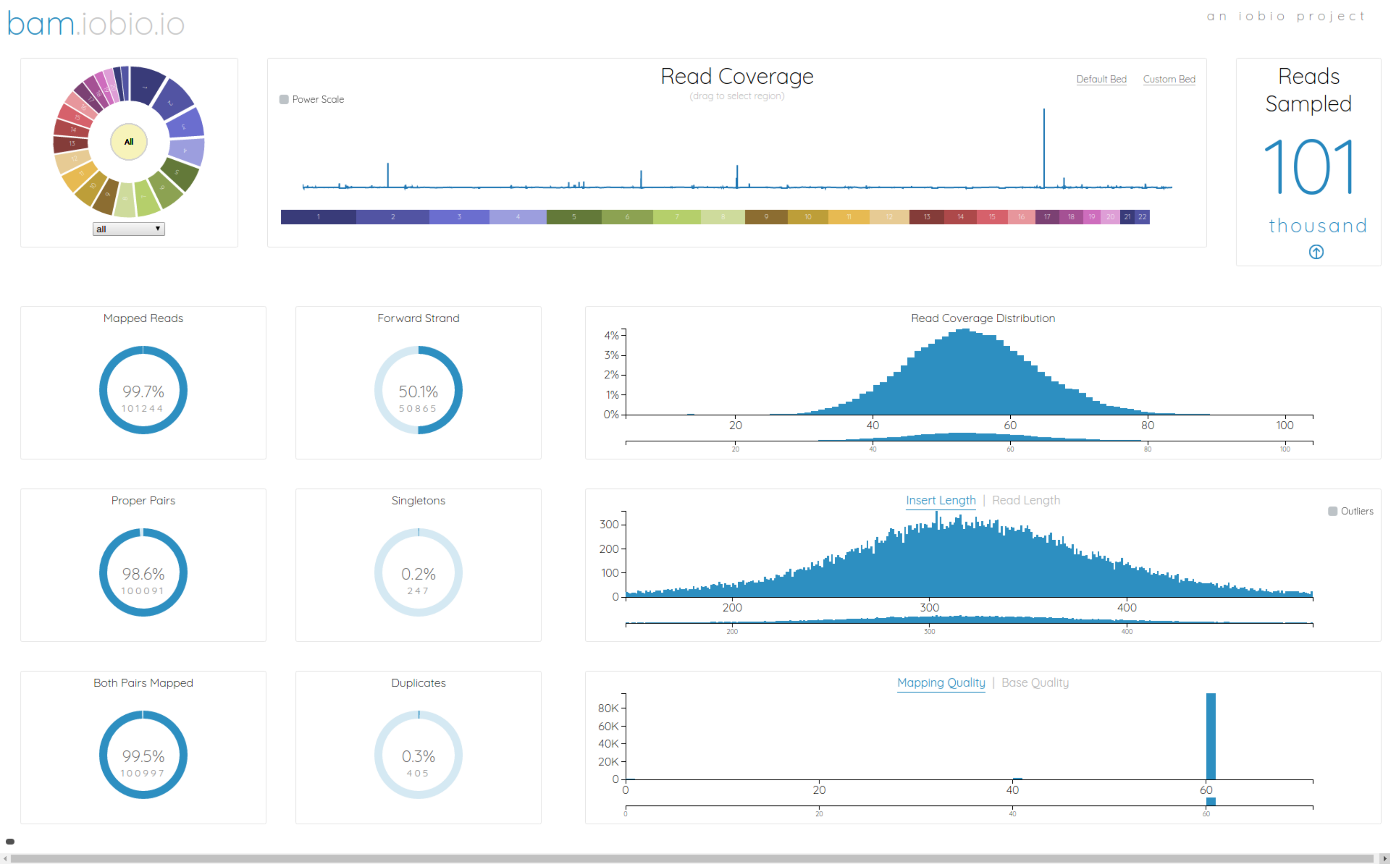
Data processing and evaluation:
Inspecting Alignment Data

Matthieu Miossec, PhD
Bioinformatics Analyst, Bioinformatics Core Group

whg

UNIVERSITY OF OXFORD

Getting Alignment Stats with bam.iobio.io



bam.iobio.io

an iobio project

Read Coverage (drag to select region)

Reads Sampled 101 thousand

Mapped Reads 99.7% 101244

Forward Strand 50.1% 50865

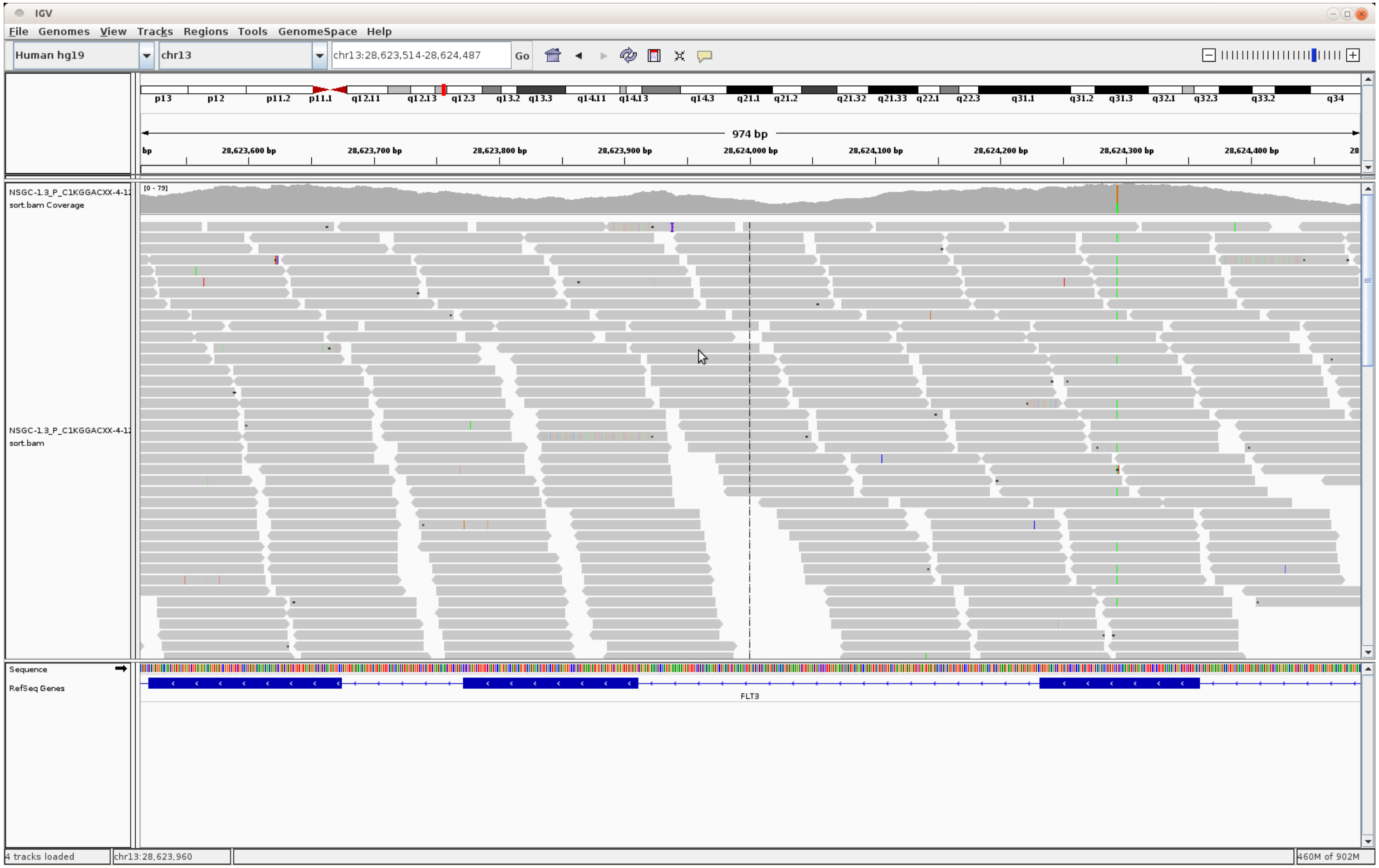
Proper Pairs 98.6% 100091

Singletons 0.2% 247

Both Pairs Mapped 99.5% 100997

Duplicates 0.3% 405

Checking a locus with IGV



Human hg19

chr13

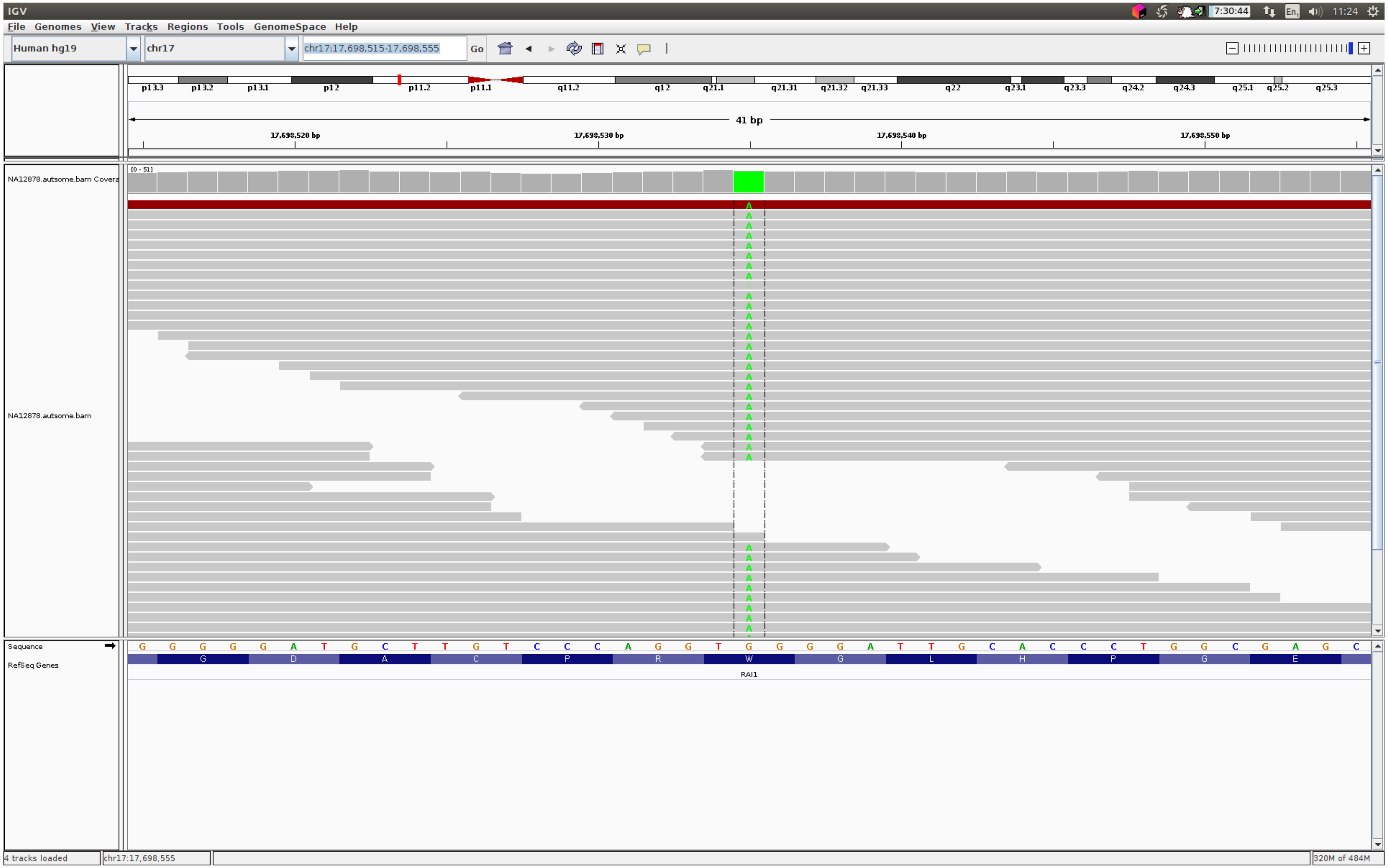
chr13:28,623,514-28,624,487

4 tracks loaded

chr13:28,623,960

460M of 902M

Checking a locus with IGV



Human hg19

chr17

chr17:17,698,515-17,698,555

4 tracks loaded

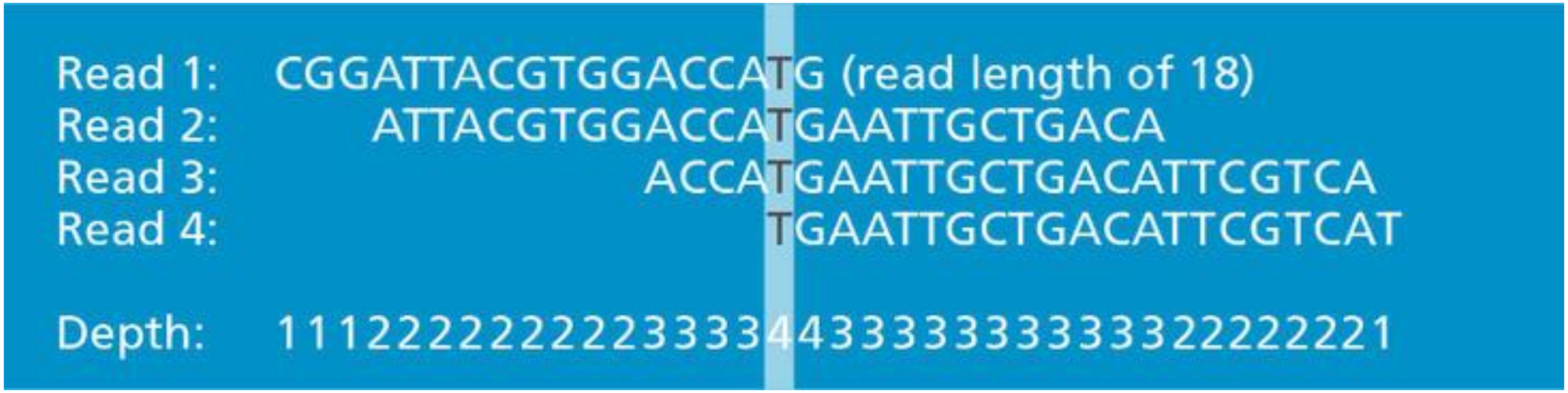
chr17:17,698,555

320M of 484M

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Go
7:30:44
11:24
p13.3 p13.2 p13.1 p12 p11.2 p11.1 q11.2 q12 q21.1 q21.31 q21.32 q21.33 q22 q23.1 q23.3 q24.2 q24.3 q25.1 q25.2 q25.3
41 bp
17,698,520 bp
17,698,530 bp
17,698,540 bp
17,698,550 bp
[0 - 51]
NA12878.autsome.bam Covera
NA12878.autsome.bam
A
Sequence
RefSeq Genes
G G G G G A T G C T T G T C C C A G G T G G G G A T T G C A C C C T G G C G A G C
G D A C P R W G L H P G E
RAI1

Read Depth/Coverage

- While base call accuracy for NGS platforms is only ~99.9%*, the high throughput nature of these platforms make it so that a region or locus can be covered by multiple reads.



```
Read 1:   CGGATTACGTGGACCATG (read length of 18)
Read 2:      ATTACGTGGACCATGAATTGCTGACA
Read 3:               ACCATGAATTGCTGACATTCGTCA
Read 4:                   TGAATTGCTGACATTCGTCAT

Depth:    1112222222223333443333333333322222221
```

  - The more reads cover a particular locus, the more accurately bases that deviate from a reference (variants) can be distinguished from error.

*Remember, Sanger has a 99.999% base call accuracy.

Questions

Thanks for listening

Any questions?