UNIVERSITY OF OXFORD

Next-Generation Sequencing Analysis

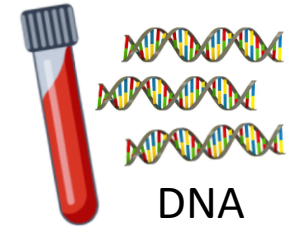
QC, Alignment and Variant Calling

Matthieu Miossec, PhD
Bioinformatics Analyst @ Bioinformatics Core

Sequencing At a Glance

UNIVERSITY OF OXFORD

- The process of determining a sequence of bases and producing a digital representation (aka. a **read** sequence) for further analysis.



DNA

RNA ➢cDNA

A FASTQ file (more on those later…)

```
@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 an ideal world, we could sequence <u>entire</u> chromosomes <u>quickly</u> and get a <u>single</u> <u>error-free</u> readout for each chromosome as our output.

This is not the world we live in…

Innovations of NGS (Recap)

UNIVERSITY OF OXFORD

- Each NGS platform uses distinct biochemical processes for sequencing, but all share important attributes:
  - These processes are happening simultaneously for distinct fragments ➔ massively parallel.
  - PCR amplification is used to turn weak bioluminescent signals generated by a small fragment, into the strong signal of a cluster of ~1000 identical fragments.
  - Sequencing-by-synthesis (SBS)

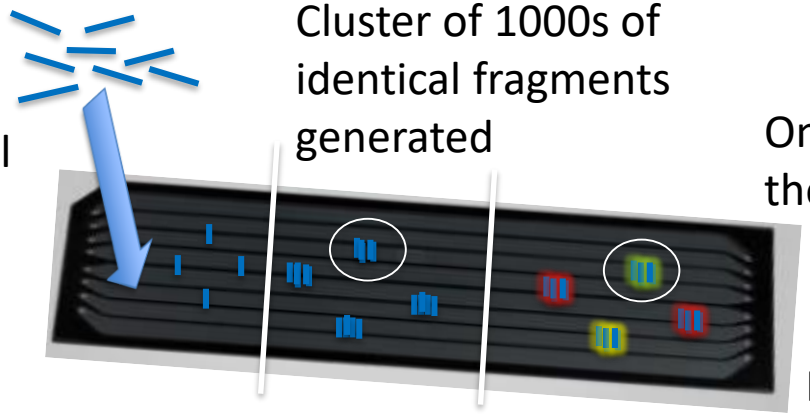
Sequencing-by-Synthesis (e.g. Illumina)

UNIVERSITY OF OXFORD

Fragments run over 2D flowcell

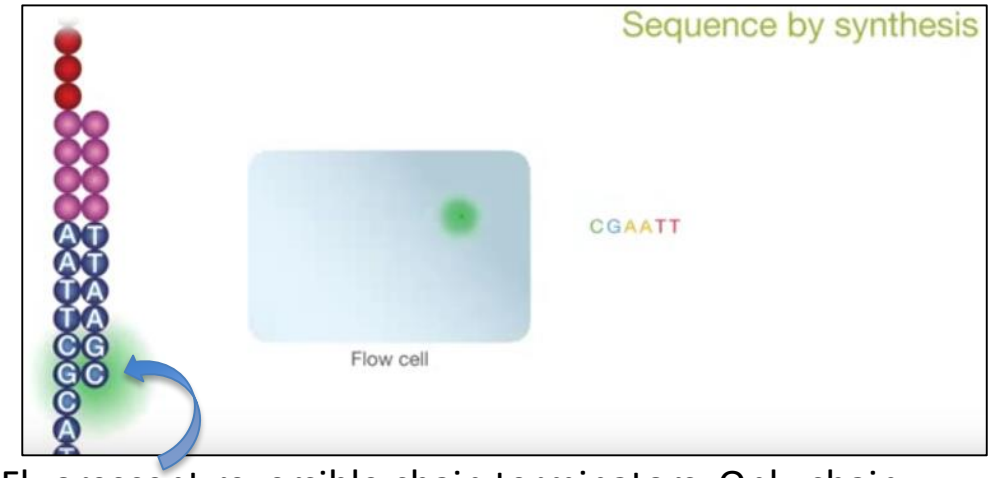
Cluster of 1000s of identical fragments generated

One base incorporation (in theory) per cycle.

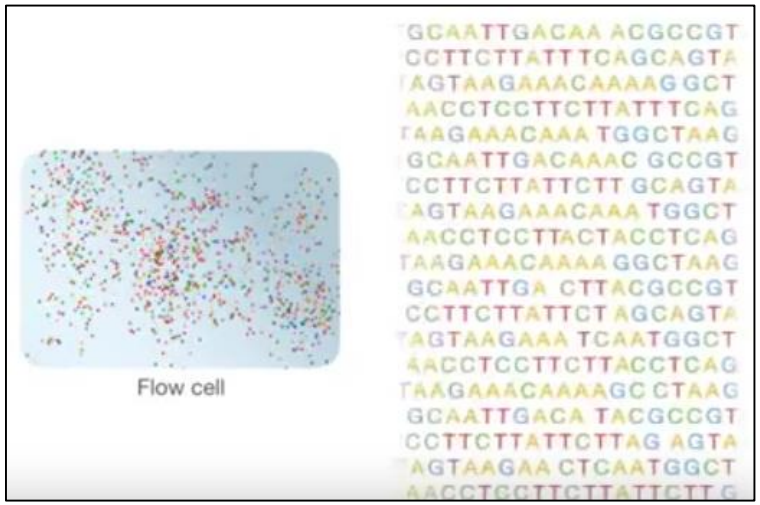


Flowcell

Captured with highly sensitive camera.



Fluorescent reversible chain terminators. Only chain terminating during a cycle, long enough for capture.

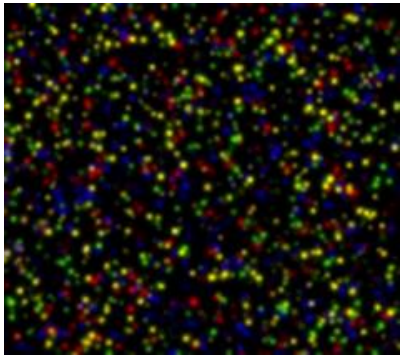


Flowcell during a given cycle. Multiple reads built in parallel.

https://youtu.be/fCd6B5HRaZ8

Limitations of NGS (short-read)

- New biochemical processes place constraints on **read length** and **accuracy**.
  - Determining the bases of a sequence (aka. **base calling**) via bioluminescence has several pitfalls:



- ❖ Signals from clusters in close proximity interfere with on another.
- ❖ Synchronicity between strands within a cluster is gradually lost with each cycle (size of reads ≡ number of cycles).
- ❖ The intensity of a signal naturally varies.
- ❖ A signal can be ambiguous where bases repeat (e.g. Did the machine detect C-C- or C-C-C- ?).

- Another name for NGS: short read sequencing.

Accuracy from Redundancy

- Base call accuracy for NGS platforms is ~99.9% (~1 error every 1000 bases).
  - Far lower than Sanger method's 99.999%...however…
  - The high throughput nature of NGS platforms means most regions or loci are covered repeatedly by multiple reads. ➔ read depth.

reads

AATAGGCGTACTTC
ATAGGCGTACTTCA
AGGCCTACTACATC

Real or sequencing error?

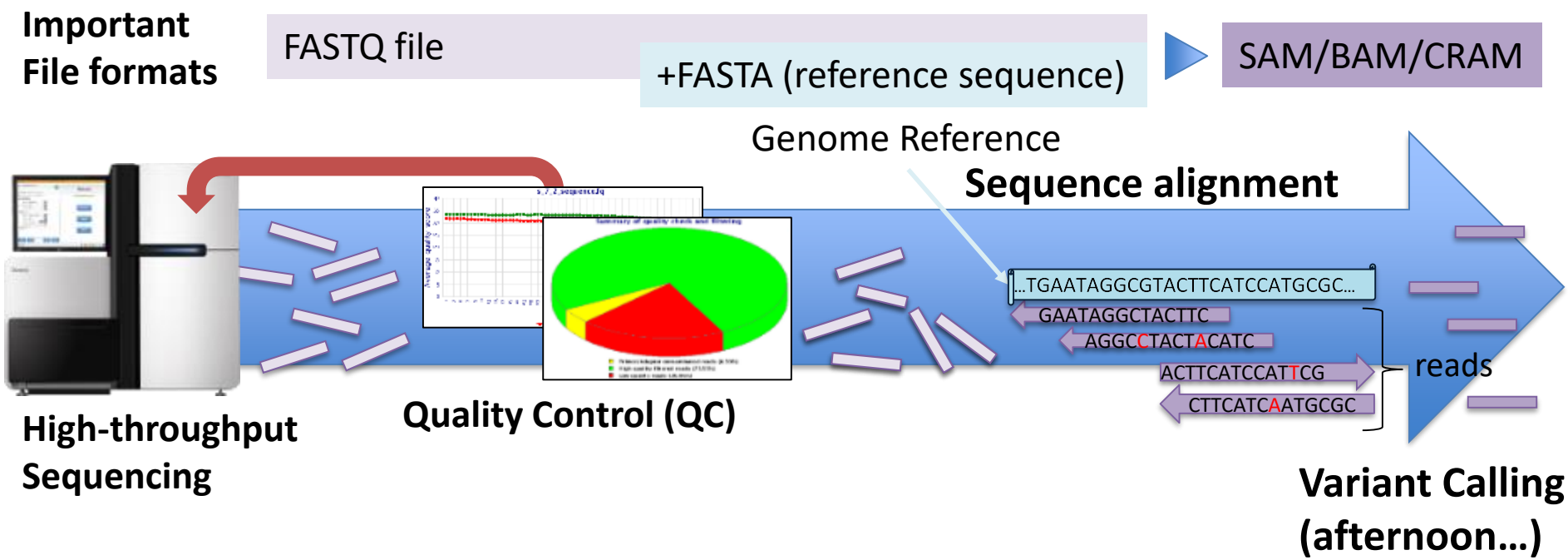
Higher **read depth**
=
Higher certainty

AATAGGCGTACTTC
ATAGGCGTACTTCA
AGGCCTACTACATC
GCCTACTTCATCCA
CTACTTCAGCCATG
TACTTCATCCATGC
…

Quality Control, Alignment and File Formats

UNIVERSITY OF OXFORD

- The first time you actually look at your data, it will most likely be in the FASTQ format.
  - Quality control and alignment performed on FASTQ.



File Format: FASTA

UNIVERSITY OF OXFORD

- The simplest sequence file formats for storing sequence data (ext: .fasta, .fa…).
  - Contains <u>at least</u> one identifier line followed by a sequence (A,T,G,Cs…and N) of any given length.
  - One file can contain several separate sequences stored one after the other, each with its own identifier (e.g. human genome reference, with a sequence per chromosome)

Let's have a look at one recurring example:

The Latest Human Genome Reference
GRCh38

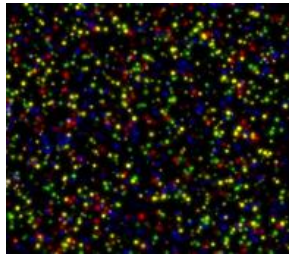
File format: FASTQ

UNIVERSITY OF OXFORD

- Builds on FASTA, but crucially adds a line for base call qualities (ext: .fastq, .fq…).
  - Base call quality is shown as a sequence of ASCII values, each single character value representing a (typically) double-digit number.
    - Each value is the quality of the base directly above it in the DNA sequence.
  - When handling sequencing data, this is likely the first file format you will encounter.
    - Each sequence is a read from the sequencer

Let's take a look at another example!

Phred base Quality Score 1/2



- The process of base calling is imperfect. The way we quantify some of that uncertainty is using **Phred quality scores**.
  - Each base call has an estimated probability $\boldsymbol{P}$ of being called incorrectly.
    (e.g. a *T* is called where a *C* should have been called)
  - These probabilities can be expressed in logarithmic form:

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

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

Phred base Quality Score 2/2

UNIVERSITY OF OXFORD

- 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

- These Phred quality scores are present in FASTQ files encoded in ASCII+(index for quality 0).
  - Phred+33 now the standard.
    (i.e. 0 is encoded by ASCII symbol 33 or '!')

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ................................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ...................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 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)
```

Quality Control

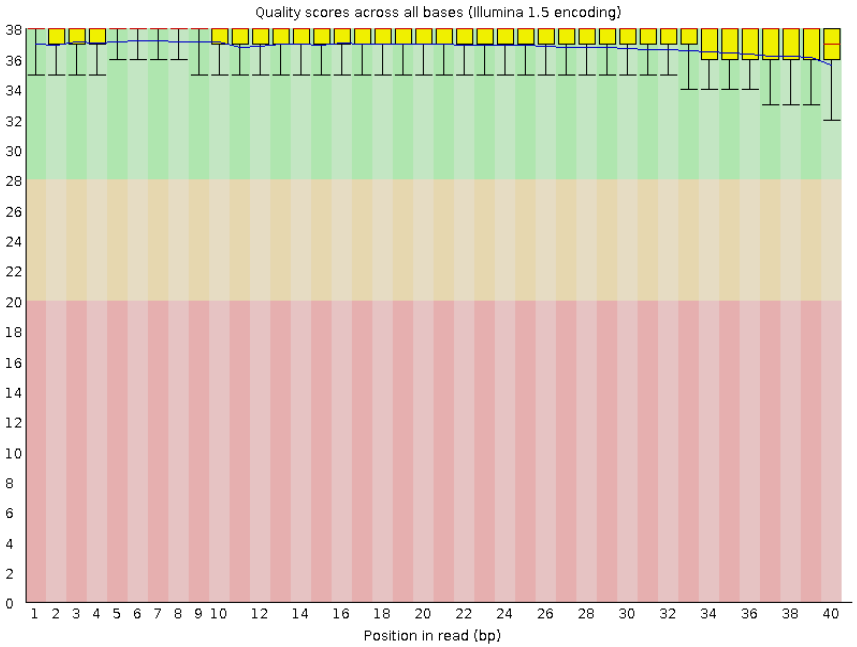
- An absolutely crucial, unescapable step.
  - Bad quality data lead to disappointing results (garbage in → garbage out).

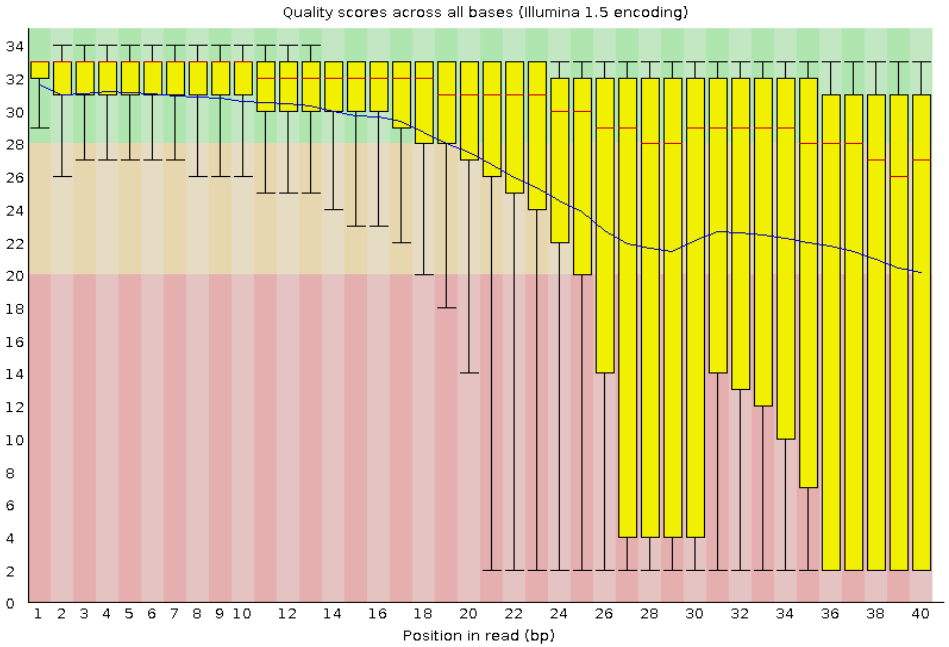
FastQC

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

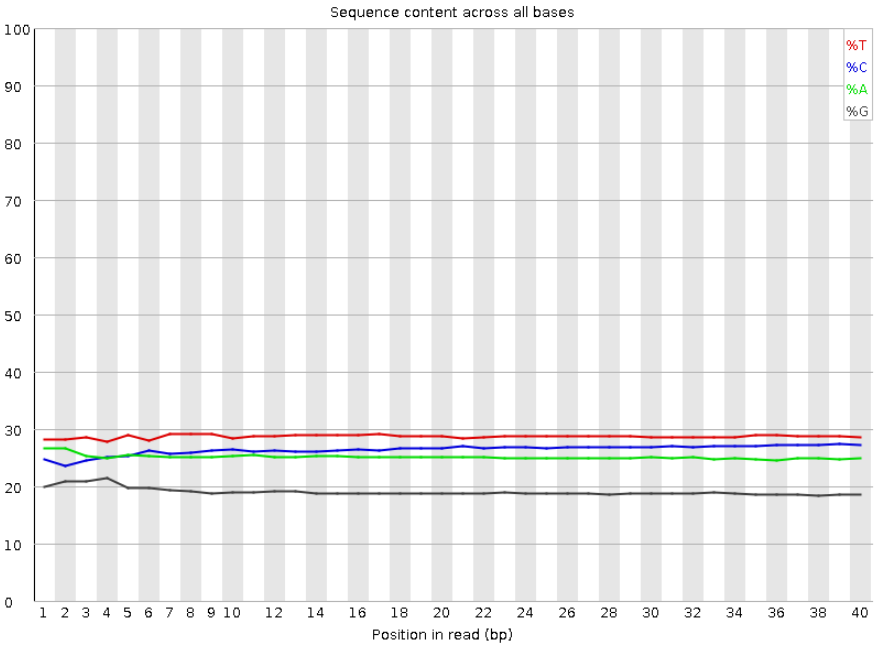
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

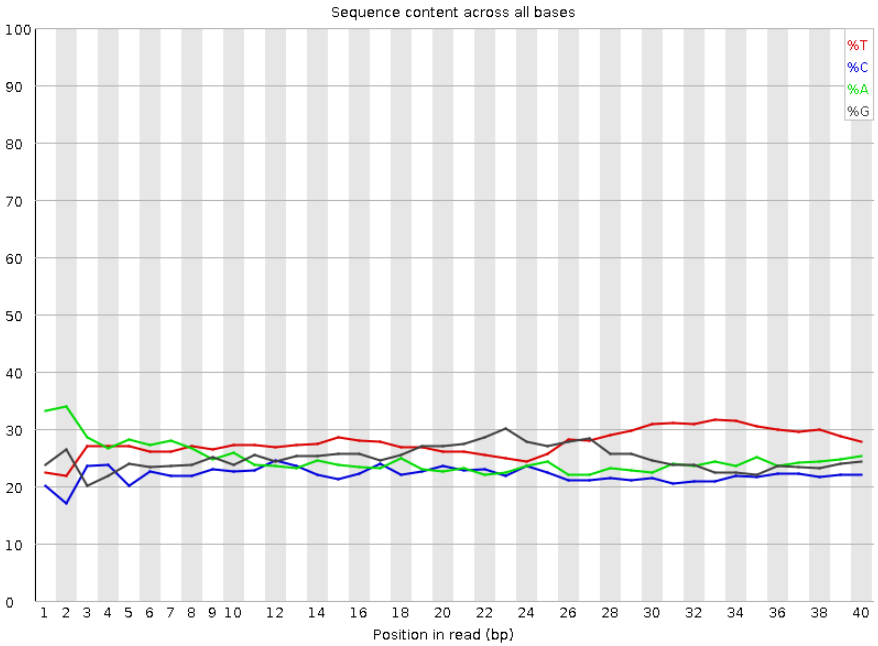
Example: Per Base Sequence Quality





Example: Per Base Sequence Content





Assembly or alignment

- We have sequenced genomes/exomes and we just performed quality control, what are our options?
  - The first genome of a species has to be assembled from scratch (*de novo* assembly), a computationally intensive operation.

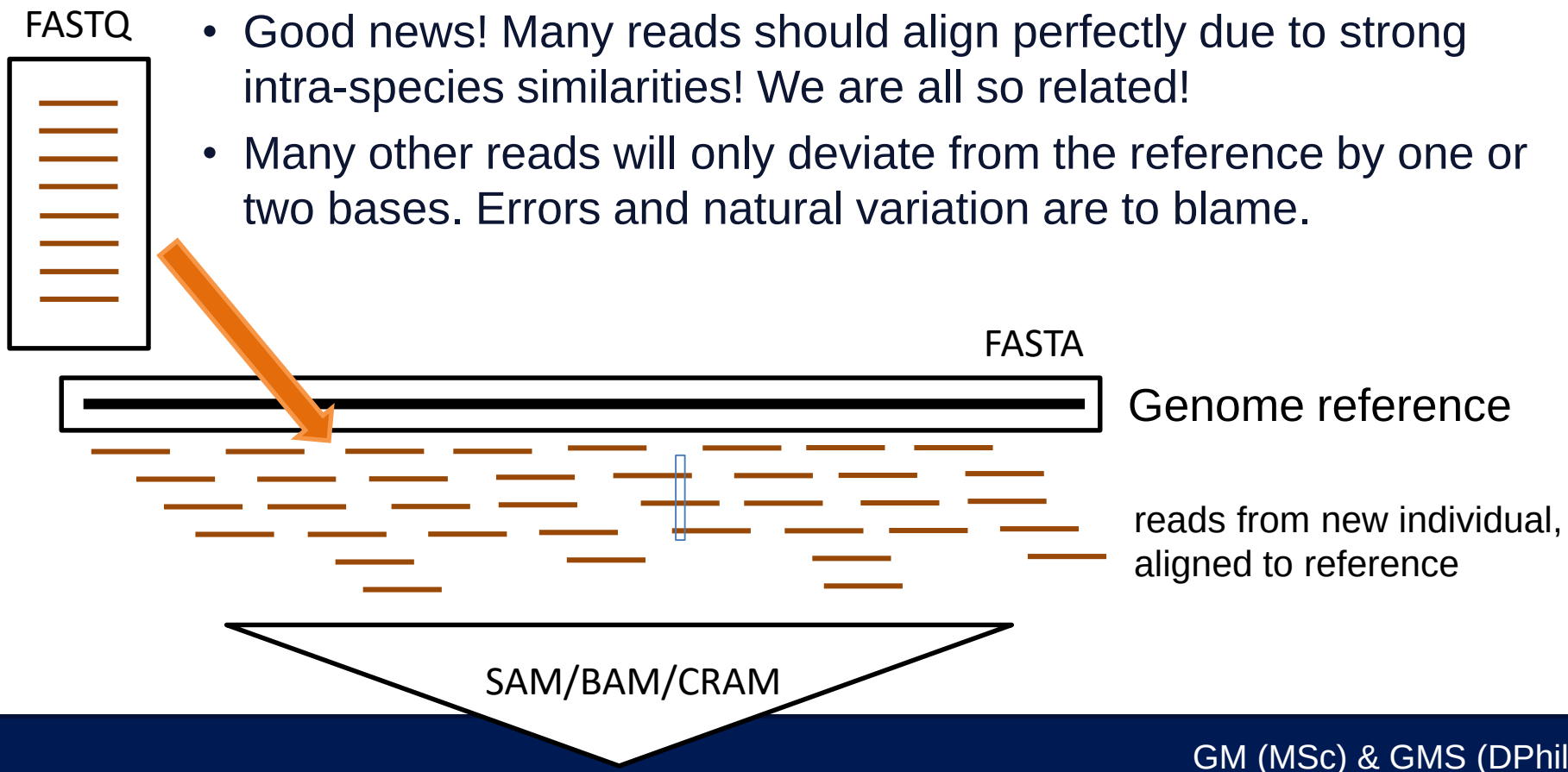
Does a public reference sequence exist for this species?

NO → *De novo* **assembly**

YES → **Alignment/mapping** against a reference (ska. comparative assembly)

Alignment to a Reference

- Once a complete species reference genome exists, we can align/map all subsequent individuals of the same species to it.



- Good news! Many reads should align perfectly due to strong intra-species similarities! We are all so related!
- Many other reads will only deviate from the reference by one or two bases. Errors and natural variation are to blame.

Challenges of Alignment

UNIVERSITY OF OXFORD

- Less computationally demanding does not mean aligning reads to a reference is simple:

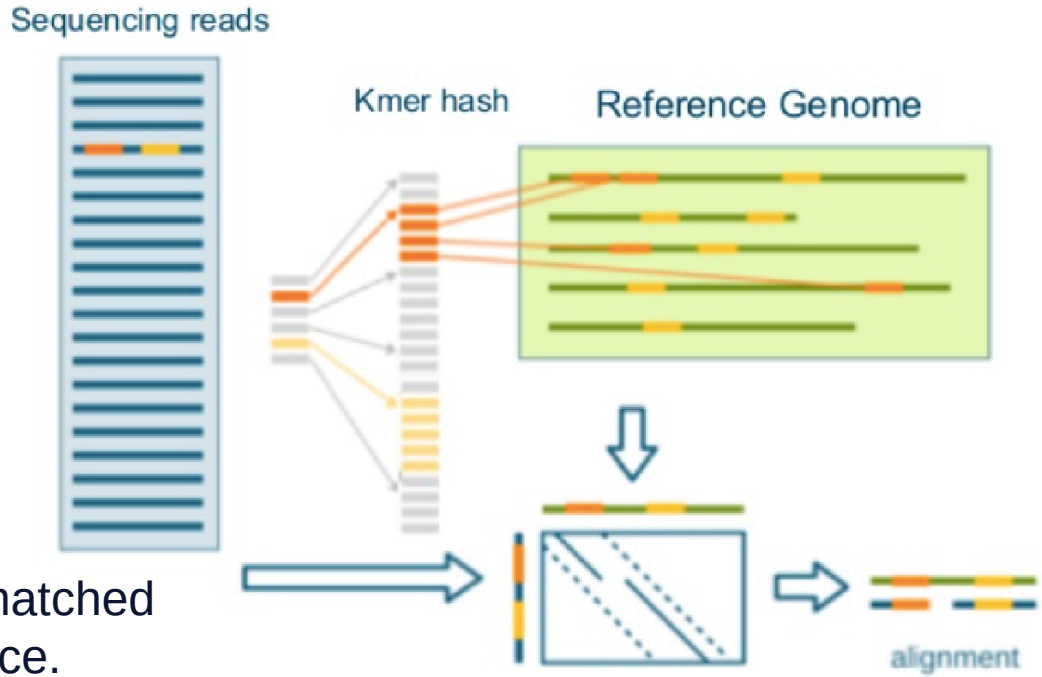
  - Millions of short reads to map to an entire genome
    ➔ data structure for rapid matching is needed.

  - The presence of both errors and individual variation complicates the alignment process.
    ➔ requires some clever dynamic programming.

  - Low complexity and repetitive regions are difficult to align to.
    ➔ paired-end reads help in some cases.

Fast Alignment using Hash

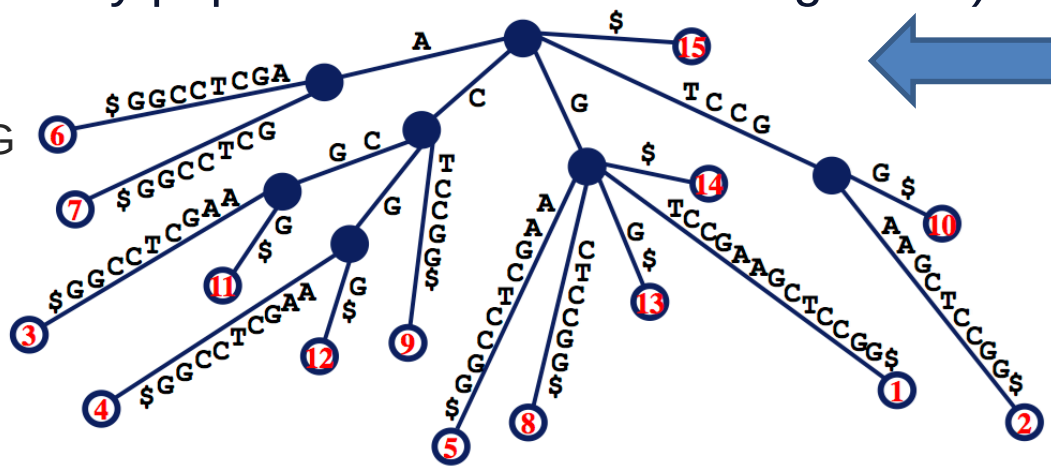
- Fast aligners largely fall into two categories based on underlying data structure used to store and compare reads and reference:
  1. Hash table
     - Used by aligners **Novoalign** and **MAQ** (also BLAST)
     - Reference genome stored as subsets of size k (aka. k-mers) in a hash.
       - Subsets of reads matched against the reference.
     - Based on perfect matches, align mismatching parts of reads.



Fast Alignment using Tries

- Fast aligners largely fall into two categories based on underlying data structure used to store and compare reads and reference:
  2. Suffix tries or arrays, or FM-index
     - Structures based on storing all possible suffixes of a sequence.
     - Used by **BWA** and **Bowtie2**.
       (both very popular methods for fast alignment).

A suffix trie for GTCCGAAGCTCCGG



GTCCGAAGCTCCGG$
TCCGAAGCTCCGG$
CCGAAGCTCCGG$
CGAAGCTCCGG$
GAAGCTCCGG$
AAGCTCCGG$
AGCTCCGG$
GCTCCGG$
CTCCGG$
TCCGG$
CCGG$
CGG$
GG$
G$
$

Deviations from the reference

UNIVERSITY OF OXFORD

- Hashes and tries are useful for exact matches, but not all reads match a region of the reference.
  - Mismatches take different forms:
    - Single nucleotide alteration.

Ref …ATGATGCCATGACTGACCCTGAT…

Read …ATGATGCCATGACTGAC**A**CTGAT…

source: variant (SNV) or base calling error.

Jointly referred to as indels

- Insertion

Ref …TCCATGTGTGACTA******CACC…

Read …TCCATGTGTGACTATTTGTCACC…

source: real insertion or region difficult to align

- Deletion

Ref …AAACTTAGTGCAACAGTGCACGAG…

Read …AAAC**AGTGCAACAGTGCACGAG…

source: real deletion or region difficult to align

Phred Quality Score Revisited

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

Basecall quality score (BQ o BASEQ). Also 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** (more in a minute…).

File format: Sequence Alignment/Mapping (SAM)

- Most popular fast-aligners (e.g. BWA, Bowtie2) take FASTQ as input and output SAM/BAM files.
  - Adds alignment information to FASTQ read data (i.e. position relative to reference, mapping quality, presence of insertions/deletions).

```
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
HCHHHHHHHHGHGHEHFHCHHHHHHHHHHHHHHFEHHHEHHHHHAFE?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 using `samtools view`!

Alignment Information

- SAM/BAM file contains the following info about each read 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
HCHHHHHHHHGHGHEHFHCHHHHHHHHHHHHHHFEHHHEHHHHHAFE?FCFFFFHEHDFFEEFEEGEGFGHHH?GDCFGGHHHF?FCGGC     NM:i:2  NH:i:2  C
C:Z:chr15 CP:i:102514823        HI:i:0
```

- Position relative to reference where a read (inc. chromosome) and its corresponding read pair (incl. relative to the first half) are mapped.
- Mapping quality (MAPQ).
- The CIGAR (Concise Idiosyncratic Gapped Alignment Report*)
- A Bitwise flag for additional information about the read.

*Definitely the worse backronym in bioinformatics

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.

e.g. 40M5I30M2D25M

ATGATGCCATGACTGACCCTGATGGTCCATGTGTGACTA***CACCACATGCTGGATAGGTGCCCGTGAAACTTAGTGCAACAGTGCACGAGATGAGGAGTG**

ATGATGCCATGACTGACCCTGATGGTCCATGTGTGACTATTTGTCACCACATGCTGTATAGGTGCCCGTGAAACAGTGCAACAGTGCACGAGATGAGGAGTG**

Preparing files for variant calling

UNIVERSITY OF OXFORD

- Once SAM files have been generated, a number of steps can be reduce file size and prepare the file for variant calling:
  - SAM files can be compressed to smaller binary files.
    ➔ BAM, no longer human-readable.
    - Crucial when sequencing hundreds of samples.
  - BAM files can then be sorted and indexed to further reduce size and speed up subsequent variant calling.
  - Duplicates are removed.
- If using GATK, one final step is needed*
  ➔ Base Quality Score Recalibration (BQSR)

*BQSR will be discontinued in the near future.

Variant calling and genotyping

UNIVERSITY OF OXFORD

- A number of reads mapped to the reference contain one or more bases deviating from the reference.
  - The task of a variant caller is distinguishing real variants ☑ from error✘ based on various supporting evidence
    - read depth, base/mapping quality scores, fits within an existing genotype configuration, larger haplotype context…

As a useful secondary goal, a variant caller can produce a genotype assignment.

- For a human sequence this is codified as:
  -0/0 or 1/1 (homozygous wildtype and variant respectively)
  -0/1 (heterozygous)

Read Depth/Coverage

UNIVERSITY OF OXFORD

- Base call accuracy for NGS platforms is usually around ~99.9%* (~1 error every 1000 bases).
  - However, the high throughput nature of these platforms translates to most regions or loci being covered by multiple reads.

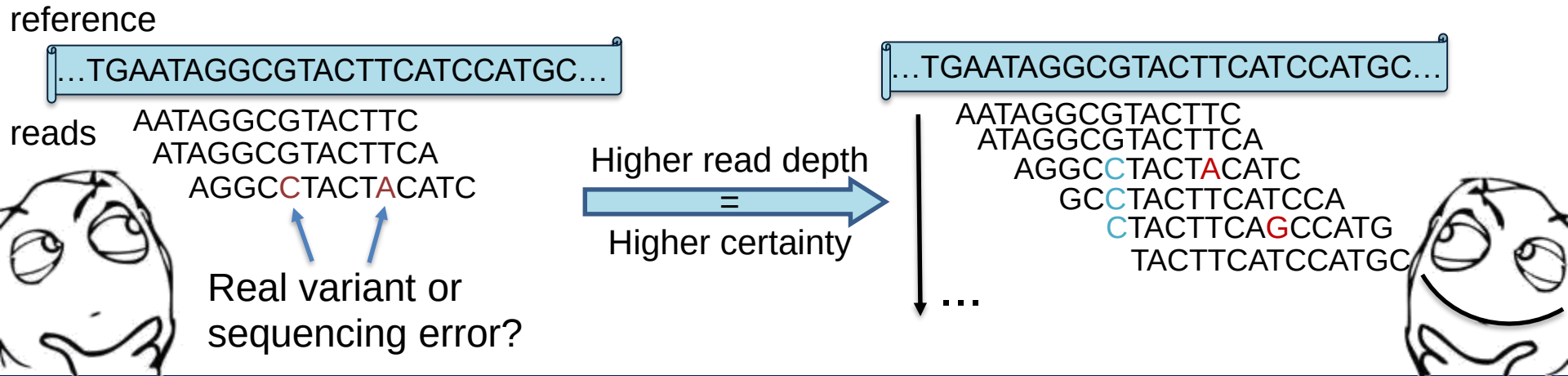
```
Read 1:   CGGATTACGTGGACCATG (read length of 18)
Read 2:       ATTACGTGGACCATGAATTGCTGACA
Read 3:                   ACCATGAATTGCTGACATTCGTCA
Read 4:                       TGAATTGCTGACATTCGTCAT

Depth:    111222222222333344333333333322222221
```

*Remember, that Sanger method has 99.999% base call accuracy.

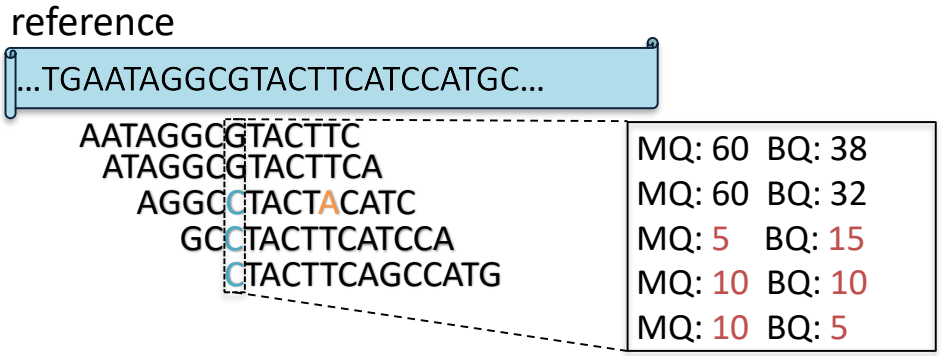
Calling Variants with Read Depth

- The quantity of reads that overlap a locus of interest ➔ **read depth**.
  - Consensus between reads gives us our first clue at the validity of a variant. More reads constitutes more evidence confirming or denying the existence of a variant at said locus.
    - Genotypes can be derived from read depth as well (homozygous variant would appear in most reads, heterozygous in about half the reads).



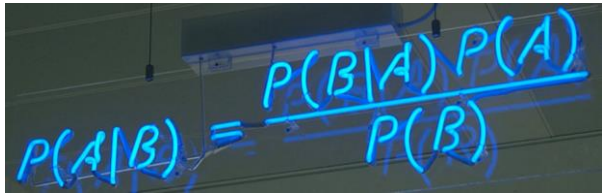
Using Quality Scores and More

- Another aspect of reads that a variant caller needs to account for is base and mapping quality.



In this example, **C** doesn't look as good as it did previously.

- The various lines of evidence are usually combined using some form of Bayesian statistics.



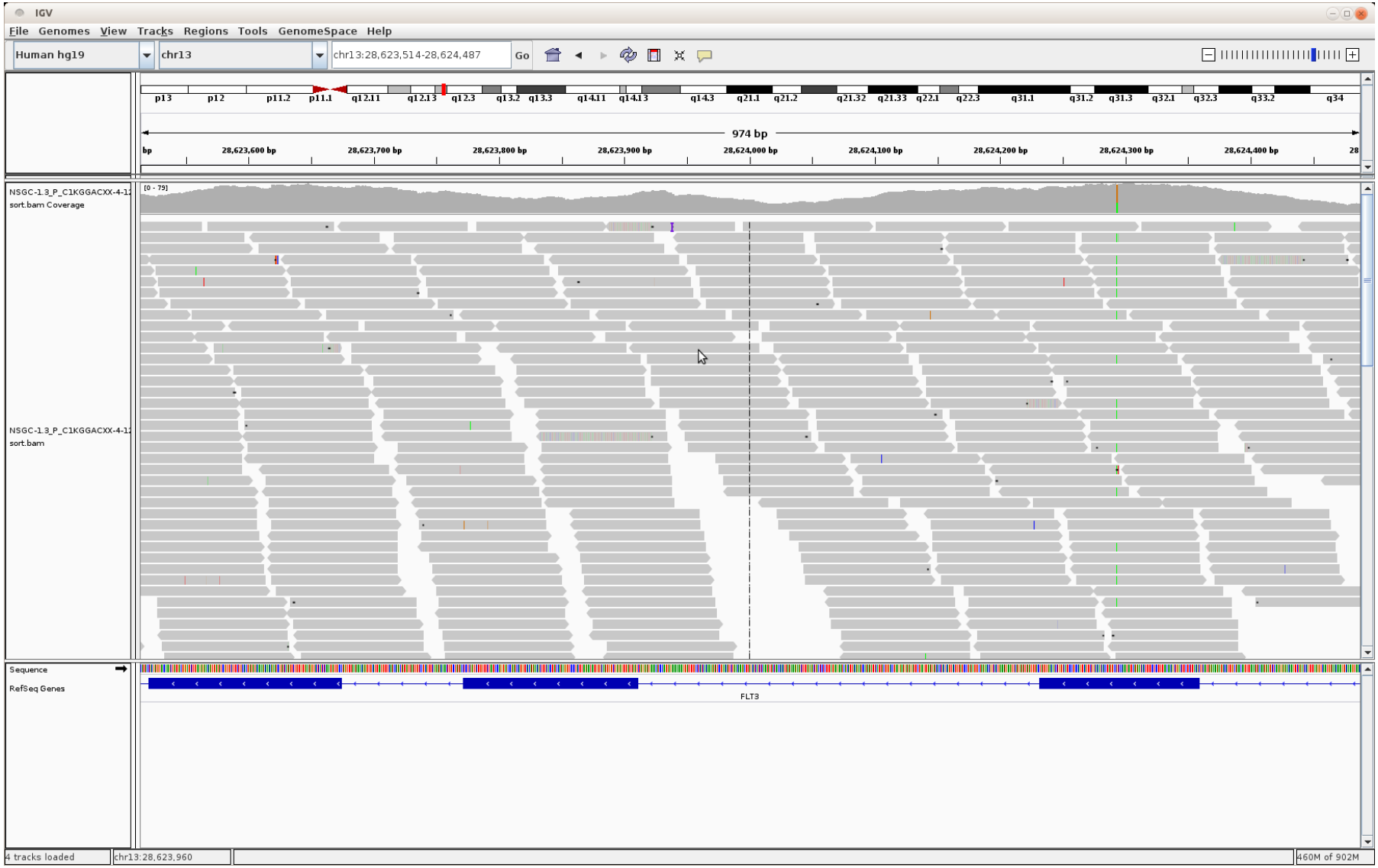
  - GATK HaplotypeCaller further incorporates local realignment and haplotype context.

VCF Format

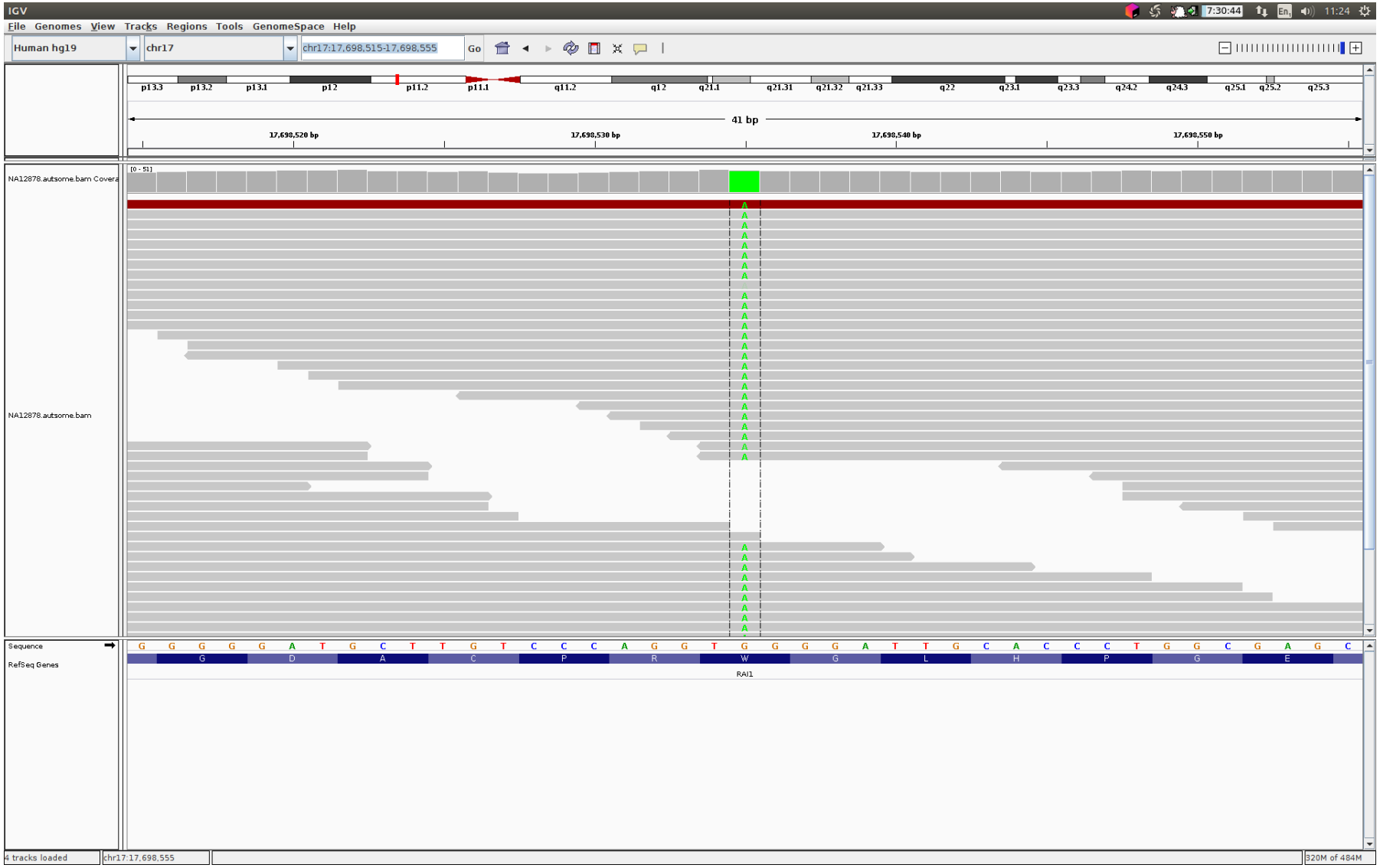
- The Variant Call Format (VCF) is designed to be human-readable.
  - We can use it for some further quality filtering
  - Filtering by genotype, particularly if working with families.

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Checking a locus with IGV

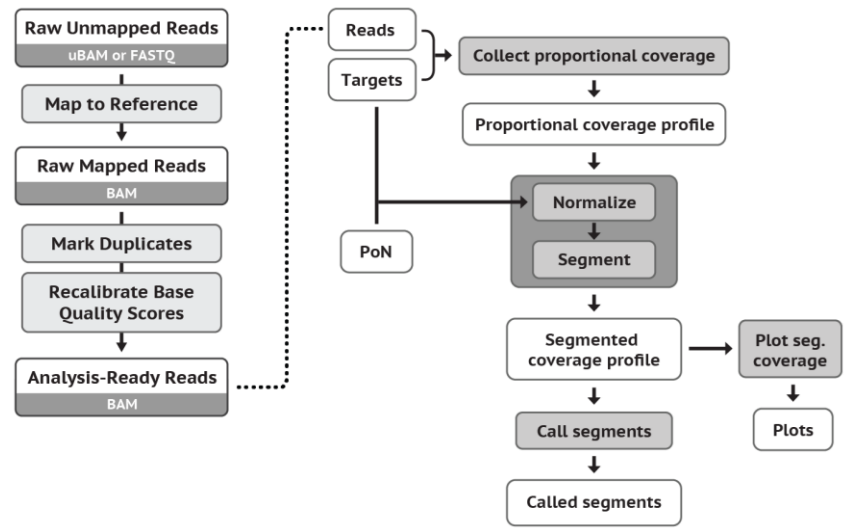


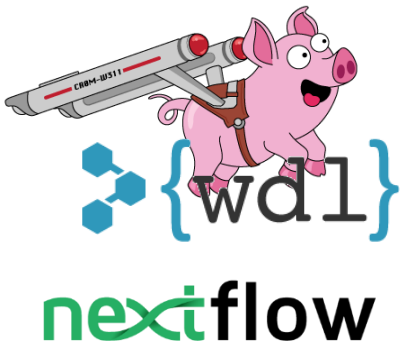
Checking a locus with IGV



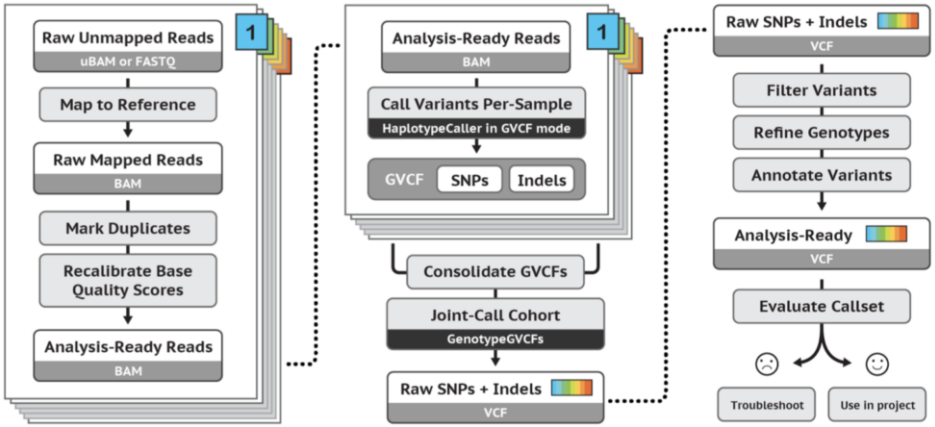
Available pipelines

- The various tools for QC/Alignment/Variant Calling can be and have been organized into pipelines.
  - The Broad Institute provide their preferred pipelines using WDL.
  - For cases in which one wants to create their own pipeline:





COMMON WORKFLOW LANGUAGE



End of Sequence

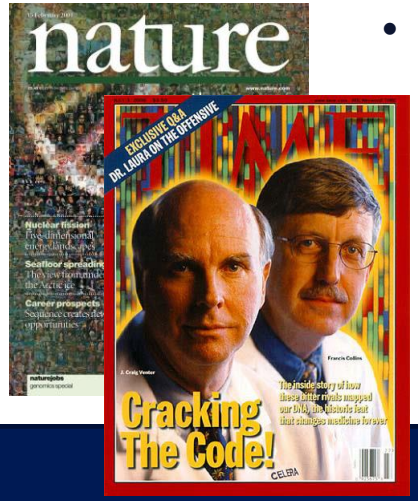
Thanks for listening!

Any questions?

Extra: Early Sequencing

- For nearly 40 years, dominant approach to sequencing ➔ **Sanger method** .
  - Originally quite <u>lab</u>orious, progressively automated.
  - Various organisms first sequenced using Sanger, incl. *Homo sapiens* with The **Human Genome Project** (HGP)
    - Begun in 1990, took 13 years to ‘complete’.
      - Complete euchromatic sequence of the human genome (~92%), used as a reference in successive projects.

Twice Nobel Prize winner **Frederick Sanger**



Extra: How does Sanger work? 1/2

UNIVERSITY OF OXFORD

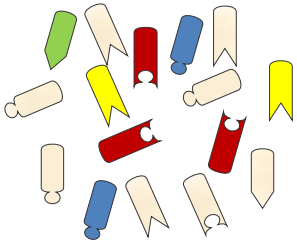
How can we determine the sequence of this fragment?

(C C A G C T G A T C A A G T A)

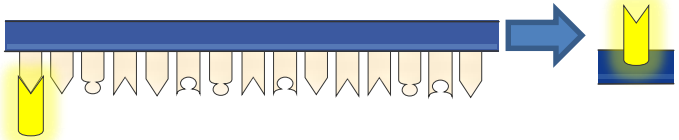
PCR amplified and denatured sequences

+

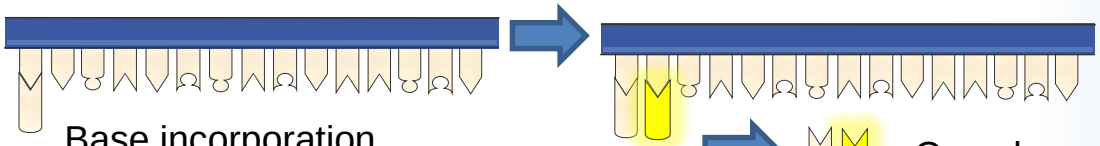
Mixture of dNTPs and fluorescently-labelled ddNTPs



Base incorporation stops when labelled ddNTP is added.

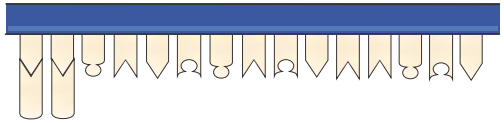


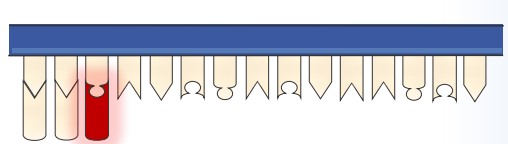
Complement with a single nucleotide.

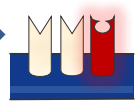


Base incorporation continues for unlabeled strand.

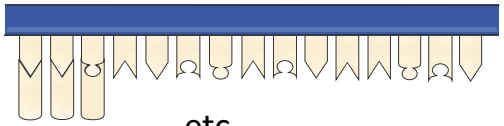
Complement with two nucleotides.



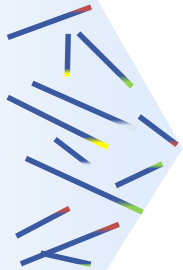




Complement with three nucleotides.



etc…

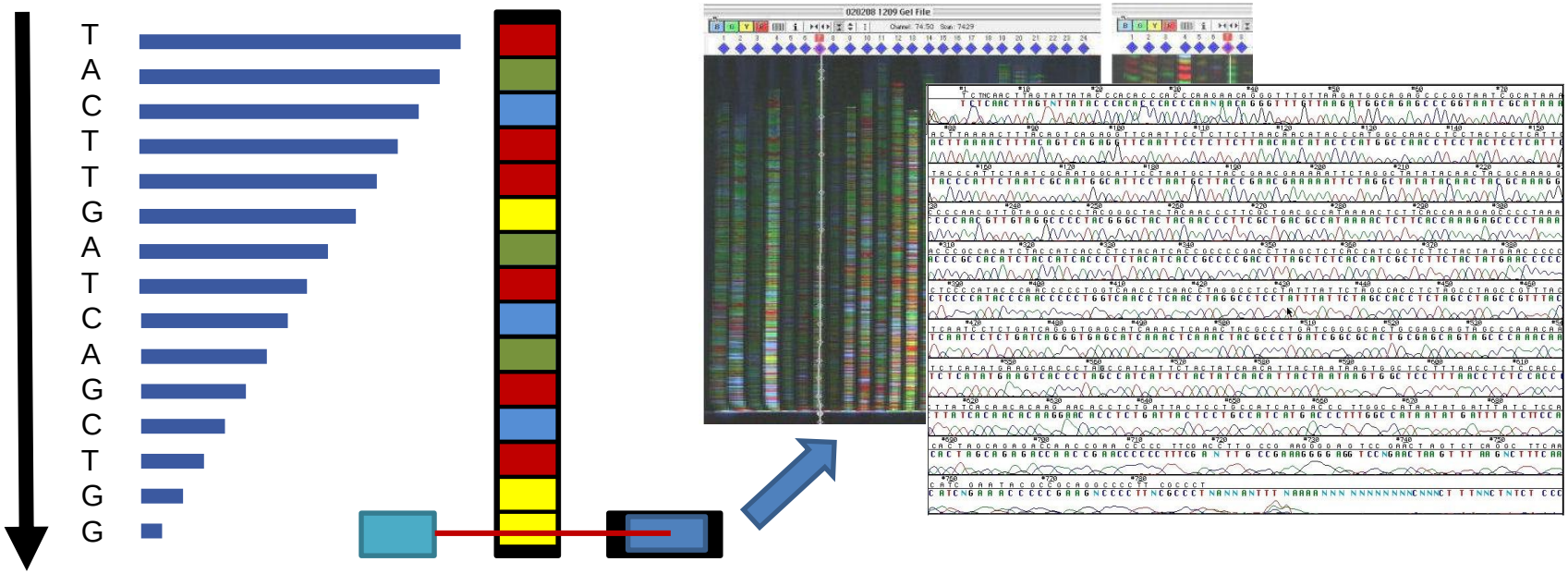


Random terminator incorporation process gives rise to multiple incomplete complement fragments where <u>length acts as a proxy for position</u> at which the chain terminating ddNTP was incorporated!

ddNTP = Dideoxynucleoside triphosphate (chain-terminating)

Extra: How does Sanger work? 2/2

- Complement fragments passed through capillary tube using electrophoresis, their final ordering determined by size.
  - The smaller a fragment the faster the pass through.
  - ddNTP fluorescence captured via laser excitation.



Extra: Sanger Method's success

- To this day, Sanger's <u>single read per base quality</u> remains unmatched.
  - For a fragment of ~500 base pairs (bp), per base accuracy ➔ **99.999%**
    - Note that's still 1 error per 100 000 bases.
  - Fragments now reach up to 1000 bp in length.

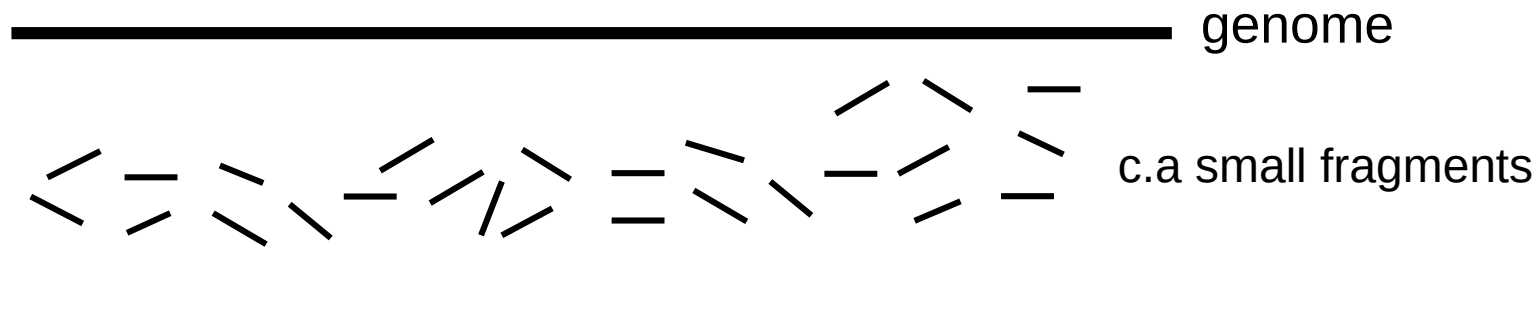
How do you sequence the 3 billion base pairs of the human genome with technology that only takes fragments of (up to*) 1000bp at a time?

*Originally much less than

Extra: Shotgun approach to Seq.

- Developed to speed up the HGP and replace slow and costly clone-by-clone approach.
  - Genome broken into small **overlapping** sequencing-ready fragments, all clonally amplified.
    - **<u>No ordering conserved in the process.</u>**
      - Computational biologists now have a **BIG** problem to solve ➔ mapping billions of reads back to the genome.

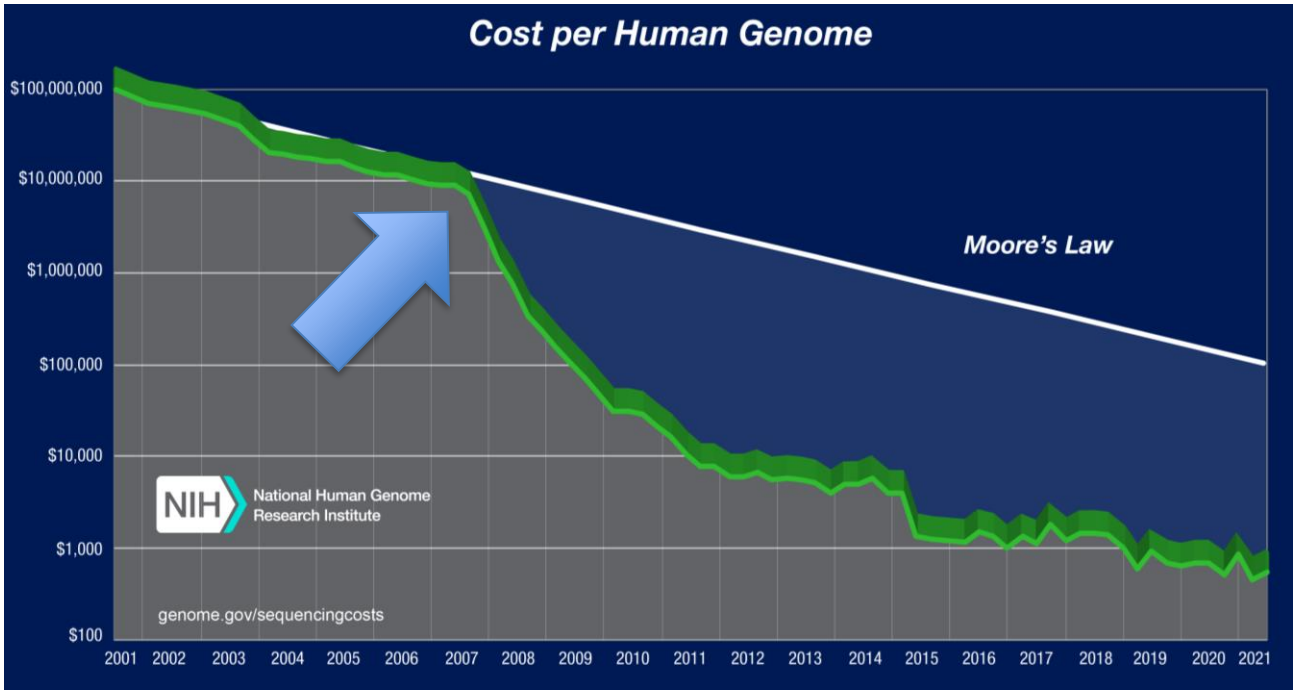
-cheaper
-faster
…
-assembly challenging



Which part of the genome did each read come from?!
Does read x align/overlap with read y?!

Extra: Next-generation Sequencing

- Mid-2000s: **high-throughput/massively parallel sequencing** (HTS/MPS) platforms released:
  - Referred to as next-generation sequencing (NGS)*.
  - A 13 year endeavor now takes days and costs ~1000$.

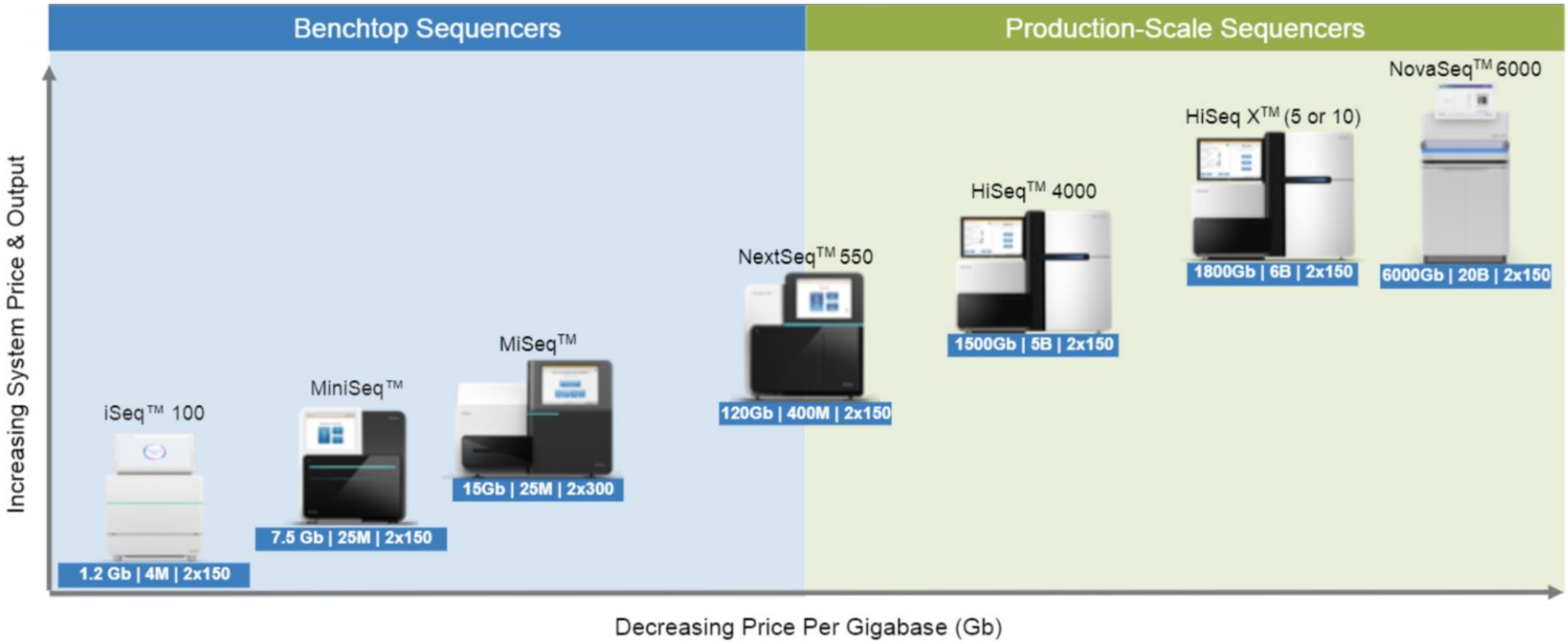


*Despite newer platforms, the name persists.

Extra: The First Three NGS Platforms

UNIVERSITY OF OXFORD

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



Today Illumina dominates the sequencing market (~80%) with a wide range of platforms tailored to different needs.

Extra: GATK's BQSR step

- Base quality scores are only as good as the error predictions a platform makes.
  - Some systematic biases have been observed.
  - These can be addressed using known variants at a population level (e.g. the dbSNP database).
    - Base changes present in the data and also present in a database treated as real.
    - Other variants treated as errors.
    - These assumptions are true enough to be a useful approximation.