An Introduction to Next-Generation Sequencing

Matthieu Miossec, PhD
Senior Bioinformatician @ Bioinformatics Core

Sequencing At a Glance

- The process of deciphering a sequence of bases and producing a digital representation → a **read**, for further analysis.

FASTQ file (more on those later…)

This is a read.

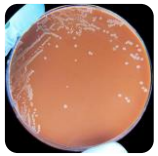
```
@SN431_1138:5:1101:1176:10390#33/2
CGGGTGTGTGACCTGGGAGAGCTGGTGCATCACGGGCGTATCTTTGCACTGCTGGCTGGTGCGCCCATGGGGAAGCATCCGCGTTGGGTCCCACATTCTC
+
=ED>>=<<ECCD0/GDF;D.IGGDCEA;<H.H=;)H:<,;B-HF--;GAEHC:<BEI.;<;8.;99F@<.DD'8@,,'D6E9;@7GB.H-(7C6',A880
@SN431_1138:5:1101:1177:20812#33/2
AGTCAGTGAGGCCTTGTCGGTAAGGACCTGGTCCTTGAGGCCTTGCCAGTGAGGCCTTGTCAGTAAGGTCCTGGTCACTGAGGCCTTGTCAGTAAGGACT
+
=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
```

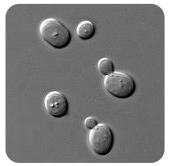
In an ideal world, we could sequence entire chromosomes quickly and get a single error-free readout for each chromosome as our output.

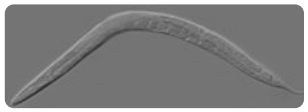
This is not the world we live in…

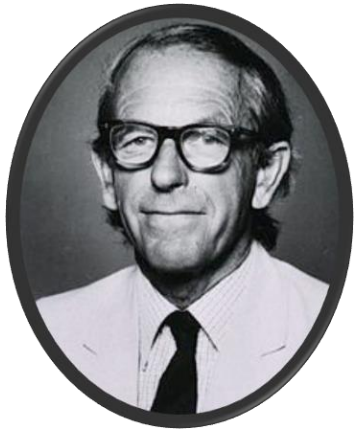
Early Sequencing

- For over 40 years, dominant approach to sequencing → **the Sanger method**.
  - Originally manual and laborious, gradually automated in the 80s.
  - Various model organisms first sequenced using Sanger
  - includes *Homo sapiens* with The **Human Genome Project** (HGP)

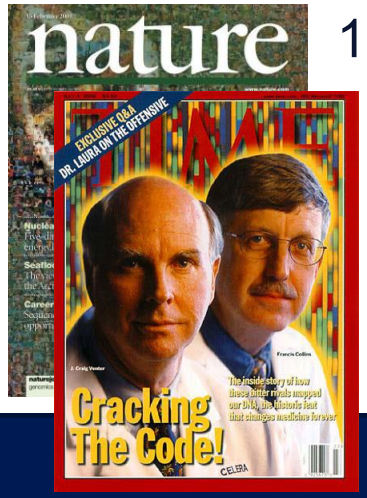








Twice Nobel Prize winner **Frederick Sanger**



1990 → 13 years to 'complete' → 2003

- Complete euchromatic sequence of the human genome (~92%), used as a reference in successive projects.

Sanger's Success

- Even today, Sanger's single read per base quality is unmatched.
  - For a fragment of ~500 base pairs (bp), the resulting read's per base accuracy → **99.999%**
    - Note that's still 1 error per 100 000 bases.
  - Reads can now reach up to 1000 bp in length.

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

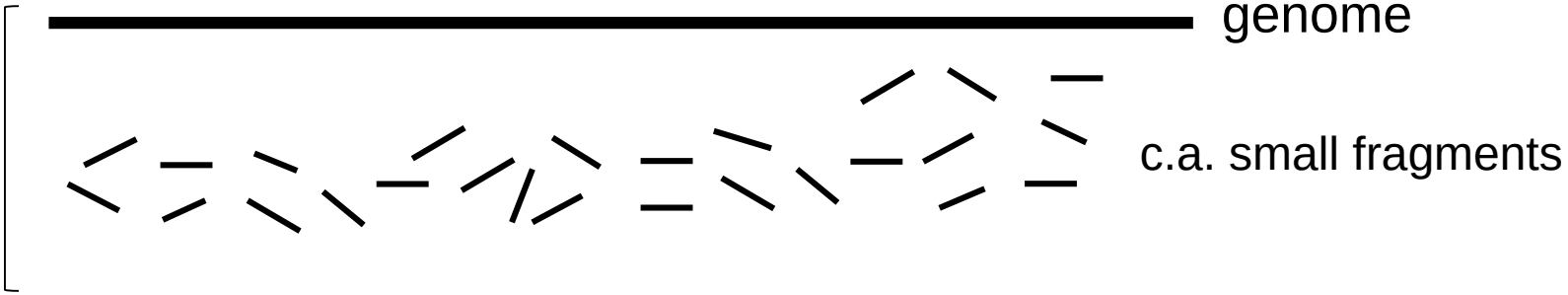
*Throughout the HGP, much less than 1000 bp!

Shotgun approach to Sequencing

- Developed to speed up the HGP and replace slow and costly clone-by-clone approach.
  - Genome broken up into small **overlapping** sequencing-ready fragments, all clonally amplified.

    😱 **No ordering conserved in the process.** 😱
    - Computational biologists now have a **HUGE** problem to solve → mapping billions of reads back to the genome.

-cheaper
-faster
…
-challenging
re-assembly



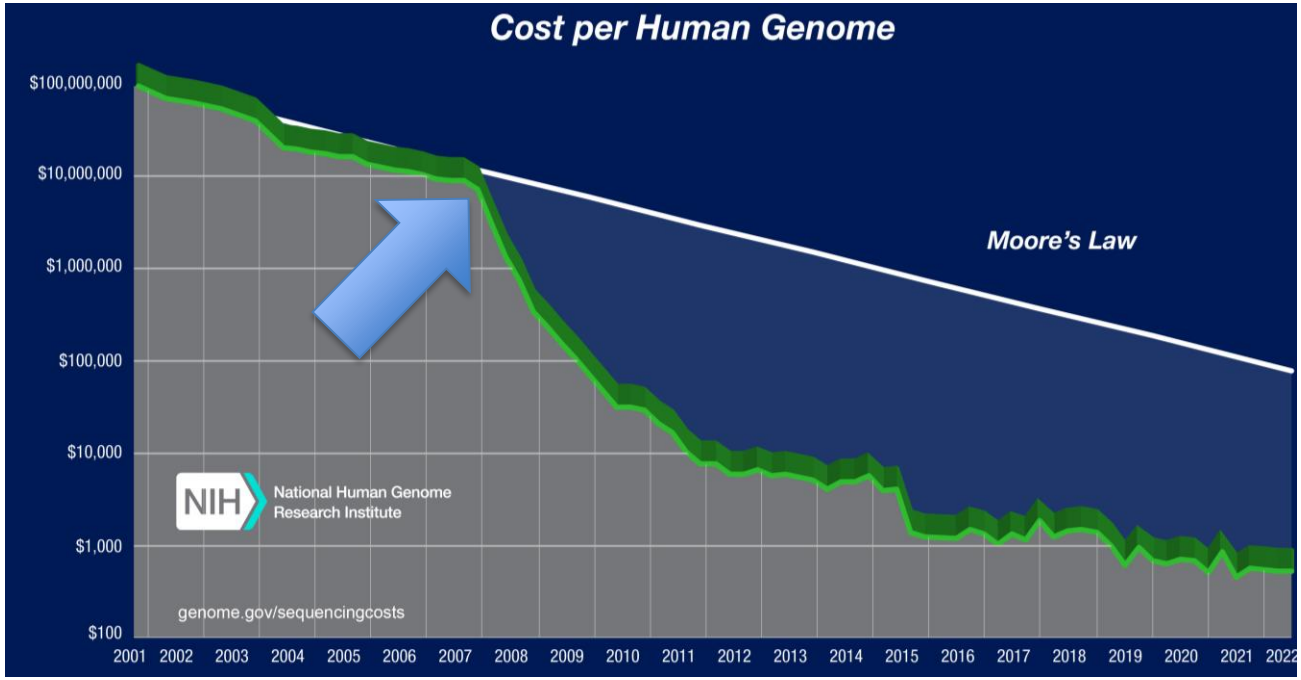
Which part of the genome did each read come from?!
Does read x align/overlap with read y?!

Assembly/mapping after shotgun

Next-generation Sequencing

UNIVERSITY OF OXFORD

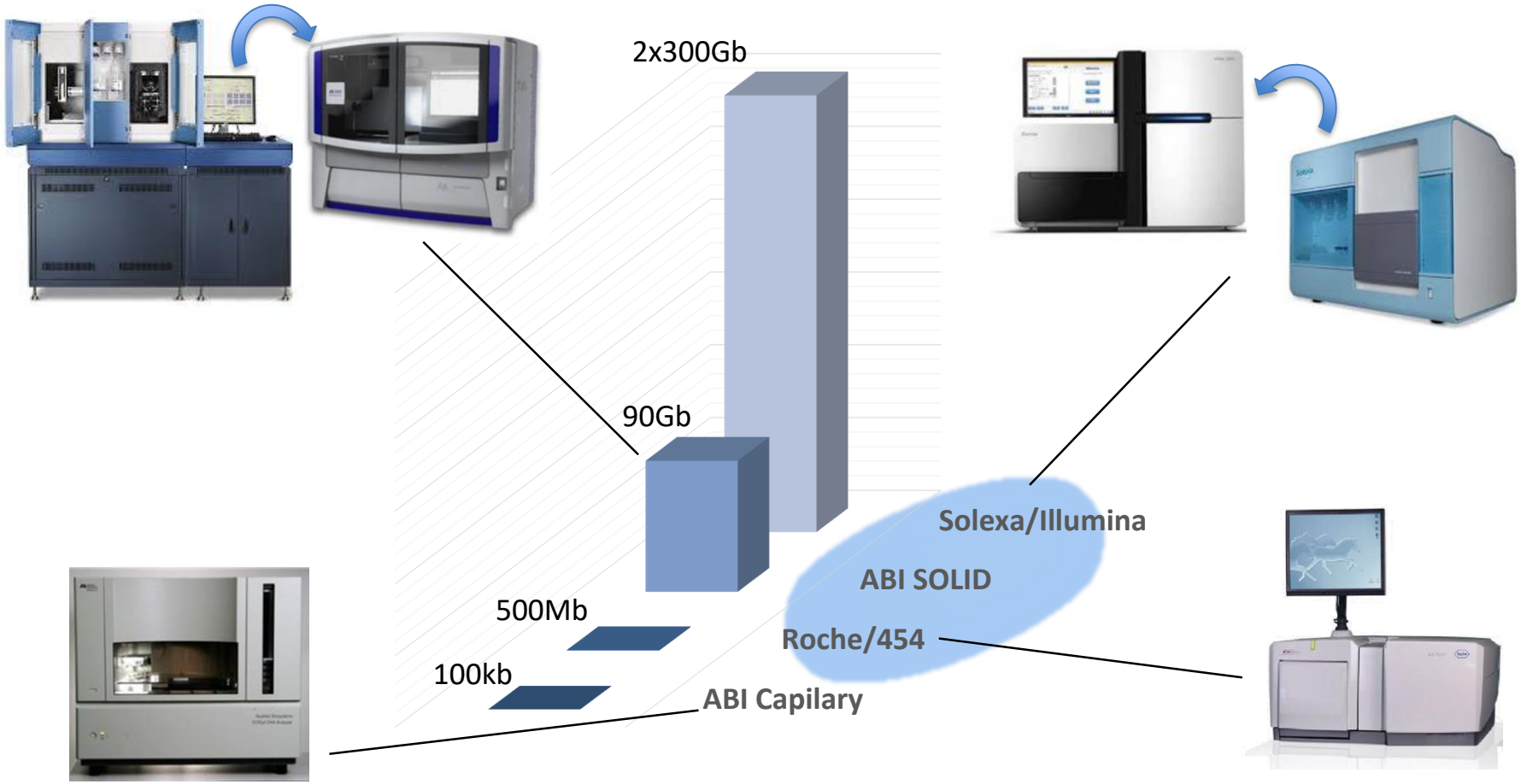
- Mid-2000s: **high-throughput/massively parallel sequencing** (HTS/MPS) platforms released:
  - Referred to as next-generation sequencing (NGS)*.
  - A 13 year endeavor now takes days and costs ~1000$.



*There is a fourth name we will get to soon…

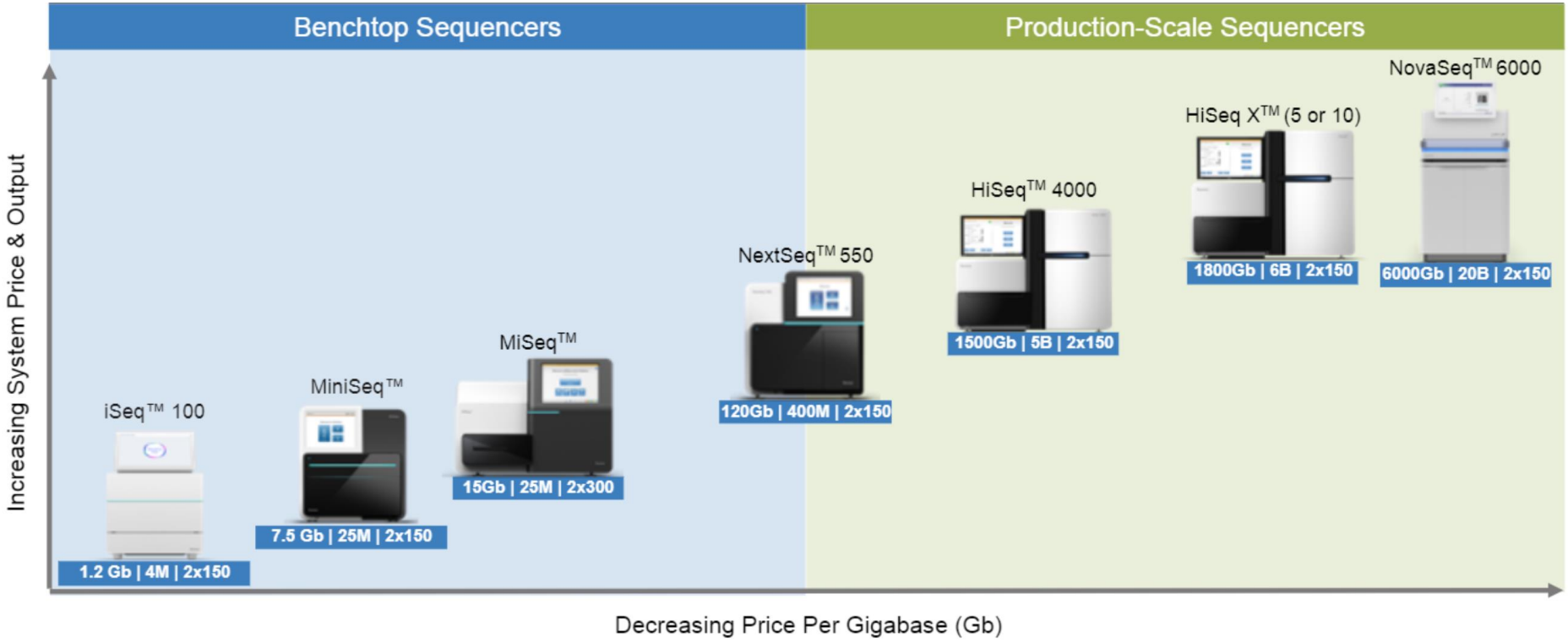
The first Three NGS Platforms

- NGS originally referred to three increasingly high-throughput sequencing platforms:



Illumina NGS Dominance

- Illumina currently dominates the market (~70%) with a wide range of platforms tailored to different needs.
  - Now risks being displaced by long-read technologies (more later).



Innovations of NGS

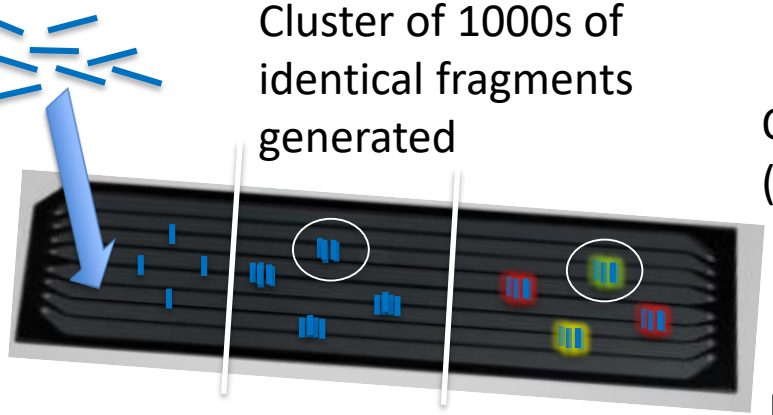
- Each NGS platform has its distinct biochemical processes for sequencing, all share important attributes:
  - Biochemical processes happening simultaneously for distinct fragments → massively parallel production of reads.
  - PCR amplification used to turn weak bioluminescent signals generated by a small fragment, into a strong signal of a cluster of ~1000 identical fragments.
  - Sequencing-by-synthesis (SBS)

Sequencing-by-Synthesis (in Illumina)

Fragments run over 2D flowcell

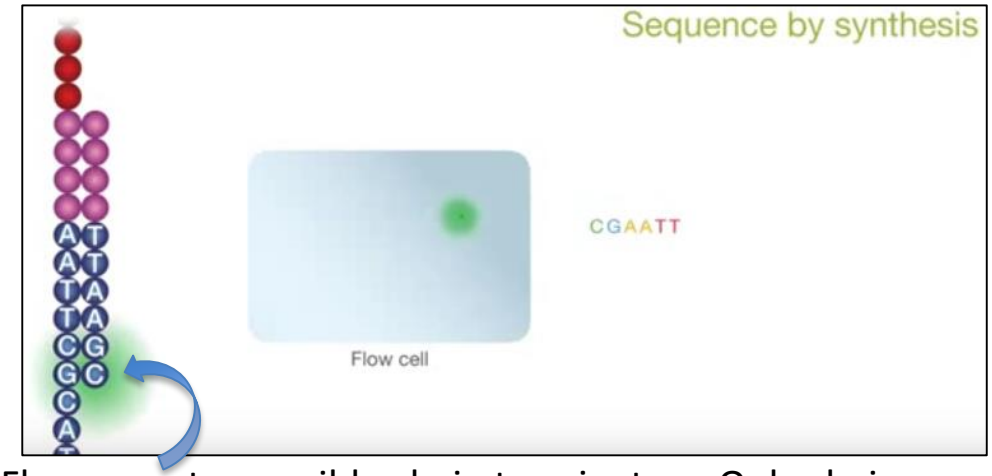
Cluster of 1000s of identical fragments generated

One base incorporated (in theory…) per cycle.

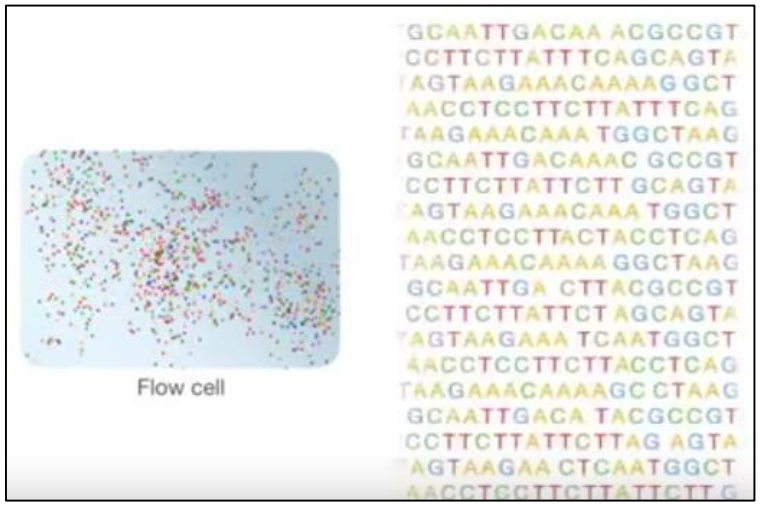


Flowcell

Captured with highly sensitive camera.



Fluorescent <u>reversible</u> chain terminators. Only chain terminating during a cycle, long enough for capture.

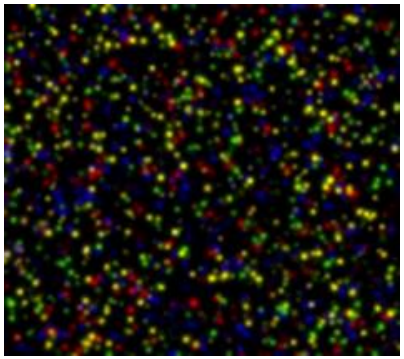


Flowcell during a given cycle. Multiple reads built in parallel.

https://youtu.be/fCd6B5HRaZ8

Limitations of NGS (short-read)

- Not all good news! New biochemical processes place constraints on **read length** and **accuracy**.
  - Determining bases of a sequence (a.k.a. **base calling**) using bioluminescence as a proxy, has several pitfalls:

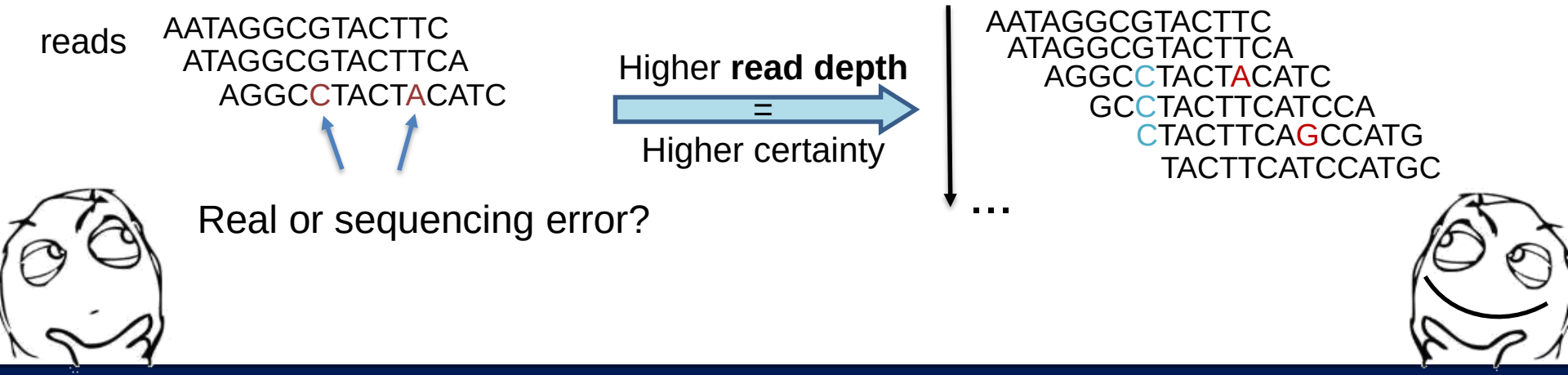


- ❖ Signals from clusters in close proximity interfere with on another.
- ❖ Synchronicity between strands within a cluster is gradually lost with each cycle (size of reads ≡ number of cycles).
- ❖ The intensity of a signal naturally varies.
- ❖ A signal can be ambiguous where bases repeat (e.g. Did the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

Another name for NGS, now in vogue:
short-read sequencing

Accuracy from Redundancy

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



Slow Start for Long-read Sequencing

UNIVERSITY OF OXFORD

- Early 2010s: New platforms appear to address some of NGS' flaws:
  - Speed, reliance on PCR amplification, lack of portability…quickly focus is on shortness of reads.
- Very limiting flaw → **low base qualities** (70-90%) 😱

Read length: up to 1.5kb (closer to 20kb now)

2010
Ion Torrent

2011
PACIFIC BIOSCIENCES®
SMRT sequencer
(single Molecule, Real-Time)

Over 100Kb reads!

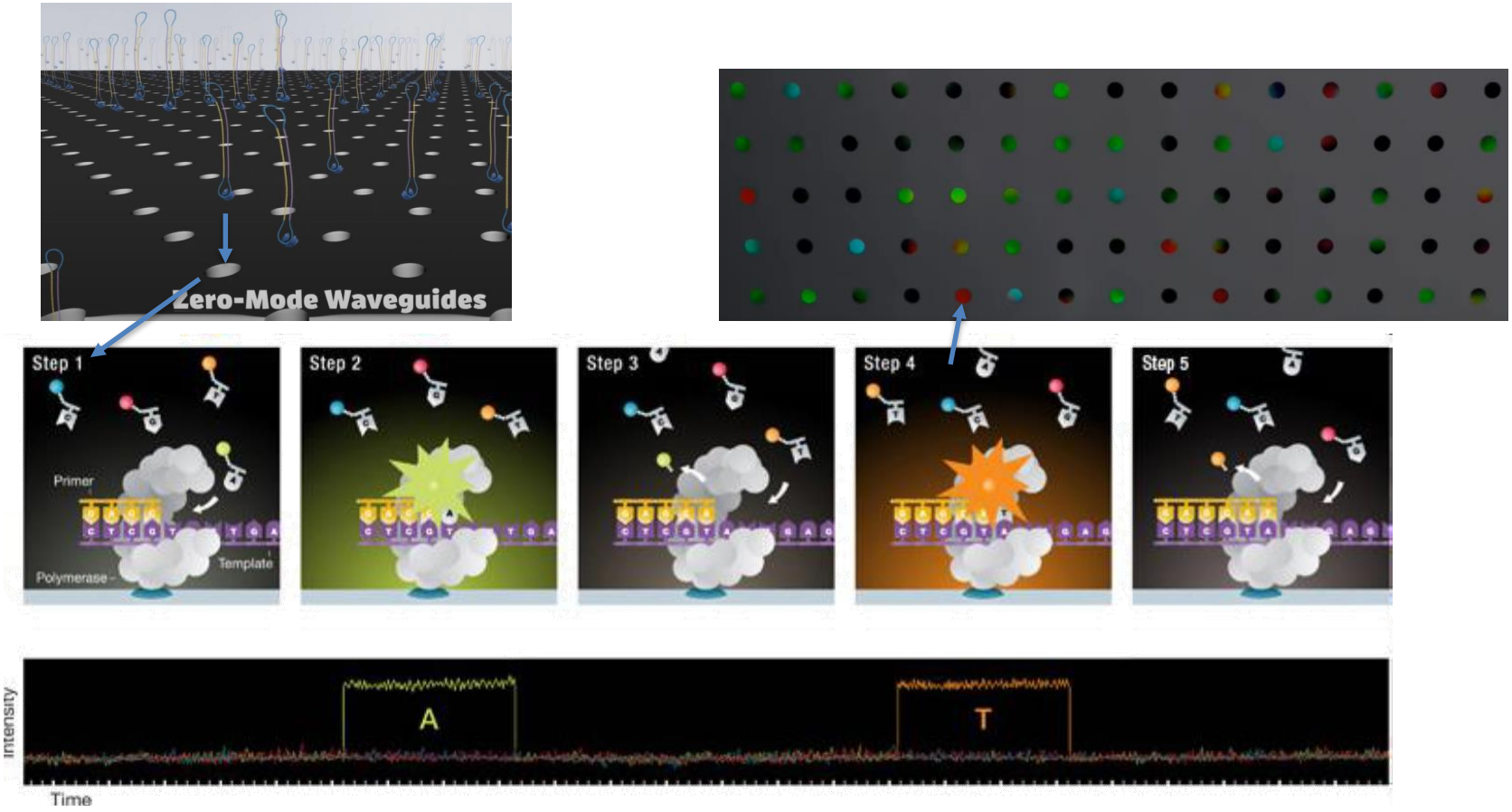
2015
Oxford NANOPORE Technologies
MinION

Good for finishing parts of the genome difficult to map using short-reads (e.g. repetitive regions, structural variation).

Dreadful for identifying point or small mutations.

PacBio's Zero-Mode Waveguides

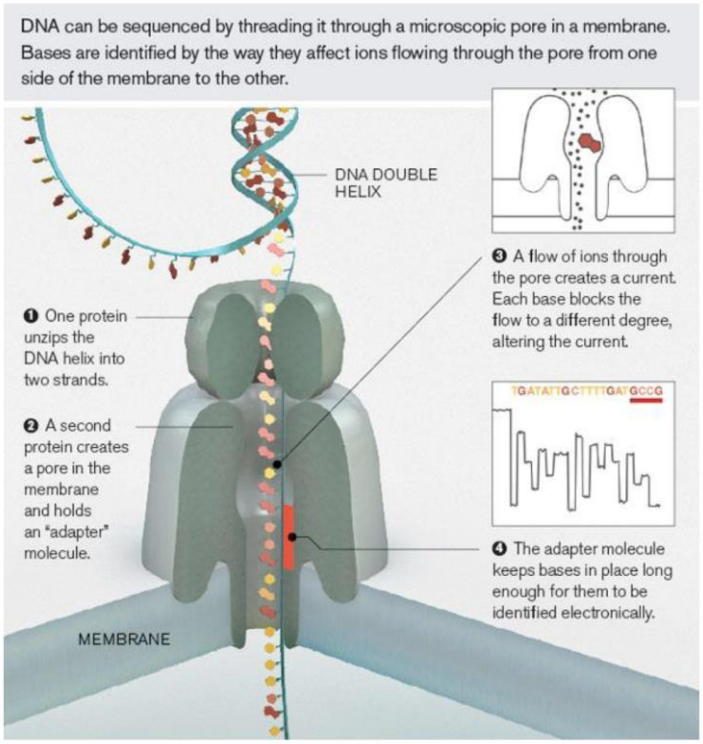
- Each well contains a distinct fragment + primer + polymerase. ~~No PCR, no chain termination~~.



https://youtu.be/_lD8JyAbwEo

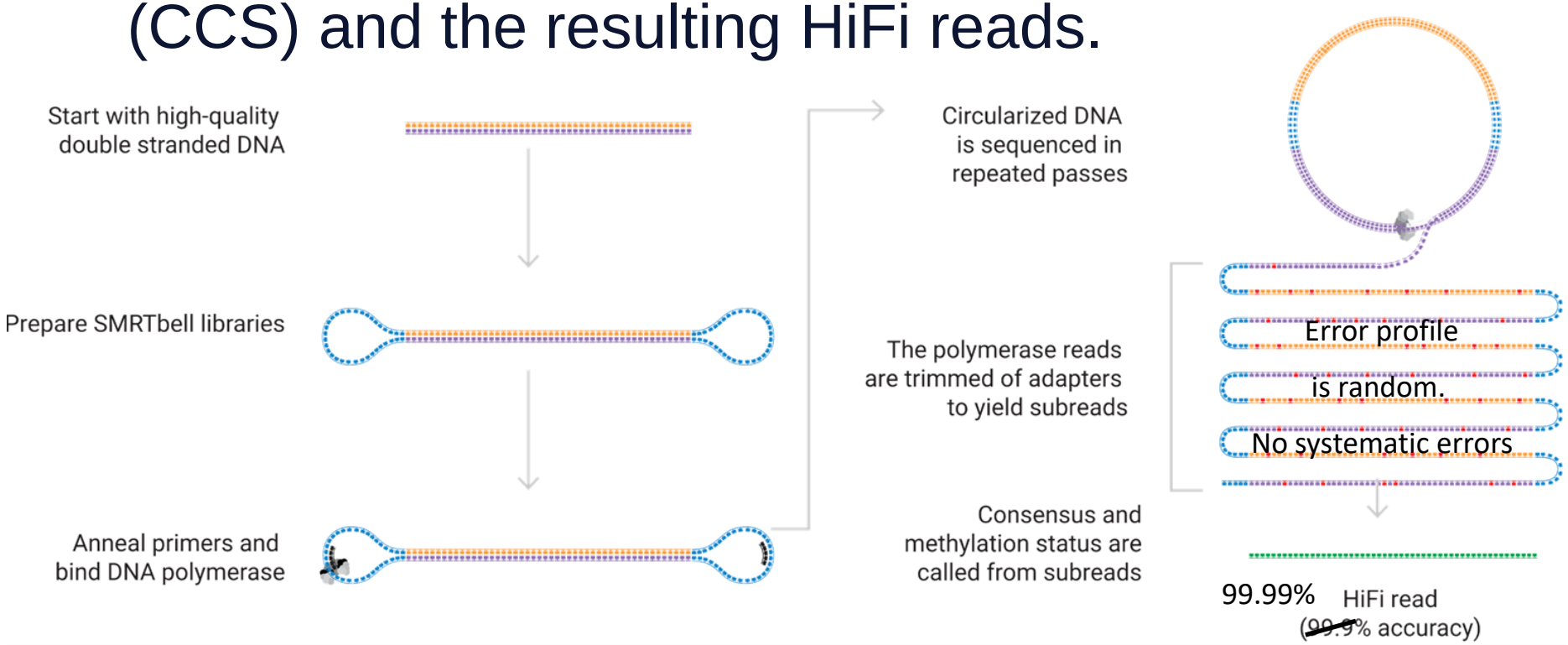
Oxford Nanopore Technology

- No more fluorescence, works via protein pore interaction. One DNA/RNA molecule at a time.
  - Detecting tiny changes in electrical current as nucleotides pass through.
- Key strength: Portability.
  - Great for work in the field like pathogen detection or metagenomics.



Long-Read Becomes Competitive!

- Late 2010s: PacBio HiFi taps into the power of redundancy with Circular Consensus Sequencing (CCS) and the resulting HiFi reads.

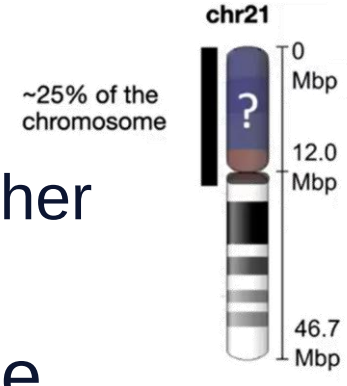


- Similar error rate than NGS, much longer reads.

https://youtu.be/_lD8JyAbwEo

Long-read Milestone: T2T

- 2003-19: Human genome only ~92% complete with incremental gap filling.
  - Centromeres, segmental duplications and other difficult regions ignored.
- 2020: First truly complete human genome
  - Combining Nanopore's very long reads and PacBio's long and HiFi/accurate reads, every chromosome entirely sequenced, telomere to telomere (T2T).



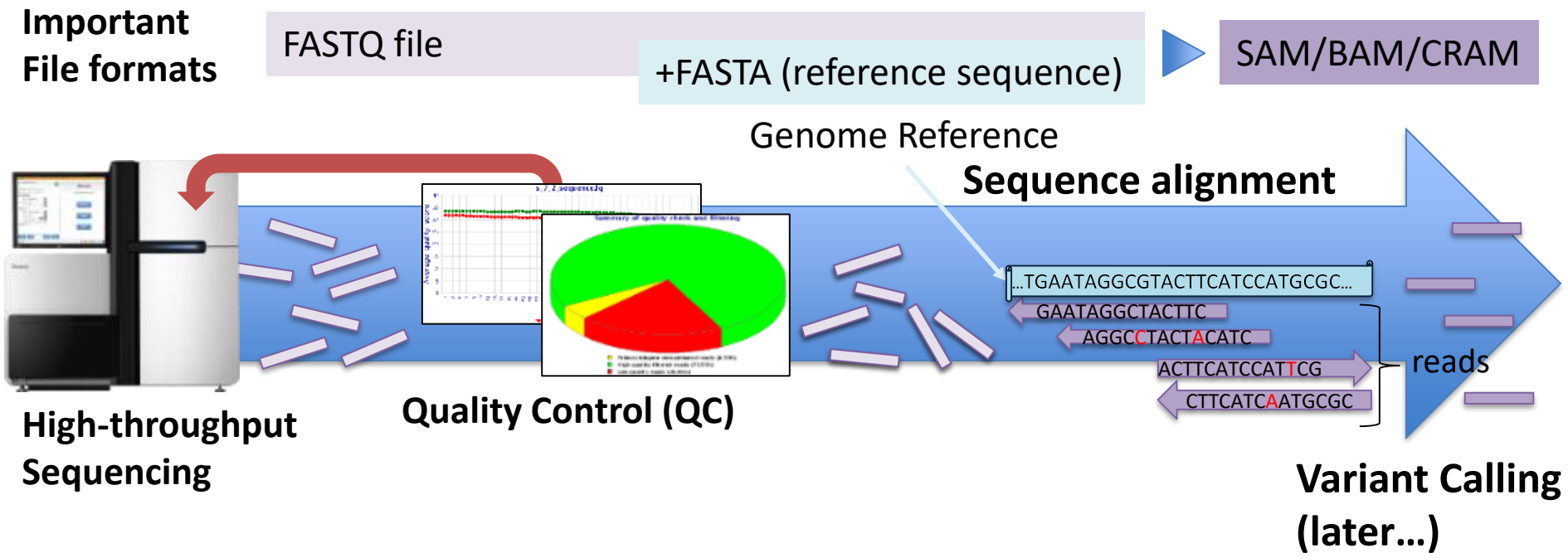
Reference publicly available now: T2T-CHM13v2.0: https://www.ncbi.nlm.nih.gov/assembly/GCA_009914755.4

TELOMERE-TO-TELOMERE CONSORTIUM

https://youtu.be/fOLlXSu5k14

Quality Control, Alignment and File Formats

- 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

- The simplest sequence file formats for storing sequence data (ext: .fasta, .fa…).
  - Contains at least 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:

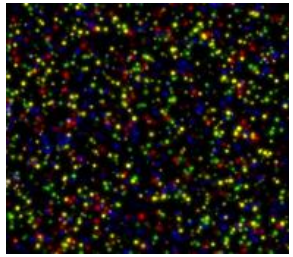
The Latest Human Genome Reference
GRCh38

File format: FASTQ

- 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 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

- 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

UNIVERSITY OF OXFORD

- 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.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
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/

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.

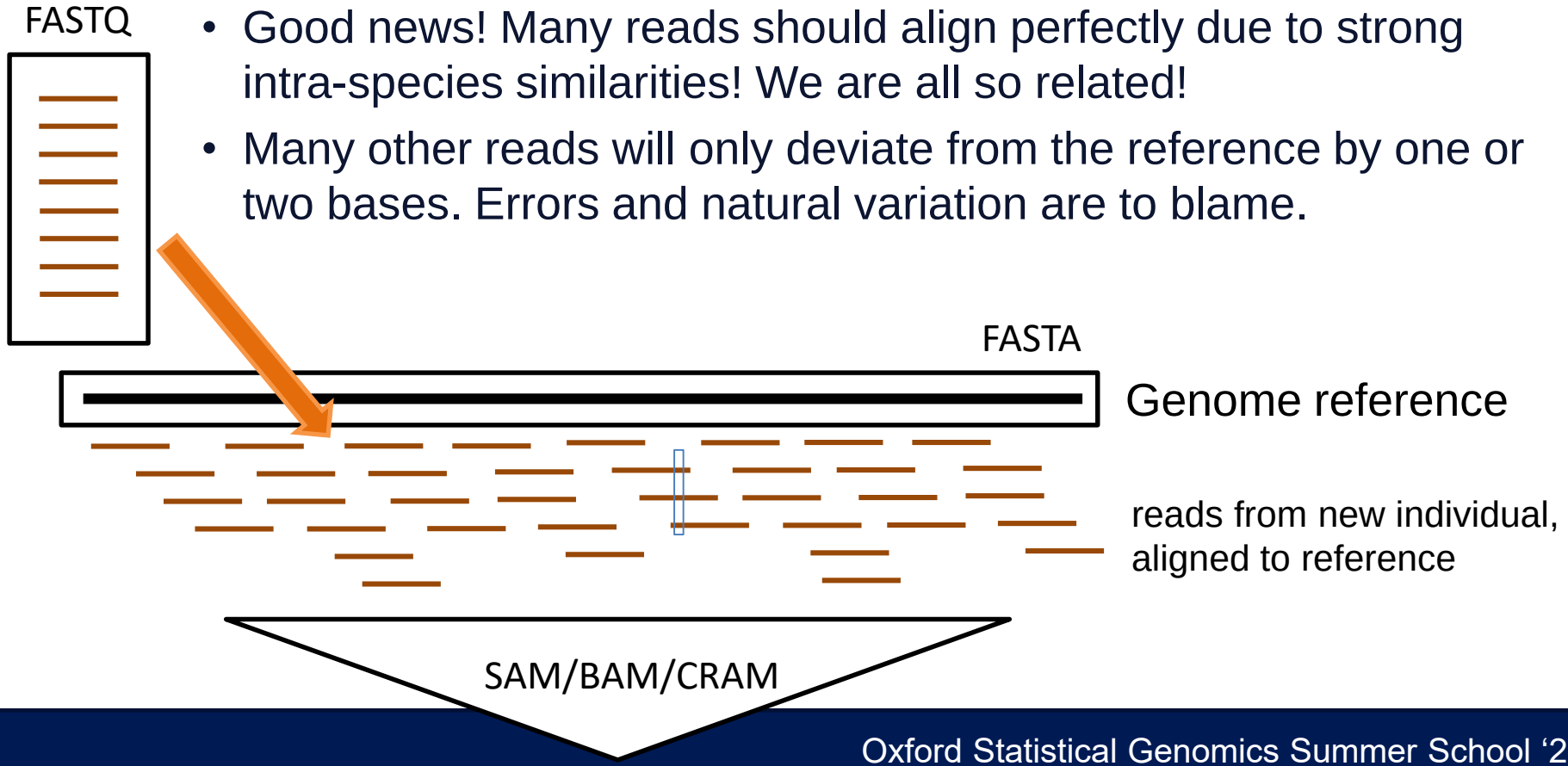
Does a public reference sequence exist for this species?

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:

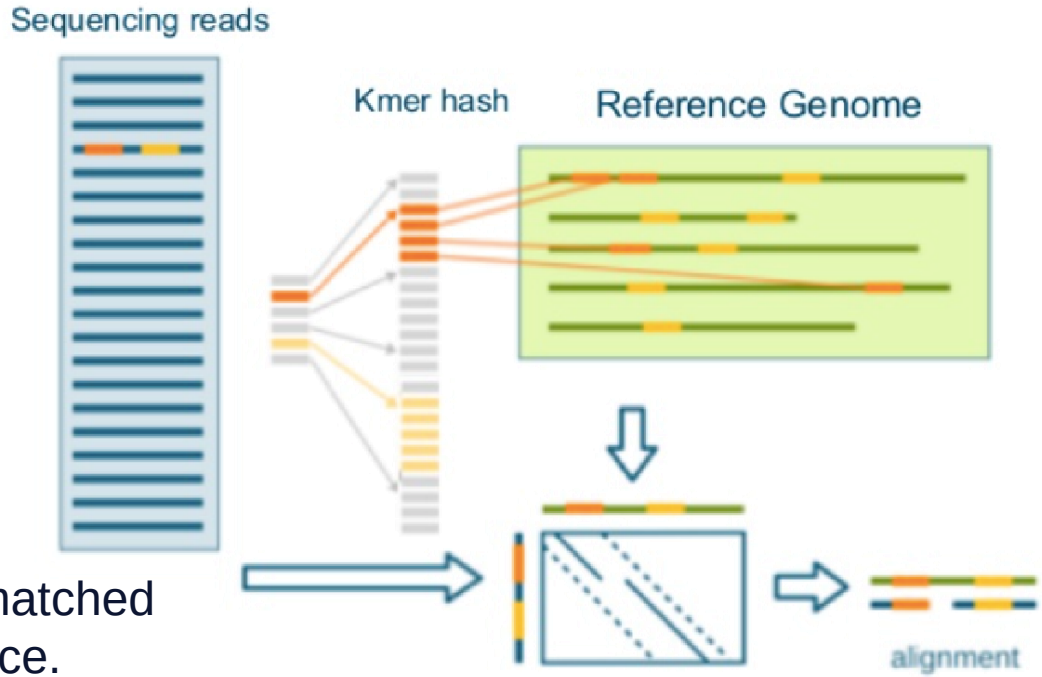
  - Millions of short reads to map to an entire genome
    ➔ data structure for rapid matching is needed.

  - 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.

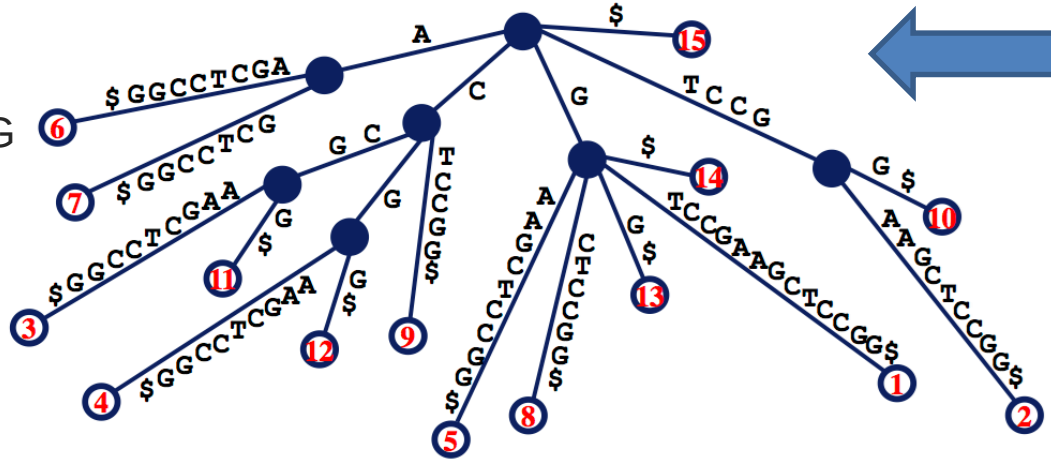


  - Based on perfect matches, align mismatching parts of reads.

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**. (dominant 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

- 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

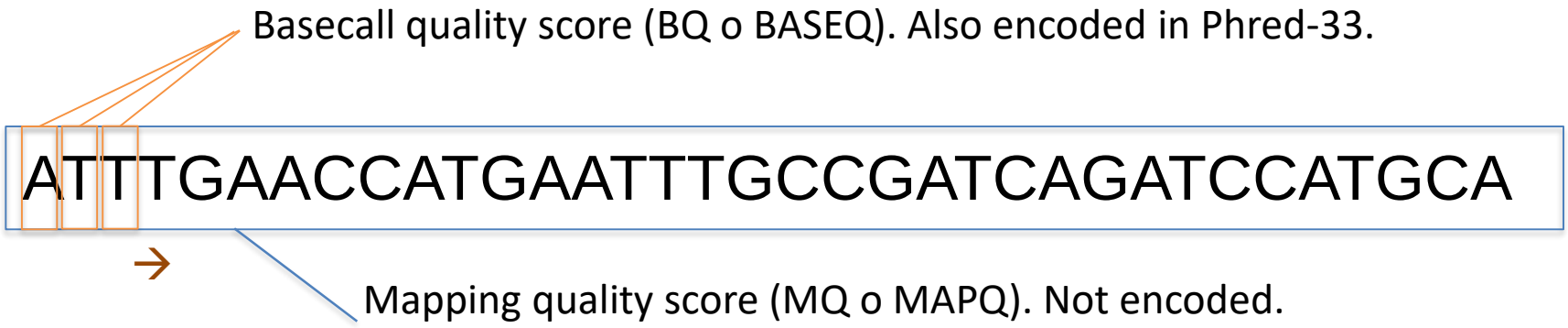
Ref …AAACTTAGTGCAACAGTGCACGAG…

Read …AAAC**AGTGCAACAGTGCACGAG…

source: real deletion or region difficult to align

Phred Quality Score Revisited

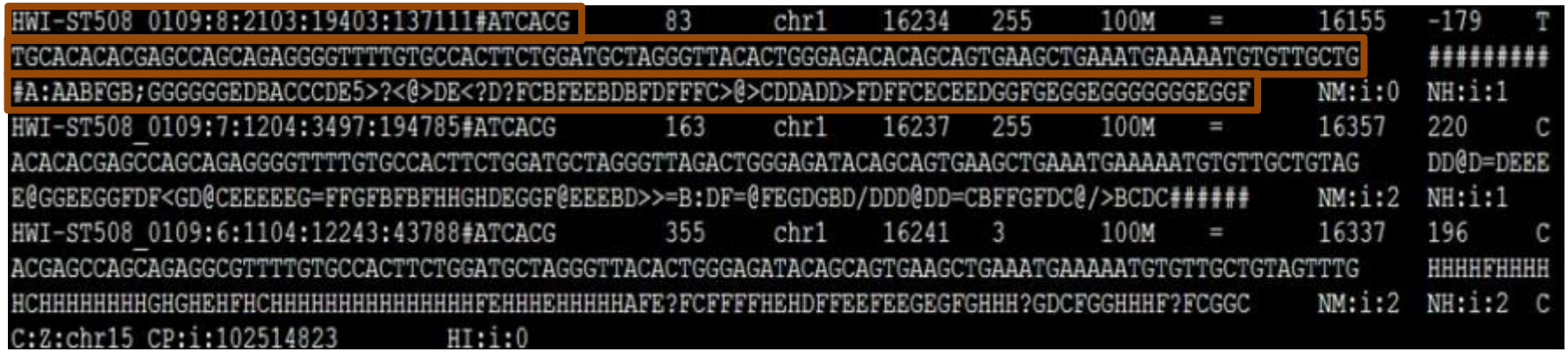
- Phred quality scores are also used to quantify **mapping** uncertainty.
  - Mapping quality is applied to a single read rather than individual bases.



  - 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)

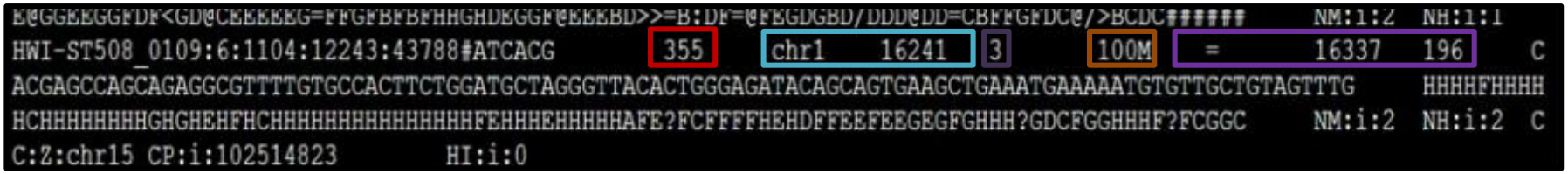
- Most popular fast-aligners (e.g. BWA, Bowtie2) take FASTQ as input and output SAM/BAM files.
  - Adds alignment information to FASTQ read data (i.e. position relative to reference, mapping quality, presence of insertions/deletions).



Let's take a look at an example using `samtools view`!

Alignment Information

- SAM/BAM file contains the following info about each read alignment.

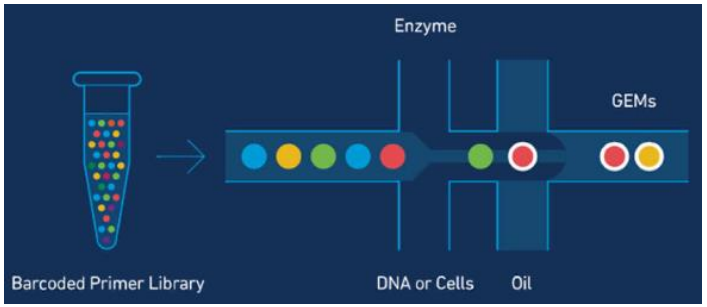


  - Position relative to reference where a read (inc. chromosome) and its corresponding read pair (incl. relative to the first half) are mapped.
  - Mapping quality (MAPQ).
  - The CIGAR (Concise Idiosyncratic Gapped Alignment Report*)
  - A Bitwise flag for additional information about the read.

*Definitely one of the worst backronyms

Plethora of Sequencing Methods

- Mainly talked about sequencing technologies as applied to DNA sequencing…and **bulk** RNA sequencing to an extent.
  - I.e. Where read depth proxy for mRNA (cDNA) quantity itself proxy for gene expression.
  - Yet… https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/ForAllYouSeqMethods.pdf
  - Includes single-cell which allows measurements on a per-cell basis.
  - More on **single-cell** RNA sequencing later today!



End of Sequence

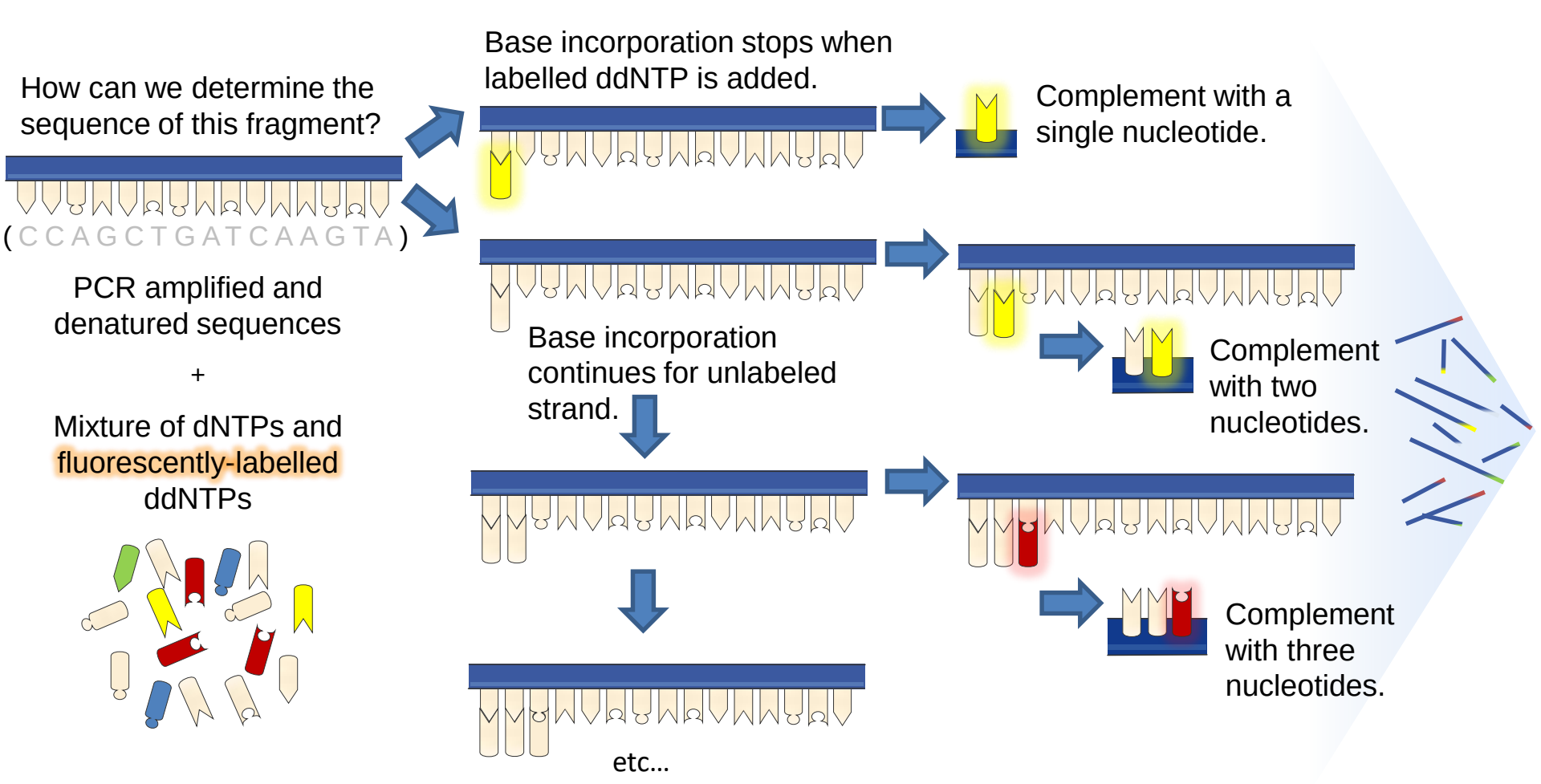
UNIVERSITY OF OXFORD

Thanks for listening!

Extra Slides Below…

UNIVERSITY OF OXFORD

How does Sanger work? 1/2

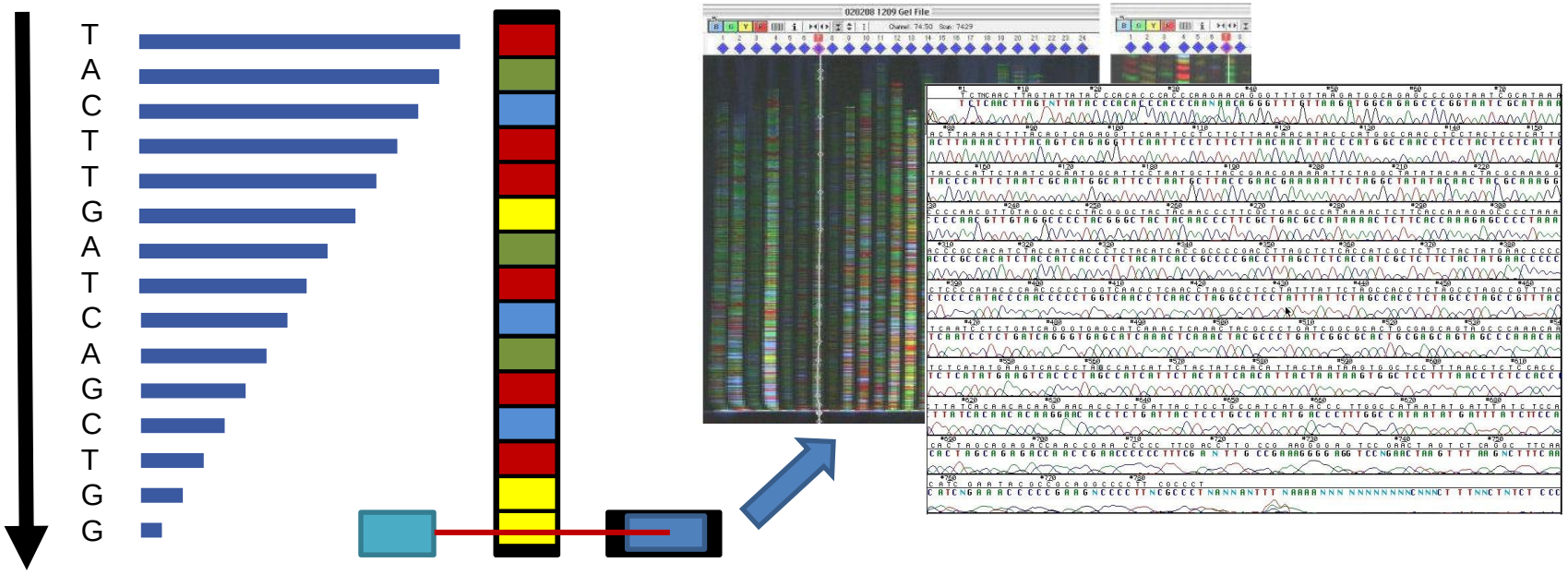


Random terminator incorporation process gives rise to multiple incomplete complement fragments where length acts as a proxy for position at which the chain terminating ddNTP was incorporated!

ddNTP = Dideoxynucleoside triphosphate (chain-terminating)

How does Sanger work? 2/2

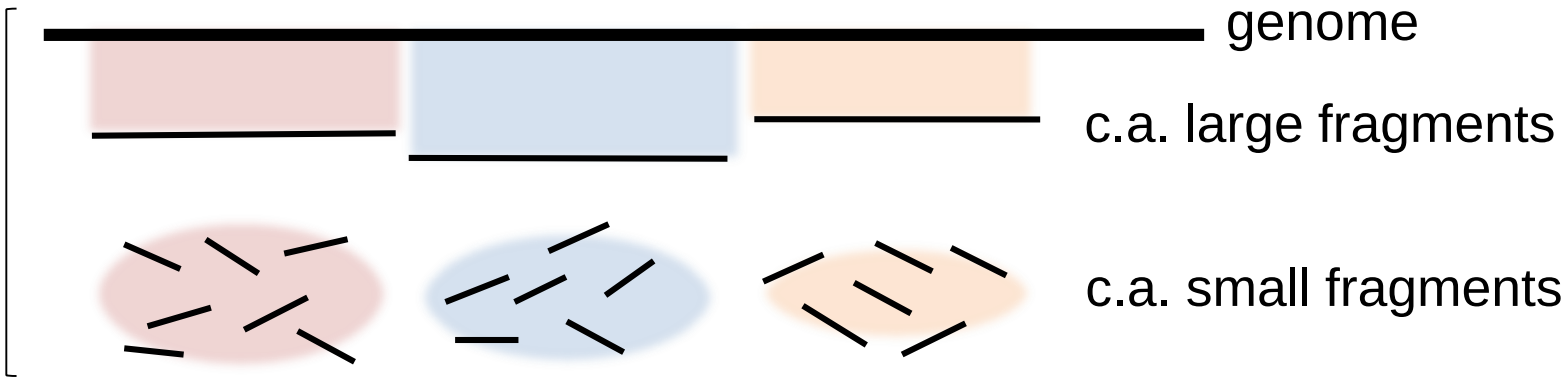
- Complement fragments passed through capillary tube using electrophoresis, their final ordering determined by size.
  - The smaller a fragment the faster the pass through.
  - ddNTP fluorescence captured via laser excitation.



Clone-by-clone Sequencing

- The HGP began by using a clone-by-clone approach to sequencing:
  - Genome broken into large 150 kbp fragments, positions of each fragment carefully recorded.
    - Fragments clonally amplified and then broken into **smaller overlapping** sequencing-ready, themselves clonally amplified.
      - Overlapping fragments over short region ➔ fairly straightforward assembly.

-laborious
-costly
-slow
…
-orderly



c.a = clonally amplified

The Human Genome Ref

- Since its first assembly, the human genome has seen 20 major updates.*
  - The last 5 version are available through either of these portals:

➢ **University of California Santa-Cruz**

https://genome.ucsc.edu

(downloads>Humans>hg38>Full data set>hg38.fa.gz)

➢ **Genome Reference Consortium**

https://www.ncbi.nlm.nih.gov/grc

(human>GRCh38.p13 (latest minor release) FTP>GCA_000001405.28_GRCh38.p13_genomic.fna.gz)

*and many many more minor updates

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**