DNA sequencing

Andrew Read

Emeritus Professor of Human Genetics

Manchester University

Boy Scientist Fails to Win Nobel Prize….

PhD Cambridge 1964

“A Method for Stepwise Degradation of RNA”

Real-world DNA Sequencing

- Sanger (dideoxy) sequencing

- “Next Generation” (massively parallel) sequencing
    - Illumina

- “Third Generation” (long read, single molecule) sequencing
    - PacBio
    - Oxford Nanopore

Why are you doing it?

- De novo sequencing / assembly

- Resequencing

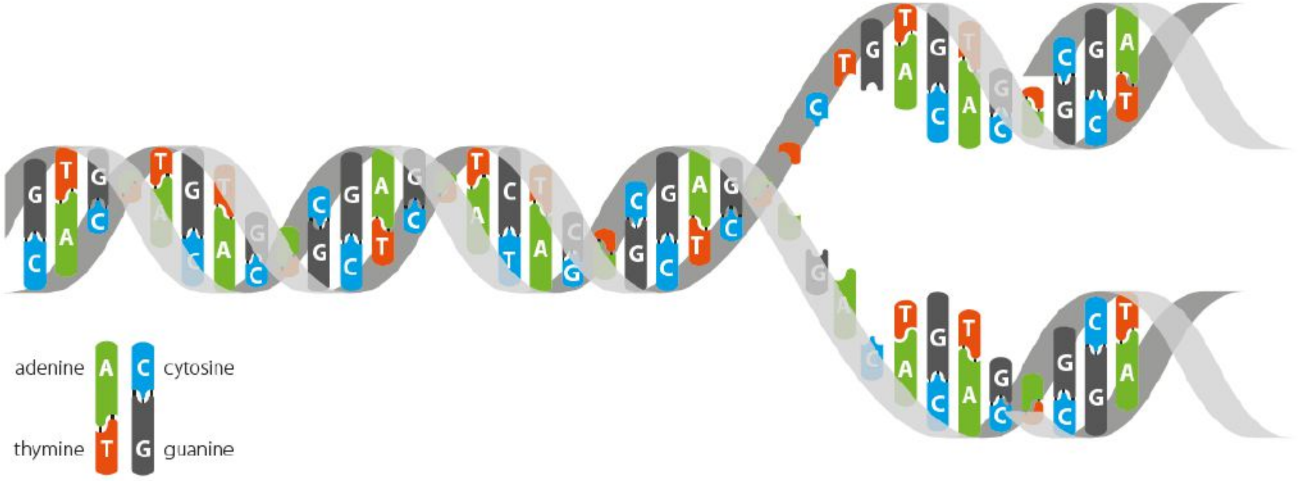
  Compare an individual's genome sequence to the Reference Sequence.

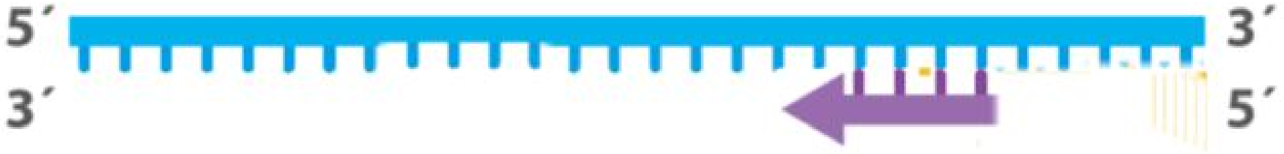
  A typical healthy individual's genome has 4-5 million differences compared to the Reference Sequence.

  - 4 - 4.5 million single nucleotide variants (SNVs or SNPs)
  - 700,000 indels (small insertions or deletions, <50bp)
  - 25,000 structural variants (insertions, deletions, rearrangements >50bp))

Eichler *New Engl J Med* **318:** 64-74; 2019

DNA synthesis





DNA polymerases never synthesise a new strand from scratch. They require a primer.

They always work by extending the primer in the 5′ – 3′ direction.

In the lab, primers are typically synthetic 15-25 base single-stranded oligonucleotides.

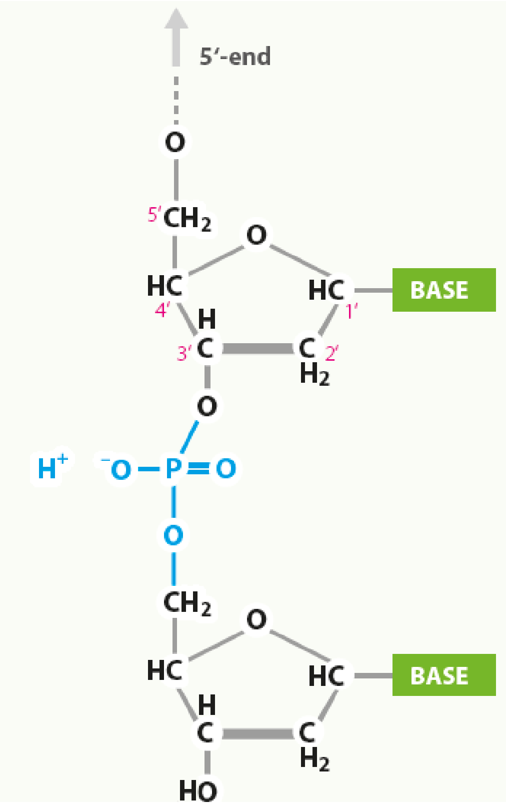
Sanger (Dideoxy) Sequencing

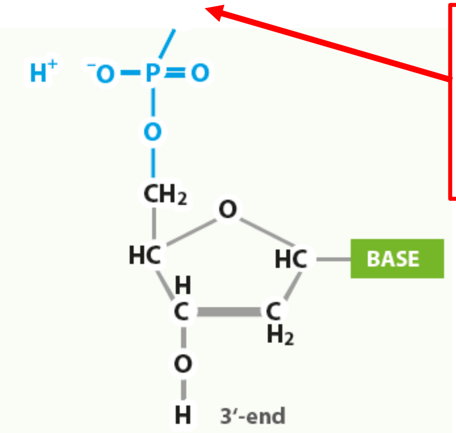
Fred Sanger 1977: Sequencing by Synthesis

- Pool of identical single-stranded DNA fragments (e.g. denatured PCR products)
- Add primer, DNA polymerase, 4 monomers

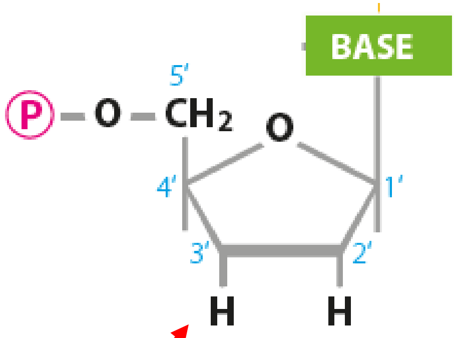
  → complementary strand synthesised
- Pool of monomers is spiked with chain-terminating dideoxy monomers

Chain growth and chain termination





DNA Polymerase attaches the next nucleotide to the 3’hydroxyl group at the end of the growing chain



A dideoxy nucleotide

There is no hydroxyl group for DNA polymerase to attach the next nucleotide, so synthesis comes to a stop

Sanger Sequencing – the Original Method

- A P^{32} or S^{35} - labelled primer is used, so that all the newly synthesised strands are radiolabelled.

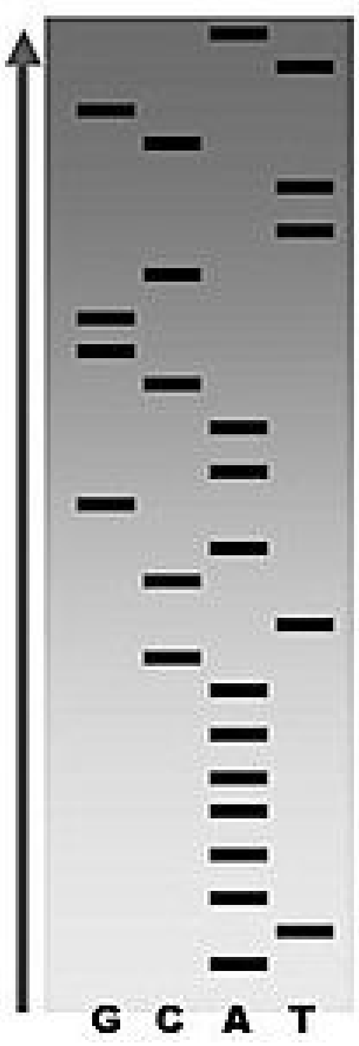
- Four separate reactions are run, in each of which one of the four monomer pools is spiked with about 1% of the corresponding chain terminator

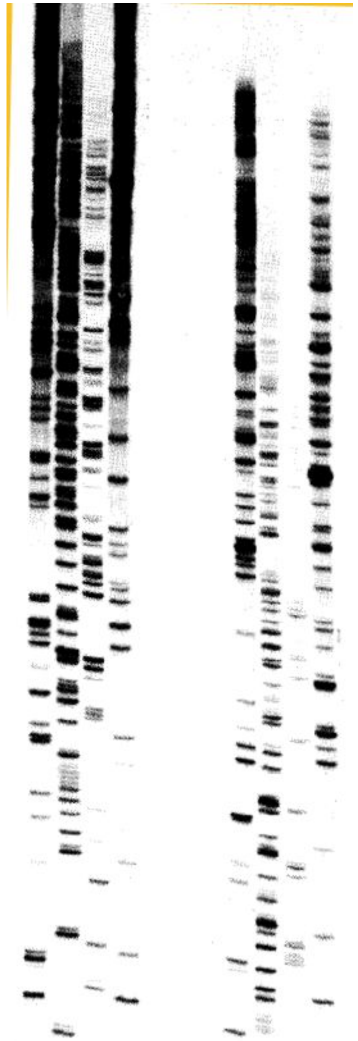
- The products are separated by gel electrophoresis and detected by autoradiography.

- Products of the four reactions are run in adjacent lanes: in this example, the reaction using the G-terminator on the left, the one with the C terminator next, etc.

The sequence can be read:

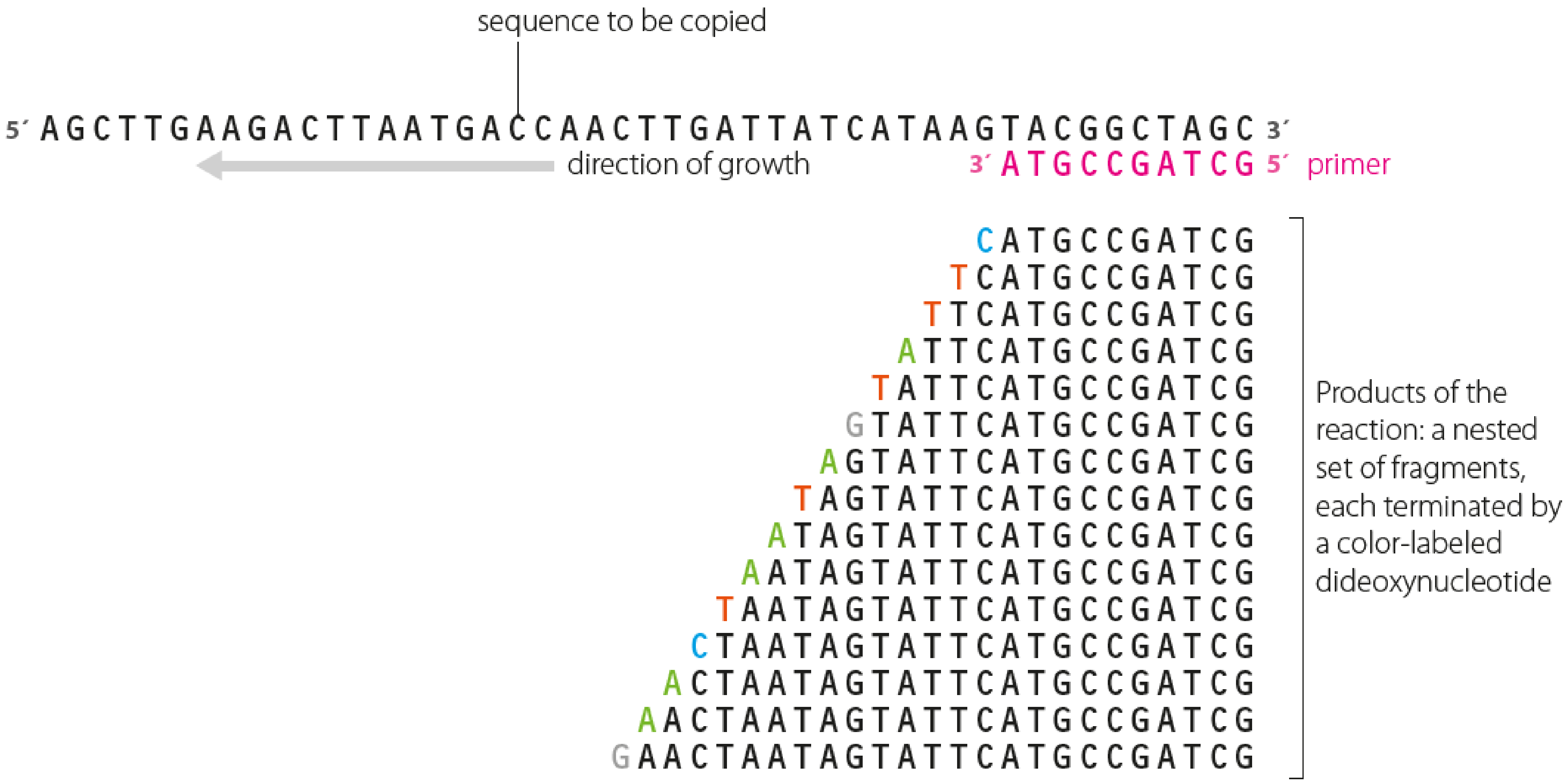
ATGCTTCGGCAAGACTCAAAAAATA





Sanger Sequencing with fluorescent labelling

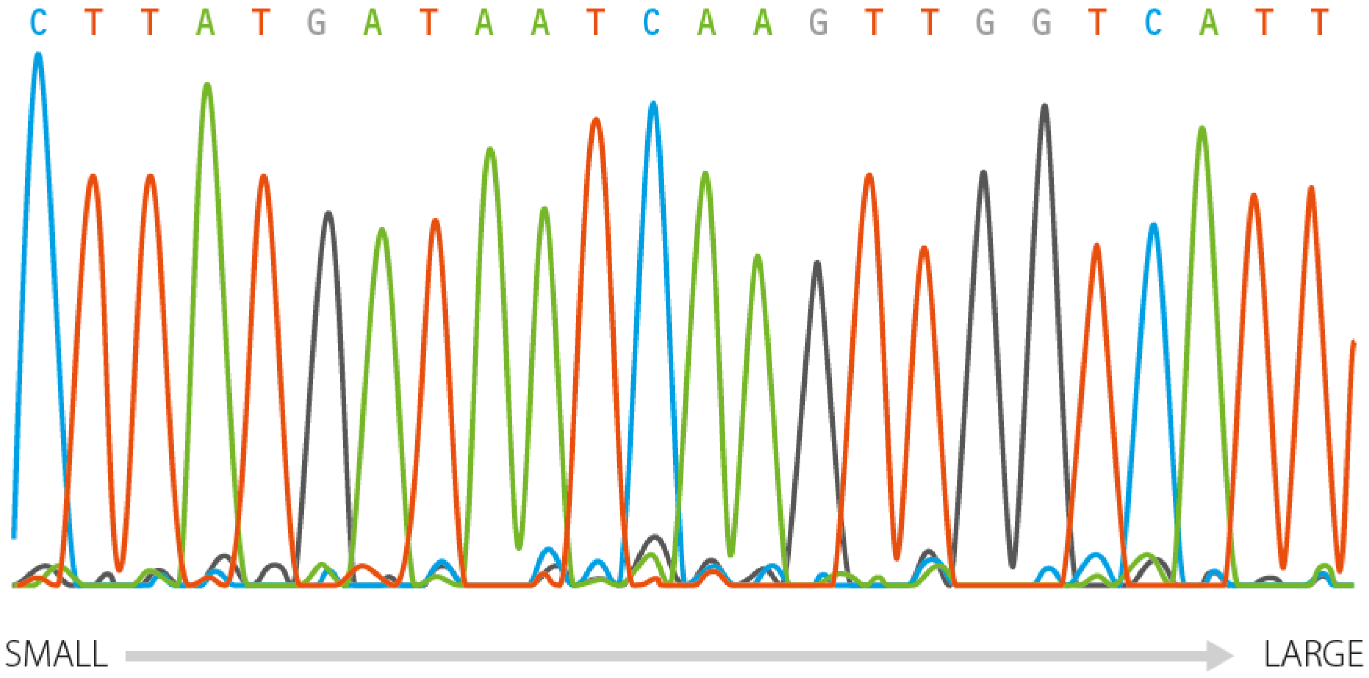
A single reaction with the four chain terminators labelled with a different colour fluorophore



An ABI automated sequencer.

Products of independent reactions run through 96 capillary electrophoresis systems, colours detected as each set of products passes lasers.

Output:



Sanger Sequencing – Advantages and Disadvantages

Advantages

- Very accurate: up to 800bp of very accurate sequence.
- Targeted: the choice of primer dictates where the sequencing starts.
- Relatively cheap and simple: most labs can do it.

Disadvantage

- Relatively low throughput: doing a whole genome would need millions of sequencing runs. The Human Genome Project used big banks of ABI sequencers.

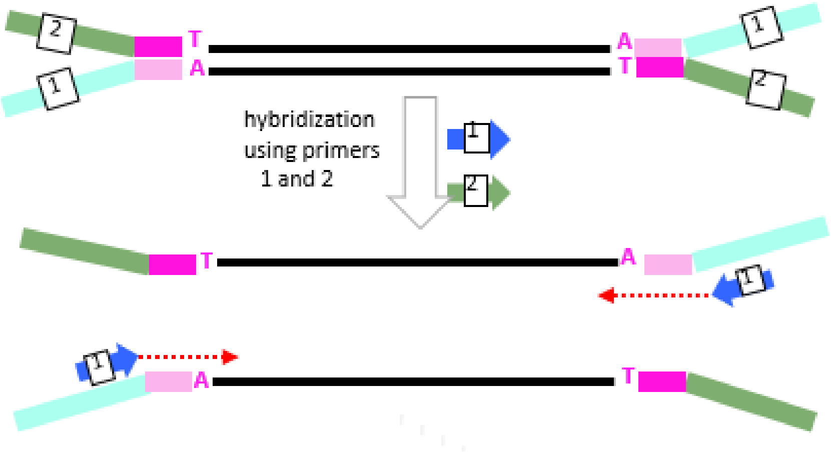
“Next Generation” massively parallel sequencing

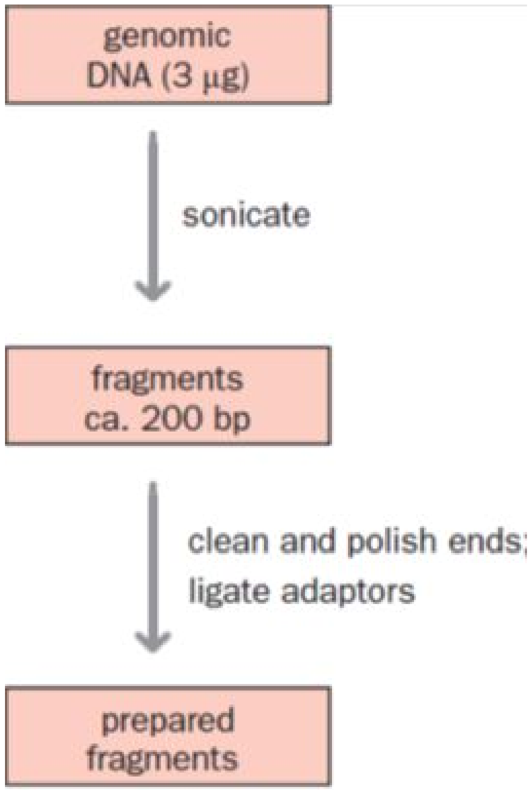
- Starting around 2005, a rash of new methods using different technologies, but all having in common that they were massively parallel – i.e. millions of individual sequencing reactions were run in parallel.
- Intense competition between companies spurred rapid technological advance, particularly regarding throughput, accuracy, cost and convenience.
- Unlike Sanger sequencing, they are non-selective – i.e. you can’t decide what gene or sequence to target, except when selecting the input DNA; they just sequence whatever you put in.
- Often need complex procedure to construct the ‘sequencing library’ of input DNA
- Many competing technologies, but main winner has been Illumina.

Illumina sequencing (1)

Library preparation:

- Randomly fragmented input DNA
- Size-selected
- Universal adapters ligated to the ends.




genomic DNA (3 μg)

sonicate

fragments ca. 200 bp

clean and polish ends; ligate adaptors

prepared fragments

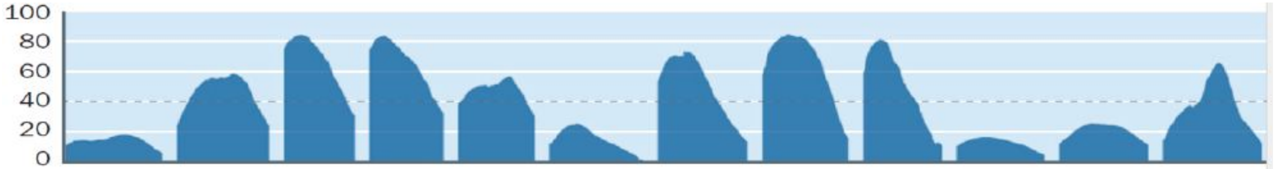
Exome vs. Whole-genome sequencing

- For most purposes interest focuses on the protein-coding sequences.

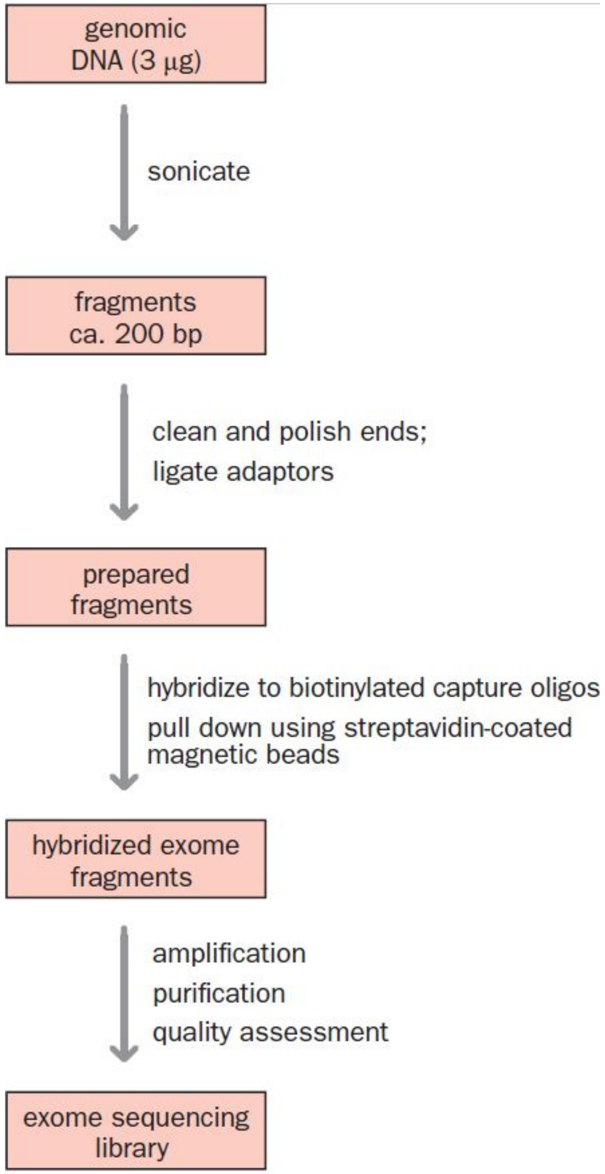
- These comprise only ca.2% of the human genome.

- Use exon capture to prepare a sequencing library consisting of just the DNA of every exon.

- Unequal representation of exons in the final library is a problem



E.g. the 12 exons of the *PSK9* gene as seen in the ExAC database.



genomic DNA (3 μg)

sonicate

fragments ca. 200 bp

clean and polish ends; ligate adaptors

prepared fragments

hybridize to biotinylated capture oligos pull down using streptavidin-coated magnetic beads

hybridized exome fragments

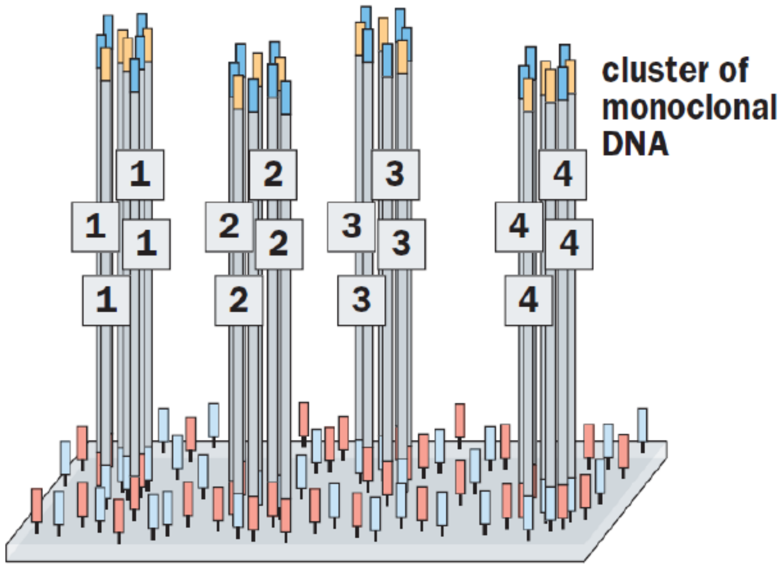
amplification purification quality assessment

exome sequencing library

Illumina sequencing (2)

Preparing for sequencing:

- Library loaded into flow cell which has millions of primers matching the adapters on the library covalently anchored to the cell plate.



HGM5 Fig 6-24

- Complex form of PCR produces millions of clusters, each consisting of many copies of the same fragment, anchored at scattered locations across the cell.

Illumina sequencing (3)

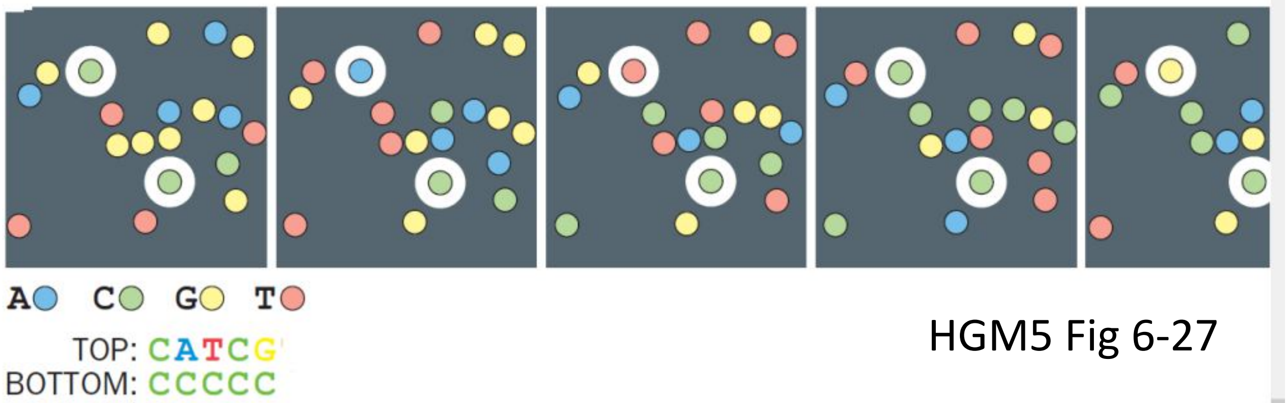
- Sequencing by synthesis like in Sanger sequencing. Chain-terminating fluorescently labelled monomers, **BUT:**

- No normal monomers, 100% terminators, so only one nucleotide added.

- Chain terminating 3’ blocking groups and fluorescent labels are removable.

- Read colour, then remove blocking groups so that synthesis can proceed to the next nucleotide

- Keep repeating and recording the colour each time.

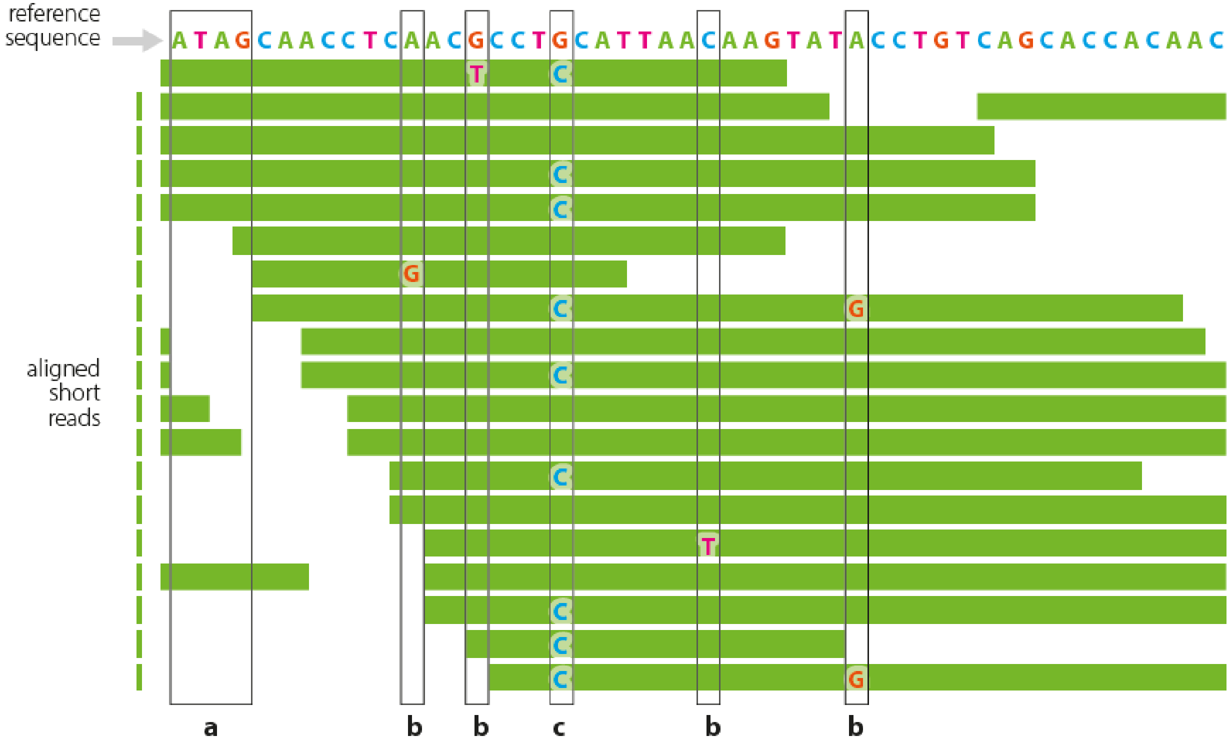


HGM5 Fig 6-27

Illumina sequencing (4)

Assembling the fragments:

- Initial file of colour photos converted into sequence of each cluster
  → millions of independent short reads (35-100 bp)
- Short reads are aligned to reference genome sequence:

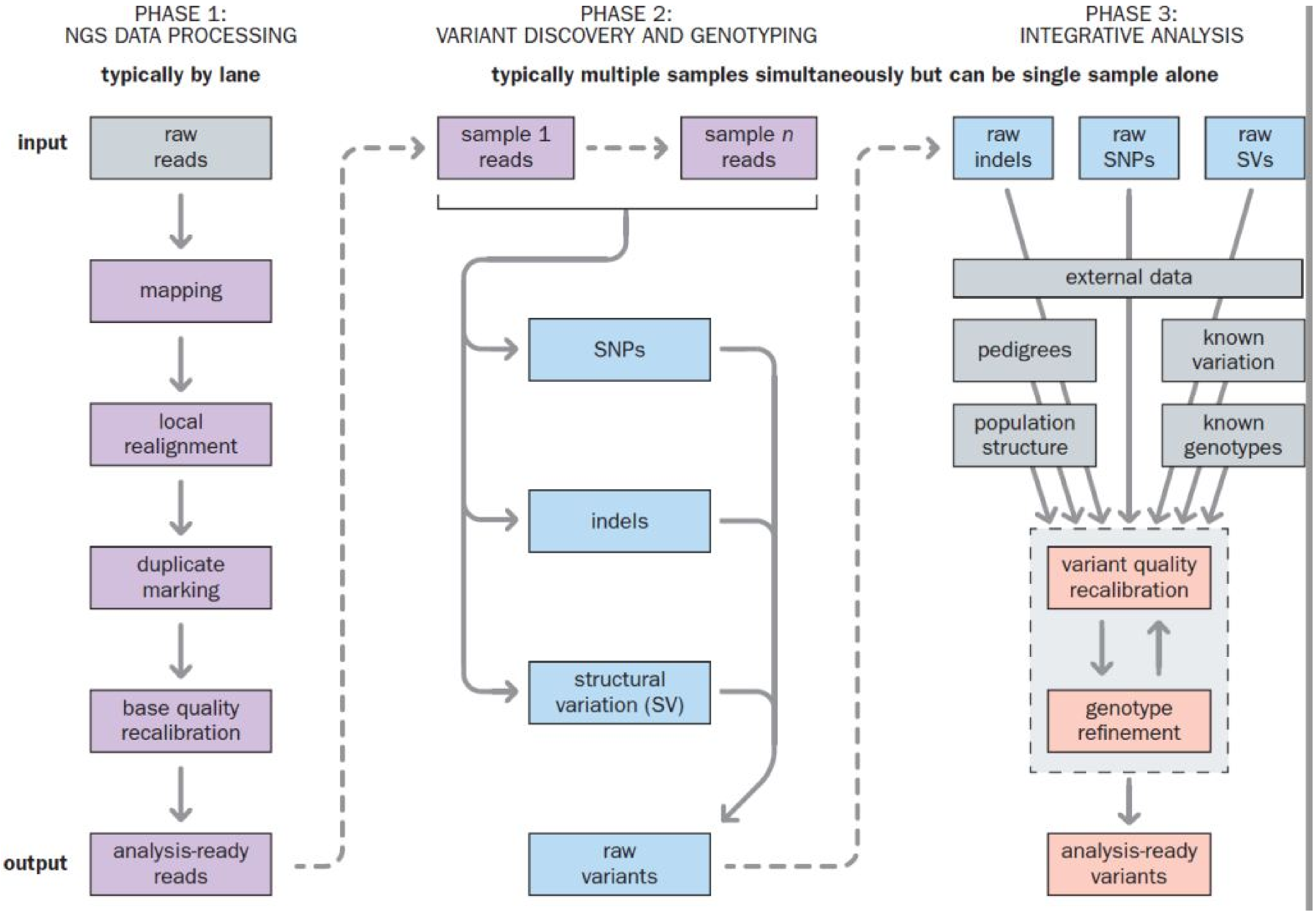


a: a region with poor coverage

b: the variants in these tracks are probably sequencing errors

c: at this position the subject is most likely heterozygous G/C

A real example would have much greater read depth.


PHASE 1:
NGS DATA PROCESSING
typically by lane
input
raw reads
mapping
local realignment
duplicate marking
base quality recalibration
output
analysis-ready reads
PHASE 2:
VARIANT DISCOVERY AND GENOTYPING
typically multiple samples simultaneously but can be single sample alone
sample 1 reads
sample *n* reads
SNPs
indels
structural variation (SV)
raw variants
PHASE 3:
INTEGRATIVE ANALYSIS
raw indels
raw SNPs
raw SVs
external data
pedigrees
known variation
population structure
known genotypes
variant quality recalibration
genotype refinement
analysis-ready variants

Illumina sequencing – advantages and disadvantages

Advantages

- Massive throughput: can easily sequence whole genomes (if you're rich enough).
- Can use selective libraries: e.g. libraries of exome (all 180,000 exons) or all genes in a disease-specific panel.

Disadvantages

- Short reads: 35 up to maybe 200 bp
- Uneven representation: some sequences amplify poorly or not at all in library preparation, especially in selective libraries.
- Alignment problems: the short reads don't cope well with repetitive sequences or structural variations (deletions, insertions etc).

https://emea.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Long Read Single-Molecule Sequencing

(PacBio, Oxford Nanopore)

Advantages

- Long reads (>10kb) can handle complex structural variants.
- Can detect epigenetic base modifications: sequence the original DNA, not a PCR copy

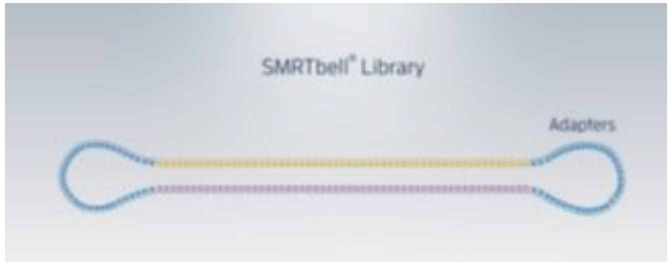
Disadvantages

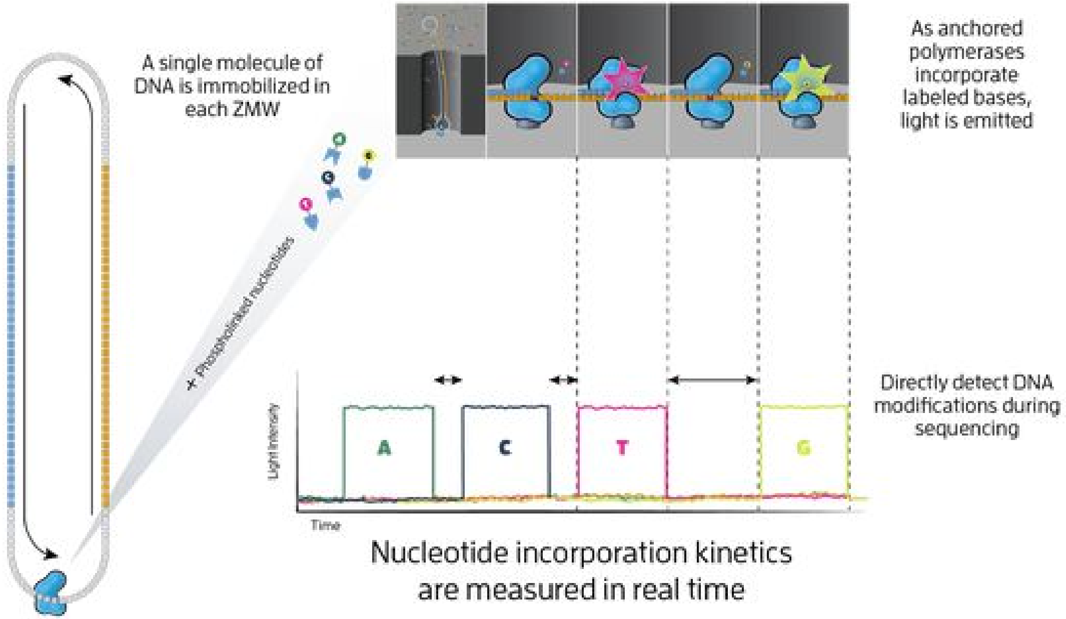
- Compared to Illumina, expensive and relatively low throughput.
- Nanopore technology (but not PacBio) has a high error rate: **4-10%.**

PacBio sequencing

- Individual molecules are located in wells (‘zero mode waveguides’) where an anchored polymerase adds labelled monomers and an optical system records the light emitted as a nucleotide is added.

- Hairpin adapters are ligated to the double-stranded DNA to form a continuous loop




A single molecule of DNA is immobilized in each ZMW
+ Phospholinked nucleotides
As anchored polymerases incorporate labeled bases, light is emitted
Light intensity
Time
A
C
T
G
Directly detect DNA modifications during sequencing
Nucleotide incorporation kinetics are measured in real time

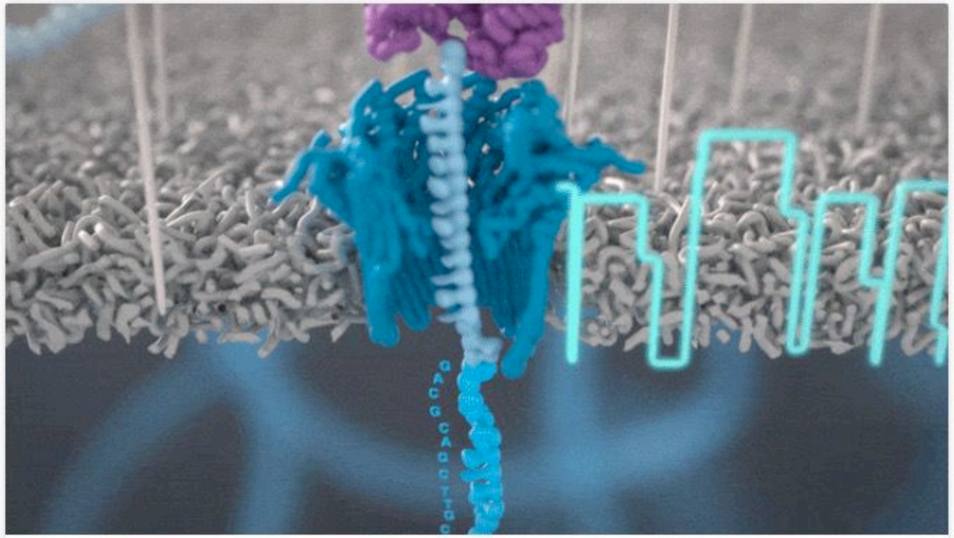
- The loop goes round and round, so that the same sequence is analysed many times. Random errors are averaged out, giving extremely high overall accuracy.

Applications of PacBio sequencing

- Up to 20,000 base pairs reads. Not a routine general sequencing method (too slow and too expensive) but probably the highest accuracy of any system.
- Use in resequencing, to sequence through repetitive regions and large structural variants that short-read sequencing (Illumina etc) can't handle.
- Typical application would be to use Illumina etc sequencing to define the straightforward parts of the input DNA, then tidy up with PacBio.
- This strategy used to produce the first complete telomere-to-telomere human chromosome sequences, 20 years after the Human Genome Project declared the sequence 'finished' (see Nature **593**:101-7; 2021).
- Also used to catalogue epigenetic changes (NB. Sanger or Illumina techniques sequence a **copy** of the input DNA, all epigenetic information is lost).

Oxford Nanopore system

- A single strand of DNA is pulled through a protein pore in an insulating membrane.
- The electric current flowing through the pore is monitored.
- Each nucleotide blocks the pore to a different degree, changing the current flowing through the pore.



-ve

Insulating membrane

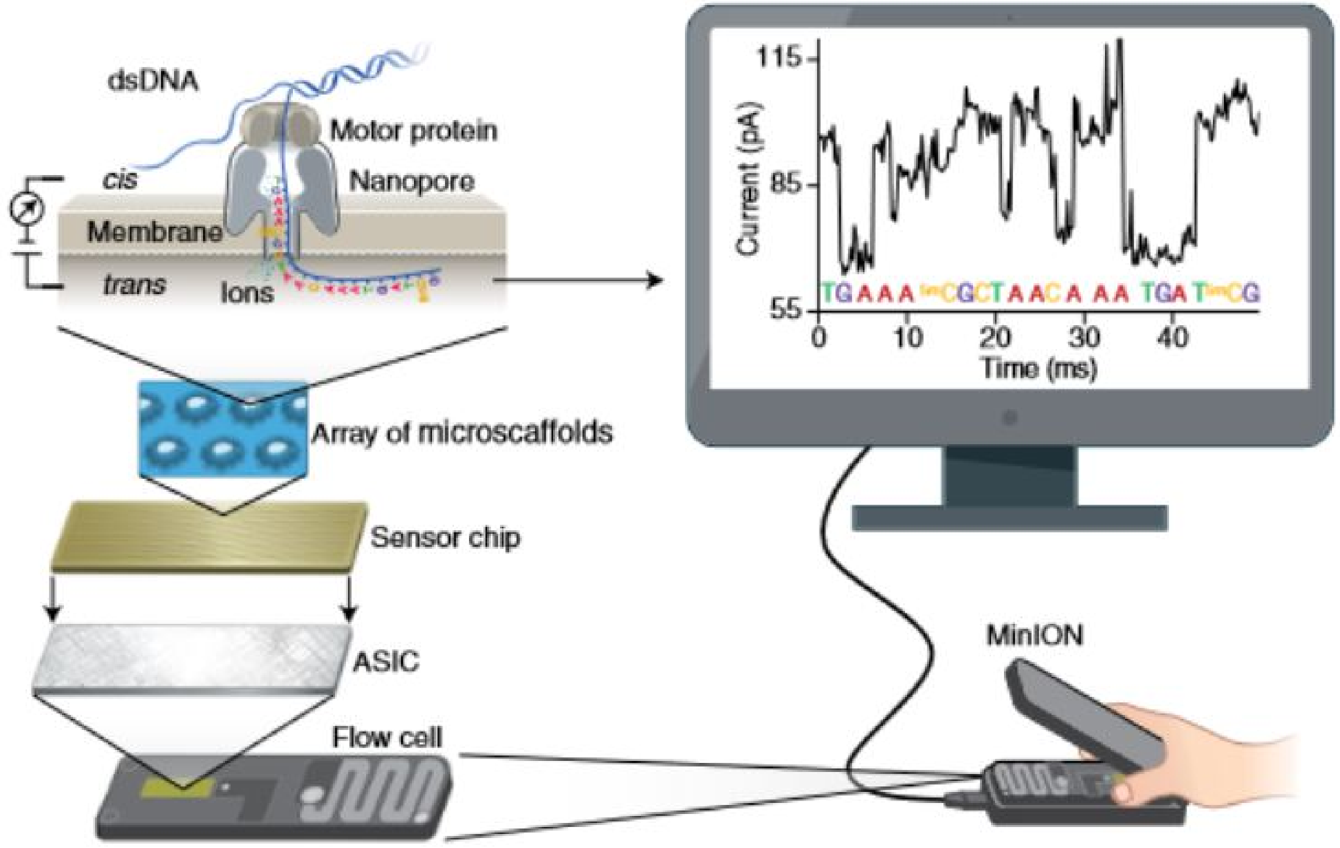
+ve

Current

Screen grab from https://nanoporetech.com/platform/technology

Nanopore sequencing

Oxford Nanopore Minion device



Wang et al, Nature Biotechnology **39:** 1348-65; 2021

Applications of Nanopore sequencing

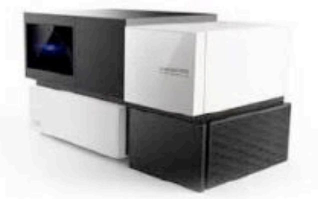
- Potentially a routine general sequencing method, replacing Illumina etc.
- Extremely long reads (>100,000 bases), portability and real-time sequence readout are the main strong points. Can sequence RNA as well as DNA.
- Much publicised success tracking Ebola virus etc outbreaks in remote locations.
- Requires relatively large amounts of input DNA. Error rate is still a major problem.
- Also used to catalogue epigenetic changes (NB. Sanger or Illumina techniques sequence a **copy** of the input DNA, all epigenetic information is lost).

See Nature **614**: 789-800; 2023 for a review

Other ‘new toys’?

Sequencing:

- *GeneReader*
- *Genius*
- *Promethlon/Gridlon*
- *QuantuMDx*
- *MGI T7*
- …



Long-read mapping:

- Bionano Genomics
- Nabsys