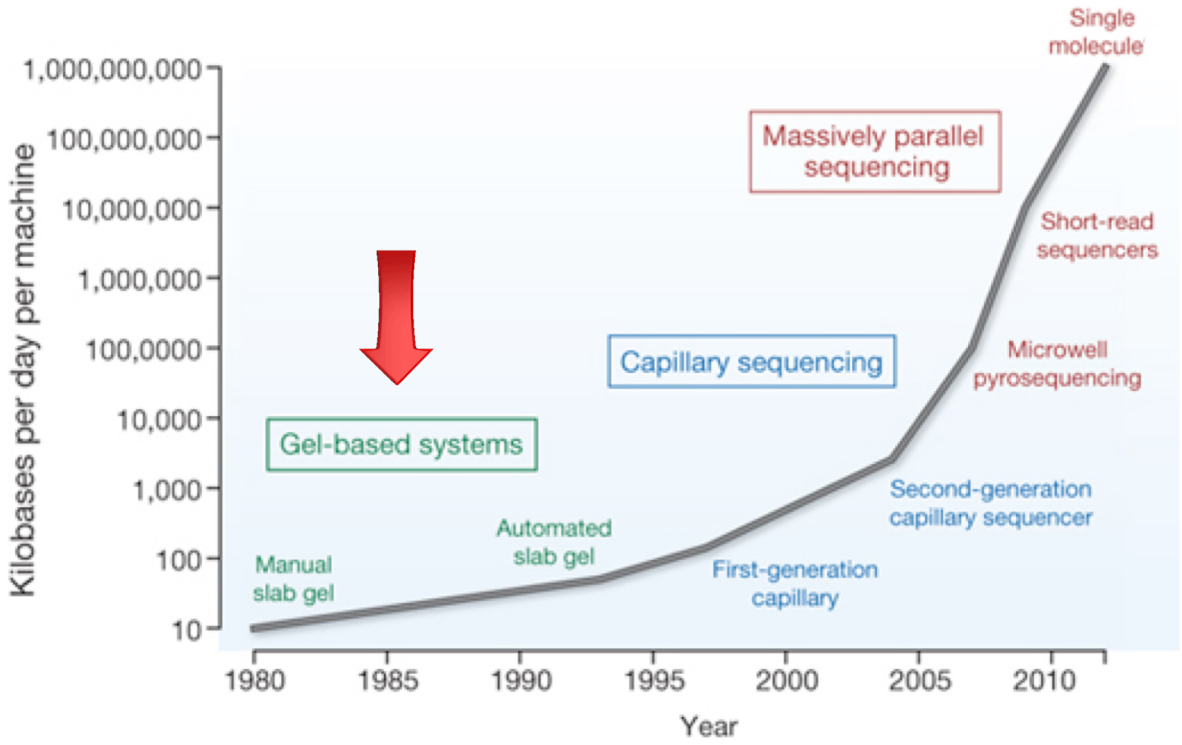
Microarrays vs RNA-seq

- Microarrays:
  - Consolidated technology
  - Well characterised bias
  - Differential expression at gene/ exon level, non-coding RNAs
  - No absolute expression levels
  - A lot of public data
  - Very cheap

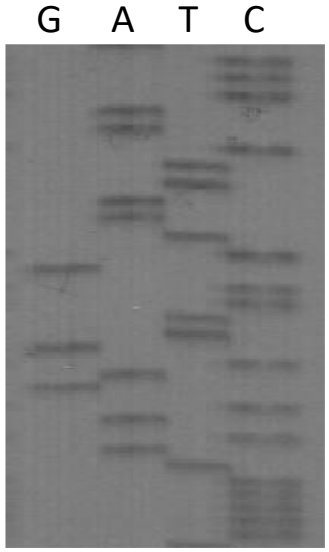
- RNA-seq:
  - Under development technology
  - Bias only partially characterized
  - Differential expression at gene level/exon level, non-coding RNAs
  - Absolute expression
  - Few public data
  - Getting cheaper
  - mRNA isoforms deconvolution
  - **Fusion products**
  - SNP
  - Computational demanding

Critical issues in red

History of DNA analysis

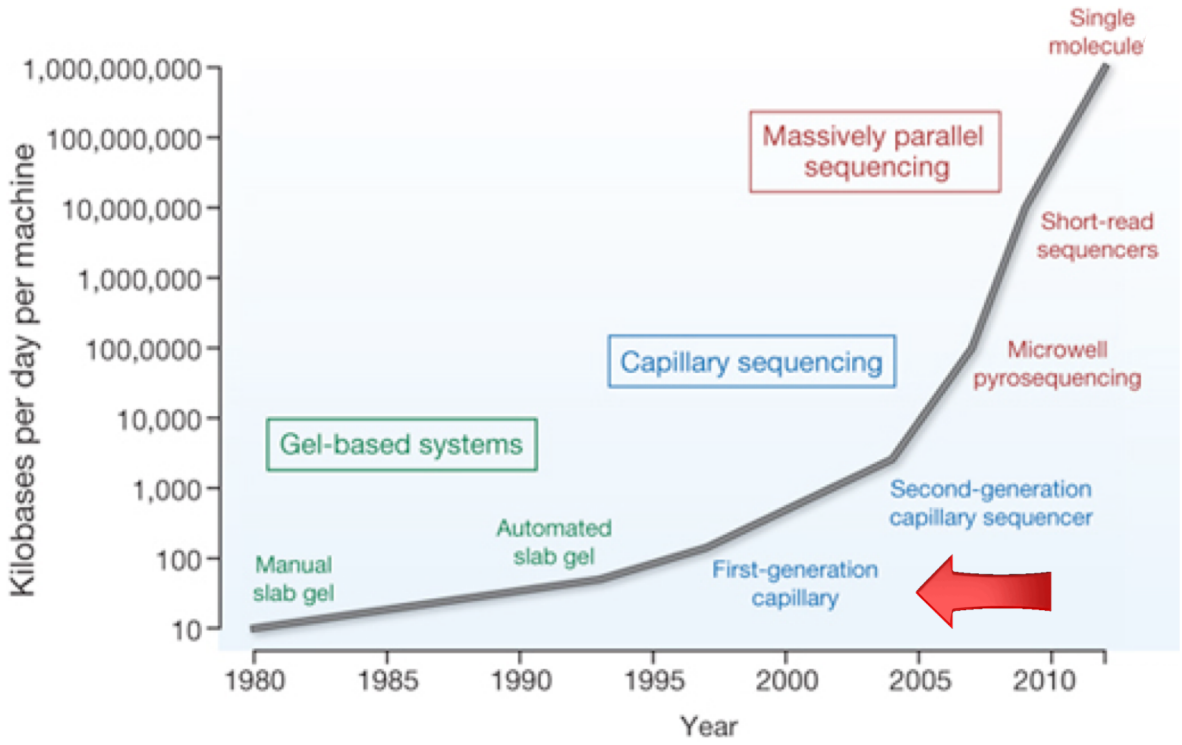


Manual slab gels

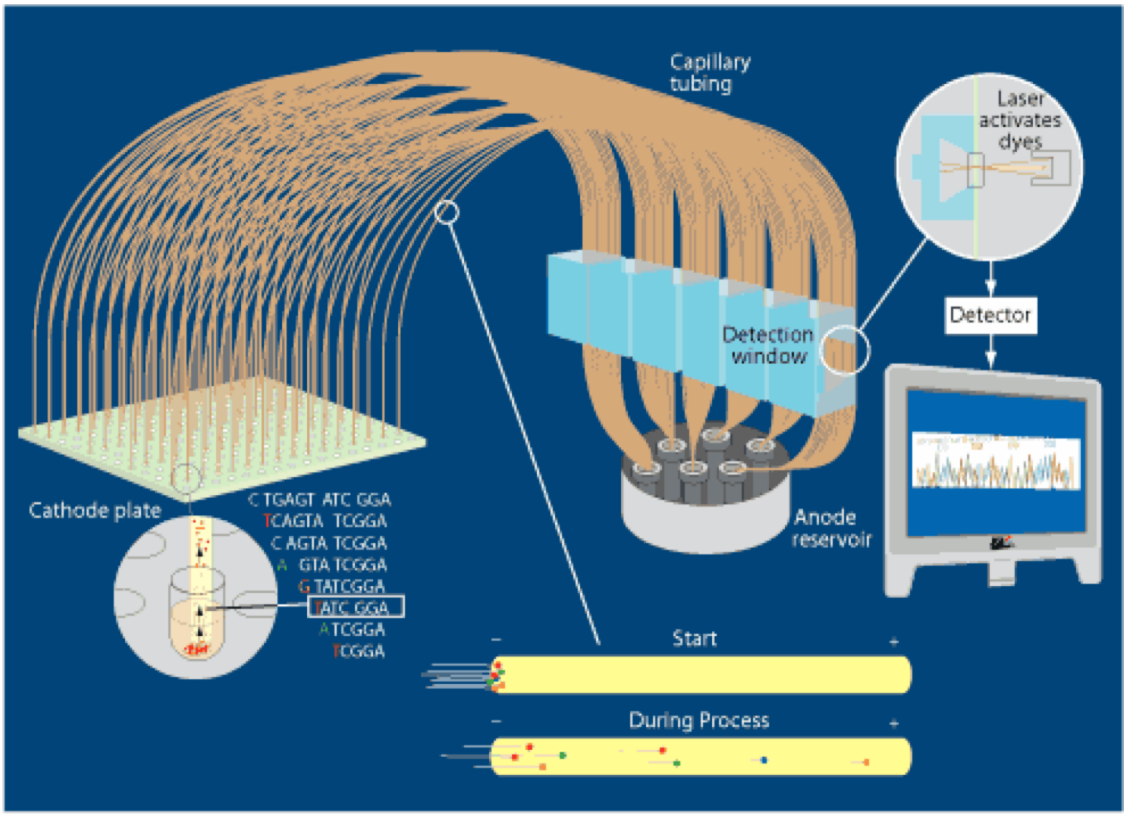


- Unfeasible the sequence of large genomes

History of DNA analysis

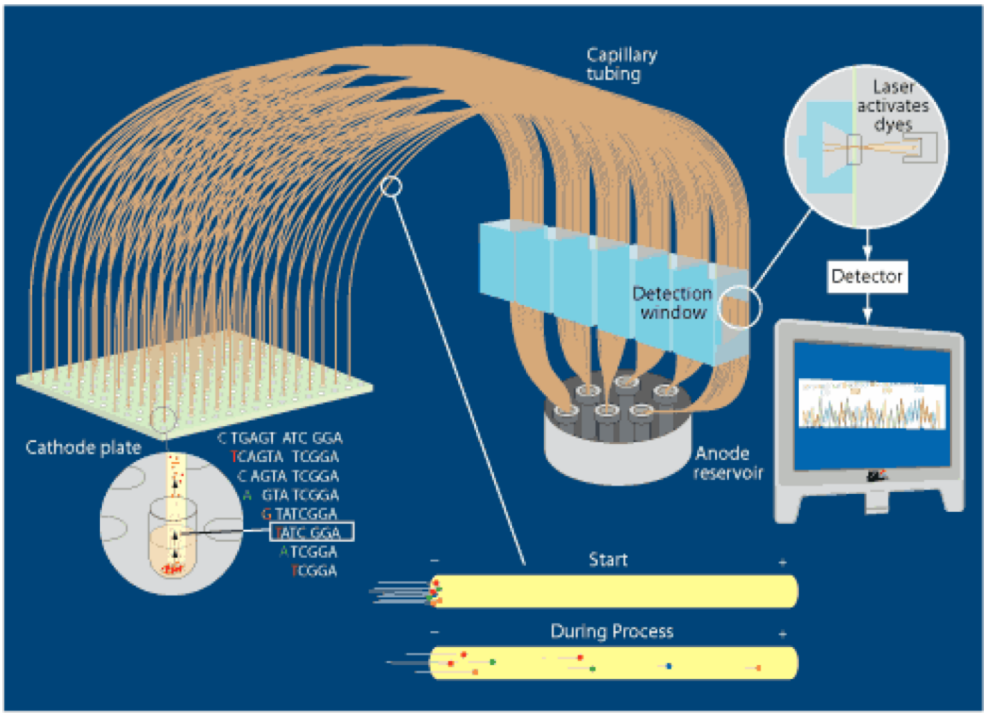


Capillary DNA sequence



In capillary sequencing machines, DNA fragments are separated by size through a long, thin, acrylic-fibre capillary (instead of anelectrophoresis gel, as with the Sanger method).

Capillary DNA sequence



1 A sample containing fragments of DNA is injected into the capillary. This is done by dipping the capillary and an electrode into a solution of the sample, and briefly applying an electric current. This causes the DNA fragments to migrate on to the end of the capillary.

2 Once the sample has been injected, the electric field is reapplied, to drive the DNA fragments through the capillary.

3 A fluorescence detecting laser, built into the machine, then shoots through the capillary fibre, causing the coloured tags on the DNA fragments, to fluoresce. Each base terminator is labelled with a different colour: A = Green, C = Blue, G = Yellow and T = Red.

4 The colour of the fluorescent bases is detected by a camera, and recorded by the sequencing machine.

5 The colours of the bases are then displayed on a computer as a graph of different coloured peaks.

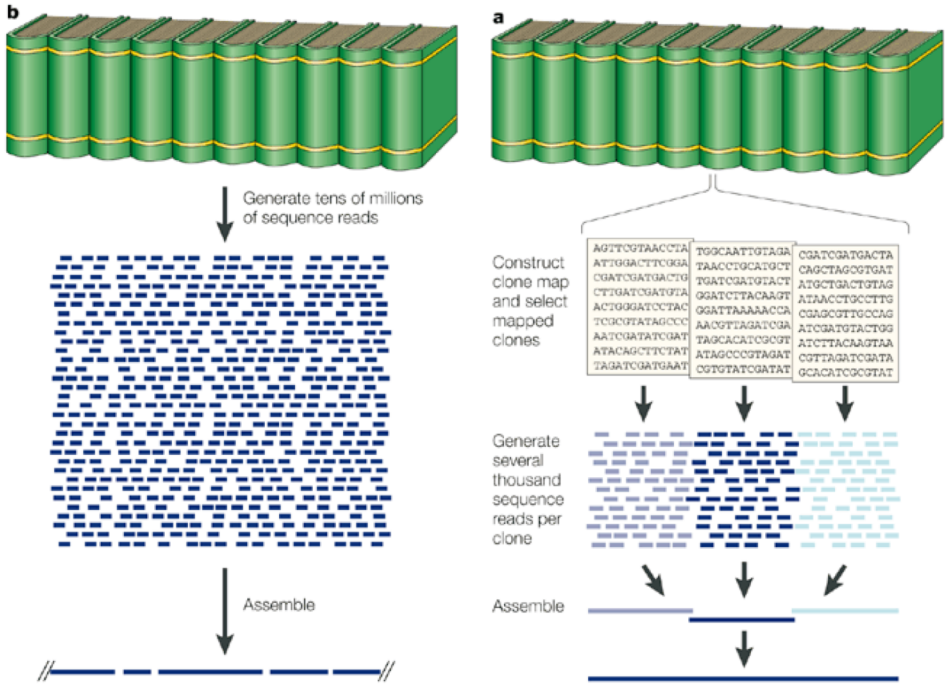
2001: G. Venter genome

5.11X

2001: The human genome project

4.5X

100 millions $



ThermoFisher SCIENTIFIC

3730XL (96 capillaries):

- Sequence length: 400 - 900 bases
- Samples/day: 3840
- Max sequence coverage: 2.1 Mb

Roche 454 SEQUENCING

GS FLX Titanium XL+:

- Sequence length: 400 - 1000 bases
- Samples/day: 3840
- Max sequence coverage: 700 Mb

ThermoFisher SCIENTIFIC

Ion Proton™:

- Sequence length: up to 200 bases
- Max sequence coverage: 10 Gb

illumina

NextSeq 500

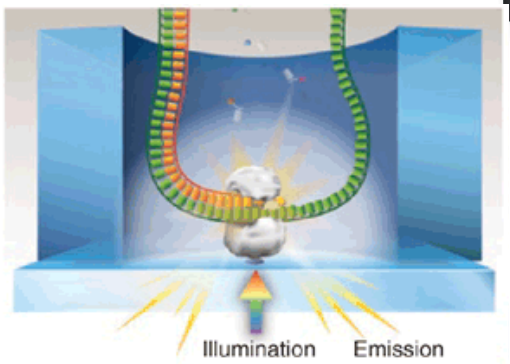
NextSeq 500:

- Sequence length: 75-300 bases
- Max sequence coverage: 120 Gb
- One 40X Human Genome in 30 hours
- ≤ 4000 $ 40 X Human Genome

NovaSeq 6000

The NovaSeq 6000 System offers output up to 6 Tb and 20 B reads in < 2 days

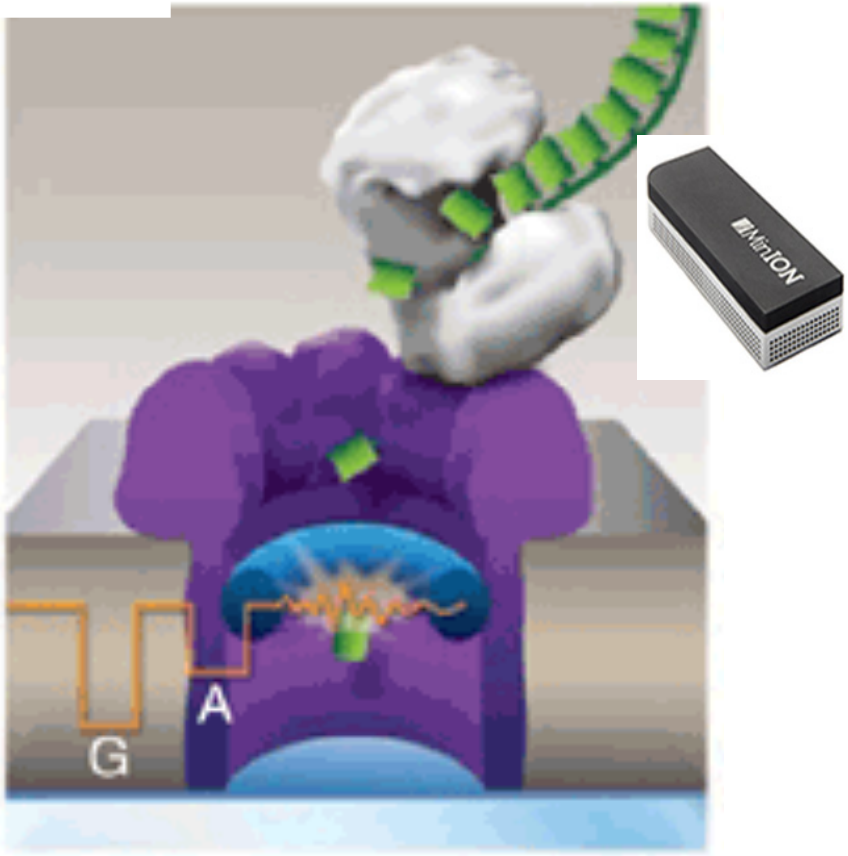
PACIFIC BIOSCIENCES™



Based on DNA polymerase

Sequel:

- 1 million ZMWs per SMRT Cell
- Sequence length: 400 - 1000 bases
- 16 SMRT
- Max sequence coverage per SMRT cell: 0.5-1Gb



Oxford Nanopore

Mini ION:

- Disposable
- 512 channels
- Sequence length: up to 300 kb
- Max sequence coverage at 70 bps/48h: 6 Gb

Porin protein embedded in the lipid surface.
The modified protein alfa-hemolysin (from bacteria) is able to recognized the 4 nucleotides thanks to the modulation of the ionic current.
Exonuclease used fo cut the bases from the DNA seqeunce

"future nanopore sensing devices linked to cloud based analyses could run anywhere on anything" - Clive Brown, CTO of Oxford Nanopore

Application	# reads/sample	Run type	# read length (bp)	Remark
Transcriptome analysis				
Tag based (SAGE/CAGE)	>10 million	Single end	20–50	
SmallRNA	>10 million	Single end	20–50	
mRNA Seq	>30 million	Paired-end	>50	Efficient exclusion of rRNA derived sequences increases the resolution of the transcripts of interest
Ribosome profiling	>20 million	Single end	20–50	
ChIP-Seq	>20 million	Single or Paired-end	≥50	Specificity of the ChIP enzyme determines the # reads needed. Low specificity ~ more background = more reads needed
De novo sequencing	30× genome coverage, preferably more.	long single-end reads and Paired-end	As long as possible	Ideal PacBio long reads. Or combination of paired-end, mate-pair and PacBio.
Meta-genomics				
Tag based (ITS, 16S)	>100,000	Paired-end, long single-end reads	As long as possible	Complexity of the specific biosphere determines both the primer pairs and/or #reads per sample. Longer reads allow for better differentiation between related species
Shotgun	>100 million	Paired-end, long single reads	As long as possible	Complexity of the specific biosphere determines the library insert size and/or #reads per sample.
Methylation analysis				
Whole genome	>400 million	Paired-end	≥100	Ideal situation: all PacBio long reads on native/ unmodified shotgun libraries.
Enrichment strategies	>50 million	Paired-end	≥100	
Infections	>25 million	Single or Paired-end	≥100	~2% of cell-free DNA from plasma is of non-human origin
Non-invasive prenatal testing	>10–20 million	Single-end	>50	Trisomy detection from cell-free fetal DNA in maternal plasma
Disease gene identification diagnostics				
Whole genome	1 billion	Paired-end	≥100	30× average coverage
Exome (50 Mb)	>60 million	Paired-end	≥100	50× average coverage, >75% on target