SEPTEMBER
2023

High throughput sequencing

Functional genomic data analysis: transcriptomics

Stéphane Le Crom
stephane.le_crom@sorbonne-universite.fr

SORBONNE UNIVERSITÉ

First generation sequencing methods

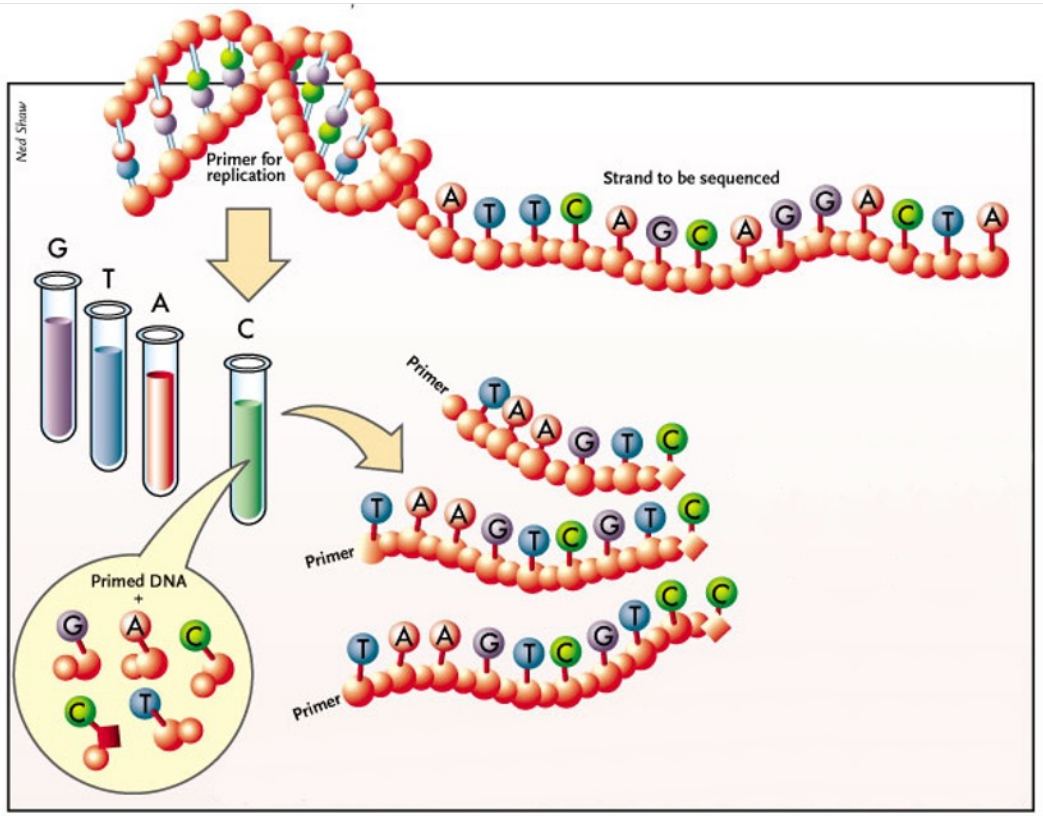
Sanger sequencing by synthesis

Method discovered in 1977 by Frédérick Sanger (nobel price 1980).

DNA polymerisation using a **complementary primer**. Elongation using **thermostable DNA polymerase** (PCR).

Addition of 4 **deoxynucleotides** (dATP, dCTP, dGTP, dTTP) and low concentrations of one of four **dideoxynucleotides** (ddATP, ddCTP, ddGTP ou ddTTP).

These ddNTP once incorporated in the newly synthesized DNA strand, block elongation. Synthesis termination is done by a statistical manner on each possible positions.



Sequence reading

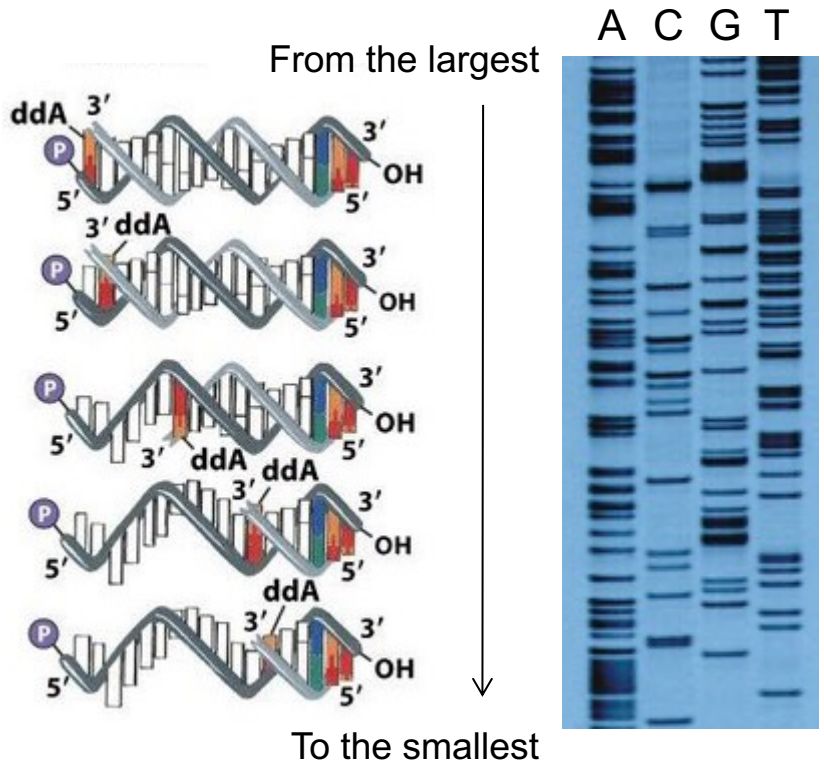
We a get a **mix of DNA fragments** terminating at each position of the sequence.

These fragments are then separated on a DNA **polyacrylamide gel electrophoresis**.

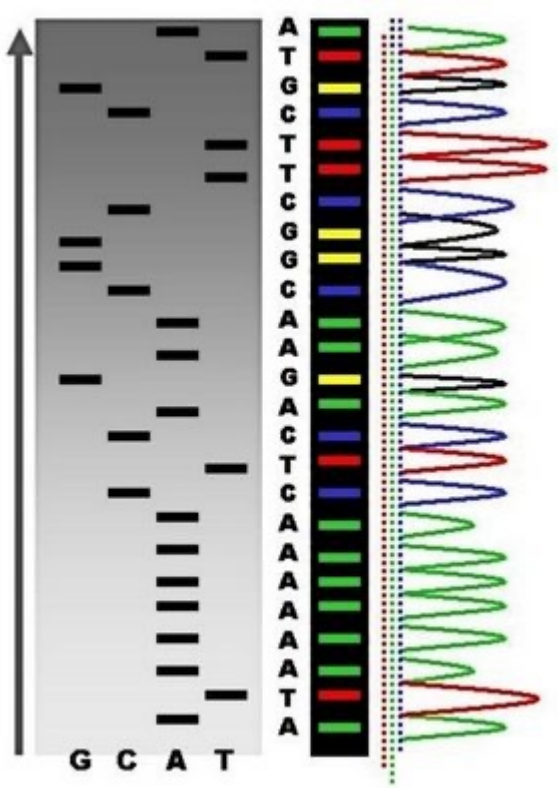
Detection of synthesize fragments is done by the **incorporation of labelling beacon** in the DNA.

At the origin this label was radioactive, attached either on the primer or on the dideoxynucleotide.

Around **1 kb of DNA** by run during **2 days**.
One read by **sample**.



Capillary sequencers

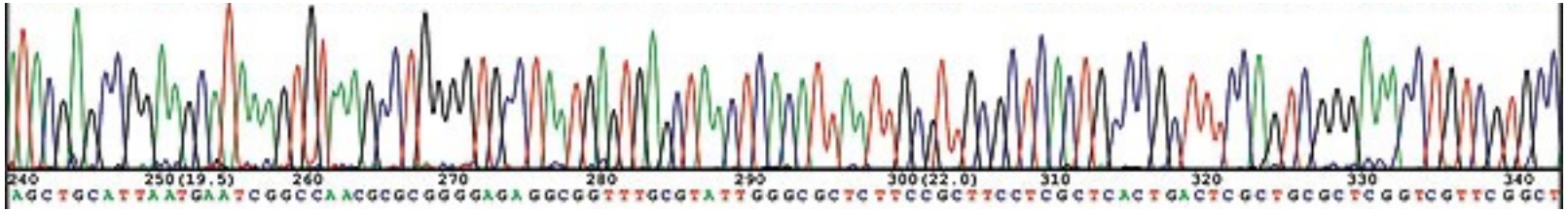


First version in the 90's thanks to the **modification of the radioactive label by a fluorescent one**.

Using **glass capillary of few micron diameters**, on 30 to 50 cm long.

The four nucleotides migrate in the same tube thanks to **four different fluorescent dyes**.

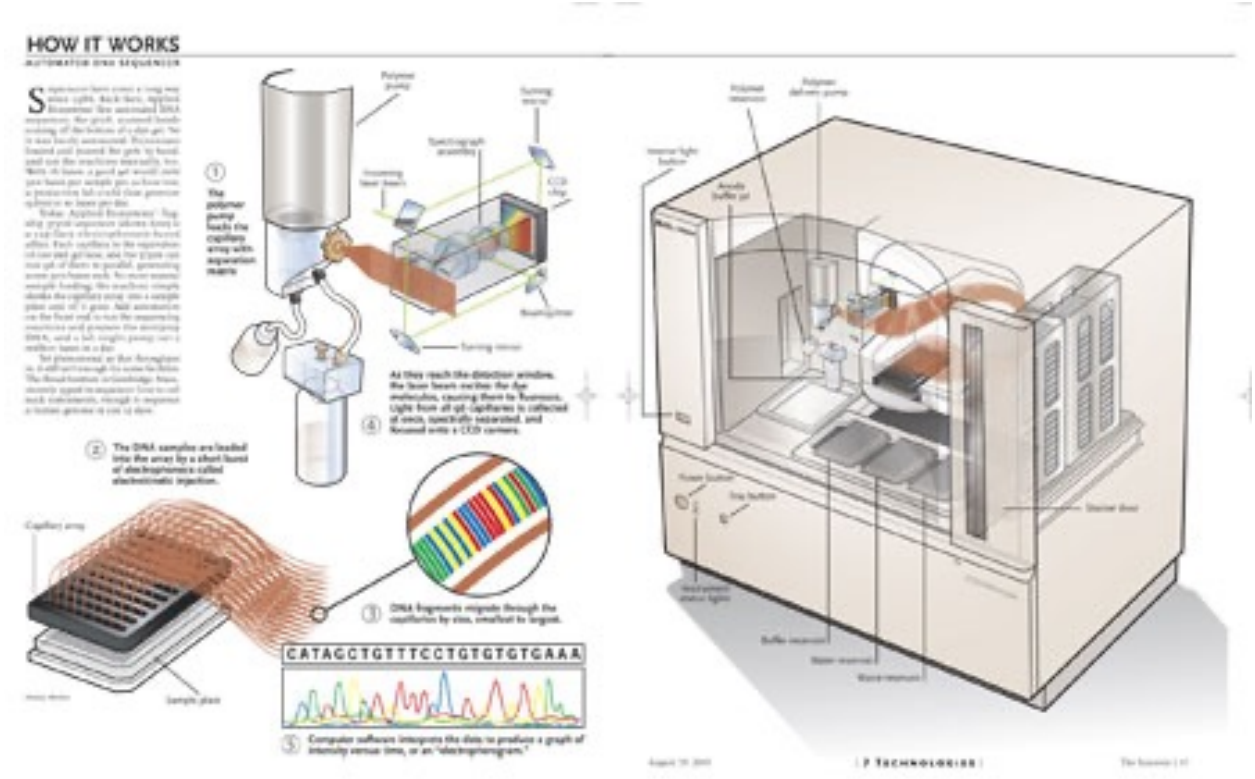
300 kb of DNA by run during **3 hours**. Several hundred sample at a time.



ABI 3730xl DNA Analyzer

96 parallel capillary (up to 50 cm) array.
768 samples, **690 kb** DNA, **3 hours** run.

At the Broad Institute (Cambridge, Massachusetts) **126 devices** were able to sequence **1 human genome** in **12 days**.



From *The Scientist*

Second generation sequencing: high throughput sequencing

The first technologies on the market

Goal: to obtain a huge number of short reads

since 1996

Applied Biosystems
ABI 3730XL
1 Mb / day

January '07 GS20
June '07 GS FLX

Roche / 454
Genome Sequencer FLX
100 Mb / run

January '07

Illumina / Solexa
Genome Analyzer
2,000 Mb (2 Gb) / run

November '07

Applied Biosystems
SOLiD
3,000 Mb (3 Gb) / run

Illumina Genome Analyzer
January 2007

From: Clive Brown <clive.Brown@solexa.com>
Date: Sun, 20 Feb 2005 16:34:46 +0100
To: Nick McCooke <Nick.McCooke@solexa.com>, Tony Smith <Tony
Swerdlow <Harold.Swerdlow@solexa.com>, John Milton <JM.Milto
<Kevin.Hall@solexa.com>, Colin Barnes <Colin.Barnes@solexa.c
<Vincent.Smith@solexa.com>, Klaus Maisinger <Klaus.Maisinger
Conversation: WE'VE DONE IT !!!!
Subject: WE'VE DONE IT !!!!

Tony Cox, Peta and I now agree - having looked at all of the PhiX174 data.

We have re-sequenced our first genome !!!!!!!

DNA library preparation

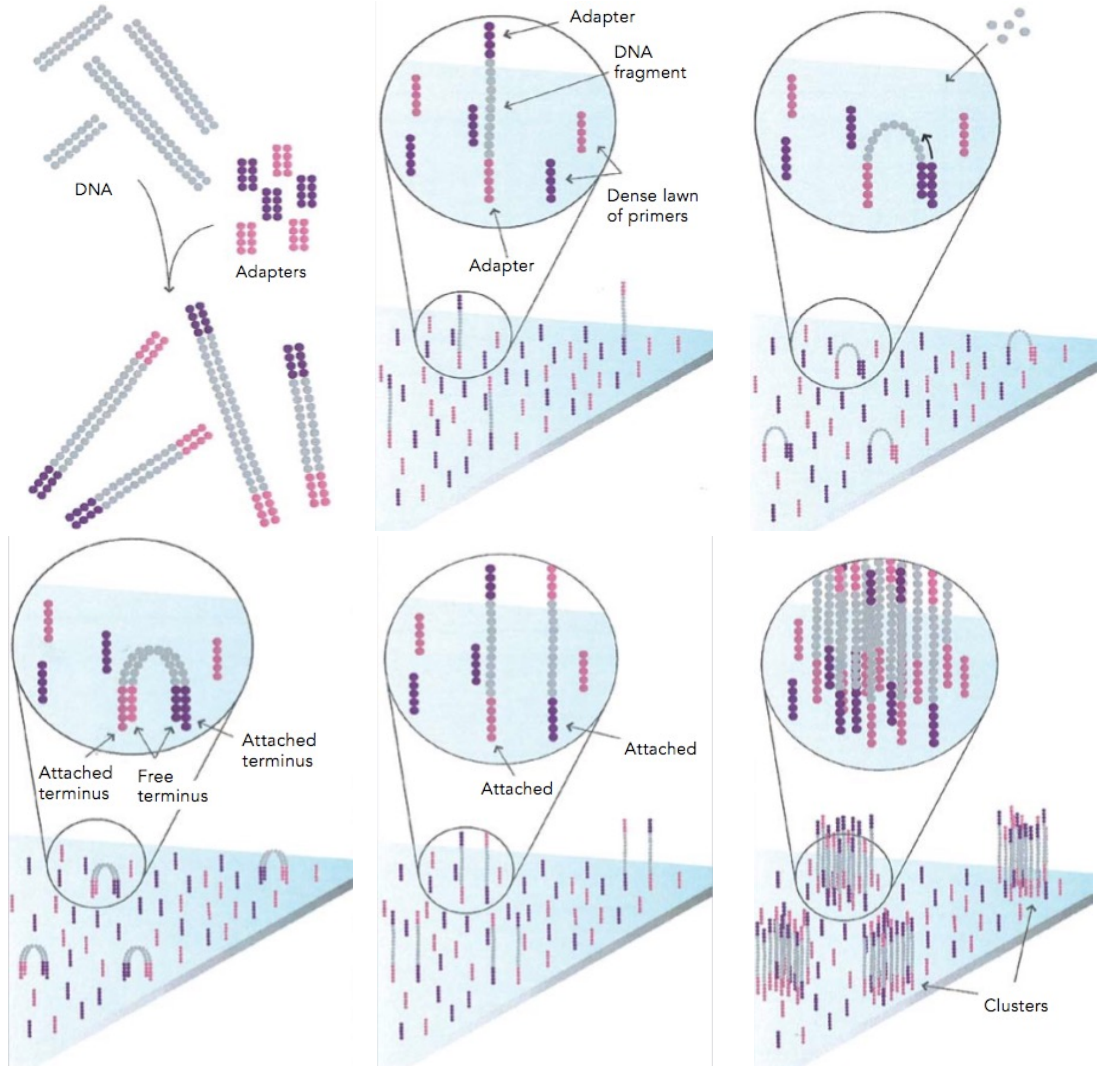
Random DNA fragmentation and size selection.

Ligation of adaptors.

DNA denaturation.

Hybridization of fragments onto the “**flowcell**” surface.

Solid phase bridge PCR.



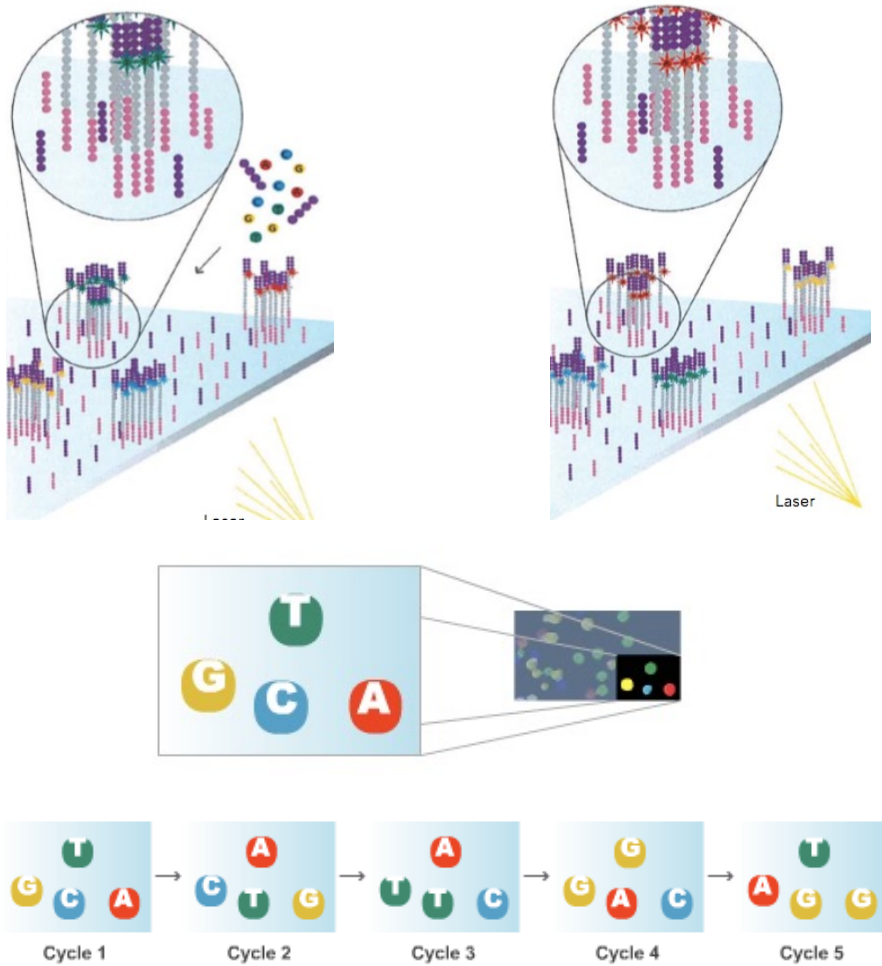
Reversible terminator sequencing

The four **reversible terminators** are added simultaneously.

Laser scanning of the flowcell surface.

Release of the blocking terminator.

Sequencing cycles are repeated one base at a time.

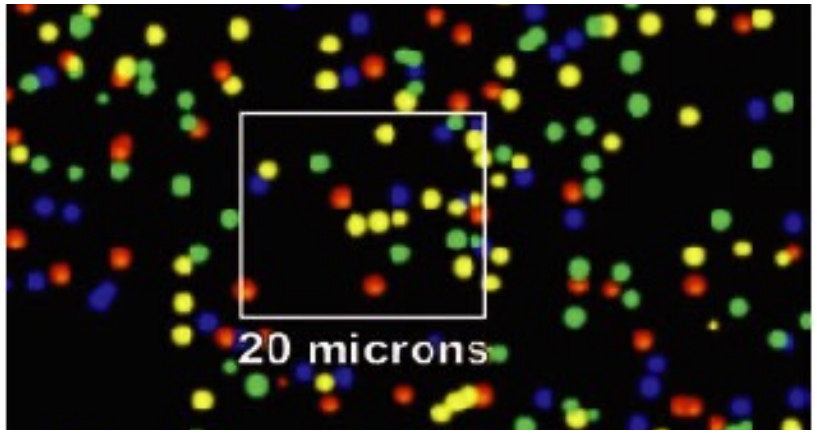


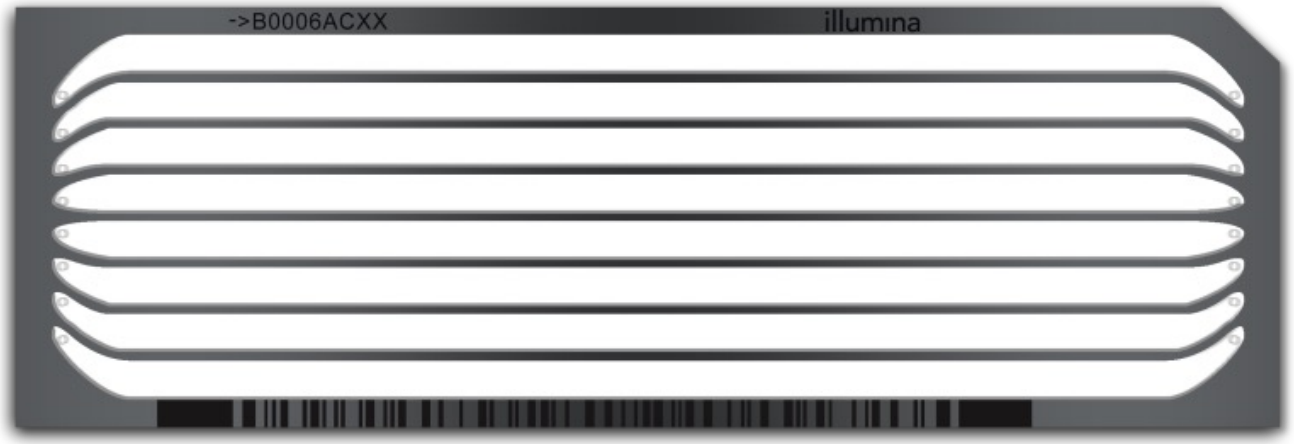
http://www.illumina.com/

Sequence analysis

Scanning at each position for all sequences (reads) in parallel.

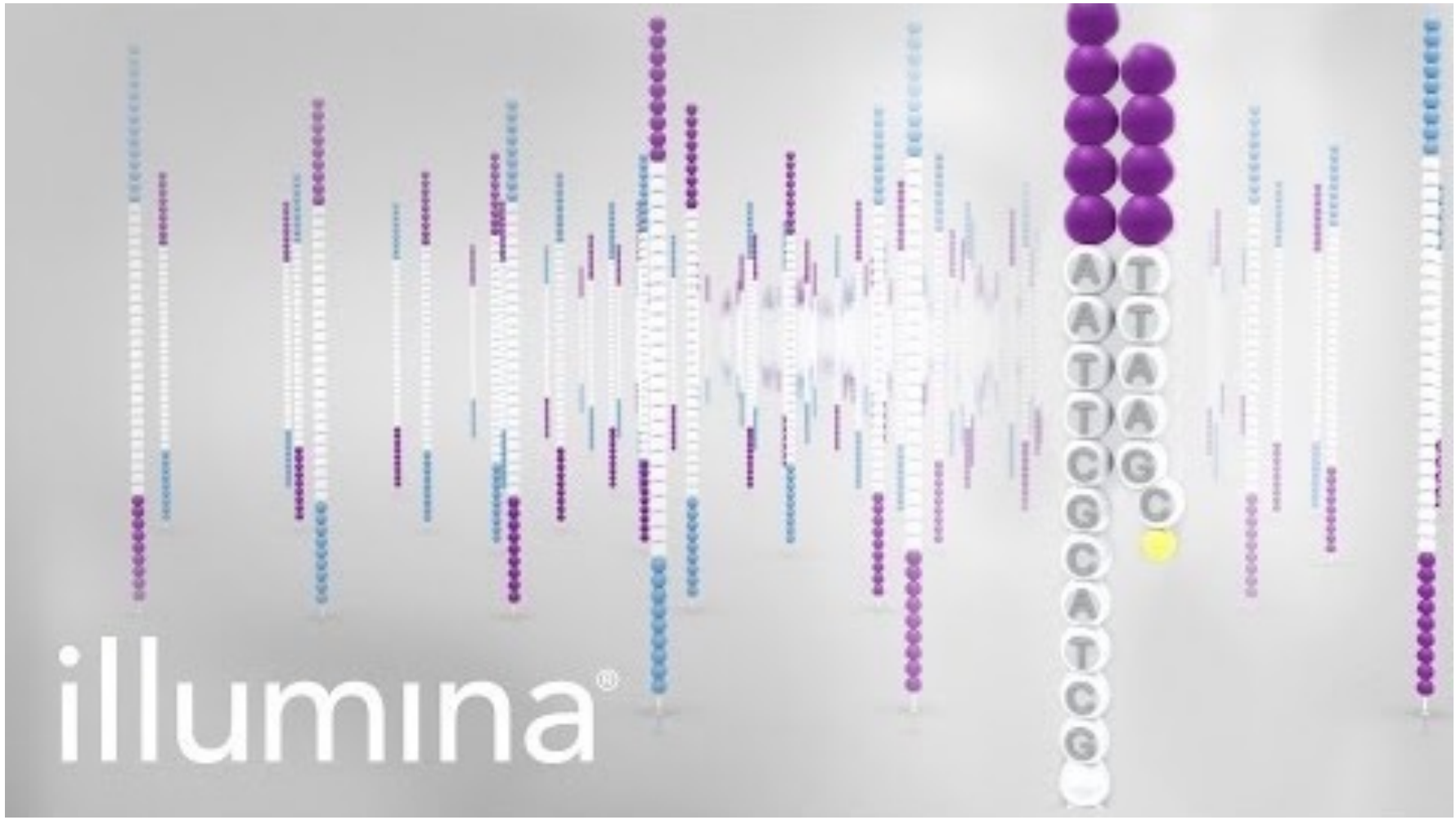
Most of the errors (99%) are sequencing errors (misincorporation).





http://www.illumina.com/

Illumina sequencing by synthesis



https://www.youtube.com/watch?v=fCd6B5HRaZ8

Specifications of the latest Illumina sequencers





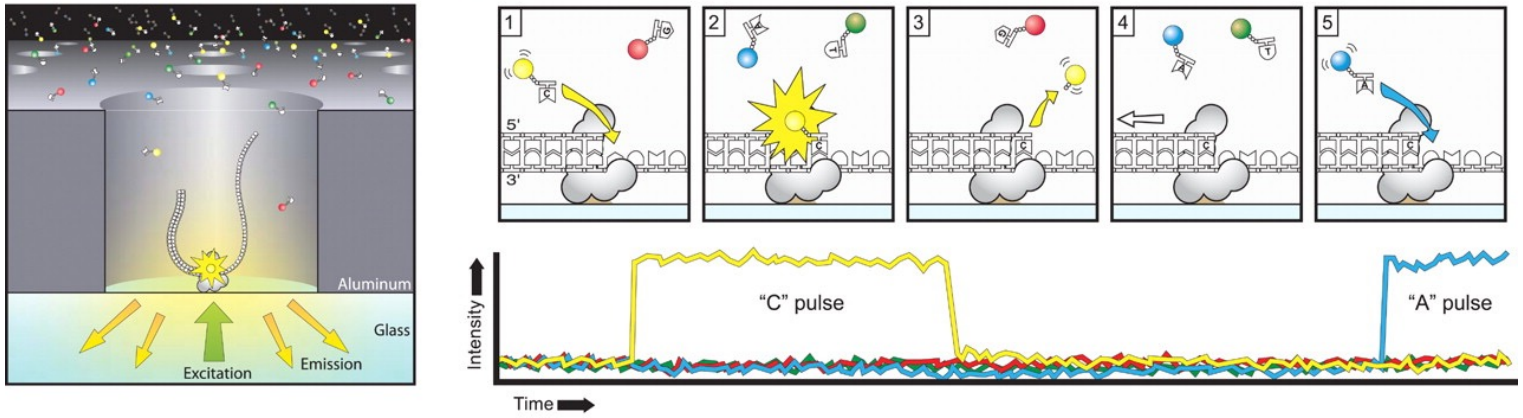
	MiniSeq	MiSeq	NextSeq 550	NextSeq 2000	NovaSeq
Run Time	24 hours	56 hours	29 hours	2 days	44 hours
Read length (bp)	2x 150	2x 300	2x 150	2x 150	2x 150
Read number	50 10^6	50 10^6	800 10^6	1 10^9	10 10^9
Ouput	7.5 Gb	15 Gb	120 Gb	300 Gb	3,000 Gb
Throughput	7 Gb/day	7 Gb/day	100 Gb/day	150 Gb/day	1,500 Gb/day

The third generation

Real time sequencing

Real time sequencing on **single molecule** thanks to **RNA polymerase immobilisation** in wells.

Each **base incorporation is measure in real time** with a CCD camera under the bottom of the plate.



Eid (2009) *Science*

Sequel specifications

8,000,000 reads / SMRT cell;

From 1 to 52 kb (**avg 20 kb**);

450 Gb by SMRT cell;

Run duration = **30 hours**.

Nanopore technology

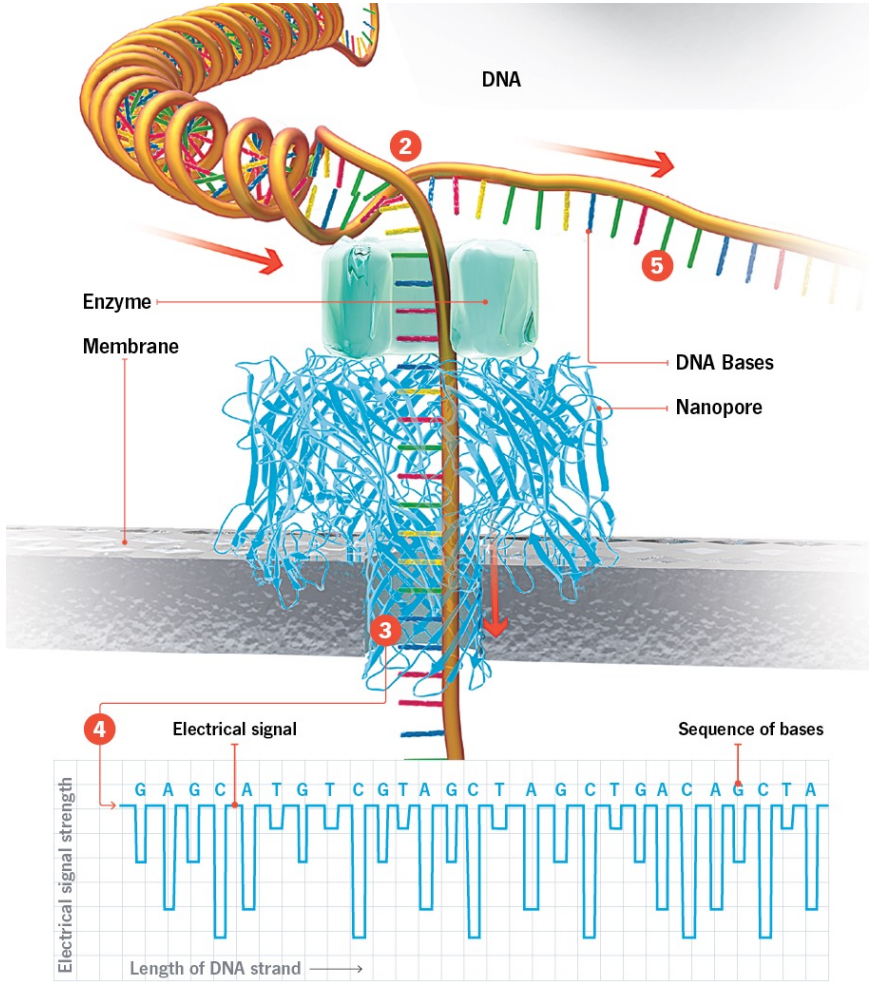
Single molecule detection system by passing single strand nucleic acids through a **nanometric pore**.

Base to base analysis in real time using **electric properties** of the nanopore.

DNA size sequencing of **kilobases.**

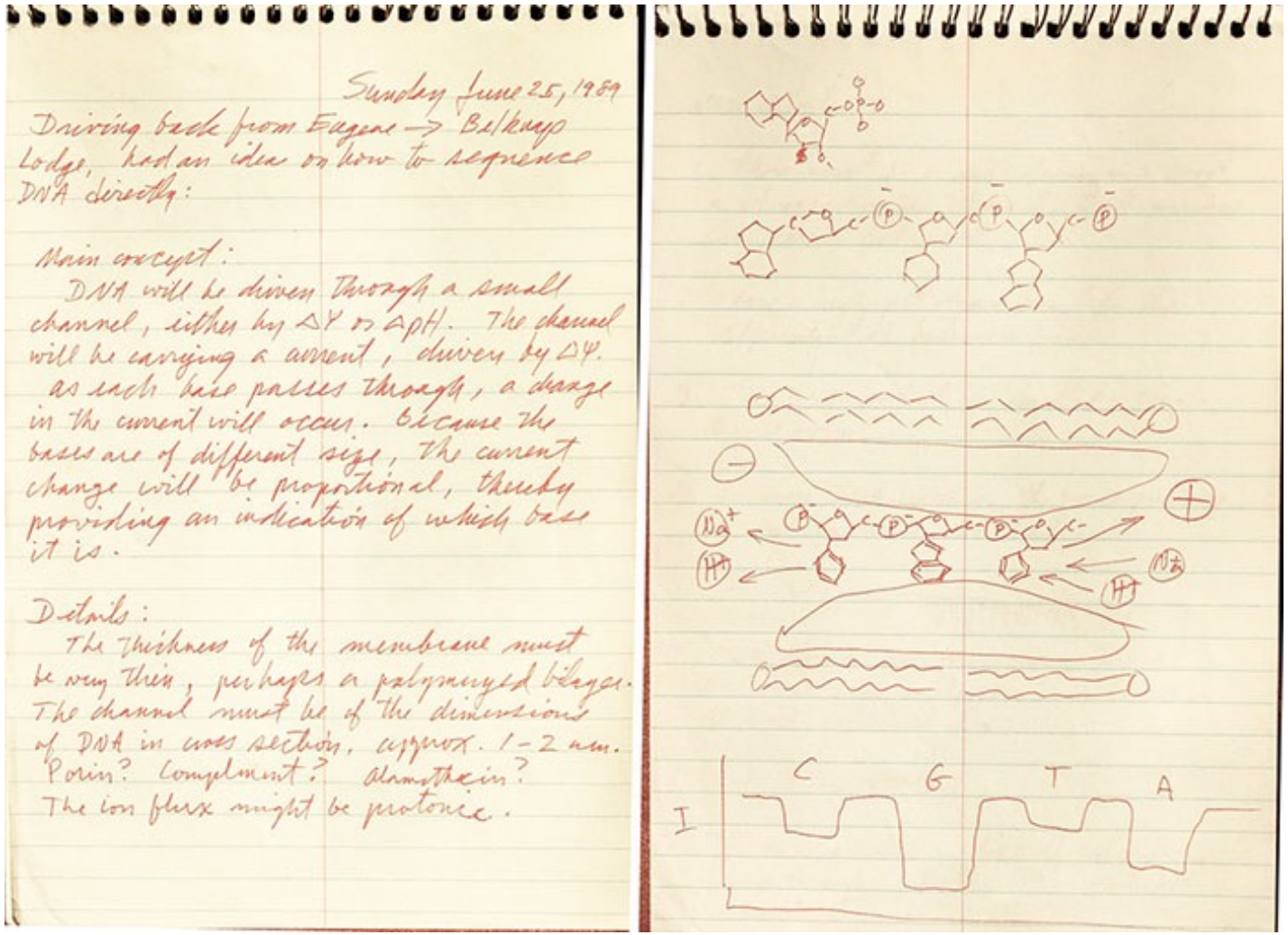
No limitation on the acid nucleic type to be detected (**DNA or RNA** even amino acids) including modifications (epigenetics).

No amplification.



Greenwood (2013) *Popular Science*

First described by David Deamer in 1989



Sunday June 25, 1989
Driving back from Eugene → Belknap Lodge, had an idea on how to sequence DNA directly:

Main concept:
DNA will be driven through a small channel, either by ΔΨ or ΔpH. The channel will be carrying a current, driven by ΔΨ. As each base passes through, a change in the current will occur. Because the bases are of different size, the current change will be proportional, thereby providing an indication of which base it is.

Details:
The thickness of the membrane must be very thin, perhaps a polymerized bilayer. The channel must be of the dimensions of DNA in cross section, approx. 1-2 nm. Porin? Complement? Alamethicin?
The ion flux might be protonic.

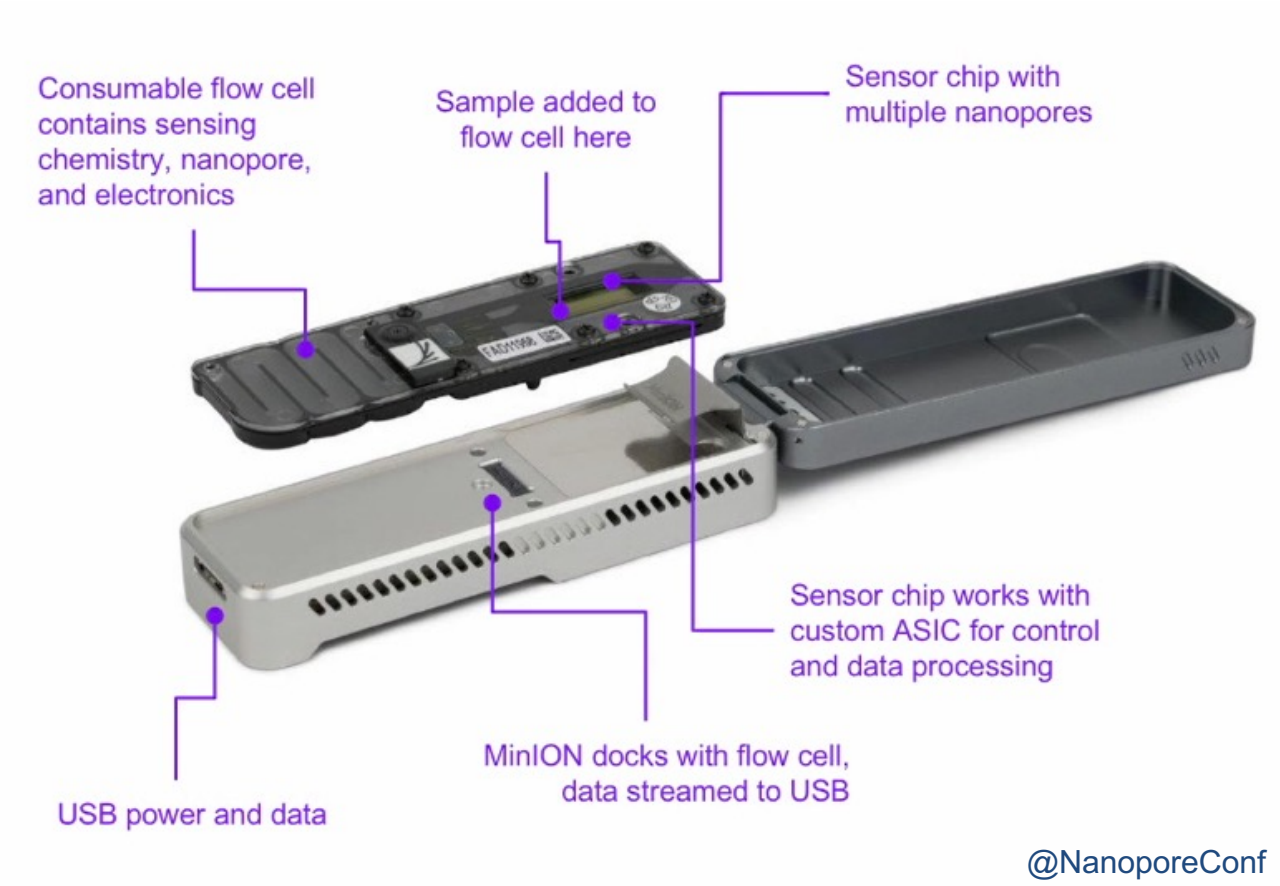
Oxford nanopore technologies

https://www.youtube.com/watch?v=RcP85JHLmnI

The MiniON flow cell

1 flow cell = 1 membrane
with 512 nanopores.

Single molecule sequencing
up to 4 Mb during up to 72 hours.



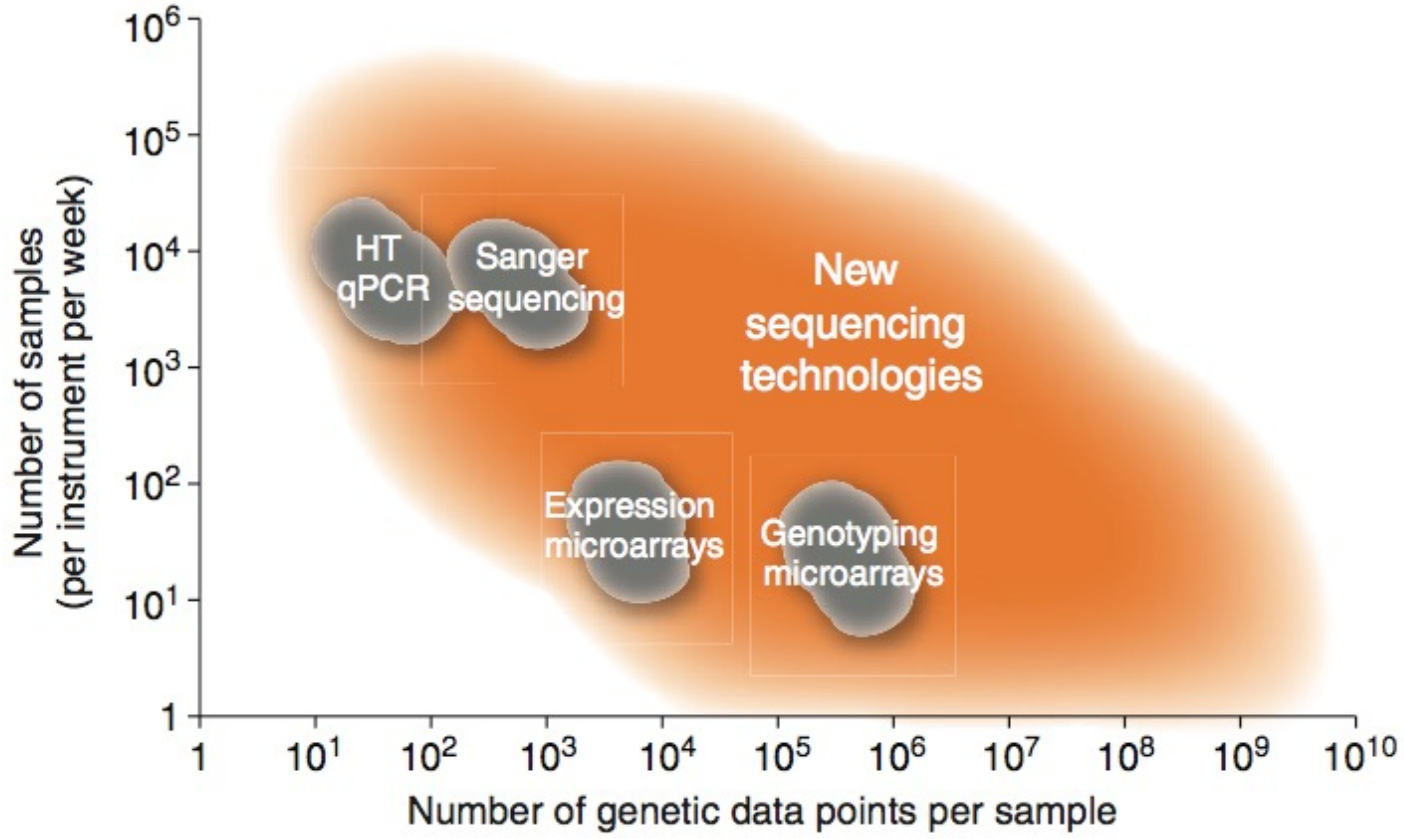
@NanoporeConf

Field genomics: portable sequencer tracks infectious disease outbreaks

Nick Loman using a MinION to sequence the Zika virus in Brazil

Applications

They cover a lot of previous existing techniques



Kahvejian et al. (2008) *Nat. Biotech.*

De novo sequencing

Quicker and cheaper sequencing than Sanger.

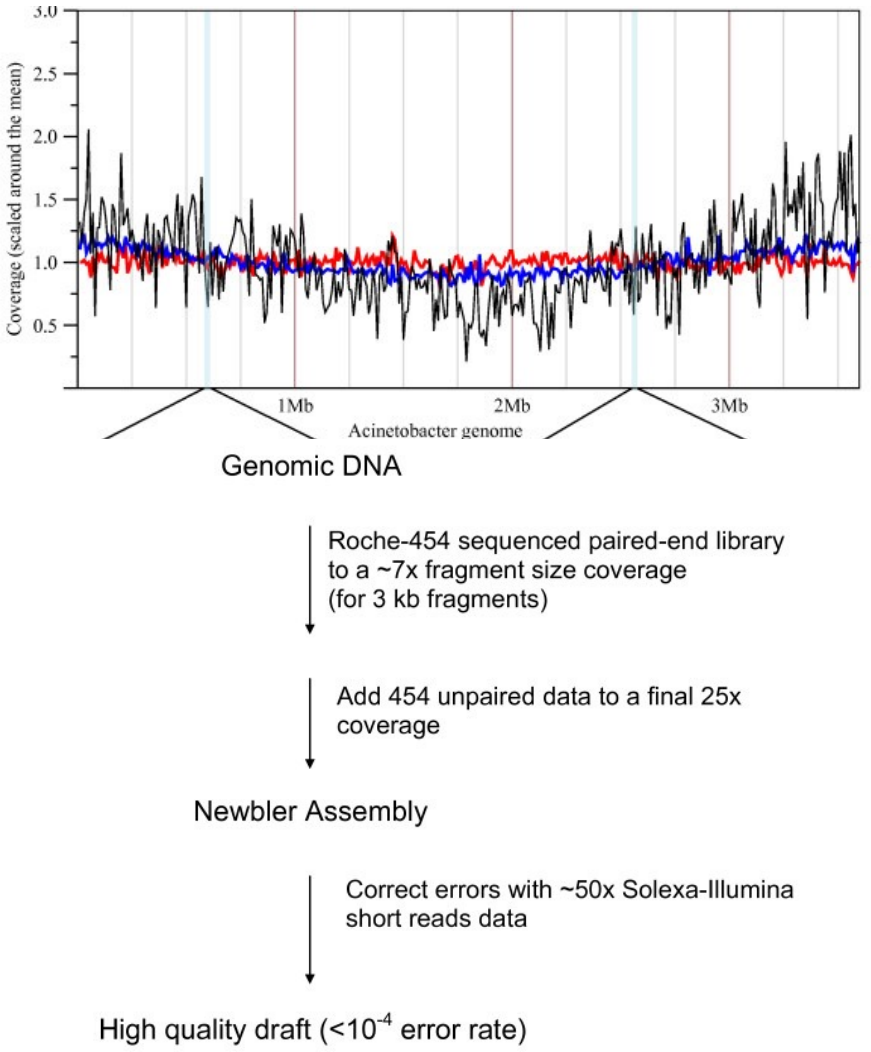
But **reads are smaller**.

Combination of different methods allow to obtain **better quality sequencing drafts**.

=> Combining 454 and Illumina.

Low error rate and **homogenous coverage** due to no cloning biases compared to the Sanger method.

Errors are not same with the two high throughput sequencing methods.



Aury et al. (2008) *BMC Genomics*

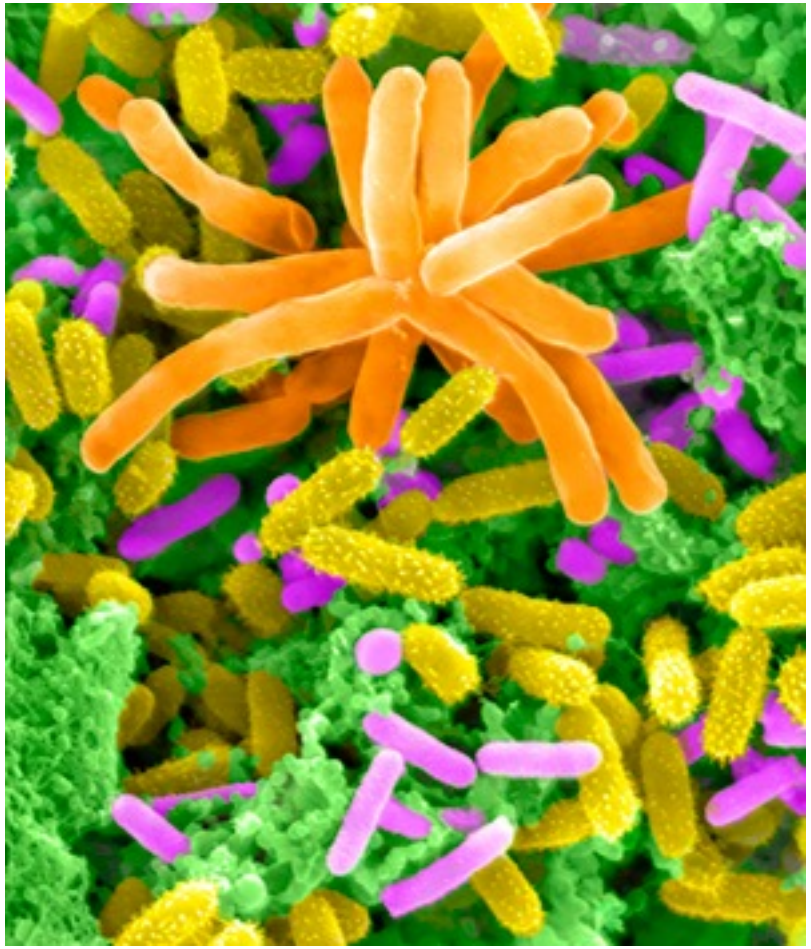
Resequencing applications

Goal: to analyze various genomes compared to a reference one.

Search for polymorphisms and structural variants in populations, mutation identification in biotechnology, organism evolution analyses, cell differentiation along time, ancient DNA discovery…

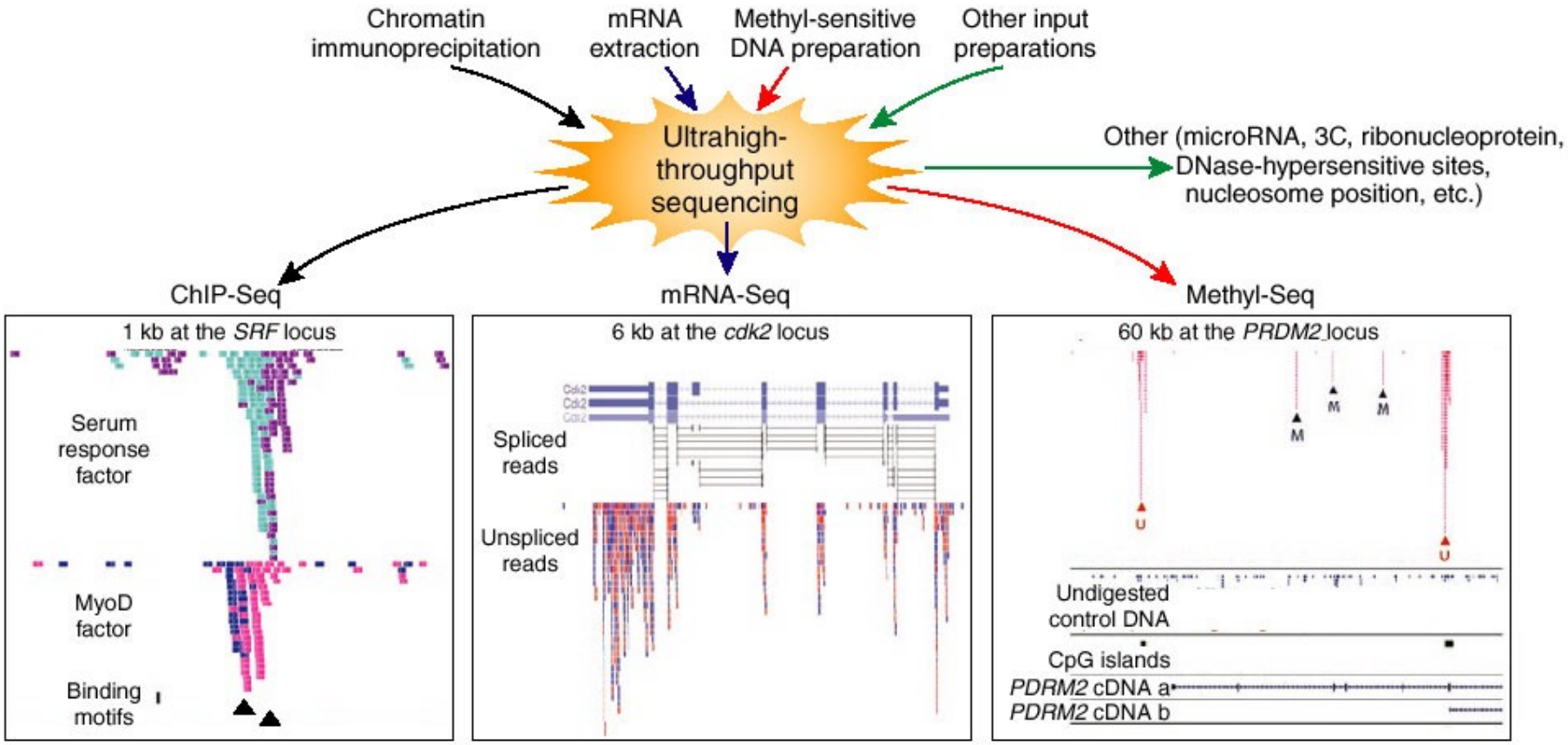
Metagenomics: genome characterisation in samples.

A wide range of applications: characterise pathogen micro-organisms in patient tissues, definition of the species found in environment samples, understand species evolution…



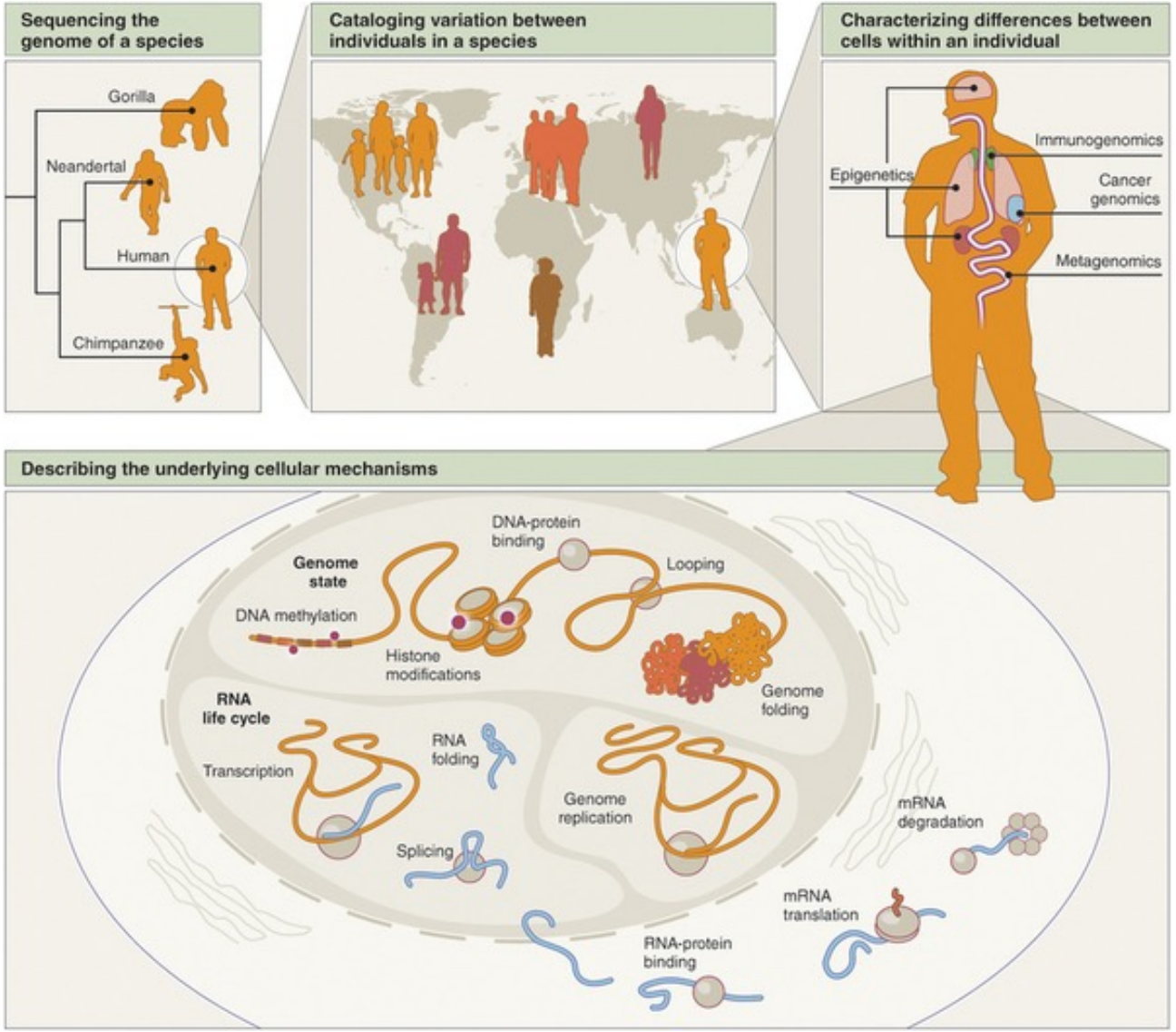
From JGI DOE

Functional genomic applications



Wold et al. (2008) *Nat. Methods*

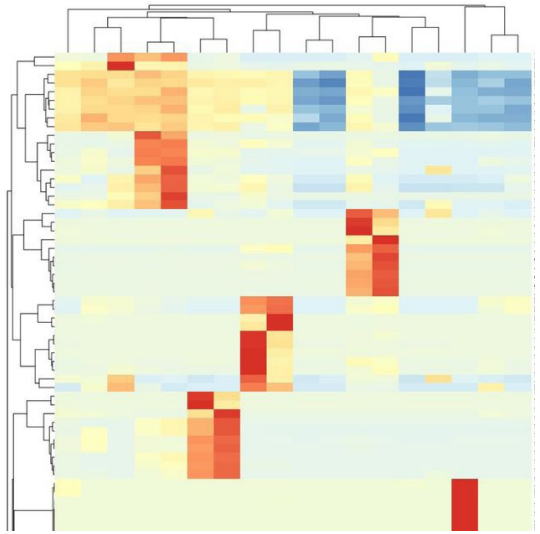
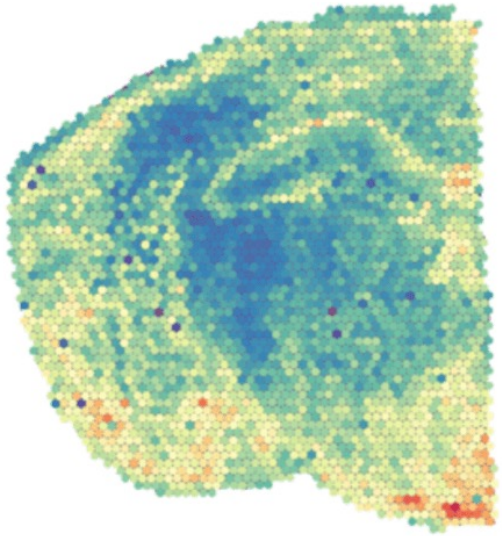
Perspectives



Shendure & Aiden (2012) *Nat. Biotech.*

From bulk to localised single-cell

2012 2019

Bulk	Single cell	Spatial
High throughput sequencing Method of the year 2009	Single cell genomics Method of the year 2013	Spatially-resolved transcriptomics Method of the year 2020
		
Martins et al. (2020)	Han Chen D. (2019)	Jorgensens M. (2021)


SORBONNE
UNIVERSITÉ