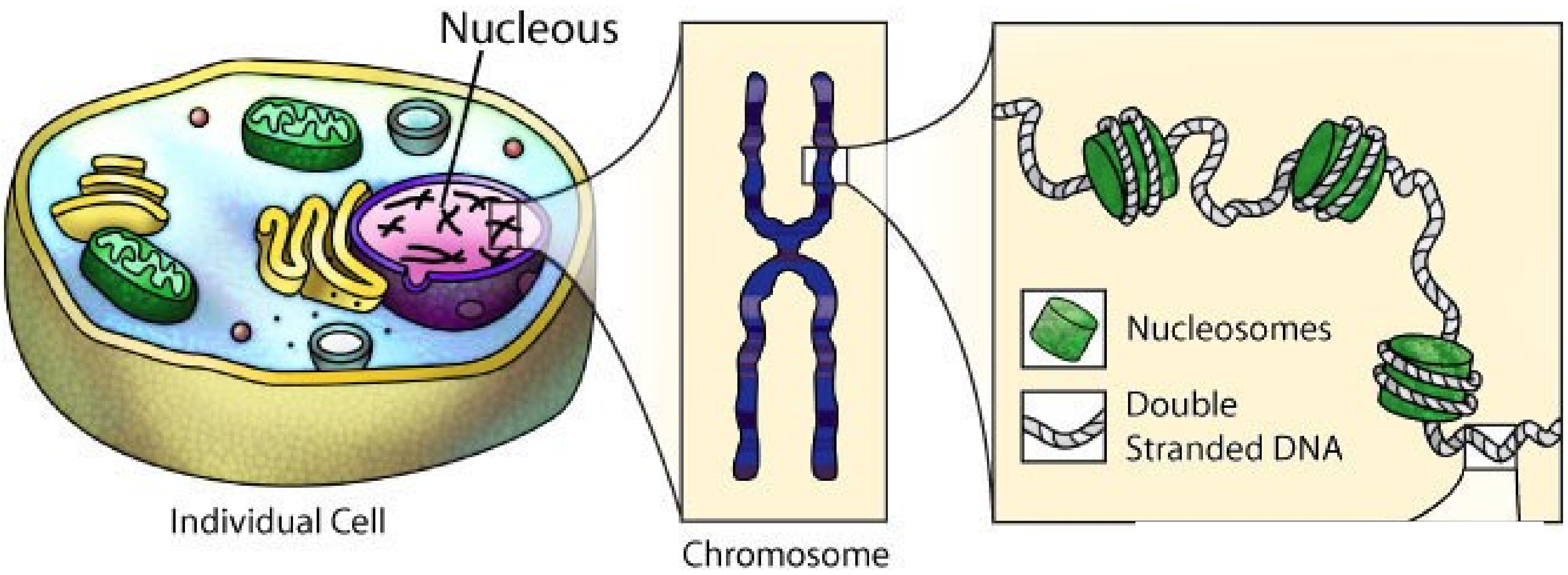
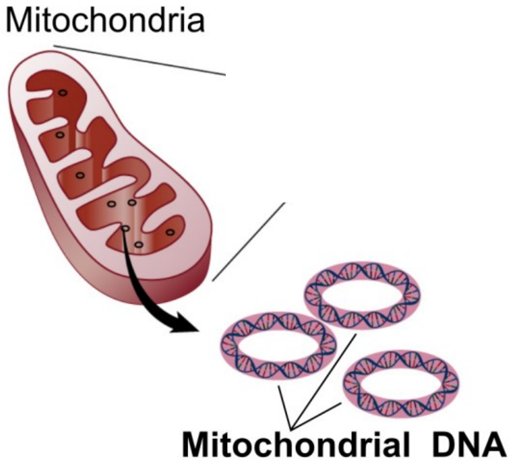
Next-generation sequencing

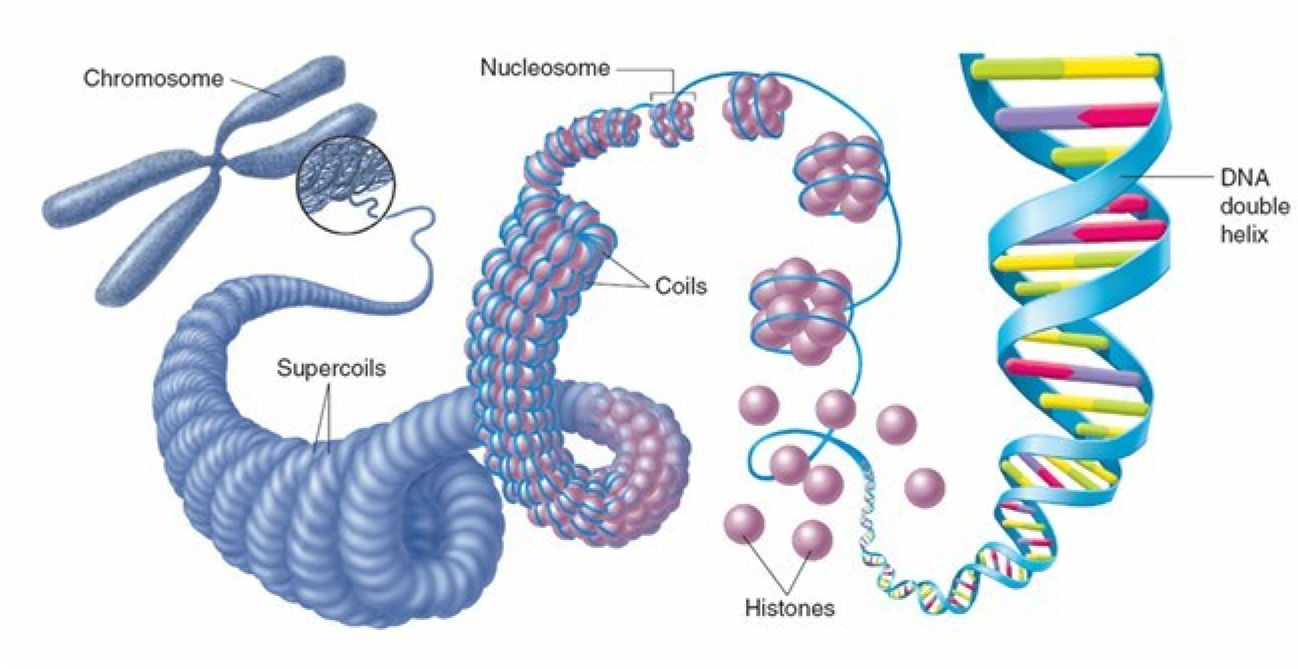
Radka Reifová





Genome = the entire set of DNA molecules in a cell

- Nucleus - chromosomes
- Mitochodria, plastids
- Plasmids in bacteria



Human genome

- 23 pairs of chromosomes
- 3 billion bp (haploid genome)
  = 3,000,000 bp
  = 3 Giga bp

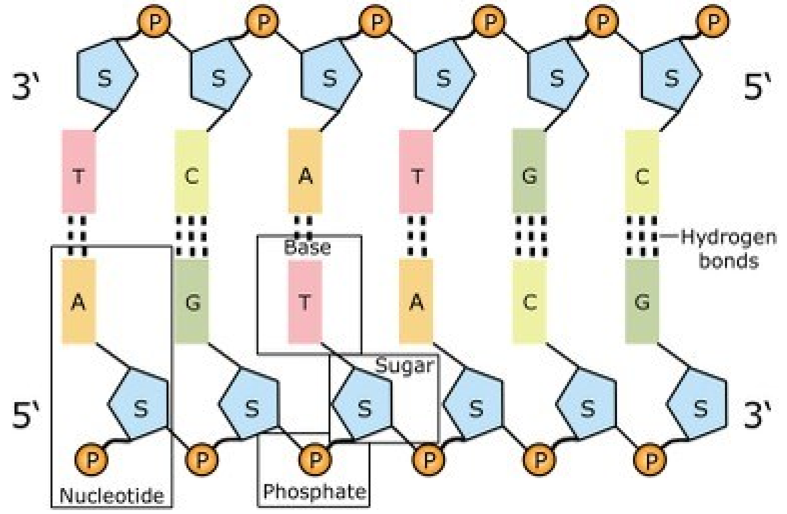
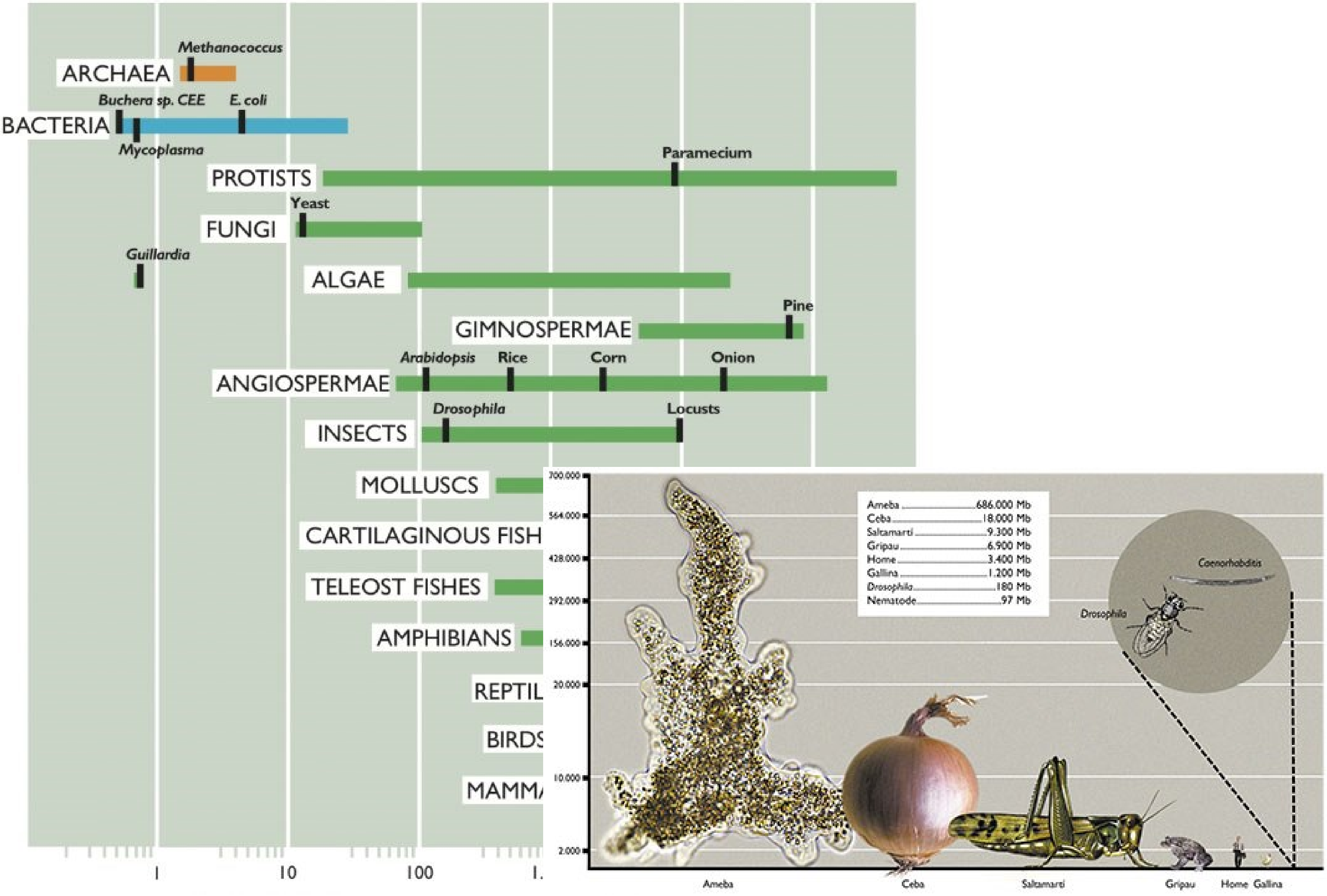


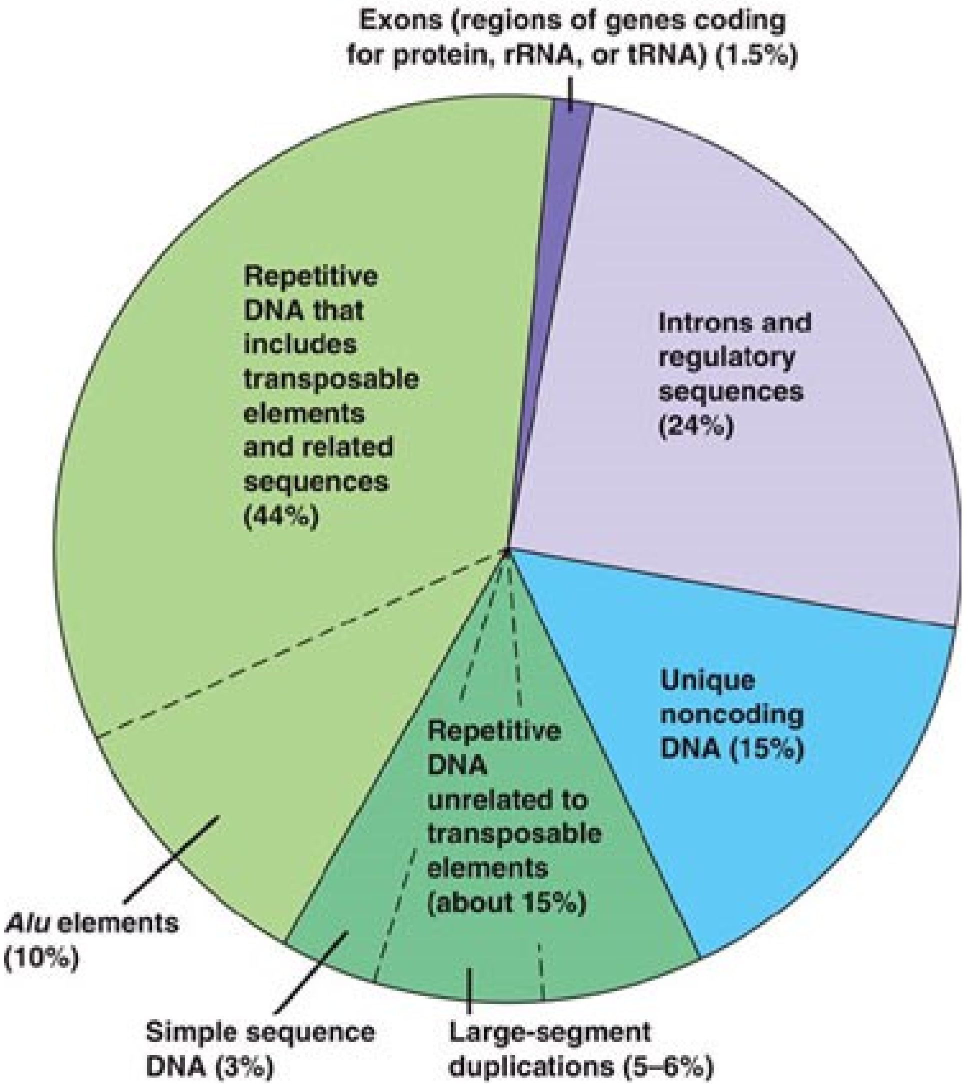
Image adapted from: National Human Genome Research Institute.

Genome size



Ameba	686.000 Mb
Ceba	18.000 Mb
Saltamarti	9.300 Mb
Gripau	6.900 Mb
Home	3.400 Mb
Gallina	1.200 Mb
Drosophila	180 Mb
Nematode	97 Mb

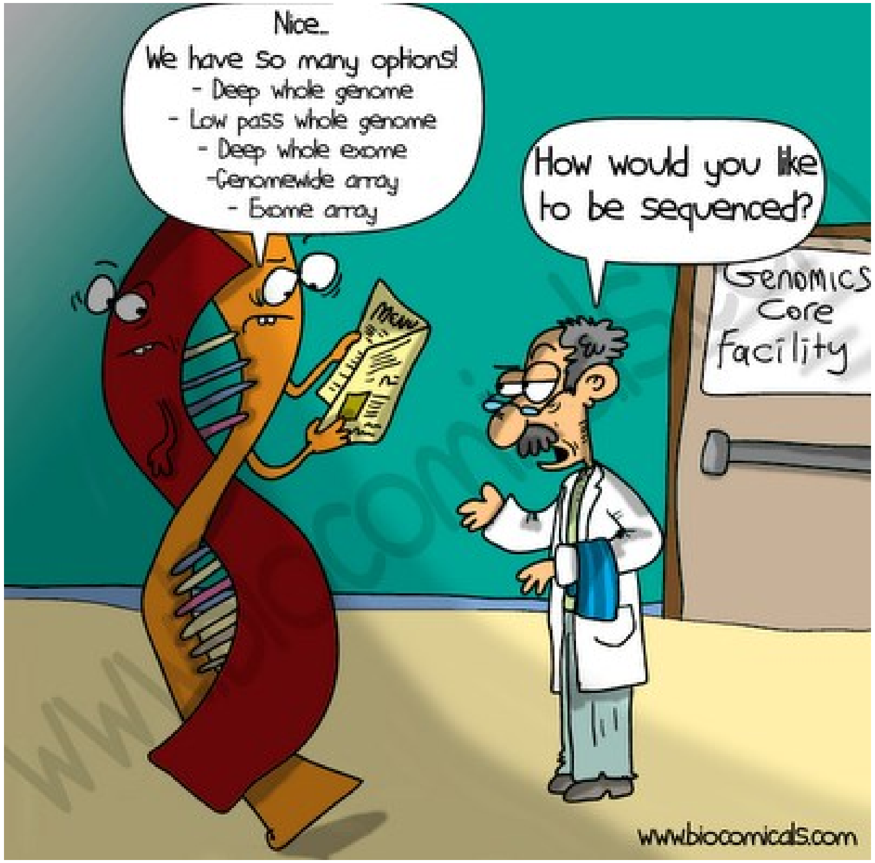
Composition of human genome



What can we sequence?

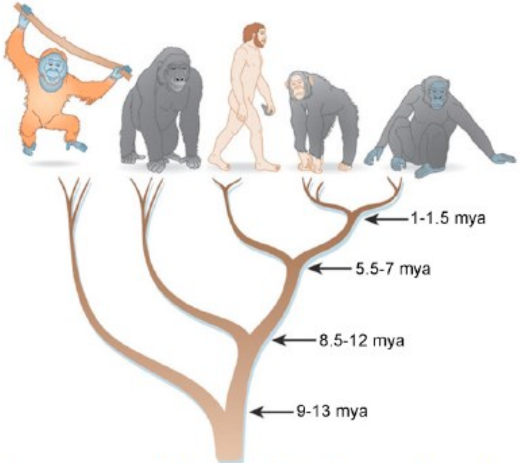
- Individual genes
- Repetitive sequences (microsatelites)
- Transcriptomes (RNA)
- Individual chromosomes or set of genes
- Whole genome
- mt DNA

………

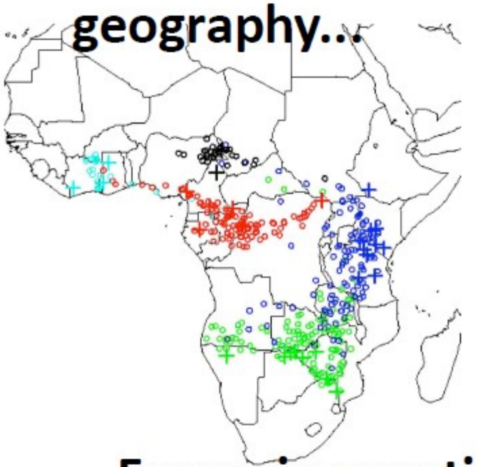


What can DNA sequence tell us?

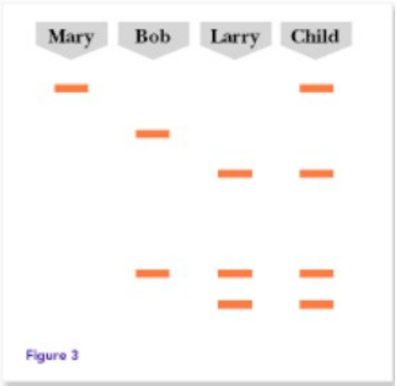
Evolution, speciation...



Populations & geography...



Paternity



Barcoding (fisheries)

Forensic genetics (criminalistics)

- DNA is kind of chemical fingerprint...

© Damian Baker/Daily Mail Australia

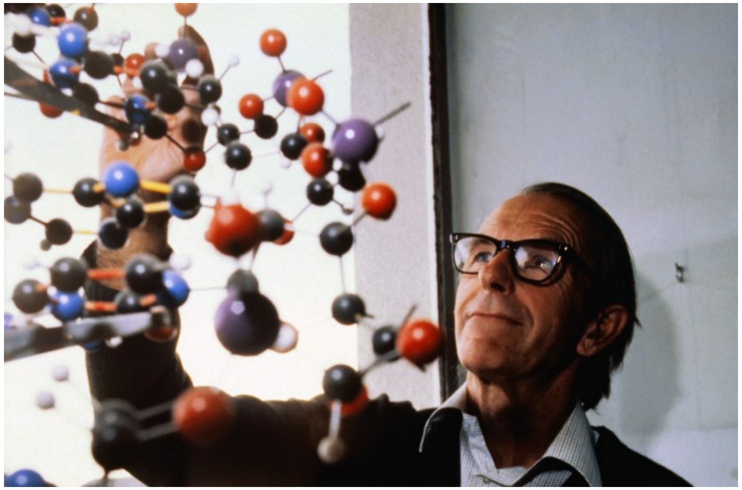
... and many many other fields!

How to sequence DNA?

DNA sequencing

Sanger sequencing – based on extension of primers and dideoxynucleotide chain termination

Maxam-Gilbert sequencing – based on chemical modification of DNA followed by cleavage at specific bases



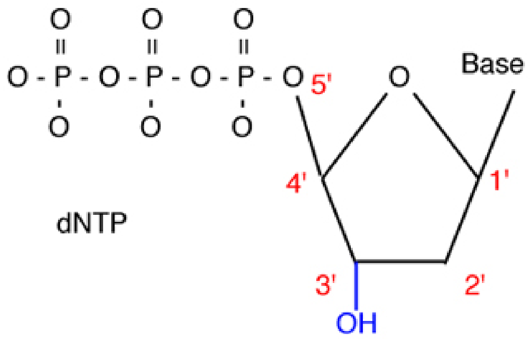
Frederic Sanger

Nobel prize

1958 – inzulin structure
1980 – DNA sequencing

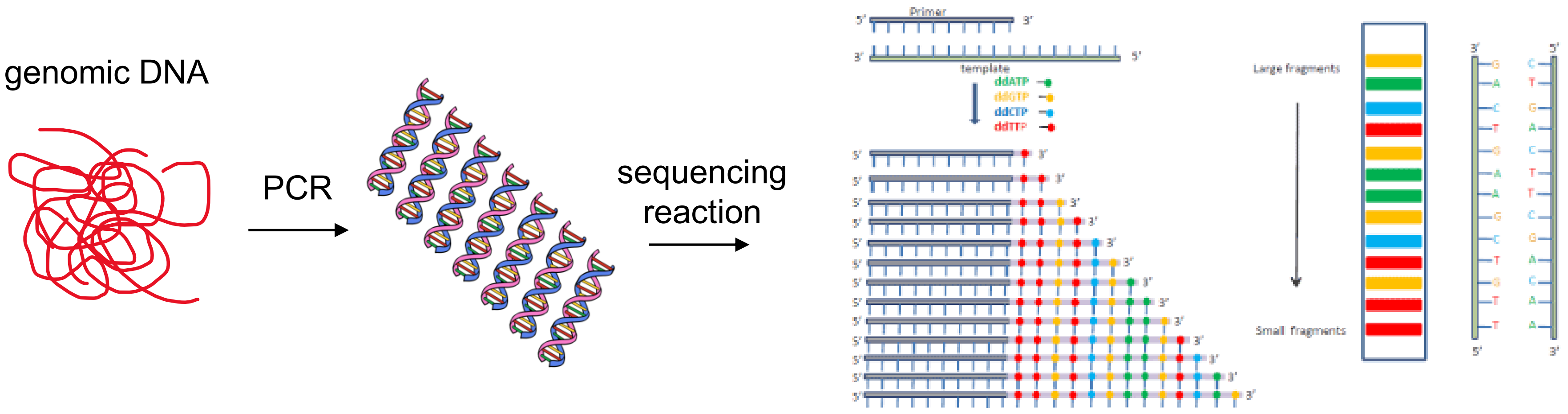
Sanger sequencing

- Sequencing of individual genes
  ONE DNA FRAGMENT IN EACH SEQUENCING REACTION



dNTP

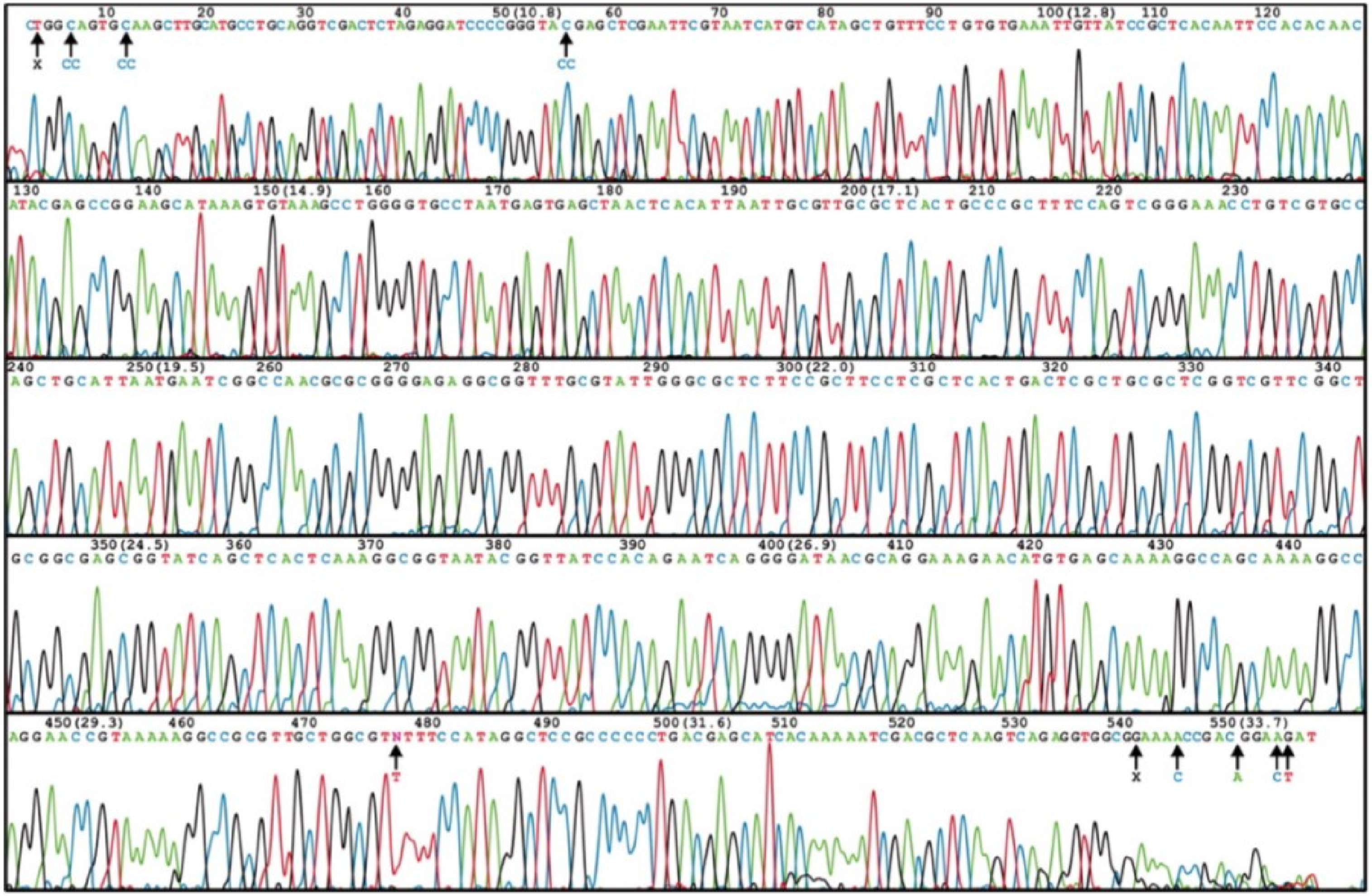
- PCR amplification of the DNA fragments
- Sequencing reaction using dideoxynucleotidtriphospates ddNTP
- Detection of fragments by electrophoresis on the gel


genomic DNA

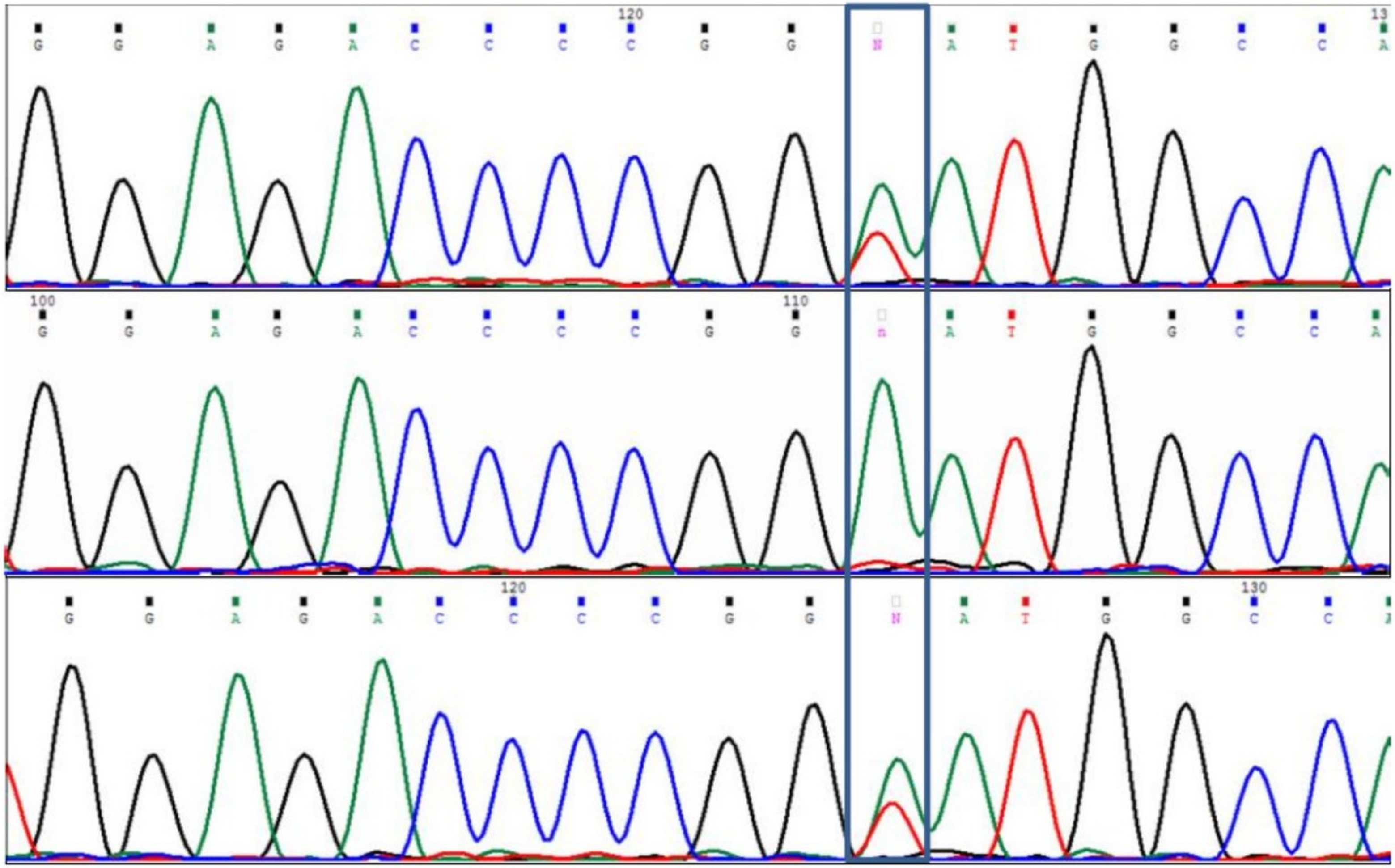
PCR

sequencing reaction

Chromatogram



- cca 500-800 bp
- Lower quality at the beginnig and end of the sequence
- Sequencing from from forward or reverse PCR primers, or other primers



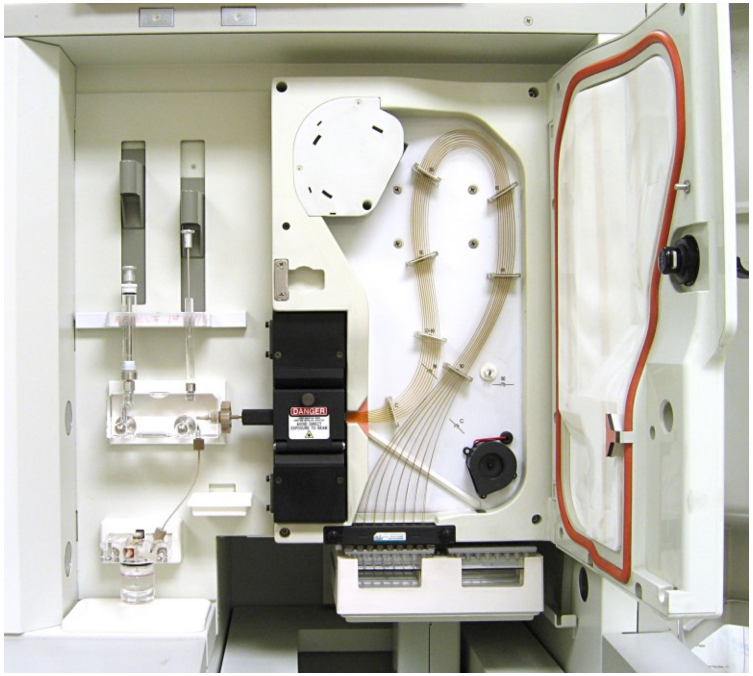
- High quality of data
- Detection of heterozygotes in diploid organisms

Sanger sequencers

Laboratory of DNA sequencing at Faculty of Science

- 4 capillary 3130 Genetic Analyzer (2007)
- 16 capillary 3130xl Genetic Analyzer (2010)
- 24 capillary 3500 Genetic Analyzer (2015)

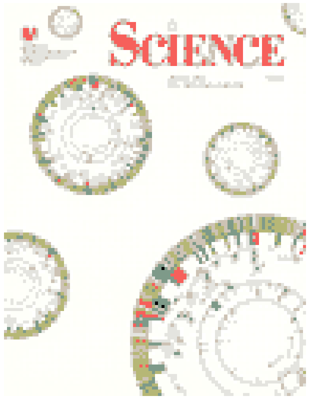
Large sequencing facilities

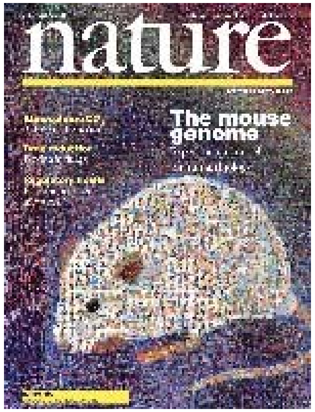


- 96 capillary DNA sequencing machines

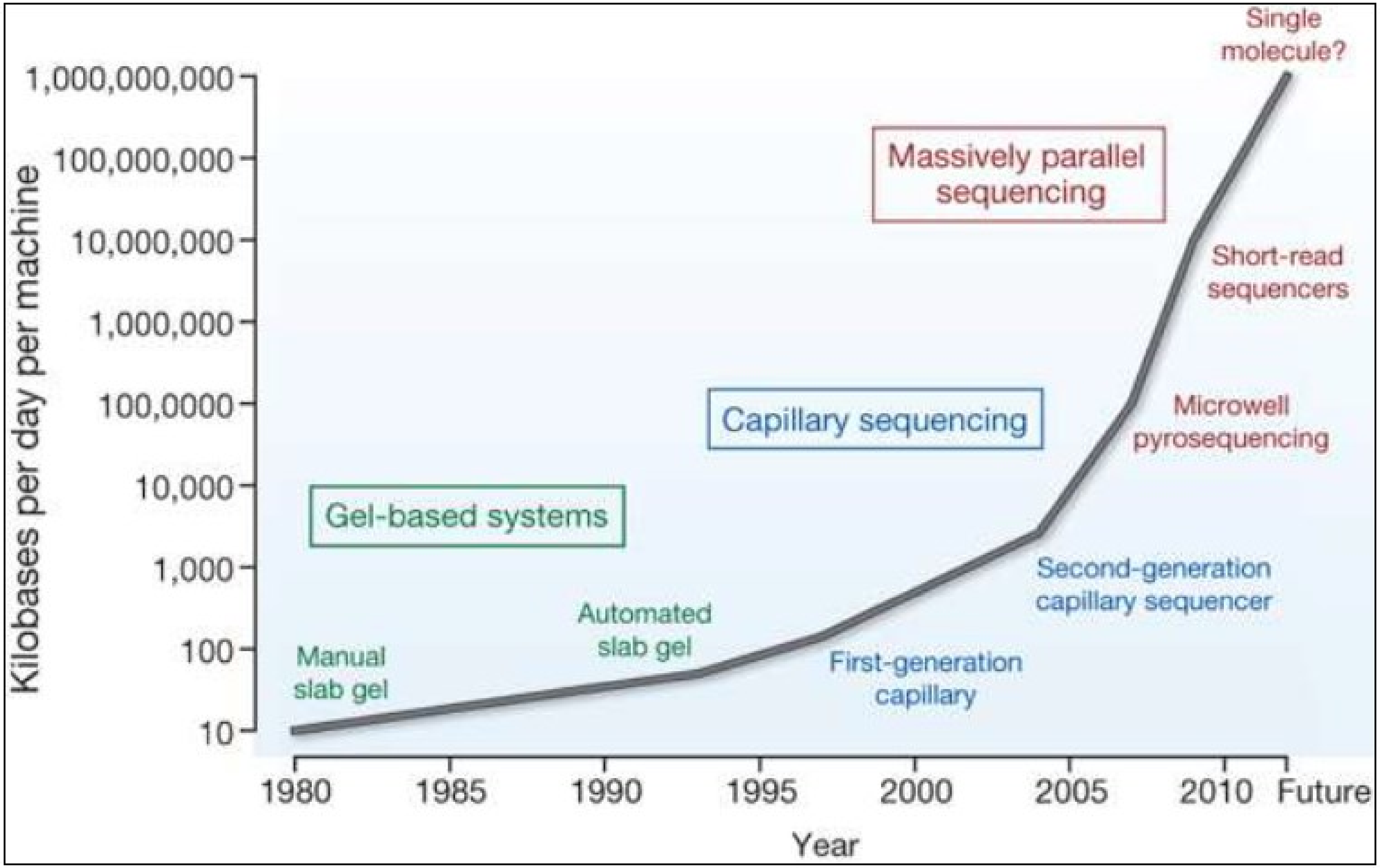
First genomes were sequenced using Sanger sequencing

- 1995 *Haemophilus influenze*
- 1996 *Saccharomyces cerevisiae*
- 1998 *Caenorhabditis elegans*
- 2000 *Drosophila melanogaster*
- 2001 *Homo sapiens*
- 2002 *Mus musculus*





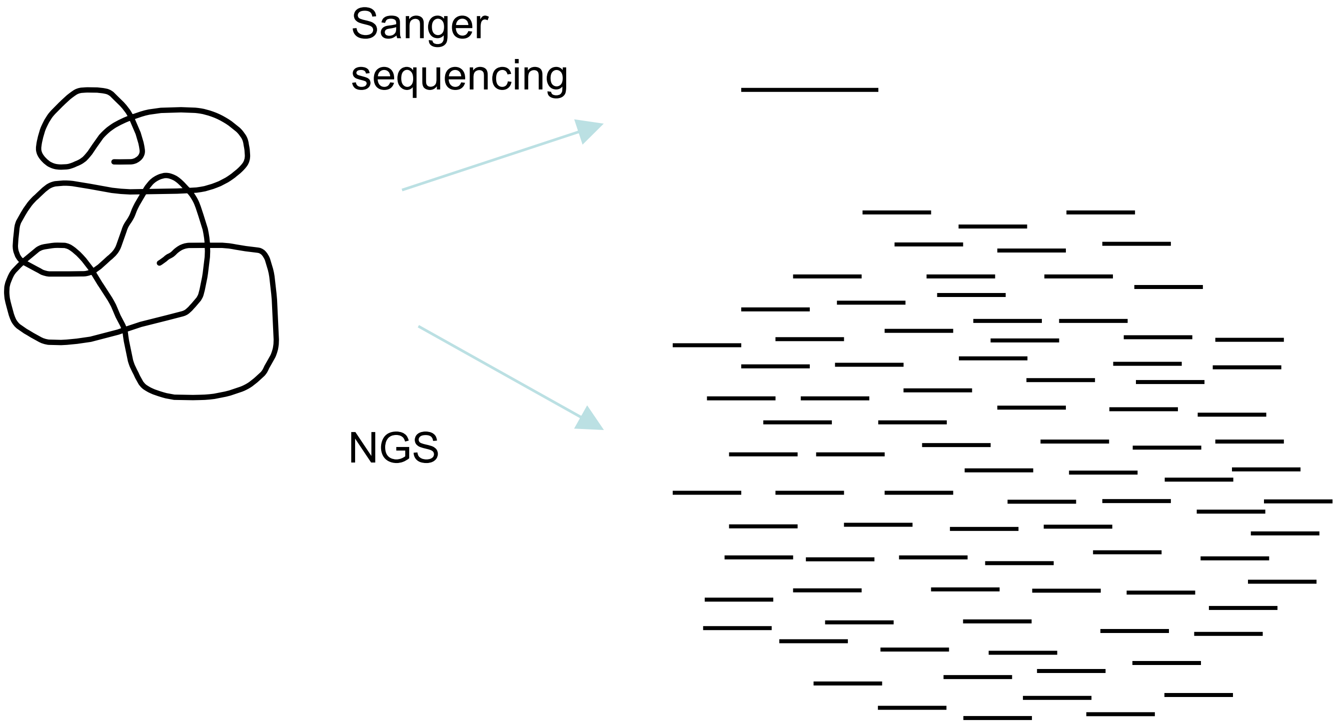
Next-generation sequencing (massively parallel sequencing)

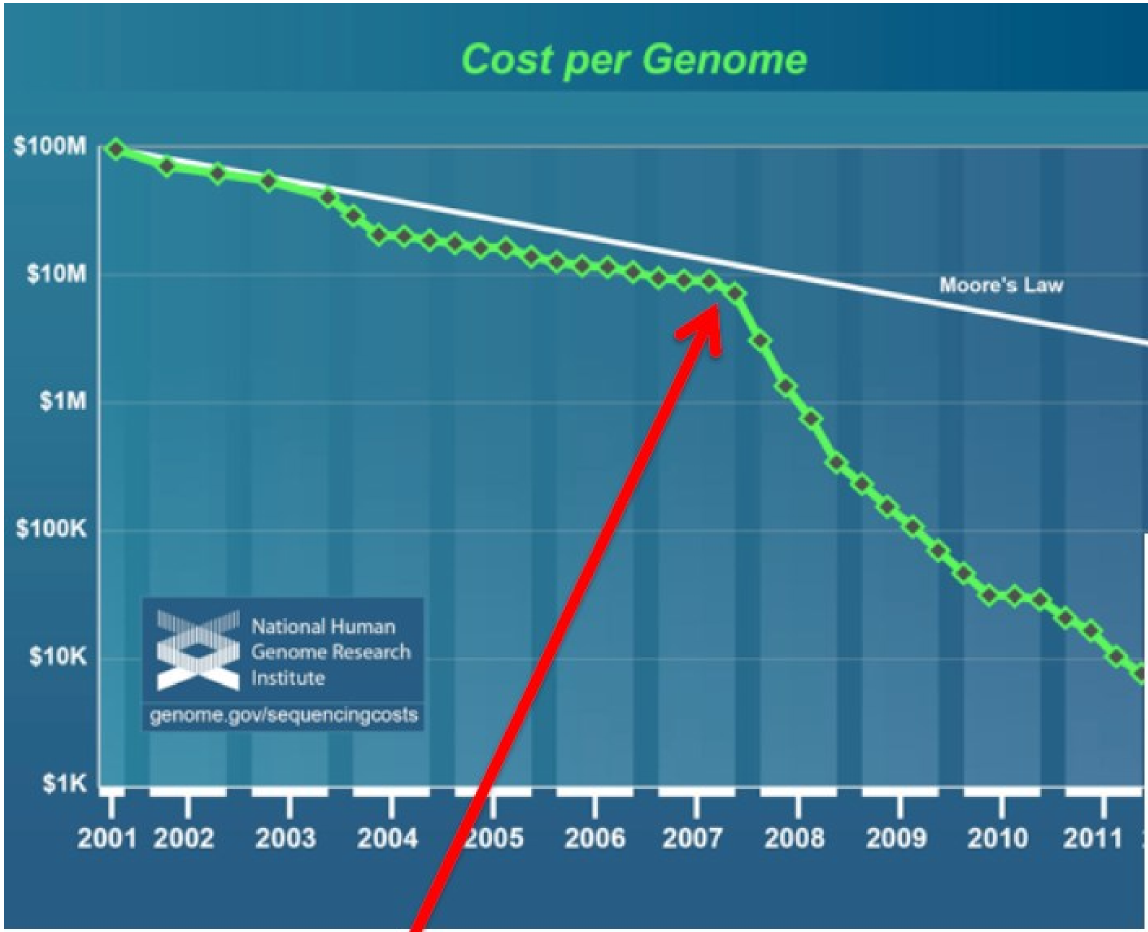


Stratton et al. 2009 Nature 458:719-724

Next-generation sequencing (massivelly parallel sequencing)

- DNA fragmentation
- PCR amplification of all fragments in a single reaction.
- Parallel sequencing of millions or billions of fragments in a single reaction
- The length of obtained sequenes (reads) usually short cca 70 – 300 bp.
- Several hundreds or thousands Gb/run.

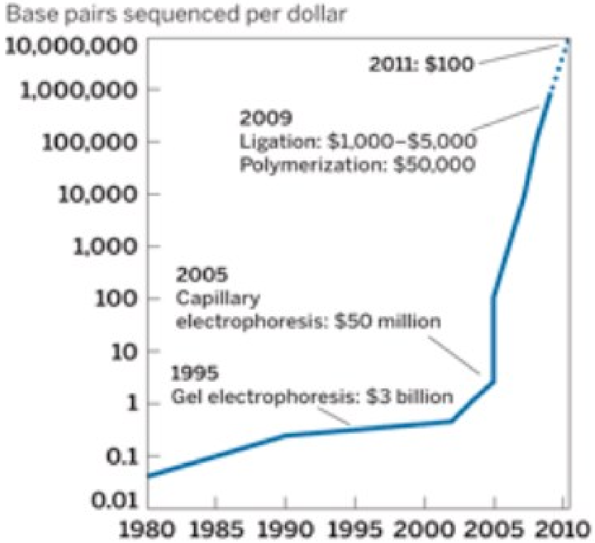




next-generation sequencing boom

A NEW 'MOORE'S LAW'

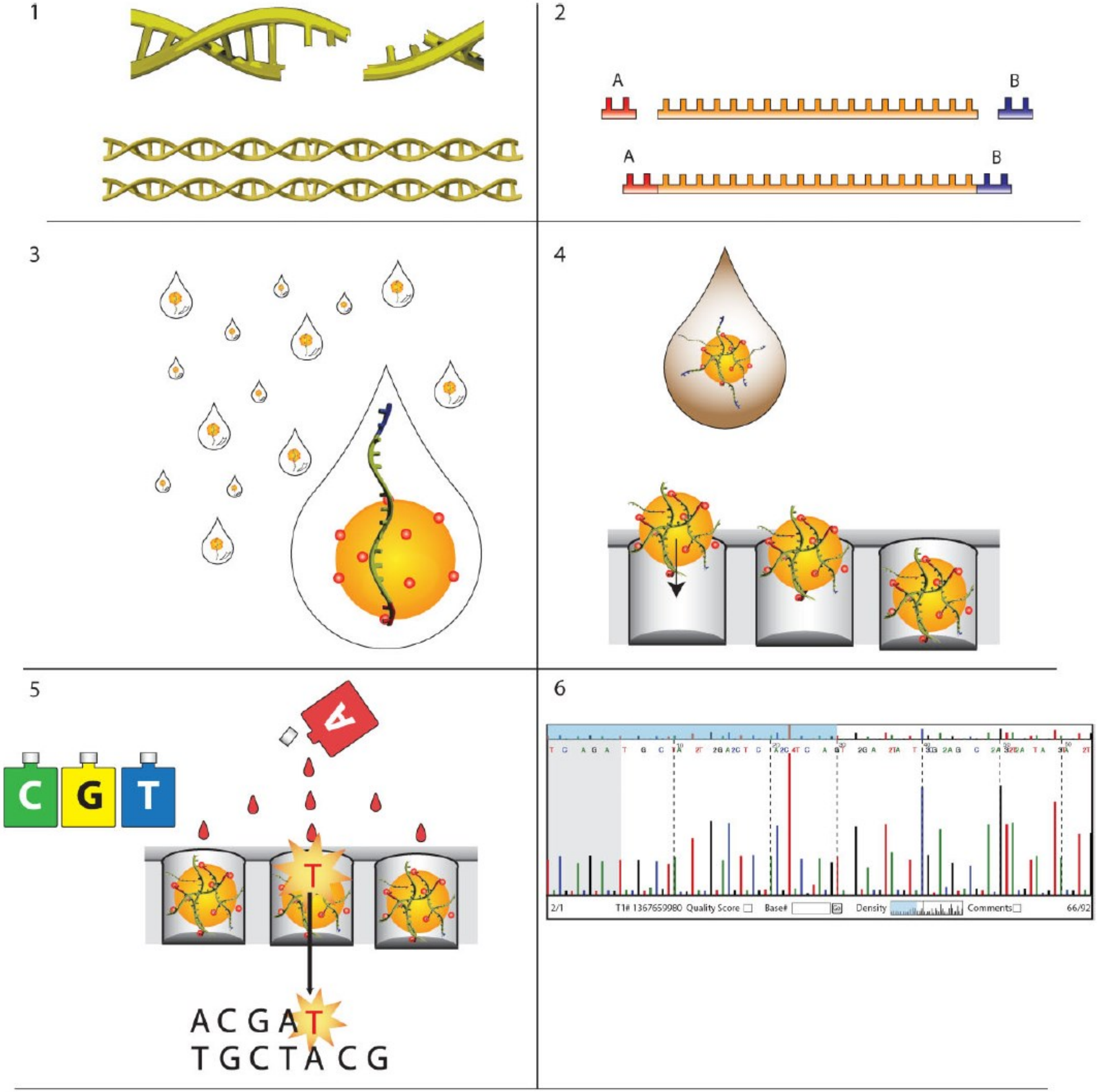
Improvements in DNA sequencing are driving down the cost of whole genomes



NOTE: Dollar figures refer to reagent costs.
SOURCE: George Church, Harvard University

454 - 2005

- emulsion PCR
- pyrosequencing



454 Genome Sequencers

FLX System

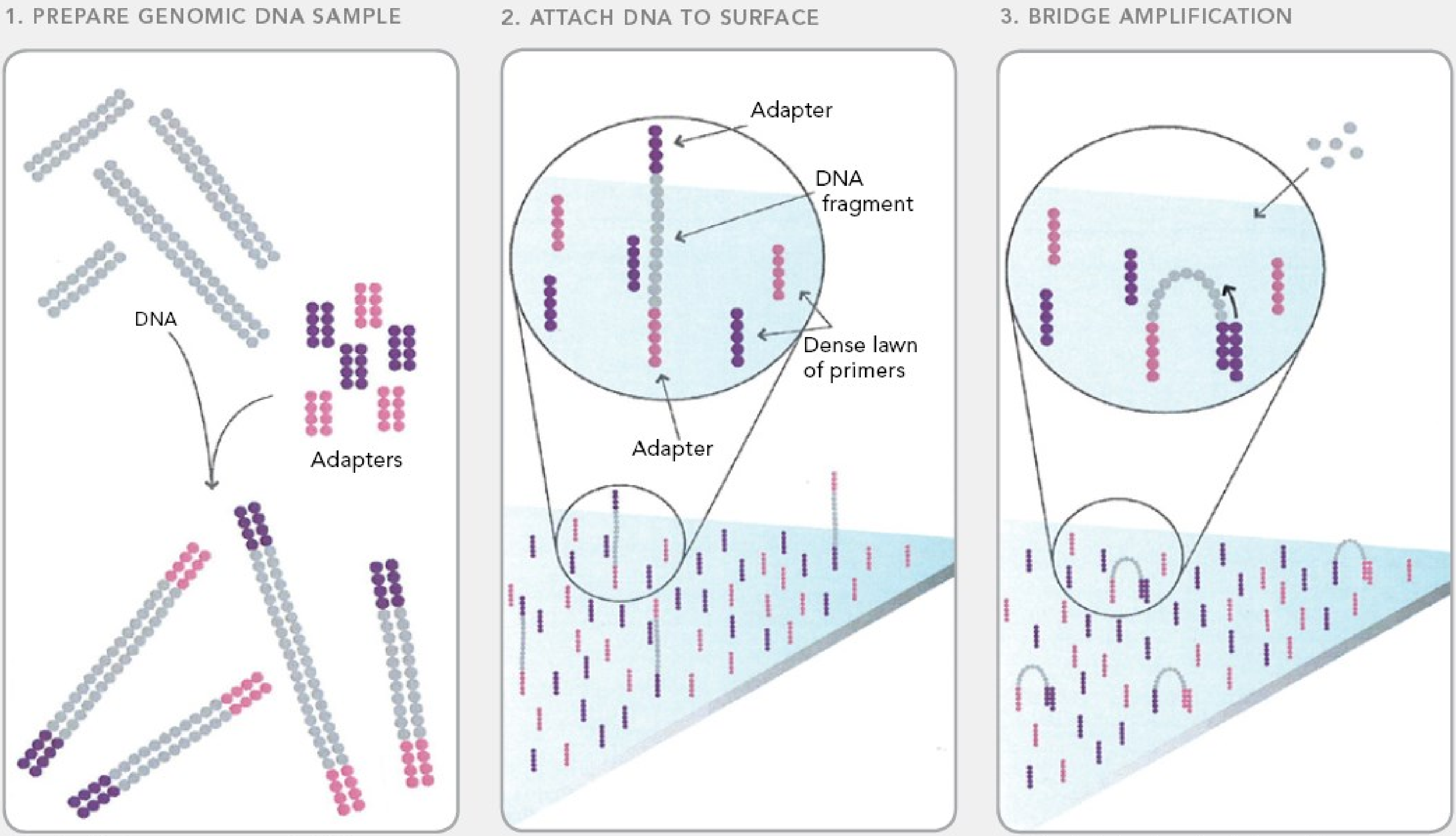
- 1 million of reads/run
- 400-650 bp/read

GS Junior

- 0.1 millions of reads/run
- 400 bp/read

Solexa (Illumina) - 2007

- bridge PCR
- Sequencing by DNA synthesis



1. PREPARE GENOMIC DNA SAMPLE

2. ATTACH DNA TO SURFACE

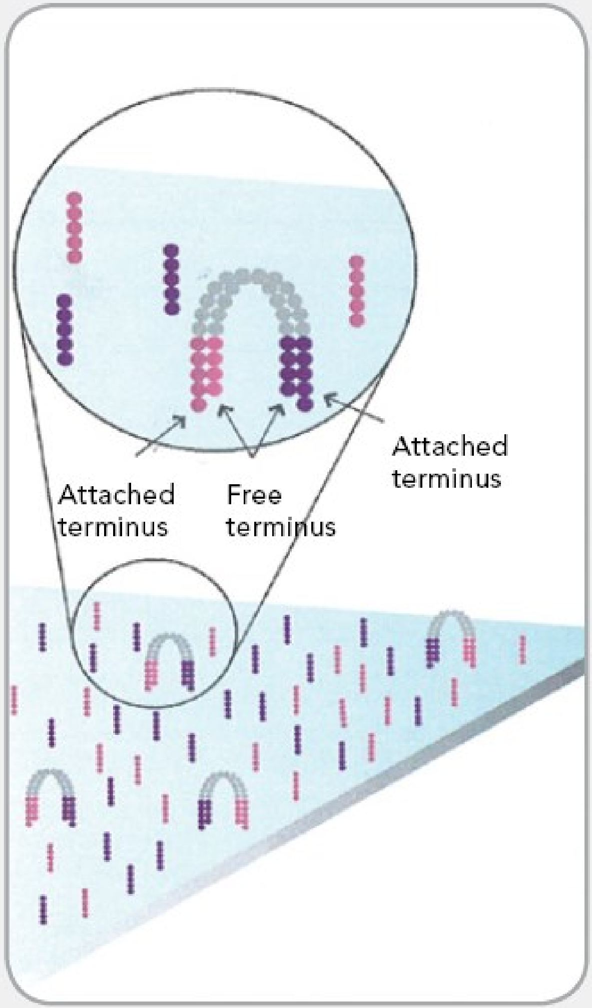
3. BRIDGE AMPLIFICATION

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

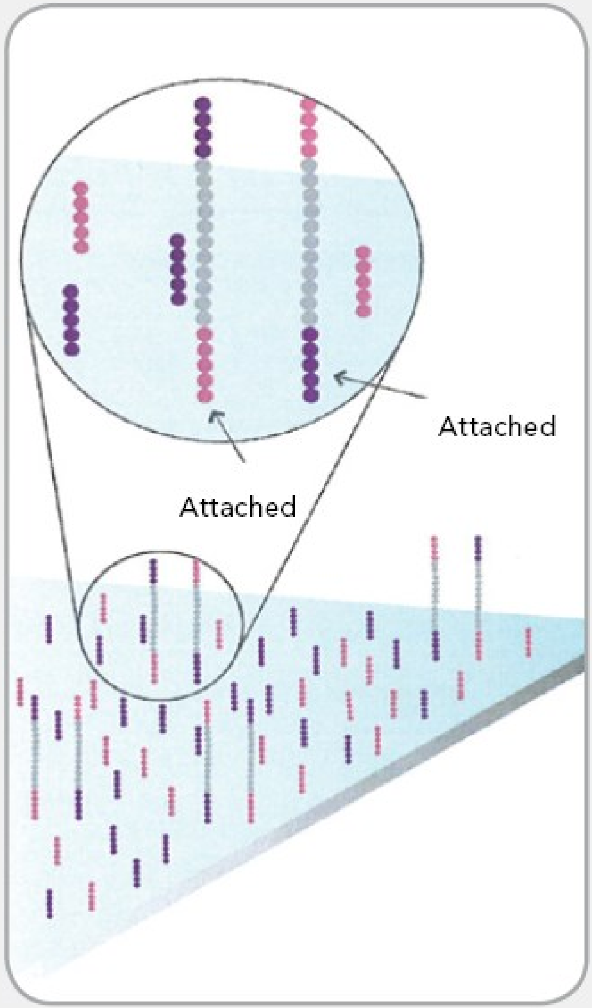
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED



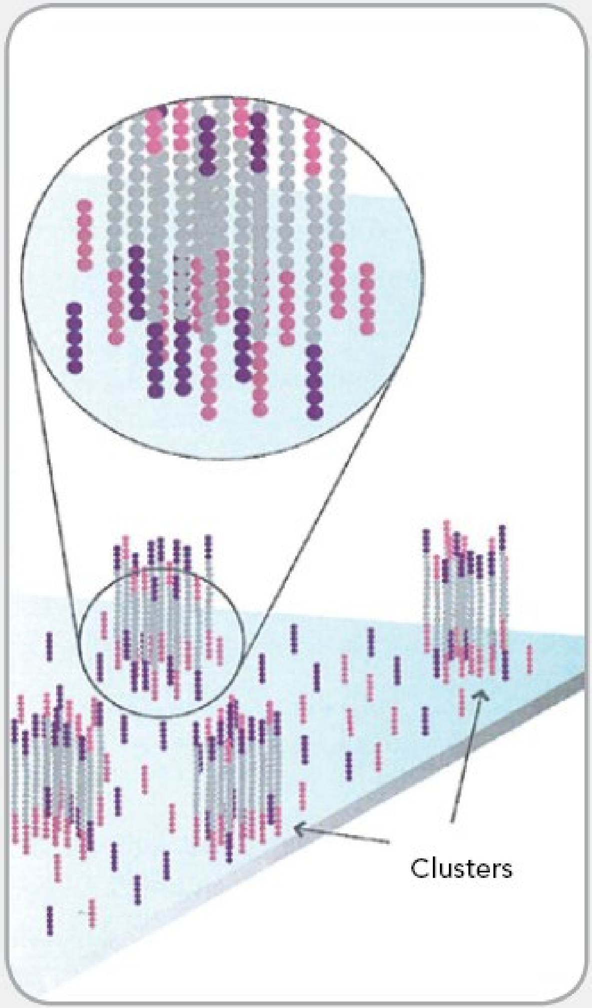
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



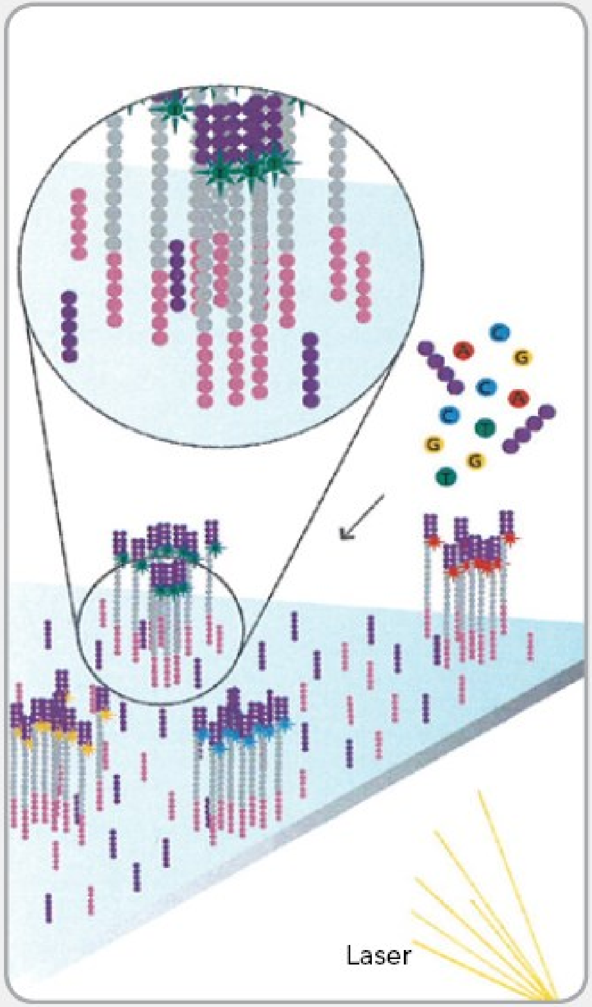
Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION



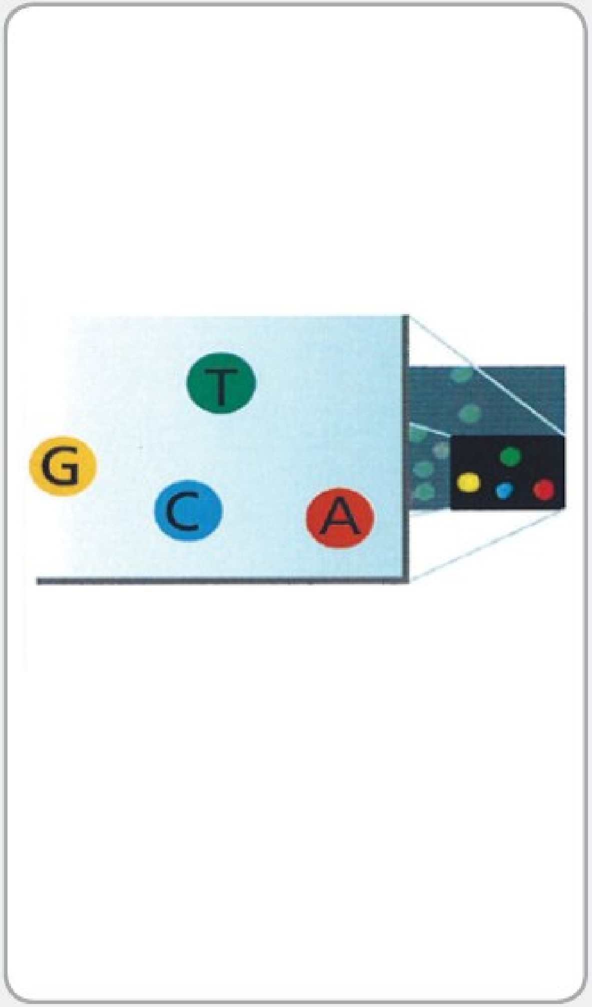
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

7. DETERMINE FIRST BASE



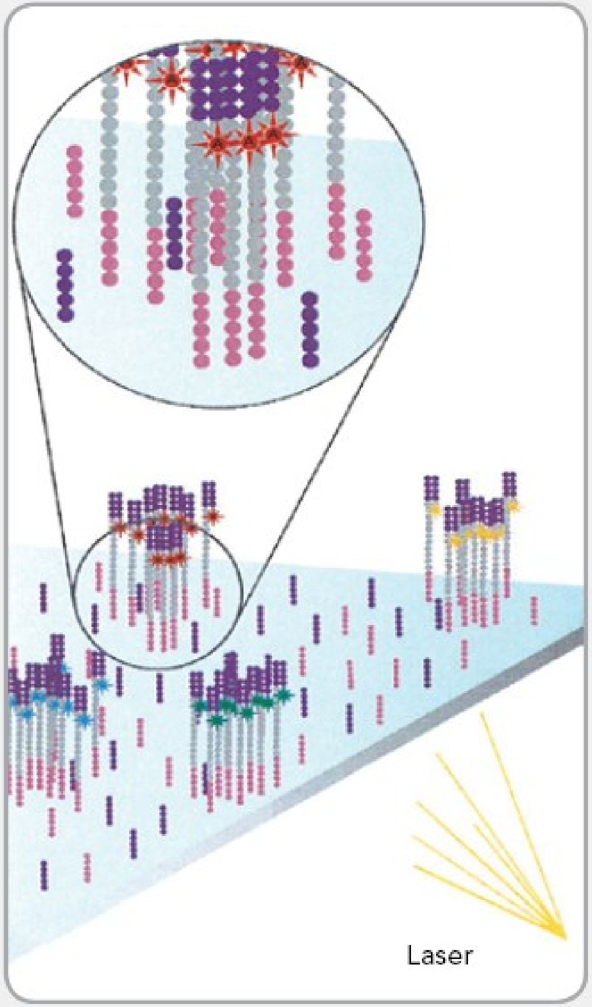
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE



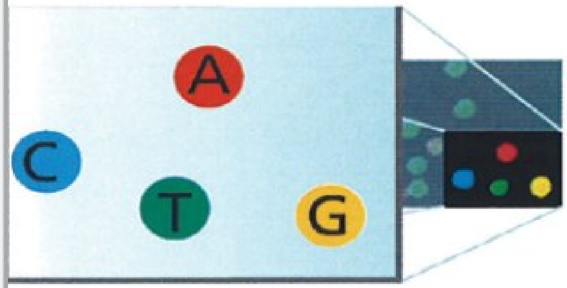
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE



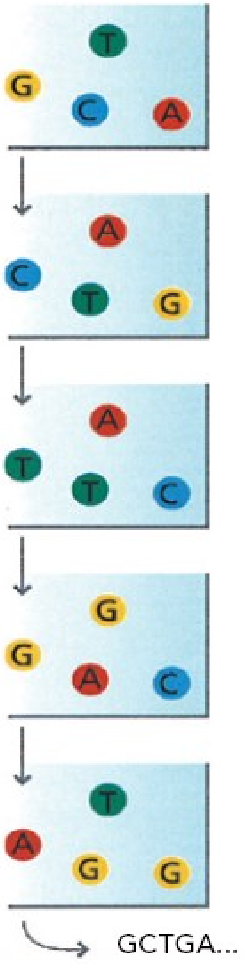
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE



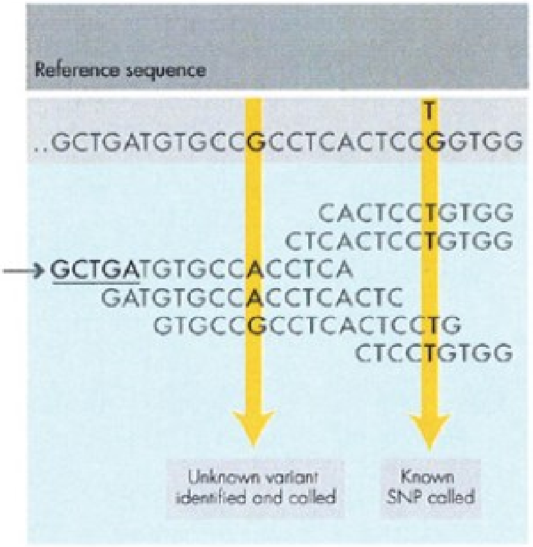
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA



Align data, compare to a reference, and identify sequence differences.

Other next-generation sequencing platforms

Solid (2008) sequencing by ligation

Ion Torrent (2010) Ion-semiconductor sequencing

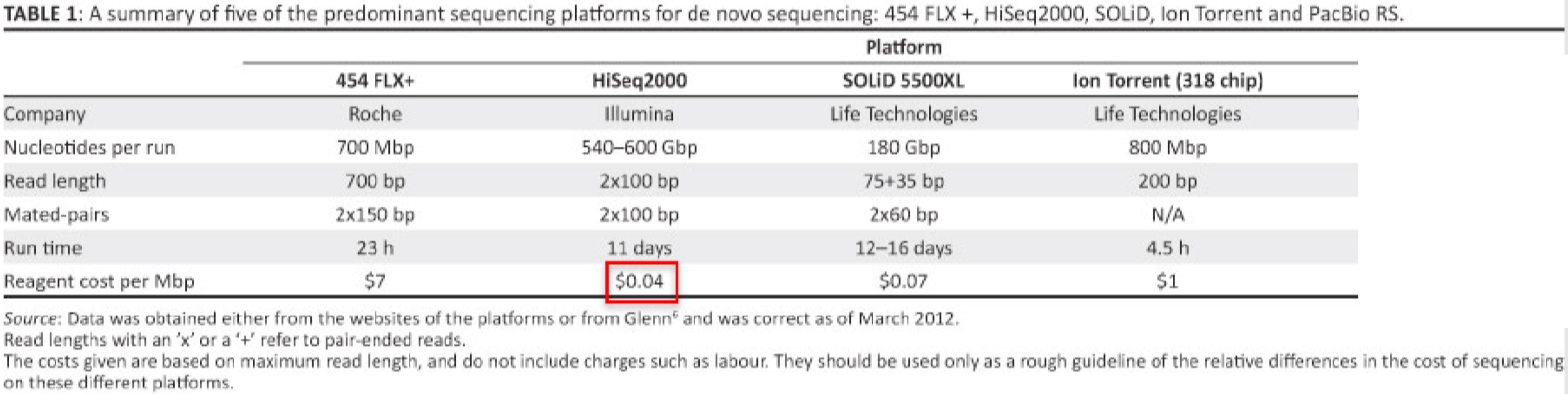


TABLE 1: A summary of five of the predominant sequencing platforms for de novo sequencing: 454 FLX +, HiSeq2000, SOLiD, Ion Torrent and PacBio RS.

	Platform			
	454 FLX+	**HiSeq2000**	**SOLiD 5500XL**	**Ion Torrent (318 chip)**
Company	Roche	Illumina	Life Technologies	Life Technologies
Nucleotides per run	700 Mbp	540–600 Gbp	180 Gbp	800 Mbp
Read length	700 bp	2x100 bp	75+35 bp	200 bp
Mated-pairs	2x150 bp	2x100 bp	2x60 bp	N/A
Run time	23 h	11 days	12–16 days	4.5 h
Reagent cost per Mbp	$7	$0.04	$0.07	$1

Source: Data was obtained either from the websites of the platforms or from Glenn[6] and was correct as of March 2012.
Read lengths with an 'x' or a '+' refer to pair-ended reads.
The costs given are based on maximum read length, and do not include charges such as labour. They should be used only as a rough guideline of the relative differences in the cost of sequencing on these different platforms.

Illumina is currently the most widely used NGS platform.
The highest sequencing output, the lowest cost per bp.
But short reads.

Illumina sequencers

MiSeq

NextSeq

HiSeq

Small genome, amplicon, and targeted gene panel sequencing.

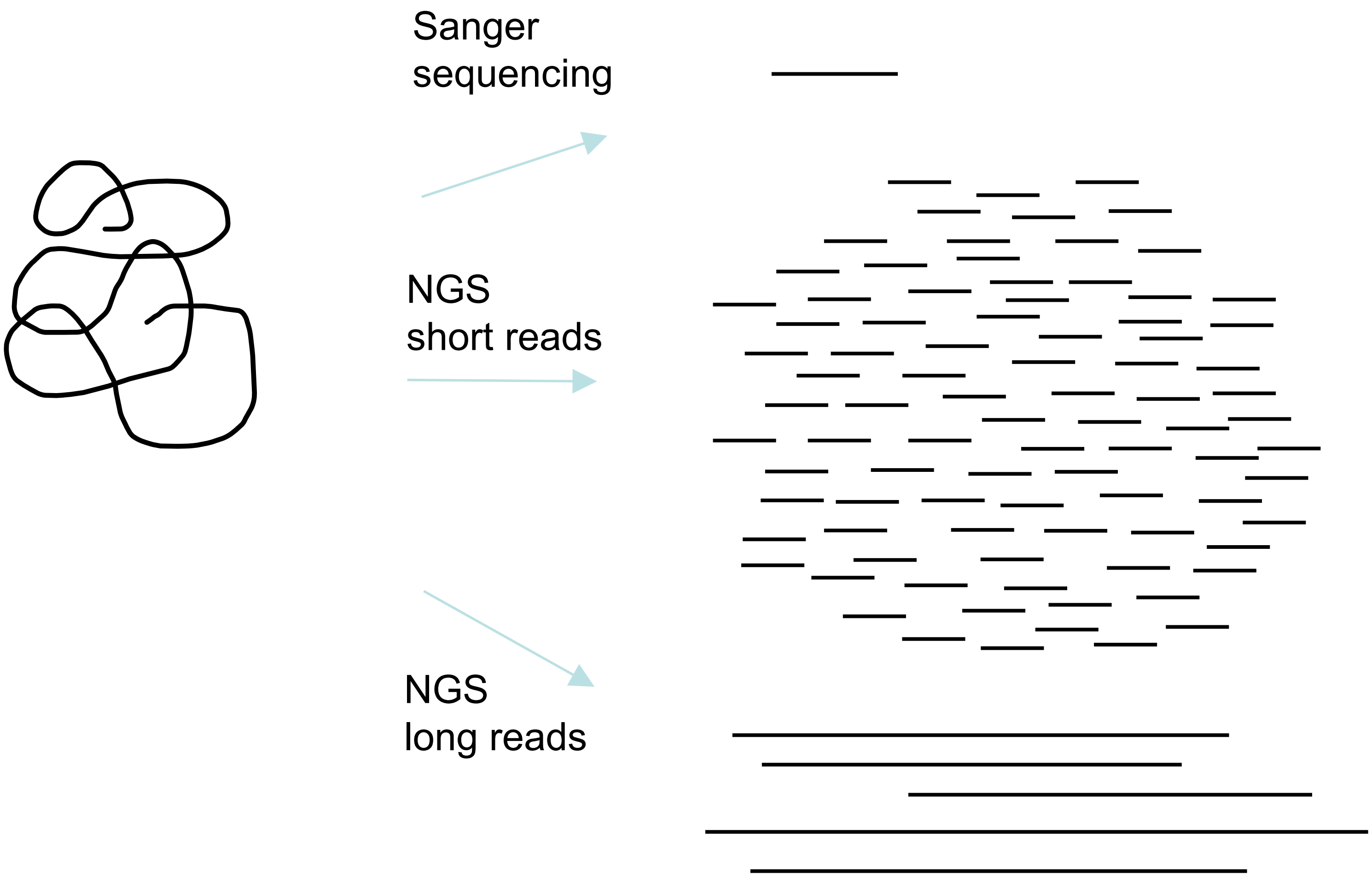
Everyday genome, exome, transcriptome sequencing, and more.

Production-scale genome, exome, transcriptome sequencing, and more.

Output Range	• 0.3-15 Gb	30-120 Gb	50-1000 Gb
Run Time	• 5-55 hours	12-30 hours	<1-6 days
Reads per Flow Cell	• 25 million	400 million	2 billion
Maximum Read Length	• 2 x 300 bp	2 x 150 bp	2 x 125 bp

http://www.illumina.com/systems/sequencing.html

„Second generation“ next-generation sequencing long reads



Sanger sequencing

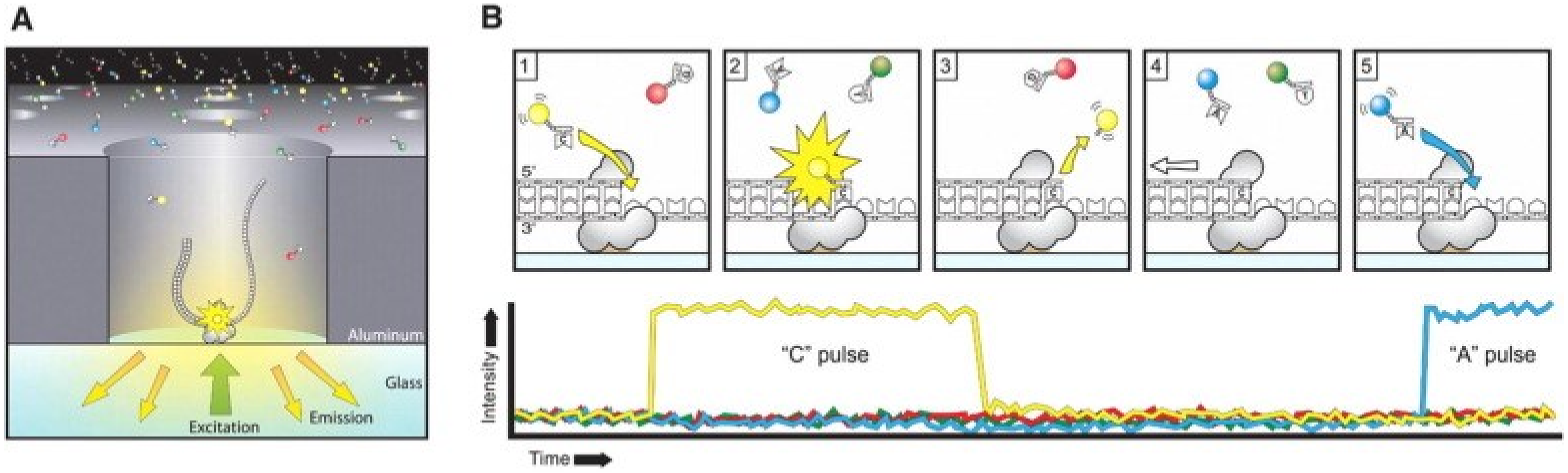
NGS short reads

NGS long reads

„Second generation“ next-generation sequencing long reads

Pacific Biosciences (2010)

- Single-molecule real-time sequencing. PCR is not needed.
- Sequencing during DNA replication. DNA polymerase uses fluorescently labelled nucleotides.
- Long reads (860-1500 bp).

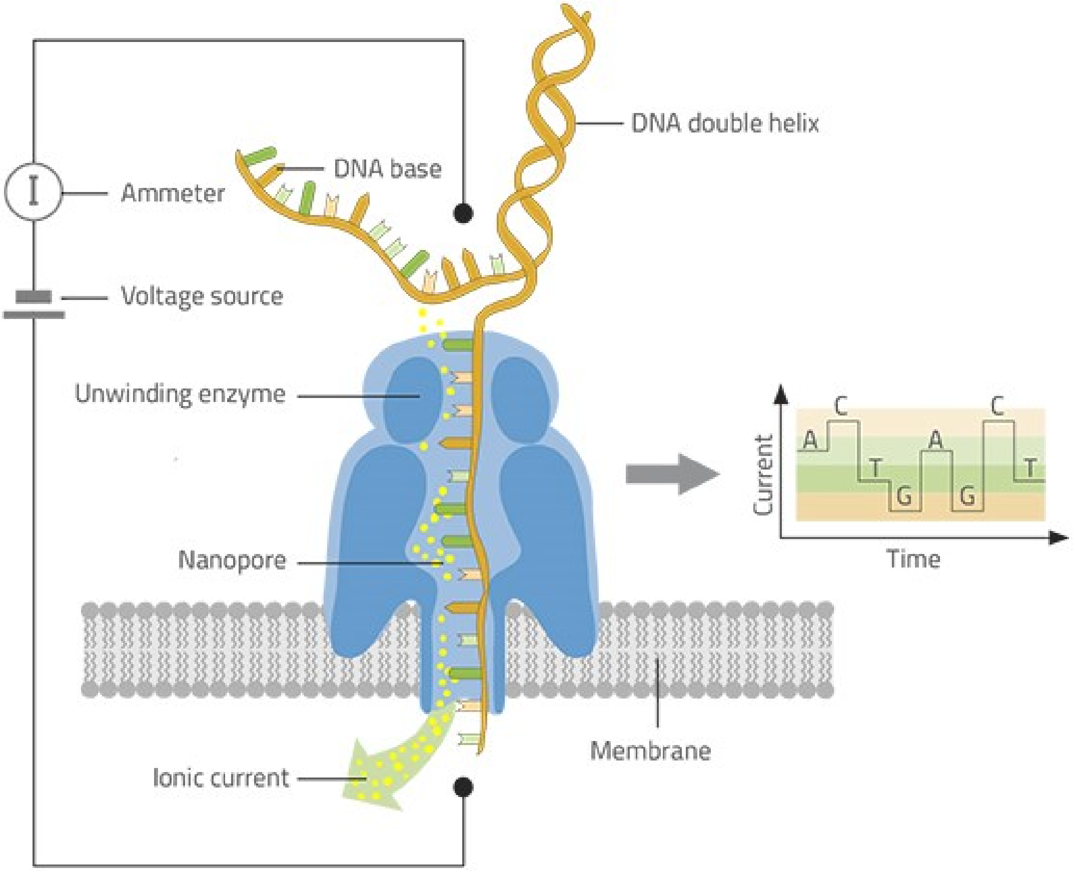


Pacific Biosciences (2010)

- High error rate (cca 15%)
- Lower sequencing output and higher cost/bp in comparison with Illumina.

Oxford Nanopore (2012)

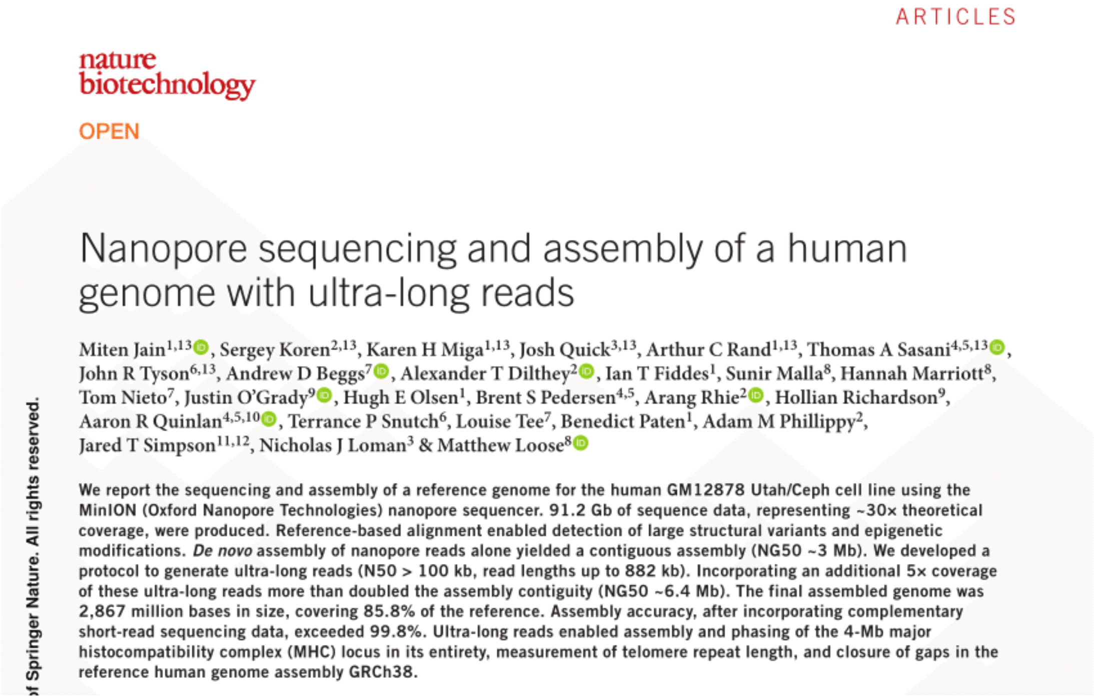
- Single-molecule sequencing. No PCR.
- Nucleotides are determined based on their conductivity during the passage through the nonopore.
- Very long reads (several tens or thousands kbp)!
- High error rate and relatively high cost/bp.



MinION USB stick sequencer

2017

Enables to assemble even repetitive sequences (e.g. MHC genes), determine length of telomeres or detect structural variants (duplications, translocations, inversion).



ARTICLES

nature biotechnology

OPEN

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O'Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30× theoretical coverage, were produced. Reference-based alignment enabled detection of large structural variants and epigenetic modifications. *De novo* assembly of nanopore reads alone yielded a contiguous assembly (NG50 ~3 Mb). We developed a protocol to generate ultra-long reads (N50 > 100 kb, read lengths up to 882 kb). Incorporating an additional 5× coverage of these ultra-long reads more than doubled the assembly contiguity (NG50 ~6.4 Mb). The final assembled genome was 2,867 million bases in size, covering 85.8% of the reference. Assembly accuracy, after incorporating complementary short-read sequencing data, exceeded 99.8%. Ultra-long reads enabled assembly and phasing of the 4-Mb major histocompatibility complex (MHC) locus in its entirety, measurement of telomere repeat length, and closure of gaps in the reference human genome assembly GRCh38.

Comparison of sequencing platforms

Instrument	Run time	Millions of Reads/run	Bases / read	Gbp/run	cost/Gb
Applied Biosystems 3730 (capillary)	2 hrs.	0.000096	650	0.000	$2,307,692.31
454 GS Jr. Titanium	10 hrs.	0.1	400	0.050	$19,540.00
454 FLX Titanium	10 hrs.	1	400	0.400	$15,500.00
Illumina MiSeq v3	55 hrs.	22	600	13.200	$109.24
Illumina NextSeq 500	30 hrs.	400	300	120.000	$33.33
Illumina HiSeq 2500 - high output v4	6 days	2000	250	500.000	$29.90
Illumina HiSeq X (2 flow cells)	3 days	6000	300	1,800.000	$7.08
Ion Torrent – PGM 318 chip	7.3 hrs.	4.75	400	1.900	$460.00
Life Technologies SOLiD – 5500xl	8 days	1410	110	155.100	$67.72
Pacific Biosciences RS II	2 hrs.	0.03	3000	0.090	$1,111.11
Oxford Nanopore MinION (forecast)	≤6 hrs.	0.1	9000	0.900	$1,000.00

http://www.molecularecologist.com/next-gen-fieldguide-2014/

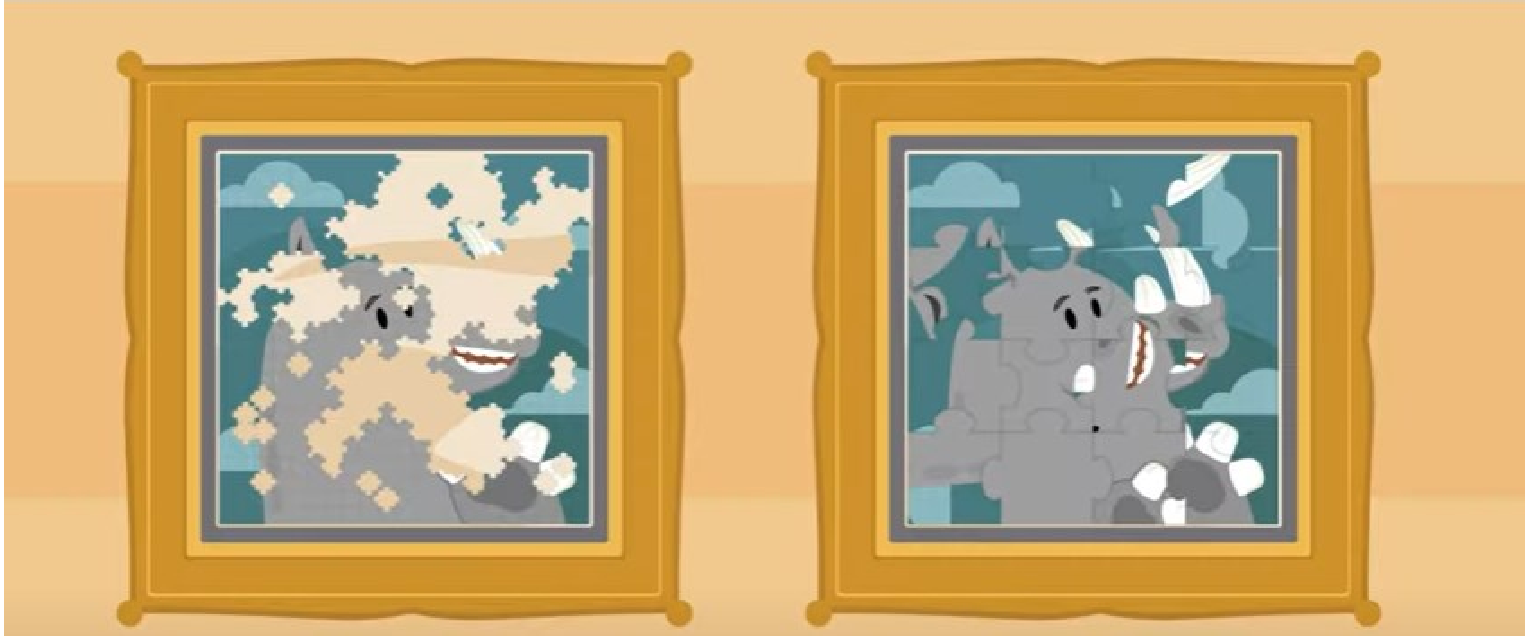
Comparison of sequencing platforms

Error rate

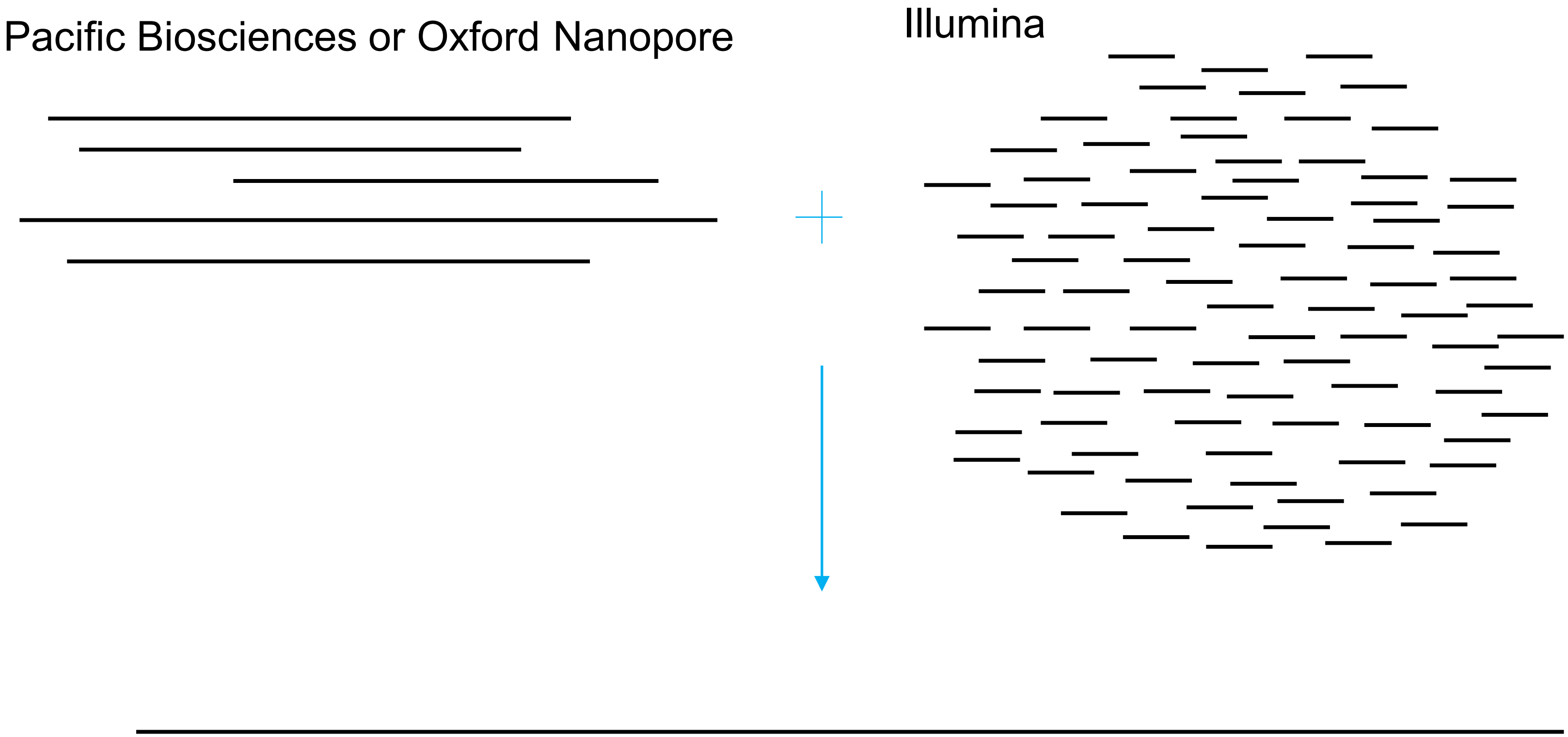
Instrument	Primary Errors	Single-pass Error Rate (%)	Final Error Rate (%)
3730xl (capillary)	substitution	0.1-1	0.1-1
454 All models	indel	1	1
Illumina All Models	substitution	~0.1	~0.1
Ion Torrent – all chips	Indel	~1	~1
SOLiD – 5500xl	A-T bias	~5	≤0.1
Oxford Nanopore	deletions	≥4*	4*
PacBio RS	Indel	~13	≤1

http://www.molecularecologist.com/next-gen-fieldguide-2014/

Short Read
Long Read



Combination of long and short reads from the same sample allows to reconstruct the high-quality genome sequence.

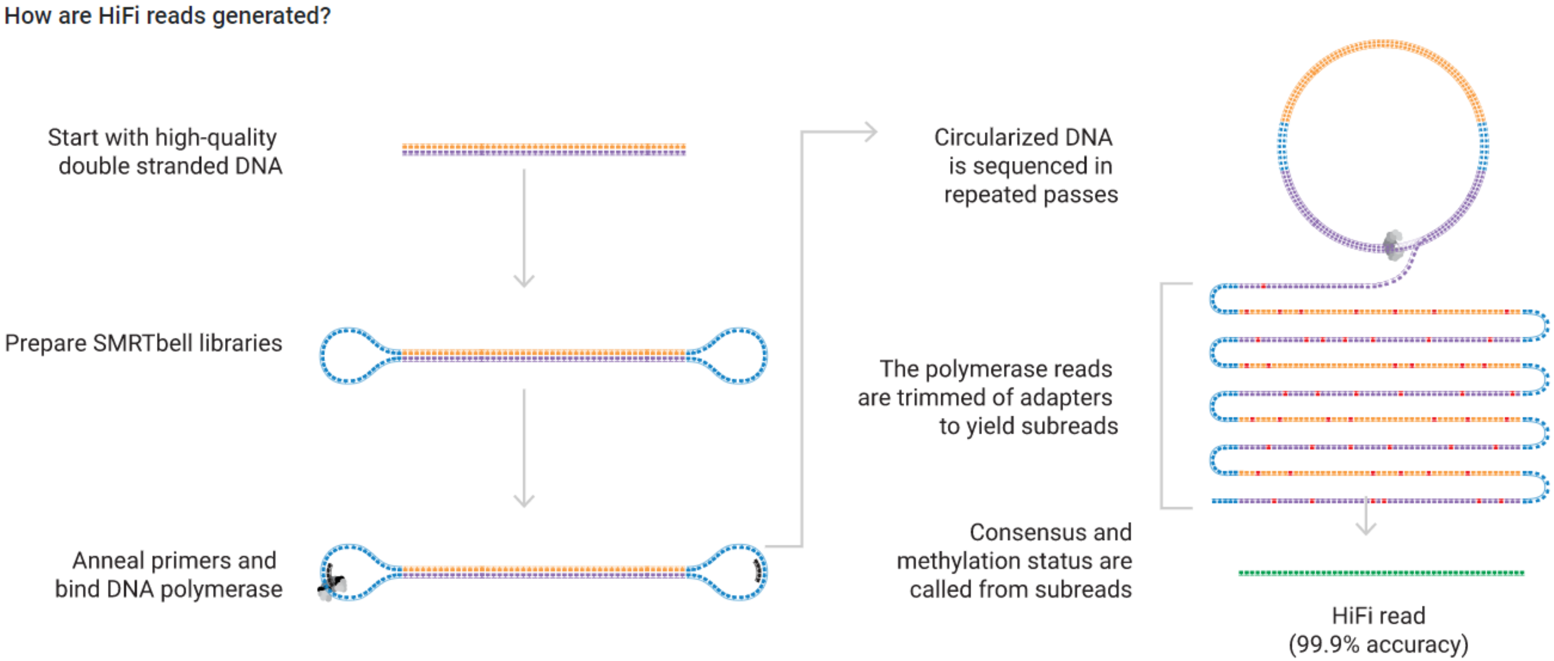


Pacific Biosciences or Oxford Nanopore

Illumina

Hi-Fi (High Fidelity) Sekvenování

Uses PacBio sequencing



How are HiFi reads generated?

Start with high-quality double stranded DNA

Prepare SMRTbell libraries

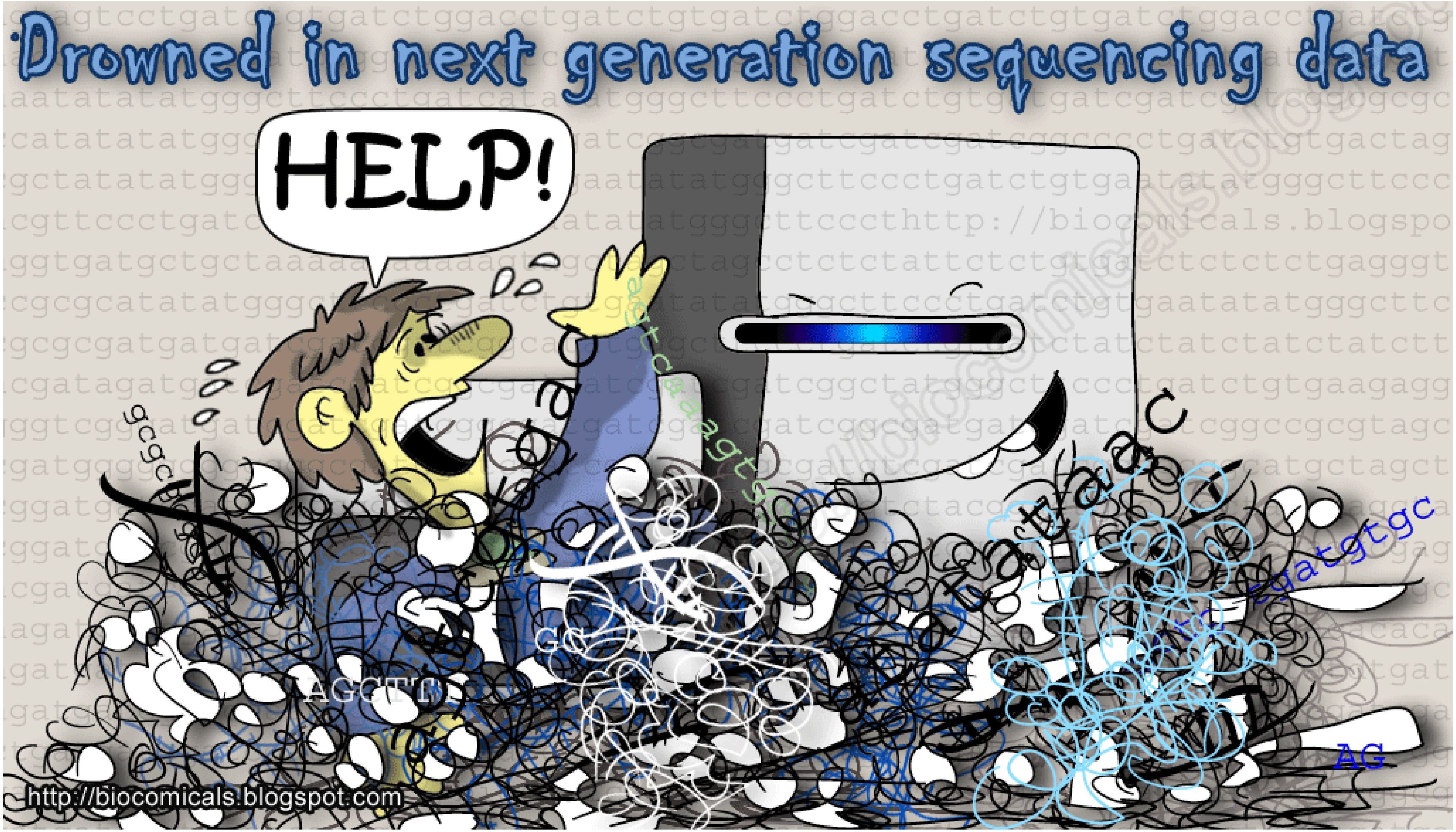
Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus and methylation status are called from subreads

HiFi read (99.9% accuracy)


Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com