
🔍 1. Sequence-Based Function Prediction

Even without knowing the 3D structure, **functional clues** can be found from the sequence alone:

- **Homology search**: Use tools like **BLAST**, **PSI-BLAST**, or **HMMER** to find similar sequences in known databases. If a similar gene/protein has a known function, it can suggest a similar role.
- **Conserved domains/motifs**: Identify known **protein domains** (via databases like **Pfam**, **InterPro**, or **CDD**) which are associated with specific functions.
- **Gene Ontology (GO) annotation**: Based on sequence similarity and curated annotations, GO terms (biological process, molecular function, cellular component) are assigned.
- **Machine learning**: Newer models use neural networks or transformers (like AlphaFold or ProtT5) to predict function from sequence patterns.

🧬 2. Structure-Based Function Prediction

Knowing the **3D structure** (via X-ray, NMR, or AlphaFold) opens up deeper insights:

- **Active site identification**: Structural motifs like catalytic triads or binding pockets hint at enzymatic or receptor activity.
- **Molecular docking**: Simulate interactions with small molecules or ligands to predict binding and reaction possibilities.
- **Comparative modeling**: Structural alignment with functionally known proteins helps infer potential roles.
- **Electrostatics & surface analysis**: Charge, hydrophobicity, or topology can hint at interaction partners or substrates.

🔬 3. Experimental Functional Validation

Computational predictions are **hypotheses**. True function is confirmed through **wet lab experiments**:

- **Gene knockouts / RNAi / CRISPR**: Remove the gene and observe phenotype.
- **Reporter assays**: Test regulatory or enzymatic activity in cells.
- **Protein-protein interaction studies**: Co-IP, yeast two-hybrid, FRET, etc.
- **In vitro assays**: Biochemical reactions to confirm enzyme activity or binding.

🌐 Databases and Tools Used

- **UniProt** – curated functional information
- **STRING** – protein interaction networks
- **KEGG / Reactome** – functional pathways
- **AlphaFold DB** – structural models with confidence scores

---

🧠 Summary

Layer	Approach	Tools / Methods
Sequence	Homology, motifs, ML	BLAST, Pfam, GO, ProtT5, DeepGOPlus
Structure	Docking, pocket analysis	PyMOL, AlphaFold, DALI, COACH
Function	Experiments, modeling	CRISPR, assays, STRING, KEGG

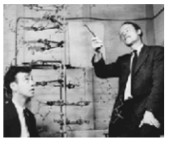
DNA sequencing is the process of determining the order of the four chemical bases (A, T, C, G) that make up DNA. This sequence reveals genetic information, helping scientists identify genes, regulatory regions, and disease-causing mutations.

In the DNA double helix, bases always pair in a specific way: adenine (A) with thymine (T) and cytosine (C) with guanine (G). This pairing allows DNA to be copied during cell division and forms the basis of sequencing techniques. The human genome has about 3 billion base pairs that guide human development and maintenance.

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

A quick history of sequencing

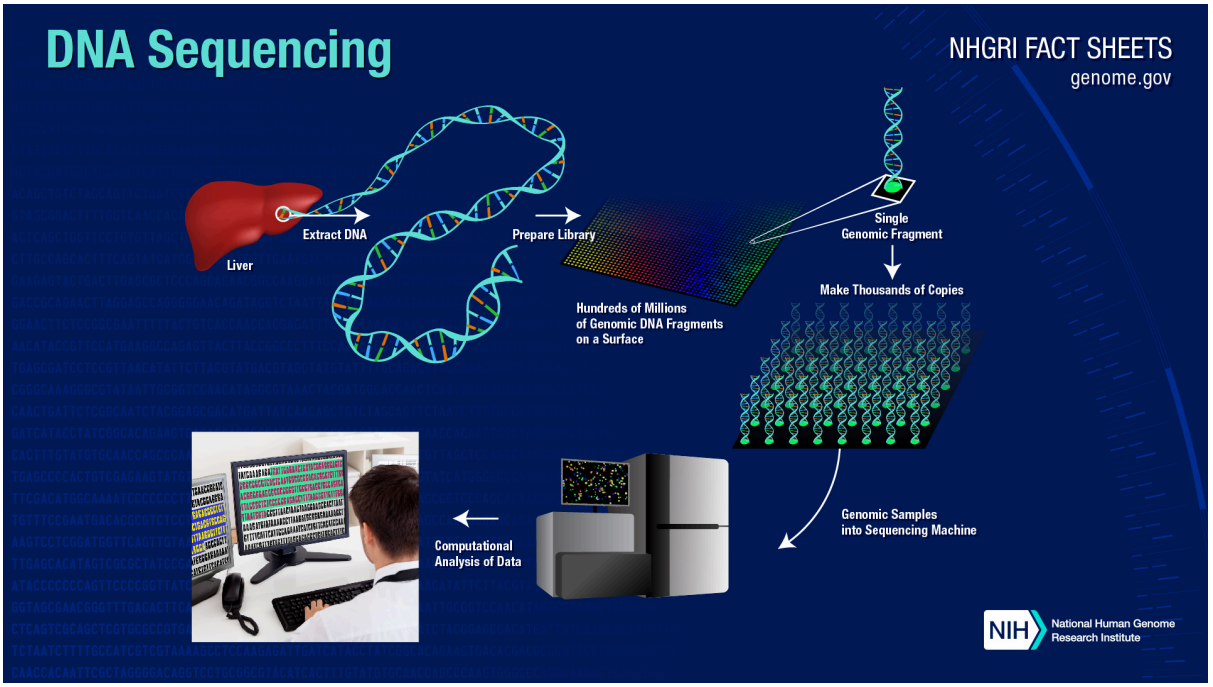
1869 – Discovery of DNA
1909 – Chemical characterisation
1953 – Structure of DNA solved
1977 – First genome (ΦX174) Sequencing by synthesis (Sanger) - Sequencing by degradation (Maxam- Gilbert)
1986 – First automated sequencing machine 1990 – Human Genome Project started
1992 – First "sequencing factory" at TIGR
1995 – First bacterial genome – *H. influenzae* (1.8 Mb) 1998 – First animal genome – *C. elegans* (97 Mb)
2003 – Completion of Human Genome Project (3 Gb) – 13 years, $2.7 bn
2005 – First "next-generation" sequencing instrument
2013 – 10,000 genome sequences in NCBI database



Since the Human Genome Project, advancements in technology have made DNA sequencing faster and cheaper. It's now common to sequence individual genes, and labs can sequence large amounts of DNA each year at a lower cost. The goal is to eventually sequence a human genome for under $1,000.

New sequencing technologies are being developed, including methods that track DNA polymerase with a fast camera and dye colors, and nanopore sequencing, where DNA strands pass through tiny pores, identifying bases based on their electrical effects.

Improved sequencing methods are advancing human health by enabling rapid comparisons of DNA, revealing links between genetics, diseases, and environmental factors. For example, in cancer care, sequencing helps doctors choose the best treatments based on a patient's specific genetic information. DNA sequencing is also being used to identify genetic causes of rare diseases, screen newborns, and study common diseases like heart disease and diabetes. Comparing DNA across species helps scientists understand development and evolution.



DNA sequencing determines the exact order of nucleotides in DNA. Before direct sequencing methods, the process was indirect, requiring DNA to be converted to RNA. This method was slow and could only sequence short DNA strands. In the 1970s, researchers like Walter Gilbert and Alan Maxam developed ways to sequence DNA more directly, with the Lac operator being the first long sequence determined.

With advances in sequencing technology, from traditional methods to newer techniques like pyro-sequencing, sequencing has become faster and cheaper, allowing for detailed genetic studies. Pyrosequencing, a key method, is based on DNA synthesis, generating light signals to determine sequences in real time.

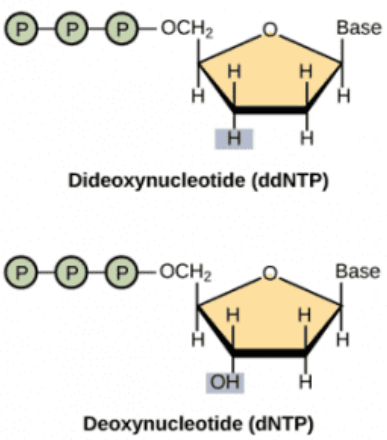
Key Methods:

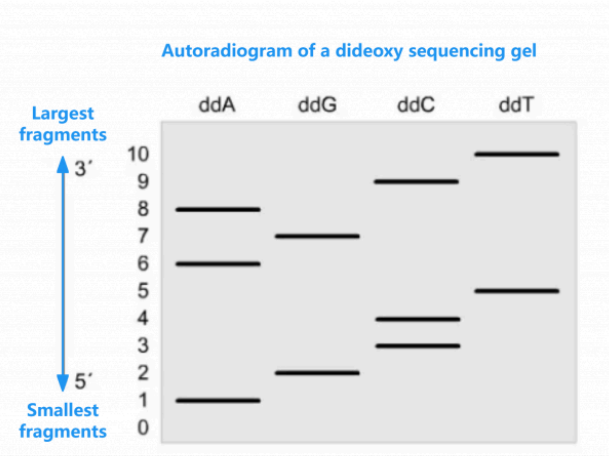
What is Sanger Sequencing? Or the Chain termination Method

Sanger sequencing, developed by Frederick Sanger in 1977, is a method for determining the order of nucleotide bases in a DNA strand, typically for smaller DNA fragments (less than 1,000 base pairs). It's highly accurate, with 99.99% base accuracy, making it the "gold standard" for validating DNA sequences, even those obtained through newer methods like next-generation sequencing (NGS). It was used in the Human Genome Project to sequence small DNA fragments.

In the presence of the four deoxynucleotide triphosphates (**dNTPs: A, G, C, and T**), the polymerase extends the primer by adding the complementary dNTP to the template DNA strand. To determine which nucleotide is incorporated into the chain of nucleotides, four dideoxynucleotide triphosphates (**ddNTPs: ddATP, ddGTP, ddCTP, and ddTTP**) labeled with a distinct fluorescent dye are used to terminate the synthesis reaction. Compared to dNTPs, ddNTPs has an oxygen atom removed from the ribonucleotide, hence cannot

form a link with the next nucleotide. Following synthesis, the reaction products are loaded into four lanes of a single gel depending on the diverse chain-terminating nucleotide and subjected to gel electrophoresis. According to their sizes, the sequence of the DNA is thus determined.





AGCCTAGACT

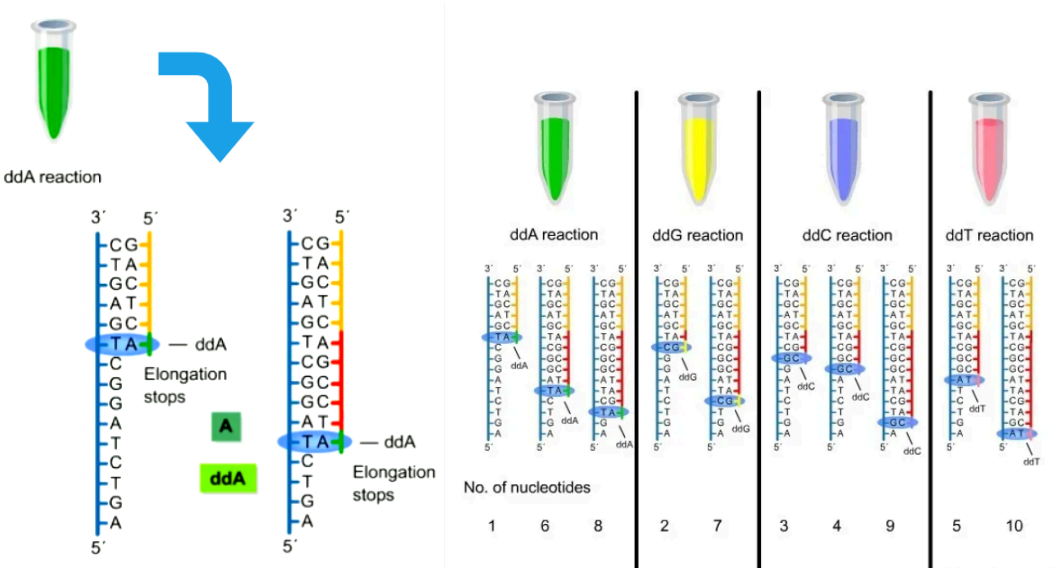
TCGGATCTGA

If the template DNA sequence is 5′-AGTCTAGGCTGAGTC-3′ and the primer is 5′-GACTC-3′, the sequencing process could generate sequences like 5′-GACTCAGCCT-3′ or 5′-GACTCAGCCTAGACT-3′, depending on where the synthesis stops at the T bases of the template strand (where ddT is used).

It's important to use the correct amount of ddT, because too much or too little can affect where the DNA strands stop. If too much ddT is used, most strands will stop early, leading to an incorrect pattern. This can cause issues in identifying the correct positions of T bases in the DNA sequence.

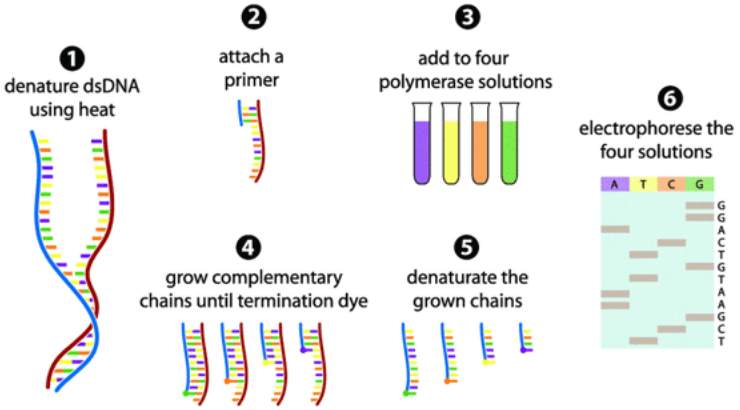
Once the DNA strands are created, they are analyzed using gel electrophoresis to determine the sequence. The lengths of the fragments help pinpoint where the adenine (A) bases match the thymine (T) bases in the template strand.

In the Sanger method, four types of dideoxynucleotide triphosphates (ddNTPs) are added, which stop the DNA synthesis at specific points, allowing the sequence to be determined based on where these stops occur.



How Sanger Sequencing Works:

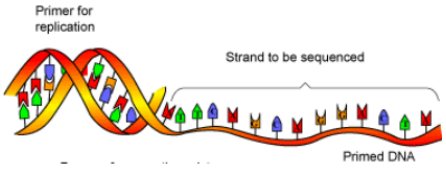
DNA is split into two single strands.
A primer is added to one strand to start DNA synthesis.
Four types of normal nucleotides (dNTPs) and four special terminator nucleotides (ddNTPs) are added. The ddNTPs stop DNA synthesis.
The DNA is synthesized, and the process randomly stops when a ddNTP is added.
The fragments are separated by size using gel electrophoresis.
The sequence is determined by reading the order of these fragments.

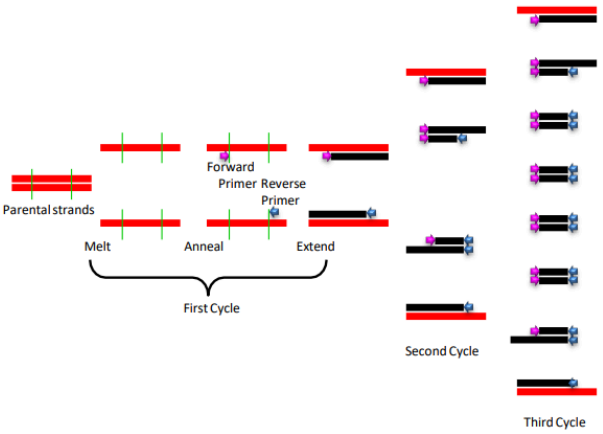


Sanger sequencing is still widely used for specific tasks like gene cloning, SNP identification, and validating other sequencing methods.

Sanger Method

- Sequence of interest is targeted via designed primers
- Massive amplification of target (e.g. by ca. 35 rounds of PCR amplification)



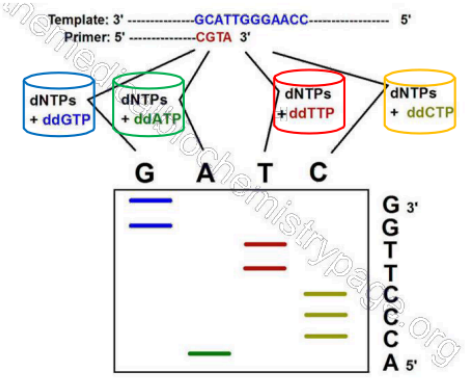


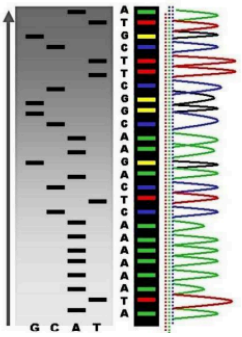
Sanger sequencing: dye-terminator sequencing

- Elongation of DNA sequences by polymerase
- Enzyme stops at a random position per copy (by **dd**NTP)
- Terminated copies are separated within a gel (smaller ones run further)
- Sequence can be read directly

→ Extremely accurate: „gold standard" (error rate ~1:100,000)

→ Slow: poorly parallelizable (60x max.)





5' TGGGAACC3'

Sanger sequencing, while accurate, suffers from limitations like low throughput, slow processing time, and high cost, especially for large-scale projects. It's also less suitable for complex mixtures or detecting rare mutations due to its low sensitivity. Furthermore, Sanger sequencing can only sequence relatively short DNA fragments (about 300 to 1000 base pairs), and the quality of the sequence degrades after a certain length.

1. **Maxam and Gilbert Method**: This technique uses chemical reactions to cleave DNA at specific bases. The fragments are then separated by size and analyzed. However, it's less popular due to its use of hazardous chemicals and slower processing time.

   Principle of Maxam-Gilbert sequencing: Four separate reactions were carried out for the degradation of bases in a single stranded DNA fragment: A+G, G, C+T and C. DNA fragments of different length are obtained following base degradation and cleavage of sugar-phosphate backbone. The products are loaded into four separate wells in polyacrylamide gel. The sequence is read from bottom to top as GTATGC. If a G is found opposite to a gap in the gel, that is confirmed to be 5-methylCytosine in the template strand.

🔍 Why use A+G and C+T combinations?

- The method relies on **chemically modifying bases** so that the DNA strand can be **cleaved at specific sites**.
- However, **modifying and cleaving only one base (e.g., A or G alone) is less efficient**, so combinations are used to **increase sensitivity and coverage**.

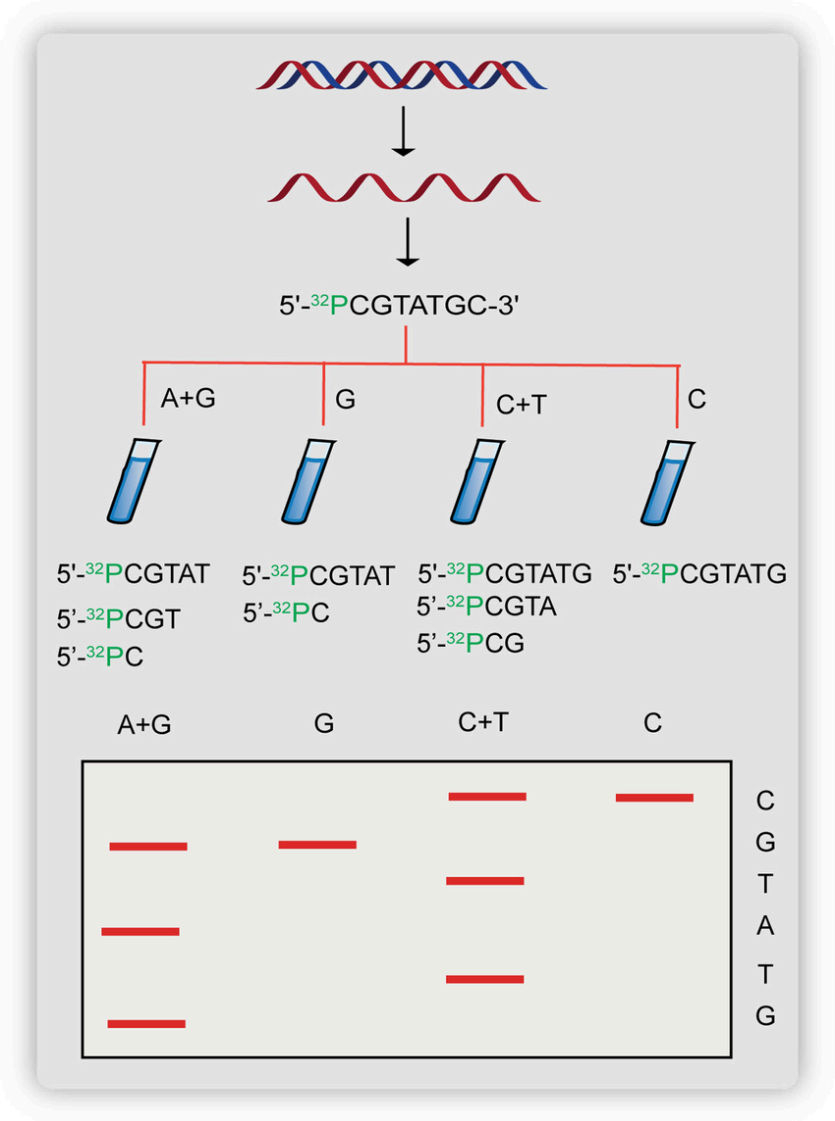
🧪 Explanation of Each Reaction:

Reaction	Modifies/Cleaves at	Purpose
G	Guanine	Identifies **only G** positions
A+G	Adenine **and** Guanine (purines)	Identifies **A and G** positions together
C	Cytosine	Identifies **only C** positions
C+T	Cytosine **and** Thymine (pyrimidines)	Identifies **C and T** positions together

---

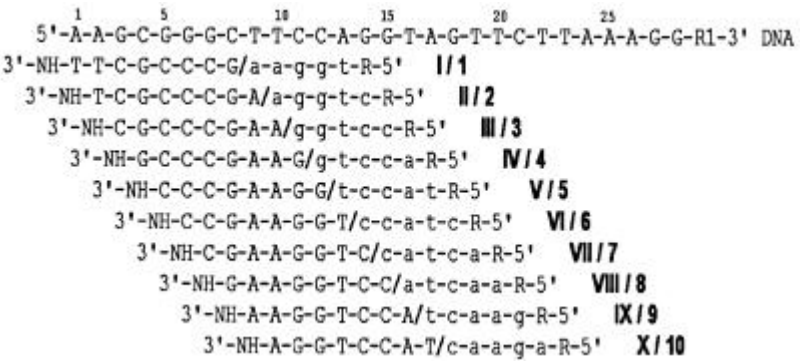
🧬 How It Helps Sequencing:

- By **comparing the A+G lane with the G-only lane**, you can **deduce which bands correspond to adenine** (A = bands in A+G but not in G).
- Similarly, **comparing C+T to C** lets you isolate **T residues** (T = bands in C+T but not in C).



The Maxam-Gilbert method, while historically significant, has several disadvantages including the use of hazardous chemicals, low throughput, and limited read length compared to modern sequencing methods. It's also technically complex and difficult to scale up, making it less practical for routine use, particularly in high-throughput applications.

2. **Hybridization Method**: This method involves hybridizing DNA to arrays of short sequences (oligonucleotides), then using a computer to analyze the pattern. It's used for detecting specific DNA sequences but can have ambiguities.



A **28-nucleotide (nt) DNA fragment** is being analyzed. Below this DNA sequence:

- **10 gel-attached 8-nucleotide (8mer) probes** are shown. These are labeled I to X and match specific sections of the DNA.
  10 short, 5-nucleotide (5mer) sequences in solution are also shown. These match the DNA immediately next to where the 8mers bind and are labeled 1 to 10. The **8mers** are attached to a **microchip surface**, while the **5mers** are **free in solution**. Each 8mer and its matching 5mer pair is separated by a slash (/) in the diagram.

Chemical Labels:

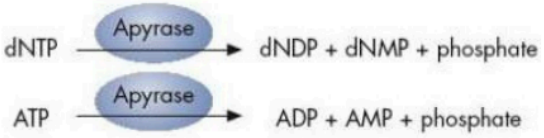
- The **5mers** have special chemical groups (called **R or R1**) for detection:
  R = -(CH2)6-NH2 is used for **mass spectrometry**.
  R = -(CH2)6-NH-TR or **R1 = FITC** is used for **fluorescence** detection.
  These labels help researchers monitor whether the DNA fragment hybridizes correctly using either **mass spectrometry** or **fluorescence**.

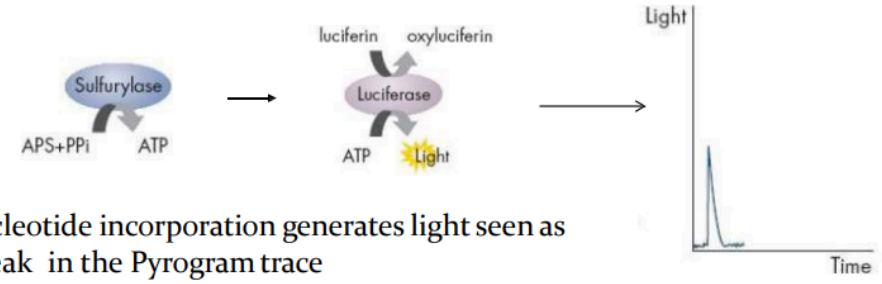
Purpose:

This setup helps detect and analyze specific DNA sequences by using a **chip-bound probe** (8mer) and a **short signal-generating probe** (5mer), making it easier to identify DNA with high accuracy.

3. **Pal Nyren's Method (Pyrosequencing)**: Pyrosequencing uses an enzymatic reaction to generate light signals that correspond to nucleotide incorporation during DNA synthesis, offering real-time results. It's efficient and has been improved for high-throughput sequencing.

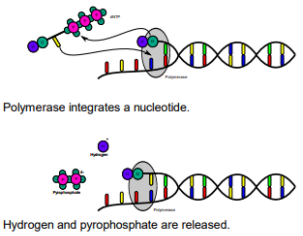
Pyrosequencing: non-electrophoretic, bioluminescence method that measures the release of inorganic pyrophosphate by proportionally converting it into visible light using a series of enzymatic reaction



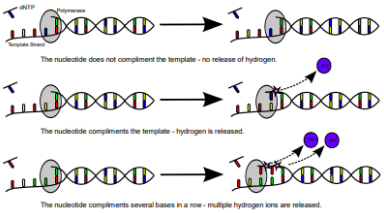


Nucleotide incorporation generates light seen as a peak in the Pyrogram trace

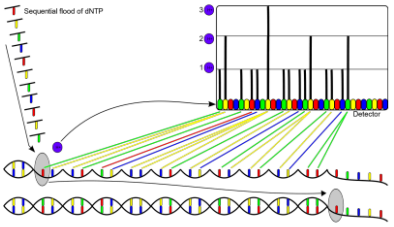
5a. Proton detection sequencing



The incorporation of deoxyribonucleotide Triphosphate into a growing DNA strand causes the release of hydrogen and pyrophosphate.

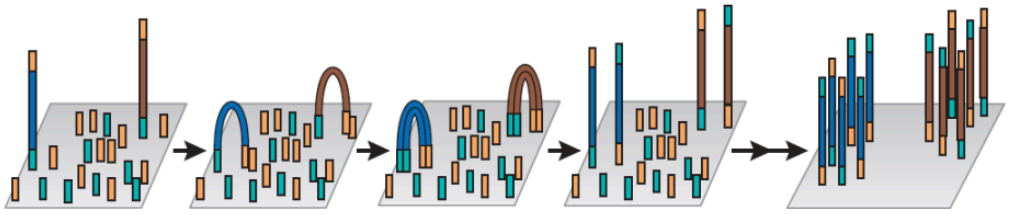


The release of hydrogen ions indicate if zero, one or more nucleotides were incorporated.

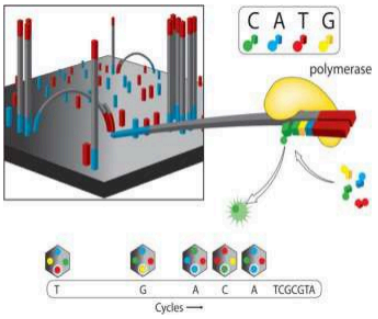


Released hydrogens ions are detected by an ion sensor. Multiple incorporations lead to a corresponding number of released hydrogens and intensity of signal.

4b. Clonal amplification by bridge PCR



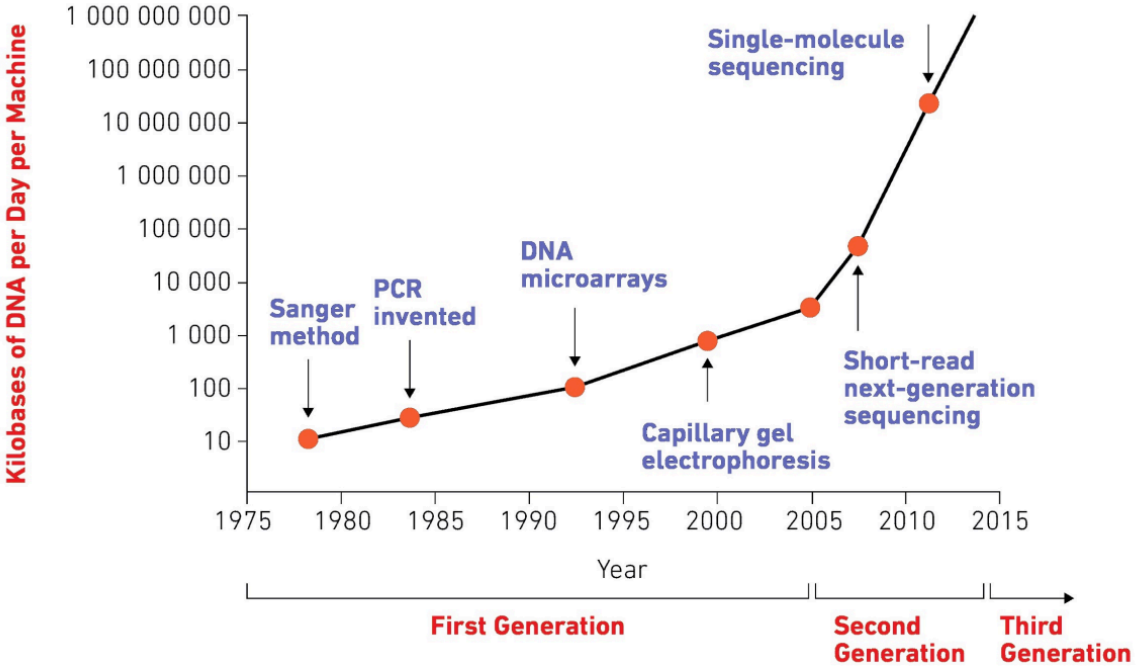
- DNA fragments are flanked with adaptors (**library**)
- A solid surface is coated with primers complementary to the two adaptor sequences
- Isothermal amplification, with one end of each "bridge" tethered to the surface
- **Clusters** of DNA molecules are generated on the chip. Each cluster is originated from a single DNA fragment, and is thus a clonal population.



Sanger Sequencing vs. NGS:

https://www.youtube.com/watch?v=jhGm37Wx8cQ

Sanger Sequencing: Best for small DNA fragments or single gene sequencing. It's more accurate but slower and more expensive for large-scale projects.
NGS: Allows sequencing of whole genomes or many genes at once, much faster and more cost-effective for large projects but less accurate per read.

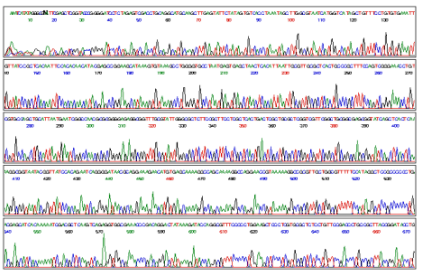


1. **Next-Generation Sequencing (NGS)**: NGS technologies, like Illumina sequencing, are more efficient, faster, and cost-effective than Sanger sequencing. They can process millions of DNA sequences simultaneously, making them popular in research and diagnostics. NGS has expanded the use of DNA sequencing in fields like medicine and forensics, though it requires strong computational tools due to short sequence lengths.

Automated DNA Sequencers

ABI 377 Plate Electrophoresis

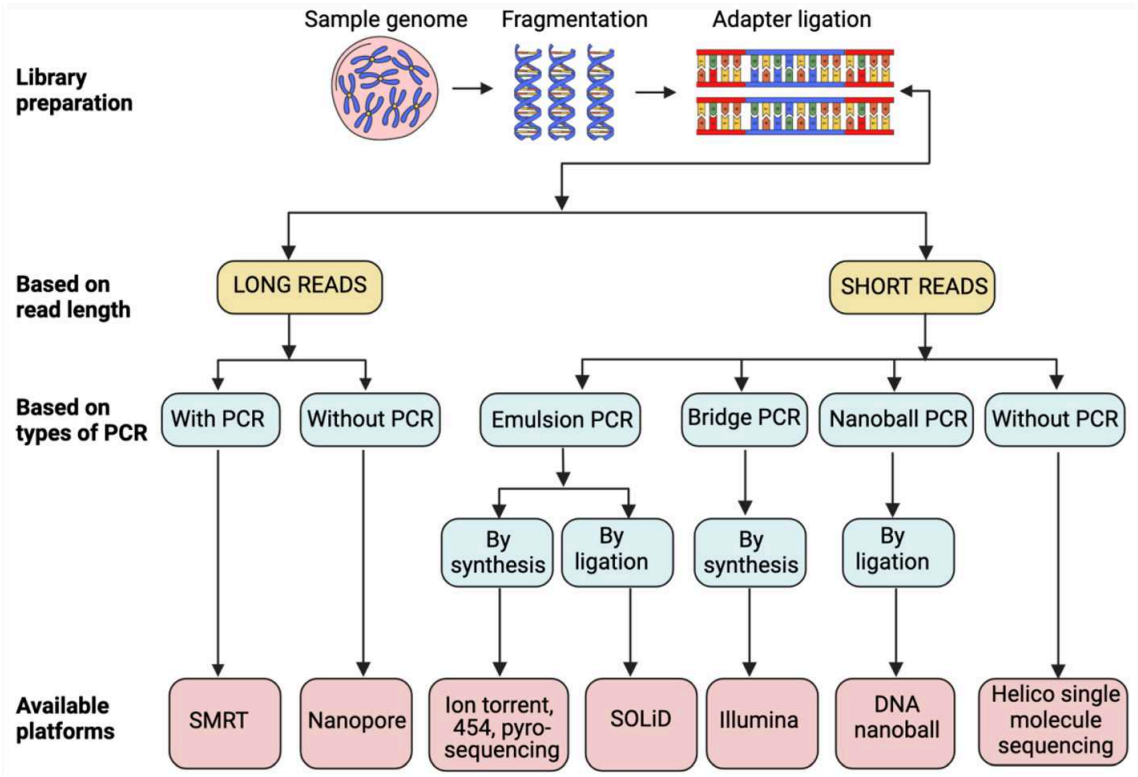
ABI 3730 xl Capillary Electrophoresis



Human Genome Project

Launched in 1989 –expected to take 15 years
Competing Celera project launched in 1998

- 1° Draft released in 2000
  - "Complete" genome released in 2003
  - Sequence of last chromosome published in 2006
- Cost: ~$3 billion
  - Celera ~$300 million



Sequencing Process:

1. **RNA Isolation**: Total RNA is extracted from the sample.

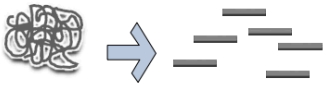
2. **Library Preparation**: RNA is enriched (e.g., ribosomal RNA depletion or poly-A selection), fragmented, and converted to cDNA for sequencing.

3. **Sequencing**: High-throughput sequencing is done using platforms like Illumina or PacBio, with long-read strategies providing full-length isoforms without assembly.

4. **Bioinformatics Analysis**: Data is processed, including quality control, alignment to reference genomes, and gene expression quantification.

NGS WORKFLOW

1. Sample extraction
2. Create DNA fragments



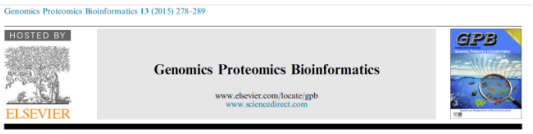
3. End repair & adapter ligation.

Add platform-specific adapter sequences to library. Adapter molecules bind every fragment to a flowcell o bead; add barcodes for multiplexing.

Third generation sequencing

The introduction of Third Generation Sequencing (TGS) circumvents the need for PCR, sequencing single molecules without prior amplification steps.

Sequence information is obtained with the use of DNA polymerase by monitoring the incorporation of fluorescently labeled nucleotides to DNA strands with single base resolution.



REVIEW

Fourth generation sequencing

Genomics Proteomics Bioinformatics 13 (2015) 4–16

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

GPB

REVIEW

Nanopore-based Fourth-generation DNA Sequencing Technology

Nanopore technology requires no amplification and uses the concept of single molecule sequencing but with the integration of tiny biopores of nanoscale diameter (nanopores) through which the single molecule passes and is identified.

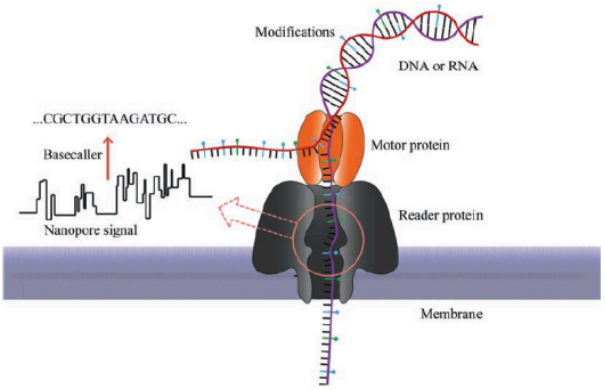
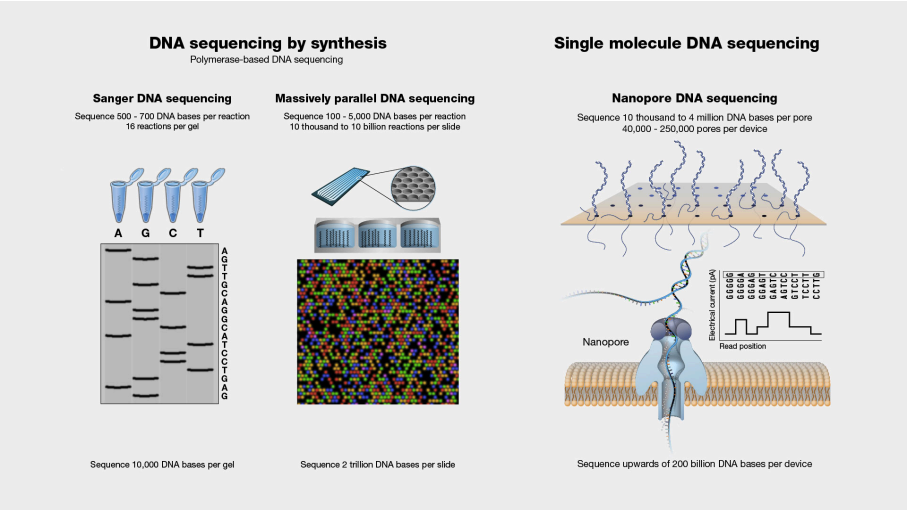
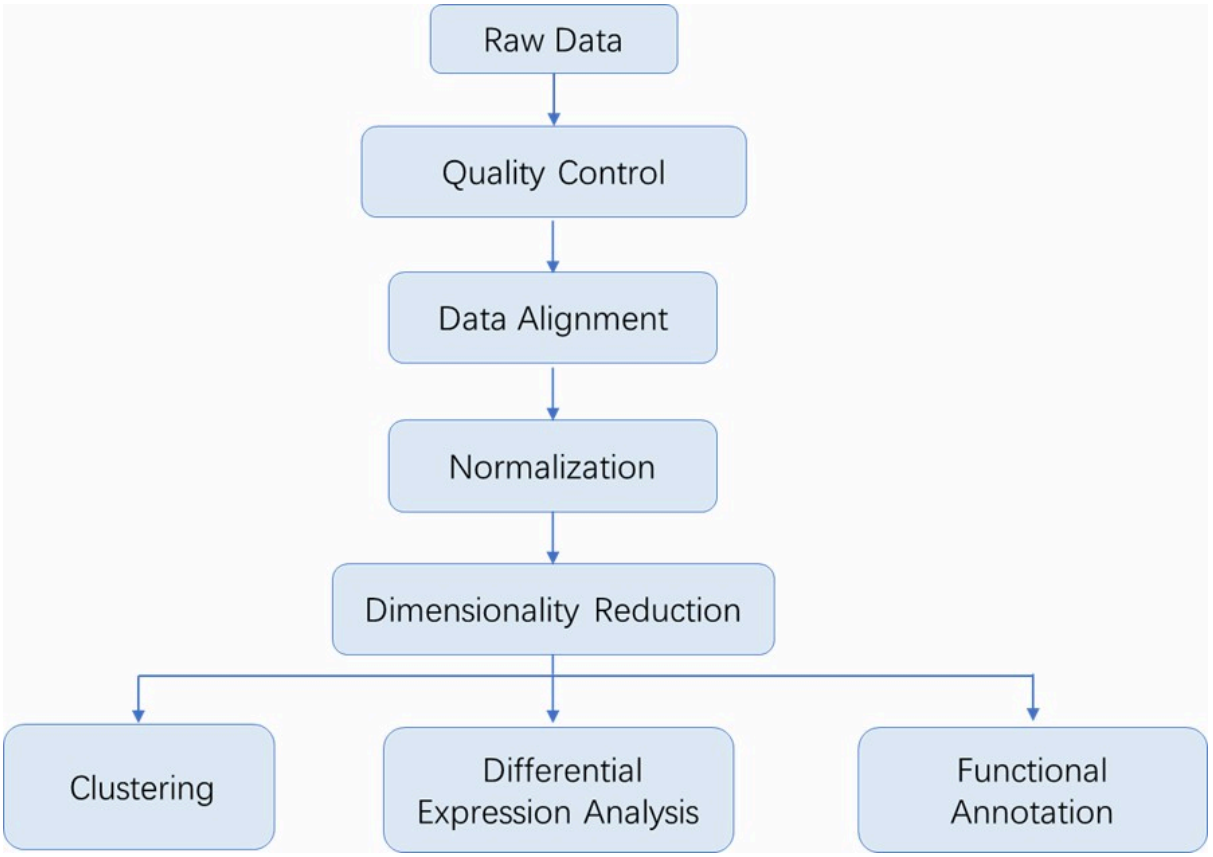


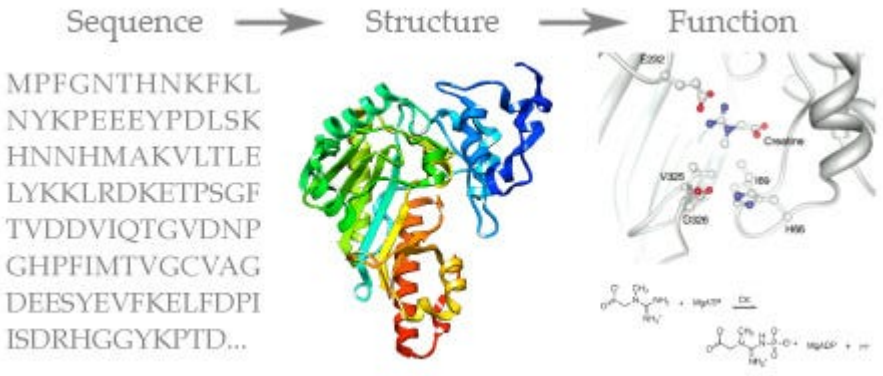
Fig. 1 Scheme for nanopore sequencing. The motor protein guides the DNA or RNA strand through the nanoscale pore provided by the reader protein. Passage of the nucleic acid molecules through a nanopore causes fluctuations of the current across the membrane. The basecaller converts the nanopore signal into the corresponding nucleic acid sequence

2. **Single-Molecule Sequencing**: This newer method, including Single-Molecule Real-Time (SMRT) and nanopore sequencing, excels at reading longer DNA sequences without needing amplification. It's becoming more popular for its ability to read long stretches of DNA.

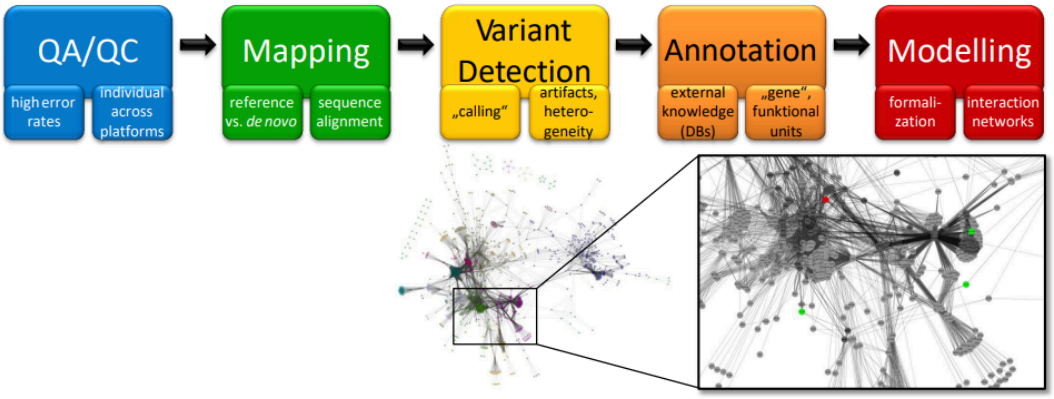


These sequencing technologies help scientists explore genetic links to diseases, traits, and biological evolution.



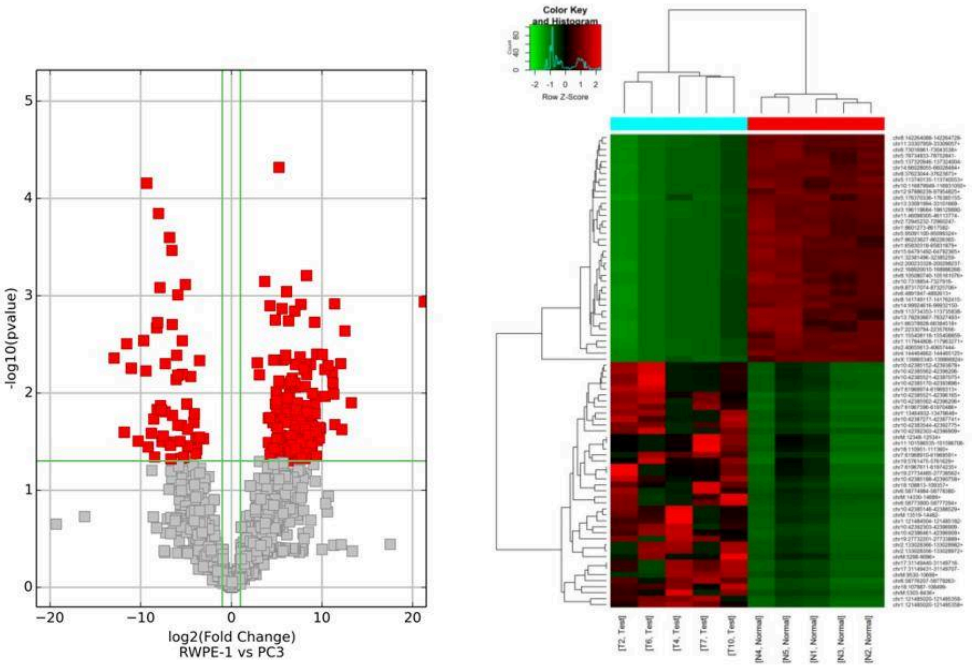


- NGS is just a technology generating data
- Scientists need assays in order to get from questions to answers
- Great variety of problems, scientific fields, target molecules, biological mechanism etc. determine the assay and the data analysis
- General scheme:



Data Analysis:

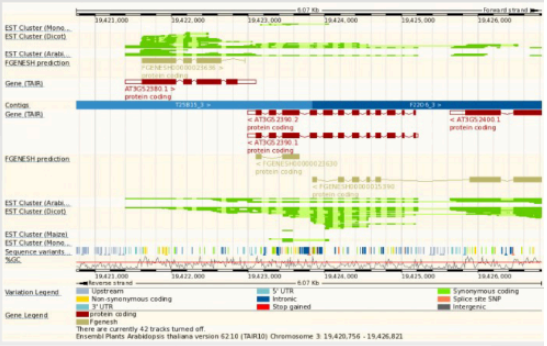
- **Differential Expression**: Identifies significant changes in mRNA, circular RNA, and long non-coding RNA with statistical thresholds (Fold Change ≥ 2, P-value ≤ 0.05).

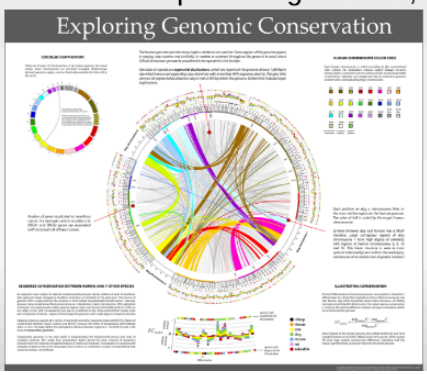


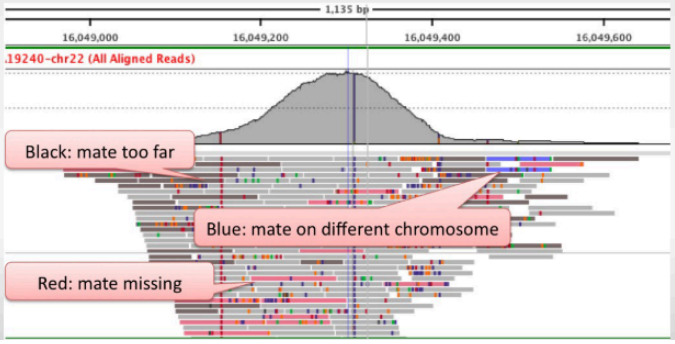
Certain circRNA/LncRNA molecules can regulate the expression of miRNA target genes by binding to miRNAs. Through the analysis of circRNA/LncRNA-miRNA-mRNA associations, we can assist in deducing the circRNA/LncRNA molecules acting as miRNA sponges and their mechanisms of action.

https://www.youtube.com/watch?v=CZeN-IgjYCo

- Main of the applications come from ***data analysis***:
  - Genome browsing and annotation
  - Exploration, interpretation and manipulation of data
  - *de novo* sequencing assembly
  - NGS ***read alignments*** and ***variation*** data visualization
  - Comparative genomics, ...







Advanced Computing Technologies

Big data analysis and NoSQL databases

- **Apache Hadoop (http://hadoop.apache.org/)** is currently *de facto* standard for ***big data* processing and analysis:**
  - **Core**: HDFS, MapReduce, HBase
  - **Spark**: SparkML, SparkR
- **NoSQL databases**, four main families of **high-performance distributed and scalable** databases:
  - *Column store*: Apache Hadoop HBase ...
  - *Document store*: MongoDB, Solr, ...
  - *Key-Value*: Redis, ...
  - *Graph*: Neo4J, ...
- New solutions for PB scale **interactive analysis**:
  - *Google Dremel* (Google BigQuery) and similar implementations: new *Hive, Cloudera Impala*
  - Nested data, and comma and tab-separated data, **SQL queries allowed**

HPG Suite

NGS pipeline, a HPC and *big data* implementation

NGS sequencer

Fastq file, up to hundreds of GB per run

QC and preprocessing

QC stats, filtering and preprocessing options

HPG Aligner, short read aligner

Double mapping strategy:
Burrows-Wheeler Transform (*GPU Nvidia CUDA*)
+
Smith-Waterman (*CPU OpenMP+SSE/AVX*)

SAM/BAM file

QC and preprocessing

QC stats, filtering and preprocessing options

HPG suite
High-Performance Genomics

More info at:

http://bioinfo.cipf.es/docs/compbio/projects/hpg/doku.php

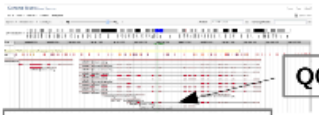
Other analysis (*HPC4Genomics consortium*)

RNA-seq (*mRNA sequenced*)
DNA assembly (*not a real analysis*)
Meth-seq
Copy Number
Transcript isoform
...

Variant calling analysis

GATK and SAM mPileup HPC Implementation.
Statistics genomic tests

VCF file



Variant VCF viewer

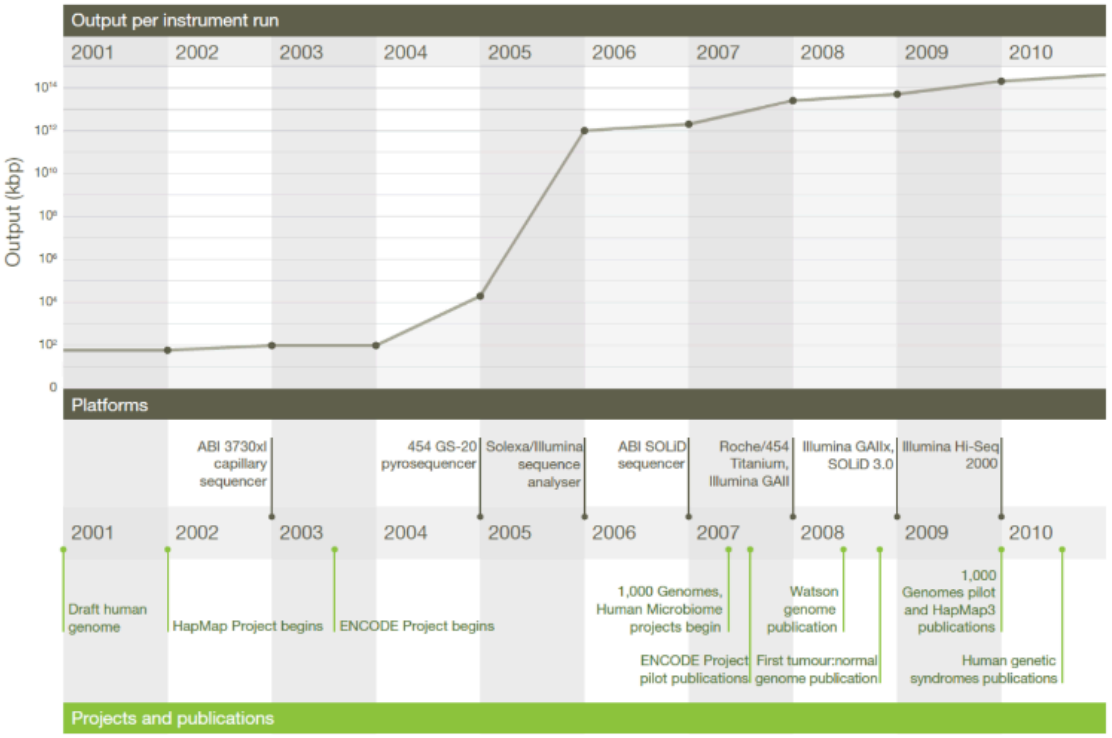
HTML5+SVG Web based viewer

QC and preprocessing

QC stats, filtering and preprocessing options

HPG Variant, Variant analysis

Consequence type, GWAS, regulatory variants and system biology information



NGS Technologies

edinburgh genomics

applied biosystems
part of life technologies

illumina

Roche 454 SEQUENCING

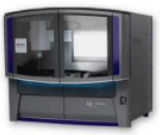
ion torrent
by life technologies

PACIFIC BIOSCIENCES

Oxford NANOPORE Technologies

Roche 454 FLX+

Illumina GAIIx



Life Tech SOLID 5500

Life Tech Ion Torrent

Helicos Heliscope

Roche 454 Junior

Illumina MiSeq

NextSeq

Illumina HiSeq

Life Tech Ion Proton

Pacific Biosciences RS

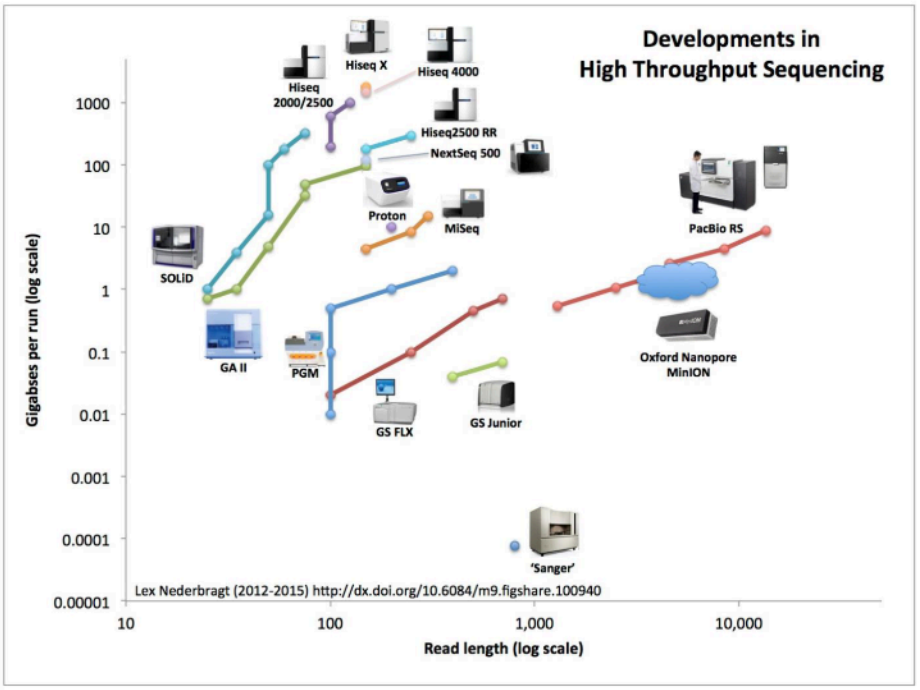
Oxford Nanopore GridIon

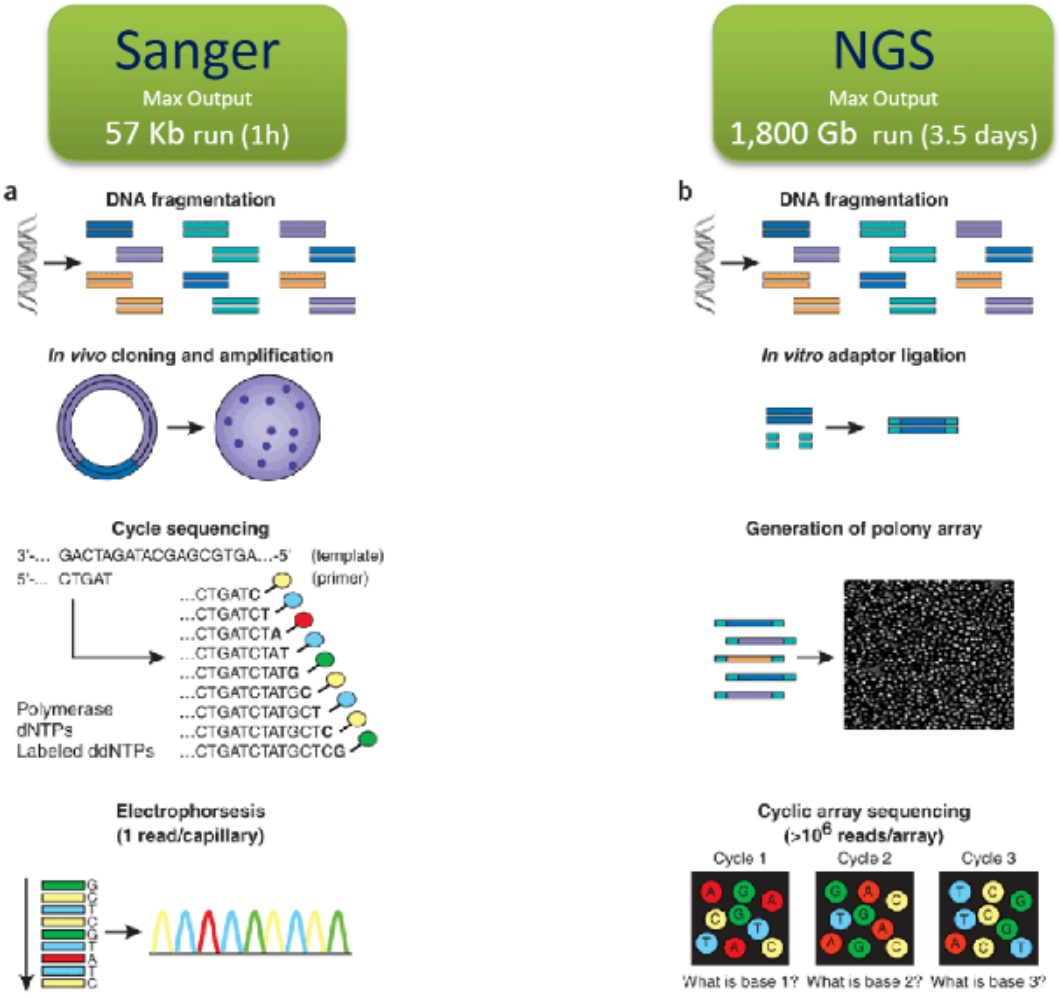
Oxford Nanopore MinION

Oxford Nanopore PromethION

Complete Genomics Revolocity

PacBio Sequel


Developments in
High Throughput Sequencing
Hiseq X
Hiseq 4000
Hiseq 2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
PacBio RS
SOLiD
GA II
PGM
GS FLX
GS Junior
Oxford Nanopore MinION
'Sanger'
Lex Nederbragt (2012-2015) http://dx.doi.org/10.6084/m9.figshare.100940
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1,000
10,000
Read length (log scale)


Sanger
Max Output
57 Kb run (1h)
NGS
Max Output
1,800 Gb run (3.5 days)
a
DNA fragmentation
In vivo cloning and amplification
Cycle sequencing
3'-... GACTAGATACGAGCGTGA...-5' (template)
5'-... CTGAT (primer)
...CTGATC
...CTGATCT
...CTGATCTA
...CTGATCTAT
...CTGATCTATG
...CTGATCTATGC
...CTGATCTATGCT
...CTGATCTATGCTC
...CTGATCTATGCTCG
Polymerase
dNTPs
Labeled ddNTPs
Electrophorsesis
(1 read/capillary)
b
DNA fragmentation
In vitro adaptor ligation
Generation of polony array
Cyclic array sequencing
(>10^6 reads/array)
Cycle 1
Cycle 2
Cycle 3
What is base 1? What is base 2? What is base 3?