Omics in human diseases

Index

- Omics data and Biological databases
- **NGS data analysis**
- Prediction and interpretation of pathogenic variants
- Protein-protein interaction networks

Course organization 2022/2023

Monday: Frontal lecture
Thursday: Frontal lecture/ guided practical activity

How to pass the exam: multiple choice quiz (50%) + results from practical activities (50%) + Bonus points, e.g. summary of previous lecture (up to 10%)

Mail: **emanuela.leonardi@unipd.it**

Methods for the interpretation of NGS and clinical data

What will we discuss…

- Next-Generation Sequencing (NGS)
- Sequencing Technologies
- NGS data analysis
- Bioinformatics resources for NGS data interpretation

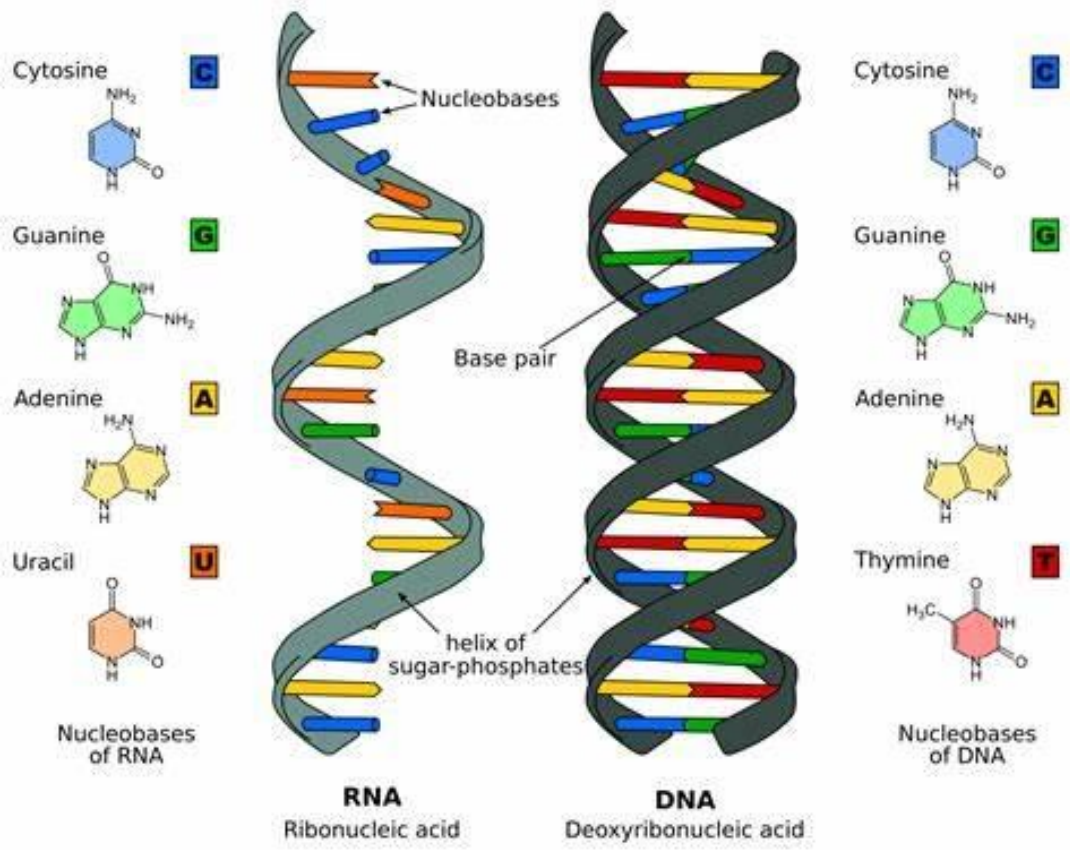
What is sequencing?

"In genetics and biochemistry, sequencing means to determine the primary structure (sometimes incorrectly called the primary sequence) of an unbranched biopolymer. Sequencing results in a symbolic linear depiction known as a sequence which succinctly summarizes much of the atomic-level structure of the sequenced molecule."

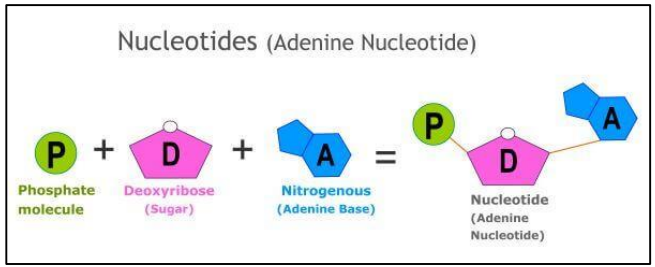
DNA sequencing is the process of determining the nucleotide order of a given DNA fragment.

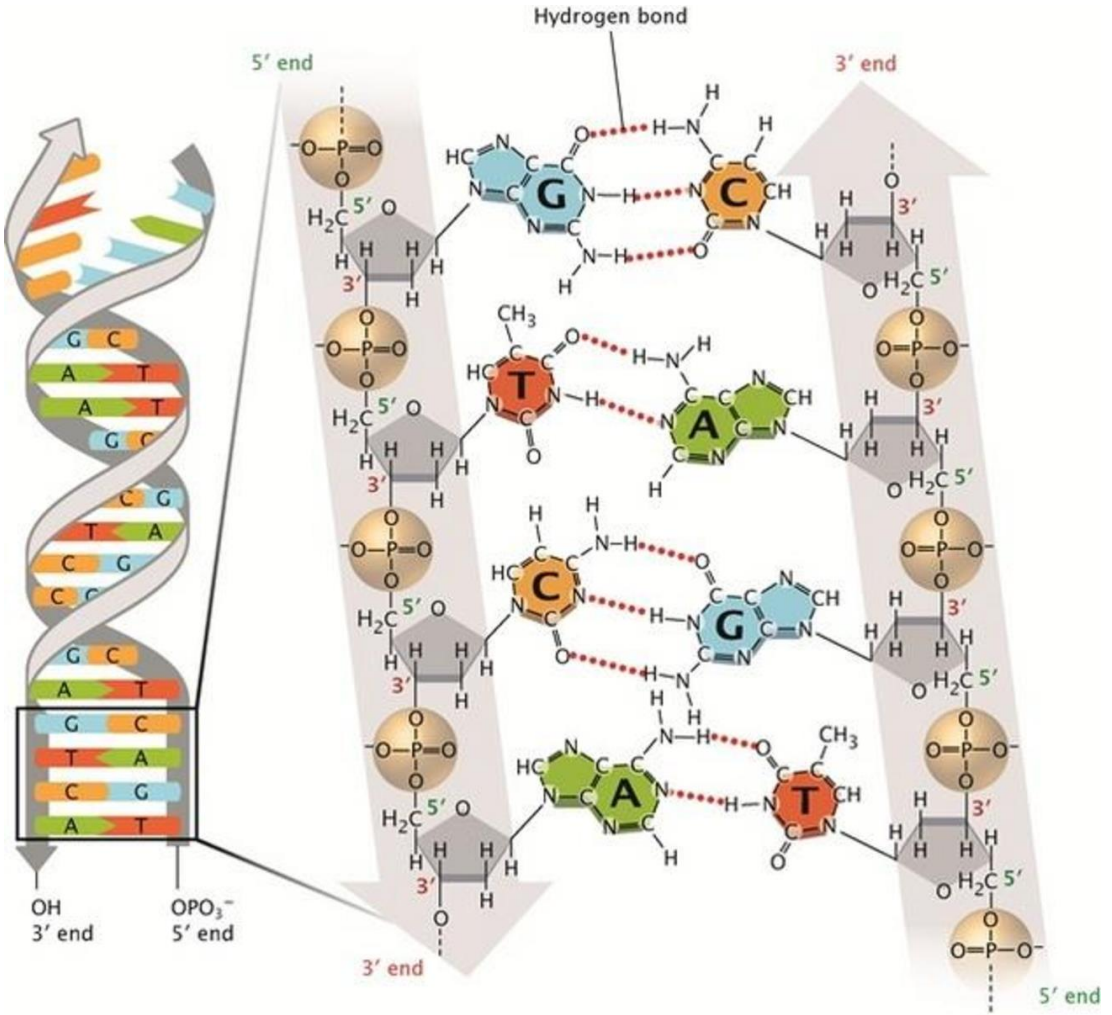
(http://en.wikipedia.org/wiki/Sequencing)

What is sequencing?



What is sequencing?





Once upon a time…

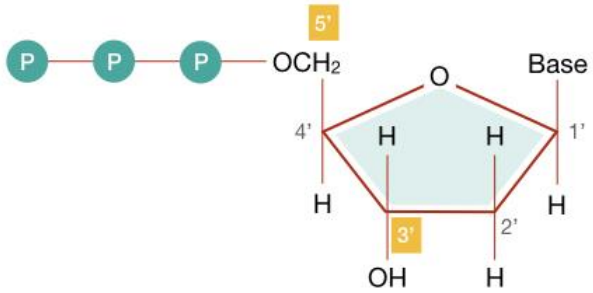
- Fredrik Sanger and Alan Coulson
- Chain Termination Sequencing (1977)
- Nobel prize 1980

Principle:

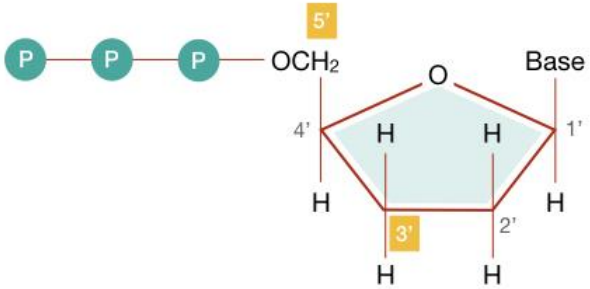
SYNTHESIS of DNA is randomly TERMINATED at different points

Separation of fragments that are 1 nucleotide different in size

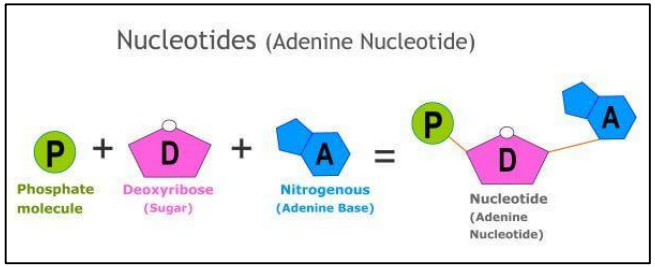
Sanger’s sequencing

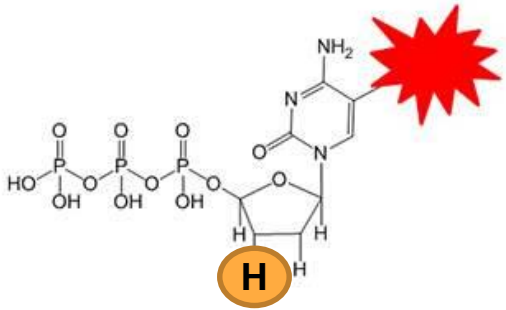


dNTP- deoxynucletide



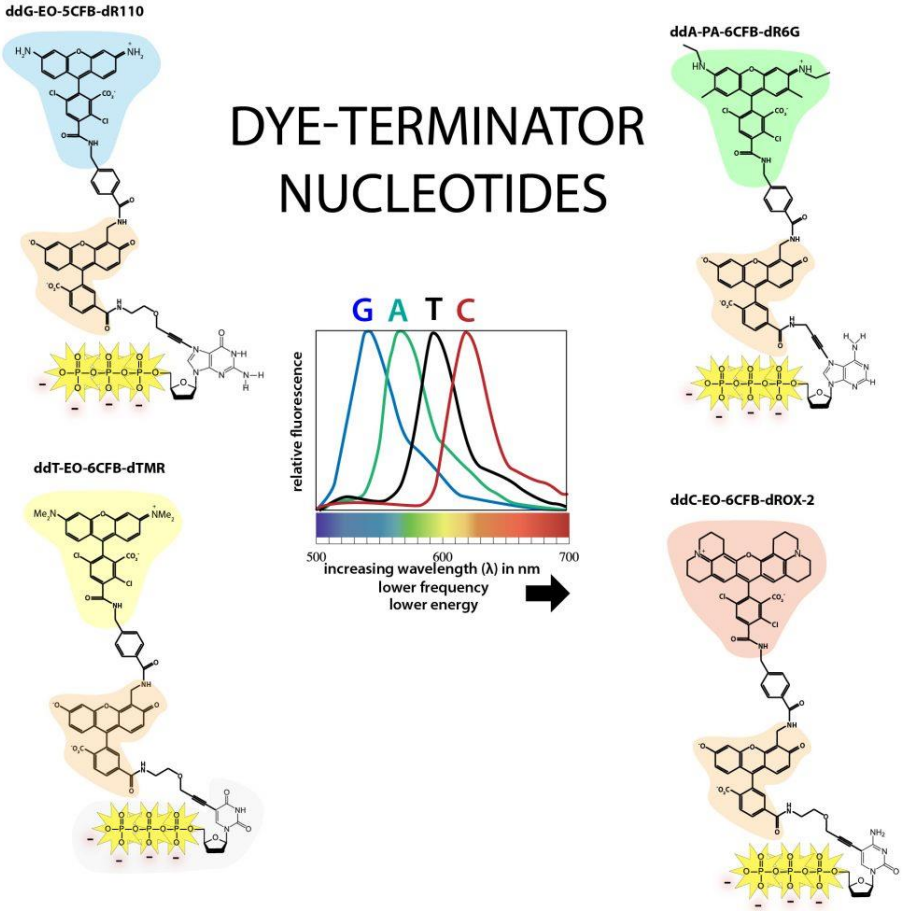
ddNTP- dideoxynucletide



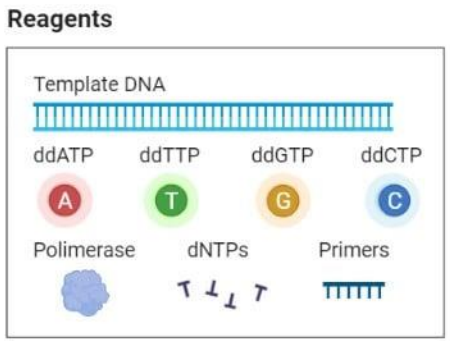


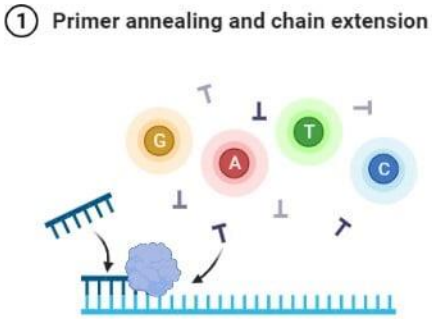
Fluorescent dye terminators

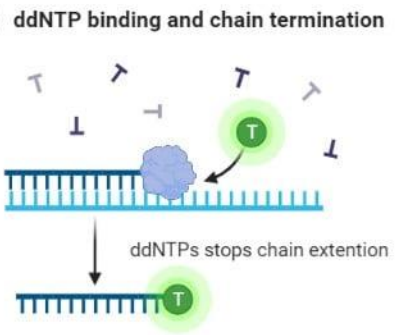
Sanger’s sequencing

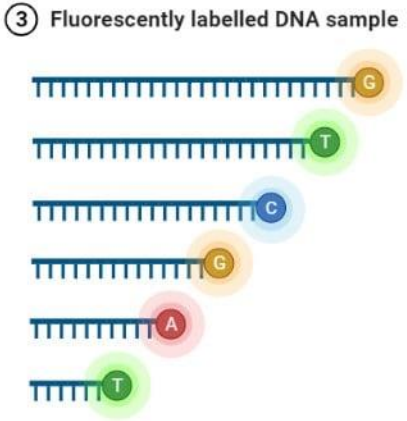


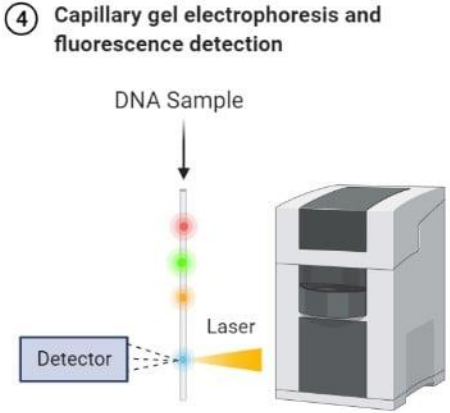
Sanger’s sequencing

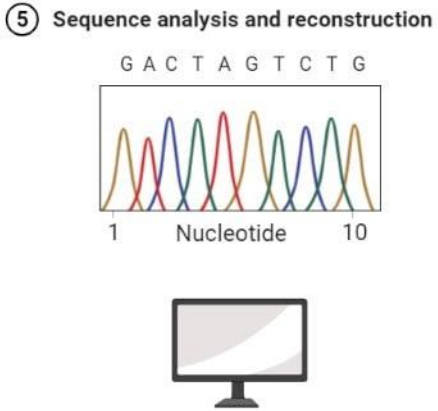












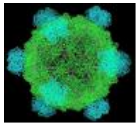
Max fragment length – 750 bp

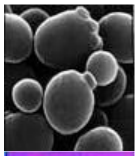
Sequencing genomes using Sanger's method

- Extract & purify genomic DNA
- Fragmentation
- Make a clone library
- Sequence clones
- Align sequences (-> contigs -> scaffolds)
- Close the gaps

Cost/Mb=1000 $, and it takes TIME

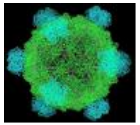
At the very beginning of genome sequencing era…

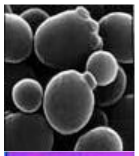




- ➔ First genome: virus X 174 - 5 368 bp (1977)
- ➔ First organism: Haemophilus influenzae - 1.5 Mb (1995)
- ➔ First eukaryote: Saccharomyces cerevisiae - 12.4 Mb (1996)
- ➔ First multicellular organism: Caenorhabditis elegans - 100 MB (1998-2002)
- ➔ First plant: Arabidopsis thaliana - 157 Mb (2000)

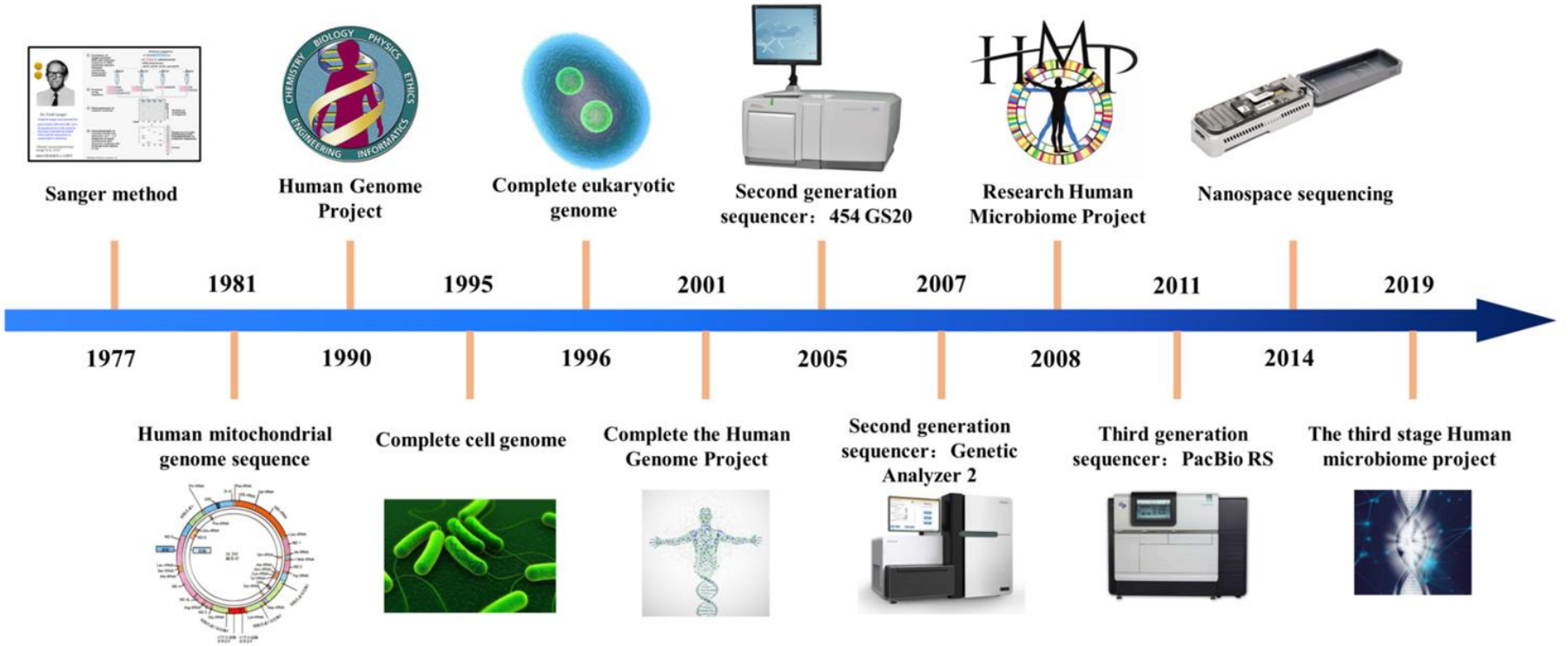
At the very beginning of genome sequencing era…





- First genome: virus X 174 - 5 368 bp (1977)
- First organism: Haemophilus influenzae - 1.5 Mb (1995) * (* .. *1995 my first year at university...*)
- First eukaryote: Saccharomyces cerevisiae - 12.4 Mb (1996)
- First multicellular organism: Caenorhabditis elegans - 100 MB (1998-2002)
- First plant: Arabidopsis thaliana - 157 Mb (2000)

Sequencing: some history



Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA (2020) Yang A., et al. Front. in Bioeng. Biotech.

Sequencing: some history

- «Rapid DNA sequencing" by Frederick Sanger (UK) in the 1970s, became the method of choice for DNA sequencing, and was worth him his 2nd Nobel Prize in chemistry in 1980
- Sanger method for sequencing DNA was used in the Human Genome Project (HGP) that produced the first reference sequence of the human genome.
- The Human Genome Project started in 1990 and was expected to take 13 years.
- A first "rough draft" was finished in 2000 and announced in a press conference by… Bill Clinton and Tony Blair!
- The complete genome was announced in 2003 (it accounted for 92% of the human genome and less than 400 gaps)

On March 31, 2022, the Telomere-to-Telomere (T2T) consortium announced that had filled in the remaining gaps and produced the **first truly complete human genome sequence**.

What's new with NGS?

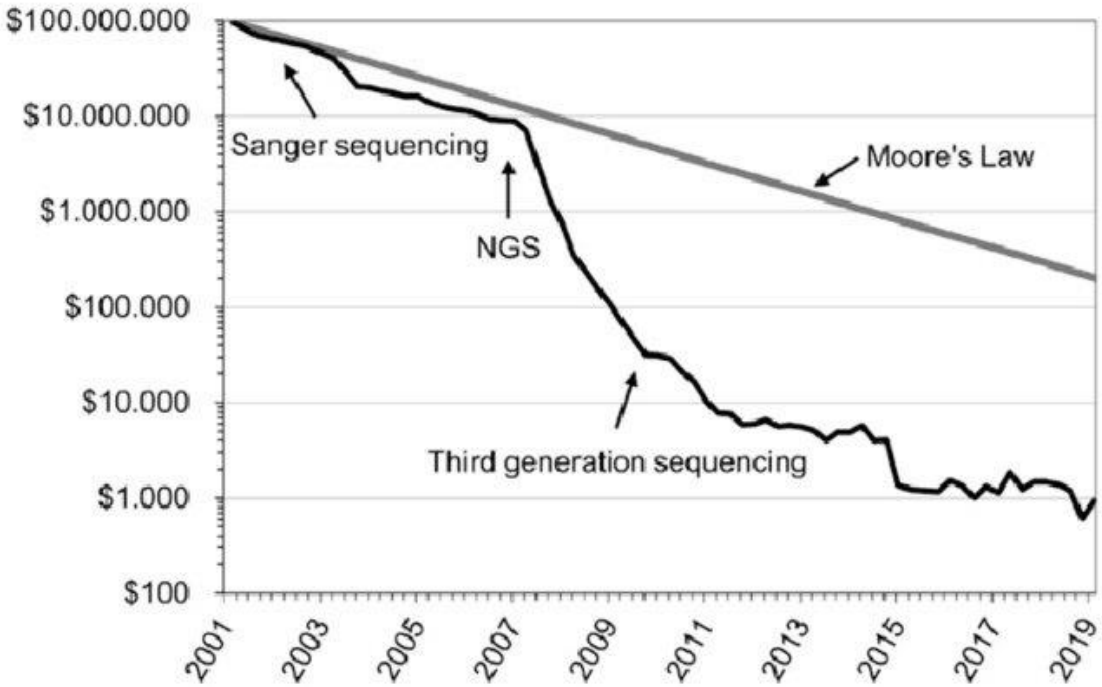
- Sequencing the whole human genome took the HGP:
  - ➔ 3.000.000.000 dollars
  - ➔ 13 years

- Sequencing a whole human genome **now** with NGS techniques takes:
  - ➔ about **1.000** dollars
  - ➔ 4-5 days

(Automatic) Sequencing is much faster and (thus) cheaper !!

https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project

NGS costs per WGS

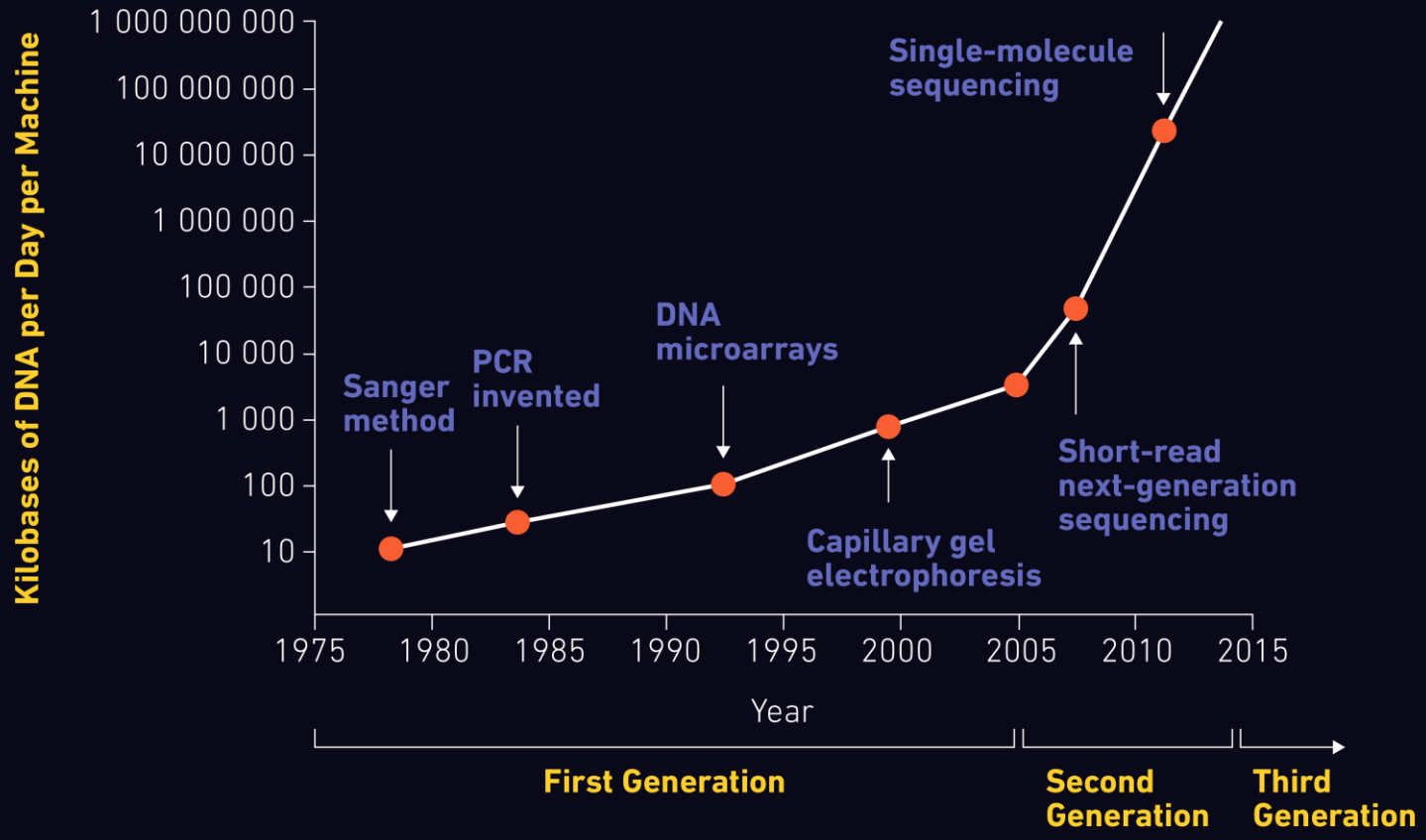


The departure of sequencing cost curve from Moore's law coincides with the emergence of next generation sequencing (NGS).

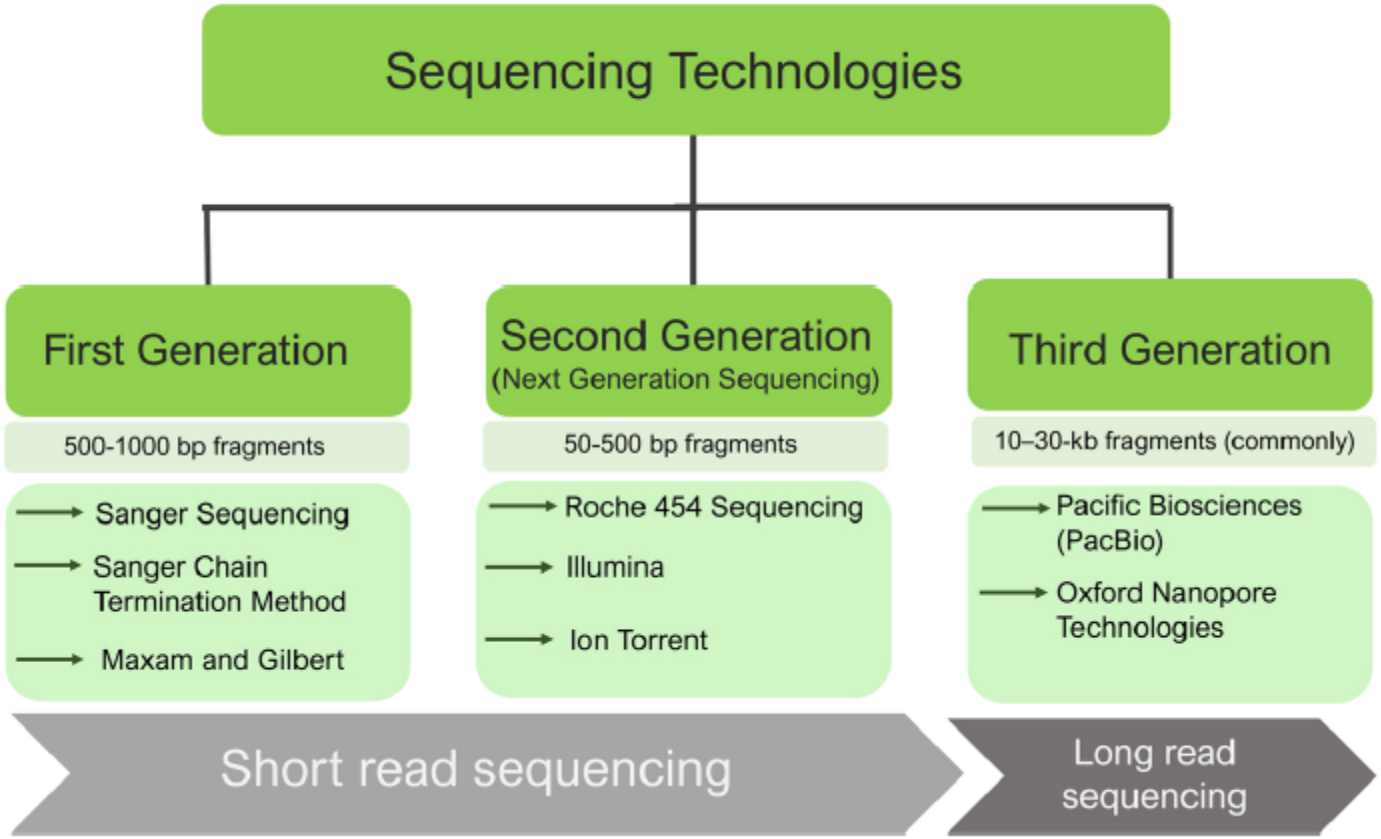
Moore's law originates in the computer hardware industry that involves doubling of 'computing power' every two years.

Technologies that follow the law are regarded as successful. It thus represents a useful relationship to compare technology advances

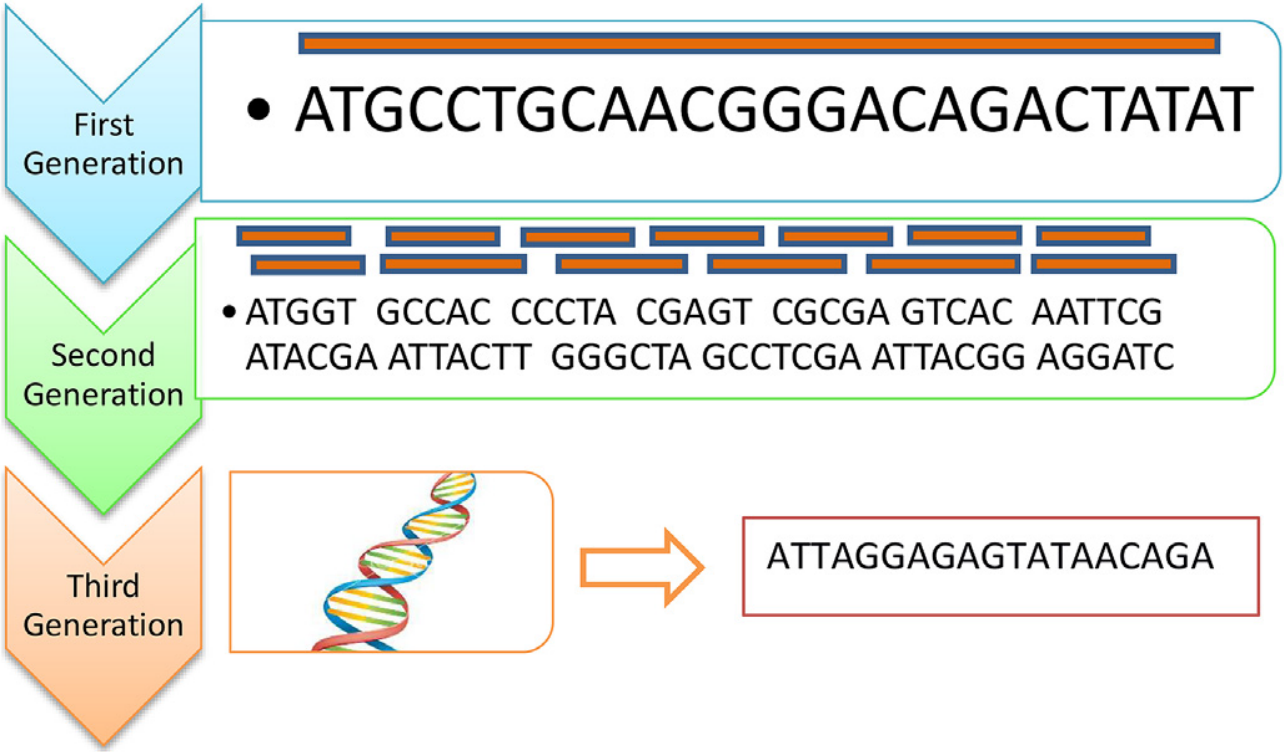
Innovations in chemistry, optics, fluidics computational hardware, and bioinformatics solutions


Kilobases of DNA per Day per Machine
1 000 000 000
100 000 000
10 000 000
1 000 000
100 000
10 000
1 000
100
10
Sanger method
PCR invented
DNA microarrays
Capillary gel electrophoresis
Short-read next-generation sequencing
Single-molecule sequencing
1975
1980
1985
1990
1995
2000
2005
2010
2015
Year
First Generation
Second Generation
Third Generation

Sequencing Technologies



Sequencing Technologies



Rucha M. Wadapurkar, 2018 Informatics in Medicine Unlocked

Differences between automatic NGS platforms

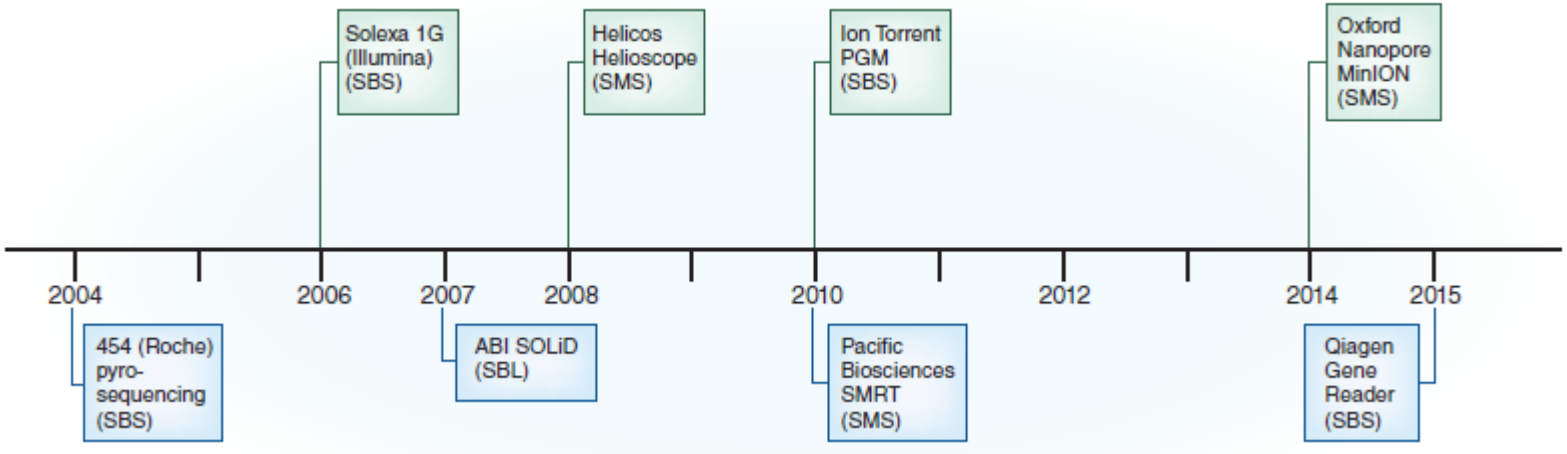
- Technology: chemistry + signal detection
- Run times vary from hours to days
- Production range from Mb to Gb
- Read length from <100 bp to > 20 Kbp
- Accuracy per base from 0.1% to 15%
- Cost per base varies

NGS technologies

Company	Platform	Amplification	Sequencing method
Roche (dismissed)	454	emPCR	Pyrosequencing
Illumina	HiSeq MiSeq	Bridge PCR	Synthesis
LifeTech	SOLiD	emPCR/ Wildfire	Ligation
LifeTech	Ion Torrent, S5 Studio	emPCR	Synthesis (pH)
Pacific Bioscience	RSII	None	Synthesis
Complete genomics	Nanoballs	None	Ligation
Oxford Nanopore	GridION	None	Flow

NGS: by synthesis and by ligation

The first NGS platforms were instruments performing sequencing by synthesis (SBS)
Commercially available SBS platform (454, Illumina, Ion Torrent and GeneReader) uses a **UNIQUE** combination of **sequencing-reaction chemistry** and **detection modality** applied in a stepwise manner to produce sequencing data

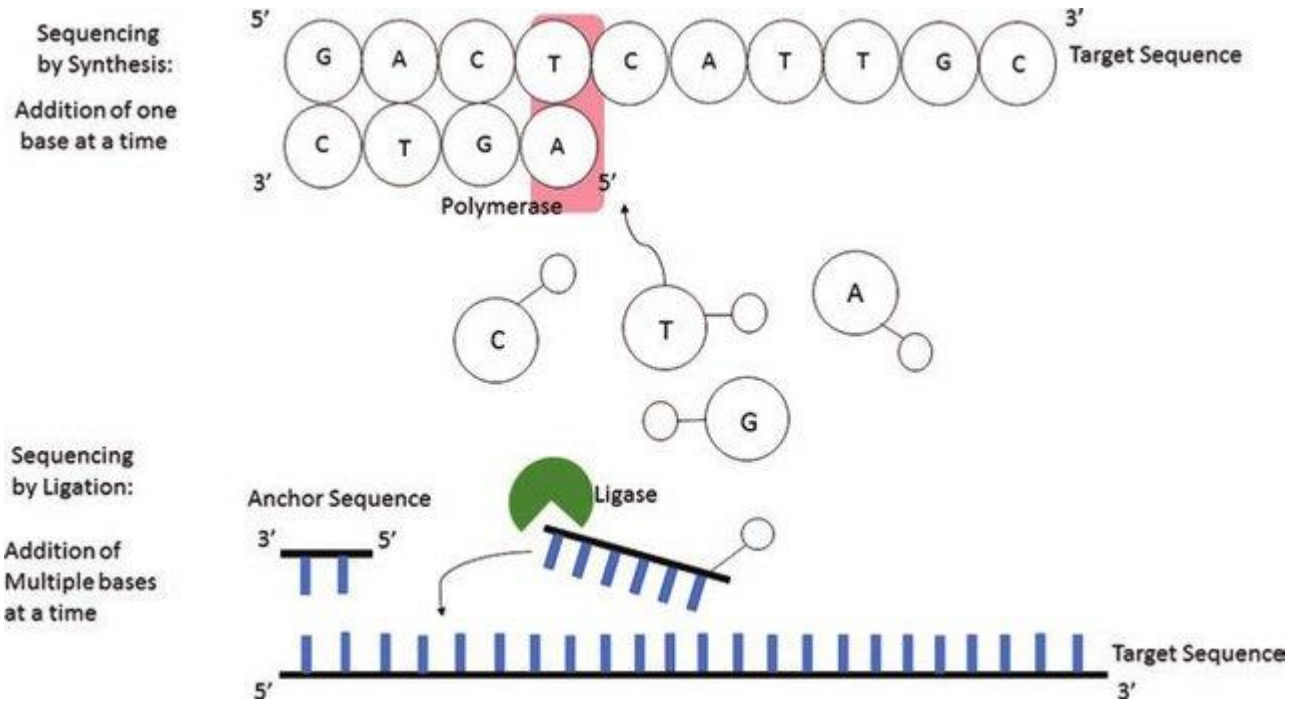


SBS, sequencing by synthesis; SMS, single-molecule sequencing; SBL, sequencing by ligation.

Because each sequencing read originates from a single library fragment, SBS data are digital, and the resulting ability to quantify read data from different preparatory inputs is one important advantage of SBS

Elaine R Mardis, 2016, Nature Protocols

Sequencing: by synthesis and by ligation



Sequencing by ligation: Instead of creating a second strand as in sequencing-by-synthesis approach, sequencing by ligation take advantage of the **mismatch sensitivity of a DNA ligase** to determine the underlying sequence of a target DNA molecule

NGS workflow

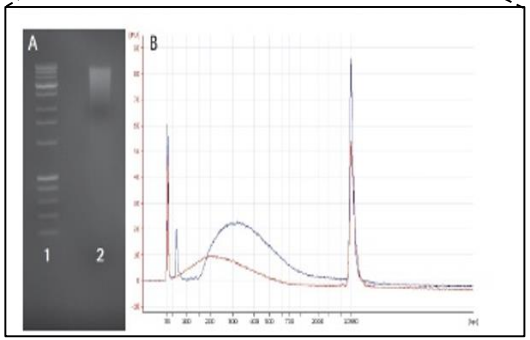
- Samples collection
- Nucleic acid extraction
- Quality Control

- Adaptor ligation/barcoding
- Size selection
- Amplification/purification
- Quality control

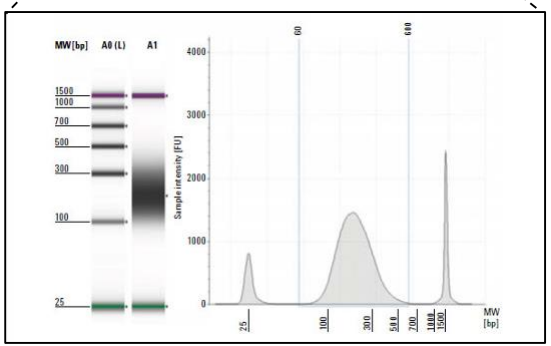
- Sequencing
- Data analysis
  - Base calling
  - Read alignment
  - Variant calling
  - Variant annotation

Making a NGS library

Ligation of sequencing adaptors, technology specific



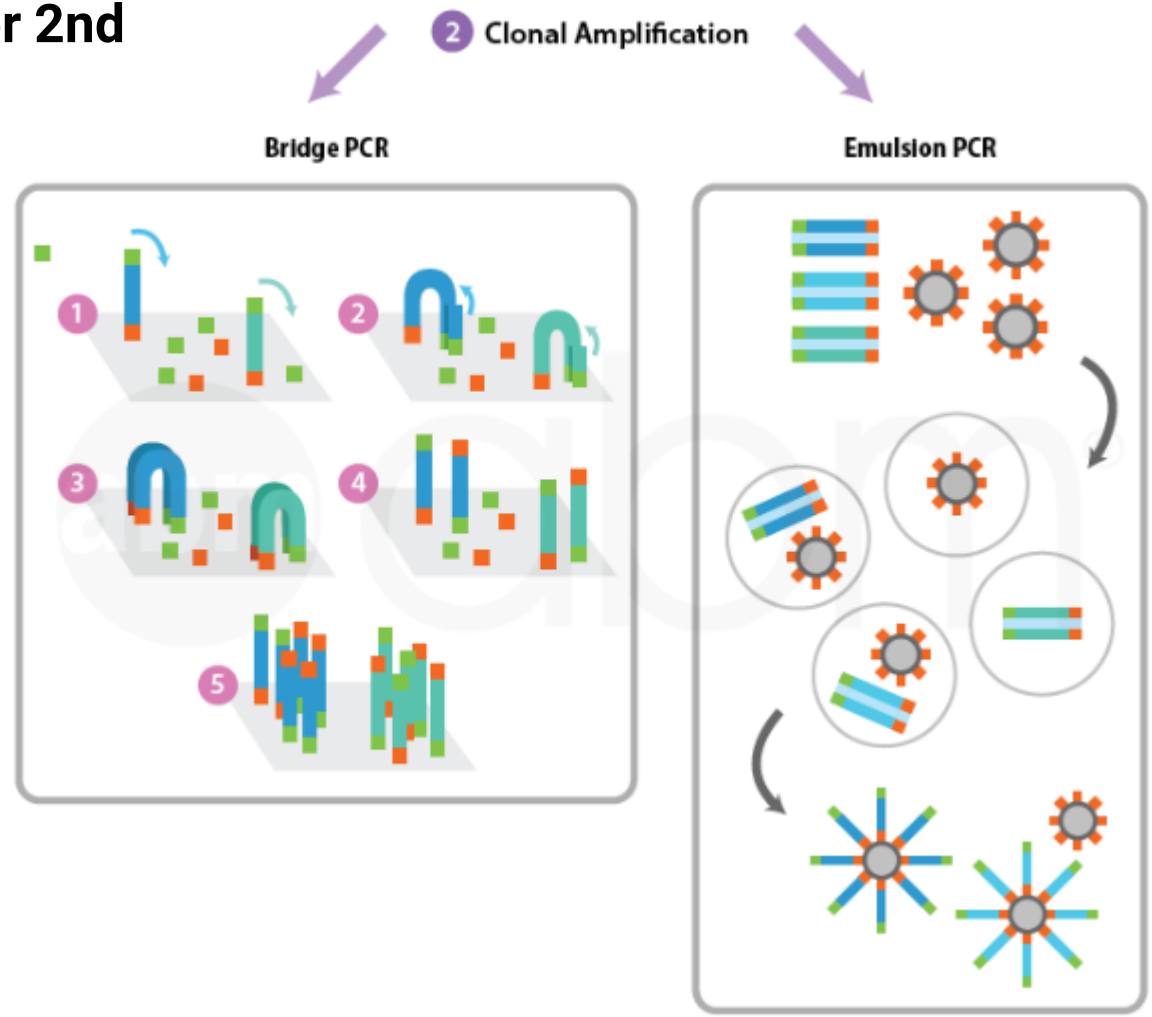
DNA QC – paramount importance



Sharing & size selection

Amplification step required for 2nd generation Sequencing

- Each library fragment is amplified *in situ* on a solid surface (either **beads** or a **flat silicon derived surface**)
- **Adapters** allow for library hybridization to the sequencing chips and provide a universal priming site for sequencing primers
- The amplification creates **clusters of DNA**, each originating from a single library fragment, **needed to produce sufficient signal strength**
- Each cluster will act as an **individual sequencing reaction**.

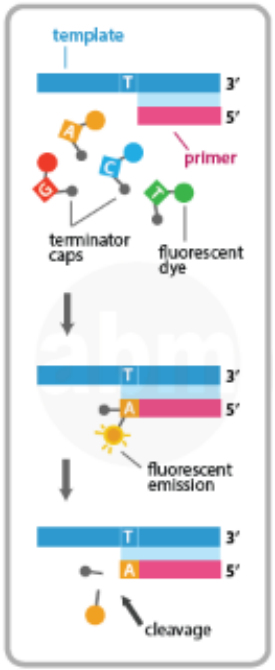


2 Clonal amplification

Bridge PCR

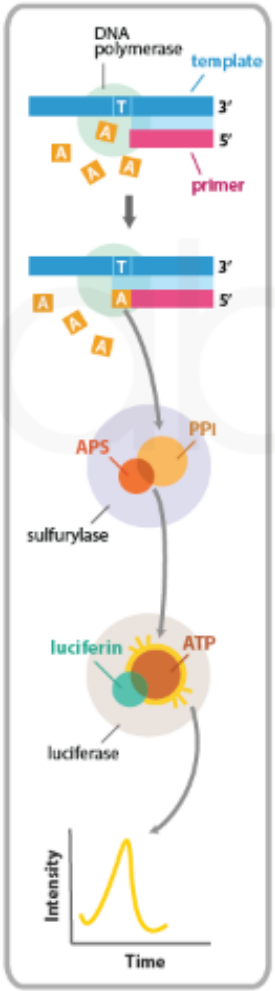
Emulsion PCR

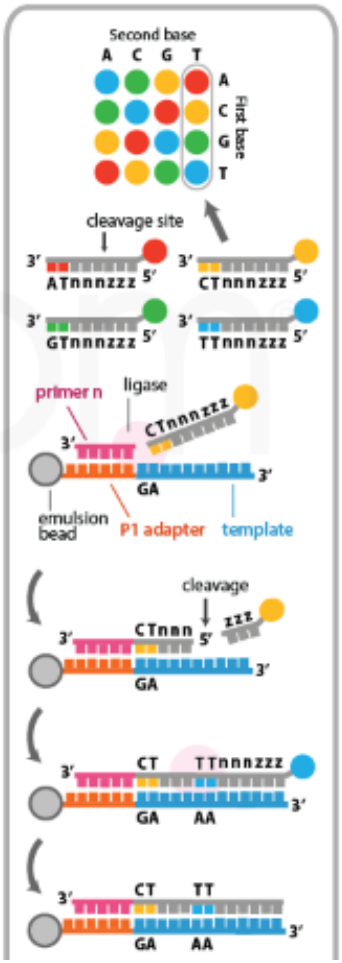
Terminator-Seq



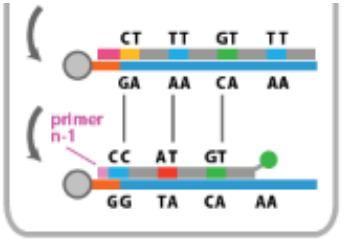
ILLUMINA

454-Roche





ABI-SOLiD



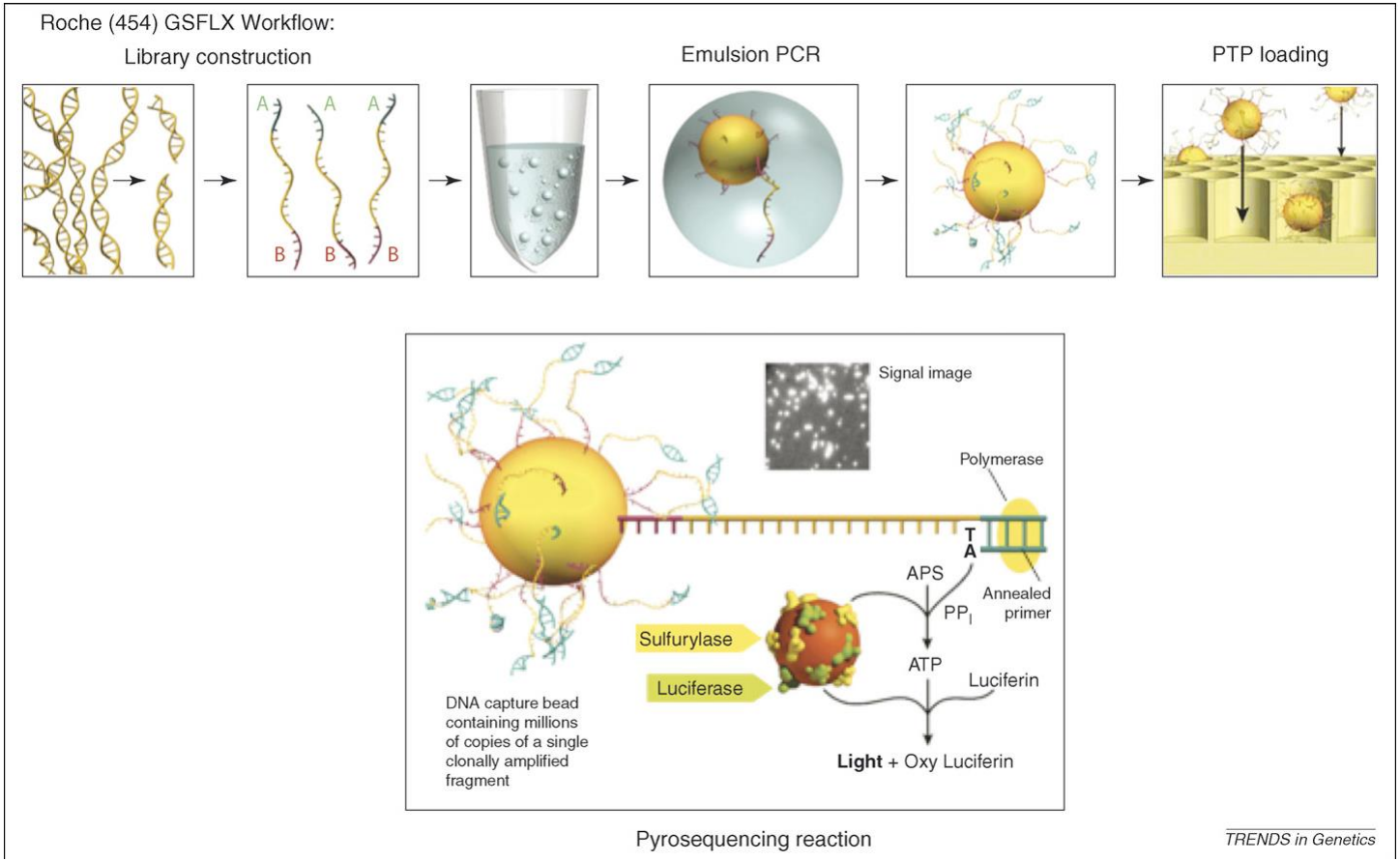
NGS Technology - Roche - Pyrosequencing

Instrument	Yield and run time	Read Length	Error rate	Error type
454 FLX+	0.9 GB, 20 hrs	700	1%	Indels
454 FLX Titanium	0.5 GB, 10 hrs	450	1%	Indels
454 FLX Jr	0.050 GB, 10 hrs	400	1%	Indels

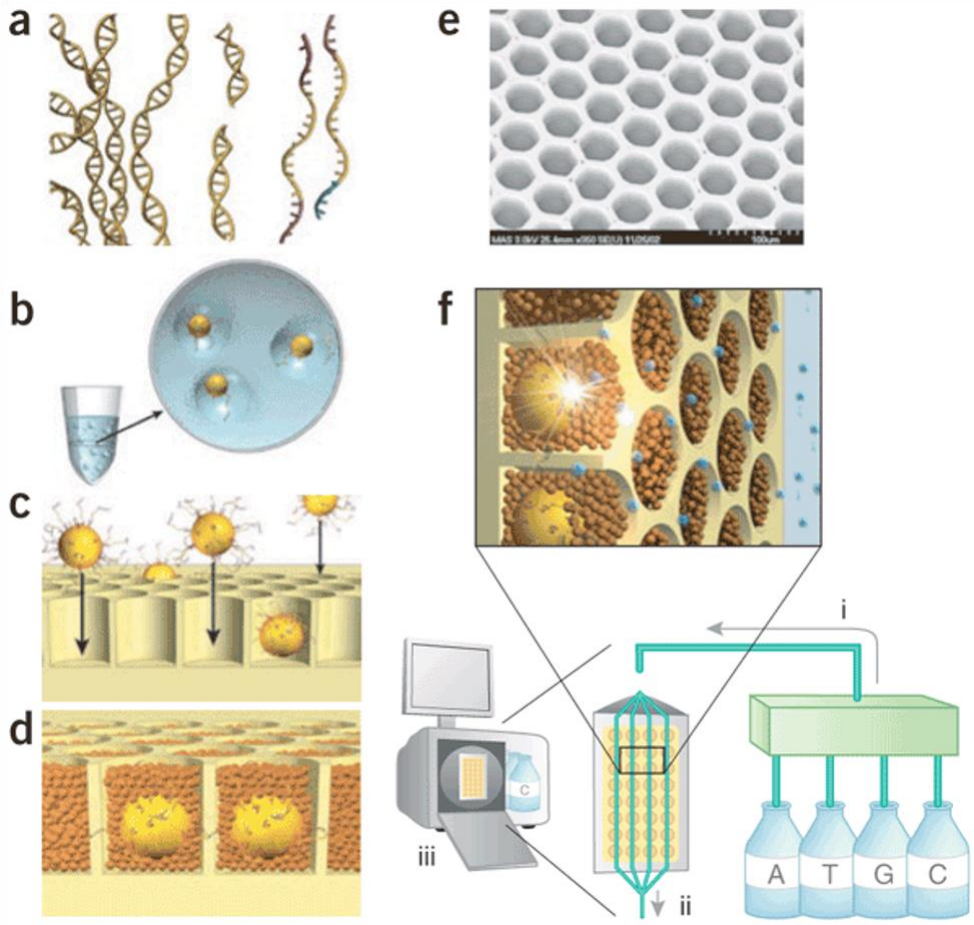
Main applications:

- Microbial genomics and metagenomics
- Targeted resequencing

454 – Roche workflow - Pyrosequencing



ROCHE - 454 Titanium GS FLX



Pyrosequencing

Is based on the "sequencing by synthesis" principle

Detects the nucleotide incorporated by a DNA polymerase.

Relies on light detection based on a chain reaction when pyrophosphate (Ppi) is released.

ILLUMINA

Benchtop Sequencers

Production-Scale Sequencers

iSeq 100

MiniSeq

MiSeq Series

NextSeq 550 Series

NextSeq 1000 & 2000

Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)					
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●	●
Exome & Large Panel Sequencing (enrichment-based)				●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)				●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)				●	●
Targeted Gene Expression Profiling	●	●	●	●	●

ILLUMINA

Benchtop sequencers

	iSeq 100	MiniSeq	MiSeq Series	NextSeq 550 Series	NextSeq 1000 & 2000
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	360 Gb *
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.2 billion *
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

ILLUMINA

Benchtop Sequencers | Production-Scale Sequencers

NextSeq 1000 & 2000

NovaSeq 6000 Series

NovaSeq X Series

Popular Applications & Methods	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)		●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome & Large Panel Sequencing (enrichment-based)	●	●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	●	●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	●	●	●
Chromatin Analysis (ATAC-Seq, ChIP-Seq)	●	●	●
Methylation Sequencing	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)	●	●	●

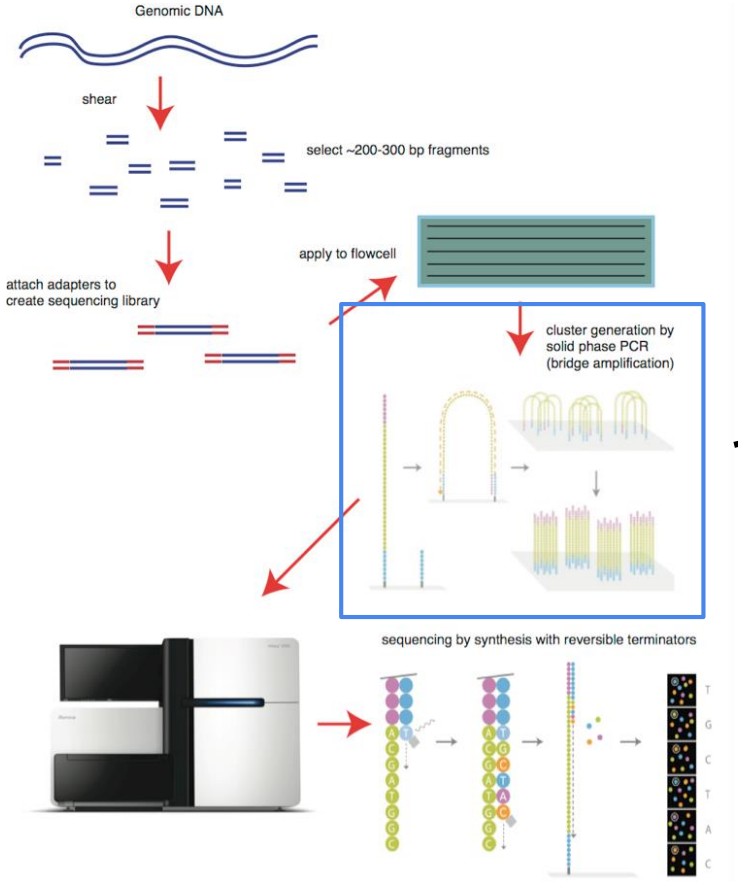
ILLUMINA

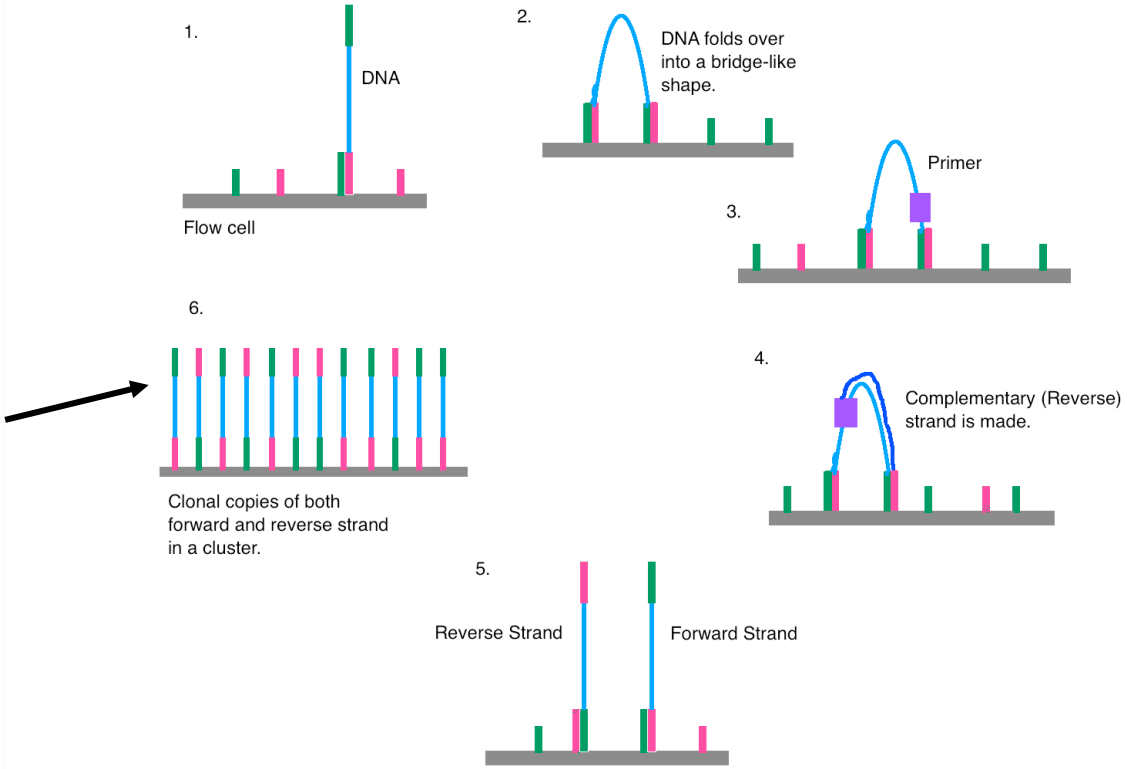
Production scale sequencers

	NextSeq 1000 & 2000	NovaSeq 6000 Series	NovaSeq X Series
Run Time	11-48 hours	~13–38 hours (dual SP flow cells)	~13–21 hours (1.5B flow cells‡)
		~13–25 hours (dual S1 flow cells)	~18–24 hours (10B flow cells‡)
		~16–36 hours (dual S2 flow cells)	~48 hours (25B flow cells‡)
		~44 hours (dual S4 flow cells)	
Maximum Output	360 Gb *	6000 Gb	16 Tb
Maximum Reads Per Run	1.2 billion *	20 billion	26 billion (single flow cells)
			52 billion (dual flow cells)
Maximum Read Length	2 × 150 bp	2 x 250 bp**	2 × 150 bp

Illumina sequencing protocol

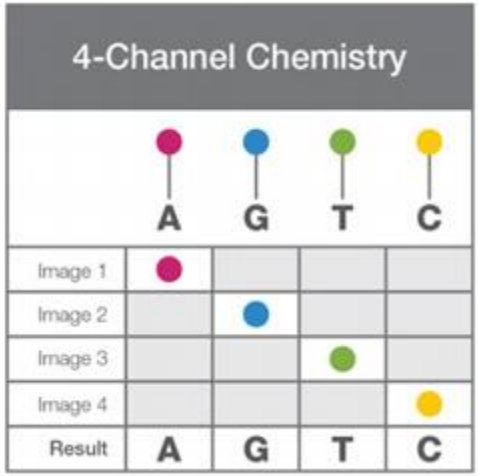
https://youtu.be/fCd6B5HRaZ8





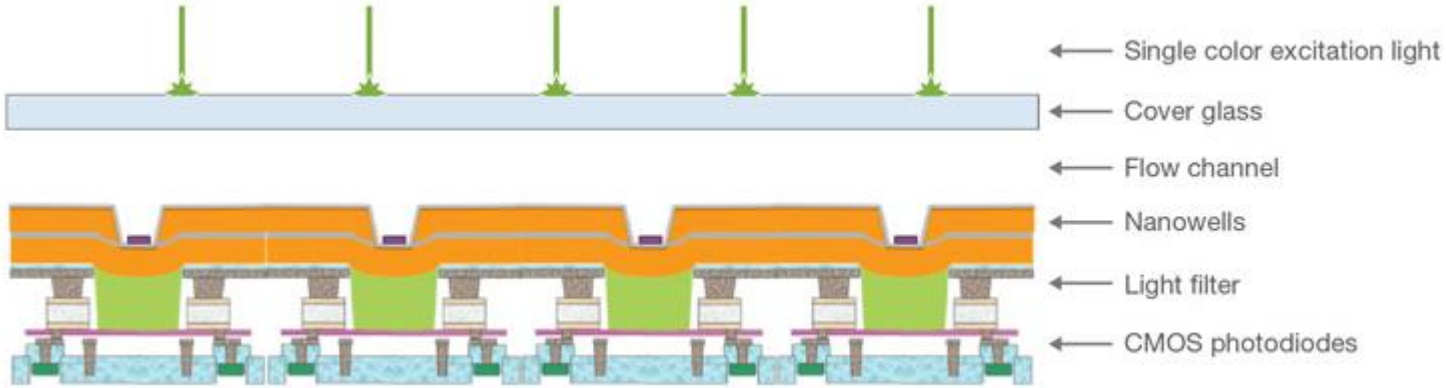
Illumina Sequencing chemistry

- Sequencing by synthesis (SBS), detects single bases as they are incorporated into growing DNA strands.
- A fluorescently labeled reversible terminator is imaged as each dNTP is added, and then cleaved to allow incorporation of the next base.
- Since all 4 reversible terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias.
- The method virtually eliminates errors and missed calls associated with strings of repeated nucleotides (homopolymers).



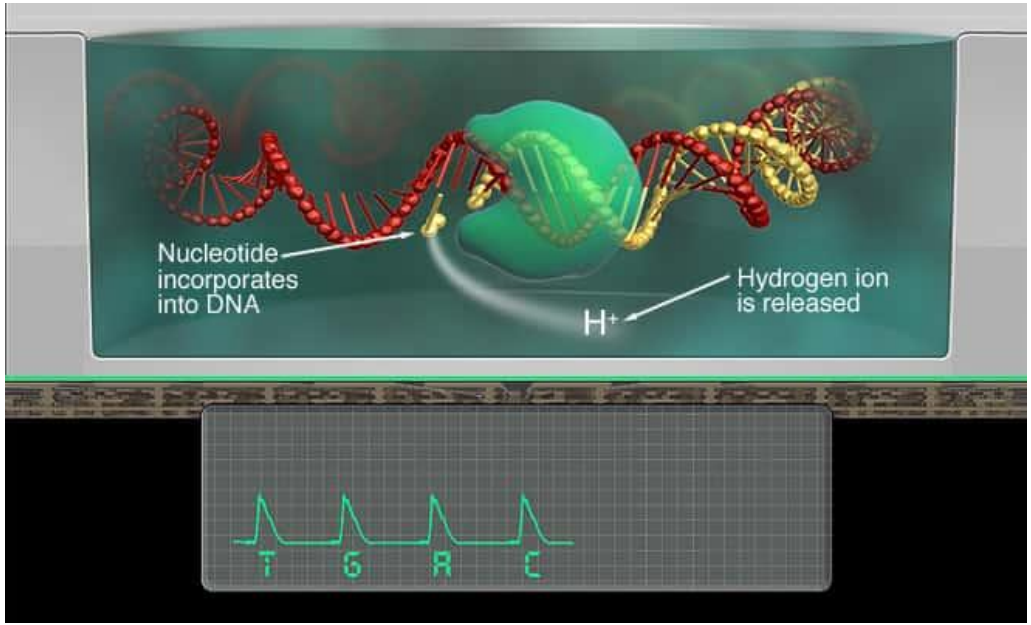
Illumina CMOS Flow Cell

https://youtu.be/pfZp5Vgsbw0



The iSeq 100 System uses a patterned flow cell with nanowells fabricated over a CMOS chip. Each well in the flow cell is aligned over a CMOS photodiode. During cluster generation, distinct DNA fragments are amplified in each well, resulting in clonal clusters. During each imaging step, the CMOS sensor detects light emissions at the bottom of each well.

Thermo Fisher Scientific - Ion Torrent



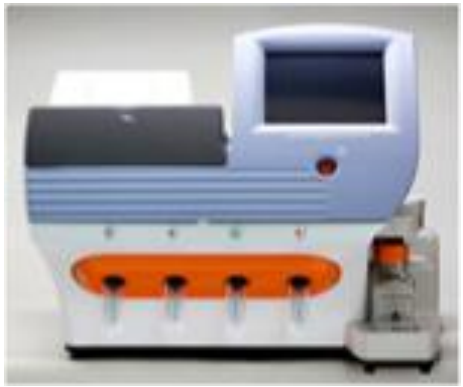
Ion Torrent™ technology directly translates chemically encoded information (A, C, G, T) into digital information (0, 1) on a semiconductor chip.

Because this is direct detection—no scanning, no cameras, no light—each nucleotide incorporation is recorded in seconds.

The potential to be **cost-effective** and provide **rapid turnaround time** and greater **flexibility** in throughput

Ion Torrent – Ion GeneStudio S5

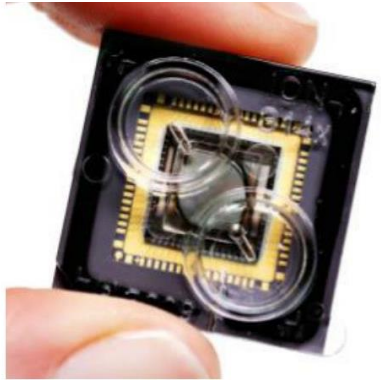
- ➔ For small scale (PGM) and large scale sequencing (Ion Gene Studio S5)
- ➔ Amplicon based library
- ➔ Rapid sequencing (run time ~ 2-4 hours)
- ➔ Measures changes in pH
- ➔ Sequencing on a chip

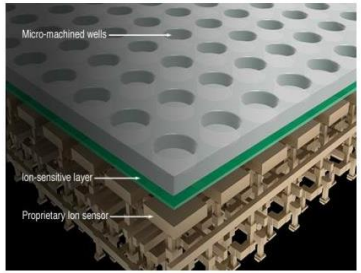


Personal Genome Machine (PGM)

Ion GeneStudio S5

Ion Torrent – Ion GeneStudio S5





- **Ion 550™ Chip** 100–130M reads
- **Ion 540™ Chip** 60–80M reads
- **Ion 530™ Chip** 15–20M reads
- **Ion 520™ Chip** 4–6M reads
- **Ion 510™ Chip** 2–3M reads

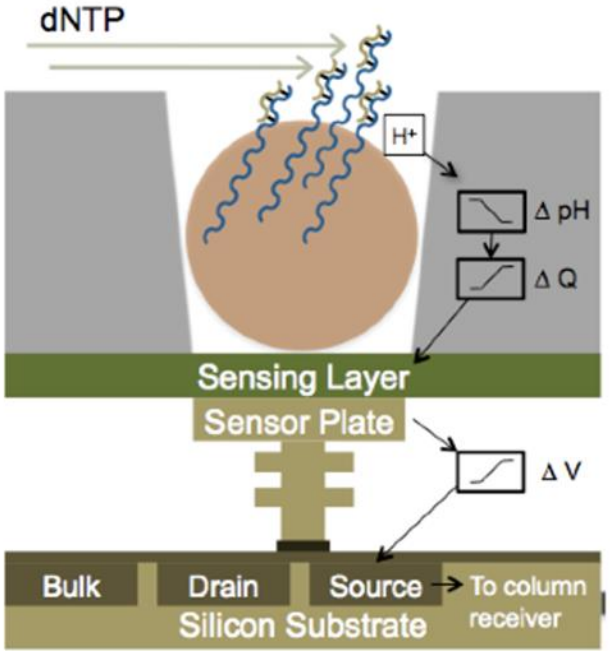
Five Ion S5™ chip options enable a sequencing throughput range of 2M to 130M reads

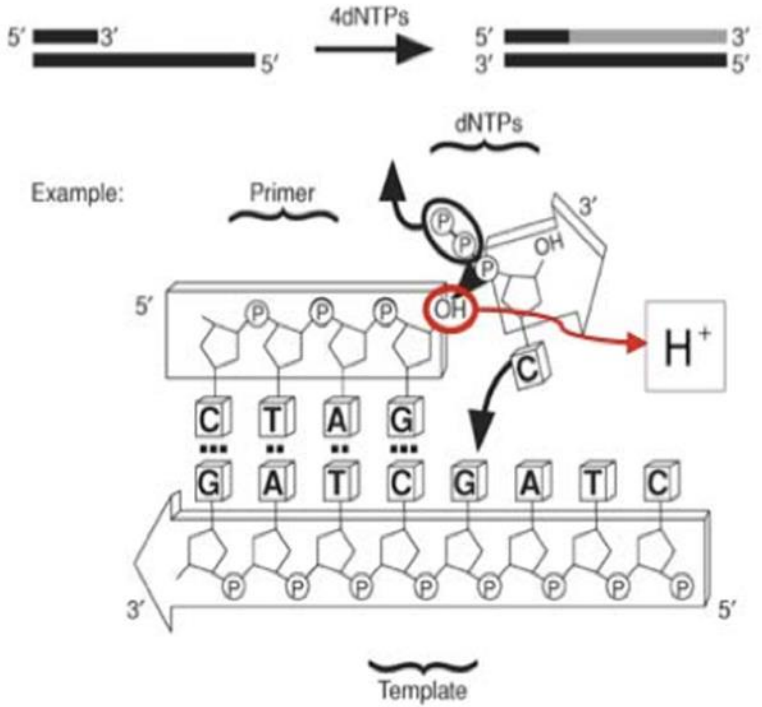
Sequencing and analysis in as little as 3 hours with the Ion GeneStudio S5 Prime System

Ion Torrent – Ion GeneStudio S5

	Ion GeneStudio S5 System	Ion GeneStudio S5 Plus System	Ion GeneStudio S5 Prime System
Max. throughput/day (chip type)	15 Gb (one Ion 540 Chip)	30 Gb (two Ion 540 chips)	50 Gb (two Ion 550 chips)
Total sequencing and analysis time at max. throughput	19 hr (Ion 540 Chip)	10 hr (Ion 540 Chip) 11.5 hr (Ion 550 Chip)	6.5 hr (Ion 540 Chip) 8.5 hr (Ion 550 Chip)
Compatible chips	Ion 510, 520, 530, 540 chips	Ion 510, 520, 530, 540, and 550 chips	

Ion Torrent - H+ ion-sensitive field effect transistors





The Ion Torrent next-generation sequencers then sequentially floods the chip with **one nucleotide after another**. If the next nucleotide that floods the chip is not a match, no **voltage change** will be recorded and no base will be called.

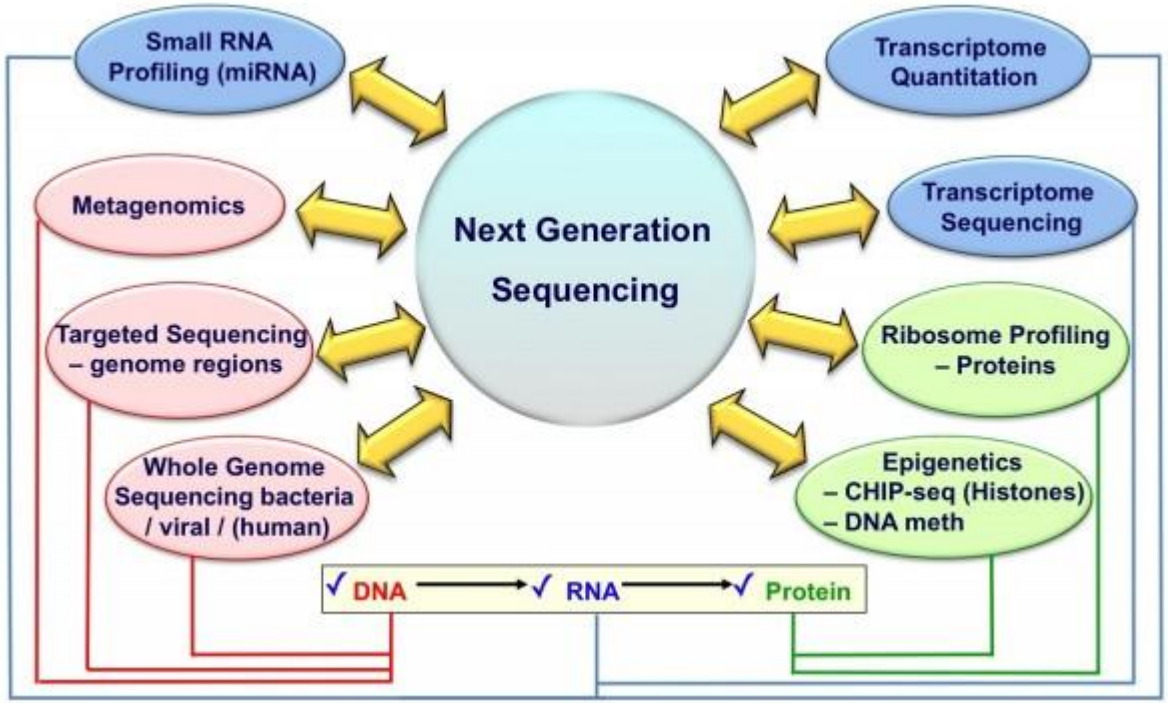
Performance evaluation

- Comparisons of WES performed on **Ion Proton** (**previous version of Ion GeneStudio S5**) and Illumina platforms revealed that the former performs with high accuracy at SNP discovery, but has a **high ratio of false positives in the identification of small indels.**
- In 2015, the company released the **HiQ sequencing chemistry** to improve accuracy of indel detection.
- Most **technical optimization studies** have focused on Illumina
- Similarly, most **bioinformatic methods** are optimized for analysis of Illumina-based data, while strategies to improve **data analysis and variants identification** on Ion Torrent platforms have not been discussed in detail so far.

Stringency	SNPs		INDELs	
	% TP retained	% FP filtered	% TP retained	% FP filtered
Low	99.2	21.2	98.9	10.2
Medium	96.1	41.9	95.1	21.2
High	89.9	68.2	89.8	40.4

doi:10.1007/s00439-016-1656-8

NGS Applications



- **Re-sequencing**: no assembly, just mapping on a known reference genome.
- **Metagenomics:** the study of genetic material recovered directly from environmental or clinical samples
- **ChIP-Seq:** Chromatin immunoprecipitation combined with DNA sequencing

Why RNA-Seq

- Sequencing the genome does not tell us which genes a cell is expressing, or the functions or processes it is carrying out at any given moment.

- To determine these, we need to work out its **gene expression profile**.

- If a gene is being used to make mRNA, it is considered **'on'**; if it is not being used to make mRNA, it is considered **'off'**.

The transcriptome are the genes that are being actively expressed at a given time.

- RNA-Seq also captures **unknown RNAs** (novel transcripts, splice variants, and gene fusions) and **Non-coding RNA** for gene regulation

Why RNA-Seq

- Investigate the **effects of different conditions on gene expression** by altering the environment to which the cell is exposed, and determining which genes are expressed.

- **Determine whether a cell is carrying out a function**. For example, certain genes are known to be involved in cell division; if these genes are active in a cell, you can tell the cell is undergoing division, or whether a cell is differentiated

- **Hypothesis generation**. If very little is known about when and why a gene will be expressed, expression profiling under different conditions can help design a hypothesis to test in future experiments. For example, if gene A is expressed only when the cell is exposed to other cells, this gene may be involved in intercellular communication. Further experiments could determine whether this is the case.

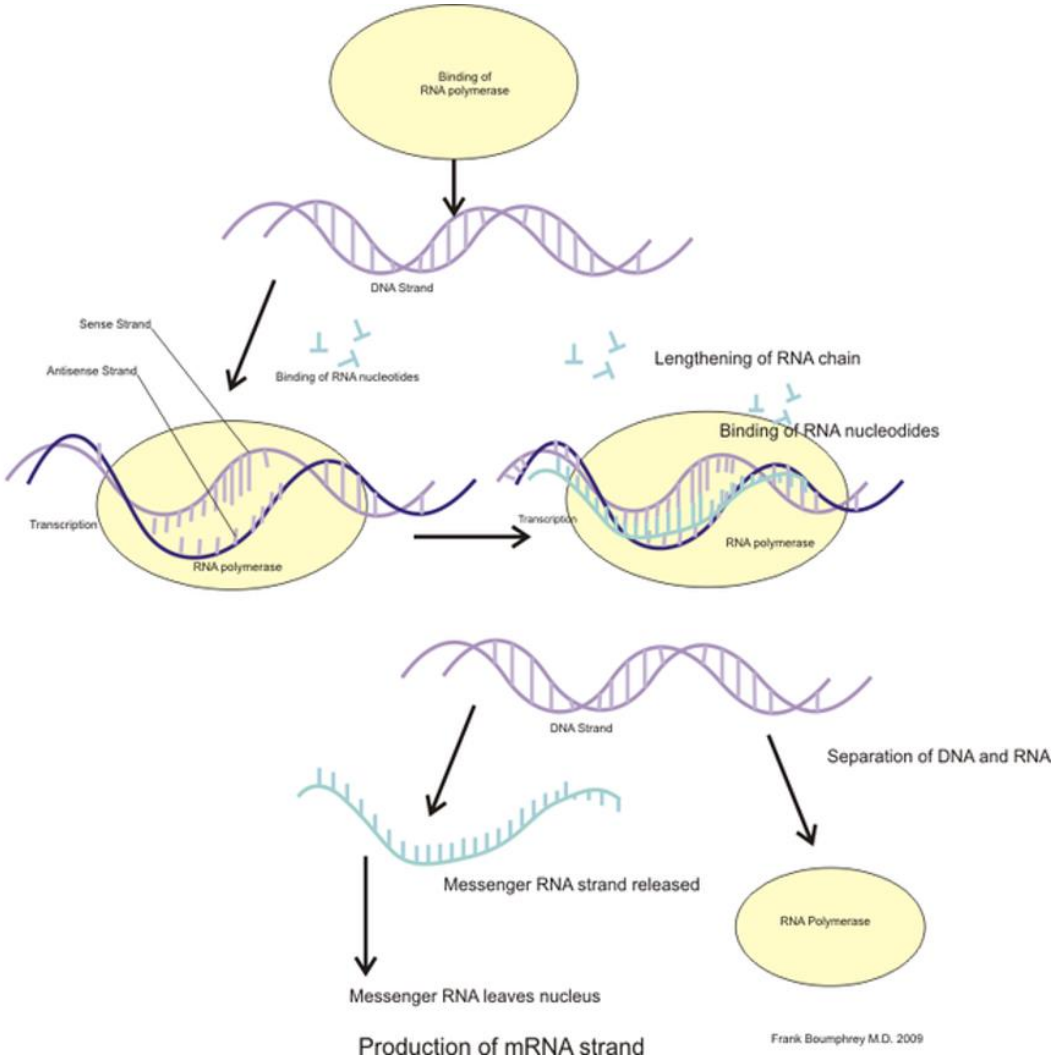
- Investigate **the effect of drug-like molecules** on cellular response. You could identify the gene markers of drug metabolism, or determine whether cells express genes known to be involved in response to toxic environments when exposed to the drug [4].

- **Diagnostic tool**. If cancerous cells express higher levels of certain genes, and these genes code for a protein receptor, this receptor may be involved in the cancer, and targeting it with a drug might treat the disease. Gene expression profiling might then be a key diagnostic tool for people with this cancer

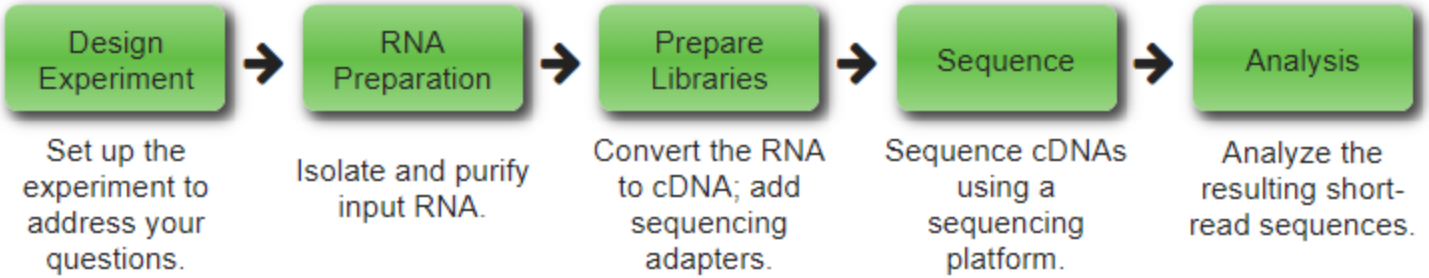
What is RNA-Seq

NGS opened a new phase in transcriptomics (aka expression profiling) thanks to:

- ➔ low requirements of nucleotide sequence product
- ➔ deep coverage

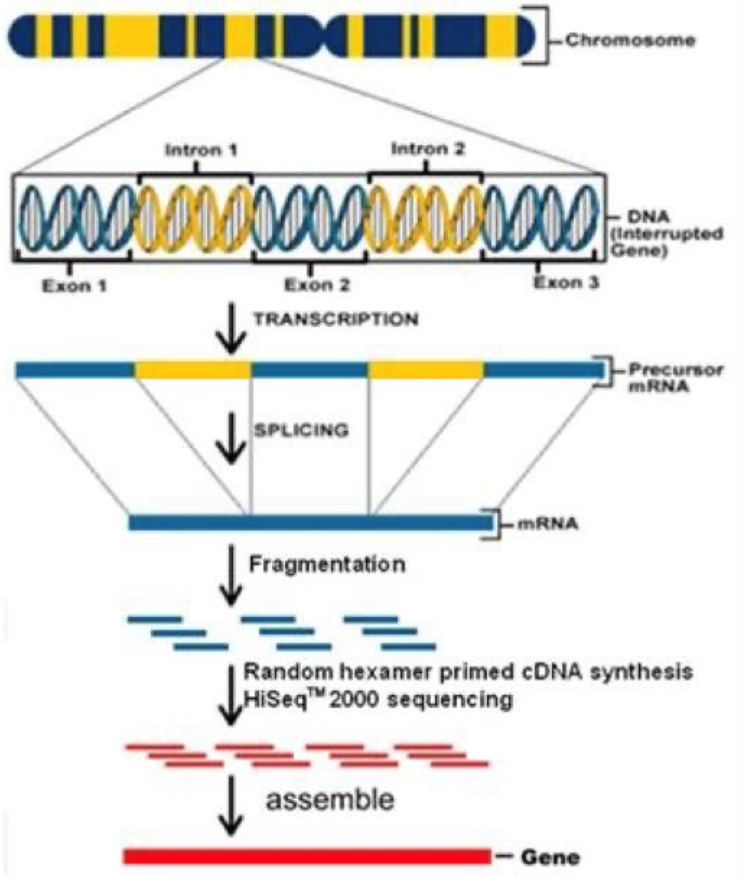


RNA sequencing workflow



RNA-seqlopedia: https://rnaseq.uoregon.edu/

RNA-Seq

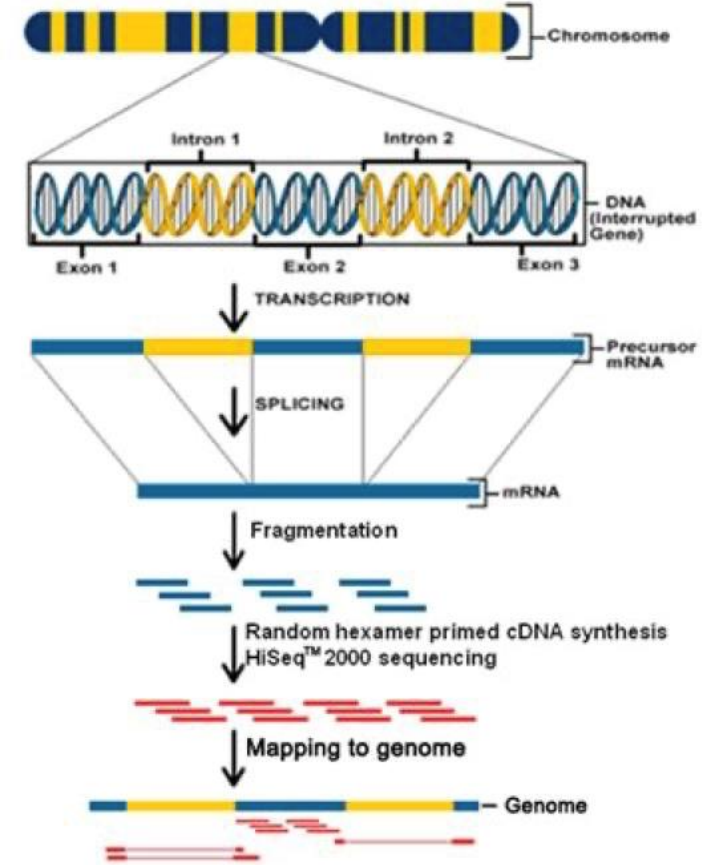


de novo transcriptome assembly

- Sequencing the transcriptome to investigate differentially expressed genes:
  - ➔ under different conditions, or
  - ➔ in different tissues
  - ➔ in different alleles

- The different expression can be in quantitative terms or in alternative splicing terms (eukaryotes only).

RNA-Seq

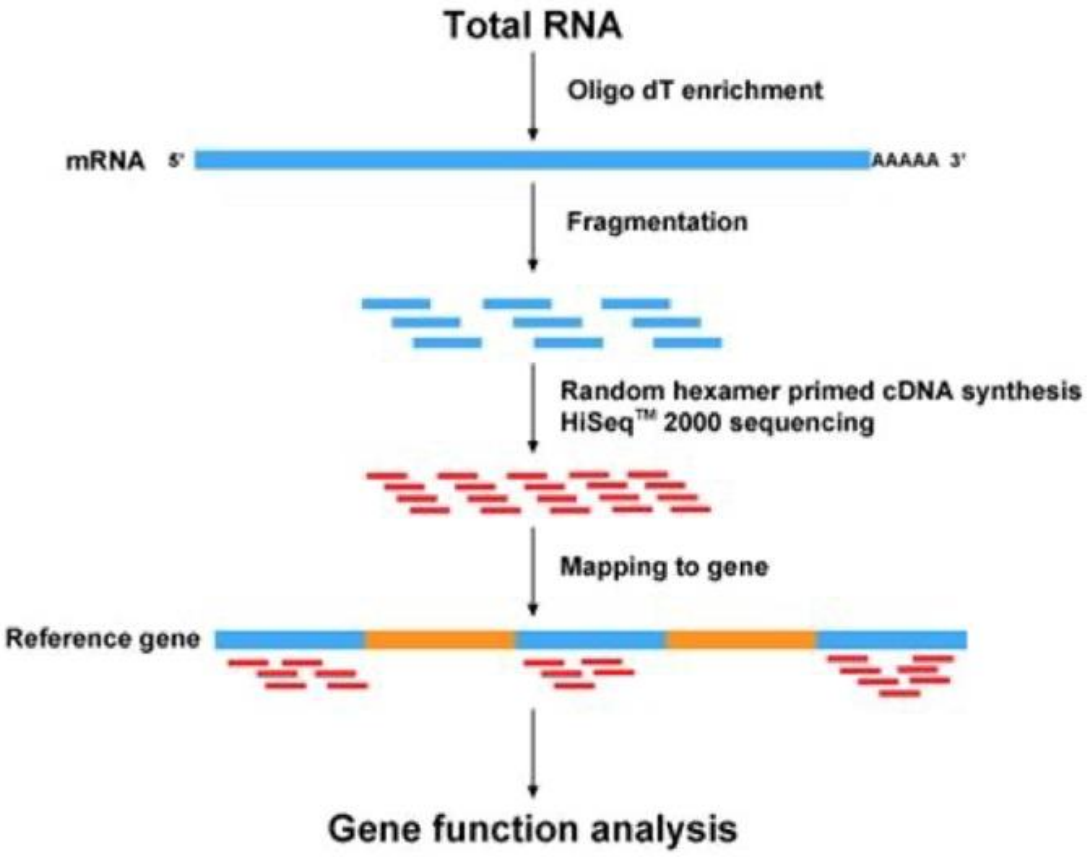


transcriptome re-sequencing

- Sequencing the transcriptome to investigate differentially expressed genes:
  - ➔ under different conditions, or
  - ➔ in different tissues
  - ➔ in different alleles

- The different expression can be in quantitative terms or in alternative splicing terms (eukaryotes only).

RNA-Seq quantification



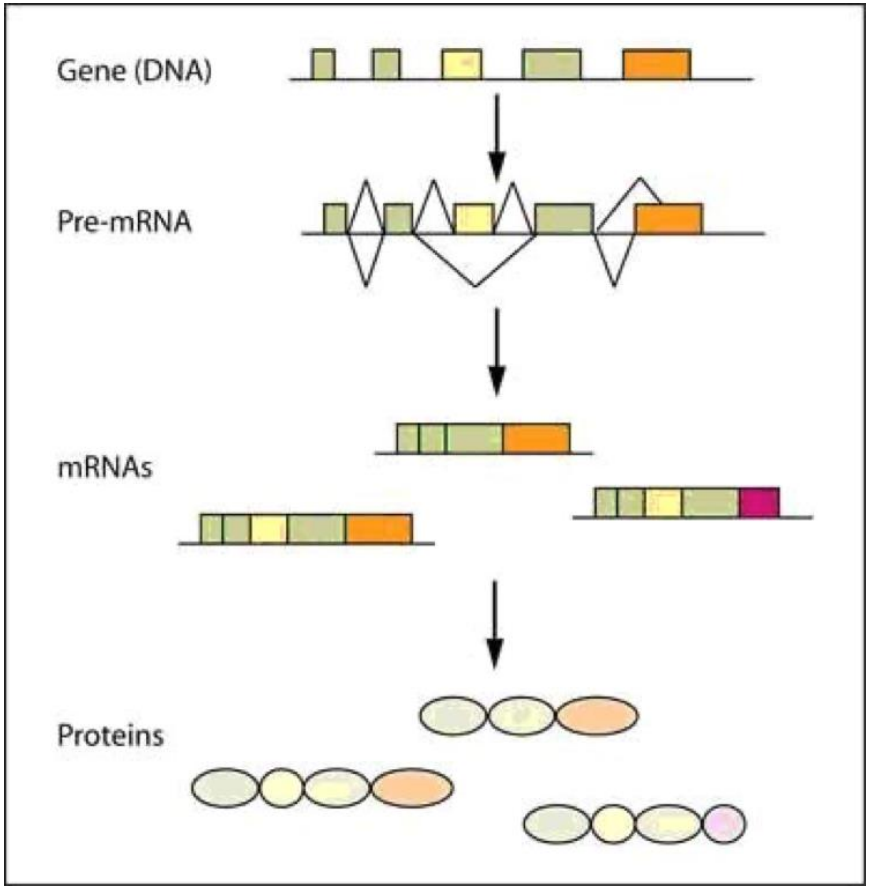
RNA-Seq

Quantitative data includes measuring differences in expression, alternative splicing, alternative TSS, and alternative polyadenylation between two or more treatments or groups.

Measure differential gene expression (DGE)

*DGE experiments must be designed to accurately measure both the **counts** of each transcript and the **variances (sampling, technical, biological)** that are associated with those numbers*

Challenge: Alternative Splicing (AS)

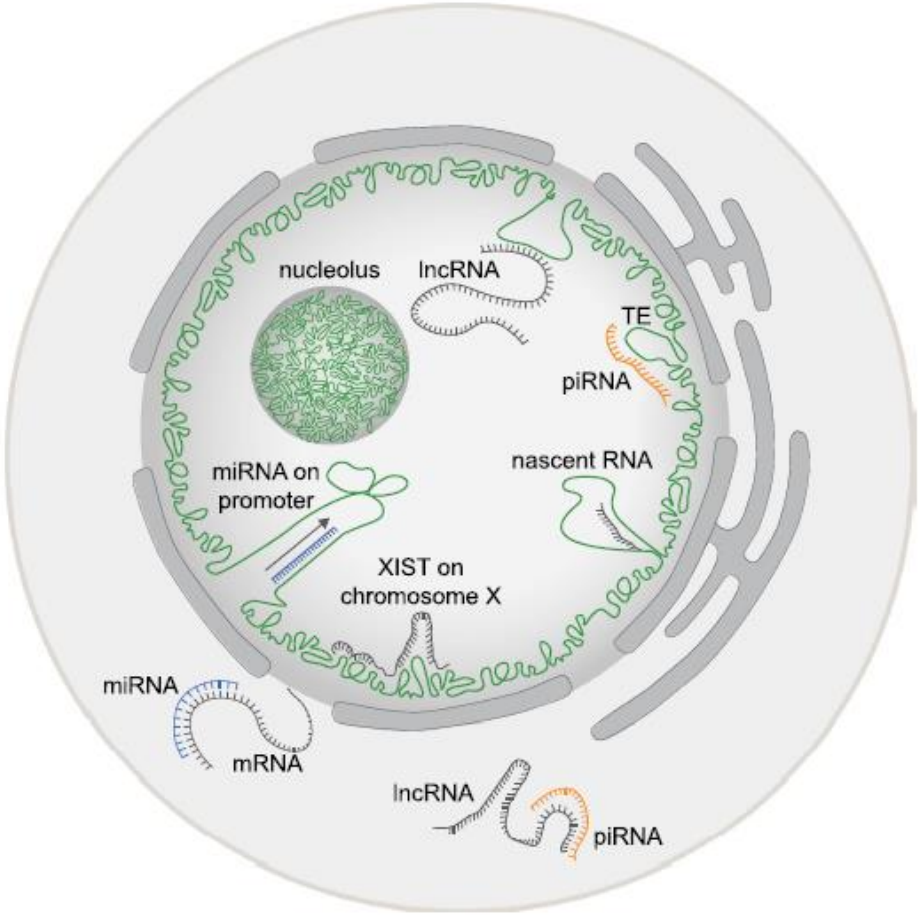


- AS is when several mRNAs can be produced from a unique pre-mRNA
- E.g. in humans there are approximately **30,000 genes** and it is estimated that 70% of human protein-coding genes undergo alternative splicing to generate up to **150,000-200,000 mRNAs** and proteins through alternative splice site usage.
- In 2008, an experiment revealed that 34% of human transcripts were not from known genes [Science v. 321 DOI: 10.1126/science.1160342]

non coding RNA: an RNA that does not encode a protein, but has other cellular functions

- ncRNA includes a wide class of regulatory RNA molecules whose function is as crucial as not yet understood.
- Discovering their sequences and (hence) genomic locations is hard because they are (mostly) small and poorly conserved over evolutionary time.
- Currently, ncRNA are mostly discovered by sequencing small RNA fragments, for which task NGS tools are ideal!
- In silico analysis of such data will be crucial for understanding it (secondary structure prediction, putative functions prediction based on learning methods).
- A new class of miRNA (or small RNA) is being discovered every day…

non coding RNA



The non-coding genome is widely accessible and transcriptionally active, producing a **multitude of RNA classes** reviewed in (*The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones* https://doi.org/10.1016/j.cell.2014.03.008)

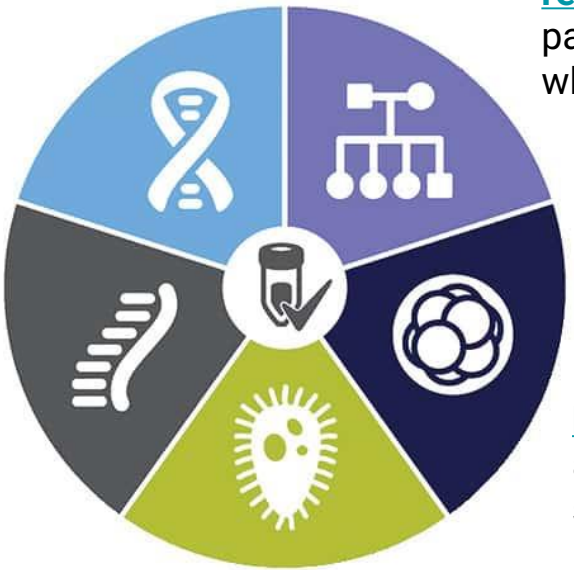
miRNA: RNA that, in complex with AGO protein, uses seed sequences near its 50 end to base pair with a target mRNA to induce deadenylation and decay or translational regulation

piRNA (PIWI-associated RNA): RNA that directs the modification of chromatin to repress transcription

NGS applications

Cancer research:
gene panels for SNPs, indels, copy number, gene expression, and gene fusion analysis

Inherited disease research:
panels for targeted gene or whole exome analysis

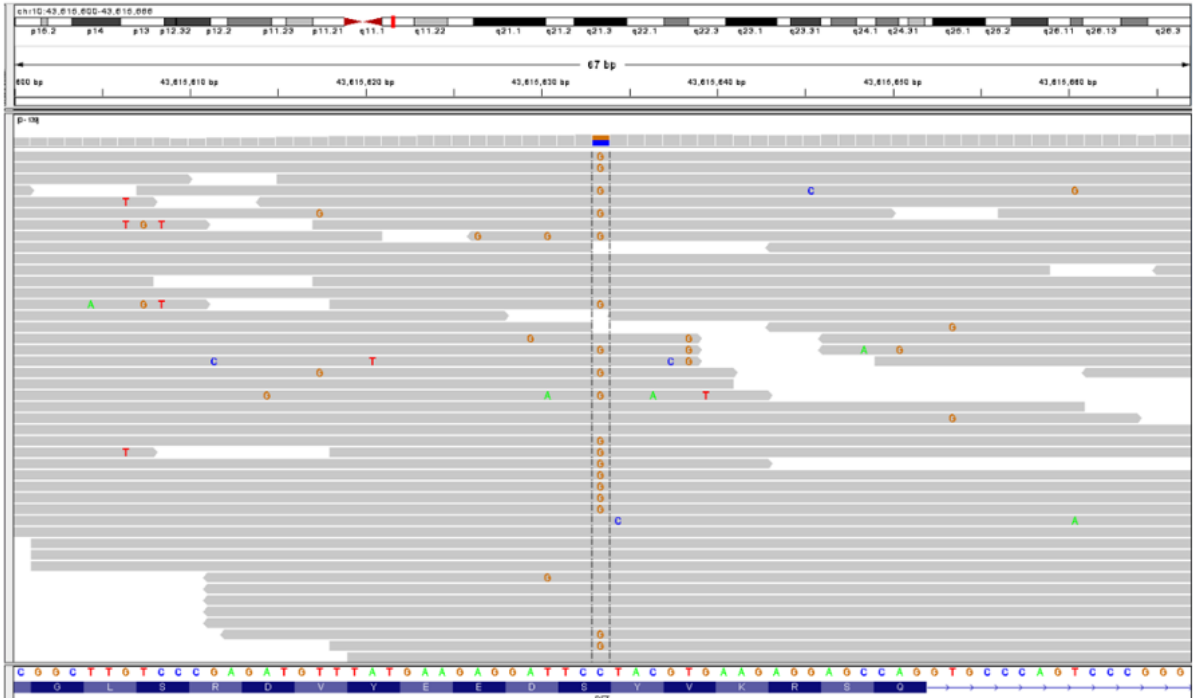


Gene expression analysis:
whole transcriptome RNA-Seq, targeted RNA sequencing, and small RNA sequencing

Reproductive health research:
aneuploidy detection, carrier screening research

Microbiology/infectious disease research:
microbial whole genomes, microbial typing, and metagenomics

NGS Applications: Examples-Inherited Conditions



→ **Discovery tool: Single gene disorders i.e. AD – Kabuki syndrome (MLL)**

→ **Causative mutations for multigenic diseases –superior to "one by one" approach of traditional sequencing**

→ **Diagnostic advancements for diseases with overlapping symptoms, multiple possible syndromes/genes**

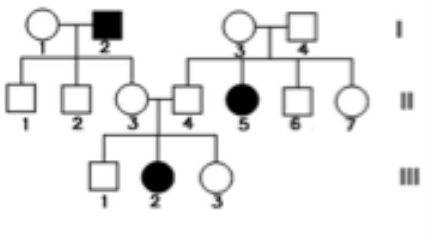
BioComputing

Inherited Conditions-Challenges and Opportunities

Challenges

Example: Monogenic disorders

- Novel missense mutations
- Germ line mosaicism
- Structural aberrations
- Imprinting effects
- Epigenetic factors



Inherited Conditions-Challenges and Opportunities

Challenges

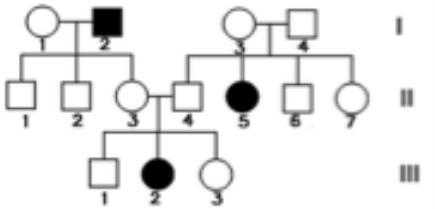
Example:
Monogenic disorders

- Novel missense mutations
- Germ line mosaicism
- Structural aberrations
- Imprinting effects
- Epigenetic factors

Opportunities

Example:
Multifactorial disease

- Risk loci more often in non-coding or inter-gene regions
- Pathogenicity of variants often unclear- less testing vs. monogenic disease
- Reference human genome cataloguing of variants = more test offerings



NGS Applications: Examples-Neoplastic Conditions

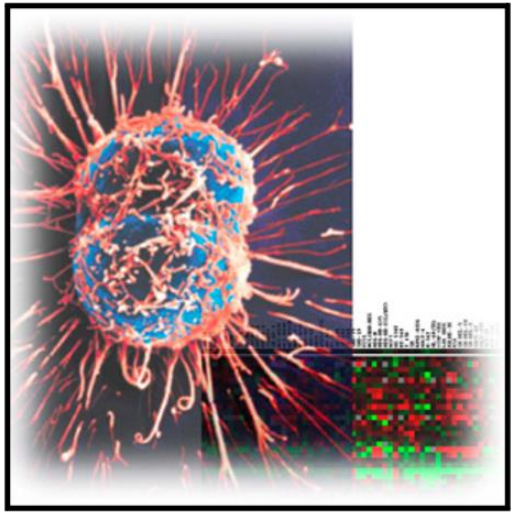
Cancer susceptibility genes

Risk assessment
Risk management

Somatic/driver mutations

Micro-RNAs

Methylation
Epigenetic changes



Patient stratification

Predictions of therapeutic response
personalized treatment

Therapeutic monitoring

Prognosis

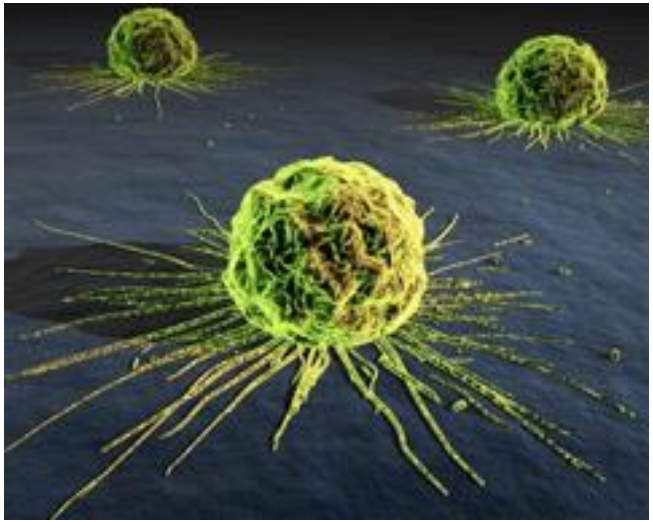
Alterations in gene expression

Molecular profiling

Tumor sub-typing

NGS Applications: Examples-Neoplastic Conditions

- Mutation panel screening
- Exome and transcriptome screening
- Genome sequencing-comparison to normal tissue/reference sample



Human genome project – reference genome and massive cataloguing of variants from different tumor sources (http://cancercommons.org, www.icgc.org and http://cancergenome.nih.gov/)

↓

Cost effective profiling of patient tumor DNA vs. mutation screening or profiling studies

NGS Analysis And Neoplastic Conditions

Quantitative nature of NGS- improvement vs. other technologies:

- Gene expression tests: Mammaprint (70 genes), Oncotype DX (21 genes) and Rotterdam signature (76 genes)... can be replaced by a single NGS analysis of signature transcripts!

Advantage:

- Germ line DNA characterization and somatic changes, transcriptome and methylation profiles - using a single, rapid and cost effective platform

NGS - Clinical Utility

- Balance of net health benefits vs. harm
- NGS –transformative for personalized treatment of disease
- Clinical indication - includes test rationale, patient population and clinical scenarios
- Principles of comparative effectiveness- requires individualized evidence-based approach for each patient

NGS - Clinical Utility: Challenges

NGS data density = frequently encountered variants of unknown significance

- Which variants are clinically actionable?
- Development of evidence-based scientific standards to evaluate utility in in different patient populations for accurate risk estimation
- Risk of over interpretation unnecessary medical action unwarranted psychological stress
- Careful selection of patients for genome sequencing and genetic counseling-crucial