Vall d'Hebron
Institut de Recerca

INTRODUCCIÓ A LES TECNOLOGIES DE 'NEXT GENERATION SEQUENCING'

Bioinformàtica per a la Recerca Biomèdica

Ricardo Gonzalo Sanz
ricardo.gonzalo@vhir.org

21/11/2018

Vall d'Hebron Hospital · Generalitat de Catalunya Departament de Salut · Institut Català de la Salut · UAB Universitat Autònoma de Barcelona · Generalitat de Catalunya Departament d'Economia i Coneixement · Institució CERCA Centres de Recerca de Catalunya · Instituto de Salud Carlos III

Vall d'Hebron
Institut de Recerca

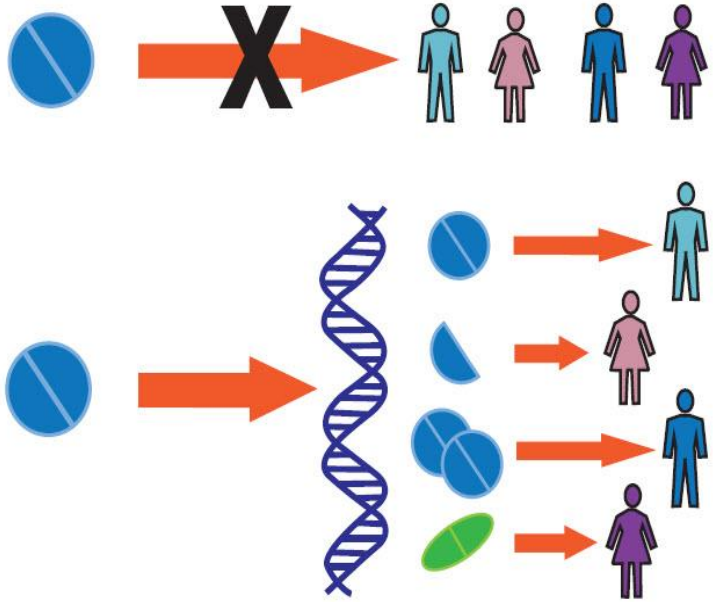
1. Introduction to NGS
2. First Generation Sequencing
3. Second Generation Sequencing
4. Third Generation Sequencing
5. Sequencing generation face to face
6. Applications of NGS techniques
7. A (very) brief introduction to DoE

Vall d'Hebron
Institut de Recerca

1. **Introduction to NGS**
2. First Generation Sequencing
3. Second Generation Sequencing
4. Third Generation Sequencing
5. Sequencing generation face to face
6. Applications of NGS techniques
7. A (very) brief introduction to DoE

Personalized medicine era

The right therapeutic strategy for the right person at the right time



https://iipm.medicinedept.iu.edu/

<u>Biomarker</u> identification:

- Diagnostic
- Susceptibility/risk (prevention)
- Prognostic (indolent vs. aggressive)
- Predictive (response)

1. Introduction to NGS

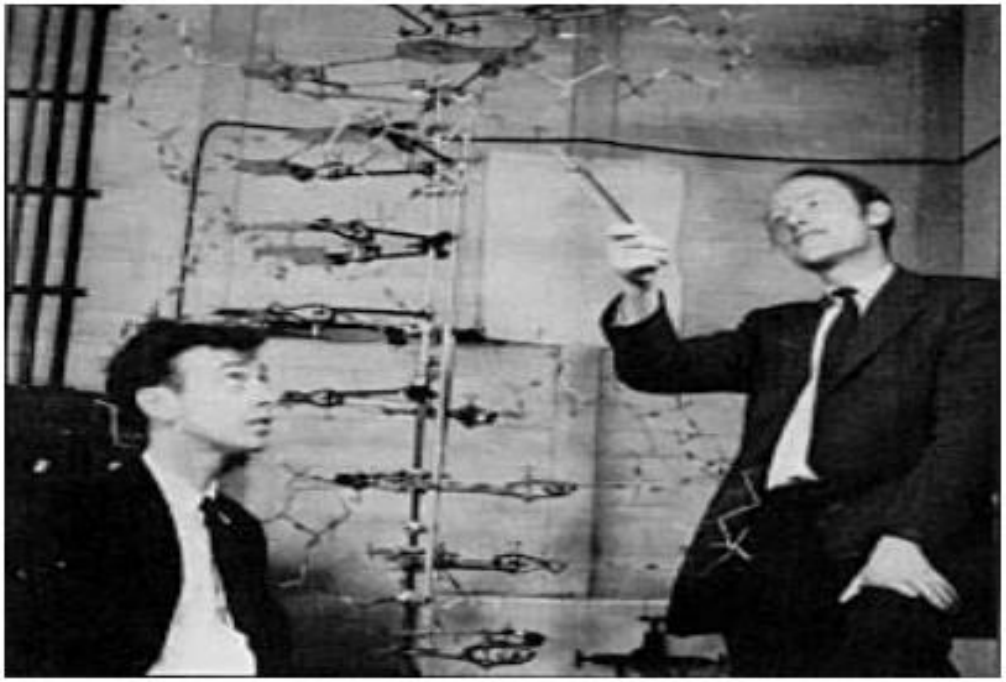
Genomics is a branch of genetics that enables the study of genomes of whole organisms.

Gregor Mendel

It differs from “classical genetics” in that it considers the hereditary material of an organism **in global** rather than **one gene** or one **gene product** at a time.

Vall d'Hebron
Institut de Recerca

Deoxyribose nucleic acid (DNA)



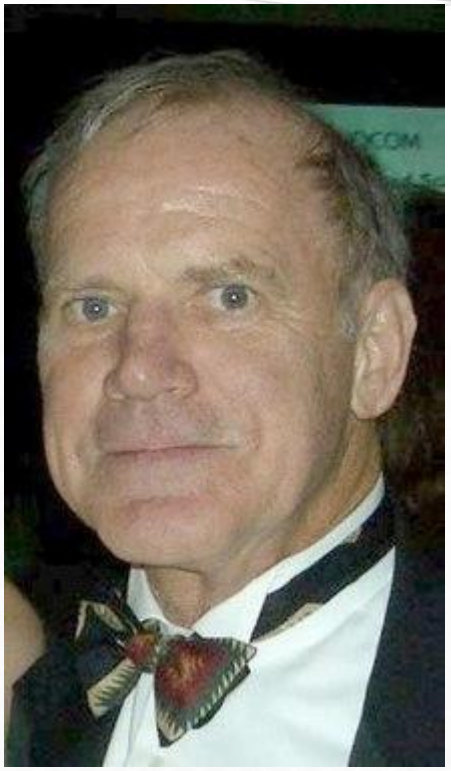
"We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest."

J.D. Watson & F. H. C. Crick. (1953). Molecular structure of Nucleic Acids. *Nature*. **171**: 737-738.

1. Introduction to NGS

Polymerasa Chain Reaction

In 1983, Kary Mullis, PhD, a scientist at the Cetus Corporation, conceived of PCR as a method to copy DNA and synthesize large amounts of a specific target DNA



PCR was awarded the 1993 Nobel Prize in Chemistry

.

Vall d'Hebron
Institut de Recerca

Genome sequencing

Genome sequencing is figuring out the order of DNA nucleotides, or bases, in a genome—the order of As, Cs, Gs, and Ts that make up an organism's DNA.

Vall d'Hebron
Institut de Recerca

Why is genome sequencing so important?

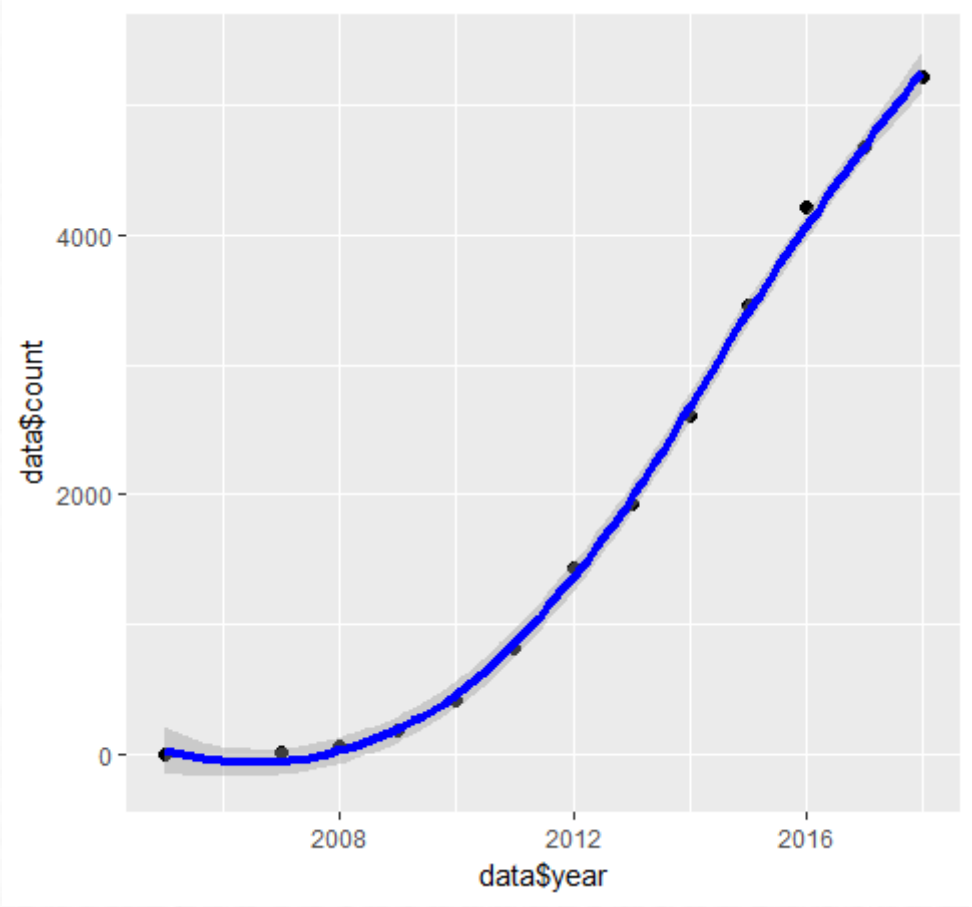
- **How the genome as a whole works:** how genes work together to direct the growth, development and maintenance of an entire organism.
- **Find genes** much more easily and quickly.
- Genes account for less than 25 percent of the DNA in the genome, and so knowing the entire genome sequence will help scientists **study the parts of the genome outside the genes**.

Human Genome Project

The Human Genome Project (HGP) was the international, collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings.

- Begin in 1990
- First draft in February 2001
- full sequence April 2003
- It catalyzes the developing and improving of sequencing techniques

1. Introduction to NGS



Vall d'Hebron
Institut de Recerca

Consortia-based projects

Table 2. Examples of Consortia-Based Projects

Initiative	Purpose	Website
1000 Genomes Project	Cataloging normal variation in diverse human populations.	www.1000genomes.org
The Encyclopedia of DNA Elements	Identifying functional genomic elements in the human genome.	www.encodeproject.org
Roadmap Epigenomics Project	Catalogue human epigenomic data with the goal of advancing basic biology and disease-oriented research.	www.roadmapepigenomics.org
Human Microbiome Project	Comprehensive characterization of the human microbiome and analysis of its role in human health and disease.	www.hmpdacc.org
Genotype-Tissue Expression Program	Characterizing gene expression and regulation in many human tissues and correlating with genetic variation and disease.	www.commonfund.nih.gov/GTEx/index
Human Immunology Project Consortium	Characterizing the diverse states of the human immune system following infection, vaccination or treatment.	http://www.immuneprofiling.org
Grand Opportunity Exome Sequencing Project	Discovery of novel genes and mechanisms contributing to heart, lung and blood disorders.	https://esp.gs.washington.edu/drupal
The Cancer Genome Atlas	Understanding the molecular basis of cancer.	www.cancergenome.nih.gov
International Cancer Genome Consortium	Describing the genomic, transcriptomic and epigenomic changes in 50 different tumor types.	www.icgc.org
Clinical Sequencing Exploratory Research Program	Develop methods as well as the legal and ethical frameworks necessary to integrate sequencing into the clinic.	www.genome.gov/27546194
Centers for Mendelian Genomics	Discovering the genes and genetic variants underlying human Mendelian disorders.	www.mendelian.org
Undiagnosed Diseases Network	Promoting the use of genomic data to elucidate the mechanisms underlying the diseases of unknown etiology.	www.commonfund.nih.gov/Diseases/index
Newborn Sequencing in Genomic Medicine and Public Health	Exploring the challenges and opportunities associated with using genomic sequence information in the newborn period.	www.genome.gov/27558493
The Pediatric Cardiac Genomics Consortium	Determining the genes responsible for congenital heart disease.	www.benchtobassinet.com
Alzheimer's Disease Sequencing Project	Identifying genes contributing to risk of developing Alzheimer's disease in multiethnic populations.	www.niagads.org/adsp

2008-2015

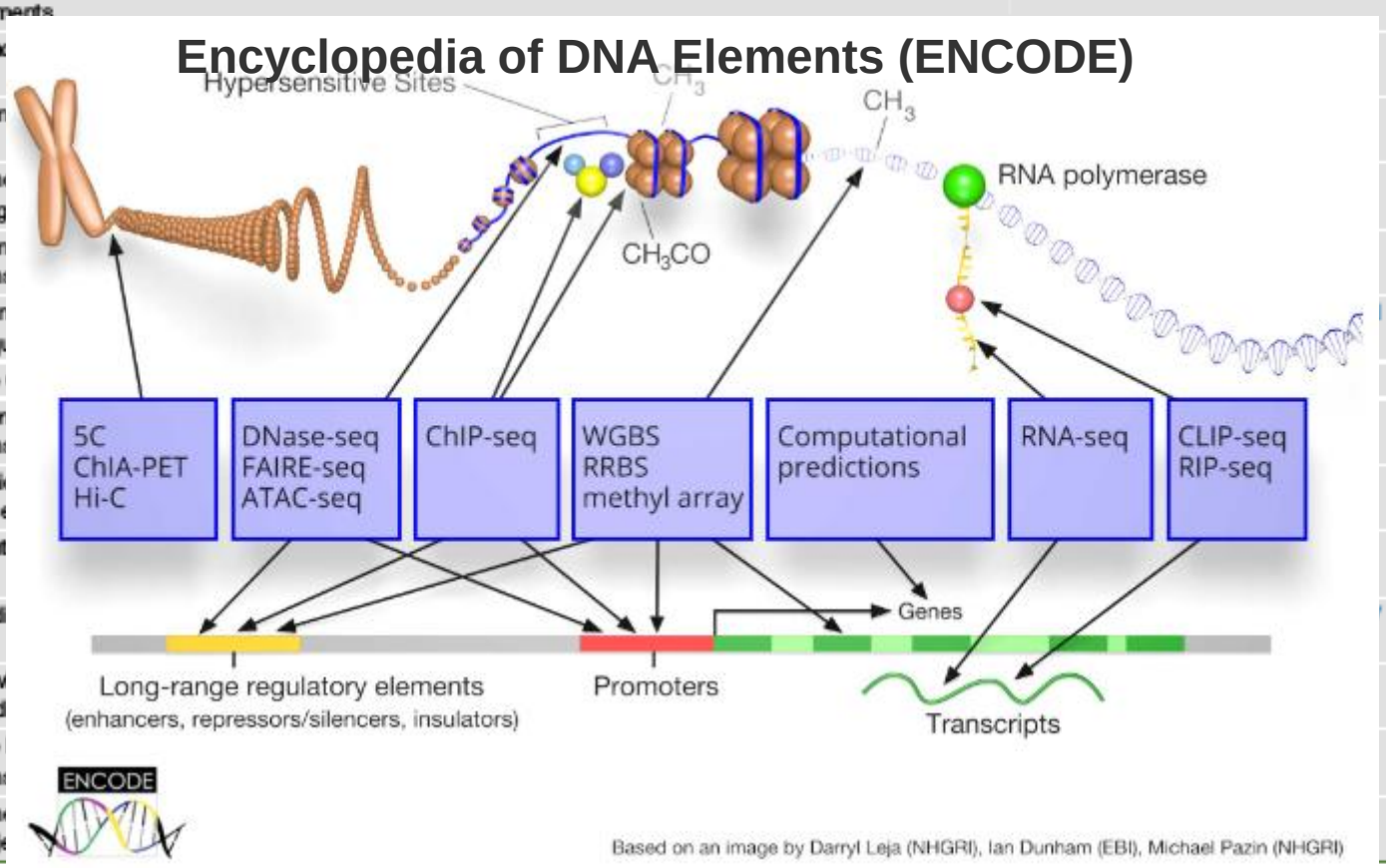
Vall d'Hebron
Institut de Recerca

Consortia-based projects

Table 2. Examples of Consortia-Based Projects

Initiative	Purpose	Website
1000 Genomes Project	Cataloging normal variation in diverse human populations.	www.1000genomes.org
The Encyclopedia of DNA Elements	Identifying functional genomic elements in the human genome.	www.encodeproject.org

Encyclopedia of DNA Elements (ENCODE)



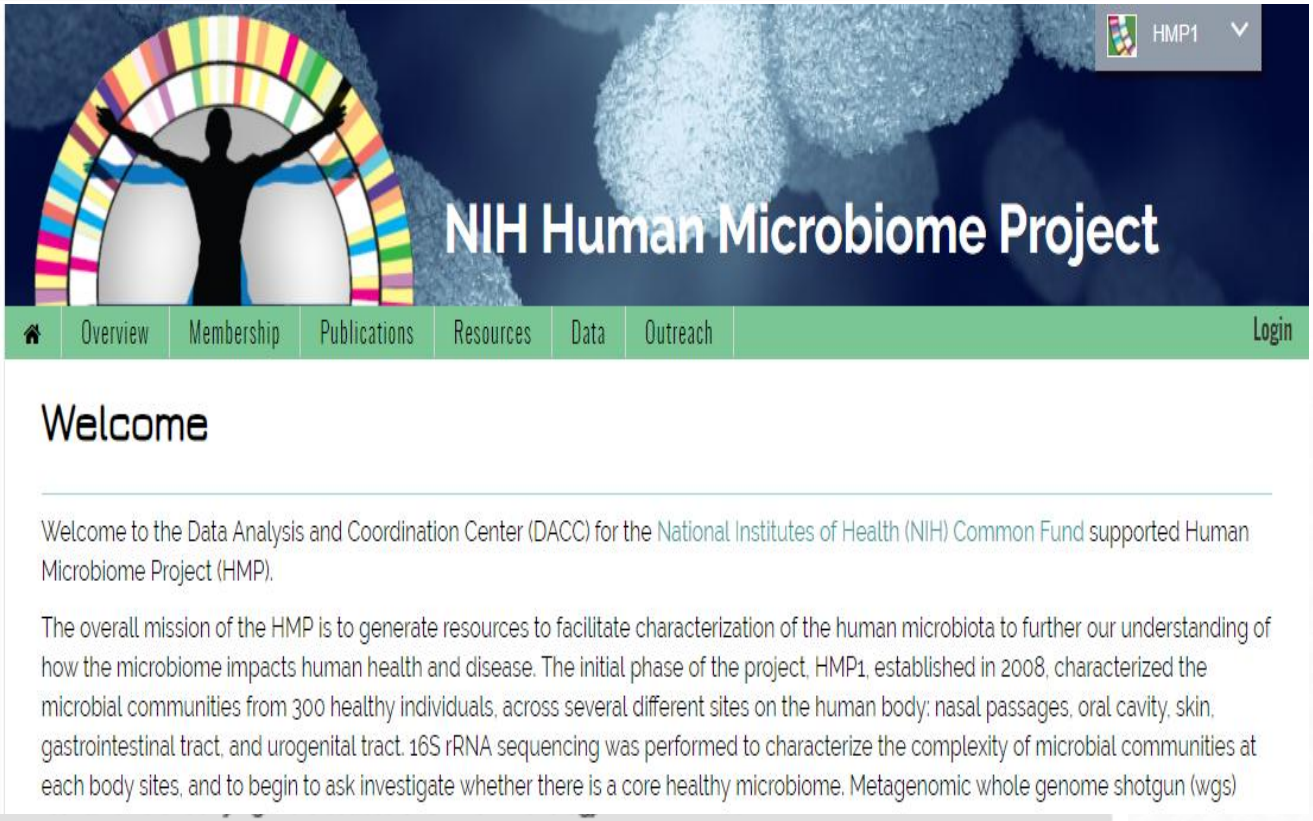
Molecular Cell 58, May 21, 2015 ©2015 Elsevier Inc.

Vall d'Hebron
Institut de Recerca

Consortia-based projects

Table 2. Examples of Consort

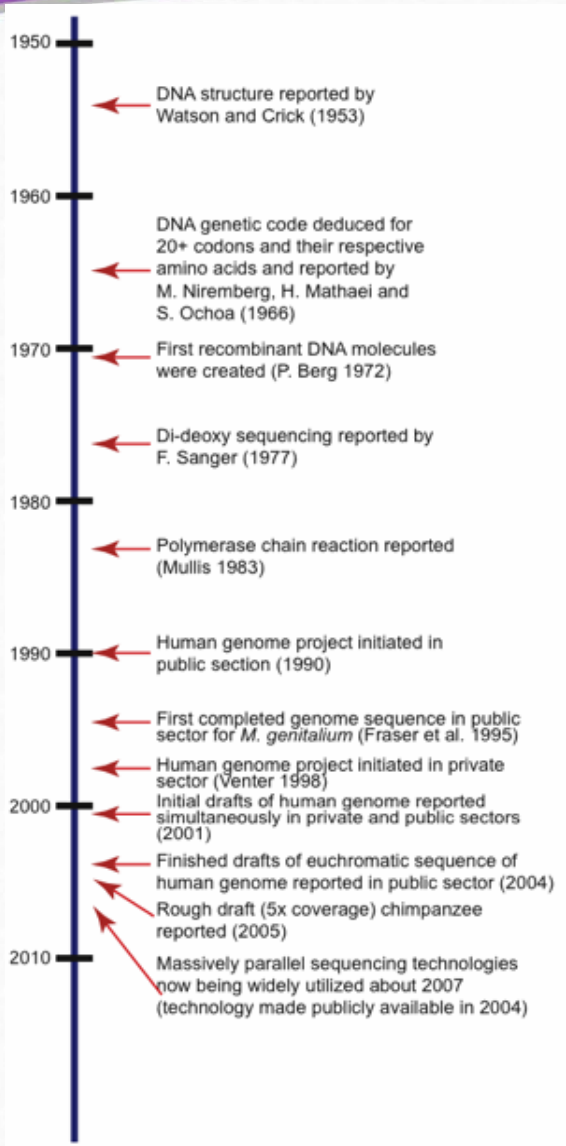
Initiative
1000 Genomes Project
The Encyclopedia of DNA Elements
Roadmap Epigenomics Project
Human Microbiome Project
Genotype-Tissue Expression Program
Human Immunology Project Consortium
Grand Opportunity Exome Sequencing Project
The Cancer Genome Atlas
International Cancer Genome Consortium
Clinical Sequencing Exploratory Research Program
Centers for Mendelian Genomics
Undiagnosed Diseases Network
Newborn Sequencing in Genomic Medicine and Public Health
The Pediatric Cardiac Genomics Consortium
Alzheimer's Disease Sequencing Project

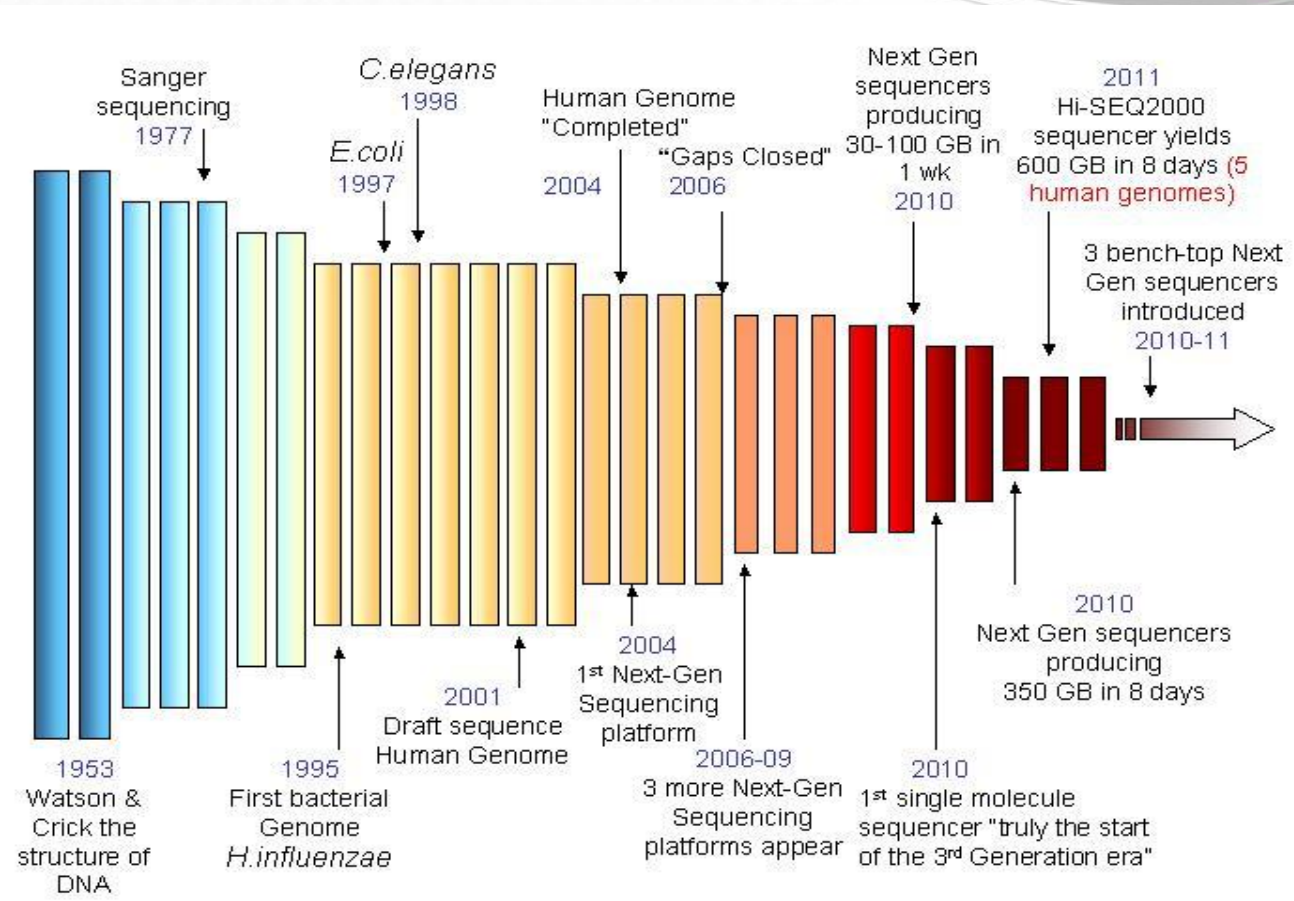


Newborn Sequencing in Genomic Medicine and Public Health	Exploring the challenges and opportunities associated with using genomic sequence information in the newborn period.	www.genome.gov/27558493
The Pediatric Cardiac Genomics Consortium	Determining the genes responsible for congenital heart disease.	www.benchtobassinet.com
Alzheimer's Disease Sequencing Project	Identifying genes contributing to risk of developing Alzheimer's disease in multiethnic populations.	www.niagads.org/adsp

Molecular Cell 58, May 21, 2015 ©2015 Elsevier Inc.

1. Introduction to NGS



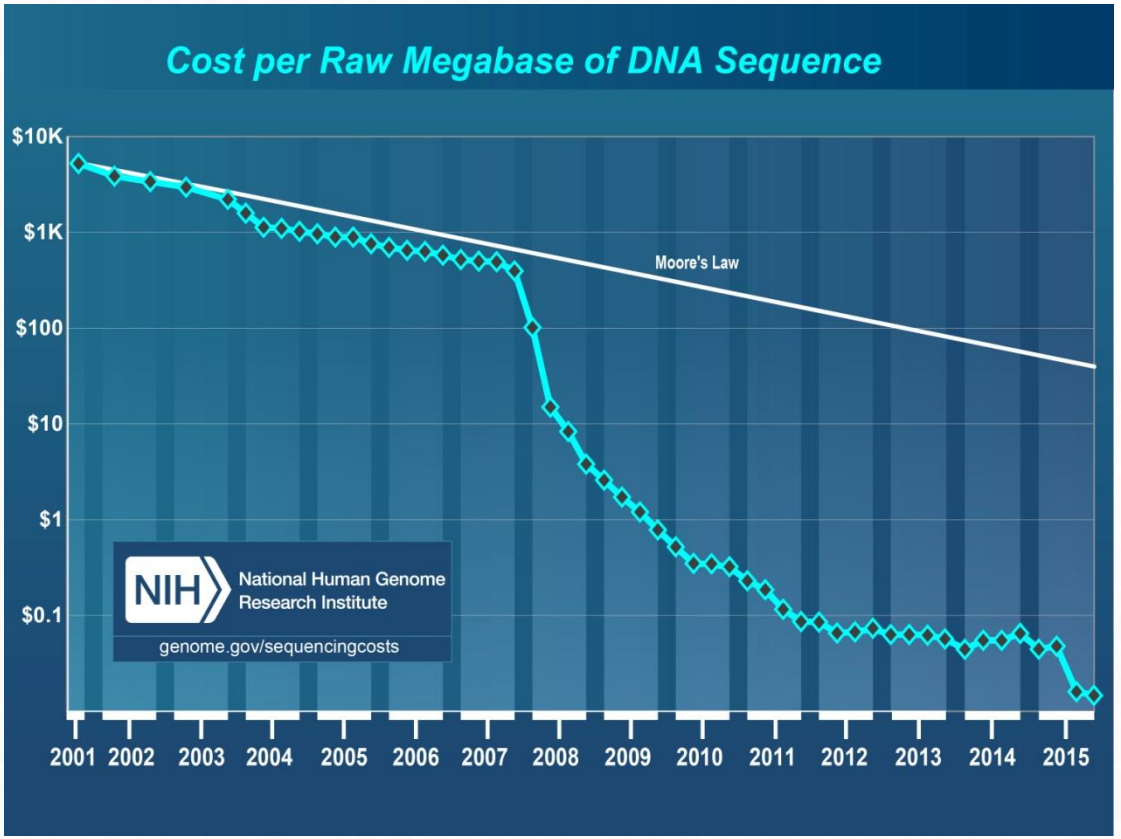


http://www.ipc.nxgenomics.org/newsletter/no11.htm

Vall d'Hebron
Institut de Recerca

Cost of sequencing

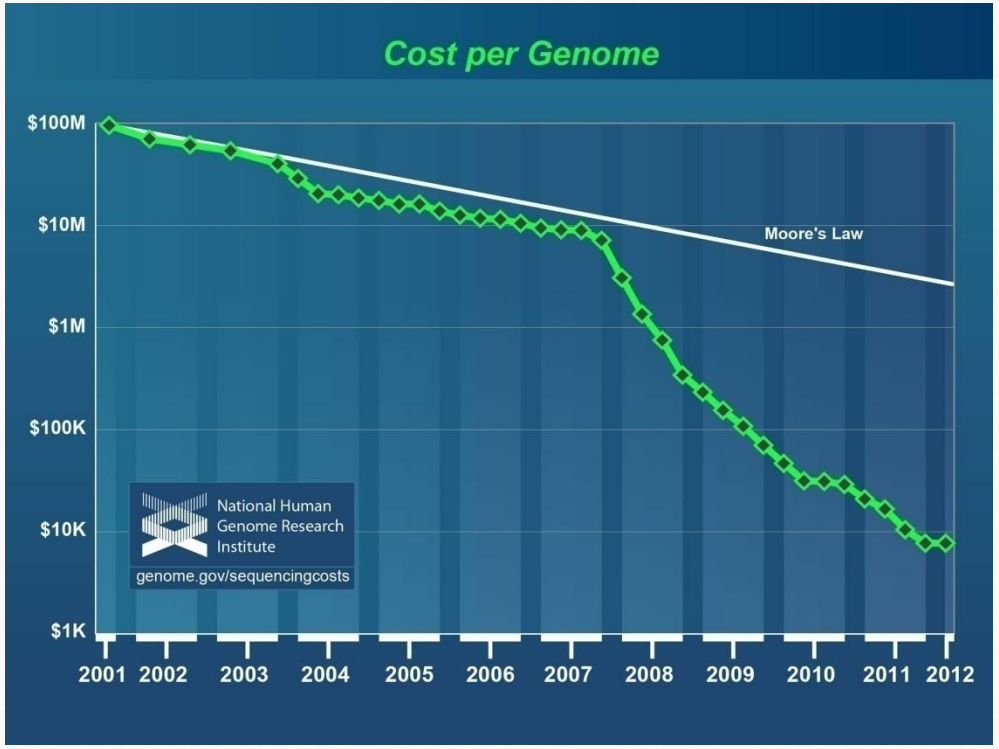
- Size of the genome
- Whole genome/exome/'de novo'
- Data quality



Moore's Law: the number of transistors in an integrated circuit doubles approximately every two years (1965).

Vall d'Hebron
Institut de Recerca

Cost of sequencing



Cost of human genome sequence:

HGP: 1 - $3x10^{12}$$

2006: $14x10^{6}$$

2016: 1000-4000$

https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

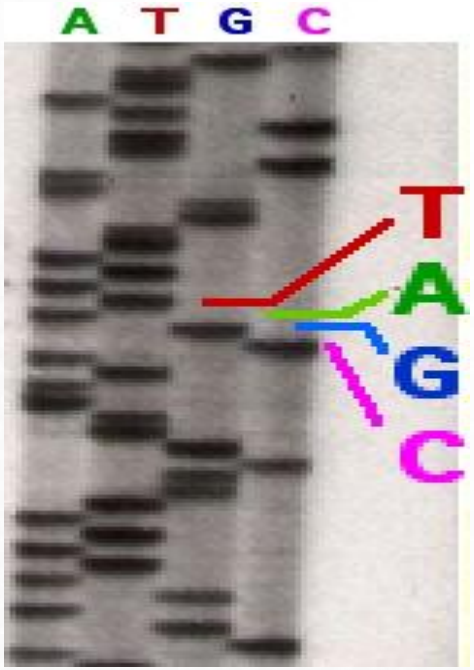
Vall d'Hebron
Institut de Recerca

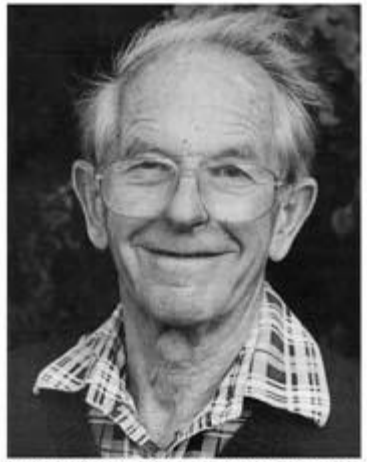
1. Introduction to NGS
2. First Generation Sequencing
3. Second Generation Sequencing
4. Third Generation Sequencing
5. Sequencing generation face to face
6. Applications of NGS techniques
7. A (very) brief introduction to DoE

2. First Generation Sequencing

Sanger sequencing

Method of **DNA sequencing** based on the selective incorporation of chain terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication. Developed by Frederick Sanger and colleagues in 1977, it was the most widely used sequencing method for approximately 25 years.





Courtesy of Dr. F. Sanger, MRC, Cambridge.
Noncommercial, educational use only.

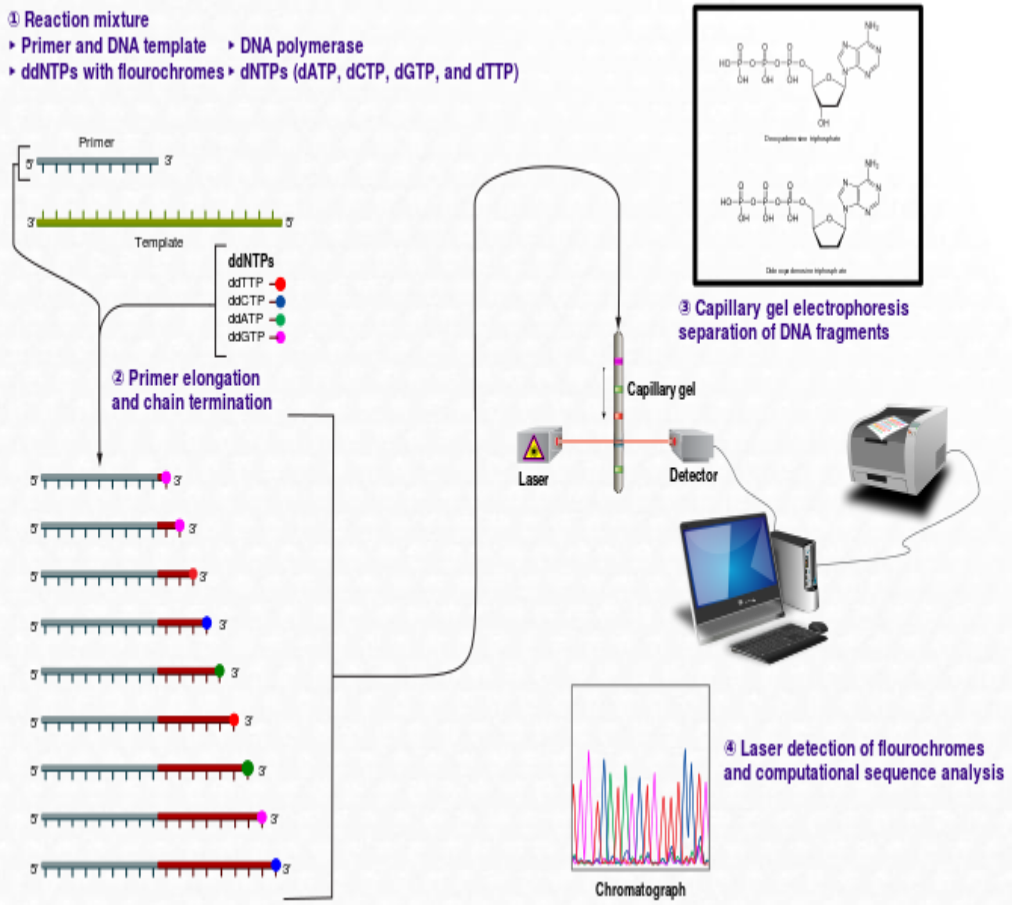
Frederick Sanger received two Nobel prizes (in the same category), for his work on protein sequencing and DNA sequencing

http://www.yourgenome.org/stories/third-generation-sequencing

2. First Generation Sequencing

Sanger sequencing. How it works?

- A DNAp enzyme is used to replicate a ssDNA. In the mix reaction, there exist normal and modified nucleotides.
- Random incorporation of modified nucleotides stops the synthesis reaction.
- Each generated fragment will have a different length that could be distinguish by gel electrophoresis.



https://www.youtube.com/watch?v=vK-HlMaitnE

<1kb read

Vall d'Hebron
Institut de Recerca

1. Introduction to NGS
2. First Generation Sequencing
3. Second Generation Sequencing
4. Third Generation Sequencing
5. Sequencing generation face to face
6. Applications of NGS techniques
7. A (very) brief introduction to DoE

Why second generation sequencing?

- Disadvantage of Sanger sequencing: **low sequence output**
  - using of gels or polymers as separation media
  - limited number of samples which could be handled in parallel
  - difficulties with automation of the sample preparation

These limitations triggered the efforts to develop new techniques

3. Second Generation Sequencing

Main characteristics of NGS:

- high speed and throughput
- shorter reads
- accuracy
- much higher degree of sequence coverage
- Huge data storage demands

3.Second Generation Sequencing

Vall d'Hebron
Institut de Recerca

2ond generation instruments

High throughput

Benchtop

ROCHE

GS Junior 454

Illumina

HiSeq 4000

HiSeq X-Ten

NovaSeq 6000 System

NextSeq Series

iSeq 100 System MiniSeq System MiSeq Series

Life Technologies

SOLID5500xl

Ion Gene Studio S5 PrimeSystem

Ion Proton

IonPGM

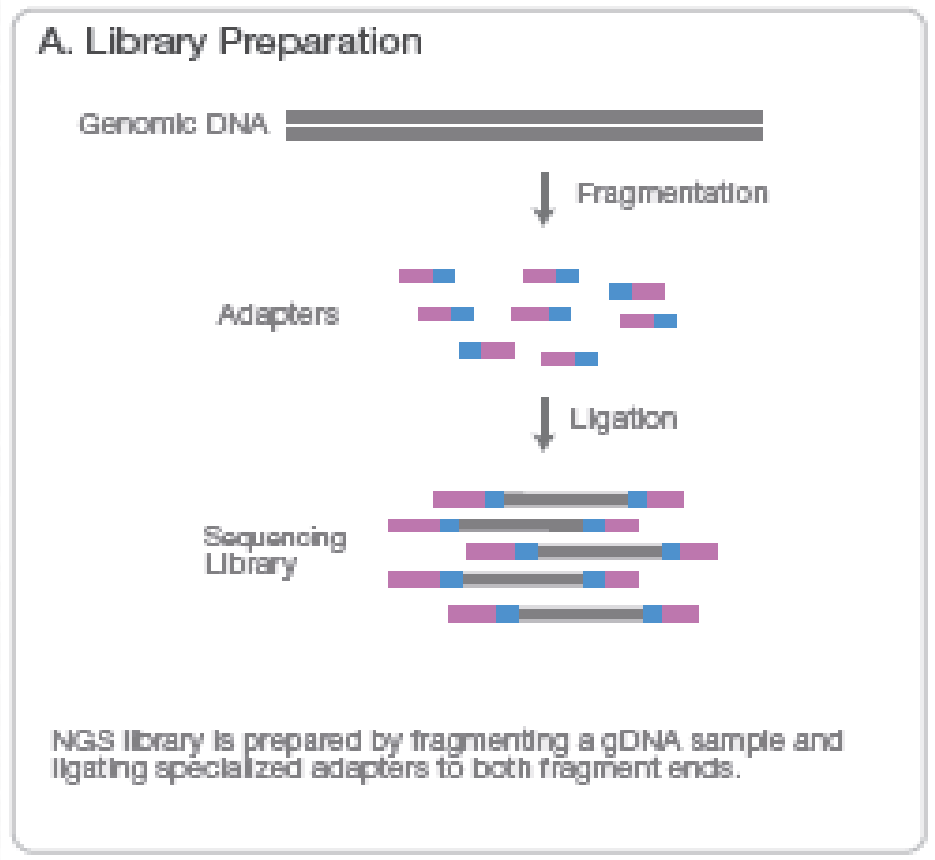
Vall d'Hebron
Institut de Recerca

Basic NGS workflow.

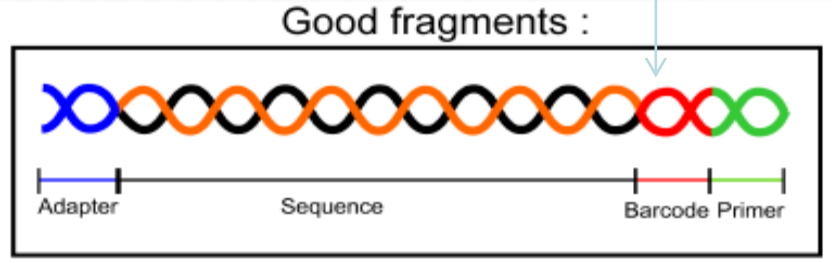
1. Library Preparation	It is prepared by random fragmentation of DNA or cDNA sample, followed adapter ligation. Adapter-ligated fragments are then PCR amplified and gel purified
2. Clonal amplification	Each DNA fragment is amplified millions of times.
3. Sequencing	The nucleotides incorporated are read by the detector

3. Second Generation Sequencing

Library preparation:



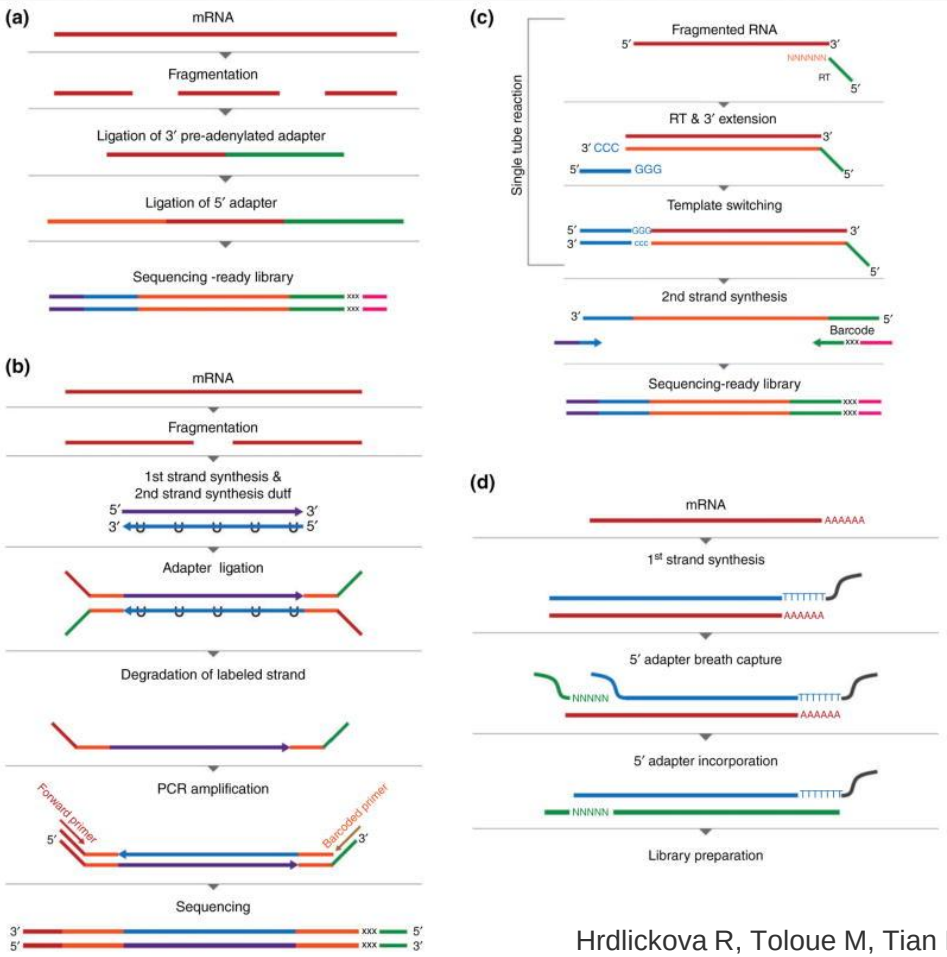
Useful for multiplexing samples



Depending on the final application and method used for sequencing, different kits are available

3. Second Generation Sequencing

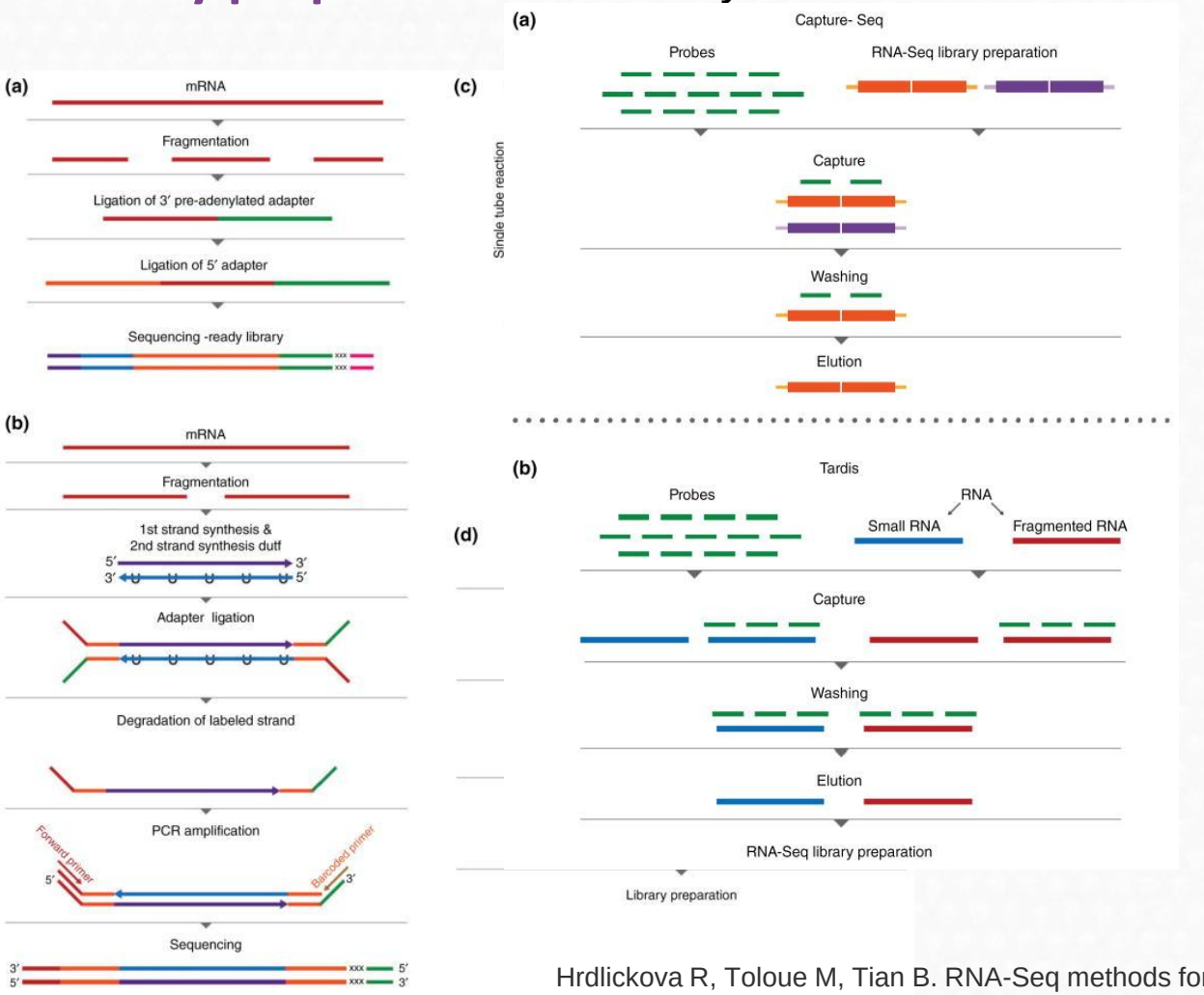
Library preparation: Too many methods / kits available



Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2016;8(1):10.1002/wrna.1364.

3. Second Generation Sequencing

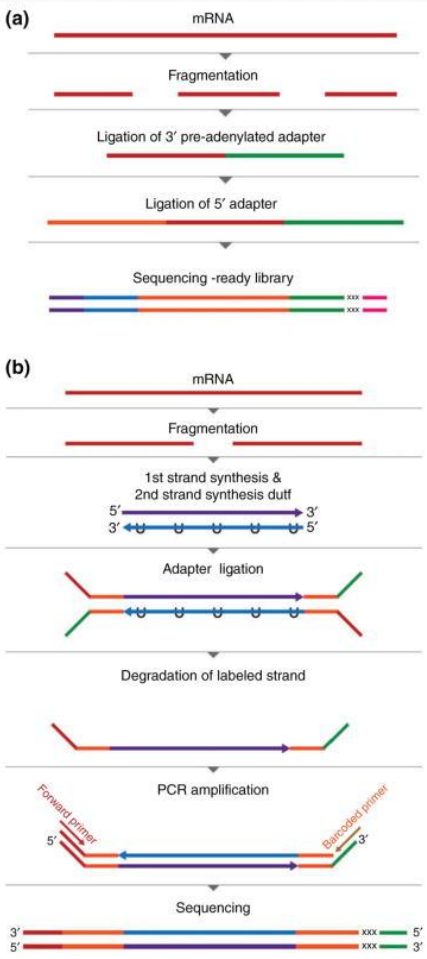
Library preparation: Too many methods / kits available

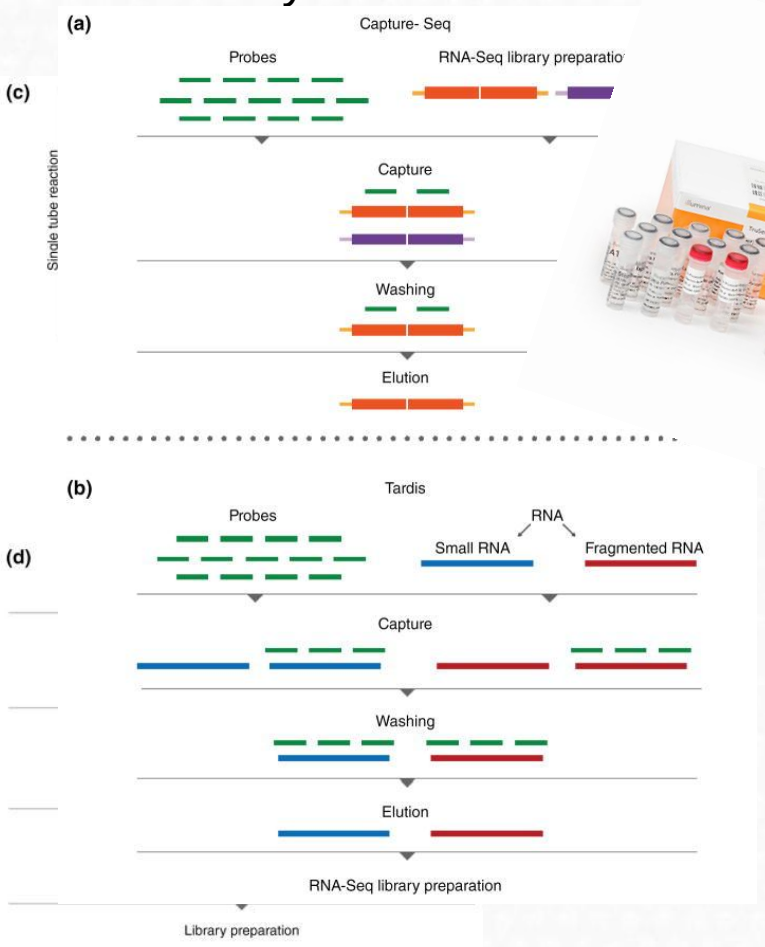


Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2016;8(1):10.1002/wrna.1364.

3. Second Generation Sequencing

Library preparation: Too many methods / kits available

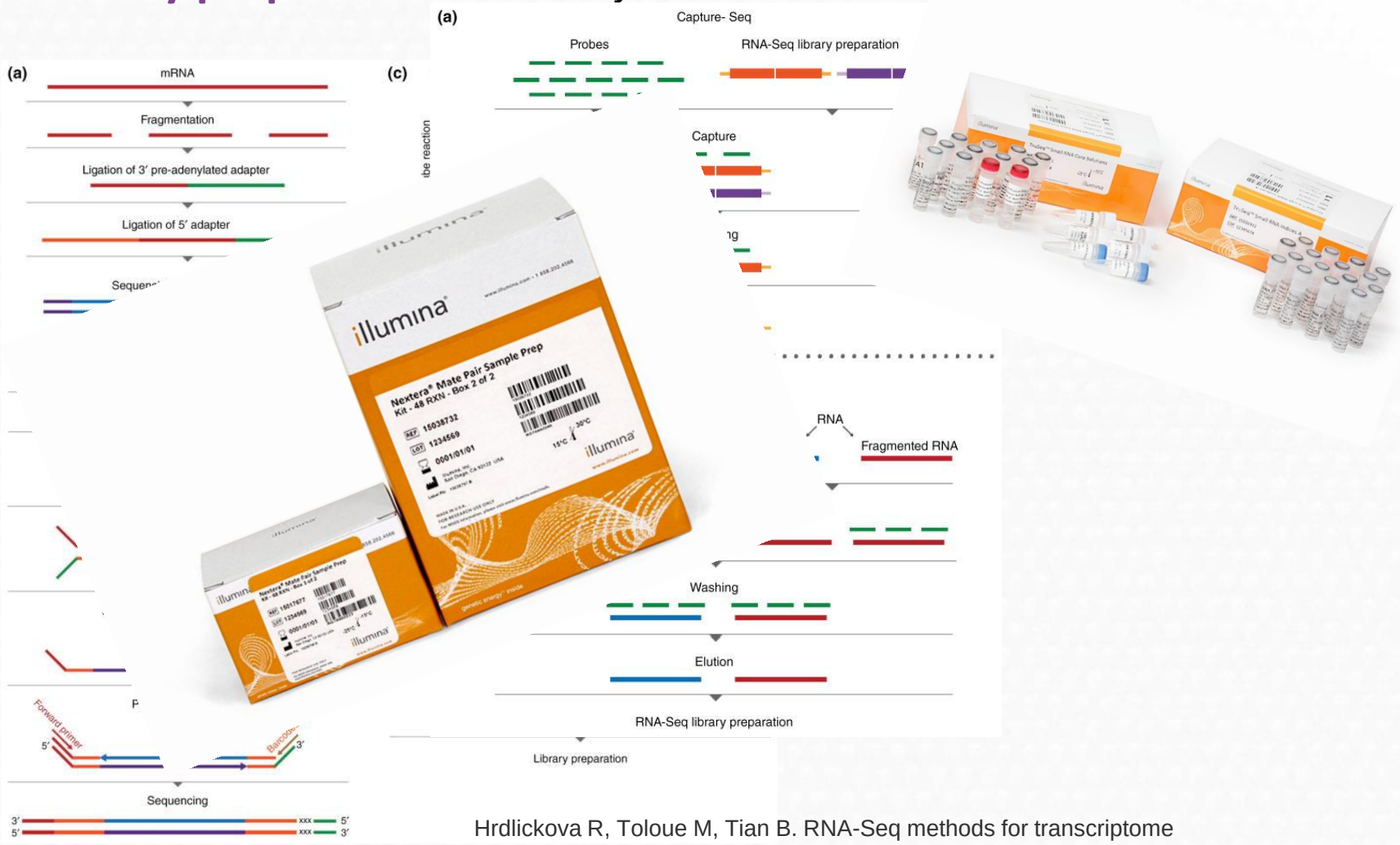




Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2016;8(1):10.1002/wrna.1364.

3. Second Generation Sequencing

Library preparation: Too many methods / kits av



Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2016;8(1):10.1002/wrna.1364.

3. Second Generation Sequencing

Library preparation: Too many methods / kits av

Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2016;8(1):10.1002/wrna.1364.

3. Second Generation Sequencing

Library preparation: Too many methods / kits av

Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2016;8(1):10.1002/wrna.1364.

3. Second Generation Sequencing

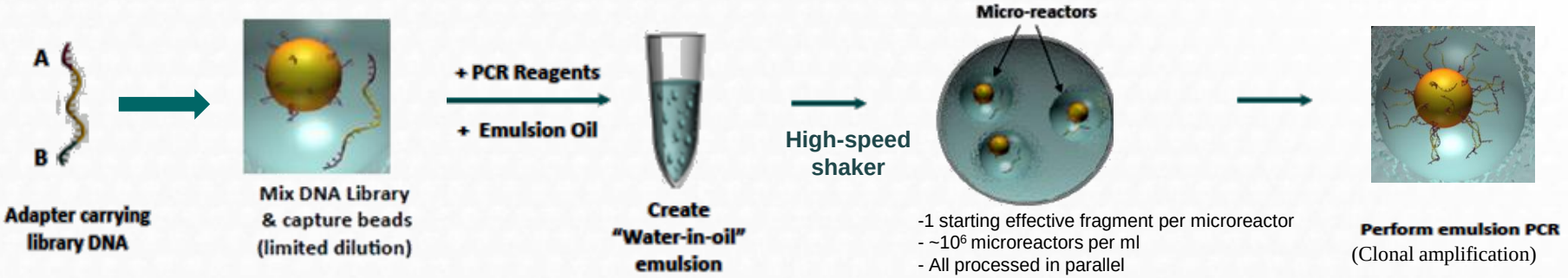
Library preparation: Too many methods / kits av

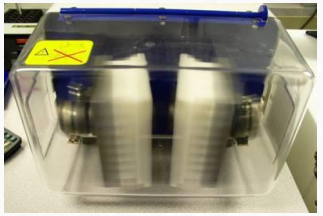
Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2016;8(1):10.1002/wrna.1364.

3. Second Generation Sequencing

Clonal amplification:

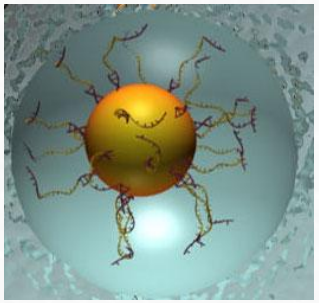
For emPCR based systems (Ion/PGM, 454)



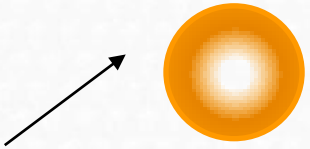


Clonal amplification:

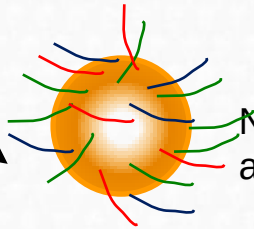
For emPCR based systems (Ion/PGM, 454)



Clonal amplification??



No empty beads



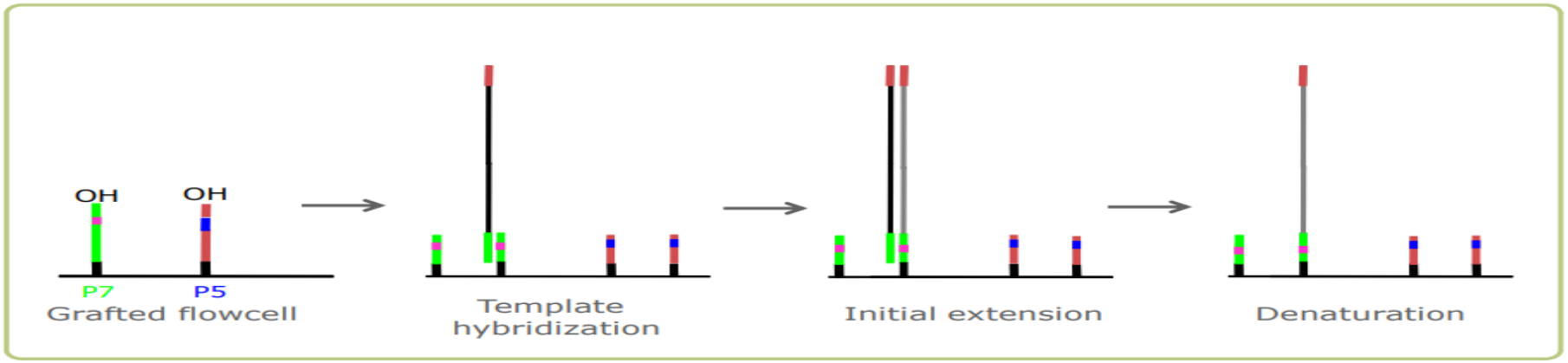
No beads containing more than one amplified fragment

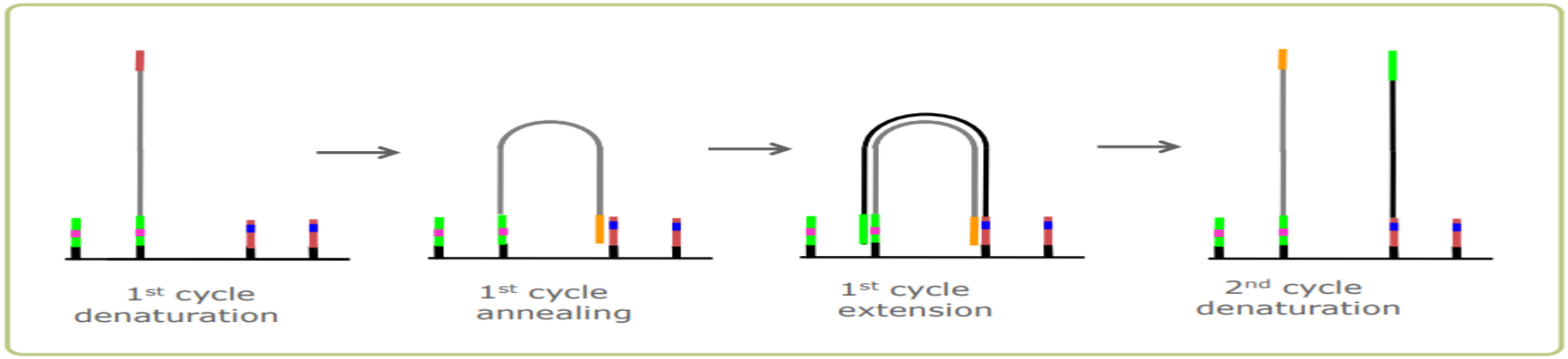
1) Titration: constant number of beads vs. different DNA starting quantities

2) Optimal enrichment: one single fragment amplified millions of times in one single bead

Vall d'Hebron
Institut de Recerca

Clonal amplification: Illumina

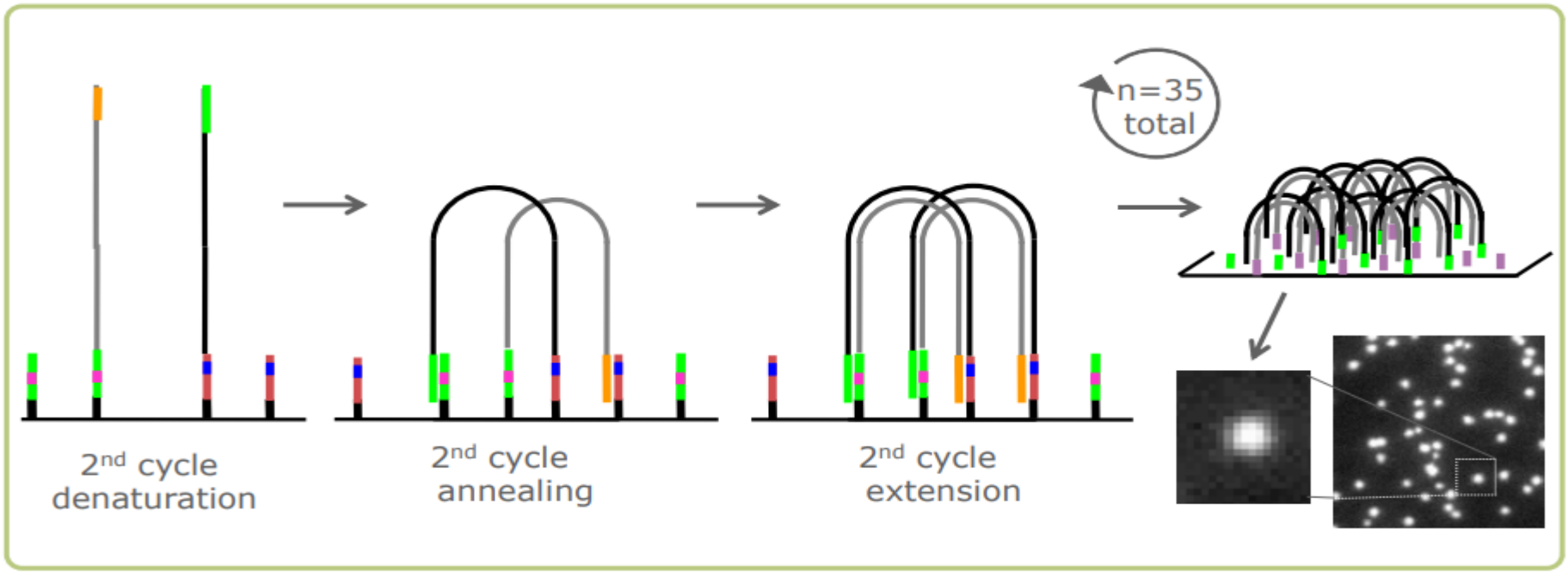




BROAD INSTITUTE illumina

3. Second Generation Sequencing

Clonal amplification: Illumina



Clonal amplification: Bridge PCR (Illumina):

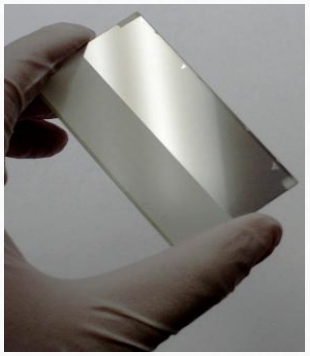
SYBR QC: Ensure successful amplification before continuing

- GOAL: Visually confirm successful cluster generation and optimal density before continuing

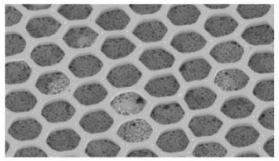
– Too Sparse: Loss of valuable real estate on flow cell
– Too Dense: Analysis problems

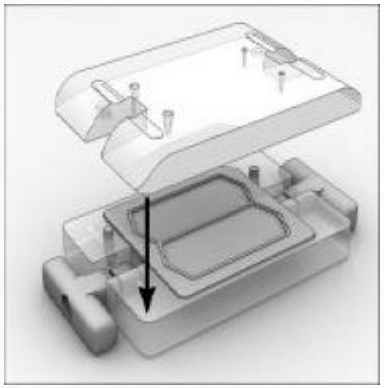
Vall d'Hebron
Institut de Recerca

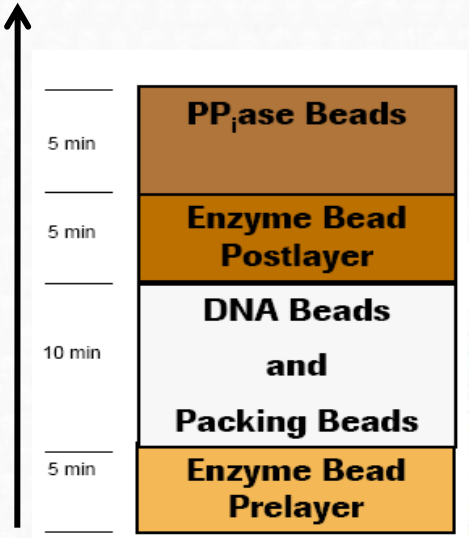
Sequencing: Roche 454 (pyrosequencing, sequencing by synthesis)

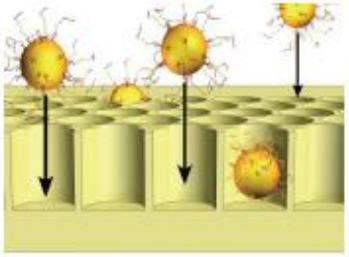


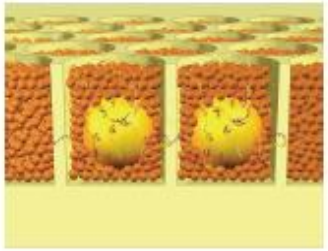
Metal coated PTP reduces crosstalk
29 µm well diameter (20/bead)
3,400,000 wells per PTP





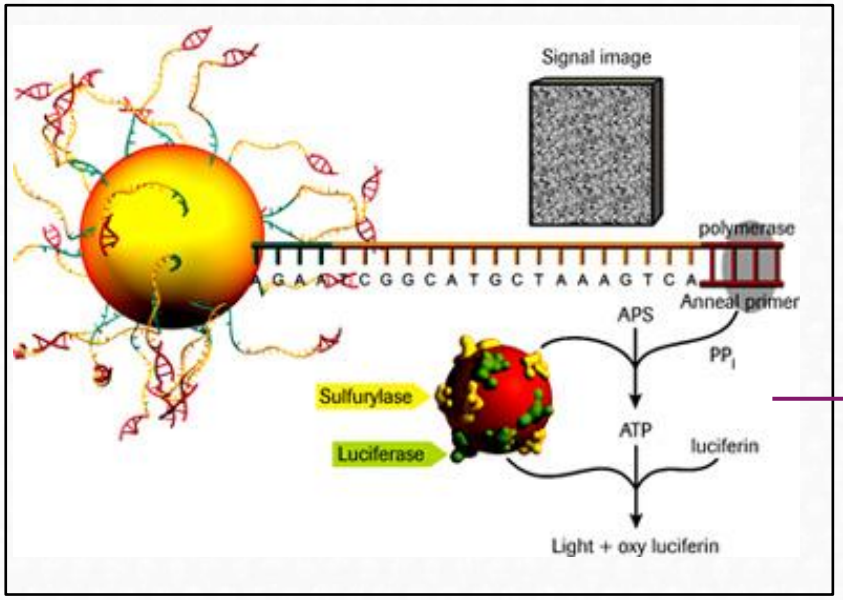




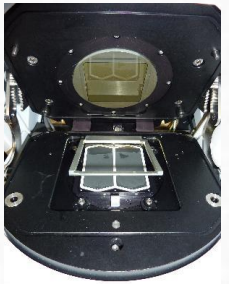


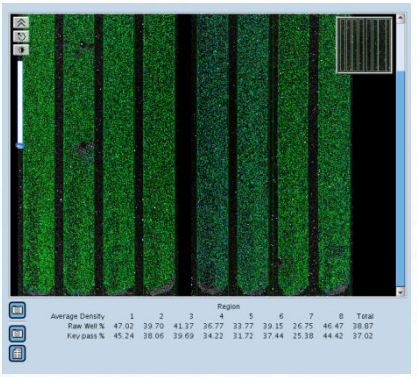
3. Second Generation Sequencing

Sequencing: Roche 454 (pyrosequencing, sequencing by synthesis)

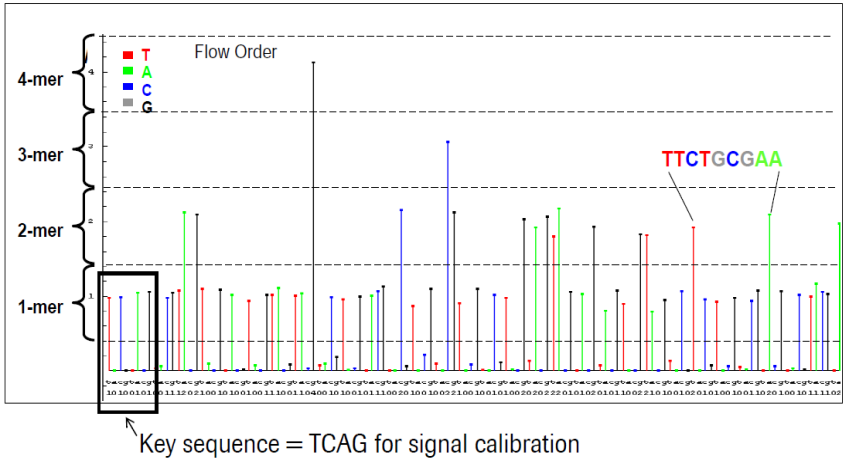


CCD Camera



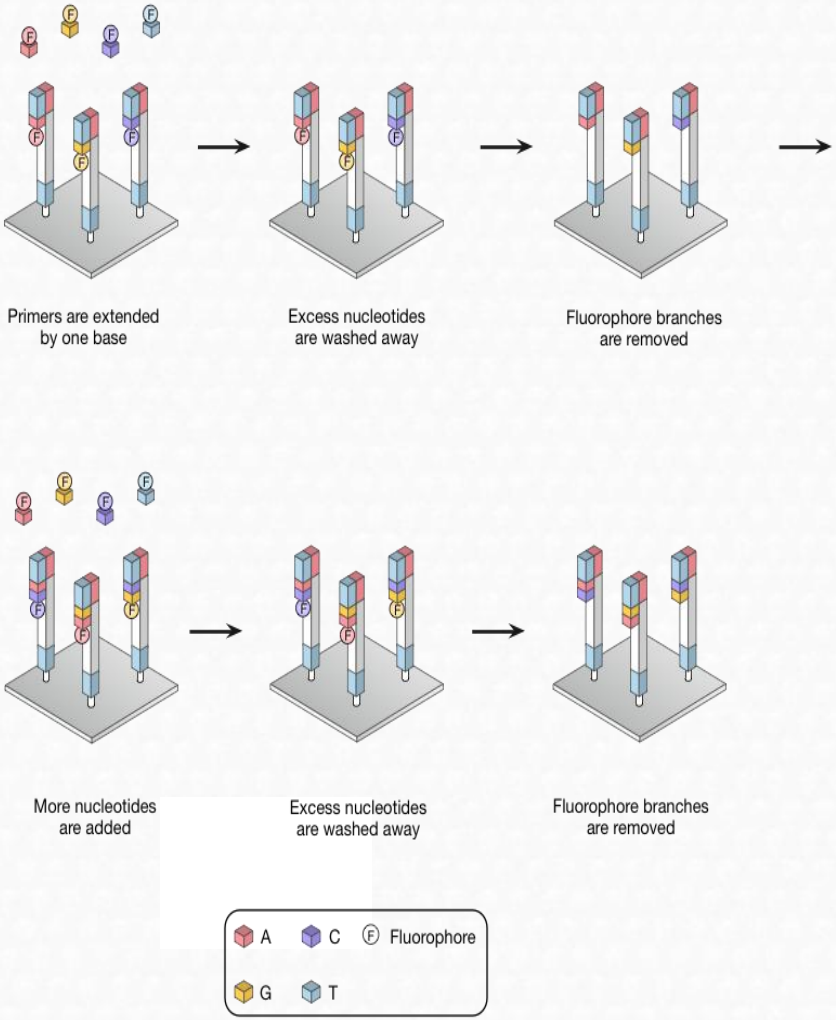


signal intensity is proportional to the number of nucleotides incorporated in the sequence



- throughput limited by the nº of wells
- errors in homopolymers (454)
- **long sequences** (up to 1000bp)
- **low throughput**, very expensive reagents
- required for some specific applications, advisable for others (*de novo* sequencing)

Sequencing: Illumina (dye terminator nt, sequencing by synthesis)



- Limited by the fragment length than can efectively "bridge"
- Labelled nucleotides are not incorporated as efficiently as native ones
- Short sequences
- Scalable set of machines suitable for nearly all the applications
- High throughput

3. Second Generation Sequencing

Illumina workflow.

https://www.youtube.com/watch?annotation_id=annotation_1533942809&feature=iv&src_vid=HMyCqWhwB8E&v=fCd6B5HRaZ8

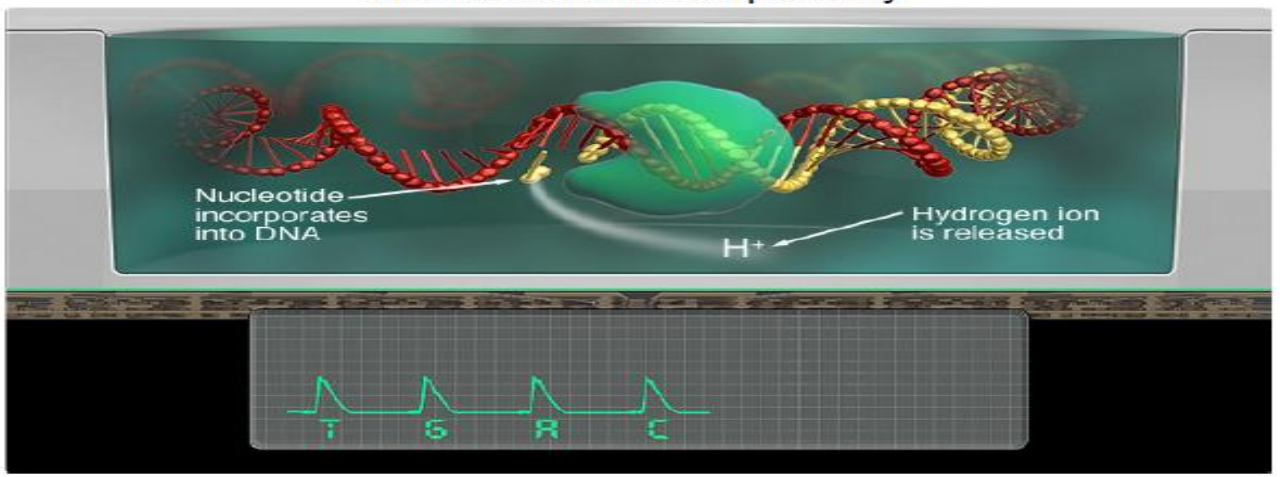
Figure 6: Sequencing Systems for Every Scale.

3. Second Generation Sequencing

Sequencing: Ion S5/PGM

Beads containing the clonally amplified library are loaded onto the chip.
The chip is run in the machine

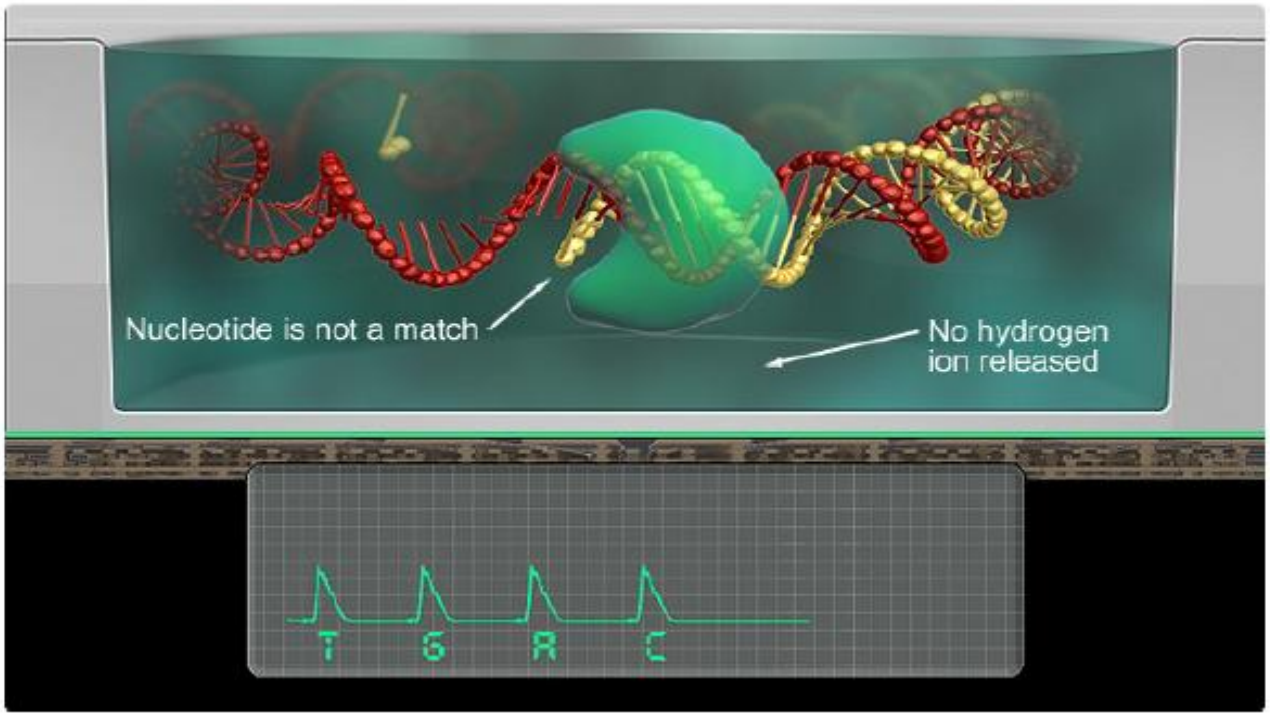
4 nucleotides flow sequentially



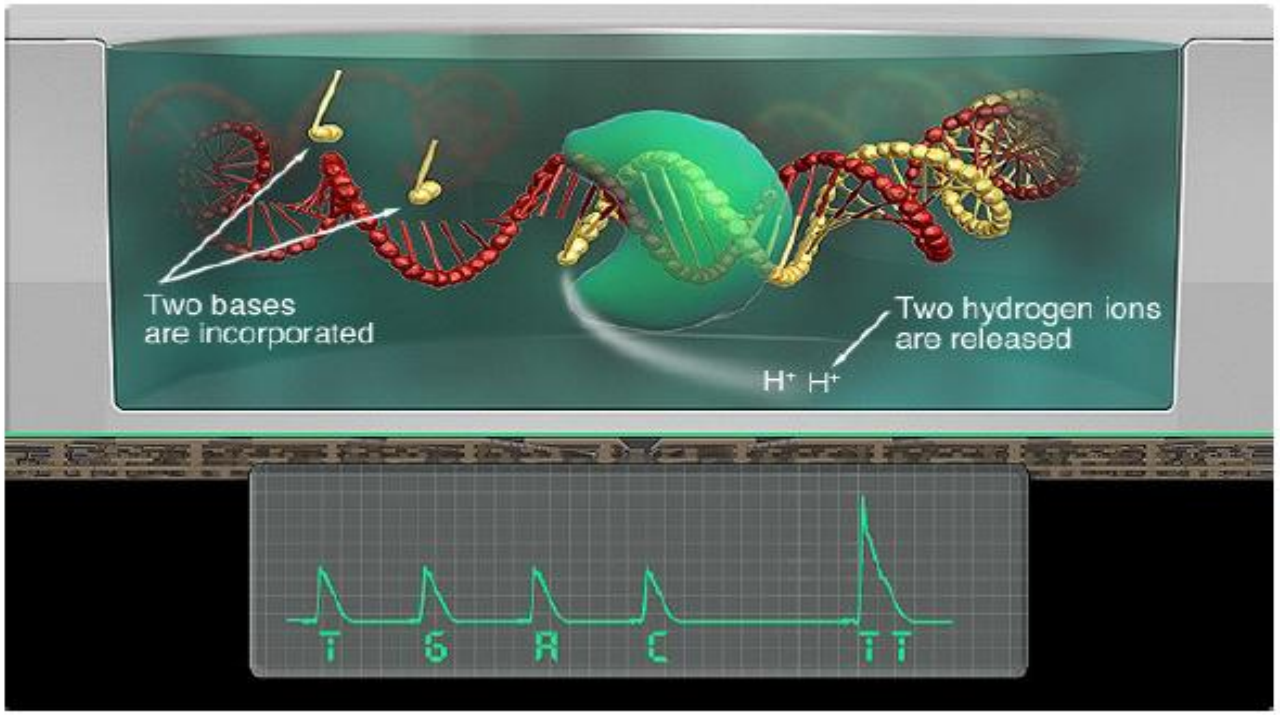
No camera, just a pH sensor

Vall d'Hebron
Institut de Recerca

Sequencing: Ion Torrent



Sequencing: Ion Torrent



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

3. Second Generation Sequencing

Small NGS platforms

GS Junior Plus (Roche)

Read Length: 700 pb
Output: 70 Mb
Running time: 18 hours

Ion Torrent PGM (LifeTechnol.)

Read length: 200 to 400 bp
Output: 30Mb – 2Gb
Running time: 2,3 - 4,4 hours

MiniSeq (Illumina)

Read Length: 2x150 bp
Output: 0.6 - 7.5 Gb
Running time: 4-24 hours

Applications:

- Amplicon sequencing
- Targeted sequencing (DNA / RNA)
- Metagenomics (16S)

Source: Supplier webpage

3. Second Generation Sequencing

High throughput NGS Platforms

GS FLX+

Read length: 700 bp
Output: 1 Mb
Running time: 23 hours

Ion Gene Studio S5 Prime System

Read length: Up to 200 bp
Output: 50Gb/day (2 chips)
Run time: 8,5 hours

NovaSeq 6000 System

Read length: 2x150 bp
Output: 6.000 Gb
Run time: 44 hours

Applications:

- Amplicon sequencing
- Targeted sequencing
- Metagenomics (16S)
- Genomes
- Exomes, transcriptome

Source: Supplier webpage

3. Second Generation Sequencing

High throughput NGS Platforms




	iSeq 100 System	MiniSeq System	MiSeq Series ⊕	NextSeq Series ⊕
Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)				●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●
Exome Sequencing				●
Targeted Gene Sequencing (amplicon, gene panel)	●	●	●	●
Whole-Transcriptome Sequencing				●
Gene Expression Profiling with mRNA-Seq				●
Targeted Gene Expression Profiling	●	●	●	
Long-Range Amplicon Sequencing*	●	●	●	
miRNA & Small RNA Analysis	●	●	●	●
DNA-Protein Interaction Analysis			●	●
Methylation Sequencing				●
16S Metagenomic Sequencing		●	●	●

Source: Supplier webpage

3. Second Generation Sequencing

High throughput NGS Platforms

NextSeq Series ✪

HiSeq 4000 System

HiSeq X Series‡

NovaSeq 6000 System

Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)	●	●	●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●		●
Exome Sequencing	●	●		●
Targeted Gene Sequencing (amplicon, gene panel)	●	●		●
Whole-Transcriptome Sequencing	●	●		●
Gene Expression Profiling with mRNA-Seq	●	●		●
miRNA & Small RNA Analysis	●	●		●
DNA-Protein Interaction Analysis	●	●		●
Methylation Sequencing	●	●		●
Shotgun Metagenomics	●	●		●

Source: Supplier webpage

3. Second Generation Sequencing

High throughput NGS Platforms

Ion GeneStudio S5 Prime System

Turnaround time: 6.5 hr*

Targeted DNA sequencing

Targeted RNA sequencing

Microbial sequencing

Simplest & fastest* workflow

Single day workflow from sample to annotated variants for gene panel sequencing featuring Ion AmpliSeq™ target technology, Ion PGM™ System, and the automated Ion Chef™ System**.

Most accurate for multiple gene panels

Up to 100% sensitivity for multiple gene panels, with Torrent Suite™ software and an improved variant calling algorithm that provides high-quality consensus accuracy for SNP detection.

Cost to buy & run

Affordable sequencing with Ion PGM™ v2 chips that dramatically reduce cost per sample and the Ion PGM™ System that is a fraction of the cost of the alternative.

	Ion 510™ Chip	Ion 520™ Chip	Ion 530™ Chip	Ion 540™ Chip	Ion 550™ Chip
Max. output (reads)	3 M	6 M	20 M	80 M	130 M
Targeted DNA sequencing ** e.g., Ion Torrent™ Oncomine™ Focus Assay	•	•	•	•	•
Small genome sequencing† e.g., Bacterial typing using Ion Xpress™ Plus Fragment Library Kit		•	•		
16S metagenomics sequencing†† e.g., Ion 16S™ Metagenomics Kit		•	•		
Exome sequencing e.g., Ion AmpliSeq™ Exome Panel				•	•
Targeted RNA sequencing e.g., Ion AmpliSeq™ made-to-order RNA panels	•	•	•	•	•
miRNA/small RNA profiling e.g., Ion Total RNA-Seq v2 Kit	•	•	•		
Targeted transcriptome sequencing e.g., Ion AmpliSeq™ Transcriptome Human Gene Expression Kit				•	•
Whole transcriptome sequencing e.g., Ion Total RNA-Seq v2 Kit				•	•
Low-pass whole genome sequencing (PGS) e.g., Ion ReproSeq™ PGS Kit	•	•	•		

Five Ion Torrent™ sequencing chips achieve 2–130 M reads per run (or 2–260 M reads per day) to enable a broad range of sequencing applications.

Source: Supplier webpage

Vall d'Hebron
Institut de Recerca

1. Introduction to NGS
2. First Generation Sequencing
3. Second Generation Sequencing
4. Third Generation Sequencing
5. Sequencing generation face to face
6. Applications of NGS techniques
7. A (very) brief introduction to DoE

Vall d'Hebron
Institut de Recerca

Single molecule sequencing

Advantages:

- Less sample preparation (no PCR)
- No amplification
  - ✓No PCR errors
  - ✓Fewer contamination issues
  - ✓No GC-bias
  - ✓Analyze every sample (unPCRable, unclonable)
  - ✓Analyze low quality DNA (forensics samples, archeological)
- Absolute quantification
- Sequence RNA directly

Vall d'Hebron
Institut de Recerca

Helicos Genetic Analysis system

Workflow similar to Illumina, but without bridge amplification:

- relative slow and expensive
- short reads

	Helicos
Read Length	35 bp
Throughput	35 Gb
Reads per run	600,000,000 - 1,000,000,000
Accuracy	97 %
Run Time	8 days

SMRT (Pacific bioscience)

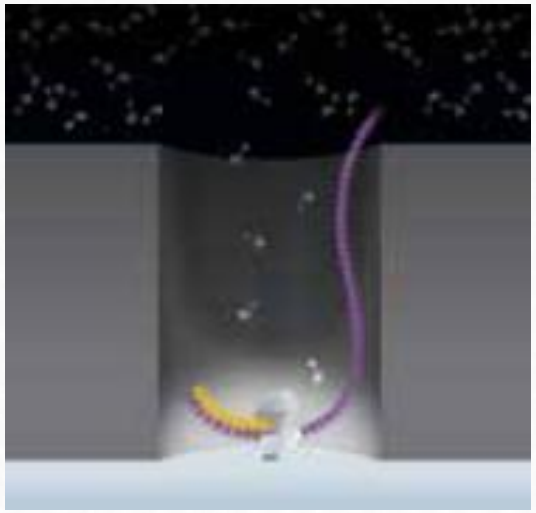
- **Quick** library construction (6 hours)
- Run time: from 30 min to 6 hours
- Sequencing by synthesis
- No library amplification
- **Long reads:** 5 Kb – 20 Kb
- **High error rate** (but improving!)

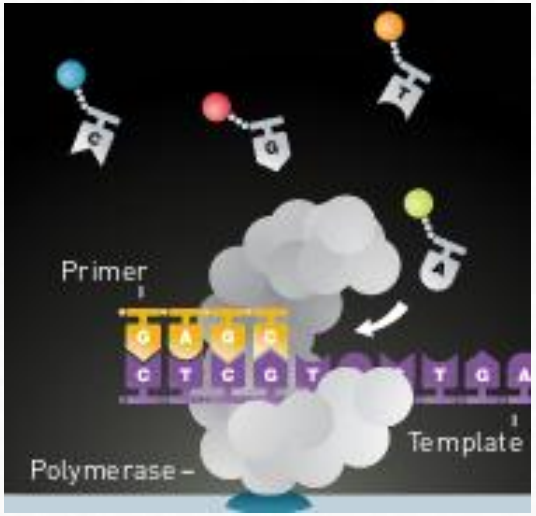
Read Lengths	20 kb
Ouput:	750 Mb – 1.25 Gb per SMRT Cell
Nº SMRT Cells:	1-16
Run Time:	6 hours
Error rate	10-15%

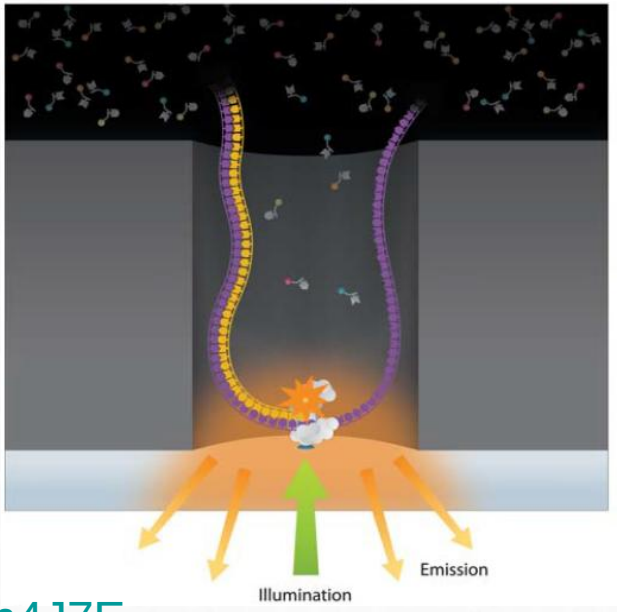
Source: Supplier webpage

4. Third Generation Sequencing

SMRT (Pacific bioscience)





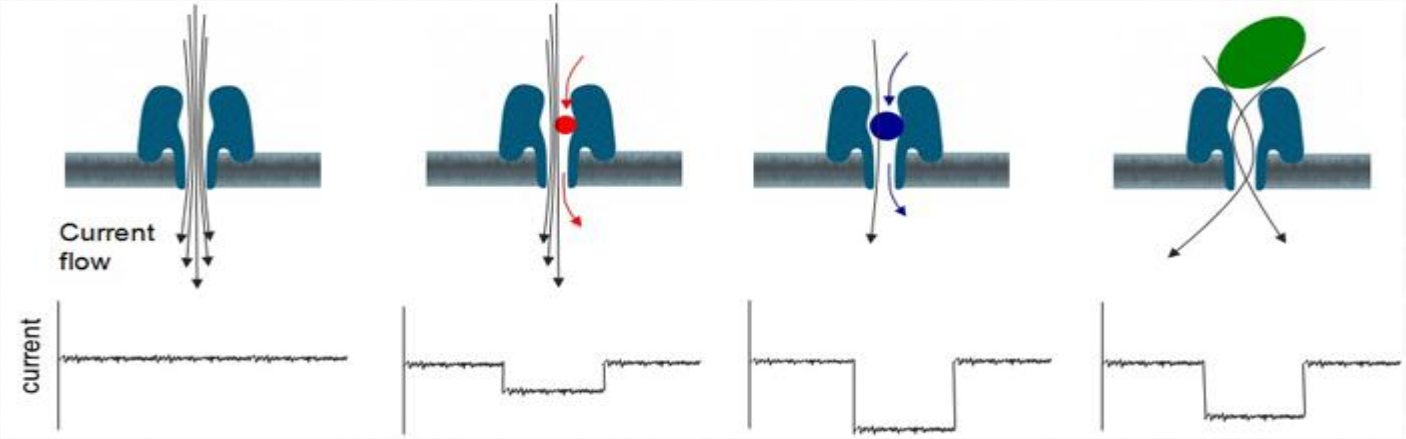


https://www.youtube.com/watch?v=GX6RSKh4J7E

4. Third Generation Sequencing

Oxford nanopore

- alpha-hemolysin
- Heptameric protein with a pore of inner diameter 1nm
- Pore diameter same scale of DNA
- Protein nanopores can be adapted.
- The company has optimised its large-scale production.
- DNA, RNA, miRNA and protein analysis
- Changes in membrane voltage are measured



4. Third Generation Sequencing

Oxford nanopore

SmidgION

MinION

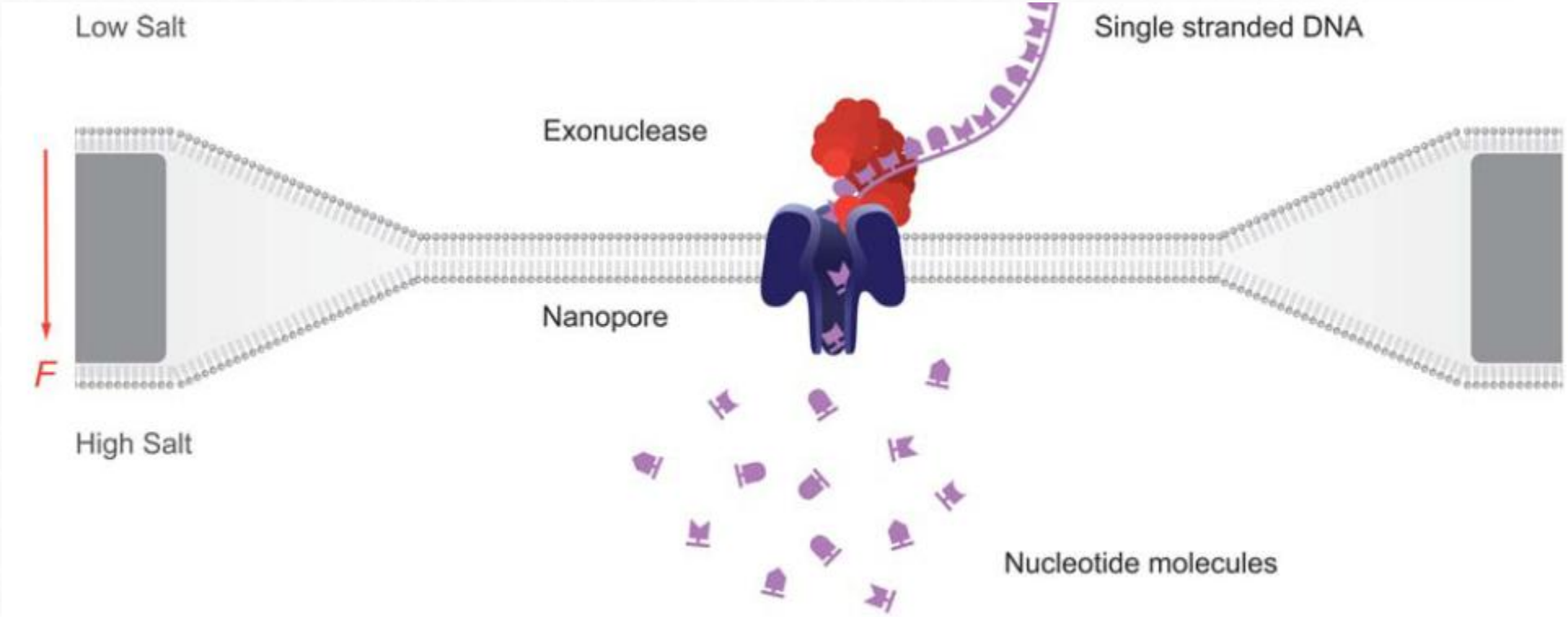
GridION

PromethION

- Simple 10 minute sample prep available
- Read length: Ultra long reads (2 Mb)
- Very fast, but high error rates.
- Easy to work in the field (Ebola viruses sequenced in Guinea 2 days after sample collection, Quick J, 2016)

4. Third Generation Sequencing

Oxford nanopore

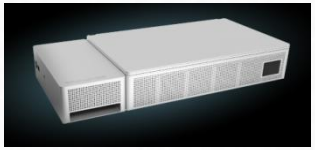


Human Molecular Genetics, 2010, Vol. 19, Review Issue 2

https://www.youtube.com/watch?v=GUb1TZvMWsw

Vall d'Hebron Institut de Recerca

3rd generation instruments

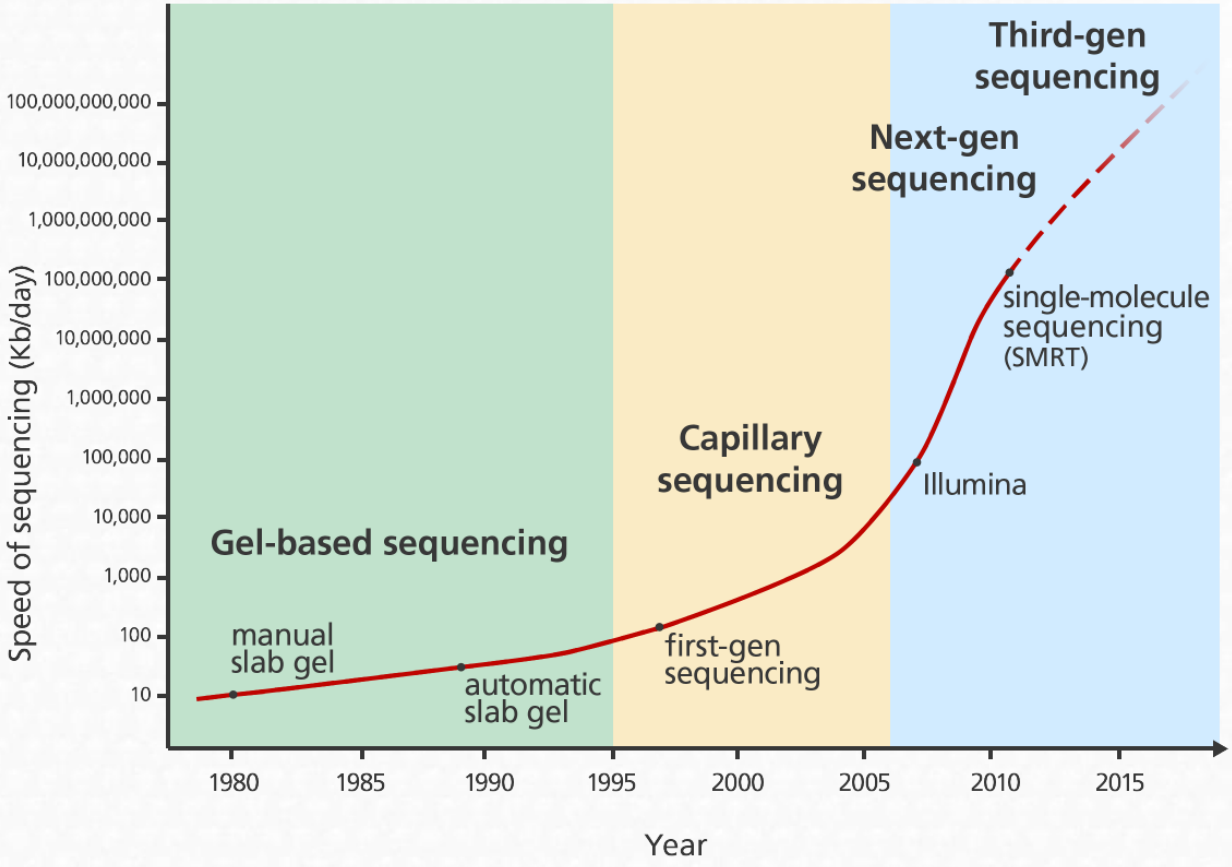
	High througput	Benchtop
Pacific Bioscience	 Sequel system	
Oxford Nanopore Technologies	 PromethION  GridION	 minION

Vall d'Hebron
Institut de Recerca

1. Introduction to NGS
2. First Generation Sequencing
3. Second Generation Sequencing
4. Third Generation Sequencing
5. Sequencing generation face to face
6. Applications of NGS techniques
7. A (very) brief introduction to DoE

5. Sequencing Generation face to face

Speed variation by technology



http://www.yourgenome.org/stories/third-generation-sequencing

5. Sequencing Generation face to face

Vall d'Hebron
Institut de Recerca

Table 1. Comparison of first-generation sequencing, SGS and TGS

	First generation	Second generation[a]	Third generation[a]
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

Human Molecular Genetics, 2010, Vol. 19, Review Issue 2

5. Sequencing Generation face to face

Vall d'Hebron
Institut de Recerca

Table 1. Comparison of first-generation sequencing, SGS and TGS

	First generation	Second generation[a]	Third generation[a]
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

Human Molecular Genetics, 2010, Vol. 19, Review Issue 2

5. Sequencing Generation face to face

Vall d'Hebron
Institut de Recerca

Table 1. Comparison of first-generation sequencing, SGS and TGS

	First generation	Second generation[a]	Third generation[a]
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

Human Molecular Genetics, 2010, Vol. 19, Review Issue 2

5. Sequencing Generation face to face

Vall d'Hebron
Institut de Recerca

Table 1. Comparison of first-generation sequencing, SGS and TGS

	First generation	Second generation[a]	Third generation[a]
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

Human Molecular Genetics, 2010, Vol. 19, Review Issue 2

5. Sequencing Generation face to face

Vall d'Hebron
Institut de Recerca

Table 1. Comparison of first-generation sequencing, SGS and TGS

	First generation	Second generation[a]	Third generation[a]
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

Human Molecular Genetics, 2010, Vol. 19, Review Issue 2

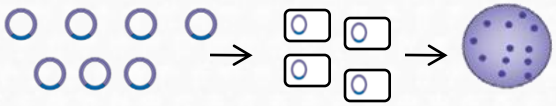
5.Sequencing Generation face to face

Vall d'Hebron
Institut de Recerca

FGS

1. DNA fragmentation.

2. Vector cloning, bacterial transformation and growth, DNA isolation and purification

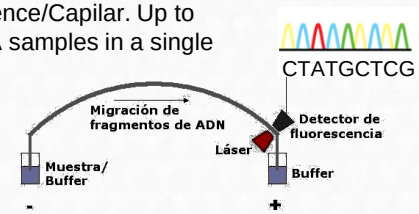


3. Sequencing (chain-terminating ddnucleotides)

ssDNA
DNA primer
Polimerase
dNTPs
Fluorescent ddNTPs

4. Image processing

Capillary electrophoresis (1 Sequence/Capilar. Up to 384 DNA samples in a single run)

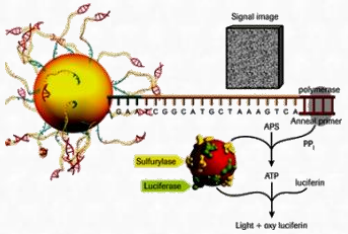


SGS

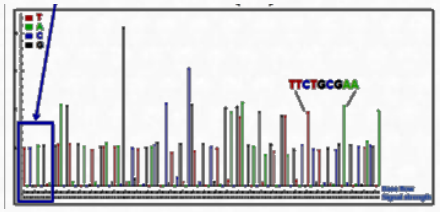
1. DNA fragmentation

2. In vitro adaptor ligation + clonal amplification

3. Massive paralel sequencing



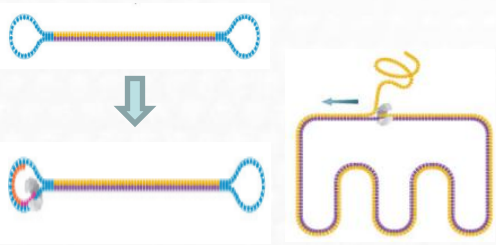
4. Image processing and data analysis



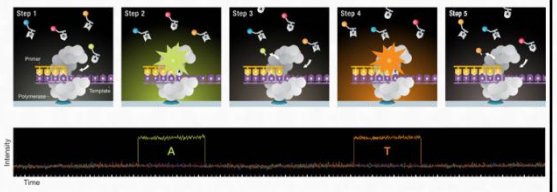
TGS

1. DNA fragmentation

2. y 3. in vitro adaptor ligation. NO AMPLIFICATION.
Massive parallel sequencing.



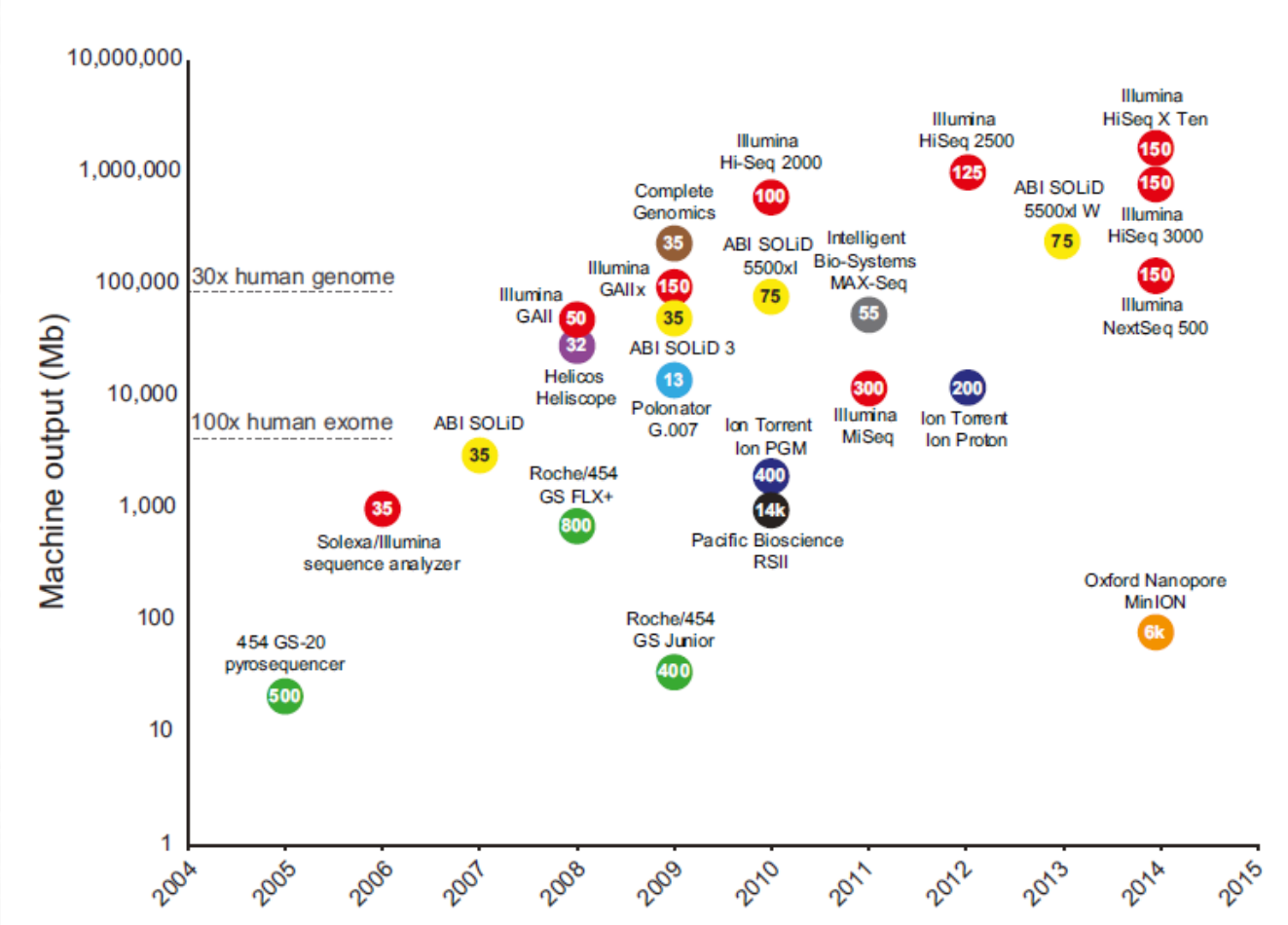
4. Image processing and data analysis.



5.Sequencing Generation face to face

Vall d'Hebron
Institut de Recerca

Release dates vs Machine outputs per run



Molecular Cell *58*, May 21, 2015 ©2015 Elsevier Inc.

Vall d'Hebron
Institut de Recerca

1. Introduction to NGS
2. First Generation Sequencing
3. Second Generation Sequencing
4. Third Generation Sequencing
5. Sequencing generation face to face
6. Applications of NGS techniques
7. A (very) brief introduction to DoE

6. Applications of NGS techniques

6. Applications of NGS techniques

Vall d'Hebron
Institut de Recerca

Table 1. Selected HTS Methods

Method	Purpose	Reference
RNA-seq	Transcript analysis	Nagalakshmi et al., 2008
Global run-on sequencing (GRO-seq)	Transcription	Core et al., 2008
Nascent-seq	Transcription	Khodor et al., 2011
Native elongating transcript sequencing (NET-seq)	Transcription	Churchman and Weissman, 2011
Ribo-seq	Translation	Ingolia et al., 2009
Replication sequencing (Repli-seq)	Replication	Hansen et al., 2010
Hi-C	Chromatin conformation	Lieberman-Aiden et al., 2009
Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)	Chromatin conformation	Fullwood et al., 2009
Chromosome conformation capture carbon copy (5-C)	Chromatin conformation	Dostie et al., 2006
Chromatin isolation by RNA purification sequencing (ChIRP-seq)	Genome localization	Chu et al., 2011
Reduced representation bisulphite sequencing (RRBS-seq)	Genome methylation	Meissner et al., 2008
Bisulfite sequencing (BS-seq)	Genome methylation	Cokus et al., 2008
DNAse-seq	Open chromatin	Crawford et al., 2006
Assay for transposase-accessible chromatin using sequencing (ATAC-seq)	Open chromatin	Buenrostro et al., 2013
Parallel Analysis of RNA structure (PARS)	RNA structure	Kertesz et al., 2010
Structure-seq	RNA structure	Ding et al., 2014
RNA on a massively parallel array (RNA-MaP)	RNA-protein interactions	Buenrostro et al., 2014
RNA immunoprecipitation sequencing (RIP-seq)	RNA-protein interactions	Sephton et al., 2011
Parallel analysis of RNA ends sequencing (PARE-seq)	microRNA target discovery	German et al., 2008
Massively parallel functional dissection sequencing (MPFD)	Enhancer assay	Patwardhan et al., 2012

Molecular Cell 58, May 21, 2015 ©2015 Elsevier Inc.

6. Applications of NGS techniques

Vall d'Hebron
Institut de Recerca

Table 2
Recommendation for data requirements for a selection of NGS applications.

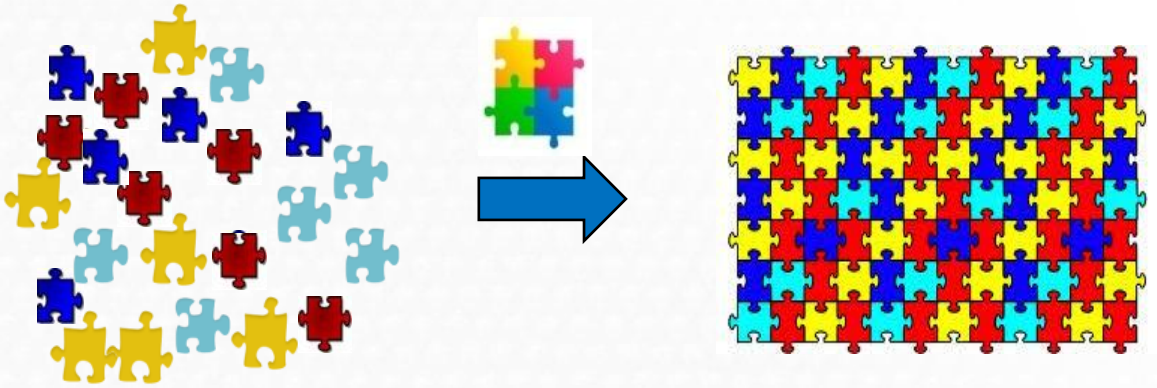
Application	# reads/sample	Run type	# read length (bp)	Remark
Transcriptome analysis				
Tag based (SAGE/CAGE)	>10 million	Single end	20–50	
SmallRNA	>10 million	Single end	20–50	
mRNA Seq	>30 million	Paired-end	>50	Efficient exclusion of rRNA derived sequences increases the resolution of the transcripts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

H.P.J. Buermans, J.T. den Dunnen / Biochimica et Biophysica Acta 1842 (2014) 1932–1941

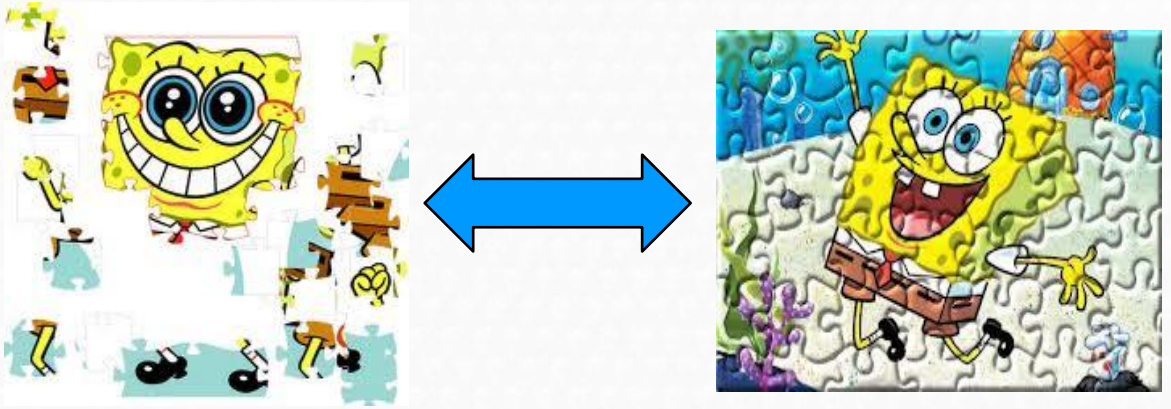
6. Applications of NGS techniques

Whole Genome sequencing

***De novo* sequencing**



Resequencing



6. Applications of NGS techniques

Vall d'Hebron
Institut de Recerca

Whole Genome sequencing

- Complete characterization of the entire genome.
- Beijing Genomics Institute
  - "Does it look cute, we'll sequence it"
  - "Does it taste good, we'll sequence it"
- GWAS studies
  - microarray-based: 2 million markers
  - Now sequence the whole genome: 3.2 billion bases
- It can be applied for plant, microbial…
- The rapid drop in sequencing costs allow researchers to sequence a genome quickly.

Vall d'Hebron
Institut de Recerca

Whole Genome sequencing

PROS

- Global genome picture, no systematic missing of information
- Useful for diseases that involve multiple genetic phenomena

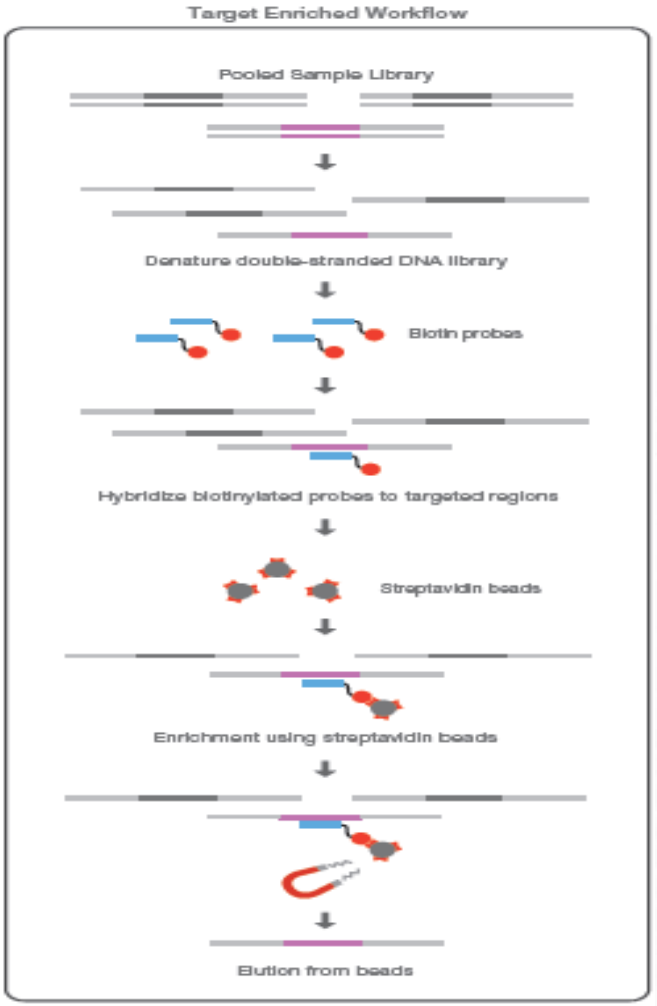
CONS:

- Can miss variants in the exonics regions due to **lower coverage**
- some regions cannot be sequenced/assembled (repetitive and GC rich regions)
- More expensive and time consuming

6. Applications of NGS techniques

Targeted sequencing

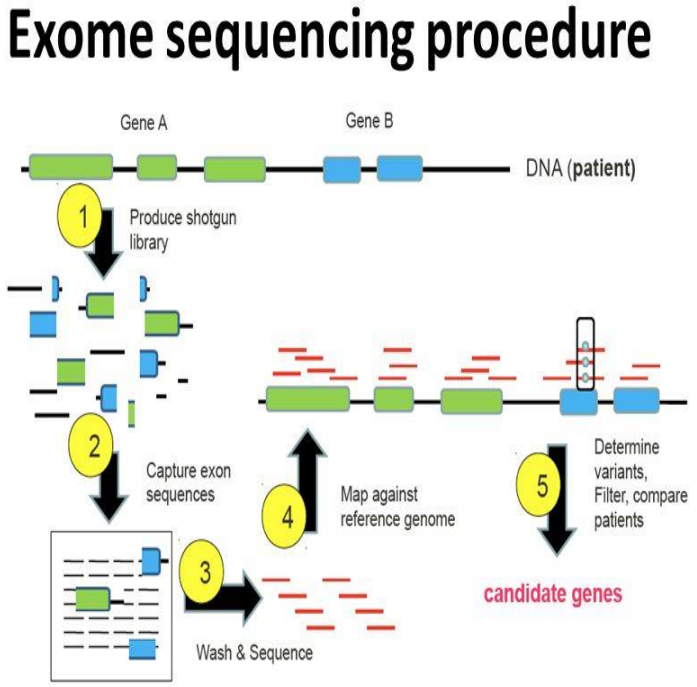
- Only a **subset of genes** or regions of the genome are isolated and sequenced.
- It allows to focus times, expenses and data analysis on specific areas of interest.
- Enables sequencing at **much higher coverage** levels.
- Target sequencing panels can be purchased with preselected content or can be custom designed.



6. Applications of NGS techniques

Exome sequencing

- Identifies **variants** across a wide range of applications
- Achieves comprehensive coverage of coding regions
- Provides a cost-effective alternative to whole-genome sequencing (4–5 Gb of sequencing per exome compared to ~90 Gb per whole human genome)
- Produces a smaller, more manageable data set for faster, easier analysis compared to whole-genome approaches



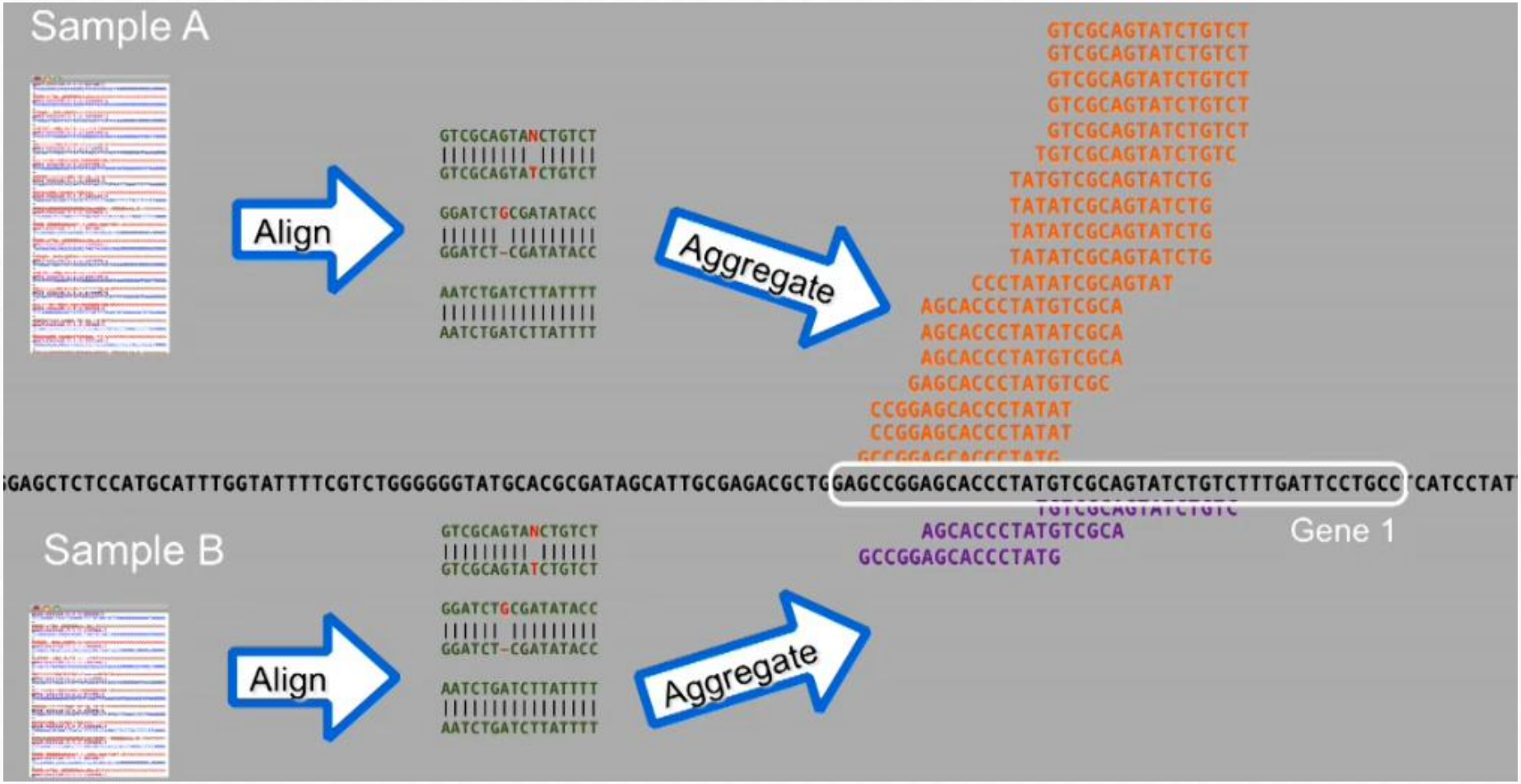
Vall d'Hebron
Institut de Recerca

RNA-seq Expression Analysis

- RNA-seq works by sequencing every RNA molecule and profiling the expression of a particular gene by counting the number of time its transcripts have been sequenced.
- Since now, performed with microarrays technology
- Sensitivity of sequence based studies is limited by the depth of sequencing
- Applications:
  - Differential gene expression analysis (DGE)
  - splice variants (resolution at base-level)
  - Detection of novel transcripts and isoforms
  - Detection of allele specific expression patterns

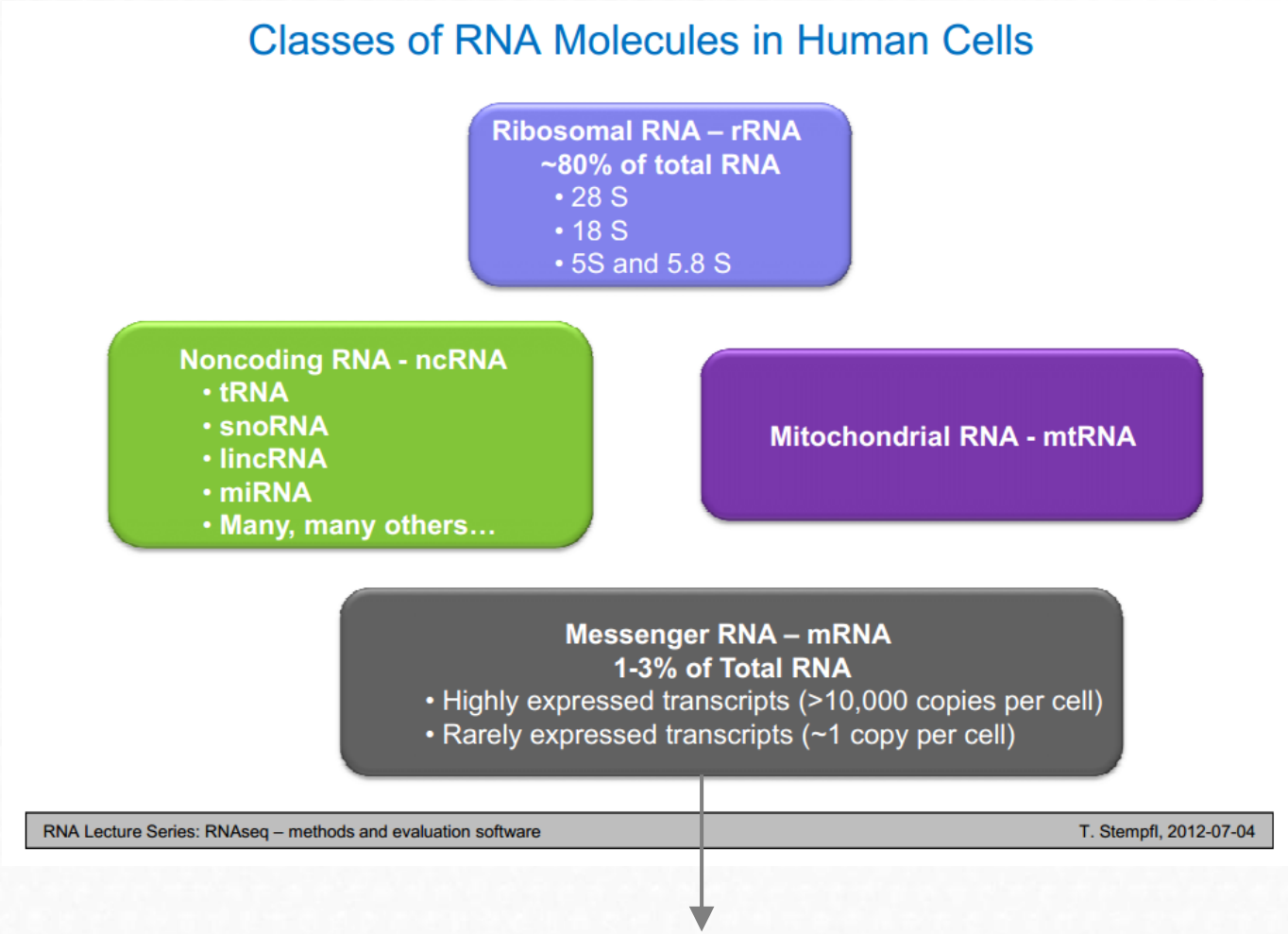
6. Applications of NGS techniques

RNA-seq Expression Analysis



Adapted from R.irizarry

6. Applications of NGS techniques

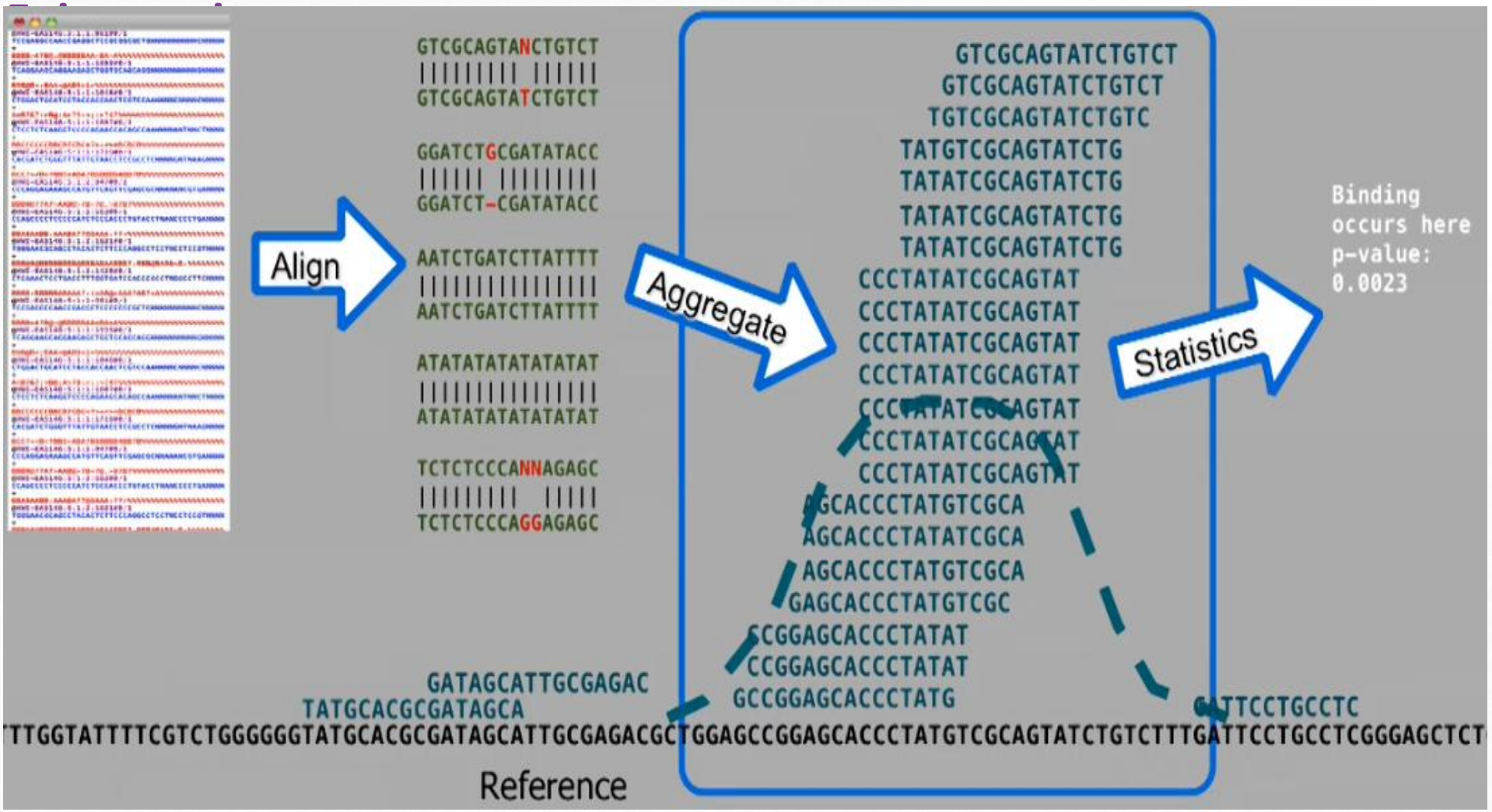


Very high dynamic range (10^5 to 10^7)

ChIP-seq

- analyze protein interactions with DNA: Transcription factors
- ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins
- used to map global binding sites precisely for any protein of interest

6. Applications of NGS techniques



Vall d'Hebron
Institut de Recerca

Single cell RNA-seq (scRNA-seq)

- Allows comparison of the transcriptomes of individual cells
- Results depend of protocol used
- Special attention to single cell purification (nuclei RNA-seq)
- Low signal of weak expressed genes
- Required number of cells increases with the complexity of the sample under investigation
- Requires specific bioinformatic approaches.

Single cell RNA-seq (scRNA-seq)

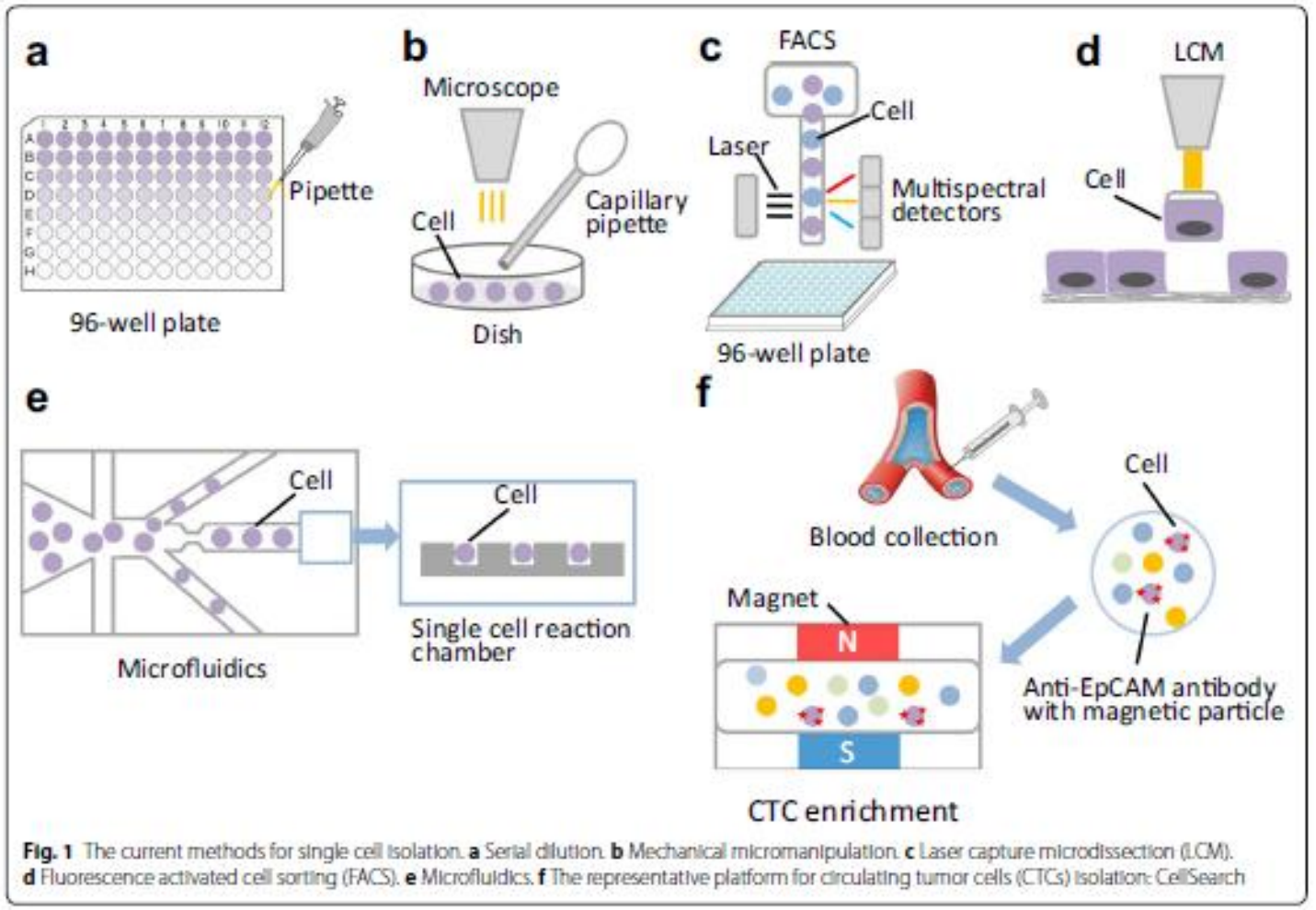


Fig. 1 The current methods for single cell isolation. **a** Serial dilution. **b** Mechanical micromanipulation. **c** Laser capture microdissection (LCM). **d** Fluorescence activated cell sorting (FACS). **e** Microfluidics. **f** The representative platform for circulating tumor cells (CTCs) isolation: CellSearch

Wang and Song *Clin Trans Med* (2017) 6:10
DOI 10.1186/s40169-017-0139-4

Vall d'Hebron
Institut de Recerca

Single cell RNA-seq (scRNA-seq)

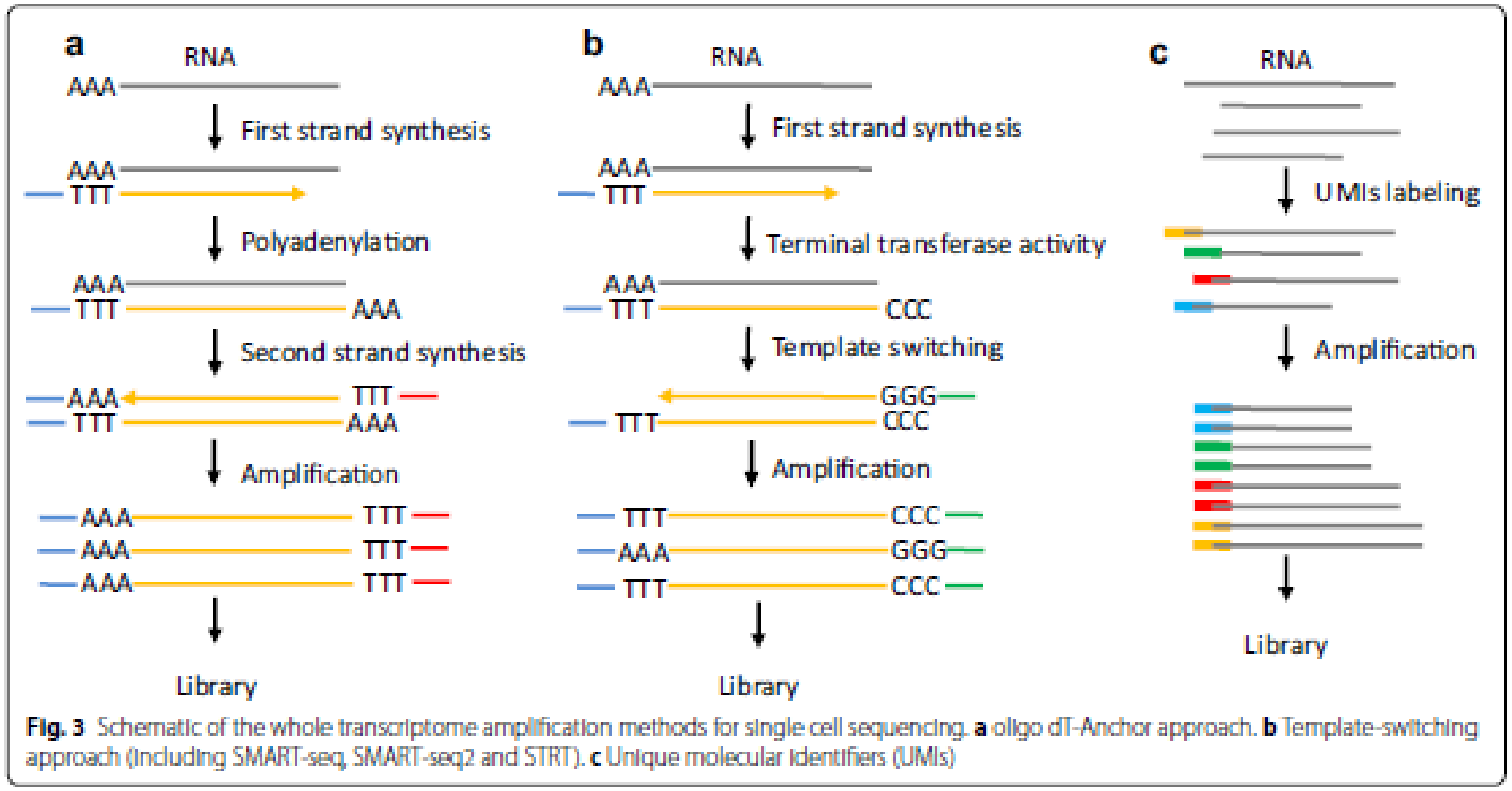


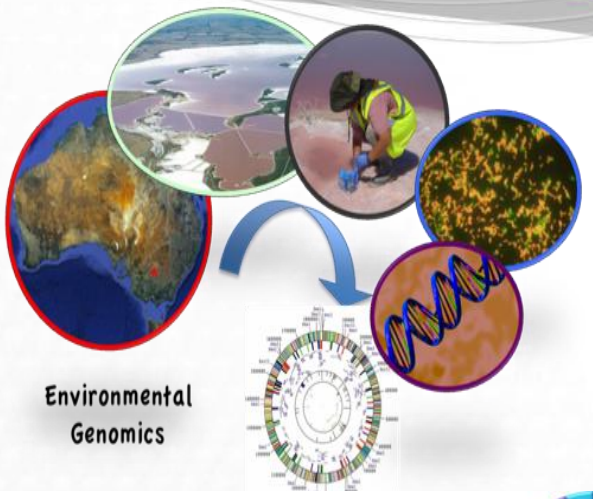
Fig. 3 Schematic of the whole transcriptome amplification methods for single cell sequencing. **a** oligo dT-Anchor approach. **b** Template-switching approach (including SMART-seq, SMART-seq2 and STRT). **c** Unique molecular identifiers (UMIs)

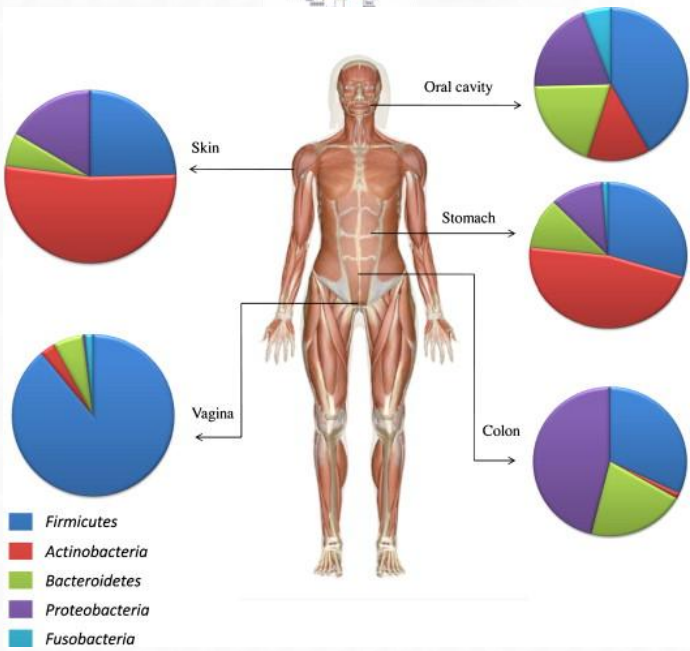
Wang and Song *Clin Trans Med (2017) 6:10*
DOI 10.1186/s40169-017-0139-4

6. Applications of NGS techniques

Metagenomics

- Is a way to make an inventory of what (DNA) is present in a sample
- Two approaches:
  - Sequence it all
  - Focus on specific conserved sequences (ribosomal genes)
- Technology facilitates the study of the consequences of environmental changes and the causes of the changes.





6. Applications of NGS techniques

Vall d'Hebron
Institut de Recerca

Towards a genomics-informed, real-time, global pathogen surveillance system

Jennifer L. Gardy[1,2] *and Nicholas J. Loman*[3]

Abstract | The recent Ebola and Zika epidemics demonstrate the need for the continuous surveillance, rapid diagnosis and real-time tracking of emerging infectious diseases. Fast, affordable sequencing of pathogen genomes — now a staple of the public health microbiology laboratory in well-resourced settings — can affect each of these areas. Coupling genomic diagnostics and epidemiology to innovative digital disease detection platforms raises the possibility of an open, global, digital pathogen surveillance system. When informed by a One Health approach, in which human, animal and environmental health are considered together, such a genomics-based system has profound potential to improve public health in settings lacking robust laboratory capacity.

doi:10.1038/nrg.2017.88
Published online 13 Nov 2017
NATURE REVIEWS | GENETICS

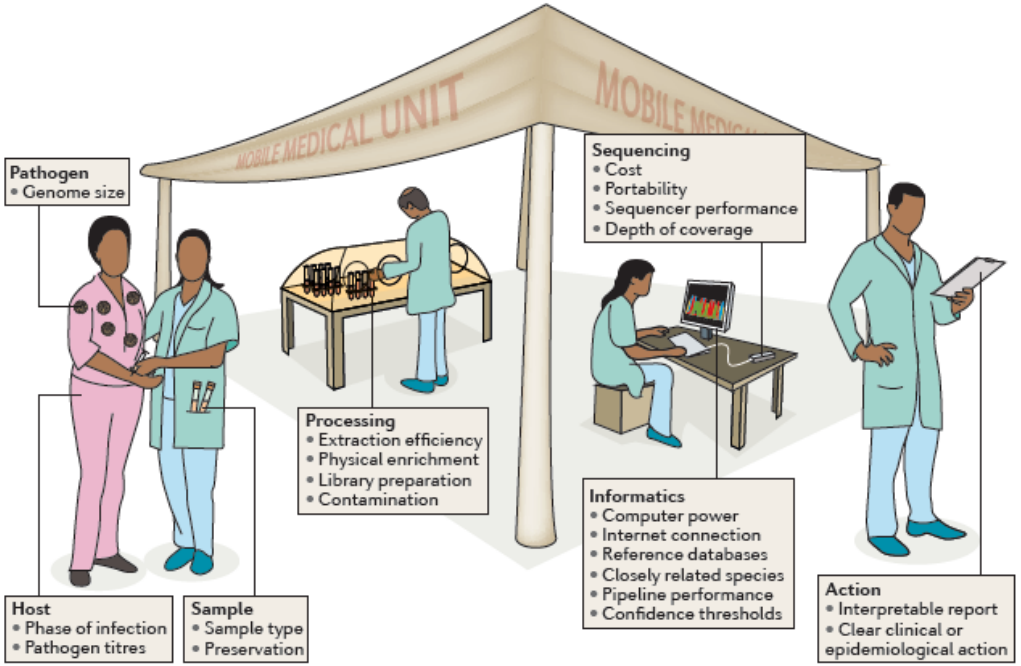


Figure 2 | **Challenges to in-field clinical metagenomics for rapid diagnosis and outbreak response**. A mobile medical unit deploying a portable clinical metagenomics platform has been established at the epicentre of an infectious disease outbreak, but the team faces challenges throughout the diagnostic process and epidemiological response. For example, in the case of Zika virus, samples, such as blood, with low viral titres, a small genome of <11 kb and transient viraemia[120] combine to complicate detection of viral nucleic acid by use of a strictly metagenomic approach. Furthermore, obtaining a sufficient amount of viral nucleic acids for genome sequencing beyond simple diagnostics requires a tiling PCR and amplicon sequencing approach[14]. Other challenges include, for example, access to a reliable Internet connection, the ability to collect sample metadata and translating genomic findings into real-time, actionable recommendations.

Vall d'Hebron
Institut de Recerca

1. Introduction to NGS
2. First Generation Sequencing
3. Second Generation Sequencing
4. Third Generation Sequencing
5. Sequencing generation face to face
6. Applications of NGS techniques
7. A (very) brief introduction to DoE

7. A (very) brief introduction to DoE

The **(statistical) design of experiments** is an efficient procedure for planning experiments so that the data obtained can be analyzed to yield valid and objective conclusions

Why are many life scientists so adverse to thinking about design?

It is common to think that time spent designing experiments would be better spent actually doing experiments

7. A (very) brief introduction to DoE

Variability types that play in an experiment:

- **Planned systematic variability**: This is the differences in response between treatments applied.

- **Noise variability:** random noise. Differences between two consecutives measures. We cannot avoid that.

- **Systematic variability not planned:** Produce a systematic variation in the results. A priori the reason is not known. It can be avoided with the *randomization* and the *local control*.

7. A (very) brief introduction to DoE

Important steps to define before begin the experiment:

- Establish the main **objectives** of the experiment. Avoid collateral problems
- Identify all the **noise** sources: Treatment, experimental errors,…
- **Allocate** each experimental unit which each treatment
- Clarify the **type of response** expected in each treatment
- Determinate the **number** of individuals in each group
- Run a **pilot study**

7. A (very) brief introduction to DoE

Important steps to define before begin the experiment:

- Establish the main **objectives** of the experiment. Avoid collateral problems
- Identify all the **noise** sources: Treatment, experimental errors,…
- **Allocate** each experimental unit which each treatment
- Clarify the **type of response** expected in each treatment
- Determinate the **number** of individuals in each group
- Run a **pilot study**

7. A (very) brief introduction to DoE

Important steps to define before begin the experiment:

- Establish the main **objectives** of the experiment. Avoid collateral problems
- Identify all the **noise** sources: Treatment, experimental errors,…
- **Allocate** each experimental unit which each treatment
- Clarify the **type of response** expected in each treatment
- Determinate the **number** of individuals in each group
- Run a **pilot study**

7. A (very) brief introduction to DoE

Important steps to define before begin the experiment:

- Establish the main **objectives** of the experiment. Avoid collateral problems
- Identify all the **noise** sources: Treatment, experimental errors,…
- **Allocate** each experimental unit which each treatment
- Clarify the **type of response** expected in each treatment
- Determinate the **number** of individuals in each group
- Run a **pilot study**
- How the **data** will be statistically analysed.

7. A (very) brief introduction to DoE

Sample	Treatment	Sex	Batch
1	A	Male	1
2	A	Male	1
3	A	Male	1
4	A	Male	1
5	B	Female	2
6	B	Female	2
7	B	Female	2
8	B	Female	2

Treatment are confounded between sex and batch

Sample	Treatment	Sex	Batch
1	A	Male	1
2	A	Female	2
3	A	Male	2
4	A	Female	1
5	B	Male	1
6	B	Female	2
7	B	Male	2
8	B	Female	1

Treatment is well balanced

7. A (very) brief introduction to DoE

LOCAL CONTROL

REPLICATION

RANDOMIZATION

7. A (very) brief introduction to DoE

What do you need to ask before starting a NGS experiment?

-What do I want to sequence? Whole genome, exome, metagenome, epigenome, RNAseq.....

-How many samples?

-Length of read required?

-Quality and quantity of starting material?

-Size of nucleic acids to sequence

-Amount of sequence needed: **coverage**

✓**(Depth of) Coverage:** how many times a particular base is sequenced.
30x = each base has been read by 30 sequences (in average)