Vall d'Hebron
Institut de Recerca

INTRODUCTION TO 'NEXT GENERATION SEQUENCING'

Esther Camacho[1], Mireia Ferrer[1], Álex Sánchez[1,2], Angel Blanco[1,2], Berta Miró[1]
[1] Unitat d'Estadística i Bioinformàtica (UEB) VHIR [2] Departament de Genètica Microbiologia i Estadística, UB

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

Vall d'Hebron
Institut de Recerca

Personalized medicine era

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

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

PHARMGKB

https://www.pharmgkb.org/

Biomarker identification:

- Diagnostic
- Susceptibility/risk (prevention)
- Prognostic
- Predictive (response)

1. Introduction to NGS

Vall d'Hebron
Institut de Recerca

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.

1. Introduction to NGS

Deoxyribose nucleic acid (DNA)

Polymerase Chain Reaction (PCR)

Human Genome Project

The Human Genome Project (HGP) 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

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

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.

1. Introduction to NGS

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**.

Cost of sequencing

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

Cost of human genome sequence:

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

2006: $14x10^{6}$$

2016: 1000-4000$

2019: 200 – 500$

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

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

Vall d'Hebron Institut de Recerca

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

Vall d'Hebron
Institut de Recerca

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

Vall d'Hebron
Institut de Recerca

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

Instruments

	High throughput	Benchtop
ROCHE	GS FLX+ 454	GS Junior 454
Illumina	NextSeq 550 Series NextSeq 2000 NovaSeq 6000 System	iSeq 100 System MiniSeq System MiSeq Series
Thermofisher	Ion Gene Studio S5 PrimeSystem	Ion Torrent Genexus 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 by **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**

Vall d'Hebron
Institut de Recerca

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 transcriptomic 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 transcriptomic analysis. *Wiley Interdiscip Rev RNA*. 2016;8(1):10.1002/wrna.1364.

Vall d'Hebron
Institut de Recerca

Clonal amplification:

For emPCR based systems (Ion/PGM, 454)

Vall d'Hebron
Institut de Recerca

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

Vall d'Hebron
Institut de Recerca

Clonal amplification: Illumina

BROAD INSTITUTE 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

Vall d'Hebron
Institut de Recerca

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.

Vall d'Hebron
Institut de Recerca

Sequencing: Ion S5/PGM

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

Vall d'Hebron
Institut de Recerca

Sequencing: Ion Torrent

Vall d'Hebron
Institut de Recerca

Sequencing: Ion Torrent

https://www.youtube.com/watch?v=zBPKj0mMcDg&ab_channel=ThermoFisherScientific

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

Vall d'Hebron
Institut de Recerca

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

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

Vall d'Hebron
Institut de Recerca

High throughput NGS Platforms

Popular Applications & Methods	iSeq 100 System	MiniSeq System	MiSeq Series	NextSeq Series
	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

Vall d'Hebron
Institut de Recerca

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

Vall d'Hebron
Institut de Recerca

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

Helicos Genetic Analysis system

Workflow similar to Illumina, but **without amplification**:

- relative slow and expensive
- short reads
- **single molecule sequencing**

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

Vall d'Hebron
Institut de Recerca

Helicos Genetic Analysis system

Workflow similar to Illumina, but **without amplification**:

- relative slow and expensive
- short reads
- **single molecule sequencing**

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

4. Third Generation Sequencing

Sequel System (Pacific bioscience)

- **Quick** library construction (< 3 hours)
- Run time: from 30 min to 30 hours
- Sequencing by synthesis
- No library amplification
- **Long reads:** 5 Kb – 20 Kb

Source: Supplier webpage

Vall d'Hebron
Institut de Recerca

Sequel System (Pacific bioscience)

- **Quick** library construction (< 3 hours)
- Run time: from 30 min to 30 hours
- Sequencing by synthesis
- No library amplification
- **Long reads:** 5 Kb – 20 Kb

Source: Supplier webpage

Vall d'Hebron
Institut de Recerca

Sequel System (Pacific bioscience)

https://www.youtube.com/watch?v=_lD8JyAbwEo&feature=emb_rel_end&ab_channel=PacBio

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

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

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)

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

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)

FGS

1. DNA fragmentation

2. In vitro adaptor ligation + clonal amplification

3. Massive paralel sequencing

4. Image processing and data analysis

SGS

1. DNA fragmentation

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

4. Image processing and data analysis.

TGS

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.

https://genohub.com/ngs-instrument-guide/

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

Depth of Coverage (DNA)

Estimate of Coverage Requirements by Application Type

Application Type	Coverage
DNA-Seq (Re-Sequencing)	30 - 80X
DNA-Seq (De novo asser	
SNP Analysis / Rearrange	0X
Exome	200X
ChIP-Seq	0X

Depth of Coverage: average number of reads that align to, or "cover," known reference bases.

30x = each base has been covered by 30 sequences (in average)

https://genohub.com/recommended-sequencing-coverage-by-application/

Depth of Coverage (RNA)

Sample Type	Reads Needed for Differential Expression (millions)	Reads Needed for Rare Transcript or De Novo Assembly (millions)	Read Length
Small Genomes (i.e. Bacteria / Fungi)	5	30 - 65	50 SR or PE for positional info
Intermediate Genomes (i.e. Drosophila / C. Elegans)	10	70 - 130	50 – 100 SR or PE for positional info
Large Genomes (i.e. Human / Mouse)	15 - 25	100 - 200	>100 SR or PE for positional info

6. Applications of NGS techniques

Vall d'Hebron
Institut de Recerca

Depth of Coverage (DNA)

Estimate of Coverage Requirements by Application Type

Application Type	Coverage
DNA-Seq (Re-Sequencing)	30 - 80X
DNA-Seq (De novo assembly)	100X
SNP Analysis / Rearrangement Detection	10 - 30X
Exome	100 - 200X
ChIP-Seq	10 - 40X

https://genohub.com/recommended-sequencing-coverage-by-application/

Depth of Coverage (RNA)

Sample Type	Reads Needed for Differential Expression (millions)	Reads Needed for Rare Transcript or De Novo Assembly (millions)	Read Length
Small Genomes (i.e. Bacteria / Fungi)	5	30 - 65	50 SR or PE for positional info
Intermediate Genomes (i.e. Drosophila / C. Elegans)	10	70 - 130	50 – 100 SR or PE for positional info
Large Genomes (i.e. Human / Mouse)	15 - 25	100 - 200	>100 SR or PE for positional info

6. Applications of NGS techniques

Vall d'Hebron Institut de Recerca

Category	Detection or Application	Recommended Coverage (x) or Reads (millions)	References
Whole genome sequencing	Homozygous SNVs	15x	Bentley et al., 2008
	Heterozygous SNVs	33x	Bentley et al., 2008
	INDELs	60x	Feng et al., 2014
	Genotype calls	35x	Ajay et al., 2011
	CNV	1-8x	Xie et al., 2009; Medvedev at al., 2010
Whole exome sequencing	Homozygous SNVs	100x (3x local depth)	Clark et al., 2011; Meynert et al., 2013
	Heterozygous SNVs	100x (13x local depth)	Clark et al., 2011; Meynert et al., 2013
	INDELs	not recommended	Feng et al., 2014
Transcriptome Sequencing	Differential expression profiling	10-25M	Liu Y. et al., 2014; ENCODE 2011 RNA-Seq
	Alternative splicing	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	Allele specific expression	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	De novo assembly	>100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
DNA Target-Based Sequencing	ChIP-Seq	10-14M (sharp peaks); 20-40M (broad marks)	Rozowsky et al., 2009; ENCODE 2011 Genome; Landt et al., 2012
	Hi-C	100M	Belton, J.M et al., 2012
	4C (Circularized Chromosome Confirmation Capture)	1-5M	van de Weken, H.J.G. et al., 2012
	5C (Chromosome Carbon Capture Carbon Copy)	15-25M	Sanyal A. et al., 2012
	ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing)	15-20M	Zhang, J. et al., 2012
	FAIRE-Seq	25-55M	ENCODE 2011 Genome; Landt et al., 2012
	DNAse 1-Seq	25-55M	Landt et al., 2012
DNA Methylation Sequencing	CAP-Seq	>20M	Long, H.K. et al., 2013
	MeDIP-Seq	60M	Taiwo, O. et al., 2012
	RRBS (Reduced Representation Bisulfite Sequencing)	10X	ENCODE 2011 Genome
	Bisulfite-Seq	5-15X; 30X	Ziller, M.J et al., 2015; Epigenomics Road Map
RNA-Target-Based Sequencing	CLIP-Seq	10-40M	Cho J. et al., 2012; Eom T. et al., 2013; Sugimoto Y. et al., 2012
	iCLIP	5-15M	Sugimoto Y. et al., 2012; Rogelj B. et al., 2012
	PAR-CLIP	5-15M	Rogelj B. et al., 2012
	RIP-Seq	5-20M	Lu Z. et al., 2014
Small RNA (microRNA) Sequencing	Differential Expression	~1-2M	Metpally RPR et al., 2013; Campbell et al., 2015
	Discovery	~5-8M	Metpally RPR et al., 2013; Campbell et al., 2015

https://genohub.com/recommended-sequencing-coverage-by-application/

6. Applications of NGS techniques

Vall d'Hebron
Institut de Recerca

Whole Genome sequencing

***De novo* sequencing**

Resequencing

Vall d'Hebron
Institut de Recerca

Whole Genome sequencing

- Complete characterization of the **entire genome**.
- 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

RNA-seq Expression Analysis

- Sequencing every **RNA** molecule and profiling the expression of a particular gene by counting the number of time its transcripts have been sequenced.
- 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**

Vall d'Hebron Institut de Recerca

RNA-seq Expression Analysis

- Sequencing every **RNA** molecule and profiling the expression of a particular gene by counting the number of time its transcripts have been sequenced.
- 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**

TGS

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

Single cell RNA-seq (scRNA-seq)

- Allows **comparison of the transcriptomes of individual cells**
- sequence the transcriptomes of **up to tens of thousands of individual cells** for a single project
- Results depend of protocol used
- Special attention to **single cell purification**
- **Low signal of weak** expressed genes
- Required **number of cells increases with the complexity** of the sample under investigation
- Requires specific **bioinformatic** approaches.

Vall d'Hebron Institut de Recerca

Single cell RNA-seq (scRNA-seq)

Different methods for isolating single cells from a cell suspension:

- **Fluorescence-activated cell sorting** (FACS)
- **Microfluidics**: allows separation and cDNA synthesis and transcriptome amplification
- **Mechanical micromanipulation/micropipetting**
- **Laser capture microdissection**

Figure 33.2.1 Single-cell isolation methods. (A) Fluorescence-activated cell sorting (FACS). Fluorescently labeled cells are exposed to a laser beam. Detectors identify cells with the desired fluorescence pattern, and the cells are sorted into plates, with one cell in each well. (B) Microdroplets. Cells and beads containing primers and reagents are enclosed in microdroplets. Downstream biochemical reactions occur within these droplets. (C) Micropipetting. Cells of interest are manually picked with a glass pipet under a microscope. (D) Laser capture microdissection (LCM). Cells on a glass slide are attached to a polymer by means of a laser beam. The polymer and cells of interest are then lifted from the slide and transferred to a microcentrifuge tube.

Current Protocols in Molecular Biology e57, April 2018
Published online April 2018 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpmb.57
Copyright © 2018 John Wiley & Sons, Inc.

Vall d'Hebron Institut de Recerca

Single cell RNA-seq (scRNA-seq)

Also different methods for cDNA transcription….

TABLE 1 | Summary of widely used scRNA-seq technologies.

Methods	Transcript coverage	UMI possibility	Strand specific	References
Tang method	Nearly full-length	No	No	Tang et al., 2009
Quartz-Seq	Full-length	No	No	Sasagawa et al., 2013
SUPeR-seq	Full-length	No	No	Fan X. et al., 2015
Smart-seq	Full-length	No	No	Ramskold et al., 2012
Smart-seq2	Full-length	No	No	Picelli et al., 2013
MATQ-seq	Full-length	Yes	Yes	Sheng et al., 2017
STRT-seq and STRT/C1	5′-only	Yes	Yes	Islam et al., 2011, 2012
CEL-seq	3′-only	Yes	Yes	Hashimshony et al., 2012
CEL-seq2	3′-only	Yes	Yes	Hashimshony et al., 2016
MARS-seq	3′-only	Yes	Yes	Jaitin et al., 2014
CytoSeq	3′-only	Yes	Yes	Fan H.C. et al., 2015
Drop-seq	3′-only	Yes	Yes	Macosko et al., 2015
InDrop	3′-only	Yes	Yes	Klein et al., 2015
Chromium	3′-only	Yes	Yes	Zheng et al., 2017
SPLiT-seq	3′-only	Yes	Yes	Rosenberg et al., 2018
sci-RNA-seq	3′-only	Yes	Yes	Cao et al., 2017
Seq-Well	3′-only	Yes	Yes	Gierahn et al., 2017
DroNC-seq	3′-only	Yes	Yes	Habib et al., 2017
Quartz-Seq2	3′-only	Yes	Yes	Sasagawa et al., 2018

Chen G, Ning B and Shi T (2019) Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front. Genet. 10:317. doi: 10.3389/fgene.2019.00317

Vall d'Hebron
Institut de Recerca

Single cell RNA-seq (scRNA-seq)

Also different methods for reads mapping….

TABLE 2 | Tools for read mapping and expression quantification of scRNA-seq data.

Tools	Category	URL	References
TopHat2	Read mapping	https://ccb.jhu.edu/software/tophat/index.shtml	Kim et al., 2013
STAR	Read mapping	https://github.com/alexdobin/STAR	Dobin and Gingeras, 2015
HISAT2	Read mapping	https://ccb.jhu.edu/software/hisat2/index.shtml	Kim et al., 2015
Cufflinks	Expression quantification	https://github.com/cole-trapnell-lab/cufflinks	Trapnell et al., 2010
RSEM	Expression quantification	https://github.com/deweylab/RSEM	Li and Dewey, 2011
StringTie	Expression quantification	https://github.com/gpertea/stringtie	Pertea et al., 2015

Chen G, Ning B and Shi T (2019) Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front. Genet. 10:317. doi: 10.3389/fgene.2019.00317

Vall d'Hebron
Institut de Recerca

Single cell RNA-seq (

Also different metho

TABLE 2 | Tools for read mapping and expression qua
of scRNA-seq data.

Tools	Category	URL
TopHat2	Read mapping	https://ccb.jhu.edu/software/tophat/index.shtml
STAR	Read mapping	https://github.com/alexdobin/STAR
HISAT2	Read mapping	https://ccb.jhu.edu/software/hisat2/index.shtml
Cufflinks	Expression quantification	https://github.com/cole-trapnell-lab/cufflinks
RSEM	Expression quantification	https://github.com/deweylab/RSEM
StringTie	Expression quantification	https://github.com/gpertea/stringtie

TABLE 4 | Differential expression analysis tools for RNA-seq data.

Methods	Category	URL	Referenes
ROTS	Single cell	https://bioconductor.org/packages/release/bioc/html/ROTS.html	Seyednasrollah et al., 2016
MAST	Single cell	https://github.com/RGLab/MAST	Finak et al., 2015
BCseq	Single cell	https://bioconductor.org/packages/devel/bioc/html/bcSeq.html	Chen and Zheng, 2018
SCDE	Single cell	http://hms-dbmi.github.io/scde/	Kharchenko et al., 2014
DEsingle	Single cell	https://bioconductor.org/packages/DEsingle	Miao et al., 2018
Cencus	Single cell	http://cole-trapnell-lab.github.io/monocle-release/	Qiu et al., 2017
D3E	Single cell	https://github.com/hemberg-lab/D3E	Delmans and Hemberg, 2016
BPSC	Single cell	https://github.com/nghiavtr/BPSC	Vu et al., 2016
DESeq2	Bulk	https://bioconductor.org/packages/release/bioc/html/DESeq2.html	Love et al., 2014
edgeR	Bulk	https://bioconductor.org/packages/release/bioc/html/edgeR.html	Robinson et al., 2010
Limma	Bulk	http://bioconductor.org/packages/release/bioc/html/limma.html	Ritchie et al., 2015
Ballgown	Bulk	http://www.bioconductor.org/packages/release/bioc/html/ballgown.html	Frazee et al., 2015

For differential expression analysis.....

Chen G, Ning B and Shi T (2019) Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front. Genet. 10:317. doi: 10.3389/fgene.2019.00317

Vall d'Hebron
Institut de Recerca

Single cell RNA-seq (scRNA-seq)

HUMAN CELL ATLAS

To create comprehensive reference maps of all human cells

https://www.humancellatlas.org/

Mapping the Human Body at the Cellular Level

Community generated, multi-omic, open data processed by standardized pipelines

4.5M CELLS | 33 ORGANS | 289 DONORS | 28 PROJECTS | 81 LABS

6. Applications of NGS techniques

Vall d'Hebron
Institut de Recerca

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.

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

Vall d'Hebron
Institut de Recerca

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

Vall d'Hebron
Institut de Recerca

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

Vall d'Hebron
Institut de Recerca

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

Vall d'Hebron
Institut de Recerca

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

7. A (very) brief introduction to DoE

Vall d'Hebron
Institut de Recerca

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

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

Vall d'Hebron
Institut de Recerca

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**