Introduction to NGS technologies

Joaquín Panadero Romero

29th February 2016

GDA
International Course on
Genomic Data Analysis

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

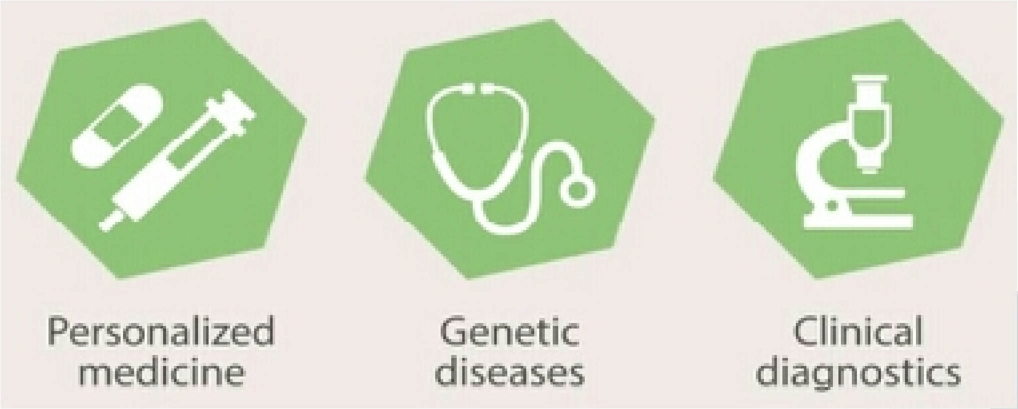
1

OUTLINE

1. Basics on the NGS technologies
2. Comparisons across NGS platforms
3. Computing infrastructure for NGS analyses
4. Tools for data analysis

2 Basic on NGS technologies

Millions of DNA molecules sequenced simultanously



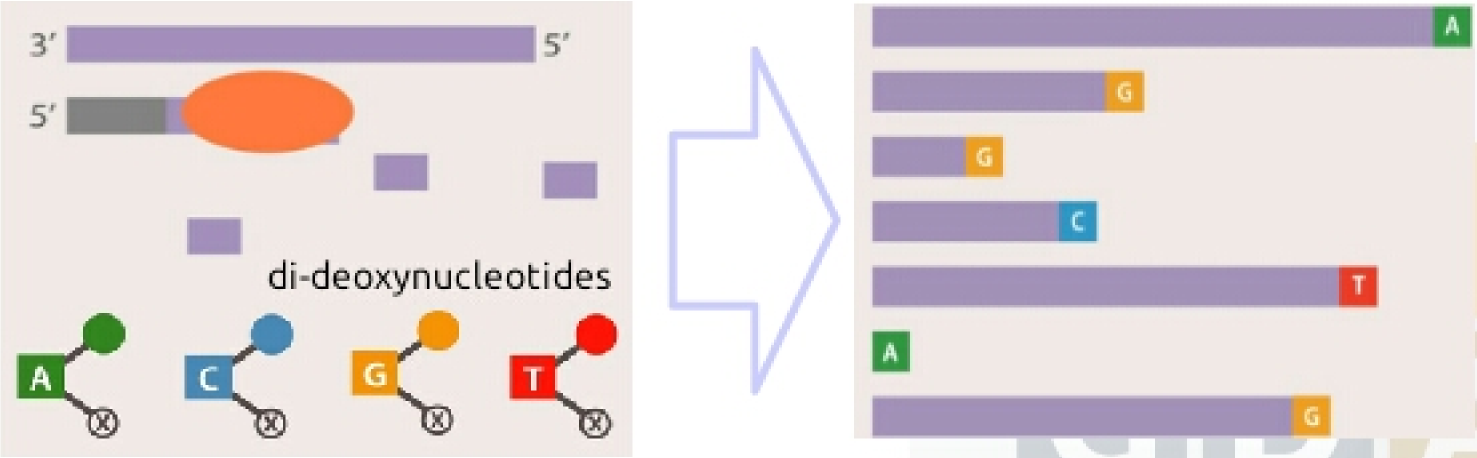
Types:

Sanger
Pyrosequencing
Sequencing by synthesis
Sequencing by ligation
Ion-Semiconductor sequencing

3 SANGER

Used nowadays in:

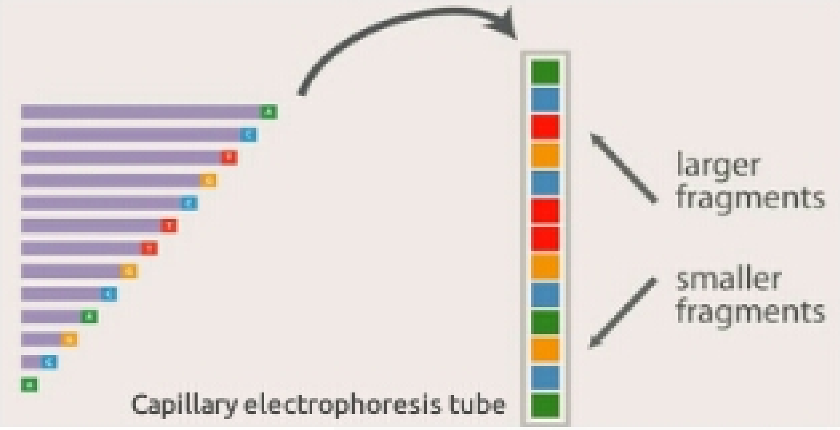
- Routine sequencing applications
- NGS data validation



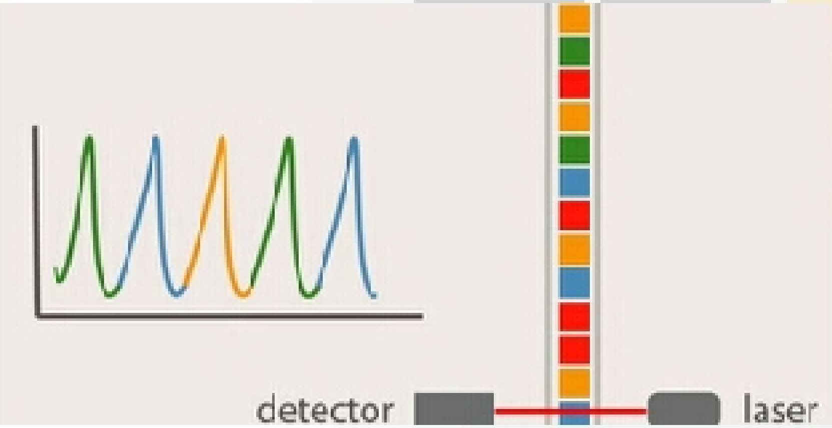
Multiple DNA fragments covering each base position

4 SANGER

DNA fragments move according their size



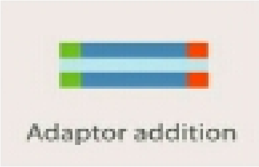
Light detected shows the base added at each position



5

Common among NGS technologies

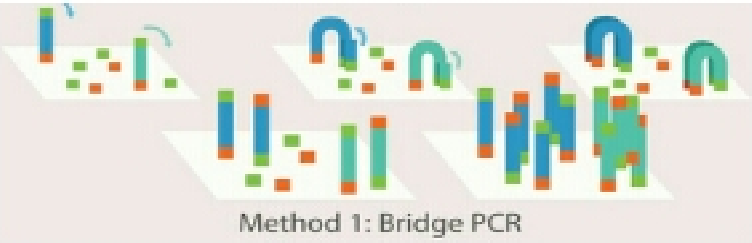
1. sample preparation


Adaptor addition

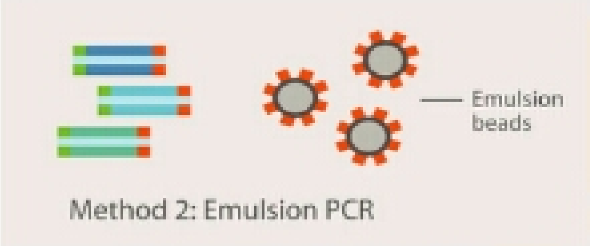
cDNA fragments ligated to adaptors at both ends

Amplification based on PCR bridges or bead emulsion

2. sequencing machine

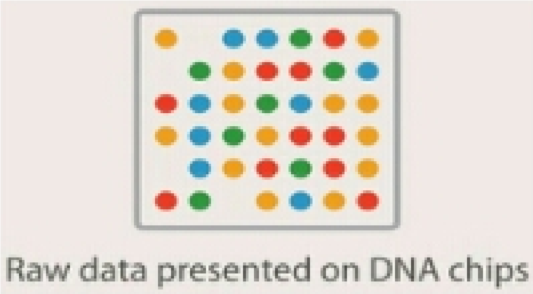

Method 1: Bridge PCR

⇔



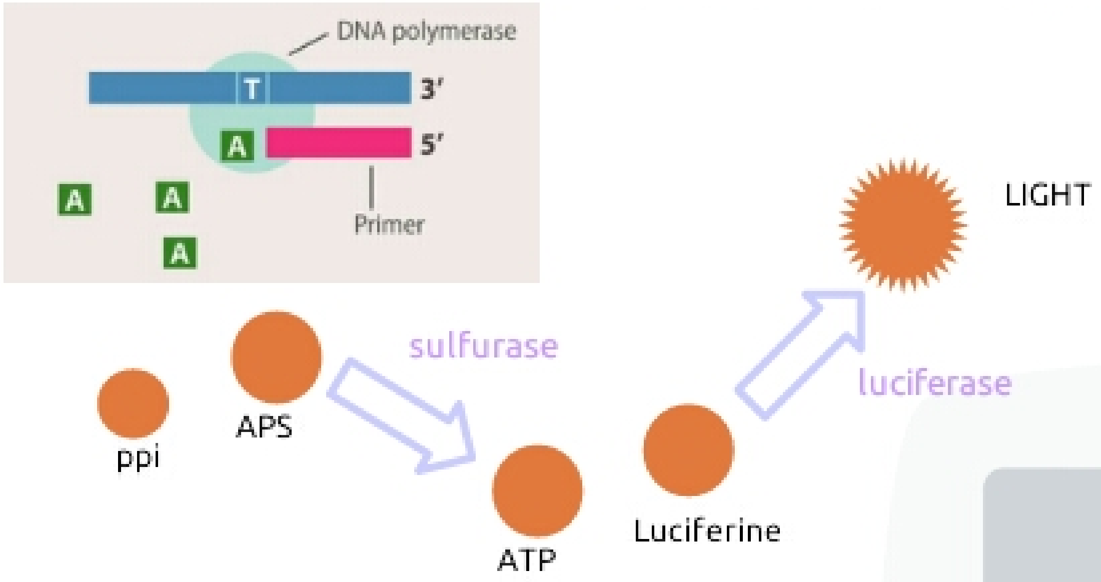
Method 2: Emulsion PCR

3. data output


Raw data presented on DNA chips

Sequencing output is provided in clusters

6 Pyrosequencing



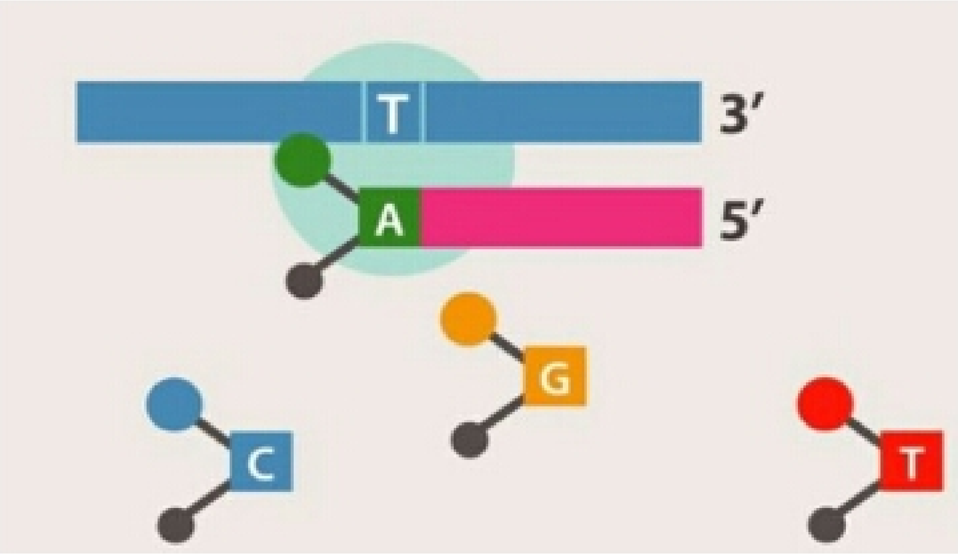
Overview

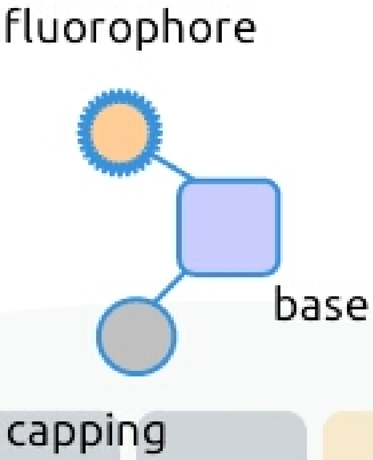
- Large reads lengths generation

- High reagent cost

- High error rate over strings of 6+ homopolymers

7 Sequencing by synthesis





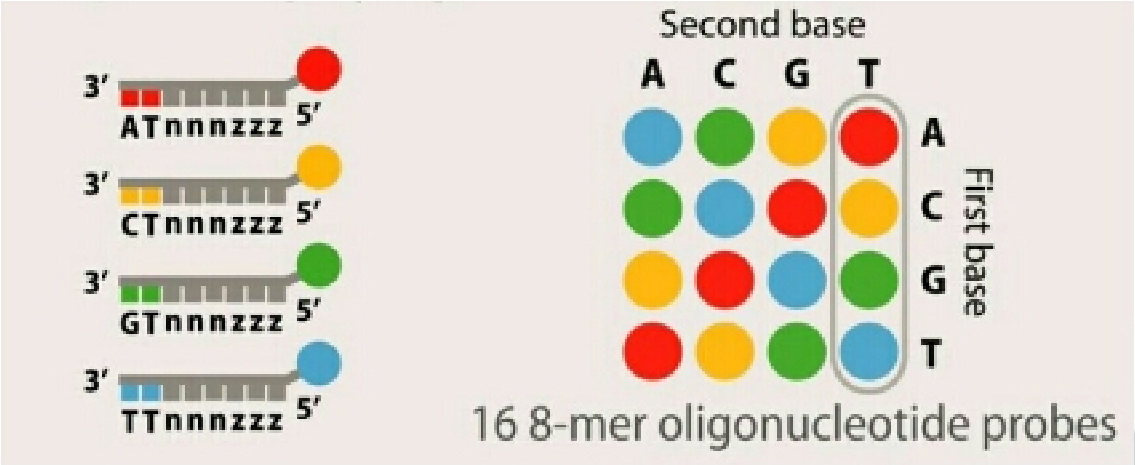
Overview

- Overcomes homopolymer issue due to terminated nucleotides

- Increased error rate with increased read lengths

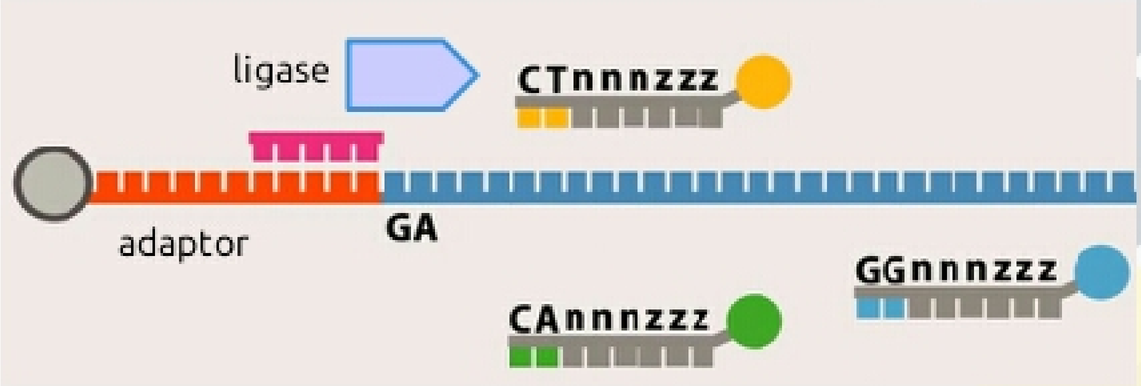
8

Sequencing by ligation



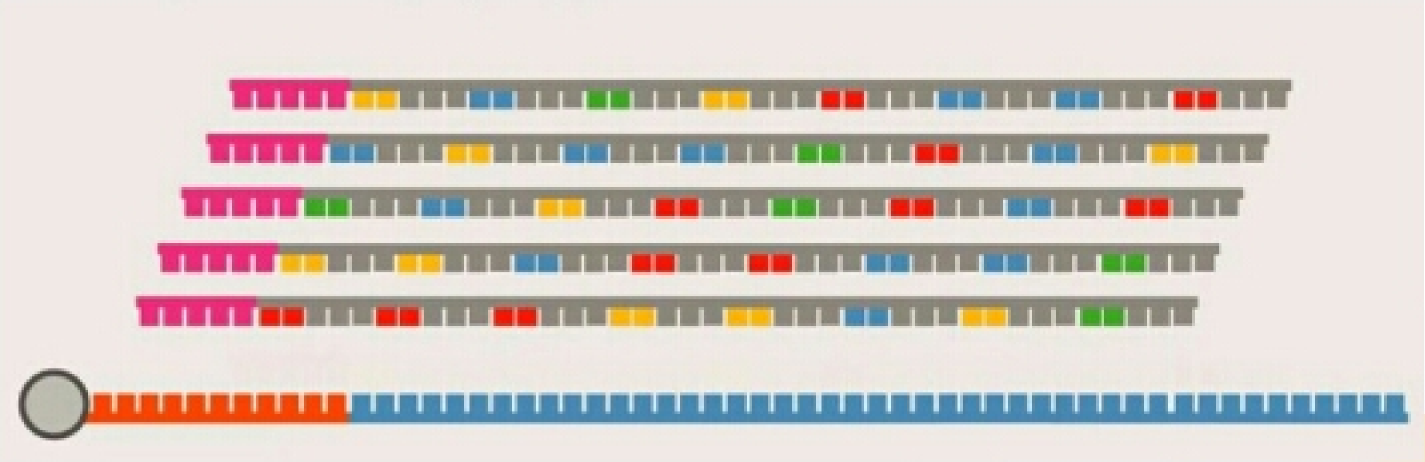
	A	C	G	T
A	🔵	🟢	🟡	🔴
C	🟢	🔵	🔴	🟡
G	🟡	🔴	🔵	🟢
T	🔴	🟡	🟢	🔵

16 8-mer oligonucleotide probes



9

Sequencing by ligation



5 x 7 ligation cicles. Each primer hibridizes one base back

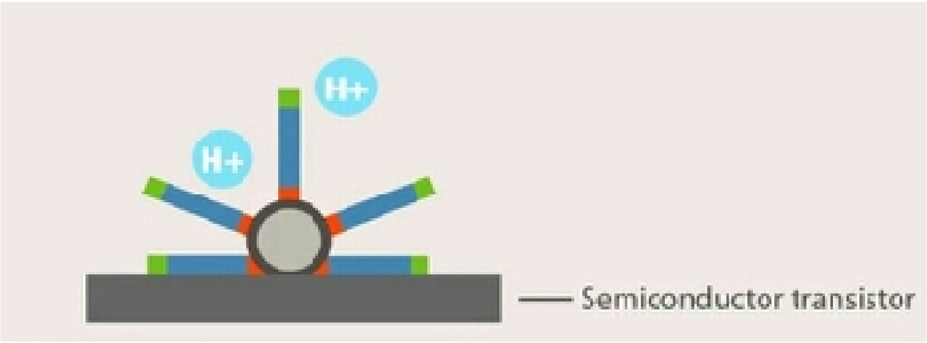
Overview

- Oligonucleotide probes used rather than DNA Polymerase

- Very short read lengths

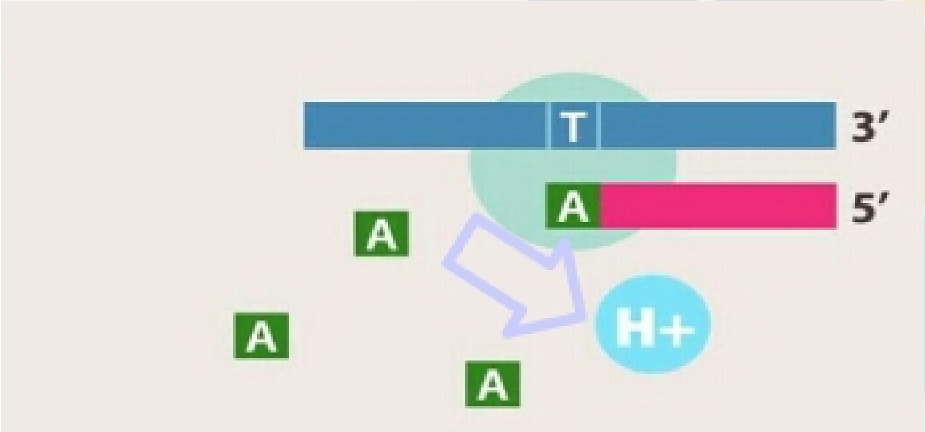
10 Ion - Semiconductor sequencing

Beads are attached to semiconductor transistors



Each time a nucleotide is added, one H+ is released

Semiconductor transistor detects changes on PH solution



11 Ion - Semiconductor sequencing

__Overview__

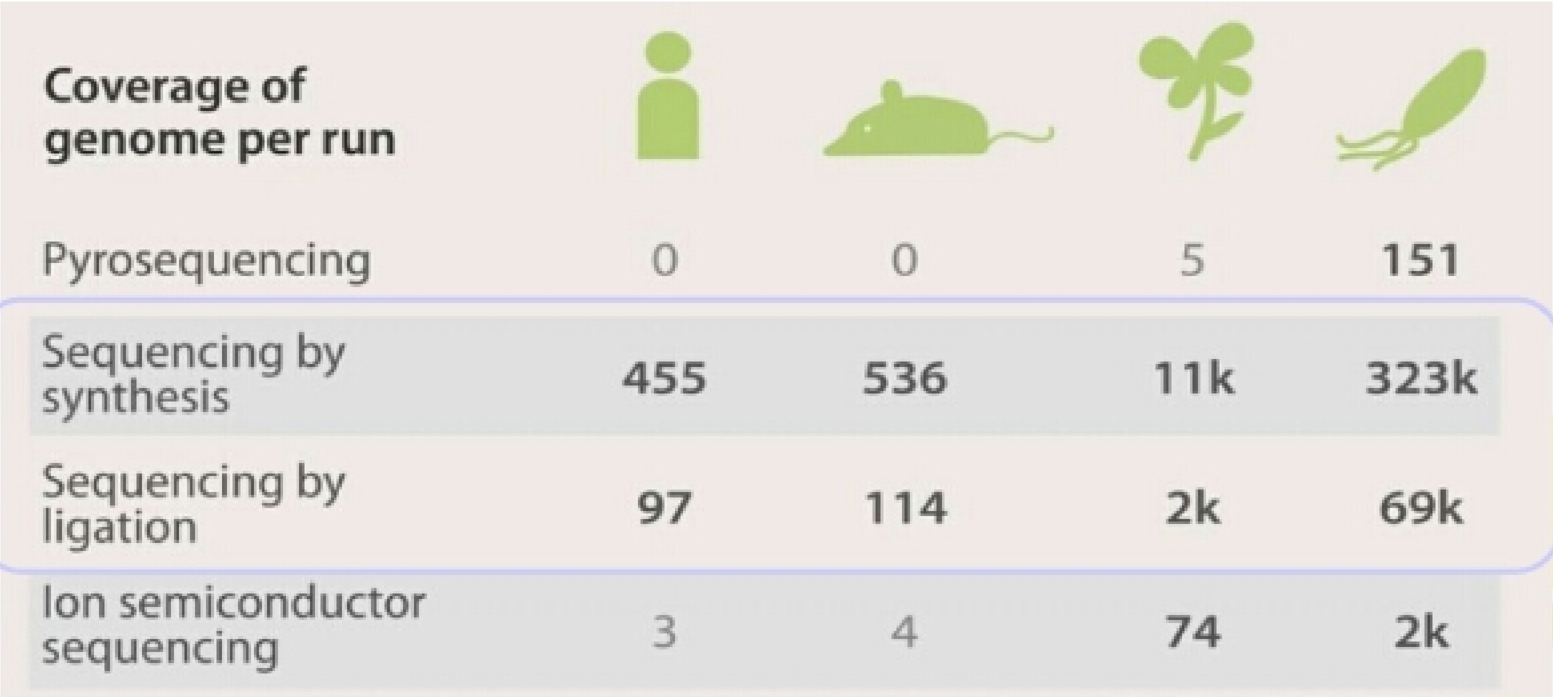
- Similar to pyrosequencing, but measures the release of H+ instead of pyrophosphate

- Most cost - effective and time - efficient

Examples of NGS systems

13

NGS comparison

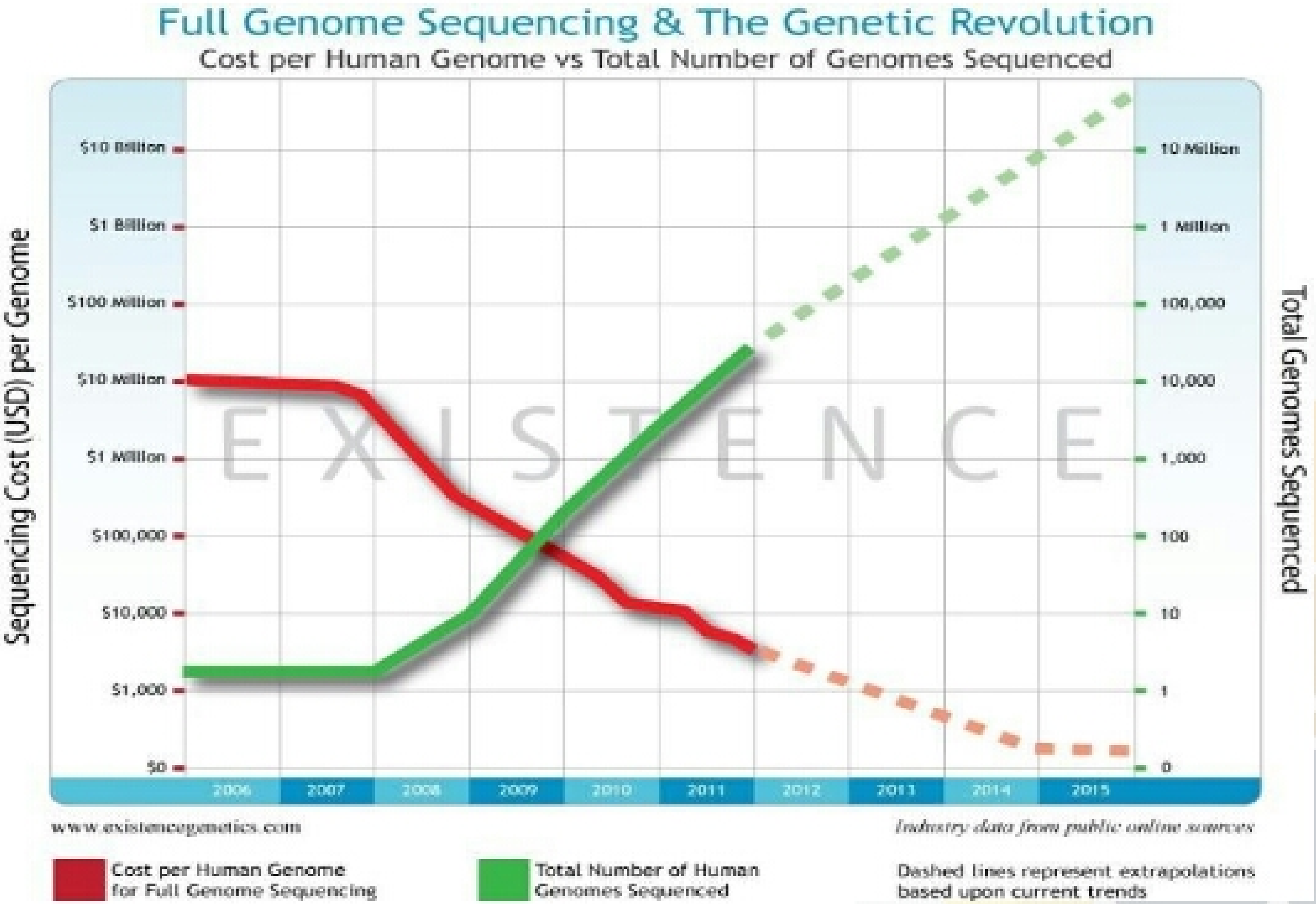


Coverage of genome per run	🧍	🐭	🍀	🦠
Pyrosequencing	0	0	5	**151**
Sequencing by synthesis	**455**	**536**	**11k**	**323k**
Sequencing by ligation	**97**	**114**	**2k**	**69k**
Ion semiconductor sequencing	3	4	**74**	**2k**

Applications

- Whole genome sequencing
- Variant Calling
- RNA-seq
- De novo sequencing and assembly
- Chip-seq
- Methyl-seq
- Metagenomics

Sequencing costs



Full Genome Sequencing & The Genetic Revolution

Cost per Human Genome vs Total Number of Genomes Sequenced

www.existencegenetics.com

Industry data from public online sources

🟥 Cost per Human Genome for Full Genome Sequencing

🟩 Total Number of Human Genomes Sequenced

Dashed lines represent extrapolations based upon current trends

Sequencing costs

- In NGS we have to process really big amounts of data, which is not trivial in computing terms

- Big NGS projects require supercomputing infrastructures

thus

we can tackle such amount of data by using specific hardware combined with software capable to deal with data generated

Computational infrastructure for NGS

Requirements:

- Conditioned data center (server rooms)
- Computing cluster (racks)
- Many computing nodes (servers)
- High performance and high capacity storage
- Fast networks (10Gb ethernet, infiniband...)
- Skilled people in computing (sysadmins and developers)

Computing cluster and storage

Distributed memory cluster

- 8 or 12 cores per node
- At least 48GB RAM per node

Fast networks

- 10 Gbit, infiniband...

Batch queue system

- sge, slurm, condor, pbs

What do we want to storage?

Raw data (fastq)
Processed data (fastq, bam, sam, vcf)
Final results (txt, excel...)

How many storage resources?
For how long?

CNAG Centro Nacional Analisis Genómico

Sequencing instruments
- 10 Illumina HiSeq2000

Informatics infrastructure
- 850 core cluster
- 7.5 petabytes, lustre filesystem
- 10 x 10 Gb link with MareNostrum

BGI Beijing Genomics Institute

Sequencing instruments
- Illumina HiSeq
- AB Solid System
- Ion Torrent

Informatics infrastructure
- 20576 cores cluster
- 17 PB (petabytes)

Alternatives: cloud computing

Pros

- flexibility
- you pay what you use
- don't need to mantain a data center

Cons

- transfer datasets through the internet is slow
- lower performance
- privacy and security concerns
- more expensive for big and long term projects

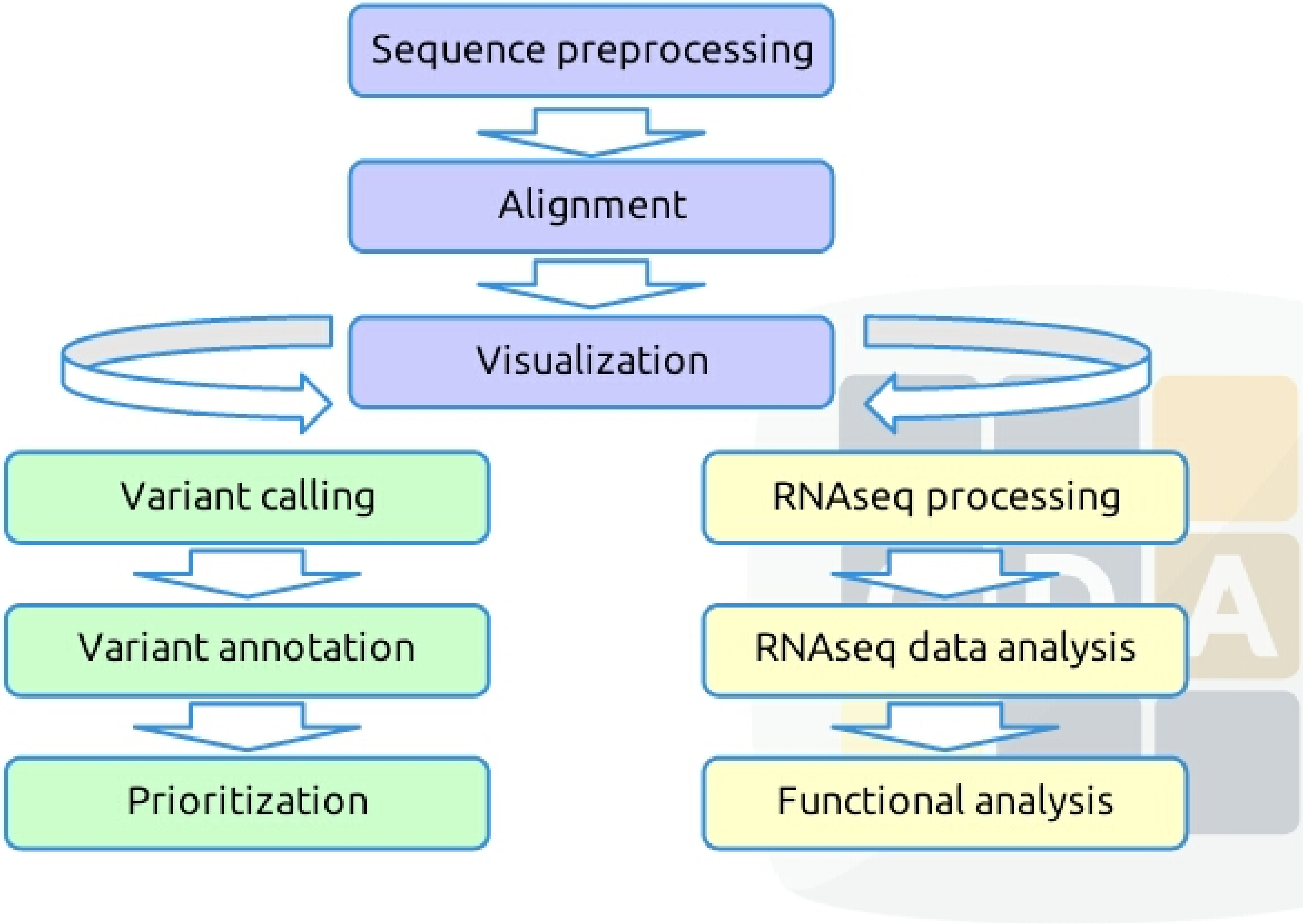
Tools on NGS data analysis

Quality	FastQC
Trimming	cutadapt
Assembly	abyss, velvet, ...
ORFs prediction	glimmer, augustus, ...
Annotation	Blast2GO
Mapping	BWA, bowtie, hpgaligner

Tools on NGS data analysis

Differential expression	babelomics, cuffdiff, bioconductor
Variant calling and Variant annotations	GATK, samtools, Annovar, BiERapp
Metagenomics	qimme, mothur
Methilation	bismark
Functional profiling	babelomics
Path signaling	hiPathia and Pathact

NGS Data Analysis Pipeline



Sequence preprocessing

Alignment

Visualization

Variant calling

Variant annotation

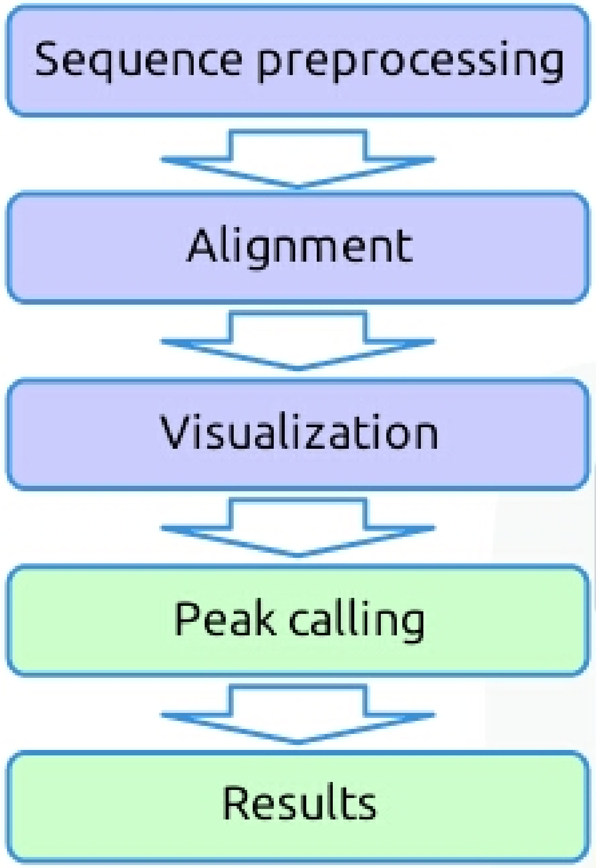
Prioritization

RNAseq processing

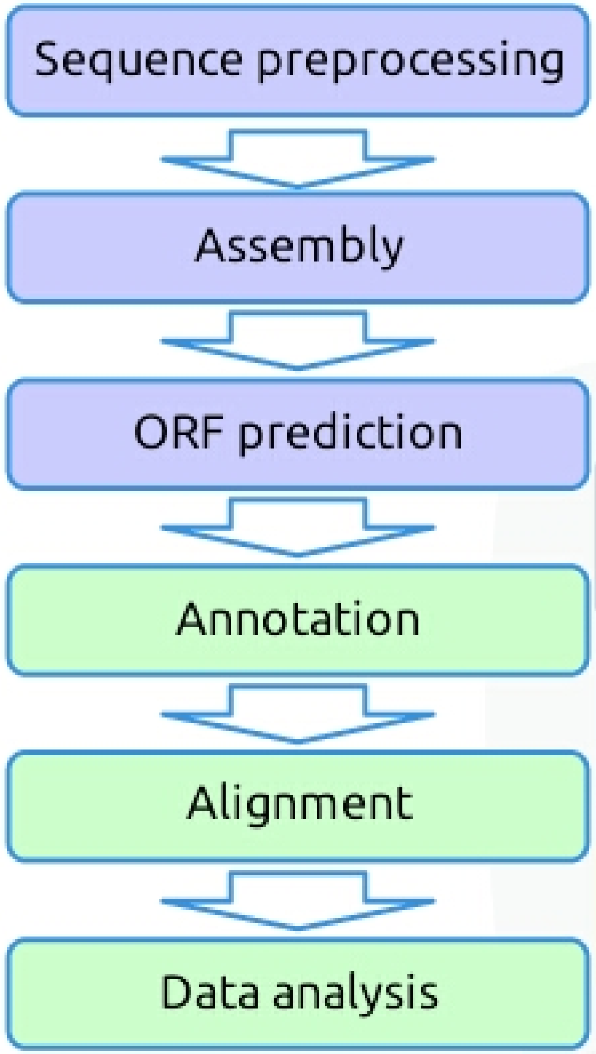
RNAseq data analysis

Functional analysis

NGS Data Analysis Pipeline



NGS Data Analysis Pipeline



Sequence preprocessing

Assembly

ORF prediction

Annotation

Alignment

Data analysis

THANKS