Introduction to NGS technologies

Valencia, 28-30 Sep 2015

Joaquín Tárraga
jtarraga@cipf.es

Genomics Data Analysis CIBERER

Outline

1) Basics on the NGS technologies

2) Some remarks on experiment design

3) Computing infrastructure for NGS analyses

Basic on NGS technologies

- Next Generation Sequencing (NGS) is a powerful platform that has enabled the sequencing of thousands to millions of DNA molecules simultaneously.
- This powerful tool is revolutionizing fields such as personalized medicine, genetic diseases, and clinical diagnostics by offering a high throughput option with the capability to sequence multiple individuals at the same time.

NGS technologies

Cost-effective
Fast
Ultra throughput
Cloning-free
Short/long reads

Next Generation Genome Sequencers

Illumina HiSeq and MiSeq

PacBio SMRT

454 GS FLX

Oxford Nanopore

GD

N

Platform Features

Feature	HiSeq2500 - Highoutput	HiSeq2500 – Rapid mode	MiSeq	PacBio RSII
Number of reads	150-180M/lane	100-150M/lane	12-15M (v2) 20-25M (v3)	50-80K/SMRT cell
Read length	2 x 100 bp	2 x 150 bp	2 x 300 bp (v3)	~ 10-20 kb
Yield per lane (PF data)	up to 35 Gb	up to 45Gb	up to 15 Gb	up to 0.4 Gb
Instrument Time	~12-14 days	~2 days	~2 days	~2 hours
Pricing per Gb	$59 (PE100)	$53 (PE150)	$108 (PE300)	$697

Basic on NGS technologies

VIDEO from Applied Biological Materials (abm)

https://www.youtube.com/watch?v=jFCD8Q6qSTM

Applications

- Whole genome sequencing
- Exome sequencing
- RNA-seq
- De novo sequencing and assembly
- Chip-seq
- Methyl-seq
- Disease panels
- ...

Some remarks on experiment design

What's your biological question ?

Roche 454 / PacBio

- Long fragments
- Errors: indels / many
- Low throughput
- Expensive

Illumina

- Short fragments
- Errors: mismatches
- High throughput
- Cheap

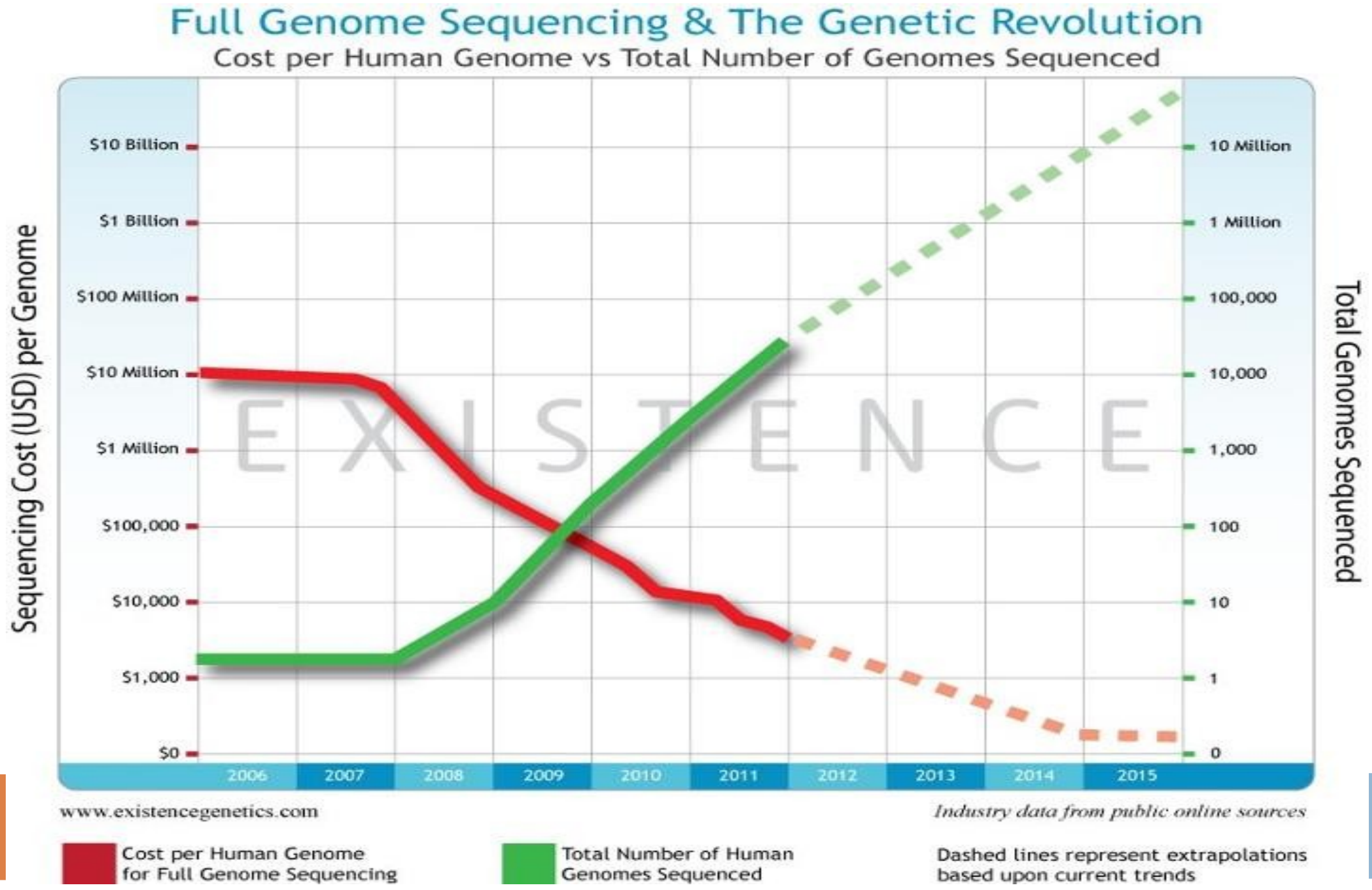
- Coverage
- Quality (Q20)

Computational infrastructure for NGS

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

- Big NGS projects require **supercomputing** infrastructures

Sequencing cost



Computational infrastructure for NGS

- Expensive and not trivial to use, 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)
    - In CNAG, about 30 staff (>50% informatics)

Computing cluster

- 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
- Many GPUs tools are being developed

Storage system in NGS

- The most important piece
- The most expensive
- Good design is really important
- Keep in mind the storage scalability
- Distributed filesystem:
  - Lustre, GPFS, Ibrix, GlusterFS,...
- Reading: http://www.bioteam.net/wp-content/u dag-xgen-storageForNGS_v3.pdf

Sequencing center examples

- Spain
  - MGP: Medical Genome Project, Seville
  - CNAG: Centro Nacional de Análisis Genómico, Barcelone
- Largest in the world
  - BGI: Beijing Genomics Institute

Medical Genome Project (MGP)

- Sequencing instruments
  - 7 GS-FLX (Roche)
  - 4 SolidTM 5500 (Applied Biosystems)
- Informatics infrastructure
  - 300 core cluster
  - 0.5 PB (petabytes), ibrix filesystem

MGP: raw data generation

- 1 solid sequencer run
  - 7 days running
  - Generates around 4 TB

- Only the four solid sequencers working full time can generate around 12 TB each week

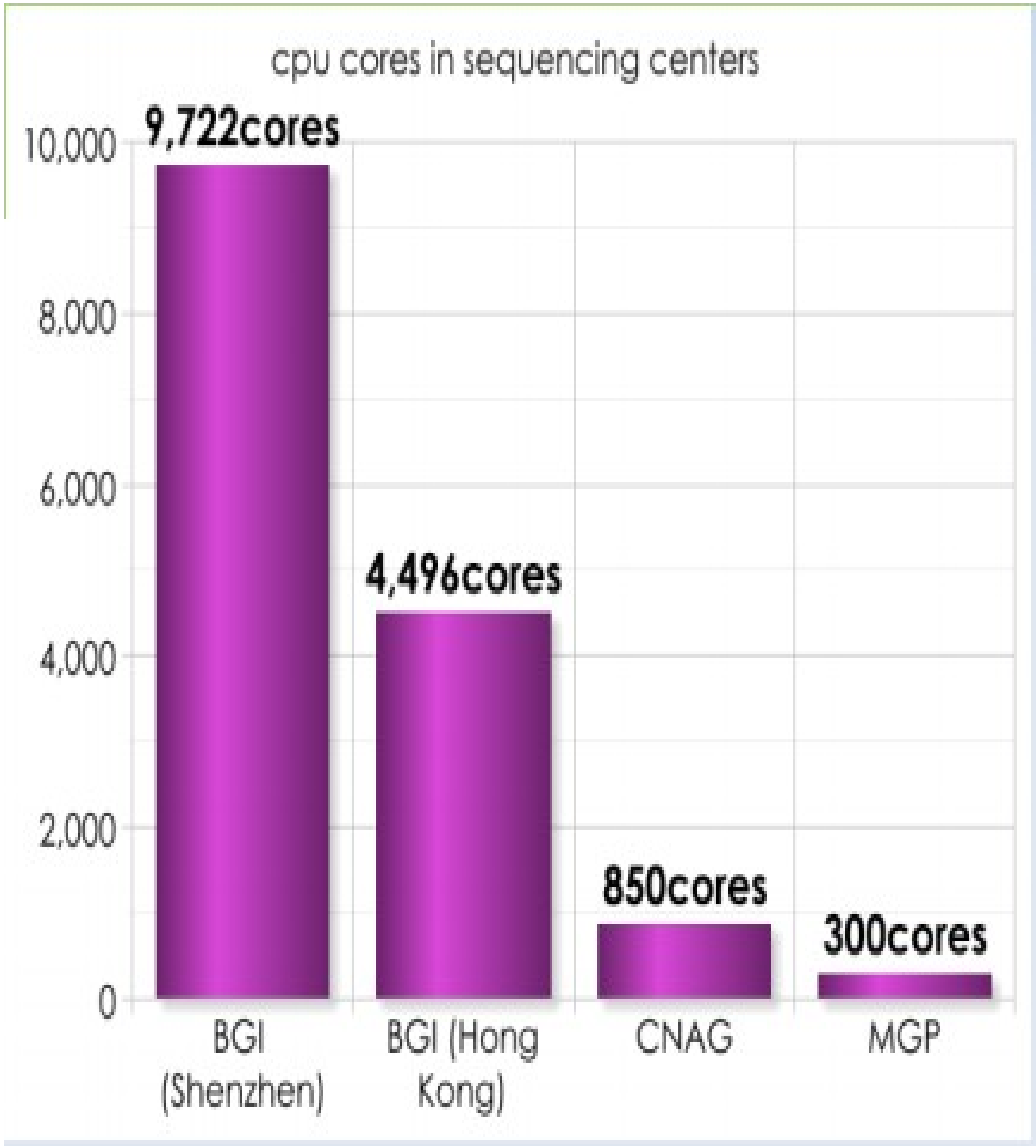
CNAG

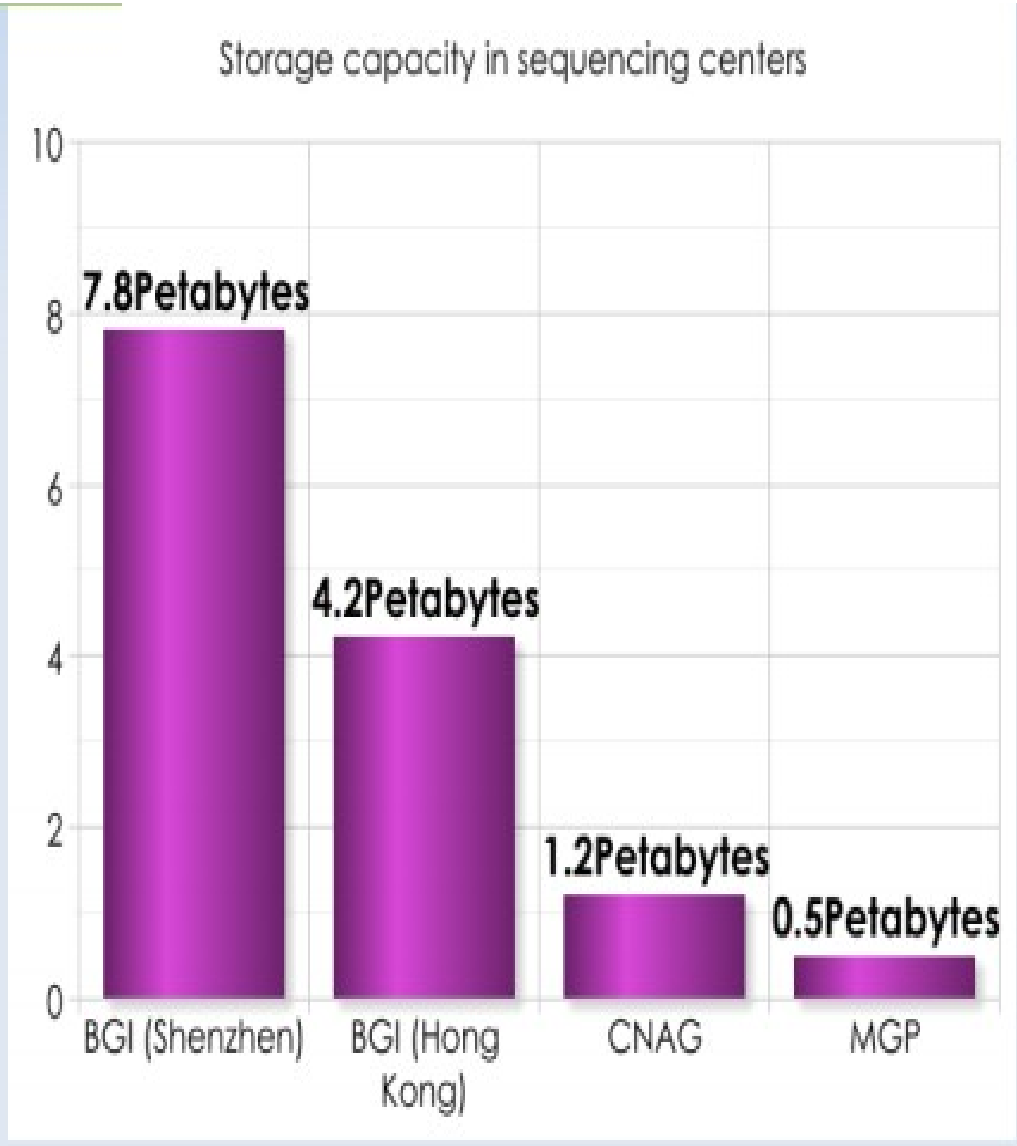
- Sequencing instruments
  - 10 Illumina HiSeq2000
- Informatics infrastructure
  - 850 core cluster
  - 1.2 PB (petabytes), lustre filesystem
  - 10 x 10 Gb/s link with MareNostrum (the most powerful supercomputer in Spain: 10240 cores)

BGI

- Sequencing instruments
  - Illumina HiSeq
  - AB Solid System
  - Ion Torrent
- Informatics infrastructure
  - 20576 cores cluster
  - 17 PB (petabytes)

Sequencing center resources





Alternatives: cloud computing

- Pros
  - Flexibility
  - You pay what you use
  - Don't need to maintain a data center
- Cons
  - Transfer big datasets over internet is slow
  - Lower performance, specially in disk read/write
  - Privacy/security concerns
  - More expensive for big and long term projects

THANKS