BIG4: Biosystematics, informatics and genomics of the big 4 insect groups- training tomorrow's researchers and entrepreneurs

Kick-Off Meeting

14-18 September 2015

Copenhagen, Denmark

BIG4
Biosystematics, Informatics, Genomics

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642241

Era7 bioinformatics

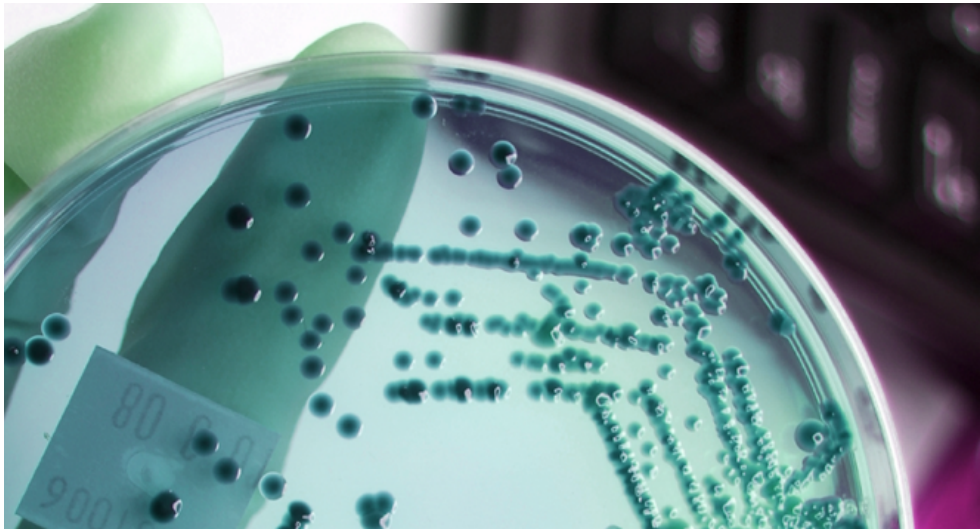
NGS equencing

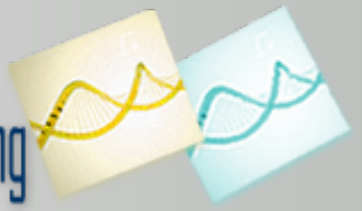
Next Generation Sequencing (NGS) Revolution

14 of September 2015

BIG4 Kick-off meeting

www.era7bioinformatics.com



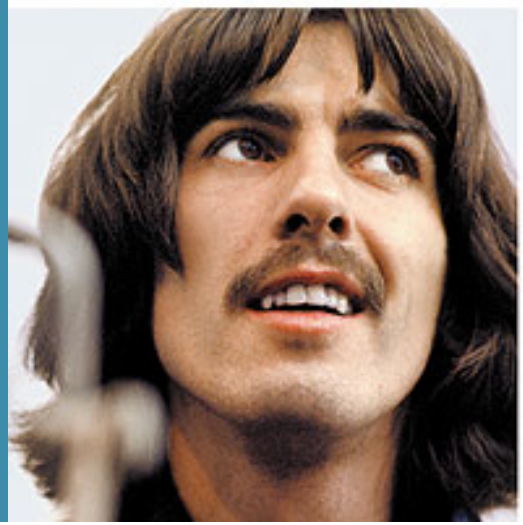


A bit of context ….

What is Era7 ?

(Some people does not understand why we are in a project related with “beatles”)



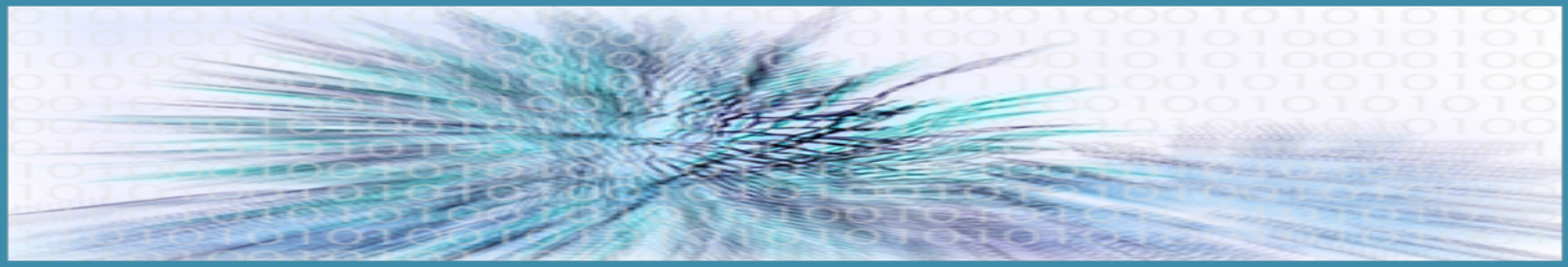


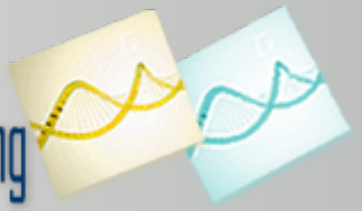

Era7 bioinformatics

NGS equencing

We think that bioinformatics is not an objective in itself but a tool that has sense to get

New Biological Knowledge





The most important thing in Era7 is
our multidisciplinary, young and great team:

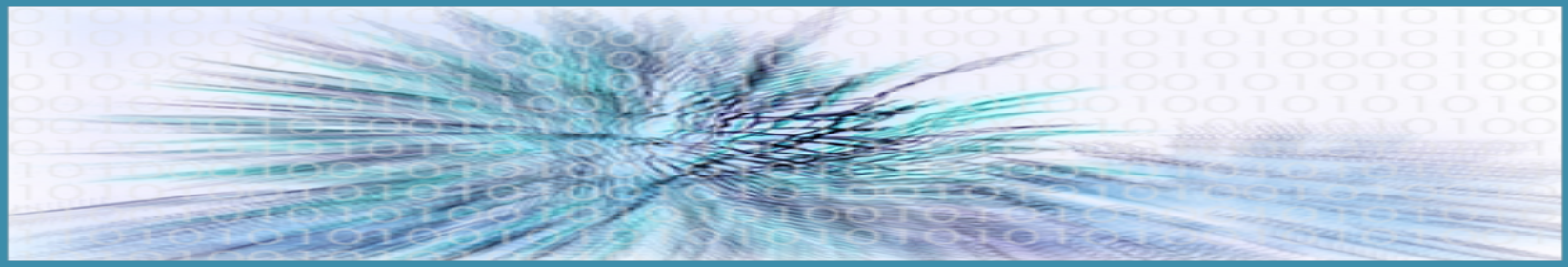
- Mathematics
- Informatics
- Bioinformatics
- Biochemistry
- Medicine

Era7 bioinformatics

NGS equencing

Covering all the spectrum and understanding different languages and partners what is needed to get

New Biological Knowledge



Era7 bioinformatics

NGS equencing

Era7 started in 2005

Granada

Era7 bioinformatics

NGS equencing

From 2010 in

Madrid

Era7 bioinformatics

NGS equencing

From 2012 in

Cambridge MA
USA

Era7 Bioinformatics Inc.

Era7 bioinformatics

NGS equencing

From 2014

A wet lab in

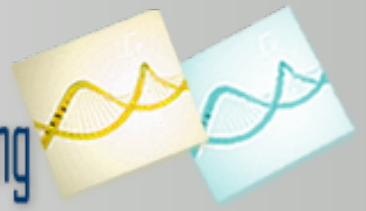
Granada

Era7 bioinformatics

NGS Sequencing

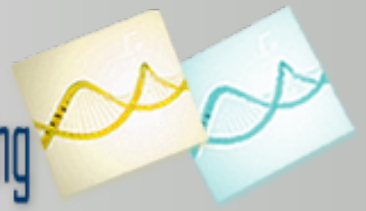
Our mission is to solve ALL the needs a customer could have about the NGS project.

Era7 bioinformatics

NGS equencing 

Innovative Business Model

- Research
- Cloud Computing
- Open Source
- Being focused in NGS



Research:

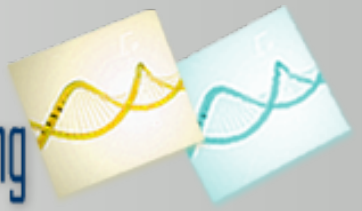
- More than 70% of our staff doing R&D
- European, national and regional projects

ohnosequences!

era7 bioinformatics R&D group

Era7 bioinformatics

NGS sequencing



Research in Era7 Bioinformatics

Research Lines:

- Algorithms for assembly
- Bacterial genome annotation
- Cloud Computing Architectures
- Graph Databases for Biological data
- Comparative genomics
- Cancer Genomics
- Genome Plasticity
- Big Data integration and visualisation
- Host Immune System and infection

Software Projects

- BG7
- MG7
- CG7
- Bio4j
- Biographika
- NEXTMICRO
- Statika
- Nispero

Cloud Computing

- Scalability (thousands of machines)
- Flexibility adapting computation to project
- Big Data Solutions
- Parallel Computation
- No need of investment:
- Lower costs
- IaaS

powered by amazon web services™

Open Source

- Freedom for our customers
- The possibility of building the best pipelines

Bio4J Graph oriented database integration (Uniprot GO Refseq)

Bio4j is a bioinformatics graph based DB including most data available in:

- UniProt (SwissProt + Trembl)
- UniRef (50,90,100)
- Gene Ontology (GO)
- RefSeq
- NCBI taxonomy
- Enzyme DB

Big Data:

2.000.000.000 relationships
400.000.000 nodes
1.000.000.000 properties

Era7 bioinformatics

NGS equencing

Bio4j project selected this year among 190 projects world wide of all areas for the

Google Summer of Code 2014

Biographika

An Open Source project about data visualisation

www.biographika.com

https://vimeo.com/127019770

Era7 bioinformatics

NGS Sequencing

Our customers are researchers from:

- Research centres
- Universities
- Hospitals
- Biotech Companies
- Pharma Companies
- Agrifood Companies

Era7 bioinformatics

NGS equencing

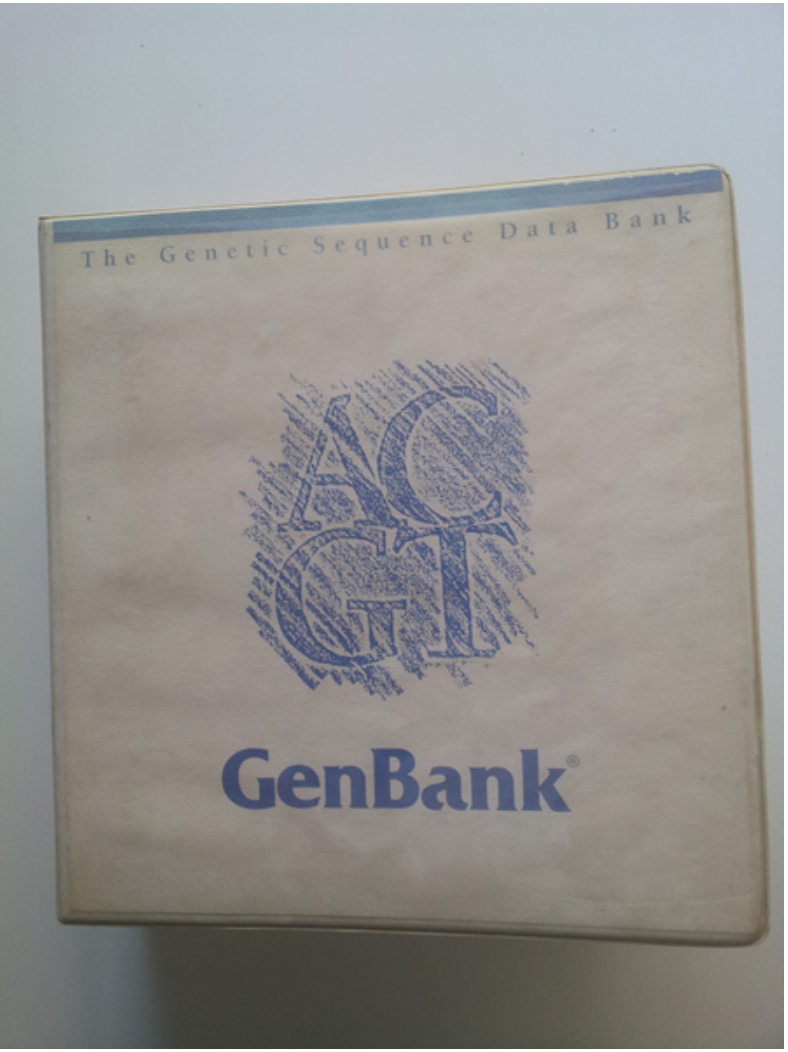
Our Customers

NGS (Next Generation Sequencing) is a revolution. It is changing the way we do

- Genomics
- Metagenomics
- Transcriptomics
- Amplicon Sequencing
- ……..

Why a Revolution ?

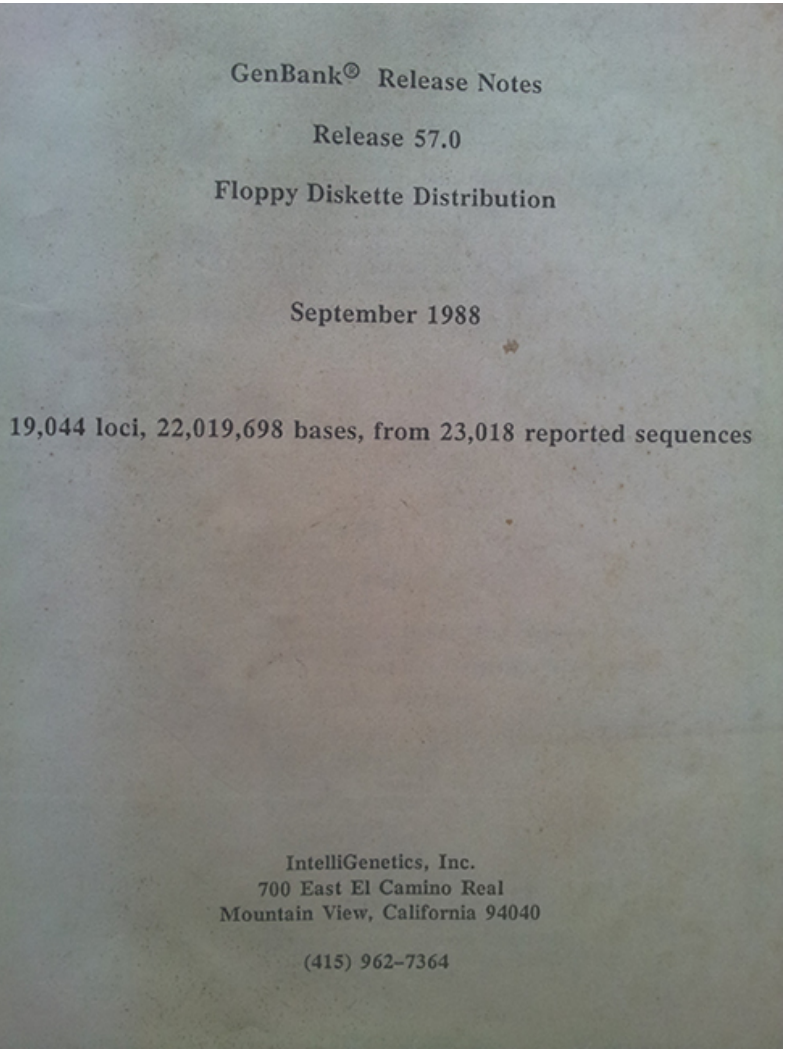
- Quicker and Easier
- Much more throughput
- Much less Price



1988

NCBI did not exist at this moment

Era7 bioinformatics
ohnosequences!



GenBank® Release Notes

Release 57.0

Floppy Diskette Distribution

September 1988

19,044 loci, 22,019,698 bases, from 23,018 reported sequences

IntelliGenetics, Inc.
700 East El Camino Real
Mountain View, California 94040

(415) 962-7364

1988

19,044	loci
23,018	sequences
22,019,698	bases

Era7 bioinformatics
ohnosequences!

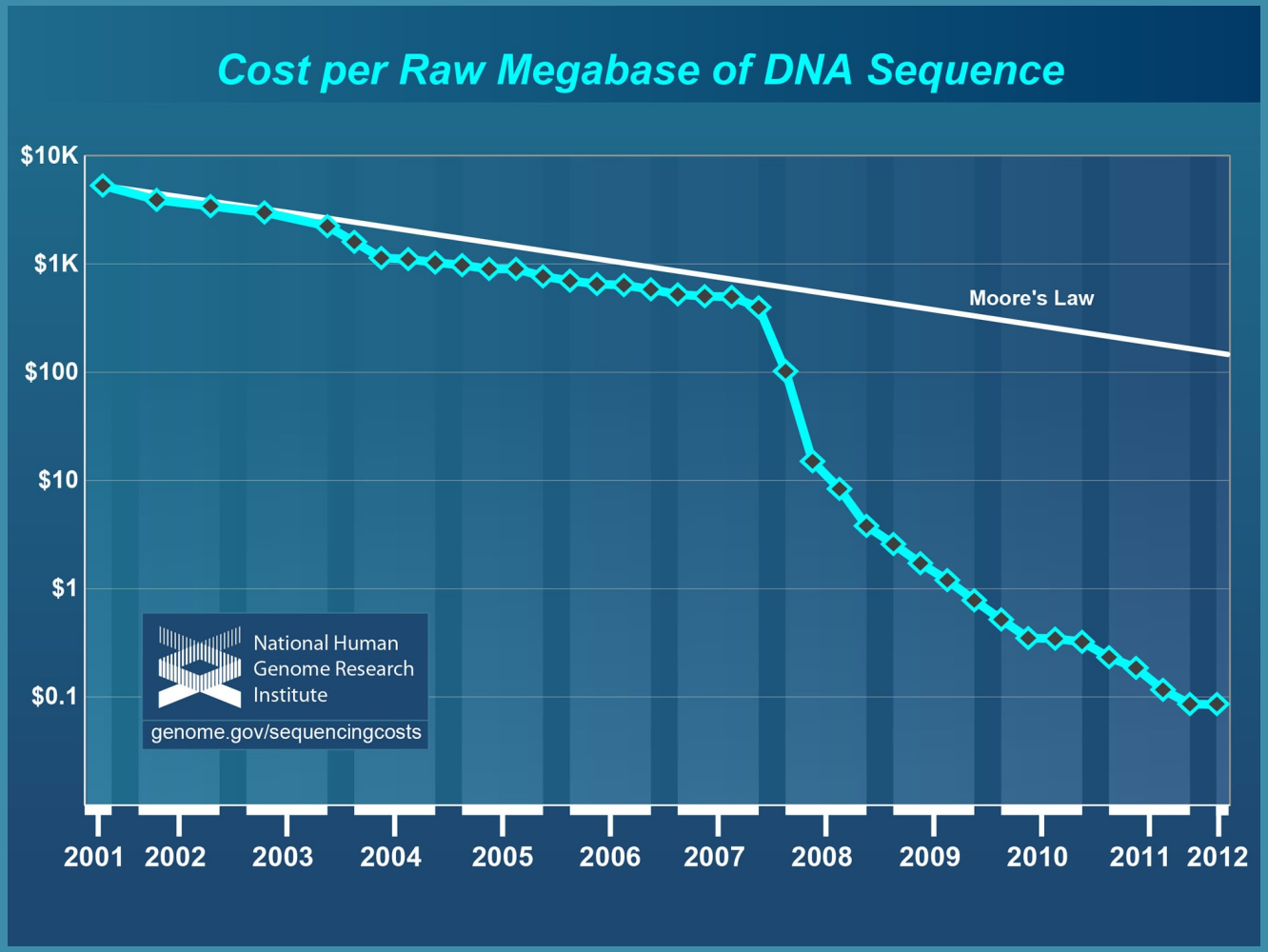
1988

360 Kb

Era7 bioinformatics
ohnosequences!

World Wide Sequencing is today more than 35 Ptb

• Kilo base	1000	nucleotides
• Megabase	1000000	nucleotides
• Gigabase	1000000000	nucleotides
• Terabase	1000000000000	nucleotides
• Petabase	1000000000000000	nucleotides



A factor of more than

1000X !!

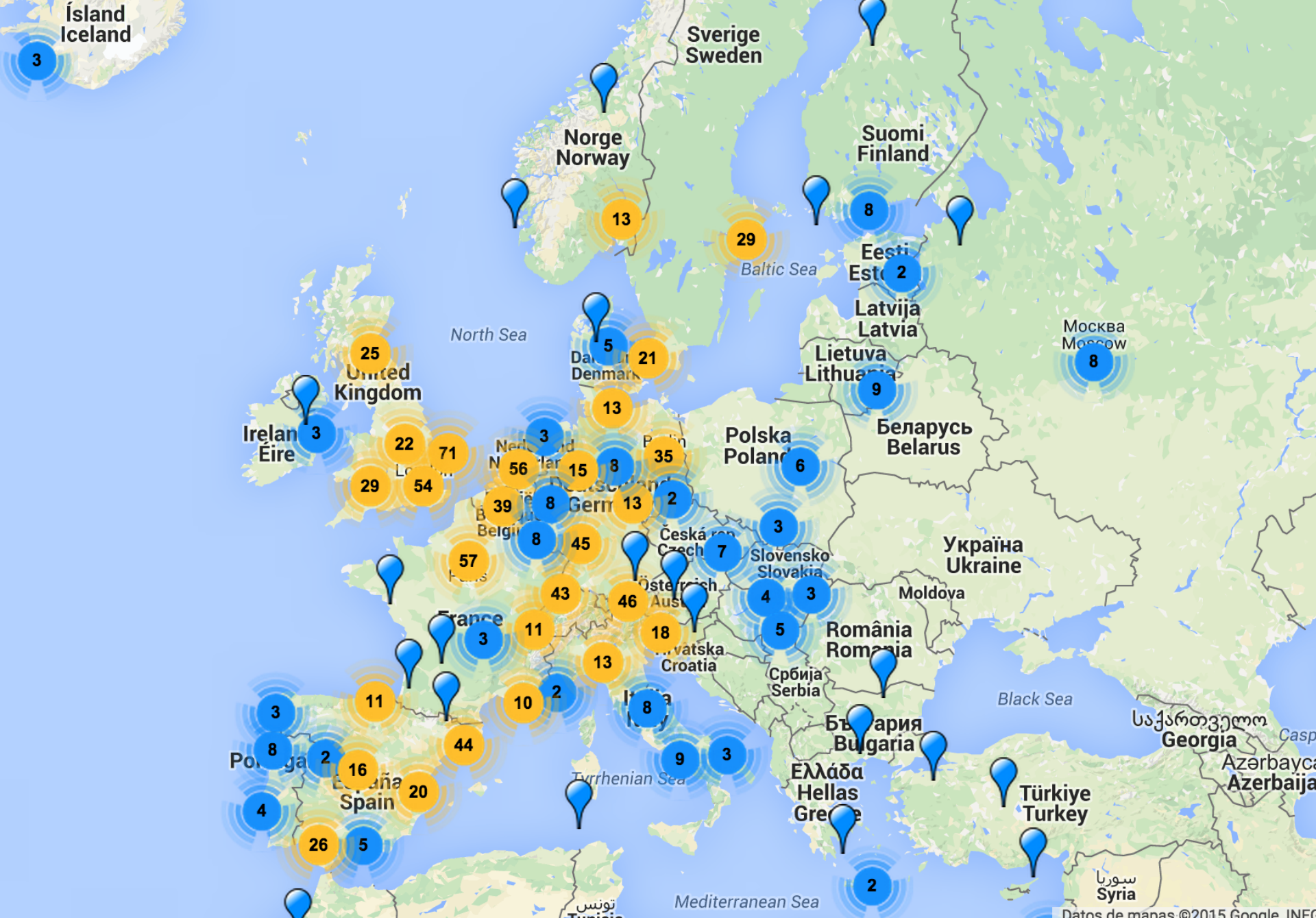
Past 7 years

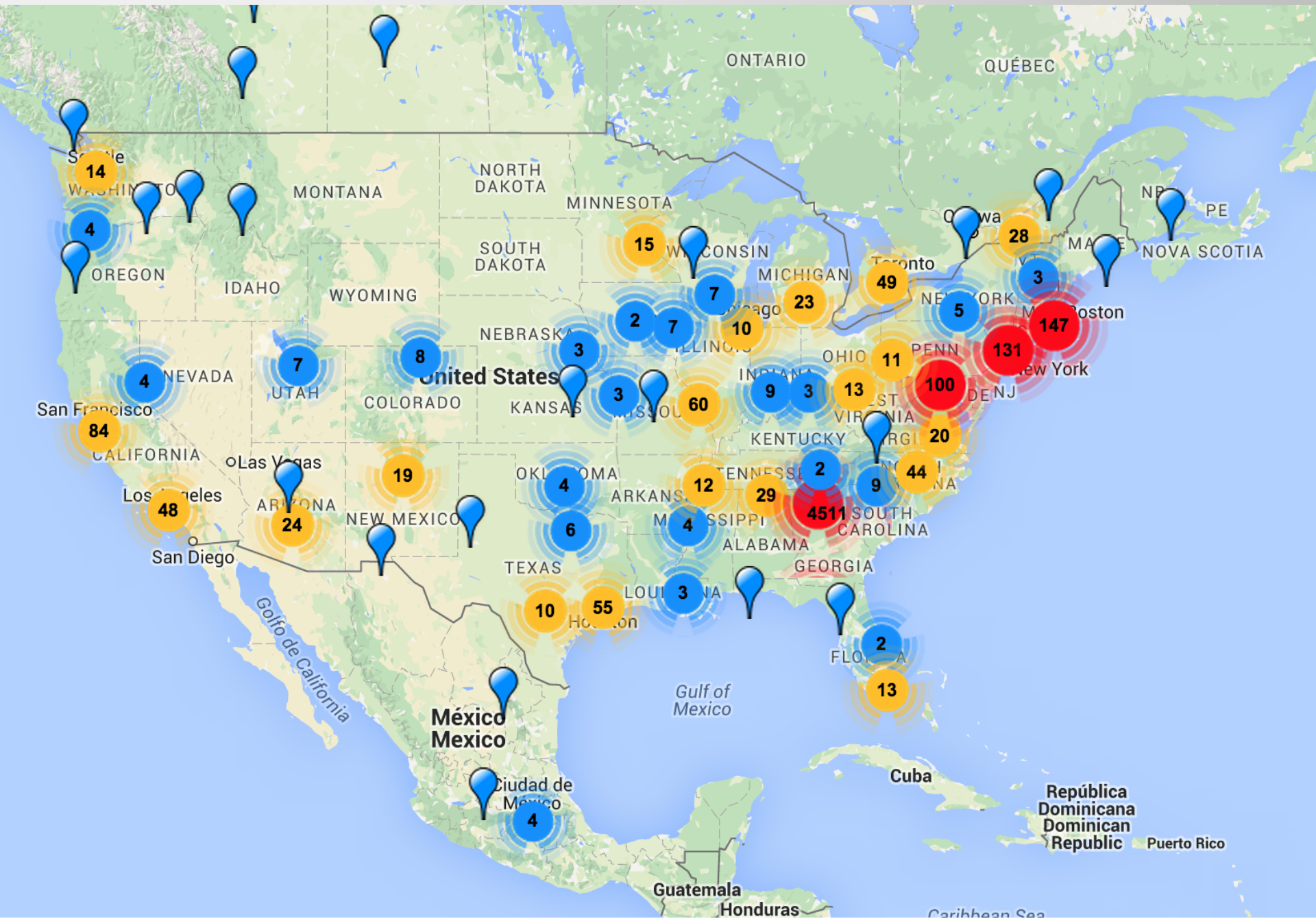
=

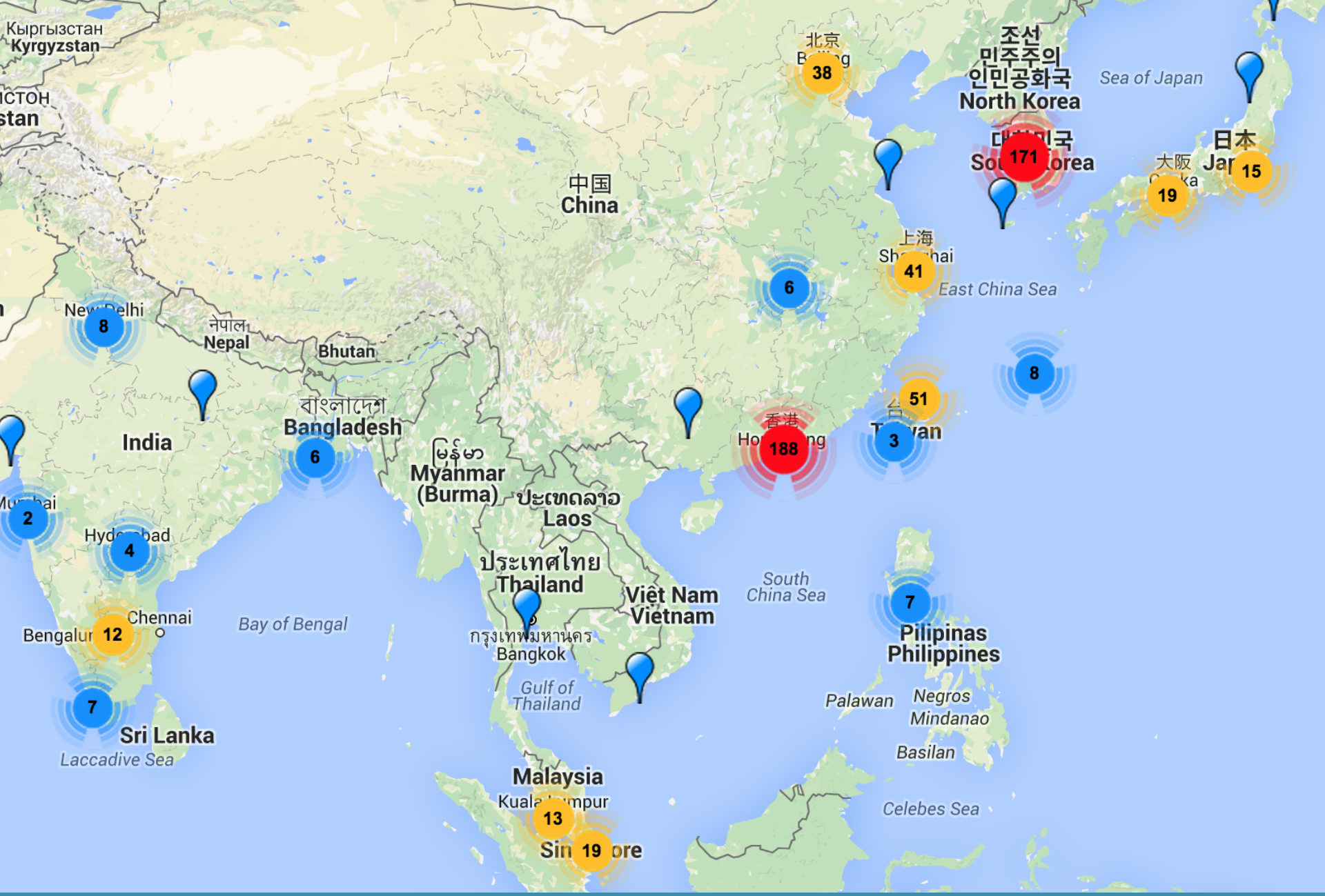
NGS

- Total amount of sequence data produced doubling approximately every seven months
- The OmicsMaps reports that there are more than 2,500 high-throughput instruments located in nearly 1,000 sequencing centers in 55 countries in universities, hospitals, and other research laboratories

Plos Biology 10.1371/journal.pbio.1002195


Ísland
Iceland
Sverige
Sweden
Norge
Norway
Suomi
Finland
Eesti
Latvija
Latvia
Lietuva
Lithuania
Москва
Moscow
Беларусь
Belarus
Polska
Poland
Україна
Ukraine
Moldova
România
Slovensko
Slovakia
Česká
Czech
Österreich
Hrvatska
Croatia
Србија
Serbia
България
Bulgaria
Ελλάδα
Hellas
Türkiye
Turkey
საქართველო
Georgia
Azərbaycan
Azerbaijan
سوريا
Syria
United Kingdom
Ireland
Éire
France
Spain
North Sea
Baltic Sea
Black Sea
Tyrrhenian Sea
Mediterranean Sea
Datos de mapas ©2015 Google


ONTARIO
QUÉBEC
NORTH DAKOTA
MONTANA
MINNESOTA
SOUTH DAKOTA
OREGON
IDAHO
WYOMING
NEBRASKA
MICHIGAN
NEVADA
UTAH
COLORADO
KANSAS
OHIO
INDIANA
KENTUCKY
San Francisco
CALIFORNIA
Las Vegas
Los Angeles
San Diego
ARIZONA
NEW MEXICO
OKLAHOMA
ARKANSAS
TEXAS
ALABAMA
GEORGIA
SOUTH CAROLINA
FLORIDA
Gulf of Mexico
Golfo de California
México
Mexico
Ciudad de México
Cuba
República Dominicana
Dominican Republic
Puerto Rico
Guatemala
Honduras
Caribbean Sea
NOVA SCOTIA
Toronto
Boston
New York
14
4
84
48
24
19
7
8
4
3
2
7
15
7
23
49
10
3
60
9
3
13
11
100
147
131
5
28
3
20
44
2
9
4511
29
12
4
6
10
55
3
2
13


Кыргызстан
Kyrgyzstan
北京
38
조선
민주주의
인민공화국
North Korea
Sea of Japan
171
日本
15
19
中国
China
上海
41
East China Sea
6
8
Nepal
Bhutan
India
বাংলাদেশ
Bangladesh
6
51
3
188
Myanmar
(Burma)
ประเทศลาว
Laos
ประเทศไทย
Thailand
Việt Nam
Vietnam
South
China Sea
2
4
12
Chennai
Bay of Bengal
Bangkok
7
Pilipinas
Philippines
Palawan
Negros
Mindanao
Basilan
Gulf of
Thailand
Sri Lanka
Laccadive Sea
Malaysia
13
19
Celebes Sea

NGS

~32,000 microbial genomes, ~5,000 plant and animal genomes, and ~250,000 individual human genomes that have been sequenced or are in progress thus far

- 100.000 genomes UK project
- 1.000.000 genomes Obama Precision Medicine Initiative
- 1.000.000 plant and animals genomes. (BGI)

(The proof of the anthropocentric vision)

Plos Biology 10.1371/journal.pbio.1002195

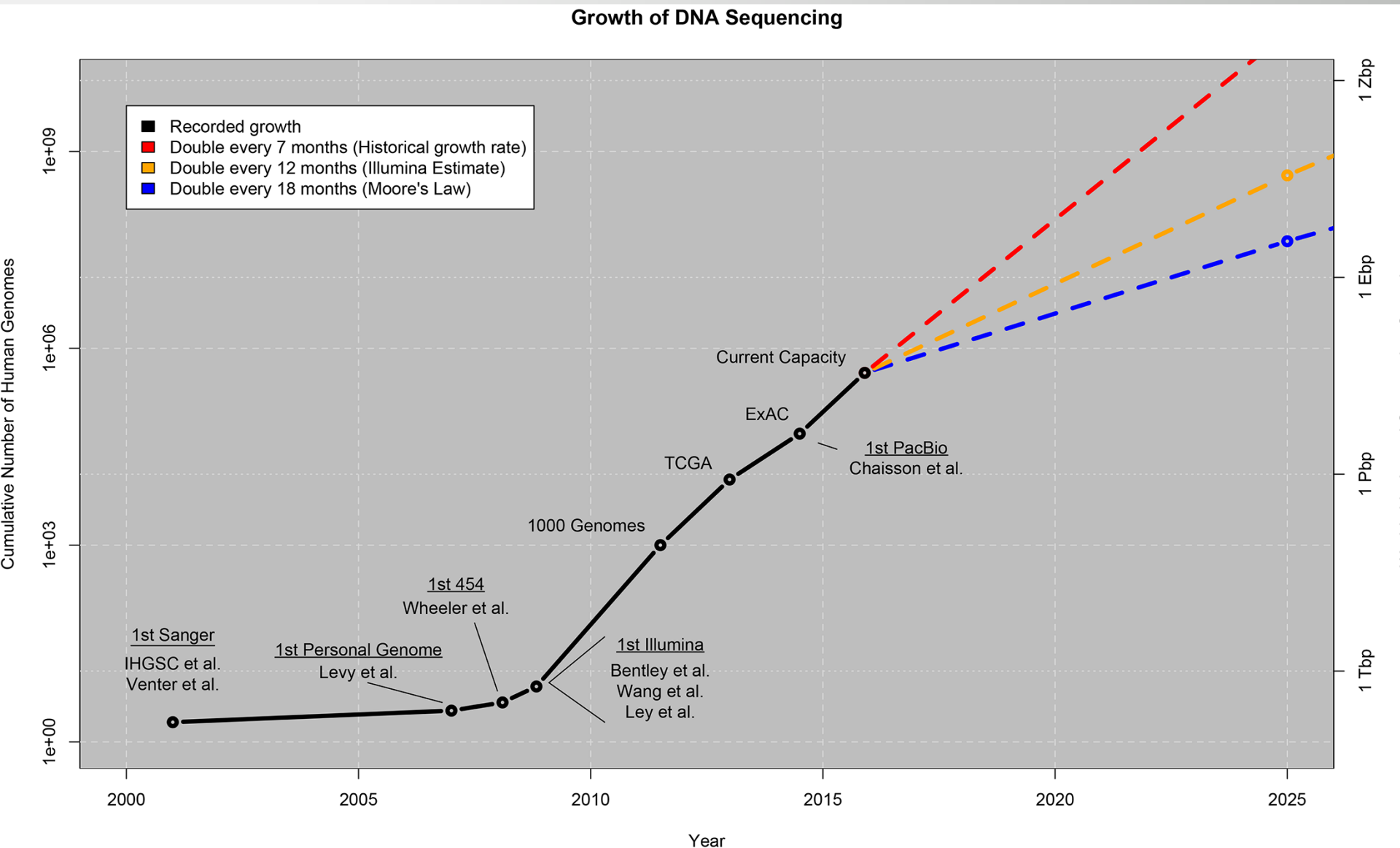
NGS

We therefore estimate between 100 million and as many as 2 billion human genomes could be sequenced by 2025 far exceeding the growth for the three other Big Data domains. Indeed, this number could grow even larger, especially since new *single-cell genome sequencing* technologies are starting to reveal previously unimagined levels of variation, especially in cancers, necessitating sequencing the genomes of thousands of separate cells in a single tumor

Plos Biology 10.1371/journal.pbio.1002195

Era7 bioinformatics

NGS Sequencing



Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

NGS Technologies:

- illumina
- Ion Torrent
- PacBio
- Oxford Nanopore
- Complete Genomics
- (SOLiD)
- (454)

illumina is the market leader

More than 90% of sequenced nucleotides has been sequenced with illumina technology

Era7 bioinformatics

NGS Sequencing

Sequencing

MiSeq
Focused Power

MiSeqDx
Focused Power

MiSeq FGx
Focused Power

NextSeq 500/550
Flexible Power

HiSeq 2500
Production Power

HiSeq 3000
Production Power

HiSeq 4000
Production Power

HiSeq X Five
Population Power

HiSeq X Ten
Population Power

Sequencing Library Prep

NeoPrep
Powerfully Simple

Arrays

HiScan
Flagship Scanner

iScan
Cutting-edge Scanner

Era7 bioinformatics

NGS sequencing

				
	MiSeq **Focused power.** Speed and simplicity for targeted and small genome sequencing.	**NextSeq 500** **Flexible power.** Speed and simplicity for everyday genomics.	**HiSeq 2500** **Production power.** Power and efficiency for large-scale genomics.	**HiSeq X*** **Population power.** $1,000 human genome and extreme throughput for population-scale sequencing.
Key applications	Small genome, amplicon, and targeted gene panel sequencing.	Everyday genome, exome, transcriptome sequencing, and more.	Production-scale genome, exome, transcriptome sequencing, and more.	Population-scale human whole-genome sequencing.

Run mode	N/A	Mid-Output	High-Output	Rapid Run	High-Output	N/A
Flow cells processed per run	1	1	1	1 or 2	1 or 2	1 or 2
Output range	0.3-15 Gb	20-39 Gb	30-120 Gb	10-180 Gb	50-1000 Gb	1.6-1.8 Tb
Run time	5-65 hours	15-26 hours	12-30 hours	7-40 hours	< 1 day - 6 days	< 3 days
Reads per flow cell†	25 Million‡	130 Million	400 Million	300 Million	2 Billion	3 Billion
Maximum read length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 125 bp	2 × 150 bp

Ion Torrent

It has some applications

Ion PGM™ Chip	Run time		Output	
	200 bp read	400 bp read	200 bp read	400 bp read
Ion 314™ Chip v2	2.3 hr	3.7 hr	30–50 Mb	60–100 Mb
Ion 316™ Chip v2	3.0 hr	4.9 hr	300–500 Mb	600 Mb–1 Gb
Ion 318™ Chip v2	4.4 hr	7.3 hr	600 Mb–1 Gb	1.2–2 Gb
Ion Proton™ Chip	Run time		Output	
	200 bp read		200 bp read	
Ion PI™ Chip	2–4 hr		Up to 10 Gb	

Era7 bioinformatics

NGS equencing

Pacific Biosciences

The advantage of long reads

Era7 bioinformatics

NGS equencing

PACIFIC BIOSCIENCES®

SISTEMATIC ERROR in illumina

RANDOM ERROR in PacBio

Era7 bioinformatics

NGS equencing

Oxford Nanopore MinION

The advantage of not being really a machine

Oxford Nanopore Long reads, still a lot of errors.

Era7 bioinformatics

NGS Sequencing

Oxford Nanopore Ebola Sequencing

NATURE NEWS

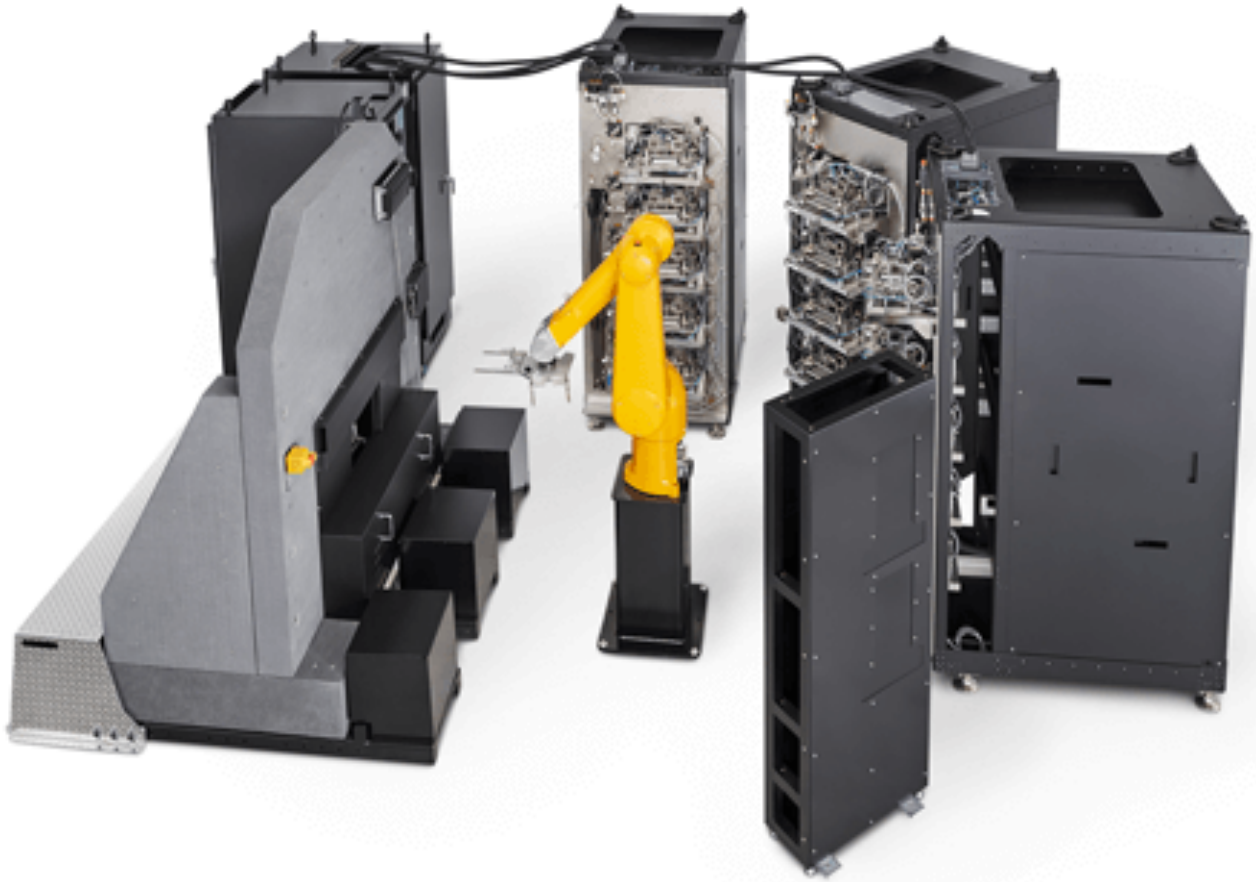
05 May 2015

Complete Genomics (BGI)

Now, you can buy machines….

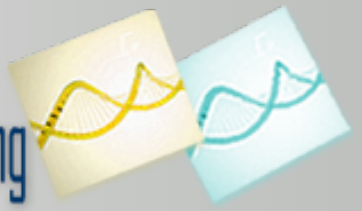
Era7 bioinformatics

NGS equencing



Revolocity: From Blod to VCF Files

BGI NEWS 05 June 2015



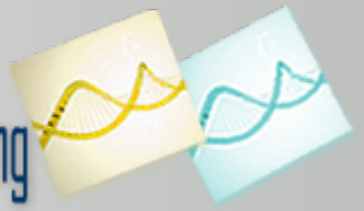
PacBio :

- Really LONG READS (8-10 kb)
- Nucleotide Modifications (no PCR)
- Large amount of DNA

PacBio for De novo Bacterial Genomics:

The Best !!!

- "Finished" Genome
- Chromosome and plasmids assembled in one experiment !!

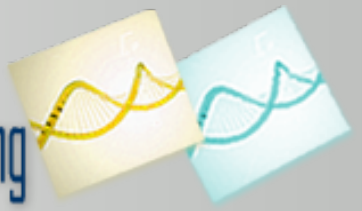


PacBio :

- To improve Complex Genome Assembly
- Impressive information about Isoforms
- Large amplicons (haplotype info)

With such amount of sequencing data

- Storage
- Analysis
- Interpretation



There is a growing need of

Bioinformatics

Some concepts:

- De novo Genome
- Resequencing
- Shotgun
- amplicon sequencing
- RNA-seq
- Exome
- Enrichment
- Targeted sequencing
- Metagenomics
- Metatranscriptomics
- Libraries
- Run, Lane, multiplexing…

Some bioinformatics concepts

- Assembly
- Mapping
- Annotation
- ……..

Some kind of applications

Era7 bioinformatics

Bacterial Genomics:

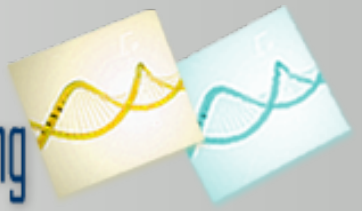
- Design
- Sequencing
- Assembly
- Annotation
- Submitting
- Comparative Genomics (SNPs and Whole Genome)
- Interactive Data Visualisation
- Databases and Cloud Solutions

Bacterial Genomics:

- illumina
- PacBio
- Strategies in a set of isolates

Era7 bioinformatics

NGS equencing



Comparative genomics

- Genome alignment
- Ortholog tables
- Strain exclusive genes and genomic regions
- SNPs
- Innovative interactive visualisation

Metagenomics:

- 16S
- Shotgun
- (Consortium)

Metagenomics:

- 16S
  - Squencing.
  - Analysis:
    - alfa – diversity
    - beta – diversity

Bacterial Metagenomics:

- 16S

  Primers

  Target regions

  Read length….

illumina, MiSeq (2x300) and now HiSeq (2x250) (454)

Metagenomics:

- Shotgun
  - Sequencing
  - Analysis:
    - Functional Profile
    - Specific Functional Profiles:

Metagenomics:

- Consortium,
    - A few species
    - Usually < 10
    - Potential assembly
    - Useful:
        - Infections
        - Industrial Biotechnolgy

Application of Metagenomics:

HUMAN MICROBIOME AND HEALTH

Microbiome influence in health:

- Microbiome of mother and child
- Microbiome of centenary individuals
- Microbiome of lean individuals

Microbiome and diseases:

- Microbiome and Cancer
  - In cancer prevention
  - In cancer pathogeny
  - Metagenomics as biomarker
- Microbiome and Obesity
- Gut Microbiome as biomarker of chronic inflammatory diseases

Application of Metagenomics:

- Microbiome and diseases:
  - Oral microbiome and cardiovascular diseases
  - Nervous System and gut microbiome
  - Aging and gut microbiome
  - Microbiome and allergy
  - Microbiome and Inflammatory Bowel Diseases
  - Microbiome and autoimmune diseases
- Microbiome and infections
  - Microbiome and antibiotic treatments
  - Study of Commensal/Pathogen microbial organisms
  - Biofilms
  - Virome study
  - Mobilome study
  - Resistome study

Application of Metagenomics:

- Microbiome and infections
  - De novo Diagnosis of nonculturable microbial pathogens
    - Microbial fungi
    - Viruses
    - Bacteria

- METAGENOMICS IN VETERINARY

  - Metagenomics in Veterinary has equivalent applications as in Human health
  - In many cases human diseases have a disease model in animals
  - Vector-borne pathogens studies

Application of Metagenomics:

ENVIRONMENTAL SCIENCES

Diversity in earth microenvironments to analyze:
- Climatic changes
- Pollution induced changes
- Seasonal changes

Plague defence

Water analysis

Waste water treatment plants

Application of Metagenomics:

METAGENOMICS IN AGRIFOOD SECTOR

- Conservation of food (contamination, toxins, ..)
- Fermentations:
  - Beer
  - Wine
  - Bread
  - Cheeses and other fermented dairy products (consortiums involved in artesanal products)
- Probiotics
- Crops:
  - Study of Rhizosphere communities:
    - Mycorhizas to improve crops
    - pathogens forming consortiums or complex biofilms
  - Bioremediation
  - Microbial studies of compost
  - Crop microbial plagues research

Application of Metagenomics:

BIOTECHNOLOGY AND DRUG DISCOVERY

- Specific habitats Diversity:
  - Searching for specific activities for bioenergy industry
  - Digestions in industrial reactors
- Research in Marine Samples
  - Searching for new active compounds:
    - Cancer research: New anticancer drugs
    - New antibiotics
    - New drugs
  - Marine Diversity:
    - Secondary metabolism cluster detection from samples of marine banks
- Metagenomes of extreme microenvironments
  - New enzymes (especially from extremophile bacteria) for advanced biotechnological uses
  - Microorganisms with special resistence to salt, temperature, toxics, pollutants,..
  - Astrobiology

Bacterial RNA – seq

- Sequencing
- Analysis

“Dual” RNA – seq:

- One experiment providing RNA – seq info:
  - Bacteria (Micro-organism)
  - Host
- Analysis:
  - Time Points
  - Deciphering networks

- Bacterial Genomics in medicine:
  - Infectious Diseases
  - Autoimmune Diseases
  - MICROBIOME

- Human (and mouse) Genome
- Human exome
- RNA-seq

Era7 bioinformatics

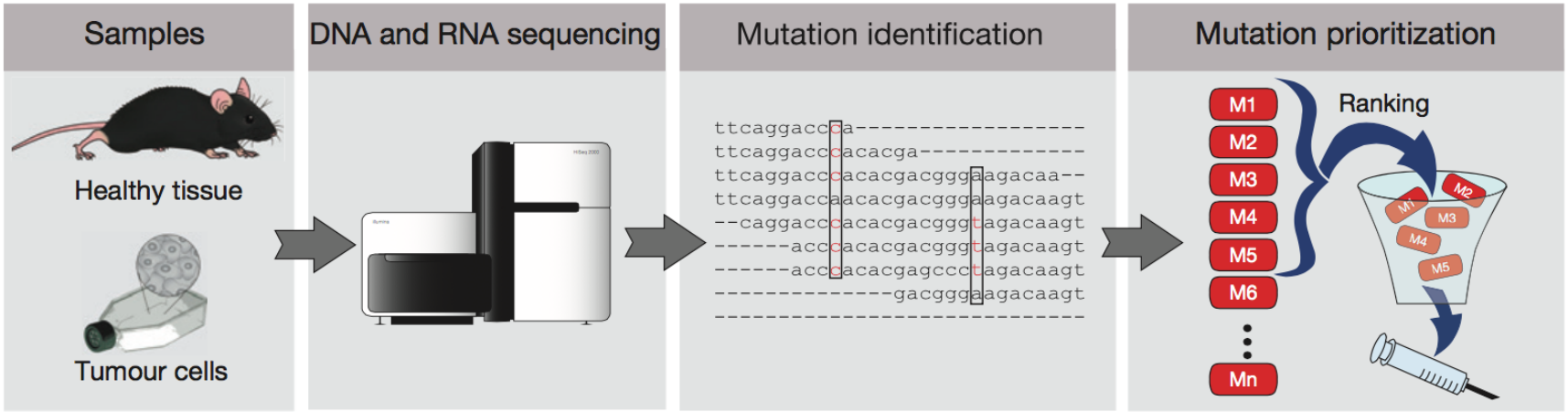
NGS equencing

Cancer Immunotherapy based on neoantigens

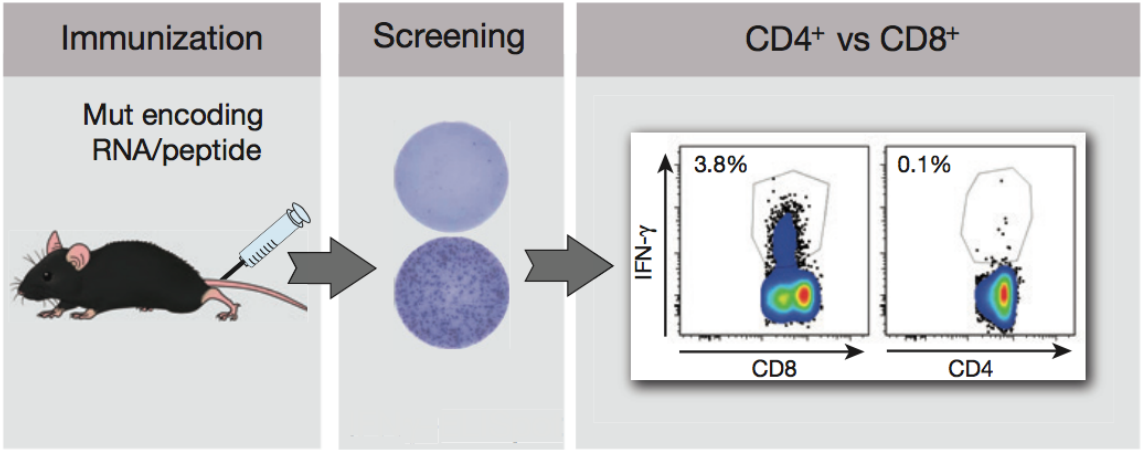
Era7 bioinformatics

NGS equencing

Mutation discovery and prioritization



Immunogenicity testing



Era7 bioinformatics

Cancer Immunotherapy based on neoantigens. NGS and Bioinformatics.

- RNA-seq analysis
- Bioinformatics selection
- Future new criteria (not only immunogenicity)
- The “real” Personalised Medicine

RNA – seq

- De novo
- With model organism

RNA – seq

- mRNA
- miRNA
- lncRNA

Era7 bioinformatics

NGS equencing

Viral Genomics and Metagenomics

Viral Amplicons (HIV etc.)

Different kinds of amplicons

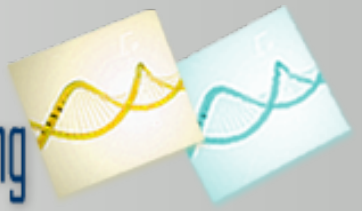
- HLA
- Different Large genes

Remember the possible advantages of PacBio Technology

Immunogenomics

- T cell Repertoires
- B cell Repertoires
- HLA

Era7 bioinformatics

NGS equencing 

Projects

- de novo Genomics
- Genomics with reference (resequencing)
- Exome
- Metagenomics
- Comparative Genomics
- Transcriptomics (mRNA, miRNA, lncRNA)
- Dual RNA-seq (Pathogen and Host)
- SNPs (and whole genome comparison)
- ChiP-Seq

- Cancer Genomics
- Big Data analysis :

  Public data sets
- Evolutionary studies
- Antibiotic resistance
- Bacteriocins
- Metabolic Engineering
- Biofuels
- Agrifood

Next future in Hospitals:

- Microbiology:
  - Metagenomics and diagnostics
    - 16S for bacteria
    - Shotgun agnostic
  - Genomics (epidemiology, low cost soon)
- Oncology
- Immunology
- Many: Microbiome(s)

RNA-seq blood analytics?

An example of use of NGS

Big4 project

Era7 bioinformatics

Thank you for your attention !

Eduardo Pareja

epareja@era7.com
info@era7.com

www.era7bioinformatics.com