Gaia: at the frontiers of astrometry

Processing massive datasets in genomics

François Artiguenave - 08 / 06 /2010

2010

Genoscope (National Sequencing Center)

- Created en 1997

- 150 people

- Service for academic collaborative projects

- Support in house R&D research projects

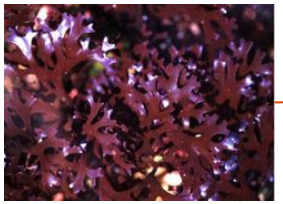
- Large-scale sequencing projects

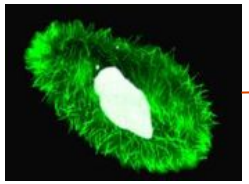
  - Part of international projects : Human Genome (K14), Arabidopsis, Rice, Medicago, Anopheles etc.
  - Main actor for Tetraodon, Oikopleura, Grape, or Paramecium, Truffle …. Banana, Coffee, Cacao, wheat 3b chromosome.

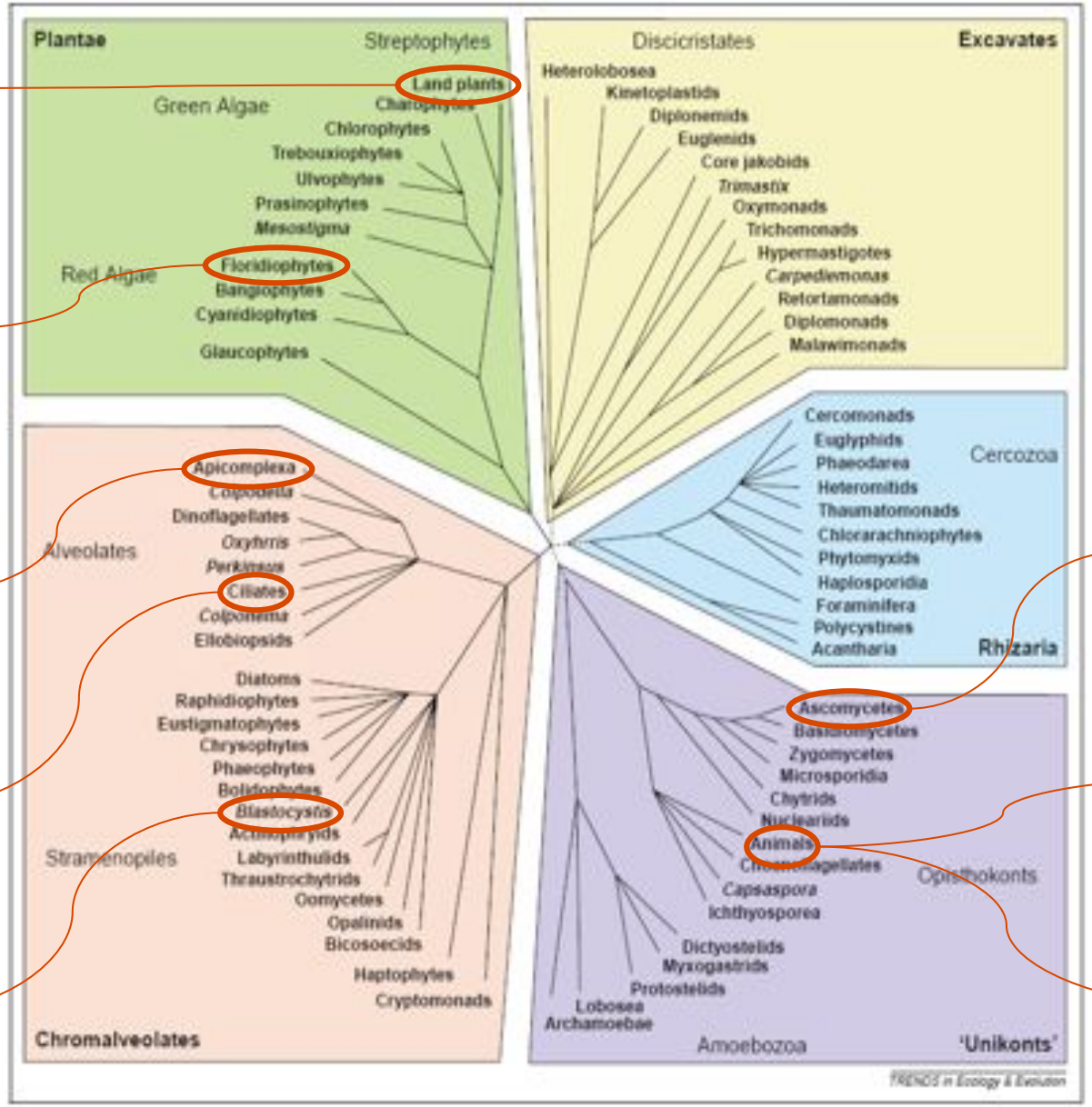
- Realized different fungi genome projects (Botrytis, Tuber) and many prokaryote sequencing projects

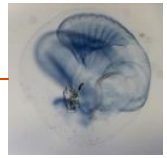
- Long standing experience in prokaryote and eukaryote genome annotation

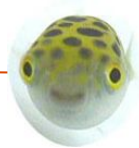
Genome projects









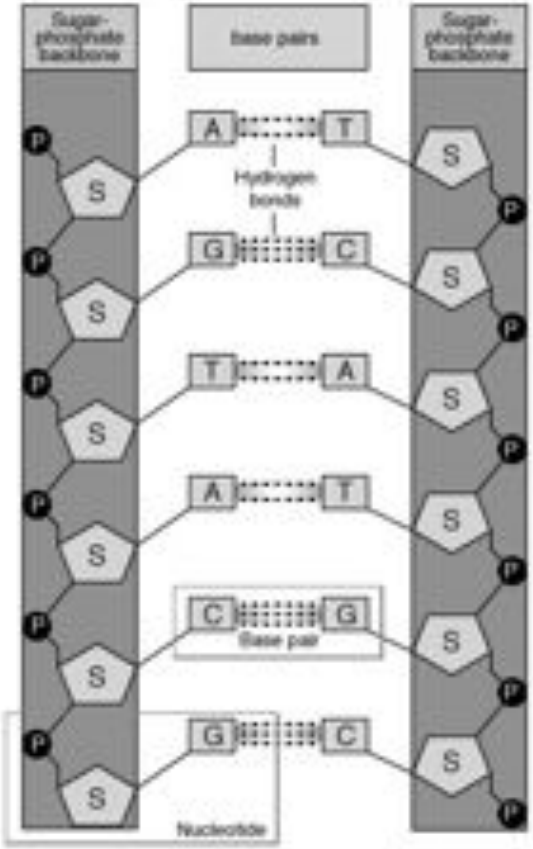


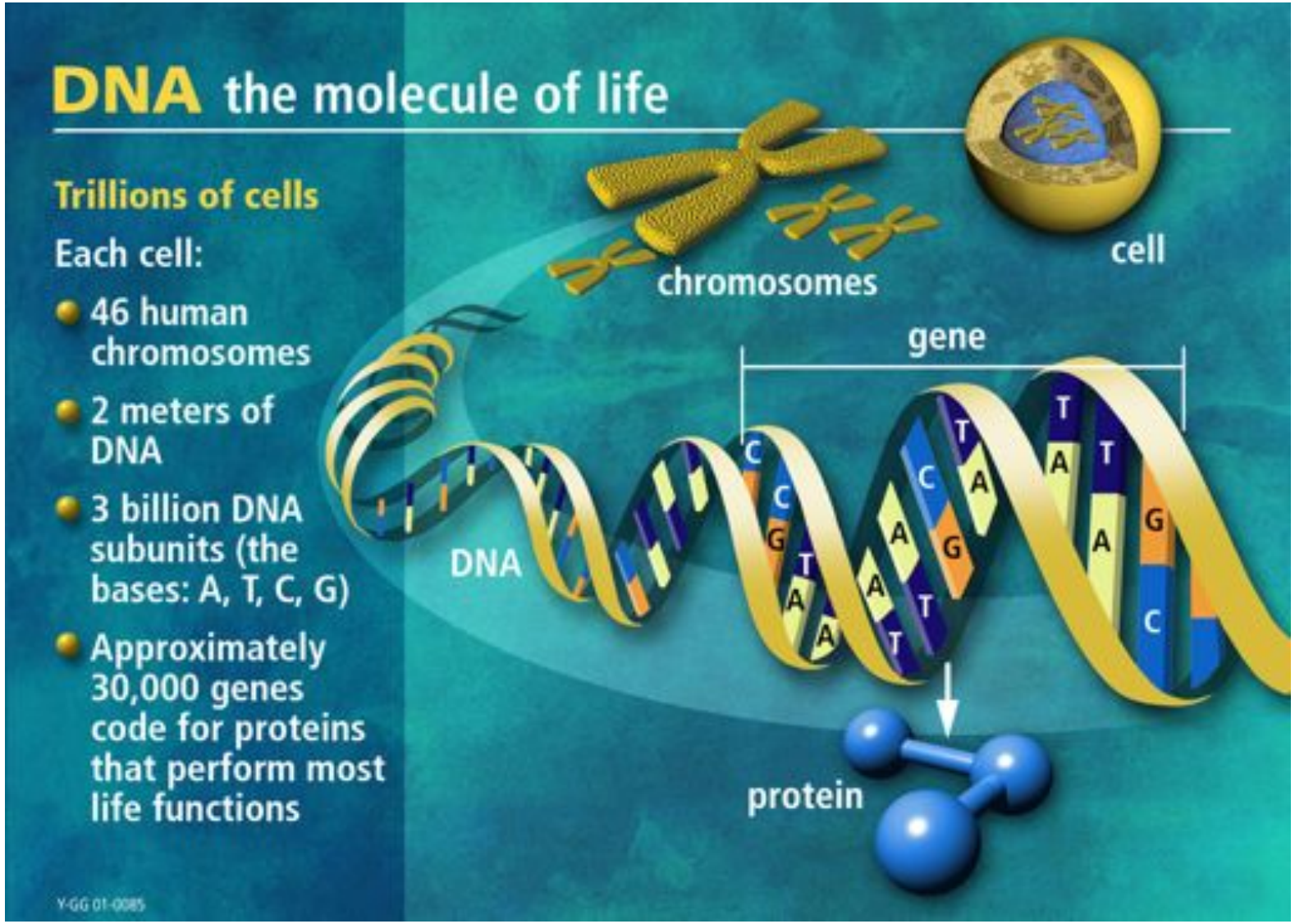
What is genomics ?

- Genomics is “the study of functions and interactions of all the genes in the genome, including their interactions with environmental factors.”

- A genome is “all the DNA contained in an organism or a cell, which includes both the chromosomes within the nucleus and the DNA in mitochondria… all our genes together.”

DNA : the molecule of life





U.S. Department of Energy Human Genome Project ~ www.ornl.gov/hgmis

Annotation of Genomes

Data Distribution

Genome browser
Web site
Submission
Formatting
etc…

Masking

known repeats

Data Collection

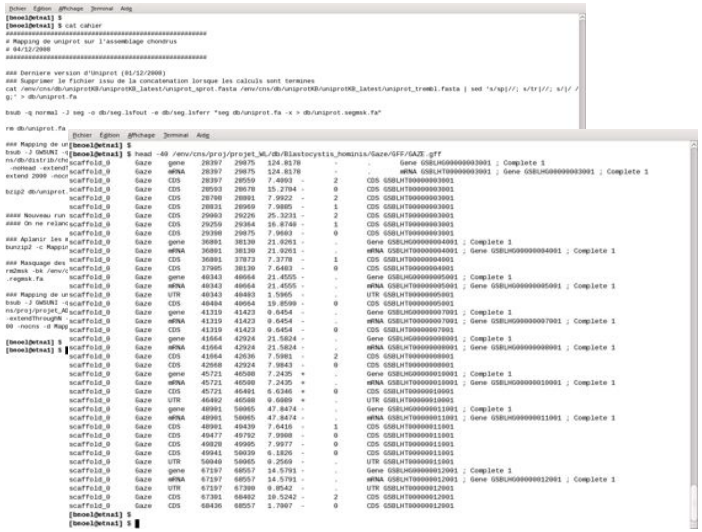
cDNAs mapping
Public protein mapping
Public ESTs mapping
Gene models *ab initio* predictions
Repeats *ab initio* detections

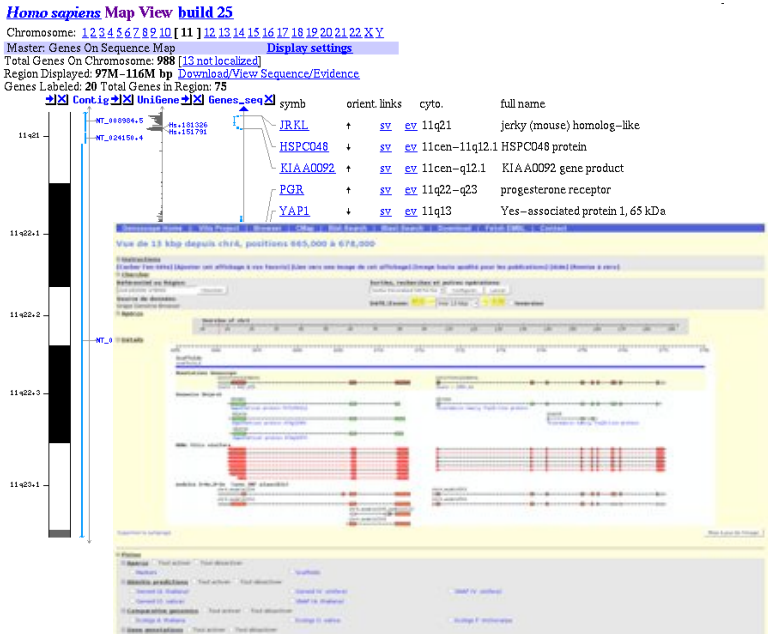
Integration

Gene models prediction

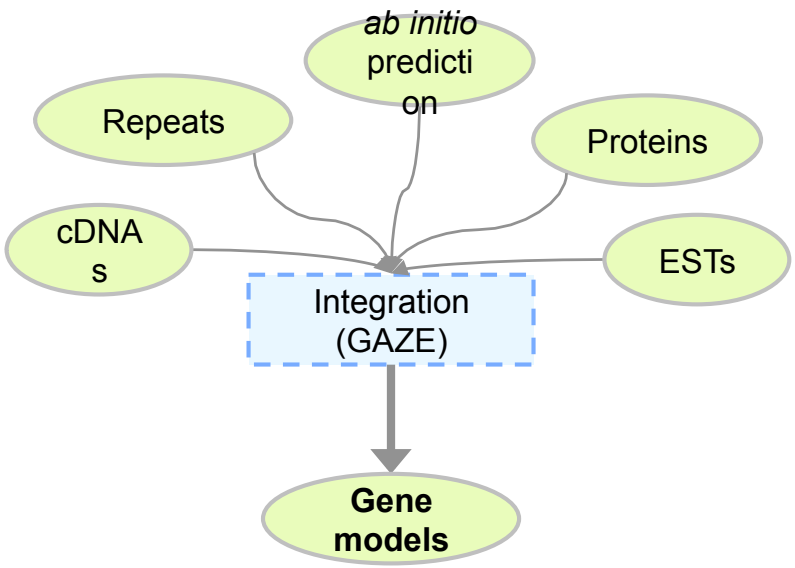
Post Annotation Analysis

Proteic domains detection
Paralogs/Orthologs definition
Enzymatic activity detection
Metabolic pathway inference





Annotation of Genomes



Data integration

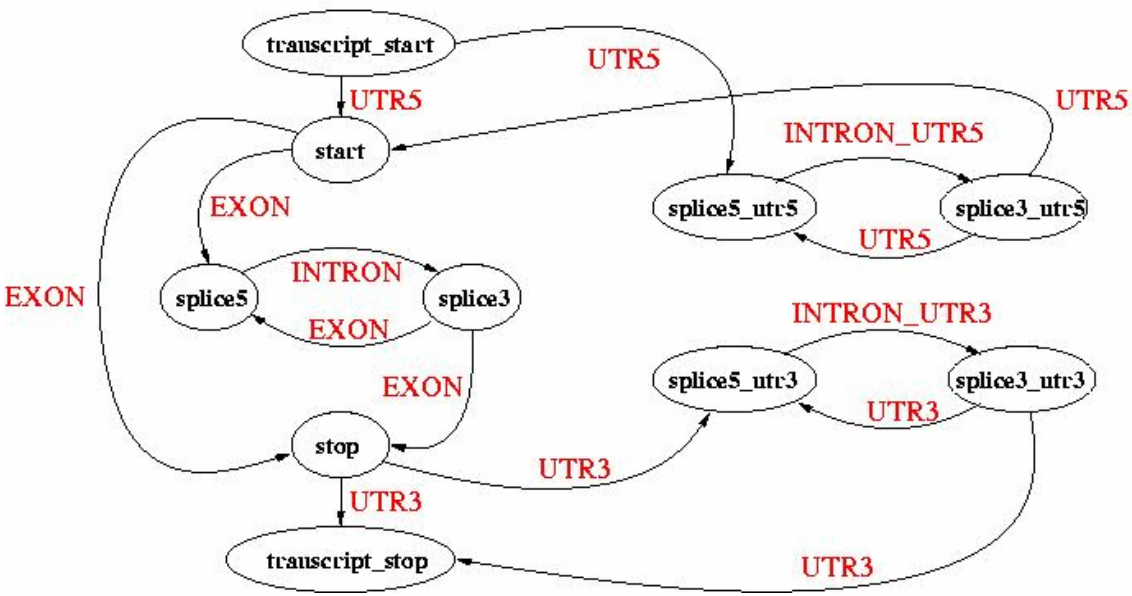
Objective : Build gene models using whole dataset (expression products, proteins, ab initio (machine learning) predictors).

Gene model represented as an automat

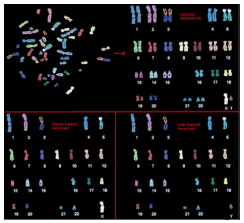
2 type of elements

segments (exons, intron,…)

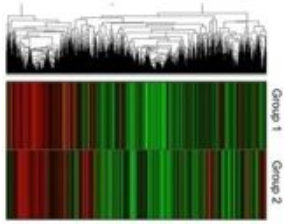
signals (start, GT/AG, stop,…)



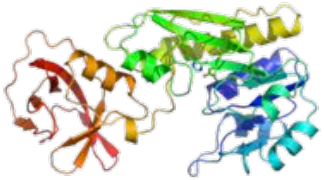
Integrative biology



Genetics/Genomics

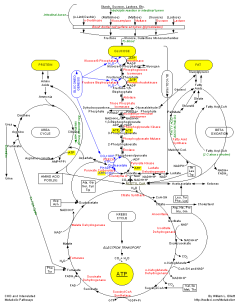


Gene Expression

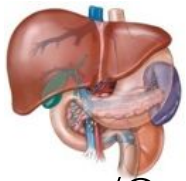


Proteomics

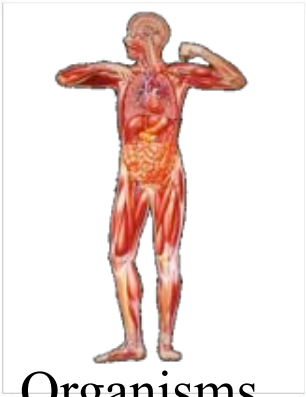
Cancer Biology



Metabolomics



Tissues/Organs



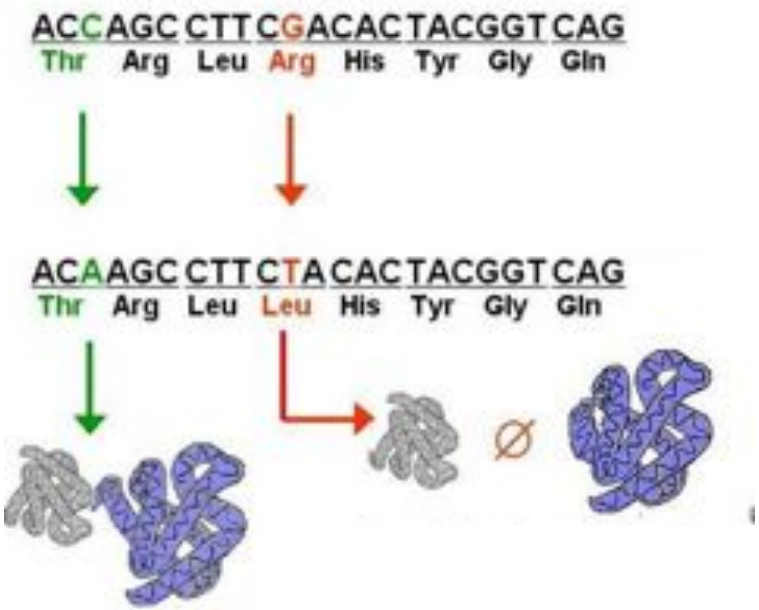
Organisms

Environment

Why study genomics ?

- From a public health perspective, genomics is the study of the gene-environment-host interaction that leads to disease — or disease prevention — in populations.

- Rare diseases
- Single gene disorders
- Public health activities
  - Newborn screening
  - Reproductive health
  - Genetic services



Contemporary Public Health Genetics

- Common diseases
- Multiple genes
- Gene/environment interactions
- Public health activities/implications
  - Chronic diseases
  - Infectious diseases
  - Environmental health
  - Epidemiology

What's new in genomics ?

- Genetic testing
  - To detect mutations
    - For disease diagnosis and prognosis
    - For the prediction of disease risk in individuals or families
- Several hundred genetic tests are in use.
  - Rare genetic disorders (muscular dystrophies, cystic fibrosis)
  - Complex conditions (breast, ovarian, and colon cancers)
- Pharmacogenomics
  - The development of drugs tailored to specific subpopulations based on genes
  - Pharmacogenomics has the potential to:
    - Decrease side effects of drugs
    - Increase drug effectiveness
    - Make drug development faster and less costly

History

1953 The double helical structure of DNA

1977 Frederick Sanger (MRC) develops methods for sequencing DNA

1980 First method to map the entire human genome based on RFLPs

1984 Complete DNA sequence of the Epstein-Barr virus, 170 kb

1985 Kary Mullis develops PCR , a technique to replicate vast amounts of DNA

1986 The first automated DNA sequencing machine

1987 Human Genome Initiative begins

1992 First physical map of chromosome 21 (D. Cohen)

First genetic maps of human (J. Weissenbach)

1995 The first sequence of a complete genome, *Haemophilus influenzae* , 1.8 Mb (C. Venter)

1996 NIH funds six groups to attempt large-scale sequencing of the human genome

1997 Genoscope is setup in France

1998 3700 capillary sequencing machines

Celera company is funded and declares that it will sequence the human genome within 3 years

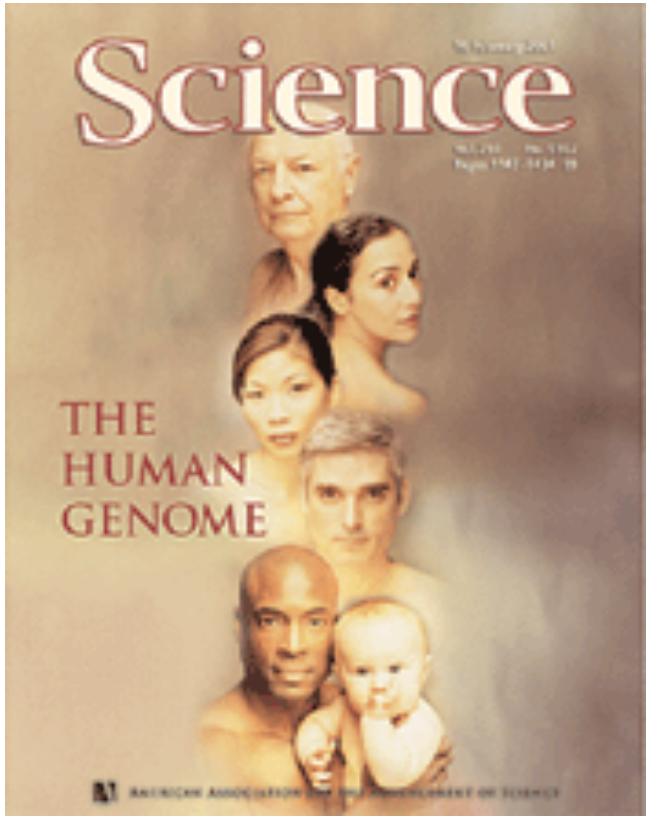
New objectives of the HGP of creating a "working draft" of the human genome by 2001

Completion date for the finished draft from 2005 to 2003.

2000 Complete genome of the first plant, *Arabidopsis thaliana* 125 Mb

2001

The HGP consortium publishes its working draft in *Nature* (15 February), and Celera publishes its draft in *Science* (16 February).



Sequencing technology evolution

1995 : The human genome was considered a 10 years world wide project

2005 Second generation of DNA sequencing machines

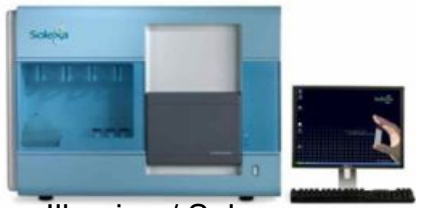
A draft of the human genome could be obtained in less than 3 months in one genome center

2011 : Third generation of DNA sequencing machines

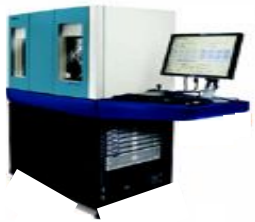
=> A human genome in 15'

Applied Biosystems
ABI 3730XL

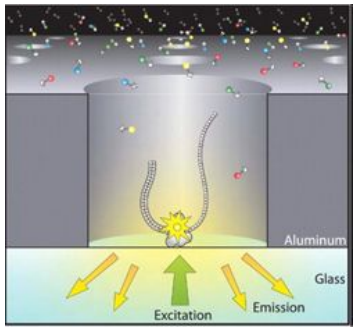
Roche / 454
Genome Sequencer FLX



Illumina / Solexa
Genetic Analyzer



Applied Biosystems
SOLiD



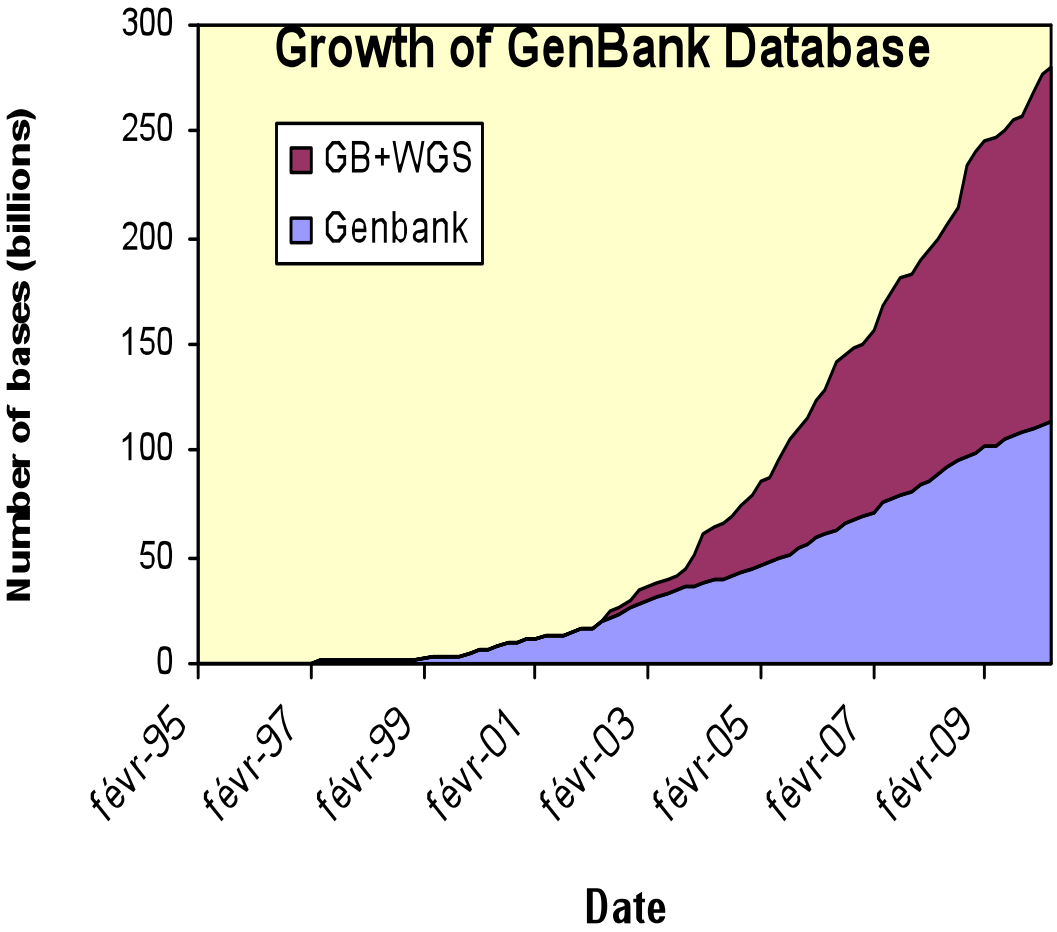
Pacific biosciences

1000 euros / Mbase **100 euros / Mbase** **10 euros / Mbase** → **<1 euros / Mbase**

2006	2007	2008	2009
Gs20 **20Mb** / run 100bp / lecture	GsFLX **100Mb** / run 250 bp / read	Titanium GS FLX **500Mb** / run 400bp / read	Titanium GS Titanium **900Mb** / run 700bp / read

	GA I **1Gb** / run 32 bp	GA II **2 à 6 Gbp** / run 36 à 108 bp /read	GA IIx **30Gb /** run 75bp / read	HiSeq **100Gb** / run 75bp / read

		SOLID **5 Gbp** / run 35 bp/read	**100 Gbp** / run 35 bp/read	**300 Gbp** / run 35 bp/read



280 Gb (+WGS)

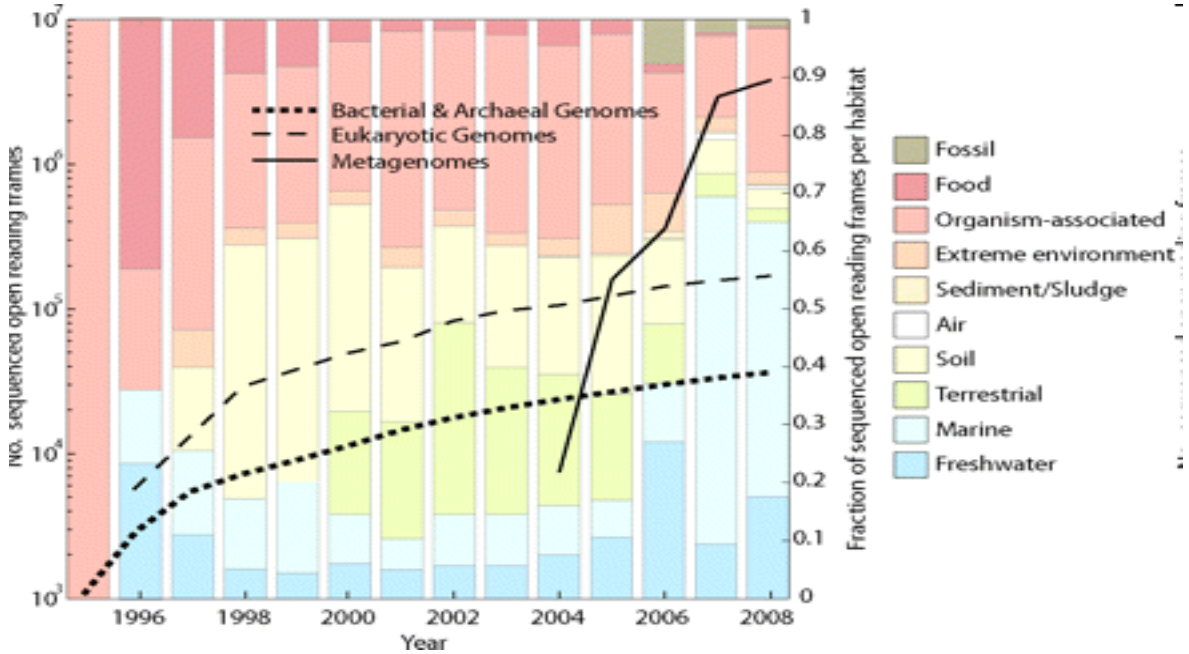
We project in 3-5 years: 100x increase in sequencing volume

Fundamental computing capabilities should increase:

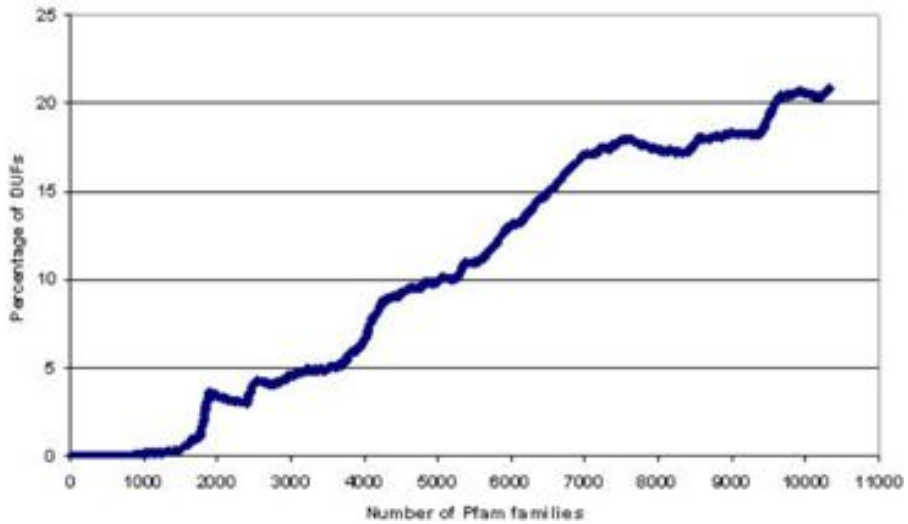
7-10x in 5 years

50-100x in 10 years

Trends in the increase of genomics data



Domains of unknown function (DUFs) inPFAM database

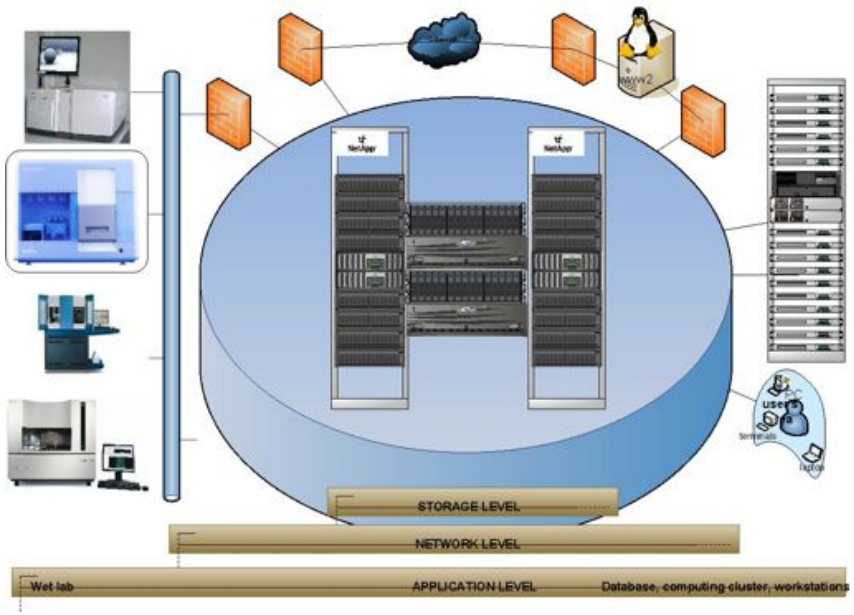


Genoscope projection : 200 à 300 Tb (2010-2011) , > 1Pb for 2011

- Computational facilities :
  - Genoscope :

Main compute cluster

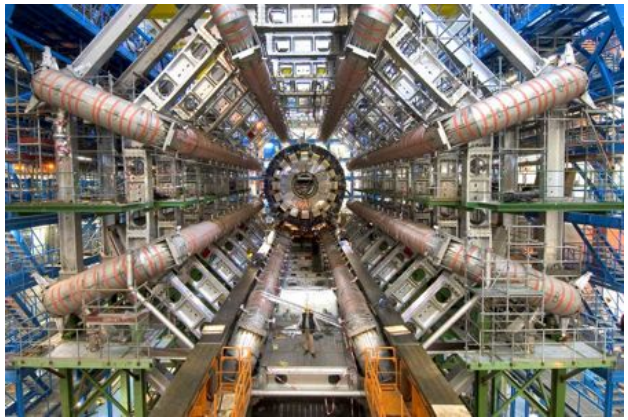
- ~40 Sun type Sunfire x4100, 330 cores,
- > 2TB memory.
- Storage : ~50 TB, increasing.



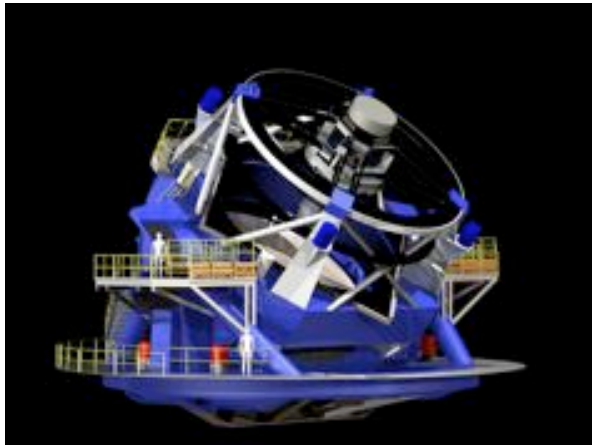
  - Genci / CCRT (CEA – Bruyères-le-Châtel)
    - Bull itanium : 47.7 Tflops, ~4000 cores
    - 23To memory
    - Storage : 420To.

CERN
Large Hadron Collider (LHC)

~10 PB/year at start
~1000 PB in ~10 years



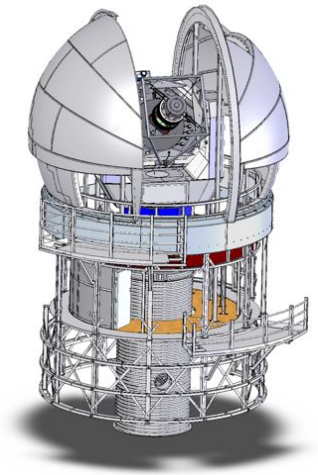
http://www.cern.ch



Large Synoptic Survey Telescope (LSST)
NSF, DOE, and private donors

~5-10 PB/year at start in 2012
~100 PB by 2025

Pan-STARRS (Haleakala, Hawaii)
US Air Force
now: 800 TB/year
soon: 4 PB/year



Data filtering at the source

CERN LHC Atlas detector generates 10^5 more data than is stored

Biology stores * 1000 than detected (Genoscope)

A need to filter the data at every stop along the way using strategy appropriate to a particular experiment/analysis

Current discussion on data normalization

Thank you