



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



The Barcelona Supercomputing Center

Romina Royo Garrido
Life Sciences Department

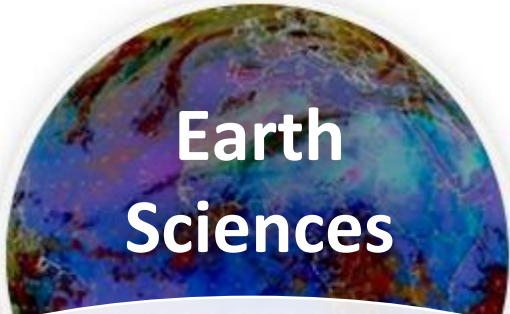
May 7th, Salamanca

Mission of BSC Scientific Departments



Computer Sciences

To influence the way machines are built, programmed and used: programming models, performance tools, Big Data, computer architecture, energy efficiency



Earth Sciences

To develop and implement global and regional state-of-the-art models for short-term air quality forecast and long-term climate applications



Life Sciences

To understand living organisms by means of theoretical and computational methods (molecular modeling, genomics, proteomics)



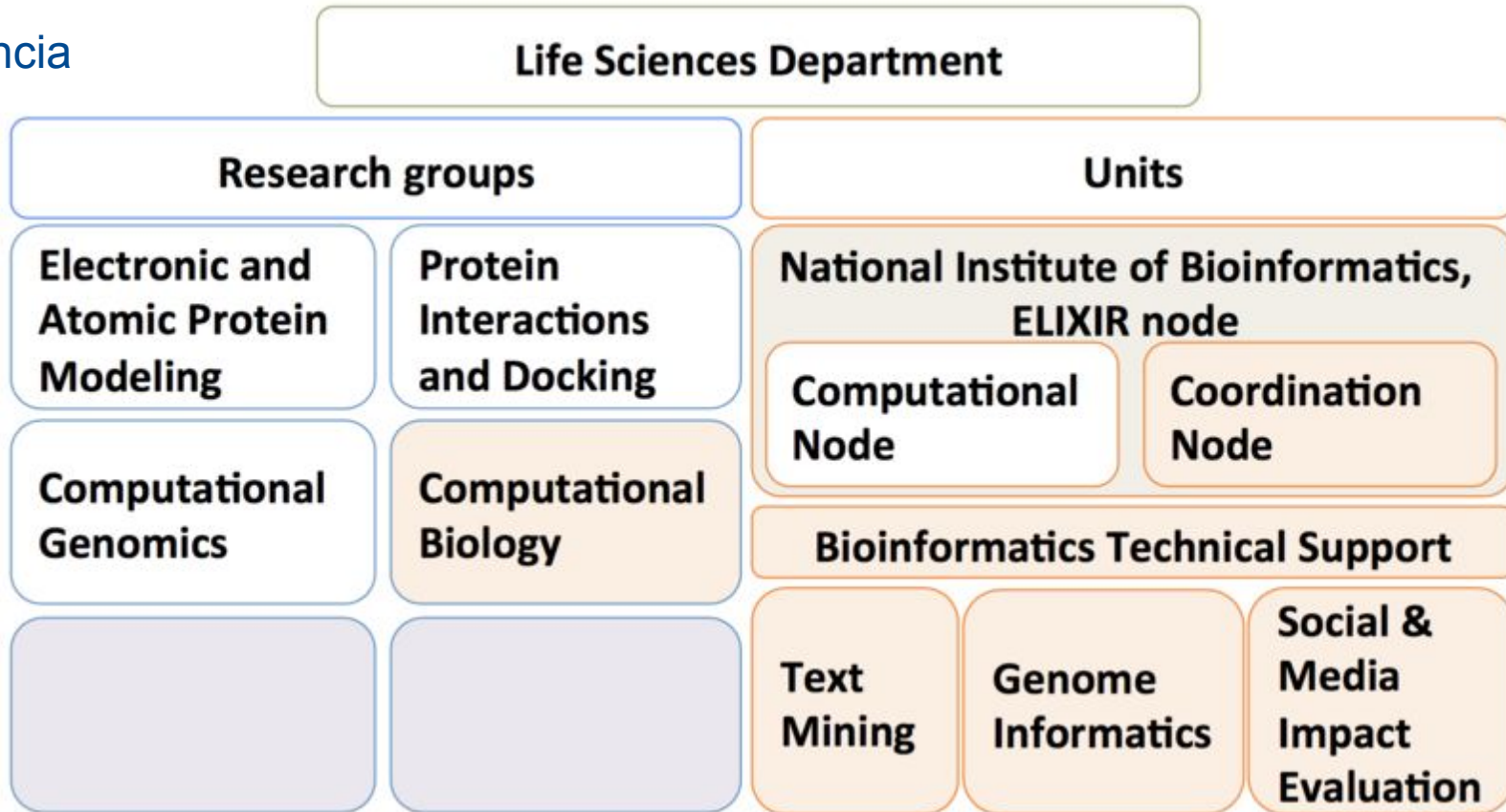
CASE

To develop scientific and engineering software to efficiently exploit super-computing capabilities (biomedical, geophysics, atmospheric, energy, social and economic simulations)

Life Sciences Department



DIRECTOR:
Alfonso Valencia



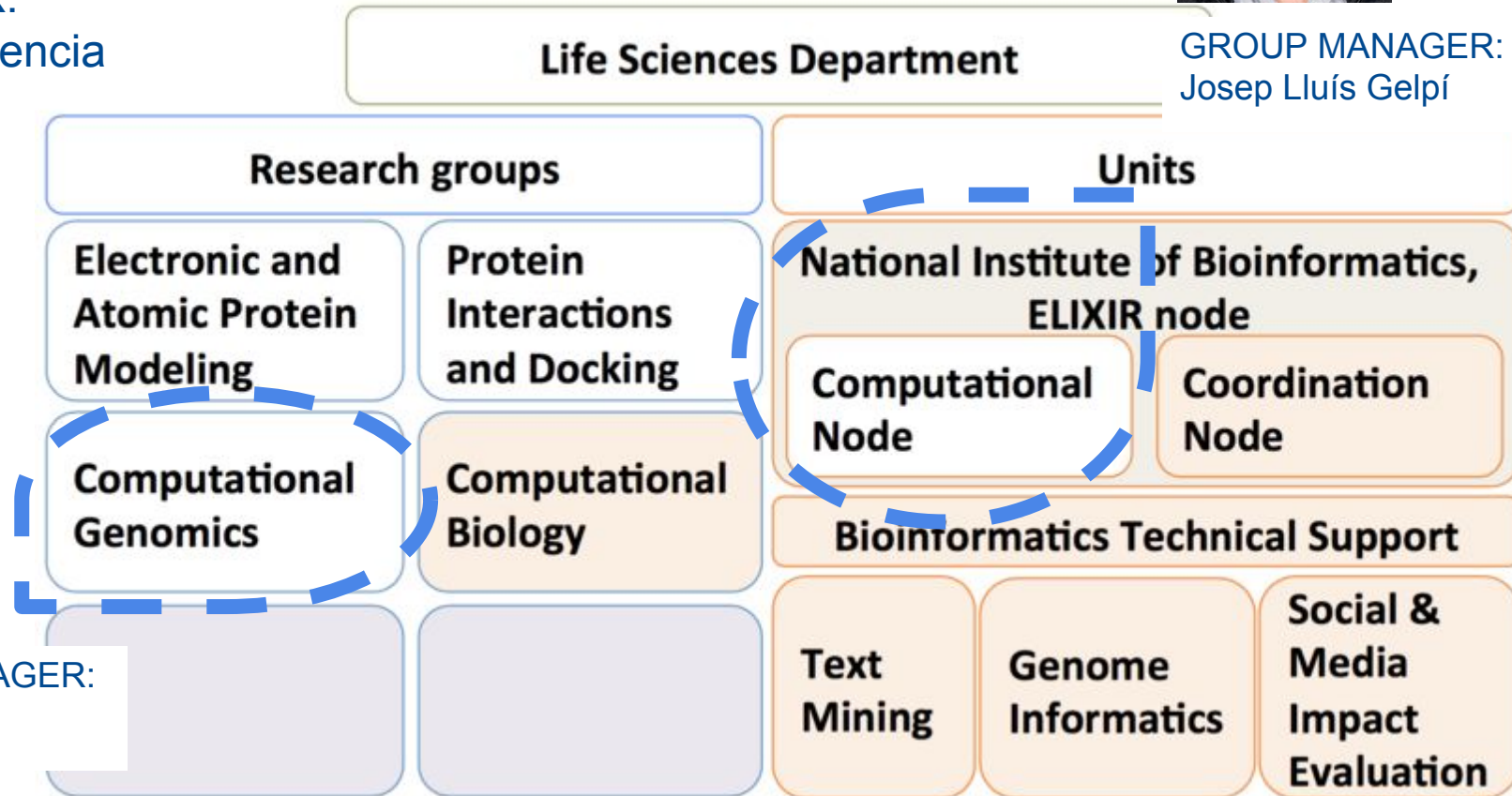
Life Sciences Department



DIRECTOR:
Alfonso Valencia



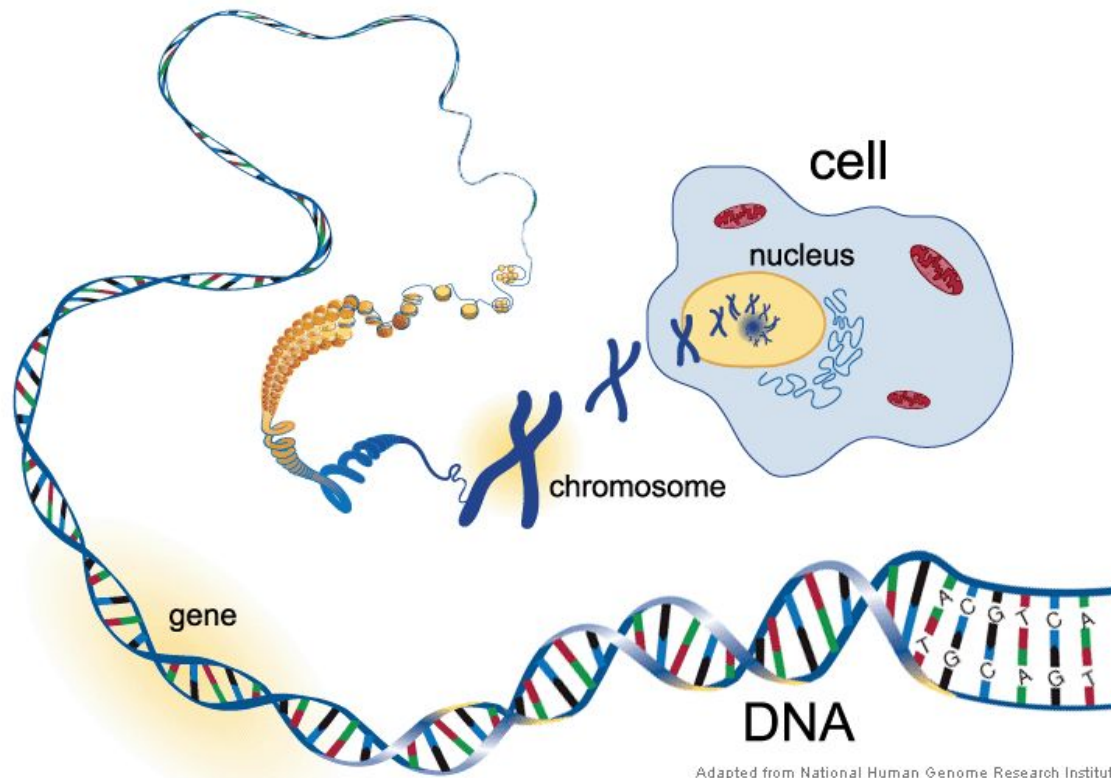
GROUP MANAGER:
Josep Lluís Gelpí



GROUP MANAGER:
David Torrents

Computational Genomics from a biological point of view

Finding the relationship between the biology of the genome and disease in humans.

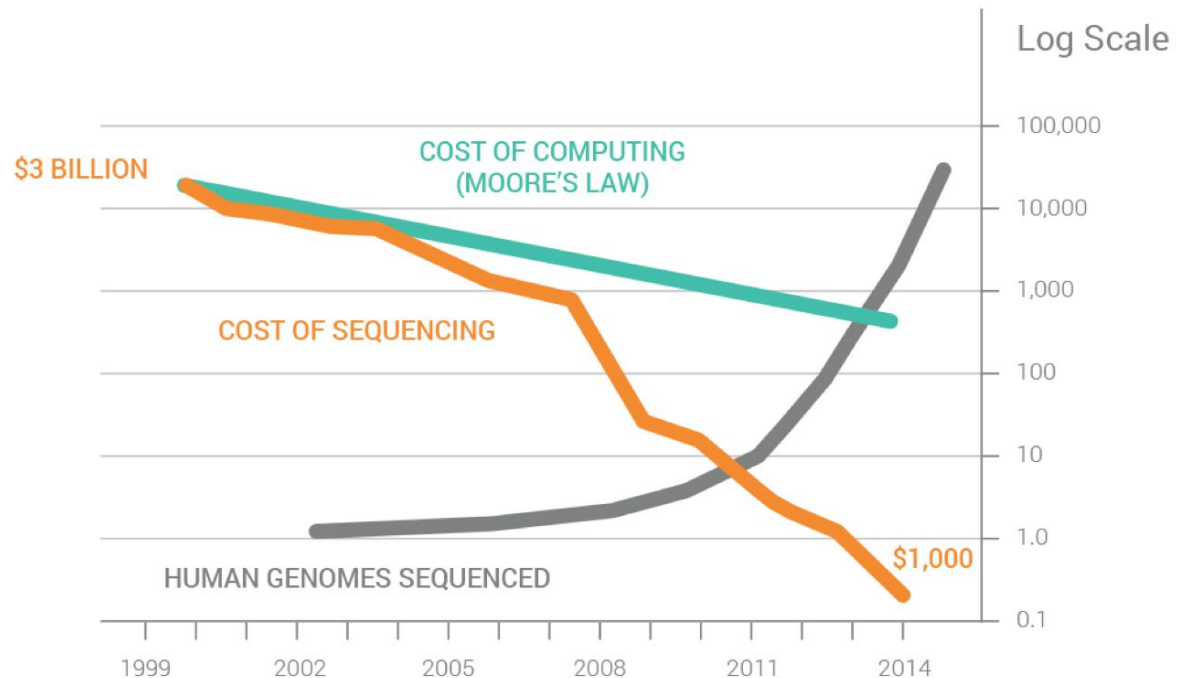


Adapted from National Human Genome Research Institute

The Sequencing Explosion



*First human genome, required 15 years to sequence and cost nearly 3 billion dollars.
In 2014, 45 human genomes can be sequenced in a single day for approximately \$1000 each.*

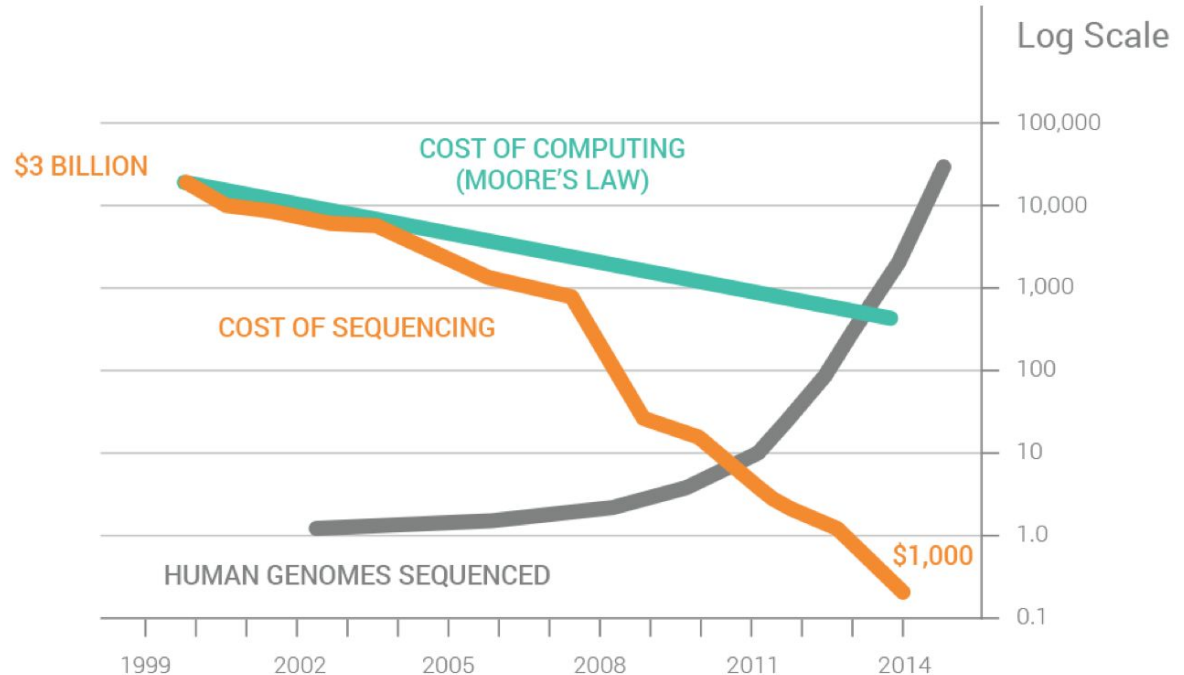
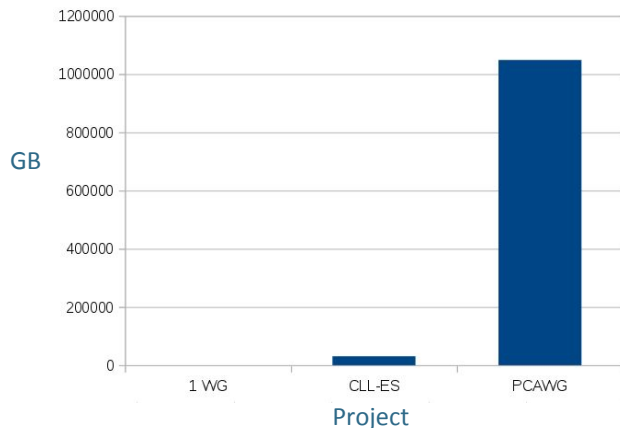


The Sequencing Explosion



**First human genome, required 15 years to sequence and cost nearly 3 billion dollars.
In 2014, 45 human genomes can be sequenced in a single day for approximately \$1000 each.**

GB of data per project



Computational Genomics from a computational point of view



Resources

- Disk storage
- Computing
- Data transfer
- Security

EGA European Genome-phenome Archive



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación

What is EGA?

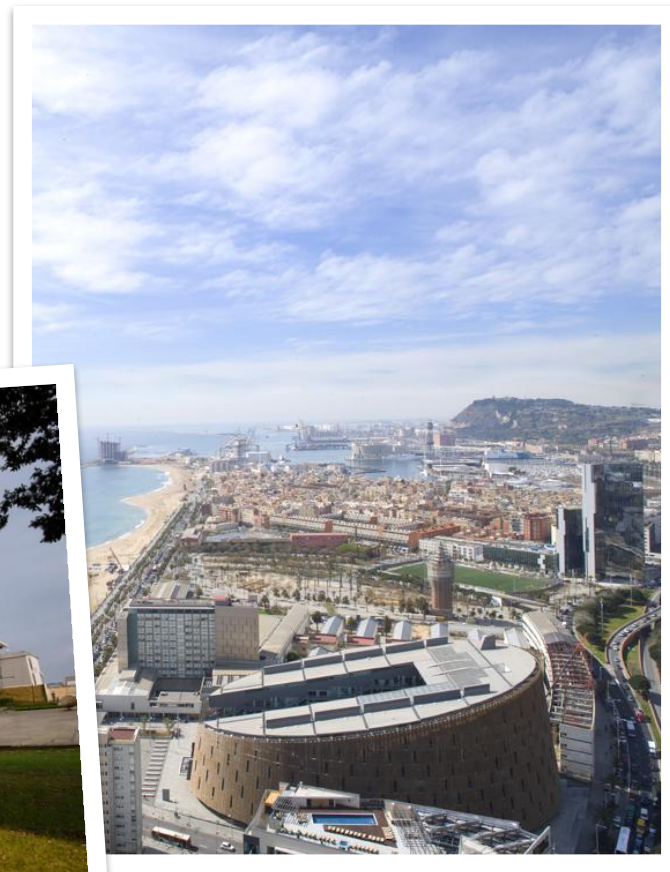


EUROPEAN
GENOME-PHENOME
ARCHIVE

10TH ANNIVERSARY



EMBL-EBI



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación

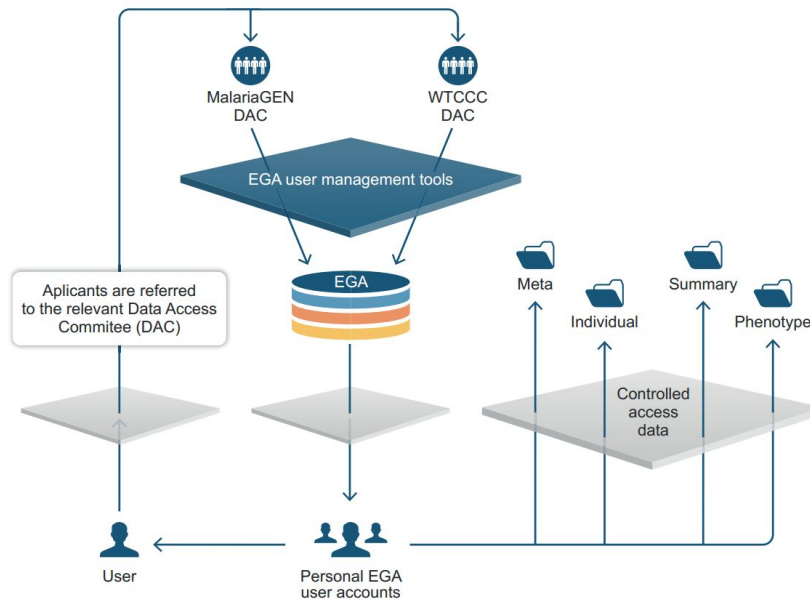


**Barcelona
Supercomputing
Center** is key infrastructure contributor
Centro Nacional de Supercomputación

What is EGA?

The EGA is a resource for permanent secure archiving and sharing of all types of potentially identifiable bio-molecular and phenotypic data resulting from biomedical research projects.

- Data is provided by research centers and health care institutions.
- Access is controlled by Data Access Committees.
- Data requesters are researchers from other research or health care institutions.



How EGA transfers data

Submission

Contact



Receive



+ encrypt
Upload



Document



SFTP



aspera

Data Access

Search



Contact DAC



Sign Data Access
Agreement (DAA)



Receive



Download
+ decrypt

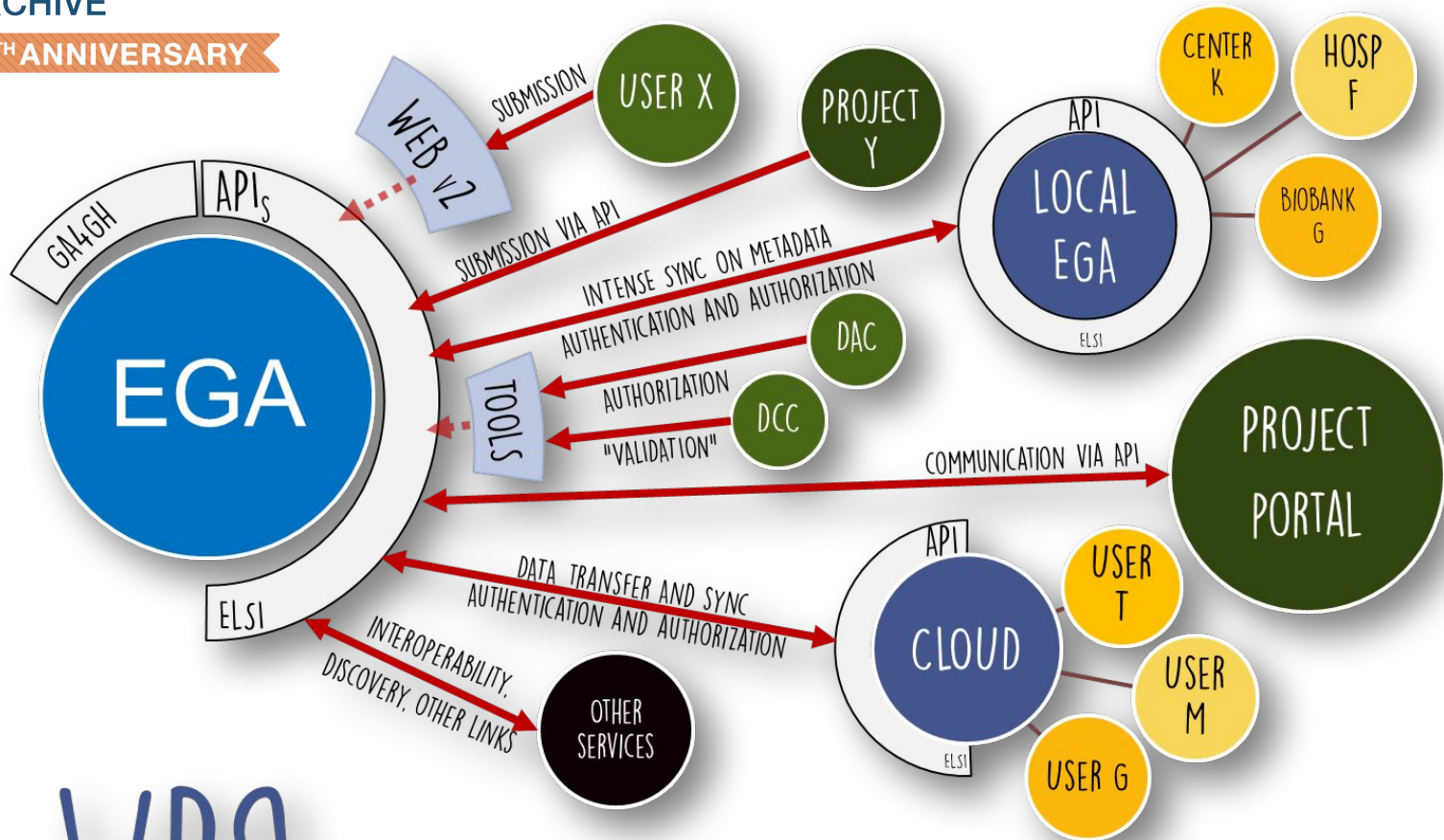


The EGA contains a variety of data



EUROPEAN
GENOME-PHENOME
ARCHIVE

10TH ANNIVERSARY



WP9

The EGA contains a variety of data



EUROPEAN
GENOME-PHENOME
ARCHIVE

10TH ANNIVERSARY

The EGA in numbers

- ~1,500 Studies
- ~3,900 Datasets
- ~800,000 Donors
- ~900 DACs
- ~ 700 Data Providers
- ~11,800 Data Requesters

The EGA in Volume

- ~4.6 Petabytes (+1PB)
- ~1.1 M files

* Updated April, 12th 2018

EUROPEAN GENOME-PHENOME ARCHIVE
10TH ANNIVERSARY

Search...

ABOUT SUBMISSION BROWSE ACCESS DOWNLOAD METADATA

Helpdesk Log in

The European Genome-phenome Archive (EGA) is a service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects.

DISEASE TYPE TECHNOLOGY SAMPLE TYPES

What is in the EGA?

Studies in the EGA by disease

Click on a column to view category subgroups

Disease Type	Number of Studies
Cancer	882
Cardiovascular	181
Infectious	84
Inflammatory	55
Neurological	52
Other	818

If applicable, a study may be included in more than one category

Latest studies

Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes - 2018-01-24

The reanalysis of existing GWAS data represents a powerful and cost-effective opportunity to gain insights into the genetics of complex diseases.

Read more →

Study 1 / 4

View Study

Published in:

Sanger

EMBL

BROWSE

- Studies
- Datasets
- DACs
- Beacon

HELP

- FTP & Aspera
- Tools
- EGA Blog
- Contact us
- Twitter

I want to access data

I have data to submit

elixir Core Data Infrastructure

The European Genome-phenome Archive (EGA) is part of the ELIXIR infrastructure. EGA is an Elixir Core Data Resource. Learn more ...

© COPYRIGHT 2018, EGA CONSORTIUM

ABOUT THE EGA - ABOUT THE CRG - ABOUT EMBL-EBI - CONTACT US - LEGAL NOTICE - SUPPORT

CRG Centre for Research in Genomics and Regeneration

EMBL-EBI

ICGC PanCancer



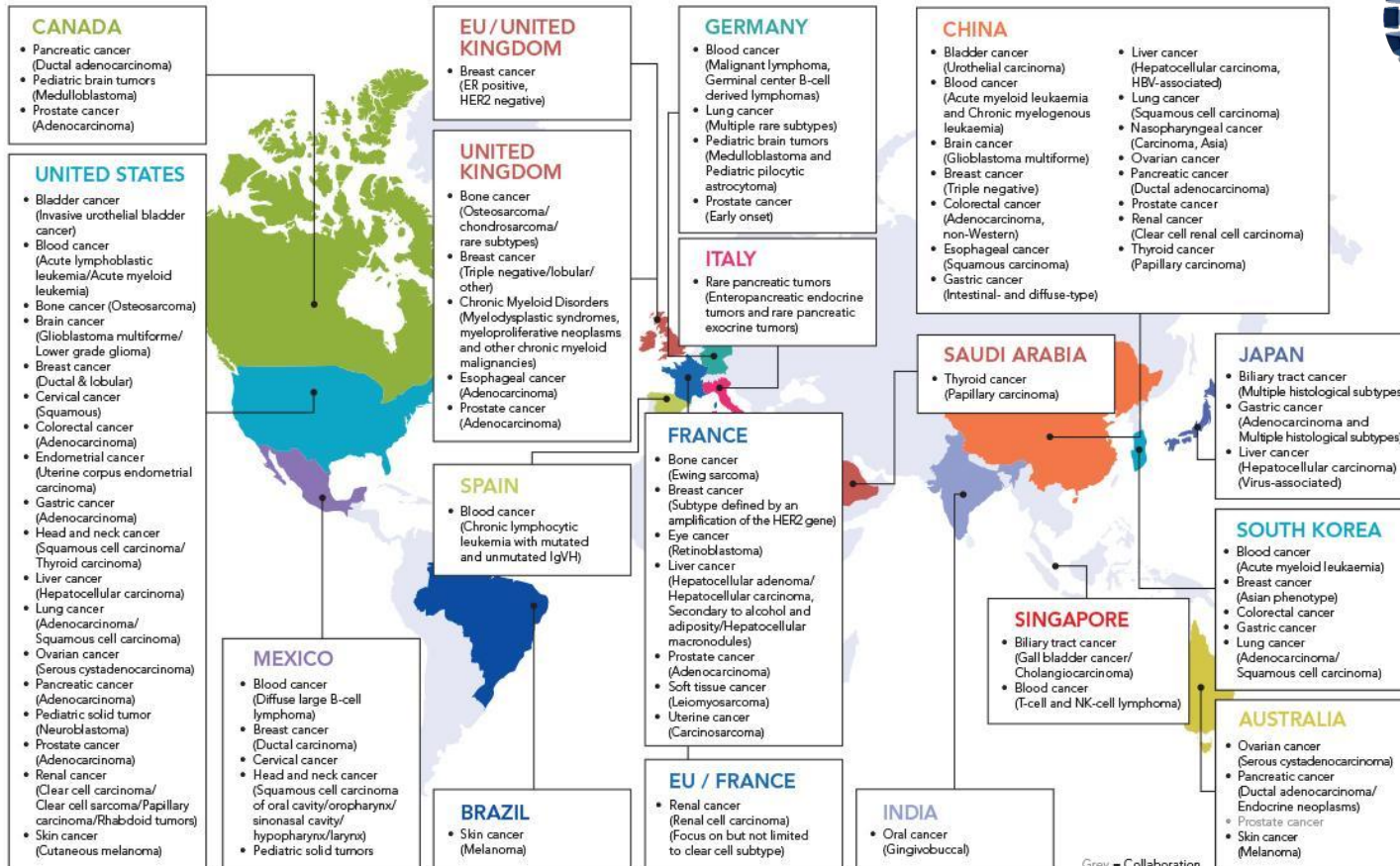
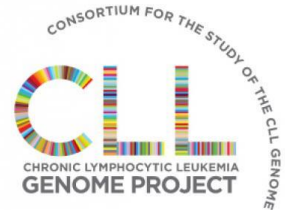
**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

International Cancer Genome Consortium



International
Cancer Genome
Consortium



Each country analyzes 500 patients of specific tumors

ICGC PanCancer Project

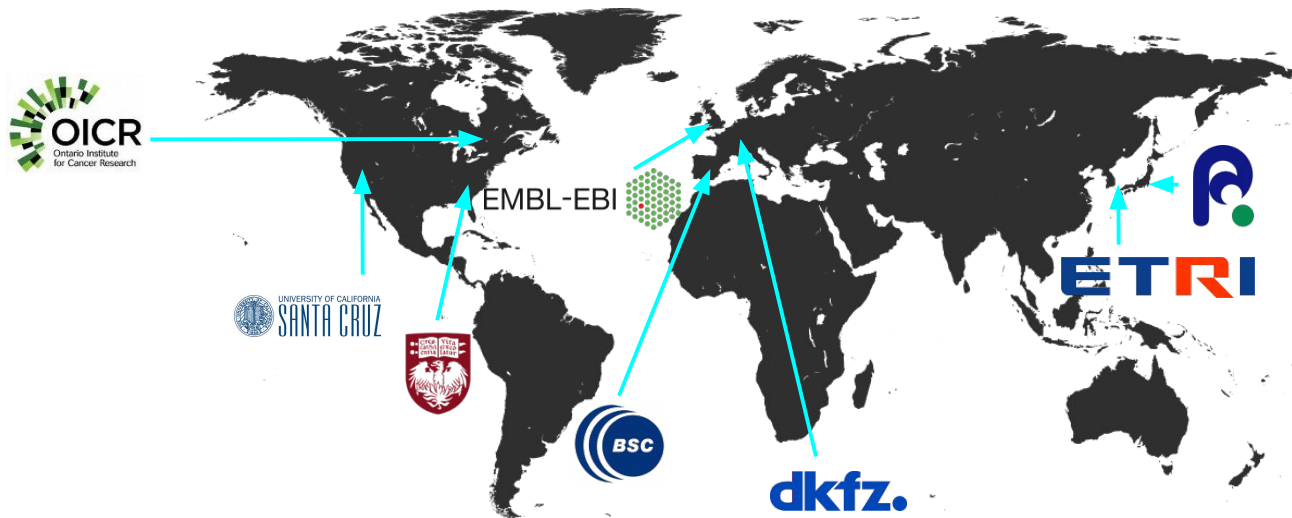


International
Cancer Genome
Consortium



PCAWG
PanCancer Analysis
OF WHOLE GENOMES

Integrated and homogeneous analysis of more than 2800 tumor-normal genome pairs of different cancer types in order to identify the genomic alterations that can lead to tumor formation so that we can explain the origin and progression of cancer.

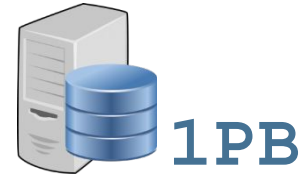


**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación

ICGC PanCancer Phases

- 1- Uploading phase
- 2- Analysis phase
- 3- Synchronizing phase
- 4- Downstream analysis

Annai-GNOS™



Data collection and organization

Data transfer

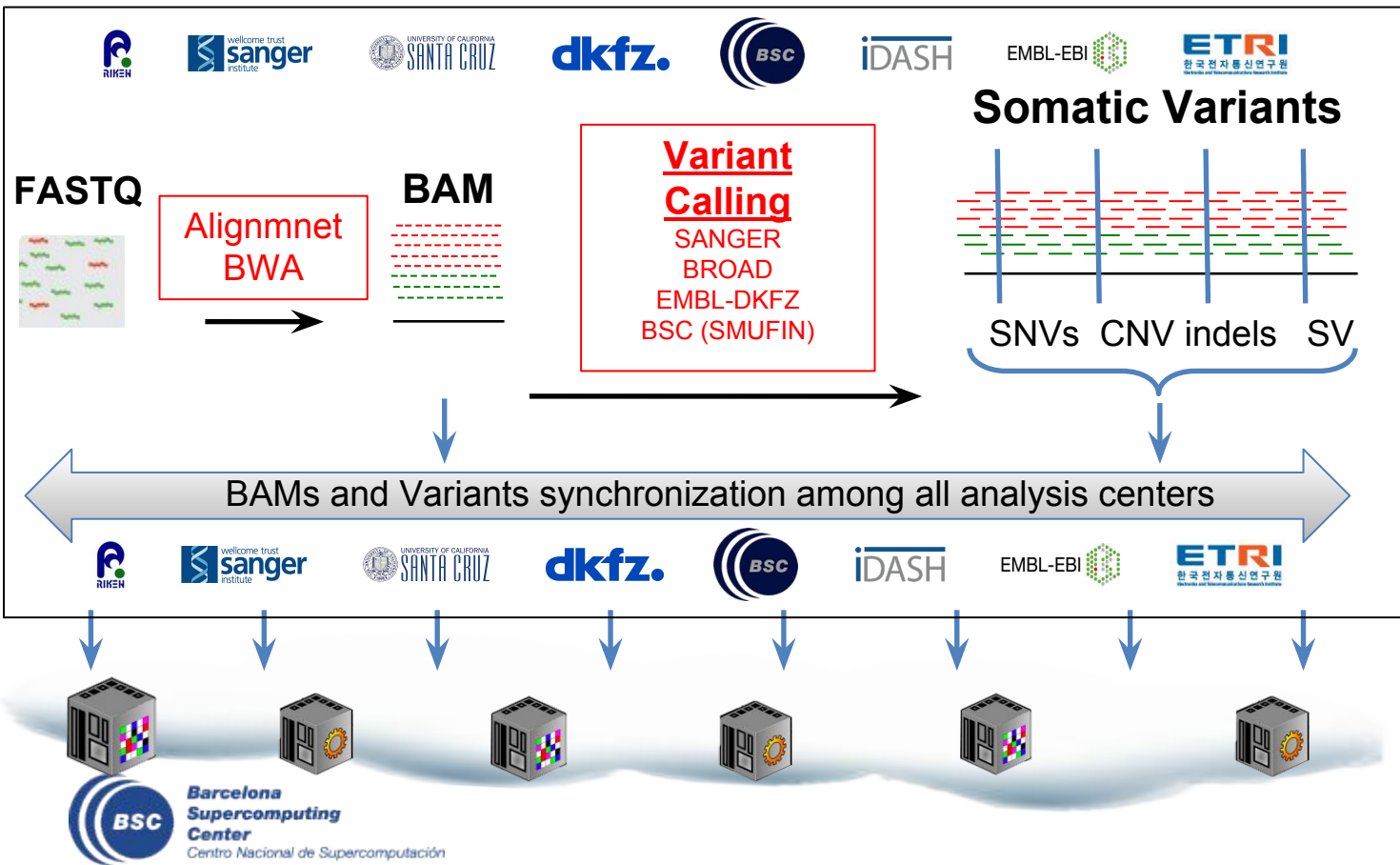
Primary analysis

Generation of BAMs

Identification of somatic variants

Data Synchronization

Data distribution to tackle specific biomedical questions



Selected Centers for the storage, analysis and distribution of PanCancer data

Data

- 2800 **Tumor-Normal WG** Pairs for more than 20 Tumor types
- ~1500 **RNAseq** samples
- ~1500 **Expression arrays**
- ~1400 **Methylation** data
- **Clinical Records**

Compute

14 **cloud and HPC environments**

- 3 commercial
- 7 openstack
- 4 HPC

~630 VMs, ~15Kcores, ~60TB of RAM

Selected Centers for the storage, analysis and distribution of PanCancer data

Data

- 2800 **Tumor-Normal WG** Pairs for more than 20 Tumor types
- ~1500 **RNAseq** samples
- ~1500 **Expression arrays**
- ~1400 **Methylation** data
- **Clinical Records**

Compute

14 **cloud and HPC environments**

- 3 commercial
 - 7 openstack
 - 4 HPC
- ~630 VMs, ~15Kcores, ~60TB of RAM

Portable tools

containerized workflows for portability between sites

Commercial cloud policies

key policy changes enabled commercial cloud usage

- NIH updated dbGaP cloud policy (March 2015)
- ICGC DACO updated ICGC cloud policy (May 2015)

Selected Centers for the storage, analysis and distribution of PanCancer data

Data

- 2800 **Tumor-Normal WG** Pairs for more than 20 Tumor types
- ~1500 **RNAseq** samples
- ~1500 **Expression arrays**
- ~1400 **Methylation** data
- **Clinical Records**

Compute

14 **cloud and HPC environments**

- 3 commercial
- 7 openstack
- 4 HPC

~630 VMs, ~15Kcores, ~60TB of RAM



Data repositories

8 sites storing and sharing data via GNOS

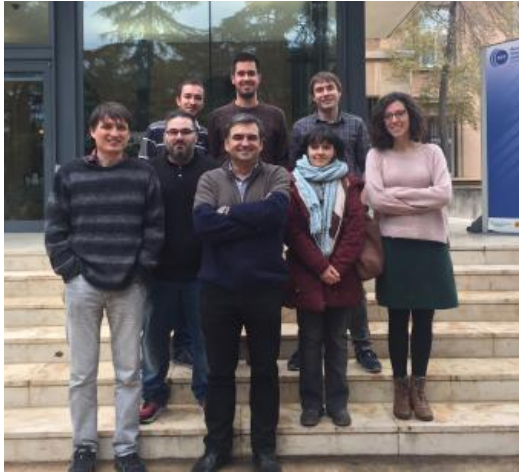
Users / Data sharing

- 580 researchers
- 130 research projects
- 16 thematic working groups

Pending challenges...

- Local EGA at each hospital? Centralized repository?
- Sharing of phenotypic and clinical data: ethical and legal issues
- Heterogeneity of methods: no standard, different results
- Scalability: there's still a lot of manual curation
- Future applications and pipelines for genomics: tightly coupled with the sequencing future
- Bioinformatics applications: Not so efficient, and moving towards virtualization

Acknowledgements



INB Computational



Computational Genomics



Operations Department



EGA Slides adapted from Jordi Rambla's



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



Thank you

romina.royo@bsc.es