ELIXIR: a pan european Research Infrastructrure for Life Science

Graziano Pesole,
ELIXIR-IIB Head of Node
CNR-IBIOM and University of Bari

IBIOM
Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari

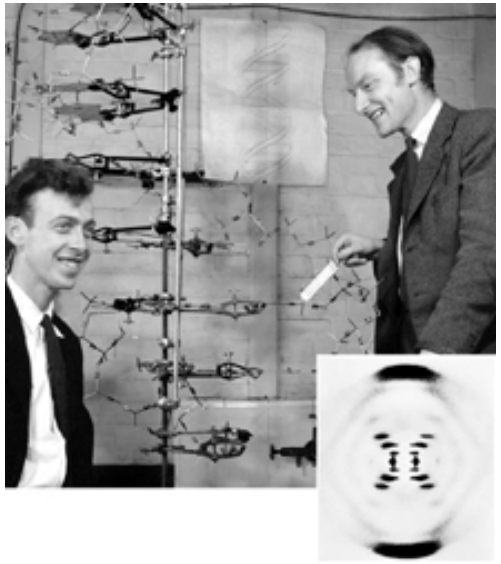
elixir ITALY

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

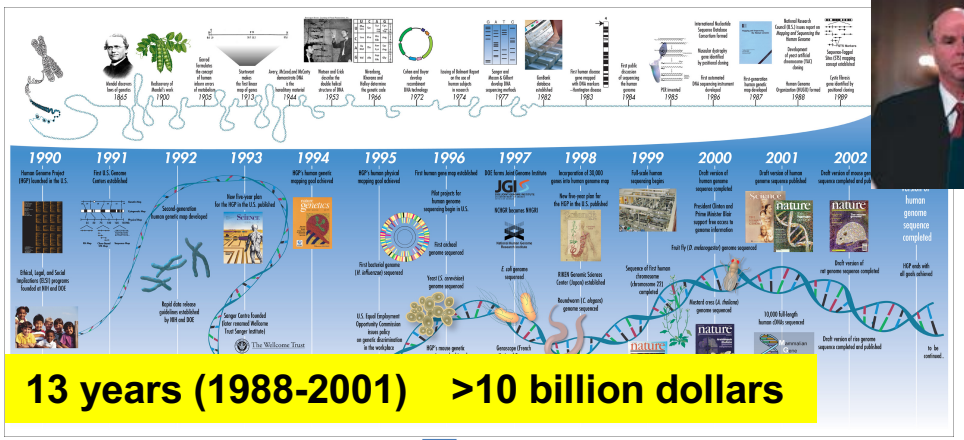
Incontro Utenti RECAS 2019
Bari, 12 July 2019

ELIXIR: DATA FOR LIFE

The development and widespread use of massive DNA sequencing systems has initiated a phase of extraordinary progress in the field of biological research, which is reaching unimaginable goals until just a few years ago.
Today the most advanced platforms allow to generate in one single experiment 6000 Gbp (6 G reads, 2x 150 bp), corresponding to 2000 human genomes in a few hours. The current cost per genome is less than 1000 euros. Programs are underway (UK, US, Japan, 1M genomes) to conduct full-scale sequencing analysis on entire populations, to unravel the cause of many genetic diseases and to develop protocols of personalized or precision medicine.



1953 – DNA structure

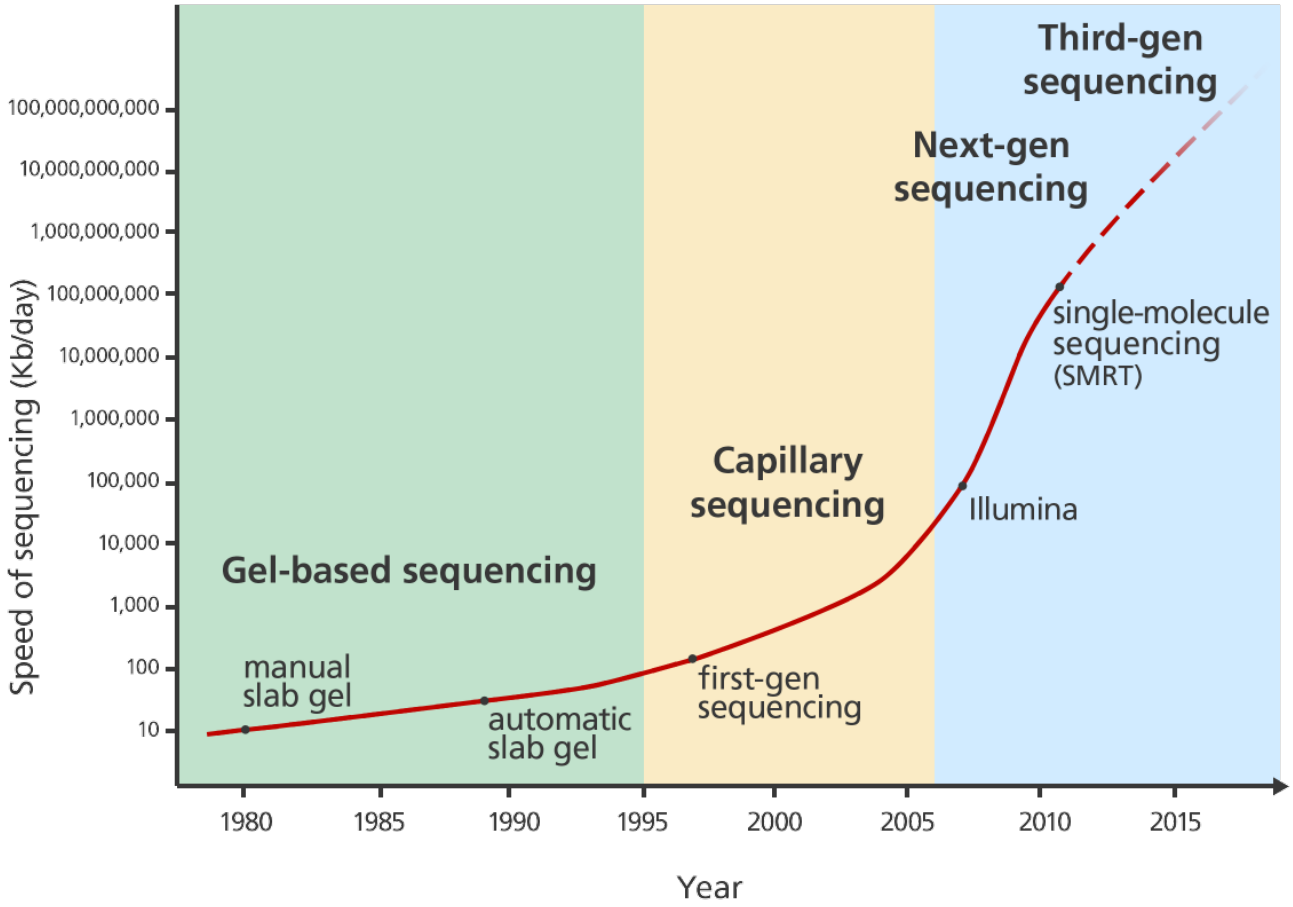


13 years (1988-2001) >10 billion dollars

Now - Few days < 1000 dollars

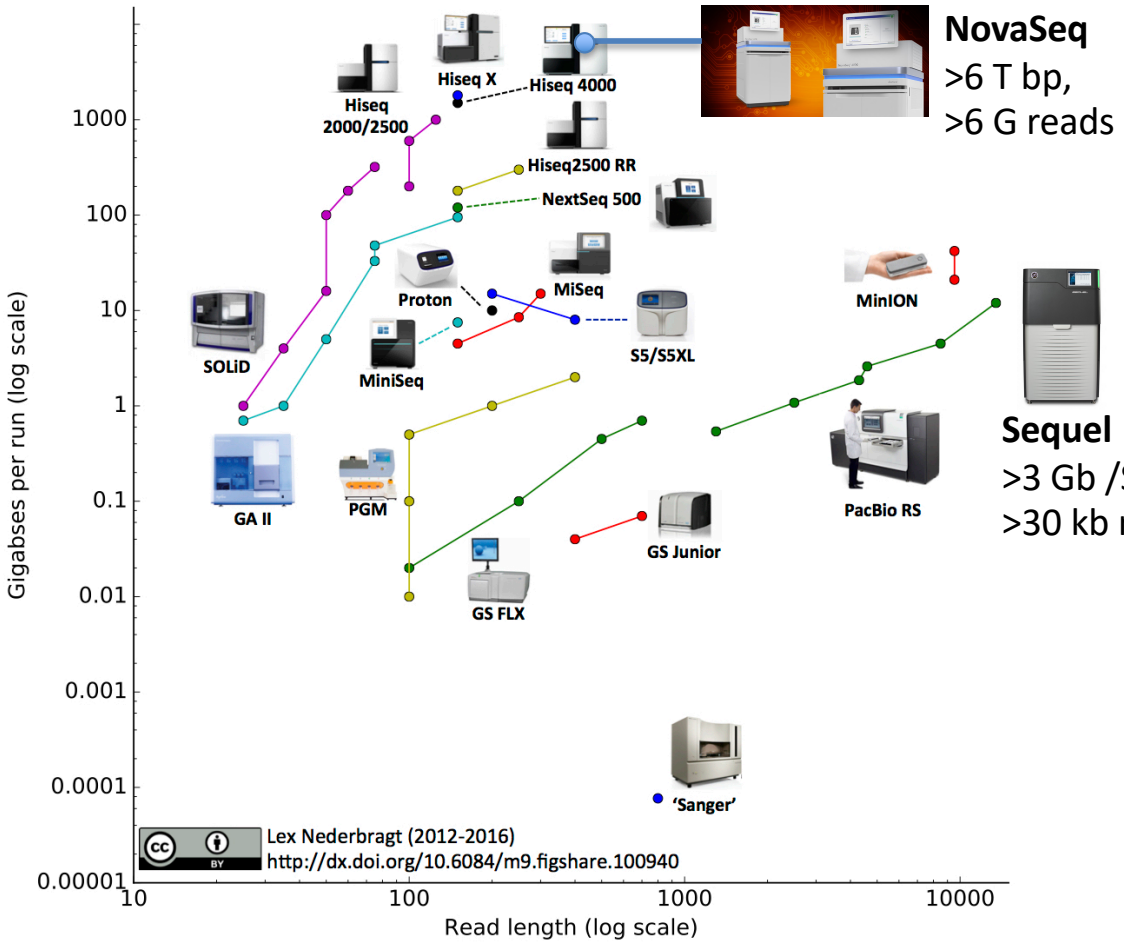
NGS revolution

The "Human Genome" project has favored an exceptional technological innovation in the field of DNA sequencing technologies that in the last decade have allowed an exponential increase in the sequencing capacity and an incredible parallel reduction in costs.



SECOND AND THIRD GENERATION MASSIVE SEQUENCING

A large number of platforms using different strategies and chemistries, and with a different throughput are progressively entering the market and third-generation systems are on the way.



NovaSeq
>6 T bp,
>6 G reads

Sequel
>3 Gb /SMRT cell
>30 kb read

Volume of Data

Cost / bp
(about 14 cent / Mb)

The data challenge: Geographic spread

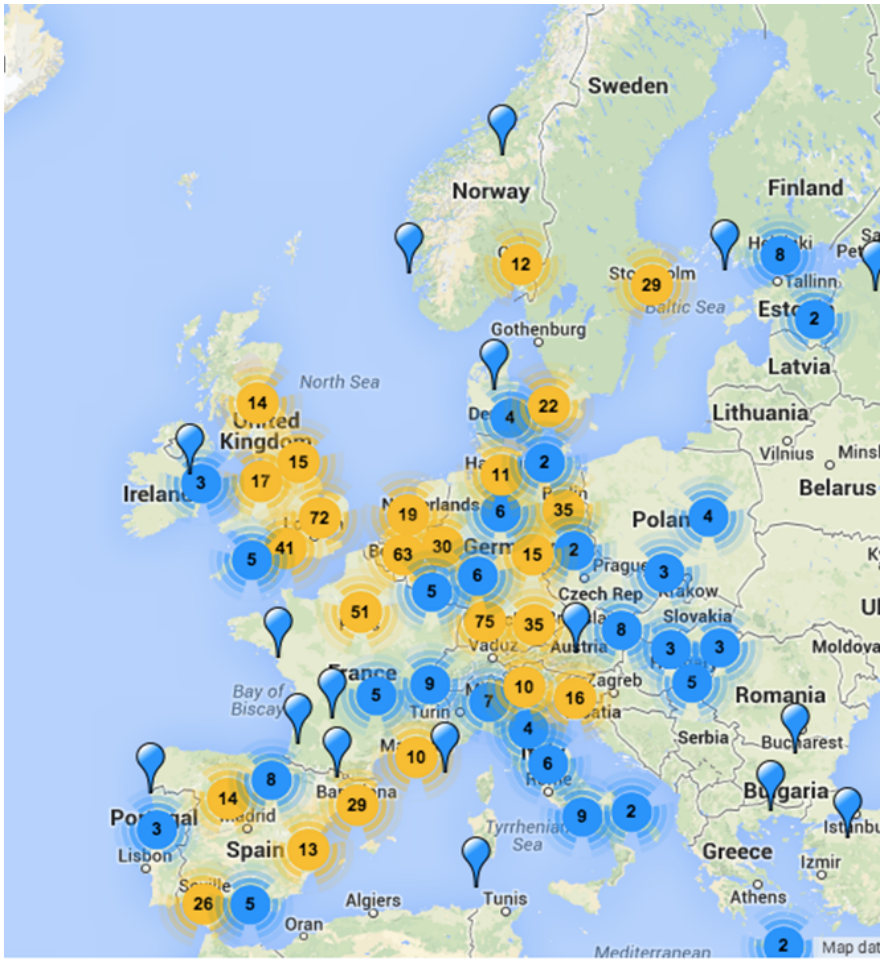
- Many data production sites across Europe

Genomics as a Big Data Science

Discipline	Duration	Size	# Devices
HEP - LHC	10 years	15 PB/year*	One
Astronomy - LSST	10 years	12 PB/year**	One
Genomics - NGS	2-4 years	0.4 TB/genome	1000's

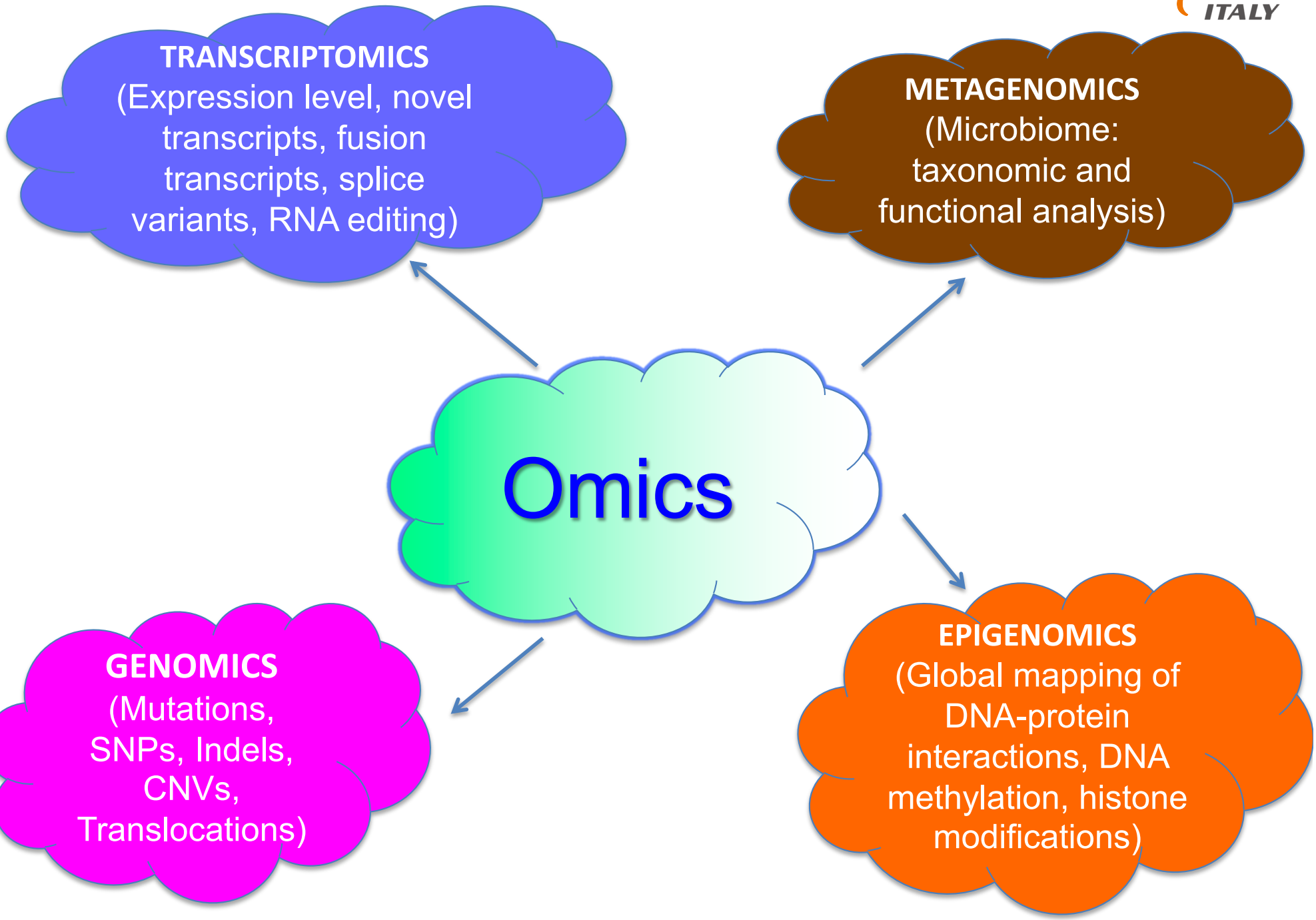
*At full capacity, the Large Hadron Collider (LHC), the world's largest particle accelerator, is expected to produce more than 15 million Gigabytes of data each year. ... This ambitious project connects and combines the IT power of more than 140 computer centres in 33 countries. Source: http://press.web.cern.ch/public/en/Spotlight/SpotlightGrid_081008-en.html

**As it carries out its 10-year survey, LSST will produce over 15 terabytes of raw astronomical data each night (30 terabytes processed), resulting in a database catalog of 22 petabytes and an image archive of 100 petabytes. Source: http://www.lsst.org/News/enews/teragrid-1004.html



Source: http://omicsmaps.com

Applications of the OMIC Technologies



BIG DATA IN BIOLOGY:
NOT ONLY GENOMICS

Nature Methods, 2019

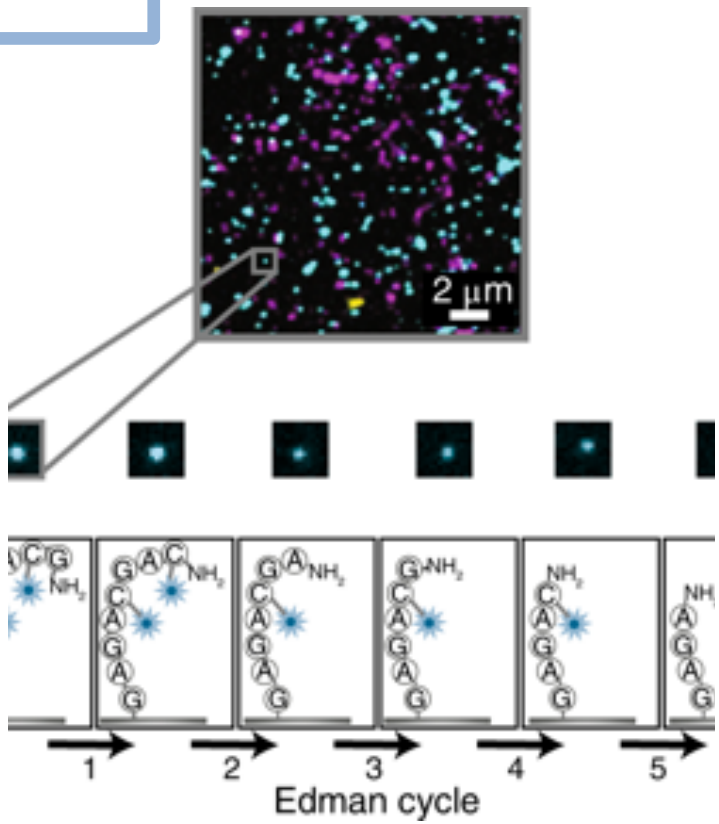
research highlights

SEQUENCING

Next-generation peptide sequencing

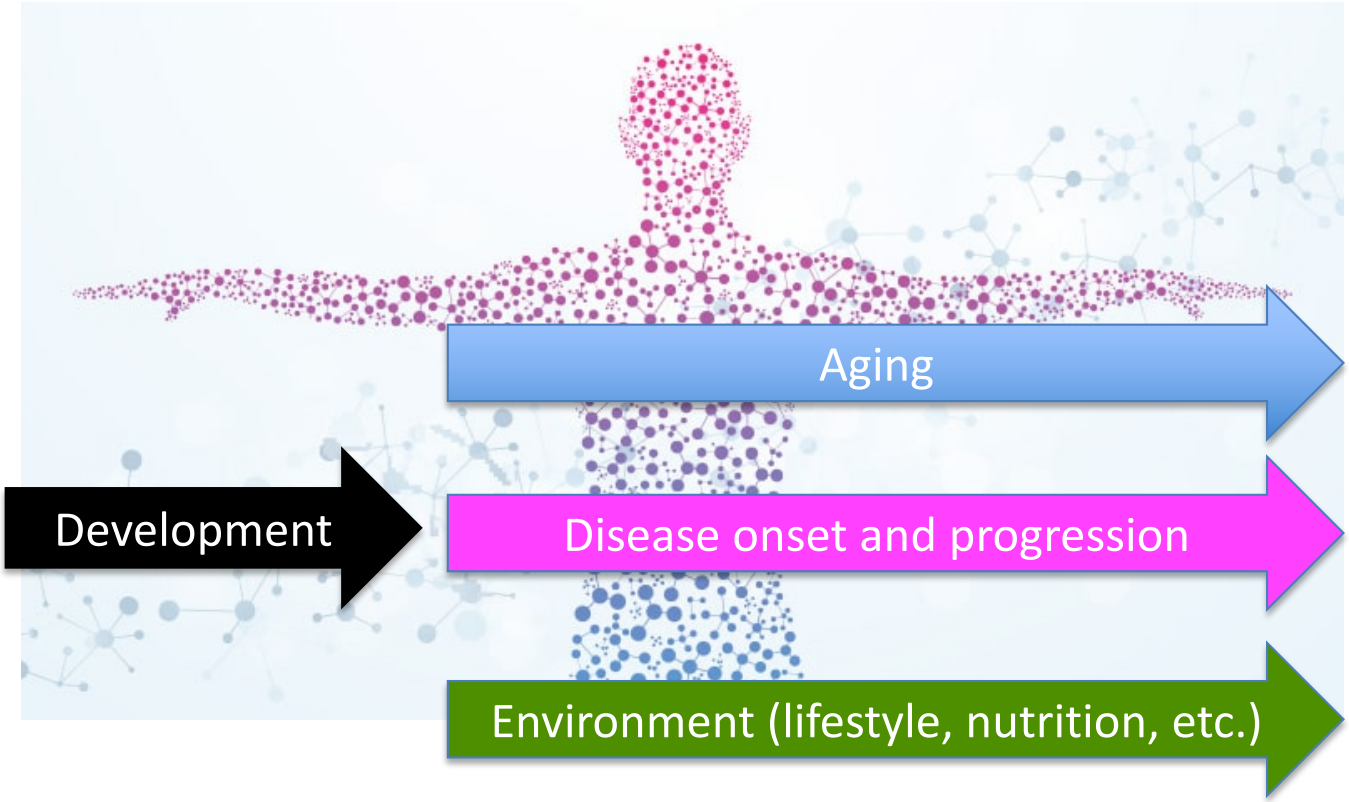
The concept of massively parallel single-molecule protein sequencing emerges.

Proteomics, Metabolomics,
and many more



BIG DATA IN BIOLOGY

A single individual is the repository of a amazing amount of data.



The current high-throughput technologies now allow large-scale omics analysis at single cell resolution

1 genome (6 Gb)
20,000 genes

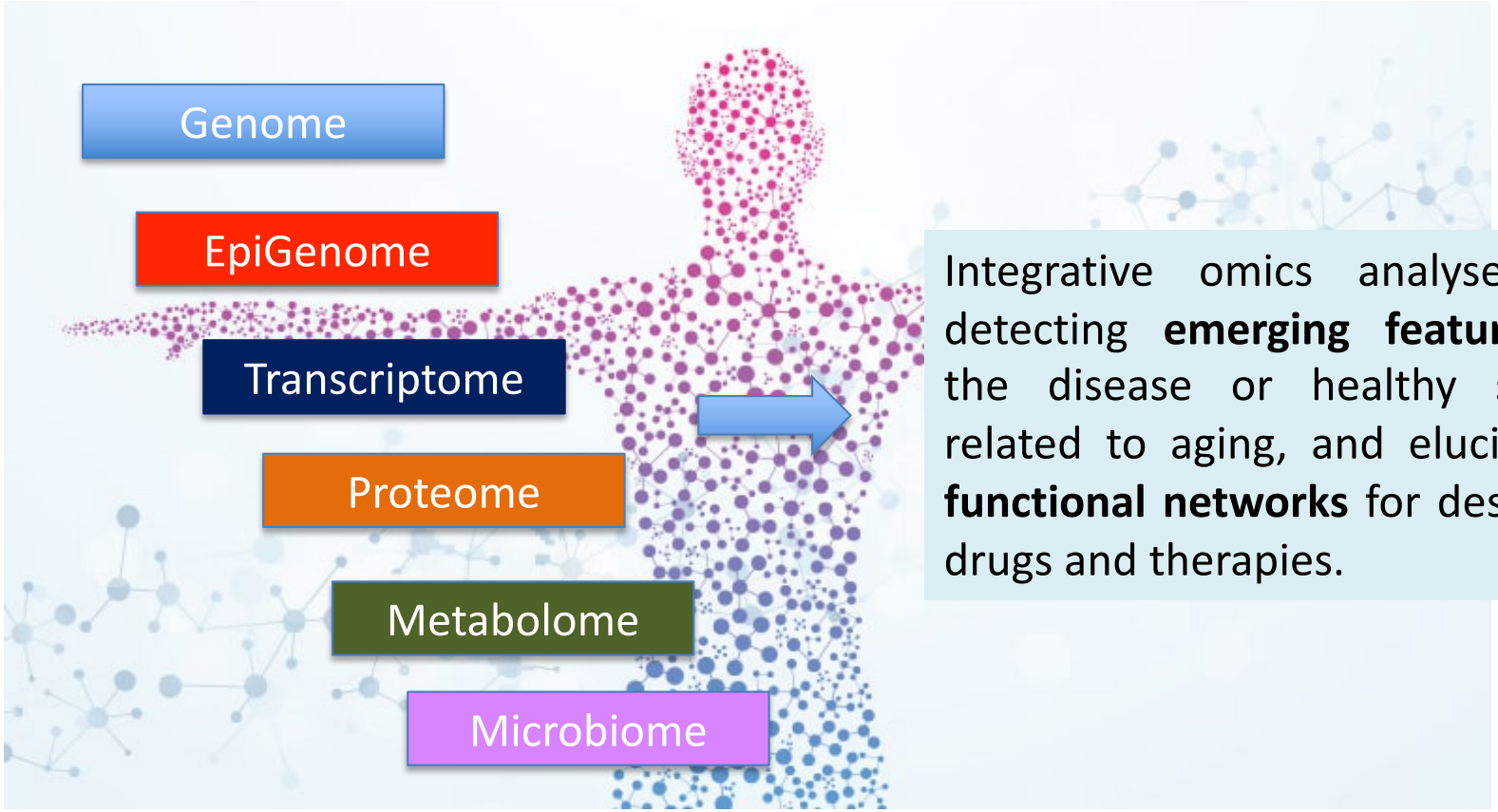
10^{13} epigenomes
10^{13} transcriptomes
10^{14} microbiomes

etc. proteome, metabolome

shutterstock.com · 428687926

Exabyte (10^{18}) scale biodata information size

BIG DATA IN BIOLOGY: DATA COLLECTION, ANALYSIS AND INTEGRATION

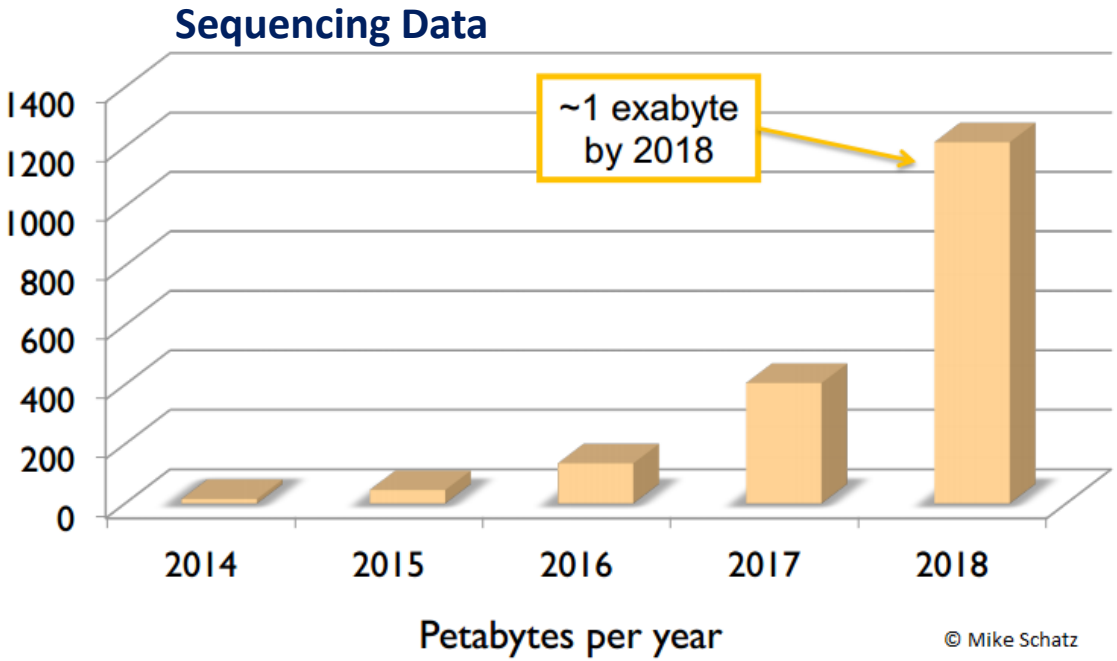


Integrative omics analyses for detecting **emerging features** of the disease or healthy status, related to aging, and elucidating **functional networks** for designing drugs and therapies.

A pan-European sustainable European infrastructure for biological information (e.g. Omics data) is thus critically needed for supporting life science research and its translation to medicine, agriculture, bioindustries and society.

NEXT GENERATION BIOINFORMATICS

The dizzying development of "omics" technologies is generating an avalanche of data that definitely sets the Life Sciences in the ranks of **BIG DATA SCIENCE**. To face the challenges of the BIG DATA it is necessary to have ICT infrastructures adequately sized in terms of storage and calculation capacity, such as standard operating procedures (POS, GLP), quality management systems, data sharing procedures and interoperability, etc.. These challenges are not sustainable by individual research groups or institutions. For this reason the "Next Generation Bioinformatics" must make use of large transnational research infrastructures.



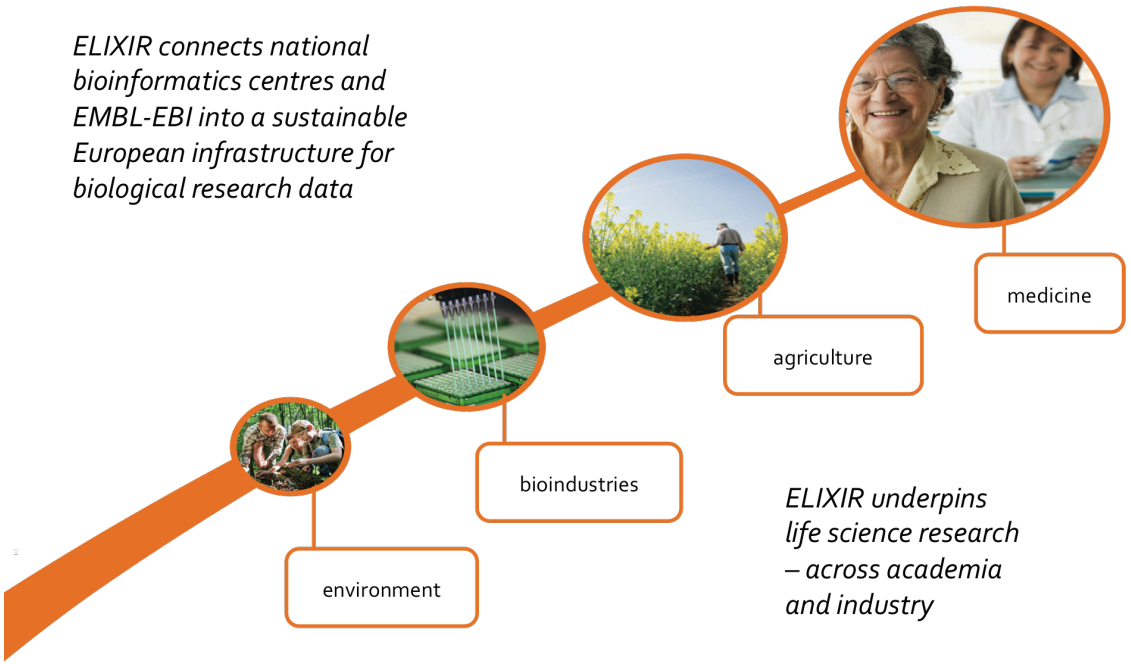
Supercomputer Marconi (CINECA)

ELIXIR: The Life Science Research Infrastructure to face the Big Data challenge in Biology

ELIXIR is an intergovernmental organisation, formally established in 2016 as a Landmark European Research Infrastructure, that brings together "bioinformatic resources" for life sciences from across Europe. These resources include **databases**, **software tools**, **training materials**, **best practices**, **cloud storage** and **supercomputers**.

The goal of ELIXIR is to coordinate these resources so that they form a single infrastructure. This infrastructure makes it easier for scientists to find and share data, exchange expertise, and agree on best practices. Ultimately, it will help them gain new insights into how living organisms work.

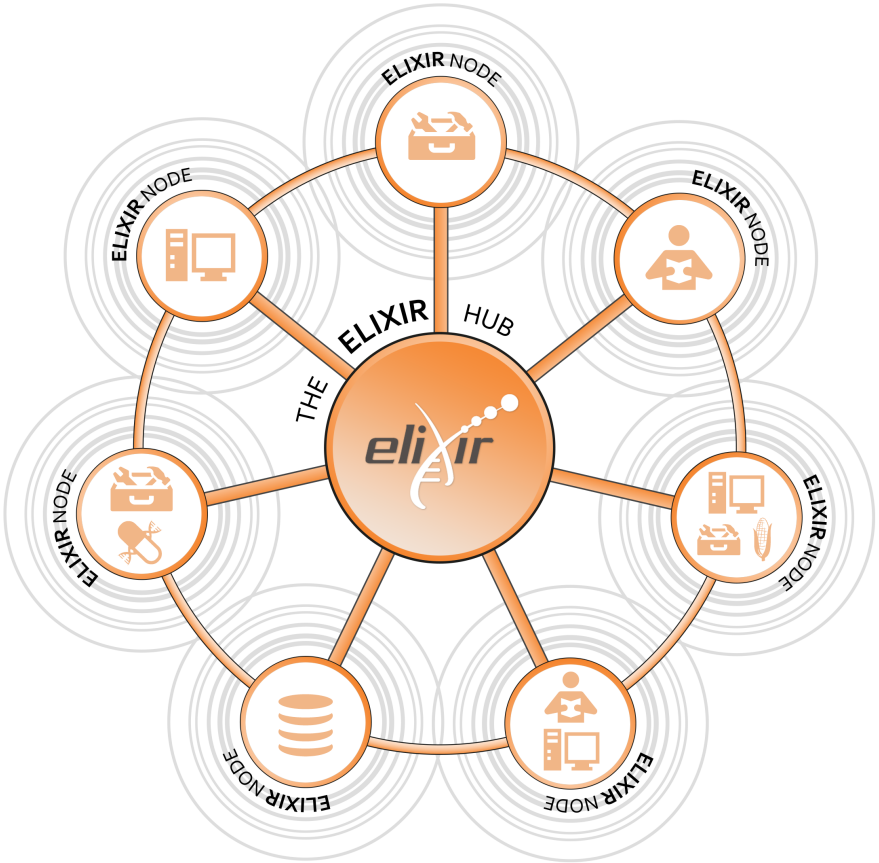
ELIXIR connects national bioinformatics centres and EMBL-EBI into a sustainable European infrastructure for biological research data

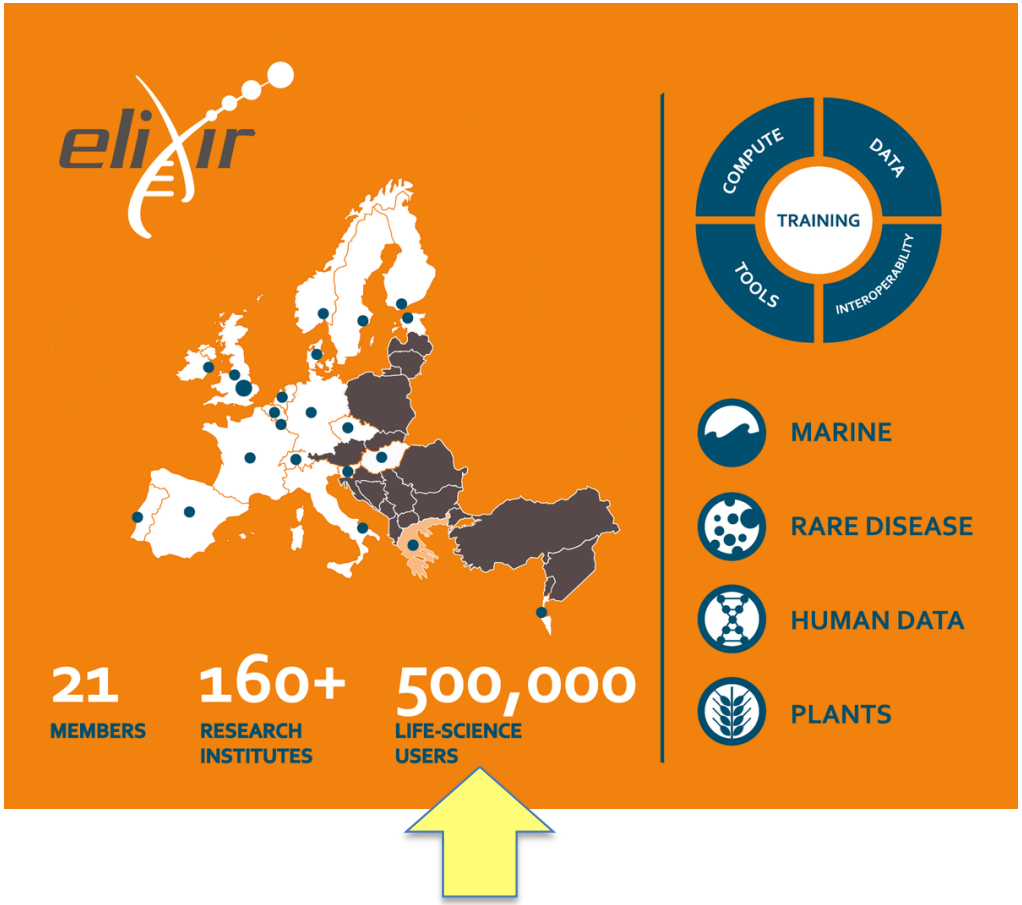


ELIXIR underpins life science research – across academia and industry

ELIXIR: A pan-european distributed Infrastructure for Bioinformatics

ELIXIR is structured as a central hub, located in the Wellcome Genome Campus (Hinxton, UK) and 23 national nodes including over 160 Research Organizations.

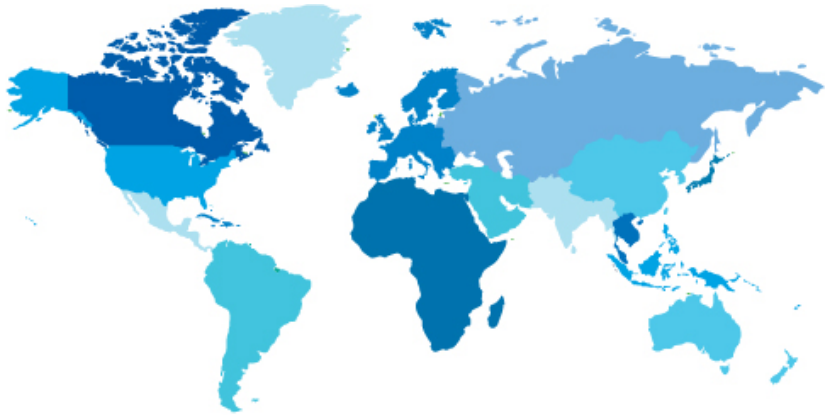




An infrastructure of global significance

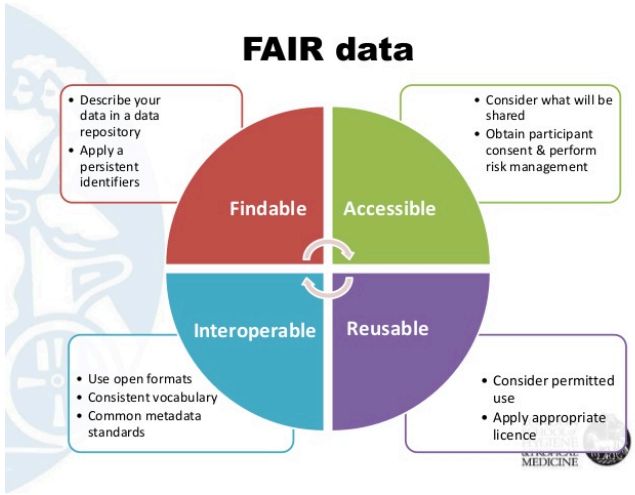
- ELIXIR put forward in G7 Group of Senior Officials report for 2015 on global research infrastructures
- 2016 ESFRI Roadmap classifies ELIXIR as a Landmark project
- Discussions initiated with Canada (Genome Canada) and Australia
- Collaboration with NIH-funded Big Data 2 Knowledge Initiative

G7 GERMANY 2015



More data challenges…

- Secure access and governance of human data
- Open data mandates of National and European funders (data FAIRification)



erc

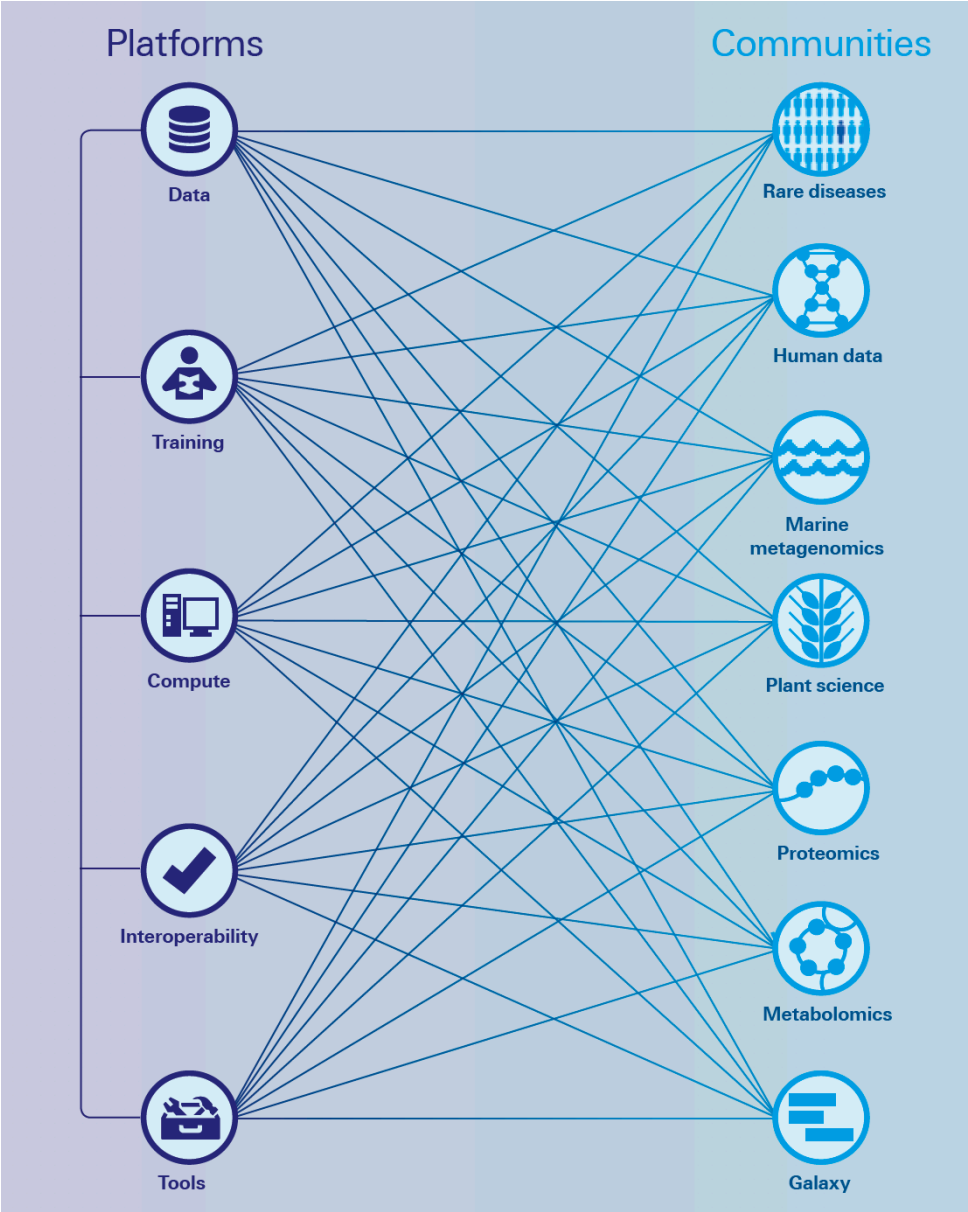
European Commission

BBSRC bioscience for the future

imi Innovative Medicines Initiative

wellcome trust

ELIXIR Organization



Five technical platforms for Compute, Data, Tools , Interoperability and Training

Complemented by seven user communities

In the 2019-23 Scientific Programme use cases evolved in "User Communities" enlarging the ELIXIR portfolio such as Proteomics, Metabolomics, Galaxy, ..

ELIXIR Services

Data deposition:
ENA, EGA, PDBe, EuropePMC, …

Data management:
Genome annotation
Data management plans

Added value data:
UniProt, Ensembl, OrphaNet, …

Data Interoperability:
BioSharing, identifiers.org and OLS

Compute:
Secure data transfer, cloud computing, AAI

Bioinformatics tools:
Bio.tools

Industry:
Innovation and SME programme
Bespoke collaborations

Training:
TeSS, Data Carpentry, eLearning

ELIXIR AAI (Authorisation and Authentication Infrastructure)

- Identification (ELIXIR ID)
- Group/role and attribute (such as researchers home organization)

- Authentication (via GEANT/eduGAIN, social media or ORCID)
- Strong step-up authentication (for sensitive services, GDPR compliant)
- Personal authorisation management (for datasets that require DAC approval)

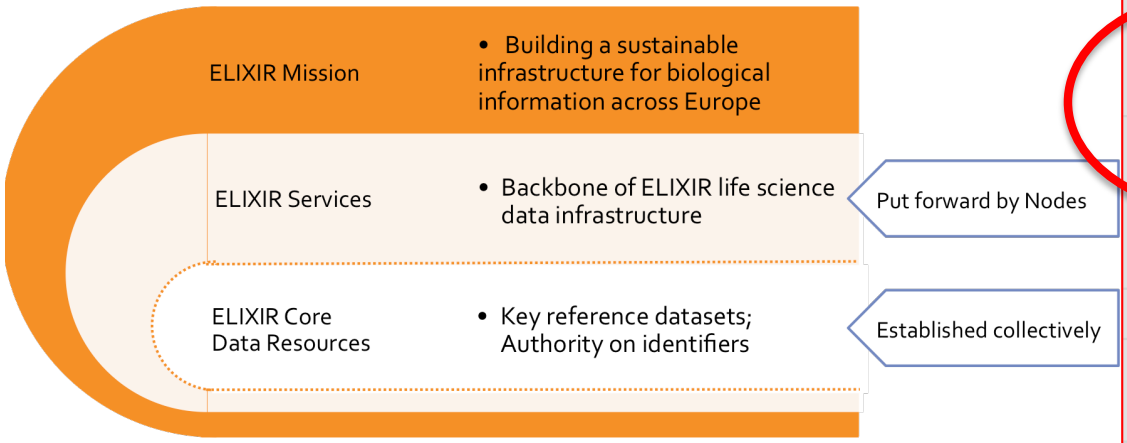
- International mutual recognition – code-of-conducts, policies
- Institutional maturation models (cf OECD)
- *Bona fide* researcher status management (e.g. restricted services)

ELIXIR Core Data Resources

ELIXIR Core Data Resources are a set of European data resources of fundamental importance to the wider life-science community and the long-term preservation of biological data.

Identification of the ELIXIR Core Data Resources involves a careful evaluation of the multiple facets of the data resources. Indicators used in the evaluation are grouped into five categories:

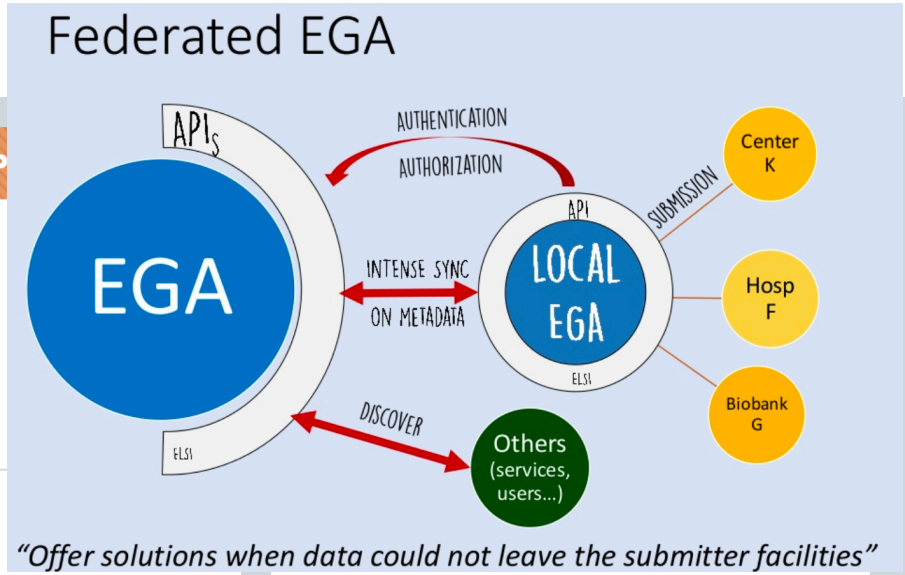
- Scientific focus and quality of science
- Community served by the resource
- Quality of service
- Legal and funding infrastructure, and governance
- Impact and translational stories

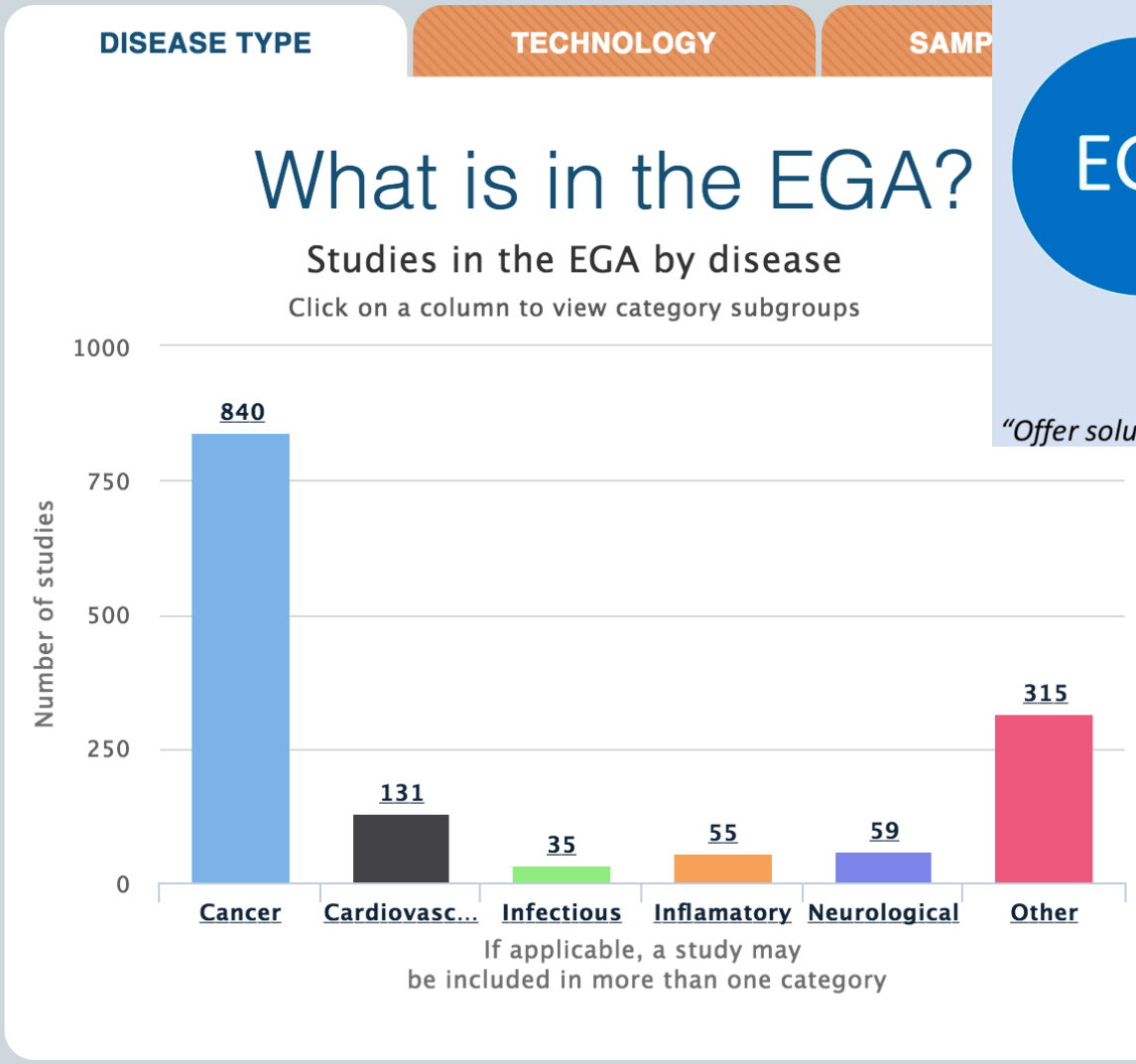


Core Data Resource	Data type
ArrayExpress	Functional Genomics Data from high-throughput functional genomics experiments.
CATH	A hierarchical domain classification of protein structures in the Protein Data Bank.
ChEBI	Dictionary of molecular entities focused on 'small' chemical compounds.
ChEMBL	Database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties and abstracted bioactivities.
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.
ENA	Nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.
Ensembl	Genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation.
Ensembl Genomes	Comparative analysis, data mining and visualisation for the genomes of non-vertebrate species.
Europe PMC	Europe PMC is a repository, providing access to worldwide life sciences articles, books, patents and clinical guidelines.
Human Protein Atlas	The Human Protein Atlas contains information for a large majority of all human protein-coding genes regarding the expression and localization of the corresponding proteins based on both RNA and protein data.
The IMEx Consortium: represented by IntAct and MINT	ntAct provides a freely available, open source database system and analysis tools for molecular interaction data. MINT focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators.
InterPro	Functional analysis of protein sequences by classifying them into families and ...ing the presence of domains and important sites. *...is is an umbrella resource to which many collaborating databases contribute. In ...nterPro as a Core Data Resource, the critical role of the constituent databases is ...ed.*
PDBe	Biological macromolecular structures.
PRIDE	Mass spectrometry-based proteomics data, including peptide and protein expression information (identifications and quantification values) and the supporting mass spectra evidence.
STRING-db	Known and predicted protein-protein interactions.
UniProt	Comprehensive resource for protein sequence and annotation data.

European Genome-Phenome Archive (EGA)

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.





represents a powerful and cost-effective opportunity to gain insights into the genetics of complex diseases. **Read more** →

Study 1 / 4

Next Study

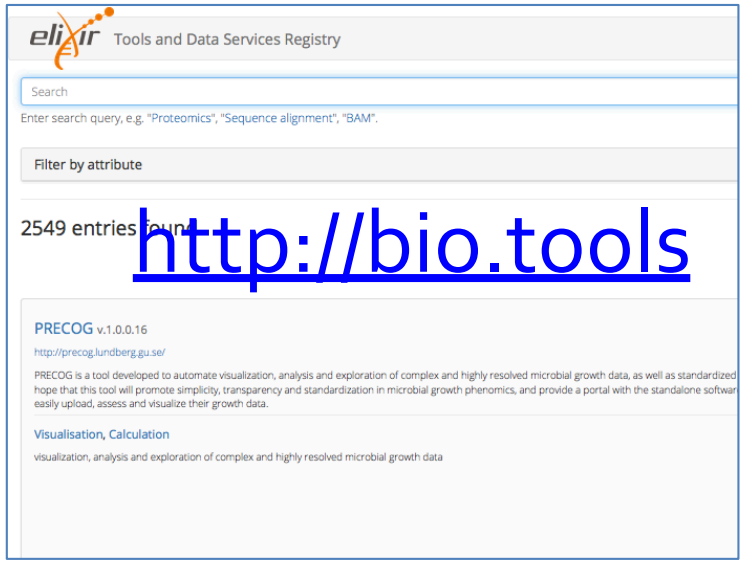
Published in:

wellcome trust sanger institute

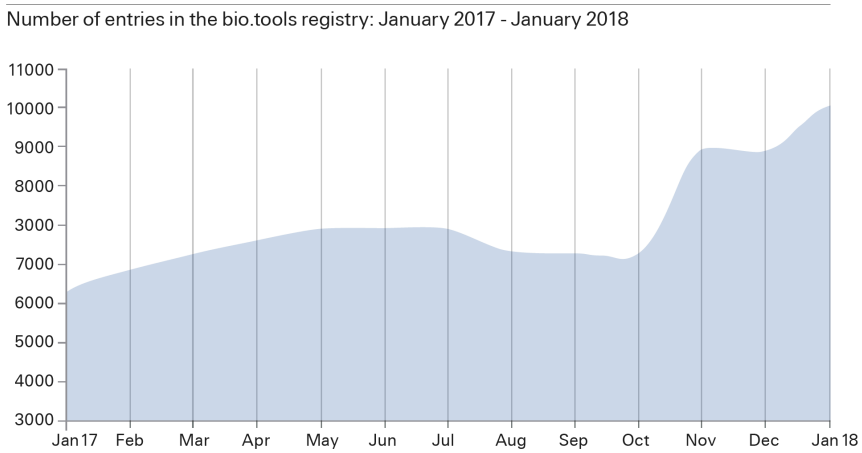
nature COMMUNICATIONS

bio.tools: ELIXIR tools and data registry

- Discovery portal for life science tools and data resources
- Easy to browse, search and update
- Based on EDAM ontology
- Over 10,000 entries and growing
  - Community-driven curation through hackathons and workshops
- Run by ELIXIR Denmark

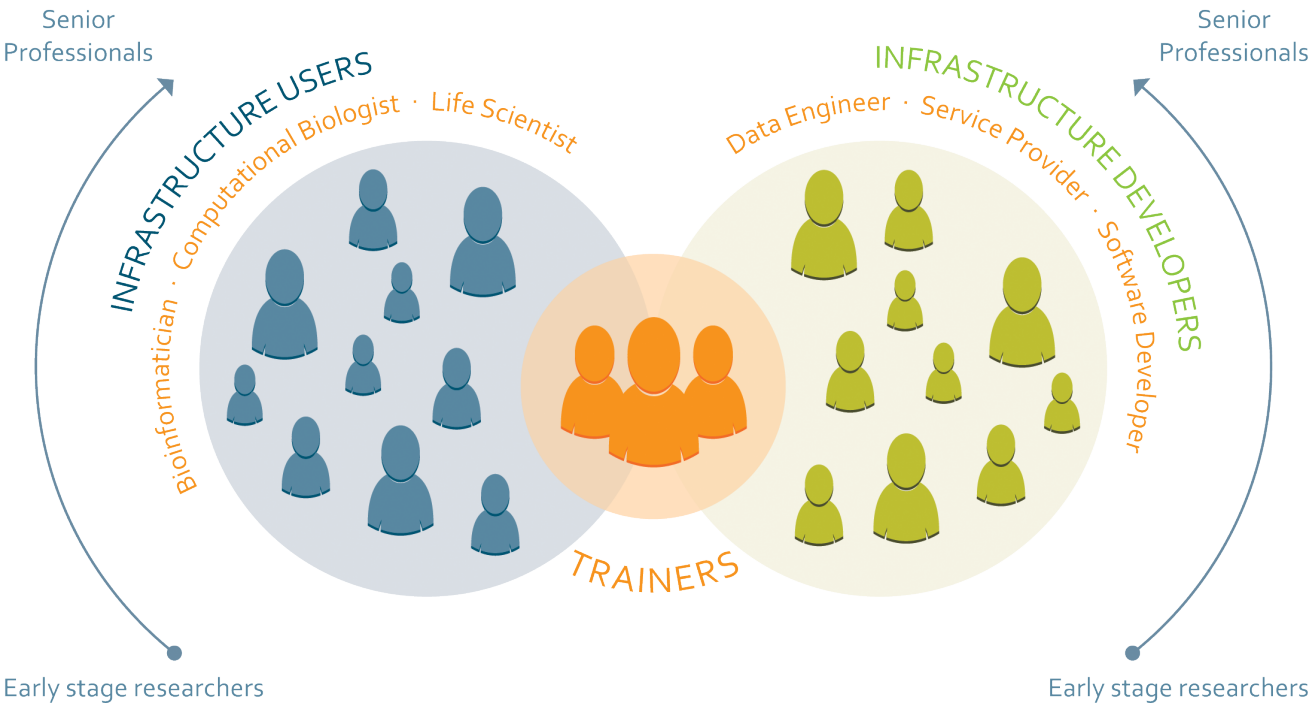


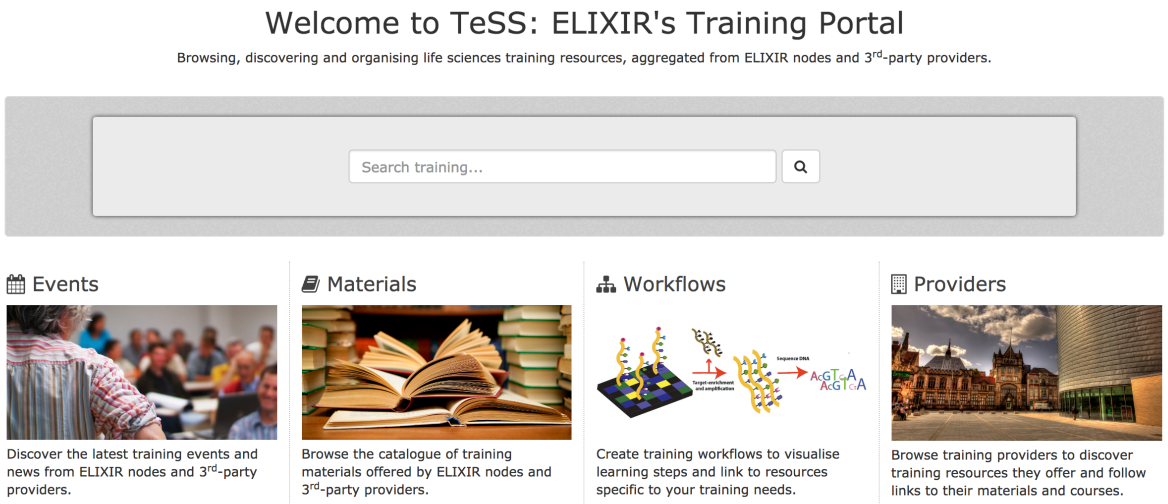
http://bio.tools



Tools and data services registry: a community effort to document bioinformatics resources.- NAR Jan 2016

ELIXIR Training programme

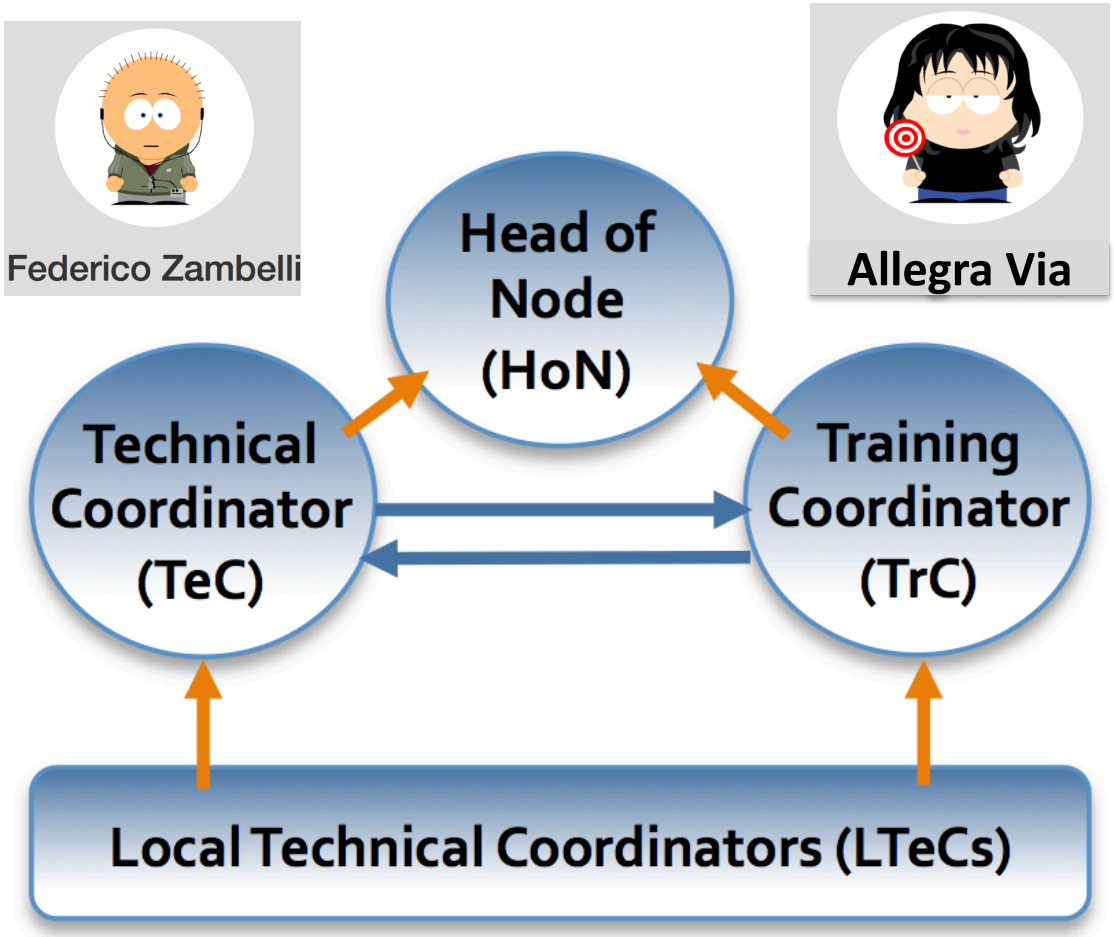




ELIXIR-Italy: a distributed ELIXIR Node

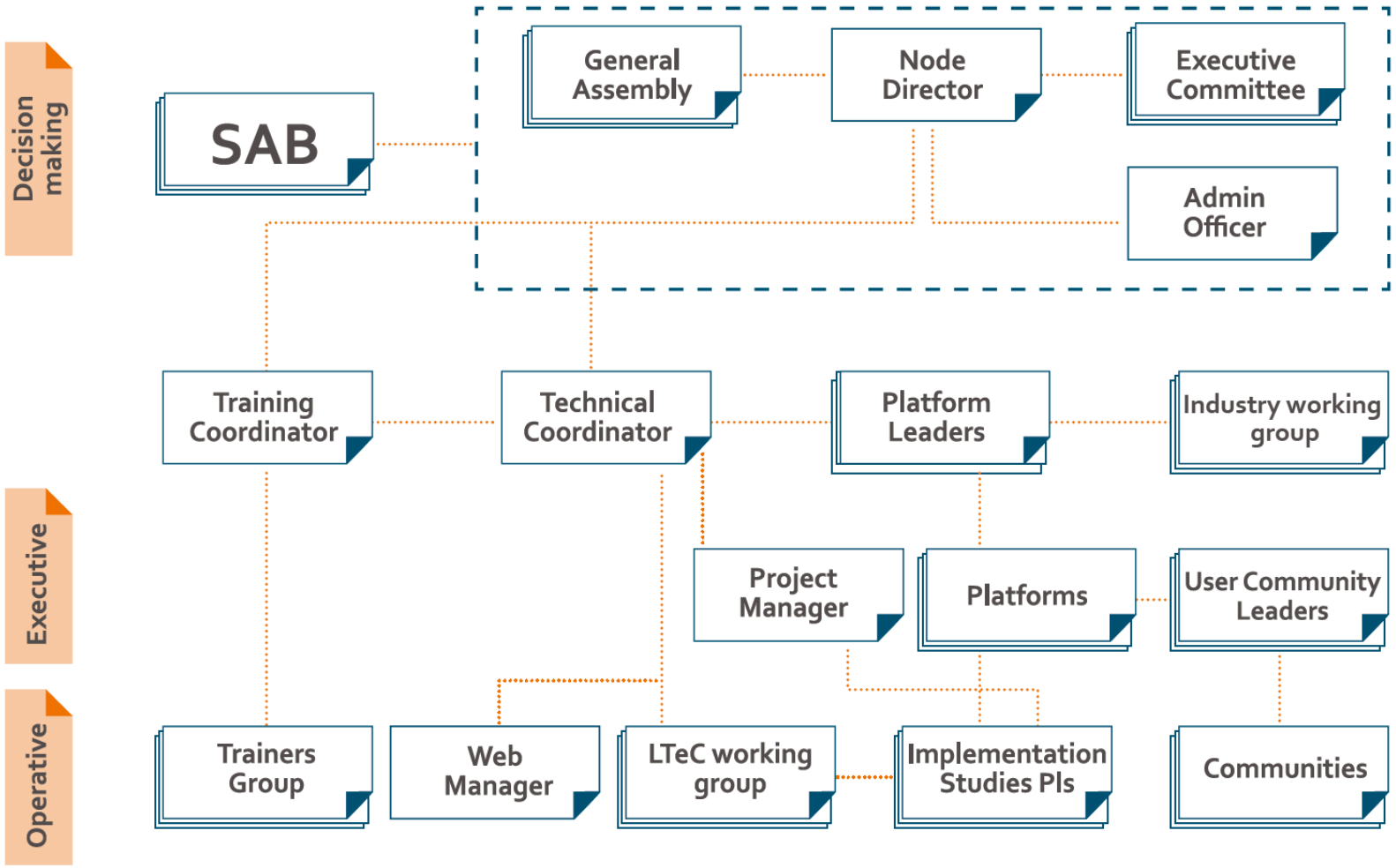
The Italian node is configured as a **Joint Research Unit** (JRU) -named **ELIXIR-IIB-** and is in charge of coordinating the delivery of existing bioinformatics services at the national level, also pursuing their integration in the ELIXIR infrastructure (ECA signed on Dec 2015). ELIXIR-IIB is led by National Research Council (CNR) of Italy and comprises other **22 partners** including several universities as well as leading high-performance computing partners such as CINECA, CRS4, GARR and INFN. ELIXIR-IIB also has strong local connections with other Italian nodes of ESFRI Biomedical Science and Environmental Science Infrastructures (e.g. **LifeWatch, BBMRI, EMBRC, MIRRI, etc**).

ELIXIR-IIB Coordination framework



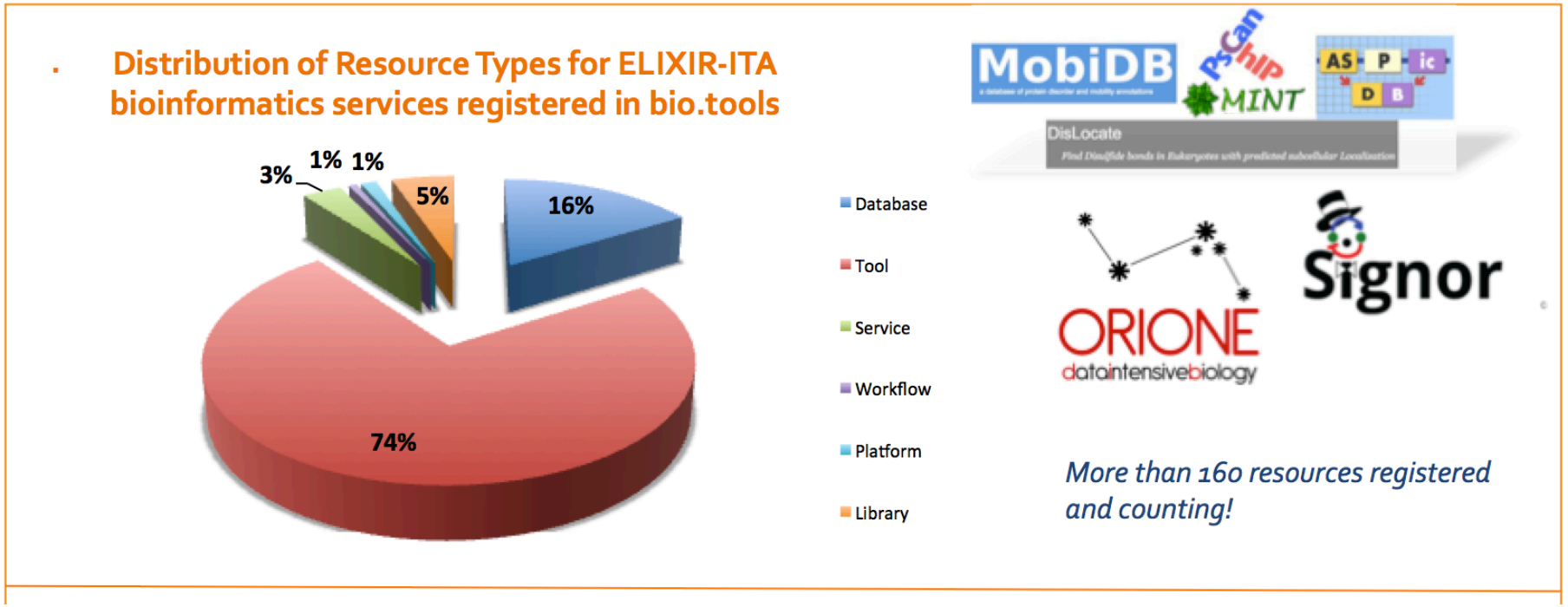
A distributed node with many partners spread all over Italy needs a continuous, quick and efficient exchange of informations in order to coordinate its activities. This is achieved through the **Local Technical Coordinators** (LTeCs) network. Each ELIXIR-IIB representative appoints a LTeC with the responsibility to keep the ELIXIR-IIB Technical and Training Coordinators and the other LTeCs informed about ongoing activities in their local institution and to report back to their institution representative. This framework helps fostering ideas and collaboration among ELIXIR-IIB members.

ELIXIR-IT
Organigram



(… about 100 FTE, in kind)

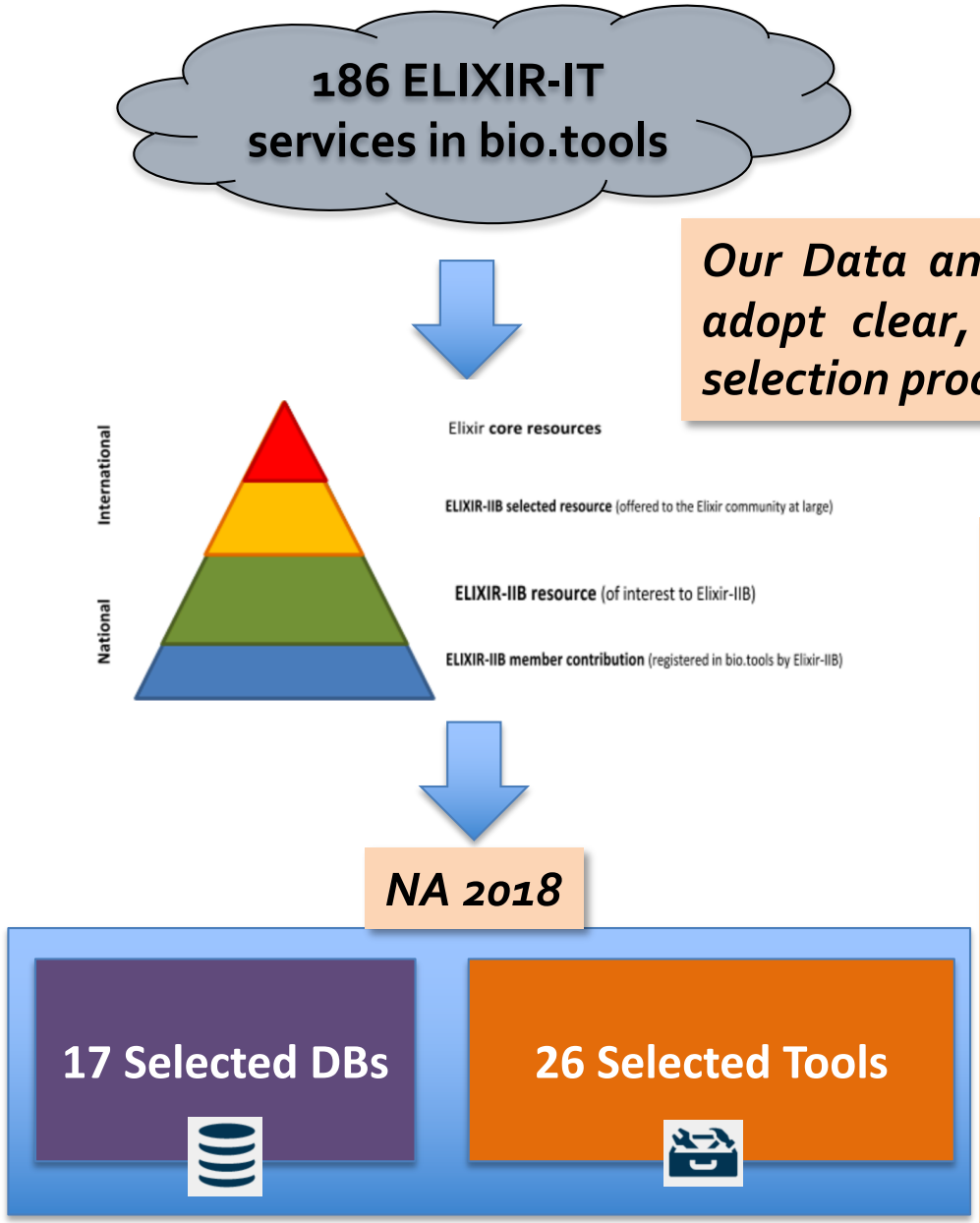
ELIXIR-IIB Service Registry



185 registered services run by **ELIXIR-IIB** members

The registry is not reserved to services run by ELIXIR members. As **ELIXIR-IIB** we aim to promote the registration of any service and database run by Italian researchers and institutions. *We **invite everybody here to contact us whether they want or need assistance in registering their services within the ELIXIR Service Registry at bio.tools** .*

Tools and Data Services Quality Management Policy



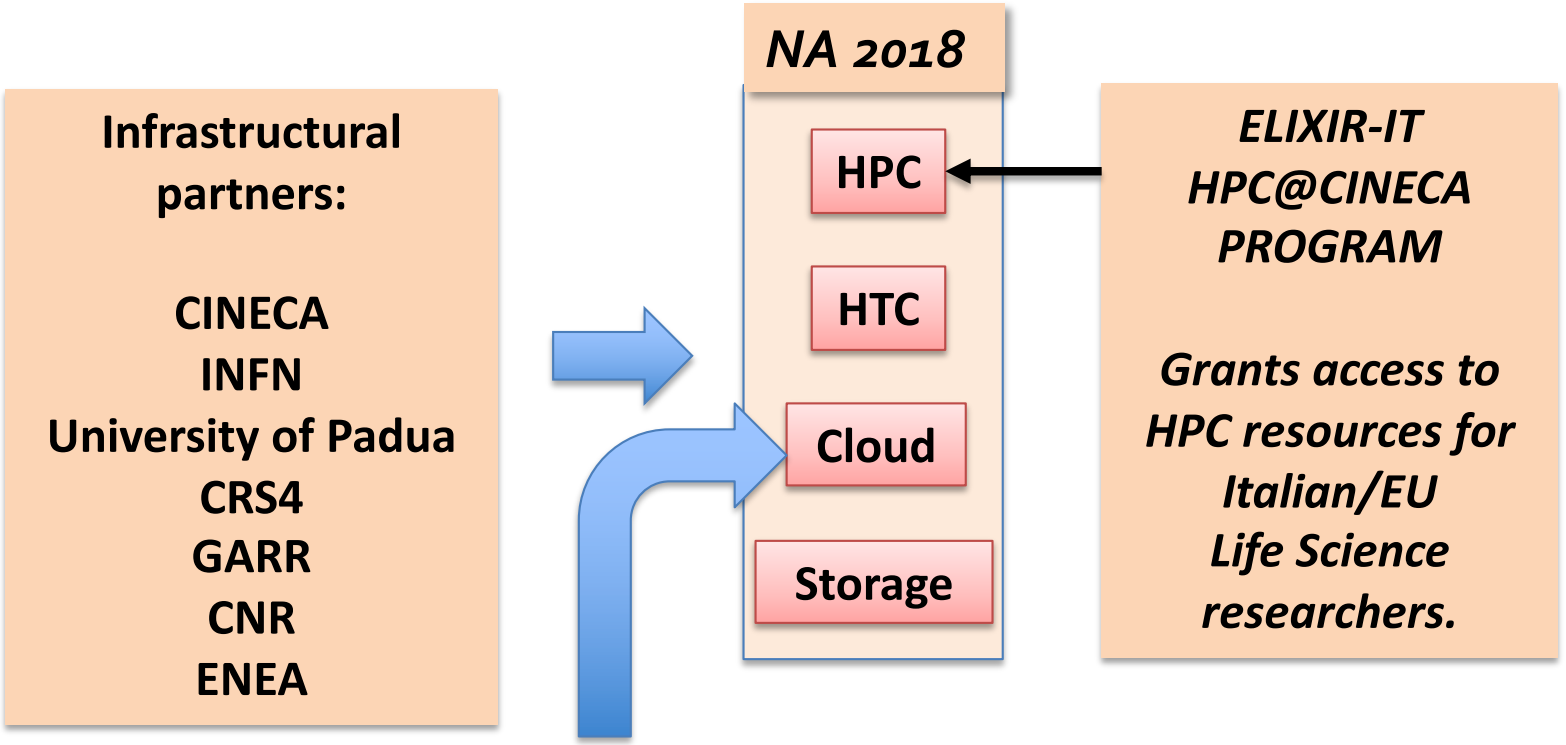
Our Data and Tools Services QM Policy allowed us to adopt clear, shared and accepted criteria during the selection process of the services to be included in the NA.

Only the two top tiers of services have been included in the NA, that are:

- ***ELIXIR Core Resources***
  - *Determined by ELIXIR through the Core Resources selection Process.*
- ***ELIXIR-IIB Selected Resources***
  - *Stringent quality criteria based on*
    - ✓ *Scientific focus*
    - ✓ *Diffusion in the community*
    - ✓ *Technical parameters*
    - ✓ *Legal framework*

Compute Services

Compute platform



*The Node aims to provide a **PaaS Layer** based on the technology developed within the **INDIGO-DataCloud H2020 Project** enabling the federation of ELIXIR Compute resources, this layer will be compatible with the already available platform both in terms of APIs and AAI. Thanks to this layer it would be possible to instantiate complex clusters of service and provide dynamic usage of the resources. The INDIGO PaaS will also provide a solution for the secure and private access to data, that could help fulfilling the requirements in terms of data protection.*

ELIXIR-IIB HPC@CINECA

- This pilot project started in April 2016. First example of service offered at Node level rather than local Node level
- HPC@CINECA offers CINECA HPC resources to the Italian (and beyond) bioinformatics community.
- Users have access to an HPC bioinformatics platform through a streamlined project review procedure.
- The base package provides 50K core hours and 5Tb of storage for six months. Special needs can be addressed.

	Total Nodes	CPU	Cores per Nodes	Memory (RAM)	Notes
Compute/login node	66	Intel Xeon E5 2670 v2 @2.5Ghz	20	128 GB	
Visualization node	2	Intel Xeon E5 2670 v2 @ 2.5Ghz	20	128 GB	2 GPU Nvidia K40
Big Mem node	2	Intel Xeon E5 2650 v2 @ 2.6 Ghz	16	512 GB	1 GPU Nvidia K20
BigInsight node	4	Intel Xeon E5 2650 v2 @ 2.6 Ghz	16	64 GB	32TB of local disk

INFN/UNIBA resources to ELIXIR ITA

INFN/UNIBA could provide access to large computing facilities with different scope and optimization. All the resources, which can be provided as both IaaS cloud and high-level cloud services, could be both accessed directly or via easy and user-friendly interfaces (e.g. **BioMas, MetaShot** for metagenomic data analysis).

13'000
Cpu Core

5.6 PByte
Storage

2.5 Pbyte
Tape

800 Cpu Core HPC
20 NVIDIA K40

Cloud resources

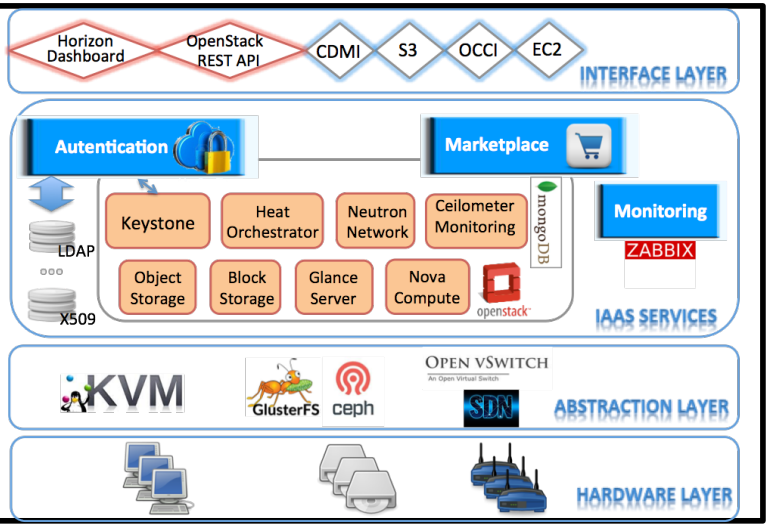
1700 CPU/Core

6.7 TB of RAM memory

270 TB of Storage

256 Public IP

10Gbit/s connection



Training Services

Training platform

The Italian ELIXIR Node regularly delivers **TRAINING** by designing, organising and delivering courses covering the following topics:

- bioinformatics tools and resources
- computational skills
- (bio)data science
- data management, annotation and analysis
- data interoperability and FAIRification
- bio.tools and bio.schemas
- ...and others

ELIXIR-IT Training Platform Guidelines

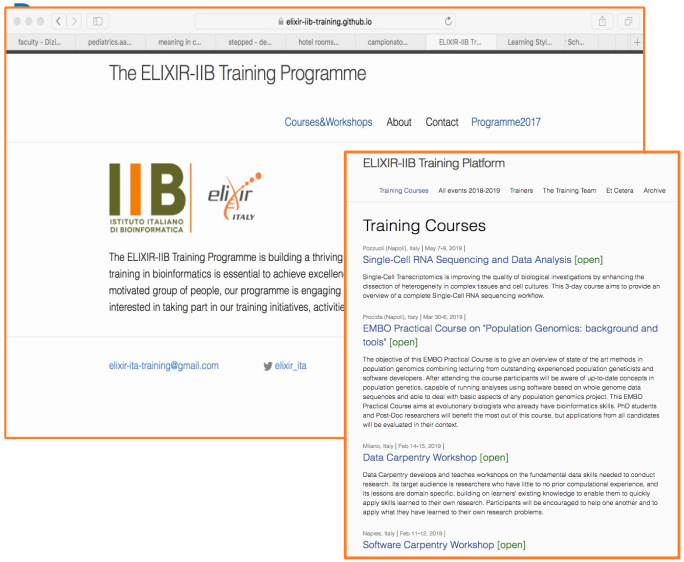
TrT Program

21
Training Courses
in 2017-18
+
Other events involving ELIXIR-IT Training (e.g. Train the Trainer, BYOD, etc...)

ELIXIR-IIB Training Platform

Supporting researchers and professionals in acquiring specialized computational and data management skills

Web



https://elixir-iib-training.github.io/website/

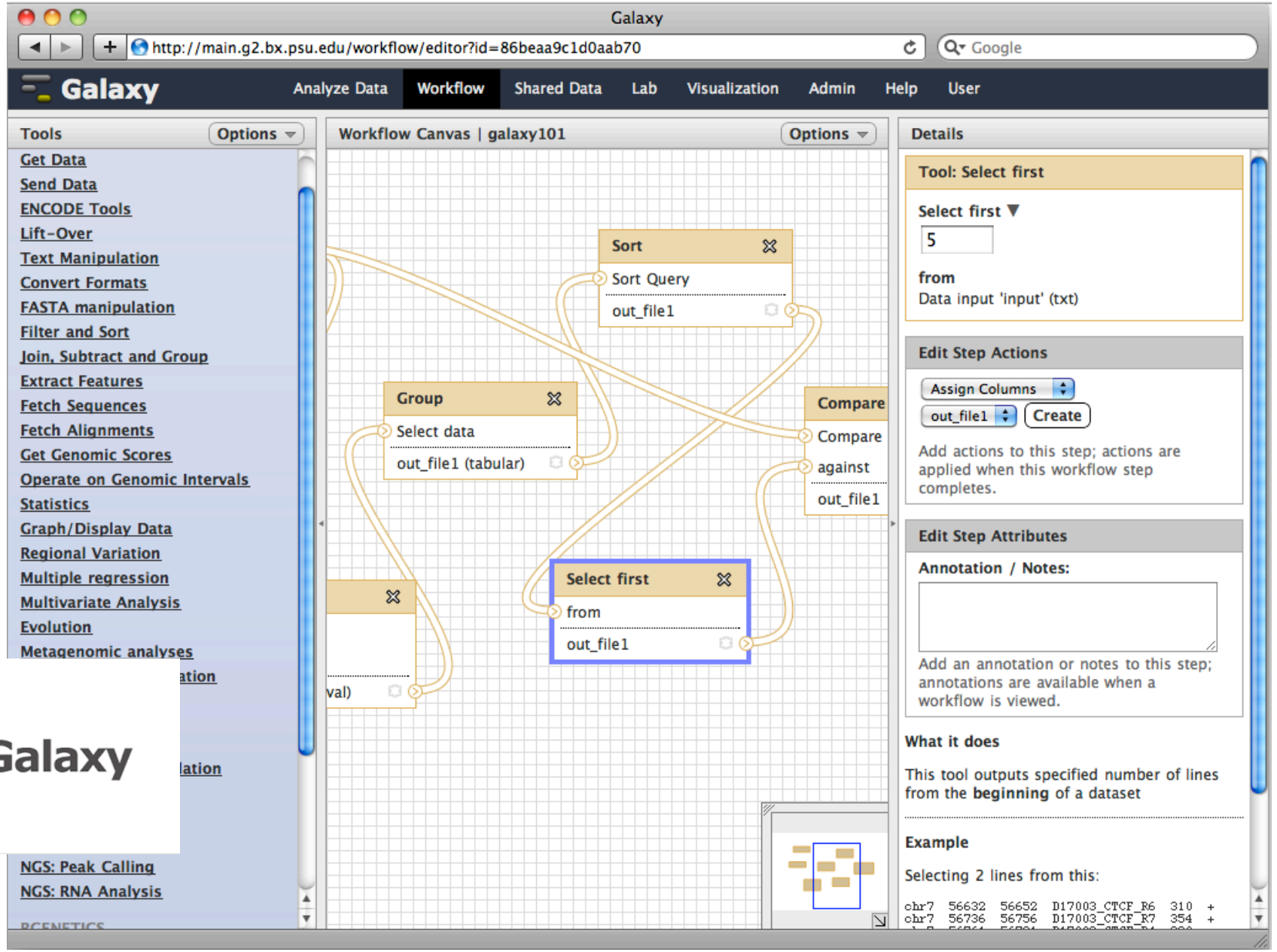
@elixir_ita

elixir.ita.training@gmail.com

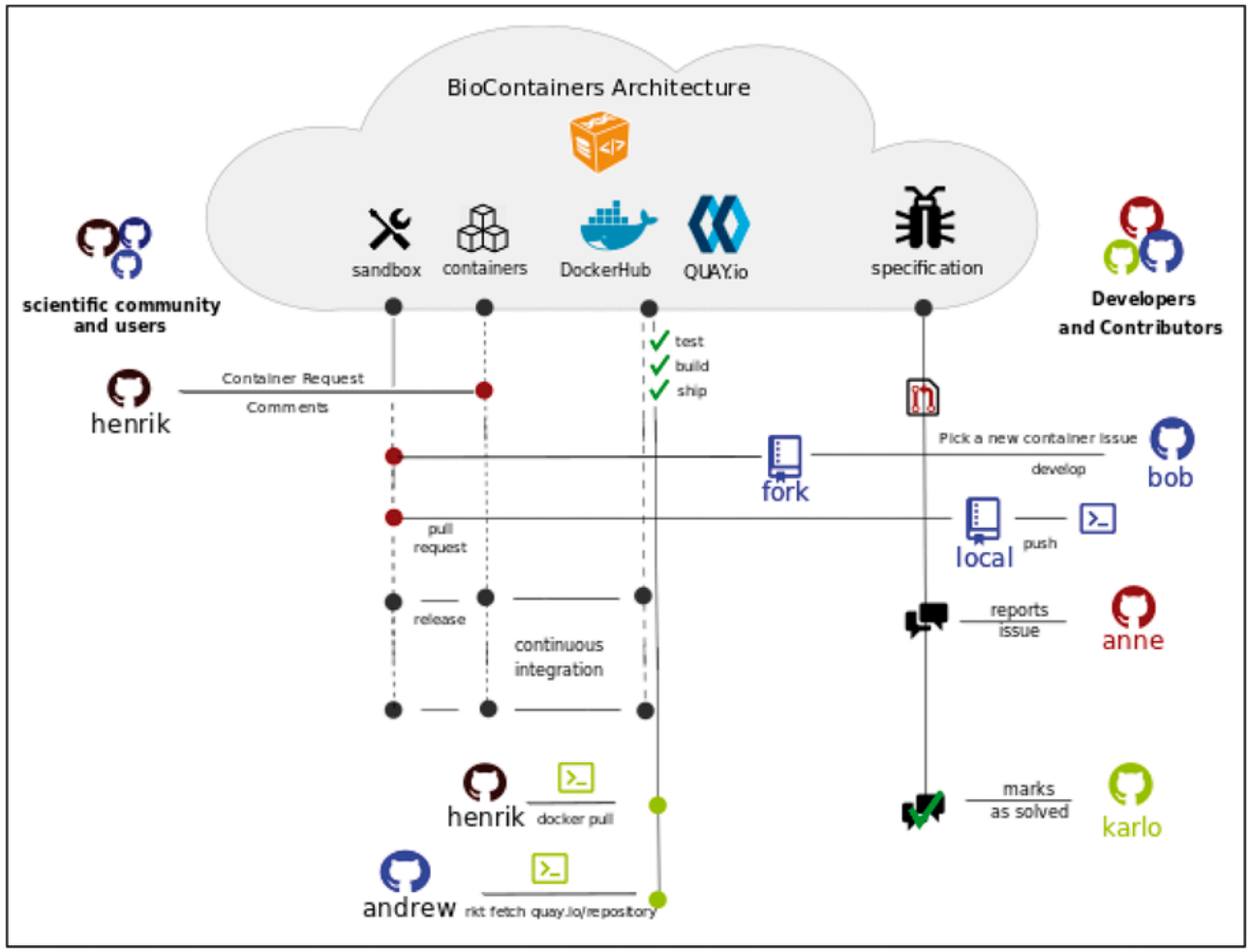
- ❑ <25 participants to allow *high interactivity* and intensive *practical* work
- ❑ trainers selected among *experts* in the course topic
- ❑ advanced *teaching techniques*

- ❑ activities with other ELIXIR Nodes:
  - Train the Trainer
  - eLearning
  - Implementation Studies training support
  - Best Practices for Software Development
  - Quality & Impact programme
  - Bring Your Own Data events for Rare Disease patient registries

Making Bioinformatic tools accessible: Galaxy



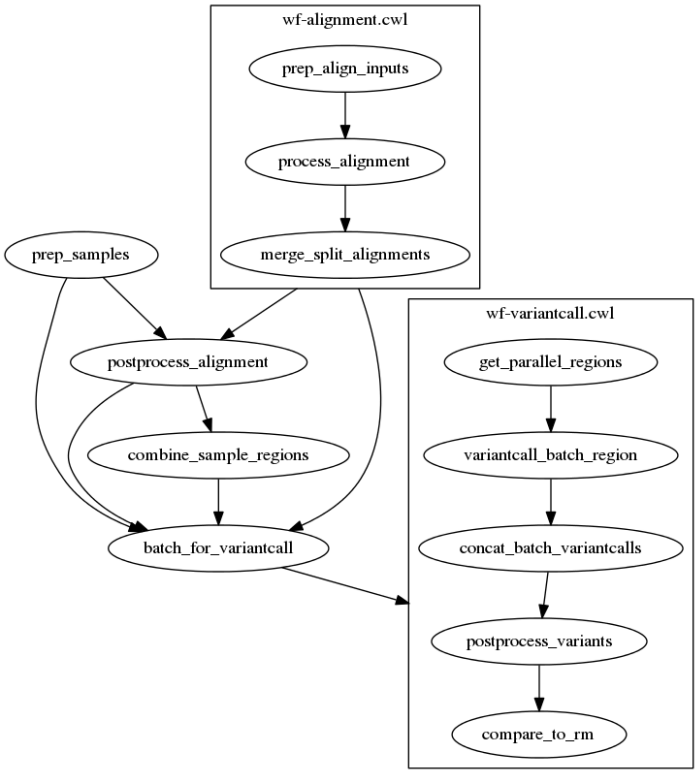
Making Bioinformatic tools reproducible: Containers



Overview of the BioContainers architecture: Users and developers can use the BioContainers infrastructure by interacting via GitHub account page. All container Dockerfiles are freely available and people are encouraged to participate submitting pull requests or asking for new containerized software. Containers can be acquired via Docker command line interface, or by downloading the Dockerfile directly from the GitHub organization

Making Bioinformatic tools reproducible: CWL

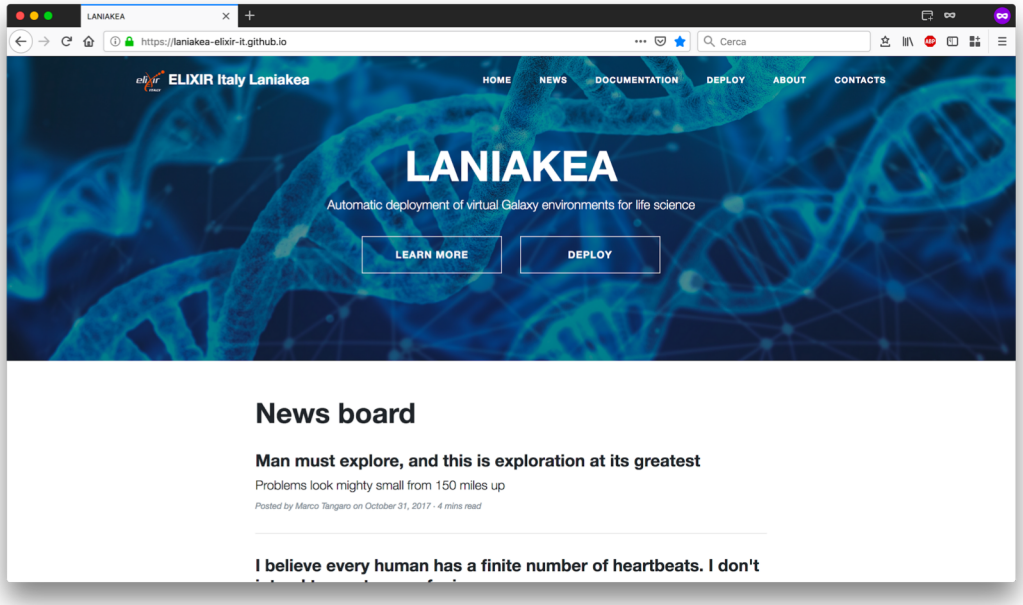
COMMON WORKFLOW LANGUAGE



The Common Workflow Language (CWL) is a specification for describing analysis workflows and tools in a way that makes them portable and scalable across a variety of software and hardware environments, from workstations to cluster, cloud, and high performance computing (HPC) environments. CWL is designed to meet the needs of data-intensive science, such as Bioinformatics.

Laniakea

LANIAKEA is a cloud Galaxy instance provider, based on INDIGO-DataCloud software catalogue. Its architecture automates the creation of Galaxy-based virtualized environments.



INDIGO - DataCloud

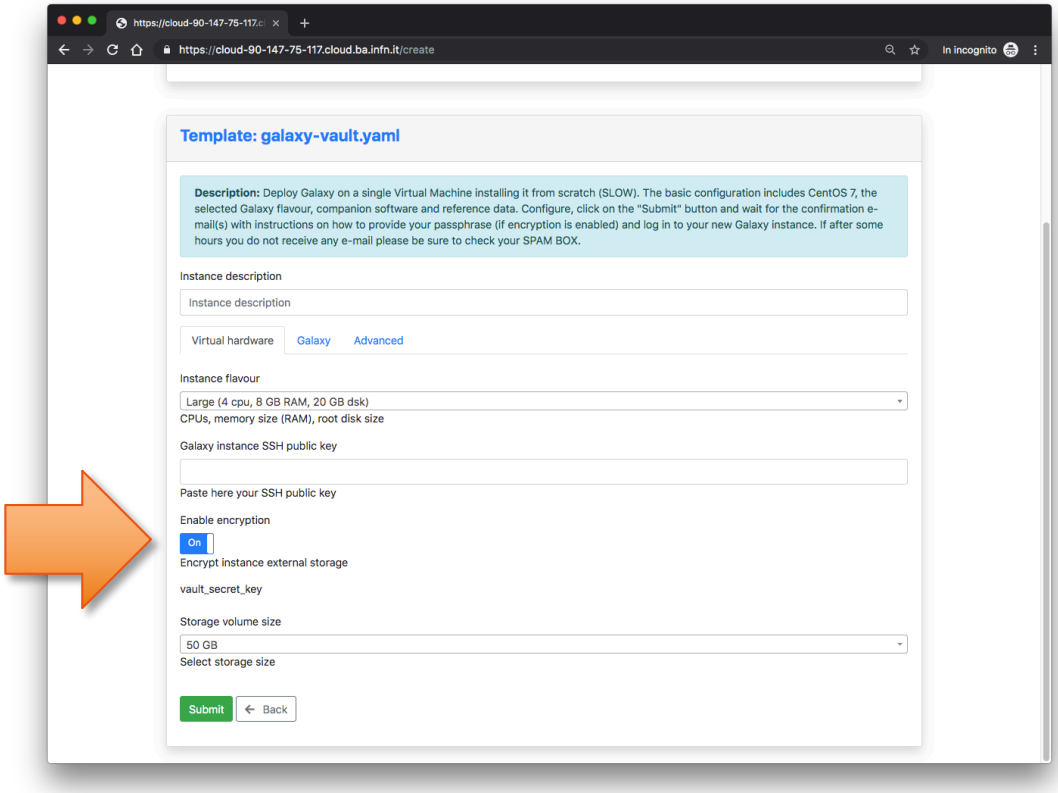
https://laniakea-elixir-it.github.io

(*) The Laniakea Supercluster (Laniakea; also called Local Supercluster or Local SCl or sometimes Lenakaeia) is the galaxy supercluster that is home to the Milky Way and approximately 100,000 other nearby galaxies [Wikipedia].

Laniakea features

- **Galaxy production environment** - The deployed Galaxy instance supports a multi-user production environment.
- **Galaxy flavours available** - Each Galaxy instance is customizable with different sets of pre-installed tools.
- **Shared reference data** - Each instance comes with reference data (e.g. genomic sequences) already available for many species (shared among all the instances). Galaxy is automatically configured to properly use them.
- **Galaxy with cluster support** for compute-intensive tasks.
- **Persistent storage for data** with and without encryption (see next slide).

Laniakea Storage encryption



Users data encryption to facilitate GDPR compliance and isolate user data.

The encryption procedure has been completely automated, allowing the user to straightforwardly encrypt storage on-demand with a strong random alphanumerical passphrase.

Once the storage is encrypted and the User can retrieve his random passphrase from Laniakea portal.

Laniakea information

The production phase of the ELIXIR-IT Laniakea@ReCaS service will start in the second half of 2019 and will be announced on ELIXIR-ITALY mailing list (http://tinyurl.com/elixir-it-ml).

Paper submitted to Biorxiv:
Laniakea: an open solution to provide Galaxy "on-demand" instances over heterogeneous cloud infrastructures.
url: https://www.biorxiv.org/content/early/2018/11/19/472464
doi: https://doi.org/10.1101/472464

Useful links:

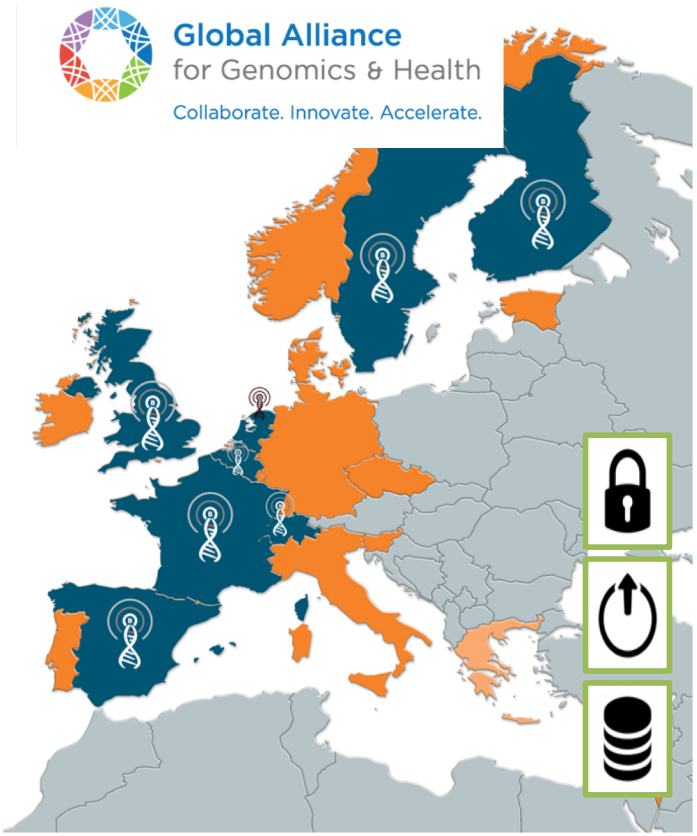
- New website: https://laniakea-elixir-it.github.io
- Documentation: http://laniakea.readthedocs.io
- GitHub code: https://github.com/Laniakea-elixir-it
- Demo video: https://www.youtube.com/watch?v=rub3skcs84Q

Genetic Data Sharing (Beacon)

Data sharing is key to the success of medical research. However, getting the right balance between protecting sensitive genomic information and making it useful to researchers has been a challenge.

A project called **The Beacon Project**, initiated by the Global Alliance for Genomics & Health (GA4GH) and funded in part by the National Human Genome Research Institute (NHGRI) represents a giant leap forward secure data sharing.

The data is stored on servers, known as beacons, at institutions participating in the project, and users can query the beacons for genomic information stored in that beacon. Essentially, the system allows a user to ask whether a specific nucleotide (an A, T, C or G) exists at a particular chromosome location in any genome in a given beacon, but keeps all other sequence data concealed. This would allow a clinician to check whether a patient's mutation had been discovered in other patients without needing access to those other patients' genomes.



www.elixir-europe.org/beacons

General Data Protection Regulation (GDPR)

European regulation on personal data protection becomes effective on 25 May 2018.

A European regulation is enforceable as **law** in all member states simultaneously (differently from directives that need to be transposed into national law).

According to GDPR managing and analysing human genetic data **must** be compliant to GDPR specific requirements.

Governance

- Data Protection Officer (DPO)
- Data Controller
- Data Processor

Elixir IIB established a working group, in collaboration with ELIXIR, to provide assistance to Italian scientist in fulfilling GDPR compliance

ICT infrastructure

- AAI requirements
- Secure Data transfer
- Encryption strategies
- ..

Illumina NovaSeq 6000

Oxford Nanopore GridION

PacBio Sequel

Orbitrap Fusion

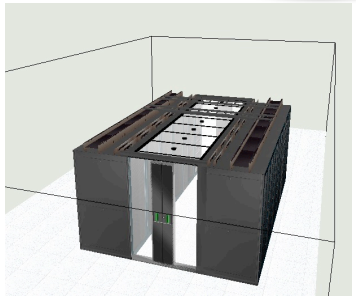
CNR.Biomics:
Centro di Eccellenza per le Scienze Omiche e la Bioinformatica

PON PIR01_00017
14.5 M€

BioNano Genomics

FACS Melody

Omnia LH100



ICT Facility
9K core – 7 Pb

10X Genomics

ELIXIR-IIB Website

Please visit elixir-italy.org for further info and finding specific resources.

ELIXIR IIB Contacts

ELIXIR	https://www.elixir-europe.org/
ELIXIR-Italy	http://elixir-italy.org/
ELIXIR-Italy Training:	http://bioinformaticstraining.pythonanywhere.com/
ELIXIR-Italy ML:	https://goo.gl/NUHMxZ
Head of Node:	g.pesole@ibiom.cnr.it
TeC:	federico.zambelli@unimi.it


THANK
YOU
return