BIG DATA & LIFE SCIENCES

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

elixir ITALY

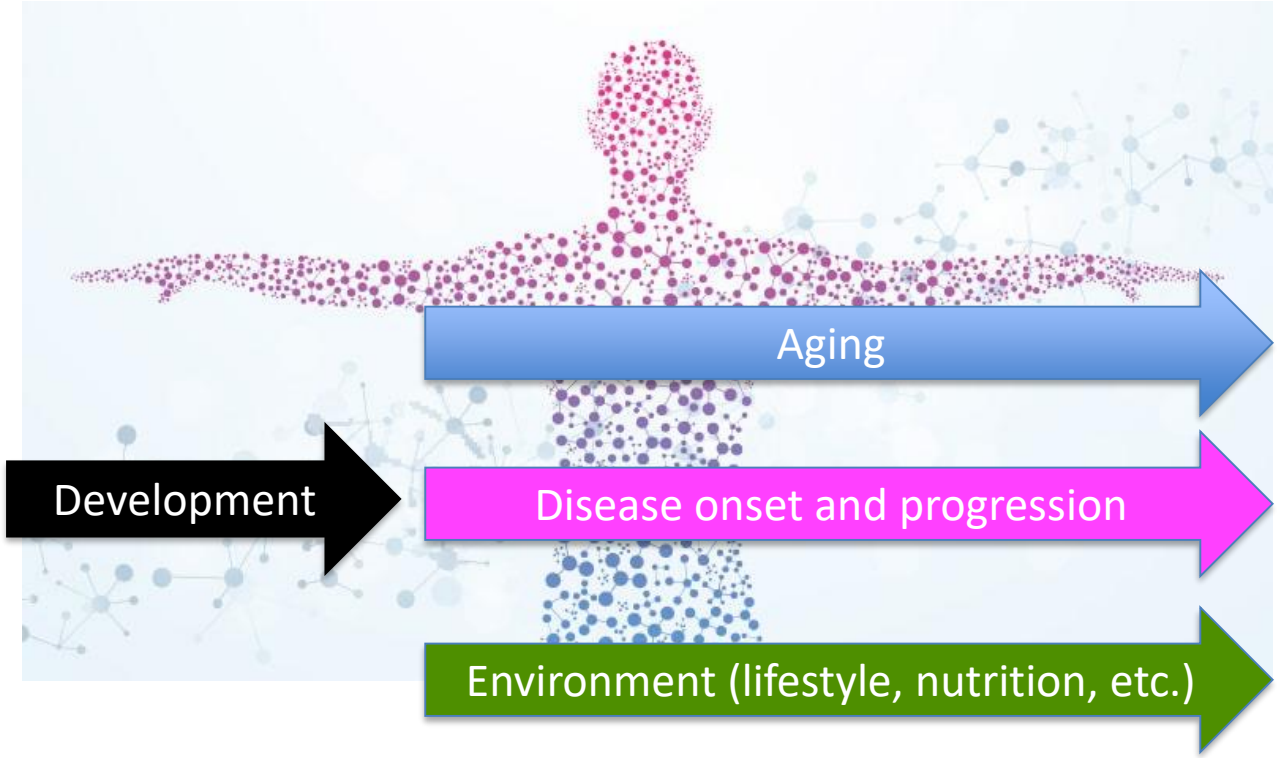
PAVIA, ITALY
25 > 27 OCTOBER
2018

CHANGES IN REGULATORY SCIENCES IN THE EU
how to move from a reactive
to a multi-stakeholder proactive attitude

ISTITUTI CLINICI SCIENTIFICI MAUGERI
Via Salvatore Maugeri, 6

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
- several microbiomes

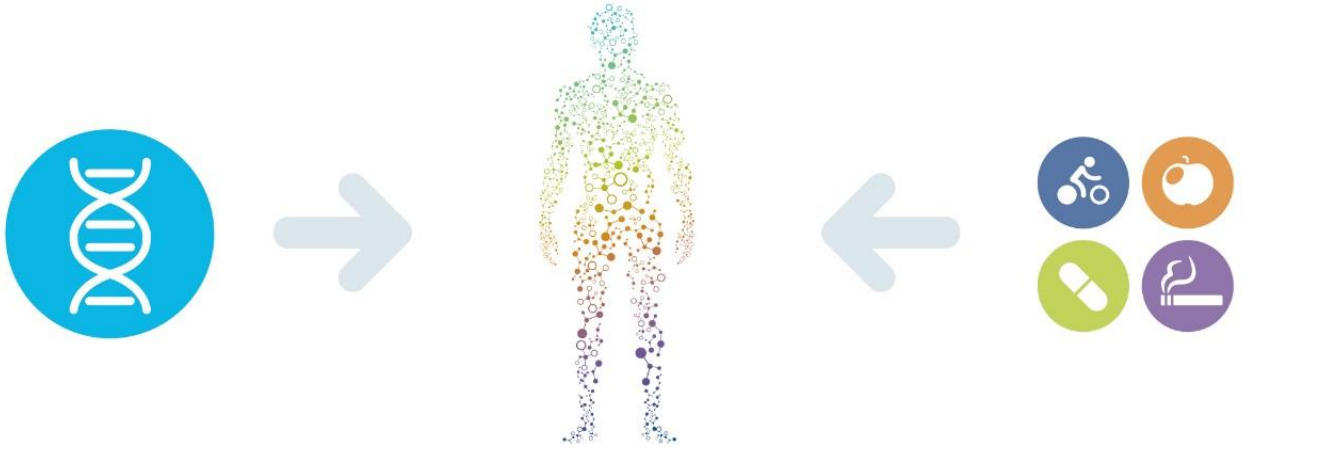
etc. proteomes, metabolomes

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

elixir ITALY

BIG DATA IN BIOLOGY

The uniqueness of each individual, and therefore his response to therapies, is determined by a combination of his genetic profile and environmental factors (diet, exercise, microbiota, etc.)

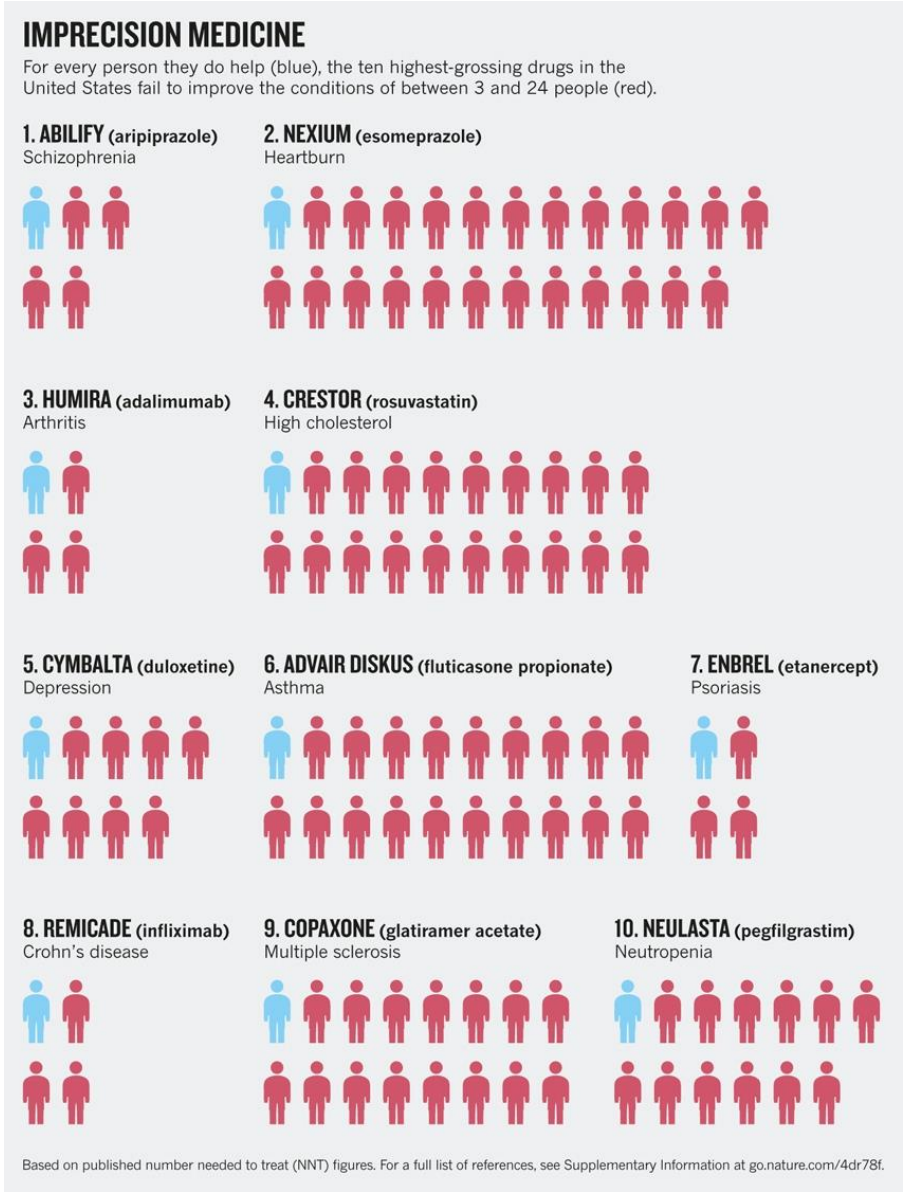


GENOMICS
Our genes can suggest what diseases we *might* be predisposed to, but it's an incomplete picture of human health.

PHENOTYPE
A snapshot of the current state of health that can be used to prevent, diagnose and treat disease or improve health.

LIFESTYLE/ENVIRONMENT
External factors like diet, exercise, medications, microbiota and even where we live influence our metabolic state.

Big Data in Biology: from Imprecision to Precision Medicine

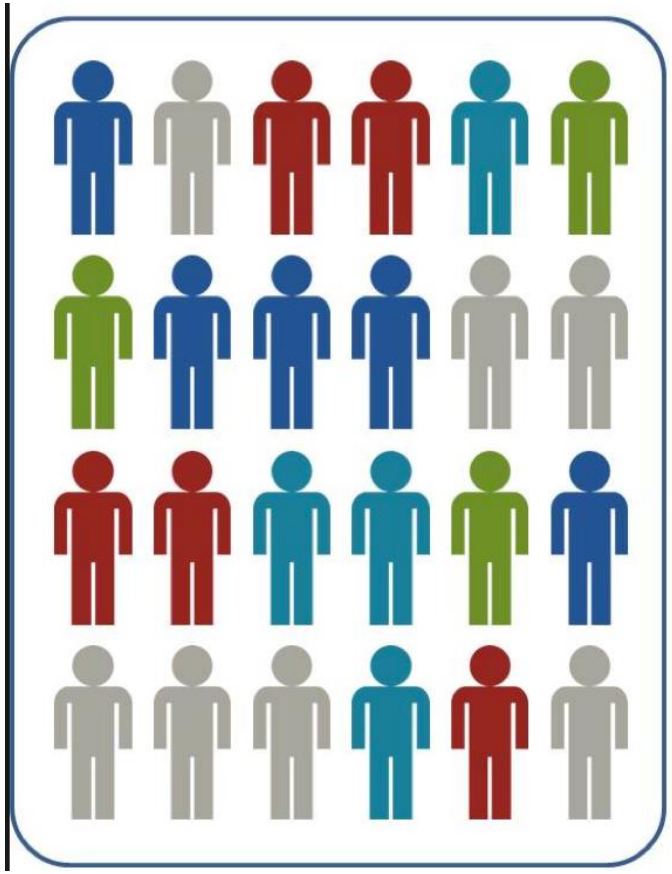


Most of the drugs that are administered to patients today are completely ineffective, and sometimes harmful. A study conducted in the USA on the 10 most expensive drugs has proved effectiveness in a percentage ranging between 2 and 25%. For example, statins administered to reduce cholesterol levels produce benefits in only 1 in 50 patients. Some drugs are also harmful to specific ethnic groups due to the fact that clinical trials are usually performed on ethnically defined individuals (e.g. caucasian people).

elixir ITALY

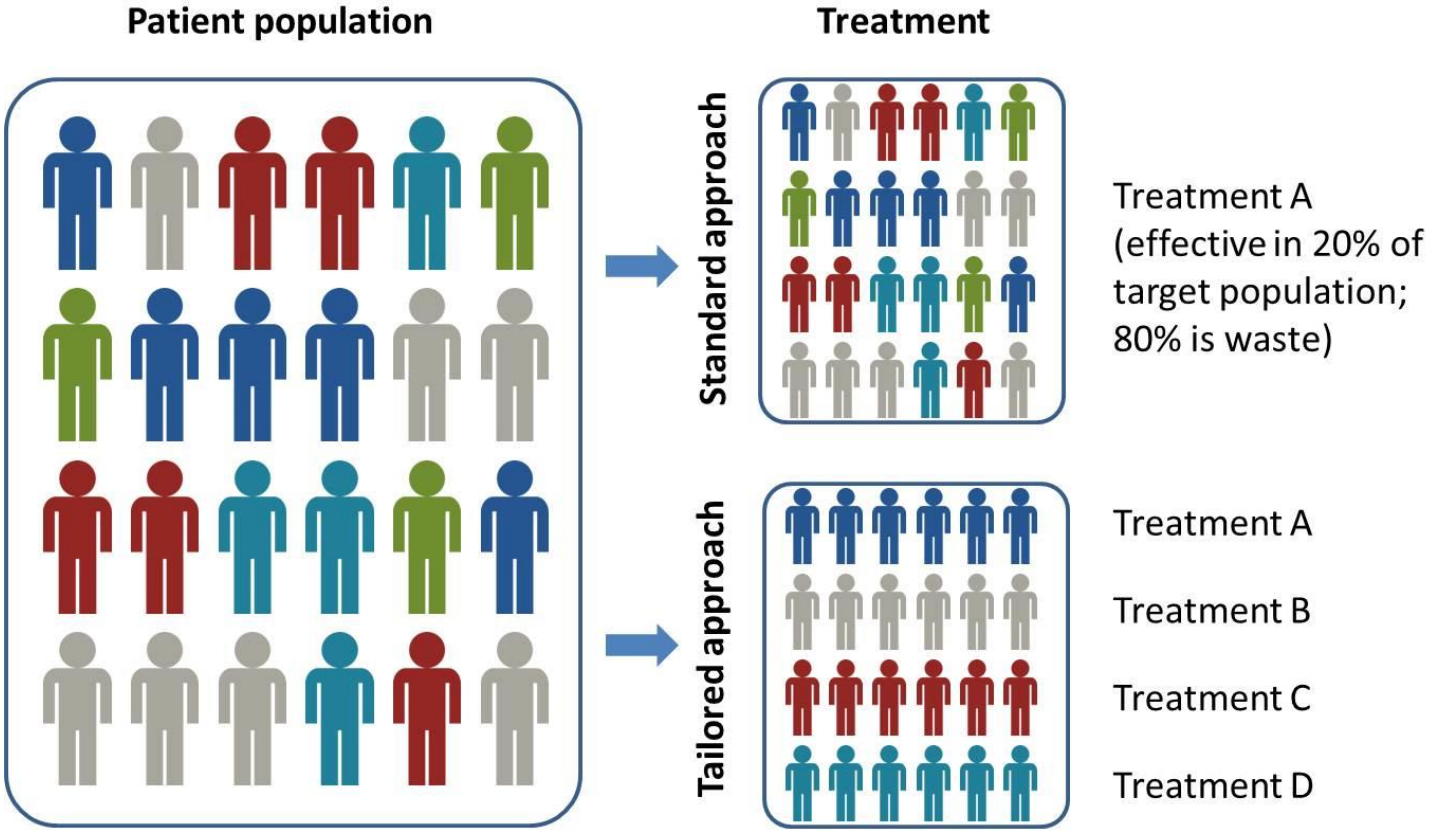
Big Data in Biology: from Imprecision to Precision Medicine

Although still unique, different individuals can be stratified into a specific class, if we consider the response to a pharmacological treatment.



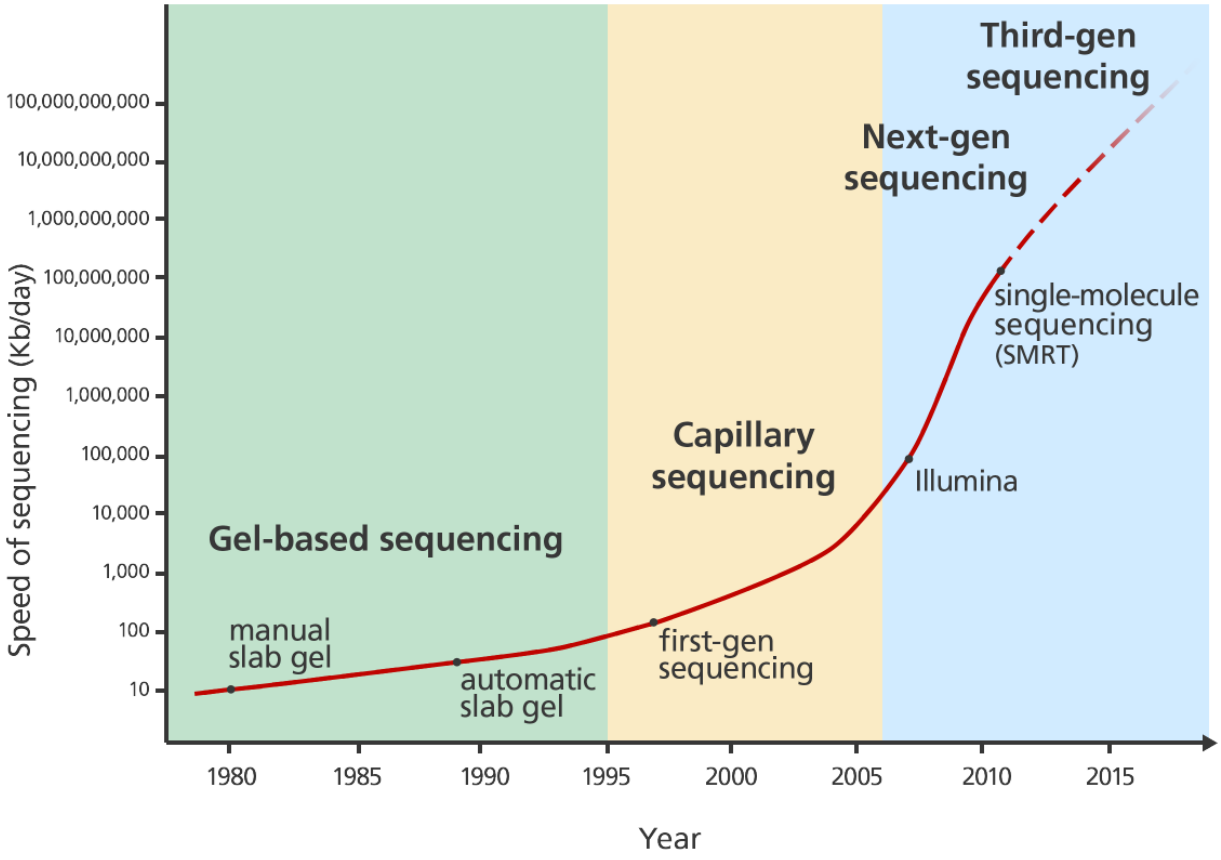
Big Data in Biology: from Imprecision to Precision Medicine

Given the classification of patients, based mainly on their genetic profile, but also on other characteristics, it is possible to put in place a much more effective and economic prophylaxis.



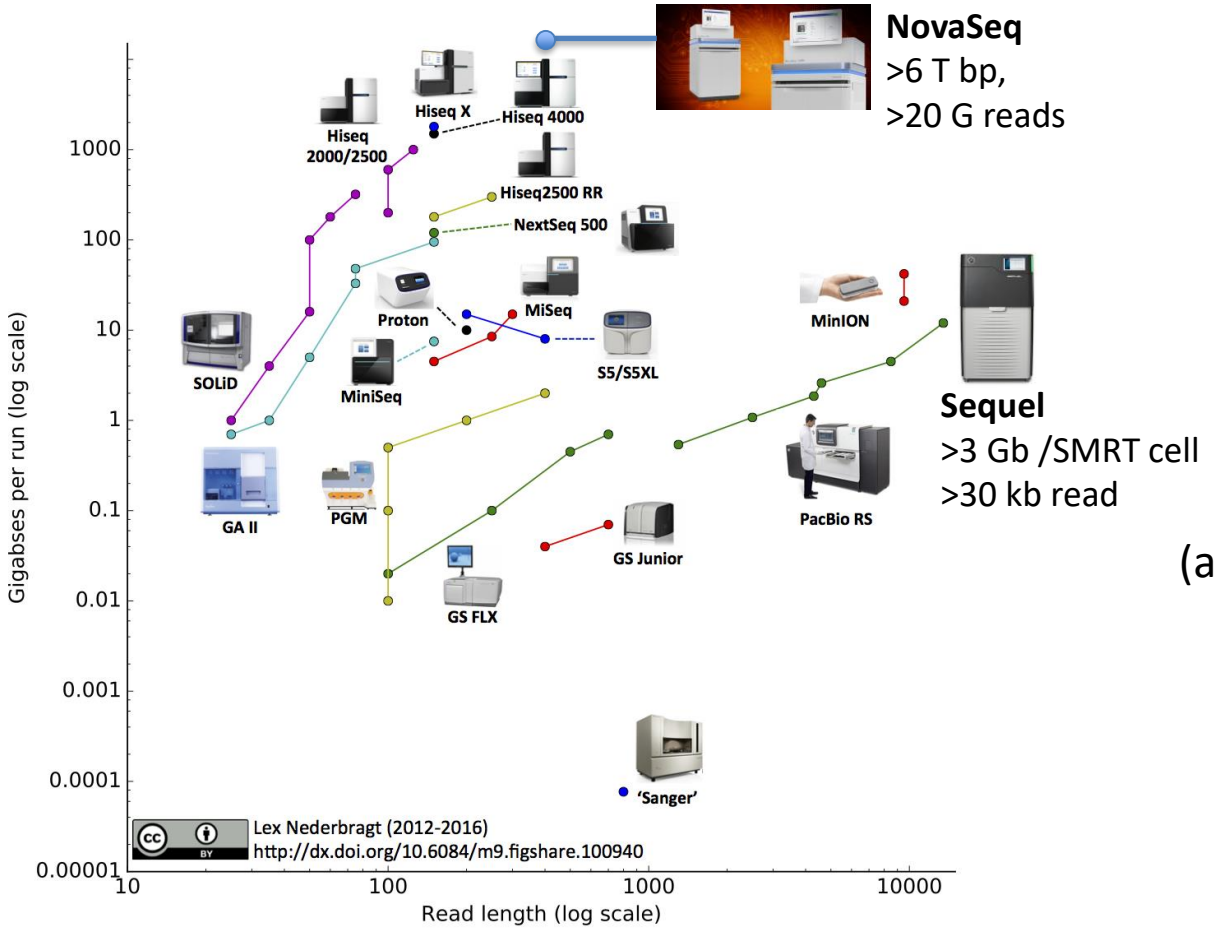
NGS REVOLUTION

The "Human Genome" project has burst exceptional technological innovations 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.
The human genome sequencing costs decreased about 10,000 fold in the last 20 years.



SECOND AND THIRD GENERATION MASSIVE SEQUENCING

The advent of high-throughput sequencing platforms has revolutionized biological research opening new amazing horizons.
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.



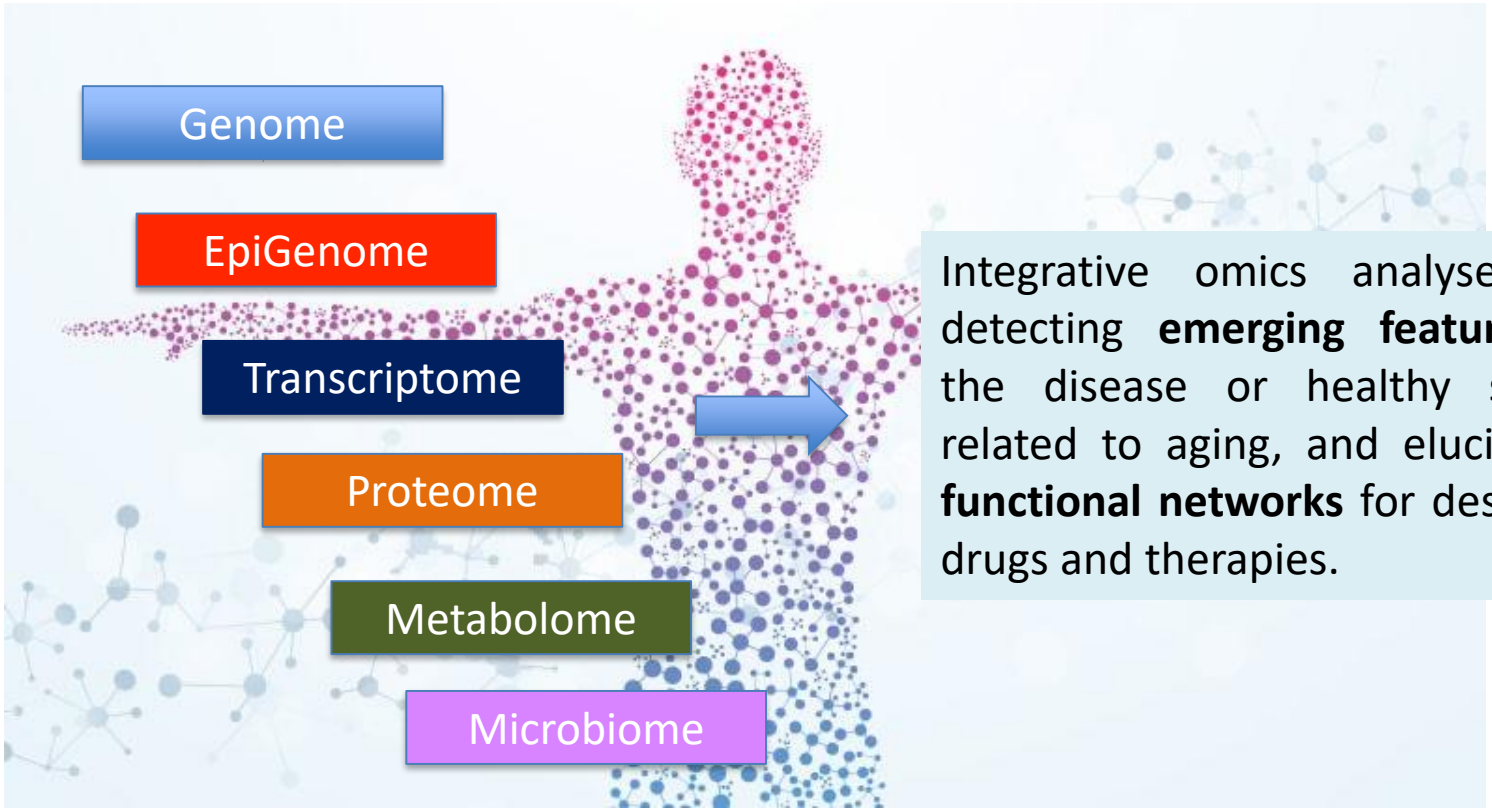
NovaSeq
>6 T bp,
>20 G reads

Sequel
>3 Gb /SMRT cell
>30 kb read

Volume of Data

Cost / bp
(about 30 cent / Mb)

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.

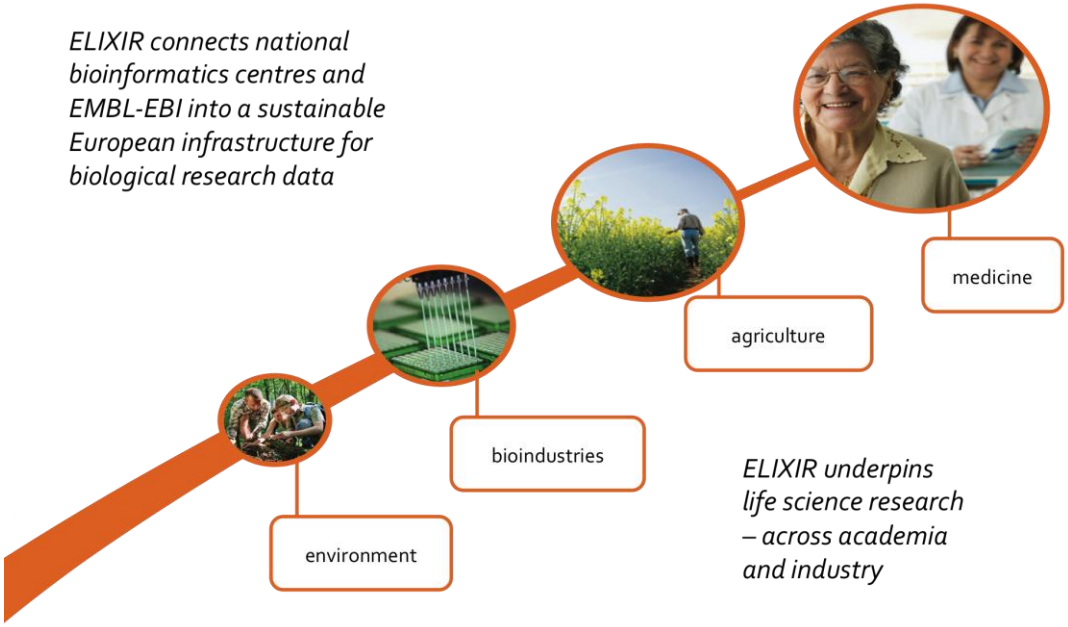
elixir ITALY

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.

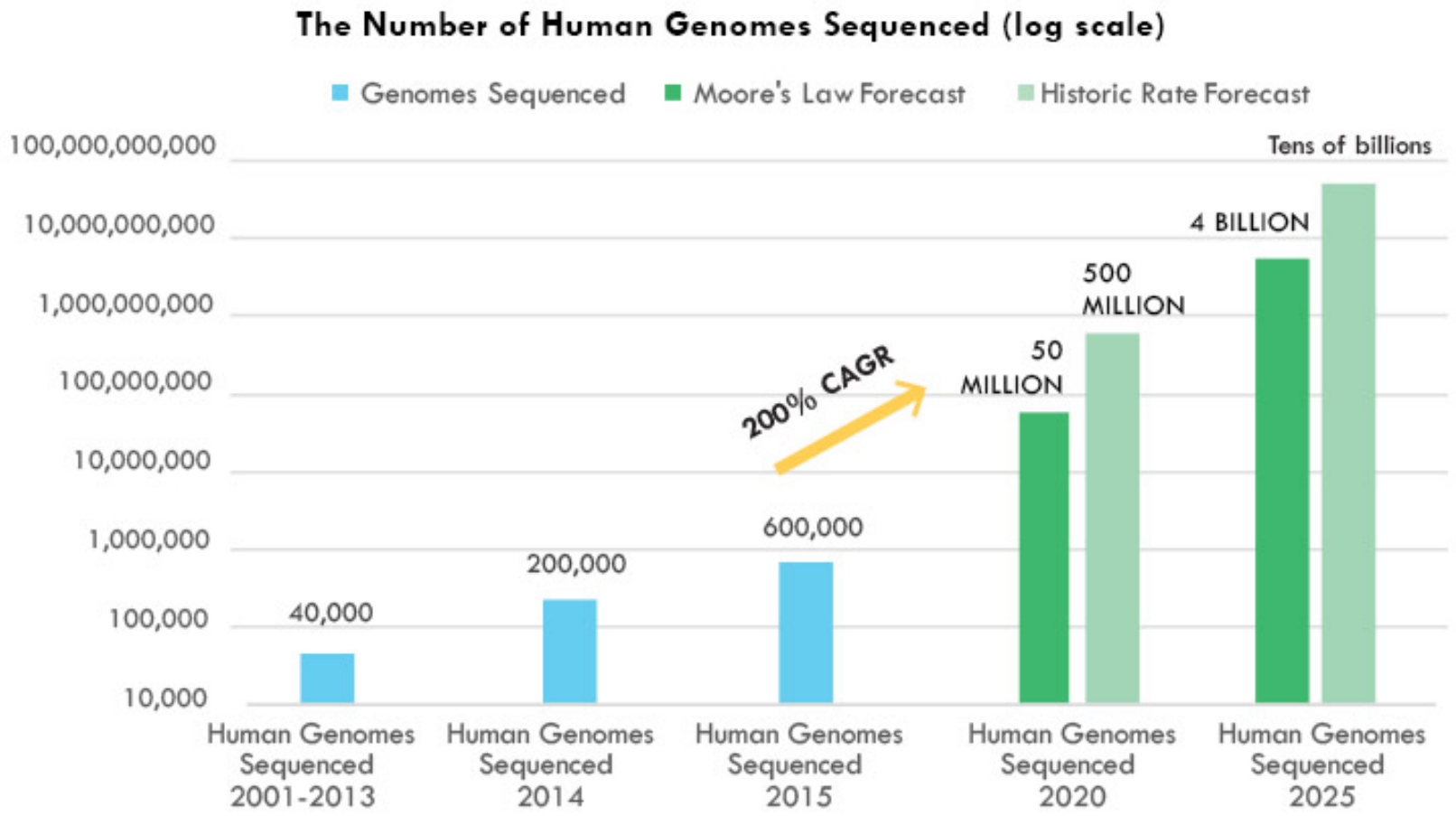
ELIXIR connects national bioinformatics centres and EMBL-EBI into a sustainable European infrastructure for biological research data



ELIXIR underpins life science research – across academia and industry

elixir ITALY

The data challenge: Data growth



Source: National Human Genome Research Institute (NHGRI), ARK Investment Management LLC

elixir ITALY

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

elixir ITALY

Diversity and fragmentation of data resources in life science

- Genomics Databases (non-vertebrate) (17.9%)
- Protein sequence databases (12.9%)
- Human Genes and Diseases (9.8%)
- Structure Databases (9.7%)
- Metabolic and Signaling Pathways (9.3%)
- Nucleotide Sequence Databases (8.8%)
- Human and other Vertebrate Genomes (7.1%)
- Plant databases (7.1%)
- RNA sequence databases (4.9%)
- Microarray and other Gene Expression Databases (4.5%)
- Other Molecular Biology Databases (3.3%)
- Immunological databases (1.8%)
- Organelle databases (1.6%)
- Proteomics Resources (1.2%)
- Cell biology (0.2%)

molecular biology data resources

http://www.oxfordjournals.org/nar/database/a/

~1700

elixir ITALY

An infrastructure of global significance

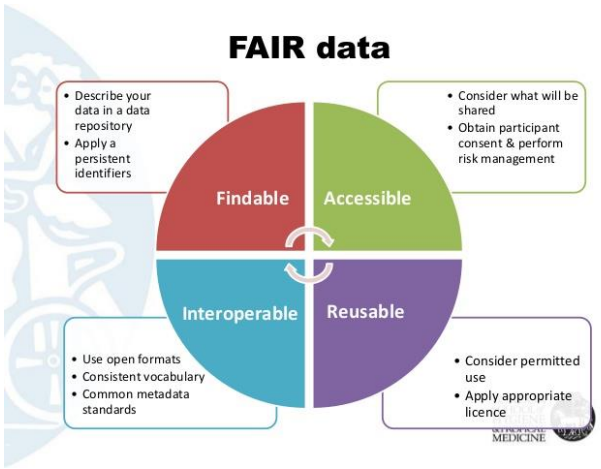
- ELIXIR put forward in G7 Group of Senior Officials report for 2015 on global research infrastructures
- 2016 ESFRI Roadmap classifies ELIXIR as a Landmark project
- Discussions initiated with Canada (Genome Canada) and Australia
- Collaboration with NIH-funded Big Data 2 Knowledge Initiative

G7 GERMANY 2015

elixir ITALY

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

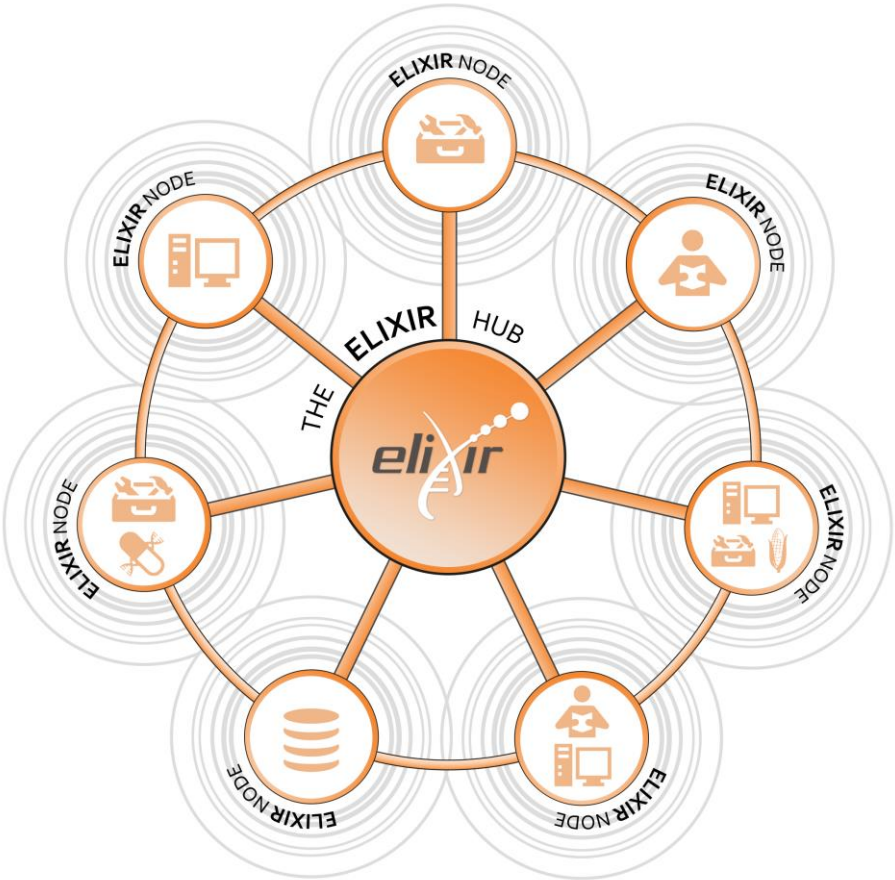
imi Innovative Medicines Initiative

wellcome trust

elixir ITALY

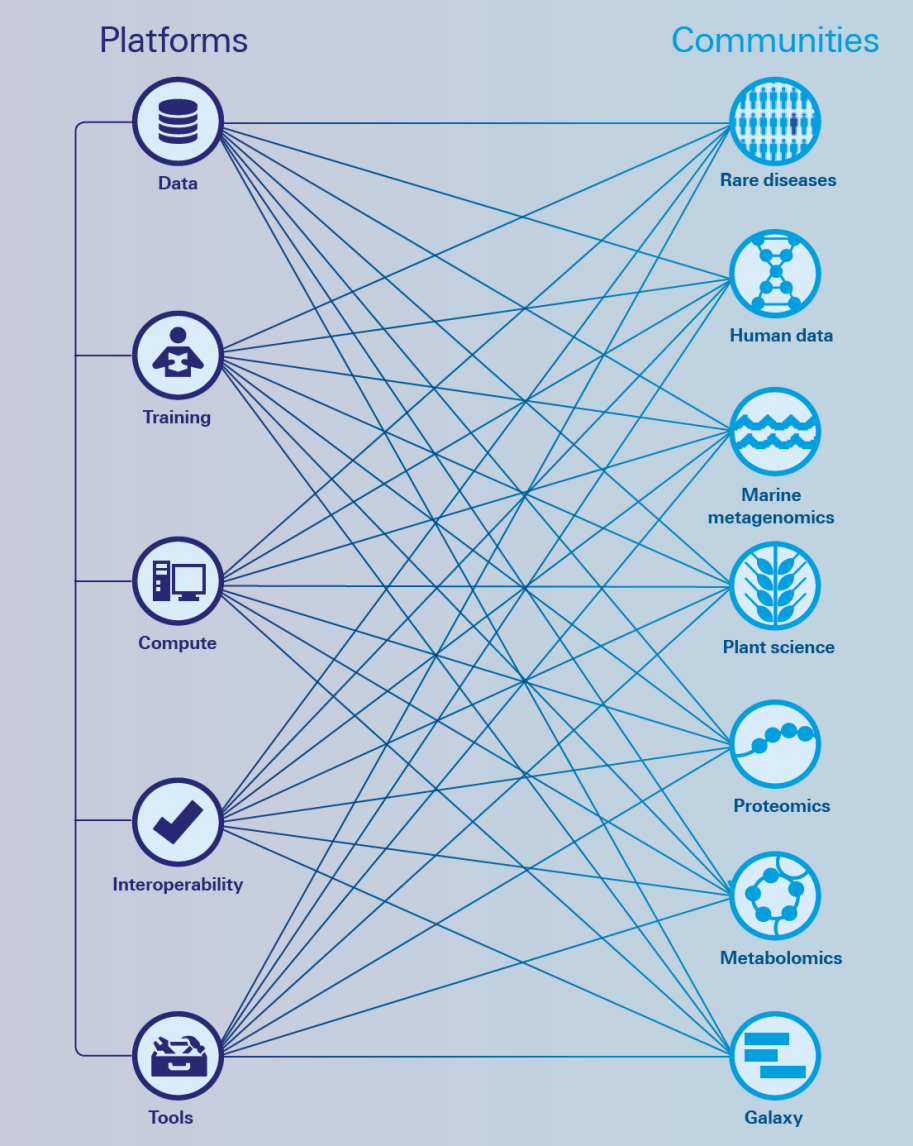
ELIXIR Organization

ELIXIR is organized as a Central Hub located in the Wellcome Trust Genome Campus in Hinxton - Cambridge (UK) and several national nodes build on national strengths and priorities



elixir ITALY

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 ITALY

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 ITALY

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)

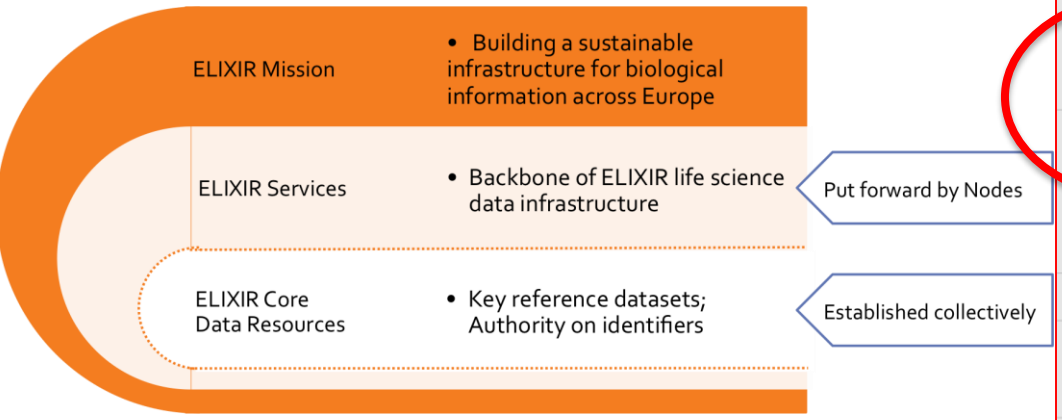
elixir ITALY

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:

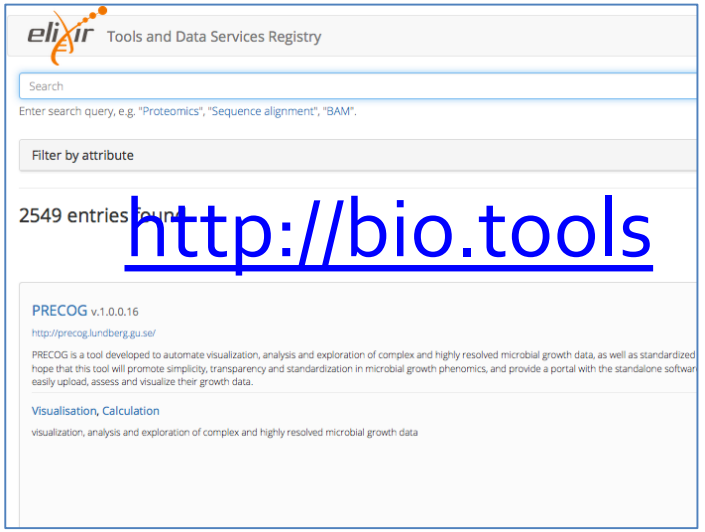
- Scientific focus and quality of science
- Community served by the resource
- Quality of service
- Legal and funding infrastructure, and governance
- Impact and translational stories

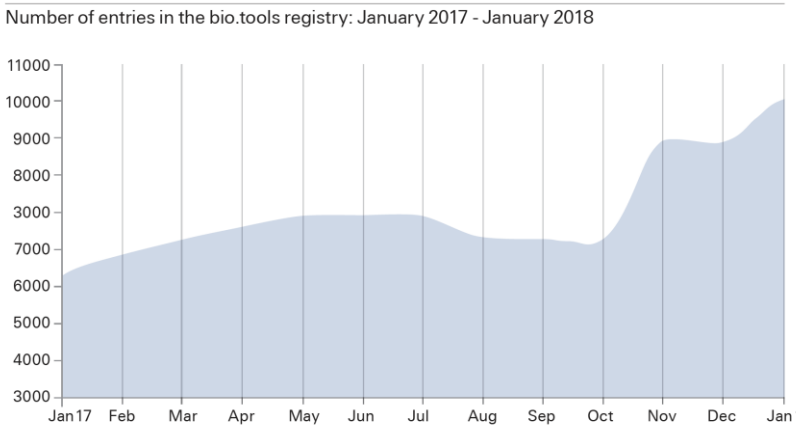


Core Data Resource	Data type
ArrayExpress	Functional Genomics Data from high-throughput functional genomics experiments.
CATH	A hierarchical domain classification of protein structures in the Protein Data Bank.
ChEBI	Dictionary of molecular entities focused on 'small' chemical compounds.
ChEMBL	Database 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.

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

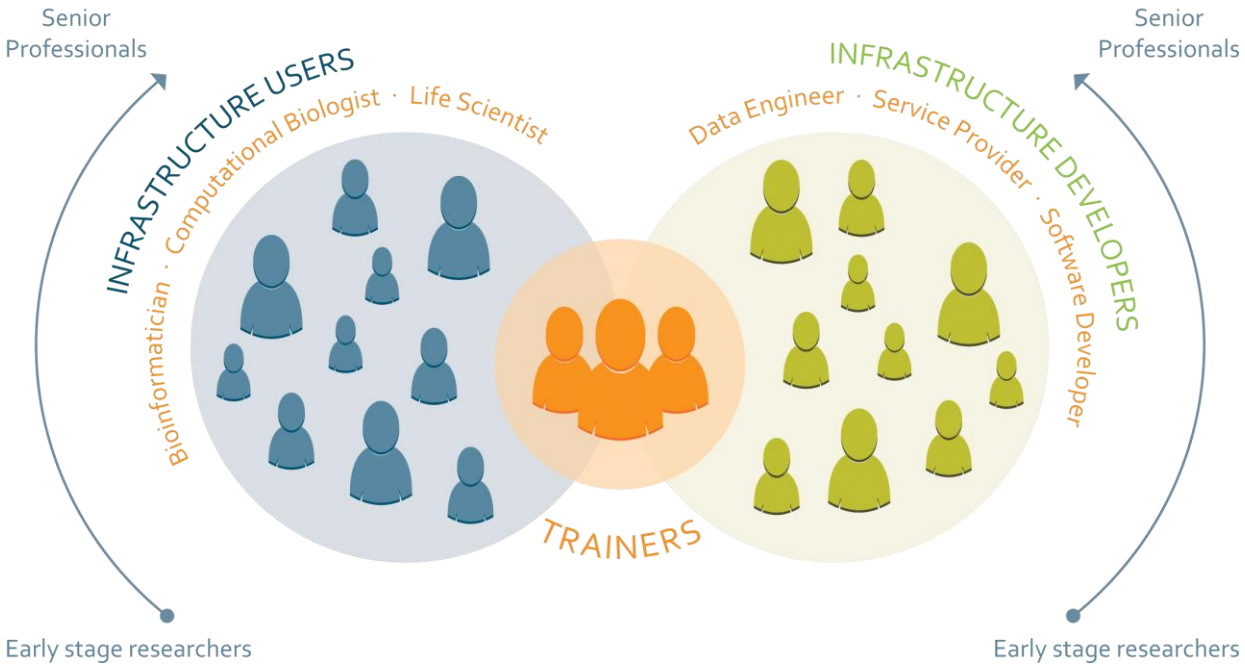


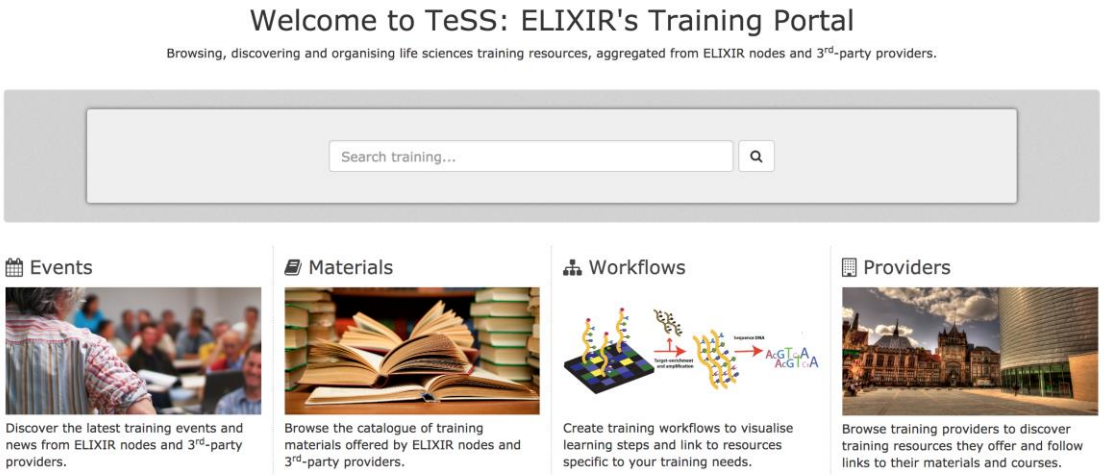


Tools and data services registry: a community effort to document bioinformatics resources.- NAR Jan 2016

elixir ITALY

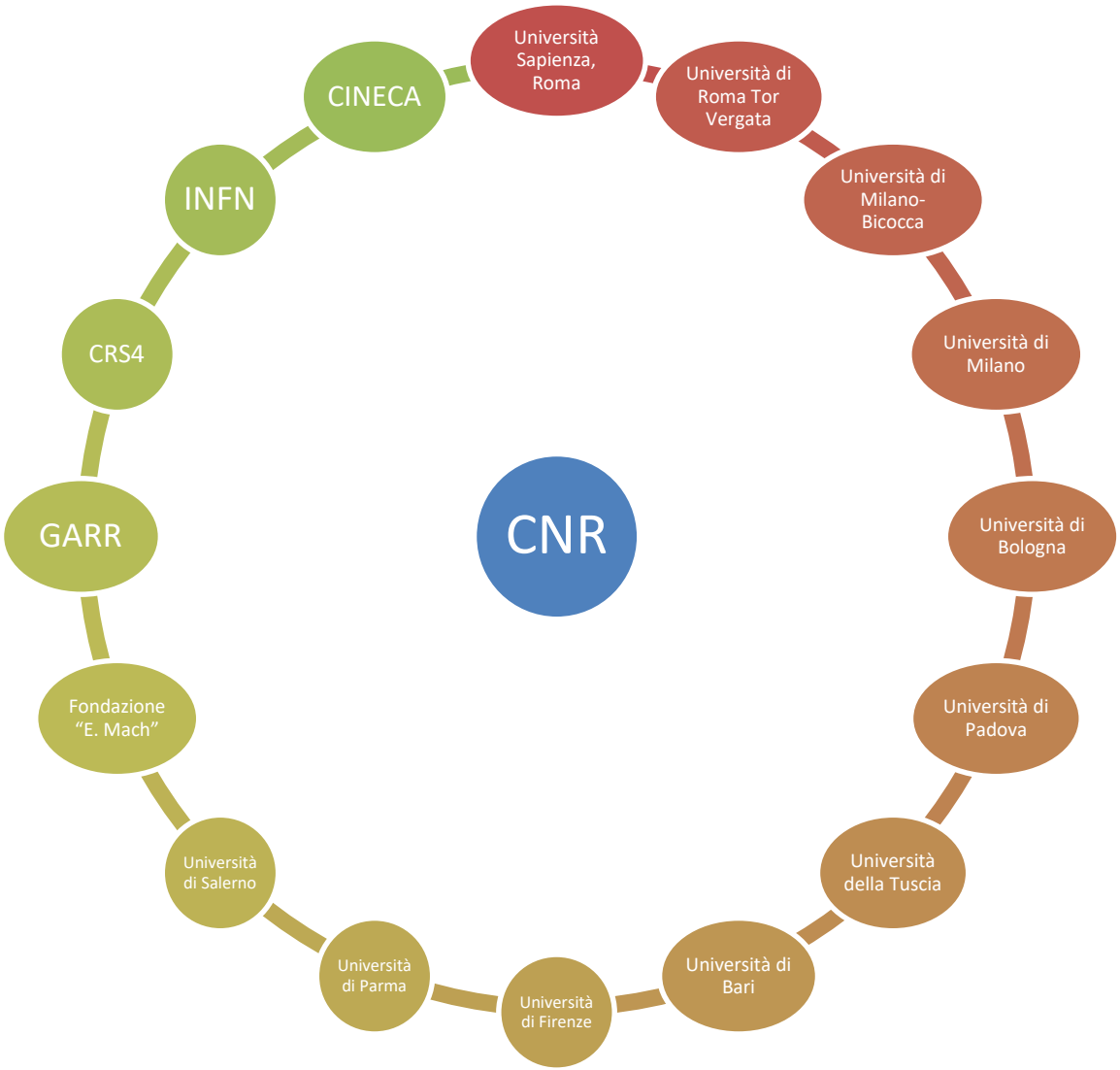
ELIXIR Training programme





elixir ITALY

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 16 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**).

A growing ELIXIR Node

Original Node Application (2012)

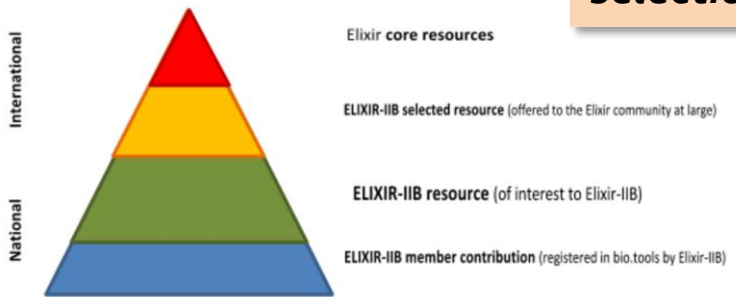
Updated Node Application (2018)

Six members joining, their application currently at ELIXIR-IT SAB. Procedure ends in December 2018.

Institution	Joined in
CNR (Lead)	2012
Sapienza Università di Roma	2012
Università di Roma Tor Vergata	2012
Università di Padova	2012
CINECA	2012
CRS4	2012
GARR	2012
Università di Milano	2013
Università di Milano-Bicocca	2013
Università di Bologna	2013
Università della Tuscia	2013
INFN	2013
Università di Bari	2014
Università di Firenze	2016
Università di Parma	2016
Università di Salerno	2016
Fondazione Edmund Mach	2016
New member (application sent and under review)	2019
New member (application sent and under review)	2019
New member (application sent and under review)	2019
New member (application sent and under review)	2019
New member (application sent and under review)	2019
New member (application sent and under review)	2019
Prospective New member (informal contacts)	2020?
Prospective New member (informal contacts)	2020?
Prospective New member (informal contacts)	2020?

Tools and Data Services Quality Management Policy

186 ELIXIR-IT services in bio.tools



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.

NA 2018

17 Selected DBs

26 Selected Tools

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

Infrastructural partners:

CINECA
INFN
University of Padua
CRS4
GARR
CNR

NA 2018

HPC

HTC

Cloud

Storage

ELIXIR-IT HPC@CINECA PROGRAM

Grants access to HPC resources for Italian/EU Life Science researchers.

*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.*

ITALY

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

ir TALY

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

excelerate

TtT Program

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

ELIXIR-IIB and Horizon 2020

Call: H2020-INFRADEV-1-2014-1

EMBRIC - European Marine Biological Research Infrastructure Cluster to promote the Blue Bioeconomy

Call: H2020-EINFRA-2014-2

- INtegrating Distributed data Infrastructures for Global ExplOitation

Call: H2020-INFRADEV-1-2015-1

ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life-sciences.

Call: H2020-INFRAEOSC-2018-2

EOSC-Life: Providing an open collaborative space for digital biology in Europe

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