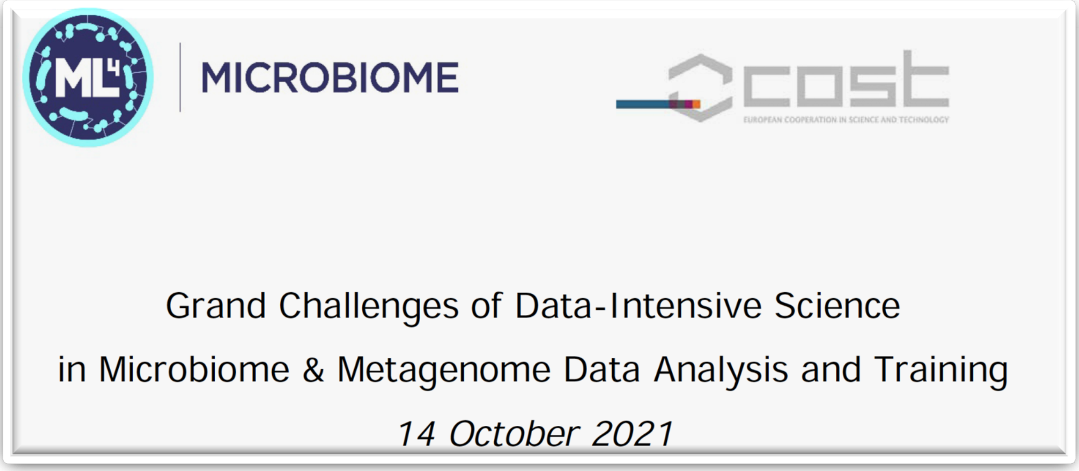
The ELIXIR CNR.BiOmics platform: an Italian infrastructure for BIG DATA production, analysis and training

IBIOM
Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari

Graziano Pesole,
ELIXIR-IIB Head of Node

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO



ML4 | MICROBIOME

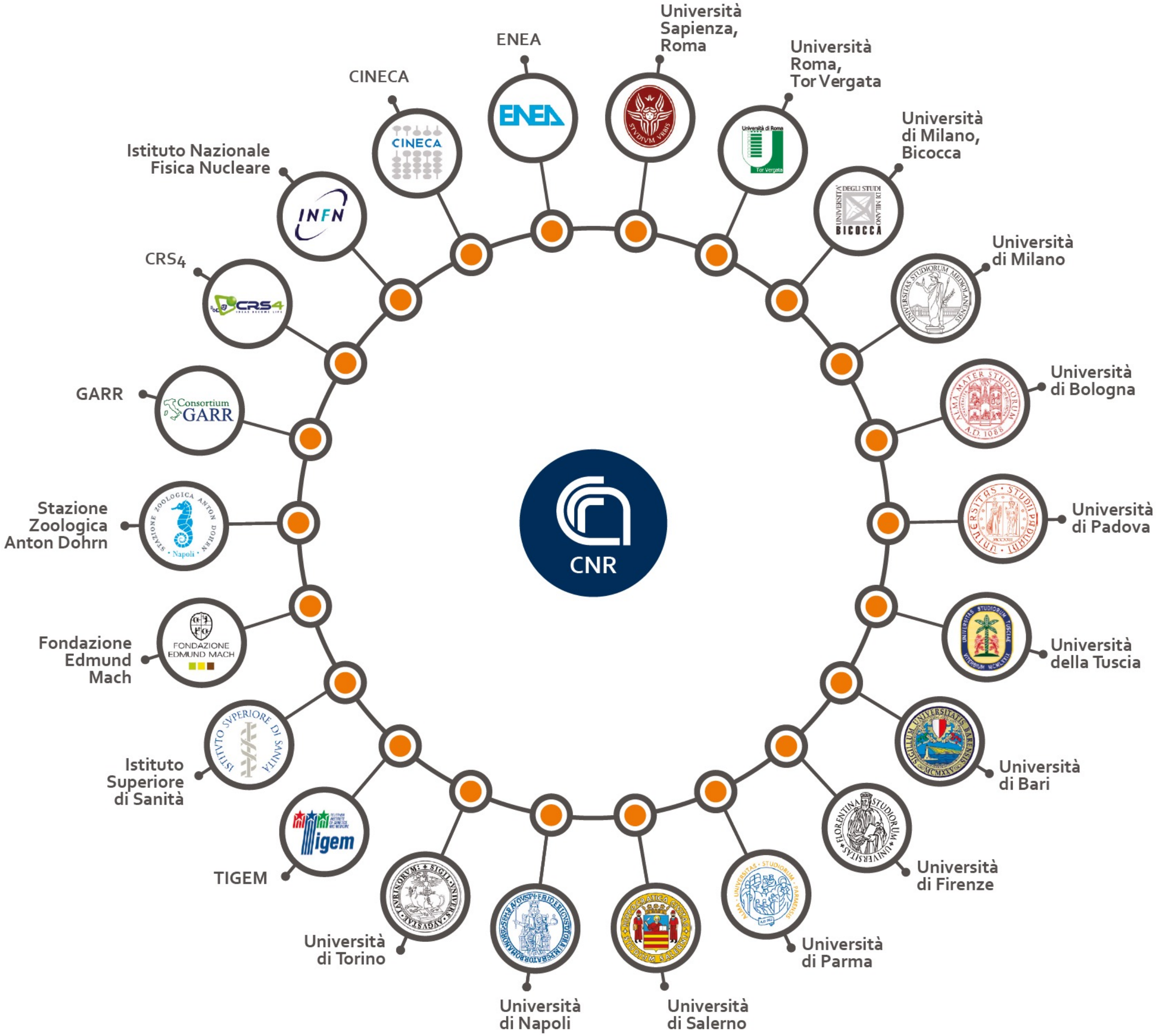
cost
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Grand Challenges of Data-Intensive Science

in Microbiome & Metagenome Data Analysis and Training

14 October 2021

ELIXIR-Italy: a distributed ELIXIR Node



The Italian node of ELIXIR (ELIXIR-IT) has been formally established as a **Joint Research Unit** (JRU) - named **ELIXIR-IIB (Infrastruttura Italiana di Bioinformatica) -** and is in charge of the coordination and delivery of existing bioinformatics services at the national level, all the while ensuring they integrate perfectly in the overall ELIXIR infrastructure.

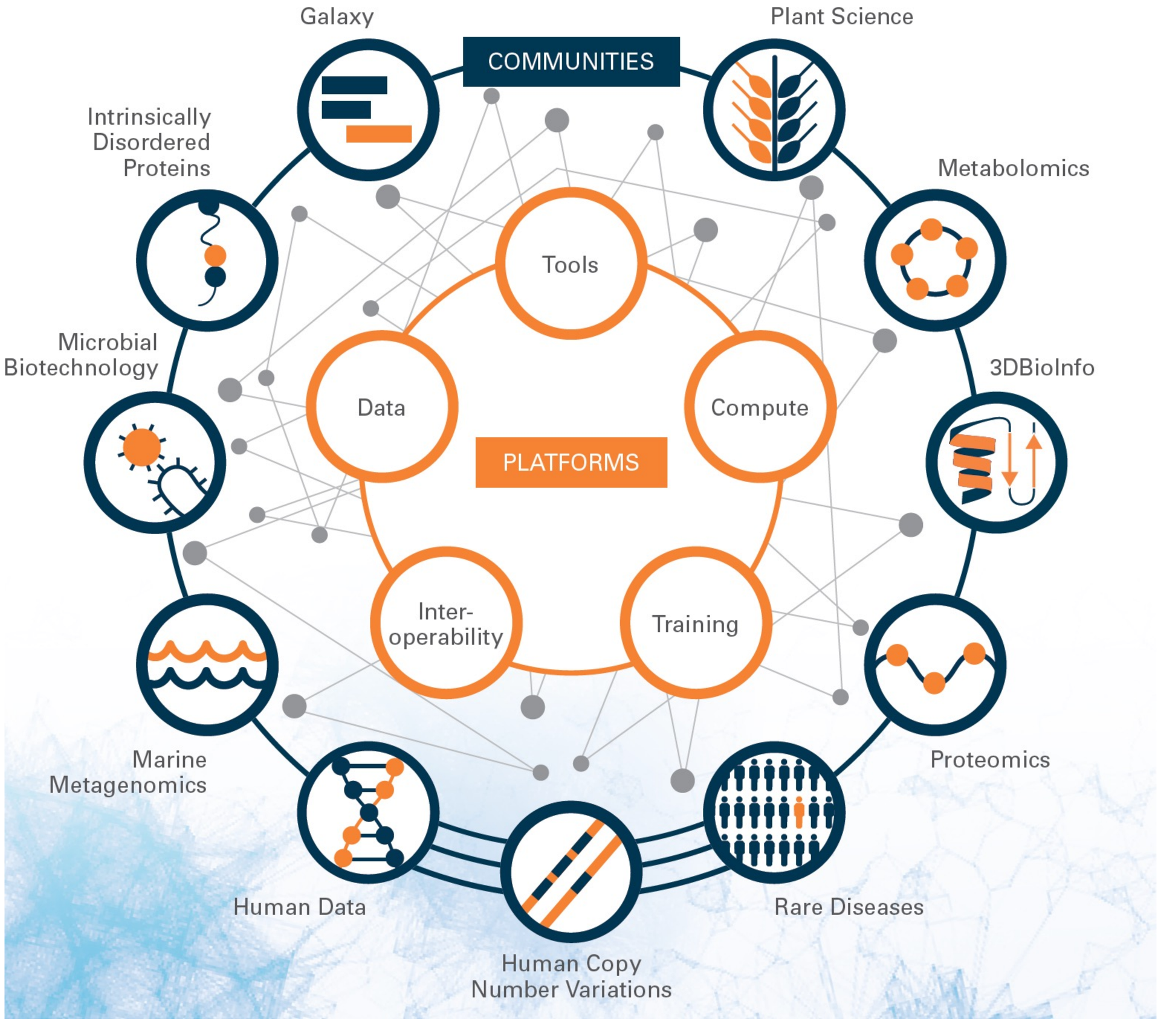
ELIXIR-IT is led by National Research Council (CNR) of Italy and currently involves **23 partners** including Universities and Research Institutions / Facilities of national relevance.

ELIXIR-IT Organisation

The national organization of ELIXIR-IT mirrors the organization of ELIXIR at European level.

ELIXIR-IT currently includes the five operational platforms (**Compute**, **Data**, **Tools**, **Interoperability** and **Training**) of ELIXIR Hub and several **Thematic communities**.

ELIXIR-IT platforms coordinate the delivery of high quality computational services for life science and drive the integration of national services within the ELIXIR infrastructure ecosystem.

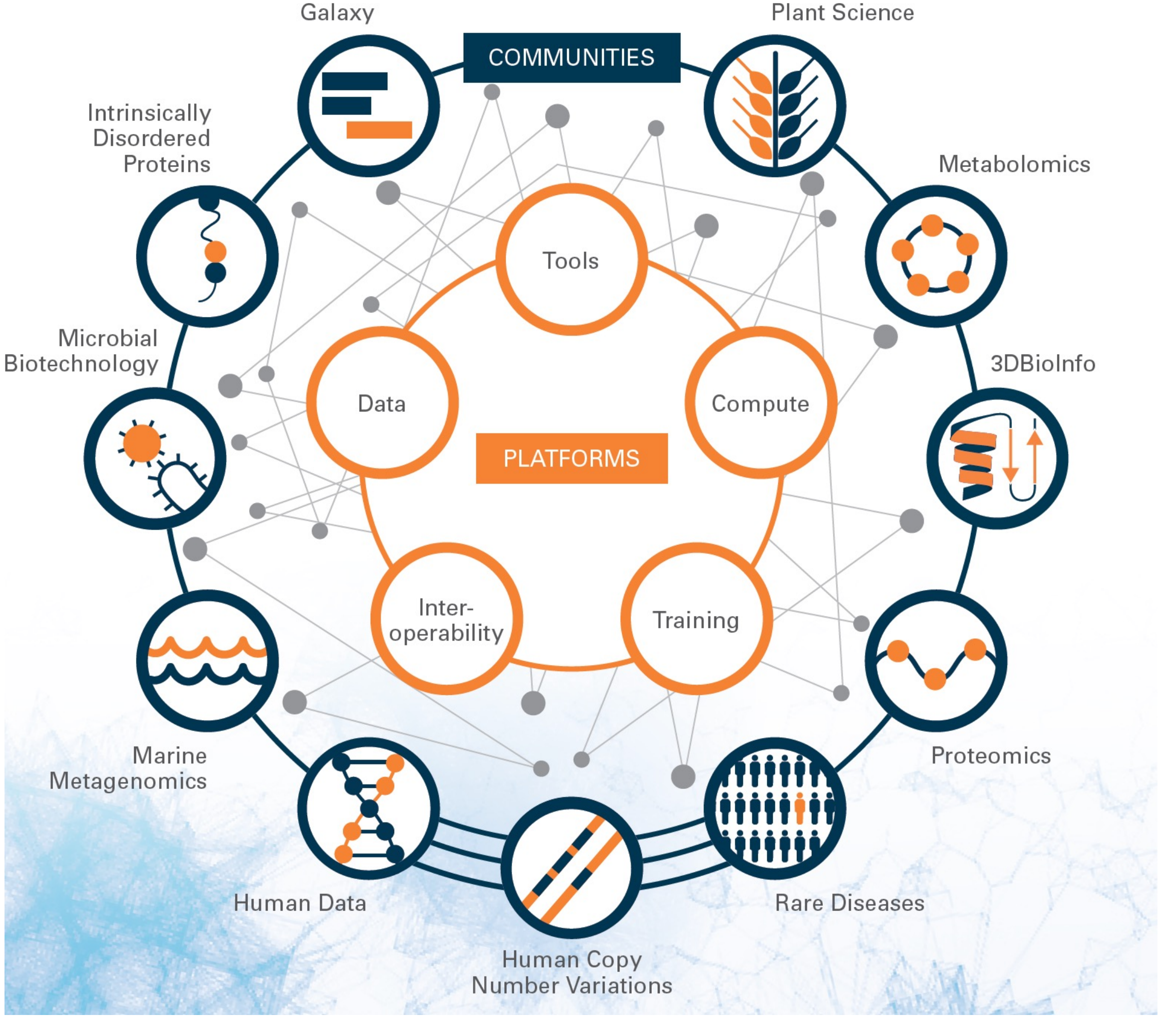


ELIXIR-IT Operations

ELIXIR-IT members contribute to the construction of the Infrastructure providing services, facilities, interoperability prescriptions, and training. All contributions should comply strict quality standards and form the **SDL** (**Service Delivery Plan**) of ELIXIR-IT which is shared in the ELIXIR Ecosystem.

Working groups have been established both at national and European level to contribute to Platforms and Communities operation and development.

ELIXIR-IT users are all interested researchers in public and private bodies. A large amount of services are free, but for some it is necessary to refund running costs.



CNR.BiOmics

BIG DATA FOR BETTER LIFE

Consiglio Nazionale delle Ricerche · Università degli Studi di Bari Aldo Moro · INFN · elixir ITALY

elixir ITALY

It is currently ongoing the **CNR.BiOmics project** ("National Research Center in Bioinformatics for Omics Sciences", PIR01_00017 14.5 M€ and CIR01_00017 2M€) which aims to strengthen the Italian node of the European Research Infrastructure ELIXIR in the southern regions.
Coordinator **Dr. Elisabetta Sbisà** https://www.cnr.it/it/pon-cnr-biomics

10X Genomics

Illumina NovaSeq 6000

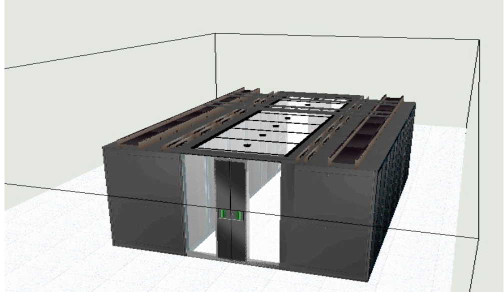
Oxford Nanopore GridION

PacBio Sequel

Thermo Scientific Orbitrap Fusion

Thermo Scientific LTQ XL™

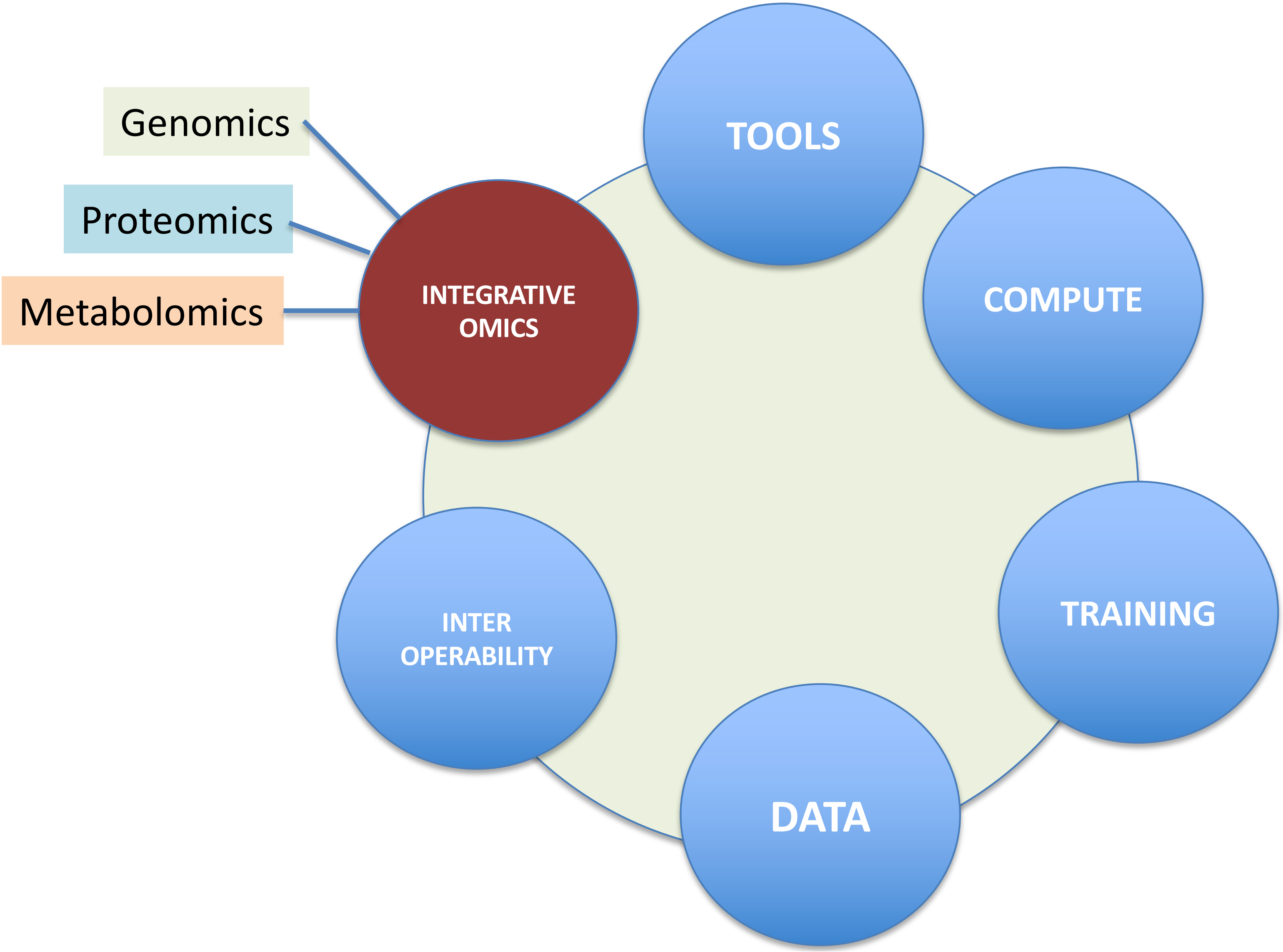
BioNano Genomics



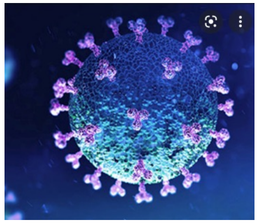
ICT Facility
10K core – 15 PB

ELIXIR-IT Platforms

The CNR.BiOmics project enabled the integration in ELIXIR-IT of a new Platform for “Integrative Omics” providing the user community with state of the art equipment for high-throughput generation of genomic, proteomic and metabolomic data.



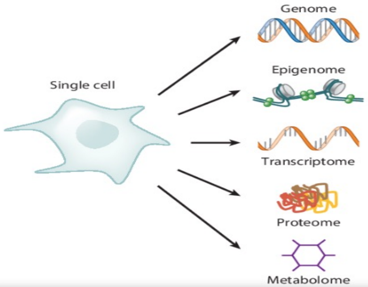
Genomics



Target Enrichment:
- Whole exome sequencing
- Pre-designed sequencing panels

Illumina NovaSeq 6000

10X Genomics

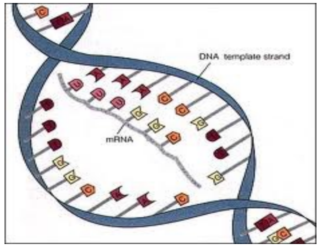


Single cell “omics”

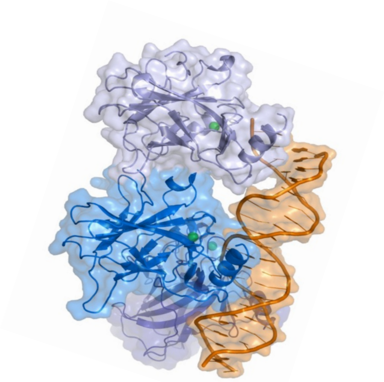
PacBio Sequel

BioNano Genomics

Oxford Nanopore GridION

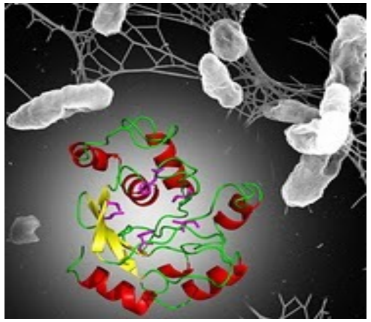


Transcriptome



DNA-Protein Interaction Analysis
- ChIp-seq
- ATAC-seq
- Spatial Organization of Chromatin
- Epigenome

Whole Genome Sequencing
- Genome assembly
- Structural variants

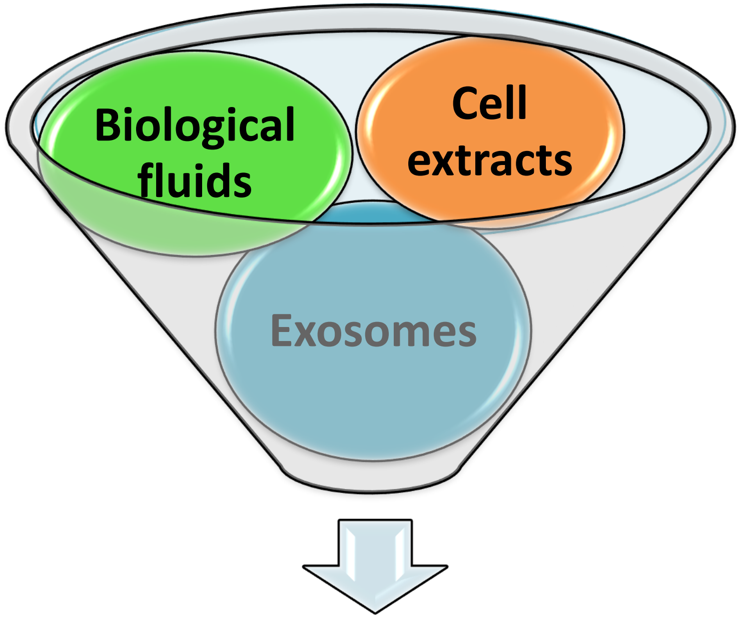


Microbiome
- Shotgun Metagenomic
- DNA Metabarcoding

CNR.BiOmics
BIG DATA FOR BETTER LIFE

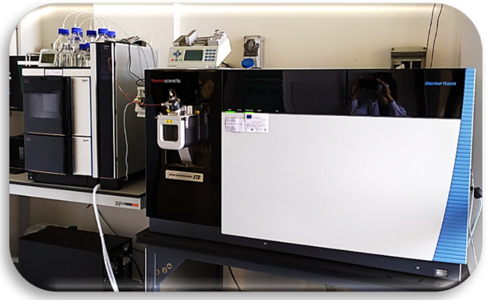
Metabolomics

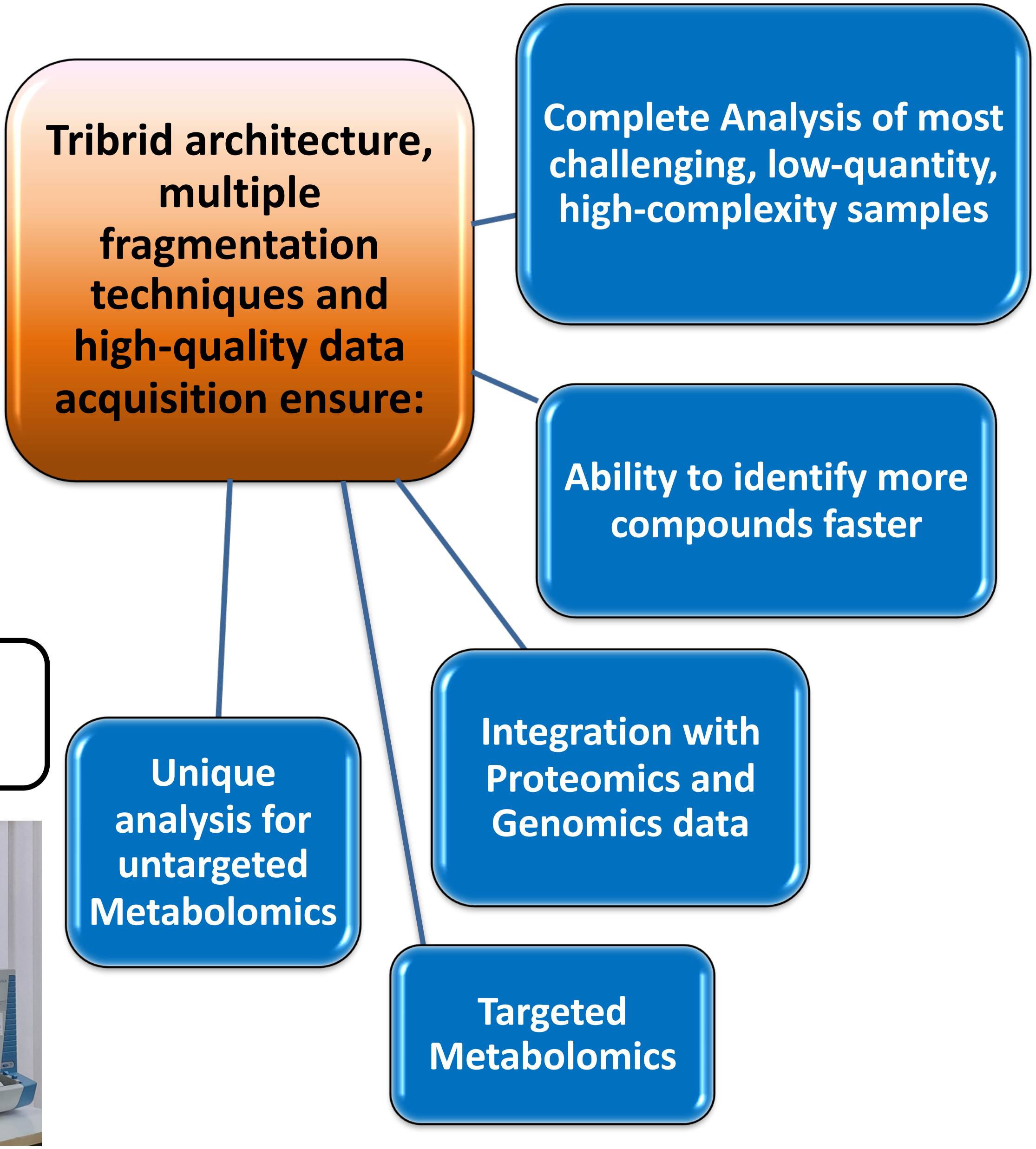
elixir ITALY



Thermo Scientific Orbitrap Fusion™

Thermo Scientific LTQ XL™



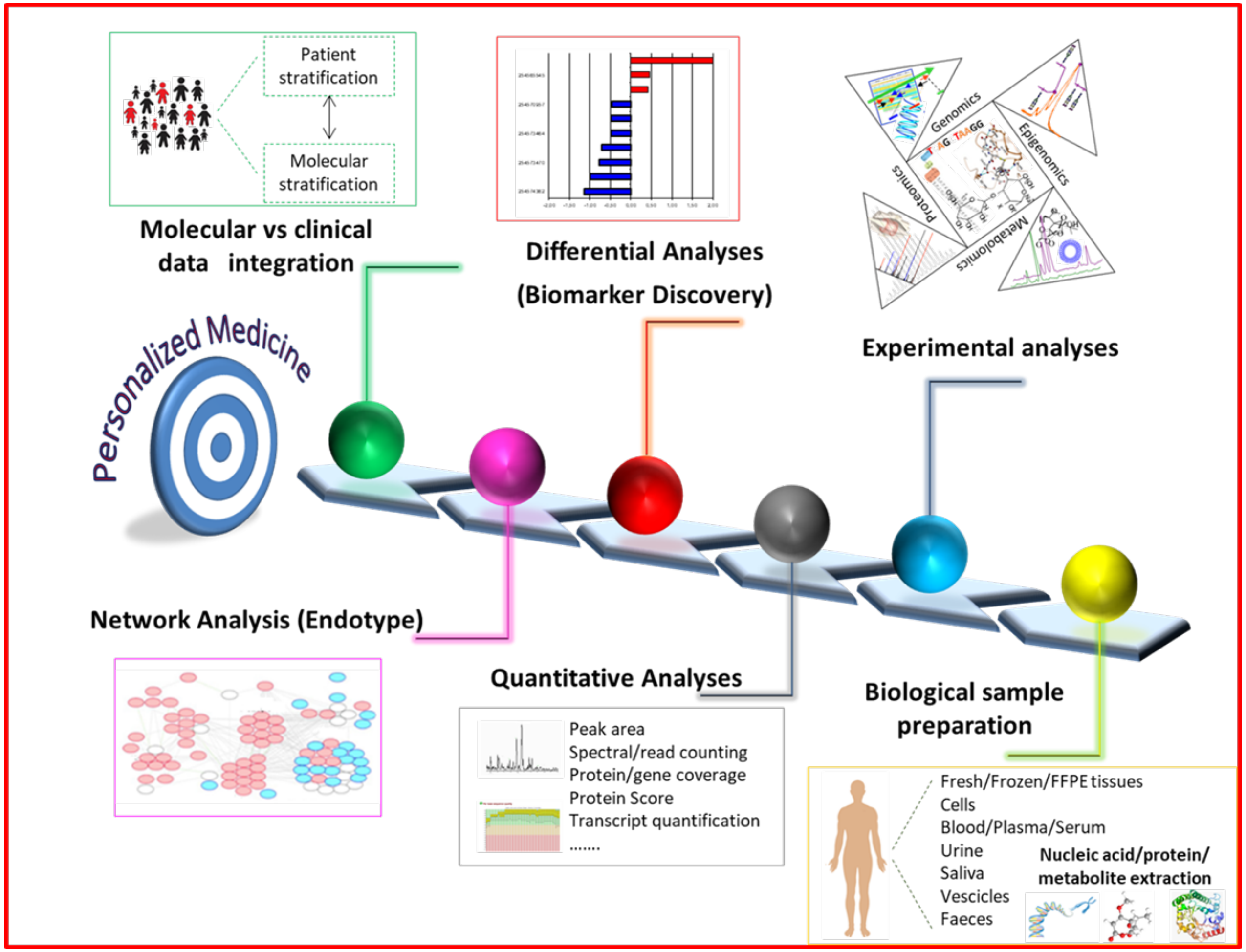


Tribrid architecture, multiple fragmentation techniques and high-quality data acquisition ensure:

- Complete Analysis of most challenging, low-quantity, high-complexity samples
- Ability to identify more compounds faster
- Integration with Proteomics and Genomics data
- Targeted Metabolomics
- Unique analysis for untargeted Metabolomics

Proteomics

The Proteomics Lab has been established as a "**data production unit**", integrated with computational expertise, to perform multidisciplinary and applicative projects ("**knowledge factories**"), in a wide range of applications, from human to plants and microbes.



Patient stratification

Molecular stratification

Molecular vs clinical data integration

Differential Analyses (Biomarker Discovery)

Experimental analyses

Network Analysis (Endotype)

Quantitative Analyses

Peak area
Spectral/read counting
Protein/gene coverage
Protein Score
Transcript quantification
....

Biological sample preparation

Fresh/Frozen/FFPE tissues
Cells
Blood/Plasma/Serum
Urine
Saliva
Vescicles
Faeces

Nucleic acid/protein/ metabolite extraction

Mass spectrometrer hybrid TripleTOF 6600+ coupled to nanoLC (Sciex/Eksigen)

Translational Research

Discovery of **biomarkers** useful for **diagnosis**, phenotyping and **therapy** monitoring

Basic Research

Elucidation of **molecular mechanisms** (**endotypes**) related to diseases and its therapies.

CNR.BiOmics
BIG DATA FOR BETTER LIFE

BioRepository

elixir ITALY

"…a LARGE repository for omics (BIG) data…"

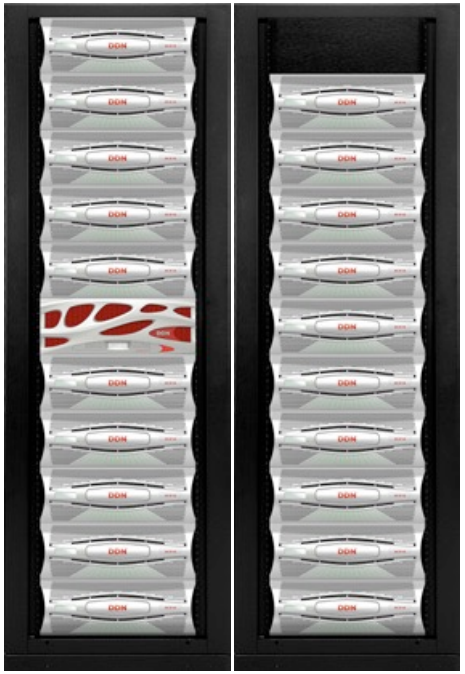
Applications

- Conservation and (mid/long-term) preservation of:
  - data and metadata produced by omics experiments
  - valuable public database
  - large datasets for bioinformatics analysis and AI applications
- Federation with research datasets to lead national and international collaborations
- Implementation of a Local Federated EGA node (European Genome-Phenome Archive)
- Secure data protection, conservation and controlled access of private and sensitive data (GDPR) conforming to Open Access and Open Science paradigms
- Geographically distributed to implement "Near Data Computing" and to guarantee efficient disaster recovery

Technical Solution

High-performance
Parallel Storage

15 PB
Bari – Milan - Naples

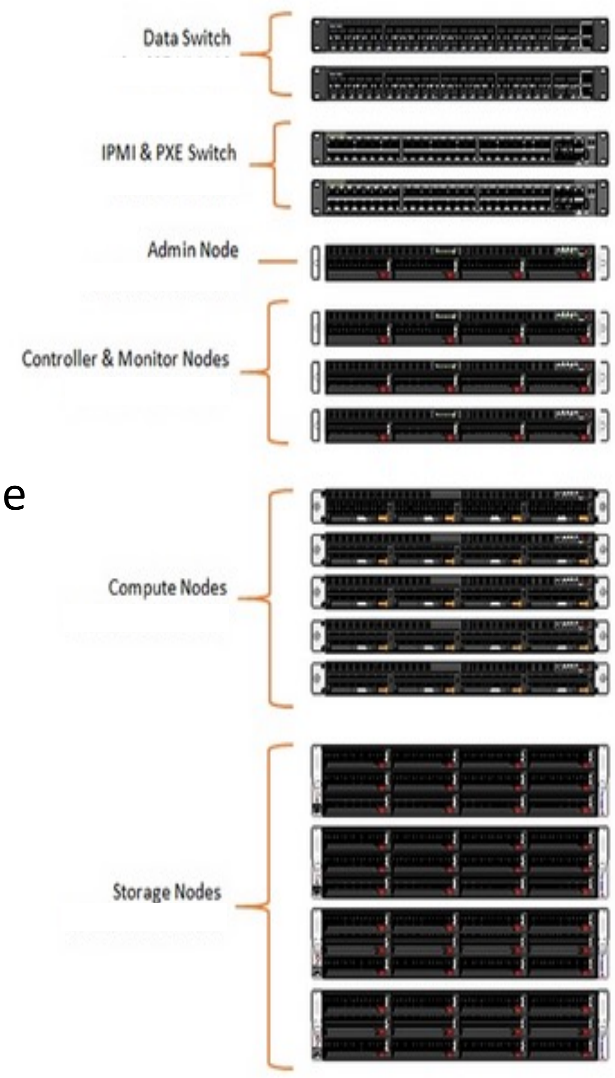


Contact:
flavio.licciulli@ba.itb.cnr.it

Cloud Computing

- A multi-institution, distributed and federated compute and storage infrastructure
- Providing most advanced technologies in the fields of Cloud and HPC
  - Able to support the latest Artificial Intelligence and Big Data analysis solutions
- Flexible and expandable in the future aiming to support future bioinformatics use cases
- The overall infrastructure will leverage on:
  - 27,000 Cpu/cores
  - About 20Pyte of disk storage
  - 20 NVIDIA V100 GPU
- Distributed over 4 different sites in Italy (CNR-Bari, INFN-Bari, CNR-Napoli, CNR-Milano)

OpenStack Cloud Platform



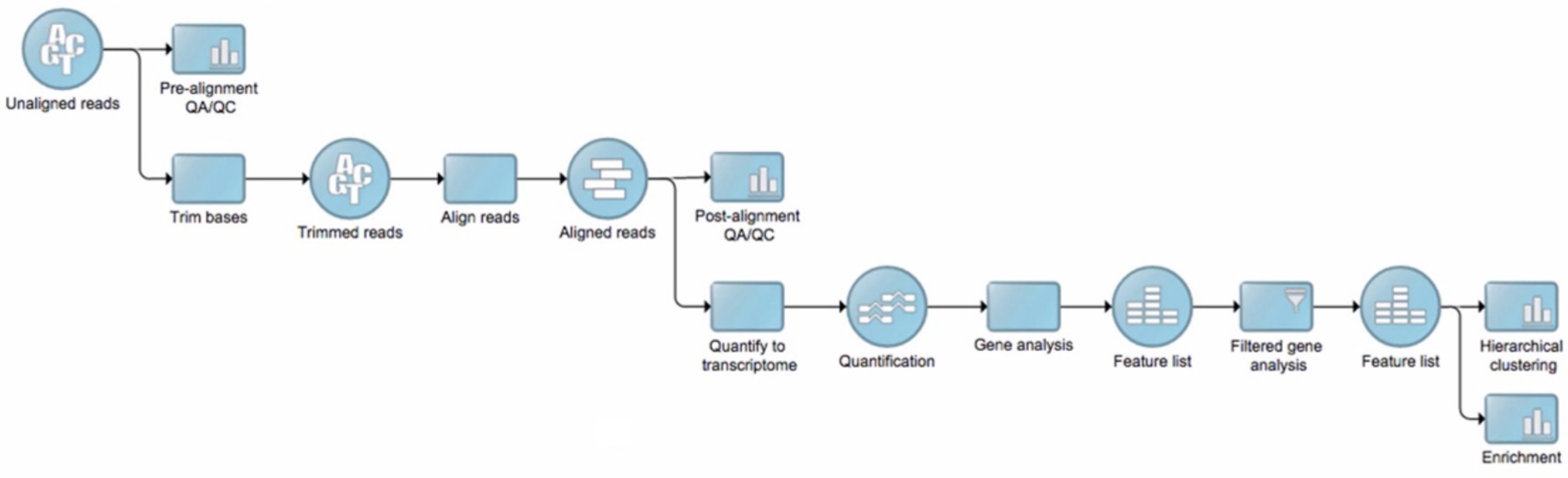
Infrastructure-as-a-Service (**IaaS**) private clouds

Contact: giacinto.donvito@ba.infn.it

Tools

The TOOLs platform will be strengthened by dedicated software:

- to automate the analysis of large NGS data (WGS, WES, RNAseq and so on) through well established pipelines;
- to facilitate the analysis of proteomic data;
- to improve the functional interpretation by the pathways analysis.



Training

- **Infrastructural investment:**
  - Training room equipment (instructor's workstation, learners' laptops, projector, printer)
  - Equipment for the production of virtual lessons/courses

- **Applications**

- Deliver training courses where needed
- Design and build the ELIXIR-IT eLearning Platform (it will be running on the ReCAs servers)
- Use the technology provided by the PON to design courses according to principles of effective learning
- Offer a portfolio of opportunities to learn bioinformatics online

ELIXIR-Italy: Access Program

ELIXIR-IT already provides a rich portfolio of computational services through its technological partners (e.g. HPC@CINECA, LANIAKEA, etc.) usually for free. The completion of the CNR.BiOmics project, with the establishment of the new "Integrative Omics" platform for data generation requires the establishment of a more structured access program, thus making possible the full exploitation of the infrastructural facilities. We plan to start by September 2021, following the model below.

1. User apply for a service filling a web form exposing the "services portfolio" including all platform services (e.g. Compute, Integrative Omics, etc.)
2. The request is evaluated for feasibility and scientific appropriateness by the specific Platform committee
3. The request approval is notified to the user and the service delivery may start (for free or under the running costs refunding, depending on the service)
4. A suitable platform team will assist users in services delivery (e.g. for the Integrative Omics platform will collaborate in the final design of the experiment and give useful indications for the preparation of the samples.

Credits

Genomics

Apollonia Tullo
Anna Maria D’Erchia
Carmela Gissi
Flaviana Marzano
Mariano Caratozzolo
Caterina Manzari

Compute

Giacinto Donvito
Flavio Licciulli
Marco Tangaro
Roberto Bellotti

Proteomics

Pierluigi Mauri
Dario Di Silvestre

Tools

Ernesto Picardi
Giorgio Grillo

Metabolomics

Sergio Giannattasio
Clara Musicco
Giuseppe Petrosillo
Bruno Fosso
Fabrizio Mastrorocco
Angelo Facchiano
Virginia Carbone

Training

Allegra Via
Francesca De Leo
Domenica D’Elia

Direction and Administration

Elisabetta Sbisà
Laura Marra

CNR referents

Cabina di regia
Dep. Biomedical Science

ELIXIR IIB Contacts

ELIXIR	https://www.elixir-europe.org/
ELIXIR-Italy	http://elixir-italy.org/

Head of Node	g.pesole@ibiom.cnr.it
TeC	federico.zambelli@unimi.it
Node Coordinator	rita.casadio@unibo.it

PLATFORMS

• Compute	giacinto.donvito@ba.infn.it
• Data	silvio.tosatto@unipd.it
• Training	allegra.via@gmail.com
• Tools	giulio.pavesi@unimi.it
• Interoperability	pierluigi.martelli@unibo.it
• Integrative Omics	elisabetta.sbisa@ba.itb.cnr.it
Genomics	a.tullo@ibiom.cnr.it
Proteomics	pierluigi.mauri@itb.cnr.it
Metabolomics	s.giannattasio@ibiom.cnr.it

elixir
ITALY
DATA FOR LIFE

THANK
YOU
return