DOI: 10.26314/GARR-Conf25-proceedings-03

BioRepository@ELIXIR-IT: a computational environment for storing and sharing human genetic data

Claudio Lo Giudice[1*], Giorgia Miniello[1*], Guido Cauli[1*], Francesco Rubino[8], Gianluca Cecinato[1], Marco Moscatelli[7], Sharon N. Cox[2], Nadina Foggetti[3], Francesca De Leo[3], Angelo S. Varvara[2], Bruno Fosso[2], Ermes Filomena[2], Pietro D'Addabbo[2], Marco A. Tangaro[3], Roberto Cilli[4], Giacinto Donvito[6], Federico Zambelli[3,5], Ernesto Picardi[2,3], Flavio Licciulli[1], Graziano Pesole[2,3]

[1]Institute of Biomedical Technologies, National Research Council, 70126 Bari, Italy, [2]Department of Biosciences, Biotechnology and Environment, University of Bari A. Moro, 70126 Bari, Italy [3]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, 70126 Bari, Italy, [4]Department of Physics, University of Bari A. Moro, 70126 Bari, Italy, [5]Department of Biosciences, University of Milan, 20133 Milan, Italy, [6]National Institute for Nuclear Physics (INFN), 70126 Bari, Italy, [7]Research Area Milan 4, National Research Council, 20054 Segrate, Italy, [8]Ruder Boskovic Institute, Department of Medicine, Bijenička cesta 54, 10000 Zagreb

(*) Contributed equally to the work and are recognized as co-first authors

Abstract. Nucleic acid sequencing is becoming more accessible, opening doors for new healthcare applications like precision medicine and pharmacogenomics. These could greatly improve treatments for conditions such as cancer and genetic diseases. However, to make the most of this, we need to address complex technical, legal, and ethical issues, regarding data management. This paper introduces BioRepository, a new integrated service by ELIXIR-IT. BioRepository is designed to manage human genetic data from its collection to its storage, supporting the use of genetic information in research and healthcare

Keywords. Genomic Data, Data Management, FAIR Data principles, Elixir, Secure Data Access

1. Introduction

Biorepositories can be defined as structured services designed to collect, store, manage, and distribute biological specimens associated data and metadata for research and clinical applications. These repositories play a pivotal role in biomedical research, enabling large-scale studies in genomics and precision medicine. By collecting high-quality omics sampling data — derived from genomics and proteomics analysis and sequencing — and linking them to relevant metadata, biorepositories can promote reproducibility, interope-

rability, and data sharing.
With the widespread adoption of cost-effective sequencing technologies, volume and complexity of genetic data have increased dramatically. This trend necessitates computing infrastructures capable of ensuring secure storage, controlled access, traceability, and compliance with regulatory frameworks and laws. The sensitive nature of genetic data, in particular, requires robust mechanisms for authentication, encryption, and enforcement of access policies.

2. BioRepository@ELIXIR-IT Service Overview

In order to meet these demands, ELIXIR-IT (1) has developed an integrated service for the management of human genetic data, encompassing the entire data lifecycle from sequencing to deposition. These services are built on a computational environment based on virtual machines (VM) infrastructure, resulting in a secure, scalable and user-friendly environment that can be tailored to the needs of researchers and clinicians. The system leverages the ReCaS (2) data center in Bari (Italy), part of the Italian Computing and Data Infrastructure (ICDI) (3) and the European Open Science Cloud (EOSC) (4). The platform adheres to the FAIR principles (Findable, Accessible, Interoperable, and Reusable) and is compliant with the European General Data Protection Regulation (GDPR) (5).

3. Core Requirements

To fully support genomic data management, the BioRepository service addresses several technical needs that are critical for secure and reliable operation. These include robust data security, scalable infrastructure, and the ability to support high-quality data processing pipelines within a reproducible and transparent framework.

Data Security

Sequenced data are encrypted and digitally signed using the CRYPT4GH (6) encryption suite, while all data transfers are secured via asymmetric-encrypted SSHv2 tunnels (using public and private keys in ED25519 format), ensuring that only authorized recipients can access the data. Digital signatures verify both data integrity and source authenticity. The platform ensures that private keys remain within a secure internal environment and that all access events are logged and auditable.

Scalability

The infrastructure is designed to support projects ranging from small cohort studies to large-scale national initiatives. It uses a high-capacity redundant Parallel Storage System (DELL Isilon – PowerScale), OpenStack (7), and Proxmox Virtual Environment (8) to manage virtualized environments. An underlying Ceph infrastructure provides distributed, fault-tolerant storage across multiple nodes, supporting high availability and elastic scalability.

Processing Quality

The infrastructure ensures that all components, from data uploading to final storage support processing reliability, traceability, and compliance with quality standards.

4. System Architecture

The infrastructure consists of two main components detailed in Fig.1. Each component is designed to address the challenges of handling sensitive data while ensuring global accessibility and full compliance with data protection regulations such as the GDPR. The key parts are represented by a secure BioRepository and a virtualized analysis environment. The BioRepository, hosted at CNR-ReCaS in Bari, offers 5 PB raw space for encrypted storage with geo-redundant backups at CNR-ITB (Milan, Italy) and CNR-ICAR (Naples, Italy). Both external (such as uploads of raw omics data from sequencer workstations and download of processed data to the recipient) and internal data transfers are protected via SSH tunnels with asymmetric keys encryption. The repository also hosts curated, versioned reference datasets which can be essential for bioinformatics pipelines.

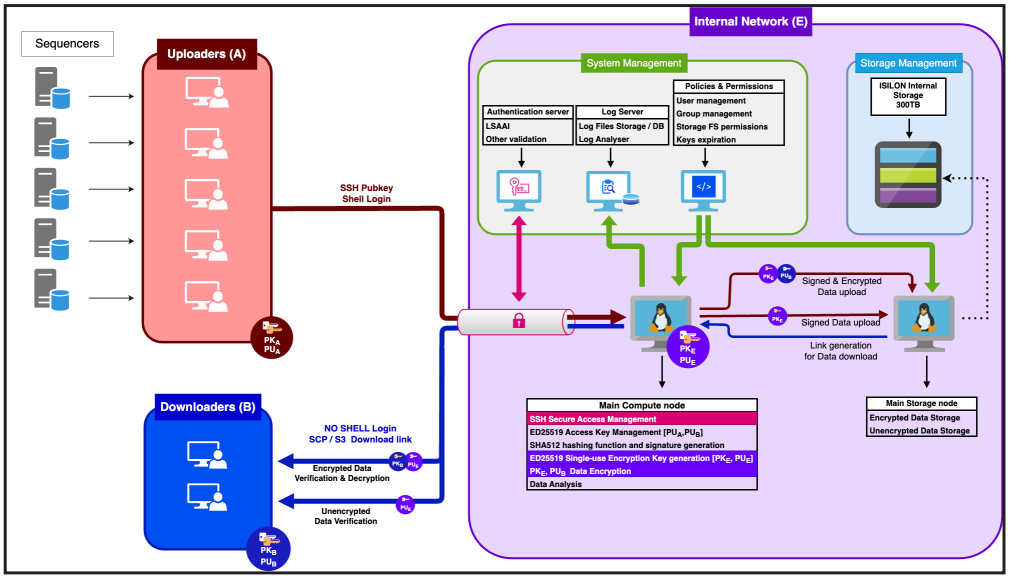


Fig. 1

BioRepository@ELIXIR-IT system architecture: visual representation of the secure data workflow across operational roles. Uploaders (A) send raw sequencing data through encrypted channels; the Internal Network (E) manages key generation, encryption, signing, logging, and storage; Downloaders (B) retrieve processed or raw data with integrity and authenticity validation. The infrastructure integrates policy enforcement, user management, and scalable encrypted storage

5. Secure Data Workflow

The system defines three operational roles: Uploaders (A), Downloaders (B), and the Internal Network (E).

- Uploaders submit raw data obtained from sequencers.
- Downloaders access processed data upon request.
- The Internal Network handles encryption, processing, and secure storage.

For each operational project, data workflow phases include:

1. Preparation: all users involved in an operational project must upload their ED25519

public keys to the system. Meanwhile, a CRYPT4GH public-private key pair PU(E) and PK(E) is created into the internal network for that operational project by the system administrator.

2. Upload: an encrypted transfer tunnel is established from the uploading workstation to the system network gateway. Data is transferred via SSH and then verified using SHA512 hash fingerprints.
3. Optionally, uploaded data can be processed using internal computing facilities, according to the specific agreement upon Institutes.
4. Hosting: sensitive data are encrypted using the download recipient public key PU(B), and digitally signed with a private key PK(E) generated into the system and unique for any different operational project, in order to grant data authenticity.
5. Download: authorized Downloaders acquire the public key PK(E) of the operational project, then they retrieve the processed data using a SSH secure tunnel. In the case of sensitive data, Downloaders can decrypt them using their own private key PK(B), while data integrity and authenticity is verified using the internal network Public Key PU(E). For non-sensitive data, the decryption part is skipped while integrity and authenticity can be verified using PU(E) as said.

Each processed dataset receives a unique identifier to support version control and traceability. All user actions are logged, and the logs are securely retained for auditing purposes. Once the download operation is complete or the intended hosting period of the operational project ceases, data and associated internal key pairs are securely deleted.

6. Infrastructure and Resources

While originally based on proprietary VMware ESXi [9], the Biorepository infrastructure currently relies on Proxmox VE, an open-source virtualization platform based on KVM/QEMU. Ceph [10] integration allows for efficient, resilient storage using OSDs, MONs, CRUSH maps, and Logical Volume Management (LVM). The system supports live migration of VMs, resource-aware scheduling, and automated failover.

7. Virtualized Analysis Environment

This environment also supports the deployment of customized, scalable bioinformatics pipelines executed on tailored VMs. VMs utilize encrypted filesystems and provide shared access to the BioRepository. High-availability configurations ensure minimal downtime, while snapshots and backup mechanisms can grant recovery and auditability of critical services.

Each VM is provisioned with:

- Up to 20 vCPUs (Intel Xeon E5/E7 with AVX-512)
- 200 GB DDR4 ECC RAM
- 500 GB local scratch storage
- 30 TB shared storage (NFS/iSCSI)

VMs can be equipped with one or two NVIDIA A100 GPUs to enable accelerated execution

of compute-intensive workflows such as Parabricks, DeepVariant, and Guppy.
Analytical environments can be replicated to support collaboration and benchmarking.

8. Software and Workflow Management

The platform supports containerized execution and environment management using:

- LXC and Docker for container isolation
- Conda and Mamba for dependency resolution and environment replication

To ensure processing consistency and traceability, bioinformatics workflows are containerized, version-controlled, and subject to rigorous quality assurance protocols. All analytical pipelines adhere to standardized sequencing data formats (e.g., FASTQ, BAM, VCF), and utilize metadata schemas aligned with the FAIR principles.
Audit trails and validation checks are embedded throughout the execution environment, ensuring reproducibility and transparency across different analyses. Containers are maintained in secure internal registries with automatic version tracking and security validation.

9. Conclusions

The BioRepository@ELIXIR-IT platform offers a secure, scalable, and standards-compliant solution for the full lifecycle management of human genetic data. By integrating open-source technologies such as Proxmox VE, Ceph, and CRYPT4GH, and by enforcing strict data protection policies, the platform ensures high levels of performance, interoperability, and trust.
Its modular and containerized architecture also supports a wide range of bioinformatics workflows, while GPU acceleration, live migration, and automated backups enhance computational efficiency and resilience. This platform serves as a forward-looking model for infrastructures supporting precision medicine, collaborative genomic research, and ethical data sharing in biomedical science.

Acknowledgments

The BioRepository service has been fully established thanks to funding from CNR.BiOmics — "National Research Center in Bioinformatics for Omics Sciences" (PON R&I 2014-2020, PIR01_00017) and PNRR ELIXIRxNextGenIT — “ELIXIR x NextGenerationIT: Consolidation of the Italian Infrastructure for Omics Data and Bioinformatics” (IR0000010).
Angelo Sante Varvara is a PhD student within the European School of Molecular Medicine (SEMM).

References

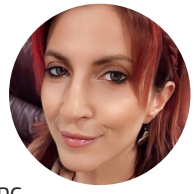
[1] https://elixir-italy.org/
[2] https://www.recas-bari.it/
[3] https://www.icdi.it/it/
[4] https://eosc.eu/
[5] https://gdpr-info.eu/

[6] https://crypt4gh.readthedocs.io/en/latest/
[7] https://www.openstack.org/
[8] https://www.proxmox.com/en/products/proxmox-virtual-environment/overview
[9] https://www.vmware.com/
[10] https://ceph.io/

Authors

Claudio Lo Giudice claudio.logiudice@cnr.it

Dr. Claudio Lo Giudice is a bioinformatician with a PhD in Cell Biology and Biotechnology. He conducted proteomic research in Finland and taught Bioinformatics at the University of Bari. Currently a technologist at CNR-ITB in Bari, he works on Linux infrastructure, scientific data management, and sensitive data handling. He is the author of REDIdb and UTRdb 2.0, databases for transcriptome and UTR region studies. His interests include bioinformatics, big data, cloud, RNASeq, and alternative splicing.



Giorgia Miniello giorgia.miniello@cnr.it

Dr. Giorgia Miniello holds a Ph.D. in Particle Physics from the University of Bari, with research conducted at the LHC-CMS on Higgs boson production and dark matter searches. She also earned a Master's in HPC Data Center Management, focusing on Big Data and monitoring systems. Currently a technologist at CNR-ITB, she develops cloud-based solutions for scientific infrastructures. Her main interests include particle physics, Big Data, HPC, and machine learning.

Guido Cauli guido.cauli@cnr.it

Student in Computer Science for Digital Businesses and graduated in Cinema, Photography and Audiovisual media. System administrator for GNU/Linux and Unix-like operating systems, mainly focusing on server and workstation management through Proxmox VE clustering and OpenStack systems, LXC and Docker container operations, enterprise networking and IT security aimed at the production, processing and secure storage of bioinformatics data.

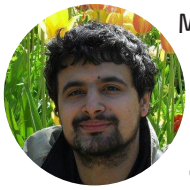
Francesco Rubino frubino@irb.hr

Francesco Rubino studied Biological Science in Bari, specialising in Bioinformatics. After experiences in industrial contexts, he defended his PhD thesis at Aberystwyth University working on ruminants' microbiota. He then worked at University of Queensland on marine microbiota. Returning to work on rumen microbiota at Queen's University Belfast, he contributed to Covid studies in wastewater for early diagnosis. Now he's at Ruder Boskovic Institute in Croazia as Research Associate.

Gianluca Cecinato gianluca.cecinato@cnr.it

Currently working at the Bari branch of the Institute for Biomedical Technologies (ITB) with the role of 'Technical Collaborator for Research Bodies (CTER)'. Responsibilities include the management of Unix-like open-source operating systems, networking, firewall administration, virtualization systems, and hardware maintenance of computing infrastructures

within the framework of the enhancement project 'ELIXIRxNextGenerationIT'.



Marco Moscatelli marco.moscatelli@cnr.it

Marco Moscatelli earned a Master's degree in Bioinformatics, during which he developed a bioinformatics platform to study the relationship between atmospheric particulate matter and human health. He attended courses on Red Hat System Administration and Ansible Essentials, which introduced him to system operations automation. He has experience in using cloud computing with Openstack and highlights the importance of this technology in providing elastic, scalable, and virtualized resources.

Sharon N. Cox sharonnatasha.cox@uniba.it

Sharon Natasha Cox, PhD in Biotechnology for Organ and Tissue Transplants, is a researcher who applies omics sciences to human health and disease. Her expertise includes, complex statistical and computational analysis of WES gene expression profiling, bioinformatics, variant identification, and the study of mtDNA–nDNA interactions in neurodegeneration. She is currently developing pipelines for long-read whole-genome and differential methylation analysis, with increasing focus on epigenomics.

Nadina Foggetti nadina.foggetti@cnr.it

A lawyer with a Ph.D. in EU and International Law, she is a Contract Professor in Cybersecurity, IT, and Biotech Law at Uniba. Technologist at CNR IBIOM and ELSI Officer for ELIXIR-IT, she has contributed to national and international projects on cybercrime, privacy, biotech, and digital law. Within ELIXIRxNextGenIT, she focuses on the Access Program, applying Open Science, FAIR data, and Open Access principles.

Francesca De Leo francesca.deleo@cnr.it

Francesca De Leo is Technology Director at CNR and Deputy Head of the ELIXIR-IT Node, based at the Institute of Biomembrane, Bioenergetics and Molecular Biotechnologies. She coordinates the ELIXIRxNextGenIT project funded by MUR and leads the Industry & Impact and Communication Offices within ELIXIR-IT. She holds a degree in Biological Sciences and a PhD in Biochemistry and Molecular Biology, with expertise in innovation, technology transfer, and research infrastructure management.

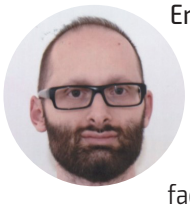
Angelo S. Varvara angelo.varvara@unimi.it

Angelo Sante Varvara is a PhD Student in Computational Biology at SEMM (European School of Molecular Medicine). Based in Bari, his expertise lies in computational analysis of human WGS and mitochondrial data, variant identification and prioritization, biological database creation and management, virtual machine administration, and biological workflow development. He is currently involved in human projects using long-read whole-genome and differential methylation analysis, with an increasing focus on rare human diseases.

Bruno Fosso bruno.fosso@uniba.it

Bruno Fosso is Associate Professor at the University of Bari “Aldo Moro”. The metagenomic investigation of host-associated microbiomes and environmental prokaryotic com-

munities is the main topic of his research activities. During the last 10 years, he developed tools and databases for both metabarcoding and shotgun metagenomics investigation of microbial communities. He participated in several Italian (MICROMAP, OMICS4FOOD) and European projects (BIOVEL, EMBRIC, LIFEWATCH, EXCELERATE and ELIXIR) and he is the coordinator of the ELIXIR-IT tools platform.

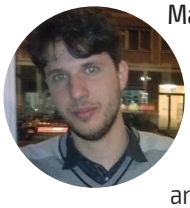


Ermes Filomena ermes.filomena@uniba.it

Bioinformatician passionate about free and open-source software, especially GNU/Linux systems. Since the beginning of his career, he has worked on NGS experiments, focusing on gene expression analysis from bulk and single-cell RNA-seq data. In the past two years, he has managed storage and primary analysis of data produced by the sequencing facility led by Prof. Pesole.

Pietro D'Addabbo pietro.daddabbo@uniba.it

Pietro D'Addabbo holds a degree in Medical Biotechnology and a PhD in Molecular Cytodifferentiation from the University "Alma Mater Studiorum" of Bologna. He has a 20-years experience in technical-scientific support and bioinformatic analysis, mainly in the field of genomic research. He is currently a fixed-term research assistant (RTDa, founded by PNRR) in Molecular Biology and lecturer in the Laboratory of Molecular Biology and Bioinformatics course at the University "Aldo Moro" of Bari.

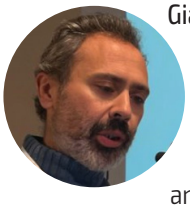


Marco A. Tangaro marcoantonio.tangaro@cnr.it

Currently a Researcher at CNR-IBIOM. Since 2015 he has been involved in the ELIXIR-IT community, developing Cloud services for bioinformatics and integrating new tools within the Galaxy workflow manager. In particular, he leads the development of the Laniakea platform, which allows the creation of on-demand Galaxy instances on the Cloud, and the UseGalaxy.it national Galaxy server.

Roberto Cilli roberto.cilli@uniba.it

Physicist and Data Analyst with experience in the information technology and services sector. Skilled in Geographic Information Systems, Machine Learning, and Spatial Analysis. Current research focuses on remote sensing—optical and SAR image processing, segmentation, registration, and quantitative metrics for decision support systems—and spatio-temporal data analysis for modeling and socio-economic applications.



Giacinto Donvito giacinto.donvito@infn.it

Giacinto Donvito, Senior Technologist, is an expert in distributed computing and cloud infrastructures for scientific research. He coordinates national and international projects in the fields of bioinformatics and omics data integration, leading the adoption of innovative technologies for the analysis, management, and enhancement of big data in life sciences and public-private partnerships.

Federico Zambelli

Federico Zambelli is an Associate Professor of Molecular Biology at the University of

Milan. His research focuses on the development of bioinformatics tools and algorithms for the analysis of sequencing data and the characterisation of gene expression regulation. As Technical Coordinator of ELIXIR-IT, he has contributed to several national projects aimed at building the technological infrastructure for biological data in Italy.

Ernesto Picardi ernesto.picardi@uniba.it

Ernesto Picardi is Full Professor of Molecular Biology at the University of Bari (Italy) and Research Associate at the Institute of Biomembranes and Bioenergetics (IBBE) of the National Research Council (CNR). His research activity focuses on bioinformatics and computational approaches to investigate co- and post-transcriptional molecular phenomena like alternative splicing and RNA editing by high-throughput sequencing technologies (including Illumina, PacBio, Oxford Nanopore). Further details are available at ORCID: http://orcid.org/0000-0002-6549-0114.

Flavio Licciulli flavio.licciulli@cnr.it

Master degree in Computer Science. Bioinformatician from 2001. Expertise in Research Data Management; expert in design and development of biological database and data integration tools; expert in application of FAIR principles for data and metadata standardization. Expert in Data Center management for the storage and processing of Life Science-oriented data. Competences in development of pipelines for the analysis of omics data.

Graziano Pesole graziano.pesole@uniba.it

Graziano Pesole is full professor of Molecular Biology in the University of Bari A. Moro and Associate Researcher of CNR-IBIOM, Director of "Consorzio Interuniversitario Biotecnologie (Trieste), Head of the Italian Node of ELIXIR, the European Research Infrastructure for Life Science (>400, h-index=84, total cites ≥30,000). His research activity is mostly focused on bioinformatics applications for the management and analysis of next generation sequencing data, also at single-cell resolution.