

Development of a state of the art computational environment for handling human genetic data

the effort of ELIXIR-IT

Claudio Lo Giudice CNR ITB Bari



Agenda

Introduction

Data
Infrastructure
Challenges

Architecture
Overview

Data Transfer
and Storage

Computational
Environment

Secure Data
Storage

Analysis
Workflow

Services
Deployment

Downstream
Services

Beacon

Fega

Conclusions

Introduction

- Nucleic acid sequencing technologies are revolutionizing various sectors, particularly healthcare
 - Applications like personalized medicine and pharmacogenomics
 - Reshaping medical treatments
- Technical, legal, and ethical challenges must be addressed



Data Infrastructure Challenges

- Sequence data from human samples require specialized handling
 - Size and ethical considerations are important factors
- ELIXIR-IT aims to develop efficient, secure solutions
 - Robust infrastructure for data storage, management, and access control is needed




Architecture Overview

- Service integrated into wider computational environment mainly hosted at ReCaS datacenter in Bari, Italy
- Key components include data transfer, storage facilities, and computational environment




CNR Milano:
~1200 CPU cores and
5 PB storage (mirror)
CNR Milano:
5 PB storage (backup)



UniMi:
INDACO update
UniBo:
IT infrastructure update
UniPd:
IT infrastructure update





INFN Bari:
4.192 CPU cores
2,1 PB storage; 10Gb network
CNR Bari:
12.320 CPU cores, 10 GPUs
7,2 PB storage; 25Gb network




UniBa (Physics Dept.)

4000 CPU cores and 16 GPUs
5.5 PB Cloud Storage
2 PB Posix Storage
20 PB Tape Library



Uniba (Physics Dept.)

5 PB Storage HPC/HTC
250 CPU cores and 8 GPUs
1.5 PB Storage Cloud



CNR (Bari) and University of Bari
Sequencing facility



Data Transfer and Storage

- Data Transfer via **SSH**
 - Transferred to BioRepository at ReCaS-Bari
- Data-at-rest Encryption and Geo-redundant Storage
 - Provided by BioRepository
- Backup Locations
 - CNR-ITB in Milan
 - CNR-ICAR in Naples



Computational Environment



Scalable and Virtualized Computational Resources

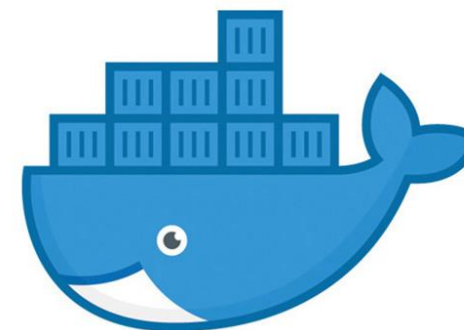
VM-based environment offers flexibility in resource allocation

State-of-the-Art Bioinformatics Tools

Deployed for efficient and accurate data analysis

Containerization and Package Management

Ensures compatibility and reproducibility of results



Secure Data Storage and Bioinformatic Reference Databases

- Encryption at the file system level of the virtual volumes used by the VMs provides secure storage for the data while they are being analyzed.
- Shared access to regularly updated bioinformatic reference databases stored in the BioRepository facility.



Analysis Workflow

Analysis Steps

- Quality Control
- Mapping
- Variant Calling and prioritization
- VCF Handling

Workflow Management Systems

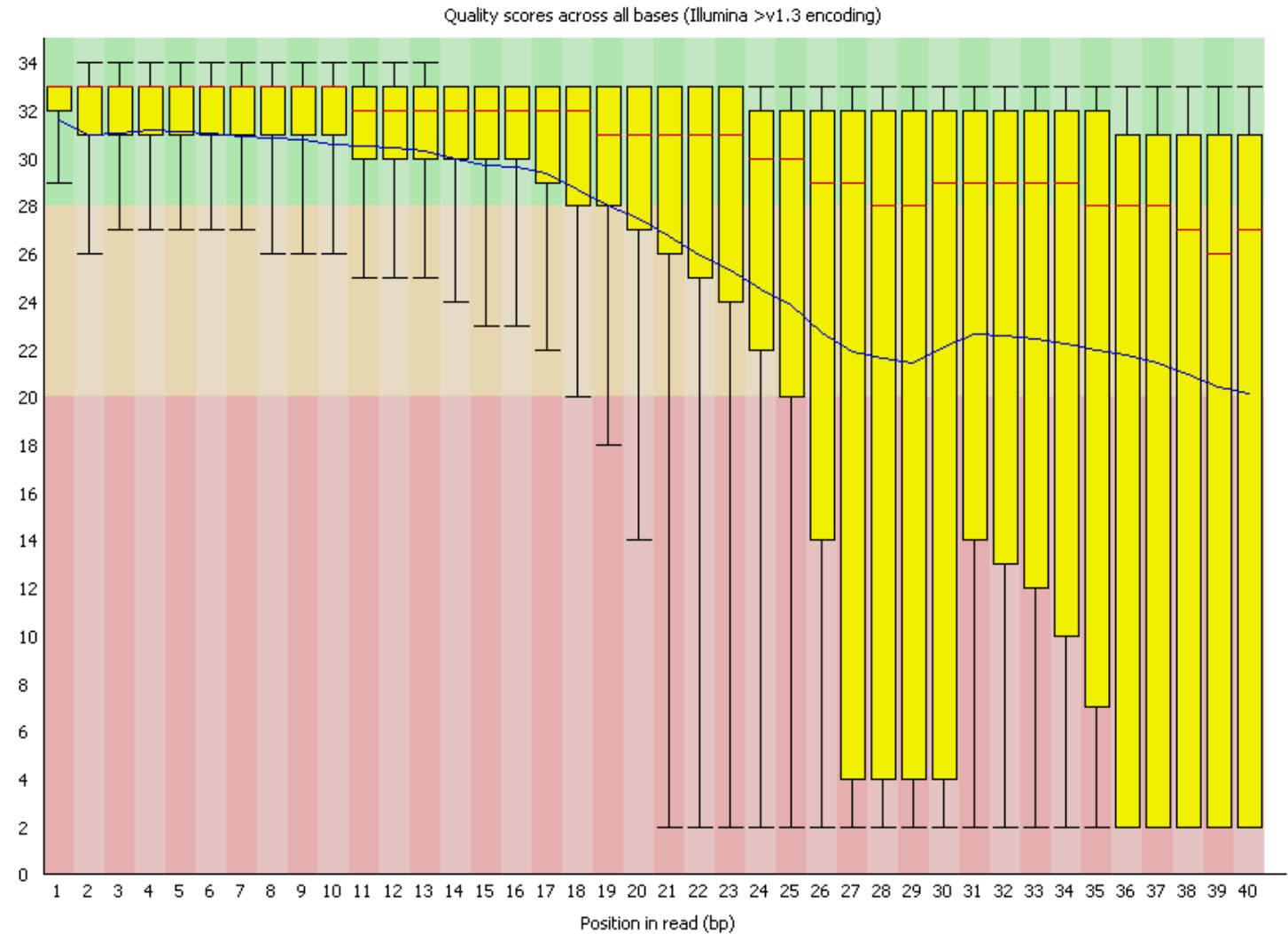
- Snakemake
- Nextflow
- Galaxy

IT Automation Engines

- Ansible

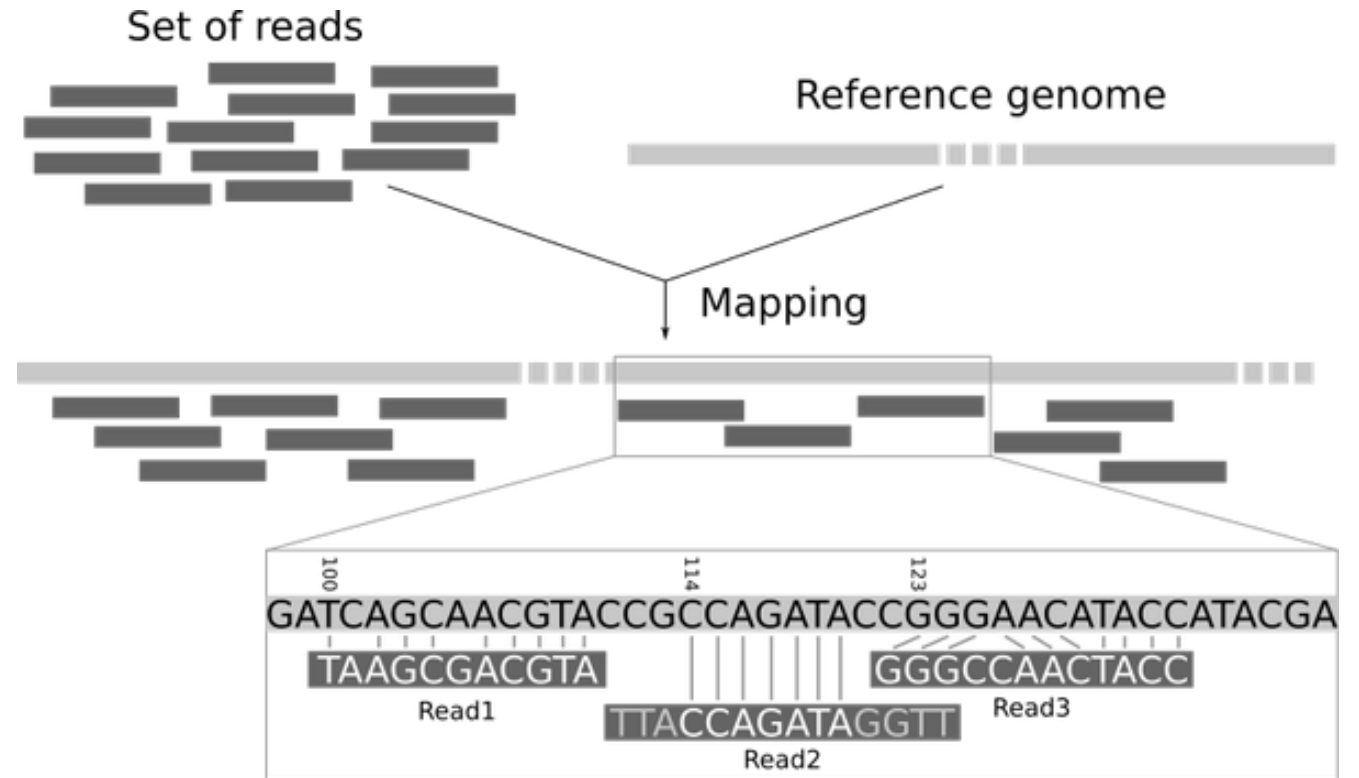
Assessing Data Quality

- Quality control is essential to ensure that the raw data is of sufficient quality.
- It involves checking for sequencing errors, base quality scores, and sequence complexity.



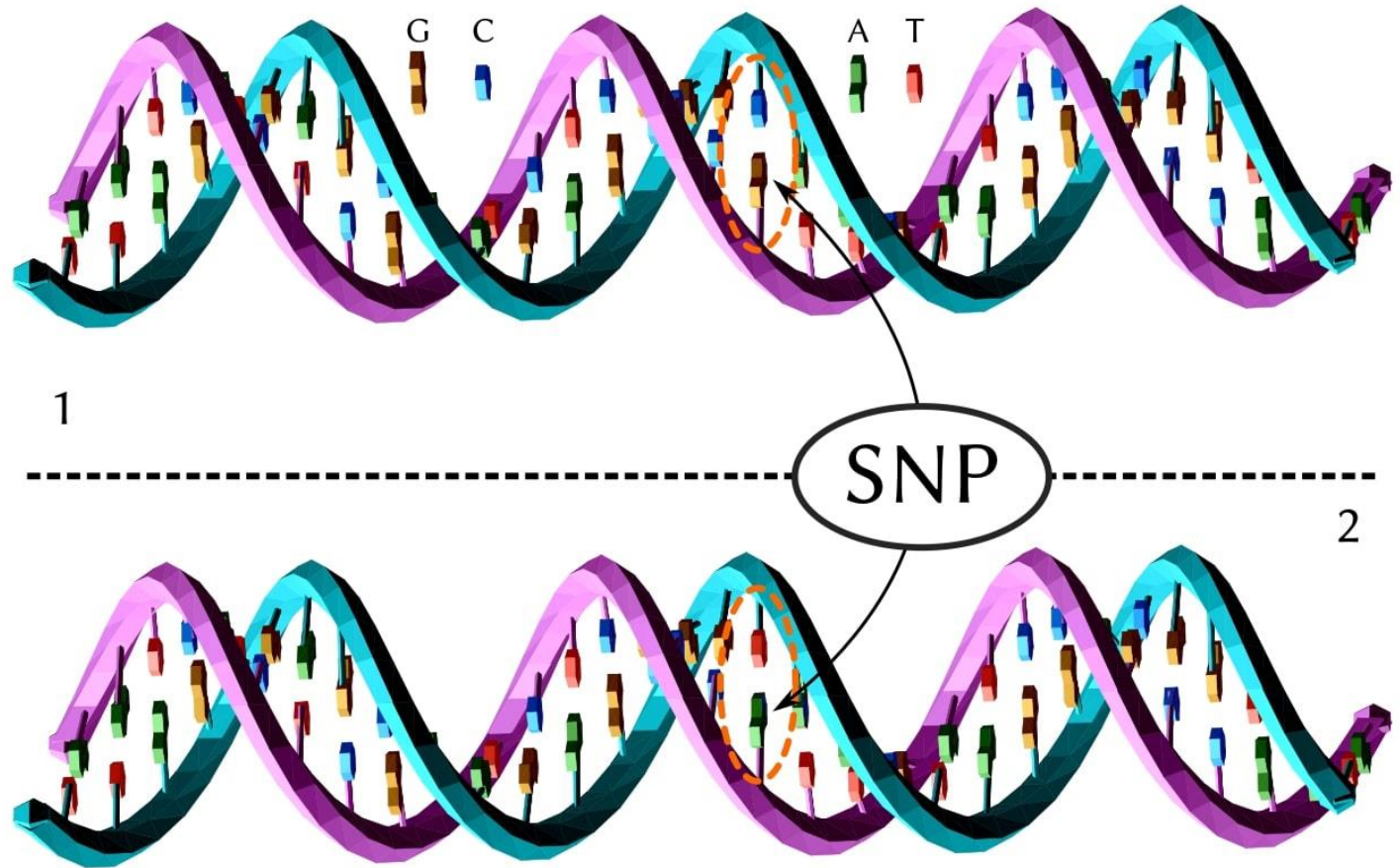
Mapping Process

- Mapping is done to align the reads obtained from the sequencing process to a reference genome.
- Various tools can be used for mapping, such as BWA, Bowtie or STAR.



Variant Calling

- Genetic variations in an individual's DNA compared to a reference genome.
- Single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) are relevant in various fields: medical research, forensic science and evolutionary biology.



Variant Call Format (VCF)

- VCF is the standard file format used to store variant information after variant calling.
- Efficient handling of VCF files is essential for downstream analysis.

```
##fileformat=VCFv4.2
##fileDate=20151002
##source=callMomV0.2
##reference=gi|251831106|ref|NC_012920.1| Homo sapiens mitochondrion, complete genome
##contig=<ID=MT,length=16569,assembly=b37>
##INFO=<ID=VT,Number=.,Type=String,Description="Alternate allele type. S=SNP, M=MNP, I=Indel">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate allele counts, comma delimited when multiple">
##FILTER=<ID=fa,Description="Genotypes called from fasta file">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HG00096 HG00097 HG00099 HG00100 HG00101 HG00102 HG00103 HG00105 HG00106 HG00107
MT 10 . T C 100 fa VT=S;AC=3 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 16 . A T 100 fa VT=S;AC=3 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 26 . C T 100 fa VT=S;AC=3 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 35 . G A 100 fa VT=S;AC=2 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 40 . TC CT 100 fa VT=M;AC=1 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 41 . C T 100 fa VT=S;AC=4 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 42 . TCC CCC,T 100 fa VT=S,I;AC=1,1 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 46 . T C 100 fa VT=S;AC=1 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 47 . G A 100 fa VT=S;AC=1 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 52 . TGG CAA 100 fa VT=M;AC=1 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 55 . TATTTT T,CATTTT,AATTTT,TTT,TTTTT 100 fa VT=I,S,S,I,I;AC=5,3,2,1,1 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 57 . T C 100 fa VT=S;AC=3 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 58 . TTT T 100 fa VT=I;AC=4 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MT 59 . T A 100 fa VT=S;AC=1 GT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Workflow management Systems



Snakemake

Snakemake is a popular workflow management system used in bioinformatics analysis.

It is known for its scalability and ease of use.



Nextflow

Nextflow is another popular workflow management system used in bioinformatics analysis.

It is known for its portability, and scalability, and for its easy integration with other tools and platforms.

Galaxy

Galaxy is a web-based workflow management system. It is known for its user-friendly interface, scalability, and ability to integrate with other tools and platforms.

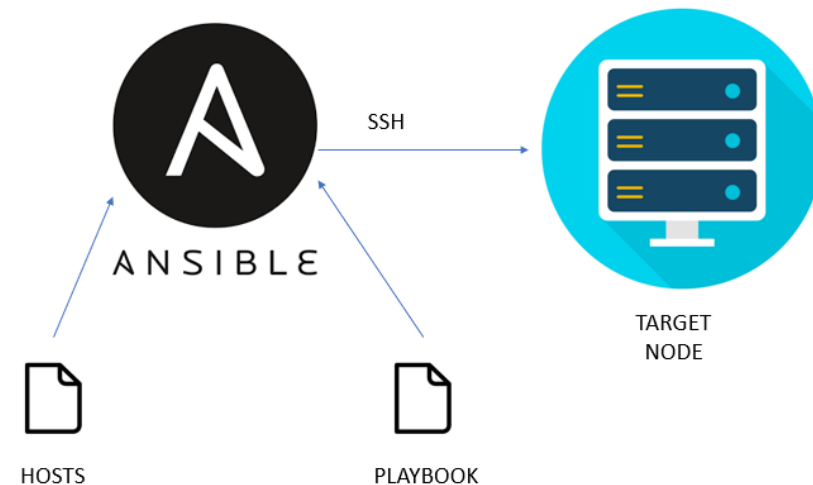
IT Automation Engines

Flexibility and Maintainability

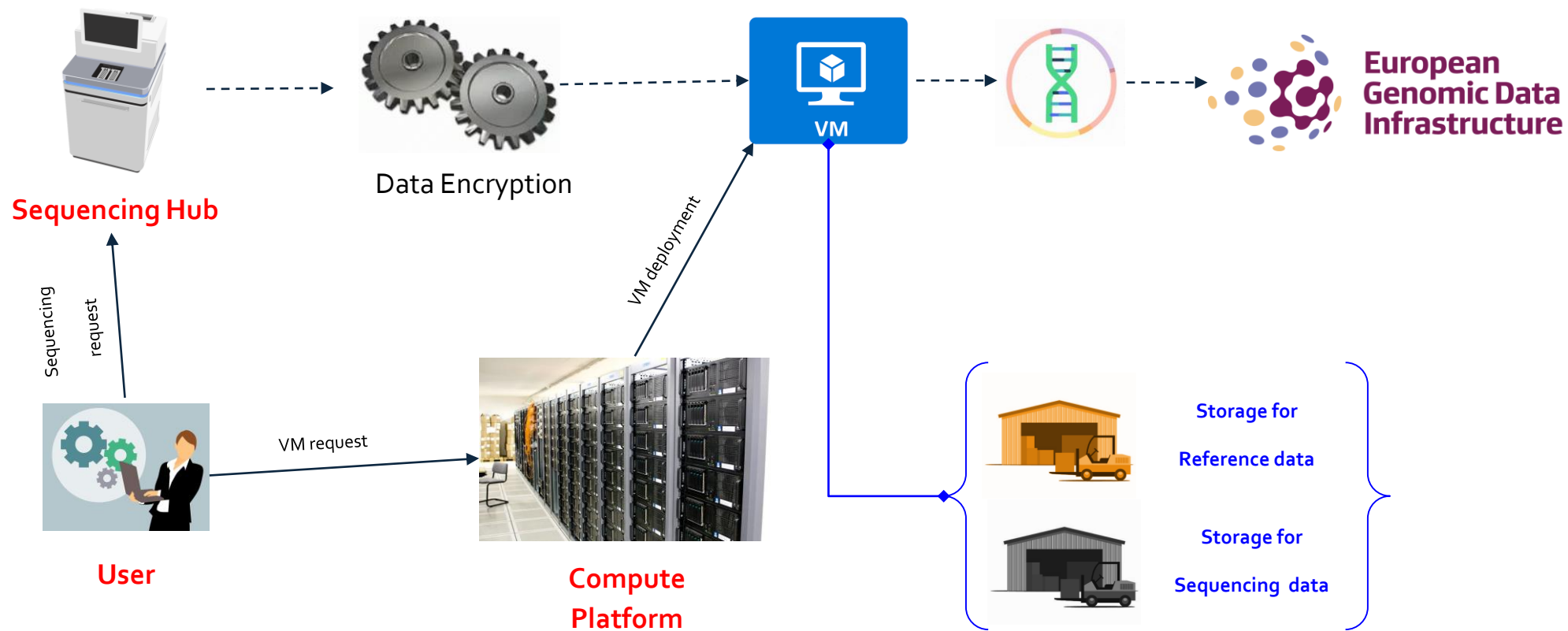
IT automation engines like Ansible ensure flexibility and maintainability of the bioinformatics analysis pipeline, making it easier to change and update as needed while reducing the risk of errors.

Infrastructure-as-Code Approach

Ansible provides an infrastructure-as-code approach that makes the bioinformatics analysis pipeline more scalable, reliable, and easy to maintain across multiple machines equipped with different operating systems.

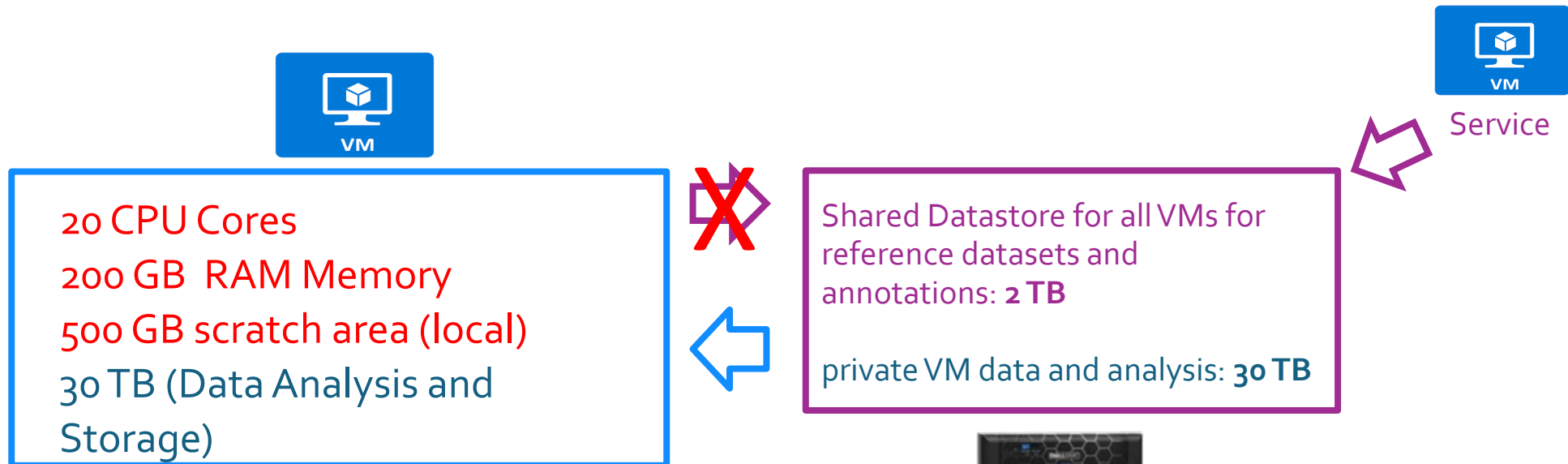


Services Deployment



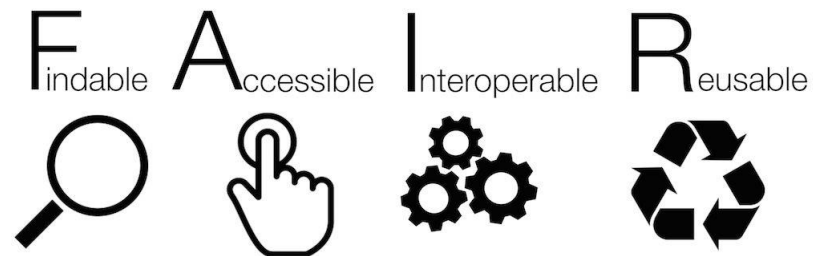
VMs configuration

VMs are configured to address specific tasks (e.g. WGS analysis) and “cloned” to be deployed to different users/projects.



High-performance
Parallel Storage

Downstream analysis



Downstream Services for Analysis

- FAIRification, deposition, and discoverability

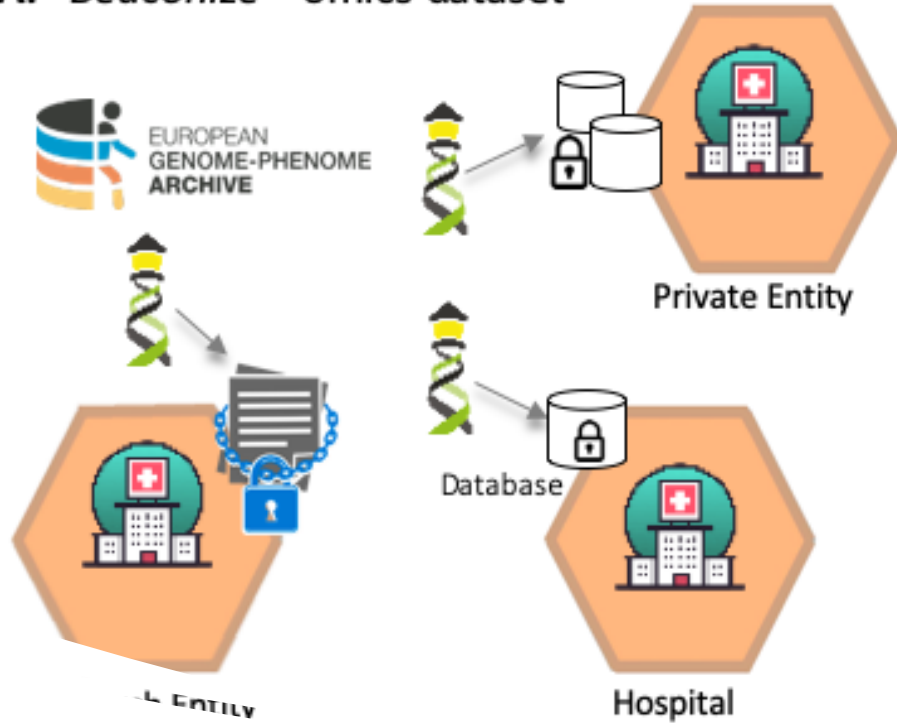
Federated Node of EGA Human Genome-Phenome Archive

- Benefits of FEGA

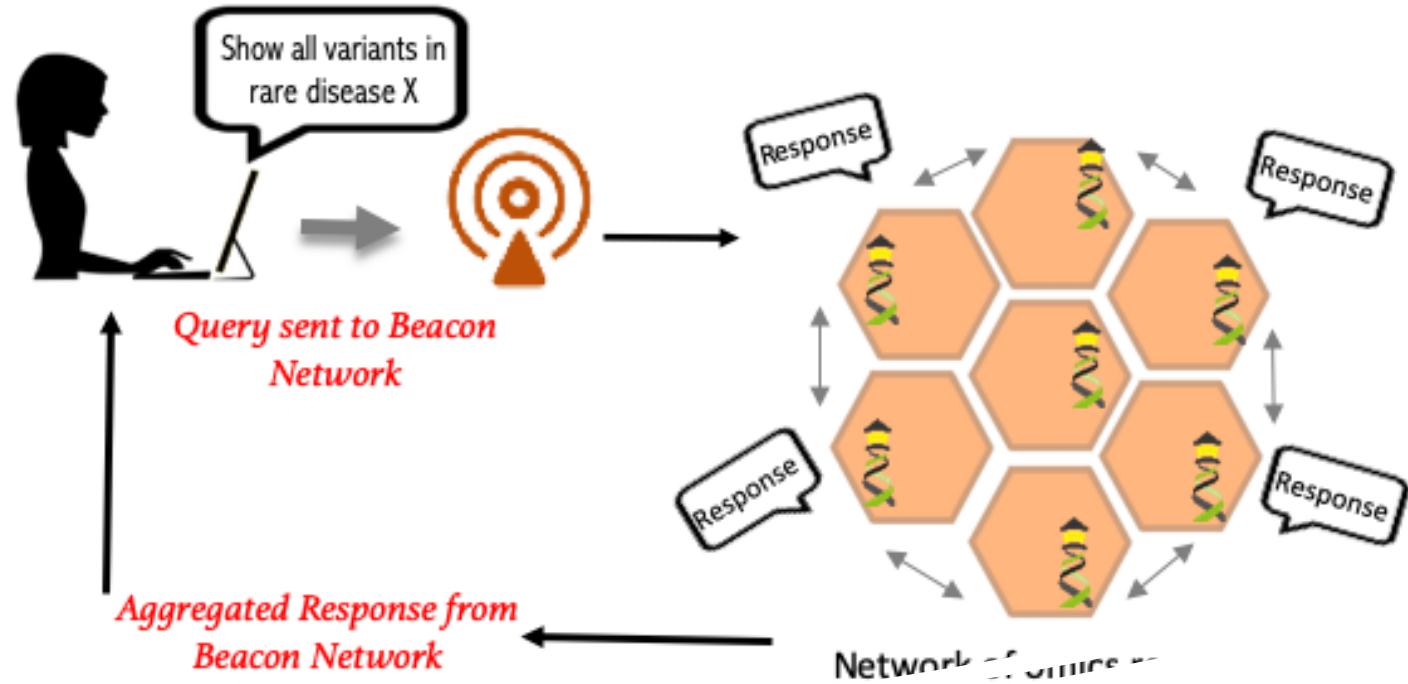
Beacon Protocol for Dataset Discoverability

- Discoverability of the datasets hosted by the FEGA node through its genomic variant-based query system.

A. "Beaconize" -omics dataset



B. Exchange of information



Beacon

The beacon protocol is a standardized method enabling institutions and databases to anonymously share information about the presence or absence of specific genetic variants within their datasets while adhering to privacy regulations. It empowers researchers to query multiple databases to ascertain the presence of a particular genetic variant in those datasets without disclosing personal information about individual subjects.

FEGA

The Federated European Genome-phenome Archive (EGA) facilitates secure storage, sharing, and analysis of genomic data across European research institutions, ensuring data privacy and compliance with ethical standards.

Integrates with the Beacon protocol, enabling researchers to query the platform to confirm the presence of genetic variants of interest within the hosted datasets, without compromising the privacy of individual data.



Conclusion

- Advancements in nucleic acid sequencing technologies offer immense potential for healthcare.
- Addressing technical, legal, and ethical challenges is crucial for realizing this potential.
- The integrated services approach presented here represents a significant step towards harnessing genetic data for research and health applications.

