Remastering data sets in life sciences

Spataro N., Carreño A., Di Tommaso P., Rambla J., Notredame C., van den Bergh E.

Cloud Computing
NextFlow
Data reproducibility

The Cloud Computing Pilot

In recent years, the technological advances in genomics have resulted in the massive generation of data and the development of various tools for their analysis. Nowadays, a limited number of computational tools represent the core of data analysis pipelines, but their continuous updating poses new challenges that the scientific community needs to face.

Challenges

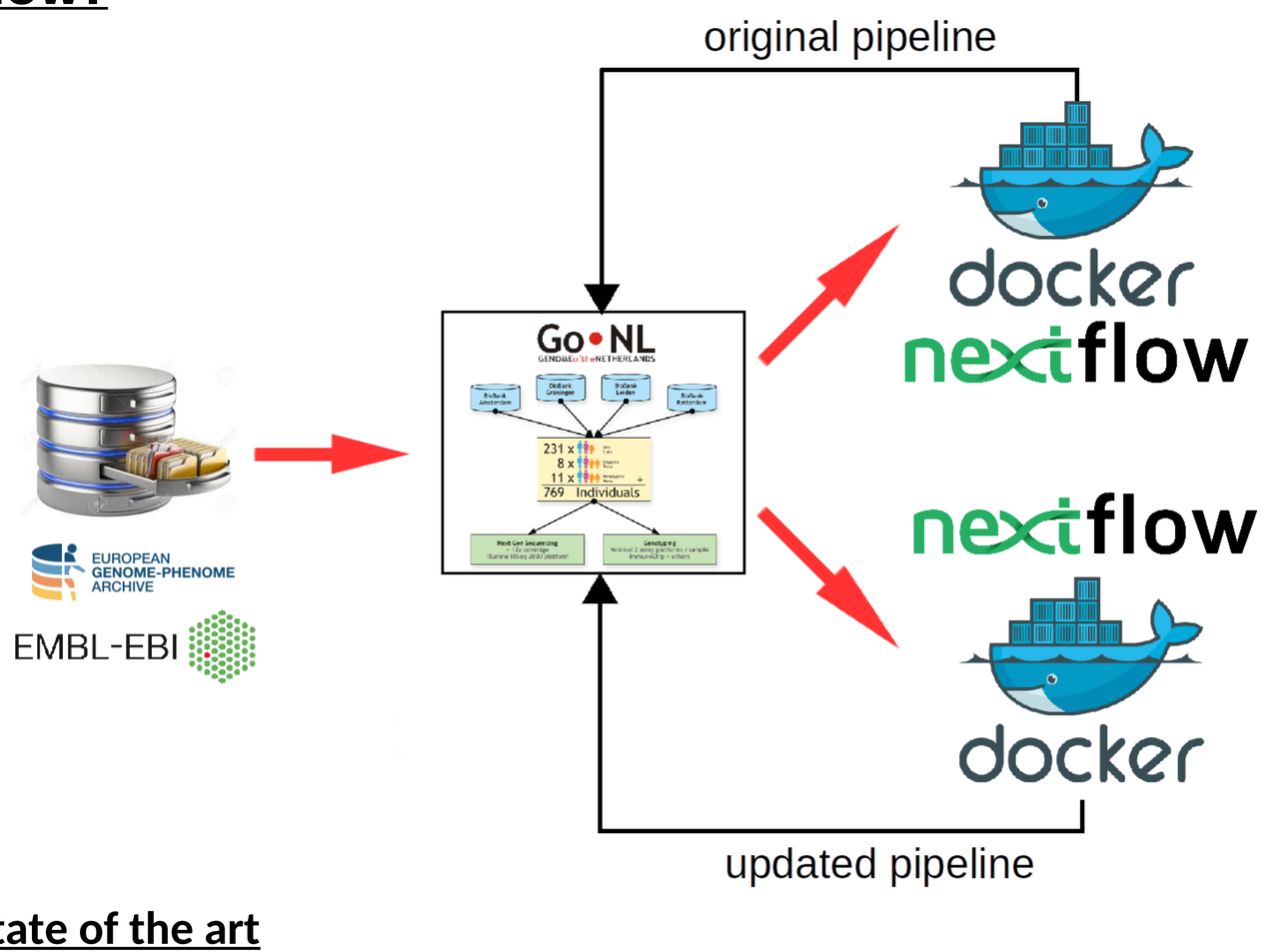
Resource obsolescence: Data are processed with reference genomes and analysis pipelines that were current at the time, but become obsolete quite fast

Resource portability: Software and tools require specific environment and dependencies that limit pipeline portability and maintaining them represents a time consuming task

Goals

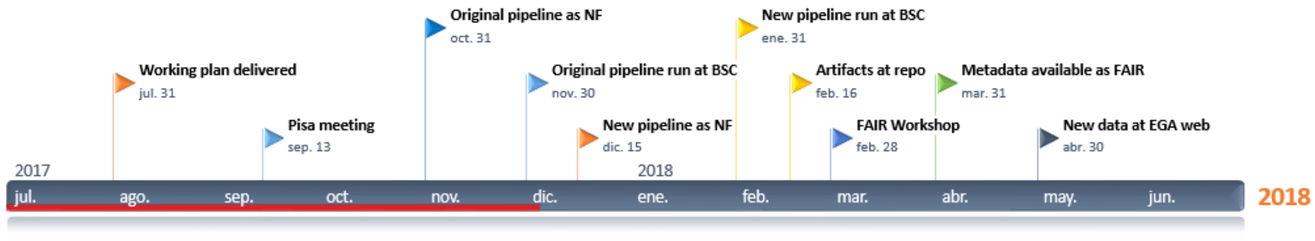
- Enable reproducibility → through the reconstruction of original analysis pipelines and the use of software containers which guarantee invariant executions and consistent results over time
- Allow portability → by packaging the original pipeline by using Nextflow workflow framework which enables the deployment across different computational environments in a portable manner
- Data re-analysis → by reproducing original pipelines and the use of updated resources

How?



- A third part dataset (GoNL project) as use case
- Reproduction of the original pipeline
- Production of an updated pipeline
- Containerized versions of both pipelines
- Test both pipelines on the use case dataset

State of the art



- ✓ Retrieve and install softwares and resources used in the original pipeline
- ✓ Generate a pipeline using the same commands and options used in the original pipeline
- ✓ Run the pipeline on a testing GoNL dataset
- ✓ Convert the pipeline as NextFlow pipeline

Technical challenges

- ✘ Availability of original softwares and resources (GATK 1.0, ancillary files from 1000Genomes Project)
- ✘ Custom files and resources used in the original pipeline
- ✘ Openness to share sensitive data within the scientific community and the associated security from resource providers