Publishable summary

The BiobankCloud project is addressing the Biobank Bottleneck problem, that is, the lack of platform support for the secure storage, analysis and interconnection of the coming massive wave of human genomic data. We are building a Platform-as-a-Service (PaaS) for Biobanking, by adding Big Data and data-intensive computing support for genomic data. Our BiobankCloud PaaS will provide the software that enables researchers to securely store and efficiently analyze up to petabytes of genomic data. Our platform is open-source and it extends Apache's Hadoop and YARN platforms, adding cloud-computing support, security, and support for genomic file formats. Our platform is also being defined within the context of the European and national regulatory frameworks for Biobank data, as well as within the context of the European data directive. We see our main research challenges as having to:

- define the regulatory framework and data model for biobank data sharing,
- develop an elastic, scalable, highly available storage infrastructure,
- develop a cross-cutting security platform that ensures data confidentiality, data integrity, and data access auditing,
- develop data-intensive computing workflows for aligning, clustering, aggregating, compressing and anonymizing sequence data,
- support the interconnection and sharing of genomic data between biobanks,
- leverage the storage and processing capacity of public clouds,
- validate our system by evaluating real-world, parallelized analysis pipelines to facilitate the biological interpretation of genomic data.

During the first 12 months of the BiobankCloud project, we have

- defined a regulatory and ethical framework for our platform;
- released a first version of the Hadoop platform with PaaS support, available for download from our website;
- developed a first scientific workflow that runs on YARN over our Hadoop platform;
- defined a security framework that we will develop on our Hadoop platform;
- and defined a framework for sharing Biobank Data between Biobanks.