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 human whole-genome sequence data. We are building Platform-as-a-Service (PaaS) support for Biobanking applications, by adding support for managing and processing sensitive genomic data to our data-intensive computing platform. 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, scalability, 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. The main research challenges we are addressing are:

- 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 24 months of the BiobankCloud project, we have

- defined and started implementing the regulatory and ethical framework for our platform;
- developed a scale-out Hadoop distribution, with a more scalable file system and highly available implementation of YARN;
- developed support for the automated deployment of our PaaS on both cloud-computing and bare-metal platforms;
- developed a second-level workflow scheduler (Hi-WAY) for YARN along with a parallel analysis platform for genomic data (Cuneiform);
- defined a security framework for our Hadoop platform and have implemented two-factor authorization and a perimeter security model based on a portal web application;
- defined and developed a first prototype of the framework for sharing Biobank Data between Biobanks.