Population-scale genomics with BiobankCloud and Hops/Hadoop

Jim Dowling
KTH - Royal Institute of Technology, Stockholm, Sweden
Dowling J (2015) *EMBnet.journal* **21**(Suppl A), e825. http://dx.doi.org/10.14806/ej.21.A.825

Recent advances in the cost, throughput, and speed of Next-Generation Sequencing (NGS) technology have resulted in huge growth in both the volume and velocity of genomic data available for processing.

Large NGS projects are now starting with the goal of population-based genomics, that is, the analysis of genomes of tens or even hundreds of thousands of individuals. Population-based genomics needs petabyte-scale data analytics platforms. BiobankCloud is an open-source framework, based on Hadoop[1], that supports the scalable and secure storage and analysis of NGS data, alongside traditional Biobank metadata. The platform scales to manage petabytes of data and it supports a number of Hadoop-based data analytics platforms for scale-out processing of NGS data, such as Cuneiform/HiWAY.

BiobankCloud supports multi-tenancy for studies with sensitive data. A project-based model for authentication has been incorporated into the Hadoop ecosystem, and studies can co-exist on the same platform without users accidentally or maliciously accessing each others study data. The multi-tenancy support is based around a user interface that brings together, in a single platform, the owners of NGS data with bioinformaticians, the researchers who typically analyze NGS data.

1 hadoop.apache.org/