Data Management and Analysis Infrastructure for the 1000 Genomes Project

Paul Flicek

European Bioinformatics Insititute

Current DNA sequencing technology enables large genome sequencing centres to match the data output of the decade-long human genome project in hours. These data volumes are challenging the bioinformatics and analysis infrastructure at the sequencing centers and at dedicated data repositories such as the EBI. This challenge is especially acute for efforts requiring data collection for combined analysis such as the 1000 Genomes Project: an international consortium including 9 sequencing centers in 4 countries to create a comprehensive catalog of human variation. The project is archiving primary data and has established a dedicated data coordinating center (DCC) to gather, validate, organise and provide data openly to researchers within and outside the project. As the project scales to full production, the DCC will also conduct a significant amount of low-level analysis. With others in the project, we've created new data formats, analysis tools, and put in place a dramatically more robust infrastructure.