

# The 1000Genomes project: A large data problem

Laura Clarke<sup>1</sup>

## Abstract

The 1000 Genomes Project is producing a deep catalogue of human variation, to provide a better baseline to underpin human genetics.

The project aims to sequence 1200 individuals from 3 main population groups in its first phase. The sequencing is being carried out at 6 different centres on both the Illumina and ABI SOLiD platforms, This will produce nearly 5000x coverage of the human genome.

This volume of data presents some unique challenges to the project from a data handling and distribution perspective. The Data Coordination Centre (DCC) for the project was setup as collaboration between the EBI and the NCBI. The DCC is closely linked to the Short Read Archive (SRA).

The DCC is responsible for ensuring data quality and availability both for the rest of the project consortium and the wider public.

The sequence data is retrieved from the SRA and the filtered to remove low quality reads and and check file syntax. The fastq files are released on the ftp site for use by the rest of the Consortium.

Groups at the Sanger Institute and Tgen provide alignments of the sequence data in BAM format. We also QC check these files again to ensure syntax and data consistency.

The data is made publicly available via both ftp and aspera on two mirrored ftp sites

<ftp://ftp.1000genomes.ebi.ac.uk/> and

<ftp://ftp-trace.ncbi.nih.gov/1000genomes/>

Once the Analysis group has provided results we make these available both on the ftp site and on the 1000genomes Ensembl style browser <http://browser.1000genomes.org/>. The SNPs visible in this browser are based on calls made on the high coverage CEU and YRI individuals. These calls area available in:

[ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot\\_data/release/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/) and

[ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot\\_data/release/](ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/)

---

<sup>1</sup>European Bioinformatics Institute Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK.