EMBL-EBI

European Bioinformatics Institute
is an Outstation of the
European Molecular Biology Laboratory.

1000 Genomes Project Resources

L. Clarke, H. Zheng-Bradley, R. Smith, E Kuleshea, I Toneva, B. Vaughan, P. Flicek and
1000 Genomes Consortium
European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Introduction

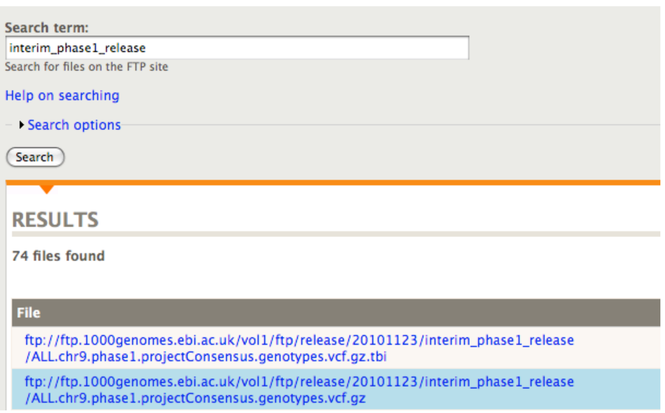
The main goal of the 1000 genomes project is to establish a comprehensive and detailed catalogue of human genome variations; which in turn will empower association studies to identify disease-causing genes. The project now has data and variant genotypes for more than 1000 individuals in 14 populations. The ftp site contains more than 120Tbytes of data in 200,000 files.

DATA TYPE	FILE FORMAT	SIZE
sequence	FASTQ	43 Tbases raw sequence
alignment	BAM	56 Tbytes of BAM files
variants	VCF	38.9M SNPs ~4.7M short indels

Discoverability

Sequence, alignment and variant data is made available as quickly as possible through the project ftp site. (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/ | ftp://ftp-trace.ncbi.nih.gov/1000genomes/). With more than 200,000 files though discovering new data can be difficult.

The ftp site has a index updated nightly. This index is searchable from our website. http://www.1000genome.org/ftpsearch



The search allows users to specify which ftp site to get paths to, to get md5 checksums and also filter out high volume results like bam and fastq files

We also have various routes for users to discover new data.

- Website http://www.1000genomes.org/announcements
- Twitter @1000genomes
- RSS http://www.1000genomes.org/announcements/rss.xml
- Email 1000announce@1000genomes.org

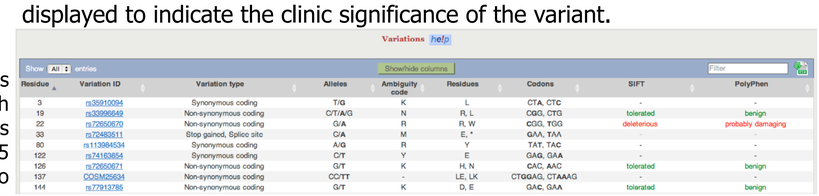
Visualization

http://browser.1000genomes.org

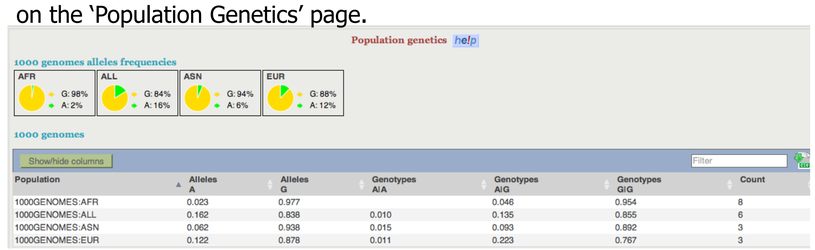
The 1000 Genomes project utilizes the Ensembl Browser to display our variant calls. We provide rapid access to project variant calls through the browser before they become available via dbSNP and DGVa.

Tracks of 1000 genomes variants by population can be viewed in the location page:

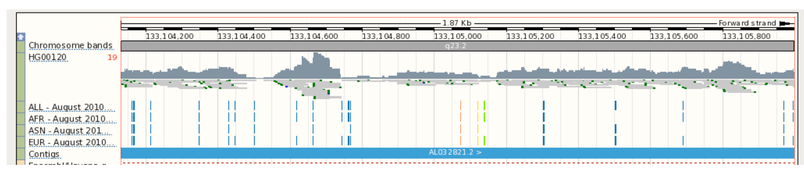
A list of variants can be obtained for any given transcript. In addition to basic information about a variant, PolyPhen and SIFT annotation are displayed to indicate the clinic significance of the variant.



Allele frequency for individual variants in different populations is displayed on the 'Population Genetics' page.



Users can Attach remote files as custom tracks. In example below, the HG00120 track is 1000 Genomes bam file added to the browser.



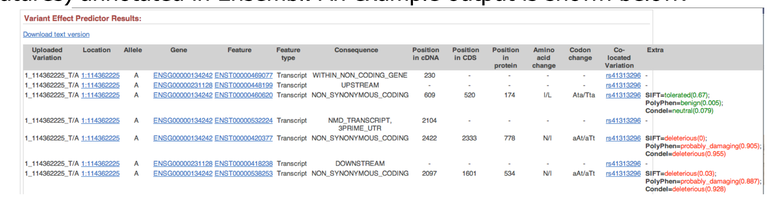
Accessibility

http://browser.1000genomes.org/tools.html

The project provides several tools to help users access and interpret the data provided.

Variant Effect Predictor

The predictor takes a list of variant positions and alleles, and predicts the effects of each of these on any overlapping features (transcripts, regulatory features) annotated in Ensembl. An example output is shown below:

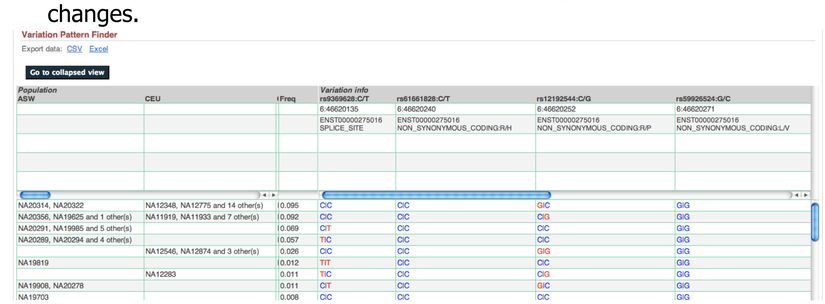


Data Slicer

Many of the 1000 Genomes files are large and cumbersome to handle. The Data Slicer allows users to get data for specific regions of the genome and to avoid having to download many gigabytes of data they don't needl samples/ populations you choose. Below is the Data Slicer input interface:

Variation Pattern Finder

- The Variation Pattern Finder (VPF) allows one to look for patterns of shared variation between individuals in the a VCF file.
- Within a vcf file different samples have different combination of variation genotypes. The VPF looks for distinct variation combinations within a user specifed region, shared by different individuals.
- The VPF only on variations that functional consequences for protein coding genes such as non-synonymous coding SNPs and splice site changes.



Acknowledgements

We would like to thank the Ensembl variation team for all their help, particularly Will McLaren and Graham Ritchie.
Funding: The Wellcome Trust

Laura Clarke
EBI
laura@ebi.ac.uk

EMBL- EBI
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD
UK

1000 Genomes
A Deep Catalog of Human Genetic Variation