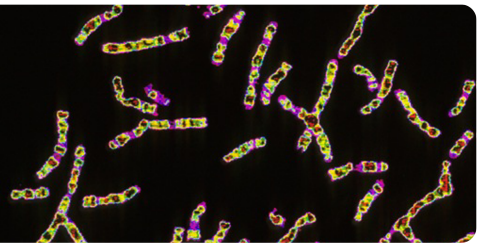
1000 Genomes

A Deep Catalog of Human Genetic Variation



How to Access the Data

Laura Clarke

October 14th 2014

Size

There was 235GB of sequence in the ENA before the 1000genomes project started.

- There are 461406 files on the ftp site
- There are 580T of data on the ftp site
- There are 26 populations
- There are 2854 samples
- There are 79072 gigabases of low coverage sequence
- 28753 x coverage in low coverage
- There are 35607 gigabases of exome sequence

There are currently 1196200GB of sequence in the ENA.

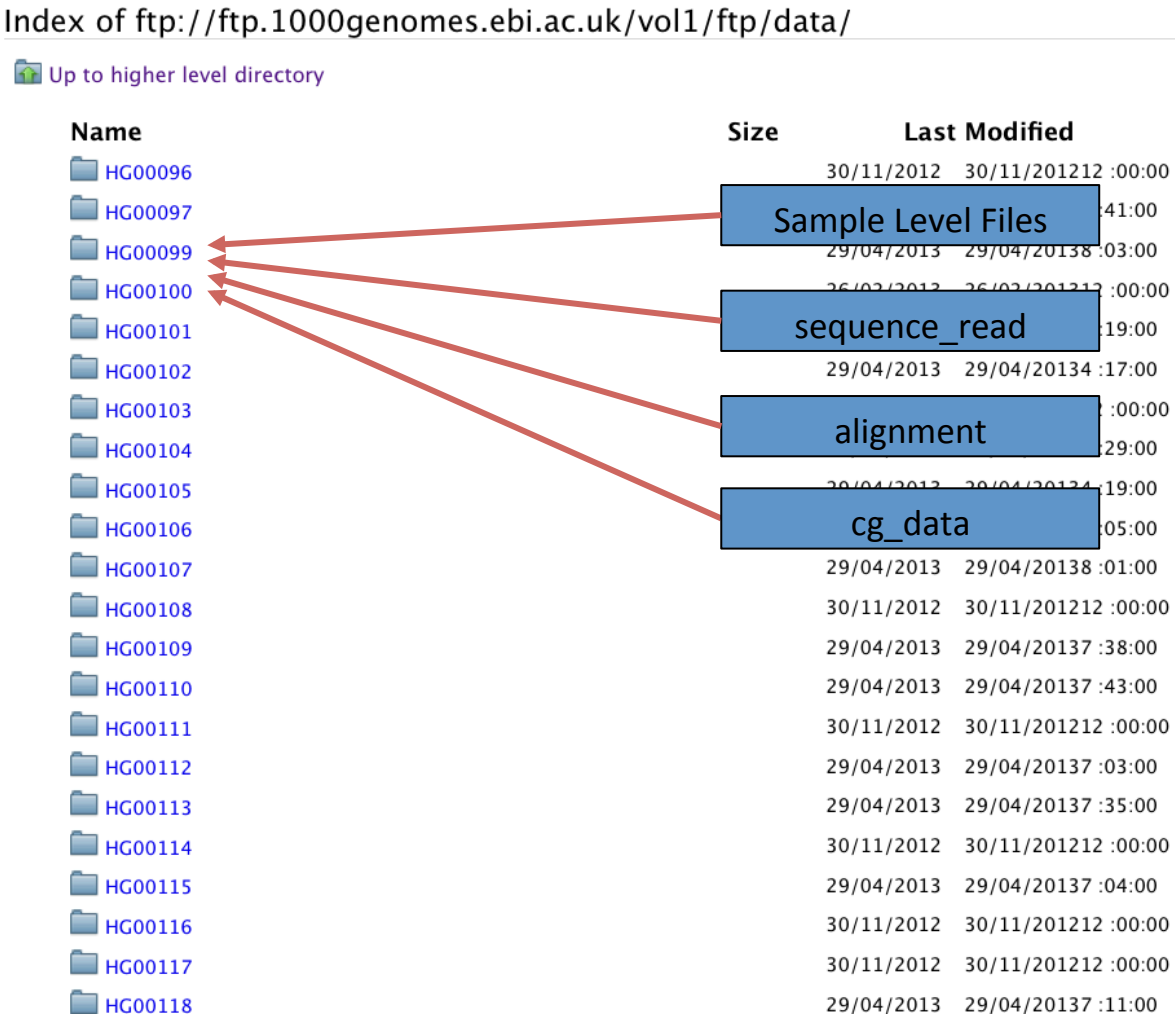
Data Availability

- FTP site: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/
  - Raw Data Files
- AWS Amazon Cloud: http://aws.amazon.com/1000genomes/
  - FTP mirror
- Web site: http://www.1000genomes.org
  - Release Announcements
  - Documentation
- Ensembl Style Browser: http://browser.1000genomes.org
  - Browse 1000 Genomes variants in Genomic Context
  - Variant Effect Predictor
  - Data Slicer
  - Other Tools

FTP Site

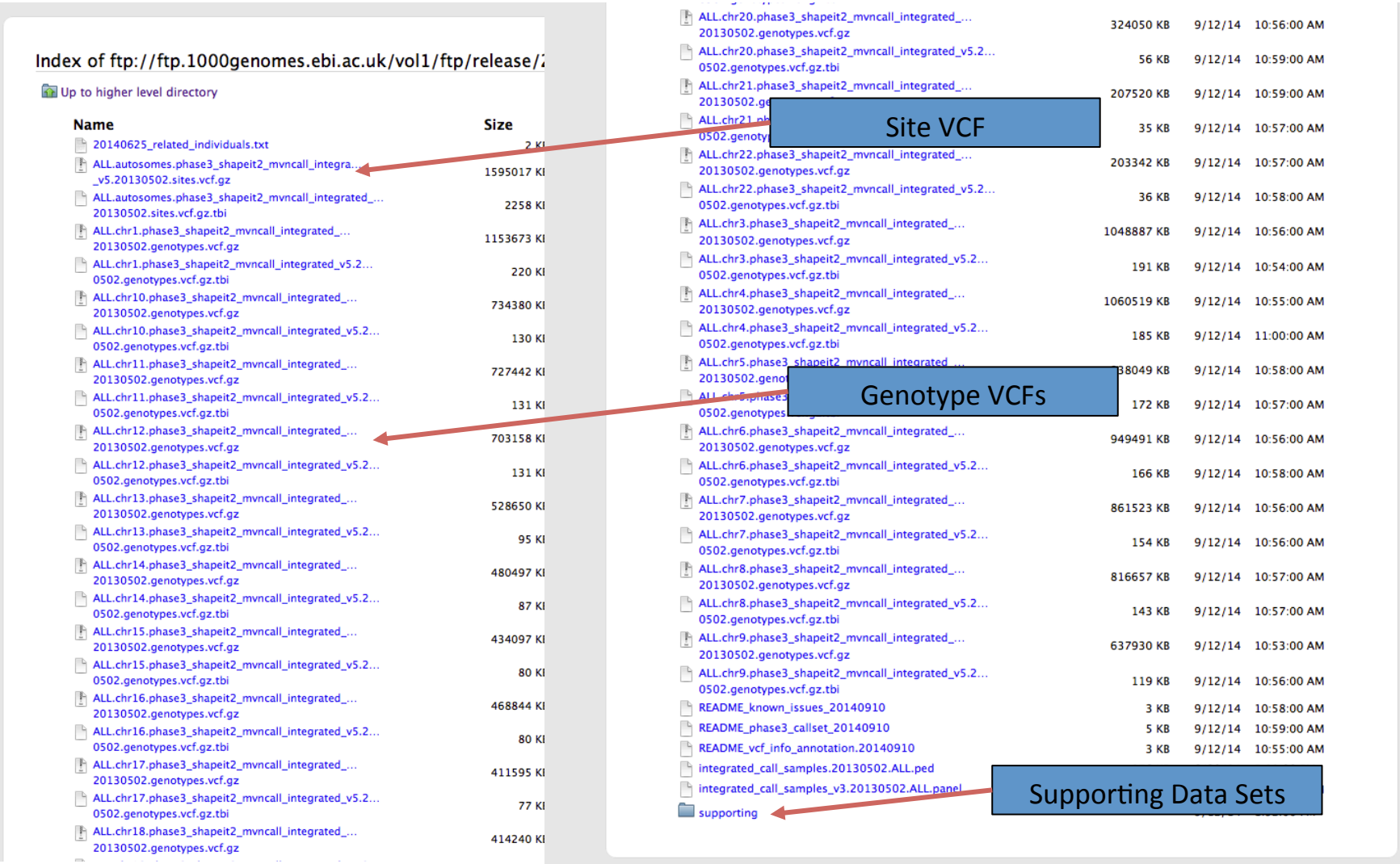
- Two mirrored ftp sites
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp
  - ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp
- NCBI site is direct mirror of EBI site
- Can be up to 24 hours out of date
- Both also accessible using aspera
  - http://asperasoft.com/
- Available via GlobusGridFTP
- EBI site has http mirror
  - http://ftp.1000genomes.ebi.ac.uk/vol1/ftp

The FTP Site: Data



1000 Genomes
A Deep Catalog of Human Genetic Variation
EMBL-EBI

FTP Site: Phase 3

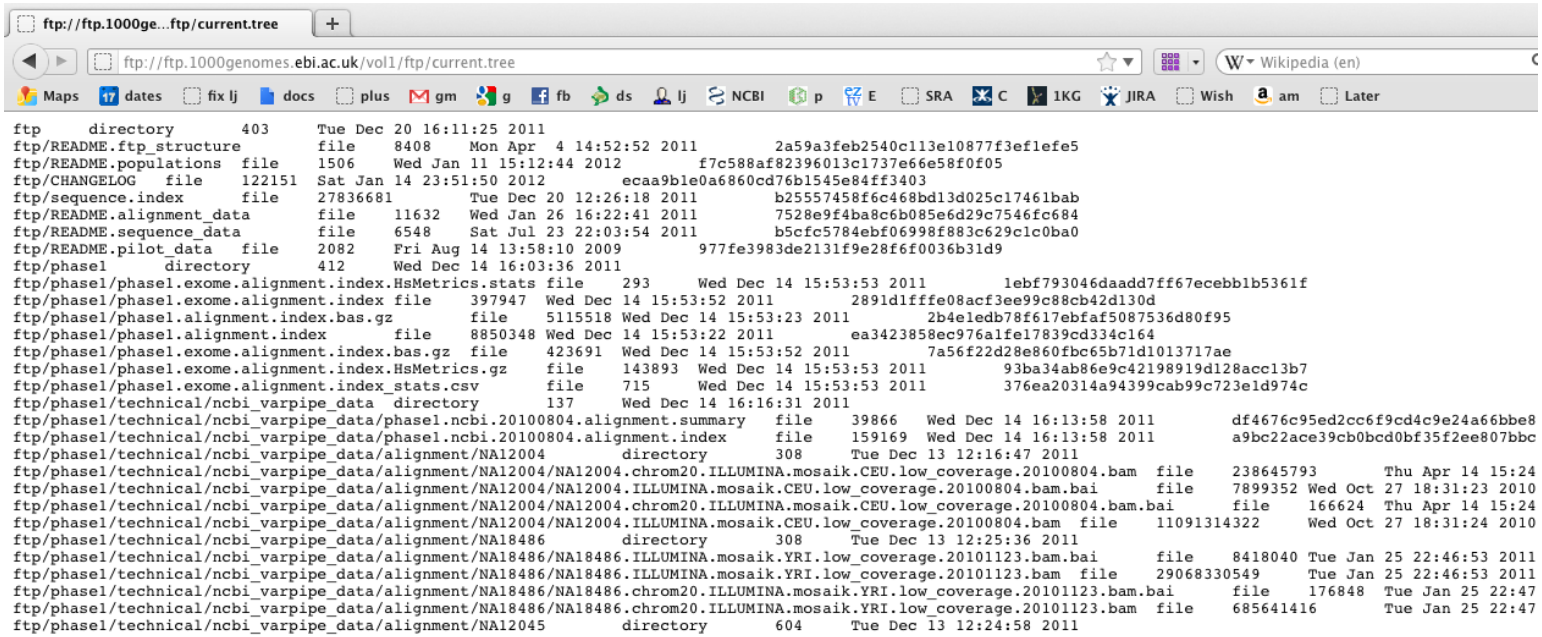


Finding Data

Current.tree file

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree

Current Tree is updated nightly so can be upto 24 hours out of date



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

- Data Slicer
- Variation Pattern Finder
- VCF to PED Converter
- Variant Effect Predictor
- Forge
- Allele Frequency Calculator

Allele Frequency Calculator

- 1000 Genomes VCF files contain genotypes for all individuals
- The individuals are from many different populations
- The Allele Frequency Calculator provides per population allele frequency and alternative allele counts.

1000 Genomes

A Deep Catalog of Human Genetic Variation

Tools | Help

Search 1000 Genomes

Go

e.g. **gene BRCA2** or **Chromosome 6:133098746-133108745**

Start Browsing 1000 Genomes data

Browse Human →
GRCh37

Protein variations →
View the consequences of sequence variation at the level of each protein in the genome.

Individual genotypes →
Show different individual's genotype, for a variant.

Browser update September 2011

based on interim Main project data from 20101123 for 1094 individuals and ensembl release 63. The data can be found on the ftp site.

Please see www.1000genomes.org for more information about the data presented here and instructions for downloading the complete data set.

- View sample data

The 1000 Genomes Browser

Ensembl-based browser provides early access to 1000genomes data

In order to facilitate immediate analysis of the 1000 Genomes Project data by the whole scientific community, this browser (based on Ensembl) integrates the SNP calls from an interim release 20101123. This data has be submitted to dbSNP, and once rsid's have been allocated, will be absorbed into the UCSC and Ensembl browsers according to their respective release cycles. Until that point **any non rs SNP id's on this site are temporary and will NOT be maintained**.

Links

1000 Genomes →
More information about the 1000 Genomes Project on the 1000 genomes main site.

Pilot browser →
This browser is based on Ensembl release 60 and represents the variant set analysed as part of A map of human genome variation from population-scale sequencing , Nature 467 ,1061.1073.

Tutorial →
The 1000 Genomes Browser Tutorial.

The 1000 Genomes Project is an international collaborative project described at www.1000genomes.org.

The 1000 Genomes Browser is based on Ensembl web code.

Ensembl is a joint project of EMBL-EBI and the Wellcome Trust Sanger Institute

Tools page

1000 Genomes

A Deep Catalog of Human Genetic Variation

We provide a number of ready-made tools for processing your data. At the moment, small datasets can be uploaded to our servers and processed online; for larger datasets, we provide an API script that can be downloaded

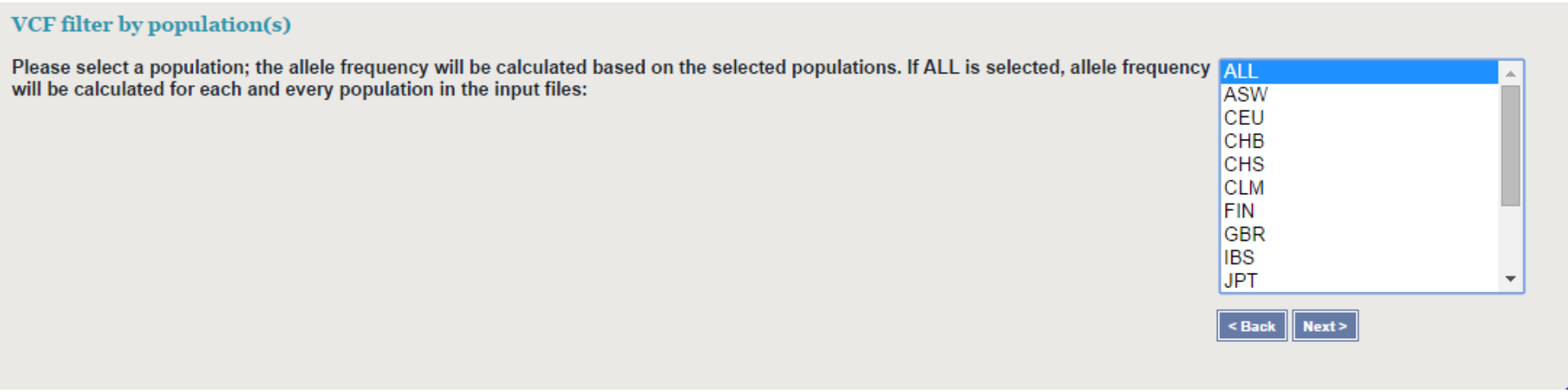
In the near future we aim to offer an intermediate service, whereby medium-to-large data sets can be submitted to a queue, similar to BLAST.

Currently available:

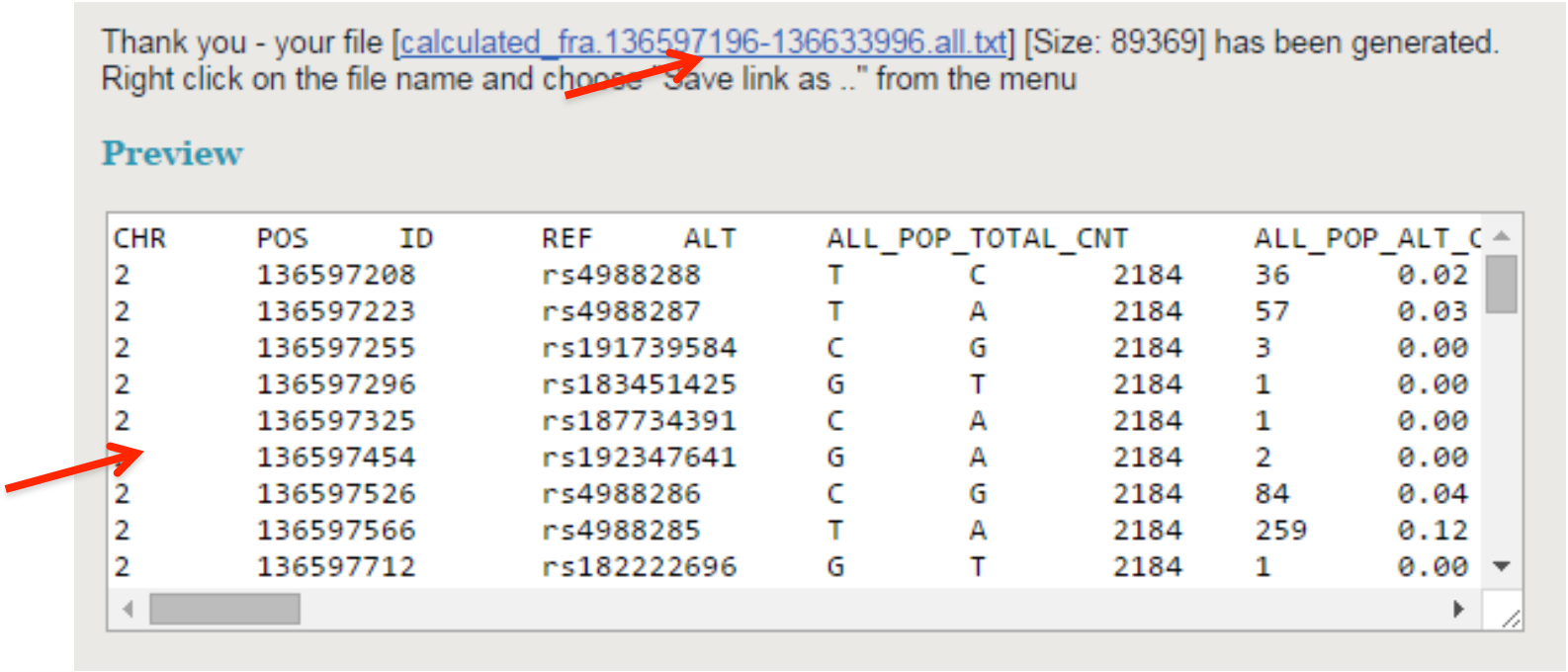
Tool	Description	
Assembly converter	Map your data to the current assembly. Accepted file formats: GFF, GTF, BED, PSL N.B. Export is currently in GFF only	Online version
ID History converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.	Online version (max 30 ids)
Variant Effect Predictor	(Formerly SNP Effect Predictor). Upload a set of SNPs in our standard format and export a file containing consequence types. Uploaded tracks can also be viewed on Location pages.	Online version (max 750 SNPs)
Data Slicer	Get a subset of data from a BAM or VCF file.	Online version (max 10K region)
Variation Pattern Finder	Identify variation patterns in a chromosomal region of interest for different individuals. Only variations with functional significance such non-synonymous coding, splice site will be reported by the tool. Click here for more extensive documentation.	Online version
VCF to PED converter	The VCF to PED converter allows users to parse a vcf file to create a linkage pedigree file (ped) and a marker information file, which together may be loaded into ld visualization tools like Haploview. Click here for more extensive documentation.	Online version
Forge analyser	The Forge tool takes a list of variants and analizes their enrichment in functional regions from the ENCODE or Roadmap Epigenome project on a tissue specific basis. The analysis requires a miniumum of 20 variants which have to be present in phase 1 of 1000 Genomes project. Click here for full tool documentation.	Online version
Allele frequency calculator	This tool takes a VCF file, a matching sample panel file, a chromosomal region, a population name and calculates population-wide allele frequency for sites within the chromosomal region defined.	Online version

Allele Frequency Calculator

Allele Frequency Calculator



Allele Frequency Calculator



Allele Frequency Calculator

CHR	POS	ID	REF	ALT	AN	AC	FREQ
2	136606360	rs4988252	A	C	198	198	1
2	136606413	rs4988251	A	C	198	198	1
2	136606741	rs2082730	G	T	198	198	1
2	136607870	rs6752360	A	T	198	198	1
2	136607871	rs6752362	A	C	198	198	1
2	136607976	rs10648294	T	TCTCTC	198	198	1
2	136608466	rs4954492	C	A	198	198	1
2	136608231	rs4954490	G	A	198	158	0.8

The Future

The International Genome Sample Resource (IGSR), a Wellcome Trust funded project that will be built on the foundation of the 1000 Genomes Project starts in 2015.

- **IGSR plans to:**
- Maintain the existing 1000 genomes data and move to GRCh38
- Collect other data sets generated on the Coriell Cell Lines including Geuvadis
- Add new populations to expand the global diversity of the variant catalog

Announcements and Contact Info

http://1000genomes.org

1000announce@1000genomes.org

http://www.1000genomes.org/1000-genomes-annoucement-mailing-list

http://www.1000genomes.org/announcements/rss.xml

http://twitter.com/#!/1000genomes

Please send questions to info@1000genomes.org

Acknowledgements

The 1000 Genomes Consortium

Paul Flicek
Holly Zheng Bradley
Bert Overduin
Emily Pritchard
Ian Streeter
Avik Datta
David Richardson

Ensembl Variation

Fiona Cunningham
Will McLaren
Laurent Gil
Anja Thormann
Sarah Hunt
The Rest of Ensembl

Forge

Ian Dunham