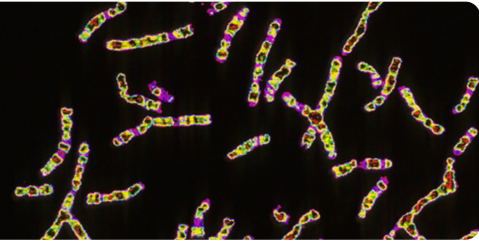


1000 Genomes

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- Can be up to 24 hours out of date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

Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/

 Up to higher level directory























Name	Size	Last Modified
 HG00096		30/11/2012 30/11/2012 12:00:00
 HG00097		
 HG00099		29/04/2013 29/04/2013 8:03:00
 HG00100		26/03/2012 26/03/2012 12:00:00
 HG00101		
 HG00102		29/04/2013 29/04/2013 4:17:00
 HG00103		
 HG00104		
 HG00105		29/04/2013 29/04/2013 4:19:00
 HG00106		
 HG00107		29/04/2013 29/04/2013 8:01:00
 HG00108		30/11/2012 30/11/2012 12:00:00
 HG00109		29/04/2013 29/04/2013 7:38:00
 HG00110		29/04/2013 29/04/2013 7:43:00
 HG00111		30/11/2012 30/11/2012 12:00:00
 HG00112		29/04/2013 29/04/2013 7:03:00
 HG00113		29/04/2013 29/04/2013 7:35:00
 HG00114		30/11/2012 30/11/2012 12:00:00
 HG00115		29/04/2013 29/04/2013 7:04:00
 HG00116		30/11/2012 30/11/2012 12:00:00
 HG00117		30/11/2012 30/11/2012 12:00:00
 HG00118		29/04/2013 29/04/2013 7:11:00

Diagram illustrating the structure of the FTP site data. Red arrows point from the 'Sample Level Files' box to folders HG00096, HG00097, and HG00099. Red arrows point from the 'sequence_read' box to folders HG00100, HG00101, and HG00102. Red arrows point from the 'alignment' box to folders HG00103, HG00104, and HG00105. Red arrows point from the 'cg_data' box to folders HG00106, HG00107, and HG00108.



FTP Site: Phase 3

Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/

Up to higher level directory

Name	Size
20140625_related_individuals.txt	2 KI
ALL.autosomes.phase3_shapeit2_mvncall_integrated_v5.20130502.sites.vcf.gz	1595017 KI
ALL.autosomes.phase3_shapeit2_mvncall_integrated_v5.20130502.sites.vcf.gz.tbi	2258 KI
ALL.chr1.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	1153673 KI
ALL.chr1.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	220 KI
ALL.chr10.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	734380 KI
ALL.chr10.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	130 KI
ALL.chr11.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	727442 KI
ALL.chr11.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	131 KI
ALL.chr12.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	703158 KI
ALL.chr12.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	131 KI
ALL.chr13.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	528650 KI
ALL.chr13.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	95 KI
ALL.chr14.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	480497 KI
ALL.chr14.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	87 KI
ALL.chr15.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	434097 KI
ALL.chr15.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	80 KI
ALL.chr16.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	468844 KI
ALL.chr16.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	80 KI
ALL.chr17.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	411595 KI
ALL.chr17.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	77 KI
ALL.chr18.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	414240 KI

ALL.chr20.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	324050 KB	9/12/14	10:56:00 AM
ALL.chr20.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	56 KB	9/12/14	10:59:00 AM
ALL.chr21.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	207520 KB	9/12/14	10:59:00 AM
ALL.chr21.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	35 KB	9/12/14	10:57:00 AM
ALL.chr22.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	203342 KB	9/12/14	10:57:00 AM
ALL.chr22.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	36 KB	9/12/14	10:58:00 AM
ALL.chr3.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	1048887 KB	9/12/14	10:56:00 AM
ALL.chr3.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	191 KB	9/12/14	10:54:00 AM
ALL.chr4.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	1060519 KB	9/12/14	10:55:00 AM
ALL.chr4.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	185 KB	9/12/14	11:00:00 AM
ALL.chr5.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	88049 KB	9/12/14	10:58:00 AM
ALL.chr5.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	172 KB	9/12/14	10:57:00 AM
ALL.chr6.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	949491 KB	9/12/14	10:56:00 AM
ALL.chr6.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	166 KB	9/12/14	10:58:00 AM
ALL.chr7.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	861523 KB	9/12/14	10:56:00 AM
ALL.chr7.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	154 KB	9/12/14	10:56:00 AM
ALL.chr8.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	816657 KB	9/12/14	10:57:00 AM
ALL.chr8.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	143 KB	9/12/14	10:57:00 AM
ALL.chr9.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz	637930 KB	9/12/14	10:53:00 AM
ALL.chr9.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz.tbi	119 KB	9/12/14	10:56:00 AM
README_known_issues_20140910	3 KB	9/12/14	10:58:00 AM
README_phase3_callset_20140910	5 KB	9/12/14	10:59:00 AM
README_vcf_info_annotation.20140910	3 KB	9/12/14	10:55:00 AM
integrated_call_samples.20130502.ALL.ped			
integrated_call_samples_v3.20130502.ALL.panel			
supporting			

Site VCF

Genotype VCFs

Supporting Data Sets

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

```
ftp directory 403 Tue Dec 20 16:11:25 2011
ftp/README.ftp_structure file 8408 Mon Apr 4 14:52:52 2011 2a59a3feb2540c113e10877f3ef1efe5
ftp/README.populations file 1506 Wed Jan 11 15:12:44 2012 f7c588af82396013c1737e66e58f0f05
ftp/CHANGELOG file 122151 Sat Jan 14 23:51:50 2012 ecaa9ble0a6860cd76b1545e84ff3403
ftp/sequence.index file 27836681 Tue Dec 20 12:26:18 2011 b25557458f6c468bd13d025c17461bab
ftp/README.alignment_data file 11632 Wed Jan 26 16:22:41 2011 7528e9f4ba8c6b085e6d29c7546fc684
ftp/README.sequence_data file 6548 Sat Jul 23 22:03:54 2011 b5cfc5784ebf06998f883c629c1c0ba0
ftp/README.pilot_data file 2082 Fri Aug 14 13:58:10 2009 977fe3983de2131f9e28f6f0036b31d9
ftp/phasel directory 412 Wed Dec 14 16:03:36 2011
ftp/phasel/phasel.exome.alignment.index.HsMetrics.stats file 293 Wed Dec 14 15:53:53 2011 1ebf793046daadd7ff67ecebb1b5361f
ftp/phasel/phasel.exome.alignment.index file 397947 Wed Dec 14 15:53:52 2011 2891dlfff08acf3ee99c88cb42d130d
ftp/phasel/phasel.alignment.index.bas.gz file 5115518 Wed Dec 14 15:53:23 2011 2b4e1edb78f617ebfaf5087536d80f95
ftp/phasel/phasel.alignment.index file 8850348 Wed Dec 14 15:53:22 2011 ea3423858ec976af1e17839cd334c164
ftp/phasel/phasel.exome.alignment.index.bas.gz file 423691 Wed Dec 14 15:53:52 2011 7a56f22d28e860fbc65b71d1013717ae
ftp/phasel/phasel.exome.alignment.index.HsMetrics.gz file 143893 Wed Dec 14 15:53:53 2011 93ba34ab86e9e42198919d128acc13b7
ftp/phasel/phasel.exome.alignment.index_stats.csv file 715 Wed Dec 14 15:53:53 2011 376ea20314a94399cab99c723e1d974c
ftp/phasel/technical/ncbi_varpipe_data directory 137 Wed Dec 14 16:16:31 2011
ftp/phasel/technical/ncbi_varpipe_data/phasel.ncbi.20100804.alignment.summary file 39866 Wed Dec 14 16:13:58 2011 df4676c95ed2cc6f9cd4c9e24a6bbe8
ftp/phasel/technical/ncbi_varpipe_data/phasel.ncbi.20100804.alignment.index file 159169 Wed Dec 14 16:13:58 2011 a9bc22ace39cb0bcd0bf35f2ee807bbc
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA12004 directory 308 Tue Dec 13 12:16:47 2011
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.chrom20.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam file 238645793 Thu Apr 14 15:24
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam.bai file 7899352 Wed Oct 27 18:31:23 2010
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.chrom20.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam.bai file 166624 Thu Apr 14 15:24
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam file 11091314322 Wed Oct 27 18:31:24 2010
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA18486 directory 308 Tue Dec 13 12:25:36 2011
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam.bai file 8418040 Tue Jan 25 22:46:53 2011
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam file 29068330549 Tue Jan 25 22:46:53 2011
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.chrom20.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam.bai file 176848 Tue Jan 25 22:47
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.chrom20.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam file 685641416 Tue Jan 25 22:47
ftp/phasel/technical/ncbi_varpipe_data/alignment/NA12045 directory 604 Tue Dec 13 12:24:58 2011
```

Tools

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

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



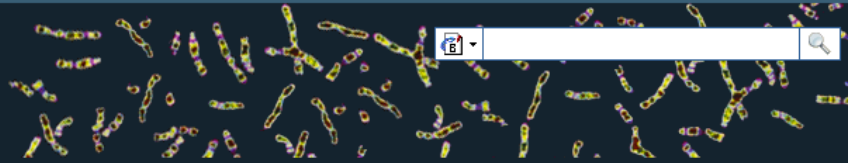
Allele Frequency Calculator

- 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

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 been 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 analyzes their enrichment in functional regions from the ENCODE or Roadmap Epigenome project on a tissue specific basis. The analysis requires a minimum 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

1000 Genomes release 15 - Sep 2014 © [EBI](#)



Allele Frequency Calculator

Provide file URL:

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/ALL.chr2.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.vcf.gz
```

Example file: [vcf](#)

Region:

e.g. 1:1-50000

Sample-Population Mapping File URL:

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/integrated_call_samples.20101123.ALL.panel
```

[What is a panel file?](#)

e.g. [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/integrated_call_samples.20101123.ALL.panel](#)

[Clear box](#)

[Next >](#)



Allele Frequency Calculator

VCF filter by population(s)

Please select a population; the allele frequency will be calculated based on the selected populations. If ALL is selected, allele frequency will be calculated for each and every population in the input files:

- ALL
- ASW
- CEU
- CHB
- CHS
- CLM
- FIN
- GBR
- IBS
- JPT

< Back Next >



Allele Frequency Calculator

Thank you - your file [\[calculated_fra.136597196-136633996.all.txt\]](#) [Size: 89369] has been generated. Right click on the file name and choose "Save link as .." from the menu

Preview

CHR	POS	ID	REF	ALT	ALL_POP_TOTAL_CNT	ALL_POP_ALT_C	
2	136597208	rs4988288	T	C	2184	36	0.02
2	136597223	rs4988287	T	A	2184	57	0.03
2	136597255	rs191739584	C	G	2184	3	0.00
2	136597296	rs183451425	G	T	2184	1	0.00
2	136597325	rs187734391	C	A	2184	1	0.00
2	136597454	rs192347641	G	A	2184	2	0.00
2	136597526	rs4988286	C	G	2184	84	0.04
2	136597566	rs4988285	T	A	2184	259	0.12
2	136597712	rs182222696	G	T	2184	1	0.00



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-announcement-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

