

The 1000 Genomes Tutorial Tools for Data Handling and Processing

Laura Clarke
17th February 2012



Introduction

This Presentation should give the user an overview of the tools and command lines users can use to gain access to 1000 genomes data sets.



Glossary

- **Pilot** : The 1000 Genomes project ran a pilot study between 2008 and 2010
- **Phase 1**: The initial round of exome and low coverage sequencing of 1000 individuals
- **Phase 2**: Expanded sequencing of 1700 individuals and method improvement
- **SAM/BAM**: Sequence Alignment/Map Format, an alignment format
- **VCF**: Variant Call Format, a variant format



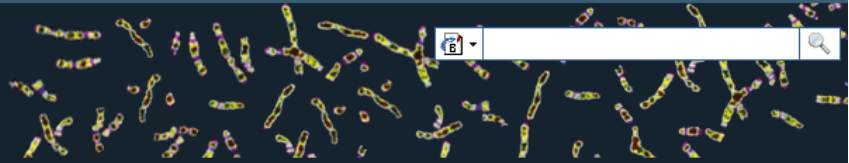
Summary

- Where to find the Tools
- Variant Effect Predictor
- Data Slicer
- Variation Pattern Finder
- VCF to PED
- API and Database access
- Data Availability
- Announcements



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](#)



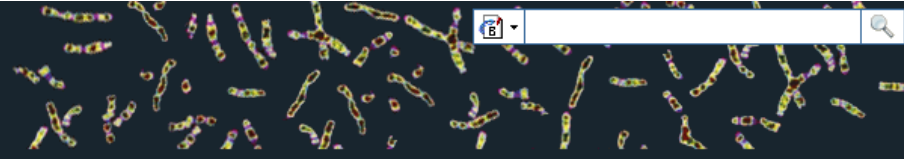
<http://browser.1000genomes.org>



Tools page

1000 Genomes

A Deep Catalog of Human Genetic Variation



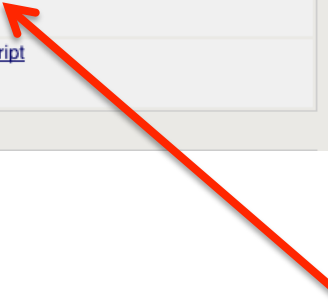
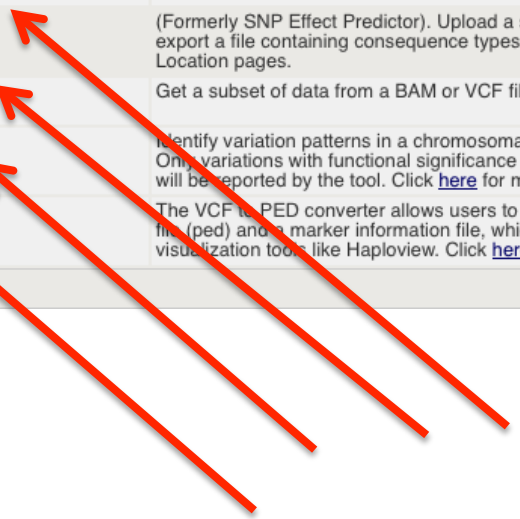
Tools | Help

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 (you will also need to [install our Perl API](#) to use these).

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	Online version	API script
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	API script
ID History converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.	Online version (max 30 ids)	API script
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)	API script
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	API script
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 Id visualization tools like Haploview. Click here for more extensive documentation.	Online version	API script



Variant Effect Predictor

- Predicts Functional Consequences of Variants
- Both Web Front end and API script
- Can provide
 - sift/polyphen/condel consequences
 - Refseq gene names
 - HGVS output
- Can run from a cache as well as Database
- Convert from one input format to another
- Script available for download from:
- [ftp://ftp.ensembl.org/pub/misc-scripts/Variant effect predictor/](ftp://ftp.ensembl.org/pub/misc-scripts/Variant_effect_predictor/)
- [http://browser.1000genomes.org/Homo_sapiens/
UserData/UploadVariations](http://browser.1000genomes.org/Homo_sapiens/UserData/UploadVariations)



Data Management

- Upload Data
- Attach DAS
- Attach Remote File
- Manage Data
- Features on Karyotype
- Data Converters
 - Assembly Converter
 - ID History Converter
 - Variant Effect Predictor**
 - Data Slicer
 - Variation Pattern Finder

Variant Effect Predictor:

This tool takes a list of variant positions and alleles, and predicts the effects of each of these on overlapping transcripts and regulatory regions annotated in Ensembl. The tool accepts substitutions, insertions and deletions as input, uploaded as a list of [tab separated values](#), [VCF](#) or Pileup format input.

Upload is limited to 750 variants; lines after the limit will be ignored. Users with more than 750 variations can split files into smaller chunks, use the standalone [perl script](#) or the [variation API](#). See also [full documentation](#)

Input file

Species:

Human (Homo sapiens): GRCh37

Name for this upload (optional):

Paste file:

Upload file:

Choose File no file selected

or provide file URL:

Input file format:

Ensembl default

Options

Get regulatory region consequences:

Type of consequences to display:

Ensembl terms

Check for existing co-located variants:

Yes

Return results for variants in coding regions only:

Show HGNC identifier for genes where available:

Show Ensembl protein identifiers where available:

Show HGVS identifiers for variants where available:

No

Non-synonymous SNP predictions (human only)

SIFT predictions:

No

PolyPhen predictions:

No

Condel consensus (SIFT/PolyPhen) predictions:

No

Frequency filtering of existing variants (human only)

Filter variants by frequency:

NB: Enabling frequency filtering may be very slow for large datasets

Filter: Exclude variants with MAF greater than 0.1 in any 1KG low coverage population

Next >

Variant Effect Predictor

- `perl variant_effect_predictor.pl -input 6_381831625_3184704.vcf -sift p -polyphen p -check_existing`
- `less variant_effect_output.txt`

```
#Uploaded_variation Location Allele Gene Feature Feature_type Consequence
cDNA_position CDS_position Protein_position Amino_acids Codons Exi
sting_variation Extra
rs138094825 6:31831667 A ENSG00000204385 ENST00000414427 Transcript
DOWNSTREAM - - - - - rs138094825 -
rs138094825 6:31831667 A ENSG00000204385 ENST00000229729 Transcript
INTRONIC - - - - - rs138094825 -
6_31832657_C/T 6:31832657 T ENSG00000204385 ENST00000229729
Transcript NON_SYNONYMOUS_CODING 1883 1862 621 R/H cGc/cAc -
PolyPhen=possibly_damaging;SIFT=deleterious
```



Data Slicing

- Use samtools to get subsections of bam files
 - **samtools view** http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/HG01375.mapped.ILLUMINA.bwa.CLM.low_coverage.20111114.bam 6:31833625-31833704
- Use tabix to get subsections of vcf files
 - **tabix -h** ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b37.vcf.gz 6:31830969-31846823 | **vcf-subset -c HG01375**
- http://browser.1000genomes.org/Homo_sapiens/UserData/SelectSlice



Data Slicing

Custom Data

Data Management

- Upload Data
- Attach DAS
- Attach Remote File
- Manage Data
- Features on Karyotype
- Data Converters
 - Assembly Converter
 - ID History Converter
 - Variant Effect Predictor
 - **Data Slicer**
 - Variation Pattern Finder
 - VCF to PED converter

i Data Slicer:

When slicing a VCF or BAM file, both the data file and its index file should be present on the web server and named correctly. The VCF file should have a ".vcf.gz" extension, and the index file should have a ".vcf.gz.tbi" extension, E.g: MyData.vcf.gz, MyData.vcf.gz.tbi. The BAM file should have a ".bam" extension, and the index file should have a ".bam.bai" extension, E.g: MyData.bam, MyData.bam.bai

Click [here](#) for more extensive documentation.

Upload files

VCF File URL:

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr1.phase1.projectConsensus.genotypes.vcf.gz
```

[Clear box](#)

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr1.phase1.projectConsensus.genotypes.vcf.gz

Region:

```
6:46620015-46620998
```

e.g. 1:1-50000

Use VCF filters (this doesn't apply to BAM files):

- None
- By individual(s)
- By population(s) *

(to filter by populations please provide URL to a Sample-Population Mapping File in the box below)

Sample-Population Mapping File URL:

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel
```

[Clear box](#)

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel

Next >



Variation Pattern Finder

- Remote or local tabix indexed VCF input
- Discovers patterns of Shared Inheritance
- Variants with functional consequences considered by default
- Web output with CSV and Excel downloads
- http://browser.1000genomes.org/Homo_sapiens/UserData/VariationsMapVCF



Variation Pattern Finder

Variation Pattern Finder:

The Variation Pattern Finder allows one to look for patterns of shared variation between individuals in the same vcf file. The finder looks for distinct variation combinations within the region, as well as individuals associated with each variation combination pattern. Only variants which have potentially functional consequences are considered, both intergenic and intronic snps are excluded. Click [here](#) for more extensive documentation.

The search will be performed on any VCF file you provided. It should be a URL for the file location. Please refer to <http://vcftools.sourceforge.net/specs.html> for VCF format specification. A URL for the latest VCF file for variation calls and genotypes released by the 1000 Genomes Project is displayed as an example below the input box. A mapping file between individual sample and population is required as well. The latest mapping file between individual sample and population released by the 1000 Genomes Project is displayed as well below the input box.

Upload files

VCF File URL:

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz
```

[Clear box](#)

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz

Sample-Population Mapping File URL:

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel
```

[Clear box](#)

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel

Region:

e.g. 6:46620015-46620998

Next >



Variation Pattern Finder

- **perl variant_pattern_finder.pl** -vcf ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_integrated_calls.20101123.snps_indels_svsvs.genotypes.vcf.gz -sample_panel_file ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel -region 6:31830969-31846823 -expand



Variation Pattern Finder Output

freq	6:31833647_[T]	6:31833660_rs6915800[G]	samples	
freq	ENST00000414427-SPLICE_SITE[],ENST00000544672-SPLICE_SITE[],ENST0000029729-SPLICE_SITE[],ENST00000375562-SPLICE_SITE[]	ENST00000414427-NON_SYNONYMOUS_CODING[R/C],ENST00000229729-NON_SYNONYMOUS_CODING[R/C],ENST00000544672-NON_SYNONYMOUS_CODING[R/C],ENST00000375562-NON_SYNONYMOUS_CODING[R/C]	samples	
0.73	REF REF	G A	YRI(3)	NA18933, NA19149, NA19098 and 0 others.
0.27	REF REF	A G	YRI(2)	NA19146, NA19198
0.18	REF REF	A A	LWK(1)	NA19372
0.09	C T	REF REF	CHB(1)	NA18592



VCF to PED

- LD Visualization tools like Haploview require PED files
- VCF to PED converts VCF to PED
- Will a file divide by individual or population
- http://browser.1000genomes.org/Homo_sapiens/UserData/Haploview



VCF to PED

Custom Data

Data Management

- Upload Data
- Attach DAS
- Attach Remote File
- Manage Data
- Features on Karyotype
- Data Converters
 - Assembly Converter
 - ID History Converter
 - Variant Effect Predictor
 - Data Slicer
 - Variation Pattern Finder
 - VCF to PED converter**

VCF to PED converter:

When providing a VCF file, both the data file and its index file should be present on the web server and named correctly. The VCF file should have a ".vcf.gz" extension, and the index file should have a ".vcf.gz.tbi" extension, E.g: MyData.vcf.gz, MyData.vcf.gz.tbi. Click [here](#) for more extensive documentation.

Upload files

VCF File URL:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz`

[Clear box](#)

e.g. `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz`

Sample-Population Mapping File URL:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel`

[Clear box](#)

e.g. `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel`

Region:

e.g. 6:46620015-46620998

[Next >](#)



VCF to PED

- **perl vcf_to_ped_convert.pl** -vcf ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_integrated_calls.20101123.snps_indels_svsvs.genotypes.vcf.gz -sample_panel_file ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel -region **6:31830969-31846823** -population **CEU**
- Output should be two files
- 6_31830969-31846823.info
- 6_31830969-31846823.ped



Haploview

- haploview



<http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview>



Access to backend Ensembl databases

- Public MySQL database at
 - `mysql-db.1000genomes.org` port 4272
- Full programmatic access with Ensembl API
 - The 1000 Genomes Pilot uses Ensembl v60 databases and the NCBI36 assembly (this is frozen)
 - The 1000 Genomes main project currently uses Ensembl v63 databases
- <http://jun2011.archive.ensembl.org/info/docs/api/variation/index.html>
- <http://www.ensembl.org/info/docs/api/variation/index.html>
- <http://www.1000genomes.org/node/517>



Data Availability

- FTP site: <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/>
 - Raw Data Files
- 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



Announcements

- <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>



Questions

Please send any questions about this presentation and any other material on our website to info@1000genomes.org



Thanks

- The 1000 Genomes Project Consortium
- Paul Flicek
- Richard Smith
- Holly Zheng Bradley
- Ian Streeter

