The 1000 Genomes Project
A History, Results and Tools
Laura Clarke
14th March 2012
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
The 1000 Genomes Project: Overview

- International project to construct a foundational data set for human genetics
  - Discover virtually all common human variations by investigating many genomes at the base pair level
  - Consortium with multiple centers, platforms, funders
- Aims
  - Discover population level human genetic variations of all types (95% of variation > 1% frequency)
  - Define haplotype structure in the human genome
  - Develop sequence analysis methods, tools, and other reagents that can be transferred to other sequencing projects
3 pilot coverage strategies

- **Trio**
  - Phased by transmission
  - Individual haploid genomes
  - Common haplotypes

- **Low coverage**
  - Statistical phasing
  - Common haplotypes

- **Exon**
  - Unphased
  - Exon variants
Main Project Design

• Based on the result of the pilot project, we decided to collect data on 2,500 samples from 5 continental groupings
  • Whole-genome low coverage data (>4x)
  • Full exome data at deep coverage (>20x)
  • A number of deep coverage genomes to be sequenced, with details to be decided
  • High density genotyping at subsets of sites using both Illumina Omni and Affymetrix Axiom

• Phase 1 Release Integrated Variant Release has been made.
Hapmap, The Pilot Project and The Main Project

- **Hapmap**
  - Starting in 2002
  - Last release contained ~3m snps
  - 1400 individuals
  - 11 populations
  - High Throughput genotyping chips

- **1000 Genomes Pilot project**
  - Started in 2008
  - Paper release contained ~14 million snps
  - 179 individuals
  - 4 populations
  - Low coverage next generation sequencing

- **1000 Genomes Phase 1**
  - Started in 2009
  - Phase 1 release has 36.6million snps, 3.8million indels and 14K deletions
  - 1094 individuals
  - 14 populations
  - Low coverage and exome next generation sequencing

- **1000 Genomes Phase 2**
  - Started in 2011
  - 1722 individuals
  - 19 Populations
  - Low coverage and exome next generation sequencing
Timeline

- **September 2007**: 1000 Genomes project formally proposed Cambridge, UK
- **April 2008**: First Submission of Data to the Short Read Archive.
- **May 2008**: First public data release.
- **October 2008**: SAM/BAM Format Defined.
- **December 2008**: First High Coverage Variants Released.
- **December 2008**: First 1000 genomes browser released
- **May 2009**: First Indel Calls released.
- **July 2009**: VCF Format defined
- **August 2009**: First Large Scale Deletions released.
- **December 2009**: First Main Project Sequence Data Released.
- **March 2010**: Low Coverage Pilot Variant Release made
- **July 2010**: Phased genotypes for 159 Individuals released.
- **October 2010**: A Map of Human Variation from population scale sequencing is published in Nature.
- **January 2011**: Final Phase 1 Low coverage alignments are released
- **May 2011**: @1000genomes appears on Twitter
- **May 2011**: First Variant Release made on more than 1000 individuals
- **October 2011**: Phase 1 integrated variant release made
- **March 2012**: Phase 2 Alignment release
Fraction of variant sites present in an individual that are **NOT** already represented in dbSNP

<table>
<thead>
<tr>
<th>Date</th>
<th>Fraction not in dbSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>February, 2000</td>
<td>98%</td>
</tr>
<tr>
<td>February, 2001</td>
<td>80%</td>
</tr>
<tr>
<td>April, 2008</td>
<td>10%</td>
</tr>
<tr>
<td>February, 2011</td>
<td>2%</td>
</tr>
<tr>
<td>Now</td>
<td>&lt;1%</td>
</tr>
</tbody>
</table>
Sequencing Data Evolution

- The Project contains data from 3 different providers and multiple platforms

<table>
<thead>
<tr>
<th>Platform</th>
<th>Min Read Length (bp)</th>
<th>Max Read Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 Roche GS FLX Titanium</td>
<td>70</td>
<td>400</td>
</tr>
<tr>
<td>Illumina GA</td>
<td>30</td>
<td>81</td>
</tr>
<tr>
<td>Illumina GA II</td>
<td>26</td>
<td>160</td>
</tr>
<tr>
<td>Illumina HiSeq</td>
<td>50</td>
<td>102</td>
</tr>
<tr>
<td>ABI Solid System 2.0</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>ABI Solid System 2.5</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>ABI Solid System 3.0</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>
1000 Genomes Project: Present & Future

• First Phase 2 sequence release 14th November 2011
• First Phase 2 alignment release now available
• First Phase 2 variant site release Summer 2012

• Sample collected expected end to June 2012
• Final Phase 3 Sequence release expected December 2012
• 2013 will represent finalization of 1000 genomes analysis results and final data releases
Pipelines for data processing and variant calling

- Tens of analysis groups have contributed
- Individual pipelines and component tools vary
- Typical main steps:
  - Read mapping
  - Duplicate filtering
  - Base quality value recalibration
  - INDEL realignment
  - Variant Site Discovery
  - Individual Genotype Assignment (sometimes part of site discovery)
  - Variant filtering / call set refinement
  - Variant reporting
Alignment Data

- The project has made more than 10 releases of Alignment Data
- Pilot Project
  - Aligned to NCBI36
  - Maq and Corona
  - Base Quality Recalibration done
- Phase 1
  - Aligned to GRCh37
  - BWA and Bfast
  - Indel Realignment
- Phase 2
  - Aligned to extended GRCh37
  - Improvements to Base Quality Recalibration
Variant Calling

- Early call sets used a single variant caller
- Intersect approach developed during pilot
- Variant Quality Score Recalibration (VQSR) developed for Phase 1
- Genotype Likelihoods assigned to help with genotype calling
- Integrated genotype calling based on individual variant call sets
- Phase 2 looks to improve site discovery and improve integration
Phase 1 analysis goal: an integrated view of human variations

- Reconstruct haplotypes including all variant types, using all datasets

Deletions

Indels

SNPs (from LC, EX, OMNI)
Deep coverage exome data is more sensitive to low-frequency variants

Allele count in 766 exomes (chr. 20, exons only)
Newly discovered SNPs are mostly at low frequency and enriched for functional variants

Enza Colonna, Yuan Chen, Yali Xue

Presentation on using the data for GWAS by Brian Howie
Phase 1 Integrated Release

- ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/
- Based on Low coverage and Exome data

- It contains
  - 38.2M SNPs
  - 1.44M Short Indels
  - 14K Deletions
  - Phased genotypes for 1094 individuals
Data Availability

- **FTP site**: [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/)
  - Raw Data Files
- **Web site**: [http://www.1000genomes.org](http://www.1000genomes.org)
  - Release Announcements
  - Documentation
- **Ensembl Style Browser**: [http://browser.1000genomes.org](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/

• EBI site has http mirror
  • http://ftp.1000genomes.ebi.ac.uk/vol1/ftp
ftp://ftp.1000genomes.ebi.ac.uk

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHANEGLOG</td>
<td>118 KB</td>
<td>05/01/2012 5:00:00</td>
</tr>
<tr>
<td>README.alignment_data</td>
<td>12 KB</td>
<td>26/01/2011 26:00</td>
</tr>
<tr>
<td>README.fish_structure</td>
<td>9 KB</td>
<td>04/04/2011 4:04</td>
</tr>
<tr>
<td>README.pilot_data</td>
<td>3 KB</td>
<td>14/07/2011 14:07</td>
</tr>
<tr>
<td>README.populations</td>
<td>2 KB</td>
<td>18/02/2010 18:02</td>
</tr>
<tr>
<td>README.sequence_data</td>
<td>7 KB</td>
<td>23/04/2011 23:04</td>
</tr>
<tr>
<td>alignment_indices</td>
<td></td>
<td>14/07/2011 14:07</td>
</tr>
<tr>
<td>changelog_details</td>
<td></td>
<td>04/04/2011 4:04</td>
</tr>
<tr>
<td>current.tree</td>
<td></td>
<td>05/01/2012 5:00:00</td>
</tr>
<tr>
<td>data</td>
<td>29933 KB</td>
<td>05/01/2012 5:00:00</td>
</tr>
<tr>
<td>phase1</td>
<td></td>
<td>14/07/2011 14:07</td>
</tr>
<tr>
<td>pilot_data</td>
<td></td>
<td>21/01/2012 21:01</td>
</tr>
<tr>
<td>release</td>
<td></td>
<td>27/12/2011 27:12</td>
</tr>
<tr>
<td>sequence.index</td>
<td>27185 KB</td>
<td>20/12/2011 20:12</td>
</tr>
<tr>
<td>sequence_indices</td>
<td></td>
<td>14/11/2011 14:11</td>
</tr>
<tr>
<td>technical</td>
<td></td>
<td>13/12/2011 13:12</td>
</tr>
</tbody>
</table>

- **Documentation**
- **Raw Data**
- **Phase 1 Data**
- **Pilot Data**
- **Release Data**
- **Technical Data**
The FTP Site: Data

Sample Level Files
- sequence_read
- alignment

Files:
- HG00104
- HG00105
- HG00106
- HG00107
- HG00108
- HG00109
- HG00110
- HG00111
- HG00112
- HG00113
- HG00114
- HG00115
- HG00116
- HG00117
- HG00118
- HG00119
- HG00120
- HG00121
- HG00122
- HG00123
- HG00124
- HG00125
- HG00126
- HG00127
- HG00128
- HG00129
- HG00130
- HG00131
## FTP Site: Release

### Date Format YYYYMMDD

- 2008_12
- 2009_02
- 2009_04
- 2009_05
- 2009_08
- 20100804
- 20101123
- 2010
- 2011
- 20110521

### Older Release Dirs

- 21/02/2009
- 07/05/2009
- 08/06/2009
- 10/08/2009
- 16/09/2010
- 16/02/2011
- 16/12/2011

### Sequence Index Dates

- 19/07/2011
- 19/07/2011
- 21/05/2011
- 19/07/2011
- 19/07/2011
- 19/07/2011
- 19/07/2011
- 19/07/2011
- 19/07/2011
FTP Site: Phase 1

Frozen Phase 1 Alignments
Finding Data

- FTP search
- [http://www.1000genomes.org/ftpsearch](http://www.1000genomes.org/ftpsearch)
- Search on the current.tree file
- Provides full ftp paths and md5 checksums
- Every page also has a website search box
http://www.1000genomes.org

**LATEST ANNOUNCEMENTS**

**Wednesday October 12, 2011**

**October 2011 Integrated Variant Set release #ICHG2011**

This October 2011 release represents an integrated set of variant calls and phased genotypes including SNPs, short INDELs and Deletions based on low coverage and exome sequencing data across 1092 individuals.

Our FAQ contains instructions on how to get smaller subsections of these files.

Data access links: EBI / NCBI

Link to additional information: README file

**Thursday June 23, 2011**

**June 2011 Data Release**

Genotypes for 1094 individuals for the May 2011 SNP calls from the 20101123 sequence and alignment release of the 1000 genomes project has now been made. This release is based on the GRCh37 assembly of the human genome and is released in the format VCF 4.0.

Our FAQ contains instructions on how to get smaller subsections of these files.

Data access links: EBI / NCBI

Link to additional information: README file
http://browser.1000genomes.org
Genes and SNPs

Line indicates number of SNPS
Each Line is One SNP
Turning on Tracks
File upload to view with 1000 Genomes data

Manage your data

- Supports popular file types:
  - BAM, BED, bedGraph, BigWig, GBrowse, Generic, GFF, GTF, PSL, VCF*, WIG
  
* VCF must be indexed
Uploaded VCF

Example:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/
ALL.wgs.phase1_release_v2.20101123.snps_indels_sv.sites.vcf.gz
Click the Gene tab, then ‘Variation Table’ or ‘Variation Image’

Gene Tab

Download as csv

Get in vcf format
Variation Image

- Gene variation zoom
Transcript Tab: Variations

Effect on Protein:
- SIFT
- PolyPhen
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 associations 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, 1081-1093.

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.
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:

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
<th>Online version</th>
<th>API script</th>
</tr>
</thead>
</table>
| 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 | 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 | API script |
| Data Slicer                 | Get a subset of data from a BAM or VCF file.                                | Online version (max 10K regions) | API script |
| 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 |
Data Slicing

**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:

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

Region:

```
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:**

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

Clear box
Thank you - your VCF file [6.31830969-31846823.ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz] [Size: 83436] has been generated. Right click on the file name and choose “Save link as ..” from the menu.

### Preview

```plaintext
##fileformat=VCFv4.0
##source=BCM:SNPTools:hapfuse
##reference=1000Genomes-NCBI37
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AP,Number=2,Type=Float,Description="Allelic Probability, P(Allele=1">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT
6 31831159 . rs3869144 C T 100 PASS . GT:AP
6 31831167 . T C 100 PASS . GT:AP
```
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:
    Variant_effect_predictor/
  - http://browser.1000genomes.org/Homo_sapiens/
    UserData/UploadVariations
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:
- 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:

Non-synonymous SNP predictions (human only)
- SIFT predictions:
- PolyPhen predictions:
- Condel consensus (SIFT/PolyPhen) predictions:

Frequency filtering of existing variants (human only)
- Filter variants by frequency:
  - Exclude variants with MAF greater than 0.1 in any 1KG low coverage population:

Next >
<table>
<thead>
<tr>
<th>Chromosome/Sample</th>
<th>Genotype</th>
<th>Transcript</th>
<th>Distance</th>
<th>Feature</th>
<th>1KG_6_3183357</th>
<th>SIFT</th>
<th>PolyPhen</th>
<th>Condel</th>
</tr>
</thead>
<tbody>
<tr>
<td>6_31833249_A/G</td>
<td>G</td>
<td>ENSG00000204385 ENST00000487680</td>
<td>Transcript</td>
<td>UPSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_31833249_A/G</td>
<td>G</td>
<td>ENSG00000204385 ENST00000414427</td>
<td>Transcript</td>
<td>DOWNSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_31833249_A/G</td>
<td>G</td>
<td>ENSG00000204385 ENST00000479777</td>
<td>Transcript</td>
<td>DOWNSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_31833249_A/G</td>
<td>G</td>
<td>ENSG00000204385 ENST00000475563</td>
<td>Transcript</td>
<td>DOWNSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>-</td>
<td>RegulatoryFeature</td>
<td>REGULATORY_REGION</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000495867</td>
<td>Transcript</td>
<td>UPSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000480384</td>
<td>Transcript</td>
<td>UPSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000491768</td>
<td>Transcript</td>
<td>UPSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000437561</td>
<td>Transcript</td>
<td>UPSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000479533</td>
<td>Transcript</td>
<td>UPSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000429729</td>
<td>Transcript</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>1625</td>
<td>1604</td>
<td>535</td>
<td>R/H</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000475562</td>
<td>Transcript</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>1673</td>
<td>1673</td>
<td>493</td>
<td>R/H</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000454672</td>
<td>Transcript</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>1544</td>
<td>1473</td>
<td>493</td>
<td>R/H</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000487680</td>
<td>Transcript</td>
<td>UPSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000414427</td>
<td>Transcript</td>
<td>DOWNSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000479777</td>
<td>Transcript</td>
<td>DOWNSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_3183357_C/T</td>
<td>T</td>
<td>ENSG00000204385 ENST00000475563</td>
<td>Transcript</td>
<td>DOWNSTREAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_31833612_C/G</td>
<td>G</td>
<td>ENSR000001987922</td>
<td>RegulatoryFeature</td>
<td>REGULATORY_REGION</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6_31833612_C/G</td>
<td>G</td>
<td>ENSG00000204386 ENST00000495807</td>
<td>Transcript</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>535</td>
<td>493</td>
<td>459</td>
<td>R/H</td>
</tr>
<tr>
<td>6_31833612_C/G</td>
<td>G</td>
<td>ENSG00000204386 ENST00000480384</td>
<td>Transcript</td>
<td>NON_SYNONYMOUS_CODING</td>
<td>535</td>
<td>535</td>
<td>459</td>
<td>R/H</td>
</tr>
</tbody>
</table>

SIFT: deleterious; PolyPhen: probably_damaging; Condel: deleterious
SIFT: deleterious; PolyPhen: possibly_damaging; Condel: deleterious
SIFT: deleterious; PolyPhen: probably_damaging; Condel: deleterious
SIFT: deleterious; PolyPhen: possibly_damaging; Condel: deleterious
SIFT: deleterious; PolyPhen: probably_damaging; Condel: deleterious
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](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.phased.projectConsensus.genotypes.vcf.gz
Clear box

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

Region:
e.g. 6:46620015-46620998
<table>
<thead>
<tr>
<th>Population</th>
<th>CEU</th>
<th>Freq</th>
<th>rs116706632:G/A</th>
<th>rs117127493:G/C</th>
<th>rs644827:T/C</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA20289, NA20296 and 13 other(s)</td>
<td>NA065</td>
<td>0.293</td>
<td>GIG</td>
<td>GIG</td>
<td>CIC</td>
</tr>
<tr>
<td>NA20127, NA19703 and 9 other(s)</td>
<td>NA125</td>
<td>0.203</td>
<td>GIG</td>
<td>GIG</td>
<td>CIC</td>
</tr>
<tr>
<td>NA20314, NA20317 and 6 other(s)</td>
<td>NA120</td>
<td>0.195</td>
<td>GIG</td>
<td>GIG</td>
<td>TIC</td>
</tr>
<tr>
<td>NA19920, NA19700 and 2 other(s)</td>
<td>0.032</td>
<td>GIG</td>
<td>GIG</td>
<td>GIG</td>
<td>CIC</td>
</tr>
<tr>
<td>NA19819, NA20281 and 2 other(s)</td>
<td>0.026</td>
<td>GIG</td>
<td>GIG</td>
<td>GIG</td>
<td>CIC</td>
</tr>
<tr>
<td>NA20291, NA20355 and 3 other(s)</td>
<td>0.016</td>
<td>GIG</td>
<td>GIG</td>
<td>GIG</td>
<td>TIC</td>
</tr>
<tr>
<td>NA19908</td>
<td>NA122</td>
<td>0.013</td>
<td>GIG</td>
<td>GIG</td>
<td>CIC</td>
</tr>
<tr>
<td>0.008</td>
<td>GIG</td>
<td>GIG</td>
<td>CIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.005</td>
<td>GIG</td>
<td>GIG</td>
<td>TIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA119</td>
<td>0.005</td>
<td>GIG</td>
<td>GIG</td>
<td>CIC</td>
<td></td>
</tr>
<tr>
<td>NA10916</td>
<td>0.004</td>
<td>GIG</td>
<td>GIG</td>
<td>CIC</td>
<td></td>
</tr>
<tr>
<td>NA19711, NA20340</td>
<td>0.003</td>
<td>GIG</td>
<td>GIG</td>
<td>CIC</td>
<td></td>
</tr>
<tr>
<td>NA119</td>
<td>0.003</td>
<td>GIG</td>
<td>GIG</td>
<td>CIC</td>
<td></td>
</tr>
</tbody>
</table>
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](http://browser.1000genomes.org/Homo_sapiens/UserData/Haploview)
VCF to PED

Upload files

VCF File URL:

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

Sample-Population Mapping File URL:

http://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 filter by population(s)

Select one or more populations from the scrollable list:

ASW
CEU
CHB
CHS
CLM
FIN
GBR
IBS
JPT
LWK

Next >

Your linkage pedigree and marker information files have been generated:
Right click on the file name and choose "Save link as .." from the menu:
Marker Information File  Linkage Pedigree File
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://www.ensembl.org/info/docs/api/variation/index.html](http://www.ensembl.org/info/docs/api/variation/index.html)
  • [http://www.1000genomes.org/node/517](http://www.1000genomes.org/node/517)
More Information

http://www.1000genomes.org/using-1000-genomes-data

Please email info@1000genomes.org with any questions
Announcements

- [http://1000genomes.org](http://1000genomes.org)
- [1000announce@1000genomes.org](mailto:1000announce@1000genomes.org)
- [http://www.1000genomes.org/1000-genomes-annoucement-mailing-list](http://www.1000genomes.org/1000-genomes-annoucement-mailing-list)
- [http://www.1000genomes.org/announcements/rss.xml](http://www.1000genomes.org/announcements/rss.xml)
- [http://twitter.com/#!/1000genomes](http://twitter.com/#!/1000genomes)
Thanks

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