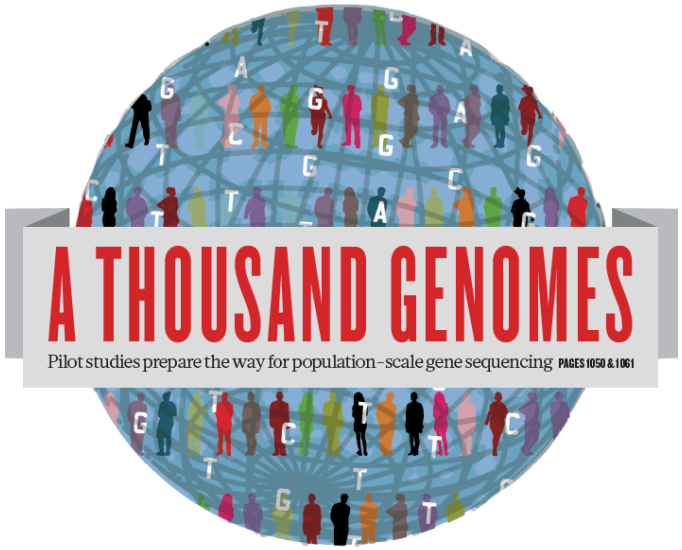


The 1000 Genomes Project: A Tutorial

Laura Clarke and Paul Flicek
EBI Training Room
24th January 2012

EMBL-EBI

Agenda

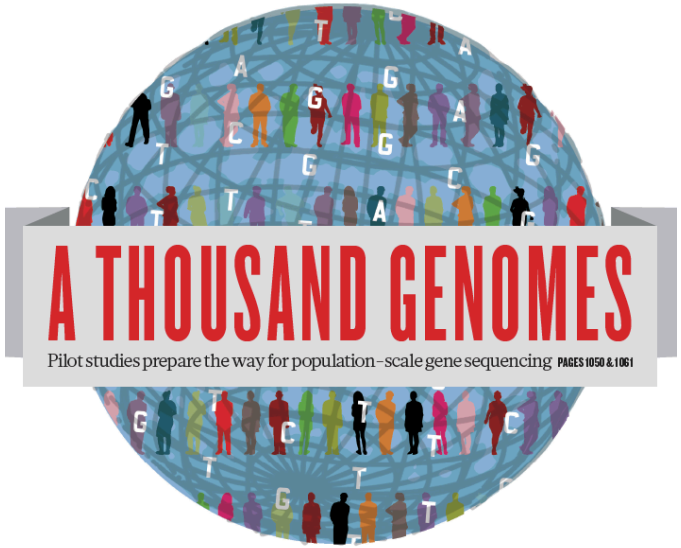
- Introduction
- Brief History of the 1000 Genomes Project, data and analysis
- The Raw Data and FTP site
- Finding Data
  - Exercise: Finding Data
- The Website and Browser
  - Exercise: Using the Browser
- The 1000 Genomes Tools
  - Exercise: Using the Tools

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

How are you using 1000 genomes data?

EMBL-EBI



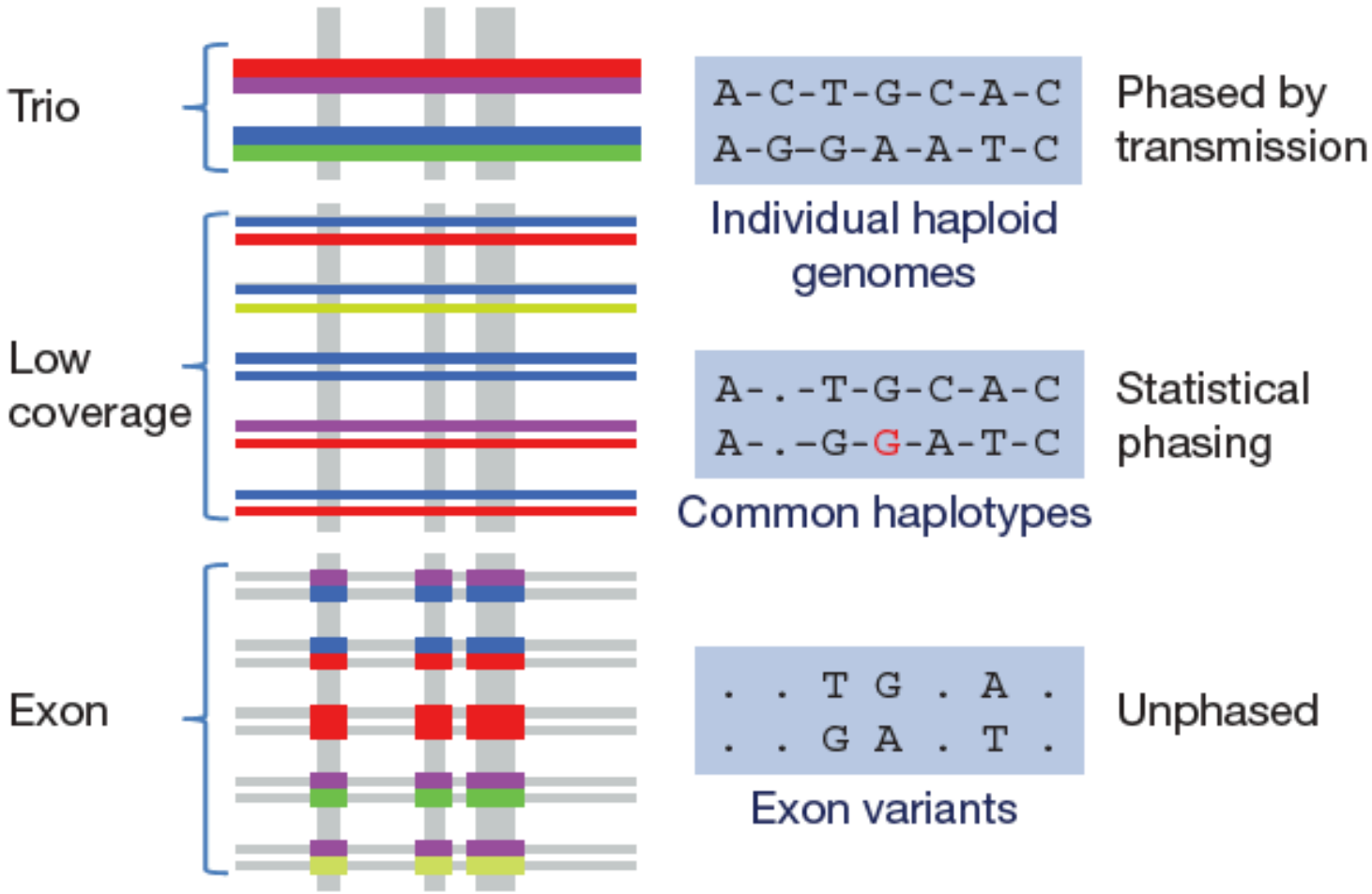
The 1000 Genomes Project: An Introduction

EMBL-EBI

The 1000 Genomes Project

- 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



A THOUSAND GENOMES

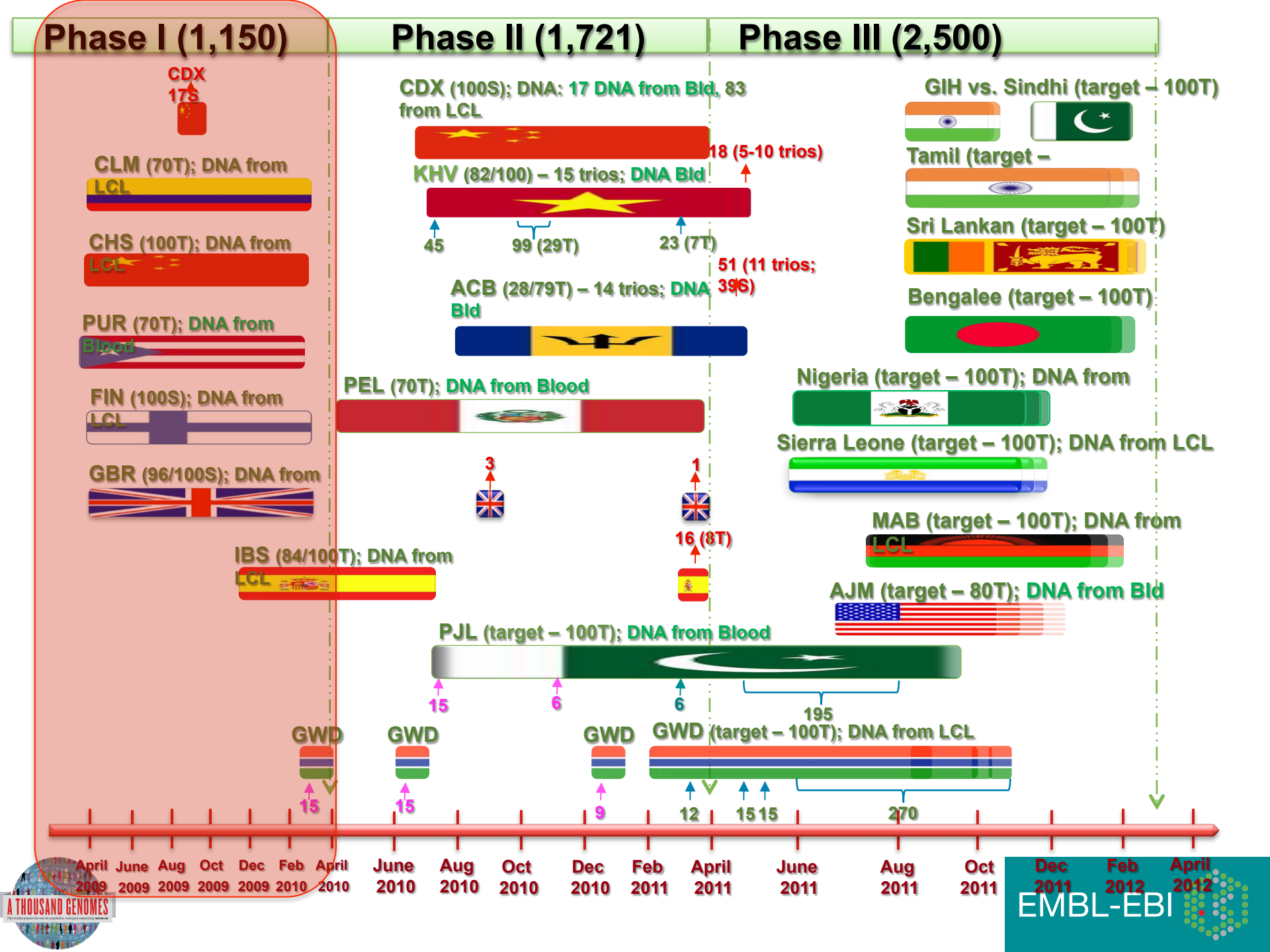
EMBL-EBI

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
- Phase 1 Release Integrated Variant Release has been made.

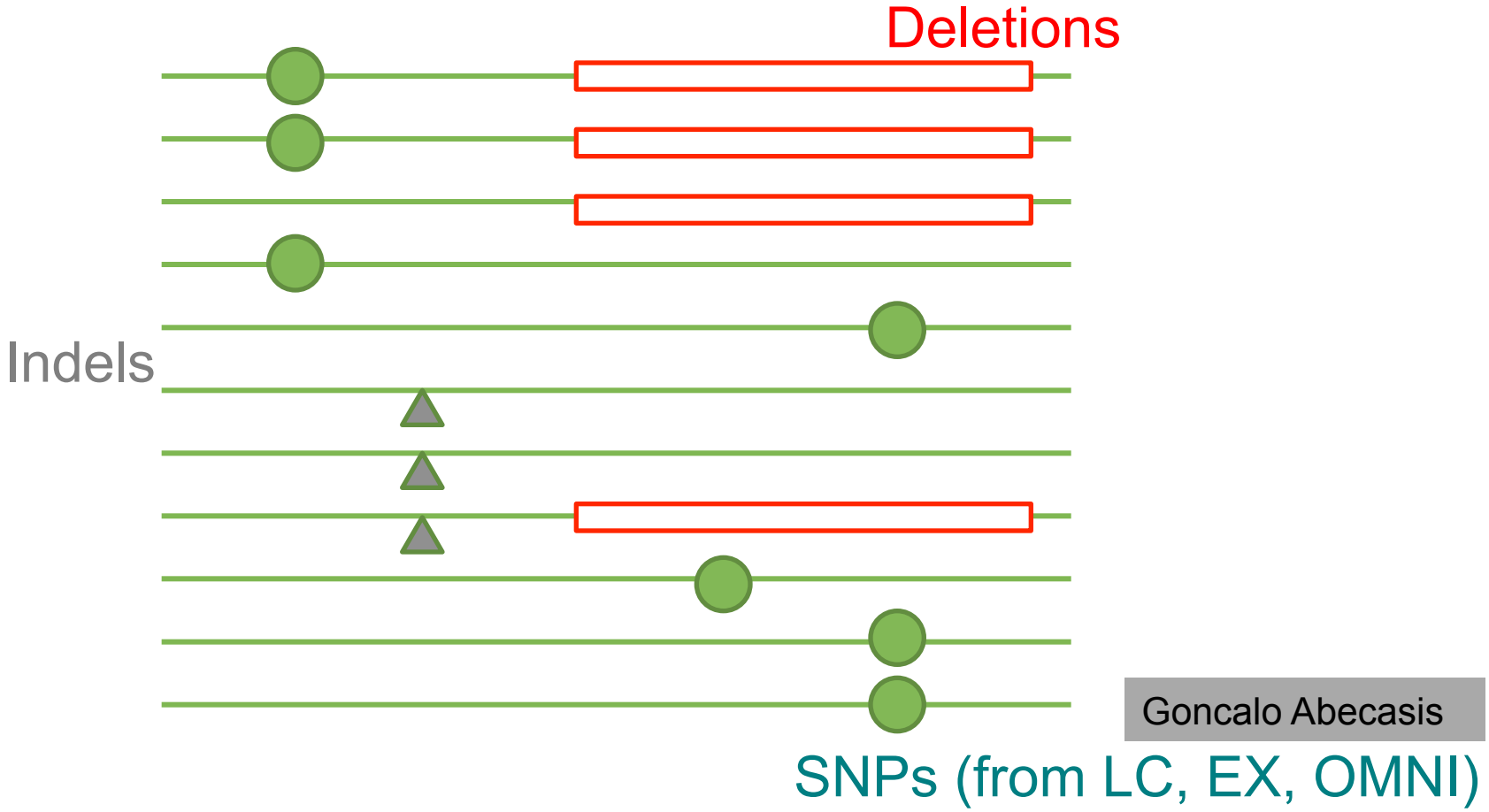
A THOUSAND GENOMES

EMBL-EBI


Phase I (1,150)
Phase II (1,721)
Phase III (2,500)
CDX 17S
CLM (70T); DNA from LCL
CHS (100T); DNA from LCL
PUR (70T); DNA from Blood
FIN (100S); DNA from LCL
GBR (96/100S); DNA from
IBS (84/100T); DNA from LCL
CDX (100S); DNA: 17 DNA from Bld, 83 from LCL
18 (5-10 trios)
KHV (82/100) – 15 trios; DNA Bld
45
99 (29T)
23 (7T)
51 (11 trios; 39S)
ACB (28/79T) – 14 trios; DNA Bld
PEL (70T); DNA from Blood
3
1
16 (8T)
PJL (target – 100T); DNA from Blood
15
6
6
195
GWD
GWD
GWD
GWD (target – 100T); DNA from LCL
15
15
9
12
15 15
270
GIH vs. Sindhi (target – 100T)
Tamil (target –
Sri Lankan (target – 100T)
Bengalee (target – 100T)
Nigeria (target – 100T); DNA from
Sierra Leone (target – 100T); DNA from LCL
MAB (target – 100T); DNA from LCL
AJM (target – 80T); DNA from Bld
April 2009
June 2009
Aug 2009
Oct 2009
Dec 2009
Feb 2010
April 2010
June 2010
Aug 2010
Oct 2010
Dec 2010
Feb 2011
April 2011
June 2011
Aug 2011
Oct 2011
Dec 2011
Feb 2012
April 2012
A THOUSAND GENOMES
EMBL-EBI

Phase 1 analysis goal: an **integrated view of human variations**

- Reconstruct haplotypes including all variant types, using all datasets

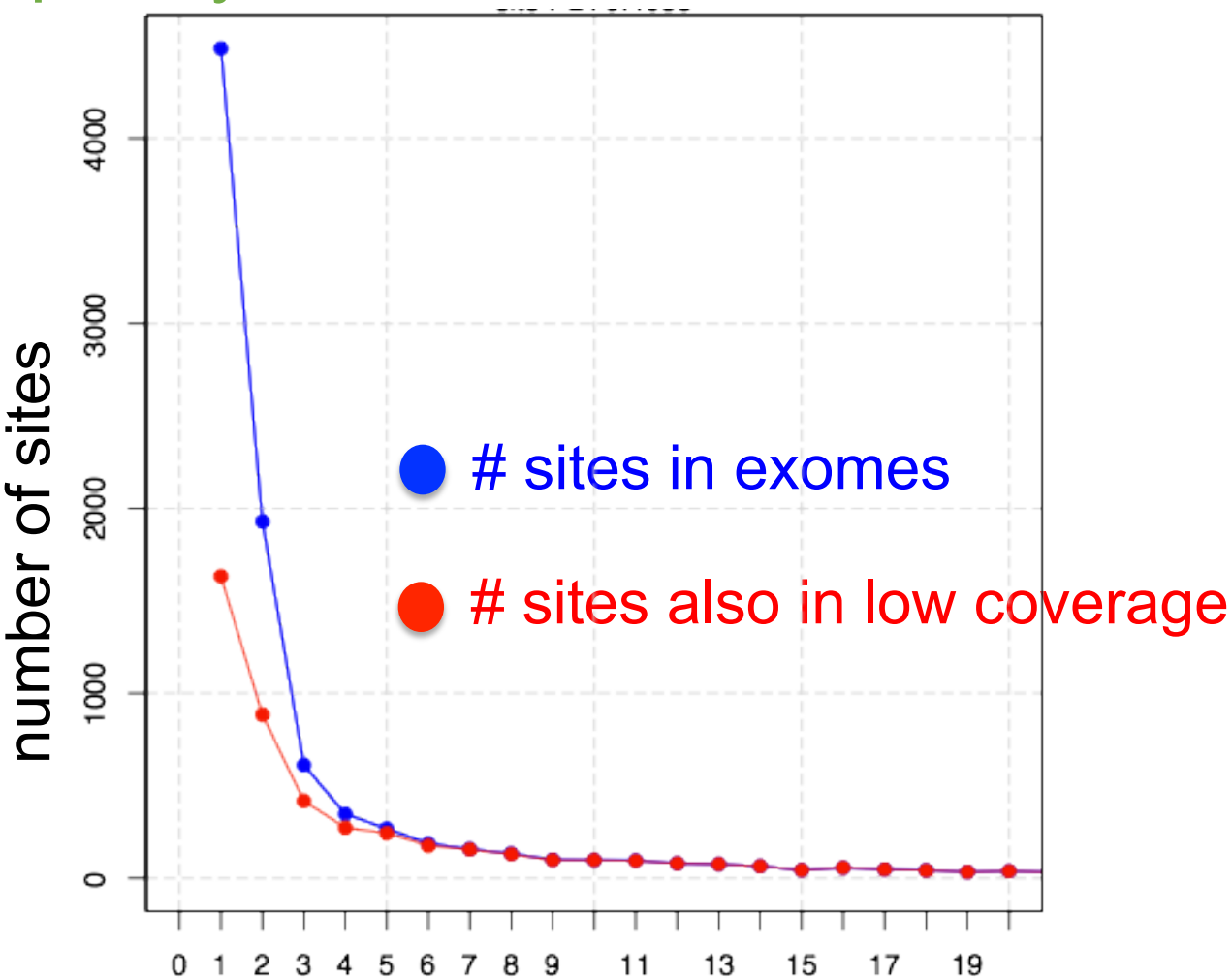


Goncalo Abecasis

A THOUSAND GENOMES

EMBL-EBI

Deep coverage exome data is more sensitive to low-frequency variants



Allele count in 766 exomes (chr. 20, exons only)

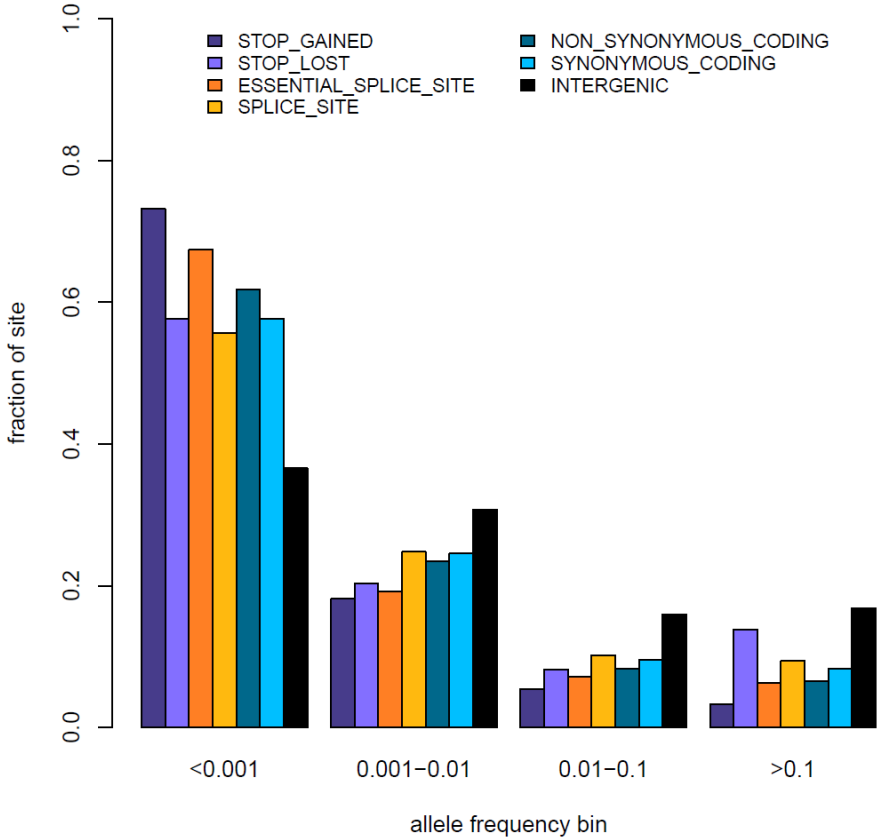
Erik Garrison

A THOUSAND GENOMES

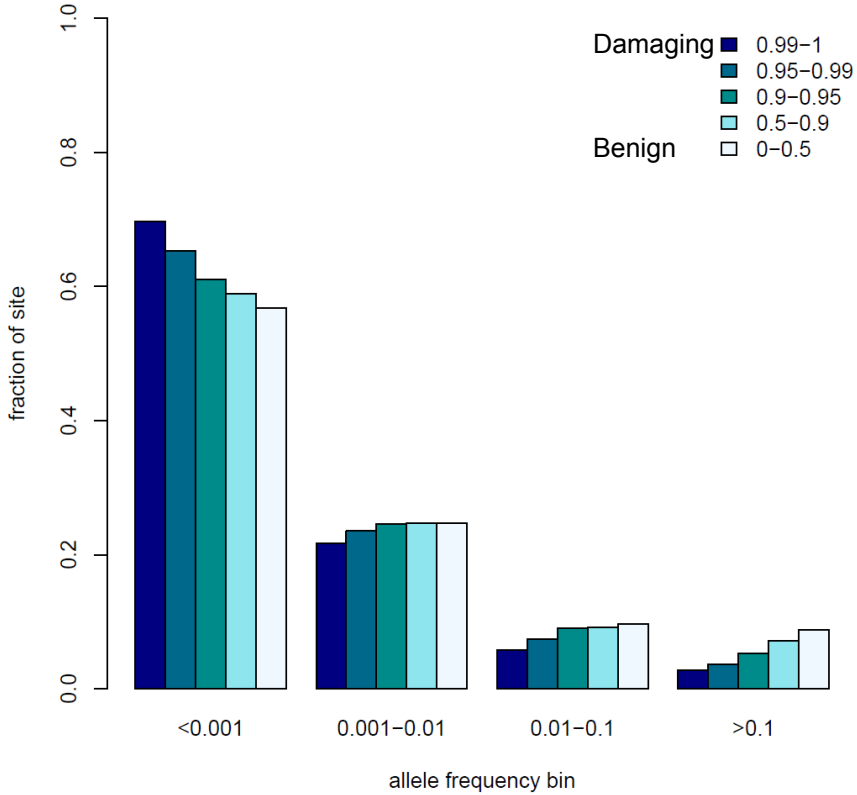
EMBL-EBI

Newly discovered SNPs are mostly at low frequency and enriched for functional variants

Functional category



Non-synonymous: Condel score



Presentation on using the data for GWAS by Brian Howie

Enza Colonna, Yuan Chen, Yali Xue

A THOUSAND GENOMES

EMBL-EBI

Fraction of variant sites present in an individual that are NOT already represented in dbSNP

Date	Fraction not in dbSNP
February, 2000	98%
February, 2001	80%
April, 2008	10%
February, 2011	2%
Now	<1%

Ryan Poplin, David Altshuler

A THOUSAND GENOMES

EMBL-EBI

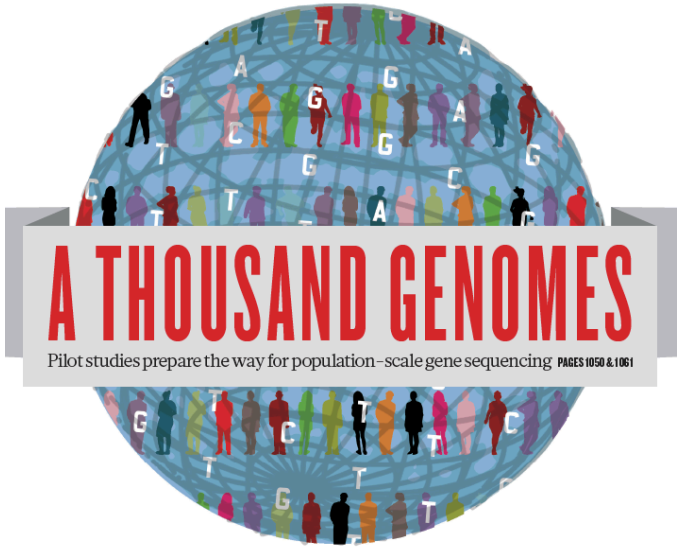
1000 Genomes Project: Present & Future

- First Phase 2 sequence release 14th November 2011
- First Phase 2 alignment release February 2012
- 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

A THOUSAND GENOMES

EMBL-EBI



The 1000 Genomes Project:
A Brief History of Data and Analysis Results

EMBL-EBI

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

A THOUSAND GENOMES

EMBL-EBI

Sequencing Data

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

Platform	Min Read Length (bp)	Max Read Length (bp)
454 Roche GS FLX Titanium	70	400
Illumina GA	30	81
Illumina GA II	26	160
Illumina HiSeq	50	102
ABI Solid System 2.0	25	35
ABI Solid System 2.5	50	50
ABI Solid System 3.0	50	50

Alignment Data

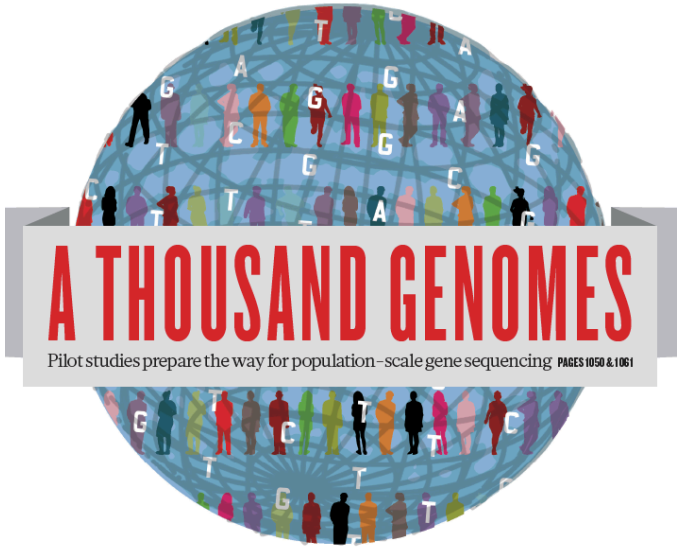
- The project has made more than 10 releases of Alignment Data
- Pilot Project
  - Aligned to NCBI36
  - Maq and Corona
  - Base Quality Recalibration done
- Phase 1
  - Aligned to GRCh37
  - BWA and Bfast
  - Indel Realignment
- Phase 2
  - Aligned to extended GRCh37
  - Improvements to Base Quality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

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



The 1000 Genomes Project: The Raw Data

EMBL-EBI

ftp://ftp.1000genomes.ebi.ac.uk

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp

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

Up to higher level directory

Name	Size	Last Modified
CHANGELOG	118 KB	05/01/2012 5/01/201212 :40:00
README.alignment_data	12 KB	26/01/2011 26/01/201112 :00:00
README.ftp_structure	9 KB	04/04/2011 4/04/201112 :00:00
README.pilot_data	3 KB	14/ 912 :00:00
README.populations	2 KB	18/02/2010 18/02/201012 :00:00
README.sequence_data	7 KB	23/ 19 :03:00
alignment_indices		14/07/2011 14/07/201110 :53:00
changelog_details		05/ 12 :40:00
current.tree	29933 KB	05/ 12 :37:00
data		04/ 8 :50:00
phase1		14/ 14 :03:00
pilot_data		27/ 012 :00:00
release		12/ 13 :18:00
sequence.index	27185 KB	20/12/2011 20/12/201112 :26:00
sequence_indices		14/11/2011 14/11/201110 :10:00
technical		13/12/2011 13/12/20111 :05:00

Documentation

Raw Data

Phase 1 Data

Pilot Data

Release Data

Technical Data

A THOUSAND GENOMES

EMBL-EBI

The FTP Site: Data

Index of ftp://ftp.1000genomes...

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

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am Later

HG00104	14/12/2011	14/12/201112 :06:00
HG00105	13/12/2011	13/12/20112 :45:00
HG00106	13/12/2011	13/12/20112 :45:00
HG00107	13/12/2011	13/12/20112 :40:00
HG00108	13/12/2011	13/12/20112 :43:00
HG00109		
HG00110	13/12/2011	13/12/20112 :43:00
HG00111	13/12/2011	13/12/20112 :36:00
HG00112		
HG00113	13/12/2011	13/12/20112 :41:00
HG00114	13/12/2011	13/12/20112 :41:00
HG00115		
HG00116		
HG00117	13/12/2011	13/12/20112 :43:00
HG00118	13/12/2011	13/12/20112 :44:00
HG00119	13/12/2011	13/12/20112 :38:00
HG00120	13/12/2011	13/12/20112 :43:00
HG00121	13/12/2011	13/12/20112 :37:00
HG00122	13/12/2011	13/12/20112 :45:00
HG00123	13/12/2011	13/12/20112 :43:00
HG00124	13/12/2011	13/12/20112 :44:00
HG00125	13/12/2011	13/12/20112 :36:00
HG00126	13/12/2011	13/12/20112 :39:00
HG00127	13/12/2011	13/12/20112 :39:00
HG00128	14/12/2011	14/12/201112 :06:00
HG00129	14/12/2011	14/12/201112 :06:00
HG00130	13/12/2011	13/12/20112 :46:00
HG00131	13/12/2011	13/12/20112 :44:00

Sample Level Files

sequence_read

alignment

A THOUSAND GENOMES

EMBL-EBI

FTP Site: Technical

Index of ftp://ftp.1000genomes...

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am Late

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

Up to higher level directory

Name	Size	Last Modified
README.reference	1 KB	12/10/2009 12/10/200912 :00:00
browser		19/12/2011 19/12/20113 :50:00
method_development		06/06/2011 6/06/201112 :00:00
ncbi_varpipe_data		
other_exome_alignments.alignment_indices		20/07/2011 20/07/201112 :00:00
pilot2_high_cov_GRCh37_bams		11/01/2012 11/01/20125 :56:00
pilot3_exon_targetted_GRCh37_bams		
qc		
reference		
retired_reference		
simulations		04/05/2010 4/05/201012 :00:00
supporting		21/12/2009 21/12/200912 :00:00
working		17/01/2012 17/01/20124 :07:00

Alternative Alignments

Reference Data Sets

Experimental Data

FTP Site: Release

Index of ftp://ftp.1000genomes...

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

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am

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

Up to higher level directory

Name	Size	Last Modified
2008_12		2 :00:00
2009_02		21/02/2009 21/02/200912 :00:00
2009_04		07/05/2009 7/05/200912 :00:00
2009_05		08/06/2009 8/06/200912 :00:00
2009_08		10/08/2009 10/08/200912 :00:00
20100804		0:00
20101123		0:00
2010_11		16/02/2011 16/02/201112 :00:00
20110521		16/12/2011 16/12/201110 :09:00

Older Release Dirs

Sequence Index Dates

20110521.sequence.index	23693 KB	19/07/2011	19/07/201112 :00:00
20110521.sequence.index.exome.stats	48 KB	19/07/2011	19/07/201112 :00:00
20110521.sequence.index.low_coverage.stats	53 KB	21/05/2011	21/05/201112 :00:00
20110521_20110719.exome.stats.csv	2 KB	19/07/2011	19/07/201112 :00:00
20110521_20110719.low_coverage.stats.csv	2 KB	19/07/2011	19/07/201112 :00:00
20110719.sequence.index	23961 KB	19/07/2011	19/07/201112 :00:00
20110719.sequence.index.exome.stats	52 KB	10/10/2011	10/10/201110 :10:00
20110719.sequence.index.low_coverage.stats	54 KB	10/10/2011	10/10/201110 :13:00
20110719_20110920.exome.stats.csv	1 KB	10/10/2011	10/10/20119 :45:00
20110719_20110920.low_coverage.stats.csv	2 KB	10/10/2011	10/10/20119 :45:00

A THOUSAND GENOMES

EMBL-EBI

FTP Site: Pilot Data

Index of ftp://ftp.1000genomes...

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am

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

Up to higher level directory

Name	Size	Last Modified	
README.alignment.index	2 KB	26/08/2009	26/08/200912 :00:00
README.bas	3 KB	27/08/2009	27/08/200912 :00:00
README.sequence.index	2 KB	22/07/2009	22/07/200912 :00:00
SRP000031.sequence.index	7365 KB	12/07/2010	12/07/201012 :00:00
SRP000032.sequence.index	2181 KB	12/07/2010	12/07/201012 :00:00
SRP000033.sequence.index	480 KB	12/07/2010	12/07/201012 :00:00
data			0:00
paper_data_sets		03/02/2011	3/02/201112 :00:00
pilot_data.alignment.index	795 KB	06/05/2010	6/05/201012 :00:00
pilot_data.alignment.index.bas.gz	1740 KB	14/06/2010	14/06/201012 :00:00
pilot_data.sequence.index	10025 KB	12/07/2010	12/07/201012 :00:00
release		20/07/2010	20/07/201012 :00:00
technical		29/07/2010	29/07/201012 :00:00

Final Paper Data

A THOUSAND GENOMES

EMBL-EBI

Data formats and key tools

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 16 2009, pages 2078–2079
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009
Advance Access publication June 8, 2009
Associate Editor: Alfonso Valencia

BAM alignment files

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2156–2158
doi:10.1093/bioinformatics/btr330

Sequence analysis Advance Access publication June 7, 2011

The variant call format and VCFtools

Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, [5]Department of Biology, Boston College, MA 02467, [6]National Institutes of Health National Center for Biotechnology Information, MD 20894, USA and [7]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

Associate Editor: John Quackenbush

VCF variant files

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 5 2011, pages 718–719
doi:10.1093/bioinformatics/btq671

Sequence analysis Advance Access publication January 5, 2011

Tabix: fast retrieval of sequence features from generic TAB-delimited files

Heng Li

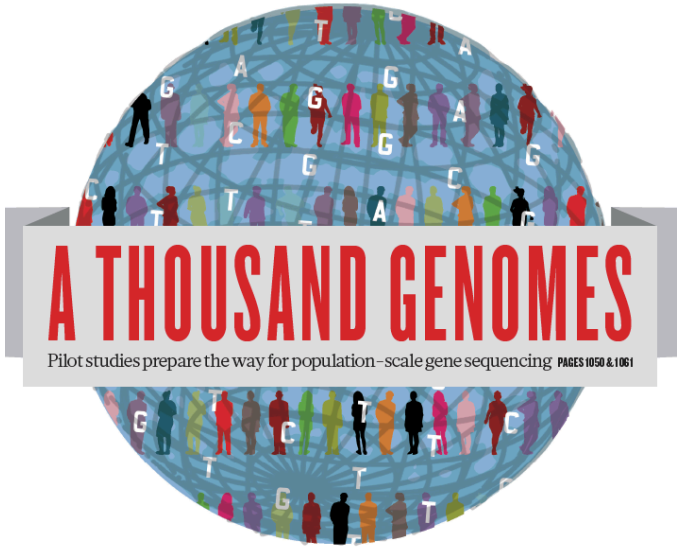
Program in Medical Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

Associate Editor: Dmitrij Frishman

All indexed for fast retrieval

A THOUSAND GENOMES

EMBL-EBI



The 1000 Genomes Project: Finding Data

EMBL-EBI

Finding Data

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

ftp://ftp.1000ge...ftp/current.tree

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

Maps dates fix lj docs plus gm g fb ds lj NCBI p E SRA C 1KG JIRA Wish am Later

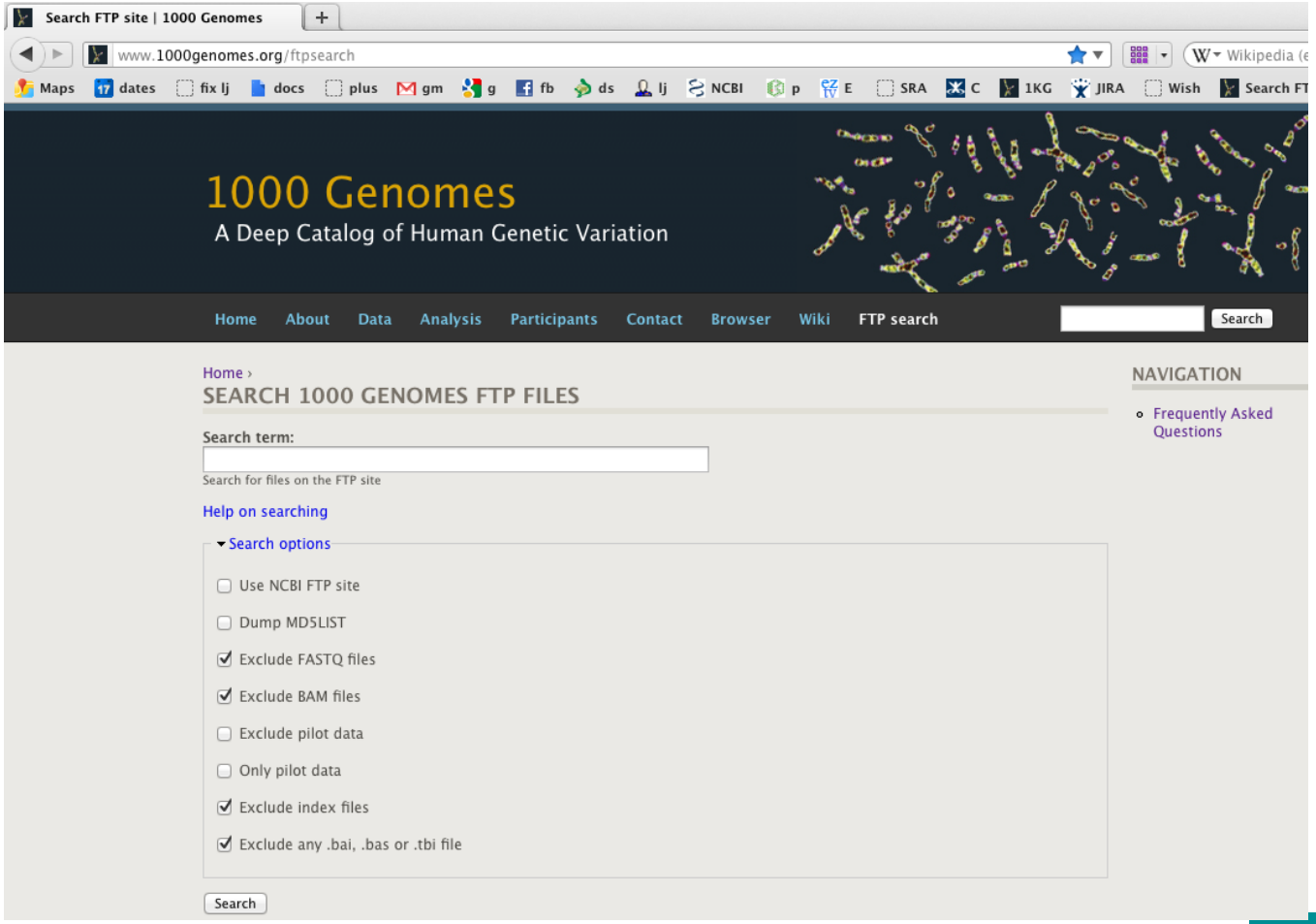
```
ftp     directory       403     Tue Dec 20 16:11:25 2011
ftp/README.ftp_structure        file    8408    Mon Apr  4 14:52:52 2011        2a59a3feb2540c113e10877f3ef1efe5
ftp/README.populations  file    1506    Wed Jan 11 15:12:44 2012        f7c588af82396013c1737e66e58f0f05
ftp/CHANGELOG   file    122151  Sat Jan 14 23:51:50 2012        ecaa9b1e0a6860cd76b1545e84ff3403
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/phase1      directory       412     Wed Dec 14 16:03:36 2011
ftp/phase1/phase1.exome.alignment.index.HsMetrics.stats file    293     Wed Dec 14 15:53:53 2011        1ebf793046daadd7ff67ecebb1b5361f
ftp/phase1/phase1.exome.alignment.index file    397947  Wed Dec 14 15:53:52 2011        2891d1fffe08acf3ee99c88cb42d130d
ftp/phase1/phase1.alignment.index.bas.gz        file    5115518 Wed Dec 14 15:53:23 2011        2b4e1edb78f617ebfaf5087536d80f95
ftp/phase1/phase1.alignment.index       file    8850348 Wed Dec 14 15:53:22 2011        ea3423858ec976a1fe17839cd334c164
ftp/phase1/phase1.exome.alignment.index.bas.gz  file    423691  Wed Dec 14 15:53:52 2011        7a56f22d28e860fbc65b71d1013717ae
ftp/phase1/phase1.exome.alignment.index.HsMetrics.gz    file    143893  Wed Dec 14 15:53:53 2011        93ba34ab86e9c42198919d128acc13b7
ftp/phase1/phase1.exome.alignment.index_stats.csv       file    715     Wed Dec 14 15:53:53 2011        376ea20314a94399cab99c723e1d974c
ftp/phase1/technical/ncbi_varpipe_data  directory       137     Wed Dec 14 16:16:31 2011
ftp/phase1/technical/ncbi_varpipe_data/phase1.ncbi.20100804.alignment.summary   file    39866   Wed Dec 14 16:13:58 2011        df4676c95ed2cc6f9cd4c9e24a66bbe8
ftp/phase1/technical/ncbi_varpipe_data/phase1.ncbi.20100804.alignment.index     file    159169  Wed Dec 14 16:13:58 2011        a9bc22ace39cb0bcd0bf35f2ee807bbc
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA12004        directory       308     Tue Dec 13 12:16:47 2011
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.chrom20.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam  file    238645793       Thu Apr 14 15:24
ftp/phase1/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/phase1/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.chrom20.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam.bai      file    166624  Thu Apr 14 15:24
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA12004/NA12004.ILLUMINA.mosaik.CEU.low_coverage.20100804.bam  file    11091314322     Wed Oct 27 18:31:24 2010
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA18486        directory       308     Tue Dec 13 12:25:36 2011
ftp/phase1/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/phase1/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam  file    29068330549     Tue Jan 25 22:46:53 2011
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.chrom20.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam.bai      file    176848  Tue Jan 25 22:47
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA18486/NA18486.chrom20.ILLUMINA.mosaik.YRI.low_coverage.20101123.bam  file    685641416       Tue Jan 25 22:47
ftp/phase1/technical/ncbi_varpipe_data/alignment/NA12045        directory       604     Tue Dec 13 12:24:58 2011
```

A THOUSAND GENOMES

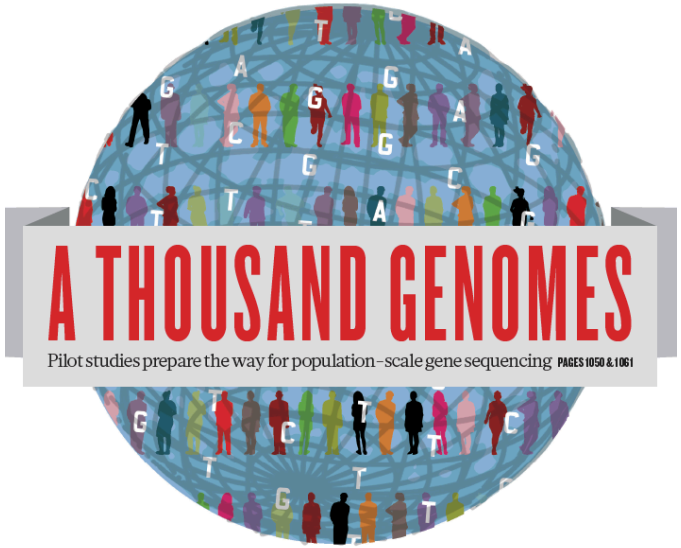
EMBL-EBI

Finding Data

- FTP search
- http://www.1000genomes.org/ftpsearch



EMBL-EBI



The 1000 Genomes Project:

Exercise 1: Finding Data on the 1000 genomes ftp site

EMBL-EBI

Finding Data: Exercise

1. Find what VCF files we have containing genotypes from the Illumina Omni platform.

   http://www.1000genomes.org/ftpsearch

2. Find the FAQ question which gives you instructions on how to get a sub-section of a VCF file. The Search Box is on the top right hand corner of any website page.

   http://www.1000genomes.org/

Finding Data: Exercise

1. Finding Omni VCF Files

Home About Data Analysis Participants Contact Browser Wiki FTP search Search

Home ›

SEARCH 1000 GENOMES FTP FILES

Search term:

omni*vcf

Search for files on the FTP site

Help on searching

Search options

- [ ] Use NCBI FTP site
- [ ] Dump MD5LIST
- [x] Exclude FASTQ files
- [x] Exclude BAM files
- [ ] Exclude pilot data
- [ ] Only pilot data
- [x] Exclude index files
- [x] Exclude any .bai, .bas or .tbi file

Search

LAURA@EBI.AC.UK

- My account
- Create content
- List content
- List users
- Manage files
- Log out
- Frequently Asked Questions

A THOUSAND GENOMES

EMBL-EBI

Finding Data: Exercise

Home About Data Analysis Participants Contact Browser Wiki FTP search

Home ›

SEARCH 1000 GENOMES FTP FILES

Search term:

omni*vcf

Search for files on the FTP site

Help on searching

Search options

Search

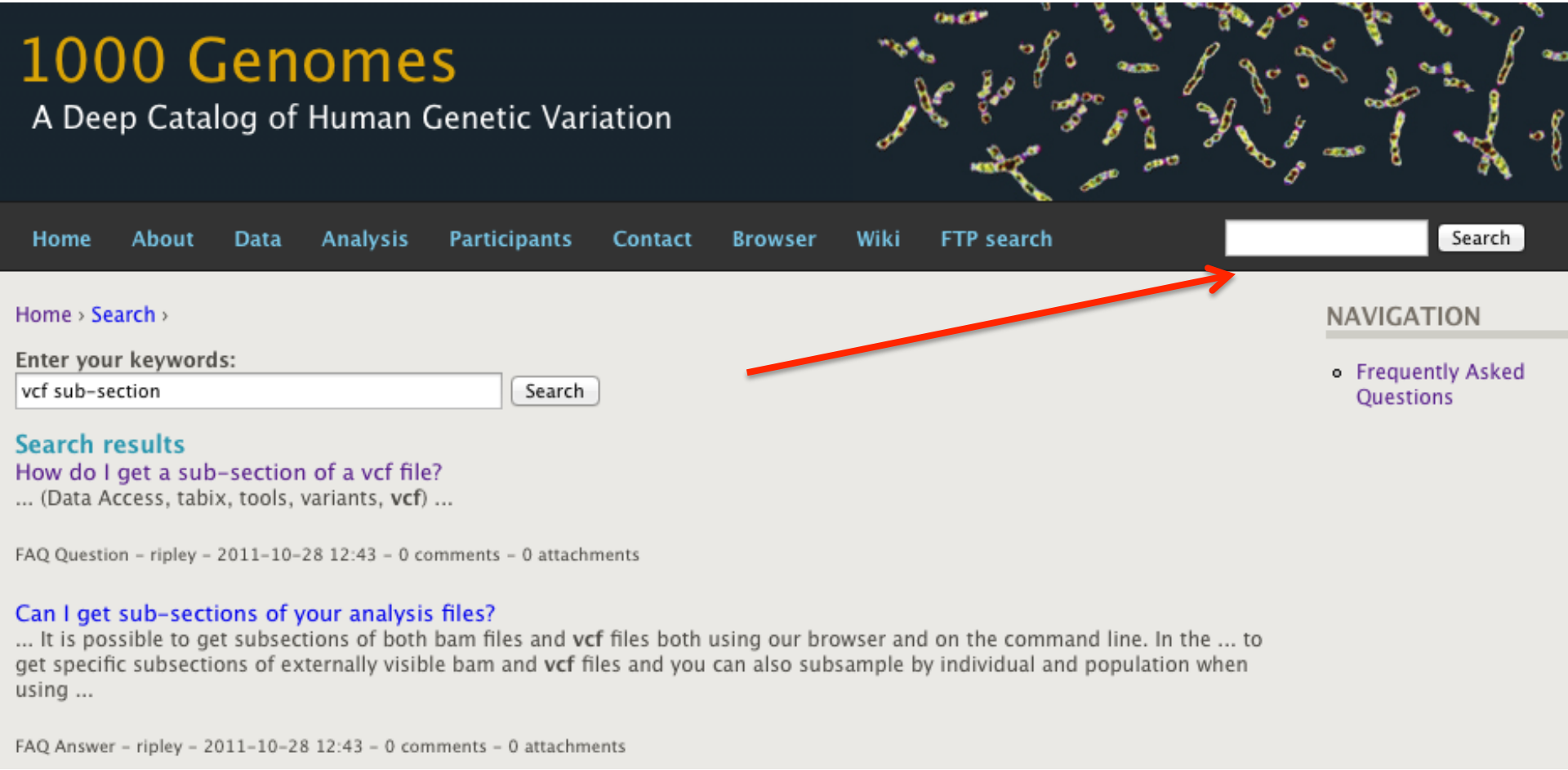
RESULTS

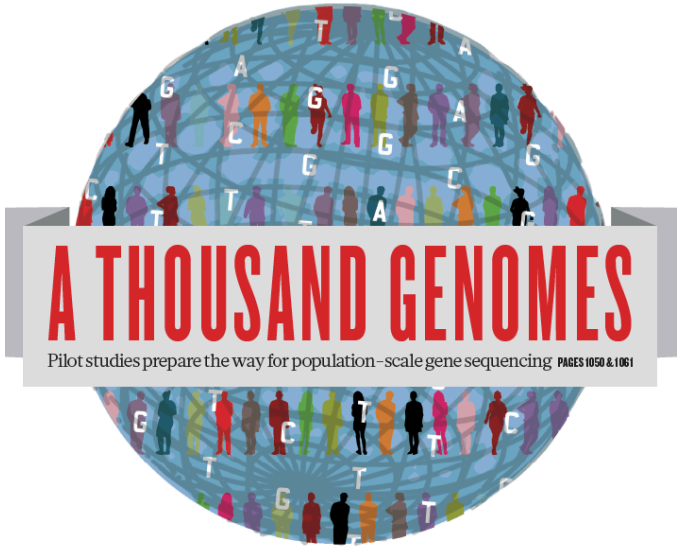
50 files found

File
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr20.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr15.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr4.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr9.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr8.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr12.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz

Finding Data: Exercise

- Finding help on getting sub-sections of VCF files



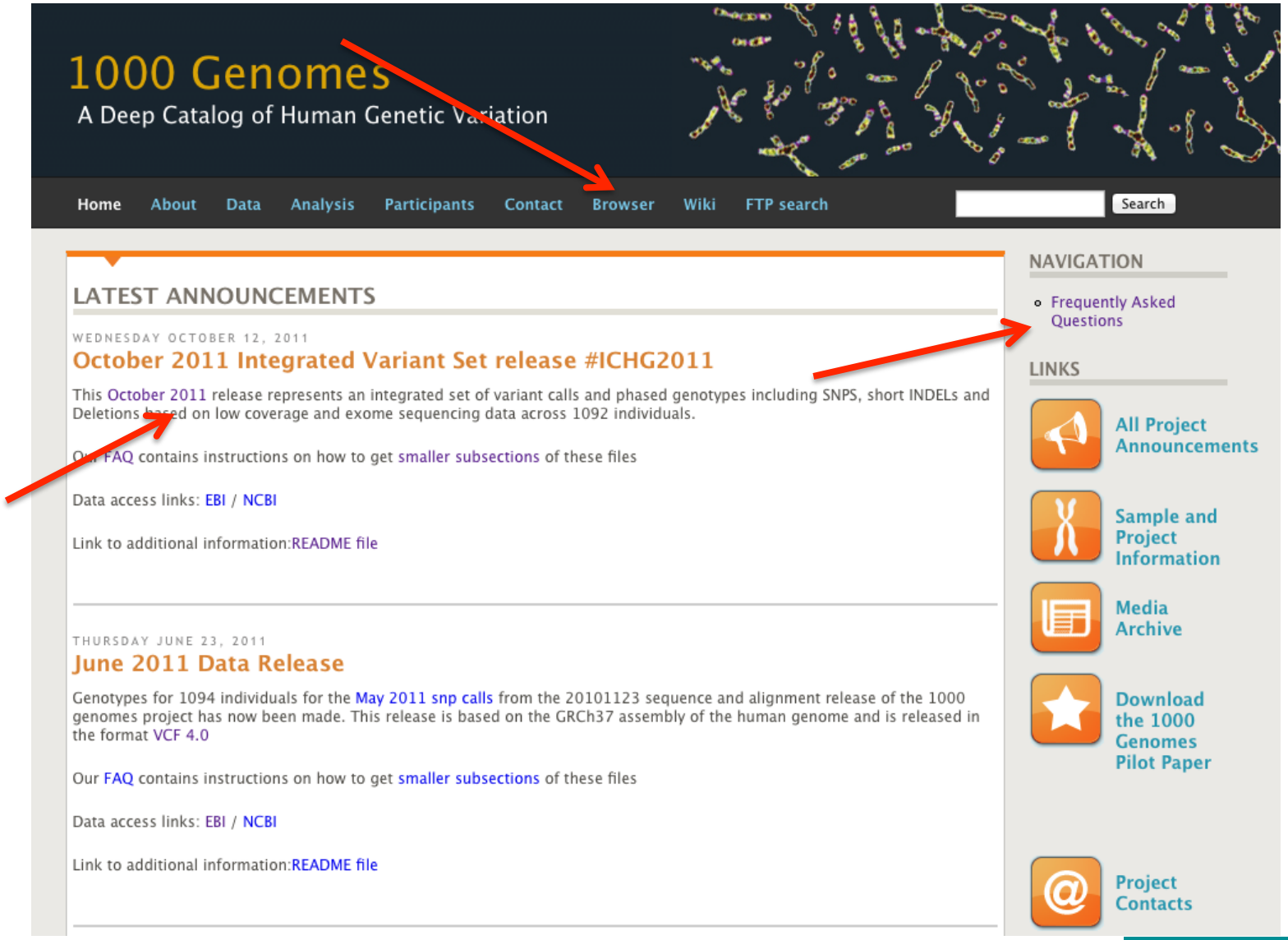


The 1000 Genomes Project:

The 1000 Genomes Website and Ensembl-style 1000 Genomes Browser

EMBL-EBI

http://www.1000genomes.org



A THOUSAND GENOMES

EMBL-EBI

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

1000 Genomes release 10 - October 2011 © EBI

About 1000 Genomes | Contact Us | Help

http://browser.1000genomes.org

A THOUSAND GENOMES

EMBL-EBI

1000 Genomes

A Deep Catalog of Human Genetic Variation

Human (GRCh37) | Location: 1:114,356,433-114,414,381 | Gene: PTPN22 | Transcript: PTPN22-001

Tools | Help

Location-based displays

- Whole genome
- Chromosome summary
- Region overview
- Region in detail
- Genetic Variation
  - Resequencing (20)
  - Linkage Data
- Markers

Configure this page

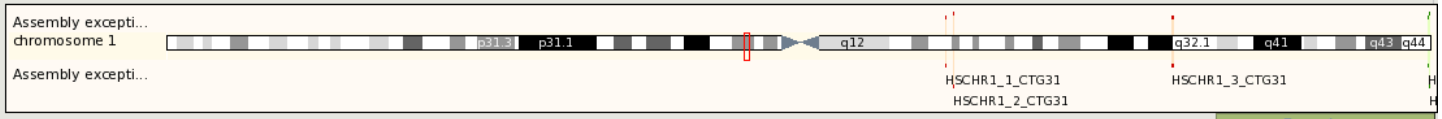
Manage your data

Export data

Get VCF data

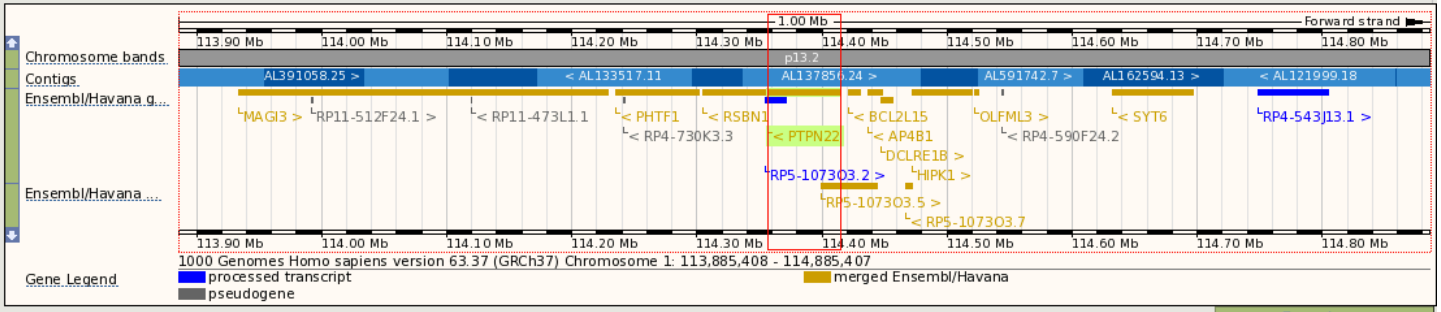
Bookmark this page

Chromosome 1: 114,356,433-114,414,381



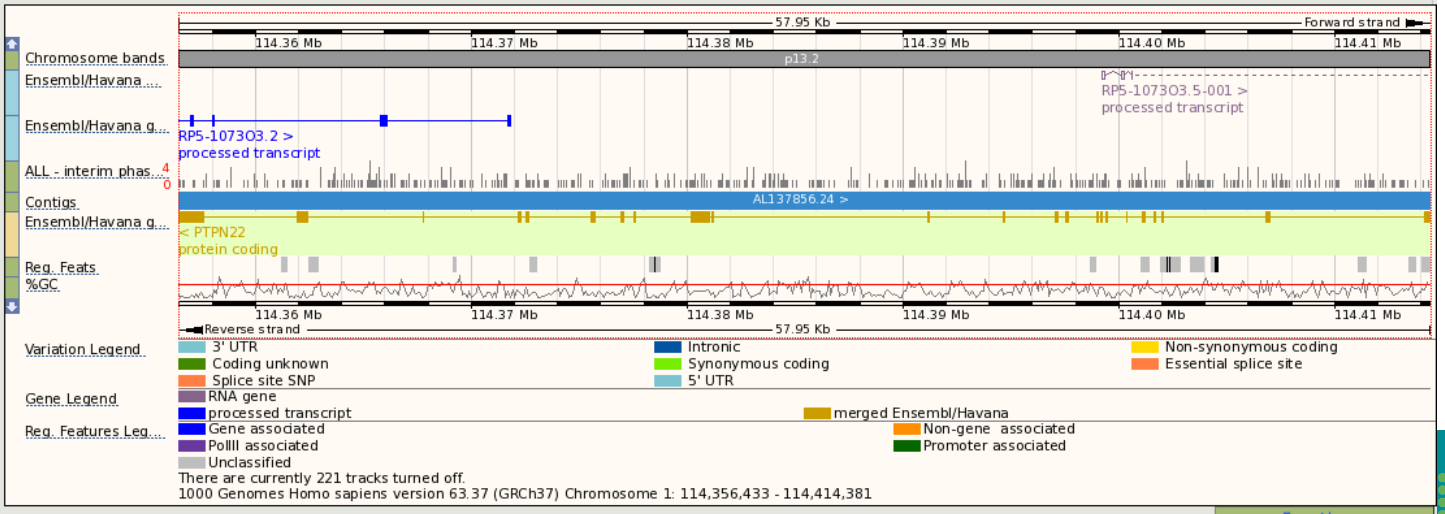
Export Image

Region in detail help



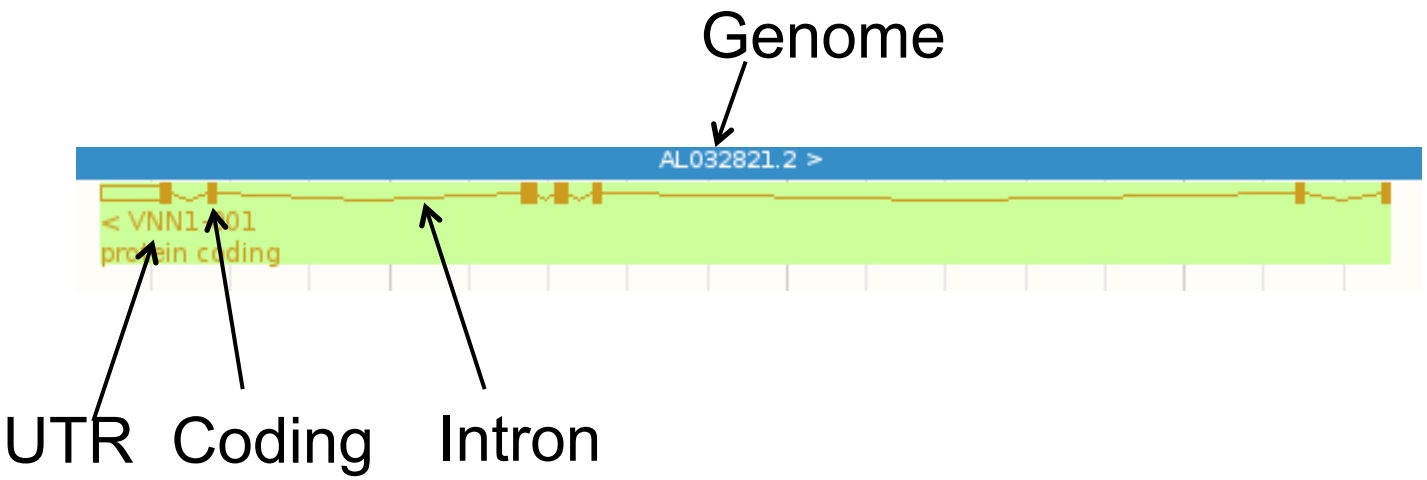
Export Image

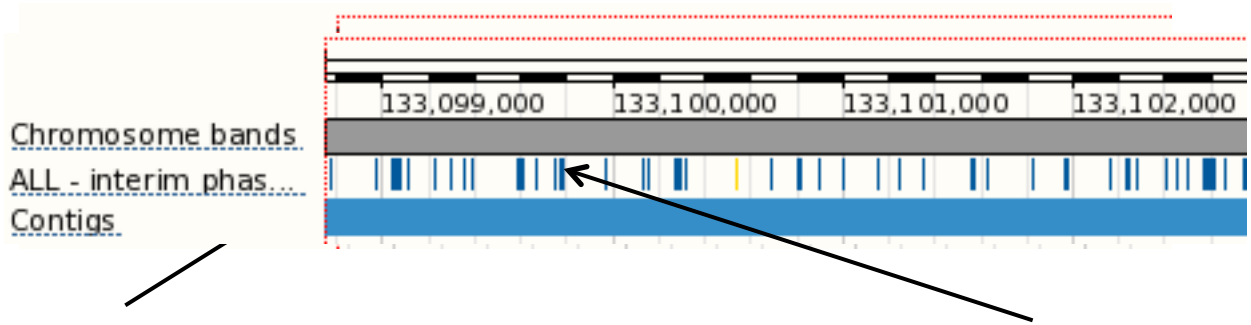
Location: 1:114356433-114414381 Go Gene: Go



Export Image

What to Look For





- Gene variation zoom

- Population

1000 Genomes

A Deep Catalog of Human Genetic Variation

Human (GRCh37) | Location: 6:74,125,388-74,126,388 | Variation: rs311685 | Tools | Help

Variation displays
- Flanking sequence
- Gene/Transcript (3)
- Population genetics (46)
- Individual genotypes (2769)
- Genomic context
- Phenotype Data
- Phylogenetic Context
- External Data

Configure this page
Manage your data
Export data
Get VCF data
Bookmark this page
Download view as CSV

Variation: rs311685

Variation class	SNP (rs311685 source dbSNP 132 - Variants (including SNPs and indels) imported from dbSNP [http://www.ncbi.nlm.nih.gov/projects/SNP/])
Synonyms	**Affy GeneChip 100K Array** SNP_A-1679873 **Affy GenomeWideSNP_6.0** AFFY_6_1M_SNP_A-8668494, SNP_A-8668494 **dbSNP** rs58378291, rs17756820, rs52794514, rs524803, rs3173186, rs11567000, rs17421786 **ENSEMBL** ENSSNP9062281 **Illumina_Human1M-duoV3** rs311685 **Uniprot** VAR_057235
Present in	1000 genomes - High coverage - Trios (1000 genomes - High coverage - Trios - CEU, 1000 genomes - High coverage - Trios - YRI),1000 genomes - Low coverage (1000 genomes - Low coverage - CEU, 1000 genomes - Low coverage - CHB+JPT, 1000 genomes - Low coverage - YRI),ALL - interim phase 1 - 1000 Genomes (AFR - interim phase 1 - 1000 Genomes, AMR - interim phase 1 - 1000 Genomes, ASN - interim phase 1 - 1000 Genomes, EUR - interim phase 1 - 1000 Genomes),ENSEMBL:Venter,HapMap
Alleles	**A/G** (Ambiguity code: **R**)
Ancestral allele	A
Location	This feature maps to 6:74125888 (forward strand) | View in location tab
Validation status	Proven by **cluster, frequency, doublehit, 1000Genome HapMap variant**
HGVS names	This feature has 4 HGVS names - click the plus to show

1000 genomes alleles frequencies

AFR	ALL	AMR	ASN	EUR
A: 69%	A: 51%	A: 54%	A: 45%	A: 42%
G: 31%	G: 49%	G: 46%	G: 55%	G: 58%

1000 genomes

Show/hide columns | Filter | CSV

Population	Alleles A	Alleles G	Genotypes A|A	Genotypes A|G	Genotypes G|G	Count
1000GENOMES:AFR	0.689	0.311	0.463	0.451	0.085	114
1000GENOMES:ALL	0.507	0.493	0.269	0.477	0.254	294
1000GENOMES:AMR	0.539	0.461	0.293	0.492	0.215	53
1000GENOMES:ASN	0.446	0.554	0.199	0.493	0.308	57
1000GENOMES:EUR	0.421	0.579	0.184	0.475	0.341	70

1000 genomes pilot

Show/hide columns | Filter | CSV

Population	ssID	Submitter	Alleles A	Alleles G	Count
1000GENOMES:pilot_1_CEU_low_coverage_panel	ss233534774	1000GENOMES	0.458	0.542	
1000GENOMES:pilot_1_CHB+JPT_low_coverage_panel	ss240577229	1000GENOMES	0.400	0.600	
1000GENOMES:pilot_1_YRI_low_coverage_panel	ss222470667	1000GENOMES	0.729	0.271	

CSHL-HAPMAP:HAPMAP-LWK	ss5253350	TSC-CSHL	0.667	0.333	0.400	0.533	0.067	6
CSHL-HAPMAP:HAPMAP-MEX	ss5253350	TSC-CSHL	0.490	0.510	0.245	0.490	0.265	13
CSHL-HAPMAP:HAPMAP-MKK	ss5253350	TSC-CSHL	0.633	0.367	0.410	0.446	0.144	20
CSHL-HAPMAP:HAPMAP-TSI	ss5253350	TSC-CSHL	0.488	0.512	0.226	0.524	0.250	21
CSHL-HAPMAP:HapMap-YRI	ss5253350	TSC-CSHL	0.708	0.292	0.487	0.442	0.071	8
SEATTLESEQ:Eight-Hapmap-Samples	ss159712995	SEATTLESEQ	unknown	unknown				

Other data (26)

- SIFT
- PolyPhen

1000 Genomes

A Deep Catalog of Human Genetic Variation

Human (GRCh37) | Location: 1:114,356,433-114,414,381 | Gene: PTPN22 | Transcript: PTPN22-001 | Tools | Help

Transcript-based displays

- Transcript summary
- Supporting evidence (22)
- Sequence
  - Exons (21)
  - cDNA
  - Protein
- External References
  - General identifiers (43)
  - Oligo probes (45)
- Ontology
  - Ontology chart (19)
  - Ontology table (19)
- Genetic Variation
  - Population comparison
  - Comparison image
- Protein Information
  - Protein summary
  - Domains & features (15)
  - Variations (46)
- External Data
- ID History
  - Transcript history
  - Protein history

Configure this page
Manage your data
Export data
Get VCF data
Bookmark this page
Download view as CSV

Transcript: PTPN22-001 (ENST00000359785)

Description protein tyrosine phosphatase, non-receptor type 22 (lymphoid) [Source:HGNC Symbol;Acc:9652]
Location Chromosome 1: 114,356,433-114,414,381 reverse strand.
Gene This transcript is a product of gene ENSG00000134242 - There are 12 transcripts in this gene

Show All entries | Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
PTPN22-001	ENST00000359785	3654	ENSP00000352833	807	Protein coding	CCDS863
PTPN22-002	ENST00000460620	1794	ENSP00000433141	179	Protein coding	-
PTPN22-004	ENST00000528414	3424	ENSP00000435176	752	Protein coding	-
PTPN22-006	ENST00000420377	2726	ENSP00000388229	795	Protein coding	-
PTPN22-007	ENST00000525799	2118	ENSP00000432674	668	Protein coding	-
PTPN22-201	ENST00000354605	2347	ENSP00000346621	691	Protein coding	CCDS864
PTPN22-202	ENST00000538253	2414	ENSP00000439372	563	Protein coding	-
PTPN22-008	ENST00000532224	2421	ENSP00000431249	135	Nonsense mediated decay	-
PTPN22-010	ENST00000529045	527	ENSP00000434932	92	Nonsense mediated decay	-
PTPN22-009	ENST00000534519	565	No protein product	-	Processed transcript	-
PTPN22-003	ENST00000484147	2258	No protein product	-	Retained intron	-
PTPN22-005	ENST00000469077	562	No protein product	-	Retained intron	-

Transcript and Gene level displays

Views in 1000 Genomes are separated into gene based views and transcript based views according to which level the information is more appropriately associated with. This view is a transcript level view. To flip between the two sets of views you can click on the Gene and Transcript tabs in the menu bar at the top of the page.

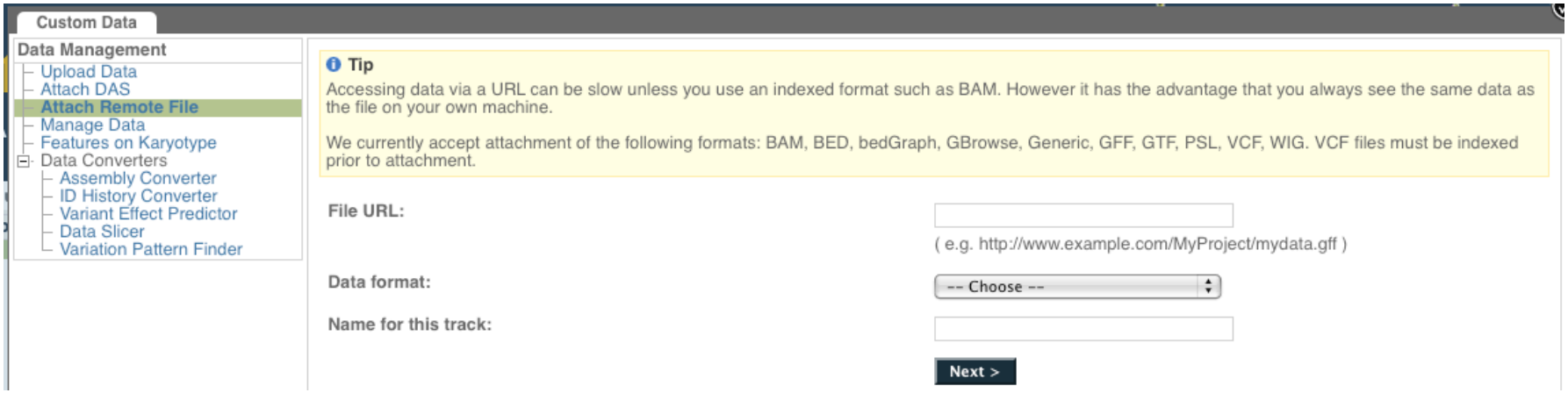
Variations help

Download view as CSV

Show All entries | Show/hide columns | Filter

Residue	Variation ID	Variation type	Alleles	Ambiguity code	Residues	Codons	SIFT	PolyPhen
16	rs74163639	Synonymous coding	G/A	R	S	AGC, AGT	-	-
49	rs61745743	Synonymous coding	A/G	R	A	GCT, GCC	-	-
71	rs74163642	Non-synonymous coding	A/G	R	V, A	GTA, GCA	deleterious	probably damaging
141	rs115552198	Non-synonymous coding	G/A	R	R, C	CGC, TGC	deleterious	probably damaging
177	1KG_1_114399013	Synonymous coding	C/T	Y	K	AAG, AAA	-	-
183	rs34590413	Stop gained	G/A	R	R, *	CGA, TGA	-	-
201	rs74163647	Non-synonymous coding	G/A	R	S, F	TCT, TTT	deleterious	probably damaging
206	rs61738614	Non-synonymous coding	A/C	M	L, R	CTT, CGT	deleterious	probably damaging
232	rs78195073	Synonymous coding	T/C	Y	G	GGA, GGG	-	-
247	rs35910094	Synonymous coding	T/G	K	L	CTA, CTC	-	-
263	rs33996649	Non-synonymous coding	C/T	Y	R, Q	CGG, CAG	tolerated	benign
266	rs72650670	Non-synonymous coding	G/A	R	R, W	CGG, TGG	deleterious	probably damaging
277	rs72483511	Stop gained, Splice site	C/A	M	E, *	GAA, TAA	-	-
324	rs113984534	Synonymous coding	A/G	R	Y	TAT, TAC	-	-
366	rs74163654	Synonymous coding	C/T	Y	E	GAG, GAA	-	-
370	rs72650671	Non-synonymous coding	G/T	K	H, N	CAC, AAC	deleterious	possibly damaging
388	rs77913785	Non-synonymous coding	G/T	K	D, E	GAC, GAA	deleterious	benign
413	1KG_1_114380784	Non-synonymous coding	T/G	K	Q, P	CAA, CCA	deleterious	benign
414	1KG_1_114380780	Synonymous coding	A/G	R	S	AGT, AGC	-	-
427	rs112873647	Non-synonymous coding	-/ATT	-	-, N	-, AAT	-	-
444	rs74163655	Non-synonymous coding	T/A	W	I, L	ATA, TTA	tolerated	benign
447	rs112191110	Non-synonymous coding	G/A	R	T, I	ACC, ATC	deleterious	probably damaging
452	rs56174946	Synonymous coding	A/G	R	F	TTT, TTC	-	-
456	rs72650672	Non-synonymous coding	G/C	S	Q, E	CAG, GAG	deleterious	possibly damaging
477	rs74163656	Synonymous coding	A/G	R	H	CAT, CAC	-	-
778	rs41313296	Non-synonymous coding	T/A	W	N, I	AAT, ATT	deleterious	probably damaging

File upload to view with 1000 Genomes data



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

* VCF must be indexed

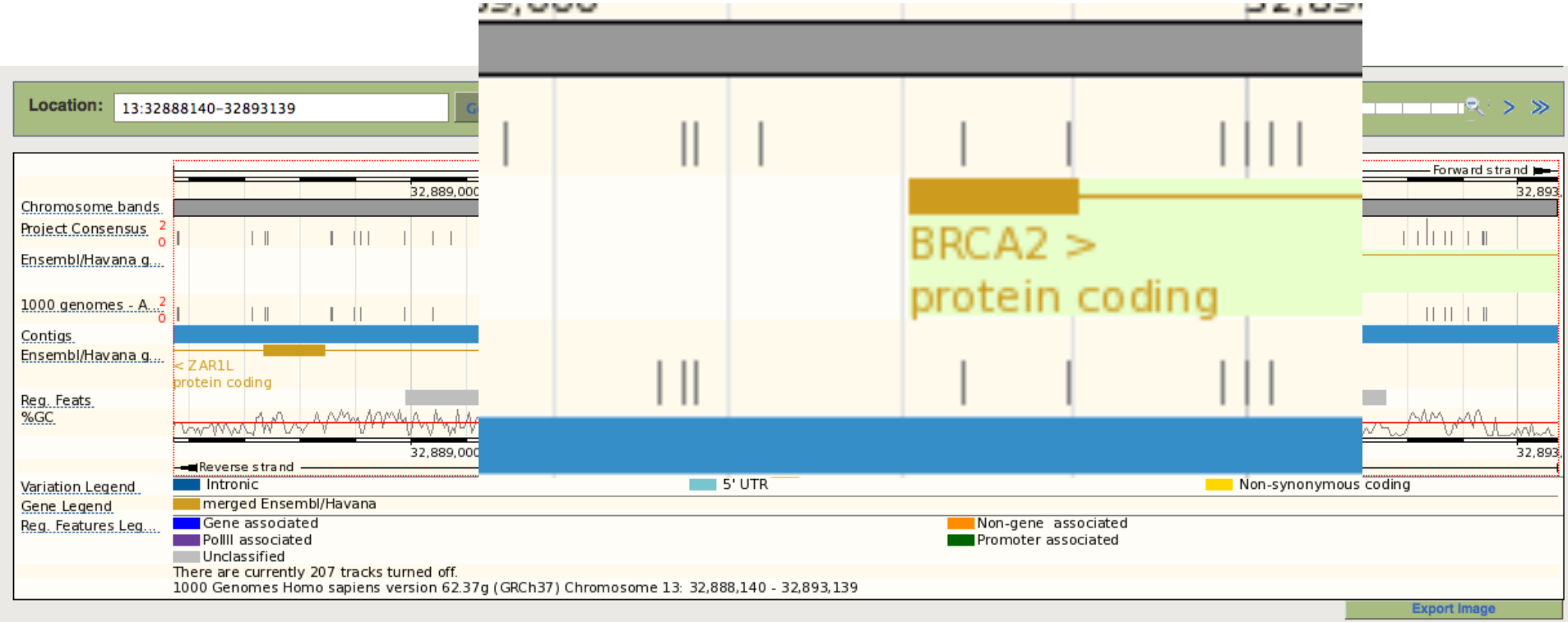
A THOUSAND GENOMES

EMBL-EBI

Uploaded VCF

Example:

Comparison of August calls and
/technical/working/20110502_vqsr_phase1_wgs_snps/
ALL.wgs.phase1.projectConsensus.snps.sites.vcf.gz



A THOUSAND GENOMES

EMBL-EBI

The Browser: Coming Soon

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Login · Register

Human (GRCh37) ▼ | Location: 9:22,125,003-22,126,003 | Variation: rs1333049

Variation displays

- Explore this variation
- Genomic context
  - Gene/Transcript (2)
- Population genetics (28)
- Individual genotypes (1737)
- Linkage disequilibrium
- Phenotype Data (8)
- Phylogenetic Context (4)
- Flanking sequence
- External Data

Configure this page

Manage your data

Export data

Bookmark this page

rs1333049 SNP

Source	dbSNP 134 - Variants (including SNPs and indels) imported from dbSNP
Alleles	Reference/Alternative: **G/C** | Ancestral: **C** | Ambiguity code: **S** | MAF: **0.40** (C)
Location	Chromosome **9:22125503** (forward strand) | View in location tab
Validation status	This variation is validated by **1000 Genomes**, **HapMap** and also cluster, doublehit, frequency, precious, submitter
Synonyms ⊞	This feature has **7** synonyms - click the plus to show
HGVS name	9:g.22125503G>C

Explore this variation help

Genomic context | Gene / Transcript | Population genetics | Individual genotypes

Linkage disequilibrium | Phenotype data | Phylogenetic context | Flanking sequence

Help with variations

YouTube videos

- SNPs and other Variations - 1 of 2
- SNPs and other Variations - 2 of 2
- Clip: Genome Variation
- BioMart: Variation IDs to HGNC Symbols

Reference materials

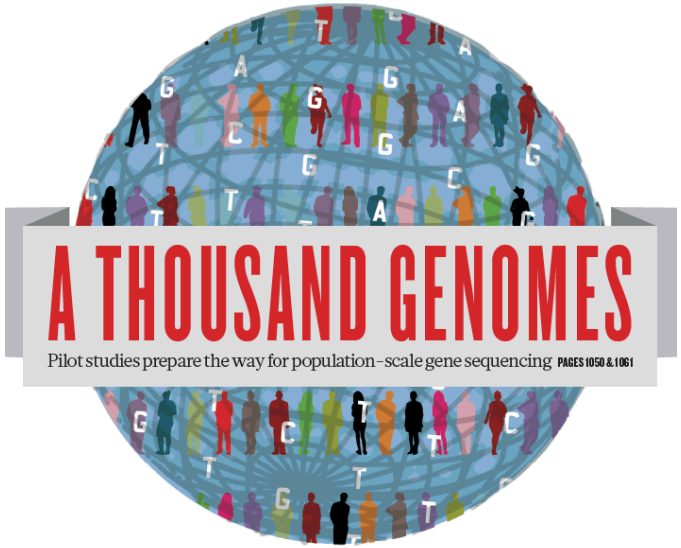
- Ensembl variation data: background and terminology
- Variation Quick Reference card

Additional resources

- Accessing variation data with the Variation API
- Genomes and SNPs in Malaria

A THOUSAND GENOMES

EMBL-EBI



The 1000 Genomes Project:

Exercise 2: Finding Variation Using the Browser

EMBL-EBI

Exercise: Finding Variation Using the Browser

- Find the variant rs45562238
  - http://browser.1000genomes.org
- Which 1000 Genomes Super Populations was it discovered in?
- What are it's allele frequencies?

1000 Genomes Browser

browser.1000genomes.org/index.html

Maps | dates | fix lj | docs | plus | gm | g | fb | ds | lj | NCBI | p | E | SRA | C | 1KG | JIRA | Wish | am | Later

1000 Genomes

A Deep Catalog of Human Genetic Variation

rs45562238

Tools | Help

Search Box

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**

s 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

1000 Genomes release 9 - September 2011 © EBI

About 1000 Genomes | Contact Us | Help

A THOUSAND GENOMES

EMBL-EBI

A Deep Catalog of Human Genetic Variation

Human (GRCh37)

Search 1000 Genomes

New Search

Configure this page

Manage your data

Export data

Get VCF data

Bookmark this page

Results Su

You searched for 'rs45562238'

Gene or Gene Product

0 entrie(s) matched your search strings.

Genetic Marker

0 entrie(s) matched your search strings.

Array Probe Set

0 entrie(s) matched your search strings.

SNP

1 entrie(s) matched your search strings.

1. dbSNP SNP: rs45562238

SNP Result

Interpro Domain

0 entrie(s) matched your search strings.

Gene Family

0 entrie(s) matched your search strings.

Sequence Aligned to Genome, eg. EST or Protein

0 entrie(s) matched your search strings.

Genomic Region, eg. Clone or Contig

0 entrie(s) matched your search strings.

A THOUSAND GENOMES

EMBL-EBI

Human (GRCh37) ▼ | Location: 6:133,013,044-133,014,044 | Variation: rs45562238 | Tools | Help

Variation displays

- Flanking sequence
- Gene/Transcript (1)
- Population genetics (11)
- Individual genotypes (2770)
- Genomic context
- Phenotype Data
- Phylogenetic Context
- External Data

Configure this page

Manage your data

Export data

Get VCF data

Bookmark this page

Variation: rs45562238

Population Genetics

Variation class	SNP (rs45562238 source dbSNP 132 - Variants (including SNPs and indels) imported from dbSN
Synonyms	**OMNI** SNP6-133055237 **Uniprot** VAR_023973
Present in	1000 genomes - Low coverage (1000 genomes - Low coverage - CEU),ALL - interim phase 1 - 1000 Genomes (AMR - interim phase 1 - 1000 Genomes, EUR - interim phase 1 - 1000 Genomes),ENSEMBL:Watson
Alleles	**T/C** (Ambiguity code: **Y**)
Ancestral allele	T
Location	This feature maps to 6:133013544 (forward strand) I View in location tab
Validation status	Proven by **cluster, frequency, 1000Genome**
HGVS names ⊞	This feature has 3 HGVS names - click the plus to show

Flanking sequence help

Flanking Sequence (reference and dbSNP)

```
AAAAAAAAAAAAAAAGATTTAGCAAATAAATTTAGCCTGTATTTGTAGGCTTCCCAGCTA
AAGTTGAATTTCAAATAAACAACAAAATTTTTAGCATACATAAATCCCCAAAAGTGTGAT
ATGCTTATACTAAAAAATCATTCATTATTTATCTAAACTGTATATTTAACTGAGTGTCTT
TATTTTATCCAGTAACTCTTCTGACATCAAAATATTACCTGTAGATAATAGCGCCCTTCC
ACAGTGTGCAGTCCGTCAAATGCCCCTAGAGCGTACACTTCATTTGGTATGTTCTCAGAC
ATTTTGTAGCTTAAATGACAGCAGAGATCTTTCTGACAAACTGTATAATTTCCTGCAACT
CCTGTGAGCTTCACAAAAGTGAATTCATCGAAAAAGACAGYGCCTTTAAATTCCTTGTTT
CCTGATGAGAGCGCTTCTATACTGCTGGCATAGGAAGTCCAGTTCACCACTGCAGAATGG
GATGGGTGGGAATCCAGTTGCGAGAGGAGGAGTTTTCCCTCTTCTGTCTTCATATCATAA
TGAAATGCTCTTGAAGAATTGGGTGCATAGATGCCACTTCCTGGAAGTAATAAGTTGAAA
ACATCATTTGAAATTGGAAAGTAAAAAGGGATTTCATAATTGTTATTACTATTTTAACCT
CATATTCACATTTTACTCTCTCTAAGTTATATGTTTTATTAATTTCAGTAAAACATGCTG
ATAAAGACAGAGAAGTGAGAGGAAATCAATTAAAGTATGAAACCAAGAAGCCAAGAAGGG
AGAGTTGTATGTAAGAAGTCA
```

(Variant highlighted)

A THOUSAND GENOMES

EMBL-EBI

Gene/Transcript

Human (GRCh37) | Location: 6:133,013,044-133,014,044 | Variation: rs45562238 | Tools | Help

Variation displays
- Flanking sequence
- Gene/Transcript (1)
- **Population genetics (11)**
- Individual genotypes (2770)
- Genomic context
- Phenotype Data
- Phylogenetic Context
- External Data

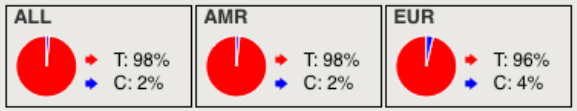
Configure this page
Manage your data
Export data
Get VCF data
Bookmark this page
Download view as CSV

Variation: rs45562238

Variation class	SNP (rs45562238 source dbSNP 132 - Variants (including SNPs and indels) imported from dbSNP [http://www.ncbi.nlm.nih.gov/projects/SNP/])
Synonyms	**OMNI** SNP6-133055237 **Uniprot** VAR_023973
Present in	1000 genomes - Low coverage (1000 genomes - Low coverage - CEU),ALL - interim phase 1 - 1000 Genomes (AMR - interim phase 1 - 1000 Genomes, EUR - interim phase 1 - 1000 Genomes),ENSEMBL:Watson
Alleles	**T/C** (Ambiguity code: **Y**)
Ancestral allele	T
Location	This feature maps to 6:133013544 (forward strand) | View in location tab
Validation status	Proven by **cluster, frequency, 1000Genome**
HGVS names ⊞	This feature has 3 HGVS names - click the plus to show

Population genetics help

1000 genomes alleles frequencies



Pie Charts

1000 genomes

Show/hide columns | Filter

Population	Alleles C	Alleles T	Genotypes CIT	Genotypes TIT	Allele count	Genotype count	Genotype detail
1000GENOMES:ALL	0.016	0.984	0.031	0.969	34 **(C)** / 2154 **(T)**	34 **(CIT)** / 1060 **(TIT)**	Show
1000GENOMES:AMR	0.017	0.983	0.033	0.967		6 **(CIT)** / 175 **(TIT)**	Show
1000GENOMES:EUR	0.037	0.963	0.073	0.927		28 **(CIT)** / 353 **(TIT)**	Show

A THOUSAND GENOMES

EMBL-EBI

1000 Genomes

A Deep Catalog of Human Genetic Variation

Human (GRCh37) | Location: 6:133,013,044-133,014,044 | Variation: rs45562238

Tools | Help

Variation displays

- Flanking sequence
- Gene/Transcript (1)
- Population genetics (11)
- Individual genotypes (2770)
- Genomic context
- Phenotype Data
- Phylogenetic Context
- External Data

Configure this page

Manage your data

Export data

Get VCF data

Bookmark this page

Download view as CSV

Variation: rs45562238

Variation class	SNP (rs45562238 source dbSNP 132 - Variants (including SNPs and indels) imported from dbSNP [http://www.ncbi.nlm.nih.gov/projects/SNP/])
Synonyms	**OMNI** SNP6-133055237 **Uniprot** VAR_023973
Present in	1000 genomes - Low coverage (1000 genomes - Low coverage - CEU),ALL - interim phase 1 - 1000 Genomes (AMR - interim phase 1 - 1000 Genomes - EUR - interim phase 1 - 1000 Genomes),ENSEMBL:Watson
Alleles	**T/C** (Ambiguity code: **Y**)
Ancestral allele	T
Location	This feature maps to 6:133013544 (forward strand) | View in location tab
Validation status	Proven by **cluster, frequency, 1000Genome**
HGVS names ⊞	This feature has 3 HGVS names - click the plus to show

Sift/PolyPhen

Gene/Transcript help

Show/hide columns

Filter

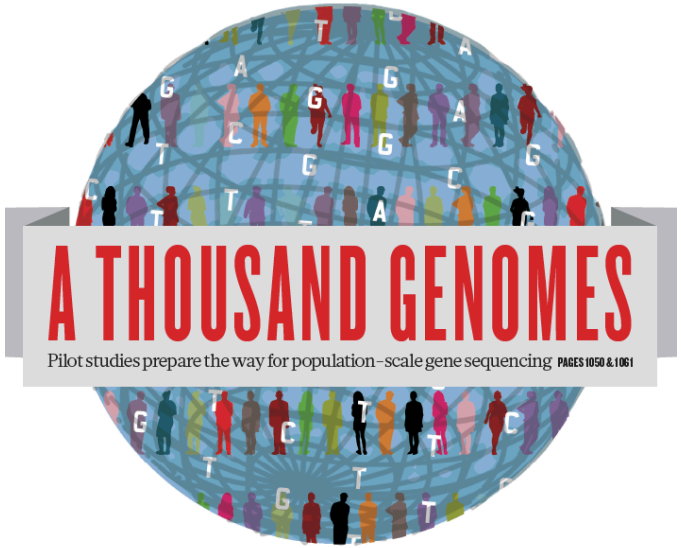
Gene	Transcript (strand)	Allele (transcript allele)	Type	HGVS names	Position in transcript	Position in CDS	Position in protein	Amino acid	Codons	SIFT	PolyPhen
ENSG00000112299	ENST00000367928 (-)	C (G)	Non-synonymous coding	ENST00000367928.4:c.1006A>G ENSP00000356905.4:p.Thr336Ala	1020	1006	336	T/A	ACT/GCT	tolerated	benign

1000 Genomes release 9 - September 2011 © EBI

About 1000 Genomes | Contact Us | Help

A THOUSAND GENOMES

EMBL-EBI



The 1000 Genomes Project:

The 1000 Genomes Tools

EMBL-EBI

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

1000 Genomes release 10 - October 2011 © EBI

About 1000 Genomes | Contact Us | Help

http://browser.1000genomes.org

EMBL-EBI

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		
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.	Online version	

1000 Genomes release 10 - October 2011 © EBI

About 1000 Genomes | Contact Us | Help

A THOUSAND GENOMES

EMBL-EBI

Data slicer for subsets of the data

1000 Genomes

A Deep

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

Tip

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.

VCF / BAM File URL:

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

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_phase1.2(

< Back Next >

A THOUSAND GENOMES

EMBL-EBI

Get VCF Button

Human (GRCh37) ▼ | Location: 6:31,830,969-31,846,823 | Gene: SLC44A4 | Tools | Help

Gene-based displays

- Gene summary
- Splice variants (9)
- Supporting evidence
- Sequence
- External references
- Regulation
- Genetic Variation
  - Variation Table
  - Structural Variation
  - Variation Image
- External Data
- ID History
  - Gene history

Configure this page

Manage your data

Export data

Get VCF data

Bookmark this page

Gene: SLC44A4 (ENSG00000204385)

Description solute carrier family 44, member 4 [Source:HGNC Symbol;Acc:13941]

Location Chromosome 6: 31,830,969-31,846,823 reverse strand.

Transcripts There are 9 transcripts in this gene

Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
SLC44A4-001	ENST00000229729	2589	ENSP00000229729	710	Protein coding	CCDS4724
SLC44A4-004	ENST00000414427	1233	ENSP00000398901	411	Protein coding	-
SLC44A4-201	ENST00000375562	2505	ENSP00000364712	668	Protein coding	-
SLC44A4-202	ENST00000544672	2634	ENSP00000444109	634	Protein coding	-
SLC44A4-002	ENST00000465707	681	No protein product	-	Processed transcript	-
SLC44A4-003	ENST00000462671	426	No protein product	-	Processed transcript	-
SLC44A4-007	ENST00000487680	392	No protein product	-	Processed transcript	-
SLC44A4-005	ENST00000475563	575	No protein product	-	Retained intron	-
SLC44A4-006	ENST00000479777	655	No protein product	-	Retained intron	-

Ensembl Variant Effector Predictor (VEP)

- Takes list of variation and annotates with respect to Ensembl features
- Returns whether the SNP has been seen in the 1000 Genomes and if it has an rs number (if one has been assigned)
- Returns SIFT, PolyPhen and Condel scores
- Extensive filtering options by MAF and populations
- Web and command line versions

A THOUSAND GENOMES

EMBL-EBI

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

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 >

Variation Pattern Finder

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

A THOUSAND GENOMES

EMBL-EBI

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

Variation Pattern Finder

Export data: CSV Excel

Go to collapsed view

Population ASW	CEU	Freq	*Variation info* rs9369628:C/T	rs61[illegible]1828:C/T	rs12192544:C/G	rs599
			6:46620135	6:46620240	6:46620252	6:466
			ENST00000275016 SPLICE_SITE	ENST00000275016 NON_SYNONYMOUS_CODING:R/H	ENST00000275016 NON_SYNONYMOUS_CODING:R/P	ENST0 NON_S
NA20314, NA20322	NA12348, N	0.095	CIC	CIC	GIC	GIG
NA20356, NA19625 and 1 other(s)	NA11919, N	0.092	CIC	CIC	CIG	GIG
NA20291, NA19985 and 5 other(s)		0.069	CIT	CIC	CIC	GIG
NA20289, NA20294 and 4 other(s)		0.057	TIC	CIC	CIC	GIG
	NA12546, N	0.026	CIC	CIC	GIG	GIG
NA19[illegible]		0.012	TIT	CIC	CIC	GIG
	NA12283	0.011	TIC	CIC	CIG	GIG
NA19908, NA20278		0.011	CIT	CIC	GIC	GIG
NA19703		0.008	CIC	CIC	CIC	GIG
NA20351		0.007	CIC	CIC	CIC	GIG
		0.006	CIC	CIC	CIG	GIG
NA19712		0.004	CIC	CIC	CIC	CIG
		0.003	CIC	CIC	GIG	GIG
		0.003	TIC	CIC	CIC	GIG
		0.002	CIC	CIC	CIC	GIG

A THOUSAND GENOMES

EMBL-EBI

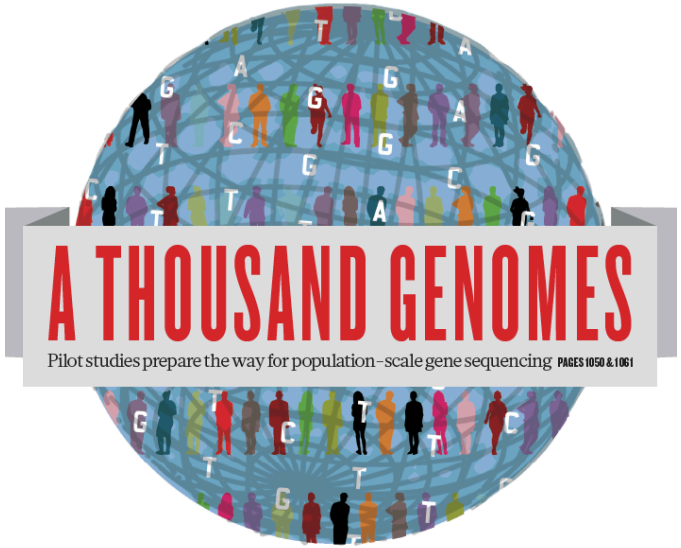
Access to backend Ensembl databases

- 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

A THOUSAND GENOMES

EMBL-EBI



The 1000 Genomes Project:

Exercise 3: Using 1000 Genomes Tools

Exercise: Using 1000 Genomes Tools

- Find the gene SLC44A4 using the search box on http://browser.1000genomes.org
- Get a VCF file for this Gene using the Get VCF button.
- Uncompress this file
  - You can get a copy at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120120_1000genomes_tutorial/6.31830969-31846823.ALL.chr6.phase1.projectConsensus.genotypes.vcf
- Use this file with the Variant Effect Predictor
  - http://browser.1000genomes.org/Homo_sapiens/UserData/UploadVariations
- Do any of the variants have deleterious effects according to SIFT or PolyPhen
- Use the example url on the page and the coordinates 6:31830700-31840700 with the Variation Pattern Finder
  - http://browser.1000genomes.org/Homo_sapiens/UserData/VariationsMapVCF

1000 Genomes

A Deep Catalog of Human Genetic Variation

SLC44A4

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.

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.

A THOUSAND GENOMES

EMBL-EBI

1000 Genomes

A Deep Catalog of Human Genetic Variation

Human (GRCh37)

Tools | Help

Search 1000 Genomes

New Search

Configure this page

Manage your data

Export data

Get VCF data

Bookmark this page

Results Summary

You searched for 'SLC44A4'

Gene or Gene Product

10 entrie(s) matched your search strings.

1. **Gene: ENSG00000204385 [Region in detail]**
   SLC44A4 - solute carrier family 44, member 4 [Source:HGNC Symbol;Acc:13941]
2. **Transcript: ENST00000229729 [Region in detail]**
3. **Peptide: ENSP00000398764 [Region in detail]**
   SLC44A4
4. **Peptide: ENSP00000392054 [Region in detail]**
   SLC44A4
5. **Peptide: ENSP00000404572 [Region in detail]**
   SLC44A4
6. **Peptide: ENSP00000398901 [Region in detail]**
   SLC44A4
7. **Peptide: ENSP00000415708 [Region in detail]**
   SLC44A4
8. **Peptide: ENSP00000400263 [Region in detail]**
   SLC44A4
9. **Peptide: ENSP00000414296 [Region in detail]**
   SLC44A4
10. **Peptide: ENSP00000399161 [Region in detail]**
    SLC44A4

A THOUSAND GENOMES

EMBL-EBI

Human (GRCh37) | Location: 6:31,830,969-31,846,823 | Gene: SLC44A4 | Tools | Help

Gene-based displays

- Gene summary
- Splice variants (9)
- Supporting evidence
- Sequence
- External references
- Regulation
- Genetic Variation
  - Variation Table
  - Structural Variation
  - Variation Image
- External Data
- ID History
  - Gene history

Configure this page

Manage your data

Export data

Get VCF data

Bookmark this page

Gene: SLC44A4 (ENSG00000204385)

Description solute carrier family 44, member 4 [Source:HGNC Symbol;Acc:13941]

Location Chromosome 6: 31,830,969-31,846,823 reverse strand.

Transcripts There are 9 transcripts in this gene

Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
SLC44A4-001	ENST00000229729	2589	ENSP00000229729	710	Protein coding	CCDS4724
SLC44A4-004	ENST00000414427	1233	ENSP00000398901	411	Protein coding	-
SLC44A4-201	ENST00000375562	2505	ENSP00000364712	668	Protein coding	-
SLC44A4-202	ENST00000544672	2634	ENSP00000444109	634	Protein coding	-
SLC44A4-002	ENST00000465707	681	No protein product	-	Processed transcript	-
SLC44A4-003	ENST00000462671	426	No protein product	-	Processed transcript	-
SLC44A4-007	ENST00000487680	392	No protein product	-	Processed transcript	-
SLC44A4-005	ENST00000475563	575	No protein product	-	Retained intron	-
SLC44A4-006	ENST00000479777	655	No protein product	-	Retained intron	-

A THOUSAND GENOMES

EMBL-EBI

Configure Page | 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

Tip

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.

VCF / BAM File URL:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1

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

Region:

6:31830969-31846823

(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

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

< Back | Next >

A THOUSAND GENOMES

EMBL-EBI

Configure Page | 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

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

```
##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  HG00096 
6       31831159         rs3869144       C       T       100     PASS    .       
6       31831167         .       T       C       100     PASS    .       GT:AP   
```

A THOUSAND GENOMES

EMBL-EBI

Custom Data

- Assembly Converter
- ID History Converter
- Variant Effect Predictor
- Data Slicer
- Variation Pattern Finder

Input file

Species: Human (Homo sapiens): GRCh37

Name for this upload (optional): SLC44A4

Paste file:

Upload file: /Users/laura/Downloads/6. Browse...

or provide file URL:

Input file format: VCF

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: Prediction only

PolyPhen predictions: Prediction only

Condel consensus (SIFT/PolyPhen) predictions: No

A THOUSAND GENOMES

EMBL-EBI

6_31833249_A/G	6:31833249	G	ENSG00000204385	ENST00000487680	Transcript	UPSTREAM	-	-	-	-	-	-	-	-
6_31833249_A/G	6:31833249	G	ENSG00000204385	ENST00000414427	Transcript	DOWNSTREAM	-	-	-	-	-	-	-	-
6_31833249_A/G	6:31833249	G	ENSG00000204385	ENST00000479777	Transcript	DOWNSTREAM	-	-	-	-	-	-	-	-
6_31833249_A/G	6:31833249	G	ENSG00000204385	ENST00000475563	Transcript	DOWNSTREAM	-	-	-	-	-	-	-	-
6_31833357_C/T	6:31833357	T	-	ENSR00000487922	RegulatoryFeature	REGULATORY_REGION	-	-	-	-	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204386	ENST00000495807	Transcript	UPSTREAM	-	-	-	-	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204386	ENST00000480384	Transcript	UPSTREAM	-	-	-	-	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204386	ENST00000491768	Transcript	UPSTREAM	-	-	-	-	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204386	ENST00000375631	Transcript	UPSTREAM	-	-	-	-	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204386	ENST00000479533	Transcript	UPSTREAM	-	-	-	-	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000229729	Transcript	NON_SYNONYMOUS_CODING	1625	1604	535	R/H	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000375562	Transcript	NON_SYNONYMOUS_CODING	1544	1478	493	R/H	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000544672	Transcript	NON_SYNONYMOUS_CODING	1673	1376	459	R/H	-	-	-	-
6_31833357_C/T	6:31833357	T	E	-	-	-	1KG_6_31833357	-						-
6_31833357_C/T	6:31833357	T	E											-
6_31833357_C/T	6:31833357	T	E	-	-	-	1KG_6_31833357	-						-
6_31833357_C/T	6:31833357	T	E	535	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=probably_damaging; Condel=deleterious						-
6_31833612_C/G	6:31833612	G												-
6_31833612_C/G	6:31833612	G	E											-
6_31833612_C/G	6:31833612	G	E											-
				493	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=possibly_damaging; Condel=deleterious						
				459	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=probably_damaging; Condel=deleterious						
				-	-	-	1KG_6_31833357	-						
				-	-	-	1KG_6_31833357	-						
				-	-	-	1KG_6_31833357	-						

A THOUSAND GENOMES

EMBL-EBI

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

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:

6:31830700-31840700

e.g. 6:46620015-46620998

Next >

A THOUSAND GENOMES

EMBL-EBI

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

Variation Pattern Finder

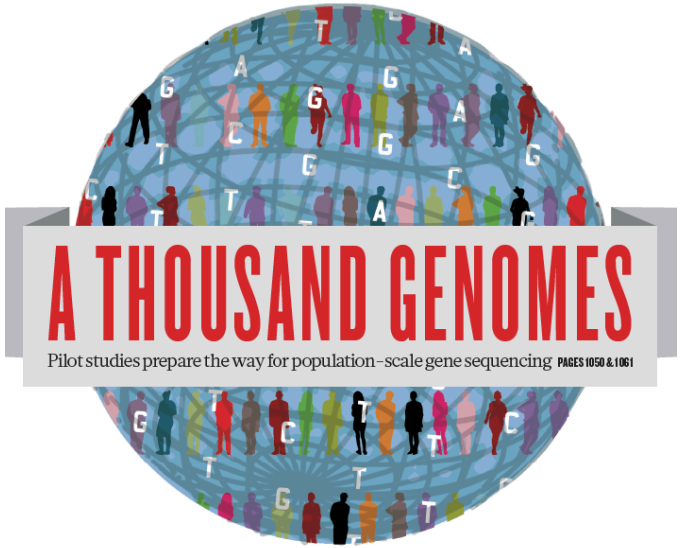
Export data: CSV Excel

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	MXL	PUR	TSI	YRI	Freq	rs3869144:C	6:31831167:	6:31831206:	6:31831301:	6:31831478:	6:31833357:	rs6915800:G	rs116706632	rs11712
2						6:31831159	6:31831167	6:31831206	6:31831301	6:31831478	6:31833357	6:31833660	6:31836976	6:31837
3						ENST000002	ENST000002	ENST000002	ENST000002	ENST000002	ENST000002	ENST000002	ENST000002	ENST000
4						ENST000003	ENST000003	ENST000003	ENST000003	ENST000003	ENST000003	ENST000003	ENST000003	ENST000
5						ENST000005	ENST000005	ENST000005	ENST000005	ENST000005	ENST000005	ENST000005	ENST000005	ENST000
6												ENST000004	ENST000004	ENST000
7	NA19758; NA	HG01048; HG	NA20787; NA	NA19116; NA	0.293	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
8	NA19741; NA	HG00640; HG	NA20801; NA	NA19189; NA	0.203	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
9	NA19654; NA	HG01085; HG	NA20588; NA	NA19236; NA	0.195	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
10	NA19747; NA	HG01102	NA20522; NA	NA19119; NA	0.032	C|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
11	NA19746		NA20826; NA	NA19172; NA	0.026	T|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
12	NA19756			NA19160; NA	0.016	C|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
13			NA20752	NA19209	0.013	T|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
14			NA20510; NA		0.008	T|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	C|G
15		HG01105	NA20783; NA		0.005	C|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|C
16	NA19682; NA				0.005	C|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|C
17					0.004	T|C	T|T	G|G	C|C	T|T	C|C	A|G	G|G	G|G
18				NA18933	0.003	C|T	T|T	G|G	C|C	T|T	C|C	G|A	G|G	G|G
19				NA19098; NA	0.003	T|C	T|T	G|G	C|C	T|T	C|C	A|G	G|G	G|G
20	NA19740				0.003	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|A	G|G
21					0.003	T|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	C|G
22					0.002	T|T	T|T	G|G	C|C	T|T	C|C	A|A	G|G	G|G
23					0.002	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
24					0.002	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
25	NA19675				0.002	C|C	T|T	G|G	C|C	C|T	C|C	G|G	G|G	G|G
26				NA19093	0.002	C|C	C|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
27		HG00734		NA18486	0.002	C|C	T|T	G|A	C|C	T|T	C|C	G|G	G|G	G|G
28					0.001	C|C	T|T	G|A	C|C	T|T	C|C	G|G	G|G	G|G
29					0.001	T|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	C|C
30	NA19678				0.001	C|C	T|T	G|G	C|C	T|C	C|C	G|G	G|G	G|G
31					0.001	C|C	T|T	A|G	C|C	T|T	C|C	G|G	G|G	G|G
32			NA20818		0.001	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
33					0.001	T|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
34				NA19095	0.001	T|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
35					0.001	T|C	T|T	G|G	C|C	T|T	C|C	G|G	A|G	G|G
36				NA19152	0.001	T|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
37					0.001	C|C	T|T	G|G	C|G	T|T	C|C	G|G	G|G	G|G
38				NA19146	0.001	C|T	T|T	G|G	C|C	T|T	C|C	G|A	G|G	G|G
39				NA19099	0.001	C|C	T|C	G|G	C|C	T|T	C|C	G|G	G|G	G|G
40	NA19752				0.001	C|T	T|T	G|G	C|C	T|T	C|C	G|G	G|A	G|G
41					0.001	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
42					0.001	C|C	T|T	G|G	C|C	T|T	C|T	G|G	G|G	G|G
43		HG01083			0.001	C|C	T|T	G|G	C|C	T|T	C|C	G|G	A|G	G|G
44														
45														
46														

7493:G/C	rs644827:T/C
009	6:31838441
00229729 NONYMOUS_CODING:Q/E	ENST00000229729 NON_SYNONYMOU
00375562 NONYMOUS_CODING:Q/E	ENST00000375562 NON_SYNONYMOU
00544672 NONYMOUS_CODING:Q/E	ENST00000544672 NON_SYNONYMOU
00414427 NONYMOUS_CODING:Q/E	
	CIC
	CIT
	TIC
	CIC
	CIC
	TIC
	CIT
	CIC
	TIC
	CIC
	CIC
	CIC
	CIT
	CIC
	CIT

A THOUSAND GENOMES

EMBL-EBI



The 1000 Genomes Project:

Finding out about New Data and using Data on Campus

EMBL-EBI

Announcements

- 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
- info@1000genomes.org

Thanks

- The 1000 Genomes Project Consortium
- Paul Flicek, Laura Clarke
- Richard Smith and Holly Zheng Bradley
- Giulietta Spudich and Denise Carvalho-Silva

A THOUSAND GENOMES

EMBL-EBI