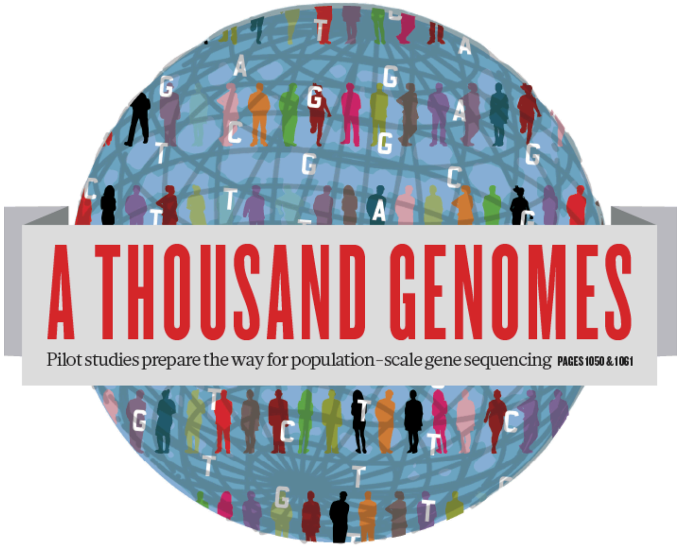

A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project: A Tutorial

Laura Clarke and Giulietta Spudich
EBI Training Room
16th February 2012

EMBL-EBI

Agenda

- Brief History of the 1000 Genomes Project, data and analysis
- The Raw Data and FTP site
- Exercise: Finding and viewing Data
- The Website and Browser
- Exercise: Using the Browser
- The 1000 Genomes Tools
- Exercise: Interacting with 1000 genomes on the command line

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

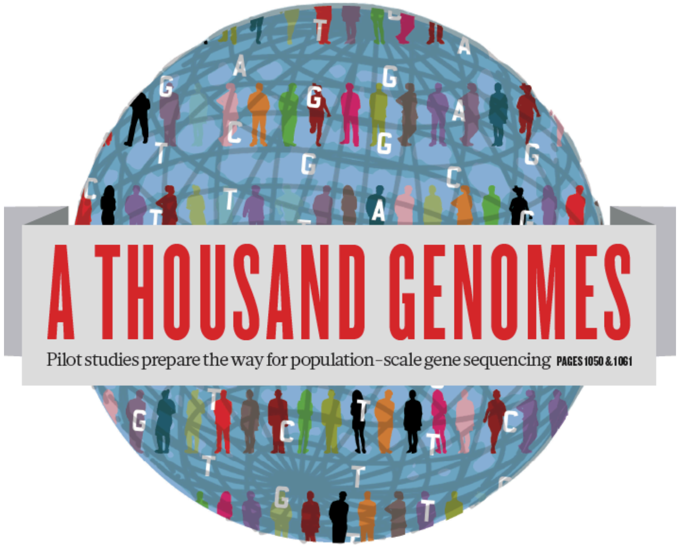
Command Line Tools

- Samtools http://samtools.sourceforge.net/
- Tabix http://sourceforge.net/projects/samtools/files/tabix/
  - (Please note it is best to use the trunk svn code for this as the 0.2.5 release has a bug)
  - svn co https://samtools.svn.sourceforge.net/svnroot/samtools/trunk/tabix
- Vcftools http://vcftools.sourceforge.net/
- The ensembl variation and core apis http://www.ensembl.org/index.html
- The variant effect predictor ftp://ftp.ensembl.org/pub/misc-scripts/Variant_effect_predictor/
- The variation pattern finder ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/variation_pattern_finder/version_1.0
- VCF to PED Converter ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/vcf_to_ped_converter/version_1.0/
- Haploview http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/downloads

Slides available online

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

How are you using 1000 genomes data?


A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

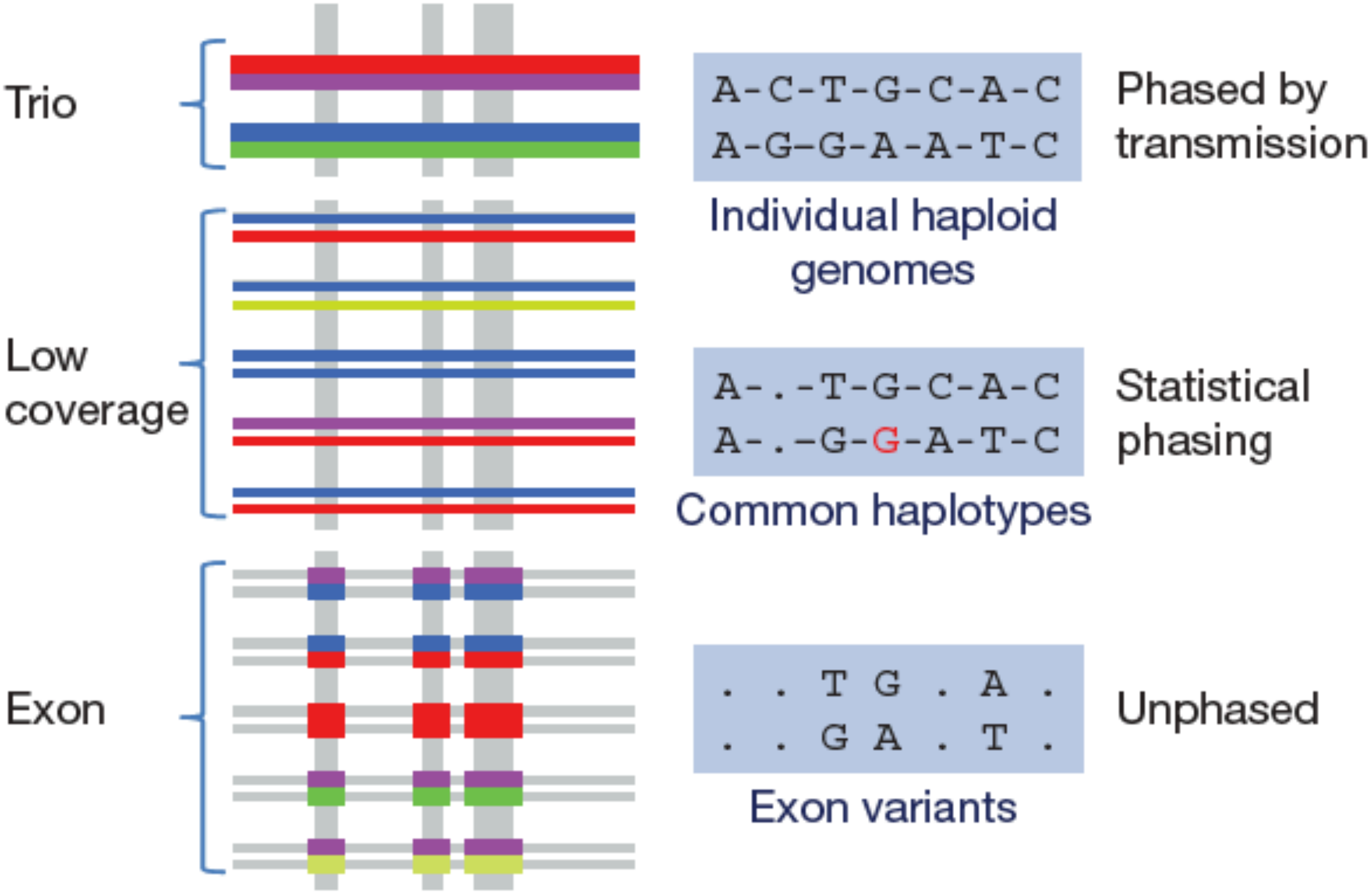
The 1000 Genomes Project: A Brief History of Data and Analysis

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



Trio

A-C-T-G-C-A-C
A-G-G-A-A-T-C

Phased by transmission

Individual haploid genomes

Low coverage

A-.-T-G-C-A-C
A-.-G-G-A-T-C

Statistical phasing

Common haplotypes

Exon

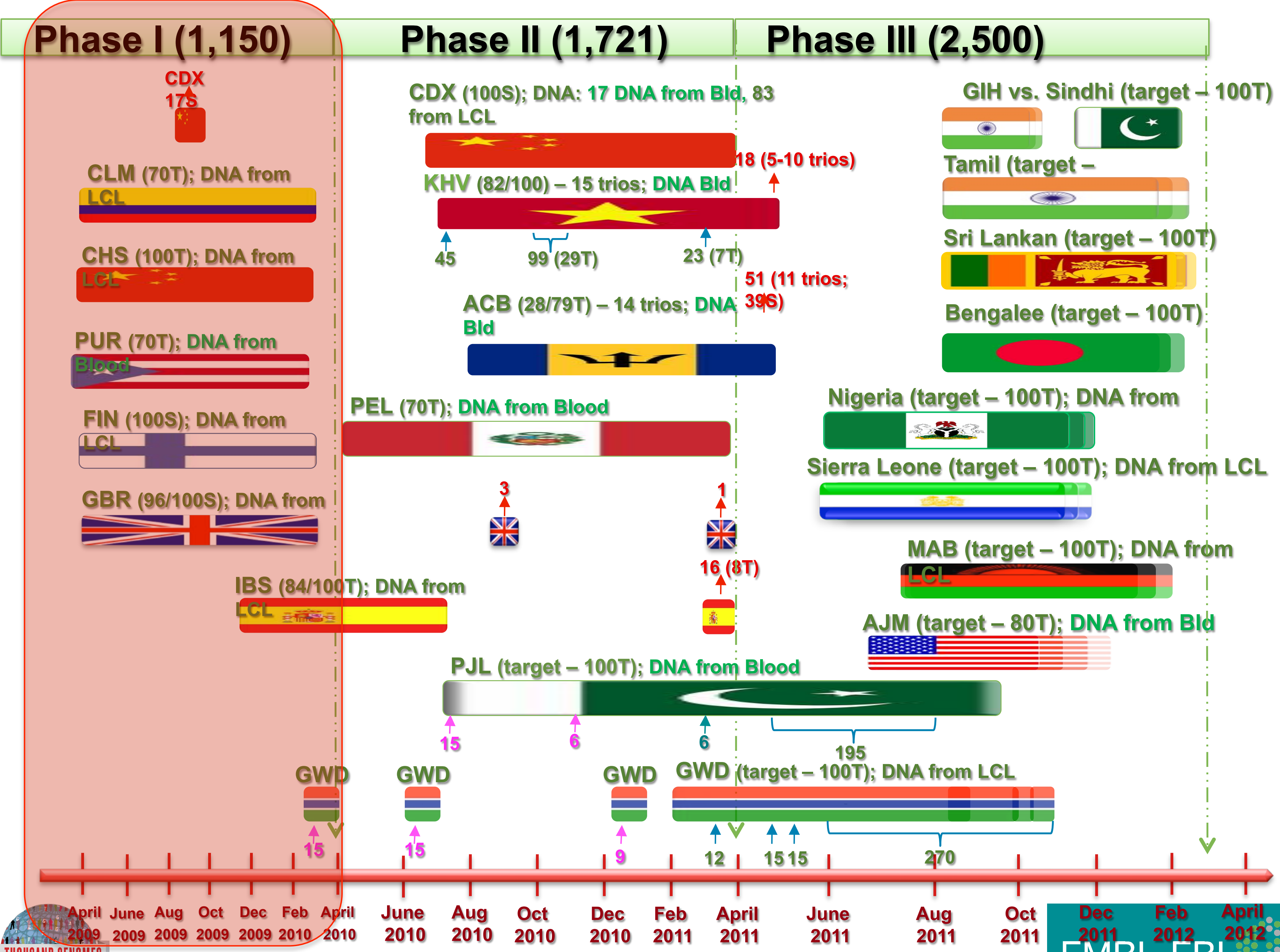
. . T G . A .
. . G A . T .

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
- Phase 1 Release Integrated Variant Release has been made.



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

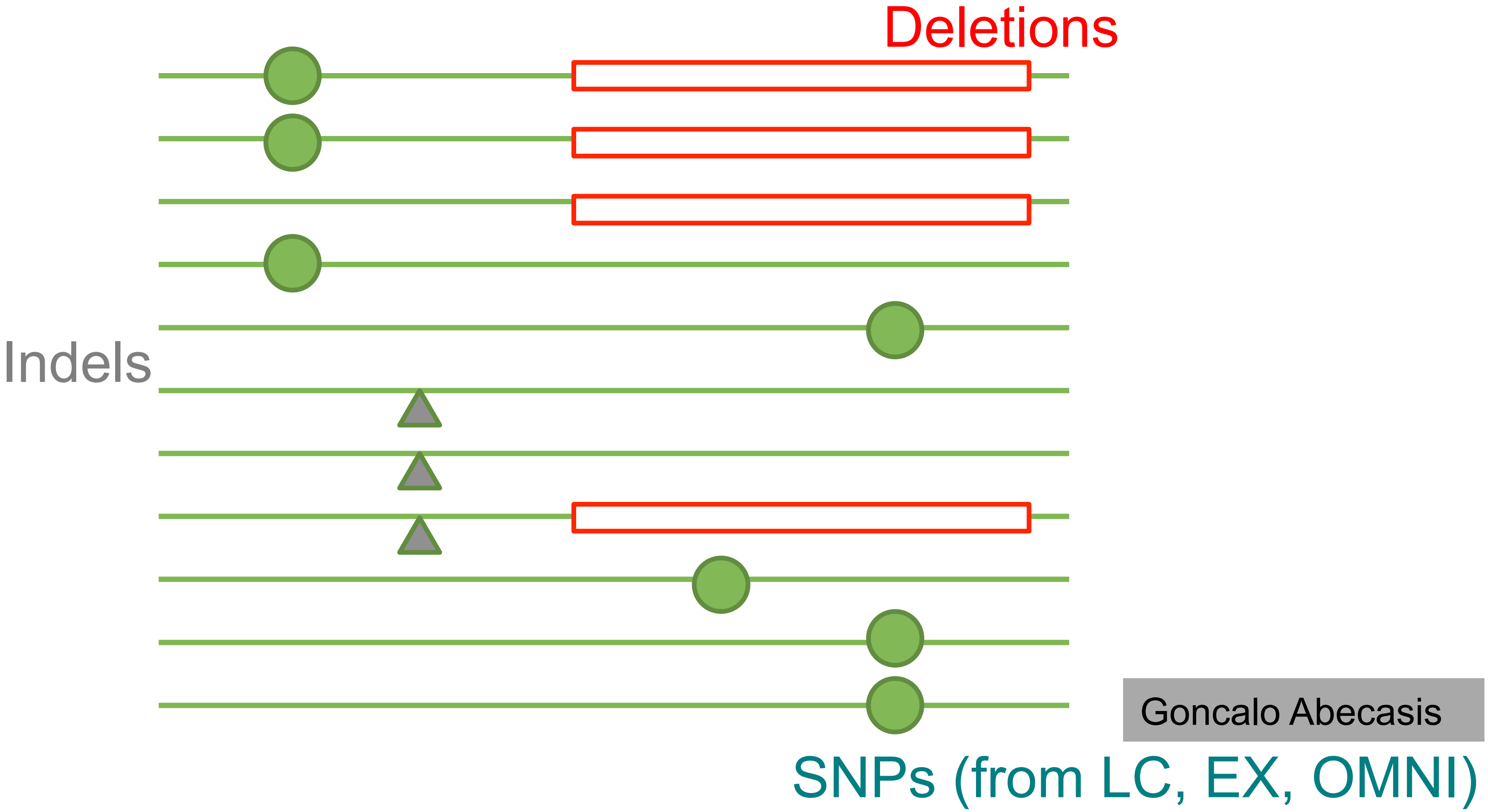
AJM (target – 80T); DNA from Bld

A THOUSAND GENOMES

EMBL-EBI

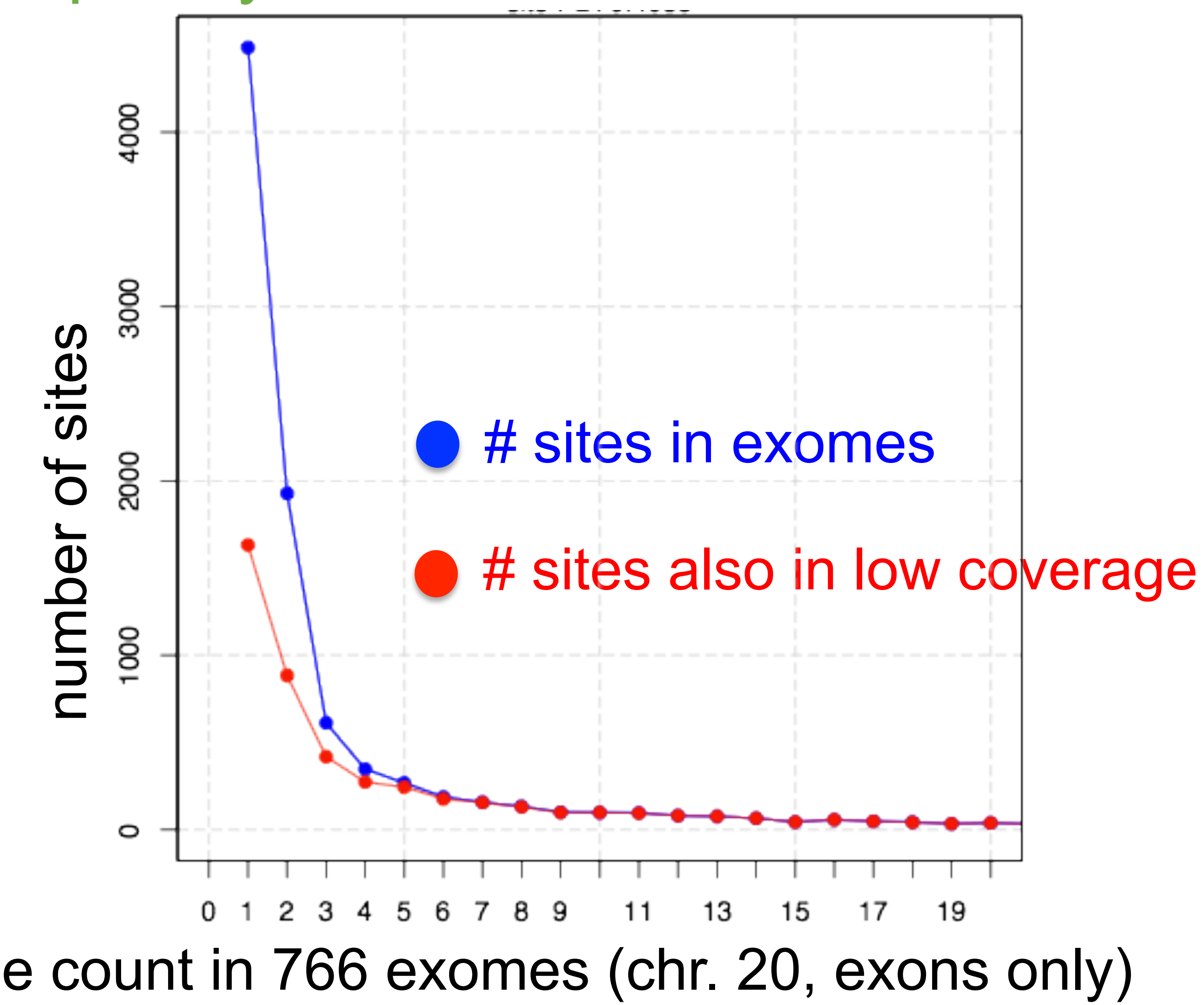
Phase 1 analysis goal: an **integrated view of human variations**

- Reconstruct haplotypes including all variant types, using all datasets



Goncalo Abecasis

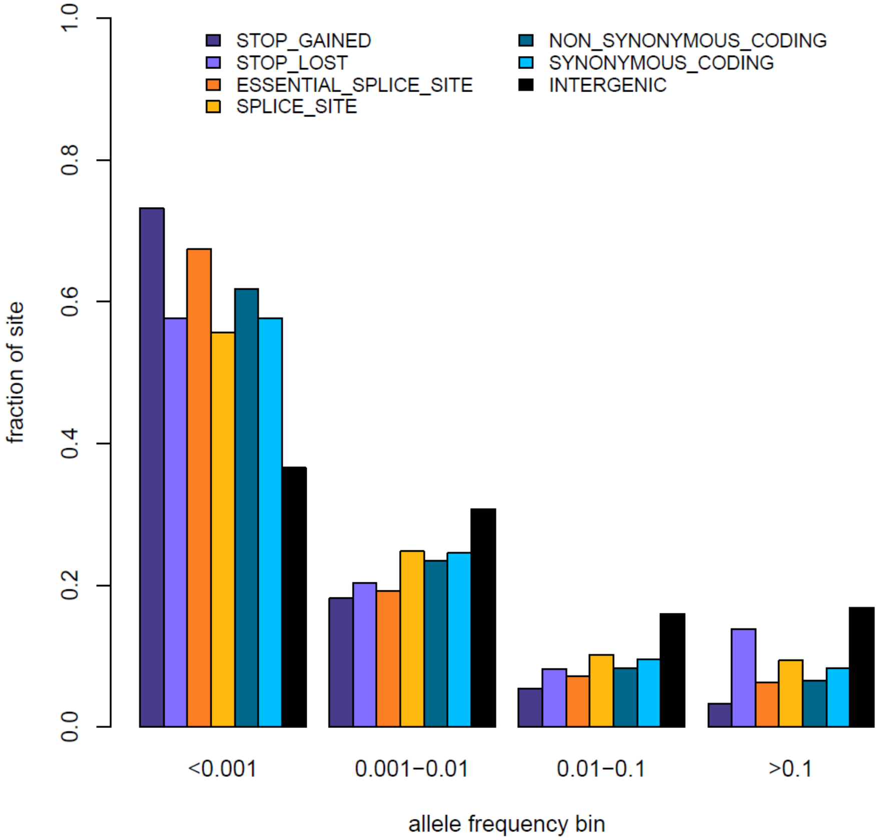
Deep coverage exome data is more sensitive to low-frequency variants



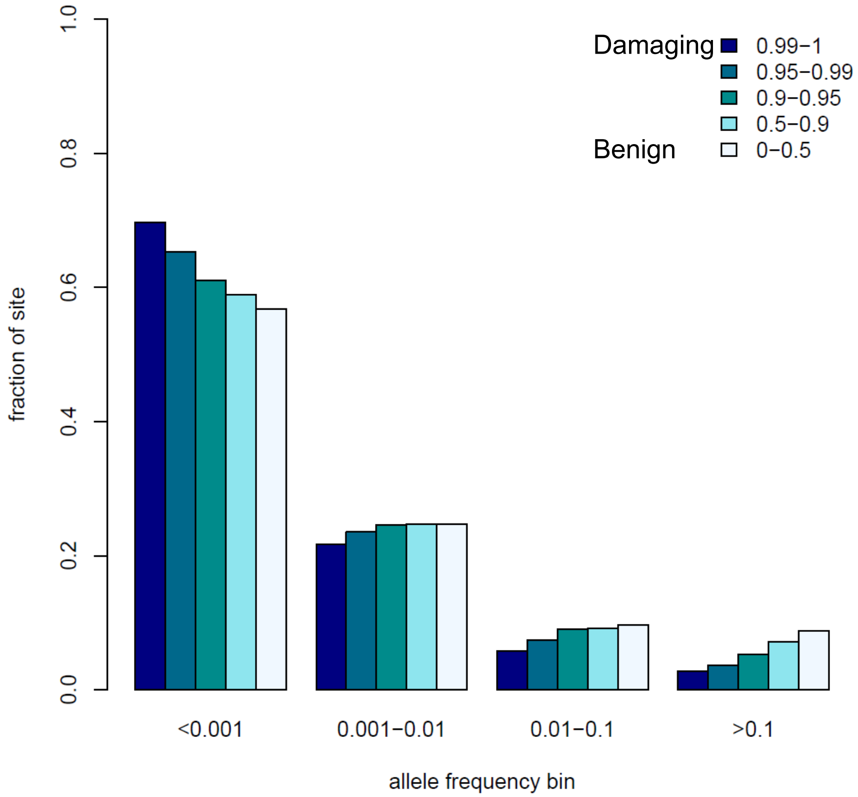
Erik Garrison

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

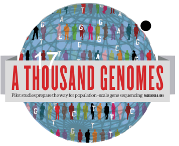
1000 Genomes Project: Present & Future

- First Phase 2 sequence release 14th November 2011
- First Phase 2 alignment release in progress
- 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

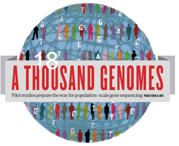
Hapmap, The Pilot Project and The Main Project

- 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 36.6millon snps, 3.8millon indels and 14K deletions
  - 1094 individuals
  - 14 populations
  - Low coverage and exome next generation sequencing
- 1000 Genomes Phase 2
  - Started in 2011
  - 1715 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



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

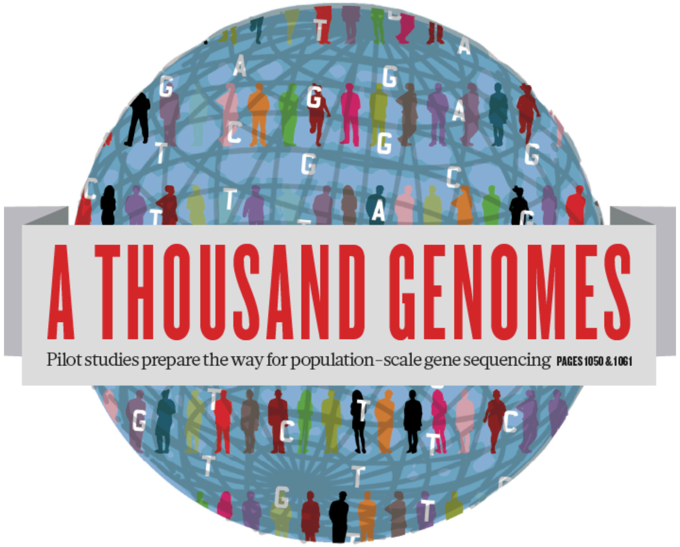
- The project has made more than 10 releases of Alignment Data
- Pilot Project
  - Aligned to NCBI36
  - Maq and Corona
  - Base Quality Recalibration done
- Phase 1
  - Aligned to GRCh37
  - BWA and Bfast
  - Indel Realignment
- Phase 2
  - Aligned to extended GRCh37
  - Improvements to Base Quality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

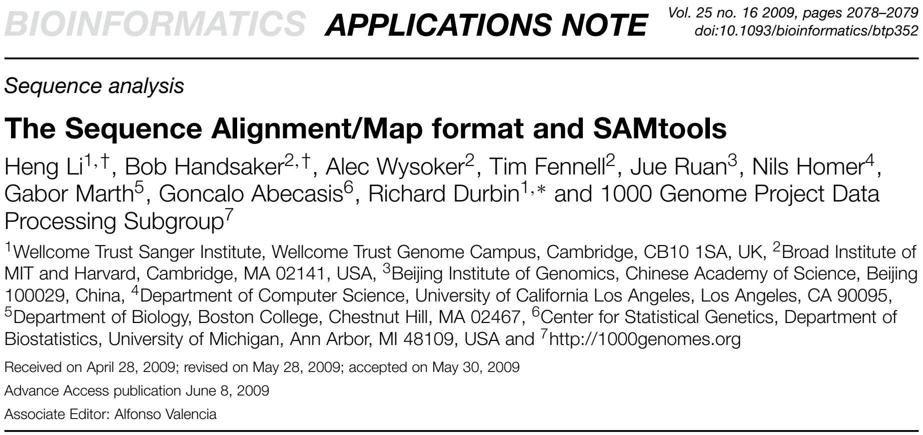

A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project: The Raw Data

What is available on the ftp site

- Sequence Data
  - Fastq files
  - @ERR050087.1 HS18_6628:8:1108:8213:186084#2/1
  - GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
  - +
  - DCDHKHKKIJGNNHIJIIKLLMCLKMAILIJH3K>HL1I=>MK.D<
- Alignment Data
  - BAM files
  - ERR052835.20962733 163 11 60239 0 100M = 60609 469
- Variant Calls
  - VCF files
  - 1 10523 . TCCG T 152 PASS VT=INDEL;RSQ=0.5246; AFR_AF=0.01
- Reference Data Sets
  - Reference genome in fasta
  - Annotation sets in bed or gtf

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]

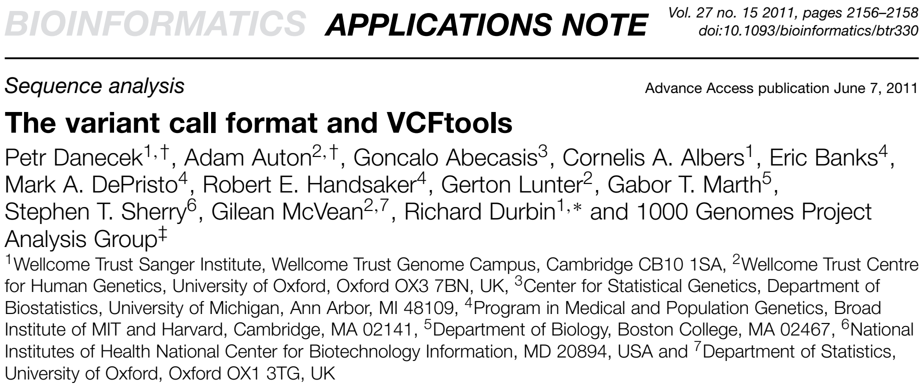
[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California 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

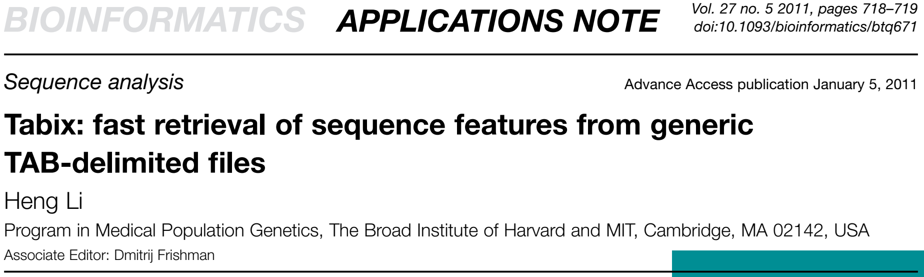
Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

[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

All indexed for fast retrieval



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

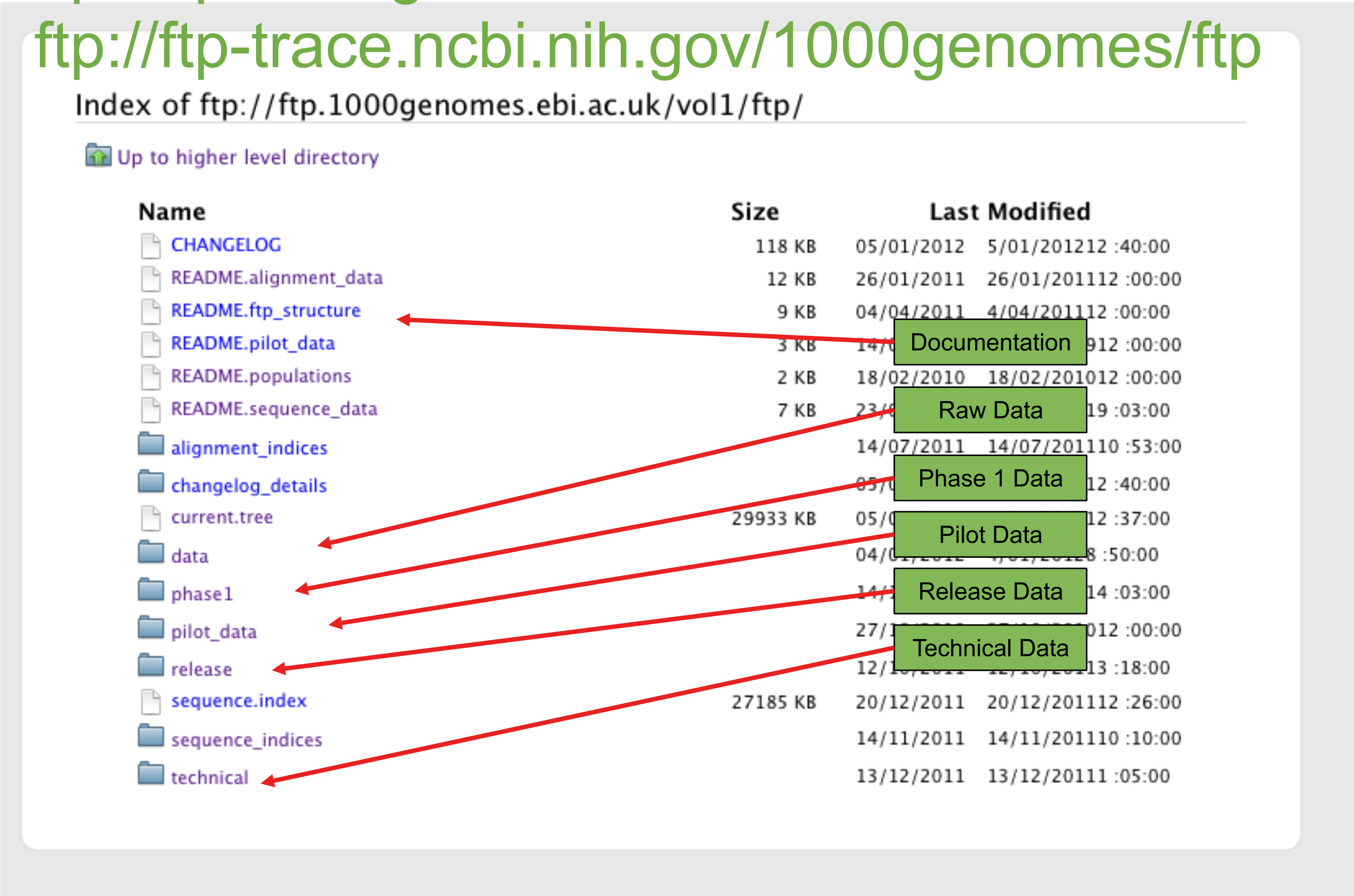
Program in Medical Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

Associate Editor: Dmitrij Frishman

A THOUSAND GENOMES

EMBL-EBI

ftp://ftp.1000genomes.ebi.ac.uk
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp

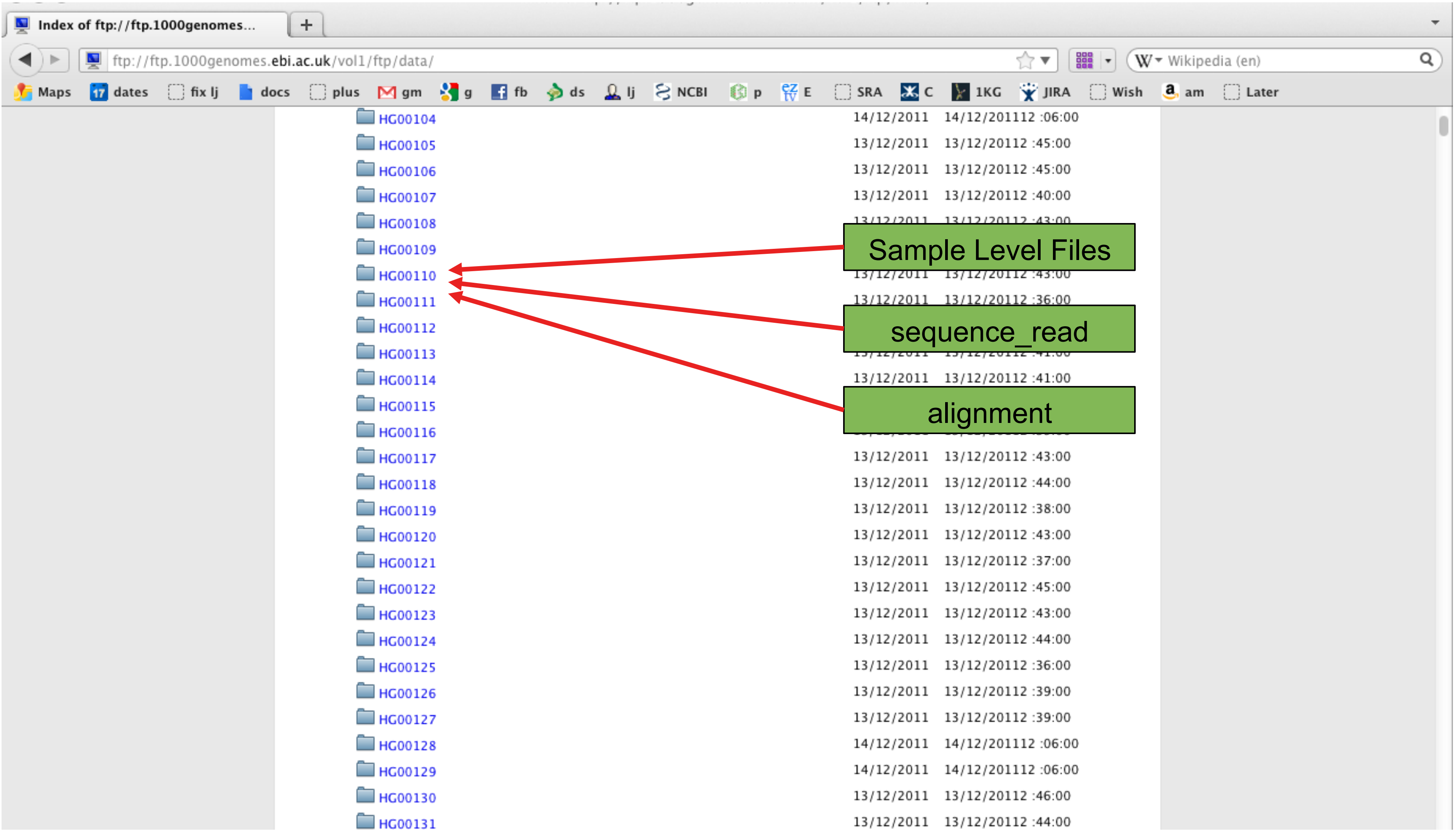


Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/

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	14/0	912 :00:00
README.populations	2 KB	18/02/2010	18/02/201012 :00:00
README.sequence_data	7 KB	23/0	19 :03:00
alignment_indices		14/07/2011	14/07/201110 :53:00
changelog_details		05/0	12 :40:00
current.tree	29933 KB	05/0	12 :37:00
data		04/01/2012	4/01/2012[illegible]8 :50:00
phase1		14/1	14 :03:00
pilot_data		27/1	012 :00:00
release		12/10/2011	12/10/2011[illegible]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

The FTP Site: Data



Index of ftp://ftp.1000genomes...
+
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/
Wikipedia (en)
Maps
dates
fix lj
docs
plus
gm
g
fb
ds
lj
NCBI
p
E
SRA
C
1KG
JIRA
Wish
am
Later

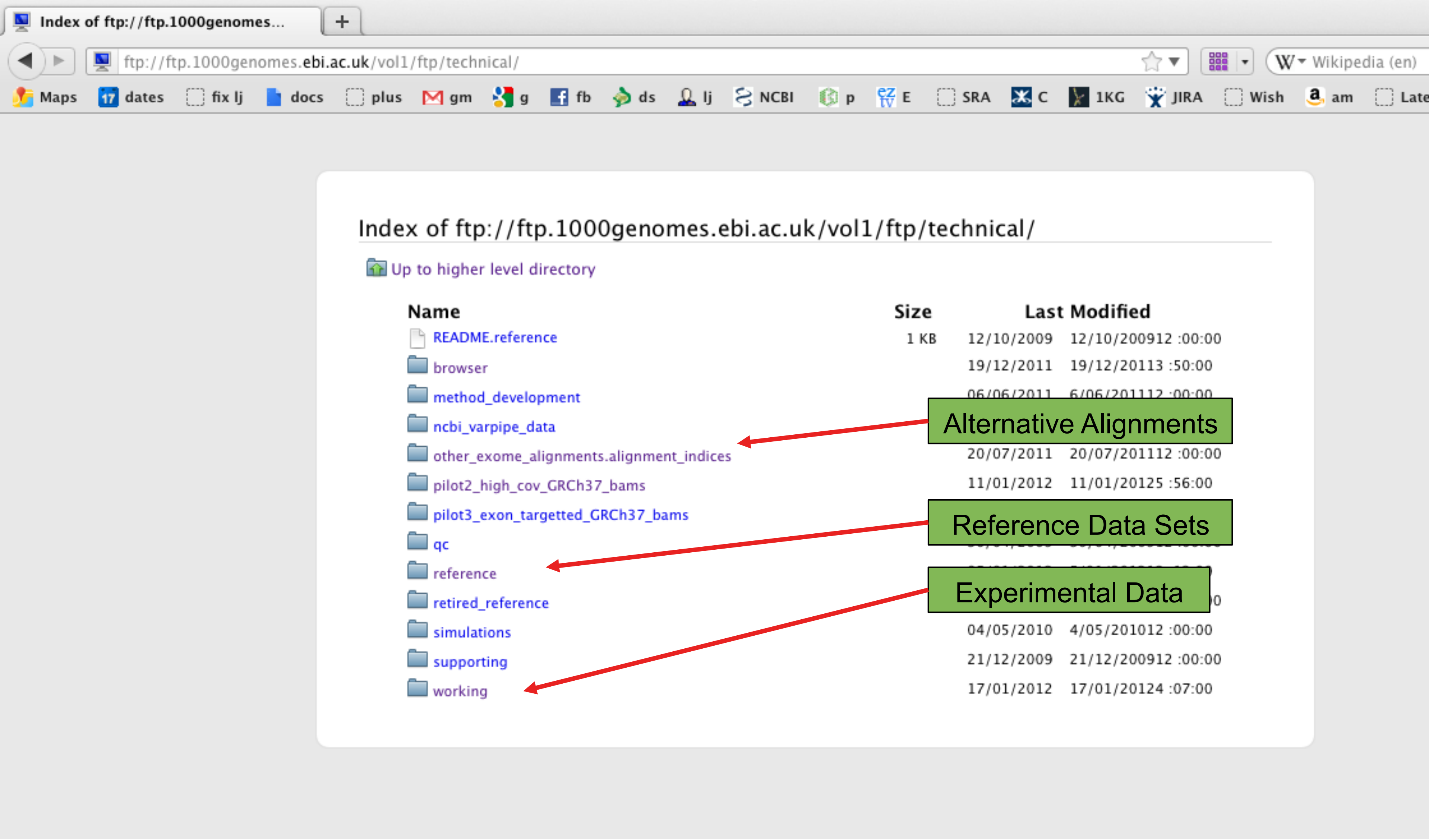
```
HG00104        14/12/2011   14/12/201112 :06:00
HG00105        13/12/2011   13/12/20112 :45:00
HG00106        13/12/2011   13/12/20112 :45:00
HG00107        13/12/2011   13/12/20112 :40:00
HG00108        13/12/2011   13/12/20112 :43:00
HG00109        [illegible]
HG00110        13/12/2011   13/12/20112 :43:00
HG00111        13/12/2011   13/12/20112 :36:00
HG00112        [illegible]
HG00113        13/12/2011   13/12/20112 :41:00
HG00114        13/12/2011   13/12/20112 :41:00
HG00115        [illegible]
HG00116        [illegible]
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

FTP Site: Technical


Index of ftp://ftp.1000genomes...
+
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/
Wikipedia (en)
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

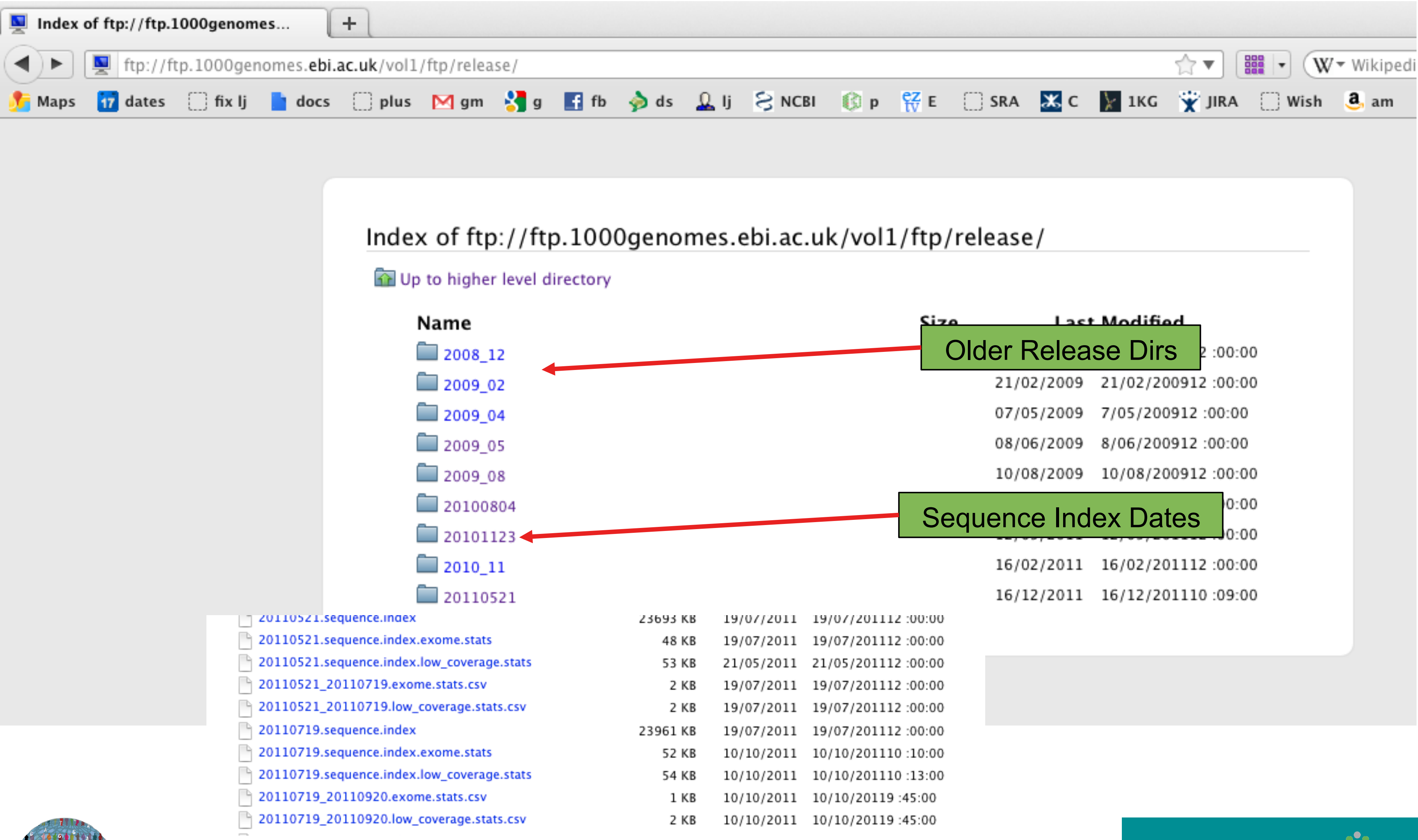
Name	Size	Last Modified	
README.reference	1 KB	12/10/2009	12/10/200912 :00:00
browser		19/12/2011	19/12/20113 :50:00
method_development		06/06/2011	6/06/201112 :00:00
ncbi_varpipe_data		[illegible]	[illegible]
other_exome_alignments.alignment_indices		20/07/2011	20/07/201112 :00:00
pilot2_high_cov_GRCh37_bams		11/01/2012	11/01/20125 :56:00
pilot3_exon_targetted_GRCh37_bams		[illegible]	[illegible]
qc		[illegible]	[illegible]
reference		[illegible]	[illegible]
retired_reference		[illegible]	[illegible]
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/
W Wikipedi
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/

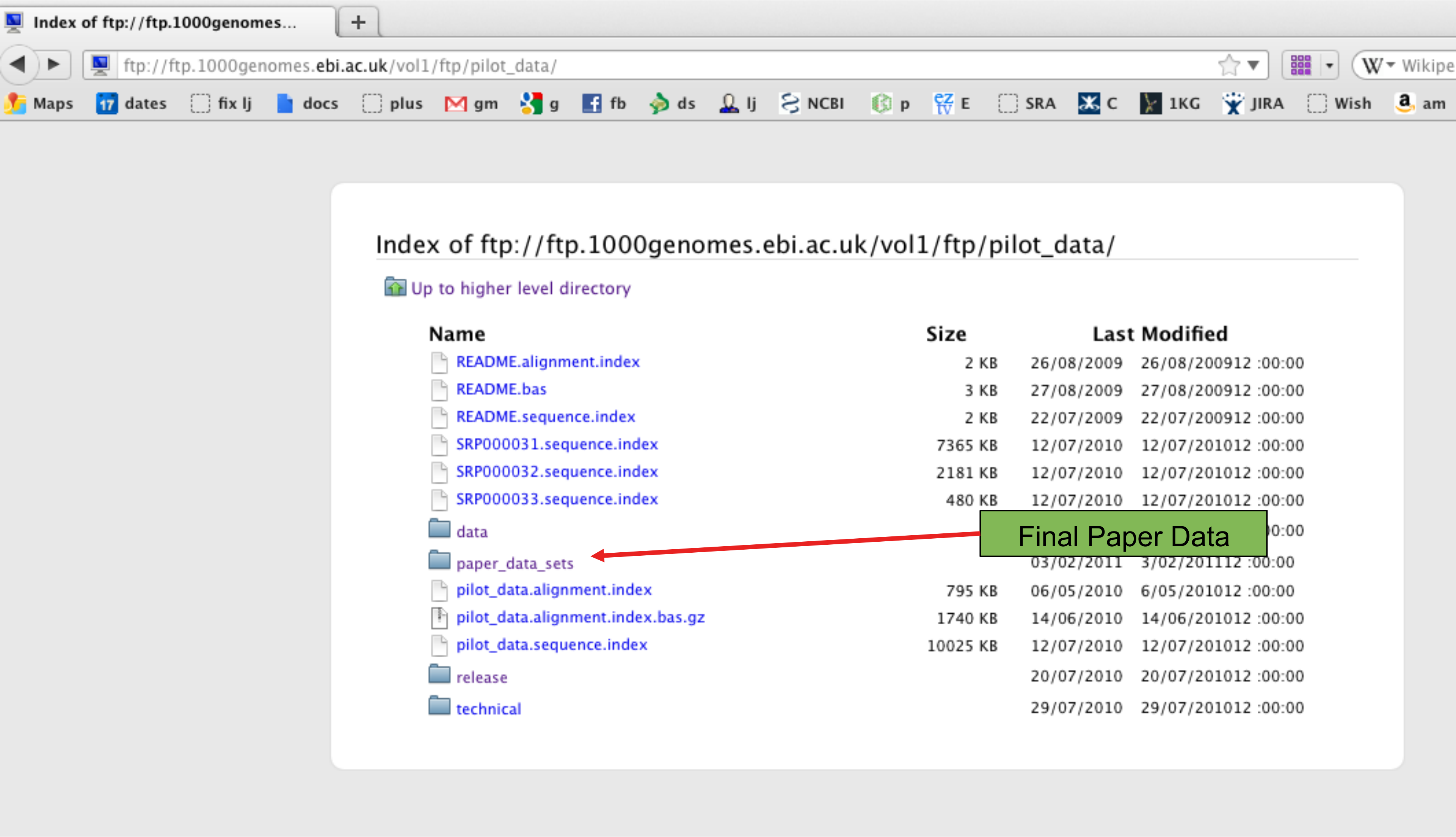
Up to higher level directory

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
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

Older Release Dirs

Sequence Index Dates

FTP Site: Pilot Data

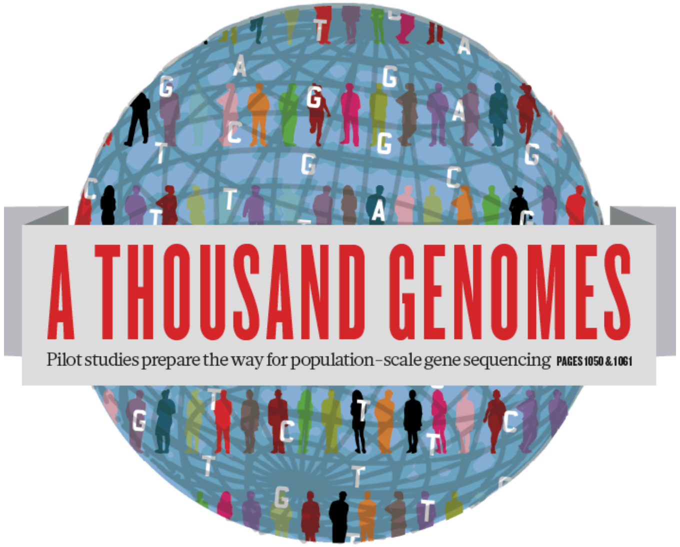

Index of ftp://ftp.1000genomes...
+
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/
W Wikipe
Maps
17 dates
fix lj
docs
plus
gm
g
fb
ds
lj
NCBI
p
E
SRA
C
1KG
JIRA
Wish
am

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

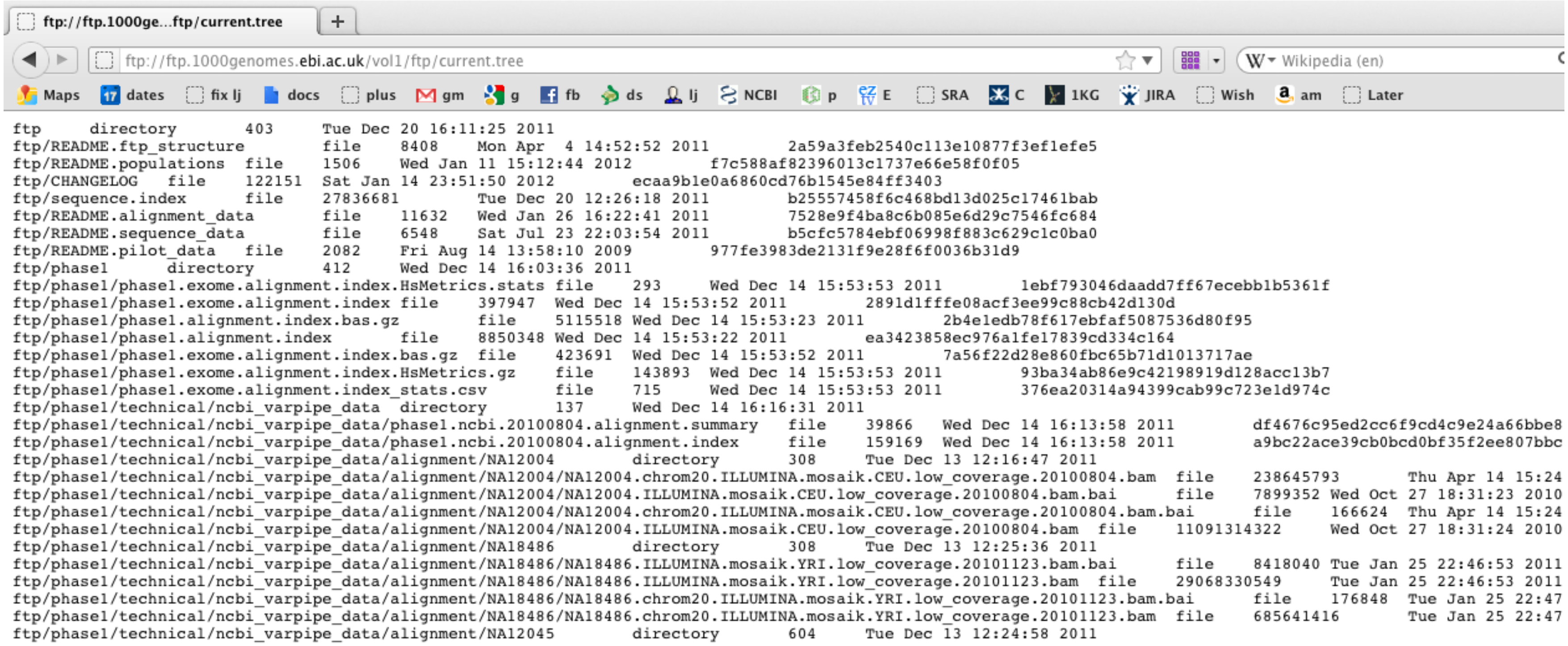
Final Paper Data


A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project: Finding Data

Finding Data

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

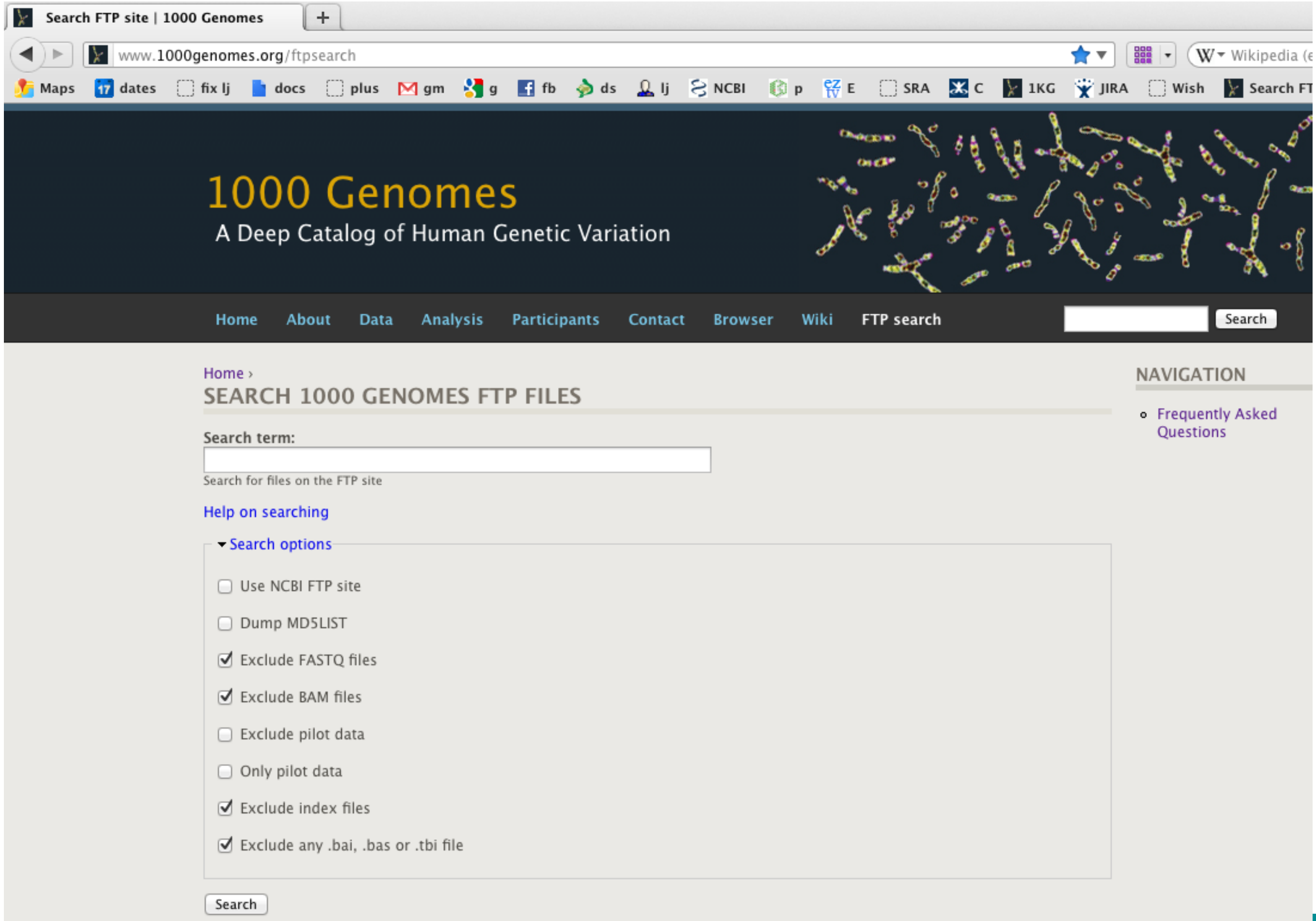

ftp://ftp.1000ge...ftp/current.tree
+
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree
Wikipedia (en)
Maps
dates
fix lj
docs
plus
gm
g
fb
ds
lj
NCBI
p
E
SRA
C
1KG
JIRA
Wish
am
Later

```
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
```

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

Finding Data

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



SEARCH 1000 GENOMES FTP FILES

Search term:

Search for files on the FTP site

Help on searching

Search options

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

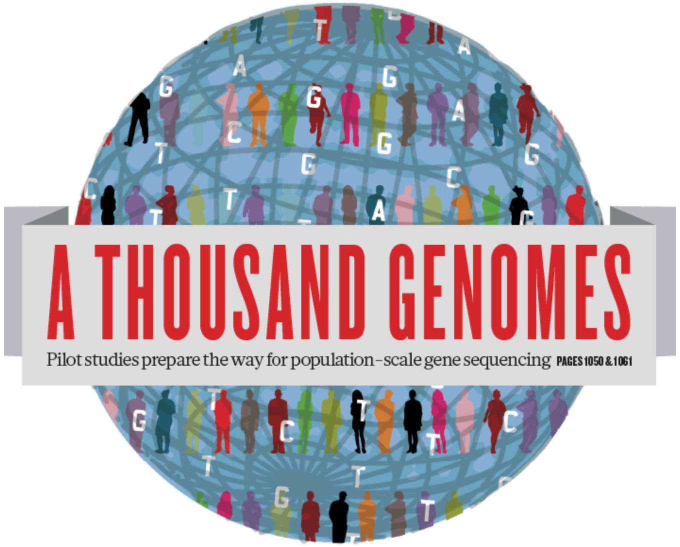
Search

Viewing Files

- All alignment and variant files are indexed so subsections can be downloaded remotely
- Use samtools to get subsections of bam files
  - **samtools view** http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/HG01375.mapped.ILLUMINA.bwa.CLM.low_coverage.20111114.bam 6:31833200-31834200
- Use tabix to get subsections of vcf files
  - **tabix -h** ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b37.vcf.gz 6:31833200-31834200
- You can also use the web Data Slicer interface to do this
  - http://browser.1000genomes.org/Homo_sapiens/UserData/SelectSlice

More Information

- Sam/Bam format
- http://samtools.sourceforge.net/
- samtools-help@lists.sourceforge.net
- VCF format
- http://vcftools.sourceforge.net/
- vcftools-help@lists.sourceforge.net


A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project:

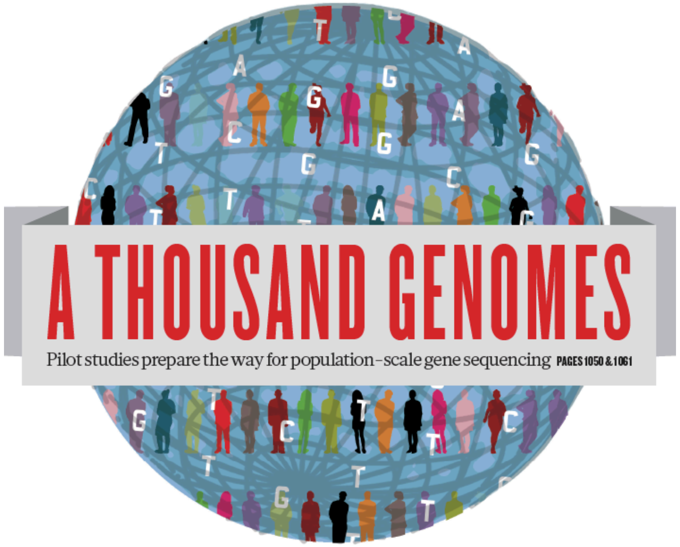
Exercise 1: Finding Data and viewing data on the 1000 genomes ftp site

Exercise: Finding and Viewing Data

- Finding data can use either ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree or http://www.1000genomes.org/ftpsearch
- Find a omni vcf file
- View the data it contains for the region 6:31831625-31834704
- Find the mapped low coverage bam file for HG01375
- View the data it contains for the same region

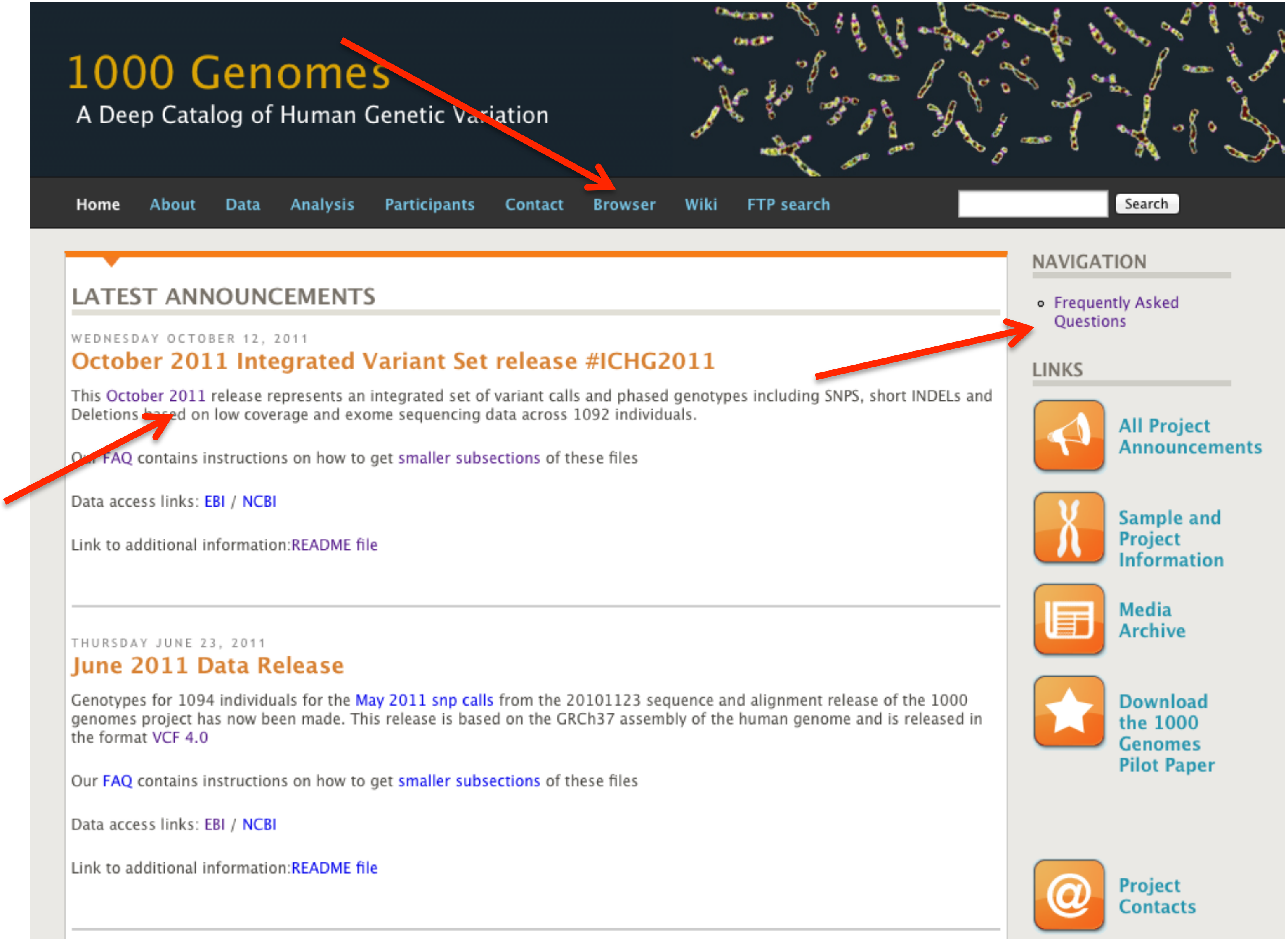
Answers

- **wget** ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree
- **grep** omni current.tree | grep vcf | grep 2012 | grep -v tbi | cut –f1
- **tabix -h** ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b37.vcf.gz 6:31831625-31834704
- **grep** HG01375 current.tree | grep low_coverage | grep mapped | grep -v bai | grep -v bas | grep -v unmapped | cut -f1
- **samtools view** http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/HG01375.mapped.ILLUMINA.bwa.CLM.low_coverage.20111114.bam 6:31831625-31834704


A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Website and Ensembl- style Browser

http://www.1000genomes.org



1000 Genomes

A Deep Catalog of Human Genetic Variation

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

NAVIGATION

- Frequently Asked Questions

LINKS

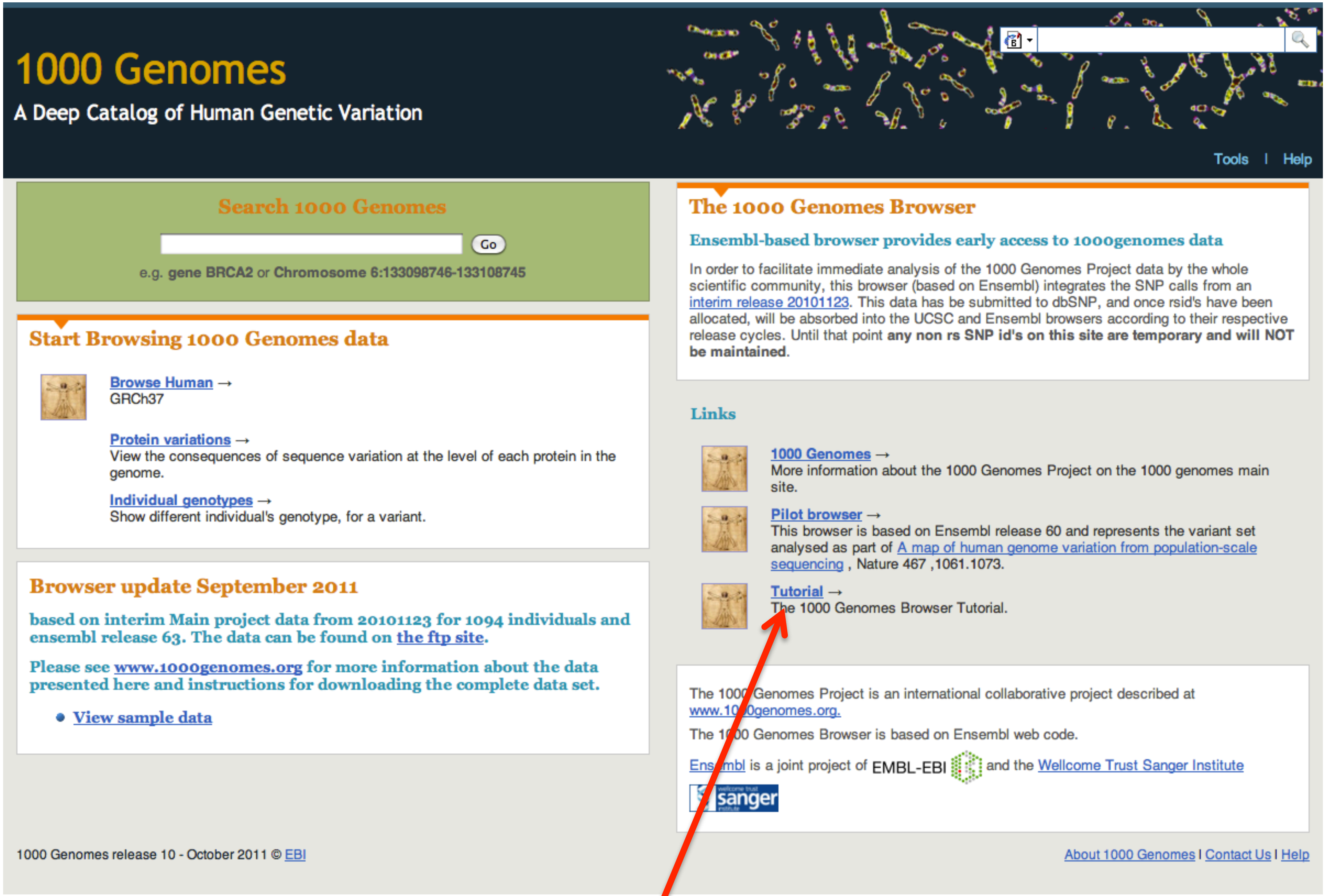
All Project Announcements

Sample and Project Information

Media Archive

Download the 1000 Genomes Pilot Paper

Project Contacts



1000 Genomes

A Deep Catalog of Human Genetic Variation

Search 1000 Genomes

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

Start Browsing 1000 Genomes data

Browse Human →
GRCh37

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

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

Browser update September 2011

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

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

- View sample data

The 1000 Genomes Browser

Ensembl-based browser provides early access to 1000genomes data

In order to facilitate immediate analysis of the 1000 Genomes Project data by the whole scientific community, this browser (based on Ensembl) integrates the SNP calls from an interim release 20101123. This data has 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

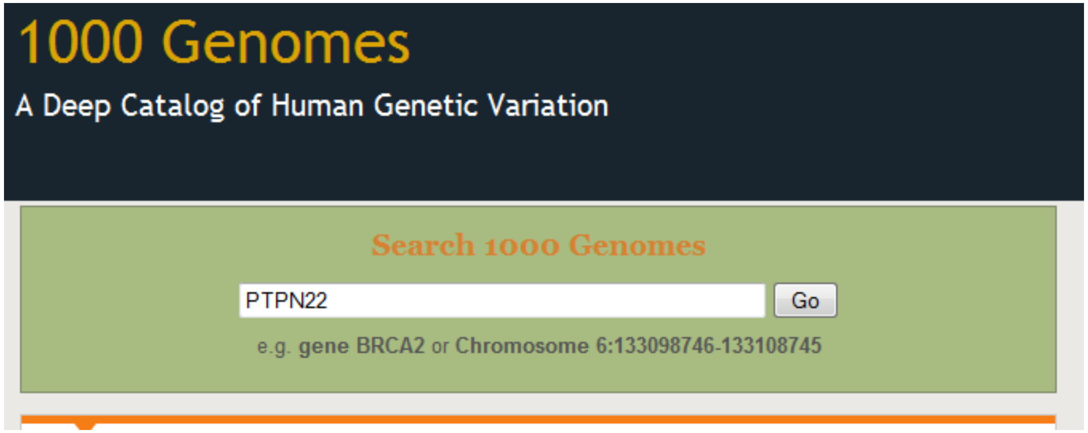
1000 Genomes release 10 - October 2011 © EBI

About 1000 Genomes | Contact Us | Help

http://browser.1000genomes.org

Searching the Browser

- *http://browser.1000genomes.org*



1000 Genomes

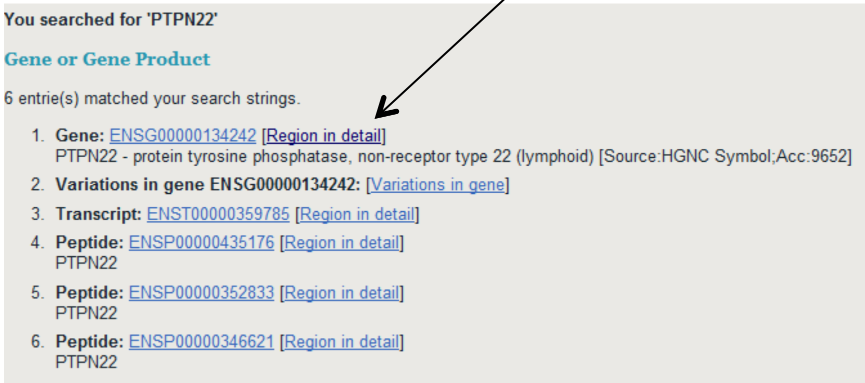
A Deep Catalog of Human Genetic Variation

Search 1000 Genomes

PTPN22 Go

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

- Search for PTPN22
- Click ‘Region in Detail’



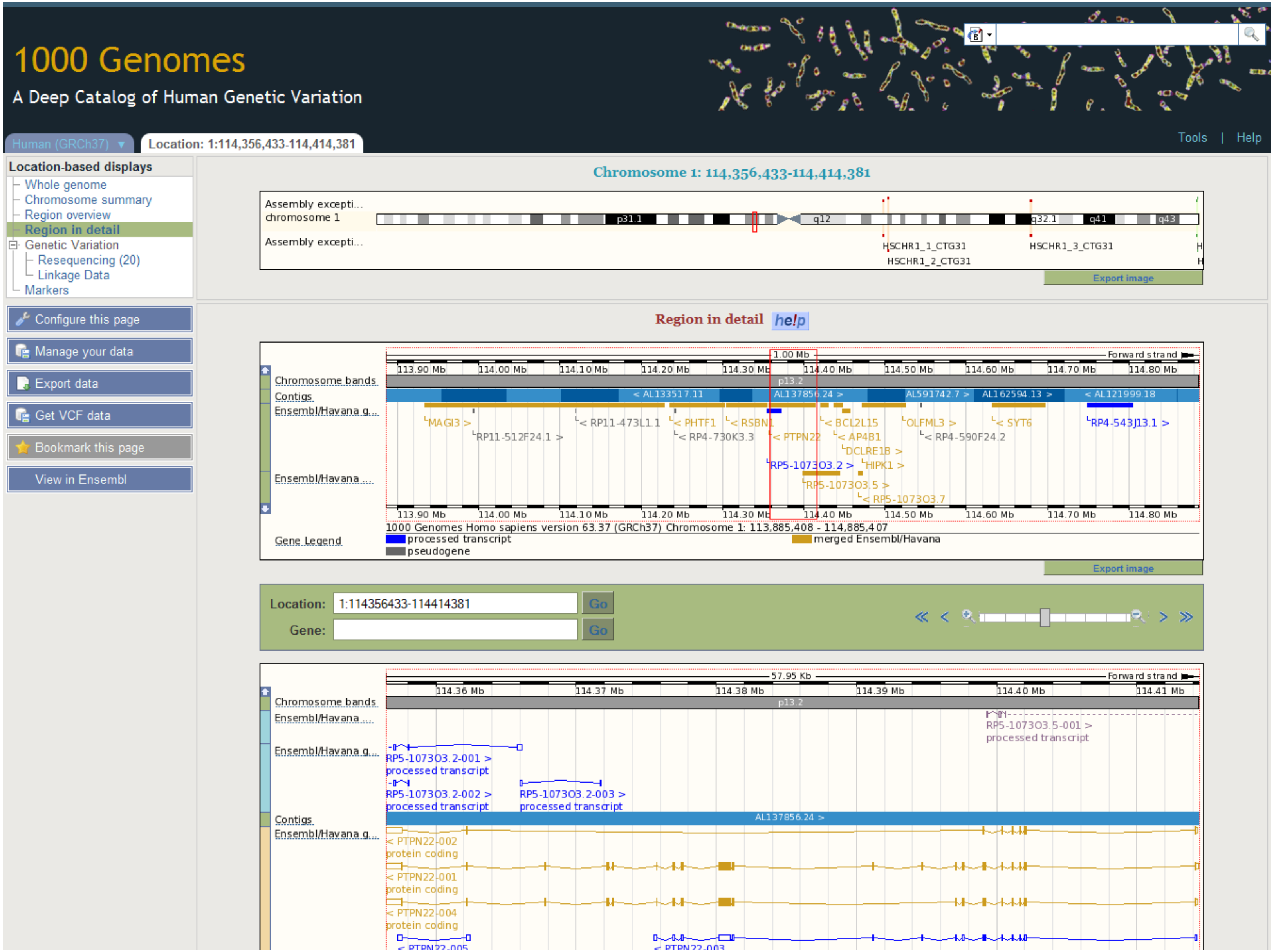
You searched for 'PTPN22'

Gene or Gene Product

6 entrie(s) matched your search strings.

1. **Gene:** ENSG00000134242 [Region in detail]
   PTPN22 - protein tyrosine phosphatase, non-receptor type 22 (lymphoid) [Source:HGNC Symbol;Acc:9652]
2. **Variations in gene ENSG00000134242:** [Variations in gene]
3. **Transcript:** ENST00000359785 [Region in detail]
4. **Peptide:** ENSP00000435176 [Region in detail]
   PTPN22
5. **Peptide:** ENSP00000352833 [Region in detail]
   PTPN22
6. **Peptide:** ENSP00000346621 [Region in detail]
   PTPN22

Region in Detail

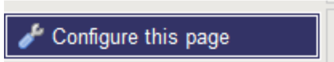


Location: 1:114,356,433-114,414,381

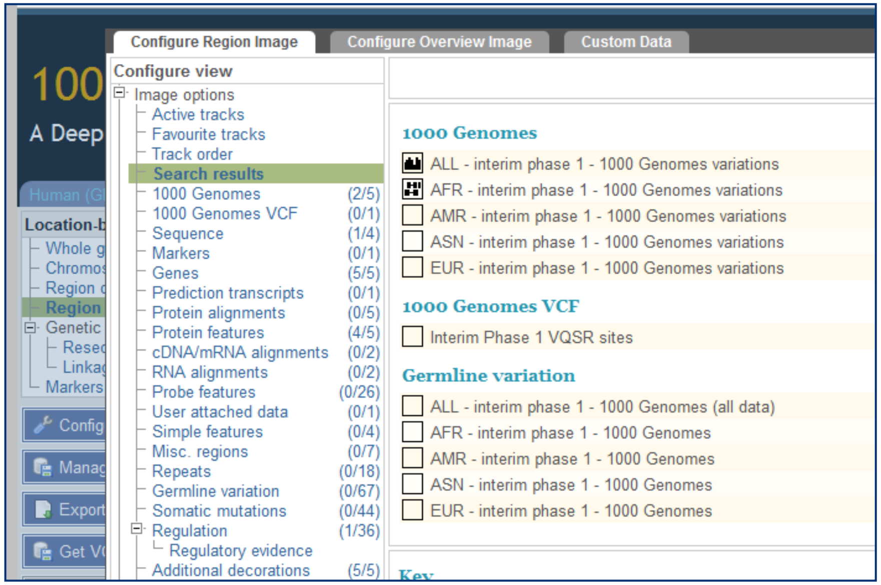
1000 Genomes Homo sapiens version 63.37 (GRCh37) Chromosome 1: 113,885,408 - 114,885,407

Location: 1:114356433-114414381

Turning on Tracks



Configure this page



Configure Region Image | Configure Overview Image | Custom Data

100
A Deep
Human (G
Location-b
Whole g
Chromos
Region c
Region
Genetic
Resea
Linkag
Markers
Config
Manag
Export
Get V

Configure view

- Image options
  - Active tracks
  - Favourite tracks
  - Track order
  - Search results
  - 1000 Genomes (2/5)
  - 1000 Genomes VCF (0/1)
  - Sequence (1/4)
  - Markers (0/1)
  - Genes (5/5)
  - Prediction transcripts (0/1)
  - Protein alignments (0/5)
  - Protein features (4/5)
  - cDNA/mRNA alignments (0/2)
  - RNA alignments (0/2)
  - Probe features (0/26)
  - User attached data (0/1)
  - Simple features (0/4)
  - Misc. regions (0/7)
  - Repeats (0/18)
  - Germline variation (0/67)
  - Somatic mutations (0/44)
  - Regulation (1/36)
    - Regulatory evidence
  - Additional decorations (5/5)

1000 Genomes

☑ ALL - interim phase 1 - 1000 Genomes variations
☑ AFR - interim phase 1 - 1000 Genomes variations
□ AMR - interim phase 1 - 1000 Genomes variations
□ ASN - interim phase 1 - 1000 Genomes variations
□ EUR - interim phase 1 - 1000 Genomes variations

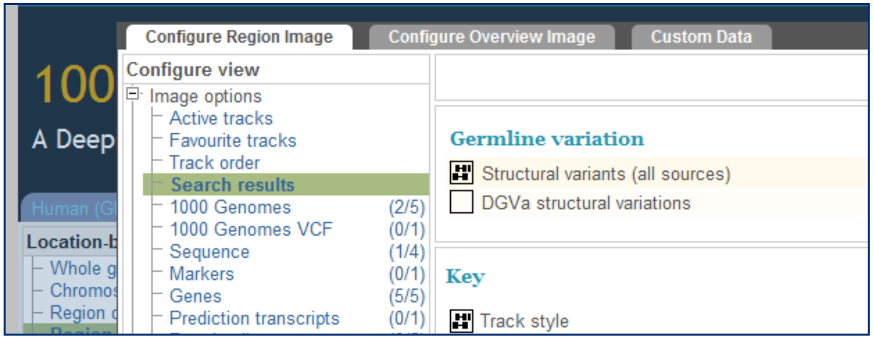
1000 Genomes VCF

□ Interim Phase 1 VQSR sites

Germline variation

□ ALL - interim phase 1 - 1000 Genomes (all data)
□ AFR - interim phase 1 - 1000 Genomes
□ AMR - interim phase 1 - 1000 Genomes
□ ASN - interim phase 1 - 1000 Genomes
□ EUR - interim phase 1 - 1000 Genomes

Key



Configure Region Image | Configure Overview Image | Custom Data

100
A Deep
Human (G
Location-b
Whole g
Chromos
Region c

Configure view

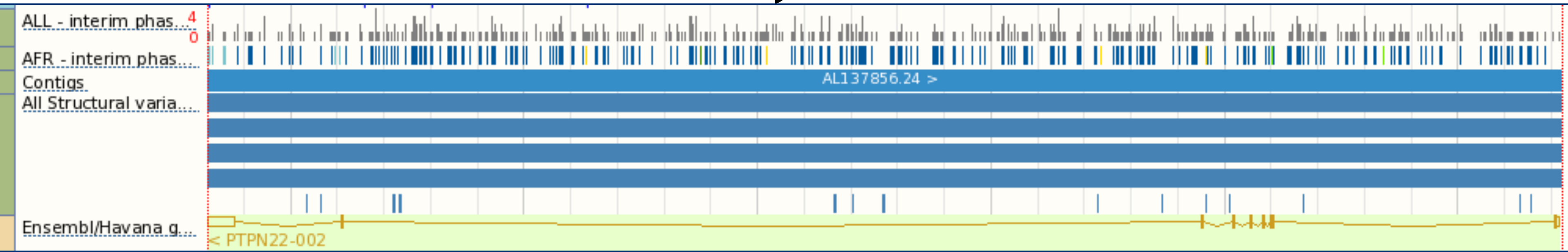
- Image options
  - Active tracks
  - Favourite tracks
  - Track order
  - Search results
  - 1000 Genomes (2/5)
  - 1000 Genomes VCF (0/1)
  - Sequence (1/4)
  - Markers (0/1)
  - Genes (5/5)
  - Prediction transcripts (0/1)

Germline variation

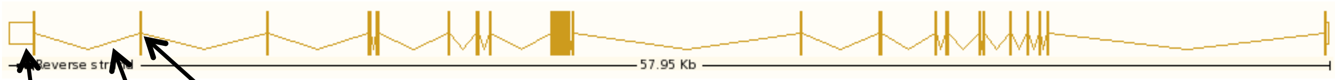
☑ Structural variants (all sources)
□ DGVa structural variations

Key

☑ Track style



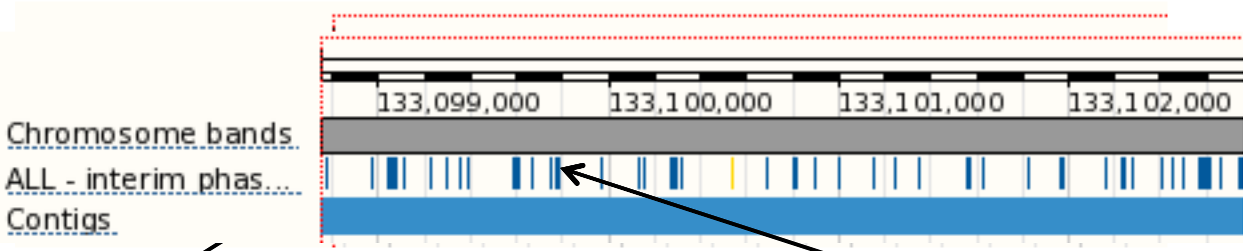
Genes and SNPs



UTR

Intron

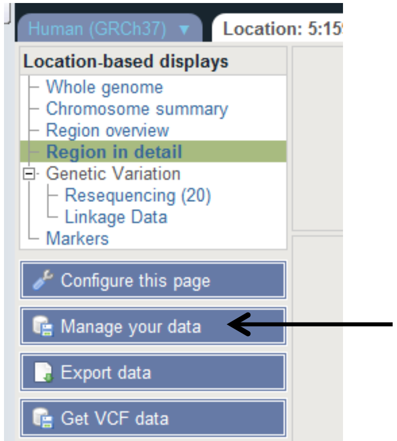
Coding



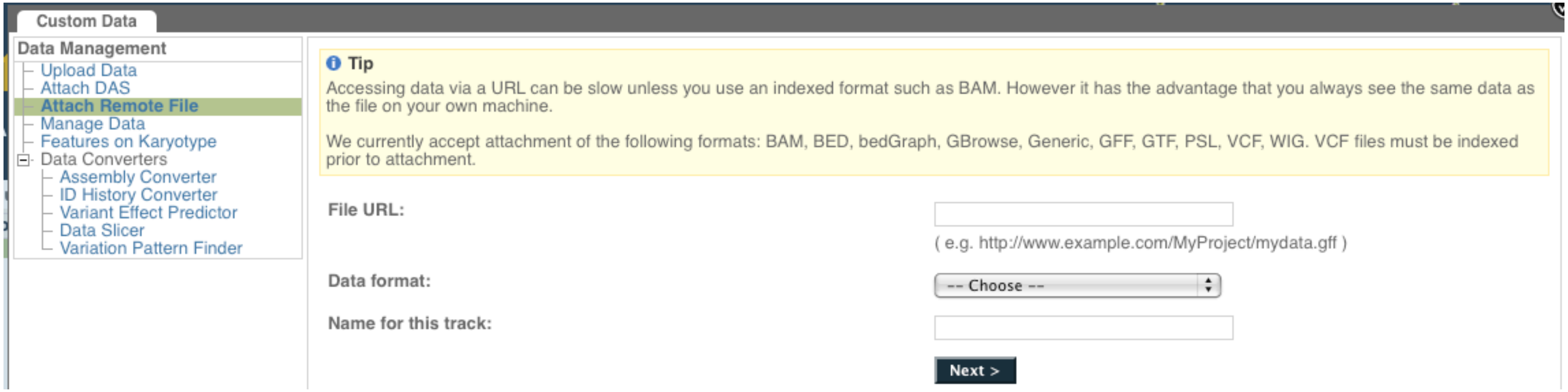
Line indicates number of SNPS

Each Line is One SNP

File upload to view with 1000 Genomes data



Manage your data



Tip

Accessing data via a URL can be slow unless you use an indexed format such as BAM. However it has the advantage that you always see the same data as the file on your own machine.

We currently accept attachment of the following formats: BAM, BED, bedGraph, GBrowse, Generic, GFF, GTF, PSL, VCF, WIG. VCF files must be indexed prior to attachment.

File URL:

(e.g. http://www.example.com/MyProject/mydata.gff)

Data format: -- Choose --

Name for this track:

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

* VCF must be indexed

Uploaded VCF

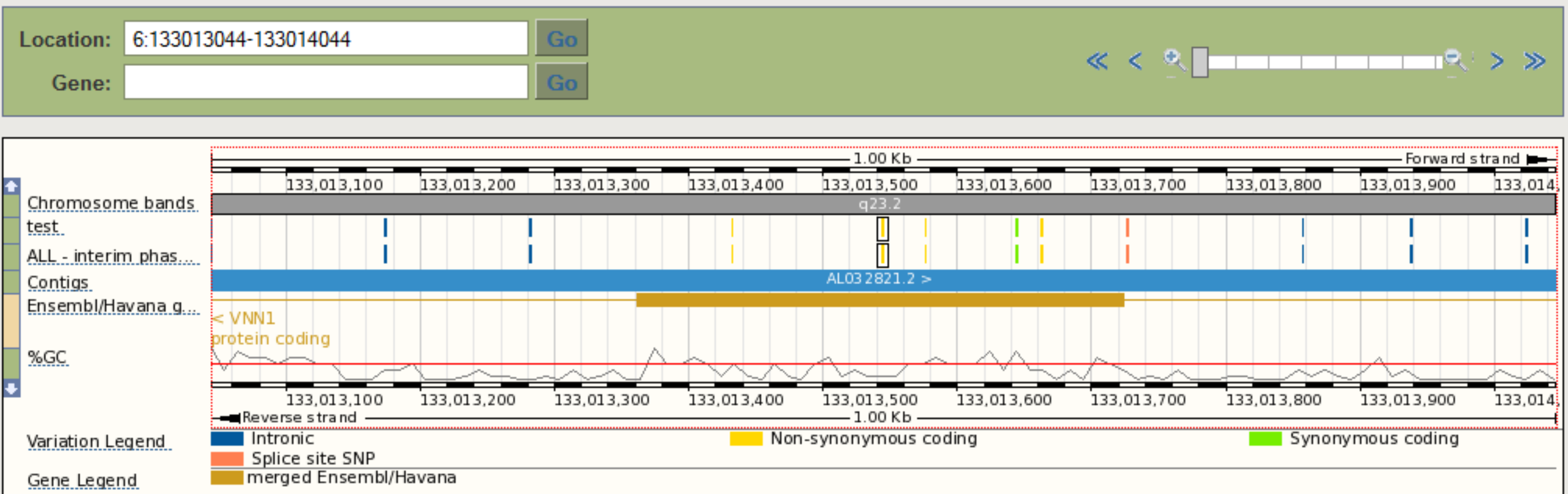
Example:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.wgs.phase1_integrated_calls.20101123.snps_indels_svs.sites.vcf.gz

OR find this at:

http://tinyurl.com/1000vcf

(but don’t use this address as the input URL- rather, copy the ftp link.)



Location: 6:133013044-133014044

Gene:

Uploaded BAM

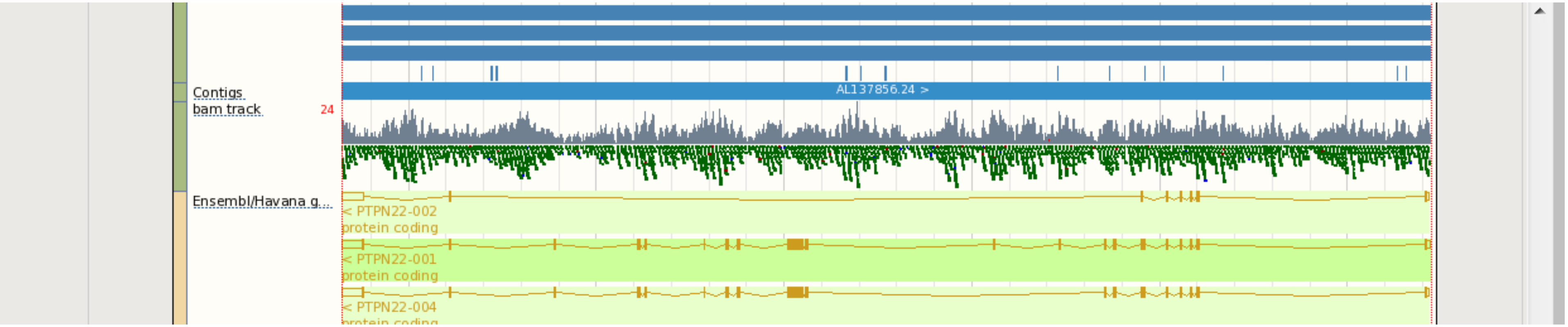
Example:

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/
HG01375.mapped.ILLUMINA.bwa.CLM.low_coverage.20111114.bam

OR find this at:

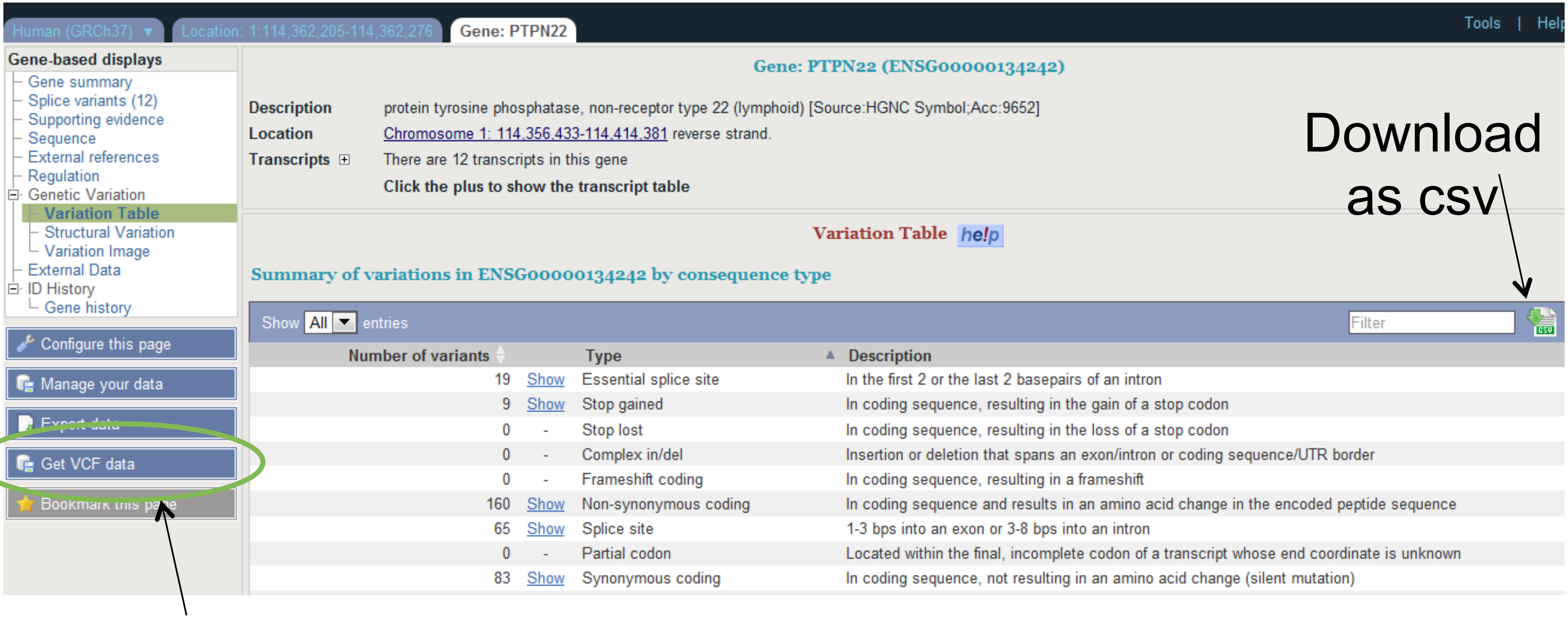
http://tinyurl.com/1000bam

(but don’t use this address as the input URL- rather, copy the ftp link.)



Back to browsing…

Click the Gene tab, then ‘Variation Table’ or ‘Variation Image’



Gene-based displays

- Gene summary
- Splice variants (12)
- Supporting evidence
- Sequence
- External references
- Regulation
- Genetic Variation
  - Variation Table
  - Structural Variation
  - Variation Image
- External Data
- ID History
  - Gene history

Gene: PTPN22 (ENSG00000134242)

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.

Transcripts ⊞ There are 12 transcripts in this gene

Click the plus to show the transcript table

Variation Table he!p

Summary of variations in ENSG00000134242 by consequence type

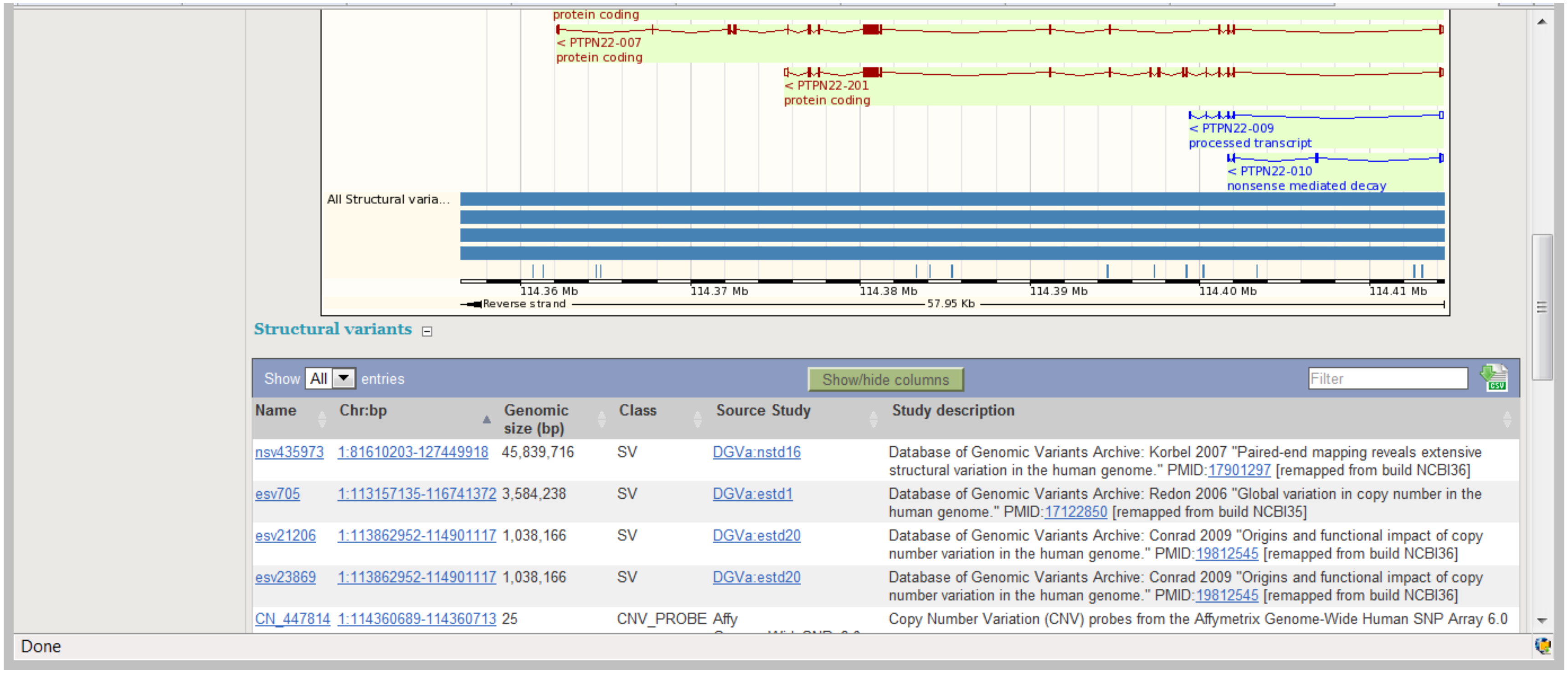
Show All ▼ entries Filter

Number of variants		Type	Description
19	Show	Essential splice site	In the first 2 or the last 2 basepairs of an intron
9	Show	Stop gained	In coding sequence, resulting in the gain of a stop codon
0	-	Stop lost	In coding sequence, resulting in the loss of a stop codon
0	-	Complex in/del	Insertion or deletion that spans an exon/intron or coding sequence/UTR border
0	-	Frameshift coding	In coding sequence, resulting in a frameshift
160	Show	Non-synonymous coding	In coding sequence and results in an amino acid change in the encoded peptide sequence
65	Show	Splice site	1-3 bps into an exon or 3-8 bps into an intron
0	-	Partial codon	Located within the final, incomplete codon of a transcript whose end coordinate is unknown
83	Show	Synonymous coding	In coding sequence, not resulting in an amino acid change (silent mutation)

Download as csv

Get in vcf format

Structural variation (in the Gene tab)



Structural variants ⊟

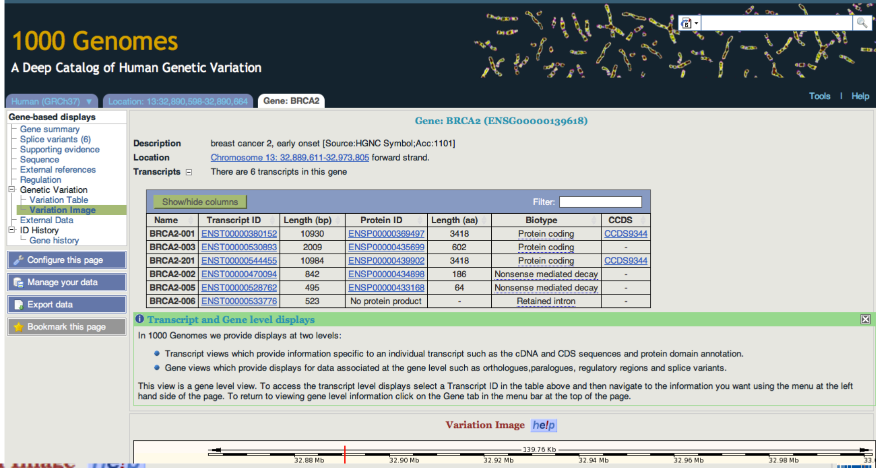
Show All ▼ entries | Show/hide columns | Filter

Name	Chr:bp	Genomic size (bp)	Class	Source Study	Study description
nsv435973	1:81610203-127449918	45,839,716	SV	DGVa:nstd16	Database of Genomic Variants Archive: Korbel 2007 "Paired-end mapping reveals extensive structural variation in the human genome." PMID:17901297 [remapped from build NCBI36]
esv705	1:113157135-116741372	3,584,238	SV	DGVa:estd1	Database of Genomic Variants Archive: Redon 2006 "Global variation in copy number in the human genome." PMID:17122850 [remapped from build NCBI35]
esv21206	1:113862952-114901117	1,038,166	SV	DGVa:estd20	Database of Genomic Variants Archive: Conrad 2009 "Origins and functional impact of copy number variation in the human genome." PMID:19812545 [remapped from build NCBI36]
esv23869	1:113862952-114901117	1,038,166	SV	DGVa:estd20	Database of Genomic Variants Archive: Conrad 2009 "Origins and functional impact of copy number variation in the human genome." PMID:19812545 [remapped from build NCBI36]
CN_447814	1:114360689-114360713	25	CNV_PROBE	Affy	Copy Number Variation (CNV) probes from the Affymetrix Genome-Wide Human SNP Array 6.0

Done

Variation Image

- Gene variation zoom



1000 Genomes

A Deep Catalog of Human Genetic Variation

Human (GRCh37) | Location: 13:32,890,598-32,890,664 | Gene: BRCA2

Tools | Help

Gene-based displays
- Gene summary
- Splice variants (6)
- Supporting evidence
- Sequence
- External references
- Regulation
- Genetic Variation
  - Variation Table
  - Variation Image
- External Data
- ID History
  - Gene history

Configure this page

Manage your data

Export data

Bookmark this page

Gene: BRCA2 (ENSG00000139618)

Description breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]

Location Chromosome 13: 32,889,611-32,973,805 forward strand.

Transcripts There are 6 transcripts in this gene

Show/hide columns | Filter:

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
BRCA2-001	ENST00000380152	10930	ENSP00000369497	3418	Protein coding	CCDS9344
BRCA2-003	ENST00000530893	2009	ENSP00000435699	602	Protein coding	-
BRCA2-201	ENST00000544455	10984	ENSP00000439902	3418	Protein coding	CCDS9344
BRCA2-002	ENST00000470094	842	ENSP00000434898	186	Nonsense mediated decay	-
BRCA2-005	ENST00000528762	495	ENSP00000433168	64	Nonsense mediated decay	-
BRCA2-006	ENST00000533776	523	No protein product	-	Retained intron	-

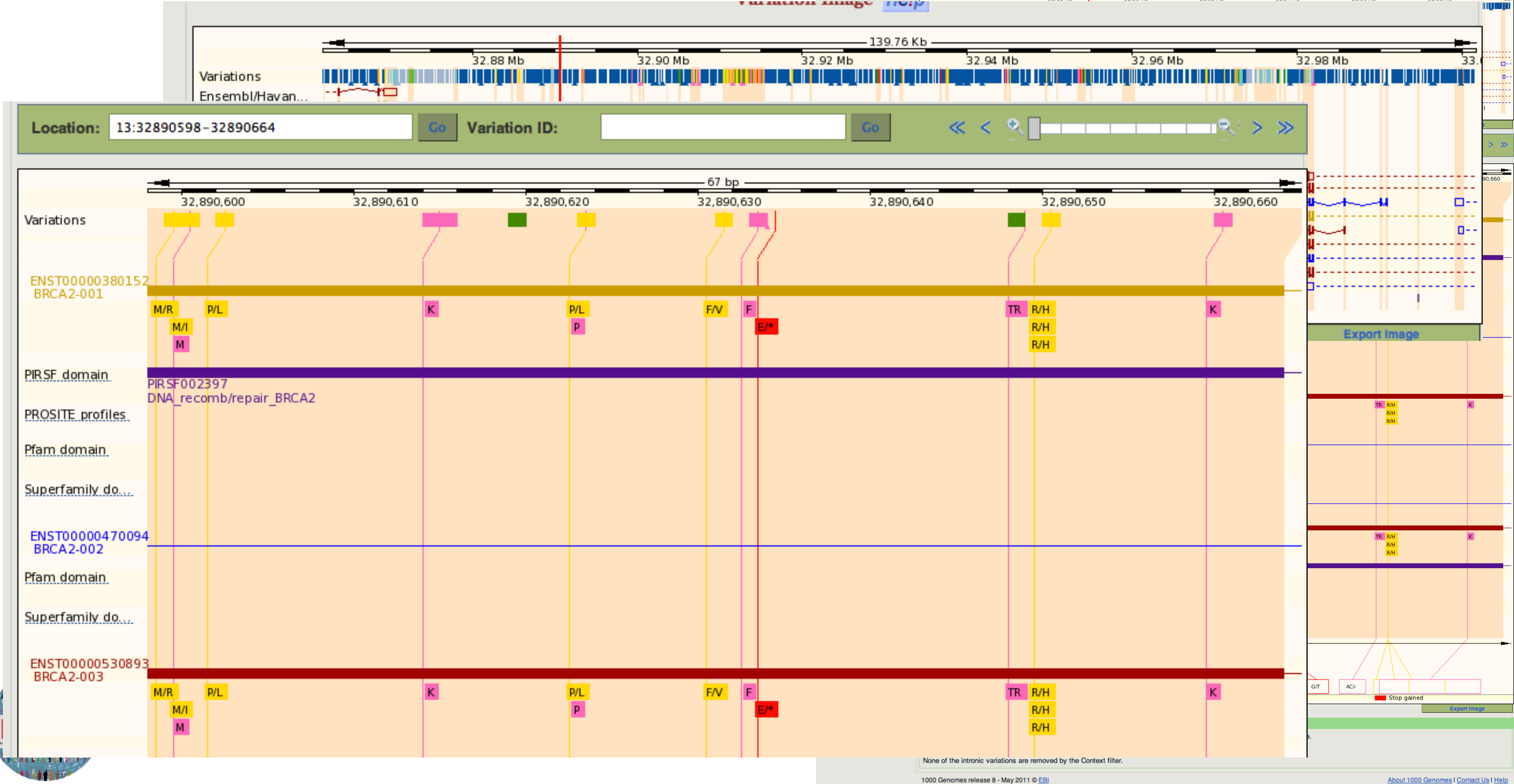
Transcript and Gene level displays

In 1000 Genomes we provide displays at two levels:

- Transcript views which provide information specific to an individual transcript such as the cDNA and CDS sequences and protein domain annotation.
- Gene views which provide displays for data associated at the gene level such as orthologues,paralogues, regulatory regions and splice variants.

This view is a gene level view. To access the transcript level displays select a Transcript ID in the table above and then navigate to the information you want using the menu at the left hand side of the page. To return to viewing gene level information click on the Gene tab in the menu bar at the top of the page.

Variation Image help



Location: 13:32890598-32890664 Go **Variation ID:** Go

None of the intronic variations are removed by the Context filter.

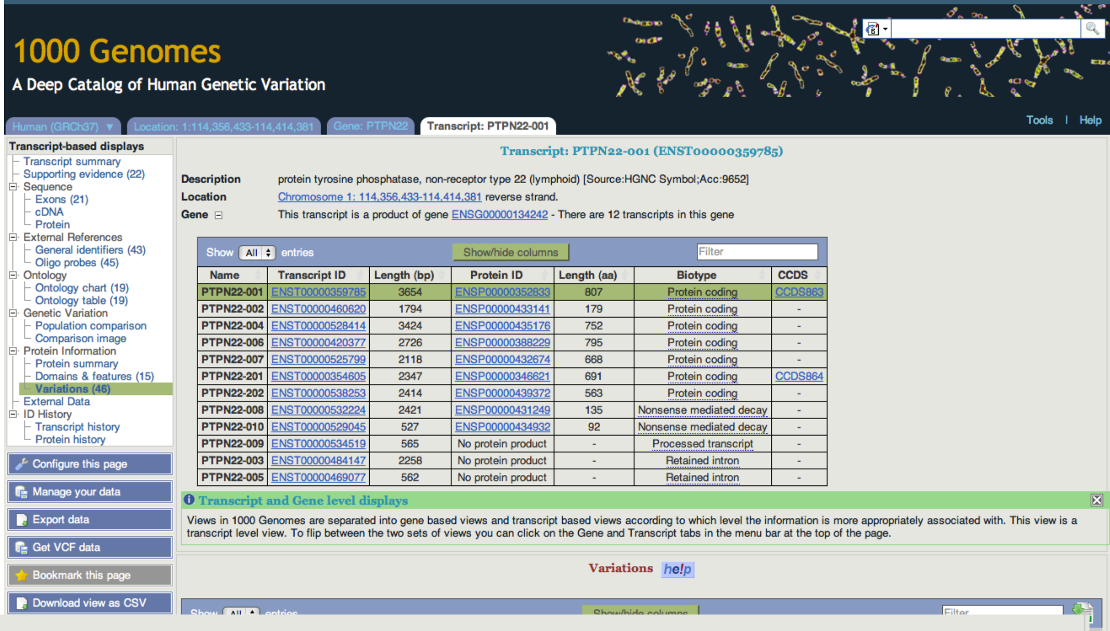
1000 Genomes release 8 - May 2011 © EBI

About 1000 Genomes | Contact Us | Help

Transcript Tab: Variations

Effect on Protein:

- SIFT
- PolyPhen



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

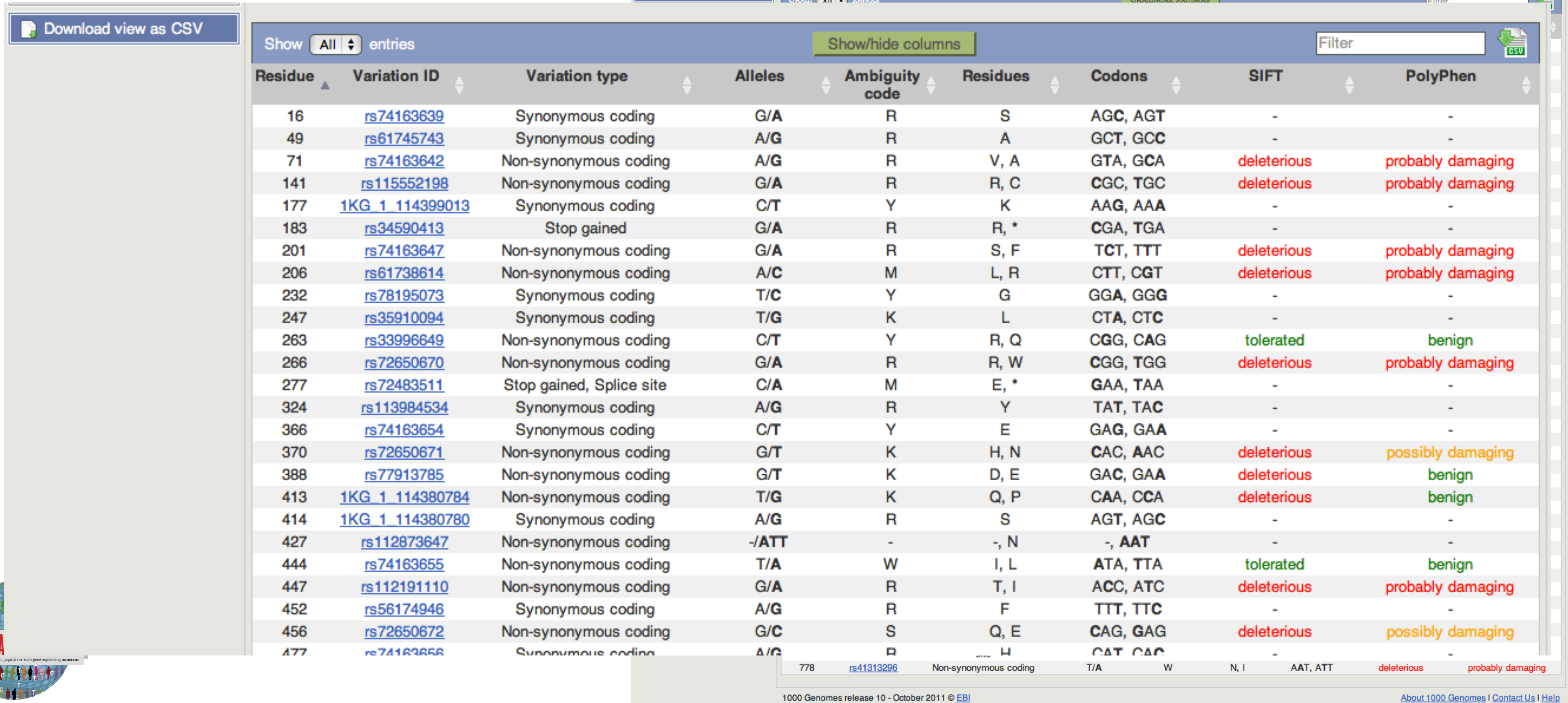
Show All entries | Show/hide columns | Filter

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



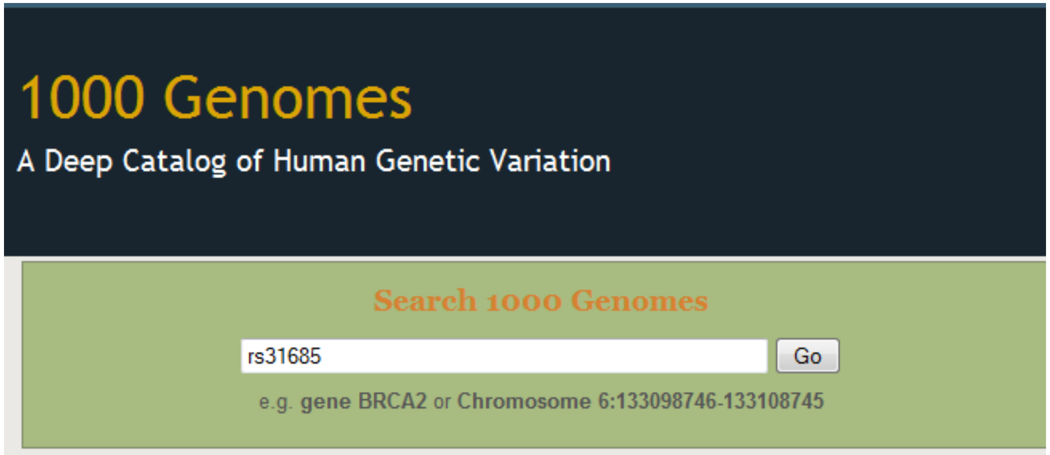
Residue	Variation ID	Variation type	Alleles	Ambiguity code	Residues	Codons	SIFT	PolyPhen
16	rs74163639	Synonymous coding	G/A	R	S	AGC, AGT	-	-
49	rs61745743	Synonymous coding	A/G	R	A	GCT, GCC	-	-
71	rs74163642	Non-synonymous coding	A/G	R	V, A	GTA, GCA	deleterious	probably damaging
141	rs115552198	Non-synonymous 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

1000 Genomes release 10 - October 2011 © EBI

About 1000 Genomes | Contact Us | Help

Start again- search for a variation (rs31685)



1000 Genomes

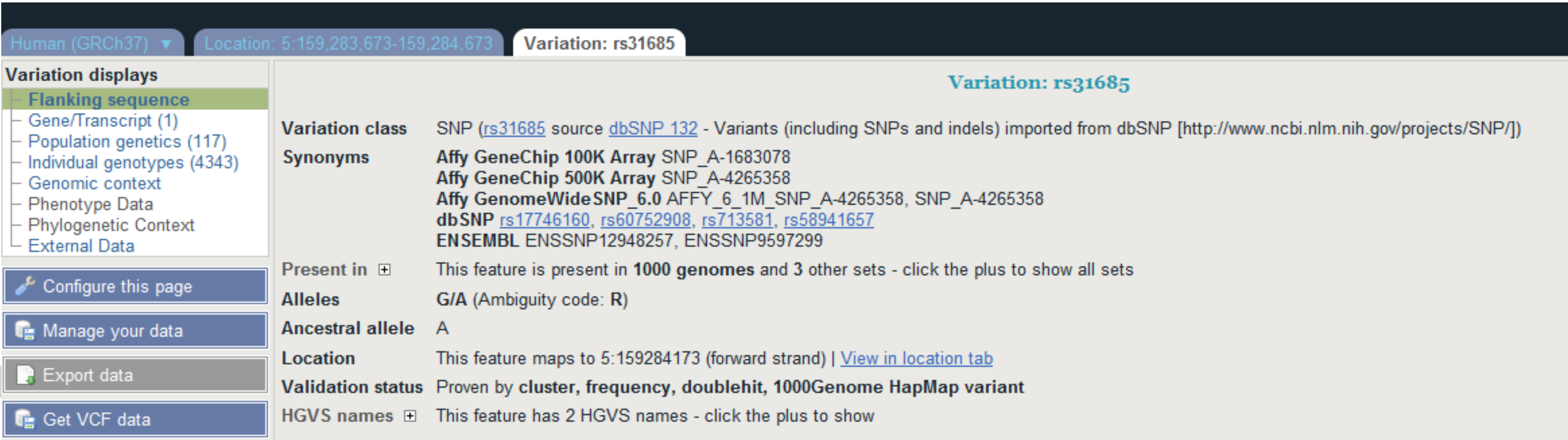
A Deep Catalog of Human Genetic Variation

Search 1000 Genomes

rs31685 [Go]

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

- The Variation tab- left hand links take you to more information

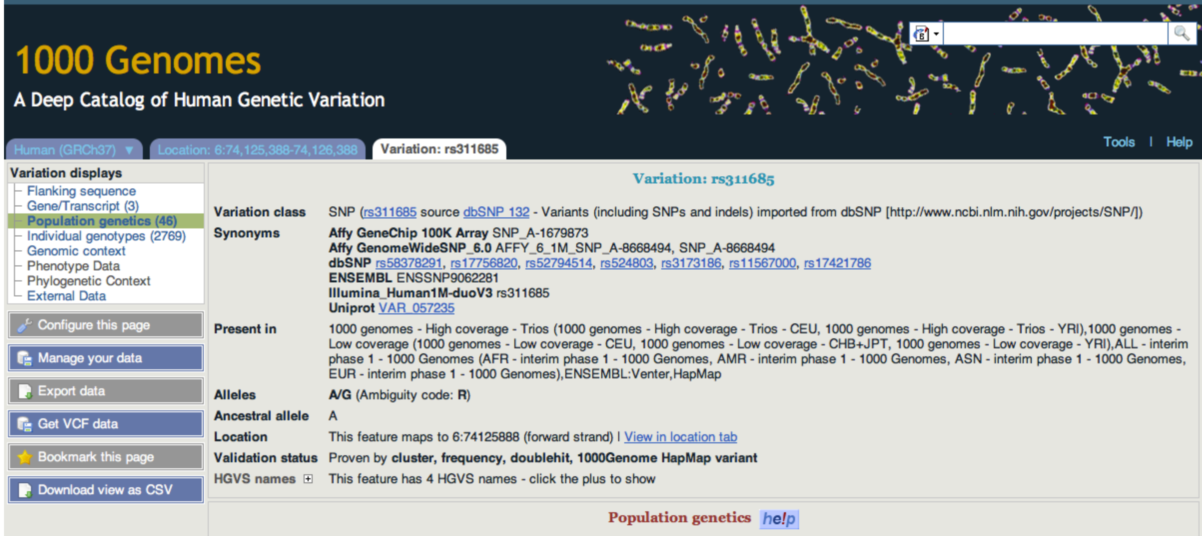


Variation displays
- Flanking sequence
- Gene/Transcript (1)
- Population genetics (117)
- Individual genotypes (4343)
- Genomic context
- Phenotype Data
- Phylogenetic Context
- External Data

Variation: rs31685

Variation class	SNP (rs31685 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-1683078 **Affy GeneChip 500K Array** SNP_A-4265358 **Affy GenomeWideSNP_6.0** AFFY_6_1M_SNP_A-4265358, SNP_A-4265358 **dbSNP** rs17746160, rs60752908, rs713581, rs58941657 **ENSEMBL** ENSSNP12948257, ENSSNP9597299
Present in ⊞	This feature is present in **1000 genomes** and **3** other sets - click the plus to show all sets
Alleles	**G/A** (Ambiguity code: **R**)
Ancestral allele	A
Location	This feature maps to 5:159284173 (forward strand) | View in location tab
Validation status	Proven by **cluster, frequency, doublehit, 1000Genome HapMap variant**
HGVS names ⊞	This feature has 2 HGVS names - click the plus to show

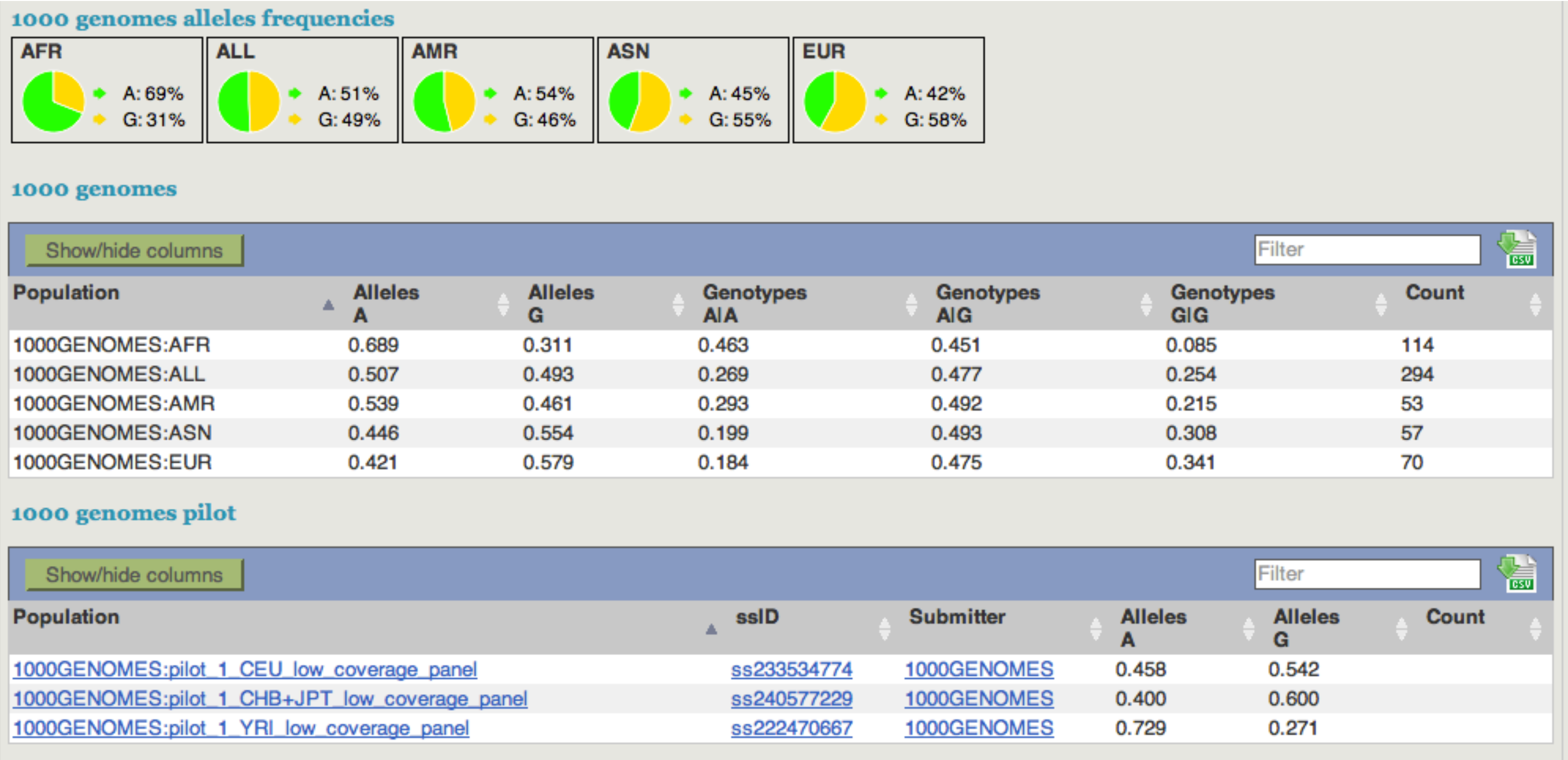
- Population



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

Population genetics help



1000 genomes alleles frequencies

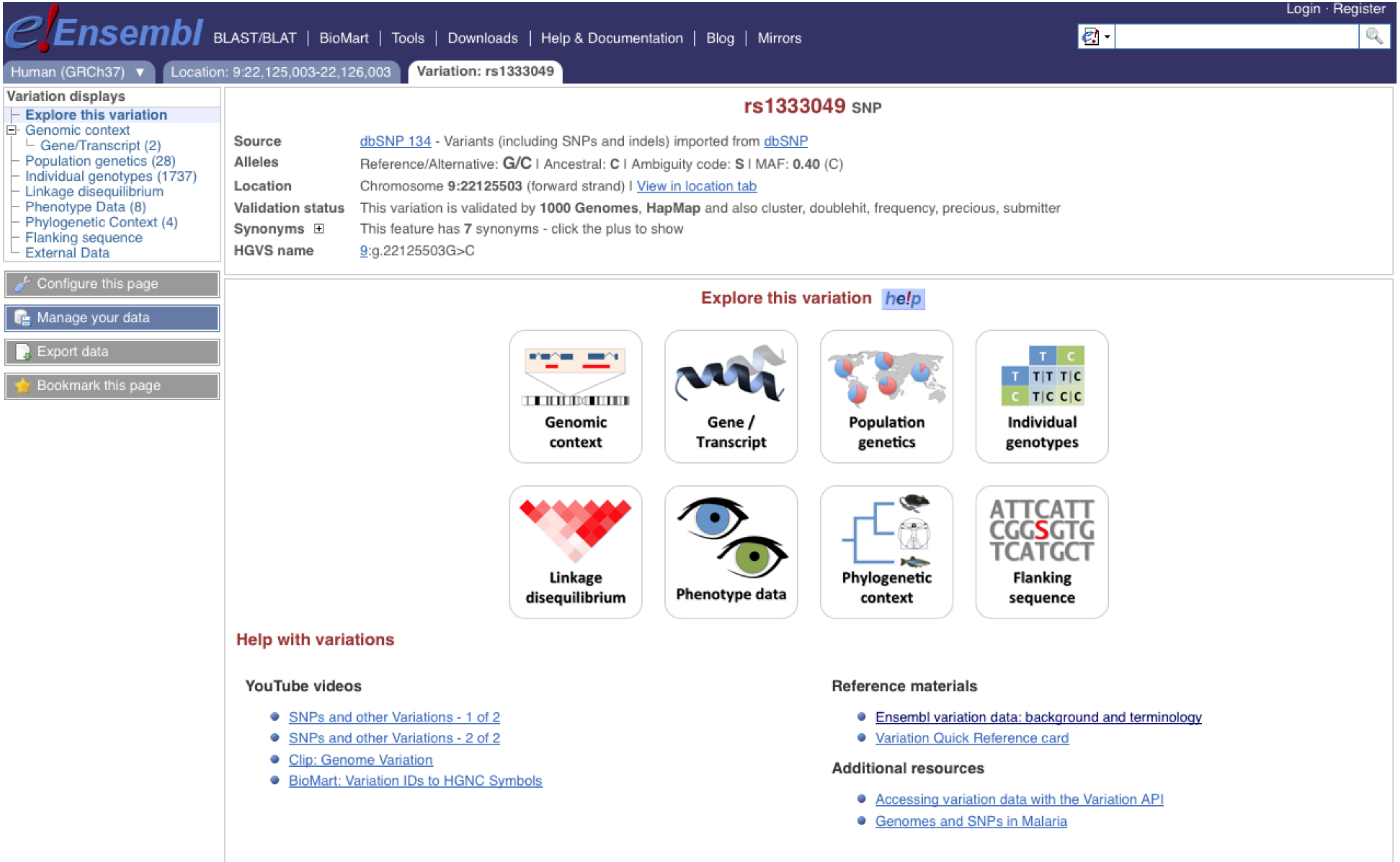
1000 genomes

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

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	

The Browser: Coming Soon



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

Help with variations

YouTube videos

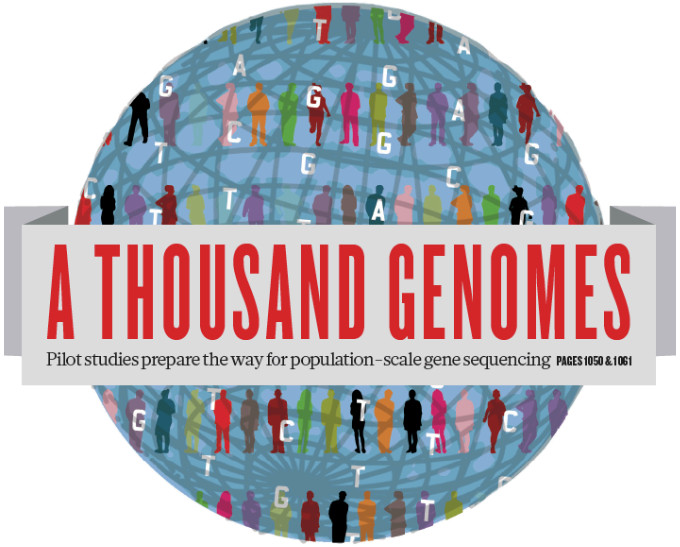
- SNPs and other Variations - 1 of 2
- SNPs and other Variations - 2 of 2
- Clip: Genome Variation
- BioMart: Variation IDs to HGNC Symbols

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
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project:

Exercise 2: Finding Variation Using the Browser

Exercise: Finding Variation Using the Browser

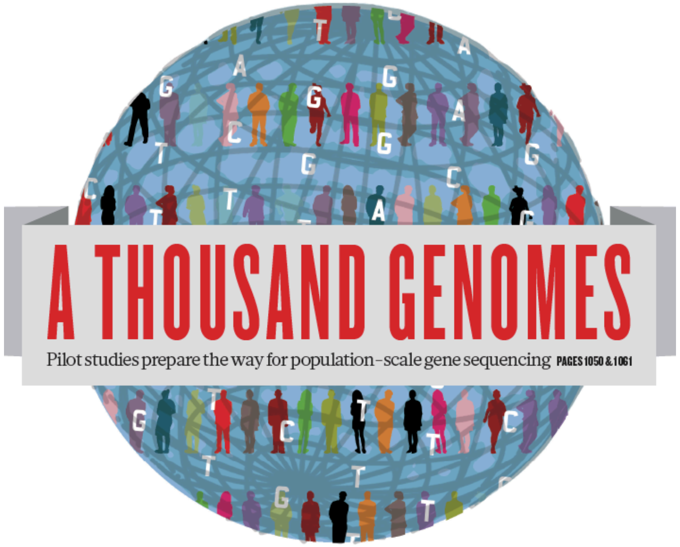
- Find the variant rs45562238

  http://browser.1000genomes.org

- In which 1000 Genomes Populations was it detected?

- What are its allele frequencies?

- In which gene is the variant found?

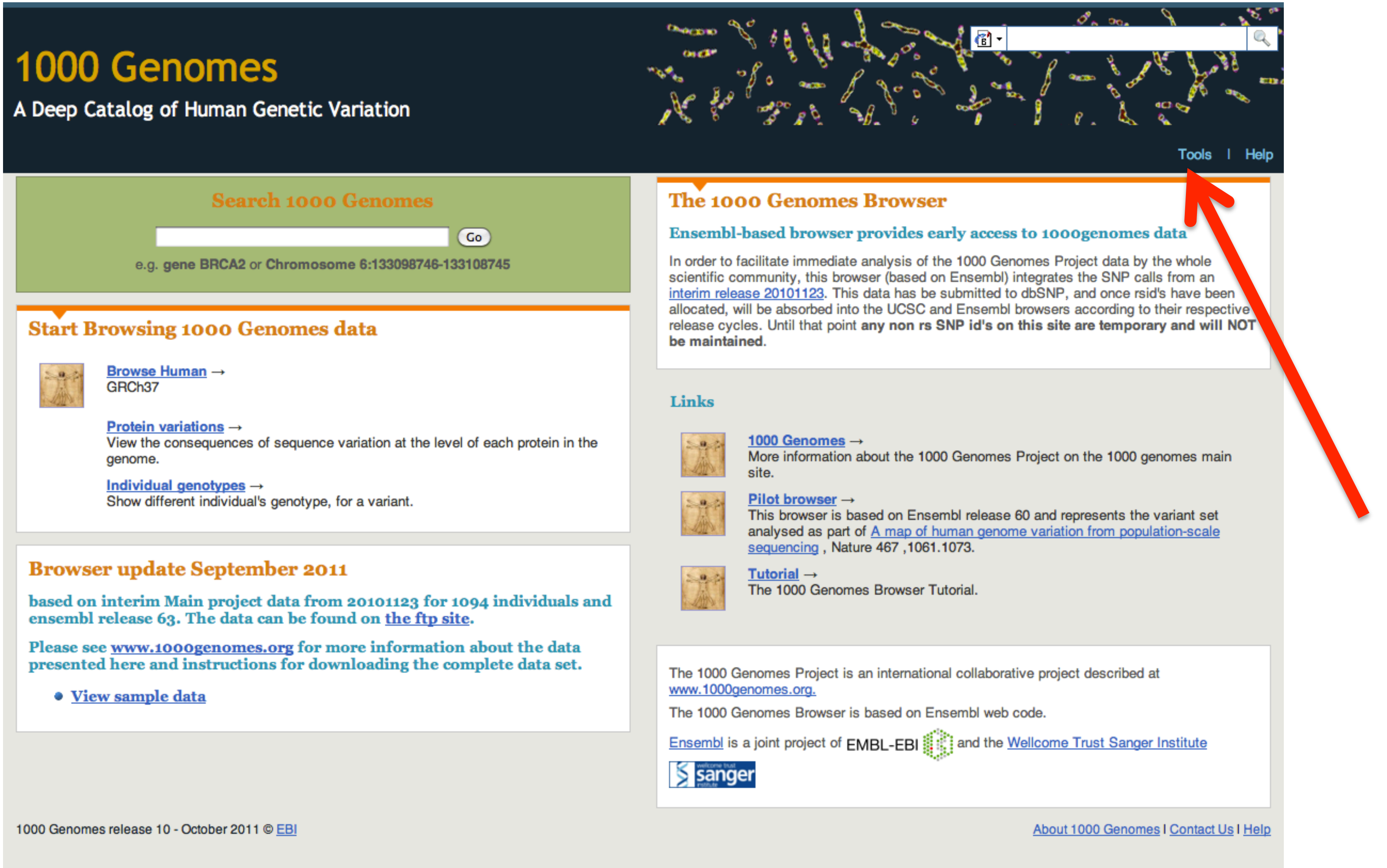

A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project:

The 1000 Genomes Tools

The 1000 Genomes Tools

- Data Slicer
- Variant Effect Predictor
- Variation Pattern Finder
- VCF to PED
- API and Database access



1000 Genomes

A Deep Catalog of Human Genetic Variation

Search 1000 Genomes

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

Start Browsing 1000 Genomes data

Browse Human →
GRCh37

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

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

Browser update September 2011

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

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

- View sample data

The 1000 Genomes Browser

Ensembl-based browser provides early access to 1000genomes data

In order to facilitate immediate analysis of the 1000 Genomes Project data by the whole scientific community, this browser (based on Ensembl) integrates the SNP calls from an interim release 20101123. This data has 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.

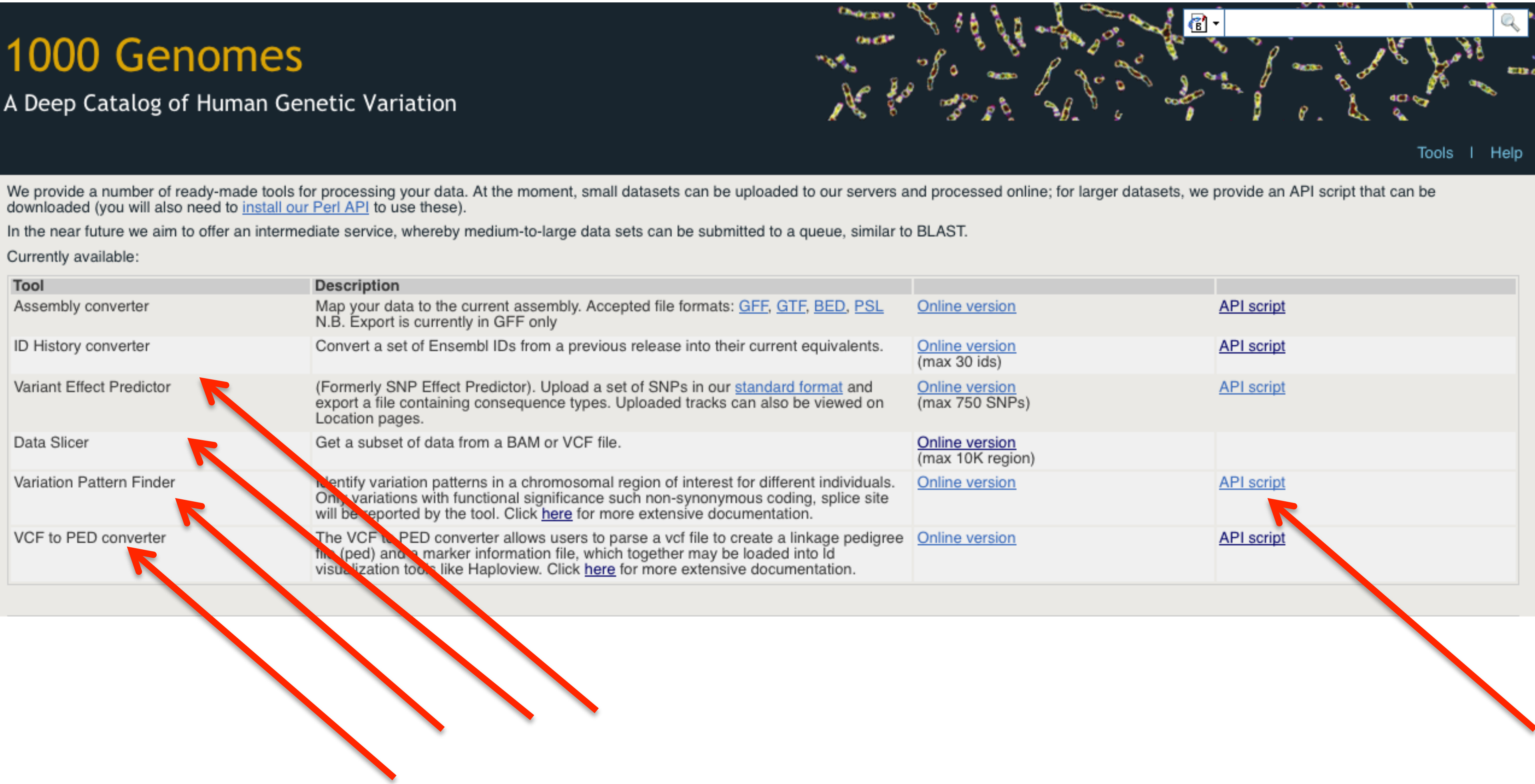
Ensembl is a joint project of EMBL-EBI and the Wellcome Trust Sanger Institute

1000 Genomes release 10 - October 2011 © EBI

About 1000 Genomes | Contact Us | Help

http://browser.1000genomes.org

Tools page



1000 Genomes

A Deep Catalog of Human Genetic Variation

We provide a number of ready-made tools for processing your data. At the moment, small datasets can be uploaded to our servers and processed online; for larger datasets, we provide an API script that can be downloaded (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. 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 ld visualization tools like Haploview. Click here for more extensive documentation.	Online version	API script

Variant Effect Predictor

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

Variant Effect Predictor

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

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

Data Slicing

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

Variation Pattern Finder

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

Variation Pattern Finder

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

Variation Pattern Finder Output

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

VCF to PED

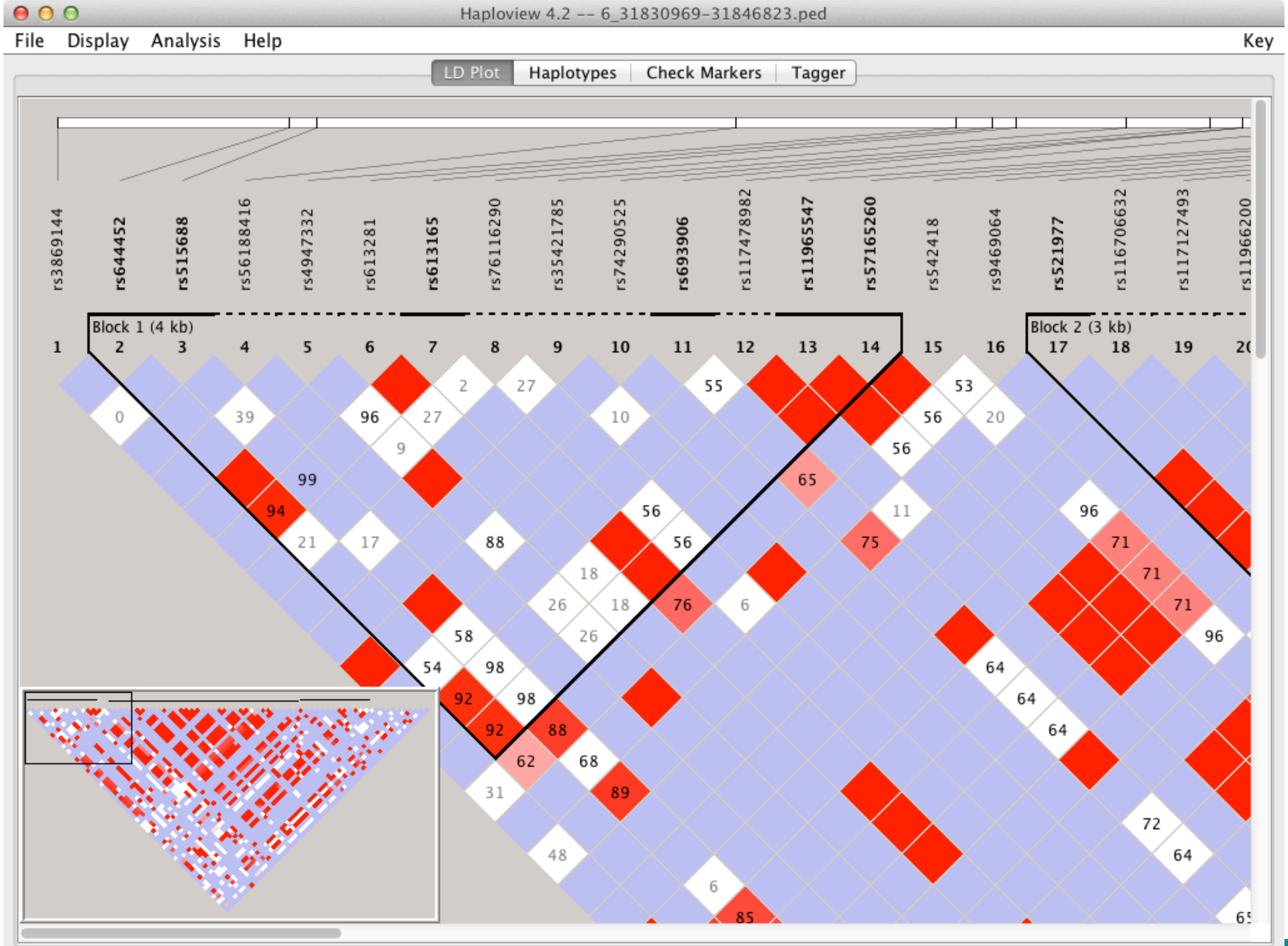
- LD Visualization tools like Haploview require PED files
- VCF to PED converts VCF to PED
- Will a file divide by individual or population
- http://browser.1000genomes.org/Homo_sapiens/UserData/Haploview

VCF to PED

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

Haploview

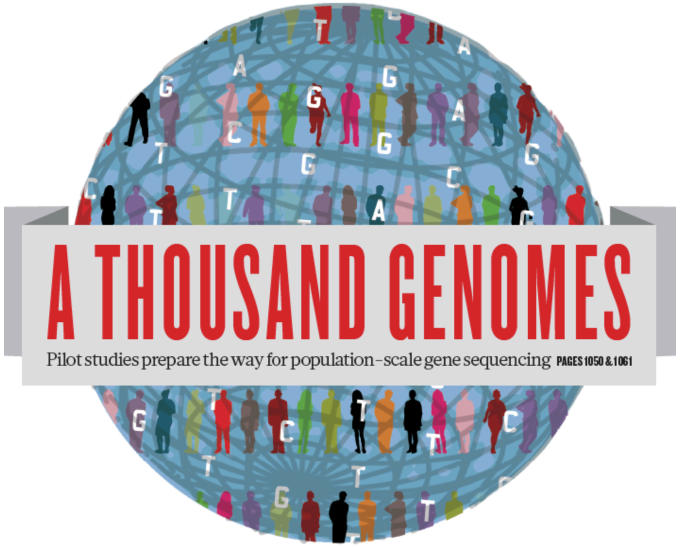
- haploview



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

Access to backend Ensembl databases

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


A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project:

Exercise 3: Using 1000 Genomes Tools

Exercise 3 Using 1000 Genomes Tools

- Get a 6:31831625-31834704 slice of our chr6 release vcf file using tabix.
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_release_v2.20101123.snps_indels_svs.vcf.gz
  - Use vcf subset to get just the genotype info for HG01375
  - Direct this into a file
- Use the retrieved vcf file with the variant effect predictor script
- Which variants have deleterious sift/polyphen consequences?
- Use the variant pattern finder to look at the pattern of variation in the same region using the remote chr6 vcf file
- Produce ped and locus information files using the vcf to ped tool for the CEU population

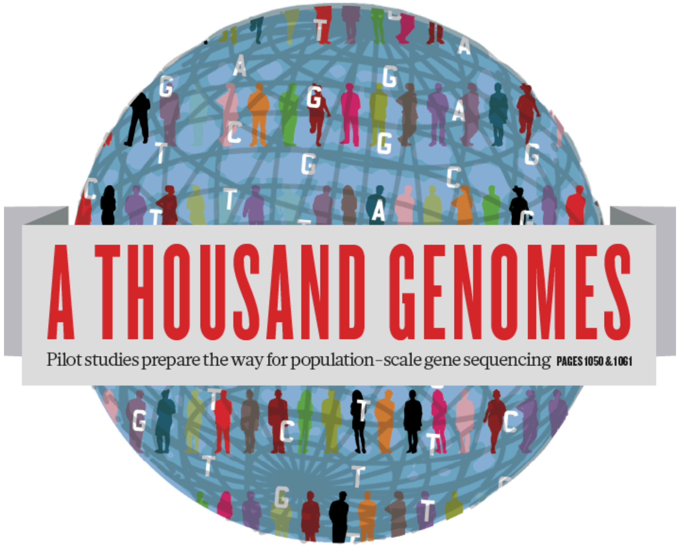
Answers

-
  ```
  tabix -h ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_release_v2.20101123.snps_indels_svs.vcf.gz 6:31831625-31834704 > 6_381831625_3184704.vcf
  ```

-
  ```
  perl variant_effect_predictor.pl -input 6_381831625_3184704.vcf -sift p -polyphen p -check_existing
  ```
  - 6_31832657_C/T, rs141954433 and rs149841290 all have deleterious sift and polyphen results

-
  ```
  perl variant_pattern_finder.pl -vcf ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_integrated_calls.20101123.snps_indels_svs.genotypes.vcf.gz  -sample_panel_file ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel -region 6:31831625-31834704  -expand
  ```

-
  ```
  perl vcf_to_ped_convert.pl -vcf ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_integrated_calls.20101123.snps_indels_svs.genotypes.vcf.gz  -sample_panel_file ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel -region 6:31830969-31846823  -population CEU
  ```


A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing PAGES 1050 & 1061

The 1000 Genomes Project:

Finding out about New Data and using Data on Campus

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

1000 Genomes Data on Campus

- @EBI
  - Email resequencing-informatics@ebi.ac.uk
- @Sanger
  - Email Jim Stalker jws@sanger.ac.uk

Thanks

- The 1000 Genomes Project Consortium
- Paul Flicek, Laura Clarke
- Richard Smith, Holly Zheng Bradley and Ian Streeter
- Giulietta Spudich and Bert Overduin