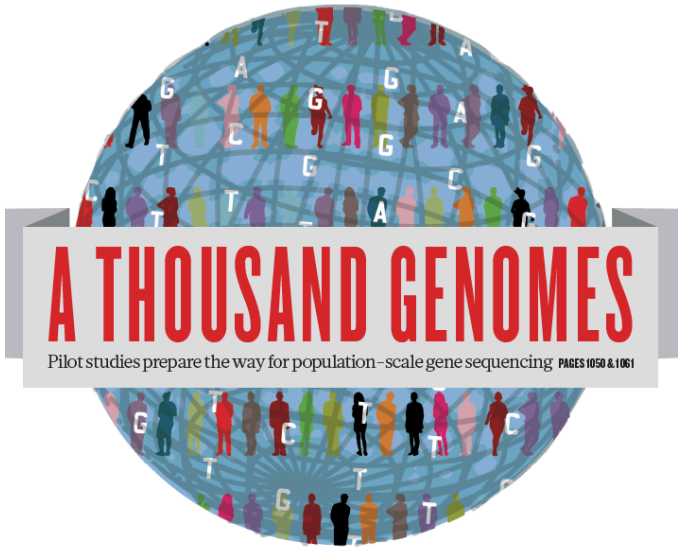


The 1000 Genomes Project

Laura Clarke

EMBL-EBI

- “After the first complete sequence of a human genome is obtained, the next challenge will be to discover and understand the function and variation of genes and, ultimately, to understand how such qualities affect health and disease. A key to this undertaking will be the availability of methods for efficient and accurate identification of genetic variation ..."

- Sydney Brenner et al. (June 2000) Nature Biotechnology

EMBL-EBI

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
- **Date Formats**: In 1000 genomes file/directory names dates are mostly represented as YYYYMMDD

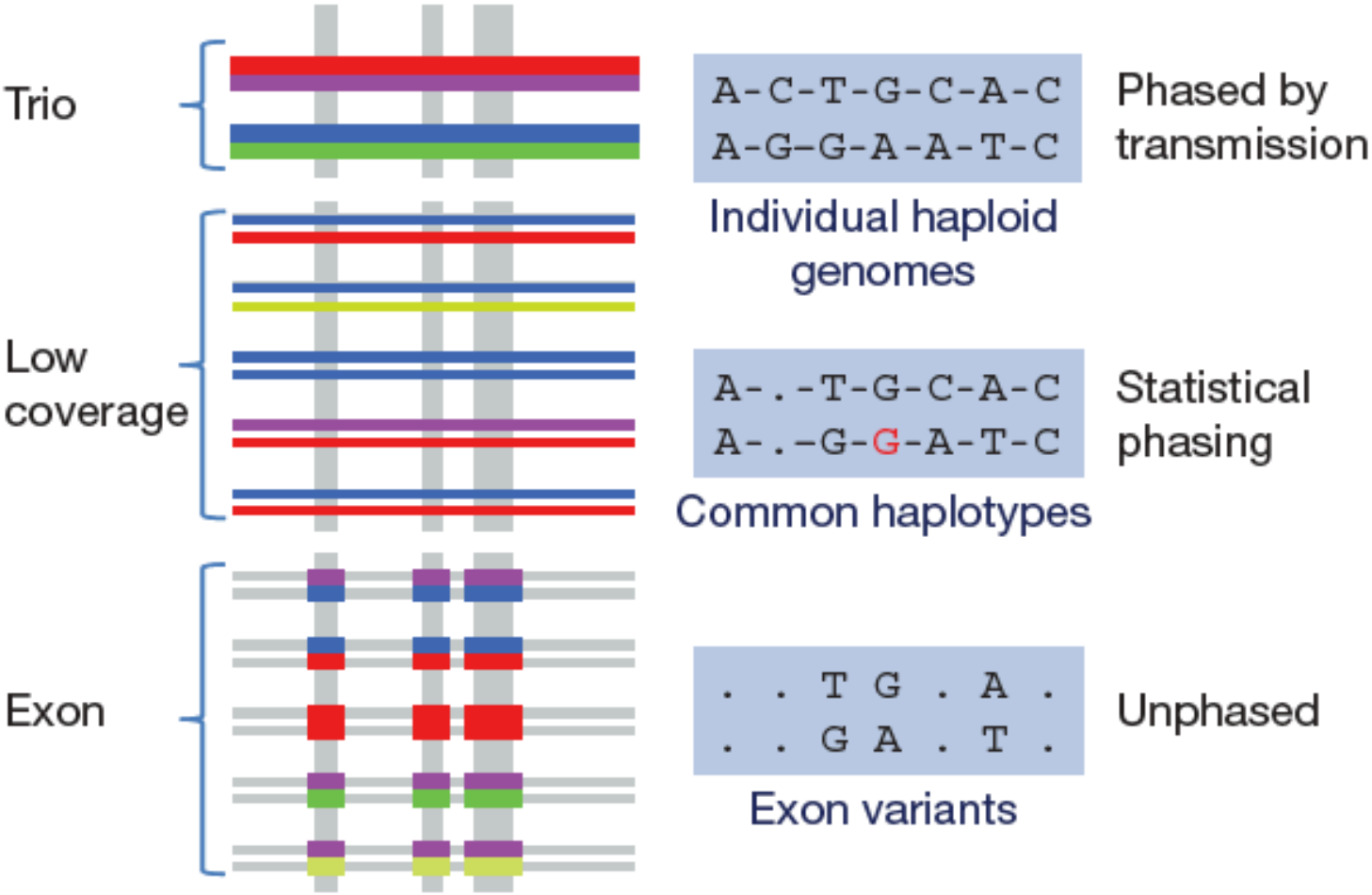
EMBL-EBI

The 1000 Genomes Project: Overview

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

EMBL-EBI

3 pilot coverage strategies

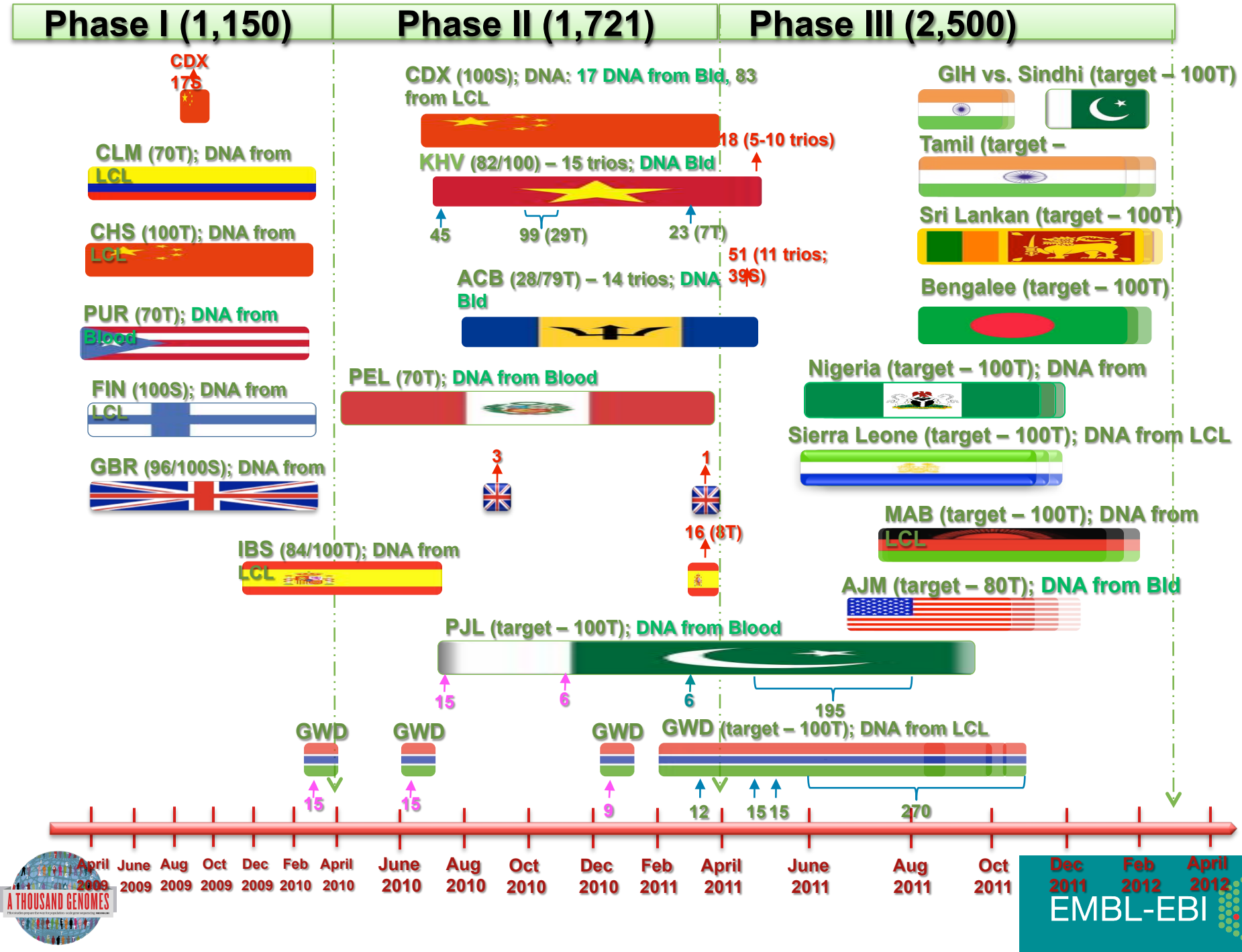


A THOUSAND GENOMES
EMBL-EBI

Main Project Design

- Based on the result of the pilot project, we decided to collect data on more than 2,500 samples from 5 continental groupings
  - Whole-genome low coverage data (>4x)
  - Full exome data at deep coverage (>20x)
  - 500 deep coverage genomes to be sequenced
  - High density genotyping at subsets of sites using both Illumina Omni and Affymetrix Axiom
- Phase 1 Release Integrated Variant Release has been made.

EMBL-EBI


Phase I (1,150)
Phase II (1,721)
Phase III (2,500)
CDX 176
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
15
GWD
15
GWD
9
GWD (target – 100T); DNA from LCL
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

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, 1.5millon indels and 14K deletions
  - 1092 individuals
  - 14 populations
  - Low coverage and exome next generation sequencing
- 1000 Genomes Phase 2
  - Started in 2011
  - 1721 individuals
  - 19 Populations
  - Low coverage and exome next generation sequencing

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

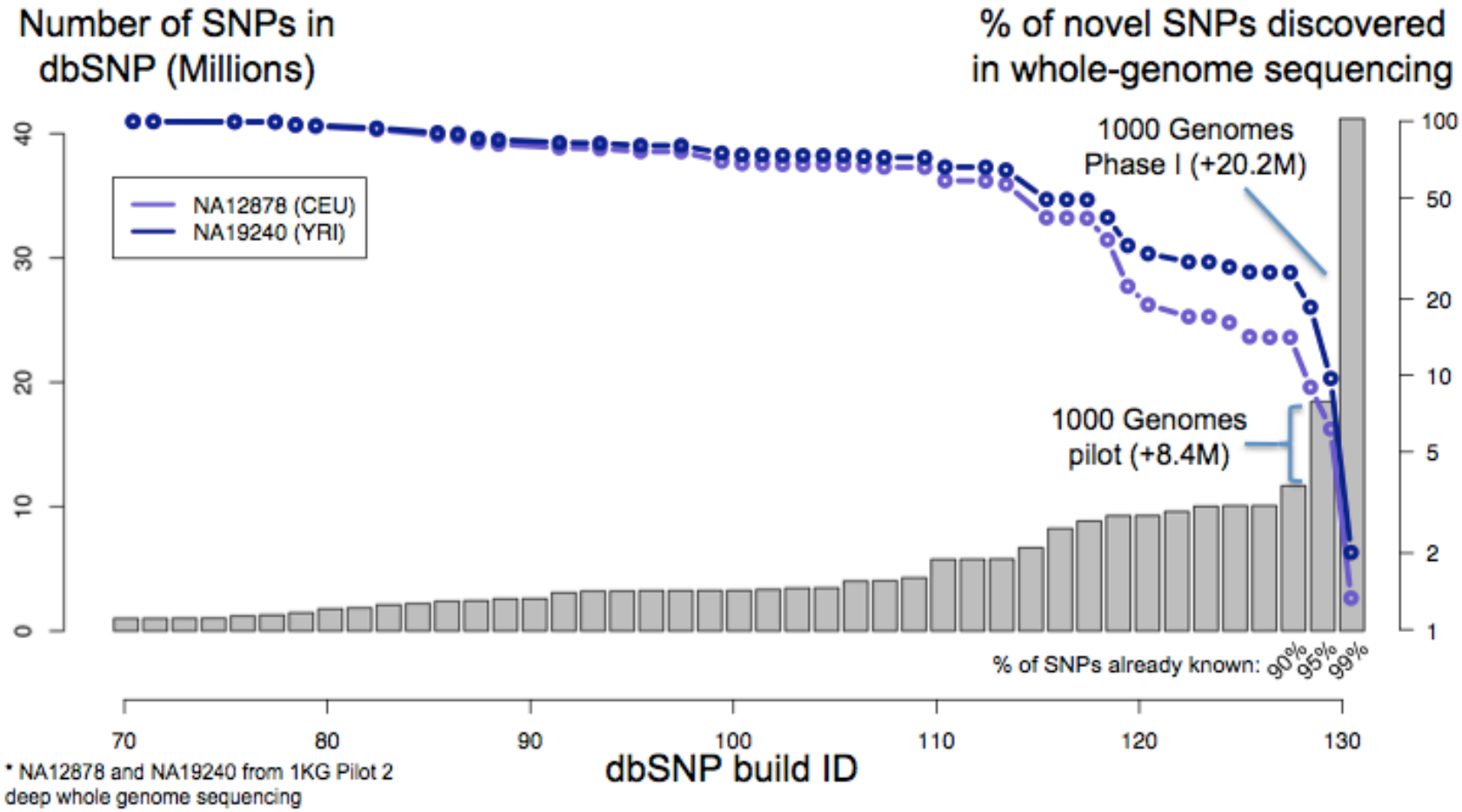
EMBL-EBI

Sequencing Data Evolution

- 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

1000 Genomes discovered 29M new SNPs; now ~99% of variation in each person is already known



Mark de Pristo

EMBL-EBI

1000 Genomes Project: Present & Future

- First Phase 2 sequence release 14th November 2011
- First Phase 2 alignment release 12th March 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

EMBL-EBI

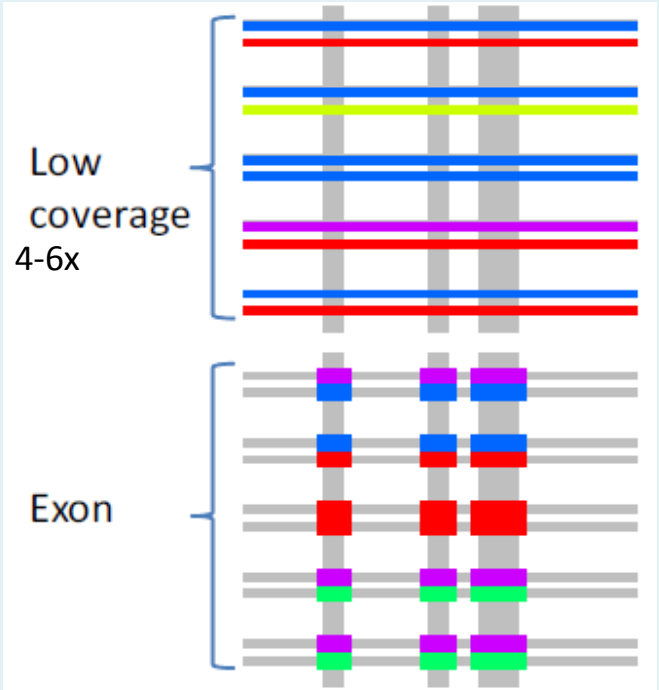
Pipelines for data processing and variant calling

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

EMBL-EBI

Approach to the Analysis

Data



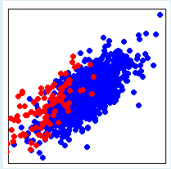
Multiple centres
Multiple technologies

Variant discovery

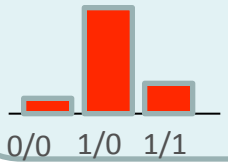
- SNPs
- Short indels
- Structural variants

Multiple Methodologies

Variant integration

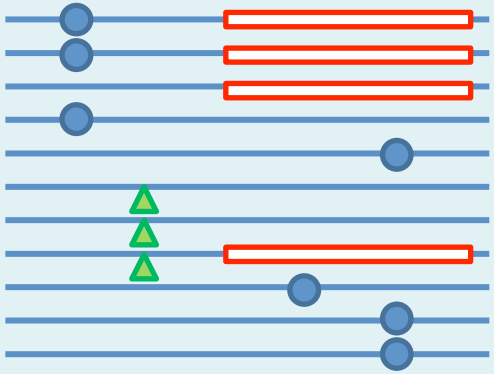


Adjudication



Likelihoods

Haplotype integration



Single release VCF

A THOUSAND GENOMES

EMBL-EBI

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

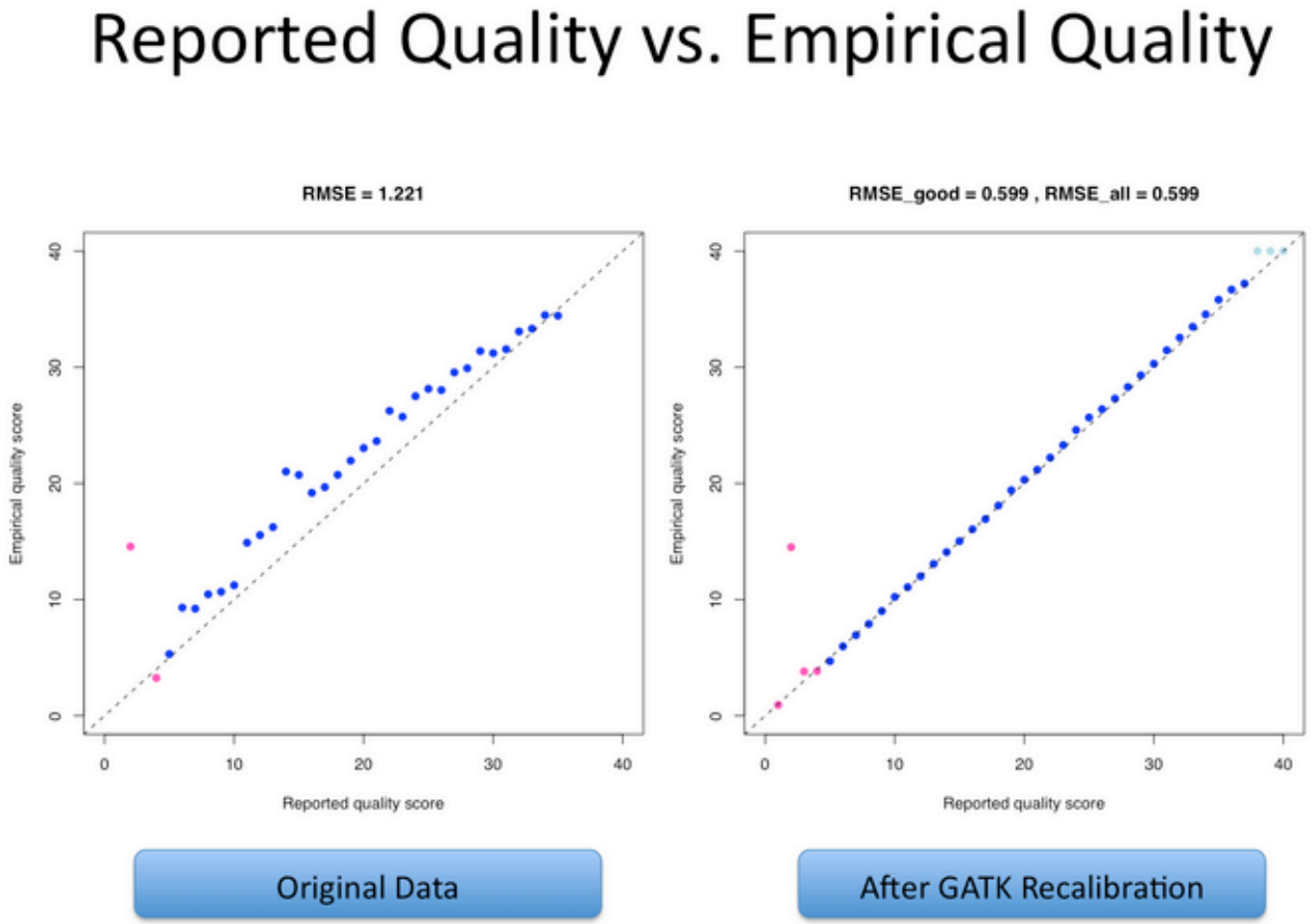
EMBL-EBI

Base Quality Score Recalibration

- 1000 Genomes Sequence Data is sourced from many different machines across many different institutes
- Each machine may assign Base Quality Values differently
- Base Quality Score Recalibration tests empirical error rates
  - Run alignment
  - Compare mismatches to know variation
- Base Qualities adjusted on basis of empirical measurements

EMBL-EBI

Base Quality Score Recalibration



Methods for integrated genotypes

Components		SNPs	INDELs	SVs
Low-Pass Genomes	**Call Sets**	BC, BCM, BI NCBI, SI, UM	BC, BI, DI OX, SI	BI, EBI, EMBL UW, Yale
	Consensus	VQSR	VQSR	GenomeSTRiP
Deep Exomes	**Call Sets**	BC, BCM, BI UM, WCMC	N/A	N/A
	Consensus	SVM	N/A	N/A
Likelihood		BBMM	GATK	GenomeSTRiP
Site Models		Variants are linearly ordered as point mutations		
Haplotyper		MaCH/Thunder with BEAGLE's initial haplotypes		

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 make site discovery more intergrated

Variant Quality Score Recalibration

- Multiple Different Variant Callers are used as part of the 1000 Genomes
- Variant Quality Score Recalibration used to define high quality variants from large input set
- Variants as points in a point cloud can be modeled using a Gaussian mixture model
- Model compared to various statistical models to define best set of variants

EMBL-EBI

VQSR consensus out performs previous merging strategy

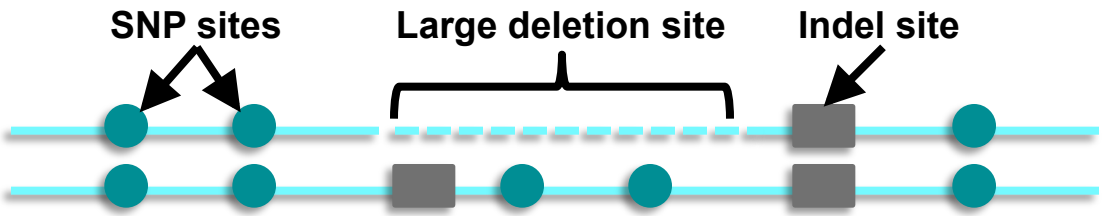
Called In	Total # variants	dbSNP % (129)	# novels	Novel ti/tv	Omni poly sensitivity	Omni mono false discovery
Union	46.26M	19.39%	37.29M	1.998	98.94% 2.09M / 2.12M	16.31% 9,739 / 59,721
2 of 6	39.11M	22.24%	30.41M	2.153	98.55% 2.09M / 2.12M	11.23% 6,707 / 59,721
3 of 6	35.69M	23.62%	27.26M	2.219	98.09% 2.08M / 2.12M	3.66% 2,184 / 59,721
4 of 6	32.55M	24.82%	24.48M	2.263	97.39% 2.06M / 2.12M	1.82% 1,085 / 59,721
5 of 6	28.45M	26.72%	20.85M	2.286	95.93% 2.03M / 2.12M	1.06% 634 / 59,721
Intersection	24.02M	27.57%	17.40M	2.317	89.23% 1.89M / 2.12M	0.76% 457 / 59,721
VQSR Project Consensus	**38.88M**	**21.92%**	**30.36M**	**2.154**	**98.41%** **2.08M / 2.12M**	**2.11%** **1,261 / 59,721**

EMBL-EBI

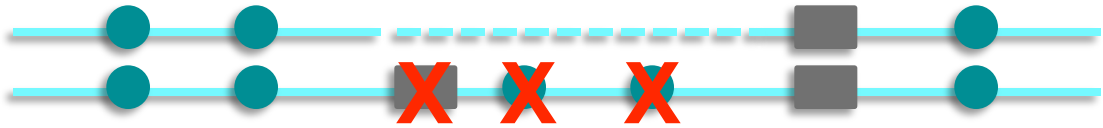
Methods for integrated genotypes

Components		SNPs	INDELs	SVs
Low-Pass Genomes	**Call Sets**	BC, BCM, BI NCBI, SI, UM	BC, BI, DI OX, SI	BI, EBI, EMBL UW, Yale
	Consensus	VQSR	VQSR	GenomeSTRiP
Deep Exomes	**Call Sets**	BC, BCM, BI UM, WCMC	N/A	N/A
	Consensus	SVM	N/A	N/A
Likelihood		BBMM	GATK	GenomeSTRiP
Site Models		Variants are linearly ordered as point mutations		
Haplotyper		MaCH/Thunder with BEAGLE's initial haplotypes		

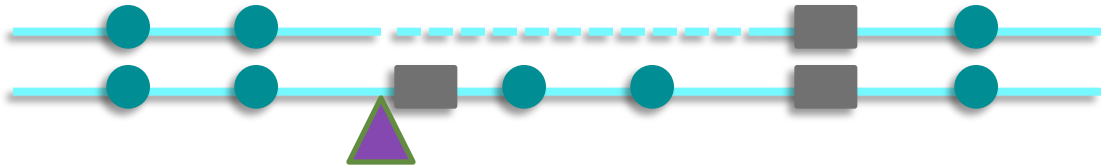
Strategies for integrating deletions with other types of variation



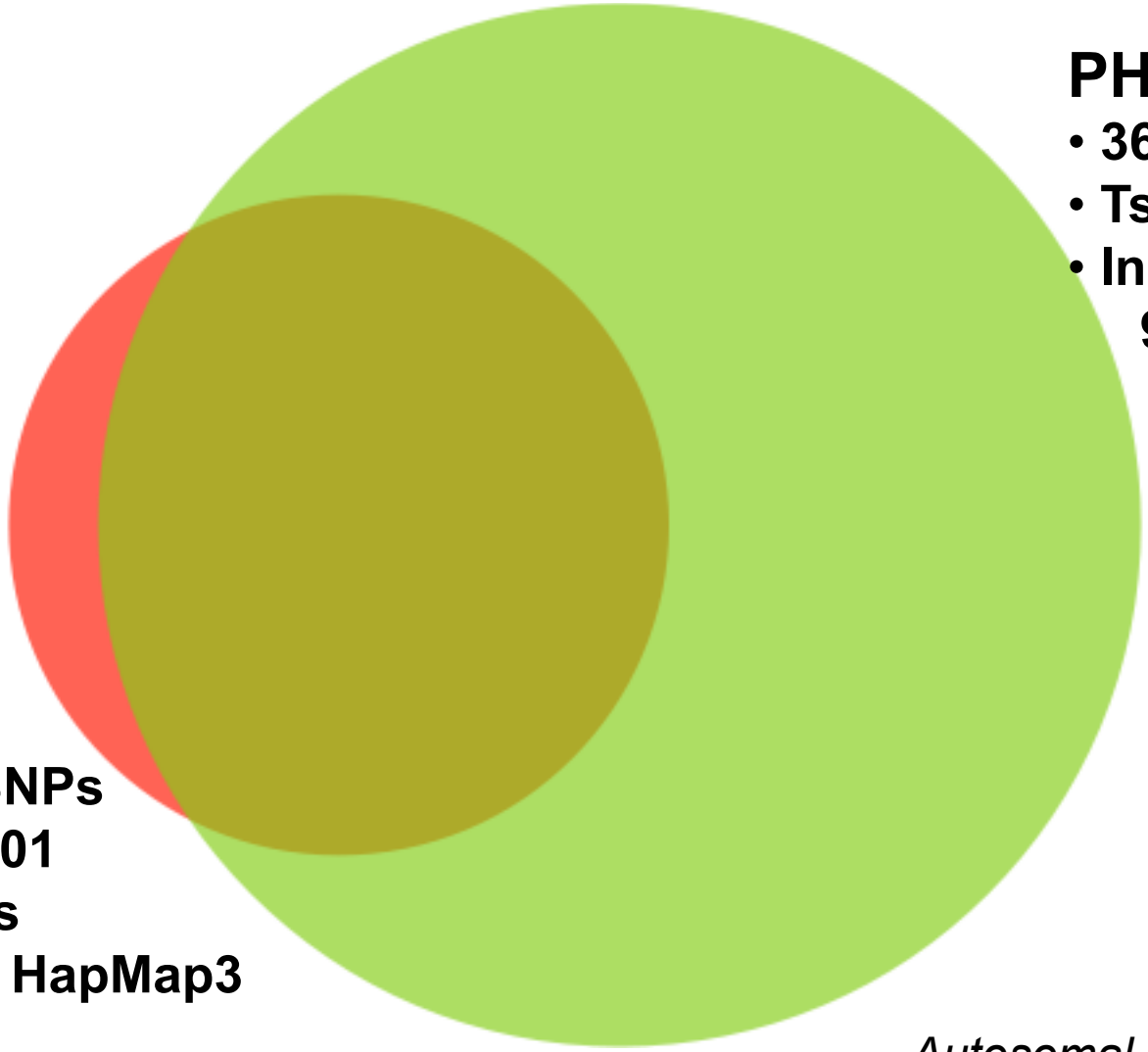
Previous Approach
Remove SNPs under SVs for imputation
(1000G pilot, Handsaker et al., 2010)



Current Approach
Treat SVs as point events
(1000 Genomes phase 1)



From PILOT to PHASE1



PHASE1
- **36.8M SNPs**
- **Ts/Tv 2.17**
- **Includes 98.9% HapMap3**

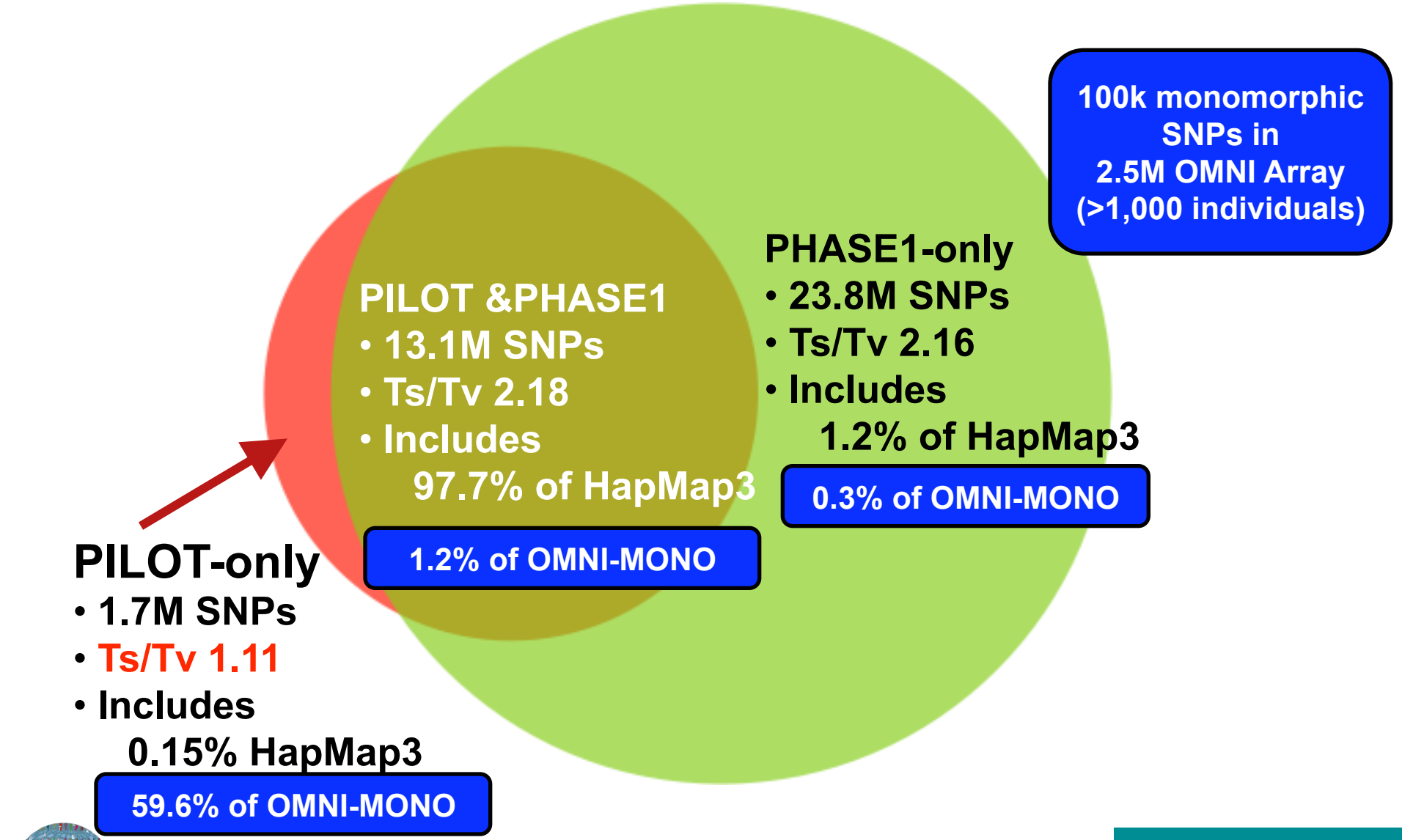
PILOT
- **14.8M SNPs**
- **Ts/Tv 2.01**
- **Includes 97.8% HapMap3**

Autosomal chromosomes only

A THOUSAND GENOMES

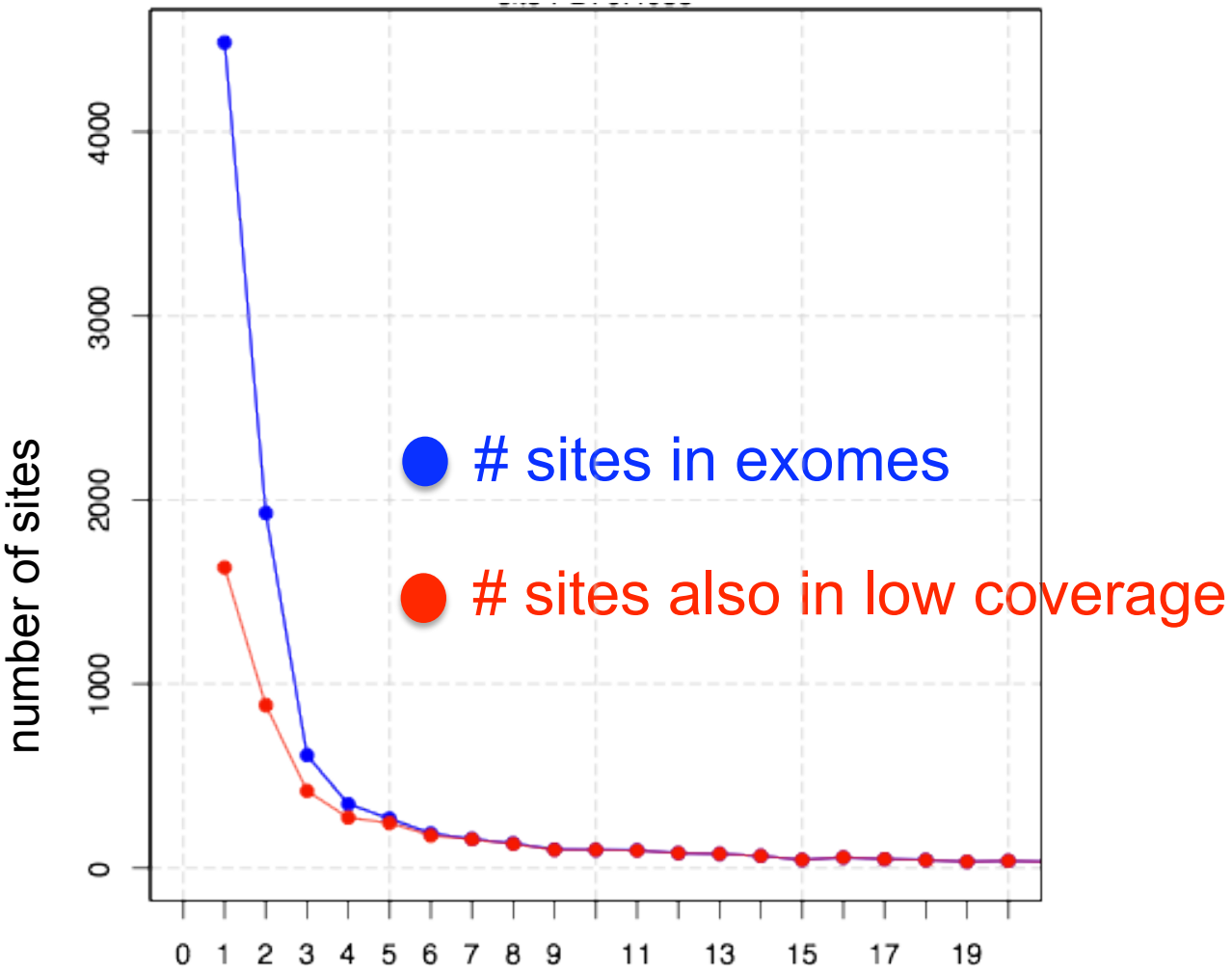
EMBL-EBI

From PILOT to PHASE1 : Improved SNP calls



OMNI-MONO information was not used in making phase1 variant calls

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



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

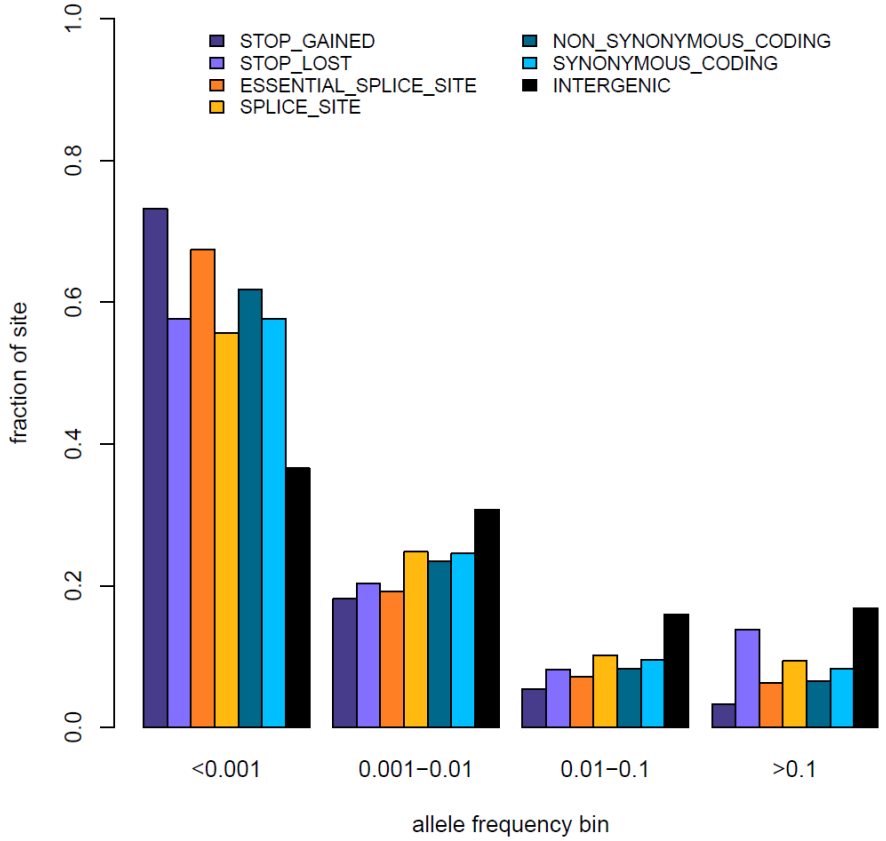
Erik Garrison

A THOUSAND GENOMES

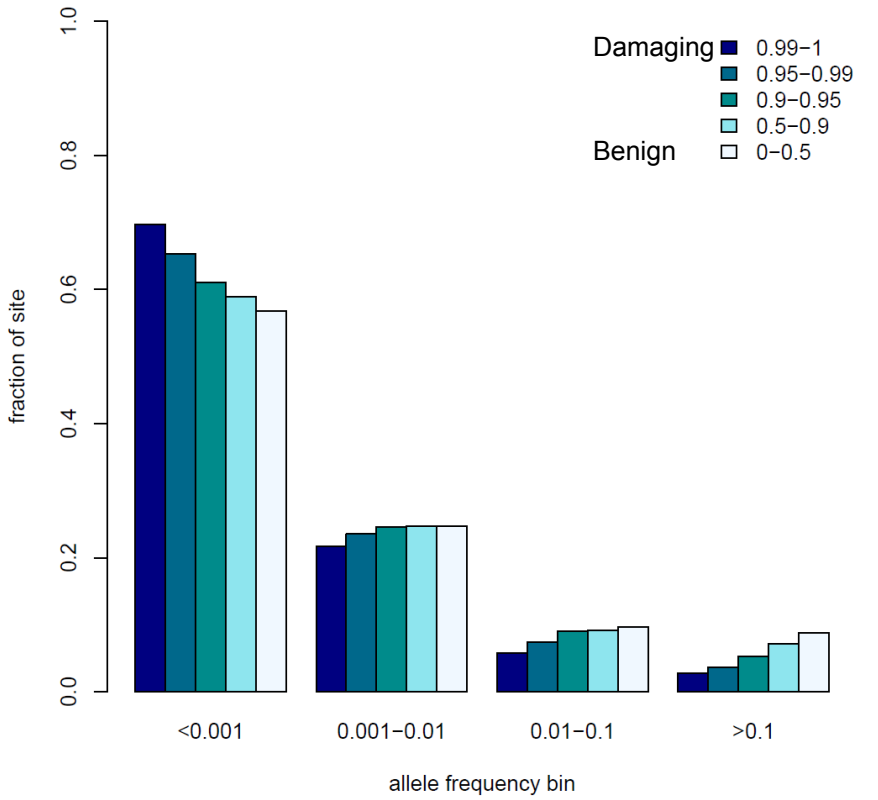
EMBL-EBI

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

Functional category



Non-synonymous: Condel score



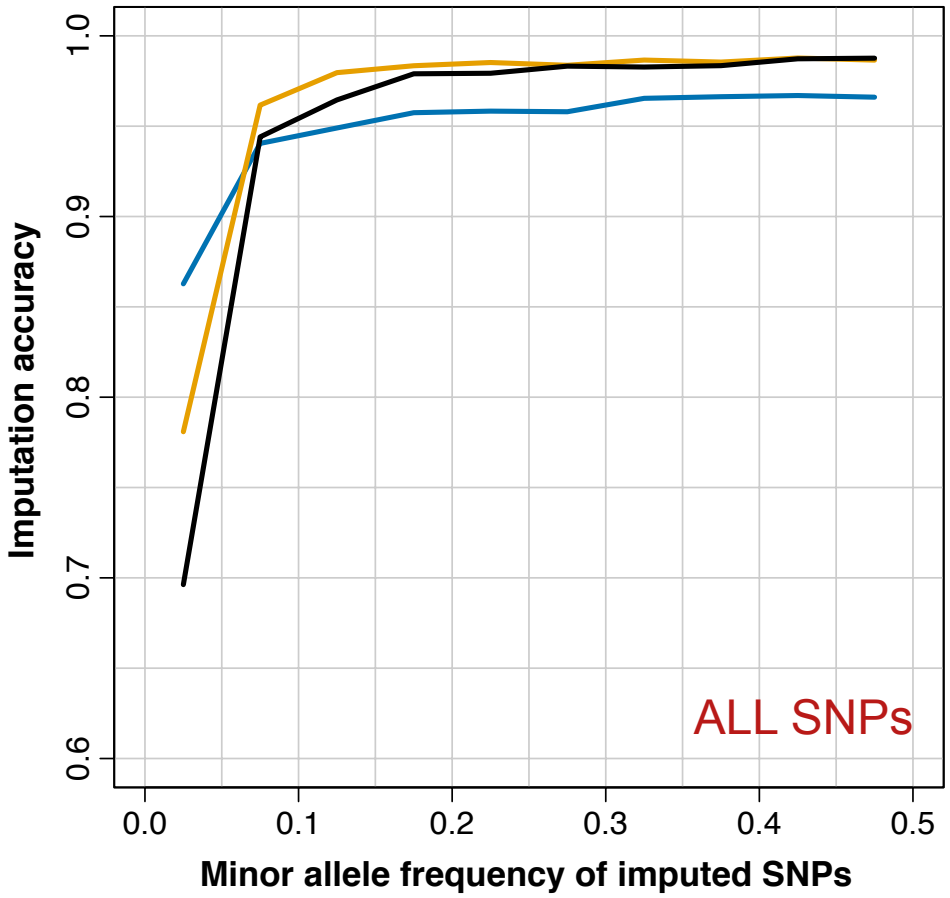
Presentation on using the data for GWAS by Brian Howie

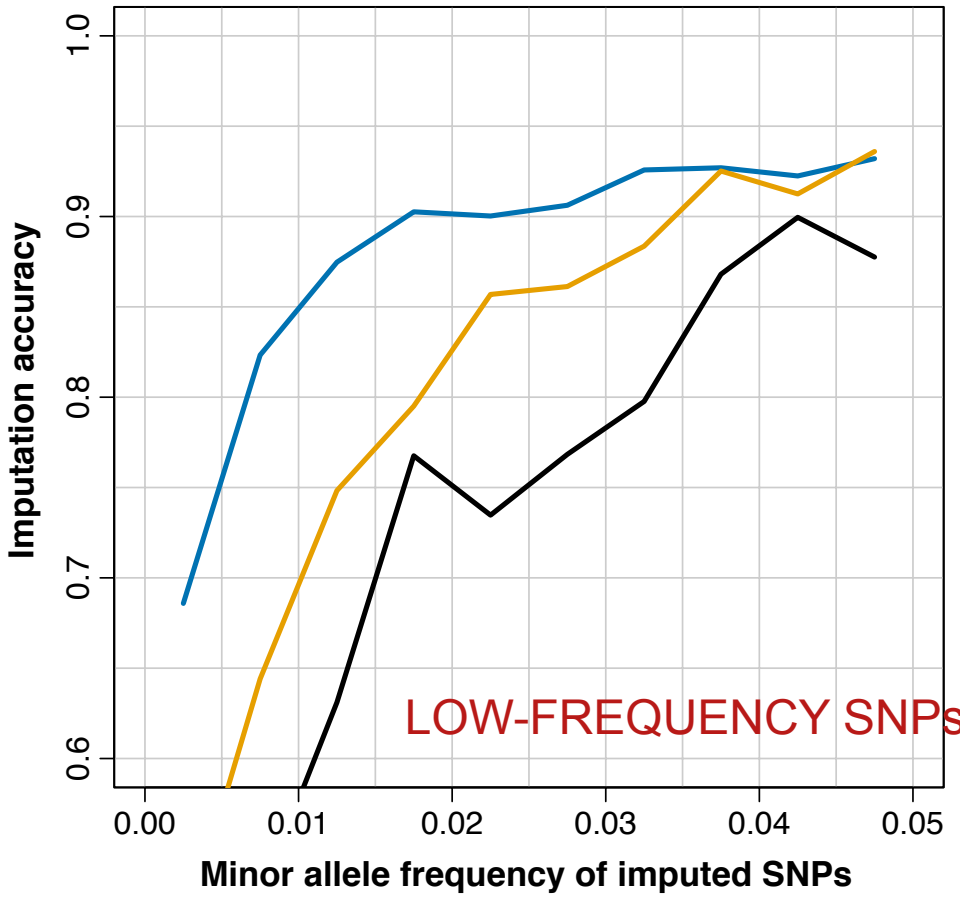
Enza Colonna, Yuan Chen, Yali Xue

A THOUSAND GENOMES

EMBL-EBI

1,000 Genomes haplotypes are highly accurate





European ancestry

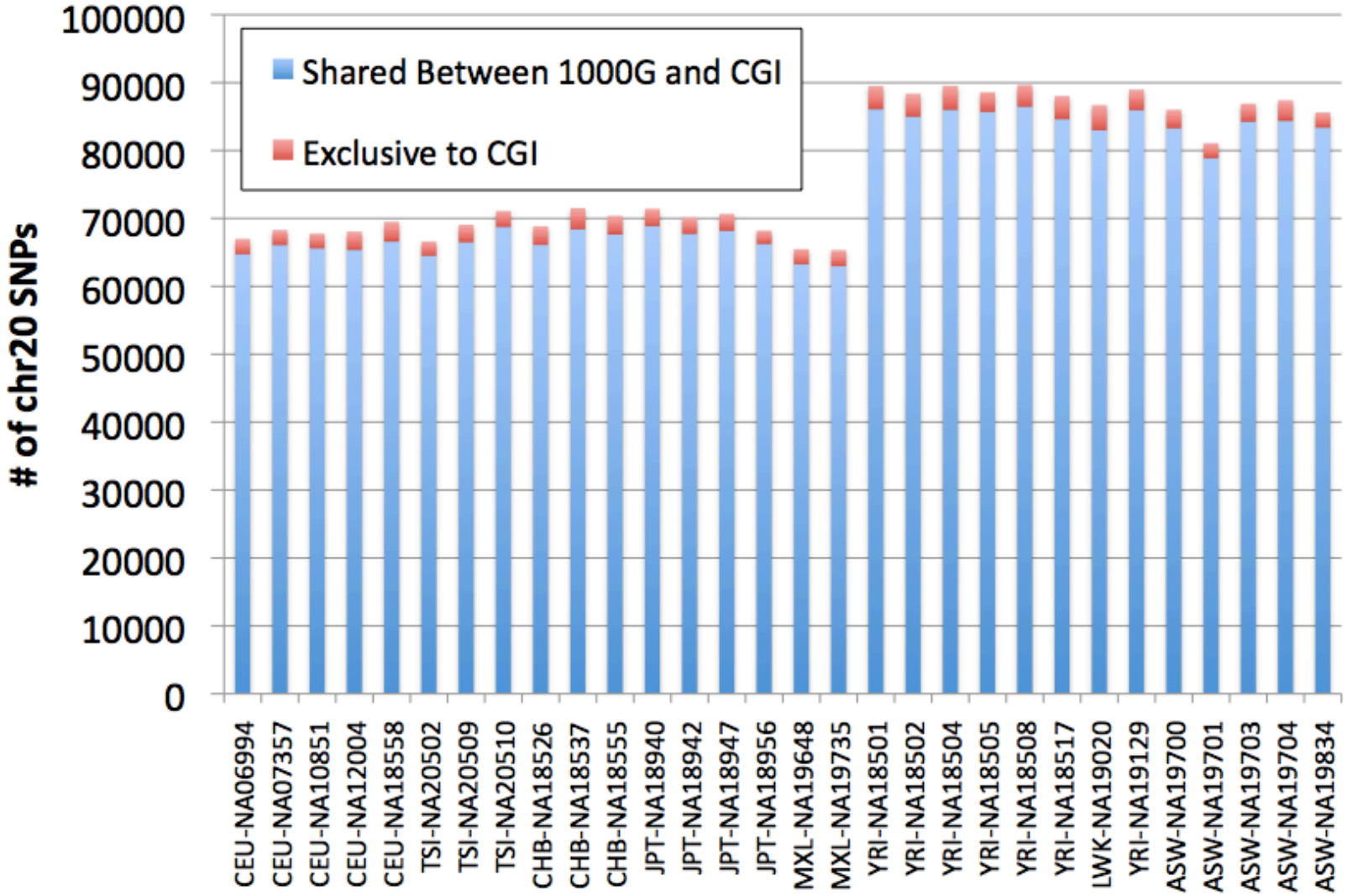
African ancestry

Admixed (Americas)

A THOUSAND GENOMES

EMBL-EBI

>96% SNPs are detected compared to deep genomes



A THOUSAND GENOMES

EMBL-EBI

Data Availability and the FTP site

EMBL-EBI

File Formats

- Sequence in Fastq
- Alignments in SAM/BAM
- Variant Calls in VCF
- Other data
  - ped
  - gff/gtf
  - bed

EMBL-EBI

More Information About BAM Files

- http://samtools.sourceforge.net/
- samtools-help@lists.sourceforge.net

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

ABSTRACT

Summary: The Sequence Alignment/Map (SAM) format is a generic alignment format for storing read alignments against reference sequences, supporting short and long reads (up to 128 Mbp) produced by different sequencing platforms. It is flexible in style,

2 METHODS

2.1 The SAM format

2.1.1 Overview of the SAM format The SAM format consists of one header section and one alignment section. The lines in the header section start with character '@', and lines in the alignment section do not. All lines

More Information About VCF Files

http://vcftools.sourceforge.net/
vcftools-help@lists.sourceforge.net

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

ABSTRACT

Summary: Tabix is the first generic tool that indexes position sorted files in TAB-delimited formats such as GFF, BED, PSL, SAM and SQL export, and quickly retrieves features overlapping specified regions. Tabix features include few seek function calls per query, data

2 METHODS

Tabix indexing is a generalization of BAM indexing for generic TAB-delimited files. It inherits all the advantages of BAM indexing, including data compression and efficient random access in terms of few seek function calls per query.

A THOUSAND GENOMES

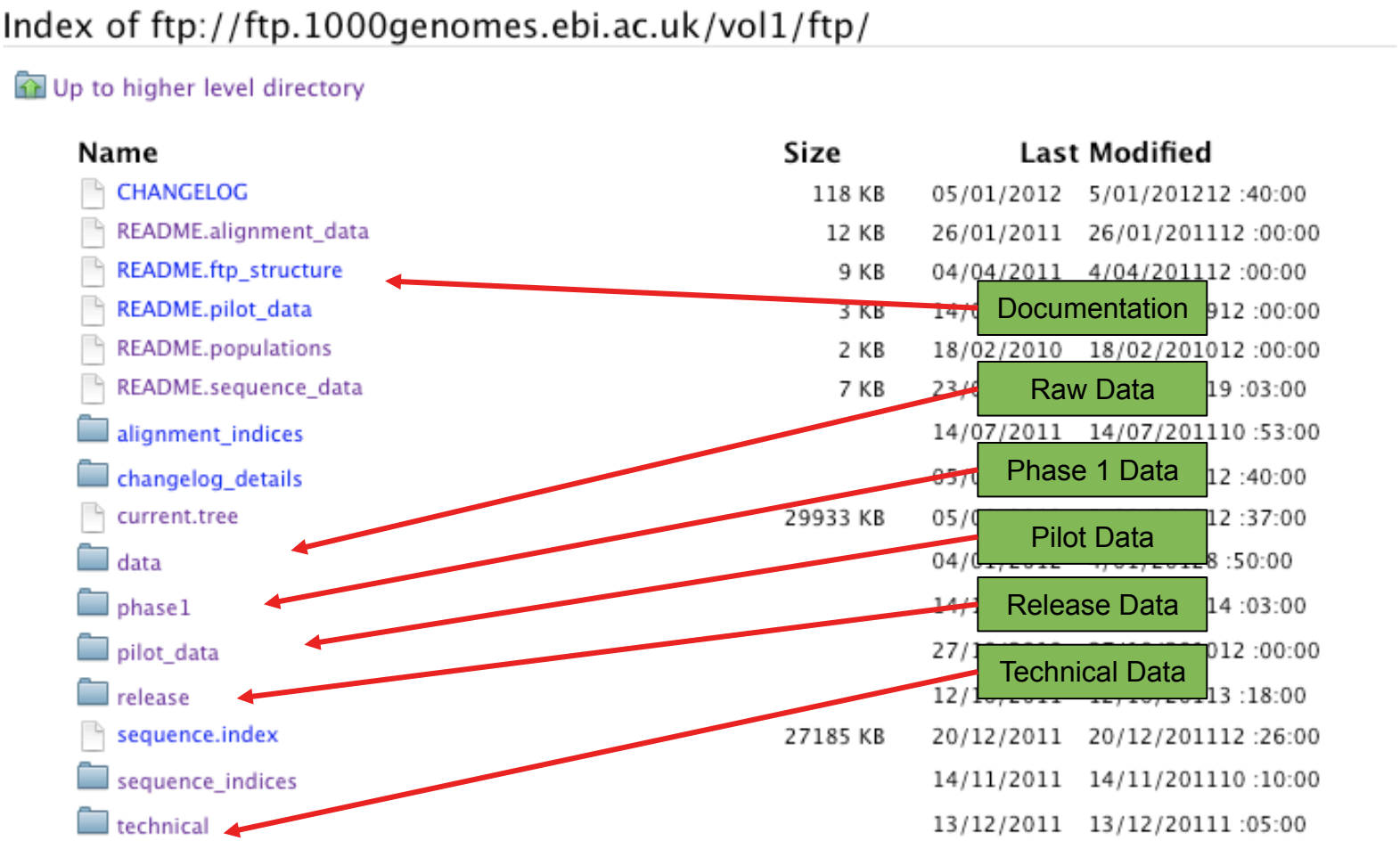
EMBL-EBI

FTP Site

- Two mirrored ftp sites
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp
  - ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp
- NCBI site is direct mirror of EBI site
- Can be up to 24 hours out of date
- Both also accessible using aspera
- http://asperasoft.com/
- EBI site has http mirror
  - http://ftp.1000genomes.ebi.ac.uk/vol1/ftp

EMBL-EBI

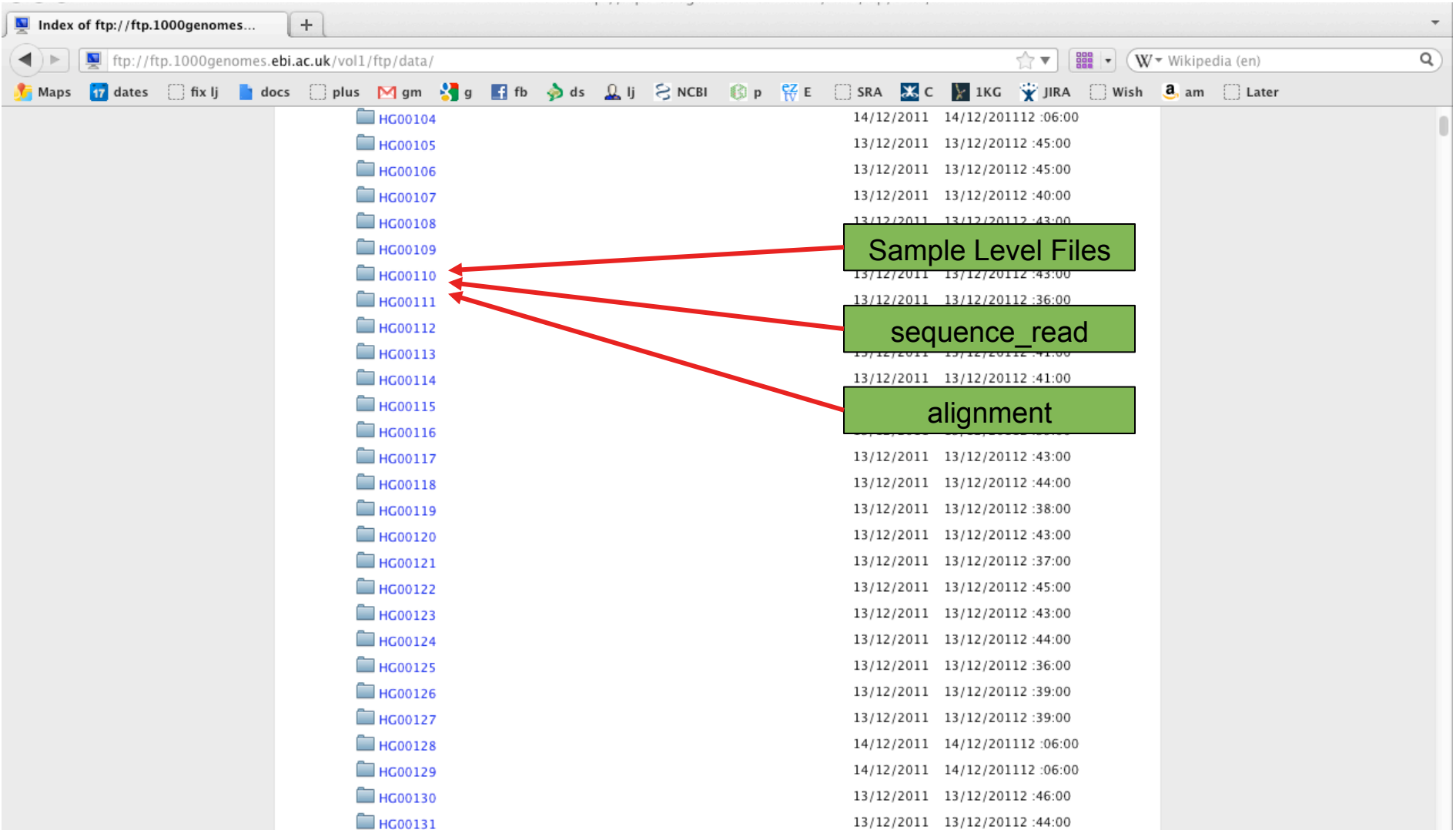
ftp://ftp.1000genomes.ebi.ac.uk
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp



A THOUSAND GENOMES

EMBL-EBI

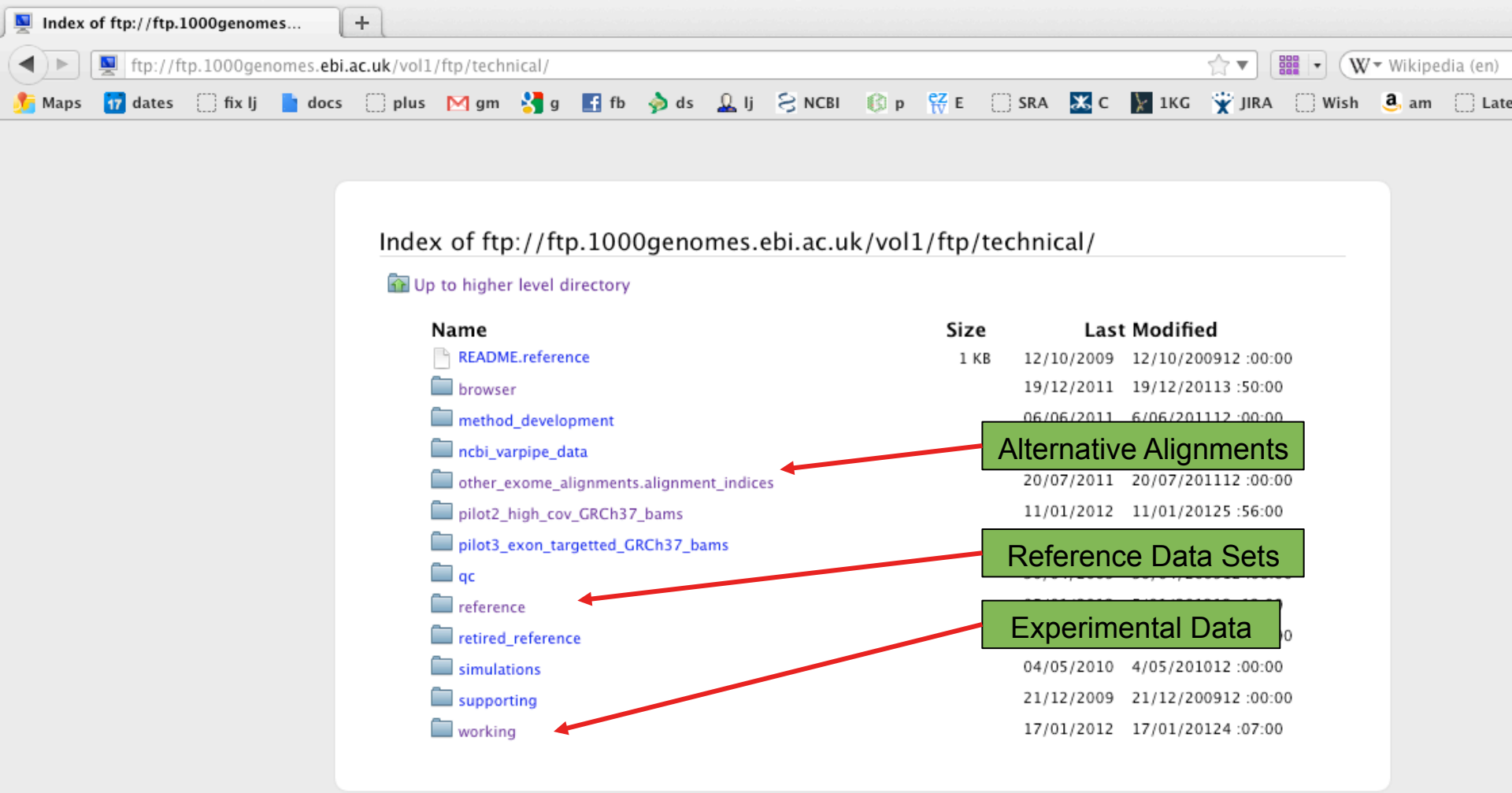
The FTP Site: Data



A THOUSAND GENOMES

EMBL-EBI

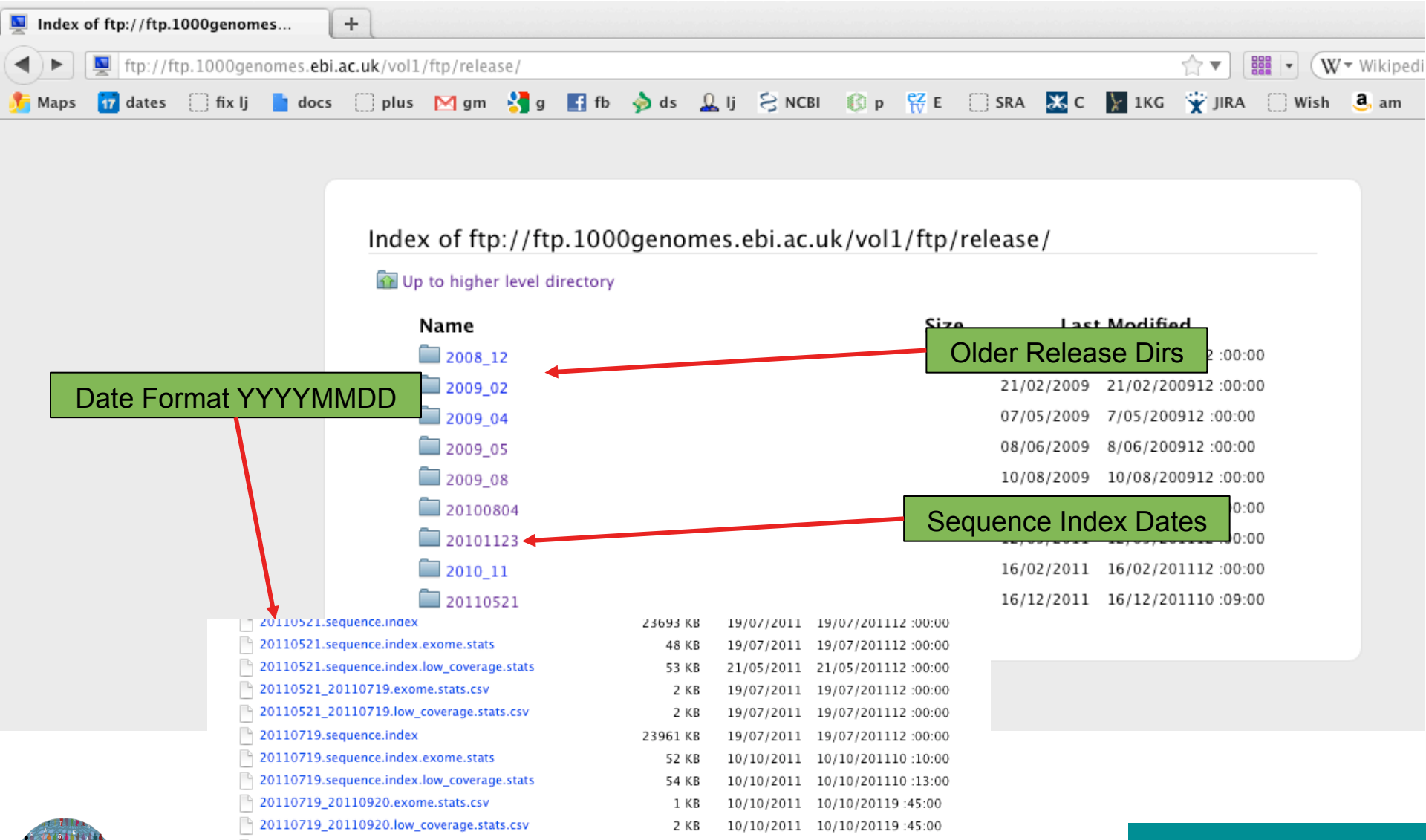
FTP Site: Technical



A THOUSAND GENOMES

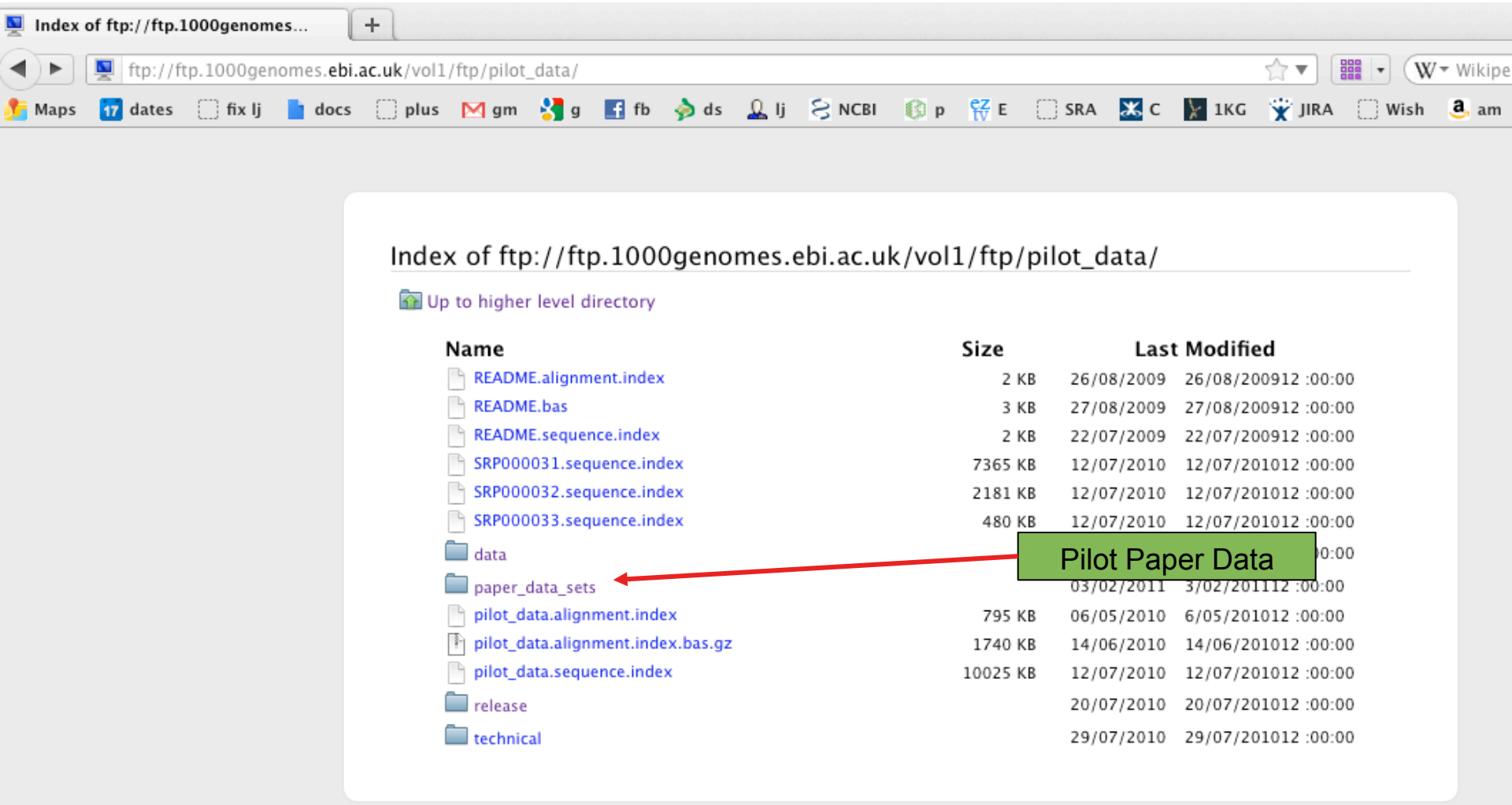
EMBL-EBI

FTP Site: Release



EMBL-EBI

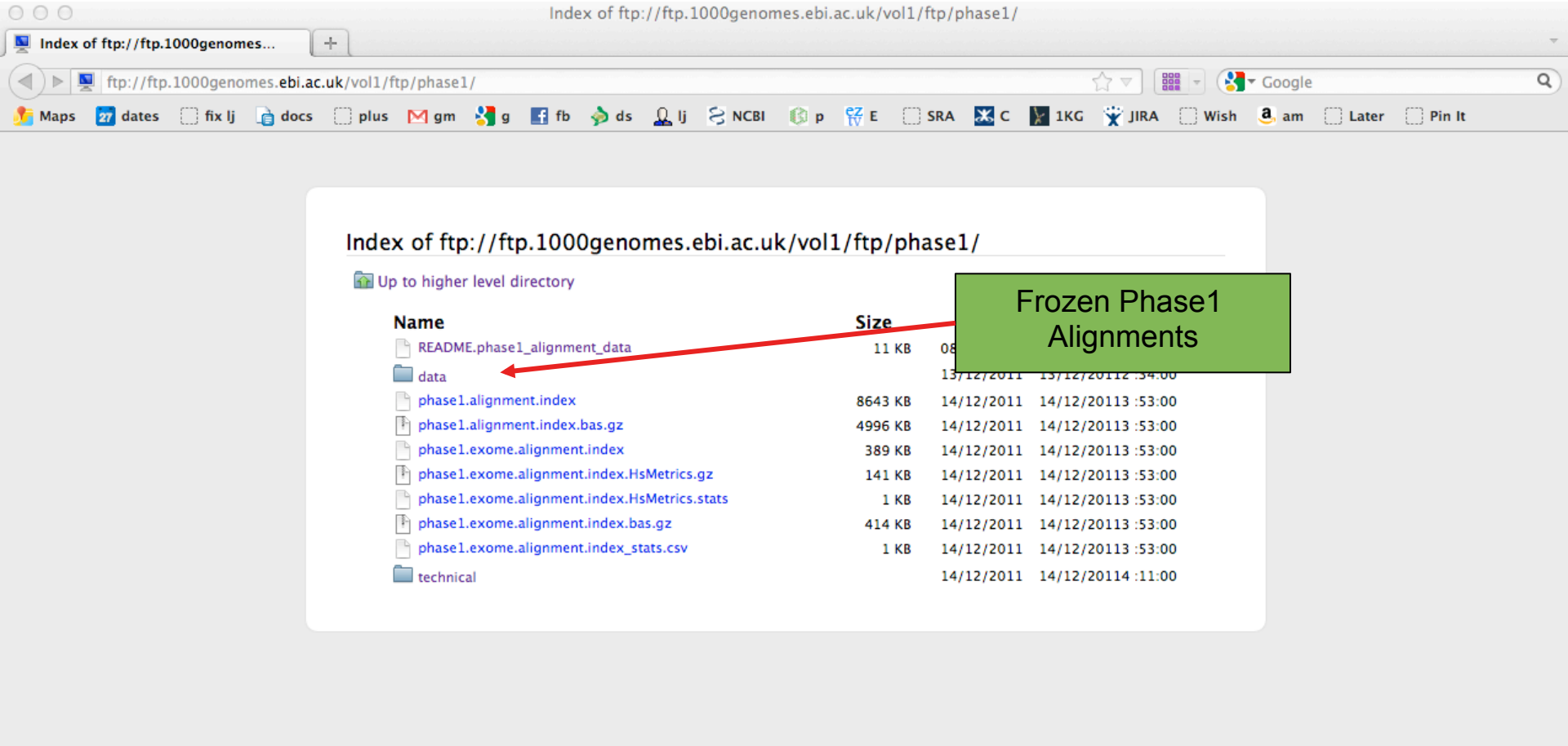
FTP Site: Pilot Data



A THOUSAND GENOMES

EMBL-EBI

FTP Site: Phase 1

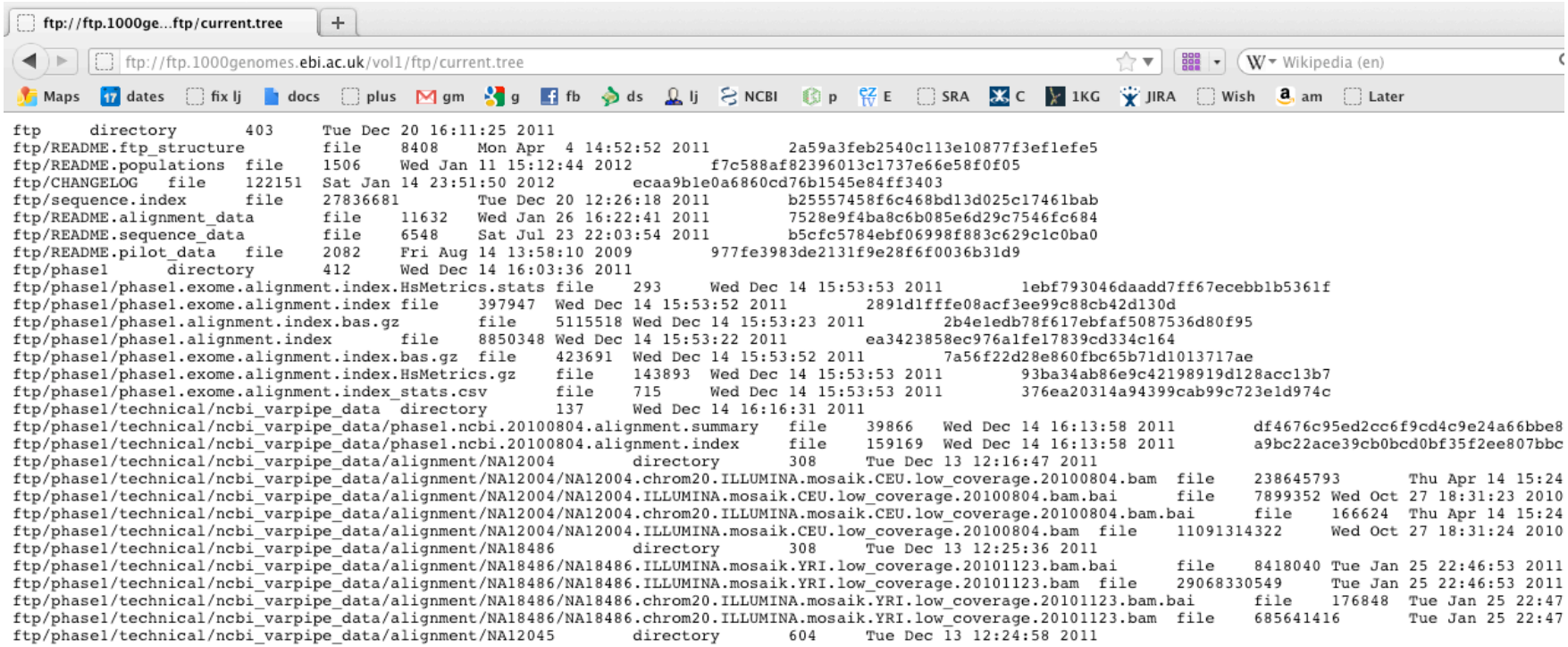


A THOUSAND GENOMES

EMBL-EBI

Finding Data

- Current.tree file
- ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree
- Current Tree is updated nightly so can be upto 24 hours out of date




```
ftp     directory       403     Tue Dec 20 16:11:25 2011
ftp/README.ftp_structure        file    8408    Mon Apr  4 14:52:52 2011        2a59a3feb2540c113e10877f3ef1efe5
ftp/README.populations  file    1506    Wed Jan 11 15:12:44 2012        f7c588af82396013c1737e66e58f0f05
ftp/CHANGELOG   file    122151  Sat Jan 14 23:51:50 2012        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
```

Finding Data

- Current tree file

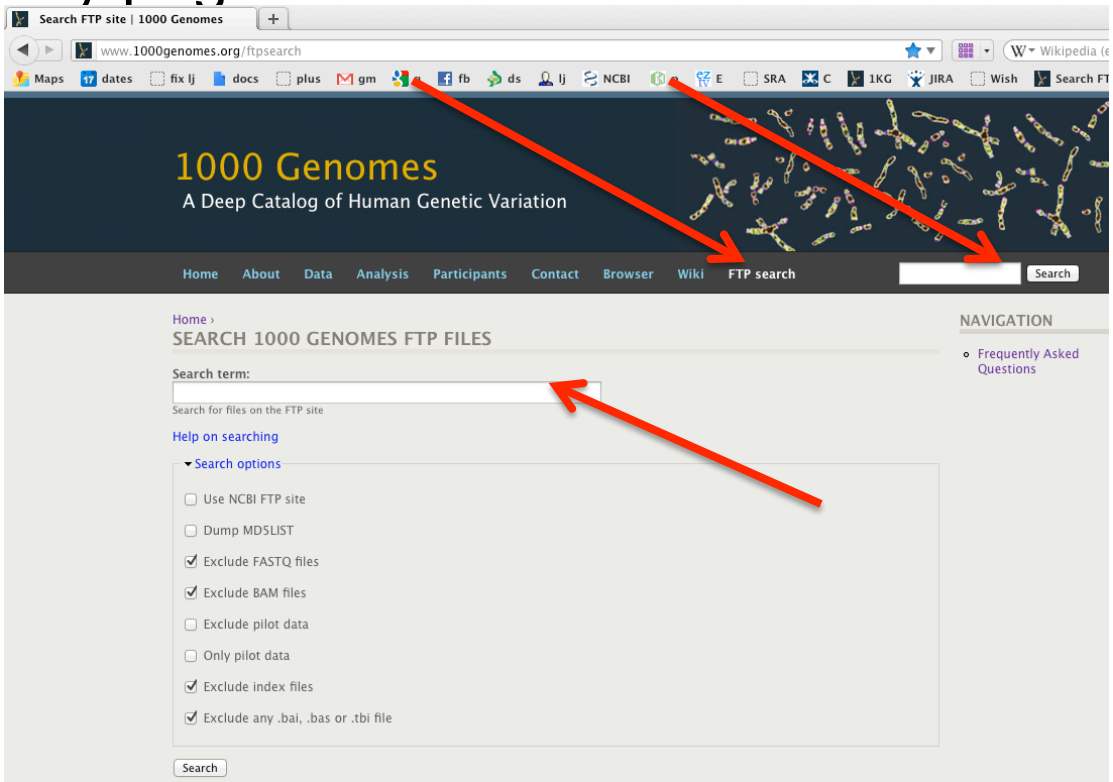
Description	Example
Relative Path	ftp/data/NA21091/alignment/ NA21091.chrom20.ILLUMINA.bwa.GIH.low_coverage. 20111114.bam
Type (file/directory)	file
Size in bytes	297914382
Last Updated Time Stamp	Thu Jan 26 00:26:52 2012
MD5 checksum	3fd679acc8c92cdc838aa0e5c1849d58

- Relative path does not contain the complete ftp path
- ftp://ftp.1000genomes.ebi.ac.uk/vol1/
- ftp://ftp-trace.ncbi.nih.gov/1000genomes/

EMBL-EBI

Finding Data

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



EMBL-EBI

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

EMBL-EBI

Exercises

1a. Find what Omni VCF files we have on our ftp site using the website ftp search. (Omni is a high throughput genotyping platform from Illumina on which all 1000 genomes samples are being genotyped)

1b. Find the most recent Omni VCF file on GRCh37 from the 31st January 2012

2. Use the Website search box found in the top right hand corner of all pages to find the FAQ question about getting subsections of VCF files.

Hints: You can use http://www.1000genomes.org/ftpsearch or ftp://ftp.1000genomes.ebi.ac.uk/vol1/current.tree and grep to answer 1a and b

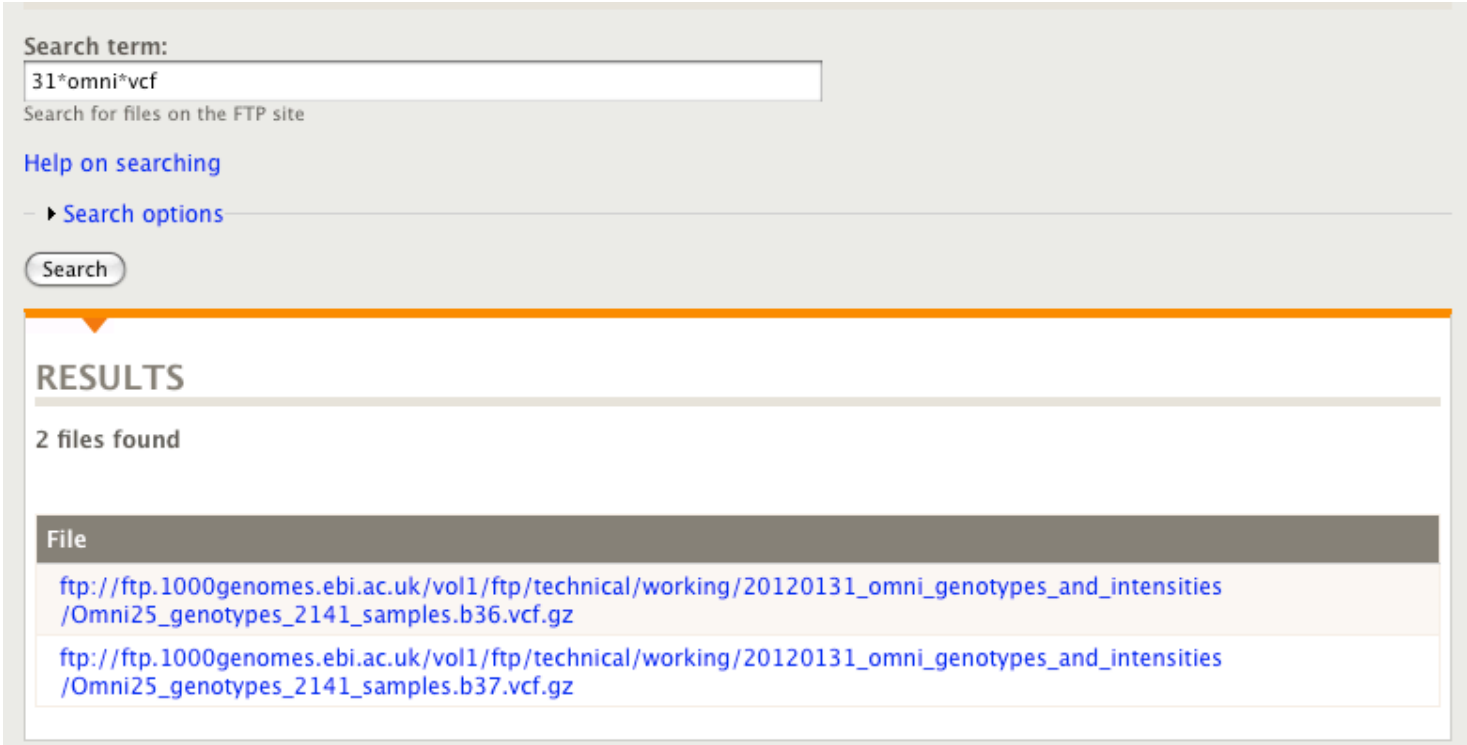
EMBL-EBI

Exercise Answers

1a. Put omni*vcf into the ftp site search box

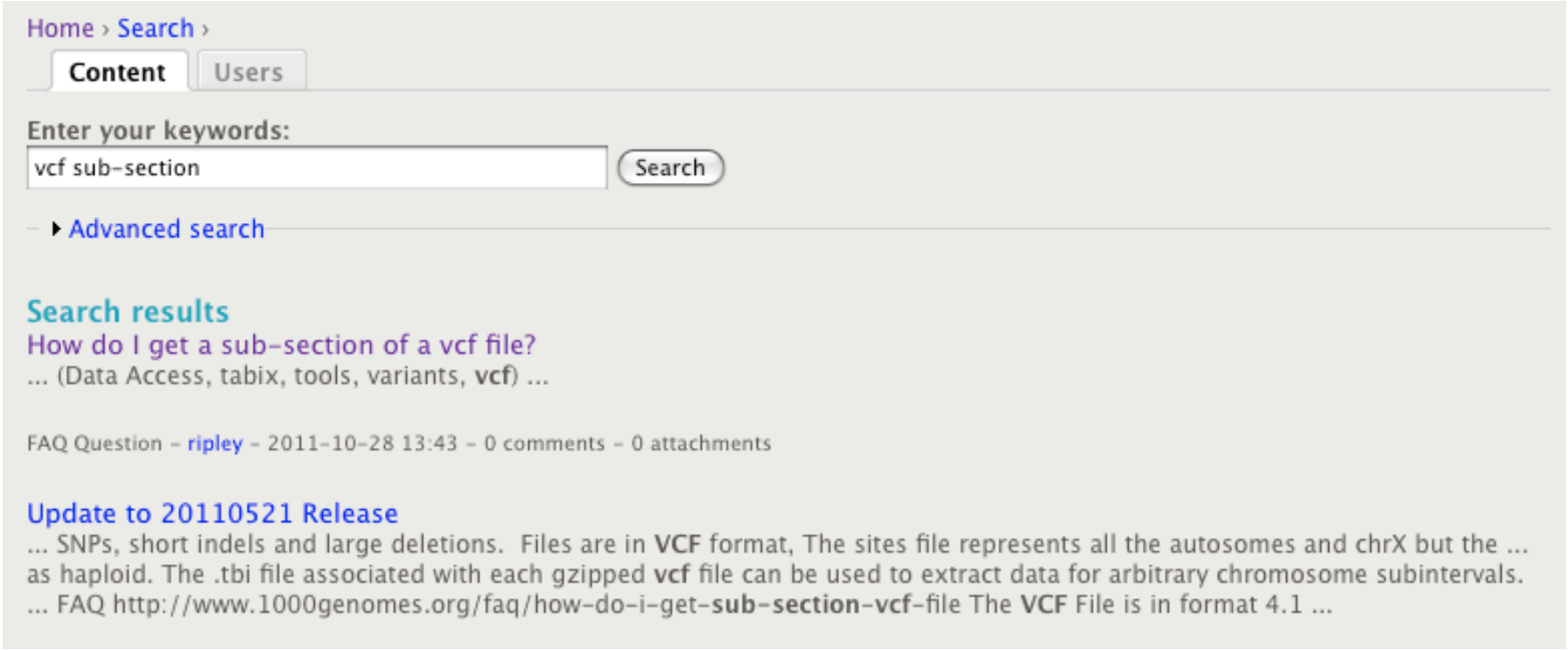
Exercise Answers, Finding Data

1b. Use 31*omni*vcf to get results. This should return 2 files. One is labeled b36 and it in NCBI36 coordinates. The other is labeled b37 and is on GRCh37



Exercise Answers, Finding Data

2. Using the box that is in the top right hand corner of every page of 1000genomes.org with the term sub-section and vcf should return the appropriate FAQ page



The 1000 Genomes Browser
http://browser.1000genomes.org

EMBL-EBI

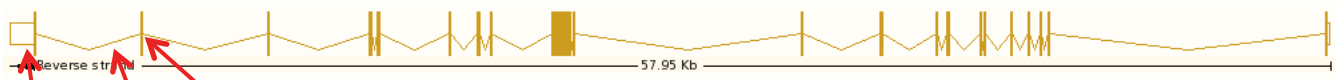
Caveats

- 1000 Genomes and Ensembl always define variants on the forward strand

- Allele strings are always reported ref/alt

EMBL-EBI

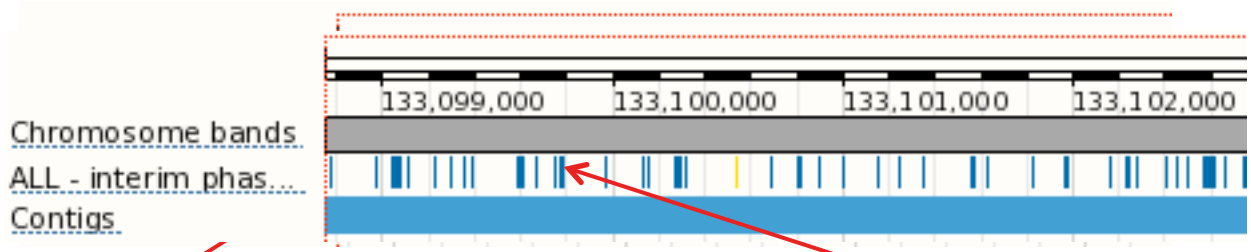
Genes and SNPs



UTR

Intron

Coding



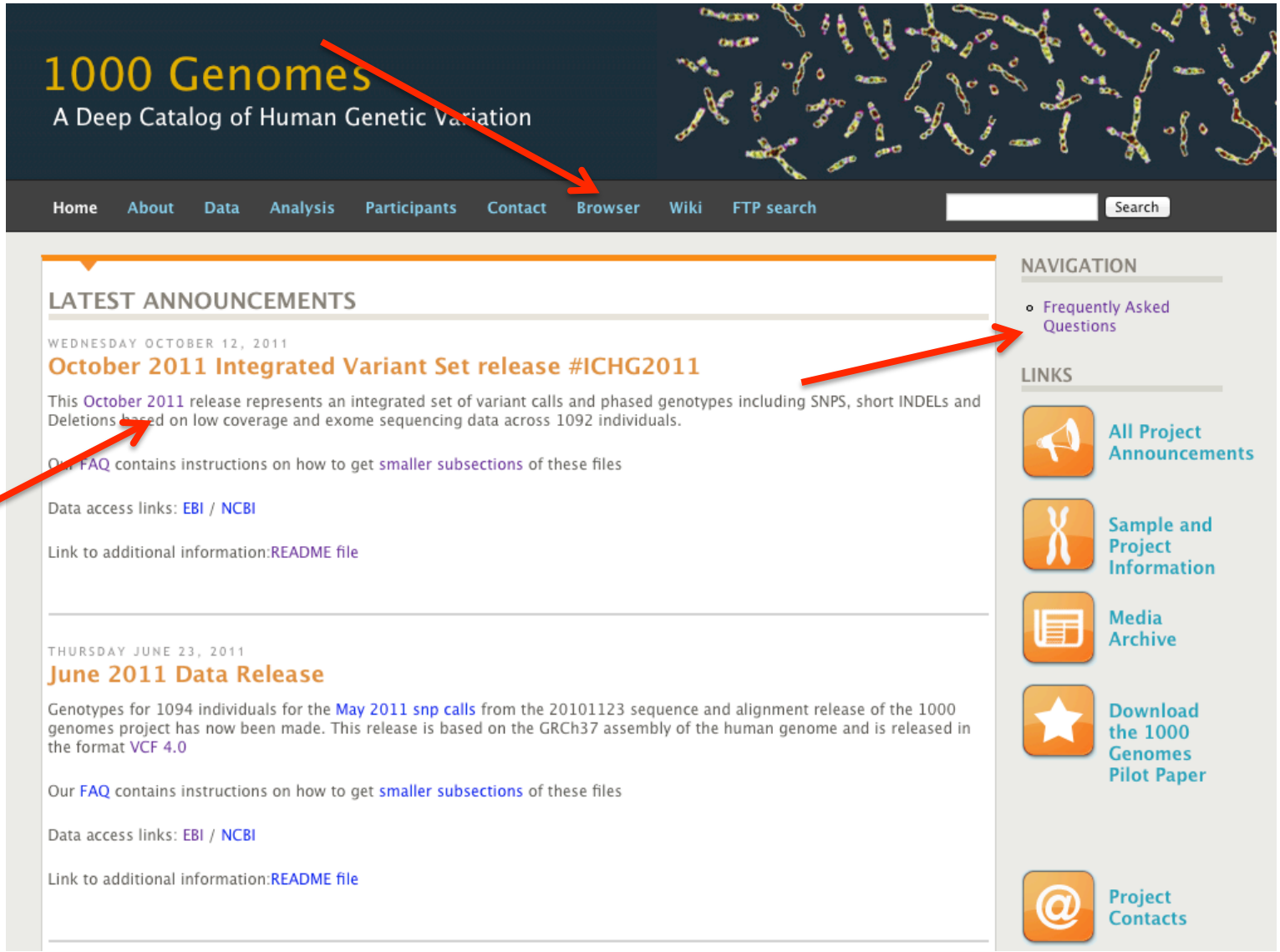
Line indicates number of SNPS

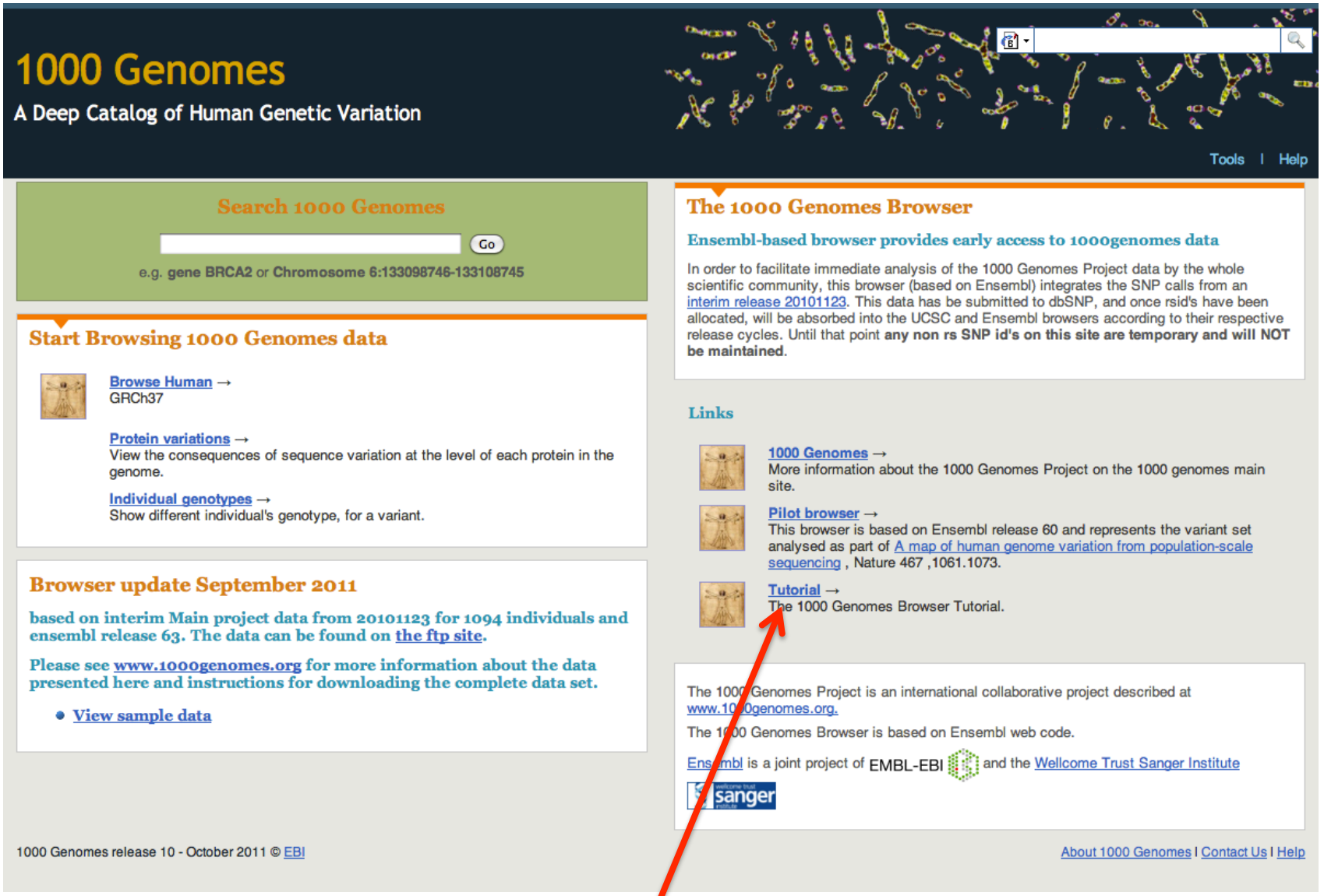
Each Line is One SNP

A THOUSAND GENOMES

EMBL-EBI

http://www.1000genomes.org





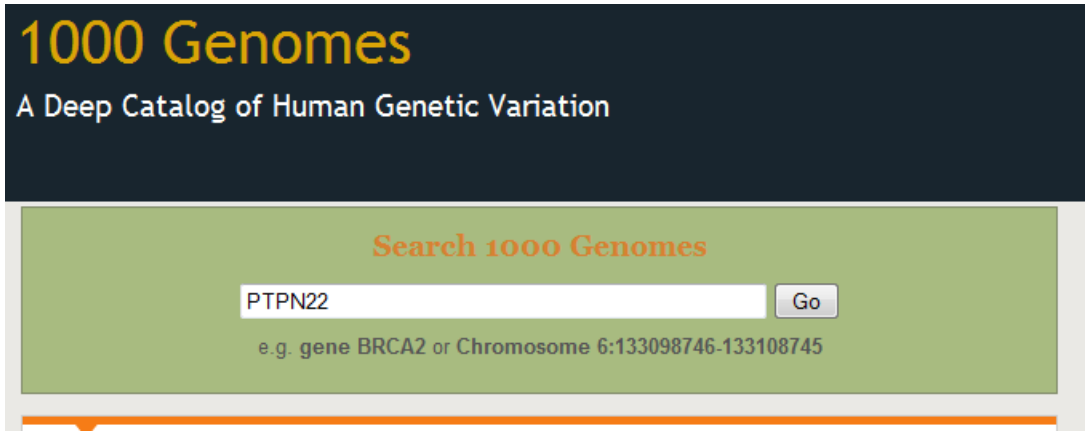
http://browser.1000genomes.org

A THOUSAND GENOMES

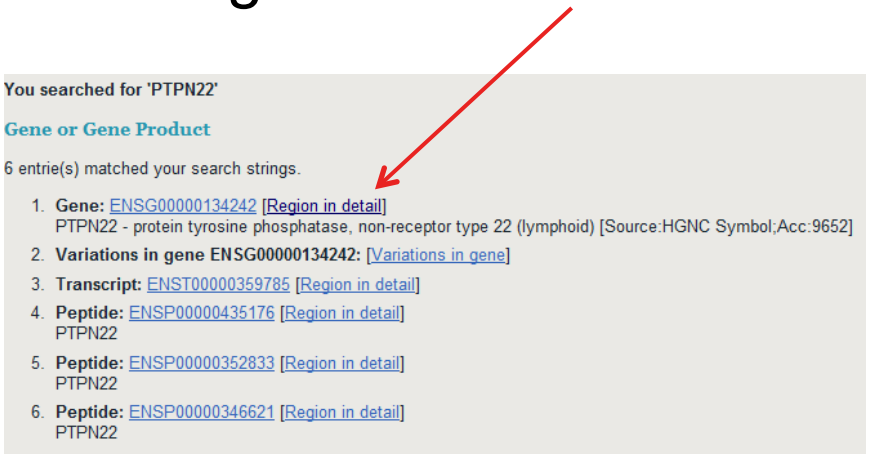
EMBL-EBI

Searching the Browser

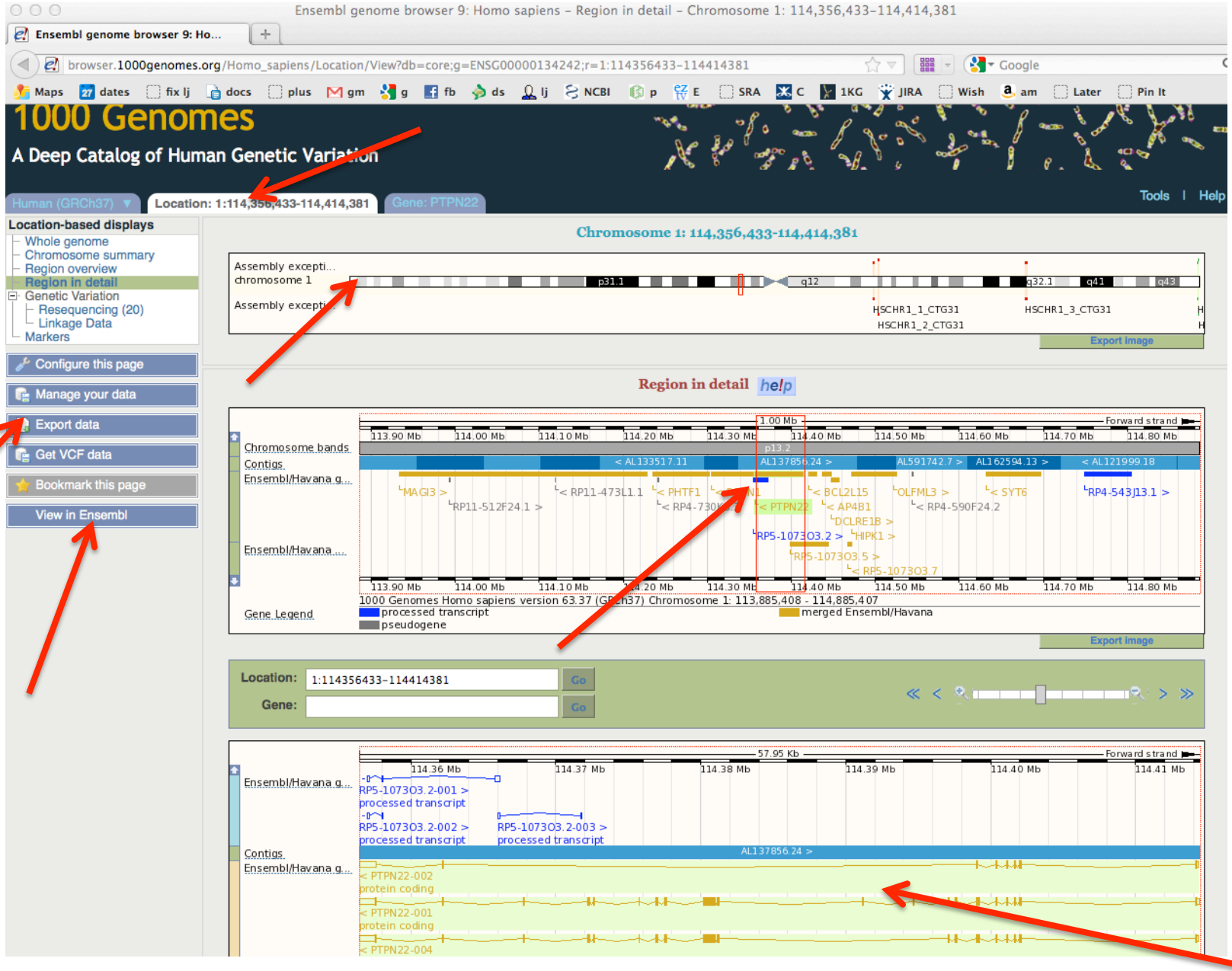
- *http://browser.1000genomes.org*



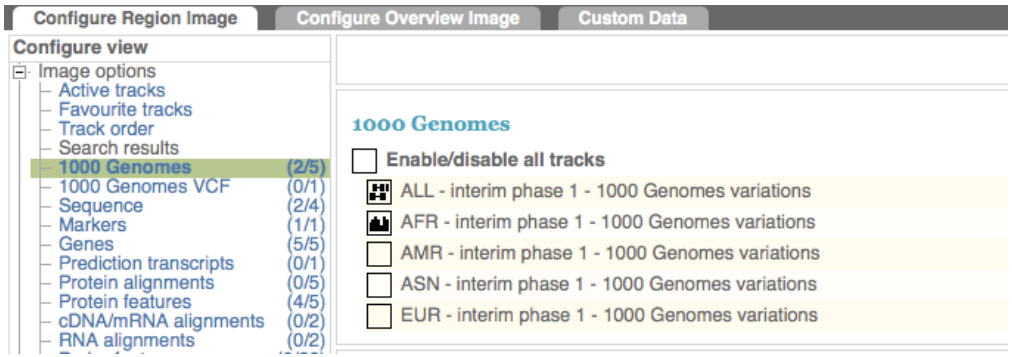
- Search for PTPN22
- Click ‘Region in Detail’

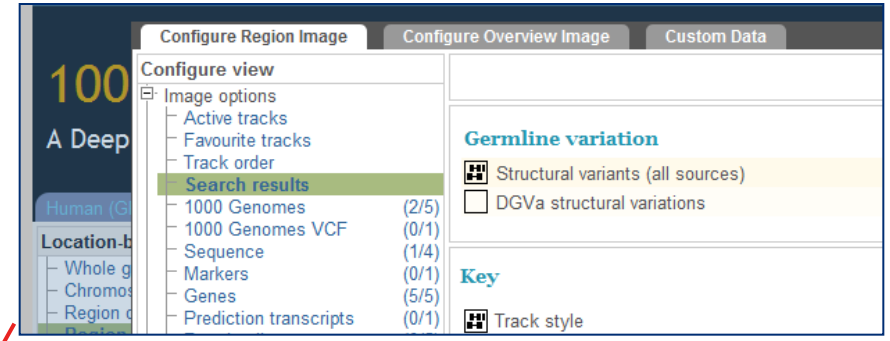


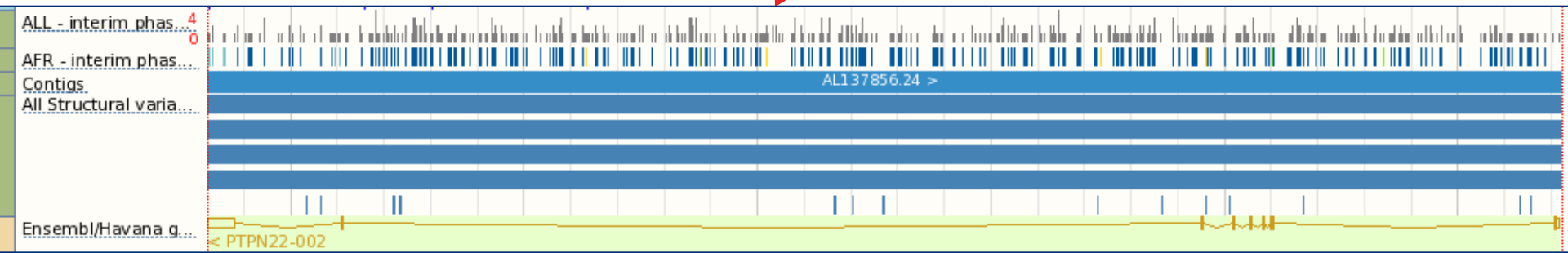
Region in Detail



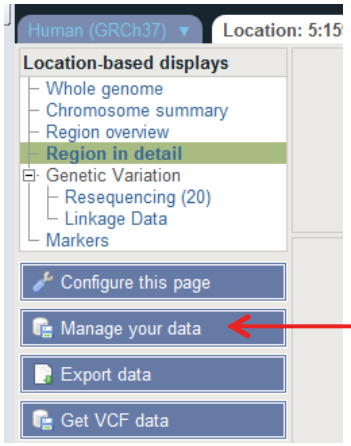
Turning on Tracks



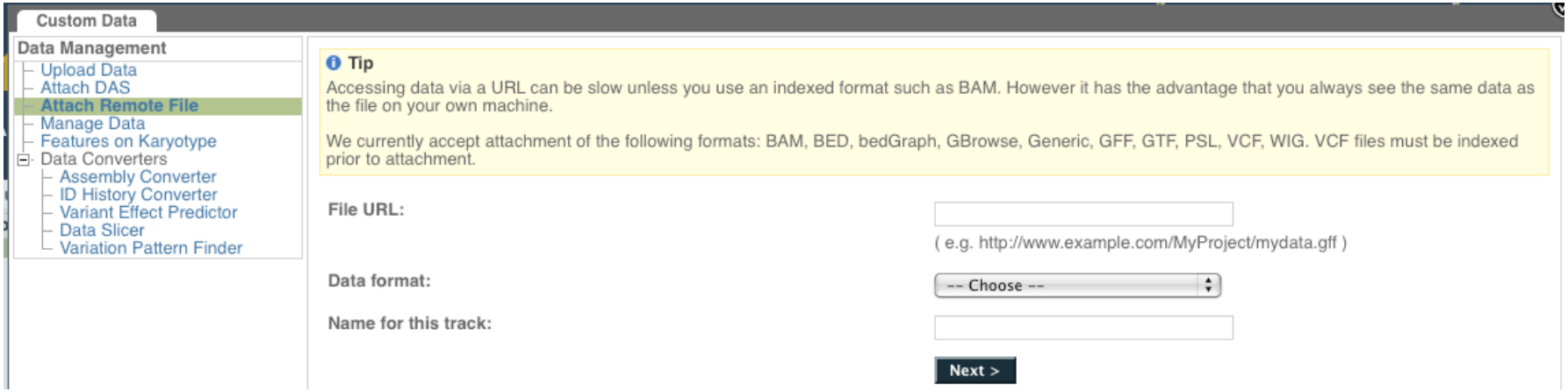




File upload to view with 1000 Genomes data



Manage your data



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

* VCF must be indexed

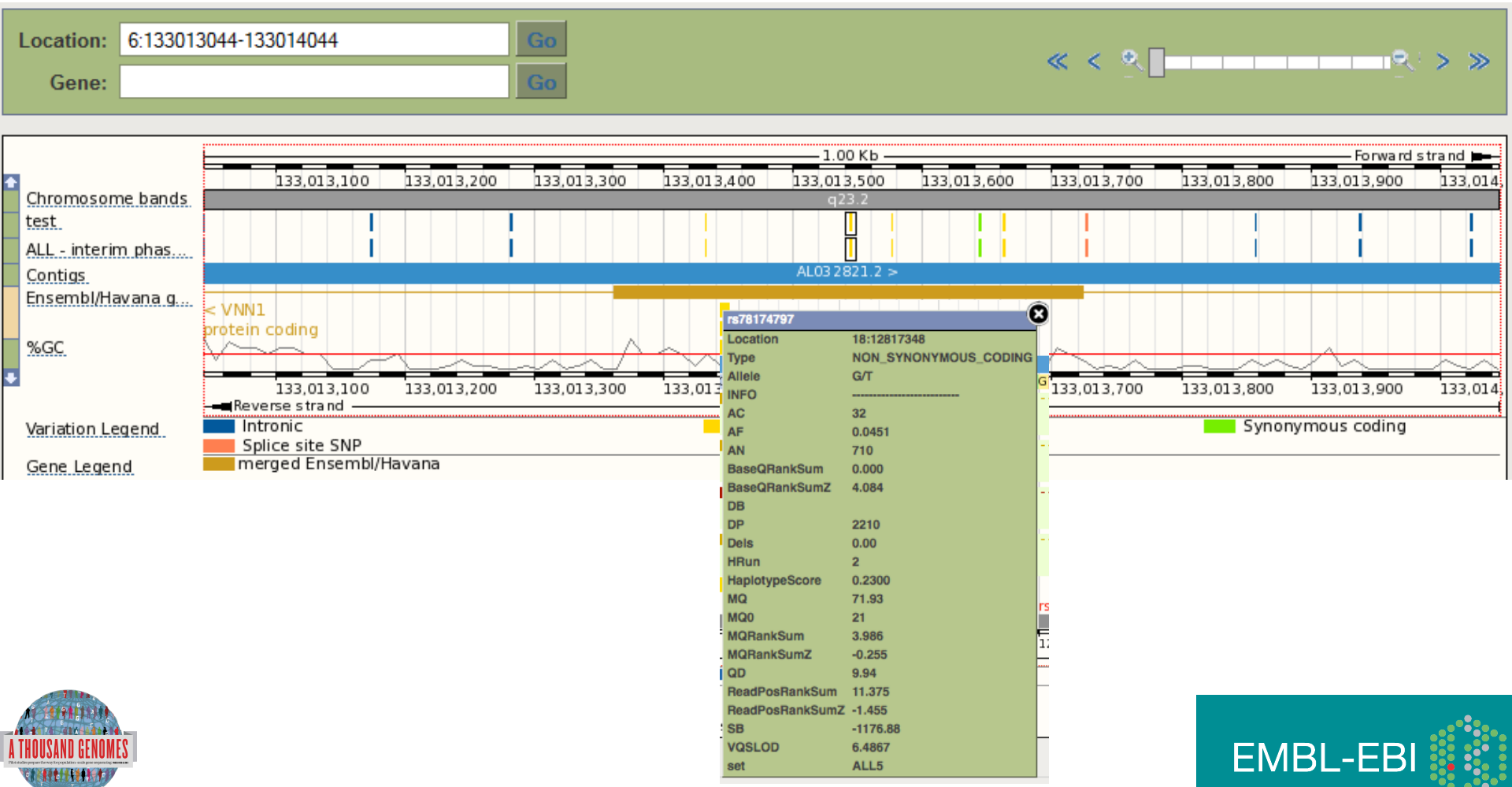
A THOUSAND GENOMES

EMBL-EBI

Uploaded VCF

Example:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.wgs.phase1_release_v2.20101123.snps_indels_sv.sites.vcf.gz



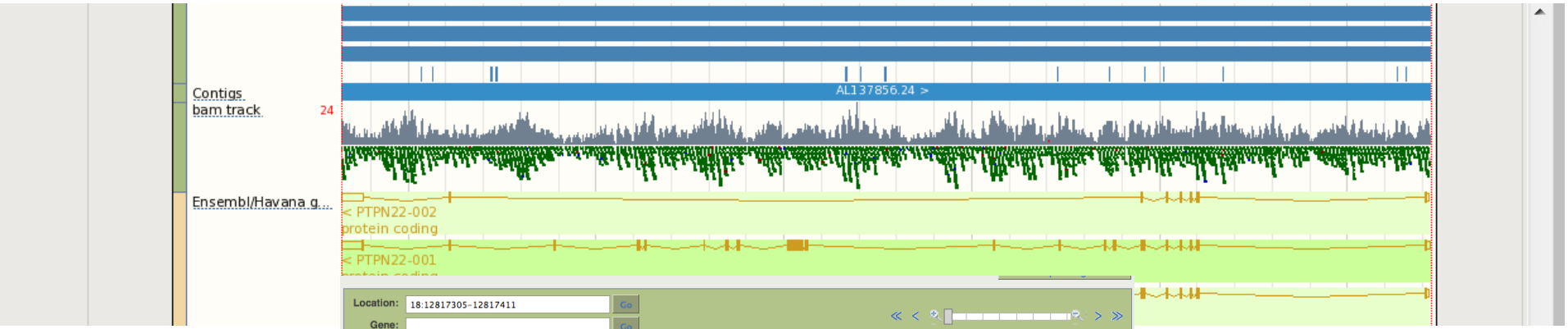
A THOUSAND GENOMES

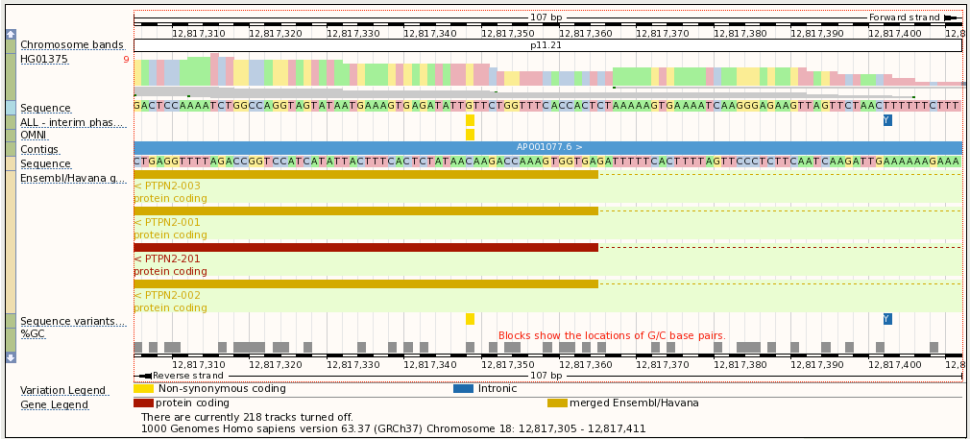
EMBL-EBI

Uploaded BAM

Example:

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



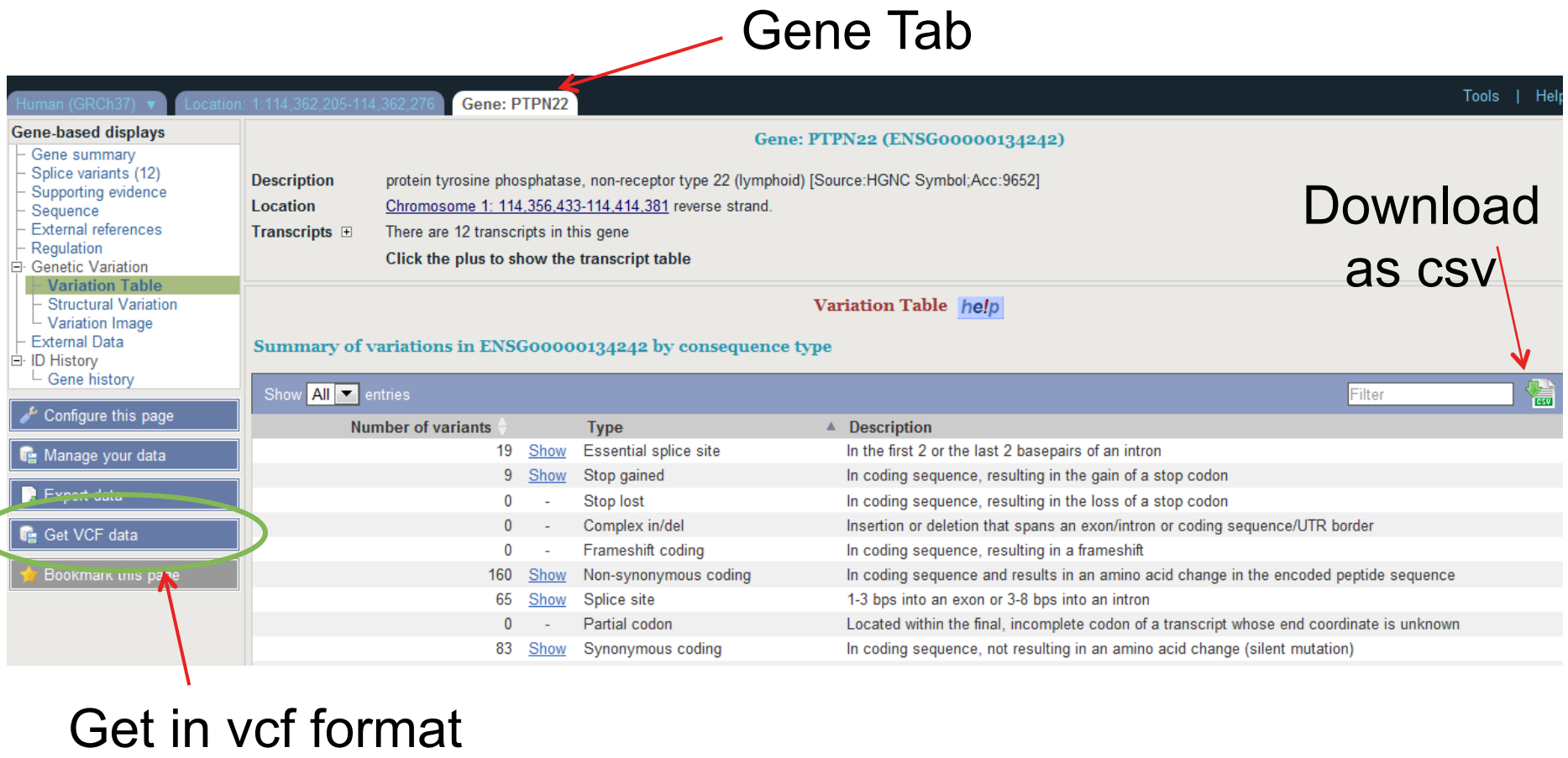


A THOUSAND GENOMES

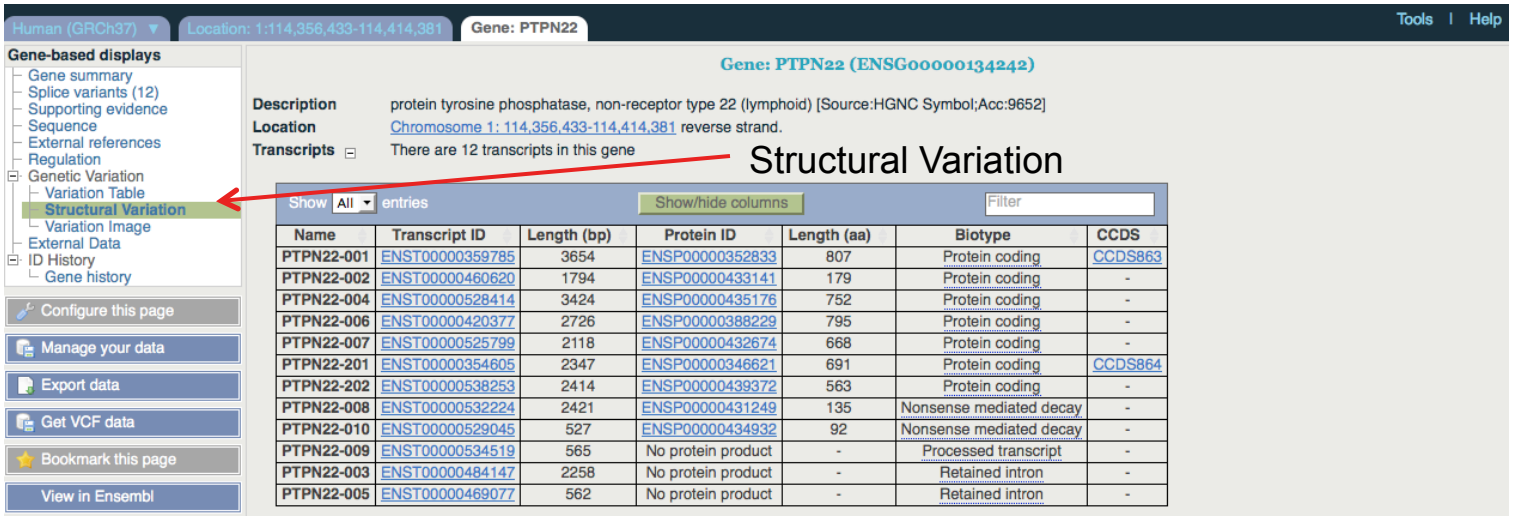
EMBL-EBI

Gene View

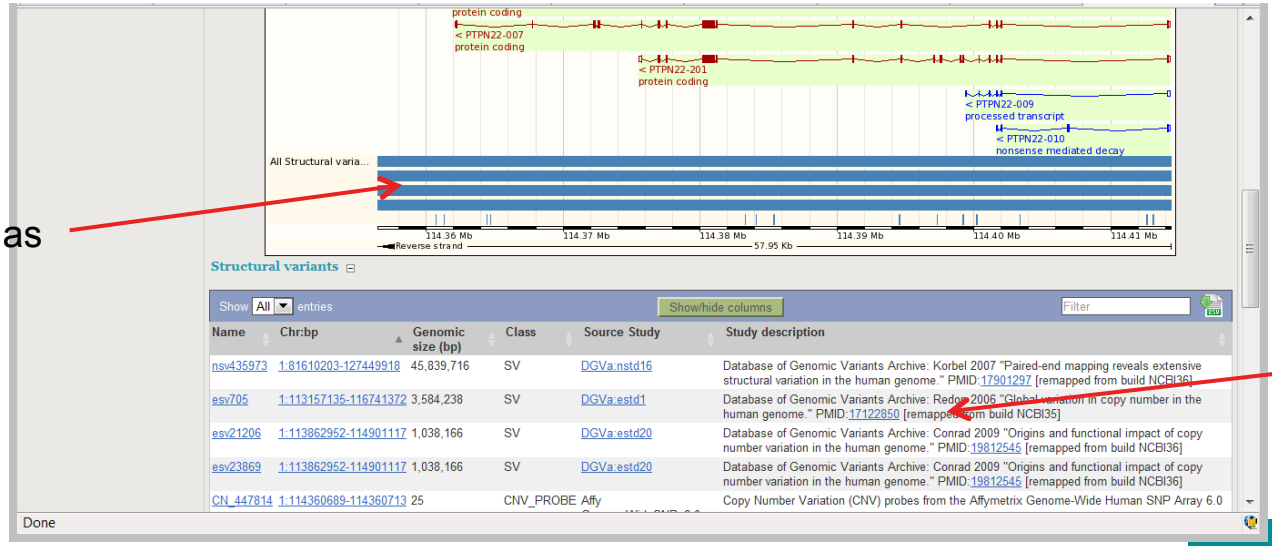
Click the Gene tab, then ‘Variation Table’ or ‘Variation Image’



Structural variation (in the Gene tab)



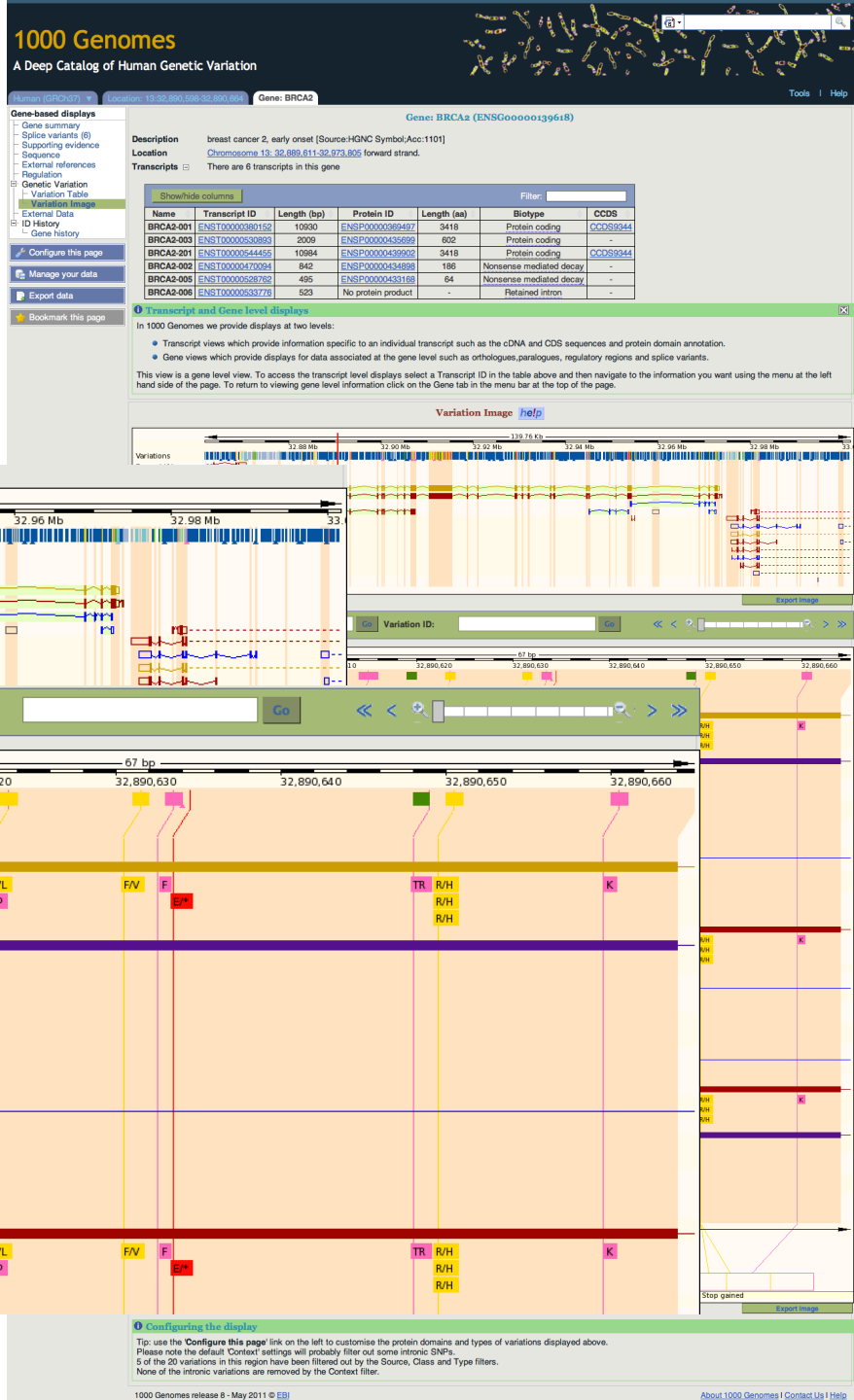
Structural Variants as boxes



Table

Variation Image

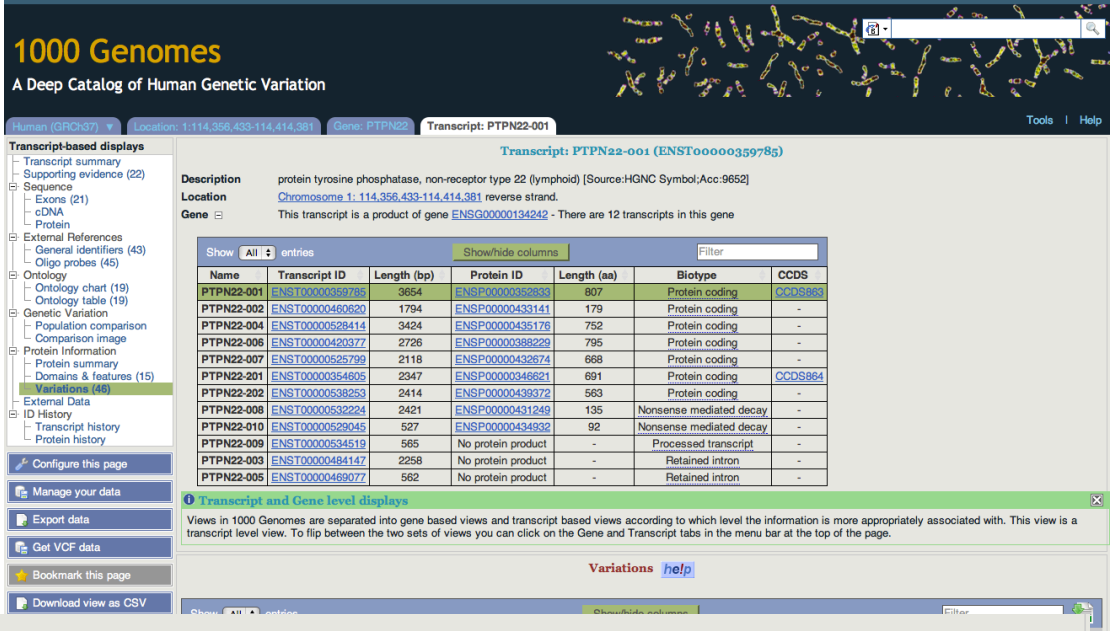
- Gene variation zoom



Transcript Tab: Variations

Effect on Protein:

- SIFT
- PolyPhen

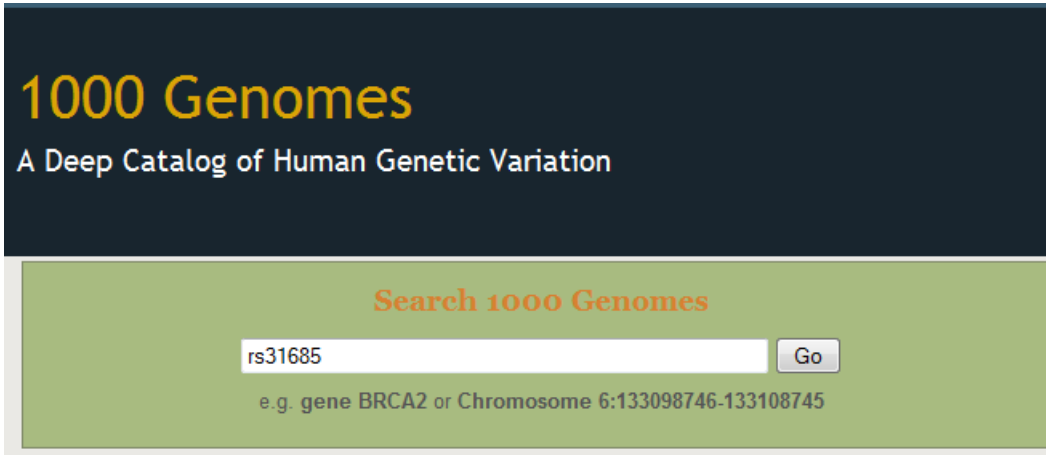


Download view as CSV

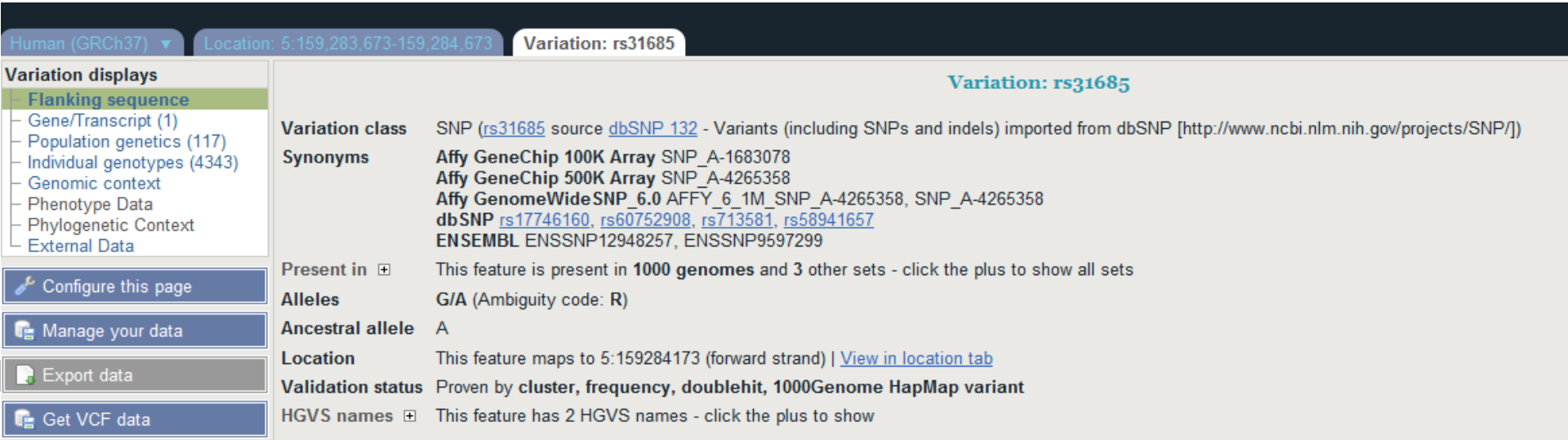
Show All entries | Show/hide columns | Filter

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
778	rs41313296	Non-synonymous coding	T/A	W	N, I	AAT, ATT	deleterious	probably damaging

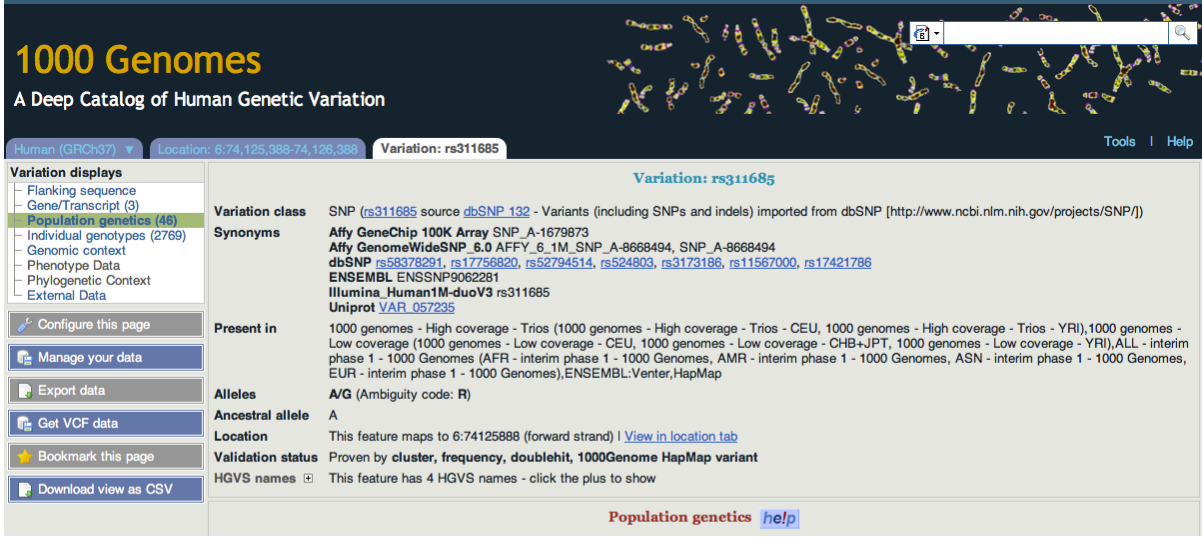
Start again- search for a variation (rs31685)



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



- Population



1000 genomes alleles frequencies

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

1000 genomes

Show/hide columns | Filter | CSV

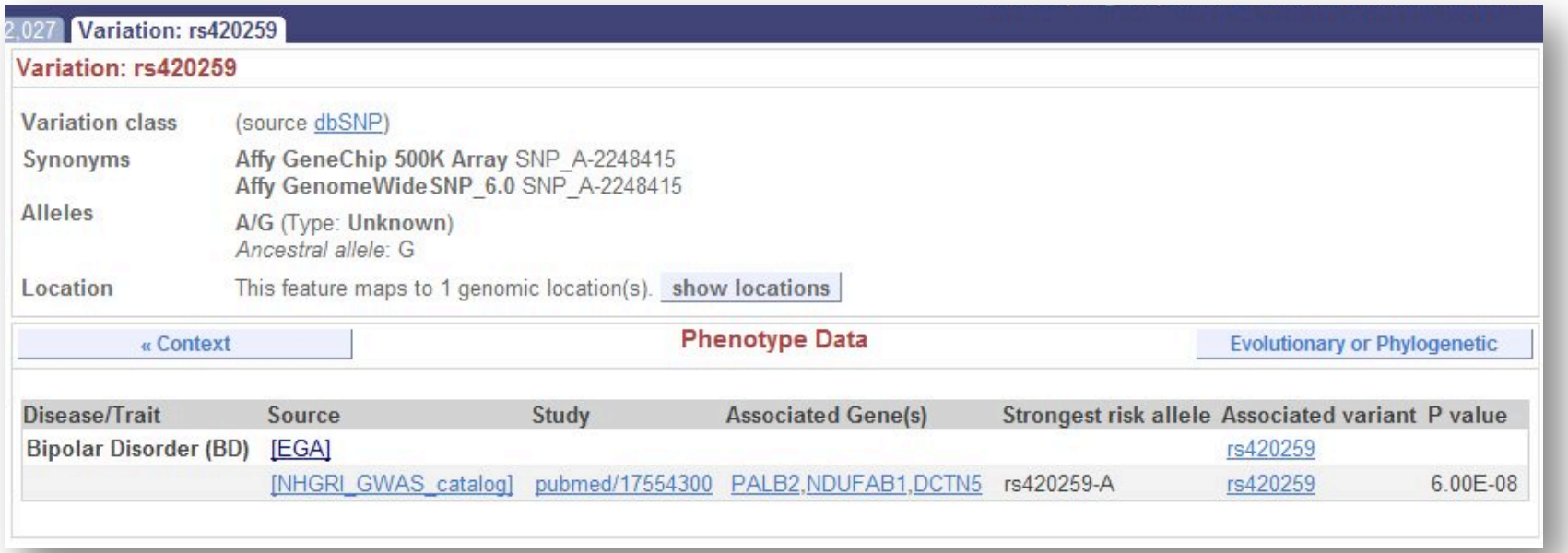
Population	Alleles A	Alleles G	Genotypes A|A	Genotypes A|G	Genotypes G|G	Count
1000GENOMES:AFR	0.689	0.311	0.463	0.451	0.085	114
1000GENOMES:ALL	0.507	0.493	0.269	0.477	0.254	294
1000GENOMES:AMR	0.539	0.461	0.293	0.492	0.215	53
1000GENOMES:ASN	0.446	0.554	0.199	0.493	0.308	57
1000GENOMES:EUR	0.421	0.579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

Phenotype for one variant



2,027 Variation: rs420259

Variation: rs420259

Variation class	(source dbSNP)
Synonyms	Affy GeneChip 500K Array SNP_A-2248415 Affy GenomeWideSNP_6.0 SNP_A-2248415
Alleles	A/G (Type: **Unknown**) *Ancestral allele*: G
Location	This feature maps to 1 genomic location(s). show locations

« Context | **Phenotype Data** | Evolutionary or Phylogenetic

Disease/Trait	Source	Study	Associated Gene(s)	Strongest risk allele	Associated variant	P value
Bipolar Disorder (BD)	[EGA]				rs420259	
	[NHGRI_GWAS_catalog]	pubmed/17554300	PALB2,NDUFAB1,DCTN5	rs420259-A	rs420259	6.00E-08

EGA	*http://www.ebi.ac.uk/ega*
NHGRI	*http://www.genome.gov/gwastudies/*
Open GWAS DB	*http://www.biomedcentral.com/1471-2350/10/6*
COSMIC	*http://www.sanger.ac.uk/genetics/CGP/cosmic/*
OMIM	*http://www.ncbi.nlm.nih.gov/omim*
HGMD-Public	*http://www.hgmd.cf.ac.uk/ac/index.php*
UniProt	*http://www.uniprot.org/*

Coming Soon Ensembl 65

Should arrive in May

EMBL-EBI

Exercise, Browser

3. Find the variant rs45562238 using http://browser.1000genomes.org.

4. In what 1000 Genomes Super Population is this variant detected?

5. What are its global allele frequencies in the 1000 Genomes Data set?

6. In which gene is the variant found?

EMBL-EBI

Exercise Answers, Browser

3

SNP

1 entrie(s) matched your search strings.

1. dbSNP SNP: rs45562238

Interpro Domain

0 entrie(s) matched your search strings.

Exercise Answers

4. In what 1000 Genomes Super Population is this variant detected?

American and European

5. What are its global allele frequencies in the 1000 Genomes Data set?

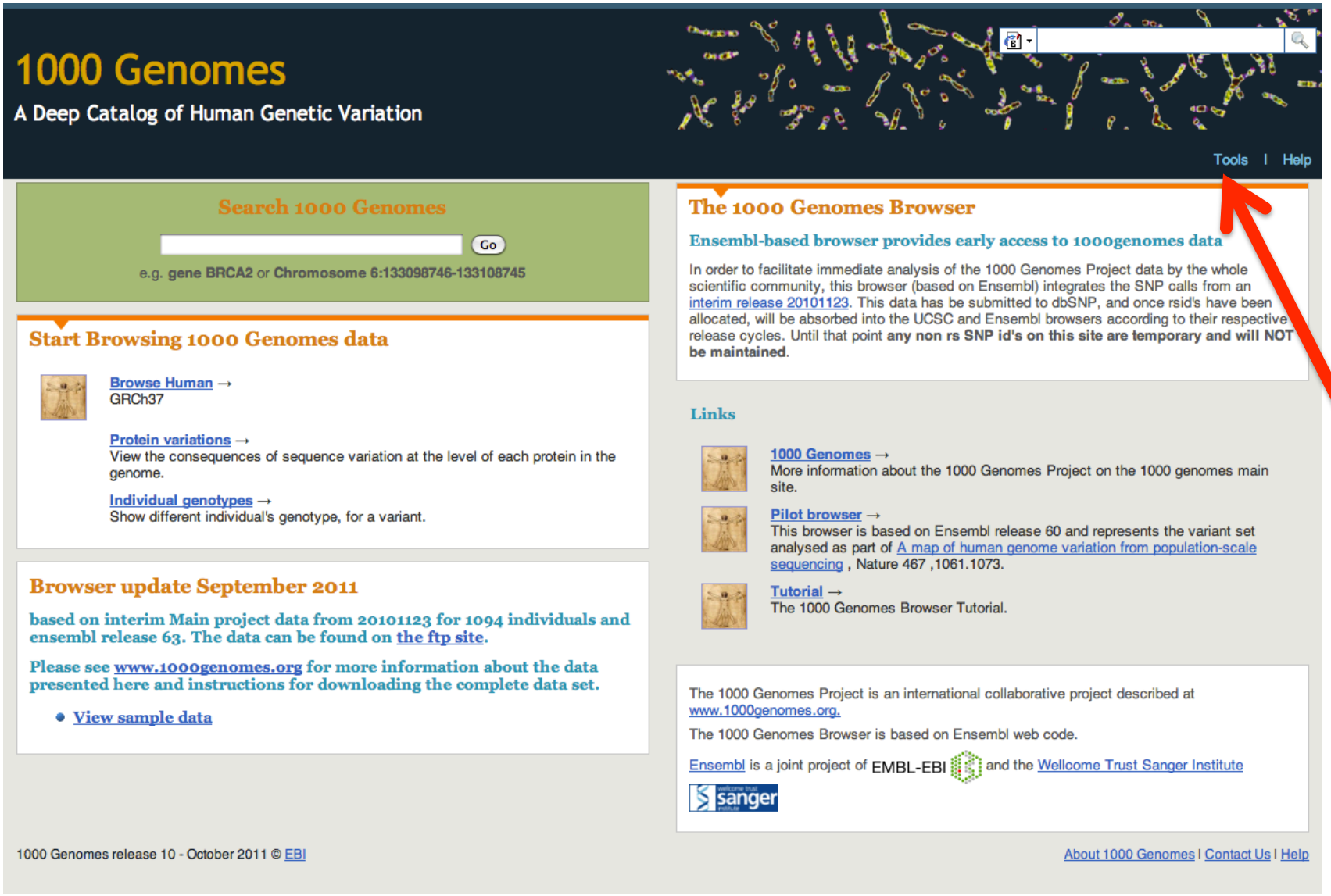
0.02 is the global allele frequency, this is also the American Allele Frequency but it rises to 0.04 in the Europeans. The absence of Asians or Africans in this chart means that the variant was not found in any of our Asian of African individuals.

6. In which gene is the variant found?

ENSG00000112299, Vanin 1

1000 Genomes Tools

EMBL-EBI



http://browser.1000genomes.org

EMBL-EBI

Tools page

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

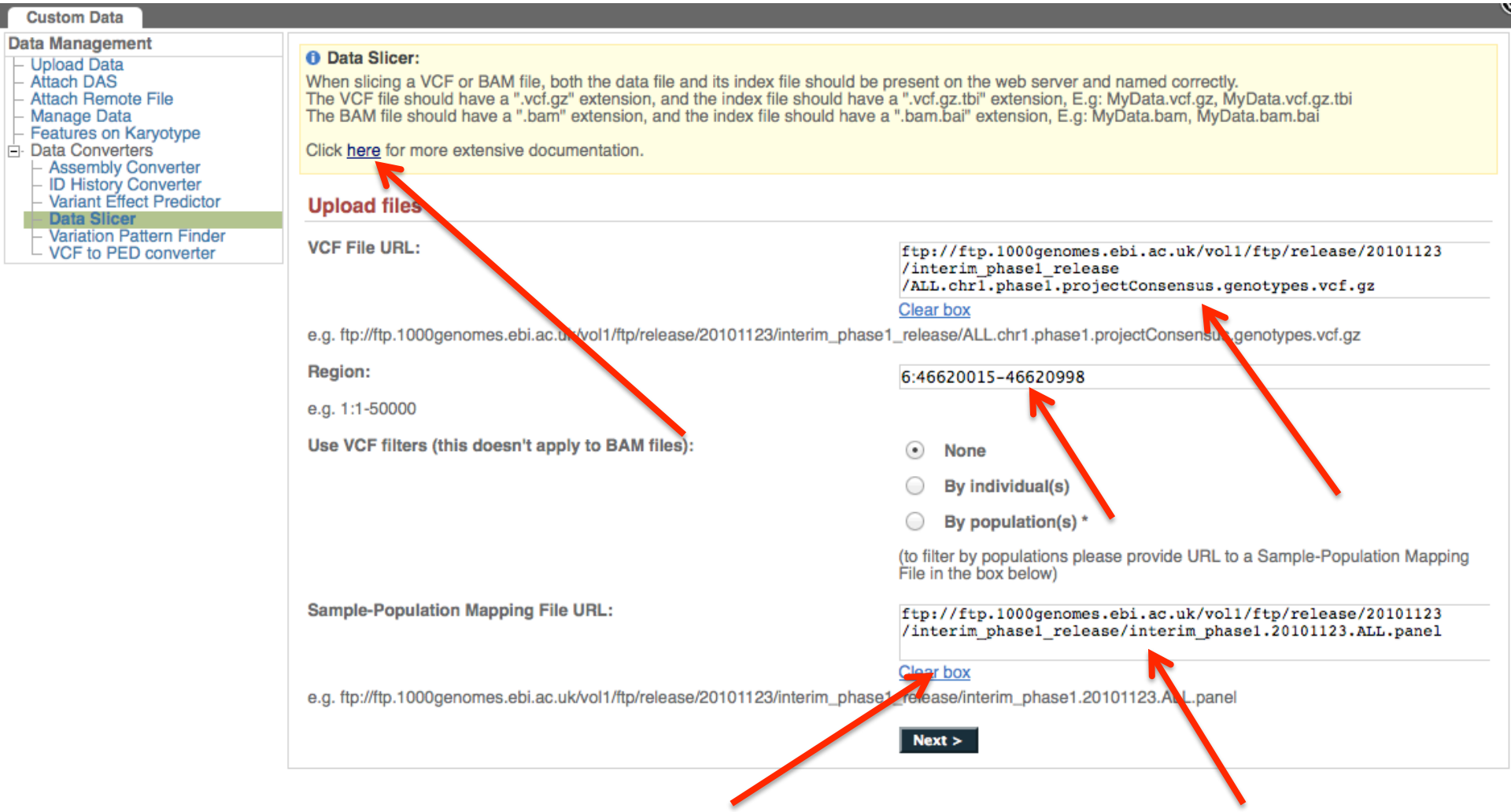
EMBL-EBI

Data Slicer

- Remote Bam or VCF files
- Genomic Location
- Returns subsection of given file
- VCF files can be subset by
  - Population
  - Individual
  - Must provide a panel file to map individual to population

EMBL-EBI

Data Slicing



A THOUSAND GENOMES

EMBL-EBI

Data Slicer Example screens

VCF filter by population(s)

Select one or more populations from the scrollable list:

ASW
CEU
CHB
CHS
CLM
FIN
GBR
IBS
JPT
LWK

< Back Next >

Thank you - your VCF file [filtered_6.31830969-31846823.ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz] [Size: 7529] 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
##source_20120302.1=/nfs/public/rw/ensembl/vcftools/bin/vcf-subset -c HG01112,HG(
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG01112 I


6       31831159        rs3869144       C       T       100     PASS    .       (
```

A THOUSAND GENOMES

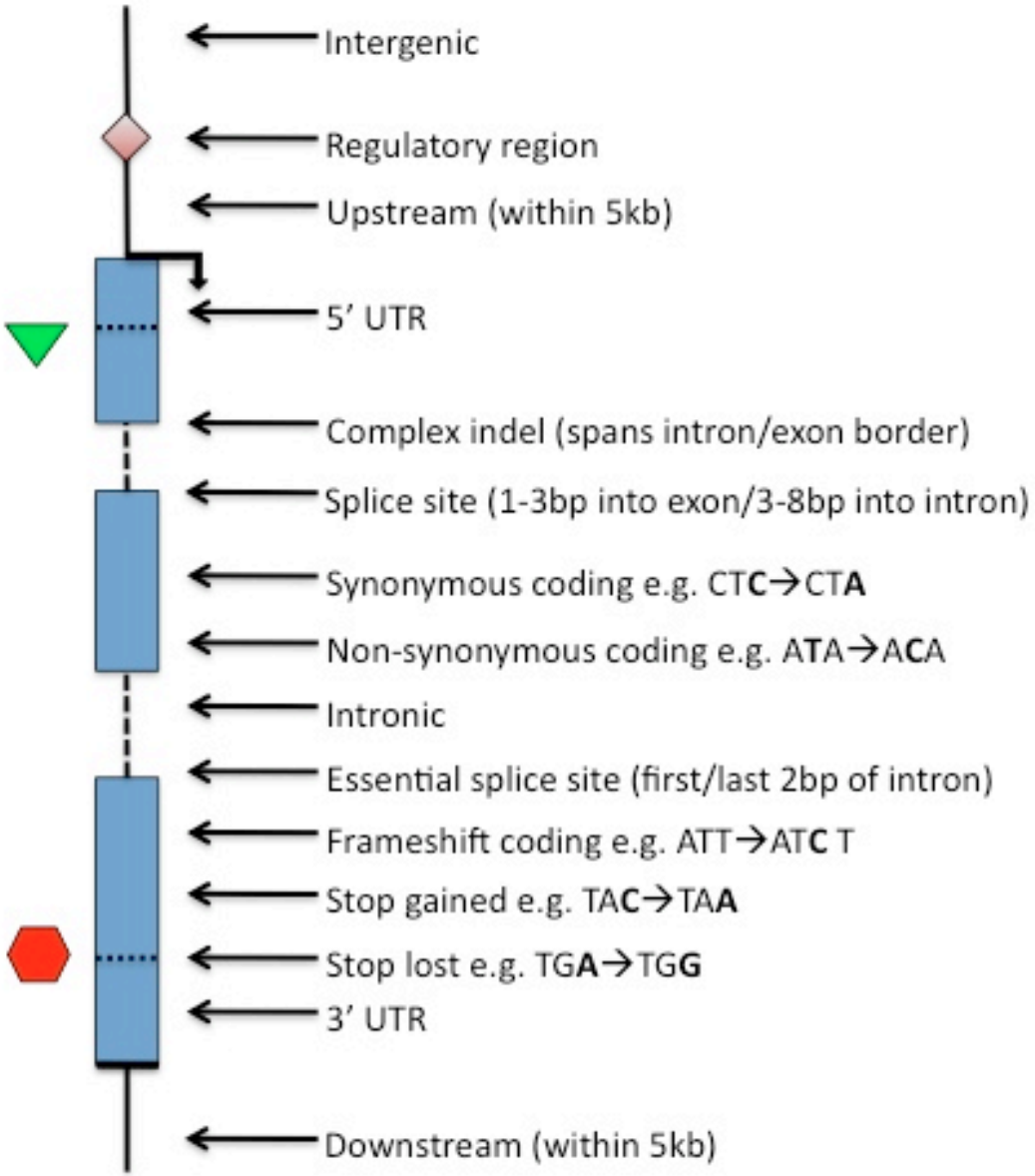
EMBL-EBI

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

EMBL-EBI

Variant Effect Predictor



Others: Within non-coding gene, Within mature miRNA, NMD transcript

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 Effect Predictor Output

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
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000375562	Transcript	NON_SYNONYMOUS_CODING	1544	1478	493
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000544672	Transcript	NON_SYNONYMOUS_CODING	1673	1376	459
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000487680	Transcript	UPSTREAM	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000414427	Transcript	DOWNSTREAM	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000479777	Transcript	DOWNSTREAM	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000475563	Transcript	DOWNSTREAM	-	-	-

0204386	ENST00000491768	Transcript	UPSTREAM	-	-	-	-	-	1KG_6_31833357	-
0204386	ENST00000375631	Transcript	UPSTREAM	-	-	-	-	-	1KG_6_31833357	-
0204386	ENST00000479533	Transcript	UPSTREAM	-	-	-	-	-	1KG_6_31833357	-
0204385	ENST00000229729	Transcript	NON_SYNONYMOUS_CODING	1625	1604	535	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=probably_damaging
0204385	ENST00000375562	Transcript	NON_SYNONYMOUS_CODING	1544	1478	493	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=possibly_damaging
0204385	ENST00000544672	Transcript	NON_SYNONYMOUS_CODING	1673	1376	459	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=probably_damaging
0204385	ENST00000487680	Transcript	UPSTREAM	-	-	-	-	-	1KG_6_31833357	-
0204385	ENST00000414427	Transcript	DOWNSTREAM	-	-	-	-	-	1KG_6_31833357	-
0204385	ENST00000479777	Transcript	DOWNSTREAM	-	-	-	-	-	1KG_6_31833357	-
0204385	ENST00000475563	Transcript	DOWNSTREAM	-	-	-	-	-	1KG_6_31833357	-

EMBL-EBI

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

EMBL-EBI

Variation Pattern Finder

Variation Pattern Finder Output

Variation Pattern Finder

Export data: CSV Excel

Go to collapsed view

CEU	C	Freq		rs12661281:T/A	6:31843711:C/T	6:31845340:C/T	rs2075798:C/A
				6:31842598	6:31843711	6:31845340	6:31846741
			DING:N/S	ENST00000229729 NON_SYNONYMOUS_CODING:D/V	ENST00000229729 SPLICE_SITE	ENST00000544672 SPLICE_SITE	ENST00000229729 NON_SYNONYMOU
			DING:N/S	ENST00000544672 NON_SYNONYMOUS_CODING:D/V	ENST00000375562 SPLICE_SITE	ENST00000544672 5PRIME_UTR	ENST00000375562 NON_SYNONYMOU
			DING:N/S	ENST00000414427 NON_SYNONYMOUS_CODING:D/V	ENST00000544672 SPLICE_SITE		ENST00000414427 NON_SYNONYMOU
					ENST00000414427 SPLICE_SITE		
					ENST00000465707 SPLICE_SITE		
					ENST00000462671 SPLICE_SITE		
NA12872, NA07000 and 1 other(s)	N	0.032		TIA	CIC	CIC	CIC
NA12874, NA12717	N	0.028		TIT	CIC	CIC	AIC
NA07346	N	0.027		TIT	CIC	CIC	CIA
	N	0.027		TIT	CIC	CIC	CIC
NA10851, NA12342 and 5 other(s)	N	0.024		AIT	CIC	CIC	CIC
NA12058, NA12273 and 1 other(s)	N	0.020		AIA	CIC	CIC	CIC
	N	0.018		TIT	CIC	CIC	CIC
	N	0.015		AIT	CIC	CIC	CIA
	N	0.014		TIT	CIC	CIC	AIA
	N	0.013		TIT	CIC	CIC	CIC
NA10847	N	0.011		TIA	CIC	CIC	AIC
NA12286, NA11892 and 2 other(s)	N	0.009		TIT	CIC	CIC	CIC

A THOUSAND GENOMES

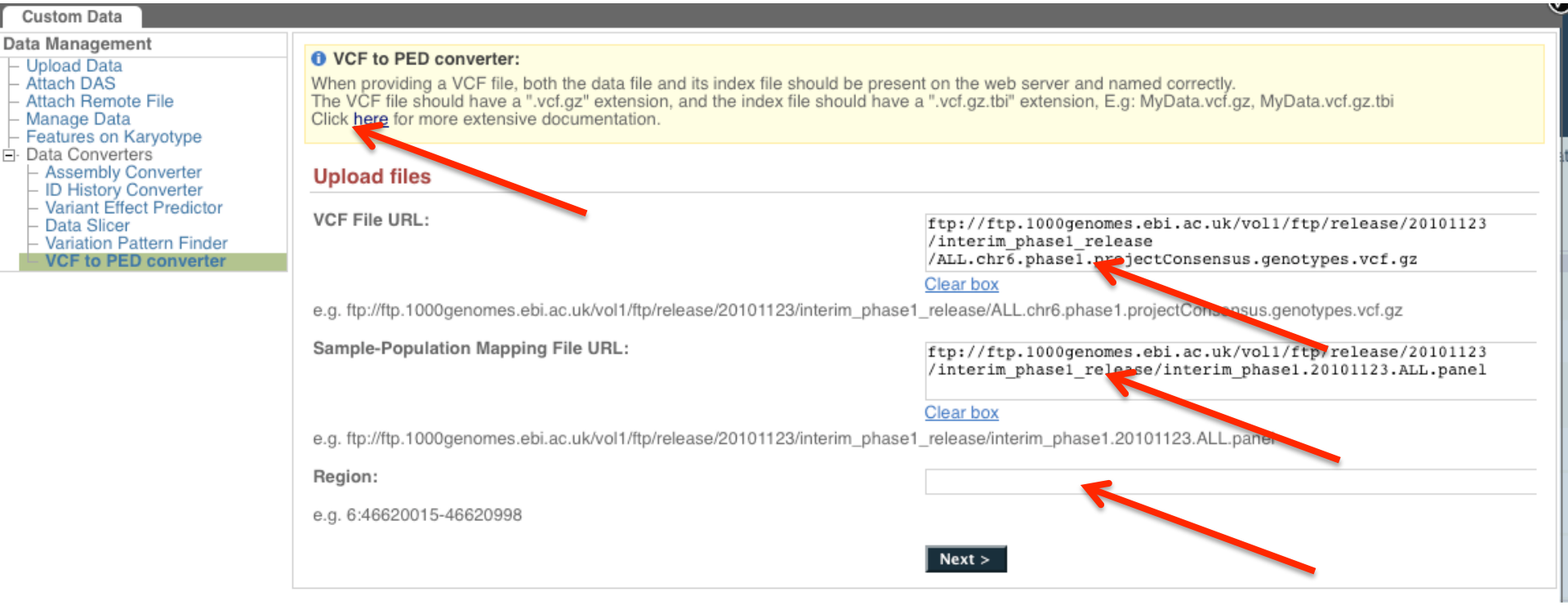
EMBL-EBI

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

EMBL-EBI

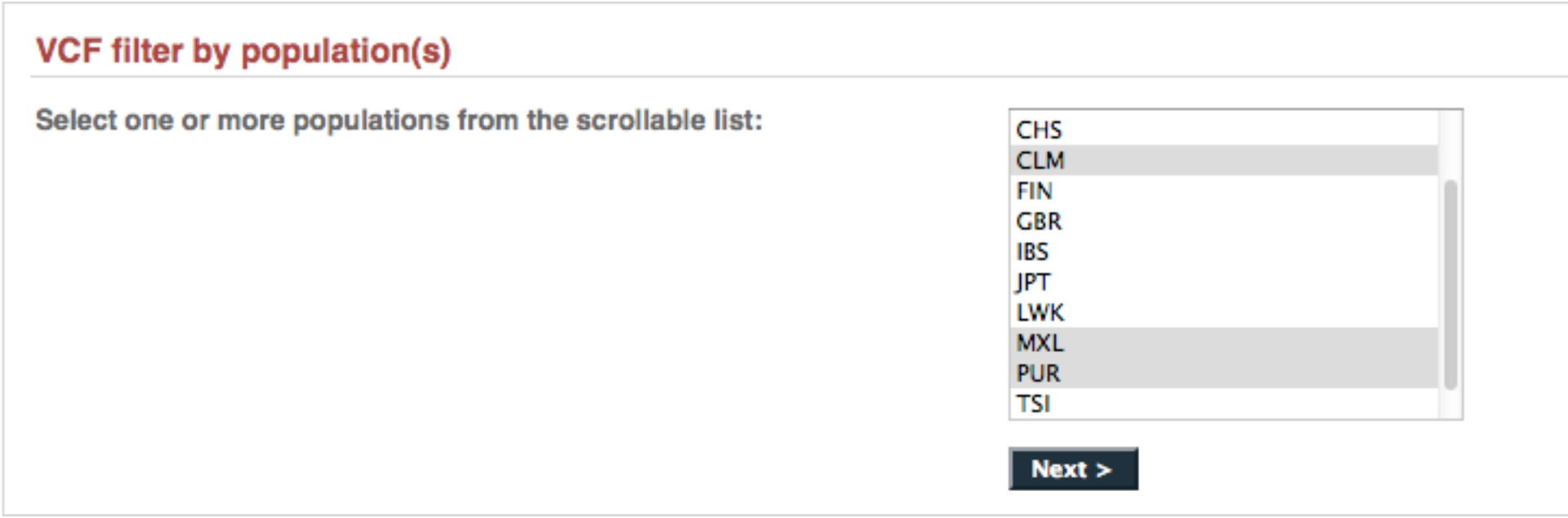
VCF to PED



A THOUSAND GENOMES

EMBL-EBI

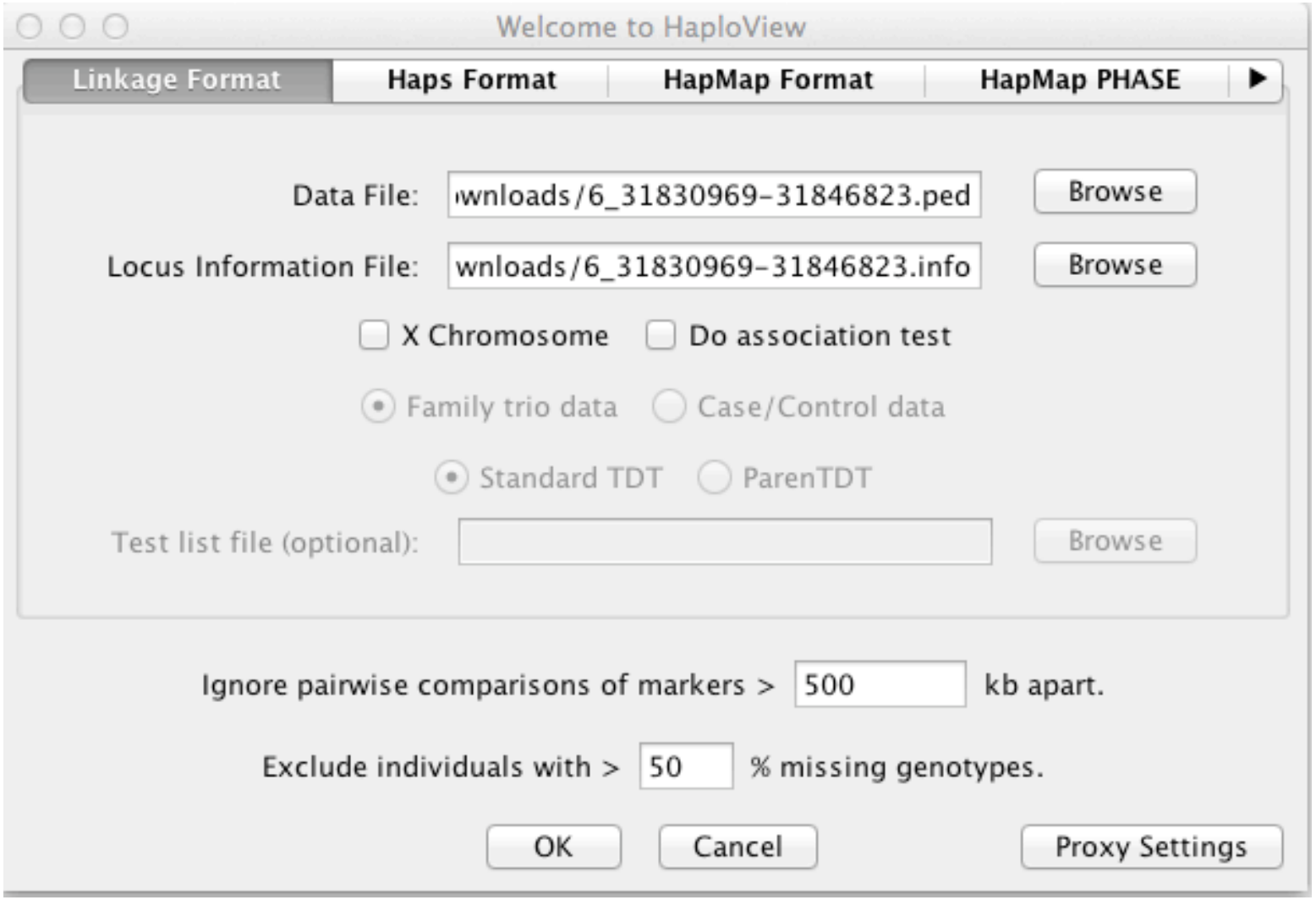
VCF to PED example output



Your linkage pedigree and marker information files have been generated:
Right click on the file name and choose "Save link as .." from the menu:
Marker Information File Linkage Pedigree File

Haplotype example input

java –jar Haploview.jar



Haploview

- haploview

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

Exercises

Use the browser to find the SLC44A4 gene.

7. Use the get VCF button in the left hand menu on the gene page to get a slice of a vcf file for this Gene.

8. Unzip this VCF file using a tool like winzip or Archive Utility.

9. Upload this VCF file to the Variant Effect Predictor.
http://browser.1000genomes.org/Homo_sapiens/UserData/UploadVariations

10. Do any of the variants have negative Sift or Polyphen predictions?

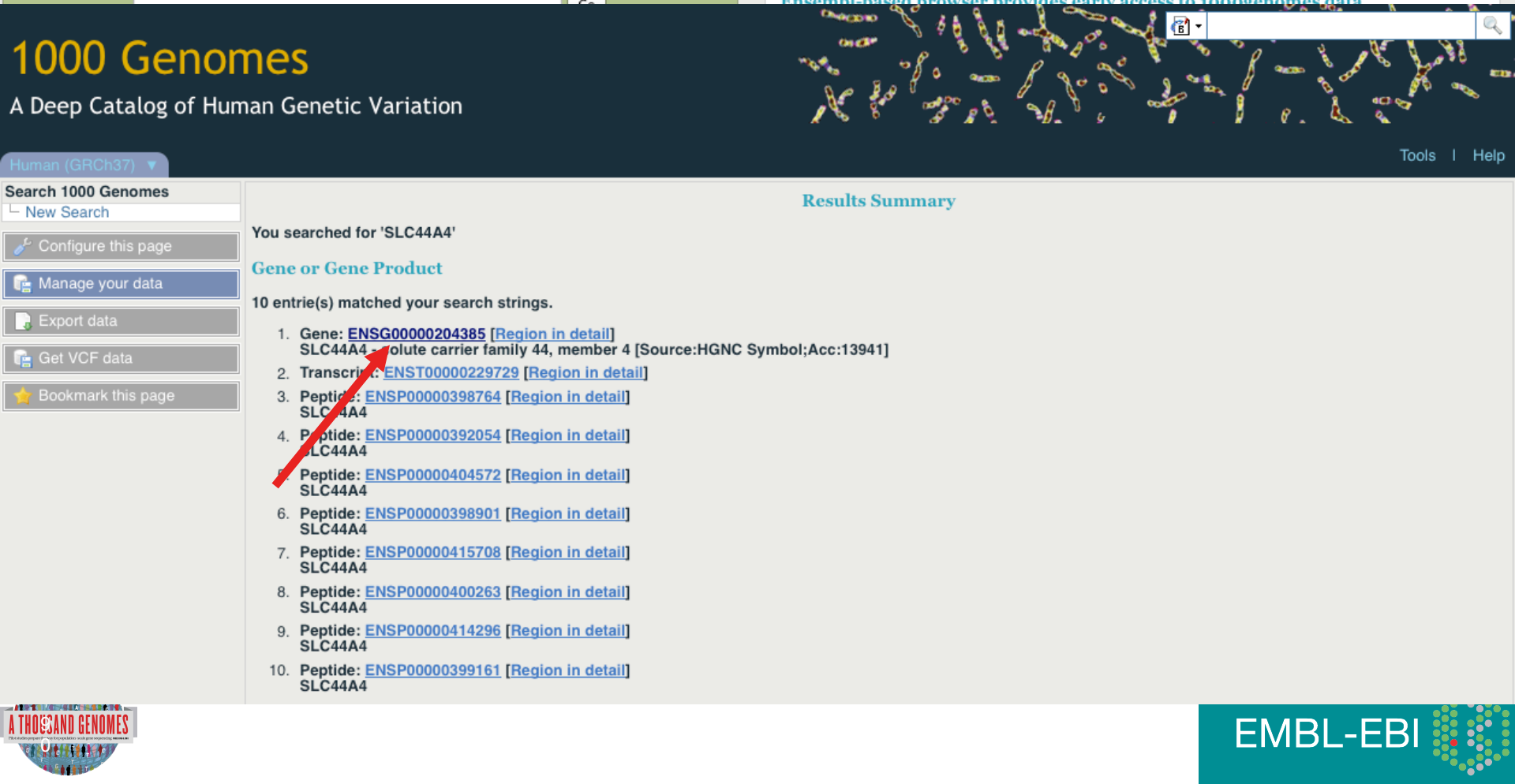
11. Using the example URLs on the Variation Pattern Finder tool menu look at the patterns of inheritance for this region: 6:31830700-31840700
http://browser.1000genomes.org/Homo_sapiens/UserData/VariationsMapVCF

12. For the same region use the VCF to PED tool to produce a ped and info file for the CEU population.

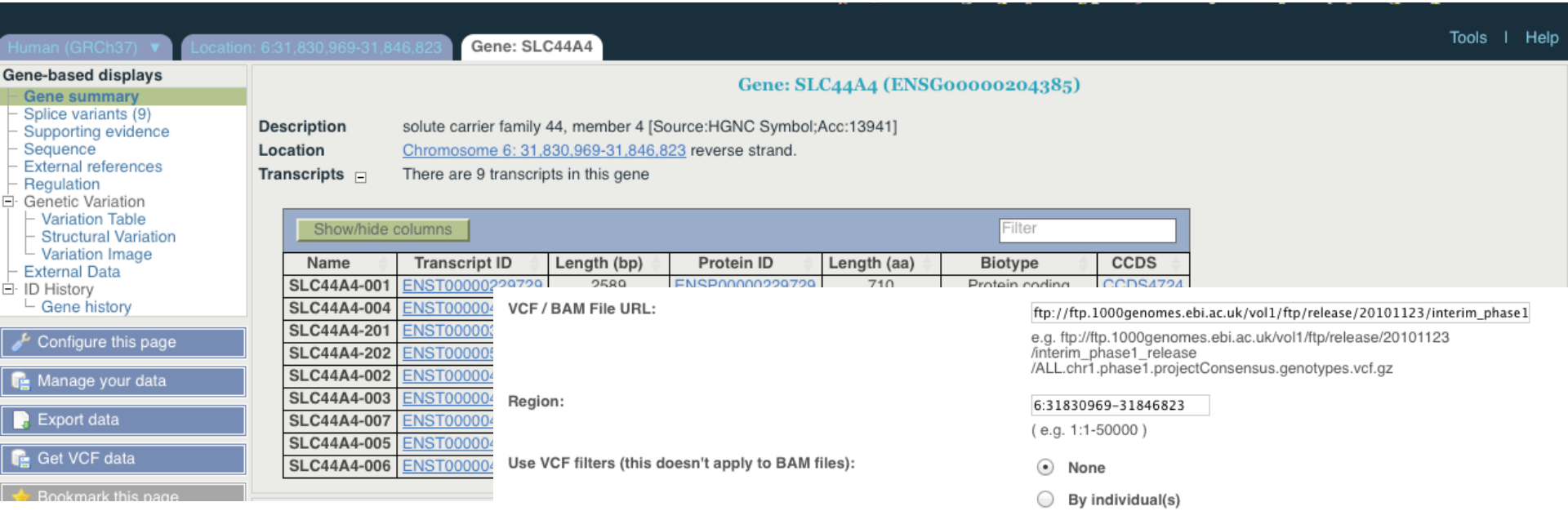
13. Look at these files in haploview.

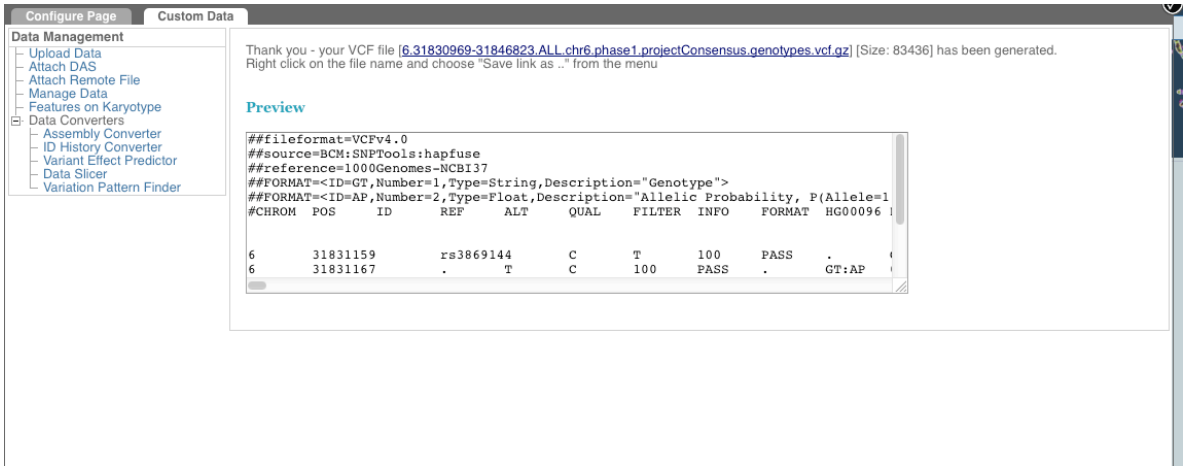
14. How many haplotype blocks does haploview think there are in this section?

Exercise Answers



Exercise Answers

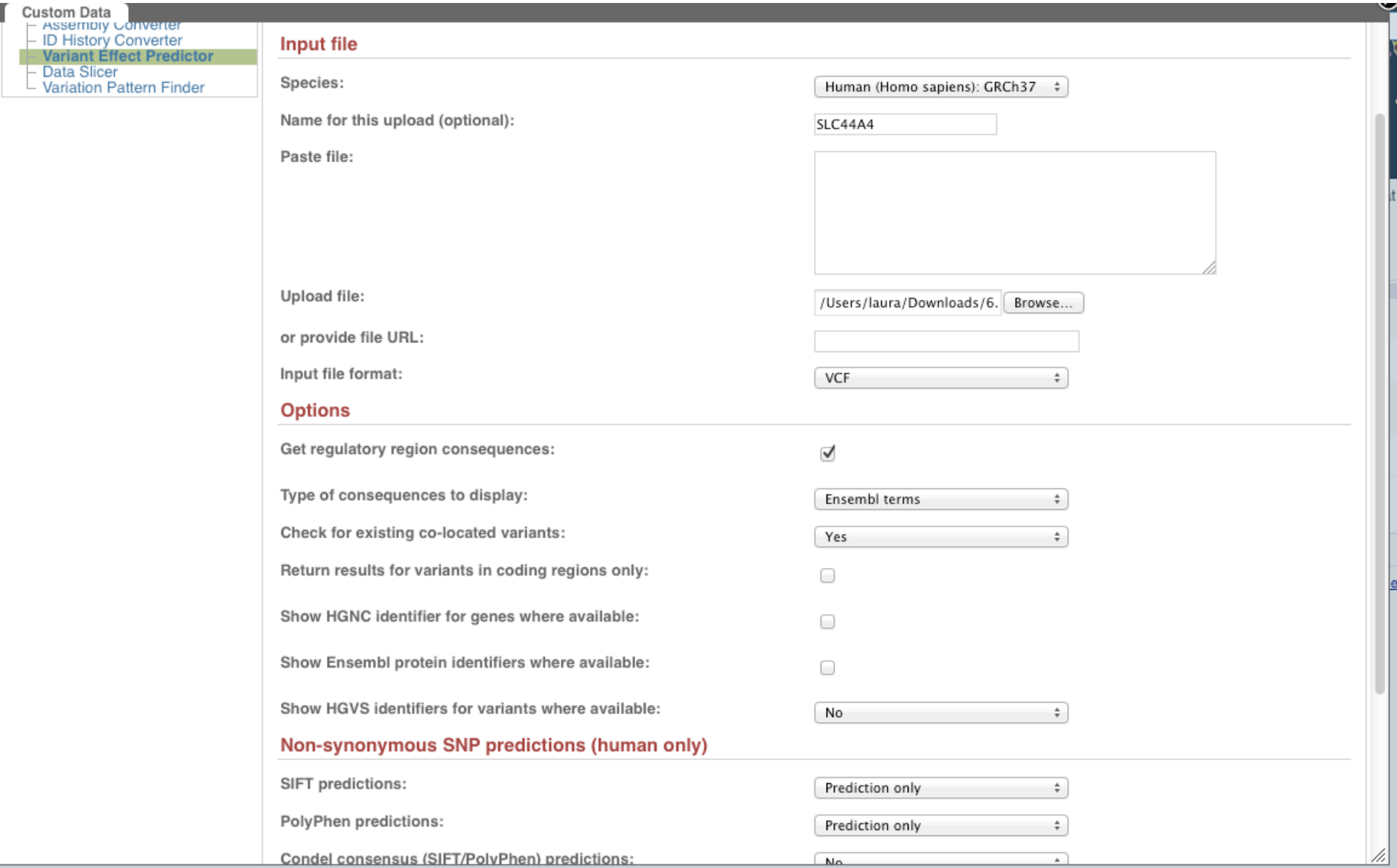




A THOUSAND GENOMES

EMBL-EBI

Exercise Answers



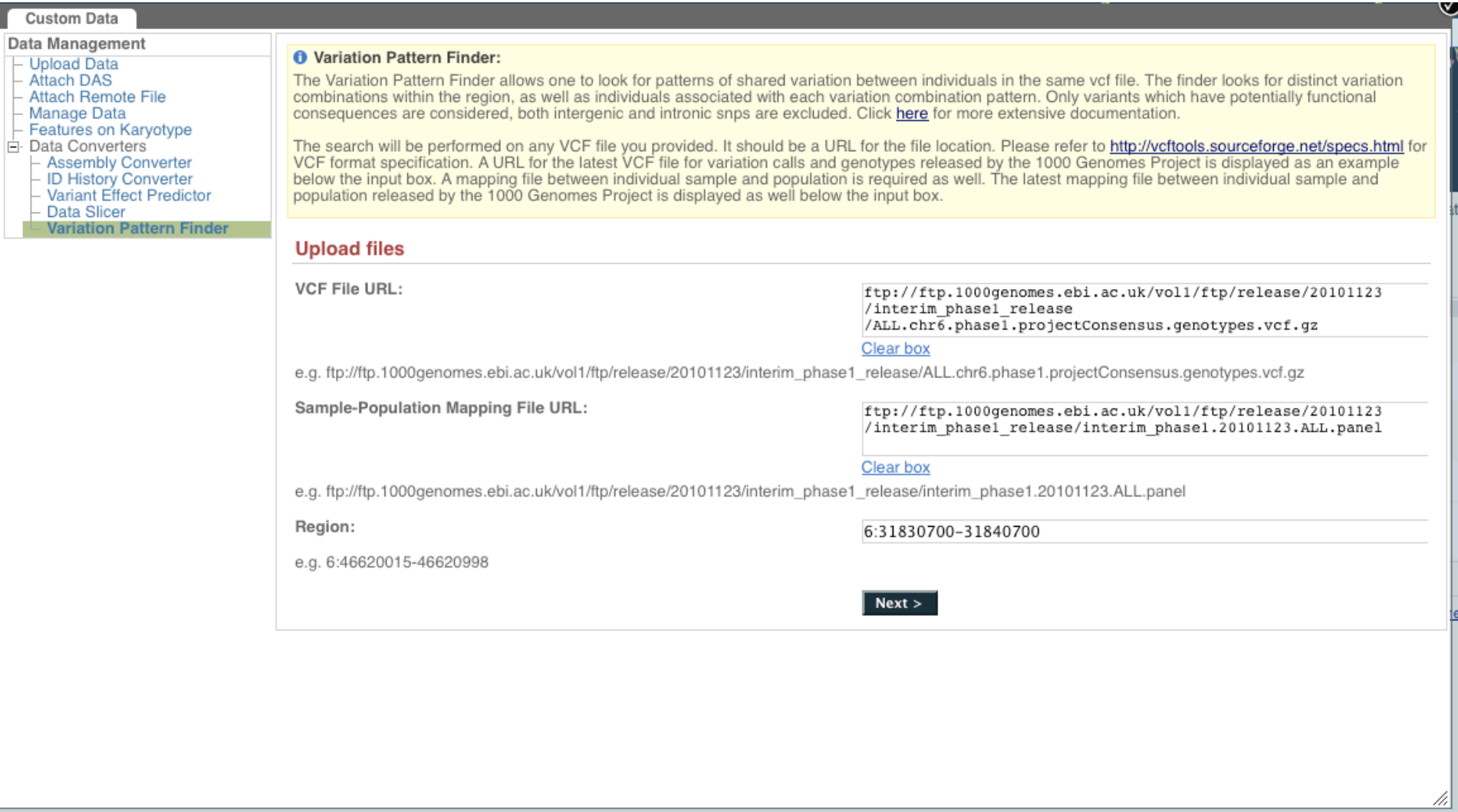
Exercise Answers

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	ENSG00	-	-	-	1KG_6_31833357	-
6_31833357_C/T	6:31833357	T	ENSG00					
6_31833357_C/T	6:31833357	T	ENSG00	-	-	-	1KG_6_31833357	-
6_31833357_C/T	6:31833357	T	ENSG00	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	ENSG00					
6_31833612_C/G	6:31833612	G	ENSG00	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	-

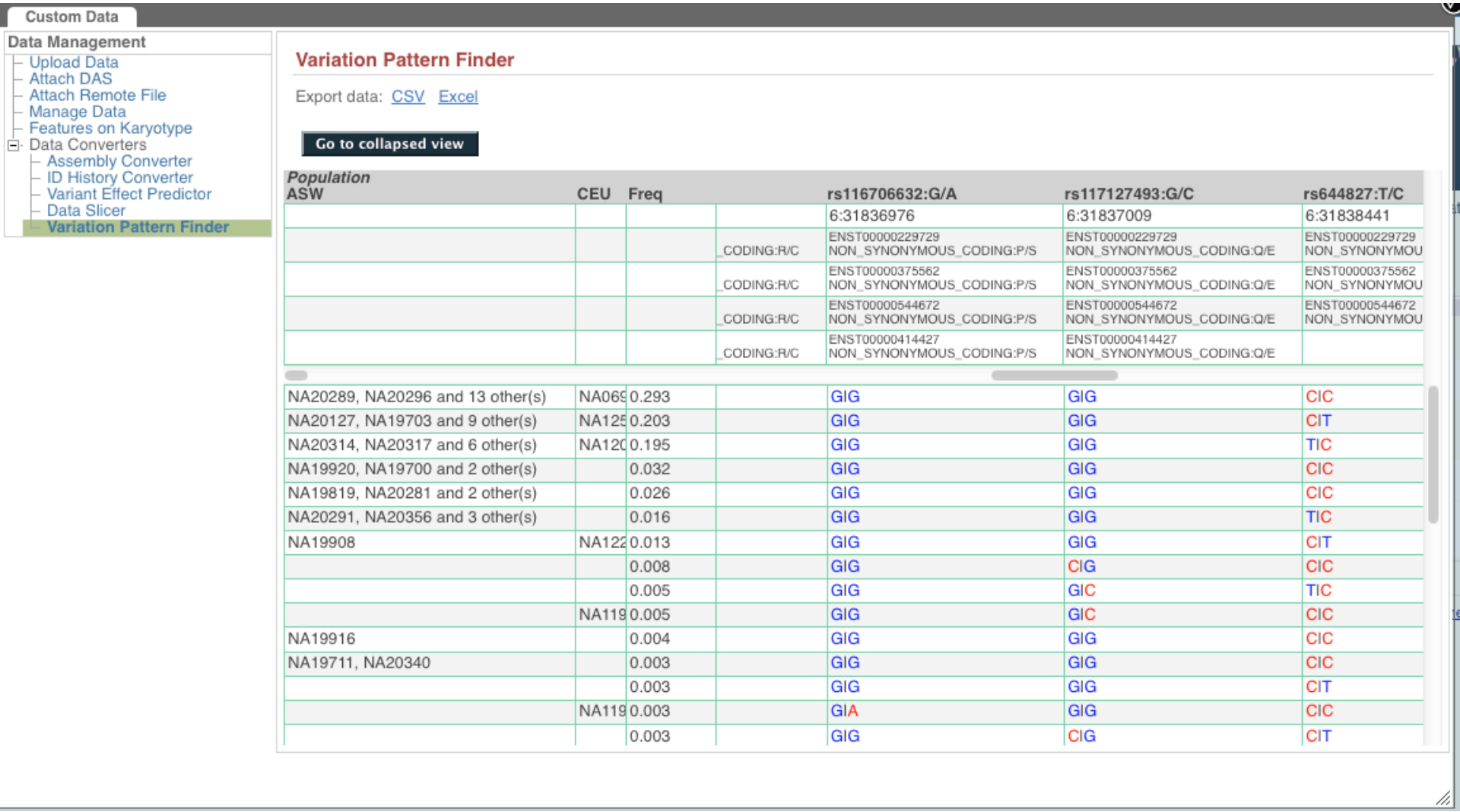
EMBL-EBI

Exercise Answers



EMBL-EBI

Exercise Answers



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		rs116706632:G/A	rs117127493:G/C	rs644827:T/C
				6:31836976	6:31837009	6:31838441
			_CODING:R/C	ENST00000229729 NON_SYNONYMOUS_CODING:P/S	ENST00000229729 NON_SYNONYMOUS_CODING:Q/E	ENST00000229729 NON_SYNONYMOU
			_CODING:R/C	ENST00000375562 NON_SYNONYMOUS_CODING:P/S	ENST00000375562 NON_SYNONYMOUS_CODING:Q/E	ENST00000375562 NON_SYNONYMOU
			_CODING:R/C	ENST00000544672 NON_SYNONYMOUS_CODING:P/S	ENST00000544672 NON_SYNONYMOUS_CODING:Q/E	ENST00000544672 NON_SYNONYMOU
			_CODING:R/C	ENST00000414427 NON_SYNONYMOUS_CODING:P/S	ENST00000414427 NON_SYNONYMOUS_CODING:Q/E	
NA20289, NA20296 and 13 other(s)	NA069	0.293		GIG	GIG	CIC
NA20127, NA19703 and 9 other(s)	NA125	0.203		GIG	GIG	CIT
NA20314, NA20317 and 6 other(s)	NA120	0.195		GIG	GIG	TIC
NA19920, NA19700 and 2 other(s)		0.032		GIG	GIG	CIC
NA19819, NA20281 and 2 other(s)		0.026		GIG	GIG	CIC
NA20291, NA20356 and 3 other(s)		0.016		GIG	GIG	TIC
NA19908	NA122	0.013		GIG	GIG	CIT
		0.008		GIG	CIG	CIC
		0.005		GIG	GIC	TIC
	NA119	0.005		GIG	GIC	CIC
NA19916		0.004		GIG	GIG	CIC
NA19711, NA20340		0.003		GIG	GIG	CIC
		0.003		GIG	GIG	CIT
	NA119	0.003		GIA	GIG	CIC
		0.003		GIG	CIG	CIT

A THOUSAND GENOMES

EMBL-EBI

Exercise Answers

VCF to PED converter:

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

Upload files

VCF File URL:

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

Clear box

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

Sample-Population Mapping File URL:

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

Clear box

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

Region:

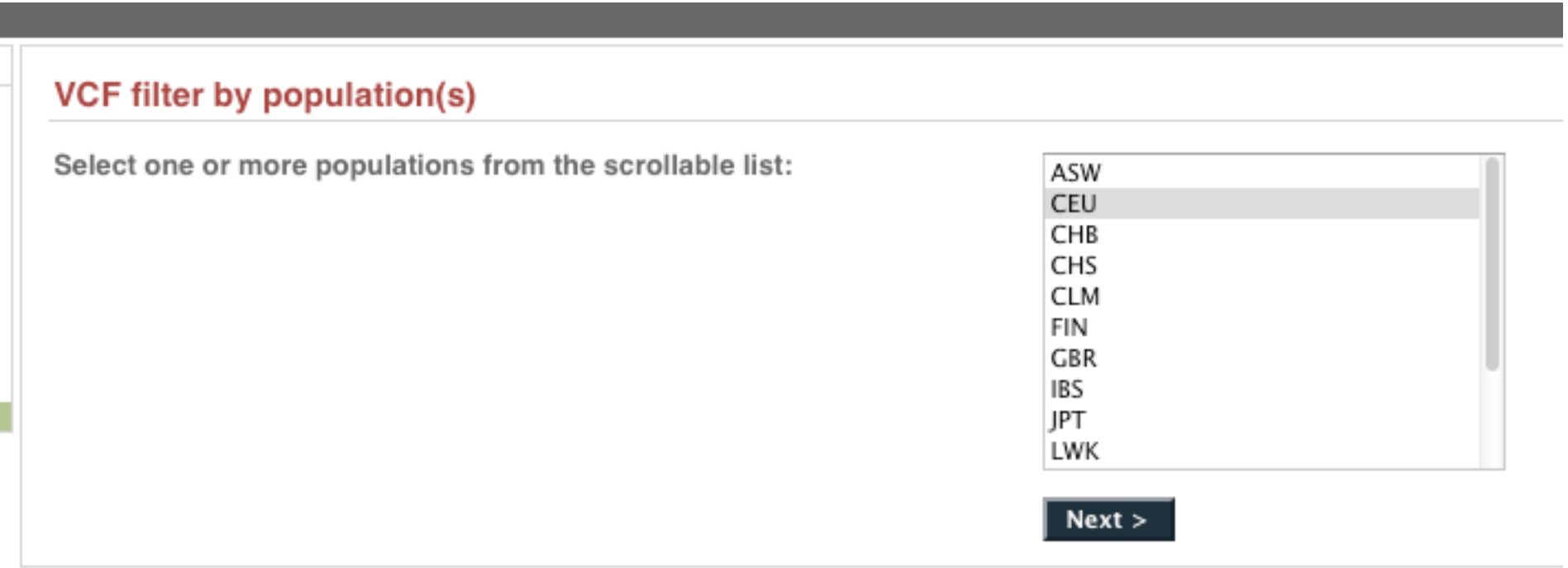
```
6:31830700-31840700
```

e.g. 6:46620015-46620998

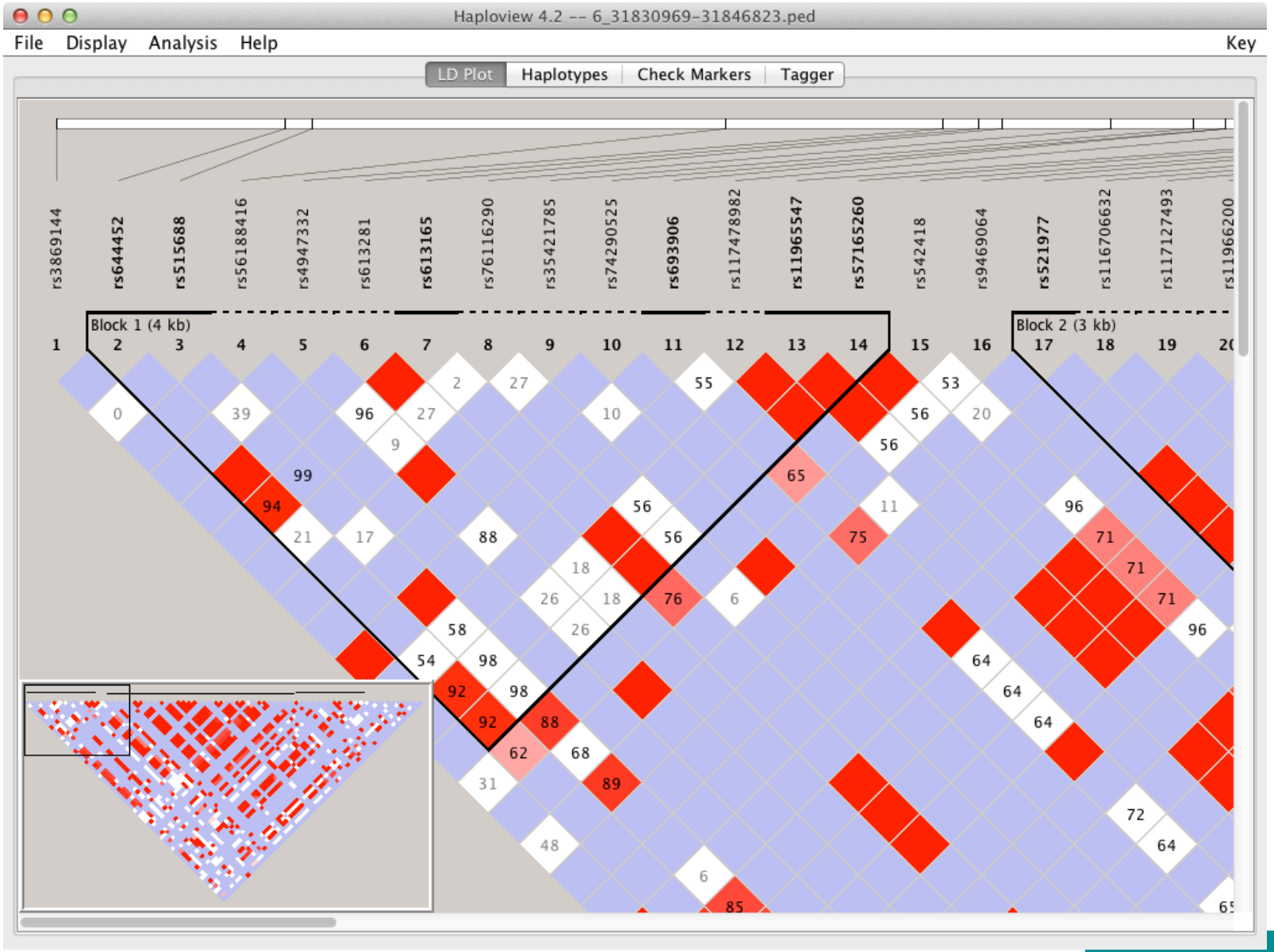
Next >

EMBL-EBI

Exercise Answers

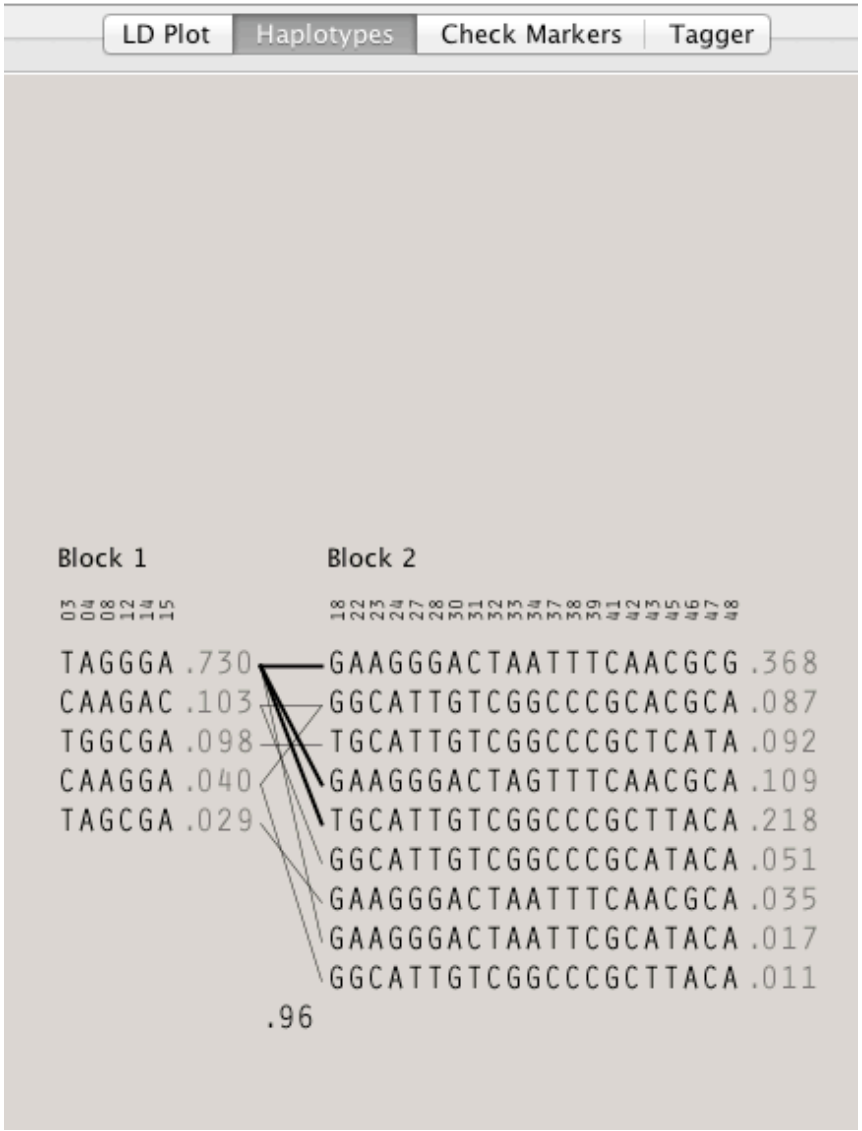


Exercise Answers



A THOUSAND GENOMES
EMBL-EBI

Exercise Answers



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

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

EMBL-EBI

Questions

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

EMBL-EBI

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

EMBL-EBI

1000 Genomes Community Meeting

- University of Michigan, Ann Arbor on the 12th and 13th of July 2012
- Showcase Advances made by the Project
- Generate Discussion about the next round of Human Genome Sequencing
- Registration closes May 15th
- http://1000gconference.sph.umich.edu/

EMBL-EBI

Thanks

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

A THOUSAND GENOMES

EMBL-EBI