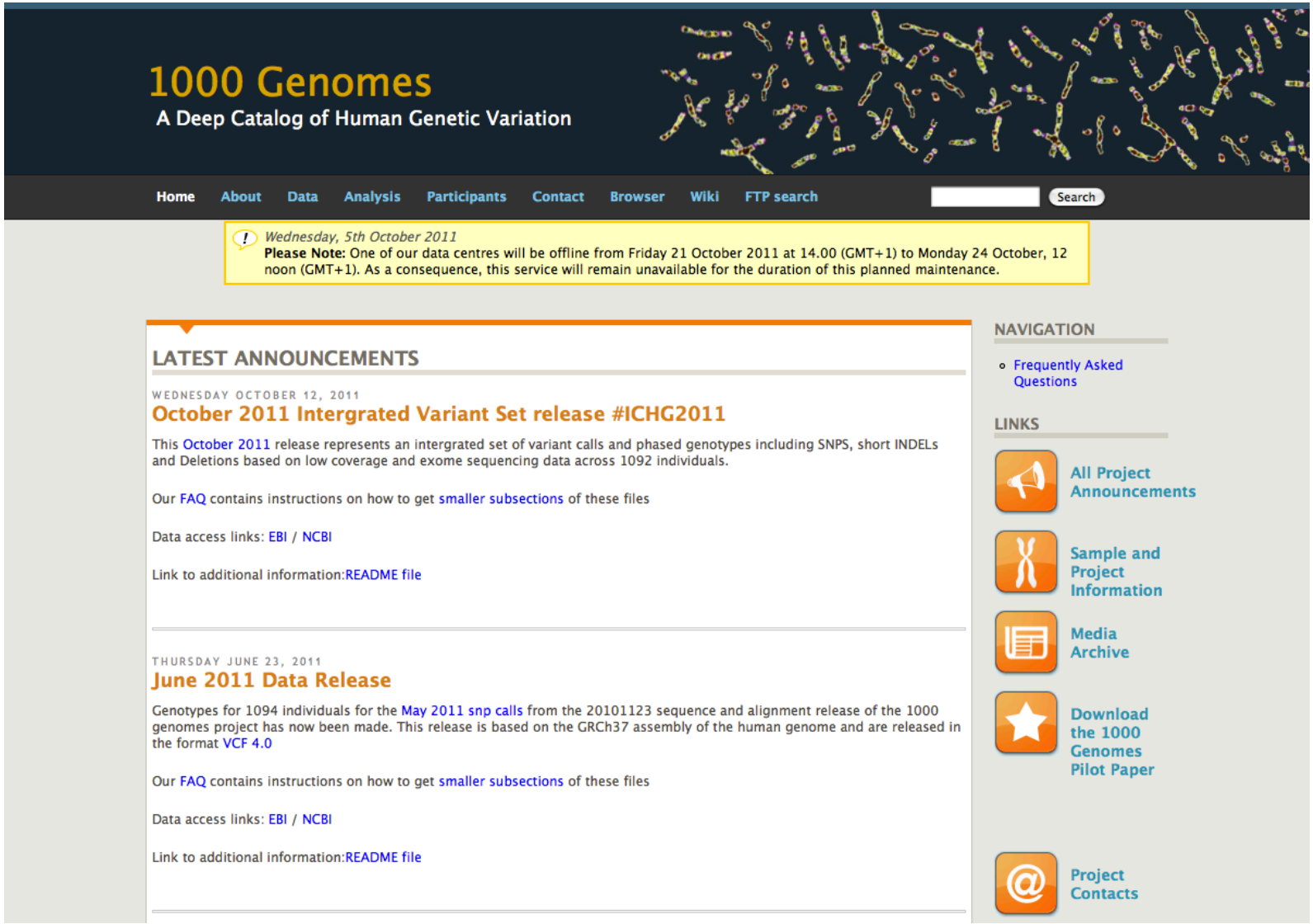
Accessing the 1000 Genomes Data

Paul Flicek
European Bioinformatics Institute

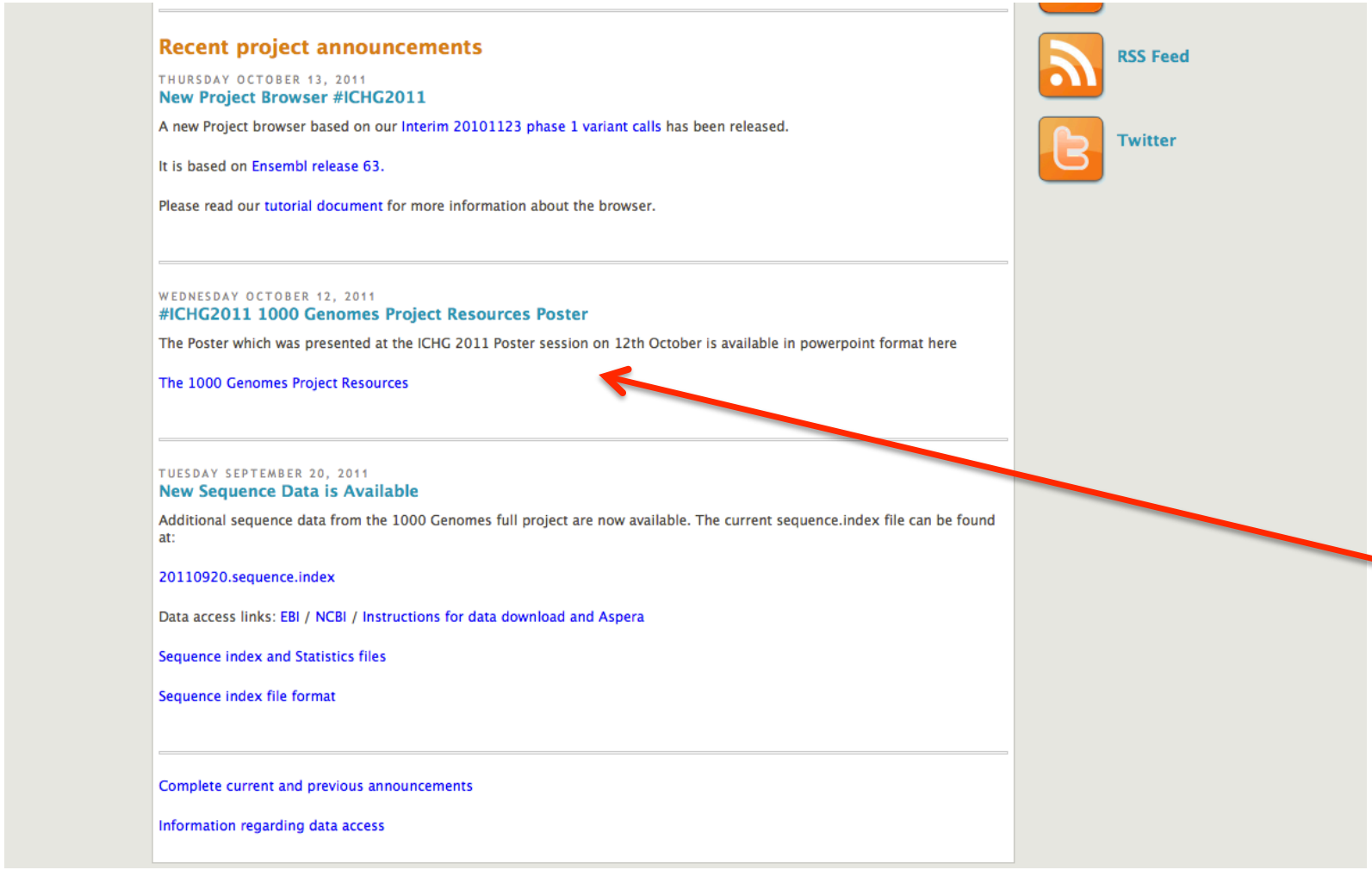
Data access

- General information
- File access
- 1000 Genomes Browser
- Tools
- Where to find help

www.1000genomes.org



www.1000genomes.org



EMBL-EBI

European Bioinformatics Institute
is an Outstation of the
European Molecular Biology Laboratory.

1000 Genomes Project Resources

L. Clarke, H. Zheng-Bradley, R. Smith, E Kuleshea, I Toneva, B. Vaughan, P. Flicek and 1000 Genomes Consortium
European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Introduction

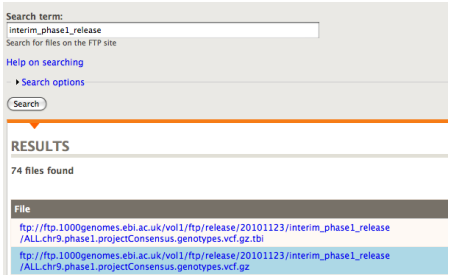
The main goal of the 1000 genomes project is to establish a comprehensive and detailed catalogue of human genome variations; which in turn will empower association studies to identify disease-causing genes. The project now has data and variant genotypes for more than 1000 individuals in 14 populations. The ftp site contains more than 120Tbytes of data in 200,000 files.

DATA TYPE	FILE FORMAT	SIZE
sequence	FASTQ	43 Tbases raw sequence
alignment	BAM	56 Tbytes of BAM files
variants	VCF	38.9M SNPs ~4.7M short indels

Discoverability

Sequence, alignment and variant data is made available as quickly as possible through the project ftp site. (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/ | ftp://ftp-trace.ncbi.nih.gov/1000genomes/). With more than 200,000 files though discovering new data can be difficult.

The ftp site has a index updated nightly. This index is searchable from our website. http://www.1000genome.org/ftpsearch



The search allows users to specify which ftp site to get paths to, to get md5 checksums and also filter out high volume results like bam and fastq files

We also have various routes for users to discover new data.

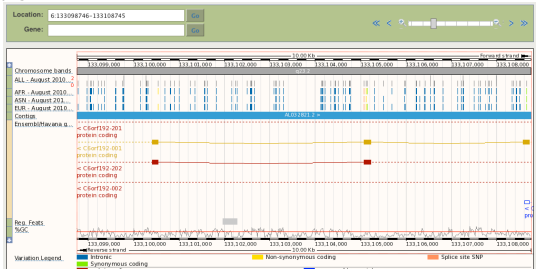
- Website http://www.1000genomes.org/announcements
- Twitter @1000genomes
- RSS http://www.1000genomes.org/announcements/rss.xml
- Email 1000announce@1000genomes.org

Visualization

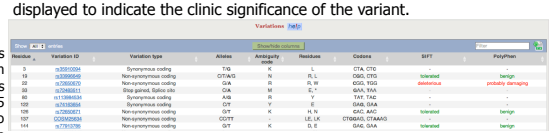
http://browser.1000genomes.org

The 1000 Genomes project utilizes the Ensembl Browser to display our variant calls. We provide rapid access to project variant calls through the browser before they become available via dbSNP and DGVa.

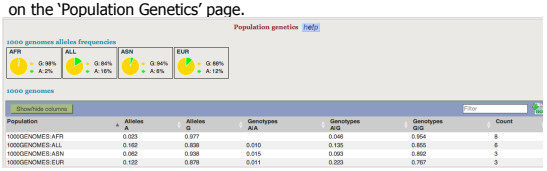
Tracks of 1000 genomes variants by population can be viewed in the location page:



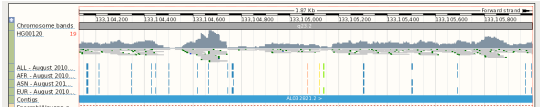
A list of variants can be obtained for any given transcript. In addition to basic information about a variant, PolyPhen and SIFT annotation are displayed to indicate the clinic significance of the variant.



Allele frequency for individual variants in different populations is displayed on the 'Population Genetics' page.



Users can Attach remote files as custom tracks. In example below, the HG00120 track is 1000 Genomes bam file added to the browser.



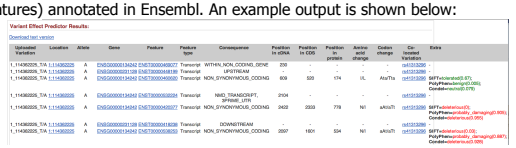
Accessibility

http://browser.1000genomes.org/tools.html

The project provides several tools to help users access and interpret the data provided.

Variant Effect Predictor

The predictor takes a list of variant positions and alleles, and predicts the effects of each of these on any overlapping features (transcripts, regulatory features) annotated in Ensembl. An example output is shown below:

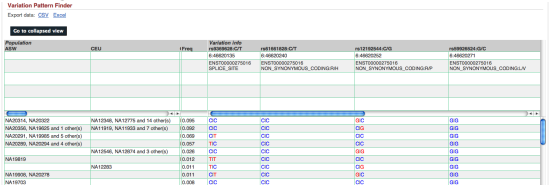


Data Slicer

Many of the 1000 Genomes files are large and cumbersome to handle. The Data Slicer allows users to get data for specific regions of the genome and to avoid having to download many gigabytes of data they don't needl samples/ populations you choose. Below is the Data Slicer input interface:

Variation Pattern Finder

- The Variation Pattern Finder (VPF) allows one to look for patterns of shared variation between individuals in the a VCF file.
- Within a vcf file different samples have different combination of variation genotypes. The VPF looks for distinct variation combinations within a user specifed region, shared by different individuals.
- The VPF only on variations that functional consequences for protein coding genes such as non-synonymous coding SNPs and splice site changes.



Acknowledgements

We would like to thank the Ensembl variation team for all their help, particularly Will McLaren and Graham Ritchie.
Funding: The Wellcome Trust

Laura Clarke
EBI
laura@ebi.ac.uk

EMBL- EBI
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD
UK

1000 Genomes
A Deep Catalog of Human Genetic Variation

1000 Genomes

A Deep Catalog of Human Genetic Variation

1000 GENOMES DATA AND SAMPLE INFORMATION

The 1000 Genomes Project is a community resource project that aims to release data rapidly for the benefit of the scientific community.

Description of data released by the project
How to Access 1000 Genomes Data
Data Release Policy
Sample Availability
Use of the Project data, presentations and publications, and authorship

DATA RELEASED BY THE 1000 GENOMES PROJECT

Sample lists and sequencing progress

A summary of sequencing done for each of the three pilot projects is available here. The list of samples and allocations is provided in a spreadsheet.

Variant Calls

The pilot variant calls are available in vcf format from EBI|NCBI

Alignments

The main project alignments are available in BAM format. A list of the files currently available can be found in the alignment index EBI|NCBI. Alignment statistics can be found in the alignment_indices directory EBI|NCBI. There is also a README which explains the alignment process and file layout

Raw sequence files

The main project raw sequence data is available in fastq format. A list of files currently available can be found in the sequence.index EBI|NCBI Sequence statistics can be found in the sequence_indices directory EBI|NCBI. There is also a README which explains the sequence processing and the file layout

NAVIGATION

- Frequently Asked Questions

LINKS

All project announcements

Files and formats

Software tools

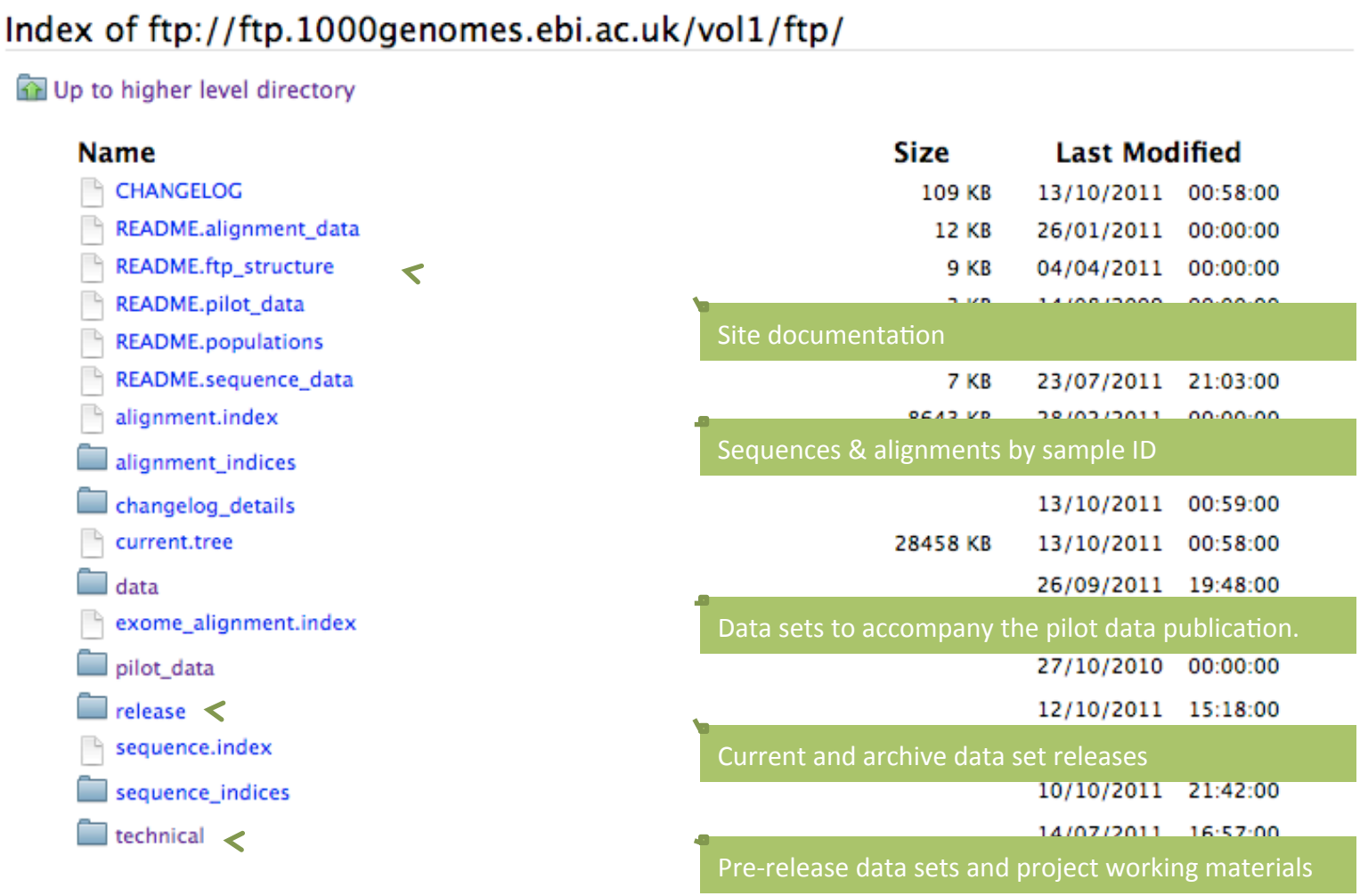
Download the 1000 Genomes Pilot Paper

Project Contacts

Data access

- General information
- File access
- 1000 Genomes Browser
- Tools
- Where to find help

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



KEY FILE FORMATS

Information on key file formants is normally provided within README files located near the relevant data on the FTP site. Information on the major formats is also collected here. Tools developed by the 1000 Genomes project or appropirate to work with the data from the project are listed on the tools page.

Index file formats
Data file formats

INDEX FILE FORMATS

Sequence index

The sequence.index file is a tab delimited file containing all the meta data you should need to download and subset the files on this ftp site by individual, library, experiment and sequencing technology.

The columns are:

1. FASTQ_FILE, path to fastq file on ftp site
2. MD5, md5sum of file
3. RUN_ID, SRA/ERA run accession
4. STUDY_ID, SRA/ERA study accession
5. STUDY_NAME, Name of stury
6. CENTER_NAME, Submission centre name
7. SUBMISSION_ID, SRA/ERA submission accession
8. SUBMISSION_DATE, Date sequence submitted, YYYY-MM-DAY
9. SAMPLE_ID, SRA/ERA sample accession
10. SAMPLE_NAME, Sample name
11. POPULATION, Sample population
12. EXPERIMENT_ID, Experiment accession
13. INSTRUMENT_PLATFORM, Type of sequencing machine
14. INSTRUMENT_MODEL, Model of sequencing machine
15. LIBRARY_NAME, Library name
16. RUN_NAME, Name of machine run
17. RUN_BLOCK_NAME, Name of machine run sector
18. INSERT_SIZE, Submitter specifed insert size
19. LIBRARY_LAYOUT, Library layout, this can be either PAIRED or SINGLE
20. PAIRED_FASTQ, Name of mate pair file if exists (Runs with failed mates will have a library layout of PAIRED but no paired fastq file)
21. WITHDRAWN, 0/1 to indicate if the file has been withdrawn, only present if a file has been withdrawn
22. WITHDRAWN_DATE, date of withdrawal, this should only be defined if a file is withdrawn
23. COMMENT, comment about reason for withdrawal
24. READ_COUNT, read count for the file
25. BASE_COUNT, basepair count for the file
26. ANALYSIS_GROUP, the analysis group of the sequence, this reflects sequencinq strategy. Current this includes low coverage, high coverage and exon targetted to reflect the 3 stratergies used by the 1000 genomes project.

Any run_id can have up to 3 files associated with it. Single runs have one file. Paired runs can have anywhere from 1 to 3 files depending on the success of the pairing.

NAVIGATION

- Frequently Asked Questions

LINKS

All project announcements
Files and formats
Software tools
Download the 1000 Genomes Pilot Paper
Project Contacts

Data formats and key tools

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

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, compact in size, efficient in random access and is the format in which alignments from the 1000 Genomes Project are released. SAMtools implements various utilities for post-processing alignments in the SAM format, such as indexing, variant caller and alignment viewer,

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 are TAB delimited. An example is shown in Figure 1b.

In SAM, each alignment line has 11 mandatory fields and a variable number of optional fields. The mandatory fields are briefly described in Table 1. They must be present but their value can be a '*' or a zero (depending

BIOINFORMATICS APPLICATIONS NOTE

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

Sequence analysis

Advance Access publication June 7, 2011

VCF variant files

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

All indexed for fast retrieval

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 compression with gzip compatibility and direct FTP/HTTP access. Tabix is implemented as a free command-line tool as well as a library in C, Java, Perl and Python. It is particularly useful for manually examining local genomic features on the command line and enables

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.

2.1 Sorting and BGZF compression

Before being indexed, the data file needs to be sorted first by sequence name and then by leftmost coordinate, which can be done with the standard Unix

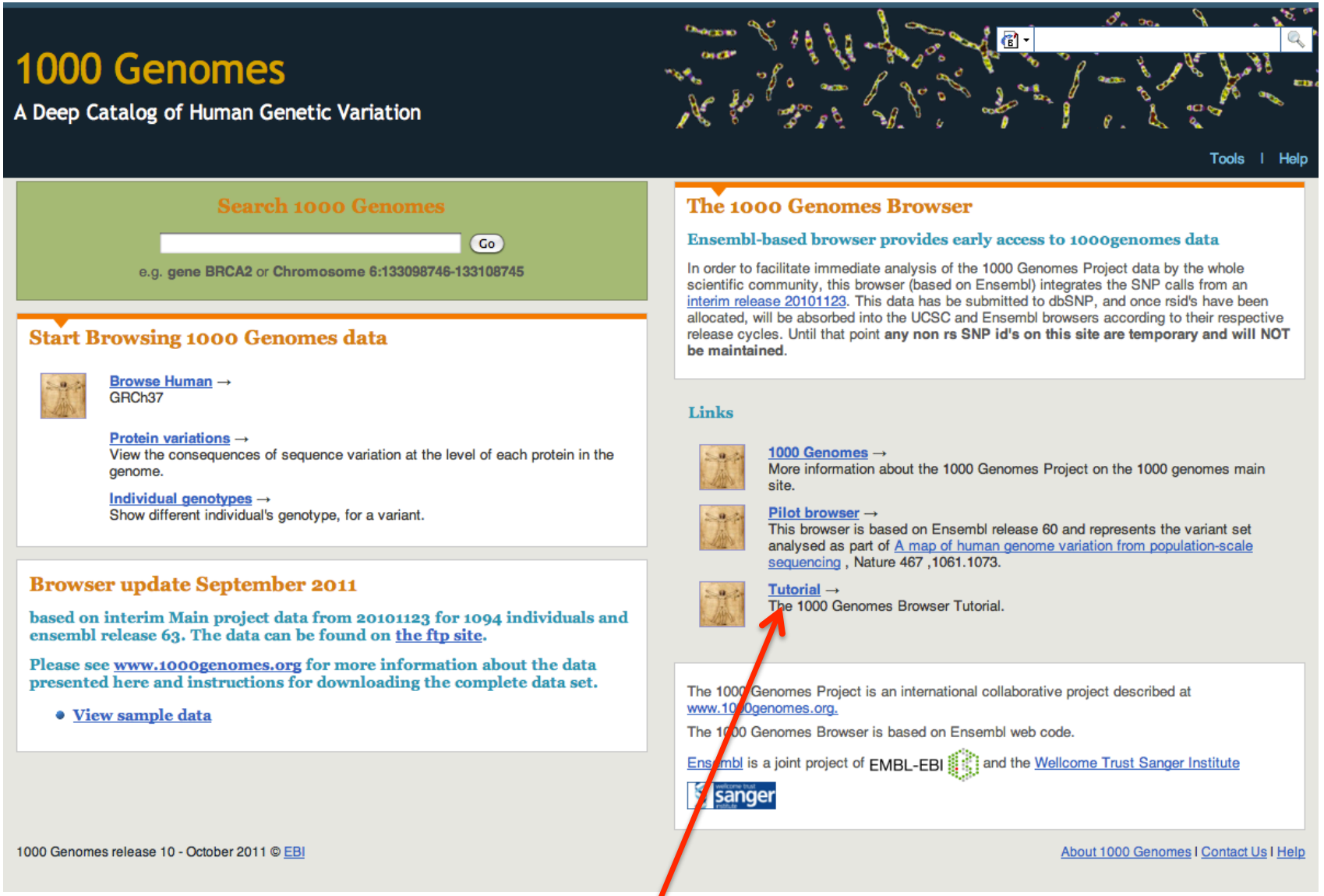
1000 Genomes is in the Amazon cloud

1KG pilot content (BAM) is available at
s3://1000genomes.s3.amazonaws.com

You can see the XML at
http://1000genomes.s3.amazonaws.com

Data access

- General information
- File access
- 1000 Genomes Browser
- Tools
- Where to find help



http://browser.1000genomes.org

1000 Genomes

A Deep Catalog of Human Genetic Variation

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

Location-based displays

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

Configure this page

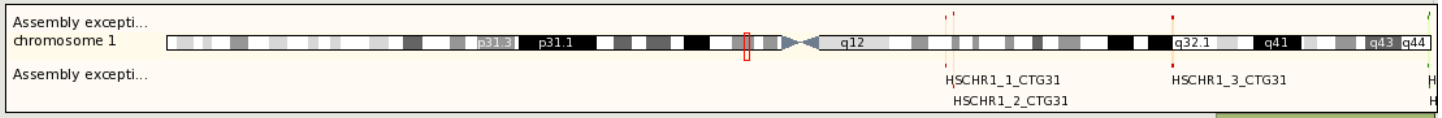
Manage your data

Export data

Get VCF data

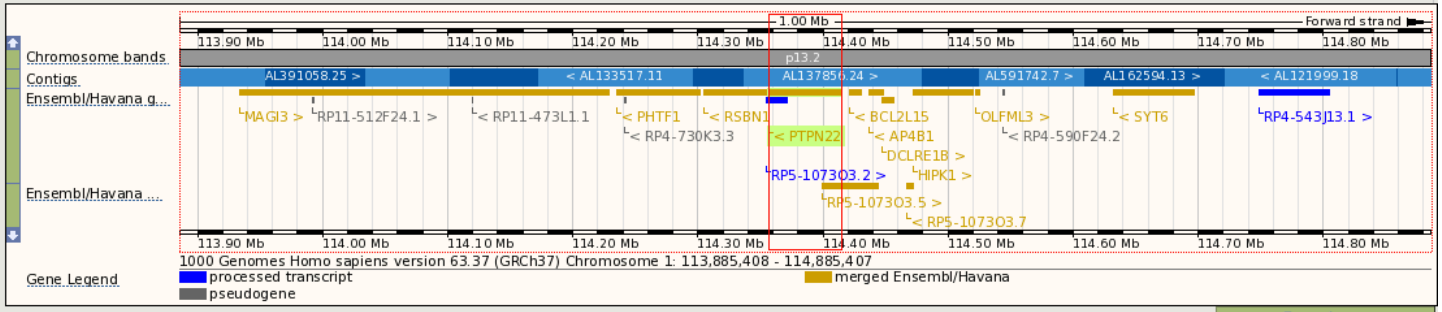
Bookmark this page

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



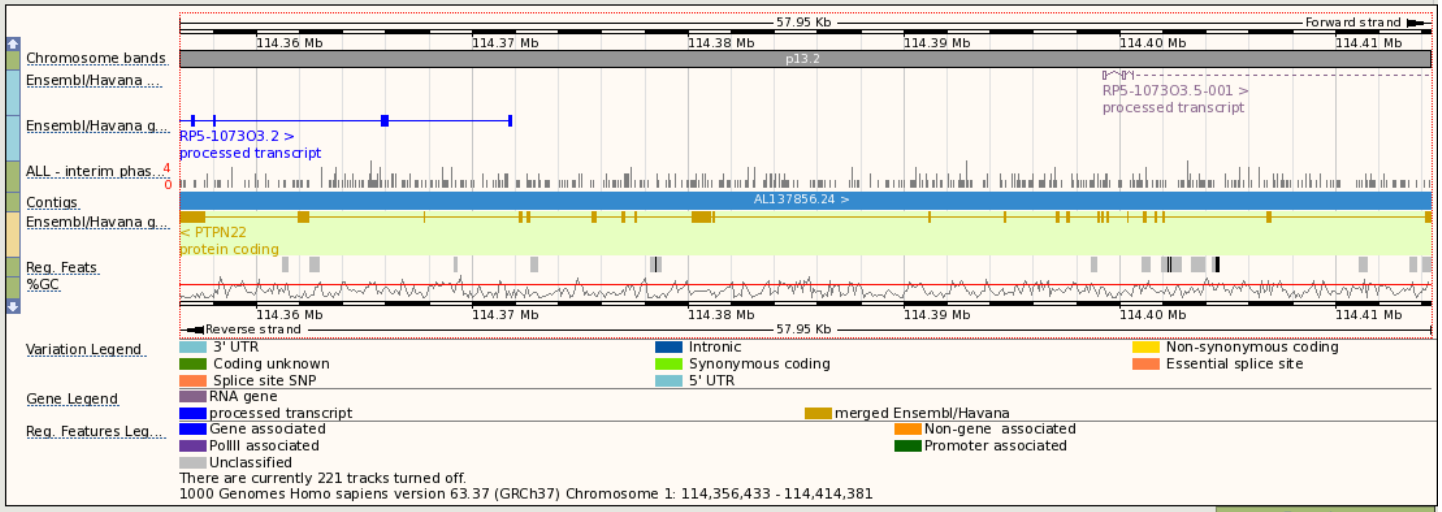
Export Image

Region in detail help



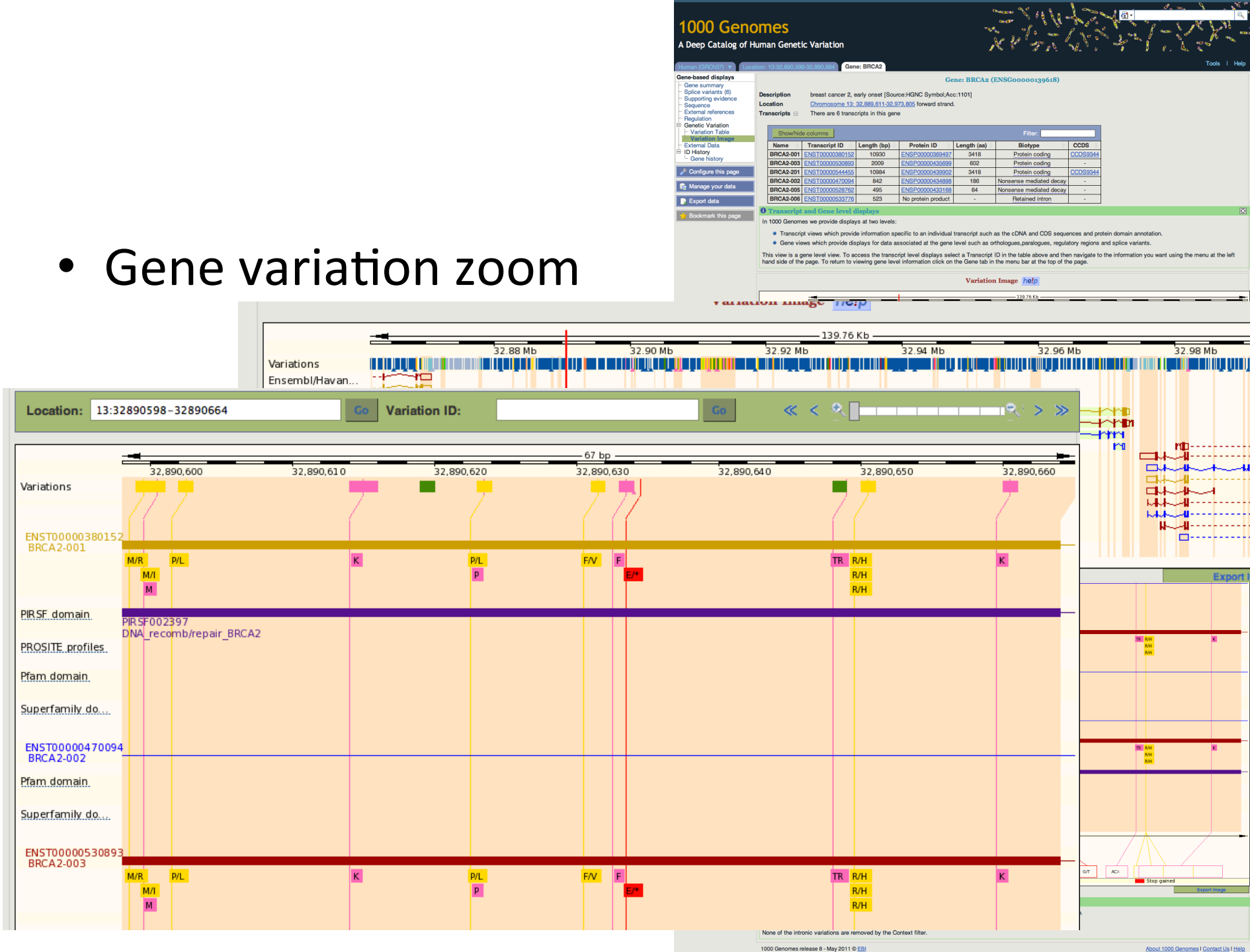
Export Image

Location: 1:114356433-114414381 Go Gene: Go

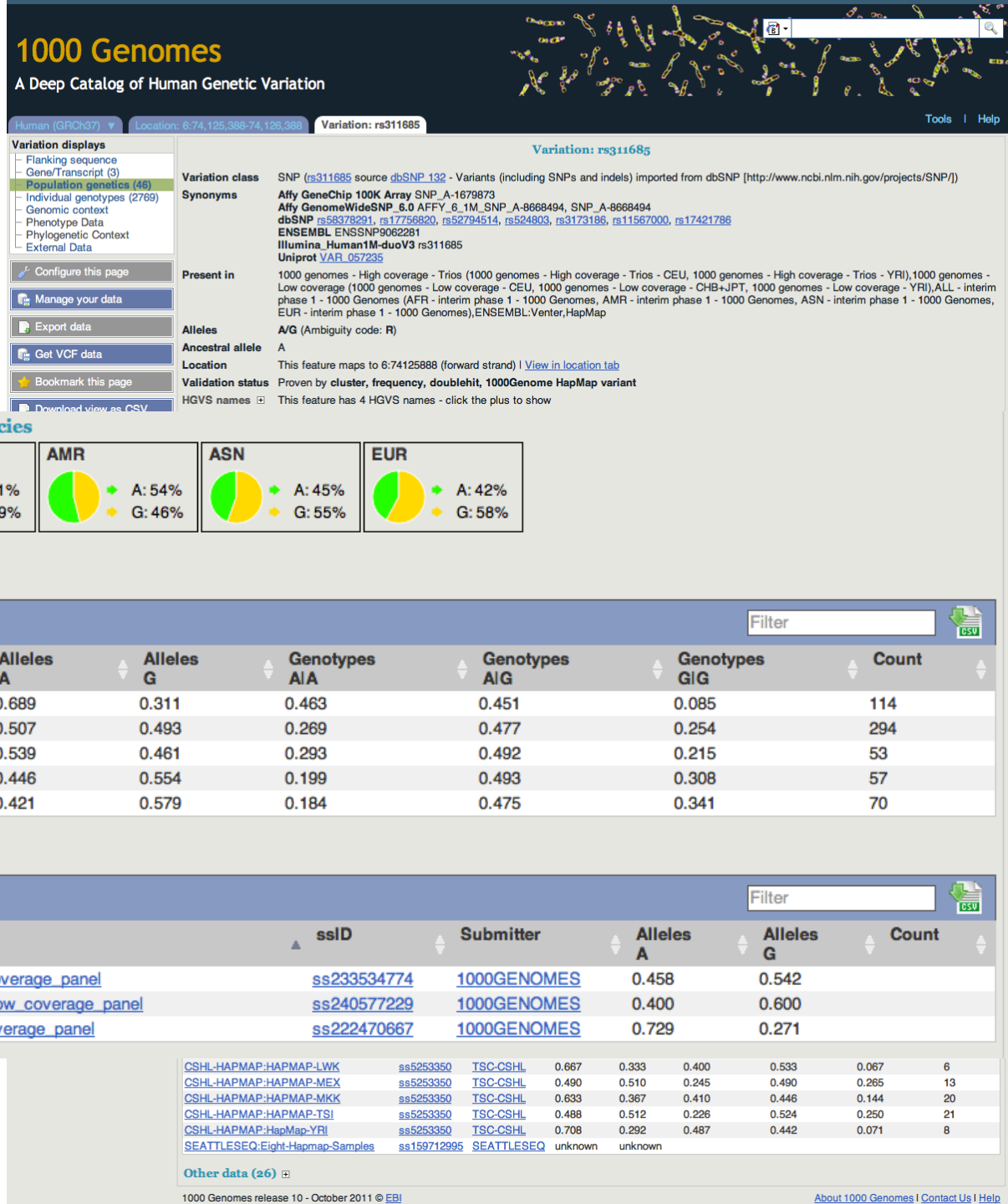


Export Image

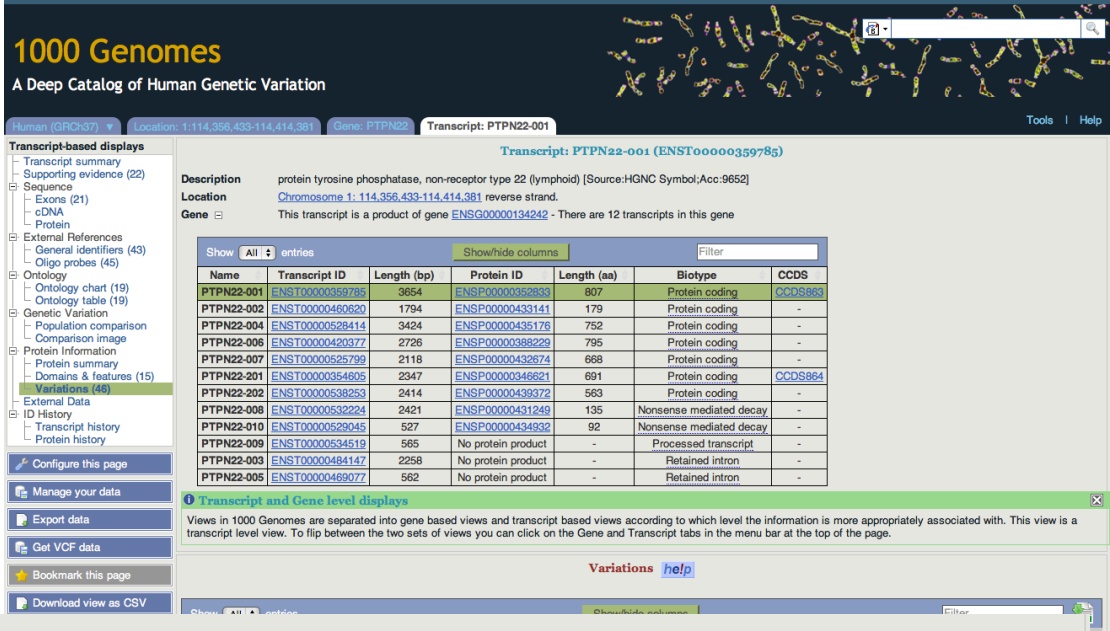
- Gene variation zoom



- Population


1000 Genomes
A Deep Catalog of Human Genetic Variation
Human (GRCh37)
Location: 6:74,125,388-74,126,388
Variation: rs311685
Tools | Help
Variation displays
Flanking sequence
Gene/Transcript (3)
Population genetics (46)
Individual genotypes (2769)
Genomic context
Phenotype Data
Phylogenetic Context
External Data
Configure this page
Manage your data
Export data
Get VCF data
Bookmark this page
Variation: rs311685
Variation class
SNP (rs311685 source dbSNP 132 - Variants (including SNPs and indels) imported from dbSNP [http://www.ncbi.nlm.nih.gov/projects/SNP/])
Synonyms
Affy GeneChip 100K Array SNP_A-1679873
Affy GenomeWideSNP_6.0 AFFY_6_1M_SNP_A-8668494, SNP_A-8668494
dbSNP rs58378291, rs17756820, rs52794514, rs524803, rs3173186, rs11567000, rs17421786
ENSEMBL ENSSNP9062281
Illumina_Human1M-duoV3 rs311685
Uniprot VAR_057235
Present in
1000 genomes - High coverage - Trios (1000 genomes - High coverage - Trios - CEU, 1000 genomes - High coverage - Trios - YRI),1000 genomes - Low coverage (1000 genomes - Low coverage - CEU, 1000 genomes - Low coverage - CHB+JPT, 1000 genomes - Low coverage - YRI),ALL - interim phase 1 - 1000 Genomes (AFR - interim phase 1 - 1000 Genomes, AMR - interim phase 1 - 1000 Genomes, ASN - interim phase 1 - 1000 Genomes, EUR - interim phase 1 - 1000 Genomes),ENSEMBL:Venter,HapMap
Alleles
A/G (Ambiguity code: R)
Ancestral allele
A
Location
This feature maps to 6:74125888 (forward strand) | View in location tab
Validation status
Proven by cluster, frequency, doublehit, 1000Genome HapMap variant
HGVS names
This feature has 4 HGVS names - click the plus to show
1000 genomes alleles frequencies
AFR
A: 69%
G: 31%
ALL
A: 51%
G: 49%
AMR
A: 54%
G: 46%
ASN
A: 45%
G: 55%
EUR
A: 42%
G: 58%
1000 genomes
Show/hide columns
Filter
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
Population
ssID
Submitter
Alleles A
Alleles G
Count
1000GENOMES:pilot_1_CEU_low_coverage_panel ss233534774 1000GENOMES 0.458 0.542
1000GENOMES:pilot_1_CHB+JPT_low_coverage_panel ss240577229 1000GENOMES 0.400 0.600
1000GENOMES:pilot_1_YRI_low_coverage_panel ss222470667 1000GENOMES 0.729 0.271
CSHL-HAPMAP:HAPMAP-LWK ss5253350 TSC-CSHL 0.667 0.333 0.400 0.533 0.067 6
CSHL-HAPMAP:HAPMAP-MEX ss5253350 TSC-CSHL 0.490 0.510 0.245 0.490 0.265 13
CSHL-HAPMAP:HAPMAP-MKK ss5253350 TSC-CSHL 0.633 0.367 0.410 0.446 0.144 20
CSHL-HAPMAP:HAPMAP-TSI ss5253350 TSC-CSHL 0.488 0.512 0.226 0.524 0.250 21
CSHL-HAPMAP:HapMap-YRI ss5253350 TSC-CSHL 0.708 0.292 0.487 0.442 0.071 8
SEATTLESEQ:Eight-Hapmap-Samples ss159712995 SEATTLESEQ unknown unknown
Other data (26)
1000 Genomes release 10 - October 2011 © EBI
About 1000 Genomes | Contact Us | Help

- SIFT
- PolyPhen

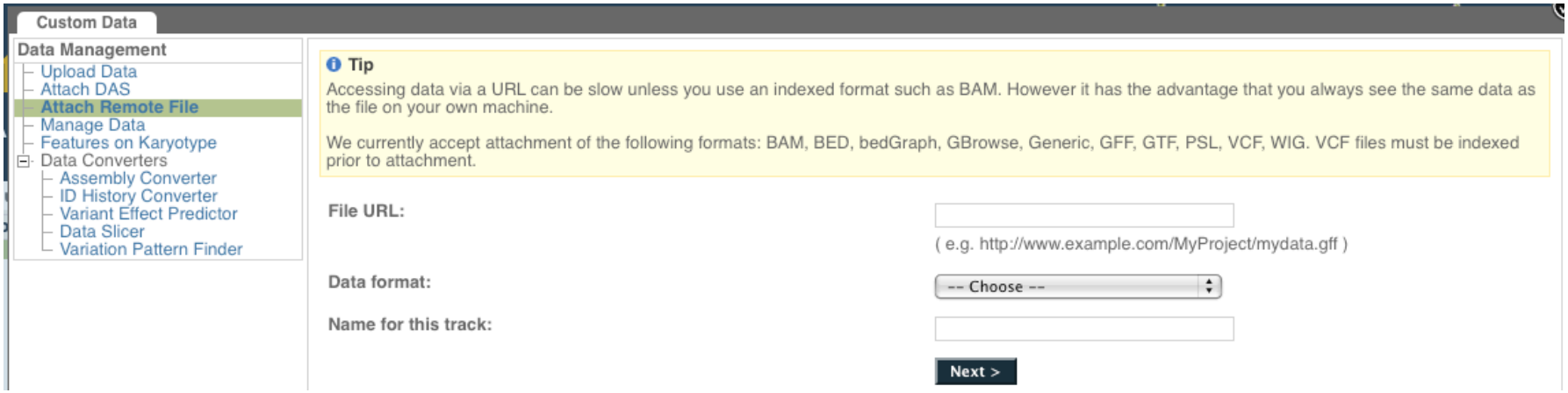


Download view as CSV

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

File upload to view with 1000 Genomes data



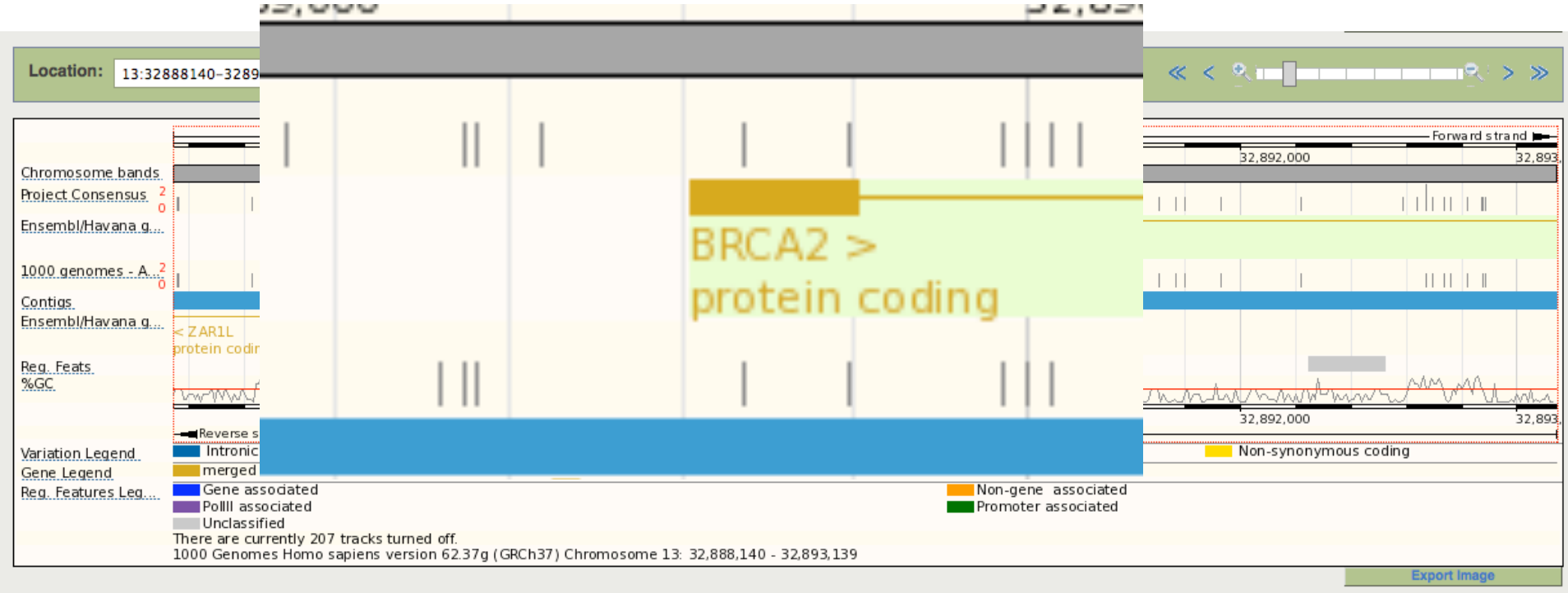
- Supports popular file types:
  - BAM, BED, bedGraph, BigWig, GBrowse, Generic, GFF, GTF, PSL, VCF*, WIG

* VCF must be indexed

Uploaded VCF

Example:

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

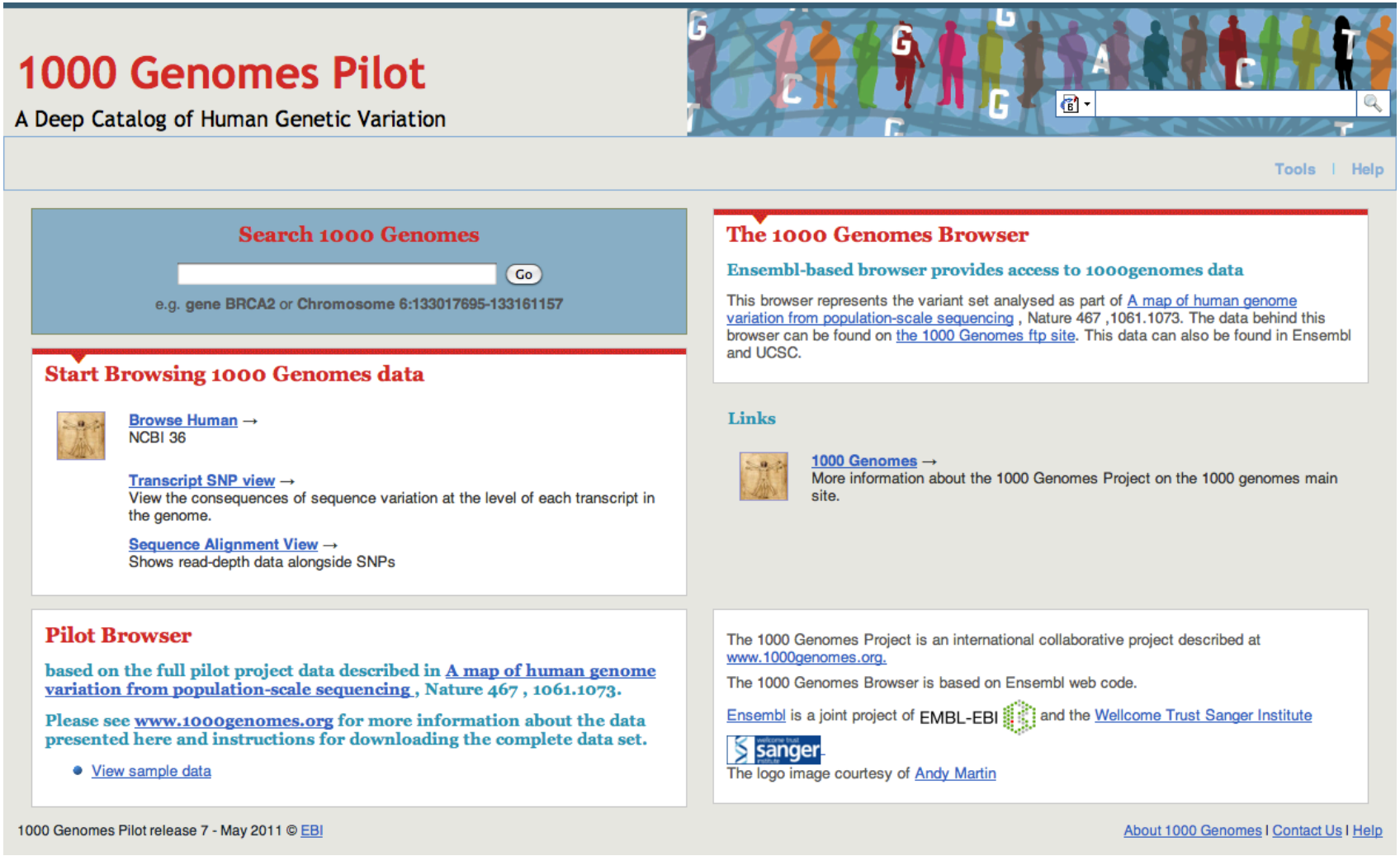


1000 Genomes Browser

- For further information on the capabilities of the browser and its use, attend the Ensembl "New Users" Workshop on Saturday at 12:30

SATURDAY, October 15

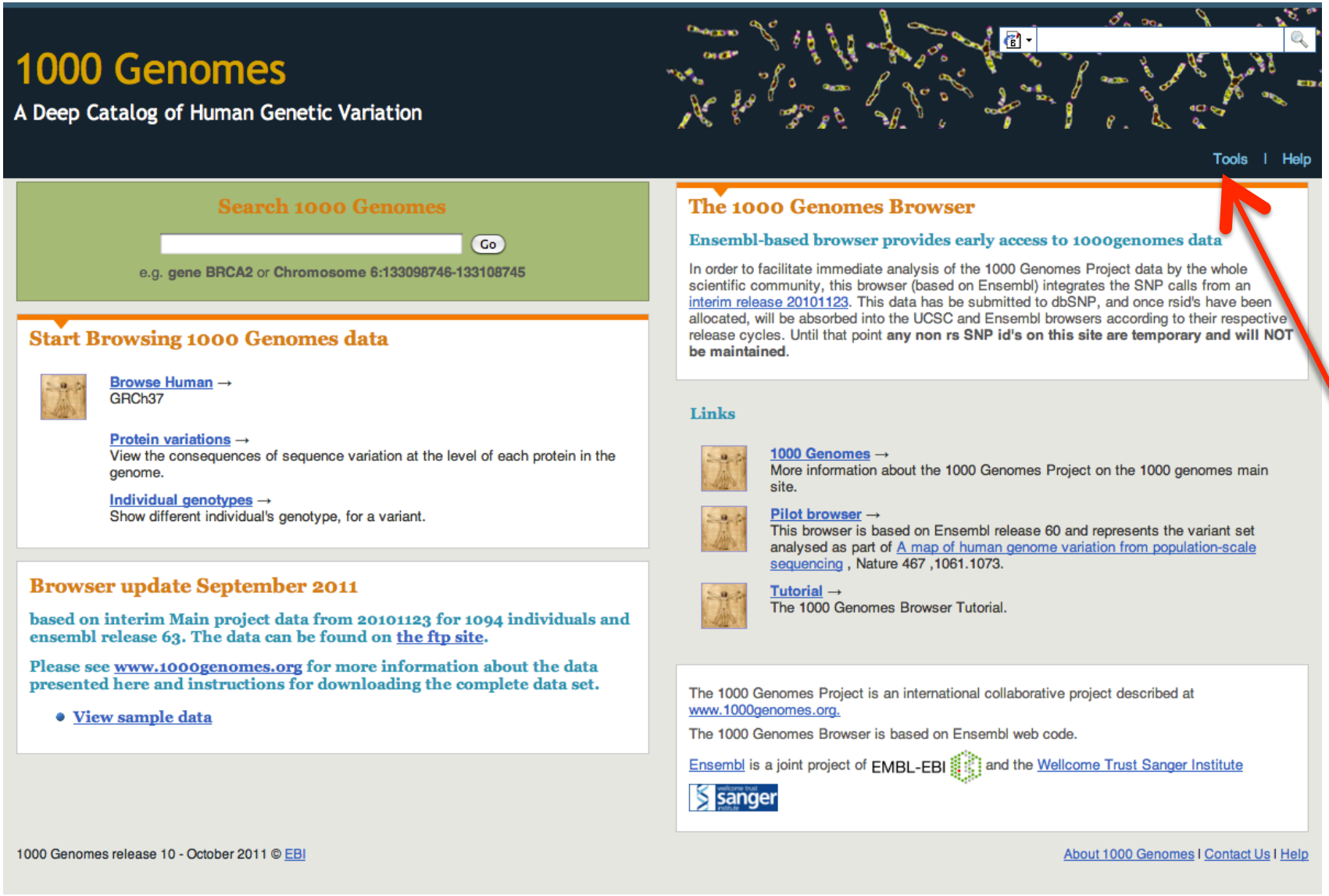
*12:30 PM - 1:30 PM		Ensembl 'New Users' Workshop: Web site and BioMart *For further information, e-mail xose@ebi.ac.uk*	Convention Center Room 524, Level 5



http://pilotbrowser.1000genomes.org

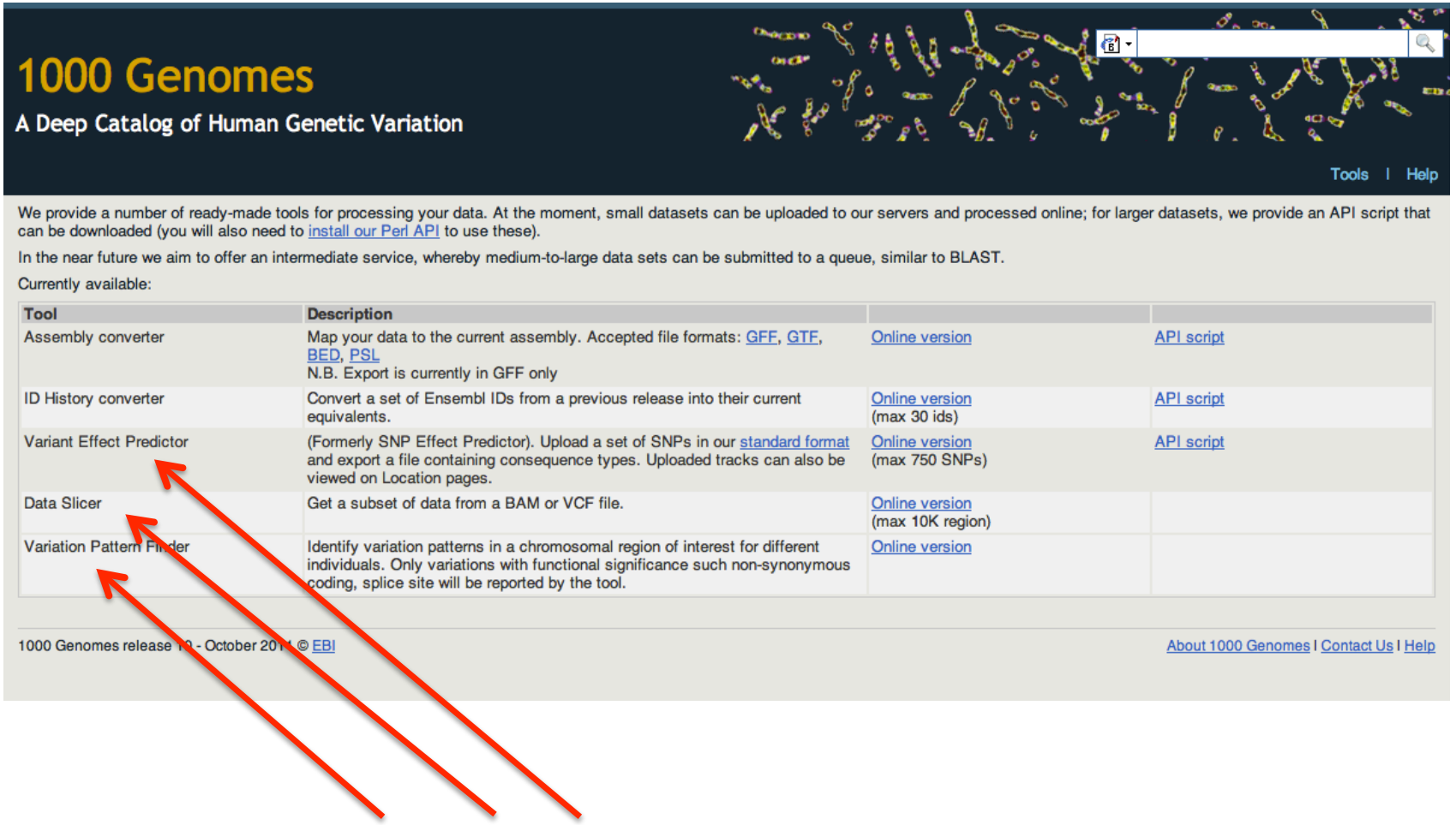
Data access

- General information
- File access
- 1000 Genomes Browser
- Tools
- Where to find help



http://browser.1000genomes.org

Tools page


1000 Genomes
A Deep Catalog of Human Genetic Variation
Tools | Help
We provide a number of ready-made tools for processing your data. At the moment, small datasets can be uploaded to our servers and processed online; for larger datasets, we provide an API script that can be downloaded (you will also need to install our Perl API to use these).
In the near future we aim to offer an intermediate service, whereby medium-to-large data sets can be submitted to a queue, similar to BLAST.
Currently available:
Tool
Description
Assembly converter
Map your data to the current assembly. Accepted file formats: GFF, GTF, BED, PSL
N.B. Export is currently in GFF only
Online version
API script
ID History converter
Convert a set of Ensembl IDs from a previous release into their current equivalents.
Online version
(max 30 ids)
API script
Variant Effect Predictor
(Formerly SNP Effect Predictor). Upload a set of SNPs in our standard format and export a file containing consequence types. Uploaded tracks can also be viewed on Location pages.
Online version
(max 750 SNPs)
API script
Data Slicer
Get a subset of data from a BAM or VCF file.
Online version
(max 10K region)
Variation Pattern Finder
Identify variation patterns in a chromosomal region of interest for different individuals. Only variations with functional significance such non-synonymous coding, splice site will be reported by the tool.
Online version
1000 Genomes release 19 - October 2011 © EBI
About 1000 Genomes | Contact Us | Help

Ensembl Variant Effector Predictor (VEP)

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

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)

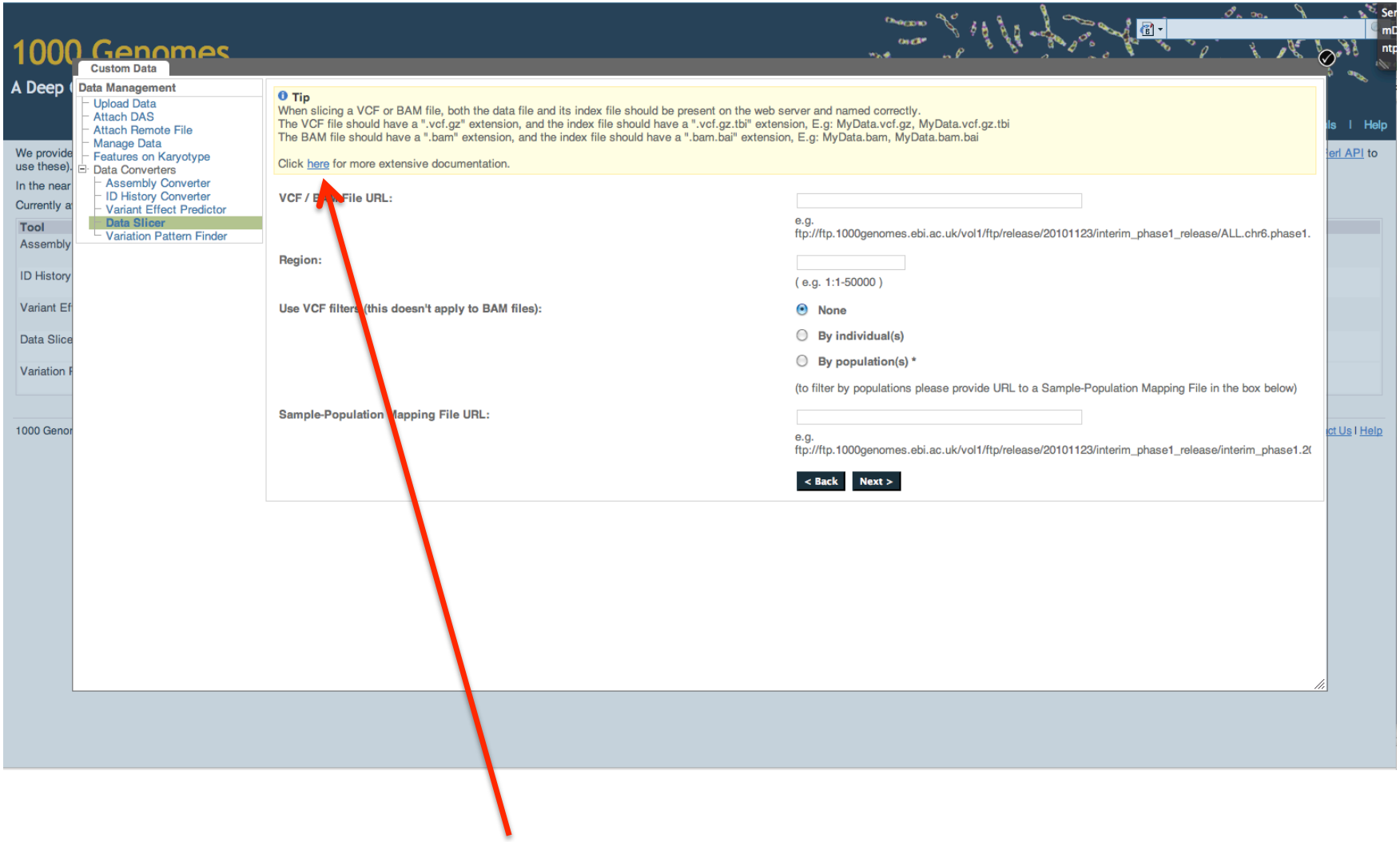
Filter variants by frequency: ☐

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

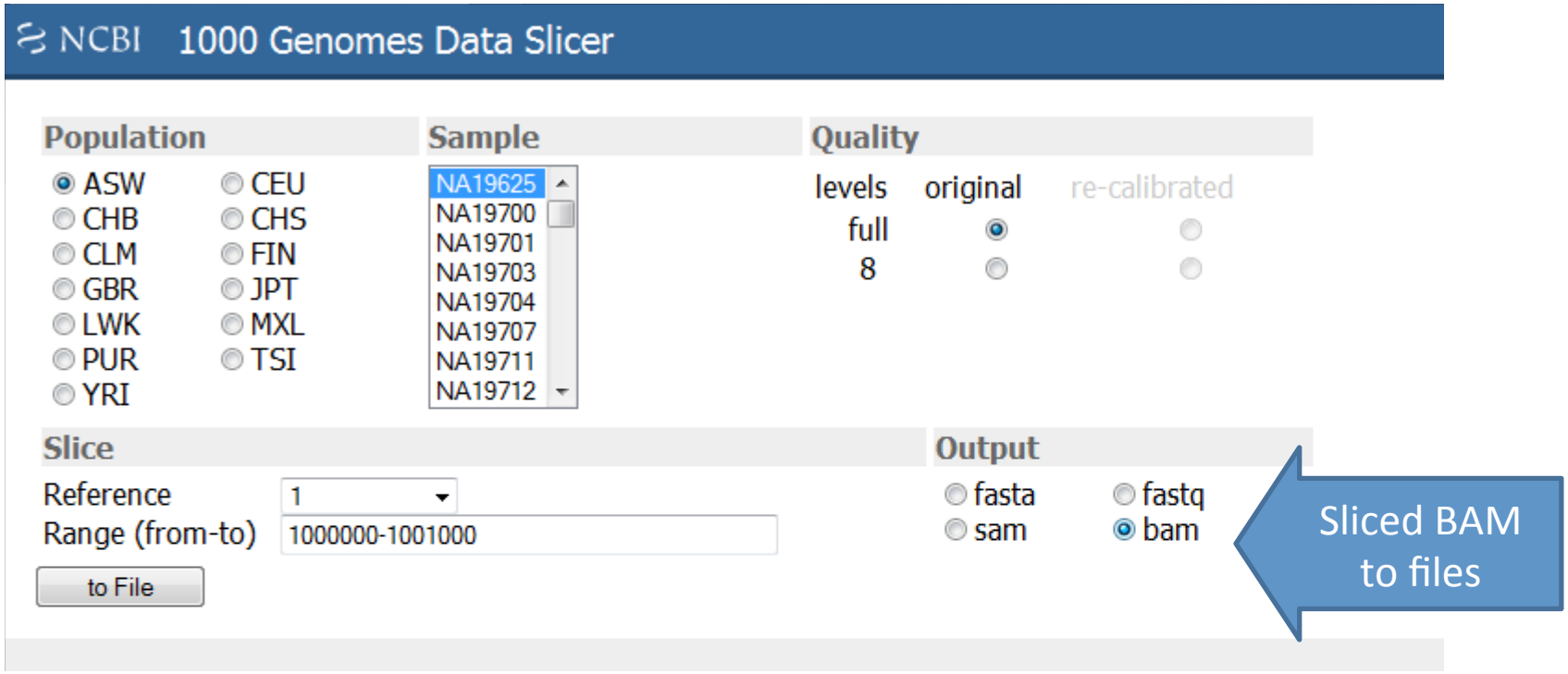
Filter: Exclude | variants with MAF greater than | 0.1 | in any 1KG low coverage population

Next >

Data slicer for subsets of the data

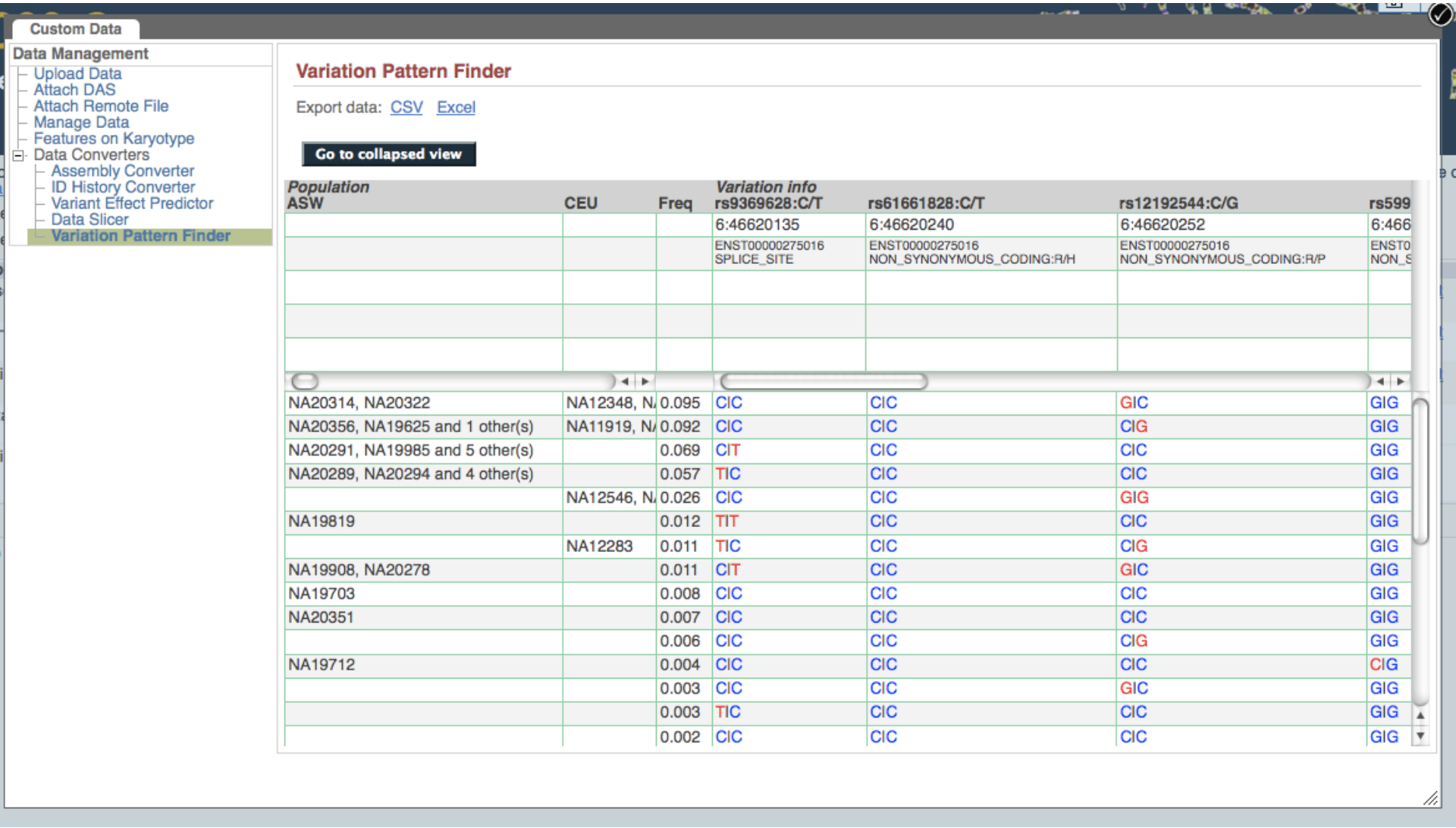


http://trace.ncbi.nlm.nih.gov/Traces/1kg_slicer/



Variation Pattern Finder

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



Population ASW	CEU	Freq	*Variation info* rs9369628:C/T	rs61661828:C/T	rs12192544:C/G	rs599
			6:46620135	6:46620240	6:46620252	6:466
			ENST00000275016 SPLICE_SITE	ENST00000275016 NON_SYNONYMOUS_CODING:R/H	ENST00000275016 NON_SYNONYMOUS_CODING:R/P	ENST0 NON_S
NA20314, NA20322	NA12348, N/	0.095	CIC	CIC	GIC	GIG
NA20356, NA19625 and 1 other(s)	NA11919, N/	0.092	CIC	CIC	CIG	GIG
NA20291, NA19985 and 5 other(s)		0.069	CIT	CIC	CIC	GIG
NA20289, NA20294 and 4 other(s)		0.057	TIC	CIC	CIC	GIG
	NA12546, N/	0.026	CIC	CIC	GIG	GIG
NA19819		0.012	TIT	CIC	CIC	GIG
	NA12283	0.011	TIC	CIC	CIG	GIG
NA19908, NA20278		0.011	CIT	CIC	GIC	GIG
NA19703		0.008	CIC	CIC	CIC	GIG
NA20351		0.007	CIC	CIC	CIC	GIG
		0.006	CIC	CIC	CIG	GIG
NA19712		0.004	CIC	CIC	CIC	CIG
		0.003	CIC	CIC	GIC	GIG
		0.003	TIC	CIC	CIC	GIG
		0.002	CIC	CIC	CIC	GIG

Access to backend Ensembl databases

- Public MySQL database at
  - mysql-db.1000genomes.org port 4272

- Full programmatic access with Ensembl API
  - More information on the use of the Ensembl API at the Ensembl "Advanced Users" Workshop tomorrow

FRIDAY, October 14

*6:15 PM - 8:00 PM		Ensembl 'Advanced Users' Workshop: API *For further information, e-mail xose@ebi.ac.uk*	Convention Center Room 524, Level 5

Data access

- General information
- File access
- 1000 Genomes Browser
- Tools
- Where to find help

- Can I map your snp coordinates between NCBI36 and GRCh37
- Can I use the 1000 genomes data for imputation?
- How are your alignments generated?
- Are input files available for using 1000 genomes data with the Beagle imputation algorithm?
- Are input files available for using 1000 genomes data with the Impute2 imputation algorithm?
- Are input files available for using 1000 genomes data with the Mach imputation algorithm?
- How can I get the allele frequency of my variant?
- How many individuals will be sequenced?
- How much disk space is used by the 1000 genomes project?
- How much sequence data has been generated for single individuals?
- Is the data for the pilot study still available?
- What Depth of Coverage was used to call the 1000 genomes snps
- What Sequencing Platforms were used for the 1000 genomes project
- What Structural variant data is available for the project?
- What are the targets for your exon targetted pilot study
- What are the targets for your whole exome sequencing?
- What do the names of your bam files mean?
- What do the names of your fastq files mean?
- What do the names of your vcf files mean?
- What does an individual have a genotype in a location where it has no sequence coverage?
- What format are your alignment files in?
- What format are your sequence files in?
- What format are your variant files in?
- What is a bas file?
- What is the difference between your data directory and the pilot_data/data directory
- What is the gender and family relationships of your samples?
- What library insert sizes where used in the 1000 genomes project
- What read lengths are being used by the project
- What tools can I use to download 1000 genomes data
- What version of vcf are your vcf files in?
- What was the source of the DNA for sequencing?
- Where are the pilot structural variants archived?
- Where are the snps for the X/Y/MT chr
- Where are your alignment files located?
- Where are your reference data sets?
- Where are your sequence files located?
- Where are your variant files located?
- Where can I get consequence annotations for the 1000 genome variants
- Where do I get the 1000 genomes data from?
- Where does the Ancestral Allele Information for your variants come from?
- Which samples are you sequencing?
- Why are the coordinates of your pilot variants different to what is displayed in Ensembl or UCSC
- Why do some of your vcf genotype files have genotypes of ./. in them?
- Why is only 85% of the genome assayable?
- Why is the Allele frequency different from Allele Count/Allele Number?
- Why is the sequence data distributed in 2 or 3 files labelled SRR_1, SRR_3 and SRR?
- Why isn't a snp in dbSNP or HapMap
- Why isn't my snp in browser.1000genomes.org

Do I need a password to access 1000 genomes data
›

Credits & Contact

- Eugene Kulesha, Iliana Toneva, Bren Vaughan
- Will McLaren, Graham Ritchie, Fiona Cunningham
- Laura Clarke, Holly Zheng-Bradley, Rick Smith

- Steve Sherry, Chunlin Xiao

For more information contact info@1000genomes.org