

# 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.1000genomes.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.

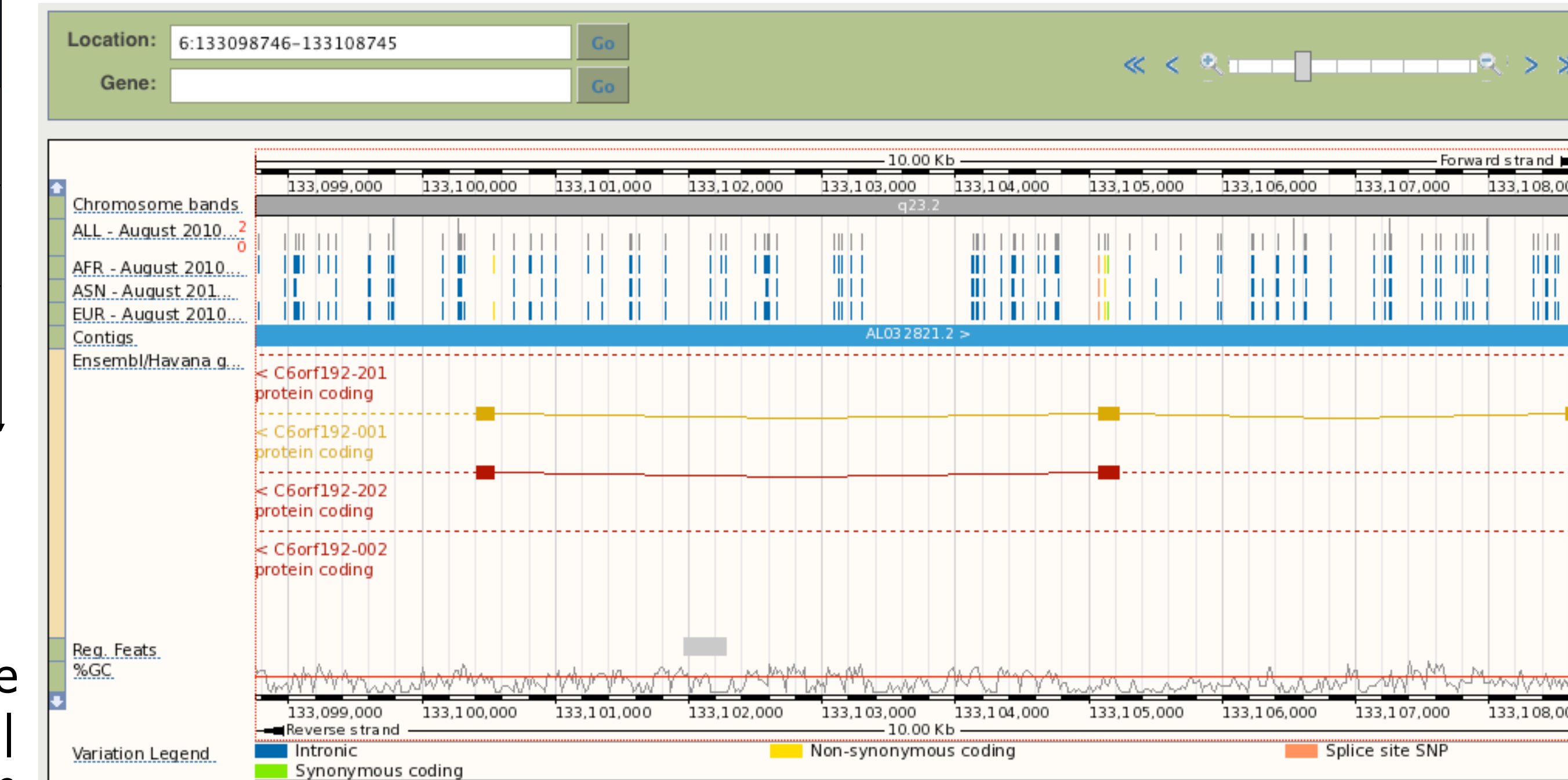
- Website <http://www.1000genomes.org/announcements>
- Twitter [@1000genomes](https://twitter.com/1000genomes)
- RSS <http://www.1000genomes.org/announcements/rss.xml>
- Email [1000announce@1000genomes.org](mailto:1000announce@1000genomes.org)

## 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.

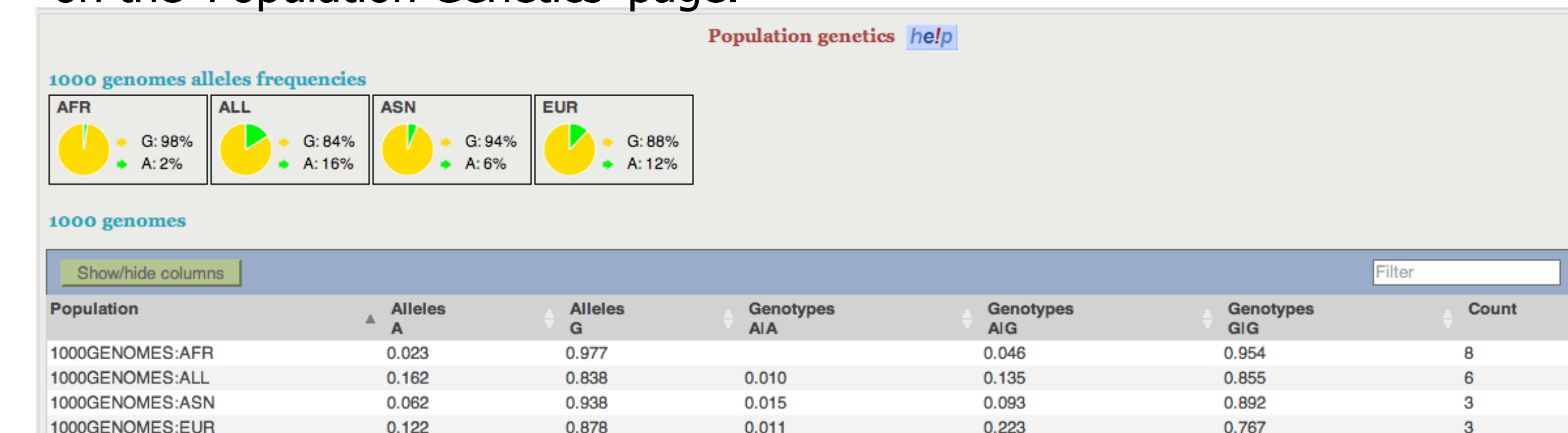
Tracks of 1000 genomes variants by population can be viewed in the location page:



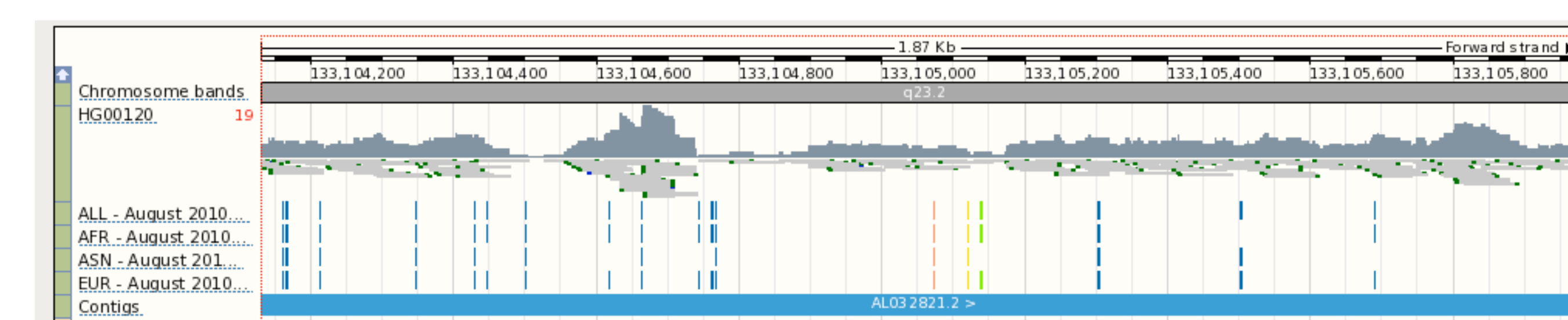
A list of variants can be obtained for any given transcript. In addition to basic information about a variant, PolyPhen and SIFT annotation are displayed to indicate the clinic significance of the variant.

Residue	Variation ID	Variation type	Alleles	Ambiguity code	Residues	Codons	SIFT	PolyPhen
3	rs35919294	Synonymous coding	T/G	K	L	CTA, CTC	-	-
19	rs30996469	Non-synonymous coding	CTT/A/G	N	R, L	CCG, CTC	tolerated	benign
22	rs72550670	Non-synonymous coding	G/A	R	R, W	CGG, TGG	deleterious	probably damaging
30	rs72483611	Stop gained, Splice site	G/A	M	E, *	GAA, TAA	-	-
80	rs11994534	Synonymous coding	A/G	R	Y	TAT, TAC	-	-
129	rs76169864	Synonymous coding	G/T	Y	E	GAG, GAA	-	-
126	rs72550671	Non-synonymous coding	G/T	K	H, N	CAC, AAC	tolerated	benign
137	CG9M29634	Non-synonymous coding	CGT/T	-	LE, LK	CTGAG, CTAAG	-	-
144	rs77913786	Non-synonymous coding	G/T	K	D, E	GAC, GAA	tolerated	benign

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.

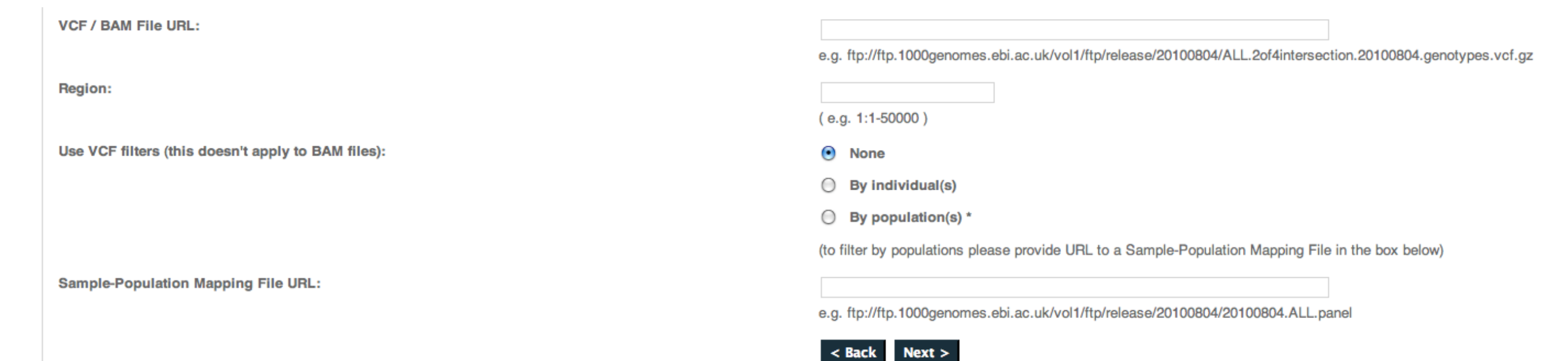
### Variation Effect Predictor

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:

Download text version	Location	Allele	Gene	Feature	Feature Type	Consequence	Position in cDNA	Position in CDS	Position in protein	Amino acid change	Codon change	Co-located Variation	Extra
	1,11436225, T/A	1,11436225	A	ENSG00000134242	ENST00000469077	Transcript WITHIN, NON_CODING_GENE	230	-	-	-	-	rs41313096	-
	1,11436225, T/A	1,11436225	A	ENSG00000134242	ENST00000468199	Transcript UPSTREAM	609	520	174	UL	Ata/Tta	rs41313096	SIFT:tolerated(0.67); PolyPhen:benign(0.005); Condel:neutral(0.079)
	1,11436225, T/A	1,11436225	A	ENSG00000134242	ENST00000352224	Transcript 5'UTR	2104	-	-	-	-	rs41313096	-
	1,11436225, T/A	1,11436225	A	ENSG00000134242	ENST00000420377	Transcript NON_SYNONYMOUS_CODING	2422	2333	778	NI	aAla/tI	rs41313096	SIFT:deleterious(0); PolyPhen:probably_damaging(0.905); Condel:deleterious(0.555)
	1,11436225, T/A	1,11436225	A	ENSG00000134242	ENST00000418928	Transcript DOWNSTREAM	-	-	-	-	-	rs41313096	-
	1,11436225, T/A	1,11436225	A	ENSG00000134242	ENST00000358253	Transcript NON_SYNONYMOUS_CODING	2097	1601	534	NI	aAla/tI	rs41313096	SIFT:deleterious(0.03); PolyPhen:probably_damaging(0.887); Condel:deleterious(0.628)

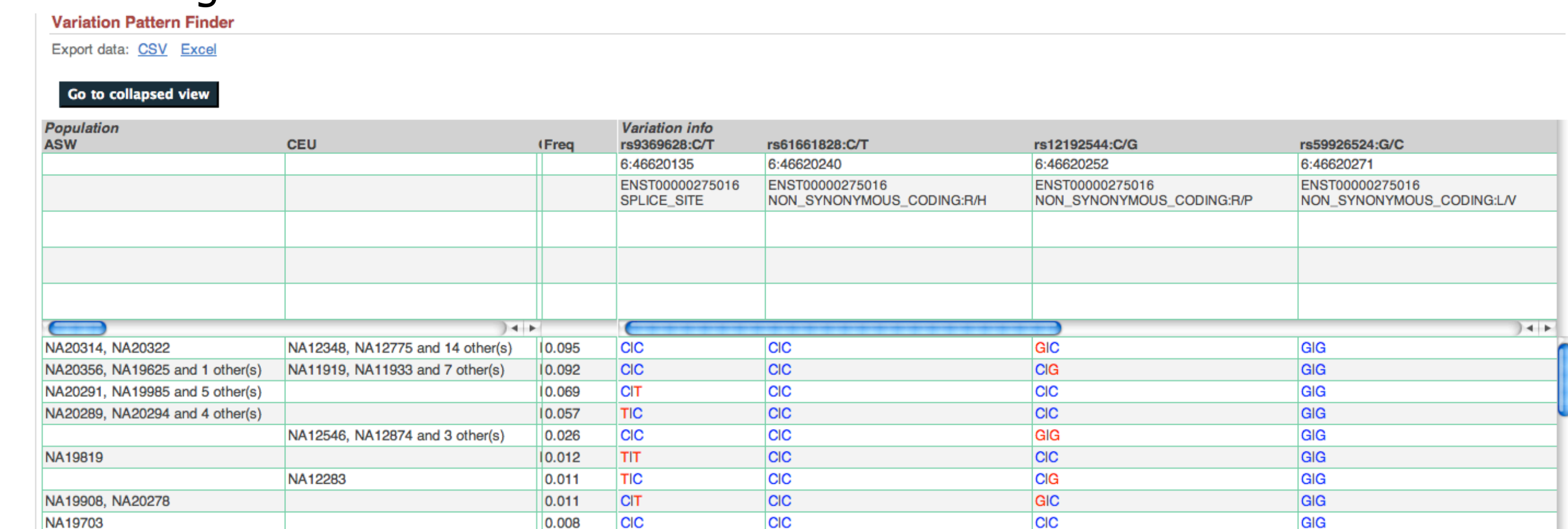
### Data Slicer

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 need 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 specified 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.



Population	CEU	AFR	ASN	EUR
rs1961828:C/T	0.092	0.092	0.092	0.092
rs12192544:C/G	0.009	0.009	0.009	0.009
rs59926524:G/C	0.057	0.057	0.057	0.057
rs12546, NA12874 and 3 other(s)	0.028	0.028	0.028	0.028
NA19819	0.012	0.012	0.012	0.012
NA19808, NA20278	0.011	0.011	0.011	0.011
NA19703	0.008	0.008	0.008	0.008

## Acknowledgements

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