L. Clarke, H. Zheng-Bradley, I. Streeter, E. Kulesha, J. Khobova, B. Vaughan, P. Flicek and The 1000 Genomes Consortium

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

The 1000 Genomes Project Data and Tools

The main goal of the 1000 Genomes Project is to establish a comprehensive catalogue of human genome variation. We now have low coverage, whole genome and exome sequence data for more than 2500 individuals across 26 different populations.

The sequence, alignment and variant data is available from both the EBI and NCBI ftp sites

- ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp
- ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp

Here we present some phase3 release stats and how to access the data and tools available from our browser (http://browser.1000genomes.org) to help the community utilise the data.

Phase 3 Data Release

Type	Count
biallelic_SNP	77,818,409
biallelic_del	1,930,866
biallelic_ins	1,058,174
multiallelic_SNP	259,371
multiallelic_del_ins	47,490
multiallelic_ins	34,911
biallelic_<CN0>	32,355
multiallelic_SNP_del	29,444
multiallelic_SNP_ins	28,011
multiallelic_del	13,765
biallelic_<INS:ME:ALU>	12,491
biallelic_<INS:ME:LINE1>	2,910
multiallelic_SNP_del_ins	1,248
biallelic_<INS:ME:SVA>	822
multiallelic_<CN>	169
biallelic_<INS:MT>	165
biallelic_<INV>	100
Total	**81,270,701**

The phase 3 release is based on 2504 unrelated individuals from 26 Populations

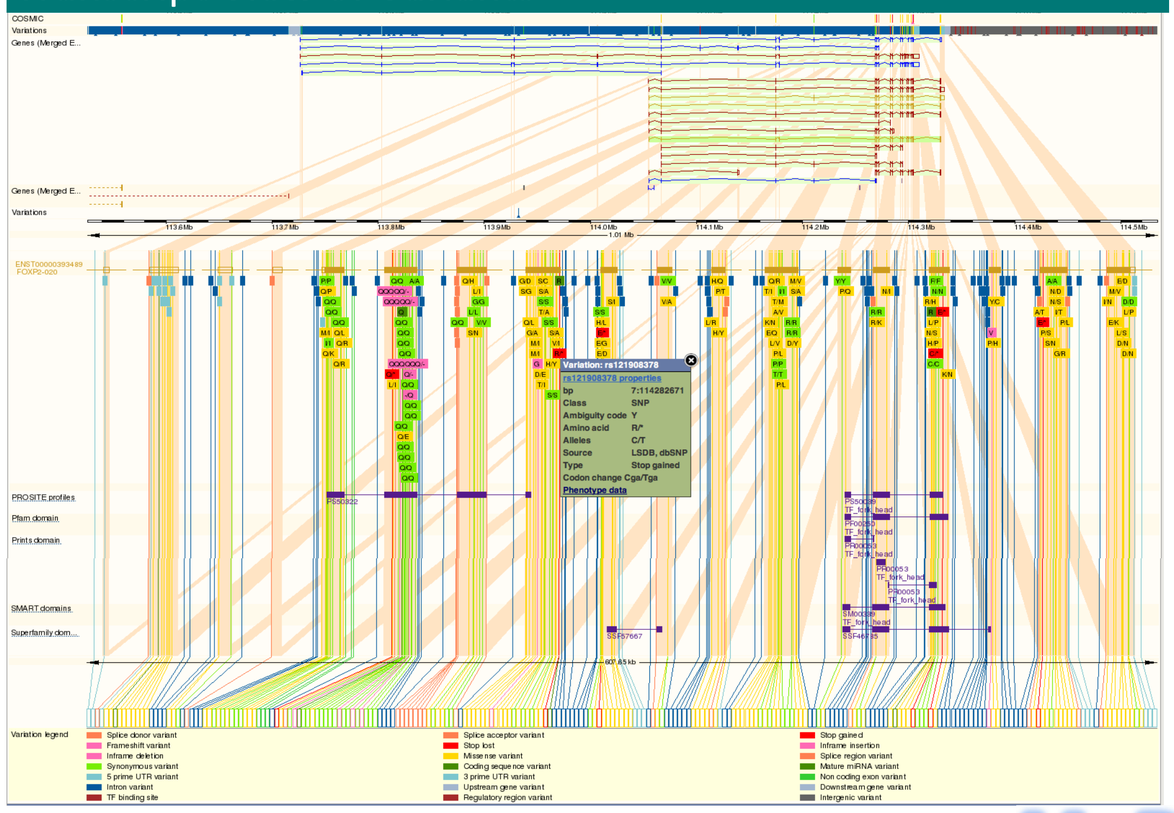
It contains phased genotypes for SNPs, short indels and large deletions like the phase 1 release. There are also new variant types including MNPs, long insertions and copy number changes.

This new release also contains multi allelic variants.

This is thanks to the new haplotype refinement pipeline which utilised ShapeIt2 and MVNCall to create this dataset

Tool	Description
The Browser	Browse our variations in genomic context alongside other annotation
The Data Slicer	Enabling download of genomic subsections of our alignments and variant files
The Ensembl Variant Effect Predictor (VEP)	Providing functional consequence annotation of given variants
The VCF to PED Converter	Converting VCF files to PED files
The Variation Pattern Finder	Finding patterns of shared inheritance between different individuals in the same VCF file
FORGE	Performs functional overlap analysis of GWAS SNPS
The Allele Frequency Calculator	Provides population specific allele frequencies for a region in a VCF file

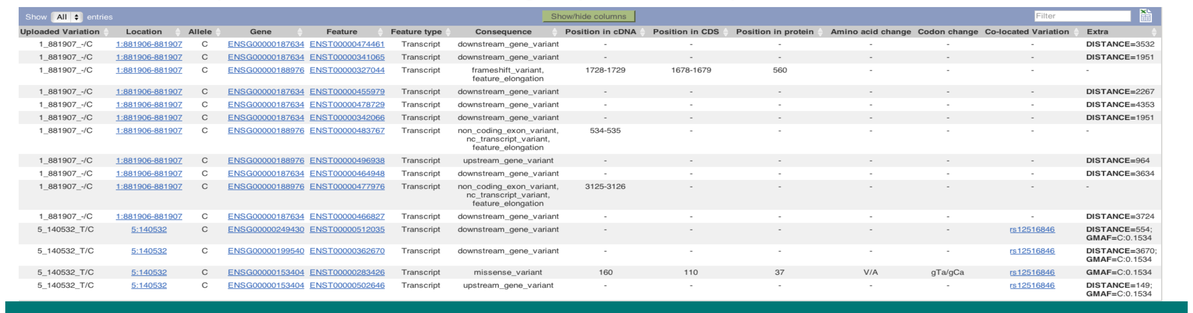
Transcript View of Variants



The Ensembl Variant Effect Predictor

The Ensembl Variant Effect Predictor allows you to provide variants and discover

- The genes affected by the variants.
- The location of the variants relative to other annotation (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- The consequence your variants on the protein sequence (e.g. stop gained, missense, frameshift)
- Known variants that match yours, and the associated minor allele frequencies
- SIFT and PolyPhen scores for changes protein sequence



Variation Pattern Finder

The Variation Pattern Finder calculates patterns of shared inheritance between individual. In a genomic region, some individuals will share distinct combinations of inheritance. The finder calculates:

- The distinct patterns of inheritance
- Which individuals have a particular pattern.
- The frequency of those patterns in each population.
- The pattern finder focuses on variants with functional consequences.

Here is example output:

ASW	CEU	CHB	CHS	CLM	FIN	GBR	JPT	LWK	MXL	PUR	TSI	YRI	Freq	rs116865371:G/A	rs149779687:T/C	rs200248570:C/T	rs9369628:C/T
														6:46620039	6:46620104	6:46620111	6:46620135
														ENST00000275016 intron	ENST00000275016 intron	ENST00000275016 intron	ENST00000275016 splice_region_variant
																	ENST00000275016 intron_variant
NA19908; NA		NA18615; NA	HG00683; HG	HG01360; HG	HG00190; HG	HG00261; HG	NA19056; NA	NA19038; NA	NA19758; NA	HG00734; HG	NA20538; NA	NA19189; NA	0.082	-	-	-	T|C
NA20289; NA	NA12283	NA18595; NA	HG00699; HG	HG01461; HG			NA18947; NA	NA19395; NA	NA19675; NA	HG00638; HG	NA20505; NA	NA18507; NA	0.074	-	-	-	C|T
NA19819		NA18612	HG00406	HG01384			NA18992	NA19317; NA	NA19782; NA	HG01108			0.012	-	-	-	T|T
		NA18557; NA	HG00404; HG				NA18953; NA						0.007	G|A	-	-	-
		NA18543					NA18976; NA						0.005	A|G	-	-	-
NA20351				HG01488						HG01095		NA18853	0.004	-	T|C	-	-
NA20412												NA19209	0.002	-	C|T	-	-
		NA18627					NA18945						0.001	A|G	-	-	C|T
NA19703													0.001	-	C|C	-	-
			HG00593										0.001	-	-	T|C	-

Allele Frequency Calculator

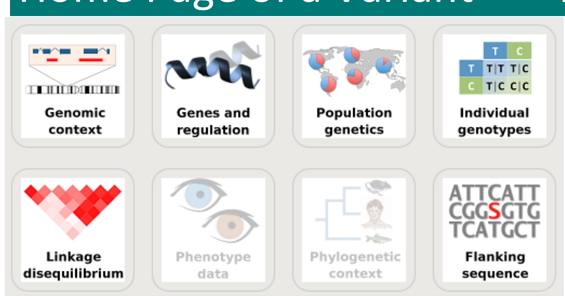
CHR	POS	ID	REF	ALT	TOTAL_CNT	ALT_CNT	FRQ
22	17004085	rs182269758	A	G	170	9	0.05
22	17004141	rs192917218	A	G	170	2	0.01
22	17004902	rs4010207	A	G	170	12	0.07
22	17004113	rs186630910	A	G	170	9	0.05
22	17004057	rs187809646	A	G	170	11	0.06
22	17003679	rs2890298	A	G	170	12	0.07
22	17004914	rs5992906	C	T	170	66	0.39
22	17004058	rs190038164	T	C	170	4	0.02
22	17004861	rs4010206	C	T	170	0	0
22	17004497	rs142081770	G	T	170	0	0
22	17004071	rs9605433	G	A	170	17	0.1
22	17004153	rs9680545	G	A	170	16	0.09
22	17004454	rs8135289	T	C	170	71	0.42
22	17004348	rs28417007	T	A	170	0	0
22	17003990	rs146134165	A	C	170	28	0.16
22	17003695	rs56873984	C	T	170	17	0.1
22	17004593	rs145217318	T	A	170	0	0
22	17003703	rs191454176	G	A	170	0	0
22	17003987	rs184186321	T	A	170	11	0.06

The Allele Frequency Calculator takes:

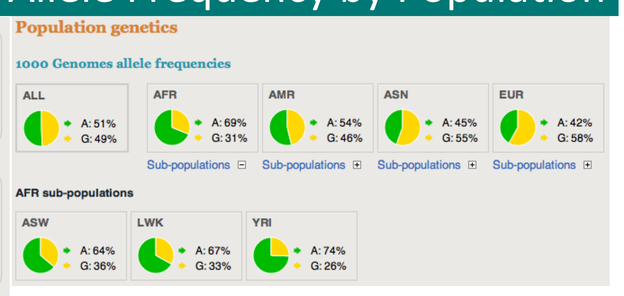
- A VCF file
- A sample to population mapping file
- A chromosomal region
- A population name

It calculates population-wide allele frequency for sites within the defined region. Here is example output.

Home Page of a Variant



Allele Frequency by Population



Variant Properties

Residue	Variation ID	Variation type	Alleles	Minor allele	Global frequency	Ambiguity code	Residues	Codons	SIFT	PolyPhen
2	COSM1548107	Missense variant	G/A	-	-	R	M, I	ATG, ATA	0.01	0.948
4	rs75306477	Synonymous variant	C/T	T	0.001	Y	P	CCC, CCT	-	-
5	rs74507296	Missense variant	A/C	-	-	M	Q, P	CAG, CCG	0.09	0.992
7	COSM173607	Synonymous variant	C/T	-	-	Y	I	ATC, ATT	-	-
11	COSM1548106	Missense variant	C/A	-	-	M	Q, K	CAA, AAA	0	0
13	rs140009033	Synonymous variant	G/A	-	-	R	Q	CAG, CAA	-	-
17	rs142280302	Synonymous variant	G/A	-	-	R	Q	CAG, CAA	-	-
32	rs80073025	Synonymous variant	A/G	G	0.001	R	Q	CAA, CAG	-	-
33	COSM1568681	Missense variant	A/G	-	-	R	Q, R	CAG, CGG	0.04	0
33	COSM23407	Missense variant	A/T	-	-	W	Q, L	CAG, CTG	0.11	0
39	COSM85073	Missense variant	A/G	-	-	R	Q, R	CAG, CGG	0.11	0
57	COSM33426	Stop gained	C/T	-	-	Y	Q, *	CAG, TAG	-	-
58	COSM1196467	Missense variant	C/A	-	-	M	L, I	CTT, ATT	0.25	0

The International Genome Sample Resource Beyond 1000 Genomes Project

The International Genome Sample Resource (IGSR), a Wellcome Trust funded project that will be built on the foundation of the 1000 Genomes Project starts in 2015.

IGSR plans to:

- Maintain the existing 1000 genomes data and move to GRCh38
- Collect other data sets generated on the Coriell Cell Lines including Geuvadis
- Add new populations to explain the global diversity of the variant catalog

Acknowledgements

We would like to thank the Ensembl variation team for all their help and the Wellcome Trust and European Molecular Biology Laboratory for their funding.

1000 Genomes
A Deep Catalog of Human Genetic Variation

http://www.1000genomes.org
info@1000genomes.org
@1000genomes