

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

# Introduction

The main goal of the 1000 genomes project is to establish a comprehensive and http://browser.1000genomes.org detailed catalogue of human genome variations; which in turn will empower The 1000 Genomes project utilizes the Ensembl Browser to display our association studies to identify disease-causing genes. The project now has data variant calls. We provide rapid access to project variant calls through the and variant genotypes for more than 1000 individuals in 14 populations. The ftp browser before they become available via dbSNP and DGVa. site contains more than 120Tbytes of data in 200,000 files. Tracks of 1000 genomes variants by population can be viewed in the location

DATA TYPE	FILE FORMAT	SIZE
sequence	FASTQ	43 Tbases raw sequ
alignment	BAM	56 Tbytes of BAM f
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. (<u>ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/</u> 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

Search term:	Th <i>e</i>
interim_phase1_release	
Search for files on the FTP site	use
Help on searching	to
- • Search options	che
Search	filte
	fast
RESULTS	

74 files found

File

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim\_phase1\_release /ALL.chr9.phase1.projectConsensus.genotypes.vcf.gz.tbi ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim\_phase1\_release /ALL.chr9.phase1.projectConsensus.genotypes.vcf.gz

### 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

# **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

Visualization



search allows ers to specify which site to get paths to get md5 ecksums and also er out high volume tq files

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.

Show All \$	entries			Show/hide colu	mns			Filter
Residue 🛓	Variation ID	Variation type	Alleles	Ambiguity code	Residues	Codons	SIFT	PolyPhen
3	rs35910094	Synonymous coding	T/G	к	L	CTA, CTC	-	-
19	rs33996649	Non-synonymous coding	C/T/A/G	N	R, L	CGG, CTG	tolerated	benign
22	rs72650670	Non-synonymous coding	G/A	R	R, W	CGG, TGG	deleterious	probably damaging
33	rs72483511	Stop gained, Splice site	C/A	M	E, *	GAA, TAA	-	-
80	rs113984534	Synonymous coding	A/G	R	Y	TAT, TAC	-	-
122	rs74163654	Synonymous coding	C/T	Y	E	GAG, GAA	-	-
126	rs72650671	Non-synonymous coding	G/T	К	H, N	CAC, AAC	tolerated	benign
137	COSM25634	Non-synonymous coding	CC/TT	-	LE, LK	CTGGAG, CTAAAG	-	-
144	rs77913785	Non-synonymous coding	G/T	к	D, E	GAC, GAA	tolerated	benign

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

1000 genomes al	leles frequenci	ASN	EUR		
← G: 98%	◆ G: 84% ◆ A: 16%	6 G: 94%	G: 88% → A: 12%		
1000 genomes					
Show/hide colum	ns				
Population		Alleles A	Alleles G	Genotypes AIA	Gen
1000GENOMES:AFF	}	0.023	0.977		0.046
		0.162	0.838	0.010	0.135
1000GENOMES:ALL					
1000GENOMES:ALL	1	0.062	0.938	0.015	0.093
USEROMES:ALL	an Atta	0.062 0.122 ach rem	0.938 0.878 Note files	0.015 0.011 as cus	stom tra
Users C HG0012	an Atta 0 track	0.062 0.122 ach rem is 1000	0.938 0.878 ote files Genome	o.o15 o.o11 as cus es bam f	stom tra file addeo
USEROMES:ALL 1000GENOMES:ASM 1000GENOMES:EUF USERS C HG0012	an Atta 0 track	0.062 0.122 ach rem is 1000	0.938 0.878 ote files Genome	as cus es bam f	stom tra file addeo
1000GENOMES:ALL 1000GENOMES:ASN 1000GENOMES:EUF USERS C HG0012	an Atta 0 track	0.062 0.122 ach rem is 1000	0.938 0.878 ote files Genome	0.015 0.011 as cus es bam f	0.093 0.223 stom tra file addeo 1.87 КЬ 133,105,000 1
Chromosome ba	an Atta 0 track	0.062 0.122 ach rem is 1000	0.938 0.878 0 ote files 0 Genome	0.015 0.011 as cus es bam f	0.093 0.223 stom tra file added
Users C HG00120	an Atta 0 track	0.062 0.122 ach rem is 1000	0.938 0.878 10te files 100 133,104,600	0.015 0.011 as cus es bam f	0.093 0.223 stom tra file added 1.87 КЬ 133,105,000 1 q23.2
Chromosome ba	an Atta 0 track	0.062 0.122 ach rem is 1000	0.938 0.878 10te files 100 133,104,600	0.015 0.011 as cus es bam f	0.093 0.223 stom tra file added
Chromosome ba HG00120	an Atta 0 track	0.062 0.122 ach rem is 1000	0.938 0.878 0 ote files 0 Genome	0.015 0.011 as cus es bam f	0.093 0.223 stom tra file added
Chromosome ba HG00120	an Atta 0 track	0.062 0.122 ach rem is 1000	0.938 0.878 0 ote files 0 Genome	0.015 0.011 as cus es bam f	0.093 0.223 stom tra file added
Chromosome ba HG00120 ALL - August 202 ASN - August 202 EUR - August 202	an Atta 0 track	0.062 0.122 ach rem is 1000	0.938 0.878 0 ote files 0 Genome	0.015 0.011 as cus es bam f	0.093 0.223 stom tra file added

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



s. In example below, the 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:

Variant Effect	Predictor Re	sults:											
Download text ver	rsion												
Uploaded Variation	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_114362225_T//	A <u>1:114362225</u>	Α	ENSG0000134242 EN	NST00000469077	Transcript	WITHIN_NON_CODING_GENE	230	-	-	-	-	rs41313296	-
1_114362225_T//	A <u>1:114362225</u>	Α	ENSG0000231128 El	NST00000448199	Transcript	UPSTREAM	-	-	-	-	-	rs41313296	
1_114362225_T//	A <u>1:114362225</u>	A	ENSG00000134242 Et	NST00000460620	Transcript	NON_SYNONYMOUS_CODING	609	520	174	I/L	Ata/Tta	<u>rs41313296</u>	SIFT=tolerated(0.67); PolyPhen=benign(0.005); Condel=neutral(0.079)
1_114362225_T//	A <u>1:114362225</u>	Α	ENSG00000134242 EN	NST00000532224	Transcript	NMD_TRANSCRIPT, 3PRIME_UTR	2104	-	-	-	-	<u>rs41313296</u>	
_114362225_T//	A <u>1:114362225</u>	A	ENSG00000134242 Et	NST00000420377	Transcript	NON_SYNONYMOUS_CODING	2422	2333	778	N/I	aAt/aTt	<u>rs41313296</u>	SIFT=deleterious(0); PolyPhen=probably_damaging(0.905); Condel=deleterious(0.955)
1_114362225_T//	A <u>1:114362225</u>	Α	ENSG0000231128 EN	NST00000418238	Transcript	DOWNSTREAM	-	-	-	-	-	rs41313296	
1_114362225_T//	A <u>1:114362225</u>	A	ENSG00000134242 Et	<u>NST00000538253</u>	Transcript	NON_SYNONYMOUS_CODING	2097	1601	534	N/I	aAt/aTt	<u>rs41313296</u>	SIFT=deleterious(0.03); PolyPhen=probably_damaging(0.887); Condel=deleterious(0.928)

### **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:

VCF / BAM File URL:
Region:
Use VCF filters (this doesn't apply to BAM files):

ample-Population Mapping File UR

### **Variation Pattern Finder**

- shared variation between individuals in the a VCF file.
- Within a vcf file different samples have different combination of variation specifed region, shared by different individuals.
- The VPF only on variations that functional consequences for protein changes.

Variation Pattern Finder						
Export data: CSV Excel						
Go to collapsed view						
Population ASW	CEU	(Freq	Variation info rs9369628:C/T	rs61661828:C/T	rs12192544:C/G	rs59926524:G/C
			6:46620135	6:46620240	6:46620252	6:46620271
			ENST00000275016 SPLICE_SITE	ENST00000275016 NON_SYNONYMOUS_CODING:R/H	ENST00000275016 NON_SYNONYMOUS_CODING:R/P	ENST00000275016 NON_SYNONYMOUS_CODING:L/V
	) 4 )	•	<b>(</b>			) 4 + (
NA20314, NA20322	NA12348, NA12775 and 14 other(s)	0.095	CIC	CIC	GIC	GIG
NA20356, NA19625 and 1 other(s)	NA11919, NA11933 and 7 other(s)	10.092	CIC	CIC	CIG	GIG
NA20291, NA19985 and 5 other(s)		0.069	CIT	CIC	CIC	GIG
NA20289, NA20294 and 4 other(s)		10.057	TIC	CIC	CIC	GIG
	NA12546, NA12874 and 3 other(s)	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

## Acknowledgements

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

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20100804/ALL.2of4intersection.20100804.genotypes.vcf.gz
(e.g. 1:1-50000)
None
By individual(s)
By population(s) *
(to filter by populations please provide URL to a Sample-Population Mapping File in the box below)
e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20100804/20100804.ALL.panel
< Back Next >

• The Variation Pattern Finder (VPF) allows one to look for patterns of

genotypes. The VPF looks for distinct variation combinations within a user

coding genes such as non-synonymous coding SNPs and splice site

