The 1000 Genomes Project
How to Access the Data?

7/11/2012

Laura Clarke

EMBL-EBI

How Much Data?

EMBL-EBI

How Much Data?

- There are 333,332 files on the ftp site
- There are 367T of data on the ftp site
- There are 26 populations
- There are 2,208 samples
- There are 51,913 gigabases of low coverage sequence
- 18877x coverage in low coverage
- There are 23,860 gigabases of exome sequence

EMBL-EBI

Access

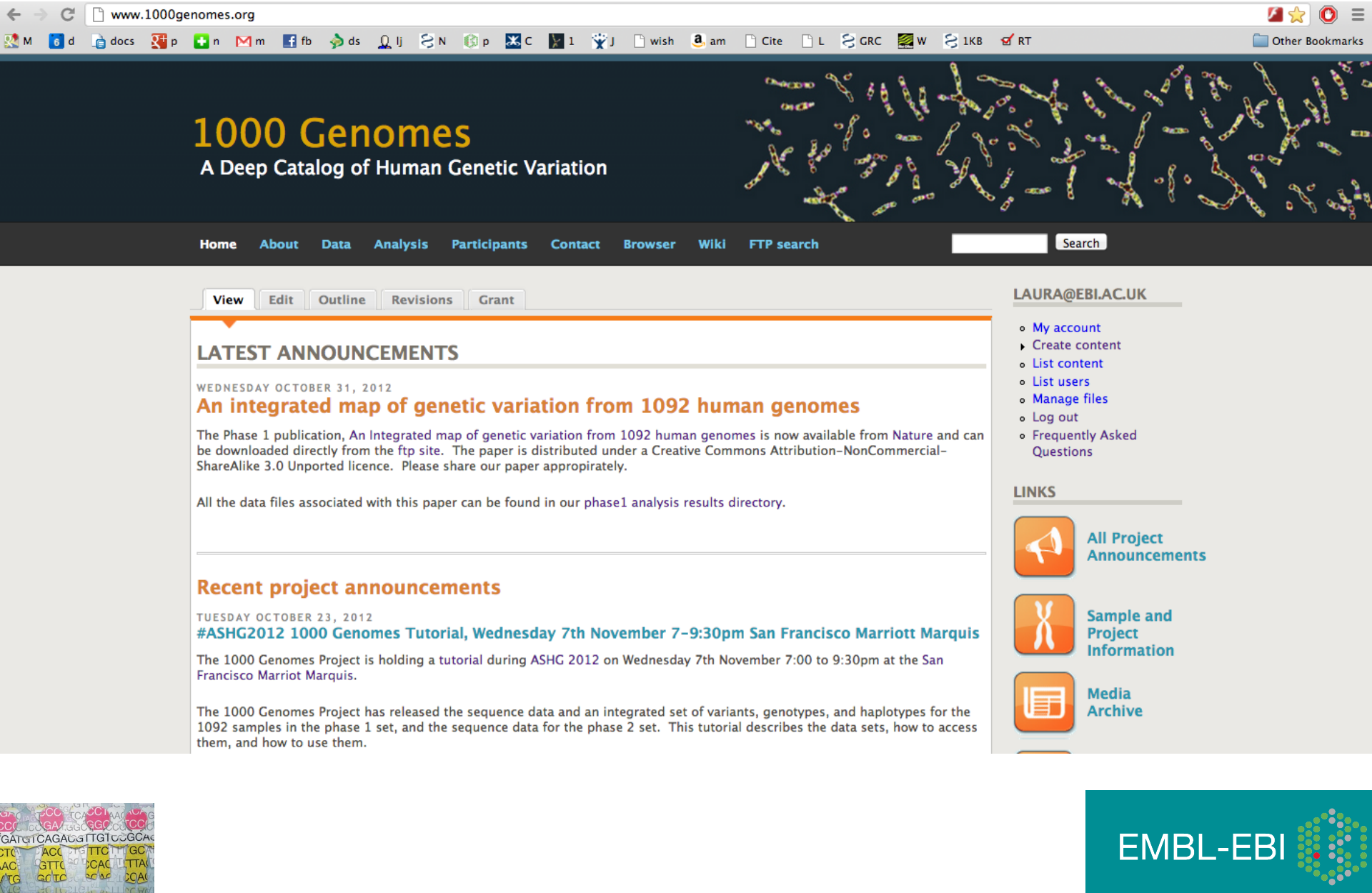
EMBL-EBI

Access

- Website
  - http://www.1000genomes.org
- 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- Can be up to 24 hours out of date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
- Amazon Cloud public data set

EMBL-EBI

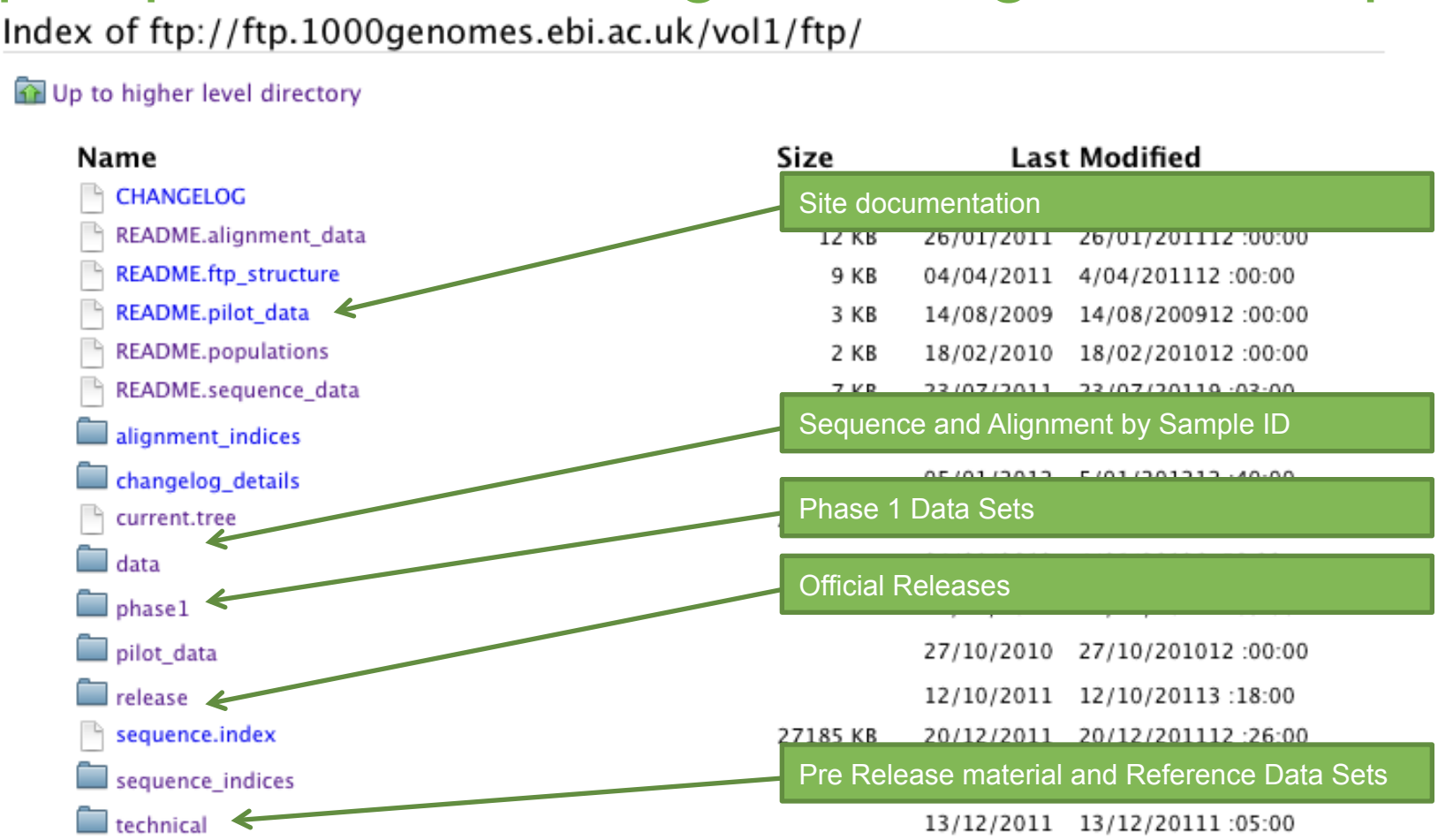
Website http://www.1000genomes.org



EMBL-EBI

ftp://ftp.1000genomes.ebi.ac.uk

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp



EMBL-EBI

Phase I complete; paper just published

ARTICLE

doi:10.1038/nature11632

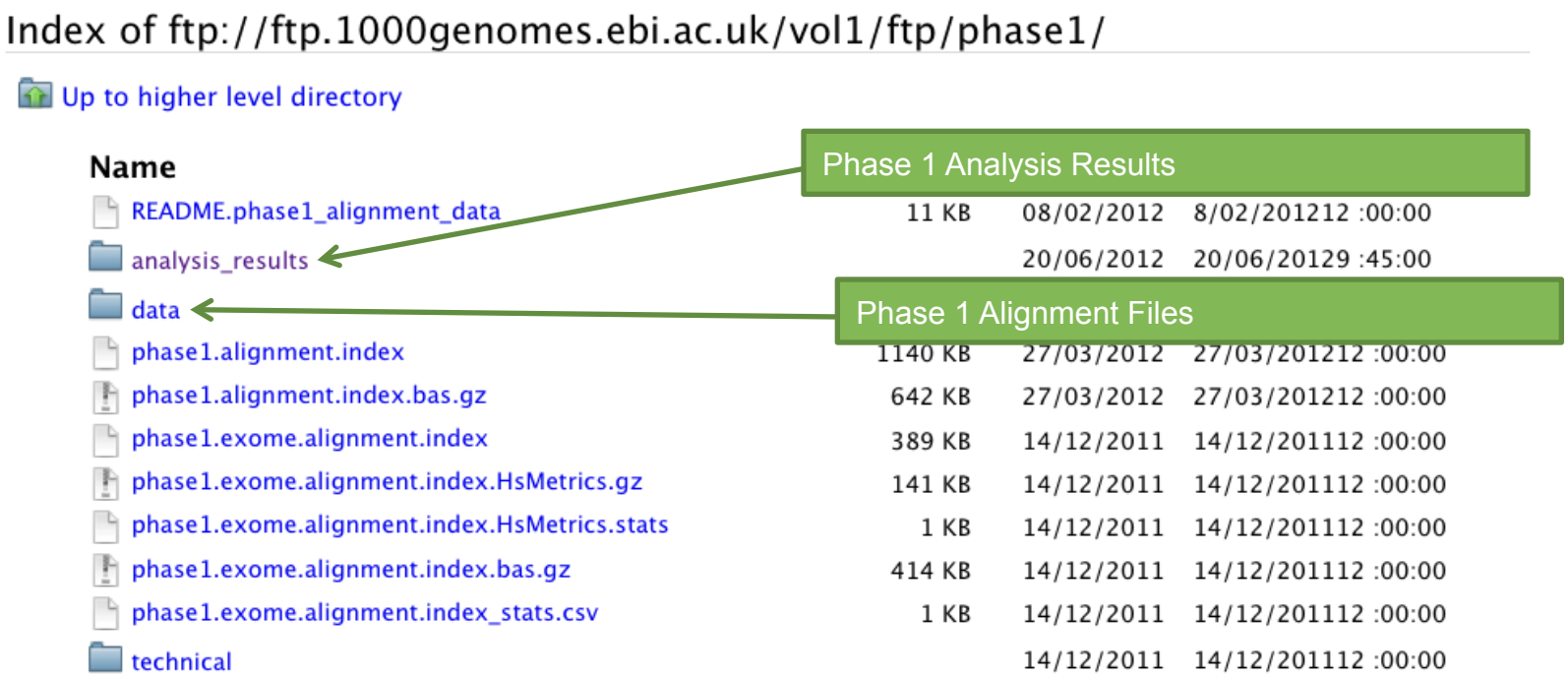
An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

Nature. 2012. 491:56-65

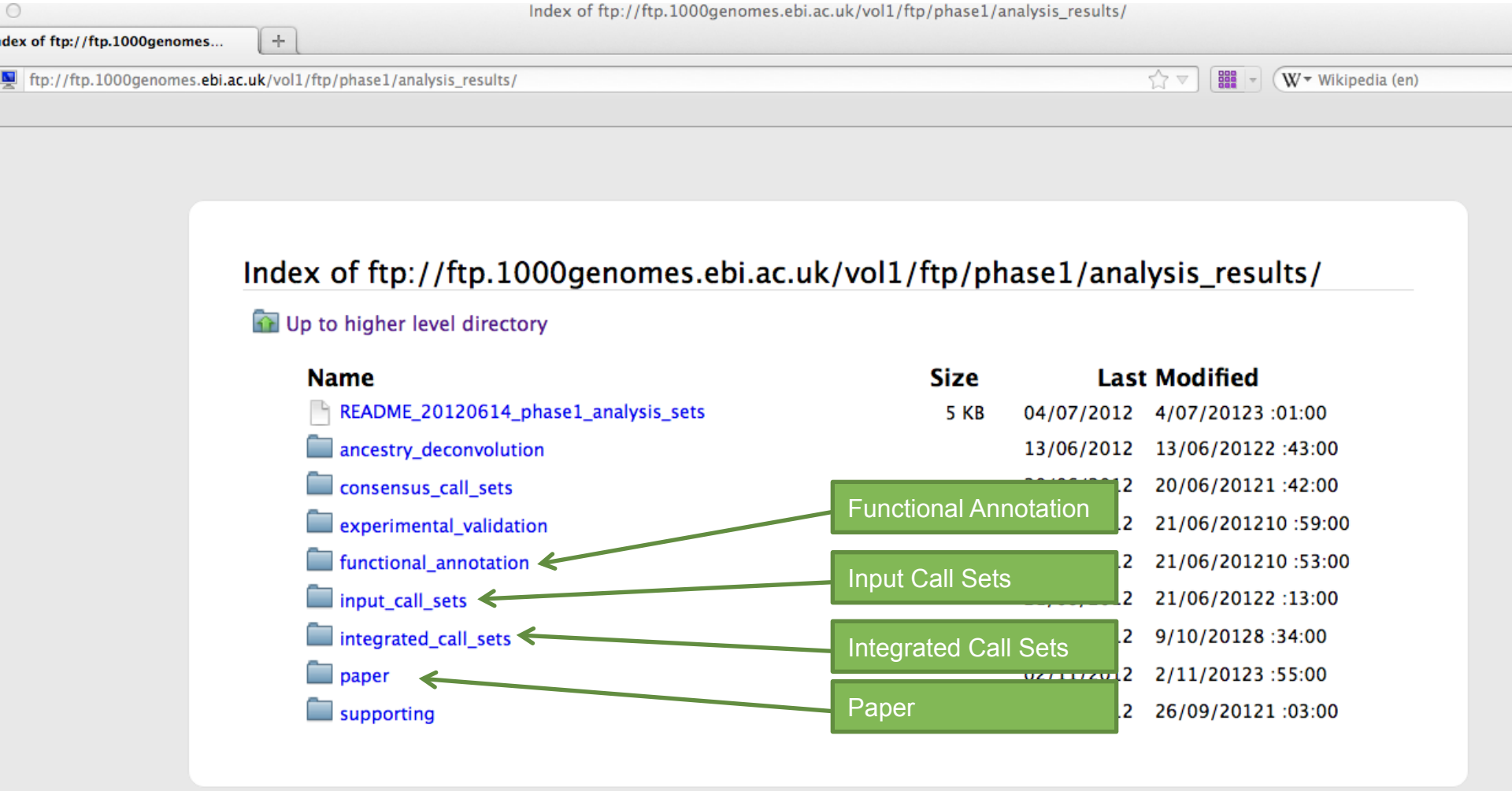
EMBL-EBI

FTP Site: Phase1 Data



EMBL-EBI

FTP Site: Phase1 Data



EMBL-EBI

Data formats and key tools

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

BAM alignment files

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

EMBL-EBI

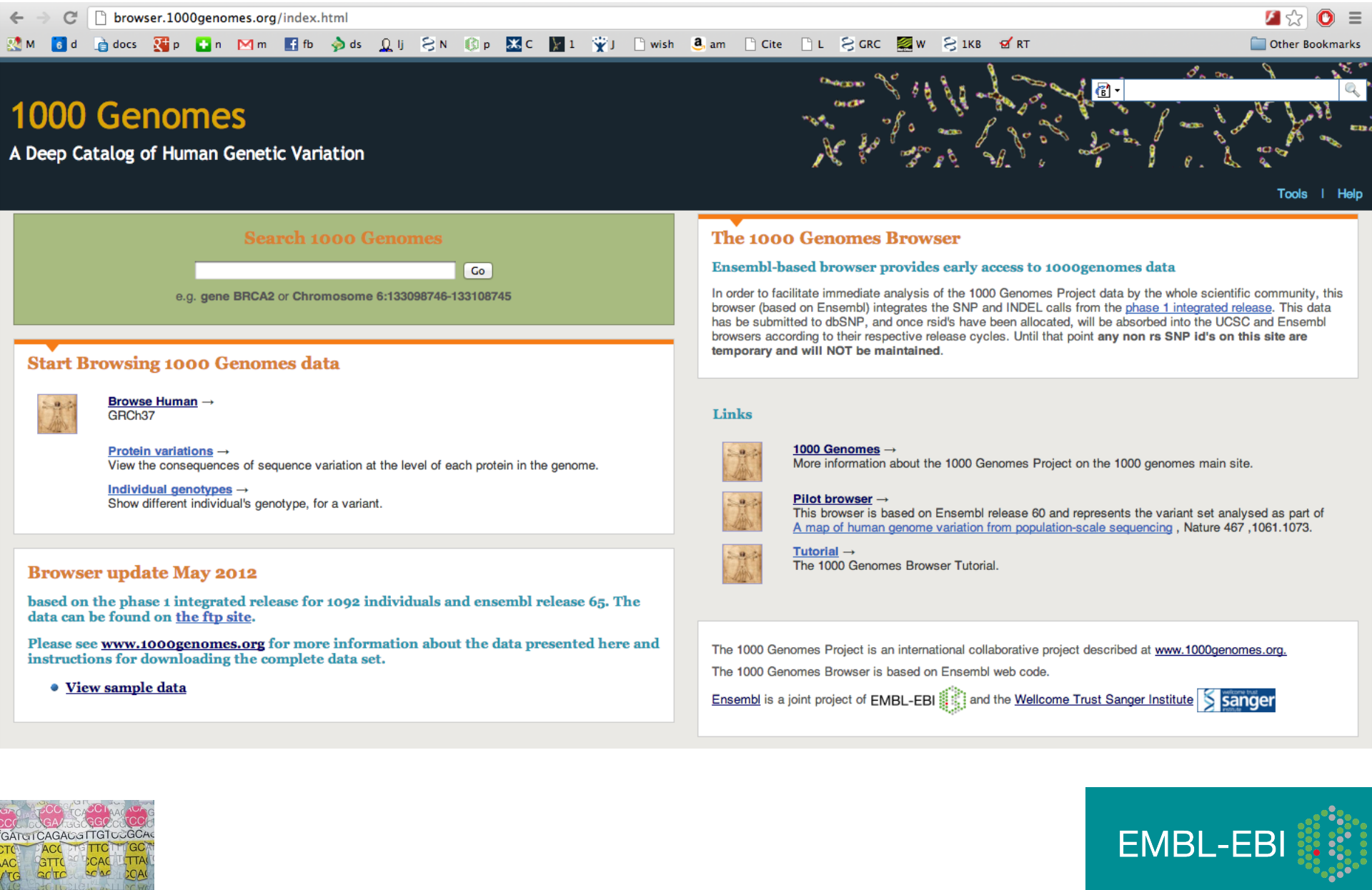
Amazon ECS

- FTP site mirror
- AMI image for our tutorial
- Preconfigured mount in CloudBioLinux
- http://www.1000genomes.org/using-1000-genomes-data-amazon-web-service-cloud

Browser

EMBL-EBI

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

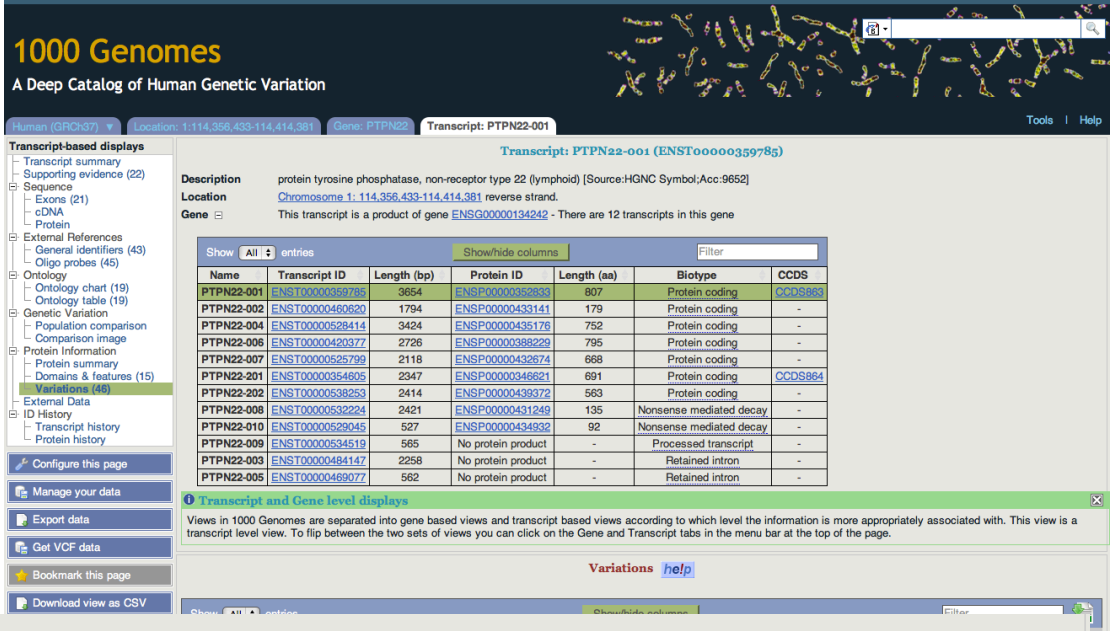


EMBL-EBI

Transcript Tab: Variations

Effect on Protein:

- SIFT
- PolyPhen



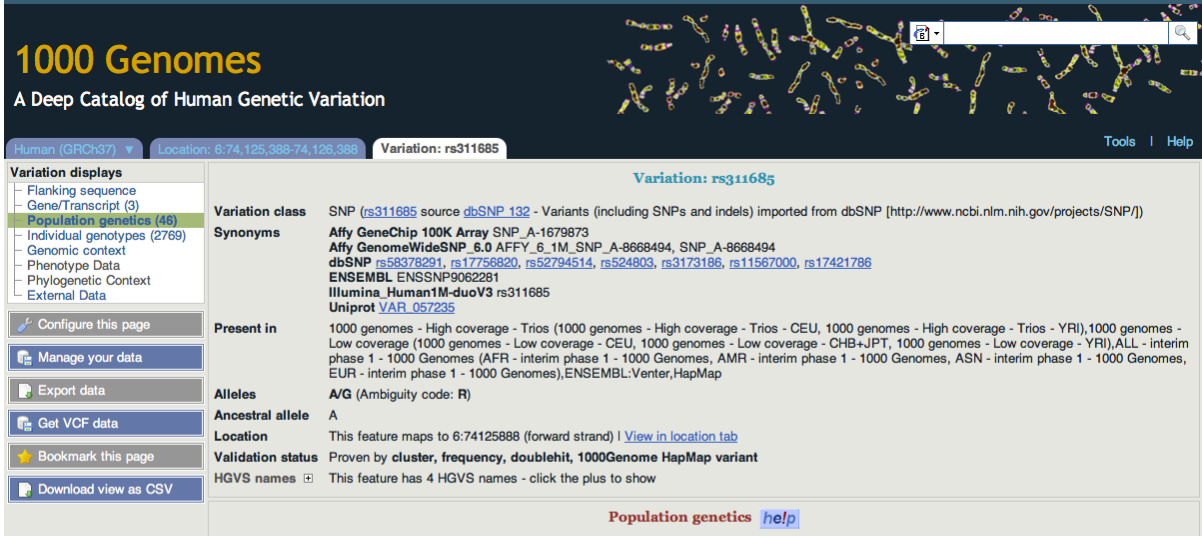
Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
PTPN22-001	ENST00000359785	3654	ENSP00000352833	807	Protein coding	CCDS863
PTPN22-002	ENST00000460620	1794	ENSP00000433141	179	Protein coding	-
PTPN22-004	ENST00000528414	3424	ENSP00000435176	752	Protein coding	-
PTPN22-006	ENST00000420377	2726	ENSP00000388229	795	Protein coding	-
PTPN22-007	ENST00000525799	2118	ENSP00000432674	668	Protein coding	-
PTPN22-201	ENST00000354605	2347	ENSP00000346621	691	Protein coding	CCDS864
PTPN22-202	ENST00000538253	2414	ENSP00000439372	563	Protein coding	-
PTPN22-008	ENST00000532224	2421	ENSP00000431249	135	Nonsense mediated decay	-
PTPN22-010	ENST00000529045	527	ENSP00000434932	92	Nonsense mediated decay	-
PTPN22-009	ENST00000534519	565	No protein product	-	Processed transcript	-
PTPN22-003	ENST00000484147	2258	No protein product	-	Retained intron	-
PTPN22-005	ENST00000469077	562	No protein product	-	Retained intron	-

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

Variation



1000 genomes alleles frequencies

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

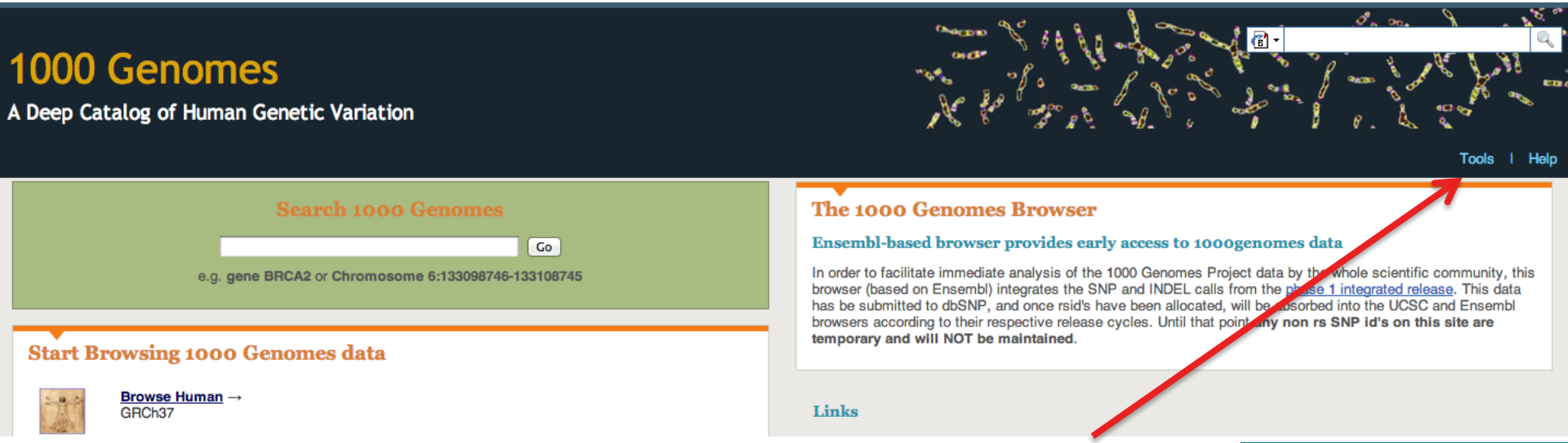
1000 genomes pilot

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	

Tools

- Data Slicer
- Variant Effect Predictor
- Variation Pattern Finder
- VCF to PED converter

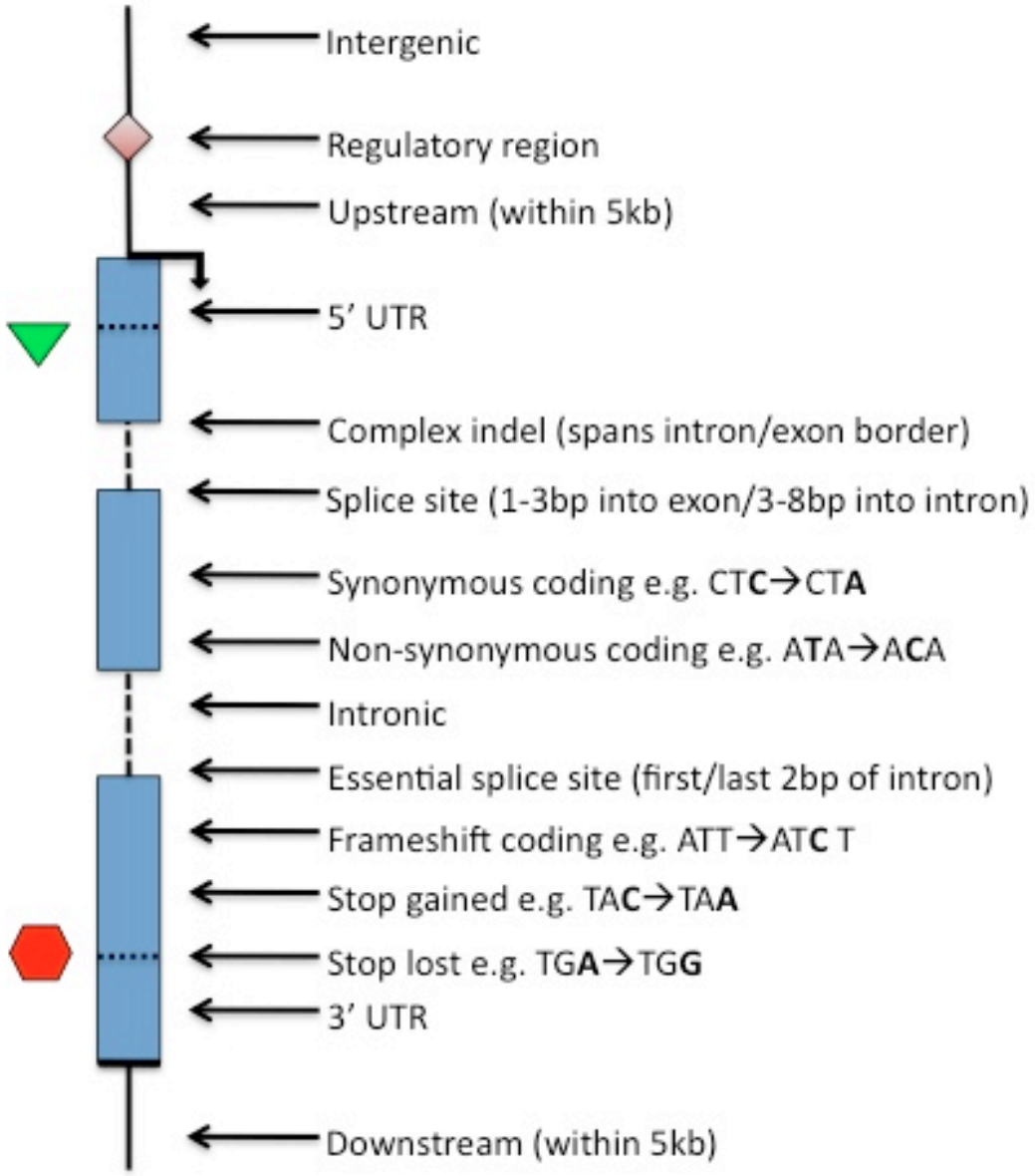


EMBL-EBI

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

Variant Effect Predictor



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

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	-	-	-	

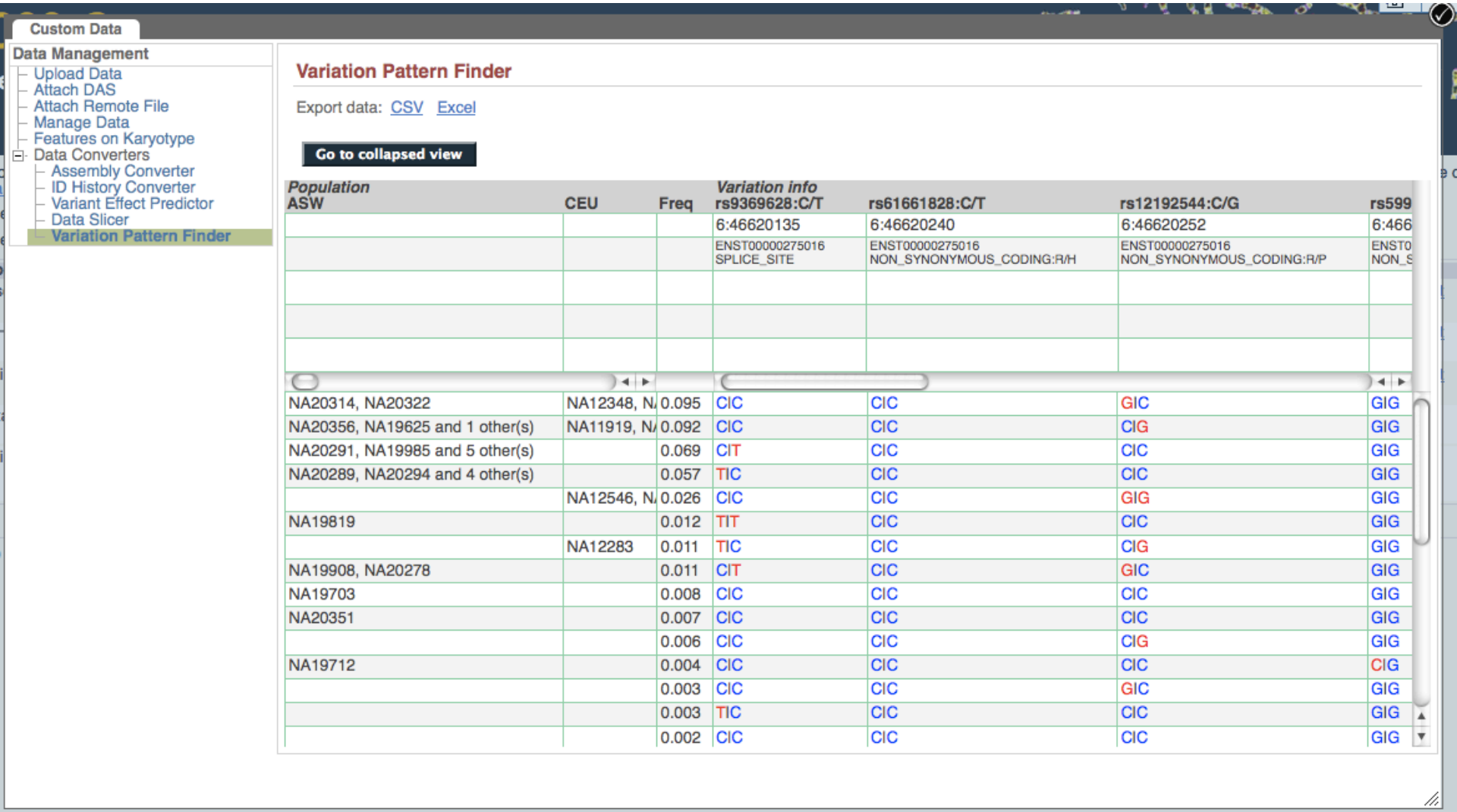
0204386	ENST00000491768	Transcript	UPSTREAM	-	-	-	-	-	1KG_6_31833357	-
0204386	ENST00000375631	Transcript	UPSTREAM	-	-	-	-	-	1KG_6_31833357	-
0204386	ENST00000479533	Transcript	UPSTREAM	-	-	-	-	-	1KG_6_31833357	-
0204385	ENST00000229729	Transcript	NON_SYNONYMOUS_CODING	1625	1604	535	R/H	cGc/cAc	1KG_6_31833357	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

- 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

EMBL-EBI



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	C|C	C|C	G|C	G|G
NA20356, NA19625 and 1 other(s)	NA11919, N/	0.092	C|C	C|C	C|G	G|G
NA20291, NA19985 and 5 other(s)		0.069	C|T	C|C	C|C	G|G
NA20289, NA20294 and 4 other(s)		0.057	T|C	C|C	C|C	G|G
	NA12546, N/	0.026	C|C	C|C	G|G	G|G
NA19819		0.012	T|T	C|C	C|C	G|G
	NA12283	0.011	T|C	C|C	C|G	G|G
NA19908, NA20278		0.011	C|T	C|C	G|C	G|G
NA19703		0.008	C|C	C|C	C|C	G|G
NA20351		0.007	C|C	C|C	C|C	G|G
		0.006	C|C	C|C	C|G	G|G
NA19712		0.004	C|C	C|C	C|C	C|G
		0.003	C|C	C|C	G|C	G|G
		0.003	T|C	C|C	C|C	G|G
		0.002	C|C	C|C	C|C	G|G

EMBL-EBI

More Help

- http://www.1000genomes.org/using-1000-genomes-data
- ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/1000genomes_browser_main_project_20110521/The_1000_Genomes_Browser_Tutorial.ensembl_65.doc
- http://www.1000genomes.org/ashg-2012-poster
- http://www.1000genomes.org/faq
- Twitter @1000genomes.org
- 1000announce@1000genomes.org
- info@1000genomes.org

EMBL-EBI

Questions?

info@1000genomes.org

EMBL-EBI