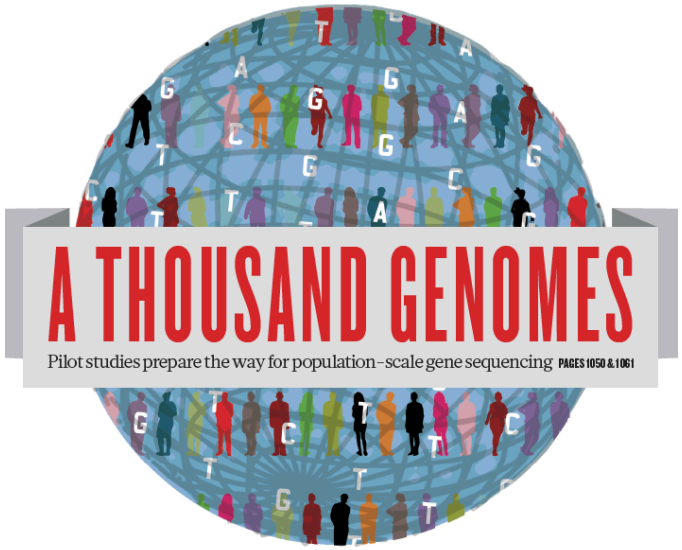


The 1000 Genomes Project: A Tutorial

Browser and Tools Exercises

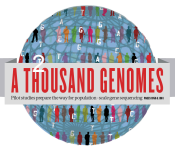
EMBL-EBI

Browser and Tools Exercises

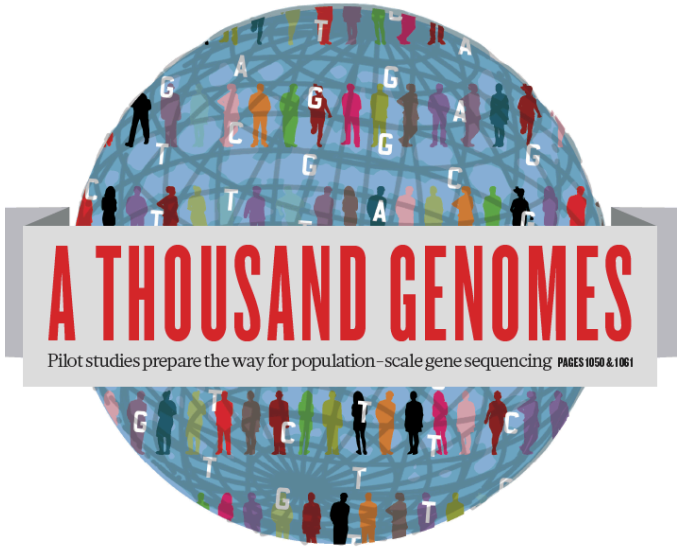
These exercises use the 1000 genomes website and browser and tools built around it.

http://www.1000genomes.org

http://browser.1000genomes.org



EMBL-EBI



The 1000 Genomes Project:

Exercise 1: Finding Data and viewing data on the 1000 genomes ftp site

EMBL-EBI

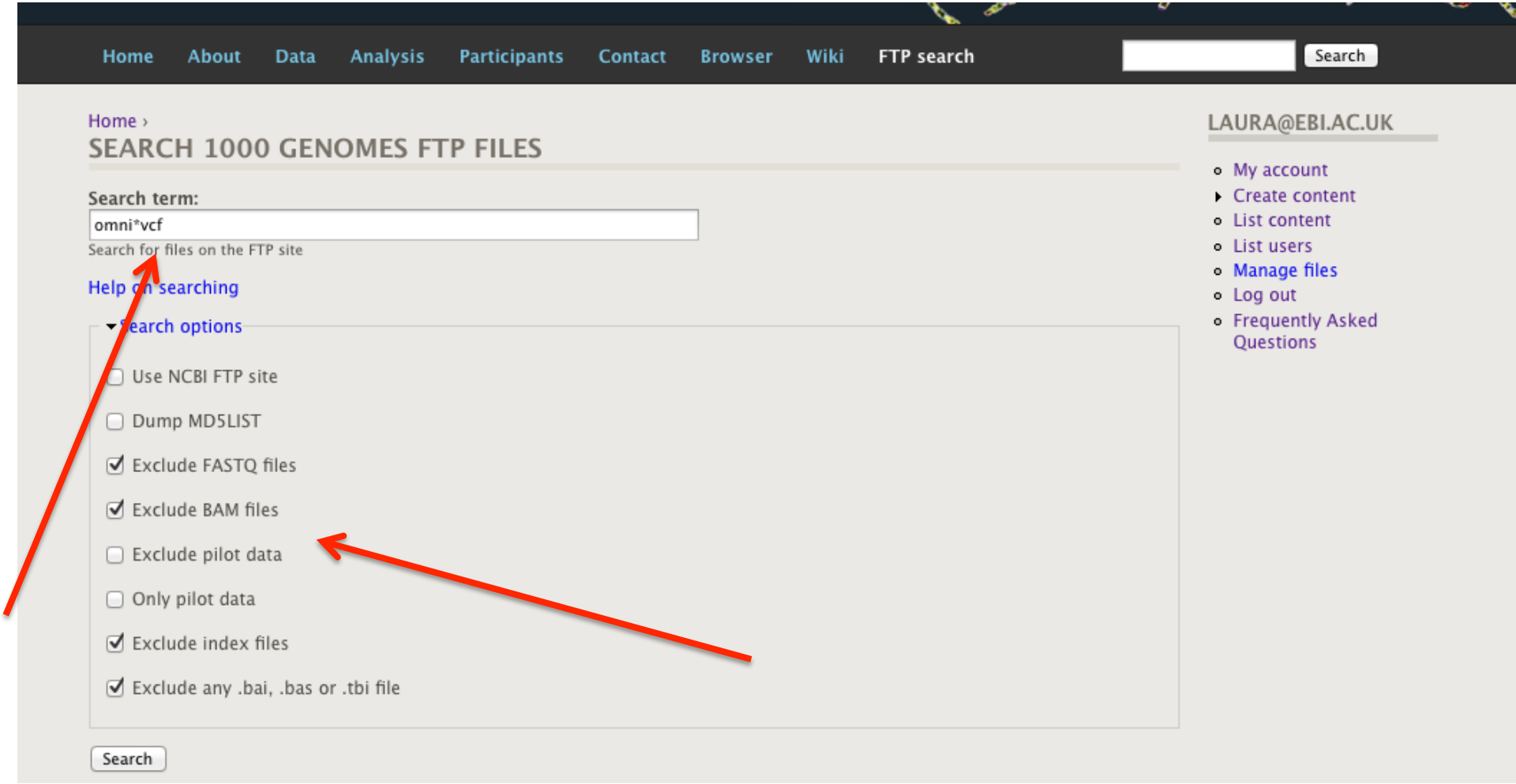
Finding Data: Exercise

1. Find what VCF files we have containing genotypes from the Illumina Omni platform.
   http://www.1000genomes.org/ftpsearch

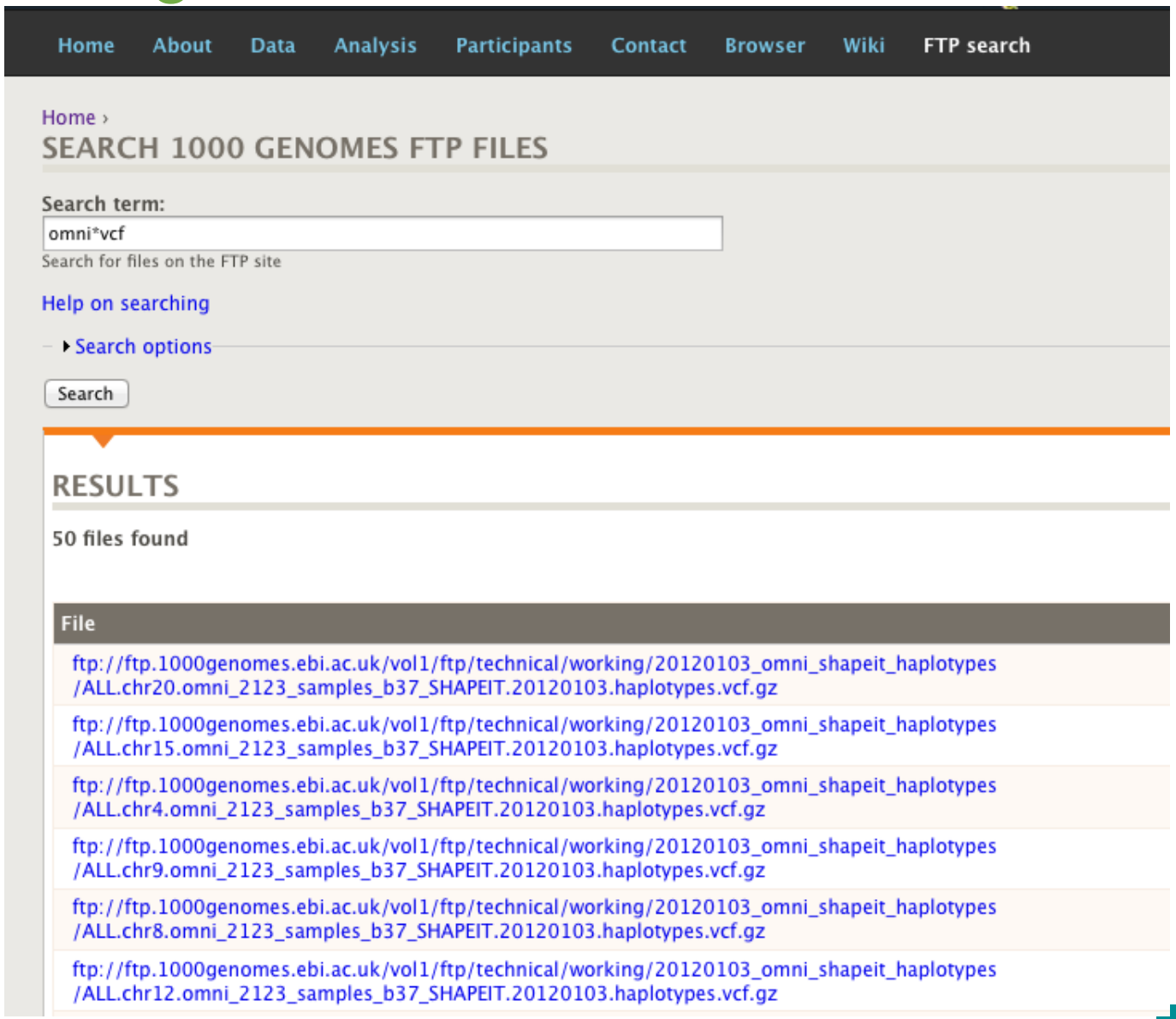
2. Find the FAQ question which gives you instructions on how to get a sub-section of a VCF file. The Search Box is on the top right hand corner of any website page.
   http://www.1000genomes.org/

Finding Data: Exercise

1. Finding Omni VCF Files

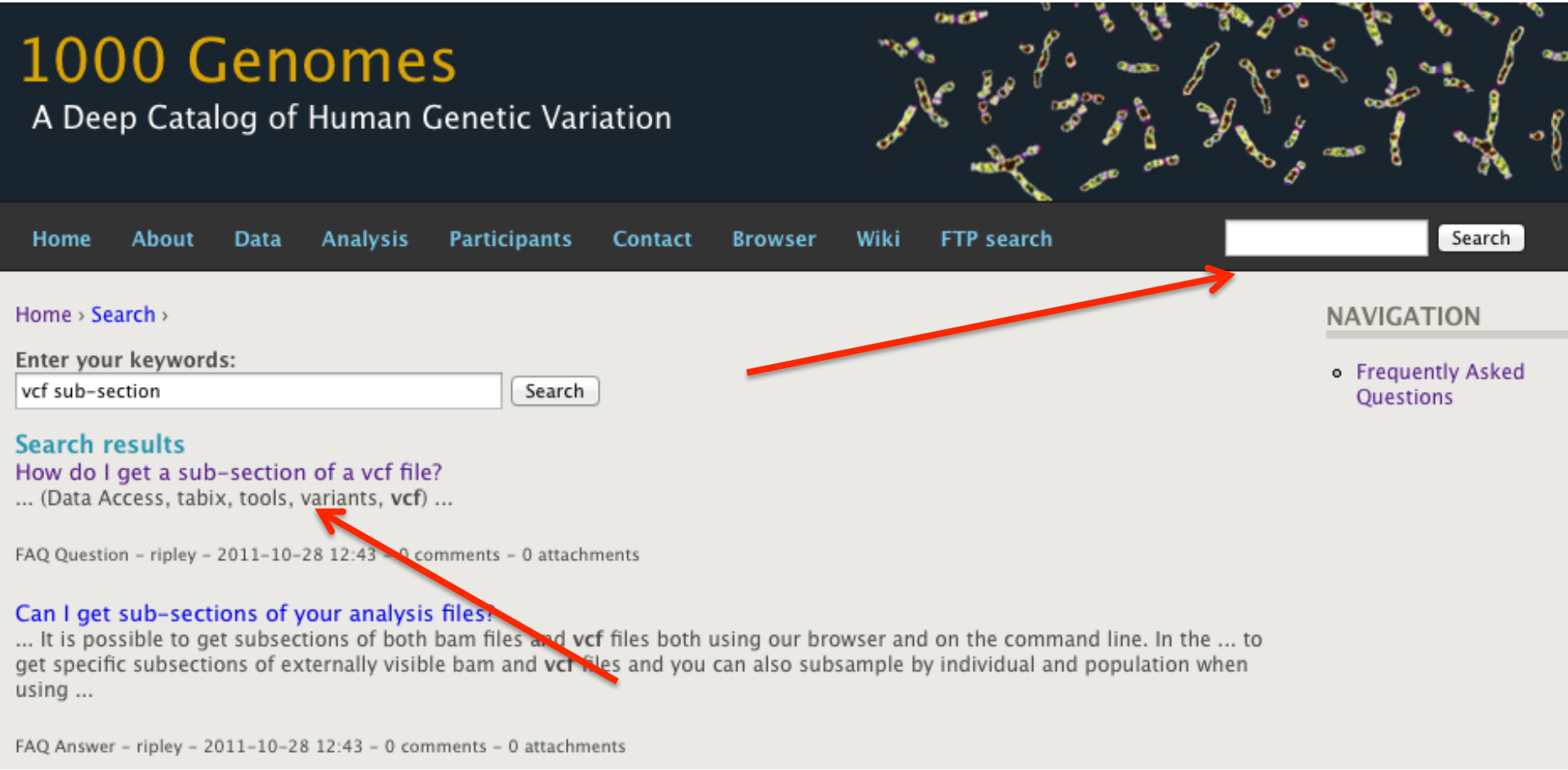


Finding Data: Exercise



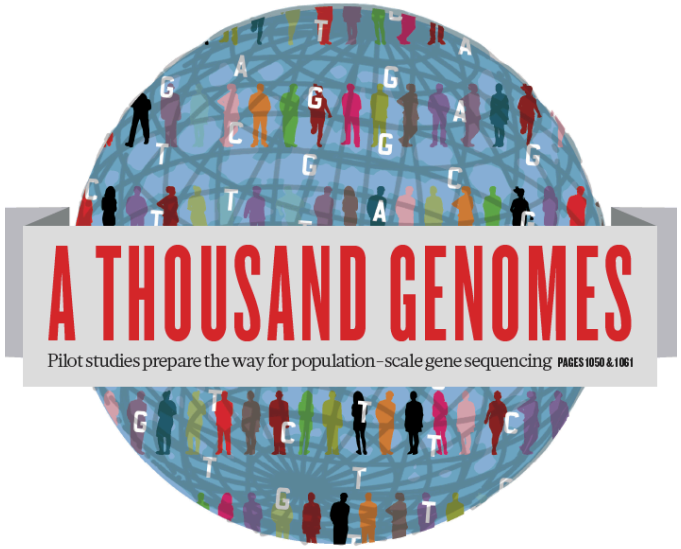
Finding Data: Exercise

- Finding help on getting sub-sections of VCF files



A THOUSAND GENOMES

EMBL-EBI



The 1000 Genomes Project:

Exercise 2: Finding Variation Using the Browser

EMBL-EBI

Exercise: Finding Variation Using the Browser

- Find the variant rs45562238

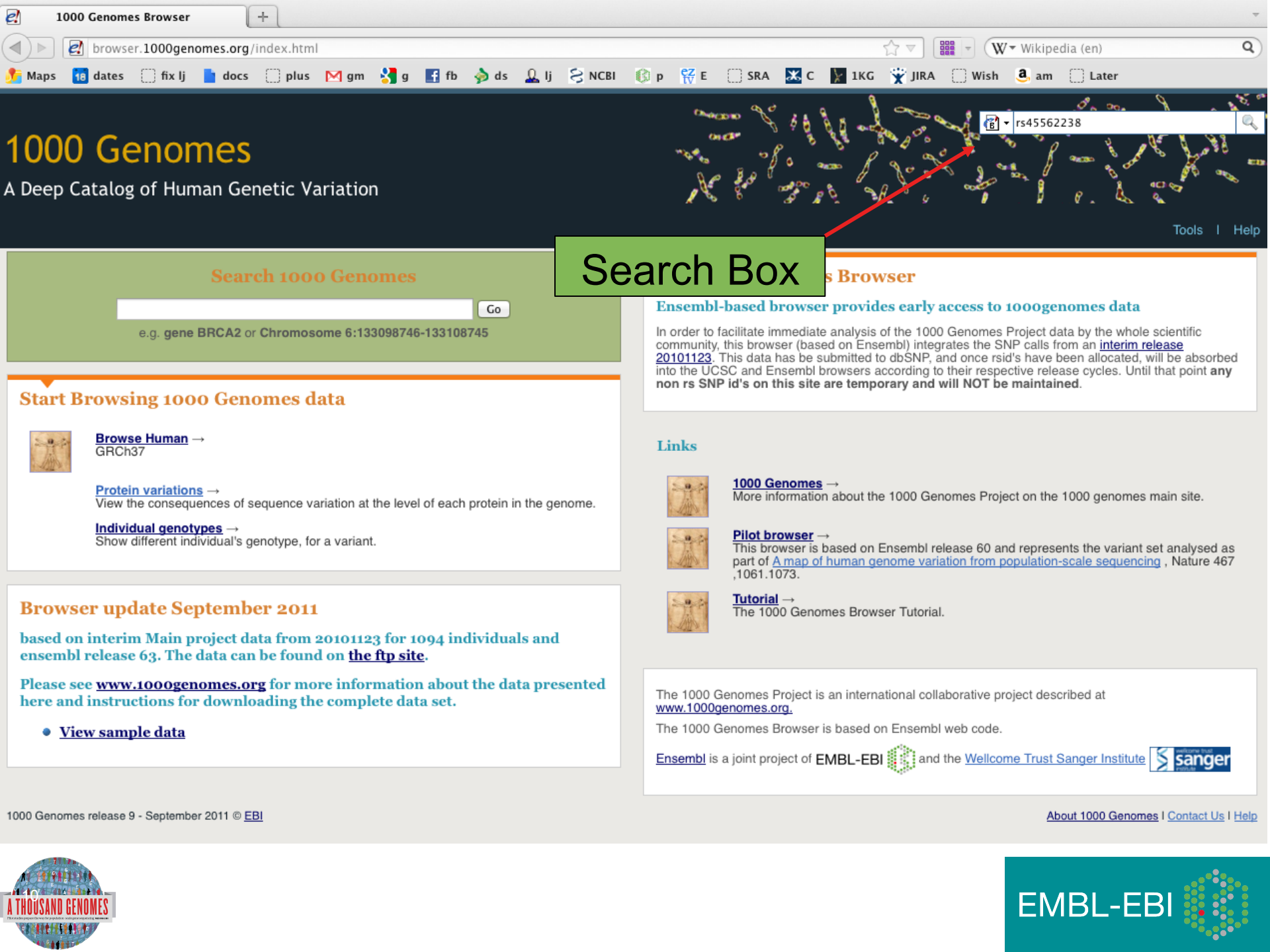
  http://browser.1000genomes.org

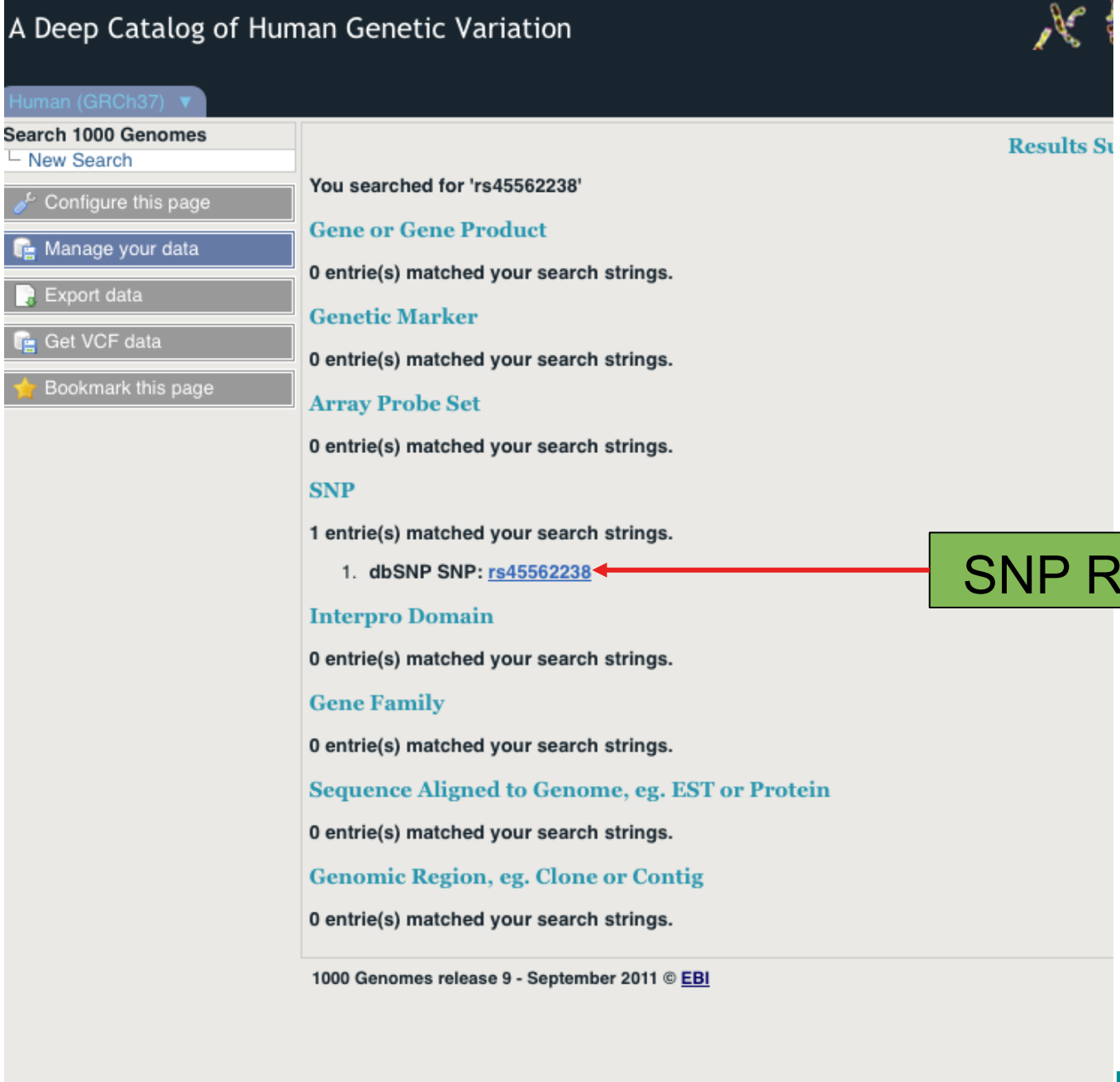
- In which 1000 Genomes Populations was it detected?

- What are its allele frequencies?

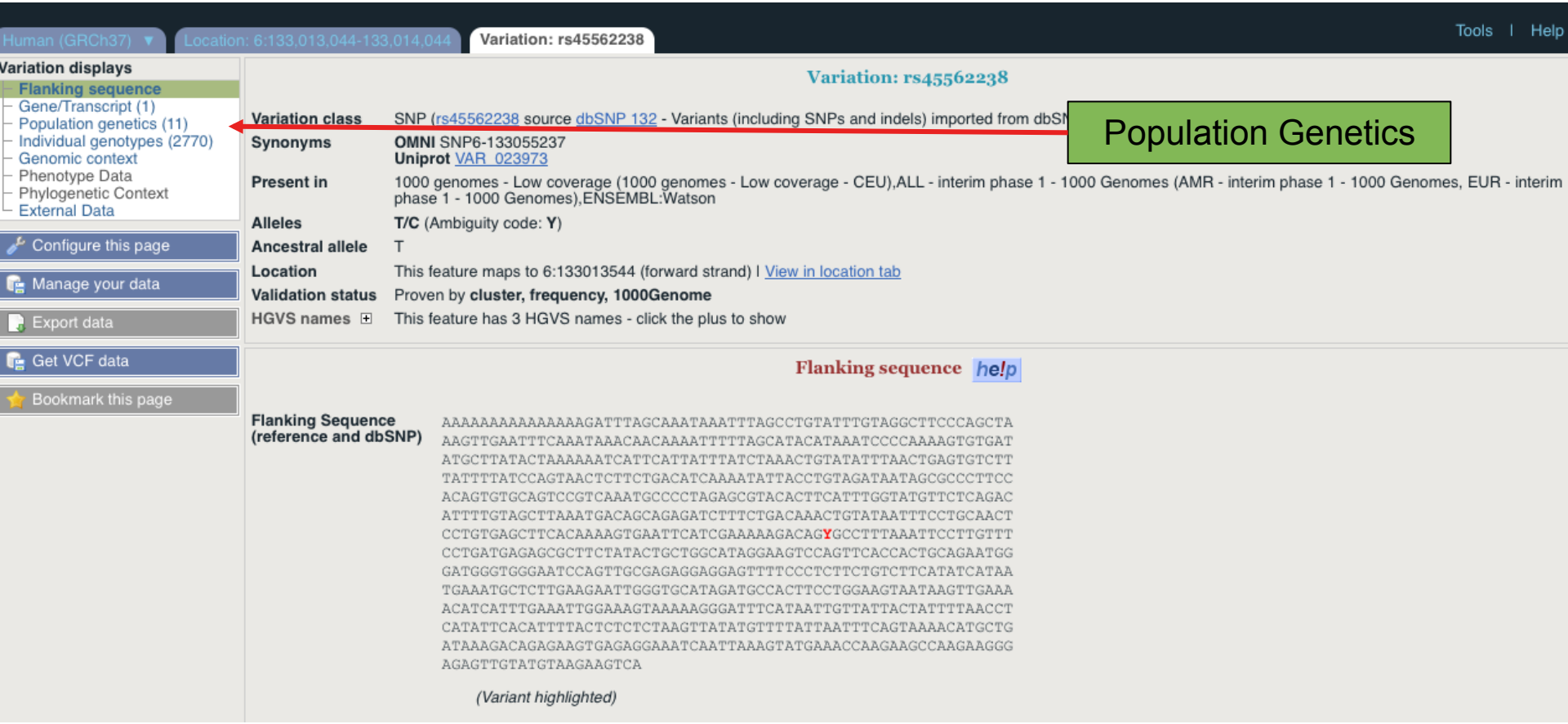
- In which gene is the variant found?

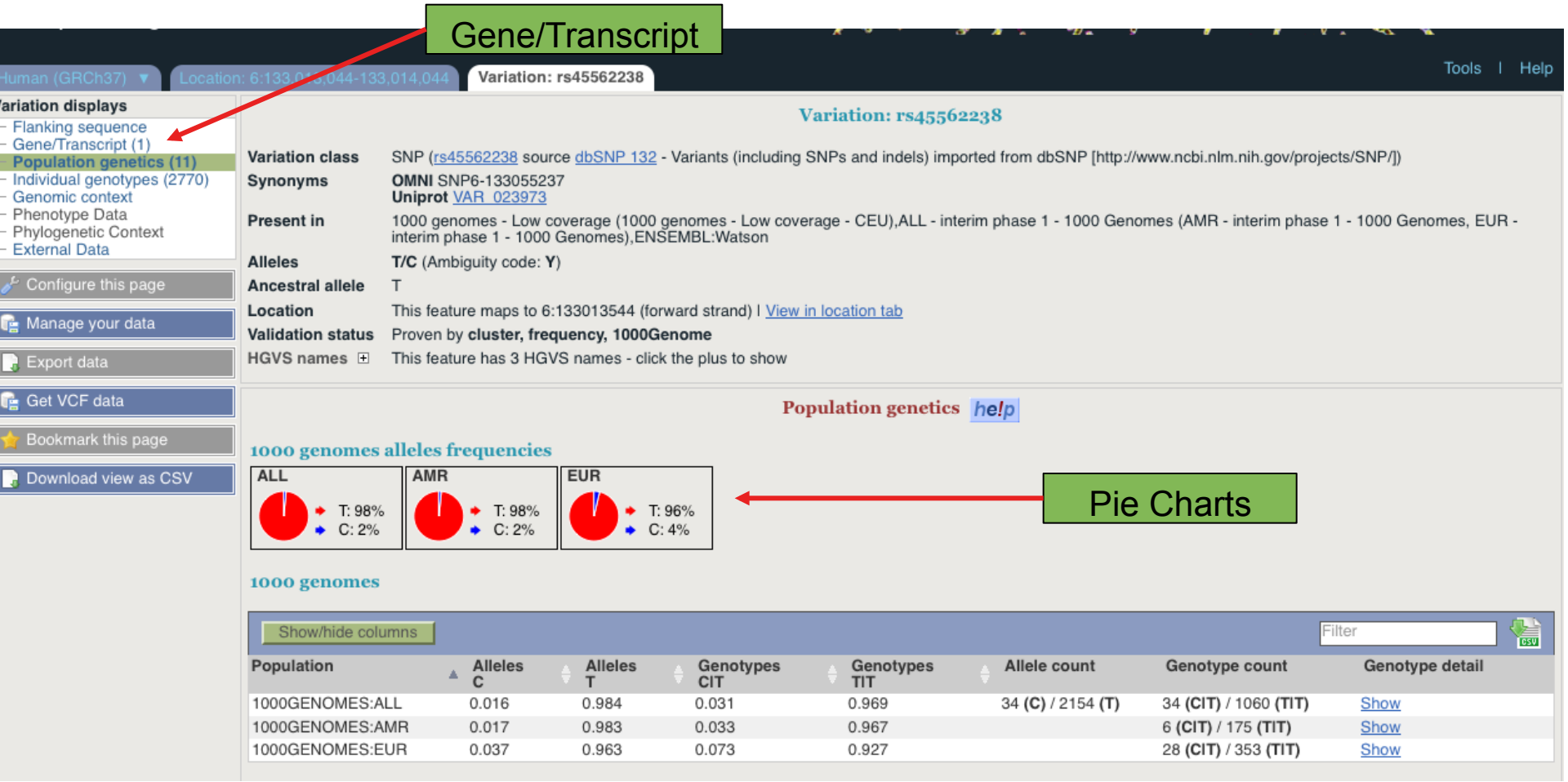
EMBL-EBI


1000 Genomes Browser
browser.1000genomes.org/index.html
Wikipedia (en)
Maps
dates
fix lj
docs
plus
gm
g
fb
ds
lj
NCBI
p
E
SRA
C
1KG
JIRA
Wish
am
Later
1000 Genomes
A Deep Catalog of Human Genetic Variation
rs45562238
Tools | Help
Search Box
Search 1000 Genomes
Go
e.g. gene BRCA2 or Chromosome 6:133098746-133108745
Start Browsing 1000 Genomes data
Browse Human →
GRCh37
Protein variations →
View the consequences of sequence variation at the level of each protein in the genome.
Individual genotypes →
Show different individual's genotype, for a variant.
Browser update September 2011
based on interim Main project data from 20101123 for 1094 individuals and ensembl release 63. The data can be found on the ftp site.
Please see www.1000genomes.org for more information about the data presented here and instructions for downloading the complete data set.
View sample data
s Browser
Ensembl-based browser provides early access to 1000genomes data
In order to facilitate immediate analysis of the 1000 Genomes Project data by the whole scientific community, this browser (based on Ensembl) integrates the SNP calls from an interim release 20101123. This data has be submitted to dbSNP, and once rsid's have been allocated, will be absorbed into the UCSC and Ensembl browsers according to their respective release cycles. Until that point any non rs SNP id's on this site are temporary and will NOT be maintained.
Links
1000 Genomes →
More information about the 1000 Genomes Project on the 1000 genomes main site.
Pilot browser →
This browser is based on Ensembl release 60 and represents the variant set analysed as part of A map of human genome variation from population-scale sequencing , Nature 467 ,1061.1073.
Tutorial →
The 1000 Genomes Browser Tutorial.
The 1000 Genomes Project is an international collaborative project described at www.1000genomes.org.
The 1000 Genomes Browser is based on Ensembl web code.
Ensembl is a joint project of EMBL-EBI and the Wellcome Trust Sanger Institute
1000 Genomes release 9 - September 2011 © EBI
About 1000 Genomes | Contact Us | Help
A THOUSAND GENOMES
EMBL-EBI



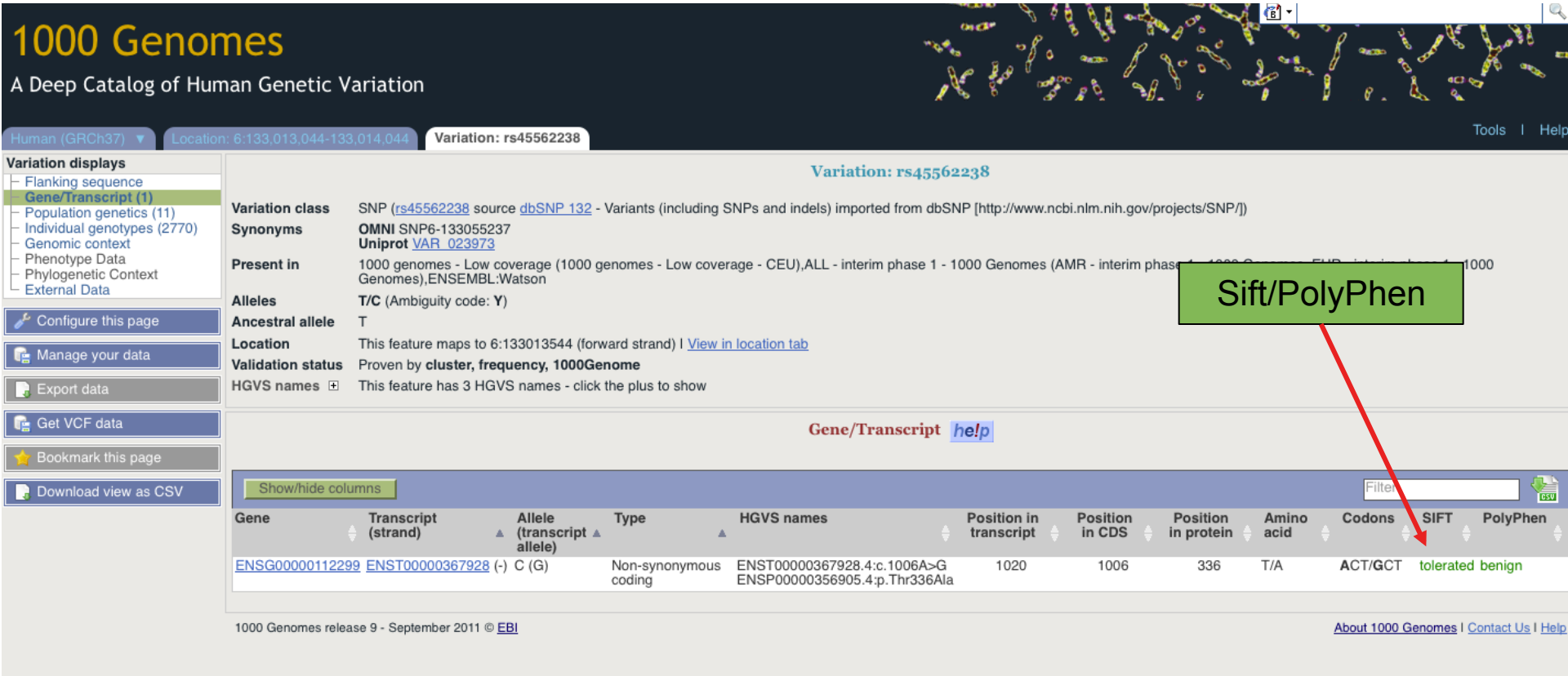
SNP Result





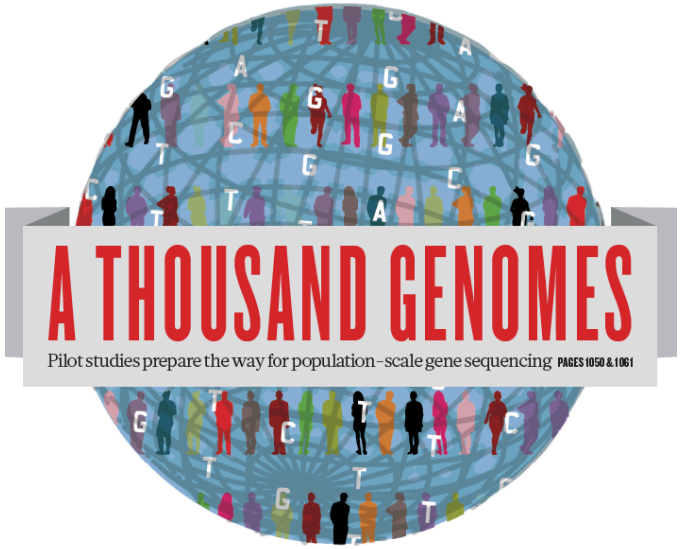
A THOUSAND GENOMES

EMBL-EBI



A THOUSAND GENOMES

EMBL-EBI



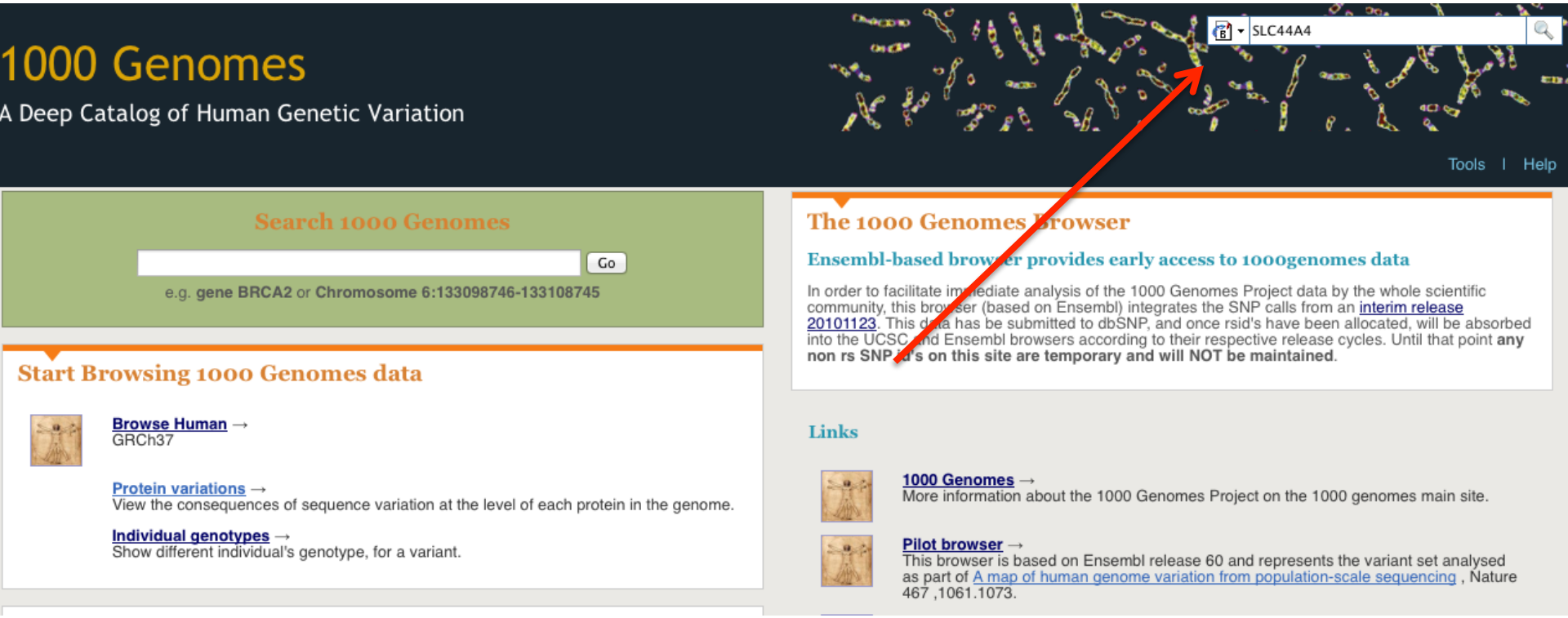
The 1000 Genomes Project:

Exercise 3: Using 1000 Genomes Tools

EMBL-EBI

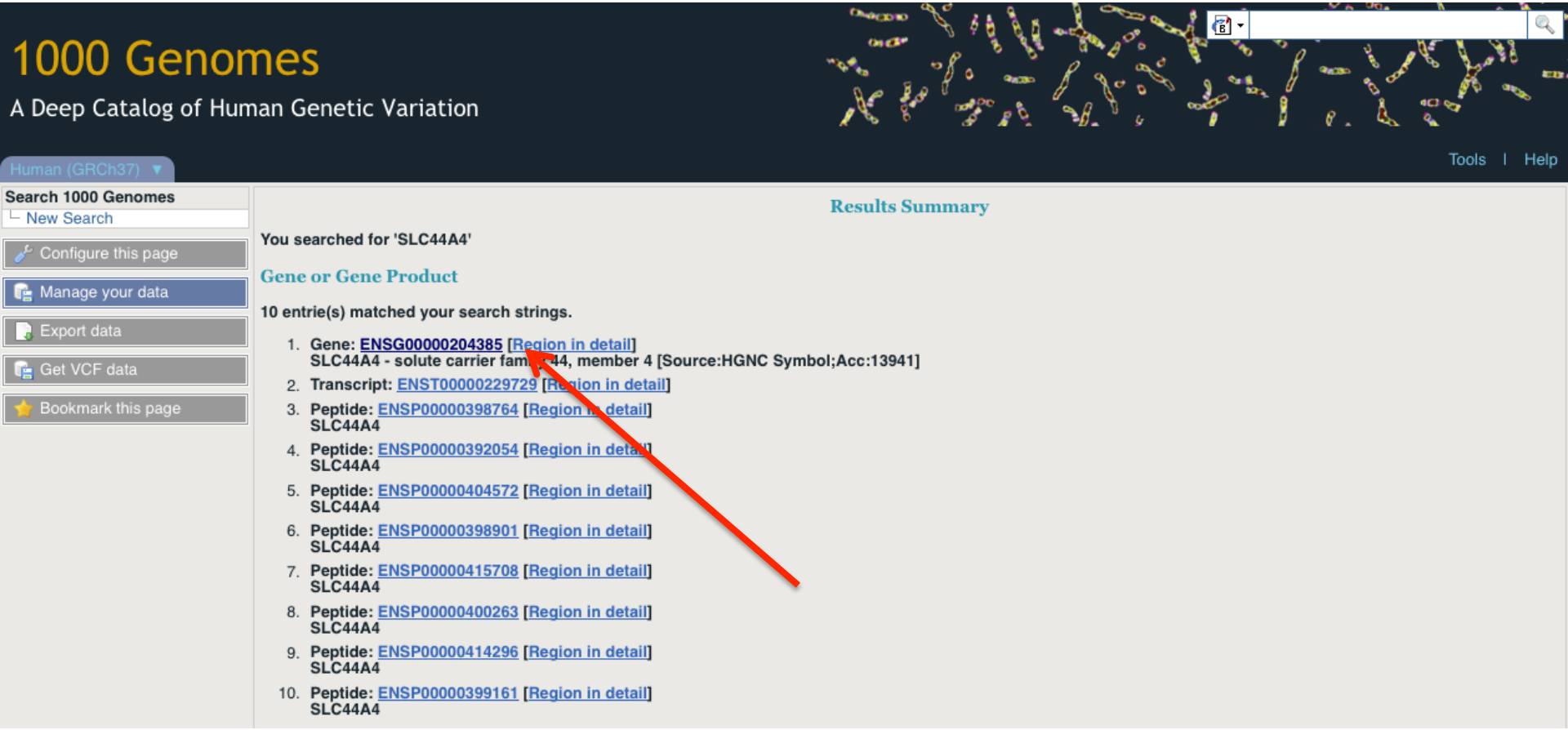
Exercise: Using 1000 Genomes Tools

- Find the gene SLC44A4 using the search box on http://browser.1000genomes.org
- Get a VCF file for this Gene using the Get VCF button.
- Uncompress this file
  - You can get a copy at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120120_1000genomes_tutorial/6.31830969-31846823.ALL.chr6.phase1.projectConsensus.genotypes.vcf
- Use this file with the Variant Effect Predictor
  - http://browser.1000genomes.org/Homo_sapiens/UserData/UploadVariations
- Do any of the variants have deleterious effects according to SIFT or PolyPhen
- Use the example url on the page and the coordinates 6:31830700-31840700 with the Variation Pattern Finder
  - http://browser.1000genomes.org/Homo_sapiens/UserData/VariationsMapVCF



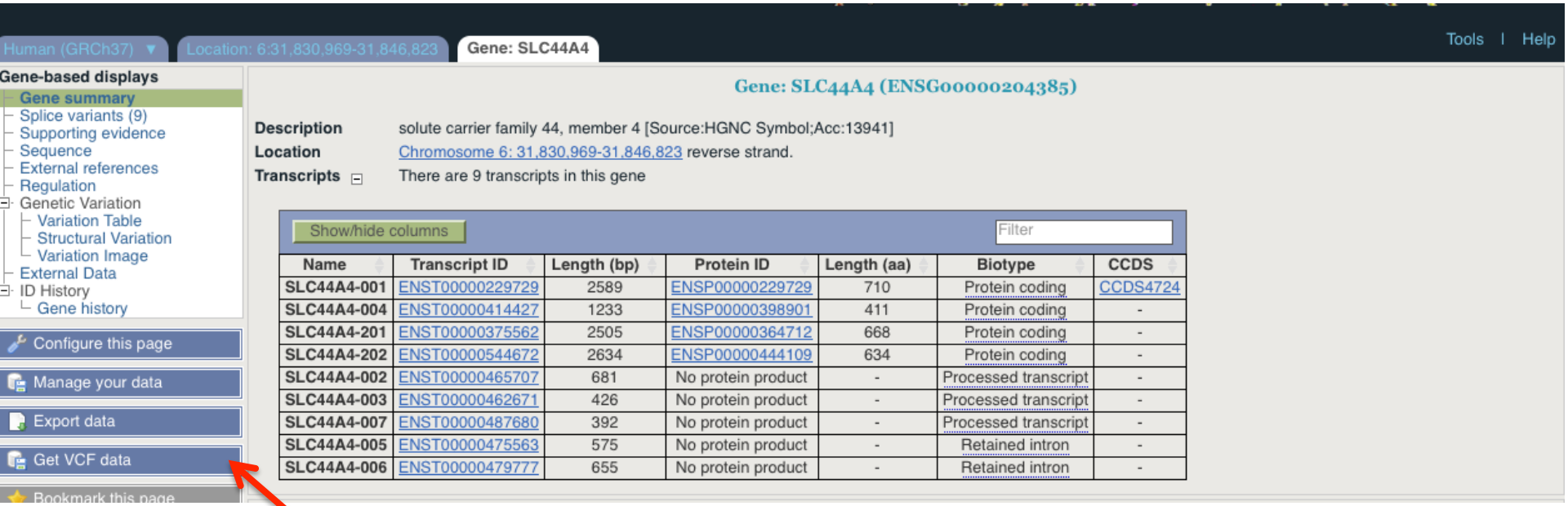
A THOUSAND GENOMES

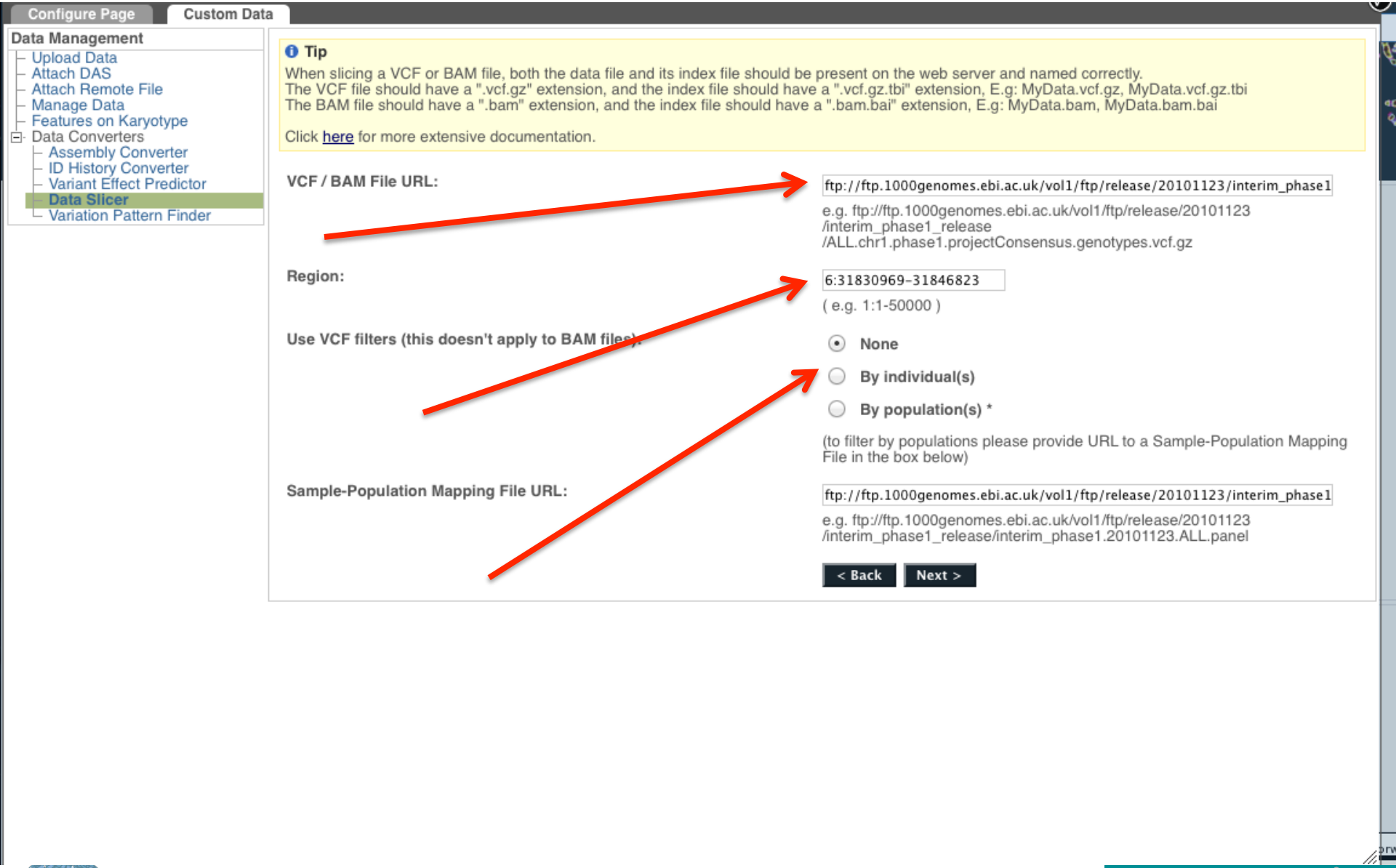
EMBL-EBI



A THOUSAND GENOMES

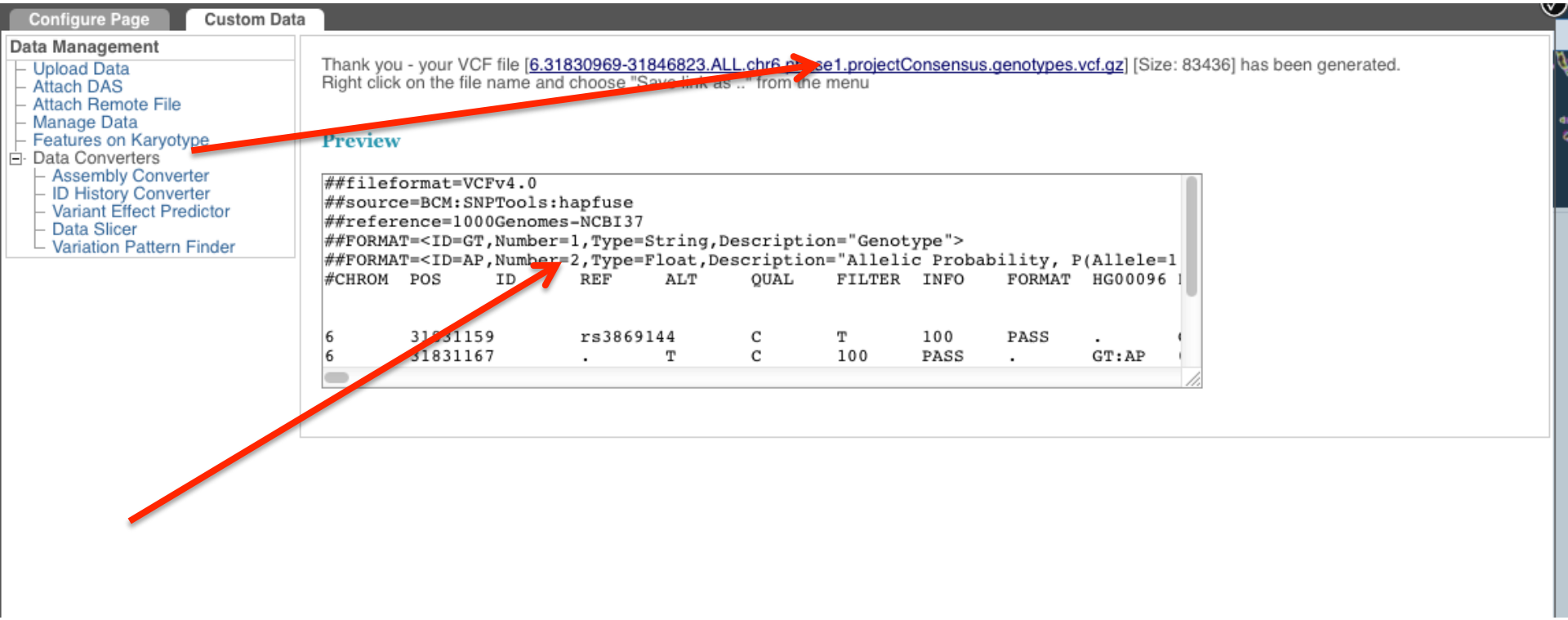
EMBL-EBI

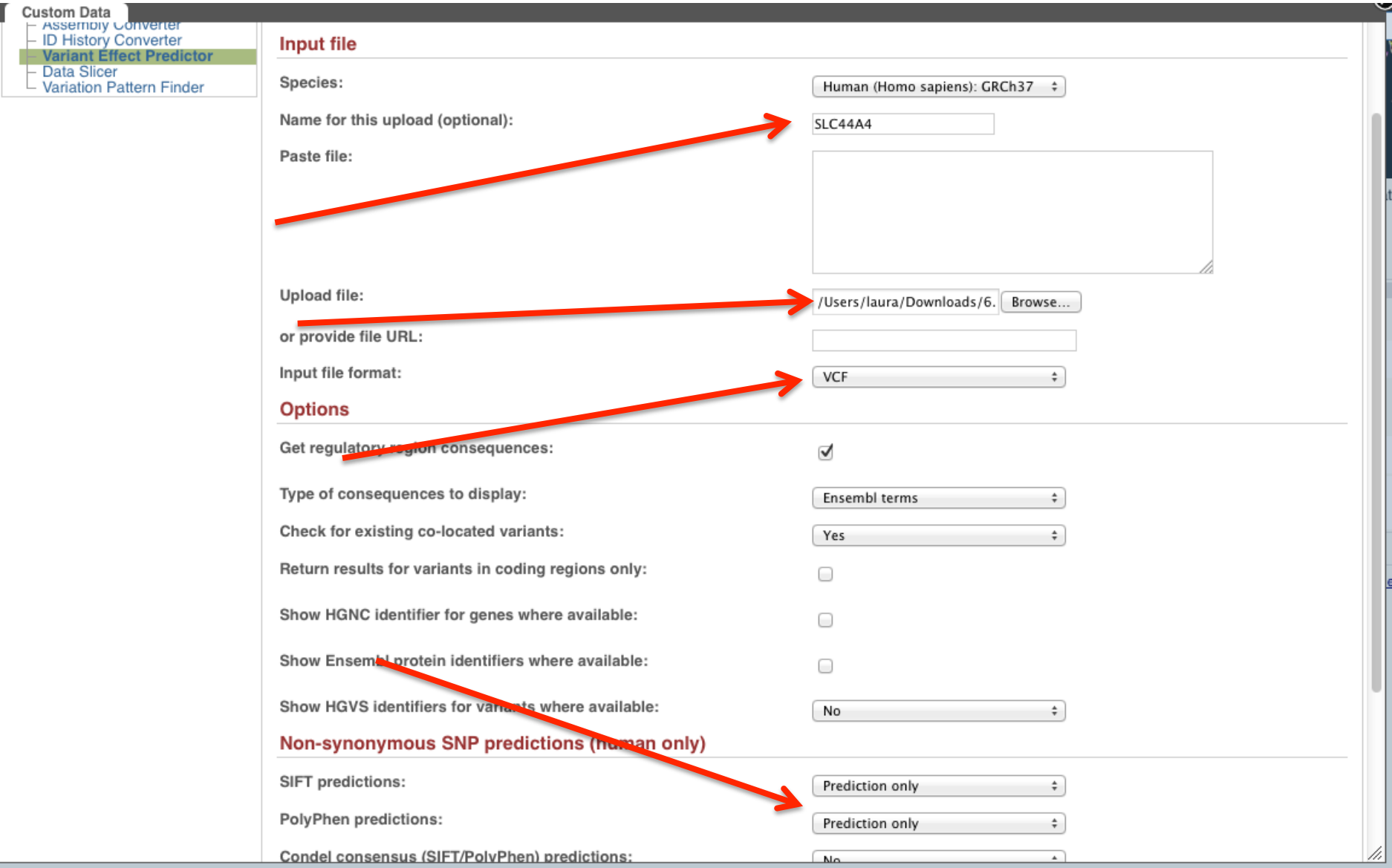

Human (GRCh37)
Location: 6:31,830,969-31,846,823
Gene: SLC44A4
Tools | Help
Gene-based displays
Gene summary
Splice variants (9)
Supporting evidence
Sequence
External references
Regulation
Genetic Variation
Variation Table
Structural Variation
Variation Image
External Data
ID History
Gene history
Configure this page
Manage your data
Export data
Get VCF data
Bookmark this page
Gene: SLC44A4 (ENSG00000204385)
Description
solute carrier family 44, member 4 [Source:HGNC Symbol;Acc:13941]
Location
Chromosome 6: 31,830,969-31,846,823 reverse strand.
Transcripts
There are 9 transcripts in this gene
Show/hide columns
Filter
Name
Transcript ID
Length (bp)
Protein ID
Length (aa)
Biotype
CCDS
SLC44A4-001
ENST00000229729
2589
ENSP00000229729
710
Protein coding
CCDS4724
SLC44A4-004
ENST00000414427
1233
ENSP00000398901
411
Protein coding
-
SLC44A4-201
ENST00000375562
2505
ENSP00000364712
668
Protein coding
-
SLC44A4-202
ENST00000544672
2634
ENSP00000444109
634
Protein coding
-
SLC44A4-002
ENST00000465707
681
No protein product
-
Processed transcript
-
SLC44A4-003
ENST00000462671
426
No protein product
-
Processed transcript
-
SLC44A4-007
ENST00000487680
392
No protein product
-
Processed transcript
-
SLC44A4-005
ENST00000475563
575
No protein product
-
Retained intron
-
SLC44A4-006
ENST00000479777
655
No protein product
-
Retained intron
-



A THOUSAND GENOMES

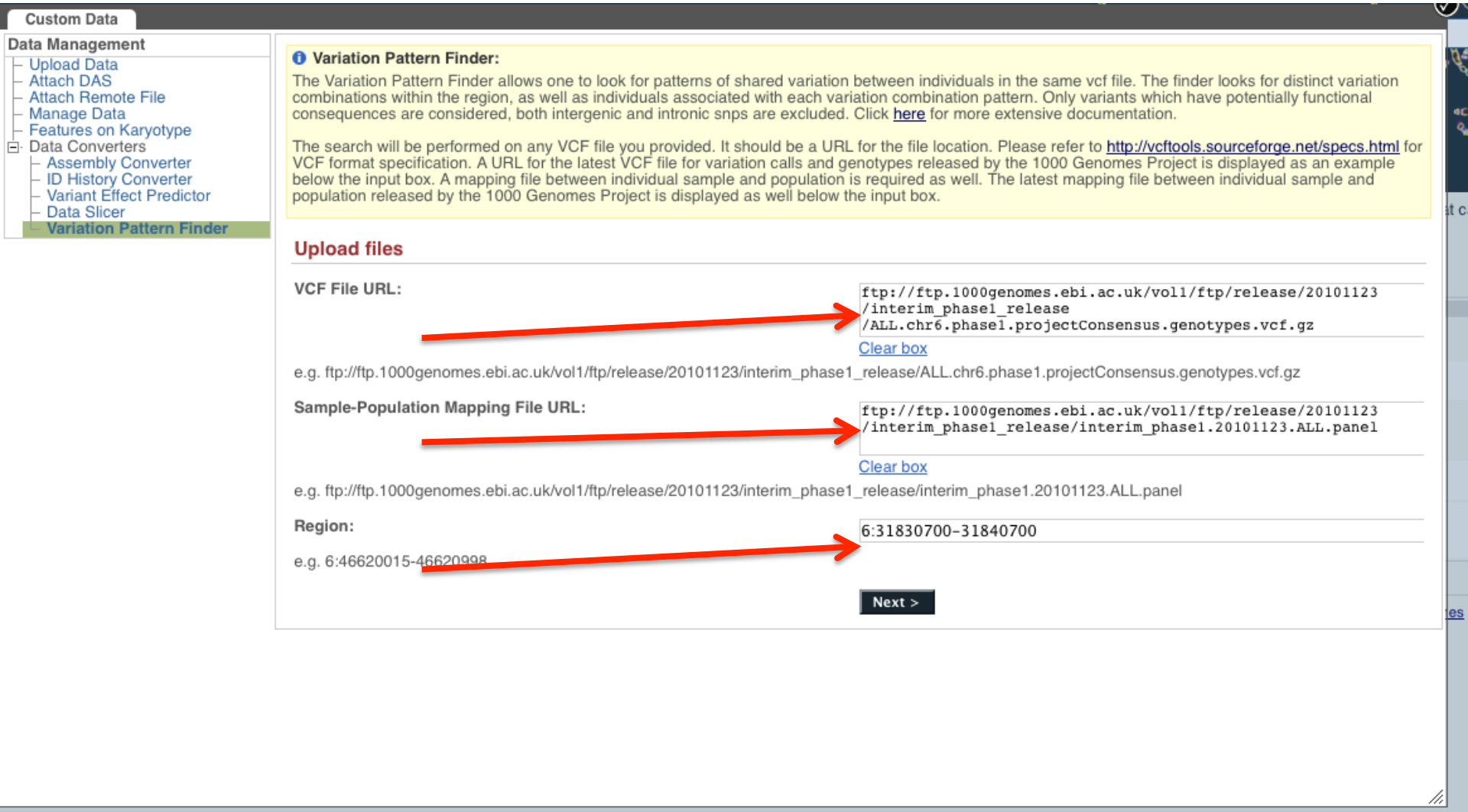
EMBL-EBI


Configure Page
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
Thank you - your VCF file [6.31830969-31846823.ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz] [Size: 83436] has been generated.
Right click on the file name and choose "Save link as .." from the menu
Preview
##fileformat=VCFv4.0
##source=BCM:SNPTools:hapfuse
##reference=1000Genomes-NCBI37
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AP,Number=2,Type=Float,Description="Allelic Probability, P(Allele=1
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HG00096
6 31831159 rs3869144 C T 100 PASS . (
6 31831167 . T C 100 PASS . GT:AP (


Custom Data
Assembly Converter
ID History Converter
Variant Effect Predictor
Data Slicer
Variation Pattern Finder
Input file
Species:
Human (Homo sapiens): GRCh37
Name for this upload (optional):
SLC44A4
Paste file:
Upload file:
/Users/laura/Downloads/6.
Browse...
or provide file URL:
Input file format:
VCF
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:
Prediction only
PolyPhen predictions:
Prediction only
Condel consensus (SIFT/PolyPhen) predictions:

6_31833249_A/G	6:31833249	G	ENSG00000204385	ENST00000487680	Transcript	UPSTREAM	-	-	-	-	-	-	-	-	-
6_31833249_A/G	6:31833249	G	ENSG00000204385	ENST00000414427	Transcript	DOWNSTREAM	-	-	-	-	-	-	-	-	-
6_31833249_A/G	6:31833249	G	ENSG00000204385	ENST00000479777	Transcript	DOWNSTREAM	-	-	-	-	-	-	-	-	-
6_31833249_A/G	6:31833249	G	ENSG00000204385	ENST00000475563	Transcript	DOWNSTREAM	-	-	-	-	-	-	-	-	-
6_31833357_C/T	6:31833357	T	-	ENSR00000487922	RegulatoryFeature	REGULATORY_REGION	-	-	-	-	-	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204386	ENST00000495807	Transcript	UPSTREAM	-	-	-	-	-	-	-	-	-
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	R/H	-	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000375563	Transcript	NON_SYNONYMOUS_CODING	1544	1478	493	R/H	-	-	-	-	-
6_31833357_C/T	6:31833357	T	ENSG00000204385	ENST00000544672	Transcript	NON_SYNONYMOUS_CODING	1673	1376	459	R/H	-	-	-	-	-
6_31833357_C/T	6:31833357	T	E	-	-	-	1KG_6_31833357	-							-
6_31833357_C/T	6:31833357	T	E												-
6_31833357_C/T	6:31833357	T	E	-	-	-	1KG_6_31833357	-							-
6_31833357_C/T	6:31833357	T	E	535	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=probably_damaging; Condel=deleterious							-
6_31833612_C/G	6:31833612	G													-
6_31833612_C/G	6:31833612	G	E												-
6_31833612_C/G	6:31833612	G	E												-
				493	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=possibly_damaging; Condel=deleterious							
				459	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=probably_damaging; Condel=deleterious							
				-	-	-	1KG_6_31833357	-							
				-	-	-	1KG_6_31833357	-							
				-	-	-	1KG_6_31833357	-							

EMBL-EBI


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
Variation Pattern Finder:
The Variation Pattern Finder allows one to look for patterns of shared variation between individuals in the same vcf file. The finder looks for distinct variation combinations within the region, as well as individuals associated with each variation combination pattern. Only variants which have potentially functional consequences are considered, both intergenic and intronic snps are excluded. Click here for more extensive documentation.
The search will be performed on any VCF file you provided. It should be a URL for the file location. Please refer to http://vcftools.sourceforge.net/specs.html for VCF format specification. A URL for the latest VCF file for variation calls and genotypes released by the 1000 Genomes Project is displayed as an example below the input box. A mapping file between individual sample and population is required as well. The latest mapping file between individual sample and population released by the 1000 Genomes Project is displayed as well below the input box.
Upload files
VCF File URL:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz
Clear box
e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz
Sample-Population Mapping File URL:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel
Clear box
e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel
Region:
6:31830700-31840700
e.g. 6:46620015-46620998
Next >

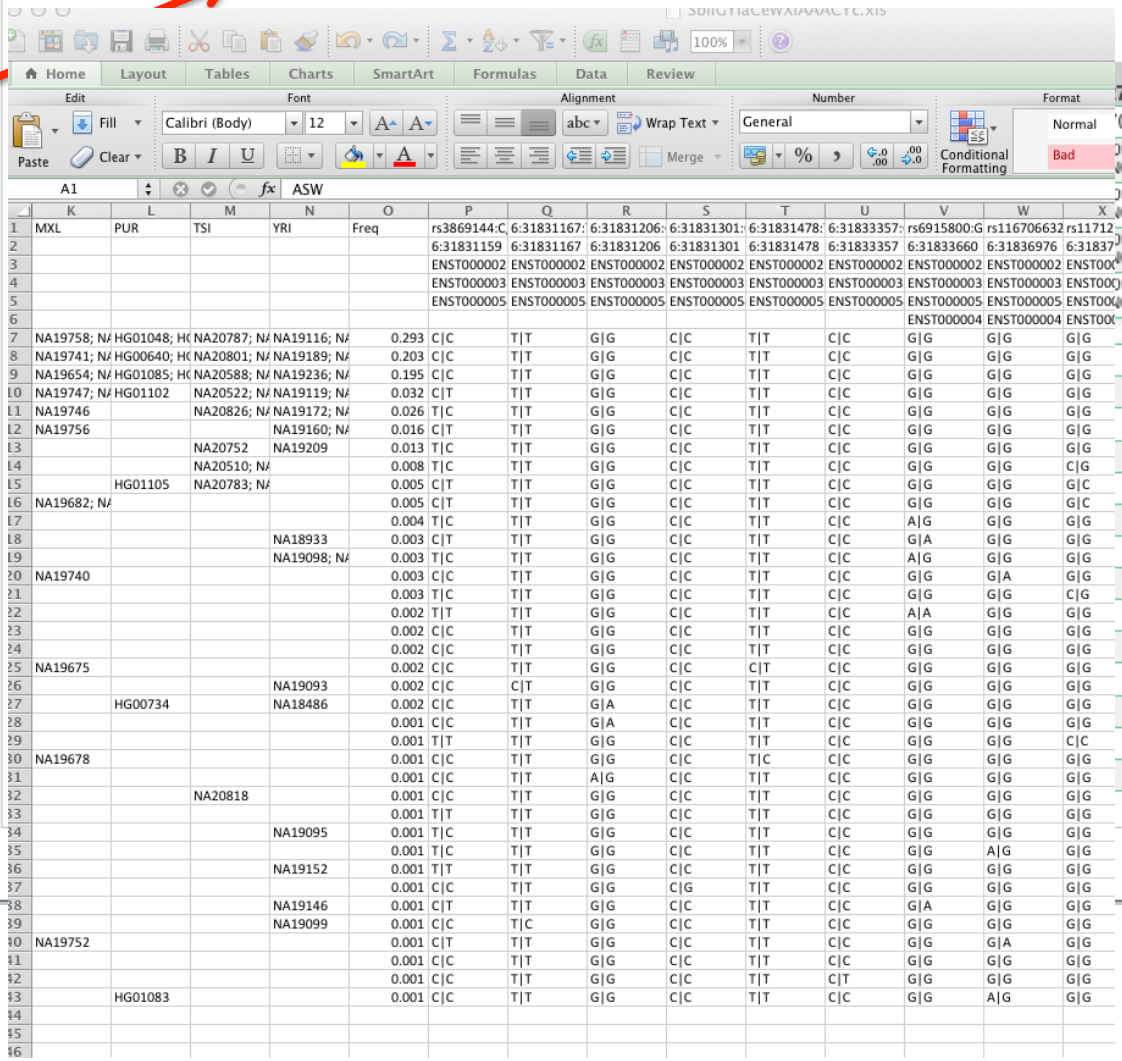
Custom Data

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

Variation Pattern Finder

Export data: CSV Excel



K	L	M	N	O	P	Q	R	S	T	U	V	W	X
MXL	PUR	TSI	YRI	Freq	rs3869144:C	6:31831167:	6:31831206:	6:31831301:	6:31831478:	6:31833357:	rs6915800:G	rs116706632	rs11712
					6:31831159	6:31831167	6:31831206	6:31831301	6:31831478	6:31833357	6:31833660	6:31836976	6:31837
					ENST000002	ENST000002	ENST000002	ENST000002	ENST000002	ENST000002	ENST000002	ENST000002	ENST000
					ENST000003	ENST000003	ENST000003	ENST000003	ENST000003	ENST000003	ENST000003	ENST000003	ENST000
					ENST000005	ENST000005	ENST000005	ENST000005	ENST000005	ENST000005	ENST000005	ENST000005	ENST000
											ENST000004	ENST000004	ENST000
NA19758; NA	HG01048; HG	NA20787; NA	NA19116; NA	0.293	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
NA19741; NA	HG00640; HG	NA20801; NA	NA19189; NA	0.203	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
NA19654; NA	HG01085; HG	NA20588; NA	NA19236; NA	0.195	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
NA19747; NA	HG01102	NA20522; NA	NA19119; NA	0.032	C|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
NA19746		NA20826; NA	NA19172; NA	0.026	T|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
NA19756			NA19160; NA	0.016	C|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
		NA20752	NA19209	0.013	T|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
		NA20510; NA		0.008	T|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	C|G
	HG01105	NA20783; NA		0.005	C|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|C
NA19682; NA				0.005	C|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|C
				0.004	T|C	T|T	G|G	C|C	T|T	C|C	A|G	G|G	G|G
			NA18933	0.003	C|T	T|T	G|G	C|C	T|T	C|C	G|A	G|G	G|G
			NA19098; NA	0.003	T|C	T|T	G|G	C|C	T|T	C|C	A|G	G|G	G|G
NA19740				0.003	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|A	G|G
				0.003	T|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	C|G
				0.002	T|T	T|T	G|G	C|C	T|T	C|C	A|A	G|G	G|G
				0.002	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
				0.002	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
NA19675				0.002	C|C	T|T	G|G	C|C	C|T	C|C	G|G	G|G	G|G
			NA19093	0.002	C|C	C|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
	HG00734		NA18486	0.002	C|C	T|T	G|A	C|C	T|T	C|C	G|G	G|G	G|G
				0.001	C|C	T|T	G|A	C|C	T|T	C|C	G|G	G|G	G|G
				0.001	T|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	C|C
NA19678				0.001	C|C	T|T	G|G	C|C	T|C	C|C	G|G	G|G	G|G
				0.001	C|C	T|T	A|G	C|C	T|T	C|C	G|G	G|G	G|G
		NA20818		0.001	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
				0.001	T|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
			NA19095	0.001	T|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
				0.001	T|C	T|T	G|G	C|C	T|T	C|C	G|G	A|G	G|G
			NA19152	0.001	T|T	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
				0.001	C|C	T|T	G|G	C|G	T|T	C|C	G|G	G|G	G|G
			NA19146	0.001	C|T	T|T	G|G	C|C	T|T	C|C	G|A	G|G	G|G
			NA19099	0.001	C|C	T|C	G|G	C|C	T|T	C|C	G|G	G|G	G|G
NA19752				0.001	C|T	T|T	G|G	C|C	T|T	C|C	G|G	G|A	G|G
				0.001	C|C	T|T	G|G	C|C	T|T	C|C	G|G	G|G	G|G
				0.001	C|C	T|T	G|G	C|C	T|T	C|T	G|G	G|G	G|G
	HG01083			0.001	C|C	T|T	G|G	C|C	T|T	C|C	G|G	A|G	G|G

7493:G/C	rs644827:T/C
009	6:31838441
00229729 ONYMOUS_CODING:Q/E	ENST00000229729 NON_SYNONYMOU
00375562 ONYMOUS_CODING:Q/E	ENST00000375562 NON_SYNONYMOU
00544672 ONYMOUS_CODING:Q/E	ENST00000544672 NON_SYNONYMOU
00414427 ONYMOUS_CODING:Q/E	
	CIC
	CIT
	TIC
	CIC
	CIC
	TIC
	CIT
	CIC
	TIC
	CIC
	CIC
	CIC
	CIT
	CIC
	CIT

A THOUSAND GENOMES

EMBL-EBI

Any questions?

- Please email info@1000genomes.org if you have any questions or feedback about this resource.

EMBL-EBI

Thanks

- The 1000 Genomes Project Consortium
- Paul Flicek, Laura Clarke
- Richard Smith, Holly Zheng Bradley and Ian Streeter
- Giulietta Spudich and Bert Overduin

A THOUSAND GENOMES

EMBL-EBI