The 1000 Genomes Web based Tutorial Exercises.

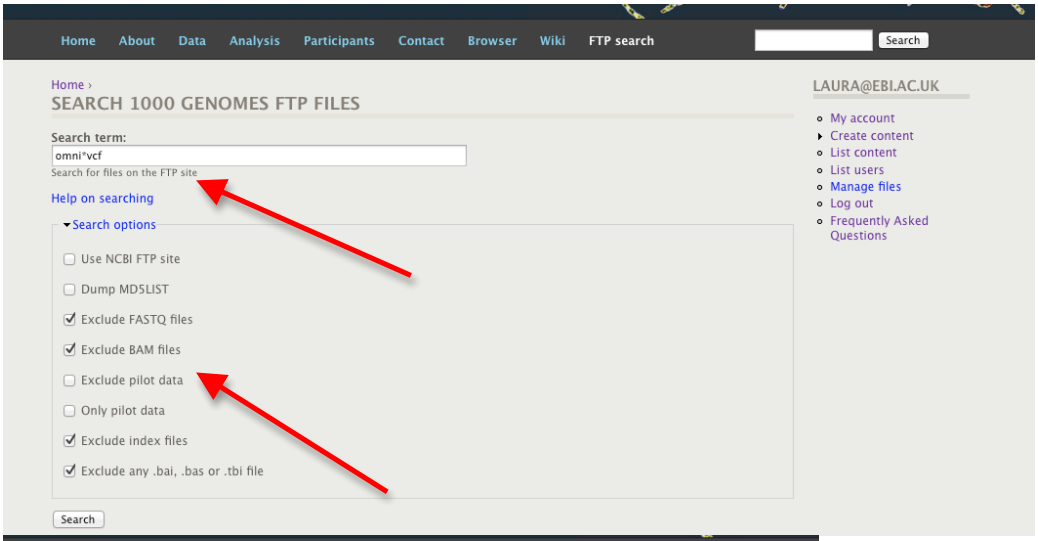
These are the answers for the Web Based Tutorial Exercises. Please note these are our recommended ways of doing these tasks but there may be other solutions too.

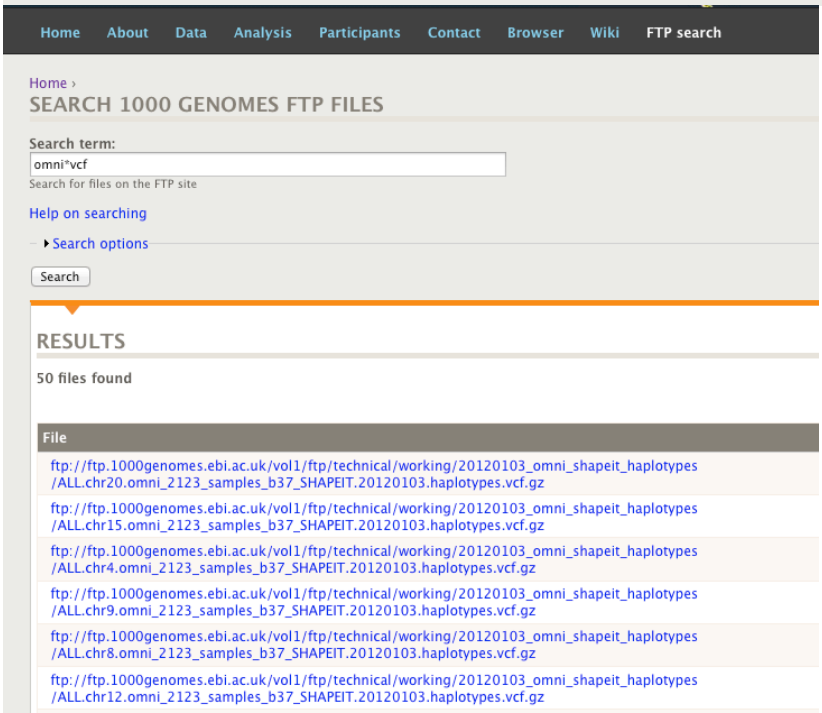
As mentioned in the Exercises document these exercises require you to have haploview installed to be able to complete them.

Haploview is available from:
http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/downloads

Finding Data

1a. Find what Omni VCF files we have on our ftp site using the website ftp search.



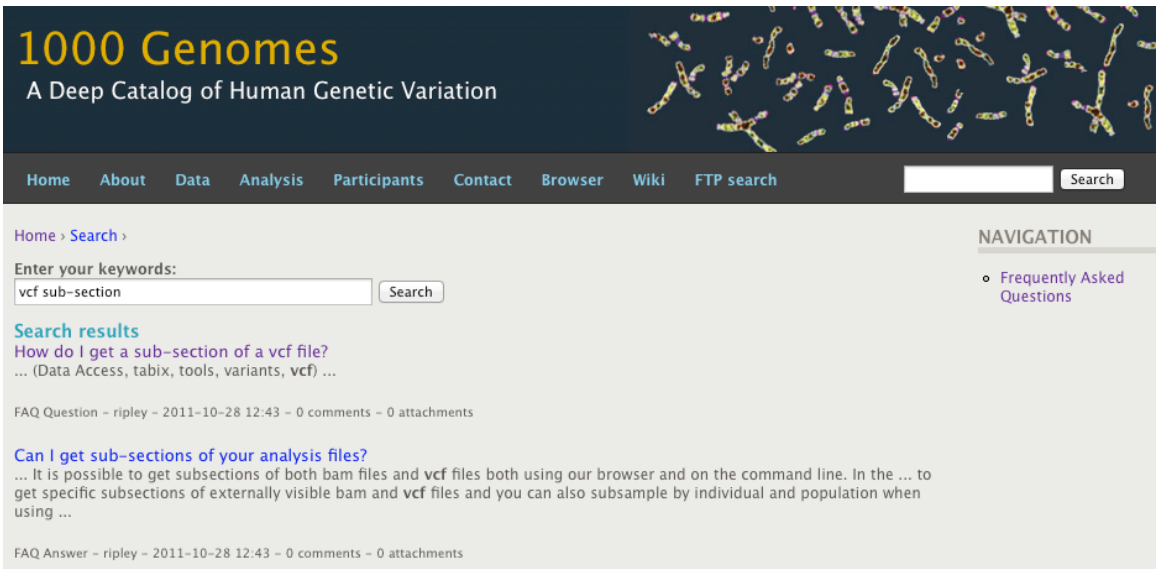


1b. Find the most recent Omni VCF file on build 37 from the 31st January 2012

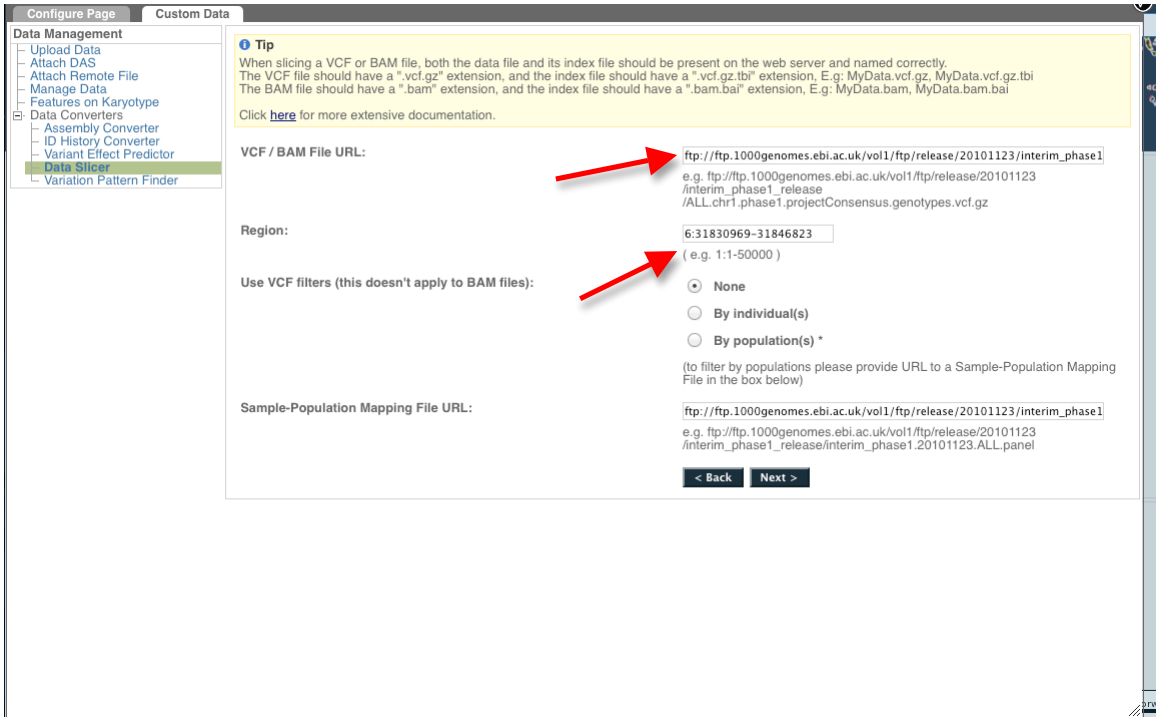
using 31*omni*vcf as a search term should give you two results, one which is b36 and one which is b37

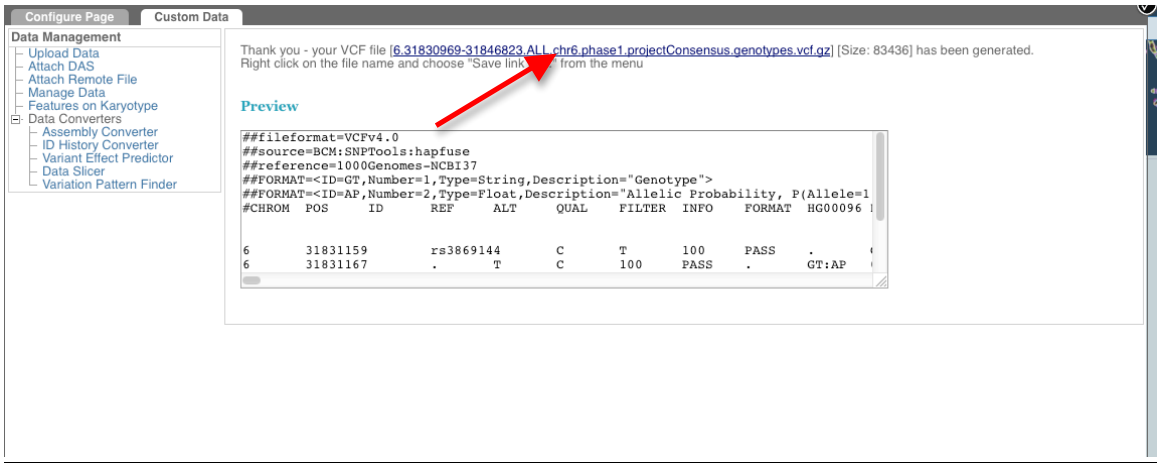
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b37.vcf.gz

2. Use the Website search box found in the top right hand corner of all pages to find the FAQ question about getting subsections of VCF files.



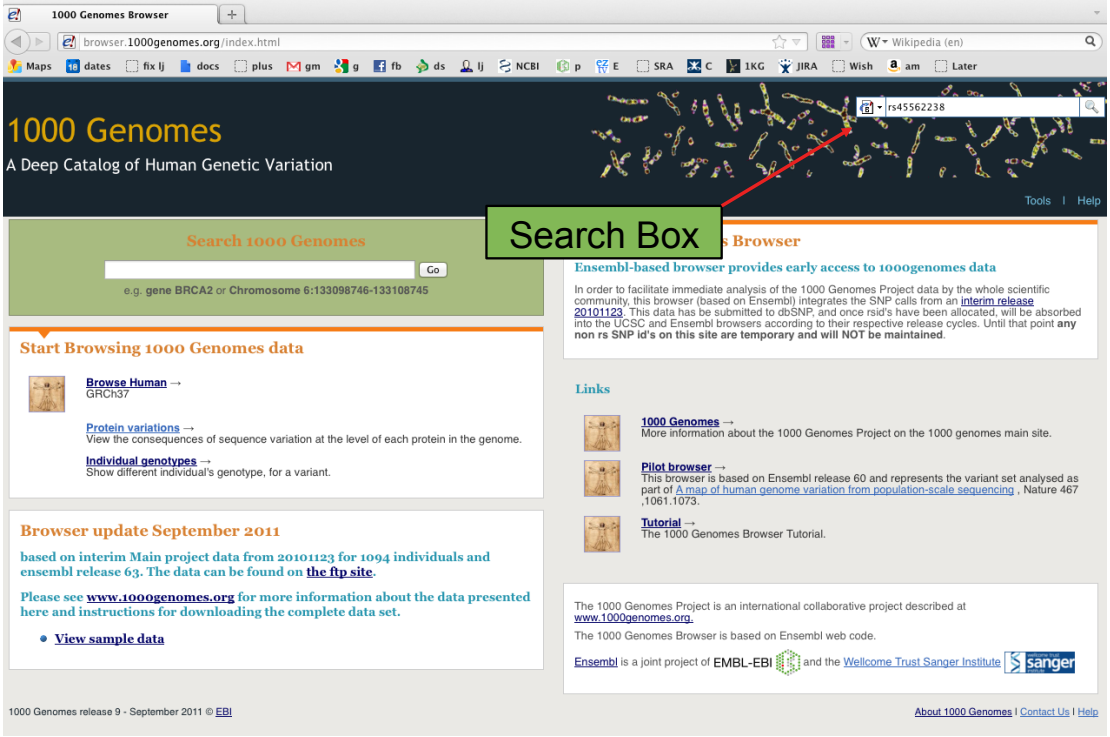
3. Use the Data slicer to get this section of the Omni VCF file 6:31830969-31846823:
http://browser.1000genomes.org/Homo_sapiens/UserData/SelectSlice

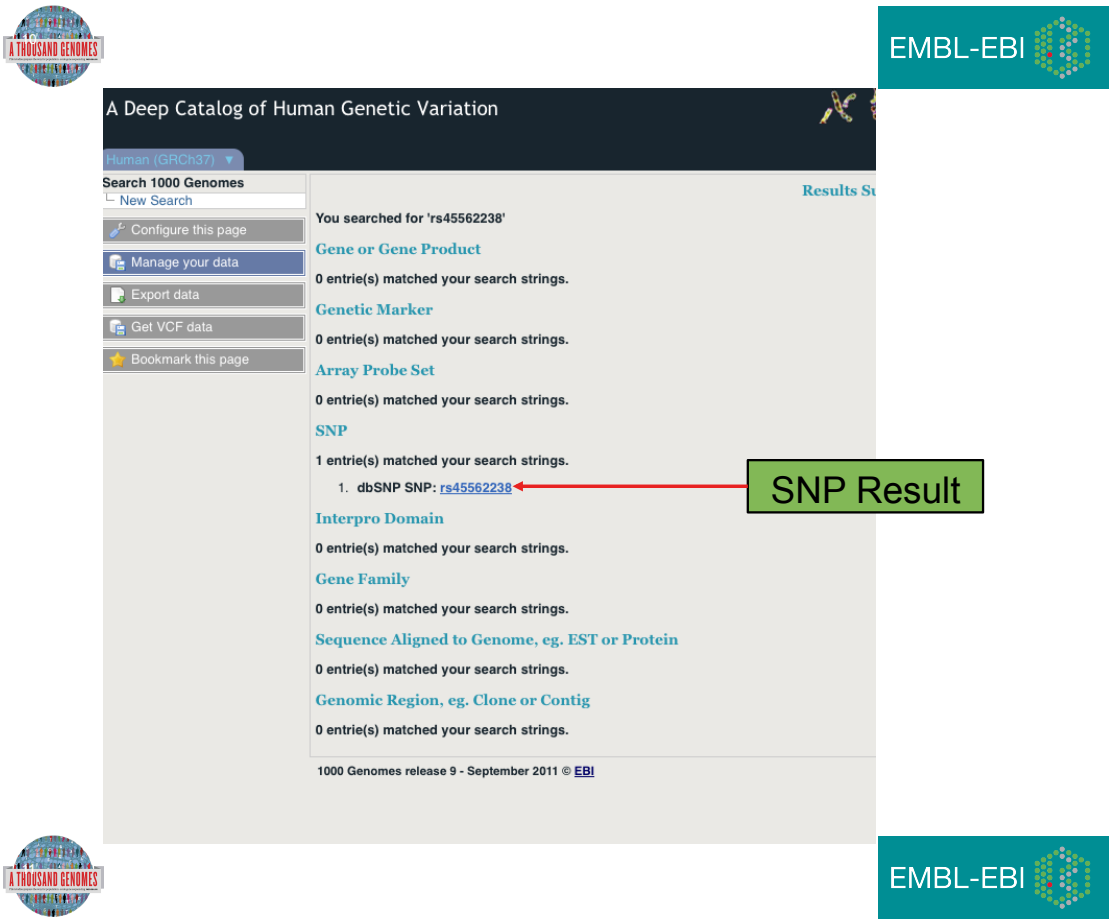




Browsing Data

4. Find the variant rs45562238 using http://browser.1000genomes.org.



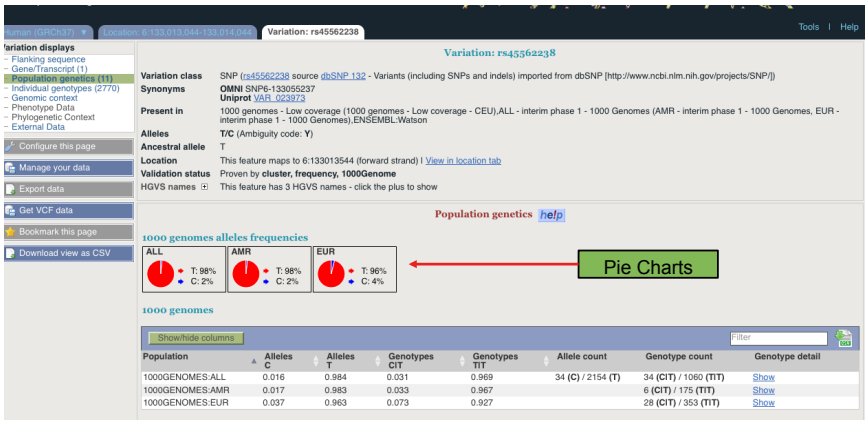


5. In what 1000 Genomes Super Population is this variant detected?

American and European

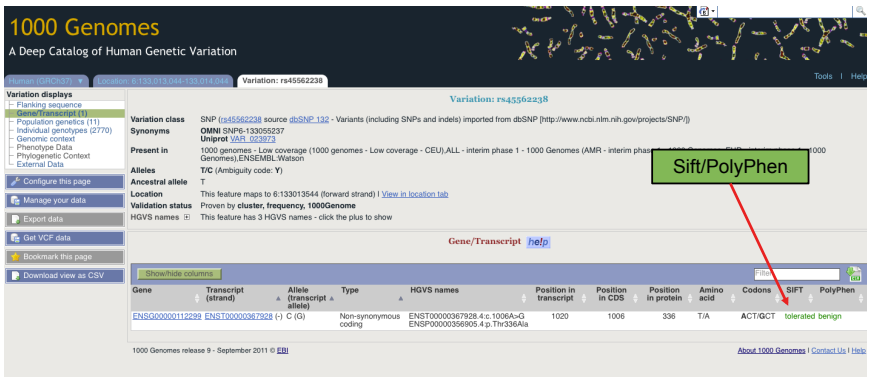
6. What are its global allele frequencies in the 1000 Genomes Data set?

0.02 is the global allele frequency, this is also the American Allele Frequency but it rises to 0.04 in the Europeans. The absence of Asians or Africans in this chart means that the variant was not found in any of our Asian of African individuals.



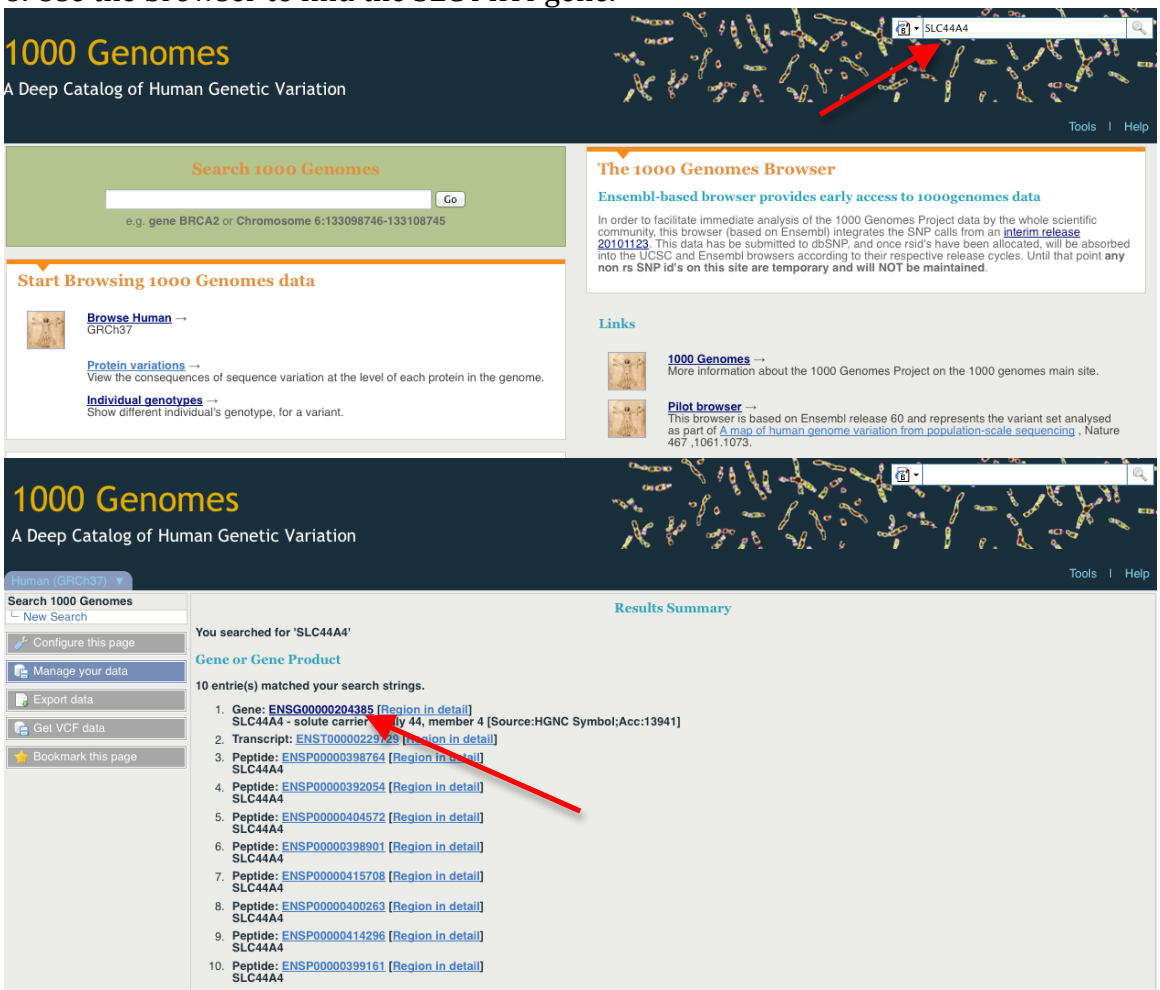
7. In which gene is the variant found?

ENSG00000112299, Vanin 1



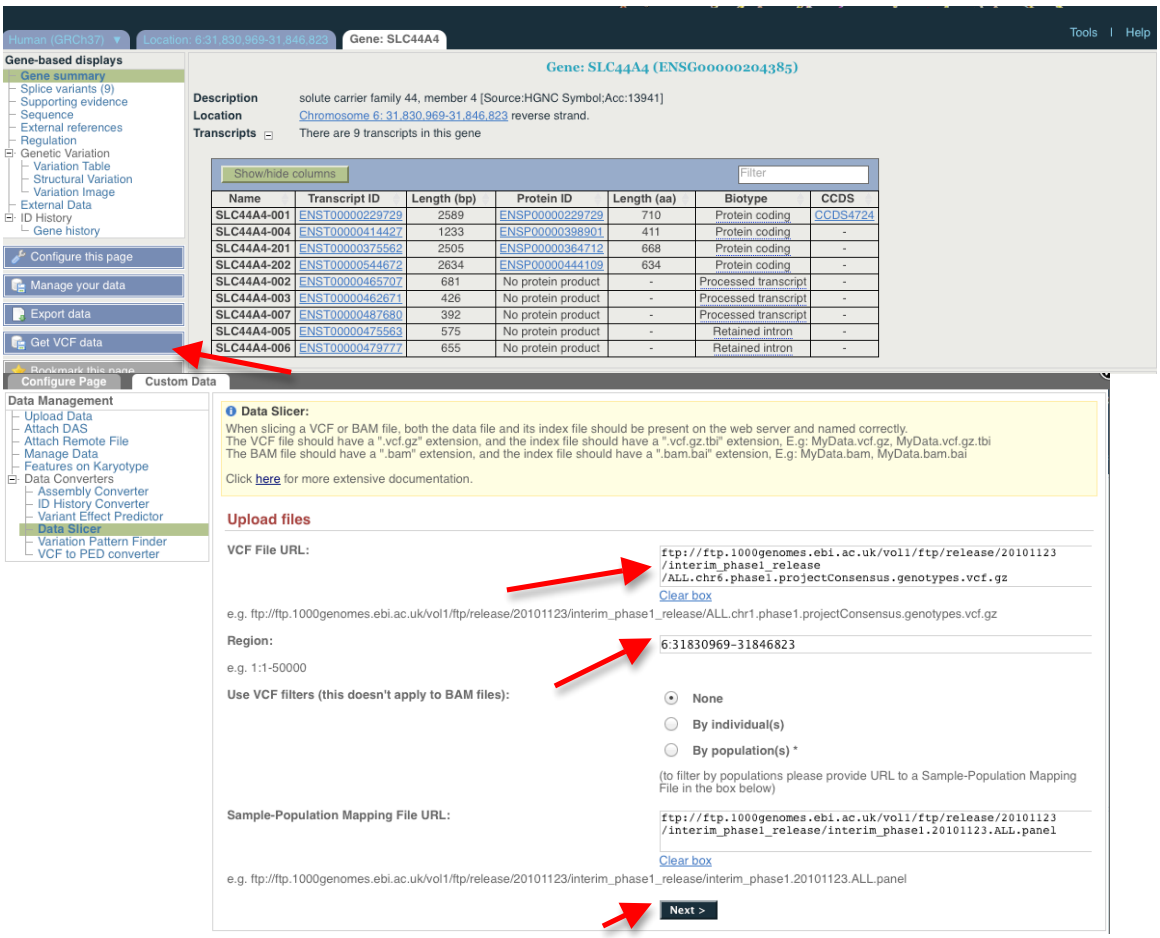
Using the 1000 Genomes Tools

8. Use the browser to find the SLC44A4 gene.

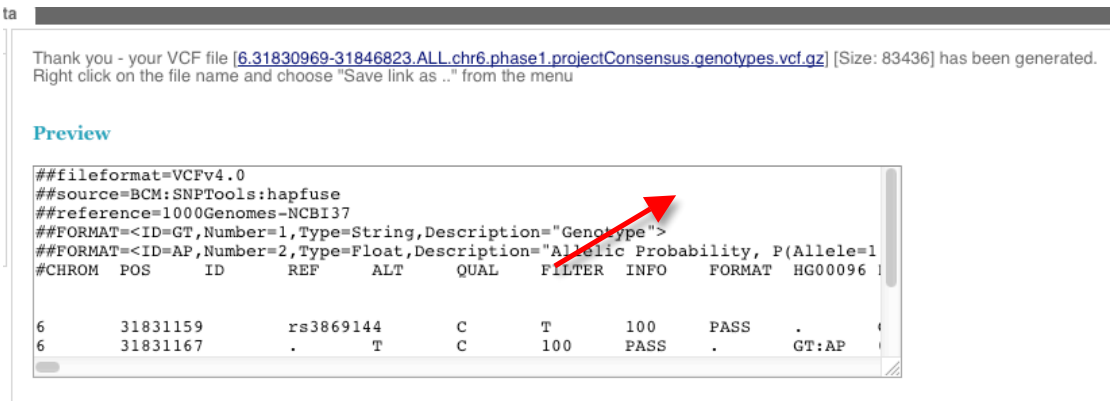


Putting the Gene name in the search box that is found in the top right hand corner of every page should lead you to the results page. You should follow the Gene name link to the Gene page.

9. Use the get VCF button in the left hand menu on the gene page to get a slice of a vcf file for this Gene.



When following the get vcf data button the form automatically fills out the input vcf file and position of the gene in the region box. If you wish to sub select a particular population or individual you would need to tick the appropriate box

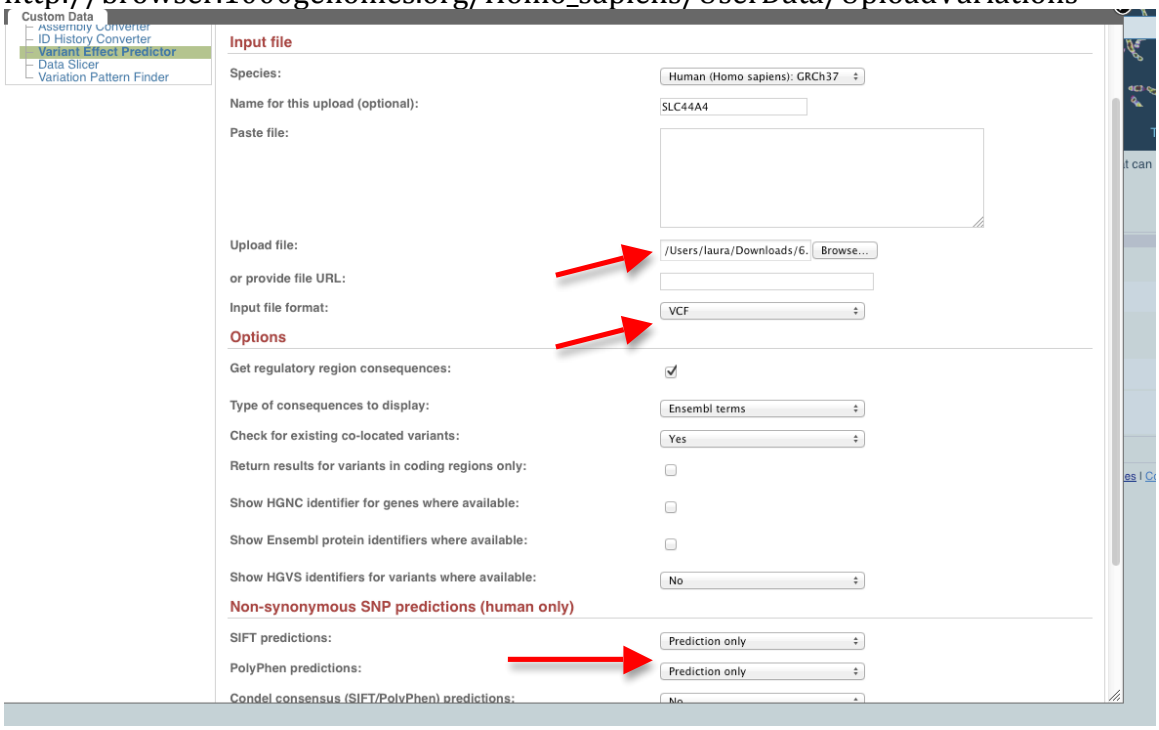


The Final page gives you a look at the top few lines of the file and a link to download the complete file

10. Unzip this VCF file using a tool link winzip or Archive Utility.

This should produce a file called 6.31830969-31846823.ALL.chr6.phase1.projectConsensus.genotypes.vcf

11a. Upload this VCF file to the Variant Effect Predictor.
http://browser.1000genomes.org/Homo_sapiens/UserData/UploadVariations



The input form asks you to browse to the location of the vcf file. You need to select vcf as the input format and to see the SIFT and PolyPhen predictions you need to select the appropriate dropdown menus.

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	ENST00000375562	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	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	-	-	-	-
6_31833612_C/G	6:31833612	G	-	ENSR00000487922	RegulatoryFeature	REGULATORY_REGION	-	-	-	-
6_31833612_C/G	6:31833612	G	ENSG00000204386	ENST00000495807	Transcript	UPSTREAM	-	-	-	-
6_31833612_C/G	6:31833612	G	ENSG00000204386	ENST00000480384	Transcript	UPSTREAM	-	-	-	-

The output from the Variation Effect Precictor gives the provided identifier for the variant (or uses the position and allele string to create one), the position of the variant, the alternative allele then information about the feature it overlaps with and that effect that causes.

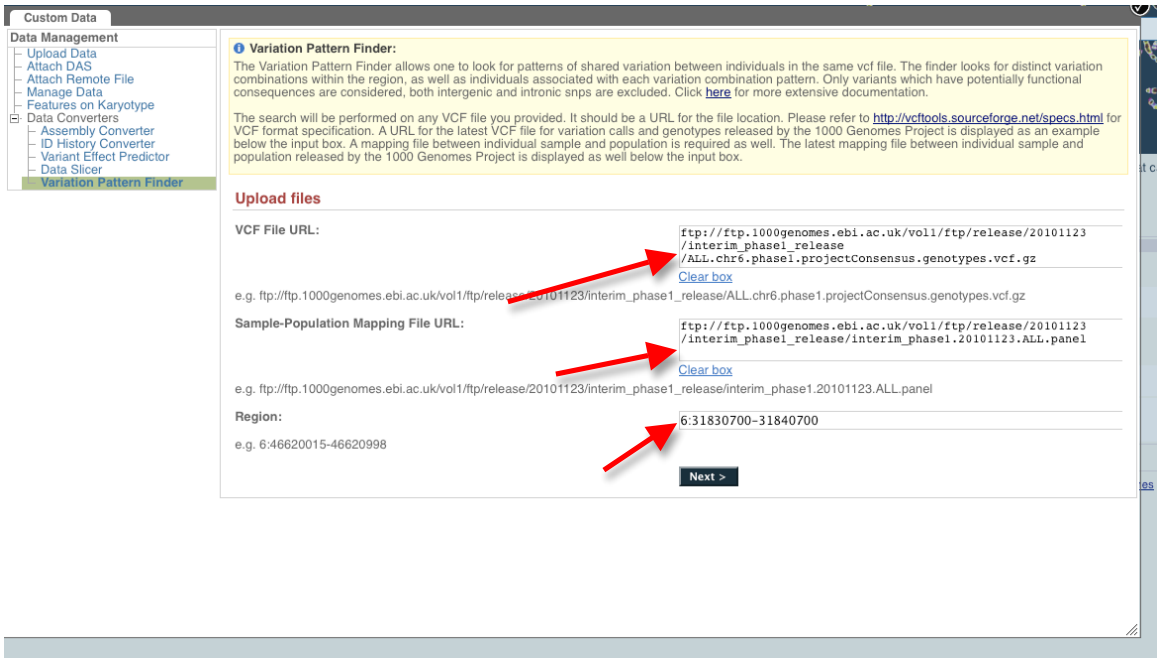
-	-	-	1KG_6_31833357	-
-	-	-	1KG_6_31833357	-
535	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=probably_damaging; Condel=deleterious
493	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=possibly_damaging; Condel=deleterious
459	R/H	cGc/cAc	1KG_6_31833357	SIFT=deleterious; PolyPhen=probably_damaging; Condel=deleterious
-	-	-	1KG_6_31833357	-
-	-	-	1KG_6_31833357	-
-	-	-	1KG_6_31833357	-

If you scroll along the page you will see additional information that you requested on the form like any variants in the Ensembl database yours overlaps with and what the SIFT and PolyPhen results are.

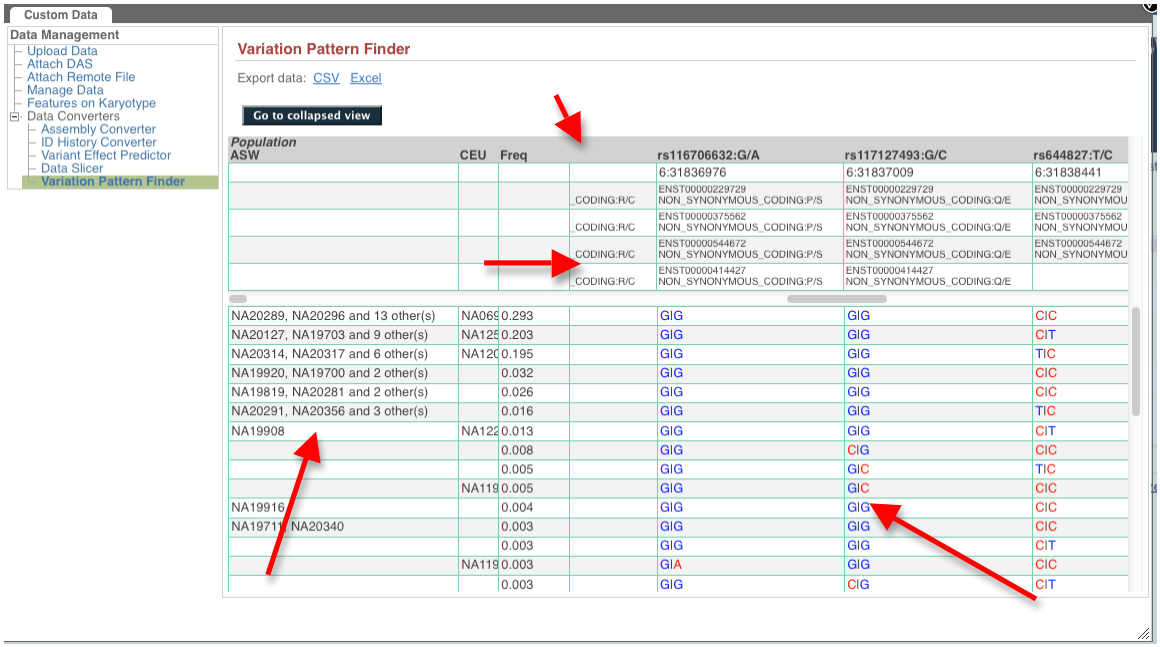
11b. Do any of the variants have negative SIFT or PolyPhen predictions?

Yes, There are several variants which have negative SIFT or PolyPhen predictions including 6_31833357_C/T which overlaps with 3 different transcripts all with deleterious non synonymous codon changes, ENST00000229729, ENST00000375562, ENST00000544672

12. Using the example URLs on the Variation Pattern Finder tool menu look at the patterns of inheritance for this region: 6:31830700-31840700
http://browser.1000genomes.org/Homo_sapiens/UserData/VariationsMapVCF

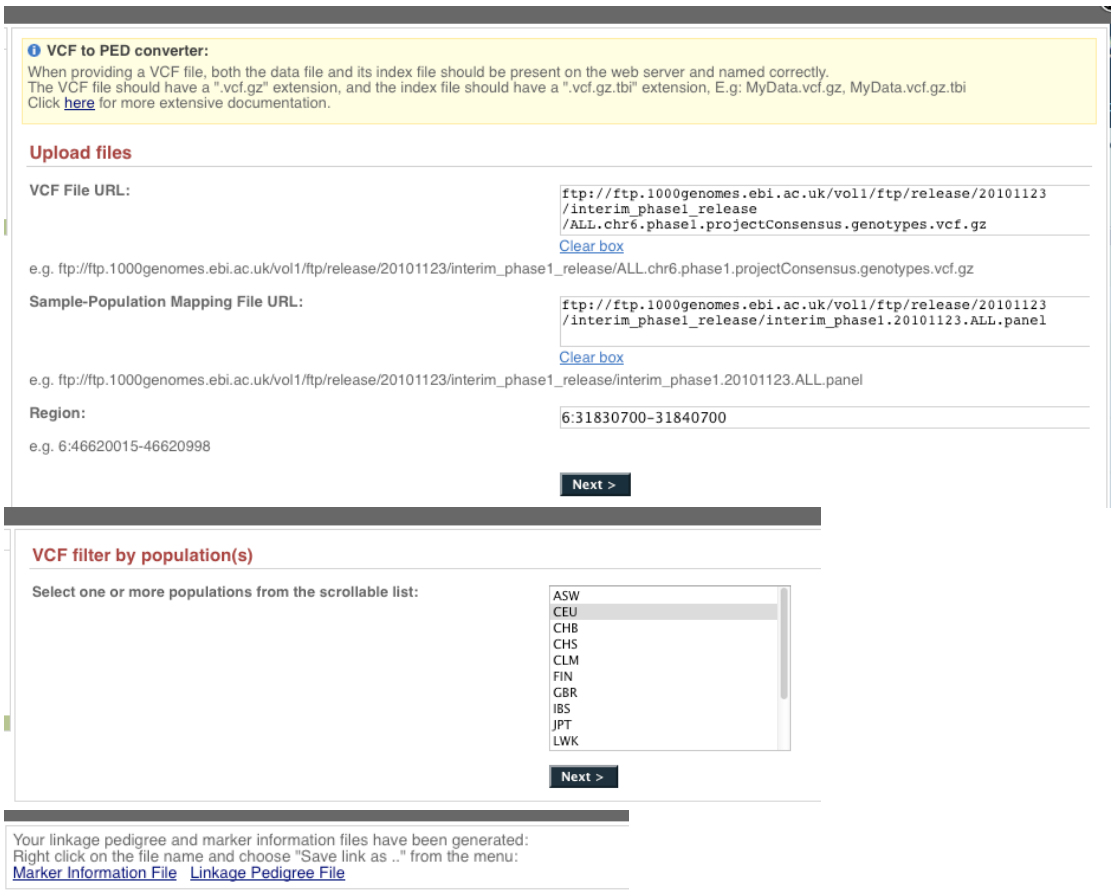


As this gene is on chr 6 the default URLs in the form when you first click on it should be fine. You need to add your region to the region box and then after clicking next you will see:



The grey headline row has population names and varianat names and alleles. The following rows contain the functional consequences of these variants. This tool only considers variants with functional consequences. The rows then contain a list of individuals who are part of that population, the global frequency this pattern occurs in and the actual pattern of genotypes in those individuals.

13a. For the same region use the VCF to PED tool to produce a ped and info file for the CEU population.



Again as we are considering a gene on chromosome 6 the default URLs in the box should work. Once you have put the region in the region box and clicked next you should see a list of populations including CEU. If you select CEU and click next you will then be presented with two links to files you can right click on and download.

13b. Look at these files in haploview.

Loading the data (ped) and locus information file (info) into haploview gives you the ability to look at the ld plot for the region.

13c. How many haplotype blocks does haploview think there are in this section?

The Haplotypes button views you a view of the haplotype blocks which exist in that region. In this case there are 2 haplotype blocks.