

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.

The image shows two screenshots of the 1000 Genomes FTP search website. The top screenshot shows the search interface with the search term 'omni*vcf' and several search options. Two red arrows point to the search term input field and the 'Exclude FASTQ files' checkbox. The bottom screenshot shows the search results page, which displays 50 files found. The results are listed in a table with a 'File' column containing the full FTP path for each file.

Home About Data Analysis Participants Contact Browser Wiki FTP search

Home >
SEARCH 1000 GENOMES FTP FILES

Search term:
omni*vcf
Search for files on the FTP site

Help on searching

Search options

- Use NCBI FTP site
- Dump MD5LIST
- Exclude FASTQ files
- Exclude BAM files
- Exclude pilot data
- Only pilot data
- Exclude index files
- Exclude any .bai, .bas or .tbi file

Search

LAURA@EBI.AC.UK

- My account
- Create content
- List content
- List users
- Manage files
- Log out
- Frequently Asked Questions

Home About Data Analysis Participants Contact Browser Wiki FTP search

Home >
SEARCH 1000 GENOMES FTP FILES

Search term:
omni*vcf
Search for files on the FTP site

Help on searching

Search options

Search

RESULTS

50 files found

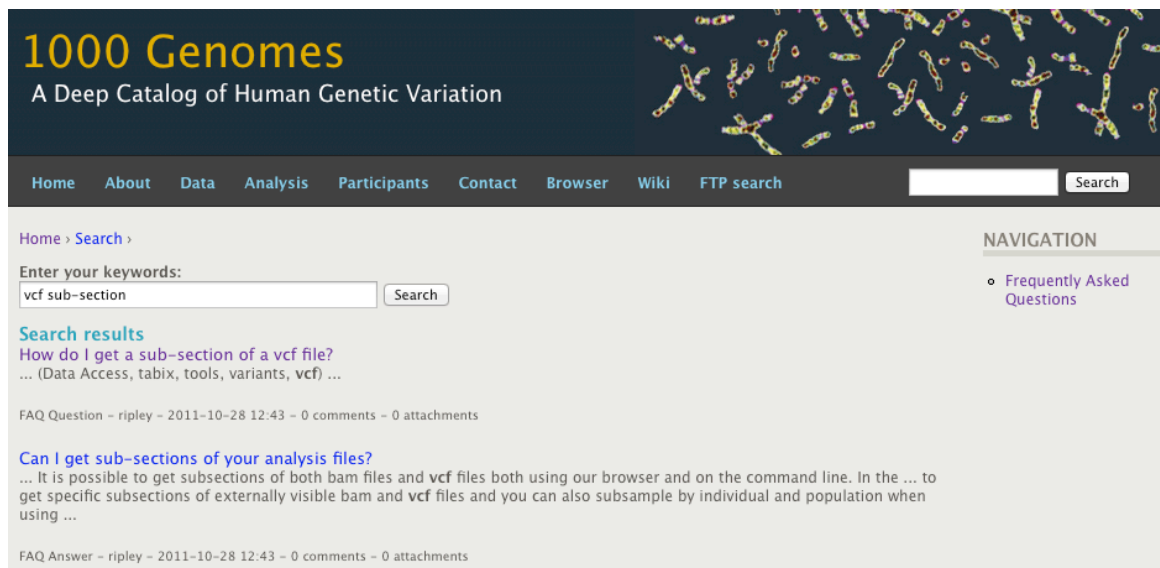
File
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr20.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr15.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr4.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr9.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr8.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr12.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz

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.



1000 Genomes
A Deep Catalog of Human Genetic Variation

Home About Data Analysis Participants Contact Browser Wiki FTP search Search

Home > Search >

Enter your keywords:
 Search

Search results
How do I get a sub-section of a vcf file?
... (Data Access, tabix, tools, variants, vcf) ...

FAQ Question - ripley - 2011-10-28 12:43 - 0 comments - 0 attachments

Can I get sub-sections of your analysis files?
... It is possible to get subsections of both bam files and vcf files both using our browser and on the command line. In the ... to get specific subsections of externally visible bam and vcf files and you can also subsample by individual and population when using ...

FAQ Answer - ripley - 2011-10-28 12:43 - 0 comments - 0 attachments

NAVIGATION
◦ [Frequently Asked Questions](#)

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

```

##fileformat=VCFv4.0
##source=BCN: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(AAllele=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

```

Browsing Data

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

1000 Genomes
A Deep Catalog of Human Genetic Variation

Search 1000 Genomes

Search Box

Ensembl-based browser provides early access to 1000genomes data

Start Browsing 1000 Genomes data

Browser update September 2011

Links

1000 Genomes release 9 - September 2011 © EBI

A Deep Catalog of Human Genetic Variation

Human (GRCh37)

Search 1000 Genomes

New Search

Configure this page

Manage your data

Export data

Get VOF data

Bookmark this page

You searched for 'rs45562238'

Gene or Gene Product

0 entrie(s) matched your search strings.

Genetic Marker

0 entrie(s) matched your search strings.

Array Probe Set

0 entrie(s) matched your search strings.

SNP

1 entrie(s) matched your search strings.

1. dbSNP SNP: [rs45562238](#)

Interpro Domain

0 entrie(s) matched your search strings.

Gene Family

0 entrie(s) matched your search strings.

Sequence Aligned to Genome, eg. EST or Protein

0 entrie(s) matched your search strings.

Genomic Region, eg. Clone or Contig

0 entrie(s) matched your search strings.

1000 Genomes release 9 - September 2011 © EBI

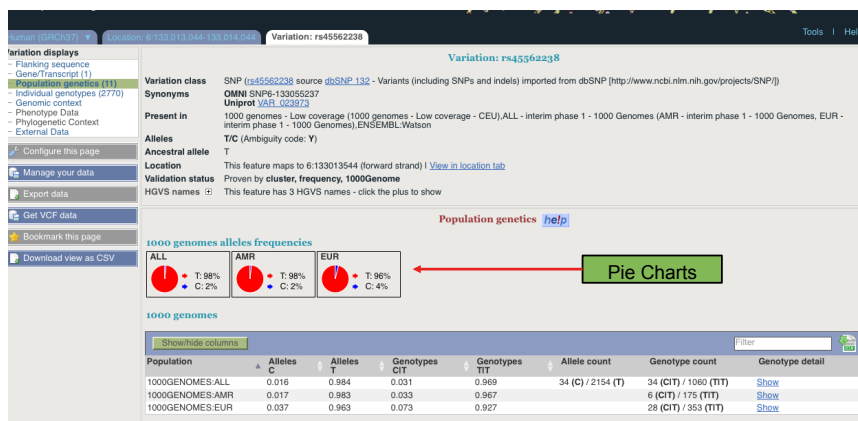
EMBL-EBI

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

American and European

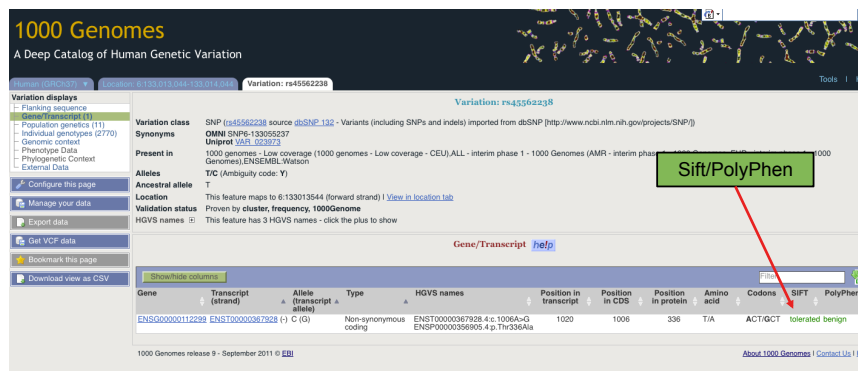
6. What are its global allele frequencies in the 1000 Genomes Data set?

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 or 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.

The image shows two screenshots of the 1000 Genomes website. The top screenshot shows the homepage with a search bar in the top right corner containing the text 'SLC44A4'. A red arrow points to this search bar. Below the search bar is a 'Search 1000 Genomes' section with a search input field and a 'Go' button. To the right is a section titled 'The 1000 Genomes Browser' with a description and 'Links' to '1000 Genomes' and 'Pilot browser'. The bottom screenshot shows the search results page for 'SLC44A4'. The search bar at the top right still contains 'SLC44A4'. On the left is a sidebar with search options. The main content area shows 'Results Summary' for 'Gene or Gene Product' and lists 10 entries that matched the search strings. A red arrow points to the first entry: 'Gene: ENSG00000204385 [Region in detail] SLC44A4 - solute carrier family 44, member 4 [Source:HGNC Symbol;Acc:13941]'. The second entry is 'Transcript: ENST00000229729 [Region in detail] SLC44A4'.

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.

Human (GRCh37) Location: 6:31,830,969-31,846,823 Gene: SLC44A4

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

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	-

Data Slicer:
 When slicing a VCF or BAM file, both the data file and its index file should be present on the web server and named correctly. The VCF file should have a ".vcf.gz" extension, and the index file should have a ".vcf.gz.tbi" extension, E.g: MyData.vcf.gz, MyData.vcf.gz.tbi. The BAM file should have a ".bam" extension, and the index file should have a ".bam.bai" extension, E.g: MyData.bam, MyData.bam.bai. Click [here](#) for more extensive documentation.

Upload files

VCF File URL: [Clear box](#)

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr1.phase1.projectConsensus.genotypes.vcf.gz

Region:

e.g. 1:1-50000

Use VCF filters (this doesn't apply to BAM files):

None
 By individual(s)
 By population(s) *

(to filter by populations please provide URL to a Sample-Population Mapping File in the box below)

Sample-Population Mapping File URL: [Clear box](#)

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel

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

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(Alele=1)>
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HG00096
6 31831159 rs3869144 C T 100 PASS . GT:AP
6 31831167 . T C 100 PASS . GT:AP
```

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	ENST00000491788	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 Predictor 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, ENST0000229729, 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

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: [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: [Clear box](#)

e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel

Region:

e.g. 6:46620015-46620998

Next >

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:

Variation Pattern Finder

Export data: [CSV](#) [Excel](#)

[Go to collapsed view](#)

Population	CEU	Freq	rs116706632:G/A	rs117127493:G/C	rs644827:T/C
ASW			6:31836976	6:31837009	6:31838441
			ENST00000229729 NON_SYNONYMOUS_CODING:P/S	ENST00000229729 NON_SYNONYMOUS_CODING:Q/E	ENST00000229729 NON_SYNONYMOUS_CODING:Q/E
			ENST00000375562 NON_SYNONYMOUS_CODING:P/S	ENST00000375562 NON_SYNONYMOUS_CODING:Q/E	ENST00000375562 NON_SYNONYMOUS_CODING:Q/E
			ENST00000544672 NON_SYNONYMOUS_CODING:P/S	ENST00000544672 NON_SYNONYMOUS_CODING:Q/E	ENST00000544672 NON_SYNONYMOUS_CODING:Q/E
			ENST00000414427 NON_SYNONYMOUS_CODING:P/S	ENST00000414427 NON_SYNONYMOUS_CODING:Q/E	ENST00000414427 NON_SYNONYMOUS_CODING:Q/E
NA20289, NA20296 and 13 other(s)	NA066	0.293	G/G	G/G	C/C
NA20127, NA19703 and 9 other(s)	NA126	0.203	G/G	G/G	C/T
NA20314, NA20317 and 6 other(s)	NA120	0.195	G/G	G/G	T/C
NA19920, NA19700 and 2 other(s)		0.032	G/G	G/G	C/C
NA19819, NA20281 and 2 other(s)		0.026	G/G	G/G	C/C
NA20291, NA20356 and 3 other(s)		0.016	G/G	G/G	T/C
NA19908	NA122	0.013	G/G	G/G	C/T
		0.008	G/G	C/G	C/C
		0.005	G/G	G/C	T/C
	NA119	0.005	G/G	G/C	C/C
NA19916		0.004	G/G	G/G	C/C
NA19711, NA20340		0.003	G/G	G/G	C/C
		0.003	G/G	G/G	C/T
	NA118	0.003	G/A	G/G	C/C
		0.003	G/G	C/G	C/T

The grey headline row has population names and variant 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.

VCF to PED converter:
When providing a VCF file, both the data file and its index file should be present on the web server and named correctly. The VCF file should have a ".vcf.gz" extension, and the index file should have a ".vcf.gz.tbi" extension, E.g: MyData.vcf.gz, MyData.vcf.gz.tbi
Click [here](#) for more extensive documentation.

Upload files

VCF File URL:
e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz
[Clear box](#)

Sample-Population Mapping File URL:
e.g. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/interim_phase1.20101123.ALL.panel
[Clear box](#)

Region:
e.g. 6:46620015-46620998

[Next >](#)

VCF filter by population(s)

Select one or more populations from the scrollable list:

- ASW
- CEU**
- CHB
- CHS
- CLM
- FIN
- GBR
- IBS
- JPT
- LWK

[Next >](#)

Your linkage pedigree and marker information files have been generated:
Right click on the file name and choose "Save link as .." from the menu:
[Marker Information File](#) [Linkage Pedigree File](#)

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.