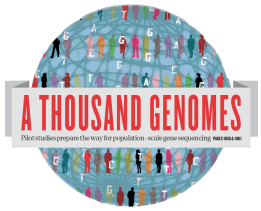


The 1000 Genomes Command line Tutorial Exercises.

These are the answers for the command line tutorial exercises. Please note this represent one-way of answering these questions. For some of the questions there are multiple correct answers.

1a. Use the current.tree file from our ftp site to find what omni vcf files are available. (Omni is a high throughput genotyping platform from Illumina on which all 1000 genomes samples are being genotyped)
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree

> grep omni current.tree
ftp/technical/working/20110329_wgs_genotypes/bcm/ALL.chr20.vqsr_site_v4_omni.20101123.genotypes.vcf.gz file 457314670 Tue May 3 14:23:41 2011 e05b0bf1f2e0dd694a26e92195c4cc91
ftp/technical/working/20110329_wgs_genotypes/bcm/ALL.chr20.vqsr_site_v4_omni.20101123.genotypes.vcf.gz.tbi file 53587 Tue May 3 14:23:41 2011 50f148df5a2cf425b8da9e4170c51e04
ftp/technical/working/20120103_omni_shapeit_haplotypes directory 3831 Tue Jan 3 15:43:57 2012
ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr19.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz file 23484317 Tue Jan 3 15:02:51 2012 55a1d0bdb004f139a5d1844dc4611b34
ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr1.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz.tbi file 178633 Tue Jan 3 15:27:55 2012 9dc46e75b76faf15a08aed0100009ca5
ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr17.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz.tbi file 57618 Tue Jan 3 15:28:22 2012 3af55a98cfc8a81c746b7105bb3ecb18
ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr5.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz.tbi file 142520 Tue Jan 3 15:27:30 2012 45a22c59eadae67f3949d914250294c5
ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr18.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz.tbi file 56884 Tue Jan 3 15:27:38 2012 28517a7168cdc9fbfecf1140fbcc4f79
ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr20.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz.tbi file 46339 Tue Jan 3 15:27:24 2012 8334495a3cd13a1b0d293faeeec3c972
ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr15.omni_2123_samples_b37_SHAPEIT.20120103.haplotypes.vcf.gz file 33379133 Tue Jan 3 15:10:16 2012 0aae64cbba827b65a6dff164c80860bd

This should produce a list of 124 files

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

>grep omni current.tree | grep 20120131 | grep b37 | grep -v tbi | cut -f1 | awk '{print "ftp://ftp.1000genomes.ebi.ac.uk/vol1/"$1}'
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b37.vcf.gz

The chain of grep commands in this command line there to filter the results just to the b37 command. The grep -v excludes the tabix index file and the final awk statement adds the ftp url to the filepath

2a. Use tabix to get a slice of the 31st January b37 Omni VCF File. Fetch a piece for the position 6:31830969-31846823

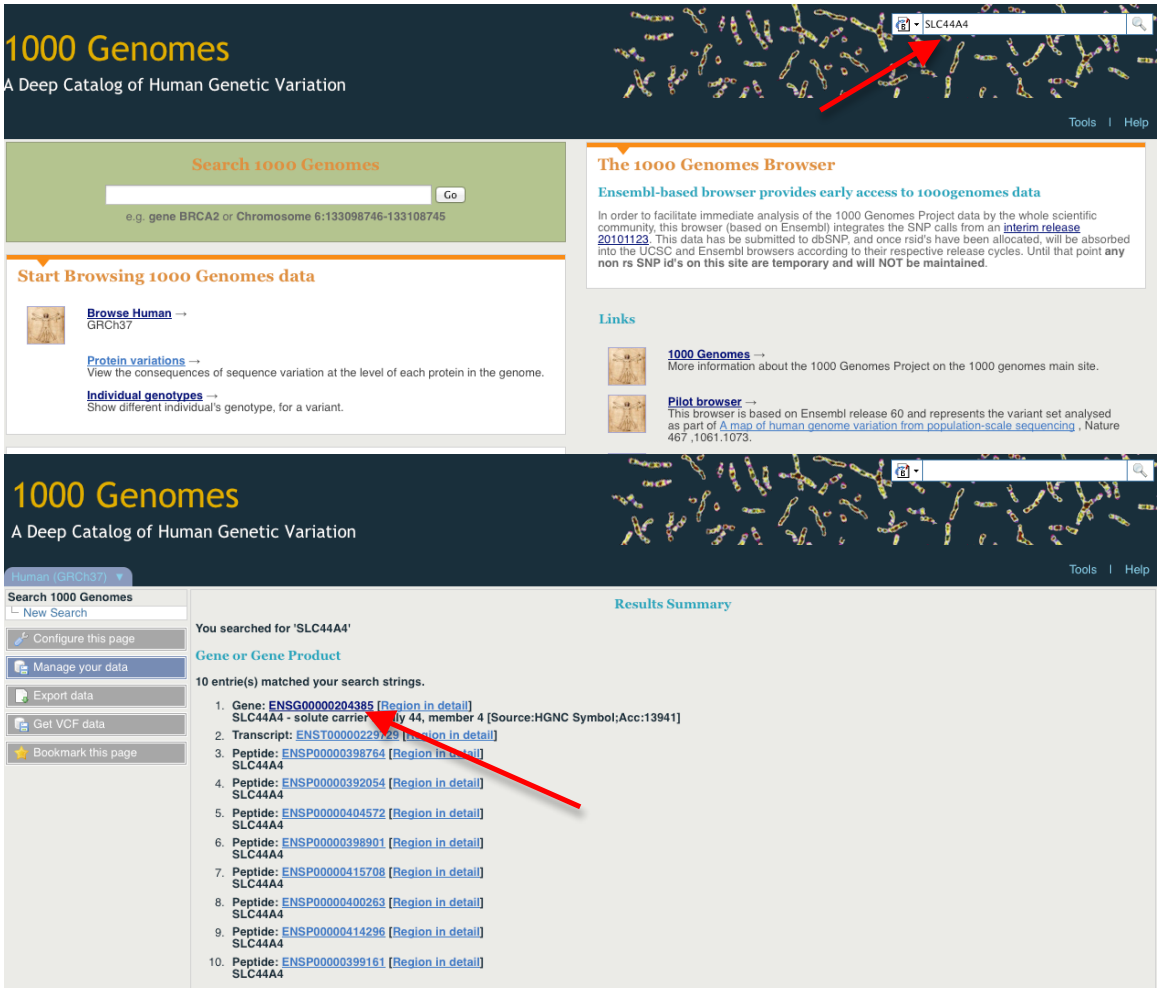
```
> tabix -h
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b37.vcf.gz 6:31830969-31846823
6	31833221	rs17207713	G	A	.	PASS
CR=99.9063;GentrainScore=0.5911;HW=1.0
6	31833504	rs34418207	G	A	.	PASS
CR=99.85981;GentrainScore=0.7957;HW=1.0
6	31834197	rs4947332	C	T	.	PASS
CR=99.95225;GentrainScore=0.7429;HW=0.80660635
6	31836151	SNP6-31944130	G	A	.	PASS
CR=99.85;GentrainScore=0.8616;HW=0.036248714
```

2b. Use vcftools vcf-subset to generate this subsection but only containing the individual HG00096

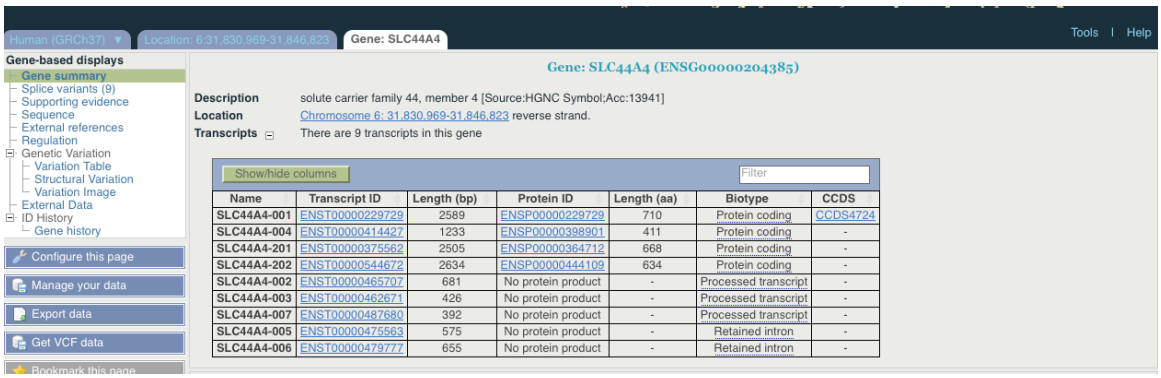
```
> tabix -h
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/Omni25_genotypes_2141_samples.b37.vcf.gz 6:31830969-31846823 | vcf-subset -c HG00096
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096
6	31833221	rs17207713	G	.	.	PASS
CR=99.9063;GentrainScore=0.5911;HW=1.0	GT:GC	0/0:0.4376
6	31833504	rs34418207	G	.	.	PASS
CR=99.85981;GentrainScore=0.7957;HW=1.0	GT:GC	0/0:0.7427
6	31834197	rs4947332	C	.	.	PASS
CR=99.95225;GentrainScore=0.7429;HW=0.80660635	GT:GC	0/0:0.7194
6	31836151	SNP6-31944130	G	.	.	PASS
CR=99.85;GentrainScore=0.8616;HW=0.036248714	GT:GC	0/0:0.8978
```

Using the 1000 Genomes Tools

3. Use the browser to find location 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.



You can also find this information using our public mysql instance

```
>mysql -hmysql-db.1000genomes.org -P4272 -uanonymous -D homo_sapiens_core_63_37g -e "select gene.gene_id, xref.display_label, seq_region.name, gene.seq_region_start, gene.seq_region_end from gene, xref, seq_region where gene.display_xref_id = xref.xref_id and gene.seq_region_id = seq_region.seq_region_id and xref.display_label = 'SLC44A4' order by name"
+---------+---------------+------+------------------+----------------+
| gene_id | display_label | name | seq_region_start | seq_region_end |
+---------+---------------+------+------------------+----------------+
|   63734 | SLC44A4       | 6    |         31830969 |       31846823 |
+---------+---------------+------+------------------+----------------+
```

4a. Find the 20111114 low coverage mapped bam file for HG01375

```
> grep 20111114 current.tree | grep "low_coverage" | grep mapped | grep HG01375  | grep -v unmapped | grep -v bas | grep -v bai | cut -f1 | awk '{print "http://ftp.1000genomes.ebi.ac.uk/vol1/"$1}';
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/HG01375.mapped.ILLUMINA.bwa.CLM.low_coverage.20111114.bam
```

4b. Use samtools to look at the HG01375 bam in the region of SLC44A4

```
> samtools view
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/HG01375/alignment/HG01375.mapped.ILLUMINA.bwa.CLM.low_coverage.20111114.bam 6:31830969-31846823
[knet_seek] SEEK_END is not supported for HTTP. Offset is unchanged.
ERR020288.66235376	163	6	31830879	60	91M	=	31831274	485
AGGACACGTCACAGGGGCGGGGCCGATTTTAAAGAGCCGGGCGCGGAAAAAAAAAGGCCGCCTGTCGTCGTGGAGAGAATGAGTCACAGAT
AFEFDGF?GHIGJJGKKJBLELCE@DJJJJGIKIIGKLBLJGAKAGJJMMJMILMDJGA>BFFDDB:AB=FIHFH<C?@CA=@3>@BFA?C	X0:i:1	X1:i:0	MD:Z:91	RG:Z:ERR020288	AM:i:37	NM:i:0	SM:i:37	MQ:i:60	XT:A:U
BQ:Z:@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@E
```

5. Use tabix to get a vcf file from our 20110521 release for the region of SLC44A4

grep 20110521 current.tree | grep release | grep chr6 | grep -v tbi | cut -f1 | awk '{print "ftp://ftp.1000genomes.ebi.ac.uk/vol1/"$1}'
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_release_v2.20101123.snps_indels_svs.vcf.gz

> tabix -h ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_release_v2.20101123.snps_indels_svs.vcf.gz 6:31830969-31846823 > 6_31830969_31846823.vcf
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HG00096
6 31831004 . TTTTG T 39 PASS AC=7;ERATE=0.0005;AN=2184;VT=INDEL;RSQ=0.8382;LDAF=0.0039;THETA=0.0029;AVGPOST=0.9984;AF=0.0032;AMR_AF=0.0028;AFR_AF=0.01;EUR_AF=0.0013 GT:DS:GL 0|0:0.000:0.00,-0.30,-7.60
6 31831159 rs3869144 C T 100 PASS LDAF=0.0646;AA=C;THETA=0.0002;AN=2184;VT=SNP;AC=141;AVGPOST=0.9988;SNPSOURCE=LOWCOV;RSQ=0.9908;ERATE=0.0002;AF=0.06;ASN_AF=0.06;AMR_AF=0.05;AFR_AF=0.13;EUR_AF=0.04 GT:DS:GL 0|0:0.000:-0.01,-1.70,-5.00
6 31831167 rs182547180 T C 100 PASS LDAF=0.0013;THETA=0.0004;AA=T;AN=2184;VT=SNP;RSQ=0.7096;SNPSOURCE=LOWCOV;ERATE=0.0003;AC=2;AVGPOST=0.9992;AF=0.0009;AFR_AF=0.0041 GT:DS:GL 0|0:0.000:-0.01,-1.66,-5.00

6. Use this vcf file with the variation effect predictor script to find which variants in this region have deleterious SIFT and PolyPhen effects

>perl variant_effect_predictor.pl -input ~/6_31830969_31846823.vcf -sift p -polyphen p --force_overwrite
2012-02-28 15:33:13 - Starting...
2012-02-28 15:33:13 - Detected format of input file as vcf
2012-02-28 15:33:13 - Read 211 variants into buffer
2012-02-28 15:33:13 - Analyzing chromosome 6
2012-02-28 15:33:13 - Reading transcript data from cache and/or database
[==========] [100%]
2012-02-28 15:33:19 - Retrieved 28 transcripts (0 mem, 0 cached, 28 DB, 0 duplicates)
2012-02-28 15:33:19 - Analyzing variants
[==========] [100%]
2012-02-28 15:33:21 - Calculating and writing output
[==========] [100%]
2012-02-28 15:33:24 - Processed 211 total variants

> less variant_effect_output.txt

rs150385253 6:31831478 C ENSG00000204385 ENST00000229729 Transcript NON_SYNONYMOUS_CODING 2080 2059 687 M/V Atg/Gtg - PolyPhen=benign;SIFT=tolerated

rs150385253 6:31831478 C ENSG00000204385 ENST00000544672 Transcript NON_SYNONYMOUS_CODING 2128 1831 611 M/V Atg/Gtg - PolyPhen=possibly_damaging;SIFT=tolerated

rs150385253 6:31831478 C ENSG00000204385 ENST00000375562 Transcript NON_SYNONYMOUS_CODING 1999 1933 645 M/V Atg/Gtg - PolyPhen=benign;SIFT=tolerated

6_31832657_C/T 6:31832657 T ENSG00000204385 ENST00000229729 Transcript NON_SYNONYMOUS_CODING 1883 1862 621 R/H cGc/cAc - PolyPhen=possibly_damaging;SIFT=deleterious

6_31832657_C/T 6:31832657 T ENSG00000204385 ENST00000544672 Transcript NON_SYNONYMOUS_CODING 1931 1634 545 R/H cGc/cAc - PolyPhen=possibly_damaging;SIFT=deleterious

6_31832657_C/T 6:31832657 T ENSG00000204385 ENST00000375562 Transcript NON_SYNONYMOUS_CODING 1802 1736 579 R/H cGc/cAc - PolyPhen=possibly_damaging;SIFT=deleterious

6_31832657_C/T has a deleterious effect when looking at both PolyPhen and SIFT predictions. There are other variants that one algorithm or the other call as damaging.

7. Use this vcf file with the variation pattern finder to look at the pattern of inheritance in this region

>perl variant_pattern_finder.pl -vcf ~/6_31830969_31846823.vcf -sample ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel -region 6:31830969-31846823

This should produce a file called chr6_31830969-31846823.txt. It is best to view this in a spreadsheet program

8. Use this vcf file with the vcf to ped converter to produce ped and info files for the CEU population

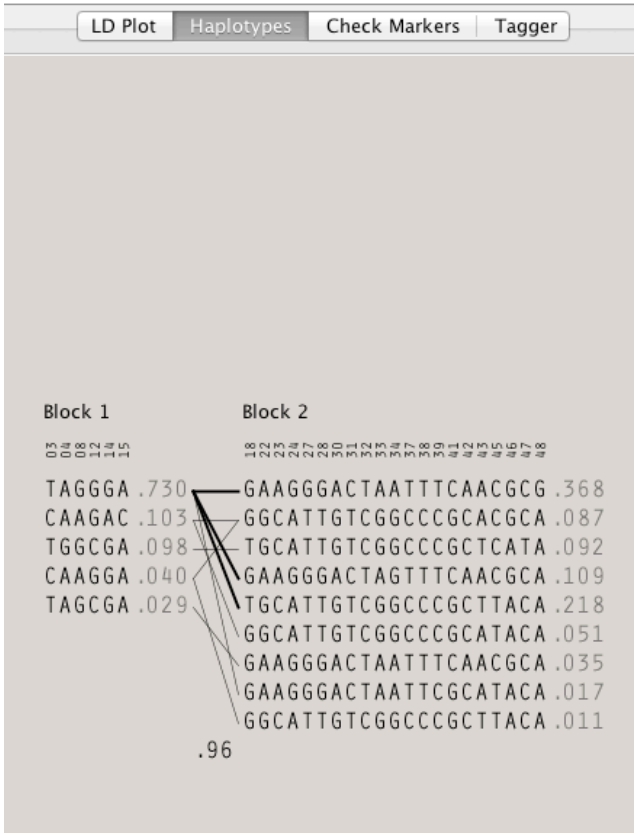
>perl vcf_to_ped_convert.pl -vcf ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr6.phase1_release_v2.20101123.snps_indels_svs.vcf.gz -sample ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel -region 6:31830969-31846823 -population CEU

Created 6_31830969-31846823.info and 6_31830969-31846823.ped

9a. 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.

9b. 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.