

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_wqs_genotypes/bcm/ALL.chr20.vqsr_site_v4_omni.20101123.geno Tue May types.vcf.qz file 457314670 3 14:23:41 2011 e05b0bf1f2e0dd694a26e92195c4cc91 ftp/technical/working/20110329_wgs_genotypes/bcm/ALL.chr20.vqsr_site_v4_omni.20101123.geno types.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_SH APEIT.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_SHA PEIT.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_SH APEIT.20120103.haplotypes.vcf.gz.tbi 57618 Tue Jan 3 15:28:22 2012 file 3af55a98cfc8a81c746b7105bb3ecb18 ftp/technical/working/20120103 omni_shapeit_haplotypes/ALL.chr5.omni_2123_samples_b37_SHA PEIT.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_SH APEIT.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_SH APEIT.20120103.haplotypes.vcf.gz.tbi 46339 Tue Jan 3 15:27:24 2012 file 8334495a3cd13a1b0d293faeeec3c972 ftp/technical/working/20120103_omni_shapeit_haplotypes/ALL.chr15.omni_2123_samples_b37_SH Tue Jan 3 15:10:16 2012 APEIT.20120103.haplotypes.vcf.gz 33379133 file 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_intensiti es/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_intensiti
es/Omni25 genotypes 2141 samples.b37.vcf.gz 6:31830969-31846823
    31833221
                 rs17207713 G
                                            PASS
6
                                   Α.
CR=99.9063;GentrainScore=0.5911;HW=1.0
6
    31833504
                 rs34418207 G
                                  Α
                                            PASS
                                        .
CR=99.85981;GentrainScore=0.7957;HW=1.0
    31834197
                 rs4947332
                            С
                                   Т
                                            PASS
6
                                      .
CR=99.95225;GentrainScore=0.7429;HW=0.80660635
                 SNP6-31944130 G
                                     А
                                              PASS
6
    31836151
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_intensiti es/Omni25_genotypes_2141_samples.b37.vcf.gz 6:31830969-31846823 | vcf-subset -c HG00096 #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HG00096 31833221 rs17207713 G PASS 6 . 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 rs4947332 31834197 С PASS 6 . . CR=99.95225;GentrainScore=0.7429;HW=0.80660635 GT:GC 0/0:0.7194 31836151 SNP6-31944130 G PASS 6 . . 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.

Human (GRCh37) V Location	n: 6:31,830,969-31,8	46,823 Gene: SL	C44A4						Tools I	Help
Gene-based displays				Cono: SI		200000004285)				
Gene summary				Gene: SL	C44A4 (ENSC	300000204385)				
 Splice variants (9) Supporting evidence 	Description	solute carrier family	44, member 4 [So	ource:HGNC Symbol;/	Acc:13941]					
- Sequence	Location	Chromosome 6: 31.830.969-31.846.823 reverse strand.								
 External references 	Transcripts There are 9 transcripts in this gene									
- Regulation										
Genetic variation Variation Table						-				
- Structural Variation	Show/hide columns Filter									
Variation Image	Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotyne	CCDS			
External Data	SI C4484-001	ENST0000229729	2580	ENISP00000220720	710	Protein coding	CCDS4724			
Gene history	SLC44A4-001	ENST00000414427	1000	ENSP0000228725	411	Protoin coding	00004724			
dene matory	SLC44A4-004	ENCT00000414427	1200	ENGRODO000350301	411	Protein couling	-			
🥕 Configure this page	SLC44A4-201	ENST0000375562	2505	ENSP00000364712	668	Protein coding	-			
	SLC44A4-202	ENS10000544672	2634	ENSP0000444109	634	Protein coding	-			
📭 Manage your data	SLC44A4-002	ENS100000465707	681	No protein product	-	Processed transcript	-			
	SLC44A4-003	ENST00000462671	426	No protein product		Processed transcript	-			
🔒 Export data	SLC44A4-007	ENST00000487680	392	No protein product	-	Processed transcript	-			
	SLC44A4-005	ENST0000475563	575	No protein product	-	Retained intron	-			
Get VCF data	SLC44A4-006	ENST00000479777	655	No protein product	-	Retained intron				
- Declarated this second										

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.

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 Т PASS 6 31831004 TTTTG 39 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:GI 0 | 0:0.000:0.00,-0.30,-7.60 6 31831159 100 PASS rs3869144 С Т LDAF=0.0646;AA=C;THETA=0.0002;AN=2184;VT=SNP;AC=141;AVGPOST=0.9988;SNPSOURCE=L OWCOV;RSQ=0.9908;ERATE=0.0002;AF=0.06;ASN AF=0.06;AMR AF=0.05;AFR AF=0.13;EUR AF 0|0:0.000:-0.01,-1.70,-5.00 =0.04 GT:DS:GL Т 6 31831167 rs182547180 С PASS 100 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 ſ 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 ſ 100%] 2012-02-28 15:33:21 - Calculating and writing output ſ 100%]

2012-02-28 15:33:24 - Processed 211 total variants

> less variant effect output.txt rs150385253 6:31831478 С ENSG0000204385 ENST00000229729 Transcript 2080 NON SYNONYMOUS CODING 2059 687 M/V Atg/Gtg -PolyPhen=benign;SIFT=tolerated rs150385253 6:31831478 С ENSG0000204385 ENST00000544672 Transcript NON SYNONYMOUS CODING 2128 1831 611 M/V Atg/Gtg -PolyPhen=possibly damaging;SIFT=tolerated rs150385253 6:31831478 С ENSG0000204385 ENST00000375562 Transcript NON SYNONYMOUS CODING 1999 1933 645 M/V Atg/Gtg -PolyPhen=benign;SIFT=tolerated 6 31832657 C/T 6:31832657 ENSG0000204385 ENST00000229729 Transcript Т NON_SYNONYMOUS_CODING cGc/cAc -1883 1862 621 R/H PolyPhen=possibly_damaging;SIFT=deleterious 6_31832657_C/T 6:31832657 ENSG0000204385 ENST00000544672 Transcript Т cGc/cAc -NON SYNONYMOUS CODING 1931 1634 545 R/H PolyPhen=possibly_damaging;SIFT=deleterious 6 31832657 C/T 6:31832657 Т 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.

000	Welcome to HaploView	
Linkage Format	laps Format HapMap Format HapMap PHA	SE 🕨
Data Fil	e: wnloads/6_31830700-31840700.ped Brows	e
Locus Information File	e: wploads (6, 31830700-31840700 info	
	Chromosome Do association test	
•	Family trio data 🔷 Case/Control data	
	● Standard TDT ○ ParenTDT	
Test list file (optional): Brows	e
Ignore pairwise	e comparisons of markers > 500 kb apart.	
Exclude in	dividuals with > 50 % missing genotypes.	
		Cattinen
	OK Cancel Proxy	Settings
File Display Analysis Help	OK Cancel Proxy S Haploview 4.2 6_31830700-31840700.ped 1000000000000000000000000000000000000	Settings
5 16 17 18 19 20 21 22	OK Cancel Proxy Prox Prox Prox Prox Prox Prox Prox Prox Prox Prox	Settings ³⁷ ³⁸ ³⁹
File Display Analysis Help 5 16 17 18 19 20 21 22 59 74 16	OK Cancel Proxy S Haploview 4.2 6_31830700-31840700.ped ID Plot Haplotypes Check Markers Tagger 23 24 25 26 27 28 29 30 31 32 33 34 35 36	Settings Kr
File Display Analysis Help 5 16 17 18 19 20 21 22 5 74 16 7	OK Cancel Proxy Haploview 4.2 6_31830700-31840700.ped ID Plot Haplovipes Check Markers Tagger 23 24 25 26 27 28 29 30 31 32 33 34 35 36	Settings Kr
File Display Analysis Help 5 16 17 18 19 20 21 22 74 12 72 72 72 72	OK Cancel Proxy Haploview 4.2 6_31830700-31840700.ped Haploview 4.2 6_31830700-31840700.ped ID Pior Haplovipes Check Markers Tagger 23 24 25 26 27 28 29 30 31 32 33 34 35 36	Settings Kr
File Display Analysis Help 5 16 17 18 19 20 21 22 74 16 74 72 72 72 72 13	OK Cancel Proxy S Haploview 4.2 6_31830700-31840700.ped Haploview 4.2 6_31830700-31840700.ped Haploview 4.2 6_31830700-31840700.ped D Piot Haplotypes Check Markers Tagger 23 24 25 26 27 28 29 30 31 32 33 34 35 36 4 4 4 4 4 4 4 4 4 4 4 4 5 36 4 4 5 36 4 4 5 36 4 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4 5 36 4	Settings Kr 37 38 39 90 90 90 90 90 90 90 90 90 9
File Display Analysis Help 5 16 17 18 19 20 21 22 74 12 72 72 72 72 72 13 63 63 72	OK Cancel Proxy S Haploview 4.2 6_31830700-31840700.ped Haploview 4.2 6_31830700-31840700.ped 1 10 Piol Haplotypes Check Markers Tagger 23 24 25 26 27 28 29 30 31 32 33 34 35 36 4	Settings Ke
File Display Analysis Help 5 16 17 BIOCK 2 (3 KD) 20 21 22 5 74 16 74 12 72 72 72 13 63 63 72 72	OK Cancel Proxy S Haploview 4.2 6_31830700-31840700.ped	Settings
File Display Analysis Help 5 16 17 018 19 20 21 22 74 16 72 74 63 63 72 72 63 63 63 63 63	OK Cancel Proxy 9 Haploview 4.2 6_31830700-31840700.ped Haploview 4.2 6_31830700-31840700.ped Image: Check Markers Tagger Tagger 23 24 25 26 27 28 29 30 31 32 33 34 35 36 23 24 25 26 27 28 29 30 31 32 33 34 35 36 4	Settings 50 50 50 50 50 50 50 50 50 50
File Display Analysis Help 5 16 17 18 19 20 21 22 74 10 74 63 63 63 63 63 63 63 63 63 63	OK Cancel Proxy 9 Haploview 4.2 6_31830700-31840700.ped Haploview 4.2 6_31830700-31840700.ped 1 3	Settings x x x x x x x x x x x x x
File Display Analysis Help 5 16 17 18 19 20 21 22 74 16 74 63 63 63 63 63 63 63 63 63 63 63 63 63	OK Cancel Proxy 9 Haploview 4.2 6_31830700-31840700.ped DPior Haplotypes Check Markers Tagger 23 24 25 26 27 28 29 30 31 32 33 34 35 36 72 72 72 72 72 72 72 72 72 72 72 57 40	Settings

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





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