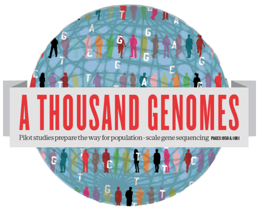
A THOUSAND GENOMES
Pilot studies prepare the way for population-scale gene sequencing

The 1000 Genomes Command line Tutorial Exercises.

These exercises should present a set of questions whose answers can be found in our tutorial slides that are found on: http://www.1000genomes.org/using-1000-genomes-data

All these questions can be answered using command line tools and data from our ftp site.

The software requirements for this tutorial are:

Samtools http://samtools.sourceforge.net/

Tabix http://sourceforge.net/projects/samtools/files/tabix/
(Please note it is best to use the trunk svn code for this as the 0.2.5 release has a bug)
svn co https://samtools.svn.sourceforge.net/svnroot/samtools/trunk/tabix

Vcftools http://vcftools.sourceforge.net/

The ensembl variation and core apis
http://www.ensembl.org/info/docs/api/index.html

Please note the following two perl scripts need either/both the vcftools perl code or the ensembl core and variation api to function properly.

The variant effect predictor ftp://ftp.ensembl.org/pub/misc-scripts/Variant_effect_predictor/
http://www.ensembl.org/info/docs/variation/vep/index.html#about

The variation pattern finder
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/variation_pattern_finder/version_1.0

VCF to PED Converter
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/vcf_to_ped_converter/version_1.0/

Haploview

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

Finding Data

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

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

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

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

Using the 1000 Genomes Tools

3. Use the browser to find location the SLC44A4 gene.

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

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

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

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

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

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

9a. Look at these files in haploview.

9b. How many haplotype blocks does haploview think there are in this section?