Using 1000 Genomes Tutorial Exercises 11th April 2012

Web Based Exercises

Finding data

1a. Find what Omni VCF files we have on our ftp site using the website ftp search. (Omni is a high throughput genotyping platform from Illumina on which all 1000 genomes samples are being genotyped)

1b. Find the most recent Omni VCF file on GRCh37 from the 31st January 2012

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.

Using the Browser

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

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

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

6. In which gene is the variant found?

Using the 1000 Genomes Tools

Use the browser to find the SLC44A4 gene.

7. 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.

8. Unzip this VCF file using a tool like winzip or Archive Utility.

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

10. Do any of the variants have negative Sift or Polyphen predictions?

11. 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

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

13. Look at these files in haploview.

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

Command Line Exercises

File formats and meta data

15. How many GRCh37 omni vcf files are in technical/working

16. Which exome sample from 20110521 has the highest percentage of targets covered at 20x or greater

17. Find the exome bam file for this sample

18. Get a slice of this exome bam file between 7:114173990-114175942

Command line tools

19. Get a slice of HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam for 7:114304000-114305000 (FoxP2)

20. Get the equivalent section of the 20110521 release chr 7 genotypes file

21. Use vcftools vcf-stats to specify which SNP transition happens most in this section

22. Use this piece with tools, the variant effect predictor, the vcf pattern finder and the vcf to ped converter

23. Are there any snps with deleterious sift/polyphen consequences?

24. What is the most common pattern of variation in this region?

25. Use haploview to discover how many different haplotype blocks does the section contain for the CEU population?