

Please note the software on these machines has not been fully tested and some of the answers to these questions might not work

Command Line Exercises

File formats and meta data

1. How many GRCh37 omni vcf files are in technical/working
2. Which exome sample from 20110521 has the highest percentage of targets covered at 20x or greater. You need to look at the 20110521.alignment.index.HsMetrics.gz file to find this
3. Find the exome bam file for this sample
4. Get a slice of this exome bam file between 7:114173990-114175942

Command line tools

5. Get a slice of HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam for 7:114304000-114305000 (FoxP2 exon)
6. Get the equivalent section of the 20110521 release chr 7 genotypes file [\[1\]](#)
7. Use vcftools vcf-subset to get the genotypes for HG00737, does HG00737 have any variant sites in this location?
8. Use this piece with tools, the variant effect predictor, the vcf pattern finder
9. Are there any snps with deleterious sift/polyphen consequences?
10. What is the most common pattern of variation in this region?
11. Use the vcf to ped script with 6:31830700-31840700 and population CEU
12. How many different haplotype blocks does the section contain?