Using 1000 Genomes Tutorial Exercises 12th April 2012

Finding and Handling Data

1. How many Omni VCF files can you find on the ftp site (Omni is a high throughput genotyping platform from Illumina on which all 1000 genomes samples are being genotyped)

```
> wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/current.tree
> grep omni current.tree | cut -f1 | grep vcf | grep -v tbi | wc –l
> 52
```

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

```
> grep omni current.tree | cut -f1 | grep vcf | grep -v tbi  | grep 20120131 | grep b37 | 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
```

3. 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 box which is in the top right hand corner of every page of 1000genomes.org with the term sub-section and vcf should return the appropriate FAQ page

4. Which exome sample from 20110521 has the highest percentage of targets covered at 20x or greater using the 20110521.exome.alignment.index.HsMetrics.gz file and PCT_TARGET_BASES_20X column

```
> zcat 20110521.exome.alignment.index.HsMetrics.gz | cut -f1,31 | sort  -k2 –n | tail –n1
> HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam   0.932651
```

5. Find the exome bam file for this sample

```
> grep  HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam current.tree  | grep -v "bam\." | awk '{print "ftp://ftp.1000genomes.ebi.ac.uk/vol1/"$1}'
>ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/HG00737/exome_alignment/HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam
```

6. Get a slice of this exome bam file between 7:114173990-114175942 (exon of FOXP2)

```
> samtools view
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/HG00737/exome_align
ment/HG00737.mapped.illumina.mosaik.PUR.exome.20110411.bam
7:114173990-114175942 | tail -n1
> SRR099984.44615561     83     7      114174990      65     76M     =
114174660      -405
GAACCATATTTGGTGTACATAGGCATAAAGAATTTTGCATAAAACCCCCTTGTGGGA
TTTTATTCATACATAGGTT
SD@GIB>BFDDHDCDBBJCAFHHJBBDDEHDBFFDCHJB<CCC4IIHHIECGCGGGAEE
E@AEBH??@H@?CFDBS    RG:Z:SRR099984  NM:i:0
OQ:Z:DE@DEE?EEBEGEDEGFHHFGHHHHGHHFHHGHHDHHHHHGHHDHHGGGH
HHHHHHHHHHHHHHGFHHHHGHHHHH
```

<u>Command Line Tools</u>

7. Get the 7:114304000-114305000 (FoxP2 exon) section of the 20110521 release chr 7 genotypes file

```
> tabix –h
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr7.phase1_
release_v3.20101123.snps_indels_svs.genotypes.vcf.gz   7:114304000-
114305000 > 20110521.vcf
```

8. Use vcftools vcf-stats to specify which SNP transition happens most in this section. You should look at the statistics for all variants/samples.

```
> vcf-stats 20110521.vcf
> 'G>A' => 5
```

9. Use this piece with tools, the variant effect predictor, the vcf pattern finder

```
>perl variant_effect_predictor.pl -input ~/20110521.vcf -sift p -polyphen p --
force_overwrite
```

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

There are several snps with deleterious effects

e.g

```
> grep SIFT variant_effect_output.txt
> rs182138317     7:114304331     A      ENSG00000128573 ENST00000393489
Transcript     NON_SYNONYMOUS_CODING   1949    1567    523     A/T     Gcc/Acc
-      PolyPhen=possibly_damaging;SIFT=deleterious
```

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

```
> perl variant_pattern_finder.pl -vcf ~/20110521.vcf -sample
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel -region 7:114304000-114305000
```

This produces a tsv file which can be view in a spreadsheet program

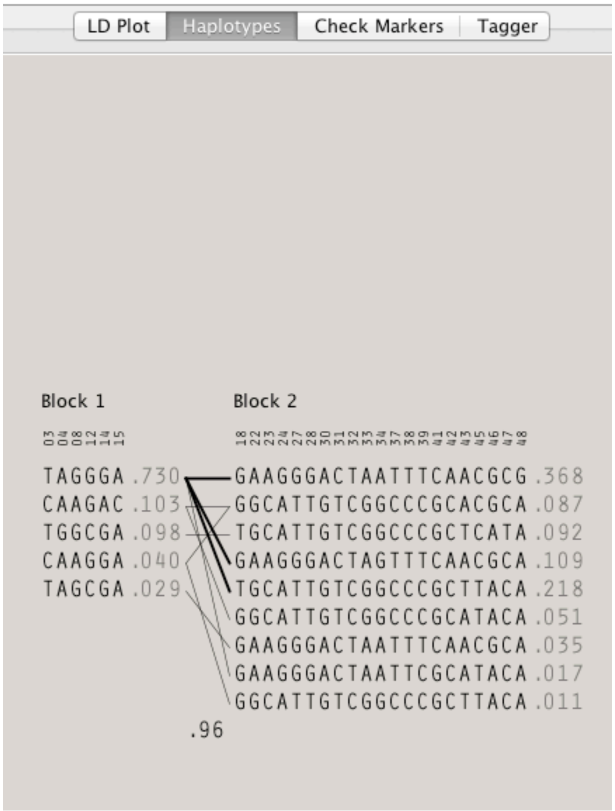
```
        7:114304563_rs1378771[C]  7:114304630_rs1378772[A] 7:114304969_rs2396765[T]
> 14.38 - - C|T - A|T - - - - - T|C TSI(30)
```

12. Use the vcf to ped script with 6:31830700-31840700 for population CEU

```
> perl vcf_to_ped_convert.pl -vcf 20110521.vcf -sample
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel -region 6:31830700-31840700  -population CEU
> ls ./
> 6_31830700-31840700.info
> 6_31830700-31840700.ped
```

13. How many different haplotype blocks does the section contain?

This region contains 2 haplotype blocks



```
Block 1                   Block 2

03 04 08 12 14 15         18 22 23 24 27 28 30 31 32 33 34 37 38 39 41 42 43 45 46 47 48

TAGGGA .730               GAAGGGACTAATTTCAACGCG .368
CAAGAC .103               GGCATTGTCGGCCCGCACGCA .087
TGGCGA .098               TGCATTGTCGGCCCGCTCATA .092
CAAGGA .040               GAAGGGACTAGTTTCAACGCA .109
TAGCGA .029               TGCATTGTCGGCCCGCTTACA .218
                          GGCATTGTCGGCCCGCATACA .051
                          GAAGGGACTAATTTCAACGCA .035
                          GAAGGGACTAATTCGCATACA .017
                          GGCATTGTCGGCCCGCTTACA .011
                    .96
```