Variant calling exercises

A pilot study is being prepared in your laboratory. This study tries to explore the variability of the different populations around the world. To complete this, the department of primary analysis has prepared the following *bam* files:

```
HG00240_GBR__chr21.bam
HG01083_PUR__chr21.bam
HG01504_IBS__chr21.bam
NA18974_JPT__chr21.bam
NA19236_YRI__chr21.bam
```

They correspond to 5 individuals from 5 different populations (GBR, IBS, YRI, PUR and JPT, see http://www.1000genomes.org/category/population/). Your goal is to perform a small study to compare the variability between selected individuals in order to extrapolate the conclusions to the rest of the world populations.

Roadmap:

1- Tailoring you bam file

Apply required preprocessing steps (only mark duplicates and filtering of low quality reads)

2- Predict variants

Call variants for prepared bam files. Also, label variants by quality.

3- Annotate variants

Use *snpEff* via command line to annotate the effect of variants.

4- Webtools

Upload your high quality variants to *wAnnovar* (http://wannovar.usc.edu/index.php) and *Cravat* (http://www.cravat.us/) to explore relevant annotations.

5- Extended exercise

Use provided R script to evaluate your data. Also, obtain conclusions the variability of your individuals (populations).

6- Important questions you need to address

- Which proportion of reads is finally filtered?
- How many variants have you found?
- How many high quality variants do you obtain?
- How many LOF variants are present in your individuals?
- Do you think it's reasonable to find LOF variants in healthy individuals?
- Is there any cancer related variant?
- Are the individuals properly spread depending on their population?

Commands guide

Remove duplicates

```
➢ samtools rmdup sample.bam sample_nodup.bam
```

Filter reads with MQ < 10

```
➢ samtools view -h -b -q 10 -o sample_nodup_q10.bam sample_nodup.bam
➢ samtools index sample_nodup_q10.bam
```

Obtain simple bam stats

```
➢ samtools flagstat sample.bam
```

Call variants

```
➢ ls *q10.bam > bam.list
➢ java -jar software/GenomeAnalysisTK.jar -T UnifiedGenotyper -nt 4 -maxAltAlleles
  1 -glm SNP -R resources/hs37d5.fa -I bam.list -o raw_variants.vcf
```

Label variants by quality

```
➢ java -jar software/GenomeAnalysisTK.jar -T VariantFiltration -R
  resources/hs37d5.fa --filterExpression "QD < 2.0 || MQ < 40.0 || FS > 60.0 ||
  HaplotypeScore > 13.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"
  --filterName "STD_FILTER" -V raw_variants.vcf -o raw_variants_labeled.vcf
```

Annotate mutation effect

```
➢ java -Xmx4g -jar software/snpEff/snpEff.jar -v GRCh37.70 -canon
  raw_variants_labeled.vcf > raw_variants_labeled_snpeff.vcf
```

Select High quality variants

```
➢ grep -v STD_FILTER raw_variants_labeled_snpeff.vcf >
  raw_variants_labeled_snpeff_hq.vcf
```

Select High quality variants with high impact

```
➢ grep -v STD_FILTER raw_variants_labeled_snpeff.vcf | grep -v MODIFIER | grep -v
  LOW > raw_variants_labeled_snpeff_hq_higheffect.vcf
```

Get some general stats

```
➢ software/plinkseq-0.10/pseq raw_variants_labeled_snpeff_hq.vcf i-stats
➢ software/plinkseq-0.10/pseq raw_variants_labeled_snpeff_hq.vcf v-stats
```

Extended exercise (R)

```
➢ software/bcftools/bcftools query -f
  '%CHROM\t%POS\t%REF\t%ALT\t%FILTER\t%INFO/EFF[\t%GT]\n'
  raw_variants_labeled_snpeff.vcf -o raw_variants_labeled_snpeff.txt
➢ Rscript software/compare_populations.r
```