Genetic variant analysis hands-on lab

Required software

BCFtools
VCFtools
VG

Before you start

1. Make sure all the required tools are in your UNIX PATH and can be called from command-line.
2. Data are in `/home/data/Desktop/data/`

Write all the commands you run and your answers in a Jupyter Notebook.

To do the assignments refer to respective tools manuals:

- VCFtools
- BCFtools
- VG

Working with VCF files

We will work on VCF file from 1000 Genome Project (1KGP) phase 3. The VCF file has been obtained from whole genome genotyping experiment. In this experiment were targeted population specific SNPs. For the genotyping was used the Omni Broad-Sanger combined chip.

If working on InfOmics server, the VCF file can be reached with
`/home/data/Desktop/data/vcf/ALL.chip.omni_broad_sanger_combined.20140818.snps.genotypes.vcf.gz`

1. **How many individuals were genotyped? How many SNPs are available in the VCF?**

2. **Restrict the analysis to chr22. How many variants were called on chr22? How many variants were mapped between positions 15000000 and 25000000 on chr22?**

3. **Write to a new VCF file the SNPs mapped on chr22. Name the output file "chr22subset". Keep all data from INFO field. If not already Gzipped, compress it. (Hint: use bgzip). Store the file in YOUR PERSONAL Desktop directory.**

4. **Retrieve the allele frequencies from the original VCF. How many SNPs has the alternative allele frequency > 0.5?**

5. **Convert the VCF file to the corresponding BCF.**

Working with VG

Here we will showcase some of the main functionalities of VG software suite. As mentioned above, for detailed description of VG's functionalities refer to the github's wiki.

If working on InfOmics server, the FASTA reference and the VCF are stored in

/home/data/Desktop/data/vg/

1. **Build a VG, using xy.fa as reference genome and enrich it with variants contained in xy.vcf.gz. Build a graph for each of the two sequences available in xy.fa. (Hint: use -R option, while calling VG). The graphs must be named x.vg and y.vg.**

2. **Retrieve the PNG image of x.vg and y.vg. (Hint: use 'vg view' in pipe with 'dot'). Do you notice any difference between the two graphs?**

3. **Index the two graphs and track the haplotypes. (Hint: study 'vg index' help).**

4. **Extract all the 8-mers from position 1 to position 150 from x.xg. Store the result in kmers.tsv. How many k-mers are found on the reference genome? How many on alternative haplotypes? List all the non reference k-mers having haplotype frequency == 2.**