Bioinformatics Workshop Exercise Handout #3

Day 2, AM: Variant Calling and Filtration
Required Dataset: hardfilter_S.vcf.gz
Required Programs: BASH, R, RStudio, BCFtools, VCFtools

GitHub and RMarkdown Exercise:

1. Navigate to https://github.com/
2. Find rflamio
3. Study the sockeye_GWAS repository
   Repositories > sockeye_GWAS
   - README.md files provide instructions to the reader about repository contents
   - Can include R Markdowns, scripts, datasets (including intermediate files), etc.
4. Download 'SockeyeVariantCalling.Rmd' and open in R Studio
5. Observe prelude, non-chunks, and chunks.
6. Observe Table of Contents on right-hand side.

Genome Analysis Toolkit Website Exercise:

1. Navigate to https://gatk.broadinstitute.org/hc/en-us.
2. Click on the workflow 'Germline short variant discovery' and glance at the sections of the documentation.
3. Click on the HaplotypeCaller tool under tool index 4.4.0.0. Notice the different sections and arguments.

GenomicsDB Workspace Exercise:

1. Navigate to the consolidate_S.sh script on my github page and we will look at the script together.
   https://github.com/rflamio/sockeye_GWAS/blob/main/VariantCalling/consolidate_S.sh

BCFtools and VCFtools Exercise:

Given the VCF file hardfilter_S.vcf.gz

1. Count the number of variants in the dataset.
2. Retain only biallelic SNPs.
3. Remove sites with a MAF < 0.2. Recode.
   1. How many sites were removed?
   2. How many sites remain?
4. Remove individual tS01. Recode.
5. Remove variants with > 5% missing data. Recode.
6. How many SNPs and individuals are in the final dataset?