Homework/Mini Project 2

W. Evan Johnson, Ph.D.
Director, Center for Data Science
Rutgers New Jersey Medical School

Due May 30, 2023

Now its time to practice what we have learned in class and learn even more! For this homework/mini project you will do a complete GWAS analysis of the Ogden's Syndrome dataset. Note that from now on your homework should be written in R Markdown, and turned in by uploading a tarball with your .Rmd, .html (from Rmarkdown, MultiQC, etc) and other outputs (e.g. .vcf) on Canvas.

DNA-sequencing and GWAS analysis

1. To access the data for Homework 2, please download the following (just normal point and click, not need to curl or wget): https://www.dropbox.com/scl/fi/ezilkcxjjhlcft20965fz/ogdens_data.tar?rlkey=jshzg6d0h7e7q58hucue2h0up&st=pa8xh3jo&dl=0
2. Use `cat` in Unix to merge the `proband_29.fq.gz` and the `proband_short.fg.gz` files – these are from the same sample
3. Use `FASTQC` and `MultiQC` to summarize the fastq files for these datasets. For which diagnostics do these data fail? Should we be concerned about the quality of these data? Why or why not?
4. Align genome sequences from these human sequencing experiment using `bwa` to the `chrX_5MB.fa` reference.
5. Process the reads through `samtools mpileup` to generate a .vcf file.
6. Complete this process using `gatk` to generate an alternative .vcf file. How do the .vcf files compare?

7. Generate a Manhattan plot, comparing the proband sequences with the sequences from the brother and uncle. List the SNPs/regions of interest.
8. Are the mother and grandmother heterozygous at these positions? Please show your work. Why is this important.