Homework 1: SNP and GWAS Analysis

100 Points Total

Due May 21, 2025

Now it's time to apply what you've learned during the first three weeks of class—and to deepen your understanding! In this assignment, you will perform a complete Genome-Wide Association Study (GWAS) using the Ogden's Syndrome dataset.

Please save all your code for each part of the assignment. Your instructor may request it, and it may also be used for awarding partial credit if needed. We strongly recommend completing the assignment in an RMarkdown document to ensure your work is well-documented and reproducible.

Submit your completed assignment via the designated homework submission site on Canvas.

DNA-sequencing and GWAS analysis

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

1. What command did you use to extract this tarball for this repository:
   (a) unzip ogdens_data.tar
   (b) tar -czf ogdens_data.tar
   (c) tar -xvf ogdens_data.tar
   (d) tar –remove-files ogdens_data.tar
2. Use an appropriate Unix command to to merge the 'proband_29.fq.gz' and the 'proband_short.fg.gz' files as these are from the same sample. The following can be used:
   (a) cat proband_29.fq.gz proband_short.fq.gz > proband_merged.fq.gz
   (b) cat proband_29.fq.gz > proband_short.fq.gz > proband_merged.fq.gz
   (c) merge proband_29.fq.gz proband_short.fq.gz > proband_merged.fq.gz
   (d) gzip proband_29.fq.gz proband_short.fq.gz > proband_merged.fq.gz
3. Use FASTQC and MultiQC to summarize the FASTQ files for these datasets. Should we be concerned about the quality of these data? Why or why not?
   (a) No—overall quality scores are strong, and the minor issues found are typical of short-read sequencing and can be ignored.
   (b) Yes—there are multiple indicators of potential problems, including duplication, GC bias, and adapter contamination.
   (c) No—FASTQC and MultiQC do not typically reveal serious quality concerns unless there is complete sequencing failure.
   (d) Yes—all samples failed the sequence quality histogram and length distribution checks.

4. For which diagnostics do these data fail, based on the MultiQC summary of the FASTQC results?
   (a) Only per-base sequence quality and sequence length distribution failed across all samples.
   (b) The samples failed in having uniformly high quality and balanced GC content.
   (c) Several samples had high duplicate read levels, unbalanced GC content, one had poor per-tile sequence quality, and one showed excessive adapter contamination.
   (d) There were no actual failed diagnostics—MultiQC flagged warnings, not failures.
5. Align genome sequences from these human sequencing experiment using 'bwa' to the 'chrX_5MB.fa' reference. Which is the appropriate unix command for this purpose?
   (a) bwa mem ../genome/chrX_5MB.fa proband_merged.fq.gz > proband.sam
   (b) bwa aln ../genome/chrX_5MB.fa proband_merged.fq.gz > proband.sam
   (c) bwa mem proband_merged.fq.gz ../genome/chrX_5MB.fa > proband.sam
   (d) bwa mem -i ../genome/chrX_5MB.fa proband_merged.fq.gz -o proband.sam
6. Process the reads through 'samtools mpileup' to generate a .vcf file. How many lines are in your final .vcf file for the proband?
   (a) 1303522
   (b) 320145
   (c) 57890
   (d) 2049831
7. Complete this process using GATK to generate an alternative .vcf file. How do the .vcf files compare? Which of the following is NOT true?
   (a) GATK and pileup agree on a small number of high-confidence variants.
   (b) Pileup reports many more candidate variants than GATK due to fewer filters.
   (c) GATK uses additional quality metrics and statistical models to reduce false positives.
   (d) GATK and pileup produce nearly identical sets of variant calls across the genome.
8. Generate a Manhattan plot, comparing the proband sequences with the sequences from the brother and uncle. List the SNPs/regions of interest. Upload a .png or .jpeg of your Manhattan plot here.
9. What is the genotype of the reference and the proband at position chrX_149913753:2939281? This is the disease-causing SNP.
   (a) Reference = A, Proband = G
   (b) Reference = A, Proband = A
   (c) Reference = G, Proband = A
   (d) Reference = G, Proband = G
10. Are the mother and grandmother heterozygous at position chrX_149913753:2939281? What is their genotype at this location, and why is this important?
   (a) Yes, both are heterozygous (A/G). This is important because the mutation is X-linked and recessive. Female carriers (mother and grandmother) can pass the mutation to male offspring, who are affected if they inherit the variant.
   (b) No, both are homozygous for the reference allele (A/A). This would suggest a de novo mutation in the proband, which is inconsistent with inheritance patterns.

(c) Yes, both are homozygous for the alternate allele (G/G). This would likely mean they are also affected, inconsistent with X-linked recessive inheritance in females.

(d) No, the mother is heterozygous (A/G) and the grandmother is homozygous reference (A/A), which would break the expected pattern of inheritance.