Statistical Genomics
M-Stat (2nd Year), First Semester
Final Semester Examination 2025-26

Date: November 28, 2025 **Time: 3 hours**

The paper carries 100 marks. Answer all questions.

1. (a) In a GWA study the goal is to find the SNPs associated with cholesterol levels measured by three quantities, i.e. HDL, LDL and Triglycerides. Accordingly, from a sample of 1000 subjects, phenotypes are measured and the SNP specific genotype data are also recorded for 560 SNPs.
   Propose a suitable Statistical approach for a single SNP GWA study. Write down the model, set of hypotheses to be tested, and the test procedure. Describe the problem of multiple testing and discuss the BH algorithm for controlling false discovery rate.
   (b) Develop a penalized regression-based approach for the same study. Mention two serious limitations of this approach. How can you handle those limitations?
   [15+5=20]

2. When John started his career as a post-doctoral fellow in the famous LA labs in the United States, he met Nancy who was about to finish her PhD from the same lab. Later Jon figured out that he might be genetically related to Nancy and discovered that his grand-father and Nancy's (maternal) grand-mother were double first cousins. A wonderful coincidence!
   Consider a biallelic locus with alleles A and a, and at the population level P(A)=0.75. Find the probability that at this locus the genotype of John and Nancy will be Aa and aa, respectively. [15]

3. (a) The sequences below represent an optimum alignment of the first 50 nucleotides from the human and sheep preproinsulin genes. Estimate the number of substitutions that have occurred in this region since humans and sheep last shared a common ancestor using the Jukes-Cantor model.

Human: ATGGCCCTGT GGATGCGCCT CCTGCCCCTG CTGGCGCTGC TGGCCCTCTG

Sheep: ATGGCCCTGT GGACACGCCT GGTGCCCCTG CTGGCCCTGC TGGCACTCTG

(b) Using the number of substitutions calculated in (a), and assuming that humans and sheep last shared a common ancestor 100 million years ago, estimate the rate at which the

sequence of the first 50 nucleotides in their preproinsulin gene has been accumulating substitutions.

(c) Would the rate of mutation be greater or less than the observed substitution rate for a sequence of a gene such as the one shown in (a)? Why?

(d) Use UPGMA (Average Linkage Method) to construct a phylogenetic tree using the following distance matrix. Did you obtain the correct tree? If not, explain why.

Species	A	B	C	D
B	3	-	-	-
C	6	5	-	-
D	9	9	10	-
E	12	11	13	9

[30]

4. Researchers believed that consumption of red meat and physical activities might play key roles in the evolution of a specific type of cancer for humans. Accordingly, a study was conducted with 3000 participants from Europe, where binary responses (yes/no) were recorded for the two factors mentioned above, and 30,000 SNPs were genotyped along with the locations of 7 important genes already reported in the literature before. Using a multivariate dimensionality reduction (MDR) approach how can you test if the consumption of red meat and physical activities interact with genes for the development of cancer? [15]

5. Paper presentation as per class assignment. [20]