Population Genetic Analyses of Genomic Data: Assignment 1

1 Exercises

1.1 Measurement of variance

A population of insects, assumed to be identical at the beginning of the experiment, is divided into two subpopulations. One subpopulation is treated with a mild insecticide, while the other population is left untreated. Each population ls left to grow for some time until at the end of the experiment, the frequency of a given allele is measured in each population. Samples of a given allele frequency within each population are taken, according to Table 1:

Table 1: Collected data.

	Sample size	Allele frequency
Selected population	124	0.43
Unselected population	176	0.34

a) Calculate F_{ST} for this allele. Show your working in doing this.

b) Suppose that you conduct a statistical analysis and obtain the result that your value of F_{ST} is significantly high. What biological conclusions would you draw from this result? You do not need to conduct the statistical analysis.

1.2 Modelling fitness in a diploid system

We consider an organism in the wild, which is susceptible to infection by a parasite. We suppose that at any given time, a fraction $0 \leq p_i \leq 1$ of individuals are infected by the parasite.

An allele at a given genetic locus conveys some protective effect against the parasite upon the organism. We suppose that at this locus, organisms have the allele A with frequency p, and the allele a with frequency $q = 1 - p$.

We suppose that the fitness of individuals depends entirely upon whether or not they are infected with the parasite, and upon their genetic composition at the given locus, as follows:

Table 2: Fitnesses of individuals.

	Genotype		
	AA	Aa	aa
Infected	a	b	c
Not infected	1	1	c

for some a, b, c satisfying $0 < a < 1$, $0 < b < 1$, and $0 < c < 1$.

a) Write down the Hardy-Weinberg proportions for the number of individuals with each genotype, AA, Aa, and aa.

b) Calculate the mean fitness of the population at these frequencies when $p_i = 0.9$, $a = 0.8$, $b = 0.9$, $c = 0.7$, and $p = 0.5$.

c) For these values of p_i, a, b, and c, find the frequency p that gives the maximal mean population fitness.

d) Given the above values for p_i, a, and c, which values of b would lead to the allele a being maintained in the population over a long period of time?

1.3 Identifying beneficial mutations in a neutral marker experiment

a) Download the Optimist software package from the Sanger Institute website address *http://www.sanger.ac.uk/resources/software/optimist/*. Instructions for using the program are contained in the file “Instructions.rtf”. The file “Colour freq.out” contains data from an evolutionary experiment. In this exercise you will analyse these data using the Optimist code.

b) Compile all of the code (you will need a C++ compiler and access to GSL code libraries) and run the initial optimization program, analyse_first_multi, to test scenarios with a single mutation. It is suggested to use the value 10 for the number of optimisations. What is the maximum likelihood produced by the code for the given data?

c) Run the systematic search algorithm on the results from your first optimization. This performs a second, more thorough, round of optimization, in which the timings of mutations are restricted to integer values. The search method requires an input file named Transfer data.out, which needs to contain the first 6 lines of the file Transfer data multi.out, produced by the first optimisation method. What is the final likelihood produced by this code? Plot the inferred trajectories, contained in the output file Model frequencies.out, against time, comparing with the observed allele frequencies, contained in the file Real frequencies.out.

d) Repeat the optimisations of steps **b)** and **c)** with two and three mutations. Note that, where there are m mutations, the top m+5 lines of the file Transfer data multi.out are required to form the file Transfer data.out, required by the systematic optimization. In each case, plot the frequencies produced by the best fitting model. How do the likelihoods change as the number of mutations is increased? On the basis of this code, how many beneficial mutations would you infer there to be?

e) One case in which beneficial mutations might be observed is if the population is adapting in the presence of an antibiotic drug. Give three examples of mechanisms via which a mutation might convey antibiotic resistance.

f) The Optimist code fits a model to data collected via manual counts of the numbers of bacteria with each marker. How could errors made in the counting process affect the result of the inference? Suppose that data were instead collected using a flow cytometer to distinguish the two markers. Describe how you would you adapt the code to this purpose? You do not need to write the code for this question.

1.4 Inference of evolutionary parameters

Suppose you are given allele frequency measurements from an experiment performed on a yeast population of size $N \sim 10^7$. The population is derived by crossing two diverged ancestral strains for a few generations. This population is then exposed to an artificial selection pressure and propagated as a haploid population i.e. with no recombination taking place after the selection experiment starts. DNA sequence data from the pool is obtained at constant time intervals once every 24 hours at a depth of 100 reads per segregating site; Each segregating site has two alleles, arbitrarily labelled 0 and 1. Table 1 shows variant allele counts from one such segregating site.

Write an appropriate equation of motion for this system (use a continuous time model). Using the equation you have written implement a likelihood function for having observed the data under different evolutionary scenarios (characterised by parameters in your model). Learn maximum likelihood parameters of your model parameters and estimate the extent of uncertainty in each one. What can be learnt about the evolution of the system from which this observation was made?

Table 3: Data for the inference problem (Q. 1.4).

time	0	1	2	3	4	5	6
$n_i^1(t)$	28	39	57	72	78	81	82