GENE0005: Advanced Computational Biology

Assignment for "Population Genetics" lectures (weeks 3-4)

This work is assessed, and all parts of the assessment must be attempted. Submit your answers via Moodle **by 11:59pm on Wednesday, February 19**.

1. **F_{ST}, coalescent, heterozygosity.**

   There are five files on Moodle, each labeled `hapmap3_r2_b36_chr20_Y.haps` and containing Single-Nucleotide-Polymorphisms (SNPs) covering chromosome 20 from population `Y`. Each such file contains phased haplotypes from individuals sampled from population `Y` as part of Phase 3 of the HapMap project (http://www.hapmap.org). Here `Y = {CEU, JPT, YRI, TSI, PopX}`, reflecting individuals with ancestry related to western Europe (`CEU`), Japan (`JPT`), Nigeria (`YRI`), Tuscany (`TSI`) and another population (`PopX`).

   You can read in the data with the following in `R`:

   ```
   ceu = t(read.table(file.choose()))
   ```

   and navigating to the folder where you have saved the file "`hapmap3_r2_b36_chr20_CEU.haps`". After doing so, `ceu` will be formatted such that each column represents a SNP, and each row is a distinct haplotype, with every two consecutive rows representing the DNA from a single diploid individual. The two possible allele types at each SNP are coded as `{0,1}`. Read in the data for the other four populations in the same manner, saving each file's data to a different variable each time (e.g. `jpt, yri, tsi, popX`).

   Answer the following questions, in the form of a maximum 3 page report including figures. (I.e. the answers to **all parts** (a)-(e) of question one should span at most 3 pages in total.) Be sure to state the assumptions underlying your responses.

   (a) For each of the five populations, calculate the **median** heterozygosity (H) across all SNPs. Are results as you expected? Why or why not?

   (b) Between every *pairing* of the five populations, calculate the **median** F_{ST} across all SNPs. (NOTE: use `median(...,na.rm=TRUE)` to remove any `NA` values when calculating the median; otherwise `median(...)` might just return `NA`.) Are results as you expected? Why or why not?

   (c) Where do you think `PopX` comes from? Justify your answer.

(d) For each of the five populations, pull off the first five SNPs (columns) of the first two individuals (rows). Build an Ancestral Recombination Graph (ARG) for this 5-SNP region that is consistent with the data.

(e) Is the resulting ARG what you expected? Why or why not? In general for a sample of ten sequences (i.e. five diploid individuals), what is the expected time until each of the coalescent events? What is the expected time until all individuals coalesce?

2. **Selection and recombination.**

For this question, you will use the file `ldsel.R`, which includes an `R` function `ldsel` that simulates a population undergoing random mating with mutation (`mu`), recombination (`rho`) and selection (`s`) at two biallelic loci. Notice that while the `wf.R` models a haploid population, to study recombination we need to consider diploid individuals and so `nall` here refers to the total number of diploid individuals. The default is `nall = 2000` individuals, which corresponds to 4000 haplotypes. The user-input value `init` is a vector that gives the relative frequency of the four haplotypes {`AB,Ab,aB,ab`} at generation 1; the default is {`0.5,0,0,0.5`}. The A allelic type at the first locus is undergoing selection, with the parameter `s` controlling the increase in fitness for haplotypes carrying A. The parameter `rho` reflects the rate of recombination between the two loci.

The output is a vector with 6 elements that gives final results at the end of `ngen` generations (default `ngen` is 500 generations), with the first four elements giving the haplotype proportions for {AB,Ab,aB,ab} and the last two elements giving $|D'|$ and r^2 between the two loci. As long as `to.plot` in the `ldsel` function is set to "yes", there will also be two plots. Examples of these plots are in the lecture notes: one gives the haplotype frequency trajectories over time, as well as the frequency trajectories of the allele frequencies of A and B. The other plot gives the values of linkage disequilibrium measures r^2 and $|D'|$ between the two loci over this same time frame. Note that unlike with `wf.R`, you can only simulate one population at a time with `ldsel`, so that the plot can illustrate the frequencies of the different haplotypes. (*Hint:* You should first set `to.plot` equal to "no" before copy n' pasting `ldsel` into R to run, anytime you want to simulate lots of populations in e.g. a `for` loop.) For simplicity, throughout this question set the population size (`nall`) and the mutation rate (`mu`) to their default values (i.e. `nall=2000`, `mu=0`).

Answer all parts to the question below, again in the form of a maximum 3 page report including figures. (I.e. the answers to **all parts** (a)-(e) of question two should span at most 3 pages in total.) Be sure to state the assumptions underlying your responses.

(a) Run `ldsel` with default settings. What do you see? What is the median time to fixation of the A allele?

(b) Now add some recombination, keeping the same default `init` and `s`. How do patterns change? Now what is the median time to fixation of the A allele?

(c) Now explore selection, in particular a scenario where a new mutation (which will be the A allele) enters the population and immediately undergoes selection, as this new mutation is advantageous for the species' survival. To do so, run `ldsel` setting values so that the selected allele A has initial frequency 0.01 (i.e. a small value, thus mimicking a newly arisen mutation). First set `rho=0`. Varying `s`, what is the pattern over time? For one value of `s`, find the median

time to fixation of the A allele.

(d) Now increase `rho` and repeat (c). How do patterns change?

(e) What kind of selection is this?