PSET 5 - 6.047/6.878/HST.507 - Fall 2018

Due Friday, November 16 at 11:59pm (submit via Stellar)

Please submit a zip file of your solution file via Stellar. Make sure you review the Stellar announcement regarding the submission format.

1 Positive selection in the human genome

In this problem, we will analyze a region of human chromosome 5 to identify single nucleotide polymorphisms (SNPs) that exhibit characteristic signatures of recent positive selection in human populations.

(a) **Long haplotype** tests are one approach to detect regions of the genome that may be under selection.

(a) Briefly explain why long haplotypes are evidence for recent selection.

(b) Why is it difficult for haplotype-based methods to identify individual polymorphism(s) under selection, especially if the selection is very strong?

Cross-population extended haplotype homozygosity (XP-EHH) is a metric used to identify long haplotypes in one subpopulation versus another. The file `XPEHH.txt` contains XP-EHH scores for a series of SNPs in populations from Europe (CEU), Africa (YRI), and Asia (JPT+CHB).

(a) Plot the scores across chromosome 5 for for the three pairwise population comparisons. You can plot XP-EHH against the position of the SNPs either in terms of their DNA sequence positions (bp) or in terms of their recombinant frequencies (cM). Which is preferable for this purpose, and why?

(b) In which subpopulation do you see the strongest evidence for natural selection? Explain your answer.

(c) How many SNPs have XP-EHH scores above 2.0 in at least one pairwise comparison?

(b) **Derived allele frequency.** For a SNP, we distinguish between the *ancestral allele*, the allele present in the common ancestor, and the *derived allele*, arising from a recent mutation and possibly under selection. One way to study the strength of selection on a derived allele is to determine how much it has spread through the population.

To approach this for a given SNP, we need to know which of its alleles is the ancestral allele, and the derived allele frequency in the modern population.

(a) One way to infer the ancestral allele is to assume that it is the base observed in a closely related species – chimpanzee is often used for humans. However, this may sometimes lead to an erroneous conclusion as to which of the SNP's alleles is the ancestral allele. Explain why, and give a back-of-the-envelope estimate for how likely this is to occur for a given SNP, considering that the mean human-chimp sequence divergence is 1.23%.

(b) The file `Derived.txt` specifies how many copies of the derived allele were found among 120 European, 120 African, and 180 Asian chromosomes for each of the SNPs we are studying.

Calculate the derived allele frequencies in the population that you concluded is under selection in part (a), and plot them across chromosome 5.

(c) How many SNPs have both the long haplotype signal above and derived allele frequency above 0.6?

(c) **Population differentiation.** A third line of evidence for recent selection is given by highly differential allele frequencies between subpopulations. One way of measuring the degree of difference is to compare the *heterozygosity* within each subpopulation to the heterozygosity of the population as a whole. If p is the frequency of an allele in a population, then the expected heterozygosity is the frequency of heterozygotes in the population at Hardy-Weinberg equilibrium, $2p(1-p)$.

The statistic F_{ST} is defined as $\frac{H_T - H_S}{H_T}$, where H_T is the heterozygosity for the total population and H_S is the average heterozygosity of the subpopulations. Roughly speaking, F_{ST} tells us how much genetic differences between subpopulations, rather than genetic diversity within subpopulations, contribute to overall genetic diversity.

(a) Assume we have a population composed of two equally sized subpopulations. The overall allele frequency in the population is p, and the allele frequencies in each subpopulation are $p + d$ and $p - d$. Derive a simple expression for F_{ST} in terms of p and d.

(b) Based on the derived allele frequencies for the human subpopulations in part (b), calculate F_{ST} for the subpopulation under selection against each of the other subpopulations. How do you estimate p and d? For each SNP, average these two pairwise F_{ST} values and plot them across chromosome 5.

(c) How many SNPs pass the above thresholds and also have an average $F_{ST} > 0.6$?

(d) **Function.** Finally, to facilitate follow-up studies, we would like to restrict our investigation to SNPs that fall within known or likely functional elements in the genome.

(a) The file `phastConsElements.txt` gives the coordinates (in bp) of regions that are evolutionarily conserved in vertebrate species, and the file `genes.gff` gives the exons and introns of known protein-coding genes in the region of chromosome 5 we are studying. How many SNPs are within conserved elements? exons? introns?

(b) Based on these annotations, how many SNPs pass the above thresholds and also lie within known or likely functional elements?

(e) Based on all the evidence we've now collected, which SNP is the best candidate target of selection? If it lies within a gene, search the internet to find the function of the gene.

2 Maximum parsimony phylogeny

In this problem we will work through the algorithms to build maximum parsimony phylogenetic trees.

(a) There are three possible unrooted trees (acyclic undirected graphs) of four nodes. For each position in the multiple alignment below, give the cost of each tree assuming the cost of a *transition* ($A \leftrightarrow G$ or $C \leftrightarrow T$) is 1 and the cost of a transversion (any other mismatch) is 2.

Give the total cost of each tree over the entire alignment and indicate the minimum cost tree.

```
AACCGG
ACTCAG
GTCCTT
GGTTCG
```

(b) Given the distance matrix below, perform UPGMA clustering. Show all intermediate trees and distance matrices.

	a	b	c	d	e
a	0	3	11	10	12
b	3	0	12	11	13
c	11	12	0	9	11
d	10	11	9	0	8
e	12	13	11	8	0

Note the pairwise distances in the resulting tree do not match the given distances. Holding the tree topology constant, give modified branch lengths to recapitulate the given distance matrix.

Is it possible to construct a tree with this topology and the correct pairwise distances? If so, explain why and give an algorithm. If not, why not?

3 (6.878) Wright-Fisher process and the coalescent

In this problem, we will explore statistical models for population genetics.

(a) First, write a program to simulate the Wright-Fisher reproduction process. Your program should accept as input the population size and number of generations to simulate.

  (a) Recall that the basic Wright-Fisher model assumes clonal reproduction of haploid individuals and that the population size remains constant. In each generation, each possible parent produces an infinite number of progeny, and a new population is selected from these childen (so the sampling is with replacement).

  (b) Now, extend your simulator to track the coalescence times of lineages of the most recent individuals. If you are observing k individuals, you should report the $k-1$ generations where coalescence occurred. Run your simulator for 1000 times with a population size of $N = 500$. Report the average and standard deviation of the first coalescence times (across 1000 trials) of $k = 2$, 3, and 4 individuals. Explain how you selected the number of generations to simulate. How would your results change if you simulated too few or too many generations?

  (c) Do your results agree with the coalescent approximation? Justify your answer (agreement or disagreement) with a brief quantitative argument and explain why you think it does or does not agree.

(b) We will now extend the simulation to approximate sexual reproduction.

  (a) Adjust your simulation so that the originating ancestors have a known gender, assigned randomly. There will F females and $M = N - F$ males (and this ratio will be constant in all generations). We will implement a simplified model of sexual reproduction where each individual is haploid and selects its chromosome from one of its two parents (without recombination). In each generation, possible parental combinations are formed by pairing all female chromosomes with all male chromosomes (assume these are only autosomal chromosomes). These pairs are then split up into haploid choices for all possible individuals. The next generation is chosen by selecting from this set of (haploid) chromosomes. Of the N individuals in this next generation, F will be female and $M = N - F$ will be male. You might verify your simulation by testing with $F = M = 250$ and comparing your results to those from the previous section.

  (b) Assume that there are $F = 100$ females and $M = 400$ males. As in the previous section, perform 1000 simulations for $k = 2$ individuals. Report the average and standard deviation of the coalescence times. Do your results agree with the coalescent approximation? Again, provide a quantitative justification and a brief explanation.

  (c) **Extra credit:** If your results do not agree with the (standard) coalescent approximation, can you extend the coalescent approximation to incorporate this gender imbalance? You might approach an answer empirically by comparing simulation results for various values of N and M or F. Hint: an extension might incorporate the expression $\frac{4MF}{N^2}$.