Biological Data Analysis (CSE 182) : Assignment 6

Logistics

Except when explicitly permitted, all of your calculations must be done by hand using first principles, and not using R, Python, or published tables. Please try to type-set your answers or write cleanly. Points will be awarded for clarity.

Population Genetics

1. Define the following terms: allele, locus, heterozygote, phasing. What is the problem with using chromosomal coordinates to define a locus? How do people get around the problem?

2. The so called ‘bread wheat’ is hexaploid (6 copies of each chromosome). Consider a locus with 4 allelic values (A, G, C, T) with frequencies 0.5, 0.25, 0.15, 0.1, respectively. (a) Compute the number of distinct possible genotypes. (b) Compute the expected number of occurrences of the genotype defined by one occurrence of A, 3 of T, and 2 of C in a sample of 10,000 individuals, assuming HW equilibrium holds. (c) Generalize part (a) to compute the number of distinct genotypes given a ploidy of n (n copies of each chromosome) and m alleles.

3. In class we learned about the $\chi^2$-test in the context of linkage disequilibrium. Here, we will try and apply it to Hardy-Weinberg equilibrium. In sampling population 1, the genotype counts at a locus were given by AA:507, Aa:390, and aa:89. In population 2, the counts were given by AA:59, Aa:42, aa:4. Our goal is to figure out for each of these populations, if HW equilibrium is violated, or satisfied.

   (a) For each of the populations, use genotype frequencies to estimate the frequency of major allele A in the population.

   (b) Devise a $\chi^2$ statistic to test the Null hypothesis that the population is in HW disequilibrium. Output the $\chi^2$ statistic value for each of the two populations. Note that you must not use R or any table for this, but show all your calculations for computing the statistic.

   (c) Compute the $p$-values for each of the two $\chi^2$ values, and decide if HW is violated or satisfied. In order to do this, you need to understand the concept of degrees of freedom. The number of degrees of freedom corresponds to the number of independent values in the computation of the statistic. Here, we have 3 genotype frequency values that we are using to compute the $\chi^2$ statistic. What is the number of independent values? Use this information, and look up a table to compute $p$-values for the two populations, and then decide if HW is satisfied or violated.

4. Draw a perfect phylogeny using the SNP matrix $M$ or suggest that none is possible. All mutations must be marked on edges and labeled, and all leaves and internal nodes must have haplotype
sequence attached to them. You should not assume that 0 is the ancestral mutation, but you
can make the infinite sites assumption. You can do this assignment by writing a program, or by
hand.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

5. Show, using a derivation, how the $\rho$ value of Linkage Disequilibrium can be interpreted as a $\chi^2$ statistic.

6. Assume that when SNPs are created, the D’ value of LD between two equals 1 (Perfect LD). Also assume that the recombination rate is $10^{-8}$bp$^{-1}$gen$^{-1}$. The table below provides the average LD values of SNPs $\ell$ bp apart for different values of $\ell$ (1Kbp is 1,000bp). We know that one of the two populations separated from the other, older, population. Given the data, say which population is older and explain your calculations? How many generations ago did the newer population split away from the older one?

| $\ell$ | Pop. 1 $|D'|$ | Pop. 2 $|D'|$ |
|---|---|---|
| 1Kbp | 0.905 | 0.819 |
| 10Kbp | 0.368 | 0.135 |
| 20Kbp | 0.135 | 0.0183 |
| 50Kbp | 0.00673 | 4.54 $\cdot 10^{-5}$ |
| 100Kbp | 4.54 $\cdot 10^{-5}$ | |