Algorithms for genetics (CSE280a) Assignment 2

January 28, 2019

Questions

1. (20pts) For the data-set provided, compute LD for all pairs of columns, using the D' measure. (a) Plot the LD as a matrix as shown in the graphic (Figure 1). Use the LD value to identify Haplotype blocks (regions with little to no recombination). (b) Use lecture notes to convert the D' statistic into an appropriate P-value. Redo the matrix by replacing the D' value with $-\log(\text{P-value})$.

2. (30pts) (a) Design and implement a no-recombination, coalescent based backward simulator that outputs a sample (a haplotype SNP matrix with n rows) evolving neutrally but with exponential growth in the population. The input to your simulator is a set of parameters: sample size n, is θ, Current population N, and with growth rate parameter α. To explain the use of growth rate, let N_t denote the population size t generations prior to the current generation. Then, $N_0 = N$, and $N_t = e^{-\alpha} N_{t-1}$.

   (b) Simulate 100 samples for each of two cases: $\alpha = 0$, and a positive (small) value of α using the parameters provided and show qualitatively and quantitatively, how the allele frequency spectrum changes in the exponentially growing population relative to the constant population size scenario.

3. (35 pts) Consider a binary SNP matrix with m columns, n rows from a neutrally evolving region with no recombinations, and scaled mutation rate θ. Assume that all mutations are of the form $0 \to 1$. Let m_i denote the number of columns with exactly i ones. Show that

$$\text{Exp}(m_i) = \frac{\theta}{i}$$

   Hint: Recall that the coalescent can be partitioned into $n-1$ time epochs $T_2, \ldots, T_n$, such that there are exactly k lineages in the time epoch T_k. Let p_{kji} be the probability that lineage j in the k-th epoch has exactly i descendants. First show that

$$p_{kji} = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} = \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}} \cdot \frac{k-1}{i}$$

   then use this to prove the desired result.

4. (15 pts) Show that the following are all valid estimates of θ

   - $\theta_{FL} = m_1$
   - $\theta_W = h_n^{-1} \sum_{i=1}^{n-1} m_i$, where $h_n = \sum_{i=1}^{n-1} 1/i$.
   - $\theta_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i m_i$
   - $\theta_\pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i) m_i$.
   - $\theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 m_i$.

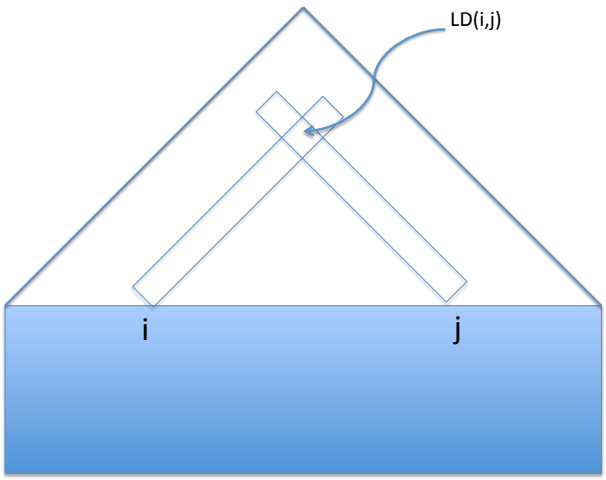

Figure 1: Pairwise LD values, color coded.