Problem 1: A Bit of Mathematica Programming

1. Program the equation to compute $r^2$.

   
   \[
   rsquared = (probs[[1]] - pi1plus*piplus1)^2 / (pi1plus*pi2plus*piplus1*piplus2);
   \]

2. Compute pairwise $r^2$ for the SNP Matrix.

   
   \[
   For[i = 1, i <= Length[hap1], i = i + 1,
   For[k = 1, k < i, k = k + 1,
   probs = ComputeProbs[Hap, k, i];
   pi1plus = probs[[1]] + probs[[2]];
   pi2plus = probs[[3]] + probs[[4]];
   piplus1 = probs[[1]] + probs[[3]];
   piplus2 = probs[[2]] + probs[[4]];
   rMatrix[[k, i]] = (probs[[1]] - pi1plus*piplus1)^2 / (pi1plus*pi2plus*piplus1*piplus2);
   ]
   ]
   \]

Problem 2: LD Interpretation

The goal of this problem is to interpret intermediate values of $D'$ and $r^2$ and to try to describe the relationship between the two quantities.

Answers.

$D'$ and $r^2$ are more correlated at the end points. We were looking for some type of thresholding for which they are more correlated. Any other types of observations about the behavior in different intervals are valuable. Validated conditions as we studied in class (such as $r^2 = 1$ or $D' = 1$) with empirical data is useful.

Other solutions provided by students of previous years:

Compliments of Bahar Erar

$D'$ and $r^2$ measure the amount of linkage disequilibrium from sort of different aspects. For a population where only two of the four possible haplotypes exists, the LD of the observed sample will be both complete ($D' = 1$) and perfect ($r^2 = 1$) since in this case the two SNPs explain the
variation in the other perfectly. However, for intermediate values, interpretations of the two measures change. $r^2$ has a wider range of possible values compared to $D'$. For example, $D' = 1$ if only three of the four possible haplotypes are observed regardless of the frequencies; however, $r^2$ can be anywhere in $[0, 1]$ depending on the frequencies. As an example, we can look at a population with haplotype frequencies $P_{AA} = 0.5$, $P_{AB} = 0$, $P_{BA} = 0.48$ and $P_{BB} = 0.02$. Generating a random sample of 10000 haplotypes from this population, we get $D' = 1$ and $r^2 = 0.018$. On the contrary, a sample from a population with $P_{AA} = 0.5$, $P_{AB} = 0$, $P_{BA} = 0.02$ and $P_{BB} = 0.48$, gives $D' = 1$ and $r^2 = 0.93$. Therefore $r^2$ explains the nature of the mutations that occurred while $D'$ is solely reflecting the amount of recombination. We can also look at how the two measures behave when all haplotypes occur with equal probability in the population. Just by running the sampling algorithm a few times, we can see that $r^2$ is much closer to 0 compared to $D'$. Furthermore, if we take smaller samples, say size of 25, both measures are inflated although $r^2$ was still closer to 0 while $D'$ varied greatly from sample to sample taking much higher values.

Compliments of Douglas McErlean

$D'$ is highest when we have complete LD, meaning at least one possible haplotype has frequency zero. Meanwhile it is lowest when all the haplotypes have the same frequency. In some sense, these situations are logical polar opposites, because in the former case there are two alleles which each have an infinite ratio between their two haplotype frequencies, whereas in the latter case the ratios between haplotype frequencies for every allele are all equal to one. This pattern seems to generalize fairly well to the intermediate case, as $D'$ climbs when as many of these ratios are as high as possible. This makes it particularly sensitive to low frequency alleles, as small fluctuations in one haplotype frequency can still change the ratio dramatically when both haplotypes are relatively rare.

Meanwhile, while $r^2$ reaches its minimum in the same case $D'$ does, its maximum occurs only when EVERY allele appears in exactly one haplotype, not just some of them. Thus, while a maximal $r^2$ implies a maximal $D'$, the converse is not necessarily true. In the intermediate case, $r^2$ seems to require a much more balanced set of ratios than does $D'$, in that if even one allele has two haplotype probabilities that are similar, $r^2$ will be very low. This makes sense, because such an allele would correlate only very weakly with its counterparts on the other locus, and $r^2$ is a correlation coefficient. As a result, $r^2$ is at its highest when a pair of haplotypes have similar probabilities if and only if they share no alleles, e.g. AB and ab can have similar probabilities, but AB and Ab cannot. This precludes $r^2$ from having a high value when an allele is rare, because bounding the frequency of allele A necessarily also bounds the probabilities of AB and Ab, and thus their difference.

Compliments of Rohan Palmer
Problem 3: Cases for LD

Construct examples (by hand or with help from the Mathematica notebook) where (a) \( r^2 = 1 \) and \( D' = 1 \), (b) \( r^2 < 1 \) and \( D' = 1 \), (c) \( r^2 < 1 \) and \( D' < 1 \).

**Answers.**

There are many answers for the following problems. Here is a sample:

(a) \( h_1 = 01 \ h_2 = 10 \)

(b) \( h_1 = 01 \ h_2 = 10 \ h_3 = 00 \)

(c) \( h_1 = 01 \ h_2 = 10 \ h_3 = 00 \ h_4 = 11 \)

Problem 4: Questions on assigned reading

1. What is the leading cause for erosion of Linkage Disequilibrium? Single point mutations and structural variation are rare events and “there is no evidence to indicate that mutation contributes significantly to the erosion of LD between SNPs.” Recombination erodes LD by combining maternal and paternal chromosomes in the formation of the gametes. The genetic material passed to offspring is a combination of two distinct haplotypes which mixes alleles and may create new haplotypes in the population.
2. What are complete LD, perfect LD, and useful LD? Complete LD is defined by $D’ = 1$ which happens when there are three or less haplotypes between two the SNPs. Perfect LD is defined by $r^2 = 1$ which implies that the information to infer a marker is completely encapsulated in the other marker. The answer to problem (3a) is an example of Perfect LD. Useful LD is related to the power required to detect an association which depend on the parameters of the GWAS (e.g. sample size). As a general rule, an $r^2 < \frac{1}{3}$ is not useful LD.

3. How do LD haplotype blocks form? Tightly packed sets of SNPs and SNPs in regions of the genome with low recombination rates do not have the opportunity to create many discrete contiguous blocks of SNPs through the process of recombination.

Student answer compliments of Brady Tang: Haplotype blocks tend to form in regions where recombination is relatively rare (as explained in 4.1, erosion of LD is generally a result of recombination). Proximal sites are also likely to exhibit greater LD in general, even if recombination rate is uniform, as they are less likely to be separated. Consequently, even in the uniform recombination rate case we will observe LD blocks.

**Problem 5: Clark phasing**

Clark method solution.

\[
\begin{array}{cccccc}
g_1 & 1 & 0 & 0 & 1 & 1 \\
g_2 & 1 & 2 & 0 & 2 & 1 \\
g_3 & 0 & 2 & 1 & 1 & 1 \\
g_4 & 1 & 2 & 1 & 0 & 2 \\
g_5 & 1 & 0 & 1 & 2 & 0 \\
g_6 & 0 & 1 & 0 & 0 & 0 \\
\end{array}
\]

The first and sixth row can be added to the resolved haplotype set becomes it contains all homozygous SNPs. We can also add the two complementary haplotypes that explain rows 3 and 5 because they contain only 1 heterozygous SNP. The resolved haplotype set is now

\[
\begin{array}{cccccc}
1 & 0 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
\end{array}
\]
we now attempt to resolve unresolved genotypes using resolved haplotypes. We can resolve
(1 2 0 2 1) with the haplotype in the first row and the complement haplotype (1 1 0 0 1). The resolved haplotype set is now

\[
\begin{array}{cccc}
1 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
\end{array}
\]

We can resolve (1 2 1 0 2) with the haplotype in the fifth row and the complement haplotype (1 1 1 0 1). The final resolved haplotype set is

\[
\begin{array}{cccc}
1 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 \\
\end{array}
\]