Problem Research Question – Linkage Disequilibrium vs. Informativeness (35%)

We present here a sample of answers supplemented with key insights from student’s solutions. Many different answers are considered “correct” as long as they correctly support or identify key aspects of each question.

2.1: The definition of LD blocks is ad-hoc and block boundaries are fluctuating between definitions. Also, many tag SNPs hit between blocks so in this sense it is difficult to use the LD concept to tag these SNPs. A fundamental difficulty with D-based LD measures is that they do not extend to multiple SNPs. Thus the following situation may be possible: a set of SNPs may individually have low LD with a SNP to be tagged, but collectively a high amount of LD. Informativeness measures can overcome these shortcomings but are more computationally expensive to compute; the dynamic programming algorithm takes exponential time in size of the window of SNPs, which make it difficult to calculate on dense arrays. Also, there would still have to be a formal definition of a block in the informativeness interpretation, which is not inherently clear.

2.2, 2.3: Examples of algorithms for LD: LD-Select, Tagger, HaploBlockFinder and other greedy heuristics based on SNPs that are in high LD with others. Examples of algorithms for MIS: The Halldorsson et. al. DP presented in class.[1]

2.4, 2.5: This was intended as a research questions to entice experimentations to search for patterns. A good idea here would be to calculate the informativeness of the SNPs that Tagger picks and vice versa for various input. Try to find correlations. For example, for SNPs that are in LD > 80%, what is the value of informativeness?

Courtesy of Fabien Wagner (2010)

The hope behind LD-based SNP tagging is that a set of SNPs that are in strong pairwise LD with the other SNPs will be able to predict the state of the entire set. It is equivalent to saying that a set of strong pairwise will lead to more informativeness. Although this is not entirely true, there are some cases in which it is. If two SNPs are in complete LD (D’=1), the informativeness (defined as the ability to predict one SNP from another) will be equal to 1. Similarly, if the SNPs are in HW equilibrium (D’=0), the informativeness will be null. One can also define a LD measure for a set of SNPs (d^2) by summing all the pairwise correlations. But as mentioned in question 1, even if all the pairwise correlations are weak, the informativeness can be high. So, in general a strong LD would tend to increase informativeness, but it is not true in all cases and can result in selecting a non-optimal set of SNPs.

2.6: The translation between informativeness and LD based tag SNP selection because they operate in different units and the algorithms are based on very different measures (statistical vs.
information theory). Although, and you did not have access to this article, there is a lemma stating \( r^2 = \text{directed informativeness}^2 \) where directed informativeness is the directed graph theoretic counterpart of the informativeness measure. We did not expect you to derive such proofs; empirical observations would be valid answers.

**Problem Haplotype Phasing by Expectation Maximization and Clark (55%)**

For full credit students had to give and support different methods for comparing haplotype phasings, including quantitative results comparing the algorithm solutions of Clark and EM to each other and the ground truth and describe how LD and haplotype diversity factor into algorithm correctness measurements. Using the block files along with the region file hinted that haplotype phasings across blocks are less reliable than within blocks. Using this intuition, the most used measure of correctness is the haplotype switch metric which counts the number of haplotype switches it takes to convert one haplotype to another. The definition is best understood by example.

Let individual 1’s inferred haplotypes be 000000 111111, and individual 1 true haplotypes be 000111 111000. Under the edit distance model, we would count 6 inconsistencies regardless of how we matched the haplotypes. But, a much more likely explanation is that we had two haplotype blocks and there was one haplotype switch error between the blocks.

1.b) We expected a large variety of answers. Examples: (1) Phase windows of the genotype separately and combine the windows. (2) Perform the EM method on a subset of the haplotypes, e.g., the set of haplotypes that are shared by at least 2 people. (3) A mixture of the Clark and EM algorithms.

**Problem Adopt a GWAS for the Midterm and Final Project (10%)**

Awarded if student selected a GWAS prior to the midterm.

**References**