Optimal haplotype block free selection of tagging SNPs for genome-wide association studies

Supplemental Material 2: Connecting Informativeness with measures of LD and Diversity

Zhang et al. (Zhang et al. 2002) also consider the problem of minimizing SNPs based on linkage disequilibrium with other SNPs. They limit their prediction to *blocks*, or regions of low recombination. They formulate the problem as one of predicting *common* haplotypes (haplotypes that occur at least twice), a measure clearly related to ours. Recall our formula for informativeness of set S' in predicting the set T

$$I(S', T) = \frac{|E(S') \cap E(T)|}{|E(T)|}$$

Let S be the set of SNPs in a haplotype block described by the matrix M_S after eliminating all haplotypes that occur exactly once. The rows of this matrix form the set of common haplotypes H. Observe that $E(S) = H \times H$, the complete graph on the vertex set H. Zhang et al. (Zhang et al. 2002) define an informative set of SNPs $S' \subseteq S$, such that

$$\exists V \subseteq S \text{ s.t. } E(S') = V \times V \text{ and } \frac{|V|}{|H|} = \beta > 0.8 \tag{1}$$

Note that if $\beta = 1$, the problem would be identical to picking $S' \subseteq S$ such that $E(S') = H \times H = E(S)$. Thus the two measures of informativeness of SNPs are expressed on the same graph, one based on vertices, and the other based on edges. One advantage of our method is that it can be used with respect to a single SNP, or a predetermined group of SNPs, which makes it comparable to other measures in population genetics.

In the population genetics community, various measures for linkage disequilibrium also measure the informativeness of a SNP with respect to another. A commonly employed measure of linkage disequilibrium between allele A_i and B_j (at loci A and B, respectively) is D_{ij} defined as

$$D_{ij} = p_{ij} - p_{i.}p_{.j} \tag{2}$$

Where $p_{i.}$ is the probability of seeing allele A_i at loci A, $p_{.j}$ is the probability of seeing allele B_j at loci B and p_{ij} is the probability of seeing allele A_i at loci A and allele B_j at loci B. Measures related to this are

$$\begin{aligned} D^2 &= \sum_{i,j} D_{ij}^2 && (3) \\ D' &= \begin{cases} D_{00}/\min(p_{0.}p_{.1}, p_{.0}p_{1.}) & D_{00} > 0 \\ D_{00}/\min(p_{0.}p_{.0}, p_{.1}p_{1.}) & D_{00} < 0 \end{cases} && (4) \end{aligned}$$

While these measures are useful in measuring linkage between SNPs, they are less useful in measuring informativeness, which is the extent to which one SNP can help predict the other. Observe for example that $|D'| = 1$, its maximum value whenever any of the four possible haplotypes is not present. Assuming the infinite-sites model of evolution, this implies that $|D'| = 1$ when there is no recombination between the two loci. This is not, however, enough for a SNP to predict the other. A more useful measure of LD that also describes informativeness is the d^2 measure (Devlin and Risch 1995; Kruglyak 1999).

Following Kruglyak (Kruglyak 1999) denote the variant allele in the target SNP t as v, and the normal allele as $+$. Then the d^2 measure for a SNP s containing alleles 0, and 1 with respect to t is estimated as

$$\left(\frac{n_{v0}}{n_v} - \frac{n_{+0}}{n_+}\right)^2 \tag{5}$$

This measure is 1 exactly when s can completely predict t, and 0 when s provides no information w.r.t. to t ($\frac{n_{v0}}{n_v} = \frac{n_{+0}}{n_+} = 0.5$). The d^2 measure does not, however, extend easily to the informativeness of a set of SNPs S' w.r.t. t. Let $|S'| = k$ so that 2^k haplotypes are possible. For each haplotype H, its informativeness w.r.t. t can be denoted by $\left(\frac{n_{vH}}{n_v} - \frac{n_{+H}}{n_+}\right)^2$. The d^2 measure for S' w.r.t. t is given by

$$\max_H \left(\frac{n_{vH}}{n_v} - \frac{n_{+H}}{n_+}\right)^2 \tag{6}$$

However, it is easy to see that this measure is actually quite restrictive. A small subset of SNPs S' could easily provide complete information about t even though the d^2 measure for any one pair of SNPs is quite low. The measure we propose is thus a better measure of informativeness, and also provides insight into linkage between two arbitrary sets of SNPs.

References

Devlin, B. and Risch, N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**: 311-322.

Kruglyak, L. 1999. Prospects for whole-genome linkage mapping of common disease genes. *Nature Genetics* **22**: 139-144.

Zhang, K., Deng, M., Chen, T., Waterman, M. S., and Sun, F. 2002. A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of*

Sciences **99**: 7335-7339.