Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci

David Clayton
Diabetes and Inflammation Laboratory
Cambridge Institute for Medical Research
Wellcome Trust/MRC Building
Addenbrooke's Hospital, Cambridge, CB2 2XY

July 24, 2001

1 What are htSNPs?

On typing a large number of SNPs within a small genomic region in European populations it is commonly found that there is rather little haplotype diversity — the observed haplotypes fall into rather few major groups with only minor differences between haplotypes within groups. Johnson *et al.* [1] suggested that linkage disequilibrium and haplotype diversity within the region can be captured by a much smaller subset of the markers, which they term "haplotype tagging SNPs" (htSNPs). This note describes one formal approach to the optimal choice of htSNPs.

2 Measuring haplotype diversity

Consider a haploid genetic system or assume that gametic phase of S linked polymorphic markers can be determined. Further assume that the loci are diallelic. Then each observation, $i = 1 \ldots N$, of a haplotype can be represented as a vector $z_i = \{z_{ij}, j = 1 \ldots S\}$ of alleles, which we will assume here are coded either 0 or 1.

Locus and haplotype diversity can be defined as the total number of differences recorded in all N^2 pair-wise comparisons between the observations [2] . For locus j, since the difference $(z_{ij} - z_{kj})$ is 0 if observations i and k are the same and ± 1 if they differ, the diversity can be written as

$$D_j = \sum_{i=1}^{N}\sum_{k=1}^{N}(z_{ij} - z_{kj})^2 = 2\left\{N\sum_{i=1}^{N} z_{ij}^2 - \left(\sum_{i=1}^{N} z_{ij}\right)^2\right\}.$$

This is $2N$ times the conventional total sum of squares in the analysis of variance. For the haplotype as a whole, the diversity is

$$D = \sum_{i=1}^{N}\sum_{k=1}^{N}(z_i - z_k)^{\mathrm{T}}(z_i - z_k) = 2\left\{N\sum_{i=1}^{N} z_i^{\mathrm{T}} z_i - \left(\sum_{i=1}^{N} z_i\right)^{\mathrm{T}}\left(\sum_{i=1}^{N} z_i\right)\right\}.$$

A candidate collection of H htSNPs classify the N observed haplotypes into $G =$, at most, 2^H groups. We may then define the *residual diversity* as the sum of the within–group diversities. Denoting the haplotype groups as $\mathcal{G}_\}, \} = \infty \ldots \mathcal{G}$, the residual diversitites for locus j and overall are

$$\begin{aligned} R_j &= \sum_{g=1}^{G} \left\{ \sum_{i \in \mathcal{G}_\}} \sum_{k \in \mathcal{G}_\}} (z_{ij} - z_{kj})^2 \right\}, \\ R &= \sum_{g=1}^{G} \left\{ \sum_{i \in \mathcal{G}_\}} \sum_{k \in \mathcal{G}_\}} (z_i - z_j)^{\mathrm{T}} (z_i - z_j) \right\} \end{aligned}$$

The former can be shown to be equal to $2N$ times the residual (or "within group") sum of sqaures in the analysis of variance. By analogy with the coefficient of determination (conventionally denoted by R^2) we can define the proportion of diversity "explained" by the set of htSNPs, at locus j and overall, by

$$\begin{aligned} P_j &= 1 - \frac{R_j}{D_j}, \\ P &= 1 - \frac{R}{D}. \end{aligned}$$

The index P_j also has an interpretation as a probability. Imagine drawing, at random, a haplotype from the population of haplotypes which have allele 0 at locus j and a second haplotype from those which have allele 1 at this locus. Then P_j is the probability that these haplotypes will differ in one or more of the htSNPs.

3 Computer programs for htSNP choice

The program `hapdiv`, written in the macro language of the statistical package *Stata*, carries out the above computations and lists, for any choice of htSNPs, the indices (D, R, P), together with the same quantities, (D_j, R_j, P_j) for every locus, j.

The companion program `htsubsets` searches all subsets of size up to a given maximum, recording the best subset according to two criteria:

1. maximum overall haplotype diversity explained, P, and
2. maximum of the minimum, over loci, of the corresponding locus-specific indices, i.e. $\min P_j$.

We would require well chosen htSNPs to score highly in both respects.

References

[1] GCL Johnson *et al.* (2001) Haplotype tagging for identification of common disease genes. *Nature Genetics* (submitted).

[2] L Escoffier (2001) Analysis of Population Subdivision. In *Handbook of Statistical Genetics*, eds Balding DJ, Bishop M, and Cannings C. Wiley: Chichester (2001).