# 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

January 10, 2002

## 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 $\pm 1$ 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 full haplotypes into $G =$, at most, $2^H$ groups defined by haplotypes of the htSNPs ("ht-haplotypes"). We may then define the *residual diversity* as the sum of the within–group diversities. Denoting the ht-haplotype groups as $\mathcal{G}_g, g = 1 \ldots G$, the residual diversitites for locus $j$ and overall are

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

The former can be shown to be equal to $2N$ times the residual (or "within group") sum of squares in the analysis of variance. By analogy with the coefficient of determination we can define the proportion of diversity "explained" (PDE) 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 locus–specific PDE, $P_j$, has an interpretation as a probability. Imagine drawing, at random, a haplotype from the population of full haplotypes which have allele 1 at locus $j$ and a second full haplotype from those which have allele 2 at this locus. Then $P_j$ is the probability that these haplotypes will differ in one or more of the htSNPs *i.e.* that they fall into different ht-haplotype groups. This measures the extent to which knowledge of the ht-haplotypes carried by a subject predicts the alleles carried at the further locus, $j$. If $P_j = 1$, then locus $j$ is perfectly predicted by the ht-haplotypes.

The overall PDE, $P$, is a weighted average of the locus–specific PDEs, with weights $p_j q_j$, where $p_j$ and $q_j$ are the frequencies of alleles 1 and 2 respectively of locus $j$. Thus, an SNP with equally frequent alleles will have the most weight in the total PDE, and loci with one common allele and one rare allele will receive less weight.

Even if the loci were in linkage equilibrium, selection of a set of htSNPs will result in reduction of residual diversity. By analogy with Cohen's kappa statistic, "chance corrected" versions of $P_j$, $P$ can be proposed. If $N_g$ represents the total number of haplotypes in the $g$-th group defined by the htSNPs, the expected percentage of haplotype diversity explained by chance is

$$C = \frac{N}{N-1} \left( 1 - \frac{T}{N^2} \right),$$

where $T$ is the total number of pair-wise comparisons within groups:

$$T = \sum_{g=1}^{G} N_g^2.$$

The chance–corrected version of $P$ is then

$$P^* = \frac{P - C}{1 - C},$$

which takes the value zero (or less) when the explanation of haplotype diversity does not exceed its chance value, $C$. A similar expression holds for the locus–specific index.

Although useful for assessing the strength of association (disequilibrium) between a set of htSNPs and those remaining, there is not a strong case for using this index to choose between different candidate subsets. An index which has a stronger claim in this respect is the *mean residual diversity*, $\overline{R} = R/T$.

# 3 Weighted analysis

Frequently haplotypes are not directly observed, and their frequencies must be inferred from phase-unknown genotype data by statistical methods.[1] We will then have a list of unique haplotypes each with a "weight", $w_i$, proportional to the probability or frequency of the haplotype in the population. The indices described above then become

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

and we replace $N$ and $N_g$ by overall and within-group sums of weights respectively. When weights represent true frequencies there is no need for any further modification. Otherwise we must omit the $N/(N-1)$ factor in the computation of the chance explanation, $C$.

# 4 Haplotype frequencies from allele frequencies

A further desirable property for the choice of a set htSNPs is their ability to predict the distribution of haplotypes. If only $1+H$ different haplotypes are observed, then their frequencies may be predictable using linear contrasts of allele frequencies of only $H$ diallelic loci *if these are correctly chosen*, although in general there will not be a unique choice. The importance of this observation lies in the ability to make use of haplotype tagging ideas when directly estimating allele frequencies by typing DNA pools rather than by typing individual subjects.

The ability to infer haplotype frequencies from allele frequencies is concerned with the solvability of linear simultaneous equations and is easily stated mathematically. We define an $(H+1) \times (H+1)$ matrix, $X$, with rows representing the observed haplotypes, first column containing any constant, and with remaining columns containing the alleles for the subset of loci considered. This subset can be said to *linearly identify* the haplotypes if the matrix $X$ is of full rank. This concept is illustrated by the following example in which there are five loci but only five different haplotypes: Since there are only five haplotypes, it is necessary only to type four loci in order to linearly identify identify the haplotypes. The set of loci $\{A, B, C, D\}$ successfully identify the haplotypes in the sense described here, while the set $\{B, C, D, E\}$ does not. This is shown by application of the rule; recoding alleles as 0/1 for simplicity, the $5 \times 5$ matrices for these two sets of loci are,

[1] The program `snphap`, distributed from this site, is one such program.

|  | Locus |  |  |  |  |
|---|---|---|---|---|---|
| Haplotype | $A$ | $B$ | $C$ | $D$ | $E$ |
| 1 | 2 | 1 | 2 | 1 | 2 |
| 2 | 1 | 2 | 1 | 2 | 2 |
| 3 | 1 | 1 | 2 | 1 | 2 |
| 4 | 1 | 1 | 1 | 2 | 2 |
| 5 | 1 | 1 | 1 | 1 | 1 |

respectively,

$$X_{\{A,B,C,D\}} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \text{and} \qquad X_{\{B,C,D,E\}} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

$X_{\{A,B,C,D\}}$ is of full rank which, rather loosely, implies that all the columns (loci) contain new information. However, $X_{\{B,C,D,E\}}$ is not of full rank, since the last column is simply the sum of the two preceding columns.

# Computer programs for htSNP choice

The program `hapdiv`, written in the macro language of the statistical package *Stata*, carries out the computations of haplotype diversity described here. It lists, for any choice of htSNPs, the indices $(\overline{D}, \overline{R}, P)$, together with the same quantities, $(\overline{D}_j, \overline{R}_j, P_j)$ for every locus, $j$.

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

1. minimum residual mean overall diversity, $\overline{R}$,

2. minimum of the maximum, over loci, of the corresponding locus-specific indices, i.e. $\max \overline{R}_j$.

3. maximum percentage of overall diversity explained, $P$, and

4. maximum of the minimum, over loci, of the corresponding locus-specific indices, i.e. $\min P_j$.

Optionally the current version of the program only considers those subsets which, for subset size $H$, linearly identify the $H+1$ most common haplotypes.

A further program `haplist` lists haplotypes in groups defined by a choice of htSNPs, identifying alleles which differ from the most common haplotype of the group.

Note that these programs are designed to aid the process of selecting htSNPs rather than to automate it. Any "optimal" subset choice found by `htsubsets` must be carefully scrutinized using `hapdiv` and `haplist` to check that there is adequate capture of information.

The size of subsets which can be searched is limited by the computer time one is prepared to expend. Currently the default maximum subset size is five. This does not imply that five htSNPs will always be sufficient to capture the haplotype diversity in a region — it often will not. This default has been chosen simply to limit computation time. If a substantially larger number of htSNPs is necessary, exhaustive subset search may not be feasible and other methods may be needed. One possibility is to do a preliminary analysis of pairwise linkage disequilibrium measures in order to suggest subdivision of the region into two or more smaller regions within which exhaustive subset search is feasible. The additional program `pwld` computes various pairwise linkage disequilibrium measures and permits a compact display of the pattern of disequilibrium. We are working to develop more formal methods to aid this process.

# References

[1] GCL Johnson *et al.* (2001) Haplotype tagging for identification of common disease genes. *Nature Genetics*, **28**, Oct 2001, pp 1–9.

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