

Haplotype Tagging Single Nucleotide Polymorphisms and Association Studies

Deborah Thompson^{a,b} Dan Stram^c David Goldgar^a John S. Witte^{a,d}

^aUnit of Genetic Cancer Epidemiology, International Agency for Cancer Research, Lyon, France; ^bCR-UK Genetic Epidemiology Unit, University of Cambridge, UK; ^cDepartment of Preventive Medicine, University of Southern California, Los Angeles, Calif., and ^dDepartment of Epidemiology and Biostatistics, University of California, San Francisco, Calif., USA

Key Words

Association · Haplotypes · Single nucleotide polymorphisms

Abstract

Objectives: Discrete blocks of low haplotype diversity exist within the human genome. The non-redundant subset of 'haplotype tagging' single nucleotide polymorphisms (htSNPs) in such blocks can distinguish a majority of the haplotypes. Several approaches have been proposed to determine htSNPs, ranging from visual inspection to formal analytic procedures. Optimal htSNPs can be estimated using a small subgroup of an association study population that have been genotyped for a dense SNP map, and it is just these htSNPs that are genotyped in the remainder of the samples. We investigated by simulation how the size of the subsample affects the power of association studies, and what type of subjects it should include. **Methods:** We used the program tagSNPs [Stram et al., Hum Hered 2003;55:27–36], which selects htSNPs to minimize the uncertainty in predicting common haplotypes for individuals with unphased genotype data. **Results:** On average, 27% of the SNPs were designated as htSNPs. Genotyping as few as 25 unphased individuals to select the htSNPs did not appear to reduce

the power of an association study, as compared with using all SNPs. For the disease models considered, selecting htSNPs based on cases, controls, or a mixture of both gave similar results. **Conclusions:** These results suggest that the genotyping effort in an association study can be substantially reduced with little loss of power by identifying htSNPs in a small subsample of individuals.

Copyright © 2003 S. Karger AG, Basel

Introduction

The increasing availability of single nucleotide polymorphisms (SNPs) throughout the human genome has led to their widespread use in association studies attempting to decipher the genetic basis of common diseases. Such studies often evaluate the association between individual SNPs and disease. More information, however, may be gained from combining SNPs into haplotypes [1–3].

Recent work suggests that the human genome contains discrete chromosomal regions with low haplotype diversity, termed 'haplotype blocks,' which are separated by recombination hotspots [e.g. 4–8]. This has prompted substantial efforts toward the development of a haplotype map of the human genome [9].

By definition, SNPs within haplotype blocks may be in strong linkage disequilibrium. Therefore, information from some SNPs within each block may be redundant; in other words, having information on one SNP provides all the information about another. The majority of the haplotypes within a block can thus be distinguished using a much smaller number of SNPs, known as ‘haplotype tagging’ SNPs, or htSNPs. Using such SNPs can drastically reduce the effort required to undertake large scale association studies. Instead of saturating an entire chromosomal region with genotypes in all study samples, an investigator can first screen for SNPs within a sub-sample of study subjects to determine the htSNPs. Then only these tagging SNPs (and possibly other promising SNPs) can be genotyped in the entire study population. The use and potential value of tagging SNPs in association studies is outlined in figure 1.

Several approaches have been suggested for identifying optimal htSNPs. These include simple visual inspection of haplotypes and analytic approaches. Johnson et al. [2] defined the optimal htSNPs as the subset that best explain the residual haplotype diversity that exists within groups of haplotypes that are classified as being the same by these htSNPs. Alternatively, Zhang et al. [10] used a dynamic programming algorithm to simultaneously divide the region into haplotype blocks and to select the tagging SNPs within each block, in such a way as to minimize the total number of htSNPs. Meng et al. [11] have suggested a method based on the spectral decomposition of the matrix of all pairwise LD coefficients between markers, and adopt a sliding-window approach to allow its application to a large set of markers. The htSNPs are chosen as those that make the largest contributions to the eigenvectors with the highest eigenvalues. Alternatively, one can also determine genotype tagging SNPs as those that give the best average prediction of the remaining variants (D. Clayton, personal communication).

Previous work has shown that the use of htSNPs can substantially reduce genotyping efforts [2, 10]. However, the number of individuals who should be included in the initial subsample for determining the htSNPs remains unclear. Some have suggested that only 20–40 phased haplotypes are needed to distinguish the common htSNPs [10]. Others have used 70 or more cases and controls of a specific ethnic group to find the htSNPs for use in an association study [12] (A. Loukola, personal communication). We provided here a more full investigation of this, using a simulation study to evaluate the power of association studies when using different size subsamples for finding haplotype tagging SNPs.

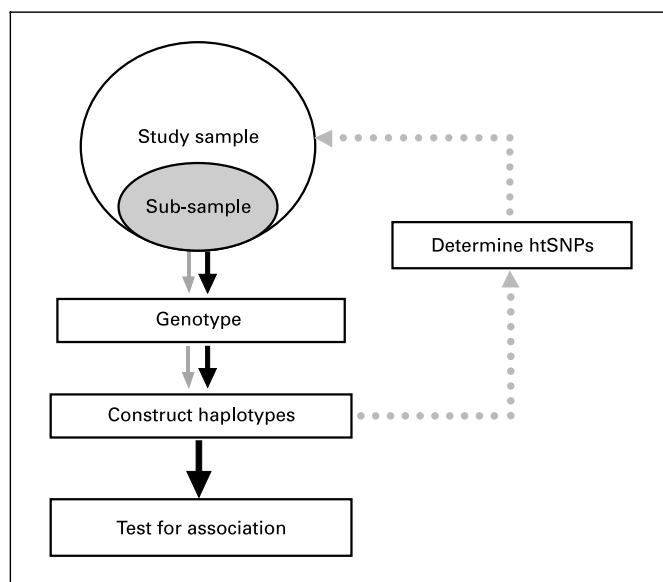


Fig. 1. Flow diagram to illustrate the design of a two-phase study based on haplotype tagging SNPs.

Methods

Our simulation study used the neutral coalescent model to generate chromosomal segments. In particular, for each replicate, a set of 10,000 chromosome segments were created by the program ‘ms’ [13]. This program uses a standard approach for generating a random genealogy and randomly placing mutations onto it, according to an infinite-sites model. Cross-over events are modeled using a finite-sites recombination model.

The parameters used in the simulation study were: an assumed effective diploid population size of 10,000 (e.g. per the International SNP Map Working Group [14]); a recombination probability between adjacent bases of 10^{-8} (i.e. 1 cM/Mb); and a neutral mutation rate of 3.8×10^{-4} (the mean diversity over 50 candidate genes, as reported by Turet et al. [15]). The scenario we aimed to model was that of a candidate gene study, rather than a genome wide association study, and so the length of the generated chromosome segments were fixed to be 29 kb, which was the average gene length, including untranslated exonic sequences [reported in 15].

The number of variants on a chromosome segment generated in this way varied between replicates, and was typically around 150. We first selected one variant to represent the disease-associated allele. This was required to be close to the center of the gene (between 40 and 60% of the genetic distance), for greater homogeneity among replicates. Disease alleles were required to be between 5 and 15% in frequency (p); if no such variant existed, then that replicate was rejected. Many of the generated variants did not adequately represent the kind of SNPs that are commonly used in association studies, either because they were too rare, or unrealistically close to one another. Therefore we next selected marker SNPs from among the remaining variants, as if these were the SNPs that had actually been typed in the samples. Unless otherwise stated, marker SNPs were

required not to be the disease variant itself; rather we were considering the situation where information from typed SNPs is used to map a nearby untyped causative variant. We considered only SNPs with a minor allele frequency of $\geq 5\%$, and required the selected marker SNPs to be spaced a minimum of 0.15 kb from each other (0.5% of the total length). In practice they were usually much more widely spaced, and there were no restrictions placed on their distance from the disease variant.

At this stage, the 10,000 simulated chromosome segments each consisted of a series of suitable marker SNPs and a disease allele that was treated as unobserved (wildtype allele *d*, deleterious allele *D*). These haplotypes were randomly paired, with replacement, to form 1000 diploid subjects. Adopting the model of a case-control study with equal numbers of cases and controls (500 of each), cases were oversampled from this population. A multiplicative codominant disease model was assumed, with relative risks (RR) β and β^2 for carriers of 1 or 2 copies of the disease allele respectively, in comparison to the risk in wildtype homozygotes. Constraining the overall incidence to be equal to the population prevalence of disease, κ , the phenocopy rate is given by $c = \kappa / (1 + p(\beta - 1))^2$. The two haplotypes carried by a case or a control were selected from among the sets of *D*-haplotypes (those carrying the disease allele) and *d*-haplotypes (those carrying the wildtype allele) with the usual probabilities:

$$\begin{aligned} P(dd | \text{case}) &= c(1 - p)^2 / \kappa & P(dd | \text{ctrl}) &= (1 - c)(1 - p)^2 / (1 - \kappa) \\ P(dD | \text{case}) &= 2c\beta p(1 - p) / \kappa & P(dD | \text{ctrl}) &= 2(1 - c\beta)p(1 - p) / (1 - \kappa) \\ P(DD | \text{case}) &= c\beta^2 p^2 / \kappa & P(DD | \text{ctrl}) &= (1 - c\beta^2)p^2 / (1 - \kappa) \end{aligned}$$

Pairs of alleles at each SNP were treated as if the phase were unknown, to mimic the anticipated situation in a population-based association study.

The analysis was first conducted using all individuals and all of the available 'typed' SNPs. Haplotype frequencies were estimated using the 'partition-ligation' E-M algorithm [16], as implemented in the tagSNPs program [12]. The same program predicts the haplotypes carried by each individual, estimating the probabilities of different haplotype assignments in ambiguous individuals (i.e., those heterozygous for more than one SNP). Briefly, for individual *i* with observed unphased multilocus genotypes G_i and unobserved true pair of haplotypes H_i , the haplotype dosage, $\delta_h(H_i)$ is defined as the number of copies of haplotype *h* in the pair H_i . The predicted haplotype dosage given the observed genotypes is computed for each individual, $E[\delta_h(H_i) | G_i]$. For a given multilocus genotype G_i this expectation, computed under an assumption of Hardy-Weinberg equilibrium, is a function only of the haplotype frequency estimates obtained from the E-M algorithm. For subjects where the haplotype pair carried by an individual is ambiguous given G_i the expectation will depend upon the estimated haplotype frequencies and will not in general take an integer value.

These predicted haplotype dosages were used as the independent variables in logistic regression [17], implemented using the software 'R' (version 1.6.2). The most common haplotype was taken as the baseline group, and all 'rare' haplotypes (defined as those with an estimated frequency $< 5\%$) were pooled to make a single category. The risk of carrying two copies of a haplotype was assumed to be the square of the risk of carrying a single copy, relative to a homozygous carrier of the most common haplotype. Power was determined from the test of the overall association between haplotypes and disease; that is, the difference between the residual deviance and the null

deviance, compared with the χ^2 distribution with degrees of freedom equal to the number of 'common' haplotypes (varies between replicates).

Following the diagram in figure 1, a random subsample of subjects was taken to represent those who would be typed for all SNPs, and hence used to identify the optimal htSNPs. The key aim of this simulation study was to investigate the effect of varying the size of this subsample. Unless otherwise stated, this group was sampled without regard to phenotype; we also considered restricting the subsample to cases or to controls.

Within this subsample of subjects, the partition-ligation E-M algorithm was used to estimate haplotype frequencies with the full set of SNPs. To quantify the predictive value of any given potential set of htSNPs a formal calculation of the expected squared correlation, R_h^2 , between estimated and true haplotype dosage (i.e., $E[\delta_h(H_i) | G_i]$ vs. $\delta_h(H_i)$) was computed assuming that only the htSNPs would be observed in G_i [12]. Under Hardy-Weinberg equilibrium, this squared correlation is solely a function of the haplotype frequency estimates. The computations were performed using the tagSNPs program. In order to compare potential candidate sets of htSNPs we used the statistic $\min R_h^2$, which is the minimum of the R_h^2 s over all common haplotypes (here, all those with $\geq 5\%$ frequency). For each *k* between 1 and the number of SNPs, the set of *k* SNPs that maximizes the $\min R_h^2$ was reported. The optimal set of htSNPs was chosen to be the optimal set of size *k*, where *k* is the smallest number of SNPs that will give a ratio $\max(\min R_h^2(k)) : \min R_h^2(\text{all}) \geq 0.9$; that is, the fewest SNPs that are at least 90% as useful for predicting haplotypes in individuals as the complete set of SNPs.

Once a set of optimal htSNPs was decided upon using the preliminary subsample, the frequencies of the haplotypes made up only of these htSNPs were estimated using genotype data from all individuals in the full case-control study, and used to compute the htSNP haplotype dosage predictions $E[\delta_h(H_i) | G_i]$ for all individuals. The predictions were used in a logistic regression as described above. For each replicate, the same simulated dataset was analyzed twice; once using the complete set of SNPs, and once using only the htSNPs. In each case, the empirical power was given by the proportion of replicates for which a significant association was reported. The significance of the difference in power between the two analyses was estimated using McNemars test for matched data (i.e., matched by simulation replicate).

We considered different relative risks, and looked at the effect of varying the size and phenotypic status of the subgroup used for htSNP detection. One thousand replicates were simulated under each combination of parameters. The majority of the analyses were based on testing the overall haplotype-disease association; as an alternative, within each of a separate set of replicates we noted the haplotype(s) that carried the disease allele, and then looked specifically for a statistically significant positive association between at least one of these haplotypes and disease. Of course, where the disease allele does not lie on any of the common haplotypes, the analysis will automatically fail to identify such an association. Simulations were performed for relative risks of 1.5, 1.75 and 2.0. We considered subsamples of sizes 500 (not shown), 200, 100, 50, 25 and 10, from a complete sample of 500 cases and 500 controls.

Results

Across 15,000 simulations, the median frequency of the disease allele was 8.6% (95% coverage 5.1–14.6%) and the median SNP frequency was 21.5% (95% coverage 13.6–30.4%). The median number of common haplotypes (frequency $\geq 5\%$) based on all SNPs was 7 (95% coverage 4–10). The median number of eligible SNPs per replicate was 34 (95% coverage 19–48), equivalent to approximately one SNP every 0.85 kb.

Using a subsample of 50, 100, or 200 individuals, the median number of htSNPs was 9 (95% coverage 6–12), and the set of htSNPs comprised between 26–27% of the possible SNPs (table 1). For subsamples of 10 or 25, slightly fewer htSNPs were estimated. The median number of common haplotypes based on just the htSNPs was 8, or 7 for a subsample of just 10 individuals (table 1).

Recall that our aim was to compare the power when using haplotypes consisting of all SNPs with the power using haplotypes constructed from the htSNPs. Naturally, the power to detect an overall haplotype-disease association based on all SNPs was higher for disease models with a higher RR (table 2). The relative power using htSNPs, as compared with using all SNPs, is shown in table 2. The empirical false positive rate (nominally 5%) was similar using all SNPs and using htSNPs (5.5% and 4.9% respectively, using a subsample of size 50, $p = 0.4$ for the difference). Using htSNPs does not appear to affect the power, except when a very small number of individuals are used to select them. It is not apparent that this pattern is dependent on the RR, at least not for RRs within the range tested. For a RR of 2.25, a tagging subsample of 10 people gave a power ratio of 0.96 ($p < 0.001$), relative to a 91.7% power using all SNPs.

The majority of analyses were based on the assumption of a relatively common disease-associated allele. We also considered a rarer variant (fixed to be between 1 and 5%

frequency; median frequency = 2.0%, 95% coverage 1.0–4.8%). The power using all SNPs was vastly reduced e.g. 18.1% for RR = 2.0. The relative power using htSNPs was more variable than for a common disease allele, but there was no trend with the number in the subsample, and the relative power was not significantly different from 1 (for a RR of 2.0: subsample = 50, relative power = 0.92, $p = 0.1$; subsample = 25, relative power = 0.93, $p = 0.3$; subsample = 10, relative power = 0.99, $p = 0.9$).

Forcing the disease variant to be among the marker SNPs (i.e., the scenario where it is, unwittingly, among the set of htSNPs typed in the full study) increased the power very slightly (67.5% for RR = 1.75), but gave similar ratios between the power using all SNPs and the power using just the htSNPs (e.g. for a RR of 1.75: subsample = 50, relative power = 1.03, $p = 0.05$; subsample = 25, relative power = 1.00, $p = 0.9$; subsample = 10, relative power = 0.95, $p = 0.03$).

Table 1. Number of estimated htSNPs for different sized tagging subsamples¹

Size of subsample used to find htSNPs	Estimated htSNPs	SNPs used as htSNPs, %	Common htSNP haplotypes ²
200	9 (6–12)	26.7 (16.0–50.0)	8 (6–11)
100	9 (6–12)	26.3 (16.3–47.8)	8 (6–11)
50	9 (6–12)	26.5 (16.3–47.6)	8 (5–11)
25	8 (5–11)	23.3 (13.3–43.5)	8 (5–10)
10	7 (3–9)	20.0 (8.9–36.4)	7 (4–9)

¹ Values quoted are the medians and 95% coverage, based on 3,000 replicates.

² 'Common haplotypes' are defined here as those with frequency $\geq 5\%$.

Table 2. Relative power of logistic regression overall test of association, comparing using htSNPs with using all SNPs. Mix of cases and controls (1,000 replicates)

RR	Power using ALL SNPs	Number in subsample; relative power (p value) ¹				
		200	100	50	25	10
1.50	35.7%	1.04 (0.26)	1.01 (0.7)	1.05 (0.2)	1.01 (0.7)	0.96 (0.3)
1.75	65.8%	0.98 (0.20)	1.00 (0.9)	1.01 (0.5)	0.99 (0.6)	0.93 (0.001)
2.0	82.5%	1.02 (0.08)	1.02 (0.06)	1.02 (0.04)	0.99 (0.7)	0.92 (<0.001)

¹ p value for the difference between power using htSNPs versus using all SNPs.

Table 3. Power of logistic regression overall test of association – cases or controls (1,000 replicates)¹

RR	Power using all SNPs	Subsample; relative power (p value)			
		50 controls	50 cases	10 controls	10 cases
1.75	65.3%	1.02 (0.2)	0.98 (0.2)	0.93 (0.002)	0.93 (0.003)
2.0	84.6%	1.00 (0.7)	0.98 (0.04)	0.98 (0.2)	0.92 (<0.001)

¹ p value for the difference between power using htSNPs versus using all SNPs.

Table 4. Power of logistic regression test of association with specific disease haplotypes (RR = 2.0, 1,000 replicates)

Number in subsample	Power using all SNPs	Subsample; relative power (p value) ¹		
		mix	cases	controls
50	76.3%	0.98 (0.27)	0.94 (<0.001)	0.93 (<0.001)
25		0.94 (0.001)	0.90 (<0.001)	0.86 (<0.001)
10		0.79 (<0.001)	0.77 (<0.001)	0.77 (<0.001)

¹ p value for the difference between power using htSNPs versus using all SNPs.

The analyses reported here used a random mixture of cases and controls to identify the optimal htSNPs. We also restricted the tagging subsample to either cases or controls, but there were no consistent differences (table 3). For the disease models considered, this is not surprising. For example, assuming Hardy-Weinberg equilibrium and using standard expressions for genotype probabilities conditional on phenotype, for $\kappa = 0.05$, $p = 0.1$ and $\beta = 2.0$, among 10 cases we would expect to see 4 disease alleles and 16 normal alleles. Among 10 controls, we would expect 2 disease alleles, and in 10 subjects randomly sampled from cases and controls, 3 disease alleles. In other words, the differences are minor.

As an alternative to the overall test of association, we also analyzed the data from an additional set of simulations by rejecting the null hypothesis only when a significant positive association was seen between disease and one or more of the haplotypes that carry the disease allele (table 4). In this case, the power using all SNPs was markedly lower than for the overall association test (for RR = 2.0, 76.3% power, as compared with 82.5%). Reducing the size of the subsample reduced the power more rapidly than when testing for an overall association. The reduction was more evident when the subsample was restricted to either cases or to controls rather than using a mixture of both. Using just controls gave a slightly lower power than using just cases, but the difference was not large.

Discussion

We found that when looking at common haplotypes and a common disease, using small numbers of individuals to distinguish haplotype tagging SNPs has little effect on power. Specifically, the power to detect moderate associations is essentially unchanged unless one reduces the subsample to 10 individuals. Since so few individuals are required to determine htSNPs, the potential benefits of using this approach – instead of fully genotyping all individuals within a study – are substantial.

For example, using a subsample of 25 individuals to obtain the htSNPs, and assuming that there are one quarter as many htSNPs as SNPs, would allow an approximately $4N/(75 + N)$ fold reduction in genotyping effort, where N is the total sample size. If n_s is the total number of SNPs, $n_s N$ genotypes would be necessary if all SNPs were used, whereas $(25n_s + 0.25n_s(N - 25))$ genotypes would be required to identify the htSNPs and then type them in the whole sample. The proportional reduction in time and expense will clearly increase with the size of the study.

In another simulation study, Meng et al. [11] reported that 50–100 individuals should be used to estimate the htSNPs. Their criteria for ‘dropping’ possible htSNPs were more stringent than those used here, with the consequence that the set of htSNPs typically contained around 60% of all the possible SNPs (i.e., a smaller reduction in

genotyping effort). They considered the consistency across replicates of the probability that each SNP was included among the htSNPs, rather than assessing the power to detect an association.

The numbers of htSNPs we estimated from tagging subsamples of sizes 50, 100 or 200 were almost identical, but fewer htSNPs were selected when there were just 25 or 10 individuals in the subsample. A smaller number of samples would be expected to contain fewer distinct haplotypes, and hence fewer htSNPs would be required to distinguish them. Correspondingly, the number of common haplotypes constructed from the htSNPs was lower for the smaller subsamples, which may explain why more true disease-haplotype associations were missed in these situations.

We found that the median number of common haplotypes was slightly higher when they were constructed from the htSNPs as opposed to all the SNPs. When using a subset of the possible SNPs, haplotypes that differ only at SNPs not in that subset will be thought of as identical. This leads to fewer distinct haplotypes being observed, and each with greater frequency. Hence, one would expect to often see a greater number of common haplotypes when considering only the htSNPs.

In a few of our simulation studies the power from using all SNPs was non-significantly lower than that from using the htSNPs. Using htSNPs is essentially equivalent to grouping haplotypes that differ only at non-htSNPs. A causal variant located on a haplotype below 5% in frequency will be unlikely to be detected, but grouping haplotypes by htSNPs may increase the frequency of its htSNP haplotype 'group' to above 5%, and hence the association will be more likely to be observed. Alternatively, a causal variant may lie on several distinct haplotypes, but if these are all considered as being the same htSNP haplotype, the chances of detecting a significant association will be increased.

The simulations were based on the 'common disease, common variant' hypothesis, as opposed to the model of a study seeking to locate a rare, high-risk mutation, or several rare mutations within the same gene. That is, the optimal htSNPs are defined as those that best differentiate 'common' haplotypes (i.e., those with a frequency above some specified cut-off), with the expectation that these cause a common disease under study. Such htSNPs, however, may not well define rarer haplotypes. Therefore, if one expected the variant(s) of interest to lie on a rarer haplotype(s), the htSNPs could be selected to define less common haplotypes (e.g., a threshold of 1% instead of 5%). This would require larger tagging subsamples to

ensure that they are all correctly detected, and would increase the number of estimated htSNPs, ultimately reducing the savings achieved by using a haplotype tagging strategy. Moreover, detecting an association with such less common haplotypes will increase the number of haplotypes in the analysis, and require a larger overall sample size.

In contrast, pooling all haplotypes below a 5% frequency in the logistic regression means that a rarer variant would either lie on one of the pooled haplotypes, or on some fraction of one of the common haplotypes, and thus would be unlikely to be detected unless it conferred a substantial risk. However, the relevant frequency is that in the tested sample, not in the general population, and in a case-control study the sample will be enriched for a truly causal variant. As anticipated, the power to detect a causal variant of around 2% frequency was much lower than for a more common variant, although using htSNPs did not significantly alter this.

The tagSNPs program used here reconstructs haplotypes under the assumption that the SNPs lie within a region of restricted haplotype diversity. We have examined the pattern of pairwise LD between the simulated SNPs for a few sample replicates, using the GOLD software [18]. For some replicates there was strong LD between the majority of pairs of SNPs, but in other cases there was evidence that the SNPs fell into two, or occasionally more, distinct blocks. This would have the effect of increasing the number of possible haplotypes and the number of SNPs required to define them. Nevertheless, since the comparison of the power using all SNPs with the power using htSNPs was based on analyzing each replicate using both methods, this should not affect the estimates of the relative power.

In our simulation studies we assumed that all individuals came from the same population. In practice, association studies may often include individuals from different ethnic groups. Allele frequencies can vary substantially between populations, as can patterns of LD and haplotype block structure [e.g. 3, 19–21]. Although haplotype blocks whose boundaries are defined by genuine recombination hotspots are likely to be consistent across populations, if the observed block structure results from factors relating to population history such as selection, bottlenecks, or population admixture then the pattern of haplotype blocks would be expected to vary between different populations [22, 23]. One must be certain that the htSNPs are distinguished within the population(s) being studied; this may lead to a larger total subsample being required to determine them. Otherwise, one may not be able to ade-

quately re-construct the haplotype-level information from the htSNPs for all individuals.

Moreover, the development of the HapMap suggests that one may be able to use outside information to determine htSNPs [9]. Again, one must be careful to make certain that such information adequately reflects the population one is studying. This may be most problematic when the disease-associated haplotype is either rare or absent in the control set, but more frequent among cases. In this situation, the htSNPs being used may not be able to distinguish the disease-associated haplotype from other, more common haplotypes, masking its effect on disease.

Furthermore, the manner in which htSNPs are selected may introduce bias into an association study. For example, if one uses a subset of cases to determine htSNPs, and if some of the selected htSNPs are rare, then including these same individuals in the association analysis may bias the estimates of association. This is because we would be selecting SNPs observed in cases only, and hence would expect to see a difference in comparison with controls that merely reflects this selection process, not causality.

While using htSNPs instead of full haplotypes can substantially decrease the genotyping effort, it does not necessarily reduce the ultimate number of comparisons. Although it may be tempting to simply look at the association between each htSNP and disease, in some situations such individual-level comparisons may be inefficient and may fail to detect a real association. Nevertheless, instead of focusing on tagging SNPs that distinguish haplotypes, one could consider those that best predict all other genotypes (D. Clayton, personal communication). By looking at genotype tagging SNPs one avoids the issue of estimating haplotypes when phase is unknown (i.e., in non family-based studies). One might anticipate that searching for haplotype or genotype tagging SNPs leads to selection of the same tagging variants.

In the case of a single causative variant that is among the htSNPs typed in the whole sample, it is possible that the variant would be observed on multiple disparate haplotypes, and hence looking at genotypes may be more efficient. However, it may well be the case that the disease-causing variant was not among the marker SNPs, or was not selected as a htSNP. It seems likely that such a variant would be in stronger LD with the haplotypes formed by the set of neighboring htSNPs than with any individual SNP. Hence, one should instead analyze the haplotypes reconstructed from the htSNPs, which will be of a similar number as would be observed if entire haplotypes were genotyped. This also removes the need to

explicitly consider the interaction terms between the htSNPs, and analysis at the haplotype level will clearly be favored if multiple variants in cis position influence the disease phenotype.

Naturally, in addition to the SNPs formally identified as being the most useful for tagging haplotypes, one should also consider typing any SNPs that have the potential to be particularly interesting, including SNPs: (1) previously reported as being potentially associated with disease; (2) at sites known to be conserved between species; (3) in promising genomic regions, or (4) with known functional relevance. Zhang et al. have recently extended their dynamic programming algorithm for block partition and htSNP selection to allow investigators to attach different weights to SNPs depending on features such as whether they lie in coding or non-coding regions [24].

The tagSNPs program imputes information for missing genotypes, but care should be taken if data from a particular SNP are missing in many subjects in the tagging subsample. If having incomplete data alters the estimated haplotype frequencies, it can potentially affect the choice of the htSNPs, hence as far as is possible it would be advisable to ensure that all individuals in the subsample have complete genotype data.

Although our results suggest that as few as 25 individuals should suffice for obtaining suitable htSNPs, one must keep in mind that this is based on results from a simulation study. A large number of the simulated variants may have been equally informative, whereby many different choices of htSNPs would give comparable results. This is unlikely to always be the case; the next step will be to use 'real' data from a range of genes to investigate whether in practice such a small tagging subsample would be adequate to reconstruct the haplotypes of interest.

Acknowledgments

This work was supported in part by NIH grants CA88164 and CA94211. J.S.W.'s work reported in this paper was undertaken during the tenure of a Visiting Scientist Award by the International Agency for Research on Cancer. D.T.'s work reported in this paper was undertaken during the tenure of a Postdoctoral Fellowship from the International Agency for Research on Cancer.

References

- 1 Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ: Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res* 2001;11:143–151.
- 2 Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA: Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001;29:233–237.
- 3 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001;68:978–989.
- 4 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001;29:229–232.
- 5 Jeffreys AJ, Kauppi L, Neumann R: Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 2001;29:217–222.
- 6 Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001;294:1719–1723.
- 7 Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lohmussaer E, Zernant J, Tonissson N, Remm M, Magi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I: A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 2002;418:544–548.
- 8 Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: The structure of haplotype blocks in the human genome. *Science* 2002;296:2225–2229.
- 9 <http://www.genome.gov/10005336>: International Consortium Launches Genetic Variation Mapping Project. NIH News Advisory 2003. Ref Type: Internet Communication
- 10 Zhang K, Calabrese P, Nordborg M, Sun F: Haplotype block structure and its applications to association studies: Power and study designs. *Am J Hum Genet* 2002;71:1386–1394.
- 11 Meng Z, Zaykin VD, Xu C-F, Wagner M, Ehm MG: Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet* 2003;73:115–130.
- 12 Stram DO, Haiman CA, Kolonel LN, Altshuler D, Henderson BE, Hirschhorn JN, Pike MC: Choosing haplotype-tagging SNPs based on unphased genotype data from a preliminary sample of unrelated subjects, the Multiethnic Cohort Study. *Hum Hered* 2003;55:27–36.
- 13 Hudson RR: Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 2002;18:337–338.
- 14 The International SNP Map Working Group: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;409:928–933.
- 15 Tiret L, Poirier O, Nicaud V, Barbaux S, Herrmann SM, Perret C, Raoux S, Francomme C, Lebard G, Tregouet D, Cambien F: Heterogeneity of linkage disequilibrium in human genes has implications for association studies of common diseases. *Hum Mol Genet* 2002;11:419–429.
- 16 Qin ZS, Niu T, Liu JS: Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet* 2002;71:1242–1247.
- 17 Breslow NE, Day NE: *Statistical Methods in Cancer Research. Volume 1 – The Analysis of Case-Control Studies*. Lyon, IARC, 1980.
- 18 Abecasis GR, Cookson WO: GOLD – graphical overview of linkage disequilibrium. *Bioinformatics* 2000;16:182–183.
- 19 Goddard KA, Hopkins PJ, Hall JM, Witte JS: Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet* 2000;66:216–234.
- 20 Zhu X, Yan D, Cooper RS, Luke A, Ikeda MA, Chang YP, Weder A, Chakravarti A: Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program. *Genome Res* 2003;13:173–181.
- 21 Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA: Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* 2003;33:518–521.
- 22 Wang N, Akey JM, Zhang K, Chakraborty R, Jin L: Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. *Am J Hum Genet* 2002;71:1227–1234.
- 23 Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Stubbaker JF, Ankener WM, Alfisi SV, Kuo FS, Camisa AL, Pazorov V, Scott KE, Carey BJ, Faith J, Katari G, Bhatti HA, Cyr JM, Derohannessian V, Elosua C, Forman AM, Grecco NM, Hock CR, Kuebler JM, Lathrop JA, Mockler MA, Nachtman EP, Restine SL, Varde SA, Hozza MJ, Gelfand CA, Broxholme J, Abecasis GR, Boyce-Jacino MT, Cardon LR: Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet* 2003;33:382–387.
- 24 Zhang K, Sun F, Waterman S, Chen T: Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *Am J Hum Genet* 2003;73:63–73.