Biostatistics (2009), **10**, 4, *pp.* 680–693
doi:10.1093/biostatistics/kxp023
Advance Access publication on July 15, 2009

SHARE: an adaptive algorithm to select the most informative set of SNPs for candidate genetic association

JAMES Y. DAI*, MICHAEL LEBLANC

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, M2-C200, Seattle, WA 98109, USA
jdai@fhcrc.org

NICHOLAS L. SMITH, BRUCE PSATY

Department of Epidemiology, Cardiovascular Health Research Unit, University of Washington, Seattle, WA 98195, USA

CHARLES KOOPERBERG

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, M2-C200, Seattle, WA 98109, USA

SUMMARY

Association studies have been widely used to identify genetic liability variants for complex diseases. While scanning the chromosomal region 1 single nucleotide polymorphism (SNP) at a time may not fully explore linkage disequilibrium, haplotype analyses tend to require a fairly large number of parameters, thus potentially losing power. Clustering algorithms, such as the cladistic approach, have been proposed to reduce the dimensionality, yet they have important limitations. We propose a SNP-Haplotype Adaptive REgression (SHARE) algorithm that seeks the most informative set of SNPs for genetic association in a targeted candidate region by growing and shrinking haplotypes with 1 more or less SNP in a stepwise fashion, and comparing prediction errors of different models via cross-validation. Depending on the evolutionary history of the disease mutations and the markers, this set may contain a single SNP or several SNPs that lay a foundation for haplotype analyses. Haplotype phase ambiguity is effectively accounted for by treating haplotype reconstruction as a part of the learning procedure. Simulations and a data application show that our method has improved power over existing methodologies and that the results are informative in the search for disease-causal loci.

Keywords: Adaptive regression; Haplotype; Multilocus analysis; SNP.

*To whom correspondence should be addressed.

1. Introduction

Owing to the availability of high-throughput genotyping technologies and the comprehensive coverage of common genetic variants by the HapMap project (The International Hapmap Consortium, 2005, 2007), association studies are widely used to dissect the genetic basis of complex diseases in a scope varying from a number of candidate genes to the whole genome. A typical association study involves initial prioritization of single nucleotide polymorphism (SNP) genotypes in a small subsample or a selection of tagSNPs derived from an existing database such as the HapMap project. These tagSNPs are subsequently genotyped for a sample of cases and controls (Smith *and others*, 2007). While the causal variants may not be interrogated directly, it is hoped that linkage disequilibrium (LD) mapping could narrow the search down to a small neighborhood around the causal variants. However, despite the explosion of genetic information available, challenges remain for statistical analyses due to the diversity of LD patterns in the human genome (The International Hapmap Consortium, 2005), the sheer number of SNPs being genotyped, and the complex nature of common disorders. Currently, the single-SNP scan and multiple-SNPs haplotype analyses are 2 commonly used approaches. The power comparison between these 2 approaches is somewhat inconclusive, as it depends on underlying disease models and local LD patterns (Morris and Kaplan, 2002; Roeder *and others*, 2005). It has been suggested that a single-SNP scan is an effective method to detect common disease alleles, while haplotype-based methods are useful to map more recent, relatively rare mutations (Lin *and others*, 2004; Schaid, 2004), though strategies to construct informative haplotypes (clusters) are far from mature. This paper pertains to adaptive SNP/haplotype analysis exploiting LD among SNPs in a candidate chromosomal region.

When many SNPs in a targeted chromosomal region are under investigation, a naive haplotype analysis using all SNPs is often ineffective due to the large number of haplotypes and hence too many degrees of freedom in an omnibus test. Instead, one may first dividing SNPs into haplotype blocks of high LD and then performing a haplotype analysis in each block (Barrett *and others*, 2005). However, the block definition itself is arbitrary, and typically, there is substantial correlation not captured between blocks. An alternative strategy is to construct a genealogical tree of haplotypes, known as a cladogram, and study the correlation between the disease phenotype and the clusters (clades) of haplotypes, thereby reducing the dimensionality of haplotype analyses (Templeton *and others*, 1987; Seltman *and others*, 2001; Molitor *and others*, 2003; Durrant *and others*, 2004; Morris, 2006). The motivation is that the causal allele should be embedded within the cladogram that describes the evolution of the sampled chromosomes. However, an accurate construction of the underlying cladogram typically relies on the assumption that there is no recombination. This is hardly true for any given region because of background recombination in the human genome, particularly for regions near or within recombination hot spots. To this end, a sliding window approach was proposed in the hierarchical clustering algorithm called CLADHC (Durrant *and others*, 2004), yet the optimal window size cannot be universal due to the diversity of local LD through the human genome. Even in an extreme scenario with complete LD, it was pointed out that cladistic approaches cannot be optimal in all disease models (Clayton *and others*, 2004) since the rule of clustering haplotypes is based solely on genotypic data.

Other strategies for multilocus analyses exist (e.g. Browning, 2006; Yu and Schaid, 2007; Li *and others*, 2007). These methods generally assume that local LD structures are somewhat contiguous, thus the order of SNP locations is critical. It is possible that SNPs that are separated apart can display strong LD, so a contiguous scan might miss signals. Similarly, multiple nonsynonymous mutations in a gene may disrupt the function of its coded protein jointly, possibly with interactions, regardless of their order in the chromosome. Furthermore, all aforementioned methods do not account for extra variability incurred by phase ambiguity in the model searching process, except the computationally intensive MCMC approach (Morris, 2006).

In this article, we propose SNP-Haplotype Adaptive REgression (SHARE), an adaptive algorithm that searches for a subset of SNPs, which fully capture genetic association in a candidate chromosomal region.

The selected set of SNPs is the most informative in a heuristic sense: adding more SNPs introduces noise and excluding any SNP in the set may lose information. Contrary to the cladistic approaches, where the clustering process depends solely on haplotypes, in our algorithm, both the trait and the genotypes guide the model selection process, and the SNP selection is irrespective of the order of the SNPs. Depending on the genealogy and the ancestral recombination among disease liability mutations and markers, the most informative set may contain a single SNP or several SNPs that lay a foundation for haplotype analyses, thereby effectively integrating a single-locus scan and a haplotype analyses into 1 unified framework. Furthermore, our algorithm stands apart from existing methods in that it accommodates phase ambiguity seamlessly by treating the inference of haplotypes as part of the procedure. The method is tailored to genetic association studies with a fair number of tagSNPs genotyped in a candidate gene approach, but, as we address in the Section 4, it can be extended to genome-wide association studies.

2. METHODS

2.1 *Rationale*

We use an example to introduce the main idea: there generally exists a subset of SNPs that are sufficient to capture genetic association. Figure 1 shows the genealogical tree of 5 genptyped SNPs, labeled as **ABCDE**, and the unscored disease susceptibility SNP **X**. The genotyped SNPs can be tagSNPs that preserve maximal LD information with minimum redundancy. The haplotypes based on all 6 SNPs are displayed as strings of 0s and 1s, labeled numerically. Depending on where the susceptibility SNP arise, different subsets of SNPs are required to differentiate haplotypes that do and do not carry disease risk. In Figure 1(a), **X** occurs before **A** in lineage, thus only 1 SNP (**A**) is sufficient to capture the disease risk. In Figure 1(b), the functional variant **X** descended from **2**, generating a new haplotype that parallels **5** and **6** in lineage. We recognize that, instead of including all haplotypes based on 5 SNPs in an analysis, if we restrict the haplotype analysis to **A**, **D**, and **E**, haplotype `100` carries an increased disease risk, while all other haplotypes do not. In this case, a cladistic approach will collapse **2, 5**, and **6** and therefore dilute the disease signal. In the presence of recombination, the adjacent SNPs could have different genealogies. Thus, the genealogy of a sample of haplotypes is usually a graph with loops rather than a tree. Figure 1(c) depicts a situation where there is recombination between **6** and **3**, occurring between the fourth and the fifth locus. A new haplotype recombinant **8** is created. The functional variant **X** later arose in **6**.

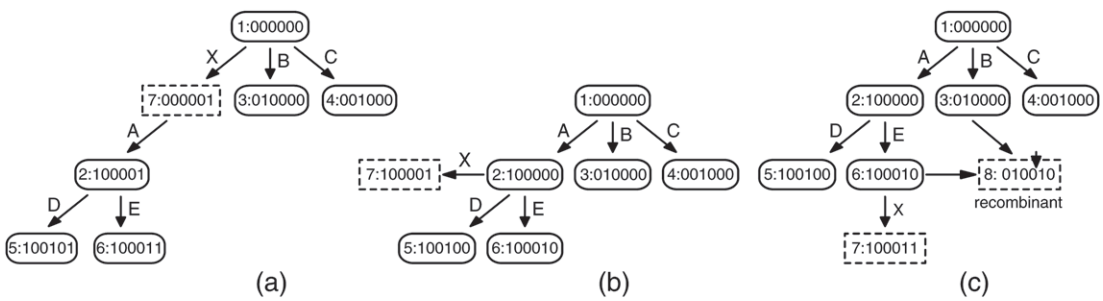

Fig. 1. An example to show that there generally exists an optimal set of SNPs for association analysis. The order of SNPs in a haplotype is **ABCDE(X)**. (a) The disease-causal locus **X** occurs before **A** in lineage. The optimal set for genetic association is just **A**. (b) The disease-causal locus **X** occurs after **A** and in parallel to **D, E**. The optimal set for genetic association is **A, D, E**. (c) The disease-causal locus **X** occurs after **E** in lineage. There is recombination between haplotypes 6 and 3, generating a recombinant **8**. The vertical arrow on the top of haplotype 8 points to the break point of recombination. The optimal set for genetic association contains **A, E** or **B, E**.

An inspection of SNPs before and after the breaking point suggests that either **A** and **E** or **B** and **E** will be adequate to discern the normal and risk-carrying haplotypes. For example, if we use SNP **A** and **E** to construct haplotypes, the haplotype 11 carries increased disease risk, while the other 3 haplotypes do not. This example sheds light on the effect of ancestral recombination on association mapping: it weakens the LD between the functional variant and the "proxy" in its lineage; in consequence, haplotypes across the breaking point become useful in mapping the functional variant. This is in the same spirit of the previous results that long haplotypes cross the recombination breaking point can help to map recent rare mutations (Lin *and others*, 2004). Note that the SNPs selected in Figure 1 are those before and after the functional variant in evolution, thus forming an evolutionary pocket surrounding the disease variant.

To find the most informative set, ideally, we would search all possible subsets of the available SNPs using, for example, the generalized Akaike information criterion,

$$\mathrm{AIC}_a = -2\ell + ap, \tag{2.1}$$

where ℓ is the likelihood for a model, a is a penalty parameter, and p is the number of parameters in the model. The best penalty parameter can be chosen by cross-validation. In reality, however, searching in all possible subsets quickly becomes infeasible as the number of SNPs gets larger than 20. We instead propose a stepwise algorithm to identify the most informative set. That is, we sequentially select the current best set by adding/deleting 1 SNP at a time to the previous best set, therefore substantially simplifying search paths. While stepwise algorithms have limitations, in the genetic context where LD structure is present in adjacent SNPs, stepwise selection is a natural choice, as opposed to more elaborative searching. The rationale is that the fundamental unit of inheritance—the haplotype is formed by sequential (stepwise) mutations during the history. Recombination shuffles around the haplotypes at breaking points (like hot spots), but the majority of genomic regions should be highly structured. With the nearly complete coverage of whole-genome common variations by the HapMap project, it is hard to imagine that an underlying disease loci does not exhibit any extra marginal association at all.

Different from a stepwise logistic regression treating SNPs as covariates (Cordell and Clayton, 2002), our algorithm iteratively constructs haplotypes based on the SNPs in current set. If we consider the sample space a population of haplotypes, our algorithm resembles recursive partitioning (Classification and regression tree [CART]; Breiman *and others*, 1984). We use the example in Figure 1(b) to illustrate this point. In Figure 1(b), a 3-SNP haplotype is best to capture the disease risk. One potential search path shown in Figure 2 is that we first find SNP **A** as the most significant SNP by a single-locus scan, next detect haplotypes constructed by **A** and **D** as the best 2-SNP haplotypes, finally, we reach the most informative set {**A, D, E**} so that a 3-SNP haplotype concentrates the disease risk. Note that adding 1 SNP actually partitions the sample space of haplotypes. For any particular haplotype, it can be sent down the tree just as an observation is being sent down in CART. While CART is effective to dissect high-order interactions, growing haplotypes is essentially refining high-order interactions between loci, as a haplotype effect is a linear combination of locus main effects and high-order interactions (Schaid, 2004).

2.2 *Notation*

For ease of exposition, we consider a sample of n unrelated affected cases and unaffected controls. Continuous traits can be accommodated using the generalized linear model framework. Let $\mathbf{Y}_i = 1$ if the ith individual is a case and $\mathbf{Y}_i = 0$ otherwise, $i = 1, 2, \ldots, n$. Let $\mathbf{G}_i = (g_{i1}, g_{i2}, \ldots, g_{ik}, \ldots, g_{iK})$ be the SNP genotypes of individual i at K loci on some chromosomal region of interest, coded as 0, 1, 2 for the number of the minor alleles at the kth locus. These SNPs could be tagSNPs selected to represent genetic polymorphisms in the targeted region and some of them may be missing for some individuals. Suppose that in addition to the genetic data, we also have information on r covariates $\mathbf{Z}_i = (z_{i1}, z_{i2}, \ldots, z_{ir})$, containing demographic and environmental factors.

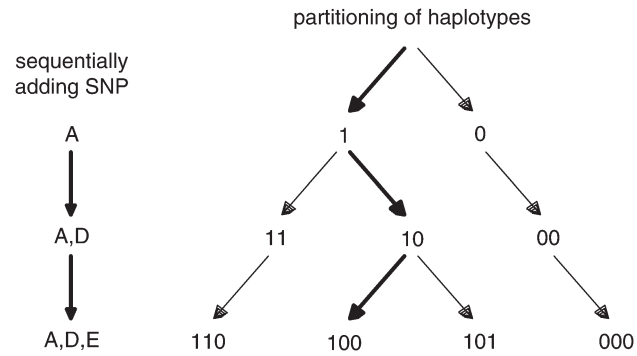

Fig. 2. The tree illustration of the sequential partition of haplotypes in Figure 1(b). The left panel shows the growing set of SNPs used in analysis and the right panel shows the partitions resulted from the haplotypes based on the current set of SNPs. The minimal set of SNPs that captures the genetic association is (A, D, E), with the disease risk concentrated on the haplotype `100`. The path leading to discovering it could be `1` $\rightarrow$ `10` $\rightarrow$ `100`. The corresponding order of SNPs in the haplotypes is $\mathbf{A} \rightarrow \mathbf{AD} \rightarrow \mathbf{ADE}$.

Let Ω_K denote the complete set of all K SNPs, and let Ω_l denote the most informative set of l SNPs that adequately captures the genetic association. By definition, $\Omega_l \subseteq \Omega_K$ and $0 \leqslant l \leqslant K$. When $l = 0$, Ω_l is an empty set, and there is no genetic association in the chromosomal region. Let $\mathbf{G}^{\Omega_k}$ be the observed genetic data on a set Ω_k. Assume that in the population, there are m distinct haplotypes $h_1^{\Omega_k}, h_2^{\Omega_k}, \ldots, h_m^{\Omega_k}$ based on SNPs in Ω_k, with (unknown) population frequencies $\mathbf{p}^{\Omega_k} = (p_1^l, p_2^l, \ldots, p_m^l)$. If Ω_k contains a single SNP, the haplotype is simply the genotype of the (single) locus. Hereafter, we generalize the definition of "haplotype" to include single-SNP genotypes. For the ith individual, let $\mathbf{H}_i^{\Omega_k} = \{\mathbf{H}_{i1}^{\Omega_k}, \mathbf{H}_{i2}^{\Omega_k}\}$ be the haplotype pair based on Ω_k. We assume that the underlying probabilistic model describing the association, denoted as $\mathcal{M}(\mathbf{H}_i^{\Omega_k}, \mathbf{Z}_i)$, is

$$\text{logit}[\Pr(\mathbf{Y}_i = 1 | \mathbf{G}_i^{\Omega_K}, \mathbf{Z}_i)] = \alpha + \beta[f(\mathbf{H}_i^{\Omega_k})] + \gamma \mathbf{Z}_i, \tag{2.2}$$

where $f(\mathbf{H}_i^{\Omega_k})$ is a function that delineates the haplotype effect model, α is the intercept, and β and γ are regression parameters for genetic and environmental effects, respectively. For instance, in an additive model, $f(\mathbf{H}_i^{\Omega_k})$ represents a vector of m integers in {0,1,2} indicating the number for each of m possible haplotypes. In the dominant models, having 1 or 2 copies of a haplotype has the same effect. In the recessive model, only having 2 copies of the causal haplotype will affect the trait. Gene–environment interactions can be added to (2.2). For the observed data, we can compute the maximal likelihood estimators of parameters in (2.2) and obtain $\widehat{\mathcal{M}}(\mathbf{H}_i^{\Omega_k}, \mathbf{Z}_i)$. Note that we have genotypes for all SNPs (Ω_K); however, we only select a subset to be used in the regression model (Ω_k). The best subset Ω_l with its associated model $\mathcal{M}(\mathbf{H}_i^{\Omega_l}, \mathbf{Z}_i)$ is selected by minimizing the prediction error. Let $\hat{p}_i = \widehat{\Pr}(\mathbf{Y} = y_i)$ be the probability of accurately predicting y_i based on $\widehat{\mathcal{M}}(\mathbf{H}_i^{\Omega_k}, \mathbf{Z}_i)$ when a new independent subject comes in. We define a loss function, namely deviance or cross-entropy (Hastie *and others*, 2001), $\mathcal{D}(y_i, \hat{p}_i) = -[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)]$, that quantifies the correctness of $\hat{p}_i$. Our goal was to minimize the expected loss (or the expected prediction error) over Ω_k to find the best set Ω_l. This can be expressed as

$$\Omega_l = \text{argmin}_{\Omega_k} E\{\mathcal{D}(y_i, \hat{p}_i | \widehat{\mathcal{M}}(\mathbf{H}_i^{\Omega_k}, \mathbf{Z}_i))\}. \tag{2.3}$$

2.3 *The algorithm*

If we search for the most informative set by 10-fold cross-validation estimates of the above objective function, the algorithm is as follows:

1. For $i = 1$ to 10,
   - In the ith training set, grow a sequence of nested sets $\Omega_0 \subset \Omega_{i1} \subset \Omega_{i2} \cdots \subset \Omega_{iM}$, where M is the largest number of SNPs in a candidate subset, specified by the investigator. Here Ω_0 indicates a model without any genetic effect.
   - In the ith training set, prune Ω_{iM} back 1 SNP at a time to obtain a sequence of nested sets $\Omega_{iM} \supset \Omega_{iM-1}{}' \supset \Omega_{iM-2}{}' \cdots \supset \Omega_0$.
   - In the ith training set, evaluate the prediction deviance for the models associated with $\Omega_0 \subset \Omega_{i1} \subset \Omega_{i2} \cdots \subset \Omega_{iM} \supset \Omega_{iM-1}{}' \supset \Omega_{iM-2}{}' \cdots \supset \Omega_1{}'$.
2. Sum up the prediction deviances from the 10 cross-validations for each model path, choose the number of SNPs, l, with the smallest prediction deviance.
3. Use all data to search for the model formed by l SNPs. If l was achieved in the growing stage, grow the subset up to l SNPs; If l was achieved in the pruning stage, grow the subset up to M SNPs and prune back to l SNPs.

Let m^{Ω_t} be the number of haplotypes given the current best subset Ω_t with t SNPs. Starting from the best single SNP, we select the next best subset containing $t+1$ SNPs, Ω_{t+1}, with the maximal statistic ϕ, defined as

$$\phi_{t+1} = \begin{cases} \dfrac{\mathcal{D}^{\Omega_t} - \mathcal{D}^{\Omega_{t+1}}}{m^{\Omega_{t+1}} - m^{\Omega_t}} & \text{if } m^{\Omega_{t+1}} \neq m^{\Omega_t}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

This involves fitting regression model (2.2) to each candidate subset, computing the maximal likelihood, and evaluating the ϕ statistic. Note that the statistic ϕ incorporates both the information from the LD between SNPs and the regression of the trait on the SNPs. If adding 1 SNP does not increase the number of unique haplotypes, that is the new SNP is in perfect LD with the rest of SNPs, there is no contribution in model fitting and thus ϕ equals 0. The maximum of ϕ_{t+1} represents the largest penalty parameter in AIC_a (2.1), so that the model with 1 extra SNP is still preferable. For all $a < \max(\phi_{t+1})$, the set associated with $\max(\phi_{t+1})$ has the minimal AIC_a among all candidate sets with $t+1$ SNPs.

If the best model size is 0, there appears to be no genetic association in the region of interest. The lower the prediction deviance of the final model compared with the null model, the more likely the association. To assess significance of the associations, we perform a permutation test to correct the over-optimism incurred by the greedy model searching process. We first compute a nominal p value for the global haplotype effect in the final model using a Wald test. We then permute the trait 1000 times regardless of the genotypic data, carry out model searching for each permuted data set, and compute the nominal p value using a Wald test. When environmental factors are present, we permute the trait within the strata defined by environmental factors. Finally, the experimentwise p value is computed by comparing the observed p value to its null distribution.

When the haplotype phase is unknown, as is usually the case, our algorithm treats phasing as a part of the learning procedure. The full-scale haplotype phasing is carried out only once using all SNPs under investigation to obtain the maximal resolution. Because each training data consist of 9/10 full data, the haplotype frequencies estimated from each training set are usually a slight modification of those estimated from the full data and hence can be computed rather quickly by an Expectation-Maximization algorithm

(Excoffier and Slatkin, 1995). One can have multiple pairs of possible haplotypes, each pair has an estimated conditional probability given all genotypes. For a model associated with Ω_k, the expected deviance for a subject can be expressed as follows:

$$E[\mathcal{D}_i^{\Omega_k}|\mathbf{G}_i^{\Omega_K}] = -\sum_{\mathbf{H}_i^{\Omega_k}\in\mathbf{G}_i^{\Omega_K}} \widehat{\Pr}(\mathbf{H}_i^{\Omega_k}|\mathbf{G}_i^{\Omega_K}, \mathbf{p}^{\Omega_K}) \log[\Pr(\mathbf{Y}_i|\mathbf{H}_i^{\Omega_k}, \mathbf{Z}_i)], \tag{2.5}$$

where $\widehat{\Pr}(\mathbf{H}_i^{\Omega_k}|\mathbf{G}_i^{\Omega_K}, \mathbf{p}^{\Omega_K})$ is the estimated conditional probability of the haplotype pair for the SNPs in the model given the all genotypes. For each $\mathbf{H}_i^{\Omega_k}$, this conditional probability is equal to the sum of a collection of $\widehat{\Pr}(\mathbf{H}_i^{\Omega_K}|\mathbf{G}_i^{\Omega_K}, \mathbf{p}^{\Omega_K})$ since each haplotype formed by Ω_k represents a cluster of haplotypes formed by Ω_K. For the final model, robust sandwich variance estimates are used to compute the nominal p value. Note that under the null hypothesis that there is no genetic effect, the estimation of haplotype frequencies is independent of the estimation of the regression parameters, so using a sandwich variance estimate yields a valid test for any global genetic effect.

The core of the SHARE algorithm is written in `C` with an `R` interface. An `R`-package can be downloaded from the first author's homepage: http://www.scharp.org/faculty/jdai as well as `CRAN` (The Comprehensive R Archive Network). Currently, a model searching process for for settings with 10–30 SNPs and 2000 subjects takes about half a second on a Dell workstation with a 3.0-GHz processor. In situations where the best model after selection is the null model, no permutation test is required as this clearly indicates nonsignificance. Otherwise, the permutation test can be speeded up by giving up early on clearly nonsignificant results (Besag and Clifford, 1991).

3. RESULTS

3.1 *Simulations*

The details of the simulations are in the *Biostatistics* online supplementary document. We compared the empirical type I errors and the power of detecting the global genetic effect in the simulated chromosomal region when a hypothesis test is performed with a type I error of 0.05. As benchmarks of our comparison, we used the single-locus scan and the naive haplotype analysis using all SNPs, assuming known phase. The haplotype score test (Schaid *and others*, 2002) was included for evaluating the impact of haplotype ambiguity when model selection is not employed. We used CLADHC as an example of cladistic approaches in the comparison. The window size for CLADHC is set to 6 SNPs throughout simulations. We performed the SHARE analysis with and without knowing haplotype phase, with the maximal number of 6 SNPs in the candidate sets and 10-fold cross-validation. For all methods except the 2 using full haplotypes, permutation tests were used to correct for multiple testing.

We first simulated 2 disease models using the empirical haplotype frequencies of the F11 gene in the PGA database. Among the 45 SNPs in this gene, 11 tagSNPs (2, 3, 5, 6, 9, 11, 22, 24, 30, 42, and 45) are observed in 800 cases and 800 controls. For our first model, we simulated an unscored functional variant, which has a minor allele frequency (MAF) around 0.05 and is strongly tagged by 2 tagSNPs. Table 1 displays the haplotype frequencies formed by the 3 SNPs. Both tagSNPs 3 and 5 are correlated with SNP 12, and the haplotype `11` formed by these 2 SNPs perfectly predicts SNP 12. When an additive disease risk is added to SNP 12 using a logistic penetrance function with odds ratios (ORs) 1.5, 1.75, and 2.0, it is anticipated that a haplotype analysis that is just using SNPs 3 and 5 will best capture the disease signal. In the upper panel of Table 2, all methods yield valid tests, as the type I errors are all within a reasonable range of 0.05. Clearly, CLADHC has the worst power since the recombination between SNP 3 and SNP 5 renders it hard to construct a correct cladogram. There is a clear advantage of SHARE over the full haplotype

Table 1. *The first model in simulations based on empirical data: 2 tagSNPs display strong LD with the unscored causal locus. TagSNPs 3 and 5 are genotyped and SNP 12 is the unscored functional locus. Haplotype* `11` *by SNPs 3 and 5 perfectly tagsSNP 12. The haplotype frequencies are estimated from 23 Americans of European descend in the PGA database*

Haplotype/SNP	3	5	12	Haplotype frequency
1	0	0	0	0.847
2	0	1	0	0.087
3	1	0	0	0.022
4	1	1	1	0.044

Table 2. *Simulations based on empirical data: a comparison of type I errors and power for various methods under 2 disease model in 500 simulations. Standard errors are given in parentheses. For the first model, we generate data for 800 cases and 800 controls with ORs 1.5, 1.75, and 2; for the second model, we generate data for 400 cases and 400 controls with ORs 1.25, 1.5, and 1.75*

	Method	Type I error	Power		
			OR = 1.5	OR = 1.75	OR = 2
Model 1†	Single-locus scan	0.048 (0.010)	0.318 (0.015)	0.648 (0.015)	0.914 (0.013)
Phase known	Full haplotype	0.052 (0.010)	0.284 (0.020)	0.590 (0.022)	0.870 (0.015)
	CLADHC	0.056 (0.010)	0.256 (0.020)	0.546 (0.022)	0.854 (0.016)
	SHARE	0.050 (0.010)	0.336 (0.021)	0.654 (0.021)	0.928 (0.012)
Phase unknown	Haplotype score	0.035 (0.008)	0.288 (0.020)	0.544 (0.022)	0.863 (0.015)
	SHARE	0.054 (0.010)	0.326 (0.021)	0.650 (0.021)	0.900 (0.013)
			OR = 1.25	OR = 1.5	OR = 1.75
Model 2‡	Single-locus scan	0.046 (0.007)	0.176 (0.017)	0.608 (0.015)	0.916 (0.012)
Phase known	Full haplotype	0.046 (0.009)	0.184 (0.017)	0.616 (0.022)	0.920 (0.012)
	CLADHC	0.062 (0.011)	0.138 (0.015)	0.548 (0.022)	0.882 (0.014)
	SHARE	0.046 (0.009)	0.182 (0.017)	0.678 (0.021)	0.952 (0.010)
Phase unknown	Haplotype score	0.050 (0.010)	0.158 (0.016)	0.586 (0.022)	0.900 (0.013)
	SHARE	0.044 (0.009)	0.190 (0.018)	0.666 (0.021)	0.942 (0.010)

†The unscored disease-causing locus is best captured by haplotypes based on 2 tagSNPs.
‡Two tagSNPs separated apart carry disease risk additively.

analysis because of SNP selection. SHARE only slightly outperforms the single-locus scan since the r^2 between tagSNP 3 and SNP 12 is fairly high (0.63). Among the final models selected by SHARE with $p < 0.05$, the median size of the best SNP set is 2. Approximately one-third of the significant models contain only 1 SNP, again because of the marginal correlation between tagSNP 3 and SNP 12. When the full haplotype test assuming phase is known and the score test with phase ambiguity are compared, it is seen that phase ambiguity diminishes the power by at most 5% (for OR = 1.75). Less impact from haplotype ambiguity was observed for SHARE. The reason is that model selection leads to far fewer SNPs being used in the final association and hence far less haplotype ambiguity needs to be resolved.

For the second model, we simulated 2 tagSNPs (5, 45) contributing independently, rather than through a haplotype effect, to the disease risk. TagSNP 5 has MAF 0.13 and tagSNP 45 has MAF 0.087. The disease risk was imposed via a logistic function with ORs 1.25, 1.5, and 1.75. We generated 400 cases and 400 controls. The bottom panel of Table 2 reveals that SHARE yields the best power regardless

of whether the phase is known or not. The cladistic approach has the worst power because a moving window of 6 SNPs will not cover both tagSNP 5 and tagSNP 45. However, a longer window does not help much since it is more likely to introduce recombination among SNPs. Neither the single-locus scan nor the full haplotype analysis outperforms SHARE since the single-locus scan does not combine the effects from the 2 loci and the full haplotype analysis uses too many parameters. Again phase ambiguity does not substantively impact the power of SHARE. Due to the small sample variation, SHARE with phase ambiguity seems to outperform SHARE without phase ambiguity when OR is 1.25, although this improvement is not statistically significant.

We now evaluate the performance of SHARE on average in a variety of models generated by sampling haplotypes based on coalescence theory (Hudson, 2002) and randomly assigning 1 relatively rare SNP (with MAF $\approx$ 0.05) to carry the disease risk. Table 3 shows the comparison of type I error and power for the various methods more than 500 simulations. The recombination rates are set to reflect regions with background recombination rate and regions with high recombination rate, such as the regions near or within hot spots. When the LD is high, the median number of common haplotypes is nearly the same as the number of tagSNPs. Because the disease locus is left out before tagSNP selection, the average maximal r^2 between any tagSNP and the underlying disease locus is merely 0.36, and in only 12% of the simulations, the maximal r^2 between tagSNPs and the underlying disease locus is larger than 0.8. In this case, the haplotype-based methods generally yield higher power than the single-locus scan, particularly when the signal is strong (OR = 2). CLADHC performs only slightly better than the naive full haplotype approach. It appears that not every tagSNP is necessarily useful in detecting association, as SHARE outperforms all other methods, having 5–15% more power, particularly when the OR is more than 1.75. Since the LD is strong, haplotype ambiguity has little effect on power even for the naive haplotype analysis. On the other hand, the lower half panel in Table 3 suggests that the high recombination rate drastically reduces the power for all methods considered. The average maximal r^2 between tagSNPs and the underlying disease

Table 3. *Simulations based on coalescence by* `ms`*: a comparison of type I errors and power for various methods in 500 simulations. Standard errors are given in parentheses. The high and low LD represent recombination rate per site per generation of* 10^{-9} *and* 10^{-7}*, respectively. The sample consists of 1000 cases and 1000 controls*

LD[#SNP†,#Hap‡]	Method	Type I error	Power		
			OR = 1.5	OR = 1.75	OR = 2
High[15,16]	Single-locus scan	0.052 (0.010)	0.254 (0.019)	0.416 (0.022)	0.592 (0.022)
Phase known	Full haplotype	0.050 (0.010)	0.242 (0.019)	0.456 (0.022)	0.648 (0.021)
	CLADHC	0.042 (0.009)	0.246 (0.019)	0.464 (0.022)	0.678 (0.021)
	SHARE	0.050 (0.010)	0.312 (0.021)	0.548 (0.022)	0.728 (0.020)
Phase unknown	Haplotype score	0.034 (0.008)	0.216 (0.018)	0.462 (0.022)	0.664 (0.021)
	SHARE	0.050 (0.010)	0.308 (0.021)	0.528 (0.022)	0.738 (0.020)
Low[15,30]	Single-locus scan	0.038 (0.009)	0.154 (0.016)	0.316 (0.021)	0.420 (0.022)
Phase known	Full haplotype	0.040 (0.009)	0.116 (0.014)	0.232 (0.019)	0.370 (0.022)
	CLADHC	0.044 (0.009)	0.146 (0.016)	0.328 (0.021)	0.510 (0.022)
	SHARE	0.042 (0.009)	0.174 (0.017)	0.382 (0.022)	0.516 (0.022)
Phase unknown	Haplotype score	0.032 (0.009)	0.084 (0.012)	0.148 (0.016)	0.322 (0.021)
	SHARE	0.048 (0.010)	0.160 (0.016)	0.354 (0.021)	0.498 (0.022)

† The median number of tagSNPs in 500 simulated data.

‡ The median number of common haplotypes in 500 simulations. The common haplotypes are defined as those with frequencies larger than 1%.

locus is only 0.25, and in only 4.4% simulations, the maximal r^2 between tagSNPs and the underlying disease locus exceeds 0.8. The full haplotype method yields much lower power than the single-locus scan because of the increased number of haplotypes. Although SHARE retains the best power power, the advantage over CLADHC is not as large as in the high LD scenario, suggesting that there is limited LD structure to be exploited in regions across recombination hot spots. Contrary to a naive haplotype analysis, there is little effect of phase ambiguity on the performance of SHARE in this scenario. Note that the power shown in Table 3 is the average of 500 different models. Overall, the model selection procedure by SHARE has more power than the other approaches, particularly in regions with high LD.

3.2 *Data application*

We used SHARE to re-analyze the data from a published case–control genetic association study (Smith *and others*, 2007). This study aimed to investigate the association of common genetic variation in 24 coagulation, anti-coagulation, fibrinolysis, and antifibrinolysis candidate genes with risk of incident non-fatal venous thrombosis in postmenopausal women. The participants were selected from a large integrated health care system in Washington State and consist of 349 cases and 1680 controls matched on age, hypertension status, and calendar year. In the original analysis, the single-locus scan and a full haplotype analysis were applied to each of 24 genes, assuming an additive genetic effect, adjusting for race and other matching variables (e.g. age and hypertension status). We illustrate our algorithm using the data on the tissue factor pathway inhibitor (TFPI) gene. The LD among five genotyped tagSNPs in TFPI gene is quite strong and the highest correlation occurs between 2 adjacent tagSNPs: `rs2192824` and `rs2300412` ($r^2 = 0.4$). This region has been shown to have a significant global haplotype effect (Smith *and others*, 2007). Figure 3 shows prediction deviances, estimated by 10-fold cross-validation, of various models

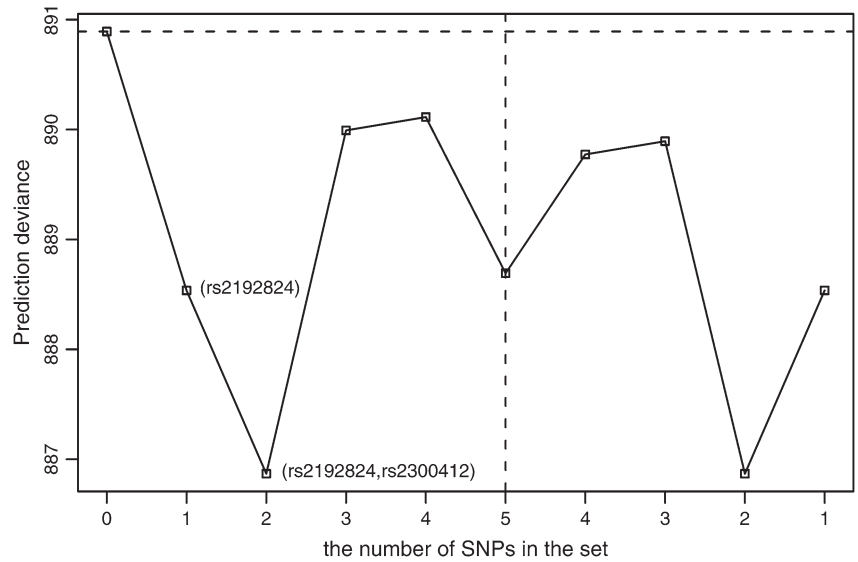

Fig. 3. The prediction deviances of different models for the TFPI gene. The horizontal axis is the number of SNPs included in the sequence of best subsets when model growing and pruning. The horizontal dashed line on the top represents the deviance of a null model without considering genetic effect. The vertical dashed line indicates the switch from model growing to pruning. The deviance is calculated from a model with haplotypes constructed from SNPs in the set. The lower the deviance is interpretted as better the model prediction.

Table 4. *The results of a SHARE analysis on 5 SNPs in the TFPI gene. SNPs 1–5 are* rs2192824, rs2300412, rs8176597, rs8176612, *and* rs3771059*, respectively. The left table shows 10 haplotypes using all 5 SNPs in the analysis. The right table shows that using SHARE, the association was narrowed down to haplotypes based on* rs2192824 *and* rs2300412

Haplotype	1	2	3	4	5	Frequency
1	0	0	0	0	0	0.2254
2	0	0	0	0	1	0.0018
3	0	0	0	1	0	0.0002
4	0	1	0	0	0	0.0593
5	0	1	0	0	1	0.2644
6	0	1	0	1	0	0.0003
7	0	1	1	0	0	0.0515
8	1	0	0	0	0	0.2989
9	1	0	0	0	1	0.0376
10	1	0	0	1	0	0.0605

Haplotype	1	2	Frequency	OR (95% confidence interval)
I	0	0	0.227	—
II	0	1	0.376	1.28 (1.02, 1.63)
III	1	0	0.397	1.43 (1.16, 1.87)

with different sets of SNPs. Lower deviances suggest better prediction accuracy. It is clear that there is a genetic effect in this gene as all models yield a smaller prediction deviance than the null model. We performed a permutation test within the strata defined by other covariates. The p value for a global null hypothesis is 0.023. It is interesting to observe that though SNP rs2192824 predicts the disease status fairly well, a haplotype model based on SNPs rs2192824 and rs2300412 further improves the prediction accuracy; inclusion of additional tagSNPs no longer helps. Inspection of the haplotype analysis using SNPs rs2192824 and rs2300412 (Table 4), we found that the increased disease risk is concentrated on haplotypes "01" and "10." This pattern implies that there might be an underlying disease-causing allele that is tagged by these 2-SNP haplotypes. We searched the SeattleSNPs database for SNPs that are not genotyped in the study. Based on 23 European descended individuals, there is a total of 54 SNPs with MAF larger than 0.05, covering a distance of 3.9 kb in this region. We found 2 SNPs (rs8676500 and rs8176531), that are in perfect LD with each other, which display a pattern of correlation to rs2192824 and rs2300412 similar to the results in the SHARE analysis. The highest pairwise r^2 between scored SNPs and unscored SNPs is 0.17, however, if we define a multilocus r^2 as in Hao *and others* (2007), the highest r^2 jumps to 0.72. Both rs8676500 and rs8176531 are located in the intron region of the TFPI gene. Although it is too early for an interpretation on this finding, this analysis gives useful hints for future studies, such as genotyping more SNPs that are correlated with rs2192824 and rs2300412 such as rs8676500 and rs8176531.

This data example shows a further benefit beyond the power enhancement that we focused on in the simulations: a parsimonious model with fewer SNPs helps us hunt for the "true" underlying "causal" mutation. The naive haplotype analysis using all 5 tagSNPs yields significant association for a global genetic effect, yet it is not clear how to pursue this analysis further based on all 10 haplotypes (shown in Table 4). On the other hand, if we are willing to take the most likely haplotype pair for subjects with haplotype phase ambiguity, we can perform a CLADHC analysis, which here suggests that the best partition of 10 haplotypes is formed by 2 clusters: a cluster with haplotypes 8 and 9 and a cluster with the remaining haplotypes. However, it is not clear how to interpret and follow-up these clusters by CLADHC.

4. DISCUSSION

Many strategies have been proposed to perform haplotype-based multilocus analyses. The majority focus on how to cluster haplotypes given a set of predefined SNPs. There are a few attempts to select SNPs

that form haplotypes. For instance, to evaluate the coverage of the Affymetrix GeneChip and Illumina BeadChip on the HapMap project data, Pe'er *and others* (2006) used multimarker predictors that capture an additional 9–25% of SNPs in the ENCODE region or in the HapMap Phase II data. The search for multimarker predictors is limited to those SNPs with strong LD. Alternatively, it has been proposed to exhaustively search for windows of contiguous SNPs that form haplotypes (Lin *and others*, 2004). The computational burden can be insurmountable for large regions and it is not clear whether only considering contiguous SNPs in a window is an effective strategy. In this article, we propose a novel strategy to select the most informative set of SNPs which in turn forms the basis for haplotype analysis. The advantage of the SHARE algorithm is its adaptivity: it exploits both the LD structure and the underlying disease-generating mechanism, and the addition or deletion of SNPs is noncontiguous. In a variety of simulation settings, SHARE consistently outperforms other existing methods.

Imputation procedures have been shown to be useful to capture the association of a phenotype and unmeasured genotypes (e.g. Nicolae, 2006; Servin and Stephens, 2007). Rather than choosing haplotypes that "tag" untyped variants, as implemented in the SHARE algorithm, these procedures impute the "missing" (untyped but known) variants based on the LD estimated from public databases (e.g. the HapMap project). For common alleles already cataloged in these databases, the imputation procedures would likely be powerful since they incorporate external information. For rare alleles (MAF $\leqslant$ 5%), the imputation methods could miss the signal (depending on the size of the database), while our method may still capture it by constructing rare haplotypes. Furthermore, if there are multiple loci with interaction effects on a disease phenotype, for example multiple nonsynonymous mutations, a haplotype analysis with model selection can be useful to integrate the joint and interactive effects of the multiple loci.

Our method is motivated, but not limited by candidate gene studies. To scale our haplotype analyses up to genome-wide association studies, our strategy is to first divide the genome into long haplotype blocks and perform an adaptive haplotype analysis in each block. We define blocks by long chromosomal regions between recombination hot spots that may stretch several hundred kbs and contain a fair number of tagSNPs ($10 \sim 50$). This is rather a loose criterion compared with the definition of haplotype blocks used in Gabriel *and others* (2002) or Wang *and others* (2002). In phase II of the HAPMAP project, 32,996 recombination hot spots were identified of which 68% are localized to a region of $\leqslant$5 kb. The spacing between adjacent hot spots is 100–200 kb on average (The International Hapmap Consortium, 2007). Permutation test on the genome-wide level would be computationally demanding. However, it is unnecessary since there should be at most a handful of blocks that suggest genetic association. In the supplementary material (available at *Biostatistics* online), we discuss strategies to reduce the computation in assessing genome-wide significance. Further details to improve SHARE to a version that can deal with genome wide association studies will be pursued in future work.

Acknowledgment

We thank Chris Carlson for helpful discussions. *Conflict of Interest:* None declared.

Funding

National Institutes of Health (grants R01 CA74841 to J.Y.D., M.L., C.K.; P01 CA53996 to M.L., C.K.; R01 CA90998 to M.L.; U01 CA125489 to M.L., C.K.; R01 HL74745 to B.P., N.L.S., C.K.; R01 HL73410 to N.L.S.; R01 HL60739 to N.L.S.; R01 HL68639 to N.L.S.; R01 HL43201 to N.L.S.; R01 HL68986 to N.L.S.). Funding to pay the Open Access publication charges for this article was provided by FHCRC Faculty Development Fund—202948.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at http://www.biostatistics.oxfordjournals.org.

REFERENCES

BARRETT, J. C., FRY, B., MALLER, J. AND DALY, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265.

BESAG, J. AND CLIFFORD, P. (1991). Sequential Monte Carlo p-value. *Biometrika* **78**, 301–304.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. AND STONE, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.

BROWNING, S. R. (2006). Multilocus association mapping using variable-length markov chains. *American Journal of Human Genetics* **78**, 903–913.

CLAYTON, D., CHAPMAN, J. AND COOPER, J. (2004). Use of unphased multilocus genotype data in indirect association. *Genetic Epidemiology* **27**, 415–428.

CORDELL, H. AND CLAYTON, D. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to hla in type 1 diabetes. *American Journal of Human Genetics* **70**, 124–141.

DURRANT, C., ZONDERVAN, K., CARDON, L., HUNT, S., DELOUKAS, P. AND MORRIS, A. (2004). Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *American Journal of Human Genetics* **75**, 35–43.

EXCOFFIER, L. AND SLATKIN, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**, 921–927.

GABRIEL, S. B., SCHAFFNER, S. G., NGUYEN, H., MOORE, J. M., ROY, J., BLUMENSTIEL, B., HIGGENS, J., DEFELICE, M., LOCHNER, A., FAGGART, M. *and others* (2002). The structure of haplotype blocks in the human genome. *Science* **21**, 2225–2229.

HAO, K., DI, X. AND CAWLEY, S. (2007). LdCompare: rapid computation of single- and multiple-marker r^2 and genetic coverage. *Bioinformatics* **23**, 252–254.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

HUDSON, R. R. (2002). Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.

LI, Y., SUNG, W. AND LIU, J. J. (2007). Association mapping via regularized regression analysis of single-nucleotide-polymorphism haplotypes in variable-sized sliding windows. *American Journal of Human Genetics* **80**, 705–715.

LIN, S., CHAKRAVARTI, A. AND CUTLER, D. J. (2004). Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nature Genetics* **36**, 1181–1188.

MOLITOR, J., MARJORAM, P. AND THOMAS, D. (2003). Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *American Journal of Human Genetics* **73**, 1368–1384.

MORRIS, A. P. (2006). A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. *American Journal of Human Genetics* **79**, 679–694.

MORRIS, R. W. AND KAPLAN, N. L. (2002). On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genetic Epidemiology* **23**, 221–233.

NICOLAE, D. L. (2006). Testing untyped alleles (TUNA)- applications to genome-wide association studies. *Genetic Epidemiology* **30**, 718–727.

PE'ER, I., DE BAKKER, P. I., MALLER, J., YELENSKY, R., ALTSHULER, D. AND DALY, M. J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics* **38**, 663–667.

ROEDER, K., BACANU, S. A., SONPAR, V., ZHANG, X. AND DEVLIN, B. (2005). Analysis of single-locus tests to detect gene/disease associations. *Genetic Epidemiology* **28**, 207–219.

SCHAID, D. J. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology* **27**, 348–364.

SCHAID, D. J., ROWLAND, C. M., TINES, D. E., ROBERT, M. J. AND POLAND, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**, 425–434.

SELTMAN, H., ROEDER, K. AND DEVLIN, B. (2001). Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *American Journal of Human Genetics* **68**, 1250–1263.

SERVIN, B. AND STEPHENS, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* **3**, e114.

SMITH, N. L., HINDORFF, L. A., HECKBERT, S. R., LEMAITRE, R. N., MARCIANTE, K. D., RICE, K., LUMLEY, T., BIS, J. C., WIGGINS, K. L., ROSENDAAL, F. R. *and others* (2007). Association of genetic variations with nonfatal venous thrombosis in postmenopausal women. *Journal of American Medical Association* **297**, 489–498.

TEMPLETON, A. R., BOERWINKLE, E. AND SING, C. F. (1987). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. basic theory and an analysis of alcohol dehydrogenase activity in drosophila. *Genetics* **117**, 343–351.

THE INTERNATIONAL HAPMAP CONSORTIUM. (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

THE INTERNATIONAL HAPMAP CONSORTIUM. (2007). A second generation human haplotype map of over 3.1 million SNP. *Nature* **449**, 851–861.

WANG, N., AKEY, J. M., ZHANG, K., CHAKRABORTY, R. AND JIN, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *American Journal of Human Genetics* **71**, 1227–1234.

YU, Z. AND SCHAID, D. J. (2007). Sequential haplotype scan methods for association analysis. *Genetic Epidemiology* **31**, 553–564.

[*Received July 18, 2008; revised February 25, 2009; accepted for publication June 22, 2009*]