Original Paper

Human Heredity

Hum Hered 2010;70:167–176
DOI: 10.1159/000308456

Received: November 6, 2009
Accepted after revision: February 2, 2010
Published online: July 30, 2010

# Fast and Robust Association Tests for Untyped SNPs in Case-Control Studies

Andrew S. Allen[a] Glen A. Satten[b] Sarah L. Bray[d] Frank Dudbridge[d, e]
Michael P. Epstein[c]

[a]Department of Biostatistics and Bioinformatics, Duke University, Durham, N.C., [b]Centers for Disease Control and Prevention, and [c]Department of Human Genetics, Emory University, Atlanta, Ga., USA; [d]MRC Biostatistics Unit, Cambridge, and [e]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

**Key Words**
Genotype imputation · Genome-wide association study · Efficient score · Case-control study

**Abstract**
Genome-wide association studies (GWASs) aim to genotype enough single nucleotide polymorphisms (SNPs) to effectively capture common genetic variants across the genome. Even though the number of SNPs genotyped in such studies can exceed a million, there is still interest in testing association with SNPs that were not genotyped in the study sample. Analyses of such untyped SNPs can assist in signal localization, permit cross-platform integration of samples from separate studies, and can improve power – especially for rarer SNPs. External information on a larger collection of SNPs from an appropriate reference panel, comprising both SNPs typed in the sample and the untyped SNPs we wish to test for association, is necessary for an untyped variant analysis to proceed. Linkage disequilibrium patterns observed in the reference panel are then used to infer the likely genotype at the untyped SNPs in the study sample. We propose here a novel statistical approach for testing untyped SNPs in case-control GWAS, based on an efficient score function derived from a prospective likelihood, that automatically accounts for the variability in the process of estimating the untyped variant. Computationally efficient methods of phasing can be used without affecting the validity of the test, and simple measures of haplotype sharing can be used to infer genotypes at the untyped SNPs, making our approach computationally much faster than existing approaches for untyped analysis. At the same time, we show, using simulated data, that our approach often has performance nearly equivalent to hidden Markov methods of untyped analysis. The software package 'untyped' is available to implement our approach.



## Introduction

The rapid improvement and reduced cost of high-throughput genotyping have enabled the use of genome-wide association studies (GWASs) in case-control studies to identify genetic variants that increase the risk for many complex diseases [1]. However, while the cost of high-throughput genotyping has steadily decreased, it is still financially impractical to genotype all existing genetic polymorphisms throughout the human genome in a large

A.S. Allen and G.A. Satten contributed equally and should each be considered as first author.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com



Accessible online at:
www.karger.com/hhe

Andrew S. Allen, PhD
Department of Biostatistics and Bioinformatics
Duke University
2424 Erwin Road, Suite 1102, Durham, NC 27710 (USA)
Tel. +1 919 668 8043, Fax +1 919 668 7049, E-Mail andrew.s.allen@duke.edu

**Table 1.** Reference sample to estimate allele frequencies at loci that are untyped in a study sample

| Sample haplotype | Reference haplotypes |
|---|---|
| 10?00 | 11**1**00 |
| | 11**1**00 |
| | 00**0**00 |
| | 00**0**00 |
| | 00**0**00 |
| | 10**1**00 |
| | 10**0**00 |
| | 10**0**00 |
| | 10**0**00 |
| | 11**1**00 |

Question mark denotes location of untyped locus in the study sample. Bold 0 or 1 denote alleles in the reference sample at the untyped locus. Underlined reference haplotypes are those with the same allelic states at those loci observed in the sample haplotype. Restricting computations to these haplotypes gives an allele frequency estimate (1/4) that is quite different from the allele frequency in the reference population (4/10).

study. The commercial platforms used for most GWASs select a set of single nucleotide polymorphisms (SNPs) for genotyping with the goal of achieving a sufficiently high SNP density to map all genomic regions that harbor disease susceptibility variants. Although the number of SNPs available on commercial platforms has been steadily increasing, there remain a large number of potentially interesting SNPs that are not included in any commercial platform. Inference using these 'untyped' variants could serve several useful purposes in a GWAS of a complex disease. Most importantly, studies investigating the value of untyped variants suggest that their incorporation into a GWAS can improve power to detect rare disease susceptibility loci without compromising the power to detect more common susceptibility loci [2, 3]. Additionally, the analysis of untyped variants could assist in the localization of the disease susceptibility signal and suggest additional SNPs for genotyping in a replication study. Finally, such analyses may facilitate cross-platform comparisons and meta-analyses of specific SNPs that may not be genotyped in all relevant study samples, especially among studies that use different commercial genotyping platforms.

Several approaches [2–9] have been developed for the statistical analysis of untyped SNP loci. These methods all use the linkage disequilibrium information in a reference panel (typically, the HapMap Project [10]) to facilitate this inference. The reference panel contains haplotype information from a reference population and includes information on SNPs that are typed in study participants as well as additional untyped SNP loci. Knowledge of the typed SNPs in the study population, as well as the haplotype structure in the reference population, is then used to make inference on the association between putative trait loci and untyped loci in the study population. Table 1 illustrates how a reference panel can be used to estimate allele frequencies at loci that are untyped in a study sample.

Nicolae [4, 5] developed TUNA, the first approach to testing for association between trait and untyped alleles (see also Zaitlen et al. [6]). For each untyped locus, TUNA selects a small set of tagSNPs from the typed SNPs and then uses the reference panel to construct haplotypes comprising the tagSNPs and the untyped locus. Comparing the frequency of the tagSNP haplotypes in cases and controls then allows comparison of allele frequencies at the untyped locus. Haplotypes must be inferred in the study population assuming Hardy-Weinberg equilibrium; at the genome-wide scale, this limits the number of tagSNPs per untyped locus that can be used. Difficulties arise when the study population has haplotypes that do not appear in the reference panel or when the untyped SNP is not well tagged by a small number of SNPs. A computationally intensive bootstrap procedure is required for inference.

Lin et al. [8] developed a retrospective likelihood [11] framework for testing and estimating the effects of untyped SNPs in case-control studies, implemented in the programs HAPSTAT and SNPMstat. This approach requires joint estimation of haplotype frequencies and association parameters among the study participants along with samples from the reference panel (e.g. HapMap), which can be slow and limits haplotypes to a small number of SNPs. Although HAPSTAT and SNPMstat have higher power than TUNA, their genome-wide application can be computationally intensive. Further, like TUNA, HAPSTAT and SNPMstat implicitly assume that untyped SNPs are well tagged by a few adjacent SNPs.

Several computationally intensive approaches have been proposed for imputing genotypes at an untyped locus using hidden Markov models [2, 3, 7, 9]. These approaches use population genetics theory [12] to impute the genotype at the untyped locus, then test for association between the imputed genotype and case-control status. Because chromosome level data are used, there is no implicit assumption that nearby SNPs tag an untyped variant, which is an important advantage, especially when considering untyped variants having low minor allele frequency (MAF). Simulations indicate that these

hidden Markov methods typically reconstruct missing genotypes with high accuracy [2, 3].

Here, we consider the problem of association testing for an untyped variant as a type of haplotype regression. We are thus able to use the efficient score framework of Allen and Satten [13] as the basis for inference. This framework allows us to explicitly address several issues. First, the tests we develop are robust to misspecification of the distribution of haplotypes given the observed genotype data. Second, estimation of this distribution requires no further variance adjustment; one can simply 'plug in' an estimate and the variance of the efficient score remains valid. These two facts allow us to use computationally efficient estimators that may not correspond to maximum likelihood estimators nor elicit simple variance formulas. In particular, our approach is able to utilize information from an entire chromosome without fitting a computationally expensive hidden Markov model. As a result, the computation time required to analyze untyped SNPs using our efficient score method is a tiny fraction of the computation time required for competing untyped methods, requiring only approximately 90 min to analyze 1.6 million untyped, unphased SNPs in a case-control dataset of 1,000 subjects on a single processor. At the same time, we show, using simulated data, that our efficient score test performs almost as well as hidden Markov methods in nearly all situations considered. Finally, the haplotype regression framework provides a natural approach to incorporating covariates into the analysis.

## A Framework for Testing Hypotheses about Association with an Unmeasured Locus

Testing for association at an unobserved locus using observed genotypes at nearby loci can be considered as a type of haplotype regression. Haplotype regression analysis of case-control data seeks to determine the effect of individual haplotypes or diplotypes (haplotype pairs) on the risk of disease, by making inference on parameters in a model for the odds of disease, given diplotype and possibly other covariates. We consider haplotypes for $L$ biallelic loci. For simplicity, we assume the alleles are labeled '0' and '1'. For an individual, let $h$ denote a pair of haplotypes (a diplotype) and let $g$ denote the genotypes at the $L$-typed loci, ignoring phase information. Let $e$ denote environmental covariates that may be confounders but are not effect modifiers. Then, the odds of disease can be written as

$$\frac{P\left[D=1|h,e;\beta,\gamma\right]}{P\left[D=0|h,e;\beta,\gamma\right]}\equiv\Theta\left(h,e|\beta,\gamma\right)=e^{X^T(h)\cdot\beta+Z^T(e)\cdot\gamma} \quad (1)$$

where $X(h)$ and $Z(e)$ are design vectors that code the genetic and environmental contributions to risk, respectively, and $\beta$ and $\gamma$ are parameters. The haplotype regression analysis seeks methods for inference about parameters $\beta$ and $\gamma$ in model (1).

To see the connection between haplotype regression and testing association at untyped loci, suppose that we let $X(h)$ be a function of reference panel data that encodes information about the chance that the haplotypes comprising diplotype $h$ would have the '1' allele at the untyped locus. For example, $X(h)$ could be the average number of '1' alleles among persons with diplotype $h$ in a reference panel where the untyped locus was in fact genotyped. For such a model, a test of the (composite) null hypothesis $H_0$: $\beta = 0$ is a test of the effect of the number of '1' alleles at the unmeasured locus.

To develop methods for inference on $\beta$ we use the (prospective) likelihood. Given (multilocus) genotypes $g$ and environmental covariates $e$, the contribution to the (prospective) likelihood from a single individual $i$ is

$$\mathcal{L}_i\left(g,e,d\right)=\sum_{h\in\mathcal{H}(g)}\frac{\Theta\left(h,e|\beta,\gamma\right)^d}{1+\Theta\left(h,e|\beta,\gamma\right)}\varphi\left(h|g,e\right), \quad (2)$$

where $\mathcal{H}(g)$ is the set of diplotypes that are consistent with genotype $g$ and $\varphi(h \mid g, e) = \Pr(H = h \mid g = g, E = e)$. The score function implied by (2) has the form

$$\boldsymbol{\mathcal{S}}=\sum_i\boldsymbol{\mathcal{S}}_i\left(g_i|e_i,d_i\right),$$

where the contribution of the $i$-th individual is

$$\boldsymbol{\mathcal{S}}_i\left(g|e,d\right)=\begin{pmatrix}\boldsymbol{\mathcal{S}}_{i,\beta}\left(g|e,d\right)\\ \boldsymbol{\mathcal{S}}_{i,\gamma}\left(g|e,d\right)\end{pmatrix}=\sum_{h\in\mathcal{H}(g)}w\left(h|e,g,d\right)\boldsymbol{\mathcal{S}}_i\left(h|e,d\right),$$

where

$$w\left(h|e,g,d\right)=\frac{\dfrac{\Theta\left(h,e|\beta,\gamma\right)^d}{1+\Theta\left(h,e|\beta,\gamma\right)}\varphi\left(h|e,g\right)}{\displaystyle\sum_{h\in\mathcal{H}(g)}\frac{\Theta\left(h,e|\beta,\gamma\right)^d}{1+\Theta\left(h,e|\beta,\gamma\right)}\varphi\left(h|e,g\right)},$$

and

$$\boldsymbol{\mathcal{S}}_i\left(h|e,d\right)=\begin{pmatrix}\boldsymbol{\mathcal{S}}_{i,\beta}\left(h|e,d\right)\\ \boldsymbol{\mathcal{S}}_{i,\gamma}\left(h|e,d\right)\end{pmatrix}=\left\{d-\frac{\Theta\left(h,e|\beta,\gamma\right)}{1+\Theta\left(h,e|\beta,\gamma\right)}\right\}\boldsymbol{U}(e,h),$$

where $\boldsymbol{U}(e, h) = (X(h)^T, Z(e)^T)^T$. The total information matrix implied by (2) has the form

$$\boldsymbol{\mathcal{I}}=\sum_i\boldsymbol{\mathcal{I}}_i\left(g_i|e_i,d_i\right),$$

where the contribution from the $i$-th individual is

$$\begin{aligned}\boldsymbol{\mathcal{I}}_i\left(g|e,d\right)&=\frac{\partial\boldsymbol{\mathcal{S}}_i\left(g|e,d\right)}{\partial(\beta,\gamma)^T}\\&=\sum_{h\in\mathcal{H}(g)}\left\{\boldsymbol{\mathcal{I}}_i\left(h|e\right)+\boldsymbol{\mathcal{S}}_i\left(h|e,d\right)\boldsymbol{\mathcal{S}}_i^T\left(h|e,d\right)\right\}w\left(h|e,g,d\right)\\&\quad-\boldsymbol{\mathcal{S}}_i\left(g|e,d\right)\boldsymbol{\mathcal{S}}_i^T\left(g|e,d\right)\end{aligned}$$

and where

$$\boldsymbol{\mathcal{I}}_i(h|e)=\frac{\Theta\left(h,e|\beta,\gamma\right)}{\left[1+\Theta\left(h,e|\beta,\gamma\right)\right]^2}\boldsymbol{U}(e,h)\boldsymbol{U}^T(e,h).$$

Under the (composite) null hypothesis $H_0$: $\beta = 0$ we have $\Theta(h, e|\beta = 0,\hat{\gamma}) = \Theta(e|\hat{\gamma})$, where $\hat{\gamma}$ is the maximum likelihood estimator for $\gamma$ when $\beta$ is fixed at 0. Factoring $\mathcal{I}$ into blocks according to whether the parameters $\beta$ or $\gamma$ are being referenced, we find that under the null hypothesis, the $\gamma\gamma$ block simplifies to

$$\mathcal{I}_{0;\gamma\gamma} = \sum_i \frac{\Theta\left(e_i|\hat{\gamma}\right)}{\left[1+\Theta\left(e_i|\hat{\gamma}\right)\right]^2} Z_i Z_i^T,$$

where $Z_i = Z(e_i)$, and the $\beta\gamma$ block simplifies to

$$\mathcal{I}_{0;\beta\gamma} = \sum_i \frac{\Theta\left(e_i|\hat{\gamma}\right)}{\left[1+\Theta\left(e_i|\hat{\gamma}\right)\right]^2} m_i Z_i^T,$$

where

$$m_i = E\left[X(h)|g_i,e_i\right] = \sum_{h\in\mathcal{H}(g_i)} X(h)\varphi\left(h|g_i,e_i\right).$$

We base all inference on the efficient score function evaluated under the (composite) null hypothesis $H_0$: $\beta = 0$

$$\tilde{\mathcal{S}}_{i,\beta}\left(g|e,d\right) = \mathcal{S}^{(0)}_{i,\beta}\left(g|e,d\right) - \mathcal{I}_{0;\beta\gamma}\mathcal{I}^{-1}_{0;\gamma\gamma}\mathcal{S}^{(0)}_{i,\hat{\gamma}}\left(g|e,d\right), \tag{3}$$

where $\mathcal{S}^{(0)}_i(g \mid e, d)$ is the score function evaluated at the (composite) null hypothesis given by

$$\begin{pmatrix}\mathcal{S}^{(0)}_{i,\beta}\left(g|e,d\right)\\ \mathcal{S}^{(0)}_{i,\hat{\gamma}}\left(g|e,d\right)\end{pmatrix} = \begin{pmatrix} d - \dfrac{\Theta\left(e|\hat{\gamma}\right)}{1+\Theta\left(e_i|\hat{\gamma}\right)}\end{pmatrix}\begin{pmatrix} m_i \\ Z_i\end{pmatrix}.$$

## Inference Based on the Efficient Score Function

Inference based on (3) explicitly accounts for the effect of estimating $\gamma$ on inference on $\beta$. However, at first glance, it does not seem to address the effect of estimation of $\varphi(h \mid g, e)$. Fortunately, the likelihood (2) is identical to that considered by Allen and Satten [13]. Therefore, if we assume a saturated (categorical) model for $\varphi(h \mid g, e)$, then (3) remains the efficient score function for $\beta$ under this extended model. Further, $\tilde{\mathcal{S}}_\beta$ has mean zero under the null hypothesis even if $\varphi(h \mid g, e)$ is misspecified. Thus, improper specification of $\varphi(h \mid g, e)$ can only affect the power and not the validity of the test. Because of these properties, the nuisance parameters $\varphi(h \mid g, e)$ can be replaced by estimates from some working models (that may or may not be correct) without affecting inference based on $\tilde{\mathcal{S}}_\beta$. These properties are important in the untyped variant problem as they allow $\varphi(h \mid g, e)$ to be estimated using computationally efficient methods which, although not necessarily statistically optimal, enable an extremely fast procedure.

Tests based on the efficient score function can be constructed as follows. We first obtain an estimate of $\varphi(h \mid g, e)$ in any convenient way. We note that explicitly modeling $\varphi(h \mid g, e)$ as a function of $e$ is difficult and may require restrictive modeling assumptions in order to arrive at an identifiable model, especially when $e$ is made up of continuous covariates. However, as noted above, a particular strength of our approach is that it is robust to misspecification of $\varphi(h \mid g, e)$. In fact, we will often ignore $e$ in specifying $\varphi(h \mid g, e)$, replacing $\varphi(h \mid g, e)$ with $\varphi(h \mid g)$ throughout, secure in the knowledge that by doing so, we will not impact the validity of our procedure. Using this estimate, we calculate $m_i$ for each study participant. If $\beta$ is a scalar, a test of the (composite) null hypothesis $H_0$: $\beta = 0$ can be constructed as

$$T = \frac{\left(\sum_i \tilde{\mathcal{S}}_{i,\beta}\right)^2}{\sum_i \tilde{\mathcal{S}}^2_{i,\beta} - \frac{1}{N}\left(\sum_i \tilde{\mathcal{S}}_{i,\beta}\right)^2} = \frac{\left(\sum_i \tilde{\mathcal{S}}^{(0)}_{i,\beta}\right)^2}{\sum_i \tilde{\mathcal{S}}^2_{i,\beta} - \frac{1}{N}\left(\sum_i \tilde{\mathcal{S}}^{(0)}_{i,\beta}\right)^2}$$

which has an asymptotic $\chi^2_1$ distribution. If $\beta$ is not a scalar, then we can test using

$$T = \left(\sum_i \tilde{\mathcal{S}}_{i,\beta}\right)^T \hat{\Sigma}^{-1}\left(\sum_i \tilde{\mathcal{S}}_{i,\beta}\right),$$

where $\hat{\Sigma}$ is the empirical variance-covariance matrix of $\tilde{\mathcal{S}}_{i,\beta}$.

A useful observation is that $\mathcal{I}_{0;\gamma\gamma}$ and $\mathcal{S}^{(0)}_{i,\hat{\gamma}}$ do not depend on the locus in question (under the (composite) null hypothesis $H_0$: $\beta = 0$, the estimate $\hat{\gamma}$ is the same for all loci). Hence, they need only be calculated once per genome. If we define

$$\mathbf{V}_i = \mathcal{I}^{-1}_{0;\gamma\gamma}\mathcal{S}^{(0)}_{i,\gamma},$$

then we need only store $V_i$ (a matrix having as many rows as elements of $\beta$ and having as many columns as elements of $\gamma$). In the most important case ($\beta$ is a scalar), $\mathbf{V}_i$ is a vector. For this case, we calculate the (row) vector $\mathcal{I}_{0;\beta\gamma}$ for each locus and then compute the efficient score as

$$\mathcal{S}^{(0)}_{i,\beta}, - \mathcal{I}_{0;\beta\gamma} \cdot \mathbf{V}_i.$$

## The Special Case of Stratified Data

The efficient score can be calculated in closed form when $\gamma$ corresponds to stratification with no additional covariates. Assume that there are $K$ strata. To simplify the presentation, we will continue to consider the case when $\beta$ is a scalar. Let $Z(e) = (I[e = 1], I[e = 2], ..., I[e = K])$. Under the (composite) null $\beta = 0$, we have

$$\sum_i \mathcal{S}^{(0)}_{i,\gamma_k}\left(g_i|e_i,d_i\right) = \sum_i \left\{d_i - \frac{e^{\gamma_k}}{1+e^{\gamma_k}}\right\} I\left[e_i = k\right],$$

so that

$$\frac{e^{\hat{\gamma}_k}}{1+e^{\hat{\gamma}_k}} = \frac{n_{dk}}{n_k},$$

where $n_{dk}$ and $n_k$ are the number of cases and total participants in the $k$-th stratum, respectively. Because each person can belong to only one stratum, $\mathcal{I}_{0;\gamma\gamma}$ is a diagonal matrix. The $k$-th element is given by

$$\mathcal{I}_{0,\gamma\gamma} = \sum_i \frac{n_{dk}n_{ck}}{n_k^2} I\left[e_i = k\right] = \frac{n_{dk}n_{ck}}{n_k},$$

where $n_{ck} = n_k - n_{dk}$. Similarly, the vector $\mathcal{I}_{0;\beta\gamma}$ has the $k$-th component

$$\mathcal{I}^k_{0;\beta\gamma} = \sum_i \frac{n_{dk}n_{ck}}{n_k^2} m_i I\left[e_i = k\right] = \frac{n_{dk}n_{ck}}{n_k}\bar{m}_k,$$

where

$$\bar{m}_k = \frac{1}{n_k}\sum_i m_i I\left[e_i = k\right].$$

Thus, the efficient score is

$$\tilde{\mathcal{S}}_{i,\beta} = \left(d_i - \frac{n_{dk_i}}{n_{k_i}}\right)\left(m_i - \bar{m}_{k_i}\right),$$

where $k_i$ denotes the stratum for the $i$-th participant. Of course, when there is only one stratum, we have

$$\tilde{\mathcal{S}}_{i,\beta} = \left(d_i - \frac{n_d}{n}\right)\left(m_i - \bar{m}\right). \quad (4)$$

Thus, the test statistic for association at an unmeasured locus is

$$T = \frac{\left\{\sum_i \left(d_i - \frac{n_{dk_i}}{n_{k_i}}\right)\left(m_i - \bar{m}_{k_i}\right)\right\}^2}{\sum_i \left(d_i - \frac{n_{dk_i}}{n_{k_i}}\right)^2 \left(m_i - \bar{m}_{k_i}\right)^2 - \frac{1}{n}\left\{\sum_i \left(d_i - \frac{n_{dk_i}}{n_{k_i}}\right)\left(m_i - \bar{m}_{k_i}\right)\right\}^2}$$

which compares with the logistic regression test for trend for a covariate $m_i$ that is separately centered in each stratum.

## Specifying $X(h)$ Using Haplotype Sharing

We consider here possible choices for $X(h)$ and the subsequent computation of

$$m_i = E\left(X(h)|g_i, e_i\right) = \sum_{h \in \mathcal{H}(g_i)} X(h)\varphi\left(h|g_i, e_i\right).$$

The design matrix $X(h)$ uses information from a reference panel comprising individuals who have available genotypes at a larger number of loci than those observed in the study population. Data on genotypes at loci that are typed in the reference panel but untyped in the study population are used to construct $X(h)$. Here, we give a general approach to constructing design matrices and advocate the use of haplotype sharing to construct $X(h)$.

Let $h_j^r$ denote the $j$-th haplotype in the reference panel comprised of loci that are typed in the study population. Let $a_j^r$ be the allele at an untyped locus of interest corresponding to the reference panel's $j$-th haplotype. We denote the individual haplotypes comprising a sample diplotype $h$ by $(h_1, h_2)$ and note, as observed by Nicolae [4, 5], that

$$\frac{\sum_j I\left(h_1 = h_j^r\right) I\left[a_j^r = 1\right]}{\sum_j I\left(h_1 = h_j^r\right)} \quad (5)$$

gives the frequency of the '1' allele at the untyped locus among reference panel haplotypes that are identical by state at the typed loci to haplotype $h_1$. This forms a reasonable estimate of the likelihood that haplotype $h_1$ contains the '1' allele at the untyped locus and suggests a particularly simple approach to specifying $X(h)$. For example, an additive model (at the untyped locus) can be specified by taking $X(h)$ to be

$$X(h) = \frac{\sum_j I\left(h_1 = h_j^r\right) I\left[a_j^r = 1\right]}{\sum_j I\left(h_1 = h_j^r\right)} + \frac{\sum_j I\left(h_2 = h_j^r\right) I\left[a_j^r = 1\right]}{\sum_j I\left(h_2 = h_j^r\right)}. \quad (6)$$

Other models, including recessive and dominant models, can similarly be constructed using equation (5).

When the additive model given by (6) is used, the proposed untyped variant test compares the average number of '1' alleles in cases to the average number of '1' alleles in controls. To see this, write the score function by summing (4) over study participants to yield

$$\tilde{\mathcal{S}}_\beta = \sum_i m_i \left(d_i - \frac{n_d}{n}\right),$$

where the term involving $\bar{m}$ can be ignored after summing over individuals. We can rewrite $\tilde{\mathcal{S}}_\beta$ as

$$\tilde{\mathcal{S}}_\beta = \sum_i m_i d_i - \frac{\sum_i d_i \sum_i m_i}{n}.$$

If we write

$$\sum_i m_i = \sum_i m_i d_i + \sum_i m_i \left(1 - d_i\right),$$

we obtain

$$\tilde{\mathcal{S}}_\beta = \frac{n_d n_c}{n}\left(\bar{m}_d - \bar{m}_c\right), \quad (7)$$

where

$$\bar{m}_d = \frac{1}{n_d}\sum_i m_i d_i$$

and

$$\bar{m}_c = \frac{1}{n_c}\sum_i m_i \left(1 - d_i\right).$$

When the additive model given by (6) is used, $\bar{m}_d$ is the average number of estimated '1' alleles among the cases, while $\bar{m}_c$ is the average number of estimated '1' alleles among the controls. Equation (7) establishes a connection between the efficient score and TUNA; for phase-certain data, TUNA also compares the average number of estimated '1' alleles in the case and control populations. Our approach generalizes this comparison to phase-uncertain data, while also easily allowing for covariates.

The approach of exactly matching sample haplotypes to reference panel haplotypes implied by (6), though intuitive, is limited to a relatively small number of tagSNPs flanking the untyped locus. Otherwise, it will often be the case that no haplotypes in the reference panel will exactly match up with sample haplotypes, and $X(h)$ will fail to be well defined. This problem can be mitigated somewhat by restricting calculations concerning a given individual's data to the largest subset of the tagSNP loci that allow exact matching between the sample and reference panel haplotypes. Unfortunately, computational demands preclude this approach from being a whole-scale remedy, and one is forced to consider cases where exact matching is likely, e.g. by severely limiting the number of tagSNPs used for each untyped locus. Moreover, simulations (see Results section below) demonstrate that a small number of tagSNPs often fail to capture the information in the sample concerning the untyped locus, leading to a loss of power. Approaches based on hidden Markov models avoid these difficulties by using chromosome-scale information when imputing genotypes. The program MACH [3] can be used to estimate $m_i$ by using the '–dosage' flag to output the expected number of '1' alleles averaged over the sampled Markov chains. However, the computational requirements can be prohibitive.

**Table 2.** Estimating untyped allele frequencies using haplotype sharing

| Subject's haplotype $h_1$ | $j$ | Reference haplotypes | $h_j^r$ | $a_j^r$ | $w(h_1, h_j^r)$ | $w(h_1, h_j^r)\, I[a_j^r = 1]$ |
|---|---|---|---|---|---|---|
| ...1010**?**0001... | 1 | ...00111**0**000... | ...00110000... | 1 | 0 | 0 |
| | 2 | ...00111**0**000... | ...00110000... | 1 | 0 | 0 |
| | 3 | ...0000**0**0000... | ...00000000... | 0 | 0 | 0 |
| | 4 | ...0000**0**0000... | ...00000000... | 0 | 0 | 0 |
| | 5 | ...0000**0**0000... | ...00000000... | 0 | 0 | 0 |
| | 6 | ...0010**1**0000... | ...00100000... | 1 | 1 | 1 |
| | 7 | ...0010**0**0000... | ...00100000... | 0 | 1 | 0 |
| | 8 | ...0010**0**0000... | ...00100000... | 0 | 1 | 0 |
| | 9 | ...0010**0**0000... | ...00100000... | 0 | 1 | 0 |
| | 10 | ...00111**0**000... | ...00111000... | 1 | 0 | 0 |
| Total | | | | | 4 | 1 |

Question mark denotes location of untyped locus in the study sample. Bold 0 or 1 denote alleles in the reference sample at the untyped locus. Underlined loci indicate regions of sharing between the subject's haplotype and the reference haplotypes about the untyped locus. An estimate of the frequency of the '1' allele at the untyped locus among individuals with the subject's haplotype using the haplotype sharing weighted estimator (1/4) differs considerably from the frequency among all reference haplotypes (4/10).

To develop a method that captures the chromosome-scale information used in the hidden Markov chain approaches but is easily calculated, we relax the exact matching between sample and reference panel haplotypes and instead characterize the likelihood that haplotype $h_1$ contains the '1' allele at the untyped locus by

$$\frac{\sum_j w\left(h_1, h_j^r\right) I\left[a_j^r = 1\right]}{\sum_j w\left(h_1, h_j^r\right)}, \tag{8}$$

where $w(h_1, h_j^r)$ is a weight function that characterizes how 'similar' $h_1$ is to $h_j^r$. Here, we take $w(h, h_j^r) = I[h_j^r \in \{h_k^r \mid \mathbb{S}(h, h_k^r) \geq \mathbb{S}(h, h_{k'}^r)\ \forall\ k'\}]$, where $\mathbb{S}(h_1, h_2)$ is the number of loci that are identical by state moving up- and downstream from the untyped loci (corresponding to the information length criterion commonly used in haplotype sharing analyses [14]). Thus, $w(h_1, h_j^r)$ selects the set of reference panel haplotypes that have the largest information length in common with $h_1$, so that (8) corresponds to the proportion of '1' alleles among this set. Table 2 illustrates how haplotype sharing can be used to estimate allele frequencies at an untyped locus.

By analogy with equation (6), we characterize $X$ by an additive model by writing

$$X(h) = \frac{\sum_j w\left(h_1, h_j^r\right) I\left[a_j^r = 1\right]}{\sum_j w\left(h_1, h_j^r\right)} + \frac{\sum_j w\left(h_2, h_j^r\right) I\left[a_j^r = 1\right]}{\sum_j w\left(h_2, h_j^r\right)}. \tag{9}$$

Other models, including recessive and dominant models, can similarly be constructed using (8).

When the observed sample data consist of unphased multilocus genotypes, computing $m_i$ requires that $X(h)$ be summed over the conditional distribution of diplotypes, given the observed genotype data $\varphi(h \mid g, e)$. Thus, we are forced to specify a 'working model' for $\varphi(h \mid g, e)$. However, since the efficient score is valid even when this working model is misspecified, we are able to choose estimators that are computationally simple, secure in the knowledge that misspecification will not affect the validity of the test. One approach, which we utilize in the simulation experiment below, is to estimate $\varphi(h \mid g, e)$ by computing full-chromosome diplotypes for each study participant using a fast phasing program (for example, *ent* [15]), and then letting $\varphi(h \mid g, e)$ be the degenerate distribution that puts all mass on the imputed diplotype.

## Simulations

To evaluate the performance of our efficient score approach for testing untyped SNPs and to compare our simple haplotype sharing-based imputation to imputation using hidden Markov models, we used simulated datasets that were previously created by Li et al. [3] to examine the performance of their hidden Markov approach MACH. Using the coalescent simulation program of Schaffner et al. [16], Li et al. [3] generated 10,000 chromosomes for a series of 100 different 1-Mb regions with linkage disequilibrium patterns similar to those of the HapMap CEU sample. For each 1-Mb region, they chose 120 chromosomes to serve as the phased haplotypes of the reference panel. Within these reference panel haplotypes, they thinned the set of SNPs to have similar density to the Phase II HapMap sample, resulting in a mean (median) number of 932 (952) SNPs per 1-Mb region. From this thinned set of SNPs, they selected a panel of 100 tagSNPs that captured approximately 78% of variants with MAF >5% in the reference panel.

Li et al. [3] used the remaining chromosomes in a given 1-Mb region to form test datasets comprised of 500 cases and 500 controls. For each region, the authors generated 20 null datasets and then 25 alternative datasets having a randomly selected SNP that

serves as the disease susceptibility locus. The susceptibility locus need not occur in either the test sample or the reference panel haplotypes. The MAF of the susceptibility variant was varied between 5 values (2.5, 5, 10, 20, 50%), and the genotype-relative risk of the variant was tuned so that the power of a single-SNP test of association at the true susceptibility locus was 80% at a type-I error rate of 0.0005 on average.

Using these simulated datasets, we investigated the power of our haplotype sharing-based efficient score approach for testing untyped SNPs using the same strategy that Li et al. [3] used to evaluate the power of MACH. In the analysis of SNPs across a 1-Mb region for a specific dataset, we tested each tagSNP in the region using a single-marker allelic test and tested each untyped SNP using our efficient score approach. From these analyses, we then identified the most significant SNP (either a tagSNP or an untyped SNP) and recorded the corresponding minimum p value. We used the minimum p values obtained for the 2,000 null datasets (20 datasets per each 1-Mb region) to establish an empirical p value threshold that led to an overall type-I error rate of 5% when applied to the most significant result in each region. Using this empirical p value threshold, we evaluated the power under each specific alternative design (categorized by the MAF of the disease susceptibility allele) as the proportion of datasets when the minimum p value across a region was smaller than the empirical p value threshold. By assessing the significance of the minimum p value across a region in this manner, we inherently adjust for the testing of multiple SNPs within each region.

We compared the power of our efficient score approach (using 4 different approaches to estimating $m_i$) to the power of simply analyzing the tagSNPs alone (ignoring the untyped SNPs for the purpose of analysis). Two approaches to estimating $m_i$ involved selecting tagSNPs for each particular untyped SNP. The first of these applied the approach of Nicolae [5] implemented in the *tuna_db* component of the TUNA software package (see Web Resources) using both the suggested options (maximum number of 4 tagSNPs per untyped SNP across a 400-kb window) as well as other options (increasing the maximum number of tagSNPs per untyped SNP by 7–10 and varying the window size between 400 kb and 1 Mb). The second approach to choosing tagSNPs was a simple ad hoc flanking strategy where the 2–4 closest SNPs on each side of the untyped SNP were used to tag the untyped variant. For both of these selection approaches, $m_i$ was computed using equation (6). The third approach was to estimate $m_i$ using equation (9) and the sharing weight detailed in the section above. Finally, we estimated $m_i$ using the estimated allelic count output by MACH. To facilitate analyses, we used the same counts previously computed and applied by Li et al. [3]. These estimates were generated from MACH assuming 100 rounds of the Markov sampler and using all available haplotypes to update a subject's genotypes [Yun Li, personal communication].

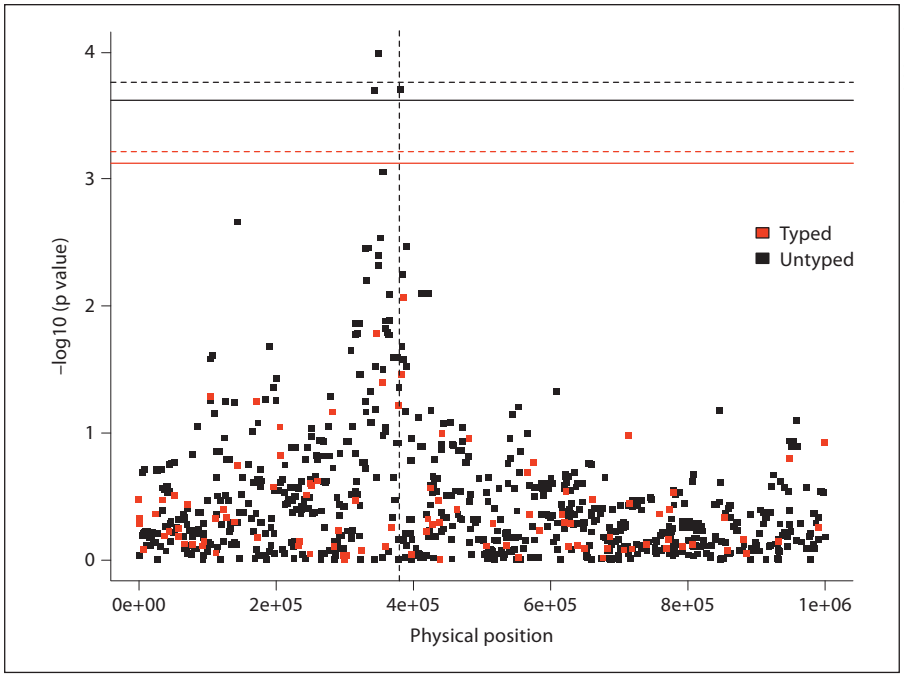
In addition to examining power, we were also interested in the ability of these four approaches to localize the disease locus. For each method, we used the physical position of the most significant SNP in each region as an estimate of the location of the disease locus. We then computed the mean squared error (average of squared differences between the estimated location and the true physical position of the disease variant) using all datasets, and also restricting to those datasets for which the method being investigated showed a significant result.

## Results

We present the power results of the simulation experiment in table 3. As can be seen from this table, common disease variants (MAF $\geq$10%) are detected with comparable power by all methods. In particular, there does not appear to be any power advantage in the untyped variant analysis over the analysis of tagSNPs only. However, when the disease variant is rare, the untyped variant approach can give a boost in power. This is true for both the analysis with MACH-computed estimates of $m_i$ and the analysis with sharing-computed estimates of $m_i$, which show a 60 and 38% increase in power, respectively, over a tagSNP-only analysis when the disease MAF is 2.5%. We note that the power of the efficient score with MACH-based estimates of $m_i$ has virtually the same power as that reported by Li et al. [3]. The flanking SNP- and TUNA-based approaches have at most a limited power advantage over the tagSNP-only analysis.

The localization results of the simulation experiment are presented in table 4. As can be seen from this table, all approaches show improved localization as the MAF at the disease locus increases. Interestingly, the TUNA-based untyped variant approach led to a poorer localization of the disease locus than the analysis using only the tagSNPs. However, the other untyped variant approaches compared favorably with the tagSNP-only approach. From table 4, we see that, when considering all simulations, calculating $m_i$ using MACH performed best; our sharing approach is roughly equivalent to flanking SNPs, with the TUNA-based method having the worst performance with regard to localization. A similar pattern is seen when restricting to localization of the causal locus in the presence of a significant finding (results not shown).

To help make these results more concrete, we illustrate the analysis (fig. 1) of one of the simulated datasets where the untyped analysis showed a significant signal while the tagSNP-only analysis did not. The significance threshold used in the simulation study involved the averaging over a large number of datasets that would, of course, not be available for a de novo analysis of this dataset. Thus, we established a region-wide significance threshold via permutation, using only the dataset at hand. To establish the threshold, we estimated the permutation distribution of the minimum p value across the region by randomly permuting the case-control status and capturing the smallest p value across the region. By repeating this procedure a large number of times (we used 10,000), we were able to precisely estimate the 5th percentile of the permutation distribution which is used to establish a sig-




Color version available online

**Fig. 1.** Example showing analysis of a simulated dataset. Black dots represent efficient score-based testing (using sharing method to estimate $m_i$) of untyped loci. Red dots represent tests of observed genotypes. Black and red solid horizontal lines represent significance thresholds used in simulation study for untyped- and typed-only analyses, respectively. Black and red dashed horizontal lines represent significance thresholds for untyped- and typed-only analyses, respectively, established by permutation applied to this dataset. Dashed vertical line denotes location of true disease locus.

**Table 3.** Power to detect disease variant with various minor allele frequencies (MAF)

| MAF at disease locus | TagSNPs only | Efficient score with $m_i$ computed by | | | |
|---|---|---|---|---|---|
| | | MACH | TUNA-based | flanking SNPs | max sharing |
| 2.5% | 0.266 | 0.432 | 0.224 | 0.326 | 0.384 |
| 5% | 0.476 | 0.580 | 0.432 | 0.538 | 0.544 |
| 10% | 0.672 | 0.730 | 0.638 | 0.698 | 0.686 |
| 20% | 0.762 | 0.768 | 0.728 | 0.766 | 0.740 |
| 50% | 0.824 | 0.806 | 0.754 | 0.812 | 0.792 |

**Table 4.** Mean square error for localizing disease variant for various minor allele frequencies (MAF)

| MAF at disease locus | TagSNPs only | Efficient score with $m_i$ computed by | | | |
|---|---|---|---|---|---|
| | | MACH | TUNA-based | flanking SNPs | max sharing |
| 2.5% | 73,783 | 58,569 | 80,633 | 69,540 | 72,248 |
| 5% | 59,736 | 40,157 | 59,650 | 47,836 | 44,992 |
| 10% | 36,477 | 27,799 | 42,535 | 31,146 | 29,728 |
| 20% | 22,134 | 14,927 | 25,372 | 19,417 | 21,140 |
| 50% | 16,589 | 15,746 | 20,695 | 17,811 | 20,320 |

Units are $(kb)^2$.

nificance threshold. These thresholds (for both the typed- and untyped-based analyses) are represented by dashed horizontal lines in figure 1. As can be seen in figure 1, testing at untyped loci offers an advantage in this dataset and identifies a significant SNP within close proximity (approx. 32 kB) of the true disease locus. Interestingly, even though the closest SNP was only approximately 600 bases away from the true disease locus, the typed-only analysis failed to detect this locus.

## Discussion

The analysis of untyped SNPs in a GWAS may have many practical benefits for gene mapping of complex diseases, including increased power for detecting rare variants and the ability to compare SNP results across different studies, utilizing different genotyping platforms. In this paper, we propose a simple and robust efficient score approach for testing untyped SNPs in a case-control GWAS. Additionally, we present haplotype sharing-based approaches that outperform tagSNP approaches but are easily calculated. We demonstrated that our approach has power for analyzing untyped SNPs nearly comparable to complex hidden Markov models, while, at the same time, yielding statistics that are computationally much simpler. As an example, the efficient score with

sharing-based estimates of $m_i$ required approximately 30 s to analyze one of the simulated 1-Mb datasets on a 32-bit laptop computer with a dual core 1.66-GHz processor and 2 GB of ram running Windows XP. This time includes the time required to phase the genotype data with the haplotyper *ent*. In contrast, the same analysis using MACH required approximately 30 min on a 64-bit desktop computer running Linux with 2 dual core 2.39-GHz AMD opteron processors and 6 GB of ram.

Two important properties of the efficient score framework allow us to take computational shortcuts without affecting the validity of the test. First, the efficient score is robust to misspecification of the imputation model, and second, the variance of the efficient score function can be estimated empirically without any additional contribution arising from the estimation of parameters used in the imputation model. These properties allow the use of computationally expedient imputation methods that may not, for example, correspond to maximum likelihood estimators. We have outlined several imputation methods and have shown that the sharing-based estimator may represent a reasonable compromise between computational speed and statistical power.

The efficient score function has additional computational advantages. First, covariates are easily included. TUNA does not allow for covariates, and it is unclear how they could be included in the TUNA framework. Further, the part of the efficient score function that accounts for the covariate effects need only be calculated once per genome and then can be used at each locus. Second, permutation testing for genome-wide significance is particularly simple using the efficient score test. If there are no covariates, then the case/control status can be permuted without recalculating $m_i$. The efficient score leads to simple Monte Carlo tests even in the presence of covariates. For stratified data, the case/control status can be permuted in each stratum, again without recalculating $m_i$. Finally, with continuous covariates, when inference is based on the efficient score function, the Monte Carlo procedure of Lin [17] can be used. The ease of permutation testing using the efficient score function contrasts strongly with HAPSTAT and SNPMstat, which require that haplotype frequency parameters be recalculated for each permutation dataset.

We also considered which method gives a better estimate of the location of an association, where the estimated corresponds to the most significantly associated SNP. When considering all situations, using $m_i$ values calculated by MACH has the best performance. However, when considering only situations where a significant association was detected, we found that our sharing approach estimates the location of the true causal SNP better than all other approaches.

We feel there are at least two reasons to develop computationally simple methods for association analysis of untyped variants. First, the computational burden of assigning genotypes at untyped loci increases both with the number of individuals and the number of SNPs in the reference panel. Computationally intensive methods are already near the limit of what can be computed in a reasonable amount of time for reference panels with the sample size and density corresponding to the first-generation HapMap, requiring cpu weeks even when running on Unix or Linux clusters. For example, we estimate that analyzing just the null datasets described here would take 35–40 days on a Linux workstation. Larger, more dense reference panels such as the 1000 Genomes Project [18] will further increase the computational burden. Second, it is not uncommon to want to run several analyses (e.g. with and without certain SNPs, or perhaps using different reference panels when trying to impute untyped variants in structured populations like Hispanic-Americans). This common modeling step, prohibitive for Markov chain Monte Carlo approaches, is easily accomplished using our method; for example, the null datasets described here were completely analyzed in 10 h. For a multiethnic population such as Hispanic-Americans, the ability to try several reference panels for imputation could result in further improvement in the power of our method compared to more computationally intensive approaches.


## Acknowledgments

We would like to thank Gonçalo Abecasis and Yun Li for the use of the simulated data. A.S.A. acknowledges support from the NIH through NIMH grant R01 MH084680. S.L.B. acknowledges support from the NIH through NIMH grant U01 MH079470. F.D. acknowledges support from UK MRC grant U.1052.00.012.00001.01. M.P.E. acknowledges support from the NIH through NHGRI grant R01 HG003618.


## Web Resources

Untyped, a software package implementing the untyped variant analysis described here, is available for download at http://www.duke.edu/~asallen/Software.html. The URLs for other web resources presented herein are as follows: International HapMap Project, http://www.hapmap.org; ent, http://dna.engr.uconn.edu/software/ent, and MACH, http://www.sph.umich.edu/csg/abecasis/mach/.

## References


1 McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn J: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356–369.

2 Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39:906–913.

3 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: Markov model for rapid haplotyping and genotype imputation in genome wide studies. Submitted.

4 Nicolae DL: Quantifying the amount of missing information in genetic association studies. Genet Epidemiol 2006;30:703–717.

5 Nicolae DL: Testing untyped alleles (TUNA) – applications to genome-wide association studies. Genet Epidemiol 2006;30:718–727.

6 Zaitlen N, Kang HM, Eskin E, Halperin E: Leveraging the HapMap correlation structure in association studies. Am J Hum Genet 2007;80:683–691.

7 Servin B, Stephens M: Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 2007;3:e114.

8 Lin DY, Hu Y, Huang BE: Simple and efficient analysis of disease association with missing genotype data. Am J Hum Genet 2008;82:444–452.

9 Browning BL, Browning SR: A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 2009;84:210–223.

10 International HapMap Consortium: A haplotype map of the human genome. Nature 2005;437:1299–1320.

11 Epstein MP, Satten GA: Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 2003;73:1316–1329.

12 Li N, Stephens M: Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 2003;165:2213–2233.

13 Allen AS, Satten GA: Robust estimation and testing of haplotype effects in casecontrol studies. Genet Epidemiol 2008;32:29–40.

14 Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F: Search for multifactorial disease susceptibility genes in founder populations. Ann Hum Genet 2000;64:255–265.

15 Gusev A, Mandoiu II, Pasaniuc B: Highly scalable genotype phasing by entropy minimization. IEEE/ACM Trans Comput Biol Bioinform 2008;5:252–261.

16 Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: Calibrating a coalescent simulation of human genome variation. Genome Res 2005;15:1576–1583.

17 Lin DY: An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 2005;21:781–787.

18 Siva N: 1000 Genomes project. Nat Biotechnol 2008;26:256.