Am. J. Hum. Genet. 75:850–861, 2004

Finding Haplotype Tagging SNPs by Use of Principal Components Analysis

Zhen Lin and Russ B. Altman

Department of Genetics, School of Medicine, Stanford University, Stanford, CA

The immense volume and rapid growth of human genomic data, especially single nucleotide polymorphisms (SNPs), present special challenges for both biomedical researchers and automatic algorithms. One such challenge is to select an optimal subset of SNPs, commonly referred as "haplotype tagging SNPs" (htSNPs), to capture most of the haplotype diversity of each haplotype block or gene-specific region. This information-reduction process facilitates cost-effective genotyping and, subsequently, genotype-phenotype association studies. It also has implications for assessing the risk of identifying research subjects on the basis of SNP information deposited in public domain databases. We have investigated methods for selecting htSNPs by use of principal components analysis (PCA). These methods first identify eigenSNPs and then map them to actual SNPs. We evaluated two mapping strategies, greedy discard and varimax rotation, by assessing the ability of the selected htSNPs to reconstruct genotypes of non-htSNPs. We also compared these methods with two other htSNP finders, one of which is PCA based. We applied these methods to three experimental data sets and found that the PCA-based methods tend to select the smallest set of htSNPs to achieve a 90% reconstruction precision.

Introduction

SNPs are the most common type of genetic variation in the human genome (Collins et al. 1997). They are stable sequence variations, in which typically two alternate nucleotide bases are observed at one position across some populations. Even though most SNPs have no observable phenotype, they occur sufficiently frequently within the genome to offer an opportunity for tracking both disease genes and population histories. Thus, they are effective markers for genomic research.

SNPs are ubiquitous in the human genome, but the precise number depends on an arbitrary cutoff for defining a polymorphism (in contrast to a mutation). For instance, ~10 million SNPs are found if they are defined as having minor-allele frequencies >1% (Kruglyak and Nickerson 2001). The dbSNP database catalogs SNPs and currently contains ~9 million distinct entries (dbSNP build 121) (Wheeler et al. 2003). With the recent development of sequencing technology, the availability of human SNP data is expanding quickly. To cope with this large amount of information, the biomedical informatics community is investigating methods to organize, summarize, and analyze SNPs (Klein et al. 2001; Ritchie et al. 2001; Hahn et al. 2003; Wheeler et al. 2003).

Sets of nearby SNPs on the same chromosome are inherited in blocks. Each block contains only a few common haplotypes (Stephens et al. 2001*a;* Gabriel et al. 2002), which are specific arrangements of alleles. Because pairs of SNPs, especially those within a block or in close proximity, are often correlated, the number of SNPs required to capture the haplotype diversity of each block can therefore be largely reduced. The correlation or association between SNPs is referred to as "linkage disequilibrium" (LD). The minimal informative subset of SNPs associated with the limited number of haplotypes in a block are often referred to as "haplotype tagging SNPs" (htSNPs) (Daly et al. 2001; Johnson et al. 2001; Patil et al. 2001; Gabriel et al. 2002).

It is sometimes possible to select htSNPs by eye for small genomic regions (Daly et al. 2001; Johnson et al. 2001). Because of the size and growth of SNP data, manual compilations of htSNPs suffer from problems of completeness and timeliness. Automatic methods for selecting htSNPs would therefore be very useful.

A number of methods for identifying htSNPs are based on searches; they systematically evaluate subsets of SNPs and use a metric to evaluate each set of candidate htSNPs. One method chooses htSNPs to discriminate all nonsingleton sequences uniquely (Patil et al. 2001). A more complex method measures the number of differences in all pairwise comparisons between sequences as the haplotype diversity and chooses htSNPs explaining the greatest amount of the total haplotype diversity (Daly et al. 2001). Several other researchers use *entropy* (*H*) to quantify haplotype diversity (Judson et al. 2002; Avi-Itzhak et al. 2003; Hampe et al. 2003) and select the set of htSNPs with the minimal size but the maximal information content retained. Others

Received May 13, 2004; accepted for publication August 31, 2004; electronically published September 23, 2004.

Address for correspondence and reprints: Dr. Russ B. Altman, Department of Genetics, Stanford University, 300 Pasteur Drive L-301, Stanford, CA 94305-5120. E-mail: russ.altman@stanford.edu

0002-9297/2004/7505-0011$15.00

choose the set of htSNPs minimizing the squared correlation between the estimated and the true value of the number of copies of haplotypes carried by a subject or alleles carried at each SNP (Chapman et al. 2003; Stram et al. 2003).

Bafna et al. (2003) proposed a completely different approach, measuring how well one SNP predicts another and how well a set of SNPs predicts a single SNP and another set of SNPs. They showed the htSNP finding problem to be *NP-complete* by use of this measure. (In computational complexity theory, *NP-complete* problems are the most difficult problems among the set of *NP* [nondeterministic polynomial-time] problems.) Along the same lines, the program BEST is based on *set theory* and recursively searches the minimal set of SNPs from which the maximum number of the other SNPs in the data set can be derived with a given function (Sebastiani et al. 2003). Meanwhile, other approaches avoid exhaustive searches by clustering SNPs on the basis of similarities of pairwise LD measures and then selecting one htSNP per cluster (Wu et al. 2003; Carlson et al. 2004).

Current computational challenges related to htSNPs fall into two areas: (1) developing efficient algorithms and (2) developing mechanisms for comparing the performance of these algorithms. *Principal components analysis* (PCA) is an efficient method of finding independent basis vectors that has been successfully applied in other application areas. We have investigated PCA-based approaches to finding htSNPs and have evaluated the selections of htSNPs according to their ability to recover genotypes of non-htSNPs. The performance in predicting genotypes of non-htSNPs on the basis of genotypes of htSNPs allows us to compare different algorithms.

PCA is a dimensionality-reduction technique for multivariate analysis (Mardia et al. 1979; Dunteman 1989; Duda et al. 2001). It is widely used in signal processing and feature recognition. It has also recently been applied to bioinformatics (Raychaudhuri et al. 2000; Troyanskaya et al. 2001), including the definition of htSNPs (Meng et al. 2003; Horne and Camp 2004). Geometrically, PCA is a procedure to rotate data such that maximum variability is projected onto orthogonal axes according to a minimum-square-error criterion. Essentially, a set of correlated variables is transformed into a substantially smaller set of uncorrelated variables (principal components) that represent most of the information in original data, where principal components are linear combinations of the original set of variables. Thus, PCA-based approaches allow us to consider all data of a gene-specific region or a haplotype block simultaneously and efficiently.

One of the challenges in using PCA on SNP data is that the principal components that are defined do not correspond to actual genotypes, since genotypes are discrete variables. Thus, we need ways to map the principal components optimally to measurable genotypes. In this article, we present the results of our (1) investigation of two different htSNP-mapping methods for PCA, (2) sensitivity analyses of these two mapping methods, and (3) comparison of PCA-based approaches with two other htSNP-finding methods through use of three publicly accessible SNP data sets. The source code of the algorithms that we implemented is available online at the authors' Web site.

Methods

We divide the htSNP-finding problem into two steps: (1) using PCA to locate principal SNP components from the sample—that is, locating eigenSNPs—and (2) mapping principal SNP components to characteristic SNPs—that is, mapping htSNPs. The second step, in particular, is approximate and can be performed in a number of different ways, two of which we will discuss.

Locating EigenSNPs

Our initial analysis focuses on the principal components of SNPs. We first create a correlation matrix to measure how each SNP contributes information to the data set. We then summarize the information compactly with principal SNP components—that is, eigenSNPs.

We start with a matrix of SNP data, $\mathbf{X}$, in which each row i from $1 \ldots n$ corresponds to a different chromosome, and each column j from $1 \ldots p$ corresponds to a SNP from the chromosome. Each X_{ij} contains two values, 1 for the rare allele and 0 for the common allele. We compute the p-dimensional mean vector μ and $p \times p$ correlation matrix $\mathbf{R}$ for the full data. Next, we compute eigenvectors $\mathbf{E} = \{e_1, \ldots, e_p\}$ and eigenvalues $\mathbf{\Lambda} = \{\lambda_1, \ldots, \lambda_p\}$, by solving the equation

$$\mathbf{R}e_j = \lambda_j e_j,\ j = 1, \ldots, p\ ,$$

and sort these eigenvectors according to decreasing eigenvalues. We choose the k eigenvectors with the largest eigenvalues, which define k eigenSNPs. Each eigenSNP s is a weighted sum of SNPs:

$$s_i = \sum_{j=1}^{p} e_{ij} x_j,\ i = 1, \ldots, k\ ,$$

where the weights are the coefficients of the eigenvector.

PCA can be based on either a covariance matrix or a correlation matrix. For this problem, a correlation matrix is attractive because it directly relates to commonly used LD measures; PCA of correlation matrices also produces less bias when variables are measured in different

units and when there are large differences in variances among variables (Mardia et al. 1979; Dunteman 1989).

Each eigenvalue λ_i is the amount of variance explained by the eigenvector e_i. The sum of variances of eigenSNPs is equal to the sum of variances of original SNPs. Therefore, the proportion of variance in the original p SNPs that k eigenSNPs accounts for is

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j} .$$

Several *rule of thumb* criteria for excluding principal components exist (Mardia et al. 1979). We choose to include just enough k components to explain 90% of the total variance. This choice can be changed depending on details of the application. Choosing a higher percentage cutoff preserves more information about SNPs. The appropriate cutoff depends on the properties of individual data sets and genotyping constraints.

Mapping htSNPs

Since eigenSNPs are mathematical abstractions and do not directly correspond to measurable quantities, they are hard to use for making decisions about which SNPs to genotype. Therefore, we map each of them to the *nearest* SNP in the original data set as an htSNP, where we define *near* as having a substantial large coefficient between the eigenSNP and the real SNP on the genome. If a SNP has a substantial large coefficient with one eigenSNP, we assume that it contributes greatly to this eigenSNP. We compare two mapping methods: (1) the greedy-discard method and (2) the varimax-rotation method.

Greedy-discard method.—We regard an eigenSNP with a small eigenvalue as being of less importance, and, consequently, the SNP that is highly correlated to it should be of less overall importance or redundant. Thus, we decompose this mapping method into two substeps. First, from the eigenvector with the smallest eigenvalue to the one with the $(p - k)$th smallest eigenvalue, we reject the SNP (1) that has the largest coefficient (absolute value) in the $(p - k)$th eigenSNP and (2) that has not been previously discarded. Then, in the reverse order, we map the retained k eigenSNPs to remaining k SNPs in the original data as k htSNPs.

Varimax-rotation method.—A SNP may have similar coefficients with several eigenSNPs, which creates difficulty in determining the eigenSNP to which this SNP most contributes. The *varimax-rotation* criterion maximizes the sum of the variances of the squared coefficients within each eigenvector. Geometrically, pairs of axes defined by PCA are rotated iteratively so that each SNP has either a high or low coefficient for a rotated eigenSNP, with perhaps some SNPs left over. The rotated solution spans the same geometric space as the original solution and explains the same amount of variance in the data as the original solution (Mardia et al. 1979; Dunteman 1989). Thus, this rotation simplifies the SNP-eigenSNP relationship and eases the interpretation, in which optimal case each SNP has a high coefficient with only one rotated eigenSNP (see further description of the varimax-rotation procedure at the authors' Web site).

We apply the varimax-rotation method on eigenvectors $\mathbf{E}$, a method detailed by Meng et al. (2003). Instead of applying the procedure to each small subset of SNPs, as they did, we apply it to all of the SNPs in the data set. We find an orthogonal transformation $\mathbf{T}$ so that rotated eigenvectors $\mathbf{E}^r = \mathbf{ET}$ will confine the influence of each SNP to a particular eigenvector, where $\mathbf{E}^r = \{e_1^r, \ldots, e_p^r\}$.

For each SNP, we compute its average coefficient for all k eigenSNPs,

$$\Gamma_i = \frac{1}{k}\sum_{j=1}^{k} |e_{ij}^r|,\ i = 1, \ldots, n ,$$

and its average coefficient for the rest of the $(p - k)$ eigenSNPs,

$$\gamma_i = \frac{1}{p-k}\sum_{j=k+1}^{p} |e_{ij}^r|,\ i = 1, \ldots, n .$$

We compare these two average coefficients and select this SNP if it has a higher average coefficient for the k eigenSNPs ($\Gamma_i > \gamma_i$), which indicates that this SNP contributes most significantly to the k eigenSNPs.

Evaluation and Comparison

If the selected htSNPs represent most of the haplotype diversity of a genomic region, we should be able to recover genotypes of non-htSNPs, given genotypes of htSNPs. Therefore, we evaluated selected htSNPs on the basis of their performances in reconstructing genotypes of remaining non-htSNPs. We used a procedure called *cross-validation,* where, each time, we partition the data set into a training set and a validation set; we use the training set to identify htSNPs and the validation set to evaluate their abilities of recovering non-htSNP genotypes.

To reduce the computation time required for validation, we used *10-fold* cross-validation for the *IBD* data set, whereas for the others we used *leave-one-out* cross-validation. For *leave-one-out* cross-validation, we run the htSNP-finding algorithm while leaving out each chromosome successively (for a total number of runs equal to the number of chromosomes in the data set)

and then validate the results for the missing chromosome, as described in next paragraph. For *10-fold* cross-validation, we randomly partition data into 10 disjoint sets of equal size; then, we run the htSNP-finding algorithm 10 separate times, each time using the training set from which a different 10% of data has been left out to be the validation set.

For each chromosome in the validation set, we use htSNP genotypes to predict non-htSNPs genotypes. The prediction of a non-htSNP genotype depends on how well it correlates with each htSNP genotype in the training set. The htSNP genotype that has the greatest correlation with the non-htSNP genotype determines the non-htSNP genotype. For instance, if (1) SNP_X has genotypes A and G at allele frequencies 70% and 30%, respectively; (2) $htSNP_Y$ has genotypes C and T also at allele frequencies 70% and 30%, respectively; and (3) genotype A is highly correlated with genotype C; then $htSNP_Y$ can reliably predict SNP_X.

When multiple htSNPs have the same correlation coefficient with a non-htSNP but their predictions on the genotype of the non-htSNP are contradictory, we fill in the genotype of the non-htSNP with its common allele. For example, if another $htSNP_Z$ has genotypes G and A at allele frequencies 60% and 40%, respectively, SNP_X correlates equally to $htSNP_Y$ and $htSNP_Z$, but $htSNP_Y$ predicts that SNP_X has genotype A whereas $htSNP_Z$ predicts that it has genotype G. Then, we predict that SNP_X has genotype A, which has an allele frequency of 70%.

Let the *variance explained* be the amount of variance that a set of k htSNPs or eigenSNPs explains,

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j} ,$$

where k is the number of htSNPs, p is the number of non-htSNPs, and λ is the eigenvalue. With the predictions on non-htSNP genotypes based on htSNP genotypes, we calculate the precision at all possible variance explained cutoffs, generating both (number of htSNPs)/variance and (number of htSNPs)/precision curves. The precision is

$$\frac{\text{number of correctly predicted alleles}}{\text{all predictions}} ,$$

which indicates the number of errors produced. We count a SNP allele as correct if the predicted genotype matches the original genotype exactly.

We defined the baseline accuracy by randomly selecting any SNPs and labeling them as htSNPs. We compared the performance of the two mapping methods based on PCA and the strategy of randomly selecting htSNPs. In addition, we compared the results with the ones produced by the application *htstep* (Johnson et al. 2001; software available from D. Clayton's Web site) and the ones based on PCA with a sliding window approach (Meng et al. 2003).

htstep method.—In this approach, the haplotype diversity is defined as the total number of differences recorded in all pairwise comparisons between chromosomes (Johnson et al. 2001). For a SNP j, the diversity is

$$D_j = \sum_{i=1}^{n} \sum_{i'=1}^{n} (x_{ij} - x_{i'j})^2 ,$$

where the difference of $(x_{ij} - x_{i'j})$ is 0 if chromosome i and i' at position j are the same and is ± 1 if they differ. The total haplotype diversity (diversity explained) is the summation over all SNPs:

$$D = \sum_{i=1}^{n} \sum_{i'=1}^{n} (x_i - x_{i'})^T (x_i - x_{i'}) = \sum_{j=1}^{p} D_j .$$

The htSNPs are a set of SNPs that explain the greatest amount of the total haplotype diversity. Once a set of k candidate htSNPs classify n chromosomes into $G = 2^k$ (at most) groups, all chromosomes within one group have all the same alleles at each htSNP position. For each group, a within-group diversity is computed. The total residual diversity (R) is the sum of the within-group diversities:

$$R = \sum_{g=1}^{G} D_g .$$

The proportion of diversity explained by the set of htSNPs is

$$P = 1 - \frac{R}{D} .$$

Therefore, htSNPs are the ones minimizing the within-group heterozygosity.

We used the original source code provided by the Clayton group (Johnson et al. 2001). Because it can be time consuming to identify the optimal htSNP set by an exhaustive search from the possible $2^p - 1$ candidate sets, we used its recommended program *htstep* with both *step-up* and *step-down* default parameters to speed up computations. With the same cross-validation strategy, we calculated the precision at every possible diversity explained cutoff, generating both (number of htSNPs)/diversity and (number of htSNPs)/precision curves.

Sliding window method.—Under the assumption that the haplotype information is important in the context of nearby markers, Meng et al. (2003) proposed a sliding window approach. They partitioned SNPs into small sets along a sliding window and recursively used PCA to filter out candidate htSNPs from each set until reaching a predetermined convergence criterion. Although this is a PCA-based method, it focuses on small windows, which leads to substantially different performance in comparison with three global PCA methods.

Given a set of p SNPs arranged according to a map order, a sliding window with a relatively small window size is moved along the map. We then apply the above-described PCA with varimax-rotation mapping procedure in each window. The selection or nonselection of a SNP is recorded in a vector $\mathbf{W}_i = \{w_{ij}; j = 1, \ldots, L\}$, where L is the window size; $w_{ij} = 1$ indicates that the jth SNP is not selected in the ith window, and $w_{ij} = 0$ otherwise. Each SNP's relative redundancy (rr) is computed by averaging its corresponding w_{ij} over all the windows in which it appears, and it is recorded in another vector $\mathbf{RR} = \{\mathrm{rr}_j; j = 1, \ldots, p\}$. A SNP is rejected as a candidate htSNP if its rr is above a predetermined threshold. Repeat the sliding window procedure on the remaining SNPs until it converges. The convergence is achieved when the difference between the number of SNPs before and after selection represents $\leqslant 5\%$ of the SNPs before selection (Meng et al. 2003).

We implemented this approach with a sliding window size of 5 and an rr threshold of 75%, as recommended in the original publication. In cross-validation experiments, we calculated the precision at every possible variance-explained cutoff, generating both (number of htSNPs)/variance and (number of htSNPs)/precision curves.

Experimental Data Sets

We used three published experimental SNP data sets for the evaluation. If a data set contained unphased diploid data, we preprocessed it by inferring haploid data by use of the PHASE 2.0.2 haplotype inference program (Stephens et al. 2001*b;* Stephens and Donnelly 2003; PHASE Web site). Though PCA can select htSNPs directly from unphased diploid data (Meng et al. 2003), not all htSNP finders can do so. We thus used phased haploid data in our analysis, to directly compare performances of different algorithms.

The first data set is the ACE (angiotensin I converting enzyme) data set from a study of 78 SNPs typed on *DCP1* (Rieder et al. 1999; Nickerson Group Web site) in 11 individuals. *DCP1* is a gene encoding ACE and stretches over a genomic region of length 24 kb. We used a total of 52 biallelic nonsingleton SNPs from 22 phased chromosomes for our analysis.

The second data set is the *ABCB1* data set from University of California–San Francisco membrane transporter gene study (Kroetz et al. 2003; Pharmacogenetics and Pharmacogenomics Knowledge Base Web site). *ABCB1* is a gene responsible for P-glycoprotein and extends over 74 kb of the genome sequence. The original diploid data contain 48 SNPs typed in 247 individuals. After applying PHASE 2.0.2, we used 27 biallelic nonsingletons from 484 phased chromosomes for our analysis.

The third data set is the *IBD* 5q31 data set from an inflammatory bowel disease study of father-mother-child trios (Daly et al. 2001; IBD5 Data Release Page). The original diploid data contain 103 SNPs covering 500 kb of the genome, typed in 387 subjects. We again used PHASE 2.0.2 to infer haplotypes, resulting in 103 biallelic nonsingletons from 774 phased chromosomes for our analysis.

We implemented the PCA greedy-discard method, the PCA varimax-rotation method, and the PCA sliding window method in Python 2.3 (Lutz and Ascher 1999) and Matlab 6.5 (The MathWorks), with Numerical Python 23.1 (Ascher et al. 2001; Numerical Python Web site) and Pymat 1.02 (Sterian 1999; Pymat Web site) libraries. We used Intercooled Stata 8.0 (StataCorp LP) to run *htstep*. All experiments were conducted on Solaris machines with 900-Mhz processors and each with at least 2 Gb of memory.

Results

We first plotted eigenvalues of the eigenSNPs from the ACE, *ABCB1,* and *IBD* data sets as Scree graphs (see fig. 1), which illustrate the set of eigenSNPs identified in each data set and their associated variance. For each method, we then identified htSNPs explaining 90% of the data variance (see details of the identified htSNPs at the authors' Web site). For instance, the PCA greedy-discard method identified 5 htSNPs in the ACE data set, 18 in the *ABCB1* data set, and 15 in the *IBD* data set. Figure 2 shows, for each method, the detailed lists of htSNPs explaining 90% of variance in the ACE data set.

We cross-validated each method on all three data sets and compared their performance. We calculated the prediction precision at every variance/diversity explained cutoff and plotted them against the corresponding number of htSNPs (see figs. 3–5). Methods that primarily reside in the lower right corner of a (number of htSNPs)/variance plot (see figs. 3*A,* 4*A,* and 5*A*) use the smallest set of htSNPs to capture the most variance. Likewise, methods that primarily reside in the upper left corner of a (number of htSNPs)/precision plot (see figs. 3*B,* 4*B,* and 5*B*) use the smallest set of htSNPs to achieve the highest prediction precision.

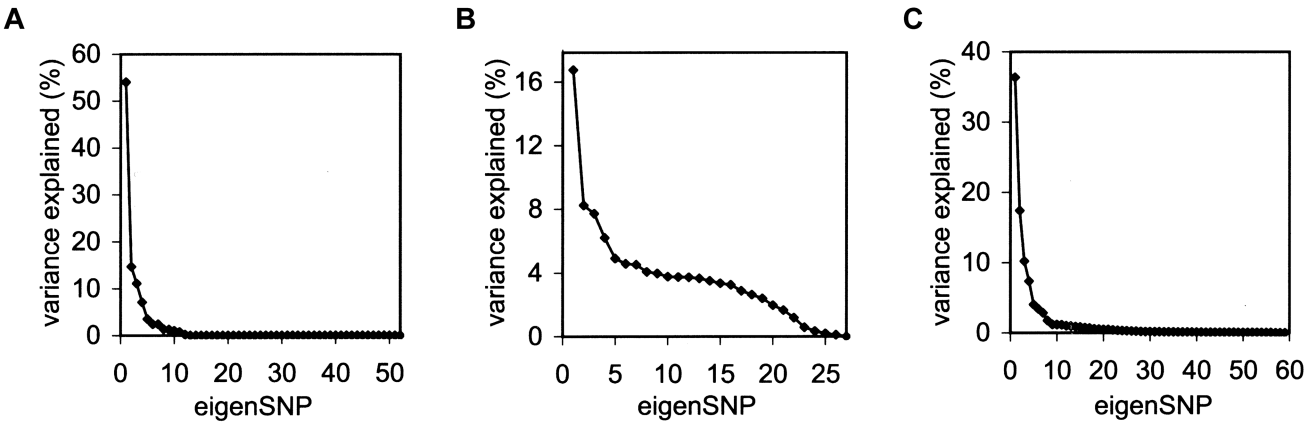

Figure 1 Eigenvalues from eigenSNPs for each experimental data set. The horizontal axes are eigenSNPs with decreasing eigenvalues. The vertical axes are the percentages of the variance in the data that each eigenSNP explains. Graphs A, B, and C show results from the ACE, *ABCB1,* and *IBD* data sets, respectively. In the ACE and *IBD* data sets, most of the variance was contained in just a small proportion of eigenSNPs. The *ABCB1* data set was an exception; the variance was split among many eigenSNPs with substantial weights. These Scree graphs are helpful in estimating the number of eigenSNPs/htSNPs required to capture data variance or haplotype diversity.

Discussion

With the enormous number of SNPs currently in the dbSNP and the rate at which we are identifying new ones, defining htSNPs that capture most of the variance in the data requires automated methods. Since LD is common among SNPs, high-throughput genotyping efforts and computer programs analyzing phenotype-genotype correlations will have to deal with correlated SNPs and can benefit from more-compact representations of SNP information.

The PCA-based approaches transform the data into orthogonal spaces and map a set of principal components back to the SNPs that contribute the most. PCA allows us to consider pairwise correlation coefficients of all SNPs at once. The PCA-based approaches efficiently found htSNPs that have better performance in recovering genotypes of non-htSNPs than those identified by the non-PCA-based approach. Because the computational complexity of PCA on a set of SNPs of size N ranges from $O(N^2)$ to $O(N^3)$, these methods have computational advantages over other methods that require exhaustive searches on all candidate htSNPs.

We evaluated the performance of all methods against a random strategy, to assess the overall value of htSNP selection for each data set, since different levels of LD may make some htSNP selection processes easier than others. For example, a random strategy performed reasonably well for ACE and *IBD* data sets but not for the *ABCB1* data set, because there is less LD in that gene. This result is explained by distributions of principal components in Scree graphs (see fig. 1). These graphs show that we can summarize most of the data variance with a much smaller subset of eigenSNPs in the ACE and *IBD* data sets. Thus, a random selection of SNPs has a high likelihood of capturing components in these data. Conversely, for the *ABCB1* data, the variance was split among many eigenSNPs with substantial weights; therefore, a random selection of htSNPs performed poorly.

We cross-validated all methods on three experimental data sets, which contain different numbers of SNPs and chromosomes from three regions of the human genome. Though the plots of results are scaled differently, they show very similar global patterns and trends across all data sets. As shown in the (number of htSNPs)/variance plots (see figs. 3*A,* 4*A,* and 5*A*), a larger set of htSNPs was generally required to explain more variance in the data. The PCA greedy-discard and varimax-rotation methods tend to choose the smallest set of htSNPs beyond the 80% variance explained cutoff. We note that the number of htSNPs selected by the sliding window approach fluctuated with the amount of variance explained. This was a result of using a small window size, so that the choices of htSNPs were sometimes trapped in local minima.

As expected (and shown figs. 3*B,* 4*B,* and 5*B*), larger sets of htSNPs had better performance in predicting genotypes of htSNPs on the basis of genotypes of non-htSNPs. When the performance of each method was compared with the random strategy of selecting the same number of htSNPs, the PCA greedy-discard and varimax-rotation methods achieved higher prediction precisions across all data sets. However, neither *htstep* nor the sliding window approach guaranteed a better performance than a random strategy.

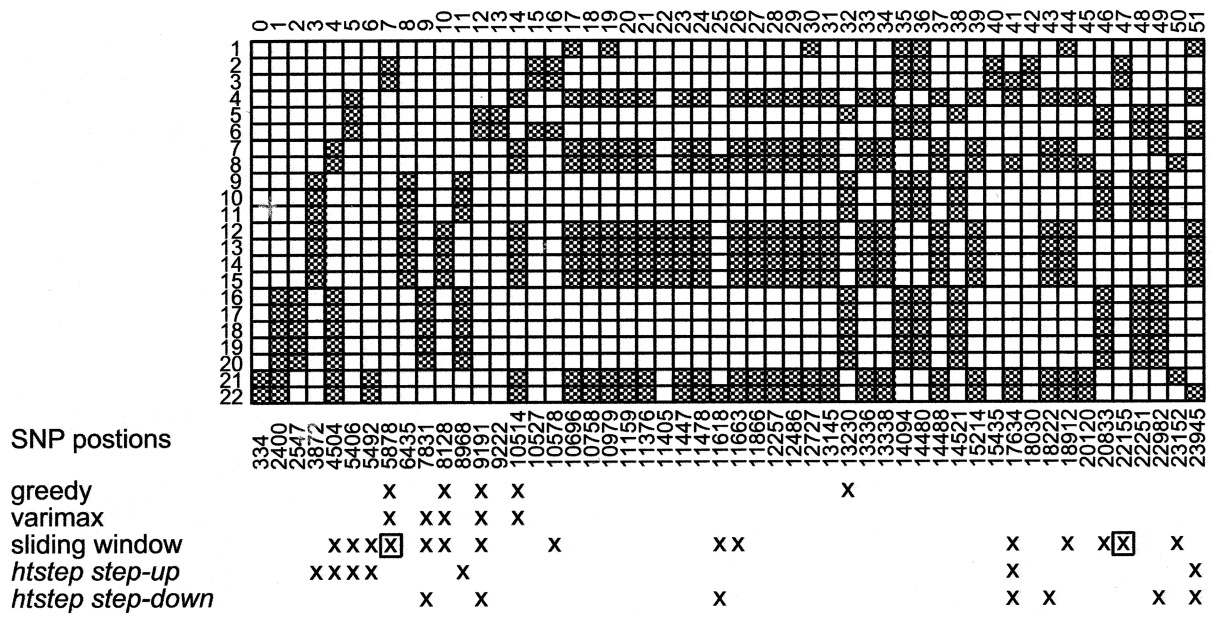

Figure 2 htSNPs explaining 90% of the variance in the ACE data set. Sequence data were genotyped in 11 individuals. A checked cell indicates a rare allele SNP; an empty cell indicates a common allele. The htSNPs from most methods successfully captured the one perfectly correlated region in data, except for the *htstep* method with the *step-up* parameter. The sliding window approach acquired the largest number of htSNPs from this region. As highlighted in black squares, two perfectly correlated SNPs that were physically distant, beyond the window size but still within 30 kb, were both selected as htSNPs under the sliding window approach.

When we chose the smallest set of htSNPs that achieved a prediction precision of 90%, the PCA greedy-discard and varimax-rotation methods outperformed others. The sliding window approach sometimes could reach a higher prediction precision, but at the expense of overselecting markers. Furthermore, when we chose only one htSNP to represent each data set, the one defined by the PCA greedy-discard and varimax-rotation methods always had the best performances. This is consistent with our expectation that, since eigenSNPs capture the true variance of data, the htSNPs retained by PCA are better choices.

When evaluating PCA-based methods to identify htSNPs, we also tried an iterative PCA method, for which we recomputed the eigenvectors each time and removed the one with the smallest eigenvalues. Unfortunately, this strategy performed poorly under all conditions.

Lastly, we examined the actual htSNPs selected by each method to explain 90% of the total variance in the ACE data set. The htSNPs from most methods successfully captured the one perfectly correlated region in the data (see the middle region in fig. 2), except for the *htstep* method with the *step-up* parameter. Failure to identify this region affected the overall prediction precision when reconstructing genotypes of non-htSNPs.

The sliding window approach selected the largest number of htSNPs from the ACE data set. Because of the small window size, this approach retained redundant information unnecessarily. For instance, as is highlighted in figure 2, two perfectly correlated SNPs that were physically distant, beyond the five-SNP window but still within 30 kb, were both selected as htSNPs.

Our experiments suggest a strategy of using either greedy discard or varimax rotation as PCA mapping methods to reach a good prediction precision with a minimal set of htSNPs. With these htSNPs as a guide, one can further prioritize the final reduced set of SNPs for high-throughput genotyping by including potentially functional SNPs, such as the ones within the 3′ and 5′ UTRs of genes; genetic regulators (enhancers, silencers); exons that are coding, noncoding, or partially coding; and alternative splicing.

In addition to the implication for cost-effective genotyping for correlation studies, defining htSNPs is important in other contexts—for example, in assessing the risk of disclosing the identities and health-related information of individuals from public-domain SNP databases. htSNP identification is particularly relevant to the identification of the independent information content in the genome, which may be a crucial factor in determining whether anonymous electronic access to individual-specific SNP data sets is appropriate (Lin et al. 2004). The htSNPs defined using PCA-based approaches can be used as an indicator of the amount of the independent information content of a SNP data set.

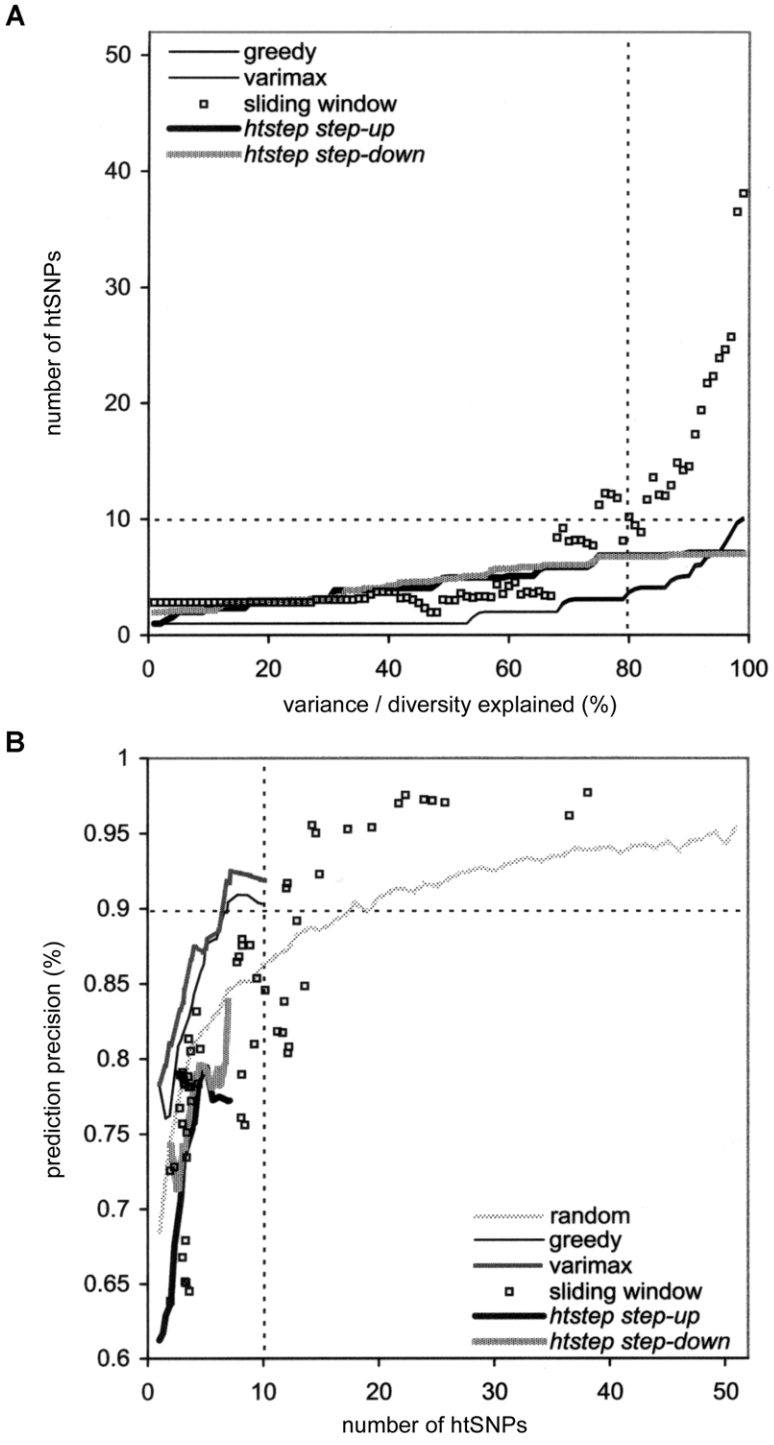

Figure 3 *Leave-one-out* cross-validation on the ACE data set. The (number of htSNPs)/variance plot (*A*) indicates the number of htSNPs selected from each method (Y-*axis*) to explain a proportion of the total variance (X-*axis*). The (number of htSNPs)/precision plot (*B*) indicates the accuracy of using the htSNP genotype to recover the non-htSNP genotype; the *Y*-axis indicates the accuracy and the *X*-axis indicates the number of htSNPs involved. Dotted horizontal and vertical lines within each graph provide a consistent check scale across different data sets. Methods that reside in the lower right corner of the (number of htSNPs)/variance plot use the smallest set of htSNPs to capture the most variance. Likewise, methods reside on the upper left corner of the (number of htSNPs)/precision plot use the smallest set of htSNPs to achieve the highest prediction precision. The (number of htSNPs)/variance plot shows that both the PCA greedy-discard method and the PCA varimax-rotation method select the smallest set of htSNPs to explain at least 80% of variance in the data. The (number of htSNPs)/precision plot shows that both the PCA greedy-discard method and the PCA varimax-rotation method select the smallest set of htSNPs to achieve a 90% precision in recovering genotypes.

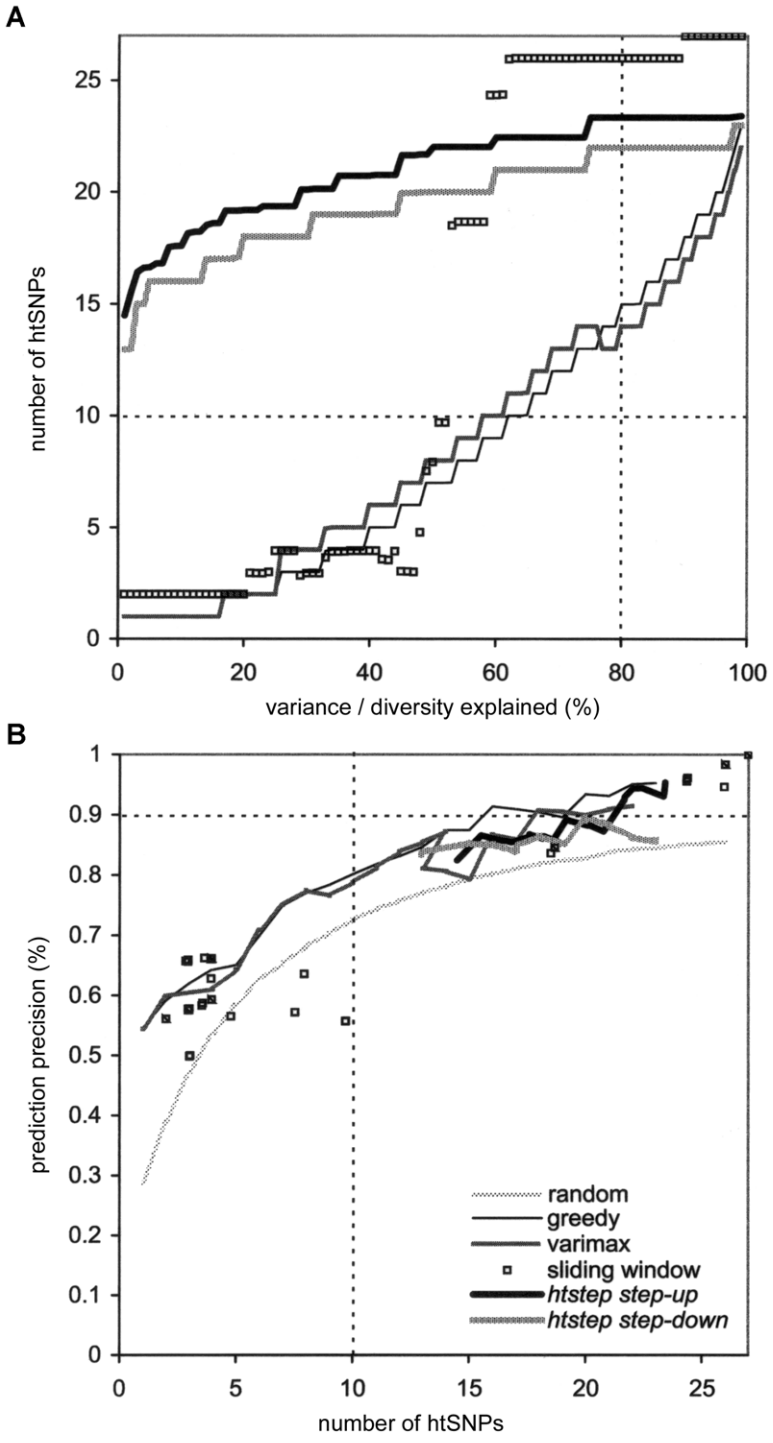

Figure 4 *Leave-one-out* cross-validation on the *ABCB1* data set. Axes in each graph are as explained in figure 3. The (number of htSNPs)/variance plot shows that both the PCA greedy-discard method and the PCA varimax-rotation method select the smallest set of htSNPs to explain at least 80% of variance in the data. The (number of htSNPs)/precision plot shows that the PCA greedy-discard method selects the smallest set of htSNPs to achieve a 90% precision in recovering genotypes.

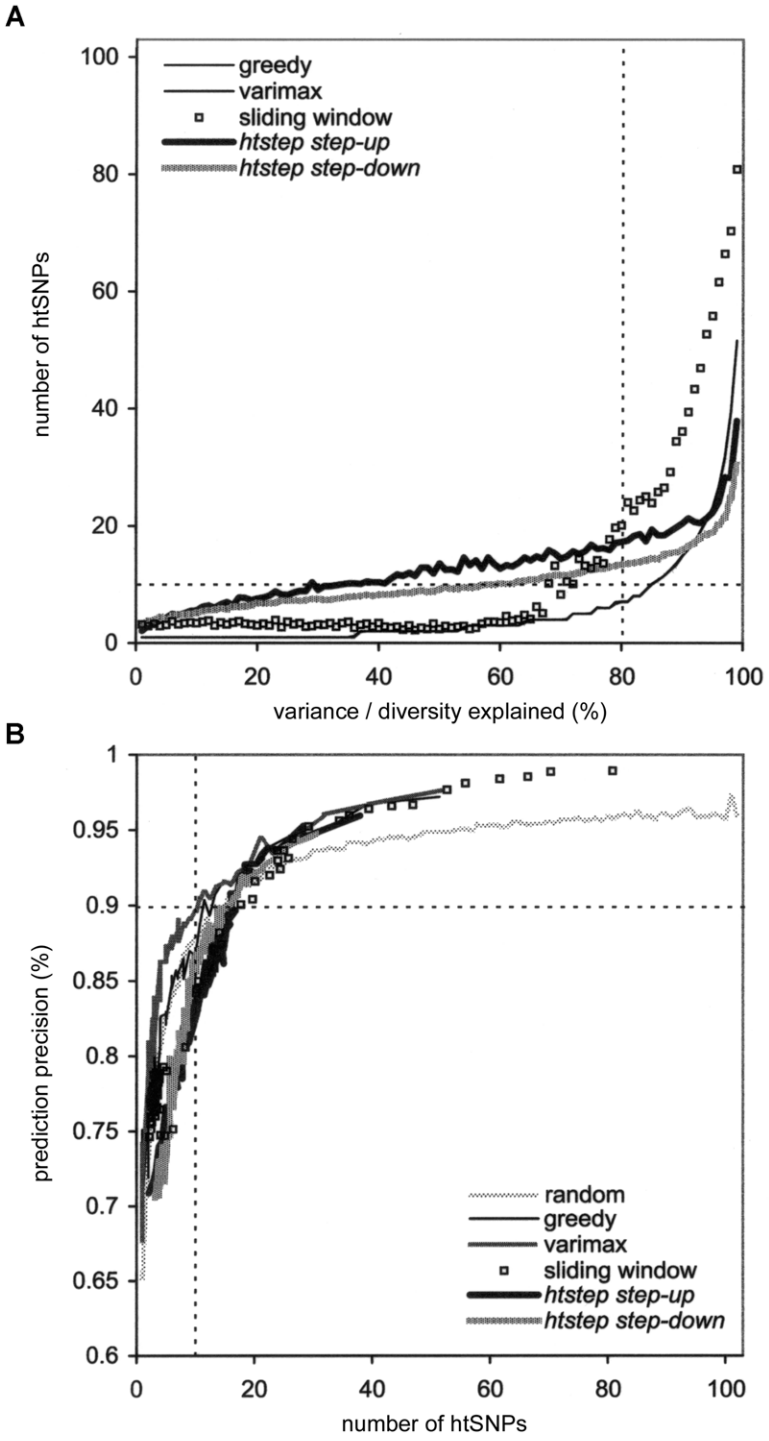

Figure 5 *10-fold* cross-validation on the IBD data set. Axes in each graph are as explained in figure 3. The plots show that both the PCA greedy-discard method and the PCA varimax-rotation method perform better than other methods. They use the smallest set of htSNPs to explain 80% of the data variance and to reach a 90% precision in recovering genotypes.

PCA-based methods can handle only chromosomes without any missing values. For unphased genotype data with missing values, we assume that other computational programs, such as SNPHAP (Johnson et al. 2001), PHASE, and HAPLOTYPER (Qin et al. 2002), can produce suitable inputs for PCA-based approaches. However, a recent report showed that inferring haplotypes can lead to a substantial loss of information (Morris et al. 2004). Because PCA-based methods can be applied to unphased genotype data directly, they may be more generally useful than methods that are limited to inferred haplotypes only. Since more-recent applications that are directly applicable to unphased genotype data have become available (Chapman et al. 2003), we are further evaluating PCA-based methods against these applications.

Because of the difficulty in using PCA with zero values, we have limited our calculation to nonsingleton SNPs in our cross-validations. In addition, we have employed a simple marginal-probability strategy for predicting genotypes when the correlated SNPs are contradictory; other methods may be able to improve performance in recovering haplotypes.

The two PCA-based methods, greedy discard and varimax rotation, used much less time to find htSNPs than the PCA sliding window approach or *htstep* in our experiments. However, we anticipate that these PCA-based methods may not scale as the size of the genomic sequence fragment increases. To deal with this challenge, we could investigate methods of integrating PCA-based methods with other computation approaches, such as the dynamic-programming algorithm (Zhang et al. 2004) and the minimal-description-length principle (Anderson and Novembre 2003).

Acknowledgments

This work is supported, in part, by National Institutes of Health/National Library of Medicine Biomedical Informatics training grant LM007033 (to Z.L.) and National Institutes of Health/National Institute of General Medical Sciences Pharmacogenetics Research Network and Database grant U01-GM61374 (to R.B.A.). We thank Jeff Chang, Brian Naughton, Mike Liang, Bernie Daigle, Art Owen, Teri Klein, and reviewers for their comments and support.

Electronic-Database Information

The URLs for data presented herein are as follows:

Authors' Web site, http://htSNP.stanford.edu/ (for supplementary material and source code)
dbSNP Home Page, http://www.ncbi.nlm.nih.gov/SNP/
D. Clayton's Web site, http://www-gene.cimr.cam.ac.uk/clayton/software/ (for *htstep*)
IBD5 Data Release Page, http://www.broad.mit.edu/humgen/IBD5/
Nickerson Group, Department of Genome Sciences, University of Washington, http://droog.mbt.washington.edu/ (for the ACE data set)
Numerical Python, http://www.pfdubois.com/numpy/
Pharmacogenetics and Pharmacogenomics Knowledge Base, http://pharmgkb.stanford.edu (for the *ABCB1* data set)
PHASE—Haplotype Reconstruction Software, https://depts.washington.edu/ventures/clickthru/ReleaseAgreement.php?raf=PHASEV2
Pymat—Python to MATLAB Interface, http://claymore.engineer.gvsu.edu/~steriana/Python/pymat.html

References

Anderson EC, Novembre J (2003) Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet 73:336–354
Ascher D, Dubois PF, Hinsen K, Hugunin J, Oliphant T (2001) Numerical python. Lawrence Livermore National Laboratory, Livermore, CA
Avi-Itzhak HI, Su X, De La Vega FM (2003) Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity. Pac Symp Biocomput 2003:466–477
Bafna V, Halldorsson BV, Schwartz R, Clark AG, Istrail S (2003) Haplotypes and informative SNP selection algorithms: don't block out information. Paper presented at Annual Conference on Research in Computational Molecular Biology, Berlin, April 10–14
Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120
Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31
Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581
Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232
Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley, New York
Dunteman GH (1989) Principal components analysis. Sage Publications, Newbury Park
Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229
Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19:376–382
Hampe J, Schreiber S, Krawczak M (2003) Entropy-based SNP

selection for genetic association studies. Hum Genet 114: 36–43

Horne BD, Camp NJ (2004) Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. Genet Epidemiol 26:11–21

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237

Judson R, Salisbury B, Schneider J, Windemuth A, Stephens JC (2002) How many SNPs does a genome-wide haplotype map require? Pharmacogenomics 3:379–391

Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, Lin Z, Liu Y, Liu S, Oliver DE, Rubin DL, Shafa F, Stuart JM, Altman RB (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project: Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J 1:167–170

Kroetz DL, Pauli-Magnus C, Hodges LM, Huang CC, Kawamoto M, Johns SJ, Stryke D, Ferrin TE, DeYoung J, Taylor T, Carlson EJ, Herskowitz I, Giacomini KM, Clark AG (2003) Sequence diversity and haplotype structure in the human ABCB1 (MDR1, multidrug resistance transporter) gene. Pharmacogenetics 13:481–494

Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27:234–236

Lin Z, Owen AB, Altman RB (2004) Genomic research and human subject privacy. Science 305:183

Lutz M, Ascher D (1999) Learning Python. O'Reilly, Sebastopol, CA

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press, New York

Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG (2003) Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet 73:115–130

Morris AP, Whittaker JC, Balding DJ (2004) Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. Am J Hum Genet 74:945–953

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723

Qin ZS, Niu T, Liu JS (2002) Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet 71:1242–1247

Raychaudhuri S, Stuart JM, Altman RB (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 2000:455–466

Rieder MJ, Taylor SL, Clark AG, Nickerson DA (1999) Sequence variation in the human angiotensin converting enzyme. Nat Genet 22:59–62

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147

Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF (2003) Minimal haplotype tagging. Proc Natl Acad Sci USA 100:9900–9905

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, et al (2001*a*) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293: 489–493

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Stephens M, Smith NJ, Donnelly P (2001*b*) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Sterian A (1999) PyMat—an interface between Python and MATLAB. Grand Valley State University, Allendale, MN

Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC (2003) Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. Hum Hered 55:27–36

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L (2003) Database resources of the National Center for Biotechnology. Nucleic Acids Res 31:28–33

Wu X, Luke A, Rieder M, Lee K, Toth EJ, Nickerson D, Zhu X, Kan D, Cooper RS (2003) An association study of angiotensinogen polymorphisms with serum level and hypertension in an African-American population. J Hypertens 21: 1847–1852

Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F (2004) Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. Genome Res 14:908–916