GENOME RESEARCH

Intra- and interpopulation genotype reconstruction from tagging SNPs

Peristera Paschou, Michael W. Mahoney, Asif Javed, Judith R. Kidd, Andrew J. Pakstis, Sheng Gu, Kenneth K. Kidd and Petros Drineas

Genome Res. 2007 17: 96-107; originally published online Dec 6, 2006;
Access the most recent version at doi:10.1101/gr.5741407

Intra- and interpopulation genotype reconstruction from tagging SNPs

Peristera Paschou,[1,4,6] Michael W. Mahoney,[2,5] Asif Javed,[3] Judith R. Kidd,[1] Andrew J. Pakstis,[1] Sheng Gu,[1] Kenneth K. Kidd,[1] and Petros Drineas[3]

[1]*Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06511, USA;* [2]*Department of Mathematics, Yale University, New Haven, Connecticut 06511, USA;* [3]*Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York 12180, USA*

The optimal method to be used for tSNP selection, the applicability of a reference LD map to unassayed populations, and the scalability of these methods to genome-wide analysis, all remain subjects of debate. We propose novel, scalable matrix algorithms that address these issues and we evaluate them on genotypic data from 38 populations and four genomic regions (248 SNPs typed for ~2000 individuals). We also evaluate these algorithms on a second data set consisting of genotypes available from the HapMap database (1336 SNPs for four populations) over the same genomic regions. Furthermore, we test these methods in the setting of a real association study using a publicly available family data set. The algorithms we use for tSNP selection and unassayed SNP reconstruction do not require haplotype inference and they are, in principle, scalable even to genome-wide analysis. Moreover, they are greedy variants of recently developed matrix algorithms with provable performance guarantees. Using a small set of carefully selected tSNPs, we achieve very good reconstruction accuracy of "untyped" genotypes for most of the populations studied. Additionally, we demonstrate in a quantitative manner that the chosen tSNPs exhibit substantial transferability, both within and across different geographic regions. Finally, we show that reconstruction can be applied to retrieve significant SNP associations with disease, with important genotyping savings.

[Supplemental material is available online at www.genome.org.]

The recent common ancestry of the human species provides a tool for the identification of genes that are involved in the susceptibility to, or protection from, common disease. However, the implementation cost of exhaustive genetic association studies comparing all human genetic variation in a very large number of cases and controls remains prohibitive. On the other hand, it has become apparent that common genetic variants such as single nucleotide polymorphisms (SNPs) contain a lot of redundant information due to the linkage disequilibrium (LD) structure of the genome (Daly et al. 2001; Goldstein and Weale 2001; Jeffreys et al. 2001; Patil et al. 2001; Stumpf 2002). This observation suggests the possibility of identifying a small set of SNPs that capture the genetic information within a specified genomic region and enables the design of cost-efficient genetic association studies. Such SNPs are commonly designated as tagging SNPs or tSNPs.

This notion motivated the HapMap project, which in phase I has released a public database of 1,000,000 SNPs, typed in four populations from three geographic regions (Africa, Europe, and East Asia) (The International HapMap Consortium 2003, 2005). It has been suggested that the populations studied in the HapMap project will serve as reference populations that will guide the selection of tSNPs for the design of genetic association studies by investigators around the world. However, the extent to which tSNPs selected in one of the HapMap populations will be predictive of unassayed SNPs in individuals from an unstudied population is an important question that has only recently been addressed by a number of studies (Ke et al. 2004; Mueller et al. 2005; Ramirez-Soriano et al. 2005; De Bakker et al. 2006; Gonzalez-Neira et al. 2006; Magi et al. 2006; Montpetit et al. 2006; Willer et al. 2006).

At the same time, a large number of methods identifying an "optimal" set of tSNPs has recently been introduced in the literature (for review, see Halldorsson et al. 2004b). Early methods necessitate haplotype inference—which is, from a computational time viewpoint, prohibitive for whole-genome studies for a large number of individuals—or rely on definitions of haplotype block boundaries, namely regions of high association between SNPs. Such methods subsequently select tSNPs based on these blocks (Johnson et al. 2001; Patil et al. 2001; Gabriel et al. 2002; Wang et al. 2002; Zhang et al. 2002, 2005; Ke and Cardon 2003; Sebastiani et al. 2003; Stram et al. 2003). No consensus "block" definition has been reached thus far, and recent studies have demonstrated marked differences in the number and length of blocks generated by different methods (Ding et al. 2005; Zeggini et al. 2005). Finally, no formal metric has been agreed upon for the quantification of the coverage provided by existing approaches and the tSNP selection problem in general (Schwartz et al. 2003; Wall and Pritchard 2003a,b). In most recent studies in the literature, this is implemented by estimating the r^2 coefficient between tagging SNPs and tagged SNPs (Chapman et al. 2003; Weale et al. 2003; Carlson et al. 2004; De Bakker et al. 2005). Although a high r^2 relationship might be a good indicator that a genetic association study will be effective, it is not clear whether such a relationship is sufficient. On the other hand, if tSNPs can be used to

Present addresses: [4]**Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupoli 68100, Greece;** [5]**Yahoo Research Labs, Sunnyvale, California 94089, USA.**
[6]**Corresponding author.**
E-mail ppaschou@mbg.duth.gr; fax 30-25510-30613.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5741407.

17:96–107 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

accurately reconstruct unassayed genotypes (or haplotypes), then it will be possible to retrieve the information retained in the data set, including r^2 relationships. The reconstruction of genotypes based on preselected tSNPs has received considerably less attention, and there is currently a dearth in methods that can efficiently address the reconstruction problem in a quantitative manner (Evans et al. 2004; Halldorsson et al. 2004a; Lin and Altman 2004).

In this study, we define the tSNPs selection problem as a reconstruction problem. Within this framework, we study a sample of ~2000 individuals from 38 populations from around the world typed for four genomic regions (Yale data set). To test our methods on a denser marker map, HapMap data from the four corresponding genomic regions were also included in our study. The data may be viewed as a table (one for each genomic region and each population) consisting of ~2000 rows, one for each individual, and a number of columns, one for each SNP site. We use a simple linear algebraic algorithm to select columns (and thus tSNPs) from this table, and we characterize the extent to which major patterns of variation of the intrapopulation data are captured by a small number of tSNPs. Next, we test the accuracy of prediction of unknown SNPs within a single population using only the tSNPs by splitting our sample into training and test sets for each of the populations. Next, we investigate the transferability of tSNPs across populations in a quantitative manner by testing the feasibility of reconstructing unknown SNPs in a previously unstudied target population using tSNPs determined in an available reference population. Finally, we test the impact of these methods on an association study using a publicly available data set (Daly et al. 2001; Rioux et al. 2001). Our algorithms are greedy, heuristic variants of recently developed randomized algorithms for extracting structure from large matrices. These randomized algorithms have provably good computational-time performance, and they are, in principle, scalable to whole-genome data analysis. Our analysis of the worldwide SNP data with these novel algorithmic tools provides an initial characterization of (1) the feasibility of intrapopulation unassayed SNP reconstruction using tSNPs, and (2) the transferability of tSNP selection for the reconstruction of unassayed SNPs for populations within and between diverse geographic regions.

Results

Data sets and characterization of linear structure in the populations

We analyzed four different genomic regions, using data both from 38 populations from around the world (Yale data set) as well as the HapMap populations (HapMap data set). For our Yale data set, a total of 248 SNPs were genotyped on ~2000 unrelated individuals (Supplemental Fig. 1; Table 1). HapMap data from the four corresponding genomic regions were also included in our study (Table 1). This provided us with the opportunity to test our methods on a denser marker map. We noticed that many of the available HapMap SNPs were actually monomorphic in at least one population. Since our aim was genotype prediction, we excluded from the analysis of the HapMap data set SNPs that were fixed in any one of the HapMap populations in order to avoid distortion of the reported errors. (Prediction for a monomorphic site will always be accurate.) This reduced the data set substantially from a total of 2731 SNPs to 1336 for all four populations, Yoruba (YRI), Europeans (CEU), Chinese (CHB), and Japanese (JPT) (Table 1).

Prior to applying our algorithms, we converted the SNP genotype data for each population and region studied to numeric data in order to process them with linear algebraic methods. Since only genotypic and not haplotypic data were available, each entry in the original data is a pair of bases that may be assumed to be ordered alphabetically. The data are converted without any information loss to numeric matrices. The (i, j)-th entry of any of these matrices is set to -1, 0, or +1, depending on whether, respectively, the i-th individual is homozygous (for one

Table 1. Yale data set and HapMap data set

Yale data set			
Region	**Chromosome (absolute positions)**	**SNPs**	**Average density**
SORCS3	10 (106,599,890–107,020,771)	53	7.94 Kb
PAH	12 (101,652,738–101,854,293)	36	5.60 Kb
HOXB	17 (43,337,796–44,477,524)	96	11.87 Kb
17q25	17 (77,751,614–78,651,254)	63	14.28 Kb

HapMap data set				
Region	**Chromosome (absolute positions)**	**SNPs (avail.)**	**SNPs (used)**	**Average density**
SORCS3	10 (106,603,000–107,021,000)	734	307	1.36 Kb
PAH	12 (101,652,100–101,854,500)	224	88	2.3 Kb
HOXB	17 (43,337,100–44,500,000)	1097	571	2.03 Kb
17q25	17 (77,751,000–78,651,500)	764	370	2.43 Kb

(Avail) Includes only SNPs that have been typed in all four HapMap populations; (used) excludes SNPs that were fixed in any of the four HapMap populations.

allele arbitrarily chosen of the two alleles) in the *j*-th SNP site, heterozygous at that site, or homozygous (for the other allele) at that site. A careful implementation of our linear algebraic algorithms allows the existence of missing entries. However, for simplicity and clarity of presentation of our algorithmic techniques, we chose to report results on matrices with no missing data. In the Yale and HapMap data sets, a small number (≤5%) of genotypes were missing, and we filled them in using the technique described in Alter et al. (2000); see Methods and the Supplemental material for details.

Linear structure in a data set is equivalent to the fact that the columns (rows) of the matrix can be expressed as linear combinations of a small number of left (resp. right) singular vectors with a small loss in accuracy (Golub and VanLoan 1989). We shall call these vectors eigenSNPs (Lin and Altman 2004). Recent results in the computer science and applied mathematics literature (Frieze et al. 2004; Drineas and Mahoney 2005, 2007; Drineas et al. 2006a,b,c) demonstrate that instead of using left (right) singular vectors, which are linear superpositions of all the columns (rows) of the matrix, a small number of actual columns (rows) might be used without any significant loss in accuracy. Since we hope to identify a small number of tSNPs that efficiently describe most of the data and also rely on a small number of individuals to do so (i.e., the HapMap subjects), this is precisely the type of structure that we hope to identify.

For each of the populations and the regions studied we computed the Singular Value Decomposition in order to determine the number of left singular vectors (eigenSNPs) that were needed to capture 90% and 99% of the spectral variance of the SNP data matrix for that population; see Methods for details. Results for each of the four genomic regions targeting 90% of the population's spectral variance are presented in Table 2 for the HapMap data set and Figure 1 for our sample of worldwide populations. (See Supplemental Table 1 and Supplemental Fig. 2 for the respective results targeting 99% of the population's spectral variance.) These data demonstrate that there exists a substantial amount of linear structure within each of the studied populations and data sets. Analysis of the HapMap genotypes in all four regions shows, as expected, that the Yoruban sample requires the highest number of eigenSNPs to capture the data, followed by the European and East Asian samples (Chinese and Japanese). For example, for the *HOXB* region (571 SNPs spanning ~1 Mb), only 11 eigenSNPs are enough to capture 90% of the spectral variance in the Yoruba and as few as six eigenSNPs suffice for the Japanese. When targeting 99% of the spectral variance of each data set, the number of eigenSNPs needed to capture the structure of the data increases on average two to three times, but still remains quite low.

The Yale data set, including 38 worldwide populations, has considerable linear structure as well. Five of our 38 populations correspond to the HapMap populations (Yoruba, European Americans, Chinese from San Francisco and Taiwan, and Japanese). Interestingly, although the four genomic regions we studied were typed at a much lower density for the 38 populations, almost the same number of eigenSNPs is needed in each case for the HapMap and our own "HapMap corresponding" populations (Table 2; Supplemental Table 1). This seems to suggest that the fundamental structure of the studied regions is accurately captured by the SNPs assayed for the Yale data set. However, testing such hypotheses further is difficult, mainly due to the fact that there is very little overlap between the SNPs typed in the Yale and HapMap samples.

Table 2. Linear structure statistics targeting 90% of the spectral variance in HapMap populations and their corresponding populations in the Yale data set

	HapMap			Yale	
	EigenSNPs	ActualSNPs		EigenSNPs	ActualSNPs
SORCS3					
YRI	6	10	Yor	6	8
CEU	5	6	Eur	5	7
CHB	3	4	SFC	3	5
			TWC	3	5
JPT	4	5	Jap	3	4
PAH					
YRI	6	10	Yor	7	9
CEU	5	6	Eur	4	6
CHB	3	4	SFC	3	5
			TWC	3	4
JPT	3	5	Jap	3	4
HOXB					
YRI	11	17	Yor	10	14
CEU	8	12	Eur	9	13
CHB	7	9	SFC	6	9
			TWC	6	7
JPT	6	8	Jap	5	8
17q25					
YRI	15	21	Yor	13	18
CEU	12	17	Eur	12	17
CHB	9	12	SFC	10	13
			TWC	8	12
JPT	9	12	Jap	9	13

In general, the amount of linear structure, as measured by the (decreasing) number of left singular vectors (eigenSNPs) required to capture the spectral variance within a population, increases as we move out of Africa to Europe, East Asia, and finally, the Americas. This is more pronounced for the two longest regions that we studied, *HOXB* and 17q25. The African Americans appear to be the most diverse population for all of the regions studied, requiring the greatest number of eigenSNPs.

Selecting tSNPs from a single population

We demonstrated that the major axes of variation in the SNP data matrices for each population could be covered with a small number of left singular vectors or eigenSNPs, which are linear combinations of the actual SNPs. We now seek to identify within each population a set of nonredundant real SNPs (tSNPs) that can retain most of the information contained in the original data matrix. Toward that end, we use the tSNPsMultiPassGreedy Algorithm (see Methods), which selects tSNPs by performing multiple passes over the data. In a pass, the "most informative" SNP (in a linear algebraic projection sense) is selected, its contribution

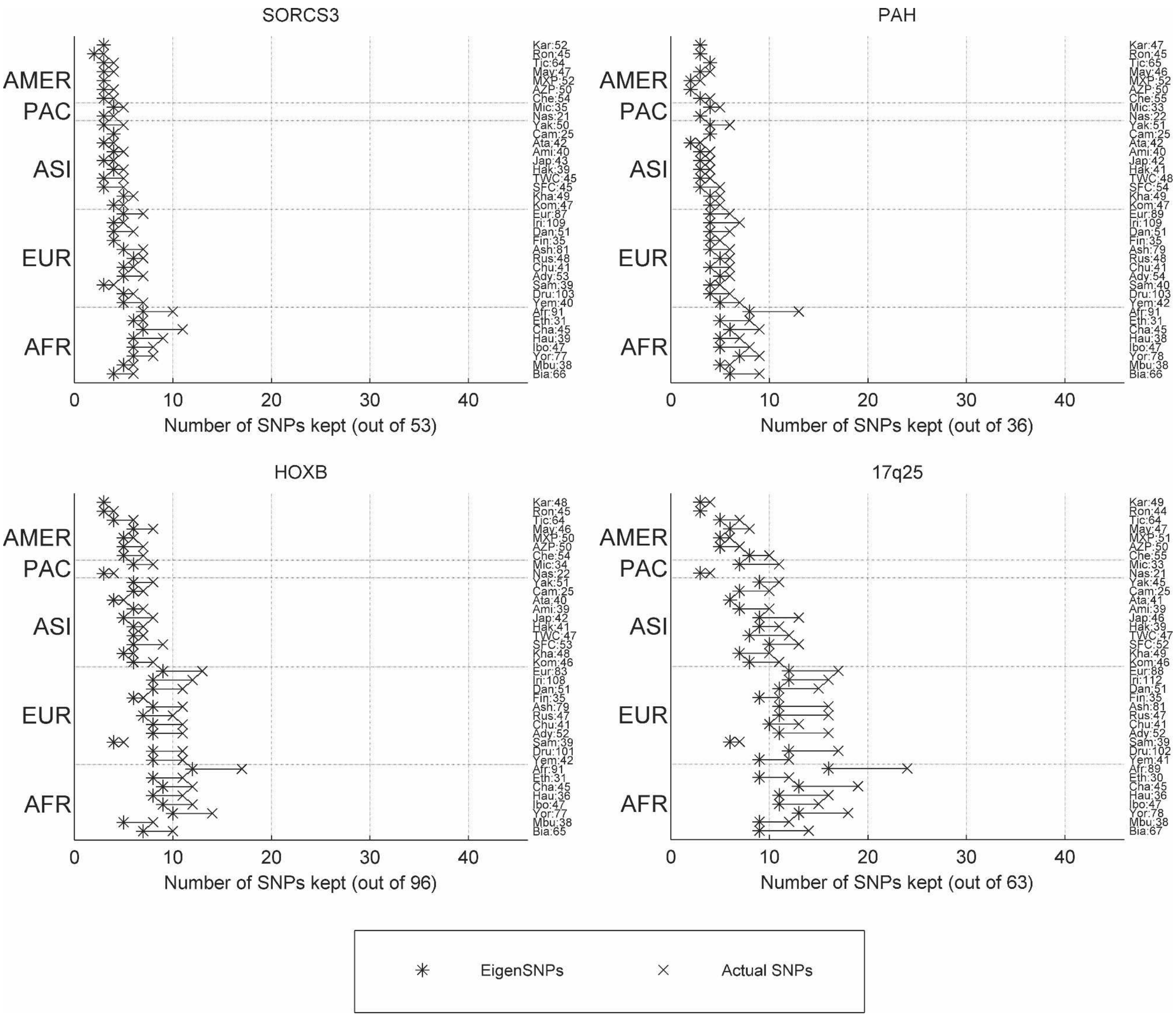

Figure 1. Number of eigenSNPs (computed with the SVD) and actual SNPs (computed with the tSNPsMultiPassGreedy algorithm) explaining 90% of each population's spectral variance. The number of individuals in each population sample is denoted next to the population's abbreviation. Populations are ordered (*bottom* to *top*) based on geographic regions (abbreviations used are shown in parentheses). **Africa**: Biaka (Bia), Mbuti (Mbu), Yoruba (Yor), Ibo (Ibo), Hausa (Hau), Chagga (Cha), Ethiopian Jews (Eth), African Americans (Afr), **South-west Asia and Europe**: Yemenites (Yem), Druze (Dru), Samaritans (Sam), Adygei (Ady), Chuvash (Chu), Russians (Rus), Ashkenazi Jews (Ash), Finns (Fin), Danes (Dan), Irish (Iri), European Americans (Eur), **Asia**: Komi (Kom), Khanty (Kha), Chinese Han-San Francisco (SFC), Chinese-Taiwan (TWC), Hakka (Hak), Japanese (Jap), Ami (Ami), Atayal (Ata), Cambodians (Cam), Yakut (Yak), **Pacific**: Nasioi (Nas), Micronesians (Mic), **America**: Cheyenne (Che), Pima-Arizona (AZP), Pima-Mexico (MXP), Maya (May), Ticuna (Tic), Rondonian Surui (Ron), Karitiana (Kar).

to data is extracted, and the procedure is repeated. This algorithm is a greedy variant of a provably accurate randomized algorithm (Drineas and Mahoney 2007).

Results are presented in Table 2 and Supplemental Table 1 for the HapMap data and Figure 1 and Supplemental Figure 2 for our 38 populations. In our linear algebraic framework, the number of eigenSNPs determined by SVD corresponds to a lower bound for the number of actual tSNPs that capture the same spectral variance in the data. We emphasize that this lower bound may not be achievable. Nevertheless, our results demonstrate that for most populations in both the HapMap and Yale data sets, a large fraction of their spectral variance can be covered by a number of actual SNPs that is not much larger than the number of eigenSNPs. We also found that the data sets that could be reconstructed from the selected tSNPs using standard least squares regression manage to retain the LD properties of each region as well as the allele frequencies for the common "tagged" SNPs (see Supplemental note and Supplemental Tables 2 and 3). We did notice, however, that in general rare SNPs appeared even less polymorphic in the reconstructed data set (data not shown). The difficulty in capturing rare variation may prove to be a general limitation of the tSNPs approach.

It is important to emphasize that at this stage tSNPs have been selected after having seen all of the genotypes for all individuals in each population. We have not performed any actual prediction in "unknown" samples, but simply established the fact that, in principle, redundancy does exist in the data, and thus it is possible to pick tSNPs that cover a certain percentage of the variance of the data.

Using tSNPs to reconstruct unassayed SNPs within a single population

We now address whether it is possible to reconstruct untyped genotypic information in individuals within a population given only a few tSNPs. For each of the populations and for each region, we split the data into training sets and test sets of three different sizes. The different training set sizes corresponded to 90%, 70%, and 50% of the population size, and the remainder of the population was used as a test set. (To get statistically significant results, 100 random splits were performed for each denomination and the results were averaged over all repetitions.) We then selected different numbers of tSNPs using the TSNPSMULTIPASSGREEDY algorithm on the training sets. We considered these tSNPs to be assayed (known) in the training sets and reconstructed the unassayed (unknown) SNPs on the test set using the RECONSTRUCTUNASSAYEDSNPS algorithm (see Methods for details).

The reconstruction error curve for the HapMap populations, using 10–60 tSNPs (in increments of 10) is shown in Figure 2. We would have expected the HapMap East Asian samples to be the easiest to predict. However, in all four regions, the highest reconstruction accuracy is achieved for the European sample. This may be due to the fact that the European sample consists of trios, while the Chinese and Japanese HapMap samples consist of unrelated individuals. As shown later in this section, in the Yale data set, where all populations consist of unrelated individuals, prediction is generally more accurate for the East Asian samples than the European samples. In most cases, the HapMap Yoruban sample is the most resistant to prediction. We discuss here our results (Fig. 2) using 70% of each population as the training set and trying to reconstruct the remaining 30%; see Supplemental Figure 3 for results using 90% of each population as the training set and trying to reconstruct the remaining 10% (results using 50% of the population as the training set were similar and are not shown). For the relatively short regions studied, *PAH* and *SORCS3*, the reconstruction error quickly drops below 10% using information from as few as 20 tSNPs of 88 and 307 SNPs, respectively. At around 40 tSNPs, in each case, the curve levels off and continues to drop with a slower rate. More tSNPs are needed for the 1-Mb regions we studied, *HOXB* and 17q25. The data set for the 17q25 region appears to be the least structured one. This may be due to the LD structure of the region or the lower density of the reference map used.

For clarity of presentation, we only show here reconstruction errors when keeping 10 or 20 SNPs for the 38 populations of

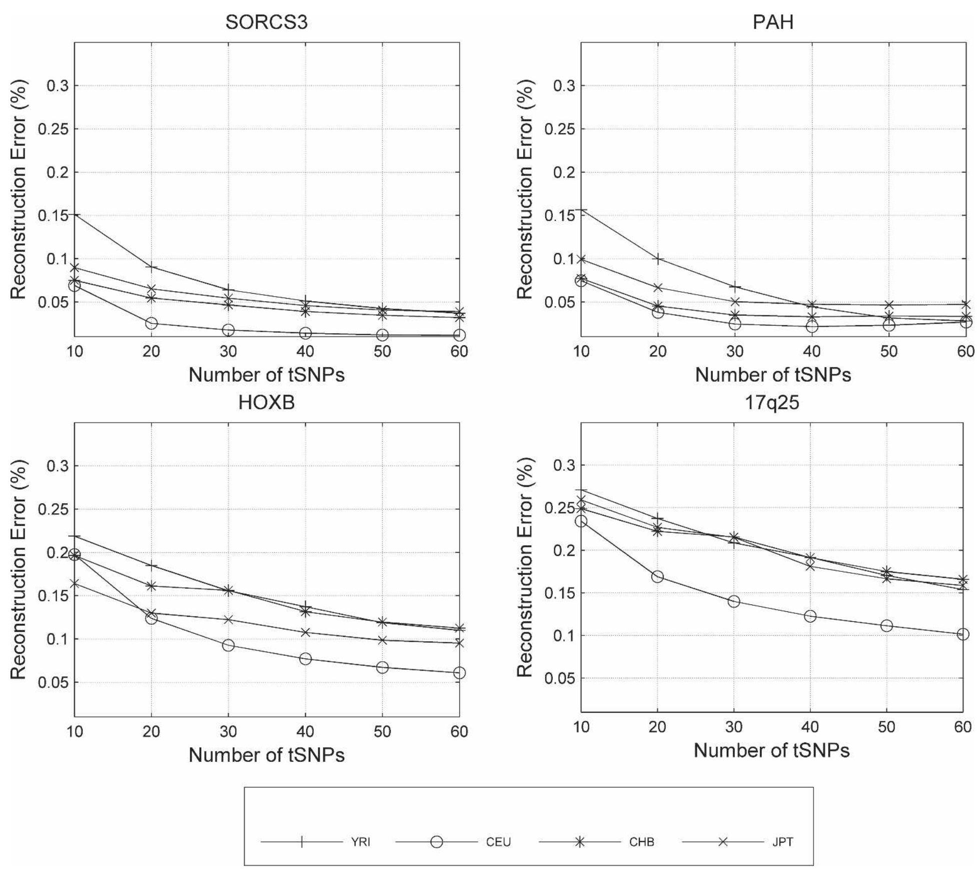

Figure 2. Intrapopulation reconstruction error (ratio of erroneously predicted entries over total number of predicted entries) for each of the four HapMap populations. The training set size is 70% of the total population size.

the Yale data set (Fig. 3, using 70% of each population as the training set, and Supplemental Fig. 4 using 90%). A gradient is again observed, with the smallest reconstruction errors achieved for the same number of tSNPs in the American Indian populations (<5%–10%) and the error increasing as we move through Asia and Europe back to Africa. This seems to follow the general pattern of migrations during human expansion out of Africa and the increasing amount of LD toward the Americas. In general, we achieve higher reconstruction accuracy in *PAH* and *SORCS3*, the shorter and more densely typed regions that we studied (reconstruction error around 10% or less for most populations, using 20 tSNPs), while 17q25, the longest and more sparsely typed region, proves to be the most difficult to reconstruct. In all four regions, the African populations show a high degree of heterogeneity and are more resistant to prediction than other populations.

Using tSNPs to reconstruct unassayed SNPs across populations

Finally, we explore the feasibility of predicting untyped SNPs in one population based on tSNPs selected on another population. Consider the following situation. We are given individuals from a reference population, typed over n SNPs. Now, a new, previously unstudied target population becomes of interest, and we seek to type a small number (say $c \ll n$) of tSNPs for this new target population and reconstruct the unassayed $n - c$ SNPs. We seek (1) tSNP selection algorithms to pick the SNPs to be assayed on the target population, given only the genotypes of all n SNPs in the reference population, and (2) tSNP reconstruction algorithms to reconstruct the unassayed SNPs, given only the genotypes of the c tSNPs in the target population and the genotypes of all n SNPs in the reference population (see Supplemental Fig. 5).

This situation represents the realistic scenario of an investigator designing a study based only on a reference population, e.g., a population studied in the HapMap project. To address this question in the Yale data set, we first assigned each of the 38 populations in turn as a reference. We then identified a set of tSNPs using the TSNPSMULTIPASSGREEDY algorithm targeting 90% and 99% of the spectral variance of the reference population and

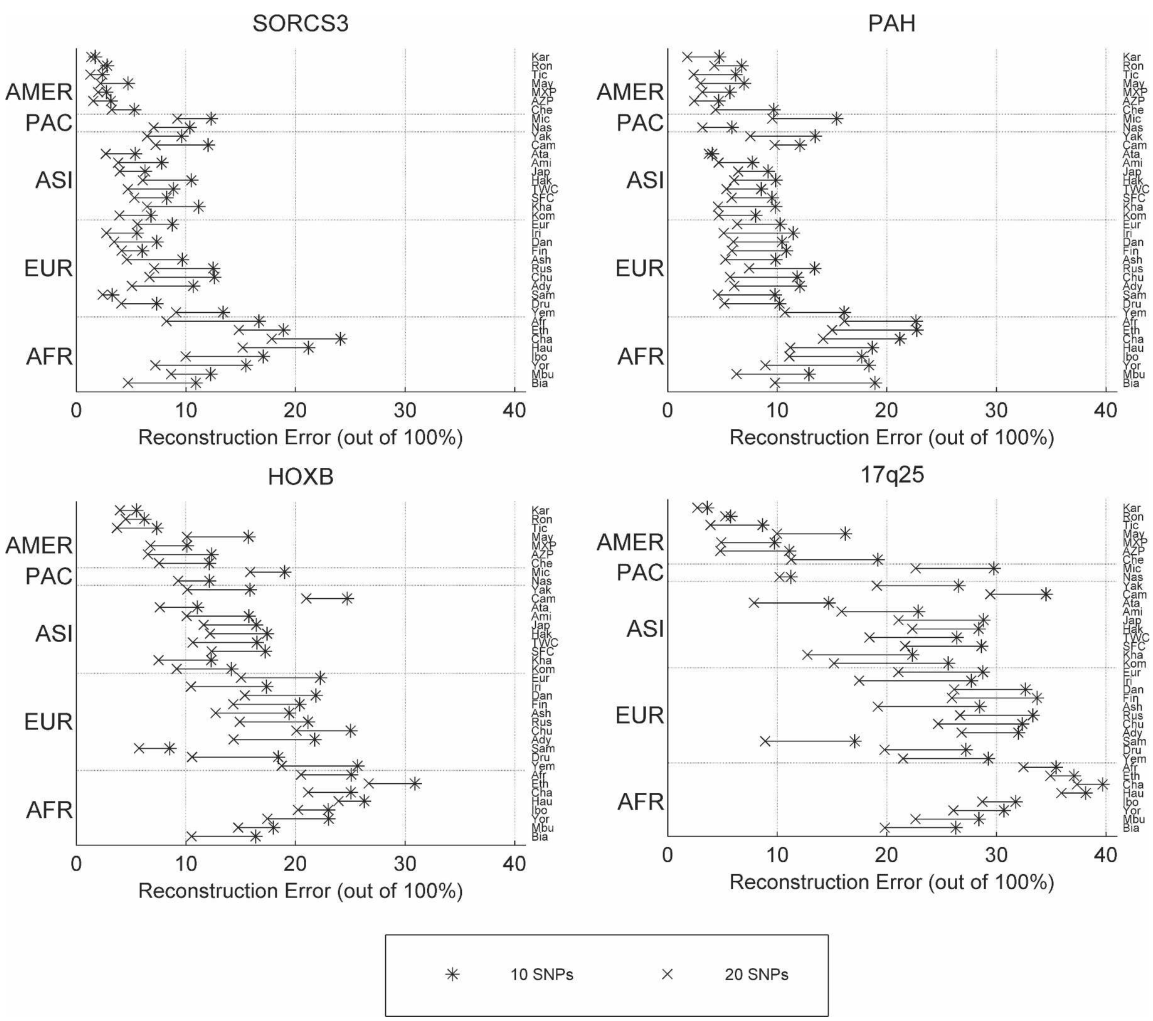

Figure 3. Intrapopulation reconstruction error (ratio of erroneously predicted entries over total number of predicted entries) for each of the 38 Yale data set populations. The training set size is 70% of the total population size. Populations are ordered (*bottom* to *top*) based on geographic regions (Africa, Europe, Asia, Micronesia, Americas).

assumed that (for each of the remaining 37 populations) these tSNPs were known. Finally, we reconstructed the unknown SNPs for all available individuals in each of the remaining 37 populations by using the ReconstructUnassayedSNPs algorithm. The same experiment was performed using the HapMap populations in the four regions that we studied. Transferability between the HapMap populations and the ones in the Yale data set could not be evaluated due to the very small overlap of the assayed SNPs in the two data sets.

Our findings for the Yale data set (Fig. 4A,B, targeting 99% of the reference population spectral variance and Supplemental Fig. 6a,b targeting 90%) suggest that there exists considerable transferability of tSNPs, mainly within the geographic boundaries of continents, but to a great extent also across them. What is particularly striking is the fact that the European populations in all four regions can be used here to predict, often with an error <10%, the majority of the Asian, Pacific, and American Indian populations. In general, moving out of Africa from West to East, populations can be used as a reasonably good reference for their more eastern neighbors, with the exception of those that are known to have remained isolated for many years, like the Samaritans or the Pacific Islanders. Interestingly, our very diverse

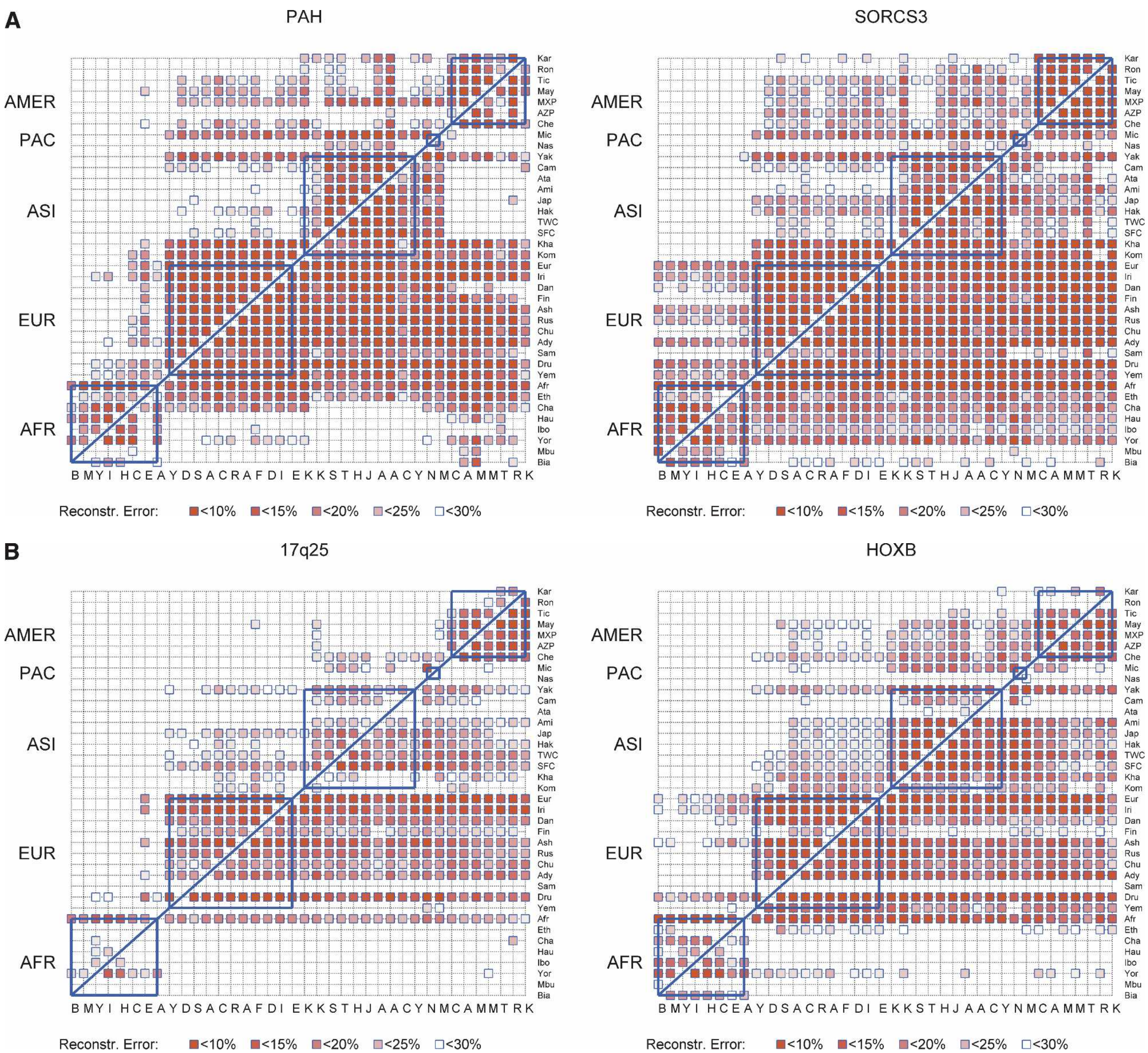

Figure 4. (*A*,*B*) Interpopulation reconstruction error (ratio of erroneously predicted entries over total number of predicted entries) for all pairs of populations. Populations are ordered (*bottom* to *top* and *left* to *right*) based on geographic regions (Africa, Europe, Asia, Micronesia, Americas). The (*i*, *j*)-th entry in the plot (*i*-th row, *j*-th column) corresponds to the reconstruction error for the *j*-th population, using the *i*-th population as reference. The SNPs to be assayed in the *j*-th population are determined by running the tSNPsMultiPassGreedy algorithm on the *i*-th population, seeking to explain 99% of the population's spectral variance. Blank entries correspond to reconstruction errors larger than 30%. The five geographic regions of our study are delimited by the blue boxes. (*A*) *PAH* and *SORCS3*; (*B*) 17q25 and *HOXB*.

sample of African Americans is the only one that can be used to predict unknown SNPs in almost all other populations in this study. This does not seem to be an artifact of the large number of selected tSNPs, since our analysis shows that even when the same number of tSNPs is selected in two reference populations from different continents, different populations will be captured in each case.

Although similar patterns are observed in all four genomic regions that we studied, the portability of tSNPs seems to be more pronounced in the short and more densely typed regions (*PAH* and *SORCS3*). The *SORCS3* region is 200 Kb longer than *PAH*. However, the structure of the region appears to be extremely homogeneous around the world. On the other hand, the 17q25 region has the least amount of tSNP transferability among populations. It has approximately the same length as the *HOXB* region that we analyzed (1 Mb), but was typed at a lower density (14.3 Kb vs. 11.9 Kb). It is not clear whether our results reflect the relatively poor marker resolution that we have for this region or the LD structure. As discussed in the next paragraph, our analysis of the HapMap genotypes for the same regions seems to support the first hypothesis.

The transferability of tSNPs among the HapMap populations (Table 3; Supplemental Table 4) seems to follow the same general principles as those shown from the analysis of the 38 populations. All four regions have been typed with markers at comparable spacing (between, on average, 1.3–2.4 Kb) and the reconstruction errors are also comparable for the same population pairs across the four regions. This depicts the effect of marker density on reconstruction accuracy.

Table 3. Interpopulation reconstruction error targeting 99% of the spectral variance of the reference population

	YRI	CEU	CHB	JPT
SORCS3				
YRI (26 SNPs)		**21.98**	**26.41**	**25.68**
CEU (16 SNPs)	31.48		**10.85**	**11.31**
CHB (12 SNPs)	46.05	**30.00**		**8.86**
JPT (14 SNPs)	36.23	**20.05**	**7.06**	
PAH				
YRI (27 SNPs)		35.20	63.11	60.01
CEU (18 SNPs)	**28.75**		**6.30**	**8.34**
CHB (14 SNPs)	55.51	32.66		**10.62**
JPT (14 SNPs)	42.00	**24.96**	**8.20**	
HOXB				
YRI (41 SNPs)		30.99	35.17	33.77
CEU (32 SNPs)	35.75		**14.67**	**13.84**
CHB (23 SNPs)	42.08	**28.04**		**11.98**
JPT (22 SNPs)	50.57	34.91	**13.58**	
17q25				
YRI (46 SNPs)		33.74	30.67	35.31
CEU (40 SNPs)	37.85		**16.39**	**16.74**
CHB (27 SNPs)	52.25	30.21		**19.00**
JPT (26 SNPs)	46.97	31.17	**16.23**	

The entries in boldface represent reconstruction error <30%.

Searching for association in a reconstructed data set

In order to further validate our methods and investigate their impact on the outcome of a real association study we used a publicly available data set previously studied for association with Crohn disease (Daly et al. 2001; Rioux et al. 2001). The data set consisted of 103 SNPs typed over 500 Kb on 5q31 for 139 family trios with one child affected with Crohn disease. First, we reproduced the results of the original study using the transmission test for linkage disequilibrium (TDT) as implemented by Haploview (Barrett et al. 2005). We chose $P \leq 2 \times 10^{-4}$ as the threshold of significance, to conform to the results reported in the original study. Eight markers were found to be associated with the disease in the original data set (see Supplemental material for association study). We then performed 100 random splits of the data in the training set (50% of the families) and the test set (the remaining families). In each trial, the training set was used as a reference for the selection of tSNPs targeting 90%–99% of its spectral variance in increments of 1%. The selected tSNPs were subsequently used to reconstruct the "unassayed" set of genotypes in the test set. We then performed the TDT for each of the reconstructed data sets, and we report here the average results over 100 runs for each target spectral variance.

As it is shown in Figure 5, using only eight of 103 SNPs, we achieve 90% reconstruction accuracy, while 41 SNPs are needed to reach 2.5% reconstruction error. As a reference for our savings success, we note here that an LD-based tSNP method, as implemented in Tagger (De Bakker et al. 2005), chooses 43 tSNPs for the same data set (capturing SNPs with r^2 threshold ≥0.8). We considered the eight markers found to be associated with Crohn disease in the original data set ($P \leq 2 \times 10^{-4}$) as ground truth, and we compared the results of our association experiments with this set of markers. Figure 5 shows the number of SNPs for which the TDT on the reconstructed data set erroneously exceeded (false positives, precision curve) or failed to reach the set threshold of significance (false negatives, recall curve). Remarkably, using only eight SNPs of 103 (10% reconstruction error) only two of the eight significant markers are missed and less than two (on average) erroneously exceed the set threshold of significance. When choosing 30 or 40 tSNPs (4% and 2.5% reconstruction error, respectively) one false negative and virtually no false positive results are found. It is true that some power is lost, however, the "false negative" SNPs of our test runs miss the mark only by very little, as revealed when plotting the TDT *P*-values for each of the eight significant SNPs using the original and the reconstructed data sets (Fig. 5). Interestingly, in almost every case, the *P*-values from the analysis of the reconstructed data sets are very close to those produced in the original tests. Furthermore, the one or two SNPs that appear as "false positives" are actually correlated to the SNPs reported in the original paper as significantly associated with the disease (data not shown). In any case, even with a 10% reconstruction error, which translates to 90% genotyping savings, the investigator would (in this example) retrieve the significant association in this chromosomal region, and could proceed to more focused genotyping in order to refine the findings of the analysis on the reconstructed data set.

Discussion

Most existing tSNP selection methods are either based on the arbitrary definition of haplotype block boundaries, or in the currently most common block-free approaches; tSNPs are picked

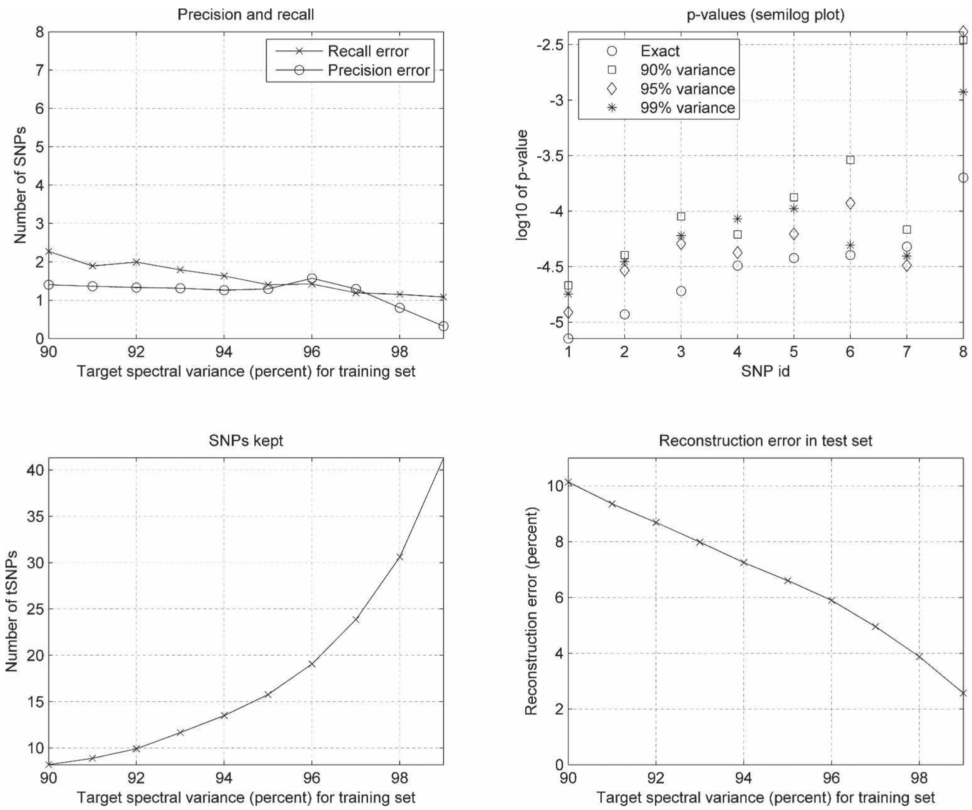

Figure 5. Genotype reconstruction for association analysis (50% training set). (*Top*, *left*) Number of SNPs for which the TDT on the reconstructed data set erroneously exceeded (false positives, precision curve) or failed to reach (false negatives, recall curve) the set threshold of significance $P \leq 2 \times 10^{-4}$. (*Top*, *right*) *P*-values for each of the SNPs that were significantly associated with the disease in the original data set (SNP id 1: IGR2063b_1, 2: IGR2060a_1, 3: IGR2055a_1, 4: IGR2096a_1, 5: IGR3081a_1, 6: IGR3096a_1, 7: IGR2198a_1, 8: IGR3236a_1), and the corresponding TDT *P*-values in reconstructed data sets targeting 90%, 95%, and 99% of the training set spectral variance ($\log_{10} 2 \times 10^{-4} \approx -3.7$). (*bottom left, right*) Number of tSNPs selected targeting 90%–99% of the training set spectral variance and reconstruction error in the test set.

based on correlations using the r^2 metric. A block-free method like the one we are using circumvents problems such as the arbitrary nature of block-length definitions and takes advantage of all existing associations, even across rigid block boundaries. If the occurrence of haplotype blocks is solely due to recombination hotspots, then SNP correlations will exist only within blocks (Goldstein and Weale 2001; Jeffreys et al. 2001). However, the formation of blocks may also be the result of the concurrent acting forces of recombination and population-specific demographic history (Wang et al. 2002; Zhang et al. 2002, 2004). On the other hand, methods that rely on r^2 estimations in order to set some SNPs (tSNPs) as proxies for others also depend on the inherent assumption that if SNP *A* is in LD with SNP *B* and SNP *B* is associated with a disease-causing variant, then SNP *A* will also be associated with the disease variant. This may not always be the case because of heterogeneity and confounding factors (Pritchard and Cox 2002; Montpetit et al. 2006; Terwilliger and Hiekkalinna 2006). Therefore, we suggest that instead of performing analysis on a set of tSNPs, one can use the next best alternative to actually having the entire data set available: an accurately reconstructed data set.

A few results (Evans et al. 2004; Halldorsson et al. 2004a; Lin and Altman 2004) in the genetics literature make an explicit attempt to evaluate their algorithms by reconstructing the "unknown" SNPs. Building upon recent results in the Computer Science and Applied Mathematics literature (Drineas and Mahoney 2005, 2007; Drineas et al. 2006a,b,c), we propose novel, scalable, linear algebraic algorithms that are useful in this context. In doing so, we show in a very large and diverse population sample that genotype reconstruction based on tSNPs is feasible and, even more interestingly, that it is possible to select tSNPs in one population in order to accurately predict unknown SNPs in a different population. Furthermore, we test the use of these algorithms in the setting of a real association study and find that significant associations with disease can be recovered in a reconstructed data set with important genotyping savings. An interesting direction for future research is to use LD tSNP selection methods and attempt reconstruction using these SNPs. Nevertheless, with this study we attempt to set the general mathematical framework for principled genotype reconstruction.

Our algorithm for tSNP selection can be readily applied to the genotypic data that current SNP typing technologies generate, without the need for the intermediate step of haplotype inference. Algorithms for tSNP selection that rely on EM-based algorithms (Excoffier and Slatkin 1995) or other haplotype inference techniques (Clark 1990; Hawley and Kidd 1995; Stephens et al. 2001; Niu et al. 2002) are computationally expensive and unlikely to be scalable to very large or whole-genome data sets.

On the contrary, given n SNPs and m individuals, our algorithm for tSNP selection scales linearly with the number of SNPs and individuals in the data. Using standard Computer Science notation, the running time of our algorithms is $O(mn)$. For reference, our algorithms ran in under 30 sec in a 2.5-GHz Pentium with 1 GB of RAM for each of the largest runs presented here, thus suggesting that extensions to much larger genome-wide SNP data sets are possible.

A few other methods motivated by linear algebra considerations, and in particular the SVD and the related Principle Components Analysis (PCA), have been previously applied to the tSNP selection problem (Meng et al. 2003; Horne and Camp 2004; Lin and Altman 2004). Lin and Altman claimed that PCA-based methods will likely be very difficult to apply on whole-genome data sets. Recent approximation algorithm results (Drineas et al. 2006b) suggest otherwise if the data are very large and if approximate solutions are adequate for the particular application.

The transferability of tSNPs among populations is a question that is beginning to be addressed by recent studies, which have either studied only a few populations or a single genomic region. Common sets of tSNPs have been defined based on the evaluation of correlations between "known" SNPs and "unassayed" ones (Ke et al. 2004; Mueller et al. 2005; Ramirez-Soriano et al. 2005; De Bakker et al. 2006; Gonzalez-Neira et al. 2006; Magi et al. 2006; Montpetit et al. 2006; Willer et al. 2006). Gonzalez-Neira et al. (2006) have recently presented a study of a worldwide sample of populations (1055 individuals) and one genomic region (1 Mb at ≈ 7 Kb density) and concluded, like we do in this study, that portability of tSNPs does exist among populations within each continental group and that tSNPs defined in Europeans are often efficient for Middle/Eastern and Central/South Asian populations. We take this kind of study one step further by studying a much larger worldwide sample (2000 individuals) and four genomic regions. Furthermore, by attempting to reconstruct untyped genotypes in our very large and diverse set of populations, we are able to quantify the amount of tSNP transferability that exists within geographic boundaries of continents, but also across them. The observed patterns of tSNP transferability reflect population relationships, histories, and migrations of ancient populations.

Even at the cost of typing extra SNPs, our study indicates that the populations used in the HapMap project will most likely serve as a good reference for extrapolation of results in other populations, especially Europeans and East Asians. Our results quantify the rather intuitive observation that given a target unstudied population, it is always better to pick a reference population from the same geographic region, since the transferability of tSNPs is significantly higher. However, we would like to note that although the tSNP selection concept in general will likely be very efficient for the analysis of common SNPs, rare variants will most probably be overlooked and different approaches should perhaps be pursued if such variants are of interest. To the extent that the common disease/common variant (Lander 1996; Chakravarti 1999) hypothesis is valid, the HapMap project and tSNP selection will prove to be powerful tools for the design of association studies.

In conclusion, we explored the extent of linear algebraic structure of genotypic data in four regions of the genome and illustrated the value of linear structure extrapolation techniques for the selection of tSNPs and reconstruction of untyped SNPs. A MatLab implementation of our algorithms and the data studied here are available at http://www.cs.rpi.edu/~javeda/CUR_ tSNPs.htm. Our results indicate that reconstruction accuracy increases with reference map density and LD of the studied region. The pattern of linear structure in our sample of worldwide populations is reminiscent of the observed LD patterns around the world and it seems possible that similar forces may have acted to shape both the LD and linear structure in such data. Further study should shed more light on the degree of correlation between the linear structure observed in a data set and the underlying LD patterns and haplotype structures. It is also possible that nonlinear structure extraction techniques will prove to be the most promising in order to elucidate in a more refined manner the genomic architecture of extremely diverse or richly structured populations.

Methods

Data sets

We present data on a total of 1979 unrelated individuals from 38 populations from around the world (Supplemental Fig. 1). ALFRED (http://alfred.med.yale.edu/), the allele frequency database, contains descriptive information and literature citations for these population samples. A total of 248 SNPs in four genomic regions were typed in all 38 populations. We also investigated the same four genomic regions (*SORCS3*, *PAH*, *HOXB*, and 17q25) using the available genotypes from the HapMap database on the four HapMap populations. We only included in our analysis data from SNPs that were polymorphic in all four populations (a total of 1336 SNPs; see Table 1 for details). Since we had selected the SNPs to be typed in the Yale samples well before the publication of the HapMap results, there was little to no overlap between the SNPs that we studied in the Yale samples and those typed in the HapMap populations. Finally, the data set that we used for validation of our algorithms in an association study is publicly available at http://www.broad.mit.edu/humgen/IBD5/ and has been described in detail (Daly et al. 2001; Rioux et al. 2001).

Encoding our data and evaluating linear structure

We transformed the raw data to numeric values, without any loss of information, in order to apply our linear algebraic algorithms. See Algorithm Encode in the Supplemental material for a precise statement of this procedure. For clarity of presentation, we filled in a (very small) number of missing entries in the Yale and HapMap data sets using the procedure described in Alter et al. (2000); for the association study data set, we did not fill in the missing data.

Our algorithms for tSNP selection and tSNP transferability take advantage of and extract linear structure in the SNP data matrix. In order to determine the extent to which the SNP data matrix has this structure, we shall use the Singular Value Decomposition (SVD) (Horn and Johnson 1985; Golub and VanLoan 1989). The SVD is a commonly used tool from Linear Algebra to extract linear or low-rank structure in data represented by a matrix. For example, it provides the mathematical foundation for the commonly used method of Principal Components Analysis (PCA). We emphasize that the SVD is used in this work only to determine the extent of linear structure in the data matrix. Our algorithms for tSNP selection and tSNP transferability will extract linear structure, but will not use the SVD.

Given an $m \times n$ matrix A, the SVD returns m pairwise orthogonal unit vectors u^i that form a complete basis for the m-dimensional Euclidean space, n pairwise orthogonal unit vectors v^i that form a complete basis for the n-dimensional Euclidean space, and $\rho = \min\{m,n\}$ singular values σ_i such that $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_\rho \geq 0$. The matrix A may be written as a sum of outer products (rank-one components) as $A = \sum_{i=1}^{\rho} \sigma_i u^i v^{i^T}$. We notice that when applied to a SNP data matrix the u^i are associated with the columns (SNPs) of this matrix and are called eigenSNPs (Lin and Altman 2004). Notice that the *i*-th singular vector corresponds to the *i*-th singular value, and thus, there exists a natural ordering of the singular vectors. One interpretation of the SVD is that keeping the top $k \leq \rho$ left singular vectors we can express all of the columns of the matrix A as linear combinations of these k left singular vectors with a small loss in accuracy. More precisely, for all $i = 1, \ldots, n$, $A^{(i)} \approx \sum_{j=1}^{k} z_{ij} u^j$; where $A^{(i)}$ denotes the *i*-th column of A as a column vector and the z_{ij} are real numbers. The z_{ij} are computed by solving least squares regression problems to minimize the Euclidean norm of the difference vector $A^{(i)} - \sum_{j=1}^{k} z_{ij} u^j$. Overall, using the top k left singular vectors we can approximate A by $A \approx A_k = U_k Z$; where Z is the $k \times n$ matrix whose entries are the z_{ij}. Standard methods from Linear Algebra can be used to show that $Z = U_k^T A$. If the difference $A - A_k$ is small, then we say that A is well-approximated by a rank- k matrix, and if $k \ll \min\{m, n\}$, then we say that A is approximately low-rank or has good linear structure. Intuitively, this means that there is significant redundancy of information in the columns of A. Algorithm 2 in the Supplemental material describes in detail how we evaluate the linear structure in our populations.

Selecting tSNPs and reconstructing genotypes

Via the SVD, we can compute a set of vectors $u^1, \ldots, u^k$ such that every column of A^X (the matrix encoding SNP data from population X) may be expressed as a linear combination of these k vectors with a small, fixed loss in accuracy. Since the columns of A^X are the SNPs that were assayed on X, one might be tempted to call the $u^1, \ldots, u^k$ tSNPs for population X. Unfortunately, the $u^1, \ldots, u^k$ are not actual SNPs (columns of A^X). Instead, they are linear combinations of actual SNPs, and in general, have no biological interpretation.

An obvious next step is to wonder whether we can find a small number of columns of A^X (namely, actual SNPs) such that expressing every column of A^X as a linear combination of these columns by solving least squares regression problems and subsequently rounding the result would return an approximation to A^X with a small number of erroneous entries. Toward that end, we slightly modified the SelectColumnsMultiPass algorithm of Drineas and Mahoney (2007). The resulting tSNPsMultiPassGreedy algorithm does not come with a provable performance guarantee, but differs from most algorithms in current genetics literature in that it is guided by strong theoretical evidence regarding its performance. See the Supplemental material for an exact description of the algorithm.

We now describe our interpopulation reconstruction algorithm. Consider the matrix A^X corresponding to the m_1 subjects of population X, and assume that we seek to predict the SNPs for all m_2 subjects of a different population Y. Assume that the subjects of X are fully assayed. We will assay a small number (say $c \ll n$) of SNPs for the subjects in Y and predict the remaining $n - c$ SNPs for every subject in Y. In order to determine which c SNPs to assay for the subjects in Y we use the tSNPsMultiPassGreedy algorithm on A^X. After assaying the selected SNPs for the subjects of Y we will get an $m_2 \times c$ matrix C^Y. The ReconstructUnassayedSNPs algorithm (which implements a *CUR*-type decomposition of a matrix) essentially performs a least-squares regression fit for the subjects of Y (Drineas et al. 2006c). See the Supplemental material for an exact description of the algorithm. The same algorithm may be used to reconstruct unassayed SNPs of individuals within a population. More specifically, given a population X, we split the individuals into two sets: a training set X_1 and a test set X_2. The ReconstructUnassayedSNPs algorithm is used with X_1 instead of X and X_2 instead of Y.

Acknowledgments

This work was funded in part by a National Science Foundation CAREER award to P.D., National Institute of Health grants GM57672 to K.K.K., and NS40025 to the Tourette Syndrome Association, and a grant from the TSA to P.P. We thank Daniel Votava for his excellent technical help. We also want to acknowledge and thank the following people who helped assemble the samples from the diverse populations and make them available to us: F.L. Black, B. Bonne-Tamir, L.L. Cavalli-Sforza, K. Dumars, J. Friedlaender, E. Grigorenko, S.L.B. Kajuna, N.J. Karoma, K. Kendler, W. Knowler, S. Kungulilo, R-B Lu, A. Odunsi, F. Okonofua, H. Oota, F. Oronsaye, M. Osier, J. Parnas, L. Peltonen, L.O. Schulz, K. Weiss, and O.V. Zhukova. In addition, some of the cell lines were obtained from the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University, Israel, and the African American samples were obtained from the Coriell Institute for Medical Research, Camden, New Jersey. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples for studies such as this. Without such participation of individuals from diverse parts of the world we would be unable to obtain a true picture of the genetic variation in our species.

References

Alter, O., Brown, P.O., and Botstein, D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.* **97:** 10101–10106.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263–265.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74:** 106–120.

Chakravarti, A. 1999. Population genetics–Making sense out of sequence. *Nat. Genet.* **21:** 56–60.

Chapman, J.M., Cooper, J.D., Todd, J.A., and Clayton, D.G. 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.

Clark, A.G. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7:** 111–122.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., and Lander, E.S. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29:** 229–232.

De Bakker, P.I.W., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., and Alshuler, D. 2005. Efficiency and power in genetic association studies. *Nat. Gen.* **37:** 1217–1223.

De Bakker, P.I., Graham, R.R., Altshuler, D., Henderson, B.E., and Haiman, C.A. 2006. Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations. *Pac. Symp. Biocomput.* 478–486.

Ding, K., Zhou, K., Zhang, J., Knight, J., Zhang, X., and Shen, Y. 2005. The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection. *Mol. Biol. Evol.* **22:** 148–159.

Drineas, P. and Mahoney, M.W. 2005. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.* **6:** 2153–2175.

Drineas, P. and Mahoney, M. 2007. A randomized algorithm for a tensor-based generalization of the SVD. In *Linear algebra and its applications*. **420:** 553–571. Elsevier.

Drineas, P., Kannan, R., and Mahoney, M.W. 2006a. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication.

SIAM J. Comput. **36:** 132–157.
Drineas, P., Kannan, R., and Mahoney, M.W. 2006b. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM J. Comput.* **36:** 158–183.
Drineas, P., Kannan, R., and Mahoney, M.W. 2006c. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM J. Comput.* **36:** 184–206.
Evans, D.M., Cardon, L.R., and Morris, A.P. 2004. Genotype prediction using a dense map of SNPs. *Genet. Epidemiol.* **27:** 375–384.
Excoffier, L. and Slatkin, M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12:** 921–927.
Frieze, A., Kannan, R., and Vempala, S. 2004. Fast Monte-Carlo algorithms for finding low-rank approximations. *J. ACM* **51:** 1025–1041.
Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. 2002. The structure of haplotype blocks in the human genome. *Science* **296:** 2225–2229.
Goldstein, D.B. and Weale, M.E. 2001. Population genomics: Linkage disequilibrium holds the key. *Curr. Biol.* **11:** R576–R579.
Golub, G.H. and VanLoan, C.F. 1989. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD.
Gonzalez-Neira, A., Ke, X., Lao, O., Calafell, F., Navarro, A., Comas, D., Cann, H., Bumpstead, S., Ghori, J., Hunt, S., et al. 2006. The portability of tagSNPs across populations: A worldwide survey. *Genome Res.* **16:** 323–330.
Halldorsson, B.V., Bafna, V., Lippert, R., Schwartz, R., DeLaVega, F.M., Clark, A.G., and Istrail, S. 2004a. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Res.* **14:** 1633–1640.
Halldorsson, B.V., Istrail, S., and DeLaVega, F.M. 2004b. Optimal selection of SNP markers for disease association studies. *Hum. Hered.* **58:** 190–202.
Hawley, M.E. and Kidd, K.K. 1995. HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.* **86:** 409–411.
Horn, R.A. and Johnson, C.R. 1985. *Matrix Analysis*. Cambridge University Press, New York.
Horne, B.D. and Camp, N.J. 2004. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genet. Epidemiol.* **26:** 11–21.
The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.
The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.
Jeffreys, A.J., Kauppi, L., and Neumann, R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29:** 217–222.
Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., DiGenova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., et al. 2001. Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29:** 233–237.
Ke, X. and Cardon, L.R. 2003. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19:** 287–288.
Ke, X., Durrant, C., Morris, A.P., Hunt, S., Bentley, D.R., Deloukas, P., and Cardon, L.R. 2004. Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum. Mol. Genet.* **13:** 2557–2565.
Lander, E.S. 1996. The new genomics: Global views of biology. *Science* **274:** 536–539.
Lin, Z. and Altman, R.B. 2004. Finding haplotype tagging SNPs by use of principal components analysis. *Am. J. Hum. Genet.* **75:** 850–861.
Magi, R., Kaplinski, L., and Remm, M. 2006. The whole genome tagSNP selection and transferability among HapMap populations. *Pac. Symp. Biocomput.* 535–543.
Meng, Z., Zaykin, D.V., Xu, C.F., Wagner, M., and Ehm, M.G. 2003. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am. J. Hum. Genet.* **73:** 115–130.
Montpetit, A., Nelis, M., Laflamme, P., Magi, R., Ke, X., Remm, M., Cardon, L., Hudson, T.J., and Metspalu, A. 2006. An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet.* **2:** 282–290.
Mueller, J.C., Lohmussaar, E., Magi, R., Remm, M., Bettecken, T., Lichtner, P., Biskup, S., Illig, T., Pfeufer, A., Luedemann, J., et al. 2005. Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am. J. Hum. Genet.* **76:** 387–398.
Niu, T., Qin, Z.S., Xu, X., and Liu, J.S. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **70:** 157–169.
Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P., et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294:** 1719–1723.
Pritchard, J.K. and Cox, N.J. 2002. The allelic architecture of human disease genes: Common disease-common variant or not? *Hum. Mol. Genet.* **11:** 2417–2423.
Ramirez-Soriano, A., Lao, O., Soldevila, M., Calafell, F., Bertranpetit, J., and Comas, D. 2005. Haplotype tagging efficiency in worldwide populations in CTLA4 gene. *Genes Immun.* **6:** 646–657.
Rioux, J.D., Daly, M.J., Silverberg, M.S., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., et al. 2001. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.* **29:** 223–228.
Schwartz, R., Halldorsson, B.V., Bafna, V., Clark, A.G., and Istrail, S. 2003. Robustness of inference of haplotype block structure. *J. Comput. Biol.* **10:** 13–19.
Sebastiani, P., Lazarus, R., Weiss, S.T., Kunkel, L.M., Kohane, I.S., and Ramoni, M.F. 2003. Minimal haplotype tagging. *Proc. Natl. Acad. Sci.* **100:** 9900–9905.
Stephens, M., Smith, N.J., and Donnelly, P. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68:** 978–989.
Stram, D.O., Haiman, C.A., Hirschhorn, J.N., Altshuler, D., Henderson, I.N., Kolonel, B.E., and Pike, M.C. 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.
Stumpf, M.P. 2002. Haplotype diversity and the block structure of linkage disequilibrium. *Trends Genet.* **18:** 226–228.
Terwilliger, J.D. and Hiekkalinna, T. 2006. An utter refutation of the "Fundamental Theorem of the HapMap". *Eur. J. Hum. Genet.* **14:** 426–437.
Wall, J.D. and Pritchard, J.K. 2003a. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Hum. Genet.* **73:** 502–515.
Wall, J.D. and Pritchard, J.K. 2003b. Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **4:** 587–597.
Wang, N., Akey, J.M., Zhang, K., Chakraborty, R., and Jin, L. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* **71:** 1227–1234.
Weale, M.E., Depondt, C., Macdonald, S.J., Smith, A., Lai, P.S., Shorvon, S.D., Wood, N.W., and Goldstein, D.B. 2003. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: Implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* **73:** 551–565.
Willer, C.J., Scott, L.J., Bonnycastle, L.L., Jackson, A.U., Chines, P., Pruim, R., Bark, C.W., Tsai, Y.-Y., Pugh, E.W., Doheny, K.F., et al. 2006. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet. Epidemiol.* **30:** 180–190.
Zeggini, E., Barton, A., Eyre, S., Ward, D., Ollier, W., Worthington, J., and John, S. 2005. Characterisation of the genomic architecture of human chromosome 17q and evaluation of different methods for haplotype block definition. *BMC Genet.* **6:** 21.
Zhang, K., Deng, M., Chen, T., Waterman, M.S., and Sun, F. 2002. A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci.* **99:** 7335–7339.
Zhang, K., Qin, Z.S., Liu, J.S., Chen, T., Waterman, M.S., and Sun, F. 2004. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res.* **14:** 908–916.
Zhang, K., Qin, Z., Chen, T., Liu, J.S., Waterman, M.S., and Sun, F. 2005. HapBlock: Haplo-type block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics* **21:** 131–134.

Received July 6, 2006; accepted in revised form November 1, 2006.