Annals of
human genetics

doi: 10.1111/j.1469-1809.2011.00673.x

Efficient Genomewide Selection of PCA-Correlated tSNPs for Genotype Imputation

Asif Javed[1,2*], Petros Drineas[2], Michael W. Mahoney[3] and Peristera Paschou[4†]

[1] *Computational Biology Center, IBM T. J. Watson Research, Yorktown Heights, NY 10598, USA*
[2] *Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*
[3] *Department of Mathematics, Stanford University, Palo Alto, CA 94305, USA*
[4] *Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupoli 68100, Greece*

Summary

The linkage disequilibrium structure of the human genome allows identification of small sets of single nucleotide polymorphisms (SNPs) (tSNPs) that efficiently represent dense sets of markers. This structure can be translated into linear algebraic terms as evidenced by the well documented principal components analysis (PCA)-based methods. Here we apply, for the first time, PCA-based methodology for efficient genomewide tSNP selection; and explore the linear algebraic structure of the human genome. Our algorithm divides the genome into contiguous nonoverlapping windows of high linear structure. Coupling this novel window definition with a PCA-based tSNP selection method, we analyze 2.5 million SNPs from the HapMap phase 2 dataset. We show that 10–25% of these SNPs suffice to predict the remaining genotypes with over 95% accuracy. A comparison with other popular methods in the ENCODE regions indicates significant genotyping savings. We evaluate the portability of genome-wide tSNPs across a diverse set of populations (HapMap phase 3 dataset). Interestingly, African populations are good reference populations for the rest of the world. Finally, we demonstrate the applicability of our approach in a real genome-wide disease association study. The chosen tSNP panels can be used toward genotype imputation using either a simple regression-based algorithm or more sophisticated genotype imputation methods.

Keywords: tSNPs, PCA, genotype imputation, HapMap 3

Introduction

Single nucleotide polymorphisms (SNPs) represent the most abundant form of variation in the human genome and are the main target of studies searching to identify susceptibility variants for common complex disorders. Without prior knowledge about the full pathways leading to a specific disease phenotype, identifying causal mutations is like looking for a genetic needle in a genomic haystack. Practical limitations of genotyping costs often restrict the number of SNPs that can be assayed for each individual participating in a study and highlight the need to prioritize these markers. Fortunately, the linkage disequilibrium (LD) structure of the human genome induces a lot of redundancy among neighboring SNPs and makes them good predictors of each other (Daly et al., 2001; Johnson et al., 2001).

Various methods have been proposed to exploit this LD structure by identifying a small set of representative markers which capture a large amount of genetic variability; for a review see Halldorsson et al. (2004b) and Stram (2004). These markers are commonly called tagging SNPs (tSNPs) and represent the remaining (tagged) SNPs. Most studies addressing sufficiently large genomic regions rely on defining blocks (or neighborhoods) of low diversity and high LD, where the neighboring SNPs are highly predictive of each other (Gabriel et al., 2002; Zhang et al., 2005). Some quantitative measure is then used to identify a few SNPs within each block which then represent the complete block. The r^2 coefficient is one widely used correlation measure. Pairwise comparisons among SNPs have been proposed and successfully implemented to ensure that most tagged SNPs are in high r^2 correlation with at least one SNP (Carlson et al., 2004)

[*] This work was completed while the author was a Ph.D. student at Rensselaer Polytechnic Institute.
[†] Corresponding author: Peristera Paschou, Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupoli 68100, Greece. Tel: +302551030658; Fax: +302551030613; E-mail: ppaschou@mbg.duth.gr

or a multimarker haplotype (DeBakker et al., 2005) in the tSNP set. Most of these methods rely on haplotype inference for the definition of blocks, which not only acts as an additional source of error, but also makes them computationally expensive and inhibits their scalability to genome-wide datasets.

Aiming to tackle the complexity of observed LD patterns, several methods have been previously proposed for the definition of haplotype blocks. Patil et al. (2001) defined a haplotype block as a segment of consecutive SNPs where at least 80% of the observed haplotypes are represented more than once. Tagging SNPs are subsequently selected within each block with the objective of uniquely distinguishing these common haplotypes and a greedy algorithm was developed in order to assign block boundaries that reduce the number of tSNPs. Zhang et al. (2002, 2004, 2005) extended this idea by formulating the objective as a mathematical problem and they proposed a dynamic programming solution in order to minimize the number of tSNPs. This algorithm is implemented in HapBlock (Zhang et al., 2005). Gabriel et al. (2002) used pairwise correlation among markers to define haplotype blocks; here two SNPs are considered strongly correlated if the 95% confidence bound of D' between them exceeds 0.7. Recall that D' is an often-used measure of frequency of recombination events and a high value implies evidence of little to no recombination between the markers. A haplotype block is defined if at least 95% of the SNP pairs within the block satisfy these criteria. This algorithm is part of the software suite made publicly available in Haploview (Barrett et al., 2005).

Principal component analysis (PCA) is a linear algebraic method which has been successfully used in multiple studies in order to study population structure and identify tSNPs in short genomic regions (Price et al., 2006; Paschou et al., 2007, 2008). However, the linear algebraic structure of the human genome has not been studied in detail in prior work. This is mainly due to the fact that applying linear algebraic methods to whole-genome data in order to select tSNPs, results in false long distance correlations. Such false correlations stem both from the existence of rare SNPs, as well as from the fact that existing genome-wide datasets are usually extremely skewed and have two or three orders of magnitude more SNPs than individuals. For example, most such datasets include a number of individuals that is in the order of thousands, whereas the number of SNPs is in the order of hundreds of thousands or millions. In order to eliminate such false correlations, prior work decomposed the genome-wide data in small windows and applied tSNP selection algorithms within each window. Meng et al. (2003) addressed these issues by using a sliding window of an arbitrary fixed size. Varimax rotation was used in each window to identify SNPs which capture the same subspace as the significant principal components and the remaining SNPs were discarded. Using the sliding window, multiple scans of the chosen SNPs were conducted to further remove the redundant markers. Meng et al. acknowledged the difficulty and advantage of incorporating LD information while determining window size, but they contended that in the absence of such a window definition, multiple scans provide a useful alternative in reducing the number of tSNPs in high LD regions. Lin and Altman got rid of the sliding window and directly applied Varimax rotation on the complete dataset (Lin & Altman, 2004). However, they conceded that their approach would not extend well to long genomic regions of varying linkage. Horne and Camp binned SNPs into LD groups correlated with each significant principle component (Horne & Camp, 2004) but ignored the relative physical position of SNPs on the chromosome. The authors focused on intragenic genetic variation and claimed that their method would capture rare variants within a gene which is more likely due to recent mutations. This claim holds true within each genic region. However, if this analysis is extended to long genomic segments, the correlations are highly likely to be an artifact of undersampling of the population.

In this study, we investigate for the first time the applicability of PCA in order to efficiently select tSNPs across the entire genome. In order to tackle this problem , we introduce a novel eigenanalysis-based definition of genomic windows which reflect the LD structure of the underlying genetic region. This approach integrates seamlessly with linear algebraic methods of tSNP selection and genotype imputation (Paschou et al., 2007). However, we also show that the tSNPs selected using our methods can easily be applied for genotype imputation using non-PCA based algorithms, such as those implemented in Beagle (Browning & Browning, 2009) or Impute (Marchini et al., 2007), providing important genotyping savings and improved accuracy, albeit at a larger computational cost. Studying autosomal data from the HapMap phase 2 database (The International HapMap Consortium, 2003, 2005), we demonstrate that our algorithms scale extremely well to genomewide tagging. Considering about 2.5 million SNPs over the entire genome, we find that as few as 13% of these SNPs for the HapMap Asian populations, 11% for the CEPH Europeans, and 24% for the Yoruba, suffice to predict the full dataset with more than 95% accuracy, while the complete analysis for each population takes only four and a half hours on commodity hardware. Analyzing seven ENCODE regions (The ENCODE Project Consortium, 2007) of the HapMap project, we compare the efficiency and accuracy of our tSNP selection approach to a popular tSNP selection algorithm, Tagger (DeBakker et al., 2005) and we demonstrate that significant savings are achieved with our window definition method, over previously described methods for block definition. Furthermore, we use the HapMap phase 3 data in an interpopulation genotype prediction study.

The Africans, who represent the birthplace of all modern humans, retain most information about human genetic variation and are good reference populations for the rest of the world. Finally, our algorithms are used in real data from a genome-wide association study in order to significantly reduce genotyping costs with minimal loss in power.

Methods

Datasets

We studied data available from the HapMap phase 2 database for Yoruban (YRI), CEPH European (CEU), Chinese, and Japanese samples (The International HapMap Consortium, 2003, 2005) (total of 270 individuals genotyped for more than three million SNPs, release 21). For the purposes of our study, the Chinese and Japanese samples were considered as a joint Asian population (ASI). We also analyzed data from seven ENCODE regions, again available from the HapMap database. The ENCODE regions have been selected for extensive genotyping on the HapMap populations and they represent one of the largest and denser genotype datasets today. In order to gauge the portability of tSNP panels and prediction coefficients, a diverse set of populations from the HapMap phase 3 dataset were also used (release 1). This dataset consists of the above mentioned four populations from phase 2 along with seven additional ones: African American (ASW), Chinese (CHD), Indian (GIH), Kenyan (LWK, MKK), Mexican (MEX), and Italian (TSI). After removing markers monomorphic in any one of the populations we were left with 1,015,780 markers genotyped for 1115 individuals.

For the association study, we analyzed a dataset made publicly available by the Corriell institute. The dataset consists of approximately 500 samples of European American ancestry genotyped for approximately 400,000 SNPs. The DNA samples come from patients with Parkinson's disease and neurologically normal controls and has been previously been described in (Fung et al., 2006). The samples are curated at the Coriell institute. Genotyping was performed using the Illumina platform. For all datasets, we only considered genotypes for autosomal SNPs in our analysis.

Encoding the Data

The proportion of missing entries in the above datasets was very small (on average less than 0.1%). As a quality control step, we excluded all SNPs with more than 10% missing entries. For each population, we omitted monomorphic SNPs from our analysis, since they are trivial to predict. After these preprocessing steps, we were left with a total of 2,273,598 SNPs for the Asian populations (out of 3,776,828), 2,421,152 SNPs for CEPH Europeans (out of 3,775,447), and 2,689,571 for the Yoruba (out of 3,685,183). For the CORIELL dataset (besides applying these filters), we analyzed only those SNPs that were common with the HapMap European population (369,627 SNPs out of 396,591). In order to simplify and speed up our computations, we filled in the (very small) number of missing entries randomly so that HWE is satisfied for each SNP. The probabilistic filling in was performed separately for each dataset, and separately in each population of the HapMap data. We then transformed the raw data to numeric values, without any loss of information, in order to apply the singular value decomposition (SVD) and extract the principal components. Consider a dataset of a population X consisting of m subjects and assume that for each subject n biallelic SNPs have been assayed. Thus, we are given a table T^X, consisting of m rows and n columns. Each entry in the table is a pair of bases, ordered alphabetically. We transform this initial data table to an integer matrix A^X which consists of m rows, one for each subject, and n columns, one for each SNP. Each entry of A^X will be -1, 0, $+1$, or empty. Let B_1 and B_2 be the bases that appear in the jth SNP (in alphabetical order). If the genotypic information for the jth SNP of the ith individual is B_1B_1 the (i, j)th entry of A^X is set to $+1$; else if it is B_1B_2 the (i, j)th entry of A^X is set to 0; else if it is B_2B_2 the (i, j)th entry of A^X is set to -1 (see the ENCODE algorithm in supplementary material for details).

Computing Low-Rank Approximation via the SVD

We will employ the SVD of matrices in order to define windows in our approach. This section briefly describes this very useful linear algebraic tool. Given m subjects and n SNPs, let the $m \times n$ matrix A denote the subject-SNP matrix encoded as described above. Then, the SVD of A returns m pairwise orthonormal vectors u^i, n pairwise orthonormal vectors v^i, and m non-negative singular values σ_i. The matrix A may be written as a sum of outer products as

$$A = \sum_{i=1}^{m} \sigma_i u^i v^{i\,T}.$$

Each triplet $(\sigma_i,\ u^i,\ v^i)$ may be used to form a principal component of A. In our setting, the left singular vectors (the u^i's) are linear combinations of the columns (SNPs) of A and will be called eigenSNPs (Lin & Altman, 2004). It is well known that keeping only the top k triplets $(\sigma_i,\ u^i,\ v^i)$ results in the best rank k approximation to A, which is denoted by A_k, and is equal to

$$A_k = \sum_{i=1}^{k} \sigma_i u^i v^{i\,T}.$$

Defining Windows for Tagging SNP Selection

Most genome-wide datasets, that are available today, are comprised of hundreds of thousands (or millions) of markers assayed for only a limited number of samples per population. Consider for example the HapMap phase 2 dataset, which consists of more than three million SNPs assayed for only 90 individuals in three populations. Recall that we consider the Chinese and Japanese samples as one Asian population. The relatively small number of available samples compared to the large number of assayed SNPs results in false structural effects in the dataset, an effect that is accentuated by the fact that more than a third of the SNPs in every population have a minor allele frequency of less than 10% (rare SNPs have a much higher probability of exhibiting false long distance correlations). As a result, one needs to define windows of consecutive SNPs in the datasets and sacrifice any correlation across different windows. This seems to be a necessary evil in datasets with many more SNPs than individual samples in order to improve accuracy. The so-called "block-free" algorithms in existing literature almost invariably require a neighborhood definition (Halldorsson et al., 2004a) when targeting long genomic regions in order to address artificial long distance correlations. Most prior window and block definitions rely on haplotype inference (Johnson et al., 2001; Gabriel et al., 2002; Stram, 2004), which is computationally expensive and acts as an additional source of error. Furthermore, these haplotype blocks do not have a direct linear algebraic interpretation thus rendering their use with PCA based methods meaningless. In these circumstances, linear algebraic tagging methods have resorted to arbitrarily fixed sized windows. This study contributes a new, simple, linear algebraic window definition specifically catered for these algorithms.

In order to introduce our window definition (see also Fig. 1 for a block diagram depiction of the proposed algorithm), we provide a simple example. Our procedure exploits the fact that consecutive SNPs are often correlated and takes two input parameters, which we will call *accuracy* and *number of eigenSNPs*. Let the accuracy be set to, say, 95% and the number of eigenSNPs to k. First, we start with just one SNP. Assume that we have already added i consecutive SNPs to our window. In order to determine whether the $(i + 1)$th SNP will be added, let A be the matrix containing the $(i + 1)$ SNPs in the current window. We then compute the best rank k approximation A_k and compare it to A. If the resulting error is less than 5%, then we add the $(i + 1)$th SNP to the current window. Otherwise, a new window starts at the $(i + 1)$th SNP (see Window Definition in supplementary material for details). It is worth noting that window breaks are always introduced at the end of a chromosome, since there is no real advantage to having windows that span multiple chromosomes. This procedure guarantees that the resulting windows will have high linear structure, since a low-rank (e.g., k) approximation to each window will result in a reconstruction accuracy of at least 95%. Notice that the *accuracy* parameter quantifies the user's need for accuracy: higher values will result in better reconstruction at the expense of selecting more tSNPs. The *number of eigenSNPs* parameter determines the amount of structure within each window. A lower number is equivalent to more stringent window definitions, and thus the SNPs within a window are strongly correlated. As a result the reconstruction error is low. However, our choice for the number of eigenSNPs also impacts the number of selected tSNPs, since a small value results in smaller windows and thus sacrifices potentially meaningful correlations across contiguous windows. These two parameters determine the tradeoff between the reconstruction accuracy and the number of selected tSNPs.

Selecting Tagging SNPs and Predicting Tagged SNPs

Once windows have been defined, we proceed to identify a small set of tSNPs which retain most of the genetic variance within a window. In pairwise correlation-based tagging each marker is represented by a single tag (Carlson et al., 2004; DeBakker et al., 2005). PCA-based methods use all the tags to predict every SNP within a window. We employ the MultipassGreedy algorithm described in (Paschou et al., 2007) in order to identify tSNPs (it is included as tSNPsMultiPassGreedy Algorithm in supplementary material for convenience). Finally, in order to predict the tagged SNPs within a window using the tSNPs we solve a least squares problem. Given a training set (where all SNPs are known) and a test set (where only the tSNPs have been assayed) we express each tagged SNP in the training set as a linear combination of the tSNPs. We then use these coefficients to predict tagged SNPs in the test set as linear combinations of the tSNPs in the test set (notice that the latter SNPs are known). A concise definition of the reconstruction algorithm is presented as ReconstructUnassayedSNPs algorithm in supplementary material. We measure and report the reconstruction accuracy of the tagged SNPs in the test set.

Results

Performance Over the HapMap Dataset

In order to validate the scalability of our approach for tSNP selection over genome-wide datasets, we analyzed the HapMap phase 2 dataset. We divided the samples in each of the three HapMap populations in a 90% training set and a 10% test set;

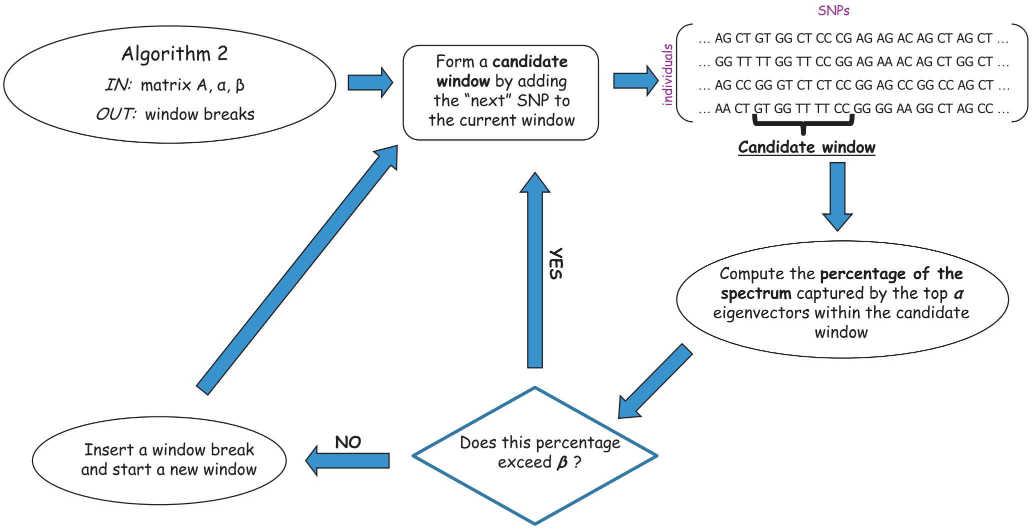
Algorithm 2
IN: matrix A, α, β
OUT: window breaks
Form a candidate window by adding the "next" SNP to the current window
SNPs
individuals
... AG CT GT GG CT CC CG AG AG AC AG CT AG CT ...
... GG TT TT GG TT CC GG AG AA AC AG CT GG CT ...
... AG CC GG GT CT CT CC GG AG CC GG CC AG CT ...
... AA CT GT GG TT TT CC GG GG AA GG CT AG CC ...
Candidate window
Compute the percentage of the spectrum captured by the top α eigenvectors within the candidate window
YES
Does this percentage exceed β ?
NO
Insert a window break and start a new window

Figure 1 A flow-chart depiction of our block-definition algorithm to define windows within a chromosome.

a 70–30% training-test split showed marginal, less than 0.2%, variation from these results (data not shown). The training set was used to define windows and identify tSNPs. Only the selected tSNPs were subsequently extracted for the samples in the test set. These tSNPs were then used to predict genotypes at the remaining SNPs. For this exceptionally large experiment, genotype imputation was performed using a simple regression-based technique which we have previously proposed (Paschou et al., 2007). This significantly reduced computational time rendering the experiment feasible. As we will show later, our tSNP selection algorithm can be easily coupled with more sophisticated genotype imputation methods, albeit at a considerable computational cost which would render the experiment described in this section, infeasible.

Figure 2 reports results over five random training-test set splits for each chromosome, for each of the three HapMap populations. The accuracy parameter for our window definition was set to either 95% or 98% and the number of eigenSNPs was set to 10, 15, or 20. In the European and Asian populations, when the 98% accuracy parameter is chosen, we typically achieve less than 5% genotype reconstruction error selecting only 10–15% of the more than 2.2 million SNPs as tagging. The genetically more diverse Yoruban population required 20–30% of the markers in order to achieve the same accuracy. In general, our results indicate that panels of carefully selected SNPs amounting to 5–10% of the total HapMap SNPs, can be used to predict unknown genotypes with more than 90% accuracy. In Supplementary Table S1 we show the number of tSNPs selected for each chromosome and each population, as well as the total number of SNPs analyzed in each case and the corresponding prediction error (parameters: 98% accuracy, 20 eigenSNPs). Chromosomes 2, 6, and 8 are the easiest to predict in all three studied populations (over 95% prediction accuracy with about 10% of SNPs used as tagging in Europeans and Asians, and 20% in Africans), while chromosomes 16, 17, and, 19 prove particularly hard to reconstruct (about 16–21% of SNPs needed in Europeans and Asians and as many as 30–32% of SNPs needed in Africans in order to reach 95% prediction accuracy, see Supplementary Table S1).

To further demonstrate the accuracy of our approach, we calculated the r^2 correlation coefficient between the true and predicted genotype value of each reconstructed SNP. As shown in Figure 3, our reconstructed or imputed SNPs show strong correlation with their actual counterparts. Even when the less stringent set of parameters is used, the r^2 correlation coefficient between the actual and the reconstructed dataset is always close to or above 0.8. As expected, varying the targeted accuracy bar provides a natural tradeoff between genotyping savings and prediction accuracy (Figs. 2 and 3).

It is interesting to consider for each chromosome the size of the windows defined by our algorithm as exhibiting a high degree of linear structure. Figure 4 shows the size of the windows identified by our algorithm and used for tSNP selection over chromosome 1 (see Supplementary Figs. S1–S3 for similar results for each autosome). Figure 4 also demonstrates how our choice of parameters influences window size. In the Yorubans, on chromosome 1, no windows greater than 500 SNPs exist and the longest window is 438 SNPs long. On the other hand, the longest window observed on chromosome 1 is 838 SNPs (about 1 Mb) in the European and 670 SNPs in the Asian populations. It is no surprise that the high

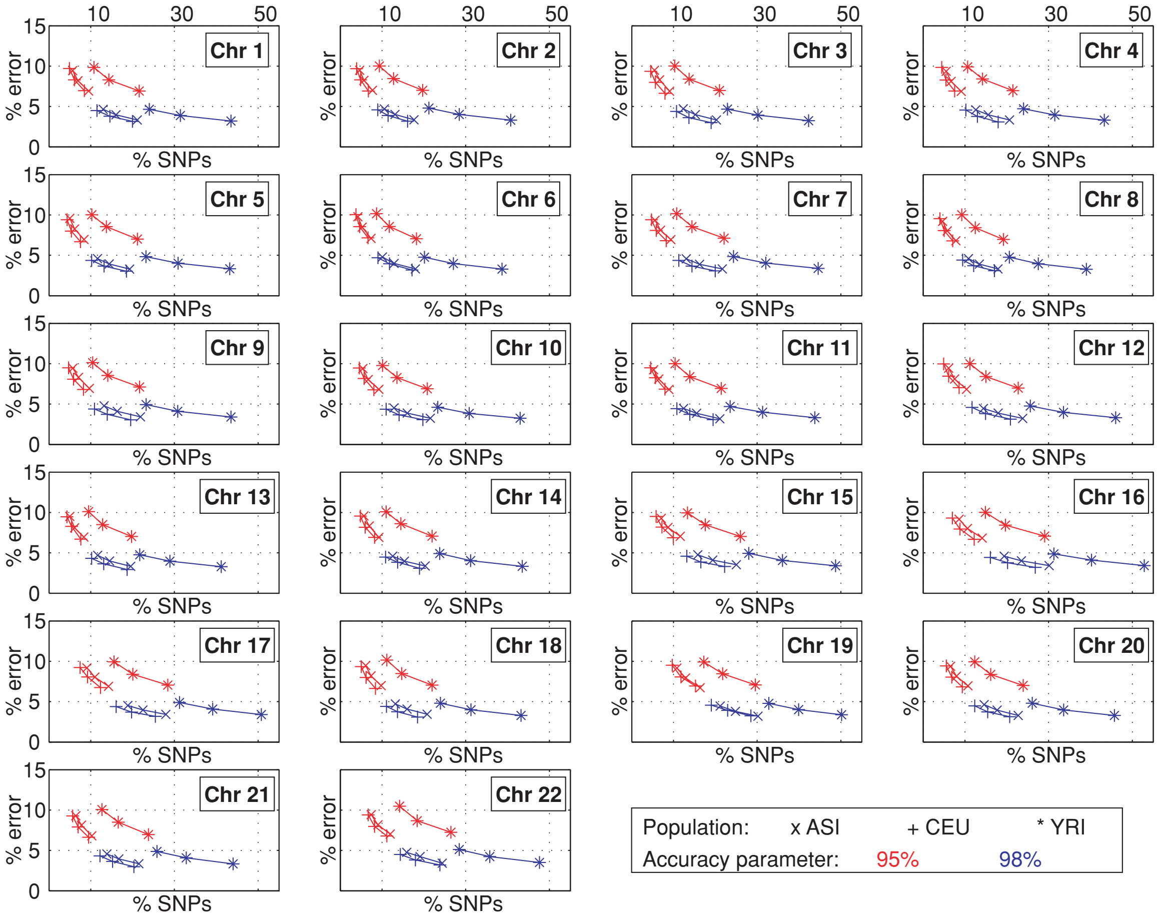

Figure 2 Efficiency-accuracy tradeoff achieved for genotype prediction using tSNPs selected with our approach for all 22 autosomes and the 3 HapMap phase 2 populations. The three points on each curve correspond to different values of the *eigenSNP* parameters (20, 15, and 10 *eigenSNPs* from left to right). The *x*-axis corresponds to the percentage of SNPs in each chromosome selected as tagging (tSNPs) and used for prediction of genotypes in the remaining SNPs.

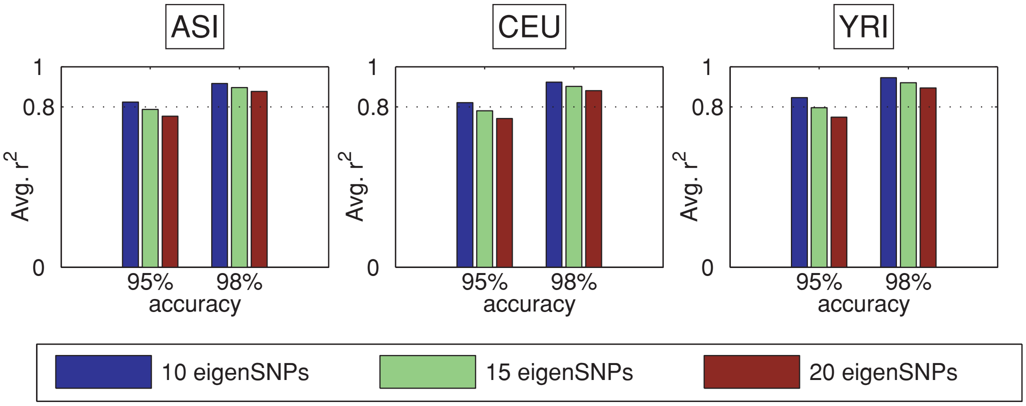

Figure 3 Average r^2 correlation between reconstructed and actual test set SNPs for chromosome 1. A line at $r^2 = 0.8$ is drawn to highlight parameter combinations which do better.

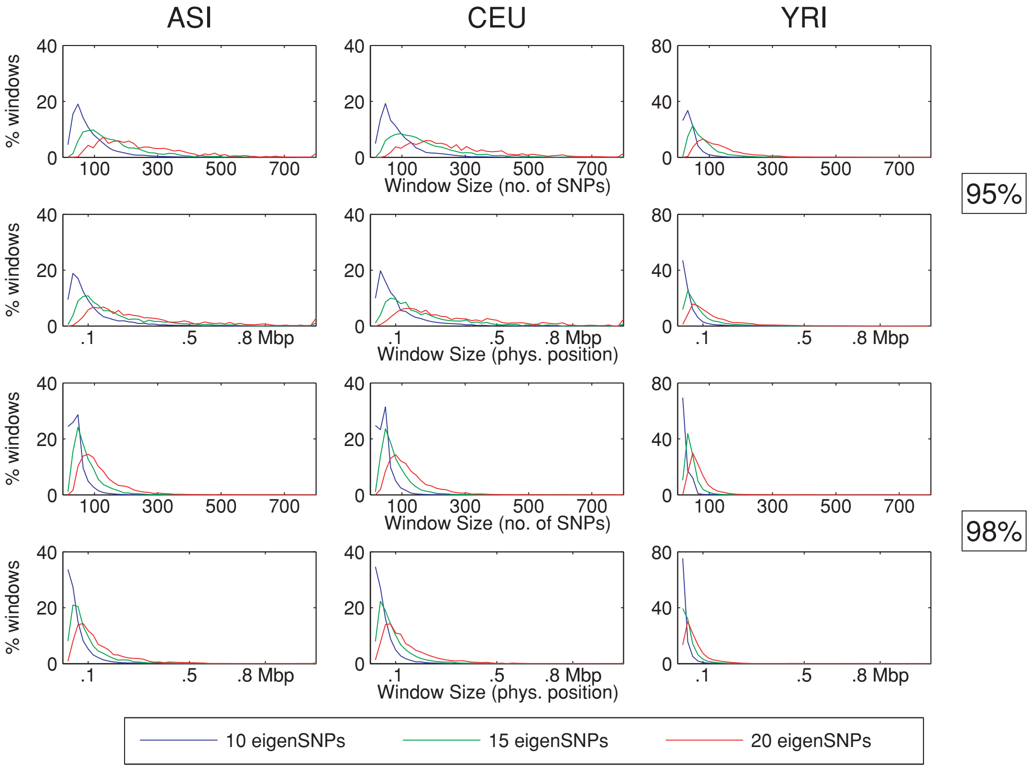

Figure 4 Histogram of window sizes for chromosome 1 in terms of number of SNPs and physical position. The top and bottom half represent 95% and 98% targeted *accuracy*, respectively. Within each subfigure, the three curves illustrate the impact of varying the *eigenSNP* parameter in our analysis.

diversity of the Yoruban population results in a high percentage of short windows. Considering the parameters of 20 eigenSNPs and 98% accuracy, about 87.2% of windows on chromosome 1 for the Yoruban population consist of less than 100 SNPs (roughly less than 100 kb), while the corresponding number for the Europeans and Asians is 47.2% and 50.6%, respectively. Equivalently, 73% of the SNPs in chromosome 1 are assigned to windows of size at most 100 kb in the Yoruban population. This number drops to 29% for the European population and 28% in the Asian populations (Supplementary Table S2). Comparable percentages are observed over the entire genome (Supplementary Table S3).

Since rare SNPs are well known to be hard to tag, we dug deeper into our results in order to evaluate our performance in this subset of SNPs. Within each window, SNPs were categorized based on their rare allele frequency (RAF) and the error for each category in the test set was computed. Our results show that in such cases, the accuracy parameter in our window definition needs to be set higher (e.g., 99%), which results in the selection of a larger number of tSNPs (see Supplementary Fig. S4). Clearly, an error greater than 5% is unreasonable for SNPs with RAF $\leq$ 5%. A trivial algorithm which simply predicts the frequent allele will give comparable performance. If rare SNPs are the focus of a study the tSNP approach for genotype prediction does not seem appropriate. Alternatively, perhaps larger datasets comprising of many more individuals should be used as reference (Browning & Browning, 2009).

Interpopulation Prediction Using the HapMap Phase 3 Dataset

We evaluated the portability of our selection of tSNPs across the 11 populations of HapMap phase 3. For this purpose, the SNPs that were polymorphic in all populations were extracted. Windows were defined and tSNPs were selected from each of the populations in turn. The selected tSNPs were considered "assayed" in the other populations and the remaining tagged SNPs were predicted (Fig. 5). As expected, populations within a continent provide a good reference panel for each other. Interestingly the Indian population is genetically much closer to the Europeans than the geographically neighboring East Asians. This could be attributed to the Himalayas acting as a deterrent for migration across the two sides, thus allowing genetic drift to act independently on each side.

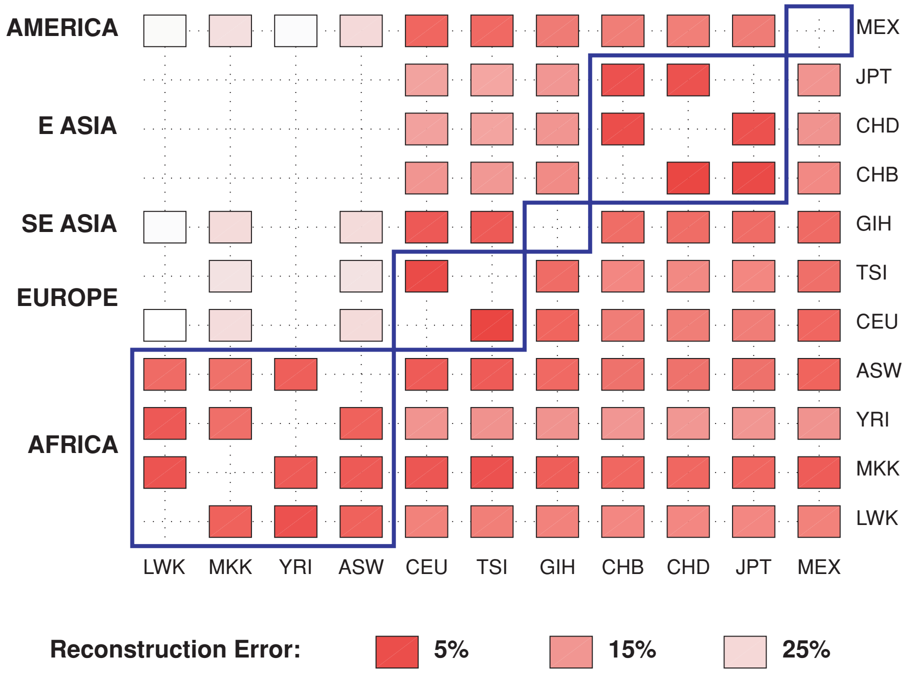

Figure 5 Interpopulation reconstruction error of HapMap phase 3 populations using parameters 98% accuracy and 20 eigenSNPs. Each population is used in turn to predict all others. Each row corresponds to a *predicting* population and each column a *predicted* population. The rows and columns are sorted based on geographic origin of the population (Africa, Europe, South-East Asia, East Asia, Americas). Darker entries represent lower reconstruction error. Entries with reconstruction error greater than 30% are left blank.

Similar results were observed in a study of Y-chromosome data for the neighboring Pakistani populations (Qamar et al., 2002) as well as in studies of genome-wide data (Li et al., 2008). The Mexican population because of its Spanish colonial history is significantly impacted by the Europeans as well.

A point of emphasis here is that, unlike previous studies (González-Neira et al., 2006; Paschou et al., 2007), tSNP panels selected using the African populations can be used as good predictors of genotypes even for populations in different continents. However, a genetically closer neighboring population would do better. These results can also be interpreted in light of the widely supported "out of Africa hypothesis," which postulates the fact that modern humans first originated in Africa and from there they migrated to the rest of the world. Short, highly correlated genetic fragments in the African populations seem to still retain the signature of these ancestral populations painting a representative picture of human genetic variation around the world. On the other hand, non-African populations lack the full breadth of this variation due to founder effects. Obviously, more recent mutations within individual populations are not reflected in the original one. These are more likely to be observed in neighboring populations due to admixture.

Comparison with Tagger

First, we compared the efficiency and accuracy of our large-scale tSNP selection approach to results obtained using Tagger, an LD-based tSNP selection algorithm (DeBakker et al., 2005). Tagger is publicly available in Haploview (Barrett et al., 2005) and has the same motivation as Carlson's LD-select (Carlson et al., 2004). A key difference is that it allows multimarker tags to improve efficiency. The size of multimarker tags in the publicly available implementation is restricted to at most three. We found that running Haploview with chromosome-wide data was computationally infeasible, so we tested the ENCODE regions of HapMap. These regions have been a particular focus of the HapMap project and have a very high SNP density. We should note that Tagger is a memory intensive algorithm, and its memory requirements varied significantly among different regions. In some cases even 20 GB of process memory were not sufficient for this algorithm. This was a key limiting factor restricting the comparison to seven of the ten ENCODE regions. On the other hand, our algorithm runs without any problems on a commodity desktop with just one gigabyte of memory.

In order to compare the two approaches, we divided the ENCODE datasets into 90% training set and 10% test set. Once again, we used the following set of parameters for our algorithm: we set the accuracy to 95% and 98% and the number of eigenSNPs to 10, 15, and 20. In each case, we reconstructed the tagged SNPs in the test set and computed the average reconstruction error. Both our algorithm and Tagger were each time run on the same training set in order to identify a set of tSNPs. If run unimpeded, Tagger selects around a third of the SNPs as tSNPs. For a fair comparison, we restricted Tagger to pick exactly the same number of tSNPs as our approach. In order to reconstruct SNPs in the test set using the Tagger tSNPs, we matched the genotypes of each tagged SNP to the genotype of its corresponding tSNP. For multimarker tags, Haploview infers haplotypes from genotype data using the partition ligation expectation maximization algorithm (Qin et al., 2002). This software is publicly available as PL-EM. We applied this implementation on neighboring tSNPs to infer the haplotypes of heterozygous tag combinations.

Figure 6 shows the accuracy and coverage achieved by each method on ENm010.7p15.2 ENCODE region. For a performance comparison on all seven ENCODE regions that we studied see Supplementary Figure S5. Since we restricted the number of tSNPs selected by Tagger to be equal to the number of tSNPs selected by our approach the coverage achieved by Tagger is not complete (SNPs exist which are not tagged by the selected subset). On the other hand, our method always provides complete coverage of the analyzed regions with a small number of tSNPs. For instance, consider the performance of each method for CEPH Europeans in the region analyzed in Figure 6. Using 2% of the total SNPs as tagging our method provides approximately 84% prediction accuracy, covering the region completely, while the Tagger tSNPs can only cover 30% of the studied SNPs, for which the prediction accuracy is approximately 94%. In fact, for the CEPH European population, Tagger restricted to the same number of SNPs as our method can only achieve 95% coverage.

Comparison with Other Block-Definition Methods

Having demonstrated the scalability as well as the accuracy of our method for tSNP selection and subsequent genotype reconstruction across the genome, we proceeded to compare the efficiency of our algorithm for the definition of windows/blocks across the genome to previously described methods for block definition. Our objective was to compare the total number as well as the size of the defined windows that are defined by each one of the methods we studied, as well as to examine the total savings in tSNP selection, when each approach is used. Once again, the ENCODE data (nine regions) was used in order to compare our method to two popular methods of block definition: HapBlock (Zhang et al., 2005) and the Gabriel et al. method (Gabriel et al., 2002) as implemented in Haploview (Barrett et al., 2005).

The algorithm of (Zhang et al., 2002) relies on phased haplotypic data and even though it was later extended to unphased data (Zhang et al., 2004), the publicly available implementation does not scale to the size of ENCODE regions (which exceed the 700 SNPs threshold). Thus, we first phased each region using Beagle (Browning & Browning, 2007) and then used the original algorithm to assign block boundaries. HapBlock parameters were set to identify blocks such that the common haplotypes account for more than 80% of the observed haplotypes; the threshold for common haplotype was set to 10%. We tested two different tSNP definitions in order to define blocks with HapBlock: the first definition required the tSNPs to capture at least 80% of the common haplotypes, whereas the second required complete coverage. In a similar manner, Haploview was used to identify SNP blocks based on the definitions of Gabriel et al. (2002). Finally, our own method for window definition was applied with the most stringent choice of parameters; the accuracy parameter was set to 98% and two different choices for the number of eigenSNPs parameter were tested (10 and 20). In all cases, once the blocks were defined, our PCA based approach was used to select tSNPs within each block.

Table 1 shows the average size of the windows (over all nine ENCODE regions) using the aforementioned methods for the three HapMap populations. Compared to other methods, the blocks defined by our PCA-based method span longer genomic segments and retain strong linear structure while maximizing the savings. This is evident in the comparison of the performance of the window definitions: even when our most stringent choice in terms of window sizes is used, our method returns longer windows both in terms of the number of SNPs within a window and the physical size of the window. Additionally, a much smaller number of tSNPs is needed in order to cover the targeted regions. In fact, when the eigenSNPs parameter for our method is set to 10 the windows defined are 3–4 times longer than those defined by other methods, and they are up to 10 times longer when this parameter is set to 20. For the Asian and European populations, this is also accompanied by savings of about 20% more in the number of selected tSNPs (out of the total number of SNPs), in comparison to the other methods we studied here. Interestingly, when the Yoruban population is studied, the savings in tSNP selection when our proposed method is used, are greater by about 50% of the total number of analyzed SNPs. Obviously, when a larger number of tSNPs is retained, the reconstruction accuracy will be higher, and there is a tradeoff between the number of tSNPs retained and the prediction

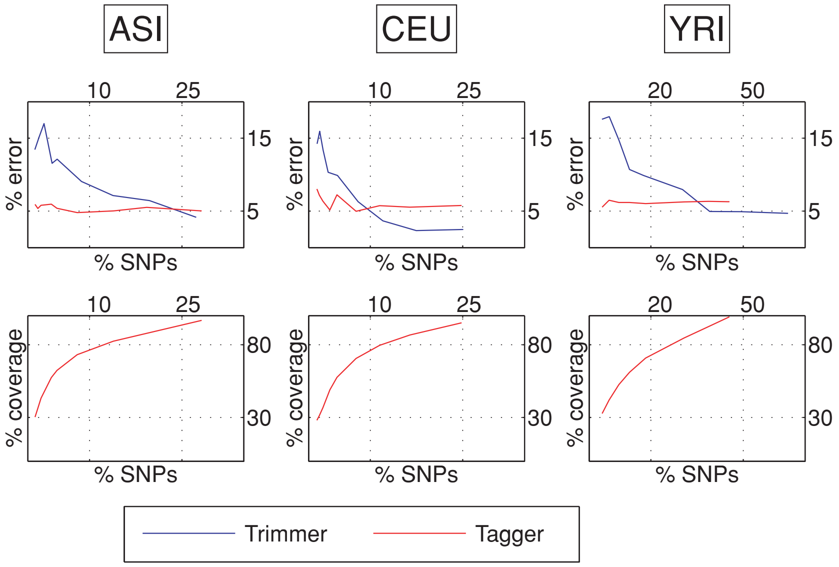

Figure 6 Performance comparison with Tagger analyzing the ENm010.7p15.2 region. Our algorithm was run with nine parameter combinations (90% , 95%, and 98% target accuracy, and 20, 15, and 10 eigenSNPs). The blue line shows percentage of SNPs needed and respective reconstruction error for each of these nine parameter combinations. In each case, Tagger was restricted to the same number of tSNPs as needed by our approach. Coverage corresponds to the percentage of total SNPs captured by Tagger. Our approach provides always perfect coverage and hence it is not plotted. The *x*-axis corresponds to the percentage of SNPs selected as tagging.

accuracy. However, as we have already shown in the previous sections, our algorithm already achieves very low errors in genotype imputation (down to less than 5% in European populations for the ENCODE regions) while also achieving great genotyping savings. In the next section, we proceed to demonstrate that the proposed method can be successfully applied in the setting of a real genomewide association study in order to considerably reduce the number of SNPs needed to uncover true associations.

Applicability on Genome-Wide Association Studies

We next validated the usefulness of our genomewide tSNP selection approach in data from a genome-wide association study. In doing so, we undertook the challenging task of identifying redundancy in a SNP panel that has been explicitly designed by Illumina in order to cover the entire genome based on a tSNP approach. At the same time, we evaluated the HapMap CEU population as a reference population for genome-wide genotype prediction in European American samples. During this experiment, we also evaluated the performance of the tSNPs selected with our method for genotype imputation using a sophisticated haplotype inference based method, as implemented in the software Beagle (Browning & Browning, 2009). Notice that in all previous experiments, genotype imputation was performed using a simple regression-based algorithm that we have previously described, which allowed us to efficiently run the exceptionally large number of comparisons already presented here.

We analyzed the Coriell Institute's publicly available dataset for the study of Parkinson's disease. The Parkinson's dataset contains 270 cases and 271 controls assayed for 396,591 SNPs (Fung et al., 2006). All the participants are of European-American ancestry and thus the HapMap CEPH Europeans were used as a reference population. This dataset contains 369,627 SNPs in common with the HapMap CEU dataset. This subset of markers from the reference HapMap European population was used to identify tSNPs and determine prediction coefficients for the tagged ones (for genotype prediction using our simple regression-based method) as well as reference haplotypes (for genotype imputation using Beagle). We targeted 98% accuracy while varying the eigenSNP parameter between 10 and 20. In the original and reconstructed datasets, we compared the Armitage trend test statistic for those SNPs that have also been genotyped for the HapMap CEU population. A *P*-value less than 10^{-4} was set as a threshold for reporting significant correlation with affection status. It should be noted that the *P*-values of (Fung et al., 2006)

Table 1 Comparison of different window/block definition methods. Two different parameter settings were used for Hapblock: (1) required the tSNPs to capture 100% of the common haplotypes and (2) required the tSNPs to capture 80% of the common haplotypes. Two different parameter settings were also used for our software (Trimmer). We indicate the average window size (over all nine ENCODE regions) as well as the percentage of tSNPs (as a fraction of the total number of available SNPs) that were selected by Algorithm 3 (see supplementary material) in each case. Clearly, our window definition results in the minimal number of tagging SNPs.

ASI	Average window length (bp)	tSNPs (%)
Hapblock (1)	7637	25
Hapblock (2)	10,044	25
Haploview	11,993	34
Trimmer ($\alpha = 98\%$, $\beta = 10$)	34,310	14
Trimmer ($\alpha = 98\%$, $\beta = 20$)	101,073	7
CEU		
Hapblock (1)	4935	31
Hapblock (2)	6816	30
Haploview	12,247	27
Trimmer ($\alpha = 98\%$, $\beta = 10$)	29,722	13
Trimmer ($\alpha = 98\%$, $\beta = 20$)	101,512	6
YRI		
Hapblock (1)	1715	69
Hapblock (2)	3025	66
Haploview	474	72
Trimmer ($\alpha = 98\%$, $\beta = 10$)	10,826	41
Trimmer ($\alpha = 98\%$, $\beta = 20$)	43,956	16

cannot be reproduced exactly as a result of our choice to fill in missing entries for illustration purposes throughout this paper.

Figure 7 demonstrates the performance of the selected tSNPs using both a simple regression-based method and the much more sophisticated algorithm implemented in Beagle. It is worth noting that, despite the relatively low density of SNPs in the reference sample (markers in the Illumina chips have been chosen to cover the entire genome), our results still uncover considerable redundancy. When the 98% accuracy and 10 eigenSNPs parameter combination is used, 62% of the SNPs are selected as tagging. Beagle takes more than 30 h in order to impute genomewide genotypes but is very successful in genotype prediction with an error of 3.2%. The regression-based method on the other hand takes less than 1 h yielding an error of 5.2%. Reflecting the more accurate prediction, Beagle produces no false positive associations while the regression-based method results in eleven false positive associations, seven of which were originally weakly associated with the disease ($P < 0.05$). When our more relaxed parameter combination is used (98% accuracy and 20 eigenSNPs), the percentage of selected tSNPs is reduced to 47% of the original dataset. Again Beagle produces more accurate predictions (4.4% prediction error) taking however more than 26 h while the regression-based method takes less than 1 h for a run over the whole genome, albeit at a cost of less accurate predictions (8.2% prediction error). Even with these higher error rates, we were able to recover significant associations with minimal loss in power, using less than half of the total SNPs. The number of false positives increases somewhat as our eigenSNP parameter is relaxed. When Beagle is used for imputation, two false positive results were found, both of which were originally weakly correlated with the disease ($P < 0.05$). Using the regression-based method resulted in 16 false positive results, 10 of which were originally weakly correlated with affection status. In any case, our results indicate that a two-step approach can be a cost-effective design for association studies. Investigating a dense map of SNPs that have been genotyped on a reference population, cases and controls are first genotyped for a carefully selected small panel of tSNPs and untyped SNPs are predicted. In the second step, imputed SNPs that are found associated with the disease are actually genotyped on the case-control sample in order to verify the prediction and eliminate false positive associations.

Discussion

The LD architecture of the human genome can be translated into linear algebraic terms and elucidated by PCA. This is the first study to undertake an evaluation of a PCA-based method for genomewide tSNP selection. At the same time, this study represents a detailed exploration of the linear algebraic structure of our genome. Dividing the genome into segments prior to further analysis is a step that cannot be circumvented with existing techniques. Here, we introduced a novel algorithm that can be used to divide the genome into contiguous windows of high linear structure. This allowed us to efficiently analyze and select tSNPs for approximately 2.5 million SNPs in four populations from three continental regions (available from the HapMap phase 2 data). Coupling our novel definition of genomic windows with a PCA-based method for tSNP selection (Paschou et al., 2007), we show that only 13%, 11%, and 24% of these SNPs suffice to accurately predict the remaining genotypes in the Asian, European, and Yoruban populations, respectively. About 25% of the genome in Europeans, and 29% in Asians can be assigned to relatively short windows of high linear structure (windows of less than 100 consecutive SNPs). In other words, as much as 70–85% of the genome accounts for particularly long and highly structured regions of over 200 kb in these populations. The situation is reversed in the African populations. In the HapMap Yoruba, only 29% of the genome can be assigned to windows that are longer than 100 SNPs (roughly 100 kb).

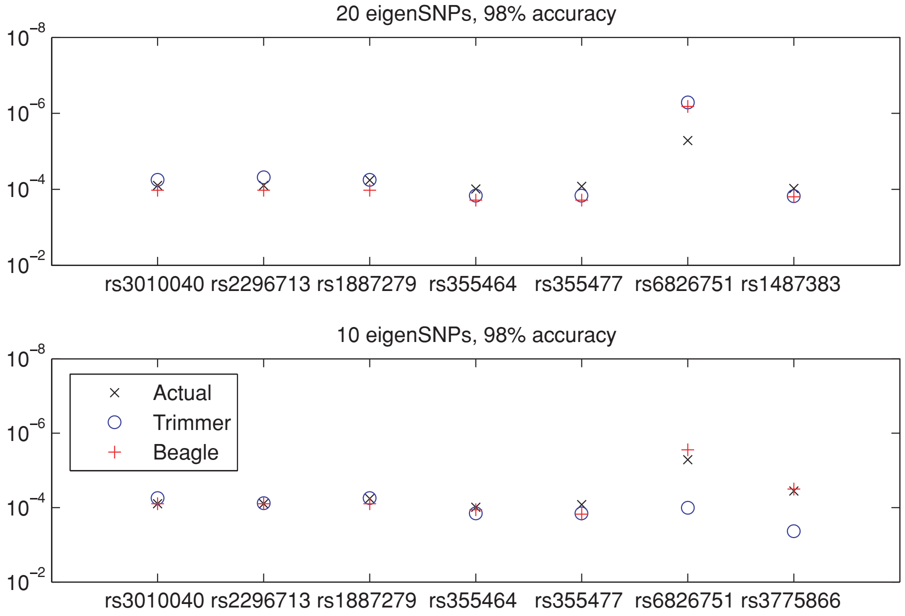

Figure 7 *P*-values of seven SNPs that are significantly associated with Parkinson's disease in the data of Fung et al. (2006). We illustrate the performance of our software (Trimmer) for two different parameter settings: 20 eigenSNPs, 98% accuracy (top) and 10 eigenSNPs, 98% accuracy (bottom). Black "×" symbols correspond to the actual *P*-value using the original data. Blue circles correspond to *P*-values of reconstructed SNPs using tSNPs selected by Trimmer and a simple regression for imputation. Finally, red "+" symbols correspond to *P*-values of reconstructed SNPs using tSNPs selected by Trimmer and Beagle for imputation. Clearly, the tagging SNPs selected by Trimmer achieve high reconstruction accuracy even using naive regression for imputation; Beagle achieves an almost perfect imputation using the SNPs selected by Trimmer (see Results section for more details).

Our approach is extremely scalable and the complete analysis took less than four and a half hours per population per set of parameters to compute over the entire 22 chromosomes. On the other hand, as also shown by our analysis here, LD-based tSNP selection algorithms are much less efficient, computationally expensive, and thus almost impractical for genomewide analyses. Using our PCA-based methods, we achieve greater genotyping savings than the LD-based Tagger software or other methods of block definition that we tested (Zhang et al., 2005; Gabriel et al., 2002).

Our algorithm takes as input the genotypic data and two parameters (accuracy and number of eigenSNPs) that determine the tradeoff between the window size (and eventually the number of selected tSNPs) and the reconstruction accuracy. In order to aid the user in the selection of these parameters, we provide a detailed analysis on different choices of parameter combinations and the resulting tradeoff between efficiency and accuracy for PCA-based tSNP selection. Based on budgetary restrictions and accuracy needs, the user can vary the parameters to sacrifice one for the other. However, we note here that our recommendation for choice of parameters is $\alpha = 20$ eigenSNPs and $\beta = 98\%$ accuracy; according to our analyses, these choices for α and β consistently return high accuracy and significant savings in terms of selected tSNPs and thus are recommended to users of our approach. All of our methods are implemented in a software (TRIMMER) that can be found at http://www.cs.rpi.edu/~javeda/genome_tSNPs.htm along with examples and instructions for the user. We would like to note that the design of nonparametric methods to divide the genome into contiguous or even noncontiguous windows of high linear structure is an interesting open problem for future research.

Our methodology of dividing the genome into contiguous fragments of significant linear structure plays a key role in high prediction accuracy across a diverse set of populations. This is vital for the African populations which are relatively difficult to predict because of their high genetic diversity. Analyzing the HapMap phase 3 dataset, we investigated the similarity in structure and the transferability of tSNPs across 5 geographic regions and 11 populations. As expected, and as has been also shown previously (Paschou et al., 2007; González-Neira et al., 2006; Huang et al., 2009), geographically neighboring populations are genetically close most of the time. Interestingly, the African populations are excellent reference populations for all of the samples that were studied here. This could reflect the ancient origins of the African populations, as well as the African origin of all modern populations around the world.

The optimal method for genotype prediction is a subject open to debate and only a small number of comparison studies have been conducted with varying results (Yu & Schaid, 2007; Pei et al., 2008). Existing sophisticated methods of genotype imputation are computationally intensive, requiring a haplotype inference step or recombination rate maps (Pei et al., 2008; Nothnagel et al., 2009). Our experiment of testing 2.5 million SNPs in three populations and splitting this dataset multiple times in training and test sets was only made possible through the use of a simple regression-based algorithm for genotype prediction. However, as we have also demonstrated here, the tSNPs selected with our methodology can be easily used for genotype imputation with more sophisticated methods than simple regression and such methods will actually produce more accurate results at the expense of a higher computational cost. We chose Beagle for our comparisons since it does not require a recombination rate map and has been shown to have comparable performance with other popular algorithms (i.e., Impute) (Marchini et al., 2007). So we expect the success of our selected tSNP panels to extend to genotype imputation using haplotype-based algorithms. It is worth noting that the PCA-based algorithms proposed here for tSNP selection and genotype prediction can be easily applied on populations that have not been studied in detail, without demanding laborious efforts for the inference of haplotypes or the construction of genetic maps.

We validated our approach in the setting of a real genomewide disease association study for Parkinson's disease. Our results indicate that the HapMap CEU population can be used as a satisfactory reference for these European American samples. Reconstructed datasets are a useful tool in order to identify candidate regions and SNPs for further analysis (Paschou et al., 2007; Huang et al., 2009; Browning & Browning, 2009). We propose a two-step approach similar to (Hirschhorn & Daly, 2005). In the first step, the participants are assayed for a carefully selected panel of tSNPs. An initial disease association study is conducted to identify SNPs correlated with the disease. Imputed markers which are found to be correlated with the disease are then genotyped for all participants in a follow-up study in order to boost the power of associations and to prune out the false correlations.

Our analysis of the HapMap phase 2 dataset reveals a few exceptionally large windows even in the African population (greater than 400 SNPs). These large windows often represent short genomic regions where a significantly higher number of SNPs have been assayed in HapMap because of their biological significance. Still, there exist windows which span very long genomic fragments. Large window sizes indicate low genetic variation in the underlying region. These highly structured windows could correspond to genic regions of low variability, indicating natural selection factors at play, as is the case for the 2 Mb region around the *LCT* gene on chromosome 2 in Europeans (Sabeti et al., 2007). Indeed, in a follow-up study, we intend to focus on such regions as candidates for selection, comparing them across different populations and investigating the possibility to uncover additional genes that are important for human adaptation to diverse environments around the world.

Next generation sequencing (NGS) technologies now offer the possibility to interrogate the entire genome including rare variants for association with disease. It is still quite expensive to apply such technologies to the large number of samples that are needed in order to identify genetic susceptibility to complex traits. However, two-stage studies with a first step of variant discovery through NGS in a smaller sample, and a second step of tSNP genotyping and genotype imputation in a large sample can be envisaged (Siu et al., 2011), as well as novel methods for testing association that will allow the use of probabilistic rather than exact genotypes (Zawistowski et al., 2010). Many questions, including the ability to impute rare variants when such datasets are used as reference, remain to be addressed. The 1000 Genomes Project aims to resequence the genomes of at least 1000 unrelated individuals across several populations (1000 Genomes Project Consortium, 2010) and will provide a valuable resource to develop the next generation of tSNPs sets that will become important tools in future genetic association studies.

Acknowledgements

PD and AJ were supported by an NSF CAREER Award to PD and EMBO Short Term Fellowship. PP was supported by a Tourette Syndrome Association, USA Award and COST Action BM0905.

References

Barrett, J. C., Fry, B., Maller, J. & Daly, M. (2005). Haploview: Analysis and visualization of ld and haplotype maps. *Bioinformatics* **21**(2), 263–265.

Browning, B. & Browning, S. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**(2), 210–223.

Browning, S. R. & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084–1097.

Carlson, C., Eberle, M., Rieder, M., Yi, Q., Kruglyak, L. & Nickerson, D. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* **74**(1), 106–120.

1000 Genomes Project Consortium 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073.

The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1 human genome by the ENCODE pilot project. *Nature* **447**, 799–816.

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.

Daly, M., Rioux, J., Schaffner, S., Hudson, T. & Lander, E. (2001). High-resolution haplotype structure in the human genome. *Nat Genet* **29**(2), 229–232.

DeBakker, P., Yelensky, R., Pe'er, I., Gabriel, S., Daly, M.& Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat Genet* **37**(11), 1217–1223.

Fung, H.-C., Scholz, S., Matarin, M., Simon-Sanchez, J., Hernandez, D., Britton, A., Gibbs, J. R., Langefeld, C., Stiegert, M. L., Schymick, J., Okun, M. S., Mandel, R. J., Fernandez, H. H., Foote, K. D., Rodriguez, R. L., Peckham, E., Vrieze, F. W. D., Gwinn-Hardy, K., Hardy, J. A. & Singleton, A. (2006). Genome-wide genotyping in Parkinson's disease and neurologically normal controls: First stage analysis and public release of data. *Lancet Neurol* **5**(11), 911–916.

Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M. & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**(5576), 2225–2229.

Gonzalez-Neira, A., Ke, X., Lao, O., Calafell, F., Navarro, A., Comas, D., Cann, H., Bumpstead, S., Ghori, J., Hunt, S., Deloukas, P., Dunham, I., Cardon, L. & Bertranpetit, J. (2006). The portability of tagsnps across populations: A worldwide survey. *Genome Res* **16**(3), 323–330.

Halldorsson, B., Bafna, V., Lippert, R., Schwartz, R., DeLaVega, F., Clark, A., & Istrail, S. 2004a. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Gen Res* **14**(8), 1633–1640.

Halldorsson, B., Istrail, S. & DeLaVega, F. 2004b. Optimal selection of SNP markers for disease association studies. *Hum Hered* **58**(3–4), 190–202.

Hirschhorn, J. & Daly, M. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**(2), 95–108.

Horne, B. & Camp, N. (2004). Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genet Epidemiol* **26**(1), 11–21.

Huang, L., Li, Y., Singleton, AB., Hardy, JA., Abecasis, G., Rosenberg, N. A. & Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* **84**, 235–250.

Johnson, G., Esposito, L., Barratt, B., Smith, A., Heward, J., DiGenova, G., Ueda, H., Cordell, H., Eaves, I., Dudbridge, F., Twells, R., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S., Clayton, D. & Todd, J. (2001). Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**, 233–237.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104.

Lin, Z. & Altman, R. (2004). Finding haplotype tagging SNPs by use of principal components analysis. *Am J of Hum Genet* **75**, 850–861.

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**(7), 906–13.

Meng, Z., Zaykin, D., Xu, C., Wagner, M. & Ehm, M. (2003). Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet* **73**(1), 115–130.

Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. & Franke, A. (2009). A comprehensive evaluation of SNP genotype imputation. *Hum Genet* **125**, 163–171.

Paschou, P., Mahoney, M. W., Javed, A., Kidd, J. R., Pakstis, A. J., Gu, S., Kidd, K. K. & Drineas, P. (2007). Intra- and interpopulation genotype reconstruction from tagging SNPs. *Genome Res* **17**(1), 96–107.

Paschou, P., Drineas, P., Lewis, J., Nievergelt, C., Nickerson, D., Smith, J., Ridker, P., Chasman, D., Krauss, R. & Ziv, E. (2008). Tracing sub-structure in the European American population with PCA-informative markers. *PLoS Genet* **4**(7), e1000114.

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., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. & Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723.

Pei, Y.-F., Li, J., Zhang, L., Papasian, C. J. & Deng, H.-W. (2008). Analyses and comparison of accuracy of different genotype imputation methods. *PLoS ONE* **3**(10), e3551.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909.

Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A., Zerjal, T., Tyler-Smith, C. & Mehdi, S. Q. (2002). Y-chromosomal dna variation in pakistan. *Am J Hum Genet* **17**(5), 1107–1124.

Qin, Z. S., Niu, T. & Liu, J. S. (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet* **71**(5), 1242–7.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., Galver, L. M., Fan, J. B., Gunderson, K., Murray,

S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Johnson, T. A., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P.,Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R. & Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918.

Siu, H., Zhu, Y., Jin, L. & Xiong, M. (2011). Implication of next-generation sequencing on association studies. *BMC Genom* **12**, 322.

Stram, D. (2004). Tag SNP selection for association studies. *Genet Epidemiol* **27**(4), 365–374.

Yu, Z. & Schaid, D. J. (2007). Methods to impute missing genotypes for population data. *Hum Genet* **122**(5), 495–504.

Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S. & Zöllner, S. (2010). Extending rare-variant testing strategies: Analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* **87**, 604–617.

Zhang, K., Deng, M., Chen, T., Waterman, M. S. & Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA* **99**, 7335–7339.

Zhang, K., Qin, Z. S., Liu, J. S., Chen, T., Waterman, M. S. & 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., Waterman, M. & Sun, F. (2005). HapBlock: Haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics* **21**, 131–134.

Supporting Information

Additional supporting information may be found in the online version of this article:

Algorithm S1 The ENCODE algorithm
Algorithm S2 Window definition
Algorithm S3 The tSNPsMultiPassGreedy algorithm
Algorithm S4 The ReconstructUnassayedSNPs algorithm
Table S1 Selection of tSNPs and prediction of tagged SNPs in each of the 22 autosomes in the HapMap populations (results shown for analysis using parameters of 20 eigenSNPs and 98% accuracy). The total number of polymorphic SNPs for each population and chromosome is also reported.
Table S2 Percentage of SNPs in chromosome 1 lying within windows of given physical size (base pairs) for the parameter combination 98% *accuracy* and 20 *eigenSNPs* in HapMap phase 2 data.
Table S3 Percentage of SNPs across all autosomes lying within windows of given size (number of SNPs) for the parameter combination 98% *accuracy* and 20 *eigenSNPs* in HapMap phase 2 data.
Table S4 Results computed using the GWAS data for Parkinson's disease. We extracted the common SNPs between the dataset under study and the HapMap phase 2 CEU data; we used the latter data to identify tSNPs and to compute prediction coefficients. The table depicts the percentage of SNPs selected as tSNPs and the error in tagged SNPs for each input parameter combination. False positives are the number of spurious associations with P value $\leq 10^{-4}$ in the reconstructed dataset.
Figure S1 An overview of our approach showing the interplay between Algorithms 2, 3, and 4.
Figure S2 Histogram of window sizes in terms of number of SNPs for all autosomes in the Asian population. Results for all six parameter combinations of accuracy and number of eigenSNPs, used in the analysis, are shown.
Figure S3 Histogram of window sizes in terms of number of SNPs for all autosomes in the European population. Results for all six parameter combinations of accuracy and number of eigenSNPs, used in the analysis, are shown.
Figure S4 Histogram of window sizes in terms of number of SNPs for all autosomes in the African population. Results for all six parameter combinations of accuracy and number of eigenSNPs, used in the analysis, are shown.
Figure S5 Prediction error distribution among SNPs with varying rare allele frequencies (RAF) in chromosome 1 datasets. The two rows represent different *accuracy* parameters used.
Figure S6 Performance comparison with Tagger analyzing the ENCODE regions. For each region, our algorithm was

run with nine parameter combinations (90% , 95% , and 98% target accuracy, and 20, 15, and 10 eigenSNPs). The blue line shows percentage of SNPs needed and respective reconstruction error for each of these nine parameter combinations. In each case, Tagger was restricted to the same number of tSNPs as needed by our approach. Coverage corresponds to the percentage of total SNPs captured by Tagger. Our approach provides always perfect coverage and hence it is not plotted. The *x*-axis corresponds to the percentage of SNPs selected as tagging. The seven subfigures correspond to (A) region ENm010.7p15.2, (B) region ENm014.7q32.33, (C) region ENr112.2p16.3, (D) region ENr113.4q26, (E) region ENr131.2q37.1, (F) region ENr213.18q12.1, and (G) region ENr232.9q34.11.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organised for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received: 11 February 2011
Accepted: 25 July 2011