Supplementary Material and Supplementary Figure Captions

Encoding and handling of missing data

In order to explain the encoding of the data, we shall focus on one population; we will refer to this population as X. Our data on X consist of m subjects; 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. Errors in the assaying procedure result in empty entries in T^X. In order to apply our algorithms, we shall transform this initial data table to an integer matrix A^X, without any information loss. A^X will consist 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 j-th SNP (in alphabetical order). If the genotypic information for the j-th SNP of the i-th 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 Algorithm Encode for a precise statement of this procedure .

A few comments are necessary to better understand our encoding procedure. First of all, notice that the two bases (out of A, C, T, G) that appear in the j-th SNP are known a priori, or may be easily inferred from the data. Also, since every entry of T^X is an alphabetically ordered pair of bases, the combination B_2B_1 never appears. Second, our choice of the integers $+1$, 0, -1 is not arbitrary, but satisfies two criteria. The first criterion is that, in the absence of a better model, we assume that for the all SNPs the distance between the base pairs B_1B_2 and B_1B_1 is equal to the distance between the base pairs B_1B_2 and B_2B_2. We also assume that the distance between the base pairs B_1B_1 and B_2B_2 is twice the above two distances. The rationale behind this choice is that, e.g., B_1B_2 and B_1B_1 are differ by one allele, whereas B_1B_1 and B_2B_2 differ in both alleles. If stronger biological assumptions are available then more informative encodings might be possible. The second criterion is that the results of the algorithms that we shall apply on the data should not change if the encoding of B_1B_1 is modified from $+1$ or -1 (and similarly for B_2B_2). This criterion is equivalent to requiring that changing the signs of the elements in the columns of A^X does not affect our analysis; this is indeed the case for all our algorithms. We emphasize that choosing the right encoding is critical for the success of linear algebraic techniques (such as Principal Components Analysis) since the correlation of SNPs is measured as a function of the numeric values used in the encoding.

In order to handle missing data we process each population individually, using a regression-based technique as described in (Alter et al. 2000). Once more we shall focus on a specific population X. First of all, we discard any individual that has more than 10% missing entries. Given A^X, in order to fill in its missing entries, we take advantage of the fact that for every population almost all individuals have no missing entries at all. We denote the set of these subjects by X_f and we denote the submatrix of A^X corresponding to these subjects by A^{X_f}. For every subject i that has missing entries we solve a least-squares regression problem in order to express the non-missing part of the i-th row of A^X as the "best" linear combination of the rows in A^{X_f}. Finally, we reconstruct the i-th row of A as a linear combination of the rows in A^{X_f} using the computed coefficients, and we fill in the missing entries in the i-th row of A^X by rounding the corresponding entries in the reconstructed i-th row to the nearest integer in $\{-1, 0, +1\}$.

```
Data   : $m \times n$ table $T^X$ of biallelic SNPs for population $X$
Result : $m \times n$ matrix $A^X$
for $j = 1 \ldots n$ do
    Let $B_1$ and $B_2$ be the bases that appear in the $j$-th SNP in alphabetical order;
    for $i = 1 \ldots m$ do
        if $T^X_{ij} == B_1 B_1$ then
            $A^X_{ij} = +1$;
        end
        if $T^X_{ij} == B_1 B_2$ then
            $A^X_{ij} = 0$;
        end
        if $T^X_{ij} == B_2 B_2$ then
            $A^X_{ij} = -1$;
        end
        if $T^X_{ij} == \textit{empty}$ then
            $A^X_{ij} = \text{empty}$;
        end
    end
end
```

Algorithm 1: The Encode algorithm.

Evaluating linear structure

In the following algorithm, δ measures how many entries of A were not exactly represented in the rounded version of A_k. If fewer than 10% of the entries of A are misrepresented we say that A is well-approximated by a rank-k matrix.

```
Data   : $m \times n$ matrix $A^X$
Result : $k$, $\delta$
Compute the SVD of $A^X$;
done = false;
$k = 0$;
while not done do
    $k = k + 1$;
    $U_k = [u^1 u^2 \ldots u^k]$;
    foreach $i = 1 \ldots n$ do
        Find $z_{ij}$ that minimize $\|A^{(i)} - \sum_{j=1}^{k} z_{ij} u^k\|_2$;
        Let $A_k^{(i)} = \sum_{j=1}^{k} z_{ij} u^k$;
    end
    Round the entries of $A_k$ to the nearest integer in $\{-1, 0, +1\}$;
    $N_{nz}$ = the number of non-zero elements in the matrix $A - A_k$;
    $\delta = N_{nz}(A - A_k)/(mn)$;
    if $\delta \leq .1$ then
        done = true;
    end
end
```

Algorithm 2: Algorithm to evaluate linear structure in SNP matrices

Since this algorithm is only used to evaluate the amount of linear structure in our population data, we do not need to provide an implementation that takes into account missing entries.

The tSNPsMultiPassGreedy algorithm

```
Data   : $m \times n$ matrix $A^X$
Result : set $S$ of tSNPs
$S = \{\}$;
$E = A^X$;
done = false;
while not done do
    for $i = 1 \dots n$ do
        $f[i] = 1 - \frac{\|P_{E^{(i)}} E\|_F^2}{\|E\|_F^2}$;
    end
    $S = S \cup \arg\min f$;
    $E = A^X - P_{A_S^X} A^X$;
    Let $\tilde{A}$ be $P_{A_S^X} A^X$ with entries rounded to the nearest integer in $\{-1, 0, +1\}$;
    $N_{nz}$ = the number of non-zero elements in the matrix $A - \tilde{A}$;
    $\delta = N_{nz}(A - \tilde{A})/(mn)$;
    if $\delta \leq .1$ then
        done = true;
    end
end
```

Algorithm 3: The tSNPsMultiPassGreedy Algorithm

In the above, $P_{E^{(i)}} E$ denotes the projection of the matrix E on the subspace spanned the the i-th column of E and similarly $P_{A_S^X}$ denotes the projection of A^X on the subspace spanned by the columns of A^X with indices in S. Computing these quantities is easy via standard Linear Algebra.

The ReconstructUnassayedSNPs algorithm

The following algorithm is used to predict unassayed SNPs.

```
Data   : $m_1 \times n$ matrix $A^X$
Result : $m_2 \times n$ matrix $\tilde{A}^Y$
Run Algorithm 3 to get a set $\mathcal{C}$ of $c$ tSNPs;
Assay the tSNPs on all subjects of $Y$ to get an $m_2 \times c$ matrix $C^Y$;
$W = m_1 \times c$ matrix whose columns are the $c$ columns of $A^X$ whose indices are in $\mathcal{C}$;
$\tilde{A}^Y = C^Y W^+ A^X$;
Round the entries of $\tilde{A}^Y$ to the nearest integer in $\{-1, 0, +1\}$;
```

Algorithm 4: The ReconstructUnassayedSNPs algorithm

Retaining r^2 and allele frequency information after tSNP selection

In order to test whether the tSNPs that we picked manage to retain the LD properties of each region, we calculated the r^2 value for all the pairwise tests (within a five-SNP sliding window) in each of the actual datasets and the datasets that could be reconstructed from the selected tSNPs using standard least squares regression. For clarity of presentation, we only describe results for the four populations of the HapMap datasets. The mean error of the comparisons between r^2 values for each pair in the exact and reconstructed datasets is very small (≤ 0.06 on average for the 90% case) as shown in Suppl. Tables 2 and 3. We also compared the actual allele frequencies for each SNP to the reconstructed allele frequencies (Suppl. Tables 2 and 3), and again found very small errors ($0.01 - 0.02$ for the 90% case), indicating that the reconstructed allele frequencies are very close to the correct ones. It is clear that in general we manage to retain the LD structure of each region as well as the allele frequency information for the common "tagged" SNPs. We did notice, however, that the method is not so powerful when it comes to rare SNPs. In our HapMap dataset, the proportion of SNPs with rare allele frequency less than 0.05 was (averaged over the four HapMap populations) 15% in the HOXB region, 9% in PAH, 7% in 17q25 and 6% in SORCS3. When these were examined separately, the reconstruction of the rare allele was erroneous more often than for common SNPs. In general, rare SNPs appeared even less polymorphic in the reconstructed dataset (data not shown). The difficulty, however, in capturing rare variation may prove to be a general limitation of the tSNPs approach.

r^2 and rarer allele frequencies (RAF) computations

We compute the LD coefficient r^2 between all pairs of SNPs (using a five-SNP sliding window) in the exact and in the reconstructed data. Let $r^2_{ex}(i,j)$ denote the r^2 value between SNPs i and j in the exact data, and let $r^2_{ap}(i,j)$ denote the r^2 value between SNPs i and j in the reconstructed data. The five-SNP sliding window means that for each i between 1 and n, j ranges between $i+1$ and $i+5$. The following statistics are measured on the exact and reconstructed data.

- Mean r^2 (exact): $\frac{1}{N}\sum_{i=1}^{n}\sum_{j=i+1,j\leq n}^{i+5}\sum_{j=i+1}^{n}\left|r^2_{ex}(i,j)\right|$
- Mean r^2 (approximate): $\frac{1}{N}\sum_{i=1}^{n}\sum_{j=i+1,j\leq n}^{i+5}\left|r^2_{ap}(i,j)\right|$
- Mean r^2 error: $\frac{1}{N}\sum_{i=1}^{n}\sum_{j=i+1,j\leq n}^{i+5}\left|r^2_{ex}(i,j)-r^2_{ap}(i,j)\right|$

In the above, n is the number of SNPs in our data and N is the appropriate normalization to compute the means. Regarding the rare allele frequencies computations, we compute (for each SNP) the frequency of the rare allele in the exact and the reconstructed genotypic data, and report (over all n SNPs) the mean error and its standard deviation.

Association study on reconstructed data

For the validation of our methods in an association study we used the dataset available on the IBD5 data release page (`http://www.broad.mit.edu/humgen/IBD5/`). In (Rioux et al. 2001) 11 SNPs in this region are reported to be significantly associated with Crohn disease in the sample of families that we also analysed here. The genotypes for two of those SNPs (IGR2078a_1 and IGR2277a_1) are not included in the publicly available dataset, while a third one (IGR2230a_1), yields results different than those reported in the paper (transmitted/untransmitted ratio is equal to 63 : 29 instead of 67 : 28 in the original paper and $P = 0.0004$ instead of 0.000063). Therefore, we concluded that the data analysed for this SNP in the original paper is not identical to what is publicly available, and we decided to not include this SNP in our analysis. For the remaining eight SNPs, both the transmitted/untransmitted allele ratios and associated P values are exactly reproduced in our analysis and results are shown in Fig. 5.

In the available dataset, 10% of the genotypes were missing. Unlike the population analysis that is presented earlier in this paper, here, we did not fill in the missing values prior to applying our methods, as our algorithms have been developed to handle the existence of missing data and we wanted to be able to reproduce the allele transmission ratios that have been reported in the original paper.

We also examined the possible distortion of Hardy-Weinberg equilibrium when reconstructing a dataset with our methods. This only occurred in a very limited number of SNPs (4.2 on average when targeting 90% of the training set spectral variance and going to virtually no errors when targeting 95% (0.6 SNPs with errors) or 99% (0.07 SNPs with Hardy-Weinberg errors) of the training set spectral variance.

Our analysis also indicates that haplotype inference will likely be accurate in a reconstructed dataset. Five of the SNPs that were significantly associated with the disease (IGR2055a_1, IGR2060a_1, IGR2063b_1, IGR2096a_1, IGR2198a_1) were close enough to build haplotypes (spanning 71Kb). In the original dataset two common haplotypes are found using PHASE (Stephens et al. 2001). Haplotype 11212 appears with frequency 46% and 22121 with frequency 48%. The residual haplotypes are very rare (frequencies less than 1%). Using the reconstructed datasets we manage to very accurately infer the observed haplotypes. When targeting 90% of the training set spectral variance the reconstruction error is approximately 10% on average and the common haplotypes are found in the reconstructed datasets with frequencies of 48% and 48.5% respectively. Targeting 99% of the training set spectral variance, haplotype 11212 is observed with frequency 47% and haplotype 22121 with frequency 48% in the reconstructed dataset.

Supplementary Tables

	HapMap			**Yale**		
		EigenSNPs	ActualSNPs		EigenSNPs	ActualSNPs
SORCS3	YRI	18	26	Yor	14	20
	CEU	11	16	Eur	13	22
	CHB	9	12	SFC	8	11
				TWC	8	11
	JPT	10	14	Jap	8	11

	HapMap			**Yale**		
		EigenSNPs	ActualSNPs		EigenSNPs	ActualSNPs
PAH	YRI	18	27	Yor	14	21
	CEU	12	18	Eur	13	19
	CHB	10	14	SFC	11	16
				TWC	10	13
	JPT	11	14	Jap	9	11

	HapMap			**Yale**		
		EigenSNPs	ActualSNPs		EigenSNPs	ActualSNPs
HOXB	YRI	31	41	Yor	23	32
	CEU	24	32	Eur	22	32
	CHB	18	23	SFC	16	24
				TWC	14	21
	JPT	18	22	Jap	14	19

	HapMap			**Yale**		
		EigenSNPs	ActualSNPs		EigenSNPs	ActualSNPs
17q25	YRI	34	46	Yor	24	34
	CEU	30	40	Eur	22	36
	CHB	22	27	SFC	18	27
				TWC	16	23
	JPT	22	26	Jap	17	24

Table 1: Linear structure statistics targeting 99% of the spectral variance in HapMap populations and their corresponding populations in the Yale dataset.

SORCS3

	Mean r^2	**r^2 error ± std**	**RAF error ± std**
	Approx. (Exact)		
YRI	0.41 (0.42)	0.06 ± 0.09	0.02 ± 0.01
CEU	0.54 (0.54)	0.05 ± 0.07	0.02 ± 0.02
CHB	0.63 (0.61)	0.09 ± 0.13	0.02 ± 0.02
JPT	0.61 (0.59)	0.07 ± 0.10	0.02 ± 0.02

PAH

	Mean r^2	**r^2 error ± std**	**RAF error ± std**
	Approx. (Exact)		
YRI	0.25 (0.23)	0.07 ± 0.15	0.01 ± 0.02
CEU	0.44 (0.42)	0.05 ± 0.08	0.01 ± 0.01
CHB	0.54 (0.49)	0.08 ± 0.11	0.01 ± 0.02
JPT	0.48 (0.41)	0.14 ± 0.18	0.02 ± 0.02

HOXB

	Mean r^2	**r^2 error ± std**	**RAF error ± std**
	Approx. (Exact)		
YRI	0.29 (0.30)	0.04 ± 0.09	0.02 ± 0.01
CEU	0.40 (0.40)	0.04 ± 0.06	0.02 ± 0.01
CHB	0.42 (0.41)	0.06 ± 0.10	0.02 ± 0.02
JPT	0.41 (0.40)	0.06 ± 0.12	0.02 ± 0.02

17q25

	Mean r^2	**r^2 error ± std**	**RAF error ± std**
	Approx. (Exact)		
YRI	0.26 (0.25)	0.05 ± 0.11	0.01 ± 0.01
CEU	0.39 (0.38)	0.04 ± 0.05	0.01 ± 0.01
CHB	0.42 (0.42)	0.06 ± 0.10	0.02 ± 0.02
JPT	0.43 (0.42)	0.05 ± 0.07	0.01 ± 0.01

Table 2: Average pairwise r^2 in exact and reconstructed data, and average pairwise r^2 error using a five SNP sliding window and targeting 90% of the spectral variance of each population. The average Rarer Allele Frequency (RAF) error is also reported.

		Mean r^2	**r^2 error ± std**	**RAF error ± std**
SORCS3		Approx. (Exact)		
	YRI	0.42 (0.42)	0.01 ± 0.02	0.003 ± 0.004
	CEU	0.55 (0.54)	0.01 ± 0.02	0.003 ± 0.005
	CHB	0.63 (0.61)	0.02 ± 0.05	0.003 ± 0.006
	JPT	0.60 (0.59)	0.02 ± 0.04	0.003 ± 0.007

		Mean r^2	**r^2 error ± std**	**RAF error ± std**
PAH		Approx. (Exact)		
	YRI	0.23 (0.23)	0.01 ± 0.02	0.003 ± 0.004
	CEU	0.42 (0.42)	0.01 ± 0.03	0.003 ± 0.005
	CHB	0.50 (0.49)	0.01 ± 0.03	0.003 ± 0.008
	JPT	0.43 (0.41)	0.02 ± 0.04	0.004 ± 0.010

		Mean r^2	**r^2 error ± std**	**RAF error ± std**
HOXB		Approx. (Exact)		
	YRI	0.29 (0.30)	0.01 ± 0.02	0.003 ± 0.004
	CEU	0.40 (0.40)	0.01 ± 0.02	0.003 ± 0.004
	CHB	0.41 (0.41)	0.01 ± 0.03	0.002 ± 0.006
	JPT	0.41 (0.40)	0.01 ± 0.03	0.002 ± 0.005

		Mean r^2	**r^2 error ± std**	**RAF error ± std**
17q25		Approx. (Exact)		
	YRI	0.25 (0.25)	0.01 ± 0.02	0.002 ± 0.004
	CEU	0.39 (0.38)	0.01 ± 0.02	0.003 ± 0.004
	CHB	0.42 (0.42)	0.01 ± 0.03	0.002 ± 0.005
	JPT	0.42 (0.42)	0.01 ± 0.02	0.002 ± 0.005

Table 3: Average pairwise r^2 in exact and reconstructed data, and average pairwise r^2 error using a five SNP sliding window and targeting 99% of the spectral variance of each population. The average Rarer Allele Frequency (RAF) error is also reported.

SORCS3

	YRI	CEU	CHB	JPT
YRI *(10 SNPs)*		32.02 (0.17)	41.14 (0.30)	39.10 (0.23)
CEU *(6 SNPs)*	36.76 (0.11)		**19.48** (0.34)	**24.41** (0.32)
CHB *(4 SNPs)*	44.24 (0.28)	31.25 (0.24)		**16.82** (0.12)
JPT *(5 SNPs)*	44.34 (0.14)	31.34 (0.25)	**12.41** (0.16)	

PAH

	YRI	CEU	CHB	JPT
YRI *(10 SNPs)*		39.87 (0.16)	45.22 (0.30)	44.33 (0.23)
CEU *(6 SNPs)*	36.48 (0.17)		**13.61** (0.23)	**23.20** (0.21)
CHB *(4 SNPs)*	56.93 (0.17)	43.12 (0.26)		**17.50** (0.20)
JPT *(5 SNPs)*	55.88 (0.15)	36.58 (0.22)	**16.05** (0.19)	

HOXB

	YRI	CEU	CHB	JPT
YRI *(17 SNPs)*		37.46 (0.19)	45.36 (0.19)	45.06 (0.19)
CEU *(12 SNPs)*	45.44 (0.32)		**23.55** (0.09)	**24.17** (0.13)
CHB *(9 SNPs)*	50.77 (0.24)	30.99 (0.05)		**17.55** (0.10)
JPT *(8 SNPs)*	52.89 (0.29)	44.33 (0.13)	**19.55** (0.05)	

17q25

	YRI	CEU	CHB	JPT
YRI *(21 SNPs)*		41.72 (0.21)	39.03 (0.35)	44.24 (0.45)
CEU *(17 SNPs)*	40.67 (0.19)		**21.76** (0.31)	**21.50** (0.23)
CHB *(12 SNPs)*	47.56 (0.48)	33.51 (0.26)		**25.57** (0.23)
JPT *(12 SNPs)*	39.08 (0.42)	31.09 (0.16)	**19.55** (0.54)	

Table 4: Interpopulation reconstruction error and, in parenthesis, mean r^2 error, targeting 90% of the spectral variance of the reference population. The entries in **boldface** represent reconstruction error less than 30%.

Supplementary Figure Captions

Supplementary Fig. 1

Populations studied around the world. Abbreviations used are indicated in parentheses.

Supplementary Fig. 2

Number of eigenSNPs (computed with the SVD) and actual SNPs (computed with the tSNPsMultiPassGreedy algorithm) explaining 99% 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 (Africa, Europe, Asia, Micronesia, Americas).

Supplementary Fig. 3

Intra-population 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 90% of the total population size.

Supplementary Fig. 4

Intra-population reconstruction error (ratio of erroneously predicted entries over total number of predicted entries) for each of the 38 Yale dataset populations. The training set size is 90% of the total population size. Populations are ordered (bottom to top) based on geographic regions (Africa, Europe, Asia, Micronesia, Americas).

Supplementary Fig. 5

Pictorial description of the relationship between the tSNPsMultiPassGreedy algorithm and the ReconstructUnassayedSNPs algorithm.

Supplementary Figs. 6a and 6b

Inter-population 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 90% of the population's spectral variance. Blank entries correspond to reconstruction errors larger than 30%. The 5 geographic regions of our study are delimited by the blue boxes. **(a)** PAH and SORCS3 **(b)** 17q25 and HOXB.