# Similarity matrices and clustering algorithms for population identification using genetic data

Daniel John Lawson[*] and Daniel Falush[†]

March 1, 2012

**Abstract**

A large number of algorithms have been developed to identify population structure from genetic data. Recent results show that the information used by both model-based clustering methods and Principal Components Analysis can be summarised by a matrix of pairwise similarity measures between individuals. Similarity matrices have been constructed in a number of ways, usually treating markers as independent but differing in the weighting given to polymorphisms of different frequencies. Additionally, methods are now being developed that better exploit the power of genome data by taking linkage into account. We review several such matrices and evaluate their 'information content'. A two-stage approach for population identification is to first construct a similarity matrix, and then perform clustering. We review a range of common clustering algorithms, and evaluate their performance through a simulation study. The clustering step can be performed either directly, or after using a dimension reduction technique such as Principal Components Analysis, which we find substantially improves the performance of most algorithms. Based on these results, we describe the population structure signal contained in each similarity matrix, finding that accounting for linkage leads to significant improvements for sequence data. We also perform a comparison on real data, where we find that population genetics models outperform generic clustering approaches, particularly in regards to robustness against features such as relatedness between individuals.

## 1 Introduction

An important goal of population genetics is to summarise the relationships between individuals from patterns in heritable molecular data. The most popular approaches for within-species analysis are model-based clustering (e.g. using STRUCTURE (45) or ADMIXTURE (2)) or forming low dimensional visual

[*] Department of Mathematics, University of Bristol, Bristol, BS8 1TW, UK

[†] Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig Germany

summaries of the variation (e.g. Principal Components Analysis, PCA (39)). Although model-based approaches appear quite dissimilar to PCA, they typically identify very similar population structure patterns (44), and have recently been shown to be deeply connected (31). By correctly accounting for the information provided in the sharing of molecular markers, the likelihood used by most popular model-based approaches (including STRUCTURE and ADMIXTURE) can be approximately represented as a similarity matrix between pairs of individuals. This similarity matrix in turn contains the same information as that used by some PCA algorithms, for example SMARTPCA (44).

Similarity matrices can therefore be related to many population genetics methods. These matrices can be used in a two-stage approach to population genetics (e.g. (41; 37; 17; 4; 25; 49; 31)) by first computing the pairwise similarities, and then perform clustering or other analyses on this summary of the dataset. This approach is appealing because it computationally efficient, and can typically be parallelised straightforwardly by using multiple processors to calculate different parts of the matrix. It is therefore an appropriate methodological paradigm for the new genomic era in which over 15 million polymorphic sites can be found within humans (1) who are relatively homogeneous compared to many species. Even computationally efficient algorithms such as ADMIXTURE become challenging to apply to this scale of data, whereas similarity-based approaches require little more computation than an equivalent PCA analysis.

This review has two main goals. The first is to discuss the choice of similarity measure, which has received little attention despite the impact it has on population identification for all similarity-based approaches. Therefore, we will not just review the main choices available, but also perform an empirical evaluation of their clustering performance. We find evidence there is a strong correspondence between the ability of clustering algorithms to correctly identify populations, and the signal-to-noise ratio within the matrix itself. Since PCA approaches can be viewed as operating on a similarity matrix, they will therefore give a significantly clearer picture of population structure if the similarity matrix is correctly chosen. Current model-based admixture models, and other approaches that can be viewed in terms of a similarity matrix, are also described by our approach in terms of the amount of population signal available in them.

When addressing different measures of similarity, we explore the benefits of accounting for linkage disequilibrium and the current methods available for this purpose. We find that such linked methods significantly outperform unlinked approaches, which implies they will be useful for PCA-like analyses. Furthermore, we find evidence that the two leading approaches are exploiting a different signal in the data, leading to the question of how a better, combined measure might be found.

The second goal focuses on the problem of population identification via clustering. We review some standard approaches to clustering, and compare these with the software FineSTRUCTURE (31). This is a similarity approach that is 'model-based' in the sense that it is theoretically equivalent to STRUCTURE under certain conditions. We again perform an empirical evaluation of the methods. We find that some simple approaches can perform very well under simulation conditions. However, only the model-based approach performed well in real data, which contained deviations from a discrete population model, e.g. due to between individual relatedness. Furthermore, the generic approaches that squeezed most information out of the data are also the most sensitive to tuning parameters that are difficult to estimate from the data, leading them to completely fail in some circumstances.

## 2 Similarity measures between genetic sequences

The aim of a genetic similarity measure is to identify pairs of individuals who are 'closely related' by assigning them higher similarity than those who are distantly related. Similarity is measured between two individuals in the sample, with the similarity matrix being formed by combining this information for all pairs of individuals. For many measures, the similarity between pairs of individuals is not just dependent on their own genetic composition but also of that of the rest of the sample. Furthermore, similarity matrices need not be symmetric, that is to say the similarity between individual $i$ and $j$ need not be the same as the similarity between individual $j$ and individual $i$. If the matrix is symmetric and other technical conditions are met then the similarity is also a distance metric (a property that holds for most but not all of the measures we consider here).

There are many possible interpretations of what it means for individuals to be closely related, leading to different matrices being constructed. A simple interpretation of relatedness is the average genetic distance (38). Assuming mutations accumulate in a clock-like manner on a segment of DNA, this distance reflects the minimum time in the past that DNA was present in a single ancestor. The time varies along each chromosome where recombination events in previous generations have spliced together maternal and paternal DNA and in doing so spliced together different genealogical trees. These times when averaged are captured by several distance measures. Two common choices are the 'identity-by-state' (IBS) and the very similar 'Allele Sharing Distance' (Table 1), both of which essentially measure the proportion of sites at which two chromosomes are the same. The measures vary in the way that the two copies of each chromosome possessed by diploid individuals are averaged. However, as explained by Witherspoon Et al, (60), this measure is sensitive to the frequencies of the SNPs used, leading to the confusing result that individuals were often 'less similar' (in this sense, when using small datasets) to individuals in their own population than to individuals in other populations.

An alternative interpretation defines relatedness in terms of differentiation via genetic drift (42; 44). Genetic drift is the change in allele frequency caused by variation between individuals in their number of offspring. On average, the absolute frequency of alleles present at intermediate frequency changes by more than when one allele is rare. As a consequence, differences between individuals at intermediate frequency SNPs provide the strongest evidence of distinct historical patterns of genetic drift. The contribution to the similarity measure is therefore normalised using the allele frequency, which can also be thought of as a correction to make the measure better reflect true information content. After normalisation, intermediate frequency and rare SNPs contribute equally to the discrimination between individuals. As for the allele sharing measure, there are a variety of implementations of the matrix that differ in minor details (Table 1).

EIGENSTRAT (44) and the 'unlinked Coancestry matrix' (31) are both effectively using the genetic drift normalisation. These matrices are important because they effectively contain the same information about population structure that is used by STRUCTURE and ADMIXTURE (under certain technical conditions outlined in the propositions of Lawson et al. (31), which approximately hold for the simulated dataset considered here). Therefore any statements about the information content of these matrices reflect the maximum that can be expected from all model-based algorithms that use the STRUCTURE likelihood.

Given data from sufficient independent genome regions, either of these distance measures should reflect differences in ancestry amongst individuals. In practise it matters which measure is used because, at least within humans, there is a low signal to noise ratio. Most variation is either shared by all populations (because a mutation arose in the long period of prehistory) or is specific to a handful of individuals within a single one (because it occurred recently as human population size expanded). Population structure is present at a range of scales and therefore there is always an advantage to using the most accurate method. As we describe below, we find that methods which normalise the variance consistently give better power to cluster closely related populations than methods that do not.

## 3 The effect of linkage

Although different copies of the same human chromosome typically differ at only 1 site in 1000, the non-repetitive content of the human genome is extremely large ($>$ 2.5 billion bases). Modern commercially available platforms can query more than a million SNPs and full genome sequencing can be used to uncover variation at more than 15 million sites (1). If all of these markers provided independent information, then any of the distance measures above would detect extremely subtle differences in ancestry. However in practise, physically close sites are genetically

Table 1: Similarity measures that do not account for linkage. See Supplementary Section S2 for application references, and for additional non-genetic measures.

| **Name**<br>(3 letter acronym) | **Equations**<br>(summed over SNPs $l = 1 \cdots L$) | **References**<br>(and programs) |
|---|---|---|
| Identity-by-State (IBS),<br>Allele Sharing Distance,<br>SNP edit distance | (a) $1 - \lvert Y_{il} - Y_{jl} \rvert$<br>(b) $Y_{il}Y_{jl} + (1 - Y_{il})(1 - Y_{jl})$<br>(c) $2(Y_{il} - 1/2)(2Y_{jl} - 1/2) + 1/2$<br>$Y_{il} = \{0, 1\}$ if haplotype $i$ {does not have/has} SNP $l$ | PLINK (46)<br>Widely used in applications, see e.g. (18) for a recent overview |
| | **Notes:** In the haploid case does not depend on the distance metric chosen (i.e. we can replace $\lvert Y_{il} - Y_{jl} \rvert$ by $\lvert Y_{il} - Y_{jl} \rvert^d$ for any $d > 0$). For diploids, the variants do have slightly different normalisation. Equation (c) shows the relation to a covariance. | |
| Covariance (COV) | (a) $(Y_{il} - \hat{f}_l)(Y_{jl} - \hat{f}_l)$.<br>$\hat{f}_l$ is the frequency of SNP $l$ | McVean (38) describes this thoroughly, although there are also many other uses |
| | **Notes:** There is a clear historical interpretation (38). This measure has been historically popular for PCA (e.g. (39)). | |
| Normalised Covariance (ESU),<br>Coancestry (CPU) | (a) $\frac{(Y_{il}-\hat{f}_l)(Y_{jl}-\hat{f}_l)}{\hat{f}_l(1-\hat{f}_l)}$<br>(b) $\frac{Y_{il}Y_{jl}}{\sum_{k \neq i} Y_{kl}} + \frac{(1-Y_{il})(1-Y_{jl})}{\sum_{k \neq i}(1-Y_{kl})}$ | EIGENSTRAT (44)<br>ChromoPainter (31) |
| | **Notes:** Equation (a) is used by EIGENSTRAT, and is almost identical (on scaling and rotation) to the ‘unlinked coancestry matrix’ (b) of ChromoPainter, both of which are approximately sufficient statistics for the STRUCTURE likelihood (See Propositions 1-4 of (31)). | |

'linked' leading to high statistical correlation, substantially reducing the amount of information available. In total a single haploid genome undergoes about 36 crossovers per generation, leading to a total recombination rate of about 36 Morgans (30). Correlations between markers exist at all genetic scales but those on the centi-Morgan scale are likely to reflect recent shared ancestry, while those at 0.01cM scale are often shared between all humans.

The number of SNPs within each 0.01cM region is highly variable, mainly because of the enormous variation in recombination rates that is observed at fine scales (40). As a result, naive application of the above distance measures to sequence data will give a disproportional weighting to low recombination regions. Furthermore, statistical approaches that assume each SNP is independent will massively overestimate the information contained in the data.

A simple approach to handling data containing linkage disequilibrium (LD) is to thin it (using e.g. PLINK (46)) so that within each region of a given genetic length there are a small and approximately similar number of SNPs. Thinning has has the attractive side-effect that the dataset might be small enough to approach using computationally intensive methods, usually STRUCTURE (4; 7). There are many possible ways of thinning data (e.g. (10; 54)), but all are bound to discard useful information and we therefore make no attempt to review or evaluate these approaches here. In an attempt to avoid completely discarding SNPs, Patterson et al. (44) implement an approach in which the SNPs are considered in order along the chromosome. The allele vector of each SNP is regressed against the $M$ previous ones, with the residuals from this regression used to represent the *additional* information provided by the extra SNP. Note however, that since the choice of $M$ is arbitrary (and fixed), the effectiveness of the approach is likely to vary between genetic regions and datasets according to SNP density, ascertainment scheme and the shape of genealogies in each region. The authors state that the approach is unlikely to work for full genome sequence data and we find that it does not substantially improve performance in our simulations (see Supplement Section S3 for details).

Recombination breaks up haplotype tracts progressively in each transmission, and therefore recent shared ancestry will result in the sharing of longer haplotype tracts (58) than older ancestry. Instead of simply attempting to correct the statistics designed for unlinked markers, a more promising approach seems to be to detect long tracts of shared ancestry directly. However, there are a number of complications involved, including the ambiguity about haplotype phase which is inherent to the data produced by most current genotypic technologies. This is most simply dealt with by running statistical phasing algorithms (51; 6; 24; 11) but these introduce a substantial number of phase switch errors into the data and are computationally intensive to apply for large datasets. Alternatively, the algorithms for detecting shared haplotype tracts can assume either complete phasing ambiguity, or estimate the level of phasing switch errors while identifying long shared tracts. However these will respectively result in a loss of statistical power or an increase the computational complexity of the algorithm. Most importantly, it is not obvious exactly which statistics about haplotype sharing capture the information most relevant to population clustering and this represents an important area for ongoing research.

A first type of approach attempts to identify tracts of high similarity (i.e. perfect or near perfect identity by state) between pairs of individuals. There are a number of approaches for identifying the very long tracts that closely related individuals share (e.g. PLINK (46), GERMLINE (19), SimWalk (52), and many others) which grew out of pedigree-based quantification of heredity. FastIBD (5) is one of the few efficient inference frameworks to search for relatively short (and hence more 'ancient') tracts of interest for population clustering. For each pair of individuals, this algorithm searches for the $k$ largest tracts that contain sufficient evidence of strong similarity - i.e. that are more similar than expected according to gene frequencies at the loci involved. The similarity measure (we call IBD) is the proportion of the total genetic map that falls within these $k$ tracts. There are a number of intricacies involved with identifying tracts and their boundaries. Furthermore, the algorithm has a number of tuning parameters whose value is likely to affect clustering performance but cannot be estimated from the data in any obvious way, the most important of which is $k$.

An alternative approach is to describe each haplotype in the sample as mixture of other haplotypes, and then to use the similarity between the mixture components to compute a distance. A method for doing this is the FastPHASE Haplotype Sharing (FHS) distance of Jakobsson et al. (28), which models haplotype structure as a mixture of $K$ pools. The similarity at each locus between a pair of individuals depends on whether they are constructed from the same pool, which is calculated using a modification of the popular program FastPHASE (51). This approach potentially captures the information provided by sharing of adjacent markers. However, the algorithm has three substantial limitations, which limit its applicability and performance (which in our simulations is comparable to the better unlinked models as described below). First, the frequencies of the haplotypes in the $K$ pools need to be calculated using a reference panel, which increases computational cost by requiring averaging over the possible choices of panel. Secondly, this approach will give equal weight to SNPs in high and low recombination rate regions and hence will give excessive weight to information in recombination cold spots. Thirdly, by modelling the population as a mix of $K$ pools, it is likely capturing the most ancient genealogical splits at each locus, while since most deep ancestry is shared, recent splits are likely to be more informative about population structure.

The final haplotype approach we consider is chromosome painting (31). Each haplotype is painted (i.e. reconstructed) using the haplotypes of each of the other individuals in the sample as possible donors. The donor in each region is interpreted as the individual with whom the haplotype shares the most recent common ancestor for that stretch of DNA. Switches between donors are interpreted as ancestral relationships changing due to historical recombination. The similarity measure (called CPL) is the number of ‘chunks’ used to reconstruct the ‘recipient’ individual from each ‘donor’ individual, and is asymmetric. By finding the nearest donor individual for each genome region, the algorithm uses information from all of the genome, including those in which there are no immediate neighbours. In regions of the genome where there is no clear nearest neighbour haplotype, the algorithm averages across the multiple individuals who might be closest. The matrix of number of chunks is called the linked coancestry matrix and is produced by the software ChromoPainter (31).

The chromosome painting approach has a number of appealing technical properties which facilitate application to extremely large datasets. Painting is performed using a version of the Hidden Markov Model of Li and Stephens (35). This has two parameters, determining switch rates and the weighting of differences between donor and recipient, both of which can be estimated from the data using the algorithm itself. The painting needs to only be run once for each individual while still appropriately reflecting statistical uncertainty about chunk assignment. Furthermore, assuming that each chunk is independent of the others (an assumption which can be relaxed in practise using a jackknifing procedure implemented within the algorithm), the elements of the coancestry matrix can be modelled statistically using the FineSTRUCTURE clustering algorithm (31) described below, which means that the clustering step also has no tuning parameters for the user to specify. A final attractive property is that if markers are treated as independent, the algorithm gives the unlinked coancestry matrix, which has appealing properties as discussed in the previous section.

These linked approaches are readily parallelisable in principle as different individuals can be independently processed. The only algorithm requiring haplotypes to be pre-phased is ChromoPainter, which has comparable running time to currently available phasing algorithms (51; 6; 24; 11). Both FastIBD and ChromoPainter require a recombination map as input but can be run using physical distances if no such map is available. However FastIBD requires a global recombination parameter to be specified, while ChromoPainter does not.

# 4 Clustering algorithms

The output of all the above methods is a similarity matrix between pairs of individuals, to which a wide variety of standard clustering algorithms can be applied. Xu and Wunch (61) list a large number of possibilities and Lee et al. (33) exam-

Table 2: Similarity measures that utilise linkage.

| **Name**<br>(3 letter acronym) | **Key features** | **References**<br>(and programs) |
|---|---|---|
| 'Ancient' Identity-by-descent (IBD) | Pairwise comparisons are independent<br>SNPs compete to be used in IBD block<br>Sum over SNPs<br>Several tuning parameters | FastIBD (5) in the program BEAGLE. Approach first appeared on `http://dienekes.blogspot.com/2012/01/clusters-galore-fastibd-edition.html` |
| | **Notes:** The number of haplotype pairs to be compared at every iteration and a 'scale factor' must be specified, which combine with an arbitrary penalty for mismatches to create a recombination scale. A fine-scale genetic map is required. As the algorithm uses stochastic estimation therefore repeated runs are required. | |
| FastPHASE Haplotype Sharing (FHS) | 'clusters' compete for similarity each SNP<br>Sum over SNPs<br>Several tuning parameters | Jakobsson et al. (28) form a similarity using the output of FastPHASE (51) |
| | **Notes:** Calculates the probability that two haplotypes originate from the same 'haplotype cluster' (the average of a number of haplotypes that look similar across a particular region). The number of reference clusters must be chosen, and multiple runs are required to average over choices of these. | |
| ChromoPainter Linked Coancestry (CPL) | Individuals compete for similarity at each SNP<br>Sum over haplotypes<br>No undetermined parameters | ChromoPainter (31), based on the painting algorithm of Li and Stephens (35). |
| | **Notes:** Calculates the probability that each other haplotype is the most recent common ancestor in the sample. All uncertain parameters can be estimated, and are global in nature so do not tend to get stuck in local modes, therefore this algorithm only needs to be run once. | |

ine the performance of some of those considered here. We focus on 'unsupervised' methods that do not use any information about the population that individuals are *apriori* expected to be in.

We will consider the four direct methods summarised in Table 3. All methods attempt to identify individuals for which the rows and columns of the similarity matrix are similar and have some method for estimating the number of populations $K$. Our first method is FineSTRUCTURE. Model-based algorithms such as ADMIXTURE (2) and STRUCTURE (45) are theoretically exploiting the same information as FineSTRUCTURE applied to the CPU matrix (31), and indeed, have approximately the same likelihood. In that paper it was verified that their performance is similar (though slightly worse due to other modelling differences) for this simulated dataset (ADMIXTURE), and a smaller, unlinked dataset generated under the same genealogical model (STRUCTURE). The FineSTRUCTURE model can also be applied to the linked CPL dataset. It assumes that individuals within a population are exchangeable, i.e. any difference between them is due to noise (coming from a Multinomial distribution). It is also the only model considered that accounts for some peculiar features of similarity matrices, such as the self-similarity (the diagonal of the matrix) being meaningless for clustering.

The second method is MCLUST (14) which fits a very similar model to fineSTRUCTURE, but instead assumes that differences between individuals in the same population arise from a Multivariate Normal distribution with unknown variance. The third method is K-Means, which places individuals in the population 'closest' to them. The final method is the hierarchical 'Unweighted Pair Group Method with Arithmetic Mean' (UPGMA (53)) which iteratively merges the closest groups.

In addition to performing clustering on the similarity matrix itself, it is possible to first use 'spectral decomposition' (usually via Principal Components Analysis) to create a smaller, less noisy matrix. This approach leads to a substantial increase in performance for all generic approaches, but introduces an extra parameter, in the form of the number of principal components retained which is in practise difficult to estimate from the data in a reliable way. The results presented immediately below are based on using the 'Tracy-Widom' (TW) criterion but no approach has been found that works well in all cases. The potential power of the spectral approach makes it very important, so we more fully discuss its use and pitfalls in the light of our results in Section 7.

# 5 Results: Empirical Evaluation of the methods

## 5.1 Similarity matrices on simulated data

The simulated data described in detail in Lawson et al. (31) were used to compare the methods. In brief, one hundred 5Mb regions of full sequence data (2.5 mil-

Table 3: Clustering algorithms used.

| **Name**<br>(2 letter acronym) | **Applicability** | **References** |
|---|---|---|
| FineSTRUCTURE (FS) | Direct only,<br>Coancestry only (i.e. CPU/CPL) | Lawson et al. (31) |
| | **Notes:** No tuning parameters. Bayesian method with a multinomial likelihood, a biologically driven prior on the similarities and a conservative prior on the number of clusters. $K$ is estimated using reversible-jump MCMC (e.g. (16)). | |
| MCLUST (MC) | Direct or Spectral,<br>Any similarity matrix | Fraley and Raftery (14) implemented the R package (15), though the general concept is much older |
| | **Notes:** Multivariate-normal likelihood letting each population have its own mean and variance. Implicitly uniform prior and uses the Bayesian Information Criterion (BIC, which is asymptotically consistent) to infer $K$. Also called 'soft K-means'. | |
| K-Means (KM) | Direct or Spectral,<br>Any similarity matrix | Hartigan and Wong (20) introduced the algorithm, widely discussed in textbooks e.g. (27) |
| | **Notes:** No explicit likelihood, uses distance-based criterion assigning individuals to their nearest population which is located at the population mean. Uses the Calinski (8) criterion to determine $K$, which compares the variance within clusters to that between clusters. Implemented in the R-package 'vegan' using the function 'cascadeKM'. Also called 'hard K-means'. | |
| Hierarchical methods, focusing on UPGMA (UP) | Direct or Spectral,<br>Any similarity matrix | AWClust (17)<br>ipPCA (25)<br>UPGMA (e.g. (50)), Minimum Variance criterion (59), Neighbour joining (e.g. PHYLIP) (13) |
| (Ward's (59) criterion, AWClust and ipPCA have been tested, see Supplementary Figure S3) | **Notes:** Covers a wide variety of methods that first form a tree, and then the tree is 'cut' to create a clustering. We use the Calinski (8) criterion to determine $K$ as for K-Means. The UPGMA and Ward approaches are implemented in the R function 'hclust'. | |

lion SNPs in total) were simulated using SFS_CODE (21) based on the estimated human genome recombination map (26). A population scenario that gives levels of polymorphism and differentiation broadly similar to that found in Europeans was simulated, with a bottleneck followed by exponential growth, generating 5 populations following a tree pattern. First populations (A,B,C) split simultaneously 3000 years ago. Then population C splits into (C1,C2) 2000 years ago, and finally population B splits into (B1,B2) 1000 years ago. 20 individuals from each population were sampled.

Example similarity matrices produced by each category of similarity measure are shown in Figure 1. These can be inspected visually to get a rough idea of how each matrix captures the population structure. A 'good' matrix will have a strong 'block' structure, with each of the 5 populations clearly distinguishable, particularly on the diagonal. The IBS measure is very fuzzy with the COV (covariance) measure only a little clearer. The ESU method is the best unlinked method, with the 'C' populations clearly distinguishable and the 'B' populations beginning to take shape. The FHS matrix looks similar to the unlinked ESU matrix, whereas the IBD and CPL methods are significantly better distinguished. Supplementary Figure S1 shows some additional matrices, including the 'linked' EIGENSTRAT matrix and the ChromoPainter unlinked matrix. These contain the 'same' information as the ESU matrix shown here, a relationship which is further illustrated using the correlation between all measures in Supplementary Figure S2. This means that we only need to further examine the quality of the similarity matrices shown in Figure 1 as they are representative of the other matrices discussed. We did check that the clustering performance on the omitted matrices closely matches that expected by this classification.

## 5.2 Scoring the similarity matrices

The primary measure of the *useful* population signal in each similarity matrix is, for our purposes, how well they facilitate clustering. However, since there are many ways of clustering and has its own peculiarities, it is of interest to describe the ratio of signal to noise more directly. We consider the average within-population similarity when compared to the between-population similarity. This ratio measures how well separable the populations are and is shown in Figure 2. When there is little data, all clusters are equally indistinguishable and the within-population similarity is equal to the between population similarity. As the amount of data increases, the populations separate. The ChromoPainter linked CPL matrix provides the largest separation, closely followed by the IBD-based method. EIGENSTRAT is the best unlinked method, just ahead of the FastPHASE FHS linked method, with the covariance and the IBS matrices both performing badly. This is consistent with our visual observations of the matrix in the previous section.

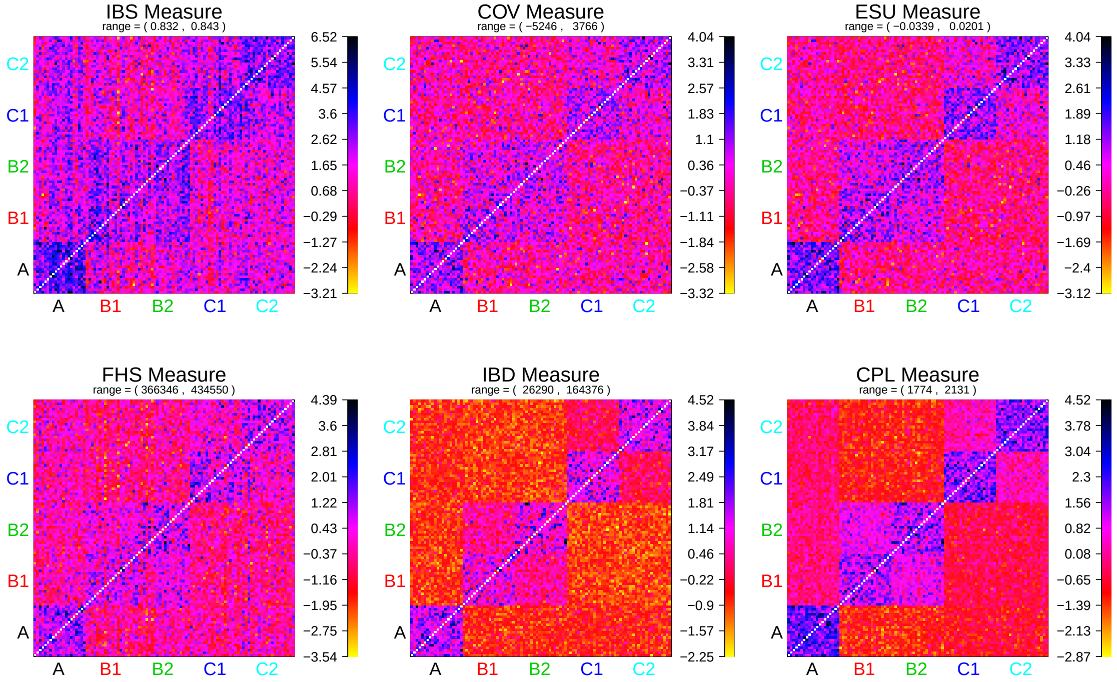

Figure 1: Visualisation of the similarity matrices as an image for one hundred 5Mb regions of simulated data. On the top row from left to right are: IBS (Identity-by-state), COV (Raw Covariance) and ESU (EIGENSTRAT 'unlinked', i.e. normalised covariance, no regression correction). The bottom row is: FHS (FastPHASE Haplotype Sharing), IBD (FastIBD Identity By Descent) and CPL (ChromoPainter Linked). The raw range is given above each matrix, but all plots are normalised by removing the diagonal, subtracting the mean and scaling the standard deviation to 1.

### 5.3 Clustering Performance

We ran each of our clustering algorithms (FineSTRUCTURE, MCLUST, K-Means, UPGMA, Spectral MCLUST, Spectral K-Means, Spectral UPGMA) on a range of the simulated data similarity matrices. States are scored by computing the average correlation with the true state as described in (31). Since direct application of MCLUST, K-Means and UPGMA to the similarity matrices was not very effective, these results are placed in Supplementary Figure S3 along with some additional hierarchical measures that perform similarly or worse relative to the clustering algorithms here.

The remaining methods are compared in Figure 3. MCLUST is better than K-Means and UPGMA in general, providing more stable estimates under most

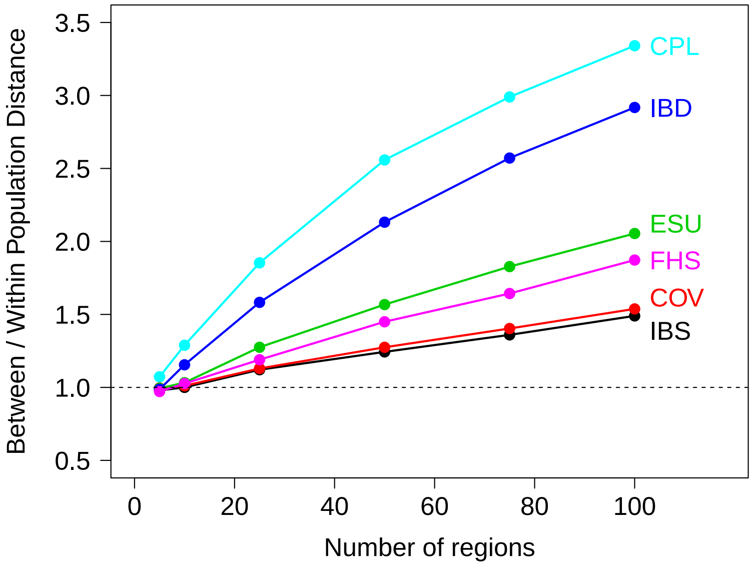

Figure 2: Mean between-population distance relative to the mean within-population distance. First, the average similarity $S_a$ within each population $a$ is calculated. Then the mean distance to $S_a$ of individuals in other populations $b \neq a$ is calculated, and averaged over all populations $a$. This is divided by the mean distance to $S_a$ for individuals within the population $a$ (again averaged over all $a$). The matrices are normalised (zero mean, diagonal removed and symmetrised by taking $YY^T$) prior to computation.

circumstances. However, K-Means and UPGMA both work on poor data, for which MCLUST fails to estimate $K$ and hence obtains a score of 0. Clustering performance is concordant with our previous observations of the quality of the similarity matrices. IBS and Covariance are the least useful measures for clustering, whilst EIGENSTRAT's normalised covariance matrix is the best unlinked method, somewhat similar to the FastPHASE performance. The linked ChromoPainter CPL matrix slightly outperforms the IBD matrix, particularly when the data is good, but both are significantly closer to the truth than the other methods. The IBD matrix consistently makes a few mistakes whereas the CPL matrix allows for perfect clustering using Spectral MCLUST.

We find similar performance between Spectral MCLUST and FineSTRUCTURE on the ChromoPainter Linked (CPL) matrix for this dataset. FineSTRUCTURE notably never splits individuals incorrectly, while MCLUST is typically bolder with the same data, for example mostly identifying the split between C1 and C2 with 50 regions which FineSTRUCTURE does not find enough evidence

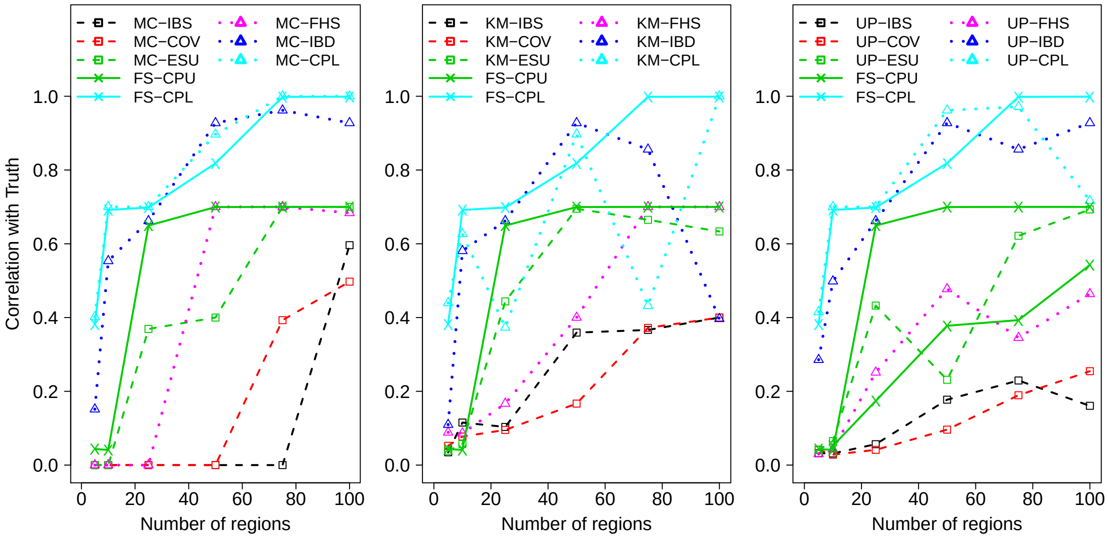

Figure 3: Correlation with the truth as a function of the number of 5Mb simulated data regions, for Spectral MCLUST (MC, left), Spectral K-Means(KM, centre) and Spectral UPGMA (UP, right), compared to fineSTRUCTURE. Shown are the clustering performance based on different similarity matrices. The unlinked methods (dashed lines) are IBS (Identity-by-state), COV (Covariance) and ESU (EIGENSTRAT Unlinked). The linked methods (dotted lines) are FHS (FastPHASE Haplotype Sharing), IBD (FastIBD Identity By Descent) and CPL (ChromoPainter Linked). FineSTRUCTURE is applied directly to the coancestry matrix only (solid lines, FS-CPU for unlinked and FS-CPL for the linked ChromoPainter Coancestry matrix), and is repeated on each plot for reference.

for. However MCLUST often creates spurious splits and in this example its performance decreases as additional regions are added, while fineSTRUCTURE gets progressively closer to the truth.

## 6 Application to HGDP data

Simulated data, however carefully constructed, lacks many of the features of real data. We therefore try out the different approaches on a subset of the HGDP (34) data. Our dataset consists of 140 individuals from East Asia and contained 500k SNPs. Although there is no ground truth in this dataset, we can consider the agreement between algorithms and attempt to interpret potential clustering problems by examination of the similarity matrices. The matrices are constructed as for the simulated data, and for illustration the PCA plots for the CPL dataset are shown in Supplementary Figure S4.

We attempted to apply the direct MCLUST and K-Means methods to this data, but these could not identify more than two populations and are not considered further. As the two leading linked similarity measures, we work primarily with the ChromoPainter CPL matrix and FastIBD matrices. The ChromoPainter unlinked (CPU) matrix (which is equivalent to the ESU matrix) is shown in Supplementary Figure S5 for reference. The choice of the number of Eigenvalues (EVs) to keep for the Spectral method does strongly matter for the the details of the clustering assignment (Supplementary Figure S6 and Supplementary Section S3), but not for the features discussed below. On the basis of the simulation results we use the Tracy-Widom (TW) statistic for choosing the number of EVs. Although the results based on the other criteria might have been preferable here (and are shown in Supplementary Figure S6), this is not evident apriori.

FineSTRUCTURE finds 17 populations in the CPL data, which correspond closely to the labels and can be validated by examination of the Coancestry matrix (Figure 4). Also shown is the highest posterior FineSTRUCTURE result, and the states found by MCLUST and K-Means. These clusterings can also be compared directly as shown in Figure 5, which additionally scores the clusterings. The score is the variance of the number of chunks copied between individuals in different populations, relative to that expected under the FineSTRUCTURE model. This can be viewed as measuring lack of uniformity within each 'block' of Figure 4, and should be close to 1 if the clustering contains no additional substructure. A technical detail to note is that that scores consistently less than 1 can arise if the effective number of independent chunks is not be the same for all pairs of individuals (as ChromoPainter estimates this without considering population structure).

The Spectral methods both find quite different population assignments to the FineSTRUCTURE results, with the MCLUST assignments in particular being less similar than might be expected from the simulation results. Two minor differences are that the spectral approaches fail to distinguish some visible splits (such as the Yi/Naxi), and do not see any structure in the Han/Han.NChina/Tujia population. Although this population does contain some clear structure, this structure might be better described by a cline (generated by admixture, visible as varying coancestry with Dai) than a division into distinct ancestry profiles and therefore it is difficult to identify precise division points. K-Means makes a mistake placing the Dai with this Han group, and MCLUST has made a serious mistake by placing 2 Naxi, 2 She and 6 Dai individuals together that have very different coancestry. The labels of individuals within this population can be visually distinguished in Figure 4, and the score from Figure 5 identifies it as an incorrect clustering.

Although MCLUST does better using other Eigenvalue retention criteria than with Tracy Widom, this does not explain the major problems with the approach.

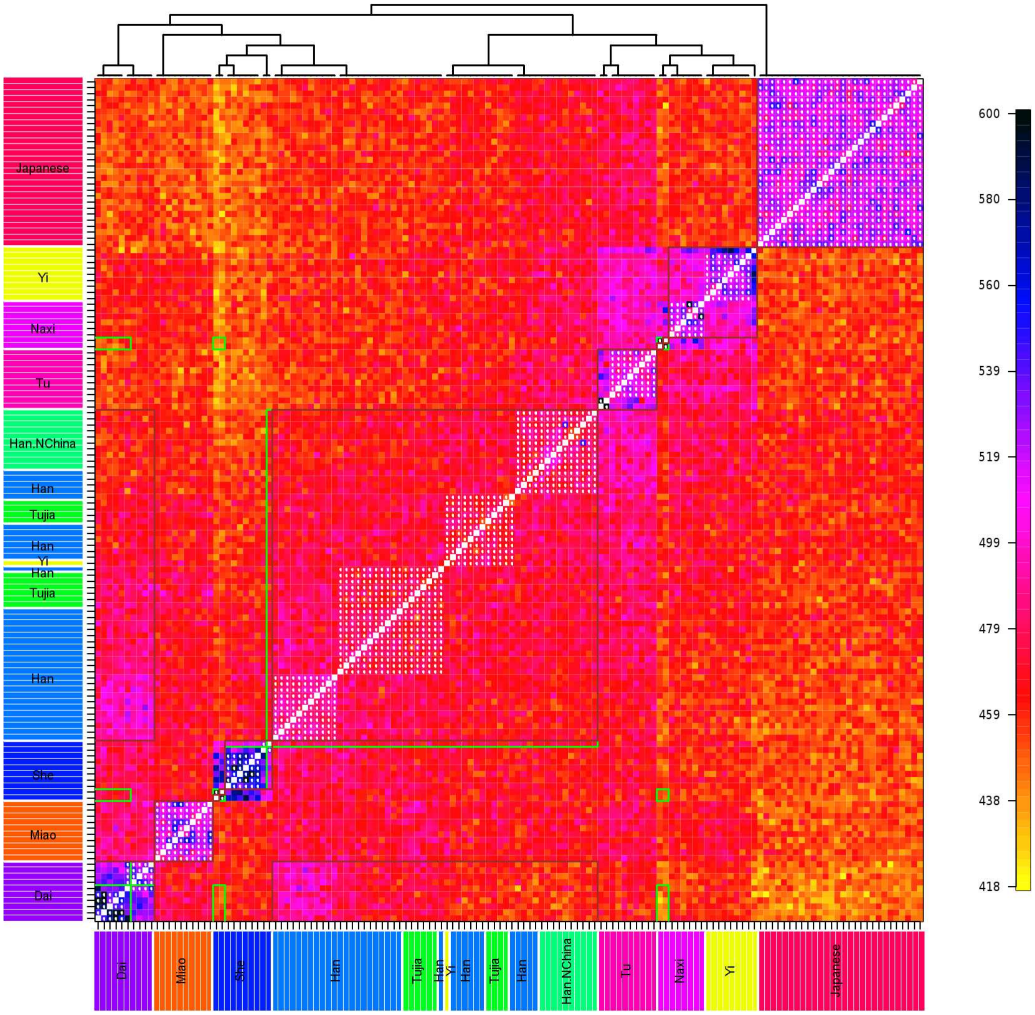

Figure 4: HGDP clustering results and coancestry matrix for the ChromoPainter linked (CPL) dataset. The main image shows the Coancestry matrix. The white dots show the pairs found in the same population in the FineSTRUCTURE Maximum-Aposteriori state. Green boxes are drawn around pairs found which coincide in the Spectral MCLUST populations, and brown boxes for Spectral K-means (both using the Tracy-Widom criterion). The ordering is formed from the FineSTRUCTURE tree, top, which has been rotated and individuals reordered within populations to make the MCLUST solution appear as close as possible to the diagonal. The Coancestry matrix has been capped at 500 to maximise contrast.

For all criteria, MCLUST places the two Naxi and She pairs together (and intermittently with the Dai). The reason for this is somewhat technical (but important) feature of the PCA decomposition. Supplementary Figure S4 shows the first 8 Eigenvectors in standard PCA plots, and Supplementary Figure S6 shows 'all components at once' using the correlation between the (top) Eigenvectors for each pair of individuals. The misplaced Naxi/She individuals are all distinguished by having a relatively low correlation with the rest of the sample. This is in turn due to their each sharing an unusually high number of chunks with another - the two Naxi are related, as are the two She, both having over 1000 chunks in common (and the next highest is within the Dai, at 600). Individuals with relatives in the sample have consistently fewer chunks from other populations. However, segments of DNA not directly shared with the relative are drawn from the same distribution as other individuals in their population. There is therefore a conflict between the population level signal and the signal from the relatives, which distorts the PCA decomposition and makes these individuals appear further away from the rest of their population than they really are.

MCLUST is fitting an incorrect population with high variance to these individuals, because they don't look similar to anyone else in the sample but do share the feature of looking *less like* the rest of the sample. Note that K-Means does not make this mistake because it does not model the variance at all. FineSTRUCTURE does not do this because it a) fits to the raw data, not the eigenvalues, and b) has a model for the variance. Related pairs of individuals are assigned to their own population, and then merged into the correct place in the tree by 'flattening' their population count.

The ChromoPainter linked matrix can be compared to the FastIBD similarity matrix, shown in Figure 6. From a visual inspection of the 'block diagonal' structure, it appears that FastIBD contains a similar strength signal to CPL, but interestingly the relationship between populations has a different emphasis. ChromoPainter found many populations with high rates of chunk sharing, for example the Dai/Han, Miao/Tu and Tu/Japanese. However, FastIBD finds *different* relationships, for example Miao/Tujia and She/Han. Although this is not a significant factor in the identification of populations as the signal is weak, the strong consistency over all individual pairs between the populations indicates that the algorithms are emphasising different but real historical ancestry signals.

The clustering results for the IBD matrix are less clear, as shown in Figure 5 (see also Supplementary Figure S7-8). The Spectral method has problems in that many Eigenvalues appear to contain a small amount of signal, leading the TW criterion to retain 35 EVs. The performance is better with the PA criterion (shown in the Figure) which retains only 5 EVs and for which comparisons between MCLUST on IBD and on the CPL matrix are mixed. They both make different clusterings of comparable quality (including the She/Naxi mistake), although it is interesting that only when using IBD are the Tujia mostly found within a single population. However, our Spectral methods are not robust enough to the choice of retained EVs to make any firm comparisons. Since FineSTRUCTURE cannot be applied to the IBD matrix, visual examination of the matrices remains the best comparison.

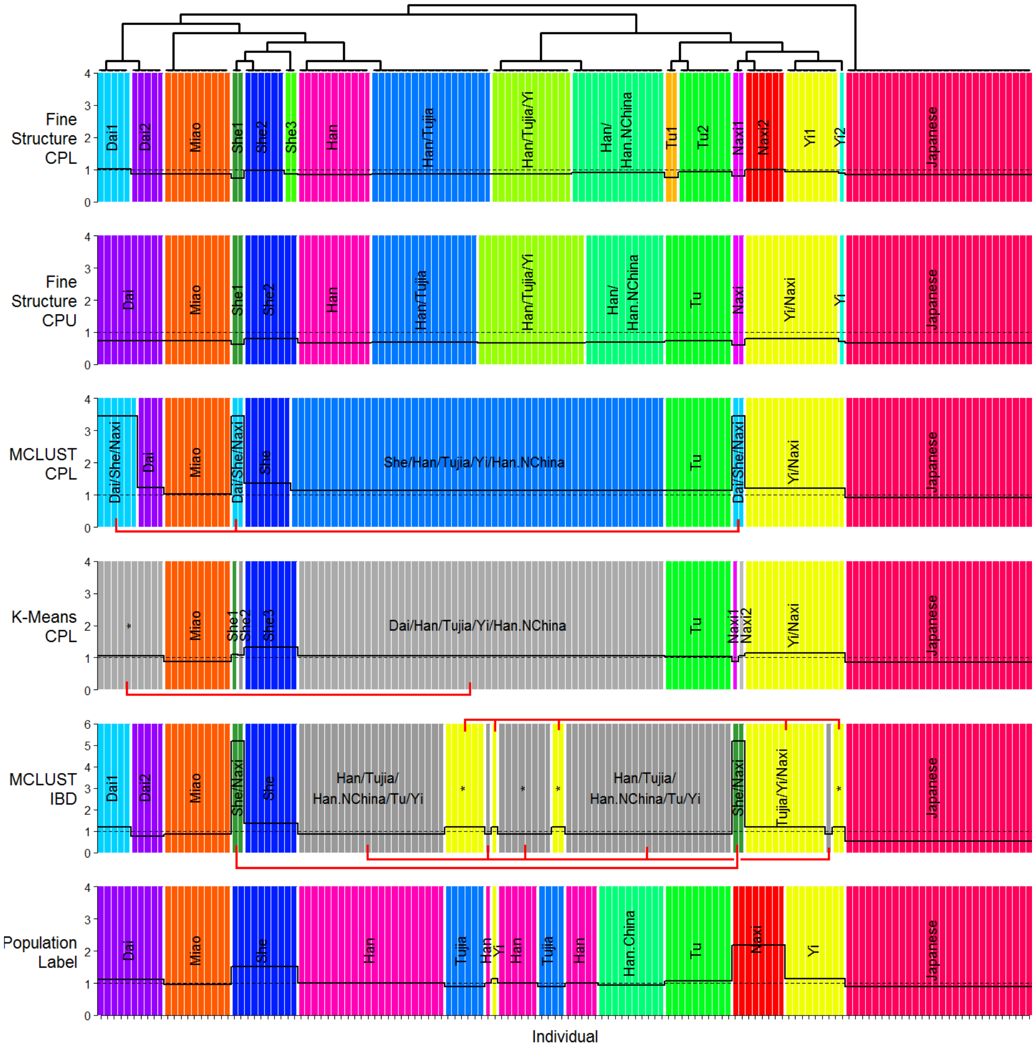

Figure 5: HGDP East Asian clustering for (top to bottom) FineSTRUCTURE on CPL data, FineSTRUCTURE on CPU, MCLUST on CPL, K-Means on CPL, and MCLUST on IBD (using the PA criterion for this data only). The self-identified population label of individuals are also shown (bottom). Identified populations are separated by white bars, and populations that are inconsistent with the FineSTRUCTURE CPL tree (above top) are linked by a red line. Each label has a unique colour, which are approximately matched. Also shown is a score measuring the ratio of observed to expected variance within a population (black curves), which if the population assignment were ideal would be one (dashed horizontal lines). This is also shown (bottom) when each unique label is assumed to form a population. Increased values imply substructure within a population. Compare with the full heatmaps (Figure 4, and Supplementary Figures S5-9), and see Supplementary Section S1 for a precise definition of the score.

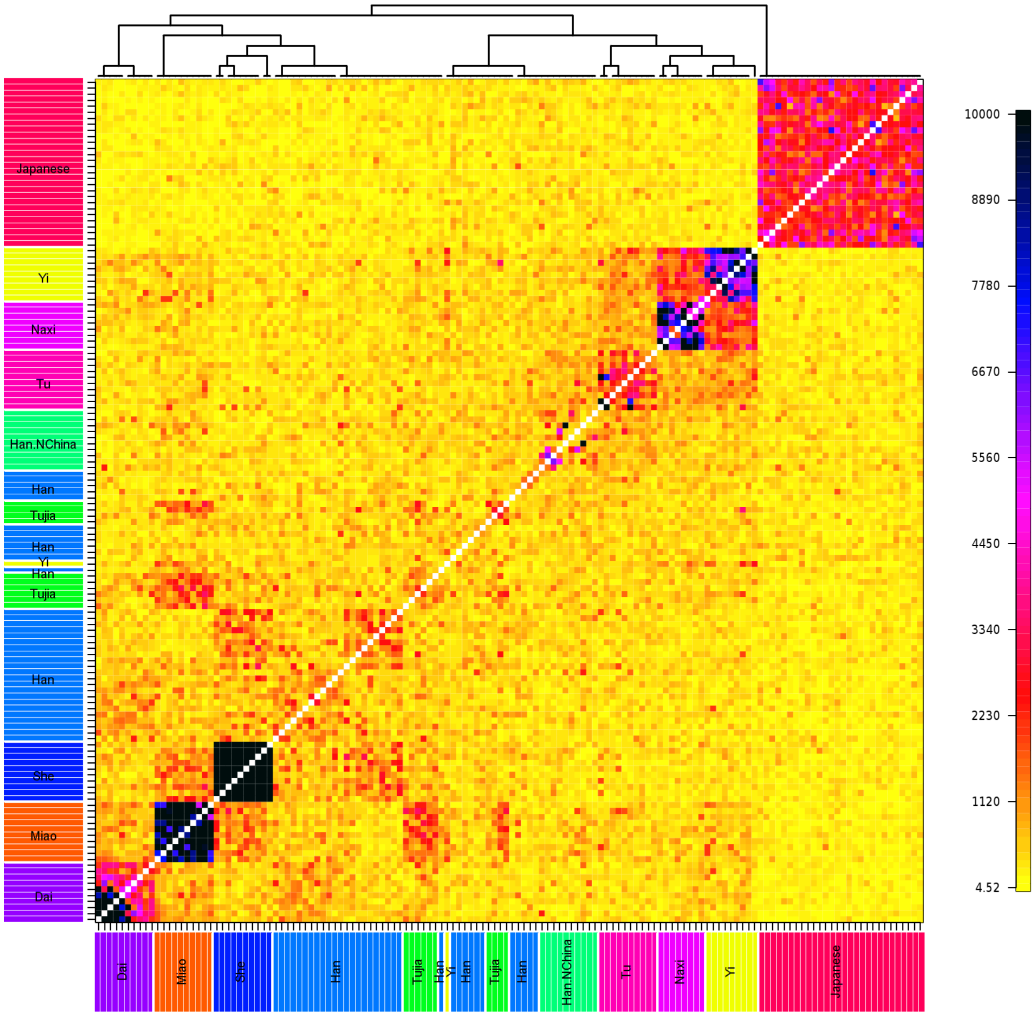

Figure 6: HGDP similarity matrix for the FastIBD dataset. The image shows the FastIBD similarity measure applied to the East Asian individuals, with other details are as Figure 4. Clusterings for the TW criterion are poor (Supplementary Figure S7) but the using the PA criterion (Supplementary Figure S8) works well. The similarity matrix has been capped at 10000 to maximise contrast.

# 7 Spectral methods

This section is concerned with the technicalities of spectral methods, which are powerful but do have several problems that are not fully resolved. We give a short review of Spectral methods, discuss some of the various approaches that have been used to estimate the number of significant Eigenvalues and describe a difference between spectral and non-spectral clustering. We conclude with the analysis of our simulated data. Readers who are focused principally on the biological application of these methods may wish to skip this section entirely.

The broad set of approaches called 'Spectral' methods all have the general goal of identifying where in the data the 'important' variation lies. In population genetics, all widely applied methods equate *important* variation with *large* variation since under the assumption of constant rates of genetic drift, individuals should be more similar within a population than between populations. The main approaches are Principal Components Analysis (PCA) (39; 44), Multidimensional Scaling (MDS) (62; 46) and Singular Value Decomposition (SVD) (3) all of which are intimately related. As discussed by McVean (38), performing SVD on the raw SNP matrix $Y'$ is equivalent to performing PCA on the covariance matrix $Y'(Y')^T$, where $Y'$ is the raw SNP matrix $Y$ with the empirical SNP frequency subtracted. MDS can be applied to a distance matrix, but (classical) MDS is otherwise equivalent to performing PCA on the covariance matrix. Other forms of MDS exist (either metric, non-metric based on ranks or graphs, and other approaches, e.g. (62)) but are not routinely applied to genetics data.

Another spectral approach is to construct a graph from the distance matrix, and perform PCA on the Graph Laplacian (see e.g. (27; 57)), an approach that has been applied to genetics (22; 63; 32). The distances are scaled to produce edge weights – two common choices are exponential scaling and to set all large lengths to have weight zero. When the complete graph is used and simple weights chosen, this method mirrors PCA very closely. For other choices of weights, a range of behaviour is possible as the algorithm can be 'tuned' to focus on a distance of interest. However, there is little theoretical understanding about how restriction to a graph might be interpreted genetically, and the addition of a free parameter makes a rigorous simulation study difficult, so we have not included graph approaches in the empirical evaluation section of this review.

## 7.1 The number of Eigenvalues to retain

Only approaches equivalent to PCA have reached significant popularity within the genetics community, so we shall focus on the approach of computing the Eigenvectors (EVs) $Z$ and Eigenvalues $\lambda$ of each similarity matrix $X$. All methods considered here must truncate the number of EVs retained. Choosing this truncation is a very difficult problem, to which there is not a definitive answer. In principle, if there were no noise in the data (for example, if we could replace samples from populations with their true population SNP frequencies) the only non-zero Eigenvalues would correspond to directions separating populations. All other Eigenvalues are non-zero in practise due to noise created from the random sampling of SNPs within individuals, and the only 'correct' way to choose the number of components is to model the expected underlying noise distribution.

Patterson et al. (44) model the Eigenvalues using the Tracy-Widom distribution (55). The Tracy-Widom distribution describes the largest Eigenvalue of the matrix $ZZ^T$, where $Z$ is a random (unstructured) matrix. Therefore, when all Eigenvalues associated with population structure have been removed, the maximum Eigenvalue should come from the Tracy-Widom distribution, allowing the construction of a statistical test for the significance of each Eigenvalue and its associated Eigenvector. In practise, to handle linked data the 'effective number of SNPs' is estimated from the Eigenvalue distribution, which additionally allows application to general similarity matrices. Whist this distribution is correct for some limit of the Normalised Covariance (ESU) data, the choice of p-value is difficult (33), and when the assumptions are violated there is no guarantee that the process will choose the correct number of Eigenvalues.

Most implemented methods use a more ad-hoc, data driven selection criterion. Elementary methods in multivariate analysis such as the Kaiser criterion (29) or the Scree 'test' (9) have been shown repeatedly to not work well (e.g. (48; 12) and also tested on our data, results not shown). Lee et al. (32) note that Tracy-Widom theory does not apply to graph Laplacians and instead use a criterion based on the 'Eigengap', i.e. the difference between Eigenvalues, but their method relies on the use of a graph Laplacian. Such an approach naturally generalises the Scree test and could potentially be developed, but a ready-to-use version is not available in the literature. Limpiti et al. (36) introduce a new criterion they call 'Eigendev' and claim increased performance relative to Tracy-Widom. Dinno (12) discusses some more applicable approaches and perform a simulation analysis on one popular method, the Monte-Carlo version of the 'Parallel Analysis' (PA) method of Horn (23) which compares the variance to that of random data. Another favoured method is the 'Minimum Average Partial' (MAP) method of Velicer (56) which looks for a drop in the relative amount of systematic variance explained (see e.g. (43) for a concise discussion of both). We consider the performance of three (TW, MAP and PA) criterion which are further described in Table 4.

## 7.2 Relating the Eigenvectors to the raw covariance

The Spectral mapping discards Eigenvectors with small Eigenvalues, which are assumed to contain only random noise. However, even if all significant Eigenvalues are retained, information is not efficiently used in our spectral approach. Because only the Eigenvectors are modelled, any information in the Eigenvalues

Table 4: Spectral decomposition

| **Name** | **Key features** | **References** |
|---|---|---|
| Minimum Average Partial (MAP) | The largest eigenvalue is successively removed from the matrix, and the average squared 'partial' correlation computed between the *remaining* eigenvectors and the original data. The MAP $K$ occurs at the minimum correlation. | (56), implemented in the R package 'psych'. |
| Parallel Analysis (PA) | Many random matrices with the same dimensions as the similarity matrix are generated, and their Eigenvalues computed. Eigenvalues that are larger in the data than the desired quantile of the random matrices are retained. | (23), implemented in the R package 'paran'. |
| Tracy-Widom (TW) | The Eigenvalue spectrum is compared to those of a theoretical random matrix. This is recursively updated as Eigenvalues are removed. The size of the theoretical matrix is calculated using the 'effective number of SNPs', calculable from the Eigenvalue spectrum itself. | (44) Implemented in the program 'twstats' from the EIGENSTRAT package. |

is lost. Although Eigenvalues are not associated with individuals and therefore don't directly impact clustering, they do describe the relative importance of each Eigenvector. This can effect clustering, as can be seen from the relationship between the similarity matrix $Z$, its Eigenvalues $\lambda$ and Eigenvector matrix $E$. $Z$ is decomposed in our method as $ZZ^T = E\text{Diag}(\lambda)E^T$, where $ZZ^T$ is the covariance of the similarity matrix. The Spectral methods are effectively clustering on the covariance of the Eigenvalues, $EE^T$ which differs by the scaling $\text{Diag}(\lambda)$. This leads to a different emphasis about which features are important in the data. Supplementary Figures S6 and S9 show the difference between these two correlations for our HGDP example. The Eigenvector representation makes individuals with relatives look less similar to their underlying population than does the similarity matrix. This property would be considered undesirable in a population genetics context. Although most spectral approaches discard the Eigenvalues as ours does, this is not required and so this problem may be solvable.

### 7.3 Simulated data: number of Eigenvalues to retain

We perform Eigenvalue decomposition of each similarity matrix $Z$ using the function 'eigen' in R, normalised by subtracting the matrix mean, removing the diagonal and symmetrising by taking $ZZ^T$. This transformation does not change the Eigenvectors of a symmetric matrix. We wish to empirically evaluate the performance of each measure from Table 4 for both the MCLUST and K-Means algorithms. By considering performance at all possible retained EVs, we can also find the 'optimum' choice which may not be found by any of the criteria. This provides a more robust view on the quality of the different criteria.

Figure 7 (with more detail in Supplementary Figure S10) explores the correlation with the truth as both the amount of data and the number of Eigenvalues retained is changed. There is a strong correspondence between the performance of all three methods and so we show only K-Means for clarity. In general the PA criterion seems to underestimate the number of Eigenvalues, the MAP criterion does a little better but is less stable, and the Tracy-Widom criterion falls somewhere in the middle. The results are consistent with those from Figure 3. As there is little empirical difference in the clustering performances between MAP and TW, we opted to work with the Tracy-Widom criterion, but note that no method dominates and that all may give misleading results under the wrong conditions.

The different similarity matrices perform very differently in this scenario. For reference, finding the splits (A,B,C) scores a correlation of 0.4 (pink) and finding the plots (A,B,C1,C2) scores 0.7 (blue). IBS (and COV) achieves weak clustering, and cannot find the C split (blue) using any criterion for the number of Eigenvalues with all 100 regions. The ESU matrix finds this split with around 50 regions, the IBD and CPL with only 10. There is a moderate amount of noise in the clustering performance as the number of EVs changes and this is usually amplified

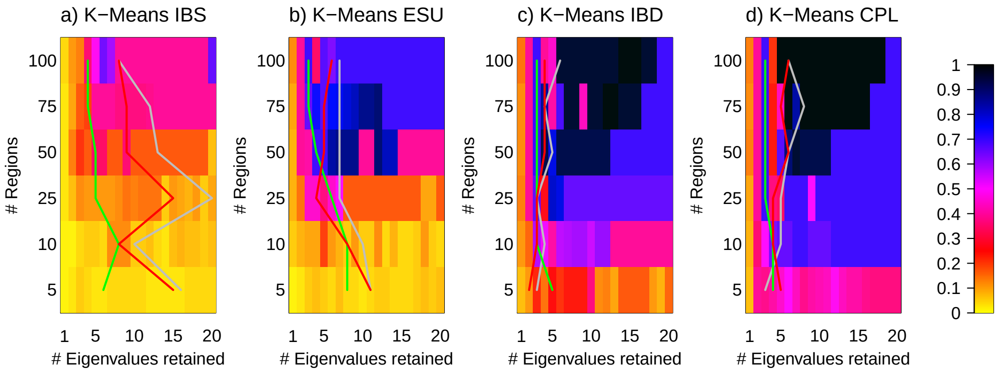

Figure 7: Correlation with the truth as a function of both the number of regions, and the number of Eigenvalues retained for fitting. A subset of K-Means matrices is shown; MCLUST and UPGMA as well as the other matrices are shown in Supplementary Figure S10. Shown are IBS (Identity-by-state, a), ESU (EIGENSTRAT Unlinked, b), IBD (Identity By Descent, c) and CPL (ChromoPainter Linked, d). The grey line corresponds to the number of Eigenvectors according to the MAP criterion, green lines to the PA criterion, and red lines to the Tracy-Widom criterion.

by applying a criterion. We note that the MCLUST performance is a little more robust to the number of EVs retained.

# 8 Discussion

Determining relationships between genetic samples represents a first step in almost all studies that hinge on patterns of genetic variation. We have reviewed the most widely used similarity/distance measures that can be constructed using genetic data, and their use in clustering algorithms identify distinct ancestry profiles. An alternative to clustering is to examine the Principal Components, which is typically done two components at a time. In our experience, visualisation via a heatmap of the ordered matrix of clusters showing the similarity between each one (Figures 1, 4 and 6) complements this traditional approach and barplots of STRUCTURE/ADMIXTURE output. The similarity heatmap is often more informative in practise since it allows variation to be assessed simultaneously at multiple different levels. Clustering the sample into 'populations' with discrete ancestry profiles also represents a useful starting point in approaches that seek to infer the historical processes that have led to differentiation between members of the sample, whether on short or on long timescales.

The distance measures can be classed into two broad categories. Unlinked methods ignore the position of markers on the chromosomes, while linked methods take chromosomal position into account and are most accurate if a detailed genetic map is available. The two unlinked distance measures considered here have a clear genetic interpretation, and might have been assumed to provide adequate descriptions of relatedness. Despite this, we observed significant differences in the information (in terms of population signal) they contain about population structure and therefore the clustering quality that they achieved. 'Raw' measures (Identity-by-State, Allele Sharing Distance, and covariance) perform relatively poorly relative to 'Normalised' measures (EIGENSTRAT's normalised covariance and ChromoPainter's unlinked Coancestry matrix).

The theoretical results of Lawson et al. (31) imply that the normalised similarity matrix approximately contains the same information as used by the fully model-based approach of STRUCTURE (45). Additionally, the likelihood of FineSTRUCTURE is approximately the same as that of STRUCTURE. This means that unlinked model-based approaches using this likelihood can do little better than the FineSTRUCTURE results on the unlinked ChromoPainter matrix, which contains less information about population structure (Figure 2) than both the FastIBD and linked ChromoPainter matrices.

The linked similarity measures require a greater investment of time and bioinformatic/computational infrastructure to run but perform significantly better than their unlinked counterparts on both our simulated and real datasets (and see also (31)), illustrating the advantages of using a model that accounts for linkage. In simulated data, they have reduced false-positive clusterings, and find more of the true clusters with much less data. FastIBD operates on a variety of the Identity-by-Descent approach, and was close behind ChromoPainter's Linked Coancestry matrix in terms of both clustering performance and our signal-to-noise score.

For real data, use of the linked similarity matrices consistently allows the identification of finer subdivisions than the unlinked methods. Comparison between the FastIBD and ChromoPainter HGDP similarity matrices implies that each is extracting a different signal about the ancestral relationships. This is contained in the distribution of shared DNA tracts, for which ChromoPainter permits only a comparison to the closest tract whereas FastIBD allows multiple relationships at each SNP. Extracting and interpreting the full genealogical signal in an efficient way continues to represent a central and unsolved problem in statistical genetics.

We examined the performance of four clustering algorithms in detail, MCLUST, K-means, UPGMA and FineSTRUCTURE. When clustering on simulated data, we found that *direct* application of generic methods (MCLUST, K-Means and UPGMA) led to inefficient use of the information in the similarity matrix. However, Spectral methods (Principal Components Analysis) could extract all the information in the data, provided that some criterion could be found that estimated the number of Eigenvectors that contained information on ancestry. We evaluated three, the MAP criterion, the PA criterion, and the Tracy-Widom criterion, but did not find strong empirical evidence for which to prefer. All could be successfully be applied under different to a range of similarity matrices, yet all showed evidence of mistakes. On simulated data, the PA criterion was weakest, discarding too many Eigenvalues, and the MAP estimate seemed less stable than the Tracy-Widom which we therefore favoured. On the HGDP results, the ordering was approximately reversed.

MCLUST and K-means perform similarly on the simulated dataset. Both attempt to solve very similar problems, minimising the root-mean square (i.e. Euclidean distance) from their population centres. The differences arise because K-means uses a 'hard' criterion, minimising this quantity directly, whereas MCLUST using a 'soft' criterion based on a multivariate normal. For many similarity measures the simulated data is drawn approximately from this multivariate normal (31) so these simple models are all close to 'correct'. They are likely to differ more strongly on data drawn from a different distribution, and indeed on the IBD measure (which has no theoretical result relating to normality) they do make different mistakes.

The HGDP data contains some departures from the typical modelling assumptions of large populations, no admixture, unrelated individuals, and random drift of SNPs. On this data we saw significant differences between the clustering provided by the model-based FineSTRUCTURE approach and the more generic methods. The presence of weak but unmodelled relationships between individuals is disastrous to the MCLUST Spectral approach which inappropriately clusters individuals with relatives in the sample together. K-means failed on the presence of admixture, providing misleading results by clustering some admixed individuals with close but clearly distinguishable populations. However, FineSTRUCTURE handles all of these problems well. It identifies relatives as such and can place them with the correct population at the hierarchical clustering stage, and places sensible boundaries on admixed populations. Constructing simulated datasets that contain these problems with a known truth would be valuable for more clearly exploring the performance of these (and other) algorithms.

There are three clear messages from the comparison study undertaken as part of this review. Firstly, for dense data, linked methods provide a moderate to strong information benefit that depends on the scale of the problem at hand. However, on HGDP density SNP data and population separation level, unlinked methods provide an adequate description when coupled with a robust inference algorithm. Secondly, there are multiple viewpoints that can be taken for understanding linked data, and each may provide different insights into the genealogical process. Thirdly, the ChromoPainter/FineSTRUCTURE pipeline is currently the best practical approach to population identification because it allows for robust and powerful model-based population identification.

We believe that there is value to examining performance on standard datasets, and therefore are making available both our simulated and HGDP datasets in PLINK, BEAGLE and PHASE formats on the website `http://www.paintmychromosomes.com` on the 'Comparisons' page. We will also provide similarity matrices for each method and example comparisons for the language R (47). Although future methods should avoid optimising to out-perform the above approaches on specific datasets, we hope that having these available for comparison will set a standard for population identification problems in genetics.

## Acknowledgements

We are very grateful to Paul Scheet for advice on running the FastPHASE Haplotype Sharing algorithm, and to Dienekes Pontikos for making us aware of the FastIBD/MCLUST approach. We are indebted to Simon Myers for deeply insightful comments on the manuscript. This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol (http://www.bris.ac.uk/acrc).

## References

1. 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073.

2. Alexander DH, Novembre J, Lange K, 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664.

3. Alter O, Brown PO, Botstein D, 2000. Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*, 97:10101–10106.

4. Biswas S, Scheinfeldt LB, Akey JM, 2009. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *Am J Hum Genet.*, 15:641–650.

5. Browning BL, Browning SR, 2011. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.*, 88:173–182.

6. Browning SR, Browning BL, 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81:1084–1097.

7. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, et al., 2010. Genome-wide patterns of population structure and admixture among hispanic/latino populations. *Proceedings of the National Academy of Sciences*, 107(Supplement 2):8954–8961.

8. Calinski T, Harabasz J, 1974. A dendrite method for cluster analysis. *Commun. Stat.*, 3:1–27.

9. Cattell RB, 1966. The scree test for the number of factors. *Multivariate Behavioral Research*, 1:629–637.

10. Cho K, Dupuis J, 2009. Handling linkage disequilibrium in qualitative trait linkage analysis using dense snps: a two-step strategy. *BMC Genet.*, 10:44.

11. Delaneau O, Marchini J, Zagury J, 2011. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9:179–81.

12. Dinno A, 2009. Exploring the sensitivity of horn's parallel analysis to the distributional form of simulated data. *Multivariate Behavioral Research*, 44:362–388.

13. Felsenstein J, 2003. *Inferring Phylogenies.* Sinauer Associates.

14. Fraley C, Raftery AE, 2002. Model-based clustering, discriminant analysis and density estimation. *J. Amer. Stat. Assoc.*, 97:611–631.

15. Fraley C, Raftery AE, 2006 (revised 2009). Mclust version 3 for r: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington.

16. Gamerman D, 1997. *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference.* Chapman and Hall, 2-6 Boundary Row, London, UK. SE1 8HN.

17. Gao X, Starmer J, 2007. Human population structure detection via multilocus genotype clustering. *BMC Genetics*, 25:34.

18. Gao X, Martin ER, 2009. Using allele sharing distance for detecting human population stratification. *Hum Hered.*, 68:182–191.

19. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, et al., 2009. Whole population, genomewide mapping of hidden relatedness. *Genome Research*, 19:318–26.

20. Hartigan JA, Wong MA, 1979. A k-means clustering algorithm. *Applied Statistics*, 28:100–108.

21. Hernandez R, 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 24:2786–2787.

22. Higham DJ, Kalna G, Kibble M, 2007. Spectral clustering and its use in bioinformatics. *J. Comp. Appl. Math.*, 204:25–37.

23. Horn JL, 1965. A rationale and a test for the number of factors in factor analysis. *Psychometrika*, 30:179–185.

24. Howie BN, Donnelly P, Marchini J, 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529.

25. Intarapanich A, Shaw PJ, Assawamakin A, Wangkumhang P, Ngamphiw C, et al., 2009. Iterative pruning pca improves resolution of highly structured populations. *BMC Bioinformatics*, 10:382.

26. International HapMap Consortium, 2007. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449:851–61.

27. Jain AK, Dubes RC, 1988. *Algorithms for clustering data.* Prentice-Hall, New Jersey.

28. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al., February 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003.

29. Kaiser HF, 1960. The application of electronic computers to factor analysis. *Edu. and Psych. Measurement*, 20:141–151.

30. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, et al., 2002. A high-resolution recombination map of the human genome. *Nature*, 31:241–272.

31. Lawson DJ, Hellenthal G, Myers S, Falush D, 2012. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8:e100245.

32. Lee AB, Luca D, Klei L, Devlin B, Roeder K, 2010. Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology*, 34:51–59.

33. Lee C, Abdool A, Huang CH, 2009. PCA-based population structure inference with generic clustering algorithms. *BMC Bioinformatics*, 10:S73.

34. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al., 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104.

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

36. Limpiti T, Intarapanich A, Assawamakin A, Shaw PJ, Wangkumhang P, et al., 2011. Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and structure. *BMC Bioinformatics*, 12:255.

37. Liu N, Zhao H, 2006. A non-parametric approach to population structure inference using multilocus genotypes. *Hum Genomics*, 2:353–64.

38. McVean G, 2009. A Genealogical Interpretation of Principal Components Analysis. *PLoS Genetics*, 5(10):e1000686.

39. Menozzi P, Piazza A, Cavalli-Sforza L, 1978. Synthetic maps of human gene frequencies in europeans. *Science*, 201:786–792.

40. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, et al., 2010. Drive against hotspot motifs in primates implicates the prdm9 gene in meiotic recombination. *Science*, 327:876–879.

41. Nakamura T, Shoji A, Fujisawa H, Kamatani N, 2005. Cluster analysis and association study of structured multilocus genotype data. *J. of Hum. Genet.*, 50:53–61.

42. Nicholson G, Smith A, Jónsson F, Gústafsson O, Stefánsson K, Donnelly P, 2002. Assessing population differentiation and isolation from single nucleotide polymorphism data. *J Roy Stat Soc B*, 64:695–715.

43. O'Connor BP, 2000. Spss and sas programs for determining the number of components using parallel analysis and velicers map test. *Behavior Research Methods, Instruments, & Computers*, 32:396–402.

44. Patterson N, Price AL, Reich D, 2006. Population Structure and Eigenanalysis. *PLoS Genetics*, 2:2074–2093.

45. Pritchard JK, Stephens M, Donnelly P, 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

46. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, et al., 2007. Plink: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81:559–75.

47. R Development Core Team, 2011. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

48. Râiche G, 2005. Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*, 19:1012.

49. Reeves PA, Richards CM, 2009. Accurate inference of subtle population structure (and other genetic discontinuities) using principal coordinates. *PLoS One*, 4:e4269.

50. Rodrigues FM, Diniz-Filho JAF, 1998. Hierarchical structure of genetic distances: Effects of matrix size, spatial distribution and correlation structure among gene frequencies. *Genetics and Molecular Biology*, 21:233–240.

51. Scheet P, Stephens M, 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, 78:629–644.

52. Sobel E, Sengul H, Weeks DE, 2001. Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Human Heredity*, 52:121–131.

53. Sokal R, Michener C, 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

54. Toussile W, Gassiat E, 2009. Variable selection in model-based clustering using multilocus genotype data. *Computational Statistics and Data Analysis*, 3:109–134.

55. Tracy C, Widom H, 1994. Level-spacing distributions and the airy kernel. *Commun Math Phys*, 159:151–174.

56. Velicer W, 1976. Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41:321–327.

57. von Luxburg U, 2007. A tutorial on spectral clustering. *Stat. Comput.*, 17:395–416.

58. Wakeley J, 2008. *Coalescent Theory: An Introduction.* Roberts & Company, Colorado.

59. Ward JH, 1963. Hierarchical grouping to optimize an objective function. *Journal of Amer. Statist. Assoc.*, 58:236–244.

60. Witherspoon DJ, Wooding S, Rogers AR, Marchani EE, Watkins WS, et al., 2007. Genetic similarities within and between human populations. *Genetics*, 176:351–359.

61. Xu R, Wunsch D, may 2005. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645 –678.

62. Young FW, Hamer RM, 1987. *Multidimensional Scaling: History, Theory and Applications*. Erlbaum, New York.

63. Zhang J, Niyogi P, McPeek MS, 2009. Laplacian eigenfunctions learn population structure. *PLoS One*, 4:e7928.