Supporting Information for: Inference of population structure using dense haplotype data

Daniel John Lawson[*], Garrett Hellenthal[†], Simon Myers[‡], Daniel Falush[§]

November 16, 2011

Contents

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

[†] Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK

[‡] Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK

[§] Environmental Research Institute, University College Cork, Ireland and Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig Germany

List of Figures

List of Tables

S1 Painting Algorithm

Li and Stephens (2003) described a likelihood based model that captures key features of the genealogical process with recombination while remaining computationally tractable for large datasets. Under the model, a chromosome is generated chunk-by-chunk by ‘copying’ from a conditional set of fixed haplotypes. In our notation, every individual consists of two haploids, each consisting of a single phased haplotype per chromosome. The L total SNPs in each haploid are listed one chromosome at a time, in order within each chromosome.

Suppose that we wish to generate a particular haploid $h_* = \{h_{*1}, ..., h_{*L}\}$, with h_{*l} the observed allele of h_* at site l, using j pre-existing donor haploids $h_1, ..., h_j$. Let $\vec{\rho} = \{\rho_1, ..., \rho_{L-1}\}$ be a vector of genetic distances, with ρ_l the population-scaled genetic distance between sites l and $l+1$ (i.e. $\rho_l = N_e g_l$, where N_e is analogous to the “effective population size” and g_l is the genetic distance in Morgans between sites l and $l+1$). (Between chromosomes, the genetic distance between the last site of the previous chromosome and the first site of the next chromosome is ∞.) Let $\vec{f} = \{f_1, ..., f_j\}$ be a vector of copying probabilities, with f_k the probability of copying from haploid h_k at any site. Let θ correspond to a per site mutation (or “imperfect copying”) parameter. The conditional probability $\Pr(h_* \mid h_1, ..., h_j; \vec{\rho}, \vec{f}, \theta)$ is structured as a Hidden Markov model. Let $\vec{Y} = \{Y_1, ..., Y_L\}$ represent the hidden state sequence vector, with Y_l the existing haploid from the set h_1,...,h_j that haploid h_* copies from at site l. Switches in the haploid being copied between Y_l and Y_{l+1} occur as a Poisson process with rate ρ_l. The transition probabilities for Y between sites l and $l+1$ are as follows (we exclude $h_1, ..., h_j$ and the parameters from the left side of equations (S1) and (S2) below for ease of reading):

$$\Pr(Y_{l+1} = y_{l+1} | Y_l = y_l) = \begin{cases} \exp(-\rho_l) + \Big(1 - \exp(-\rho_l)\Big) f_{y_{l+1}} & \text{if } y_{l+1} = y_l; \\ \Big(1 - \exp(-\rho_l)\Big) f_{y_{l+1}} & \text{otherwise,} \end{cases} \tag{S1}$$

The observed state sequence component of the Hidden Markov Chain, the probability of observing a particular allele given the haploid that h_* is copying from at a given SNP, allows for "imperfect" copying:

$$\Pr(h_{*l} = a|Y_l = y) = \begin{cases} 1.0 - \theta & h_{yl} = a; \\ \theta & h_{yl} \neq a. \end{cases} \tag{S2}$$

Here h_{kl} refers to the allelic type of haploid k at SNP l. To calculate $\Pr(D) \equiv \Pr(h_* \mid h_1, ..., h_j; \vec{\rho}, \vec{f}, \theta)$, a summation is performed over all permutations of the copying process, i.e. a summation over all possible y, which can be accomplished efficiently using the forward algorithm (e.g. Rabiner 1989).

For all analyses presented here, we fix the mutation parameter θ to Watterson's estimate (Watterson 1975), as used by Li and Stephens (2003), i.e.

$$\theta = \frac{1}{2}\frac{\left(\sum_{i=1}^{j} 1/i\right)^{-1}}{j + \left(\sum_{i=1}^{j} 1/i\right)^{-1}}$$

for j total haploids. We fix each g_l by taking the build 36 genetic distance estimates from the HapMap website (`http://www.hapmap.org`), which were calculated using Phase II genotypes and averaging values across the three HapMap populations as described by the International HapMap Consortium (2007). We also fix each f_k to be $1/j$ for $k = 1, ..., j$, allowing for equal *a priori* probability of copying from each conditional haploid.

Calculating expected number of chunks copied:

The average number of chunks copied to a haploid $*$ is a random variable denoted $\hat{x}_i = \mathbb{E}_{l=1\cdots L}(X_{il})$, where X_{il} is the probability that a given locus l is a new haplotypic segment copied from individual i. To calculate $\hat{x}_1, ..., \hat{x}_j$, the posterior expected number of chunks for which haploid h_* copies from each of $h_1, ..., h_j$, respectively, we calculate $\hat{f}_{k,l}$, the probability haploid h_* is copying from haploid h_k at site l given at least one "switch" has occurred between $l-1$ and l. Again excluding parameters for ease of reading, let $\alpha_{kl} = \Pr(h_{*1}, ..., h_{*l}, Y_l = h_k)$ and $\beta_{kl} = \Pr(h_{*(l+1)}, ..., h_{*L} \mid Y_l = h_k)$. Then

$$\begin{aligned} \hat{x}_k &= \tfrac{\alpha_{k1}\beta_{k1}}{\Pr(D)} + \textstyle\sum_{l=1}^{L-1}(\tfrac{1}{\Pr(D)})\Big[\alpha_{k(l+1)}\beta_{k(l+1)} - \alpha_{kl}\beta_{k(l+1)}\Pr(h_{*(l+1)}|Y_{l+1} = h_k)\exp(-\rho_l)\Big] \\ &= \tfrac{\alpha_{k1}\beta_{k1}}{\Pr(D)} + \textstyle\sum_{l=1}^{L-1}\hat{f}_{k,l}. \end{aligned} \tag{S3}$$

Note that we later drop the 'hat' notation for convenience, and form the matrix of all haplotype recipients $*$ as x_{ij}. Each row of x_{ij} corresponds to the vector $\hat{x}$ calculated above.

We calculate α_{kl} for $k = 1, ..., j$ in the following manner (Rabiner 1989):

1. $\alpha_{k1} = \Pr(h_{*1} \mid Y_1 = h_k) f_k$
2. $\alpha_{kl} = \Pr(h_{*l} \mid Y_l = h_k)\Big(\big[\sum_{i=1}^{j} \alpha_{i(l-1)}\big] f_k \big(1 - \exp(-\rho_l)\big) + \exp(-\rho_l)\alpha_{k(l-1)}\Big)$ for $l = 2, ..., L$.

We calculate β_{kl} for $k = 1, ..., j$ in the following manner (Rabiner 1989):

1. $\beta_{kL} = 1.0$
2. $\beta_{kl} = \big[\sum_{i=1}^{j} \beta_{i(l+1)} f_i \Pr(h_{*(l+1)} \mid Y_{l+1} = h_i)\big]\big(1 - \exp(-\rho_l)\big) + \exp(-\rho_l) \Pr(h_{*(l+1)} \mid Y_{l+1} = h_k)\beta_{k(l+1)}$ for $l = 1, ..., (L-1)$.

Calculating expected lengths of copied chunks:

To calculate $\hat{l}_1, ..., \hat{l}_j$, the posterior expected length (in Morgans) of the total genome for which haploid h_* copies from each of $h_1, ..., h_j$, respectively, we calculate the following (let $\Pr_h \equiv \Pr(h_{*(l+1)} \mid Y_{l+1} = h_k)$):

$$\begin{aligned} \hat{l}_k \quad = \quad & \tfrac{1}{\Pr(D)} \textstyle\sum_{l=1}^{L-1} g_l \Big[\alpha_{kl}\beta_{k(l+1)}\Big(\exp(-\rho_l) + (1.0 - \exp(-\rho_l))f_k\Big)\Pr_h \\ & +(1/2)\Big[\alpha_{kl}\beta_{kl} + \alpha_{k(l+1)}\beta_{k(l+1)} - 2\alpha_{kl}\beta_{k(l+1)}\Big(\exp(-\rho_l) + (1.0 - \exp(-\rho_l))f_k\Big)\Pr_h\Big]\Big]. \end{aligned} \tag{S4}$$

Note that this involves the approximation that at most only one change point occurs between neighbouring sampled sites. To get the expected length of *each* chunk copied from donor h_k, we divide equation (S4) by equation (S3) (i.e. $\hat{l}_k/\hat{x}_k$).

Calculating expected number of mutations:

To calculate $\hat{m}_1, ..., \hat{m}_j$, the posterior expected number of SNPs for which haploid h_* copies with mutation (i.e. emission) from each of $h_1, ..., h_j$, respectively, we calculate the following (let $\mathrm{I}_{[h_{*l} \neq h_{kl}]}$ be an indicator that the allelic type carried by h_* does not match the allelic type carried by h_k at SNP l):

$$\hat{m}_k \quad = \quad \tfrac{1}{\Pr(D)} \textstyle\sum_{l=1}^{L-1} \alpha_{kl}\beta_{kl}\mathrm{I}_{[h_{*l} \neq h_{kl}]}. \tag{S5}$$

Using the E-M algorithm to estimate the scaling parameter N_e:

One can take a fixed N_e for calculating $\vec{\rho}$, or use the Expectation-Maximisation (E-M) algorithm to find a local maximum of N_e in the following manner. Start with an initial value of N_e (we take $N_e = 400,000/j$), and at each iteration of the E-M replace N_e with:

$$N_e^* = \frac{\sum_{l=1}^{L-1} \left([\sum_{k=1}^{j} \hat{f}_{k,l}][\rho_l]/[1.0 - \exp(-\rho_l)] \right)}{\sum_{l=1}^{L-1} g_l}, \tag{S6}$$

where ρ_l and each $\hat{f}_{k,l}$ are calculated using the previous value of N_e. In analyses presented here, we used 10 iterations of E-M to get our final estimate of N_e.

Using the E-M algorithm to estimate the mutation parameter θ

One can take a fixed θ for calculating (S2), or use the E-M to find a local maximum of θ in the following manner. Start with an initial value of θ (we start with Watterson's estimate of θ), and at each iteration of the E-M replace θ with:

$$\theta^* = \frac{\sum_{l=1}^{L} \left(\sum_{i=1}^{j} \alpha_{il} \beta_{il} I_{[h_{*l} \neq h_{il}]} / \Pr(D) \right)}{L}. \tag{S7}$$

Here $I_{[h_{*l} \neq h_{il}]}$ is an indicator that the allele h_{*l} carried by the recipient is not equal to allele h_{il} carried by donor haploid i at SNP l, and each α_{il}, β_{il} and $\Pr(D)$ are calculated using the previous value of θ.

S2 Partition posterior probability evaluation

Here we define the full model for the coancestry matrix of expected copying counts x (dropping the 'hat' notation). Each row of x is distributed according to Multinomial likelihood $F(\cdot)$ as defined in Equation 1 of the main text:

$$x|\eta, P = \prod_{i=1}^{N} x_i | P_{q_i} \sim \prod_{i=1}^{N} F(\cdot | P_{q_i}), \tag{S8}$$

where N is the number of individuals, P_{q_i} is the row of P corresponding to the population q_i containing individual i, K is the number of populations and η is the assignment of individuals to populations. Population membership q_i can be thought of as induced by η, as is the set of individuals found in a population S_a. A Dirichlet Process Prior (e.g. Teh 2010) is placed on η, which (approximately, for the purposes of exposition) means that for large $K^* \to \infty$ (and not generally equal to K), the probability of the number of individuals assigned to each population $\mathbf{n}$ (which is related, but not equal to $\{S_a\}_{a=1 \cdots K}$) follows $\mathbf{n} \sim$ Multinomial(G) with $G \sim$ Dirichlet$(\alpha/K^*, \cdots, \alpha/K^*)$. Note that in this view, many of these populations will be empty, leaving a finite number K of occupied populations.

There are many representations of a Dirichlet Process, with a common choice being $\{P_1, \cdots, P_N\} \sim \mathrm{DP}(\alpha, G_0)$, where G_0 is the the 'base distribution', i.e. we sample parameters P_a from G_0, but obtain clustering by assigning the same parameters to multiple individuals. However, we choose an alternative description that suppressed G_0 which is simpler in our case.

The representation we find most natural is the joint assignment distribution induced on $\{\eta, K\}$, where K is the number of populations observed in our sample. This takes the form (Huelsenbeck and Andolfatto 2007):

$$p(\eta, K|\alpha, N) = \alpha^K \frac{\prod_{a=1}^{K} \Gamma(|S_a|)}{\prod_{i=1}^{N} (\alpha + i - 1)}, \tag{S9}$$

where there are N individuals, and α is the 'concentration parameter' determining the number of occupied populations expected under the Dirichlet Process. In this case we can write the distribution of each probability vector P_a:

$$\{P_a, \cdots, P_K\}|\eta = P|\eta \sim \prod_{a=1}^{K} \text{Dirichlet}(\beta_a), \tag{S10}$$

which is conjugate to F (and note that β_a is a vector of length K). This representation avoids the need to explicitly manage a G_0 that is itself a function of the number of populations K as is the case in our model. Note that we are free to use any distribution here in principle; this choice of Dirichlet distribution is not related to our use of a Dirichlet Process Prior.

From Equation S9, for fixed N and α the prior on η can be written:

$$\eta \sim p(\eta) \propto \alpha^K \prod_{b=1}^{K} \Gamma(|S_b|), \tag{S11}$$

so that when $\alpha = 1$ all possible assignments are given equal prior weight. This allows us to control K in principle (though in practice the likelihood term overwhelms the prior on K), and applies the usual Bayesian penalty for having additional parameters (via additional populations), leading to low K solutions being favoured in the posterior. We wish to calculate the probability of a particular partition η:

$$P(\eta|x) \propto P(\eta) \prod_{a=1}^{K} L(x_{S_a}|\eta) \tag{S12}$$

where $L(x_{S_a})$ is the likelihood of all the individuals in population a:

$$L(x_{S_a}) = \prod_{m \in S_a} P_{<m,S_a}(x_m) = \int \prod_{m \in S_a} F(x_m|P_m) dH_{<m,S_a}(P_m), \tag{S13}$$

where $P_{<m,S_a}(x_m)$ is the probability of the data row x_m given the data for subset $(1, \cdots, m-1)$ of individuals in S_a, with an incremental probability distribution

over P_a called (abusively) P_m. This is split up as the integral over the likelihood $F(x_m|P_m)$ of the probability of the parameters given the previous individuals data, $dH_{<m,S_a}(P_m)$. Conjugacy allows the incremental probability to be written as:

$$dH_{<m,S_a}(P_m) = \text{Dirichlet}\left(P_m; \left\{\beta_{ab} + d^{S_a}_{<m,b}\right\}_{b=1,\cdots,K}\right), \tag{S14}$$

where β_{ab} is the prior given by Equation 2 of the main text and $d^{S_a}_{<m,b}$ are the counts from population S_b to population S_a for the individuals $[[1, \cdots, m-1]]$. The final posterior follows from Eq. 3.13 of Lange (2002):

$$P(\eta|x) \propto \alpha^K \prod_{a=1}^{K} \Gamma(|S_a|) \frac{\Gamma(\beta_a)}{\Gamma(d_a + \beta_a)} \prod_{b=1}^{K} \frac{\Gamma(\beta_{ab} + x_{ab})}{\Gamma(\beta_{ab})\hat{n}^{x_{ab}}}. \tag{S15}$$

S3 MCMC implementation

There are 4 moves, 2 of which are based on the SAMS scheme from Dahl (2003) and Pella and Masuda (2006), each of which is chosen with equal probability on each iteration.

SAMS proposal: from an initial state η two different individuals are chosen at random. If they are in the same population, it is split to form η'; if they are in different populations they are merged to form η'. On a split of population c (consisting of individuals S_c), the two individuals are each placed in new populations a and b and then each other individual in c is moved to i (either a or b) with probability:

$$P(m, S_i) = \frac{|S_i| \int F(x_m|p_m) dH_{<m,S_i}(p_m)}{|S_a| \int F(x_m|p_m) dH_{<m,S_a}(p_m) + |S_b| \int F(x_m|p_m) dH_{<m,S_b}(p_m)}, \tag{S16}$$

where $\int F(x_m|p_m)dH_{<m,S}(p_m)$ is the incremental probability of adding an individual to the population. Because in our model all individuals must be assigned to populations, we approximate it using the source population only and use a rejection step to account for the discrepancy:

$$\int F(x_m|p_m) dH_{<m,S_a}(p_m) \approx \frac{P(a, \{i = 1, \cdots, m\})P(c, \{i = 1, \cdots, m\})}{P(a, \{i = 1, \cdots, m-1\}P(c, \{i = 1, \cdots, m-1\}}, \tag{S17}$$

The notation $i = 1, \cdots, m$ refers to the fact that these individuals have been moved to population a from population c previously when generating the proposal. The incremental probability is calculated as in Equation S15:

$$P(a) = \frac{\Gamma(\beta)}{\Gamma(d_a + \beta)} \prod_{b=1}^{K} \frac{\Gamma(\beta_b + x_{ab})}{\Gamma(\beta_b)\hat{n}^{x_{ab}}}, \tag{S18}$$

and $P(b)$ is defined similarly. The probability of a given pair of split populations a and b from a single population c is therefore:

$$P(a,b|m \in S_c) = \prod_{m=1}^{|S_c|} \frac{|S_{q_m}|P(q_m)}{|S_a|P(a) + |S_b|P(b)}. \tag{S19}$$

Once all individuals from population c have been placed, the new state η' is accepted with probability:

$$\min(1, P(\eta')/P(\eta)P(a,b|m \in S_c)). \tag{S20}$$

A merge of two populations a and b in η similarly forms a new state η' and is accepted with probability

$$\min(1, P(a,b|m \in S_c)P(\eta')/P(\eta)). \tag{S21}$$

'merge-and-split' (MAS) proposal: following the same strategy as above, but first forces a merge and then a split. If we call the initial two populations d and e in state η, they are merged to form an intermediate state c and resplit (according to the procedure above) to form populations a and b in state η'. If $\eta = \eta'$ the state is accepted (though we count a rejection for acceptance rate purposes) else the state η' is accepted with probability:

$$\min(1, P(a,b|m \in S_c)P(\eta')/P(\eta))P(c,d|m \in S_c). \tag{S22}$$

individual proposal: move an individual to a new population. First choose an individual i at random. If $|S_{q_i}| > 1$ propose a new state η' with i moved to a population chosen uniformly from $(1, \cdots, K) \neq q_i$ and accept η' with probability:

$$\min(1, P(\eta')/P(\eta)). \tag{S23}$$

parameter proposal: moves all hyperparameters independently.

- Delta: set $\delta' = \delta x$ with $x \sim U(-1,1)$. Accept with probability $P(\eta|\delta')\Gamma(\delta'; k_\delta, \theta_\delta)/P(\eta|\delta)\Gamma(\delta; k_\delta, \theta_\delta)$, where the prior for δ is a gamma distribution with specified parameters; $(k_\delta, \theta_\delta) = (2, 0.01)$ throughout, providing a wide tailed distribution.

- F: set $f' = fx$ with $x \sim U(-1,1)$. Accept with probability $P(\eta|f')\Gamma(f'; k_f, \theta_f)/P(\eta|f)\Gamma(f; k_f, \theta_f)$, where the prior for f is a gamma distribution with specified parameters; $(k_\delta, \theta_\delta) = (2, 0.01)$ throughout.

S4 Theory linking PCA, STRUCTURE and fineSTRUCTURE

Introduction

In this section, we give (in the form of propositions, later backed up empirically) a detailed technical description of results described in the main paper regarding the links between different approaches to infer and analyse population structure. All results in this section refer to *unlinked* markers and we assume *haploid* data, though all results also extend to the diploid case. In a number of situations that we will highlight, we have used these results to naturally extend to the linked case.

Since the subsequent material is somewhat detailed and technical, we begin with a summary of the results, which all relate to *unlinked* markers:

Proposition 1. We can approximately regard our coancestry matrix (see main text) as a rescaled version of a matrix commonly used to perform principal components analysis (PCA) on genetic data (Price et al. 2006).

Proposition 2. For all forms of the normal-approximation likelihood (covering a wide class of models) the PCA matrix contains all of the information available on population structure. The form of this likelihood demonstrates a close link between model-based analyses such as that performed by STRUCTURE (Pritchard et al. 2000), and PCA analyses of the type referred to in Proposition 1. Specifically, we show that for large datasets, the PCA matrix of Proposition 1 forms a set of sufficient statistics for the "STRUCTURE likelihood" – i.e. the likelihood of the data used by STRUCTURE's "no linkage" model, and by other similar software applications. Thus, in practice we expect that almost all the information accessible to even model-based approaches is contained within the PCA matrix, and hence also within our coancestry matrix. The permitted population structure is very general and therefore includes both discrete population models and continuous models capturing admixture.

Proposition 3. We derive an asymptotic approximation of the likelihood from Proposition 2 that takes a particularly simple form under the assumption of a large number of individuals and small drift.

Proposition 4. Provided population structure is not very strong, and again for large datasets, the multinomial likelihood form that we use in fineSTRUCTURE gives the same asymptotic likelihood as that used by STRUCTURE (Pritchard et al. 2000) found in Proposition 3. The proof also leads to an explicit rescaling of the multinomial likelihood (our "c" factor), and implies an algorithm to infer c in general, whose efficacy we test via simulation in a range of scenarios.

The results of Propositions 1 and 4, in particular, naturally motivate our approaches for analysing linked data, which we describe within the relevant sections.

For clarity, this section has slightly modified notation. We begin by defining quantities used throughout this section.

Notes on notation

We will make heavy use of vector notation. Matrices and vectors will be in bold, and scalar quantities will be in italics. For a matrix $\mathbf{M}$ of dimension $M \times N$, M_{ij} selects an element, $\mathbf{M}_{i*}$ selects the ith row and $\mathbf{M}_{*j}$ selects the jth column. We define $\mathbf{1}$ as the (appropriate length) vector containing all ones, hence $\mathbf{M1} = \mathbf{M}_{.*}$ is the vector of row sums, $\mathbf{1}^T\mathbf{M} = \mathbf{M}_{*.}$ is the vector of column sums, and $\mathbf{1}^T\mathbf{M1} = M_{..}$ is the sum over all elements of $\mathbf{M}$. Where ambiguity is possible we denote the length of one-vectors by subscripts e.g. $\mathbf{1}_L$. Note that unlike in the main text we do not use lower case to denote elements.

Definitions and Initial Assumptions

As in the main text, suppose we have data for N individuals, drawn from K populations, and let q_i be the population of individual i. Let L be the total number of segregating sites (after appropriate filtering). Let n_a be the total number of individuals from each population a.

We define $\mathbf{D}$ to be the *raw data matrix*. Specifically, elements D_{il} are binary, taking the value 1 if haplotype i carries that form at SNP l and 0 otherwise. We assume no missing data for the purposes of our derivations. We also assume all sites are biallelic, with sites with one or fewer occurrences of 0 or 1 removed and regarded as uninformative.

We define $\mathbf{X}$ to be the observed *coancestry matrix* for these data (see main text, where this is called x), where X_{ij} is the total *expected* number of chunks – each consisting of exactly one SNP in the unlinked case – that individual i copies from individual j. We view this expectation as calculated using the Li and Stephens (2003) model for genetic data applied with an infinite recombination rate, and arbitrarily small mutation rate in this setting, and thus this matrix is determined by the data $\mathbf{D}$. By the definition of the Li and Stephens algorithm, we have:

$$X_{ij} = \sum_{l=1}^{L} \left[\frac{D_{il}D_{jl}}{\sum_{k \neq i} D_{kl}} + \frac{(1 - D_{il})(1 - D_{jl})}{\sum_{k \neq i}(1 - D_{kl})} \right].$$

(Note that if a non-zero mutation rate is instead used, the matrix obtained is simply a rescaled version of this, with a constant term added, except at SNPs with seen only once in the sample, which we are assuming have been removed from the data.)

We define the *parameter matrix* $\mathbf{P}$ to be the underlying donor matrix, where $L\mathbf{P}$ gives the expected values for each entry of $\mathbf{X}$ and so P_{ij} gives the fraction of chunks that individual i is expected to copy from individual j. In particular, note we alter notation slightly from the text and view $\mathbf{P}$ as an $N \times N$ matrix (which will have a block-like structure if individuals are assigned to discrete populations).

We define the *Eigenstrat PCA matrix,* $\mathbf{E}$ to be the matrix used for principal components analysis in the Eigenstrat approach (Price et al. 2006). To construct this matrix, viewing mutations as corresponding to columns in the data matrix, columns are first scaled to have zero mean and unit variance, and subsequently the rescaled data matrix is multiplied by its transpose, to give the following:

$$E_{ij} = \sum_{l=1}^{L} \frac{(D_{il} - \widehat{f_l})(D_{jl} - \widehat{f_l})}{\widehat{f_l}(1 - \widehat{f_l})}.$$

Here, we define $\widehat{f_l}$ to be the observed sample frequency of the "1" allele at SNP l:

$$\widehat{f_l} = \frac{1}{N} \sum_{i=1}^{N} D_{il}$$

We will make use of several assumptions, which are summarized here and more fully defined when they are introduced.

A1. **No linkage disequilibrium** between loci within populations**.** Loci are independent conditional on underlying population assignments such that $cov(\mathbf{D}_{*l}, \mathbf{D}_{*l'}) = 0$ within populations for $l \neq l'$.

A2. **Large sample size.** We only require the leading order contribution in terms of N^{-1} where N is the number of individuals sampled.

A3. **Normally distributed drift.** Individuals are sampled in such a way that the drift can be approximated by a normal distribution. In population models, this means that the sample contains many individuals per population, and that rare SNPs are excluded.

A4. **Large number of loci.** We only require the leading order contribution in terms of L^{-1} where L is the number of loci.

A5. **Weak drift.** We only require the leading order contribution of the drift, defined in Section S4.2.

A6. (Technical assumption). The distribution of the (weighted) average frequency across all sampled individuals does not contain information on population structure.

A7. **More loci than individuals.** The number of individuals N is small in relation to the number of loci L.

A8. (Technical assumption). We require to integrate out the ancestral SNP frequencies, for which we assume that any prior on these is weak compared to the likelihood.

S4.1: The coancestry matrix and the PCA matrix

The main result of this section is the following proposition.

PROPOSITION 1: THE PCA AND COANCESTRY MATRICES ARE RELATED

For haploid variation data where sites are treated as unlinked (A1), the off-diagonal elements of the Eigentrat PCA matrix $\mathbf{E}$ and the observed coancestry matrix **X** are approximately related by the following equation:

$$E_{ij} = (N-1)X_{ij} - L + O(N^{-1}), \qquad i \neq j.$$

Proof:

Let $\widehat{f_l}' = \frac{1}{N-1}\sum_{k \neq i} D_{kl}$ be the empirical frequency of all SNPs excluding individual i. From the definition:

$$\begin{aligned} X_{ij} &= \sum_{l=1}^{L}\left[\frac{D_{il}D_{jl}}{\sum_{k\neq i} D_{kl}} + \frac{(1-D_{il})(1-D_{jl})}{\sum_{k\neq i}(1-D_{kl})}\right] \\ &= \frac{1}{N-1}\sum_{l=1}^{L}\left[\frac{D_{il}D_{jl}}{\widehat{f_l}'} + \frac{(1-D_{il})(1-D_{jl})}{1-\widehat{f_l}'}\right] \\ &= \frac{1}{N-1}\sum_{l=1}^{L}\left[\frac{(D_{il}-\widehat{f_l}')(D_{jl}-\widehat{f_l}')}{\widehat{f_l}'(1-\widehat{f_l}')} + 1\right] \end{aligned}$$

$$= \frac{L}{N-1} + \frac{1}{N-1}\sum_{l=1}^{L}\left[\frac{(D_{il}-\widehat{f}_l)(D_{jl}-\widehat{f}_l) + \frac{1}{N-1}(D_{il}-\widehat{f}_l)^2}{\widehat{f}_l(1-\widehat{f}_l) - \frac{1}{N-1}(D_{il}-\widehat{f}_l)^2}\right]$$

$$= \frac{L}{N-1} + \frac{1}{N-1}\sum_{l=1}^{L}\left[\frac{(D_{il}-\widehat{f}_l)(D_{jl}-\widehat{f}_l)}{\widehat{f}_l(1-\widehat{f}_l)}\right] + \mathrm{O}(N^{-2})$$

$$= \frac{L}{N-1} + \frac{E_{ij}}{N-1} + \mathrm{O}(N^{-2})$$

The proposition follows on rearrangement for E_{ij}.■

Note that we can write $\mathbf{X} \approx (\mathbf{E} - diag(\mathbf{E}))/(N-1)$.

Using properties of the PCA matrix, we can then show that this means the Eigenstrat PCA matrix **E** and the observed coancestry matrix **X** have similar eigenvectors.

COROLLARY 1: THE PCA AND COANCESTRY MATRICES HAVE THE SAME EIGENVECTORS

For haploid variation data where sites are treated as unlinked, assuming large population size (A2) the PCA matrix **E** and the observed coancestry matrix **X** have approximately identical eigenvectors.

Proof:
First note that

$$E(E_{ii}) = E\left[\sum_{l=1}^{L}\frac{\left(D_{il}-\widehat{f_l}\right)^2}{\widehat{f_l}\left(1-\widehat{f_l}\right)}\right] = E\left[\sum_{l=1}^{L}\frac{{D_{il}}^2-2D_{il}\widehat{f_l}+\widehat{f_l}^2}{\widehat{f_l}\left(1-\widehat{f_l}\right)}\right] = \sum_{l=1}^{L}\frac{\widehat{f_l}-2\widehat{f_l}^2+\widehat{f_l}^2}{\widehat{f_l}\left(1-\widehat{f_l}\right)} = L,$$

since $E\left({D_{il}}^2\right) = E(D_{il}) = g_l$, the ancestral frequency of the SNP. Additionally, assuming that drift is small, we can replace the denominator $\widehat{f_l}\left(1-\widehat{f_l}\right)$ by $g_l(1-g_l)$ and the result follows. (Note that in many sensible models $g_l = \widehat{f_l}$ as N becomes large, so the result may hold even when drift is moderate).

Now note that the row and column sums of **E** are identically zero, so trivially the N-vector **1** is an eigenvector with eigenvalue 0. Similarly the same eigenvector has eigenvalue 1 for the co-ancestry matrix **X**. Let **v** be any *other* eigenvector of **E** with eigenvalue λ. We must then have $\sum_{i=1}^{N} v_i = 0$ and so

$$\begin{aligned}\mathbf{Xv} &= \frac{1}{N-1}\left[\mathbf{Ev} - diag(E_{11}, E_{22}, \cdots, E_{NN})\mathbf{v}\right] + \frac{L}{N-1}\sum_{i=1}^{N} v_i + O(N^{-1}) \\ &\approx \frac{1}{N-1}[\lambda - L]\mathbf{v} + O(N^{-1}).\end{aligned}$$

Therefore **v** is also an eigenvector of **X** with transformed eigenvalue $\frac{1}{N-1}[\lambda - L]$ (when N is large, hence the approximation) as required.■

Discussion of Proposition 1:
The above proposition suggests that a principal components approach based on the coancestry matrix should yield results comparable to those from using standard approaches. We confirmed this fact in the main text. Note that although the proposition excludes the diagonal of the PCA matrix, in practice since each row or column sum of this matrix is exactly zero, we do not expect there to be (much) information from these diagonals. In our coancestry matrix, each row and column automatically sums to 1, with

diagonal entries set to zero, and so the information "lost" using this approach, relative to the standard PCA approach, is the value of the row and column sums in the PCA matrix. However, under weak assumptions we have shown that asymptotically the expected value of each element along the diagonal of the PCA matrix is L, so the off-diagonal sums in the PCA matrix have expected value $-L$, and add little or no information about population structure.

In implementing these ideas in practice, we modified our matrix slightly to ensure that eigenvalues were ranked equivalently between our approach and the PCA matrix itself (with the most historically relevant eigenvalues taking the largest values). By corollary 1, the eigenvalues of $\mathbf{X}$ are shifted and scaled relative to those from the PCA matrix, with one large eigenvalue equal to 1. To fix this, we first removed the shift by setting diagonal elements equal to the column sums. In the unlinked case, this is identically equal to $\frac{1}{N-1}L$ for each column, and adding these diagonal values trivially leaves eigenvectors unchanged and increases eigenvalues by this constant, removing the shift. To rescale the large eigenvalue to the value zero, we next subtract column means from each entry of the matrix. Again, in the unlinked case this simply removes $\frac{1}{N-1}L$ from each entry of the matrix, so has no effect on either the eigenvalues or eigenvectors, apart from the large eigenvalue corresponding to the *N*-vector $\mathbf{1}$. Finally, the resultant matrix $\mathbf{X}'$will have eigenvalues that are simply rescaled by a factor $\frac{1}{N-1}$ relative to the PCA matrix.

Our new PCA approach obviously extends trivially to using the equivalent coancestry matrix in the *linked* case. For this case, we made one small final modification, to account for the fact that in the linked case our coancestry matrix need not be exactly symmetric, with the resultant drawback that left and right eigenvectors will differ slightly. To fix this, we note that if $\mathbf{G}$ is a general symmetric matrix, then the eigenvectors of $\mathbf{G}\mathbf{G}^T$are identical. Motivated by this fact, and following other PCA approaches, we performed PCA in the general case on the matrix $\mathbf{X}'\mathbf{X}'^T$. In the linked case, this symmetrisation appeared to improve results slightly – other approaches we tried did not yield obvious improvements.

S4.2: The coancestry matrix and model-based approaches to inferring structure

Modelling population structure using a normally distributed drift matrix

In general, population structure between separated groups is often modelled using the concept of genetic drift between populations (for details, see e.g. Pritchard et al. (2000), Nicholson et al. (2002), Patterson et al. (2006)). We begin by defining g_{la} to be the frequency of the mutation at the lth locus in population a. Several models assume a joint prior distribution on g_{la} with some shared mean f_l and variance matrix $f_l(1-f_l)\mathbf{G}$. Conceptually, f_l can be thought of as representing the frequency of the mutation in an ancestral population before population-specific genetic drift results in new frequencies in each separate population group. The matrix $\mathbf{G}$ defines the covariance structure of this drift between populations. This allows, in general, for correlated drift among populations. Let us go further and assume as in the original Nicholson et al. (2002) formulation (which approximates genetic diffusion models) that

$$\mathbf{g}_l = (g_{l1}, g_{l2}, \cdots, g_{lK}) \sim N(f_l \mathbf{1}_K, f_l(1-f_l)\mathbf{G})$$

This "Normal drift" approximation is expected to be quite accurate, provided drift $\mathbf{G}$ is relatively small and (as below) provided it is applied to SNPs that are not at frequency very close to 0 or 1, a condition reasonably appropriate for many datasets involving ascertained SNPs, and which can be easily imposed more generally.

In all that follows, we use only the above assumption regarding drift, and thus our results apply across a fairly general range of settings. In particular, allowing for correlated drift means that if populations successively split, in a tree-like structure, the results still follow. Further, the no-linkage admixture model used by, for example, the program STRUCTURE (Pritchard et al. 2000) can be thought of as simply using an appropriate choice of $\mathbf{G}$ (with each individual representing a population).

Now at an individual level, define f_{li} to be the frequency for SNP l in the population q_i to which individual *i* belongs: $f_{li} = g_{lq_i}$. From the above, we have immediately that

$$\mathbf{f}_l | f_l = (f_{l1}, f_{l2}, \cdots, f_{lN}) | f_l \sim N(f_l \mathbf{1}_N, f_l(1-f_l)\mathbf{H})$$

where $\mathbf{H}$ is the individual-level matrix giving drift (from $\mathbf{G}$) between pairs of individuals. (This matrix is singular in general, but that will not affect our analysis.) We will formulate our likelihoods in terms of $\mathbf{f}_l$. (A notation reminder: $\mathbf{f}_l$ is the vector of population SNP frequencies, and f_l is the ancestral frequency).

<u>Constructing an approximate likelihood</u>

For the remainder of this section, we study inference under normally distributed drift, using an approximation to the likelihood of the data. For this approximation to be valid, we assume (A3) that the number n_a of individuals sampled from each population a, and corresponding population-specific allele frequencies f_{la} are such that we may use a Normal approximation to the binomial sampling likelihood of the observed data. In practice, this means n_a should be large, at least ≥20, and f_{la} should not be very close to 0 or 1.

Later, we will need an additional technical assumption (A6) that mean allele frequencies averaged over all populations are uninformative for population structure. Provided that drift is weak we expect that typically these can contain at most weak information about the underlying population assignments, justifying this assumption.

We spend the remainder of this section deriving a form for the approximate likelihood, under these two (main) assumptions.

PRELIMINARY TO PROPOSITION 2: SINGLE LOCUS LIKELIHOOD

Under assumptions A1-3, and assuming also that a weak prior is placed on the ancestral SNP frequency (A8), the likelihood for a single locus l can be expressed in the following form:

$$L(\mathbf{D}_{*l}) \propto \|\boldsymbol{\Sigma}\|^{-\frac{1}{2}}[\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}]^{-\frac{\mathbf{1}}{2}} exp\left[-\frac{1}{2\hat{f}_l(1-\hat{f}_l)}\left(\mathbf{D}_{*\boldsymbol{l}}-\hat{f}_l\mathbf{1}\right)^{\boldsymbol{T}}\left(\boldsymbol{\Sigma}^{-\mathbf{1}} - \frac{\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}1^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}}{\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}}\right)\left(\mathbf{D}_{*\boldsymbol{l}}-\hat{f}_l\mathbf{1}\right)\right].$$

Where $\boldsymbol{\Sigma} = \mathbf{I}+\mathbf{H}$ is the (non-singular) covariance matrix.

Proof and derivation:
Conditional on the underlying allele frequencies, the observed allele counts in a (haploid) individual i from population a are independent among individuals, and simply Bernoulli with mean $f_{li} = g_{la}$ and variance $f_{li}(1 - f_{li})$. By the assumptions, we may approximate the likelihood by behaving as if data are taken from a Normal distribution. If drift is weak, the overall sample mean frequency $\hat{f}_l$ can be used to approximate f_{li} in the variance term and we obtain a joint distribution for the data vector $\mathbf{D}_{*\boldsymbol{l}}$ for all N individuals, at SNP l:

$$\mathbf{D}_{*\boldsymbol{l}}|\, f_{l1}, f_{l2}\cdots, f_{lK} \dot{\sim} N\left(\mathbf{f}_l, \hat{f}_l\left(1-\hat{f}_l\right)\mathbf{I}\right).$$

Making the similar approximation

$$\mathbf{f}_l = (f_{l1}, f_{l2}, \cdots, f_{lN}) \sim N(f_l\mathbf{1}_{\mathbf{N}}, f_l(1-f_l)\mathbf{H}) \dot{\sim} N\left(f_l\mathbf{1}_{\mathbf{N}}, \hat{f}_l\left(1-\hat{f}_l\right)\mathbf{H}\right),$$

integrating out the population frequencies we have by properties of Normal distributions:

$$\mathbf{D}_{*l}|f_l \dot{\sim} N\left(f_l\mathbf{1}_{\mathbf{N}}, \hat{f}_l(1-\hat{f}_l)[\mathbf{I}+\mathbf{H}]\right).$$

Using the notation $\boldsymbol{\Sigma} = \mathbf{I}+\mathbf{H}$, the likelihood is then:

$$L(\mathbf{D}_{*\boldsymbol{l}}|f_l) = \left[2\pi\hat{f}_l(1-\hat{f}_l)\right]^{-N/2}\|\boldsymbol{\Sigma}\|^{-1/2} exp\left[-\frac{1}{2\hat{f}_l(1-\hat{f}_l)}(\mathbf{D}_{*\boldsymbol{l}}-f_l\mathbf{1})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*\boldsymbol{l}}-f_l\mathbf{1})\right].$$

Define

$$\mathrm{a} = \frac{\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}}{\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}}.$$

Let $Y = \mathbf{a}^{\boldsymbol{T}}\mathbf{D}_{*l}$. Clearly Y is normally distributed:

$$Y|f_l \sim N\left(\boldsymbol{a}^{\boldsymbol{T}} f_l \mathbf{1}_{\mathbf{N}}, \widehat{f_l}\left(1-\widehat{f_l}\right)\mathbf{a}^{\boldsymbol{T}}\boldsymbol{\Sigma}\mathbf{a}\right) \sim N\left(f_l, \frac{\widehat{f_l}\left(1-\widehat{f_l}\right)}{\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}}\right).$$

Thus, Y has mean f_l and variance which decreases as we increase sample size.

We can rewrite the exponent term in our likelihood:

$$\begin{aligned}
&(\mathbf{D}_{*\boldsymbol{l}} - f_l\mathbf{1})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*\boldsymbol{l}} - f_l\mathbf{1}) = (\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y + \mathbf{1}Y - f_l\mathbf{1})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y + \mathbf{1}Y - f_l\mathbf{1}) \\
&= (\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y)^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y) + 2(\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y)^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{1}Y - \mathbf{1}f_l) + (\mathbf{1}Y - \mathbf{1}f_l)^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{1}Y - \mathbf{1}f_l) \\
&= (\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y)^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y) + 2(\mathbf{D}_{*\boldsymbol{l}})^{\boldsymbol{T}}\left(\mathrm{I} - \mathbf{a}1^{\boldsymbol{T}}\right)\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{1}(Y - f_l) + (Y - f_l)\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{1}(Y - f_l) \\
&= (\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y)^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y) + 2(\mathbf{D}_{*\boldsymbol{l}})^{\boldsymbol{T}}(\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{1} - \boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1})(Y - f_l) + \mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}(Y - f_l)^{\mathbf{2}} \\
&= (\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y)^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*\boldsymbol{l}} - \mathbf{1}Y) + \mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}(Y - f_l)^{\mathbf{2}}
\end{aligned}$$

and substituting back in we find:

$$\begin{aligned}
L(\mathbf{D}_{*l}|f_l) &= \left[2\pi\widehat{f_l}(1-\widehat{f_l})\right]^{-N/2}\|\boldsymbol{\Sigma}\|^{-1/2} exp\left[-\frac{1}{2\widehat{f_l}(1-\widehat{f_l})}((\mathbf{D}_{*l} - \mathbf{1}Y)^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*l} - \mathbf{1}Y)\right. \\
&\qquad\qquad \left. + \mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}(Y - f_l)^{\mathbf{2}})\right] \\
&= \left[2\pi\widehat{f_l}(1-\widehat{f_l})\right]^{-(N-1)/2}\|\boldsymbol{\Sigma}\|^{-1/2}[\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}]^{-\mathbf{1}/\mathbf{2}} exp\left[-\frac{1}{2\widehat{f_l}(1-\widehat{f_l})}(\mathbf{D}_{*l} - \mathbf{1}Y)^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*l}\right. \\
&\qquad\qquad \left. - \mathbf{1}Y)\right] \times L(Y|f_l)
\end{aligned}$$

where only the final term in the likelihood depends on f_l.

Intuitively, we view Y as a weighted mean of the individual allele counts, specifically as an estimator of the ancestral allele frequency f_l. We have shown that Y is a sufficient statistic for estimating f_l. Further we can then write

$$L(\mathbf{D}_{*\boldsymbol{l}}|f_l) = L(\mathbf{D}_{*\boldsymbol{l}}|Y)L(Y|f_l).$$

The ancestral allele frequency at an individual SNP is often <u>not</u> of direct interest in inferring structure, and is typically integrated out of the likelihood as a nuisance parameter. Therefore, suppose that we have placed some prior distribution h on f_l. (For example, the original implementation of STRUCTURE uses a $\mathrm{U}(0,1)$ prior for h.) We will also suppose that in most practical settings, this prior is relatively diffuse while the data are more informative, so that over the support of $L(Y|f_l)$ we may view h as unvarying: $h(f_l) \approx h(y)$. We now may approximately integrate out the ancestral allele frequency to give unconditionally:

$$L(\mathbf{D}_{*l}) = L(\mathbf{D}_{*l}|Y = y)\int_{-\infty}^{\infty} L(y|f_l)\, h(f_l) df_l \approx L(\mathbf{D}_{*l}|Y = y)h(y)\int_{-\infty}^{\infty} L(y|f_l)\, df_l$$

$$= L(\mathbf{D}_{*l}|Y = y)h(y)$$

In our setting, we are interested in inferring structure. Thus, terms that concern us in the likelihood are only those that depend on this structure, which is wholly characterised by the covariance matrix $\boldsymbol{\Sigma}$ and so up to a constant of proportionality:

$$L(\mathbf{D}_{*l}|f_l) \propto \|\boldsymbol{\Sigma}\|^{-1/2}[\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}]^{-1/2}exp\left[-\frac{1}{2\widehat{f_l}(1-\widehat{f_l})}(\mathbf{D}_{*l}-\mathbf{1}Y)^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*l}-\mathbf{1}Y)\right].$$

In the large-sample size setting, this likelihood approximates the truth and contains almost all available information about population structure (ignoring information present in the overall average allele frequency Y at a given SNP).

We must take only one more simplifying step, and that is to rewrite the above in terms of $\widehat{f_l}$ rather than Y. Specifically, note that:

$$\begin{aligned}
(\mathbf{D}_{*l}-\mathbf{1}Y)^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{D}_{*l}-\mathbf{1}Y) &= (\mathbf{D}_{*l})^{\boldsymbol{T}}\left(\mathrm{I}-\mathbf{a}\mathbf{1}^{\boldsymbol{T}}\right)\boldsymbol{\Sigma}^{-\mathbf{1}}\left(\mathrm{I}-\mathbf{1}\mathrm{a}^{\boldsymbol{T}}\right)(\mathbf{D}_{*l}) \\
&= (\mathbf{D}_{*l})^{\boldsymbol{T}}\left(\mathrm{I}-\frac{\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}}\right)\boldsymbol{\Sigma}^{-\mathbf{1}}\left(\mathrm{I}-\frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}}{\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}}\right)(\mathbf{D}_{*l}) \\
&= (\mathbf{D}_{*l})^{\boldsymbol{T}}\left(\boldsymbol{\Sigma}^{-\mathbf{1}}-\frac{\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}\mathbf{1}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}}{\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}}\right)(\mathbf{D}_{*l}) \\
&= \left(\left[\boldsymbol{I}-\frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{\mathbf{1}^{\mathbf{T}}\,\mathbf{1}}\right]\mathbf{D}_{*l}\right)^{\boldsymbol{T}}\left(\boldsymbol{\Sigma}^{-\mathbf{1}}-\frac{\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}\mathbf{1}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}}{\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}}\right)\left(\left[\boldsymbol{I}-\frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{\mathbf{1}^{\mathbf{T}}\,\mathbf{1}}\right]\mathbf{D}_{*l}\right) \\
&= \left(\mathbf{D}_{*l}-\widehat{f_l}\mathbf{1}\right)^{\boldsymbol{T}}\left(\boldsymbol{\Sigma}^{-\mathbf{1}}-\frac{\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}\mathbf{1}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}}{\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}}\right)\left(\mathbf{D}_{*l}-\widehat{f_l}\mathbf{1}\right)
\end{aligned}$$

where in the last line, $\widehat{f_l}$ is simply the overall (unweighted) sample mean frequency. Finally, we have an approximate likelihood contribution for SNP l:

$$\begin{aligned}
L(\mathbf{D}_{*l}) \propto \|\boldsymbol{\Sigma}\|^{-\frac{1}{2}}[\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}]^{-\frac{\mathbf{1}}{2}}exp\Bigg[-\frac{1}{2\widehat{f_l}(1-\widehat{f_l})}\left(\mathbf{D}_{*l}-\widehat{f_l}\mathbf{1}\right)^{\boldsymbol{T}}\Bigg(\boldsymbol{\Sigma}^{-\mathbf{1}} \\
-\frac{\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}1^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}}{\mathbf{1}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\,\mathbf{1}}\Bigg)\left(\mathbf{D}_{*l}-\widehat{f_l}\mathbf{1}\right)\Bigg].\ \blacksquare
\end{aligned}$$

Given a set of drift parameters, and observed data $\mathbf{D}_{*l}$ for SNP l, the likelihood can readily be calculated for inference purposes. However, the more useful feature of this approximation is that it makes it straightforward to combine information across loci, which we show leads to a strong link between certain forms of the principal components matrix, and model-based approaches using likelihoods of the general form discussed above.

Linking the PCA/coancestry matrix and the model-based approach

Under our simplifying assumptions, we can now obtain a likelihood for the entire dataset by multiplying likelihoods across (unlinked) loci. The result is given by the following proposition, which easily extends to the genotype case.

PROPOSITION 2

Suppose we have haploid variation data where sites are treated as unlinked (A1), that we sample $n_1, n_2, \cdots, n_K$ individuals from each of K underlying populations, where each n_k is large, and that we exclude very rare SNPs (A2-3). Consider a general model of normally distributed allele frequency drift at each locus, with an individual level drift covariance matrix $\mathbf{H}$. If the prior placed on the ancestral allele frequency is weak (A8), and if we assume there is no information about drift from the overall allele frequency of each mutation in the sample (A6), then the likelihood of the data $\mathbf{D}$ can be approximated solely in terms of the "Eigenstrat" PCA matrix $\mathbf{E}$ from Proposition 1:

$$L(\mathbf{H}\,|\mathbf{D}) \propto exp\left(\frac{L}{2}tr[\log(\boldsymbol{\Sigma}^{-1})] - \frac{L}{2}\log((\boldsymbol{\Sigma}^{-1})_{..}) - \frac{L}{2}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} E_{ij}\left(\Sigma_{ij}^{-1} - \frac{\Sigma_{i.}^{-1}\Sigma_{j.}^{-1}}{\Sigma_{..}^{-1}}\right) - \sum_{i=1}^{N} E_{ii}\right]\right)$$

where $\boldsymbol{\Sigma} = \mathbf{I} + \mathbf{H}$.

Hence in this model, *all* available information regarding the underlying population structure is contained in the PCA matrix.

Proof:

The stated assumptions mean that the result derived in Preliminary Proposition 2 holds. Hence, multiplying the likelihood across L independent sites we obtain:

$$L(\mathbf{H}\,|\mathbf{D}) \propto \|\boldsymbol{\Sigma}\|^{-\frac{L}{2}}[(\boldsymbol{\Sigma}^{-1})_{..}]^{-\frac{L}{2}}exp\left[-\sum_{l=1}^{L}\frac{1}{2\hat{f}_l(1-\hat{f}_l)}\left(\mathbf{D}_{*l} - \hat{f}_l\mathbf{1}\right)^{\boldsymbol{T}}\left(\boldsymbol{\Sigma}^{-1} - \frac{\Sigma_{i.}^{-1}\Sigma_{j.}^{-1}}{\Sigma_{..}^{-1}}\right)\left(\mathbf{D}_{*l} - \hat{f}_l\mathbf{1}\right)\right]$$

$$\propto \|\boldsymbol{\Sigma}\|^{-\frac{L}{2}}[(\boldsymbol{\Sigma}^{-1})_{..}]^{-\frac{L}{2}}exp\left[-\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{l=1}^{L}\frac{1}{2\hat{f}_l(1-\hat{f}_l)}\left(D_{il} - \hat{f}_l\right)\left(\boldsymbol{\Sigma}^{-1} - \frac{\Sigma_{i.}^{-1}\Sigma_{j.}^{-1}}{\Sigma_{..}^{-1}}\right)_{ij}\left(D_{jl} - \hat{f}_l\right)\right]$$

$$\propto \|\boldsymbol{\Sigma}\|^{-\frac{L}{2}}[(\boldsymbol{\Sigma}^{-1})_{..}]^{-\frac{L}{2}}exp\left(-\frac{L}{2}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}E_{ij}\left(\Sigma_{ij}^{-1}-\frac{\Sigma_{i.}^{-1}\Sigma_{j.}^{-1}}{\Sigma_{..}^{-1}}\right)-\sum_{i=1}^{N}E_{ii}\right]\right)$$

(noting that the last term in the exponent is a constant, so does not affect proportionality: this is included to avoid a large term in the exponent for $i = j$), from which the required result is immediate by properties of determinants. ■

This result is general for the study of datasets containing mainly common mutations. It applies in both models incorporating discrete population structure, and also in no-linkage admixture models, where sample size is in a sense "large", in that the Normal approximation to the likelihood still applies. An implication is that apart from the weak information contained in the average allele frequencies of mutations across groups, and neglecting information from occasional variants that reach loss or fixation in some groups (expected to occur for stronger drift), the Eigenstrat style PCA matrix contains all the available information about population structure. Thus in these models, we expect model-based approaches to generally only succeed if there is a signal of structure from the PCA approach, as has been observed previously in practice (S5).

S4.3: Asymptotic behaviour of models of population structure, with weak drift

If population structure is strong, we anticipate that all reasonable approaches are likely to identify it. Of more interest is then the setting where population structure is weak (in a sense we will specify precisely below). In this section, we show that in the setting of weak structure the likelihood of a particular underlying drift matrix **H** has a particularly simple approximate form. This likelihood is a function of our coancestry matrix. Further, we show that the likelihood form is approximately equivalent to that obtained by assuming that an appropriate rescaling of the coancestry matrix yields a multinomial distribution. Thus, in the weak drift case and for unlinked sites, the likelihood form we use in the main text (with an appropriate choice of scaling parameter) approximates the full likelihood of the data. (In the next section we will show how we generalise these ideas to the linked case, and describe our approach for estimating c.)

PROPOSITION 3: ASYMPTOTIC BEHAVIOR FOR SUBTLE POPULATION STRUCTURE

Assume that the conditions of Proposition 2 hold (A1-3,A6,A8), and also assumptions A4 and A5, where (A5) is precisely that the drift is small in the sense that order $\mathbf{H}^3$ and higher terms are negligible. The likelihood given by Proposition 2 can be simplified to

$$L(\mathbf{H}\,|\mathbf{D}) \propto exp\left(-\frac{L}{4}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[Z_{ij}-Q_{ij}\right]^{2}\right),$$

Where $\mathbf{Q}$ is the row and mean zero'd drift matrix with $Q_{ij} = H_{ij} - \frac{1}{N}H_{i\cdot} - \frac{1}{N}H_{\cdot j} + \frac{1}{N^2}H_{\cdot\cdot}$, and $Z_{ij} = \frac{1}{L}\left(E_{ij} + \frac{L}{N}\right)$ is the per-locus PCA matrix.

Proof:

We will approximate terms in the likelihood shown in Proposition 2. Recalling $\mathbf{\Sigma} = \mathbf{I} + \mathbf{H}$, under assumption A5 we then have correct to second order:

$$
\begin{aligned}
\mathbf{\Sigma^{-1}} &= \mathbf{I} - \mathbf{H} + \mathbf{H^2} + \cdots \\
\Sigma_{\cdot\cdot}^{-1} &= N - H_{\cdot\cdot} + H_{\cdot\cdot}^{2} + \cdots \\
\log((\mathbf{\Sigma^{-1}})_{\cdot\cdot}) &= \log N - \frac{1}{N}H_{\cdot\cdot} + \frac{1}{N}H_{\cdot\cdot}^{2} - \frac{1}{2N^2}[H_{\cdot\cdot}]^2 + \cdots \\
\log(\mathbf{\Sigma^{-1}}) &= -\mathbf{H} + \frac{1}{2}\mathbf{H^2} + \cdots \\
tr[\log(\mathbf{\Sigma^{-1}})] &= -tr[\mathbf{H}] + \frac{1}{2}tr[\mathbf{H^2}] + \cdots
\end{aligned}
$$

We start with the likelihood from Proposition 2:

$$
\begin{aligned}
L(\mathbf{H}\,|\mathbf{D}) \propto exp\Bigg(&\frac{L}{2}tr[\log(\mathbf{\Sigma^{-1}})] - \frac{L}{2}\log((\mathbf{\Sigma^{-1}})_{\cdot\cdot}) \\
&- \frac{L}{2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}E_{ij}\left(\Sigma_{ij}^{-1} - \frac{\Sigma_{i\cdot}^{-1}\Sigma_{j\cdot}^{-1}}{\Sigma_{\cdot\cdot}^{-1}}\right) - \sum_{i=1}^{n}E_{ii}\right]\Bigg)
\end{aligned}
$$

Again correct to second order and after a little simplification we find:

$$
\frac{L}{2}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}E_{ij}\left(\Sigma_{ij}^{-1} - \frac{\Sigma_{i\cdot}^{-1}\Sigma_{j\cdot}^{-1}}{\Sigma_{\cdot\cdot}^{-1}}\right) - \sum_{i=1}^{N}E_{ii}\right] \approx \frac{L}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}E_{ij}\left(-H_{ij} + H^2{}_{ij} - \frac{1}{N}H_{i\cdot}H_{j\cdot}\right).
$$

Note that viewing the PCA matrix as estimating the drift, and evaluating the expected asymptotic values for this matrix without drift as N becomes large, we have $Z_{ij} = \frac{1}{L}\left(E_{ij} + \frac{L}{N}\right) = O(H_{ij})$ in large datasets, unless $i = j$. For the $i = j$ case write $\frac{1}{L}\left(E_{ii} + \frac{L}{N}\right) = 1 + Z_{ii}$. Again, for large datasets, $Z_{ii} = O(H_{ii})$ is small. Then, correct to second order in H_{ij} we have:

$$\sum_{i=1}^{N}\sum_{j=1}^{N}E_{ij}\left(-H_{ij}+H^2{}_{ij}-\frac{1}{N}H_{i.}H_{j.}\right)$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\left(LZ_{ij}-\frac{L}{N}\right)\left(-H_{ij}+H^2{}_{ij}-\frac{1}{N}H_{i.}H_{j.}\right)$$

$$+L\sum_{i=1}^{N}\left(-H_{ii}+H^2{}_{ii}-\frac{1}{N}H_{i.}H_{i.}\right)$$

$$\approx L\left[\left(\sum_{i=1}^{N}\sum_{j=1}^{N}-Z_{ij}H_{ij}\right)+\frac{1}{N}H_{..}+\frac{1}{N^2}[H_{..}]^2-\frac{1}{N}[\mathbf{H}^2]_{..}-tr[\mathbf{H}]+tr[\mathbf{H}^2]-\frac{1}{N}[H_{..}]^2\right]$$

Substituting into the likelihood:

$$L(\mathbf{H}\,|\mathbf{D})\propto exp\left(\frac{L}{2}\left(-tr[\mathbf{H}]+\frac{1}{2}tr[\mathbf{H^2}]\right)+\frac{L}{2}\left(\frac{1}{N}H_{..}-\frac{1}{N}[\mathbf{H}^2]_{..}+\frac{1}{2N^2}[H_{..}]^2\right)\right.$$

$$-\frac{L}{2}\left[\left(\sum_{i=1}^{N}\sum_{j=1}^{N}-Z_{ij}H_{ij}\right)+\frac{1}{N}H_{..}+\frac{1}{N^2}[H_{..}]^2-\frac{1}{N}[H^2]_{..}-tr[\mathbf{H}]\right.$$

$$\left.\left.+\,tr[\mathbf{H}^2]-\frac{1}{N}[H_{..}]^2\right]\right)$$

and after simplification, we obtain

$$L(\mathbf{H}\,|\mathbf{D})\propto exp\left(-\frac{L}{2}\left(\frac{1}{2}tr[\mathbf{H^2}]\right)+\frac{L}{2}\frac{1}{N}[\mathbf{H}^2]_{..}-\frac{L}{2}\left(\frac{1}{2N^2}[H_{..}]^2\right)-\frac{L}{2}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}-Z_{ij}H_{ij}\right]\right)$$

$$L(\mathbf{H}\,|\mathbf{D})\propto exp\left(-\frac{L}{4}\left[tr[\mathbf{H^2}]-\frac{2}{N}[\mathbf{H}^2]_{..}+\frac{1}{N^2}[H_{..}]^2-2\sum_{i=1}^{N}\sum_{j=1}^{N}Z_{ij}H_{ij}\right]\right).$$

We now substitute $Q_{ij}=H_{ij}-\frac{1}{N}H_{i.}-\frac{1}{N}H_{.j}+\frac{1}{N^2}H_{..}$, which clearly has zero row and column means, so Q_{ij} is the relative drift among individuals. Further, by expansion of the desired result it is clear that all other terms from ${Q_{ij}}^2$ cancel, and ${Z_{ij}}^2$ is a constant independent of the parameters so can be included or excluded in the proportionality. The result immediately follows. ■

Discussion of Proposition 3:

This is the key result; to second order the likelihood only depends on the data through the usual principal components matrix $\mathbf{E}$ which is transformed to give Z_{ij}, and the transformed drift matrix $\mathbf{Q}$. Hence only the relative drift $\mathbf{Q}$ is identifiable from data, and the absolute overall value of drift cannot be inferred. Further, the likelihood behaves asymptotically as if the transformed entries Z_{ij} in the principal components matrix are independent and normally distributed, with mean $\mathbf{Q}$ and variance $\frac{2}{L}$.

We have used a series of approximations in deriving this result, which suppose essentially that the sample size is large, while drift is small. Examining the latter assumption in more detail, it can be seen that our approximations require the setting $\frac{H_{..}}{N} \ll 1$, so the average drift is small compared to $\frac{1}{N}$. By Price et al. (2009), overall structure is strong if

$$\frac{H_{..}}{N^2} \gg \sqrt{\frac{1}{LN}}$$

In this setting, our assumptions may not hold but we expect (and see in practice) all competitive approaches to perform well and identify structure. In the case where structure is much weaker, then for some non-negative k we can consider:

$$\frac{H_{..}}{N} \sim k\sqrt{\frac{N}{L}}$$

which will be considerably less than 1 provided $N \ll L$, as is usually the case in current genetic studies where there are many more markers than individuals. Nevertheless, it is clearly important to evaluate the performance of both our approximation to the likelihood, and the resulting inference framework, via simulation, which we do extensively in Section S6. The results verify excellent agreement between the theory and observed results.

S4.4: fineSTRUCTURE model

The likelihoods as written above produce a dimension reduction by avoiding the need to consider SNPs individually. However, this likelihood is not particularly convenient to work with directly – neither is it straightforward to extend to incorporate LD information. It is more natural to attempt to perform inference based on our coancestry matrices, which give expected counts, and *do* extend immediately to the LD case, while still giving a dimension reduction. We showed above that these matrices also relate closely to the PCA matrix, which our approximate likelihoods are defined in terms of.

For count data (and by extension our *expected* count data matrix), a natural model is the multinomial distribution, which is a member of the exponential family of distributions, enabling the use of computationally convenient conjugate prior choices. Although we have expected counts, we may nevertheless consider a model of the same form, where individuals are ordered. In general, we allow ourselves to multiply the count matrix by a constant $1/c$ before applying the likelihood. We can view this as calculating an *effective* number of loci across the genome.

In this section, we first attempt to identify a relationship between parameters in this multinomial likelihood, and the STRUCTURE model. The parameter in a multinomial model, for group m, is the probability an observation falls in this bin. Suppose we consider a general multinomial distribution for our setting, where we have a mean probability P_{ij} that individual i copies a SNP (or a chunk) from individual j. It is immediate from Proposition 1 that if N is large, $i \neq j$, the expected count proportion for individual i from individual j is expressible in terms of the expectation of the PCA matrix and to order N :

$$
\begin{aligned}
P_{ij} &= E\left(\frac{1}{L}X_{ij}\right) \\
&= \frac{1}{N-1} + \frac{1}{N-1}E\left[\frac{1}{L}E_{ij}\right] \\
&= \frac{1}{N} + \frac{1}{N-1}E[Z_{ij}] \\
&= \frac{1}{N} + \frac{1}{N-1}Q_{ij}.
\end{aligned}
$$

so we can relate this parameter to the individual by individual relative drift matrix $\mathbf{Q}$. Note we could conceptually also include the case $i = j$, even though we have disallowed self-copying in practice.

PROPOSITION 4: ASYMPTOTIC BEHAVIOUR OF THE MULTINOMIAL LIKELIHOOD

Assuming A1-6 and additionally (A7) that the whilst both the population size N and number of loci L are large, additionally $L \gg N$

$$L(\mathbf{X};\mathbf{P}) = \prod_{i=1}^{N} L(\mathbf{X}_i;\mathbf{P}_i),$$

$$L(\mathbf{X}_i;\mathbf{P}_i) = \left[\binom{\mathrm{L}}{X_{i1},\cdots,X_{in}} \prod_{j=1,j\neq i}^{N} {P_{ij}}^{X_{ij}}\right]^{1/\mathrm{c}},$$

has the same asymptotic form found in Proposition 3, with (for haploids) the choice $c = 2/(N-2)$.

Proof and derivation:

The expected total number of chunks copied by individual i from each other individual j is large provided drift is small and the number of loci is large compared to the number of individuals.

We again employ the central limit theorem, this time relying on the fact that the number of *loci* is large. We approximate the joint distribution of the counts $\mathbf{X}_i$ using a multivariate normal distribution of dimension $N-1$ (because $X_{ii} = \mu_{ii} = 0$ this entry is therefore removed from the likelihood). $\mathbf{X}_i$ is the count vector, $\boldsymbol{\mu}_i = L\mathbf{P}_i$ is the expected number of counts, and $\mathbf{\Sigma^i}$ is the model covariance matrix for row i with the other rows j. Since $\mathbf{\Sigma^i} = diag(L\mathbf{P}_i)$, we can avoid having to use the generalised inverse directly, and have the standard likelihood form:

$$L(\mathbf{X};\mathbf{P}) \propto \left(\prod_{i,j=1,j\neq i}^{N} {P_{ij}}^{-1/2c}\right) exp\left(\frac{-1}{2c}\sum_{i,j=1,j\neq i}^{N}\frac{\left(X_{ij}-LP_{ij}\right)^2}{LP_{ij}}\right)$$
$$= \left(\prod_{i,j=1,j\neq i}^{N} {P_{ij}}^{-1/2c}\right) exp\left(\frac{-L}{2c}\sum_{i,j=1,j\neq i}^{N}\frac{\left(\widehat{P_{ij}}-P_{ij}\right)^2}{P_{ij}}\right),$$

where $\widehat{P_{ij}} = X_{ij}/L$ is the empirical frequency of copying from individual i from individual j. Then since $\sum_{j=1,j\neq i}^{N} P_{ij} = \sum_{j=1,j\neq i}^{N} \widehat{P_{ij}} = 1$,

$$exp\left(\frac{-L}{2c}\sum_{j=1,j\neq i}^{N}\frac{\left(\widehat{P_{ij}}-P_{ij}\right)^2}{P_{ij}}\right) = exp\left(\frac{-L}{2c}\sum_{j=1,j\neq i}^{N}\frac{\widehat{P_{ij}}^2}{P_{ij}} + \frac{L}{2c}\right).$$

Using Proposition 1, and noting that $E\left(E_{ij}/L\right) = O(N^{-1})$:

$$\widehat{P_{ij}}^2 \approx (N-1)^{-2}\left(\frac{E_{ij}}{L}+1\right)^2 = (N-1)^{-2}\left(\frac{2E_{ij}}{L}+1+O(N^{-2})\right)$$
$$\approx (N-1)^{-2}\left(2Z_{ij}+1-2/\mathrm{N}\right)$$

Substituting for Q_{ij}, and discarding lower order terms in N, we have:

$$L(\mathbf{X};\mathbf{P}) \dot{\propto} \left(\prod_{i,j=1,j\neq i}^{N} {P_{ij}}^{-1/2c}\right) exp\left(\frac{-L}{2Nc}\sum_{i,j=1}^{N}\frac{1+2Z_{ij}-\frac{1}{N}}{1+Q_{ij}-\frac{1}{N}} + \frac{L^2}{2c}\right)$$

where Z_{ij} is defined as above, and where in the final line we include an additional term in the exponent corresponding to Z_{ii}, assuming that the relative contribution of this single within individual term to the likelihood is small if N is large.

The first term in this likelihood can be ignored for two reasons. Firstly, we can incorporate this term into the prior distribution on $\mathbf{P}_i$ since it does not depend on the data or the number of loci or individuals except for a constant of proportionality, for example by fitting the prior variance in copying fractions based on the data. Secondly, the contribution of this term is small. Substituting $P_{ij} = N^{-1}(1 + N(N-1)^{-1}Q_{ij})$ we have:

$$\left(\prod_{i,j=1,j\neq i}^{N} {P_{ij}}^{-\frac{1}{2c}}\right) = N^{\frac{N(N-1)}{2c}} exp\left(\frac{-1}{2c}\sum_{i,j=1,j\neq i}^{N} \log\left(1+\left[\frac{N}{N-1}\right]Q_{ij}\right)\right)$$

$$= N^{\frac{N(N-1)}{2c}} exp\left(\frac{-1}{2c}\sum_{i,j=1,j\neq i}^{N}\left[\left[\frac{N}{N-1}\right]Q_{ij} - \frac{1}{2}\left[\frac{N}{N-1}\right]^2 {Q_{ij}}^2 + \cdots\right]\right)$$

$$\approx N^{\frac{N(N-1)}{2c}} exp\left(\frac{1}{4c}\left[\frac{N}{N-1}\right]^2 \sum_{i,j=1}^{N} {Q_{ij}}^2\right).$$

(Here we have discarded higher order terms in N and $\mathbf{Q}$, and noted that rows of $\mathbf{Q}$ sum to 0). The constant term can be incorporated into the proportionality, and the term inside the exponent is small compared to the remaining term in the likelihood. Choosing $c = \frac{2N}{(N-1)^2} = \frac{2}{N-2} + O(N^{-3})$, we see that for small drift ($L \gg N$) we can discard the above term, leading to:

$$L(\mathbf{X};\mathbf{P}) \propto exp\left(\frac{-L}{4}\frac{(N-1)^2}{N^2}\sum_{i,j=1}^{N}\frac{1+2Z_{ij}-\frac{1}{N}}{1+Q_{ij}-\frac{1}{N}}\right)$$

$$= exp\left(\frac{-L}{4}\frac{(N-1)^2}{N^2}\left[\sum_{i,j=1}^{N}\frac{1+2\left[\frac{N}{N-1}\right]Z_{ij}}{1+\left[\frac{N}{N-1}\right]Q_{ij}}\right]\right)$$

$$\approx exp\left(\frac{-L}{4}\frac{(N-1)^2}{N^2}\left[\sum_{i,j=1}^{N}\left(1+2\left[\frac{N}{N-1}\right]Z_{ij}\right)\left(1-\left[\frac{N}{N-1}\right]Q_{ij}+\left[\frac{N}{N-1}\right]^2{Q_{ij}}^2\right)\right]\right)$$

$$\approx exp\left(\frac{-L}{4}\frac{(N-1)^2}{N^2}\left[N^2+\sum_{i,j=1}^{N}\left[\frac{N}{N-1}\right]^2 Q_{ij}{}^2-2\sum_{i,j=1}^{N}\left[\frac{N}{N-1}\right]^2 Q_{ij}Z_{ij}\right]\right).$$

On simplification and discarding higher than second order terms in the drift we have

$$L(\mathbf{X};\mathbf{P})\propto exp\left(-\frac{L}{4}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[Z_{ij}-Q_{ij}\right]^2\right)$$

which precisely matches the form we derived in Proposition 3. ■

Thus by choosing $c=\frac{2}{N-2}$, our multinomial likelihood can be viewed as approximating the likelihood used by STRUCTURE run on the same data, at least in the case of a large number of unlinked loci, many individuals, and weak drift. Extension to the diploid case is trivial by substituting $N_{hap}=2N$ into $=\frac{2}{N_{hap}-2}$, giving $c=\frac{1}{N-1}$ in this case.

Discussion of Proposition 4:

The above results motivate the extension that modelling the coancestry matrix as multinomial may be appropriate even in the case where linkage disequilibrium is present, particularly when the number of loci is large, provided an appropriate value is chosen for c. In the unlinked case, the value $c=\frac{2}{N-2}$ (for haploids) can be viewed as an adjustment for the fact that the variance of the entries in the count matrix is overestimated by the multinomial likelihood. Indeed, if structure is weak, the true counts variance is approximately a factor $N-2$ lower than that given in the multinomial model. Further, the symmetry of the matrix gives an additional factor 2 for the off-diagonal terms yielding $c=\frac{2}{N-2}$. Noting that the multinomial distribution is generally approximated by a multivariate normal, we suggest that in general the multinomial model may still give a reasonable approximation, provided we use a value of c equal to twice the ratio of the correct underlying variance of the coancestry matrix to the "multinomial" estimated variance (at least if structure is "weak" – in the strong structure case results are insensitive to the value of c).

In the general case it is not possible to analytically identify c, so we instead estimate the true underlying variance of the number of chunks copied using a bootstrapping approach, calculate the required ratio using the genome-wide copying fractions, and substitute in this value of c into the likelihood. The above theory demonstrates that this approach will work in the simple unlinked loci case at least. In Section S6 we demonstrate that the bootstrap approach leads to the same value of c as the theory in the no-linkage case, and additionally works well in the linkage case using our copying model.

S5 Simulation Procedure

Genetic recombination maps were produced as described by the International HapMap Consortium (2007). Each map corresponds to the following regions of the genome (in cM): 6.946, 12.265, 3.423, 8.391, 2.888, 2.140, 8.708, 3.323, 8.531 and 11.764. Each is a cumulative distribution function describing the relative rate of recombination in the 5Mb region, along with an overall recombination rate `<rho>`.

For each of the 10 genetic maps, we generated 20 regions of length 5Mb by running the program SFS_CODE (Hernandez 2008) 20 times with the command:

```
 sfscode 5 1 -Td 0 0.3133
-TS 0.087084 0 1 -TS 0.087084 0 2 -TS 0.094777 1 3 -TS 0.102469 2
4 -TE 0.110162 -Tg 0 26.861714 -N 5000 -n 100 -A -L 999 ... -l p 1 --rho F
<recmapfile> <rho>
```

where `...` is 998 entries of `5003` followed by `5009`. This is a trick to create a region of exactly 5Mb consisting of 999 linked regions at distance 1 from each other each of approximate length 5000 bases (the remaining bases are gaps). This split is required for efficient simulation. This generates 20 individuals from each of 5 populations with a split structure described in Figure 2A of the main text, using a model with exponential growth following a bottleneck; consult the SFS_CODE manual for details. This generates a sample of 100 individuals per population; the first n from each were sampled where a smaller number was required. The output of each of the 200 runs was converted to phase format using a script written in R (R Development Core Team 2009). We then used ChromoPainter to perform painting on the output of each region independently in order to get 200 coancestry matrices. Coancestry (i.e. chunk copy count) matrices are summed, and when less than 200 regions are used they are ordered to give an even contribution from the different genetic maps.

Note that although we do not use them here, chunk length and mutation count matrices are available. These can be combined across runs as follows: chunk length matrices are averaged with weights given by the number of counts, to give the average length of a chunk. Mutation matrices are averaged with weights given by (length matrix times count matrix) to obtain mutation rates (proportional to) per site. Our software includes 'ChromoCombine', a tool to combine multiple ChromoPainter files as described above which is helpful for parallelization of large (e.g. genomic) datasets.

S6 Empirical validation of c

S6.1 Calculation of c

We first segment the genome into contiguous segments of constant *number of chunks d*. The number of chunks donated to individual i from j in segment k is x_{ijk}, and d is chosen such that x_{ijk} is approximately independent (for different k and conditional on i and j). This means that different individuals may have a different number of segments if they have different patterns of recombination. In practice, we found that $d = 100$ works well for the HGDP data, but due to the high LD present in the linked simulation data, there were only an average of 20 chunks per region. We therefore took the whole region to be a segment in this case and computed c using the full 200 region dataset. Then we compute $s_{ij} = \sum_k (x_{ijk})$ and $s^2_{ij} = \sum_k (x^2_{ijk})$. If individual i has R_i segments in total, we can calculate the theoretical variance for x_{ij} by first estimating the rate of inheriting from each other individual $\hat{P}_{ij} = s_{ij} / \sum_j s_{ij}$ and substituting into the multinomial variance:

$$V_T(x_{ij}; P_{ij}) \approx V_T(x_{ij}; \hat{P}_{ij}) = \sum_j s_{ij} \hat{P}_{ij}(1 - \hat{P}_{ij})/R_i$$

The empirical variance is:

$$V_E(x_{ij}) = \frac{s^2_{ij}}{R_i - 1} - \frac{(s_{ij})^2}{R_i(R_i - 1)}$$

This leads (with correction for the known overcounting factor of 2) to the estimate of c:

$$c_{ij} = 2\frac{V_E(x_{ij})}{V_T(x_{ij}; \hat{P}_{ij})}$$

and we simply take the mean value as our estimate:

$$c = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1 \neq i}^{N} c_{ij}$$

Note that we provide a helper program called ‘ChromoCombine’ that calculates this, and which can easily use the two options of d described. It also handles summation of multiple files in case of parallelization was used for processing individuals and/or chromosomes separately.

S6.2 Validation

In this section we present evidence of the effect of varying the rescaling factor ‘c’ on inference. Note that we view c as a summary of the data in the same way as the coancestry matrix X, and not as a parameter - it is therefore not appropriate to perform inference for it in the standard Bayesian way.

For interpretation of the empirical evaluation presented, we note that when c is 'too large', the effective number of chunks is reduced and therefore any mistakes in population assignment will tend to be under-split, i.e. we will not distinguish efficiently between similar populations. When c is 'too small' our model believes it has more independent chunks than is true and therefore will tend to over-split populations. The smallest c that does not over-split is called efficient, and larger c are called conservative.

We start with the unlinked model in the case where there is no population structure. Provided population sizes are large, we expect the theoretical results derived in Section S4 above to hold. Specifically, the theoretical prediction (for the approximately correct data likelihood) in the case of unlinked data is (Proposition 4 of Section S4.4):

$$F(x|p) = \prod_{i=1,j=1}^{N} \left(\frac{P_{q_i q_j}}{\hat{n}_{q_j}} \right)^{x_{ij}(n-1)} \tag{S24}$$

which is equal to Equation 1 of the main text:

$$F(x|p) = \prod_{i=1,j=1}^{N} \left(\frac{P_{q_i q_j}}{\hat{n}_{q_j}} \right)^{x_{ij}/c} \tag{S25}$$

when $c = 1/(n-1)$.

For this section we have generated datasets containing 15000 non-rare ($> 5\%$ allele frequency) unlinked SNPs (at varying N) under the same splitting scenario as the main text. Simulation for each SNP was by a) generating the 'ancestral frequency' f with $p(f) \approx 1/f$ (since this is not a probability distribution, we first choose which of 20 bins in the range 0 and 1 the SNP is from, then sample f conditional on this), then b) applying a normally distributed drift matrix for population level drift Σ, giving population level frequency vector $\mathbf{g} \sim \text{MVN}(\mathbf{f}, f(1-f)\Sigma)$, and c) sampling individuals SNP values according to this frequency. (SNPs with empirical frequency below the 5% threshold were resampled). The covariance matrix for the drift was:

$$\Sigma = \begin{pmatrix} 0.02 & 0 & 0 & 0 & 0 \\ 0 & 0.02 & 0.015 & 0 & 0 \\ 0 & 0.015 & 0.02 & 0 & 0 \\ 0 & 0 & 0 & 0.02 & 0.01 \\ 0 & 0 & 0 & 0.01 & 0.02 \end{pmatrix}$$

The theoretical prediction is compared to our empirical estimate of c on this dataset in Figure S1 which shows that our theoretical understanding of c is correct, i.e. that the correlation with the truth is 1 at the predicted value and that both the theoretical and empirical estimates of c are approximately efficient and equal for large N. The empirical estimate is conservative for small N.

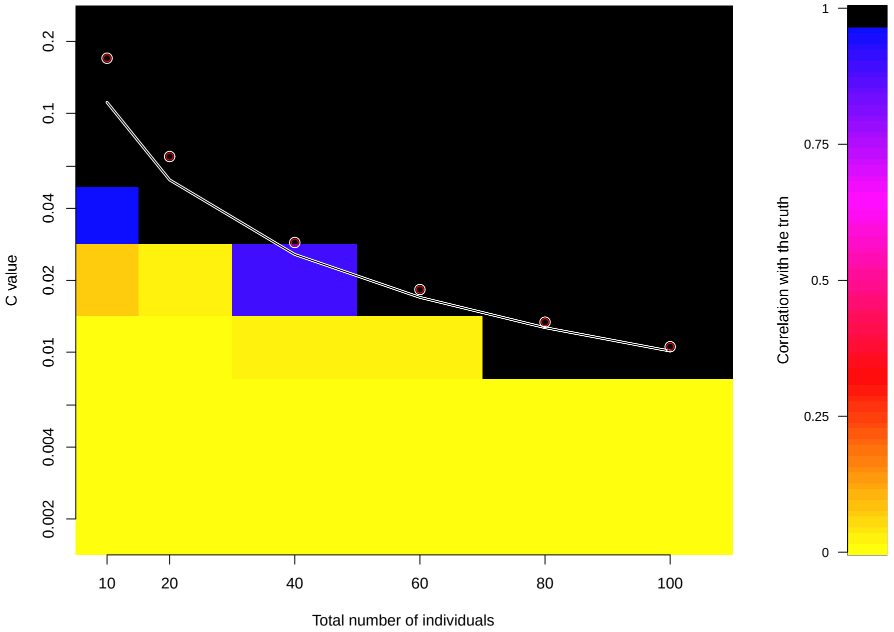

Figure S1: Correlation with the truth for 15000 non-rare ($> 5\%$ allele frequency) unlinked SNPs with a varying number of individuals and with varying chunk scaling c, when there is no true population structure. Black indicates perfect correlation, which is always achieved at the theoretical (black line) and empirical estimated (dots) values of c. (Note that at $N = 40$ the correlation is perfect at the theoretical value of c, but not at $c = 0.02$, the nearest point on the grid.)

Our empirical estimate of c is also applicable in the case of linked data, whether using our linked or unlinked models. For the linked simulated data described in the Results section of the main text, we perform a similar scan of c and N to check that our algorithm is computing an appropriate value of c in realistic circumstances. Figure S2 shows these results.

The value of c estimated by the empirical method again appears to be conservative for small N and approximately efficient for large N. In both cases, the 'truth' (if obtainable from the data) is still obtained for a very wide range of $c' > c$ i.e. greater than the estimated value. This demonstrates that exact specification of c is not an important issue for many practical purposes.

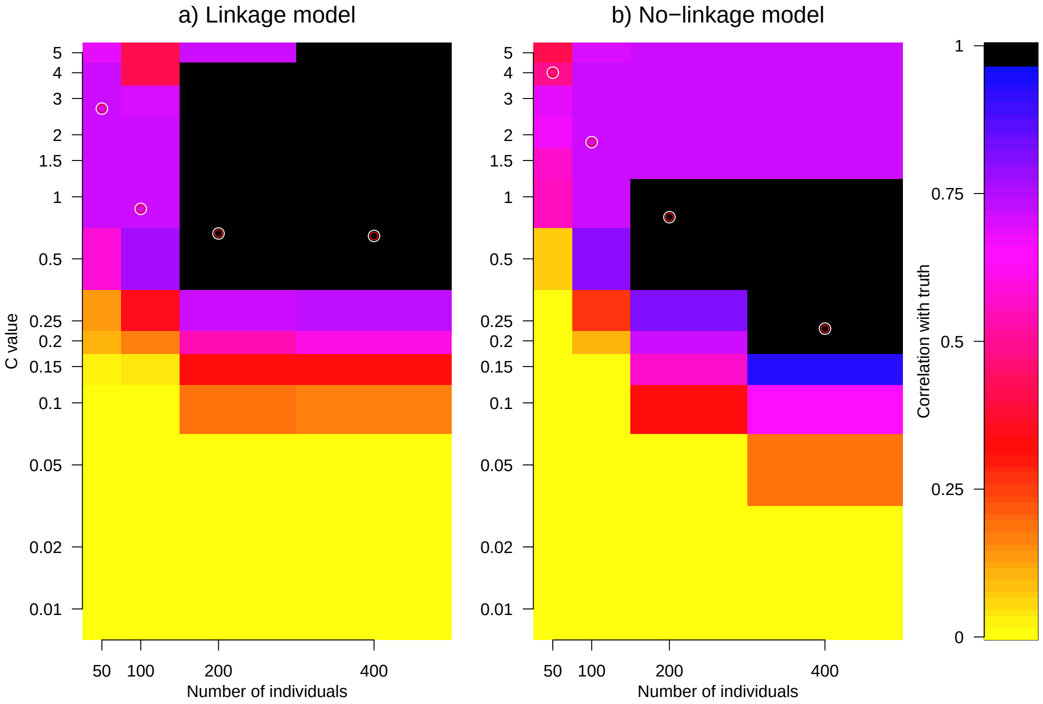

Figure S2: Correlation with the truth for linked data with a varying number of individuals and with varying chunk scaling c, with the 5 populations described in Figure 2 of the main text (and 150 regions of data). Left (a) is for the linked model, Right (b) is for the unlinked model. The empirical estimated values of c are shown as dots.

Note also that the value of c is significantly larger for linked data (in both the linkage and no-linkage models) than for the case of unlinked data. When the unlinked model is used, correlations between neighbouring loci due to linkage disequilibrium increase the variance between regions. When the linked model is used, c values do not fall substantially below 1, even for very large population sizes. Intuition behind the different behaviour of the linked and unlinked models comes from considering the uncertainty in chunk assignment. For the unlinked model, the number of haplotypes which a particular allele is identical to increases linearly with sample size. For the linked model, each addition individual in the sample has the chance of having a haplotype that is a still better match than any preceding haplotype. For this reason, the uncertainty of assignment of each haplotype does not change substantially as additional individuals are added.

We now describe a scenario in which neither the theoretical nor the empirical estimate of c work well; this is because there is not a single suitable value of c for which our model holds in this case. This is the case of unlinked markers with *strong* differentiation between populations, large numbers of markers and large sample sizes (Figure S4). Here the estimated value of c gives confident assignment of incorrect splits. The model predictions break down because individuals within the same population all share SNPs with individuals in other populations. If genetic drift is strong at individual SNPs, then sharing this coancestry can give inappropriately weighted information that the individuals are related to each other. In other words, the assumptions of Section S4, Propositions 3-4 do not hold.

It is important to note that this problem is less dramatic in linked data, and essentially does not arise in the linked model. To see why, we note that these correlations arise when an (unlinked) SNP is found in a individual that is common in *another* population but rare/absent otherwise. Such SNPs can only arise through strong drift, are not excluded because they are not rare overall, and are interpreted as overly strong evidence of shared ancestry. As N becomes large with population sizes $n_a \propto N$, most SNPs provide $O(N^{-1})$ information on population level copying proportions (which is why $c = O(N^{-1})$). However, strongly drifted SNPs provide $O(1)$ evidence because they are shared with a high fraction of another population, and not with any other individual. For (truly) linked data, such a SNP will be down-weighted due to the average level of correlation between nearby SNPs, so even in the unlinked model we will infer a larger value of c. For the linked model, all chunks are approximately unique and therefore provide $O(1)$ information per chunk, so a strongly drifted SNP will not have a dramatic influence since all *chunks* are already 'strongly drifted'. The success of our algorithm for simulated linked data also supports this argument.

For the strong drift and unlinked case, we have developed an alternative algorithm in which the likelihood is modified so that these correlations are accounted for. We do this in effect by considering only within-population counts as important, so that when considering a merge move, the between-population counts are normalised to have the same mean. We re-normalise using:

$$x'_{ij} = x_{ij} - \frac{\sum_{k \in q_i} x_{kj}}{|q_i|} - \frac{\sum_{l \in q_j} x_{il}}{|q_j|} + 2\frac{\sum_{k \in q_i} \sum_{l \in q_j} x_{kj}}{|q_i||q_j|} \tag{S26}$$

where q_i is the index of the population for individual i, and $|q_i|$ is the number of individuals in that population. This corresponds to ensuring that all row and column sums copying from population b to population a are equal. This is illustrated in Figure S3 which shows the coancestry heatmap for the unnormalized and normalised cases, as well as the difference between them. The coancestry heatmaps are visually very similar, but the undesired correlation structure is clearly visible in this difference. Some individuals have an elevated number of donated chunks to all individuals *within a specific population*, leading to a 'striped' pattern. The bottom plots show the same thing but where we consider a potential merged population (merging the most recent split). It is clear that the presence of population structure is preserved under this procedure because the two populations have a different profile *within the population being merged*. Therefore the standard likelihood applied to both the merged and split matrices can correctly identify both populations due to their different rates of copying within and between, and additionally it is not mislead by the correlated copying from other populations. However, were the only distinction between populations B1 and B2 the copy rate from some third population, this would be normalised out.

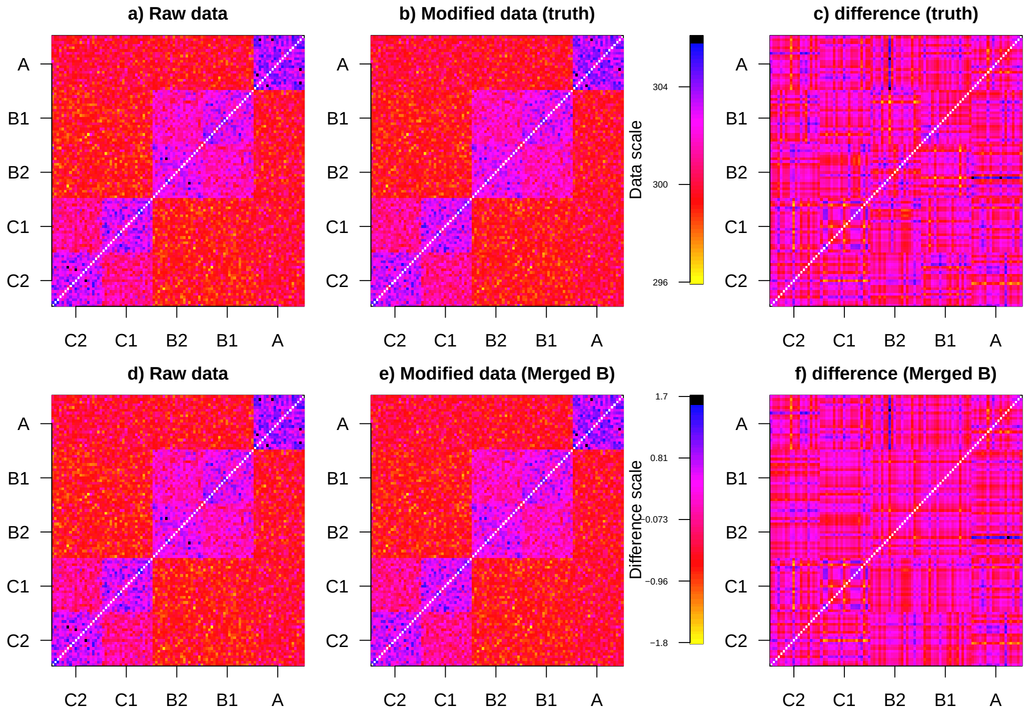

Figure S3: Left: the raw coancestry matrix for the same scenario as simulated in the main text but with 15000 unlinked SNPs. Centre: the renormalized coancestry matrix based on the true population distribution. Right: The difference, highlighting the correlated nature of the error terms for the coancestry matrix (there are differences for the merged B1 and B2 populations only). Top: These matrices based on the 'true' population structure given by the labels. Bottom: These matrices based on merging the most recent split, setting $B = (B1, B2)$.

Note that our simple likelihood is not a well defined entity under this modification, because the data depends on the population assignment. There is however an implicit likelihood induced by the modification of the data which is well defined and is correctly comparable both within states of a given number of populations K and for states of differing K, provided that we consider moving an individual

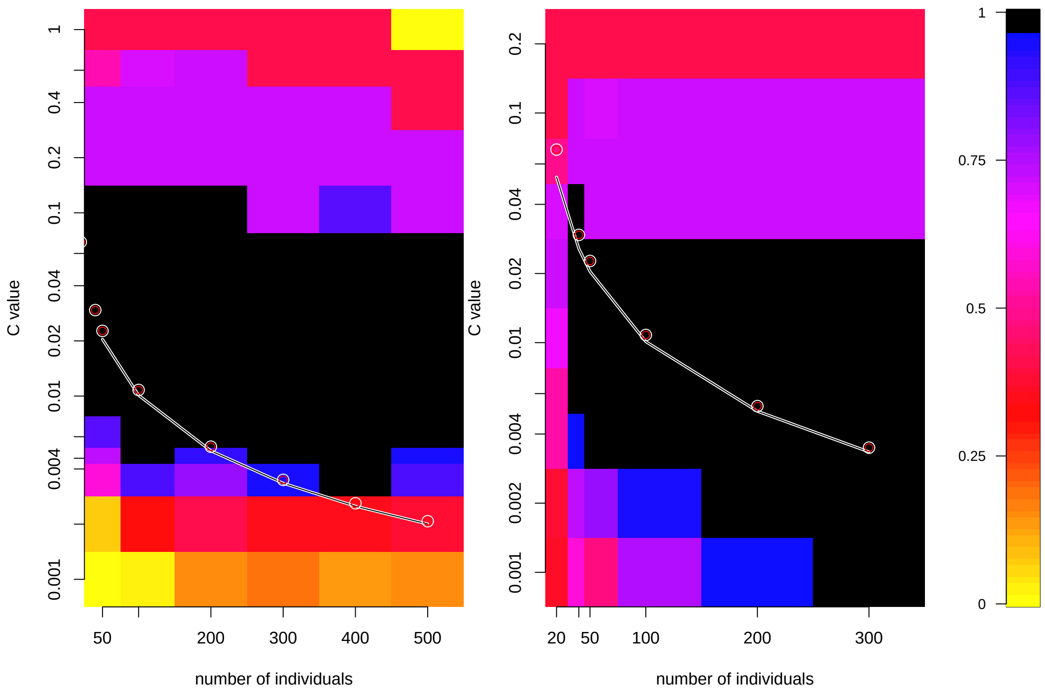

Figure S4: Demonstration of how our model breaks down in the presence of strong population structure and *unlinked* data, and our method for fixing this. This figure shows the correlation with the truth for 15000 non-rare ($>5\%$ allele frequency) unlinked SNPs under the simulation demographic model described in the main text. Left: results for the raw data. Right: results for the modified data matrix x' as described above.

between two populations as creating a merged state between the two populations (which defines the normalisation), and creating a split state corresponding to the move.

Although this procedure is more robust than the use of the raw coancestry matrix, it is not recommended for general use because firstly, it discards information about a split that comes from differential chunk counts from other populations, and secondly, it is not a clearly defined model. We have tested this procedure on the HGDP data (unlinked model, results not shown) and obtain broadly similar results to those quoted in the main text with some subtle splits lost: for example, the Tuscan/Italian split is not fully supported. We recommend using it as a conservative check if the value of c is very low (say less than 0.05) and there is strong structure in the dataset.

S7 Comparison to STRUCTURE

We have shown that in theory, and in the unlinked model case, STRUCTURE and fineSTRUCTURE are using approximately the same data and the same model, under certain limiting conditions. It is important to assess how these conditions apply in practice. Figure S5 shows the correlation with the truth, as the number of SNPs changes, for both fineSTRUCTURE and STRUCTURE for N=100 individuals sampled from the same population structure as described in the main text for the unlinked case. From this figure two things are evident. Firstly, at low SNP numbers, STRUCTURE outperforms fineSTRUCTURE by a small margin. However, as the number of SNPs increases, STRUCTURE does not keep improving its performance due to two effects. Firstly, it becomes very difficult to mix the SNP frequencies with the other parameters, and so the MCMC sampling becomes poor. We can see this by starting STRUCTURE both at the truth and from random starting locations; for large numbers of SNPs it fails to find even an adequate K=3 solution (we here show the best solution found in several runs). Secondly, the prior (F-model) STRUCTURE uses assumes independent drift for all populations, and scales with the number of SNPs. Therefore the correlated drift observed in this population scenario looks equally unlikely in the model regardless of the number of SNPs, and even when started at the truth STRUCTURE favours lower values of K. Although fineSTRUCTURE also does not have explicit correlated drift in the prior, the prior does not scale with the number of SNPs and therefore the data can overwhelm any prior structure placed on the coancestry matrix. This leads to slightly conservative splitting at all scales, as we must have positive evidence of a split, hence the very abrupt change from a K=3 to a K=4 solution (and similarly for K=5).

From the theory, we would expect that as the number of individuals increases, fineSTRUCTURE tends towards the STRUCTURE performance at lower SNP counts. The message from this comparison is that the loss of information in performing the summary step is not high for datasets with hundreds of markers, but that if few, genuinely unlinked markers are used, the STRUCTURE model is preferable. For larger numbers of markers, fineSTRUCTURE is to be preferred even if the markers are unlinked.

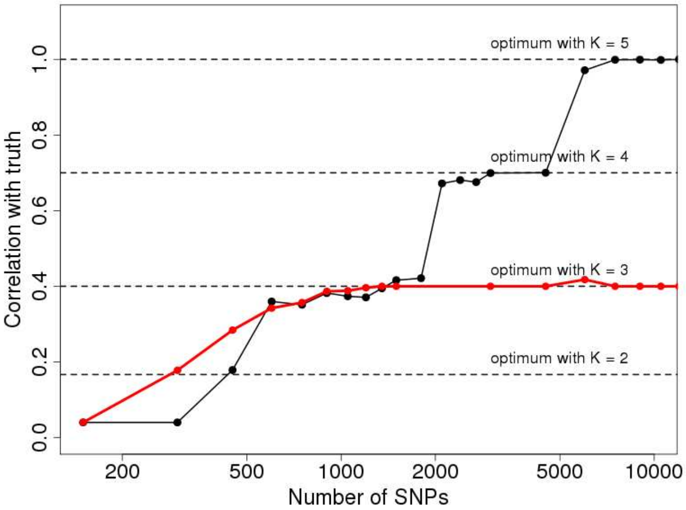

Figure S5: Correlation with truth for (black) fineSTRUCTURE and (red) STRUCTURE as a function of the number of unlinked SNPs. Data are simulated as above, with all SNPs having minor frequency > 0.05. The fineSTRUCTURE results are based on the unlinked model as described above, and the STRUCTURE results are based on the no-admixture model using the 'F model' prior started at the best possible configuration for a particular K. Optimal correlations are obtained at this configuration when there is no uncertainty in the assignment. Note that the scale is logarithmic to emphasise the behaviour with few SNPs.

S8 ADMIXTURE linked simulations analysis

For the linked simulations we have compared our results with the program ADMIXTURE (Alexander, Novembre, and Lange 2009). ADMIXTURE computes the same likelihood as STRUCTURE but performs maximum-likelihood analysis, i.e. it does not perform MCMC sampling and does not apply a prior. This makes is significantly faster and avoids many mixing problems, and is easily applicable to the HGDP dataset. To perform this analysis, we took the same phased haplotype data used as input for ChromoPainter and converted it to PLINK format (Purcell, Neale, Todd-Brown, Thomas, Ferreira, Bender, Maller, Sklar, de Bakker, Daly, and Sham 2007) (PLINK version 1.07, downloaded from `http://pngu.mgh.harvard.edu/~purcell/plink/`) using which we extracted the SNPs with minor frequency > 0.01. The filesizes were too large for manipulation within PLINK with 200 regions and therefore we used a minor frequency cutoff of 0.02 in this case. We then ran ADMIXTURE for various numbers of populations K and for varying number of regions.

We here show the details of the ADMIXTURE results since direct comparison between the methods is not possible, fineSTRUCTURE being an MCMC based no-admixture model and ADMIXTURE reporting only maximum-likelihood admixture results. The correlation reported in the paper is created by forcing the admixture solution to choose the most likely population for each individual; however, performing the correlation on the admixed solution does not change the results qualitively.

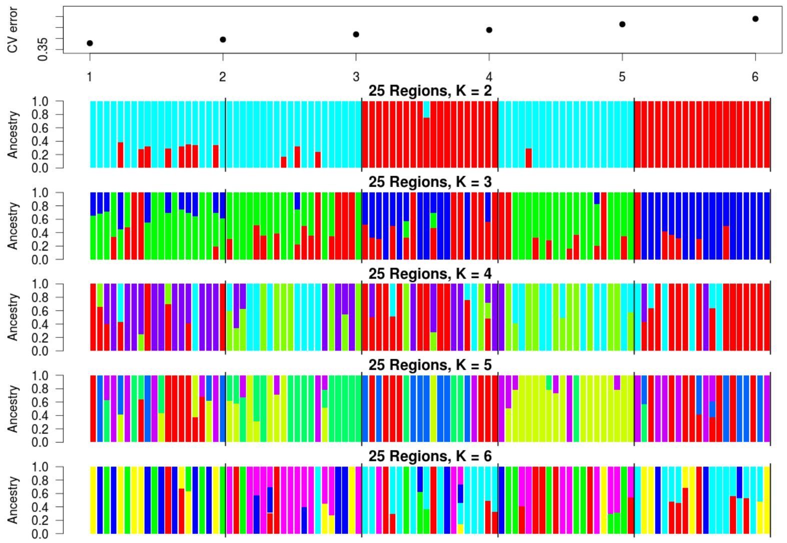

Figure S6: ADMIXTURE results for simulated data at 25 linked regions. Top: cross-validation error (lower is better). True populations are separated by a black line. The maximum correlation with truth is obtained at K=3.

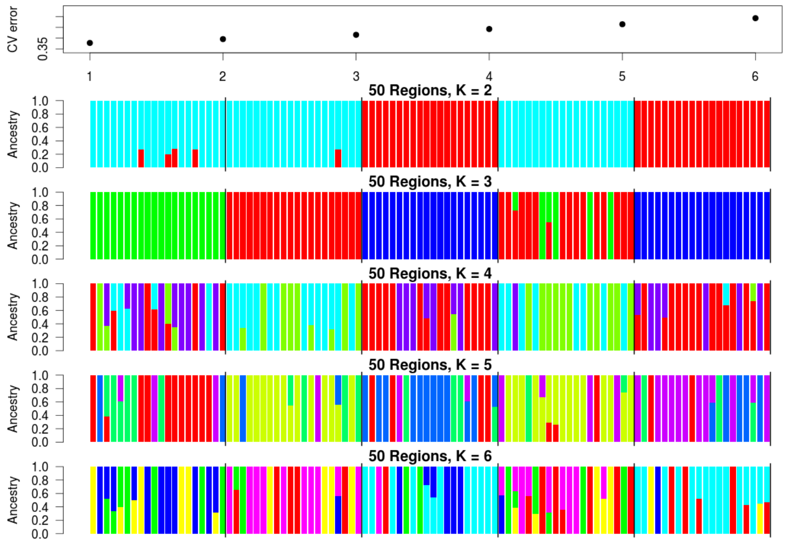

Figure S7: ADMIXTURE results for simulated data at 50 linked regions. Top: cross-validation error (lower is better). True populations are separated by a black line. The maximum correlation with truth is obtained at K=3.

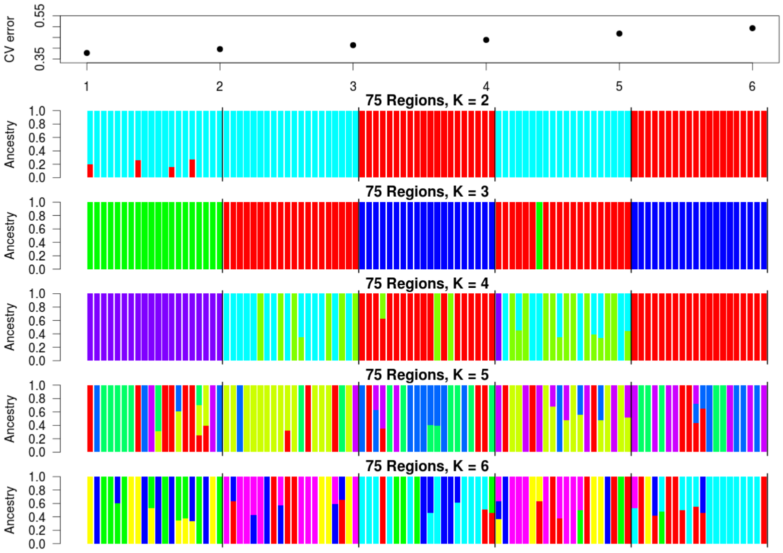

Figure S8: ADMIXTURE results for simulated data at 75 linked regions. Top: cross-validation error (lower is better). True populations are separated by a black line. The maximum correlation with truth is obtained at K=3.

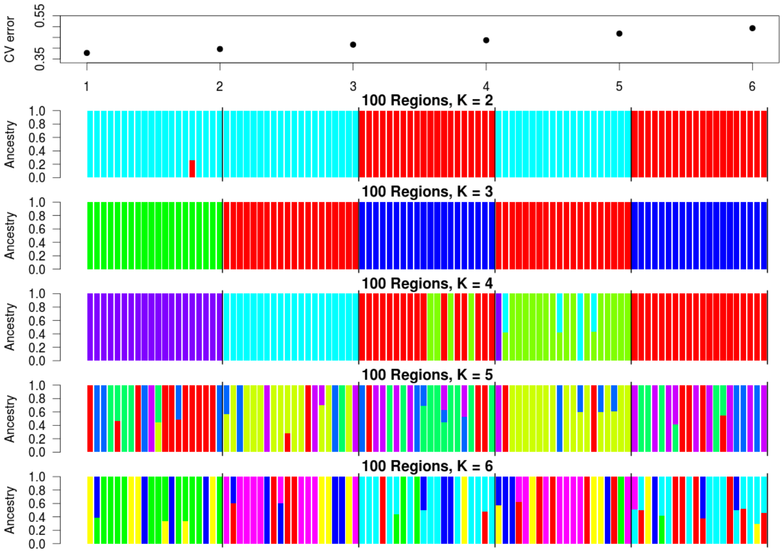

Figure S9: ADMIXTURE results for simulated data at 100 regions. Top: cross-validation error (lower is better). True populations are separated by a black line. The maximum correlation with truth is obtained at K=4.

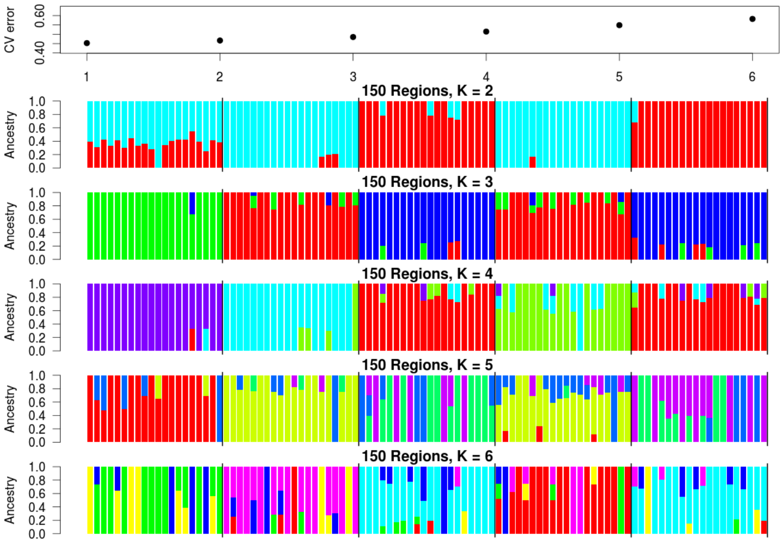

Figure S10: ADMIXTURE results for simulated data at 150 regions. Top: cross-validation error (lower is better). True populations are separated by a black line. The maximum correlation with truth is obtained at K=4.

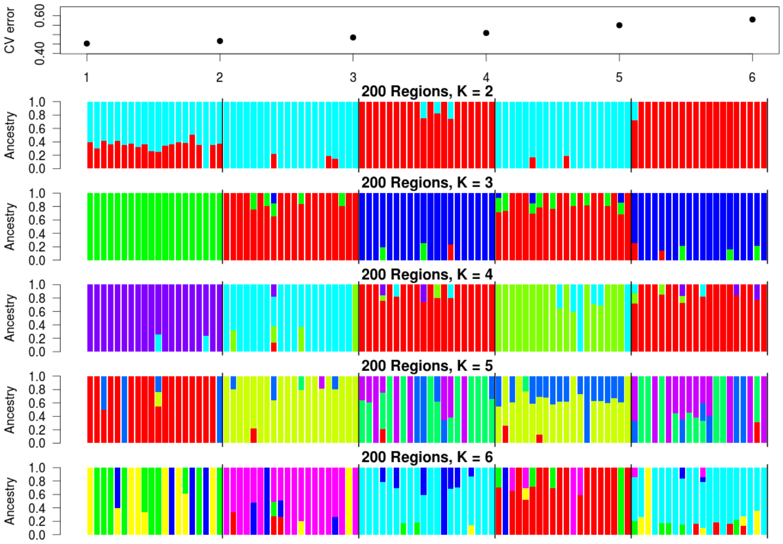

Figure S11: ADMIXTURE results for simulated data at 200 regions. Top: cross-validation error (lower is better). True populations are separated by a black line. The maximum correlation with truth is obtained at K=4.

S9 ADMIXTURE HGDP Europe analysis

For the European HGDP dataset we have compared our results with the program ADMIXTURE (Alexander, Novembre, and Lange 2009). ADMIXTURE computes the same likelihood as STRUCTURE but performs maximum-likelihood analysis, i.e. it does not perform MCMC sampling and does not apply a prior. This makes it significantly faster and avoids many mixing problems, and is easily applicable to the HGDP dataset. To perform this analysis, we took the same phased haplotype data used as input for ChromoPainter and converted it to PLINK format (Purcell, Neale, Todd-Brown, Thomas, Ferreira, Bender, Maller, Sklar, de Bakker, Daly, and Sham 2007) (PLINK version 1.07, downloaded from `http://pngu.mgh.harvard.edu/~purcell/plink/`) using which we extracted the SNPs variable in Europe with minor frequency > 0.01 and merged the chromosome 1–22 data leaving 585420 SNPs. We then ran ADMIXTURE for various numbers of populations K.

The results of this analysis are shown in Figure S12. The ADMIXTURE analysis requires that the user specify the value of K. In the standard (Bayesian) STRUCTURE approach, this can be estimated by computing the marginal probability of the data given the model with a particular K. The Bayesian approach with this model is however not feasible in the HGDP dataset. There is no robust way to compute the marginal probability in the maximum-likelihood setting and therefore ADMIXTURE instead tries to minimise the 'cross-validation error', that is, the error in predicting the value of a SNP under a cross validation scheme. In general this should be high both when the model is too simple or when it is overfitted. However, the HGDP dataset has a large number of uninformative SNPs and as Figure S13 shows the cross-validation error is minimised at $K = 1$. This occurs in part because the between-population variance is small compared to the within-population variance and hence adding population structure doesn't aid prediction. We are interested in explanatory power rather than prediction, and therefore have decided in the paper to show $K = 7$ which seems to capture many of the features in our fineSTRUCTURE analysis without obvious overfitting. We expect that the cross-validation error varies across SNPs and that higher K could be obtained by appropriate restriction to only informative SNPs.

As a check, we also used PLINK to perform linkage-based trimming of SNPs as suggested in the ADMIXTURE manual; this left a dataset of 121613 SNPs and provided broadly the same results (not shown). This shows that linkage disequilibrium has not dramatically distorted the analysis. We have discussed the results in the main text; we note here that the results as K is increased demonstrate an interesting pattern of successive population identification that correlate with our identification of populations that have drifted since admixture.

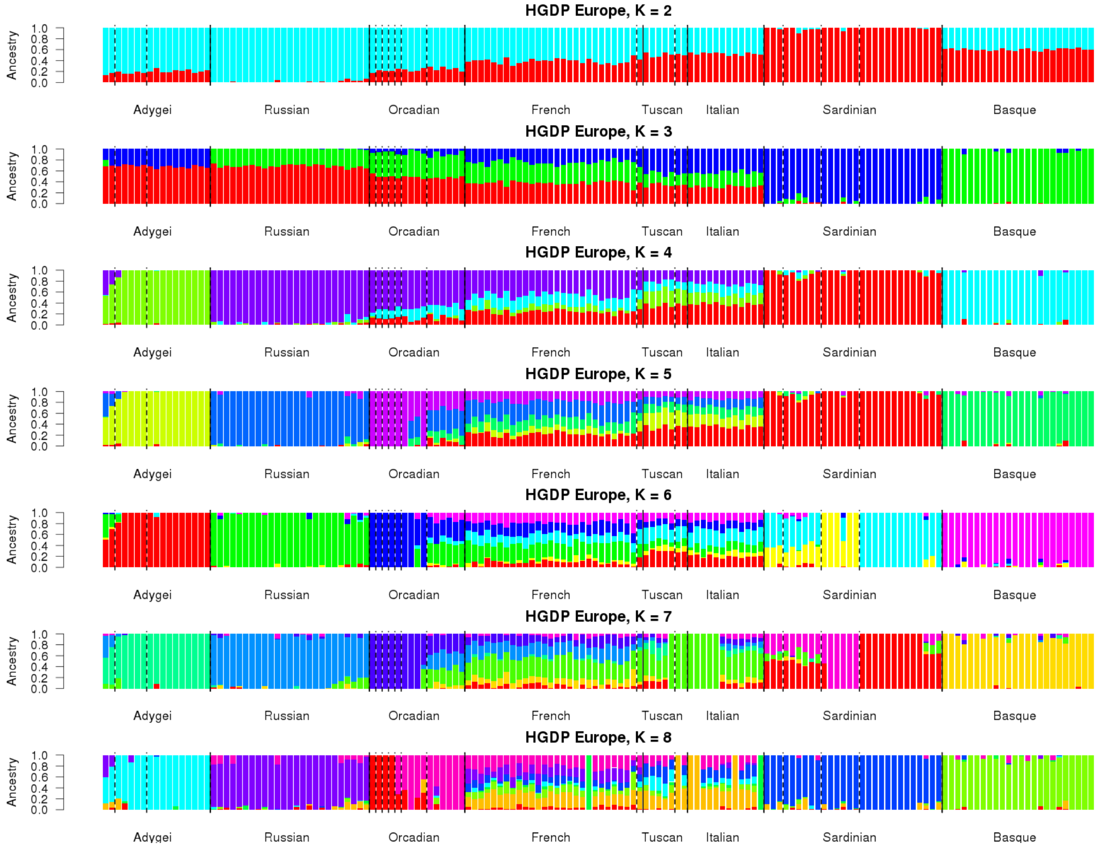

Figure S12: ADMIXTURE results for the HGDP Europe dataset for a range of K as described in the text. Dashed lines separate fineSTRUCTURE populations, solid lines separate labelled populations. fineSTRUCTURE agrees with all labelled populations with the exception of the Tuscan/French.

S10 Results for HGDP data

The full coancestry matrix for the world is given in Figure S14. Note that the colour scale is non-linear and that small changes produce a large colour change at the lower end of the scale; this does however correspond roughly to meaningful changes since these colours are present at higher frequency in the matrix, and therefore small variation at this scale may be picked up by the model as it can involve many individuals.

We now focus on the Continental analysis. Continents are defined according to Table 1. Subcontinents are defined in Figure 4 of the main text. Note that the same groupings appear in Figure S15, but at a different height due to a different number of sub-populations found in each and therefore cannot be 'cut' at a single

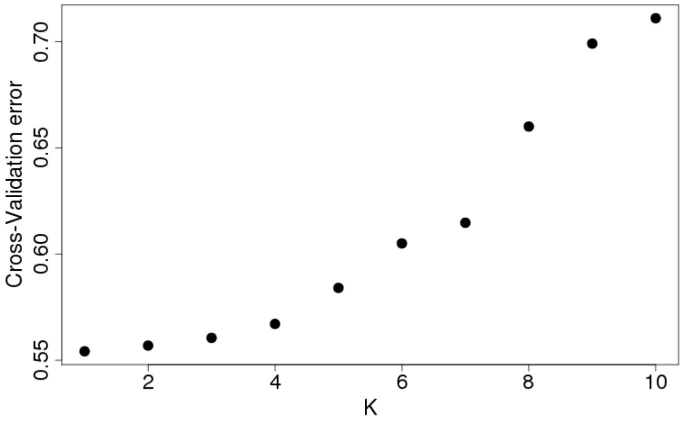

Figure S13: ADMIXTURE cross validation error as a function of K.

Continent	Populations
Africa	San, BiakaPygmy, BantuSouthAfrica, BantuKenya, MbutiPygmy, Yoruba, Mandenka
America	Colombian, Pima, Surui, Maya, Karitiana
Central South Asia	Makrani, Uygur, Brahui, Burusho, Sindhi, Balochi, Hazara, Pathan, Kalash
East Asia	Cambodian, Mongola, Oroqen, Xibo, Yi, Tu, Naxi, Daur, Hezhen, Han, Tujia, She, Japanese, Yakut, Dai, Lahu, Han.NChina, Miao
Europe	Adygei, French, Tuscan, Italian, Sardinian, Russian, Orcadian, Basque
Middle East	Mozabite, Bedouin, Palestinian, Druze
Oceania	Melanesian, Papuan

Table 1: List of populations assigned to 'continents' for PCA.

height.

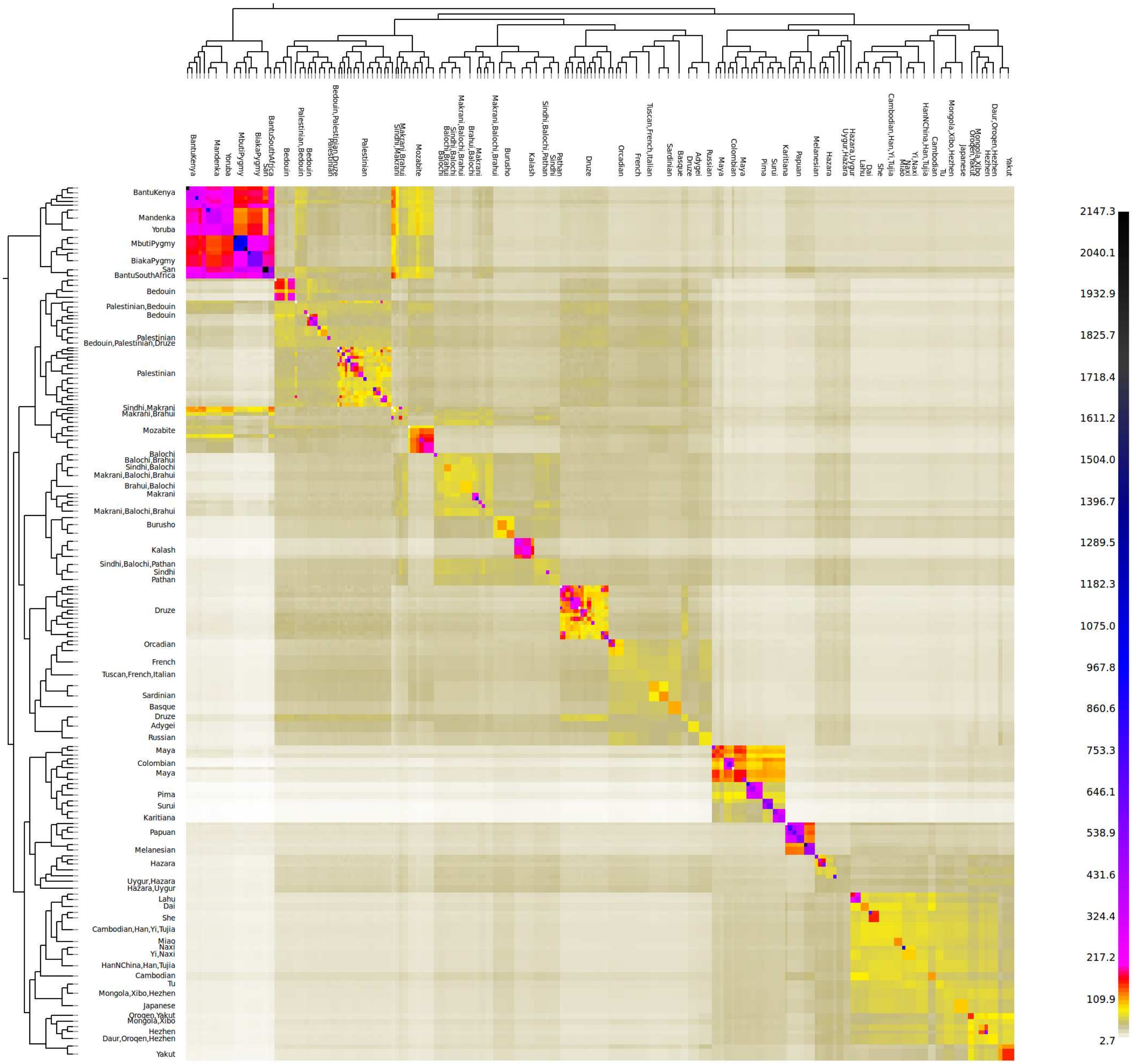

Figure S14: Whole world HGDP coancestry matrix. Some population labels are omitted for clarity; this has only been done when the neighbouring population contains the same labels and the exact distribution is recoverable from the tree and Figure 4 of the main text. The colour scale is non-linear, and population sizes have been square-rooted for clarity.

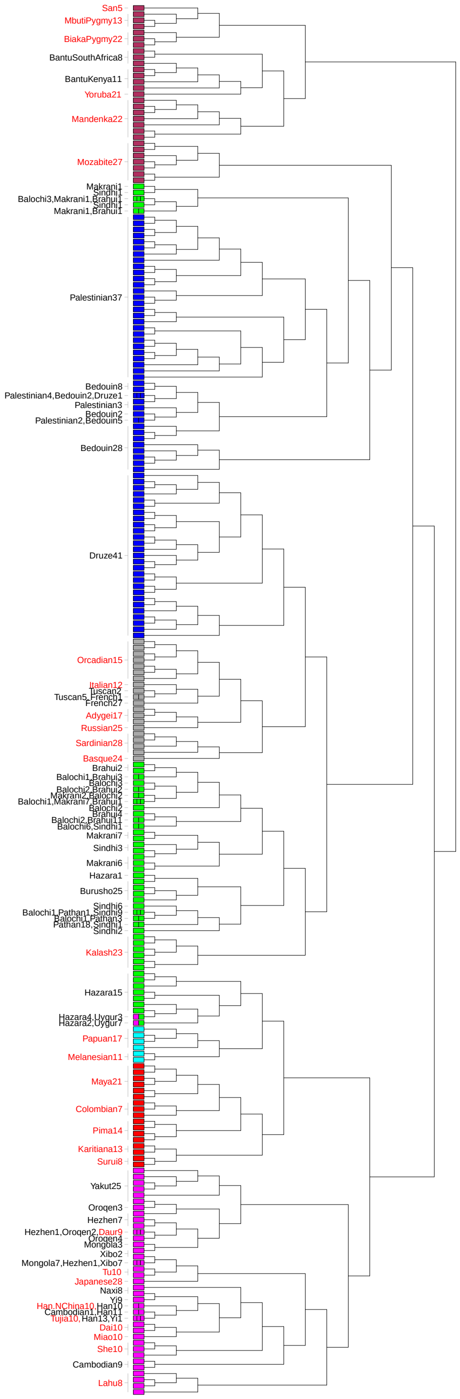

Figure S15: Tree for all populations found using inference in separate ‘sub-continents’ as detailed in Figures S16 - S24. The interpretation is the same as Figure 4 of the main text (except that probabilities have been removed for clarity).

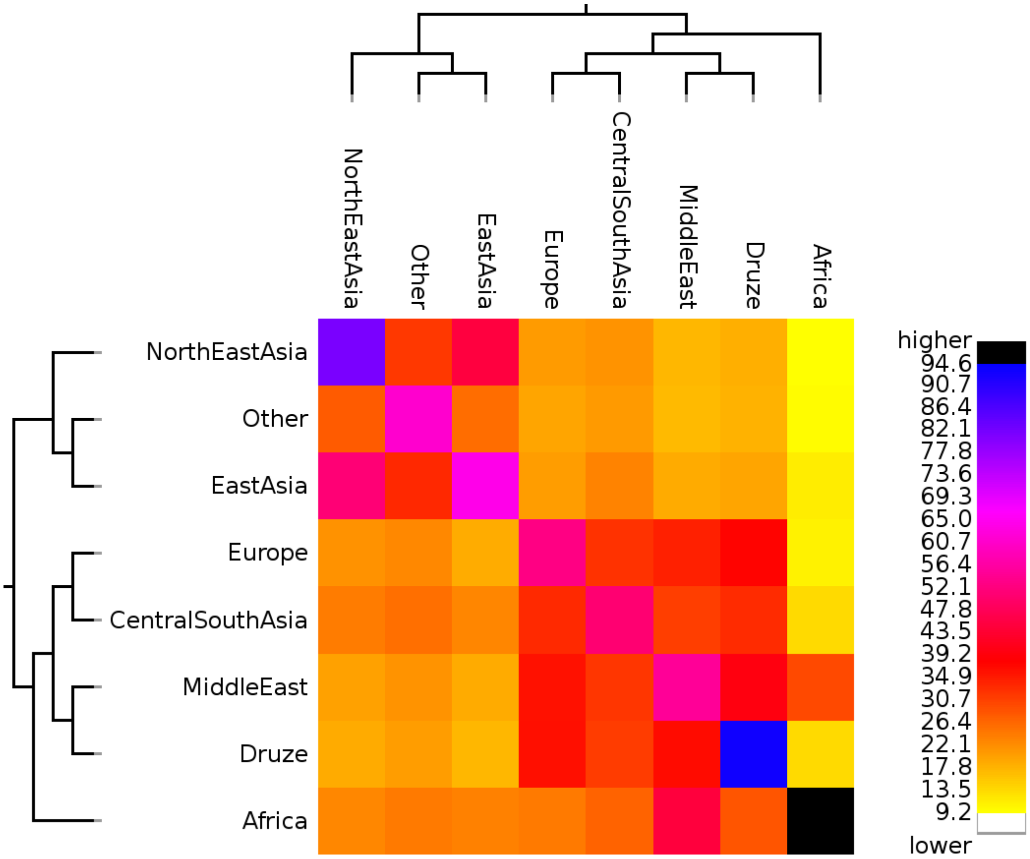

Figure S16: 'Sub-continental' coancestry matrix, for groupings as defined in Figure 4 of the main text. Recipient groups are on the left. Note that Africa has been capped, and copies 232 chunks to itself.

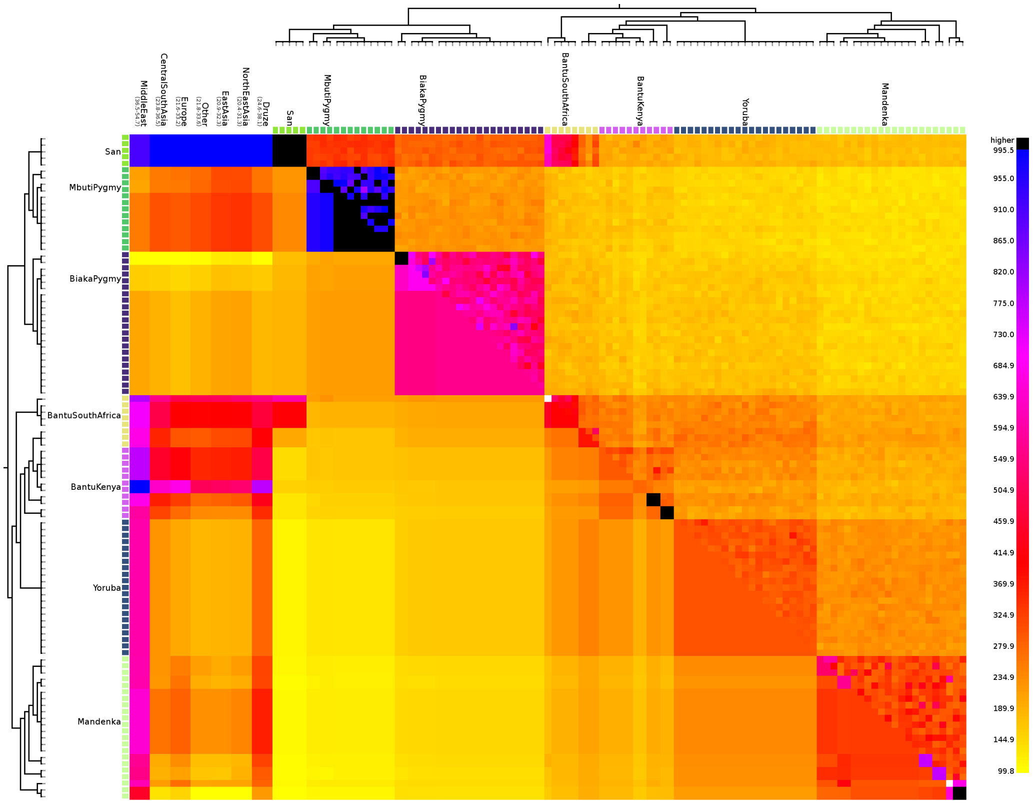

Figure S17: 'Sub-continent' of Africa coancestry matrix showing (bottom left) the Population coancestry matrix and (top right) the Individual coancestry matrix.

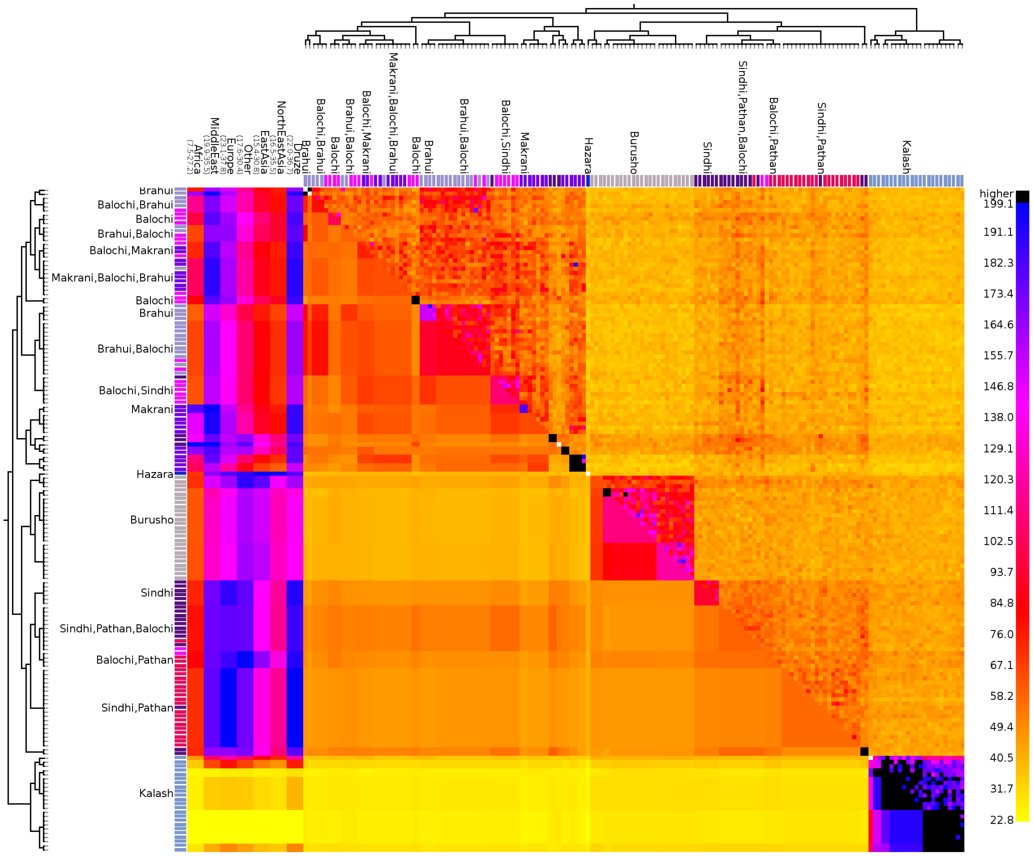

Figure S18: 'Sub-continent' of CentralSouthAsia coancestry matrix showing (bottom left) the Population coancestry matrix and (top right) the Individual coancestry matrix.

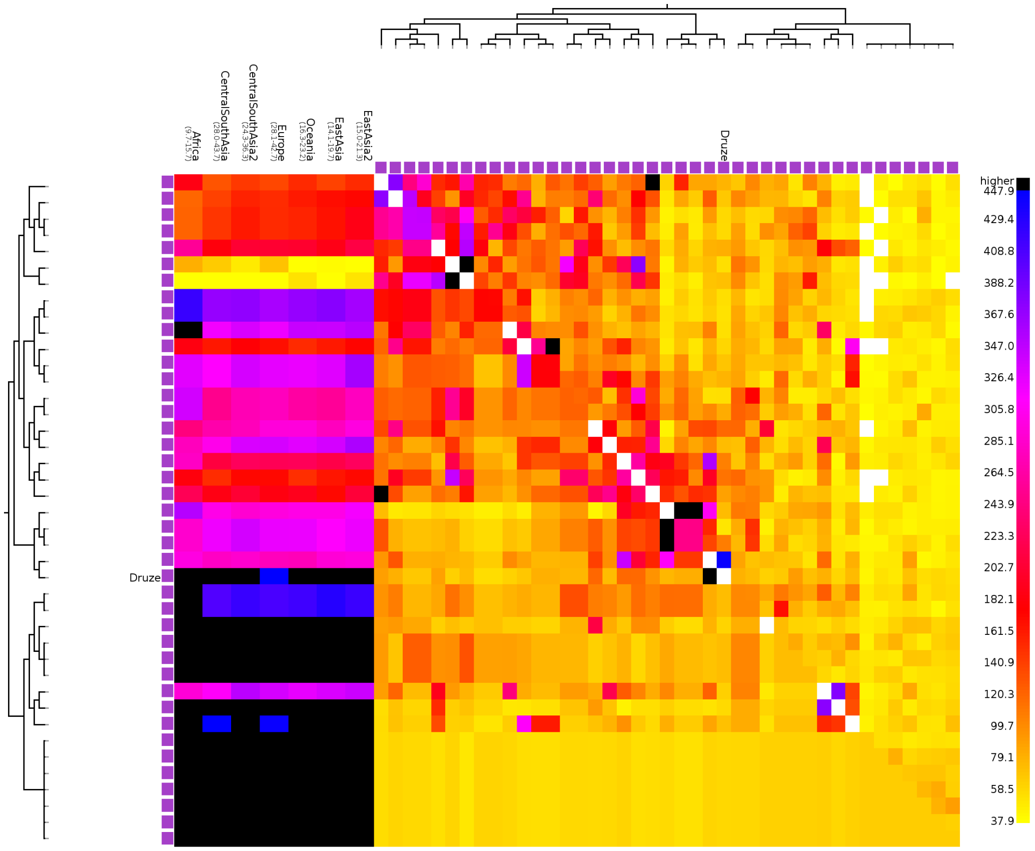

Figure S19: 'Sub-continent' of Druze coancestry matrix showing (bottom left) the Population coancestry matrix and (top right) the Individual coancestry matrix.

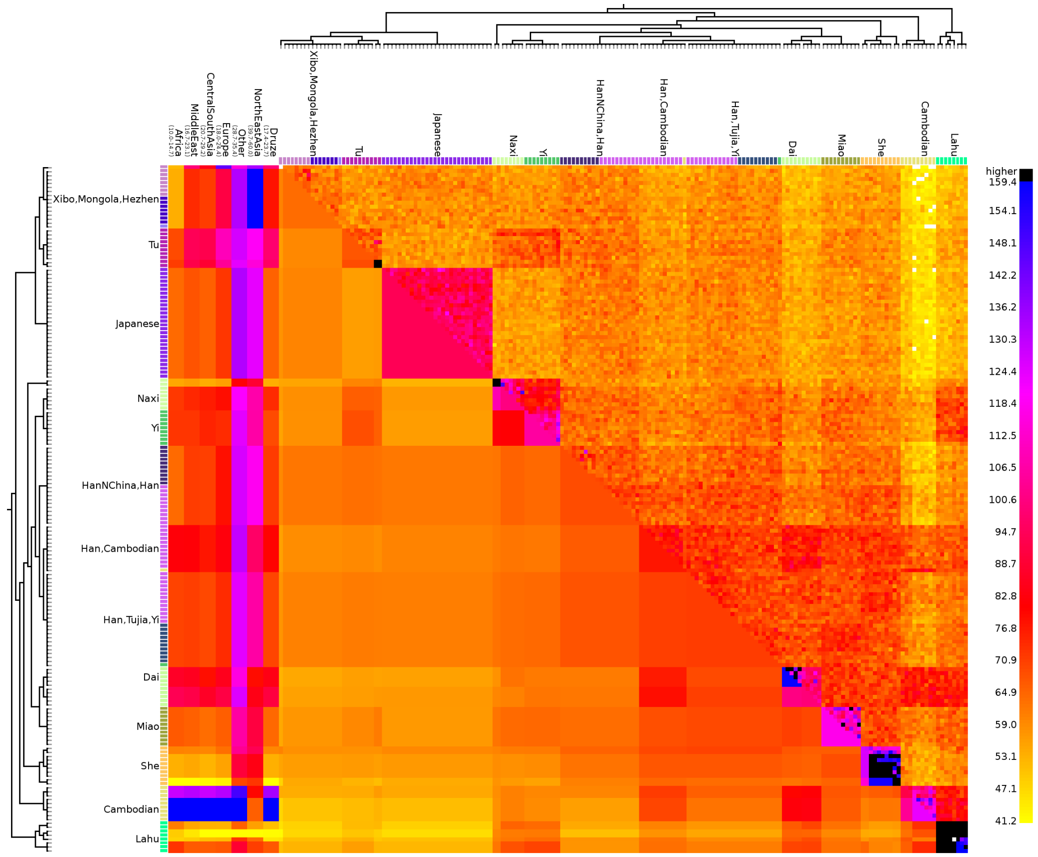

Figure S20: 'Sub-continent' of EastAsia coancestry matrix showing (bottom left) the Population coancestry matrix and (top right) the Individual coancestry matrix.

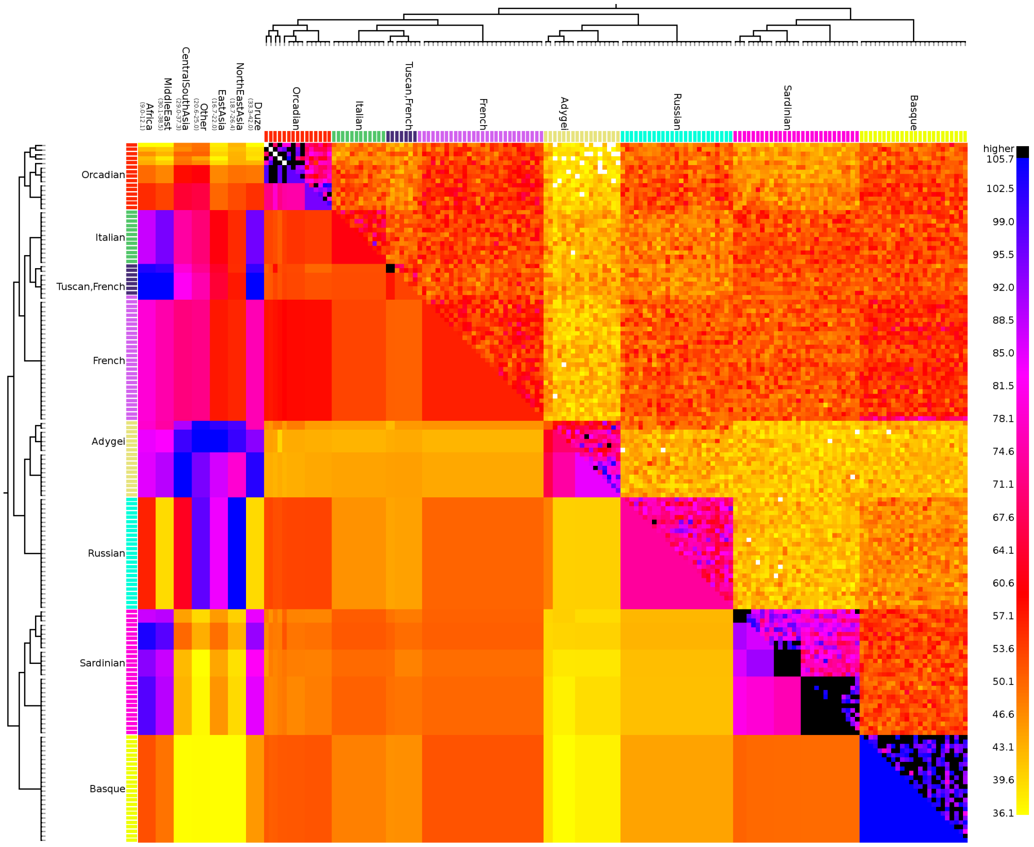

Figure S21: 'Sub-continent' of Europe coancestry matrix showing (bottom left) the Population coancestry matrix and (top right) the Individual coancestry matrix.

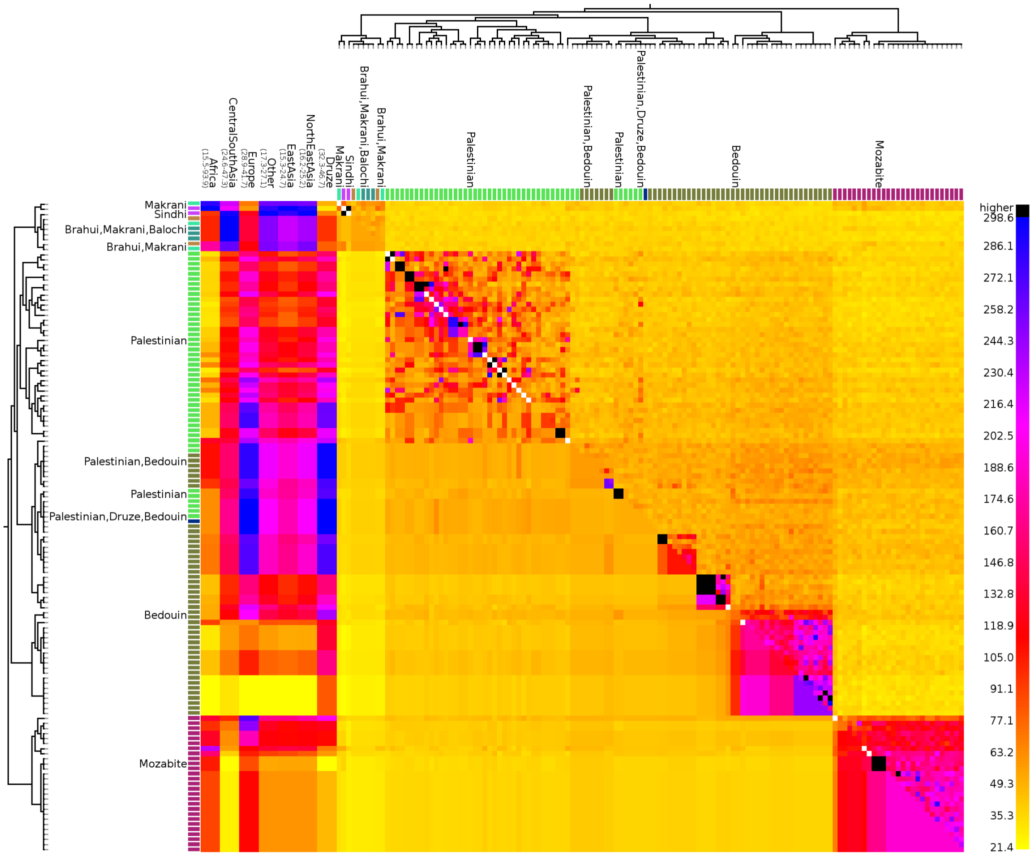

Figure S22: 'Sub-continent' of MiddleEast coancestry matrix showing (bottom left) the Population coancestry matrix and (top right) the Individual coancestry matrix.

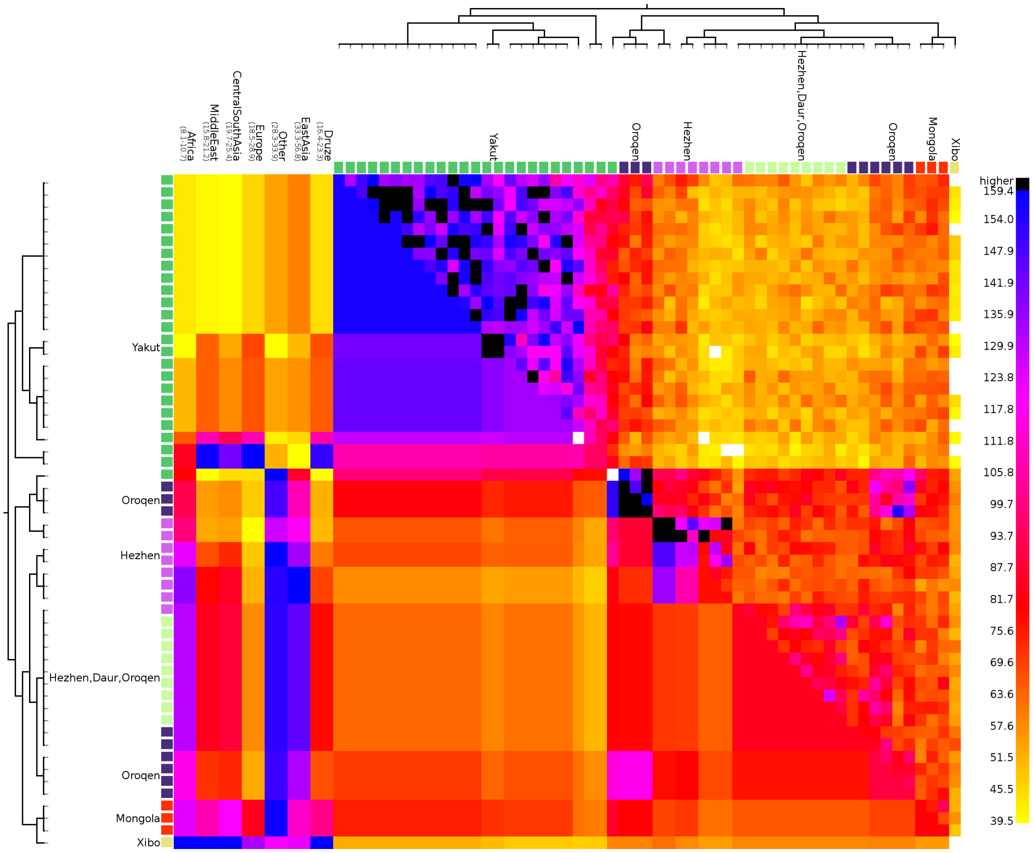

Figure S23: 'Sub-continent' of NorthEastAsia coancestry matrix showing (bottom left) the Population coancestry matrix and (top right) the Individual coancestry matrix.

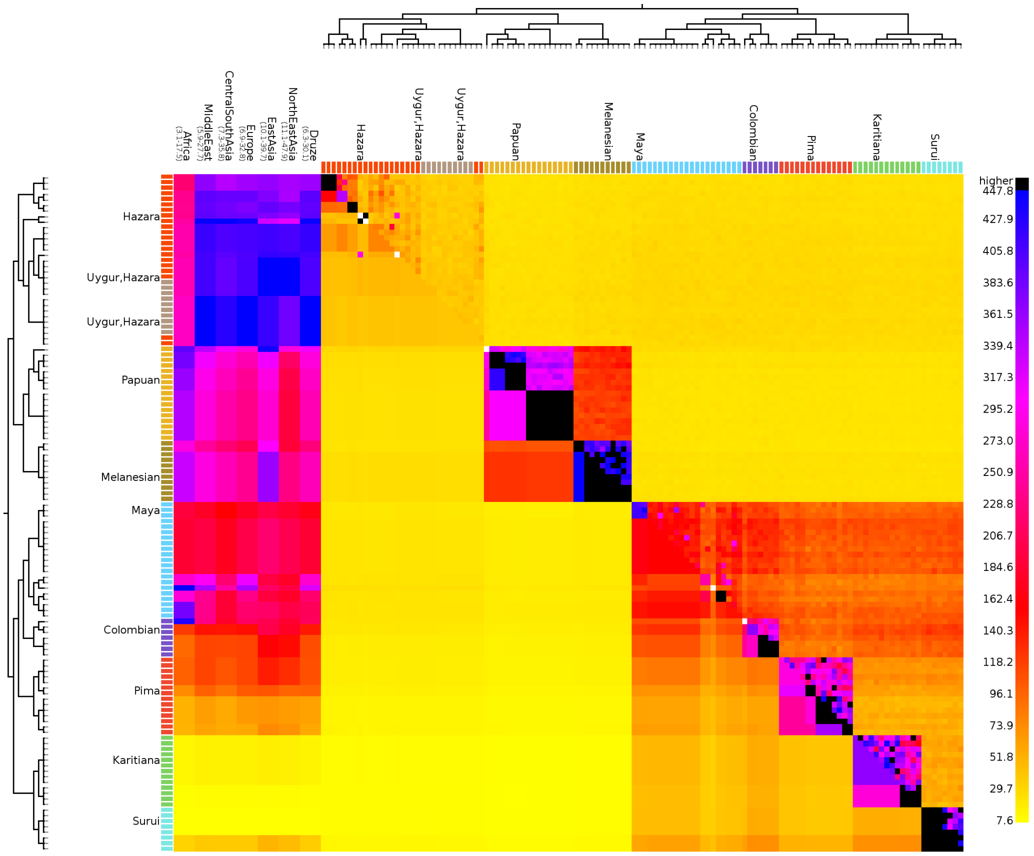

Figure S24: ‘Sub-continent’ of Other populations (America, Oceania and some Asian individuals) coancestry matrix showing (bottom left) the Population coancestry matrix and (top right) the Individual coancestry matrix.

S11 Convergence results

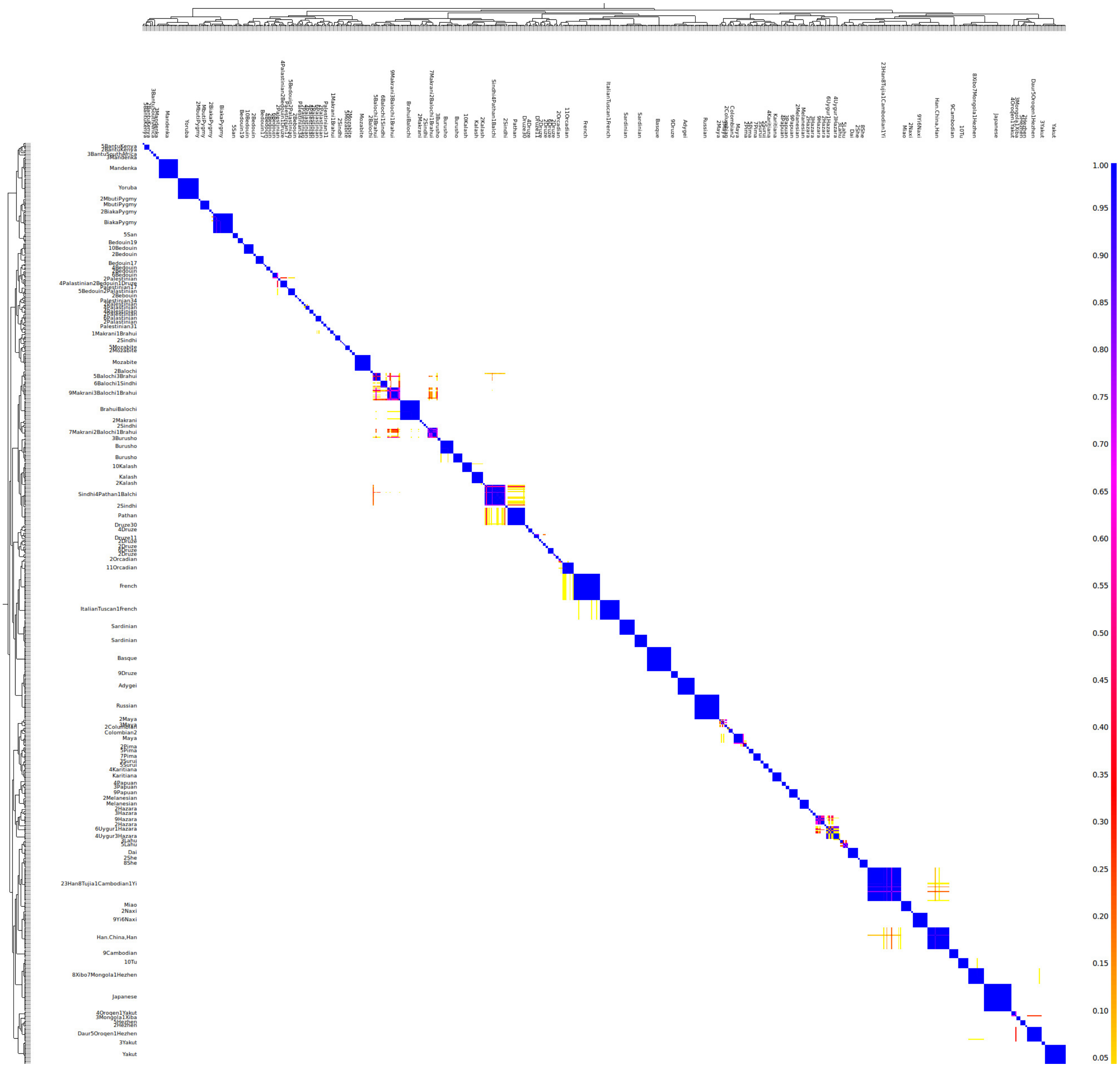

Figure S25: Whole HGDP pairwise coincidence matrix showing (bottom left) the run 1 and (top right) run 2. It is recommended to view this figure online and use zoom tools.

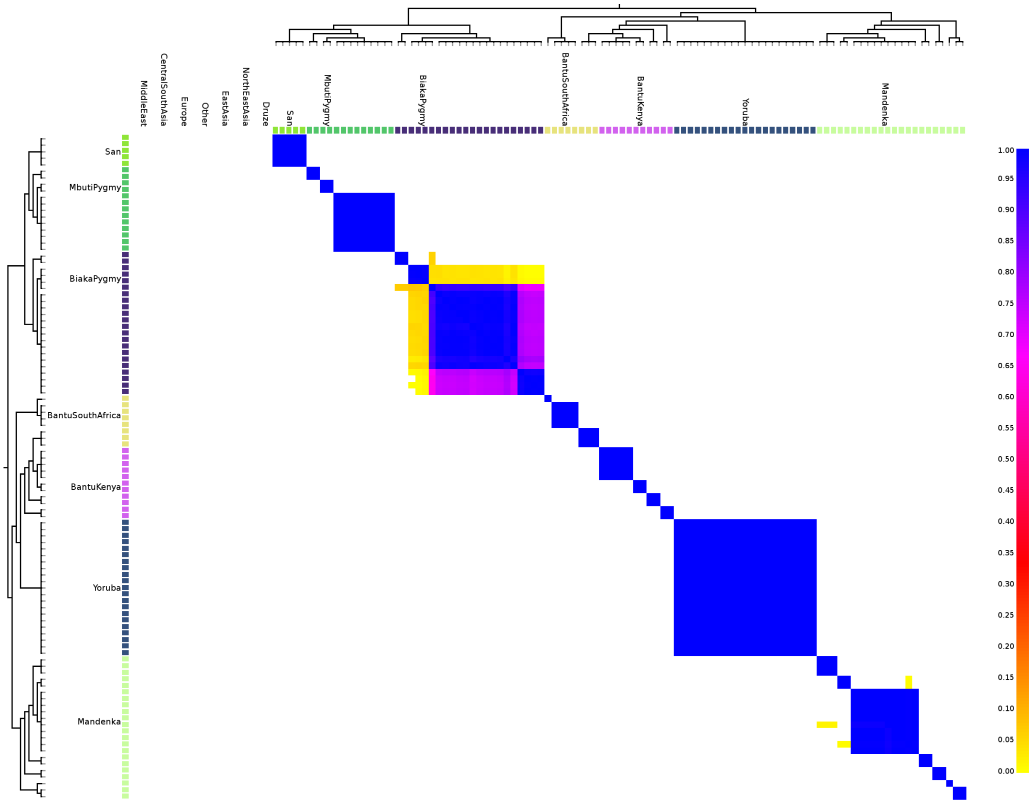

Figure S26: Africa pairwise coincidence matrix showing (bottom left) the run 1 and (top right) run 2.

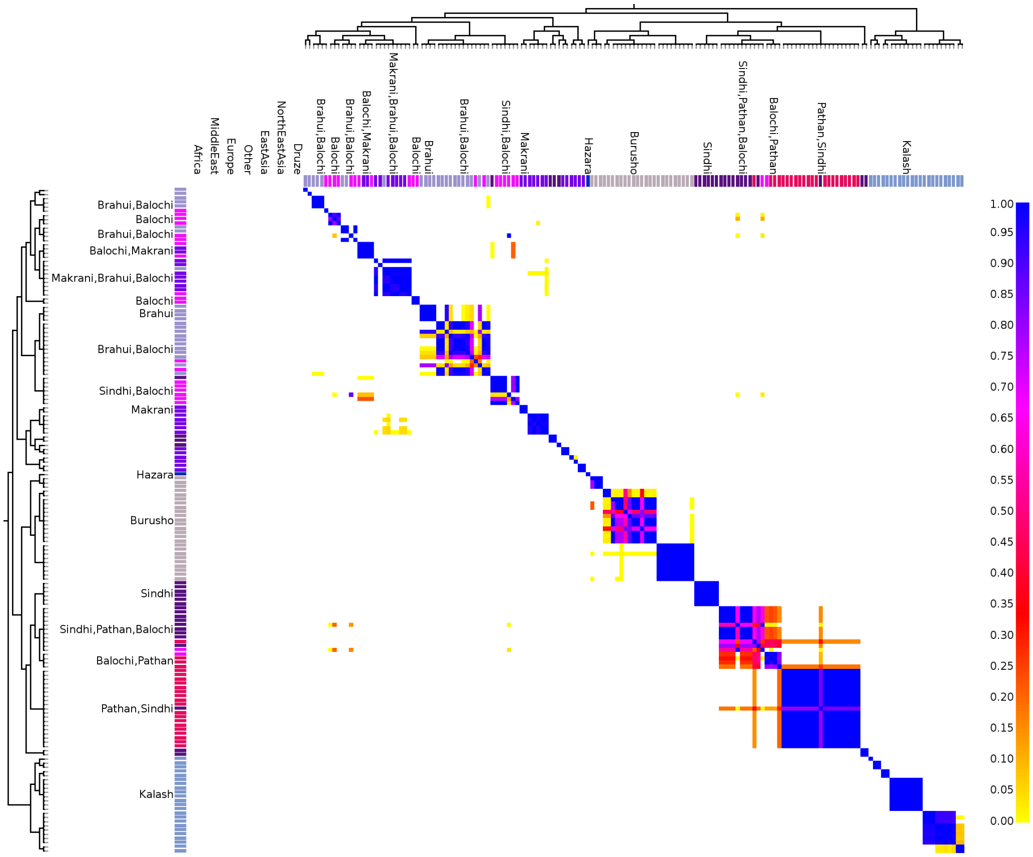

Figure S27: CentralSouthAsia pairwise coincidence matrix showing (bottom left) the run 1 and (top right) run 2.

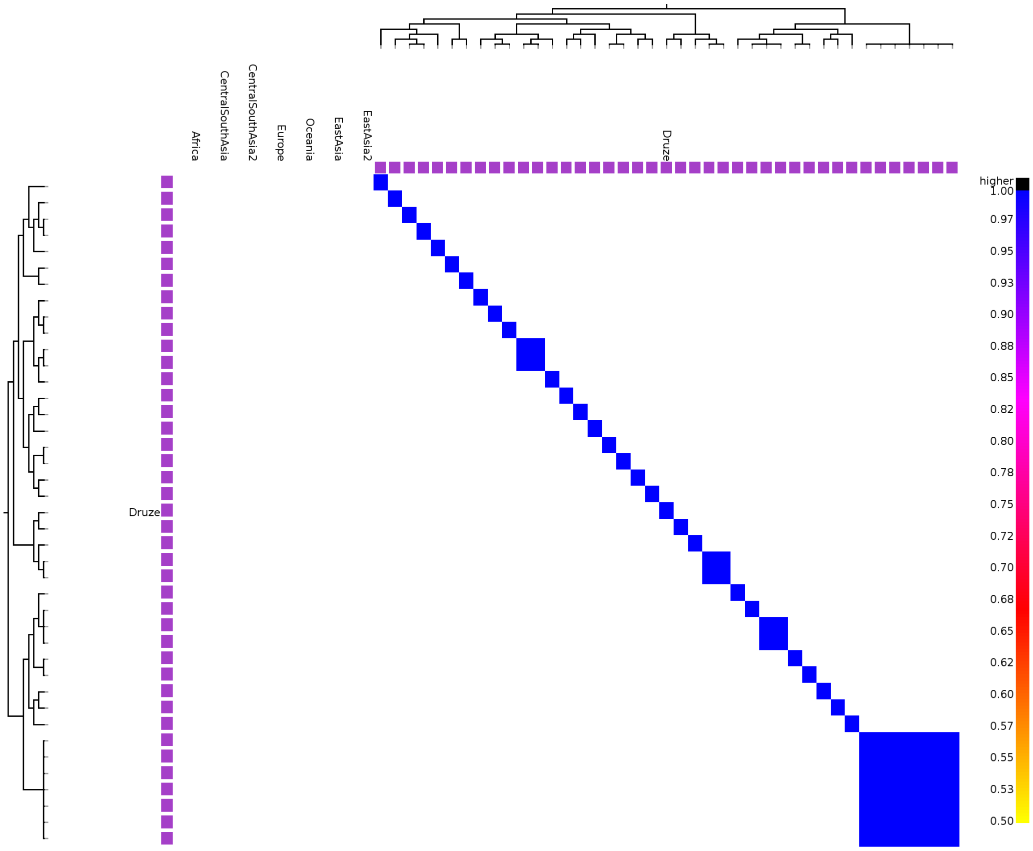

Figure S28: Druze pairwise coincidence matrix showing (bottom left) the run 1 and (top right) run 2.

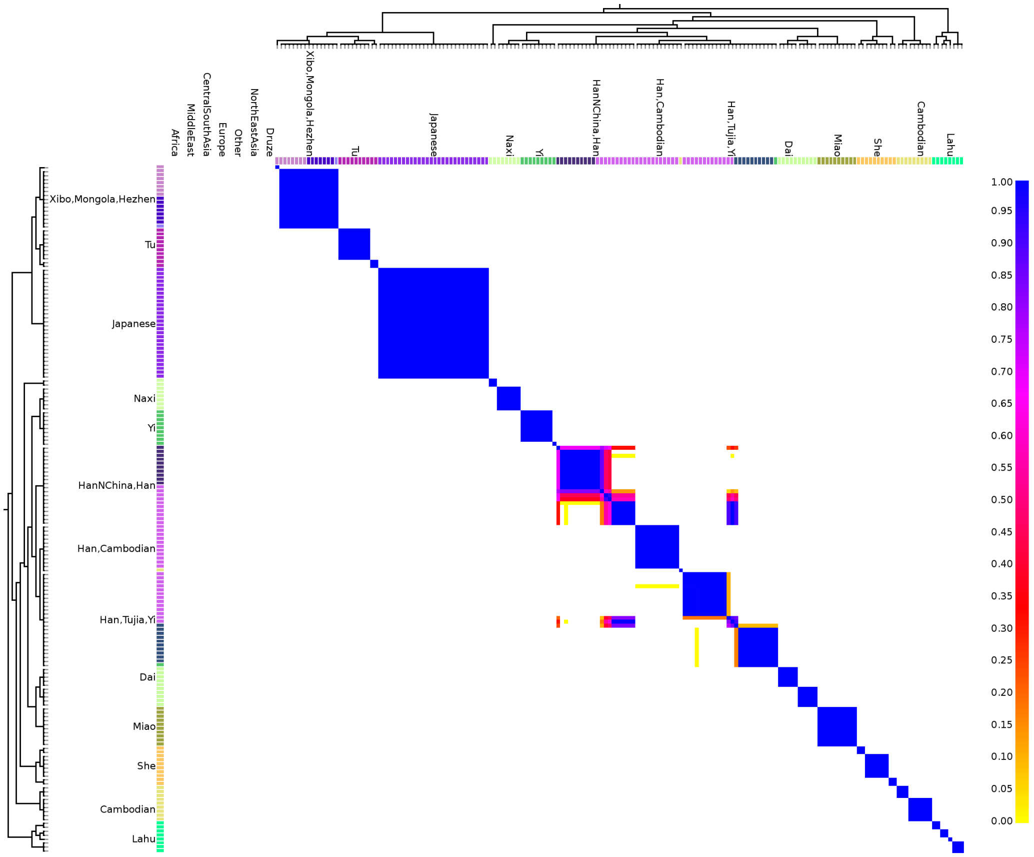

Figure S29: EastAsia pairwise coincidence matrix showing (bottom left) the run 1 and (top right) run 2.

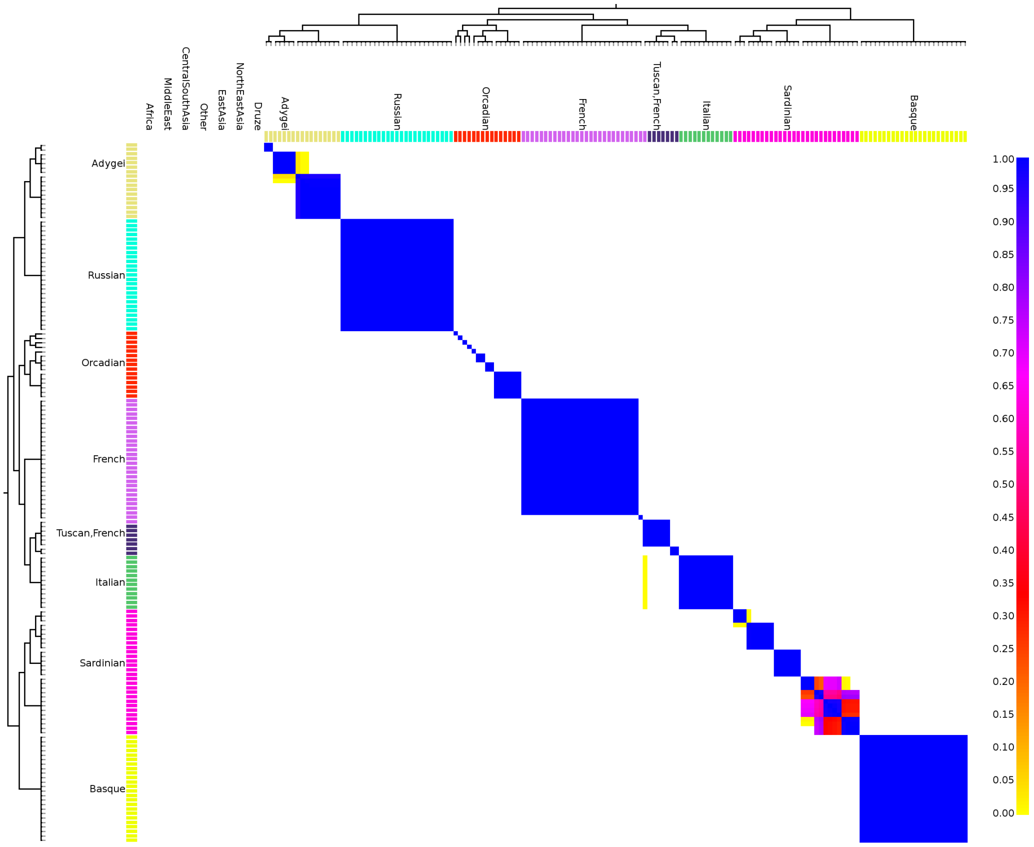

Figure S30: Europe pairwise coincidence matrix showing (bottom left) the run 1 and (top right) run 2.

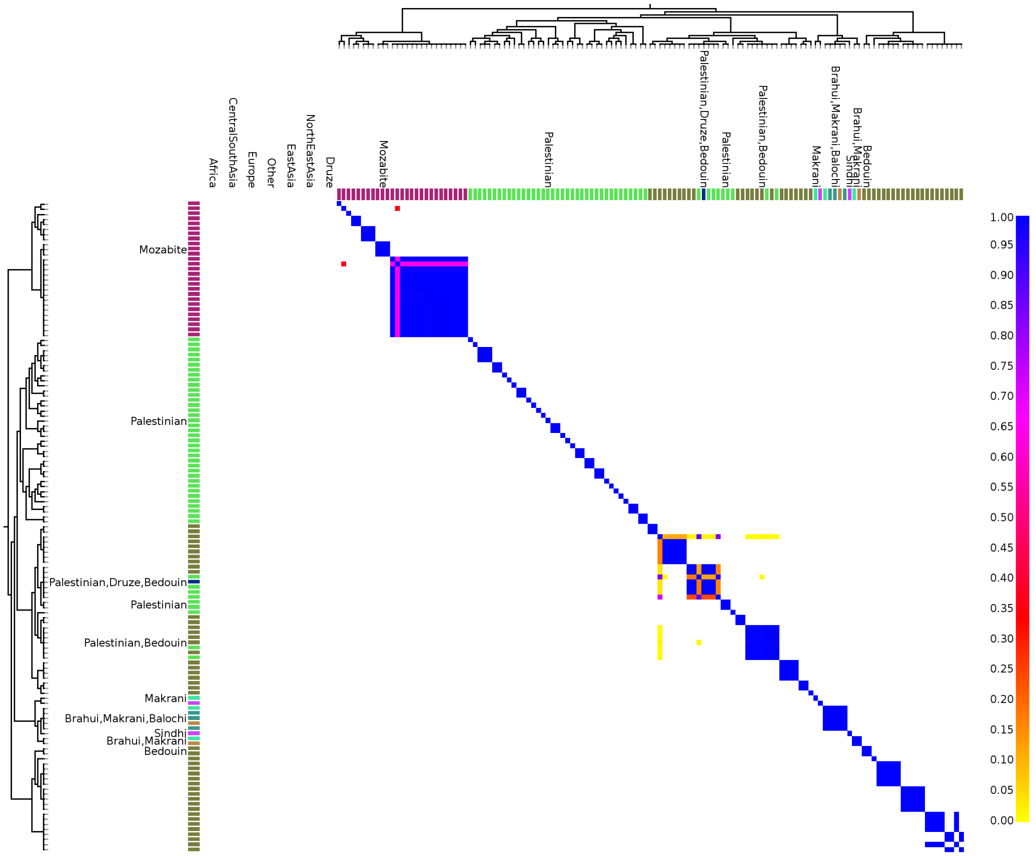

Figure S31: MiddleEast pairwise coincidence matrix showing (bottom left) the run 1 and (top right) run 2.

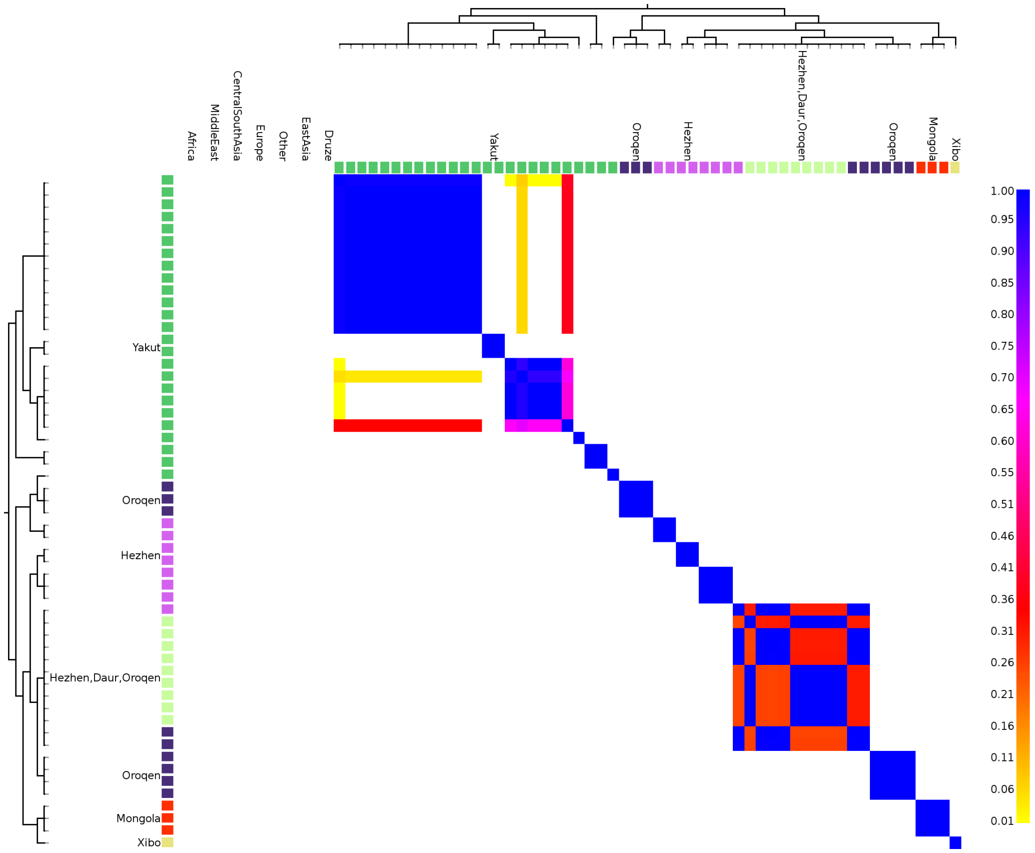

Figure S32: NorthEastAsia pairwise coincidence matrix showing (bottom left) the run 1 and (top right) run 2.

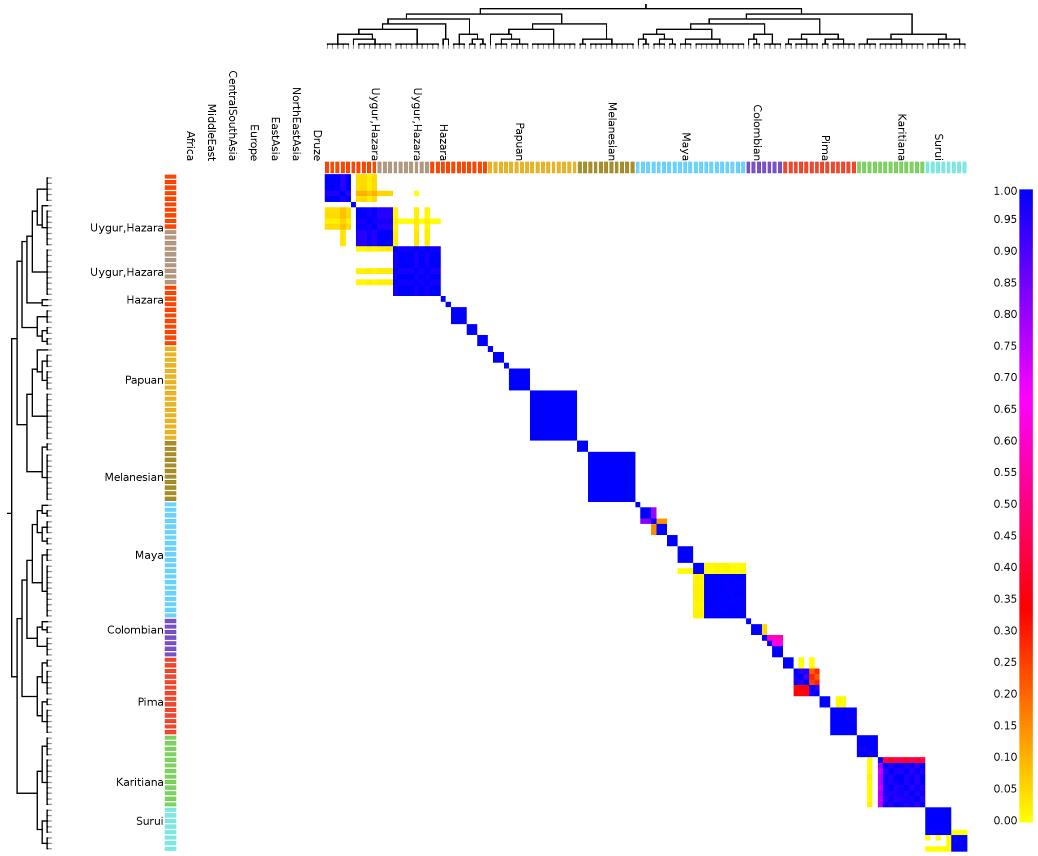

Figure S33: Other populations pairwise coincidence matrix showing (bottom left) the run 1 and (top right) run 2.

S12 Principal Components Analysis for Continents

Note that these are performed on ‘continents’, i.e. pre-defined groupings of individuals based on labels.

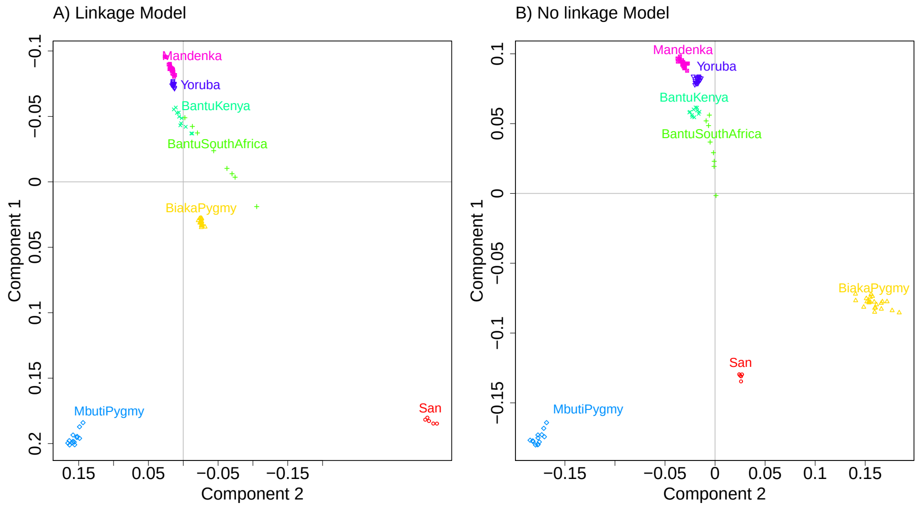

Figure S34: PCA (first 2 components) for the continent of Africa.

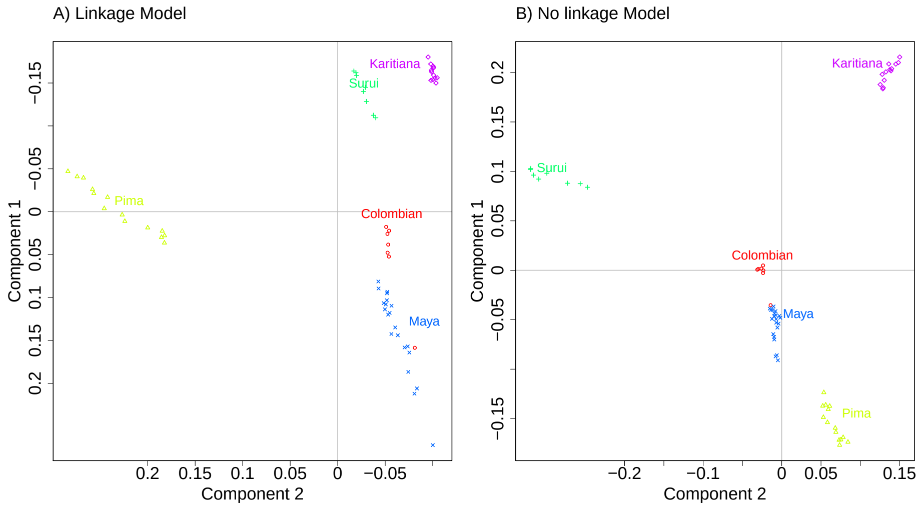

Figure S35: PCA (first 2 components) for the continent of America.

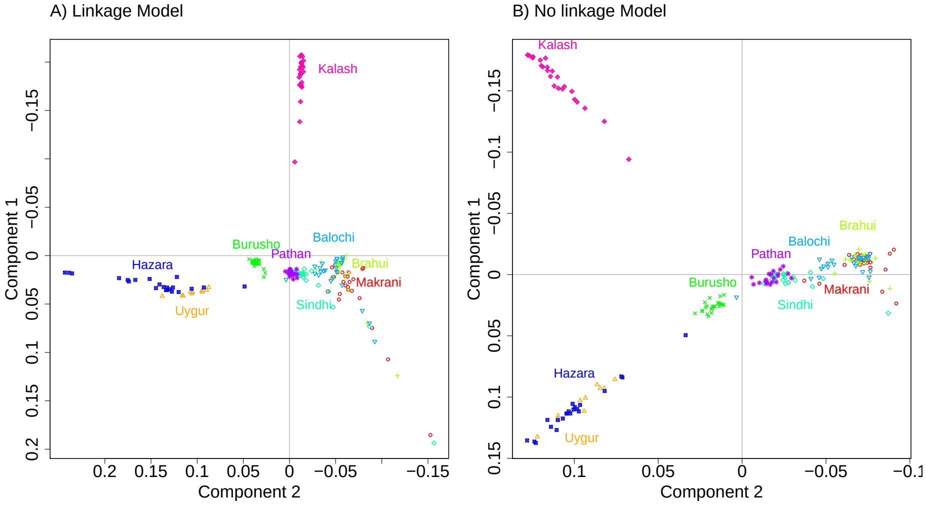

Figure S36: PCA (first 2 components) for the continent of CentralSouthAsia.

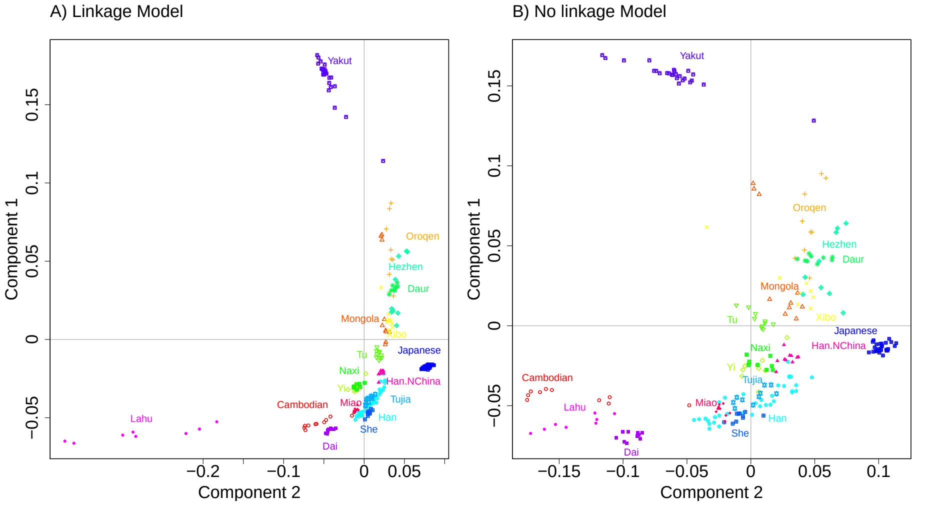

Figure S37: PCA (first 2 components) for the continent of EastAsia.

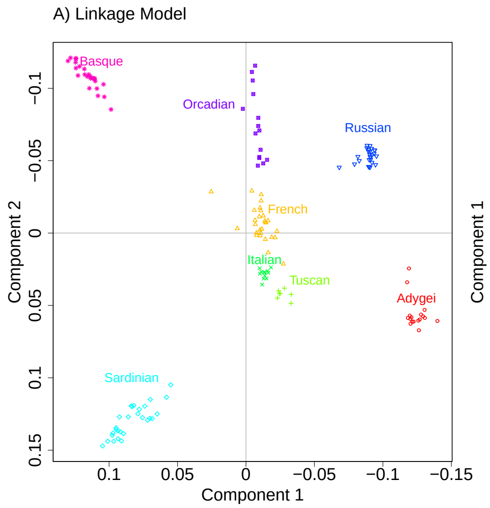

Figure S38: PCA (first 2 components) for the continent of Europe.

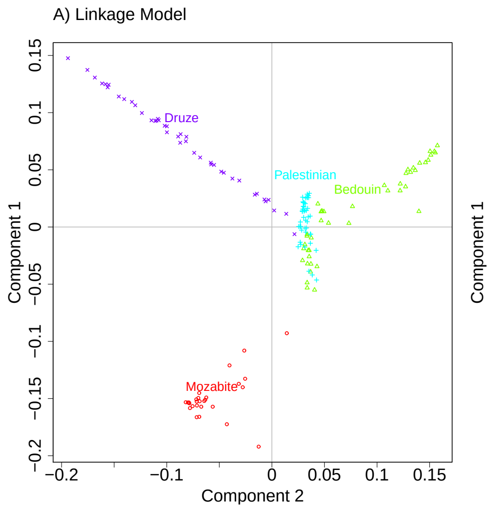

Figure S39: PCA (first 2 components) for the continent of MiddleEast.

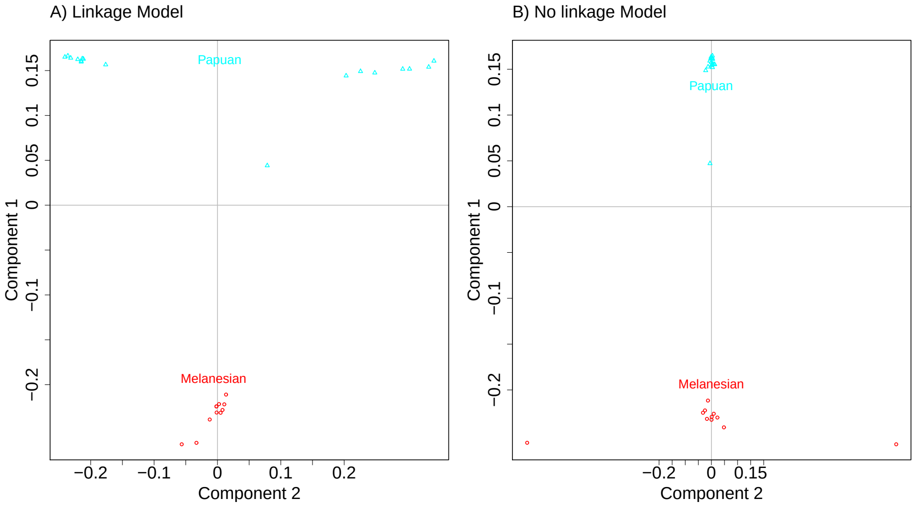

Figure S40: PCA (first 2 components) for the continent of Oceania.

References

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

Dahl, D. B., 2003 An Improved Merge-Split Sampler For Conjugate Dirichlet Process Mixture Models. Technical Report 1086, University of Wisconsin – Madison. `http://www.stat.tamu.edu/~dahl/papers/sams/tr1086.pdf`.

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

Huelsenbeck, J. P. and P. Andolfatto, 2007 Inference of Population Structure Under a Dirichlet Process Model. Genetics **175:** 1787–1802.

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

Lange, K., 2002 *Mathematical and statistical methods for genetic analysis*. Springer.

Li, N. and M. Stephens, 2003 Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. Genetics **165:** 2213–2233.

Myers, S., R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman, T. S. MacFie, G. McVean, and P. Donnelly, 2010 Drive Against Hotspot

Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. Science **327:** 876–879.

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

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

Pella, J. and M. Masuda, 2006 The Gibbs and split–merge sampler for population mixture analysis from genetic data with incomplete baselines. Can. J. Fish. Aquat. Sci. **63:** 576–596.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38:** 904–909.

Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers, 2009 Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS Genetics **5:** e1000519.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of Population Structure Using Multilocus Genotype Data. Genetics **155:** 945–959.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. de Bakker, M. Daly, and P. Sham, 2007 PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics **81:** 559–75.

R Development Core Team, 2009 R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0.

Rabiner, L., 1989 A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77:** 257–286.

Teh, Y. W., 2010 Dirichlet Process. In C. Sammut and G. Webb (Eds.), *Encyclopedia of Machine Learning*, pp. 280–287. Springer.

Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Winckler, W., S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald, R. E. Bontrop, G. A. T. McVean, S. B. Gabriel, D. Reich, P. Donnelly, and D. Altshuler, 2005 Comparison of Fine-Scale Recombination Rates in Humans and Chimpanzees. Science **308:** 107–111.