Supplementary Information: Novel probabilistic models of spatial genetic ancestry with applications to stratification correction in genome-wide association studies

Anand Bhaskar[*†‡§] Adel Javanmard[*†¶] Thomas A. Courtade[‖] David Tse[‖**]

1 Probabilistic model and localization algorithm

Suppose that we are given genotypes from n individuals at p SNPs distributed across the geographic region under study. We denote by $x_{i\ell} \in \{0, 1, 2\}$ the observed number of alleles at SNP ℓ in individual i for $\ell = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$. Further, let X be the $n \times p$ genotype matrix, where the (i, ℓ) entry is $x_{i\ell}$.

In order to capture the spatial structure of the genotype matrix, we let $\mathbf{z}_i$ be the geographical location of individual i and for each SNP ℓ, we view allele frequency q_ℓ as a function of location $\mathbf{z}_i$, i.e., $q_\ell(\mathbf{z}_i)$. Note that $\mathbf{z}_i$ are unobserved ancestry coordinates with the implicit assumption that random mating and localized migration has been occuring between proximate locations.

We define a general flexible probabilistic model of allele frequencies that generalize several previously developed parametric models of spatial genetic variation such as SPA (Yang et al. 2012), SCAT (Wasser et al. 2004) and SpaceMix (Bradburd et al. 2016). In our model, we consider an arbitrary stochastic process over the geographical region under consideration. The allele frequencies $\{q_\ell\}_{\ell=1}^p$ for different SNPs are independent sample paths drawn from this stochastic process.

Throughout, we use the shorthand $q_{i\ell} \equiv q_\ell(\mathbf{z}_i)$ to represent the allele frequency for SNP ℓ conditional on $\mathbf{z}_i$. Assuming Hardy-Weinberg equilibrium, genotypes are generated by binomial sampling as

$$x_{i\ell} \mid q_{i\ell} \sim \mathsf{Binomial}(2, q_{i\ell}) \,.$$

Before presenting our localization algorithm, we provide a brief overview of the PCA method for recovering geographic ancestry. We explain the rationale behind PCA from a perspective that motivates our algorithm and clarifies its superiority over PCA.

1.1 Why PCA?

We denote the genotypes for individual i by $x^i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$. This can be viewed as a representation of individuals in the p-dimensional space. In this way, the localization task seeks for

[*]These authors contributed equally to this work and are ordered alphabetically.
[†]Correspondence should be addressed to A.B. (abhaskar@stanford.edu) or A.J. (ajavanma@marshall.usc.edu).
[‡]Department of Genetics, Stanford University, Stanford, CA 94305
[§]Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305
[¶]Marshall School of Business, University of Southern California, Los Angeles, CA 90089
[‖]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720
[**]Department of Electrical Engineering, Stanford University, Stanford, CA 94305

an embedding of individuals from the p-dimensional space into the two-dimensional geographical region.

PCA with two principal components gives the best rank two approximation of the genotype matrix X in the following sense.[1] Define

$$X_* = \arg\min_{Y \in \mathbb{R}^{n\times 2}} \|XX^\mathsf{T} - YY^\mathsf{T}\|_F \,, \tag{1}$$

where for a matrix $A = (a_{i,\ell})$, $\|A\|_F = (\sum_{i,\ell} a_{i,\ell}^2)^{1/2}$ indicates the Frobenius norm. The solution to (1) is given by the top singular vectors of X. Specifically, let $X = U\Sigma V^T$ be the singular value decomposition of X where $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_n$. The solutions to (1) are given by $X_* = U_2\Sigma_2 Q_2$, where U_2 denotes the first two left singular vectors, $\Sigma_2 = \mathrm{diag}(\sigma_1, \sigma_2)$, and Q_2 is an arbitrary 2×2 orthogonal matrix (i.e. $Q_2 Q_2^T = I$).

We next recall the following identity that for a given set of points relates their pairwise inner products to their pairwise distances. Consider the centering matrix $L = I_{n\times n} - \mathbf{1}\mathbf{1}^\mathsf{T}$, where $I_{n\times n}$ is the identity matrix of size n and $\mathbf{1} = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$. For a vector v, Lv centers the entries of v by subtracting the mean of the entries of v. Further, we denote by D the squared distance matrix $D_{ij} = \|x^i - x^j\|^2$. Using this notation, we have the following,

$$LXX^\mathsf{T}L = -\frac{1}{2}LDL \,.$$

A common preprocessing step for PCA is to center each column in the genotype matrix. This centered genotype matrix is precisely LX. It is then straightforward to see that PCA provides a lower dimensional representation of points $\{x^i\}_{i=1}^n$ such that their squared distance matrix D_* solves the following optimization problem,

$$D_* = \arg\min_{\tilde{D} \in \mathcal{D}} \|LDL - L\tilde{D}L\|_F \,. \tag{2}$$

Here $\mathcal{D}$ is the set of squared distance matrices for all possible two-dimensional embeddings of the points $\{x^i\}_{i=1}^n$. In other words, PCA seeks a low-dimensional representation of the n points that best approximates *all* pairwise distances.

However, an important question that is unanswered is the following:

> *How are the spatial distances between individuals reflected in their genotype information?*

The PCA approach merely assumes that for any two individuals, their genotype distance is a good approximation of their spatial distance and hence it returns the embedding of individuals on the map that best preserves all pairwise genotype distances. However, a more profound answer to the above question requires a model that relates genetic distances to spatial distances. The PCA approach to ancestry localization lacks such a model. In the following, we use our proposed probabilistic model for the allele frequencies to answer the above question.

According to our model, the spatial allele frequencies at each SNP ℓ come from some spatial stochastic process. Our model posits that the underlying spatial processes are second-order stationary, in the sense that for each SNP ℓ, $\mathbb{E}(q_\ell(\mathbf{z})) = \mu_\ell$ for all locations $\mathbf{z}$, and the allele frequency covariance functions $\mathrm{Cov}(q_\ell(\mathbf{z}), q_\ell(\mathbf{z}'))$ depend solely on $\mathbf{z} - \mathbf{z}'$ as follows,

$$\mathrm{Cov}(q_\ell(\mathbf{z}), q_\ell(\mathbf{z}')) = \mathbb{E}[(q_\ell(\mathbf{z}) - \mu_\ell)(q_\ell(\mathbf{z}') - \mu_\ell)] := \eta(\mathbf{z} - \mathbf{z}') \,. \tag{3}$$

[1] In general, the top k principal components give the best rank k approximation in a similar sense.

Note that while the processes for different SNPs can have different means μ_ℓ, they share the same covariance function $\eta(\cdot)$. The implicit structure imposed by the second-order stationarity are used by our localization algorithm GAP, which consists of three main steps:

(1) Construct consistent estimators of $\eta(\cdot)$ and $\mu(\cdot)$ using the genotype information from p SNPs ($p \gg n$).

(2) Use $\eta(\cdot)$ and $\mu(\cdot)$ functions to approximate *local* spatial distances between the individuals.

(3) Find a *global* embedding of individuals on the geographical map that respects the estimated *local* distances.

1.2 GAP algorithm

In the following, we discuss the details of each step.

Step (1)**: Estimating mean and autocorrelation.** Estimates for $\eta(\cdot)$ and $\mu(\cdot)$ functions are given by Theorem 1 below.

Theorem 1. *Consider the proposed probabilistic model for the allele frequency functions and define the following quantities:*

$$\hat{\mu}_\ell = \frac{1}{2n}\sum_{i=1}^{n} x_{i\ell}\,, \tag{4}$$

$$\hat{\eta}_{i,j} = \frac{1}{p}\sum_{\ell=1}^{p}\left(\frac{x_{i\ell}}{2}-\hat{\mu}_\ell\right)\left(\frac{x_{j\ell}}{2}-\hat{\mu}_\ell\right), \tag{5}$$

$$\hat{\eta}_0 = \frac{1}{p}\sum_{\ell=1}^{p}\left(\frac{1}{2n}\sum_{i=1}^{n}(x_{i\ell}^2 - x_{i\ell}) - \hat{\mu}_\ell^2\right). \tag{6}$$

Let $\mathbf{1} = (1/\sqrt{n},\ldots,1/\sqrt{n})^\mathsf{T}$ *and let* $K \in \mathbb{R}^{n\times n}$ *with* $K_{ij} = \eta(\mathbf{z}_i - \mathbf{z}_j)$*. Further, set* $\kappa := \mathbf{1}^\mathsf{T} K \mathbf{1}$*. Then with probability at least* $1-(n+1)^{-2}$*, the following statements are true:*

$$|\hat{\eta}_{i,j} - \eta(\mathbf{z}_i - \mathbf{z}_j)| \le 5\sqrt{\frac{2\log(n+1)}{p}} + 16\sqrt{\frac{\kappa}{n}} + \frac{8}{n}\,, \qquad \forall 1 \le i \ne j \le n\,, \tag{7}$$

$$|\hat{\eta}_0 - \eta(0)| \le 5\sqrt{\frac{2\log(n+1)}{p}} + 16\sqrt{\frac{\kappa}{n}} + \frac{8}{n}\,. \tag{8}$$

Remark 1.1. *The estimates in* (5) *and* (6) *are consistent, i.e., as the number of individuals* n *increases indefinitely, the resulting sequence of estimates converges in probability to the quantities of interest, provided that* $\log(n+1)/p \to 0$ *and* $\kappa/n \to 0$*. (Note that* κ *is bounded by the spectral radius of* K*.)*

We provide the proof of Theorem 1 in §3.

Step (2)**: Estimating local spatial distances.** The next step consists in showing how the local spatial distances can be inferred from functions $\eta(\cdot)$ and $\mu(\cdot)$. To do so, we write the Taylor expansion of $\eta(\cdot)$ around the origin,

$$\eta(\mathbf{z}_i - \mathbf{z}_j) - \eta(\mathbf{0}) = \nabla\eta(\mathbf{0})^\mathsf{T}(\mathbf{z}_i - \mathbf{z}_j) + \frac{1}{2}(\mathbf{z}_i - \mathbf{z}_j)^\mathsf{T}\nabla^2\eta(\mathbf{0})(\mathbf{z}_i - \mathbf{z}_j) + O(d_{ij}^3)\,, \tag{9}$$

where $d_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|$ represents the spatial distance between individuals i and j. Further, $\nabla\eta$ and $\nabla^2\eta$ respectively denote the gradient and the Hessian of the $\eta(\cdot)$ function. Recall that the autocorrelation function of a stationary distribution achieves its maximum at zero and therefore $\nabla\eta(\mathbf{0}) = \mathbf{0}$. Further, $\nabla^2\eta(\mathbf{0})$ is negative semidefinite. We let J be a square root of $(-1/2)\nabla^2\eta(\mathbf{0})$. For 'local distances' where d_{ij} is small enough we can neglect the higher order term $O(d_{ij}^3)$ in (9) and therefore,

$$\eta(\mathbf{0}) - \eta(\mathbf{z}_i - \mathbf{z}_j) \approx \|J(\mathbf{z}_i - \mathbf{z}_j)\|^2 . \tag{10}$$

Using our estimates from the previous step we obtain

$$\hat{\eta}_0 - \hat{\eta}_{i,j} \approx \|J(\mathbf{z}_i - \mathbf{z}_j)\|^2 . \tag{11}$$

We correct for the transformation J using some anchor individuals whose locations are known apriori.[2] Hence, we obtain consistent estimates for local pairwise distances.

It is worth noting that the above argument fails if d_{ij} is large because the higher order term $O(d_{ij}^3)$ cannot be neglected in our estimation procedure. We thus employ a threshold value τ and only use the estimated distances in (11) for individuals i and j for which $\hat{d}_{ij} = (\hat{\eta}_0 - \hat{\eta}_{i,j})^{1/2} \leq \tau$. When some of the estimated local distances $\hat{\eta}_0 - \hat{\eta}_{i,j}$ are negative, we shift all of the estimates by the smallest constant which makes them non-negative. We discuss the procedure for choosing this threshold τ in §1.3.

Step (3)**: Global embedding.** The final step is finding a 'global' embedding of individuals from their estimated local pairwise distances. There has been a great deal of research on this task as it appears in various applications such as network localization (Shang et al. 2003; Patwari et al. 2005) and reconstruction of protein conformations from NMR measurements. It is also directly related to dimensionality reduction of high dimensional data under the topic of manifold learning. Several interesting algorithms have been proposed in the literature for this task. Probably the most well-known is the ISOMAP algorithm (Tenenbaum et al. 2000). It first estimates the missing pairwise distances by computing the shortest path between all pairs of nodes, via local distances. It then applies multidimensional scaling (MDS) to infer the locations from the pairwise distances. Some other methods for this task are Locally linear embedding (LLE) (Saul and Roweis 2003), Laplacian eigenmap (Belkin and Niyogi 2002), Hessian eigenmap (Donoho and Grimes 2003), and Locally rigid embedding (Singer 2008). Another group of algorithms formulate the localization task as a non-convex optimization problem and then consider different convex relaxations to solve it. A famous example of this type is the relaxation to semidefinite programming (SDP) (Biswas and Ye 2004; Alfakih et al. 1999; Weinberger and Saul 2006; Javanmard and Montanari 2013).

One can use any of the above proposed methods for this step. In the remainder of this paper, we use the ISOMAP algorithm to infer the locations from the estimated local distances. For the reader's convenience, we summarize the steps of ISOMAP below.

Let $\hat{d}_{ij} = (\hat{\eta}_0 - \hat{\eta}_{i,j})^{1/2}$ be estimated pairwise distances. Construct a graph G with n nodes such that i and j are connected by an edge of weight $\hat{d}_{ij}$ if they are within the local distance threshold, i.e., $\hat{d}_{ij} \leq \tau$. The steps of ISOMAP follow:

(1) Compute pairwise shortest paths in the (weighted) graph G.

(2) Let D_τ be the matrix of squared shortest paths distances in G.

[2]PCA also reconstructs locations only up to an orthogonal transformation. In particular, if X_* is a solution to the optimization problem (1), then X_*Q_2 for any 2×2 orthogonal matrix Q_2 is also a solution.

(3) Let (u_1, u_2) and (σ_1, σ_2) be the top two eigenvectors and eigenvalues of $(-1/2)LD_\tau L$, where $L = I_{n\times n} - \mathbf{1}\mathbf{1}^\mathsf{T}$ is the centering matrix.

(4) Return the estimated locations $\mathbf{z}_i = (\sqrt{\sigma_1}u_{1,i}, \sqrt{\sigma_2}u_{2,i})$, for $i = 1, 2, \ldots, n$.

Remark 1.2. *It is straightforward to verify that output of* GAP *is unaltered if we relabel the alleles at any SNP. In other words, for any SNP ℓ, if we replace the genotypes 0 and 2 for all individuals at that SNP,* GAP *returns the same locations.*

1.3 Choosing the local distance threshold τ

We describe two strategies for choosing the distance threshold τ that we use in order to estimate spatial distances from genetic distances in (11). When we have the true sampling locations for the individuals in the dataset, we can use a subset of these known locations as training data for choosing τ. In particular, in the simulations for GAP in §1.5 and in the simulations for the association testing procedure SCGAP in §2.1, we used the known locations of a random subset of 20% of the samples as training data, and chose the value of τ that minimized the spatial reconstruction RMSE on the training set. For evaluation on the real datasets where we have the sampling coordinates for each subpopulation, we used a leave-one-out cross-validation procedure on the training set to choose the value of τ. For the simulation scenarios, we optimized τ using the RMSE on the training set instead of performing leave-one-out cross-validation for computational reasons. Figure S1 shows an example of the dependence of the spatial reconstruction RMSE (of the entire dataset) on the choice of τ in the isotropic covariance decay model. In general, different covariance decay models η will exhibit different dependence of the spatial reconstruction RMSE on the threshold τ, and one can tune this parameter using the kind of training and cross-validation procedures that are commonly employed in machine learning.

The second-order stationarity assumption of our model, i.e. $\mathrm{Cov}(q_\ell(\mathbf{z}), q_\ell(\mathbf{z}')) = \eta(\mathbf{z} - \mathbf{z}')$, implicitly assumes that the covariance decay function η is the same across space. However, we expect different covariance decay functions in different geographic regions (Ramachandran and Rosenberg 2011; Jay et al. 2013) due to geographic barriers, historical migrations, and other factors that introduce spatial heterogeneity. The distance threshold τ is used to determine the regime in which the second-order Taylor expansion of the η function given in (9) can be considered to be valid. As a result, when the estimated covariances $\hat{\eta}_{ij}$ and distances $\hat{d}_{ij}$ change, say due to refocusing an earlier analysis on a subset of the samples, it would make sense to retune the local distance threshold τ via the above cross-validation procedure using samples from the relevant geographic region.

When we do not have any individuals with known locations, as is the case with the Northern Finland Birth Cohort GWAS dataset analyzed in §2.2, we use the following procedure for picking τ. For any given threshold τ, D_τ is the squared shortest paths matrix produced in the second step of the ISOMAP algorithm. Let $\tilde{D}_\tau$ denote the pairwise squared distance matrix of the two-dimensional embedding produced in the fourth step of the ISOMAP algorithm. We choose τ to maximize the value of $||L\tilde{D}_\tau L||_*/||LD_\tau L||_*$, where $\|A\|_*$ is the nuclear norm of the matrix A and is given by the sum of the singular values of A.

1.4 Relation to previous spatial models

Here, we show that several previously proposed spatial genetic models and ancestry localization algorithms can be viewed as a special case of our probabilistic model and algorithm.

1. SpaceMix (Bradburd et al. 2016): This model posits that the distribution of alleles among individuals comes from a spatial process such that the covariance function F between normalized allele frequencies for individuals i and j has an exponential decay with respect to

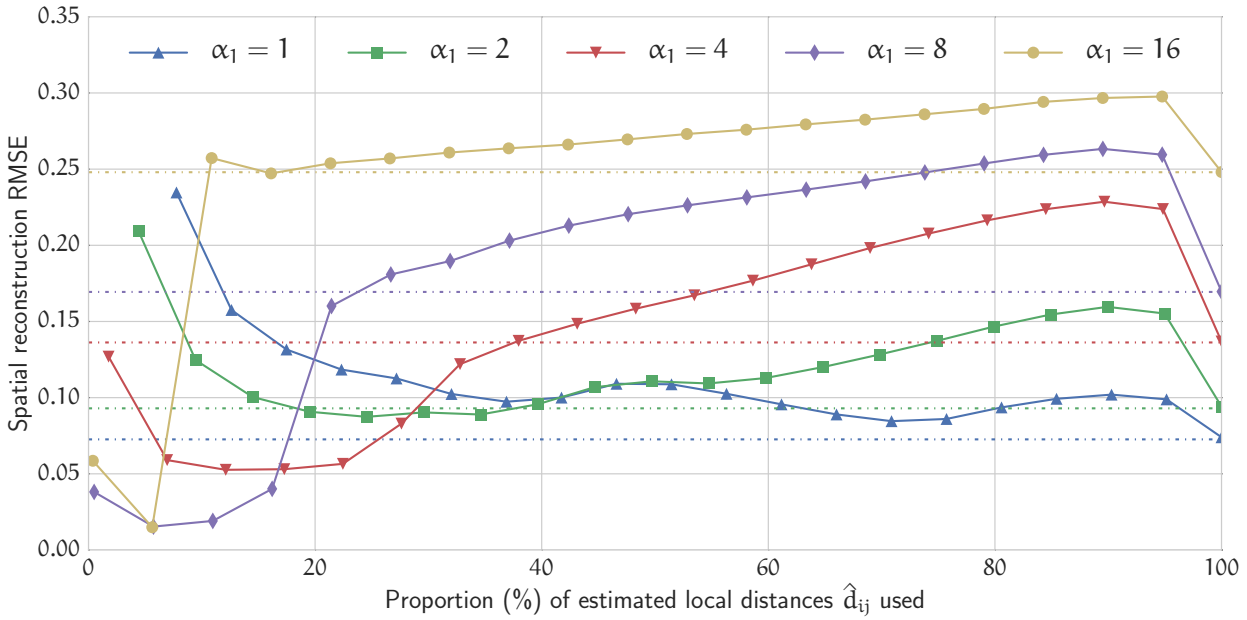


Fig. S1: **Sensitivity of the spatial reconstruction RMSE to the local distance threshold τ for the isotropic covariance decay model.** The simulations were performed with $n = 2{,}000$ individuals sampled from the unit square $[0,1]^2$, with $p = 50{,}000$ SNPs according to the isotropic covariance decay model with parameter combinations $\alpha_0 = \alpha_2 = 1$ and different choices of α_1. Solid lines indicate the spatial reconstruction RMSE of GAP as a function of the distance threshold τ, while the horizontal dashed lines indicate the reconstruction RMSE of PCA. In order to put the different ranges for τ for each parameter setting of α_1 on the same scale, the x-axis is measured using the percentage of the estimated $\binom{n}{2}$ genetic distances $\hat{d}_{ij} \leq \tau$ which are used by GAP.

their spatial distance:

$$F(\mathbf{z}, \mathbf{z}') = \frac{1}{\alpha_0} \exp(-(\alpha_1 \|\mathbf{z} - \mathbf{z}'\|)^{\alpha_2}) \,. \tag{12}$$

This is clearly a special case of our probabilistic model since $F(\mathbf{z}, \mathbf{z}')$ is a function of $\mathbf{z} - \mathbf{z}'$.

2. SCAT (Wasser et al. 2004): For the case of two alleles at each locus (similar to the setting considered in the present paper), this model is based on writing the allele frequencies as

$$q_\ell(\mathbf{z}) = \frac{1}{1 + \exp(\theta_\ell(\mathbf{z}))} \,, \tag{13}$$

where the $\theta_\ell(\cdot)$ values for different SNPs ℓ are assumed to be independent Gaussian processes. For each ℓ, θ_ℓ is a Gaussian spatial process with $\mathbb{E}(\theta_\ell(\mathbf{z})) = \mu_\ell$ and covariance kernel $K_{\theta_\ell}(\mathbf{z}, \mathbf{z}') = (1/\alpha_0) \exp(-(\alpha_1 \|\mathbf{z} - \mathbf{z}'\|)^{\alpha_2})$. Note that process θ_ℓ is translation invariant. More specifically, for any collection of locations $\{\mathbf{z}_i\}_{i=1}^n$, the distribution of $(\theta_\ell(\mathbf{z}_1 + \delta), \cdots, \theta_\ell(\mathbf{z}_n + \delta))$ is invariant to δ. This property is preserved after applying any one-to-one deterministic function, and in particular, the logistic function. Therefore, the process $q_\ell(\mathbf{z})$ is also translation invariant. As a result, the covariance of allele frequencies $\mathrm{Cov}(q_\ell(\mathbf{z}), q_\ell(\mathbf{z}'))$ only depends on $\mathbf{z} - \mathbf{z}'$ and can be written as $\mathrm{Cov}(q_\ell(\mathbf{z}), q_\ell(\mathbf{z}')) = \eta(\mathbf{z} - \mathbf{z}')$ for some function η. This is clearly a special case of our probabilistic model.

3. SPA (Yang et al. 2012): In the SPA model, allele frequencies are given by a logistic function

$$q_\ell(\mathbf{z}) = \frac{1}{1 + \exp(-\langle \mathbf{a}_\ell, \mathbf{z} \rangle - b_\ell)} \,, \tag{14}$$

where $\mathbf{a}_\ell$ and b_ℓ are coefficients for SNP ℓ. Under such a model, the allele frequencies at each SNP ℓ are constant along lines perpendicular to the vector $\mathbf{a}_\ell$. The directional covariance decay model introduced in the Simulation results section of the main text also possesses this property. In the directional covariance decay model, the allele frequency at SNP ℓ and location $\mathbf{z}$ is given by $q_\ell(\mathbf{z}) = 1/(1 + \exp(G_\ell(\mathbf{z})))$ where $G_\ell(\cdot)$ is a sample path from a Gaussian spatial process with mean 0 and covariance kernel $K(\mathbf{z}, \mathbf{z}') = (1/\alpha_0) \exp(-(\alpha_1 |\langle \mathbf{u}, \mathbf{z} - \mathbf{z}' \rangle|)^{\alpha_2})$. For any two locations $\mathbf{z}, \mathbf{z}'$ such that $\mathbf{z} - \mathbf{z}' \perp \mathbf{u}$, we have $K(\mathbf{z}, \mathbf{z}') = 1/\alpha_0$. Further, $K(\mathbf{z}, \mathbf{z}) = K(\mathbf{z}', \mathbf{z}') = 1/\alpha_0$. In words, $G_\ell(\mathbf{z})$ and $G_\ell(\mathbf{z}')$ have equal variance and are perfectly correlated, therefore $G_\ell(\mathbf{z}) = G_\ell(\mathbf{z}')$ almost surely. This argument shows that the lines perpendicular to the direction vector $\mathbf{u}$ are level sets for the allele frequency. This is also apparent from Figure 1 in the main text.

4. PCA (Price et al. 2006; Novembre et al. 2008): We next show that under our probabilistic model for allele frequencies, GAP asymptotically *always* dominates PCA. Specifically, if we choose the local distance threshold τ to be large enough, then GAP and PCA return the same outputs in the asymptotic regime $n \to \infty$, and hence PCA can be viewed as a special case of GAP. In Tables S3–S10, the ratio of the RMSE of GAP to the RMSE of PCA exceeds 1 by a very small amount for some parameter combinations, which is due to the effect of finite sample size. However, this effect of the finite sample size is already very small for $n \geq 2{,}000$. To corroborate our claim, recall that PCA estimates locations using the two top eigenvectors of $LXX^\mathsf{T}L$, where L is the centering matrix $L = I - \mathbf{1}\mathbf{1}^\mathsf{T}$, with $\mathbf{1} = (1/\sqrt{n}, \ldots, 1/\sqrt{n})^\mathsf{T}$ the unit norm vector with equal entries. Often, the columns of the centered genotype matrix are normalized to have unit variance before applying PCA (This is also done in our simulations.) In the asymptotic regime $n \to \infty$, the normalization factors for all columns concentrate at

$\eta(\mathbf{0})$ and as such PCA uses the (scaled) two top eigenvectors of $(1/\eta(\mathbf{0}))LXX^\mathsf{T}L$. On the other hand, in Step (1) of GAP, the estimates $\hat{\eta}_{i,j}$ can be written as the (i,j) entry of $LXX^\mathsf{T}L$. Let $\widehat{D} = (\hat{d}_{ij}^2)$ where $\hat{d}_{ij} = (\hat{\eta}_0 - \hat{\eta}_{i,j})^{1/2}$ are the estimated local spatial distances in Step (2). We thus have the matrix representation $\widehat{D} = \hat{\eta}_0 \mathbf{1}\mathbf{1}^\mathsf{T} - LXX^\mathsf{T}L$. If τ is chosen to be larger than the range of pairwise distances $\hat{d}_{ij}$, all of them will be treated as local distances and no thresholding occurs. Therefore, in Step (3) the constructed graph G is a complete graph and the squared shortest path distances $\widehat{D}$ are indeed the squared local distances D. The ISOMAP employed in the last step reduces to PCA applied to

$$-\frac{1}{2}L\widehat{D}L = -\frac{1}{2}L(\hat{\eta}_0\mathbf{1}\mathbf{1}^\mathsf{T} - LXX^\mathsf{T}L)L = \frac{1}{2}LXX^\mathsf{T}L\,,$$

where the last equality holds because $L\mathbf{1} = \mathbf{0}$ and $L^2 = L$. It is now clear that GAP and PCA are the same procedure in this case (up to a scaling factor which is corrected for using some individuals with known locations).

1.5 Ancestry localization simulations

As described in the main text, we considered two sets of simulation scenarios to model isotropic and direction-dependent decay rates for the allele frequency covariance. For both simulation scenarios, we simulated $n = 2{,}000$ individuals at $p = 50{,}000$ SNPs. The true geographic origin $\mathbf{z}_i$ of individual i was simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square. This distribution lets us smoothly interpolate between dense sampling of individuals in the interior of the space to dense sampling at the boundaries (Figures 1(a) and 1(d) in the main text), with $\beta = 1$ representing uniform sampling. We considered $\beta \in \{0.25, 0.5, 1, 2, 4\}$. The spatial allele frequencies at each SNP were generated by applying the logistic function to sample paths from a spatial Gaussian process. The genotypes of each individual i were then drawn according to a binomial distribution from the allele frequencies at their geographic origin $\mathbf{z}_i$.

- **Isotropic covariance decay**: The allele frequency $q_\ell(\mathbf{z}_i)$ of SNP ℓ at location $\mathbf{z}_i$ is given by $q_\ell(\mathbf{z}_i) = 1/(1+\exp(Z_{\ell,i}))$, where $Z_{\ell,\cdot}$ is an n-dimensional Normal random variable with mean $\mathbf{0}$ and covariance $\mathrm{Cov}(Z_{\ell,i}, Z_{\ell,j}) = \exp(-(\alpha_1\|\mathbf{z}_i - \mathbf{z}_j\|)^{\alpha_2})/\alpha_0$. Such covariance decay models have been previously used by Wasser et al. (2004) and Bradburd et al. (2016). Figure S2A in the main text shows example allele frequency surfaces drawn from this model.
- **Directional covariance decay**: Given a unit norm direction vector $\mathbf{u}_k \in \mathbb{R}^2$, the allele frequency $q_\ell(\mathbf{z}_i)$ of SNP ℓ at location $\mathbf{z}_i$ is given by $q_\ell(\mathbf{z}_i) = 1/(1+\exp(Z_{\ell,i}))$, where $Z_{\ell,\cdot}$ is an n-dimensional Normal random variable with mean $\mathbf{0}$ and covariance $\mathrm{Cov}(Z_{\ell,i}, Z_{\ell,j}) = \exp(-(\alpha_1|\langle \mathbf{u}_k, \mathbf{z}_i - \mathbf{z}_j\rangle|)^{\alpha_2})/\alpha_0$. Figure S2B in the main text shows example allele frequency surfaces of this form. Such models can be viewed as a generalization of the SPA model of Yang et al. (2012) (see §1.4). In the simulations, we sampled 100 different direction vectors $\mathbf{u}_k$ from a von Mises distribution, which is a circular analogue of the Normal distribution. For each such direction vector $\mathbf{u}_k$, we simulated 500 SNPs, which will have level sets of equal allele frequency in directions perpendicular to $\mathbf{u}_k$.

For each parameter combination in the above simulation scenarios, we simulated 10 random datasets, and used PCA and our algorithm GAP to infer the spatial coordinates $\mathbf{z}_i$. PCA can estimate the coordinates up to an orthogonal transformation, while GAP estimates coordinates up to the invertible linear transformation J in (11). We use the true geographic locations of a random subset of 20% of the simulated individuals to rescale the coordinates inferred by PCA and GAP. As a measure of inference accuracy, we use the root mean squared error (RMSE) between the inferred

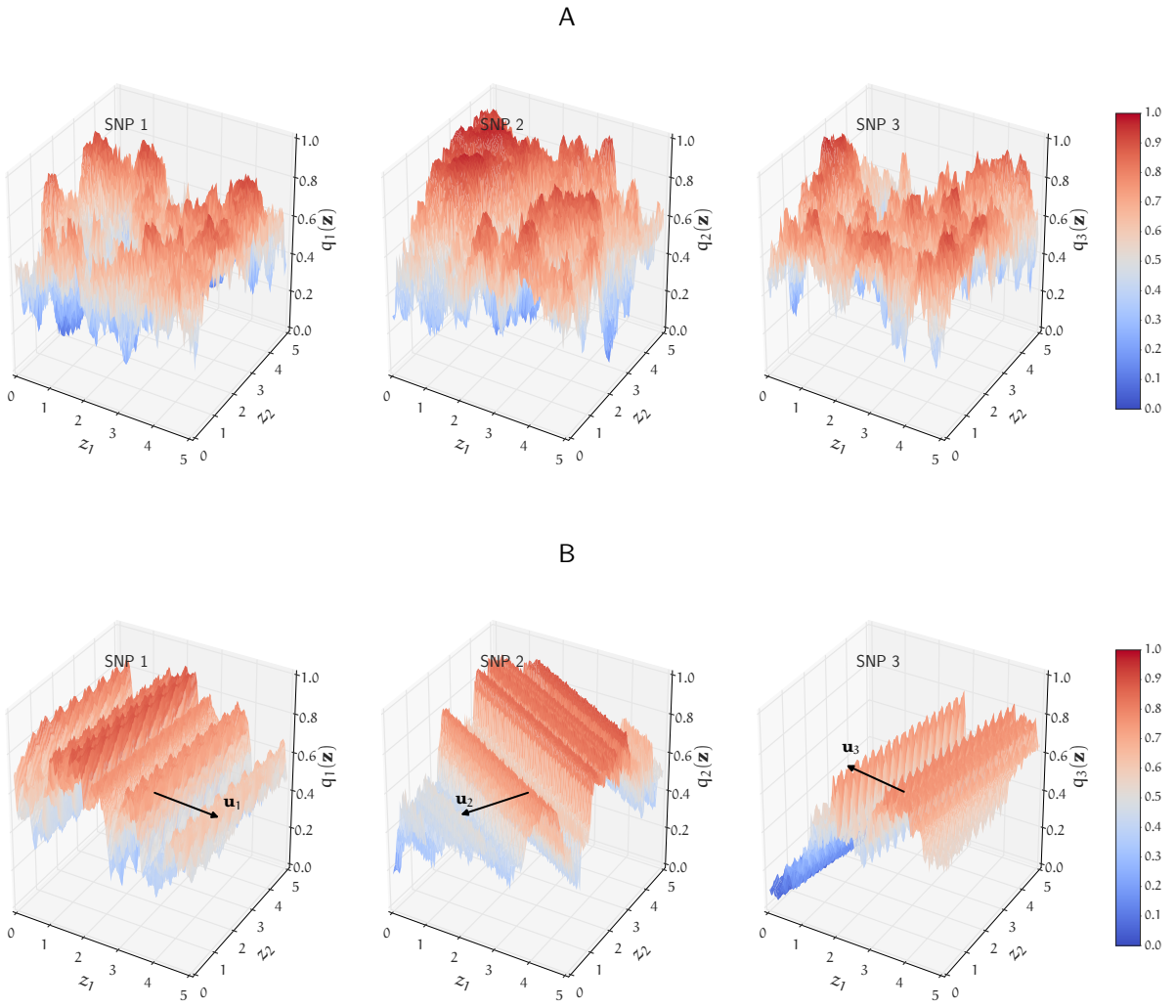


Fig. S2: **Sample spatial allele frequencies from our probabilistic model.** Each figure corresponds to an allele frequency covariance function η in the underlying collection of spatial stochastic processes over the two-dimensional space represented by a 5×5 grid.
(A) **Isotropic covariance decay.** The allele frequency surface is generated by a logistic function applied to a Gaussian process. In particular, the allele frequency $q_\ell(\mathbf{z})$ of SNP ℓ at location $\mathbf{z}$ is given by $q_\ell(\mathbf{z}) = 1/(1 + \exp(G_\ell(\mathbf{z})))$, where $G_\ell(\cdot)$ is a sample path from a stationary Gaussian process with mean 0 and covariance kernel $K(\mathbf{z}, \mathbf{z}') = \exp(-(\alpha_1 \|\mathbf{z} - \mathbf{z}'\|)^{\alpha_2})/\alpha_0$. In this example, $\alpha_0 = \alpha_1 = \alpha_2 = 1$. Such families of allele frequency functions are also used in previous probabilistic models such as SCAT (Wasser et al. 2004) and SpaceMix (Bradburd et al. 2016).
(B) **Directional covariance decay.** The allele frequency $q_\ell(\mathbf{z})$ of SNP ℓ at location $\mathbf{z}$ is given by $q_\ell(\mathbf{z}) = 1/(1 + \exp(G_\ell(\mathbf{z})))$, where $G_\ell(\cdot)$ is a sample path from a stationary Gaussian process with mean 0 and covariance kernel $K(\mathbf{z}, \mathbf{z}') = \exp(-(\alpha_1 |\langle \mathbf{u}, \mathbf{z} - \mathbf{z}' \rangle|)^{\alpha_2})/\alpha_0$. This form for the Gaussian process kernel leads to level sets of equal allele frequency in directions perpendicular to $\mathbf{u}$. In this example, $\alpha_0 = \alpha_1 = \alpha_2 = 1$ and the directions $\mathbf{u}$, shown as black arrows, were randomly chosen for each SNP. Such allele frequency functions can be viewed as a generalization of the logistic allele frequency surfaces considered in the SPA model (Yang et al. 2012).

locations $\hat{\mathbf{z}}_i$ and the true locations $\mathbf{z}_i$ as follows,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{z}_i - \hat{\mathbf{z}}_i\|^2}\,. \tag{15}$$

In order to choose the threshold parameter τ that is used when estimating the local spatial distances from the genetic distances, we picked a uniform grid (of 20 points) over the quantiles of the $\binom{n}{2}$ estimated pairwise genetic distances $\hat{d}_{ij}$. We picked the value of τ which minimized the reconstruction RMSE over the aforementioned random subset containing 20% of the samples whose locations were assumed known. For most parameter combinations in both the simulation models, and for the range of sampling distribution parameters β, the RMSE of GAP is substantially lower than that of PCA (Tables S1–S10).

			$\beta = 1$			
α_2	α_1	$\frac{\text{RMSE GAP}}{\text{RMSE PCA}}$	RMSE PCA	RMSE GAP	τ	Proportion of distances used
	1	1.010	0.0879	0.0888	0.2507	100.0%
	2	**0.978**	0.1030	0.1008	0.2305	17.2%
0.5	4	**0.699**	0.1001	0.0700	0.2349	11.4%
	8	**0.414**	0.1151	0.0477	0.2416	6.2%
	16	**0.307**	0.1372	0.0421	0.2612	5.8%
	1	1.018	0.0716	0.0729	0.2700	100.0%
	2	**0.880**	0.0929	0.0818	0.2599	24.4%
1	4	**0.359**	0.1285	0.0461	0.2584	11.7%
	8	**0.094**	0.1733	0.0163	0.2587	5.7%
	16	**0.096**	0.2554	0.0245	0.2816	5.7%
	1	1.028	0.0555	0.0570	0.2806	100.0%
	2	**0.857**	0.0983	0.0843	0.2800	33.5%
1.5	4	**0.212**	0.1647	0.0349	0.2610	12.2%
	8	**0.100**	0.2842	0.0285	0.2595	5.7%
	16	**0.100**	0.3114	0.0311	0.1963	0.5%

Table S1: **Isotropic covariance decay model**: Comparison of the localization accuracy of GAP and PCA for simulated datasets with $n = 2{,}000$ samples and $p = 50{,}000$ SNPs. The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square. In this table, $\beta = 1$, which is equivalent to sampling the individuals uniformly from the unit square. The allele frequency for individual i at locus ℓ is given by $q_{i\ell} = 1/(1+\exp(Z_{i\ell}))$, where $Z_{\cdot,\ell}$ is an n-dimensional multivariate Gaussian random variable with mean 0 and covariance between the i-th and j-th entries given by $\exp(-(\alpha_1\|\mathbf{z}_i - \mathbf{z}_j\|)^{\alpha_2})/\alpha_0$. In these simulations, $\alpha_0 = 1$. The columns for PCA and GAP show the root mean squared error (RMSE) in the reconstruction accuracy for PCA and MDS, respectively. The column τ indicates the threshold on the genetic distance that was used when applying GAP. This threshold τ was optimized using the known ancestral locations of a random subset of 20% of the simulated points. The last column of the table indicates the percentage of entries in the pairwise genetic distance matrix less than the threshold value τ.

α_2	α_1	$\frac{\text{RMSE GAP}}{\text{RMSE PCA}}$	RMSE PCA	RMSE GAP	τ	Proportion of distances used
			$\beta = 0.5$			
0.5	1	**0.925**	0.0605	0.0560	0.2223	23.4%
	2	**0.774**	0.0650	0.0503	0.2371	23.1%
	4	**0.681**	0.0697	0.0475	0.2326	6.9%
	8	**0.448**	0.0718	0.0321	0.2609	12.1%
	16	**0.444**	0.0826	0.0367	0.2633	6.2%
1	1	**0.966**	0.0623	0.0602	0.2309	37.1%
	2	**0.593**	0.0770	0.0457	0.2524	24.4%
	4	**0.205**	0.1090	0.0224	0.2293	6.8%
	8	**0.200**	0.1174	0.0234	0.2544	5.9%
	16	**0.072**	0.3645	0.0263	0.2290	1.2%
1.5	1	1.027	0.0567	0.0583	0.2817	100.0%
	2	**0.466**	0.0921	0.0429	0.2797	36.9%
	4	**0.195**	0.1336	0.0261	0.2451	11.6%
	8	**0.071**	0.3815	0.0269	0.2619	5.9%
	16	**0.102**	0.4187	0.0425	0.2144	0.9%

Table S2: **Isotropic covariance decay model**: The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square, with $\beta = 0.5$. The other simulation settings are as described in Table S1.

α_2	α_1	$\frac{\text{RMSE GAP}}{\text{RMSE PCA}}$	RMSE PCA	RMSE GAP	τ	Proportion of distances used
			$\beta = 0.25$			
0.5	1	**0.772**	0.0476	0.0367	0.2005	15.1%
	2	**0.590**	0.0524	0.0309	0.2212	15.2%
	4	**0.491**	0.0715	0.0351	0.2232	9.3%
	8	**0.252**	0.1190	0.0300	0.2445	9.1%
	16	**0.221**	0.1938	0.0428	0.2628	8.4%
1	1	1.010	0.0534	0.0540	0.2758	100.0%
	2	**0.545**	0.0532	0.0290	0.2169	15.2%
	4	**0.228**	0.1736	0.0396	0.2462	14.0%
	8	**0.132**	0.3293	0.0433	0.2616	8.5%
	16	**0.067**	0.4390	0.0293	0.2455	3.1%
1.5	1	**0.991**	0.0464	0.0460	0.2315	40.3%
	2	**0.499**	0.0795	0.0397	0.2409	22.7%
	4	**0.097**	0.3640	0.0352	0.2522	12.8%
	8	**0.165**	0.3683	0.0609	0.2678	7.6%
	16	**0.089**	0.4571	0.0407	0.2711	4.3%

Table S3: **Isotropic covariance decay model**: The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square, with $\beta = 0.25$. The other simulation settings are as described in Table S1.

α_2	α_1	$\frac{\text{RMSE GAP}}{\text{RMSE PCA}}$	RMSE PCA	RMSE GAP	τ	Proportion of distances used
			$\beta = 2$			
0.5	1	1.009	0.0891	0.0899	0.2508	100.0%
	2	1.008	0.0940	0.0947	0.2617	100.0%
	4	**0.910**	0.1064	0.0968	0.2319	11.4%
	8	**0.657**	0.1241	0.0816	0.2358	5.8%
	16	**0.347**	0.1377	0.0478	0.2295	1.1%
1	1	1.020	0.0663	0.0677	0.2647	100.0%
	2	1.014	0.0859	0.0871	0.3076	100.0%
	4	**0.695**	0.1182	0.0821	0.2683	17.5%
	8	**0.202**	0.1725	0.0349	0.2554	5.9%
	16	**0.085**	0.2438	0.0207	0.2401	1.5%
1.5	1	1.030	0.0483	0.0498	0.2662	100.0%
	2	1.018	0.0757	0.0771	0.3356	100.0%
	4	**0.468**	0.1416	0.0663	0.2869	18.8%
	8	**0.092**	0.2205	0.0203	0.2513	5.7%
	16	**0.066**	0.2845	0.0187	0.2318	1.5%

Table S4: **Isotropic covariance decay model**: The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square, with $\beta = 2$. The other simulation settings are as described in Table S1.

α_2	α_1	$\frac{\text{RMSE GAP}}{\text{RMSE PCA}}$	RMSE PCA	RMSE GAP	τ	Proportion of distances used
			$\beta = 4$			
0.5	1	1.011	0.0775	0.0783	0.2448	100.0%
	2	1.009	0.0809	0.0816	0.2550	100.0%
	4	1.006	0.0922	0.0927	0.2723	100.0%
	8	**0.815**	0.0994	0.0810	0.2340	5.8%
	16	**0.467**	0.1131	0.0528	0.2247	1.1%
1	1	1.020	0.0540	0.0551	0.2441	100.0%
	2	1.017	0.0694	0.0706	0.2986	100.0%
	4	**0.958**	0.0940	0.0900	0.2759	18.8%
	8	**0.417**	0.1257	0.0524	0.2489	6.3%
	16	**0.105**	0.1798	0.0189	0.2561	3.7%
1.5	1	1.029	0.0369	0.0380	0.2419	100.0%
	2	1.025	0.0538	0.0552	0.3239	100.0%
	4	**0.962**	0.0956	0.0920	0.2848	18.0%
	8	**0.225**	0.1576	0.0355	0.2742	10.5%
	16	**0.108**	0.2030	0.0219	0.2786	6.3%

Table S5: **Isotropic covariance decay model**: The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square, with $\beta = 4$. The other simulation settings are as described in Table S1.

κ	α_1	$\beta = 1$ RMSE GAP / RMSE PCA	RMSE PCA	RMSE GAP	τ	Proportion of distances used
0.1	1	1.026	0.0680	0.0697	0.2383	100.0%
	2	1.021	0.0808	0.0825	0.2779	100.0%
	4	**0.644**	0.1105	0.0711	0.2561	23.5%
	8	**0.345**	0.1392	0.0480	0.2397	6.0%
	16	**0.257**	0.1597	0.0410	0.2575	5.9%
1	1	1.024	0.0660	0.0676	0.2367	100.0%
	2	1.032	0.0884	0.0912	0.2387	36.1%
	4	**0.655**	0.1062	0.0695	0.2617	23.3%
	8	**0.362**	0.1395	0.0505	0.2369	6.2%
	16	**0.167**	0.1552	0.0259	0.2621	5.8%
10	1	**0.512**	0.2962	0.1518	0.1817	23.6%
	2	**0.518**	0.2956	0.1532	0.2135	19.3%
	4	**0.345**	0.2980	0.1028	0.2186	8.9%
	8	**0.411**	0.3067	0.1260	0.2252	6.4%
	16	**0.356**	0.3142	0.1118	0.1762	0.6%

Table S6: **Directional covariance decay model**: Comparison of the localization accuracy of GAP and PCA for simulated datasets with $n = 2{,}000$ samples and $p = 50{,}000$ SNPs. The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square. The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square, with $\beta = 1$. In this table, $\beta = 1$, which is equivalent to sampling the individuals uniformly from the unit square. 100 different directions vectors $\mathbf{u}_k$ were chosen from a von Mises distribution with mean direction $(1, 0)^T$ (i.e. the x-axis) and dispersion parameter κ. For each such direction vector $\mathbf{u}_k$, 500 SNPs were simulated such that they have level sets of equal frequency along directions perpendicular to $\mathbf{u}_k$. The allele frequency for individual i at SNP ℓ for direction vector $\mathbf{u}_k$ is given by $q_\ell(\mathbf{z}_i) = 1/(1 + \exp(Z_{\ell,i}))$, where $Z_{\ell,\cdot}$ is an n-dimensional multivariate Gaussian random variable with mean $\mathbf{0}$ and covariance between the i-th and j-th entries given by $\exp(-(\alpha_1|\langle \mathbf{u}_k, \mathbf{z}_i - \mathbf{z}_j \rangle|)^{\alpha_2})/\alpha_0$. In these simulations, $\alpha_0 = \alpha_2 = 1$. Bold values indicate those parameter combinations where the root mean squared error (RMSE) of GAP is lower than that of PCA.

$\beta = 0.5$						
κ	α_1	$\frac{\text{RMSE GAP}}{\text{RMSE PCA}}$	RMSE PCA	RMSE GAP	τ	Proportion of distances used
0.1	1	1.024	0.0648	0.0664	0.2366	100.0%
	2	**0.836**	0.0710	0.0593	0.2340	36.2%
	4	**0.549**	0.0984	0.0540	0.2304	13.6%
	8	**0.340**	0.0896	0.0305	0.2320	6.3%
	16	**0.102**	0.2766	0.0281	0.2603	6.3%
1	1	1.021	0.0663	0.0677	0.2357	100.0%
	2	**0.885**	0.0780	0.0690	0.2437	40.8%
	4	**0.479**	0.0803	0.0384	0.2430	18.6%
	8	**0.255**	0.0854	0.0218	0.2320	6.4%
	16	**0.109**	0.2747	0.0299	0.2558	6.0%
10	1	**0.360**	0.3584	0.1291	0.1750	27.6%
	2	**0.272**	0.3571	0.0970	0.1927	17.4%
	4	**0.221**	0.3542	0.0783	0.1936	7.8%
	8	**0.278**	0.3579	0.0994	0.2049	6.3%
	16	**0.359**	0.3562	0.1279	0.1574	0.7%

Table S7: **Directional covariance decay model**: The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square, with $\beta = 0.5$. The other simulation settings are as described in Table S6.

$\beta = 0.25$						
κ	α_1	$\frac{\text{RMSE GAP}}{\text{RMSE PCA}}$	RMSE PCA	RMSE GAP	τ	Proportion of distances used
0.1	1	1.016	0.0433	0.0441	0.2384	100.0%
	2	**0.979**	0.0523	0.0512	0.1943	17.5%
	4	**0.559**	0.0511	0.0286	0.2194	14.4%
	8	**0.102**	0.2807	0.0286	0.2248	8.3%
	16	**0.115**	0.3723	0.0428	0.2618	8.7%
1	1	1.013	0.0425	0.0431	0.2401	100.0%
	2	**0.865**	0.0613	0.0531	0.1917	17.1%
	4	**0.553**	0.0740	0.0409	0.1964	9.9%
	8	**0.140**	0.2284	0.0321	0.2180	7.8%
	16	**0.117**	0.2678	0.0313	0.2596	8.3%
10	1	**0.303**	0.4099	0.1243	0.1441	18.7%
	2	**0.193**	0.4172	0.0806	0.1650	12.7%
	4	**0.318**	0.4083	0.1300	0.1447	4.9%
	8	**0.192**	0.4163	0.0798	0.2159	8.0%
	16	**0.313**	0.3580	0.1120	0.1921	2.3%

Table S8: **Directional covariance decay model**: The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square, with $\beta = 0.25$. The other simulation settings are as described in Table S6.

κ	α_1	$\frac{\text{RMSE GAP}}{\text{RMSE PCA}}$	RMSE PCA	RMSE GAP	τ	Proportion of distances used
			$\beta = 2$			
	1	1.025	0.0627	0.0643	0.2286	100.0%
	2	1.022	0.0780	0.0797	0.2732	100.0%
0.1	4	1.014	0.0969	0.0983	0.3022	100.0%
	8	**0.495**	0.1382	0.0684	0.2456	6.4%
	16	**0.222**	0.1574	0.0349	0.2571	6.0%
	1	1.025	0.0676	0.0693	0.2256	100.0%
	2	1.022	0.0750	0.0766	0.2735	100.0%
1	4	1.014	0.0968	0.0982	0.3118	100.0%
	8	**0.521**	0.1350	0.0703	0.2626	11.6%
	16	**0.166**	0.2412	0.0400	0.2519	5.9%
	1	**0.487**	0.2297	0.1119	0.1886	22.2%
	2	**0.533**	0.2298	0.1226	0.2354	24.8%
10	4	**0.474**	0.2372	0.1124	0.2542	15.7%
	8	**0.372**	0.2453	0.0913	0.2376	7.3%
	16	**0.295**	0.2585	0.0763	0.2022	1.0%

Table S9: **Directional covariance decay model**: The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square, with $\beta = 2$. The other simulation settings are as described in Table S6.

κ	α_1	$\frac{\text{RMSE GAP}}{\text{RMSE PCA}}$	RMSE PCA	RMSE GAP	τ	Proportion of distances used
			$\beta = 4$			
	1	1.024	0.0506	0.0518	0.2161	100.0%
	2	1.023	0.0625	0.0640	0.2642	100.0%
0.1	4	1.016	0.0765	0.0778	0.2991	100.0%
	8	**0.750**	0.1014	0.0761	0.2512	12.1%
	16	**0.363**	0.1285	0.0467	0.2577	6.9%
	1	1.023	0.0532	0.0544	0.2188	100.0%
	2	1.022	0.0603	0.0616	0.2646	100.0%
1	4	1.015	0.0776	0.0788	0.2983	100.0%
	8	**0.751**	0.1018	0.0764	0.2480	12.2%
	16	**0.254**	0.1879	0.0477	0.2551	7.2%
	1	**0.553**	0.1702	0.0941	0.1944	23.3%
	2	**0.534**	0.1743	0.0930	0.2408	25.8%
10	4	**0.540**	0.1757	0.0948	0.2721	23.2%
	8	**0.491**	0.1840	0.0903	0.2726	9.1%
	16	**0.404**	0.1978	0.0800	0.2298	2.1%

Table S10: **Directional covariance decay model**: The geographic locations $\mathbf{z}_i$ of the individuals are simulated by sampling each coordinate according to a Beta(β, β) distribution from the unit square, with $\beta = 4$. The other simulation settings are as described in Table S6.

1.6 Evaluation on the real datasets

Each of the three real datasets we analyzed (Human Origins, GLOBETROTTER, and POPRES) contained latitude and longitude coordinates at the subpopulation level. When applying GAP to each dataset, we used the genotype samples from a random subset of 20% of the subpopulations along with their true sampling locations in order to pick the local genetic distance threshold τ and to rescale the inferred locations to latitude-longitude coordinates. Recall from (11) that GAP infers locations up to translation and the unknown 2×2 invertible matrix J. Similarly, when applying PCA to each dataset, we used the same random subset of subpopulations to estimate the translation and rescaling of the principal components.[3]

In particular, let $\mathcal{D}$ denote the subpopulations in the full dataset and $\mathcal{T} \subset \mathcal{D}$ the random subset of training subpopulations, where $|\mathcal{D}| = m$ and $|\mathcal{T}| = m_0 = \lceil m/5 \rceil$. Let $S_j \subset \{1, \ldots, n\}$ be the set of individuals in subpopulation j in the sample, $1 \leq j \leq m$. Letting the inferred coordinates from GAP or PCA be denoted by $\hat{\mathbf{z}}_i$ for individual i, the mean inferred coordinates $\hat{\mathbf{y}}_j$ for each subpopulation j is computed by averaging the inferred coordinates over the individuals in the subpopulation,

$$\hat{\mathbf{y}}_j = \frac{1}{|S_j|} \sum_{i \in S_j} \hat{\mathbf{z}}_i \,.$$

If the true latitude-longitude sampling coordinates of population j is denoted by $\mathbf{y}_j$, we are interested in estimating the 2×2 coordinate rescaling matrix A^* and translation vector $\mathbf{b}^*$ which minimizes the following objective function measuring reconstruction error,

$$A^*, \mathbf{b}^* = \underset{\substack{A \in \mathbb{R}^{2\times 2} \\ \mathbf{b} \in \mathbb{R}^2}}{\arg\min} \sum_{j \in \mathcal{T}} \frac{|S_j|}{n} \|\mathbf{y}_j - (A\hat{\mathbf{y}}_j + \mathbf{b})\|^2 \,. \tag{16}$$

We can solve (16), for example, via the weighted least squares estimator for linear regression. To choose the value of the genetic distance parameter τ when applying GAP, we do a grid search over τ using the leave-one-out cross-validation error over the subpopulations in the training set of subpopulations $\mathcal{T}$. The RMSE of GAP and PCA on the full dataset is computed by transforming the inferred subpopulation coordinates using the estimated translation and rescaling in (16),

$$\text{RMSE} = \sqrt{\sum_{j \in \mathcal{D}} \frac{|S_j|}{n} \|\mathbf{y}_j - (A^*\hat{\mathbf{y}}_j + \mathbf{b}^*)\|^2} \,. \tag{17}$$

Since the reconstruction error in (17) depends on the training subset $\mathcal{T}$, we performed the above procedure with 100 randomly drawn subsets of training populations and report some statistics of the reconstruction RMSE in Table S11.

1.7 Application to the POPRES dataset

The POPRES dataset is an aggregation of 5,918 individuals with self-reported ancestry from several studies (Nelson et al. 2008; Preisig et al. 2009; Kooner et al. 2008). The dataset we analyzed contained individuals genotyped at 457,297 SNPs. We filtered SNPs deviating from Hardy-Weinberg equilibrium and thinned SNPs with linkage disequilibrium r^2 greater than 10% in sliding windows of 50 SNPs. This left us with a dataset of 77,678 SNPs. We selected 1,217 individuals from Europe

[3] From equation (1), we see that we should only have to transform the PC coordinates by a translation and a 2×2 rotation/reflection matrix. However, we provide PCA the same degrees of freedom in rescaling locations to latitude-longitude coordinates as we do for GAP.

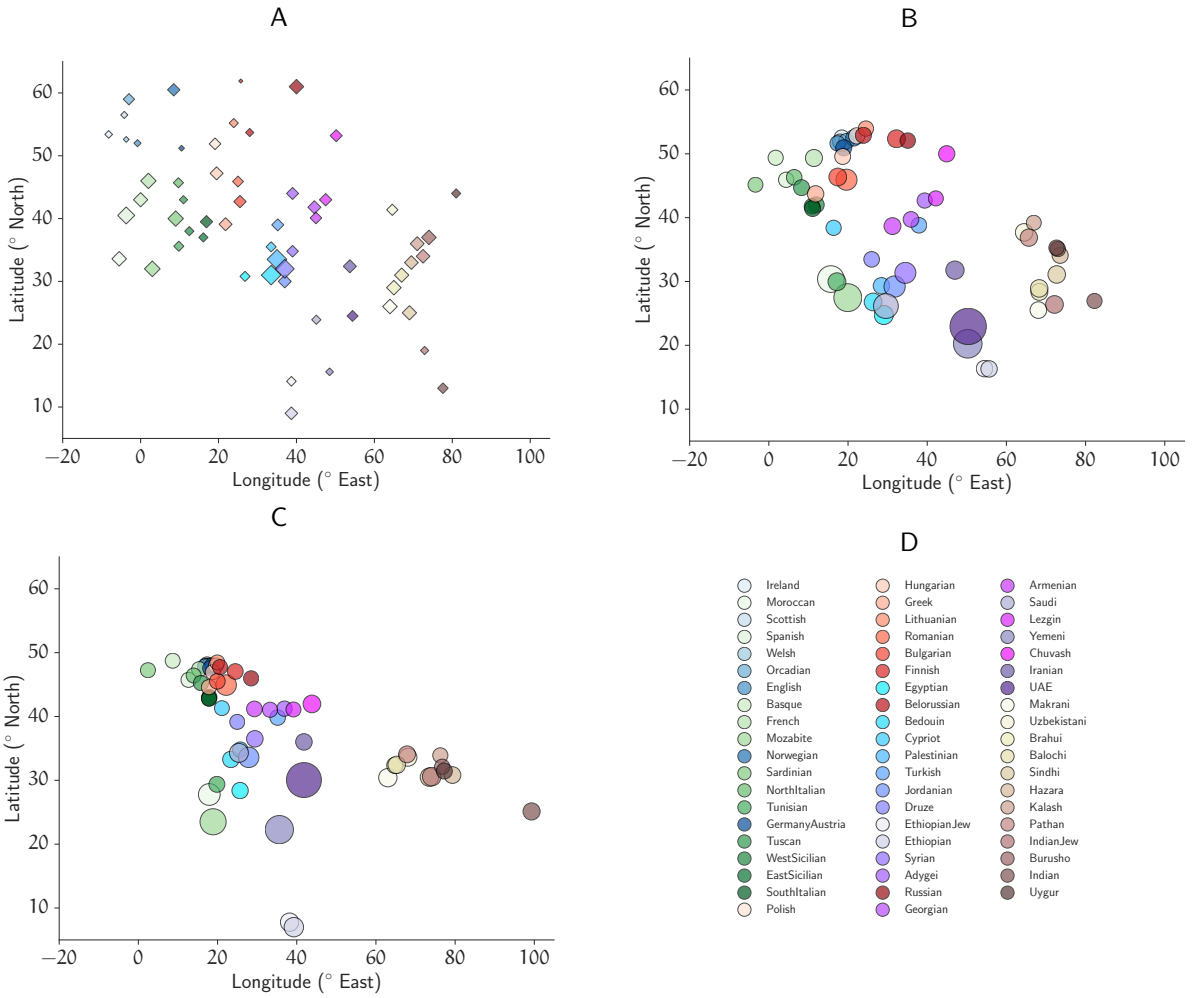


Fig. S3: PCA and GAP visualization of the populations in the GLOBETROTTER dataset. We analyzed 59 populations from Europe, the Middle East, North and East Africa, and Western, Central and South Asia. (A) True sampling locations, (B) GAP reconstructed locations, (C) PCA reconstructed locations, and (D) population legends. The diamonds are placed at the sampling locations for each population, while the circles are placed at the mean inferred location of the samples in each population. PCA tends to localize individuals from Southern Europe, North Africa, and the Middle East closer together, while GAP is better at separating them.

Dataset	RMSE GAP / RMSE PCA		RMSE PCA		RMSE GAP	
	Median	90% CI	Median	90% CI	Median	90% CI
Human Origins	0.69	(0.28, 0.89)	16.5°	(14.1°, 44.2°)	11.5°	(9.7°, 14.9°)
GLOBETROTTER	0.90	(0.65, 1.23)	15.2°	(13.0°, 26.1°)	13.5°	(11.5°, 25.8°)
POPRES	0.44	(0.20, 1.09)	18.0°	(8.6°, 41.4°)	7.5°	(5.3°, 24.4°)
POPRES (MAF $\geq$ 10%)	1.09	(0.60, 3.18)	5.2°	(3.6°, 16.2°)	6.0°	(4.1°, 18.2°)

Table S11: **Spatial assignment accuracy of PCA and GAP on the Human Origins, GLOBETROTTER, and POPRES datasets.** The reconstruction RMSE statistics are based on 100 random subsamples of 20% of the subpopulations in the full datasets. Using the known sampling locations of the subpopulations in the training sample, we rescaled the inferred coordinates from PCA and GAP in order to learn latitude-longitude coordinates for each subpopulation in the test set. For the POPRES dataset, we found that both PCA and GAP performed significantly better at spatial reconstruction if we used only the SNPs with minor allele frequency at least 10% (also see Figure S4).

who reported all four grandparents belonging to the same country in Europe and whose reported primary language matched the country of origin of their grandparents. This filtering was performed to avoid picking individuals that might be recently admixed, and is similar to the filters applied in previous works analyzing this dataset (Novembre et al. 2008; Yang et al. 2012). Using the true sampling locations of 20% of the subpopulations to assign spatial coordinates to a test set of subpopulations, GAP has median RMSE 7.47° while PCA has median RMSE of 17.97°, where the assignment was performed using 100 random training/test splits of the full dataset. We also analyzed the data after discarding SNPs with minor allele frequency below 10%. In this setting, the spatial assignment accuracy of GAP and PCA are quite similar, with median RMSE of 5.99° and 5.15°, respectively. The visualizations produced by PCA and GAP are very similar (Figure S4), and closely recapitulate the geography of Europe as has been previously observed by Novembre et al. (2008).

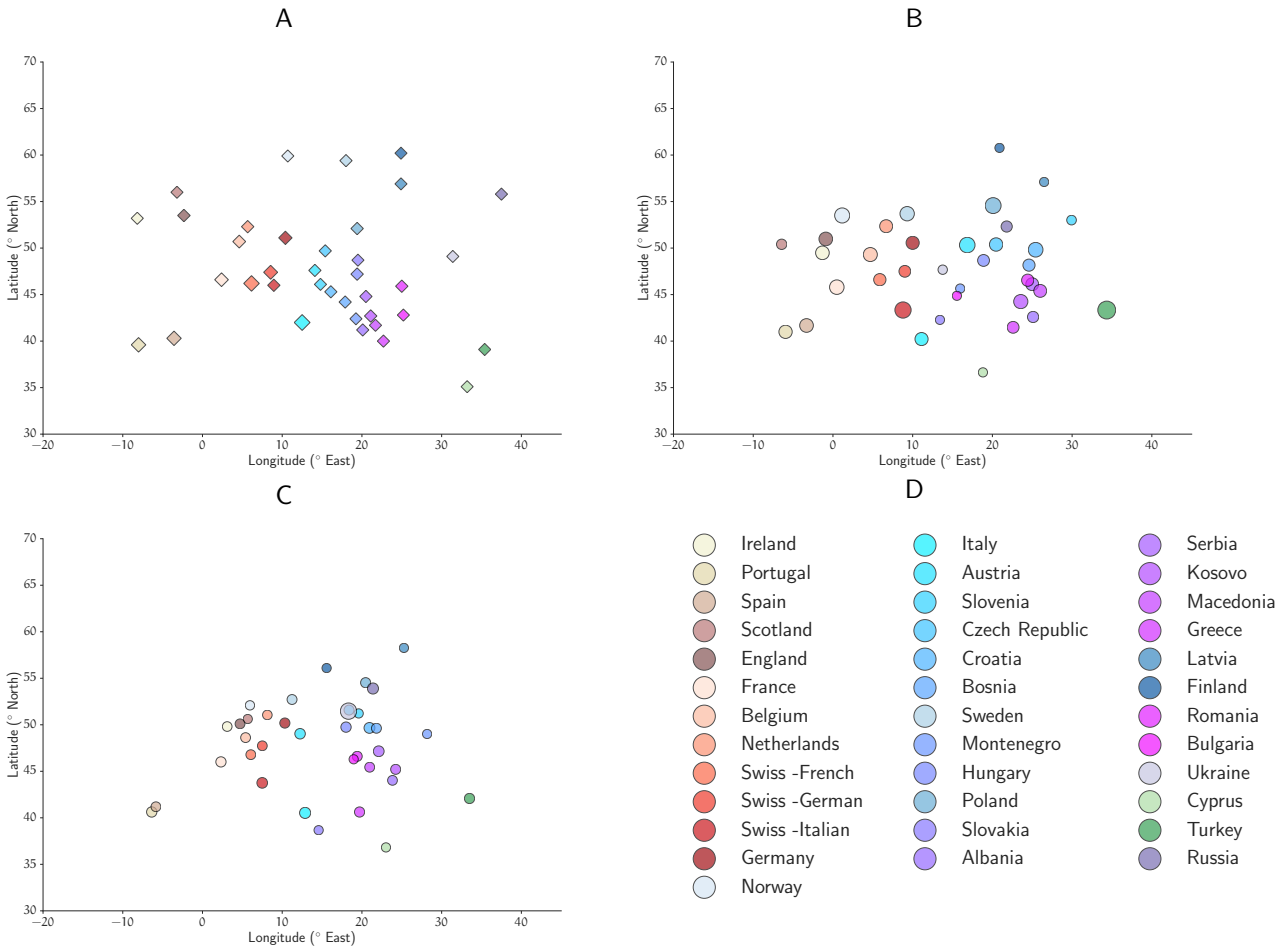


Fig. S4: **PCA and GAP visualization of the POPRES dataset.** (A) True sampling locations, (B) GAP reconstructed locations, (C) PCA reconstructed locations, and (D) population legends.

2 Association testing procedure

We present a statistical framework for testing associations between genotype and trait (either binary or quantitative) in the presence of population structure. Ancestry, on account of influencing genetic variation, can induce correlations between genotypes. Furthermore, ancestry is also correlated with the phenotype, due to varying trait prevalence with geography, ancestry-biased sampling, etc. In such cases, association tests can be confounded by population stratification, and the genotype and trait will be statistically dependent even when there is no genetic basis for the trait.

We correct for population stratification by conditioning the genotype on the confounding variable, which in our setting are the ancestry coordinates $\mathbf{z}_i$. As shown schematically in Song et al. (2015, Figure 1), this conditioning removes the statistical dependence between genotype $x_{i\ell}$ and confounder $\mathbf{z}_i$. To be more concrete, we consider a retrospective model to describe the effect of population structure and the trait on the genotype. For a particular SNP ℓ and an individual i with trait y_i, genotype $x_{i\ell}$ is generated according to the following model,

$$x_{i\ell} \mid y_i, \mathbf{z}_i \sim \mathsf{Binomial}(2, \theta_{i\ell})$$
$$\theta_{i\ell} = \frac{\kappa_\ell R_\ell^{y_i} q_\ell(\mathbf{z}_i)}{1 - q_\ell(\mathbf{z}_i) + \kappa_\ell R_\ell^{y_i} q_\ell(\mathbf{z}_i)}\,, \tag{18}$$

where R_ℓ is the genetic risk factor of the SNP ℓ on the trait, with the effect of population structure encoded in the allele frequency $q_\ell(\mathbf{z}_i)$. Note that for a binary trait $y_i \in \{0, 1\}$, by setting $\kappa_\ell = 1$ for all SNPs, the model reduces to the one studied by Price et al. (2006). Further, with the change of variables $a_\ell = \log \kappa_\ell$ and $b_\ell = \log R_\ell$, the model (18) can be rewritten as,

$$x_{i\ell} \mid y_i, \mathbf{z}_i \sim \mathsf{Binomial}(2, \theta_{i\ell})$$
$$\theta_{i\ell} = \mathrm{logit}^{-1}(a_\ell + b_\ell y_i + \mathrm{logit}(q_\ell(\mathbf{z}_i)))\,, \tag{19}$$

with $\mathrm{logit}(x) = \log(x/(1-x))$ for $0 < x < 1$. This inverse regression model has been put forth by Song et al. (2015) as a means of correcting for population structure and environmental confounders under fairly general assumptions. Under model (18), the genotypes $x_{i\ell}$ depend on the confounder $\mathbf{z}_i$ only through the allele frequency $q_\ell(\mathbf{z}_i)$. Hence, by conditioning on the allele frequencies $q_\ell(\mathbf{z}_i)$, the genotype becomes independent of ancestry location, $x_{i\ell} \perp\!\!\!\perp \mathbf{z}_i \mid q_\ell(\mathbf{z}_i)$. This conditioning also allows us to ignore the dependency between the trait y_i and ancestry $\mathbf{z}_i$ in our model. In practice, we do not know the underlying allele frequencies $q_\ell(\mathbf{z}_i)$, and must estimate it by making some assumptions which we describe shortly.

Given a particular trait of interest, a SNP ℓ is considered non-associated if $R_\ell = 1$, and considered associated otherwise. We are thus interested in testing between the following null and alternate hypotheses,

$$H_{0,\ell} : R_\ell = 1\,, \tag{20}$$
$$H_{A,\ell} : R_\ell \neq 1\,. \tag{21}$$

The retrospective model in (18) enjoys both computational and statistical advantages over the prospective model (main text, equations (4) and (5)). From a statistical point of view, (18) allows unbiased testing under fairly general assumptions about the phenotype distribution and about the ancestry and environmental confounding variables (see Theorem 1 of Song et al. (2015)). From a computational viewpoint, the association tests in (18) for different SNPs can be performed separately. This is in contrast to the prospective model where the trait is modeled as a linear combination of genotypes, environmental effects and random noise variation, and where a principled testing

procedure might require joint estimation over all SNPs. The retrospective model thus allows conceptually simpler, statistically valid and much more efficient and parallelizable association testing algorithms than procedures based on prospective models.

We next describe our testing procedure which consists of three steps: (1) estimation of ancestry coordinates $\hat{\mathbf{z}}_i$ using GAP; (2) estimation of spatial allele frequencies $q_\ell(\hat{\mathbf{z}}_i)$; and (3) estimation of the risk factor R_ℓ and intercept term κ_ℓ. Below, we elaborate on each of these steps.

Step (1): estimation of ancestry coordinates $\hat{\mathbf{z}}_i$. We use our localization algorithm GAP on the genotype matrix X as described in §1.2 to estimate the ancestry coordinates $\hat{\mathbf{z}}_i$ for each individual in the sample. The threshold parameter τ for local distances used in our algorithm GAP can be chosen using the strategies described in §1.3. As we will see in the next step, we do not need to rescale the ancestry coordinates by the dilation matrix J of (11).

Step (2): estimation of $q_\ell(\hat{\mathbf{z}}_i)$. We start with a naive estimate of the allele frequencies and then apply a kernel to smooth the allele frequencies over space. Using the genotype data, we compute the initial estimate $\xi_{i\ell}$ of the unknown allele frequency $q_\ell(\mathbf{z}_i)$ as,

$$\xi_{i\ell} = \frac{x_{il}}{2}. \tag{22}$$

We then refine these estimates by making the assumption that the allele frequency function $q_\ell(\mathbf{z})$ vary smoothly over space $\mathbf{z}$. Without such an additional assumption on the allele frequency function, the estimation problem is not well defined. We smooth the crude estimates of allele frequencies in (22) via an exponential kernel interpolation to get estimates $\hat{q}_{i\ell}$ for $q_\ell(\mathbf{z}_i)$ as follows.

$$\hat{q}_{i\ell} = \frac{\sum_{j=1}^n \xi_{j\ell} \exp(-\frac{1}{2}\|H^{-1}(\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_j)\|^2)}{\sum_{j=1}^n \exp(-\frac{1}{2}\|H^{-1}(\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_j)\|^2)}, \tag{23}$$

where H is a 2×2 bandwidth matrix. We use Scott's rule (Scott 1979) for selecting the bandwidth, where H is chosen so that $HH^T = n^{-1/3}\Sigma$ and Σ is the 2×2 covariance matrix of the estimated locations $\hat{\mathbf{z}}_i$. Finally, we threshold $\hat{q}_{i\ell}$ by one to ensure that they are valid probabilities. Note that with the choice of the bandwidth matrix in (23), we only need to estimate the ancestry coordinates $\hat{\mathbf{z}}$ up to the invertible 2×2 linear transformation J in (11).

Step (3): estimation of R_ℓ and κ_ℓ. Given an estimate of the allele frequencies $\hat{q}_{i\ell}$ from Step (2), we use Newton's method to estimate the risk factor R_ℓ (under the alternate hypothesis $H_{A,\ell}$) and intercept term κ_ℓ (under both the null and alternate hypotheses).

Given y_i and $x_{i\ell}$, the log-likelihood of the model (18), in terms of $\theta_{i\ell}$, is given by

$$\mathcal{L} = \sum_{i=1}^n x_{i\ell} \log \theta_{i\ell} + (2 - x_{i\ell}) \log(1 - \theta_{i\ell}). \tag{24}$$

Define the entries of the scaled gradient of the log-likelihood function, $F_1(R_\ell, \kappa_\ell) := R_\ell \partial\mathcal{L}/\partial R_\ell$, and $F_2(\kappa_\ell, \kappa_\ell) := \kappa_\ell \partial\mathcal{L}/\partial\kappa_\ell$.

After some algebraic manipulation we get,

$$F_1(R_\ell, \kappa_\ell) = \sum_{i=1}^n \Big(x_{i\ell} y_i - \frac{2 y_i q_{i\ell} \kappa_\ell R_\ell^{y_i}}{1 - q_{i\ell} + \kappa R_\ell^{y_i} q_{i\ell}} \Big), \tag{25}$$

$$F_2(R_\ell, \kappa_\ell) = \sum_{i=1}^n \Big(x_{i\ell} - \frac{2 \kappa_\ell R_\ell^{y_i} q_{i\ell}}{1 - q_{i\ell} + \kappa_\ell R_\ell^{y_i} q_{i\ell}} \Big). \tag{26}$$

Under the alternate hypothesis $H_{A,\ell}$, we obtain the maximum-likelihood estimate of R_ℓ by simultaneously solving $F_1(R_\ell, \kappa_\ell) = F_2(R_\ell, \kappa_\ell) = 0$ using Newton's method. The initial value we used for Newton's method is $R_\ell = \kappa_\ell = 1$ which corresponds to a non-associated SNP. We similarly compute the maximum likelihood estimate for κ_ℓ under the null hypothesis, and compute a p-value from the log-likelihood ratio using the χ^2_1 distribution.

2.1 Association test simulations

We simulated genotype data for $n = 2{,}000$ individuals at $p = 50{,}000$ SNPs sampled uniformly from the unit square using the isotropic and direction-dependent allele frequency covariance decay models described in §1.5. We used the same set of parameter combinations as we did for the ancestry localization simulations. We then generated quantitative phenotype data using the following linear model,

$$y_i = \alpha + \sum_{\ell=1}^{p} \beta_\ell x_{i\ell} + \lambda_i + \varepsilon_i, \tag{27}$$

where y_i is the phenotype and $x_{i\ell} \in \{0, 1, 2\}$ is the genotype at SNP ℓ for individual i, α is an intercept term, β_ℓ is the effect size of SNP ℓ, and λ_i and ε_i are the ancestral and random environmental/noise contributions respectively to the phenotype of individual i. We randomly selected 10 SNPs to be causal, with the effect sizes β_ℓ drawn from a standard normal distribution. The ancestry contribution λ_i to the phenotype was set to the first component (x-coordinate) of the sampling location $\mathbf{z}_i$, and the environmental/noise contribution was drawn from a standard normal distribution. We then rescaled the genotypic (i.e. $\sum_{\ell=1}^{p} \beta_\ell x_{i\ell}$), ancestry (i.e. λ_i), and environment/noise contributions (ε_i) so that they accounted for 20%, 10%, and 70% respectively of the variance of the phenotype in the sample.

To perform our association test, we first inferred the ancestry coordinates using GAP, where we used 20% of the data points as anchors with known sampling locations. To choose the threshold parameter τ for estimating local distances from the genetic distances, we swept over a range of τ which minimized the localization error over the subset of anchor data points. We then applied our association test described in the previous section with the inferred locations of all data points. Note that our association test only needs the ancestry coordinates up to the coordinate transformation matrix J given in (11), and hence this matrix does not need to be estimated.

We find that for parameter combinations where GAP has lower reconstruction RMSE in inferring ancestry coordinates compared to PCA, there is a concomitant increase in power to detect associations. Moreover, using the ancestry coordinates inferred from GAP in our association test performs almost as well as an oracle that has the true ancestry coordinates (Figures S5–S10 and Tables S12–S13).

α_2	α_1	SCGAP	PCA coordinates	True coordinates	GCAT ($d = 6$)
0.5	1	0.5937	0.5989	0.6031	0.5958
	2	0.6266	0.6268	0.6313	0.6233
	4	0.6556	0.6535	0.6535	0.6545
	8	0.5759	0.5650	0.5788	0.5574
	16	0.5867	0.5755	0.5840	0.5833
1	1	0.6139	0.6143	0.6205	0.6066
	2	0.6046	0.6045	0.6152	0.6080
	4	0.5881	0.5669	0.5929	0.5621
	8	0.6100	0.5802	0.6104	0.5568
	16	0.6345	0.6015	0.6354	0.5820
1.5	1	0.6345	0.6348	0.6369	0.6310
	2	0.5521	0.5474	0.5613	0.5517
	4	0.6122	0.5489	0.6117	0.5458
	8	0.5841	0.5483	0.5842	0.5250
	16	0.6043	0.5446	0.6072	0.5443

Table S12: **Association tests on data simulated under the isotropic covariance decay model.** Area under the ROC curve conditional on the FP rate being $\leq 10^{-3}$ for SCGAP and for our allele frequency estimation procedure applied to ancestry coordinates inferred by PCA (column 3) and to the true ancestry coordinates (column 4). For comparison, we also show the performance of GCAT using $d = 6$ latent factors used for estimating the allele frequencies. The genotype data were simulated according to the isotropic covariance decay model (see Figure S2A for an example) with the same parameter combinations as in Table S1. Each parameter combination row corresponds to 40 simulated datasets with $n = 2{,}000$ individuals and $p = 50{,}000$ SNPs, where 10 SNPs were chosen to have non-zero effects, with their effect sizes drawn from a standard normal distribution. The genotypic, ancestry, and environmental contribution to the phenotypic variance were set to 20%, 10% and 70% respectively.

κ	α_1	SCGAP	PCA coordinates	True coordinates	GCAT ($d = 6$)
0.1	1	0.6309	0.6311	0.6318	0.6329
	2	0.5968	0.5962	0.6044	0.5974
	4	0.5551	0.5477	0.5664	0.5372
	8	0.5867	0.5659	0.5863	0.5683
	16	0.5931	0.5676	0.5920	0.5569
1	1	0.5920	0.5922	0.5972	0.5995
	2	0.6150	0.6158	0.6263	0.6164
	4	0.5678	0.5565	0.5800	0.5471
	8	0.5970	0.5705	0.6015	0.5467
	16	0.5873	0.5679	0.5907	0.5514
10	1	0.6130	0.6097	0.6144	0.6131
	2	0.5677	0.5754	0.5732	0.5787
	4	0.5825	0.5687	0.5863	0.5780
	8	0.5858	0.5807	0.5918	0.5772
	16	0.5929	0.5807	0.5920	0.5791

Table S13: **Association tests on data simulated under the directional covariance decay model.** Area under the ROC curve conditional on the FP rate being $\leq 10^{-3}$ for SCGAP and for our allele frequency estimation procedure applied to ancestry coordinates inferred by PCA (column 3) and to the true ancestry coordinates (column 4). For comparison, we also show the performance of GCAT using $d = 6$ latent factors used for estimating the allele frequencies. The genotype data were simulated according to the directional covariance decay model (see Figure S2B for an example) with the same parameter combinations as in Table S6. Each parameter combination row corresponds to 40 simulated datasets with $n = 2{,}000$ individuals and $p = 50{,}000$ SNPs, where 10 SNPs were chosen to have non-zero effects, with their effect sizes drawn from a standard normal distribution. The genotypic, ancestry, and environmental contribution to the phenotypic variance were set to 20%, 10% and 70% respectively.

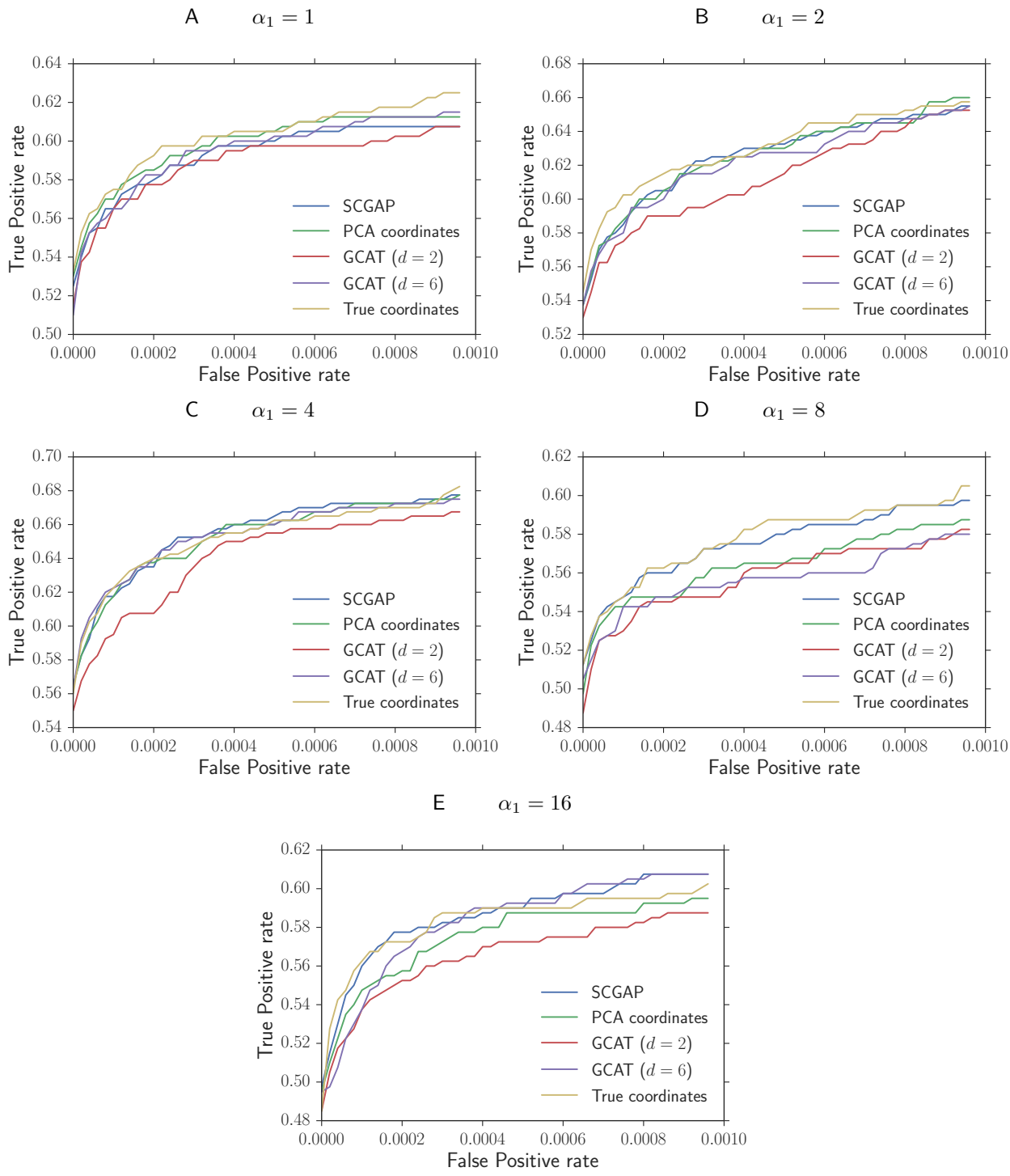


Fig. S5: **Isotropic covariance decay.** ROC curves for our association testing procedure with ancestral locations inferred using GAP, PCA, or using the true locations. We also compared our results with the GCAT method, which uses a latent factor model with d factors to estimate the allele frequencies for each individual at each locus. Genotypes were drawn according to the isotropic covariance decay model with $\alpha_0 = 1$ and $\alpha_2 = 0.5$, with the different choices of α_1 given in the panel captions. These are the same simulation parameters used in Table S1.

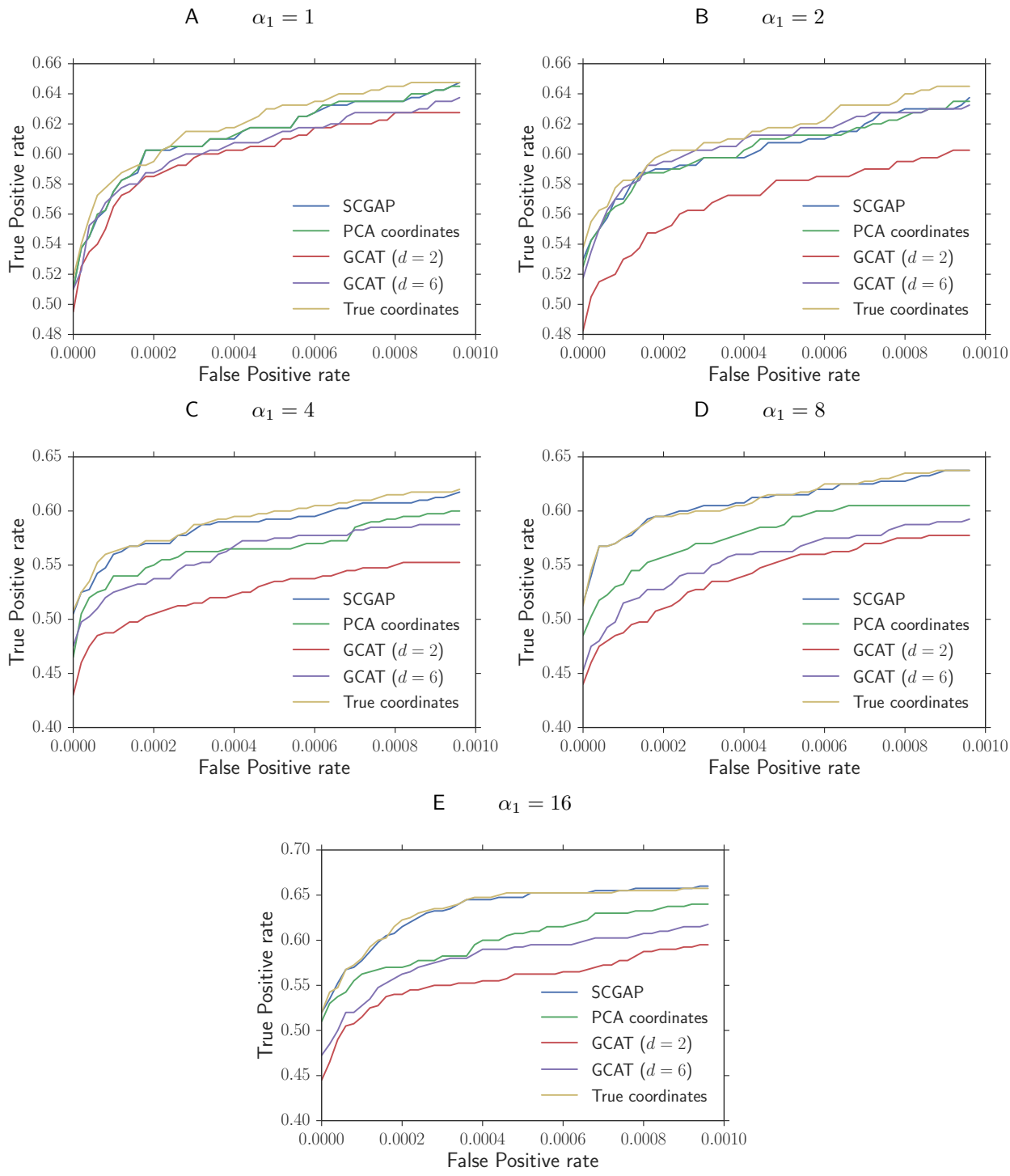


Fig. S6: Same simulation scenario as in Figure S5, with $\alpha_2 = 1$.

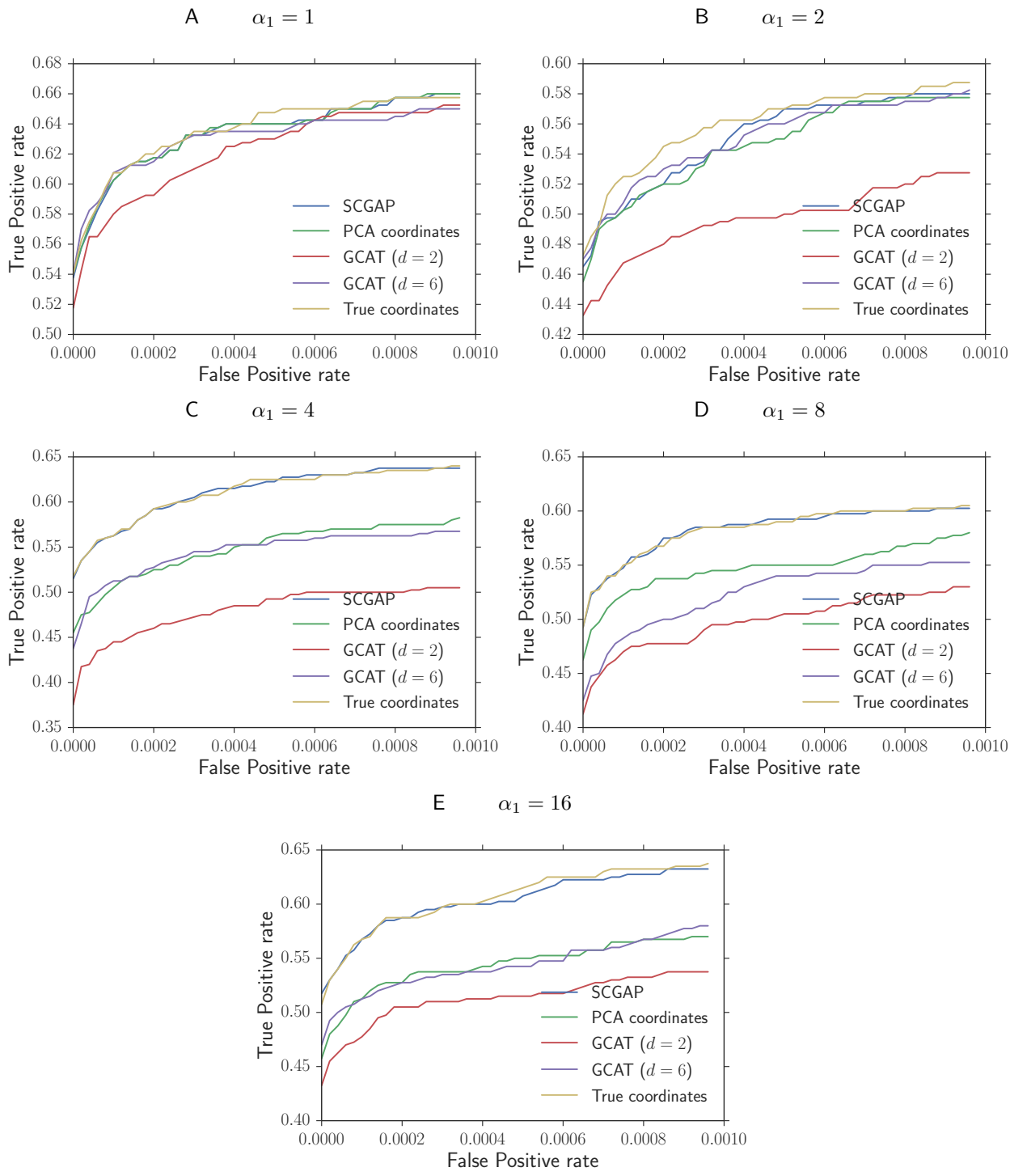


Fig. S7: Same simulation scenario as in Figure S5, with $\alpha_2 = 1.5$.

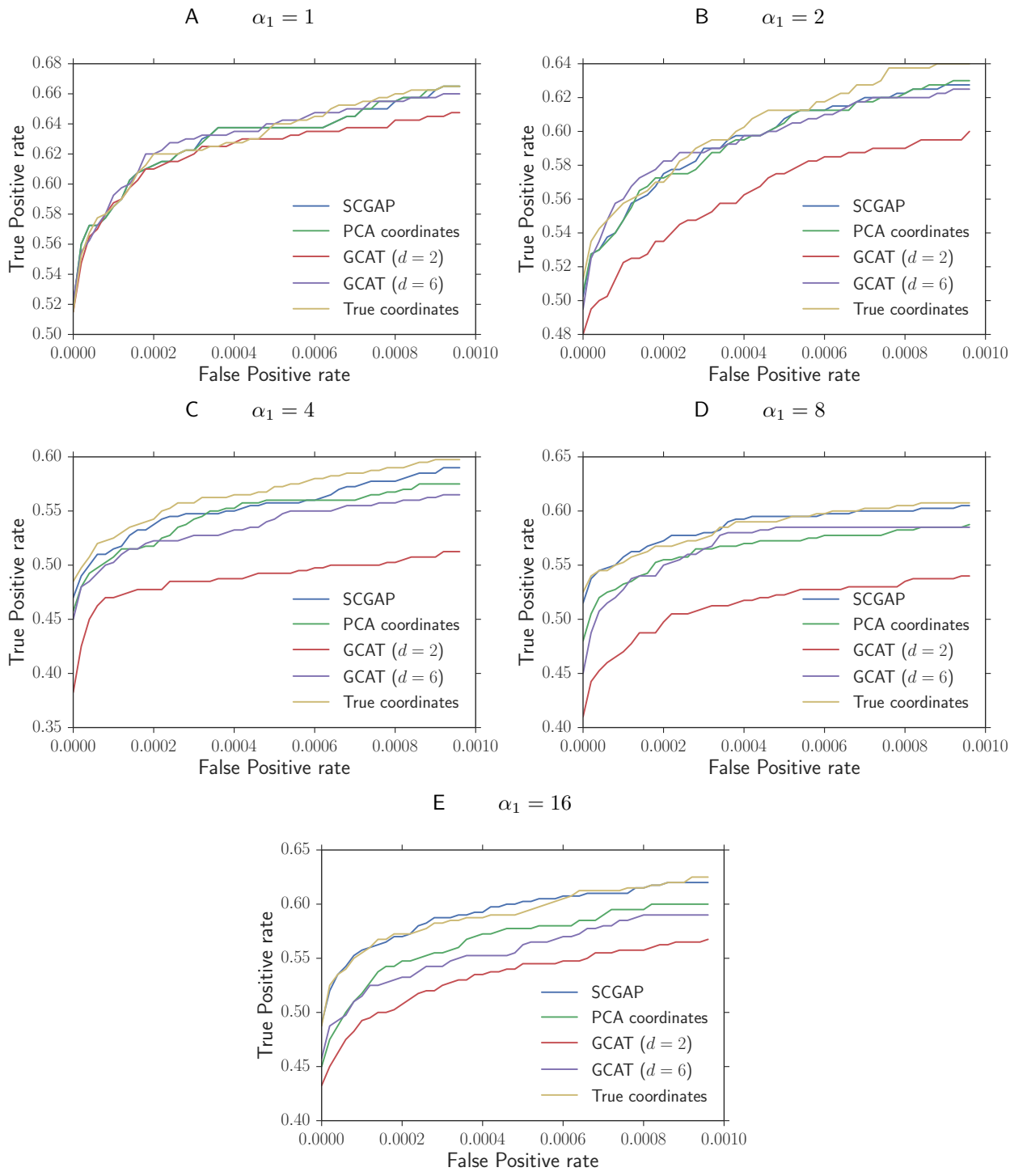


Fig. S8: **Directional covariance decay.** ROC curves for our association testing procedure with ancestral locations inferred using GAP, PCA, or using the true locations. We also compared our results with the GCAT method, which uses a latent factor model with d factors to estimate the allele frequencies for each individual at each locus. Genotypes were drawn according to the directional covariance decay model with $\alpha_0 = \alpha_2 = 1$ and $\kappa = 0.1$, with the different choices of α_1 given in the panel captions. These are the same simulation parameters used in Table S6.

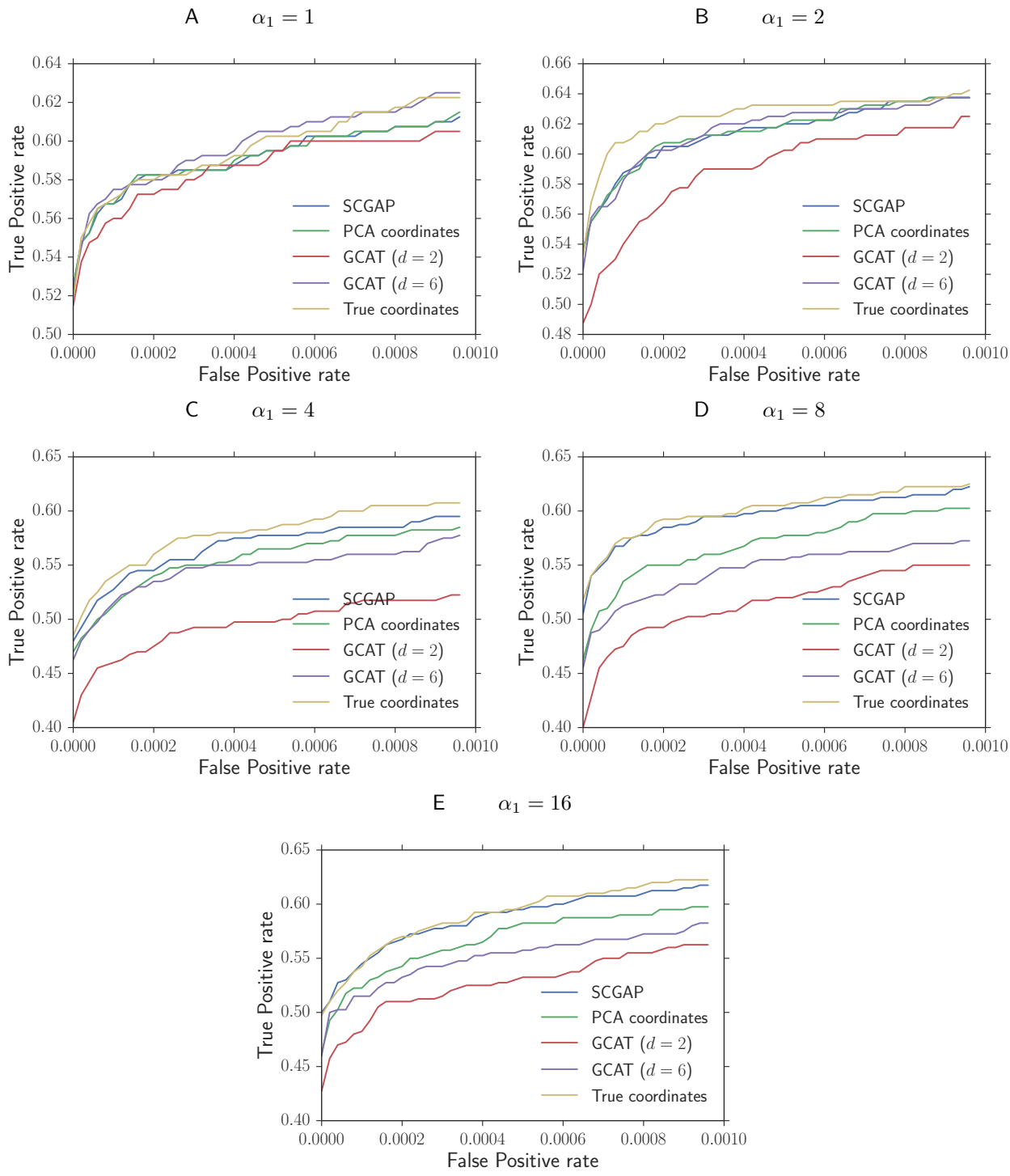


Fig. S9: Same simulation scenario as in Figure S8, with $\kappa = 1$.

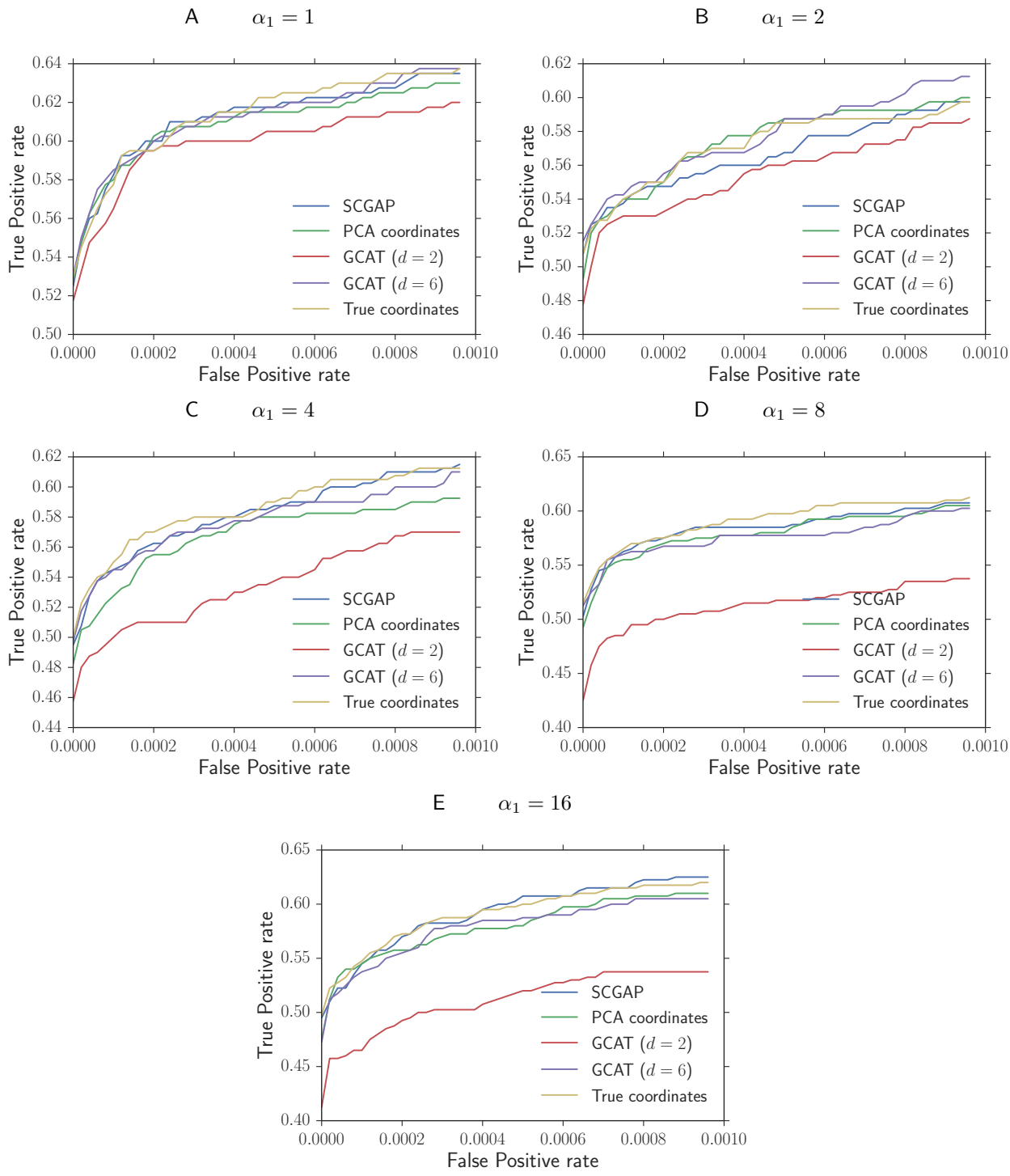


Fig. S10: Same simulation scenario as in Figure S8, with $\kappa = 10$.

2.2 NFBC dataset

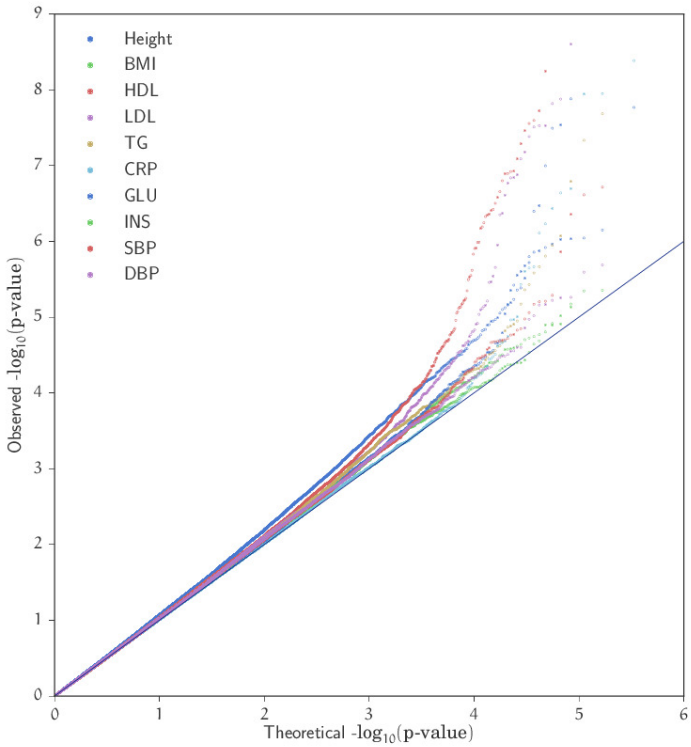
The original dataset contained 364,590 SNPs from 5,402 individuals. After filtering individuals and SNPs using the same criteria for missing genotypes and deviation from Hardy-Weinberg equilibrium as described in Song et al. (2015), we were left with 335,143 SNPs and 5,246 individuals. We added features for known confounders such as sex, oral contraceptive use, pregnancy status, and fasting status according to the procedure described in the first analysis of this dataset by Sabatti et al. (2009). We performed a Box-Cox transform on the median 95% of trait values to make the distribution of traits as close to a normal distribution as possible.

We applied our localization algorithm GAP on the genotype data to estimate two spatial ancestry coordinates for each individual. Since we did not have the ancestral or birth locations of the individuals in the sample, we could not optimize the threshold τ for estimating local spatial distances from the genetic distances as we had done in the simulations. Instead, we picked the threshold τ as described in §1.3.

Trait	Abbreviation	SCGAP+GC	GCAT+GC	LMM+GC	PCA+GC	Uncorrected+GC
Height	Height	1	1	0	0	0
Body mass index	BMI	0	0	0	0	0
HDL cholesterol levels	HDL	5	4	4	2	4
LDL cholesterol levels	LDL	4	4	3	3	3
Triglyceride levels	TG	2*	2	3	2	2
C-reactive protein	CRP	2*	2*	2	2	2
Glucose levels	GLU	3	3	2	2	2
Insulin levels	INS	0	0	0	0	0
Diastolic blood pressure	DBP	0	0	0	0	0
Systolic blood pressure	SBP	0	0	0	0	0

Table S14: **Number of significant loci discovered by SCGAP and several other association-testing approaches on the Northern Finland Birth Cohorts (NFBC) dataset.** The log-likelihood ratios from each method were corrected using genomic control, denoted by "+GC".The genome-wide significance level was set to 7.2×10^{-8}.
*Result when the trait was not transformed using the Box-Cox transformation. Under the transformation, one locus was significant.

A



B

Phenotype	GCIF
Height	1.0794
BMI	1.0521
HDL	1.0527
LDL	1.0515
TG	0.9983
CRP	1.0614
GLU	1.0049
INS	1.0462
SBP	1.0312
DBP	1.0485

Fig. S11: (A) Plot of the observed versus expected log p-values of SCGAP for the 10 metabolic traits in the Northern Finland Birth Cohorts (NFBC) dataset. See Table S15 for the most significant p-values for each trait. (B) Genomic control inflaction factor estimated for each of the 10 quantitative traits in the NFBC dataset using SNPs spaced 250kb apart.

Height

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs2814982	6	34654538	1.697e-08	7.101e-08
rs6719545	2	218160079	7.032e-07	2.147e-06
rs2815005	6	34746825	8.971e-07	2.684e-06
rs2744972	6	34767032	9.203e-07	2.748e-06
rs2814983	6	34699185	9.344e-07	2.786e-06
rs2814993	6	34726871	1.078e-06	3.175e-06
rs2814985	6	34656274	1.197e-06	3.496e-06
rs4911494	20	33435328	1.324e-06	3.833e-06
rs6088813	20	33438595	1.380e-06	3.981e-06
rs6058154	20	33049495	1.892e-06	5.316e-06

Body mass index (BMI)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs5957365	23	119412593	4.390e-06	6.394e-06
rs6567030	18	54679876	4.476e-06	6.515e-06
rs10247383	7	63536549	6.667e-06	9.579e-06
rs1001729	6	2540477	1.223e-05	1.722e-05
rs12658762	5	18615363	1.263e-05	1.777e-05
rs6638764	23	8171611	1.558e-05	2.177e-05
rs2408165	16	61259627	1.691e-05	2.356e-05
rs4953198	2	45248172	1.792e-05	2.493e-05
rs17207196	7	74939001	1.876e-05	2.606e-05
rs8050136	16	52373776	1.971e-05	2.733e-05

HDL cholestorol levels (HDL)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs1532624	16	55562980	0.000e+00	0.000e+00
rs7499892	16	55564091	1.110e-16	2.220e-16
rs1532085	15	56470658	2.351e-12	5.686e-12
rs9989419	16	55542640	8.050e-11	1.723e-10
rs1800961	20	42475778	5.643e-09	1.044e-08
rs7120118	11	47242866	1.882e-08	3.341e-08
rs2167079	11	47226831	2.525e-08	4.436e-08
rs255049	16	66570972	2.759e-08	4.833e-08
rs415799	15	56478046	3.438e-08	5.977e-08
rs255052	16	66582496	5.116e-08	8.774e-08

LDL cholestorol levels (LDL)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs646776	1	109620053	1.419e-12	4.361e-12
rs693	2	21085700	2.010e-11	5.512e-11
rs754524	2	21165046	2.477e-09	5.525e-09
rs6754295	2	21059688	1.317e-08	2.734e-08
rs6728178	2	21047434	1.509e-08	3.115e-08
rs207150	1	55579053	2.946e-08	5.910e-08
rs11668477	19	11056030	2.948e-08	5.913e-08
rs3923037	2	21011755	3.078e-08	6.164e-08
rs4844614	1	205941798	4.107e-08	8.123e-08
rs754523	2	21165196	6.584e-08	1.276e-07

Triglyceride levels (TG) (untransformed)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs1260326	2	27584444	2.053e-08	2.318e-08
rs10096633	8	19875201	4.597e-08	5.160e-08
rs780094	2	27594741	1.608e-07	1.788e-07
rs5939593	23	92156237	8.433e-07	9.268e-07
rs673548	2	21091049	1.125e-06	1.233e-06
rs676210	2	21085029	1.551e-06	1.697e-06
rs6728178	2	21047434	2.140e-06	2.337e-06
rs3923037	2	21011755	2.592e-06	2.825e-06
rs4825619	23	117987171	3.688e-06	4.010e-06
rs6754295	2	21059688	4.327e-06	4.699e-06

C-reactive protein (CRP) (untransformed)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs2794520	1	157945440	4.086e-09	5.607e-09
rs1169300	12	119915608	1.115e-08	1.503e-08
rs2464196	12	119919810	1.131e-08	1.525e-08
rs2592887	1	157919563	1.998e-07	2.563e-07
rs2650000	12	119873345	2.277e-07	2.915e-07
rs2098930	3	153371624	3.664e-07	4.652e-07
rs6624381	23	68197466	5.836e-07	7.349e-07
rs735396	12	119923227	7.723e-07	9.679e-07
rs10035541	5	7592712	2.147e-06	2.644e-06
rs12093699	1	157914612	2.236e-06	2.752e-06

Table S15: The top 10 most significant SNPs found by SCGAP for each of the 10 traits in the Northern Finland Birth Cohort dataset. We report the p-values both before and after genomic control adjustment.

Glucose levels (GLU)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs560887	2	169471394	3.279e-11	5.947e-11
rs3847554	11	92308474	1.575e-10	2.742e-10
rs2971671	7	44177862	3.367e-10	5.749e-10
rs2908290	7	44182662	1.306e-08	2.029e-08
rs1387153	11	92313476	2.882e-08	4.387e-08
rs563694	2	169482317	3.188e-08	4.840e-08
rs2166706	11	92331180	9.990e-08	1.473e-07
rs1447352	11	92362409	3.354e-07	4.793e-07
rs7121092	11	92363999	3.982e-07	5.666e-07
rs758989	7	44169531	1.584e-06	2.176e-06

Insulin levels (INS)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs521184	8	41720842	7.293e-06	9.437e-06
rs5985850	23	28409093	9.537e-06	1.226e-05
rs5943445	23	28411095	1.205e-05	1.541e-05
rs6502762	17	3819013	1.257e-05	1.606e-05
rs7241379	18	64306982	2.042e-05	2.579e-05
rs998223	2	64824633	2.260e-05	2.847e-05
rs6126645	20	50745422	2.895e-05	3.626e-05
rs6526679	23	27468439	3.672e-05	4.574e-05
rs932052	12	62081496	3.968e-05	4.934e-05
rs2037206	18	64323734	3.974e-05	4.942e-05

Triglyceride levels (TG)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs1260326	2	27584444	4.484e-09	5.903e-09
rs780094	2	27594741	6.890e-08	8.704e-08
rs10096633	8	19875201	1.174e-07	1.471e-07
rs673548	2	21091049	6.148e-07	7.515e-07
rs676210	2	21085029	7.149e-07	8.719e-07
rs6728178	2	21047434	9.045e-07	1.099e-06
rs2304130	19	19650528	9.208e-07	1.119e-06
rs6754295	2	21059688	1.726e-06	2.077e-06
rs3923037	2	21011755	6.887e-06	8.118e-06
rs12805061	11	116058235	7.972e-06	9.377e-06

C-reactive protein (CRP)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs2794520	1	157945440	6.284e-14	8.360e-14
rs12093699	1	157914612	8.348e-12	1.058e-11
rs2592887	1	157919563	1.260e-08	1.487e-08
rs1811472	1	157908973	4.596e-08	5.358e-08
rs402681	4	104634397	9.352e-07	1.059e-06
rs7694802	4	104621696	1.973e-06	2.219e-06
rs7178765	15	23672266	4.585e-06	5.114e-06
rs10107791	8	101040128	5.874e-06	6.535e-06
rs340468	4	104637688	7.626e-06	8.464e-06
rs6701469	1	199265442	9.789e-06	1.084e-05

Systolic blood pressure (SBP)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs782588	2	55695144	1.916e-07	2.523e-07
rs782586	2	55689669	2.414e-07	3.164e-07
rs782602	2	55702813	4.374e-07	5.669e-07
rs2627759	2	55706845	1.363e-06	1.729e-06
rs10496050	2	55659817	5.122e-06	6.337e-06
rs1754154	1	43243353	6.118e-06	7.544e-06
rs9656787	8	104301120	6.223e-06	7.671e-06
rs782606	2	55740106	6.696e-06	8.243e-06
rs782652	2	55716279	8.643e-06	1.059e-05
rs782637	2	55747751	1.049e-05	1.281e-05

Diastolic blood pressure (DBP)

RSID	Chr	Pos	SCGAP	SCGAP+GC
rs5928929	23	35549454	2.031e-06	2.894e-06
rs6942973	7	3134277	2.528e-06	3.580e-06
rs472594	1	226668261	5.430e-06	7.521e-06
rs5927821	23	31656866	5.510e-06	7.629e-06
rs952061	12	100502356	5.948e-06	8.217e-06
rs1079199	11	6384682	6.771e-06	9.320e-06
rs2094147	23	33025497	6.873e-06	9.455e-06
rs7783562	7	106704674	9.493e-06	1.294e-05
rs808127	23	8458979	1.057e-05	1.436e-05
rs4548444	1	204956761	1.381e-05	1.861e-05

Table S15 continued

Phenotype	RSID	Chr	Pos	Nearest gene	SCGAP+GC	Replication Study
Height	rs2814982	6	34654538	C6orf106	7.101e-08	Weedon et al. (2008)*
HDL	rs1532624	16	55562980	CETP	0.000e+00	Aulchenko et al. (2009)
HDL	rs1532085	15	56470658	LIPC	5.686e-12	Willer et al. (2013)
HDL	rs1800961	20	42475778	HNF4A	1.044e-08	Willer et al. (2013)
HDL	rs7120118	11	47242866	NR1H3	3.341e-08	—
HDL	rs255049	16	66570972	LCAT	4.833e-08	Willer et al. (2013)
TG	rs1260326	2	27584444	GCKR	2.318e-08	Willer et al. (2013)
TG	rs10096633	8	19875201	LPL	5.160e-08	Aulchenko et al. (2009)
LDL	rs646776	1	109620053	CELSR2	4.361e-12	Aulchenko et al. (2009)
LDL	rs693	2	21085700	APOB	5.512e-11	Aulchenko et al. (2009)
LDL	rs207150	1	55579053	USP24	5.910e-08	—
LDL	rs11668477	19	11056030	LDLR	5.913e-08	Willer et al. (2013)
CRP	rs2794520	1	157945440	CRP	5.607e-09	Dehghan et al. (2011)
CRP	rs1169300	12	119915608	HNF1A	1.503e-08	Dehghan et al. (2011)
GLU	rs560887	2	169471394	G6PC2	5.947e-11	Dupuis et al. (2010)
GLU	rs3847554	11	92308474	MTNR1B	2.742e-10	Dupuis et al. (2010)
GLU	rs2971671	7	44177862	GCK	5.749e-10	Dupuis et al. (2010)

* SNP rs2814993 is reported to be associated with height, and is 72kb from rs2814982 with LD $r^2 = 0.56$ in the 1000 Genomes CEU samples (see also Song et al. (2015)).

Table S16: Most significantly associated SNPs at each locus that were detected by SCGAP with genomic control, and the replication studies on different datasets which have also reported these associations.

3 Proofs

3.1 Proof of Theorem 1

We use the shorthand $q_{i\ell} := q_\ell(\mathbf{z}_i)$ to lighten the notation. By applying the triangle inequality, for $1 \le i < j \le n$,

$$\begin{aligned}|\hat{\eta}_{i,j} - \eta(\mathbf{z}_i - \mathbf{z}_j)| \le &\left|\frac{1}{p}\sum_{\ell=1}^{p}\left(\frac{x_{i\ell}}{2} - \mu_\ell\right)\left(\frac{x_{j\ell}}{2} - \mu_\ell\right) - \mathbb{E}[(q_{i\ell} - \mu_\ell)(q_{j\ell} - \mu_\ell)]\right| \\ &+ \frac{1}{2p}\sum_{\ell=1}^{p}(x_{i\ell} + x_{j\ell})\,|\hat{\mu}_\ell - \mu_\ell| \\ &+ \frac{1}{p}\sum_{\ell=1}^{p}(\hat{\mu}_\ell + \mu_\ell)|\hat{\mu}_\ell - \mu_\ell|\,.\end{aligned} \tag{28}$$

Define the following events $\mathcal{A}_{i,j}$ for $1 \le i < j \le n$,

$$\mathcal{A}_{i,j}: \quad \left|\frac{1}{p}\sum_{\ell=1}^{p}\left(\frac{x_{i\ell}}{2} - \mu_\ell\right)\left(\frac{x_{j\ell}}{2} - \mu_\ell\right) - \mathbb{E}[(q_{i\ell} - \mu_\ell)(q_{j\ell} - \mu_\ell)]\right| < t \tag{29}$$

We note that conditional on $q_{i\ell}$ and $q_{j\ell}$, the genotypes $x_{i\ell}$ and $x_{j\ell}$ are independent. Therefore,

$$\mathbb{E}\Big[\left(\frac{x_{i\ell}}{2} - \mu_\ell\right)\left(\frac{x_{j\ell}}{2} - \mu_\ell\right)\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\left(\frac{x_{i\ell}}{2} - \mu_\ell\right)\left(\frac{x_{j\ell}}{2} - \mu_\ell\right)\Big|q_{i\ell}, q_{j\ell}\Big]\Big] = \mathbb{E}[(q_{i\ell} - \mu_\ell)(q_{j\ell} - \mu_\ell)]\,.$$

Further, the genotypes $x_{i\ell}$ are independent for different SNPs ℓ. By applying Hoeffding's inequality, we obtain $\mathbb{P}(\mathcal{A}_{i,j}^c) \le 2e^{-2pt^2}$ for any fixed pair i and j.

To bound the second term in (28), we write,

$$\frac{1}{2p}\sum_{\ell=1}^{p}(x_{i\ell} + x_{j\ell})|\hat{\mu}_\ell - \mu_\ell| \le \frac{2}{p}\sum_{\ell=1}^{p}|\hat{\mu}_\ell - \mu_\ell| \tag{30}$$

Note that the summands $|\mu_\ell - \hat{\mu}_\ell|$ are independent. We define the event $\mathcal{E}$ as,

$$\mathcal{E}: \quad \frac{1}{p}\sum_{\ell=1}^{p}|\mu_\ell - \hat{\mu}_\ell| - \frac{1}{p}\sum_{\ell=1}^{p}\mathbb{E}(|\mu_\ell - \hat{\mu}_\ell|) < t\,. \tag{31}$$

By applying Hoeffding's inequality, we obtain $\mathbb{P}(\mathcal{E}^c) \le e^{-2pt^2}$.

We next bound $\mathbb{E}(|\mu_\ell - \hat{\mu}_\ell|)$. For each ℓ, $1 \le \ell \le p$, define the events $\mathcal{B}_\ell$ and $\mathcal{C}_\ell$ as follows,

$$\mathcal{B}_\ell: \quad \left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{i\ell}}{2} - q_{i\ell}\right)\right| < t$$

$$\mathcal{C}_\ell: \quad \left|\frac{1}{n}\sum_{i=1}^{n}q_{i\ell} - \mu_\ell\right| < t\,.$$

Note that conditional on the allele frequencies $q_{i\ell}$, the genotypes $x_{i\ell}$ are independent across index i. We can apply Hoeffding's inequality again to get $\mathbb{P}(\mathcal{B}_\ell^c \mid \{q_{i\ell}\}_{1\le i\le n}) \le 2e^{-2nt^2}$. Hence, $\mathbb{P}(\mathcal{B}_\ell^c) \le 2e^{-2nt^2}$ as well. Bounding the probability of $\mathcal{C}_\ell$ requires more work since the summands $q_{i\ell}$ are

dependent. We construct $K \in \mathbb{R}^{n\times n}$ with $K_{ij} = \eta(\mathbf{z}_i - \mathbf{z}_j)$. Let $\mathbf{1} = (1/\sqrt{n}, \ldots, 1/\sqrt{n})^\mathsf{T}$, $\mathbf{q}_\ell^{\text{cen}} = (q_{1\ell} - \mu_\ell, q_{2\ell} - \mu_\ell, \ldots, q_{n\ell} - \mu_\ell)^\mathsf{T}$ and set $\tilde{\mathbf{q}}_\ell^{\text{cen}} = K^{-1/2}\mathbf{q}_\ell^{\text{cen}}$. We write

$$\frac{1}{n}\sum_{i=1}^{n}(q_{i\ell} - \mu_\ell) = \frac{1}{\sqrt{n}}\mathbf{1}^\mathsf{T}K^{1/2}\tilde{\mathbf{q}}_\ell^{\text{cen}}\,. \tag{32}$$

Since the coordinates of $\tilde{\mathbf{q}}_\ell^{\text{cen}}$ are mean zero and are uncorrelated, using Chebyshev's inequality and (32), we get

$$\mathbb{P}(\mathcal{C}_\ell^c) \le \mathbb{P}(|\mathbf{1}^\mathsf{T}K^{1/2}\tilde{\mathbf{q}}_\ell^{\text{cen}}| > t\sqrt{n}) \le \frac{1}{nt^2}(\mathbf{1}^\mathsf{T}K\mathbf{1})\,.$$

We are now ready to bound $\mathbb{E}(|\mu_\ell - \hat{\mu}_\ell|)$. Using the notation $\kappa := \mathbf{1}^\mathsf{T}K\mathbf{1}$, we have

$$\mathbb{P}\Big(|\hat{\mu}_\ell - \mu_\ell| > 2t\Big) \le \mathbb{P}(\mathcal{B}_\ell^c) + \mathbb{P}(\mathcal{C}_\ell^c) \le 2e^{-2nt^2} + \frac{1}{nt^2}\kappa\,, \tag{33}$$

where the first inequality follows from triangle inequality. Hence,

$$\begin{aligned}
\mathbb{E}(|\mu_\ell - \hat{\mu}_\ell|) &= \int_0^\infty \mathbb{P}(|\mu_\ell - \hat{\mu}_\ell| > t) \\
&\le s + \int_s^\infty \left(2e^{-nt^2/2} + \frac{4\kappa}{nt^2}\right)\mathrm{d}t \\
&\le s + \frac{2}{n} + \frac{4\kappa}{sn}\,,
\end{aligned}$$

for any $s > 0$. Choosing $s = 2\sqrt{\kappa/n}$, we arrive at

$$\mathbb{E}(|\mu_\ell - \hat{\mu}_\ell|) \le \frac{2}{n} + 4\sqrt{\frac{\kappa}{n}}\,. \tag{34}$$

Using equations (30), (34) and recalling definition (31) we conclude that on event $\mathcal{E}$ the following is true,

$$\frac{1}{2p}\sum_{\ell=1}^{p}(x_{i\ell} + x_{j\ell})|\hat{\mu}_\ell - \mu_\ell| < 2t + \frac{4}{n} + 8\sqrt{\frac{\kappa}{n}}\,. \tag{35}$$

By a similar argument, on event $\mathcal{E}$ we have the following bound on the third term in equation (28):

$$\frac{1}{p}\sum_{\ell=1}^{p}(\hat{\mu}_\ell + \mu_{j\ell})|\hat{\mu}_\ell - \mu_\ell| < 2t + \frac{4}{n} + 8\sqrt{\frac{\kappa}{n}}\,. \tag{36}$$

Combining (29), (35), and (36), we obtain that on event $\mathcal{A}_{ij} \cap \mathcal{E}$, the following is true,

$$|\hat{\eta}_{i,j} - \eta(\mathbf{z}_i - \mathbf{z}_j)| \le 5t + \frac{8}{n} + 16\sqrt{\frac{\kappa}{n}}\,. \tag{37}$$

We next proceed to bound $|\hat{\eta}_0 - \eta(\mathbf{0})|$. Note that $\mathbb{E}[x_{i\ell}^2 - x_{i\ell}] = 2\mathbb{E}[q_{i\ell}^2]$. For $1 \le i \le n$, define the event $\mathcal{D}_i$ as,

$$\mathcal{D}_i: \quad \frac{1}{p}\sum_{\ell=1}^{p}\left(\frac{x_{i\ell}^2 - x_{i\ell}}{2} - \mathbb{E}[q_{i\ell}^2]\right) < t\,. \tag{38}$$

Recalling that $x_{i\ell}$ are independent for different ℓ, using Hoeffding's inequality, we obtain $\mathbb{P}(\mathcal{D}_i^c) \leq 2e^{-2pt^2}$. On the event $\cap_{i=1}^n \mathcal{D}_i \cap \mathcal{E}$, we have,

$$
\begin{aligned}
|\hat{\eta}_0 - \eta(0)| &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{p}\sum_{\ell=1}^{p}\Big(\frac{x_{i\ell}^2 - x_{i\ell}}{2} - \mathbb{E}[q_{i\ell}^2]\Big)\right| + \left|\frac{1}{p}\sum_{\ell=1}^{p}(\hat{\mu}_\ell^2 - \mu_\ell^2)\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{p}\sum_{\ell=1}^{p}\Big(\frac{x_{i\ell}^2 - x_{i\ell}}{2} - \mathbb{E}[q_{i\ell}^2]\Big)\right| + \frac{1}{p}\sum_{\ell=1}^{p}(\hat{\mu}_\ell + \mu_\ell)\,|\hat{\mu}_\ell - \mu_\ell| \\
|\hat{\eta}_0 - \eta(0)| &\leq 3t + \frac{4}{n} + 8\sqrt{\frac{\kappa}{n}}\,, \qquad (39)
\end{aligned}
$$

where we used (36) in the last inequality. Finally, by utilizing the union bound over $\mathcal{A}_{i,j}$, $\mathcal{D}_i$, and $\mathcal{E}$ for all $1 \leq i < j \leq n$, equations (37) and (39) are true with probability at least

$$
1 - (n^2 - n)e^{-2pt^2} - 2ne^{-2pt^2} - e^{-2pt^2} \geq 1 - (n+1)^2 e^{-2pt^2}\,. \qquad (40)
$$

Choosing $t = \sqrt{2\log(n+1)/p}$ gives the desired result.

References

Alfakih, A. Y., Khandani, A., and Wolkowicz, H., 1999. Solving Euclidean distance matrix completion problems via semidefinite programming. *Computational Optimization and Applications*, **12**:13–30.

Aulchenko, Y. S., Ripatti, S., Lindqvist, I., Boomsma, D., Heid, I. M., Pramstaller, P. P., Penninx, B. W., Janssens, A. C. J., Wilson, J. F., Spector, T., *et al.*, 2009. Loci influencing lipid levels and coronary heart disease risk in 16 european population cohorts. *Nature genetics*, **41**(1):47–55.

Belkin, M. and Niyogi, P., 2002. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, **15**:1373–1396.

Biswas, P. and Ye, Y., 2004. Semidefinite programming for ad hoc wireless sensor network localization. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, IPSN '04, pages 46–54, New York, NY, USA. ACM.

Bradburd, G., Ralph, P., and Coop, G., 2016. A spatial framework for understanding population structure and admixture. *PLoS Genetics*, **12**(1):e1005703–e1005703.

Dehghan, A., Dupuis, J., Barbalic, M., Bis, J. C., Eiriksdottir, G., Lu, C., Pellikka, N., Wallaschofski, H., Kettunen, J., Henneman, P., *et al.*, 2011. Meta-analysis of genome-wide association studies in¿ 80 000 subjects identifies multiple loci for c-reactive protein levels. *Circulation*, **123**(7):731–738.

Donoho, D. L. and Grimes, C., 2003. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(10):5591–5596.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., Wheeler, E., Glazer, N. L., Bouatia-Naji, N., Gloyn, A. L., *et al.*, 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, **42**(2):105–116.

Javanmard, A. and Montanari, A., 2013. Localization from incomplete noisy distance measurements. *Foundations of Computational Mathematics*, **13**(3):297–345.

Jay, F., Sjödin, P., Jakobsson, M., and Blum, M. G., 2013. Anisotropic isolation by distance: the main orientations of human genetic differentiation. *Molecular Biology and Evolution*, **30**(3):513–525.

Kooner, J. S., Chambers, J. C., Aguilar-Salinas, C. A., Hinds, D. A., Hyde, C. L., Warnes, G. R., Pérez, F. J. G., Frazer, K. A., Elliott, P., Scott, J., *et al.*, 2008. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nature Genetics*, **40**(2):149–151.

Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., Briley, L. P., Maruyama, Y., et al., 2008. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, **83**(3):347–358.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., *et al.*, 2008. Genes mirror geography within Europe. *Nature*, **456**(7218):98–101.

Patwari, N., Ash, J. N., Kyperountas, S., Moses, R., and Correal, N., 2005. Locating the nodes: cooperative localization in wireless sensor networks. *IEEE Signal Processing Magazine*, **22**:54–69.

Preisig, M., Waeber, G., Vollenweider, P., Bovet, P., Rothen, S., Vandeleur, C., Guex, P., Middleton, L., Waterworth, D., Mooser, V., *et al.*, 2009. The PsyCoLaus study: methodology and characteristics of the sample of a population-based survey on psychiatric disorders and their association with genetic and cardiovascular risk factors. *BMC Psychiatry*, **9**(1):9.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**(8):904–909.

Ramachandran, S. and Rosenberg, N. A., 2011. A test of the influence of continental axes of orientation on patterns of human gene flow. *American Journal of Physical Anthropology*, **146**(4):515–529.

Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., *et al.*, 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics*, **41**(1):35–46.

Saul, L. K. and Roweis, S. T., 2003. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *The Journal of Machine Learning Research*, **4**:119–155.

Scott, D. W., 1979. On optimal and data-based histograms. *Biometrika*, **66**(3):605–610.

Shang, Y., Ruml, W., Zhang, Y., and Fromherz, M. P., 2003. Localization from mere connectivity. In *Proceedings of the 4th ACM International Symposium on Mobile ad hoc networking & computing*, pages 201–212. ACM.

Singer, A., 2008. A remark on global positioning from local distances. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(28):9507–11.

Song, M., Hao, W., and Storey, J. D., 2015. Testing for genetic associations in arbitrarily structured populations. *Nature Genetics*, **47**(5):550–554.

Tenenbaum, J. B., Silva, V., and Langford, J. C., 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, **290**(5500):2319–2323.

Wasser, S. K., Shedlock, A. M., Comstock, K., Ostrander, E. A., Mutayoba, B., and Stephens, M., 2004. Assigning african elephant DNA to geographic region of origin: applications to the ivory trade. *PNAS*, **101**(41):14847–14852.

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., Perry, J. R., Stevens, S., Hall, A. S., *et al.*, 2008. Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics*, **40**(5):575–583.

Weinberger, K. Q. and Saul, L. K., 2006. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *Proceedings of the 21st national conference on Artificial Intelligence*, volume 2, pages 1683–1686. AAAI Press.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., *et al.*, 2013. Discovery and refinement of loci associated with lipid levels. *Nature genetics*, **45**(11):1274–1283.

Yang, W.-Y., Novembre, J., Eskin, E., and Halperin, E., 2012. A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics*, **44**(6):725–731.