Master of Science

A method to conciliate the application of shared genomic data and privacy

다기관 유전체 연구에서
유전체 정보 공유의 한계와 그 해결책

The Graduate School

of the University of Ulsan

Department of Medicine

Kim, Kun Hee

A method to conciliate the application of shared genomic data and privacy

다기관 유전체 연구에서
유전체 정보 공유의 한계와 그 해결책

Supervisor : Jin Kyung Roh

A Master's Thesis

Submitted to

the Graduate School of the University of Ulsan

In partial fulfillment of the requirements

for the degree of

Master of Science

by

Kim, Kun Hee

Department of Medicine

Ulsan, Korea

February 2020

A method to conciliate the application of shared genomic data and privacy

다기관 유전체 연구에서 유전체 정보 공유의 한계와 그 해결책

This certifies that the Master’s Thesis of Kim, Kun Hee is approved.

Committee Chair Professor Kyuyoung Song

Committee Member Professor Jin Kyung Roh

Committee Member Professor Buhm Han

Department of Medicine

Ulsan, Korea

February 2020

ABSTRACT

The advance in high-throughput genomic technology facilitated the generation of genomic data at an ever-decreasing cost. Aggregation of genomic data is indispensable for the progress of biomedical research. Sharing these data sets yields unbiased and novel findings through an increased sample size. However, a growing concern is the ability to protect the genetic privacy of the data originators. Therefore, full discloser of genetic information of samples is often limited, whereas summary-level statistics are shared among institutions. Although some researches proposed cryptographical or statistical approaches to moderate tension between the application of genomic data and privacy, most of them were kept within achieving a single determinate application.

We present a framework that reconciles privacy and genomic analysis by sharing additional distance information along with the summary-level genetic data of samples. In this work, we describe our framework which is built upon the concept of multilateration, a localization technique for wireless sensor networks in which spatial coordinates of a node with an unknown position are inferred by distances measured from the node to several reference nodes at known positions.

We show that through our framework, certain types of genomic analysis can be achieved, such as identifying sample overlaps and close relatives, decomposing ancestry, and mapping of geographical origin, without disclosing personal genomes.

Key words: genomic privacy, GWAS summary-level statistic, population genetics, sample overlap, genetic relatedness

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Recent technological development allowed accumulation of genetic information at an unprecedented rate. Gathering this genetic information led the expansion of our biomedical knowledge, where it has been used to elucidate genetic architecture. Recent studies suggest that the analysis of millions of samples is required to predict genetic architecture of complex traits from genetic data [1, 2)]. Collecting cohorts at such scales is typically beyond the reach of individual investigators and cannot be achieved without combining different sources. Therefore, it has been a common practice to combine genetic information from multiple institutions to maximize statistical power for the discovery of risk alleles of minute effect. However, at most times only summary-level statistics of a cohort is allowed to be shared in order to protect genetic information of data originators.

Although sharing of summary statistics concealed the genetic information of data originator, it undermines the statistical power of the analysis result as well as hinders various genomic analysis. For one, sample-level quality control has been difficult when combining multiple cohorts. The existence of genetically closely related samples, which includes sample overlap, can induce correlations between the summary statistics and inflate the false positive rate of meta-analyses [3)]. Also, it encumbers the application of genetic information to various genomic analysis, such as field of population genetic where the analysis is preposterous to perform only with summary statistics.

Several recent studies have proposed methods to overcome privacy issue in genetic information sharing. Some proposed cryptographic method to secure genetic information of data originators. He et al. [4)] have proposed a secure method for detecting the genetic relatives using genotype data using the ‘fuzzy’ encryption, where each individual has a public key and a private key. Public key for each individual is accessible by all the other individuals and the private key for each individual is hidden from all the other individuals. They show if two individuals are genetically related their secure method can detect them while not leaking any information. However, their methods were limited to achieving a single application and could not be applied to other genetic analysis.

Others have proposed statistical methods to assuage limitation of sharing of summary level statistics. Chen et al.[5] have proposed four metrics for monitoring and improving the quality of large-scale GWAMA based on summary statistics. Their F_{st}-derived genetic distance measure can detect cohort-level outlier by placing them on a geographical map, λ_{meta} can provide information on sample overlap and heterogeneity between cohorts by utilizing the estimated allelic effect sizes and their standard errors, and PPSR, in which the score is generated for each individual, can detect various degree of relatives. However, many of their metrics can only work to the quality control of cohort-level, not individual-level. Also, their PPSR score has limited to the achieve only one application.

We aim to define a framework that has versatile applications regarding summary statistics of multiple cohorts. In this work, we propose a new framework to reconcile the limitation of summary level statistics and the genetic research. Our framework uses localization technique of multilateration, which utilizes publically open genomic data such as 1000Genomes [6] and Hapmap projects [7] to allow multiple application of summary level statistics, including detection of genetically related samples to the population genetics. We show that our framework can detect genetically related samples up to first-degree relatives as well as overlapping samples, and reconstruct principal component map of samples to infer geographic origin of samples as well as decomposing samples' ancestry.

RESULTS

1. Genetically related sample detection

Our method builds upon multilateration, a localization technique for wireless sensor networks in which spatial coordinates of a node with an unknown position are inferred by measuring the distances from the node to several reference nodes at known positions [8)]. For example, in the GPS navigation system of an aircraft, the distances from the aircraft to satellites are calculated from time lags in transmitted radio signals. These distances are then used to calculate the aircraft's position (Fig. 1a).

The information of an individual to multiple reference individuals, which we call *"distance vector"* in the continuing manuscript (Fig. 1b), we aim to make inference on the genetic relatedness of any two individuals despite that itself is inscrutable to reconstruct target individual's genetic information

First, we devised a statistical model that utilize distance vector to determine if two individuals were sample overlap. Also, we show that the statistic we devised in the first application can also distinguish close relatives. Finally, we show that metric for measuring genetic distance need not be confined to Euclidean distance but other metric, such as genetic relatedness, can be used.

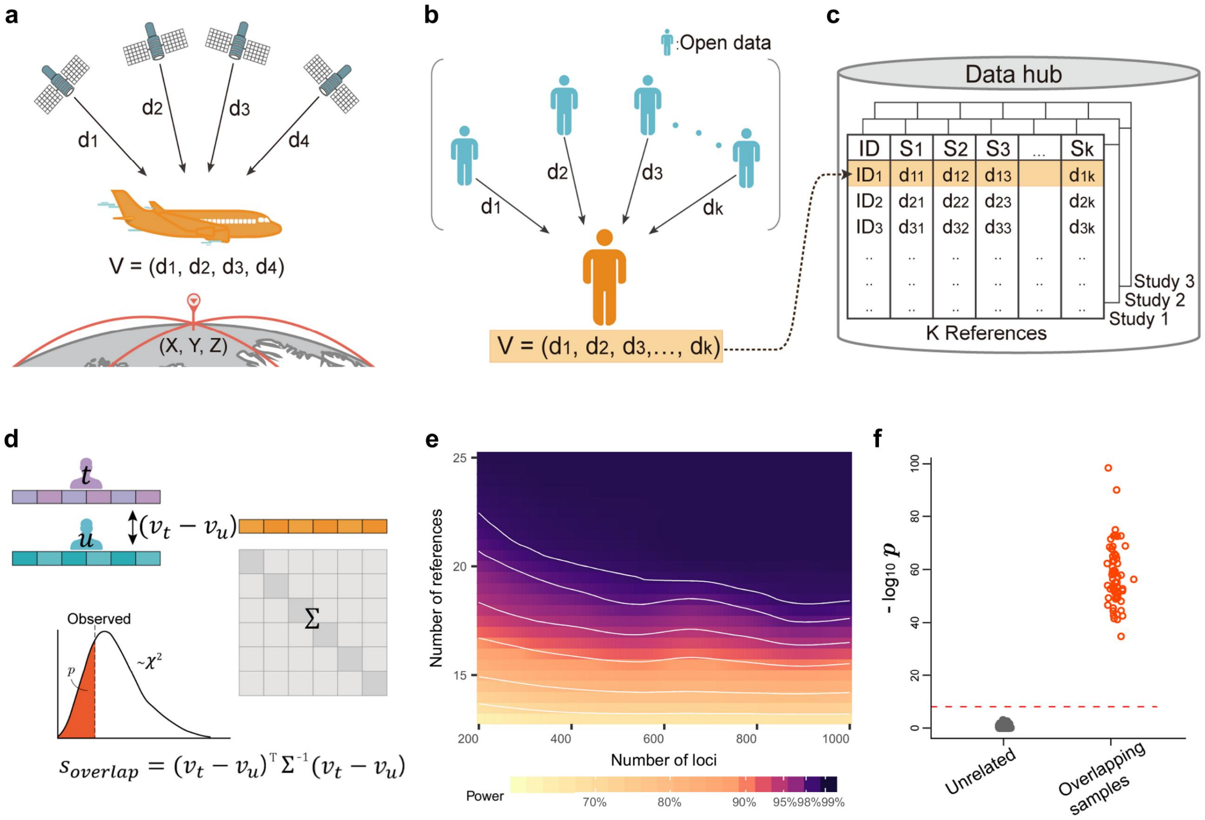


Fig. 1. Genomic GPS and its application to sample overlap detection.

a, The concept of conventional GPS. Distances to satellites are used to compute an aircraft's location. b, The concept of genomic GPS. The genetic distances of an individual to reference individuals in public datasets are calculated to create a distance vector. c, Distance vectors can be shared, for example, by using a public data hub. d, Sample overlap detection using distance vectors. The distance vectors of two individuals are compared using a statistic that follows a χ^2 distribution. e, The power of the sample overlap detection method as a function of the number of loci and the number of reference individuals used to calculate the genetic distances. f, *P*-values of the sample overlap detection method for overlapping pairs and unrelated pairs in the simulations using the WTCCC data.

1.1. Overlapping sample detection

1.1.1. Simulation

Our method which compare two distance vectors using a statistic s_{overlap} that follows a χ^2 distribution with the number of references as degrees of freedom is briefly illustrated in Fig. 1d. Details of method is described in **MATERIALS AND METHODS**.

To show our asymptotic distribution follows a χ^2 distribution with the number of references as degrees of freedom (df), we generated genotypes at 1,000 loci for 1,000,000 pairs of samples under the null hypothesis, to check the false positive rate. The 20 reference individuals were generated to calculate s_{overlap}. To avoid bias due to a specific reference dataset, we regenerated reference set for each pair of target samples. Fig. 2 shows that the empirical null distribution of our test statistic well matched to the asymptotic χ^2 distribution with 20 df. We estimated the false positive rate at a given significance threshold α as the proportion of simulated pairs with *P*-value $\leq \alpha$. The false positive rate was well controlled at varying thresholds from 0.05 to 0.0001 (Table 1).

In order to use the asymptotic distribution for *P*-value calculation, we will need a sufficiently large *N*. We found that we need a large *N* (>100) as well as a large ratio of *N*/*K* (>20) for accurate approximations (Fig. 3). Fortunately, these requirements are easy to meet in practical situations. Most of the currently available genotyping platforms provide >1,000 independent loci after LD pruning, thereby satisfying *N* >100 and *N*/*K* >20 requirements easily if we assume we use several tens of reference individuals (*K*).

To check the power, we simulated the alternative hypothesis that the pair of individuals was a sample overlap. While varying the number of loci (N) and the number of reference individuals (K), we checked the power using 100,000 simulated pairs of individuals. We used the significance threshold $\alpha = 10^{-7}$. The power at each ordered pair of (*N*, *K*) is shown in Fig. 1e. The result showed that power is dependent on both *N* and *K*, but the major determinant is *K*.

Based on our power and false positive rate results, we chose to use *N*=1,000 and *K*=30 (or

N=500 and *K*=20) in various simulations below. The reasons for this choice are the following: (1) We have >20 reference individuals in each population of the HapMap or the 1000Genomes data. (2) We will need a ratio of *N* >*K*×20 for an appropriate control of false positive rate. (3) Most of genotyping platforms will provide >1,000 independent SNPs after pruning. Note that in many real situations, we can increase *N* and *K* larger than these choices, which will increase power as shown in Fig. 1e.

1.1.2. Simulation with WTCCC dataset

We performed real data-based analysis using the Wellcome Trust Case Control Consortium (WTCCC) data [9)] by designing studies with overlapping samples. Our method could detect overlapping samples with perfect sensitivity (100%) and specificity (100%) (Fig. 1f), when using 50 randomly selected samples from the 1000Genomes data [6)] as reference.

We performed additional analysis to exclude the possibility that our method worked well because of some characteristics that these samples happened to have. We sampled 1,000 unrelated pairs and 1,000 overlapping samples from the WTCCC dataset. We used the same 1,000Genomes data as reference and used the same threshold. In this additional analysis, the distinction by our method was still perfect, showing that the results were not affected by sampling bias (Fig. 4).

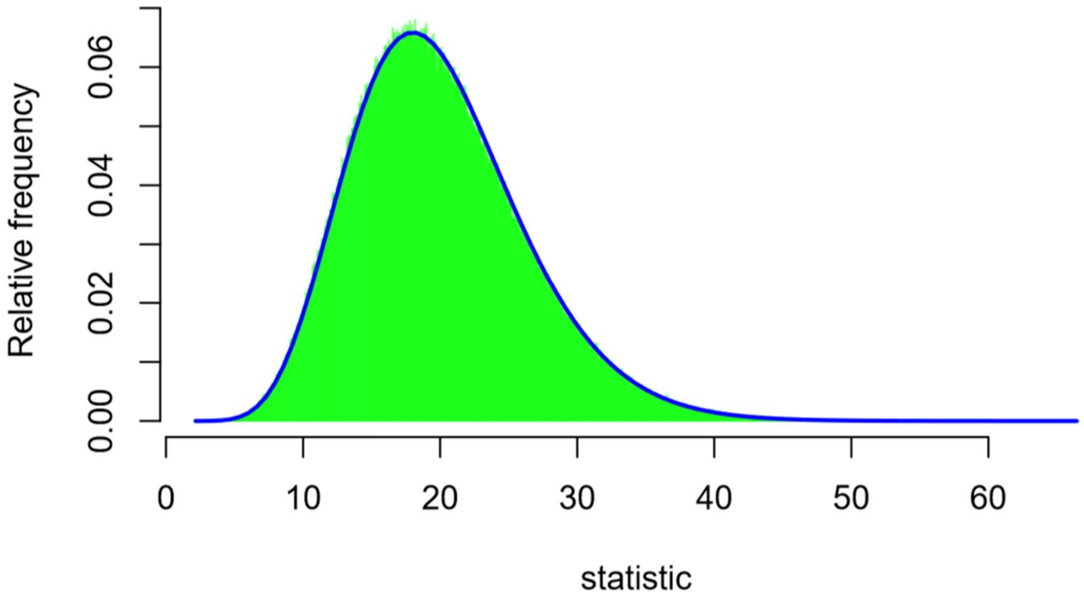


Fig. 2. The empirical distribution of s_{overlap}.

We show the histogram of 1,000,000 statistics simulated under the null hypothesis of unrelated pairs, assuming 20 reference individuals and 1,000 SNPs (K=20 and N=1,000). Blue line denotes the probability density function of the chi-square distribution with 20 df.

Table 1. False positive rate of the sample overlap detection statistic

We simulated 1,000,000 unrelated pairs of individuals. We used 1,000 loci and 20 reference individuals (N=1,000 and K=20). With a given threshold α, the false positive rate was estimated as the proportion of simulations with P-value $\leq \alpha$.

Threshold	False positive rate
0.05	0.049189
0.01	0.009932
0.005	0.00492
0.001	0.001017
0.0005	0.000527
0.0001	0.000108

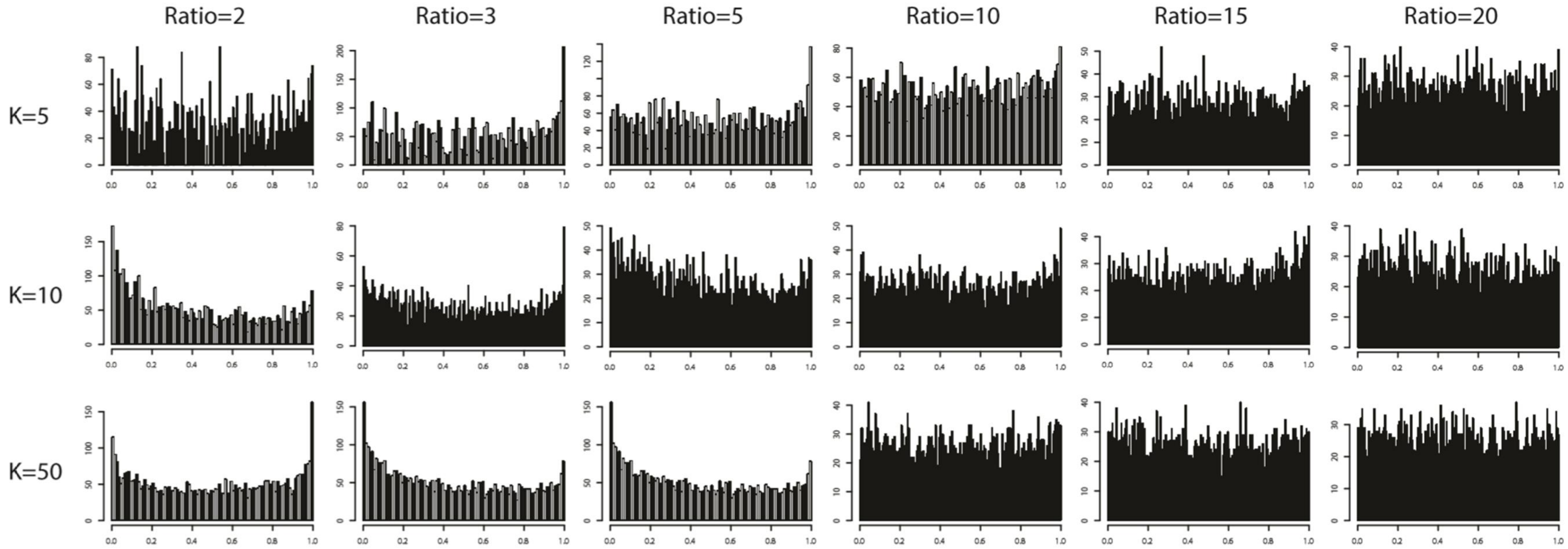


Fig. 3. *P* value distribution of the sample overlap detection method for different *N* and *K*.

We examined the validity of the asymptotic approximation of *P*-values for different numbers of loci (*N*) and reference individuals (*K*). A valid approximation will give us uniformly distributed *P*-values under the null hypothesis that individuals are unrelated. For each simulation, we generated 5,000 unrelated pairs. The ratio denotes *N* to *K* ratio, which is the number of SNPs divided by the number of references (*N*/*K*). At lower ratio (*N*/*K*<20), the distributions often showed peaks at one or both ends. At higher ratio (*N*/*K*>20), the distributions were closer to being uniform.

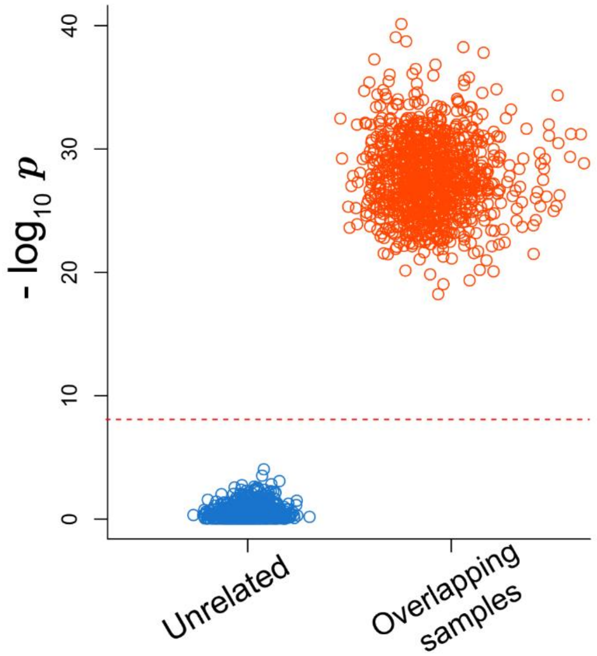


Fig. 4. *P* values of the sample overlap detection method for unrelated and overlapping samples in an additional real-data-based simulation.

To avoid possible sample selection bias in the real-data-based simulations using the WTCCC data, we performed an additional analysis to select 1,000 unrelated pairs and 1,000 overlapping pairs from the same dataset. The p-values of our overlapping sample detection method were distinct for the two groups, showing that the results in Fig. 1f were not driven by the bias.

1.2. Identification of relatives

In terms of genetic contents, overlapping samples are no different from twins. If the tested pairs are close relatives, it is possible that our statistic s_{overlap} can have a smaller expected value than unrelated pairs. In that sense, our method may possibly be used to distinguish relatives. We performed simulations to examine if our statistic can distinguish relatives, and if yes, to what extent. To this end, we simulated pairs of individuals that are relatives of different degrees. We gradually increased the degree from twins (equivalent to overlapping samples) to 1st degree, 2nd degree, 3rd, and unrelated. For each relation, we simulated 100,000 pairs and calculated their s_{overlap} using 30 reference individuals.

We then examined whether close relatives were distinguishable using our statistic. Density plot of our statistic under different degrees of relationships is shown in Fig. 5a. As described previously, the two clusters for twin pairs (in other words, overlapping samples) and unrelated pairs showed clear distinction. The 1st degree relatives formed a cluster that was located between the two clusters. This cluster was clearly distinctive from the twins but had some overlap with unrelated pairs.

We calculated the posterior probability of relationship as the density of the distribution of a specific relationship divided by the sum of densities of all relationships, assuming a uniform prior (Fig. 5b). Then we predicted the relationship of each pair as the relationship with the highest posterior probability. Given the true relationship being the 1st degree, 79% of simulated pairs were predicted to be the 1st degree (Fig. 5c). Thus, we can say that the 1st degree relatives were generally distinguishable, although the distinction may not be perfect in some cases. In contrast, starting from the 2nd degree relatives, the cluster had much larger overlap with unrelated pairs. For example, given the 2nd degree relatives, only 39% of simulated pairs were predicted to be the 2nd degree.

Next, we measured how precisely our predicted relationship represents true relationship. Out of all pairs predicted to be each degree of relationships, we measured the proportion of the correct prediction (thus, “precision”). The precisions were 100% for sample overlap, 65% for 1st degree relatives, 36% for 2nd degree relatives, 35% for 3rd degree relatives, and 51% for unrelated pairs (Fig. 6). Thus, we can say that for the 2nd or higher degree relationships, our method may not have sufficient distinctive power to predict the relationship correctly. The detailed result of prediction performance is described in Fig. 6.

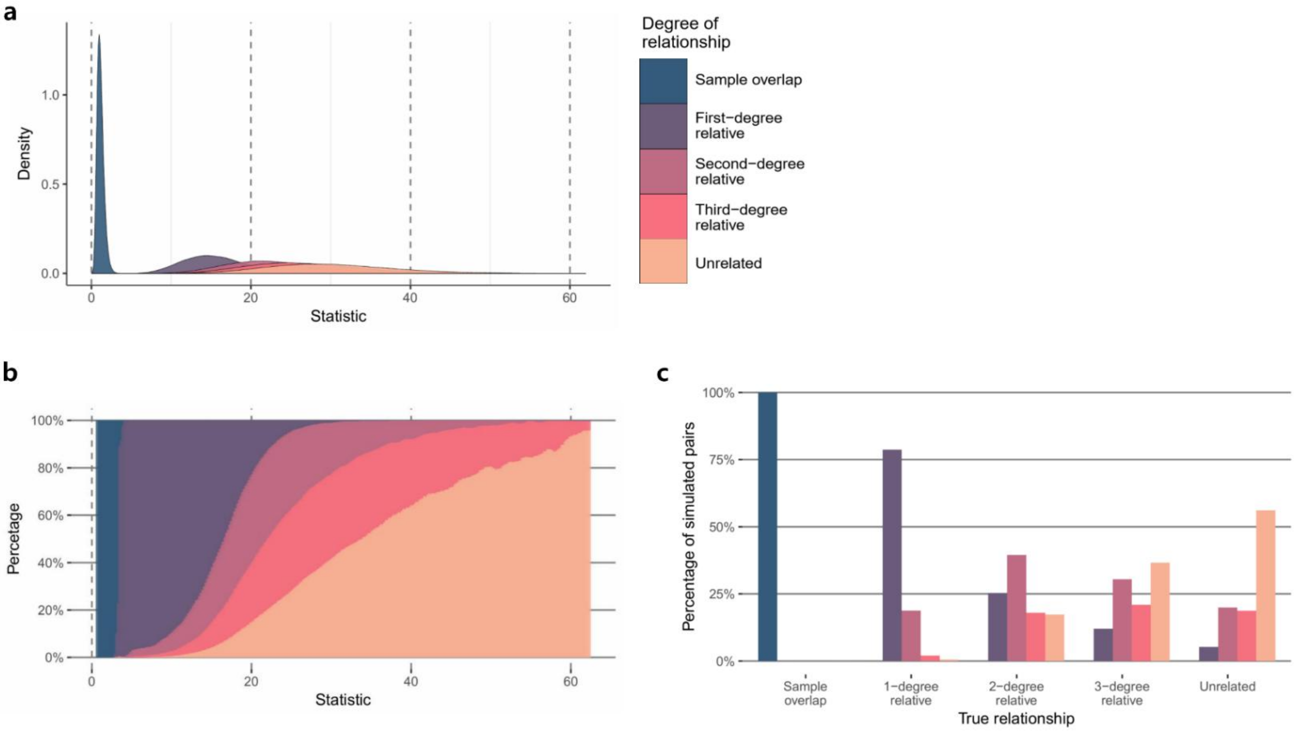


Fig. 5. Distribution of the sample overlap detection statistic for different degrees of relatives. We obtained the distribution of our sample overlap statistic under different relationships of pairs: sample overlap (or twins), 1st degree relatives, 2nd degree relatives, 3rd degree relatives, and unrelated pairs. We assumed 1,000 loci and 30 reference individuals (*N*=1000 and *K*=30). a, The density of statistic for differing degrees of relatives. b, The posterior probability of being in each category given the statistic, which was calculated as the probability density of the category divided by the sum of the densities of all categories. c, The proportion of correct assignment. This is the proportion of samples of a specific relationship that was correctly assigned to that relationship after determining the most likely relationship based on the posterior probability.

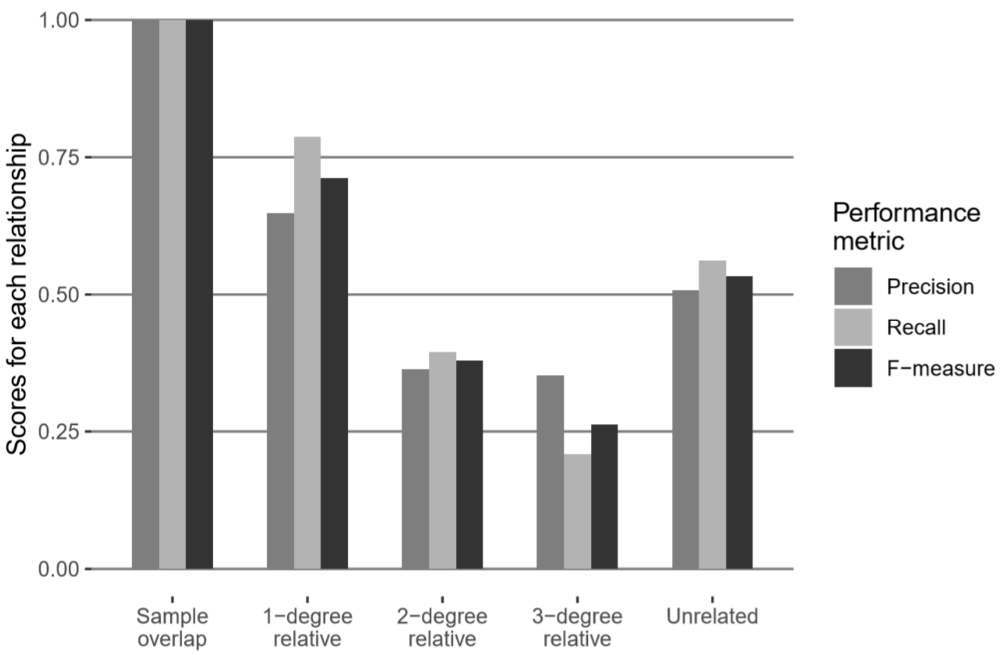


Fig. 6. Precision and recall of relationship prediction using sample overlap detection statistic. We simulated 100,000 pairs assuming each of 5 relationships. We assumed 1,000 loci and 30 reference individuals (*N*=1000 and *K*=30). We calculated their sample overlap detection statistics and predicted the most likely relationship based on the statistics. We measured performance of our predictions using 3 metrics; precision was measured as the proportion of true assignments out of total number of predicted assignments to a specific relationship, recall was measured as the proportion of correct assignments out of total simulated pairs of a specific true relationship, and F-measure was calculated as $2 \times \frac{precision \ \times \ recall}{precision \ + \ recall}$.

1.3. Using genetic relatedness as metric

Although we had been using the squared Euclidean distance as our metric of genetic distance for sample overlap detection, other metrics for distance can also be used. The genetic relatedness is a commonly used measure of genetic distance in quantitative genetics, for example for calculating the genetic relationship matrices (GRMs) in heritability estimation [10)]. When genetic relatedness was used as measure for distance, the result was concordant to when Euclidean distance was used as measure. We simulated 10,000 pairs under the null hypothesis (pairs are unrelated) and 10,000 pairs under the alternative hypothesis (pairs are sample overlaps). The *P*-values under the null hypothesis were uniformly distributed (Fig. 7a). Also, the sample overlap formed a clearly distinguishable cluster in the histogram of statistics (Fig. 7b). These results showed that the genetic relatedness can effectively be used for sample overlap detection by using the empirically estimated covariance matrix.

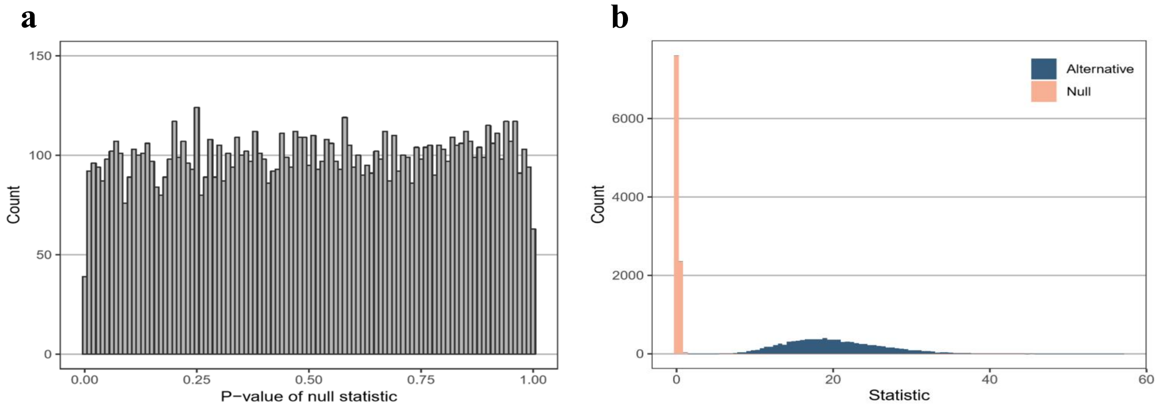


Fig. 7. Characteristics of the sample overlap detection method when using genetic relatedness as the distance measure.

a, *P*-value distribution under the null hypothesis. 10,000 pairs were simulated under the null hypothesis where pairs were unrelated. Their distance vectors were calculated as the genetic relatedness to the 20 simulated reference individuals (*N*=500, *K*=20). X-axis is the *P*-value acquired from the lower tail of χ^2_{20} distribution. b, Density plot of the statistic under the null and the alternative hypotheses. 10,000 pairs were simulated under both hypotheses (null hypothesis: the pairs were unrelated, and alternative hypothesis: the pairs were sample overlap or identical twins). Blue bars show the density of statistics under the null hypothesis and red bars show the density of statistics under the alternative hypothesis.

2. Principal component map construction

Distance vectors can be used for several population genomic analyses. First, they can be used for constructing the spatial structure in the principal component (PC) map. As Novembre et al. [11] showed, two-dimensional PC plot can approximate the geographical origins of individuals. Typically, actual genotype data of individuals are required to calculate PCs. We developed a method to approximate the PC map using only distance vectors without requiring actual genotype data.

2.1. Reconstructing spatial map of POPRES dataset

To show the distance vector can reconstruct PC map, we used POPRES dataset [12]. This dataset includes 1387 samples from 36 European countries. Population codes used in figures are described in Table 2.

2.1.1. POPRES as a reference set

The first analysis was to use POPRES for both target samples and reference. We subsampled 40% from each population of POPRES data and used them as our reference set. The rest (60%) of the data was used as our target samples. We calculated distance vectors of 815 target samples and approximated their position on the PC map of reference samples.

The approximated PC map based on distance vectors (Fig. 8a) greatly resembled the original PC map based on actual genotypes (Fig. 8b). The samples from same geographic regions were clustered together and the populations were distinguishable. Populations geographically adjacent were found near each other and populations geographically apart were found far from each other. Also, some populations in our result corresponded to the geographic outline of Europe. Spanish and Portuguese samples (population codes: SP and PT in Fig. 8a) formed a shape similar to Iberian Peninsula, Italian samples (population code: IT in Fig. 8a) formed a shape similar to Italian peninsula, and English and Irish samples (population codes: UK and IE in Fig. 8a) formed a shape similar to two main islands of United Kingdom.

The accuracy of the spatial structure mapping can depend on the number of variants used. We gradually decreased the number of SNPs used in this analysis by subsampling SNPs. As expected, the resolution of mapping was reduced as fewer SNPs were used (Fig. 9). In particular, the resolution drastically decreased when the number of SNPs was reduced from 50,000 to 10,000.

2.1.2. 1000Genomes as a reference set

Then, we used 305 European samples from the 1000Genomes data [6)] as our reference. Fig. 9 shows that the constructed PC map based on distance vectors, which roughly resembled the European map. However, the resolution decreased when compared to the PC map based on genotype data (Fig. 1a of Novembre et al. [11)]) or when compared to our analysis using subsamples of POPRES as reference (Fig. 8a). For example, it was hard to distinguish Eastern Europe and Russian populations, where they were entangled with central Europe population. The lower resolution was expected, because our method uses the PC map of reference data as "anchors" and therefore depends on how much variability of the target samples the reference data represents. Indeed, the 1000Genomes data lacked references from Eastern Europe and Russia . We expect that the resolution of this analysis will keep increasing as more diverse and ample reference datasets are built.

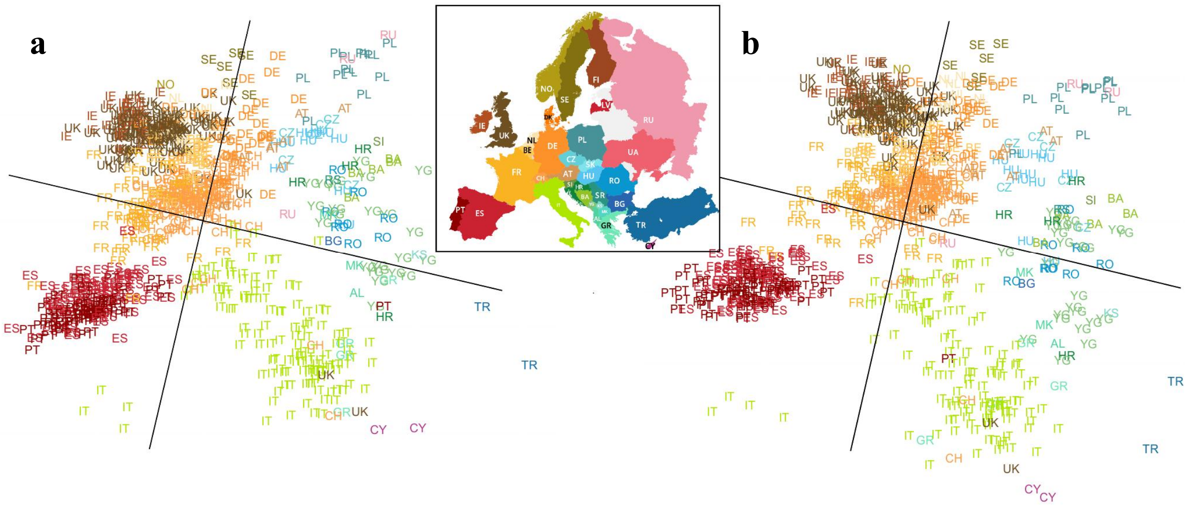


Fig. 8. Comparison of two dimensional mapping methods of the Europeans in the POPRES data.

a, Two-dimensional mapping of the Europeans in the POPRES data using only distance vectors. We mapped a subset (60%) of the POPRES individuals and used the rest of the individuals (40%) as references. See Table 2 for the abbreviated population names. b, Mapping result of the same individuals using actual genomic data (the top two PCs).

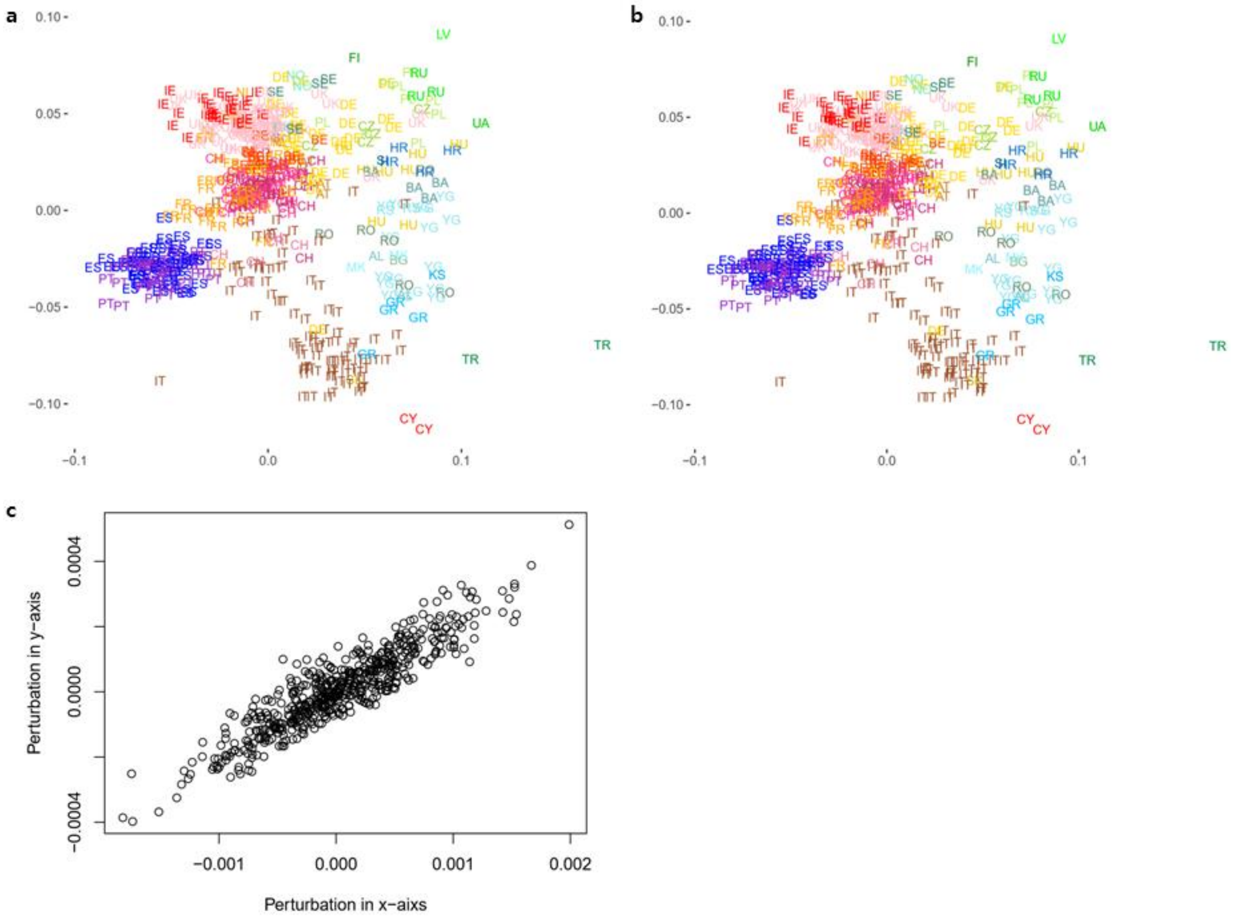


Fig. 9. Changes in the top two principal components of the reference set when one additional sample is added to the reference set.

a, In the POPRES analysis where we used 40% of the samples as reference, we plotted the reference samples' top two principal components. In the **MATERIALS AND METHODS**, we call this space $\mathcal{P}$. b, To map a sample based on distance vector, we added target sample to the reference set and regenerated the principal components. Because of this addition, the reference samples' positions were slightly distorted. In the **MATERIALS AND METHODS**, we call this space $\mathcal{P}'$. c, Between subfigure a and b, we compared the location of the same reference datapoint and measured the difference (perturbation) in location. We repeated this comparison for all reference datapoints. The comparison was done after adjusting for the rotation to align the two figures as much as possible.

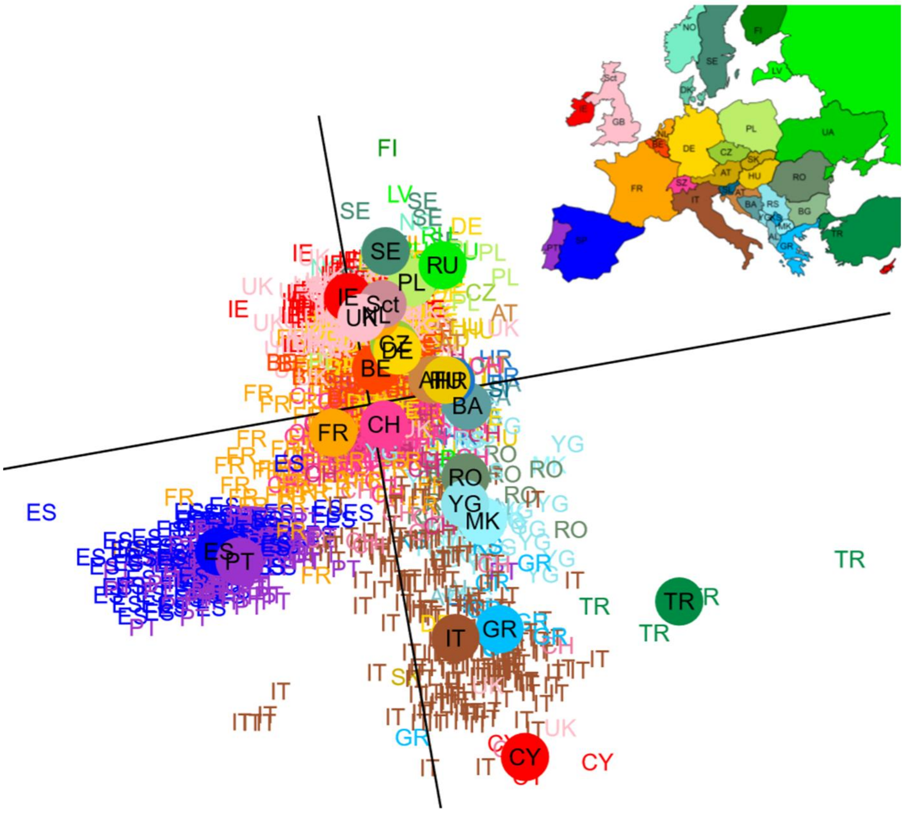


Fig. 10. Two dimensional mapping of the Europeans in the POPRES data using distance vectors calculated with 1000Genomes dataset.

We reconstructed PC map of entire 1396 samples from POPRES dataset using our method. 201 European samples from 1000Genomes dataset were used as references to calculate distance vectors (reference set only includes 1000Genomes with population code of GBR, TSI, and IBS).

2.2. Ancestry estimation

In this simulation, we evaluated another application of distance vector to population genomic analysis, ancestry estimation for admixed individual. Using the 1000Genomes data [6], we simulated admixed individuals from two populations, gradually varying the proportions. When the two populations were genetically distant, Japanese (JPT) and British (GBR) population data from the 1000Genomes, the estimated proportion was close to the true proportion (r^2 = 0.98, Fig. 11a). When two populations were genetically close, we can expect that decomposing the proportion will be more difficult. We simulated admixed individuals of British (GBR) and Toscani in Italia (TSI) from 1000Genomes. The estimation was less accurate but showed high correlation to the true proportion (r^2 = 0.86, Fig. 11a).

We then combined data for three populations (CHS, GBR, and YRI from the 1000Genomes data) in varying proportions. For comparison, we applied an existing method that can assign individuals to population groups or decompose the ethnic composition of an individual, ADMIXTURE [13]. Both ADMIXTURE and our method gave estimations that were highly concordant with the true proportions (Fig. 11b).

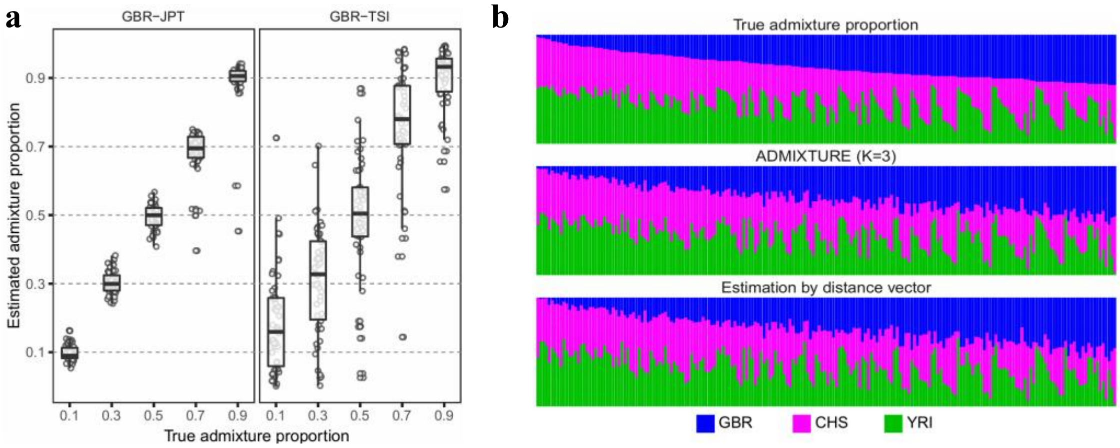


Fig. 11. Estimation of admixture proportion using distance vectors.

a, We simulated admixed individuals from two distant populations (GBR: British in England and Scotland and JPT: Japanese in Tokyo, Japan) and two close populations (GBR and TSI: Toscani in Italia) using the 1000Genomes data. b, Admixture of three populations (GBR, CHS: Southern Han Chinese, and YRI: Yoruba in Ibadan, Nigeria). The proportions were estimated using distance vectors and ADMIXTURE.

3. Genomic information is inscrutable through distance vector

Because our method conceals the genotype data of an individual and only shares a distance vector to reference data, it is important that the genotype data should not be recovered by the information in the distance vector. That is, we don't want the exact position of the individual in the *N*-dimensional space to be identified or specified by the distance vector. We describe mathematical proofs and simulations related to this issue. Note that we use the term "unidentifiability" to describe the condition that the actual genotypes cannot be recovered, but not the condition that the individual identity cannot be identified. The latter can have different meanings depending on the context; if one has distance vectors of a group of individuals and wants to identify if a new individual is in the group, it is certainly possible by comparing distance vectors as described in our application of overlapping sample detection.

We derived a mathematical proof showing that in *N*-dimensional space, and with *K* reference nodes with known positions, an unknown node's coordinates can be unequivocally identified if $K>N$. Perhaps more importantly, we derived another proof showing that an unknown node's coordinates can never be exactly specified if $K<N-1$ (**MATERIALS AND METHODS**). This was encouraging because it suggested that the distances to known nodes convey limited information under this condition and can be safely shared without disclosing the actual location.

If we imagined that genotypes were real numbers, it is theoretically impossible to reconstruct the genotypes as long as $N \gg K$, as shown by our proof. Unfortunately, genotype data resides in a very restricted space, $\{0,1,2\}^N$. Nevertheless, the search space is still large enough to prevent data reconstruction in practice. We designed a greedy algorithm that tries to reconstruct the genotype data given a distance vector and reference (**MATERIALS AND METHODS**) and applied it to simulated data. To avoid local optima, we allowed multiple restarts of the algorithm to find the best possible solution.

We then ran greedy algorithm 1,000 times to obtain 1,000 solutions. Note that each solution was the best selection among 1,000 candidate local optima. Given 1,000 solutions, we were able to measure overall accuracies of the prediction. For comparison, we generated another set of 1,000 solutions simply by random generation based on the allele frequencies.

We first evaluated the per-individual accuracy. The average per-individual accuracy was 39.4% in random samples and was 40.3% in greedy solutions (Fig. 12a). The *t*-test *P*-value was significant (*P*<1E-15), which means that the greedy algorithm was able to push the genetic contents toward the target sample. However, the small magnitude difference in accuracy (which is only 0.9%) shows that it is unlikely that the greedy algorithm can achieve 100% accuracy to recover the genome. Then, we measured the per-SNP accuracy. Unlike the per-individual accuracy, the *t*-test *P*-value was not significant (*P*=0.09). This shows that at the SNP level, the prediction by the greedy algorithm is not much better than the random prediction based on the allele frequency. The distributions are shown in Fig. 12b.

Finally, we checked if the risk score for a disease would be different between the two groups. We generated random weights of SNPs from the uniform distribution (0,1) and normalized them to have a sum of 1. We calculated the weighted average of the allele dosage (risk score) for individuals. When we examined the between-group difference of risk score between random samples and greedy solutions, *t*-test was significant (*P*=0.003). This was not surprising because the risk score can be considered per-individual information, and the per-individual accuracy showed difference above. However, the absolute magnitude of score was similar between the two groups; the mean of risk score was 0.989 in random samples and 0.986 in greedy solutions. This small difference shows that it is difficult to extract the risk information of an individual from the distance vector alone. The distributions are shown in Fig. 12c (where the blue dashed line indicates the risk score of the target sample).

Although simulations in the previous section showed that it is difficult to recover the gnomic data from distance vectors, the interpretation is not conclusive because we used a simple greedy algorithm. One may argue that if an opponent (who wants to breach privacy) develops a better algorithm, the genome can be revealed. Instead of implementing every possible algorithm as our opponent, which is impossible, we show a simple analysis on the complexity.

If we use 1,000 SNPs, our search space is of the size 3^{1000}. Apparently, this is a very large space. A widely used group of cryptographic hashing algorithms is a group called Secure

Hashing Algorithm 2 (SHA-2), which includes the popular SHA-256 [14)]. The complexity of SHA-256 is ~2^{255}. Thus, the search space of our problem is $\frac{3^{1000}}{2^{255}} = 10^{400}$ times bigger. Although the likelihood in our search space is not uniform due to the knowledge of allele frequencies, we can always reduce this skewness by selecting only common SNPs. Moreover, it is easy for us to increase the number of SNPs beyond 1,000. Overall, the search space is large enough such that it is highly unlikely that the summarized information in distance vector can allow us to recover the exact solution.

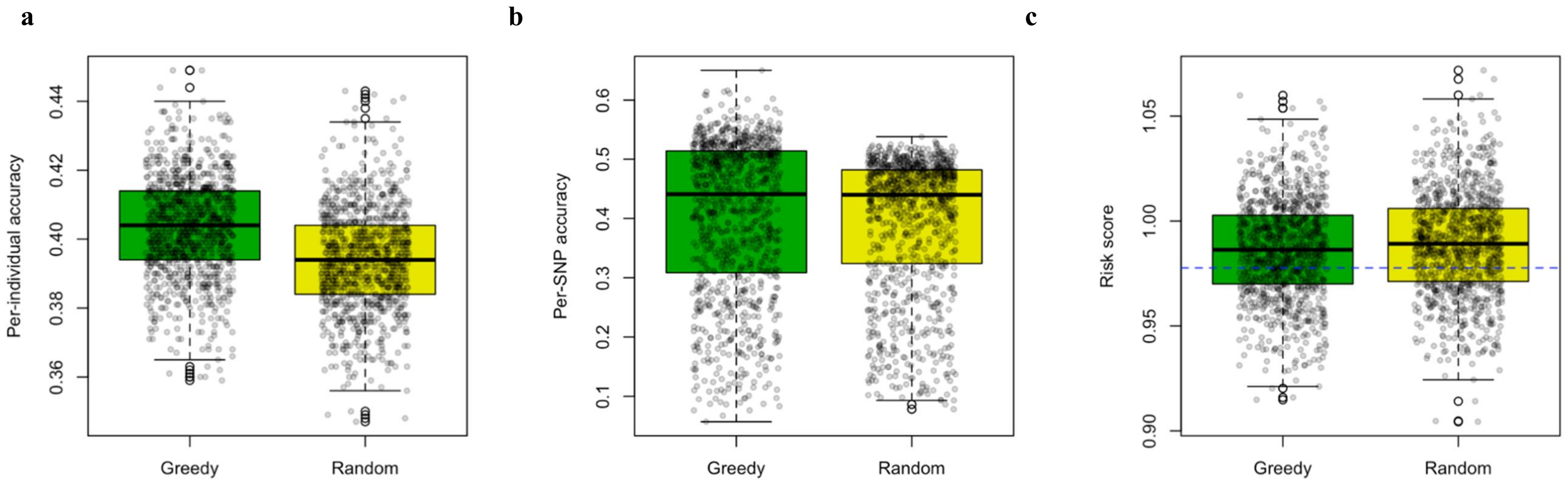


Fig. 12. Comparison of greedy algorithm solutions and random solutions.

We designed a greedy algorithm that tries to reconstruct the genotypes of a target sample given distance vector and reference data (see **Supplementary Note**). We simulated a target sample and reference individual data, assuming 30 reference individuals and 1,000 SNPs (K=30 and N=1,000) and applied the algorithm to the data. We also constructed random solutions based on allele frequencies. We compared the greedy and random solutions (1,000 solutions each) to the target sample genotypes and measured accuracy. a, Per-individual prediction accuracy. b, Per-SNP prediction accuracy. c, Per-individual risk score accuracy for a simulated set of risk weights of SNPs. The blue dash line is the risk of the target sample.

MATERIALS AND METHODS

1. Basic concept of the Method

Suppose that an individual's genomic sequence has *N* loci. A common form of the locus is single nucleotide polymorphism (SNP), where each locus can have a value of 0, 1, or 2 (the count of the reference allele). Given *N* SNPs, we define an *N*-dimensional space, where an individual's SNP data specifies the individual's position. The distance between two individuals in this space represents the genetic distance between the two. Assume that we have a target individual t whose position in this space is unknown and a reference set of multiple (K) individuals with known positions, such as samples from the HapMap [7)] or 1000Genomes [6)]. We can calculate the distances between the target individual *t* and those of the reference individuals to obtain a length-*K* vector, v_t, which we call a distance vector. The i^{th} element of the distance vector represents the distance between t and the i^{th} reference individual (Fig. 1b). A distance vector can be shared among institutions or researchers for a number of purposes, as discussed in the main article, without disclosing the individual's genotype data (Fig. 1c).

2. Statistical modeling

2.1. Statistic $s_{overlap}$

Distance vectors can be used for detecting sample overlaps. Consider two target individuals t and u whose distance vectors to K satellites (v_t and v_u) are known. To detect if t and u are a sample overlap, we calculate a statistic

$$s_{\text{overlap}} = (v_t - v_u)^T \Sigma^{-1} (v_t - v_u), \qquad (1)$$

where Σ is the covariance matrix of $v_t - v_u$. Under the condition that loci are independent, our statistic follows a χ^2 distribution with K degrees of freedom (df) under the null hypothesis that the two individuals are unrelated.

For measuring the distance between the two individuals, differing metrics can be used. Below, we use the squared Euclidean distance. If we use the squared Euclidean distance, Σ can be analytically calculated. Let $X_{t,n} \in \{0,1,2\}$ be the reference allele count of individual *t* at SNP

n. The squared Euclidean distance between individuals t and u is $D_{t,u} = \sum_{n=1}^{N}(X_{t,n} - X_{u,n})^2$. However, a different metric such as the genetic relatedness can also be used for this statistic, which we describe in section **2.3**.

2.1.1. Covariance matrix Σ

Below we derive the closed form of the covariance matrix, step by step.

Two individuals.

Suppose that we have two individuals A and B. Suppose that we have a single locus (N=1). For simplicity, we will use the same letters (A and B) to refer to the reference allele count of A and B. Given that A and B can have a value of 0, 1, or 2, the squared Euclidean distance, $(A-B)^2$, can have a value of 0, 1, or 4. Under the null hypothesis that the two individuals are unrelated, each individual has 0, 1, or 2 with probability p^2, $2p(1-p)$, and $(1-p)^2$ where p is the population frequency of the non-reference allele. If we enumerate all possible cases and their probabilities,

A	B	$(A-B)^2$	Probability
0	0	0	p^4
0	1	1	$2p^3(1-p)$
0	2	4	$p^2(1-p)^2$
1	0	1	$2p^3(1-p)$
1	1	0	$4p^2(1-p)^2$
1	2	1	$2p(1-p)^3$
2	0	4	$p^2(1-p)^2$
2	1	1	$2p(1-p)^3$
2	2	0	$(1-p)^4$

We can summarize the table with respect to $(A-B)^2$,

$(A-B)^2$	Probability
0	$p^4 + 4p^2(1-p)^2 + (1-p)^4$
1	$2p^3(1-\mathrm{p}) + 2p^3(1-\mathrm{p}) + 2p(1-p)^3 + 2p(1-p)^3$
4	$p^2(1-p)^2 + p^2(1-p)^2$

From this table, we can compute the mean and variance of $(A-B)^2$:

$$E[(A-B)^2] = -4p^2 + 4p \tag{2}$$

$$Var[(A-B)^2] = 8p^4 - 16p^3 + 4p^2 + 4p$$

Now consider that we have *N* independent loci. Now, A and B refer to the size-N vector of allele counts in individuals A and B. Because we assume that all loci are independent, simply adding the locus-wise mean and variance for all N loci gives us the mean and variance of the squared Euclidean distance:

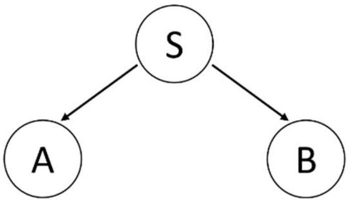
$$E[\|A-B\|^2] = \textstyle\sum_{n=1}^{N} -4p_n^2 + 4p_n$$

$$Var[\|A-B\|^2] = \textstyle\sum_{n=1}^{N} 8p_n^4 - 16p_n^3 + 4p_n^2 + 4p_n$$

where p_n is the population frequency of the non-reference allele at the *n^th^* locus. If N is large, due to the central limit theorem, $\|A-B\|^2$ follows a normal distribution with the mean and variance specified above.

Two individuals and a reference individual.

Now suppose a situation that the distance between A and B cannot be measured, but the distances from A and B to a third person, our reference individual S, are known. This relation is described in the diagram below:



We call this relationship as a triad relationship, where there are two terminal vertices with degree of 1 connected to a single vertex with degree of 2. (Here, degree refers to the number of edges connected to a vertex). In our situation, the distance between each terminal vertex (A or B) and the connection vertex (S) is known. Let $D_{SA} = (S-A)^2$ be the squared Euclidean distance between S and A. Similarly, we define $D_{SB} = (S-B)^2$ and $D_{AB} = (A-B)^2$. We know D_{SA} and D_{SB}, but we do not know D_{AB}.

Our approach is to use $D_{SA} - D_{SB}$ as our measure to check whether A and B are sample overlap. If A and B have the same genetic composition (sample overlap), $D_{SA} - D_{SB}$ will obviously become zero, or close to zero even with genotyping errors. Assume that we only have a single locus. To derive the mean and variance of $D_{SA} - D_{SB}$, we enumerate all possible cases:

S	A	B	D_{SA}	D_{SB}	$D_{SA} - D_{SB}$	Probability
0	0	0	0	0	0	p^6
0	0	1	0	1	-1	$2p^5(1-p)$
0	0	2	0	4	-4	$p^4(1-p)^2$
0	1	0	1	0	1	$2p^5(1-p)$
0	1	1	1	1	0	$4p^4(1-p)^2$
0	1	2	1	4	-3	$2p^3(1-p)^3$

0	2	0	4	0	4	$p^4(1-p)^2$
0	2	1	4	1	3	$2p^3(1-p)^3$
0	2	2	4	4	0	$p^2(1-p)^4$
1	0	0	1	1	0	$2p^5(1-p)$
1	0	1	1	0	1	$4p^4(1-p)^2$
1	0	2	1	1	0	$2p^3(1-p)^3$
1	1	0	0	1	-1	$4p^4(1-p)^2$
1	1	1	0	0	0	$8p^3(1-p)^3$
1	1	2	0	1	-1	$4p^2(1-p)^4$
1	2	0	1	1	0	$2p^3(1-p)^3$
1	2	1	1	0	1	$4p^2(1-p)^4$
1	2	2	1	1	0	$2p(1-p)^5$
2	0	0	4	4	0	$p^4(1-p)^2$
2	0	1	4	1	3	$2p^3(1-p)^3$
2	0	2	4	0	4	$p^2(1-p)^4$
2	1	0	1	4	-3	$2p^3(1-p)^3$
2	1	1	1	1	0	$4p^2(1-p)^4$
2	1	2	1	0	1	$2p(1-p)^5$
2	2	0	0	4	-4	$p^2(1-p)^4$
2	2	1	0	1	-1	$2p(1-p)^5$
2	2	2	0	0	0	$(1-p)^6$

We can summarize the table with respect to $D_{SA} - D_{SB}$,

$D_{SA} - D_{SB}$	Probability
0	$p^6 + 4p^4(1-p)^2 + p^2(1-p)^4 + 2p^5(1-p) + 2p^3(1-p)^3$ $+ 8p^3(1-p)^3 + 2p^3(1-p)^3 + 2p(1-p)^5$ $+ p^4(1-p)^2 + 4p^2(1-p)^4 + (1-p)^6$
1	$2p^5(1-p) + 4p^4(1-p)^2 + 4p^2(1-p)^4 + 2p(1-p)^5$
3	$2p^3(1-p)^3 + 2p^3(1-p)^3$
4	$p^4(1-p)^2 + p^2(1-p)^4$
-1	$2p^5(1-p) + 4p^4(1-p)^2 + 4p^2(1-p)^4 + 2p(1-p)^5$
-3	$2p^3(1-p)^3 + 2p^3(1-p)^3$
-4	$p^4(1-p)^2 + p^2(1-p)^4$

From this table, we obtain

$$E[D_{SA} - D_{SB}] = 0$$

$$Var[D_{SA} - D_{SB}] = 24p^4 - 48p^3 + 20p^2 + 4p$$

Again, consider that we have N independent loci. Then, $D_{AB} = \|A - B\|^2$. Because we assume that all loci are independent, simply adding the locus-wise mean and variance for all N loci gives us the mean and variance of the difference in squared Euclidean distances.

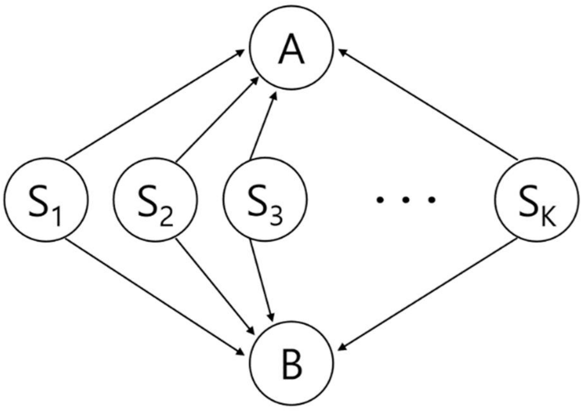
$$E[D_{SA} - D_{SB}] = 0$$

$$Var[D_{SA} - D_{SB}] = \sum_{n=1}^{N} 24p_n^4 - 48p_n^3 + 20p_n^2 + 4p_n$$

Due to the central limit theorem, if N is large, $D_{SA} - D_{SB}$ will follow a normal distribution with the mean and variance specified above.

Two individuals and multiple reference individuals.

To detect sample overlap, we use a set of reference individuals instead of one. Suppose that we use *K* references. The relation is described in the diagram below:



Distances for all edges are known. Thus, we can define distance vectors, referred as V_A and V_B. Distance vector, V_A (or V_B), is a size-K vector in which the i^{th} element contains distance measured between the sample A (or B) to the i^{th} reference individual. That is,

$$V_A = \left(D_{S_1A}\ ,D_{S_2A}\ ,\dots,D_{S_KA}\right) = ((S_1 - A)^2,(S_2 - A)^2,\dots,(S_K - A)^2)$$

$$V_B = \left(D_{S_1B}\ ,D_{S_2B}\ ,\dots,D_{S_KB}\right) = ((S_1 - B)^2,(S_2 - B)^2,\dots,(S_K - B)^2)$$

If A and B are overlapping samples, each element of $V_A - V_B$ will be zero or close to zero even with measuring errors. Thus, we can use the difference of the two distance vectors ($V_A - V_B$) as for our statistic to test sample overlap. To this end, we needed to de-correlate elements of $V_A - V_B$ using covariance matrix. Let Σ be the covariance of $V_A - V_B$ under the null hypothesis that A and B are unrelated. Let $V_A(i)$ be the ith element of V_A. Then we can calculate the (i, j)th element of Σ as follows $(i \neq j)$.

$$\Sigma_{ij}= cov(V_A(i) - V_B(i), V_A(j) - V_B(j))$$

$$= cov(V_A(i), V_A(j)) - cov(V_A(i), V_B(j)) - cov(V_B(i), V_A(j)) + cov(V_B(i), V_B(j))$$

Since D_{S_iA} and D_{S_jB} are uncorrelated, $cov(V_A(i), V_B(j)) = 0.$ Thus,

$$= cov(V_A(i), V_A(j)) + cov(V_B(i), V_B(j))$$

Since we assume that A and B are from the same general population,

$= 2 * cov(V_A(i), V_A(j))$

$= 2 * cov(D_{S_iA}, D_{S_jA})$

Assume that the reference individuals are from the general population. Then, $S_i \rightarrow A \leftarrow S_j$ triad relationship is no different from $A \leftarrow S \rightarrow B$ in terms of covariance.

$= 2 * cov(D_{SA}, D_{SB})$

This quantity can be obtained by enumerating all possibilities:

S	A	B	D_{SA}	D_{SB}	$D_{SA}D_{SB}$	Probability
0	0	0	0	0	0	p^6
0	0	1	0	1	0	$2p^5(1-p)$
0	0	2	0	4	0	$p^4(1-p)^2$
0	1	0	1	0	0	$2p^5(1-p)$
0	1	1	1	1	1	$4p^4(1-p)^2$
0	1	2	1	4	4	$2p^3(1-p)^3$
0	2	0	4	0	0	$p^4(1-p)^2$
0	2	1	4	1	4	$2p^3(1-p)^3$
0	2	2	4	4	16	$p^2(1-p)^4$
1	0	0	1	1	1	$2p^5(1-p)$
1	0	1	1	0	0	$4p^4(1-p)^2$
1	0	2	1	1	1	$2p^3(1-p)^3$
1	1	0	0	1	0	$4p^4(1-p)^2$

1	1	1	0	0	0	$8p^3(1-p)^3$
1	1	2	0	1	0	$4p^2(1-p)^4$
1	2	0	1	1	1	$2p^3(1-p)^3$
1	2	1	1	0	0	$4p^2(1-p)^4$
1	2	2	1	1	1	$2p(1-p)^5$
2	0	0	4	4	16	$p^4(1-p)^2$
2	0	1	4	1	4	$2p^3(1-p)^3$
2	0	2	4	0	0	$p^2(1-p)^4$
2	1	0	1	4	4	$2p^3(1-p)^3$
2	1	1	1	1	1	$4p^2(1-p)^4$
2	1	2	1	0	0	$2p(1-p)^5$
2	2	0	0	4	0	$p^2(1-p)^4$
2	2	1	0	1	0	$2p(1-p)^5$
2	2	2	0	0	0	$(1-p)^6$

We can summarize the table with respect to $D_{SA}D_{SB}$,

$D_{SA}D_{SB}$	Probability
0	$p^6 + 2p^5(1-p) + p^4(1-p)^2 + 2p^5(1-p) + p^4(1-p)^2 + 4p^4(1-p)^2 + 4p^4(1-p)^2 + 8p^3(1-p)^3 + 4p^2(1-p)^4 + 4p^2(1-p)^4 + p^2(1-p)^4 + 2p(1-p)^5 + p^2(1-p)^4 + 2p(1-p)^5 + (1-p)^6$
1	$4p^4(1-p)^2 + 2p^5(1-p) + 2p^3(1-p)^3 + 2p^3(1-p)^3 + 2p(1-p)^5 + 4p^2(1-p)^4$
4	$2p^3(1-p)^3 + 2p^3(1-p)^3 + 2p^3(1-p)^3 + 2p^3(1-p)^3$
16	$p^4(1-p)^2 + p^2(1-p)^4$

From this table, we obtain

$$E[D_{SA}D_{SB}] = 12p^4 - 24p^3 + 10p^2 + 2p$$

Since we know $E[D_{SA}] = E[D_{SB}] = -4p^2 + 4p$,

$$cov(D_{SA}, D_{SB}) = E[D_{SA}D_{SB}] - E[D_{SA}]E[D_{SB}]$$

$$= 12p^4 - 24p^3 + 10p^2 + 2p - (-4p^2 + 4p)^2$$

$$= -4p^4 + 8p^3 - 6p^2 + 2p$$

Again, consider that we have *N* independent loci. Let A, B, and S refer to the size-N vector of allele counts in individuals A, B, and S. Because we assume that all loci are independent, simply adding the locus-wise covariance for all N loci gives us the covariance in a triad relationship.

$$cov(D_{SA}, D_{SB}) = \sum_{n=1}^{N} - 4p_n^4 + 8p_n^3 - 6p_n^2 + 2p_n$$

Therefore, we have

$$\Sigma_{ij} = 2 * cov(D_{SA}, D_{SB})$$

$$= \sum_{n=1}^{N} 2(-4p_n^4 + 8p_n^3 - 6p_n^2 + 2p_n)$$

$$= \sum_{n=1}^{N} - 8p_n^4 + 16p_n^3 - 12p_n^2 + 4p_n$$

Now we completely specified Σ.

2.1.2. Significance threshold

There are two ways to set the significance threshold for our sample overlap test. The first is to consider repeated comparisons to minimize family-wise error rate. For example, if we test every possible pair of 1,000 samples, the threshold will be $\alpha = 0.05/\binom{1000}{2} \approx 1 \times 10^{-7}$ by the Bonferroni correction. Another way is to let our method to choose an appropriate threshold. As shown in Fig. 1f, our statistic s_{overlap} forms distinctive clusters for unrelated individuals and overlapping samples. A threshold corresponding to the middle value of these two clusters can provide an appropriate balance between specificity and sensitivity. However, this value will depend on the specific set of SNPs used for the analysis, their allele frequencies, and the reference individual dataset. Our software implementation includes a module that, given all this information, simulates a large number of pairs of unrelated and overlapping samples, figures out the mean and variance of the statistic for the two clusters, and chooses an appropriate middle value automatically.

2.2. Simulation

2.2.1. Genotype generation

For simulated genotypes, we randomly chose the reference allele frequency p_i for each locus i from the uniform distribution in the range $(0.05, 0.95)$. Given p_i, each simulated individual was assigned an allele count $g \in \{0, 1, 2\}$ drawn from $Binom(2, p_i)$, with stringent genotyping error rate ϵ. In most of our simulation analysis, we used ϵ as 0.1, or otherwise noted.

For simulating genetically related individuals, we randomly selected loci whose the ratio of their number is proportional to given degree of relation. For example, when we simulate alternative hypothesis, the ratio of shared loci between twin is 1. Therefore all loci will be selected. If simulating for 1st degree relation, 0.5 of loci will be selected. Then, we assigned genotypes generated for selected loci to both individuals. Else, we assigned seperatedly generated genotypes for loci that were not chosen.

2.2.2. Real data simulation using WTCCC

To simulate sample overlap detection with real data, we used 1,963 samples of Type 1 diabetes(T1D) cases and 1,480 samples of the 1958 British Birth Cohort(58C) controls in the Wellcome Trust Case Control Consortium (WTCCC) data [9)]. We split T1D cases and 58C controls into 3 disjoint case/control studies. Then we randomly chose individuals and added them to the other studies, such that each pair of studies would have 10~30 overlapping samples.

In the end, Study A(# of sample=1,180) and Study B(# of sample=1,158) overlapped in 25 samples, Study A(1,180) and Study C(1,158) overlapped in 13 samples, and Study B(1,158) and Study C(1,158) overlapped in 13 samples. In total, the 3 studies included 3,496 samples, when double-counting overlapping samples.

We utilized 50 randomly selected samples from the 1000Genomes data [6)] as our reference set (see "Section 6. Datasets used in the Analysis"). We pruned 1000Genomes SNPs based on physical distance (retaining those $>$ 5Mb) and applied quality control (call rate $>$ 0.95) to obtain 161,235 independent SNPs. Given 7,790 overlapping SNPs between the WTCCC data and the 1000Genomes pruned independent SNPs data. Missing alleles were imputed with random selection from $\{0,1,2\}$. We compared the total of 4,073,844 pairs of individuals across three groups. To correct multiple comparison, we used the significance threshold 1.2×10^{-8} ($\approx 0.05/4{,}073{,}844$).

2.3. Using Genetic relatedness as metric for distance

Let $X_{t,n} \in \{0,1,2\}$ be the reference allele count of individual *t* at SNP *n*. Let p_i be the reference allele frequency of locus i, which can be obtained from independent reference data or the sample data itself. Let $p_n = (\sum_{b=1}^{B} X_{b,n} + 1)/(2\mathrm{B} + 2)$ [15)], where B is the number of data used. We then standardize $X_{t,n}$ such that $\bar{X}_{t,n} = (X_{t,n} - 2p_n)/\sqrt{2p_n(1 - p_n)}$ [10)]. The genetic relatedness between individuals *t* and *u* is defined as,

$$G_{t,u} = \frac{1}{N}\sum_{n=1}^{N} \bar{X}_{t,n}\bar{X}_{u,n} \quad {}^{10)}.$$

If we use the genetic relatedness as our metric, we need a different way of obtaining the covariance matrix Σ, because it is not easy to analytically derive Σ as we have done for the squared Euclidean distance metric. Instead, we can empirically estimate Σ. Given the set of SNPs and their allele frequencies, we simulated 100,000 random pairs and calculated distance vectors to a fixed set of 20 references using 500 loci (*K*=20 and *N*=500). We estimated the covariance matrix from the 100,000 vectors of $(v_t - v_u)$. We then used this as the estimated covariance matrix Σ.

3. Principal component map construction

The core idea of our approach is to estimate the location of our samples on the PC map of reference data. The procedure begins with generating the PC map of the reference data first. The reference individuals in this map serve as “fixed anchors”. Then, we put each target individual onto this map using a distance vector. Finally, we erase the reference individuals, which leaves us the PC map of target individuals.

Consider that we have K reference individuals. We first approximate the coordinates of the K references by applying principal component analysis (PCA) to their GRM. The top two eigenvectors (PCs) from the PCA can be plotted in a two-dimensional (2D) space, $\mathcal{P}$. For each target individual, we approximate the position of the individual in $\mathcal{P}$ as follows. First, we calculate the target’s distance vector to K references based on the genetic relatedness metric. For this application, the use of genetic relatedness metric is natural because PCA is closely related to GRM. Then, we construct the GRM for the K+1 individuals (the reference set and the target) by appending the target’s distance vector to the rows and columns of the GRM of the references. We apply PCA to this size-(K+1)×(K+1) GRM to obtain a new principal component map of K+1 individuals in a new 2D space, $\mathcal{P}'$. The positions of the K references in $\mathcal{P}'$ are similar to their positions in $\mathcal{P}$ (after adjusting for rotation); however, they are not identical, because adding one more datapoint in the PCA can distort the positions of the other points (Fig. **S6**). Because of this subtle difference between $\mathcal{P}$ and $\mathcal{P}'$, although we know the target sample’s location in $\mathcal{P}'$, we need a procedure to project that point from $\mathcal{P}'$ to $\mathcal{P}$. Interestingly, we can apply another layer of “multilateration technique” to overcome this slight difference between $\mathcal{P}$ and $\mathcal{P}'$. Using the map in $\mathcal{P}'$, the 2-D Euclidean distances between the target and the references can be calculated to create a distance vector. (Note that this distance vector is not for the genetic distance used for GRM, but for the 2-D distance on the PC map). Using the standard multilateration technique, this distance vector can be used to map the approximate position of the target on $\mathcal{P}$. In particular, the least-square minimization method is used as follows. Let (x, y) be the unknown position of the target individual in $\mathcal{P}$. Let (x_k, y_k) be the coordinate of the k^{th} reference in $\mathcal{P}$. Let r_k be the distance between the sample and the k^{th} reference calculated in $\mathcal{P}'$. Let $\hat{r}_k$ be the distance between the sample and the k^{th} reference in $\mathcal{P}$. We minimize the function

$$F(x,y) = \sum_{k=1}^{K} (\hat{r}_k - r_k)^2 = \sum_{k=1}^{K} \left(\sqrt{(x - x_k)^2 + (y - y_k)^2} - r_k\right)^2$$

to approximate (x, y) in $\mathcal{P}$. This procedure is repeated for each target individual. After all repetitions, the approximated PC map of target individuals is obtained by removing reference data from the plot.

3.1. Reconstructing spatial map of POPRES

3.1.1. POPRES as a reference set

We used 40% from each population of POPRES data as the reference individuals and rest 60% as our target sampls. This splitting gave us 572 reference individuals and 815 target samples. Missing genotypes were assigned with value $2p_i$, where p_i was the allele frequency inferred from the reference individuals. A total of 197,146 variants were used to calculate genetic relatedness between any two individuals. For this application, we did not prune SNPs, similar to Novembre et al. [11)].

3.1.2. 1000Genomes as a reference set

Out of 1,092 samples in the 1000Genomes phase I dataset [6)], we selected 305 European samples from populations TSI, GBR, and IBS. We excluded CEU because the population was from USA, and FIN which was underrepresented in POPRES. We used all 1,387 POPRES samples as our target samples. Because the reference and target samples were independent datasets, we screened for variants shared between the two datasets, matched strands and matched reference alleles. Missing genotypes were assigned with value $2p_i$, where p_i was the allele frequency inferred from the reference individuals. A total of 196,350 shared variants were used to calculate genetic relatedness to generate distance vector.

3.1.3. Map rotation

For this real data analysis, we rotated the maps to facilitate the visual comparison to the physical map of the Europe. For this purpose, we used the following method described in

Novembre et al. [11]. We searched for the rotation angle θ which maximizes the function:

$$f(\theta) = Cor(lat, x'(\theta, v_1, v_2)) + Cor(long, y'(\theta, v_1, v_2))$$

where Cor is the correlation function, lat and $long$ are the vectors of the latitude and longitude of each reference according to their geographic origin, $x'$and y' are functions for 2D rotation about a point, and v_1 and v_2 are the approximated coordinates of the samples. Specifically, when a point located at (x, y) is rotated about the origin with a rotation angle of θ, the new location of the point (x', y') will be

$$x' = xcos\theta - ysin\theta$$

$$y' = ycos\theta + xsin\theta$$

3.2. Estimating ancestry proportion

Another population genomic analysis application that distance vectors can be used for is ancestry estimation. We developed a method to use distance vectors to infer the ancestry composition of an individual. Our method works in a supervised way, requiring the reference data of candidate populations. The method is built upon the PC map construction described in section 3. The idea is to approximate the location of a target individual in the PC map of the reference populations. Then, we measure the Euclidean distance of the individual to the centroid of each population in the PC map. The ancestry proportion is estimated as being inversely proportional to these distances. For example, if we use the CEU and YRI of 1000Genomes as our reference, and the distance to each population was 10 and 5, then the ancestry proportion is estimated as 1/10:1/5=1:2. If we use two candidate populations, we use the 1 dimensional Euclidean distance (difference in PC1) (Fig. 13). If we use more than two candidate populations, we use the 2 dimensional Euclidean distance in the PC1-PC2 map (Fig. 14).

3.2.1. Admixed individual generation

We used Hapgen2 software [16] to generate haplotypes of admixed individuals with multiple ancestries. For each population, we randomly selected 50 samples to be use as ancestry pools.

The rest of the population data was used as the reference to calculate distances to. For admixed haplotype generation, the ratio to the desired admixture proportion were sampled from ancestry pools, then submitted to Hapgen2.

3.2.2. ADMIXTURE

For comparison to our method, we used ADMIXTURE [13)], which also decompose individual ancestries from genotype dataset using maximum likelihood estimation. Different from our method, ADMIXTURE calculates ancestry directly from genotype data of individuals. Notwithstanding prerequisite genotype data, it runs in an unsupervised way without need of training or reference data.

We used same genotype data for distance vector generation of our method and ADMIXTURE. We placed genotypes for target individuals as well as the references together as input and ran in a semi-supervised fashion by giving the correct number of populations as input (K=3) This way, the reference individuals from the three populations form three distinct clusters in the ADMIXTURE algorithm so that the target individual can be proportionally assigned to each cluster. Again, to avoid data re-use, we used 50 samples from each reference population (CHS, GBR, and YRI) for data generation using Hapgen2 and used the rest as the reference data for our method or for ADMIXTURE

Table 2. Population distribution of the POPRES dataset. The abbreviated population code was used in Fig. 8-10

Geoscheme for Europe	Population name	Population code	Total	Population total
Eastern Europe	Bulgaria	BG	2	
	The Czech Republic	CZ	11	
	Hungary	HU	19	
	Poland	PL	22	
	Romania	RO	14	
	Russia	RU	6	
	Slovakia	SK	1	
	Ukraine	UA	1	
	Turkey	TR	4	
	Cyprus	CY	4	84
Southern Europe	Albania	AL	3	
	Bosnia	BA	9	
	Montenegro	YG	41	
	Kosovo	KS	2	
	Macedonia	MK	4	
	Servia	RS	3	
	Croatia	HR	8	
	Greece	GR	8	
	Italy	IT	219	
	Portugal	PT	128	
	Slovenia	SI	2	
	Spain	ES	136	563
Western Europe	Austria	AT	14	
	Belgium	BE	43	
	France	FR	91	
	Germany	DE	71	
	Netherlands	NL	17	
	Switzerland	CH	222	458
Northern Europe	Denmark	DK	1	
	Finland	FI	1	
	Ireland	IE	61	
	Latvia	LV	1	

Table 2. (Continue)

	Norway	NO	3	
	Sweden	SE	10	
	United Kingdom	UK	200	
	UK-Scotland	Sct	5	282
Total				1387

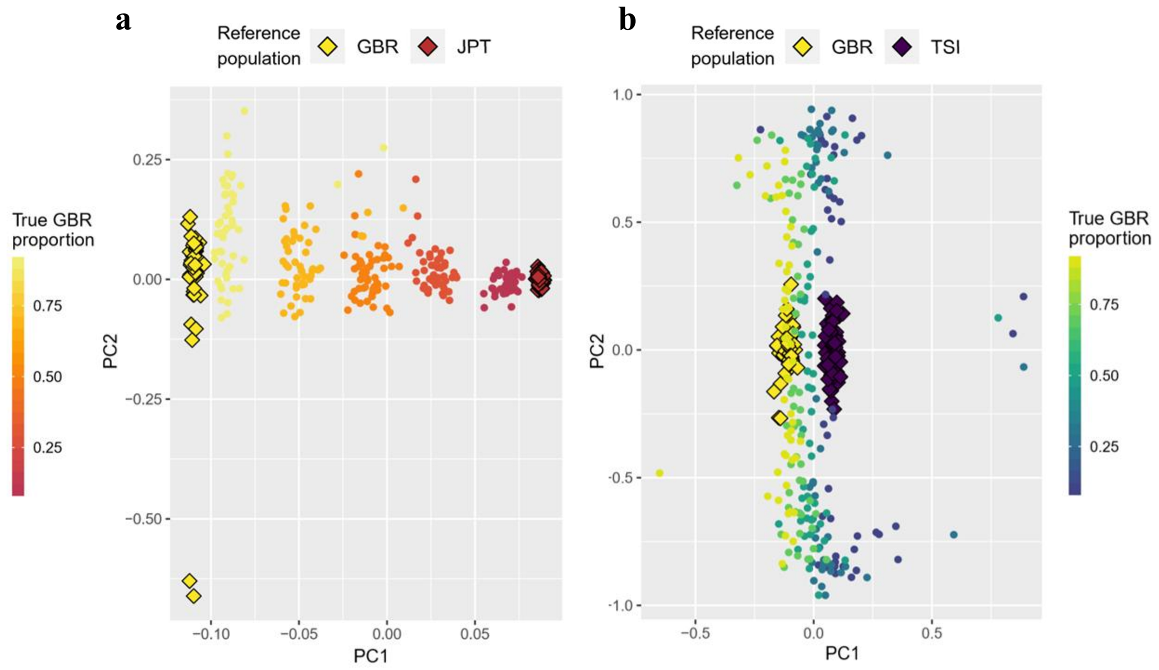


Fig. 13. Two dimensional PC map of the admixed samples as well as references used in the two-population ancestry estimation.

Given the two reference populations, we used the subset of each population to generate the admixed samples and used the rest as the reference data for our method (denoted as diamonds here). To estimate the ancestry proportion of a sample, the PC1 distance was measured from the sample to each reference cluster, of which the inverse proportions were our estimates. **a**, The two populations were distant (GBR and JPT). **b**, The two populations were close (GBR and TSI).

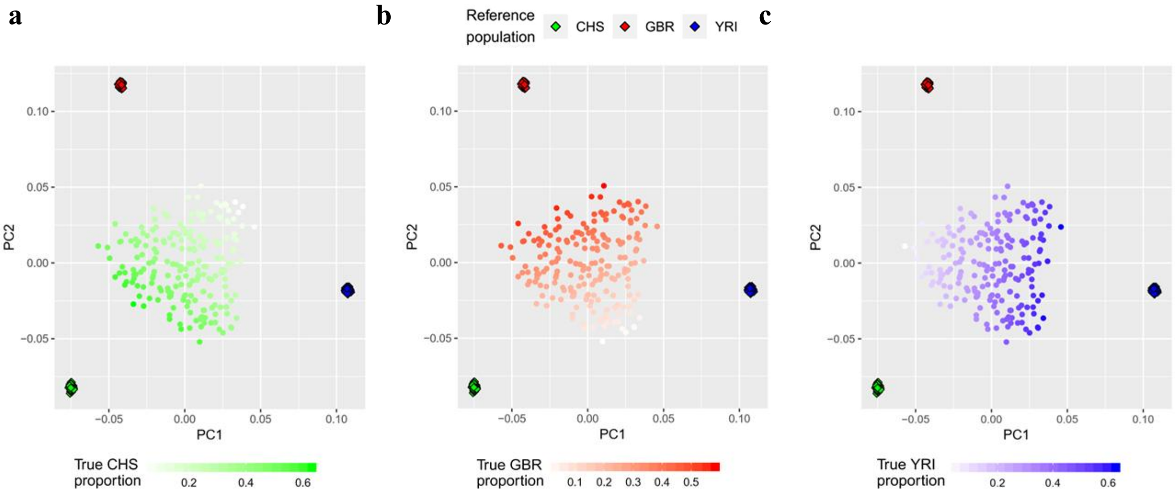


Fig. 14. Two dimensional PC map of the admixed samples as well as references used in the three-population ancestry estimation

Given the three reference populations, we used the subset of each population to generate the admixed samples and used the rest as the reference data for our method (denoted as diamonds here). To estimate the ancestry proportion of a sample, the Euclidean distance in the PC1-PC2 space was measured from the sample to each reference cluster, of which the inverse proportions were our estimates. **a**, True proportion of CHS ancestry is colored. **b**, True proportion of GBR ancestry is colored. **c**, True proportion of YRI ancestry is colored.

4. Unidentifiability

4.1. Mathematical proofs

We present several mathematical proofs of unidentifiability that a person's distance vector does not identify genotypes the person. These proofs make a general assumption that the coordinates of a point are real-valued. This assumption is appropriate for typical situations that multilateration is applied (e.g. aircraft navigation).

Condition under which point is identifiable

We prove that under certain conditions, the target point can be uniquely determined. Let $p_0, \ldots, p_k$ be $k+1$ distinct points in the n-dimensional Euclidean space $\mathbb{R}^n$. Assume a point $x \in \mathbb{R}^n$ whose precise location is not known but the distance $r_i := d(x, p_i)$ for each i is known. Note that if $r_i + r_j < d(p_i, p_j)$ for some $i \neq j$, then no x satisfies the given data. Also if $r_i + r_j = d(p_i, p_j)$ for some $i \neq j$, then the possible position of x is already uniquely determined by only p_i, p_j. This means that the interesting case to analyze is when $r_i + r_j > d(p_i, p_j)$ for all $i \neq j$. Therefore, we suppose the following assumption.

$$r_i + r_j > d(p_i, p_j), \forall i \neq j.$$

An *affine hyperplane* of $\mathbb{R}^n$ is an $(n-1)$-dimensional affine subspace of co-dimension 1. In other words, one can think of a linear subspace of $\mathbb{R}^n$ of dimension $n-1$ and take a parallel copy by translating the subspace along the normal direction. An $(n-1)$-dimensional space obtained this way, we call it affine hyperplane or shortly just hyperplane (since this is the only type of hyperplane we will consider here).

For $p \in \mathbb{R}^n$ and $R > 0$, let $S_R(p)$ denote the set of points in $\mathbb{R}^n$ whose distance from p is precisely R. It is clear that $S_R(p)$ is a $(n-1)$-dimensional sphere with radius R and center p.

For each i, let S_i denote the sphere $S_{r_i}(p_i)$. By definition, x lies in the intersection of these sets S_i, i.e., $x \in \cap_{i=0}^{n} S_i$. Note that for any $i \neq j$, S_i is different from S_j. In a special case when S_i and S_j are tangent to each other, i.e, the intersection is just one point, x is already uniquely determined. Hence, we assume it is not the case. In this case, $S_i \cap S_j$ lies on the unique hyperplane $H_{i,j}$.

Now we prove the following claim:

Proposition 1. *If* $\mathrm{k} = \mathrm{n}$ *and* $\mathrm{p}_0, \ldots, \mathrm{p}_\mathrm{n}$ *are affinely independent (i.e.,* $\mathrm{p}_1 - \mathrm{p}_0, \mathrm{p}_2 - \mathrm{p}_0, \ldots, \mathrm{p}_\mathrm{n} - \mathrm{p}_0$ *are linearly independent), then the position of* x *is uniquely determi ned.*

Proof.

For each $i = 0, \ldots, n$, say H_i is the unique hyperplane containing the intersection between S_0 and S_i. It is clear that the vector $u_i := p_i - p_0$ is a normal vector to H_i. Since $x \in \cap_{i=0}^{n} S_i$, we also have $x \in \cap_{i=1}^{n} H_i$. Now it suffices to show that $\cap_{i=1}^{n} H_i$ has dimension 0, i.e., just a single point.

We first consider the following general lemma.

Lemma 1. *Suppose that there are* m *hyperplanes in* $\mathbb{R}^\mathrm{n}$ *whose normal vectors are* v_i *and their intersection is nonempty. Then the intersection of those hyperplanes has dimension of at least* $\mathrm{n} - \mathrm{m}$*, and exactly* $\mathrm{n} - \mathrm{m}$ *when* v_i *are linearly independent.*

Let $W_1, \ldots, W_m$ be affine subspaces in $\mathbb{R}^n$ where $\dim W_i = d_i$, and $\dim(W_1 \cap \ldots \cap W_\ell) = i_\ell$. Also let $W = W_1 \cap \ldots \cap W_m$. Since we are assuming W is nonempty, $W_1 \cap \ldots \cap W_\ell$ is nonempty for all ℓ.

Note that $\dim S + \dim T - \dim(S \cap T) \leq n$ for any affine subspaces S and T of $\mathbb{R}^n$. Applying this to W_1 and W_2, we have $d_1 + d_2 - i_2 \leq n$, hence $i_2 \geq d_1 + d_2 - n$. Applying this again to $W_1 \cap W_2$ and W_3, one gets $i_2 + d_3 - i_3 \leq n$, i.e., $i_3 \geq i_2 + d_3 - n \geq d_1 + d_2 + d_3 - 2n$. By induction, one can easily show that $\dim W = i_m \geq d_1 + \cdots + d_m - (m-1)n$.

Because W_i are hyperplanes, $\dim W_i = n - 1$ for all i. Hence, $\dim W \geq m(n-1) - (m-1)n = n - m$. Thus, we proved that the intersection has dimension of at least $n - m$.

Now let v_i be a normal vector to W_i for each i, and assume that they are linearly independent. For any vector w contained in W, w is orthogonal to each v_i. Thus, $v_1, \ldots, v_m$ are independent vectors in $W^{\perp}$, which gives us $\dim W + m \leq n$. Since we already have the inequality in the other direction, this implies that $\dim W$ is precisely $n - m$. This completes the proof of *Lemma 1.* □

Now, let's come back to the proposition. Since $u_1, \ldots, u_n$ are linearly independent, $\dim H$ is precisely $n - n = 0$ when $H = \cap_{i=1}^{n} H_i$. Hence, it is just a single point. This proves *Proposition 1.* □

Summary: This proof shows that, in an n-dimensional space, if we have $n + 1$ reference datapoints to measure distances to, the target point can be exactly specified in general. For example, in the 3-dimensional space, we will need 4 satellites to specify the location of an aircraft. (In practice, 3 satellites can suffice because of the additional condition that the aircraft has a lower altitude than satellites.)

Condition under which point is unidentifiable

We prove that under certain conditions, the target point cannot be uniquely determined. We argue that if $n \geq k + 1$, the exact location of x cannot be identified.

Since $x \in S := \cap_{i=0}^{k} S_i$, we want to obtain the lower bound of $dimS$. Given $k + 1$ points in $\mathbb{R}^n$, we can consider every possible pair of points. Each pair of $k + 1$ points gives us a hyperplane where x must belong. Imagine that we have an intersection H' of all $\binom{k+1}{2}$ hyperplanes. $S = \cap_{i=0}^{k} S_i$ will be a sphere that resides in H' and will have one less dimension than H'. This remark about the dimension of S compared with the dimension of H' will be justfied after we first explain how to understand H' below.

First of all, what is the dimension of H'? It turns out that to obtain the intersection H', it suffices to consider only k hyperplanes. Write our $k + 1$ points as $p_0, \ldots, p_k$ and define $u_i = p_i - p_0$ for each i as before. Then the hyperplane determined by $p_i \ and \ p_j$ has $p_i - p_j$ as a normal vector, but this vector is already in the span of u_i and u_j. This means when we intersect $H_i \cap H_j$ with the hyperplane plane determined by p_i and p_j, the dimension does not go down, i.e., $H_i \cap H_j$ is contained in the hyperplane determined by p_i and p_j. By definition, $H_i \cap H_j$ is precisely the set of vectors which is orthogonal to both u_i and u_j, hence they are already orthogonal to $u_i - u_j$. From this, we conclude that the intersection of $\binom{k+1}{2}$ hyperplanes obtained from the $k+1$ given points is the same as the intersection of H_i's we considered in the proof of the *Proposition 1.*

Now note that by definition, H_1 is the smallest affine subspace containing the sphere $S_0 \cap S_1$, hence the dimension of $S_0 \cap S_1$ is one lower than the dimension of H_1. Recall that we are working on the mild assumption that $r_i + r_j > d(p_i, p_j), \forall i \neq j$. This is equivalent to saying that the sphere S_i and S_j intersect non-tangentially (or, transversally). In practice, it is also safe to assume further that the spheres we get as the intersection of S_i's also intersect transversally. Not only such a condition is satisfied with probability 1, also it is always guaranteed if we perturb our data by introducing an error term as in the next section (lattice case). Under this assumption, the spheres $S_0 \cap S_1$ and $S_0 \cap S_2$ intersect transversally, hence $H_1 \cap H_2$ is the smallest affine subspace containing the sphere $(S_0 \cap S_1) \cap (S_0 \cap S_2) = S_0 \cap S_1 \cap S_2$. By induction, we conclude that $\cap_{i=1}^{k} H_i$ is the smallest affine subspace containing $\cap_{i=0}^{k} S_i$. We just showed above that $H' = \cap_{i=1}^{k} H_i$, hence we know that $\cap_{i=0}^{k} S_i$ has one less dimension than H'.

Recall that by Lemma 1 the intersection of all H_i's has dimension of at least $n-k$ provided that $n-k \geq 0$ and the intersection is nonempty (which follows from our assumption that x exists). Note that $x \in \cap_{i=0}^{k} S_i \subset \cap_{i=1}^{k} H_i$. Since $\cap_{i=0}^{k} S_i$ is a $(n-k-1)$-dimensional sphere in the affine space $\cap_{i=1}^{k} H_i$ of dimension $n-k$, as long as $n-k \geq 1$, the set of possible locations of x is more than one point. Hence, the final conclusion is that if $n \geq k+1$, either x cannot exist or x exists but its location cannot be uniquely determined under the assumption that $r_i + r_j > d(p_i, p_j)$ for all $i \neq j$.

Lattice case

The proofs above depend on the assumption that the coordinates are real-valued. If the points are in a more constrained space, the conditions for identifiability and unidentifiability may change. Here, we consider a lattice space (e.g. all integers). Assume that our data set $\{p_0, \ldots, p_k, x\}$ is contained in some lattice L in $\mathbb{R}^n$. Let H be as in the previous section. As we saw before, $dimH \geq n-k$. Hence, if $n-k>0$, H itself does not determine the exact location of x.

The set of possible locations of x is now $H \cap L$. So, as long as $n-k<n$, i.e., $k>0$, H may intersect L only at one point. This causes a problem for our method, since when $H \cap L$ is a single point, the exact location of x is completely determined, and this phenomenon happens very often. In all practical purposes, it is enough to assume that L is the integral lattice $\mathbb{Z}^n$. We will assume this for the rest of this section.

Mathematically speaking, a generic (randomly chosen) affine subspace will miss the lattice L. As an instructive example, we describe the situation in dimension 2, i.e, let's consider $\mathbb{R}^2$ with the integral lattice $\mathbb{Z}^2$. For a straight line in $\mathbb{R}^2$ to intersect more than one point in $\mathbb{Z}^2$, it must have a rational slope. On the other hand, the set $\mathbb{Q}$ of rational numbers has Lebesgue measure zero as a subset of $\mathbb{R}$ (said differently, a randomly chosen real number is irrational with probability one). Hence, a random straight line would intersect $\mathbb{Z}^2$ at most one point.

Here is our suggestive solution for the above problem; instead of using $r_i = d(x, p_i)$, we perturb the given data by introducing an error term. More precisely, choose a small positive number ϵ and define $d_{i'} = r_i + \epsilon_i$ where ϵ_i is a number chosen randomly in the interval $(-\epsilon, \epsilon)$. The actual values of r_i and ϵ_i are hidden and only the value of $d_{i'}$ is given to the user of the data set. Then H_i is replaced by the ϵ_i-neighborhood of H_i, and at the end H is replaced by μ-neighborhood of H, call it H_μ where $\mu = \min\{\epsilon_i\}$. If one can show $H_\mu \cap L$ contains infinitely many points, then we can overcome the problem we described above.

Obviously, if μ is arbitrarily small, this still does not hold. Fortunately, the following mathematical statement is true for an obvious reason. There exists $\epsilon_0 > 0$ such that as long as $\mu \geq \epsilon_0$, then $H_\mu \cap L$ contains infinitely many points. Hence, one can take a positive number ϵ bigger than ϵ_0, and choose each ϵ_i in the set $(-\epsilon, -\epsilon_0) \cup (\epsilon_0, \epsilon)$. For instance, in $\mathbb{R}^n$, it would be enough to take ϵ_0 to be $\sqrt{N}$. But this choice of ϵ_0 is quite large, and one might want to choose ϵ_0 as small as possible. While we do not fully resolve this optimization problem, we note that a very famous theorem of Hurwitz [17)] implies that one can take ϵ_0 to be $1/\sqrt{5}$ in dimension 2, so it is plausible that one might be able to take a quite small number as ϵ_0 in more general case.

Now we explain the situation in dimension 2 with more details. Say we choose a line with irrational slope r. Without loss of generality, let's assume this line actually passes through the origin and call it P. Then P is the set of solutions of the equation $y = rx$. If $(a, b) \in \mathbb{Z}^2$ is contained in the region B bounded by two lines $y = rx + \epsilon$ and $y = rx - \epsilon$, then surely (a, b) lies in the $\epsilon -$neighborhood of P. On the other hand, $(a, b) \in B$ is equivalent to $|ar - b| < \epsilon$. Alternatively, one can write $|r - \frac{b}{a}| < \frac{\epsilon}{a}$. Hurwitz's theorem says $|r - \frac{b}{a}| < \frac{1}{\sqrt{5}a^2}$ is satisfied by infinitely many $(a, b) \in \mathbb{Z}^2$. In particular, $(1/\sqrt{5})$-neighborhood of P intersects $\mathbb{Z}^2$ at infinitely many points. Hence, in dimension 2, it is enough to take ϵ_0 to be $1/\sqrt{5}$.

4.2. Simulation through greedy algorithm

Although we provided a mathematical proof about the unidentifiability condition, that only applies to the spaces of real values. The lattice space case is not directly relevant to our context, because the allele count in genome data is in a highly constrained integer space, $\{0,1,2\}^n$. Therefore, we performed extensive simulations to show that in practice, the original genome data is not recoverable given distance vector information.

To this end, we designed a greedy algorithm that aims to recover the original genome.

Algorithm 1: Greedy algorithm

```
INPUT: Reference data $R$, distance vector $v$, and allele frequencies of the SNPs $\{p_i\}$
OUTPUT: A prediction for target sample $t$
REPEAT 1,000 times
    Randomly generate a sample $t$' based on $\{p_i\}$
```

```
        Calculate distance vector $v$’ between $t$’ and $R$
        Calculate sum of squared error (SSE) between $v$’ and $v$
        WHILE True:
            Randomly select a SNP $j$
            Try changing SNP $j$ to two alternating forms (e.g. 0 → 1, 0 → 2)
            If an alternating form reduces SSE, accept the change
            If no change reduces SSE for a long period (1,000 trials), break loop
        Record solution $t$’
    END REPEAT
    Choose the best solution with the minimum SSE among 1,000 repeats.
```

Starting from the random genotypes, the algorithm stepwisely moves the space to find the solution whose distance vector most resembles the input distance vector. The algorithm quits if no improvement is obtained for a prolonged time. To avoid local optimum, the algorithm restarts with a new starting point 1,000 times and chooses the best solution at the final step.

We assumed that we have 1,000 SNPs. We randomly decided the frequencies of the SNPs from the uniform distribution (0.3, 0.7). We generated 30 reference individuals (R) and one target individual t based on the frequencies. We calculated the distance vector v (squared Euclidean distances) between R and t.

We measured accuracy as the proportion of the correct allele count (0/1/2) in the data. Let $\left[X_{ij}\right]$ be 1000×1000 matrix where X_{ij} is 1 if the ith solution’s jth SNP is correct and 0 otherwise. To obtain the vector of per-individual accuracy, we averaged each row of this matrix. To obtain the vector of per-SNP accuracy, we averaged each column of this matrix.

5. Datasets used in the analysis

5.1. POPRES data

The samples we used in the PC map analysis were taken from the POPRES data set, which includes nearly 6,000 subjects of African-American, East Asian, South Asian, Mexican, and European origin [12]. The data were genotyped using the Affymetrix 500K SNP panel. For our analysis, we used a subsample of European individuals from the London Life Sciences Population (LOLIPOP) study [18], which comprises mainly European individuals sampled in London, and from the CoLaus study [19], which includes a broad set of European individuals sampled from Lausanne, Switzerland. **Table S2** summarizes the population distribution for the 1,387 individuals used in the final sample. POPRES data is accessible via dbGaP Study accession number phs000145.v4.p2 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v4.p2)

5.2. WTCCC data

The WTCCC dataset includes ~3,000 shared controls and ~2,000 cases for each of seven diseases (including type 1 and type 2 diabetes, and Crohn's disease), with a total of ~14,000 cases on genotypes for up to 500,000 SNPs [9]. Access to WTCCC data is provided at European Genome-phenome Archive (EGA) with accession number EGAD00000000001 for 1958 British Birth Cohort and EGAD00000000008 for Type 1 Diabetes (T1D) samples (https://www.ebi.ac.uk/ega/datasets). The data were genotyped using the Affymetrix 500K SNP panel. We used ~1,500 controls from the 1958 British Birth Cohort and ~2,000 cases of type 1 diabetes. After a quality control process, 1,480 controls and 1,963 cases were used for our sample overlap detection simulation.

5.3. 1000Genomes data

1000Genomes data [6] were used for our real data simulations. We used the phase 1 dataset which comprised whole genome sequencing and exome sequencing of 1,092 samples from 14 populations. The genotype data for 1000Genomes Phase 1 was downloaded from the webpage of PLINK 1.9 (http://www.cog-genomics.org/plink/1.9/resources). For the real-data-based simulation for sample overlap detection, 50 randomly selected samples were used as reference set. For the PC map analysis, 201 samples of the British (GBR), Tuscan (TSI), and Spanish (IBS) populations were used as a reference set to infer the spatial structure of the POPRES data. For ancestry estimation analysis, we used the phase 3 dataset of which the haplotype data were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.In the simulation of admixed samples from two populations, we used 302 samples of British (GBR), Tuscan (TSI), and Japanese (JPT). In the simulation of admixed samples from three populations, we used 304 samples of British (GBR), Chinese (CHS), and African (YRI).

DISCUSSION

Here we presented a method that aims to achieve a balance between data sharing and privacy protection. It allows sharing of information to a degree sufficient for approximating relative genetic distance of an individual from either another individual or a group. Identification of closer relatives and population genomic analyses, such as ancestry decomposition and geographical origin mapping, are possible. Importantly, though, the shared information conceals individual genotypes, making it extremely difficult to reconstruct the personal genomes.

Our method is subject to several limitations. First, distance vector equivocally distinguishes relation distant than 2nd degree. In Fig. 6, both precision and recall rate of 2nd degree relation and 3rd degree relation were below 40%. The result shows the distance vector does not have sufficient power to detect distant relationship. However, such a disclosure could be considered a leak of information in some situations. In those situations, alternatives such as fuzzy encryption of He et al. [4] can be considered for detecting sample overlaps. Second, our method relies on the composition of reference samples to reconstruct PC map of samples. The reconstructed map of samples is obtained by estimating the position of each samples on the PC map of references. Hence, choice of reference set determines the result of reconstructed map of samples. For example, in the two-dimensional mapping of the Europeans in the POPRES data using distance vectors, if we have used reference set consisted of Asian, European, and African samples instead of intra-European samples, the geographic resemblance to the result would not have been clearly distinctive as in Fig. 8. Finally, our method works in a supervised way, requiring the reference data of candidate populations. Since ancestry estimation was based on estimating the position of each samples on the PC map of references, inaccurate selection of candidate populations may lead to spurious estimation. However, the comparison between results of a few trial-and-errors using differently comprised set of references can give a rough estimation of composition of samples.

In conclusion, we have presented a novel technique that applies multilateration to genomic data. Our method allows sharing distance vectors with other investigators or institutions, enabling certain types of genomic analysis while making it difficult to reconstruct the personal genomes. We expect that our approach will find interesting applications in the future in addition to those described herein.

REFERENCES

1. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J-H. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nature Genetics. 2013;45:400.

2. Park J-H, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature Genetics. 2010;42(7):570-5.

3. Kim EE, Lee S, Lee CH, Oh H, Song K, Han B. FOLD: a method to optimize power in meta-analysis of genetic association studies with overlapping subjects. Bioinformatics (Oxford, England). 2017;33(24):3947-54.

4. He D, Furlotte NA, Hormozdiari F, Joo JWJ, Wadia A, Ostrovsky R, et al. Identifying genetic relatives without compromising privacy. Genome research. 2014;24(4):664-72.

5. Chen GB, Lee SH, Robinson MR, Trzaskowski M, Zhu ZX, Winkler TW, et al. Across-cohort QC analyses of GWAS summary statistics from complex traits. Eur J Hum Genet. 2016;25(1):137-46.

6. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature: The Author(s); 2015. p. 68–74.

7. The International HapMap Consortium. A haplotype map of the human genome. Nature2005. p. 1299-320.

8. Mantilla-Gaviria I, Leonardi M, Galati G, Balbastre J. Localization algorithms for multilateration (MLAT) systems in airport surface surveillance. Signal, Image and Video Processing. 2014;9:1-10.

9. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. In: Consortium TWTCC, editor. Nature2007. p. 661-78.

10. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. American journal of human genetics. 2011;88(1):76-82.

11. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456(7218):98-101.

12. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. American journal of human genetics. 2008;83(3):347-58.

13. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009;19(9):1655-64.

14. National Institute of Standards and Technology. FIPS 180-2: Secure Hash Standard. Federal Information Processing Standards Publication 180-2, U.S. Department of Commerce; 2002.

15. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS genetics. 2006;2(12):e190-e.

16. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics (Oxford, England). 2011;27(16):2304-5.

17. Hurwitz A. Ueber die angenäherte Darstellung der Irrationalzahlen durch rationale Brüche. Mathematische Annalen. 1891;39(2):279-84.

18. Kooner JS, Chambers JC, Aguilar-Salinas CA, Hinds DA, Hyde CL, Warnes GR, et al. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. Nature Genetics. 2008;40(2):149-51.

19. Firmann M, Mayor V, Vidal PM, Bochud M, Pécoud A, Hayoz D, et al. The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC cardiovascular disorders. 2008;8:6.

국문요약

저효율, 고비용으로 생산되던 유전자정보는 대용량 데이터 처리 기술의 발전과 가격 하락으로 인해 높은 효율과 저비용으로 대량 생산이 가능해졌다. 이러한 대량의 유전자정보는 이전에는 찾지 못했던 질환에 대한 새로운 원인유전변이를 보여줄 수 있을 것으로 기대를 모았다. 그러나 한 기관에서 새로운 원인유전변이를 발견할 수 있을 만큼 많은 수의 표본을 모으는 것은 여전히 쉽지 않으므로 기관간의 협력을 통한 연구가 필수적이다. 유전자 정보 또한 정보 제공자의 개인정보로 취급되기 때문에 현재 대다수의 기관은 협력 연구를 요청하는 기관에게 원 유전자 정보 대신 각 마커에 대한 요약 정보만을 제공하고 있다. 기관간에 공유되는 요약 정보는 원 유전자 정보에 비해 굉장히 한정적인 정보를 제공한다. 개개인에 대한 유전자형이 요구되는 연구방법을 적용할 수 없을 뿐더러 기관 간의 데이터를 표본레벨에서 비교 분석하기에도 어렵다. 현재 요약 정보를 더 다양한 유전체 연구에 활용할 수 있도록 하는 연구들이 많이 진행 되었지만 대다수의 연구들이 한가지의 활용에 그치는 성과를 보였다.

이 연구에서 우리는 개인의 유전자 정보를 암호화하여 마커에 대한 요약 정보만으로는 진행할 수 없었던 분석들이 가능함을 보이려고 한다. 자동차나 항공 네비게이션에서 일반적으로 사용되는 ‘다변 측정’이라는 기술을 사용하여 유전자 정보를 암호화하였다. 일반적으로 GPS 네비게이션에서는 사물과 인공위성 간의 거리를 측정하여 사물의 위치를 특정한다. 이 개념을 사람의 유전자정보에 도입하여 사람 간 유전적 거리를 측정하여 거리 정보로 암호화를 하고 이 암호화된 정보만을 공유하도록 하여 다기관 협력 연구를 촉진 시킬 수 있는 알고리즘을 개발하였다.

이 알고리즘을 실제 유전체 데이터에 적용하는 실험을 진행한 결과, 정보를 암호화한 상태로 기관 간에 공유했을 때 다양한 유전체 연구가 실제로 가능하다는 것을 증명할 수 있었다. 이번 연구에서는 암호화 된 거리 정보를 이용해 데이터 간에 중복되는 샘플을 찾아내고, 가까운 친척관계에 있는 샘플을 찾을 수 있었다. 또한 혼혈인의 혼혈 비율을 유추하거나, 유럽인들의 유전자 지도를 만드는 등의 인류유전학적 연구 결과에서도 암호화된 정보는 비암호화된 정보와 거의 비슷한 정확도를 보여주었다.