Supplementary Information for "Privacy-Aware Estimation of Relatedness in Admixed Populations"

Su Wang[1], Miran Kim[2], Wentao Li[3], Xiaoqian Jiang[3], Han Chen[1,4], Arif Harmanci[1,*]

[1] Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA.
[2] Department of Computer Science and Engineering and Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology, Ulsan, 44919, Republic of Korea.
[3] Center for Secure Artificial intelligence For hEalthcare (SAFE), School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, 77030, USA.
[4] Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA.
[*] Corresponding Author

On Distance and Correlation-based Kinship Estimators

We discuss the motivation behind the formulations of the distance and correlation-based kinship estimators. For the sake of simplicity, we first present the formulations using homogeneous ancestry and discuss extension to the heterogeneous case.

Distance-based Kinship Estimator. The distance-based estimator utilizes the expectation of the distance between the genotypes of the query subjects i and j over the variants, i.e.,

$$E_k\left(\left(G_{i,k}^{(q)} - G_{j,k}^{(q)}\right)^2\right)$$

where expectation $E_k(\cdot)$ indicates that the expectation is computed over the variants. This expectation can be written conditional on the IBD sharing state:

$$E_k\left(\left(G_{i,k}^{(q)} - G_{j,k}^{(q)}\right)^2\right) = P(IBD_{i,j} = 0) \times E_k\left(\left(G_{i,k}^{(q)} - G_{j,k}^{(q)}\right)^2 \mid IBD_{i,j} = 0\right) +$$

$$P(IBD_{i,j} = 1) \times E_k\left(\left(G_{i,k}^{(q)} - G_{j,k}^{(q)}\right)^2 \mid IBD_{i,j} = 1\right) +$$

$$P(IBD_{i,j} = 2) \times E_k\left(\left(G_{i,k}^{(q)} - G_{j,k}^{(q)}\right)^2 \mid IBD_{i,j} = 2\right)$$

An important observation is that the expected genotype distances conditioned on the IBD sharing events can be formulated as:

$$E_k\left(\left(G_{i,k}^{(q)} - G_{j,k}^{(q)}\right)^2 \mid IBD_{i,j} = 0\right) = E_k\left(4 \times \mu_k \times (1 - \mu_k)\right)$$

$$E_k\left(\left(G_{i,k}^{(q)}-G_{j,k}^{(q)}\right)^2 \mid IBD_{i,j}=1\right)=E_k\left(2\times\mu_k\times(1-\mu_k)\right)$$

$$E_k\left(\left(G_{i,k}^{(q)}-G_{j,k}^{(q)}\right)^2 \mid IBD_{i,j}=2\right)=0$$

Above, μ_k denotes the alternate allele frequency of the k^{th} variant and the expectation of the mean allele frequencies is computed over the variants. These relationships can be derived using the probabilities conditioned on IBD state. For example, given k^{th} variant with allele frequency μ_k, the expected genotype distance can be written as:

$$E\left(\left(G_{i,k}^{(q)}-G_{j,k}^{(q)}\right)^2 \mid IBD_{i,j}=0\right)=2^2\times P\left(G_{i,k}^{(q)}=0,G_{i,k}^{(q)}=2 \mid IBD_{i,j}=0\right)+$$

$$2^2\times P\left(G_{i,k}^{(q)}=2,G_{i,k}^{(q)}=0 \mid IBD_{i,j}=0\right)+1^2\times P\left(G_{i,k}^{(q)}=0,G_{i,k}^{(q)}=1 \mid IBD_{i,j}=0\right)+$$

$$1^2\times P\left(G_{i,k}^{(q)}=1,G_{i,k}^{(q)}=0 \mid IBD_{i,j}=0\right)+1^2\times P\left(G_{i,k}^{(q)}=1,G_{i,k}^{(q)}=2 \mid IBD_{i,j}=0\right)+$$

$$1^2\times P\left(G_{i,k}^{(q)}=2,G_{i,k}^{(q)}=1 \mid IBD_{i,j}=0\right)$$

Replacing the genotype probabilities conditioned on $IBD_{i,j}=0$, into above, we get:

$$E\left(\left(G_{i,k}^{(q)}-G_{j,k}^{(q)}\right)^2 \mid IBD_{i,j}=0\right)=8\mu_k^2\times(1-\mu_k)^2+4\mu_k(1-\mu_k)^3+4\mu_k^3(1-\mu_k)$$
$$=4\mu_k(1-\mu_k)\left(\mu_k^2+2\mu_k(1-\mu_k)+(1-\mu_k)^2\right)=4\mu_k(1-\mu_k)$$

Other conditional expectations can be derived using a similar formulation as above. Plugging these into the formulation of $E_k\left(\left(G_{i,k}^{(q)}-G_{j,k}^{(q)}\right)^2\right)$, we get

$$E_k\left(\left(G_{i,k}^{(q)}-G_{j,k}^{(q)}\right)^2\right)=4\times E\left(\mu\times(1-\mu)\right)\times\delta_{i,j}^0+2\times E\left(\mu\times(1-\mu)\right)\times\delta_{i,j}^1$$

where we used $P\left(IBD_{i,j}=0\right)=\delta_{i,j}^0$ and $P\left(IBD_{i,j}=1\right)=\delta_{i,j}^1$. Rearranging the terms, we get

$$\frac{E_k\left(\left(G_{i,k}^{(q)}-G_{j,k}^{(q)}\right)^2\right)}{2\times E_k(\mu_k\times(1-\mu_k))}=2\times\delta_{i,j}^0+\delta_{i,j}^1$$

We also know $\delta_{i,j}^0+\delta_{i,j}^1+\delta_{i,j}^2=1$ and, by definition, $\phi_{i,j}=\left(0.5\times\delta_{i,j}^2+0.25\times\delta_{i,j}^1\right)$. Combining these two relations, we get:

$$2-\frac{E_k\left(\left(G_{i,k}^{(q)}-G_{j,k}^{(q)}\right)^2\right)}{2\times E_k\left(\mu_k\times(1-\mu_k)\right)}=2\times\delta_{i,j}^2+\delta_{i,j}^1=4\times\phi_{i,j}$$

Rearranging terms, we get the final distance estimator:

$$\phi_{i,j}^{(Dist)}=\frac{1}{2}-\frac{1}{4}\times\frac{\sum_k\left(G_{i,k}^{(q)}-G_{j,k}^{(q)}\right)^2}{2\times\sum_k\left(\mu_k\times(1-\mu_k)\right)}$$

In the final estimator, the expectations are estimated using means over the variants. Equation (13) of main text modifies this estimator by integrating individual specific allele frequencies in the distance estimation (numerator) and the denominator.

Correlation-based Kinship Estimator. The correlation-based estimator relies on an expected covariance statistic:

$$E_k\left(\frac{\left(0.5 \cdot G_{i,k}^{(q)} - \mu_k\right)\left(0.5 \cdot G_{j,k}^{(q)} - \mu_k\right)}{\mu_k \times (1-\mu_k)}\right) = E_k(CorrStat_{i,j,k})$$

where the expectation is calculated over the variants. Similar to above case, we formulate the expectation conditional on the IBD state at each variant:

$$E_k(CorrStat_{i,j,k}) = P(IBD=0) \times E_k(CorrStat_{i,j,k} \,|IBD=0) +$$

$$P(IBD=1) \times E_k(CorrStat_{i,j,k} \,|IBD=1) +$$

$$P(IBD=2) \times E_k(CorrStat_{i,j,k} \,|IBD=2)$$

We use the following observations to simplify the above formulation:

$$E_k(CorrStat_{i,j,k} \mid IBD=0) = 0$$

$$E_k(CorrStat_{i,j,k} \mid IBD=1) = 0.25$$

$$E_k(CorrStat_{i,j,k} \mid IBD=2) = 0.50$$

Plugging these into expectation and using the relationship $\phi_{i,j} = \left(0.5 \times \delta_{i,j}^2 + 0.25 \times \delta_{i,j}^1\right)$, we get:

$$E_k\left(\frac{\left(0.5 \cdot G_{i,k}^{(q)} - \mu_k\right)\left(0.5 \cdot G_{j,k}^{(q)} - \mu_k\right)}{\mu_k \times (1-\mu_k)}\right) = \delta_{i,j}^1 \times 0.25 + \delta_{i,j}^2 \times 0.50 = \phi_{i,j}$$

Thus, the correlation-based kinship estimator can be derived using mean of the covariance statistic over all variants:

$$\phi_{i,j}^{(corr)} = \sum_k \frac{\left(\left(0.5 \cdot G_{i,k}^{(q)} - \mu\right) \times \left(0.5 \cdot G_{j,k}^{(q)} - \mu\right)\right)}{\mu_k \times (1-\mu_k)}$$

Equation (12) of the main text replaces the allele frequencies with individual-specific allele frequencies.

Comparison of Distance and Correlation-based Estimators. An important distinction between the above derivation is the starting point of the estimators. Distance-based estimator utilizes the convergence of the mean squared-distance between the variant genotypes to the expected kinship value and therefore loses the covariance information. This information becomes important to estimate the deviations around mean. For example, the enrichment of excess homozygous genotypes is indicative of ancestral inbreeding events and can be used to estimate the inbreeding coefficient. Correlation-based estimators are useful for detecting these because they rely on convergence of the covariance between the genotype signals.

Both estimators rely on Hardy-Weinberg Equilibrium (HWE) to hold for the analyzed variants and will give biased estimates when variants do not satisfy HWE.

Extended Background on Genomic Privacy and Kinship Estimation

Decreasing cost of sequencing has contributed to a massive increase in the number of available genetic data [1,2]. Coupled with recreational usage of genetic data, genetic data has become increasingly prevalent in clinic and daily life [3,4]. There are, however, challenges around the usage of genetic data. For example, as the genealogy databases are growing in size, genetic surveillance has also taken off and is actively used by law enforcement to solve cold cases [5,6]. This is mainly accomplished by searching for genetic data recovered from crime scenes and using publicly available genealogy databases to identify relatives. Genetic data is sensitive in nature and can be used to re-identify individuals very easily. Even a a handful genetic variants from an individual can be used to reidentify them within a large cohort [7–10]. This method is very effective for identifying individuals and can be used to identify a large portion of the population [11]. In addition, numerous computational "attacks" that can enable reidentification of individuals have been described in previous studies. These include linking attacks [12–14], genotype reconstruction attacks [15–17], attacks on genealogy databases [18], membership and phenotype inference attacks[19–23], and model inversion attacks [24]. Much of these attacks implicate and create discrimination and stigmatization risks to individuals themselves and their families [25–28]. To counter these concerns, several laws have been enacted. European Union's General Data Protection Regulation (GDPR) is currently the most extensive regulation on the 3rd parties regarding personal data sharing and storage.

One of the main usages of genetic data is identifying relatedness and relatives. In principle, parent-child and siblings share approximately half of their genetic information. The sharing patterns decrease as the relatedness degree decreases, e.g. grandparents, cousins, etc. This sharing stems from inheritance of DNA through random assortment and homologous chromosome recombination in meiosis. Based on the expected value of the realized kinship value, we can find relatives using marker genotypes [29]. This biological phenomenon has far outreaching impact and is used extensively in population genetics [30] and forensics [31,32]. Genetic relatedness or kinship [33,34] is an important quantity that is central to many fields such as behavioral science [35], human evolution [36], animal and plant breeding [37,38]. Of note, kinship estimation is essential in linkage mapping studies [39] and kinship matrices are also used to model the polygenic effects in association studies [40–42] for the correction of cryptic relatedness that is known to create confounding effects and bias effect sizes [41,43]. While pedigree information provides an exact value of expected kinship, systematic estimation and correction of relatedness among samples can be more beneficial in large studies [44]. Also, even in pedigrees, genetic kinship estimation can provide more exact estimates of variation [45] among individuals and should be more preferred than using reported pedigree information that can contain manual curation errors. With the growing awareness to increase diversity in the field of population genetics, there is a need to correct for the biases that are caused by admixed populations. Moreover, non-random mating, i.e., assortative mating, among similar ancestral groups [46,47] may bias estimates of relationship. General methods that assume random mating or simple homogeneous populations are not effective in appropriately estimating kinship and may impact downstream analysis and interpretations.

The kinship between two individuals, denoted by ϕ, is the probability that two alleles at a random position in the genomes of the individuals are identical-by-descent (IBD), i.e., they are inherited from the same

ancestor. Kinship coefficient is closely related to other metrics such as the inbreeding coefficient [48] and IBD-sharing probabilities [49], which are essential for estimating population-level genetic information. Multiple methods have been proposed for estimating kinship and related statistics using marker genotypes. Methods that make use of marker genotypes infer the realized kinship by estimating IBD using pairwise comparison of Identical-by-state (IBS) marker genotypes and developing statistics based on the expected matches and mismatches [50].

Allele frequency-based approaches include maximum-likelihood (ML) and method-of-moments (MoM) approaches. ML approaches rely on joint probabilistic modeling of observed genotypes and maximizing the likelihood (ERSA [51] and RelateAdmix [52]) and have higher computational requirements than MoM estimators, which relate IBS frequencies to IBD estimates by matching mean IBS markers under Hardy-Weinberg equilibrium. Among these, ML methods [53] seem to perform slightly better in comparison to MoM estimators for identifying relatives albeit higher computational requirements. Of note, genomic relationship matrix (GRM), which is also used in plink [54] package, has been used to quantify relatedness, although it may be sensitive to variant selection [55]. Similarly, GRAF utilizes efficient metrics and is used to detect close relatives and duplicate individuals in the protected datasets in dbGAP [56].

KING [57] utilizes very efficient estimators for kinship and IBD sharing probabilities derived from IBS statistics and a novel formulation of genotype distances in terms of kinship estimates. However, KING underestimates kinship coefficients when there is population admixture in the compared individuals. REAP [58] and PC-Relate [59] make use of estimation of population admixture [60] and individual-specific allele frequencies, which are used to correct biases in kinship estimates in admixed populations. In general, REAP, PC-Relate, and GCTA's estimation methods of kinship are inherently very similar based on a conditional genotype correlation metric while KING provides a different approach based on a formulation of kinship in terms of squared genotype distance. KING's efficiency derives from the fact that genotype distances can be formulated as bitwise operations that can be very efficiently computed. On the other hand, KING suffers from underestimation of kinship for distantly related and unrelated individuals, as also discussed in Manichaikul et al. [57].

New methods are developed for estimating kinship statistics directly from next-generation-sequencing datasets such as NGSRemix [61], SEEKIN [62], lcmKin [63,64] , and LASER [65]. These methods make use of read level information to extract kinship information by taking the variant accuracy into account, which is important when for low-coverage samples. There are also methods that derive kinship and relatedness using efficient IBD-segment matching between individuals and quantify kinship directly from match genomewide statistics such as FastIBD [66], RAFFI [67], IBDKin [68]. One of the drawbacks of these methods is that they require phased genotype data, which may incur high computational costs while estimating IBD statistics on large populations. However, most of the large scale genotype data are distributed after phasing and can be processed with these tools.

Numerous methods have been proposed for privacy-aware analysis of ancestry and admixture. Kinship statistics are sensitive information as they can be used to detect relatives in 3rd party databases without consent of the owners. Similarly, population-level inbreeding estimates can cause marginalization and stigmatization risks [69–71]. It is therefore important to consider privacy risks while estimating and reporting kinship for underrepresented and historically isolated populations. PREMIX [72] computes admixture rates in a privacy-preserving manner using now deprecated SGX-based extensions, using an EM step to optimize admixture rates. He et al. proposed using an efficient genome sketching technique and

combined it with cryptographic evaluation to search for relatives using marker genotype datasets [73]. Similar sketching techniques have been proposed for finger print and relative search analysis [74]. Dervishi et al. proposed privacy-aware kinship estimation by integrating local differential privacy and genotypic data hiding [75], which may hinder the utility of genetic data and may provide ad-hoc privacy guarantees. While these methods are promising, the impact of admixture is not generally taken into account, in addition, the effectiveness of the methods is tested only for one kinship statistic that provides partial information about relatedness.

Here, we present SIGFRIED, a projection-based approach to utilize existing reference genotype datasets for estimating admixture rates for each individual and use these estimates for kinship and related statistics [65] in admixed populations. SIGFRIED can also perform secure kinship estimation to provide confidentiality to genotype data. Projection-based on Principal Component Analysis (PCA) is used extensively to estimate population structure. For example, PC-Relate [59] first performs PCA on the genotype matrix of unrelated individuals and estimates individual-specific allele frequencies of variants using a base average frequency and a residual estimate of the ancestry-specific allele frequency component. REAP [58] depends on estimation of individual-specific allele frequencies using external tools such as frappe [76] or ADMIXTURE [77].

SIGFRIED takes a 2-step approach to decrease computational requirements while making use of publicly available reference panels: (1) We estimate admixture rates using a non-linear function of projections on the reference panel. Usage of PCA and reference populations with a "distance-based" estimation of admixture has shown promise in previous studies [78,79]. We capitalize on these and propose a similar approach as input to kinship statistic estimation. (2) The predicted admixture rates are used to estimate individual-specific allele frequencies and are integrated into computation of kinship, inbreeding, and IBD sharing probabilities. In comparison to previous methods, SIGFRIED imposes less computational burden without the requirement of a full PCA or more complex EM-based admixture estimates using the genotype matrix for estimation of the kinship and IBD-sharing probabilities, which are prohibitively challenging in secure implementations. Rather, we show that when the existing reference panel is concordant with the ancestry of individuals, projection-based admixture estimates can be used for accurate kinship estimates. Thus, SIGFRIED uses admixture rates only as intermediary information. After establishing the accuracy of the projection-based approach, we focus on the privacy-aware implementation. One of the main advantages of SIGFRIED's approach is that it renders itself well for efficient and flexible privacy-aware computations because of its modular approach. We formulated and implemented a secure federated kinship estimation among 2-sites wherein genetic data is kept confidential while kinship statistics are estimated. Our implementation relies on homomorphic encryption [80], which enables processing encrypted genotype data directly without ever being decrypted and therefore provides provable security guarantees on the genetic data. Overall, these results highlight the utility of existing population panels for secure estimation of kinship statistics. From privacy-preserving analysis perspective, this can enable circumventing a full PCA computation – which is prohibitively challenging for large genotype datasets – and enable secure and accurate analysis of relatedness and population-level inbreeding under different scenarios.

Secure Computation of Zero-IBD-Sharing Probabilities ($\delta_{i,j}^0$). We again assume that 2 sites would like to compute zero-IBD sharing probabilities without sharing genotype data in plaintext form. We first decompose computation of the zero-IBD sharing probabilities into 4 components:

$$\delta^0_{i,j} = \frac{\langle I^{(1,AA)}_{i,\cdot}, I^{(2,aa)}_{j,\cdot}\rangle + \langle I^{(1,aa)}_{i,\cdot}, I^{(2,AA)}_{j,\cdot}\rangle}{\langle \mu^2_{i,\cdot}, (1-\mu^2_{j,\cdot})\rangle + \langle \mu^2_{j,\cdot}, (1-\mu^2_{i,\cdot})\rangle} \quad (1)$$

In (11), $I^{(1,AA)}_{i,k}$ denotes an indicator variable that takes on a value of 1 if $G^{(1)}_{i,k} = 2$ (i.e., AA) and is 0 otherwise. $I^{(1,aa)}_{i,k}, I^{(2,aa)}_{i,k}$, and $I^{(2,AA)}_{i,k}$ are similarly defined. To compute $\delta^0_{i,j}$ from (1), it is necessary to share indicator matrices and individual-specific allele frequencies between sites. The indicator variables explicitly describe the homozygous genotypes and they must be encrypted. As in estimation of $\phi_{i,j}$, we assume Site-2 homomorphically encrypts the indicator variable matrices, $\check{I}^{(2,AA)}_{j,k}$, $\check{I}^{(2,aa)}_{j,k}$ and sends them to Site-1. We assume allele frequencies from Site 2 are sent in plaintext format without encryption. Next, Site-1 securely computes the numerator in (1): $\langle I^{(1,AA)}_{i,\cdot}, \check{I}^{(2,aa)}_{j,\cdot}\rangle + \langle I^{(1,aa)}_{i,\cdot}, \check{I}^{(2,AA)}_{j,\cdot}\rangle$. The denominator is computed in plaintext format on Site-1 using the allele frequencies from the two sites. Site 1 sends the encrypted zero-IBD probabilities, $\check{\delta}^0_{i,j}$, to Site-2. Site-2 decrypts $\check{\delta}^0_{i,j}$ and shares the results with Site-1. As with distance-based kinship estimation, the numerator and denominator in (1) can be computed in parallel on the two sites to decrease the computational load on one site.

Discussion on Advantages and Limitations of SIGFRIED

Kinship and related statistics are essential in many genetic studies and they are sensitive for individual and group-level privacy. Here, we presented SIGFRIED, an efficient, accurate, and secure method that utilizes projection on existing reference panels. SIGFRIED balances accuracy and efficiency to ensure that the final algorithm can be implemented with secure primitives. While projection on existing population panels has been utilized previously by other methods, SIGFRIED utilizes projection to circumvent computations that are otherwise hard to implement in the secure domain, such as performing full secure PCA or computationally intensive expectation-maximization iterations. From this perspective, we view SIGFRIED as a private-by-design methodology wherein the privacy considerations are balanced against efficiency and accuracy and these are reflected in each step of the method. Projection does not explicitly require reference panel genotypes, and only reference population centroid coordinates, allele frequencies, and PCs are necessary for the projection. Since the reference genotypes are not explicitly shared, we believe the centroids and PCs create minimal risk for reference panels under restricted access (i.e. TOPMed [81]).

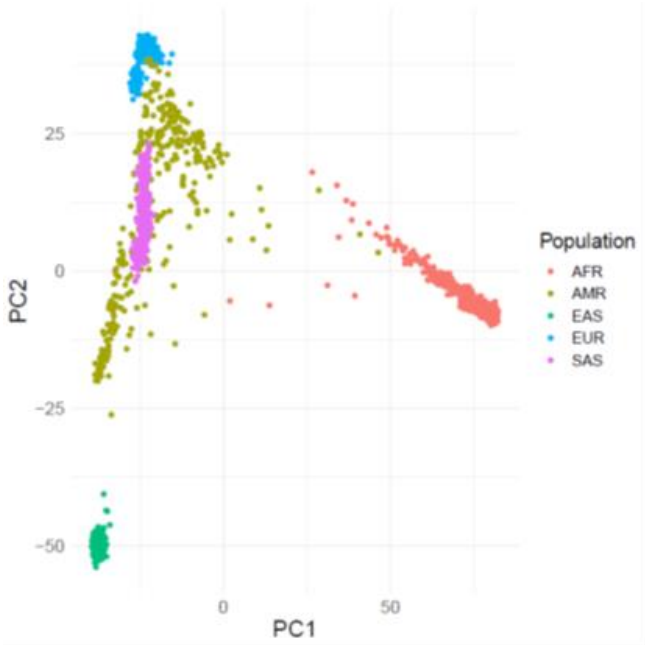
While we presented a specific privacy-preserving scenario with a proof-of-concept implementation for a 2-site federated estimation of kinship, the implementation and the scenarios can be differently setup and framed to expand to more than 2 sites and also for utilizing an outsourcing service for kinship estimation. The outsourcing can be performed by an untrusted entity because sensitive data is encrypted and cannot be used to infer any information by an unauthorized party. When deployed on a highly scalable but untrusted environment such as AWS or Google Cloud, the performance can be tuned as desired. Also, SIGFRIED implements kinship estimation in a flexible manner using modular steps and decomposable functions. This is beneficial for optimizing privacy-vs-performance in different scenarios. The flexibility is important because new protocols can choose to encrypt only certain parts of the intermediate statistics to ensure that performance is optimized and security requirements are met according to local regulations and patient or participant consent. For instance, the individual-specific allele frequencies are averages of population-specific allele frequencies weighted by admixture rates. As such, they are highly aggregated function of genotypes and can be deemed safe to share in plaintext form.

SIGFRIED has several limitations that warrant future research. First, we presented two distance-to-admixture mapping functions that may not be optimal for certain scenarios and may need re-parametrizations. Also, new functions for mapping distances to admixture rates can provide less biased results. For example, it is likely that other functions such as *logit* can be explored for further optimizations in accuracy and secure implementation. Second, SIGRFIED relies on a representative set of reference populations and a-priori knowledge of the query dataset, which may be limiting factor in certain cases, especially when the query samples are of unknown origin or are members of underrepresented populations. We foresee that the increase in the number and diversity of available reference panels (i.e. TOPMed Project) will make the reference panels more complete and inclusive. It is, however, still necessary to generate the reference panel centroids in the most optimal way, which is a direction that should be studied further. One example of this is PCAir[60] method, which estimates the principal components by selecting a subset of the query individuals who are most likely unrelated and uses these to compute the principal components. Although this approach cannot be efficiently implemented in secure primitives as it is, it can be used to build a more accurate centroid estimation method in SIGFRIED. However, PCAir requires a metric to define the unrelated that is generated by an external tool [82] (such as KING-Robust), which may create a circularity problem as kinship estimation depends on identifying unrelated individuals. It is worth noting, however, that identification of unrelated query individuals is generally a much easier task than the exact estimation of kinship. Third, the performance of secure federated kinship estimation may be prohibitive for very large sample sizes. To get around this limitation, the secure computation of kinship and IBD-sharing probabilities can be more efficiently performed with the use of simpler encryption techniques, which can provide better performance. The performance can further be improved using smaller number of variants depending on kinship distance that is required from the estimation – for example, 1st and 2nd degree relatives can be identified with smaller number of variants, which can improve the secure estimation performance.

SIGFRIED has several advantages over other approaches. For example, numerous other methods rely on large sample sizes that are representative of the underlying populations (e.g., PC-Air or ADMIXTURE) or existence of phased genotype data (IBDKin and RAFFI). On the other hand, SIGFRIED relies on an existing reference population panel and can work effectively even in small sample sizes. Also, the estimates of kinship statistics do not rely on the query genotype data because SIGFRIED's individual-specific allele frequencies do not change when query data is subsetted or extended by the addition of new samples. This is not necessarily true for other methods that estimate admixture and allele frequency parameters from query genotype datasets.

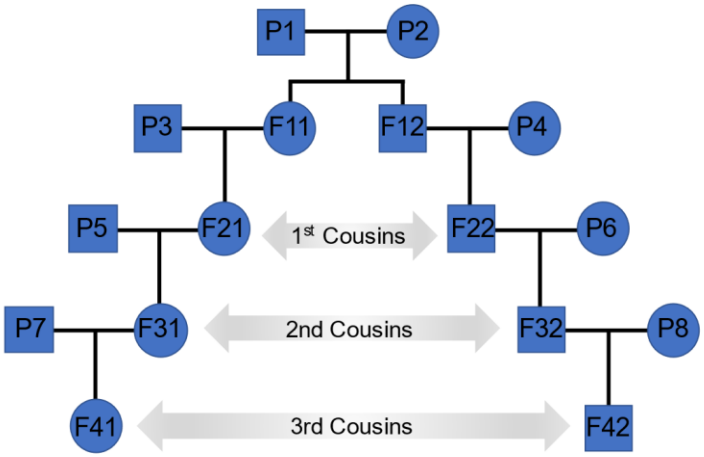
Supplementary Figures

Fig S1.

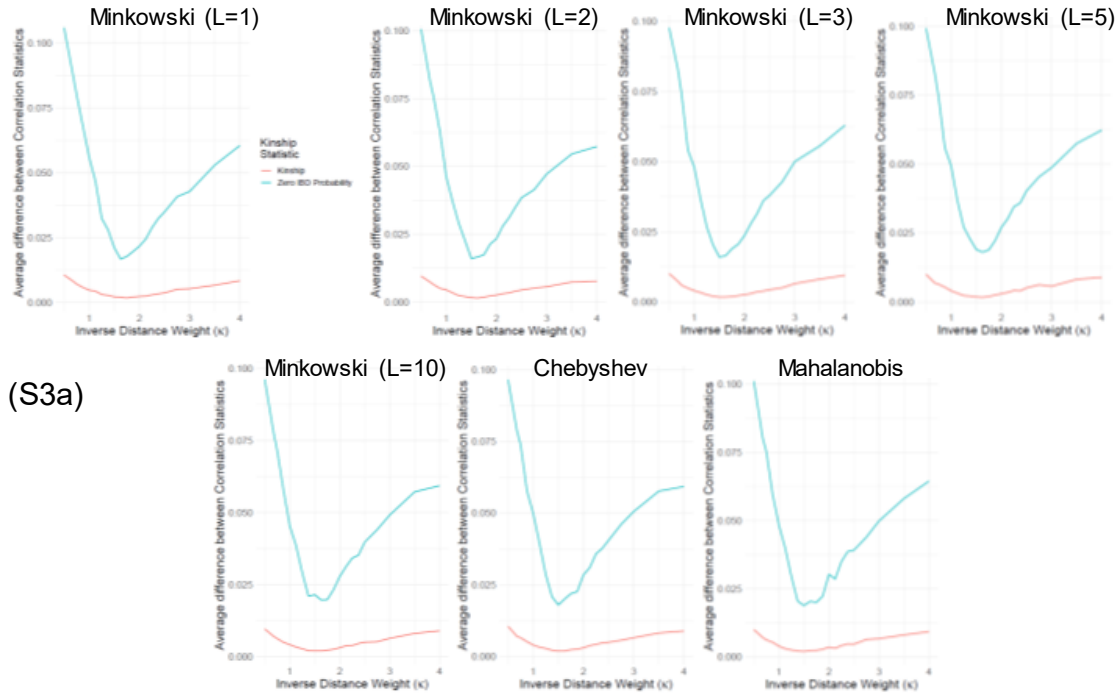


Supplementary Figure 1. Projection of the 2,504 individuals in the 1000 Genomes Project on top 2 components of the genotype matrix. Each dot represents an individual and colors indicate the population of each individual. It should be noted 2 components are used for illustration purposes. The number of components that SIGFRIED uses for admixture estimation step can be changed by the user.

Fig S2. Pedigree structure used for simulations



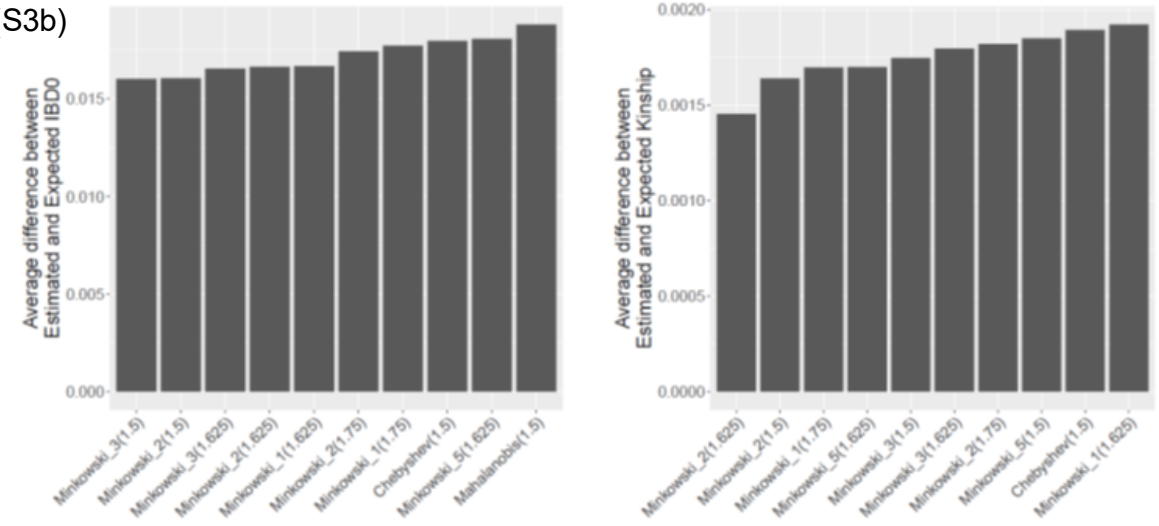
Supplementary Figure 2. The pedigree structure that is used for simulations. Individuals named with P1-8 are the founders. The 1st, 2nd, and 3rd-degree cousins are indicated by grey arrows.



(S3a)

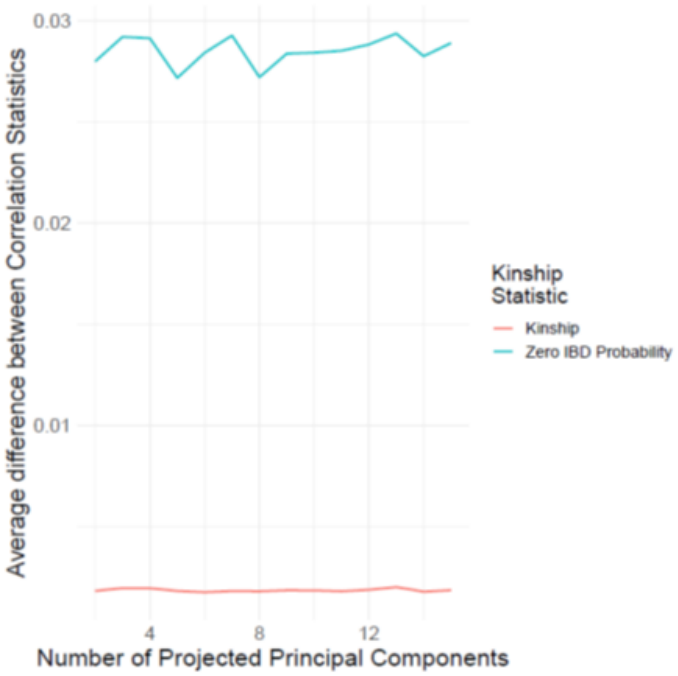
Supplementary Figure 3a. The average absolute difference between REAP-ADMIXTURE and correlation-based kinship statistics with changing κ in inverse-distance to admixture mapping function. Red line shows the difference in kinship coefficient (ϕ_{ij}) and cyan line shows the difference in zero-IBD probability, δ_{ij}^{0}. Each plot shows the difference in statistics for a distinct distance metric that is utilized in admixture estimation. The distance metric is indicated at the top of the plot, namely Minkowski, Chebyshev, and Mahalanobis distances. For Minkowski distance metric, the power term (L) that is used for the corresponding plot is indicated in the parenthesis.

(S3b)



Supplementary Figure 3b. The average absolute difference between REAP-ADMIXTURE and correlation-based kinship statistics with changing κ in inverse-distance to admixture mapping function for different distance metrics. Each barplot corresponds to a distance metric (indicated on the x-axis) and distance weight parameter shown in the parenthesis on the x-axis. Left plot shows the difference in IBD0 probability estimated by REAP-ADMIXTURE and SIGFRIED and right plot shows the kinship difference between the methods.

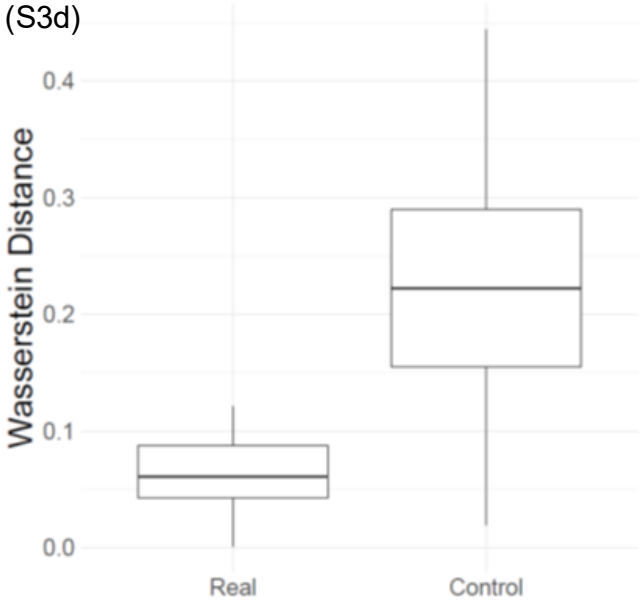
(S3c)



Supplementary Figure 3c. The average absolute difference REAP-ADMIXTURE and correlation-based kinship statistics with changing number of components (K) used in distance estimation.

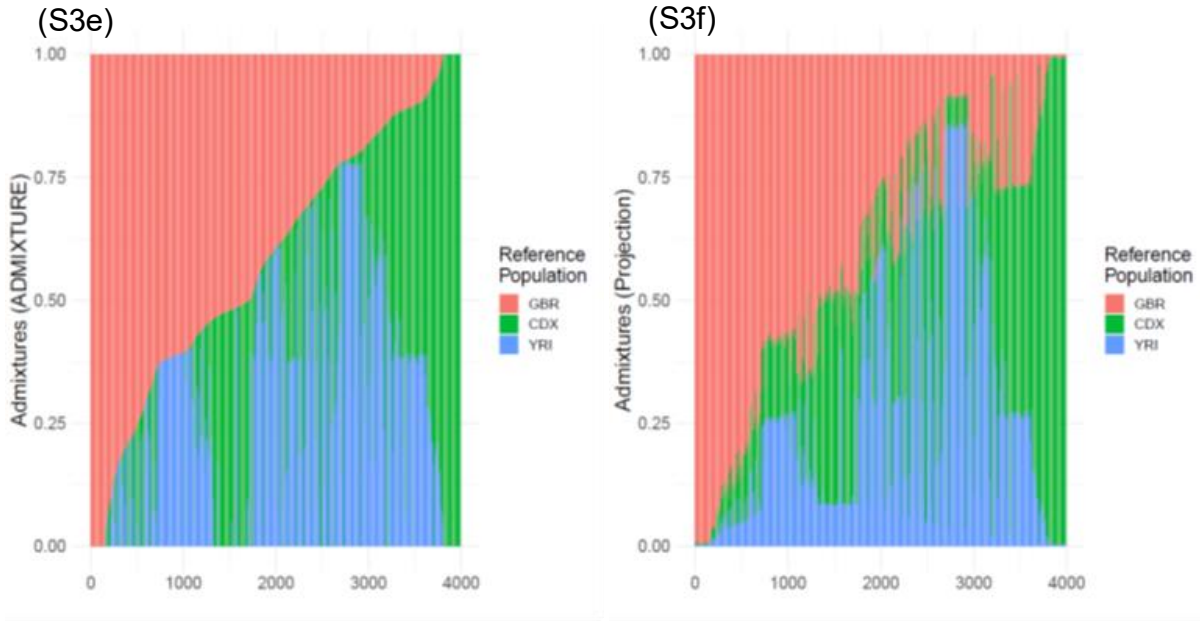
Fig. S3. Admixture Concordance

(S3d)



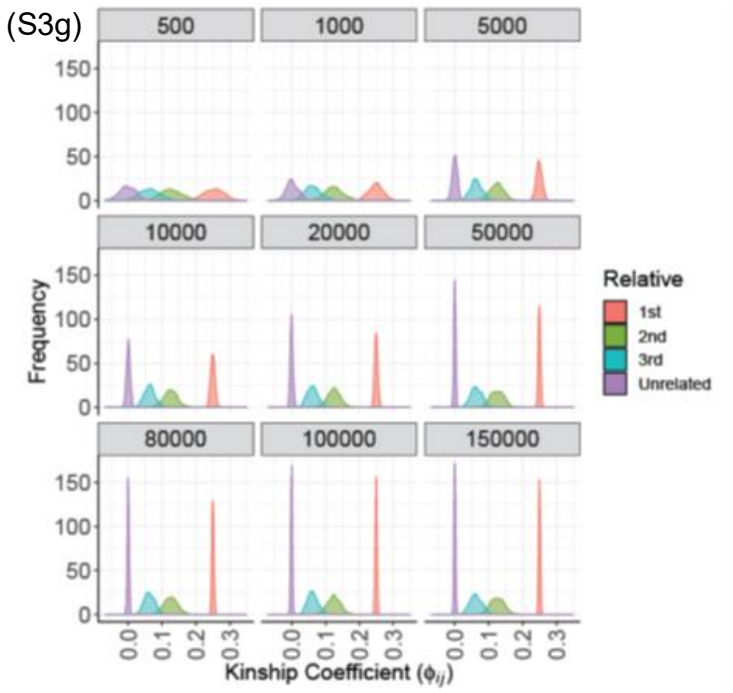
Supplementary Figure 3d. The distribution of Wasserstein distance between ADMIXTURE-predicted rates and projection-based rates (marked with “Real”) and distribution of Wasserstein distance between ADMIXTURE-predicted rates and uniform assigned rates (marked with “Control”).

Fig. S3. Admixture Concordance



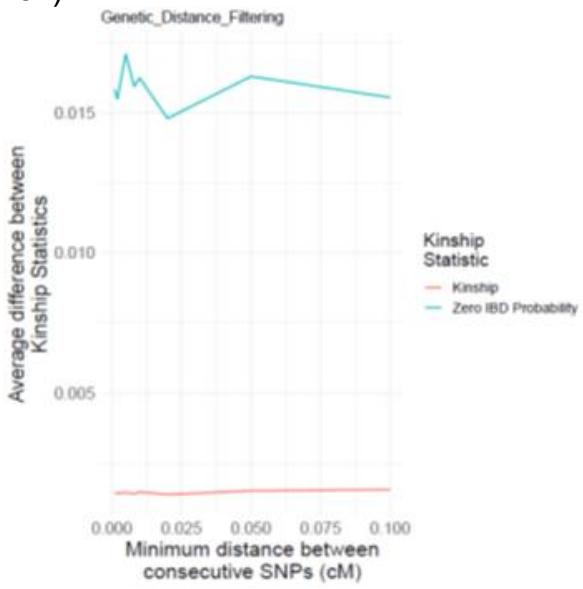
Supplementary Figure 3e, 3f. (S3e) The distribution of assigned admixture rates to the 4000 non-founder individuals in the simulated pedigrees by ADMIXTURE using data in Fig. S3d. Each column corresponds to an individual and the length of colored bars indicate y-axis the admixture fractions of each individual in the corresponding column. The colors in each column indicate different ancestries: Red is GBR, Green is CDX, and Blue is YRI. **(S3f)** The distribution of assigned admixture rates by projection-based admixture estimation**.**

Fig. S3. Number of Variants

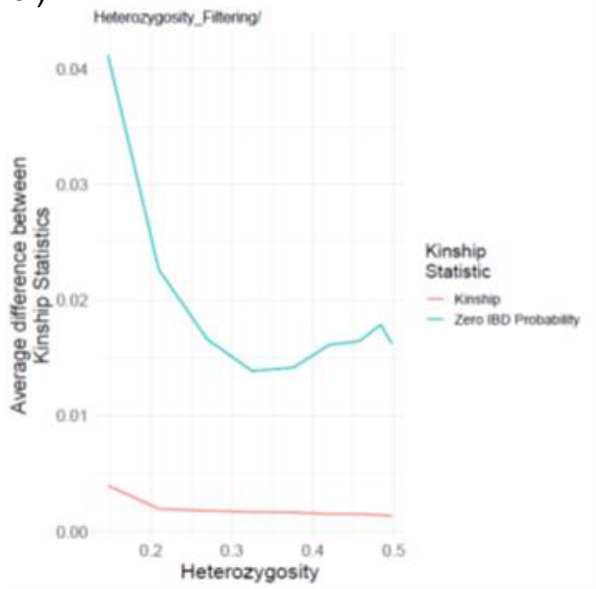


Supplementary Figure 3g. The kinship coefficient (x-axis) distribution with different number of variants that are uniformly subsampled from the 1000 Genomes variants. Each plot shows a kinship distribution generated using number of variants indicated at the label.

(S3h)

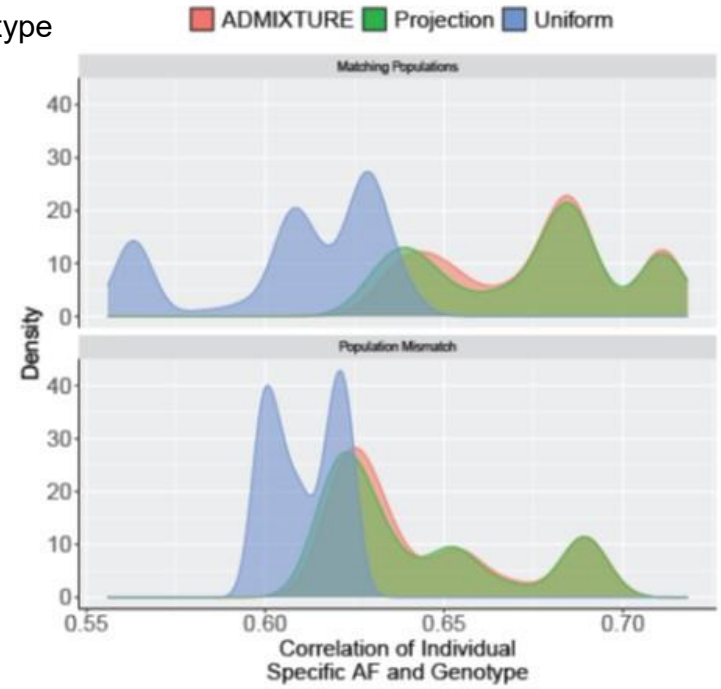


(S3i)



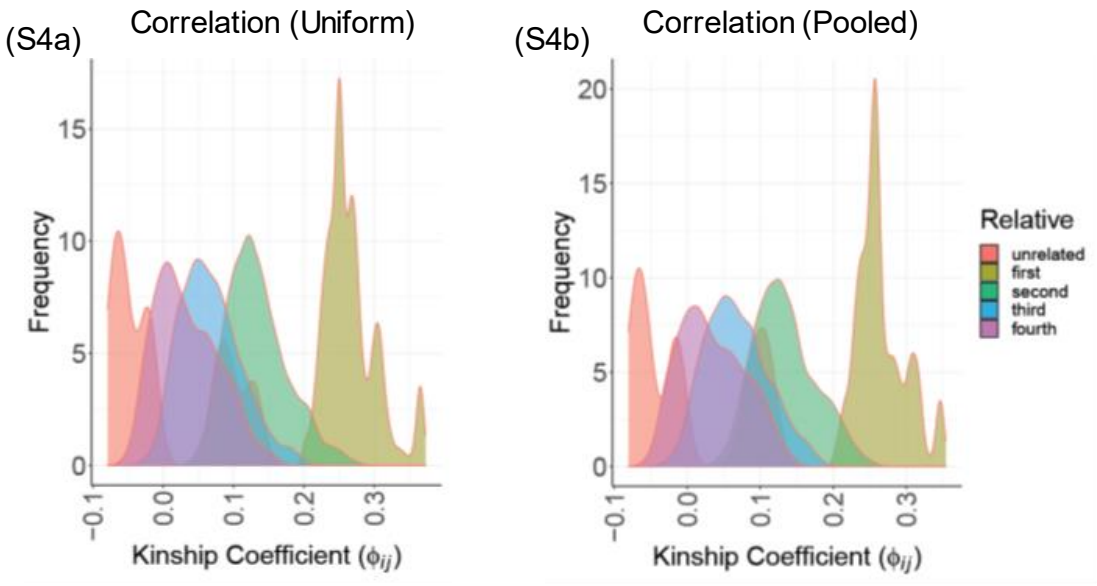
Supplementary Figure 3h, 3i. Impact of genetic distance and heterozygosity of the selected variants on kinship statistic estimations. **(3h)** The average absolute difference between REAP-ADMIXTURE and correlation-based kinship statistics with changing genetic distance between consecutive variants. X-axis shows the genetic distance between consecutive variants (in centiMorgans). Cyan curve shows difference in probability of IBD0 estimates and red curve shows the difference in kinship estimates. **(3i)** The kinship statistic difference with changing heterozygosity of the variants (x-axis) used in kinship estimation.

(S3j) Personal AF vs Genotype Correlations



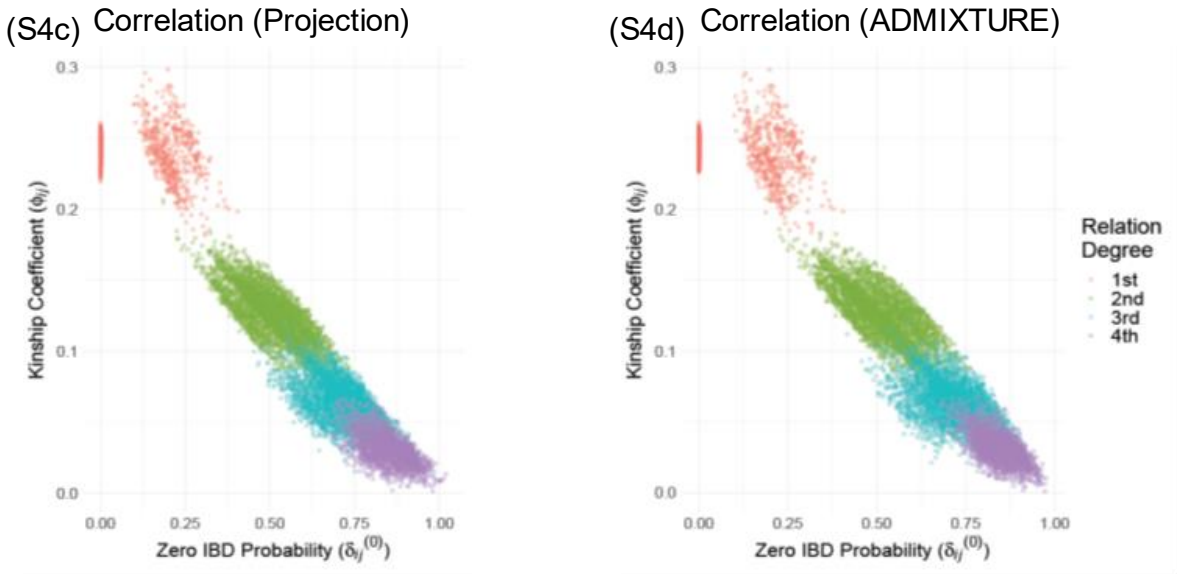
(Supplementary Figure S3j) The distribution of Pearson correlation between the individual-specific allele frequency and genotype for 50 pedigrees using matching (top) and non-matching (bottom) pedigree and reference populations. Colors indicate the method used to estimate admixture rates used in estimation of individual specific allele frequencies.

Fig. S4: Heterozygous Ancestry



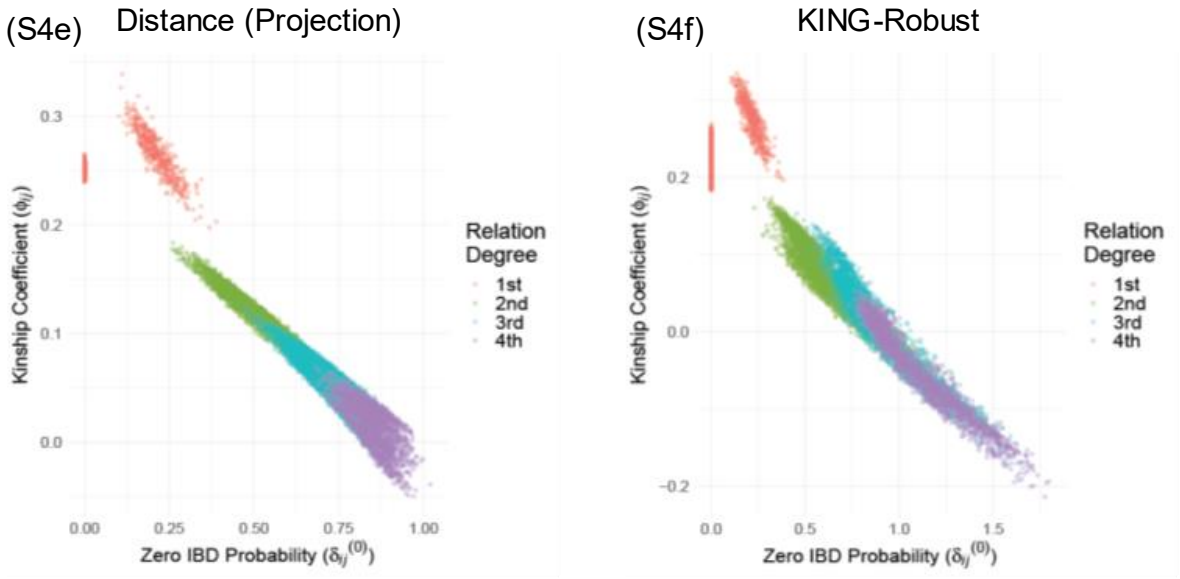
(Supplementary Figure 4a, 4b). **(S4a)** Distribution of correlation-based kinship estimates using uniform admixture assignments for every sample. **(S4b)** Distribution of correlation-based kinship estimates using all populations in assignment of individual specific allele frequencies.

Fig. S4: Het. Kinship and Pr(IBD=0)



(Supplementary Figure 4c, 4d). **(S4c)** Correlation-based Kinship estimates (Projection) vs Zero-IBD probability. **(S4d)** Correlation-based Kinship estimates (ADMIXTURE) vs Zero-IBD probability.

Fig. S4: Het. Kinship and Pr(IBD=0)



(Supplementary Figure 4e, 4f) (S4e) Distance-based Kinship estimates (Projection) vs Zero-IBD probabilities. **(S4f)** KING-Robust Kinship estimates vs Zero-IBD probabilities.

References

1. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 2016;17: 53.

2. Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biology. 2011. p. 125. doi:10.1186/gb-2011-12-8-125

3. Evans JP. Recreational genomics; what's in it for you? Genet Med. 2008;10: 709–710.

4. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013. doi:10.1038/gim.2013.73

5. Wickenheiser R. Forensic genealogical searching and the golden state serial killer. Forensic Science International: Synergy. 2019;1: S9–S10.

6. Wickenheiser RA. Forensic genealogy, bioethics and the Golden State Killer case. Forensic Sci Int Synerg. 2019;1: 114–125.

7. Visscher PM, Hill WG. The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS Genet. 2009;5: e1000628.

8. Wei YL, Li CX, Jia J, Hu L, Liu Y. Forensic Identification Using a Multiplex Assay of 47 SNPs. J Forensic Sci. 2012;57: 1448–1456.

9. Pakstis AJ, Speed WC, Fang R, Hyland FCL, Furtado MR, Kidd JR, et al. SNPs for a universal individual identification panel. Hum Genet. 2010;127: 315–324.

10. Yousefi S, Abbassi-Daloii T, Kraaijenbrink T, Vermaat M, Mei H, van 't Hof P, et al. A SNP panel for identification of DNA and RNA specimens. BMC Genomics. 2018;19. doi:10.1186/s12864-018-4482-7

11. Kaiser J. We will find you: DNA search used to nab Golden State Killer can home in on about 60% of white Americans. Science. 2018. doi:10.1126/science.aav7021

12. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat Methods. 2016;13: 251–256.

13. Harmanci A, Gerstein M. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nat Commun. 2018;9. doi:10.1038/s41467-018-04875-5

14. Gürsoy G, Emani P, Brannon CM, Jolanki OA, Harmanci A, Strattan JS, et al. Data Sanitization to Reduce Private Information Leakage from Functional Genomics. Cell. 2020;183: 905-917.e16.

15. Gürsoy G, Lu N, Wagner S, Gerstein M. Recovering genotypes and phenotypes using allele-specific genes. Genome Biol. 2021;22: 263.

16. Paige B, Bell J, Bellet A, Gascón A, Ezer D. Reconstructing genotypes in private genomic databases from genetic risk scores. J Comput Biol. 2021;28: 435–451.

17. Ayoz K, Ayday E, Cicek AE. Genome reconstruction attacks against genomic data-sharing beacons. Proc Priv Enhancing Technol. 2021;2021: 28–48.

18. Edge MD, Coop G. Attacks on genetic privacy via uploads to genealogical databases. Elife. 2020;9. doi:10.7554/eLife.51810

19. Chen J, Wang WH, Shi X. Differential privacy protection against membership inference attack on machine learning for genomic data. Pac Symp Biocomput. 2021;26: 26–37.

20. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. 2017 IEEE Symposium on Security and Privacy (SP). IEEE; 2017. doi:10.1109/sp.2017.41

21. Almadhoun N, Ayday E, Ulusoy Ö. Inference attacks against differentially private query results from genomic datasets including dependent tuples. Bioinformatics. 2020;36: i136–i145.

22. Humphries T, Oya S, Tulloch L, Rafuse M, Goldberg I, Hengartner U, et al. Investigating membership inference attacks under data dependencies. arXiv [cs.CR]. 2020. Available: http://arxiv.org/abs/2010.12112

23. Hagestedt I, Humbert M, Berrang P, Lehmann I, Eils R, Backes M, et al. Membership inference against DNA methylation databases. 2020 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE; 2020. doi:10.1109/eurosp48549.2020.00039

24. Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15. New York, New York, USA: ACM Press; 2015. doi:10.1145/2810103.2813677

25. Ayday E, Humbert M. Inference attacks against kin genomic privacy. IEEE Secur Priv. 2017;15: 29–37.

26. Humbert M, Ayday E, Hubaux J-P, Telenti A. Addressing the concerns of the lacks family: quantification of kin genomic privacy. Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13. 2013. doi:10.1145/2508859.2516707

27. Telenti A, Ayday E, Hubaux JP. On genomics, kin, and privacy. F1000Res. 2014. doi:10.12688/f1000research.3817.1

28. Samani SS, Huang Z, Ayday E, Elliot M, Fellay J, Hubaux JP, et al. Quantifying genomic privacy via inference attack with high-order SNV correlations. Proceedings - 2015 IEEE Security and Privacy Workshops, SPW 2015. 2015. pp. 32–40.

29. Helfer BS, Fremont-Smith P, Ricke DO. The genetic chain rule for probabilistic kinship estimation. bioRxiv. bioRxiv; 2017. p. 202879. doi:10.1101/202879

30. Bérénos C, Ellis PA, Pilkington JG, Pemberton JM. Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. Mol Ecol. 2014;23: 3434–3451.

31. Jobling MA, Gill P. Encoded evidence: DNA in forensic analysis. Nat Rev Genet. 2004;5: 739–751.

32. Kayser M, de Knijff P. Erratum: Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet. 2012;13: 753–753.

33. Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 2015;16: 33–44.

34. Goudet J, Kay T, Weir BS. How to estimate kinship. Mol Ecol. 2018;27: 4121–4135.

35. Fisher RM, Cornwallis CK, West SA. Group formation, relatedness, and the evolution of multicellularity. Curr Biol. 2013;23: 1120–1125.

36. Uyenoyama MK. Inbreeding and the evolution of altruism under kin selection: Effects on relatedness and group structure. Evolution. 1984;38: 778.

37. Madsen T, Stille B, Shine R. Inbreeding depression in an isolated population of adders Vipera berus. Biol Conserv. 1996;75: 113–118.

38. Wellmann R, Bennewitz J. Key genetic parameters for population management. Front Genet. 2019;10: 667.

39. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet. 1998;63: 259–266.

40. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42: 348–354.

41. Choi Y, Wijsman EM, Weir BS. Case-control association testing in the presence of unknown relationships. Genet Epidemiol. 2009;33: 668–678.

42. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004;36: 512–517.

43. Kirkpatrick B, Bouchard-Côté A. Correcting for cryptic relatedness in genome-wide association studies. arXiv [q-bio.QM]. 2016. Available: http://arxiv.org/abs/1602.07956

44. Wang J. Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? Theor Popul Biol. 2016;107: 4–13.

45. Gao S, Donohue B, Hatch KS, Chen S, Ma T, Ma Y, et al. Comparing empirical kinship derived heritability for imaging genetics traits in the UK biobank and human connectome project. Neuroimage. 2021;245: 118700.

46. Sebro R, Hoffman TJ, Lange C, Rogus JJ, Risch NJ. Testing for non-random mating: evidence for ancestry-related assortative mating in the Framingham heart study. Genet Epidemiol. 2010;34: 674–679.

47. Risch N, Choudhry S, Via M, Basu A, Sebro R, Eng C, et al. Ancestry-related assortative mating in Latino populations. Genome Biol. 2009;10: R132.

48. Rousset F. Inbreeding and relatedness coefficients: what do they measure? Heredity (Edinb). 2002;88: 371–380.

49. Meuwissen TH, Goddard ME. Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol. 2001;33: 605–634.

50. Wang B, Sverdlov S, Thompson E. Efficient estimation of realized kinship from single nucleotide polymorphism genotypes. Genetics. 2017;205: 1063–1078.

51. Huff CD, Witherspoon DJ, Simonson TS, Xing J, Watkins WS, Zhang Y, et al. Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Res. 2011;21: 768–774.

52. Moltke I, Albrechtsen A. RelateAdmix: a software tool for estimating relatedness between admixed individuals. Bioinformatics. 2014;30: 1027–1028.

53. Ramstetter MD, Dyer TD, Lehman DM, Curran JE, Duggirala R, Blangero J, et al. Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. Genetics. 2017;207: 75–82.

54. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575.

55. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res (Camb). 2009;91: 47–60.

56. Jin Y, Schäffer AA, Sherry ST, Feolo M. Quickly identifying identical and closely related subjects in large databases using genotype data. PLoS One. 2017;12: e0179106.

57. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26: 2867–2873.

58. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet. 2012;91: 122–138.

59. Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free estimation of recent genetic relatedness. Am J Hum Genet. 2016;98: 127–148.

60. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet Epidemiol. 2015;39: 276–293.

61. Nøhr AK, Hanghøj K, Garcia-Erill G, Li Z, Moltke I, Albrechtsen A. NGSremix: a software tool for estimating pairwise relatedness between admixed individuals from next-generation sequencing data. G3 (Bethesda). 2021;11. doi:10.1093/g3journal/jkab174

62. Dou J, Sun B, Sim X, Hughes JD, Reilly DF, Tai ES, et al. Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. PLoS Genet. 2017;13: e1007021.

63. Lipatov M, Sanjeev K, Patro R, Veeramah K. Maximum likelihood estimation of biological relatedness from low coverage sequencing data. bioRxiv. 2015. p. 023374. doi:10.1101/023374

64. Li H, Glusman G, Hu H, Shankaracharya, Caballero J, Hubley R, et al. Relationship estimation from whole-genome sequence data. PLoS Genet. 2014;10: e1004144.

65. Wang C, Zhan X, Liang L, Abecasis GR, Lin X. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. Am J Hum Genet. 2015;96: 926–937.

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

67. Naseri A, Shi J, Lin X, Zhang S, Zhi D. RAFFI: Accurate and fast familial relationship inference in large scale biobank studies using RaPID. PLoS Genet. 2021;17: e1009315.

68. Zhou Y, Browning SR, Browning BL. IBDkin: fast estimation of kinship coefficients from identity by descent segments. Bioinformatics. 2020;36: 4519–4520.

69. Kang JTL, Goldberg A, Edge MD, Behar DM, Rosenberg NA. Consanguinity rates predict long runs of homozygosity in Jewish populations. Hum Hered. 2016;82: 87–102.

70. Garrison NA. Genomic justice for native Americans: Impact of the Havasupai case on genetic research. Sci Technol Human Values. 2013;38: 201–223.

71. After Havasupai litigation, Native Americans wary of genetic research. Am J Med Genet A. 2010;152A: fmix.

72. Chen F, Dow M, Ding S, Lu Y, Jiang X, Tang H, et al. PREMIX: PRivacy-preserving EstiMation of Individual admiXture. AMIA Annu Symp Proc. 2016;2016: 1747–1755.

73. He D, Furlotte NA, Hormozdiari F, Joo JWJ, Wadia A, Ostrovsky R, et al. Identifying genetic relatives without compromising privacy. Genome Res. 2014;24: 664–672.

74. Robinson M, Glusman G. Genotype fingerprints enable fast and private comparison of genetic testing results for research and direct-to-consumer applications. Genes (Basel). 2018;9: 481.

75. Dervishi L, Wang X, Li W, Halimi A, Vaidya J, Jiang X, et al. Facilitating federated genomic data analysis by identifying record correlations while ensuring privacy. arXiv [cs.CR]. 2022. Available: http://arxiv.org/abs/2203.05664

76. Cheng JY, Mailund T, Nielsen R. Fast admixture analysis and population tree estimation for SNP and NGS data. Bioinformatics. 2017;33: 2148–2155.

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

78. Li Y, Byun J, Cai G, Xiao X, Han Y, Cornelis O, et al. FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. BMC Bioinformatics. 2016;17: 122.

79. Byun J, Han Y, Gorlov IP, Busam JA, Seldin MF, Amos CI. Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure. BMC Genomics. 2017;18: 789.

80. Gentry C. A FULLY HOMOMORPHIC ENCRYPTION SCHEME. PhD Thesis. 2009; 1–209.

81. Kowalski MH, Qian H, Hou Z, Rosen JD, Tapia AL, Shan Y, et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. 2019;15. doi:10.1371/journal.pgen.1008500

82. Matthew P. Conomos, Stephanie M. Gogarten, Lisa Brown, Han Chen, Thomas Lumley, Ken Rice, Tamar Sofer, Adrienne Stilp, Timothy Thornton, Chaoyu Yu. pcair: PC-AiR: Principal Components Analysis in Related Samples in GENESIS: GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. 28 Jan 2021 [cited 14 Apr 2022]. Available: https://rdrr.io/bioc/GENESIS/man/pcair.html