Privacy-Aware Kinship Inference in Admixed Populations using Projection on Reference Panels

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

Abstract

Estimation of genetic relatedness, or kinship, is used occasionally for recreational purposes and in forensic applications. While numerous methods were developed to estimate kinship, they suffer from high computational requirements and often make an untenable assumption of homogeneous population ancestry of the samples. Moreover, genetic privacy is generally overlooked in the usage of kinship estimation methods. There can be ethical concerns about finding unknown familial relationships in 3rd party databases. Similar ethical concerns may arise while estimating and reporting sensitive population-level statistics such as inbreeding coefficients for the concerns around marginalization and stigmatization. Here, we make use of existing reference panels with a projection-based approach that simplifies kinship estimation in the admixed populations. We use simulated and real datasets to demonstrate the accuracy and efficiency of kinship estimation. We present a secure federated kinship estimation framework and implement a secure kinship estimator using homomorphic encryption-based primitives for computing relatedness between samples in 2 different sites while genotype data is kept confidential.

Introduction

Genetic relatedness or kinship between two individuals 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 [1,2]. The kinship coefficient is closely related to other metrics such as the inbreeding coefficient [3] and IBD-sharing probabilities [4], which are essential for estimating population-level genetic information. Kinship estimates are central in behavioral science [5], human evolution [6], linkage mapping studies [7], and association studies [8–10] for the correction of biases caused by cryptic relatedness [9,11]. Numerous computational methods are developed to estimate kinship from marker genotypes but privacy and ethical concerns are sidelined. Kinship statistics are sensitive to individual privacy as they can be used to detect relatives in 3rd party databases without the consent of the owners, for example, by law

enforcement [12,13]. Similarly, population-level inbreeding estimates can cause marginalization and stigmatization risks [14–16]. In addition, it is well known that genetic data is very identifying due to its high dimensionality [17–20] and numerous "attacks" have demonstrated that databases can be linked [21–23] to reveal sensitive information. Similarly, genotypes can be recovered [24–26] and sensitive phenotypes can be inferred [27–31] using a small number of marker genotypes. Much of these attacks implicate and create discrimination and stigmatization risks to individuals and their families [32–35]. Therefore genetic kinship estimation presents numerous unaccounted challenges regarding individual and kin privacy [32,34,36].

Kinship estimation methods can be broadly divided into four categories [37]. Moment estimators such as KING [38], REAP [39], plink [40], GCTA [41], GRAF [42], and PC-Relate [43] use identical-by-state (IBS) markers and genotype distances to estimate expected kinship statistics. Maximum-likelihood methods (Such as RelateAdmix [44] and ERSA [45]) use expectation-maximization (EM) to jointly estimate the kinship statistics. Recent methods (such as RAFFI [46], IBDKin [47]) use fast algorithms to search for IBD matches from phased genotypes and estimate kinship from shared IBD estimates. There are also methods that estimate kinship from next-generation sequencing data, which are especially useful from low-coverage sequencing approaches (NGSRemix [48], LASER [49], SEEKIN [50]). While most methods can accurately estimate kinship for individuals with homogeneous ancestry, this is not a tenable assumption in admixed populations[2,51]. Moreover, non-random mating, i.e., assortative mating, among similar ancestral groups [52,53] may bias estimates of kinship. Methods that assume random mating or simple homogeneous populations are not effective in appropriately estimating kinship and may impact downstream analysis and interpretations. Several methods have been proposed for privacy-aware analysis of ancestry and admixture. PREMIX [54] computes admixture rates in a privacy-preserving manner using SGX-based extensions, which are currently deprecated on consumer-side processors. He et al. combined a genome sketching technique with cryptographic evaluation to search for relatives [55]. Similar sketching techniques have been proposed for fingerprint and relative search analysis [56]. Dervishi et al. proposed privacy-aware kinship estimation by integrating local differential privacy and genotypic data hiding [57], which may hinder the utility of genetic data. While these methods are promising, the impact of admixture is not generally taken into account, and the methods are evaluated 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 [49] in admixed populations. The modular formulation of SIGFRIED enables an efficient secure implementation. Usage of component analysis and reference populations with a "distance-based" estimation of admixture has shown promise in previous studies [58,59]. We capitalize on these and propose an efficient approach to estimate kinship, inbreeding, and IBD sharing probabilities. In comparison to previous methods, SIGFRIED imposes less computational burden without the requirement of compute-intensive admixture estimates, which are prohibitively challenging in secure implementations. We 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 [60], which enables processing encrypted genotype data directly without ever being decrypted and therefore provides provable security guarantees on the genetic data.

Results

We first describe an overview of SIGFRIED's estimation approach present accuracy results. After the new kinship estimation results, we formulate the secure implementation and present the secure collaborative kinship analysis scenario.

Kinship Estimation using Projection-based Admixture Estimates

Figure 1 summarizes the kinship estimation approach by SIGFRIED. Kinship estimation takes a query genotype matrix that contains S individuals for which $S \times S$ kinship related statistics are computed. SIGFRIED utilizes principal components and representatives computed from a reference population panel that contains S_{tot} individuals and n_{ref} populations. The reference panel genotype is first decomposed into components and for each population, we compute a representative sample. Given the query genotype matrix for S individuals, we project the genotypes to the top components of the reference panel computed in the previous step. We next compare each sample to the representatives and assign admixture rates using a non-linear function of genotypes. We finally assign the allele frequencies, μ, using the admixture rates. These operations have efficient secure implementations in homomorphic encryption [60] and can be justifiably used in this scenario [61].

Kinship Coefficients. We implement two kinship coefficients. First is the correlation metric, $\phi^{(\text{Corr.})} = \rho(G|\mu)$, between the individuals. The second kinship metric we use is a novel genotype distance-based metric, $\phi^{(\text{Dist.})} = \Delta(G|\mu)$, which integrates individual-specific allele frequencies. For a privacy-aware implementation, the distance and correlation-based can be computed using different strategies. Sites must share the genotypes and allele frequencies. Allele frequencies do not immediately reveal genetic information but they correlate significantly with actual genotypes and may need to be encrypted. These statistics can also be computed in parallel and the final kinship statistic can be aggregated at each site locally.

Zero-IBD Sharing Probability. We also report the moment estimator for zero IBD-sharing probability among individuals, which is derived from the expected number of zero identical-by-state (IBS) values:

$$\delta^0_{i,j} = \frac{\text{Number of IBS} = 0 \text{ between } i \text{ and } j}{\text{Expected Number of IBS} = 0 \text{ between } i \text{ and } j}$$

where $\delta^0_{i,j}$ denotes the probability of zero-IBD sharing among individuals i and j.

Inbreeding Coefficient. The inbreeding coefficient for each individual can be estimated from the correlation-based kinship estimator using the established relationship between kinship and inbreeding coefficients:

$$h_i = \left(2 \times \phi^{(\text{Corr.})}_{ii} - 1\right)$$

where $\phi^{(\text{Corr.})}_{ii}$ denotes the self-kinship coefficient and h_i denotes the inbreeding coefficient for i^{th} individual.

Parameter Selection

We evaluated the impact of the number of variants in the estimation of kinship statistics. For this, we simulated 50 homogeneous pedigrees and computed kinship statistics using SIGFRIED within each pedigree using an increasing number of variants from 500 variants up to 150,000 variants. As the number

of variants is increasing, the variance of kinship estimates decreases for each respective degree of relatedness. Figure 2c shows that adding more than 50,000 variants does not provide much change in the variance of the estimated kinship. Qualitatively, as small as 20,000 variants are sufficient for distinguishing 1st and 2nd-degree relatives.

Comparison of Methods

We compared the correlation and distance-based kinship estimators under homogeneous and heterogeneous pedigree scenarios. We mainly focused on comparing the approaches of SIGFRIED with REAP (with ADMIXTURE tool) and KING-Robust. For SIGFRIED, we use the correlation-based estimator and the projection-based admixture rate estimation to compute individual-specific allele frequencies. We also compare correlation-based and distance-based estimators using allele frequencies estimated by assuming uniform admixture rates over the reference populations, and by using the pooled reference as a single panel to estimate the variant allele frequencies.

Kinship Estimates in Pedigrees from Same Ancestry

We simulated 500 independent pedigrees where the founding members are randomly selected from a single European population among The 1000 Genomes Project samples. Within each pedigree, we computed the kinship and zero-IBD sharing probabilities between all pairs of members using KING-Robust[38], REAP, and the distance and correlation-based kinship and zero-IBD sharing probability statistics. For SIGFRIED's projection-based admixture estimates, we used 3 populations from the 1000 Genomes Project as the reference populations to ensure that the admixture estimation step is not trivially applied to a single reference. Overall, we observed that all correlation-based and distance-based methods performed similarly to assign the expected kinship and zero-IBD sharing probability estimates for different levels of kinship (Fig. 3). One observation is that distance-based estimators provide tighter estimates of kinship (Fig. 3c, d), compared to the correlation-based estimators (Fig. 3a, b). Considering that distance-based estimators also have lower computational requirements, these results suggest that they may be more suitable than correlation-based estimators for samples with homogeneous ancestries.

Kinship Estimates in Pedigrees from Admixed Ancestry

We next tested the estimation of kinship in admixed ancestries. In the simulation, the founders were selected randomly from populations of European, East Asian, and African descent in The 1000 Genomes Project. For admixed ancestries, we compared the correlation-based estimator using the admixture rates estimated by ADMIXTURE and also with a uniform assignment of admixtures that is equally distributed among 3 reference populations as a control method. We also compared the distance-based estimator with projection-based admixture rates and KING-Robust. In comparison, projection-based estimators and ADMIXTURE-based estimators provide the most accurate results for relatives up to 4th degree (Fig. 4a). KING-Robust underestimates the kinship coefficient, especially for unrelated individuals. Our distance-based estimator largely corrects the negative and heterogeneous trend of KING-Robust. The distribution of kinship coefficients (Fig. 4) indicates that the correlation-based estimators provide single exact peaks around the expected kinship values (Fig. 4b, c). Our novel distance-based estimator exhibits single peaks except for unrelated individuals, for which there is a second peak in negative values. On the other hand, KING-Robust exhibits a fairly high deviation from the expected values with no clear peaks (Fig. 4d, e), which demonstrates the advantage of using a modified distance metric. A similar heterogeneous distribution of kinship is observed for correlation-based estimators that use the pooled reference sample or uniformly assigned admixture rates (Fig. 4f, g).

Time and Memory Requirements

We next compared the time and memory requirements of the estimators. To compare the resource requirements of the methods, we estimated the memory and time requirements of SIGFRIED, REAP-ADMIXTURE, and KING-Robust. For all methods, we measured the total time required for admixture estimation, and kinship statistic computations and also the peak memory required for these steps. Overall, KING-Robust runs the fastest and uses the smallest amount of memory (Fig. 5a, b). REAP-ADMIXTURE runs the slowest wherein the majority of time is spent on the estimation of the admixture rates by ADMIXTURE. SIGFRIED runs at least 3 times faster than REAP-ADMIXTURE's workflow. To test the way that methods scale with the number of reference populations, we compared the resource usage by increasing the number of reference populations (Fig. 5a, b). REAP-ADMIXTURE's runtime exhibits approximately linear increase in the number of reference populations. On the other hand, SIGFRIED shows a sublinear increase. This indicates that for large admixed populations SIGFRIED's projection-based approach can provide good accuracy with less computational resource requirements.

Secure Federated Estimation of Kinship Statistics in Two-Site Setting

One of the main advantages of SIGFRIED over previous approaches is enabling privacy-aware kinship estimation in different scenarios due to its modular formulation. We focus on a 2-site collaborative scenario (such as genealogy companies or two institutions working under different regulations) where the sites aim at computing the pairwise kinship statistics among the collective set of individuals in two sites but they cannot share genotype data in plaintext format because of local privacy requirements. We also assume that the sites behave honestly without collusions or malicious data manipulations [62]. This scenario is illustrated in Figure 6a. The two sites have the genotypes matrices $G^{(1)}$ and $G^{(2)}$ for S_1 and S_2 individuals. The main task is to compute the kinship coefficients between all pairwise comparisons of S_1 and S_2 individuals among the sites. We assume the sites utilize the same reference panels to perform projection-based estimation of admixtures and the individual specific AFs for each individual locally.

Secure Computation of Correlation-based Kinship Coefficient. We first compute a normalized genotype matrix for each site, $\Gamma^{(a)}$, which denotes the normalized genotype matrix for site a by correcting with respect to allele frequencies. $\phi_{i,j}$ is computed from the normalized genotype matrices. An important observation is that normalized genotype matrices in each site can be computed locally and do not depend on the other site's private information. However, this still requires the sites to share the normalized genotype matrices in plaintext format with each other. It is therefore necessary to protect at least one of the matrices by encryption (Fig. 6a). We make use of *homomorphic encryption* to secure the data [61], which enables the processing of the encrypted data without decrypting it. In this setup, both sites compute the normalized matrices and Site-2 homomorphically encrypts and sends its encrypted genotype matrix to Site-1 (or vice versa). We denote the encrypted normalized genotype matrix of Site-2 with $\check{\Gamma}^{(2)}$. After Site-1 receives encrypted genotypes, it computes the kinship coefficient securely using $\Gamma^{(1)}$ and $\check{\Gamma}^{(2)}$. Finally, the computed kinship estimates are sent back to Site-2, which decrypts and shares the kinship coefficient matrix with Site-1.

Other Kinship Statistics. Other kinship statistics such as zero-IBD sharing probabilities and distance-based kinship estimator can be securely calculated using an approach similar to as described above under different scenarios.

Time and Memory requirements. We implemented a 2-Site kinship estimation using the SEAL library [63]. We used the CKKS encryption scheme with default security settings that satisfy 128-bit security requirements [64]. We used 100 simulated individuals and used 15,000 variants whereby each individual's genotypes fit into multiple ciphertexts. This proof-of-concept implementation finished the computation kinship coefficients in under 1 minute using less than 4 gigabytes of main memory, which includes encryption, encoding, evaluation, decoding, and decryption time. Overall, we observed that the maximum absolute difference between plaintext and encrypted kinship coefficients is 10^{-9}, which practically does not cause differences in analysis of relatedness. The secure computations can be extended to an optimized version of secure federated kinship statistics in multi-site with and without an untrusted outsourcing entity such as a cloud-based server (Fig. 6b). As the data encryption is implemented in our protocol, even untrusted entities can be used in federated kinship estimation for making use of large cloud-based scaling for improved performance [65,66].

Discussion

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 is efficiently 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. From this perspective, we view SIGFRIED as a private-by-design methodology wherein the privacy considerations are balanced against efficiency and accuracy. Projection does not explicitly require reference panel genotypes. Since the reference genotypes are not explicitly shared, this creates minimal risk for reference panels under restricted access (i.e. TOPMed [67]).

While we presented a specific privacy-preserving scenario for a 2-site federated estimation of kinship, the implementation and the scenarios can be differently set up 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 using modular steps and decomposable functions. This is beneficial for optimizing privacy-vs-performance in different scenarios. The modularity 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 are highly aggregated functions of genotypes and can be deemed safe to share in plaintext form.

Methods

Variant Selection and Simulations

For simulations, we filtered the variants in The 1000 Genomes Project by first selecting the variants with minor allele frequencies greater than 5% on the autosomal chromosomes 1 through 22. Next, we divided the samples with respect to their assigned populations and used these as population specific reference panels. Simulations are performed by sampling variants for each individual with respect to allele frequencies and the relatedness.

Projection-based Estimation of Kinship Statistics

Figure 1 summarizes the kinship estimation approach by SIGFRIED. Kinship estimation takes a query genotype matrix, $G_{N\times S}$, that contains the genotypes of N variants for S individuals. The output is $S \times S$ matrix of kinship related statistics. The kinship statistics are calculated and implemented by the formulations that are presented in the Results Section using correlation and distance-based statistics.

Source Code and Data Availability

Source code and Datasets will be made available upon publication of this manuscript.

Figure Legends

Figure 1. **(a)** Block diagram illustrates the steps for computation of kinship coefficients.

Figure 2. The distribution of assigned admixture rates by ADMIXTURE **(a)** and by projection-based admixture estimation **(b)**. **(c)** The kinship coefficient (x-axis) distribution with different number of variants. Each plot shows a kinship distribution generated using number of variants indicated at the label.

Figure 3. Scatter plots of kinship coefficients in 500 pedigrees from homozygous ancestries. **(a)** kinship coefficient (y-axis) versus Zero-IBD sharing probabilities (x-axis) by SIGFRIED. **(b)** REAP estimates. **(c)** Distance-based estimates. **(d)** KING-Robust estimates.

Figure 4. **(a)** Average kinship coefficient estimated by each method for heterogeneous populations. **(b)** Distribution of SIGFRIED kinship estimates. **(c)** Distribution of REAP estimates. **(d)** Distribution of distance-based kinship estimates. **(e)** Distribution of kinship estimates from KING-Robust. **(f,g)** Distribution of correlation-based kinship estimates using uniform and all-population admixture assignments for every sample.

Figure 5. Time and memory requirements of kinship estimation. **(a)** Time requirements (y-axis) by different methods. **(b)** Memory usage (y-axis) by kinship estimation methods.

Figure 6. Illustration of secure kinship and IBD-Sharing probability estimation for 2-Site collaboration. **(a)** Site-2 computes the normalized genotypes $\Gamma_{i,j}^{(2)}$ and sends them to Site-1 after encrypting them with the public key. Site-1 also compute the normalized genotype matrix, $\Gamma_{i,j}^{(1)}$. After receiving the encrypted genotype matrix from Site-2, Site-1 securely estimates the encrypted kinship ($\check{\phi}_{i,j}$) and other statistics. Site-1 sends the encrypted matrices to Site-2, which decrypts the kinship statistics and shares them with Site-2. **(b)** Illustration of a secure kinship estimation with an outsourcing server. The two sites compute normalized genotype matrices. The sites encrypt genotype matrices and send them to the kinship estimation server. The server pools all encrypted data and securely estimates kinship statistics among all samples. The encrypted kinship statistics are then sent to each site, each of which decrypt the kinship statistics.

References

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

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

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

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

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

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

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

8. 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.

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

10. 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.

11. 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

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

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

14. 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.

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

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

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

18. 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.

19. 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.

20. 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

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

22. 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

23. 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.

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

25. 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.

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

27. 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.

28. 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

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

30. 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

31. 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

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

33. 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

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

35. 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.

36. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nat Genet. 2020;52: 646–654.

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

38. 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.

39. 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.

40. 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.

41. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88: 76–82.

42. 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.

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

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

45. 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.

46. 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.

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

48. 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

49. 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.

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.

55. 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.

56. 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.

57. 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

58. 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.

59. 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.

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

61. Cheon JH, Kim A, Kim M, Song Y. Homomorphic encryption for arithmetic of approximate numbers. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017. pp. 409–437.

62. Dhir R, Patel AA, Winters S, Bisceglia M, Swanson D, Aamodt R, et al. A multidisciplinary approach to honest broker services for tissue banks and clinical data: a pragmatic and practical model. Cancer. 2008;113: 1705–1715.

63. Benaissa A, Retiat B, Cebere B, Belfedhal AE. TenSEAL: A library for encrypted tensor operations using Homomorphic Encryption. arXiv [cs.CR]. 2021. Available: http://arxiv.org/abs/2104.03152

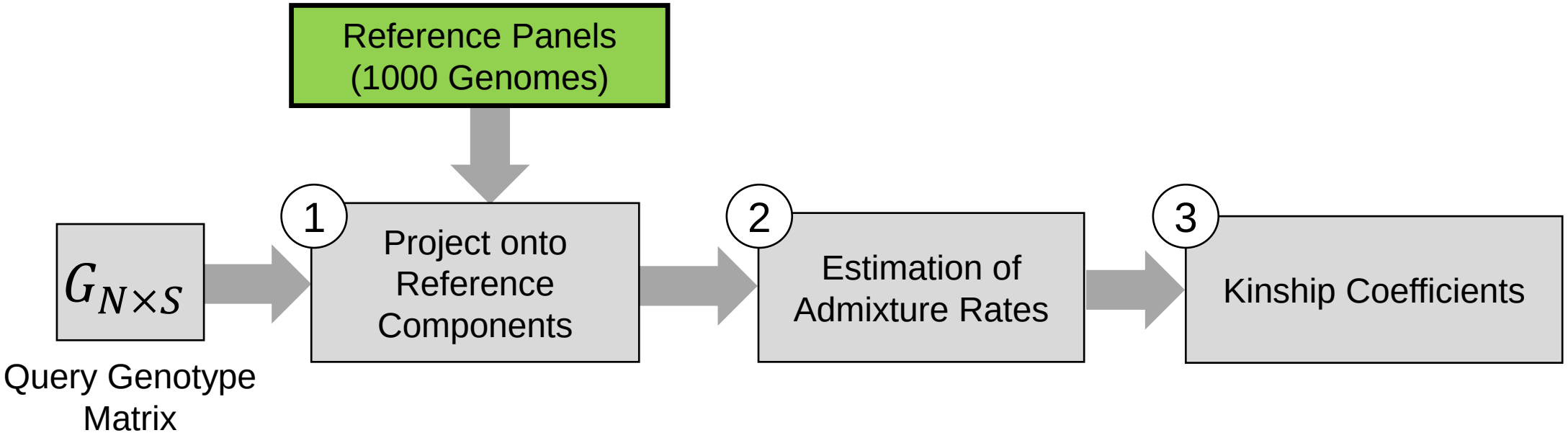
64. Albrecht M, Chase M, Chen H, Ding J, Goldwasser S, Gorbunov S, et al. Homomorphic Encryption Standard. 2018 [cited 18 Apr 2022]. Available: http://homomorphicencryption.org/wp-content/uploads/2018/11/HomomorphicEncryptionStandardv1.1.pdf

65. Kim M, Harmanci AO, Bossuat J-P, Carpov S, Cheon JH, Chillotti I, et al. Ultrafast homomorphic encryption models enable secure outsourcing of genotype imputation. Cell Systems. 2021;12: 1108-1120.e4.

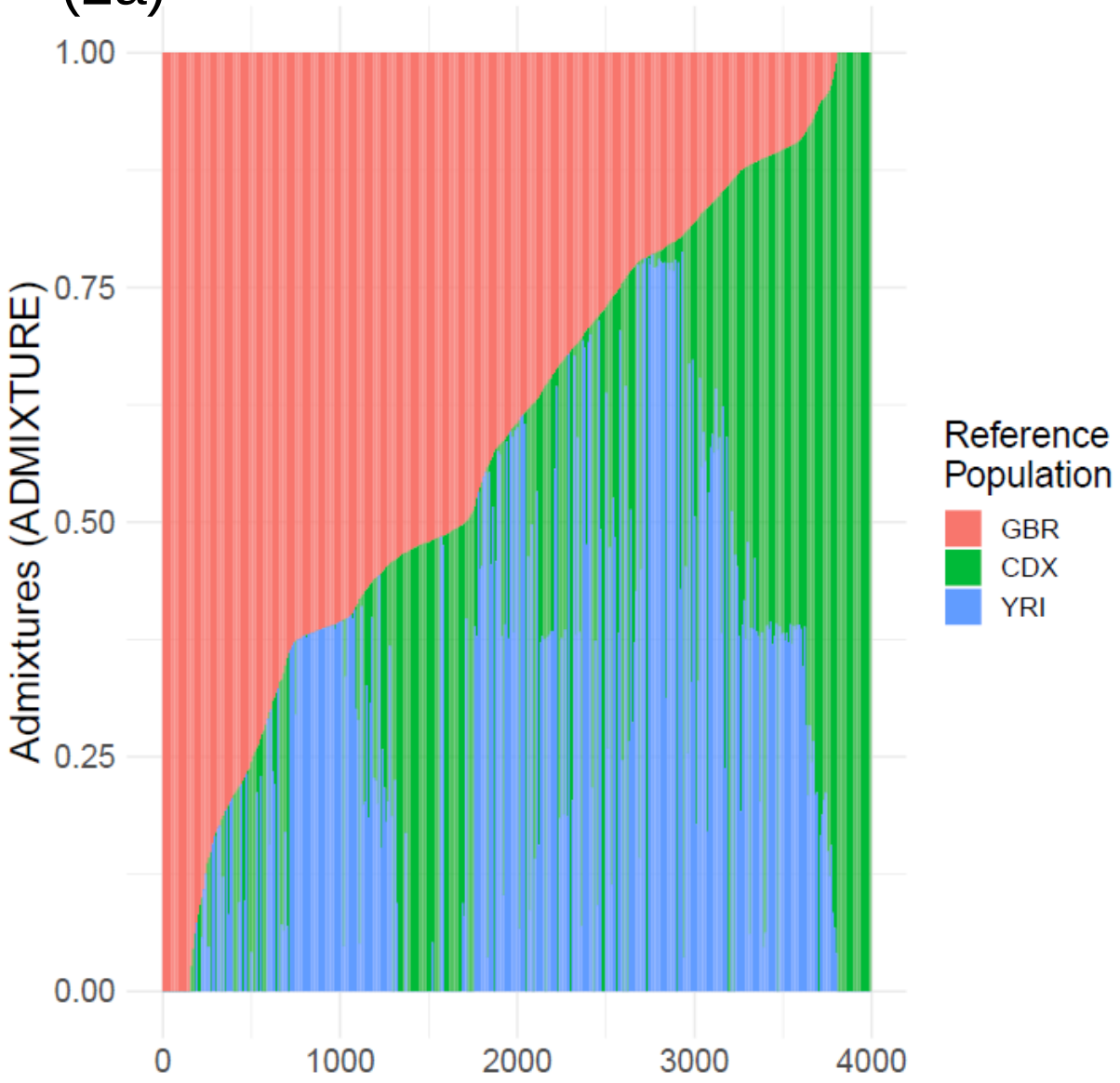
66. Harmanci AO, Kim M, Wang S, Li W, Song Y, Lauter KE, et al. Open Imputation Server provides secure Imputation services with provable genomic privacy. bioRxiv. 2021; 2021.09.30.462262.

67. 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

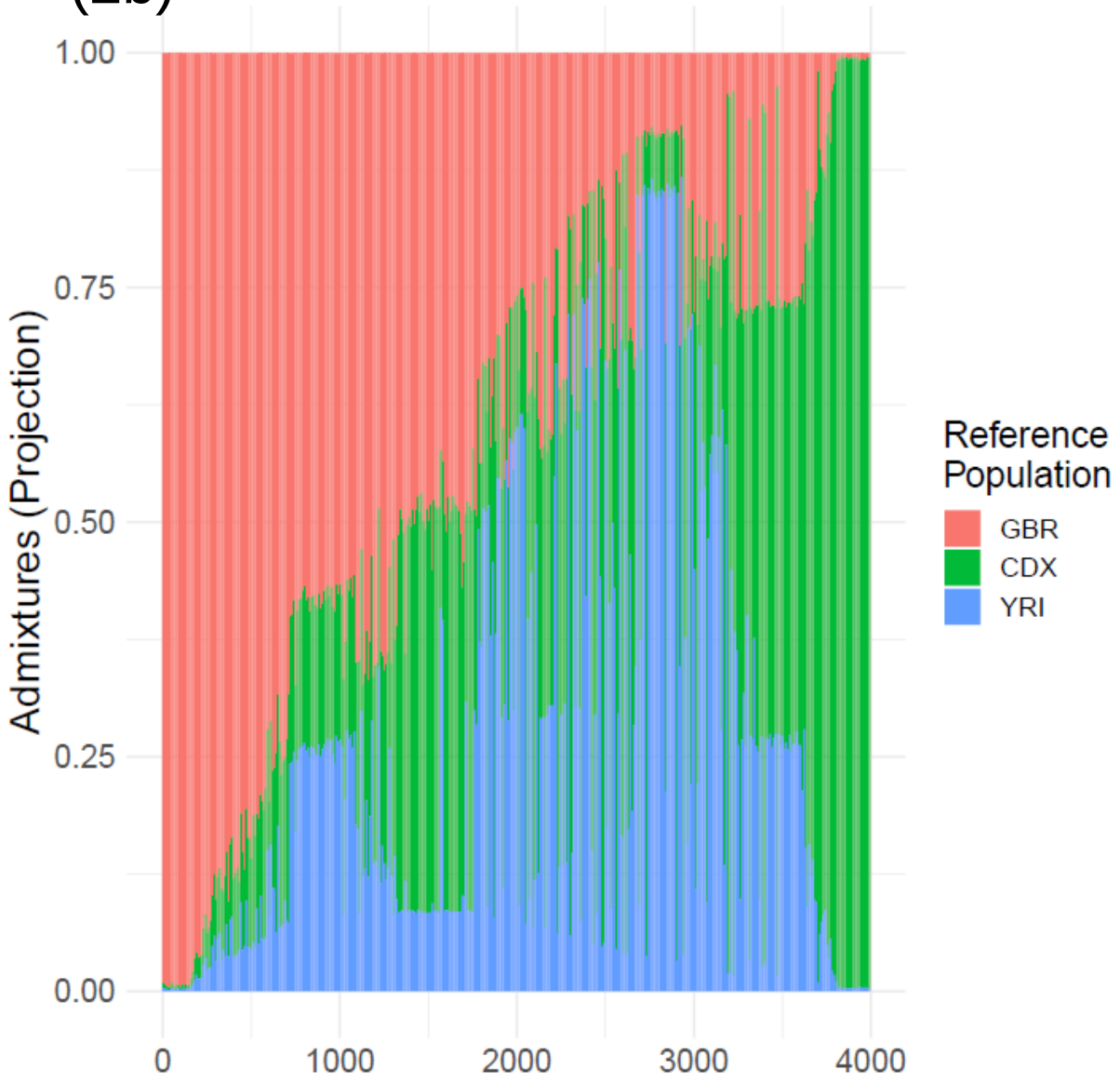
Fig 1: Overview of Kinship Estimation



(2a)


1.00
0.75
0.50
0.25
0.00
Admixtures (ADMIXTURE)
0
1000
2000
3000
4000
Reference Population
GBR
CDX
YRI

(2b)


1.00
0.75
0.50
0.25
0.00
Admixtures (Projection)
0
1000
2000
3000
4000
Reference Population
GBR
CDX
YRI

(2c)

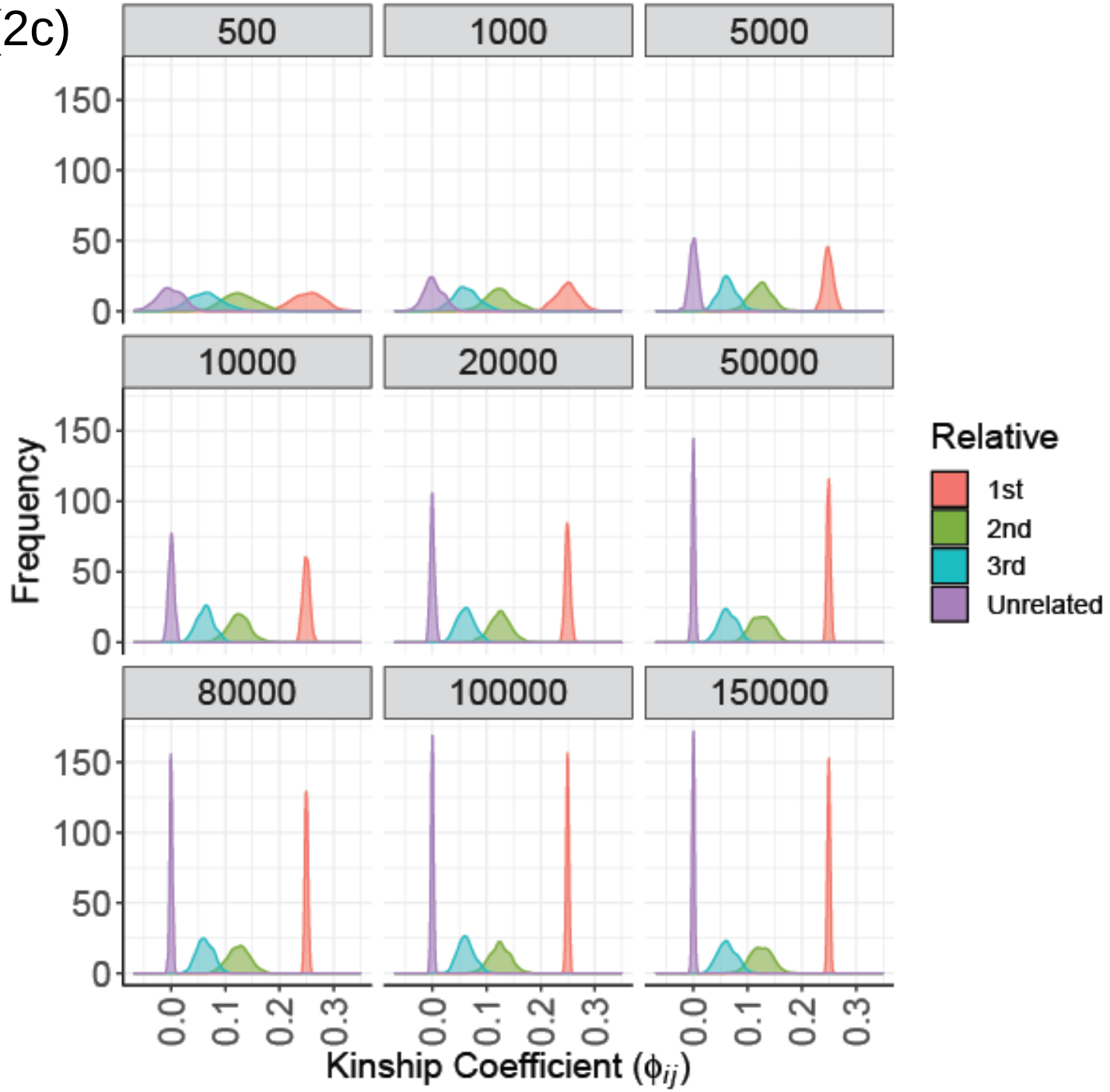

500
1000
5000
10000
20000
50000
80000
100000
150000
Frequency
150
100
50
0
0.0
0.1
0.2
0.3
Kinship Coefficient (ϕ_{ij})
Relative
1st
2nd
3rd
Unrelated

Fig. 3: Homozygous Ancestry

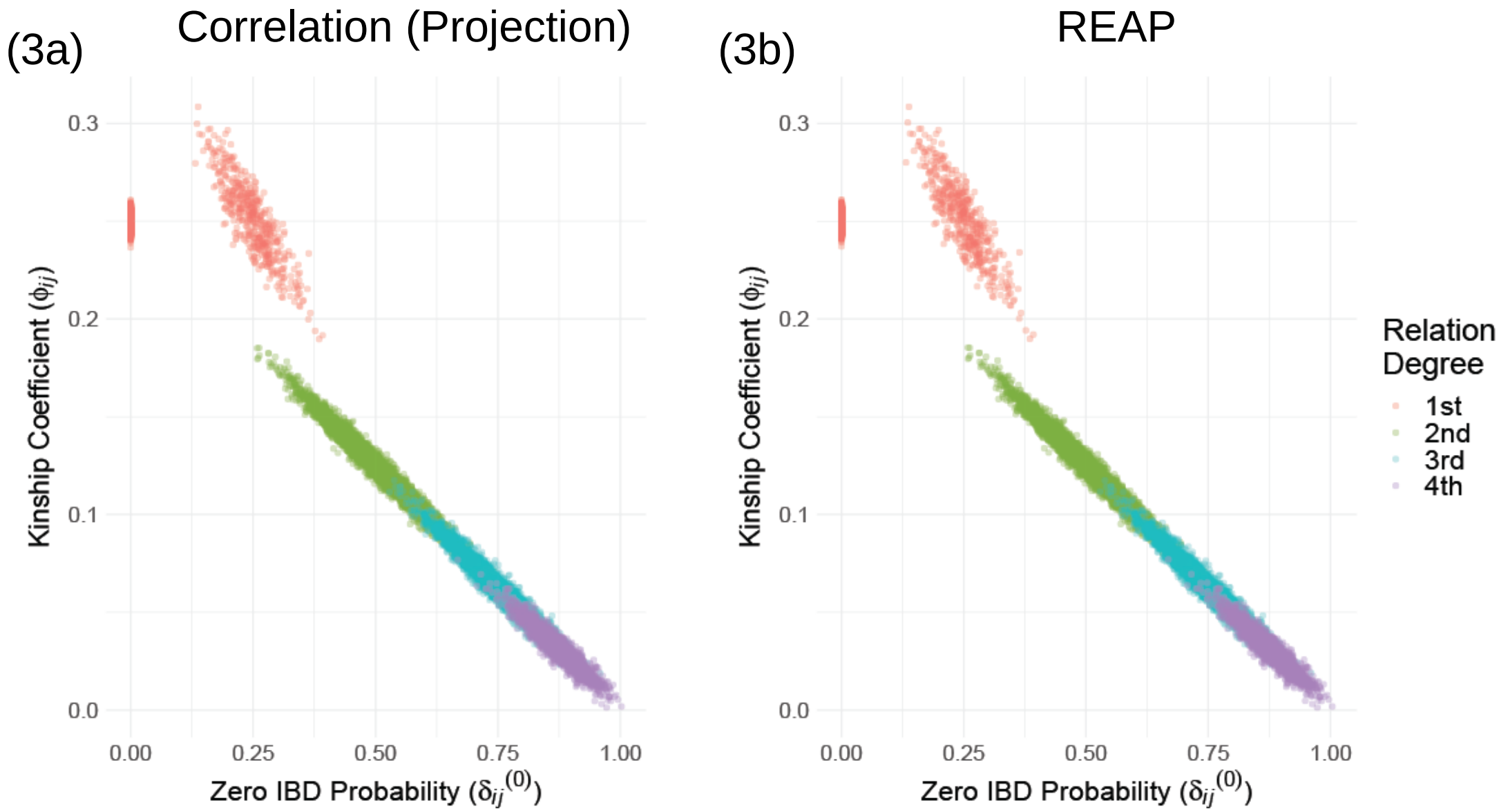


Fig. 3: Homozygous Ancestry

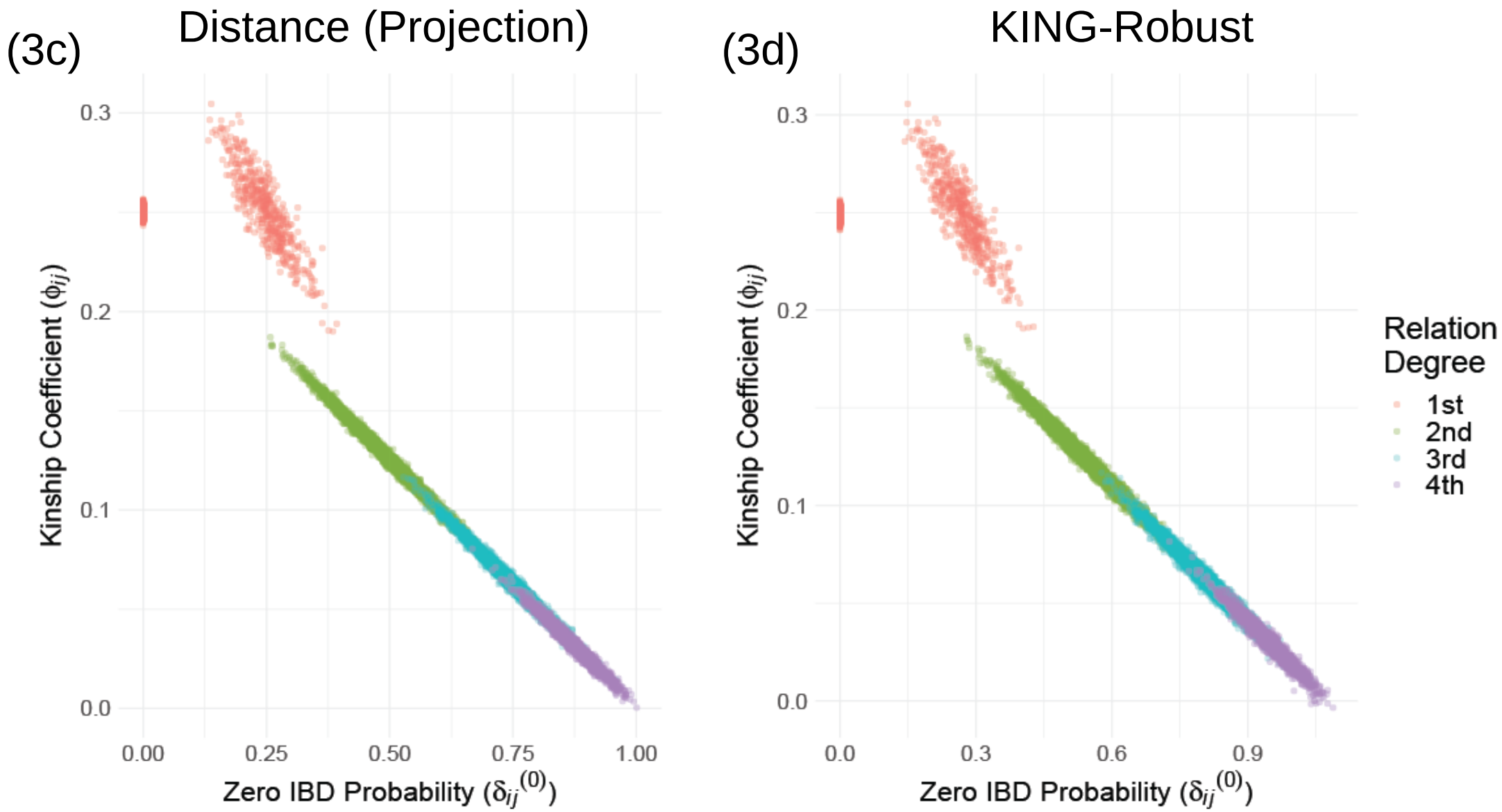


Fig. 4: Heterozygous Ancestry

(4a)

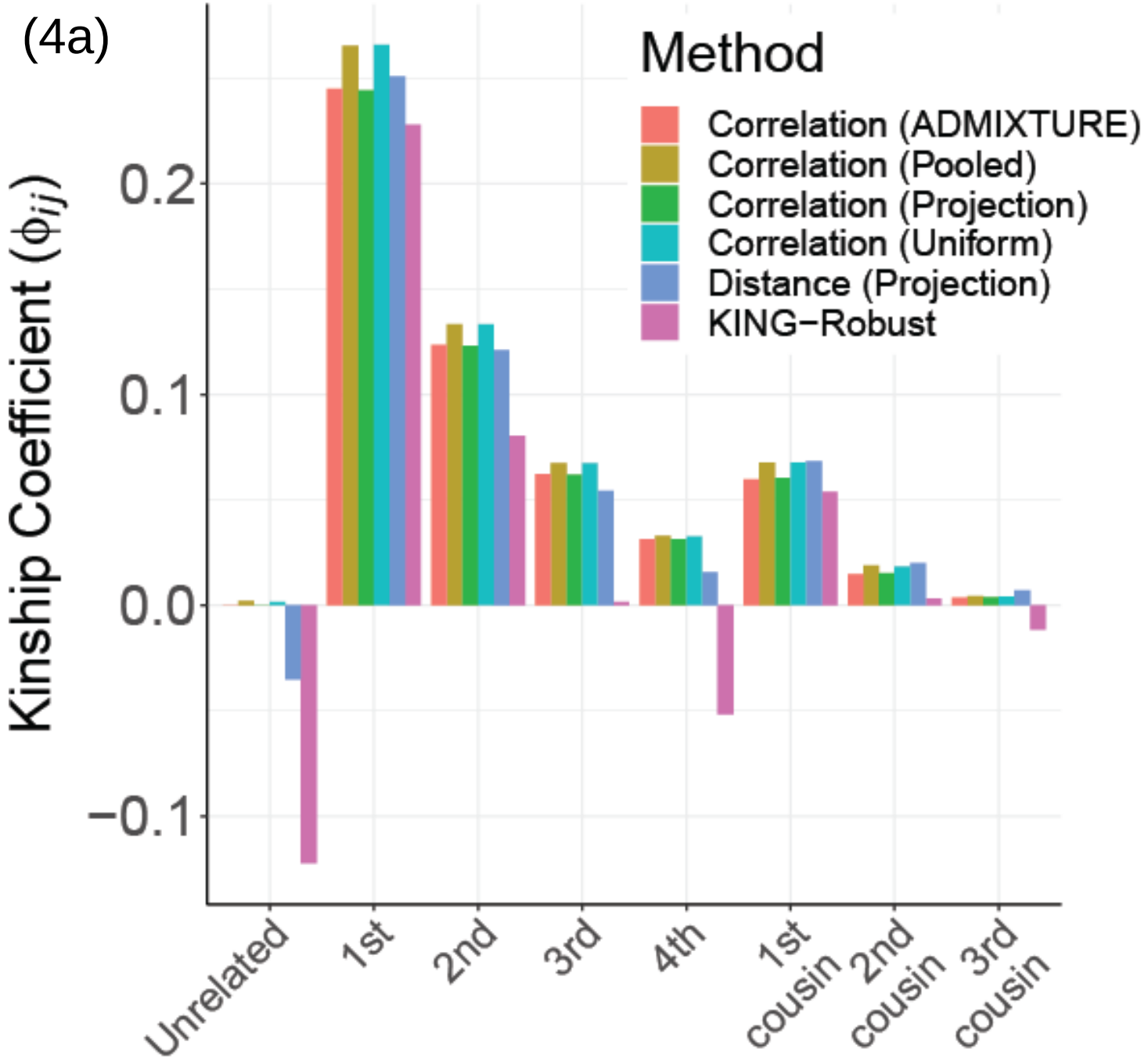


Fig. 4: Heterozygous Ancestry

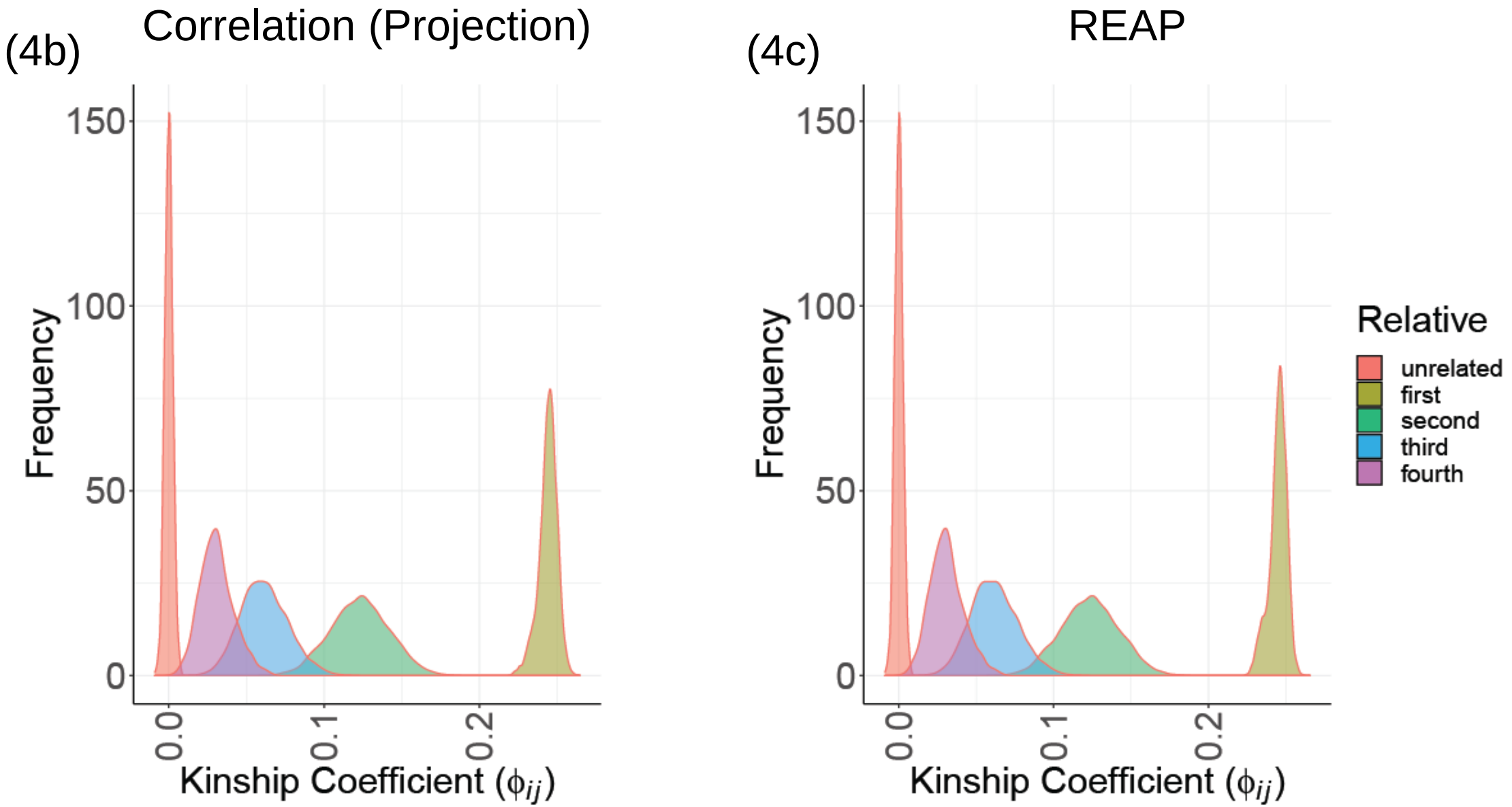


Fig. 4: Heterozygous Ancestry

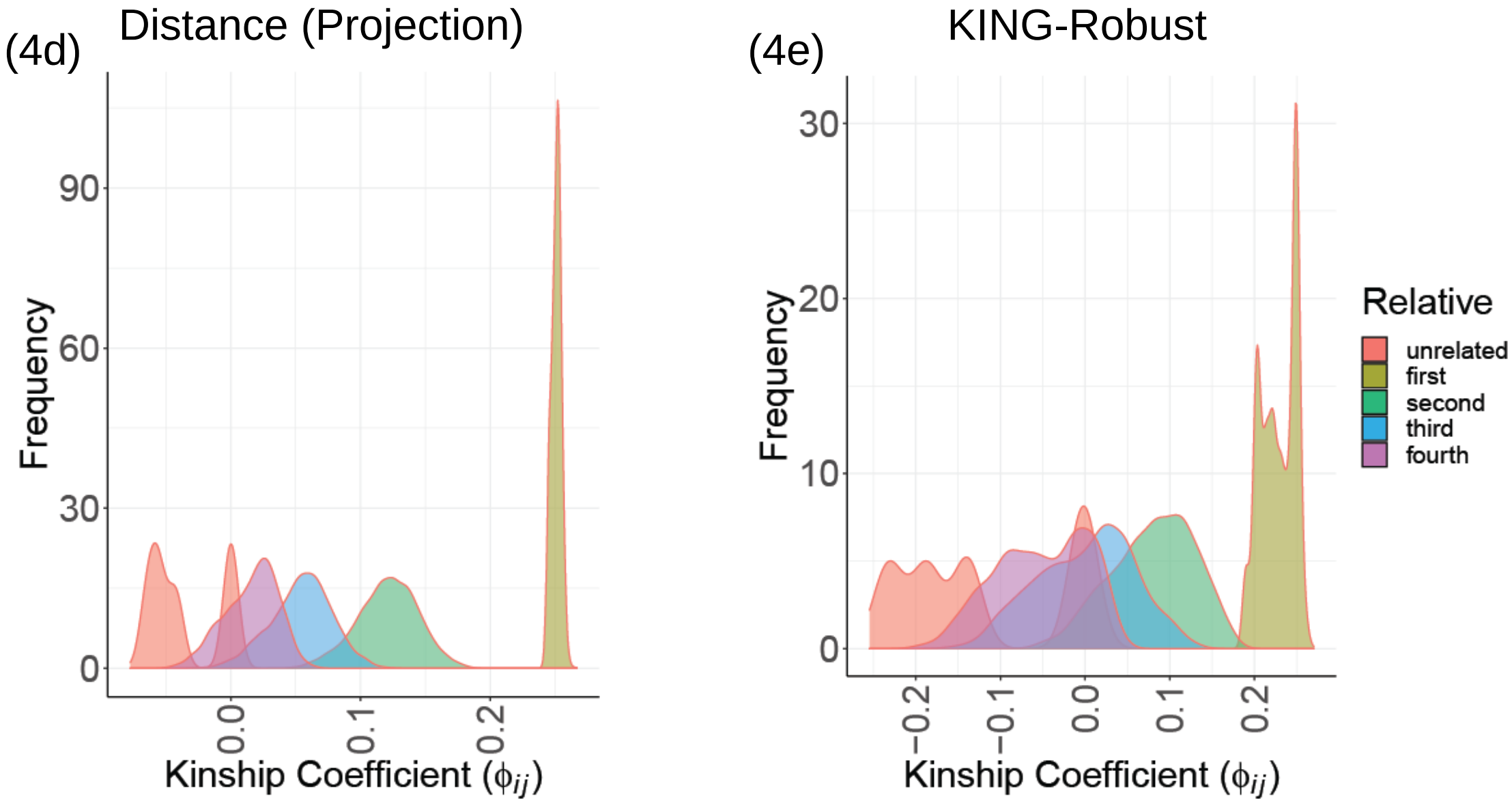
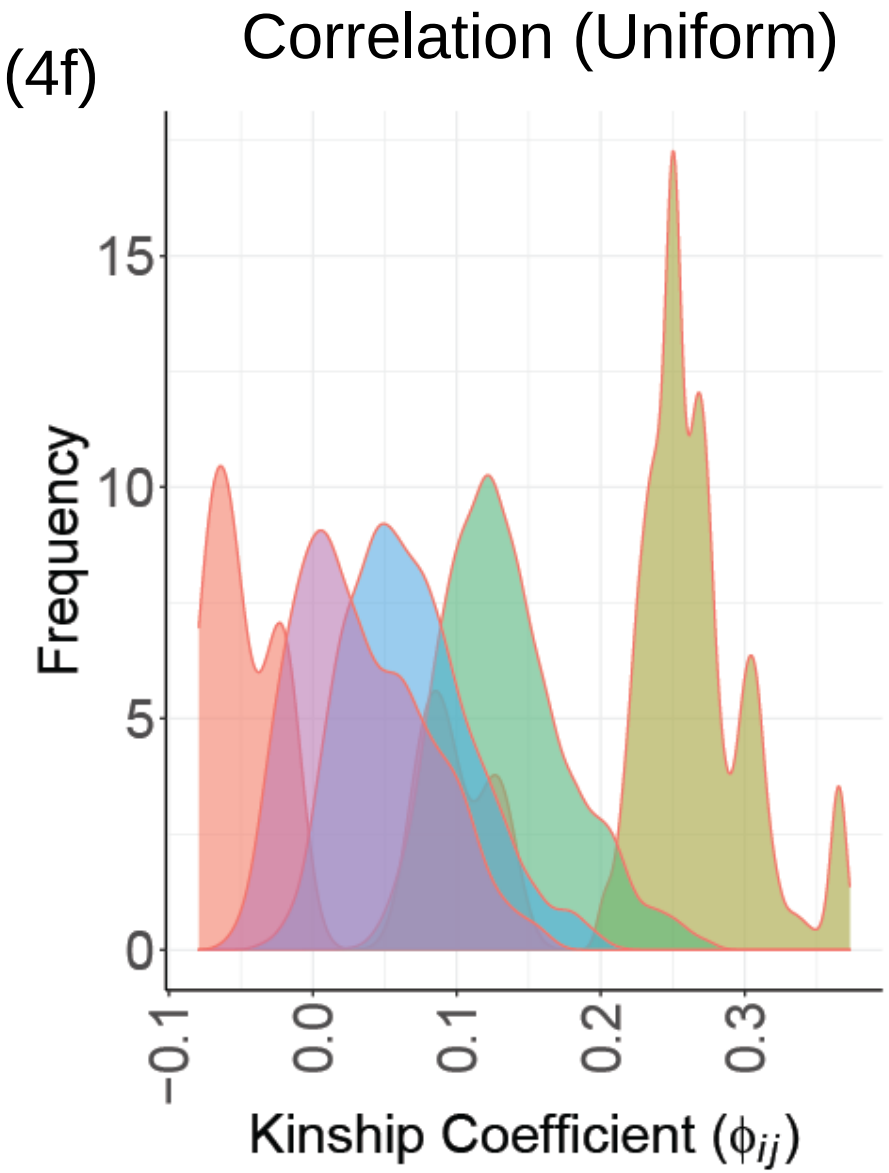


Fig. 4: Heterozygous Ancestry

(4f)



(4g)

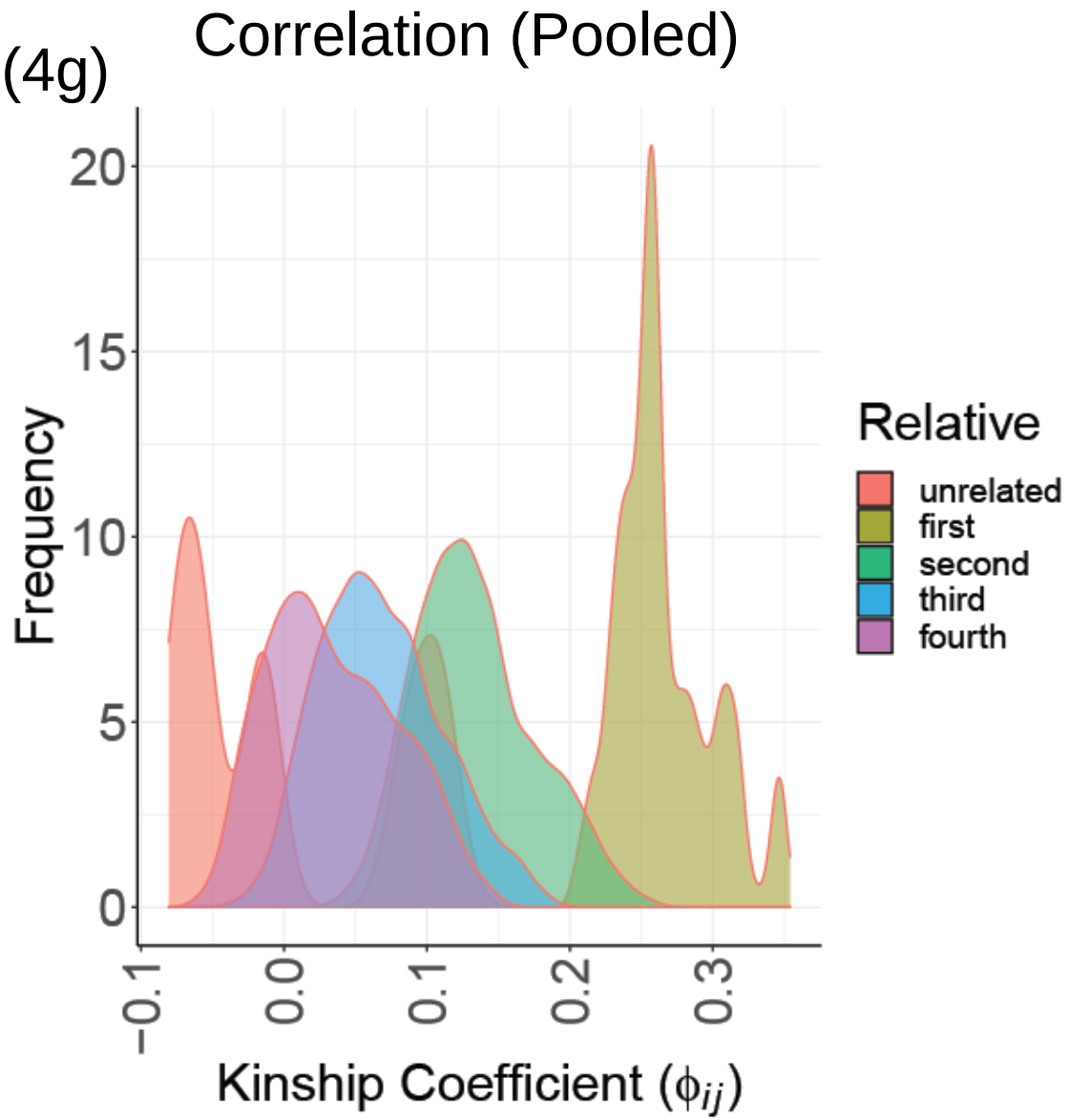
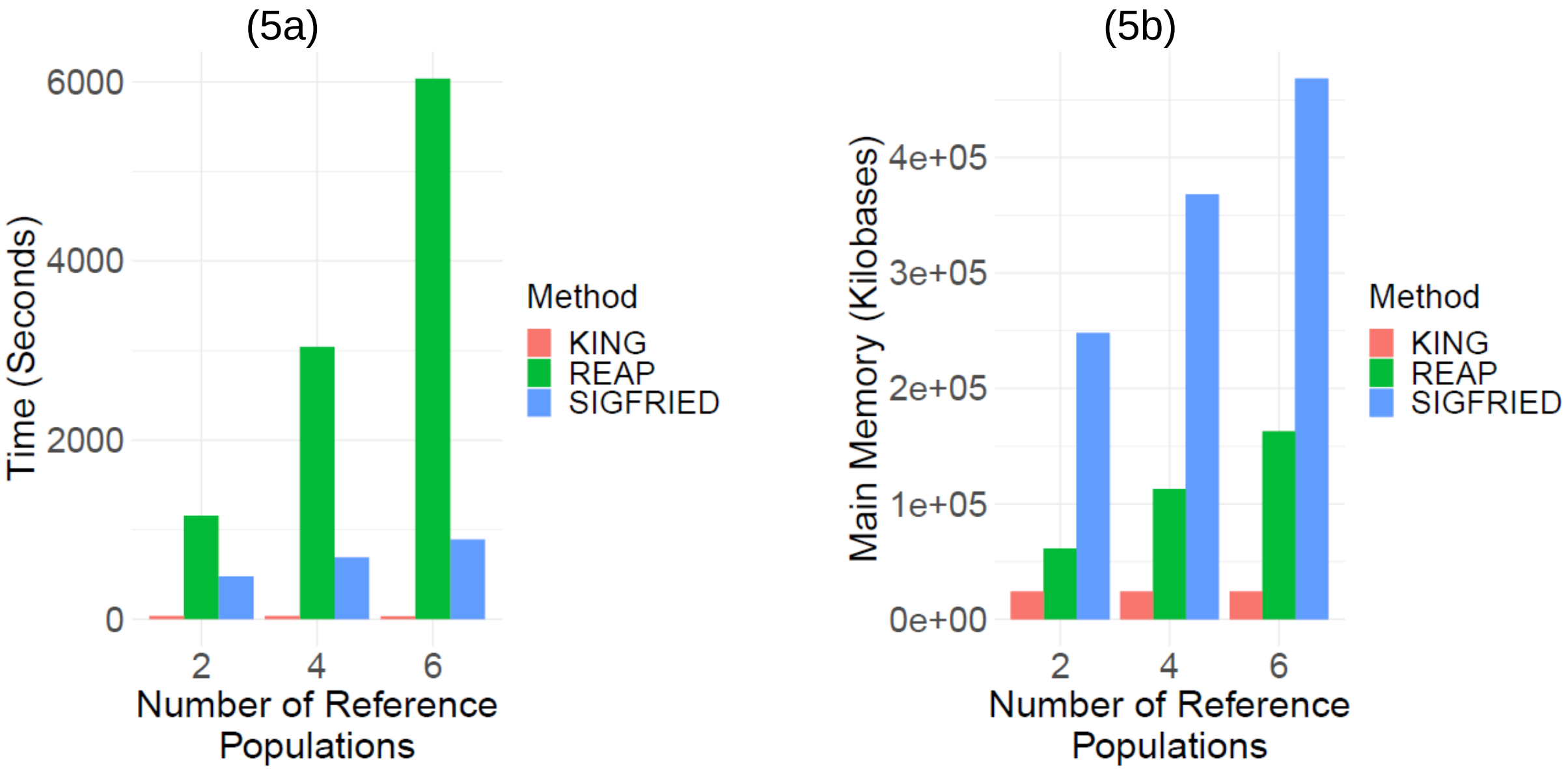


Fig. 5: Time and Memory Usage



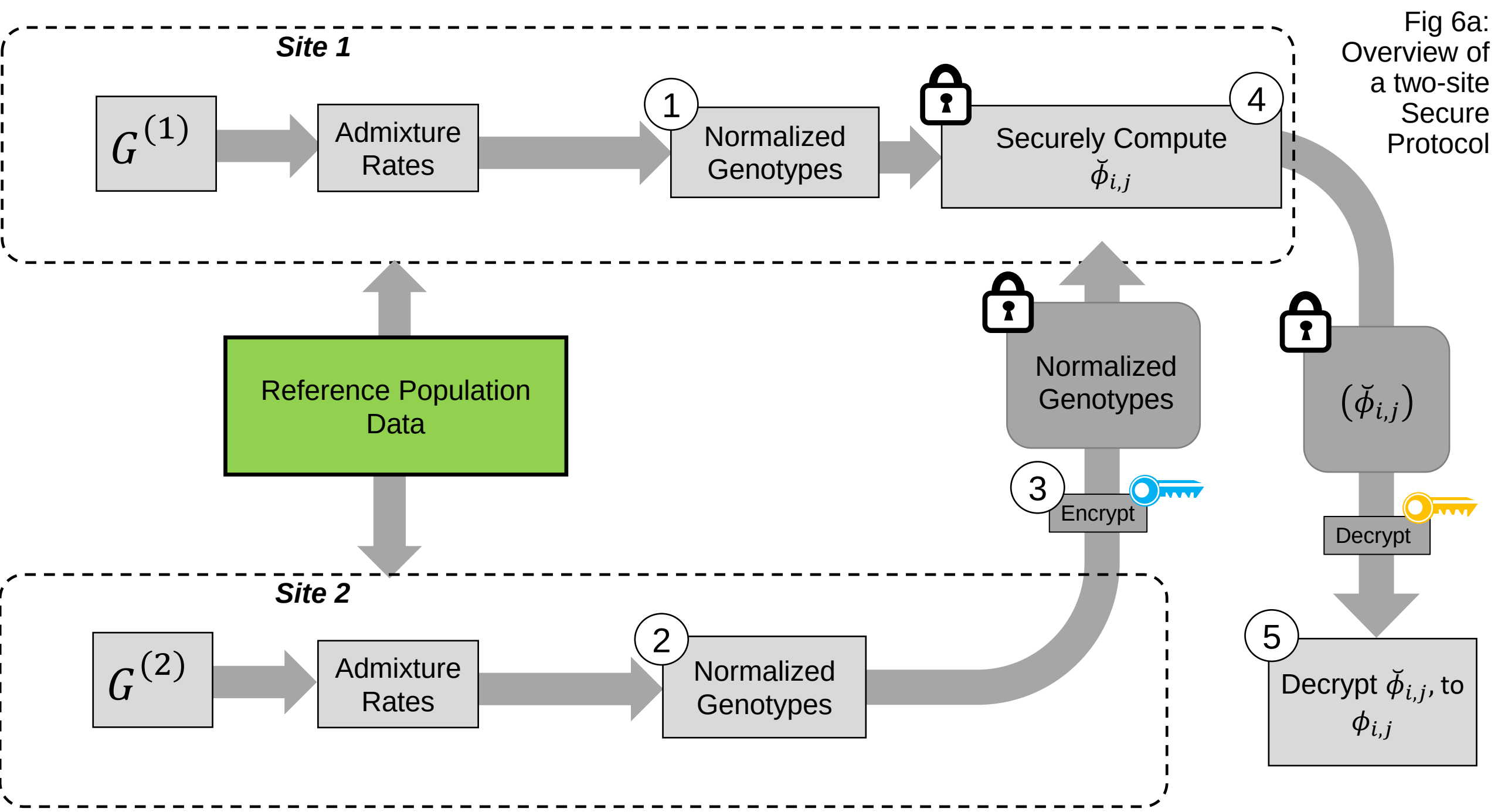


Fig 6a: Overview of a two-site Secure Protocol

Fig 6b: Utilization of an Outsourcing Site