Evaluating the impact of dropout and genotyping error on SNP-based kinship analysis with forensic samples

Stephen D. Turner, VP Nagraj, Matthew Scholz, Shakeel Jessa, Jianye Ge, August Woerner, Meng Huang, Mike Coble, Bruce Budowle

Background: Inferring familial relationships between individuals using genetic data is a common practice in population genetics, medical genetics, and forensic genetic genealogy (FGG). Sequencing and microarray technology have enabled rapid profiling of millions of single nucleotide polymorphisms (SNPs) with near-perfect accuracy. With these new methods, investigators have improved on one of the most significant challenges in forensic analysis: attribution and identification of the source or close relatives of DNA samples from unknown donors. SNP-based kinship analyses using genome-wide relatedness measures or identity-by-descent (IBD) segment approaches are commonly used in FGG analysis, but the impact of genotyping error and missing data on these approaches typically seen in forensic samples has not been fully characterized.

Objective: The goal of this study was to evaluate the accuracy of genome-wide relatedness methods and IBD segment approaches for FGG in the presence of challenges commonly encountered with forensic data: high level of dropout (low call rate) and increased genotyping error.

Technical Approach: We simulated genome-wide SNP genotyping data in large, complex pedigrees where the true underlying relationships were known, simulating genotyping error from 0-10% and missing data from 0-50% from a panel of >500k SNPs on a commonly used microarray in FGG. We benchmarked the performance of the KING estimator as a genome-wide relatedness method, and IBIS and hap-IBD segment approaches. We developed an R package to assist with benchmarking and analysis (below).

Results: When genotyping error is low, the IBD segment methods outperformed genome-wide relatedness methods for close relationships and are more accurate at distant relationship inference. However, with increasing genotyping error (1-5%), methods that do not rely on IBD segment detection are more robust and outperform IBD segment methods. Reduced call rate had little impact on either class of methods.

Conclusions: IBD segment methods are extremely sensitive to genotyping error in forensic samples, resulting in a large drop in accuracy compared with non-IBD segment methods. This can result in missed relationship identification in FGG when using low-quality/degraded samples results in genotyping error.

What's next? We are currently conducting similar FGG methods benchmarking after low-pass whole genome sequencing (LPWGS) and imputation.

Paper (*Frontiers in Genomics* 2022):

doi.org/10.3389/fgene.2022.882268

Software (skater R package):

https://cran.r-project.org/package=skater
https://github.com/signaturescience/skater

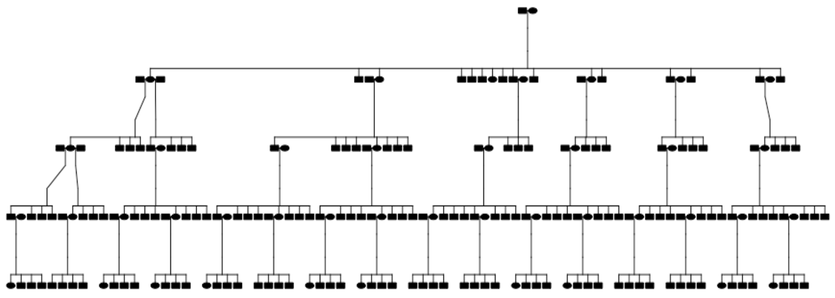

Fig 1: Pedigree structure for each of the simulations conducted. Pedigrees were simulated using ped-sim. Relationships include parent-child (1st degree), full sibling (1st degree), avuncular (2nd degree), grandparent-grandchild (2nd degree), first cousin (3nd degree), great-great-grandparent/child (4th degree), grand-avuncular (4th degree), second cousin (5th degree), third cousin, first cousin once removed, first cousin twice removed, second cousin once removed, etc.

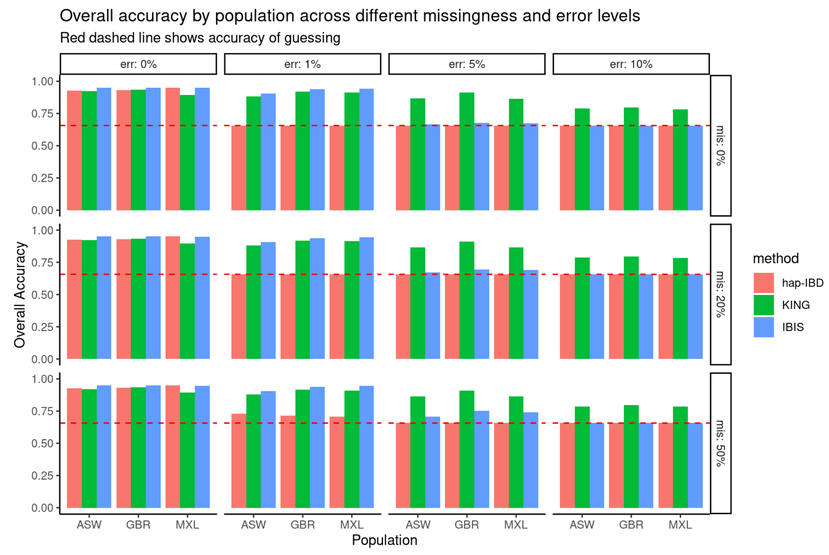

Fig 2: Overall classification accuracy using default parameters. Panels show genotyping error increasing in panels left-to-right and missing data rates increasing panels going top-to-bottom. Individual bars within each panel show the classification accuracy within each simulated population. This graphic shows roughly equivalent accuracy with zero error, but decreased accuracy for both IBD segment methods in comparison to KING with higher genotyping error.

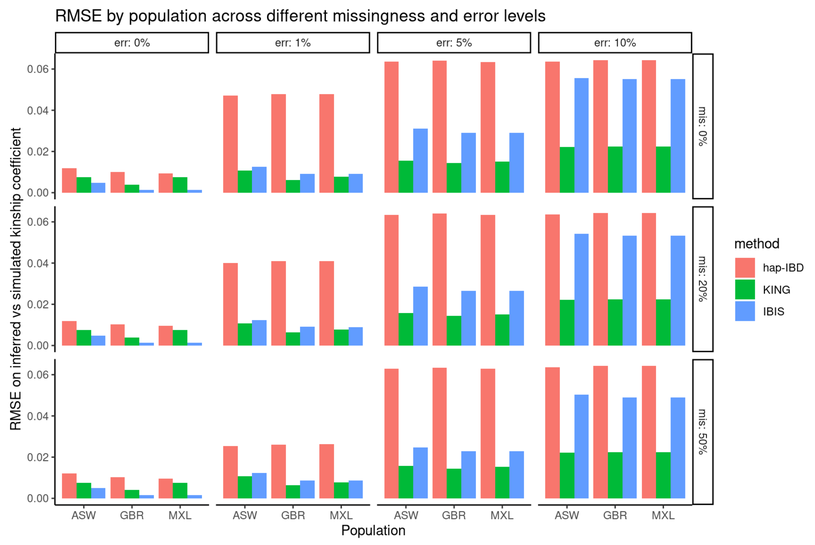

Fig 3: RMSE comparing the inferred versus simulated kinship. Panels show genotyping error increasing in panels left-to-right, and missing data rates increasing in panels going top-to-bottom. Individual bars within each panel show the classification accuracy within each simulated population (ASW, GBR, and MXL).

Fig 4: Detected versus simulated kinship coefficient. X-axis shows the actual simulated kinship coefficient (color-coded by truth relationship degree). Y-axis shows the kinship coefficient inferred using KING (top), IBIS (middle), and Hap-IBD (bottom) using default parameters at different error and missingness levels for simulated relationships from GBR founders. Error increases in panels going left-to-right. Missing data increase in panels top-to-bottom. This shows that for all methods error degrades the detected kinship coefficient, with degradation most notable in high-error simulations (5-10%). Any error completely degrades hap-IBD's ability to detect IBD segments using default parameters, and even small errors severely affect hap-IBD's performance. IBIS is severely impacted by large error, while small amounts of error impact IBIS's ability to accurately assess full siblings, which will share on average about 25% of the genome IBD2. The KING robust estimator suffers the least performance degradation.

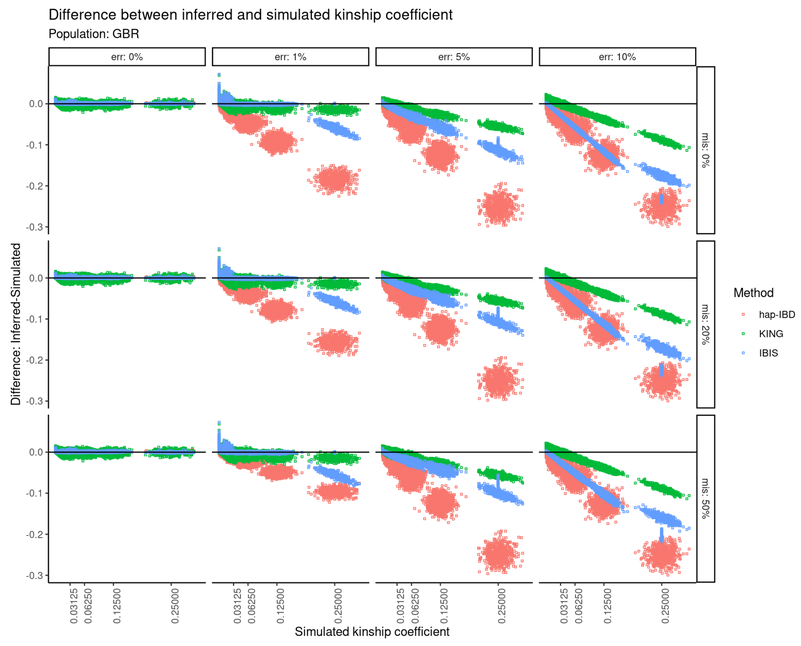

Fig 5: Difference between inferred kinship coefficient versus true simulated kinship coefficient for three different methods using default parameters at different error and missingness levels for simulated relationships from GBR founders. Error increases in panels going left-to-right. Missing data increase in panels top-to-bottom. Each point represents a pair of simulated individuals. Red=hap-IBD; green=KING; blue=IBIS.

Tool	Parameter	Description	Selected	Total runs/tool
IBIS	-mL	Minimum size of IBD kept, cM	2,7	12
	-mt	Min markers for IBD segment	64, 10, 2	
hap-IBD	min-output	Min output segment length	2,7	4
	min-markers	Min markers for IBD segment	100, 64	

Table 1: Selected Parameters for IBD tools. The KING method does not have parameters that can be tuned; however, IBIS and hap-IBD can be run with different settings to detect IBD segments within different constraints. Permissive parameters were not able to rescue IBIS or hap-IBD performance with high error rates.