Synopsis of Snedecor et al., "Fast and accurate kinship estimation using sparse SNPs in relatively large database searches" (1)

Overview

- Development of a windowed kinship algorithm that supports kinship searching in large SNP databases such as GEDmatch PRO, with ~10,000 SNP markers or less
- Windowed kinship searching is robust to loss of loci detection and heterozygosity for kinship estimation making this method appropriate for low-quality and quantity forensic casework-type samples

Background

Forensic investigative genetic genealogy (FIGG) traditionally relied on dense single nucleotide polymorphism (SNP) profiles to search for relatives against online genealogy databases of known profiles generated with microarrays or from whole genome sequencing (WGS). This requires a segment-based or "segment matching" approach to estimate kinship, by searching for contiguous blocks of identical shared alleles and estimating the total centimorgan (cM) distance of those shared segments, which are identical by descent among biological relatives. Segment matching requires hundreds of thousands of SNPs, which is why microarrays or WGS are typically used.

Segment-based approaches, while robust for detecting distant relationships, generally require large quantities of intact DNA, often unavailable in forensic samples. Additionally, SNPs that are physically linked on a chromosome are more likely to be inherited together, making part of the information used in segment matching redundant. Fewer, well-chosen SNPs can be successfully used for sensitive and specific kinship detection with a different kind of algorithm, based on a windowed kinship approach. The ForenSeq™ Kintelligence Kit is a PCR-based 10,230 SNP multiplex, with markers selected for maximum kinship information and without clinically relevant loci or disease markers. Profiles generated using this multiplex can be used to search databases utilizing the windowed kinship algorithm.

Method for kinship SNP selection

Of the 10,230 ForenSeq™ Kintelligence SNPs, 9867 are kinship-informative SNPs selected from popular direct-to-consumer (DTC) microarrays to ensure marker overlap with on-market methods. 153,000 SNPs that overlap with the various DTC microarrays were filtered by frequency for robust representation across global populations, and SNPs with minor allele frequencies (MAF) < 10 % or > 90 % were excluded. The resulting 72,000 SNPs were evaluated using the windowed kinship algorithm. The final 10,230 SNPs selected are maximally spaced across the genome to minimize linkage effects and have no reported significance in ClinVar. For more details on the full selection process, see the Materials and Methods section in Snedecor et al (1).

Windowed Kinship algorithm for sparse SNP marker set

These sparser data, as compared to microarray content, can be generated in one MiSeq FGx run but are less informative with segment-based matching. Both segment matching and windowed kinship rely on the basic principle that distant relatives share contiguous blocks of identical SNPs. Segment matching looks for long stretches of identical SNP alleles, whereas windowed kinship looks for segments as blocks of DNA with a high kinship coefficient. When querying a Kintelligence profile in GEDmatch PRO, the windowed kinship algorithm locates these shared segments using kinship coefficients in "windows" across the genome. The windowed kinship algorithm, a modification of the PC-AiR and PC-Relate tools for genetic relatedness inference ("whole genome kinship" approach) (2), can be used to estimate genetic relatedness in the absence of reference population allele frequencies and to calculate relatedness with kinship coefficients even in situations of admixture, endogamy and consanguinity. Kinship coefficients are calculated in windows across the genome, reducing the possibility of false associations while maintaining the sensitivity of relationship detection, which results in an estimation of shared cMs. Simulated and empirical data in this study demonstrate that the windowed kinship approach with ForenSeq™ Kintelligence profiles with >9000 SNPs performs comparably to segment matching accurately identifying all first, second and third-degree relationships. Additionally, fourth- and fifth-degree relationships are identified with 99% and 55% sensitivity, respectively, with no false kinship associations [Figure 1, Supplemental Table 4, (1)].

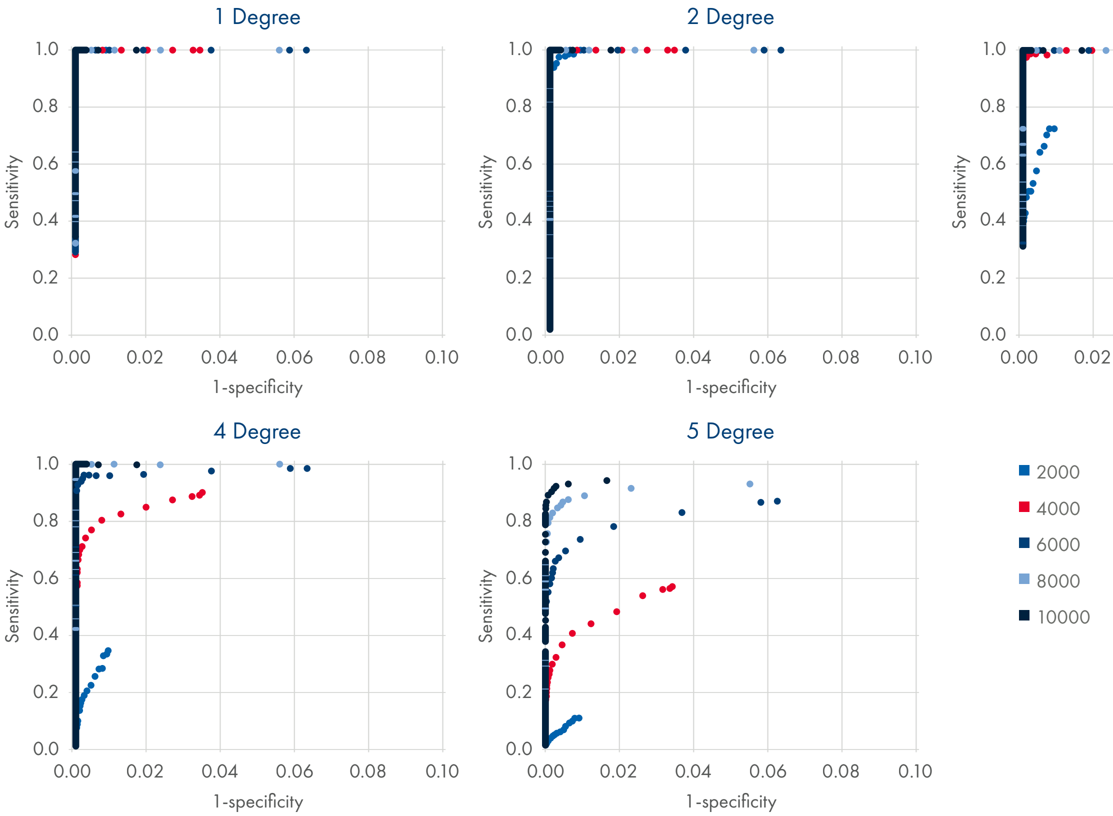

Figure 1. Performance of windowed kinship on varying levels of locus drop out.
Data are plotted for five locus call rates: 2000, 4000, 6000, 8000 and 10000. Overall, performance for first, second and third degrees was observed to be steadily maintained when over 4000 SNPs were typed. For fourth-degree, 8000 SNPs typed gave comparable performance to the full set.

Windowed Kinship algorithm performance with forensic samples

A major reason for choosing a sparser SNP set over array-based approaches is that forensic samples are prone to degradation and locus drop out, making segment matching more difficult. To assess the performance of the windowed kinship algorithm on partial profiles, a GEDmatch test set of 2954 SNP profiles with varying degrees of relationships was used. Random subsets of loci (2000-8000 loci) were selected as missing and evaluated compared to the total ~10,000. Out to fourth degree, SNP locus call rates greater than 6000–8000 generated similar results to those from full 10,230 SNP profiles with a drop in sensitivity of fourth-degree relationships to 66% from 99% for 6000 SNPs

(Figure 1). Fifth-degree relationships were detected with sensitivities of 17–55% for 6000–8000 SNPs (Figure 1). For close relationships (first to third), performance was maintained down to a 4000 SNP locus call rate. Thresholds are recommended to reduce the number of false associations, but higher sensitivity at the higher degree relationships can be obtained with the auxiliary higher false association rate [See Supplemental Table 4 (1)].

Total SNP heterozygosity percentage and heterozygote balance can be used for the assessment of profiles from challenging samples. The performance of the windowed kinship algorithm on samples with lower-than-expected heterozygosity (5-100% loss of sister allele in heterozygous genotypes) was therefore also investigated. Using default kinship thresholds for the windowed algorithm, sensitivity was maintained for first to fourth-degree relationships and approximately half of the fifth-degree relationships were detected when loss of sister allele was less than 10% (40.5–45% heterozygosity), as shown in Table 1. When 20% of sister alleles were not called (36–40% heterozygosity), kinship performance was maintained for first to third degrees with detection of 87% of fourth and 35% of fifth-degree relationships. At a 40% loss (27–30% heterozygosity), performance was maintained for first and second degrees, 87% of third-degree, and 35% of fourth-degree relationships. At greater sister allele loss, only first-degree relationships were captured at a high percentage. A highly degraded, low-input mock casework sample with 7% heterozygosity was able to detect third-degree relatives (1). Crucially, specificity was similar across all levels of heterozygous allele call rates, indicating that the loss of sister alleles did not introduce false associations.

Table 1.
Performance of windowed kinship on varying levels of sister allele drop out.

Detection of relationships by degree was calculated for varying levels of sister allele drop out. Overall, performance for first-, second- and third-degrees was observed to be steadily maintained when heterozygosity is greater than 36%

Percentage loss of sister alleles	Observed Heterozygosity	Sensitivity for Detection of Relationships (Degree)					1-Specificity	False Associations in 1.5M
		1	2	3	4	5		
0%	45-50%	100%	100%	100%	99%	55%	0.00000%	0
5%	42.8-47.5%	100%	100%	100%	98%	51%	0.00000%	0
10%	40.5-45%	100%	100%	100%	97%	49%	0.00004%	1
20%	36-40%	100%	100%	100%	87%	35%	0.00004%	1
40%	27-30%	100%	100%	86%	30%	5%	0.00000%	0
60%	18-20%	87%	30%	5%	3%	0%	0.00000%	0
80%	9-10%	52%	7%	0%	0%	0%	0.00000%	0
100%	0%	48%	5%	0%	0%	0%	0.00000%	0

* Data from a figure in Snedecor et al., reformatted into a table (1)

Conclusion

The windowed kinship algorithm applied to data generated from 10,230 SNPs supports near-perfect detection of relationships extending to the fourth degree in a large database with a high degree of specificity. Samples with reduced locus call rates or loss of sister alleles in heterozygotes can still be used to accurately detect relationships to third-degree, and more distant relationships can be detected with lower thresholds with the understanding that there can also be higher false associations. Using simulated and real SNP profiles, comparable performance was observed for the windowed kinship algorithm and 10,230 SNPs as compared to the segment matching approach and hundreds of thousands of SNPs. The approach described in this article can be considered a powerful tool for investigative lead generation in forensic casework and unidentified human remains investigations, especially with low-input and low-quality samples.

To learn more about ForenSeq Kintelligence, scan this QR code or visit **www.qiagen.com/Kintelligence.**

References:

1. Snedecor J, Fennell T, Stadick S, Homer N, Antunes J, Stephens K, et al. Fast and accurate kinship estimation using sparse SNPs in relatively large database searches. Forensic Science International: Genetics. 2022;61:102769. doi: https://doi.org/10.1016/j.fsigen.2022.102769.
2. Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free estimation of recent genetic relatedness. The American Journal of Human Genetics. 2016;98(1):127-48.