Fast and Accurate Kinship Estimation Using Sparse SNPs in Relatively Large Database Searches

1. Abstract

Forensic genetic genealogy (FGG) has primarily relied upon dense single nucleotide polymorphism (SNP) profiles from forensic samples or unidentified human remains queried against online genealogy database(s) of known profiles generated with SNP microarrays or from whole genome sequencing (WGS). In these queries, SNPs are compared to database samples by locating contiguous stretches of shared SNP alleles that allow for detection of genomic segments that are identical by descent (IBD) among biological relatives (kinship). This segment-based approach, while robust for detecting distant relationships, generally requires DNA quantity and/or quality that are sometimes not available in forensic casework samples. By focusing on SNPs with maximal discriminatory power and using an algorithm designed for a sparser SNP set than those from microarray typing, performance similar to segment matching was reached even in difficult casework samples. This algorithm locates shared segments using kinship coefficients in “windows” across the genome. The windowed kinship algorithm is a modification of the PC-AiR and PC-Relate tools for genetic relatedness inference, referred to here as the “whole genome kinship” approach, that control for the presence of unknown or unspecified population substructure. Simulated and empirical data in this study, using DNA profiles comprised of 10,230 SNPs (10K multiplex) targeted by the ForenSeq^{TM} Kintelligence Kit demonstrate that the windowed kinship approach performs comparably to segment matching for identifying first, second and third degree relationships, reasonably well for fourth degree relationships, and with fewer false kinship associations. Selection criteria for the 10K SNP PCR-based multiplex and functionality of the windowed kinship algorithm are described.

Key Words: forensic genetic genealogy, investigative genetic genealogy, GEDmatch, ForenSeq Kintelligence, extended kinship, windowed kinship algorithm, PCR-based FGG profiles

2. Introduction

Forensic genetic genealogy (FGG), also known as investigative genetic genealogy (IGG), refers to investigative lead generation using dense single nucleotide polymorphism (SNP) profiles from unidentified human remains or crime scene samples that are queried against direct-to-consumer (DTC) genealogical database(s) comprised of known, reference SNP profiles to associate with various degree relatives. FGG has gained interest from the forensic and law enforcement community as a tool to consider when CODIS searching and other means have been exhausted [1]. A large SNP profile generated from microarray analysis is expected to have better discriminatory power than the current battery of forensically relevant short tandem repeat (STR) loci. However, microarray technology requires DNA of quality and input [2] that may not be available from crime scenes or human remains such as skeletal remains.

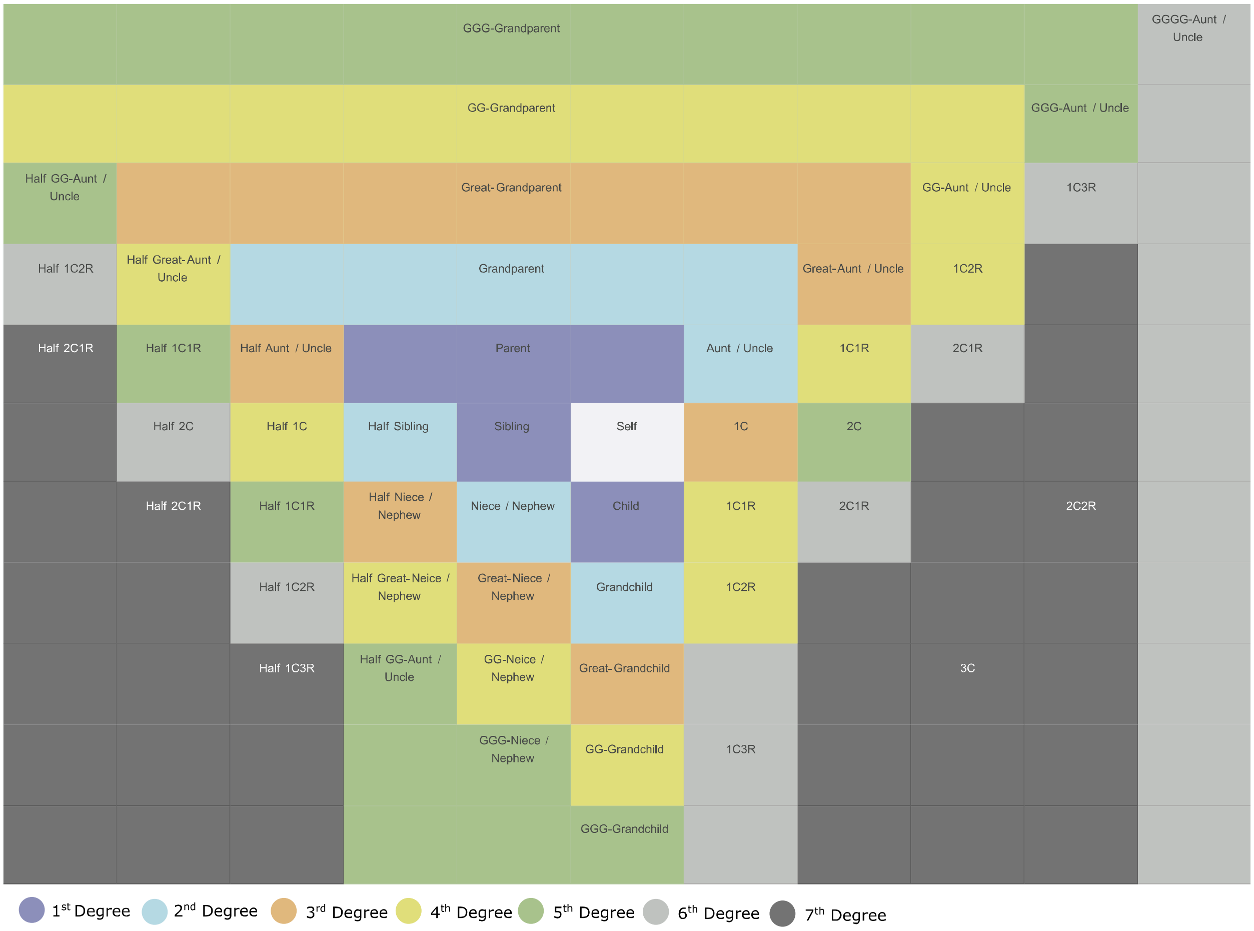


Fig. 1: Examples of degrees of human genetic relationships (adapted from DNA Painter, https://dnapainter.com/).

In addition, for investigative purposes, identifying more distant relationship matches can require a substantially higher effort than for closer relationships. Fig. 1 shows example relationships out to seventh degree. With each increase in degree the number of possible family trees increases significantly. Many genealogy investigations focus on third degree or closer relationships due to burden and inefficiencies that can occur when distance extends to fourth degree or beyond [3]. A polymerase chain reaction (PCR) based FGG typing system that targets sufficient kinship SNPs with high sensitivity of detection of first, second and third degrees relatives, and good detection of fourth and fifth degree relationships, can assist to address the technology gap between microarray and WGS SNP methods regarding sample quality and quantity, personal health information, time and cost.

Generally, FGG has used a “segment matching” approach to estimate kinship by finding contiguous blocks (usually numbering in the hundreds) of identical shared alleles and estimating the total centimorgan (cM) distance covered by those segments [4][5][6]. Segment matching across the genome requires many hundreds of thousands of SNPs thus the use of microarrays or WGS on forensic samples. SNPs that are physically linked on a chromosome are more likely to be inherited together (identical by descent (IBD)), therefore much of the information used in segment matching is redundant and intentionally so. However, as is the case with identity by state (IBS) methods , fewer SNPs can be successfully used for sensitive and specific kinship detection when they provide enough information [7].

A SNP hybridization capture panel used a similar approach using a limited SNP panel [8] and also explored alternative approaches to segment matching in order to evaluate kinship [9].

Forensic genetics has relied upon PCR for decades and can be used to target kinship informative SNPs for FGG. A targeted, forensic PCR assay and analytical software that recovers SNP allele calls from low level, damaged and/or partially degraded forensic DNA samples in a manner sufficient for FGG query was developed. With this strategy, DNA sample analyses may be conducted in operational laboratories using desktop sequencers followed by genealogical database query using a companion kinship inference method. This study describes selection criteria for 10,230 high value SNPs targeted by the ForenSeq Kintelligence™ Kit (Verogen, Inc., San Diego CA), referred to here as the 10K multiplex, and a windowed kinship algorithm to accurately locate and classify kinship out to fourth degree relatives. Of these loci, 9,867 are kinship informative SNPs selected from the Infinium CytoSNP-850K BeadChip and Global Screening Array (Illumina, Inc., San Diego, CA) and filtered using the Genome Aggregation Database (gnomAD) v3.0, the Single Nucleotide Polymorphism database (dbSNP) v151 and GEDmatch for robust representation across global populations. The SNPs are maximally spaced across the genome to minimize linkage effects and have no reported significance in ClinVar [10] (Fig. 2). The remaining 363 SNPs can be used to inform biogeographical ancestry, identity, hair and eye color, or biological sex. Identity SNPs were included in order to allow cross checking of kinship using a previously validated assay (ForenSeq DNA Signature™). The companion windowed kinship algorithm was built upon PC-AiR [11][7][12] and PC-Relate [13] methods, referred to here as the whole genome kinship method, with an additional windowing component. This windowed kinship algorithm also relies on the concept of segment matching (*i.e.*, that distant relatives share contiguous blocks of identical SNPs) and locates segments as blocks of highly scored kinship rather than stretches of identical SNP allele calls to provide even higher performance for FGG.

Simulated pedigrees and real microarray profiles from GEDmatch were used to assess performance of the windowed algorithm. Additionally, two known pedigrees were analyzed using the ForenSeq Kintelligence Kit to assess further kinship estimation performed on real DNA samples using the windowed kinship algorithm. To use GEDmatch microarray profiles as knowns for true relationships, expected degrees of relationship were set using segment matching information since multiple, real extended pedigrees were not available. The 10K SNPs for the 10K FGG multiplex were selected from the GEDmatch test set and the windowed kinship approach was compared to the PC-AiR/PC-Relate whole genome kinship method out to fifth degree relationships.

3. Materials and Methods

3.1. SNP Reference Data for Algorithm Testing

1000 anonymized query samples were selected at random from GEDmatch[1], to generate a test set of SNP profiles with varying degrees of relationship (see Fig. S1 for country of origin for test set and Fig. S2 for GEDmatch country of origin). For each query sample a single sample (if found) was selected for a set of varying total shared cM ranges calculated by the GEDmatch one-to-many tool (2787-3600, 1083-2787, 326-1083 and 0-326 cM) and was added to the target set. A target set of 2,954 samples (including the original set of 1000 query samples) was compiled. Since most donors of GEDmatch samples are

[1] Research purposes, in accordance with Terms of Service

unrelated, the test set was developed to ensure that there was a sufficient set of related samples representing each degree of interest (Table S1).

To evaluate sensitivity and specificity within a particular degree of relationship, the test set was filtered to all sample pairs with the expected shared cM range for that relationship (as shown in Table S1) and all pairs that had zero shared cM. Pairs of samples that had fewer than 9,000 mutually called loci in the 10K SNP multiplex (see Section 4.1 for details) were not considered (see Fig. S3 for overlapping loci counts in the test set). See Table S2 for sample pair totals for different relationship levels.

For simulated pedigrees, genotype data from the 1000 Genomes Project (1KGP) (Phase 3 build 20130502) [14][15] were used as pedigree founders. The original set of 2,504 samples was then filtered to remove relatives using the windowed kinship method and the 10K SNP loci. Sample pairs with > 100 shared total cM and a longest shared segment >30 cM were removed, which reduced the set to 1,851 founder samples. Ped-sim [16] was used to generate 200 pedigrees from these founders using the Poisson model and sex averaged map from Bhérer *et al*. [17]. Relationships were simulated as follows: sibling (first degree), half sibling (second degree), first cousin (third degree), half cousin (fourth degree) and second cousin (fifth degree) (see Fig. S7 for pedigrees). For each relationship degree, there were 200 true relationships and 79,600 unrelated pairings from 400 total samples. To determine how many matching SNPs could be expected for each relationship degree, 1000 independent sample pairs per degree were generated, some with overlapping founder samples. Pairs that shared founders were not compared to each other.

Two known, extended pedigrees were also used to test the 10K SNPs and the windowed kinship algorithm. All samples obtained for testing with Kintelligence were obtained after volunteers signed an informed consent form authorizing the use of de-identified samples for research use publication. One pedigree (n = 26 individuals) included relatives out to the sixth degree (see Fig. S8) uploaded on the public GEDmatch database. Relatives in GEDmatch were marked with their known relationships and anonymized. Since profiles on genealogy databases have been generated by different arrays over time, this evaluation provided a real-world example of performance on DTC data. A “self” reference buccal sample (V024) was typed using the ForenSeq Kintelligence kit, MiSeq FGx sequencer and Universal Analysis Software 2.6, and kinship analysis was performed against the entire GEDmatch database. The second pedigree (n = 15 individuals) was generated from gDNA from buccal swabs for relatives out to the fifth degree (see Fig. S10) typed with the ForenSeq Kintelligence kit.

The V004, V016, V017, V018, V019, V020, V021, and V024 samples consisted of contemporary buccal swabs extracted with the QIAamp DNA Investigator kit (Qiagen, CA), according to the manufacturer’s instructions. DNA quantification was performed using the Quantifluor® ONE dsDNA System (Promega, WI). To degrade V016, 16.8 ng of DNA was placed in each of 4 PCR tubes. All 4 replicates were subjected to continuous cycles of 98 °C for 1 hour, and 4 °C for 10 min for 24 hours, followed by an indefinite 4 °C hold. The DNA replicates were then centrifuged in a tabletop centrifuge for 1 min at maximum speed in order to concentrate any liquid particles to the bottom of the tube. To maximize recovery of DNA, 15 µL of water was used for each replicate, by pipetting the sides of the tubes 10 times, followed by vortexing and centrifugation to allow the DNA samples to be collected in the bottom of the tubes. Samples were quantified using the Quantifluor® ONE dsDNA (Promega, WI). To simulate ante-mortem samples, 1 ng input of each sample was amplified using the ForenSeq™ Kintelligence kit. To simulate post-mortem samples, degraded and/or low input (< 1 ng) samples were amplified using the ForenSeq™ Kintelligence kit, according to the manufacturer's instructions, in following manner: the degraded V016 replicate was amplified with 1 ng input, and the V016 replicate amplified with 250 pg input. All libraries were sequenced using the MiSeq FGx™ reagent kit and the MiSeq FGx™ instrument.

3.2. 10K SNP PCR-Based Multiplex Design

To maximize the value of SNPs in the Kintelligence multiplex, locus selection criteria were considered (see Fig. 2). First, SNPs were selected that are well represented in genetic genealogy databases like GEDmatch. As a quality control measure the frequencies represented in GEDmatch were assessed for general agreement (within three fold of) the population frequencies reported by the Genome Aggregation Database (gnomAD) for European ancestry [18] since this group represents the geographical location of majority of samples in GEDmatch. SNPs were selected that have demonstrated variability within all major human subpopulations (see S2_gnomad.genomes.r3.0.kintelligence_filtered.vcf.zip for population frequencies). With a rare allele at a biallelic SNP, most individuals will be homozygous for the reference allele which is not generally informative for kinship inference in large databases. Common SNP alleles increase the chances for informative differences and similarities between samples. SNPs designated as benign/likely benign in ClinVar were selected [10]. No SNPs with any clinical significance in ClinVar were included. 9,867 kinship informative SNPs that met the selection criteria were included in the ForenSeq Kintelligence multiplex design and are maximally spaced (cM) along each autosome to minimize the effects of physical linkage thereby maximizing the informational value of each individual locus (see Fig. S5 for cM distances in final multiplex).

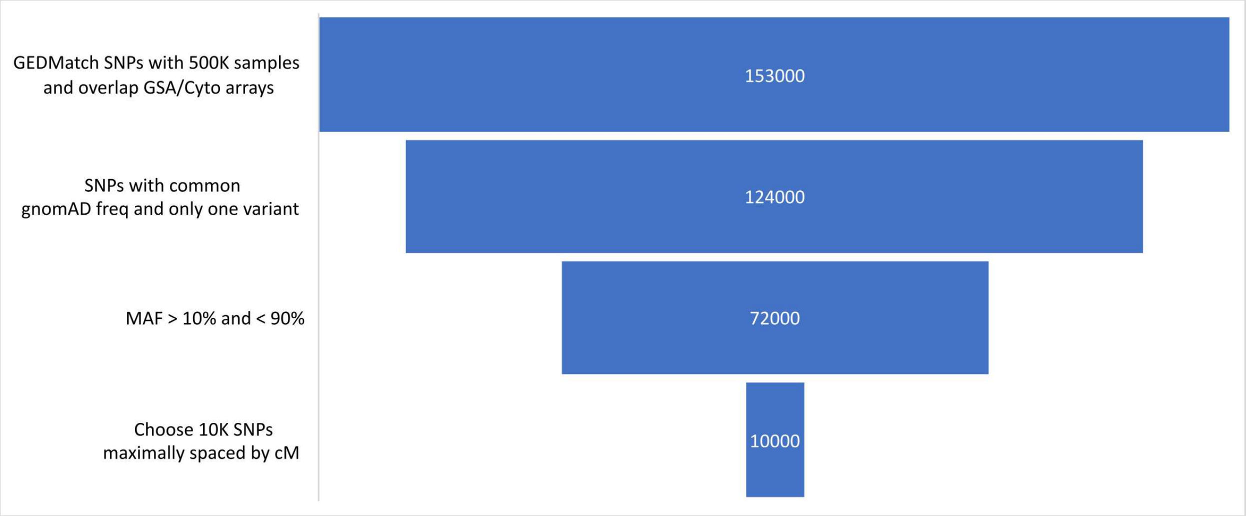


Fig. 2: Method for kinship SNP selection. The overall selection strategy as well as the number of SNPs that remained after each stage of filtering are shown. SNPs that were well represented among DTC microarrays were prioritized then limited to SNPs with gnomAD European allele frequencies that were approximated (within three fold) those observed in the GEDmatch database. SNPs with minor allele frequencies (MAF) < 10% or > 90% were excluded. The resulting 72,000 SNPs were evaluated using the windowed kinship algorithm. 9,867 maximally spaced (cM) kinship informative SNPs were optimized in a PCR-based multiplex.

The kinship informative SNP selection method (Fig. 2) was as follows:

1. Find intersection of SNP content on Infinium Global Screening Array (GSA) and Infinium CytoSNP-850K (Cyto).
2. Filter SNPs in GEDmatch to those in the GSA and Cyto list.
3. Keep SNPs with > 500,000 profiles in GEDmatch (~40% of the GEDmatch database at the time the 10K multiplex was designed in July 2020)
4. Filter to SNPs in gnomAD where GEDmatch frequency is within three-fold of gnomAD EUR allele frequency. This is intended to only capture gross discrepancies between gnomAD and GEDmatch, the frequencies can be quite divergent and still be considered (i.e., 16% gnomAD and 45% GEDmatch would still be included, while 14% to 45% would not.)
    - Enforce gnomAD minor allele frequency (MAF) frequency between 0-50% since some "minor alleles" in gnomAD are actually the major allele. Functionally, this means that if a SNPs reported frequency is > 50%, we use the frequency 100-reported frequency
    - Calculate ratio between GEDmatch and gnomAD if GEDmatch is larger or vice versa
    - If ratio >= 3, discard
5. Remove loci with more than one gnomAD SNP within three-fold of the GEDmatch frequency. This is both because GEDmatch only retains one allele per locus, and because genotypes from arrays may be untrustworthy in triallelic situations. For example, a microarray which is probing for A and C may call an A/G as A/A.
6. Remove SNPs where a population (GEDmatch or nine subpopulations in gnomAD in Table S3) have MAF < 10% or > 90%.
7. Choose N SNPs from the remaining set as follows:
    - Divide N among autosomes relative to their length in cM
    - Compute average spacing for each chromosome in cM
    - Window across the chromosome as follows:
        1. Find next SNP on the chromosome
        2. Pull all SNPs within 70% of the average cM spacing
        3. Pick SNP with the most samples in GEDmatch and the MAF closest to 50%
        4. Discard SNPs within 30% of the average cM spacing downstream of the chosen SNP
        5. Repeat

3.3. Statistical Methods

3.3.1. PC-AiR with Modified Unrelated Set Selection

Model based ancestry estimation methods are less accurate in the presence of genetic relatedness as they cannot distinguish between ancestral groups and clusters of more recent relatives [19]. The PC-AiR [11] method consists of two steps: 1) select a maximally ancestrally diverse set of unrelated samples from a source set; and 2) perform principal component analysis (PCA) on the ancestry representative subset and predict components of variation for all remaining individuals based on genetic similarities. PC-AiR defines a method for identifying a set of unrelated samples that works well for modest sample sets but does not scale well. In a database with n samples, the algorithm must perform n^2 comparisons to remove each related sample. For smaller datasets, this approach is acceptable to maximize ancestral divergence. For relatively large databases a pairwise comparison approach becomes infeasible. Consider, if a database of 1.5M has a thousand related samples then (1.5 million)^2 * 1000 or 2.25 * 10^{15} calculations are required.

Alternatively, relatives can be assessed, and samples discarded to generate an "unrelated" sample set, which can be searched in a much less computationally demanding fashion. Beginning with samples that have the fewest total related samples to minimize data loss, samples can be added iteratively to the unrelated set while relatives are immediately removed from consideration. A more stringent kinship statistic can also be used to find relatives under the assumption that since there is a larger initial dataset, removal of more potential relatives from consideration can be tolerated and helps to ensure that the final set does not contain relatives or if so minimally. Also, samples with a high number (>9000 for the 10K multiplex) of called loci can be considered in the chosen multiplex. The following algorithm was developed as a modification of the PC-AiR method:

1. Remove all samples with >= 5% missing data from the SNP set being used.
2. Compute KING-Robust [11] kinship coefficient between all pairs of samples N. This kinship coefficient for individuals i and j are denoted as φ_{ij} and is defined as the probability that a random allele selected from i and a random allele selected from j at a locus are identical by descent (IBD). Use a relatedness threshold $\tau_{\varphi 1}$= 0.01 to determine whether the pair of samples are expected to be IBD. Use a relatedness threshold of $\tau_{\varphi 2}$= 0.025 for ancestry divergent samples.
    a. Call φ_{ij} with kinship coefficient > $\tau_{\varphi 1}$ as related
    b. Call φ_{ij} with kinship coefficient < -$\tau_{\varphi 2}$ as ancestrally divergent.
3. Initialize two subsets U= ∅ and R= ∅ where ∅ is the empty set.
4. For all $i \in N$
    a. $r_i =$ the set of all j where $\varphi_{ij} > \tau_{\varphi 1}$ for $j \in N$ and $j \neq i$. r_i is the set of all relatives for i.
    b. $d_i =$ the set of all j where $\varphi_{ij} < -\tau_{\varphi 2}$ for $j \in N$ and $j \neq i$. d_i is the set of ancestrally divergent relatives for i
5. Rank all samples $i \in N$ by $|r_i|$ in ascending order.
6. For samples in N with the same $|r|$, sort by $|d|$ in descending order.
7. Iterate through ranked samples and for $i \in N$
    a. If $i \notin R$, $U = U \cup i$ and $R = R \ \cup r_i$.
    b. If $i \in R$ continue to next iteration.

Calculating pairwise kinships is still O(n^2), however the windowed kinship algorithm performs that step only once per model build instead of after every removal of a relative as in the unmodified PC-AiR. Once the unrelated set has been determined, the principal components are determined from the set U using the original PC-AiR method.

3.3.2. PC-Relate and Windowed PC-Relate

Many current methods for kinship inference either assume that pairs of samples came from a homogenous population or require that samples be categorized by sub-population. PC-Relate [13] uses principal components from PC-AiR and partitions genetic correlations into two separate components: a component for the sharing of alleles that are IBD from recent common ancestors and another component for allele sharing due to more distant common ancestry.

Assuming the top PC components from PC-AiR correctly capture the population structure of the samples, those components can be used to estimate the expected allele frequencies based on an individual's ancestral background using a linear regression model rather than using a static population frequency. As described by Conomos *et al.* regarding PC-Relate [13] for a particular SNP s and an individual i, $\hat{\mu}_{is}$ can be calculated which represents the specific expected population SNP frequency for this individual's background as a substitute for $\hat{p}_s$ which is simply the global expected frequency for that SNP determined from a population database.

Once the SNP frequencies have been estimated for each individual it is straightforward to estimate the kinship coefficient ϕ_{ij} for individuals i and j for a set of SNPs S. Let g_{is} be the number of reference alleles an individual has at SNP s.

$$\widehat{\phi_{ij}} = \frac{\Sigma_{s\epsilon S}\ \ (g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\in S}[\,\hat{u}_{is}(1-\hat{u}_{is})\hat{u}_{js}(1-\hat{u}_{is})]^{1/2}}$$

The estimator $\widehat{\phi_{ij}}$ measures the scaled residual genetic covariance between i and j after conditioning on their respective ancestries. Overall, this measurement of kinship can work well. For the FGG use case it has limitations at distant relationships. With the 10K SNP multiplex, the expected number of IBD SNPs for a fifth degree relationship is approximately 300, assuming approximately 0.3 cM between SNPs and 100 total shared cM. Thus, even random fluctuations in overall allele sharing can be above the threshold for detecting a distant relative, which was clear when comparing GEDmatch segment matching against the whole genome kinship coefficient at more distant degrees of relationship.

It is well understood that physically linked genomic regions are more likely to be from inherited DNA which is clustered in contiguous blocks that are reduced in size with each generation. Conversely random allele sharing is in general spread throughout the genome. Segment matching used in GEDmatch and Ancestry.com [20], rely upon this basic concept. A similar approach was taken here by calculating "windows" of kinship across the genome to find shared kinship segments and boost specificity in estimating the more distant relationships.

Given a set of SNPs $S = \{s_0..s_n\}$ and a window size l a window of SNPs is defined at index k as $w_k = \{s_k..s_{k+l}\}$. The windowed kinship approach is as follows:

1. Enumerate all possible windows $W = \{w_0..w_{|S|-l}\}$. Windows must be contained within a single chromosome.
2. Given an individual i and an individual j
3. Calculate kinship across all windows. For $k = \{0..|W|\}$

$$\widehat{\phi_{ijk}} = \frac{\Sigma_{s\epsilon w_k}\ \ (g_{is} - 2\hat{u}_{is})(g_{js} - 2\hat{u}_{js})}{4\sum_{s\epsilon w_k}[\,\hat{u}_{is}(1-\hat{u}_{is})\hat{u}_{js}(1-\hat{u}_{is})]^{1/2}}$$

From here locate IBD segments as follows:

1. Create an empty set P to contain all windows with kinship above threshold t.
2. Given an individual i and an individual j
   a. Iterate through $k = \{0..|W|\}$.
   b. If $\widehat{\phi_{ijk}} \geq t$ add w_k to P
   c. If $\widehat{\phi_{ijk}} < t$ continue
3. Merge windows that have overlapping genomic positions
   a. Iterate through P.

   b. If the current window overlaps with the next window, remove next window from P and reset last index in current window with values from next.
   c. Repeat until there are no overlapping windows remaining.

4. Remove all windows where the fraction of SNPs at which the individuals *i* and *j* share at least one allele is lower than a threshold value *f (e.g. f=0.95).* In windows of true IBD the fraction should be 1 in the absence of genotyping error. However false kinship signal can be generated when many SNPs share no alleles but many others share both alleles.

To find total shared cM, a two-pass approach was taken, first identifying segments with stretches of SNPs with at least one shared allele (half match) and the second, within those segments, stretches of SNPs that have two shared alleles (full match). Half matching segments have $t = 0.22$ while full matching segments have $t = 0.44$. These are reduced from the theoretical values of 0.25 and 0.5 under a strict kinship definition; in the windowed kinship algorithm the thresholds are set slightly lower to allow for genotyping error. When calculating total shared cM, first degree relationships can be distinguished as they mostly consist of half matches and consanguineous or self matches and a higher degree of full matches than more distant relationships.

4. Results and Discussion

4.1. Evaluating Kinship Informative SNP Multiplex Size

Most genetic genealogy databases use a segment matching approach. Segment matching identifies long stretches of matching SNPs, relying on the fact SNPs that are IBD are inherited in contiguous physical blocks. Since large numbers of SNPs are queried, missing or incorrect SNP calls can have minimal effect on segment matching. For FGG, a 10K PCR-based SNP multiplex was designed to provide maximum kinship information with minimal locus content and without clinically relevant loci or disease markers (Fig. 2). These sparser data, as compared to microarray content, can be generated in one MiSeq FGx run but are less informative for kinship if standard segment matching were used. A companion, windowed kinship algorithm was developed that maximizes kinship resolution from the 10K SNP multiplex. This method starts with the same core concept as segment matching, namely identifying contiguous blocks of shared DNA. Then, rather than simply counting matching SNP allele calls, the kinship coefficient described in Conomos *et al*. with PC-Relate [13] is used as a criterion of genetic relatedness. By calculating kinship coefficients in windows across the genome, the discriminatory power of fewer SNPs was enhanced by controlling for background frequencies and population substructure (see Section 3.2).

As shown in Fig. 2, 72,000 SNPs met the locus selection criteria for a PCR-based FGG multiplex. Testing of multiplexes with varied SNP numbers was performed in combination with the windowed kinship algorithm in order to balance the number of SNPs with the ability to detect third degree relatives with high sensitivity. For example, a 20K SNP multiplex and the 10K multiplex were tested and compared using genotype data simulated by ped-sim on 1KGP founder samples for detection of kinship of degrees one through five. Based on the observed fractions of shared alleles from these simulated data (Fig. S4), the 10K and 20K SNP sets enabled significant separation between sample pairs representing third degree. The 10K and 20K SNP sets were then tested using the same simulated data with the windowed kinship approach directly. Out to the third degree, receiver operating characteristic (ROC) curves were nearly identical for the 10K and the 20K SNP sets (and could reach 100% for both sensitivity and specificity). Sensitivity in this instance means the percentage of total related pairs above the scoring thresholds and specificity means the percentage of unrelated samples above the scoring thresholds based on total shared cM and longest shared cM segment. A receiver operating characteristic (ROC) curve with an L shape that aligns closely to the upper left-hand corner indicates that adjusting the thresholds follows a predictable pattern and that there exists at least one threshold with 100% sensitivity and specificity (or close to it). As shown in Fig. 3, the ROC curve for the 10K SNP set achieves 98% sensitivity and 100% specificity for the fourth degree simulated data and performs less well on fifth degree simulated data. Performance with the 20K SNP set was better for fifth degree as expected but even in that case perfect performance was not achieved even in the best-case scenario of a full profile (all loci called). Lowering the threshold of the kinship coefficient will increase the sensitivity of the 10K multiplex comparable to the 20K multiplex with the expected decrease in specificity. Since this multiplex is intended for use with low input, low quality and/or degraded samples, the number of loci is a tradeoff between overall coverage and number of possible SNP calls. Clean, high-input samples already can use existing microarray technologies to provide more SNPs. The 10K SNP multiplex can be considered to provide a practical tool for generating investigative genetic leads extending into the fourth degree (*e.g.*, first cousin once removed (1C1R)). After targeting the 10K SNP set using multiplex PCR, MiSeq FGx v3 sequencing reagents can produce 50M paired end reads, supporting a run configuration comprised of a negative control, a positive DNA control, and one forensic sample with up to 25M reads.

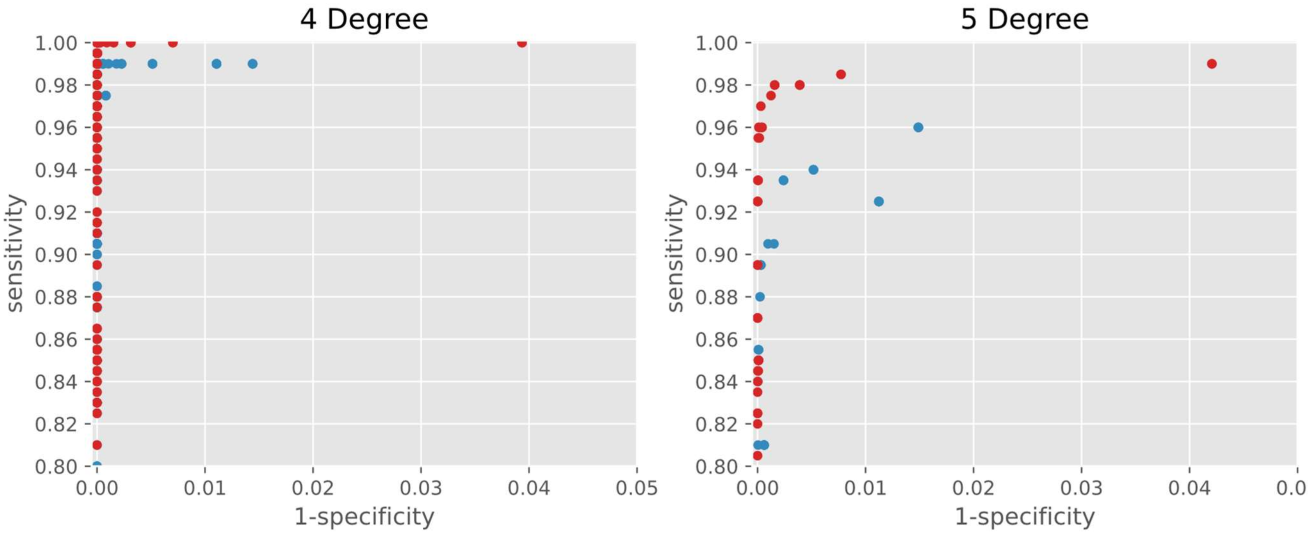


Fig. 3: Comparison between 10K (blue dots) and 20K (red dots) SNP multiplexes of sensitivity of detection and specificity of relationship degree estimation using simulated data from ped-sim. 400 true pairs and 76,000 unrelated pairs generated per degree. For degrees one, two and three, functionally identical sensitivity and specificity were observed (100% for both sensitivity and specificity) for 10K and 20K SNPs. At fourth and fifth degree, an increase in sensitivity was observed with the 20K SNP set. Sensitivity was observed at 92.5% for 20K and 76% for 10K SNPs for fifth degree kinship, and with no false associations.

Since FGG uses relatively large databases (*i.e.*, >1 million samples), evaluating the potential for false associations in the context of a list of potential candidate kinship associations can be helpful to operational settings. The term "false associations" is used here to describe pairs of samples that are above the chosen thresholds that do not actually share a relationship. These have the potential to increase operational time. As the size of a genealogy database increases, the potential also increases for unrelated sample pairs to have larger total shared cM or kinship coefficients than true relatives in a query return. As of June 2, 2022, GEDmatch contained approximately 1.5M autosomal microarray profiles[2] and the FamilyTreeDNA database approximately 1.2M[3]. Thus, even with a specificity of 99.99%, there is potential for hundreds of candidate hits to be returned that are not actual relatives. Microarray-based DNA profiles that comprise a known pedigree extending to sixth degree relationships were used to assess limitations of the windowed kinship approach on a 1.5M sample database. One sample (V024) from the known pedigree was selected as the person of interest ("self") and typed with the 10K multiplex (ForenSeq Kintelligence kit). The full database of 1.5 million profiles was searched using the windowed kinship algorithm described in Section 3.3 and the default thresholds implemented in the GEDmatch Pro™ (see Table S4). (Note: The GEDmatch Pro portal is dedicated to support FGG comparisons for investigative lead generations in criminal casework.) This GEDmatch test query simulated a workflow for unidentified human remains cases. All relationships out to fifth degree (2C) were detected; both sixth degree relationship pairs (2C1R) fell below the default thresholds for total shared cM, longest segment (cM) used by GEDmatch Pro for overlapping SNPs >9,000 (Table 1). Whole genome kinship is included for comparison purposes. Fifth degree kinship was associated to a synthetic profile generated from Native American genomic segments that had been uploaded to GEDmatch (confirmed by the user who initially uploaded the profile). Thus a false positive rate of 1/1,500,000 was achieved.

The false positive profile was later identified to have many contiguous missing sections of the genome which were incorrectly being identified as extensions to segments of kinship. Later revisions to the algorithm (deployed to GEDmatch Pro in June of 2022) address this issue and the false positive is removed.

[2] https://www.gedmatch.com/

[3] https://www.familytreedna.com/why-ftdna

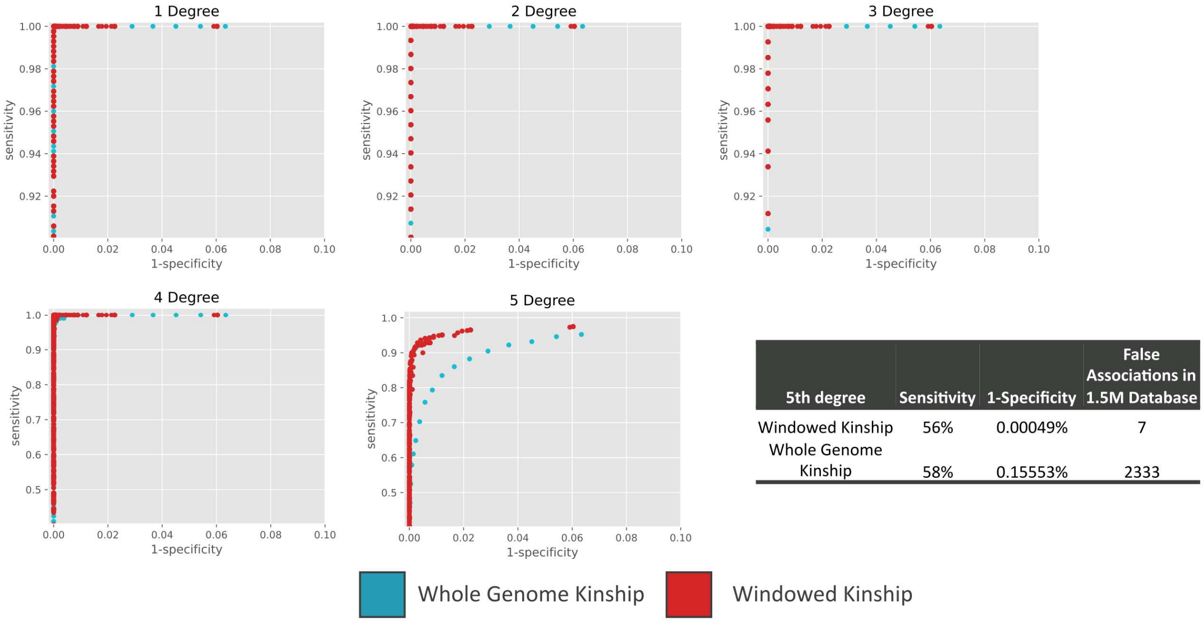
Degree	Id	Relationship	Shared cM	Overlapping SNPs	Longest Segment cM	Whole Genome Kinship
1st degree	6363	sibling	**2562.1**	**9737**	**185.9**	**0.2330**
1st degree	4250	sibling	**2528.2**	**9716**	**200.1**	**0.2411**
2nd degree	6318	niece/nephew	**1498.3**	**9570**	**122.4**	**0.1304**
3rd degree	8818	1st cousin	**791.9**	**9796**	**79.8**	**0.0764**
3rd degree	8319	1st cousin	**609.4**	**9732**	**103.2**	**0.0536**
3rd degree	9555	1st cousin	**502.1**	**9715**	**63.9**	**0.0506**
3rd degree	0603	1st cousin	**904.5**	**9765**	**103.1**	**0.0772**
4th degree	2661	1st cousin 1X removed	**272.9**	**9738**	**79.6**	**0.0304**
5th degree	6100	2nd cousin	**210.1**	**9792**	**83.9**	**0.0188**
6th degree	2491	2nd cousin 1X removed	119.9	9759	66.3	0.0169
6th degree	9608	2nd cousin 1X removed	111.2	9766	48.6	0.0128
Unrelated	8504	Unrelated	**151.9**	**8069**	**88.5**	**-0.0045**

Table 1: GEDmatch query results for a 10K SNP profile from sample V024 and a known pedigree using windowed and whole genome kinship algorithms. Shared cM and Longest Segment cM are calculated from windowed kinship and whole genome kinship from the standard PC-Relate algorithm. Bolded values are higher than GEDmatch Pro default thresholds. The fifth degree relationship that was detected was a false association to a synthetic profile.

4.2. Windowed Kinship vs Whole Genome Kinship

The windowed kinship algorithm is a modification of the PC-AiR and PC-Relate tools for inference of genetic relatedness that use a whole genome kinship approach. Performance of the 10K SNP multiplex with the whole genome kinship approach achieved detection of associations out to the third and was improved for more distant relationships by implementing the windowed kinship approach. For close relatives, the associations detected by whole genome kinship and windowed kinship are the same as there are many overlapping SNPs across the entire genome. For more distant relatives such as second cousins once removed, the number of matching genotypes at the 10K SNP loci for related sample pairs and for unrelated sample pairs is similar. In the samples shown in Fig. S9, for the second cousin match V024 and 9608 there are 3,956 fully matching genotypes. For the unrelated pair there are 3,956 fully matching genotypes. Related and unrelated pairs can therefore produce similar whole genome kinship values. However, in the related pair of samples V024 and 9608 (see Fig. S8 for pedigree), there are distinct segments of kinship which are not the case in the V024 and unrelated sample. Given that true relatives have regions of the same SNP allele calls contiguously on a chromosome rather than randomly distributed throughout the genome, there is a much higher chance of being related if they share SNPs in contiguous blocks (even if two samples have the same number of overlapping SNPs). This concept is the same as that for segment matching, *i.e.*, it is much more likely to find a segment of shared relationship than for SNPs to randomly match through the genome in true relatives.

To compare general performance of windowed kinship versus the whole genome kinship method on SNPs of the 10K set, the GEDmatch test sample set was used that contains profiles of putative relatives based on standard segment matching (see Section 3.1). Only pairs of profiles with >9,000 mutually called loci were used so that aggregate statistics were comparable. ROC curves were generated for whole genome kinship and windowed kinship methods (Fig. 4). Thresholds for windowed kinship were tested between zero and 3300 cM in steps of five for total shared cM, and between zero and 50 cM in steps of two for longest shared cM segment. Thresholds for whole genome kinship were tested between zero and 0.5 in steps of 0.01. A ROC curve that hugs the upper left-hand corner of the graph represents ability to resolve relationship classes and was observed for first through third degrees when either the windowed kinship approach or the whole genome kinship approach was used. These data indicate that 100% sensitivity and specificity can be achieved from either method in this range of relatedness. At fourth and fifth degree relationships, differences in sensitivity and specificity were observed between the two algorithms. At fifth degree in particular, the discriminatory power of the windowed kinship approach was higher than with the whole genome kinship method. From a practical perspective, given a database of 1.5 million samples and cM thresholds that support approximately 50% sensitivity for fifth degree relatives, seven false associations would be expected using windowed kinship as compared to more than 2,000 false associations using the whole genome kinship method alone.



5th degree	Sensitivity	1-Specificity	False Associations in 1.5M Database
Windowed Kinship	56%	0.00049%	7
Whole Genome Kinship	58%	0.15553%	2333

Fig. 4: ROC curves for whole genome kinship (blue squares) vs windowed kinship (red squares) methods on a test sample set in GEDmatch comprised of the 10K SNP set. For fourth and fifth degree relationships, windowed kinship significantly improved sensitivity and specificity. When thresholds for shared total cM (for the windowed kinship method) and kinship coefficient (for the whole genome method) are set to give approximately the same sensitivity for fifth degree relationships, more false associations are detected with the whole genome kinship method (>2300); see table inset.

4.3. Estimated Shared cM from Windowed Kinship vs from GEDmatch Segment Matching

Segment matching algorithms used by DTC genetic genealogy companies output an aggregate metric of total shared cM. Since this metric is widely used, there are several tertiary tools that can be used to interrogate genetic kinship associations by looking at shared cM values. For example, the Shared cM Project provides an aggregate of shared cM values across many degrees of relationships, facilitating determination of what types of relationships correspond to which ranges of shared cM values.[4] Even though the mechanism of windowed kinship is not the same as segment matching, the windowed method can provide matching segments across the genome, and output total shared cM as a kinship metric.

Estimates of shared cM from segment matching and windowed kinship were compared. One difficulty with such an evaluation is that different genetic genealogy companies use different cM maps which can lead to divergent measurements (see Fig. S11). The windowed kinship implementation in GEDmatch Pro uses newer maps from Bherer *et al.* [17] that have a total sex-averaged cM across the autosomes of 3,342 cM while GEDmatch segment matching uses an older cM map that has a total of 3,586 cM. Thus, *on average* the estimates from windowed kinship are expected to be approximately 7% lower than those with the GEDmatch segment approach. However, since the differences are unevenly distributed, they can be higher or lower depending on the shared segments between two samples.

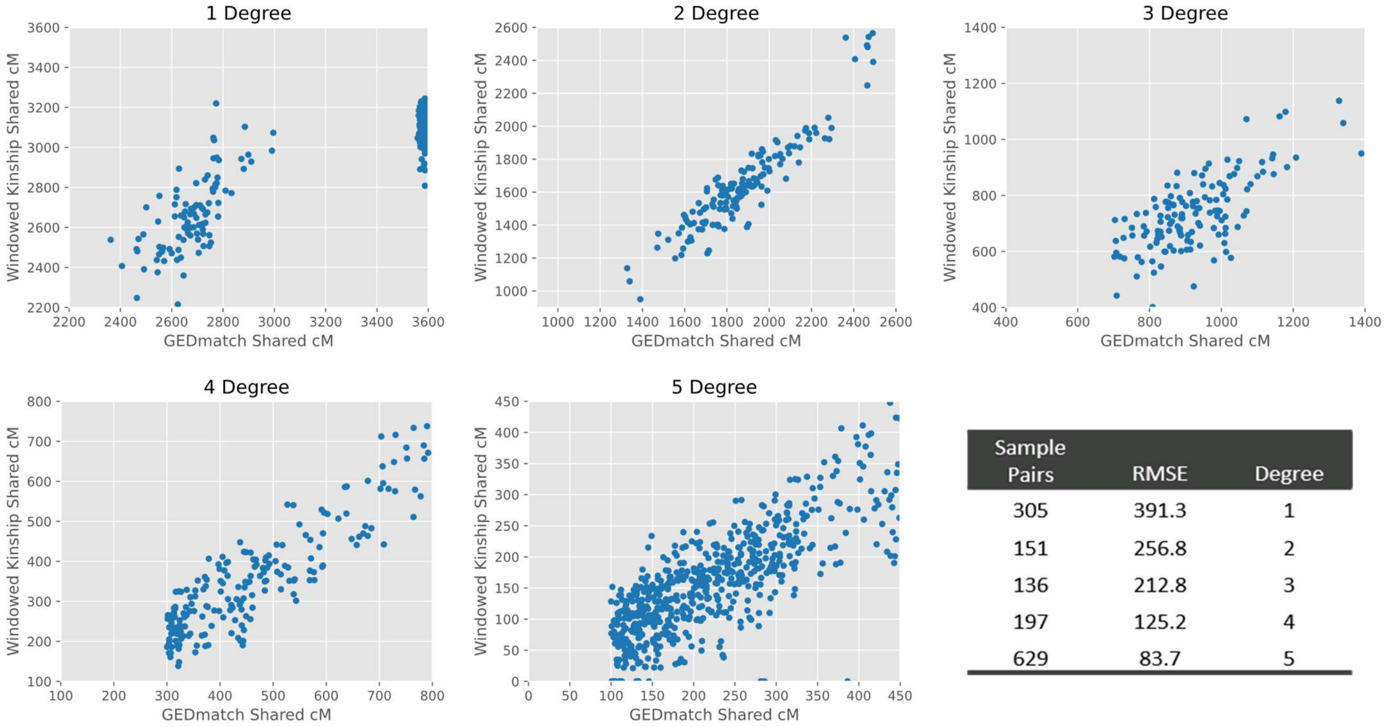
One other issue with comparing the shared cM metric is that GEDmatch only considers half-matches in its one-to-many tool[5]. When there is at least one allele in common at a single biallelic locus, then half-matching considers that as a match between samples. For example, if there is a locus with a heterozygous call in one sample and a homozygous call in another sample, then that locus is considered a half-match since either allele from the heterozygote can match to the homozygote. For a full match, each sample must be heterozygotic or must be homozygous for the same allele to be considered matching. As relationships get more distant, it is more likely that segments of shared kinship will be comprised of more half-matches than full matches, which is sufficient as a first pass when conducting database searching. However, a "self-match", *i.e.*, two samples from the same individual, has a maximum cM value of 3,586, same as a first degree relative. Since windowed kinship considers full matching, a self-match is represented by a number closer to 6,642 cM. Therefore, samples with values from windowed kinship greater than 3,342 cM will not be the same as what is reported from segment matching in GEDmatch. To control for this effect in this study, all GEDmatch test pairs with an estimated shared cM of > 3,600 by the windowed kinship method were not considered. (thus, numbers of first degree sample pairs differ between Fig. 5 and Table S2)

With these caveats in mind, concordance and differences between cM estimates from these two methods were compared. As shown in Fig. 5, the estimates of total shared cM between the windowed kinship approach and GEDmatch segment matching were similar, although variability between them was observed. Interestingly, there are many GEDmatch first degree hits with values close to the maximum possible shared cM values that have a wider spread for windowed kinship. The windowed kinship estimates fall within the first degree shared cM ranges. These data indicate that differences in the total shared cM values may be observed for close relatives compared to values generated with the GEDmatch

[4] https://thegeneticgenealogist.com/

[5] https://classic.gedmatch.com/Documents/Qdocs.pdf

segment approach. This difference may be due to segment matching less aggressively filtering segments than the windowed kinship approach thus is not an impediment to conducting FGG.



Sample Pairs	RMSE	Degree
305	391.3	1
151	256.8	2
136	212.8	3
197	125.2	4
629	83.7	5

Fig. 5: Comparison of estimated shared cM from the GEDmatch segment approach (x-axis) and the windowed kinship approach (y-axis) for 1,420 sample pairs. The inset table shows the root mean squared error (RMSE) of the windowed shared cM vs the GEDmatch shared cM estimates for the first through fifth degrees. These two approaches use different reference maps to estimate cM distances such that the overall estimated total cM in the GEDmatch segment approach differs from windowed kinship on average by ~7%. In general, observed estimations from the windowed kinship approach were slightly lower than from segment matching. 208 first degree relationships with estimated shared cM close to the maximum possible value from GEDmatch (3,600 cM) showed lower values from windowed kinship, though still within range of a first degree relationship.

4.4. Windowed Kinship Performance with Forensic Case-Type Samples

4.4.1. Windowed Kinship Performance with Partial Profiles (<10K Kinship Informative SNPs)

Some loci in targeted assays of forensic samples or unidentified human remains may not be detected (*e.g.*, data below an analytical threshold or no data detected) such that partial profiles are generated due to DNA degradation, damage and/or PCR inhibition. To assess performance of the kinship algorithms for samples with different levels of missing loci, the GEDmatch truth set described in Section 3.1 was used. The GEDmatch test set used standard segment matching on microarray data to locate relatives and then the test SNPs were filtered to the 10K SNP set and used for kinship inference. In this evaluation, from the set of 10K SNPs, random subsets of loci were selected and marked as missing from the input profiles of the GEDmatch test set. Between 2000 and 8000 loci were removed in this fashion and evaluated, equivalent to 80-20% SNP locus call rates.

ROC curves were used to evaluate these data as different levels of missing loci can be recognized when kinship is estimated. For example, if the specificity in estimating relationship degree is reduced when a certain number of the 10K loci are untyped, then kinship thresholds can be adjusted dynamically to account for it. As shown in Fig. 6, at lower levels of missing loci and out to fourth degree relationships the ROC curves were sharply upper and leftward, indicating high sensitivity and specificity. Out to fourth degree, SNP locus call rates greater than 60-80% generated similar results to those from full 10K SNP profiles. For close relationships (first to third), performance was maintained down to a 40% SNP locus call rate. Although similar performance can be achieved with large numbers of missing loci, the kinship thresholds necessary to achieve that performance can differ. Thus, it is important to use different thresholds based on how many SNPs are shared between samples, *e.g.*, if a pair of samples has 6,000 SNPs typed in common, a higher windowed kinship threshold can be used than for a pair of samples with 9,000 overlapping SNPs (see Table S4 for thresholds set for windowed kinship in the GEDmatch Pro implementation).

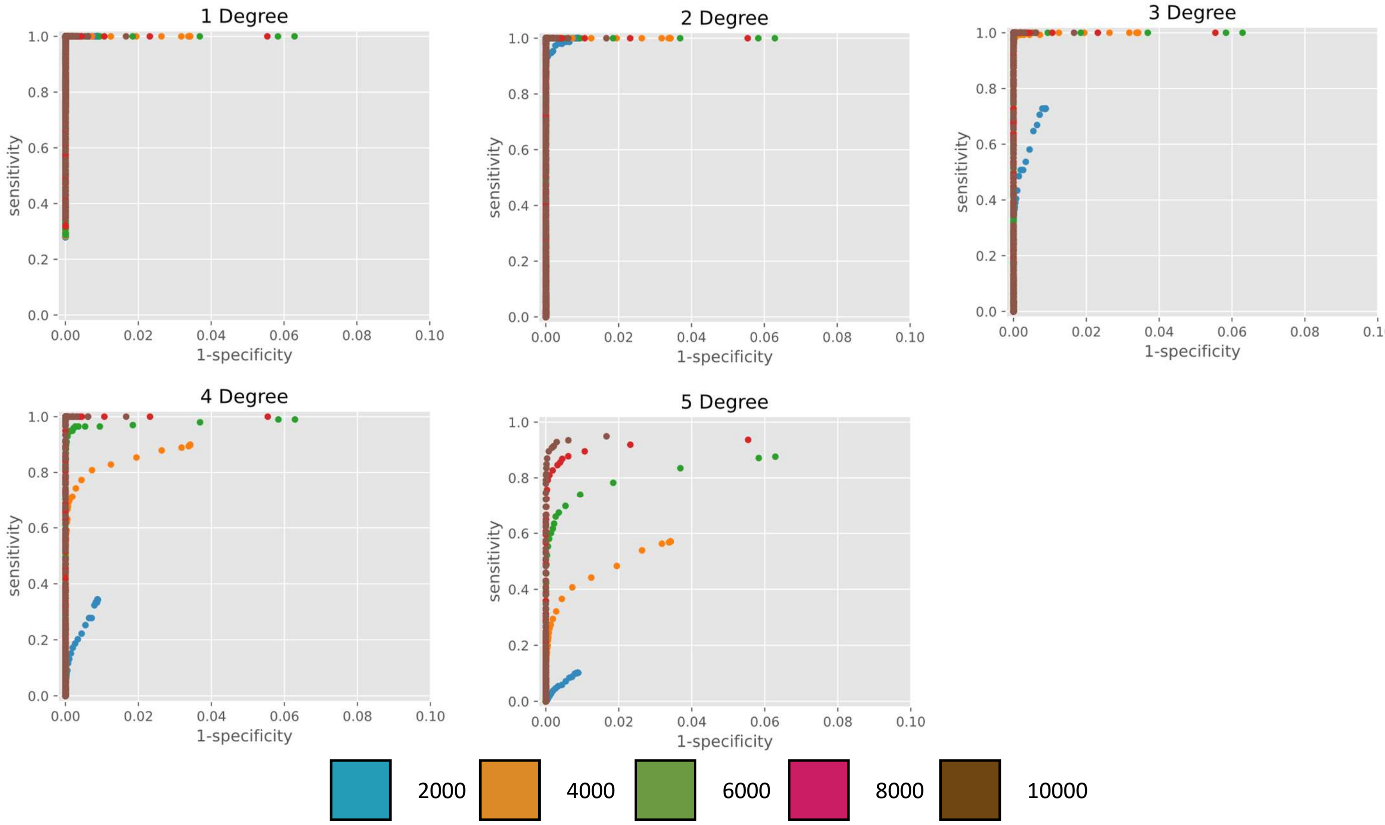


Fig. 6: Performance of windowed kinship in GEDmatch test set with call rates between 20-100% for the 10K SNP set. Data are plotted for five locus call rates as follows: 20% (2K), 40% (4K), 60% (6K), 80% (8K) and 100% (10K). Overall, performance for first, second and third degrees was observed to be steadily maintained when 80%, 60% or 40% of the 10K SNPs were typed. For fourth degree, 80% of the 10K SNPs were observed to give comparable performance to the full 10K set, and 60% was sufficient to make some kinship analyses (approximately 60% sensitivity vs 100% sensitivity for the full profile at the same specificity).

4.4.1.1. Windowed Kinship Performance using Partial SNP Allele Call Rates

Total heterozygosity at SNP loci in a human DNA sample and quantitative balance between heterozygous alleles can be used as quality metrics for SNP genotyping and particularly for assessment of profiles from challenging samples, including those in forensic casework [21]. These metrics can indicate the likelihood that only one of the sister alleles in a true heterozygote were detected and may be called as homozygous. As sample quality degrades and input DNA template is reduced, certainty in homozygous SNP calling can be affected. Forensic genetics casework employs methods and tools to assist in this regard, such as use of stochastic thresholds [22][23]. For the windowed kinship algorithm, whether similar threshold(s) are necessary to disqualify SNP data outright from proceeding with FGG or whether the algorithm was robust to some missing alleles was investigated.

To evaluate how loss of sister alleles affects windowed kinship performance, the GEDmatch test set was used. As with the previous evaluations, the truth set was generated from segment matching on whole microarray profiles, and the samples were filtered to the 10K SNP set. Different percentages of heterozygous loci were changed to homozygous reference (ref) or alternative (alt) calls. Ref calls generally refer to the more prevalent allele in a reference population while alt calls refer to the less prevalent, or "minor" allele. For the SNP locus rs6690515 as an example, a G is considered "ref" while A is considered "alt". Converting a G/A call to a G/G call, changes a heterozygote to a homozygous ref call. The ref allele is represented as 0 and the alt allele as 1 when the actual nucleotide is not germane.

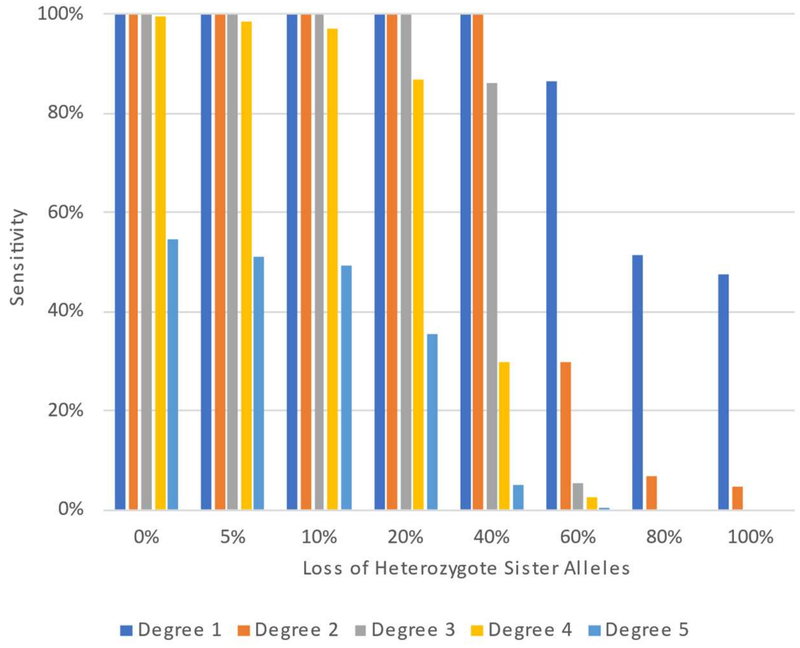
An example of the simulation strategy used in this study is as follows: Consider a simulation of 5% of sister alleles at heterozygotes among the 10K SNP set. The transition probabilities of the genotypes from the original profile are shown in Table 2. The transition table provides the percentage of a heterozygous locus modified to a homozygous call in the test case simulations of allele non-detection. For example, if an input sample has a locus with a starting genotype of 0/0, the test profile will also have a genotype of 0/0 since the probability that 0/0 transitions to 0/0 is 100%. However, if the starting genotype is a 0/1 genotype, the chance was 95% to remain 0/1 and 2.5% chance to become 1/1 or 0/0, indicating non detection of a sister allele. Essentially, this emulates cases where the second allele in a heterozygote is below the analytical threshold and therefore, calling a heterozygous call as a homozygote erroneously.

Table 2: Transition probabilities for 5% lack of detection of sister alleles at heterozygous SNPs as used in simulation studies. ref allele (0), alt allele (1).

Original Genotype	Test Genotype	Probability
0/0	0/0	100.0%
0/1	0/0	2.5%
0/1	0/1	95.0%
0/1	1/1	2.5%
1/1	1/1	100.0%
./.	./.	100.0%

Ranges of missing sister allele calls between 5 and 100% were tested. Whereas with missing loci it is trivial to determine how many are missing, it is more difficult to quantify sister allele loss in an unknown sample since it can depend on factors inherent to the sample and to the subpopulation of origin. It is likely then more illustrative to analyze performance using the default windowed kinship thresholds than all possible thresholds (see Fig. S6 for full ROC). Using the default kinship thresholds for the windowed algorithm as implemented in GEDmatch Pro (Table S3), sensitivity was observed to be maintained for first to fourth degree relationships when loss of sister allele detection was less than 10%. When 20% of sister alleles were not called, kinship performance was maintained within the first to third degrees. At a

40% loss performance was maintained within the first and second degrees and at greater sister allele loss only first degree were captured (Fig. 7). Crucially, specificity was similar across all levels of heterozygous allele call rates, indicating that the loss of sister alleles did not introduce false associations.



Loss of Sister Alleles	1-Specificity	False Associations in 1.5M
0%	0.00000%	0
5%	0.00000%	0
10%	0.00004%	1
20%	0.00004%	1
40%	0.00000%	0
60%	0.00000%	0
80%	0.00000%	0
100%	0.00000%	0

Fig. 7: Performance of the windowed kinship method on the GEDmatch test set after simulating loss of sister alleles (between 5 and 100%) at heterozygous sites of the 10K SNP set. Kinship thresholds are based on default settings of the windowed kinship algorithm implementation in GEDmatch Pro for profiles with > 9,000 SNPs typed in common (overlapping) (i.e., 140 shared cM total, 30 cM longest segment). Simulated losses were evenly distributed between ref and alt alleles, i.e., a heterozygote in the fully typed profile had a 2.5% chance to become homozygous alt and a 2.5% chance to become homozygous ref. 1-specifitiy indicates the chance of incorrectly classifying an unrelated association as related, sensitivity indicates the percentage of total true associations found above threshold.

In addition to simulated data, a known pedigree (see Fig. S11) containing relatives extending to the fifth degree was used. The person of interest (“self”) sample (V016) was heat treated to emulate partial DNA degradation and windowed kinship metrics generated from two PCR template inputs (1 ng and 250 pg for Kintelligence library preparation for sequencing) and compared. In order to test the limits of the system, one sample was run at higher plexity (12 samples in a run) than recommended by the manufacturer and also used 250 pg input. For these empirical samples, the expected associations out to third degree passed GEDmatch Pro thresholds for the 1 ng sample and second degree for the 250 pg sample (7% heterozygosity).

As represented by total shared cM values (Fig. 8) as sister allele non-detection increases, the overall estimated shared cM value dropped. For example, in a comparison of sample V004 with sample V016 with 1 ng input they fall within the expected shared cM range for a first degree hit with 3076.6 (see Table S1). The same sample compared to a V016 sample with 250 pg input only showed a shared cM value of 1561.097, which is significantly lower than would be expected for a first degree candidate hit.

				V016							
			Input	1 ng				250 pg			
			% Hetero	45%				7%			
			False Associations in 1.5M database	0				0			
Degree	Relationship	HQ Sample	% Hetero	nsnp	Shared cM	Longest cM Segment	Kinship Coefficient	nsnp	Shared cM	Longest cM Segment	Kinship Coefficient
0	Self	V016	44%	9366	**6595.395**	**260.312**	0.507	5736	**3324.178**	**259.144**	0.551
1	MO	V004	44%	9226	**3088.442**	**216.827**	0.247	5729	**1526.438**	**215.512**	0.271
2	AU	V017	45%	9253	**1663.061**	**151.056**	0.137	5729	**747.132**	**162.051**	0.151
3	C	V018	45%	9166	**739.054**	**63.518**	0.060	5726	185.690	61.256	0.067
4	1C1R	V019	44%	9248	124.744	35.951	0.013	5726	0.000	0.000	0.010
5	2C	V020	44%	9244	64.965	33.837	0.000	5735	0.000	0.000	-0.007
N/A	N/A	NA24385	44%	9179	0.000	0.000	-0.006	5732	0.000	0.000	-0.017

Fig. 8: Example showing shared cM for sample V016 as a mock casework sample within a known pedigree. All samples were typed for the 10K SNP set using the ForenSeq Kintelligence Kit. V016 (self) was heat treated to emulate DNA degradation. In order to test the limits of the system, one sample was run at higher plexity (12 samples in a run) than recommended by the manufacturer and also used 250 pg input. Bolded cells indicate which values are above the thresholds used in GEDmatch Pro. Samples were also searched against full 1.5M GEDmatch database and no false associations were found above thresholds. These are currently 140 total shared cM and 37 longest cM segment for samples with more than 9000 overlapping SNPs and 180 total shared cM and 37 longest cM segment for samples with 6000 overlapping SNPs. "nsnp" indicates the total number of SNPs shared between the two samples in the pair.)

5. Conclusions

The windowed kinship algorithm applied to data generated from the 10K SNP multiplex supports near perfect detection of relationships extending to the third degree in a large database with a high degree of specificity even in samples with reduced locus call rates or lack of detection of sister alleles in heterozygotes. Using simulated and real GEDmatch SNP profiles, comparable performance was observed for the windowed kinship algorithm and the 10K SNP set as compared to the segment matching approach that uses hundreds of thousands of SNPs. In real degraded samples the ForenSeq Kintelligence system can identify relationships robustly out to the 3rd degree. For forensic samples, the approach described herein can be considered as a powerful tool for investigative lead generation in forensic casework and unidentified human remains investigations that can be readily transferred and implemented into operational settings under an insourced or outsourced FGG SNP typing model.

Acknowledgements

Special thanks to Bruce Budowle for editing and review assistance and to John Hayward and Wilbon Davis for reviewing sections relating to GEDmatch.

References

[1] United States Department of Justice, “Interim Policy Forensic Genetic Genealogical DNA Aanalysis and Searching,” pp. 1–8, 2019, [Online]. Available: https://www.justice.gov/olp/page/file/1204386/download.

[2] J. H. de Vries *et al.*, “Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy,” *Forensic Sci. Int. Genet.*, vol. 56, no. November 2021, 2022, doi: 10.1016/j.fsigen.2021.102625.

[3] J. Ge and B. Budowle, “Forensic investigation approaches of searching relatives in DNA databases,” *J. Forensic Sci.*, vol. 66, no. 2, pp. 430–443, 2021, doi: 10.1111/1556-4029.14615.

[4] C. A. Ball *et al.*, “Discovering genetic matches across a massive, expanding genetic database,” *AncestryDNA Matching White Pap.*, pp. 1–34, 2016, [Online]. Available: https://www.ancestry.com/dna/resource/whitePaper/AncestryDNA-Matching-White-Paper.

[5] C. Morimoto *et al.*, “Pairwise kinship analysis by the index of chromosome sharing using high-density single nucleotide polymorphisms,” *PLoS One*, vol. 11, no. 7, pp. 1–17, 2016, doi: 10.1371/journal.pone.0160287.

[6] C. Morimoto, S. Manabe, S. Fujimoto, Y. Hamano, and K. Tamaki, “Discrimination of relationships with the same degree of kinship using chromosomal sharing patterns estimated from high-density SNPs,” *Forensic Sci. Int. Genet.*, vol. 33, no. November 2017, pp. 10–16, 2018, doi: 10.1016/j.fsigen.2017.11.010.

[7] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. M. Chen, “Robust relationship inference in genome-wide association studies,” *Bioinformatics*, vol. 26, no. 22, pp. 2867–2873, 2010, doi: 10.1093/bioinformatics/btq559.

[8] A. Tillmar, K. Sturk-Andreaggi, J. Daniels-Higginbotham, J. T. Thomas, and C. Marshall, “The FORCE panel: An all-in-one SNP marker set for confirming investigative genetic genealogy leads and for general forensic applications,” *Genes (Basel).*, vol. 12, no. 12, 2021, doi: 10.3390/genes12121968.

[9] D. Kling and A. Tillmar, “Forensic genealogy—A comparison of methods to infer distant relationships based on dense SNP data,” *Forensic Sci. Int. Genet.*, vol. 42, no. June, pp. 113–124, 2019, doi: 10.1016/j.fsigen.2019.06.019.

[10] M. Landrum *et al.*, “ClinVar,” no. Md, 2013.

[11] M. P. Conomos, M. B. Miller, and T. A. Thornton, “Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness,” *Genet. Epidemiol.*, vol. 39, no. 4, pp. 276–293, 2015, doi: 10.1002/gepi.21896.

[12] S. M. Gogarten *et al.*, “GWASTools: An R/Bioconductor package for quality control and analysis of genome-wide association studies,” *Bioinformatics*, vol. 28, no. 24, pp. 3329–3331, 2012, doi: 10.1093/bioinformatics/bts610.

[13] M. P. Conomos, A. P. Reiner, B. S. Weir, and T. A. Thornton, “Model-free Estimation of Recent Genetic Relatedness,” *Am. J. Hum. Genet.*, vol. 98, no. 1, pp. 127–148, 2016, doi: 10.1016/j.ajhg.2015.11.022.

[14] P. H. Sudmant *et al.*, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, no. 7571, pp. 75–81, 2015, doi: 10.1038/nature15394.

[15] A. Auton *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015, doi: 10.1038/nature15393.

[16] M. Caballero *et al.*, "Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives," *PLoS Genet.*, vol. 15, no. 12, pp. 1–29, 2019, doi: 10.1371/journal.pgen.1007979.

[17] C. Bherer, C. L. Campbell, and A. Auton, "Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales," *Nat. Commun.*, vol. 8, 2017, doi: 10.1038/ncomms14994.

[18] S. Gudmundsson *et al.*, "Variant interpretation using population databases: Lessons from gnomAD," *Hum. Mutat.*, no. November 2021, 2021, doi: 10.1002/humu.24309.

[19] T. A. Thornton and J. L. Bermejo, "Local and global ancestry inference and applications to genetic association analysis for admixed Populations," *Genet. Epidemiol.*, vol. 38, no. SUPPL.1, 2014, doi: 10.1002/gepi.21819.

[20] C. A. Ball *et al.*, "AncestryDNA matching white paper: discovering genetic matches across a massive, expanding genetic database," *AncestryDNA*, no. March, pp. 1–46, 2016.

[21] S. Zhang *et al.*, "Parallel Analysis of 124 Universal SNPs for Human Identification by Targeted Semiconductor Sequencing," *Sci. Rep.*, vol. 5, no. September, pp. 1–9, 2015, doi: 10.1038/srep18683.

[22] T. R. Moretti, A. L. Baumstark, D. A. Defenbaugh, K. M. Keys, J. B. Smerick, and B. Budowle, "Validation of Short Tandem Repeats (STRs) for Forensic Usage: Performance Testing of Fluorescent Multiplex STR Systems and Analysis of Authentic and Simulated Forensic Samples," *J. Forensic Sci.*, vol. 46, no. 3, p. 15018J, 2001, doi: 10.1520/jfs15018j.

[23] P. Gill, R. Puch-Solis, and J. Curran, "The low-template-DNA (stochastic) threshold-Its determination relative to risk analysis for national DNA databases," *Forensic Sci. Int. Genet.*, vol. 3, no. 2, pp. 104–111, 2009, doi: 10.1016/j.fsigen.2008.11.009.

Supplementary Material

GEDmatch Test Set Characteristics

Table S1: Expected shared cM ranges per degree of relationship in GEDmatch. cM ranges shown are based on DNAPainter[6]. If a pair of samples falls into more than one range (i.e., 400 shared cM overlaps with the ranges for fourth and fifth degrees) evaluation of both relationship degree possibilities may be advantageous. First cousin (1C), first cousin once removed (1C1R), second cousin (2C), great great grandchild (GG-Grandchild), great great great grandchild (GGG-Grandchild).

Degree of relationship	GEDmatch Shared cM range	Relationship examples
1	2300-3600	Sibling, Parent, Child
2	1300-2500	Half Sibling, Niece
3	700-1400	1C, Great Grandchild
4	300-800	1C1R, GG-Grandchild
5	100-450	2C, GGG-Grandchild

[6] https://dnapainter.com/tools/sharedcmv4

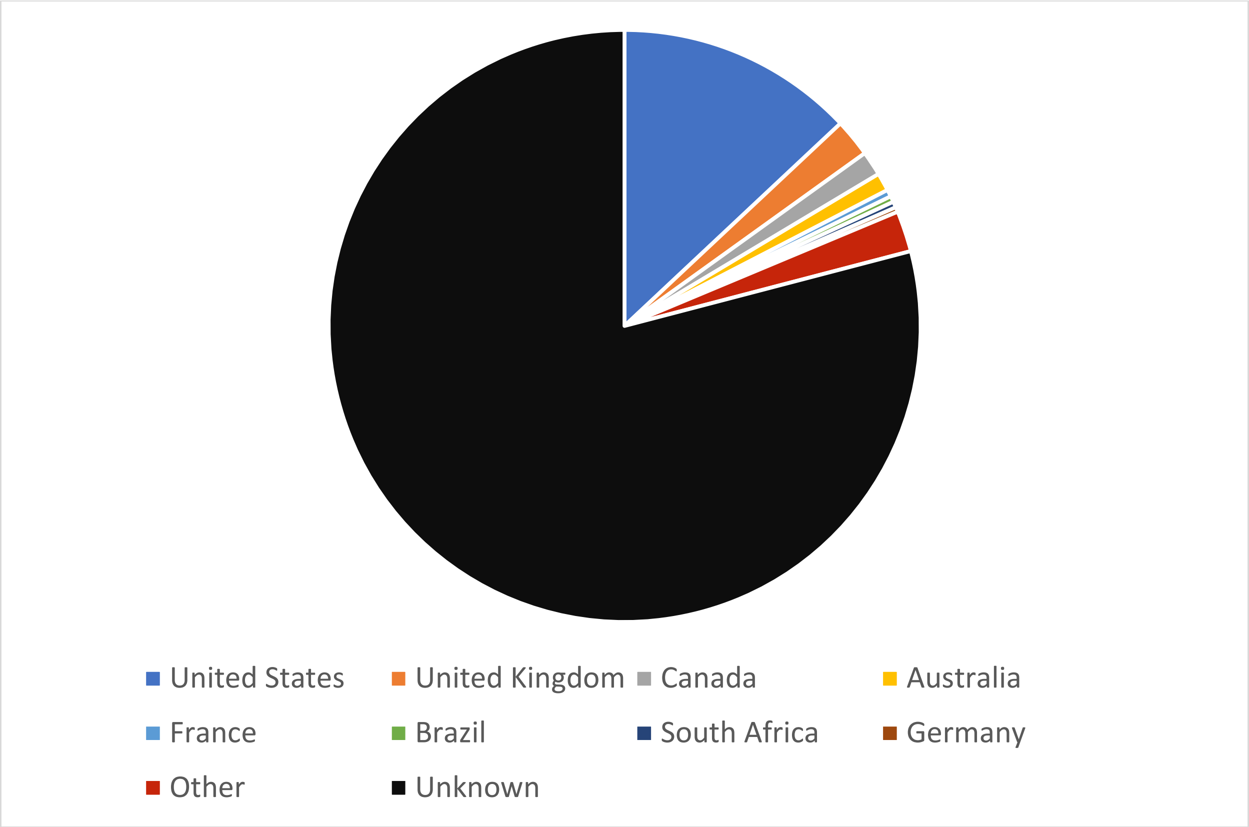
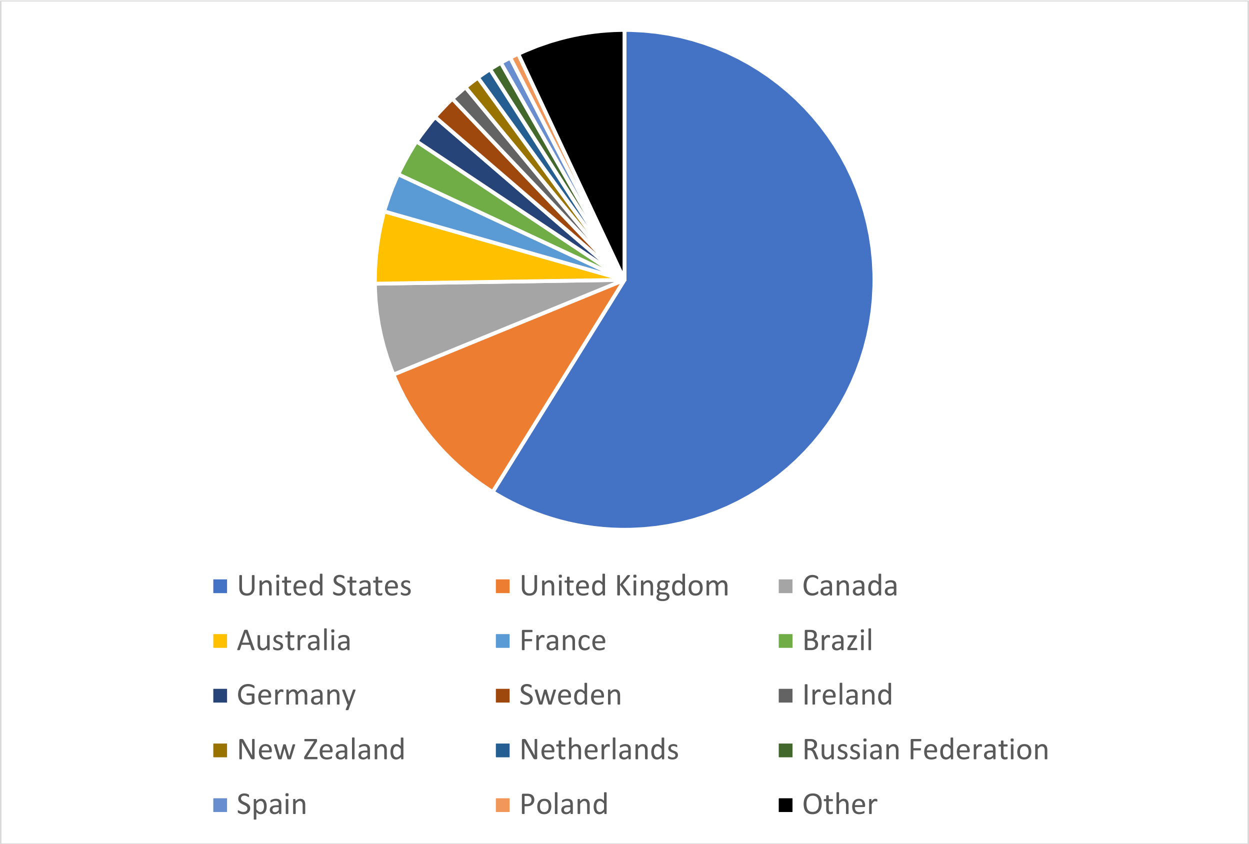


Fig. S1: Country of origin for GEDmatch test samples based on ip-address (when available as of January 1st 2022).



659

660 *Fig. S2: Country of origin for GEDmatch database as of January 1t 2022 based on ip-address (when available)*

661

Table S2: Observed shared cM ranges per degree of relationship in GEDmatch for the test set of 2,363,983 sample pairs. Pairs are limited to results where 9000 of the same loci are called for both kits.

Degree	Shared Cm Range	Kit Pairs
1	2300-3600	425
2	1300-2500	151
3	700-1400	136
4	300-800	198
5	100-450	629
Not Related	0	2362589

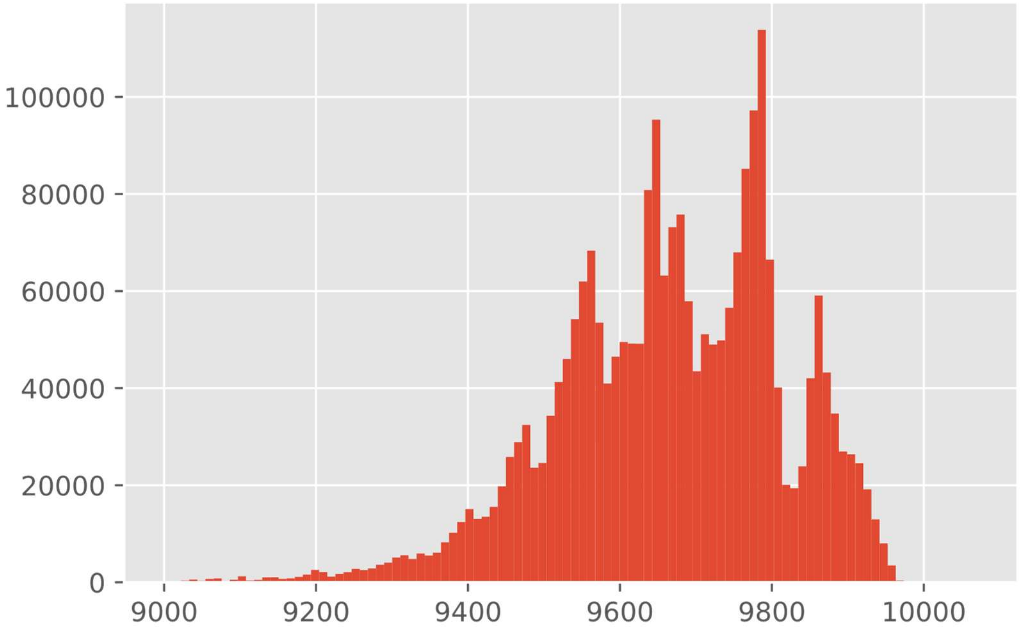


Fig. S3: GEDmatch test set. Overlapping passing SNPs for pairs of samples from 10K SNP multiplex.

SNP Multiplex Design

Table S3: : gnomAD population frequencies used during selection.

gnomAD Population	Description
afr	African-American/African ancestry
ami	Amish ancestry
amr	Latino ancestry
asj	Ashkenazi Jewish ancestry
eas	East Asian ancestry
fin	Finnish ancestry
nfe	Non-Finnish European ancestry
oth	Other ancestry
sas	South Asian ancestry

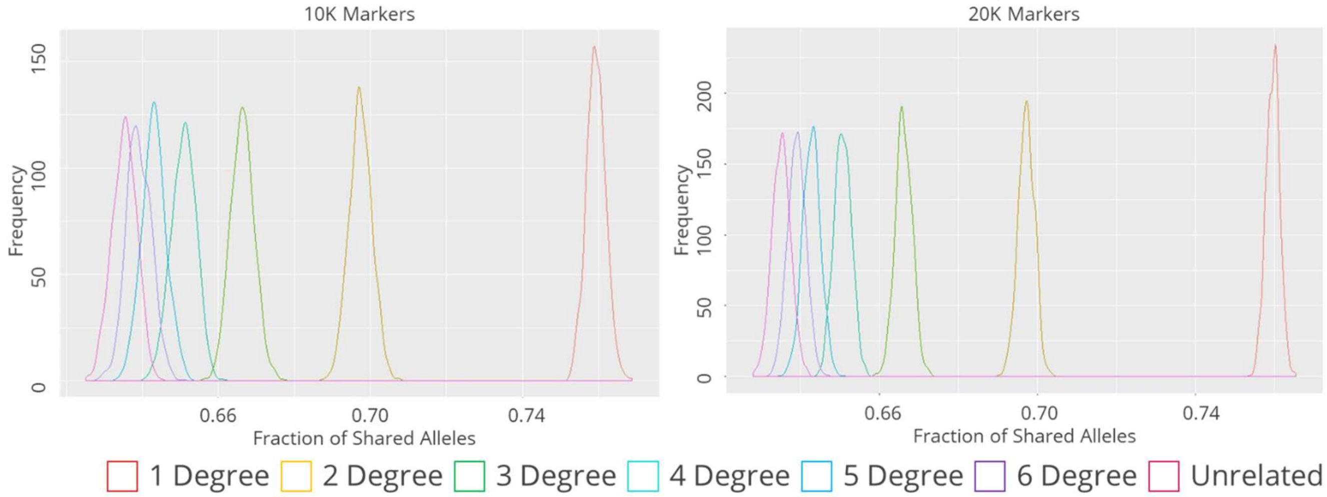


Fig. S4: Comparison of shared allele fractions between 10K (left) and 20K (right) SNP multiplexes from ped-sim simulations across kinship relationships from first to sixth degree and unrelated. 1,000 sample pairs were generated per degree of relationship.

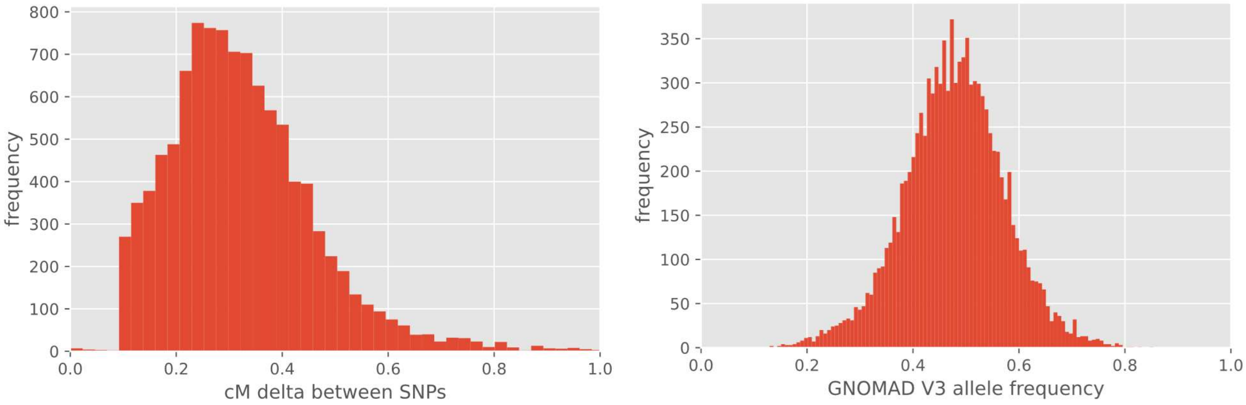


Fig. S5: SNP characteristics for of the 10K SNP set (ForenSeq Kintelligence). The left chart displays a histogram of the cM distances between loci; the right chart displays a histogram of the gnomAD allele frequency for the SNPs in the multiplex. A minimum of 0.1 cM was required for the kinship SNPs in the 10K multiplex; loci shown here that are below that value are SNPs informative for biogeographical ancestry, phenotype estimations or identity informative SNPs from the ForenSeq™ DNA Signature Prep Kit .

GEDmatch Pro Kinship Statistic Thresholds

Table S4: Windowed Kinship algorithm default thresholds as implemented in GEDmatch Pro and their influence on sensitivity of detection in first through fifth order relationships. Threshold values are based on the estimated false association (FA) rate in a search of the entire database.

Overlapping SNPs	Shared cM Total	Longest Segment cM	Sensitivity					1-Specificity	FAs in 1.5M Database
			1st	2nd	3rd	4th	5th		
9000	**140**	**30**	100%	100%	100%	99%	55%	0	0
8000	**150**	**30**	100%	100%	100%	99%	41%	0	0
6000	**180**	**30**	100%	100%	100%	66%	17%	0	0

Lack of Detection of Sister Alleles

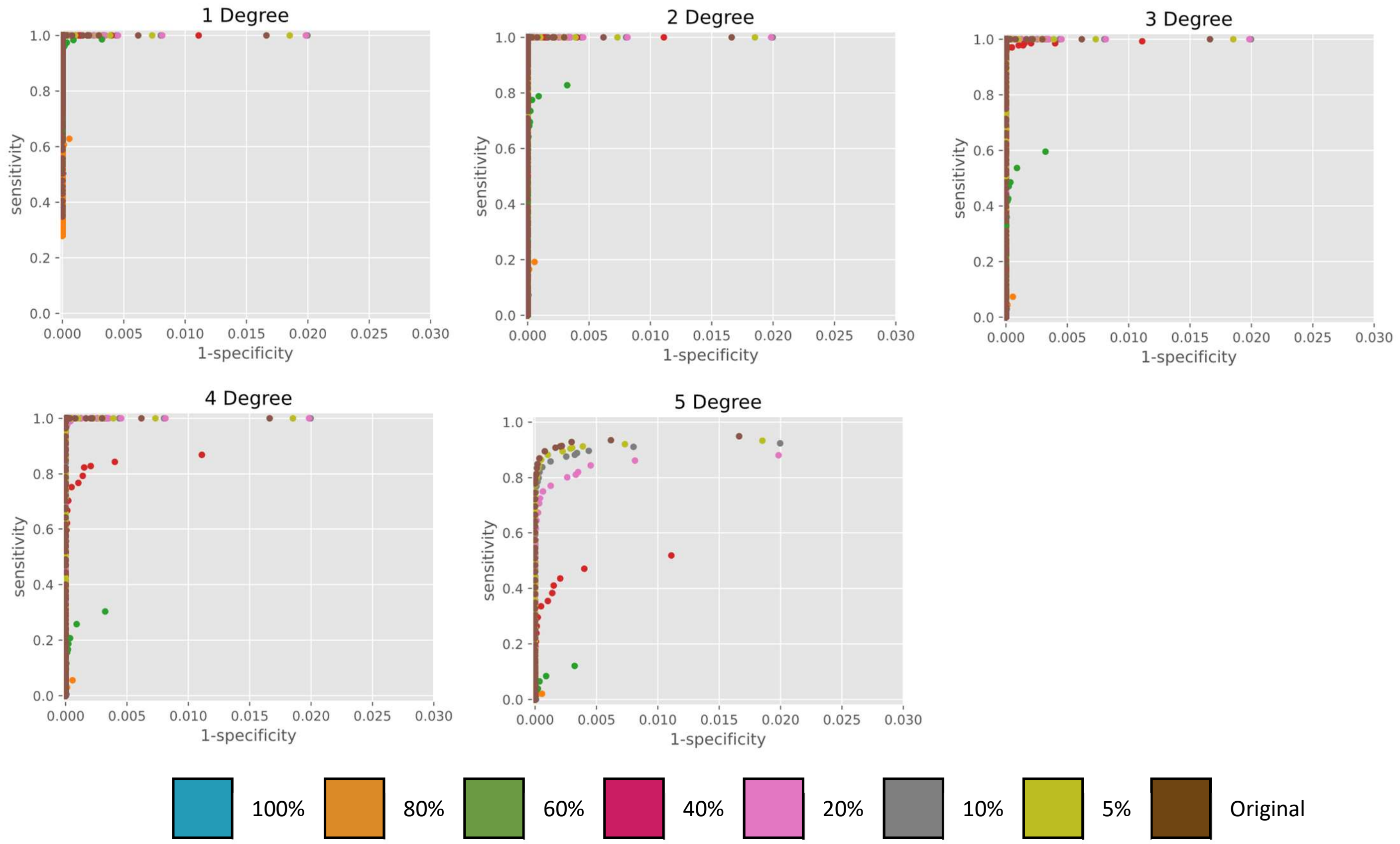


Fig. S6: ROC curves based on GEDmatch test set using simulated losses of heterozygote sister alleles between 5 and 100%.

Ped-sim Simulation

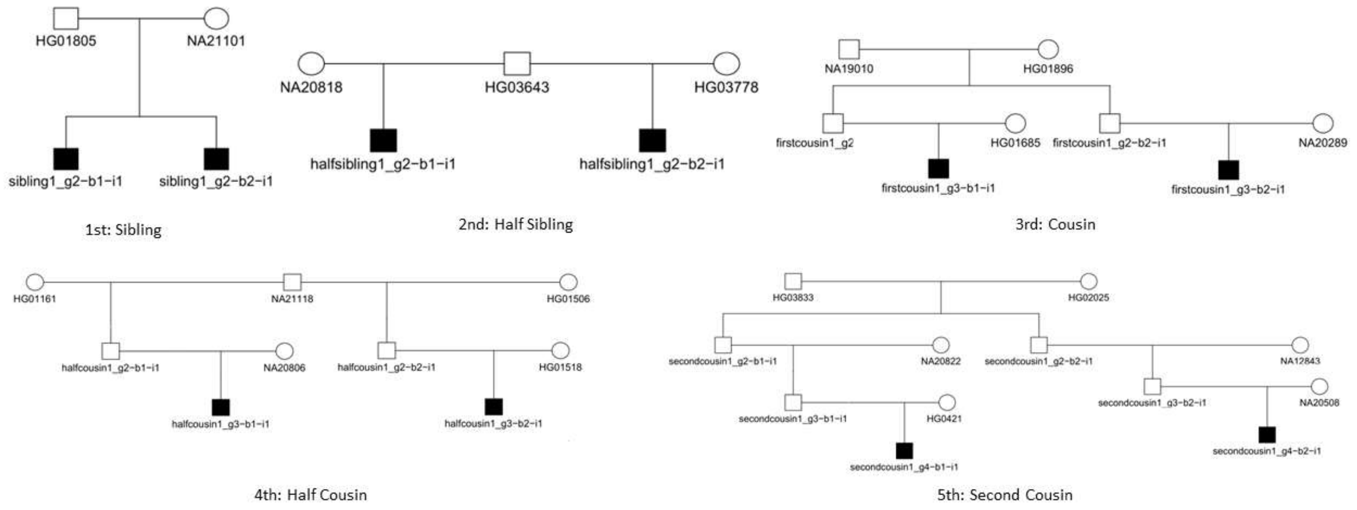


Fig. S7: Example pedigree for first through fifth relationship degrees using 1,842 founders from 1000 Genome Project samples. An example of each degree of a single pedigree simulated by ped-sim is shown. The biological sex of the founding samples was

ignored, and sex averaged linkage maps were used in the simulations. Only the darkened samples are output by ped-sim and used in evaluation scripts for this study. Pairs of samples from the same pedigree were considered true relatives, pairs of samples across pedigrees are considered unrelated.

Known Kinship SNP Pedigrees

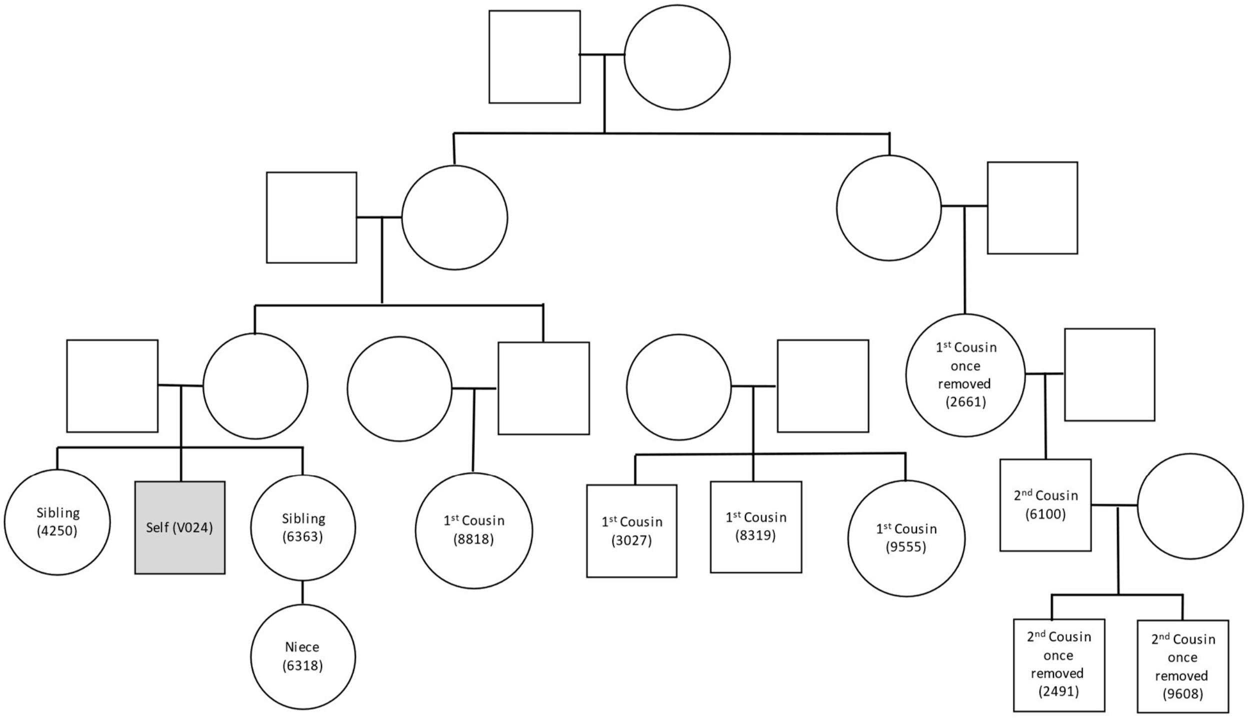
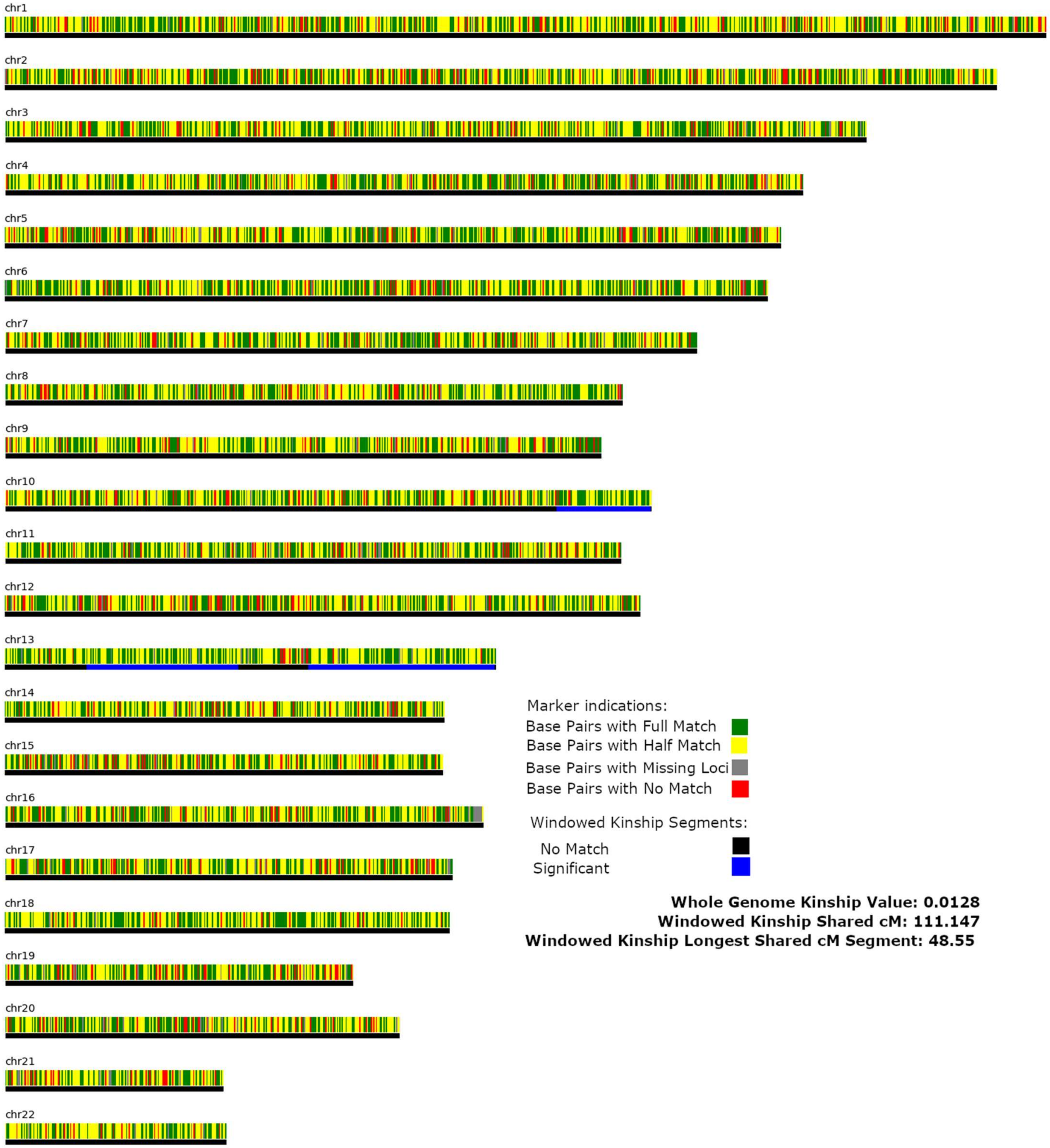


Fig. S8: Known extended pedigree from GEDmatch. Sample V024 was assigned as the person of interest and typed for the 10K SNP set using the ForenSeq™ Kintelligence kit; all other sample profiles are from microarray typing.

Related One to One Comparison, 2nd Cousin once removed



Unrelated One to One Comparison

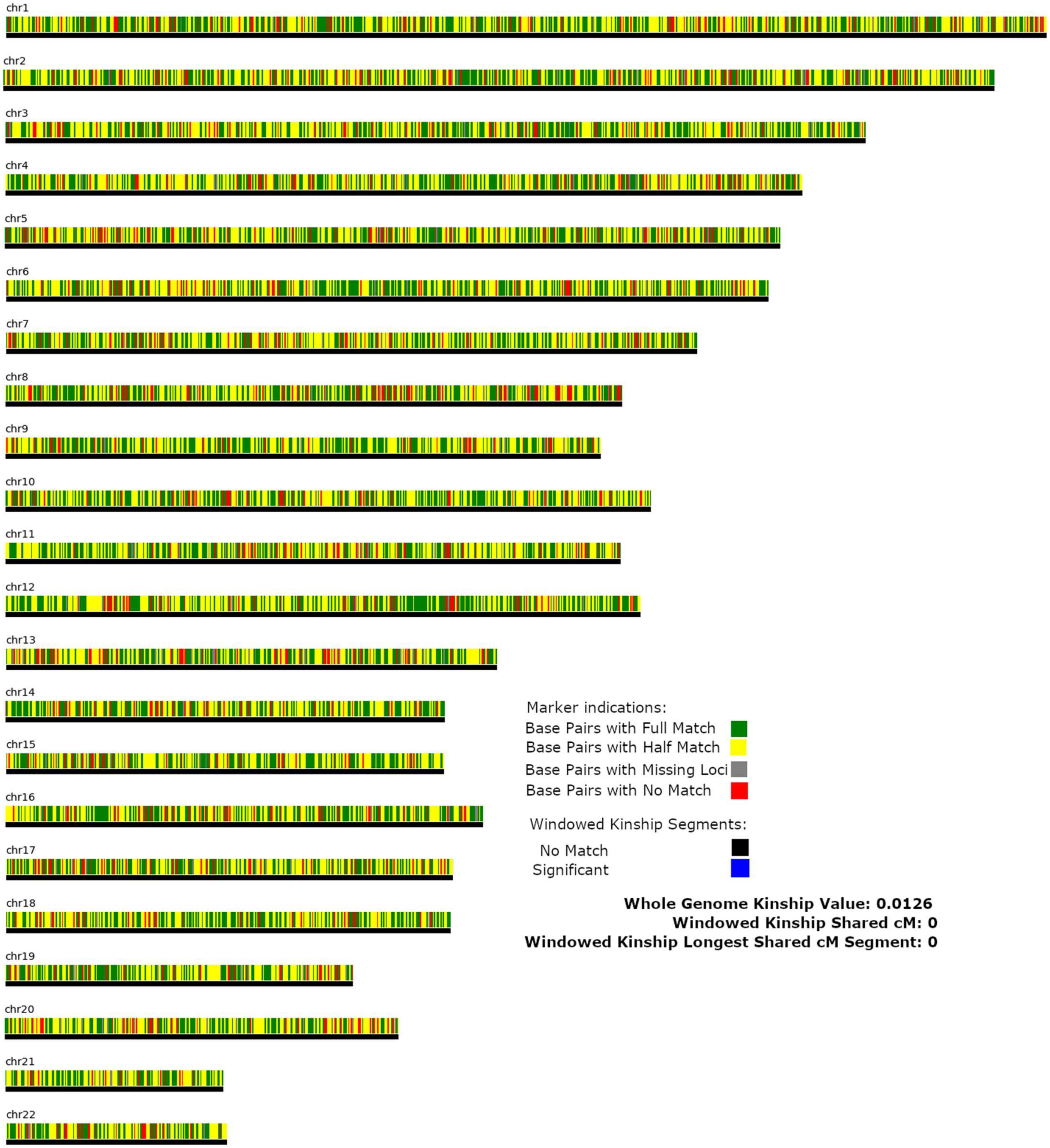


Fig. S9: Visual display of matching SNPs across genome. The upper panel show sample of interest V024 as compared to sample 9608 a 2nd cousin once removed from a known pedigree. The lower panel shows the sample of interest V024 as compared to an unrelated sample. The sample of interest was typed using the 10K SNP multiplex. Whole genome kinship cannot necessarily, and did not in this example, distinguish unrelated and related pairs since the overall number of "matching SNPs" is similar in both

scenarios shown. Since windowed kinship uses locus proximity and searches for segments of shared kinship, it better distinguishes more distant relationships from 10K SNP data.

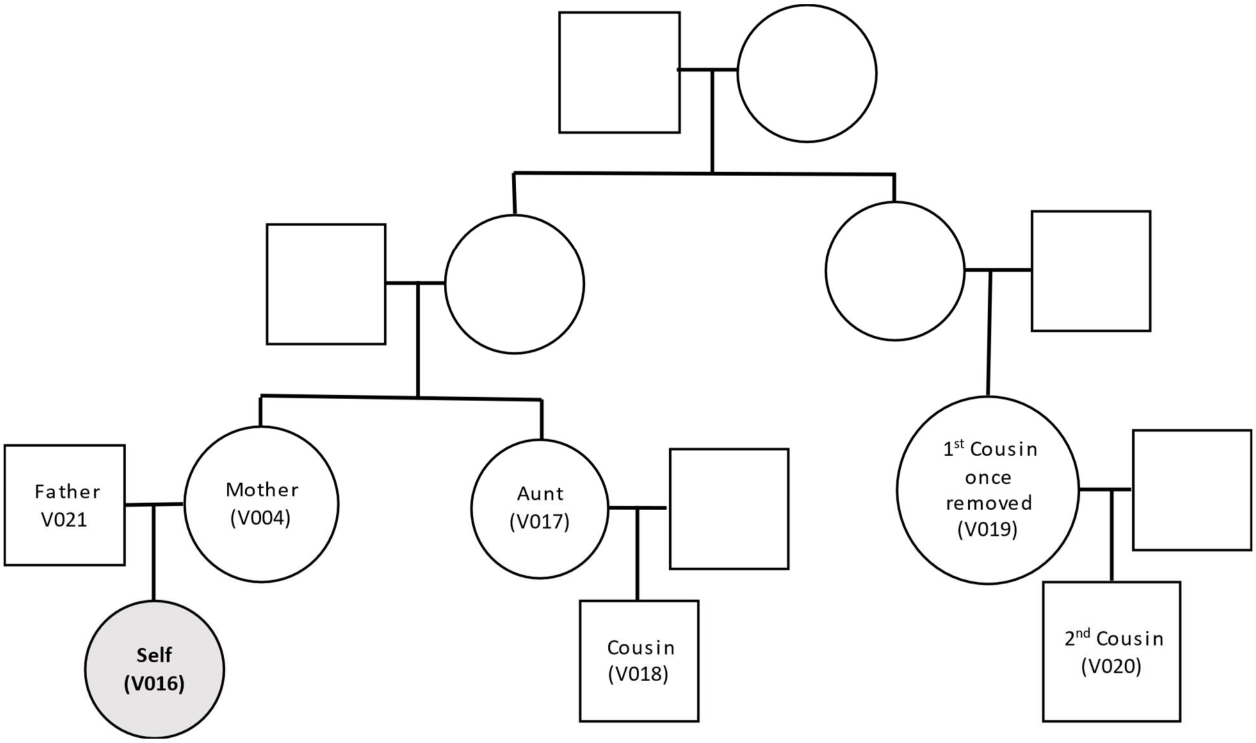
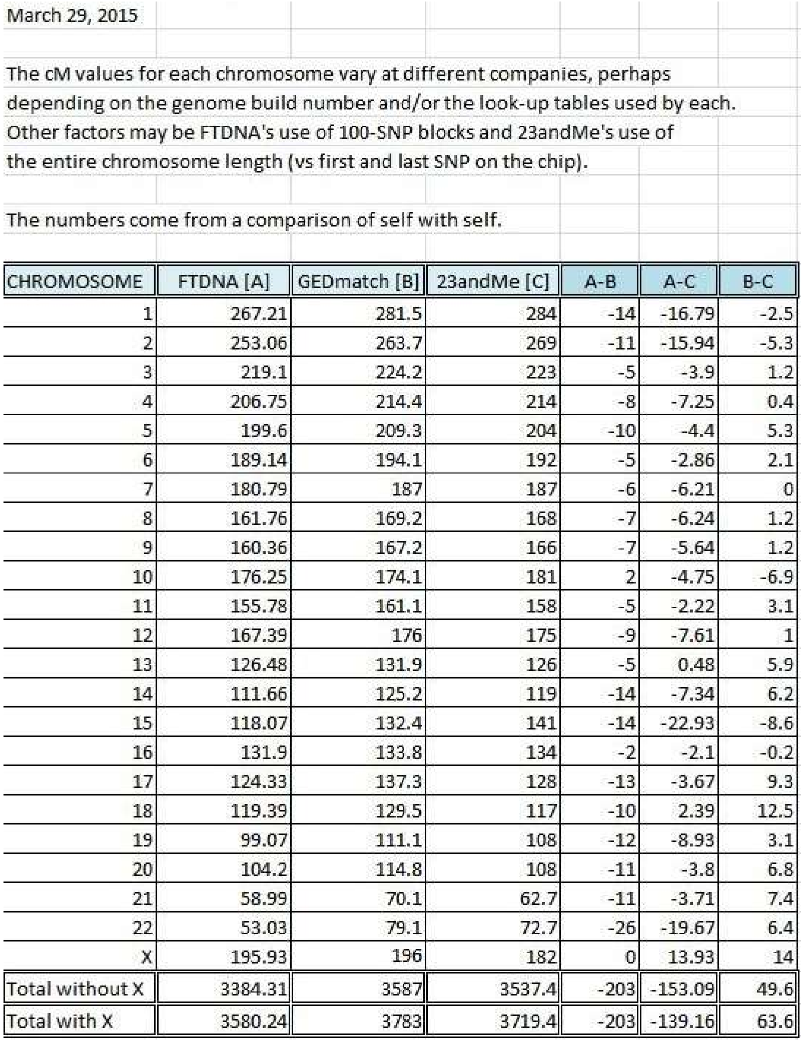


Fig. S10: Known family pedigree used for mock casework study. All samples in labeled nodes were typed at the 10K SNP set using the ForenSeq Kintelligence Kit. Sample V016 was partially degraded, 250 pg template was used and read counts were reduced by increasing sample numbers per MiSeq FGx run to simulate a challenging case-type sample for FGG database query using the windowed kinship algorithm.



March 29, 2015

The cM values for each chromosome vary at different companies, perhaps depending on the genome build number and/or the look-up tables used by each. Other factors may be FTDNA's use of 100-SNP blocks and 23andMe's use of the entire chromosome length (vs first and last SNP on the chip).

The numbers come from a comparison of self with self.

CHROMOSOME	FTDNA [A]	GEDmatch [B]	23andMe [C]	A-B	A-C	B-C
1	267.21	281.5	284	-14	-16.79	-2.5
2	253.06	263.7	269	-11	-15.94	-5.3
3	219.1	224.2	223	-5	-3.9	1.2
4	206.75	214.4	214	-8	-7.25	0.4
5	199.6	209.3	204	-10	-4.4	5.3
6	189.14	194.1	192	-5	-2.86	2.1
7	180.79	187	187	-6	-6.21	0
8	161.76	169.2	168	-7	-6.24	1.2
9	160.36	167.2	166	-7	-5.64	1.2
10	176.25	174.1	181	2	-4.75	-6.9
11	155.78	161.1	158	-5	-2.22	3.1
12	167.39	176	175	-9	-7.61	1
13	126.48	131.9	126	-5	0.48	5.9
14	111.66	125.2	119	-14	-7.34	6.2
15	118.07	132.4	141	-14	-22.93	-8.6
16	131.9	133.8	134	-2	-2.1	-0.2
17	124.33	137.3	128	-13	-3.67	9.3
18	119.39	129.5	117	-10	2.39	12.5
19	99.07	111.1	108	-12	-8.93	3.1
20	104.2	114.8	108	-11	-3.8	6.8
21	58.99	70.1	62.7	-11	-3.71	7.4
22	53.03	79.1	72.7	-26	-19.67	6.4
X	195.93	196	182	0	13.93	14
Total without X	3384.31	3587	3537.4	-203	-153.09	49.6
Total with X	3580.24	3783	3719.4	-203	-139.16	63.6

Fig. S11: Comparison of total cM values per chromosome for FamilyTreeDNA, GEDmatch and 23andMe for one DNA sample. Sourced from ISOGG wiki.[7]

[7] https://isogg.org/wiki/CentiMorgan