Research paper

Forensic genealogy—A comparison of methods to infer distant relationships based on dense SNP data

Daniel Klinga,⁎, Andreas Tillmarb, c

a Department of Forensic Sciences, Oslo University Hospital, PB. 4950 Nydalen, NO-0424, Oslo, Norway
b Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
c Department of Clinical and Experimental Medicine, Faculty of Health Sciences, Linköping University, Linköping, Sweden

ARTICLE INFO

Keywords:
Forensic genealogy
SNP
Forensic statistics
Classification
Identity by descent

ABSTRACT

The concept forensic genealogy was discussed already in 2005 but has recently emerged in relation to the use of public genealogy databases to find relatives of the donor of a crime stain. In this study we explored the results and evaluation of searches conducted in such databases. In particular, we focused on the statistical classification that entails from the search and study the variation observed for different relationship classes. The forensic guidelines advocate the use of the likelihood ratio (LR) as a mean to measure the weight of evidence, which requires exact formulation of competing hypotheses. We contrast the LR approach with alternative approaches relying on identical by state (IBS) measures to estimate the total length of shared genomic segments as well as identical by descent (IBD) coefficients for a pair of individuals.

We used freely accessible data from the 1000 Genome project to perform extensive simulations, generating data for a number of distinct relationships. Specifically we studied some overarching relationship classes and the performance of the above-mentioned evaluative approaches to classify a known pair of relatives into each class.

The results indicate that the traditional LR approach as a single source of classification is as good as, and in some cases even better than, the alternative approaches. In particular the true classification rate is higher for some distant relationship. However, the LR approach is both computer-intensive and sensitive to population frequencies as well as genetic maps (positions of the markers). We further showed that when combining different classification approaches, a lower false classification rate was achieved while still maintaining a high true classification rate.

1. Introduction

Recently a number of cold criminal cases [1–4] and missing person cases have been solved1, or given new leads, by what is known as forensic genealogy approaches. The approaches involve using high density SNP data, say more than 600,000 markers, in combination with large public databases to trace the relatives of the unknown donor [5,6]. Particularly, so called direct-to-consumer (DTC) companies market tests that will analyze more than 600,000 autosomal SNP markers on high-density microarrays2. Users may subsequently upload their raw genotype data to third-party companies, primarily GED-match,3 essentially making their data publicly available for further processing [5]. Progress in DNA sequencing technologies [7,8] have further significantly facilitated the processing of biological samples with low amount or degraded DNA, in turn yielding output that can be used for searches in DTC associated databases.

Although the genealogy approach to solving crimes or finding missing persons has prevailed in several cases, critical concerns have been raised to the use (or misuse) of big DNA data and public databases for criminal investigations [4,9]. These opinions include issues related to ethical and jurisprudence aspects as well as more theoretical thoughts about the statistical framework and methods used for relationship inference based on DNA data [10–12]. Although ethical and legal issues are of paramount importance, the present study focused on the latter topic. Methods (both in terms of laboratory and interpretation) used within forensic applications have high demands on quality and validity, due to its use and potential consequences. This entails that methods should have solid scientific foundations, be thoroughly

⁎ Corresponding author.
E-mail address: rmdakl@ous-hf.no (D. Kling).
1 See for instance http://dnadoeproject.org.

https://doi.org/10.1016/j.fsigen.2019.06.019
Received 28 January 2019; Received in revised form 15 May 2019; Accepted 24 June 2019
Available online 28 June 2019
© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).
validated and assured to be of high quality before the methods are applied in routine forensic casework [13].

Within forensics genetics, database searches for relatives of an unknown donor have been thoroughly studied [14–23] and are used as a tool for carefully selected routine cases in several countries and jurisdictions. This application is, however, restricted to DNA databases containing only convicted individuals (or traces with unknown origin) and the fact that the number of genetic markers, in this context, are limited (say around 20) [14,24]. This limitation entails that, generally, only first degree relatives (e.g. parent, child, full sibling) can be detected. Fig. 1A illustrates the concept with a common set of 23 forensically relevant STR markers. To be able to explore relationships beyond first degree relatives denser marker sets are needed.

Pedigrees spanning generations can be built based on genealogical records. However, the amount of shared chromosomal segments through common ancestry is quickly diluted as the number of generations increase. For instance two individuals with a pair of common ancestors three generations back (2nd cousins) will share an average of 3.12% of their total DNA identical by descent (IBD) which amounts to a total genetic length of roughly 212 cM. Previous studies have shown the potential to detect 2nd to 9th cousins [25] by using dense sets of SNP markers while Frazer et al. suggests that the background relatedness may be as close as 3rd cousins [26]. Due to a large proportion of randomness in which genetic material is transmitted from a parent to a child, the relative variation from the expected degree of DNA sharing increases as the degree of relatedness increases which could, potentially, lead to misclassifications [27–29].

This study aims to investigate possibilities and limitations with current established statistical methods for relationship classification based on dense SNP data. The results should highlight important features of the assumptions inherent in current models and could also point out the direction for further research areas. Three different approaches to infer IBD measures between pairs of individuals were selected (see Fig. 1C), all with a theoretical framework, mathematical definition as well as being employed in a number of applications. First, we explore the Likelihood approach, well known in forensic and medical genetics, to measure the weight of the evidence [30–34]. Briefly, the likelihoods are calculated as the conditional probability of observing genetic marker data for a set of individuals given a precise hypothesis about their relatedness. Important parameters for the likelihood calculations are population estimates of allele frequencies, genetic maps as well as information about the association of alleles (gametic equilibrium).

Secondly, we turn the attention to two different methods based on the identity by state (IBS) proportions for the persons of interest and condition on those states to infer identical by descent parameters. We define the KING method, previously described by Mannichaikul et al. [27], by simply counting the number of shared alleles identical by state for each marker and average over a large number of markers yielding a measure of the degree of relationship. Several implementations exist and detailed information can be found elsewhere [27,35] and in the section of this paper. We further define the Segment approach. Briefly, this method measures segments along the chromosomes where a pair of individuals shares at least one allele along the complete segment [36–39]. The length of each segment is commonly measured in centiMorgan (cM) and the total length of all shared segments provides a measure of the relationship. Versions of this approach are used by several of the DTC companies to infer degrees of relatedness between individuals.

To evaluate the performance of the methods, we focused on classification rates. We define classification as assigning a degree of relatedness to a pair of individuals (see Fig. 1B). Essentially this equates to determining the number of generations separating individual A from individual B. In forensic genealogy, classifications are used to roughly indicate where the suspect or missing person is located in the pedigree.
of the relative with whom a match is obtained. Consequently, misclassifications could potentially lead to false accusations, false leads in criminal investigations and overly time-consuming labor. We studied false classification rates (e.g. test say cousin when sibling is the true degree of relatedness) as well as true classification rates (e.g. test say cousin when cousin is the true degree of relatedness), both of which are essential in familial searches [17,19,40,41]. Furthermore, if a cost can be measured for each misclassification we can further tune the search parameters to minimize the total overall cost using a mathematical framework as described by Tillmar et al [42]. Apart from the behavior of each model we studied how these rates were influenced by the number of DNA markers included and the degree of relatedness.

2. Material and methods

2.1. Reference data

We used genotype data from the 1000 Genomes project (Phase 3, build 20130502) [43,44]. Annotated marker data was obtained via ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502. For the purpose of this study only the individuals with UK ancestry were extracted (GBR) encompassing in total 90 unrelated individuals. We further constrained the number of genetic markers to the SNPs covered in the current microarray chip from the direct-to-consumer (DTC) company Ancestry.com to better mirror the procedure used in genealogy searches. In total, roughly 580,000 autosomal SNP markers were extracted (exact list may be obtained from the corresponding author upon request). For the resulting markers we obtained allele frequencies from the previously mentioned GBR individuals. Genetic positions were obtained from Rutger’s repository [45], available at http://compgen.rutgers.edu/rutgers_mapst.shtml, or alternatively interpolated for markers without a defined position.

2.2. Statistical methods

All methods covered in this study provide different means to estimate the genetic relationship between a pair of individuals and have been thoroughly investigated in previous studies. See Fig. 1C for a simple illustration of the approaches. In essence, the Likelihood approach computes the probability of observing the data (i.e. IBS states) given each possible inheritance pattern (i.e. IBD states) while, in contrast, the other two methods (KING and Segment approach) infers measures of IBD using the IBS states. A more detailed description of each approach follows below.

2.2.1. Likelihood approach

The likelihood approach is well known to forensic scientists, but also in medical genetics and has traditionally been the preferred approach when solving disputes of relatedness [30,46] or when doing linkage analysis [47–49]. Briefly, the method computes the conditional probability of observing some genetic marker data for a set of individuals and some precise hypothesis about the relationship between the individuals. Likelihoods for mutually exclusive hypotheses can be compared using a likelihood ratio (LR). Gjerst et al. [30] describes some guidelines for its use and interpretation in forensic genetics. Some key parameters in this approach are population allele frequencies, genetic positions of the markers as well as population substructures. This entails that the approach is sensitive to estimation of certain population parameters but also has the potential to find relatives with a few rare shared variants.

We used the Lander-Green algorithm [33] implemented in the software Merlin [47] to compute likelihoods. Briefly, the algorithm works by considering markers as steps in a hidden Markov chain and generally has a complexity linear to the number of markers. Using a Bayesian approach, likelihoods can be translated to posterior probabilities indicating how probable each relationship hypothesis is. Assuming each hypothesis is equally likely a priori, we get

$$Pr(H_i) = \frac{L_i}{\sum_j L_j}$$

Where Pr(H_i) is the posterior probability of hypothesis i and L_i is the likelihood of hypothesis x.

In relations to the likelihood approach, previous studies have demonstrated that dense sets of genetic markers tend to favor the genetically closest relationship [29,50,51] and inflate LOD scores in linkage analysis [34,52]. The problem has been attributed to association of alleles (linkage disequilibrium) causing an increased degree of allele sharing at several adjacent genetic markers, i.e. redundant information, while most traditional models do not accurately model this concept. Measures are needed to alleviate the problem; we refer to the process as marker pruning. The most naive approach is to select markers based on their inter-distance, i.e. we specify a minimum genetic distance, measured in centiMorgans (cM), for two markers to be included in our subset. Several previous studies have demonstrated that association of alleles generally extends shorter distances but varies considerably across the genome and between populations [53–57]. Based on data from the previously referenced studies we used 0.15 cM (roughly 150 kb) as a threshold for the distance between any two markers in our pruned subset. The pruning process continues along each chromosome and results in thinned sets of markers. In addition a threshold on the minimum population allele frequency (MAF) is implemented to filter markers with little expected information, i.e. low MAF. We used 0.2 to exclude non-informative markers with 0.5 representing markers with maximum information. A more refined approach uses measures of correlation, for instance r^2, between pairs of markers and their alleles. We used 0.2 as a threshold on the correlation, representing a fairly strict value and thus exclude rather than include markers if allelic association is present. In this study, we combined the approaches mentioned above to create pruned marker sets for likelihood computations. The present study does not encompass a thorough evaluation of these thresholds, we refer to Kling et al. [29,51] for more exploratory studies. We further used an efficient implementation in the Merlin software [34,47] allowing us to perform efficient likelihood computations on clusters of tightly linked markers.

2.2.2. Segment approach

The second approach exploits the fact that stretches of DNA are inherited unchanged through generations, without the interference of recombinations. Variants of the segment approach are commonly used by so called direct-to-consumer (DTC) companies to estimate the relationship between pairs of individuals [58]. The algorithm is essentially a version of homozygous haplotyping, described by Miyazawa et al. [59], in which pairs of chromosomes are compared (see Fig. 1) and the number of consecutive markers where a pair of individuals has identical homozygous genotypes is counted. In the present study, we extended this approach by also considering heterozygous genotypes similar to algorithms employed by DTC companies (also referred to as half-identical genotypes). Briefly, our algorithm compares the genotypes of two individuals by iterating through the list of markers and starts a new shared segment if at least one allele is identical for the two genotypes in consideration. A segment is terminated only if opposite homozygote genotypes are detected. A segment is further accepted, i.e. defined as IBD, if it exceeds the thresholds described below. We used the total length of accumulated shared segments, in cM, as a measure of the relationship. Several studies have previously explored the variation of shared DNA between pairs of relatives [38,60].
As a consequence of the randomness in which DNA is transmitted through generations, the probability that distant relatives share any segments in the genome identical by descent (IBD) will decrease inversely proportional with the number of generation separating the two individuals. We studied the sensitivity with regards to two thresholds employed to estimate the length of the shared segment between pairs of individuals. Specifically we studied the minimum cM (varying from 3 to 8) for a segment to be inferred as IBD and the minimum number of SNPs in each such segment (varying from 200 to 700 SNPs).

There are more intricate algorithms to compute shared segments, for instance the white paper from DTC company Ancestry.com [58] describes a method whereby chromosomes are first phased, i.e. assigning haplotypes, and subsequently matched. Indeed, the phasing is only used to find the start (or end) of shared segments but is reportedly better to detect more distant relationships. Furthermore, Al-Khudhair et al. [61] describe the use of so-called very rare genetic variants (vrgV), to explore distant relationships. There is certainly more to it than merely identifying shared segments; however the current study will not explore these extensions, mainly due to the fact that the DTC company GEDmatch currently employs a similar segment approach as described above.8

2.2.3. KING approach
The final approach embodies the ideas described by Manichaikul et al. [27] and utilizes observed identical by state (IBS) numbers to infer two different IBD measures by averaging over a large number of markers. Consider a pair of individuals, indexed with i and j. For non-inbred relationship their genetic relatedness can be described using three IBD markers. Considers a method whereby chromosomes are first phased, i.e. assigning haplotypes, and subsequently matched. Indeed, the phasing is only used to find the start (or end) of shared segments but is reportedly better to detect more distant relationships. Furthermore, Al-Khudhair et al. [61] describe the use of so-called very rare genetic variants (vrgV), to explore distant relationships. There is certainly more to it than merely identifying shared segments; however the current study will not explore these extensions, mainly due to the fact that the DTC company GEDmatch currently employs a similar segment approach as described above.8

2.2.3.1. Simulations
Extensive simulations were employed to generate data; the approach is illustrated in Fig. 2. Briefly, the process starts by randomly drawing sets of haplotypes, two for each founder of the pedigree. Phased haplotypes were available through the 1000 Genome project [43,44]. As noted previously, the referenced study employed a combination of sequencing technologies and dense SNP microarrays to perform genotyping. Phasing (i.e. obtaining information about specific haplotypes) was previously performed using the software SHAPEIT [63], see Choi et al. [64] for a recent discussion on the error rates of such phasing algorithms. We specifically used individuals (N = 90) with UK ancestry (GBR) yielding, in total, a pool of 180 phased haplotypes.

The process continues by performing gene dropping [65] whereby the haplotypes (in Fig. 2 illustrated for a single marker pair) is subject to crossovers and transmitted to the children (descendants). The probability for crossovers is modeled using genetic maps [45] and Kosambi’s equation [66] to convert genetic positions into recombination rates. Mutations were not considered, a reasonable simplification when using SNP markers. Throughout the simulations, information about phase, i.e. paternal and maternal allele information, is kept whereas in the final step such details are removed. Note, information about identical by descent status for each marker is stored during the simulations but discarded at the end.

Finally, for each of the simulated pairs of individuals, the statistical metrics previously described were computed. Note, these computations use identical by state status for each marker.

2.3. Generating data

2.3.1. Simulations
Extensive simulations were employed to generate data; the approach is illustrated in Fig. 2. Briefly, the process starts by randomly drawing sets of haplotypes, two for each founder of the pedigree. Phased haplotypes were available through the 1000 Genome project [43,44]. As noted previously, the referenced study employed a combination of sequencing technologies and dense SNP microarrays to perform genotyping. Phasing (i.e. obtaining information about specific haplotypes) was previously performed using the software SHAPEIT [63], see Choi et al. [64] for a recent discussion on the error rates of such phasing algorithms. We specifically used individuals (N = 90) with UK ancestry (GBR) yielding, in total, a pool of 180 phased haplotypes.

8 See instance https://isogg.org/wiki/Autosomal_DNA_statistics.

2.3.2. Unrelated individuals
To compare DNA data from unrelated individuals we used an all-to-all comparison approach, where each pair of individuals in our data set were compared and the statistic metrics described above where computed for each comparison. In line with the results described in previous studies [61,67], pairs of distant relatives were detected (2nd to 3rd degree relatedness) in the reference data from 1000 Genomes. As the subsequent analyses assume an unrelated set of individuals, we removed these outliers, defined as 3rd degree relationship or closer, from the entailing simulations.

2.4. Classification of relationship classes
Throughout this study we considered some simple non-inbred relationship classes. Specifically we defined S1-1, S2-2, S3-3, S4-4 as full siblings, first cousins, second cousins and third cousins, respectively. Fig. 1B illustrates the relationship classes and Skare et al. [28] contains further details. In terms of average genome sharing, S2-2 is identical to S1-3, for instance, and our methods will therefore not be able to distinguish such identical relationships.

Furthermore, in order to assign a relationship class for a pair of individuals, we used average and expected metrics from previous studies [27,68], summarized in Table 1. We fitted a simple logarithmic regression model (see Supplementary Fig. 1) by assigning integers according to increasing degree of relatedness, i.e. 1 corresponds to S1-1 and 5 corresponds to Unrelated. In the entailing classification, the output was rounded to the closest integer coding for the different relationship classes (see Table 1).

Secondly, for the likelihood approach we performed classifications based on the relationship where the likelihood was maximized. No constraints were implemented with regards to the relative difference in likelihoods in the classifications, which entails that a pair of individuals could have a very slight difference in the likelihoods for two relationship classes but still be classified in one of them.

2.5. Expected number of relatives
For practical reasons (e.g. the size of a candidate list) it might be important to know the expected number of relatives that an individual has for a given degree of relationship. There are several factors that will influence this number. The most important, in terms of computing the average number of expected relatives, is the number of children born per generation and family, a figure that has varied over time and varies among different populations. We adopted the model presented by Henn et al. [25], but instead of using a mathematical expression with a fixed value for the number of children per family, we extended the model by allowing the number of children to vary from family to family and from generation to generation, given a discrete probability distribution (Poisson distribution [69], with a mean corresponding to the mean number of births per family) and further assumed a constant generation
time of 25 years, we implemented the model in a simulation approach where the variation in the number of relatives given the above mentioned probability function could be studied. We further estimated the expected number of 1st cousins down to the number of 3rd cousins from 10,000 simulations assuming 1 to 5 children per family and generation.

Furthermore, we used historical records of birth rates from UK⁶ and Swedish⁷ records in order to estimate the number of relatives for individuals born in 1955 or in 2005. Due to changes in birth rates through history, such estimates are expected to vary depending on the year of birth.

### 3. Results

The results are divided as follows; first we describe some explorative analyses where the fundamental properties of the Likelihood, Segment and KING approaches are illustrated separately. We use simulations to generate data, further outlined in Material and Methods. Secondly we turn the attention to classification and compare and evaluate the potential of the different approaches to classify pairs of individuals with unknown degree of relationship into the correct classes. Finally, we present results on the expected number of relatives for a given individual within each relationship class.

---

**Table 1**

Relationships and classification ranges. The table describes relationship classes and values (extracted from other studies) used for classifications.

<table>
<thead>
<tr>
<th>Relationship class</th>
<th>Integer</th>
<th>Average shared segments (cM)</th>
<th>Kinship coefficient **</th>
<th>Pr(IBD = 0) **</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1-1</td>
<td>1</td>
<td>2629 cM</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>S2-2</td>
<td>2</td>
<td>874 cM</td>
<td>0.063</td>
<td>0.75</td>
</tr>
<tr>
<td>S3-3</td>
<td>3</td>
<td>233 cM</td>
<td>0.016</td>
<td>0.938</td>
</tr>
<tr>
<td>S4-4</td>
<td>4</td>
<td>74 cM</td>
<td>0.004</td>
<td>0.984</td>
</tr>
<tr>
<td>Unrelated</td>
<td>5</td>
<td>10 cM</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

* Reference values (in cM) extracted from the Shared cM Project [68] and are based on empirical data.

** Rounded reference values extracted from Manichaikul et al. [27] based on theoretical derivations.

*** Value for unrelated is here defined as somewhere in the range of 5-7th cousins.

---

**Fig. 2.** Illustration of the simulation procedure implemented in this study. The figure illustrates how founders are drawn from a pool of phased haplotypes. Haplotypes are further transmitted through the pedigree using a method known as gene dropping whereby the haplotypes are subject to recombinations based on genetics maps. In the final step only the individuals we are interested in are retained. Phase information is known during the simulation process (highlighted using colors and pipes), whereas in the final step this information is removed.
3.1. Distribution of the relationship metrics

3.1.1. Likelihoods and posterior probabilities

The Likelihood approach computes the conditional probability of observing some genetic marker data for a set of individuals given some hypothesis about relatedness between the individuals. In essence this equates to computing the probability of the IBS states given each possible IBD state between a set of individuals. The approach incorporates population parameters such as specific allele frequencies, recombination rates etc. The main challenge for the Likelihood approach, as implemented in this study, is to find the most suitable parameter settings for the pruning procedure, further detailed in Material and methods. Briefly we construct a reduced set of SNP markers where the aim is to prune redundant information. Specifically, we removed all SNP markers with minor allele frequency, computed based on a UK population, lower than 0.2. Secondly, we selected only markers with a minimum distance of at least 0.15 cM, roughly equal to in average 150 kb physical distance. The remaining set of markers was further subjected to a third filter where a sliding window of 1 cM was used to prune markers with a correlation (measured through $r^2$) higher than 0.2. Secondly, we explored the potential of combining the different methods into a joint classification approach. We define a true classification as assigning another degree of relationship than the true relationship. Whereas when using the individual approaches, a false classification is straightforward, combining different approaches may imply statistical significance. However, this suggests that false classifications are expected for these relationship classes.

The mean values for the total length of shared segments from our simulations corresponded well with the empirical observations in the Shared cM Project (version 3.0) [68], corroborating our simulation model. For pairs of unrelated individuals, as previously pointed out, longer stretches of shared segments were identified, potentially suggesting unknown distant relatedness in the reference data [61,67].

3.1.3. Average identical by descent measures

Finally we explored what we call the KING approach. The approach rely on dense sets of SNP markers to provide estimates of the true IBD states between a pair of individuals [27]. The distributions for the estimated kinship coefficient (also referred to as half-relatedness) and Pr (IBD = 0) are shown in Fig. 5. Similar to the segment approach, the distributions for the relationship classes S1-1 and S2-2 were clearly separated for both metrics, whereas for the distant relationship classes (S3-3, S4-4 and Unrelated) a higher degree of overlap was observed (see Supplementary Fig. 4).

The obtained mean values for the kinship coefficient and for the Pr (IBD = 0) from our simulations corresponded well with the theoretical expectations (Table 3).

3.2. Classification

Classifications are used in a variety of fields. This study will cover how different approaches can be used to classify a pair of individuals into a certain degree of relationship. Genealogists use classifications to know where and how far to trace relatives in a pedigree. It entails that we need studies on how well these classification approaches perform; both in term of accuracy but also on the risk of making false predictions with regards to the degree of relationship.

In the following section, classifications were first performed using each individual method (Likelihood, Segment and KING approaches). Secondly, we explored the potential of combining the different methods into a joint classification approach. We define a true classification as correctly assigning the expected degree of relationship to a pair of individuals (including the degree unrelated) and conversely a false classification as assigning another degree of relationship than the true relationship. Whereas when using the individual approaches, a false classification is straightforward, combining different approaches may induce undetermined (unclassified) results. Therefore, another classification group is necessary; we refer to this as Undefined. An undefined classification is neither false nor true, but the methods cannot unanimously determine the class.
3.2.1. Performance of individual classification approaches

Classification rates, for the different individual methods, are shown in Tables 4–7, where the true relationship class is indicated in each column. Among the tested methods the Likelihood approach produced the highest true classification rates for all relationship classes (apart from Unrelated), followed by the Segment approach and lastly the KING classifiers (Tables 4–7 and Fig. 6). Classification of the closest relationships (S1-1 and S2-2) had a high precision for all individual methods, with 100% correct classifications aside from the kinship classifier with a 99.8% true classification rate for the S2-2 relationship. For the more distant relationship classes (S3-3 and S4-4) the success rate declined considerably for all methods, where only the Segment and Likelihood approach had a true classification rate above 90% for the S3-3 class. For the unrelated individuals, the segment approach was by far most accurate, with 97.1% true classifications. Supplementary Fig. 5 provides an alternative representation of the true classification rates divided into methods instead of relationships.

3.2.2. The power of combining different approaches

To improve on the individual classifiers, combinations of those were tested to evaluate the improved classification rates. As indicated previously, an additional class (Undefined) was introduced to cover the classifications that will be undetermined. Table 8 illustrates the classification rates when combining all four classifiers (Likelihood, Segment, KING (Pr(IBD = 0) and kinship coefficient). This requires all classifiers to assign the same degree of relatedness to a given pair of individuals. If the classifiers are ambiguous in terms of the assigned degree of relatedness, the class is undefined. As expected the true classification rates will decrease, i.e. performance of the individual approaches in terms of true classifications can never be exceeded. Table 8 illustrates that in particular for the S4-4 relationship class, combining all four classifiers results in only roughly 15% true classifications, whereas 77% is assigned as undefined classifications. This further suggests a high lack of agreement between the methods with regards to this relationship class. Furthermore, the false classification will decrease, rendering a number of undefined classifications. In particular, this applies to the S3-3 and S4-4 relationship classes.

![Fig. 4.](image1.png)

**Fig. 4.** Polygon plot showing the average length of shared cM between pairs of simulated individuals (n = 1000) and the relationship classes covered in this study (S1-1 (top), S2-2, S3-3, S4-4 and Unrelated (bottom)). Each point on the x-axis represents a different threshold used to determine shared segments, [x,y], where x represents the minimum cM and y represents the minimum number SNP markers needed to define a segment as IBD. In addition to the mean, a pointwise 95% confidence interval is plotted.

![Fig. 5.](image2.png)

**Fig. 5.** Scatter plot (upper) with 95% confidence ellipses of the kinship coefficient and Pr(IBD = 0) for 1000 simulated pairs of individuals representing the different relationship classes (S1-1, S2-2, S3-3, S4-4 and unrelated). Violin plots (middle) illustrating the distribution of the Pr(IBD = 0) estimates (range from 0 for monozygotic twins to 1 for unrelated). Violin plots (lower) illustrating the distribution of the kinship coefficient (range from 0.5 for monozygotic twins to 0 for unrelated).

### Table 2

Fractions within each relationship class that will have any shared segments identical by state for each given threshold in cM based on 1000 simulations for each relationship class. Note, this does not necessarily equate to the probability that any segments are shared identical by descent.

<table>
<thead>
<tr>
<th>Relationship class</th>
<th>3 cM</th>
<th>4 cM</th>
<th>5 cM</th>
<th>6 cM</th>
<th>7 cM</th>
<th>8 cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1-1, S2-2, S3-3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S4-4</td>
<td>0.986</td>
<td>0.982</td>
<td>0.977</td>
<td>0.965</td>
<td>0.945</td>
<td>0.93</td>
</tr>
<tr>
<td>Unrelated</td>
<td>0.277</td>
<td>0.210</td>
<td>0.168</td>
<td>0.14</td>
<td>0.119</td>
<td>0.099</td>
</tr>
</tbody>
</table>
Table 3
Comparison of theoretical values for Kinship coefficient and Pr(IBD = 0) and values obtained from simulations.

<table>
<thead>
<tr>
<th>Relationship class</th>
<th>Kinship coefficient (theoretical)</th>
<th>Kinship coefficient (mean from simulations)</th>
<th>Pr(IBD = 0) (theoretical)</th>
<th>Pr(IBD = 0) (mean from simulations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1-1</td>
<td>0.250</td>
<td>0.251</td>
<td>0.250</td>
<td>0.248</td>
</tr>
<tr>
<td>S2-2</td>
<td>0.0625</td>
<td>0.0624</td>
<td>0.750</td>
<td>0.754</td>
</tr>
<tr>
<td>S3-3</td>
<td>0.0156</td>
<td>0.0157</td>
<td>0.938</td>
<td>0.942</td>
</tr>
<tr>
<td>S4-4</td>
<td>0.00390</td>
<td>0.0044</td>
<td>0.964</td>
<td>0.986</td>
</tr>
<tr>
<td>Unrelated</td>
<td>&lt; 0.001</td>
<td>0.00096</td>
<td>&gt; 0.99</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Table 4
Classification rates for the Likelihood approach. Each column represents the relationship that has been simulated and the rates in each row represent the classifications. A subset of 21,517 markers were extracted using the criteria, MAF > 0.2, mincM > 0.15, r2 < 0.2.

Table 5
Classification rates for the Segment approach. Each column represents the relationship that has been simulated and the rates in each row represent the classifications. Segments were called based on the criteria, minSNP > 700.

Table 6
Classification rates for the KING approach (Kinship coefficient). Each column represents the relationship that has been simulated and the rates in each row represent the classifications.

Table 7
Classification rates for the KING approach (Pr(IBD = 0)). Each column represents the relationship that has been simulated and the rates in each row represent the classifications.

3.3. Expected number of relatives

For various reasons, estimates of the number of expected relatives for different relationship classes are important. Such numbers will obviously depend on the number of children born per family, a number which in turn will vary both between families and also between different generations. We performed simulations in which we applied a model for the variation of the number of children born per family (outlined in Material and methods).

The estimated number of relatives, for different degrees of cousins, is illustrated in Fig. 8 and Supplementary Table 1. As expected, the number of relatives increases exponentially with the number of children per family and generation. The variation (measured as the coefficient of variation, CV) in the number relatives, caused by allowing the number of children per family to vary, was in the same range for the different assumption of children per family but decreased as the degree of relationship decreased. As an example, assuming a mean of three children per family and generation, on average any given individual will have 93 2nd cousins (S3-3) with a 95% interval of such figure lying between 33 and 152 (Supplementary Table 1).

Furthermore, the number of relatives is generally lower for an individual born in 2005 compared with an individual born in 1955. Our simulations show that, for example, the number of 3rd cousins (S4-4) for an individual born in 2005 is around 200 while for an individual born in 1955 the number of 3rd cousins are on average 4 times higher.

4. Discussion

In a number of applications we need to establish the biological relationship between two (or more) individuals. The focus in the forensic field has historically been on paternity cases but recently also involved other topics like for instance kinship testing in immigration case for family reunifications [70,71] and familial searching aiming to find potential donors of crime scene samples [14,15,17,19,21–23]. The current practice is primarily based on a limited set of carefully chosen short tandem repeat (STR) markers with the power to resolve most
However, progress and recent developments in DNA typing technologies have provided access to whole genome data, even from small amounts of DNA, such as crime stain samples or ancient human remains [7,8,75,76]. The availability and awareness to large genotype sets have further spurred the demand for bioinformatics, with a particular focus on the classification of biological relationship between individuals. In essence, classification is the ability to assign a given relationship to its true class. The objective of this study was to evaluate the currently established likelihood approach in forensics [30] where hypotheses are clearly stated with identical by state (IBS) methods presented by other researchers and practitioners [27,39,58]. In particular we compared the classification power for each method separately, and the power of combining different classifiers to improve classification rates for different relationship classes. Each bar represents the results from 1000 classifications for each relationship class and each classification method. When a single method is used, the false classification rate is simply given as 1 minus the true classification rate.

### Table 8
Classification rates using the combination of all four classifiers (Likelihood, Segment, KING (Pr(IBD = 0) and kinship coefficient)).

<table>
<thead>
<tr>
<th>True relationship</th>
<th>S1-1</th>
<th>S2-2</th>
<th>S3-3</th>
<th>S4-4</th>
<th>Unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S2-2</td>
<td>0</td>
<td>0.997</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S3-3</td>
<td>0</td>
<td>0</td>
<td>0.755</td>
<td>0.034</td>
<td>0.001</td>
</tr>
<tr>
<td>S4-4</td>
<td>0</td>
<td>0</td>
<td>0.005</td>
<td>0.153</td>
<td>0.003</td>
</tr>
<tr>
<td>Unrelated</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.042</td>
<td>0.698</td>
</tr>
<tr>
<td>Undefined</td>
<td>0</td>
<td>0.003</td>
<td>0.24</td>
<td>0.771</td>
<td>0.298</td>
</tr>
</tbody>
</table>

Fig. 6. True classification rates for different relationship classes. Each bar represents the results from 1000 classifications for each relationship class and each classification method. When a single method is used, the false classification rate is simply given as 1 minus the true classification rate.

Fig. 7. Classification rates (Right: True and Left: False) using different combinations of the classification methods described in the text. Each bar represents the results from 1000 classifications for each relationship class and each classification method. Average – The average degree of relationship assigned from all four classifiers. At least two – Classify if at least two classifiers points to the same relationship (and the other two either assigns different relationships or identical as the first two does). At least three – Classify if at least three classifiers points to the same relationship. All four – classify only if all four classifiers points to the same relationship. For comparison, classification based on the Segment approach alone has also been added.
classification.

Recently, attention has been drawn to this topic through the successful application of dense genetic marker panels and public genealogy databases to trace the unknown donor in a number of highly profiled crime cases [2,6,77]. However, concerns have been raised relating to the premature use of this approach [9–11,78,79]. In forensics, there is generally a high demand before scientific methods are implemented in routine case work. Important characteristics include scientific foundation, high accuracy and thoroughly covered validation studies [80]. Having said that, classifications, as described in this study will only be provided as investigative leads to the relevant authorities and never taken to court. This entails lower demands in terms of false classifications, i.e. false leads, but still requires a fundamental understanding of the methods used to perform such classifications.

In this study we restricted the classification to some overarching relationship classes (Full siblings, 1st cousins, 2nd cousins, 3rd cousins and Unrelated). Although practical applications will most certainly involve other classes, for instance 1st cousins once removed; for an exploratory study like ours, it is far more important to analyze the fundamental properties of the different approaches. We argue that the conclusions drawn from this study are applicable also for more distant relationships as well as intermediary relationship classes in terms of performance of the different approaches.

When each individual classifier was used, the Likelihood approach performed best in terms of classifying relatives, whereas the Segment approach performed best when classifying unrelated individuals (see Fig. 6). Somewhat counterintuitive is the greater performance of the Likelihood approach using far less of the available DNA markers. The answer is probably at least two-fold; first all markers contain some degree of redundant information, i.e. knowing the state of one marker reveals some information about adjacent marker(s). Secondly, markers with low minor allele frequency are expected to, on average, contribute with little information about relatedness. Our careful pruning procedure filters markers implicitly on these two criteria thus leaving a reduced subset with maximum information. Furthermore, the Likelihood approach incorporates information about rare shared variants. There are theoretical scenarios that may be conceived where the Likelihood approach is the only method able of identifying relatives. For instance, if a single, extremely rare, variant is shared at one marker, the conditional probability that two unrelated individuals share this variant is exceedingly small. In turn, the likelihood approach will provide strong evidence towards some degree of relatedness whereas the other approaches, relying on average sharing across the entire genome; will point in the direction of unrelated.

Population genetic properties such as substructure, differences in allele frequency distributions and consanguinity will have an impact both in relation to the thresholds used to compute the metrics described in this study, but also to determine appropriate cutoffs for the classifiers. As we have pointed out previously, the Likelihood approach is more sensitive to population specific factors [51], which in turn may lead to a higher degree of false classifications unless these factors are accounted for. For the particular SNP markers described in this study there is, however, an abundance of reference data through projects like the 1000 Genomes project [43,44]. For populations with an increased level of consanguinity, for instance isolated communities, the classification ranges have to be adjusted accordingly as we expect a higher degree of so called background IBD patterns.

A topic not discussed so far is genotyping errors and their potential impact on the analyses (see for instance Pompeanon et al. [81] for a review and Bilton et al. [82] for a more recent application). As with other forensic DNA typing methods, there is a high demand on the quality of the genotyping (i.e. results from the analyses performed at the lab). Genotyping methods are constantly improving and the error rate is generally low, given that sample quality is not compromised [83,84]. The latter may well be the case if the application is massive genotyping of biological traces. The Segment approach would in theory be most vulnerable to genotyping errors since an occurrence could potentially terminate a segment prematurely. In contrast, the other approaches are less susceptible to genotyping errors since they rely on average whole genome sharing and not stretches of uninterrupted shared DNA. To mitigate this, most direct-to-consumer companies, relying on the Segment approach, implements a less stringent search where errors are allowed to a certain degree [58].

This study has explored the possibilities and limitations of three fundamentally different approaches to infer the degree of relationship for a pair of individuals. We have demonstrated that the methods perform well, at least for the relationships below 5th degree (S4-4). Even so, there are a number of improvements worthy of further investigation. Firstly, we showed that combining the outcomes from multiple approaches reduced the false classification rates, however, at the cost of considerably decreasing the true classification rate. Secondly, instead of reporting an exact degree of relationship, ranges of relationships could be reported (for instance, instead of classes S2-2, S3-3 etc, have classes spanning multiple relationships like S2-2 to S4-4). This will increase the true classification rate but with the trade-off of a less precise relationship classification, see Supplementary Fig. 7. Thirdly, as mentioned in the Methods section, the Segment approach has already been subject to several suggested extensions [58,61].
Furthermore, we can explore the distribution of individual shared segments instead of the total length of shared genomic segments, see Supplementary Fig. 8, potentially containing an extra depth of information. Fourthly, the Likelihood approach could be further tuned with respect to incorporating more markers and potentially implement threshold on the likelihood when classifying a pair of individuals as relatives. To select the most informative markers and to exclude markers with redundant information, we performed a marker pruning based on minor allele frequency, marker distance and degree of linkage disequilibrium between adjacent markers. We used thresholds from previous studies where the sensitivity with regards to inferring relationships has been thoroughly studied [28,29,51].

Finally, in terms of deciding on the best classifier, costs may be associated with the different outcomes. For example, it might be worse to misclassify a pair of unrelated individuals as S2-2 compared to the opposite (S2-2 as unrelated). As an example it would be possible to establish a threshold representing a theoretical or simulated maximum value for unrelated pairs in order to avoid such misclassifications. This could be added to the classification model and, in combination with the expected number of relatives, be used to further tune the search parameters and minimize the total overall cost using a mathematical framework as described by Tillmar et al. [42].

5. Concluding remarks

In light of recent event relating to finding the unknown donor of a biological stain or a missing person, public genealogical databases has emerged as a tool in the investigations. This study revealed that current methods, mostly relying on versions of the so called Segment approach whereby shared chromosomal segments are counted, have some desirable properties. However, the results also revealed that the traditional approach in forensic genetics, namely the likelihood ratio, to measure the weight of evidence, performs well and for some instances even better than the segment approach. We show that the combined use of multiple approaches decreased the false classification rate.

It is crucial that forensic practitioners are aware of the progress and possess a fundamental understanding of the behavior and limitations of the statistical approaches if they are to assist in future endeavors related to forensic genealogy.

Acknowledgement

The authors gratefully thank two anonymous reviewers for their constructive criticism which improved the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2019.06.019.

References


[19] D.J. Balding, M. Kruijver, K. Slooten, J.C. Guerrini, A. McGuire, S. Sankararaman, J. Volz, et al., Finding the unknown donor of a biological stain or a missing person, public genealogical databases has emerged as a tool in the investigations. This study revealed that current methods, mostly relying on versions of the so called Segment approach whereby shared chromosomal segments are counted, have some desirable properties. However, the results also revealed that the traditional approach in forensic genetics, namely the likelihood ratio, to measure the weight of evidence, performs well and for some instances even better than the segment approach. We show that the combined use of multiple approaches decreased the false classification rate.

It is crucial that forensic practitioners are aware of the progress and possess a fundamental understanding of the behavior and limitations of the statistical approaches if they are to assist in future endeavors related to forensic genealogy.

Acknowledgement

The authors gratefully thank two anonymous reviewers for their constructive criticism which improved the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2019.06.019.

References


