Parameters and Pedigrees in Forensic Genetics and Statistics

Parametre og pedigreer i rettsgenetikk og statistikk

Philosophiae Doctor (PhD) Thesis

Navreet Kaur

Faculty of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences

Ås 2017

Norwegian University
of Life Sciences

Thesis number 2017:46

ISSN 1894-6402

ISBN 978-82-575-1447-1

Summary

With the rapidly evolving DNA technology today, there is a constant need of more robust statistical methods for analyzing the data. The sequencing techniques are improving, making more genetic markers available, and we are able to analyze even smaller samples of degraded DNA gathered from crime scenes. Many of the traditional and commonly used statistical methods need therefore to be updated.

When a crime is committed and a suspect is found, two competing hypotheses are generally presented. The custom in forensic statistics has for long been to present competing hypotheses verbally. The prosecutor may suggest the hypothesis H_p: "the suspect contributed to the stain", whereas the defense attorney may suggest the hypothesis H_d: "an unrelated person contributed to the stain". However, giving a more statistical presentation of the problem can be beneficial as the statistical tools used to test the hypotheses then can be more sophisticated. In particular, by giving the problem a *parametric* form we are able to present the problem in a more conventional statistical framework. Using parametric models makes it possible to apply already well-known mathematical and statistical models for evaluating the hypotheses, and we are able to get an alternative understanding of the problem. For instance, when making kinship inference, a parametric formulation of the problem facilitates more generalized alternative hypotheses, and we no longer need to test a specific relation versus unrelatedness: the alternative can be any other relation.

This thesis aims at describing different parametric approaches for forensic applications. The thesis considers both pure kinship cases and forensic crime cases, and cases where these two subfields of forensics overlap. We deal with complex DNA mixture problems and present methods for identifying the contributors to the mixture. We also study kinship cases where mixtures appear, and suggest methods for determining the relation between the mixture contributors. Methods for relationship inference based on statistical estimation of the parameters is also presented, and we make use of statistical theory that deserve attention in a forensic framework.

Sammendrag

DNA teknologien utvikler seg i en stor fart, og med dette tempoet trengs det stadig nye og mer robuste statistiske metoder for å analysere data. Sekvenseringsteknikkene bedres også og fører til at vi i dag har langt flere genetiske markører tilgjengelig. Med teknologien tilgjengelig i dag kan man analysere selv mindre mengder med degradert DNA i kriminalsaker. Vi trenger derfor nye og bedre tilpassede statistiske metoder.

Etter at en kriminell handling har funnet sted presenteres det ofte to hypoteser. I rettsgenetisk statistikk har det i lang tid vært vanlig å presentere slike hypoteser verbalt. Aktor kan for eksempel foreslå hypotesen H_p: "mistenkte bidro til DNA-sporet", mens forsvaret har følgende hypotese H_d: "en urelatert person bidro til DNA-sporet". En tradisjonell matematisk statistisk formulering av problemet kan være fordelaktig. Mer spesifikt vil en *parametrisk* tilnærming åpne for at vi kan bruke velkjente matematiske og statistiske metoder for å teste hypotesene. Dette vil også gi oss en alternativ forståelse av problemet. I slektskapsanalyser vil for eksempel en parametrisk fremstilling gi oss muligheten til å gi mer generelle alternative hypoteser i den forstand at vi ikke lenger trenger å teste en spesifikk relasjon versus ubeslektet: den alternative hypotesen kan være generell.

Denne avhandlingen har som mål å beskrive slike parametriske metoder innen rettsgenetikk og statistikk. Avhandlingen tar for seg både rene slektskapssaker og kriminalsaker, samt saker dere disse to feltene innen rettsgenetikk overlapper. Vi tar opp problemer med komplekse DNA blandinger og presenterer metoder for å identifisere bidragsyterne til blandingen. Vi ser også nærmere på slektskapssaker der DNA blandinger inngår, og studerer metoder for å bestemme familierelasjonen mellom bidragsyterne. Metoder for slektskapsidentifisering basert på statistisk estimering av parametere presenteres også, og vi tar i bruk statistisk teori som fortjener oppmerksomhet i en rettsgenetisk sammenheng.

List of papers

I. N. Kaur, A.E. Fonneløp, and T. Egeland, *Regression models for DNA-mixtures.* Forensic Science International: Genetics 11 (2014): 105-110.

II. N. Kaur, M. M. Bouzga, G. Dørum, and T. Egeland, *Relationship inference based on DNA mixtures.* International Journal of Legal Medicine 130.2 (2016): 323-329.

III. G. Dørum, N. Kaur, M. Gysi, *Pedigree based relationship inference from complex DNA mixtures*. International Journal of Legal Medicine (2017): 1-13.

IV. N. Kaur, M.D. Vigeland, G. Storvik, T. Egeland, *Relationship inference: Estimation and Model Selection*. Manuscript

Acknowledgements

Without the support and guidance of my supervisor, professor Thore Egeland, the work presented in this thesis could not reach its completion. Thore, your deep knowledge and experience in the field of forensics and statistics, combined with your passion for sharing has been a true gift for me over the last years. Thank you for taking your time to guide me and to persuade me to learn different aspects of forensics, and for your genuine thoughtfulness and patience when I needed time to digest the knowledge you so kindly shared with me.

A warm gratitude to my co-supervisor Guro Dørum. Guro, we started out as two unfamiliar statisticians, but our travels brought us together. We have walked miles after miles together, and even cycled in stilettos to a gala dinner. Both of us like to make our own way, and it has been a great pleasure to walk this path with you by my side. Thank you for not only supervising me, but for also being my friend.

Geir O. Storvik and Magnus D. Vigeland, thank you for joining my team and giving both me and our work together a different perspective. My gratitude to all you hardworking souls at the forensic institute in Oslo, for always opening your doors whenever i needed biological input and help.

A special thanks goes to you, Trygve Almøy, for being there for me whether I had a statistical question or a personal problem. You are a true mentor and friend. No one explains type I and II errors better than you. Thank you, Are Aastveit, for unknowingly asking the question I at that time was afraid to answer, but that followed me throughout these years; do you really want this PhD? I think the answer is clear know.

My family. My mother and father for unconditionally supporting me, even in times you did not agree in my decisions. My sisters, Diddi and Puneet, for showing your little sister not to just go with the flow, but to find your own way and to follow your dreams. My dear DJ, for always being there for me and the rest of the family. My best friend and niece, Nena Alina, for being yourself and cheering my day with your great smile and laughter.

Thank you Amar, for showing me the importance of focus both in happy and troubled times, and for reminding me of the power of staying calm. For that I will always be grateful. And to you Anahita, for being a friend I can always call, and for being by my side, even in times I didn't know you where there.

Navreet Kaur

Carrots! What would my PhD life been without you? Chris, Walther, David, Guro, Athena and Theresa. You are all awesome! Thank you for the ISFG conferences and Euroforgen meetings, for your laughter, for John B. and of course, for carrots. Cheers to all our great memories, and for many, many more to be made!

Finally, a big thank you to everyone at Biostatistics @ NMBU. For always letting me have my party hat on, and for letting Guro, Hilde and me carry out all our insane ideas; Bollywood dancing, arranging the cowboy party of the century, ice-fishing, paintball, curling, food festivals and all the other crazy adventures we have been on together. Did anyone say statisticians are boring? Biostatistics @ NMBU, we rock!

Ås, April 2017
Navreet Kaur

Contents

1 Introduction

The kinship part of this thesis is motivated by cases where the family relationship between individuals is questioned and evaluated using DNA evidence. Searching for family roots and getting to know ones ancestral heritage is for many individuals important for identity purposes. With the diversity we see in different public groups, kinship analyses are not that straightforward, and we need to consider several aspects while reconstructing the family pedigrees. The population may for instance be subject to inbreeding as individuals may choose to mate with individuals of the same origin [33]. Traditionally in paternity testing, the hypothesis stating that a man is the biological father of a child is compared to the alternative hypothesis that the alleged father is unrelated. This alternative of unrelatedness may be too restrictive, and the parametric approach of this thesis allows for more general alternatives. Similar problems appear in other contexts like disaster victim identification. Again, the conventional formulations of the problem may limit the evaluation approaches, and more alternatives should be considered. If we turn towards forensic casework based on DNA mixtures, family relationship between the contributors (those implicated in the case as perpetrators or victims) may not easily be accounted for using existing methods and implementations. Methods and a freely available implementation for handling such cases (the `R` package `relMix`) are presented in this thesis.

The DNA technology has had an enormous progress over the last years [11], and the advances have far ranging implications including cold cases being reopened and solved. We are able to create DNA profiles using tiny amounts of often degraded samples, and the profiling is just a step towards solving the case. There are, however, some commonly known challenges in forensic casework and kinship testing that we always need to consider, even with the improved technology. Artifacts like dropout (a common problem for low template DNA samples), drop-in, silent alleles, mutations and population stratification are some examples that we need to address in connection with the statistical analysis. Commonly used statistical methods today do include such artifacts, however, there is no doubt that we need to develop the statistical methods according to the evolving DNA technology. In this thesis we propose a different

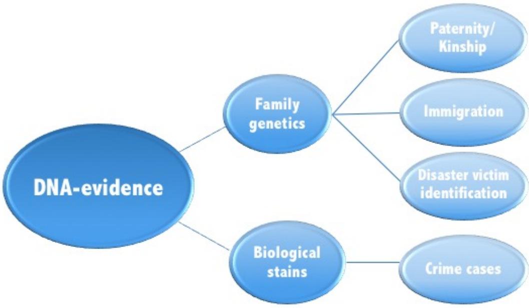


Figure 1: The figure shows how the field of forensics is divided when DNA-evidence is found. On one side we have family genetics, including kinship cases, immigration cases and disaster victim identification/missing person identification, while on the other side we find forensic crime cases based on mixtures.

perspective for solving kinship and crime cases, namely through statistical parametrization.

Figure 1 gives an overview of how the field of forensics often is divided. The problems met in this thesis will touch both family genetics and general forensic crime cases. We are in other words working in a cross-over between these two forensic fields. As an example, for papers II and III of this thesis, we could draw a line from the subfield of "Biological stains" to "Paternity/Kinship" and "Disaster victim identification" in Figure 1. The parametric approaches we present in the papers of this thesis rely on many well-known statistical theories that deserve more attention and should be explored further for forensic applications.

To understand how forensic casework is solved, we need a proper understanding of forensic DNA profiling. Some basic biological and statistical background is therefore required, and in the following sections we aim at guiding the reader through some of the biological and statistical concepts used in the papers included in this thesis.

1.1 Genetic background

1.1.1 The DNA: chromosome, genes and alleles

DNA is today associated with the well known "double helix" as discovered by Francis Crick, James Watson and Rosalind Franklin in 1953 [40]. But where is the DNA found? We use Figure 2 to give an illustration of some of our essential building blocks. The figure shows a random cell of an individual. The DNA is found in the nucleus of the cell and consist of about $3 \cdot 10^9$ base pairs, packed into chromosomes. The human DNA consists of 23 pairs of chromosomes, where 22 of these are autosomal pairs, and the last pair is known as the sex chromosome (denoted XY for males and XX for females). If we imagine that we pull out the DNA strands making up the chromosomes, the strands turn out to be twisted double helical structures. A closer inspection here shows that each DNA strand consists of the letters A (adenine), T (thymine), C (cytosine), and G (guanine), known as bases. These are the building blocks of our genes; the basic units of inheritance, storing our genetic code. Only a small fraction of the DNA strands are coding regions with genes. The major part is noncoding. The chromosomes of a pair are inherited one from each parent. A specific location in the chromosome is called a genetic marker or a locus. Loci that show variation between individuals are chosen as genetic markers to differentiate between individuals. Most of the forensic markers are positioned in the none coding regions of the chromosomes. A variant of a specific marker is called an allele.

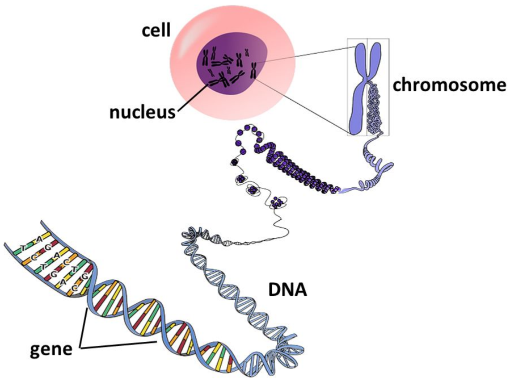


Figure 2: Essential building blocks.

1.1.2 Genetic markers - STR and SNP

The advances in forensic DNA profiling is without doubt highly related to the development in use of genetic markers. Triggs et al. [57] describe three major stages of technological advancement for finding genetic markers, namely the multilocus, single-locus and STR stages. Short tandem repeat (STR) markers are most commonly used in forensic casework today, and are a subclass of VNTR (variable number of tandem repeats) markers. STRs were introduced for investigatory purposes around 20 years ago, and are constantly subject to progressive development [30]. The characteristic of STR markers is that they consist of repeated units of short sequences, usually between 2 and 6 base pairs. In other words, such markers consist of short DNA sequences, like for instance "ACGA", which are repeated a specific number of times. The sequence "ACGA" is called the motif, and it is the number of times the motif is repeated that designates the allele name. If the motif "ACGA" is repeated, say, 16 times, this gives us the allele name "16".

The main advantage of STR markers is that they are highly polymorphic. Intuitively, a high variation in the alleles is desirable for human identification. The resulting DNA profile is often regarded as identifying. Forensic trace samples is frequently of poor quality with low DNA levels, often degraded, and may be found as mixture profiles of two or more individuals. It is therefore of importance to use markers that can be amplified regardless of poor quality, and STRs are considered to be easy to amplify using polymerase chain reaction (PCR), even in small quantity. Even though shorter markers (see SNPs below) perform better on degraded DNA, STRs are still the major tool even for analyzing degraded trace samples.

Other classes of genetic markers are also proving to be very useful in forensic casework. Single nucleotide polymorphic (SNP) markers is one such class of markers. SNPs are differences in one base occurring at single positions in the DNA, and can be described as short binary markers. These markers present most of the common human genomic variation. However, as SNPs are biallelic markers, these are not as informative as STR markers per locus. As an example, Tillmar et al. [55] show that 52 SNPs are as informative as 11 STR markers in a kinship case testing for paternity versus an uncle-nephew relation. Still, SNP markers have desirable properties that are of interest in forensic use; they are theoretically more resistant to degradation since a smaller target region is needed to recover information from DNA. They are are also more reluctant to mutations, and may therefore be considered more stable for kinship testing. Our first paper make use of SNP markers, where the Illumina GoldenGate(R) 360 SNP test panel is used. This panel is hardly used for forensic problems, however, the main focus of the mentioned paper and our thesis is on the statistical methods and applications.

The use of SNPs over STR markers in forensic applications has been a topic of discussion over the recent years, and is discussed in papers like [12], [9] and [47]. STR makers have a solid scientific foundation [11], and it is most unlikely that SNPs will replace STR makers fully. SNPs are today an important supplement to STR markers.

1.1.3 Mendel, inheritance and pedigrees

Gregor Mendel established several rules of inheritance in the mid 1800s, and his work revolutionized the science of genetics. After breeding various pea plants and establishing pure breeding lines, he cross-bred the pea lines and followed the result of their outcome for some generations. He observed that the traits followed a specific pattern, as illustrated in figure 3, where yellow and green peas are cross-bred. The first generation gave pure yellow peas, indicating that yellow was *dominant*. However, in the following generation the *recessive* green peas reappeared, and the overall ratio of dominant to recessive trait was found to be 3:1 in his studies. The paper [19] explains Mendelian inheritance and its forensic relevance using simple urn models.

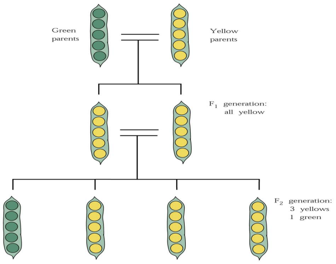

Green parents
Yellow parents
F_1 generation: all yellow
F_2 generation: 3 yellows 1 green

Figure 3: An illustration of Mendel's pea plant experiment, see [63]. Note that we already here have a family pedigree, as we discuss further in Figure 4.

Inheritance may be defined as a trait transferred genetically. Mendel's second law states that alleles for separate traits are passed on independently of one another from parents to offspring. This law was later proven wrong, and Thomas H. Morgan (awarded the Nobel Prize in Physiology and Medicine 1933) and others demonstrated that genes are carried on chromosomes. The unit of the distance between the genes is Morgan, or the more commonly used centi Morgan (cM).

Genetic linkage occurs when there is dependence in the inheritance pattern in a pedigree, i.e. alleles at different loci are not transmitted independently through the pedigree. This thesis will not concern linkage analysis, and the interested reader is referred to [53] and [54]. We mention Mendel's experiment here as his work also has great impact on general pedigree analysis and inheritance. In figure 4 we see two different family pedigrees. Generally in pedigrees, females are presented by circles and the males are presented by squares. The pedigree to the right shows a first-cousin mating (between individuals (5) and (8)), denoted by a double line, and we say that the son (9) is *inbred*. Figure 4 is made using the `R` library `paramlink`, see [23]. In human genetics, several additional symbols are used. We have symbols denoting individuals affected by a disease, individuals who are dead, individuals who are carriers etc. See Ziegler et al. [64] for a complete list of plotting symbols.

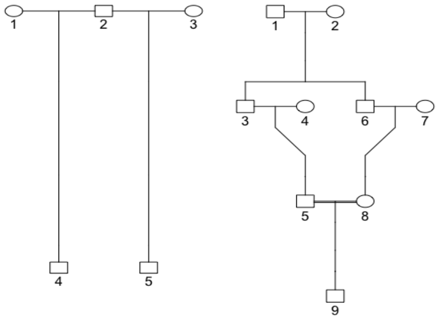


Figure 4: Two pedigrees showing two different families. The pedigree to the left displays two maternal half-brothers (individuals (4) and (5)), while the pedigree to the right involves inbreeding.

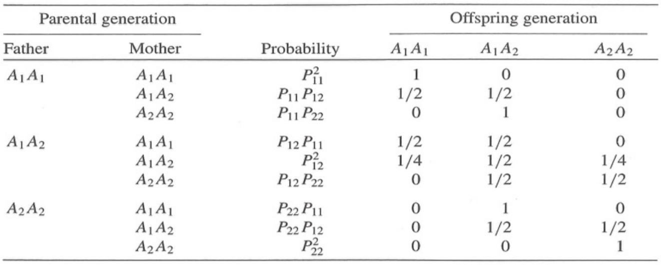
1.1.4 Population genetics

Population genetics concerns the study of genetic variation within populations and between. It is a broad subfield of genetics, and we will in the following paragraphs present the most essential population genetic effects that are needed to understand this thesis. There is a large literature on population genetics focusing on the examination and modeling of variation in the frequencies of alleles within and between populations, over space and time. From a forensic point of view [4] and [26] provides a relevant introduction.

Hardy-Weinberg equilibrium A population is said to be in Hardy-Weinberg equilibrium (HWE) if the two alleles at a particular locus are statistically independent of each other. In other words, what allele we inherit from one parent is independent of what we inherit from the other at a particular locus in HWE. More commonly we say that the allele and genotype frequencies remain constant over generations in the population. There are five underlying assumptions for HWE as described in [64], namely random mating, no selection or migration, no mutation, no population stratification (see next paragraph), and infinite population size.

Due to independence between the alleles, statistical calculations will be simplified if a population is in HWE. From a practical point of view, it is sufficient to estimate allele frequencies as genotype frequencies can be derived when HWE applies. Fung et al. [26] explains in detail the steps for finding the genotype frequencies under HWE conditions. Assume we have an autosomal locus with two alleles, A_1 and A_2. Then there are three possible genotypes, given by A_1/A_1 (sometimes also denoted A_1A_1), A_1/A_2 and A_2/A_2, with corresponding genotype proportions P_{11}, P_{12} and P_{22}. The allele frequencies for A_1 and A_2 is then given by $p_1 = P_{11} + P_{12}/2$ and $p_2 = P_{22} + P_{12}/2$. Further, we have that genotype frequencies of the offsprings of the second generation will be given by p_1^2 for homozygotes (i.e. A_1/A_1), $2p_1p_2$ for heterozygotes (A_1/A_2), and p_2^2 for homozygotes (A_2/A_2) . Figure 5 shows the possible outcomes of a standard mother-father-child trio from [26].

Population substructure (θ-correction) To account for population stratification and relatedness, the θ parameter is commonly used. In paternity cases for instance, Hardy-Weinberg will not apply in cases where the parents are related in a way not specified by the pedigree. By including the θ parameter, we essentially correct for relatedness of alleles with common ancestry. Consider an allele A_1 with frequency p_{A_1} and assume that we have sampled n alleles, where x of these alleles are of type A_1. With the



Parental generation			Offspring generation		
Father	Mother	Probability	A_1A_1	A_1A_2	A_2A_2
A_1A_1	A_1A_1	P_{11}^2	1	0	0
	A_1A_2	$P_{11}P_{12}$	1/2	1/2	0
	A_2A_2	$P_{11}P_{22}$	0	1	0
A_1A_2	A_1A_1	$P_{12}P_{11}$	1/2	1/2	0
	A_1A_2	P_{12}^2	1/4	1/2	1/4
	A_2A_2	$P_{12}P_{22}$	0	1/2	1/2
A_2A_2	A_1A_1	$P_{22}P_{11}$	0	1	0
	A_1A_2	$P_{22}P_{12}$	0	1/2	1/2
	A_2A_2	P_{22}^2	0	0	1

Figure 5: Table from Fung et al. [26] giving outcomes of random mating in an infinite population.

coancestry coefficient θ, the probability that the next allele will be of type A_1 is given by

$$\frac{x\theta + (1-\theta)p_{A_1}}{1+(n-1)\theta}.$$

See [4] for further details. The paper [7] gives estimates of θ for a wide range of populations.

IBD and IBS *Identical-by-descent* and *identical-by-state* are two related concepts that are important to have in mind while reconstructing pedigrees and family relations. Figure 6 gives an illustration of the concept. As explained in [24], an allele in one individual is said to be identical by descent to an allele in another individual if it derives from the same ancestral allele within the specified pedigree. In figure 6, individuals 3 and 4 are brothers. We say that 3 and 4 share two alleles IBD if both alleles in each brother derive from the same ancestral alleles (as they do in the first marker), they share one allele IBD if only one allele is derived from the same ancestral allele (illustrated in the second marker), and they share zero alleles IBD if none of the alleles derive from the same ancestral allele (third marker). Identical by state (IBS) on the other hand refers to allele sharing (identical alleles) and does not require the shared allele to derive from the same ancestor. For the brothers in figure 6, assume the parents are not genotyped. Then the IBD status is no longer known. The three markers now correspond to IBS being 2, 1 and 0.

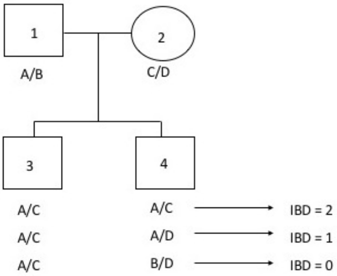


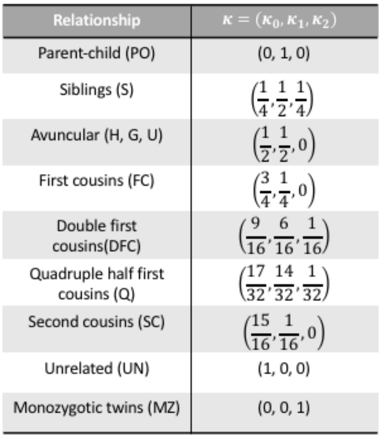
Figure 6: Figure illustrating the concept of identical-by-descent (IBD).

The κ parameter and the relationship triangle The concept of IBD can be used to identify specific non-inbred pairwise relationships, by means of the κ parameters, given by the vector $\kappa = (\kappa_0, \kappa_1, \kappa_2)$. Inbred relations would require nine parameters as explained in Jacquard (see [24], [27]). In the vector κ, κ_i is the probability that two individuals share 0, 1 or 2 alleles IBD, hence $i = 0, 1, 2$. We have that $\sum_{i=0}^{2} \kappa_i = 1$. The most common relationships in terms of κ parameters are given in the table on the left-hand side of figure 7. It is explained in [52] that we have the restriction that $\kappa_1^2 \geq 4\kappa_0(1 - \kappa_0 - \kappa_1)$, hence the valid area for our κ parameters is the white area beneath the dashed line illustrated in the plot on the right-hand side of figure 7. In other words, we have that pairwise relations can be described by the two-dimensional space given by

$$K^* = \{(\kappa_0, \kappa_2) : \kappa_0, \kappa_2 \in [0,1], \kappa_1^2 \geq 4\kappa_0(1 - \kappa_0 - \kappa_1)\} \tag{1}$$

See section 1.2.1 for an example on calculating the likelihood for a pairwise relation based on κ parameters.

Coefficient of kinship and inbreeding Studies on how generations are affected by mating between related individuals have for many years been a topic of discussion both in human genetics and in population structure studies [33, 61, 62]. The kinship coefficient between a pair of individuals is of particular interest in this area as human geneticists often measure relationships through this numerical value. The coefficient of kinship ψ between two individuals A and B measures the proportion of IBD alleles,



Relationship	$\kappa = (\kappa_0, \kappa_1, \kappa_2)$
Parent-child (PO)	(0, 1, 0)
Siblings (S)	$\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$
Avuncular (H, G, U)	$\left(\frac{1}{2}, \frac{1}{2}, 0\right)$
First cousins (FC)	$\left(\frac{3}{4}, \frac{1}{4}, 0\right)$
Double first cousins(DFC)	$\left(\frac{9}{16}, \frac{6}{16}, \frac{1}{16}\right)$
Quadruple half first cousins (Q)	$\left(\frac{17}{32}, \frac{14}{32}, \frac{1}{32}\right)$
Second cousins (SC)	$\left(\frac{15}{16}, \frac{1}{16}, 0\right)$
Unrelated (UN)	(1, 0, 0)
Monozygotic twins (MZ)	(0, 0, 1)

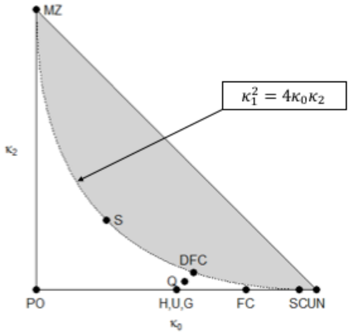


Figure 7: The table to the left shows some well-known pairwise relationships given in terms of κ parameters, while the figure to the right gives an illustration of these relations. The figure is plotted using the function `IBDtriangle` of the `R` package `paramlink` [23]. The valid domain for the κ parameters is the white area under the curve given by $\kappa_1^2 = 4\kappa_0\kappa_2$. Note that the term avuncular encompasses the three relations halfsiblings, grandparent-grandchild and uncle/aunt - niece/nephew.

and is the probability that a randomly chosen allele in A is IBD to a randomly chosen allele from B. For non-inbred individuals the parameter is

$$\psi = \frac{2\kappa_2 + \kappa_1}{4}.$$

This coefficient is also of interest as we operate with one single value and summarize pairwise relationships through one single parameter, compared to the two-dimensional setting we have using the three κ_i parameters presented in the previous paragraph. However, this parameter reduction is not always beneficial, as some relations no longer are distinguishable using ψ. For instance, using the κ_i values given in figure 7, we find $\psi = \frac{2\cdot 0+1}{4} = \frac{1}{4}$ for the parent-child relation. For siblings, we also find $\psi = \frac{2\cdot\frac{1}{4}+\frac{1}{2}}{4} = \frac{1}{4}$. Although these relations are located far from each other as is evident from the plot in Figure 7 (see PO and S), they are presented with the same value using the kinship coefficient.

The coefficient of kinship ψ and the *inbreeding* coefficient, f, are two related con-

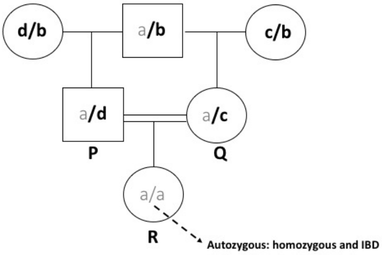


Figure 8: Figure used to explain the relationship between the kinship coefficient ψ and the inbreeding coefficient f. Recall that the double line between individuals P and Q indicates that their child R is inbred.

cepts, as explained in [54]. We use Figure 8 to explain the relation between these two concepts. Two individuals P and Q are paternal halfsiblings, and both individuals have inherited an allele (a) from their father. In other words, P and Q have one allele identical by descent, and we denote their coefficient of kinship by $\psi_{P,Q}$. The halfsiblings mate (double line), and the pedigree is expanded by including an individual R. If R inherits the same allele from her father and mother, we say that R is autozygous, i.e., a homozygote individual with alleles that are copies of the identical ancestral gene, as a result of a consanguineous mating. The probability of R being autozygous is the inbreeding coefficient of R, denoted f_R. Hence, the inbreeding coefficient of R and the kinship coefficient of P and Q, assumed to be non-inbred, are exactly the same. The following reasoning, also included in paper IV, explains this relation step by step:

$$\begin{aligned}\psi = \psi_{P,Q} &= Pr(\text{random allele of P is IBD with random allele of Q})\\ &= Pr(\text{R receives IBD alleles from her parents})\\ &= Pr(\text{R is autozygous})\\ &= f_R.\end{aligned}$$

So far we have discussed DNA marker data from individuals, based on genotypes from a reference sample of good quality. We will discuss problems that may occur for degraded DNA (dropout) and artifacts like drop-in, silent alleles, and mutations later. Also, papers II and III of the thesis use mixture DNA profiles and this will be addressed in section 1.4.1. We now turn towards statistical methods more specifically.

1.2 Statistical methods

Below we introduce some fundamental statistical methods and concepts for our applications. Some more standard methods, like multiple linear regression briefly reviewed and used in paper I, are not discussed here.

1.2.1 Likelihoods

Likelihood inference can be understood from different perspectives, and detailed explanations can be found in several basic statistical textbooks, like [49] and [18]. Assume we have independent and identically distributed data, $y_1, y_2, ..., y_n$, following a distribution described by the function $f_y(y; \phi)$. Here ϕ is an unknown parameter that we want to estimate from the data. If we let L be the joint probability distribution function of the observations $y_1, y_2, ..., y_n$, then

$$\begin{aligned} L &= f_{y_1,y_2,...,y_n}(y_1, y_2, ..., y_n; \phi) \\ &= f_y(y_1; \phi) \cdots f_y(y_n; \phi) \\ &= \prod_{i=1}^{n} f_y(y_i; \phi). \end{aligned}$$

We can look at the function L as a function of the *data*, that is, $L = L(y_1, ..., y_n; \phi)$. From this perspective, the parameter ϕ is a fixed value and the dataset of y_i's are considered as variables. However, in order to *estimate* unknown parameters from a set of data, it is beneficial to rather look at L as *a function of the parameter* ϕ and consider the y_i's as fixed. The function L is then presented by

$$L = L(\phi) = L(\phi; y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f_y(y_i; \phi),$$

and $L(\phi)$ is defined as the *likelihood function*.

The likelihood is found in several applications, presented in different forms. As an example, which will be expanded on in paper IV of the thesis, we go back to the context of the κ parameter presented in section 1.1.4.

Example 1. If the genotypes or data of two individuals 1 and 2 are given by g_1 and g_2, respectively, the likelihood function for one marker will be given by

$$L(\kappa) = \kappa_0 \mathrm{UN}(p_{g_1}, p_{g_2}) + (1 - \kappa_0 - \kappa_2)\mathrm{PO}(p_{g_1}, p_{g_2}) + \kappa_2 \mathrm{MZ}(p_{g_1}, p_{g_2}) \qquad (2)$$

Here, *UN*, *PO* and *MZ* are abbreviations of "unrelated","parent offspring", and "monozygotic twins", respectively. We have that *UN* is the probability of the genotype given that the individuals share no alleles IBD, *PO* is the probability of the genotypes given that the individuals share one allele IBD, and *MZ* is the probability of the genotype given that the individuals share two alleles IBD.

Consider two individuals with the genotypes $g_1 = 1/1$ and $g_2 = 1/2$, with corresponding genotype frequencies p_1, p_2. Then *UN* $= p_1^2 \cdot 2p_1p_2$, *PO* $= \frac{1}{2}p_1 \cdot 2p_1p_2$ and *MZ* $= 0$. From the likelihood presented above, we find that likelihood must be given by

$$\begin{aligned} L(\kappa) &= \kappa_0 \times p_1^2 \cdot 2p_1p_2 + (1 - \kappa_0 - \kappa_2) \times \frac{1}{2}p_1 \cdot 2p_1p_2 + \kappa_2 \times 0 \\ &= \kappa_0 \times 2p_1^3p_2 + (1 - \kappa_0 - \kappa_2) \times p_1^2p_2. \end{aligned}$$

For the unrelated case (*UN*), we have that $\kappa = (1, 0, 0)$ and so $L(\kappa) = 2p_1^3p_2$.

1.2.2 Estimation - Maximum Likelihood

We say that the maximum likelihood estimate $\hat{\phi}$ is the value of ϕ that maximizes the likelihood function, $L(\phi)$. That is, for any $\hat{\phi}$ where

$$L(\hat{\phi}) \geq L(\phi), \text{for all } \phi \neq \hat{\phi},$$

$\hat{\phi}$ is said to be the maximum likelihood estimator of ϕ.

1.2.3 Optimization

Optimization procedures differ from application to application. In this thesis, optimization is met in paper IV in the context of κ parameters and the relationship triangle in Figure 7. In our application we want to estimate the κ parameters in order to estimate relations between individuals.

With n independent markers, we have that the *log* likelihood function is given by

$$l(\kappa) = \sum_{i=1}^{n} \log(L_i(\kappa)),$$

where $L_i(\kappa)$ is given in (2) and $(\kappa_0, \kappa_2) \in K^*$ as in (1). The problem is that we are working with non-linear constraints. To get hold of the problem, we first *reparametrize* using

$$\alpha = \frac{\kappa_0 \kappa_2}{(1-\kappa_0-\kappa_2)^2} \leq \frac{1}{4}.$$

This gives

$$\kappa_2 = 1 - \kappa_0 - \frac{\sqrt{\kappa_0^2 + 4\alpha\kappa_0(1-\kappa_0)} - \kappa_0}{2\alpha} \tag{3}$$

By use of α, the point (κ_0, κ_2) is transformed to the point (κ_0, α), and we solve the problem by optimizing over $(\kappa_0, \alpha) \in [0,1] \times [0, 1/4]$, before transforming back to κ_2 using equation (3). The standard maximum likelihood theory, involving asymptotic normality and optimality of estimators, does not apply when the parameter is on the boundary as we comment on i paper IV.

1.2.4 Parametric bootstrap

Bootstrapping is a wide area of statistics, and there are several different bootstrapping methods, see [16]. Parametric bootstrapping has been used in this thesis both for creating confidence regions of the estimates. The essential idea is as follows: Given genotype data on two individuals whose relation is in question, an estimate κ^* is obtained from the data. Then the likelihood function (1) is used to generate a table describing the joint genotype probabilities of the two individuals for each marker. This table can then be used to simulate marker data B times from which we get the bootstrap estimates $\hat{\kappa}_1, \ldots, \hat{\kappa}_B$.

There exist several bootstrapping methods for creating confidence intervals or regions as described in [16]. We use the *percentile method* independently for the parameters κ_0 and κ_2 truncated to the interval $[0, 1]$. Note that the problems with parameter values on the boundary mentioned previously for maximum likelihood estimates also apply to bootstrap estimates as discussed in [2]. The confidence ellipses in paper IV ignore the boundary issues and assumes that $(\hat{\kappa_0}, \hat{\kappa_2})$ follows a bivariate normal distribution where the mean vector and covariance matrix is estimated from the bootstrap samples. We have used the implementation in the `R` library `ellipse` which is based on [43].

1.3 Statistics in a forensic context

1.3.1 Likelihood ratio

The likelihood in section 1.2.1 is presented in a mathematical manner including parameters, and by doing so we are able to develop the theory further and include the theory of maximum likelihood to estimate the parameters. In a forensic context, however, the likelihoods usually take a more verbal form, and we also include the hypotheses in question when stating the likelihoods. We say that the likelihood is the probability of the data, conditioned on a given hypothesis (H) and some information I (like allele frequencies) common to all hypotheses, see [6], [25]. We define the likelihood as

$$L = P(\text{data}|H, I).$$

When a crime is committed and DNA samples are gathered at the crime scene, it is of interest to calculate the *weight-of-evidence*. We will in the following denote the DNA evidence by E. It is generally accepted and also recommended that the weight-of-evidence should be summarized by the likelihood ratio (LR). See Neyman et al. [44] for a justification from a statistical point of view, Gjertson et al. [31] for kinship cases (in such cases the LR is sometimes referred to as the paternity index) and Gill et al. [28] for crime cases.

In court, two competing hypotheses stated by the prosecutor (H_p) and the defense attorney (H_d) may in crime cases typically be

H_p : The the person of interest (suspect) contributed to the evidence (E)

H_d : An unrelated man contributed to the evidence (E)

The likelihood ratio (LR) where I is omitted in the notation is then given by

$$LR = \frac{P(\text{data}|H_p)}{P(\text{data}|H_d)} = \frac{P(E|H_p)}{P(E|H_d)}.$$

The likelihood ratio is also applied as weight-of-evidence in kinship cases. It is then usual to rather state the hypotheses as H_1 versus H_2.

1.3.2 Parametric formulations of the hypotheses

A core idea of this thesis is to formulate parametric statistical models and to state the hypotheses in terms of the parameters in the model; this is the standard statistical approach. Paper I presents a crime example, where we use linear regression. The parameter β corresponds to the fraction contributed from the suspect or person of interest (POI). Obviously, the hypothesis "POI did not contribute" is equivalent to $\beta = 0$ and the alternative hypothesis "POI contributed" is equivalent to $\beta > 0$.

For a kinship example, discussed in paper IV, the standard paternity case may be formulated as $\kappa_1 = 1$ ('paternity') versus $\kappa_1 < 1$. This latter alternative is much more general than the verbal 'unrelated'. We use this parametric approach to expand on the case presented in Example 1:

Example 2. Recall the relations in terms of κ parameters given in Figure 7. If we want to test the hypothesis of a parent-child relation (PO) between two individuals versus unrelatedness (UN), we can formulate the hypotheses in terms of κ parameters, where

$$H_1\text{: } \kappa = (0, 1, 0) \text{ versus}$$
$$H_2\text{: } \kappa = (1, 0, 0)$$

We have that the LR for evaluating these parametric hypotheses is formulated by

$$LR = \frac{P(\text{data}|H_1)}{P(\text{data}|H_2)} = \frac{L(\kappa = (0,1,0))}{L(\kappa = (1,0,0))}.$$

If we turn to the the likelihood function found in Example 1: for two individuals with genotypes $g_1 = 1/1$ and $g_2 = 1/2$ we found $L(\kappa) = \kappa_0 2p_1^3 p_2 + (1 - \kappa_0 - \kappa_2)p_1^2 p_2$. This gives the LR

$$LR = \frac{0 \times 2p_1^3 p_2 + 1 \times p_1^2 p_2}{1 \times 2p_1^3 p_2 + 0 \times p_1^2 p_2} = \frac{1}{2p_1}.$$

Note that this LR could have also been obtained intuitively by looking at the genotypes $g_1 = 1/1$ and $g_2 = 1/2$ of the individuals in question:

$$LR = \frac{P(\text{child} = 1/2 \mid \text{father} = 1/1)}{P(\text{child} = 1/2)} = \frac{p_2}{2p_1 p_2} = \frac{1}{2p_1}.$$

1.3.3 p-values

If the alternative hypothesis is not clearly specified, the classical likelihood ratio approach of forensics may not apply. In such situations, one should look at other ways for evaluating the evidence based on classical hypothesis testing. Assume that some DNA evidence is available and that two competing hypotheses, H_1 and H_2, are suggested. As an example, consider two persons that may want to document that they are related, whatever that means. One may then formulate the hypotheses $H_1 : \theta \leq \theta_0$ versus $H_2 : \theta > \theta_0$ for the previously defined θ parameter. One could use $\theta_0 = 0$, or some larger value, say 0.05, in case we would like to demonstrate relatedness beyond the background value. We could calculate a test statistic, for instance a likelihood ratio as defined in Garcia-Magariños et al. [27] by $\Delta = \frac{\sup_{\theta \in H_1} L(\theta)}{\sup_{\theta \in H_1 \cup H_2} L(\theta)}$, or some other test-statistic. However, it remains to calculate a critical value T_0 so that we reject whenever $\Delta \leq T_0$. Alternatively we can calculate

$$p - \text{value} = P(\Delta \leq \Delta^* \mid H_1),$$

where Δ^* is the observed test-statistic. Intuitively, Δ is the ratio of the maximum likelihood under H_1 divided by the maximum over all values of the parameter. This explains why we reject for small Δ values or, equivalently and more common, for large values of $-2\log(\Delta))$.

If we assume that H_1 is true, then the p-value is informally defined as the probability of the observed test static or something more extreme under H_1. We use the p-value to decide whether or not H_1 should be rejected, by comparing the p-value to a chosen significance level, α. If the p-value is less than the given significance level (common values to use are $\alpha = 0.05$ and $\alpha = 0.01$), H_1 is rejected. A more theoretical statistical understanding of the concept may be found in statistical textbooks, like [18] and [49].

The use of p-values for evaluating the strength of DNA evidence has been a topic of discussion in the forensic community. There are those who promote the use of p-values as a supplementary understanding in evaluating the evidence (like Gill et al. [29]), and those who oppose the use of p-values as these in many cases may be misused due to wrong understanding of the concept. Dørum et al. [20] for instance, present p-values for complex DNA profiles were several individuals are involved. The p-value is presented as a supplement to the likelihood ratio, giving a scaled version of the LR. This view of the p-value, as a scaled test statistic or a map to the interval [0,1], is presented in the much cited book by Box et. al [5]. Kruivjer et al. [39], however, followed up on [20] with a paper recommending not to use p-values for evaluating the strength of DNA evidence. They mention different pitfalls, like for instance the prosecutor's fallacy, i.e. wrongly interpreting the p-value as the probability of the alternative hypothesis H_2 being true. They also refer to Goodman [32], discussing how commonly p-values are misinterpreted in scientific research. Their basic point is that all relevant information from the data is contained in the LR. There is another substantial problem with p-values or conventional testing of null hypotheses not mentioned in [39]. This framework is designed for non-symmetric situations: It is more important to avoid falsely rejecting the null hypothesis than failing to reject a null hypothesis which should be rejected. Clearly, p-values need to be handled carefully. However, as the promoters of p-values argue, these may give useful information when handled correctly. Also, we point out here that when we are not able to state an appropriate alternative hypothesis, the verbally based likelihood ratio may not work. In this thesis p-values only appear in paper I. As we elaborate on in the discussion, the reason is that the mentioned paper is motivated by [36] which uses p-values extensively.

1.3.4 Bayesian approach

The LR may be used in a Bayesian framework. In this context, we are able to interpret a given DNA evidence E relative to other types of evidences, and we instead look at which of the two hypotheses in question, H_p and H_d, are most likely given the evidence. This is known as the *posterior* probability, i.e. $P(H_p|E)$ and $P(H_d|E)$. Using Bayes' theorem, we convert the LR to a posterior probability, given by

$$P(H_p|E) = \frac{P(E|H_p)P(H_p)}{P(E|H_p)P(H_p) + P(E|H_d)P(H_d)} = \frac{LR \cdot P(H_p)}{LR \cdot P(H_p) + P(H_d)}, \tag{4}$$

where the last equality is obtained by dividing the numerator and denominator by $\text{Pr}(E|H_d)$.

If we have several competing hypotheses, $H_1, H_2, ..., H_k$, the posterior probability may be presented as

$$\text{Pr}(H_i|E) = \frac{\text{Pr}(E|H_i)\text{Pr}(H_i)}{\sum_{j=1}^{k} \text{Pr}(E|H_j)\text{Pr}(H_j)}.$$

Commonly, a so-called flat prior is used, such that $\text{Pr}(H_p) = \text{Pr}(H_d) = 0.5$. Using the flat prior, we find that the relation in (4) is given by

$$\text{Pr}(H_p|E) = \frac{\text{Pr}(E|H_p)}{\text{Pr}(E|H_p) + \text{Pr}(E|H_d)} = \frac{LR}{LR + 1}.$$

However, having prior information may provide useful details that may reduce or increase the LR if we use the Bayesian framework. As explained in Egeland et al. [24], if say 1000 persons are missing after a large scale disaster, and 10 of these are reported as missing females, the prior probability of an unidentified person will be $1/(1000 + 1)$. However, for families missing only a female, this probability will be $1/(10 + 1)$, and zero for the remaining 990 families missing a male. Clearly, including this information in (4) will provide substantial increase or decrease of posterior probability, hence finding a more reliable LR.

It is also possible to write Bayes theorem on *odds form* as

$$\frac{\Pr(H_1|\text{data})}{\Pr(H_2|\text{data})} = \frac{\Pr(\text{data}|H_1)}{\Pr(\text{data}|H_2)} \times \frac{\Pr(H_1)}{\Pr(H_2)}.$$

This expression clearly demonstrates how the LR modifies our prior belief, as we verbally may state

posterior odds = LR × prior odds.

The paper [14] discusses the relationship between likelihood ratios and posterior odds in different settings.

1.4 Understanding the DNA profile

The use of forensic DNA profiling has been through a great journey and expansion since it was first introduced in the mid 1980s by Sir Alec Jeffreys, [8]. The profiling techniques have evolved rapidly, and forensic scientists from both biological and mathematical sides are continuously working to develop the technologies and make the analyzing methods more robust. After all, a slight error may result in a perpetrator going free, or an innocent person being convicted. DNA typed evidence is based on scientific findings and is therefore considered to provide objective information in crime cases. Forensic DNA profiling is widely recognized as the foremost method for forensic identification, and the technique has even been referred to as "a gold standard for truth telling" [3]. With the constantly improving DNA profiling techniques, cold cases are reopened and solved [60].

The creation of a DNA profile includes several technological steps. After evidence material is gathered from a crime scene, DNA cells are separated from other cell material by DNA extraction (the extraction stage). This is followed by a polymerase chain reaction (PCR) amplification where copies of the STR regions are created. Finally, the capillary electrophoresis stage is reached, where the STR markers are separated and electropherograms are made, presenting allelic peaks giving a visual understanding of the DNA profile. We will not go through the technological steps behind creating a DNA profile, however we recommend [10] to the interested reader. Figure 9 gives an example of an electropherogram, using the ESX17 marker kit (used in Norway).

A profile from a crime scene will typically be of poor quality and hence requires more caution than a profile in a standard kinship case, where the profile most likely will be complete. However, whether or not the profile is complete, there are many considerations that need to be taken while creating a DNA profile, and we discuss some of these in the following sections.

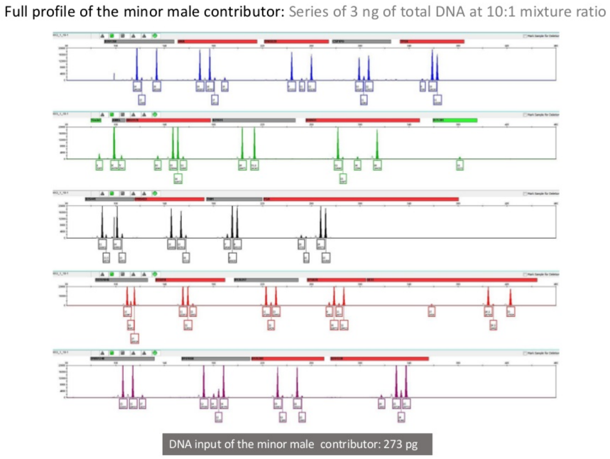


Figure 9: Figure showing an electropherogram (epg) from a two-person mixture.

1.4.1 Mixtures

A DNA mixture refers to a DNA sample where more than one individual has contributed to the stain. A typical sign of a mixture is when the electropherogram shows more than two peak heights at a single marker. The minimum number of individuals contributing to the mixture can therefore be estimated by counting the peaks at the marker with the maximum number of peak heights. Note that this intuitive method fails for SNPs as we will never estimate more than one contributor. However, better estimates are available using maximum likelihood, see Egeland et al. [21] for SNPs and Haned et al. [34] for STRs. Important examples of cases where DNA mixtures often are found are rape cases and murder cases. DNA mixtures have generally not been considered a major problem in kinship cases. However, as paper II and III of this thesis show, mixture problems may occur in kinship cases as well, and need to be handled even more carefully in such cases as the allele peaks may overlap due to shared alleles between relatives.

Figure 9 shows a typical two-person mixture. The contributor with the larger peak heights is referred to as the *major* component, whereas the contributor with the lower peak heights is referred to as the *minor* component. Of course, cases in which both individuals contribute in more or less equal amounts may also occur.

A detailed explanation of how two-person mixtures and higher order mixtures can be detected and handled is explained in detail in [15]. The same paper also explains that higher order mixtures cause computational problems, and suggests that in some circumstances it could be better to lower the dimensionality of the mixture by assuming the presence of a known individual, and subtract this profile from the mixture. In paper I of this thesis, however, we present a method for handling high-order mixtures without needing to specify the number of contributors or lowering the dimension.

Whether the mixture consists of two persons or is more complex, there are several considerations that need to be taken into account while interpreting the mixture. *Stutters* and *heterozygote imbalance* are artifacts that may appear in the electropherogram while handling low-level DNA samples, and may confuse the DNA interpretation. Appearing due to strand slippage, stutter bands typically lack one repeat unit relative to the main allele [59]. Heterozygote imbalance is caused by stochastic effects during the PCR amplification process. The imbalance occurs when the alleles are not amplified with equal peak heights (as one should expect) during the PCR amplification. Figure 10 gives an example of a stutter and heterozygote imbalance. The same figure also gives an example of allelic dropout and drop-in. We will return to these issues in section 1.4.3 and also discuss silent alleles, mutations and population stratifications in more detail in the following sections. The important point for now is that we need to handle artifacts that may appear in the electropherogram as contributing components in the mixture, or vice versa, artifacts causing contributing components to miss out from the electropherogram. The work presented in this thesis does not involve heteozygote imbalance or stutters.

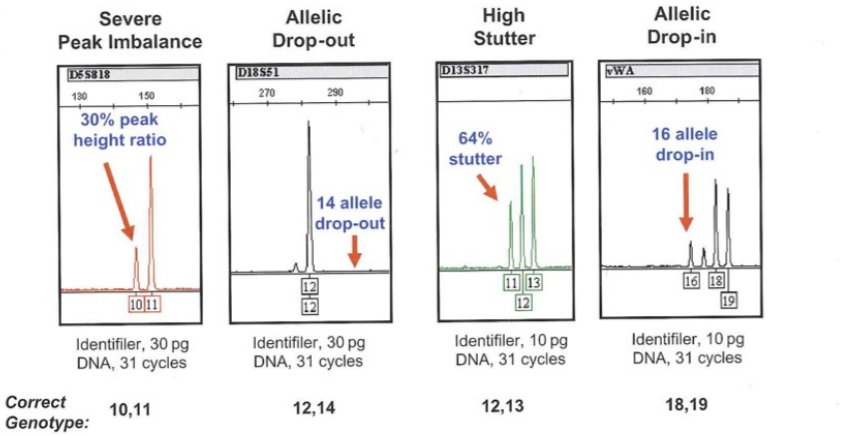


Figure 10: Figure showing heterozygotic imbalance, allelic dropout and drop-in, and stutters, see [10]

1.4.2 Mutations

Changes in DNA sequences are called mutations. The mutation may occur on the somatic level, meaning that the change in the DNA only impacts on the individual level, or in the germ line, impacting future generations as the mutation then occurs in the sex cells. Mutations in the germ line are more severe for kinship cases, as the mutation may effect pedigrees and relations that are questioned.

Mathematically, mutations are accounted for using a mutation matrix,

$$M = \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,K} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ m_{K,1} & m_{K,2} & \cdots & m_{K,K} \end{pmatrix}$$

Each element $m_{i,j}$ in the matrix presents the probability that allele i ends up as allele j. Hence, the diagonal elements are the probabilities of no mutation. There exist several mutation models, and the simplest is the 'equal' mutation model, where the probability of mutating from one allele to another is equal for all alleles. The 'stepwise' mutation model, see [17] and [13] for a mathematical presentation, is an other model where each mutation probability $m_{i,j}$ in the matrix M can be expressed as

$$m_{i,j} = \begin{cases} 1 - R & \text{if } i = j, \\ k_i r^{|i-j|} & \text{if } i \neq j. \end{cases}$$

The parameter R is the mutation rate and k_i are constants defined such that each row in the matrix M sums to 1. The underlying assumption for the stepwise mutation model is that the alleles are considered as repeats or steps [58], and "larger steps" are more unlikely than smaller steps. There exist extensions of the stepwise model, discussed in [38] and [24]. This model distinguishes between integer mutations (like a mutation from 9 to 10) and the rarer mutations between integers and non-integer alleles (like 9 to 9.3 or 9.3 to 9).

Software like `Familias` provides options for handling mutation problems computationally. There is both a Windows version of this software (see [38]) and an `R` version (see chapter 5 of [24]). The latter `R` implementation is used in our `relMix` software presented in paper III. To look at a practical example, consider a parent-child case. As explained in section 1.3.1, to test the hypotheses of whether or not an alleged father is the biological father of a child, we need to calculate the likelihood ratio. For such parent-child cases, there exist a general likelihood ratio formula. Assume that the parent's genotype is a/b and that the child's genotype is c/d. Here the alleles a, b, c and d may or may not differ. Then the likelihood ratio including mutations is generally given by

$$LR = \frac{1}{4} \frac{(m_{a,c} + m_{b,c})p_d + (m_{a,d} + m_{b,d})p_c}{p_c p_d},$$

where p is the allele frequency. We have used this formula to check implementations in our papers. For the 'equal' mutation model, the above LR is simplified even more as we then have $m_{i,i} = R$ and $m_{i,j} = 1 - R/(n-1)$ if $i \neq j$ and n is the number of alleles. If the alleged father and the child do not share any alleles, the LR accounting for mutations will be reduced to

$$LR = \frac{1}{2} \frac{m(p_c + p_d)}{p_c p_d},$$

where $m = 1 - R/(n-1)$.

1.4.3 Drop-in, dropout and silent alleles

Dropout and drop-in was introduced in section 1.4.1. The electropherograms in Figure 10 gave an example of how both terms may cause a misleading understanding of a DNA profile. Recall that problems with dropout is often observed when we work with degraded and low-template DNA.

Drop-ins are observed as additional allele peak heights in the electropheorgram, and appear as a result of sporadic addition in the DNA sample. Generally drop-ins are by definition restricted to one or two alleles in one profile, such that if multiple alleles are observed at more than two loci, these sample are more likely to contain information from an additional individual [29].

Dropouts on the contrary refer to failure of detecting alleles (one or both) at a locus. For diallelic markers we use the term *allelic* dropout when there is loss of one single allele, while the term *locus* dropout is used when both alleles are missing. If dropouts appear, heterozygous markers may falsely be assumed to be homozygous.

Dropouts may also be confused with silent alleles. Both dropouts and silent alleles may appear when an allele in the sample fail to amplify during the PCR reaction. The difference, however, is that dropouts are considered as a random, stochastic effect, and do not occur if the DNA sample is of good quality. Silent alleles on the other side are inherited and may effect several contributors in a family pedigree.

How to account for drop-ins and dropout in DNA profiling have been a topic of discussion over the last years, and Gill et al. [29] give a set of recommendation on how these effects can be handled.

1.5 Implementation

For the papers included in this thesis, three different `R` packages have been developed that are freely available. We here give a short summary of these libraries.

Package betamix The package `betamix` is introduced in paper I. This package may be used for regression analysis on DNA mixtures, and contains two functions; `sim.mod` and `reg1`. Using the function `sim.mod`, data for a number of SNP markers are simulated, which further can be scaled and standardized. Data is returned on a format convenient for regression analysis, for which the function `reg1` can be used. With this function, the proportion contributed from an individual to the mixture is estimated and a p-value is computed. The scaling coefficients are computed and data is returned. The package has been recompiled to work for the current `R` version 3.3.3 and is available from the webpage: `arken.nmbu.no/~theg/betamix_1.1.zip` (updated link compared to paper I).

Package relMix This package is first introduced in paper II, and is later expanded on in paper III. The package is used for for relationship inference based on mixtures and missing reference profiles, and calculates likelihoods for such cases by including drop-in and dropout, mutations, silent alleles and theta correction. The package uses the `R` version of `Familias` [38]. The implementation of the likelihood including dropout and drop-in presented in `relMix` is based on Equations (2.1) and (2.2) of Slooten [50], originally described in the appendix of Haned et al. [35]. The package is freely available at CRAN R, and also comes with a user-friendly graphical user interface (GUI) under function named `relMixGUI()`.

Package IBDest2 In this package, maximum likelihood estimates of IBD coefficients (the κ parameters) are obtained with nonlinear constraints. The functions presented in this package are based on the `R` library `paramlink` [23]. We handle three different cases: 1) Standard - estimates are only restricted to the relationship triangle, see Figure 7. 2) Constrained - estimates are constrained to the permissible region (white area of Figure 7). 3) BIC - we use the Bayesian Information Criteria to find the estimate. Furthermore, parametric bootstrap is implemented so that we can simulate for a pedigree with an arbitrary κ and a confidence ellipse is estimated and drawn. The package is available from `http://familias.name/IBDest2_1.0.zip` and is used in the fourth paper of this thesis.

2 Paper summaries

Figure 11 gives a visual understanding of how the papers in this thesis are ordered. Paper I concerns mixture cases, and we discuss how contributors may be detected. Papers II and III handle mixture problems in kinship cases. Paper IV concerns estimation of relations in kinship cases. In the following sections we summarize the main points of each paper.

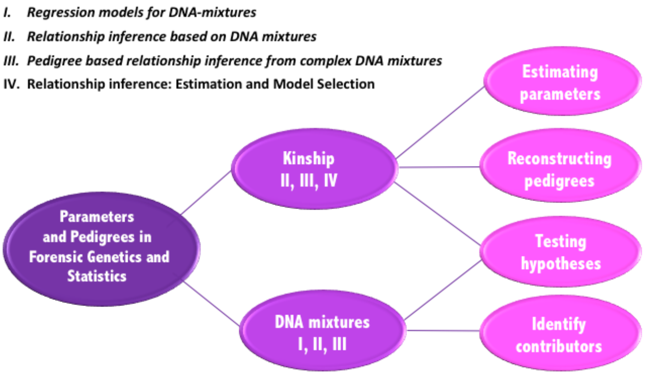

I. Regression models for DNA-mixtures
II. Relationship inference based on DNA mixtures
III. Pedigree based relationship inference from complex DNA mixtures
IV. Relationship inference: Estimation and Model Selection
Parameters and Pedigrees in Forensic Genetics and Statistics
Kinship II, III, IV
DNA mixtures I, II, III
Estimating parameters
Reconstructing pedigrees
Testing hypotheses
Identify contributors

Figure 11: The figure summarizes some of the main aspects met in the four papers of this thesis.

Paper I – Regression models for DNA-mixtures

The paper deals with DNA mixtures involving several contributors, and presents a parametric approach for detecting contributors to mixtures. The conventional methods used in forensics casework are often based on a limited number of STR markers. The paper suggests use of SNP markers as power may be increased. Moving away from the conventional verbal presentation of the hypotheses testing for whether the suspect contributed to the mixture or not, parametric hypotheses are presented, where a person is said to contribute to the mixture if and only if his contribution fraction (denoted by the parameter β) is greater than zero. A regression model is presented based on this contribution fraction β. The model does not require the number of contributors of the mixture to be known, as the contribution from the unknown contributors is replaced by expected values from the population frequencies. Data from 25 controlled, blinded experiments are used to test the model, with contributors to the mixtures varying between 2-5 and their contribution fractions range in the interval (0.01, 0.99), see [22]. These fraction were accurately estimated by the regression analyses, with no false positives, and some false negatives for the small contribution fractions of 0.1 or lower.

Paper II – Relationship inference based on DNA mixtures

The paper was developed while handling a rape case involving DNA mixtures and missing reference profiles. The scenario is as follows: a rape resulted in an unwanted pregnancy, and an abortion was performed. A suspect was later found, and a paternity test was ordered. However, the fetus material obtained from the abortion came in form of a mixture with the mother of the unborn child, and for some reason the victim (the mother) refused to give her reference DNA. Conventional methods for paternity testing did no longer apply, and new methods were needed. Whereas the main emphasis for solving cases involving DNA mixtures often is to determine the contributors to the mixture, we here instead focus on the relationship between the contributors to the mixture. Statistical methods that may handle general relationship inference based on DNA mixtures are presented. The basic idea is that likelihood calculations for mixtures can be decomposed into a series of kinship problems. The development of the `R` library `relMix` started with this paper. The software was, however, extended and we refer to paper III for the updated version.

Paper III – Pedigree based relationship inference from complex DNA mixtures

This paper extends on paper II of this thesis. The calculations have been extended to additionally account for dropout and drop-in as well as mutations, silent alleles and population substructure. An improved version of the `relMix` package is presented, both as a user-friendly graphical user interface (GUI) and as several command line functions in `R`. The motivational example for this paper is as for the previous paper a paternity test where the child's DNA profile only is available as a mixture with the mother's profile. More specifically, the improved method here is developed based on non-invasive prenatal paternity testing cases, where a blood sample taken from a pregnant woman is analysed with next generation sequencing. A highly unbalanced mixture and a very low amount of foetal DNA make dropout and drop-in likely. Whether the aim is to identify the contributors to a mixture who may be related, or to determine the relationship between individuals based on a DNA mixture, both types of problems can be handled by the method and software presented here. We focus on paternity cases in most of the examples in the paper, however, we do emphasize that our software can handle all types of relationships between individuals in a mixture, and the hypotheses may involve any number of relatives. Simulation study shows that the ability to identify true trios is drastically reduced if there is dropout in the data that is not accounted for. The method is also demonstrated on data from a real prenatal paternity case as proof of concept.

Paper IV – Relationship inference: Estimation and model selection

In this paper, we take the parametric framework of identity-by-descent (IBD) probabilities further. The methods and implementations of this paper are relevant whenever parameters describing the relationship between two non-inbred individuals are needed, as their relation may be described by a point $\kappa = (\kappa_0, \kappa_1, \kappa_2)$ in the IBD relationship triangle. Based on these κ parameters we formulate parametric hypotheses suggesting a certain relation versus another specified relation. Hence, we no longer need to state unrelatedness as an alternative hypothesis which is conventionally done when formulating such hypotheses verbally. We expand on already known methods for estimating κ from genetic markers, and take a deeper look into the estimation properties of parameters found on the boundary of the permissible area in the relationship triangle. The main novelty of the paper is that we introduce optimization with non-linear constraints and model selection based on the Bayesian Information Criterion (BIC) to get hold of the boundary issues. Also, we introduce parametric bootstrapping in order to create confidence regions for the estimated κ parameters. The kinship coefficient ψ is also introduced for practical purposes, and plotting methods are presented to visualize the estimated relations.

3 Discussion

In this thesis we have discussed statistical approaches to be applied in both kinship cases and forensic crime cases. Traditional testing of hypotheses in forensic genetics differ from most other applications of statistics as verbal formulations of the hypotheses are used. The main point in paper I and IV is that parametrical models are formulated and that hypotheses are expressed using the parameters of the model. This is the standard approach of applied statistics.

In [36] it was claimed that "mixtures where an individual contributes less than 0.1% of the total genomic DNA" could be handled. The paper [22] critically examined the statistical methods of [36] and stated: "We conclude that it is not possible to reliably infer the presence of minor contributors to mixtures following the approach suggested in [36]". The purpose of paper I was to present appropriate methods for identification of contributors to a mixture. The basic idea is the previously mentioned parametric approach: the hypothesis "POI contributed" is reformulated as $\beta > 0$, where the fraction POI contributes is β. This formulation makes classical statistical theory available. For instance, the statistical power of the test can be studied in the conventional way. In paper I a simple regression model was used. Obviously, more complex models may be needed in future cases. For instance, if more markers are used, the resulting dependence (linkage disequilibrium) must be modeled. However, we emphasize that the specific model is not the main message of paper I, but rather the parametric formulation.

There is a large literature on pairwise relationships. The paper [45] puts these kinship cases into context. Paper IV of this thesis addresses kinship problems and builds on the work of Elizabeth Thompson starting with [51]. Here, verbal statements of questioned kinship relations are replaced by parametric versions with the advantages mentioned above. The classical paternity framework, testing the alleged father against an unrelated man, is restrictive. It may also be problematic if a close relative of the alleged father may be the biological father. This restriction is removed with the parametric formulation as explained further in [27]. Whereas [27] discusses asymptotic distributions of test statistics, paper IV uses simulation and parametric bootstrapping to estimate parameters and confidence regions.

Paper II and the extension, paper III, address cases where both mixtures and individual profiles are available. Some of the contributors may be related and the objective is to determine who contributed to the mixture or to infer relationships between the contributors of the mixture. For these papers, classical verbal formulations of the hypotheses are used. In terms of computations, the likelihoods can be reduced to sums of kinship likelihoods. Dropout and drop-in are modeled using the general formulation of the likelihood in [50]. The paper [42] presents an alternative approach.

There is an `R` library for each of the four papers. The one mentioned in paper II is replaced by the more general and user friendly version of `relMix` discussed in paper III. We expect this package to have some general interest and it has been used for casework not mentioned in the papers. The other libraries, `betamix` and `IBDest2` probably need further testing and some extensions to be of practical general use, but they serve to check the examples of the papers.

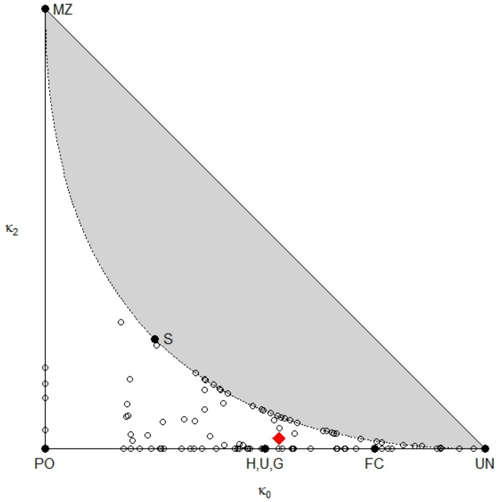


Figure 12: The red diamond indicates the true relationship. We have simulated 100 times, each with 10 markers. See Example 3.1 of paper IV for further details.

In [36], p-values are used extensively, and mainly for this reason we also use p-values in paper I as we are comparing and discussing findings. The use of p-values was discussed in [39] based on [20]. However, much of the controversy around p-values exists due to danger of misinterpretations of the conclusions in court. Sometimes p-values are related to fact finding and not interpretation of evidence, and then the issue is less controversial. Also, if it is impossible to formulate a few specific hypotheses verbally, classical hypothesis testing in parametric models may be needed.

Finally, we like to mention scopes for future work. Most importantly, methods and implementation must be developed and implemented to accommodate new data sources. In several respects more data is needed. For instance, mixture cases with say four or more contributors, are likely to be beyond the reach of analyses based on conventional kits. Similarly, it may be difficult to distinguish close family relationships as paper IV shows. Even testing siblings against half-siblings is difficult based on the limited number of unlinked markers available as discussed in [41]. Figure 12 is based on Example 3.1 of paper IV. Only ten markers are used. The 100 simulated points from the κ (indicated with a red diamond) clearly show that there is not enough data to reliably estimate the relationship. More data is also needed to deal with inbred relationships. There are data sources available of forensic relevance that is not discussed in this thesis, as X-chromosomal markers [56, 46], Y-chromosomal markers [1, 37], and mtDNA [48]. In summary, methods and software must be further developed to solve problems of practical importance while being tailored for new data sources.

4 References

[1] M. M. Andersen, A. Caliebe, A. Jochens, S. Willuweit, and M. Krawczak. Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Science International: Genetics*, 7(2):264–271, 2013.

[2] D. W. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000.

[3] J. Aronson. *Genetic witness: Science, law, and controversy in the making of DNA profiling*. Rutgers University Press, 2007.

[4] D. J. Balding and C. D. Steele. *Weight-of-evidence for Forensic DNA Profiles*. John Wiley & Sons, 2015.

[5] G. E. Box, W. G. Hunter, and J. S. Hunter. *Statistics for experimenters*. Wiley, NY, 1978.

[6] J. Buckleton, C. Triggs, and C. Champod. An extended likelihood ratio framework for interpreting evidence. *Science & Justice*, 46(2):69–78, 2006.

[7] J. Buckleton, J. Curran, J. Goudet, D. Taylor, A. Thiery, and B. Weir. Population-specific F ST values for forensic STR markers: A worldwide survey. *Forensic Science International: Genetics*, 23:91–100, 2016.

[8] J. S. Buckleton, J.-A. Bright, and D. Taylor. *Forensic DNA evidence interpretation*. CRC press, 2016.

[9] B. Budowle and A. Van Daal. Forensically relevant SNP classes. *BioTechniques: The international journal of life science methods*, 44(5):603, 2008.

[10] J. M. Butler. *Advanced topics in forensic DNA typing: methodology*. Academic Press, 2011.

[11] J. M. Butler. *Advanced topics in forensic DNA typing: interpretation*. Academic Press, 2014.

[12] J. M. Butler, M. D. Coble, and P. M. Vallone. STRs vs. SNPs: thoughts on the future of forensic DNA testing. *Forensic science, medicine, and pathology*, 3(3): 200–205, 2007.

[13] A. Caliebe, A. Jochens, M. Krawczak, and U. Rösler. A Markov chain description of the stepwise mutation model: local and global behaviour of the allele process. *Journal of Theoretical Biology*, 266(2):336–342, 2010.

[14] A. Caliebe, S. Walsh, F. Liu, M. Kayser, and M. Krawczak. Likelihood ratio and posterior odds in forensic genetics: Two sides of the same coin. *Forensic Science International: Genetics*, 28:203–210, 2017.

[15] T. Clayton, J. Whitaker, R. Sparkes, and P. Gill. Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International*, 91(1):55–70, 1998.

[16] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge University Press, 1997.

[17] A. P. Dawid, J. Mortera, and V. L. Pascali. Non-fatherhood or mutation?: A probabilistic approach to parental exclusion in paternity testing. *Forensic science international*, 124(1):55–61, 2001.

[18] J. L. Devore and K. N. Berk. *Modern mathematical statistics with applications*. Cengage Learning, 2012.

[19] G. Dørum and M. M. Bouzga. Urns and forensics. *CHANCE*, 28(1):4–11, 2015.

[20] G. Dørum, Ø. Bleka, P. Gill, H. Haned, L. Snipen, S. Sæbø, and T. Egeland. Exact computation of the distribution of likelihood ratios with forensic applications. *Forensic Science International: Genetics*, 9:93–101, 2014.

[21] T. Egeland, I. Dalen, and P. F. Mostad. Estimating the number of contributors to a DNA profile. *International journal of legal medicine*, 117(5):271–275, 2003.

[22] T. Egeland, A. E. Fonneløp, P. R. Berg, M. Kent, and S. Lien. Complex mixtures: A critical examination of a paper by Homer et al. *Forensic Science International: Genetics*, 6(1):64–69, 2012.

[23] T. Egeland, N. Pinto, and M. D. Vigeland. A general approach to power calculation for relationship testing. *Forensic Science International: Genetics*, 9: 186–190, 2014.

[24] T. Egeland, D. Kling, and P. Mostad. *Relationship Inference with Familias and R: Statistical Methods in Forensic Genetics*. Academic Press, 2015.

[25] I. W. Evett and B. S. Weir. *Interpreting DNA evidence: statistical genetics for forensic scientists*. Sinauer, 1998.

[26] W. K. Fung and Y.-Q. Hu. *Statistical DNA forensics: theory, methods and computation*. John Wiley & Sons, 2008.

[27] M. García-Magariños, T. Egeland, I. López-de Ullibarri, N. L. Hjort, and A. Salas. A parametric approach to kinship hypothesis testing using identity-by-descent parameters. *Statistical applications in genetics and molecular biology*, 14(5):465–479, 2015.

[28] P. Gill, C. H. Brenner, J. S. Buckleton, A. Carracedo, M. Krawczak, W. Mayr, N. Morling, M. Prinz, P. M. Schneider, and B. Weir. DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures. *Forensic Science International*, 160(2):90–101, 2006.

[29] P. Gill, L. Gusmão, H. Haned, W. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P. Schneider, et al. DNA commission of the International Society of Forensic Genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. *Forensic Science International: Genetics*, 6(6):679–688, 2012.

[30] P. Gill, H. Haned, O. Bleka, O. Hansson, G. Dørum, and T. Egeland. Genotyping and interpretation of STR-DNA: low-template, mixtures and database matches - twenty years of research and development. *Forensic Science International: Genetics*, 18:100–117, 2015.

[31] D. W. Gjertson, C. H. Brenner, M. P. Baur, A. Carracedo, F. Guidet, J. A. Luque, R. Lessig, W. R. Mayr, V. L. Pascali, M. Prinz, et al. ISFG: recommendations on biostatistics in paternity testing. *Forensic Science International: Genetics*, 1(3): 223–231, 2007.

[32] S. Goodman. A dirty dozen: twelve p-value misconceptions. In *Seminars in hematology*, volume 45, pages 135–140. Elsevier, 2008.

[33] J. Hajnal, M. Fraccaro, J. Sutter, and C. Smith. Concepts of random mating and the frequency of consanguineous marriages [and discussion]. *Proceedings of the Royal Society of London B: Biological Sciences*, 159(974):125–177, 1963.

[34] H. Haned, L. Pene, J. R. Lobry, A. B. Dufour, and D. Pontier. Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? *Journal of Forensic Sciences*, 56(1): 23–28, 2011.

[35] H. Haned, K. Slooten, and P. Gill. Exploratory data analysis for the interpretation of low template DNA mixtures. *Forensic Science International: Genetics*, 6(6): 762–774, 2012.

[36] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*, 4(8):e1000167, 2008.

[37] M. Kayser. Forensic use of Y-chromosome DNA: a general overview. *Human Genetics*, pages 1–15, 2017.

[38] D. Kling, A. O. Tillmar, and T. Egeland. Familias 3–extensions and new functionality. *Forensic Science International: Genetics*, 13:121–127, 2014.

[39] M. Kruijver, R. Meester, and K. Slooten. p-values should not be used for evaluating the strength of DNA evidence. *Forensic Science International: Genetics*, 16:226–231, 2015.

[40] B. Maddox. The double helix and the 'wronged heroine'. *Nature*, 421(6921): 407–408, 2003.

[41] L. Mayor and D. Balding. Discrimination of half-siblings when maternal genotypes are known. *Forensic Science International*, 159:141–147, 2006.

[42] J. Mortera, C. Vecchiotti, S. Zoppis, and S. Merigioli. Paternity testing that involves a DNA mixture. *Forensic Science International: Genetics*, 23:50–54, 2016.

[43] D. Murdoch and E. Chow. A graphical display of large correlation matrices. *The American Statistician*, 50(2):178–180, 1996.

[44] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. In *Breakthroughs in statistics*, pages 73–108. Springer, 1992.

[45] M. Nothnagel, J. Schmidtke, and M. Krawczak. Potentials and limits of pairwise kinship analysis using autosomal short tandem repeat loci. *International journal of legal medicine*, 124(3):205–215, 2010.

[46] M. Nothnagel, R. Szibor, O. Vollrath, C. Augustin, J. Edelmann, M. Geppert, C. Alves, L. Gusmão, M. Vennemann, Y. Hou, et al. Collaborative genetic mapping of 12 forensic short tandem repeat (STR) loci on the human X chromosome. *Forensic Science International: Genetics*, 6(6):778–784, 2012.

[47] C. Phillips. Applications of autosomal SNPs and indels in forensic analysis. *Forensic DNA Analysis: Current Practices and Emerging Technologies*, page 279, 2013.

[48] M. Prinz, A. Carracedo, W. Mayr, N. Morling, T. Parsons, A. Sajantila, R. Scheithauer, H. Schmitter, and P. M. Schneider. DNA Commission of the International Society for Forensic Genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI). *Forensic Science International: Genetics*, 1(1):3–12, 2007.

[49] J. Rice. *Mathematical statistics and data analysis*. Nelson Education, 2006.

[50] K. Slooten. Familial searching on DNA mixtures with dropout. *Forensic Science International: Genetics*, 22:128–138, 2016.

[51] E. A. Thompson. The estimation of pairwise relationships. *Annals of Human Genetics*, 39(2):173–188, 1975.

[52] E. A. Thompson. A restriction on. the space of genetic relationships. *Annals of Human Genetics*, 40(2):201–204, 1976.

[53] E. A. Thompson. *Pedigree analysis in human genetics*. Johns Hopkins University Press, 1986.

[54] E. A. Thompson. Statistical inference from genetic data on pedigrees. In *NSF-CBMS regional conference series in probability and statistics*, pages i–169. JSTOR, 2000.

[55] A. O. Tillmar and P. Mostad. Choosing supplementary markers in forensic casework. *Forensic Science International: Genetics*, 13:128–133, 2014.

[56] A. O. Tillmar, P. Mostad, T. Egeland, B. Lindblom, G. Holmlund, and K. Montelius. Analysis of linkage and linkage disequilibrium for eight X-STR markers. *Forensic Science International: Genetics*, 3(1):37–41, 2008.

[57] C. M. Triggs, J. S. Buckleton, and S. J. Walsh. *Forensic DNA Evidence Interpretation*. CRC Press, 2004.

[58] A. M. Valdes, M. Slatkin, and N. Freimer. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics*, 133(3):737–749, 1993.

[59] P. S. Walsh, N. J. Fildes, and R. Reynolds. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. *Nucleic Acids Research*, 24(14):2807–2812, 1996.

[60] R. Williams and P. Johnson. Inclusiveness, effectiveness and intrusiveness: issues in the developing uses of DNA profiling in support of criminal investigations. *The Journal of Law, Medicine & Ethics*, 33(3):545–558, 2005.

[61] S. Wright. The genetical structure of populations. *Annals of Eugenics*, 15(1): 323–354, 1949.

[62] S. Wright. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, pages 395–420, 1965.

[63] B. Yoshida. Mendelian genetics, 2016. URL `http://www.grossmont.edu/people/bonnie-yoshida-levine/online-lectures/genetics-mendel.aspx`.

[64] A. Ziegler, I. R. König, and F. Pahlke. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an E-learning platform*. John Wiley & Sons, 2010.

Paper I

Forensic Science International: Genetics 11 (2014) 105–110

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Regression models for DNA-mixtures

Navreet Kaur [a,*], Ane Elida Fonneløp [b], Thore Egeland [a,b]

[a] IKBM, Norwegian University of Life Sciences, Ås, Norway
[b] Norwegian Institute of Public Health, Oslo, Norway

ARTICLE INFO

Article history:
Received 29 August 2013
Received in revised form 26 February 2014
Accepted 2 March 2014

Keywords:
Forensics
DNA-mixtures
Regression models

ABSTRACT

This paper deals with the statistical interpretation of DNA mixture evidence. The conventional methods used in forensic casework today use something like 16 STR-markers. Power can be increased by rather using SNP-markers. New statistical methods are then needed, and we present a regression framework. The basic idea is that the traditional forensic hypotheses, commonly denoted H_D and H_P, are replaced by parametric versions: a person contributes to a mixture if and only if the fraction he contributes is greater than 0. This contributed fraction is a parameter of the regression model. The regression model uses the peak heights directly and there is no need to specify or estimate the number of contributors to the mixture. Also, drop-in and drop-out pose no principal problems.

Data from 25 controlled blinded experiments were used to test the model. The number of contributors varied between 2 and 5, and the fractions contributed ranged from 0.01 to 0.99. The fractions were accurately estimated by the regression analyses. There were no false positives (i.e., in no cases were non-contributors declared to contributors). Some false negatives occurred for fractions of 0.1 or lower. Simulations were performed to test the model further. The analyses show that useful estimates can be obtained from a relatively small number of SNP-markers. Reasonable results are achieved using 300 markers which is close to the 313 SNPs in the controlled experiment. Increasing the number of SNPs, the analyses demonstrate that individuals contributing as little as 1% can reliably be detected, which suggests that cases beyond the reach of conventional forensic methods today can be reported.

1. Introduction

The topic of this paper is the evaluation of DNA mixture evidence which refers to cases where there are, or could be, several contributors to a biological stain. The basic question is whether a specific individual has contributed to the mixture and we present new statistical methods which are tested on data from a controlled (blinded and randomised) experiment.

When analysing DNA-mixtures from a crime scene, the tradition has been to use STR analysis in forensic case work. By use of electropherograms, the DNA-mixtures are characterized by markers showing more than two peaks [1]. Instead of using the conventional STR-markers, we present an approach based on SNP-markers. Such markers have been studied previously in forensic contexts in e.g., [2–5], but typically aiming for kinship applications rather than mixture interpretation. But as SNP-markers are diallelic, the mixtures are not that easily recognized and proper statistical methods are required. Still, using SNPs in forensic case work can be very helpful, mainly because a much larger set of markers will be available. This in turn can be useful to handle mixtures where many contributors are involved, and also to extend the forensic case work so that individuals contributing a very small amount (close to 0) can be detected. The indicated forensic applications are those we have in mind for the methods developed in this paper. However, statistical methods for DNA-mixtures are relevant also for *pooled data* typically used in Genome Wide Analysis Studies (GWAS). DNA from a large number of individuals are then mixed to be able to estimate allele frequencies from one sample. A widely cited paper [6] presented statistical methods designed to determine contributors to a mixture with both pooling and forensic applications. In GWAS, there is typically a large number of individuals contributing to the pooled sample, whereas for forensic cases, the number of contributors will generally be small, say up to 5. Also, the contribution amount is assumed to be equal for all contributors in a pooled sample, which typically will not be the case in a forensic setting. Last, there are issues related to the amount and the quality of the DNA obtained from the crime scene; degradation or inhibition may lead to DNA profiles of poor quality.

[*] Corresponding author. Tel.: +47 64966391.
E-mail address: navreet.kaur@nmbu.no (N. Kaur).

http://dx.doi.org/10.1016/j.fsigen.2014.03.002
1872-4973/© 2014 Elsevier Ireland Ltd. All rights reserved.

Homer [6] claimed that “mixtures where an individual contributes less than 0.1% of the total genomic DNA” could be handled. The paper [7] critically examined the statistical methods of [6] and stated “We conclude that it is not possible to reliably infer the presence of minor contributors to mixtures following the approach suggested in Homer et al. (2008)”.

Clearly, more robust methods are required to handle DNA-mixtures in forensic casework. We here present a new statistical method to resolve DNA-mixtures based on SNP-markers, where the number of contributors do not need to be specified. This is done by including a term accounting for the expected contribution from unknown contributors. Testing whether a person has contributed to a DNA-mixture is reformulated in terms of a parameter: a person contributes to a mixture if and only the proportion he contributes is greater than 0. While this may appear as a trivial statement, it has wide ranging implications. The tradition of forensic genetics is to formulate hypotheses using verbal statements. This contradicts virtually all other areas dealing with statistical testing of hypotheses. There are several advantages to the parametric approach. In our context it is important to realise that this approach provides access to standard statistical methods and implementations.

2. Data and methods

2.1. Data

The data were collected by performing twenty-five controlled experiments, where DNA-mixtures were made from a number of contributors varying between two and five, as explained in [8]. Information on the number of contributors was not used or available during data analyses. We used the Illumina GoldenGate(R) 360 SNP test panel. SNPs not on the autosomes were removed, as were monomorphic SNPs, leaving 313 markers for the analyses. The alleles are denoted by 1 and 2, and their relative frequencies have been found using the Utah residents with Northern and Western European ancestry from the CEPH collection in the HapMap database. The contributors were randomly chosen among five reference persons (denoted F, D, B, H and C), with contribution proportion ranging from 0.01 to 0.99. Information on the contributors was kept blinded until the analyses were completed. Table 1 shows excerpts of the data. Line 8 of Table 4 gives an example of a two-person mixture, where individuals D and F by design contribute a fraction of 0.5 each to the mixture Blind8.

2.2. Method motivation

To motivate the statistical method, we start by looking at Fig. 1. The figure gives a simple picture of a DNA mixture where DNA from two individuals (a victim and a suspect) is mixed in different fractions. The victim (solid area) only has a peak for allele 1, corresponding to the victim being homozygous 1/1, whereas the suspect (shaded area) only has a peak for allele 2. As a result, the DNA mixture has one peak for each allele, but a larger peak height for allele 1 as the victim contributes with a larger fraction than the suspect. For the statistical method we will assume that the genotype of one potential contributor, typically the suspect, is available and summarized by the number of 1-alleles (denoted by x) and the peak heights for each allele (denoted by y). See Table 2 for a summary of the genotype for the suspect in question.

Table 1
Excerpts of data from the DNA-mixture named Blind8. The y column gives the signal strength for the allele and the SNP indicated. Data are shown for two SNPs (denoted 1 and 4) and two of the five persons (F and D) involved in the experiment. Only x, the number of 1-alleles, varies between individuals. For instance, individual F has genotype 2/2 for SNP 1.

Mixture	SNP	y	x	Allele	Freq.	Person
Blind8	1	12,929	0	1	0.49	F
Blind8	1	19,691	2	2	0.49	F
Blind8	1	12,929	2	1	0.49	D
Blind8	1	19,691	0	2	0.49	D
Blind8	4	17,962	2	1	0.65	F
Blind8	4	13,888	0	2	0.65	F
Blind8	4	17,962	1	1	0.65	D
Blind8	4	13,888	1	2	0.65	D

To investigate whether the suspect did contribute to the mixture, we can formulate two basic hypotheses:

- H_0 : the suspect did not contribute to the mixture,
- H_1 : the suspect contributed to the mixture.

Letting β_1 denote the fraction contributed by the suspect, the hypotheses correspond to

$$\begin{aligned} &\bullet H_0 : \beta_1 = 0, \\ &\bullet H_1 : \beta_1 > 0. \end{aligned} \tag{1}$$

If we know let β_2 denote the fraction contributed by the victim, we must have that $\beta_1 + \beta_2 = 1$. The number of 1-alleles for the suspect and the victim is given by x_1 and x_2, respectively, and therefore the total expected signal of the DNA-mixture for allele 1 is

$$E(y) = \beta_1 x_1 + \beta_2 x_2.$$

The model we will develop only assumes that the genotype of the suspect is available. Therefore the contribution from the victim is replaced by the *expected* contribution

$$\begin{aligned} E(y) &\approx \beta_1 x_1 + \beta_2 E(x_2) \\ &= \beta_1 x_1 + \beta_2 \mu, \end{aligned}$$

where μ is estimated from population data as explained in the next section. In terms of regression analysis, the signal y can now be expressed as

$$y = \beta_1 x_1 + (1 - \beta_1)\mu + noise. \tag{2}$$

The regression model given in (2) allows for statistical inference on our hypothesis that the suspect did not contribute to the mixture ($H_0 : \beta_1 = 0$). As mentioned in the introduction, the parametric formulation of the hypothesis corresponds to the approach most widely used to test hypotheses in statistics; the tradition in forensic genetics deviates by using the verbal statements. We return to a discussion of the pros and cons of the two approaches and also discuss more precise formulations of the hypotheses. A basic idea of the present paper is, however, to explore the parametric approach.

2.3. Statistical model

2.3.1. Basic model

Using the parametric representation of the hypothesis in (1), we now extend model (2) to involve several contributors. Let $i = 1, \ldots, n$ be an index for the SNP, $j = 1, 2$ indicate the allele, and let x_{kij} give the number of 1-alleles from individual k ($k = 1, \ldots, K$). Let β_k be the proportion contributed by individual k, so that $\sum_{k=1}^{K} \beta_k = 1$. With an error term ϵ_{ij}, standard multiple regression theory gives us the following regression model for calculating the peak height y:

$$y_{ij} = \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_K x_{Kij} + \epsilon_{ij} \tag{3}$$

For the applications we have in mind, the total number of contributors is unknown and only the suspect is genotyped. We replace the contribution from the unknown contributors by the

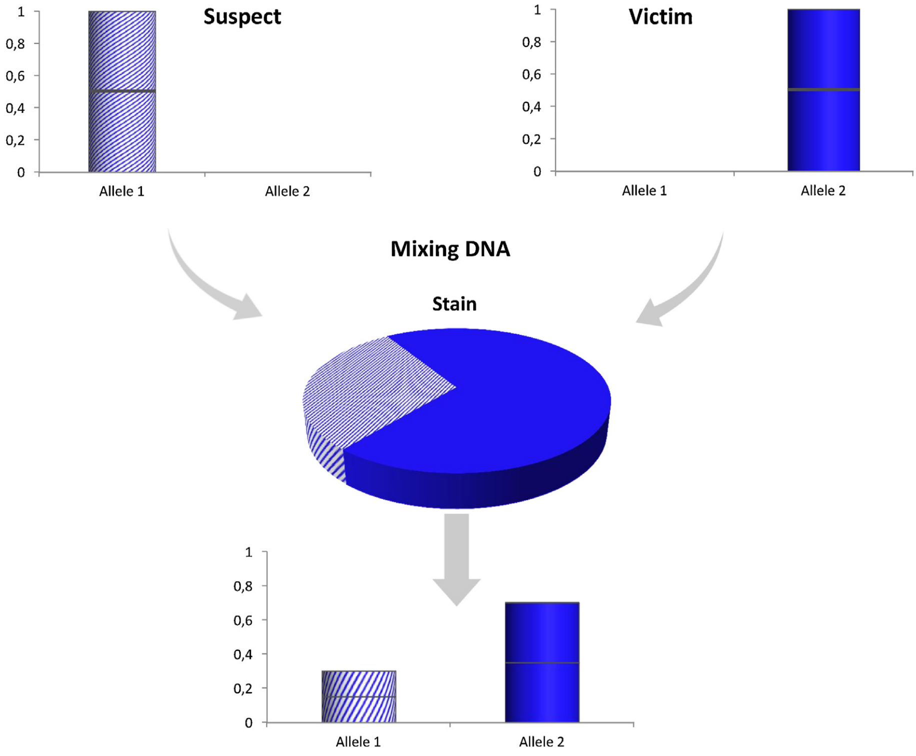


Fig. 1. DNA-mixture: two individuals contributing to a mixture with different amounts.

expected counterpart. From the definition of mathematical expectation and theta-correction as explained in [9] we find for any contributor k

$$E(x_{kij}) = 1 \cdot 2p_i(1-p_i)(1-\theta) + 2(\theta p_i + (1-\theta)\cdot p_i^2) = 2p_i,$$

where p_i is the probability of allele '2'. The total signal sign in the DNA-mixture can now be expressed by

$$\begin{aligned} y_{ij} &\approx \beta_1 x_{1ij} + E(\beta_2 x_{2ij} + \cdots + \beta_K x_{Kij}) + \epsilon_{ij} \\ &= \beta_1 x_{1ij} + (\beta_2 + \cdots + \beta_K)\mu_{ij} + \epsilon_{ij} \\ &= \beta_1 x_{1ij} + (1-\beta_1)\mu_{ij} + \epsilon_{ij}, \end{aligned} \tag{4}$$

giving us a similar model as in Eq. (2). If we let $z_{ij} = y_{ij} - \mu_{ij}$ and $u_{ij} = x_{ij} - \mu_{ij}$, the regression model in (4) may be rewritten in terms of a simple linear regression model, which can be used to test the null hypothesis (1):

$$z_{ij} = \beta u_{ij} + \epsilon_{ij}. \tag{5}$$

We refer to (5) as our basic model. As opposed to other statistical models for mixtures like in [10], our model does not require the number of contributors to be estimated. This, and the fact that the contribution amount from unknown contributors is replaced by expected values from population frequencies, makes our model more robust to handle mixture cases.

Table 2
Peak heights for a mixture and genotypes for the suspect for two SNP markers.

SNP	y (peak height)	x_1	Allele	Individual
1	0.7	2	1	Suspect
1	0	0	2	Suspect
2	1	1	1	Suspect
2	1	1	2	Suspect

2.3.2. Assumptions

Several assumptions have been made for our regression model. We have assumed that the residual terms ϵ_{ij} are independent and normally distributed with constant variance σ^2. The estimate of the slope β is reasonable without these assumptions as it can be considered a least square estimate. Normality, constant variance, independence is needed foremost when p-values are calculated. In any case, assumptions can be checked based on the residuals. The normality assumption may not be so important given the large sample in view of the central limit theorem. The simplest version of this theorem requires the error terms to be independent and identically distributed but both assumptions can be relaxed to some extent. Linkage and linkage disequilibrium (LD) may imply dependence and may therefore potentially lead to assumptions being violated. However, this is no problem for our data as using only 313 markers allows the distance between markers to be large enough to avoid LD and linkage.

Allele frequencies may vary between populations and this may cause problems for calculations. For this reason it was important to perform controlled, blinded experiments which allow the accuracy of the approach to be verified. Moreover in Section 3.2 we look at some simulations that consider cases where the p_i may vary and differ from the valued used in the calculation.

p-Values are used to test hypotheses in line with most other applied areas. As a large number of hypotheses are tested in our controlled experiment, we have corrected for multiple testing. This can be done in several ways, we have used Bonferroni correction which allows for dependence between the hypotheses tested.

2.3.3. Scaling

The specific model typically depends on the data and may therefore need adjustment depending on the chosen set of markers. From our data, it was apparent that the signal strength corresponding to allele 1 and 2 differed. This can be seen by considering markers for which all contributors are heterozygous. The peaks differ much more than would be expected from random fluctuations. For instance, the ratio of the mean peak height for allele 2 to allele 1 for our data is 3.0. (Note that this could be checked without breaking the code, i.e., revealing the true contributors and their fractions contributed of the blinded experiment.) To correct for the differing values of y for two alleles, we scale the data based on the model:

$$y_{ij} = c_j[\beta_1 x_{1ij} + (1-\beta_1)\mu_{ij}] + \epsilon_{ij}. \qquad (6)$$

Observe that $E(y_{ij}) = c_j\mu_{ij}$ and so the moment estimate becomes

$$\hat{c}_j = \frac{\bar{y}_{.j}}{\bar{\mu}_{.j}}, \quad \text{where } \bar{y}_{.j} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}, \quad \text{and} \quad \bar{\mu}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\mu_{ij}.$$

Therefore, if we let $z_{ij} = y_{ij}/\hat{c}_j - \mu_{ij}$ and $u_{ij} = x_{ij} - \mu_{ij}$, the regression model in (6) may be written in the generic form (5).

2.4. Simulations and implementation

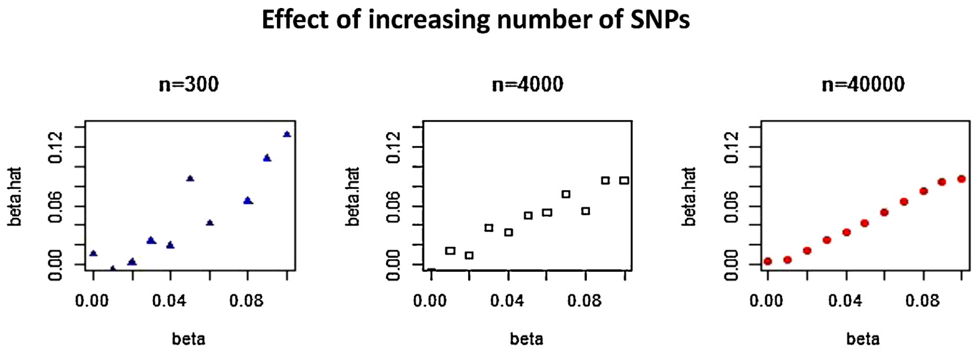
The regression model was tested by a simulation algorithm that generates a set of data from model (6). The algorithm tests how well the model performs under different conditions by varying the contribution amount β and the number of SNPs. To make the data realistic, the mean and standard deviation of peak height was simulated to resemble the data described previously.

For all numerical calculations we have used the freely available `R` package (http://cran.r-project.org/) and also functions in our R package `betamix` which is freely available from `arken.umb.no/~theg/betamix_1.1.zip`. The data described in Section 2.1 is available as part of the package. Throughout, key computations of the paper appear as documented examples of `betamix` and for some central results we provide pointers to the relevant functions. Note that we reduce the analysis to standard statistical models and so a great number of programs can do similar analyses for most cases.

3. Results

3.1. Simulation experiment

Fig. 2 displays the effect of increasing the number of SNPs. The fraction contributed, β, ranged from 0 to 0.10. Clearly, the figures show that there is a pronounced effect from increasing the number of SNPs from 300 to 4000.



β	$\hat{\beta}$ (n=300)	p-val	$\hat{\beta}$ (n=4000)	p-val	$\hat{\beta}$ (n=40000)	p-val
0	0.011	0.758	-0.007	0.454	0.004	0.175
0.01	-0.006	0.863	0.014	0.139	0.006	0.074
0.02	0.002	0.944	0.009	0.336	0.015	$< 10^{-5}$
0.03	0.024	0.500	0.038	0.000	0.025	$< 10^{-16}$
0.04	0.019	0.556	0.033	0.002	0.032	$< 10^{-16}$
0.05	0.086	0.019	0.050	$< 10^{-5}$	0.042	$< 10^{-16}$
0.06	0.041	0.266	0.053	$< 10^{-5}$	0.052	$< 10^{-16}$
0.07	-0.014	0.708	0.071	$< 10^{-5}$	0.064	$< 10^{-16}$
0.08	0.064	0.087	0.055	$< 10^{-5}$	0.075	$< 10^{-16}$
0.09	0.107	0.004	0.086	$< 10^{-5}$	0.084	$< 10^{-16}$
0.10	0.132	7.1e-05	0.086	$< 10^{-5}$	0.087	$< 10^{-16}$

Fig. 2. The figures show the effect of increasing the number of SNPs (n). The estimated β-values and corresponding p-values are found in the tables under each figure.

Table 3
Testing robustness: results for estimating $\hat{\beta}$ with two different allele frequencies, p_i. β is estimated twice; when p_i has a standard deviation term (sd) 0.01 and 0.05. The number of SNPs n was set to 4000 for all simulations.

β	p_i	$\hat{\beta}$ (sd = 0.01)	$\hat{\beta}$ (sd = 0.05)
0.05	0.5	0.049	0.047
0.05	0.2	0.049	0.044
0.10	0.5	0.098	0.095
0.10	0.2	0.097	0.089
0.20	0.5	0.200	0.190
0.20	0.2	0.200	0.180
0.30	0.5	0.300	0.290
0.30	0.2	0.290	0.270
0.40	0.5	0.390	0.380
0.40	0.2	0.390	0.350
0.50	0.5	0.490	0.480
0.50	0.2	0.490	0.440

Contribution amounts close to 0 seem to be hard to detect in all three cases. The p-values are high and for $\beta = 0$ the model may give negative estimated values. This may be handled by adding the restriction $\hat{\beta} = \max(\hat{\beta}, 0)$. Note that this restriction does not effect the p-values.

3.2. Testing robustness: allele frequencies

To investigate the robustness of the model, calculations were done to test how the model handles uncertainty in the allele frequencies. By adding a standard deviation term to the allele frequencies p_i, we tested how the model works when the estimated allele frequencies differ from the true allele frequencies. The results are given in Table 3. With two different standard deviations for the allele frequency p_i (0.01 and 0.05), β is estimated twice for 4000 SNP markers. We see that the estimated contribution amount $\hat{\beta}$ are reasonably close to the true β. The p-values were also found to be close to 0 for all cases (data omitted), suggesting that our statistical model may handle uncertainty in the allele frequencies.

3.3. Data analysis: the controlled, blinded experiment

The analyses of the data presented in Section 2.1 are summarised in Table 4. To exemplify, note that reference person F contributes a proportion of 0.1 by design for experiment 'Blind1' (first row of Table 4). The estimated proportion is 0.09, with p-value at 0.0046. In other words we find that the model accurately estimates the proportion contributed. There is no reason to doubt the assumptions of the model based on the standard checking (the documentation of the function `reg1` in `betamix` includes the commands needed to check the assumptions).

Next we consider the complete Table 4 (the documentation of the function `makeTable` in `betamix` includes commands for the coming analyses). A total of 125 comparisons are made. Consider first the conventional significance level of $\alpha = 0.05$. Then, the false negative fraction (corresponding to $\beta > 0$, $p - \text{value} > \alpha$) is 0.04 (five cases); all occurring for designed fractions of 0.05 or less. The false positive fraction (corresponding to $\beta = 0$, $p - \text{value} < \alpha$) is also 0.04 (five cases) with all occurring for designed fractions of 0.05 or less. It can be argued that it is reasonable to correct for multiple testing. The Bonferroni correction corresponds to a significance level of 0.05/125 = 0.0004. With this level, no false positive remains, while the false negative fraction increases to 0.096; all corresponding to designed proportions of 0.1 or less.

4. Discussion

In the previous sections, a regression model for analysing DNA mixtures has been presented and exemplified based on simulated data as well as a controlled experiment. However, the general approach is not restricted to SNP-markers. The parametric formulation of hypotheses applies equally well for STR-markers. Similarly, by replacing the contributions from unknown contributors by the corresponding expected value, there is no need to specify or estimate the number of contributors. However, the expected contribution depends on allele frequencies. It may be hard to estimate allele frequencies accurately and also allele frequencies

Table 4
Results from running the statistical model on the 25 controlled experiments (indicated by the 25 rows). For each reference person (F, D, B, H and C) we have three columns, giving the true contribution amount (F for reference person F), the estimated contribution amount (est.F for reference person F) and the corresponding p-value (p.F) found in the analysis.

	F	est.F	p.F	D	est.D	p.D	B	est.B	p.B	H	est.H	p.H	C	est.C	p.C
1	0.10	0.09	0.0046	0.30	0.27	0.0000	0.30	0.33	0.0000	0.30	0.40	0.0000	0.00	0.01	0.8529
2	0.10	0.08	0.0021	0.23	0.23	0.0000	0.23	0.31	0.0000	0.23	0.31	0.0000	0.23	0.22	0.0000
3	0.10	0.10	0.0019	0.45	0.42	0.0000	0.45	0.50	0.0000	0.00	0.08	0.0131	0.00	0.05	0.0730
4	0.10	0.13	0.0009	0.90	0.87	0.0000	0.00	0.05	0.3747	0.00	0.04	0.3327	0.00	0.05	0.2347
5	0.20	0.16	0.0000	0.20	0.21	0.0000	0.20	0.28	0.0000	0.20	0.28	0.0000	0.20	0.19	0.0000
6	0.33	0.30	0.0000	0.33	0.33	0.0000	0.33	0.37	0.0000	0.00	0.05	0.1056	0.00	0.03	0.3021
7	0.25	0.21	0.0000	0.25	0.24	0.0000	0.25	0.32	0.0000	0.25	0.36	0.0000	0.00	0.02	0.5499
8	0.50	0.49	0.0000	0.50	0.51	0.0000	0.00	0.05	0.2340	0.00	0.05	0.1637	0.00	0.03	0.4324
9	0.45	0.40	0.0000	0.25	0.26	0.0000	0.30	0.37	0.0000	0.00	0.06	0.0321	0.00	0.04	0.2144
10	0.50	0.41	0.0000	0.10	0.13	0.0000	0.15	0.24	0.0000	0.25	0.37	0.0000	0.00	0.00	0.9013
11	0.70	0.63	0.0000	0.10	0.14	0.0000	0.01	0.09	0.0334	0.09	0.16	0.0000	0.10	0.12	0.0002
12	0.80	0.76	0.0000	0.20	0.24	0.0000	0.00	0.04	0.3598	0.00	0.04	0.3158	0.00	0.02	0.4927
13	0.80	0.73	0.0000	0.05	0.11	0.0040	0.15	0.21	0.0000	0.00	0.05	0.1859	0.00	0.02	0.6434
14	0.30	0.23	0.0000	0.10	0.11	0.0001	0.25	0.32	0.0000	0.35	0.46	0.0000	0.00	0.00	0.9184
15	0.20	0.17	0.0000	0.25	0.25	0.0000	0.30	0.38	0.0000	0.15	0.25	0.0000	0.10	0.12	0.0000
16	0.01	0.05	0.2448	0.99	0.94	0.0000	0.00	0.05	0.3276	0.00	0.04	0.3094	0.00	0.05	0.2229
17	0.40	0.40	0.0000	0.60	0.59	0.0000	0.00	0.05	0.2843	0.00	0.04	0.2154	0.00	0.02	0.4993
18	0.15	0.12	0.0000	0.20	0.20	0.0000	0.25	0.30	0.0000	0.30	0.39	0.0000	0.10	0.11	0.0001
19	0.10	0.08	0.0058	0.20	0.18	0.0000	0.30	0.33	0.0000	0.40	0.49	0.0000	0.00	0.00	0.9796
20	0.20	0.18	0.0000	0.30	0.30	0.0000	0.50	0.54	0.0000	0.00	0.08	0.0088	0.00	0.05	0.0732
21	0.30	0.30	0.0000	0.70	0.67	0.0000	0.00	0.04	0.3448	0.00	0.03	0.3227	0.00	0.03	0.4273
22	0.05	0.03	0.1949	0.24	0.23	0.0000	0.24	0.32	0.0000	0.24	0.32	0.0000	0.24	0.23	0.0000
23	0.05	0.04	0.1514	0.32	0.28	0.0000	0.32	0.34	0.0000	0.32	0.43	0.0000	0.00	0.01	0.6585
24	0.05	0.05	0.1567	0.48	0.44	0.0000	0.48	0.51	0.0000	0.00	0.07	0.0401	0.00	0.05	0.1529
25	0.05	0.08	0.0609	0.95	0.88	0.0000	0.00	0.05	0.3322	0.00	0.03	0.4700	0.00	0.04	0.3057

may differ depending on ethnicity. In fact, the STR-markers used in forensics are typically chosen to have uniform frequencies across populations; this may not be the case for the SNP-markers we have used. The experience from the controlled experiment and simulations indicate that the results are robust.

For practical reasons, we used a relatively small number of markers for the experiment. Also, we have emphasized that our approach is not specifically linked to one set of markers. Hopefully, the statistical methods presented in this paper will be used to analyze data coming from different platforms in the future. In particular, NGS approaches are gaining attraction as explained in several chapters of [11]. Also, insertion deletion polymorphisms (Indels) markers have been demonstrated to be promising [12]. The sequence balance obtained from each allele for NGS and Indels are likely to be better than that obtained from GWAS methods. Therefore, the approach of the present paper is likely to be well suited. The scaling presented in Section 2.3.3 can only take care of systematic differences in signal strength from the two alleles. Obviously, more experience with small amounts of (possibly degraded) DNA is needed.

The regression analysis requires that the residuals are independent. If a large number of markers is used, say more than 4000, the independence assumption may be dubious and there may be a need for more sophisticated modelling. Also, Fig. 2 shows some tendency towards underestimation of β for 40,000 simulations. Moreover, if there are family relations, linkage becomes an issue.

Artefacts like drop-out and drop-in give no principal problems for our regression approach. Drop-out and drop-in become a part of the noise. Obviously, if the noise dominates the signal, the resulting output from the model will be of little use. These effects can be studied using simulation; performing experiments that mimics real forensic cases involving degraded DNA is more difficult.

The hypotheses have been formulated in terms of parameters and tested based on the conventional *p*-value rather than the likelihood ratio or random match probability used in most forensic applications. Obviously, there are alternatives to *p*-values also within our framework. The problem can be set up in a Bayesian framework with priors on the hypotheses and then the posteriors can be calculated. Also likelihood ratios can be calculated. However, we have preferred the simple approach most people are used to from other applications. It can be argued that the cut-off to use for *p*-values, the significance level, is arbitrary. While this is true, the problem is more pronounced for *LR* values. Several publications have recommended verbal translations for *LR*-values, but we are not aware of any justification for a particular value; why $LR \geq 10{,}000$ or $LR \geq 100{,}000$. The advantage of *p*-values are that they are scaled and can be interpreted as probabilities. However, with some few recent exceptions including [6,13], forensic evidence has not been summarised using *p*-values. For this reason, the use of *p*-values could well be challenged in court applications. However, we maintain that *p*-values are relevant for several reasons. In a large number of cases, numerical evidence is not presented to the court. This applies typically to complex cases. Rather the reporting officer bases his opinions on the calculations done and in this case *p*-values may be better suited than LR-s. A similar comment applies to non-court applications. Then, typically there is a general understanding of *p*-values whereas LR-s are less well known and, as indicated above, there is no justification for the thresholds adopted.

A *p*-value can be small because there is a large effect or because the sample size is large. Therefore significance can be reached when the effect is minor just because the sample size is large. In our case, any fraction above 0 leads to the same conclusion so this problem is not as pronounced for the forensic applications we have in mind. It could be helpful to accompany the *p*-value with a confidence interval for the fraction contributed. The length of the confidence interval reflects the amount of markers used. The β estimates are of interest in their own right. In forensic cases, they can give some indication of consistency, different sets of markers should give similar estimates. For pooled data, β values can be used to check that the different contributors contribute roughly similar amounts. Otherwise, estimates of allele frequencies from pooled data need to be adjusted.

Some extensions have already been discussed. There are other directions for future work. Most importantly, we would like to extend the model and the implementation to be able to handle several genotyped individuals. This extension seems straightforward, but there is a need for more experiments, preferably with different marker sets, before conclusions can be drawn. However, based on the experience so far, we find the model and the approach to be promising.

Acknowledgement

The work leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 285487 (EUROFORGEN-NoE).

References

[1] J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, FL, USA, 2005.

[2] P.M. Schneider, Beyond STR-s: the role of diallelic markers in forensic genetics, Transfus. Med. Hemother. 39 (3) (2012) 176–180.

[3] C. Børsting, E. Rockenbauer, N. Morling, Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard, Forensic Sci. Int. Genet. 4 (1) (2009) 34–42.

[4] C. Phillips, R. Fang, D. Ballard, M. Fondevila, C. Harrison, F. Hyland, E. Musgrave-Brown, C. Proff, E. Ramos-Luis, B. Sobrino, et al., Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel, Forensic Sci. Int. Genet. 1 (2) (2007) 180–185.

[5] L. Voskoboinik, A. Darvasi, Forensic identification of an individual in complex DNA mixtures, Forensic Sci. Int. Genet. 5 (5) (2011) 428–435.

[6] N. Homer, S. Szelinger, M. Redman, Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, PLoS Genet. 4 (8) (2008) e1000167.

[7] T. Egeland, A.E. Fonneløp, P.R. Berg, M. Kent, S. Lien, Complex mixtures: a critical examination of a paper by Homer et al., Forensic Sci. Int. Genet. 6 (1) (2012) 64–69.

[8] A.E. Fonneløp, Applicability of High-Density Genome-Wide SNP Arrays in Forensics, (Master's Thesis), Norwegian University of Life Sciences, Department of Animal and Aquacultural Sciences, 2010, http://arken.umb.no/~theg/master-ElidaF2010-1.pdf.

[9] D.J. Balding, R.A. Nichols, A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, Genetica 96 (1995) 3–12.

[10] M.W. Perlin, B. Szabady, Linear mixture analysis: a mathematical approach to resolving mixed DNA samples, Forensic Sci. Int. Genet. 46 (6) (2001) 1372–1378.

[11] J.G. Shewale, R.H. Liu, Forensic DNA Analysis: Current Practices and Emerging Technologies, CRC Press, Boca Raton, Florida, U.S., 2013.

[12] C. Phillips, Applications of autosomal SNPs and indels in forensic analysis, Forensic Sci. Rev. 1 (2) (2012).

[13] G. Dørum, Ø. Bleka, P. Gill, H. Haned, L. Snipen, S. Sæbø, T. Egeland, Exact computation of the distribution of likelihood ratios with forensic applications, Forensic Sci. Int. Genet. 9 (2013) 93–101.

Paper II

Int J Legal Med (2016) 130:323–329
DOI 10.1007/s00414-015-1276-1

ORIGINAL ARTICLE

Relationship inference based on DNA mixtures

Navreet Kaur[1] · Mariam M. Bouzga[2] · Guro Dørum[1] · Thore Egeland[1]

Received: 29 June 2015 / Accepted: 13 October 2015 / Published online: 5 November 2015

Abstract Today, there exists a number of tools for solving kinship cases. But what happens when information comes from a mixture? DNA mixtures are in general rarely seen in kinship cases, but in a case presented to the Norwegian Institute of Public Health, sample DNA was obtained after a rape case that resulted in an unwanted pregnancy and abortion. The only available DNA from the fetus came in form of a mixture with the mother, and it was of interest to find the father of the fetus. The mother (the victim), however, refused to give her reference data and so commonly used methods for paternity testing were no longer applicable. As this case illustrates, kinship cases involving mixtures and missing reference profiles do occur and make the use of existing methods rather inconvenient. We here present statistical methods that may handle general relationship inference based on DNA mixtures. The basic idea is that likelihood calculations for mixtures can be decomposed into a series of kinship problems. This formulation of the problem facilitates the use of kinship software. We present the freely available `R` package `relMix` which extends on the `R` version of `Familias`. Complicating factors like mutations, silent alleles, and θ-correction are then easily handled for quite general family relationships, and are included in the statistical methods we develop in this paper. The methods and their implementations are exemplified on the data from the rape case.

Keywords DNA mixtures · Kinship analysis · Unknown reference profiles · Likelihood ratios

Introduction

While solving kinship and paternity cases, it is rather unusual to make use of DNA mixtures. In such cases, buccal swabs or personal items are normally used as reference samples, generally collected or taken at a specific time, and mixtures are therefore rarely seen. However, in a case handled by the Norwegian Institute of Public Health, a paternity case was to be solved based on a DNA mixture from a mother and a fetus, and data from a specified, genotyped man. The alleged father was a suspect after a rape case, which resulted in an unwanted pregnancy followed by an abortion. Sample DNA was extracted from a uterine curettage, and the only available DNA from the fetus came in

✉ Navreet Kaur
navreet.kaur@nmbu.no

Mariam M. Bouzga
mariam.bouzga@fhi.no

Guro Dørum
guro.dorum@nmbu.no

Thore Egeland
Thore.Egeland@nmbu.no

[1] Norwegian University of Life Sciences, 1432 Aas, Oslo, Norway

[2] Department of Forensic Biology, Norwegian Institute of Public Health, PB 4404 Nydalen, 0403, Oslo, Norway

form of a mixture with the mother, the victim. Reference data from the mother was unavailable as the mother refused to give her reference sample. New methods and software was therefore needed.

The scenario presented may seem unusual, and of course, courts in different countries may have the judicial rights to act in a different manner. However, the example here also illustrates another set of problems that may arise while handling mixture samples; the samples may show that there is a relationship between the persons involved and it may then be of interest to determine the relationship between the contributors to the mixtures, or relatives of the contributors.

There is a large literature on DNA mixtures and crime cases, and the general methods are summarized in textbooks like [5, 8]. Most cases and most papers assume the contributors to be unrelated, but there are exceptions [7, 9]. Whereas these papers focus on determining contributors to the mixture, our focus is on determining the relationships between contributors to a mixtures (or relatives of contributors). From a statistical point of view, some of the calculations needed to solve the problems we address can also be solved based on methods and implementations presented in the mentioned papers.

Software packages for DNA identification, like DNA-view [1] and Familias 3 [2], are available for general genetic testing and relationship inference. But our main example case presents a more complicated scenario than cases usually handled by these softwares: not only is the reference from the mother unavailable, but the reference from the child is also found in a mixture with the mother. In relationship inference, there is also a tradition of including mutations, θ-corrections, and silent alleles. The models developed in this paper and associated freely available software makes it possible to handle DNA mixtures and at the same time incorporate such complicating factors. We present the R [7] package relMix. As far as we know, there is no freely available software that may handle relationship inference involving mixtures in complex pedigrees that also accounts for mutations, θ-correction, and silent alleles.

Material and methods

Data

As mentioned, the cases we have in mind will be based both on reference profiles (typically of good quality) and DNA mixtures. We develop the methods for discrete data (i.e., only allele designations are used), as is common for relationship inference. It is sufficient to explain our method and approach for one marker as we assume that the markers are *independent*, i.e., markers are assumed to be in *linkage disequilibrium* and *unlinked*. The latter assumption is not needed for our main example, but is generally required for larger pedigrees. Throughout this paper, we denote the mixture sample by E and let the reference profiles for the N individuals involved be denoted by $g_1, g_2, ..., g_N$.

Statistical methods

Basic method

In general, there could be several hypotheses $H_1, \ldots, H_T$, each corresponding to a specific family relationship. For our main case, only two competing hypotheses will be considered, and the hypotheses are H_1: the alleged father is the father of the child, and H_2: the alleged father is an unrelated man. Figure 1 gives a picture of the relationship and the hypotheses to be tested. The hypotheses specify the number of contributors. Typically, K of these will not be genotyped, and in our methods we condition on their genotypes $u_1, \ldots, u_K$. In the main example, the mother and the child are not genotyped and therefore $K = 2$.

The evidence is then summarized by a likelihood ratio (LR), given by

$$LR = \frac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_2)}. \tag{1}$$

The basic calculation involves computing

$$\begin{aligned} P(\text{data} \mid H_t) &= P(E, g_1, \ldots, g_N \mid H_t) \\ &= \sum_{u_1,\ldots,u_K \in A_{H_t}} P(u_1, \ldots, u_K, g_1, \ldots, g_N \mid H_t), \end{aligned} \tag{2}$$

where A_{H_t} is the set of possible genotypes for the untyped individuals under hypothesis H_t. Note that E is a subset of the union of alleles in A_{H_t} and those from the known contributors. Each term of the above formula can be calculated using software implementing likelihood calculations for pedigrees. Pedigrees can then be arbitrary, possibly involving inbreeding, and artefacts like mutation, θ-correction, and silent alleles can be accommodated as explained in the next section. As we assume that markers are independent, the overall likelihood is obtained by multiplying across markers.

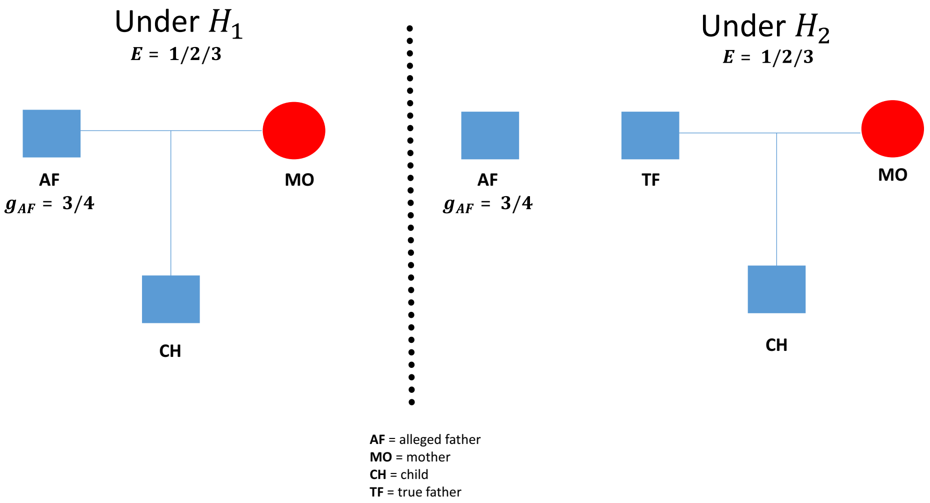


Fig. 1 Pedigrees for the main case. H_1: the alleged father (AF) is the father of the child (*left pedigree*), and H_2: the alleged father is an unrelated man (*right pedigree*)

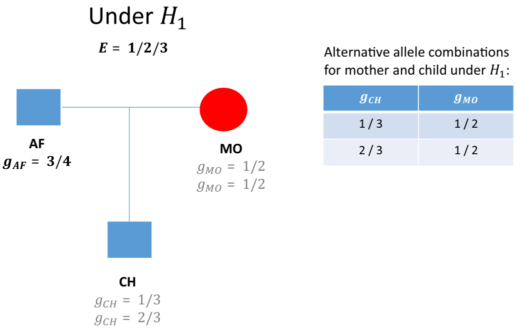
Example 1: an illustration

Consider the relationship shown in Fig. 1, where $E = 1/2/3$ and the genotype of the alleged father is $g_{\text{AF}} = 3/4$. In order to be consistent with the alleles found in the mixture and the alleged father (AF) being the father, the alternative allele combinations for the child and the mother must be as shown in Fig. 2 (mutations are for now disregarded). Let A_{H_1} be the set of all possible genotypes of the mother and the child under H_1. From Fig. 2, we see that the set A_{H_1} must be given by

$$A_{H_1} = \{(g_{\text{CH}}, g_{\text{MO}})\} = \{(1/3, 1/2), (2/3, 1/2)\}.$$

Following the generic form given in Eq. 2, we can now show that

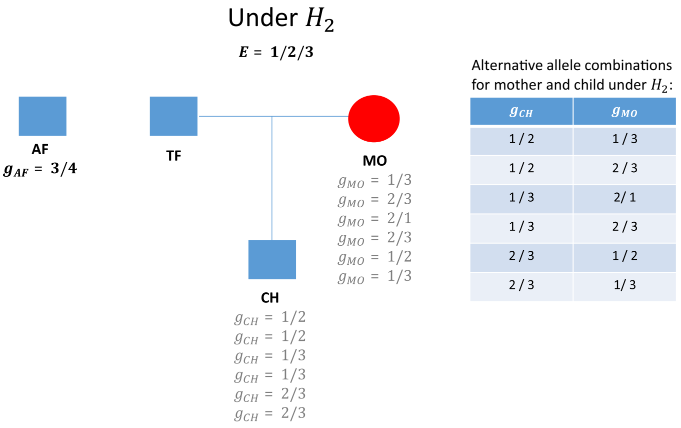
$$\begin{aligned}
P(\text{data} \mid H_1) &= P(E, g_{AF} \mid H_1) \\
&= \sum_{g_{CH}, g_{MO} \in A_{H_1}} P(g_{CH}, g_{MO}, g_{AF} \mid H_1) \\
&= \sum_{g_{CH}, g_{MO} \in A_{H_1}} P(g_{CH} \mid g_{MO}, g_{AF}, H_1) \\
&\quad \times P(g_{MO}) P(g_{AF}) \\
&= P(g_{CH} = 1/3 \mid g_{MO} = 1/2, g_{AF} = 3/4) \\
&\quad \times P(g_{MO} = 1/2) P(g_{AF} = 3/4) \\
&+ P(g_{CH} = 2/3 \mid g_{MO} = 1/2, g_{AF} = 3/4) \\
&\quad \times P(g_{MO} = 1/2) P(g_{AF} = 3/4) \\
&= 2 p_1 p_2 p_3 p_4.
\end{aligned}$$



Alternative allele combinations for mother and child under H_1:

g_{CH}	g_{MO}
1 / 3	1 / 2
2 / 3	1 / 2

Fig. 2 Pedigree under H_1 with alternative allele combinations of mother and child



Alternative allele combinations for mother and child under H_2:

g_{CH}	g_{MO}
1 / 2	1 / 3
1 / 2	2 / 3
1 / 3	2 / 1
1 / 3	2 / 3
2 / 3	1 / 2
2 / 3	1/ 3

Fig. 3 Pedigree under H_2 with alternative allele combinations of mother and child

Let now A_{H2} be the set of possible genotypes of the mother and the child under H_2. Finding A_{H2} is a bit more complicated. For $E = 1/2/3$ and $g_{\mathrm{AF}} = 3/4$, we have that both the mother and the child must be heterozygous. Figure 3 shows the six alternative allele combinations under H_2, giving

$$\begin{aligned}A_{H_2} = \{&(1/2, 1/3), (1/2, 2/3), (1/3, 2/1), (1/3, 2/3),\\ &(2/3, 1/2), (2/3, 1/3)\}.\end{aligned}$$

We find that the denominator in the likelihood ratio is given by

$$\begin{aligned}P(\text{data} \mid H_2) &= P(E, g_{AF} \mid H_2)\\ &= \sum_{g_{MO}, g_{CH} \in A_{H2}} P(g_{CH}, g_{MO}, g_{AF} \mid H_2)\\ &= \sum_{g_{MO}, g_{CH} \in A_{H2}} P(g_{CH} \mid g_{MO}, H_2) P(g_{MO}) P(g_{AF})\\ &= 12 p_1 p_2 p_3^2 p_4.\end{aligned}$$

Finally, we find that the likelihood ratio in this case is given by

$$LR = \frac{P(E, g_{\mathrm{AF}} \mid H_1)}{P(E, g_{\mathrm{AF}} \mid H_2)} = \frac{2 p_1 p_2 p_3 p_4}{12 p_1 p_2 p_3^2 p_4} = \frac{1}{6 p_3}.$$

Mutations, θ-correction, silent alleles

We made several assumptions to derive the likelihood ratio in Eq. 1. We here generalize the method presented by including mutations, θ-correction and silent alleles in our basic model.

Mutations A mutation model is specified by a mutation matrix,

$$M = \begin{bmatrix} m_{11} & \dots & m_{1K} \\ m_{21} & \dots & m_{2K} \\ \vdots & \ddots & \vdots \\ m_{K1} & \dots & m_{KK} \end{bmatrix}$$

where m_{ij} is the probability that allele i in the parent end up as allele j in the child. Generally, we have that $0 \leq m_{ij} \leq 1$ and $\sum_{i=1}^{K} m_{ij} = 1$. There are four mutation models available, denoted 'Equal,' 'Proportional,' 'Stepwise,' and 'Custom' [2, 3]. Mutations involving silent alleles are not modeled in the models we present. Choosing the 'Equal' model, the probability of mutation from one allele to another is set to be equal for all alleles, given that a mutation occurs. With a 'Proportional' mutation model, the probability of mutation of an allele is proportional to its frequency. The 'Stepwise' model divides the mutations into two types: all alleles adding or subtracting an integer to the allele, and all

Table 1 The possible genotypes and likelihoods for the silent case when AF is the father (H_1)

	MO	CH	AF	$P(data \mid H_1)$
1	A/A	A/A	A/A	p_A^4
2	A/A	A/A	A/S	$p_A^3 p_S$
3	A/A	A/S	A/A	0
4	A/A	A/S	A/S	$p_A^3 p_S$
5	A/S	A/A	A/A	$p_A^3 p_S$
6	A/S	A/A	A/S	$p_A^2 p_S^2$
7	A/S	A/S	A/A	$p_A^3 p_S$
8	A/S	A/S	A/S	$2 p_A^2 p_S^2$

Table 2 Results from using our statistical methods on the case data

Marker	E	g_{AF}	Basic LR	$LR_{\theta = 0.01}$	$LR_{mut = 0.1\ \%}$
D3S1358	14/16/18	15/18	1.1821	1.1477	1.1816
TH01	6/7/9	9/9.3	1.1827	1.1482	1.1818
D21S11	28/30.2	28/30	1.7611	1.6310	1.7633
D18S51	13/15/16	14/16	1.4355	1.3740	1.4351
D2S1338	20/23/25	18/23	1.8772	1.7535	1.8751
D1S1656	11/12	12/13	1.8669	1.7326	1.8745
VWA	16/18	16/17	1.1573	1.1246	1.1574
D8S1179	12/13	12/16	1.5394	1.4736	1.5450
FGA	21/26	21/22	1.8216	1.6801	1.8260
D19S434	12/13/15	13/14	0.6859	0.6851	0.6867
Total LR			27.5745	17.1158	27.8528
Total posterior			0.9650	0.9448	0.9653

We find the basic LR, the LR correcting for mutations (mutation rate is 0.1 % for both males and females), and the LR including kinship correction (with $\theta = 0.01$)

alleles adding or subtracting a fractional amount, see [3]. By using the 'Custom' model the user may specify the mutation model herself. Different mutation matrices follow these models, and for the 'Equal' model the mutation matrix is given by

$$M = \begin{bmatrix} 1-R & \frac{R}{N-1} & \cdots & \frac{R}{N-1} \\ \frac{R}{N-1} & 1-R & \cdots & \frac{R}{N-1} \\ \vdots & & \ddots & \vdots \\ \frac{R}{N-1} & \frac{R}{N-1} & \cdots & 1-R \end{bmatrix}$$

θ-correction To account for population stratification and relatedness, we can introduce a θ-parameter to our calculations. In paternity cases for instance, Hardy-Weinberg will not apply in cases where the parents are related in a way not specified by the pedigree. The θ-correction essentially corrects for relatedness of alleles with common ancestry. To see how the θ-correction works, consider an allele A with frequency p_A and assume that we have sampled n alleles, where x alleles are of type A. With coancestry coefficient θ, the probability that the next allele will be of type A is

$$\frac{x\theta + (1-\theta)p_A}{1+(n-1)\theta},$$

see [12].

Silent alleles Silent alleles may be present in cases where the individuals tested for relatedness are (apparently) homozygous. The alleles may fail to amplify, and individuals are mistakenly assumed to be homozygous. The term null alleles has also been used for such cases. However, silent alleles can be modeled by modifying our calculations: we let the silent allele frequency and the frequencies of the other alleles sum to 1. If there is a possibility of a silent allele S with frequency p_S, an apparently homozygous A/A could be A/S and the genotype probability is $P(A/S) = p_A^2 + 2p_A p_S$. The following example shows how silent alleles may be handled for a paternity case.

Example 2—a silent case

Consider first a paternity case where the alleged father (AF), mother (MO), and child (CH) all have genotype A/-, and so there is a silent allele S with frequency p_S.

From Table 1,

$$\begin{aligned} L_1 &= P(data \mid H_1) = p_A^4 + 4p_A^3 p_S + 3p_A^2 p_S^2 \\ &= p_A^2(p_A^2 + 4p_A p_S + 3p_S^2) \end{aligned}$$

When A_F is not the father, a similar argument gives

$$L_2 = P(data \mid H_2) = p_A^2(p_A^2 + 3p_A p_S + p_S^2)(p_A + 2p_S)$$

From this,

$$LR = \frac{p_A^2 + 4p_A p_S + 3p_S^2}{(p_A^2 + 3p_A p_S + p_S^2)(p_A + 2p_S)} = 6.82.$$

However, for the mixture problem, the answer will be slightly different as we then also could have $g_{CH} = S/S$ or $g_{MO} = S/S$ (but not both).

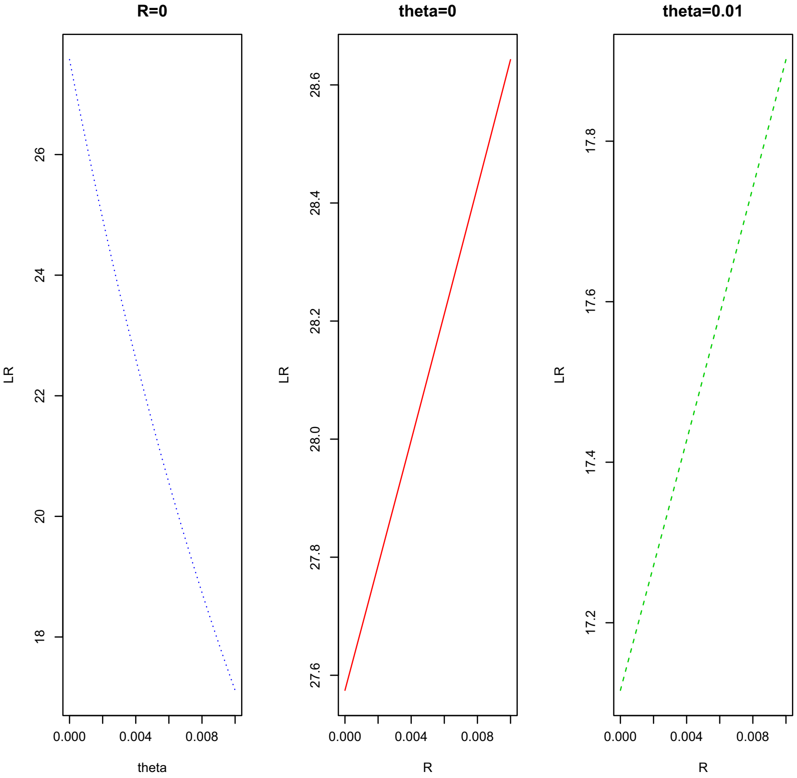


Fig. 4 Plots showing how the LR is affected for 1) different θ values with mutation rate $R = 0$ (*plot to the left*), 2) different mutation rates (R) and $\theta = 0$ (*plot in the middle*), and 3) when the mutation rate changes and θ is kept at 0.01 (*plot to the right*)

Results

In our introductory case, data was first collected using the SGM Plus kit, giving data for ten genetic short tandem repeat (STR)-markers. The data was later expanded by analyzing the fetus material using ESX17, giving 16 genetic markers.

Table 2 contains the results for our main example. The calculations are done using the package `relMix` developed in `R`, and is freely available with documentation from http://arken.umb.no/~nkaur/relMix_1.0.zip. We found the basic LR, the LR with mutations incorporated (mutation rate of 0.1 % for both males and females), and a LR that considers θ-correction ($\theta = 0.01$). The total basic LR for all markers in the data was found to be 27.6, suggesting that it is 27.6 times more likely to observe the data given H_1 compared to H_2. In other words, the total basic LR value supports the use of H_1 (see "Discussion" section). The mutations do not seem to affect the LR much, but if we correct for general kinship, the change in LR is more evident. If more markers were available, including the mother's genotype, clearer conclusions could be reached. Silent alleles are not considered here as there is no homozygosity for any of the markers found in the data. Figure 4 shows how the LR is affected for different mutation rates and θ-values.

Discussion

In this paper, we have presented statistical methods that may handle general relationship inference involving DNA mixtures. The methods are based on likelihood calculations, and the evidence is summarized by calculating the likelihood ratio (LR) comparing two hypotheses. Whereas the main emphasis for solving cases involving DNA mixtures often is to determine the contributors to the mixture, we here instead focus on the relationship between the contributors to the mixture. We draw conclusions based on likelihood ratios. The methods developed were used on the data from our main case to test H_1: the alleged father is the father of the child, versus H_2 : the alleged father is an unrelated man. In the "Results" section, we found the LR for this case to be

27.6, and we drew a conclusion in support of H_1 (the alleged father being the father of the child). A conclusion based on such a small LR is not obvious, but a discussion on this is not a topic for this paper.

We have shown that our calculations can be extended to consider complicating factors like mutations, θ-correction, and silent alleles. For the data at hand, including a mutation rate of 0.1 % did not influence the total likelihood ratio noticeably (see Table 2). However, if the genotype of the alleged father (g_{AF}) for marker D19S434 had been, say 14/14, the total LR using the basic model would be 0. This emphasizes the importance of allowing for mutations. Including a θ-correcting factor had a greater impact on the LR than the mutation rate. We note that ignoring kinship in the calculations leads to overestimation of the evidence, giving larger LR values for $\theta = 0$ than $\theta > 0$ (see Fig. 4). We therefore suggest that the θ-correction should be included in the calculation. We have also modified our calculations so that silent alleles (S) may be handled. However, the presence or absence of silent alleles may be determined using a kit of different primers [5, 6].

Our main example involved a simple pedigree with a child, mother, and an alleged father. Our methods may also handle complicated pedigrees, possibly involving inbreeding. The family relations could be more complex; for instance, the DNA profile of the alleged father may not be available, but rather genotypes of some of his known relatives may be available. The methods presented here may handle such more complicated cases, and is implemented in the software. Also, several alternative hypotheses can be tested and compared. For instance, it could be that one would like to test an alternative hypothesis stating that the alleged father's brother is the father of the child.

The framework presented can be extended in several directions. Other complicating factors and artefacts like dropouts/dropins, linkage, and linkage disequilibrium can be included. Another suggestion could be to look at a continuous model, extending the model presented to incorporate peak heights of the genotyped data using a continuous approach, see [13]. An alternative to our approach is to use Bayesian networks. [11] describes how complex problems of relationship testing using DNA profiles can be modeled using Bayesian networks instead. The Bayesian network can further be extended to handle mutations, linkage, and linkage disequilibrium between STR markers as described in [10].

References

1. Brenner C DNA-view. http://dna-view.com/dnaview.htm
2. Kling D, Tilmar A, Egeland T (2014) Familias 3—extensions and new functionality. Forensic Sci Int Genet 13:121–127
3. Mostad P, Egeland T (2015) Familias, Probabilities for Pedigrees Given DNA Data. http://cran.r-project.org/web/packages/Familias, R version 2.2
4. Butler JM (2011) Advanced Topics in Forensic DNA Typing: Methodology. Academic Press
5. Buckleton JS, Triggs CM, Walsh SJ (2005) Forensic DNA evidence interpretation. CRC press
6. Team R (2014) Core: R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012, Open access available at: http://cran.r-project.org
7. Egeland T, Dørum G, Vigeland MD, Sheehan NA (2014) Mixtures with relatives: a pedigree perspective. Forensic Sci Int Genet 10:49–54
8. Fung WK, Hu Y-Q (2008) Statistical DNA forensics: theory, methods and computation. Wiley, Statistics in practice
9. Hu Y-Q, Fung WK, Choy YT (2011) Interpreting DNA mixtures with relatives of a missing suspect. In: Remote Sensing, Environment and Transportation Engineering (RSETE), 2011 International Conference on. IEEE, pp 7649–7652
10. Kling D, Egeland T, Mostad P (2012) Using object oriented bayesian networks to model linkage, linkage disequilibrium and mutations between STR markers. PloS one 7(9): e43873
11. Dawid AP, Mortera J, Vicard P (2007) Object-oriented Bayesian networks for complex forensic DNA profiling problems. Forensic Sci Int 169(2):195–205
12. Balding DJ (2005) Weight-of-evidence for Forensic DNA Profiles. Statistics in practice. Wiley
13. Steele CD, Balding DJ (2014) Statistical evaluation of forensic DNA profile evidence. In: Annual Review of Statistics and Its Application, vol 1, pp 361–384

Paper III

Int J Legal Med (2017) 131:629–641
DOI 10.1007/s00414-016-1526-x

ORIGINAL ARTICLE

Pedigree-based relationship inference from complex DNA mixtures

Guro Dørum[1] · Navreet Kaur[2] · Mario Gysi[1]

Received: 14 July 2016 / Accepted: 28 December 2016 / Published online: 19 January 2017

Abstract We present a general method for analysing DNA mixtures involving relatives that accounts for dropout and drop-in, mutations, silent alleles and population substructure. Whether the aim is to identify the contributors to a mixture who may be related, or to determine the relationship between individuals based on a DNA mixture, both types of problems can be handled by the method and software presented here. We focus on the latter scenario, motivated by non-invasive prenatal paternity testing where the profile of the child is available only in the form of a mixture with the mother's profile. Relationships are represented by pedigrees and can include kinship between more than two individuals. The software is freely available as a graphical user interface in the R package `relMix`.

Keywords DNA mixtures · Kinship · Likelihood ratio · Dropout · Drop-in · Mutations · Non-invasive prenatal paternity testing · NGS

Introduction

There is an increasing demand for analysis of DNA mixtures that involve relatives, both in criminal cases and in relationship inference, and so there is a need for methods and software that can handle this type of cases. Since relatives are likely to share more alleles than unrelated individuals, the result of ignoring this relationship may be an overestimation of the weight of evidence. One example of this type of cases is non-invasive prenatal paternity testing. In this new application, foetal cell-free DNA that is present at low levels in a pregnant woman's blood is analysed by sequencing STRs using massive parallel sequencing (MPS) [5]. As the vast majority of the blood's cell-free DNA originates from the mother herself, foetal DNA can only be accessed through a highly unbalanced mother-child mixture. As an example, Lo et al. [11] measured on average 3.4 % foetal fraction of total cell free DNA in early pregnancy and 6.2 % in late pregnancy, but this may vary significantly. Another challenge is that cell-free DNA is heavily degraded, and the abundance of the child's DNA is very low, so artefacts such as dropout and drop-in alleles are likely to appear.

DNA mixtures with relatives have previously been discussed in the literature [4, 13, 14]. These publications do however only address pairwise relationships. Egeland et al. [3] described a method to handle mixtures with general family relationships but did not consider artefacts such as dropout and drop-in that may result in partial profiles. In addition, the aforementioned papers focus on determining the contributors to a mixture. Mortera et al. [12] presented a mixture-based paternity case analysed with an approach that included deconvolution of the mixture followed by kinship testing with the resulting profiles.

Kaur et al. [9] introduced a pedigree-based approach to relationship inference-based on DNA mixtures. Here, we present an extension of their work to also account for dropout

✉ Guro Dørum
guro.dorum@irm.uzh.ch

[1] Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland

[2] Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Aas, Norway

and drop-in. In addition, we introduce a user friendly software. Our approach can handle general relationships described by pedigrees, with any number of contributors to the mixture and any number of relatives. The kinship calculations are based on the R version of Familias (http://www.familias.name), which allows incorporation of complicating factors like mutations, silent alleles and population substructure. The model can also handle multiple replicates. While our main focus is determining relationships between individuals based on mixtures, we also present an example where the aim is to determine the contributors to a mixture in a crime case setting. The software is freely available in the R package relMix that can be found on CRAN (https://cran.r-project.org), both in the form of a graphical user interface and as command line functions for more flexible use.

We present detailed calculations for some simple examples and demonstrate the effect of dropout and drop-in in mixture-based paternity cases in a simulation study. Finally, the method is used to analyse real data in a prenatal paternity case.

Methods

Motivational example

As a motivational example, we will use a fictional prenatal paternity case. A DNA profile of the child (CH) is only available in the form of a mixture with the mother (MO). Since the mother is the major contributor, we can assume that all her alleles are in the mixture (i.e. no dropout), while some of the child's alleles may have dropped out from the mixture with a certain probability. Reference profiles for the mother and the alleged father (AF) are available. We consider the two hypotheses

H_1: AF is the father of CH
H_2: Someone unrelated to AF is the father of CH

Let E denote all available evidence; in this case, the mixture alleles and the reference profiles of MO and AF. We summarise the evidence by computing the likelihood ratio

$$LR = \frac{P(E \mid H_1)}{P(E \mid H_2)}. \tag{1}$$

The model

Each likelihood in Eq. 1 can loosely be described as involving terms of the kind $P(\text{mixture} \mid \text{contributor genotypes})$ and $P(\text{genotypes} \mid \text{relationship})$. The first term concerns only the probability of the mixture conditioned on the genotypes of the contributors, while the second term concerns the probability of the genotypes conditioned on the relationship. In the following sections, we present calculations for the mixture term and the kinship term.

Mixture model

We adopt the mixture model described in Haned et al. [6, Appendix]. This is a semi-continuous model that accounts for dropout and drop-in in the likelihood calculations but does not consider peak heights. We will try to stay close to the notation in Slooten [13]. For simplicity, we will concentrate on one marker, but since the markers are assumed to be independent, the total likelihood is simply the product of the per-marker likelihoods.

Let $g_i = (a_{i,1}, a_{i,2})$ denote the genotype of mixture contributor i. Define the vector $\mathbf{g} = (g_1, ..., g_n)$ to contain the genotypes of all n contributors to the mixture, where n is assumed known. We further define the mixture as a random variable denoted by $\mathcal{M}$. Each contributor i has a specific dropout probability $0 \le d_i \le 1$ for a heterozygous allele, and a dropout probability D_i for a homozygous allele, where it is usually assumed that $D_i \le d_i^2$. For convenience, we will use $D_i = d_i^2$ throughout the paper, so we only have to specify one dropout probability per contributor. Let the vector $\mathbf{d} = (d_1, ..., d_n)$ contain dropout probabilities for all n contributors. Further, we define a drop-in parameter c. An allele a drops in with probability cp_a, where p_a is the frequency of allele a. Since the frequencies for all alleles in a locus sum to 1, we can regard c as the expected number of drop-in alleles per locus. Let $n_{i,a} = \{0, 1, 2\}$ denote the number of times allele a is observed in the genotype of contributor i. For each allele a in the locus, the probability that it will not appear in the mixture is

$$P(a \notin \mathcal{M} \mid \mathbf{g}, \mathbf{d}, c) = (1 - cp_a) \prod_i d_i^{n_{i,a}}, \tag{2}$$

and the probability that it will appear in the mixture is thus

$$P(a \in \mathcal{M} \mid \mathbf{g}, \mathbf{d}, c) = 1 - (1 - cp_a) \prod_i d_i^{n_{i,a}}. \tag{3}$$

Note that a dropout probability of 0 for a contributor that has the allele a assures that the mixture will contain this allele with probability 1. The probability of observing a set M of mixture alleles is

$$P(\mathcal{M} = M \mid \mathbf{g}, \mathbf{d}, c) = \prod_{a \notin M} P(a \notin M \mid \mathbf{g}, \mathbf{d}, c) \cdot \prod_{a \in M} P(a \in M \mid \mathbf{g}, \mathbf{d}, c). \tag{4}$$

By considering all alleles a in the locus, we account for the probability that an allele that does not appear in any of the genotypes may have dropped in. Replicates are assumed to be conditionally independent given the parameters and the genotypes of the contributors.

Include kinship and sum over unknowns

If there are contributors in the mixture with unknown genotype, e.g. the child in the motivational example, we need to consider a set of possible genotypes U for these individuals. Let n_k and n_u be the number of known and unknown individuals in the mixture, respectively. Define the vectors $\mathbf{g_K} = (g_1, ..., g_{n_k})$ and $\mathbf{g_U} = (g_1, ...g_{n_u})$ to contain the genotypes of the known and unknown contributors. Let $u \in U$, then $\mathbf{g_U} = u$ is one possible set of genotypes for the unknown contributors. Further, let $\mathbf{g_A} = (g_1, ..., g_{n_A})$ be a vector of genotypes for the n_A additional genotyped individuals in the pedigree who are not part of the mixture. For the kinship part, we need to consider the probability $P(\mathbf{g_A}, \mathbf{g_K}, \mathbf{g_U} \mid H_j)$, where H_j specifies the relationship between all individuals. We can now model each likelihood in Eq. 1 to include both the probability of the mixture and the kinship as

$$P(E \mid H_j) = \sum_{u \in U} P(\mathcal{M} = M \mid \mathbf{g_K}, \mathbf{g_U} = u, \mathbf{d}, c) \cdot P(\mathbf{g_A}, \mathbf{g_K}, \mathbf{g_U} = u \mid H_j). \quad (5)$$

Note that different H_j's may specify different contributors and family relationships, in which case also the genotype vectors will change depending on the hypothesis. This can be specified by adding the subscript j to these vectors.

Model demonstrated on motivational example

We will use the motivational example to illustrate the model. Assume a diallelic marker with alleles 1 and 2, and frequencies p_1 and p_2. The mother's genotype is $g_{MO} = 1/1$ and the alleged father's genotype is $g_{AF} = 1/1$. The observed mother-child mixture is $M = 1/2$. We have $\mathbf{g_K} = (g_{MO})$, $\mathbf{g_A} = (g_{AF})$ and $\mathbf{g_U} = (g_{CH})$, where the set of possible genotypes for the child is $U = \{1/1, 1/2, 2/2\}$. We set the mother's dropout probability to 0 since her DNA is present in high quantity, while for the child, we assume dropout probability d, and hence $\mathbf{d} = (0, d)$.

As an example, consider $\mathbf{g_U} = (1/2)$. According to Eq. 2, the probabilities of not observing alleles 1 and 2 in the mixture is

$$P(1 \notin \mathcal{M} \mid \mathbf{g_K}, \mathbf{g_U}=(1/2), \mathbf{d}, c) = (1 - cp_1) \cdot 0^2 \cdot d = 0,$$
$$P(2 \notin \mathcal{M} \mid \mathbf{g_K}, \mathbf{g_U}=(1/2), \mathbf{d}, c) = (1 - cp_2) \cdot 0^0 \cdot d = d - dcp_2.$$

and the probabilities of observing these alleles in the mixture is thus

$$P(1 \in \mathcal{M} \mid \mathbf{g_K}, \mathbf{g_U}= (1/2), \mathbf{d}, c) = 1,$$
$$P(2 \in \mathcal{M} \mid \mathbf{g_K}, \mathbf{g_U}= (1/2), \mathbf{d}, c) = 1 - d + dcp_2.$$

Since alleles 1 and 2 are the only two alleles for this marker, and they both appear in the mixture, the probability of the mixture according to Eq. 4 is

$$P(\mathcal{M} = 1/2 \mid \mathbf{g_K}, \mathbf{g_U} = (1/2), \mathbf{d}, c) = 1 - d + dcp_2.$$

Moving on to the kinship part, the probability under each hypothesis is

$$P(\mathbf{g_A}, \mathbf{g_K}, \mathbf{g_U}= (1/2) \mid H_1) = 0,$$
$$P(\mathbf{g_A}, \mathbf{g_K}, \mathbf{g_U} = (1/2) \mid H_2) = p_1^4 p_2.$$

Since the father is 1/1, he cannot be the father of the child (unless we consider mutations, which we will do in the next section).

The calculations for all three genotypes in U are summarised in Table 1. Note that the genotype 2/2 strictly could have been omitted from the table since it has likelihood 0 under both hypotheses, but we have included it as a generalisation for the next section where also mutations will be considered. Finally, the likelihood ratio becomes

$$LR = \frac{P(E \mid H_1)}{P(E \mid H_2)} = \frac{cp_2 \cdot p_1^4}{p_1^4 \cdot [cp_1 p_2 + (1 - d + dcp_2) \cdot p_2]} = \frac{c}{cp_1 + 1 - d + dcp_2}.$$

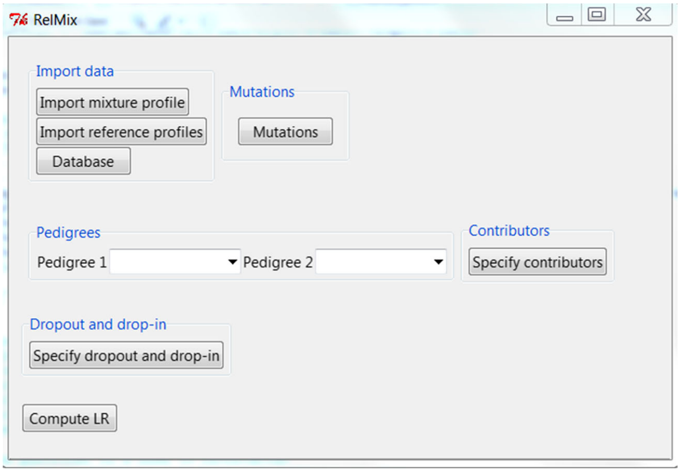
Software

The model described in the previous section is implemented in the R package relMix, freely available from http://cran.r-project.org/web/packages/relMix. Figure 1 shows a screen shot of the graphical user interface. The mixture profile, reference profiles and allele frequencies are read from files. Silent allele and minimum allele frequency can be specified. There are three built-in mutation models: 'equal', 'proportional' and 'stepwise' [2]. For flexibility with regard to the hypotheses, user-defined pedigrees can be supplied, and dropout probabilities can be specified per contributor. More details can be found in the user vignette that comes with the R package.

Table 1 Possible genotypes for the child in the motivational example, where $g_{MO} = 1/1$, $g_{AF} = 1/1$, and $M = 1/2$, with corresponding probability of kinship and probability of mixture. Mutations and silent alleles are not accounted for

$\mathbf{g_U}$	$P(\mathbf{g_A}, \mathbf{g_K}, \mathbf{g_U} \mid H_1)$	$P(\mathbf{g_A}, \mathbf{g_K}, \mathbf{g_U} \mid H_2)$	$P(\mathcal{M} = 1/2 \mid \mathbf{g_K}, \mathbf{g_U}, \mathbf{d}, c)$
1/1	p_1^4	p_1^5	cp_2
1/2	0	$p_1^4 \cdot p_2$	$1 - d + dcp_2$
2/2	0	0	$1 - d^2 + d^2 cp_2$

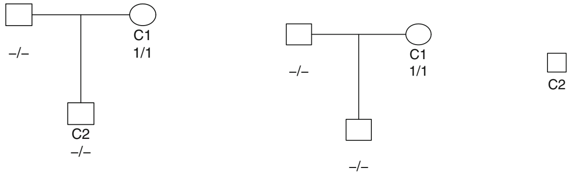
Fig. 1 The RelMix GUI



Mutations and silent alleles

In the previous section, we only considered dropout and drop-in as possible explanations for inconsistencies in the data. We will show how to incorporate mutations and silent alleles together with dropout and drop-in, by application to an example. For convenience, we have chosen to disregard population substructure in all examples in this paper, although theta correction is implemented in the method. Assume a two-person mixture with one known (C1) and one unknown (C2) contributor. We consider the hypotheses illustrated in Fig. 2. Note that these hypotheses are formulated differently compared to the motivational example. The main focus is now the contributors to the mixture (which may be related) rather than the relationship between the individuals. This is all just formalities, however, the procedure is the same as before. In this example, $\mathbf{g_K} = (g_{C1})$ and $\mathbf{g_U} = (g_{C2})$, while $\mathbf{g_A}$ is empty since there are no additional genotyped individuals involved. Consider a diallelic marker with alleles 1 and 2 and frequencies p_1 and p_2. The known contributor has genotype $g_{C1} = 1/1$, but may also be $1/s$ if we include a silent allele with frequency p_s. The mixture is $M = 1$. We assume possible dropout in both contributors. To factor in mutations, we use a mutation rate r and a mutation model that assumes all mutations to be equally likely, but that mutation to and from a silent allele is not possible [2]. All possible genotypes for the contributors, with corresponding kinship probability and mixture probability, are presented in Table 2. The silent allele is indifferent to both drop-in and dropout since we cannot see it in the mixture. Observe that if the mutation rate r, the silent allele frequency p_s and the dropout and drop-in values d and c are all set to 0, it reduces to a mixture model without artefacts, and the only possibility is that both contributors are 1/1.

Table 3 gives the values of the formulas in Table 2 when $p_1 = 0.4$, $p_2 = 0.5$ and $p_s = 0.1$. Kinship probabilities are calculated both without considering mutations, and with an equal probability mutation model with mutation rate 0.1. Mixture probabilities are calculated both without



(a) H_1: C1 and her child. (b) H_2: C1 and an unrelated individual.

Fig. 2 Example with mutations, silent allele, dropout and drop-in. H_1 and H_2 disagree on whether the second, unknown contributor C2 is related to C1 or not. (**a**) H_1: C1 and her child. (**b**) H_2: C1 and an unrelated individual

Table 2 Example with mutations, silent allele and dropout/drop-in. The first two columns give the possible genotypes for the two contributors, followed by the kinship probabilities under H_1 and H_2, and finally the mixture probability. We assume identical dropout probability d for both contributors

$\mathbf{g}_\mathbf{K}$	$\mathbf{g}_\mathbf{U}$	$P(\mathbf{g}_\mathbf{K}, \mathbf{g}_\mathbf{U} \mid H_1)$	$P(\mathbf{g}_\mathbf{K}, \mathbf{g}_\mathbf{U} \mid H_2)$	$P(\mathcal{M} = 1 \mid \mathbf{g}_\mathbf{K}, \mathbf{g}_\mathbf{U})$
1/1	1/1	$p_1^3(1-r)$	p_1^4	$(1-d^4(1-cp_1))(1-cp_2)$
1/s	1/1	$p_1^2 p_s(1-r)$	$2p_1^3 p_s$	$(1-d^3(1-cp_1))(1-cp_2)$
1/1	1/s	$p_1^2 p_s(1-r)$	$2p_1^3 p_s$	$(1-d^3(1-cp_1))(1-cp_2)$
1/s	1/s	$p_1 p_s(p_s(1-r)+p_1)$	$4p_1^2 p_s^2$	$(1-d^2(1-cp_1))(1-cp_2)$
1/1	s/s	0	$p_1^2 p_s^2$	$(1-d^2(1-cp_1))(1-cp_2)$
1/s	s/s	$p_1 p_s^2$	$2p_1 p_s^3$	$(1-d(1-cp_1))(1-cp_2)$
1/1	1/2	$p_1^2(p_2(1-r)+p_1 r)$	$2p_1^3 p_2$	$(1-d^3(1-cp_1))d(1-cp_2)$
1/s	1/2	$p_1 p_s(p_2(1-r)+p_1 r)$	$4p_1^2 p_s p_2$	$(1-d^2(1-cp_1))d(1-cp_2)$
1/1	2/2	$p_1^2 p_2 r$	$p_1^2 p_2^2$	$(1-d^2(1-cp_1))d^2(1-cp_2)$
1/s	2/2	$p_1 p_s p_2 r$	$2p_1 p_2^2 p_s$	$(1-d(1-cp_1))d^2(1-cp_2)$
1/1	2/s	$p_1^2 p_s r$	$2p_1^2 p_2 p_s$	$(1-d^2(1-cp_1))d(1-cp_2)$
1/s	2/s	$p_1 p_s(p_2+p_s r)$	$4p_1 p_2 p_s^2$	$(1-d(1-cp_1))d(1-cp_2)$

dropout/drop-in, and with a dropout probability of 0.1 and a drop-in value of 0.05. Table 3 can be used to compute several likelihood ratios for comparison. For example, if we do not consider any artefacts, the LR is

$$\mathrm{LR}_{\mathrm{simple}} = \frac{0.064}{0.0256} = 2.5.$$

Note that the allele frequencies for p_1 and p_2 are still assumed to be 0.4 and 0.5, respectively, although the silent allele frequency is removed. With only dropout and drop-in and no mutations and silent alleles, the possible genotypes for C2 are 1/1, 1/2 or 2/2, while C1 is limited to 1/1. Adding this extra uncertainty in the model reduces the LR to

$$\mathrm{LR}_{\mathrm{drop}} = \frac{0.064 \cdot 0.975 + 0.08 \cdot 0.097 + 0 \cdot 0.01}{0.0256 \cdot 0.975 + 0.064 \cdot 0.097 + 0.04 \cdot 0.01} = 2.22.$$

With mutations and silent allele, but no dropout and drop-in, the possible genotypes for C2 are 1/1, 1/s or s/s, while C1 can be 1/1 or 1/s. The LR is

$$\mathrm{LR}_{\mathrm{silMut}} = \frac{0.0576]+0.0144+0.0144+0.0196+0.004}{0.0256+0.0128+0.0128+0.0064+0.0016+0.0008} = 1.83.$$

Finally, if all artefacts are taken into account, the LR is further reduced to $\mathrm{LR}_{\mathrm{all}} = 1.68$. The last result is confirmed with the RelMix software as shown in Fig. 3.

Results

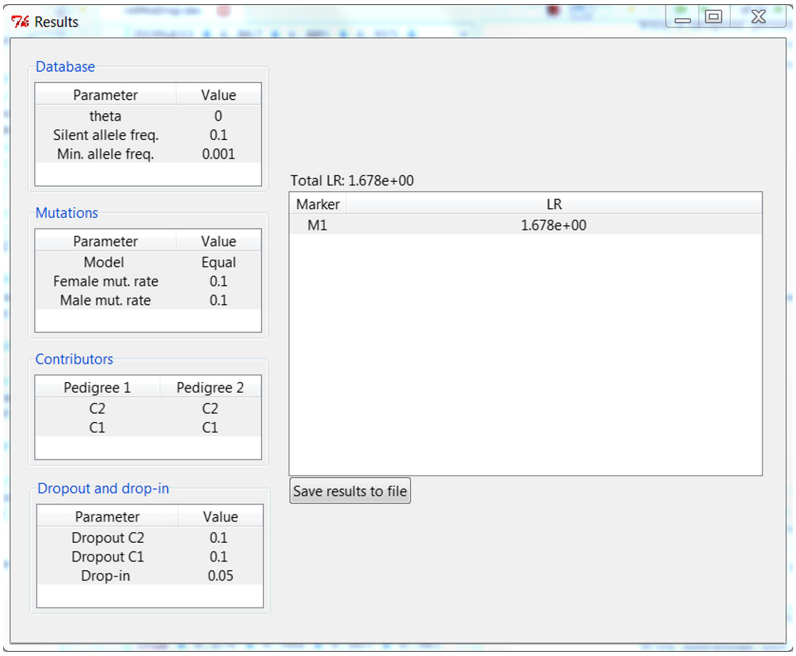
Paternity scenarios

We consider five different scenarios of the motivational example in Section "Motivational example" to illustrate the effect of dropout, drop-in and mutations on the likelihood

Table 3 Values corresponding to the formulas in Table 2 with $p_1 = 0.4$, $p_2 = 0.5$ and $p_s = 0.1$

$\mathbf{g}_\mathbf{K}$	$\mathbf{g}_\mathbf{U}$	$P(\mathbf{g}_\mathbf{K}, \mathbf{g}_\mathbf{U} \mid H_1)$		$P(\mathbf{g}_\mathbf{K}, \mathbf{g}_\mathbf{U} \mid H_2)$	$P(\mathcal{M} = 1 \mid \mathbf{g}_\mathbf{K}, \mathbf{g}_\mathbf{U})$	
		$r = 0$	$r = 0.1$		$d = 0, c = 0$	$d = 0.1, c = 0.05$
1/1	1/1	0.064	0.0576	0.0256	1	0.975
1/s	1/1	0.016	0.0144	0.0128	1	0.974
1/1	1/s	0.016	0.0144	0.0128	1	0.974
1/s	1/s	0.020	0.0196	0.0064	1	0.965
1/1	s/s	0.000	0.0000	0.0016	1	0.965
1/s	s/s	0.004	0.0040	0.0008	1	0.879
1/1	1/2	0.080	0.0784	0.0640	0	0.097
1/s	1/2	0.020	0.0196	0.0320	0	0.097
1/1	2/2	0.000	0.0080	0.0400	0	0.010
1/s	2/2	0.000	0.0020	0.0200	0	0.009
1/1	2/s	0.000	0.0016	0.0160	0	0.097
1/s	2/s	0.020	0.0204	0.0080	0	0.088

Fig. 3 Computation with RelMix confirms the result $\mathrm{LR}_{\mathrm{all}} = 1.68$ computed from Table 3. The result window displays the parameter values on the *left hand side* and the computed LR on the *right hand side*



Database

Parameter	Value
theta	0
Silent allele freq.	0.1
Min. allele freq.	0.001

Mutations

Parameter	Value
Model	Equal
Female mut. rate	0.1
Male mut. rate	0.1

Contributors

Pedigree 1	Pedigree 2
C2	C2
C1	C1

Dropout and drop-in

Parameter	Value
Dropout C2	0.1
Dropout C1	0.1
Drop-in	0.05

Total LR: 1.678e+00

Marker	LR
M1	1.678e+00

ratio. We will ignore silent alleles for now to limit the number of parameters. The scenarios are presented in Fig. 4. We assume a single marker with alleles {1, 2, 3} with corresponding frequencies {0.2, 0.3, 0.5}. The mutation model considers all mutations to be equally likely with a mutation rate of 0.1, which in most cases is unrealistically high, but it illustrates the effect of accounting for mutations. Figure 5 shows the LR as a function of the dropout probability for each scenario.

In scenario 1, no drop-in, dropout or mutations are needed to explain paternity, and the LR decreases with increasing dropout. The inclusion of mutations represents another source of uncertainty, and reduces the LR somewhat. With a dropout probability close to 1, both the child's alleles are likely to have dropped out and we have no information, as seen by the LR approaching 1. There is no effect of the choice of drop-in value.

In scenario 2, a dropout or mutation is the only explanation for paternity. As a consequence, the LR increases with increasing dropout probability and is slightly higher if we also include mutations. For high dropout values, the effect of including mutations is small.

Scenario 3 requires a drop-in or mutation to explain paternity. When $c = 0$, mutation is the only explanation, and the LR is not affected by the dropout probability. When $c = 0.05$, however, the LR actually increases with increasing dropout probability. If we look at the likelihoods for each hypothesis separately (Fig. 6a), we see that both hypotheses show decreasing likelihood as the dropout probability increases (as expected), but H_1 decreases more slowly than H_2.

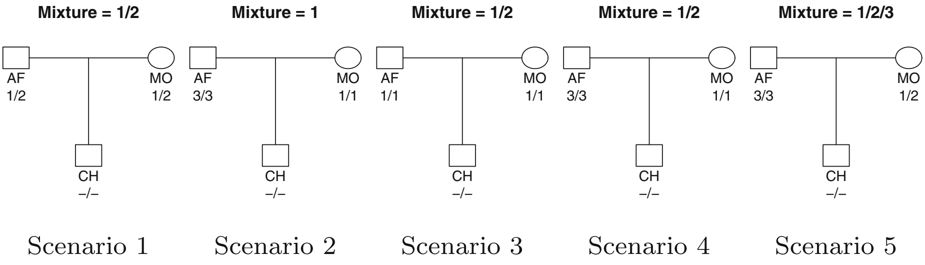
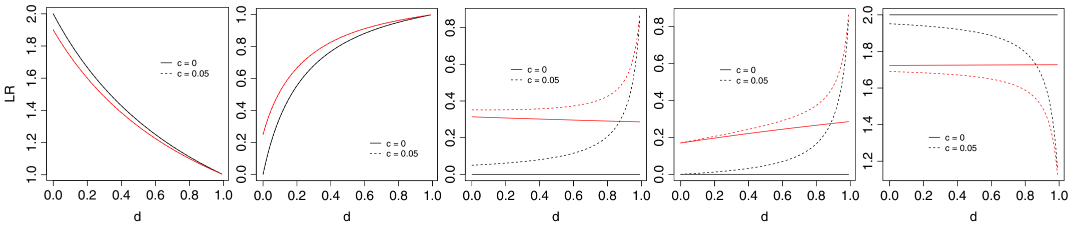


Fig. 4 Five scenarios for the motivational example. The mother-child mixture alleles are given above the pedigree



Scenario 1 Scenario 2 Scenario 3 Scenario 4 Scenario 5

Fig. 5 LR for the five scenarios in Fig. 4 as a function of the dropout probability. *Black lines* equal mutation rate $r = 0$, *red lines* equal $r = 0.1$. *Solid line* equals drop-in parameter $c = 0$ and *dashed line* equals $c = 0.05$. Silent alleles are not accounted for here

In scenario 4, either a combined dropout and drop-in incidence or a mutation is required to explain paternity. The LR is thus 0 when $c = 0$ unless we allow for mutations. A high dropout probability and inclusion of mutations increases the LR.

In scenario 5, there are three alleles in the mixture, and dropout is impossible unless there has also been a drop-in. When $c = 0$, the LR is indifferent to dropout if mutations are not accounted for. When mutations are included, there is a very weak increase in the LR with increasing dropout. The likelihood for each hypothesis (Fig. 6b) shows that the likelihood for H_1 decreases more rapidly than for H_2.

Simulations of true and false trios

To investigate the method's ability to differentiate between true and false paternities in prenatal paternity cases, we simulated 3,000 cases similar to the motivational example. Twenty-two real markers (part of prototype 24-plex STR panel from Thermo Fisher) were used to simulate genotype data where the alleged father was the true father of the child (true trios). Genotypes were simulated conditional on the pedigree with the `markerSim` function found in the `R` package `paramlink` (http://cran.r-project.org/web/packages/paramlink). A mixture including dropout and drop-in was generated from the genotypes of the mother and child with the `relMix` function `generateMix`. Each of the child's alleles would drop out with probability $d_{\text{true}} = [0, 0.1, 0.5, 0.9]$. We used a drop-in value of 0.1, which means that we expect 1 drop-in allele per 10 markers. A likelihood ratio comparing H_1 (paternity) and H_2 (not paternity) was computed with various choices of dropout values $d_{\text{LR}} = [0, 0.1, 0.5, 0.9]$. Note the difference between d_{LR} used in the LR computations and d_{true} used in the simulations. Calculations were done using the correct drop-in value of 0.1.

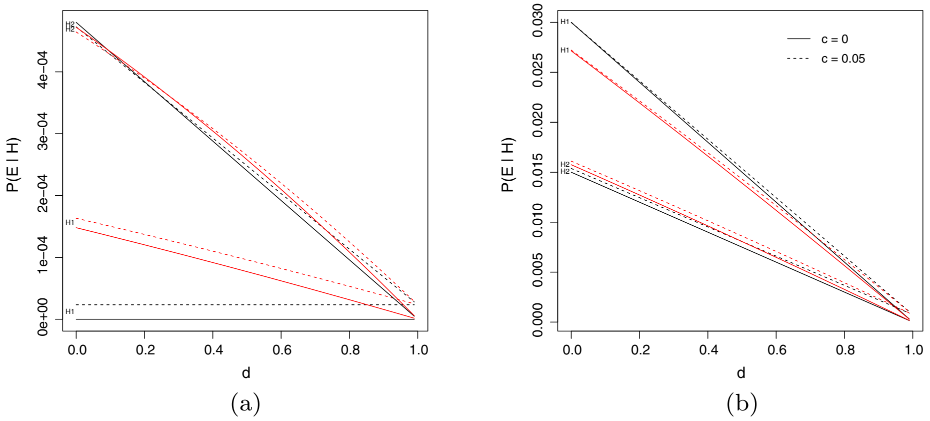
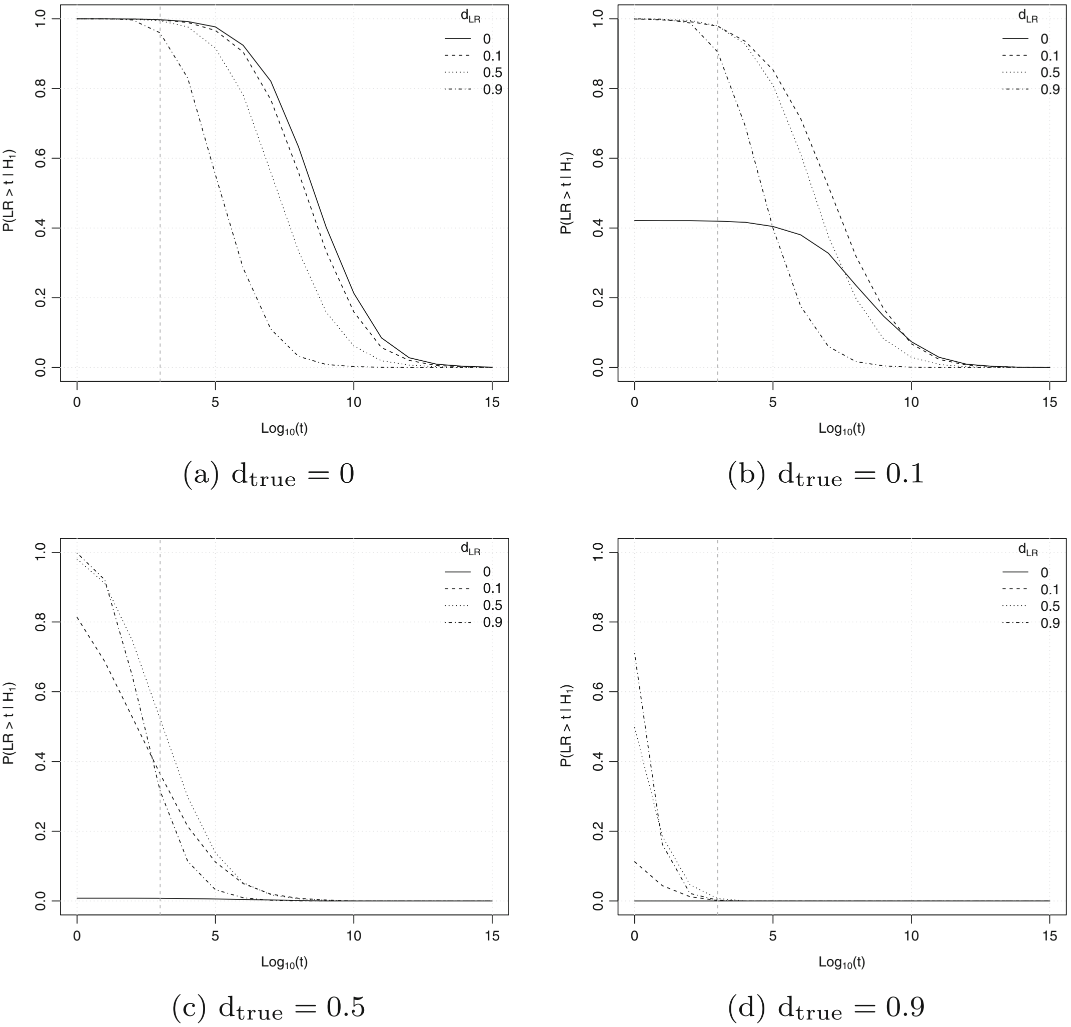


Fig. 6 Likelihood for each hypothesis in scenarios 3 and 5, with mutation rate 0 (*black lines*) and 0.1 (*red lines*), and drop-in parameter $c = 0$ (*solid line*) and $c = 0.05$ (*dashed line*). (**a**) Scenario 3: the *top four lines* are likelihoods for H_2; the *lower lines* are likelihoods for H_1. (**b**) Scenario 5: the *top four lines* are likelihoods for H_1; the *lower lines* are likelihoods for H_2

From the simulated data, we can compute $P(LR > t \mid H_1)$, i.e. the probability that the LR will exceed a threshold t if the alleged father is the true father of the child. We will refer to this as an exceedance probability. The probability of exceeding a threshold t if the alleged father and child are unrelated, $P(LR > t \mid H_2)$, is also of interest. However, doing simulations under H_2 (false trios) is challenging because events where an unrelated man would get a high LR by chance are rare, and a very large number of simulations would be required. Therefore, we used importance sampling as described in Kruijver [10] to compute exceedance probabilities under H_2 by using the simulations done under H_1. We estimated $\alpha = P(LR > t \mid H_2)$ as

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} I\left(LR_i > t\right) \cdot \mathcal{W}(LR_i) \qquad (6)$$

where I is the indicator function, LR_i is the likelihood ratio computed from simulation $i = 1, ..., N$ under H_1,



(a) $\mathrm{d_{true}} = 0$

(b) $\mathrm{d_{true}} = 0.1$

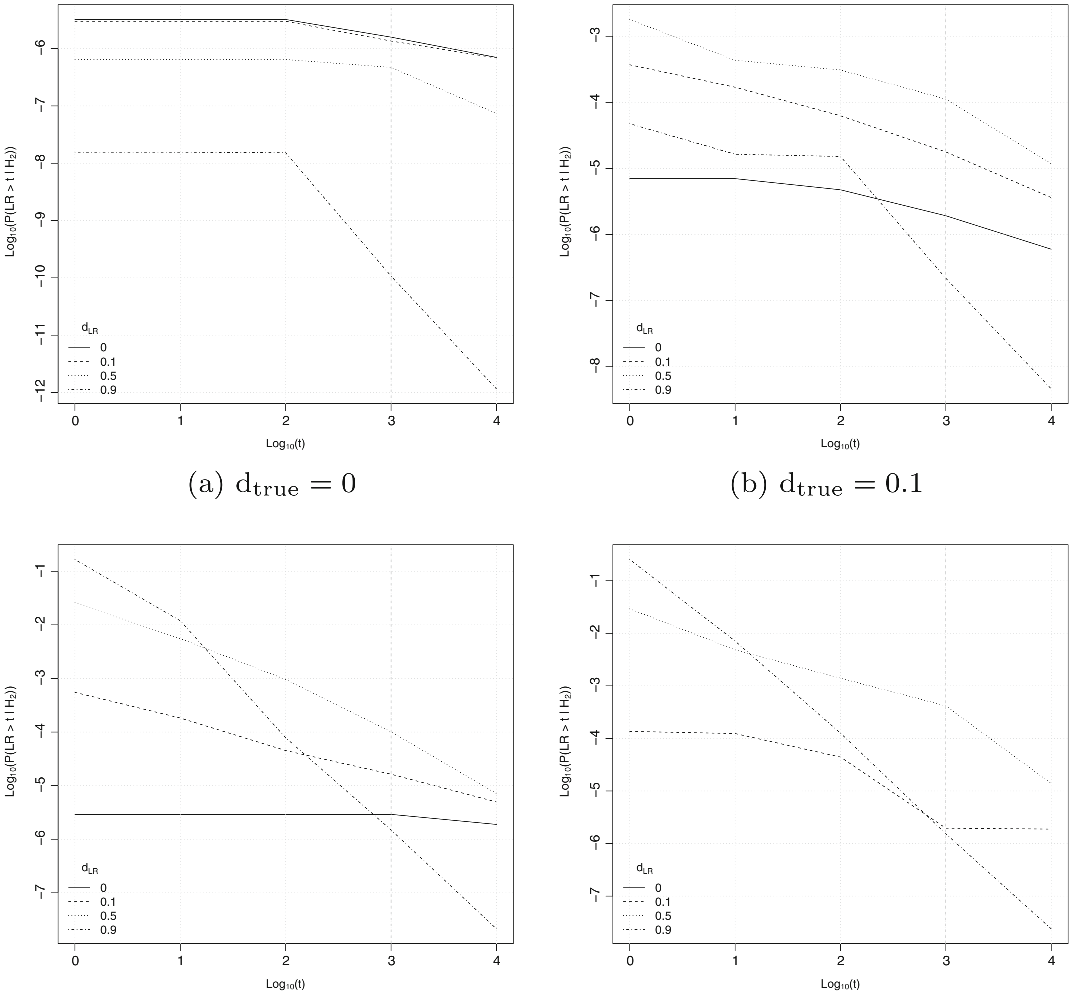
(c) $\mathrm{d_{true}} = 0.5$

(d) $\mathrm{d_{true}} = 0.9$

Fig. 7 Probability that LR will exceed threshold t for true trios with different choices of $\mathrm{d_{LR}}$. Each plot shows a different dropout level in the data ($\mathrm{d_{true}}$). The *dashed vertical line* corresponds to cut-off value $t = 1000$. In general, the highest exceedance probability is achieved when the correct dropout probability is used, and the exceedance probability decreases with increasing dropout in the data. Not accounting for dropout when there is dropout in the data gives the lowest exceedance probabilities. In fact, when the true dropout level is high (0.5 or 0.9), $\mathrm{d_{LR}} = 0$ gives exceedance probability 0 for all values of t. (**a**) $\mathrm{d_{true}} = 0$. (**b**) $\mathrm{d_{true}} = 0.1$. (**c**) $\mathrm{d_{true}} = 0.5$. (**d**) $\mathrm{d_{true}} = 0.9$

and $\mathcal{W}(LR_i)$ is a weight that compensates for the sampling bias. The weight indicates how much more likely we are to observe LR_i under H_2 compared to the hypothesis that we sample from. Note that $\mathrm{d}_{\mathrm{true}}$, the simulation dropout value, may differ from d_{LR}, the dropout value assumed in the calculation of LR_i. Therefore, we define a third hypothesis, the sampling hypothesis H_3, which is equal to H_1 except with dropout value $\mathrm{d}_{\mathrm{true}}$. The weight $\mathcal{W}(LR_i)$ corresponds to the likelihood ratio comparing H_2 and H_3.

In Fig. 7, the exceedance probability for true trios is plotted as a function of the threshold t. The dashed vertical line indicates a commonly used cut-off value of $LR = 1000$, which corresponds to a 'probability of paternity' of 99.9 % if we assume equal prior probability for the two hypotheses. In general, we achieve the highest exceedance probability when the correct dropout probability is used, while completely ignoring dropout when there is dropout in the data severely reduces the exceedance probability. The



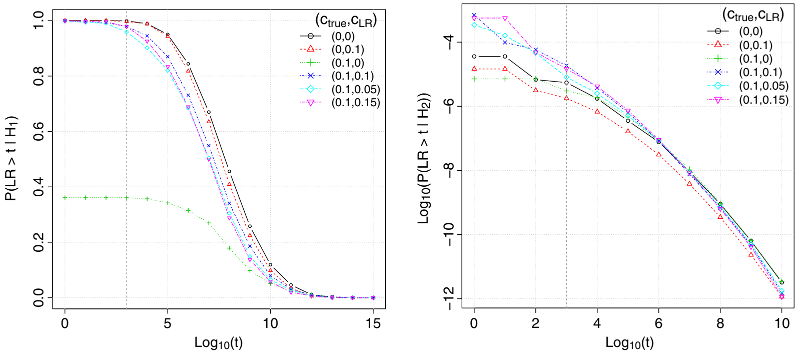
(a) $\mathrm{d}_{\mathrm{true}} = 0$

(b) $\mathrm{d}_{\mathrm{true}} = 0.1$

(c) $\mathrm{d}_{\mathrm{true}} = 0.5$

(d) $\mathrm{d}_{\mathrm{true}} = 0.9$

Fig. 8 Probability that LR will exceed threshold t for false trios with different dropout levels in the data ($\mathrm{d}_{\mathrm{true}}$). The *dashed vertical line* represents the cut-off value $t = 1000$. When $\mathrm{d}_{\mathrm{true}} = 0.9$, $\mathrm{d}_{\mathrm{LR}} = 0$ gives exceedance probability 0 for all t. (**a**) $\mathrm{d}_{\mathrm{true}} = 0$. (**b**) $\mathrm{d}_{\mathrm{true}} = 0.1$. (**c**) $\mathrm{d}_{\mathrm{true}} = 0.5$. (**d**) $\mathrm{d}_{\mathrm{true}} = 0.9$



(a) True trios (b) False trios

Fig. 9 Exceedance probabilities for (**a**) true and (**b**) false trios when the drop-in values vary. Each *curve* represents a combination of the true drop-in, c_{true}, and the drop-in value used in the LR calculation, c_{LR}. Both d_{true} and d_{LR} are kept fixed at 0.1. The *dashed vertical line* indicates $LR = 1000$

exceedance probability decreases when the true dropout rate increases.

Figure 8 shows the exceedance probability for false trios, which we may refer to as the false positive rate. With no actual dropout in the data ($d_{\text{true}} = 0$), a model with $d_{\text{LR}} = 0.9$ results in the lowest false positive rate, but also in the lowest exceedance probabilities for true trios (Fig. 7a). Similarly, when the dropout level in the data is high ($d_{\text{true}} = 0.9$), a model that ignores dropout ($d_{\text{LR}} = 0$) gives exceedance probability 0 for false trios but also for true trios (Fig. 7d). With such a high dropout level in the data, the probability of obtaining an LR above 1 is about 0.25 with $d_{\text{LR}} = 0.9$; however, the false positive rate quickly decreases as the threshold increases. In general, these simulations indicate a low probability of obtaining an LR above 1000 for false trios, even when the dropout value is misspecified and there is a high level of dropout in the data.

The above simulations only demonstrate the effect of dropout. To study the effect of having drop-in in the data and of misspecifying the drop-in value, we did some additional simulations where we kept both d_{true} and d_{LR} fixed at 0.1, and let the drop-in value vary. Let c_{true} and c_{LR} denote the drop-in value used in the simulations and in the LR calculations, respectively. The plots in Fig. 9 show the exceedance probabilities for true and false trios with different combinations of c_{true} and c_{LR}. The plots are based on 1000 simulations.

The highest exceedance probabilities for the true trios is observed when there is no drop-in in the data and we do not account for it ($c_{\text{true}} = 0$ and $c_{\text{LR}} = 0$). The result of misspecifying the drop-in value is that the exceedance probabilities are somewhat reduced; however, the effect is rather small and especially for higher thresholds. An exception is when there is drop-in in the data and this is not accounted for ($c_{\text{true}} = 0.1$ and $c_{\text{LR}} = 0$). The plot for false trios show that the largest false positive rates are observed when there is drop-in in the data. Again, the exception is when the drop-in is not accounted for. For higher thresholds, there is little effect both of having drop-in present and of the drop-in value used.

Real data

We consider a real case from the non-invasive prenatal paternity study in Gysi et al. [5]. Cell-free DNA was extracted from a mother's blood sample at 16 weeks of pregnancy. STRs were amplified with the prototype 24-plex STR

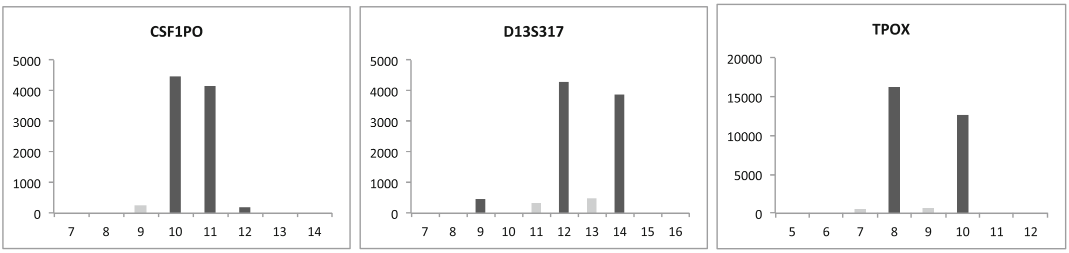


Fig. 10 Profiles for some loci in the real case. The *x*-axis indicates allele calls and the *y*-axis the number of reads. *Grey bars* indicate peaks that were filtered out as stutter

Table 4 Real data example showing the genotypes of the mother, alleged father and the mixture of mother and child, in addition to the full profile of the child. The 22 loci are part of the prototype 24-plex STR panel from Thermo Fisher

Marker	MO	AF	Mixture	CH
CSF1PO	10/11	12/12	10/11/12	11/12
D10S1248	16/16	14/15	16	14/16
D13S317	12/14	8/9	9/12/14	9/12
D14S1434	10/13	14/14	10/13/14	10/14
D16S539	12/12	12/14	12/14	12/14
D19S433	13/14	12/16	12/13/14	12/13
D1S1656	15/16	11/16	15/16	16/16
D1S1677	13/13	12/14	13	12/13
D2S1338	17/25	16/24	16/17/25	16/25
D2S1776	9/12	9/9	9/12	9/12
D2S441	10/14	10/10	10/14	10/10
D3S1358	16/19	14/17	14/16/19	14/16
D4S2408	8/9	9/10	8/9/10	8/10
D5S2500	14/18	17/17	14/17/18	14/17
D5S818	10/12	11/11	10/12	10/11
D6S1043	11/12	11/11	11/12	11/11
D6S474	14/14	14/18	14/18	14/18
D7S820	9/11	8/14	8/9/11	8/9
D8S1179	10/13	12/13	10/13	13/13
D9S2157	7/14	11/13	7/11/14	7/11
TH01	7/9.3	6/6	6/7/9.3	6/9.3
TPOX	8/10	8/11	8/10	8/10

panel from Thermo Fisher and sequenced on the Ion PGM. As the vast majority of cell-free DNA is of maternal origin, the STR profile shows a mixture of mother and child with the mother being the major component. Figure 10 shows the profiles of some of the loci in the case.

A reference sample from the child was taken after birth to evaluate the prenatal mixture profile; however, we will ignore it in the likelihood ratio computations to simulate a real scenario. Table 4 shows the profiles of the alleged father, mother, mother-child mixture and child (post-natal screening). Genotypes are listed for 22 out of 24 STR markers that were previously selected to perform best for this type of extremely unbalanced mixtures. Stutters were filtered based on sequence specific stutter ratios [17] estimated from preliminary data (unpublished). An analytical threshold of 0.005 (relative to the total number of reads per marker) was used. Since we know the genotype of the child, we can observe dropouts in the markers D10S1248, D1S1677 and D5S818. Without knowing the child's genotype, the inconsistencies between the profiles can be explained by both dropout and mutation, and for the marker D5S818 also a silent allele. Without peak height information, we cannot determine whether the child's maternal alleles have dropped out, and neither for the paternal alleles that are masked by maternal alleles. Three out of 16 paternal alleles not masked by a maternal allele have dropped out, and we can use $3/16 = 0.19$ as a rough estimate of the dropout probability. With the applied analytical threshold, there are no visible drop-in alleles.

To compute likelihood ratios, we used allele frequencies from Hill et al. [7, 8]. Figure 11 shows the LR as a function of the dropout probability d. The LR was calculated first without drop-in, then with drop-in $c = 0.05$, and finally, also with mutations and silent allele accounted for. We used a stepwise mutation model with mutation rate $r = 0.001$

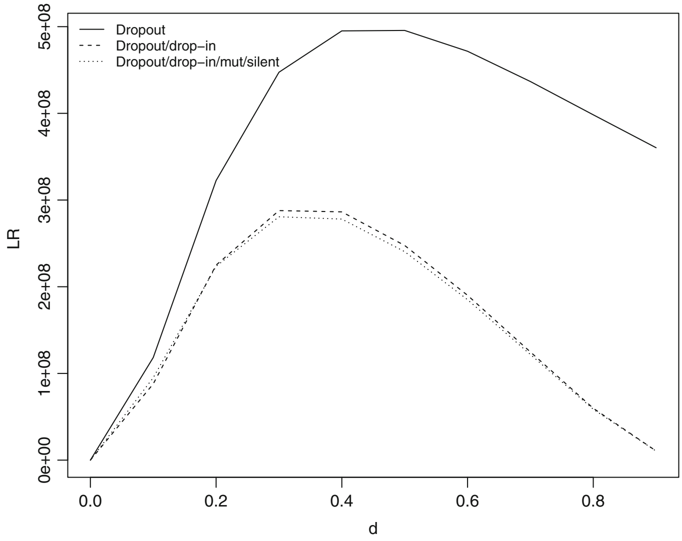


Fig. 11 Total LR in the real data example as a function of the dropout probability d with different artefacts included in the model: drop-in $c = 0.05$, stepwise mutation model with rate $r = 0.001$ and range 0.5, and silent allele frequency 0.01

Table 5 LR per marker in the real data example with various values of the dropout probability d, mutation rate r and silent allele frequency p_s. A drop-in value of $c = 0.05$ is used. The last column gives the LR for the regular paternity case (with no artefacts) where the full profile of the child is available

Marker	$r = 0, p_s = 0$		$r = 0.001, p_s = 0.01$		Regular
	$d = 0$	$d = 0.19$	$d = 0$	$d = 0.19$	
CSF1PO	2.705	2.686	2.658	2.639	2.777
D10S1248	0.000	0.638	0.001	0.624	1.679
D13S317	6.348	6.286	6.405	6.342	6.446
D14S1434	2.541	2.523	2.500	2.483	2.611
D16S539	18.73	18.53	18.87	18.65	19.00
D19S433	6.867	6.801	6.915	6.848	7.078
D1S1656	1.752	1.413	1.709	1.393	3.684
D1S1677	0.000	0.495	0.001	0.488	6.119
D2S1338	13.18	13.04	13.28	13.13	13.37
D2S1776	2.837	2.103	2.784	2.079	2.837
D2S441	2.215	1.799	2.186	1.784	4.750
D3S1358	4.629	4.588	4.669	4.626	4.688
D4S2408	2.074	2.062	2.092	2.080	2.105
D5S2500	2.747	2.727	2.697	2.677	2.819
D5S818	0.000	0.346	0.060	0.383	2.809
D6S1043	1.875	1.608	1.858	1.597	3.374
D6S474	5.537	5.485	5.584	5.531	5.587
D7S820	3.408	3.380	3.436	3.408	3.471
D8S1179	1.157	1.102	1.142	1.092	1.517
D9S2157	1.668	1.661	1.683	1.675	1.686
TH01	4.136	4.100	4.009	3.975	4.247
TPOX	0.870	0.908	0.863	0.903	0.870
Total	0.0E + 00	2.1E + 08	4.7E + 02	2.1E + 08	2.1E + 12

and range 0.5 [2], and a silent allele frequency of 0.01. The LR decreases when a drop-in probability is introduced. There is little effect of also including mutations and silent alleles. Dropout alone is sufficient to explain paternity in this case, and the inclusion of drop-in, mutations and silent allele introduces more uncertainty in the model. Independent of the inclusion of artefacts, the LR appears sufficiently large for all reasonable values of the dropout probability.

Table 5 lists the LRs per marker with different values of dropout, drop-in, mutation rate and silent allele frequency. The last column in the table gives the LR for the regular paternity case (with no artefacts) where the reference profile for the child is available.

Discussion

We have presented a method and software for calculation of relationship inference based on mixtures. The method can account for artefacts such as dropout and drop-in, mutations, silent alleles and population substructure. The software is freely available in the R package relMix, both as a graphical user interface and as several command line functions.

The primary motivation for the paper was paternity cases where the child's DNA profile is only available as a mixture with the mother's profile, and there may be dropout and drop-in in the mixture. An example is non-invasive prenatal paternity testing based on cell-free DNA. The highly unbalanced mixture and the very low amount of foetal DNA makes dropout and drop-in likely. With the model presented here, a high LR supporting paternity of an alleged father was calculated in a real prenatal paternity case from a mixture with three visible dropouts. Through simulations, it could further be shown that true trios have a high probability of achieving an LR above 1000 (corresponding to a ”probability of paternity” of 99.9 % when assuming equal prior probabilities) and that likelihood ratios that support paternity for false trios are very unlikely. Together with a sufficiently high LR for the real data example, this indicates that the challenging mixtures obtained from typing cell-free DNA in a pregnant woman's blood sample can be handled with this method, even if dropout and drop-in do occur. Although paternity cases have been the focus in most examples presented here, we do emphasise that our software can handle all types of relationships between the individuals in the mixture, and the hypotheses may involve any number of relatives.

Our simulation study shows that the ability to identify true trios is drastically reduced if there is dropout in the data that is not accounted for. Approaches for estimating the dropout probability are not discussed here, but several methods for dropout estimation in capillary electrophoresis based data exist [15, 16]. In the real data example, the STRs were sequenced using massive parallel sequencing. There may be additional factors that influence the dropout probability in MPS data; however, this is a topic beyond the scope of this paper. We further note that dropout in prenatal paternity cases may depend on the stage of pregnancy. In addition, Ashoor et al. [1] found a decrease in the foetal fraction of total cell free DNA with increasing maternal weight. In the real data example, we could count the number of visible dropout alleles as a minimum estimate of the dropout probability since we knew the child's full profile, but this information would usually not be available. One approach to deal with the unknown dropout probability is to do a sensitivity analysis by calculating the LR for a range of dropout values to see how it varies. This was done in the real data example and showed that the LR was sufficiently high for all reasonable values of the dropout probability.

In the real prenatal paternity case, stutters were filtered based on sequence specific stutter ratios estimated from preliminary data. More experiments are needed to obtain precise estimates for MPS data. In mixtures where one contributor is in large excess, defining stutter ratios precisely is crucial to be able to call minor alleles in stutter position of a major allele. Two of the three dropouts were in stutter positions of a maternal allele and it is not known whether these foetal alleles dropped out or were masked by a stutter. van der Gaag et al. [17] show that stutter ratios mainly depend on the number of uninterrupted repeats of an STR allele and therefore on the STR sequence. Massive parallel sequencing thus enables lowering the threshold to call an allele in stutter position of a major allele. We emphasise that the prenatal paternity case data is only an example of data that can be analysed with our model, and the technical details of MPS data is therefore not the main focus here.

The mixture model we have used is a semi-continuous model that includes dropout and drop-in, but does not use information about peak heights. This model could be replaced by a fully continuous model that also incorporates peak heights and stutter. There are examples of the use of continuous models [12, 14]. However, incorporating general relationships that also consider artefacts such as mutations and silent alleles into a continuous mixture model does not appear trivial. Another possible extension of the model could be to include linkage and linkage disequilibrium.

Acknowledgments GD has received funding support from the European Union Seventh Framework Programme, EUROFORGEN-NoE (FP7/2007-2013) under grant agreement no. 285487. The authors would like to thank Thore Egeland for helpful comments and two anonymous reviewers who helped improve the paper.

References

1. Ashoor G, Poon L, Syngelaki A, Mosimann B, Nicolaides KH (2012) Fetal fraction in maternal plasma cell-Free DNA at 11-13 weeks' gestation: effect of maternal and fetal factors. Fetal Diagn Ther 31(4):237–243. doi:10.1159/000337373
2. Egeland T, Kling D, Mostad P (2016) Relationship inference with familias and R: statistical methods in forensic genetics Academic Press
3. Egeland T, Dørum G, Vigeland MD, Sheehan NA (2014) Mixtures with relatives: a pedigree perspective. Forensic Sci Int Genet 10:49–54. doi:10.1016/j.fsigen.2014.01.007
4. Fung WK, Hu YQ (2008) Statistical DNA forensics theory, methods and computation. Wiley, England
5. Gysi M, Arora N, Sulzer A, Voegeli P, Kratzer A (2015) Non-invasive prenatal paternity testing with STRs: a pilot study. Forensic Science International: Genetics Supplement Series 5:e291 – e292. doi:10.1016/j.fsigss.2015.09.115
6. Haned H, Slooten K, Gill P (2012) Exploratory data analysis for the interpretation of low template DNA mixtures. Forensic Sci Int Genet 6(6):762 – 774. doi:10.1016/j.fsigen.2012.08.008
7. Hill CR, Butler JM, Coble MD (2006) Allele frequencies for 26 MiniSTR loci with U.S. Caucasian, African American, and Hispanic populations. http://www.cstl.nist.gov/biotech/strbase/NISTpop.htm
8. Hill CR, Duewer DL, Kline MC, Coble MD, Butler JM (2013) U.S. population data for 29 autosomal STR loci. Forensic Sci Int Genet 7(3):e82 – e83. doi:10.1016/j.fsigen.2012.12.004
9. Kaur N, Bouzga MM, Dørum G, Egeland T (2015) Relationship inference based on DNA mixtures. Int J Legal Med 2:323–329. doi:10.1007/s00414-015-1276-1
10. Kruijver M (2015) Efficient computations with the likelihood ratio distribution. Forensic Sci Int Genet 14:116 – 124. doi:10.1016/j.fsigen.2014.09.018
11. Lo YMD, Tein MS, Lau TK, Haines CJ, Leung TN, Poon PM, Wainscoat JS, Johnson PJ, Chang AM, Hjelm NM (1998) Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. Am J Hum Genet 62(4):768 – 775. doi:10.1086/301800
12. Mortera J, Vecchiotti C, Zoppis S, Merigioli S (2016) Paternity testing that involves a DNA mixture. Forensic Sci Int Genet 23:50–54. doi:10.1016/j.fsigen.2016.02.014
13. Slooten K (2016) Distinguishing between donors and their relatives in complex DNA mixtures with binary models. Forensic Sci Int Genet 21:95 – 109. doi:10.1016/j.fsigen.2015.12.001
14. Taylor D, Bright J-A, Buckleton J (2014) Considering relatives when assessing the evidential strength of mixed DNA profiles. Forensic Sci Int Genet 13:259–263
15. Tvedebrink T, Eriksen PS, Mogensen HS, Morling N (2009) Estimating the probability of allelic drop-out of STR alleles in forensic genetics. Forensic Sci Int Genet 3(4):222 – 226. doi:10.1016/j.fsigen.2009.02.002
16. Tvedebrink T, Eriksen PS, Asplund M, Mogensen HS, Morling N (2012) Allelic drop-out probabilities estimated by logistic regression - Further considerations and practical implementation. Forensic Sci Int Genet 6(2):263 – 267. doi:10.1016/j.fsigen.2011.06.004
17. van der Gaag KJ, de Leeuw KJ, Hoogenboom J, Patel J, Storts DR, Laros JF, de Knijff P (2016) Massively parallel sequencing of short tandem repeats—population data and mixture analysis results for the PowerSeq system. Forensic Sci Int Genet 24:86 – 96. doi:10.1016/j.fsigen.2016.05.016

Paper IV

Relationship inference: Estimation and Model Selection

Navreet Kaur[*], Geir Storvik[†], Magnus Dehli Vigeland[‡]and Thore Egeland[*]

Abstract

The methods and implementations of this paper are relevant to describe and test the relationship between two individuals. Forensics is one of several important applications where family relationships are questioned. Traditionally, both in crime cases and in kinship cases, two competing hypotheses are presented verbally. The hypotheses may for instance specify who contributed to a mixture versus an unrelated man or, in a kinship case, a specific relation between two individuals versus unrelatedness. However, the alternative parametric formulations of hypotheses are relevant in forensic case work. The alternative hypothesis can be completely general when testing a relation (for instance, there is no need to restrict attention to'unrelated'), and a parametric representation facilitates applications of well-known statistical theory. In this paper, we take the parametric framework based on IBD (*identity-by-descent*) further. An allele in one individual is IBD to an allele in another individual if the ancestral origin is the same within a specified pedigree. Any pairwise relationship of non–inbred individuals correspond to a point $\kappa = (\kappa_0, \kappa_1, \kappa_2)$ in the so-called IBD triangle where $\kappa_i, \ i = 0, 1, 2,$ is the probability that the individuals share i alleles IBD. Methods for estimating κ from genetic markers are well-known, but *boundary* issues (the parameters are not inner points of the valid domain) present challenges. The main novelty of the paper is that we address the problem by optimization with non-linear constraints and model selection based on the Bayesian Information Criterion (BIC). Moreover, we calculate parametric bootstrap confidence regions for the IBD parameters and the

[*] Faculty of Chemistry, Bioltechnology and Food Science. Norwegian University of Life Sciences. Aas, Norway

[†] Department of Mathematics. University of Oslo. Oslo, Norway

[‡] Department of Medical Genetics. Oslo University Hospital and University of Oslo. Norway

kinship coefficient. These can be used for testing purposes. Plotting methods are presented in order to visualize the relations and their location in the valid domain. The methods are implemented in the freely available `R` library `IBDest2` which is based on `paramlink`.

Keywords: Kinship estimation; IBD triangle; constrained maximum likelihood; model selection; BIC

1 Introduction

The statistical analysis following kinship tests in forensic genetics traditionally relies on a likelihood based approach. Verbal hypotheses are presented suggesting a specific relationship between two individuals versus an alternative hypothesis, typically stating unrelatedness, and likelihood ratios are calculated based on these hypotheses. The paper [13] presents pairwise kinship analysis in the forensic context.

However, papers [9] and [6] discuss parametric representations of such hypotheses. The papers show the relevance and usefulness of a parametric approach for statistical inference in forensic genetics. It is worthwhile to expand on such parametric approaches and explore these in new directions, as this paper aims to do. A parametric presentation invites for a larger frame of statistical tools that can more directly be used for forensic applications. A classic verbal hypothesis presentation for testing paternity could for instance be H_1: “A is the biological father of B”, with the alternative hypothesis stating unrelatedness, i.e., H_2: “A and B are unrelated”. In [6], a parametric formulation for such kinship testing is presented, and the hypotheses and the model is formulated in terms of the identity-by-descent (IBD) parameters $\kappa = (\kappa_0, \kappa_1, \kappa_2)$, where κ_i is the probability that the individuals share i alleles IBD. Turning back to our paternity case, a parametric presentation of the same hypotheses would be H_1: $\kappa_1 = 1$ versus H_2: $\kappa_0 = 1$.

The κ-coefficients were introduced by [3] and are used to specify the relationship between any two non–inbred individuals. The main advantages of a parametric approach in terms of the κ-parameters is that the alternative hypothesis can be quite general: the alternative hypothesis above may be H_2: $\kappa_1 < 1$.

Furthermore, a parametric formulation allows relationship testing to follow the classical framework of hypothesis testing. With such a parametric presentation, [6] argues that a proper distribution for the test statistic can be obtained, hence mathematical approaches that earlier have been out of reach in forensic research are made applicable. In other words, parametric inference as described above follows the classical approach of applied statistics: a model is specified and hypotheses are formulated in terms of the parameters of the model. We can study the power, i.e., the probability of rejecting H_1 given that the alternative is true, as a function of the parameters of interest.

In the following we take the results presented in [6, 14, 16] a step further and explore different directions of this application. We present a method for estimating the κ-parameters based on maximum likelihood theory presented in [14], and hence we estimate the relations in question. Further we introduce the corresponding confidence regions of the estimates, based on parametric bootstrapping methods. With such intervals we are able to assess the uncertainty of estimates and these supplement visual inspection of plots. We study so-called boundary issues, and find methods for handling situations where the relations in question are found on the boundary of the valid domain of all possible relationships [15]. These methods are based on model selection and optimization. Available methods, like the asymptotic theory of [6] and implementation of kinship estimates, as the `R` library `Relatedness`, are limited to SNP markers and therefore of limited forensic relevance.

The method of choice in many statistical applications for estimation is generally based on the likelihood function. Intuitively, the maximum likelihood estimate is the value of the parameter that maximizes the likelihood. This maximum likelihood estimator has many desirable properties, like asymptotic normality (i.e., the distribution approaches a normal distribution as the number of observations goes to infinity) and optimality, provided that some regularity conditions hold. One such regularity condition is that the parameter should not be on the boundary of the valid domain. In our case, the valid domain refers to the 'relationship triangle', illustrated in the leftmost panel of Figure 1. As seen from the figure, many of the most well-known pairwise relationships are found on the boundary of the triangle. This boundary problem complicates the properties of the maximum likelihood estimates and also the use of simulation based approaches like the bootstrap methods, see [1]. In order to deal with such complicating factors, we use an optimization and model selection approach to find the most appropriate parameter estimate. Model selection is done based on the Bayesian Information Criterion (BIC).

By controlling the boundary issues, we aim for more reliable confidence regions for our κ estimates in order to state how closely or remotely related two individuals are. Also, the estimated κ then corresponds to one or more existing pedigrees as shown in [16]. In other words, given a specific κ, there exists at least one pedigree with this κ.

The kinship coefficient ψ is another measure of relatedness and is often used in human genetics. When there is no inbreeding, $\psi = (2\kappa_2 + \kappa_1)/4$.

The paper is organized in the following way: In Section 2 we review the

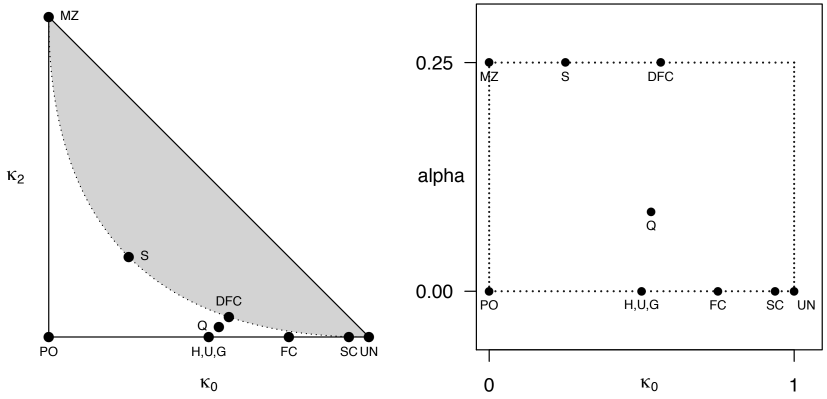


Figure 1: The IBD triangle to the left shows some common relationships, see Table 1. The dashed curve is given by $\kappa_1^2 = 4\kappa_0\kappa_2$, and the valid domain for κ is the white area under the dashed line. The plot to the right illustrates the transformation explained in Section 2.3.

concept of identity-by-descent (IBD), the IBD coefficients κ and the kinship coefficient ψ. We describe how these parameters can be estimated by maximum likelihood methods, and discuss how boundary issues may be solved by reparametrisation, constrained optimisation and model selection. In Section 2.4 we turn to simulations and present methods for finding confidence regions of estimates based on parametric bootstrapping. In Section 3 we give examples for simulated and real data. Finally, recommendations, results and limitations are discussed in Section 4.

2 Methods

Traditionally two competing hypotheses are formulated and compared using the likelihood ratio. As a motivational example, we will look at two individuals A and B who are questioned half siblings. The classical presentation of two competing hypotheses is verbal and may for instance be

$$H_1 : \text{A and B are half siblings,}$$
$$H_2 : \text{A and B are unrelated.}$$

Given genotype data on A and B, the likelihood ratio is

$$LR = \frac{Pr(\text{data}|H_1)}{Pr(\text{data}|H_2)}.$$

In the following sections, we will move away from this classical verbal presentation, and turn towards a parametric representation based on identity-by-descent theory and κ parameters, together with the kinship coefficient ψ.

2.1 The IBD coefficients

An allele in one individual is said to be identical by descent to an allele in another individual if the allele derives from the same ancestral allele within the specified pedigree [17, 5]. Given two non-inbred individuals, let Z be the number of IBD alleles at some autosomal locus. We define the IBD coefficients $\kappa_i = Pr(Z = i),\ i = 0, 1, 2$. Note that $\kappa_0 + \kappa_1 + \kappa_2 = 1$. As discussed in [14] and [15], all pairwise relations fall within the domain

$$K^* = \{(\kappa_0, \kappa_2) : \kappa_0, \kappa_2 \in [0, 1], (1 - \kappa_0 - \kappa_2)^2 \geq 4\kappa_0\kappa_2\} \tag{2.1}$$

The likelihood function for one marker can be written ([17], p. 42)

$$\begin{aligned} L(\kappa) &= \kappa_0 P(G \mid Z = 0) + \kappa_1 P(G \mid Z = 1) + \kappa_2 P(G \mid Z = 2) \\ &= \kappa_0 \mathrm{UN}(p_{g_1}, p_{g_2}) + (1 - \kappa_0 - \kappa_2)\mathrm{PO}(p_{g_1}, p_{g_2}) + \kappa_2 \mathrm{MZ}(p_{g_1}, p_{g_2}), \end{aligned} \tag{2.2}$$

where $G := (g_1, g_2)$ are the genotypes. The functions *UN*, *PO* and *MZ* correspond to the terms for 'unrelated', 'parent offspring', and 'monozygotic twins', respectively. The dependence on g_1 and g_2 and the corresponding frequencies p_{g_1} and p_{g_2} is omitted in the notation $L(\kappa)$. Throughout we assume Hardy Weinberg Equilibrium.

The IBD probabilities specify different relationships, and some common relationships are given in Table 1 in terms of the three κ parameters. Using these relations, the more traditional representation of verbal hypotheses may now take a parametric form. Turning back to our motivational example, we may instead state the hypotheses

$$\begin{aligned} H_1 &: \kappa_0 = \kappa_1 = \frac{1}{2} \\ H_2 &: \kappa_0 \neq \frac{1}{2} \text{ or } \kappa_1 \neq \frac{1}{2}. \end{aligned}$$

Note that the alternative hypotheses now is quite general compared to the conventionally adopted and restrictive alternative stating unrelatedness.

2.2 The kinship coefficient - ψ

The coefficient of kinship ψ between two individuals A and B measures the amount of IBD sharing allele, and is the probability that a randomly chosen allele in A is IBD to a randomly chosen allele from B.

Table 1: IBD probabilities for some pairwise relationships. The term avuncular refers to three relationships which are indistinguishable based on unlinked autosomal markers; halfsiblings (H), grandparent-grandchild (G) and uncle/aunt-nephew/niece (U)

Relationship	$\kappa = (\kappa_{\mathbf{0}}, \kappa_{\mathbf{1}}, \kappa_{\mathbf{2}})$	$\psi = \frac{2\kappa_2+\kappa_1}{4}$
Parent-child (PO)	$(0, 1, 0)$	$\frac{1}{4}$
Siblings (S)	$(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$	$\frac{1}{4}$
Avuncular (H, G, U)	$(\frac{1}{2}, \frac{1}{2}, 0)$	$\frac{1}{8}$
First cousins (FC)	$(\frac{3}{4}, \frac{1}{4}, 0)$	$\frac{1}{16}$
Double first cousins (DFC)	$(\frac{9}{16}, \frac{6}{16}, \frac{1}{16})$	$\frac{1}{8}$
Quadruple half first cousins (Q)	$(\frac{17}{32}, \frac{14}{32}, \frac{1}{32})$	$\frac{1}{8}$
Second cousins (SC)	$(\frac{15}{16}, \frac{1}{16}, 0)$	$\frac{1}{64}$
Unrelated (UN)	$(1, 0, 0)$	0
Monozygotic twins (MZ)	$(0, 0, 1)$	$\frac{1}{2}$

The kinship coefficient is directly related to the inbreeding coefficient, f, as explained in [17], and made precise below where R is a (possibly hypothetical) child of P and Q:

$$\begin{aligned} \psi = \psi_{\mathrm{P,Q}} &= Pr(\text{random allele of P is IBD with random allele of Q}) \\ &= Pr(\text{R receives IBD alleles from her parents}) \\ &= Pr(\text{R is autozygous}) = f. \end{aligned}$$

Values of κ and ψ for some common relationships are shown in Table 1. We will be comparing different ways of estimating κ and thus ψ. The mean squared error,

$$MSE(\hat{\psi}) = E(\hat{\psi} - \psi)^2 = var(\hat{\psi}) + \left(E\hat{\psi} - \psi\right)^2, \tag{2.3}$$

is a widely used criterion for comparison of estimators. The estimator with the smallest MSE is considered the best. Our applications are complicated by the mentioned boundary problem, but we will still use MSE to supplement visual inspection of plots.

2.3 Optimization and model selection

The log likelihood function of n independent markers is

$$l(\kappa) = \sum_{i=1}^{n} \log(L_i(\kappa)), \tag{2.4}$$

where $L_i(\kappa)$ is given in Equation (2.2). Recall that $(\kappa_0, \kappa_2) \in K^*$, see (2.1) and Figure 1. Maximum likelihood estimation of κ amounts to maximizing the expression (2.4). In our case this is complicated by the boundary conditions illustrated in Figure 1; without these boundary conditions, the maximum likelihood (ML) estimates are optimal and asymptotically normal. In other words, if the true κ is on the boundary of K^*, then the standard ML theory does not work. One can, however, argue that in some applications it is reasonable to ignore constraints beyond $\kappa_i \geq 0$ and $\kappa_0 + \kappa_1 + \kappa_2 = 1$. In our implementation and examples the resulting estimates are referred to as 'Standard'. The point is that the *realised* IBD pattern for a pair of individuals may not fall in the permissible region and one may therefore choose not to correct or constrain the estimates. The leftmost panel of Figure 3 shows examples of such estimates.

If the objective is to estimate the true pedigree relating a pair of individuals, it is reasonable to constrain estimates to the permissible region. In Section 2.3.1 this is approached via model selection, below more standard constrained optimisation is discussed. There are two numerical approaches for finding maximum values in the legal domain. The first is to use constrained optimization. We have implemented this using the `R` package `maxLik` [8]. There are some disadvantages to this approach: Only linear constraints are possible. To accomodate the non-linearity of the valid domain, we reparametrize it and let

$$\alpha = \frac{\kappa_0 \kappa_2}{(1 - \kappa_0 - \kappa_2)^2}. \tag{2.5}$$

From the definition of the valid domain given in (2.1), it follows that $0 \leq \alpha \leq \frac{1}{4}$. The only valid solution for the above equation is

$$\kappa_2 = 1 - \kappa_0 - \frac{\sqrt{\kappa_0^2 + 4\alpha\kappa_0(1 - \kappa_0)} - \kappa_0}{2\alpha}. \tag{2.6}$$

We may then optimize over $(\kappa_0, \alpha) \in [0, 1] \times [0, 1/4]$, before transforming back to κ_2 using (2.6). In addition we check the boundary, ∂K^*, i.e., we also maximize along the boundary. In other words, transformation is only relevant for interior points of K^* since we will check the border ∂K^* separately. The reparametrization and transformation from κ-space to α-space explained above is illustrated in Figure 1. We started with half siblings, i.e., $(\kappa_0, \kappa_1, \kappa_2) = (0.5, 0.5, 0)$. Using the formula for α, we find

$$\alpha = \frac{\kappa_0 \kappa_2}{\kappa_1^2} = \frac{0.5 \cdot 0}{0.5^2} = 0,$$

and so we find the same half sibling relationship as $(\kappa_0, \alpha) = (0.5, 0)$. Note that the transformation (2.5) is not defined when $\kappa_0 + \kappa_2 = 1$ as is the case for UN and MZ. On the right hand side of Figure 1 we have plotted these points as the continuous limits: For MZ, the limit along the stapled line of the plot to the left, for UN the limit along the x-axis. We return to this issue in the discussion.

2.3.1 Model selection

In order to take into account that the estimation of κ includes boundary values, we can look at our problem in terms of model selection: A solution on the boundary involves only one parameter, and may be preferred to an interior point even if the interior maximum of the likelihood is larger. In this case, when the objective is to find the true model or pedigree, it is reasonable to use the Bayesian Informative Criterion (BIC) [2],

$$BIC = -2\ln(\hat{L}) + C\ln(n),$$

where $\hat{L}$ is the maximum likelihood value, $C = 2$ for an interior point and $C = 1$ on the boundary, while n is the number of markers. With model selection, the procedure for estimating κ is to first select the best model using BIC and thereafter estimate κ within this model. This includes both the vertices as well as the boundary lines. BIC is a consistent model selection criterion, which means that as n increases we will get the right model. Within a model, it follows from standard ML theory that the estimates are consistent, so combined we get a consistent estimator. There are of course other model selection criteria that may be used, like for instance Akaike's Information Criterion (AIC). However, AIC tries to select the model that most adequately describes the unknown reality, while BIC tries to find the true model among the set of candidates. We are in the latter situation, and therefore use BIC.

2.4 Parametric bootstrap and confidence regions

Simulation is an valuable tool also for the current application. For one thing, it makes it possible to compare estimators and assess their uncertainty. Based on simulations, the parameters and the 2.5% and 97.5% percentiles can be estimated. These intervals are relevant when the pedigrees are known and the approach is exemplified in Example 3.2. More realistically, the family relationship is not known, and then we can use parametric bootstrap as follows: First an estimate κ^* is obtained from the data. Then the likelihood function (2.2) can be used to generate a table describing the joint genotype probabilities of the two individuals for each marker. This table can then be used to simulate marker data B times from which we get the bootstrap estimates $\hat{\kappa}_1, \ldots, \hat{\kappa}_B$. There exists several bootstrapping methods for creating confidence intervals or regions as described in [4]. We use the percentile method independently for the parameters κ_0 and κ_2 truncated to the interval $[0, 1]$ as exemplified in e.g. Table 4. Alternatively, we calculate a confidence ellipse. The ellipses are based on the asymptotic bivariate normal distribution for $(\hat{\kappa_0}, \hat{\kappa_2})$ where the mean vector and covariance matrix are estimated from the bootstrap samples. The next section gives details on implementation. Note that the problems with parameter values on the boundary mentioned previously for ML estimates also apply to bootstrap estimates as discussed in [1]. In particular, the confidence ellipse does not account for the boundary problems and is likely to be most reliable for inner points of the IBD triangle and a large number of markers.

2.5 Implementation

The methods and examples of this paper are implemented and documented in the `R` library `IBDest2` available as `http://familias.name/IBDest2_1.0.zip`. In several respects this library is a wrapper based on `paramlink`. The main extensions relate to model selection, constrained estimation and optimisation using the `maxLik` [8] library. Also, in some respects `paramlink` and other libraries like `Relatedness` are restricted to SNP markers. Confidence ellipses have been estimated using the library `ellipse` based on methods described in [12]. We have constrained the ellipses to the valid domain.

3 Results

Three examples are presented below. The first is based on published data and details the three different estimation methods. Then, in the second example, we simulate from known pairwise relationships using available marker databases. This allows us to compare the accuracy of the various estimates visually and also from estimates of the parameters describing the relationships. Finally, parametric bootstrap with confidence ellipses are demonstrated on simulated data for varying number of markers. In all cases, markers are assumed to be independent.

Table 2: Genotypes for two possible half-siblings on the SGMPlus loci, see [11].

Locus	$g_{1,1}$	$g_{1,2}$	$g_{2,1}$	$g_{2,2}$
D2S1338	22	23	17	23
D3S1358	17	18	16	18
FGA	22.2	24	22	23
D8S1179	14	17	12	14
TH01	8	8	8	9
VWA	18	18	17	18
D16S539	12	12	12	13
D18S51	13	16	13	16
D19S433	13.2	16.2	14	14
D21S11	30.2	35.2	29	31.2

Table 3: Comparison of estimates in Example 3.1. All estimates except 'Standard' reparametrise the model. The last line of the table is the one with lowest BIC value of the preceding four lines, in this case for $\kappa_2 = 0$.

	$\hat{\kappa_0}$	$\hat{\kappa_1}$	$\hat{\kappa_2}$	loglik	AIC	BIC	$\hat{\psi}$
Standard	0.532	0.403	0.065	-79.188	162.376	167.586	0.133
Constrained	0.532	0.444	0.024	-79.249	162.376	167.586	0.123
On curve	0.581	0.363	0.057	-79.249	160.498	163.103	0.119
$\kappa_0 = 0$	0.000	0.986	0.014	-2114.080	4230.159	4232.764	0.254
$\kappa_2 = 0$	0.527	0.473	0.000	-79.217	160.433	163.038	0.118
BIC	0.527	0.473	0.000	-79.217	160.433	163.038	0.118

Example 3.1. Genotype data for two individuals from [11] are presented in Table 2. The first line of Table 3 gives the 'Standard' estimate obtained by unrestricted optimisation of the log likelihood function. Next follows the estimate constrained to the permissible region based on reparametrisation. The subsequent three lines list the estimates on the border of the permissible region. The BIC values are minised for $\kappa_2 = 0$ and therefore these values are reproduced as the BIC choice in the last line of the table. The rightmost column gives the estimate of the kinship coefficient based on the κ–values. All three estimates indicate that the individuals could be half siblings. However, statistical tests or confidence regions are needed for a more formal conclusion as next. Figure 2 shows the result of 100 samples from the constrained

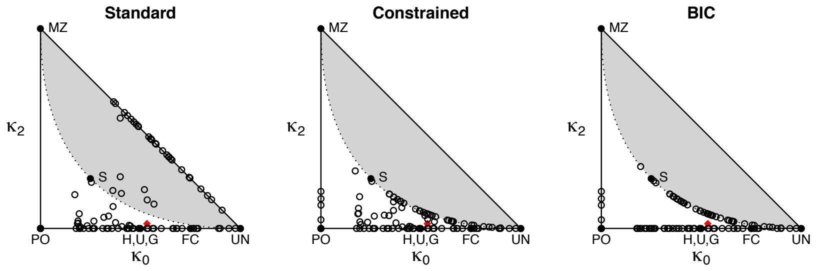


Figure 2: Figure for last part of Example 1 based on 100 samples from $\hat{\kappa} = (0.532, 0.444, 0.024)$ (red diamond).

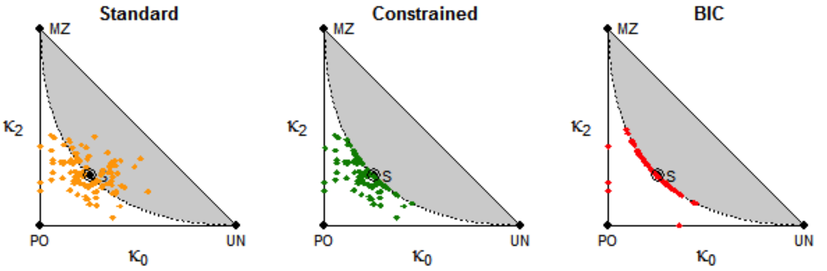


Figure 3: Marker data is simulated for the 35 markers of the database `NorwegianFrequencies` assuming a full-sib relationship (S), i.e., $\kappa_0 = \kappa_2 = 0.25$. Plots for the three estimation methods are presented.

estimate $\hat{\kappa} = (0.532, 0.444, 0.024)$. The variability is great and indicates that 10 markers are insufficient for a reliable estimate.

Example 3.2. The purpose of this example is to exemplify in some detail and compare also the different methods of estimating κ and also ψ. We do 100 simulations using marker data from the 35 loci available as `NorwegianFrequencies` in the `R` library `Familias`. Consider first a full sib relationship. The simulations are shown in Figure 3 for the three methods 'Standard', 'Constrained' and 'BIC'. Note that the penalising term draws the BIC estimate towards the boundaries. Next we consider quadruple half first cousins. As opposed to the previous example, this relationship is an interior point of the triangle. From Figure 4 we see that the BIC estimates are again drawn towards the boundary. The simulations are summarised in Table 4 for the κ estimates. Table 5 shows estimates and MSE for ψ for the above

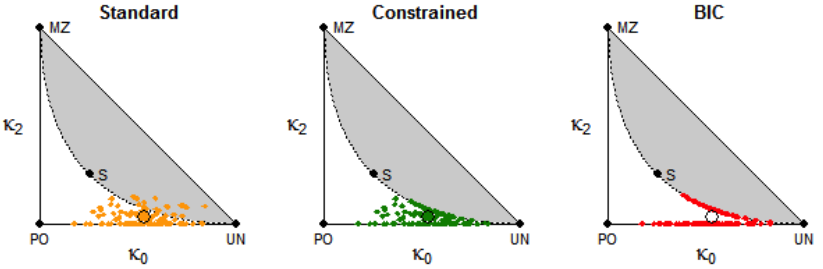


Figure 4: Marker data is simulated for the 35 markers of the database `NorwegianFrequencies` assuming a Q relationship, i.e., $\kappa_0 = 17/32$, $\kappa_2 = 1/32$. Plots for the three estimation methods are presented.

Table 4: Estimates and percentiles for κ estimates for the sibling ($\kappa = (0.25, 0.50, 0.25)$) and Q ($\kappa = (0.53125, 0.43750, 0.03125)$) cases of Example 2, see also Figures 3 and 4.

	Siblings			Q		
	$\hat{\kappa_0}$	$\hat{\kappa_1}$	$\hat{\kappa_2}$	$\hat{\kappa_0}$	$\hat{\kappa_1}$	$\hat{\kappa_2}$
Standard	0.2451	0.4940	0.2610	0.5239	0.4423	0.0338
2.5%	0.0258	0.2676	0.1043	0.2623	0.1736	0.0000
97.5%	0.4564	0.6952	0.4422	0.8007	0.7185	0.1284
Constrained	0.2198	0.5373	0.2429	0.5171	0.4540	0.0288
2.5%	0.0276	0.4662	0.1043	0.2636	0.2249	0.0000
97.5%	0.3714	0.6951	0.4229	0.7668	0.7195	0.1010
BIC	0.2438	0.4972	0.2590	0.5235	0.4477	0.0288
2.5%	0.0437	0.4460	0.1328	0.2606	0.2251	0.0000
97.5%	0.3713	0.6197	0.4256	0.7666	0.7394	0.1221

relationships and two more.

Example 3.3. In the last example we make a more serious, but not definitive, attempt at determining the best of the three estimates ‘Standard’, ’Constrained’ and ’BIC’. We simulate 100 times with 10 and 25 markers. There are 10 alleles with frequencies proportional to $i/10$, $i = 1, \dots, 10$. For real data, we would have estimated κ^* and the parametric bootstrap would be based on this estimate. In this simulation study, our point of departure is κ corresponding to the sib case and in the second case, $\kappa = (\frac{1}{8}, \frac{6}{8}, \frac{1}{8})$, abbreviated MI below. Recall that there exists at least one pedigree with this κ according to [16]. We use parametric bootstrap to sample genotypes. Table 6 shows $MSE(\hat{\psi})$ for four different relationships. Figures 5–8 display the

Table 5: Estimates and percentiles for the ψ estimates in Example 2.

	PO ($\psi = 0.25$)		Sibs ($\psi = 0.25$)		HS ($\psi = 0.125$)		Q ($\psi = 0.125$)	
	mean	MSE	mean	MSE	mean	MSE	mean	MSE
Stand	0.2568	0.0002	0.2540	0.0017	0.1295	0.0012	0.1275	0.0013
Constr	0.2547	0.0002	0.2558	0.0017	0.1296	0.0012	0.1279	0.0013
BIC	0.2567	0.0001	0.2538	0.0017	0.1285	0.0011	0.1263	0.0012

simulations. Finally, we tried with 100 markers (more markers than can be assumed independent), see Figures 9 and 10.

Table 6: $MSE(\hat{\psi})$ for full sibs (FS), half sibs (HS), $\kappa = (\frac{1}{8}, \frac{6}{8}, \frac{1}{8})$ (MI) and quadruple half first cousins (Q). The number of markers is 10 and 25 as indicated.

	Standard	Constrained	BIC
FS:10	0.00530	0.00523	0.00528
FS:25	0.00223	0.00223	0.00227
HS:10	0.00445	0.00443	0.00432
HS:25	0.00196	0.00194	0.00183
MI:10	0.00273	0.00270	0.00274
MI:25	0.00124	0.00126	0.00146
Q:10	0.00486	0.00487	0.00481
Q:25	0.00160	0.00160	0.00148

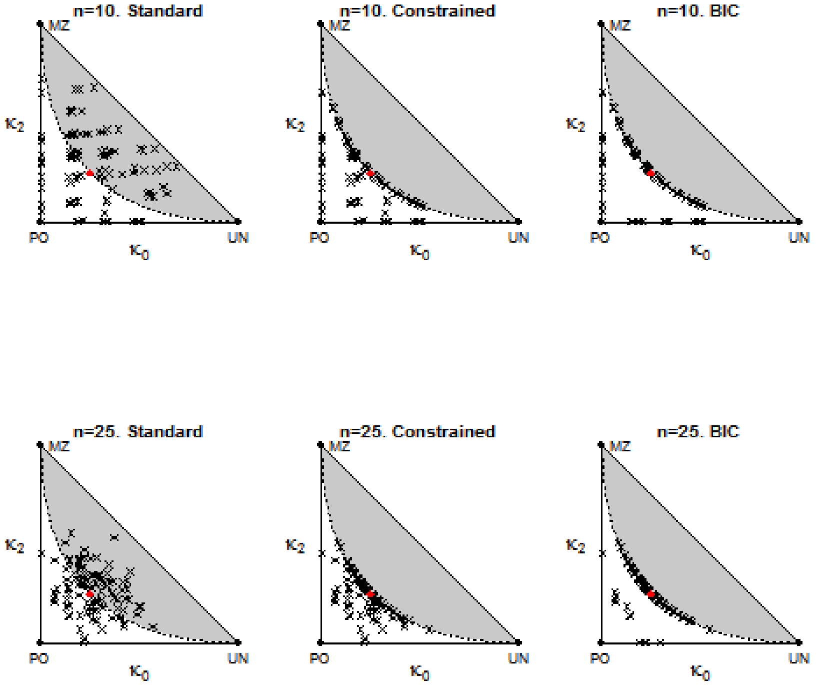


Figure 5: The sibs case in Example 3.3.

4 Discussion

Studies on how generations are affected by matings between related individuals have for many years been of interest both in human genetics and in population structure studies [20, 19, 7]. In this paper we have taken the parametric presentation of IBD parameters for estimating relationships, and suggest a parametric formulation of the hypotheses using κ coefficients. We estimate the relations in question by via the κ coefficients, and find the corresponding confidence intervals and regions (ellipses) using parametric bootstrapping. Boundary challenges appear as many of the common relations are found on the boundary of the valid domain. We have used constrained optimisation based on reparametrisation. As mentioned in connection with Figure 1, the transformation is not continuous. This may possibly lead to numerical instability even if the boundaries are checked separately. Further work should be done to search for and analyse other transformations. It is also relevant

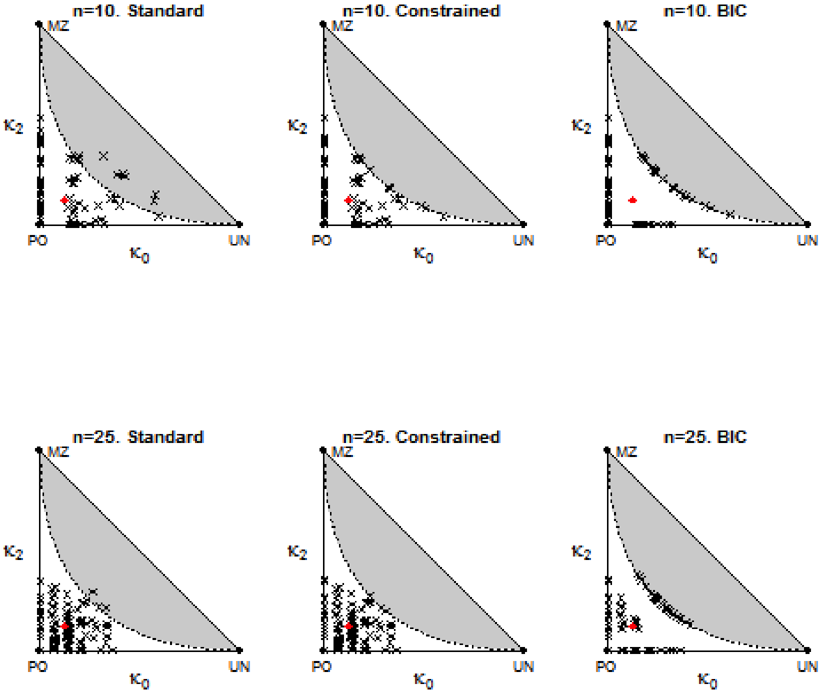


Figure 6: The relationship corresponding to $\kappa = (\frac{1}{8}, \frac{6}{8}, \frac{1}{8})$ discussed in Example 3.3.

to compare the estimators of this paper to alternatives, also for other applications. There is a large literature, [18] is a recent paper. Furthermore, we also tried model selection with the Bayesian Information Criteria (BIC). Using BIC, we first find the best suitable model and thereafter estimate the κ parameters and the kinship coefficient ψ using this model.

Our methods apply to pairwise relations, and it is not straightforward to extend to relationships involving more that two individuals. The problem is that many parameters are then needed as described and exemplified in [6]. Extending the methods to allow for inbreeding is similarly complicated as the number of parameters needed increases substantially. We have assumed markers to be independent (no linkage or disequilibrium) throughout. The examples, see for instance Figure 3, indicate that for some relationships sufficiently accurate estimates are beyond reach based even on 35 markers, corresponding roughly to the information content conventionally available for

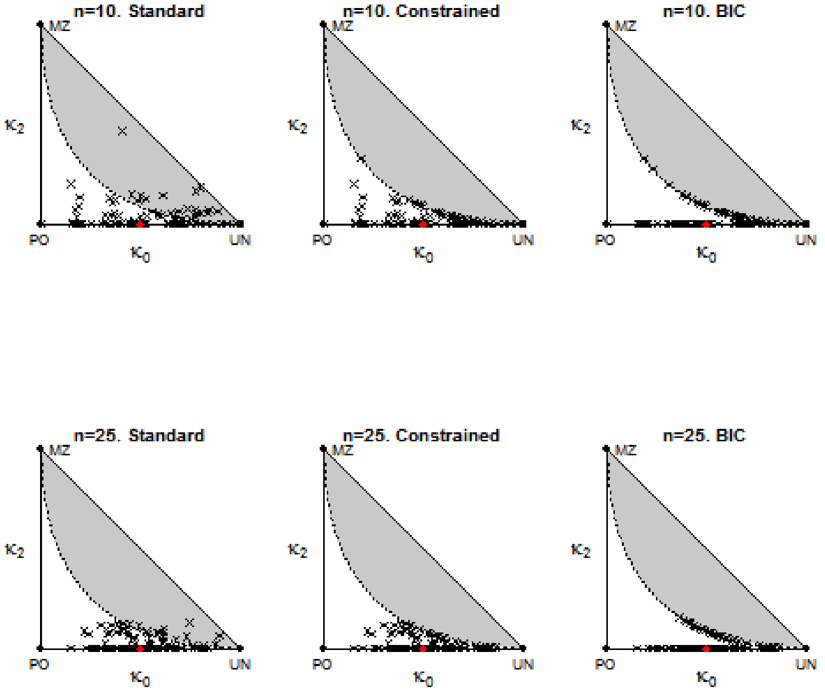


Figure 7: The half sibs case in Example 3.3.

forensic markers.

Asymptotic challenges relevant for κ estimators are discussed in [6]. The main point here is that the likelihood ratio, denoted by Λ in [6], will converge towards different distributions according to the location of κ. For κ values located in the interior of the valid domain, $-2\log\Lambda \xrightarrow{d} \chi^2_2$, while for κ on the boundary, the specific position of κ along the boundary will decide which distribution $-2\log\Lambda$ will converge towards. We will not review specific details here, however $-2\log\Lambda$ converges to a mixture of a discrete and continuous distribution. The asymptotic approach of [6] is limited to SNP markers, and it may be difficult and impractical to extend to forensically relevant markers. The confidence ellipses apparently work well when there are many markers and the relationship correspond to an inner point of the IBD triangle as in Figure 10. In other cases, see e.g. Figures 5–9, the confidence regions have not been included as we are not confident that they are reliable. Methods need to be developed further to obtain more reasonable confidence

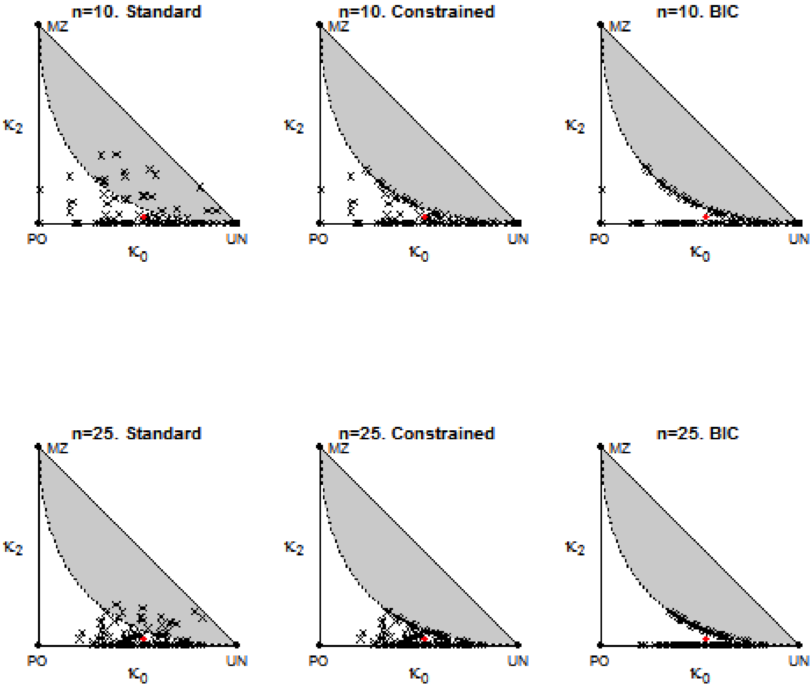


Figure 8: The quadruple half first cousin case in Example 3.3.

regions.

The code has been developed in `R`, and use both existing functions in the package `paramlink`, as well as newly developed functions for the paper found in package `IBDest2`. There exist other packages in `R` for estimating parameters describing pairwise relationships, like for instance package `Demerelate`, see [10], but the constrained maximum likelihood we present and implement appears novel in this context and is importantly not restricted to SNP markers. The examples show that BIC tends to estimate relationships on the boundary of the IBD triangle where many of the well known pedigrees are. Based on the plots and to a lesser extent $MSE(\hat{\psi})$, we recommend the 'Constrained' estimate if the objective is to find the underlying pedigree.

In summary, we hope that this paper has demonstrated the relevance of the parametric approach. We acknowledge the limitations, particularly the restriction to pairwise and also non–inbred relationships. However, it is important to provide and assess methods supplementing the classical verbally

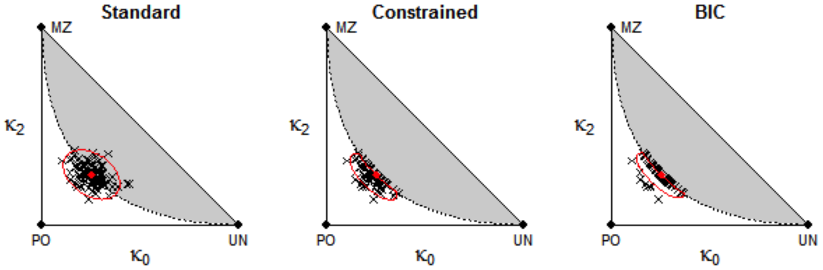


Figure 9: Simulations and confidence ellipses using 100 markers for the sib case, see Example 3.3.

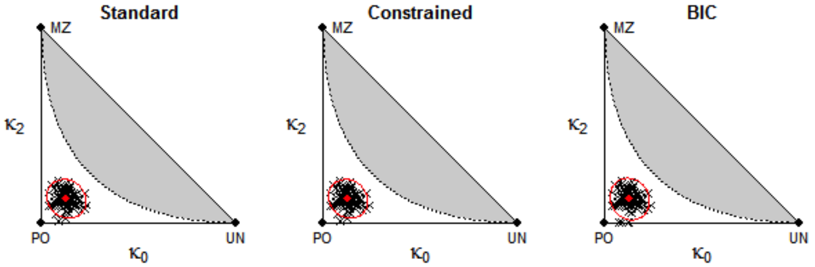


Figure 10: Simulations and confidence ellipses using 100 markers for $\kappa = (\frac{1}{8}, \frac{6}{8}, \frac{1}{8})$, see Example 3.3.

based approach of forensics.

References

[1] D. W. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000.

[2] G. Claeskens and N. L. Hjort. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge, 2008.

[3] C. W. Cotterman. *A calculus for statistico-genetics*. PhD thesis, The Ohio State University, 1940.

[4] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

[5] T. Egeland, D. Kling, and P. Mostad. *Relationship Inference with Familias and R: Statistical Methods in Forensic Genetics.* Academic Press, 2015.

[6] M. García-Magariños, T. Egeland, I. López-de Ullibarri, N. L. Hjort, and A. Salas. A parametric approach to kinship hypothesis testing using identity-by-descent parameters. *Statistical applications in genetics and molecular biology*, 14(5):465–479, 2015.

[7] J. Hajnal, M. Fraccaro, J. Sutter, and C. Smith. Concepts of random mating and the frequency of consanguineous marriages [and discussion]. *Proceedings of the Royal Society of London B: Biological Sciences*, 159(974):125–177, 1963.

[8] A. Henningsen and O. Toomet. `maxLik`: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3):443–458, 2011.

[9] N. Kaur, A. E. Fonneløp, and T. Egeland. Regression models for DNA-mixtures. *Forensic Science International: Genetics*, 11:105–110, 2014.

[10] P. Kraemer and G. Gerlach. `Demerelate`: calculating inter-individual relatedness for kinship analysis based on co-dominant diploid genetic markers using R. *Molecular Ecology Resources*, 2017.

[11] M. Kruijver, R. Meester, and K. Slooten. p-Values should not be used for evaluating the strength of DNA evidence. *Forensic Science International: Genetics*, 16:226–231, 2015.

[12] D. Murdoch and E. Chow. A graphical display of large correlation matrices. *The American Statistician*, 50(2):178–180, 1996.

[13] M. Nothnagel, J. Schmidtke, and M. Krawczak. Potentials and limits of pairwise kinship analysis using autosomal short tandem repeat loci. *International journal of legal medicine*, 124(3):205–215, 2010.

[14] E. A. Thompson. The estimation of pairwise relationships. *Annals of Human Genetics*, 39(2):173–188, 1975.

[15] E. A. Thompson. A restriction on the space of genetic relationships. *Annals of Human Genetics*, 40(2):201–204, 1976.

[16] E. A. Thompson. *Pedigree analysis in human genetics.* Johns Hopkins University Press, 1986.

[17] E. A. Thompson. Statistical inference from genetic data on pedigrees. In *NSF-CBMS regional conference series in probability and statistics*, pages i–169. NSF-CBMS regional conference series in probability and statistics, 2000.

[18] B. S. Weir and J. Goudet. A unified characterization of population structure and relatedness. *bioRxiv*, pages 1–37, 2016.

[19] S. Wright. The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354, 1949.

[20] S. Wright. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, pages 395–420, 1965.