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

Philosophiae Doctor (PhD)
Thesis 2021:48

Relationship inference and inbreeding in forensic genetics

Slektskap og inngifte i rettsgenetikk

Hilde Kjelgaard Brustad

Relationship inference and inbreeding in forensic genetics

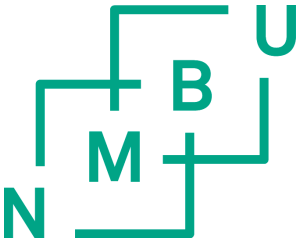
Slektskap og inngifte
i rettsgenetikk

Philosophiae Doctor (PhD) Thesis

Hilde Kjelgaard Brustad

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

Ås 2021



Thesis number 2021:48
ISSN 1894-6402
ISBN 978-82-575-1819-6

Acknowledgements

Thore, what a journey this have been! I am forever grateful for your guidance and support. You have shown such an interest in my project, which has been a great motivation throughout the last three years. I appreciate the way you have introduced me to the forensic community, both in Norway and abroad. You have been patient with me, but you have challenged me. Thank you for coping with my stubbornness in discussions. The last year has been out of the ordinary, but our digital meetings on a weekly basis have made everything easier.

I want to thank Magnus Dehli Vigeland for interesting discussions. Your attention to details and the effort you put into the finishing product are impressing. You have had a significant impact on my writing skills. Co-supervisor Geir Storvik, your strong statistical knowledge is inspiring. I am grateful for the interest you have shown in my project, even though forensics is not your field of expertise. Thank you for asking the right questions.

Guro Dørum, Hilde Vinje and Elida Fonneløp, thank you all for making time to proofread the thesis.

The research group of Bioinformatics and Applied Statistics at NMBU, you have been a joy to be a part of.

I want to thank my family, for always telling me to "just go for it". Without your support I would not have started on this PhD adventure.

And last, but not least, Thomas, my rock in life! Thank you for letting me follow my path. You are the most patient man I know. During ups and downs, you stand steady, keeping track of our girls and making life as easy as possible.

List of papers

I. H. K. Brustad and T. Egeland. The impact of ignoring inbreeding in pairwise kinship evaluations. *Forensic Science International: Genetics Supplement Series*, 7(1):462-464, 2019.

II. H. K. Brustad, M. D. Vigeland and T. Egeland. Pairwise relatedness testing in the context of inbreeding: expectation and variance of the likelihood ratio. *International Journal of Legal Medicine*, 135(1):117-129, 2021.

III. H. K. Brustad, M. Colucci, M. A. Jobling, N. A. Sheehan and T. Egeland. Strategies for pairwise searches in forensic kinship analysis. Submitted.

IV. H. K. Brustad, G. O. Storvik, M. D. Vigeland, T. Egeland. Estimating realised pairwise relatedness with bootstrap confidence. Manuscript

Summary

Inference of relatedness between individuals is of interest in fields as disparate as animal breeding, archaeology, medical genetics and genealogy. In the forensic context, emphasised in this thesis, applications include paternity cases, pedigree reconstruction and disaster victim identification.

We focus on two approaches to kinship inference: The likelihood ratio (LR) framework and estimation of relatedness parameters. The former compares two hypotheses, while the latter estimates the most likely set of relatedness parameters to describe the observations. We restrict attention to relatedness between pairs of individuals.

Statistical properties of the LR for non-inbred relationships are already derived in the literature. We extend this work to allow for inbred individuals, by modelling pairwise relationships through the Jacquard coefficients.

Further, we investigate LR testing in the context of database searches, with focus on a search type called blind search. A blind search comprises a large number of pairwise comparisons and the expected performance of this search type needs to be addressed. In particular, we demonstrate how to decide on optimal LR thresholds to control the Family Wise Error Rate.

A Bayesian framework for kinship testing can be useful, particularly when there are more than two hypotheses. We demonstrate this by including informative prior probabilities to help distinguish between hypotheses.

Maximum likelihood estimation of relatedness is addressed in the thesis. The uncertainty of the estimate is investigated through parametric and non-parametric bootstrapping. We discuss the difference between these methods and how they attempt to mimic the random process that created the genetic data of two individuals.

Sammendrag

Avhandlingen handler om testing og estimering av familierelasjoner mellom individer. Det er mange anvendelser, f.eks. innen oppdrett, arkeologi, medisinsk genetikk og slektskapsforskning. Anvendelsesområder i en rettsgenetisk sammenheng er blant andre farskapstesting, rekonstruksjon av slektstrær og identifikasjon av omkomne etter katastrofer.

I denne avhandlingen fokuserer vi på to tilnærminger til inferens av slekskap: LR ('Likelihood ratio') rammeverket og estimering av slektskapsparametre. Den første tilnærmingen sammenligner likelihood for to hypoteser, mens den andre fremgangsmåten søker etter det mest sannsynlige slektskapet blant alle mulige slektskap. Analysene i denne avhandlingen er begrenset til parvise slektskap.

Statistiske egenskaper for LR for ikke-inngiftede individer er utledet i litteraturen. I avhandlingen utvider vi dette arbeidet til slektskap generelt, ved bruk av Jacquard koeffisienter for modellering av slektskap.

Videre undersøker vi bruk av LR testing for databasesøk. Vi konsentrerer oss om en søketype kalt 'blind search', der alle par i en database blir testet for visse slektskap. Potensielt mange tester blir utført i et slikt søk og det er derfor behov for å evaluere de statistiske egenskapene til slike søk. Vi demonstrerer hvordan vi kan finne en optimal terskelverdi for hver LR test, slik at 'Family Wise Error Rate' holdes på et akseptabelt nivå.

En Bayesisk tilnærming til slektskapstesting kan være nyttig spesielt når det er mer enn to hypoteser. Vi demonstrerer dette ved bruk av informative a priori sannsynligheter for de forskjellige hypotesene.

Avhandlingens siste del omhandler estimering av slektskapsparametre. Parametrisk og ikke-parametrisk bootstrap blir brukt til å undersøke usikkerheten i estimatene. Vi diskuterer forskjellen mellom disse to bootstrapmetodene og hvordan de forsøker å etterligne den stokastiske prosessen som gir opphav til de genetiske dataene.

Contents

1 Introduction

Inference of relatedness between individuals is important in several scientific disciplines. Animal breeding, plant breeding, medical genetics and archaeological studies are some examples. This thesis focuses on relatedness inference in forensic genetics. Forensic genetics refers to cases involving DNA evidence in legal settings. The genetic similarity between the genomes of individuals contains information about how the individuals are related. Statistical methods are applied to infer relationships, based on observations of DNA data. Examples of applications of relatedness inference in forensic genetics are paternity testing, disaster victim identification (DVI), pedigree reconstruction and immigration cases. This thesis limits its attention to pairwise relatedness, i.e., relationships between pairs of individuals.

It is important to distinguish between pedigree relationships and the realised relatedness. The former refers to the pedigree connecting the individuals. Legislation in the society are written on the basis of pedigree relationships for instance when specifying that siblings are not allowed to marry.

One measure of relatedness is the level of identity-by-descent (IBD) between genomes [1]. The pedigree relationship quantifies the expected level of IBD sharing between individuals, with respect to a pedigree connecting them. This differs from the realised relatedness that describes the actual genetic similarity between individuals [2]. For instance, a pair of first cousins are genetically unrelated if there is no IBD sharing for the markers investigated.

Two statistical methods for kinship inference are likelihood ratio (LR) testing and IBD estimation. The LR framework compares the likelihood of two hypotheses, formulated as two relationships. The analysis is limited to these two alternatives. In a standard paternity case, the hypothesis that a man is the father of a child is compared to the hypothesis that they are unrelated. However, the truth may be that the alleged father is an uncle of the child. The two tested hypotheses and the true relationship are expressed parametrically by the IBD coefficients. The expectation and variance of the LR as a function of these IBD coefficients are derived in the literature. The work in this thesis extends these formulas to also apply for inbred individuals.

Database searches are important in a forensic context. Blind search is a type of search where all individuals in a database are compared and an LR is computed for each pair [3]. A blind search is used as an initial step in e.g. a DVI case. Many LRs are computed in a blind search. The number of false positives, also called type I errors, thus needs to be controlled. An approach for choosing the optimal LR threshold in a multiple testing scenario needs to be applied.

Estimation aims to find the most likely relationship among all possible alternatives. Maximum likelihood (ML) estimation finds the level of IBD sharing between two individuals that most likely explains the observed DNA data. A measure of uncertainty should accompany an estimate. Thompson [4], Milligan [5] and Anderson and Weir [6] study estimation of IBD based parameters. The latter assesses the uncertainty of the estimate through non-parametric bootstrapping. However, the literature does not seem to contain a proper discussion of the use of different bootstrapping methods in this forensic context and the topic is therefore explored in this thesis.

A common assumption in kinship testing is that individuals are non-inbred. This is often an adequate assumption. However, investigators may need tools that take inbreeding into account, for instance in cases where incest is suspected. An important focus of this thesis is therefore modelling and inference of inbred relationships.

1.1 Genetics

1.1.1 DNA, chromosomes and recombination

The human genome consists of about $3 \cdot 10^9$ base pairs, made from the nucleotides adenine (A), thymine (T), cytosine (C) and guanine (G). They form the well-known double helix structure of the DNA. Most of the human genome is the same from individual to individual, but we target the areas that vary between individuals. The DNA strand is divided into 23 pairs of chromosomes. Chromosome pairs 1 to 22 are autosomal chromosomes. Chromosome pair 23 contains the sex chromosomes. Females have two X-chromosomes and males have an X- and a Y-chromosome. The DNA is located in the nucleus of our cells. An egg cell and a sperm cell, called gametes, contain one chromosome from each chromosome pair. The DNA of a child thus roughly consists of half of the mother's DNA and half of the father's DNA. Figure 1 illustrates how chromosome pair 1 can be inherited from two parents to a child. Each parent has two copies of chromosome 1, shown with different colours. Crossovers occur when a gamete is produced. The result is that the child has a recombined version of chromosome pair 1 from the mother and a recombined version from the father. This is shown by the mix of colours along the chromosomes of the child.

The genetic distance d between two points along a chromosome is defined as the expected number of crossovers between the points. This unit is called Morgan (M), named after Thomas H. Morgan (1866-1945). A distance of 1 cM between two points corresponds to an average of 0.01 crossovers between the points. The human genome is about 30M. The female genome recombines more than the male genome, and hence, the female genome is longer.

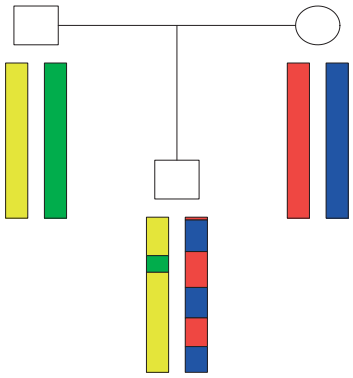


Figure 1: Illustration of recombination along a chromosome, as explained in Section 1.1.1.

1.1.2 Genetic markers

A specific location or segment of the genome is called a locus or a genetic marker. Different parts of the DNA have different functions. Some parts code for diseases. The genetic markers used in forensic applications are located in so-called non-coding regions of the genome. Two main types of markers used in forensics, STRs and SNPs, are explained next. The short tandem repeat (STR) markers, also called microsatellites, consist of repeating sequences of bases, usually from two to six bases. In general, the number of repeats makes an allele. For instance, if the sequence of bases ATCA is repeated 10 times in a row, the allele is called 10. Different number of repeats at a locus are possible, creating a set of possible alleles for a marker. Diploid organisms have two alleles at a locus, one for each chromosome of a chromosome pair. These two alleles define the genotype of a locus. A DNA profile consists of genotypes from several loci. The alleles of the STR markers and the possible genotypes vary between individuals, making them suitable for identification purposes.

Single nucleotide polymorphisms (SNPs) is another group of genetic markers. A SNP marker is made from alteration of a single base pair of the DNA. In general, the SNPs are diallelic, i.e., there are two possible alleles at each locus [7]. Because the SNPs are diallelic, a SNP marker contains less information about an individual than an STR marker [8]. However, there are more SNPs in the genome than STRs. An advantage with the SNP markers is that they are more robust against degradation, due to their short length, compared to an STR marker.

STRs are the standard genetic markers used in forensic applications. The markers are thoroughly investigated through the years and standards regarding the markers are agreed

on [9]. A forensic DNA profile is typically genotyped for 16 to 35 autosomal STR markers. They are chosen such that linkage between them are vanishingly small. A combination of marker types can be needed for difficult cases, like resolving distant relationships [8].

Artefacts like drop-ins, drop-outs, stutters, silent alleles and genotyping error affect the quality of the DNA profile [10]. In this thesis, these artefacts are not modelled.

1.1.3 Hardy-Weinberg equilibrium, linkage disequilibrium and θ-correction

The meaning of the term *population* may differ depending on the application. We will describe a population as a set of people from a specific geographical region, or a socially enclosed group of individuals. Some genotypes and alleles are more commonly observed than others in a population. If a population is in Hardy-Weinberg Equilibrium (HWE), the genotype frequencies stay the same from one generation to the next. The assumptions underlying HWE are an infinite large population with random mating, no selection, migration, population stratification or mutation. Consider a locus with alleles a and b with population allele frequencies p_a and p_b, respectively. The possible genotypes in the population are a/a, a/b and b/b, with corresponding genotype frequencies p_a^2, $2p_ap_b$ and p_b^2.

If the assumptions for HWE are not met, the above genotype frequencies do not apply. Population stratification, for instance caused by a high level of background relatedness, cause an excess of homozygous individuals. This can be adjusted for by a θ-correction. The co-ancestry coefficient θ is defined on the interval $[0, 1]$. A population in HWE has $\theta = 0$. The level of background relatedness is typically in the interval $\theta = [0.01, 0.03]$ [11]. The so-called sampling formula adjusts the allele frequencies, depending on the number of observed alleles of a type [3]. The probability of observing allele a as the first allele is p_a. Let b_j be the number of times allele a is observed among the $j-1$ previously observed alleles. The probability that allele a is sampled as the j'th allele is

$$p_a' = \frac{b_j\theta + (1-\theta)p_a}{1 + (j-2)\theta}.$$

In practice, the assumptions for HWE are never met completely. For instance, a population always has a finite size.

Another population based concept is linkage disequilibrium (LD). LD refers to non-random allelic association across loci in a population [12]. If a haplotype, i.e., a specific combination of alleles at two or more loci, is observed more frequently in a population than expected when assuming independence, the population is in LD.

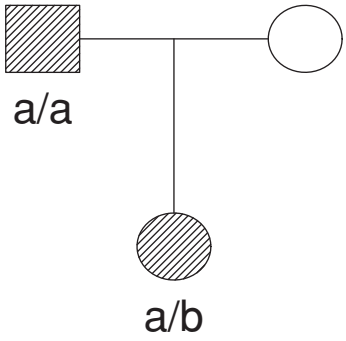


Figure 2: Pedigree for the example in Section 1.1.4. Father and daughter (the shaded individuals) are genotyped for one marker.

1.1.4 Inheritance of alleles and linkage

Gregor Mendel established in the mid 19th century a set of principles for how traits are inherited from a parent to its offspring. He experimented with cross breeding of pea plants, recording traits as colour, size and shape. He concluded that each parent pass on an element to an offspring. The trait observed corresponds to one of the elements, they are not blended in the offspring. Furthermore, different traits are inherited independently of each other. The latter proved to not necessarily be correct. The modern understanding of Mendelian inheritance says that there is a 50% chance that either of the alleles at a locus is passed on from a parent to a child. Consider the pedigree in Figure 2. The genotypes of the father is a/a, with allele frequency p_a and genotype frequency p_a^2, assuming the population is in HWE. The daughter has the genotype a/b. The probability that allele a is passed on from the father to the daughter is 1. The mother is not genotyped, and thus, the probability of observing allele b is its population allele frequency p_b. The joint probability of the genotypes of the father and daughter is $p_a^2 p_b$. Two loci located on different chromosomes are regarded as independent of each other. Alleles at loci located on the same chromosome are in general not inherited independently, however, if the loci are located on different ends on the chromosome, they are regarded as independent. Chromosomes are passed on in segments. Assume that the maternal allele at a locus of a parent is passed on to a child. The maternal allele at a second locus, close to the first locus, is then likely to also be passed on the the child. Linkage between loci is modelled by the recombination rate ρ. Loci on different chromosomes are unlinked, hence $\rho = 0.5$. Fully linked loci have $\rho = 0$, meaning that no crossover occurs between the loci. Two loci are recombined if an odd number of

crossovers have occurred. Haldane proposed a function to relate the genetic distance d to the recombination rate ρ [13]. The function

$$\rho = \frac{1}{2}(1 - e^{-2d})$$

is called Haldane's mapping function and models the number of crossovers along the genome as a Poisson process where subsequent crossovers occur independently of each other. This is not necessarily true, because the probability of a crossover event typically increases with increasing distance from the last event.

1.1.5 Mutations

A mutation is the result of a change in the sequence of bases at a locus. A mutation can be somatic or in the germline. A germline mutation is present in the reproductive cells of an individual and is passed on from a parent to an offspring. This is not the case for somatic mutations. When computing pedigree likelihoods, mutations are accounted for through a mutation matrix M. Element $m_{i,j}$ gives the probability that allele i of a parent ends up as allele j in the child. Stationary mutations models leave the population allele frequencies unchanged from generation to generation [14]. The parametric form of the likelihood function (2) does not model mutations. An expression for the likelihood of a parent-offspring relationship that accounts for mutations is given in (3). Mutations are discussed further in Section 1.3.

1.2 Relatedness

1.2.1 What does it mean to be related?

The definition of relatedness differs depending on the application. Legislation in the society builds on a non-genetic definition of relatedness. For instance, in many countries an adopted child and a biological child are not distinguished between by the law. In forensic genetics, relatedness is associated with sharing of genetic material. A *pedigree relationship* specifies the possibility that individuals share genetic material. Siblings (S), for instance, have the possibility to share genetic material from their parents. If we search far enough back in generations, most individuals will be related. In this thesis, a pedigree relationship refers to a pedigree limited to a few generations. The inference of pedigree relationship between individuals enables the results from statistical analyses to be used in legal settings.

From a pedigree point of view, two individuals are expected to share a certain amount of genetic material. However, the actual amount of genome sharing between two individuals

can differ from the expected. We call this the *realised relatedness*.

1.2.2 Identity-by-descent coefficients

Two homologous alleles are identical by state (IBS) if they are the same version of the allele. If the IBS alleles are inherited from a common ancestor, i.e., the alleles are autozygous, they are IBD. Browning et al. distinguish between ancient IBD, recent IBD and familial IBD, depending on the definition of common ancestor [15]. For the applications in this thesis, IBD is defined within a given pedigree of only a few generations, i.e., familial IBD. The idea is that closely related individuals share more of their genomes IBD than more distantly related individuals.

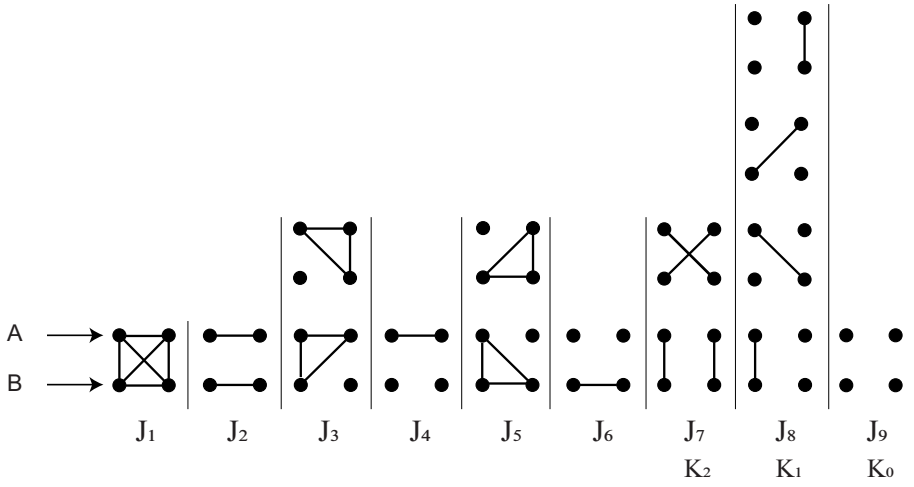


Figure 3: The 15 possible IBD configurations for the four alleles of the individuals A and B. Dots denote alleles and lines connect IBD alleles. Ignoring the difference between maternal and paternal alleles reduces the 15 configurations to the nine Jacquard states $J_1, \cdots, J_9$. Only the three last states are possible for outbred individuals, denoted as the IBD states K_0, K_1 and K_2.

There are several coefficients quantifying IBD sharing between and within individuals. The coefficient of inbreeding f, introduced by Wright in the early 1920's, defines the probability that two alleles at a locus are IBD [1]. If the parents of a child are unrelated, then the child is outbred, i.e., $f = 0$ for the child. The kinship coefficient φ between two individuals is the probability that a random allele from one individual is IBD to a random homologous allele from another individual. While f quantify how inbred an individual is, φ describes the relationship between two individuals.

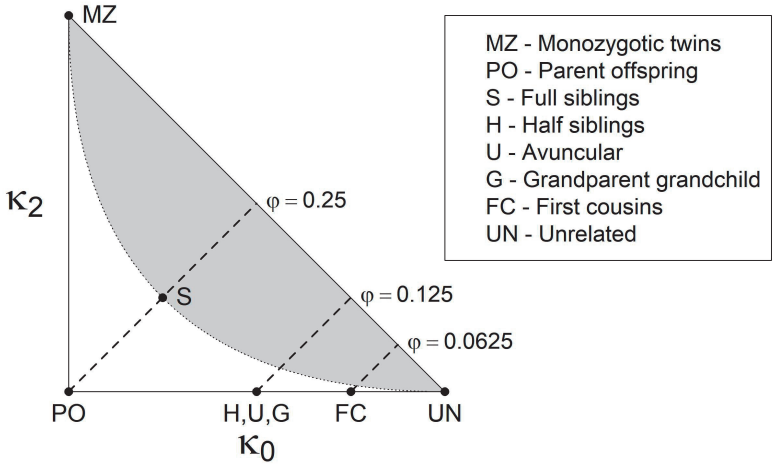


Figure 4: The IBD triangle with location of some common relationships. The dashed lines show how the kinship coefficient φ relates to the IBD coefficients $\boldsymbol{\kappa}$. The white area is the admissible region for $\boldsymbol{\kappa}$. The same figure occurs in Paper IV.

The four homologous alleles of two individuals must be in one of the 15 unique IBD states shown in Figure 3, often reduced to the nine states $J_1, \ldots, J_9$ called the Jacquard states [16]. For outbred individuals, only the last three states J_9, J_8 and J_7, denoted K_0, K_1 and K_2, respectively, are possible. The so-called k-coefficients for two individuals were introduced by Cotterman [17]. A more frequently used variant of these are the IBD coefficients $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$ [4]. These IBD coefficients define the probabilities that two outbred individuals share zero, one or two alleles IBD, i.e., $\kappa_i = P(K_i)$ for $i = 0, 1, 2$. Without considering actual genetic material, the coefficients are computed based on a pedigree structure and define the expected proportion of the genomes in the different IBD states. Because $\kappa_0 + \kappa_1 + \kappa_2 = 1$, the coefficients can be visualised in the IBD triangle by the coordinates (κ_0, κ_2). Figure 4 shows the location of $\boldsymbol{\kappa}$ for some common pedigree relationships. The pedigree based IBD coefficients are restricted by the inequality $\kappa_1^2 \geq 4\kappa_0\kappa_2$, limiting the coefficients to be located in the white area of the triangle [18].

The Jacquard coefficients $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_9)$ extend the IBD coefficients to also model inbred relationships [16]. Similar to $\boldsymbol{\kappa}$, Jacquard coefficients are defined as $\Delta_i = P(J_i)$ for $i = 1, \ldots, 9$. The coefficients sum to one. A restriction of the Jacquard coefficient similar to the restriction for $\boldsymbol{\kappa}$ is not known to the author.

The Jacquard coefficients are related to the kinship coefficient through the equation

$$\varphi = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8.$$

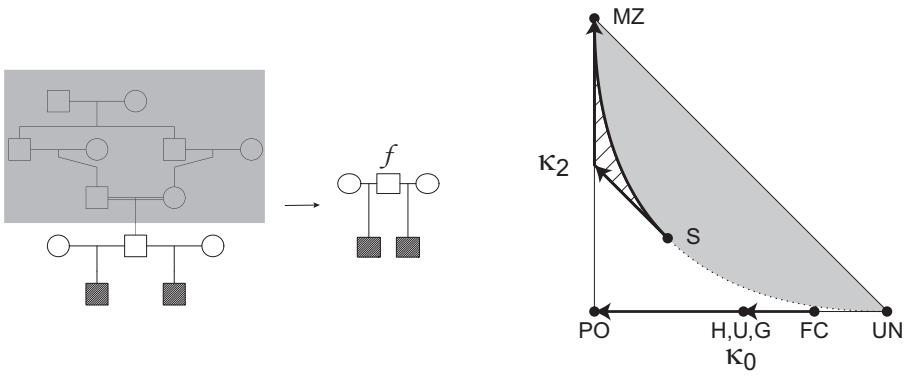


Figure 5: The concept of modelling background relatedness by founder inbreeding, exemplified by the figure presented in Paper III. The common father of the half siblings is inbred. The IBD triangle shows how $\boldsymbol{\kappa}$ changes with increasing f, as presented in Paper II.

For outbred relationships, this equation reduces to

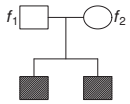
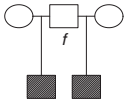
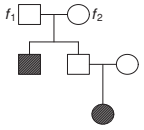
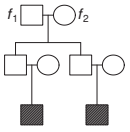
$$\varphi = \frac{1}{2}\kappa_2 + \frac{1}{4}\kappa_1.$$

The dashed lines of Figure 4 show points in the IBD triangle with the same kinship coefficient. For instance, siblings and parent-offspring have the same kinship coefficient $\varphi = 0.25$. Table 1 shows φ for some common relationships.

1.2.3 Founder inbreeding

The θ-correction accounts for background relatedness on a population level. The genotype frequencies of all founders of a pedigree are affected. By assigning an inbreeding coefficient to one or more founders of a pedigree, background relatedness can be modelled on a more local level [19, 20]. Figure 5 shows the situation. We are interested in the relationship between a pair of half siblings. The common father is inbred, his parents are cousins. The pedigree above the father can be replaced by the inbreeding coefficient $f = 0.0625$. The coefficient f contains the information needed to calculate the IBD coefficients of the half siblings. Even though the common father is inbred, the half siblings are outbred. Table 1 shows φ and $\boldsymbol{\kappa}$ as a function of f for some common relationships. These functions, as presented in Paper II, are visualised in Figure 5. As f increases from 0 to 1, the IBD coefficients of half siblings (H), avuncular (U) and grandparent-grandchild (G) moves from the point $(0.5, 0)$ to $(0, 0)$, the point of PO. A first cousin (FC) relationship moves along the

Table 1: Relatedness coefficients as functions of founder inbreeding, in a selection of common relationships. The same table occurs in Paper II.

	Rel.	φ	$\varphi(f)$	$\boldsymbol{\kappa}$	$\boldsymbol{\kappa}(f)$
	S	$\frac{1}{4}$	$\frac{1}{4}(1+\frac{f_1+f_2}{2})$	$(\frac{1}{4},\frac{1}{2},\frac{1}{4})$	$\kappa_0(f_1,f_2)=\frac{1}{4}(1-f_1)(1-f_2)$ $\kappa_1(f_1,f_2)=\frac{1}{2}(1-f_1f_2)$ $\kappa_2(f_1,f_2)=\frac{1}{4}(1+f_1)(1+f_2))$
	H	$\frac{1}{8}$	$\frac{1}{8}(1+f)$	$(\frac{1}{2},\frac{1}{2},0)$	$\kappa_0(f)=\frac{1}{2}(1-f)$ $\kappa_1(f)=\frac{1}{2}(1+f)$ $\kappa_2(f)=0$
	U	$\frac{1}{8}$	$\frac{1}{8}(1+\frac{f_1+f_2}{2})$	$(\frac{1}{2},\frac{1}{2},0)$	$\kappa_0(f_1,f_2)=\frac{1}{2}(1-\frac{f_1+f_2}{2})$ $\kappa_1(f_1,f_2)=\frac{1}{2}(1+\frac{f_1+f_2}{2})$ $\kappa_2(f_1,f_2)=0$
	FC	$\frac{1}{16}$	$\frac{1}{16}(1+\frac{f_1+f_2}{2})$	$(\frac{3}{4},\frac{1}{4},0)$	$\kappa_0(f_1,f_2)=\frac{1}{4}(3-\frac{f_1+f_2}{2})$ $\kappa_1(f_1,f_2)=\frac{1}{4}(1+\frac{f_1+f_2}{2})$ $\kappa_2(f_1,f_2)=0$

first axis, from $(0.75, 0)$ to $(0.5, 0)$. The genetic similarity between siblings becomes more and more similar to MZ as the inbreeding coefficient of each founder f_1 and f_2 increases.

If the parent of a child is inbred with a coefficient of inbreeding f, the pairwise relationship between the parent and child becomes inbred. The IBD coefficients are then invalid. The Jacquard coefficients for this relationship is $\boldsymbol{\Delta}(f) = (0, 0, f, 0, 0, 0, 0, 1-f, 0)$.

1.2.4 Realised relatedness

Mendelian inheritance and recombination between loci make the actual proportion of IBD sharing between individuals different from the expected values specified in Table 1. It is possible (but very unlikely) for two first cousins to not share any part of their genomes IBD, even though $\kappa_0 = 0.75$. The cousins are then genetically unrelated.

Denote the realised relatedness between two individuals by $\boldsymbol{\Delta}_R = (\Delta_1^R, \ldots, \Delta_9^R)$, or $\boldsymbol{\kappa}_R = (\kappa_0^R, \kappa_1^R, \kappa_2^R)$ for outbred relationships. They define the proportion of the genomes of two individuals to be in the different IBD sharing states.

Two pairs of individuals with the same realised relatedness can have different underlying IBD patterns. One pair may share long but few IBD segments, while another pair may share many short IBD segments. The realised relatedness $\boldsymbol{\kappa}_R$ is not restricted as the IBD coefficients and can therefore be located in the inadmissible region of the IBD triangle.

The realised relatedness for PO is by definition always $\boldsymbol{\kappa}_R = (0, 1, 0)$. For unrelated individuals (UN), the realised relatedness is always $\boldsymbol{\kappa}_R = (1, 0, 0)$. Pedigree relationships with $\kappa_2 = 0$, such as H, U, G and FC, have by definition a realised relatedness with $\kappa_2^R = 0$. Hill and Weir [2] derived expressions for the variation in the realised relatedness for a variety of pedigree relationships.

1.3 Likelihood model of relatedness

The Elston-Stewart algorithm is a powerful tool for computation of pedigree likelihoods, given a set of genetic data [21]. The algorithm is efficient for large pedigrees with relatively few markers. The Lander-Green algorithm computes pedigree likelihoods by applying a hidden Markov chain to model linkage between markers [22]. This algorithm is computationally heavy for pedigrees with many pedigree members. Both the Elston-Stewart algorithm and the Lander-Green algorithm compute likelihoods for a pedigree structure. The work in this thesis considers a parametric approach to computation of the likelihood of the realised relatedness between pairs of individuals.

Let $\mathcal{G}_1, \ldots, \mathcal{G}_M$ denote a set of M non-identically distributed discrete random variables, where variable j corresponds to locus j. The sample space of $\mathcal{G}_j$ is the set of all possible combinations of joint genotypes at locus j. The joint probability mass function of the M independent random variables, conditional on the realised relatedness $\boldsymbol{\Delta}_R$, is given by

$$\mathrm{P}((\mathcal{G}_1, \ldots, \mathcal{G}_M) \mid \boldsymbol{\Delta}_R) = \prod_{j=1}^{M} \sum_{i=1}^{9} \Delta_i \mathrm{P}(\mathcal{G}_j \mid \mathrm{J}_i). \tag{1}$$

The probabilities $\mathrm{P}(\mathcal{G}_j \mid \mathrm{J}_i)$ are given in Table 2.

Example 1: The purpose of this example is to show how (1) is used to compute the distribution of joint genotypes at a locus. Assume a SNP marker with alleles a and b and allele frequencies $p_a = p_b = 0.5$. There are nine possible combinations of genotypes for a pair of individuals. The left table of Table 3 shows the probability of each joint genotype conditioned on $\boldsymbol{\kappa}_R = (0, 1, 0)$. The probability that two individuals are homozygous for

Table 2: The probability $\mathrm{P}(\mathcal{G} \mid \mathrm{J}_i)$ of a pair of genotypes $\mathcal{G} = (g_A, g_B)$, conditioned on the Jacquard state J_i. The letters a, b, c and d represent different alleles, with population frequencies p_a, p_b, p_c and p_d, respectively. This is the same table as in Paper II.

$\mathcal{G}$	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
$a/a, a/a$	p_a	p_a^2	p_a^2	p_a^3	p_a^2	p_a^3	p_a^2	p_a^3	p_a^4
$a/a, b/b$	0	$p_a p_b$	0	$p_a p_b^2$	0	$p_a^2 p_b$	0	0	$p_a^2 p_b^2$
$a/a, a/b$	0	0	$p_a p_b$	$2p_a^2 p_b$	0	0	0	$p_a^2 p_b$	$2p_a^3 p_b$
$a/a, b/c$	0	0	0	$2p_a p_b p_c$	0	0	0	0	$2p_a^2 p_b p_c$
$a/b, a/a$	0	0	0	0	$p_a p_b$	$2p_a^2 p_b$	0	$p_a^2 p_b$	$2p_a^3 p_b$
$b/c, a/a$	0	0	0	0	0	$2p_a p_b p_c$	0	0	$2p_a^2 p_b p_c$
$a/b, a/b$	0	0	0	0	0	0	$2p_a p_b$	$p_a p_b(p_a + p_b)$	$4p_a^2 p_b^2$
$a/b, a/c$	0	0	0	0	0	0	0	$p_a p_b p_c$	$4p_a^2 p_b p_c$
$a/b, c/d$	0	0	0	0	0	0	0	0	$4p_a p_b p_c p_d$

allele a is

$$\begin{aligned}\mathrm{P}(\mathcal{G} = (a/a, a/a) \mid \boldsymbol{\kappa}_R = (0,1,0)) &= 0 \cdot p_a^4 + 1 \cdot p_a^3 + 0 \cdot p_a^2 \\ &= 0.125.\end{aligned}$$

The joint genotype distribution for the same marker conditioned on $\boldsymbol{\kappa}_R = (0.25, 0.5, 0.25)$ is shown in the right table of Table 3. The probability of the genotypes $(a/a, a/a)$ is

$$\begin{aligned}\mathrm{P}(\mathcal{G} = (a/a, a/a) \mid \boldsymbol{\kappa}_R = (0.25, 0.5, 0.25)) &= 0.25 \cdot p_a^4 + 0.5 \cdot p_a^3 + 0.25 \cdot p_a^2 \\ &= 0.1406.\end{aligned}$$

Table 3: Joint genotype distributions from Example 1, for a SNP markers with alleles a and b and allele frequencies $p_a = p_b = 0.5$. a) Conditioned on $\boldsymbol{\kappa}_R = (0,1,0)$. b) Conditioned on $\boldsymbol{\kappa}_R = (0.25, 0.5, 0.25)$.

a)

	a/a	a/b	b/b
a/a	0.125	0.125	0.000
a/b	0.125	0.250	0.125
b/b	0.000	0.125	0.125

b)

	a/a	a/b	b/b
a/a	0.1406	0.0938	0.0156
a/b	0.0938	0.3125	0.0938
b/b	0.0156	0.0938	0.1406

Equation (1) defined the probability distribution of the joint genotypes at a locus. Consider next the situation where we have observed the genotypes of two individuals. The individuals A and B are genotyped for M independent loci. Denote the joint genotypes of the DNA profiles by $G = (G_1, \ldots, G_M)$, where the genotypes at locus j is $G_j = (g_{A,j}, g_{B,j})$. The likelihood of the realised relatedness $\boldsymbol{\Delta}_R$ when we have observed the DNA profiles G is

$$\mathrm{L}(\boldsymbol{\Delta}_R \mid G) = \prod_{j=1}^{M} \sum_{i=1}^{9} \Delta_i \mathrm{P}(G_j \mid \mathrm{J}_i). \tag{2}$$

The likelihood function (2) ignores linkage between markers. Multi-locus IBD coefficients [23, 24] would have to be considered if the loci are linked.

As mentioned previously, mutations of alleles from one generation to another are not modelled in the likelihood. For mutations to be modelled, information about the pedigree structure is needed. The realised relatedness $\boldsymbol{\Delta}_R$ contains no information of the pedigree connecting the individuals. Furthermore, even though a mutation has occurred, this does not change the underlying IBD state at that particular segment of the genome. Mutation rates are typically low, and the error introduced when ignoring mutations in pedigree likelihood computations are vanishingly small [25]. However, if a mutation has occurred from a parent to a child, the likelihood (2) of $\boldsymbol{\kappa}_R = (0, 1, 0)$ would be 0, even though the true realised relatedness is $\boldsymbol{\kappa}_R = (0, 1, 0)$. Fortunately, a simple expression for the likelihood in this case exists [3]. Let the observed alleles at a single locus of the parent and child be $g_{PA} = (a/b)$ and $g_{CH} = (c/d)$, respectively, with corresponding allele frequencies p_a, p_b, p_c and p_d. Let m_{ij} denote the probability that allele i from the parent mutates to allele j observed for the child. The likelihood of $\boldsymbol{\kappa}_R = (0, 1, 0)$ that models mutations is given by the formula

$$\begin{aligned} \mathrm{L}(PO \mid g_{PA}, g_{CH}, M) = & 2^{-(I(a=b)+I(c=d))} \\ & p_a p_b((m_{ac} + m_{bc})p_d + (m_{ad} + m_{bd})p_c), \end{aligned} \tag{3}$$

where $I(i = j) = 1$ if $i = j$ and zero otherwise. The matrix M denotes the 2×2 mutation matrix with elements m_{ij}.

1.4 Likelihood ratio framework for kinship testing

The traditional approach to kinship testing in forensic genetics is the LR framework [26]. Two hypotheses are stated and the likelihoods of the hypotheses are compared. In pairwise kinship testing, the hypotheses are formulated as two relationships. Let H_P (P for prosecution) and H_D (D for defence) denote the hypotheses. The relationships stated by each hypothesis is expressed as a set of Jacquard coefficients, denoted $\boldsymbol{\Delta}_P$ and $\boldsymbol{\Delta}_D$. The LR comparing these hypotheses is expressed as

$$\mathrm{LR}(\mathrm{H}_P, \mathrm{H}_D \mid G) = \frac{L(\boldsymbol{\Delta}_P \mid G)}{L(\boldsymbol{\Delta}_D \mid G)}. \tag{4}$$

The set of hypotheses are not necessarily exhaustive. Only two relationships are compared, neither may be the true relationship between the individuals. To avoid infinite LRs, it is

required that

$$\mathrm{P}(G \mid \mathrm{H}_P) > 0 \Rightarrow \mathrm{P}(G \mid \mathrm{H}_D) > 0. \tag{5}$$

This means that all DNA profiles possible to observe under H_P are also possible to observe under H_D. This assumption is typically valid in applications relevant for this work.

1.4.1 Expectation and variance of the likelihood ratio

Slooten and Egeland [27] derived expressions for the expectation and variance of the LR, as a function of the IBD coefficients for outbred individuals. These expressions were later extended to condition on a true relationship that differs from the two hypotheses considered by the LR [28]. Paper II of this thesis extends the latter to apply for general pairwise relationships modelled by the Jacquard coefficients.

Consider the setup

$$\begin{aligned} \mathrm{H}_P: &\ \ \boldsymbol{\Delta} = \boldsymbol{\Delta}_P = (\Delta_1^P, \ldots, \Delta_9^P) \\ \mathrm{H}_D: &\ \ \boldsymbol{\Delta} = \boldsymbol{\Delta}_D = (\Delta_1^D, \ldots, \Delta_9^D) = (0, \ldots, 0, 1) \\ Truth: &\ \ \boldsymbol{\Delta} = \boldsymbol{\Delta}_T = (\Delta_1^T, \ldots, \Delta_9^T). \end{aligned} \tag{6}$$

We are concerned with the LR comparing H_P: $\boldsymbol{\Delta}_P$ against H_D: $\boldsymbol{\Delta}_D$ when the true source of the data is the relationship $\boldsymbol{\Delta}_T$. For simplicity, H_D states unrelated. The LR computed from a set of DNA profiles G is denoted $\mathrm{LR}(\boldsymbol{\Delta}_P, \boldsymbol{\Delta}_T \mid G)$.

The likelihood ratio can be considered a random variable, denoted by $\mathcal{LR}_{\boldsymbol{\Delta}_P, \boldsymbol{\Delta}_T}$ for the setup in (6). Let $\mathcal{G}$ be defined as in Section 1.3, for a single locus. The distribution of $\mathcal{G}$ is determined by $\boldsymbol{\Delta}_T$ and is independent of $\boldsymbol{\Delta}_P$ and $\boldsymbol{\Delta}_D$. The expectation of $\mathcal{LR}_{\boldsymbol{\Delta}_P, \boldsymbol{\Delta}_T}$ is

$$\mathrm{E}(\mathcal{LR}_{\boldsymbol{\Delta}_P, \boldsymbol{\Delta}_T}(\mathcal{G})) = \boldsymbol{\Delta}_P B_9 \boldsymbol{\Delta}_T^t. \tag{7}$$

where the matrix B_9 is given in Table 4. The constant L is the number of alleles at a locus and t denotes the vector transpose. Paper II shows the details of the derivation. The variance is computed through the general identity $\mathrm{Var}(\mathcal{X}) = \mathrm{E}(\mathcal{X}^2) - \mathrm{E}(\mathcal{X})^2$. The first expectation is computed as

$$\mathrm{E}(\mathrm{LR}^2) = \sum_{i=1}^{9} \Delta_i^P \boldsymbol{\Delta}_P B_i \boldsymbol{\Delta}_T^t,$$

Table 4: Elements of the symmetric matrix B_9, given as $\mathrm{E}(\mathrm{LR}(\mathrm{J}_i, \mathrm{J}_j))$. Each row represents J_i, a Jacquard state assumed by H_P, while each column presents J_j, the true Jacquard state. The number of alleles at a locus is denoted L and p denotes the population allele frequencies. The same table as in Paper II.

	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
J_1	$\sum \frac{1}{p^2}$	$\sum \frac{1}{p}$	$\sum \frac{1}{p}$	L	$\sum \frac{1}{p}$	L	$\sum \frac{1}{p}$	L	1
J_2		L^2	L	L	L	L	L	1	1
J_3			$\frac{1}{2}(L + \sum \frac{1}{p})$	L	L	1	L	$\frac{L+1}{2}$	1
J_4				L	1	1	1	1	1
J_5					$\frac{1}{2}(L + \sum \frac{1}{p})$	L	L	$\frac{L+1}{2}$	1
J_6						L	1	1	1
J_7							$\frac{L(L+1)}{2}$	$\frac{L+1}{2}$	1
J_8								$\frac{L+4}{3}$	1
J_9									1

where element (j, k) of matrix B_i is

$$\sum_G \frac{\mathrm{P}(G \mid \mathrm{J}_j)}{\mathrm{P}(G \mid \mathrm{J}_9)} \frac{\mathrm{P}(G \mid \mathrm{J}_k)}{\mathrm{P}(G \mid \mathrm{J}_9)} \mathrm{P}(G \mid \mathrm{J}_i).$$

The variance of the LR then becomes

$$\mathrm{Var}(\mathcal{LR}_{\boldsymbol{\Delta}_P, \boldsymbol{\Delta}_T}(\mathcal{G})) = \sum_{i=1}^{9} \Delta_i^P \boldsymbol{\Delta}_P B_i (\boldsymbol{\Delta}_T)^t - \left(\boldsymbol{\Delta}_P B_9 (\boldsymbol{\Delta}_T)^t\right)^2.$$

The expected LR for outbred relationships is independent of the allele frequencies at a locus. This means that regardless of what population the tested individuals belong to, the expected LR is the same as long as the length of the allele ladder, L, coincides. However, the variance of the LR depends on the allele frequencies.

The expected LR for inbred relationships is in general dependent of the allele frequencies, as seen from the terms in Table 4. The variance also depends on the allele frequencies.

Example 2: Consider a seemingly standard paternity setting. We want to evaluate the expectation and variance of the LR that compares H_P: The individuals are father and child to H_D: The individuals are unrelated. The true relationship between the individuals is parent-offspring, but the father is inbred with a coefficient of inbreeding f_T. The situation is parametrically represented by

$$\begin{aligned} \mathrm{H}_P: &\quad \boldsymbol{\Delta} = \boldsymbol{\Delta}_P = (0,0,0,0,0,0,0,1,0) \\ \mathrm{H}_D: &\quad \boldsymbol{\Delta} = \boldsymbol{\Delta}_D = (0,0,0,0,0,0,0,0,1) \\ Truth: &\quad \boldsymbol{\Delta} = \boldsymbol{\Delta}_T(f_T) = (0,0,f_T,0,0,0,0,1-f_T,0). \end{aligned} \tag{8}$$

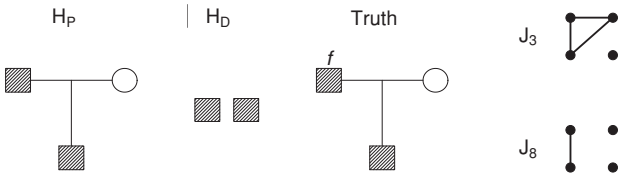


Figure 6: Hypotheses corresponding to Example 2 in Section 1.4.1 and presented in the results of Paper II. The hypothesis H_P: Parent-offspring is compared to H_D: Unrelated. The true relationship is given by the rightmost pedigree, where the father is inbred. The Jacquard states J_3 and J_8 are the only states with non-zero probability for the inbred pedigree.

Figure 6 shows the hypotheses and the true relationship in terms of pedigrees and the corresponding Jacquard states with non-zero probability. Most of the Jacquard coefficients in this setup are zero, and the expectation of the LR, computed by (7), nicely reduces to

$$E(\mathcal{LR}) = \frac{L+1}{2} f_T + \frac{L+3}{4}(1 - f_T).$$

In this special case, the expectation is independent of the allele frequencies. The expected LR for a standard paternity case, i.e. $f_T = 0$, is $(L+3)/4$. By some derivation (see Paper II for details), the variance of the LR is expressed as

$$\begin{aligned}\mathrm{Var}(\mathcal{LR}) = \quad & \frac{3L+s}{4} f_T + \big(\frac{5L+3}{8} + \frac{s-L}{16}\big)(1 - f_T) \\ & - \big(\frac{L+1}{2} f_T + \frac{L+3}{4}(1 - f_T)\big)^2,\end{aligned}$$

where $s = \sum 1/p_i$, the sum of the inverse of the allele frequencies at the locus. Figure 7 shows how the expected LR for the setup (8) increases with increasing f_T, for markers with 2, 10 and 50 alleles. The variance is shown by the grey areas in the figure, assuming uniform allele frequencies p = 1/L. If the inbreeding of the father is not accounted for in the LR, we would expect that the trust in a PO relationship increases as the true inbreeding increases.

1.4.2 The optimal LR threshold

We next discuss topics related to the interpretation of LR and the applications in paper III in the thesis. How to interpret and report evidence based on LR is, and has been for years,

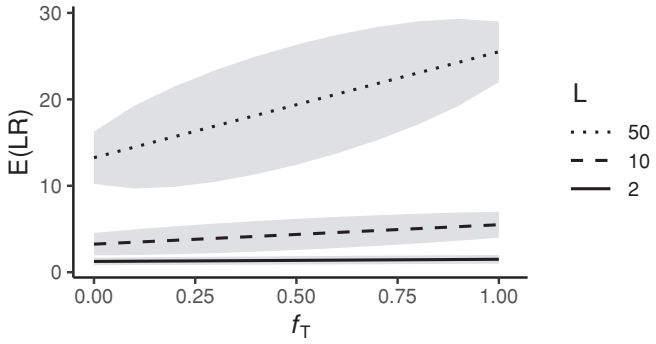


Figure 7: Expectation (lines) and variance (grey areas) of the LR for Example 2 in Section 1.4.1. Figure from Paper II. The different lines show expectation for loci with 2, 10 and 50 alleles. Uniform allele frequencies for each marker are used.

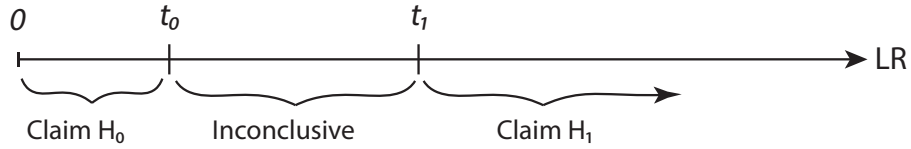


Figure 8: How a conclusion is drawn in an LR test, depending of the value of the LR, as discussed in Section 1.4.2.

a topic of discussion. The paper [29] and the references therein present different views. Verbal scales, e.g. stating that there is very strong or maybe extremely strong support of a proposition, depending on the size of the LR, are often used [30].

Figure 8 shows how a conclusion is drawn from an LR test. Let $t_0 < t_1$ denote two LR thresholds. If LR $\geq t_1$, the evidence supports H_P, i.e, we claim H_P. If LR $< t_0$ we claim H_D. For $t_0 \leq$ LR $< t_1$, the test is inconclusive. Then more data may be needed to reach a conclusion. For the work in this thesis, we assume $t_0 = t_1 = t$, so that a conclusion is always made. An important question is how to decide on an appropriate threshold.

When doing an LR test, the probability of claiming the wrong hypothesis is of concern. Let the false positive rate (FPR) and true positive rate (TPR) for a threshold t be defined as

$$\text{FPR}(t) = \text{P}(\text{LR} \geq t \mid \text{H}_D) \text{ and } \text{TPR}(t) = \text{P}(\text{LR} \geq t \mid \text{H}_P).$$

The FPR measures the probability of doing a type I error in an LR test, i.e., the probability of falsely claiming H_P. The TPR is the ability to detect the true hypothesis. Ideally, we want FPR $= 0$ and TPR $= 1$. A high threshold makes the FPR small. Unfortunately, a

high threshold makes the TPR small as well.

Kruijver describes an algorithm that computes the exact probability distribution of the LR efficiently for up to about 10 STR markers [31]. Beyond this number of markers, the distribution of the LR appears to be beyond the reach of exact calculation.

There exists an upper limit of the FPR given by the identity FPR $\leq 1/t$ [32]. In some cases, this inequality is all that is needed. However, we generally approximate FPR through simulation.

A group of methods used to decide on an optimal threshold are described next [33]. These methods are based on the receiver operating characteristic (ROC) curve of the LR. The relationship between FPR and TPR is often visualised by a ROC curve , as shown in Figure 9. The TPR is plotted as a function of the FPR, the latter often on a logarithmic scale. Each point along the ROC curve corresponds to a given LR threshold.

The Youden Index J aims to maximise the objective function

$$\mathrm{J}(t) = \mathrm{TPR}(t) - \mathrm{FPR}(t).$$

The goal is to maximize the distance between the ROC curve and the dashed diagonal line of Figure 9. Alternatively, the Concordance probability method aims to maximise

$$\mathrm{CZ}(t) = \mathrm{TPR}(t)\big(1 - \mathrm{FPR}(t)\big).$$

This means a maximisation of the area of the dotted rectangle shown in the ROC figure. The last ROC based method presented in paper III and summarised below, finds a threshold that minimises the Euclidean distance from the ROC curve to the point $(0, 1)$, shown by the long-dashed line in Figure 9. The point $(0, 1)$ can be seen as an optimal point i.e., FPR $= 0$ and TPR $= 1$. The optimal threshold is found by minimising

$$\mathrm{ER}(t) = \sqrt{\mathrm{FPR}(t)^2 + \big(1 - \mathrm{TPR}(t)\big)^2}. \tag{9}$$

The consequence of falsely claiming a relationship and not detecting the true relationship may not weigh equally, depending on the application. For instance, a false identification of a family member in a missing person case may be more severe than a false positive when removing related samples from a data base. A weight w may therefore be added as in Equation (7) of Paper III. However, deciding on an appropriate weight w may by difficult without specifying the costs associated with the two possible errors.

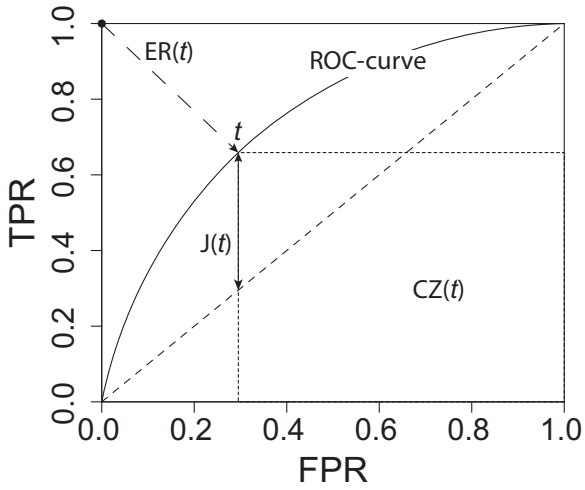


Figure 9: Figure showing a conceptual ROC curve with ROC based objective functions described in Section 1.4.2. The solid curve illustrates a ROC curve. Each point along the curve corresponds to an LR threshold with corresponding FPR and TPR. Figure adapted from [33].

1.5 The Bayesian approach to kinship testing

The likelihood presented in (2) reflects the probability of observing the data we have, conditional on a hypothesis. The LR in (4) compares the likelihood of two specific hypotheses. In a Bayesian framework, we flip the situation around and evaluate the probabilities of the hypotheses themselves, conditional on a set of observations. The general form of Bayes' theorem, involving a set of k hypotheses and the observations G, is

$$\mathrm{P}(\mathrm{H}_i \mid G) = \frac{\mathrm{P}(G \mid \mathrm{H}_i)\mathrm{P}(\mathrm{H}_i)}{\sum_{j=1}^{k} \mathrm{P}(G \mid \mathrm{H}_j)\mathrm{P}(\mathrm{H}_j)}, \text{ for } i = 1, \ldots, k. \tag{10}$$

Bayes' theorem converts the prior probabilities $\pi_i = \mathrm{P}(\mathrm{H}_i)$ for $i = 1, \ldots, k$, to posterior probabilities $\mathrm{P}(\mathrm{H}_i \mid G)$ through the evaluation of genetic data. As opposed to the likelihoods and the LR, the posterior probabilities take values between 0 and 1, making them easier to interpret. For instance, a posterior of 0.7 tells us that there is a 70% chance that H_P is true given the data and the prior. However, a prior has to be assigned to each hypothesis. If a flat prior is used, i.e., $\pi_i = 1/k$ for $i = 1, \ldots, k$, the posterior probabilities simply scale the likelihoods.

By only considering the two hypotheses, and assuming a flat prior $\mathrm{P}(\mathrm{H}_P) = \mathrm{P}(\mathrm{H}_D) =$

0.5, Bayes' theorem reduces to

$$\mathrm{P}(\mathrm{H}_P \mid G) = \frac{\mathrm{P}(G \mid \mathrm{H}_P)}{\mathrm{P}(G \mid \mathrm{H}_P) + \mathrm{P}(G \mid \mathrm{H}_D)} = \frac{\mathrm{LR}}{\mathrm{LR}+1},$$

also called the Essen-Möller index W [34]. By assuming a flat prior of 0.5, we imply that H_P and H_D are the only possible hypotheses.

An alternative form of (10) for two hypotheses is Bayes' theorem on odds form, given as

$$\frac{\mathrm{P}(\mathrm{H}_P \mid G)}{\mathrm{P}(\mathrm{H}_D \mid G)} = \frac{\mathrm{P}(G \mid \mathrm{H}_P)}{\mathrm{P}(G \mid \mathrm{H}_D)} \times \frac{\mathrm{P}(\mathrm{H}_P)}{\mathrm{P}(\mathrm{H}_D)}.$$

In words, this can be stated as

$$\text{Posterior odds} = \mathrm{LR} \times \text{prior odds}.$$

With this setup, separate priors for the hypotheses are not needed, only the ratio between them. The LR updates the prior belief in the hypotheses, and the result is the posterior odds. A posterior odds of 0.25 implies that H_D is four times more probable than H_P. A posterior odds of 2 tells that H_P is twice as probable as H_D.

The frequentist would evaluate the hypothesis through the LR, solely based on the genetic observations. A Bayesian combines the genetic and non-genetic evidence of a case. The Bayesian framework also enables more hypotheses to be compared without having to decide on a reference hypothesis. Regardless of using the Bayesian or frequentist framework, an investigator is no closer to inferring the truth if the truth is not among the considered hypotheses.

1.6 Blind search and the problem of multiple testing

A search through a DNA database can have different purposes. The aim may be to look for a direct match, i.e., to identify the individual a DNA sample originates from. Alternatively, we may want to do a familial search, i.e., search for relatives of an individual [35]. In both search applications, a single DNA profile is compared to all profiles in a database.

The purpose of a blind search is to detect pairwise relationships between DNA samples in a database. Pairwise comparison between all DNA samples are performed, computing an LR for each comparison. This can be an initial step in a DVI case [36, 37].

Let n denote the number of DNA samples in a database. A blind search among these samples results in $N = n(n-1)/2$ pairwise comparisons. We want to do a hypothesis test comparing hypothesis H_1 to H_0. This means that we perform N LR tests, concluding with

	Claim H_0	Claim H_1	Total
H_0 true	TN	FP	N_0
H_1 true	FN	TP	N_1
Total	W_0	W_1	N

Table 5: Summary of a multiple testing scenario, as described in Section 1.6. Adapted from [38]. Only W_0, W_1 and N are observed in a real setting when the truth for each test is unknown.

H_1 if the LR is a above a given threshold and H_0 if the LR is below the threshold. The outcome of a multiple testing scenario can be summarised as in Table 5. In a real case we only observe the number of LRs below and above a given LR threshold, denoted W_0 and W_1, respectively. The number of type I errors are denoted FP. The number of tests where the true relationship H_1 is detected is denoted TP. The number of times H_0 is correctly and wrongly claimed are denoted TN and FN. The number of pairs related according to the two hypotheses are denoted N_0 and N_1. We want FP and FN to be as close as possible to 0, while TN and TP stay as close as possible to 1. The possibly large number of comparisons in a blind search, may cause the probability of type I errors to be large even if the FPR for a single LR test is very small.

The Family Wise Error Rate (FWER) controls the amount of type I errors in a multiple testing scenario [38]. The FWER, denoted by α, is defined as the probability of getting at least one false positive out of N test. For N independent tests,

$$\alpha = \mathrm{P}(FP \geq 1) = 1 - (1 - FPR)^N,$$

where FPR, the false positive rate for a single LR test, is assumed to be the same for each test.

Paper III shows that the LRs in a blind search are not necessarily independent. Every DNA profile is part of several comparisons, causing the LRs to be correlated. As an example, consider two siblings A and B and an individual C unrelated to the siblings. We want to compute the LR comparing H_1: Siblings, against H_0: Unrelated. The genomes of A and B have a degree of similarity. If the LR comparing A and C is low, it is also likely that the LR comparing B and C is low.

An upper limit of the FWER for N potentially dependent tests is given by the Bonferroni bound

$$\alpha \leq N \cdot \mathrm{FPR} = \frac{n(n-1)}{2}\mathrm{FPR}.$$

An upper limit of the FPR, requiring α to stay below a given value is then computed

by

$$\text{FPR}_\alpha \leq \frac{2\alpha}{n(n-1)}.$$

An optimal threshold for the blind search is then found by, for instance, minimising (9), with the constraint $\text{FPR} \leq \text{FPR}_\alpha$.

An alternative approach to control the type I errors in a multiple testing setting is through the False Discovery Rate (FDR) [39]. The FDR is defined as the expectation of the ratio between the number of false positives and the number of tests where H_1 is claimed, i.e.,

$$\text{FDR} = \text{E}\left(\frac{FP}{W_1}\right).$$

Benjamini and Hochberg [39] derive decision rules to control the FDR, based on p-values for the test. Kruijver et al. argue against applying p-values in LR testing in forensic genetics [40].

1.7 Maximum likelihood estimation of relatedness

In the kinship testing framework presented so far, a limited set of hypotheses are evaluated and compared. Instead of a set of hypotheses, we can evaluate the likelihood function (2) for all possible values of $\boldsymbol{\Delta}_R$ and choose the most likely set of coefficients to describe the observations. This is ML estimation of $\boldsymbol{\Delta}_R$ and amounts to finding the estimate $\widehat{\boldsymbol{\Delta}}_R$ that maximises (2) in the parameter space

$$\Omega = \{(\Delta_1^R, \ldots, \Delta_9^R) \mid \sum_{i=1}^{9} \Delta_i^R = 1, 0 \leq \Delta_1^R, \ldots, \Delta_9^R \leq 1\} \in \mathbb{R}^9.$$

Thompson [4] describes ML estimation of the IBD coefficients. Milligan [5] compares different methods of estimation (e.g. likelihood estimators and method-of-moments estimator), by reduction of the IBD coefficients to the kinship coefficient. His work shows that the ML estimator is more biased than the other estimators, especially for boundary points, but has a smaller variance. Anderson and Weir [6] propose an approach for ML estimation in subdivided populations. The mentioned papers do not distinguish between the realized and the pedigree relationship. The estimate of $\boldsymbol{\kappa}$ is constrained to the admissible region of the IBD triangle. This affects both the bias and variance estimates.

1.7.1 Asymptotic properties of the ML estimator

The ML estimator has many desirable asymptotic properties, given that a set of regularity conditions are fulfilled [41]. Let $\mathcal{X}_i$ for $i = 1, \ldots, n$, denote n independent identically distributed (i.i.d.) random variables with probability mass function $f(x, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a p-dimensional parameter vector. Three of the regularity conditions are:

1. The parameter set defines the probability mass function, i.e., $f(x, \boldsymbol{\theta}) \neq f(x, \boldsymbol{\theta}')$ whenever $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$.

2. The sample space of $\mathcal{X}$ does not depend on $\boldsymbol{\theta}$.

3. The true value of $\boldsymbol{\theta}$ is an interior point in the parameter space Ω.

The full set of regularity conditions are given in [41]. If all regularity conditions hold, then $\widehat{\boldsymbol{\theta}}$ is a unique solution of $\frac{\delta}{\delta \boldsymbol{\theta}} l(\boldsymbol{\theta}) = 0$, and $\widehat{\boldsymbol{\theta}}$ converges in probability to the true parameter value $\boldsymbol{\theta}$ as n increases. The function $l(\boldsymbol{\theta})$ denotes the logarithm of the likelihood function. If this holds, then

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) \xrightarrow{D} N_p\big(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta})\big),$$

where $\mathbf{I}$, the Fisher information, is the $p \times p$ covariance matrix of the gradient of $\log f(\mathcal{X}, \boldsymbol{\theta})$. The elements of $\mathbf{I}$ are given as

$$I_{jk} = \operatorname{cov}\big(\frac{\delta}{\delta \theta_j} \log f(\mathcal{X}, \boldsymbol{\theta}), \frac{\delta}{\delta \theta_k} \log f(\mathcal{X}, \boldsymbol{\theta})\big),$$

for $j, k = 1, \ldots, p$. Furthermore, a lower bound on the variance of unbiased estimators of θ_k is given by

$$\operatorname{Var}\big(\widehat{\theta}_k\big) \geq \frac{\big(\mathbf{J}^{-1}(\widehat{\boldsymbol{\theta}})\big)_{kk}}{n},$$

where $\mathbf{J}^{-1}(\widehat{\boldsymbol{\theta}})$ is the observed Fisher information.

Consider now the application in this thesis. Observations used in the ML estimation are not i.i.d., the possible joint genotypes at the different loci have distinct probability distributions. Thompson [4] stated that under certain conditions "$\mathrm{P}(G \mid \boldsymbol{\kappa}) = \mathrm{P}(G \mid \boldsymbol{\kappa}')$ if and only if $\boldsymbol{\kappa} = \boldsymbol{\kappa}'$" for a single locus. The regularity condition 1 thus holds under certain conditions. The regularity condition 3 does not hold. The Jacquard coefficients for the most common pedigree relationships are located on the boundary of the parameter space. Thus, $\boldsymbol{\Delta}_R$ is likely to be a boundary point as well. Regularity condition 2 holds if condition 3 holds (which it does not). If $\kappa_0^R = 0$, meaning that the parameter is a boundary point, then some joint genotypes are impossible to observe, and the sample space of $\mathcal{G}$ changes

with the value of $\boldsymbol{\Delta}_R$ or $\boldsymbol{\kappa}$. The asymptotic properties of the ML estimator therefore do not apply. It is also worth to notice that the number of markers analysed is typically low and the statistical properties of the ML estimator may therefore differ from the asymptotic ones, even though the regularity conditions were to be fulfilled. Alternative methods for assessing the uncertainty of the parameter estimate thus have to be applied as discussed next.

1.8 Bootstrapping

Bootstrapping is used to assess the uncertainty of a parameter estimate when explicit formulas are not available [42]. The observations of the data set X used in the estimation is governed by a probability distribution F. Associated with F is a parameter of interest θ. This parameter is estimated by $\widehat{\theta} = T(X)$. Bootstrapping mimics the stochastic process that created the original data, through the approximation $\widehat{F}$. Bootstrap data sets $X_1^*, \ldots, X_B^*$ are sampled from $\widehat{F}$. A bootstrap estimate $\widehat{\theta}^* = T(X^*)$ is computed for each bootstrap data set. For a consistent bootstrap method, the sample variance of the bootstrap estimates $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ approximates the variance of the original parameter estimate $\widehat{\theta}$. Increasing the observations in the data set X, decreases the variance of $\widehat{\theta}$. Increasing B makes the sample variance of the bootstrap estimates closer to the variance of $\widehat{\theta}$. However, how well $\widehat{F}$ approximates F affects how well $\widehat{\theta}$ can be approximated.

The observations in our applications are two DNA profiles genotyped for M loci. The parameter we want to estimate is the nine-dimensional coefficient $\boldsymbol{\Delta}_R$ or the three-dimensional $\boldsymbol{\kappa}_R$, possibly reduced to the one-dimensional φ_R. Paper IV discusses parametric and non-parametric bootstrap in light of the applications in this thesis. The bootstrapping procedures differ in how bootstrap data sets are created, i.e., how they approximate F.

1.8.1 Non-parametric bootstrap

Bootstrap data sets are created by sampling observations with replacement from the original data set. For our applications, this would be to randomly draw M loci from the original DNA profiles, and do ML estimation on this new set of genotypes. The stochastic process of selection of loci is reflected in the sample variance of the non-parametric bootstrap estimates.

Each locus can appear several times in a bootstrap data set and different sets of loci are used in each bootstrap estimate. The non-parametric way of resampling bootstrap data sets may thus appear unnatural, particularly since markers typically are far from i.i.d.

1.8.2 Parametric bootstrap

The loci are considered fixed when performing parametric bootstrap. A bootstrap data set is created by sampling genotypes for each locus independently, conditioned on the estimate $\widehat{\mathbf{\Delta}}_R$, i.e., sampling genotypes $\mathcal{G}$ from the distribution

$$\mathrm{P}((\mathcal{G}) \mid \widehat{\mathbf{\Delta}_R}) = \sum_{i=1}^{9} \widehat{\Delta}_i \mathrm{P}(\mathcal{G} \mid \mathrm{J}_i). \tag{11}$$

Imagine a big set of all possible joint genotypes for a locus, with the value of $\widehat{\mathbf{\Delta}}_R$ as a parameter for this set. By performing parametric bootstrap, we draw at random joint genotypes from this set, with probabilities given by (11). The parametric bootstrap data sets may well joint genotypes that are not present in the original DNA profiles.

There may be problems regarding the use of non-parametric bootstrap, as argued in the last paragraph of Section 1.8.1. Parametric bootstrap seems to be the better bootstrapping procedure in this application. However, simulations indicate that the difference between the results of the two procedures is small. Non-parametric bootstrap should therefore not be discarded as useful. Also, it is simple to implement.

1.8.3 Bootstrap confidence intervals and regions

A confidence interval or region quantifies the uncertainty of a parameter estimate. Often, construction of confidence intervals are based on pivotal statistics, i.e., statistics with a known probability distribution, where the distribution is independent on the parameter of interest. As a general example, consider the random variable $\mathcal{X}$, with unknown expectation μ and variance σ. We have a set of n independent observations $X = (x_1, \ldots, x_n)$ of $\mathcal{X}$. Let $\overline{X}$ and s denote the sample mean and standard deviation, respectively. The random variable

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

is then student-t distributed with $n-1$ degrees of freedom. Let $t_{1-\alpha/2}$ denote the $1-\alpha/2$ quantile of this distribution, so that $P(-t_{1-\alpha/2} < t < t_{1-\alpha/2}) = 1-\alpha$. A $(1-\alpha)100\%$ confidence interval for μ is then $(\overline{X} - t_{1-\alpha/2} \cdot s/\sqrt{n}, \overline{X} + t_{1-\alpha/2} \cdot s/\sqrt{n})$.

Consider next that we perform bootstrapping on the data set X, to assess the uncertainty of the estimate $\widehat{\mu} = \overline{x}$. If the bootstrap procedure properly approximates the distribution of the data, then $\mathrm{E}(\widehat{\mu}^*) = \widehat{\mu}$ and $\mathrm{Var}(\widehat{\mu}^*) = \mathrm{Var}(\widehat{\mu})$, where $\widehat{\mu}^*$ denotes a bootstrap estimate. Because the expectation of the bootstrap estimate differs from μ, the confidence interval above can not be applied directly.

Methods for constructing univariate bootstrap confidence intervals are thoroughly reviewed in the literature [42]. The percentile method constructs a confidence interval from the percentiles of the empirical distribution of the bootstrap estimates, i.e., the interval between the $\alpha/2$ and $1-\alpha/2$ percentile of the ordered set of bootstrap estimates. The percentile-t method uses an pivotal statistic approximated from the bootstrap estimates, similar to the t-statistic above. The bias-corrected and accelerated confidence interval adjust for bias and skewness in the bootstrap estimates [42]. Anderson and Weir [6] assess the uncertainty of the estimate of the kinship coefficient through non-parametric bootstrap and construct a confidence interval by the percentile method.

Construction and visualisation of confidence regions for multivariate parameters are more difficult. Different approaches are described in the literature [43, 44]. For instance, a likelihood-based 95% confidence region is a region where all parameter values inside the regions are more likely than the parameters outside the region and covers the true parameter in 95% of the times it is constructed.

1.9 Implementation

The computations in this thesis are performed using the programming language R. The implementation builds on the *ped suite* [45], a collection of R packages for pedigree analysis, freely available from CRAN. These R packages contain functionality for construction, handling and plotting of pedigrees and marker data, simulation of independent markers, computation of pedigree likelihoods, estimation of relatedness parameters and bootstrapping.

The implementations for Paper II and Paper III are collected in two R packages available from the author. A parametric version of the likelihood function is implemented by the author, facilitating computation of likelihood for all possible values of relatedness parameters. The simulation of linked markers in Paper I is performed using the software Merlin [46].

2 Paper summaries

The work in this thesis consists of four papers. Paper I motivates and exemplify how ignoring inbreeding can lead to false conclusions in kinship cases. Paper II elaborates on statistical properties of the LR between pairs of individuals, and extends results from the literature to also apply for inbred relationships. The use of LRs in kinship testing is the topic of Paper III. Strategies for determining optimal LR thresholds for the multiple testing scenario called blind search are discussed, and the use of a Bayesian framework for kinship testing is shown. In Paper IV, attention is moved from the LR framework to estimation of IBD parameters among all possible alternatives. The uncertainty of the parameter estimate is investigated through bootstrapping.

2.1 Paper I

In this paper, the two frameworks for kinship inference are studied; a parametric approach and a LR approach. The aim is to investigate through simulations how ignoring inbreeding between individuals can lead to false conclusions in kinship cases.

For the parametric approach, an inbred grandparent-grandchild relationship and an inbred first cousin relationship are considered. Marker data is repeatedly simulated on the individuals of interest and the IBD coefficients between the individuals are estimated. When ignoring the inbreeding, the inbred grandparent-grandchild relationship may be mistaken as a sibling relationship. The first cousin relationship may be difficult to distinguish from half siblings, grandparent-grandchild or avuncular.

Furthermore, we analyse the LR comparing a sibling relationship to an outbred grandparent-grandchild relationship. The true relationship between the individuals is an inbred grandparent-grandchild relationship. The results show that ignoring inbreeding can lead to the false inference of a sibling relationship instead of grandparent-grandchild.

2.2 Paper II

Motivated by the preliminary simulation study of paper I, this second paper studies statistical properties of the LR for inbred relationships. Explicit formulas for the expectation and variance of the LR comparing two outbred relationships, conditioned on a third true relationship between the individuals, are previously derived in the literature. We extend these derivations to apply for pairwise relationships in general, not only outbred.

The expectation of the LR for outbred individuals are independent of population allele frequencies. This is not necessarily the case for inbred relationship, according to the for-

mulas we derive. The variance of the LR depends on the allele frequencies, regardless of the presence of inbreeding or not. Low frequent alleles at a marker increase the variance of the LR.

The effect of this increased variance becomes evident in a simulation study performed in the paper. Repeated simulations of genetic data on a half sibling relationship with founder inbreeding is performed. The LR comparing a full sibling relationship with founder inbreeding to unrelated is computed for each simulation. The mean of the LRs deviates from the expected value. This deviation is less for markers with uniform allele frequencies than for markers with varying allele frequencies. These results show how exact expressions can give insight to properties of the LR that simulations do not.

2.3 Paper III

Paper III evaluates the LR framework in a multiple testing setup. A blind search amounts to pairwise kinship testing among all combinations of a set of DNA samples. An LR is computed for each comparison. By a parametric implementation of the likelihood ratio, the hypotheses tested can be any relationship, outbred or inbred. Furthermore, we do not restrict our attention to autosomal markers, but also exemplify how X-chromosomal markers can be useful in specific kinship cases.

Even though the probability of doing a type I error in a single LR test is vanishingly small, the possibly large amount of comparisons in a blind search makes the overall probability of doing type I errors substantial. This motivates finding optimal LR thresholds in a blind search setting.

We apply the Family Wise Error Rate to control the probability of getting at least one false positive in the blind search. We show that the LRs in the blind search are not necessarily independent of each other. Each individual are present in several of the pairwise comparisons, possibly making the LRs dependent. Thus, an upper limit for the FWER is given by a Bonferroni bound. This upper bound for the Family Wise Error Rate sets an upper limit for the false positive rate (FPR) of a single LR computation in the blind search. An optimal LR threshold is then determined through estimation of error rates from simulated data.

The paper also applies a Bayesian framework to pairwise kinship testing. The Bayesian framework enables us to compare more hypotheses than the two tested in the LR. Furthermore, we show how X-chromosomal markers can distinguish maternal and paternal half siblings.

2.4 Paper IV

This last paper shifts focus from the LR framework to estimation of the realised pairwise relationship among all possible alternatives. The Jacquard coefficients, or the IBD coefficients for outbred relationships, can be estimated through a maximisation of the likelihood function. We distinguish between the pedigree relationship and the realised relatedness between individuals. The latter is estimated through the maximum likelihood estimation of the Jacquard coefficients.

The uncertainty in the estimate needs to be addressed. The asymptotic properties of a ML estimator can not be assumed to apply for the estimate of the realised relatedness since most of the common pedigree relationships are located on the boundary of the parameter space, making it likely that the realised relatedness is located on the boundary as well. Bootstrapping is instead applied to approximate the variance of the estimate and construct confidence regions for the coefficients.

The paper contains a fundamental discussion of how to define the data and of the probability distribution of the data used in the ML estimation. This discussion leads to a review of the parametric and non-parametric bootstrap. We argue that parametric bootstrap is more reasonable to use because it better mimics the stochastic process that generated the data. However, the two methods seem to give similar results, especially for interior points of the parameter space.

3 Discussion

This thesis focuses on statistical methods for inference of relationship between pairs of individuals. We investigate relationships in general, not limiting attention to outbred individuals. Paper I contains a preliminary study of the effect of ignoring inbreeding in pairwise kinship evaluations, motivating the work that follows in this thesis. The two main approaches to kinship inference in forensic applications are based on the LR framework and estimation of IBD parameters respectively. Paper II and III focus on the LR framework while Paper IV addresses estimation.

Slooten and Egeland viewed the LR as a random variable [27, 28]. They derived formulas for the expectation and variance of the LR, taking into account that the true relationship between the individuals may differ from the tested hypotheses. The work in Paper II extends these expressions to also apply for inbred individuals. For outbred relationships, the expectation of the LR is independent of the allele frequencies. We show that this is not generally the case for inbred relationships. The variance depends on the allele frequencies in both cases, and becomes high for loci with rare alleles. The effect of this higher variance is seen in simulated data. The mean LR over many simulations is considerably lower than the expected value. This difference becomes smaller when allele frequencies are more evenly distributed. However, for a locus with many alleles, the allele frequencies become small even though they are evenly distributed. The expected LR increases as a function of the number of alleles at a locus. The difference between the expected and observed LR is therefore evident for loci with many alleles, even though the alleles have equal frequencies.

The relevance of exact results when simulation easily provides results, can always be questioned. Nothnagel et al. evaluate the discriminatory power of the likelihood ratio by simulations and exact expressions [47]. Typically, exact expressions require more restricted assumptions. So, what's the point of putting great effort into deriving exact results? A general answer is that the exact counterparts may provide understanding not easily grasped from simulations. The effect of parameters on results are more easily seen from a formula than from simulation. In the problems addressed in paper II, the relevance of exact expressions are justified since a large number of Monte Carlo simulations may be needed to provide exact results. In fact, the required number of simulations may be prohibitively large. This is most easily seen from an example. The result $\mathrm{E}(\mathrm{LR} \mid \mathrm{H}_D) = 1$ holds under the reasonable regularity condition (5). However, for a paternity case it is extremely unlikely that unrelated individuals share an allele for all markers. Let p be the probability that two unrelated people fit as parent-offspring. The expected number of simulations needed to get a non-zero value of the LR is $1/p$ and one could easily have $1/p = 10^{12}$. Direct Monte Carlo

simulation will therefore not work, importance sampling as described in Kruijver [31] may work. For this reason, the results in Paper II gives important knowledge of the nature of the LR, that simulations are not able to give.

Paper III investigates the use of LRs in a multiple testing scenario. The goal is to establish a framework for deciding on appropriate LR thresholds. A blind search compares all DNA samples in a data base in a pairwise manner. The possibly large number of pairwise comparisons requires a framework for controlling the overall number of type I errors.

As exemplified in Paper III, the LRs in a blind search are not necessarily independent of each other. An upper limit for the FWER is therefore given by Bonferroni bound. This bound is valid for dependent tests and is a function of the FPR of a single LR test. The optimal LR threshold is derived through ROC-based methods. These methods exploit the relation between the FPR and TPR. As discussed in Paper III, the consequence of doing a false positive and a true positive may not weigh equally, depending on the application. For instance, if the objective is to remove related individuals from a database, doing a type I error is not problematic. Falsely identifying an individual in a missing person case, however, is not wanted. The work in Paper III shows the importance of evaluating the overall error rate of the blind search when deciding on appropriate LR thresholds.

The Bayesian approach to kinship testing is exemplified and discussed in Paper III. The focus is on a single pair of individuals. Posterior probabilities for several hypotheses are investigated, not only comparing two hypotheses as in the LR framework. The posteriors are on a scale between zero and one, making them easy to interpret. However, a prior probability, preferably objective, needs to be specified for each hypothesis. A flat prior simply scales the likelihood of each hypothesis, adding no extra information. Other informative priors can be specified, adding non-genetic information.

The discussion relating to Bayesian and frequentist approaches in forensic applications mirrors those in other areas. For instance, some labs report the posterior probability, or Essen-Moller's W [48], while others, probably most, give the LR as recommended by Gjertson et al. [26]. The arguments favouring a Bayesian approach, or a Bayesian supplement to only reporting LRs, may be stronger for the applications of Paper III than for conventional kinship cases. As pointed out above, posteriors are easier to interpret when many tests are performed. Also, the blind search is frequently explorative. Often further analyses are performed before conclusions are drawn and these final conclusions need not rely on subjective priors.

Some kinships are not possible to distinguish from autosomal markers alone. For instance, the likelihood of two sisters being maternal half siblings and paternal half siblings are identical using autosomal markers. However, these relationships have different IBD

coefficients for X-chromosomal markers, and hence, their likelihood differs. This is exemplified in Paper III, investigating the posterior probabilities of the relationships. It should be noted that markers on the X-chromosome are not independent, violating the requirements of both the LR and the Bayesian framework. A combination of different marker types can therefore give valuable information in specific kinship cases.

The blind search procedure discussed in Paper III is widely used. The implementations we are aware of, like Familias, https://familias.no, are limited to some few selected relationships and independent autosomal markers. A useful continuation of Paper III would be to provide a software implementation for general pairwise relationships. Also, including X-chromosomal markers would be helpful. The X-chromosomal example in Paper III is a proof of principle as independence markers and linkage equilibrium are assumed. A main challenge for an implementation allowing for dependent markers relates to computational speed. As pointed out, the number of comparisons is typically large, and an efficient implementation would be required. Further work is also needed to test the procedure described to derive optimal LR thresholds. Testing on simulated data is easy, but not sufficient. Real test data is needed. A challenge is then that the true relatedness is not known. However, extended sets of markers, beyond the ones typically used in forensics could be used to get closer to the true relationship.

Estimation of the Jacquard coefficients makes us able to investigate all possible pairwise relationships [4, 5]. Paper IV explores the uncertainty of the estimate through parametric and non-parametric bootstrap. As far as we know, parametric bootstrap has not previously been used in applications similar to those we address. In our view, the parametric bootstrap appears more reasonable than the non-parametric version, as this way of simulating data mimics the process that generated the data better. However, in many cases the two versions of the bootstrap produce similar results as exemplified in Paper IV. A problem is that we do not know the true distribution of the parameter estimate and we can therefore not conclude that one of the bootstrap procedures is more accurate than the other. The coverage probabilities presented in Paper IV depends on the type of confidence interval/region used. The bootstrapping procedures may perform differently for other types of intervals/regions. Paper IV gives an important discussion of the use of bootstrap methods in forensic kinship analysis.

Finally, we mention assumptions that are not specific to the applications of this thesis. Independence between markers is required in both the LR and estimation framework. Multilocus IBD coefficients have to be used when markers are linked [23, 22]. Standard forensic STR markers are close to independent. However, a higher number of markers may be needed to reduce the uncertainty of the ML estimate to an acceptable level. The markers

then become linked and linkage equilibrium could also be violated. The effect of violation of HWE is often seen in the genotype frequencies of a population. Deviation from HWE can in the likelihood equation (2) be accounted for by changing the joint genotype frequencies in Table 2. This is also discussed in the section on founder inbreeding and θ-correction in Paper IV. Anderson and Weir [6] propose an alternative to Table 2 that models structured populations.

We have mentioned some topics for future work, particularly in connection with Paper III. Regarding the other papers of the thesis, we consider Paper IV to be the one that would be most interesting to expand on. There are several hard problems that need to be addressed. For instance, a Bayesian approach would be interesting to explore. Also, estimating parameters and construction of confidence intervals, when the parameter is at the border of the parameter space remains a challenge, not only for the forensic applications of this thesis.

References

[1] S. Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56:330–338, 1922.

[2] W. G. Hill and B. S. Weir. Variation in actual relationship as a consequence of mendelian sampling and linkage. *Genetics Research*, 93(1):47–64, 2011.

[3] T. Egeland, D. Kling, and P. Mostad. *Relationship inference with Familias and R: statistical methods in forensic genetics*. Academic Press, 2015.

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

[5] B. G. Milligan. Maximum-likelihood estimation of relatedness. *Genetics*, 163(3):1153–1167, 2003.

[6] A. D. Anderson and B. S. Weir. A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, 176(1):421–440, 2007.

[7] D. Kling, T. Egeland, A. Tillmar, and L. Prieto. *Mass identifications. Statistical methods in forensic genetics*. Elsevier Academic press, 2021.

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

[9] J. M. Butler. *Advanced Topics in Forensic DNA Typing: Methodology: Methodology*. Academic Press, 2011.

[10] P. Gill, Ø. Bleka, O. Hansson, C. Benschop, and H Haned. *Forensic Practitioner's Guide to the Interpretation of Complex DNA Profiles*. Elsevier Academic Press, 2020.

[11] J. Buckleton, C. M. Triggs, and Walsh S. J. *Forensic DNA evidence interpretation*. CRC press, 2005.

[12] M. Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

[13] A. Ziegler and I. R. König. *A statistical approach to genetic epidemiology: With access to e-learning platform by Friedrich Pahlke*. Wiley Online Library, 2010.

[14] I. Simonsson and P. Mostad. Stationary mutation models. *Forensic Science International: Genetics*, 23:217–225, 2016.

[15] B. L. Browning and S. R. Browning. A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics*, 88(2):173–182, 2011.

[16] A. Jacquard. Genetic information given by a relative. *Biometrics*, pages 1101–1114, 1972.

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

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

[19] M. D. Vigeland. Relatedness coefficients in pedigrees with inbred founders. *Journal of Mathematical Biology*, 81:185–207, 2020.

[20] M. D. Vigeland and T. Egeland. Handling founder inbreeding in forensic kinship analysis. *Forensic Science International: Genetics Supplement Series*, 7(1):780–781, 2019.

[21] R. C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Human Heredity*, 21(6):523–542, 1971.

[22] E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84(8):2363–2367, 1987.

[23] E. A. Thompson. Two-locus and three-locus gene identity by descent in pedigrees. *Mathematical Medicine and Biology: A Journal of the IMA*, 5(4):261–279, 1988.

[24] K. Lange and J. S. Sinsheimer. Calculation of genetic identity coefficients. *Annals of Human Genetics*, 56(4):339–346, 1992.

[25] T. Egeland, N. Pinto, and A. Amorim. Exact likelihood ratio calculations for pairwise cases. *Forensic Science International: Genetics*, 29:218–224, 2017.

[26] 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(3-4):223–231, 2007.

[27] K. Slooten and T. Egeland. Exclusion probabilities and likelihood ratios with applications to kinship problems. *International Journal of Legal Medicine*, 128(3):415–425, 2014.

[28] T. Egeland and K. Slooten. The likelihood ratio as a random variable for linked markers in kinship analysis. *International Journal of Legal medicine*, 130(6):1445–1456, 2016.

[29] C. E. H. Berger and K. Slooten. The LR does not exist. *Science & Justice*, 56(5):388–391, 2016.

[30] J. S. Buckleton, J. Bright, and D. Taylor. *Forensic DNA evidence interpretation*. CRC press, 2018.

[31] M. Kruijver. Efficient computations with the likelihood ratio distribution. *Forensic Science International: Genetics*, 14:116–124, 2015.

[32] K. Slooten and T. Egeland. Exclusion probabilities and likelihood ratios with applications to mixtures. *International Journal of Legal Medicine*, 130(1):39–57, 2016.

[33] M. Rota and L. Antolini. Finding the optimal cut-point for gaussian and gamma distributed biomarkers. *Computational Statistics & Data Analysis*, 69:1–14, 2014.

[34] E. Essen-Möller. Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen. *Mitteilungen d. Anthrop. Ges. in Wien*, 68:2–53, 1938.

[35] M. Kruijver, R. Meester, and K. Slooten. Optimal strategies for familial searching. *Forensic Science International: Genetics*, 13:90–103, 2014.

[36] B. Bertoglio, P. Grignani, P. Di Simone, N. Polizzi, D. De Angelis, C. Cattaneo, A. Iadicicco, P. Fattorini, S. Presciuttini, and C. Previderè. Disaster victim identification by kinship analysis: the Lampedusa October 3rd, 2013 shipwreck. *Forensic Science International: Genetics*, 44:102–156, 2020.

[37] L. Olivieri, D. Mazzarelli, B. Bertoglio, D. De Angelis, C. Previderè, P. Grignani, A. Cappella, S. Presciuttini, C. Bertuglia, P. Di Simone, and other. Challenges in the identification of dead migrants in the Mediterranean: the case study of the Lampedusa shipwreck of October 3rd 2013. *Forensic Science International*, 285:121–128, 2018.

[38] A. C. Tamhane, Y. Hochberg, and C. W. Dunnett. Multiple test procedures for dose finding. *Biometrics*, pages 21–37, 1996.

[39] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: Series B (Methodological)*, 57(1):289–300, 1995.

[40] 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.

[41] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[42] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[43] P. Hall. On the bootstrap and likelihood-based confidence regions. *Biometrika*, 74(3):481–493, 1987.

[44] S. Ghosh and A. M. Polansky. Smoothed and iterated bootstrap confidence regions for parameter vectors. *Journal of Multivariate Analysis*, 132:171 – 182, 2014.

[45] M. D. Vigeland. *Pedigree analysis in R*. Academic Press, 2021.

[46] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1):97–101, 2002.

[47] 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.

[48] T. Egeland, B. Kulle, and R. Andreassen. Essen-Möller and identification based on DNA. *Chance*, 19(2):27–31, 2006.

Paper I

Forensic Science International: Genetics Supplement Series 7 (2019) 462–464

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

The impact of ignoring inbreeding in pairwise kinship evaluations

Hilde Kjelgaard Brustad*, Thore Egeland

Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Chr. M. Falsens vei 1, 1433 Aas, Norway

ARTICLE INFO

Keywords:
Relatedness
Inbreeding
IBD
Likelihood ratio

ABSTRACT

Inbreeding is often ignored in forensic kinship inference, either because the exact genealogies are unknown, or because current methods are inadequate for inbred pedigrees. In many situations, this may be a reasonable simplification, but in other, it can lead to wrong inference of relatedness between two individuals. In order to quantify the impact of ignoring inbreeding, we simulate marker data for a selection of inbred pedigrees, and subject them to pairwise kinship inference using standard noninbred assumptions.

1. Introduction

Pairwise kinship evaluations are of great importance in forensics. Usually, the individuals of interest are assumed to be noninbred and pedigree founders are assumed to be unrelated and noninbred. In many cases, these are reasonable assumptions. But what happens when the assumptions are not valid?

We will look at two approaches to pairwise kinship inference, a parametric and a likelihood ratio (LR) framework. Through simulations of genetic data from inbred individuals, we apply these approaches to kinship inference when ignoring inbreeding, and show how this affects the results.

2. Method

2.1. Parametric approach to kinship inference

The parametric approach to kinship evaluations aims to estimate values of relatedness parameters from genetic data. The relatedness between individuals can be measured by the identity-by-descent (IBD). A coarse measure of relatedness, based on the concept of IBD is the kinship coefficient φ [1]. A more refined measure of relatedness in terms of IBD status is the IBD coefficients $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$ [2]. The IBD coefficients are only defined for pairwise noninbred individuals. The Jacquard coefficients $\Delta = (\Delta_1, \Delta_2, \ldots, \Delta_9)$ [3] extend the IBD coefficients to pairwise inbred individuals. One way to study the impact of ignoring inbreeding is to compare estimates of the kinship coefficients based on IBD or Jacquard coefficients. Only the latter appropriately accommodates inbreeding.

2.2. LR approach to kinship inference

The LR describes how much more likely genetic data is explained by one hypothesis, H_P (stated as a pedigree), compared to another hypothesis, H_D. Problems arise when the true relationship between the individuals of interest is not as stated by neither of the hypotheses.

2.3. Implementation

Simulation of STR markers, estimation of IBD coefficients, and LR computations, were performed with the R package forrel, which is part of the pedtools suite of packages of pedigree analysis. (https://github.com/magnusdv/forrel). Simulations of linked markers has been performed using the software Merlin [4]. Mutations, dropouts and silent alleles are ignored for both linked and unlinked simulations.

3. Results

3.1. Parametric approach

Consider the pedigrees in Fig. 1. Twenty simulations of a set of 10,000 SNP markers, with uniform allele frequencies, for the individuals of interest (colored) are performed for each of the two pedigrees. Marker data are simulated as linked markers, with recombination rate $\theta = 0.01$. Independence between markers is assumed when performing maximum likelihood estimation of IBD and Jacquard coefficients.

The estimation does not consider any knowledge of the pedigree connecting the individuals. Fig. 1c shows estimation of $\boldsymbol{\kappa}$ from simulated data for the two different pedigrees. The cluster of blue points corresponds to estimates with data from the inbred

* Corresponding author.
E-mail address: hilde.brustad@nmbu.no (H.K. Brustad).

https://doi.org/10.1016/j.fsigss.2019.10.052
Received 16 September 2019; Accepted 7 October 2019
Available online 31 October 2019

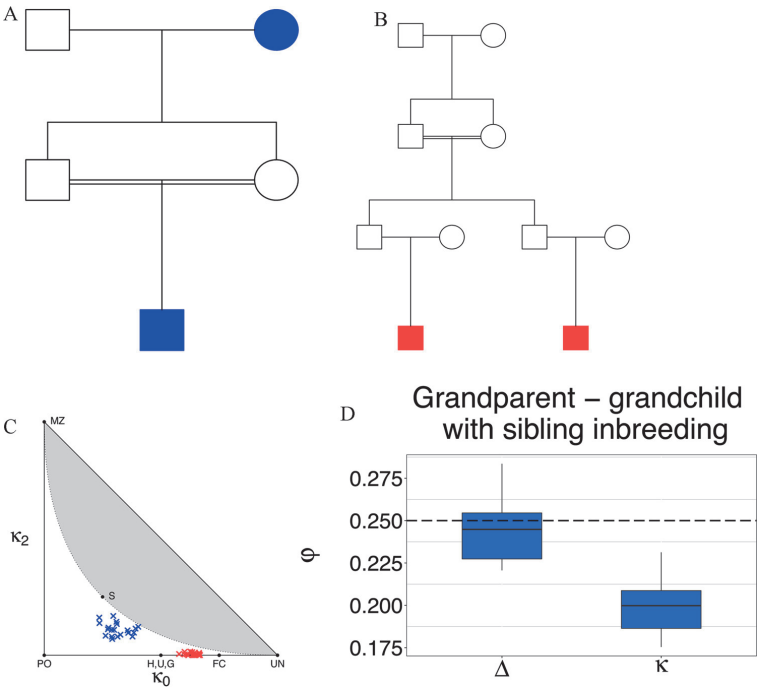


Fig. 1. IBD coefficients estimation and kinship coefficient computation. (For interpretation of the references to color in this figure citation, the reader is referred to the web version of this article.)

grandparent–grandchild pedigree. It is important to note that κ is in fact undefined for this relationship, since the grandchild is inbred. Estimates are located towards values of a sibling relationship (S), rather than a grandparent–grandchild relationship (G). The first cousins in Fig. 1b are not inbred, and so κ is defined and located at (0.652, 0) in the IBD triangle. Estimates of κ (red points) seems to be quite accurate. However, due to inbreeding, the values of the IBD coefficients between the cousins are different than what expected from a first cousin relationship, and this relationship between the individuals is likely to be rejected if the founder inbreeding is ignored.

To compare estimates of κ for inbred relationships and Δ, we calculate φ from both sets of coefficients. Fig. 1d shows φ for the inbred grandparent–grandchild relationship. Correct calculations using Δ give values of φ closer to the true coefficient ($\varphi = 0.25$, dashed line, which coincides with φ for a PO or S relationship), compared to calculations based on κ.

3.2. LR approach

For the LR approach, 1000 simulations of a set of 13 unlinked STR markers are performed, for the two individuals of interest.

Two pedigrees with inbreeding are considered. First, consider the inbred grandparent–grandchild relationship in Fig. 2a, with genetic data simulated on the grandmother and the grandson. Let H_P assumed a sibling relationship and H_D a grandparent–grandchild relationship. Based on these hypotheses, computations of the LR do not consider any inbreeding. The density plot in Fig. 2a shows that the median of the logarithm of LR is greater than zero, such that we have more trust in H_P, which states a sibling relationship, to explain the data, rather than H_D. An increase in the number of evaluated markers, yield an increase in the LR.

Next, consider a first cousin relationship with inbred founders [5], as in Fig. 2b. Genetic data are simulated on the two cousins, for increasing values of founder inbreeding. Let H_P assume a half sibling relationship, and H_D a first cousin relationship. The leftmost boxplot of Fig. 2b shows LR for $f = 0$. The logarithm of LR is negative, and we have more trust in H_D, a first cousin relationship, to be the origin of the data. As f increases, the LR becomes positive, such that the data is more likely to be explained by H_P, a half sibling relationship, rather than the first cousin relationship.

4. Discussion

In order to obtain reasonable accurate estimates, a higher amount of data needs to be considered, which makes the use of unlinked markers difficult. We have simulated linked markers, but estimates of parameters do not account for the linkage. We conjecture that properly accounting for linkage will only influence point estimates moderately for most applications. However, the impact on uncertainty estimates, confidence intervals, may be more substantial. This will be explored in future work.

5. Conclusion

A parametric approach to kinship inference aims to determine the relationship between two individuals through estimation of appropriate parameters, without any prior knowledge or assumptions of the relatedness between the individuals. In the LR framework, only the two stated hypotheses can be compared. As the examples presented above show, ignoring inbreeding in kinship evaluations can lead to wrong inference in pairwise kinship evaluations. Understanding and modeling inbreeding is important also for applications like disaster-victim

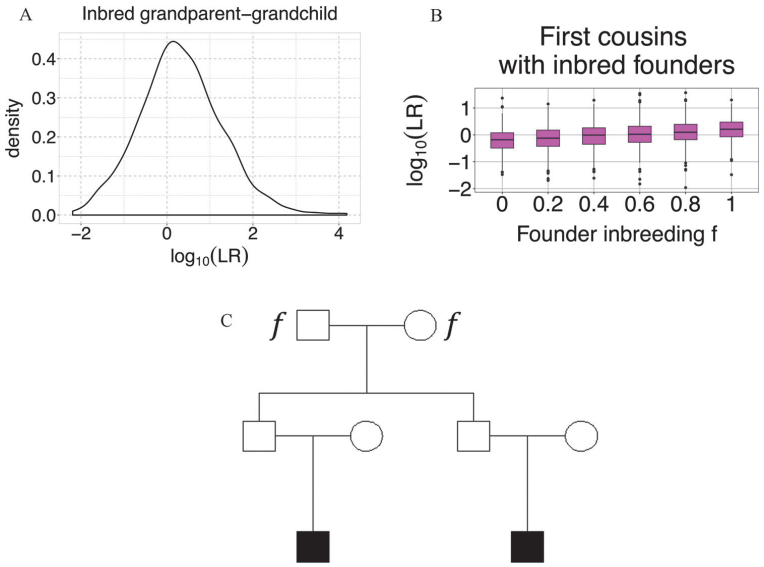


Fig. 2. LR computations.

identification.

Acknowledgments

Thanks to pedtools-suite developer Magnus Dehli Vigeland for the software and comments on the manuscript.

References

[1] S. Wright, Coefficients of inbreeding and relationship, Am. Nat. 56 (1922) 330–338.
[2] C.W. Cotterman, A Calculus for Statistico-Genetics, The Ohio State University, 1940 (Ph.D. thesis).
[3] A. Jacquard, Genetic information given by a relative, Biometrics 28 (1972) 1101–1114.
[4] G.R. Abecasis, et al., Merlin rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (1) (2001) 97.
[5] M.D. Vigeland, Relatedness coefficients in pedigrees with inbred founders, (2019) (submitted for publication).

Paper II

International Journal of Legal Medicine (2021) 135:117–129
https://doi.org/10.1007/s00414-020-02426-6

ORIGINAL ARTICLE

Pairwise relatedness testing in the context of inbreeding: expectation and variance of the likelihood ratio

Hilde Kjelgaard Brustad[1] · **Magnus Dehli Vigeland**[2] · **Thore Egeland**[1]

Received: 13 January 2020 / Accepted: 10 September 2020 / Published online: 28 September 2020

Abstract

In this paper we investigate various effects of inbreeding on the likelihood ratio (LR) in forensic kinship testing. The basic setup of such testing involves formulating two competing hypotheses, in the form of pedigrees, describing the relationship between the individuals. The likelihood of each hypothesis is computed given the available genetic data, and a conclusion is reached if the ratio of these exceeds some pre-determined threshold. An important aspect of this approach is that the hypotheses are usually not exhaustive: The *true* relationship may differ from both of the stated pedigrees. It is well known that this may introduce bias in the test results. Previous work has established formulas for the expected value and variance of the LR, given the two competing hypotheses and the true relationship. However, the proposed method only handles cases without inbreeding. In this paper we extend these results to all possible pairwise relationships. The key ingredient is formulating the hypotheses in terms of Jacquard coefficients instead of the more restricted Cotterman coefficients. While the latter describe the relatedness between outbred individuals, the more general Jacquard coefficients allow any level of inbreeding. Our approach also enables scrutiny of another frequently overlooked source of LR bias, namely background inbreeding. This ubiquitous phenomenon is usually ignored in forensic kinship computations, due to lack of adequate methods and software. By leveraging recent work on pedigrees with inbred founders, we show how background inbreeding can be modeled as a continuous variable, providing easy-to-interpret results in specific cases. For example, we show that if true siblings are subjected to a test for parent-offspring, moderate levels of background inbreeding are expected to inflate the LR by more than 50%.

Keywords Kinship analysis · Inbred founders · IBD triangle · Jacquard coefficients · Likelihood ratios

Introduction

The conventional approach to forensic kinship testing includes formulating two hypotheses and calculating a likelihood ratio (LR) based on genetic data from genotyped individuals. Practice differs between countries and laboratories, but typically the LR or some version of it is included when the case is reported. The conclusion based on the LR may be flawed when the true pedigree connecting the individuals of interest differs from the pedigrees considered by the hypotheses. As an example, consider a standard paternity case, where the prosecution asserts that a certain man is the father of a child, while the defense claims that the man and the child are unrelated. The truth, on the other hand, may be that the man is the child's uncle. A special case of incorrect hypotheses occurs when inbreeding is not accounted for. For example, if the alleged father is inbred, and this is ignored when formulating the hypotheses, this may significantly bias the LR. One aim of this paper is to investigate and quantify this effect.

Slooten and Egeland derived explicit equations for the expected value and variance of the LR [1]. They also extended this to cases where the true relationship differs from those stated in the hypotheses [2]. However, in both of these works only non-inbred individuals were considered.

✉ Hilde Kjelgaard Brustad
hilde.brustad@nmbu.no

Magnus Dehli Vigeland
m.d.vigeland@medisin.uio.no

Thore Egeland
thore.egeland@nmbu.no

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

[2] Department of Medical Genetics, University of Oslo, PB 4956 Nydalen, 0424 Oslo, Norway

An important contribution of this paper is the extension of these results to general pairwise relationships. In particular, we show that exact expressions for the expected value and variance of the LR can be obtained also in cases with inbreeding. The expressions are in general more involved than in the non-inbred case, and not as easy to interpret. However, we derive interesting and practical results in important special cases.

A parametric approach to modeling background inbreeding in kinship testing was recently introduced [3], employing the concept of inbred founders [4]. To exemplify, consider a pair of paternal half siblings, whose father is assigned an inbreeding coefficient f. As f increases from 0 to 1, the relationship between the half siblings becomes genetically indistinguishable from that between parent and child. We extend the theoretical framework of [1, 2] to pedigrees with inbred founders. As a result, the impact of background inbreeding on the expectation and variance of the LR can be studied based on exact expressions. In cases where the amount of inbreeding is unknown, we can still provide guidance on the expected values for the LR. Our approach conveniently allows a continuous range of possible true alternatives rather than a discrete set of specific alternatives. To arrive at explicit results of practical interest, we restrict attention to pairwise relationships. Furthermore, as in the work of Slooten and Egeland, we ignore mutations, dropouts, and silent alleles and we assume Hardy-Weinberg Equilibrium (HWE). However, we explain how deviation from HWE can be modeled by the so called theta (θ) correction.

R scripts and functions used to obtain numerical results in this paper are gathered in a R library (see the "R implementation" section). Pedigree likelihoods and marker simulations are performed with the forrel package [3].

This paper is organized in the following manner: After establishing some terminology and notation we review the main results of [2] regarding the expected value and variance of the LR for non-inbred pairs of individuals. We then proceed to extend these results to general pairwise relationships, including relationships in pedigrees with background inbreeding. Several worked examples follow, including a simulation study comparing our formulas with real-life results. Finally, we discuss some consequences of this work and how it relates to other aspects of forensic genetics.

Definitions and notation

A central concept for measuring genetic relatedness is that of *identity by descent* (IBD). Two alleles are said to be IBD relative to a given pedigree if they are identical by state and originate from the same ancestral allele within the pedigree [5].

Coefficients of inbreeding and kinship

The *coefficient of inbreeding* f, introduced by Wright [6], is the probability that an individual is autozygous at a given autosomal locus, i.e., that the two homologous alleles are IBD. This is the same as the *kinship coefficient* φ between the parents of the same individual, defined as the probability that a random allele from the mother is IBD to a random allele from the father at the same locus.

Founders of a pedigree are conventionally assumed to be unrelated and non-inbred. Following [3] we relax the second assumption, allowing an arbitrary inbreeding coefficient f to be assigned to any founder individual. For a given pedigree with N founders, we denote the set of founder inbreeding coefficients by $\boldsymbol{f} = (f_1, f_2, \ldots, f_N)$.

Background inbreeding in human populations is normally low, but may exceed 5% in certain cases [7, 8]. In forensic case work inbreeding is common, ranging from consanguineous marriages between cousins, $f = 1/16$ or lower, to incestuous relationships between siblings or parent-child, both with $f = 1/4$. In breeding applications values closer to 1 may occur.

Jacquard coefficients and likelihood of a pedigree

The kinship coefficient is a coarse measure of relatedness; for instance, it has the same value for a parent-child relationship as for full siblings. A more refined measure is given by the nine *Jacquard coefficients* [9] $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_9)$, also called the *condensed identity coefficients*. These are the expected relative frequencies of the

Jacquard states $J_1, \ldots, J_9$ are depicted in Fig. 1. Alleles within each individual are unordered, and hence, several IBD configurations can correspond to the same Jacquard state. Furthermore, $\boldsymbol{\Delta}$ is related to φ through

$$\varphi = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8.$$

The likelihood of two individuals being related according to $\boldsymbol{\Delta}$, given their genotypes $G = (g_1, g_2)$ at a marker may be expressed by conditioning on the Jacquard state:

$$L(\boldsymbol{\Delta} \mid G) = \sum_{i=1}^{9} \Delta_i P(G \mid J_i). \tag{1}$$

The conditional probabilities $P(G \mid J_i)$ are listed in Table 1. These probabilities are found by direct calculations; for instance, $P((aa, aa) \mid J_1) = p_a$ since J_1 dictates that all four alleles are IBD.

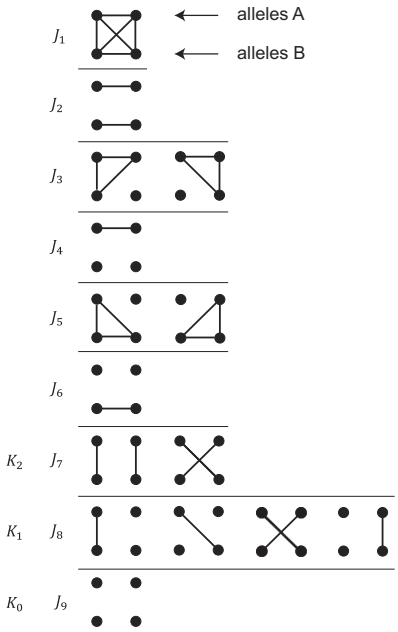


Fig. 1 The Jacquard states $J_1, \ldots, J_9$ representing all possible IBD patterns among the four alleles of two individuals at an autosomal locus. Each row of dots represents the two alleles of an individual. Connected dots indicate IBD. The states J_9, J_8, and J_7 do not involve inbreeding and are sometimes denoted K_0, K_1, and K_2

IBD coefficients and inbred founders

For two non-inbred individuals, the first six Jacquard coefficients are zero, and Δ_9, Δ_8, and Δ_7 reduce to the IBD coefficients $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$ introduced by Cotterman [10]. They give the probabilities that, at a given autosomal locus, the individuals share zero-, one-, and two-allele IBD, respectively. Note that $\kappa_0 + \kappa_1 + \kappa_2 = 1$, so $\boldsymbol{\kappa}$ can be represented in a two-dimensional triangle with axes κ_0 and κ_2. Thompson [11] showed that the IBD coefficients are restricted to $\kappa_1^2 \geq 4\kappa_0\kappa_2$. This gives rise to an inadmissible region for the parameters, in gray in Fig. 2.

Although the IBD coefficients are only defined for non-inbred individuals, other members of the pedigree can be inbred. For example, a pair of half siblings remain outbred even if their shared parent is inbred. However, this inbreeding will affect the relatedness coefficients. Table 2 lists the kinship and the IBD coefficients for some common relationships, as functions of the founder inbreeding. The effects are visualized in Fig. 2. In the half sibling example, the genetic relationship approaches that of parent-child, as the founder inbreeding increases towards 1. Similarly, the IBD coefficients of full siblings with inbred parents may fall anywhere in the lightly shaded region towards the point of monozygotic twins.

Review of previous results

We next review the main results of [2] relevant for our work. In particular we restate the explicit formulas for the expectation and variance of the LR in the case of non-inbred individuals.

The likelihood ratio as a random variable

We consider a kinship test involving genetic data from two non-inbred individuals. Two hypotheses H_P and H_D about the relationship are to be compared using the LR. For our purposes, each hypothesis corresponds to a point in the IBD triangle, denoted by $\boldsymbol{\kappa}_P$ and $\boldsymbol{\kappa}_D$ respectively. However, the evidence may be generated from another pedigree, corresponding to a third point $\boldsymbol{\kappa}_T$. We therefore have the

Table 1 The conditional probability $P(G \mid J_i)$ of a pair of genotypes $G = (g_1, g_2)$, given a Jacquard state J_i

G	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
(aa, aa)	p_a	p_a^2	p_a^2	p_a^3	p_a^2	p_a^3	p_a^2	p_a^3	p_a^4
(aa, bb)	0	$p_a p_b$	0	$p_a p_b^2$	0	$p_a^2 p_b$	0	0	$p_a^2 p_b^2$
(aa, ab)	0	0	$p_a p_b$	$2p_a^2 p_b$	0	0	0	$p_a^2 p_b$	$2p_a^3 p_b$
(aa, bc)	0	0	0	$2p_a p_b p_c$	0	0	0	0	$2p_a^2 p_b p_c$
(ab, aa)	0	0	0	0	$p_a p_b$	$2p_a^2 p_b$	0	$p_a^2 p_b$	$2p_a^3 p_b$
(bc, aa)	0	0	0	0	0	$2p_a p_b p_c$	0	0	$2p_a^2 p_b p_c$
(ab, ab)	0	0	0	0	0	0	$2p_a p_b$	$p_a p_b(p_a + p_b)$	$4p_a^2 p_b^2$
(ab, ac)	0	0	0	0	0	0	0	$p_a p_b p_c$	$4p_a^2 p_b p_c$
(ab, cd)	0	0	0	0	0	0	0	0	$4p_a p_b p_c p_d$

The symbols a, b, c, and d represent different alleles, with population frequencies p_a, p_b, p_c, and p_d respectively

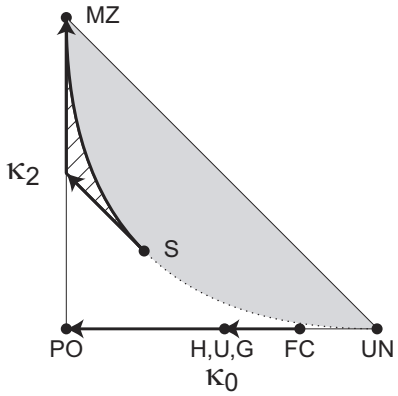


Fig. 2 The IBD triangle with location of some common relationships. The gray area is inadmissable. The arrows illustrate the effect of founder inbreeding in the cases given in Table 2. PO, parent-child; MZ, monozygotic twins; S, siblings; H, half siblings; U, avuncular; G, grandparent grandchild; FC, first cousins; UN, unrelated

following setup, comprising the competing hypotheses and the true relationship:

$$\begin{aligned} H_P: &\quad \boldsymbol{\kappa} = \boldsymbol{\kappa}_P = (\kappa_0^P, \kappa_1^P, \kappa_2^P) \\ H_D: &\quad \boldsymbol{\kappa} = \boldsymbol{\kappa}_D = (\kappa_0^D, \kappa_1^D, \kappa_2^D) = (1, 0, 0) \\ Truth: &\quad \boldsymbol{\kappa} = \boldsymbol{\kappa}_T = (\kappa_0^T, \kappa_1^T, \kappa_2^T). \end{aligned}$$

Reflecting standard practice, we will always use *unrelatedness* as the defense hypothesis, i.e., $\boldsymbol{\kappa}_D = (1, 0, 0)$. It should be noted, however, that this is not a theoretical requirement for the methods presented here.

The concept of the likelihood ratio as a random variable was discussed by Slooten and Egeland [1]. We review the basics here, presented in a slightly simpler notation sufficient for our purposes.

Denote by K_i, $i = 0, 1, 2$, the event that the individuals share exactly i alleles IBD. As shown in Fig. 1, K_0, K_1, and K_2 are identical to the Jacquard states J_9, J_8, and J_7 respectively. For fixed $\boldsymbol{\kappa}_P$ the likelihood ratio for a given pair of genotypes $G = (g_1, g_2)$ can be written as

$$\begin{aligned} LR(G) = \frac{P(G \mid H_P)}{P(G \mid H_D)} &= \frac{P(G \mid \boldsymbol{\kappa}_P)}{P(G \mid \boldsymbol{\kappa}_D)} \\ &= \sum_{i=0}^{2} \kappa_i^P \frac{P(G \mid K_i)}{P(G \mid K_0)}. \end{aligned} \tag{2}$$

Note that the final transition was obtained by applying (1) in both the numerator and denominator. The probabilities $P(G \mid K_i)$ are given in Table 1.

Now, viewing the genotypes as a random variable $\mathcal{G}$, we define the random variable $\mathcal{LR} = LR(\mathcal{G})$. Note that the distribution of $\mathcal{G}$ is completely determined by $\boldsymbol{\kappa}_T$ (assuming HWE), hence the distribution of $\mathcal{LR}$ is determined by $\boldsymbol{\kappa}_P$ and $\boldsymbol{\kappa}_T$. If these parameters are clear from the context, we will suppress them in our notation; otherwise, we write

Table 2 Relatedness coefficients as functions of founder inbreeding, in a selection of common relationships

	Relationship	φ	$\varphi(\boldsymbol{f})$	$\boldsymbol{\kappa}$	$\boldsymbol{\kappa}(\boldsymbol{f})$
	S	$\frac{1}{4}$	$\frac{1}{4}(1+\frac{f_1+f_2}{2})$	$(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$	$\kappa_0(f_1,f_2) = \frac{1}{4}(1-f_1)(1-f_2)$ $\kappa_1(f_1,f_2) = \frac{1}{2}(1-f_1 f_2)$ $\kappa_2(f_1,f_2) = \frac{1}{4}(1+f_1)(1+f_2))$
	H	$\frac{1}{8}$	$\frac{1}{8}(1+f)$	$(\frac{1}{2}, \frac{1}{2}, 0)$	$\kappa_0(f) = \frac{1}{2}(1-f)$ $\kappa_1(f) = \frac{1}{2}(1+f)$ $\kappa_2(f) = 0$
	U	$\frac{1}{8}$	$\frac{1}{8}(1+\frac{f_1+f_2}{2})$	$(\frac{1}{2}, \frac{1}{2}, 0)$	$\kappa_0(f_1,f_2) = \frac{1}{2}(1-\frac{f_1+f_2}{2})$ $\kappa_1(f_1,f_2) = \frac{1}{2}(1+\frac{f_1+f_2}{2})$ $\kappa_2(f_1,f_2) = 0$
	FC	$\frac{1}{16}$	$\frac{1}{16}(1+\frac{f_1+f_2}{2})$	$(\frac{3}{4}, \frac{1}{4}, 0)$	$\kappa_0(f_1,f_2) = \frac{1}{4}(3-\frac{f_1+f_2}{2})$ $\kappa_1(f_1,f_2) = \frac{1}{4}(1+\frac{f_1+f_2}{2})$ $\kappa_2(f_1,f_2) = 0$

$\mathcal{LR}_{\boldsymbol{\kappa}_P,\boldsymbol{\kappa}_T}$. In the special case when H_P equals the truth, i.e., $\boldsymbol{\kappa}_P = \boldsymbol{\kappa}_T$, we may simplify $\mathcal{LR}_{\boldsymbol{\kappa}_P,\boldsymbol{\kappa}_T}$ to $\mathcal{LR}_{\boldsymbol{\kappa}_P}$.

Throughout, we assume the following condition to hold

$$P(G \mid H_P) > 0 \Rightarrow P(G \mid H_D) > 0. \tag{3}$$

In the present context, it means that all DNA profiles that can occur under H_P, can also occur under H_D. In our examples H_D specifies unrelated individuals, and then (3) holds. The condition also holds for mutation models provided all elements of the mutation matrix are positive. We do not model mutations in the work presented here, as practical exact expression are then no longer available. However, the implementation allows for general mutation models. Without (3), likelihood ratios could be infinite, i.e., not defined.

Expected likelihood ratio

The expectation of $\mathcal{LR}$ may be found by summing over all possible genotypes G in the standard way:

$$E(\mathcal{LR}) = \sum_G P(G) LR(G), \tag{4}$$

where $P(G) = P(G \mid \boldsymbol{\kappa}_T) = \sum_i \kappa_i^T P(G \mid K_i)$. An exact expression for $E(\mathcal{LR})$ when $\boldsymbol{\kappa}_P = \boldsymbol{\kappa}_T$ was first derived in [1] and extended in [2] to apply when $\boldsymbol{\kappa}_P \neq \boldsymbol{\kappa}_T$. For the latter situation it was shown that, for a single marker with L alleles,

$$E(\mathcal{LR}) = \boldsymbol{\kappa}_P \cdot A_0 \cdot (\boldsymbol{\kappa}_T)^t, \tag{5}$$

where t denotes the vector transpose, and

$$A_0 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \frac{L+3}{4} & \frac{L+1}{2} \\ 1 & \frac{L+1}{2} & \frac{L(L+1)}{2} \end{pmatrix}. \tag{6}$$

Importantly, the expected value depends only on the number of alleles, not on the allele frequencies. Furthermore, the expectation is symmetric in $\boldsymbol{\kappa}_P$ and $\boldsymbol{\kappa}_T$, so that

$$E(\mathcal{LR}_{\boldsymbol{\kappa}_P,\boldsymbol{\kappa}_T}) = E(\mathcal{LR}_{\boldsymbol{\kappa}_T,\boldsymbol{\kappa}_P}). \tag{7}$$

Variance of the likelihood ratio

To derive the variance of $\mathcal{LR}$ we apply the general formula $\text{var}(\mathcal{X}) = E(\mathcal{X}^2) - E(\mathcal{X})^2$. Since the last term follows from Eq. 5, all that remains is to find the first term. Some notation is needed:

$$\begin{aligned} s_1 &= \tfrac{1}{16} \sum_{a<b} \left(\tfrac{p_a}{p_b} + \tfrac{p_b}{p_a} \right), \\ s_2 &= \sum_{a<b} \tfrac{1}{2 p_a p_b}, \\ s_3 &= \sum_a \tfrac{1}{p_a}, \\ s_4 &= \tfrac{1}{4} \sum_{a<b} \left(\tfrac{1}{p_b} + \tfrac{1}{p_a} \right), \\ s_5 &= \sum_a \tfrac{1}{p_a^2}. \end{aligned}$$

Furthermore, supplementing the matrix A_0 given in Eq. 6, we define matrices A_1 and A_2 by

$$A_1 = \begin{pmatrix} 1 & \frac{L+3}{4} & \frac{L+1}{2} \\ \frac{L+3}{4} & \frac{5L+3}{8} + s_1 & \frac{L(L+7)}{8} + 2s_1 \\ \frac{L+1}{2} & \frac{L(L+7)}{8} + 2s_1 & s_3 + s_4 \end{pmatrix} \tag{8}$$

$$A_2 = \begin{pmatrix} 1 & \frac{L+1}{2} & \frac{L(L+1)}{2} \\ \frac{L+1}{2} & \frac{L(L+7)}{8} + 2s_1 & s_3 + s_4 \\ \frac{L(L+1)}{2} & s_3 + s_4 & s_2 + s_5 \end{pmatrix} \tag{9}$$

It was shown in [2] that

$$E(\mathcal{LR}^2) = \sum_{i=0}^{2} \kappa_i^P \boldsymbol{\kappa}_P A_i (\boldsymbol{\kappa}_T)^t;$$

hence, the complete variance expression becomes

$$\text{var}(\mathcal{LR}) = \sum_{i=0}^{2} \kappa_i^P \boldsymbol{\kappa}_P A_i (\boldsymbol{\kappa}_T)^t - \left(\boldsymbol{\kappa}_P A_0 (\boldsymbol{\kappa}_T)^t \right)^2. \tag{10}$$

Contrary to the expected LR, the variance of the LR depends on the allele frequencies.

Example: paternity testing

This example serves as an illustration of the above described expected LR and the corresponding hypotheses. Consider a paternity case, where a man is claimed to be the father of a child (H_P). The truth is that a brother of the alleged father is the true father of the child. The hypotheses and the true relatedness are in terms of the IBD coefficients given as

$$\begin{aligned} H_P: &\quad \boldsymbol{\kappa} = \boldsymbol{\kappa}_P &&= (0, 1, 0) \\ H_D: &\quad \boldsymbol{\kappa} = \boldsymbol{\kappa}_D &&= (1, 0, 0) \\ Truth: &\quad \boldsymbol{\kappa} = \boldsymbol{\kappa}_T &&= (\tfrac{1}{2}, \tfrac{1}{2}, 0). \end{aligned} \tag{11}$$

Figure 3 illustrates the hypotheses in terms of pedigrees, and as points in the IBD triangle. Equation (5), with IBD coefficients as in Eq. 11, simplifies to

$$E(\mathcal{LR}) = \frac{L+7}{8}. \tag{12}$$

The variance of $\mathcal{LR}$ becomes

$$\text{var}(\mathcal{LR}) = \tfrac{7L+9}{16} + \tfrac{s_1}{2} - \left(\tfrac{L+7}{8} \right)^2.$$

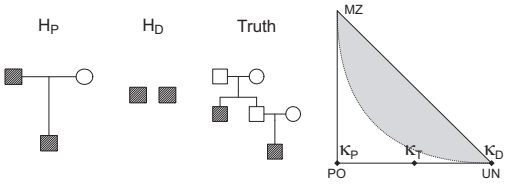


Fig. 3 Pedigrees and location of IBD coefficients $\boldsymbol{\kappa}_P$, $\boldsymbol{\kappa}_D$, and $\boldsymbol{\kappa}_T$ for a paternity case when the true relationship is avuncular

In the special case $L = 2$, and allele frequencies q and $1 - q$, the variance expression reduces to

$$\text{var}(\mathcal{LR}) = \frac{11}{64} + \frac{1}{32}\frac{(1-q)^2 + q^2}{q(1-q)}.$$

This expression is minimal when $q = 0.5$ and becomes infinitely large when q or $1 - q$ approaches 0. If no assumption is made for L, but all alleles are assumed equally frequent, the variance reduces to

$$\text{var}(\mathcal{LR}) = \frac{L(L+12)}{64} - \frac{13}{64}. \tag{13}$$

Table 3 exemplifies these formulas for various realistic values of L, and compares the results with the corresponding values if H_P was true.

Likelihood ratio for general pairwise relationships

In this section we extend the results reviewed above to relationships between any pairs of individuals. In particular we now allow inbreeding. For this to work we must pass from the IBD coefficients to the full set of Jacquard coefficients. For details regarding derivations of the results (see the Appendix).

Expected likelihood ratio

We use the same setup for kinship testing as introduced previously, but in order to allow general inbreeding, we now formulate our hypotheses using Jacquard coefficients,

$$\begin{aligned} H_P: \quad \boldsymbol{\Delta} = \boldsymbol{\Delta}_P \quad &= (\Delta_1^P, \ldots, \Delta_9^P) \\ H_D: \quad \boldsymbol{\Delta} = \boldsymbol{\Delta}_D \quad &= (\Delta_1^D, \ldots, \Delta_9^D) = (0, \ldots, 0, 1) \\ Truth: \quad \boldsymbol{\Delta} = \boldsymbol{\Delta}_T \quad &= (\Delta_1^T, \ldots, \Delta_9^T). \end{aligned}$$

Note that the defense hypothesis still corresponds to unrelatedness. We are interested in the likelihood ratio comparing H_P with H_D when the genotypes are generated by a pedigree with the Jacquard coefficients $\boldsymbol{\Delta}_T$. Equation (1) implies that

$$\begin{aligned} LR(G) &= \frac{P(G \mid \boldsymbol{\Delta}_P)}{P(G \mid \boldsymbol{\Delta}_D)} \\ &= \sum_{i=1}^{9} \Delta_i^P \frac{P(G \mid J_i)}{P(G \mid J_9)}. \end{aligned} \tag{14}$$

As shown in the Appendix, the expected LR is

$$E(\mathcal{LR}_{\boldsymbol{\Delta}_P, \boldsymbol{\Delta}_T}) = \boldsymbol{\Delta}_P B_9 (\boldsymbol{\Delta}_T)^t, \tag{15}$$

where B_9 is the symmetric 9×9 matrix given in Table 4, whose elements are $E(\mathcal{LR}_{J_i, J_j})$, for $1 \le i, j \le 9$. As opposed to the non-inbred case, we see that the expected value in general depends on the allele frequencies.

Variance of the likelihood ratio

In the Appendix matrices $B_1, \ldots, B_9$ are defined and it is shown that

$$E(\mathcal{LR}^2) = \sum_{i=1}^{9} \Delta_i^P \boldsymbol{\Delta}_P B_i (\boldsymbol{\Delta}_T)^t. \tag{16}$$

From this we obtain the variance formula

$$\begin{aligned} &\text{var}(\mathcal{LR}) = \\ &\sum_{i=1}^{9} \Delta_i^P \boldsymbol{\Delta}_P B_i (\boldsymbol{\Delta}_T)^t - \left(\boldsymbol{\Delta}_P B_9 (\boldsymbol{\Delta}_T)^t\right)^2. \end{aligned} \tag{17}$$

Pairwise relationships with inbred founders

As previously explained, a set of inbreeding coefficients $\boldsymbol{f}$ can be assigned the founders of a pedigree to model background inbreeding. The Jacquard coefficients of any pair of pedigree members are then functions of $\boldsymbol{f}$. It follows that the formulas for expectation and variance of $\mathcal{LR}$

Table 3 Expectation and variance of $\mathcal{LR}$ in the paternity example of Fig. 3, for loci with 2, 10, and 50 alleles

Truth	$\boldsymbol{\kappa}_P$	$\boldsymbol{\kappa}_T$	E[LR]	$L = 2$	$L = 10$	$L = 50$
PO	(0, 1, 0)	(0, 1, 0)	$\frac{L+3}{4}$	1.250 (0.188)	3.250 (1.686)	13.250 (9.188)
U	(0, 1, 0)	$(\frac{1}{2}, \frac{1}{2}, 0)$	$\frac{L+7}{8}$	1.125 (0.234)	2.125 (3.234)	7.125 (48.230)

The variances are computed assuming uniform allele frequencies. The bottom row (U) shows the values when the true pedigree is uncle-nephew, as analyzed in the main text. For comparison, the top row shows the corresponding numbers when H_P is true

Table 4 Elements of the symmetric matrix B_9, given as $E(\mathcal{LR}_{J_i,J_j})$

	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
J_1	$\sum \frac{1}{p_a^2}$	$\sum \frac{1}{p_a}$	$\sum \frac{1}{p_a}$	L	$\sum \frac{1}{p_a}$	L	$\sum \frac{1}{p_a}$	L	1
J_2		L^2	L	L	L	L	L	1	1
J_3			$\frac{1}{2}(L+\sum \frac{1}{p_a})$	L	L	1	L	$\frac{L+1}{2}$	1
J_4				L	1	1	1	1	1
J_5					$\frac{1}{2}(L+\sum \frac{1}{p_a})$	L	L	$\frac{L+1}{2}$	1
J_6						L	1	1	1
J_7							$\frac{L(L+1)}{2}$	$\frac{L+1}{2}$	1
J_8								$\frac{L+4}{3}$	1
J_9									1

Each row represents J_i, a Jacquard state assumed by H_P, while each column presents J_j, the true Jacquard state

involving such pedigrees remain as in Eqs. 15 and 17, except that the parameters $\boldsymbol{\Delta}_P$ and $\boldsymbol{\Delta}_T$ must be updated.

Specifically, let $\boldsymbol{f}_P$ be a vector of founder inbreeding coefficients in the pedigree assumed by H_P, and $\boldsymbol{f}_T$ similarly in the true pedigree. The expectation and variance of $\mathcal{LR}$ in this situation are then given by

$$E(\mathcal{LR}_{\boldsymbol{\Delta}_P(\boldsymbol{f}_P),\boldsymbol{\Delta}_T(\boldsymbol{f}_T)}) = \boldsymbol{\Delta}_P(\boldsymbol{f}_P)B_9(\boldsymbol{\Delta}_T(\boldsymbol{f}_T))^t$$

and

$$\begin{aligned}&\mathrm{var}(\mathcal{LR}_{\boldsymbol{\Delta}_P(\boldsymbol{f}_P),\boldsymbol{\Delta}_T(\boldsymbol{f}_T)})\\ &= \sum_{i=1}^{9} \Delta_i^P(\boldsymbol{f}_P)\boldsymbol{\Delta}_P(\boldsymbol{f}_P)B_i(\boldsymbol{\Delta}_T(\boldsymbol{f}_T))^t\\ &\quad - \left(\boldsymbol{\Delta}_P(\boldsymbol{f}_P)B_9(\boldsymbol{\Delta}_T(\boldsymbol{f}_T))^t\right)^2.\end{aligned}$$

Note that the matrices B_i only depend on L and the allele frequencies, and therefore are unchanged by founder inbreeding.

Remark 1 It should be emphasized that the formulas (15) and (17) are needed only when at least one of the tested individuals are inbred in some of the involved pedigrees. If both are non-inbred, the simpler expressions (5) and (10) using IBD coefficients suffice. Importantly, this remains true if *other* members of the pedigree are inbred, as long as this does not lead to inbreeding in the tested individuals. In particular, founder inbreeding may be accounted for in Eqs. 5 and 10 simply by replacing $\boldsymbol{\kappa}_P$ and $\boldsymbol{\kappa}_T$ by $\boldsymbol{\kappa}_P(\boldsymbol{f}_P)$ and $\boldsymbol{\kappa}_T(\boldsymbol{f}_T)$ respectively.

Founder inbreeding and θ correction

The conventional approach to background relatedness in forensics is the so called θ correction [12]. In an inbred population, the composition of genotypes do not follow the Hardy-Weinberg principle, implying that the frequencies given in Table 1 no longer hold. The following approach compensates for this by adjusting the allele frequencies. Without loss of generality we can assume that alleles observed are sampled sequentially. The probability that allele i is sampled as the jth allele is given by the *sampling formula*

$$p_i' = \frac{b_j\theta + \bar{\theta}p_i}{1+(j-2)\theta}, \tag{18}$$

where $\bar{\theta} = 1-\theta$ and b_j denotes the number of alleles of type i among the $j-1$ previously sampled. Note that for pairwise cases, the likelihood can be written

$$L(\boldsymbol{\Delta}(\boldsymbol{f}) \mid G,\theta) = \sum_{i=1}^{9} \Delta_i(\boldsymbol{f})P(G \mid J_i,\theta), \tag{19}$$

where $P(G \mid J_i,\theta)$ is calculated using Eq. 18. The matrices $B_1, ..., B_9$ then change with θ, modifying the expectation and variance of the LR. This emphasises a fundamental difference between founder inbreeding and θ correction: $\boldsymbol{f}$ modifies the relationship itself, while θ only impacts the genotype probabilities.

Example: θ correction and founder inbreeding in a paternity case

This example compares θ correction to founder inbreeding. Consider first the hypothesis H_D: A and B are unrelated. Assume both individuals are homozygous a/a. Equation (18) gives the likelihood

$$L_\theta(H_D) = p_a(\theta + \bar{\theta}p_a)\,\frac{2\theta+\bar{\theta}p_a}{1+\theta}\,\frac{3\theta+\bar{\theta}p_a}{1+2\theta}.$$

If rather than using θ correction, we assign an inbreeding coefficient f to A, the likelihood becomes

$$L_f(H_D) = (fp_a + (1-f)p_a^2)p_a^2.$$

Consider next the hypothesis H_{P1}: A is the father of B. Equation (18) now gives

$$L_\theta(H_{P1}) = p_a(\theta + \bar{\theta} p_a)\frac{2\theta + \bar{\theta} p_a}{1+\theta}$$

and so the LR with θ correction is

$$LR_\theta = \frac{L_\theta(H_{P1})}{L_\theta(H_D)} = \frac{1+2\theta}{3\theta + \bar{\theta} p_a}.$$

The inbreeding coefficient approach gives

$$L_f(H_{P1}) = (f p_a + (1-f)p_a^2)p_a$$

and $LR_f = 1/p_a$. Note that the LR does not depend on f and that this is true for all genotype combinations for A and B. The LRs for other genotype combinations for A and B with θ correction are given in Table 10.8 in [13].

To illustrate (19) consider the hypothesis H_{P2}: A and B are paternal half siblings whose father is inbred. Table 2 then gives $\Delta_8 = \frac{1}{2}(1+f)$ and $\Delta_9 = \frac{1}{2}(1-f)$, and by Eqs. 18 and 19 we may write down the likelihood for any genotype combinations. For instance, when A is homozygous a/a and B homozygous b/b the likelihood is

$$L(f,\theta) = \frac{1}{2}(1-f)p_a(\theta + \bar{\theta} p_a)\,\frac{\bar{\theta} p_b}{1+\theta}\,\frac{\theta + \bar{\theta} p_b}{1+2\theta}.$$

The LR comparing H_{P2} with A and B being unrelated becomes $\frac{1}{2}(1-f)$. If A and B share alleles, the LR will depend also on θ.

R implementation

Utilities to perform the computations in this paper are provided in a R library named InbredLR, available from the first author, building on several packages in the *ped suite*, notably pedprobr and forrel [3]. The core of InbredLR are functions that compute the expectation and variance of the likelihood ratio for pairwise relationships. The user can specify the parameters ($\boldsymbol{\kappa}$, $\boldsymbol{f}$ or $\boldsymbol{\Delta}$) or specify the pedigrees, possibly with inbred founders. A function for simulating marker data to estimate the distribution of LR is also provided, as well as a function for visualizing pedigrees H_P and H_D and the true pedigree and location of the corresponding IBD coefficients in the IBD triangle.

Results

Paternity case for siblings with inbred founders

Consider two individuals who claim to be related as parent and offspring. Their true relationship is siblings and their parents coefficients of inbreeding are $\boldsymbol{f}_T = (f_1, f_2)$. Figure 4 shows the case. This example can be relevant

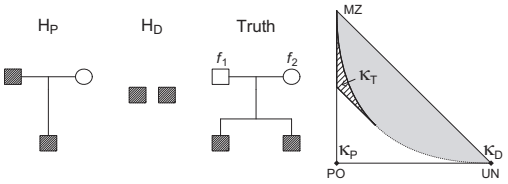


Fig. 4 Hypotheses involved in "Paternity case for siblings with inbred founders" and the location of the corresponding IBD coefficients $\boldsymbol{\kappa}_P$, $\boldsymbol{\kappa}_D$, and $\boldsymbol{\kappa}_T$ in the IBD triangle

for family reunion cases, where a parent-child relationship would give right to residence permit, whereas a sibling relationship would not. In [14] such a case is considered. H_P and H_D and their true relationship are in terms of the IBD coefficients given as

$$\begin{array}{ll} H_P: & \boldsymbol{\kappa} = \boldsymbol{\kappa}_P = (0,1,0) \\ H_D: & \boldsymbol{\kappa} = \boldsymbol{\kappa}_D = (1,0,0) \\ Truth: & \boldsymbol{\kappa} = \boldsymbol{\kappa}_T(\boldsymbol{f}_T), \end{array} \tag{20}$$

where $\boldsymbol{\kappa}_T(\boldsymbol{f}_T) = \boldsymbol{\kappa}_T(f_1, f_2)$ are as in the first row of Table 2. Keeping in mind Remark 1, we apply (5) to find the expected LR:

$$E(\mathcal{LR}) = \frac{L-1}{8}(f_1+f_2) + \frac{L+3}{4}. \tag{21}$$

Figure 5 plots $E(\mathcal{LR})$ as a function of the inbreeding level (assuming $f_1 = f_2$), for a single locus with $L = 2$, 10 and 50 alleles.

Without founder inbreeding, $E(\mathcal{LR}) = (L+3)/4$. Interestingly, this is the same as the expectation if H_P was true, i.e., if the two individuals were in fact father and son (see first row of Table 3). The variance of $\mathcal{LR}$ differs between the two cases, however (not shown here).

As the background inbreeding of the true sibling pedigree increases, $E(\mathcal{LR})$ increases. The expected LR of the

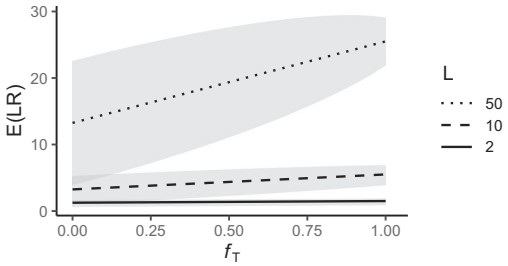


Fig. 5 $E(\mathcal{LR})$ as function of background inbreeding level f_T (assuming $f_1 = f_2$), for $L = 2$, 10, and 50 alleles, for the paternity case in Fig. 4. The shaded area shows one standard deviation below and above $E(\mathcal{LR})$, for uniform allele frequencies

paternity case (and hence the trust in H_P) is therefore higher if the true relatedness is siblings with background inbreeding, rather than the tested parent-child relationship. The variance of $\mathcal{LR}$ decreases moderately for increasing founder inbreeding. For increasing number of alleles L, the slope of the expected LR increases.

The following calculation gives a simple approximation of the inflation in the expected LR caused by background inbreeding. Suppose $f_1 = f_2 = f$, and write (21) as $\mu_0 + \mu_f$, where $\mu_0 = \frac{1}{4}(L+3)$ is the expected LR without founder inbreeding, and $\mu_f = \frac{1}{4}(L-1)f$ is the expected contribution caused by founder inbreeding. Note that $\mu_0 + \mu_f = (1 + \frac{\mu_f}{\mu_0})\mu_0$, and that for $L \geq 5$ we have $\frac{\mu_f}{\mu_0} = \frac{L-1}{L+3} f \geq \frac{1}{2} f$. This implies that with N independent markers, the total LR has expectation

$$[(1 + \tfrac{\mu_f}{\mu_0})\mu_0]^N \geq (1 + \tfrac{1}{2}f)^N \mu_0^N \geq (1 + \tfrac{1}{2}fN)\mu_0^N.$$

This means that a background inbreeding level f will inflate the expected LR by at least $\frac{1}{2}fN$. For example, if $N = 20$ and $f = 0.05$, the inflation rate is greater than 50%.

Siblings and half siblings with founder inbreeding

Distinguishing between siblings and half siblings can be difficult based on unlinked markers. Mayor and Balding address the problem in [15], with focus on the number of loci needed. If the shared parent of the half siblings has inbreeding coefficient $f_T > 0$, the problem becomes even more interesting.

Consider the situation shown in Fig. 6. The hypotheses are

$$\begin{aligned} H_P: &\quad \boldsymbol{\kappa} = \boldsymbol{\kappa}_P(\boldsymbol{f}_P) \\ H_D: &\quad \boldsymbol{\kappa} = \boldsymbol{\kappa}_D = (1,0,0) \\ Truth: &\quad \boldsymbol{\kappa} = \boldsymbol{\kappa}_T(f_T), \end{aligned} \tag{22}$$

where $\boldsymbol{f}_P = (f_1, f_2)$ are the parental inbreeding coefficients in the H_P pedigree and $\boldsymbol{\kappa}_P(\boldsymbol{f}_P)$ and $\boldsymbol{\kappa}_T(f_T)$ are as in the first and second rows of Table 2, respectively.

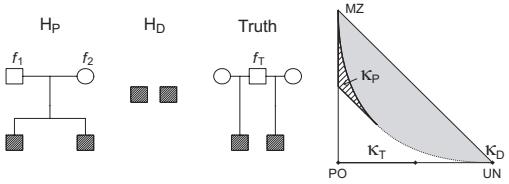


Fig. 6 The hypotheses involved in "Siblings and half siblings with founder inbreeding" and the location of the corresponding IBD coefficients $\boldsymbol{\kappa}_P$, $\boldsymbol{\kappa}_D$, and $\boldsymbol{\kappa}_T$ in the IBD triangle

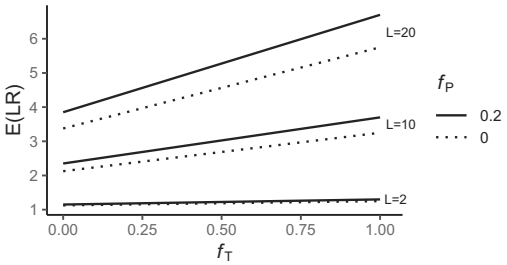


Fig. 7 $E(\mathcal{LR})$ for the case in Fig. 6 as functions of background inbreeding level f_T, for $f_P = 0$ (dashed line) and $f_P = 0.2$ (solid line), and $L = 2$, 10, and 20

This setup facilitates for modeling background inbreeding in both the true pedigree and in H_P. Equation (5) gives

$$E(\mathcal{LR}) = \frac{L-1}{8}\left(\frac{(f_1+f_2)(f_T+1)}{2} + f_T\right) + \frac{L+7}{8}. \tag{23}$$

In Fig. 7, the expectation of $\mathcal{LR}$ is shown as a function of founder inbreeding f_T of the true half sibling pedigree, for H_P stating sibling pedigree with founder inbreeding $\boldsymbol{f}_P = 0$ and 0.2 (assuming $f_1 = f_2$), and $L = 2$, 10 and 20 alleles at a locus. For increasing values of f_T, $E(\mathcal{LR})$ increases, for all values of $\boldsymbol{f}_P$, and the evidence in favor of a sibling relationship becomes stronger.

Consider next the situation when $f_1 = f_2 = 0$. H_P then assumes a sibling relationship without inbred founders. Figure 8 shows $E(\mathcal{LR})$ (dashed line) and LR computations from 1000 sets of simulated data, as a function of f_T. The solid line gives the mean value of the simulated LR. The expected LR increases slightly as founder inbreeding increases. For Fig. 8a this seems to fit well with the mean values of the LRs from simulated data. These simulation assumes 13 loci, each of 3 alleles with allele frequencies 0.4, 0.3 and 0.3. In Fig. 8b, on the other hand, there is a substantial difference between $E(\mathcal{LR})$ and the mean of the simulated LRs. These simulations use 13 CODIS markers with allele frequencies ranging from 0.0003 to 0.5378 (allele frequencies are available as a part of the R library InbredLR, see the "R implementation" section). Alleles with low frequencies will more seldom be present in the simulations. The expected LR only depends on the number of alleles at a locus, but because of the rare alleles, the simulations give in practice a lower number of alleles at these loci. The simulations in Fig. 8c use the same markers, but with uniform allele frequencies for alleles at a locus. The expectation of the LR is independent of the allele frequencies and is therefore not changed, but now the mean of the simulated LRs is closer to the expected value. Even though $E(\mathcal{LR})$ is independent of the allele frequencies, the

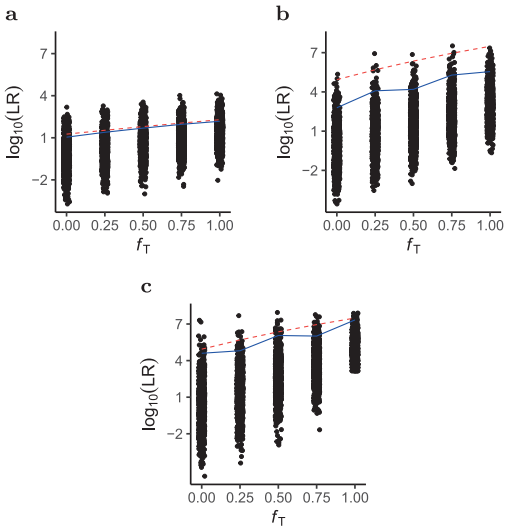


Fig. 8 Simulations of LR for the case in Fig. 6. Each figure shows 1000 LR values, for five values of f_T, each calculated from a simulation of a complete set of genotypes for 13 loci. Solid lines show mean of simulated LR. Dashed lines show $E(\mathcal{LR})$. **a** Loci with 3 alleles with frequencies 0.4, 0.3 and 0.3. **b** CODIS loci with realistic allele frequencies. **c** CODIS loci with uniform allele frequencies

variance is not, and small allele frequencies increase the variance.

Finally, we offer an approximation of the inflation in the expected LR due to background inbreeding. For simplicity, we assume $f_1 = f_2 = 0$ so that H_P states a normal sibling relationship. From Eq. 23 the expected LR is $\mu_0 = \frac{1}{8}(L+7)$ if $f_T = 0$. On the other hand, if $f_T > 0$, the expected contribution to the LR is $\mu_f = \frac{1}{8}(L-1)f_T$. For $L \geq 5$ we have $\frac{\mu_f}{\mu_0} \geq \frac{1}{3}f_T$, and it follows that

$$(\mu_0 + \mu_f)^N = [(1 + \tfrac{\mu_f}{\mu_0})\mu_0]^N \geq (1 + \tfrac{1}{3}fN)\mu_0^N.$$

A background inbreeding level of f_T will inflate the expected LR by at least $\frac{1}{3}f_T N$. For example, with $N = 20$ and $f_T = 0.05$, the inflation rate is greater than 33%.

Paternity case with inbreeding

Consider a paternity case with hypotheses as shown in Fig. 9. The alleged father is indeed the true father and has inbreeding coefficient f. We will analyze the consequences

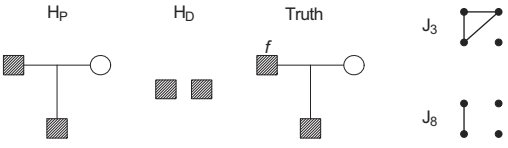


Fig. 9 The hypotheses in a paternity case with inbreeding. To the far right are the Jacquard states with nonzero probability in the true relationship

of ignoring the inbreeding in H_P. The hypotheses are parameterized in the following way:

$$\begin{aligned} H_P: \quad \boldsymbol{\Delta} &= \boldsymbol{\Delta}_P = (0,0,0,0,0,0,0,1,0) \\ H_D: \quad \boldsymbol{\Delta} &= \boldsymbol{\Delta}_D = (0,0,0,0,0,0,0,0,1) \\ Truth: \quad \boldsymbol{\Delta} &= \boldsymbol{\Delta}_T(f_T) \\ &= (0,0,f_T,0,0,0,0,1-f_T,0). \end{aligned}$$

The expression for the expected LR simplifies considerably since most elements of $\boldsymbol{\Delta}_P$ and $\boldsymbol{\Delta}_T(f_T)$ are zero. Equation (15) gives

$$E(\mathcal{LR}) = \frac{L+1}{2}f_T + \frac{L+3}{4}(1-f_T),$$

and we see that $E(\mathcal{LR})$ increases linearly from $(L+3)/4$ to $(L+1)/2$ as f_T goes from 0 to 1.

Consider next the variance. For brevity, we define

$$h(i,j,k) = E(\mathcal{LR}_{J_i,J_k} \cdot \mathcal{LR}_{J_j,J_k}). \tag{24}$$

Note that $h(i,j,k)$ is invariant under permutations of i,j,k. Equation (16) gives

$$\begin{aligned} E(\mathcal{LR}^2) &= \Delta_3^T h(8,8,3) + \Delta_8^T h(8,8,8) \\ &= f_T h(8,8,3) + (1-f_T)h(8,8,8). \end{aligned}$$

Slooten and Egeland [1] derived the term not involving inbreeding, i.e.,

$$h(8,8,8) = \tfrac{5L+3}{8} + \tfrac{s_3-L}{16}.$$

To derive the remaining term we condition on the zygosity of the son. If he is homozygous a/a, the father must also be a/a (recall that we are conditioning on Jacquard state J_3). Conversely, if the son is heterozygous a/b, the father is equally likely to be a/a or b/b. This gives

$$\begin{aligned} h(8,8,3) &= \sum_a p_a^2 \frac{1}{p_a}\frac{1}{p_a} \\ &\quad + \sum_{a<b} 2p_a p_b \left(\frac{1}{2}\left(\frac{1}{2p_a}\right)^2 + \frac{1}{2}\left(\frac{1}{2p_b}\right)^2 \right) \\ &= L + \frac{1}{4}\sum_{a \neq b} \frac{p_b}{p_a^2} = \frac{3L+s_3}{4}. \end{aligned}$$

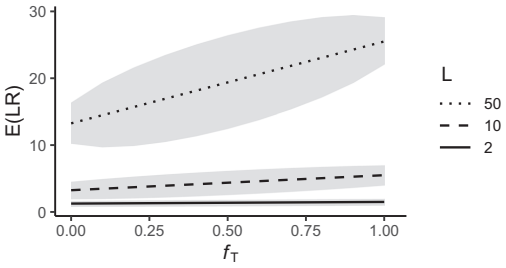


Fig. 10 $E(\mathcal{LR})$ as a function of f_T in the paternity case in Fig. 9, for a single marker with $L = 2$, 10 and 50 alleles. Shaded area shows one standard deviation below and above $E(\mathcal{LR})$, for uniform allele frequencies

In summary,

$$\begin{aligned}&\text{var}(\mathcal{LR}) = \\ &\frac{3L+s_3}{4}f_T + \left(\frac{5L+3}{8} + \frac{s_3-L}{16}\right)(1-f_T) \\ &-\left(\frac{L+1}{2}f_T + \frac{L+3}{4}(1-f_T)\right)^2. \end{aligned} \tag{25}$$

This is a concave function with respect to f_T. Figure 10 shows $E(\mathcal{LR})$ and one standard deviation on each side as a function of founder inbreeding f_T, for different number of alleles at a locus.

Discussion

In testing theory, the formulation of hypotheses is crucial. Kinship problems, as considered in this paper, are no exception. The convention of kinship testing is to compare two specific relationships using the LR. In most applications other than kinship problems, the hypotheses together span many, if not all, alternatives. For instance, a common example is testing of HWE against *all* possible deviations from HWE. In forensic genetics, H_P: "paternity" is typically tested only against H_D: "unrelated," not all other alternatives. For this reason, it becomes essential to study what happens when the truth is neither of these hypotheses.

A pairwise non-inbred relationship can be presented by a point in the IBD triangle (see Fig. 2), or in general by the Jacquard coefficients (see Fig. 1). We have presented two ways of expressing the hypotheses and the true relationship; (i) through the Jacquard coefficients, and (ii) background relatedness or founder inbreeding. These approaches let us investigate the LR for a continuous range of relationships and values of background relatedness. In both cases, the impact on the LR has been studied by deriving exact expressions for its mean and variance. In the latter case, the required formula follows rather directly by extending results in [1] and [2]. Explicit formulas for the expected LR has been derived for several sets of relationships. In the case of Jacquard coefficients, the explicit formulas are complicated to derive, and they depend on allele frequencies. An exact expression is given also for the variance. However, as the variance depends on allele frequencies, simple closed formulas can only be derived in special cases. For general applications we rely instead on the exact numerical implementation freely available in the R library InbredLR accompanying this paper.

Equipped with the results of this paper, we can address the following question when presented with a standard LR comparing two completely specified hypotheses H_P and H_D: What if the true relationship between the individuals is not as stated by H_P? Or this slightly different question: What if the true relationship is restricted to some particular region of the IBD triangle. Obviously, the LR can be re-evaluated to reflect the new specifications. However, the exact expressions for expectation and variance of the LR can in some cases directly allow for statements valid for a continuous range of alternatives. For instance, regions obtained by varying founder inbreeding have been displayed in Fig. 2. Assume a LR has been reported in a paternity case and that inbreeding in the father has been ignored. It is then useful to know that accounting for inbreeding would imply increase in the expected LR. This finding could be essential as there may not be data available to estimate the inbreeding coefficient for the father. Hence, exact LR calculation is not feasible.

Because the definition of "common ancestor" sometimes differs, there is a slight difference in the definition of IBD in the literature. The paper [16] gives three definitions of IBD: ancient IBD, recent IBD, and familial IBD. Our definition of IBD goes in the category of familial IBD, where "common ancestor" is restricted to a given pedigree.

The conventional approach to background relatedness in forensics is the so called theta (θ) correction [12]. Typical values are $\theta \in (0.01, 0.03)$. The θ parameter applies on a population level. The genotype probabilities of all founders in the pedigree are modified compared with what HWE would give. Our approach does not model relatedness between founders, but offers a richer model of inbreeding, since individual inbreeding coefficients can be specified for each founder.

Several authors (see, e.g., [2] and the references therein) have discussed reporting the logarithm of the LR rather than the LR. Nice expressions like the ones presented for the expectation and the variance are then no longer available. In most cases, the LR is reported on the original scale. In some circumstances, as for paternity cases, the LR may be 0, and

then, the logarithm is not defined. Many papers including [17] study the distribution of $\mathcal{Z} = \log(\mathcal{LR})$ by simulation. Equipped with the exact expressions of this paper, $\mathcal{Z}$ could be analyzed without resorting to simulation, since the mean and variance of $\mathcal{Z}$ can be derived from the counterparts for the LR. However, if some allele frequencies are close to 0, $\mathcal{Z}$ is not well approximated by a normal distribution for a realistic number of markers. The reason for this is the large variance when allele frequencies are small. For instance, (25) shows an example where the expression for the variance include terms of the form $1/p_a$ and these become large whenever the allele frequency p_a is small. A similar problem related to small allele frequencies is discussed in the result section. This demonstrates that the center of the $\log(\mathcal{LR})$ distribution, calculated from the expectation of $\mathcal{LR}$, can be inaccurate. However, this criticism applies to the use of $\mathcal{LR}$ instead of $\log(\mathcal{LR})$ in general, and not specifically to the expectations. We maintain that results like the ones presented for the expectation and variance have considerable theoretical interest, but should be used with caution in practice.

This paper has mainly addressed the likelihood ratio and its properties. The exclusion probability (EP), the probability that genotypes will be incompatible with a claimed relationship, is also an important statistic. The impact of founder inbreeding on EP is discussed in [3].

Figure 4 illustrates a case where the true inbred relationship is not known, and Fig. 5 shows the corresponding expected LR for a single marker. Increasing the number of markers will, in this paternity case, increase the inflation of the expected LR. This means that adding more markers to the LR computation will not solve the problem. In general, with a sufficient number of markers, the Jacquard, IBD, or inbreeding coefficients can be estimated accurately, and the true relationship detected. If such additional marker data is not available, the impact of inbreeding can be studied as exemplified by a paternity case with unknown inbreeding earlier in the discussion and as illustrated in, e.g., Fig. 5. As addressed in the "Introduction" section, different scenarios can be investigated and LR results can be evaluated in light of the analyses of these scenarios.

The present paper does not consider linked markers. For independent loci, the inbreeding coefficients contain sufficient information to compute the Jacquard coefficients needed in our formulas for LR. While a similar approach is conceivable also for linked markers, this would involve multi-locus coefficients, which is outside the scope of this work.

Funding Open Access funding provided by Norwegian University of Life Sciences.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval None required as no data from humans are used.

Appendix: Expectation and variance of LR

Below we derive the expressions for the expectation and variance of $\mathcal{LR}$ in the general pairwise case. Let J_i denote Jacquard state i and Δ_i^P and Δ_i^T the probabilities of J_i according to the relationship stated by H_P and the true relationship respectively. $\mathcal{LR}_{\boldsymbol{\Delta}_P,\boldsymbol{\Delta}_T}$ is then defined as the likelihood ratio comparing H_P:$\boldsymbol{\Delta}_P$ with H_D:$\boldsymbol{\Delta}_D$ when the marker data comes from the relationship $\boldsymbol{\Delta}_T$. Similarly, $\mathcal{LR}_{J_i,J_j}$ denotes the likelihood ratio comparing Jacquard state J_i with unrelated, i.e., J_9 when the marker data are generated by J_j.

Equation (15) follows by combining (1), (14), and (4)

$$\begin{aligned} E(\mathcal{LR}) &= \sum_G \left(\sum_{j=1}^{9} \Delta_j^T P(G \mid J_j) \sum_{i=1}^{9} \Delta_j^P \frac{P(G|J_i)}{P(G|J_9)} \right) \\ &= \sum_{i=1}^{9} \sum_{j=1}^{9} \Delta_i^P \Delta_j^T \left(\sum_G \frac{P(G|J_i)}{P(G|J_9)} P(G \mid J_j) \right) \\ &= \sum_{i=1}^{9} \sum_{j=1}^{9} \Delta_i^P \Delta_j^T E\left(\mathcal{LR}_{J_i,J_j}\right) \\ &= \boldsymbol{\Delta}_P B_9 \boldsymbol{\Delta}_T^t. \end{aligned} \tag{26}$$

In the case of no inbreeding, i.e., $\Delta_1 = \cdots = \Delta_6 = 0$, the above expression reduces to (5). The part of the 9×9 matrix B_9 corresponding to (J_7, J_8, J_9) coincides with the matrix given in Eq. 6. Since $E(\mathcal{LR}_{J_i,J_j}) = E(\mathcal{LR}_{J_j,J_i})$, B_9 is symmetric. The elements of B_9 are found by direct calculation. For instance, entry (1, 1) equals

$$E(\mathcal{LR}_{J_1,J_1}) = \sum_a \frac{p_a}{p_a^4} p_a = \sum_a \frac{1}{p_a^2}.$$

Since the expectation has been calculated, to derive the variance it remains only to find

$$
\begin{aligned}
&E(\mathcal{LR}^2)\\
&= \sum_G \left(\sum_{k=1}^{9} \Delta_k^T P(G \mid J_k) \sum_{i=1}^{9} \Delta_i^P \frac{P(G \mid J_i)}{P(G \mid J_9)} \right.\\
&\quad \left. \sum_{j=1}^{9} \Delta_j^P \frac{P(G \mid J_j)}{P(G \mid J_9)} \right)\\
&= \sum_{i=1}^{9} \sum_{j=1}^{9} \sum_{k=1}^{9} \Delta_i^P \Delta_j^P \Delta_k^T E\left(\mathcal{LR}_{J_j,J_i} \mathcal{LR}_{J_k,J_i}\right)\\
&= \sum_{i=1}^{9} \Delta_i^P \boldsymbol{\Delta}_P B_i \boldsymbol{\Delta}_T^t.
\end{aligned}
$$

The matrices $B_1, \ldots, B_9$ are symmetric 9×9 matrices. The simplest of these matrices is B_9, given in Table 4. In general, B_i consists of the elements $\{E(\mathcal{LR}_{J_j,J_i} \mathcal{LR}_{J_k,J_i})\}_{j,k=1,\ldots,9}$. The values for $i, j, k = 7, 8, 9$ have been provided in the "Review of previous results" section. Entry (j, k) of B_i is

$$\sum_G \frac{P(G \mid J_j)}{P(G \mid J_9)} \frac{P(G \mid J_k)}{P(G \mid J_9)} P(G \mid J_i). \tag{27}$$

All matrices can in principle be found from the above expression, but exact calculations by hand become unpractical and exact numerical calculation is more reasonable.

References

1. Slooten KJ, Egeland T (2014) Exclusion probabilities and likelihood ratios with applications to kinship problems. Int J Legal Med 128(3):415–425
2. Egeland T, Slooten KJ (2016) The likelihood ratio as a random variable for linked markers in kinship analysis. Int J Legal Med 130(6):1445–1456
3. Vigeland MD, Egeland T (2019) Handling founder inbreeding in forensic kinship analysis. Forensic Science International: Genetics Supplement Series. https://doi.org/10.1016/j.fsigss.2019.10.175
4. Vigeland MD (2020) Relatedness coefficients in pedigrees with inbred founders. J Math Biol 81:185–207. https://doi.org/10.1007/s00285-020-01505-x
5. Thompson EA (2000) Statistical inference from genetic data on pedigrees. IMS
6. Wright S (1922) Coefficients of inbreeding and relationship. The American Naturalist 56:330–338
7. Buckleton J, Curran J, Goudet J, Taylor D, Thiery A, Weir BS (2016) Population-specific FST values for forensic STR markers: a worldwide survey. Forensic Science International: Genetics 23:91–100. https://doi.org/10.1016/j.fsigen.2016.03.004
8. Pemberton TJ, Rosenberg NA (2014) Population-genetic influences on genomic estimates of the inbreeding coefficient: a global perspective. Human Heredity 77(1-4):37–48. https://doi.org/10.1159/000362878
9. Jacquard A (1972) Genetic information given by a relative. Biometrics 28(4):1101–1114
10. Cotterman CW (1940) A calculus for statistico-genetics. Dissertation, The Ohio State University
11. Thompson EA (1976) A restriction on the space of genetic relationships. Ann Hum Genet 40(2):201–204
12. Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96(1–2):3–12
13. Buckleton J, Triggs CM, Walsh SJ (2005) Forensic DNA evidence interpretation. CRC Press, Florida
14. Gorlin JB, Polesky HF (2000) The use and abuse of the full-sibling and half-sibling indices. Transfusion 40(9):1148–1149
15. Mayor LR, Balding DJ (2006) Discrimination of half-siblings when maternal genotypes are known. Forensic Sci Int 159(2–3):141–147
16. Browning BL, Browning SR (2011) A fast, powerful method for detecting identity by descent. The American Journal of Human Genetics 88(2):173–182
17. Nothnagel M, Schmidtke J, Krawczak M (2010) Potentials and limits of pairwise kinship analysis using autosomal short tandem repeat loci. Int J Legal Med 124(3):205–215

Paper III

Strategies for pairwise searches in forensic kinship analysis

Hilde Kjelgaard Brustad[*], Margherita Colucci[†], Mark A. Jobling[†], Nuala A. Sheehan[‡], Thore Egeland[*]

April 8, 2021

Corresponding author

Hilde Kjelgaard Brustad, hilde.brustad@nmbu.no

Highlights

- Search for general pairwise relationships.
- Finding optimal thresholds for likelihood ratios.
- Correcting for multiple testing in relationship searching.
- Bayesian version of blind search.

[*]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1432 Aas, Norway

[†]Department of Genetics & Genome Biology, University of Leicester, University Road, Leicester LE1 7RH UK

[‡]Department of Health Sciences, University of Leicester, University Road, Leicester LE1 7RH UK

Abstract

Testing kinship between pairs of individuals is central to a wide range of applications. We focus on cases where many tests are done jointly. Typical examples include cases where DNA profiles are available from a burial site, a plane crash or a database of convicted offenders. The task is to determine the relationships between DNA profiles or individuals. Our approach generalises previous methods and implementations in several respects. We model general, possibly inbred, pairwise relationships which is important for non-human applications and in archaeological studies of ancient inbred populations. Furthermore, we do not restrict attention to autosomal markers. Some cases, such as distinguishing between maternal and paternal half siblings, can be solved using X-chromosomal markers. When many tests are done, the risk of errors increases. We address this problem by building on the theory of multiple testing and show how optimal thresholds for tests can be determined. We point out that the likelihood ratios in a blind search may be dependent so multiple testing methods and interpretation need to account for this. In addition, we show how a Bayesian approach can be helpful. Our examples, using simulated and real data, demonstrate the practical importance of the methods and implementation is based on freely available software.

Keywords

Kinship testing; Blind search; LR thresholds; Inbred relationships; X chromosomal markers

1 Introduction

Inferring the relationship between pairs of individuals is central to many forensic applications. Examples include mass fatality incidents, which can be the result of accidental catastrophes like air crashes with a list of known victims [1] or shipwrecks without passenger lists [2, 3]. Other applications are natural disasters like tsunamis, where the number of victims is unknown [4] and terrorism-related events [5]. The aim is to link DNA samples from the scene to putative victims (e.g. individuals reported missing since the event) and is known as disaster victim identification (DVI). There are various other important applications like searching for relationships among individuals in mass graves of archaeological relevance [6, 7, 8]. We may also check databases collected to estimate population statistics like allele frequencies. Duplicates and close relatives should be excluded prior to the statistical analysis or estimates of allele frequencies could be biased [9].

As these cases involve unidentified DNA samples, a first step in the investigation is to screen the data for related samples. This initial step is referred to as a blind search [10]. It is helpful to first position the topics that we are addressing in the wider context of database searching. Assume that there is a

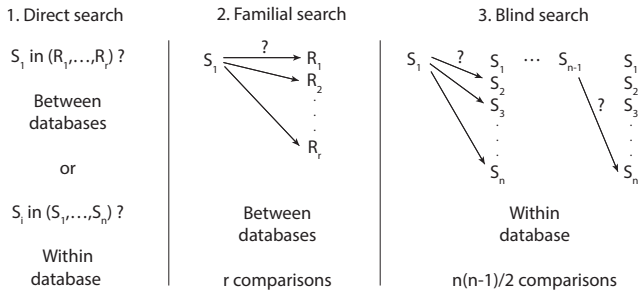


Figure 1: Different database searches. 1. Direct search: Search for direct matches between or within databases. 2. Familial search: Search for related individuals between databases. 3. Blind search: search for related individuals within databases.

case database of DNA profiles. This could comprise profiles obtained from a crime scene, a disaster site or a burial site. In addition, there may be a reference database of DNA profiles like a national database of convicted offenders. There are various searches that can be performed to detect pairwise relationships as illustrated in Figure 1:

1. Search for duplicates, i.e., direct search, performed within or between the databases. If this is done within a database, the objective is to merge identical samples. A search between databases corresponds to the widely discussed database search problem [11].

2. Familial searching involves searching between databases [12]. A selected DNA profile is compared to the profiles of a database with the aim of detecting close kin relationships, such as parent-offspring or sibling rather than a direct match.

3. Blind search. This is a search within a database and is the topic of this paper to be discussed below.

In a blind search, comparisons are performed among all pairs of DNA samples. A likelihood ratio (LR), comparing the relationship specified by H_1 to the one specified by H_0, is computed for each pair. The LRs summarise the statistical DNA evidence. For pre-specified threshold values t_0 and t_1, small values of $LR < t_0$ provide little evidence to reject H_0, and are often interpreted as supporting H_0, while large values of $LR > t_1$ favour H_1. A blind search typically involves a large number of comparisons. If there are n profiles in the database, the number of comparisons is $n(n-1)/2$, e.g. 4950 comparisons for 100 profiles.

The implications of this high number of pairwise comparisons in a blind search are of key concern in this paper. Also, it is not obvious how the thresholds t_0 and t_1 should be specified. Conventional thresholds used in paternity testing, for example, may not apply. The false positive rate $\text{FPR} = \text{P}(\text{LR} > t_1 \mid \text{H}_0)$ and false negative rate $\text{FNR} = \text{P}(\text{LR} < t_0 \mid \text{H}_1)$ should both be close to 0. Even if these error rates are small for each comparison, the probability that errors occur when many comparisons are done may be considerable. Determination of thresholds and optimisation of search strategies have been discussed in connection with database searches and familial searching [13]. The classical statistical theory of multiple testing [14] is also relevant.

Current implementations of blind search are limited to fairly simple outbred pedigree structures connecting the two individuals of interest. For example, Familias [10], a freely available kinship software package, accommodates parent offspring (PO), sibling (S), half sibling (H), first cousin (FC) and second cousin (SC) [15, 16]. We model general pairwise relationship, possibly with inbreeding, using the Jacquard coefficients [17]. By including X-chromosomal markers, some additional relationships can be addressed. For instance, paternal and maternal half sisters can be distinguished.

Prior, non-DNA, information can sometimes be important. For instance, two individuals of the same age cannot possibly constitute a parent-offspring pair even if the DNA profiles suggest otherwise. To formally include prior information, we require a Bayesian approach. In the Bayesian framework we start out with a set of prior probabilities, reflecting our belief in the hypotheses, before considering any genetic data. Our belief in each hypothesis is then updated by incorporating the DNA information. Informative priors can contribute additional information to the genetic data and this will be reflected in the posterior probabilities. A more general prior distribution for pedigrees has been discussed elsewhere [18].

Our paper is structured as follows. We first review the parametric representation of relationships and the corresponding parametric likelihood and likelihood ratio, for both autosomal and X-chromosomal markers. A review of the Bayesian approach to kinship testing is given, before we return to the likelihood ratio and its properties. These properties are then incorporated when presenting the theory for evaluating the performance of a blind search. We then introduce the data used in the results section and give a brief description of our implementation. We provide several examples and conclude with a discussion of the challenges and the advantages of the work we present.

2 Review of previous results

2.1 Relatedness coefficients

Two homologous alleles are *identical by descent* (IBD) if they are identical by state (IBS) and inherited from a common ancestor. IBD is therefore defined with reference to a specified pedigree. The idea is that closely related individuals

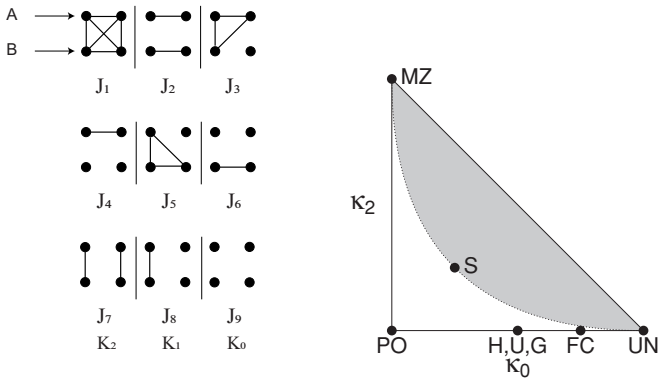


Figure 2: Left: Jacquard states $J_1, \ldots, J_9$. Dots denote alleles, and lines connect IBD alleles. Right: IBD triangle, with location of some common relationships. Abbreviations: MZ - monozygotic twins, PO - parent offspring, S - full siblings, H - half siblings, U - avuncular, G - grandparent grandchild, FC - first cousins, UN - unrelated.

share more of their genetic material IBD than more distantly related individuals.

The simplest measure of pairwise relationships is the kinship coefficient, φ, defined as the probability that a random allele at a locus from one individual is IBD to a random allele at the same locus from another individual. This is the same as the inbreeding coefficient f of a child of these two individuals [19].

The Jacquard coefficients [17] provide a description of general pairwise relationships. The four alleles of two individuals are in one of the nine Jacquard states J_i for $i = 1, \ldots, 9$ (see left panel of Figure 2). The probability that the alleles at a locus are in the different Jacquard states are given by the Jacquard coefficients, $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_9)$, where $\Delta_i = \mathrm{P}(J_i)$. The coefficients sum to one.

The first six Jacquard states model inbreeding in one or both of the individuals. The only possible IBD states for two outbred individuals are J_9, J_8 and J_7, referred to as the IBD states K_0, K_1 and K_2, respectively. Thus, for two outbred individuals, the Jacquard coefficients reduce to the IBD coefficients [20], $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$, where $\kappa_i = \mathrm{P}(K_i)$. Since $\sum_{i=0}^{2} \kappa_i = 1$, the coefficients can be visualised in the IBD triangle, with coordinates (κ_0, κ_2). Figure 2 shows the IBD triangle with the location of some common pedigree relationships. Thompson [21] showed that the coefficients satisfy the inequality $\kappa_1^2 \geq 4\kappa_0\kappa_2$, which creates an inadmissible region, shown in grey in Figure 2. This means that it is not possible to construct a pedigree connecting two individuals with IBD coefficients in the inadmissible region.

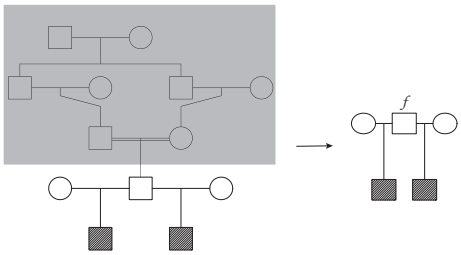


Figure 3: Figure showing the concept of founder inbreeding, as described in Section 2.1.1. The shaded part showing the first cousin relationship, is modelled by the inbreeding coefficient f.

2.1.1 Relatedness coefficients and founder inbreeding

By assigning a coefficient of inbreeding to one or more of the founders of a pedigree, background relatedness can be modelled [22]. Inbreeding of a pedigree founder (or several founders) affects the genetic relationship between other members of the pedigree [23], but does not necessarily make the pedigree members of interest inbred. For example, if it is suspected that two individuals are paternal half-siblings and the paternal grandparents are first cousins, as depicted in Figure 3, the common father has an inbreeding coefficient $f = 1/16$. The IBD coefficients for these half siblings are given by $\boldsymbol{\kappa} = (0.469, 0.531, 0)$ in contrast with $\boldsymbol{\kappa} = (0.5, 0.5, 0)$ for the non-inbred setting with $f = 0$. It can be shown that there is some finite pedigree with founder inbreeding that corresponds to each admissible point in the IBD triangle [24].

2.2 The likelihood function

Our data comprise pairs of DNA profiles, genotyped at m unlinked loci. For a single pair of individuals, A and B, let $G_j = (g_{A,j}, g_{B,j})$ denote their respective genotypes at locus j for $j = 1, \ldots, m$. The likelihood of $\boldsymbol{\Delta}$, i.e., the probability of observing the data $G = (G_1, \ldots, G_m)$ assuming $\boldsymbol{\Delta}$ to be true, is

$$\mathrm{L}(\boldsymbol{\Delta}) = \prod_{j=1}^{m} \sum_{i=1}^{9} \Delta_i \mathrm{P}(G_j \mid \mathrm{J}_i). \tag{1}$$

The probabilities $\mathrm{P}(G_j \mid \mathrm{J}_i)$ are given in Table 9 in the Appendix. For outbred individuals, the likelihood of $\boldsymbol{\kappa}$ is

$$\mathrm{L}(\boldsymbol{\kappa}) = \prod_{j=1}^{m} \sum_{i=0}^{2} \kappa_i \mathrm{P}(G_j \mid \mathrm{K}_i), \tag{2}$$

	$\mathrm{P}(G \mid \mathrm{K}_0)$	$\mathrm{P}(G \mid \mathrm{K}_1)$	$\mathrm{P}(G \mid \mathrm{K}_2)$	$\mathrm{L}(\boldsymbol{\kappa})$
$G_1 = (ab, ac)$	0.06	0.03	0	$0.06 \cdot \kappa_0 + 0.03 \cdot \kappa_1$
$G_2 = (bc, bb)$	0.011	0.004	0	$0.011 \cdot \kappa_0 + 0.004 \cdot \kappa_1$
$G_3 = (aa, bc)$	0.03	0	0	$0.03 \cdot \kappa_0$

Table 1: Likelihood of $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$, when observing genotypes for two individuals, for three unlinked loci, as described in the example of Section 2.3.

where the probabilities $P(G_j \mid \mathrm{K}_i)$ for $i = 0, 1, 2$ correspond to the last three columns of Table 9.

2.3 Parametric representation of the likelihood ratio

The likelihood ratio (LR) quantifies how much more likely it is that a set of genetic data is explained by one hypothesis H_1 than by another hypothesis H_0. In our applications, each hypothesis states a pairwise relationship, expressed by a set of relatedness coefficients $\boldsymbol{\Delta}$ (or $\boldsymbol{\kappa}$ for outbred relationships). The LR that compares (1) for two sets of coefficients $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_0$ is

$$\mathrm{LR}(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_0) = \frac{\mathrm{P}(G \mid \mathrm{H}_1)}{\mathrm{P}(G \mid \mathrm{H}_0)} = \frac{\mathrm{L}(\boldsymbol{\Delta}_1)}{\mathrm{L}(\boldsymbol{\Delta}_0)}. \tag{3}$$

The hypotheses H_1 and H_0 are not necessarily exhaustive, meaning that there may be other hypotheses that better explain the data.

Example: The purpose of this example is merely to illustrate how LRs can be easily computed for different sets of IBD coefficients using the representation in (3).

Consider two individuals genotyped at three loci. Each locus has three alleles a, b and c, with population frequencies 0.5, 0.3 and 0.2, respectively. The genotypes at each locus are given in the first column of Table 1. The likelihood of $\boldsymbol{\kappa}$ for each locus are given in the last column.

When comparing siblings, $\boldsymbol{\kappa}_1 = (0.25, 0.5, 0.25)$ against unrelated $\boldsymbol{\kappa}_0 = (1, 0, 0)$, the LR becomes

$$\begin{aligned} \mathrm{LR}(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_0) = & \frac{(0.06 \cdot 0.25 + 0.03 \cdot 0.5)}{0.06} \\ & \cdot \frac{(0.011 \cdot 0.25 + 0.004 \cdot 0.5)}{0.011} \\ & \cdot \frac{(0.03 \cdot 0.25)}{0.03} = 0.054. \end{aligned} \tag{4}$$

Similarly, we find the LR for half siblings (or avuncular or grandparent grandchild), $\boldsymbol{\kappa}_1 = (0.5, 0.5, 0)$, against unrelated, $\boldsymbol{\kappa}_0 = (1, 0, 0)$, to be 0.256. The probabilities in the middle three columns of Table 1 are independent of the tested relationships.

2.4 Properties of the LR

For specified thresholds $t_0 < t_1$, an $\text{LR} < t_0$ essentially supports H_0, while an $\text{LR} \geq t_1$ favours H_1. More data may be required to conclude when $t_0 \leq \text{LR} < t_1$ [25]. For simplicity, we will assume $t_0 = t_1 = t$, so that a conclusion can always be drawn.

When $\text{LR} \geq t$, but H_0 is true, we have a *false positive* (FP). If $\text{LR} \geq t$ and H_1 is true, we have a *true positive* (TR). We define the *false positive rate* (FPR) and the *true positive rate* (TPR) as

$$\text{FPR} = \text{P}(\text{LR} \geq t \mid \text{H}_0), \quad \text{TPR} = \text{P}(\text{LR} \geq t \mid \text{H}_1). \tag{5}$$

The TPR measures the ability to detect the relationship, while the FPR is the probability of falsely declaring a relationship. The relationship between FPR and TPR is often visualised by a receiver operating characteristic (ROC) curve [26]. Figure 4 in Section 3.3 illustrates the concept of a ROC curve.

2.5 The Bayesian approach to kinship testing

A frequentist approach to evaluating kinship is based on the LR reflecting the probabilities of the data we have observed under two specified hypotheses. An alternative approach is provided by a Bayesian framework.

Instead of just testing one hypothesis H_1 against H_0, we consider a set of hypotheses H_i, $i = 1, \ldots, k$, each against H_0. With some prior belief in each hypothesis $\pi_0, \ldots, \pi_k$, Bayes' theorem expresses the posterior probability of each hypothesis as

$$\text{P}(\text{H}_i \mid \text{data}) = \frac{\text{LR}_i \pi_i}{\sum_{j=0}^{k} \text{LR}_j \pi_j}, \text{ for } i = 0, \ldots, k\,, \tag{6}$$

where LR_i is the likelihood ratio when H_i is compared against H_0 [10]. In fact, the denominator in the LR cancels out, so (6) actually compares the likelihood of each hypothesis against all the other hypotheses jointly.

Just as for LRs, we cannot infer anything about the true relationship between the individuals as this might not be one of the hypotheses considered. For a flat prior, the posterior probabilities do not add any information to the data and are simply scaling the likelihoods (or LRs) relative to each. More informative priors, on the other hand, can contribute additional information and this will be reflected in the posterior probabilities. For example, the three relationships half-sibling (H), avuncular (U) and grandparental (G) all have the same IBD coefficients and identical likelihoods. They are hence indistinguishable in the traditional frequentist setting and in a Bayesian setting using flat priors. Age information can easily be incorporated into the Bayesian approach and may yield different posterior probabilities.

3 Methods for blind search

3.1 The likelihood ratio for X-chromosomal markers

X chromosomal markers are increasingly used in forensic applications to supplement or replace autosomal markers for some cases of practical importance [27]. One such example is shown in Figure 8. The females B and C are *paternal* half sibs while C and D are *maternal* half sibs. The distinction between maternal and paternal is captured by X-chromosomal markers but not by autosomal markers. The paternal half sibs share an allele IBD inherited from their father. The Jacquard coefficients and the likelihood calculation can be modified to apply for independent X-chromosomal markers (details omitted). Obviously, the sex of the individuals in the pair matters. As an example note that there are only two possibilities, or two states, for a pair of males: either they share an allele IBD or they do not.

Since the number of unlinked markers on the X chromosome is limited, linkage and linkage disequilibrium become an issue [28]. We will ignore such dependence in Example 5.5. However, relevant findings that take dependence into account can be checked using the freely available software FamLinkX [29].

3.2 Estimation of FPR and TPR

The true positive and false positive rates are determined by the hypotheses considered, number of loci, properties of each locus and the LR threshold. These rates can be calculated numerically using the algorithm described in [30]. However, this method only works for a small number of markers, say up to 10. In practice, we therefore resort to simulation. We denote estimates of FPR and TPR by $\widehat{FPR}$ and $\widehat{TPR}$, respectively.

Typically TPR is close to 1 and FPR close to 0. These values are generally poorly estimated from direct Monte Carlo simulation. For instance, assume FPR = 0.00001. Then 1 of 100000 simulations is expected to give a false positive. The conventional number of simulations in the range 100-10000 is therefore likely to return an estimate of 0. Kruijver [30] describes several methods for estimating small probabilities in forensic applications. One of these is importance sampling, which we use to estimate FPR in the results section. Details about importance sampling are given in Appendix B.

3.3 Optimal LR threshold

Intuitively we seek a threshold for LR that minimises the number of errors. Several approaches for choosing the optimal threshold have been suggested and compared [31]. We will focus on ROC-based methods.

Figure 4 shows a general ROC curve. Each point on the curve corresponds to a threshold t, with corresponding values for FPR and TPR. The value t in the figure corresponds roughly to FPR = 0.4 and TPR = 0.6. The upper left corner corresponds to FPR = 0 and TPR = 1 and is therefore called the optimal

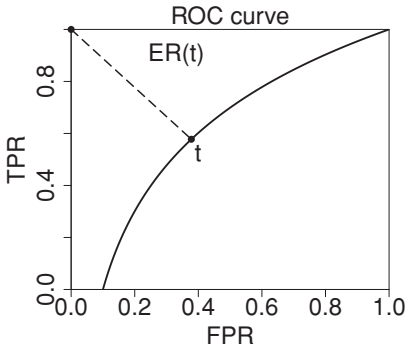


Figure 4: Figure showing the concept of a ROC curve with a corresponding threshold. The rate TPR is plotted as a function of FPR. Each point on the curve corresponds to an LR threshold t. The dashed line shows the Euclidean distance (unweighted) from the optimal point (0,1) to the ROC curve, given by (7).

point. Consequently, it is reasonable to choose a threshold that minimises the weighted Euclidean distance between the ROC curve and the point (0,1),

$$\mathrm{ER}(t) = \sqrt{(w\,\mathrm{FPR}(t))^2 + (1 - \mathrm{TPR}(t))^2}. \tag{7}$$

In our examples the weight $w = 1$. It may be that one of the errors, typically a false positive, is more important to avoid than the other, a false negative. The relative importance of errors can be modeled by using other values of w. Because we do not know the exact values of FPR and TPR, they are replaced by their estimates.

3.4 The problem of multiple testing in blind search

When doing a blind search among n DNA profiles, we compute one LR for each pair of DNA profiles, leaving us with a total of $N = n(n-1)/2$ LRs, $\mathrm{LR}_1, \ldots, \mathrm{LR}_N$. If the true hypothesis is known in each case, the result of the search (or any other multiple testing scenario) can be summarised as shown in Table 2. In practice, the truth will only be known for simulated data.

Assume that the only possibilities are the relationship stated by H_1 or the null hypothesis H_0, such that $N_0 + N_1 = N$. The number of type I errors or false positives is FP, while the number of false negatives is FN. Ideally, we want FP = 0 and FN = 0. However, this is not realistic. For a sufficiently large threshold t we will never reject H_0 and there will be no false positives, i.e. FP = 0. Similarly, there will be no false negatives, FN = 0 (which means

	Claim H_0	Claim H_1	Total
H_0 true	TN	FP	N_0
H_1 true	FN	TP	N_1
Total	W_0	W_1	N

Table 2: The statistics of a blind search summarised, as described in Section 3.4. Only W_0, the number of LRs below t, and W_1, the number of LRs above t, are observed. Adapted from [32].

TP $= N_1$), for a sufficiently small threshold. The challenge is to make both FP and FN acceptably small, or equivalently, make FP as small as possible and TP as close to N_1 as possible.

Even if the probability of a false positive is very small for a single pairwise comparison, the fact that there are so many tests in a blind search could lead to a substantial probability of at least one false positive. Approaches to analyze and control these false positives in a multiple testing setting have to be applied. The Family Wise Error Rate (FWER) [14] is often used for this purpose.

FWER is defined as the probability of getting at least one false positive out of N tests [32]. Let α denote the FWER. For N independent tests,

$$\alpha = \mathrm{P}(\mathrm{FP} \geq 1) = 1 - (1 - \mathrm{FPR})^N,$$

where the FPR, as defined in (5), is assumed to be the same for each test. As we illustrate in the results section, the pairwise tests in a blind search are not independent and so we use the Bonferroni bound

$$\alpha \leq N \cdot \mathrm{FPR} = \frac{n(n-1)}{2}\mathrm{FPR}.$$

Thus, to obtain an α below a given value, we choose a threshold so that

$$\mathrm{FPR}_\alpha \leq 2\alpha/(n(n-1)) \tag{8}$$

for a fixed sample size n. Figure 5 plots FPR_α as a function of α, for a blind search with 5, 10, 50, 100 and 200 individuals. The red vertical line is located at $\alpha = 0.05$. The aim is to find the threshold t that minimizes $\mathrm{ER}(t)$ given in (7), with the constraint that $\mathrm{FPR} \leq \mathrm{FPR}_\alpha$.

4 Data and implementation

4.1 Real data and simulations

The DNA profiles evaluated in Section 5.3 and 5.4 are from 65 individuals of Northern European origin (Germany) forming 8 pedigrees, with a variety of declared kinships up to 7th degree (considered as number of meioses between each person [33]). Most founders were not genotyped, and pedigree sizes ranged from 5 to 17, with an average of 9 members per family. Genotyping was done

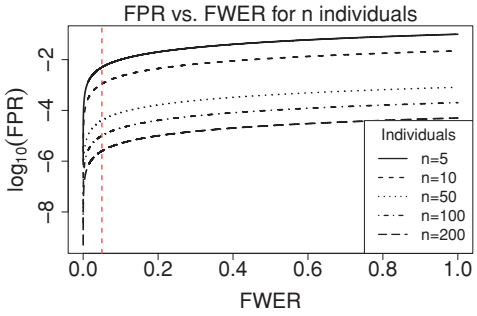


Figure 5: Highest acceptable value of FPR, as a function of α, given by (8) in Section 3.4. Plotted for blind search with 5, 10, 50, 100 and 200 individuals. The vertical line is located at $\alpha = 0.05$.

via massively parallel sequencing using the ForenSeq™ DNA Signature Prep kit (Verogen Inc., San Diego, CA, USA) and will be discussed in full elsewhere. Samples were collected with informed consent. For the purposes of the current study, we consider only the length-based genotypes from 27 autosomal STRs contained in Plex B of this kit. Allele frequencies are based on the European dataset in PopSTR (http://spsmart.cesga.es/popstr.php [34, 35]) and downloaded from the Familias website (https://familias.no/download).

The performance analysis in Section 5.2, that leads to the blind search in the following section, is based on simulated data assuming the same set of loci as the real data, i.e. the 27 autosomal STR loci described above. This set of STR markers is also used in the simulations for the last example in the results section.

To demonstrate the use of X-chromosomal markers, data are simulated based on 12 X-chromosomal STR markers included in the kit "Investigator Argus X-12", with frequencies taken from an Argentinian database [36]. This is the most widely used kit for forensic applications.

4.2 Implementation

The analyses in this paper are all performed using R code developed by the authors. The code is available from the first author on request. The main engine of the code is an implementation of the parametric version of the likelihood function. This efficiently computes likelihoods for a series of relationships and convert to LRs and posterior probabilities. The code builds on the R libraries *pedtools*, *ribd*, *forrel* and *pedmut* developed by Magnus Dehli Vigeland, freely available from CRAN.

4.3 Assumptions

All the equations above are based on (1) which is only valid under certain assumptions. Firstly, the population is assumed to be in Hardy-Weinberg Equilibrium (HWE) and in Linkage Equilibrium (LE). Secondly, mutations are ignored. Mutation rates are usually small, and the errors induced by ignoring them in likelihood calculations are typically negligible [37]. However, for a parent offspring (PO) relationship, i.e. $\boldsymbol{\kappa} = (0, 1, 0)$, the likelihood will be zero if the two samples have genotypes at any locus that are incompatible with this hypothesis, e.g. $g_A = aa$ and $g_B = bb$. For this special case, there is a simple formulation of the likelihood that incorporates mutation (see [10]). We ignore allele drop-ins and drop-outs, null alleles and genotyping errors.

5 Results

The first example shows that the LRs in a blind search are not independent. The second example demonstrates how to evaluate the performance of a blind search such as we present in the third example. We then carry out a blind search on X-chromosomal markers before showing how inbreeding can be analysed.

5.1 Correlation between LRs in a blind search

In this example, we show by simulation a case where the LRs of a blind search are correlated. Consider the pedigree in Figure 6 and the hypotheses H_1 stating a sibling relationship and H_0 unrelated. Let $LR_{1,3}$ denote the likelihood ratio when individual 1 is compared to 3 and define $LR_{2,3}$ analogously. We use 10 independent loci, each with 10 alleles and equal allele frequencies of 0.1. Note that the LRs are random variables. We simulate 1000 sets of DNA profiles for the three shaded individuals of the pedigree in Figure 6. The values of $LR_{1,3}$ and $LR_{2,3}$ are computed for each simulation. The results are shown in the scatter plot in Figure 6, the red line denoting a regression line.

The estimated correlation between the logarithmic values of $LR_{1,3}$ and $LR_{2,3}$ is 0.484. This shows that the LRs are not independent. In other words, the outcome of different comparisons cannot be interpreted independently if one individual is involved in several comparisons. We elaborate on the implications of this correlation in the discussion.

5.2 From FWER to choice of LR threshold

In Section 5.3, we carry out a blind search among 65 individuals, genotyped for a set of 27 STR loci. Here, we present the preliminary evaluation required to obtain optimal LR thresholds for that search.

The first step is to decide on an acceptable value of α. From this value of α we can decide on an upper limit of the FPR and then the corresponding optimal LR threshold. For a blind search of $n = 65$ individuals, with the requirement

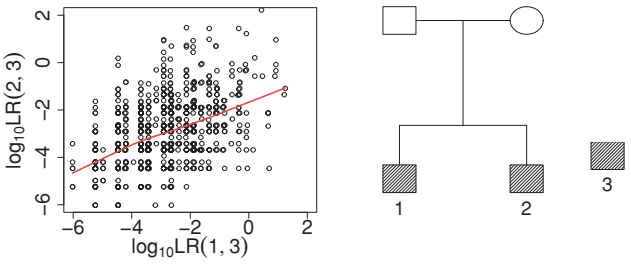


Figure 6: Figure corresponding to the correlation discussion in Section 5.1. Left: Scatter plot of LRs of simulated data for two siblings and an unrelated individual. Red line shows regression line. Right: Pedigree used for simulation of data, identifying the id labels 1, 2 and 3 in left panel.

that $\alpha \leq 0.05$, Equation (8) gives an upper limit for the false positive rate of $\text{FPR}_{0.05} = 2.404 \cdot 10^{-5}$.

The next step is to analyse how the FPR and TPR relate to each other for this particular set of markers. This depends on what hypotheses we test in the blind search. In the next example, we consider the hypotheses H_1: PO, H_2: S, H_3: H/U/G and H_4: FC, all against H_0: UN. We therefore consider these hypotheses when estimating FPR and TPR.

Figure 7 shows ROC curves for the different hypotheses. The values for $\widehat{\text{FPR}}$ and $\widehat{\text{TPR}}$ are estimated from simulated data, as described in Section 3.2. For H_1: PO, we only obtained estimates of FPR smaller than 10^{-7}, with a corresponding estimated TPR of 0.999 or higher. This shows that the LR comparing PO to UN is high when the true relationship is PO and low otherwise. Parent offspring and unrelated individuals are easily distinguished as expected and so we have omitted this curve from the graph.

The ROC curves show the estimated properties of a single computation of the LR, for the respective hypotheses, for this specific set of STR markers. The curves do not depend on the number of individuals in the blind search.

The last step in the performance analysis is to identify the optimal threshold, by minimizing $\text{ER}(t)$, with the constraint $\widehat{\text{FPR}} \leq \text{FPR}_\alpha$. The highest optimal thresholds for $\alpha = 0.05$ and $\alpha = 0.1$ are listed in Table 3.

5.3 Blind search with real data

In this example we do a blind search on the data described in Section 4.1. The data set contains 65 DNA profiles. A blind search among these profiles results in 2080 pairwise comparisons. We want to test the hypotheses H_1: PO, $\boldsymbol{\kappa}_1 = (0, 1, 0)$, H_2: S, $\boldsymbol{\kappa}_2 = (0.25, 0.5, 0.25)$, H_3: H/U/G, $\boldsymbol{\kappa}_3 = (0.5, 0.5, 0)$, H_4:

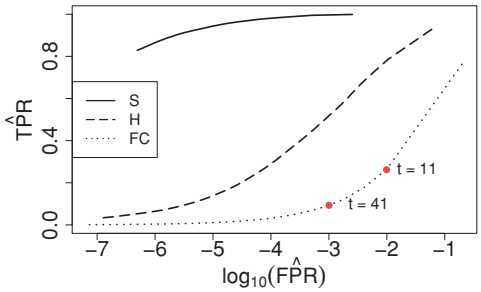


Figure 7: ROC curves for the analysis performed in Section 5.2. The hypothesis H_1 stating S, H and FC and H_0 unrelated, using 27 STR markers. ROC curves from simulated data. A threshold of 11 corresponds to an estimated false positive rate of about 0.01 and an estimated true positive rate of about 0.26 for FC.

$\alpha = 0.05 \Rightarrow \text{FPR}_\alpha = 2.404 \cdot 10^{-5}$

	t	$\widehat{\text{FPR}}$	$\widehat{\text{TPR}}$
PO	65531	$5.439 \cdot 10^{-8}$	1.000
S	771	$2.296 \cdot 10^{-5}$	0.961
H	2501	$2.365 \cdot 10^{-5}$	0.184
FC	421	$1.689 \cdot 10^{-5}$	0.014

$\alpha = 0.1 \Rightarrow \text{FPR}_\alpha = 4.808 \cdot 10^{-5}$

	t	$\widehat{\text{FPR}}$	$\widehat{\text{TPR}}$
PO	65531	$5.439 \cdot 10^{-8}$	1.000
S	311	$4.713 \cdot 10^{-5}$	0.972
H	1551	$4.770 \cdot 10^{-5}$	0.232
FC	251	$4.421 \cdot 10^{-5}$	0.022

Table 3: Optimal thresholds for different relationships, with corresponding $\widehat{\text{FPR}}$ and $\widehat{\text{TPR}}$, for the analysis performed in Section 5.2. For $\alpha = 0.05$ (left table) and $\alpha = 0.1$ (right table), for blind search with $n = 65$ individuals.

FC, $\boldsymbol{\kappa}_4 = (0.9375, 0.0625, 0)$, against H_0: UN, $\boldsymbol{\kappa}_0 = (1, 0, 0)$. In the previous section, we obtained optimal thresholds for blind searches with these hypotheses (Table 3). A step wise mutation model is implemented in the evaluation of PO.

Table 4 summarises the blind searches performed on the real data. This table is possible to construct because we know the true relationship for each pair from the pedigree information. In practice, only the sum of the last three rows (for each relationship) would be known.

For PO, we are left with a list of 47 hits. 43 of these are true PO, while 4 of the 47 hits are pairs of individuals with another relationship. 3 pairs with true PO relationship are not detected. By lowering the threshold, the remaining 3 pairs could have been detected.However, the probability of obtaining false positives increases by decreasing the threshold. For S, only one true sibling pair is not detected and there is only one false positive. However, the list of hits contains 66 pairs of individuals, 53 of these having another relationship.

	PO	S	H/U/G	FC
N_1	46	13	64	21
TP	43	12	12	0
FP	0	1	2	0
H_1 Claimed, other true	4	53	57	59

Table 4: Results of the blind search among $n = 65$ individuals in Section 5.3 with $\alpha = 0.05$ where N_1 denotes the total number of pairs in the sample with the tested relationship, TP is the number of these pairs with a LR above the threshold, and FP is the number of unrelated individuals with a LR above the threshold. The last row gives the number of other (differently related) pairs who also have a LR above the threshold

.

We conclude that the summary in Table 4 is consistent with the performance evaluation shown in Table 3. PO can easily be distinguished from UN. The more distant the tested relationship, the lower the power to distinguish it from unrelated. With the obtained optimal thresholds, the number of false positives stays low as desired. For each hypothesis tested, the list of pairs warranting further investigation comprises those in the final row of Table 4, i.e., those who do not have the tested relationship and who are also not unrelated.

5.4 Analysis of posterior probabilities

The result of each of the blind searches performed in Section 5.3 is a list of pairs with a LR above the threshold. Some pairs of individuals may appear in several of the lists, while other pairs may not be present in any of the lists. In this example, we turn to Bayesian analysis to further investigate specific pairs.

Table 5 shows LR values for 7 pairs from the above blind search. Values above the LR thresholds are in bold font. The rightmost column gives the true relationship. Only the first two pairs have LR values above the thresholds given in the left table of Table 3 corresponding to $\alpha = 0.05$. For pairs 3 to 7, the LRs are low, some below 1, indicating that a UN relationship is more plausible than the alternative hypothesis.

	PO	S	H/U/G	FC	UN	True
1	$\mathbf{5.181{\cdot}10^{10}}$	$\mathbf{1.205{\cdot}10^{8}}$	$\mathbf{1.825{\cdot}10^{7}}$	$\mathbf{5.593{\cdot}10^{4}}$	1	PO
2	353.460	$\mathbf{1.544{\cdot}10^{8}}$	$\mathbf{3.886{\cdot}10^{5}}$	$\mathbf{5.189{\cdot}10^{3}}$	1	S
3	0	0.681	57.572	20.519	1	H
4	0	$5.017{\cdot}10^{-3}$	4.156	4.0984	1	U
5	0	0.030	13.269	16.916	1	G
6	0	$1.115{\cdot}10^{-4}$	0.163	1.375	1	FC
7	0	$1.821{\cdot}10^{-6}$	0.022	0.349	1	UN

Table 5: LR values for seven pairs of the blind search in Section 5.3. Values for H, U and G are the same and shown in the column H/U/G. Values smaller than 10^{-6} are set to 0.

Next we calculate posterior probabilities to see if it is possible to infer a relationship for the different pairs. LR thresholds are not required for this. Table 6 shows posterior probabilities for the different hypotheses, with flat prior probabilities, i.e., $\pi_i = 1/7$ for $i = 0, ..., 6$. The highest probability for each pair is in bold and corresponds to the true relationship for several of the pairs. For example, the LRs comparing S, H/U/G and FC against UN for the second pair were all above the relevant LR thresholds. The posterior probability of S is close to 1, now making it possible to correctly infer this relationship. For pairs 3, 4 and 5, the highest posterior probabilities are just below 0.3. Even though the corresponding relationship is the most probable, a posterior probability of 0.3 is maybe not high enough to allow firm conclusions to be drawn.

	PO	S	H/U/G	FC	UN	True
1	**0.997**	0.002	0.0004	$1.076 \cdot 10^{-6}$	0	PO
2	$2.272 \cdot 10^{-6}$	**0.993**	0.002	$3.336 \cdot 10^{-5}$	0	S
3	0	0.0003	**0.295**	0.105	0.005	H
4	0	$2.86 \cdot 10^{-4}$	**0.237**	0.233	0.057	U
5	0	$5.15 \cdot 10^{-4}$	0.230	**0.293**	0.017	G
6	0	$3.90 \cdot 10^{-5}$	0.057	**0.480**	0.349	FC
7	0	$1.29 \cdot 10^{-6}$	0.0162	0.246	**0.706**	UN

Table 6: Posterior probabilities, computed from the LR values of Table 5, when applying a flat prior, i.e., $\pi_i = 1/7$ for $i = 0, ..., 6$, as described in Section 5.4. Values for H, U and G are the same and shown in the column H/U/G. Probabilities smaller than 10^{-6} are set to 0.

The relationships H, U and G are indistinguishable in the parametric framework presented in Section 2. Also posterior probabilities with a flat prior as in Table 6 can not differentiate between them. Additional information, preferably objective, needs to be considered.

Suppose now that we have knowledge of how many pairs of the different relationships are present among the DNA profiles. This could be the case in a plane crash with a known passenger list. The number of pairs of the different relationships are given in row 2 (N_1) of Table 4. There are 1867 unrelated pairs. The prior probabilities are $\pi_0 = 0.898$ (UN), $\pi_1 = 0.022$ (PO), $\pi_2 = 0.006$ (S) , $\pi_3 = 0.0020$ (H), $\pi_4 = 0.016$ (U), $\pi_5 = 0.013$ (G) and $\pi_6 = 0.010$ (FC).

Posterior probabilities using these more informative priors are shown in Table 7. The prior probability of a PO relationships is $\pi_1 = 0.022$, i.e., there is a chance of 2.2% that a pair of individuals has a PO relationship. The corresponding posterior probability for the first pair is 0.999. The genetic data give such strong support to PO, that even though the prior probability is low, the posterior probability of this relationship is approximately 1.

In this blind search (as in most other blind searches), most pairs of individuals are unrelated, making the prior probability of UN close to 1 and the others low. This requires the LRs for the other relationships to be high in order to be supported by the posterior probabilities. For the relationships H/U/G and FC,

	PO	S	H	U	G	FC	UN	True
1	**0.999**	$6.57{\cdot}10^{-4}$	$3.06{\cdot}10^{-5}$	$2.52{\cdot}10^{-4}$	$2.07{\cdot}10^{-4}$	0	0	PO
2	$8{\cdot}10^{-6}$	**0.988**	$7.65{\cdot}10^{-4}$	0.006	0.005	$5.36{\cdot}10^{-5}$	0	S
3	0	0.001	0.038	**0.317**	0.259	0.072	0.312	H
4	0	$2.94{\cdot}10^{-5}$	0.007	0.062	0.051	0.038	**0.841**	U
5	0	$1.26{\cdot}10^{-4}$	0.017	0.143	0.117	0.116	**0.608**	G
6	0	0	$3.42{\cdot}10^{-4}$	0.003	0.002	0.015	**0.979**	FC
7	0	0	$4.78{\cdot}10^{-5}$	$3.95{\cdot}10^{-4}$	$3.23{\cdot}10^{-4}$	0.004	**0.995**	UN

Table 7: Posterior probabilities with informative priors, as described in Section 5.4. Probabilities smaller than 10^{-6} are set to 0.

the LR of the true relationships against UN is typically low. The LR of UN against UN is always 1. The combination of priors and LRs makes the posterior probability of UN high while the posterior probability of the true relationship remains low.

For this reason, this particular set of prior probabilities, even though objective, does not help us to distinguish between the H, U and G relationships in these data.

5.5 Blind search with X-chromosomal markers

Because a male has only one X-chromosome, paternal half sisters (HSP) must inherit the same X-chromosome from their common father. Their second X-chromosomes, inherited from their respective mothers, are not IBD (since their mothers are unrelated), and hence, the IBD coefficients for a HSP relationship are $\boldsymbol{\kappa} = (0, 1, 0)$. The IBD coefficients for maternal half siblings (HSM), whether considering X-chromosomal or autosomal markers, are $\boldsymbol{\kappa} = (0.5, 0.5, 0)$. In the following example, we show with simulated data how X-chromosomal markers can distinguish between HSP and HSM.

We simulated genotypes for 12 X-chromosomal STR markers, for the shaded individuals in Figure 8. Table 8 presents the average posterior probabilities over 100 simulations, for the relationships PO, S, HSP, HSM and UN, for the six possible comparisons between the individuals A, B , C and D. A flat prior $\pi_i = 1/5$ for $i = 0, \ldots, 4$ is assumed.

The evidence in favour of C-D being HSM, shown in bold in Table 8, could not be obtained using autosomal markers. Since we are using a flat prior, the LR comparing maternal to paternal half sibs can be found from the posterior probability ratio, $0.81327/0.01916 = 42.4$. This value may not be decisive on its own, but supplements other evidence. Note that HSP cannot be distinguished from PO using X-chromosomal markers alone as the row for the comparison A-C confirms. Age information, autosomal marker data or other non-DNA data may solve such cases.

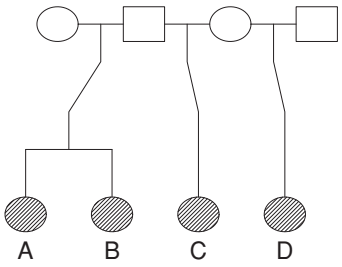


Figure 8: Pedigree connecting the individuals of the analysis in Section 5.5. Marker data are simulated for the four daughters to demonstrate blind search with X-chromosomal markers.

	PO	S	HSP	HSM	UN
A-B	0.039	0.921	0.039	0.001	0.000
A-C	0.475	0.039	0.475	0.012	$2 \cdot 10^{-5}$
A-D	0.000	0.000	0.000	0.154	0.846
B-C	0.471	0.045	0.471	0.012	$2 \cdot 10^{-5}$
B-D	0.000	0.000	0.000	0.146	0.854
C-D	0.019	0.001	0.019	**0.813**	0.147

Table 8: Posterior probabilities averaged over 100 simulations for the comparisons between the four daughters in Figure 8.

.

5.6 Half siblings with inbred founder

Computation of LRs and posterior probabilities is restricted to a limited set of predefined pedigree relationships in many current software implementations. The parametric form of the LR given in (3) enables us to compute LRs and do blind search for any pairwise relationship. In this example we show how background inbreeding can be modelled and how this can be taken into account in the Bayesian framework.

Assume a set of DNA profiles among which we want to do a blind search. The number of profiles is not important. The pedigrees connecting the individuals are unknown, but we know that the individuals come from a population where inbreeding is common. We consider the hypotheses H_1: PO, H_2: S, H_3: H and H_4: H with founder inbreeding $f = 0.25$, all against H_0: UN. The relationship in H_4 is shown in Figure 9. Individuals A and B are outbred paternal half

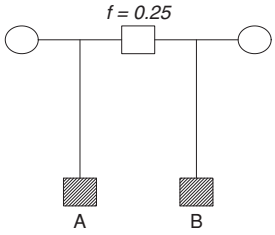


Figure 9: Half sibling pedigree with founder inbreeding assumed in the analysis in Section 5.6.

siblings, with the father being inbred with an inbreeding coefficient $f = 0.25$. This value of f corresponds to extreme inbreeding where the parents of the father are siblings. The IBD coefficients for the half sibling relationship are $\boldsymbol{\kappa} = (0.375, 0.625, 0)$.

We consider one pair with true relationship H_4. A total of 100 simulations of DNA profiles for this pair is performed. LRs and posterior probabilities, with a flat prior $\pi_i = 1/5$, $i = 0, \ldots, 4$ are computed for each simulation. Mean values of the posterior probabilities for the hypotheses $\mathrm{H}_0, ..., \mathrm{H}_4$, are: $\bar{p}_1 = 0.017$, $\bar{p}_2 = 0.094$, $\bar{p}_3 = 0.374$, $\bar{p}_4 = 0.495$ and $\bar{p}_0 = 0.019$. It can be seen that the mean posterior probability of hypothesis $\mathrm{H}(f)$ is about 0.5, making it possible to distinguish it from the half sibling relationship without inbreeding.

The coefficient of inbreeding in this example is quite high. Lower values of f makes the pair genetically more similar to half siblings without inbreeding, and distinguishing these relationships becomes harder without additional information. This high degree of inbreeding may be more relevant for non-human applications.

6 Discussion

The topic of this paper is blind search, a procedure used to search for pairwise relationships among a set of unidentified DNA profiles. Each pairwise comparison is similar to a kinship test performed, for instance, to resolve a paternity dispute. In the paper, we focus mainly on issues related to multiple testing. For this reason we will not discuss Hardy-Weinberg equilibrium and other assumptions that our applications share with other applications in forensic genetics. For instance, it is not obvious how evidence from different DNA sources like autosomal markers and X-chromosomal markers should be combined. However, this challenge is no different for a blind search than for a kinship test and is therefore not addressed here.

Case workers must decide on how the results of a blind search should be evaluated and reported. The context, or specific application, is obviously not irrelevant. In a DVI application, a false identification is likely to be a more serious error than missing an identification. To account for this, the metric for determining the threshold in (7) allows a weight to be specified which would penalise false identifications. Other applications, such as screening a database for relatives prior to estimating allele frequencies, may not require a weighting for errors. If costs can be specified for the possible errors, optimal decision rules can be derived as explained in Chapter 8.1 of [10] . However, there is hardly ever an objective way to balance the two errors that can occur and so specification of weights or costs may not be a viable option. We have used the unweighted form of the metric throughout.

We only presented one method to determine an optimal threshold based on the distance illustrated in Figure 4 although several alternatives are available [31]. Results using different approaches were practically identical for the examples we presented and so we chose not to discuss the thresholds based on the other metrics.

Figure 6 shows that the LRs from a blind search may be correlated when the same individual is involved in two comparisons. This has several implications. In particular, the results of different comparisons cannot be interpreted independently. Intuitively, we may get a high LR if unrelated individuals A and B happen to share a rare allele. Another individual C, who is a close relative of A, is likely to share this allele IBD with A and so we can also expect a high LR when comparing B and C. Importantly, the methods used to control the overall error rate must allow for dependence and for this reason we used the Bonferroni bound (8). Furthermore, a blind search will not necessarily provide a globally consistent 'solution' in the sense that the LRs may support impossible combinations of relationships, like one individual having two mothers. Finally, the true relationship may not be among the alternatives considered. This is also true for the Bayesian approach.

A Bayesian interpretation might seem more appropriate than the frequentist alternative for blind search applications than for a kinship case. The alternative, based on the LR, is designed to deal with only two hypotheses. If there are several hypotheses, a reference hypothesis must be specified. The posterior probabilities reported using a Bayesian approach make comparison of several competing hypotheses simpler as they are between 0 and 1. However, as always, a prior is needed for the Bayesian approach and the choice of prior may be crucial. If DNA is of poor quality, leading to few markers being typed, or if the competing hypotheses specify relationships that are very close to each other, conclusions may hinge on the choice of the prior.

An important aspect of this paper is the use of the parametric representation of relationships. This enables us to investigate any admissible pairwise relationship between two outbred individuals. By defining founder inbreeding in a pedigree structure, as shown in Figure 3, background inbreeding can also be modelled [22]. Rather than proposing specific alternative relationships, we could simply estimate the coefficients describing the relationship. In the out-

bred case, these estimates can be plotted in the IBD triangle in Figure 2 which would indicate where these relationships lie in relation to the well known relationships. For instance, pairs with estimates close to (0.25, 0.25) could be classified as siblings.

Throughout, we have restricted attention to pairwise testing. In principle, the blind search can be extended to search for relationship between triplets. However, the parametric approach based on the Jacquard coefficients then becomes impractical. The number of parameters needed to describe the relationship between three individuals increases, from 2 to 15 in the outbred case.

Issues to do with reporting DNA evidence are currently of key interest as evidenced by the so-called "DNA database controversy" (see [11] and references therein). The main message of this paper is that there are also problems related to multiple testing in kinship analyses which cannot be ignored.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

We thank Burkhard Rolf for DNA samples. Margherita Colucci was supported by a BBSRC-MIBTP (grant no. BB/M01116X/1) iCASE studentship, sponsored by DNA Worldwide.

References

[1] B. Olaisen, M. Stenersen, and Mevåg M. Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nature Genetics*, 15:402–405, 1997.

[2] B. Bertoglio, P. Grignani, P. Di Simone, N. Polizzi, D. De Angelis, C. Cattaneo, A. Iadicicco, P. Fattorini, S. Presciuttini, and C. Previderè. Disaster victim identification by kinship analysis: the Lampedusa October 3rd, 2013 shipwreck. *Forensic Science International: Genetics*, 44:102–156, 2020.

[3] L. Olivieri, D. Mazzarelli, B. Bertoglio, D. De Angelis, C. Previderè, P. Grignani, A. Cappella, S. Presciuttini, C. Bertuglia, P. Di Simone, and other. Challenges in the identification of dead migrants in the Mediterranean: the case study of the Lampedusa shipwreck of October 3rd 2013. *Forensic Science International*, 285:121–128, 2018.

[4] C. H. Brenner. Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities. *Forensic Science International*, 157(2-3):172–180, 2006.

[5] C. H. Brenner and B. S. Weir. Issues and strategies in the DNA identification of World Trade Center victims. *Theoretical population biology*, 63(3):173–178, 2003.

[6] T. J. Parsons, R. M. L. Huel, Z. Bajunović, and A. Rizvić. Large scale DNA identification: The ICMP experience. *Forensic Science International: Genetics*, 38:236–244, 2019.

[7] S. Palomo-Díez, Á. Esparza-Arroyo, M. Tirado-Vizcaíno, J. Velasco Vázquez, A. M. López-Parra, C. Gomes, C. Baeza-Richer, and E. Arroyo-Pardo. Kinship analysis and allelic dropout: a forensic approach on an archaeological case. *Annals of Human Biology*, 45(4):365–368, 2018.

[8] S. Palomo-Díez, A. M. López-Parra, C. Gomes, C. Baeza-Richer, A. Esparza-Arroyo, and E. Arroyo-Pardo. Kinship analysis in mass graves: evaluation of the Blind Search tool of the Familias 3.0 Software in critical samples. *Forensic Science International: Genetics Supplement Series*, 5:e547–e550, 2015.

[9] T. J. Pemberton, C. Wang, J. Z. Li, and N. A. Rosenberg. Inference of unexpected genetic relatedness among individuals in hapmap phase iii. *The American Journal of Human Genetics*, 87(4):457–464, 2010.

[10] T. Egeland, D. Kling, and P. Mostad. *Relationship inference with Familias and R: Statistical methods in forensic genetics*. Academic Press, 2015.

[11] G. Storvik and T. Egeland. The DNA database search controversy revisited: bridging the Bayesian–frequentist gap. *Biometrics*, 63(3):922–925, 2007.

[12] F. R. Bieber, C. H. Brenner, and D. Lazer. Finding criminals through DNA of their relatives. *Science*, 2006.

[13] M. Kruijver, R. Meester, and K. Slooten. Optimal strategies for familial searching. *Forensic Science International: Genetics*, 13:90–103, 2014.

[14] J. D. Storey et al. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

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

[16] T. Egeland, P. F. Mostad, B. Mevåg, and M. Stenersen. Beyond traditional paternity and identification cases: selecting the most probable pedigree. *Forensic Science International*, 110(1):47–59, 2000.

[17] A. Jacquard. Genetic information given by a relative. *Biometrics*, pages 1101–1114, 1972.

[18] N. A. Sheehan and T. Egeland. Structured incorporation of prior information in relationship identification problems. *Annals of Human Genetics*, 71(4):501–518, 2007.

[19] S. Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645):330–338, 1922.

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

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

[22] M. D. Vigeland and T. Egeland. Handling founder inbreeding in forensic kinship analysis. *Forensic Science International: Genetics Supplement Series*, 7(1):780–781, 2019.

[23] H. K. Brustad and T. Egeland. The impact of ignoring inbreeding in pairwise kinship evaluations. *Forensic Science International: Genetics Supplement Series*, 7(1):462–464, 2019.

[24] M. D. Vigeland. Relatedness coefficients in pedigrees with inbred founders. *Journal of Mathematical Biology*, 81:185–207, 2020.

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

[26] T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[27] N. Pinto, L. Gusmão, and A. Amorim. X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial. *Forensic Science International: Genetics*, 5(1):27–32, 2011.

[28] D. Kling, A. Tillmar, T. Egeland, and P. Mostad. A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations. *International Journal of Legal Medicine*, 129(5):943–954, 2015.

[29] D. Kling, B. Dell'Amico, and A. O. Tillmar. FamLinkX–implementation of a general model for likelihood computations for X-chromosomal marker data. *Forensic Science International: Genetics*, 17:1–7, 2015.

[30] M. Kruijver. Efficient computations with the likelihood ratio distribution. *Forensic Science International: Genetics*, 14:116–124, 2015.

[31] M. Rota and L. Antolini. Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers. *Computational Statistics & Data Analysis*, 69:1–14, 2014.

[32] A. C. Tamhane, Y. Hochberg, and C. W. Dunnett. Multiple test procedures for dose finding. *Biometrics*, pages 21–37, 1996.

[33] M. Steffens, C. Lamina, T. Illig, T. Bettecken, R. Vogler, P. Entz, E. Suk, M. R. Toliat, N. Klopp, A. Caliebe, et al. SNP-based analysis of genetic substructure in the German population. *Human Heredity*, 62(1):20–29, 2006.

[34] C. Phillips, L. Fernandez-Formoso, M. Garcia-Magarinos, L. Porras, T. Tvedebrink, J. Amigo, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, A. Freire-Aradas, et al. Analysis of global variability in 15 established and 5 new european standard set (ess) strs using the ceph human genome diversity panel. *Forensic Science International: Genetics*, 5(3):155–169, 2011.

[35] J. Amigo, C. Phillips, T. Salas, L. F. Formoso, Á. Carracedo, and M. Lareu. pop.str — an online population frequency browser for established and new forensic strs. *Forensic Science International: Genetics Supplement Series*, 2(1):361–362, 2009.

[36] M.G. García, C. I. Catanesi, G. A. Penacino, L. Gusmão, and N. Pinto. X-chromosome data for 12 STRs: Towards an Argentinian database of forensic haplotype frequencies. *Forensic Science International: Genetics*, 41:e8–e13, 2019.

[37] T. Egeland, N. Pinto, and A. Amorim. Exact likelihood ratio calculations for pairwise cases. *Forensic Science International: Genetics*, 29:218–224, 2017.

A Table of genotype probabilities

Table 9: The conditional probability $\mathrm{P}(G \mid \mathrm{J}_i)$ of a pair of genotypes $G = (g_A, g_B)$, given a Jacquard state J_i. The symbols a, b, c and d represent different alleles, with population frequencies p_a, p_b, p_c and p_d respectively.

(g_A, g_B)	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
(aa, aa)	p_a	p_a^2	p_a^2	p_a^3	p_a^2	p_a^3	p_a^2	p_a^3	p_a^4
(aa, bb)	0	$p_a p_b$	0	$p_a p_b^2$	0	$p_a^2 p_b$	0	0	$p_a^2 p_b^2$
(aa, ab)	0	0	$p_a p_b$	$2p_a^2 p_b$	0	0	0	$p_a^2 p_b$	$2p_a^3 p_b$
(aa, bc)	0	0	0	$2p_a p_b p_c$	0	0	0	0	$2p_a^2 p_b p_c$
(ab, aa)	0	0	0	0	$p_a p_b$	$2p_a^2 p_b$	0	$p_a^2 p_b$	$2p_a^3 p_b$
(bc, aa)	0	0	0	0	0	$2p_a p_b p_c$	0	0	$2p_a^2 p_b p_c$
(ab, ab)	0	0	0	0	0	0	$2p_a p_b$	$p_a p_b (p_a + p_b)$	$4p_a^2 p_b^2$
(ab, ac)	0	0	0	0	0	0	0	$p_a p_b p_c$	$4p_a^2 p_b p_c$
(ab, cd)	0	0	0	0	0	0	0	0	$4p_a p_b p_c p_d$

B Importance sampling

Importance sampling is a method that can be used to approximate small probabilities. We first introduce the indicator function,

$$I(\text{LR} > t) = \begin{cases} 1, & \text{if LR} \geq t, \\ 0, & \text{if LR} < t. \end{cases}$$

The expectation of I becomes

$$\begin{aligned} \text{E}(I(\text{LR} \geq t)) &= 0 \cdot \text{P}(\text{LR} < t) + 1 \cdot \text{P}(\text{LR} \geq t) \\ &= \text{P}(\text{LR} \geq t) \\ &= \text{FPR} \end{aligned}$$

It is therefore valid to say that $\text{FPR} = \text{E}(I(\text{LR} \geq t))$. Then consider the expression for the expected value in a more general sense. The value of the function I is dependent on the value of the LR, which is a function of the genotypes G of the DNA profiles. The probability distribution of G is governed by the relationships that has generated the data. For this consideration, we assume that this relationship is either H_0 or H_1. Denote by X the values that I can take on. We then have

$$\text{E}(I(\text{LR} \geq t)) = \sum_j X_j \cdot \text{P}(G_j \mid \text{H}_0) \approx \frac{1}{N} \sum_{i=1}^{N} I(\text{LR}_i^{\text{H}_0} \geq t).$$

In the last expression, the expected value is estimated by the sample mean of I, from a set of N simulations. The genotypes G, and then also X, are distributed according to H_0, which is indicated by the superscript of the LR. Then, consider the opposite probability distribution, $\text{P}(G \mid \text{H}_1)$, where the genotypes are distributed according to H_1. As long as $\text{P}(G_j \mid \text{H}_0) = 0$ whenever $\text{P}(G_j \mid \text{H}_1) = 0$, we can write

$$\text{E}(I(\text{LR} \geq t)) = \sum_j X_j \cdot \frac{\text{P}(G_j \mid \text{H}_0)}{\text{P}(G_j \mid \text{H}_1)} \text{P}(G_j \mid \text{H}_1) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{I(\text{LR}_i^{\text{H}_1} \geq t)}{LR_i^{\text{H}_1}}.$$

Using this method, the LR is sampled under the wrong hypothesis (H_1), instead of the desired hypothesis (H_0). The bias this introduces is adjusted for by the weight LR^{H_1}. An estimate of FPR is then

$$\widehat{\text{FPR}} = \frac{1}{N} \sum_{i=1}^{N} \frac{I(LR_i^{\text{H}_1} \geq t)}{LR_i^{\text{H}_1}}.$$

Paper IV

Hilde K. Brustad*, Geir O. Storvik, Magnus D. Vigeland, and Thore Egeland

Estimating realised pairwise relatedness with bootstrap confidence

Abstract: Estimation of pairwise relatedness from genetic marker data was established several decades ago, and continues to be an important tool both in forensics and other fields of genetics. Typical implementations use maximum likelihood methods to estimate various relatedness coefficients, which describe the relationship in different ways. In this work we investigate bootstrap approaches to assessing the uncertainty of such estimates. In particular, we compare non-parametric bootstrap, where existing marker data are sampled with replacement, to parametric bootstrap, where marker data are sampled from the likelihood function. Differences and similarities between the methods are discussed and examples based on simulated and real data are given. In many cases both methods give similar results. However, we argue that from a theoretical point of view parametric bootstrap is more reasonable to use.

Keywords: Pairwise relatedness, Identity-by-descent (IBD), Bootstrap, Confidence regions

1 Introduction

Relatedness is of interest in areas as disparate as human disease, gene mapping, plant breeding and ecology [1]. In forensic applications, the focus has traditionally been on the pedigree relationship connecting individuals [2], ranging from standard paternity testing to disaster victim identification cases [3].

We restrict attention to pairwise relationships, i.e., how two individuals are related. The traditional approach to relationship inference in forensic genetics is to compare competing hypotheses by means of a likelihood ratio (LR). This approach

***Corresponding author: Hilde K. Brustad,** Faculty of chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Norway
Geir O. Storvik, Department of Mathematics, University of Oslo, Norway
Magnus D. Vigeland, Department of Medical Genetics, University of Oslo, Norway
Thore Egeland, Faculty of chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Norway

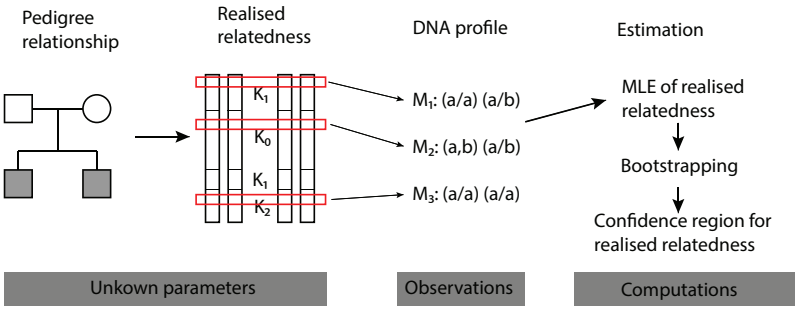


Fig. 1: The pedigree relationship to the left gives rise to a IBD pattern between the genomes of two individuals, summarised as the realised relatedness between two individuals. We only observe genotypes at discrete points along the genome, and from this we estimate the realised relatedness.

limits the analysis to a finite set of alternatives. A parametric LR approach that considers all possible alternatives has been suggested [4].

An alternative approach to kinship testing is to estimate appropriate relatedness parameters. This approach is the focus of this work. Figure 1 summarises our framework and approach. The panel to the left describes the *pedigree relationship* between two brothers. The identity-by-descent (IBD) parameters describes relationships between pairs of individuals. Two homologous alleles are IBD if they are identical by state (IBS) and inherited from the same ancestor allele. Mendel's laws implies that non-inbred full brothers share 0, 1 or 2 alleles IBD at a random autosomal locus, with probabilities 0.25, 0.50 and 0.25 respectively. Hence, these numbers are also the expected fractions of the autosomal chromosomes sharing the respective number of alleles. However, in a given pair of individuals, the actual IBD-sharing proportions – referred to as the *realised relatedness* – may differ substantially from the expected values. The realised relatedness cannot be observed directly, but can be estimated by means of genetic *markers*, as indicated by the two middle panels of Figure 1. It should be noted that even with complete sequence data from both individuals, the observed patterns of IBS will rarely determine IBD status absolutely, hence there will always be some uncertainty in the estimation of realised relatedness.

The aim of this work is to assess the uncertainty of the estimated realised relatedness through bootstrapping and confidence regions. The literature in general lacks a methodical discussion about the use of bootstrapping in the current context.

Maximum likelihood (ML) estimation of pairwise relatedness has been reviewed by Thompson [5], Milligan [6], and Anderson and Weir [7]. Milligan reduced estimates of the more refined relatedness coefficients to the one-dimensional kinship coefficient, and evaluated the estimator by means of bias, standard error and root mean square error. Anderson and Weir [7] also assessed the uncertainty in the kinship coefficient. They apply non-parametric bootstrap (resampling with replacement) across loci. A confidence interval was then computed from the bootstrap samples.

The choice made by previous authors to focus on the kinship coefficient is sensible, as it is easier to evaluate than the nine-dimensional identity coefficients of Jacquard [8]. However, these coefficients offer much more information about a relationship than the single kinship coefficient. We aim to show that bootstrap methods are well-suited to assess the uncertainty of higher-dimensional coefficients. Moreover, we offer a comparison between parametric and non-parametric bootstrap. We argue that the parametric bootstrap, which apparently has not been previously used for kinship applications, in certain situations outperforms the non-parametric version.

2 Review of previous results

In the following sections, we review the definitions of standard relatedness coefficients, and how these are estimated.

2.1 Pairwise relatedness coefficients

The kinship coefficient φ between two pedigree members is the probability that a random allele from one individual is IBD to a random allele of another individual, at the same autosomal locus. Building on work by Cotterman [9], Thompson defined the coefficients $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$ as a more refined measure of relatedness, valid for outbred individuals, where κ_i is the probability of sharing exactly i alleles IBD at a random autosomal locus [5]. In symbols we write $\kappa_i = \mathrm{P}(\mathrm{K}_i \mid \text{pedigree})$, for $i = 0, 1, 2$, where K_i is the event of sharing i alleles IBD (see Figure 2). Because $\kappa_0 + \kappa_1 + \kappa_2 = 1$, the coefficients can be visualised by the point (κ_0, κ_2) in the *IBD triangle* of Figure 3.

Thompson [10] showed that $\boldsymbol{\kappa}$ is restricted by the inequality $\kappa_1^2 \geq 4\kappa_0\kappa_2$, defining an inadmissible region in the IBD triangle (shown in grey). It is impossible to construct a pedigree where the relationship between to pedigree members has

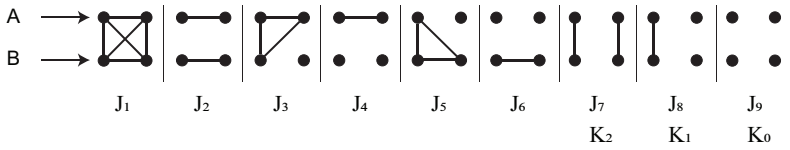


Fig. 2: The nine possible Jacquard states J_i for $i = 1, 2, \ldots, 9$ between the four alleles of the individuals A and B. Dots denote alleles, while lines connecting the dots indicate IBD alleles. The last three states J_9, J_8 and J_7 are for outbred individuals reduced to the IBD states K_0, K_1 and K_2, respectively.

IBD coefficients in the inadmissible region. The IBD coefficients relate to the kinship coefficient through the equation $\varphi = \frac{1}{4}\kappa_1 + \frac{1}{2}\kappa_2$, illustrated by the dashed lines in Figure 3.

Jacquard described inbred relationships by introducing the nine *condensed identity coefficients* $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \ldots, \Delta_9)$ [8]. These coefficients, also called the *Jacquard coefficients* are defined as the expected relative frequencies of the nine IBD configurations (or states) $J_1, \ldots, J_9$ shown in Figure 2, i.e. $\Delta_i = P(J_i \mid \text{pedigree})$, where $i = 1, 2, \ldots, 9$. When both individuals are outbred, the states $J_1, \ldots, J_6$ cannot occur, and $\boldsymbol{\Delta}$ reduces to $\boldsymbol{\kappa}$. The Jacquard coefficients relate to the kinship coefficient through the equation

$$\varphi = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8. \tag{1}$$

2.2 Realised relatedness

The realisation of a pedigree relationship, i.e., the actual proportion of the genomes of two individuals in the different IBD-sharing states, may differ from the IBD-sharing probabilities associated with their pedigree relationship [11]. We call these actual IBD sharing proportion the realised relatedness, denoted by $\boldsymbol{\Delta}_R$, or $\boldsymbol{\kappa}_R$ for two outbred individuals. Some pairs of individuals may have genomes that share long but few segments IBD, while other individuals have genomes that share short but many, segments IBD, even though the realised relatedness is the same. The IBD coefficients describing the realised relatedness, $\boldsymbol{\kappa}_R$, is only restricted by $\kappa_2^R \leq 1 - \kappa_0^R$.

Hill and Weir [12] viewed the realised relatedness, given a pedigree relationship, as a stochastic variable and derived formulas for the variation and skewness of the realised relatedness for a series of outbred pedigree relationships. They showed that

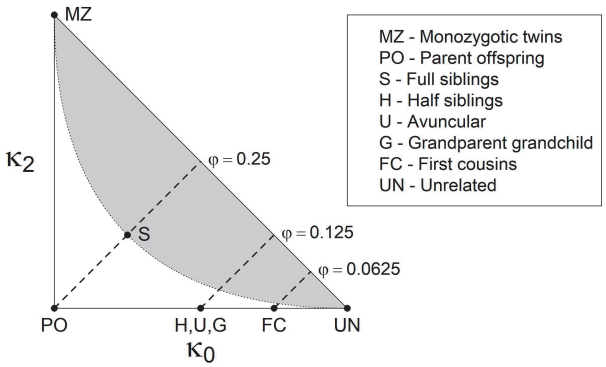


Fig. 3: IBD triangle with location of some common relationships. Dashed lines show how φ is related to the IBD coefficients. The inadmissible region for pedigree relationships is shown in grey.

the realised IBD-sharing proportions between two genomes, for a given pedigree relationship, vary considerably.

The corners of the IBD triangle represent the only (outbred) relationships with no variance in realised relationship. For example, any outbred parent-offspring (PO) relationship has $\boldsymbol{\kappa}_R = \boldsymbol{\kappa} = (0, 1, 0)$. More generally, the realised value of any pedigree-based coefficient of 0, is also 0. For example, all relationships located on the bottom edge of the IBD triangle, including half siblings (H), avuncular (U), grandparent-grandchild (G) and first cousins (FC), must satisfy $\kappa_2^R = 0$. The above observations may be summarised as follows

$$\Delta_i = 0 \Rightarrow \Delta_i^R = 0,\ i = 1, \ldots, 9. \tag{2}$$

2.3 Estimation of realised relatedness

Based on a set of DNA profiles, we are able to perform a maximum likelihood (ML) estimation of the relatedness coefficients. Let $G = (g_1, \ldots, g_M)$ denote two DNA profiles genotyped at M independent loci, where $g_j = (g_{A,j}, g_{B,j})$ is the genotypes of individual A and B at locus j. The likelihood of $\boldsymbol{\Delta}_R$, based on the observations G, is given by the equation

$$L(\boldsymbol{\Delta}_R) = \prod_{j=1}^{M} \left(\sum_{i=1}^{9} \Delta_i^R \mathrm{P}(g_j \mid \mathrm{J}_i) \right). \tag{3}$$

The probabilities $\mathrm{P}(g_j \mid \mathrm{J}_i)$ are given in Table 4. The estimated realised relatedness, $\widehat{\boldsymbol{\Delta}}_R$, is the maximum of (3) in the parameter space

$$\Omega = \{(\Delta_1^R, \ldots, \Delta_9^R) \mid \sum_{i=1}^{9} \Delta_i^R = 1, 0 \leq \Delta_1^R, \ldots, \Delta_9^R \leq 1\} \in \mathbb{R}^9. \tag{4}$$

ML estimation of the IBD coefficients was described by Thompson in 1975 [5], and investigated further and compared to other estimation methods by Milligan [6]. See also other references therein. These papers evaluate the estimation by reducing $\boldsymbol{\kappa}$ to the one-dimensional kinship coefficient φ. Milligan concludes that the ML estimator has the smallest variation, but is more biased under some conditions than the other estimation methods he investigates. Anderson and Weir [7] also assess the bias in the estimation of the kinship coefficient. The papers by Milligan and Anderson and Weir do not distinguish between pedigree relationships and realised relatedness. However, this does not affect the bias analysis for some relationships, like PO and UN, because the realized relatedness for these relationships, in the outbred case, are always $\varphi_R = 0.25$ and $\varphi_R = 0$, respectively, as a consequence of (2).

Milligan points out that because the ML estimate is restricted to the biological meaningful range, this introduces a bias in the estimate of relationships close to the boundary. Many of the common and important pedigree relationships are located on the border of the parameter space, and the corresponding realised relatedness will also be located on the boundary as implied by (2). The limited number of independent loci available is also an important contributor to bias, affecting both interior and boundary points.

Given a set of regularity conditions, the ML estimate has many desirable asymptotic properties [13]. One of the regularity conditions is that the parameter should be part of the interior of the parameter space. In e.g. a forensic casework setting, we are not able to know if the realised relatedness between two individuals fulfill this requirement. Consequently, an explicit formula to quantify the uncertainty of the ML estimate is not available for our applications. We therefore estimate the uncertainty by bootstrapping [14].

The likelihood model (3) is only valid under certain conditions. We assume throughout that the loci are independent and that the population is in Hardy-Weinberg Equilibrium (HWE). Deviation from HWE affects the probabilities in Table 4. Anderson and Weir [7] propose an estimator suitable for structured populations, or equivalently, populations not in HWE. Mutations are not modeled by (3). For mutations to be meaningful in this context, a pedigree would have to be defined. In addition, artifacts like drop-ins, drop-outs, silent alleles and genotyping errors are not modeled.

3 Methods

Bootstrapping [14] is a group of simulation based methods used to assess the uncertainty of a parameter estimate when an explicit formula for the uncertainty is not available. The general idea behind bootstrapping is the following:

Suppose a set of observations G are generated from some unknown probability distribution F. Associated with this distribution is our parameter of interest $\mathbf{\Delta}_R$, which we estimate by $\widehat{\mathbf{\Delta}}_R = T(G)$. We generate bootstrap data sets $G_1^*, \ldots, G_B^*$ from $\widehat{\mathrm{F}}$, an approximation of F, aimed to mimic the stochastic process that generated the original data. Bootstrap estimates are calculated from these new data sets, i.e., $\widehat{\mathbf{\Delta}}_b^* = T(G_b^*)$ for $b = 1, \ldots, B$. The function T that estimates $\mathbf{\Delta}_R$ is not explicitly known and estimation is therefore performed numerically.

The true variance of the bootstrap estimate is estimated by the sample variance of the bootstrap estimates. We want the true variance of the bootstrap estimates to approximate the variance of the original parameter estimate. There are two sources of error regarding this approximation. One source is that the sample variance of the bootstrap estimates is computed from a limited number of bootstrap estimates, B. Increasing B will reduce the difference between the sample variance and the true variance of the bootstrap estimates. The second source of error is subjected to how well the bootstrap method approximates the distribution F. Increasing the observations in the original data set G can contribute to decreasing this error.

Parametric and non-parametric bootstrap are two bootstrapping procedures that differ in the way bootstrap data sets are generated. In other words, $\widehat{\mathrm{F}}$ differ.

The validity of the bootstrap procedure depends on many of the same regularity conditions as the ML estimator, the most important one being the requirement that the parameter should be an interior point in the parameter space.

3.1 Non-parametric bootstrap

In non-parametric bootstrap, bootstrap data sets are made by resampling data with replacement from the observation we have, i.e., randomly draw M joint genotypes with replacement from our two DNA profiles.

Only the joint genotypes of the two original DNA profiles are possible to observe in the bootstrap data sets. The same locus may appear several times in a bootstrap sample, opposite to what is the case for the original DNA profiles. This would not have been a problem if the markers had been independent identically distributed (i.i.d). However, the markers are non-i.i.d and the question is then if

the distribution of the bootstrap estimates resembles the true distribution of the parameter estimator.

3.2 Parametric bootstrap

An alternative to the non-parametric bootstrap is parametric bootstrap. Bootstrap data sets are obtained by sampling new joint genotypes at each locus individually, from the probability distribution

$$\mathrm{P}(\mathcal{G} \mid \widehat{\boldsymbol{\Delta}}_R) = \sum_{i=1}^{9} \widehat{\Delta}_i^R \mathrm{P}(\mathcal{G} \mid \mathrm{J}_i), \tag{5}$$

where $\mathcal{G}$ is a random variable taking values in the joint genotypes that are possible to observe at a locus. As an example, consider the case $\widehat{\boldsymbol{\kappa}} = (0, 1, 0)$. Genotypes for the bootstrap data sets are sampled from the distribution $\mathrm{P}(\mathcal{G} \mid \kappa_1 = 1) = P(\mathcal{G} \mid K_1)$. From Table 4 we see that only joint genotypes with at least one allele IBS are possible to sample.

After obtaining the original estimate $\widehat{\boldsymbol{\Delta}}_R$, the original DNA profiles are no longer used. Hence, given a set of markers and corresponding allele frequencies, the distribution of the parametric bootstrap estimates are equal for cases where the original estimates $\widehat{\boldsymbol{\Delta}}_R$ are the same.

3.3 Confidence regions

The uncertainty of a parameter estimate can be assessed by a confidence interval, or a confidence region for a multi-dimensional parameter.

We first focus on the one-dimensional kinship coefficient φ, which relates to the Jacquard coefficients through (1). Anderson and Weir [7] apply non-parametric bootstrap and approximate a 95% confidence interval for φ_R as the interval covered by the middle 95% of the bootstrap estimates, often called a percentile bootstrap confidence interval. By this method, the bootstrap confidence interval is constructed by the percentiles of the empirical distribution of the bootstrap estimates. A $(1 - \alpha)100\%$ bootstrap confidence interval for φ_R is given by the interval

$$\big(\varphi*_{(\alpha/2)}, \varphi*_{(1-\alpha/2)}\big),$$

where $\varphi*_{(1-\alpha/2)}$ is the $1 - \alpha/2$ percentile of the empirical distribution of the bootstrap estimates. The one-dimensional confidence intervals we present in the result section are given by this interval. To visualize a confidence region in nine-

dimensions, as would be the case for $\boldsymbol{\Delta}_R$, is not practical. We therefore evaluate each Jacquard coefficient marginally by use of the percentile method.

Estimation of $\boldsymbol{\kappa}_R$ is nicely visualised in the IBD triangle. Instead of a confidence interval, we construct a confidence region in two-dimensions. Different bootstrap confidence regions are described in the literature [15, 16]. We construct a $(1-\alpha)100\%$ confidence region as an ellipse containing $(1-\alpha)100\%$ of the bootstrap estimates. The center of the ellipse corresponds to the mean of the bootstrap estimates $(\overline{\kappa}_0^*, \overline{\kappa}_2^*)$, where $\overline{\kappa}_i^* = \sum_{b=1}^B \widehat{\kappa}_{i,b}^*/B$. The ellipse is rotated according to the eigenvectors of the 2×2 covariance matrix of the bootstrap estimates. The radius in the direction of the first eigenvector is the $1-\alpha$ percentile of $\|\widehat{\kappa}_{0,b}^* - \overline{\kappa}_0^*\|$ for $b = 1, \ldots, B$. The radius in the direction of the second eigenvector is computed by substituting κ_0 with κ_2.

We use the average Euclidean distance between the bootstrap estimates and the ML estimate to compare the bootstrap methods. Define the random variable $\mathcal{D}$ as the Euclidean distance between the estimate $\widehat{\boldsymbol{\Delta}}_R$ and the parameter value $\boldsymbol{\Delta}_R$, i.e.,

$$\mathcal{D}(\mathrm{F}) = \|\widehat{\boldsymbol{\Delta}}_R(\mathrm{F}) - \boldsymbol{\Delta}_R(\mathrm{F})\|.$$

with expectation $\mathrm{E}(\mathcal{D}(\mathrm{F})) = \delta$. As indicated above, the parameter and its estimator is a function of F, the true distribution of the data. We estimate F by $\widehat{\mathrm{F}}$ through the bootstrap procedure. An approximation of $\mathcal{D}$ is then given as

$$\mathcal{D}(F) \approx \mathcal{D}(\widehat{F}) = \|\widehat{\boldsymbol{\Delta}}_R^* - \widehat{\boldsymbol{\Delta}}_R\|.$$

We estimate δ by

$$\widehat{\delta} = \frac{1}{B}\sum_{b=1}^{B} d_b, \tag{6}$$

where d_b is the Euclidean distance between bootstrap estimate b and the ML estimate. In the general case the distance will include both bias and variance. If the bootstrap estimates are unbiased, i.e., $\mathrm{E}(\widehat{\boldsymbol{\Delta}}^*) = \widehat{\boldsymbol{\Delta}}_R$, then a large value of δ indicates a large variance in the bootstrap estimates, and consequently a large variance of the original estimate.

The estimate of the relatedness coefficients will always be in the parameter space. The confidence region may extend outside the parameter space, which may not be desirable. A reparametrisation of the parameter space, making all estimates a part of the interior of the parameter space is theoretically possible. However, this is problematic, since many common pedigree relationships are in fact located on the border of the parameter space. Our solution to the problem is simply to truncate the confidence region. This approach resembles what is often done in applications as when a negative lower bound for the confidence intervals of a variance is replaced by 0.

3.4 Data and Implementation

The results in this paper are obtained using the programming language R. Most of the implementation is based on the *ped suite* libraries *ribd*, *forrel* and *ibdsim2* freely available from CRAN and described in the book [17]. Computation of pedigree coefficients, and visualisations in the IBD triangle, are provided by *ribd*. Simulation of marker genotypes are done in *forrel*, either by gene dropping through the pedigree or parametrically using the likelihood distribution. This library also provides maximum likelihood estimation of relatedness coefficients, and both parametric and non-parametric bootstrapping. Functionality for simulation of the IBD patterns in a pedigree is available in the library *ibdsim2*. Additional functionality for plotting and computation of confidence regions is available from the first author.

A set of 23 STR markers and allele frequencies based on the Norwegian population, available from https://familias.no/download, was used in Section 4.1 and for the computations for Figure 5 in Section 4.2. Genotype data in Section 4.4 are published in [18].

4 Results

We present the following results and examples: First, we give an example with simulated DNA profiles of two siblings. The goal is to compare expected and realised coefficients and parametric and non-parametric bootstrap. In the second example, we investigate and compare the mean deviation in the parametric and non-parametric bootstrap estimates as a function of the number of markers for different outbred relationships. We then investigate how the bootstrap methods behave with misspecified allele frequencies. At the end, estimation of the realised relatedness and bootstrapping for a inbred relationship, with both real and simulated data, is investigated.

4.1 From sibling pedigree to $\widehat{\boldsymbol{\kappa}}_R$ with confidence region

Consider two full siblings, with pedigree coefficients $\boldsymbol{\kappa} = (0.25, 0.5, 0.25)$, corresponding to the green asterisk in Figure 4. Using the R package *ibdsim2* a realistic, genome-wide IBD pattern between the siblings was simulated. This resulted in a realised relatedness of $\boldsymbol{\kappa}_R = (0.216, 0.588, 0.196)$, shown by the black triangle in the figure. Note that in a real case, both $\boldsymbol{\kappa}$ and $\boldsymbol{\kappa}_R$ would normally be unknown. Next, DNA profiles for 23 independent STR markers, were simulated for the siblings,

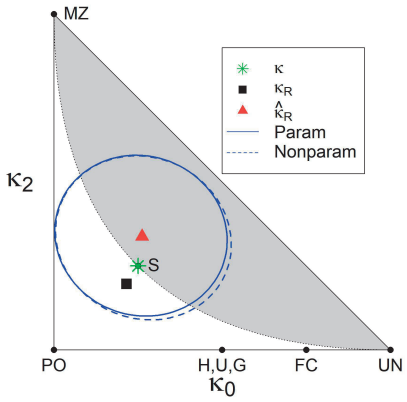


Fig. 4: Illustration for the example in Section 4.1 showing estimated realised relatedness with 95% confidence region for both parametric (solid line) and non-parametric bootstrap (dashed line). The pedigree relationship for siblings is indicated by a green asterisk, while a black square shows realised relatedness. The red triangle shows the parameter estimate $\widehat{\boldsymbol{\kappa}}_R$. Computed from 1000 bootstrap estimates.

conditionally on the IBD pattern already obtained. These profiles gave the estimate $\widehat{\boldsymbol{\kappa}}_R = (0.262, 0.4, 0.338)$, corresponding to the red triangle in Figure 4.

To assess the uncertainty of the estimate $\widehat{\boldsymbol{\kappa}}_R$ we performed parametric and non-parametric bootstrap, in both cases using $B = 1000$ replicates. The results are illustrated by the solid and dashed blue ellipses in Figure 4, which show 95% confidence regions for $\boldsymbol{\kappa}_R$. We note that the two bootstrapping procedures give very similar output in this case. In particular, both regions cover the true parameter $\boldsymbol{\kappa}_R$. The realised relatedness of parent-offspring (PO), half sibling (H), avuncular (U), grandparent-grandchild (G), first cousin (FC) relationships and unrelated (UN) individuals are by definition located along the first axis of the IBD triangle, i.e, outside the confidence regions.

4.2 Mean deviation of bootstrap estimates

The results in this section compare the two bootstrap procedures by assessing the average Euclidean distance between the estimate $\widehat{\boldsymbol{\kappa}}_R$ and the bootstrap estimates. Two DNA profiles, genotyped for 23 STR markers described in Section 3.4, are simulated, and an estimate of δ, as given in (6), is computed from $B = 1000$

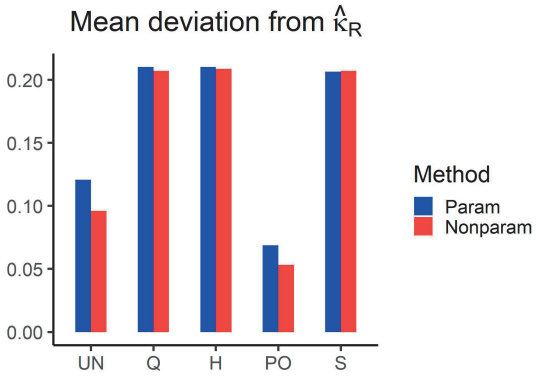


Fig. 5: Average euclidean distance between $\widehat{\boldsymbol{\kappa}}_R$ and bootstrap estimates for different $\boldsymbol{\kappa}_R$, as described in Section 4.2, for parametric (blue) and non-parametric (red) bootstrap.The values are the average over 200 simulated DNA profiles with 23 STR markers, each bootstrapped with $B = 1000$. The five set of bars correspond to UN: $\boldsymbol{\kappa}_R = (1, 0, 0)$, Q: $\boldsymbol{\kappa}_R = (0.531, 0.438, 0.031)$, H: $\boldsymbol{\kappa}_R = (0.5, 0.5, 0)$, PO: $\boldsymbol{\kappa}_R = (0, 1, 0)$ and S: $\boldsymbol{\kappa}_R = (0.25, 0.5, 0.25)$.

bootstrap estimates. This is then repeated 200 times, each time simulating new DNA profiles, and the average of $\widehat{\delta}$ over these 200 simulations is computed. Figure 5 shows this average distance for five different values of the realised relatedness: UN: $\boldsymbol{\kappa}_R = (1, 0, 0)$, Q (quadruple half first cousins): $\boldsymbol{\kappa}_R = (0.531, 0.438, 0.031)$, H: $\boldsymbol{\kappa}_R = (0.5, 0.5, 0)$, PO: $\boldsymbol{\kappa}_R = (0, 1, 0)$ and S: $\boldsymbol{\kappa}_R = (0.25, 0.5, 0.25)$. The blue and red bars correspond to parametric and non-parametric bootstrap, respectively. Table 1 shows the mean of the bootstrap estimates, averaged over the 200 simulations, for the two bootstrapping methods.

From Figure 5 we see that there is a difference between the bootstrap methods at the corner points UN and PO. Parametric bootstrap has a higher mean deviation than the non-parametric method. Furthermore, we see from Table 1 that the parametric method seems to be more biased than the non-parametric for PO. This indicates that the methods behave differently for this boundary point. However, the bootstrapping methods do not differ much for the boundary point H. The height of the bars in Figure 5 are about the same, as well as the means in Table 1. The bootstrap procedures seem to behave similarly for the interior points Q and S.

Next, we compare the mean deviation for parametric and non-parametric bootstrap, as the number of markers increase from 5 to 200. Each marker has 10 alleles, with allele frequencies randomly sampled between 0 and 1, and scaled to sum to 1.The comparisons are performed for the relationships UN, H, PO, and

Tab. 1: Mean of 1000 bootstrap estimates, averaged over 200 simulations for realised relatedness as explained in caption of Figure 5 and in Section 4.2, for both parametric and non-parametric bootstrap.

Rel.	Method	$\overline{\kappa}_0^*$	$\overline{\kappa}_1^*$	$\overline{\kappa}_2^*$
UN	Param	0.891	0.086	0.022
UN	Nonparam	0.909	0.074	0.017
Q	Param	0.540	0.398	0.061
Q	NonParam	0.537	0.407	0.056
H	Param	0.509	0.442	0.049
H	Nonparam	0.505	0.454	0.040
PO	Param	0	0.941	0.059
PO	Nonparam	$1.090 \cdot 10^{-6}$	0.952	0.048
S	Param	0.244	0.512	0.244
S	Nonparam	0.244	0.512	0.244

S. The blue and red line in Figure 6 show mean deviation for parametric and non-parametric bootstrap, respectively, with $B = 1000$.

For UN and PO and few markers, the mean deviation is somewhere between 0.1 and 0.2. The curves flattens out at about 50 markers, with a mean deviation of about 0.05. For H and S, the parametric and non-parametric bootstrap perform similarly. The mean deviation is higher for the low number of markers, and flattens out to a value of about 0.1 for 100 markers and more.

The findings shown in Figure 5 and 6 support each other: The bootstrapping methods behave similarly for the interior point S and the boundary point H, but differ for the corner points PO and UN.

4.3 Coverage probabilities for φ_R and misspecified allele frequencies

It is important for the likelihood model that the correct allele frequencies are used. If individuals originate from two different populations or the population has a strong subpopulation structure, accurate specification of allele frequencies is difficult. Frequencies of alleles based on ancient DNA [18] may also be hard to specify.

Table 2 shows coverage probabilities, mean values, and estimated bias for the estimate of the kinship coefficient $\widehat{\varphi}_R$, using parametric and non-parametric bootstrap. The coverage probabilities are computed from 200 simulations. For each simulation a 95% percentile bootstrap confidence interval is constructed from $B = 1000$ bootstrap estimates. The kinship coefficient is estimated through

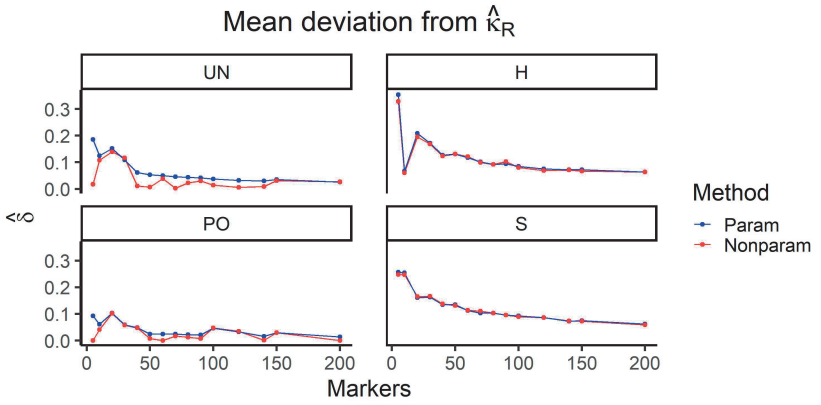


Fig. 6: The estimate $\hat{\delta}$ from 1000 bootstrap estimates, as a function of the number of markers, as described in Section 4.2. Each marker has 10 alleles with random allele frequencies. Blue and red lines show parametric and non-parametric bootstrap, respectively.

estimation of the IBD coefficients. The analysis is performed for the same coefficients as in Figure 5, i.e., UN, Q, H, PO and S. The corresponding kinship coefficients are $\varphi_R = 0$ for UN, $\varphi_R = 0.125$ for Q and H and $\varphi_R = 0.25$ for PO and S. The 23 STR markers described previously were used.

The two bootstrap methods have about the same coverage probabilities. The realised relatedness corresponding to Q and S have coverage less than 95%, while the coverage probability for H coincides with the specified coverage of 0.95. The coverage is about 1 for PO and UN. Importantly, there is a difference in the coverage probabilities for IBD coefficients that have the same kinship coefficients.

The same analysis is repeated, but a wrong set of random allele frequencies are used in the estimation and bootstrapping. Table 3 shows the results. The mean and bias is about the same for the two methods, however, the coverage probabilities differ between the methods. The estimated coverages are in both cases quite far from the target values. Non-parametric bootstrap shows lower coverage than the parametric when the allele frequencies are misspecified. Each non-parametric bootstrap estimate is computed with the original genotype data, but with the wrong allele frequencies. In the parametric situation, the original estimate is computed with the wrong allele frequencies. The parametric bootstrap estimate is then computed with simulated genotypes and the allele frequencies used to simulate them. Furthermore, it is worth noticing that the coverage probabilities for UN is close to zero for misspecified allele frequencies. The differences in coverage

Tab. 2: Mean, estimated bias and coverage probabilities for φ_R, for the realised relatedness described in the caption of Figure 5, for parametric and non-parametric bootstrap. The simulations are explained in Section 4.3.

Rel	Boot	φ_R	Mean	Bias	CovProb
UN	Param	0.000	0.0328	0.0328	0.98
UN	Nonparam	0.000	0.0271	0.0271	0.98
Q	Param	0.125	0.1303	0.0053	0.90
Q	Nonparam	0.125	0.1299	0.0049	0.92
H	Param	0.125	0.1350	0.0100	0.95
H	Nonparam	0.125	0.1337	0.0087	0.95
PO	Param	0.250	0.2648	0.0148	1.00
PO	Nonparam	0.250	0.2620	0.0120	1.00
S	Param	0.250	0.2500	0.0000	0.92
S	Nonparam	0.250	0.2501	0.0001	0.93

probabilities between the bootstrap methods in this case is based on one misspecified model. Further studies are needed before more firm conclusions can be drawn.

4.4 Jacquard coefficients with confidence intervals

In this example, we shift our attention to the Jacquard coefficients. We start with a data set from the ancient Egypt. The work of Hawass et al. [18] resulted in DNA profiles from several mummies, including the remains of Tutankhamun and his presumed grandfather Amenhotep III. The pedigree in Figure 7 shows how the two mummies are assumed to be related. Tutankhamun's parents were siblings, leaving him with a quite high degree of inbreeding. The Jacquard coefficients describing this pedigree relationship is $\boldsymbol{\Delta} = (0, 0, 0, 0, 0.125, 0.125, 0.125, 0.5, 0.125)$.

The DNA profiles are genotyped for 8 STR loci. Estimation of the Jacquard coefficients with only 8 markers, results in quite high uncertainty. The left panel of Figure 7 shows the ML estimate of $\boldsymbol{\Delta}_R$, with 95% marginal confidence intervals constructed with parametric bootstrap (blue) and non-parametric bootstrap (red). The black dots denote the point estimate for each coefficient, $\widehat{\boldsymbol{\Delta}}_R = (0, 0, 0, 0, 0, 0.160, 0.030, 0.809, 0)$. Parametric bootstrap creates in general larger intervals than non-parametic bootstrap. The coefficients Δ_4 and Δ_6 correspond to the Jacquard states J_4 and J_6 in Figure 2. These states indicate IBD alleles within each individual, meaning that an individual is inbred. The first of them, Δ_4 is estimated to 0, with a marginal confidence interval of zero width, indicating that one of the individuals are not inbred. The other, Δ_6, is estimated to 0.160, indicating that the other individual is inbred. The value 0 is contained in the

Tab. 3: Mean, bias and coverage probabilities for φ_R, for the realised relatedness described in caption of Figure 5, for parametric and non-parametric bootstrap with misspecified allele frequencies. The simulations are explained in Section 4.3

Rel	Boot	φ_R	Mean	Bias	CovProb
UN	Param	0.000	0.1178	0.1178	0.08
UN	Nonparam	0.000	0.1167	0.1167	0.04
Q	Param	0.125	0.1903	0.0653	0.62
Q	Nonparam	0.125	0.1901	0.0651	0.59
H	Param	0.125	0.2030	0.0780	0.42
H	Nonparam	0.125	0.2023	0.0773	0.38
PO	Param	0.250	0.2861	0.0361	0.76
PO	Nonparam	0.250	0.2856	0.0356	0.66
S	Param	0.250	0.2948	0.0448	0.76
S	Nonparam	0.250	0.2949	0.0449	0.72

confidence interval. Furthermore, it is worth to notice that parametric bootstrap gives rise to confidence intervals for the coefficients Δ_1, Δ_2 and Δ_3, unlike for the non-parametric bootstrap.

Next, we use simulated data and look at how the CIs change with the number of markers, for the two different bootstrap methods. Two DNA profiles are simulated conditionally on $\boldsymbol{\Delta}_R = (0, 0, 0, 0, 0.125, 0.125, 0.125, 0.5, 0.125)$, are shown by the black squares in Figure 8. One simulation is performed for a set of 10 markers, one for 20 markers and one for 50 markers. Each marker has 10 alleles with allele frequencies sampled at random between 0 and 1. From each of these simulations, parametric and non-parametric bootstrap are performed. Figure 8 shows marginal CIs for 10, 20 and 50 markers, indicated with the color red, green and blue, respectively. The left and right panel of the figure show results for parametric and non-parametric bootstrap, respectively. The colored circles show the estimate $\widehat{\boldsymbol{\Delta}}_R$ for each of the three simulated sets of DNA profiles.

As expected, the width of the confidence intervals decreases as the number of loci increases. Parametric and non-parametric bootstrap give about the same confidence intervals, except for the coefficients estimated to 0. In general, parametric bootstrap creates intervals for these coefficients, while the non-parametric version does not. This coincides with the results shown in Figure 7.

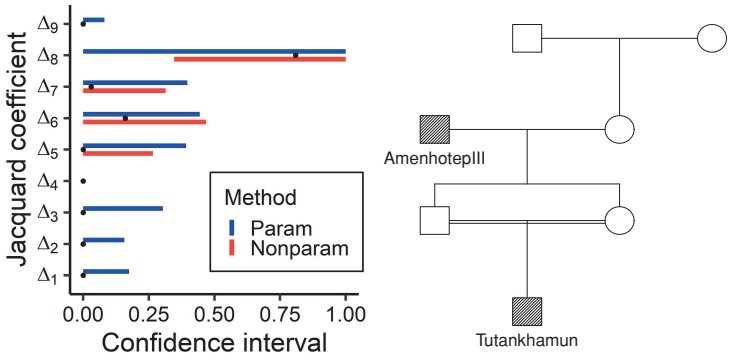


Fig. 7: Estimate of $\boldsymbol{\Delta}_R$ between Amenhotep III and Tutankhamun, genotyped at 8 loci, with 95% marginal CI for each coefficient, from 1000 bootstrap estimates, as described in Section 4.4. Black dots show the point estimate $\widehat{\boldsymbol{\Delta}}$.

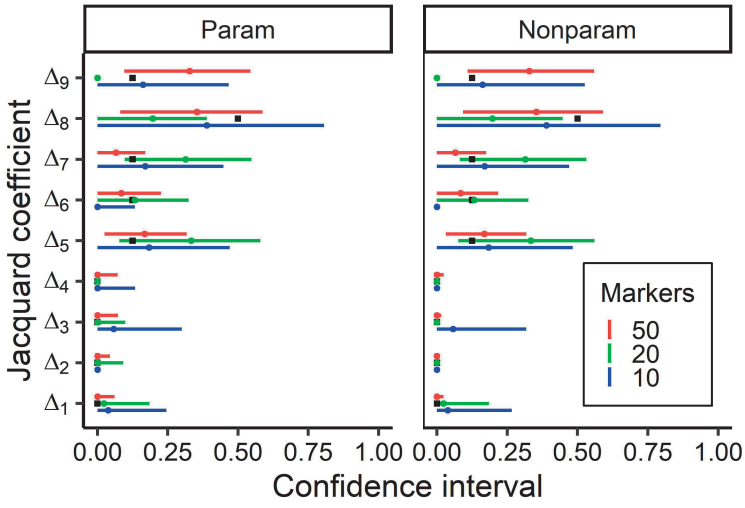


Fig. 8: The Figure shows 95% marginal CIs from 1000 parametric (left) and non-parametric (right) bootstrap estimates, as described in Section 4.4. The realised relatedness is $\boldsymbol{\Delta}_R = (0, 0, 0, 0, 0.125, 0.125, 0.125, 0.5, 0.125)$, corresponding to the black squares. The colored dots denote the estimate $\widehat{\boldsymbol{\Delta}}_R$ for the different number of markers.

5 Discussion

The work in this paper discusses the use of bootstrap methods to assess the uncertainty in the maximum likelihood estimate of the kinship coefficient, IBD coefficients and Jacquard coefficients. We estimate the realised relatedness between pairs of individuals, i.e., the proportion of the genomes of two individuals in the different IBD states. A pair of individuals may fail to prove a certain kinship if their realised relatedness differs substantially from the pedigree counterpart. The realised relatedness is not subjected to the same constraints in relatedness coefficients as the pedigree relationship. Thus, the likelihood of the realised relatedness is maximized in a bigger parameter space than previously done in the literature [5, 7, 6].

Which bootstrap procedure to apply leads to a discussion on how to define the distribution of the data. Two definitions seem reasonable to consider:

i The genotype at each possible locus along the genome is fixed. The loci considered in the estimation are random. The population, for which $\boldsymbol{\Delta}_R$ is a property, is the two genomes we consider.
ii The loci used in the estimation are fixed, but the genotypes observed at each locus are random. All pairs of individuals with the same $\boldsymbol{\Delta}_R$ are the population.

In definition i), loci considered in the estimation are drawn randomly. One set of loci, with corresponding genotypes, gives one estimate, while another set of loci will give another estimate. This is then the variation we want to capture. This is in several aspects equivalent to survey sampling. A problem with definition i) is how to define the sample space of the data. Is it the set of all unlinked STR markers along the genome? The size of this set is small, maybe limited to the loci for which the DNA profiles are genotyped, and if so, it is not possible to obtain another estimate and there is no variation. Furthermore, definition i) does not comply with the likelihood function used in the ML estimation. By this, we mean that the probability of each joint genotype along the genome do not sum to one. If the joint genotypes along the genomes were the only possible genotypes to observe, they should sum to one.

By definition ii), the loci to consider are fixed. The sample space at a locus is the set of all joint genotypes possible to observe. For instance, for a marker with two alleles a and b, all possible joint genotypes for this locus, not considering ordered pairs, is (aa, aa), (bb, bb), (aa, bb), (aa, ab), (bb, ab) and (ab, ab). The distribution (5) sums to one over all ordered pairs of joint genotypes at the locus. Then, conditioned on $\widehat{\boldsymbol{\Delta}}_R$, one set of DNA profiles, genotyped for M loci is more, or less, probably to observe than another profile.

Of the two definitions above, the second one, definition ii), seems to be the proper way to define the data and their probability distribution, when we use (3) for the ML estimation. The source of the variation in the estimate of $\boldsymbol{\Delta}$ is the possibility to observe different genotypes at a locus. We therefore argue that parametric bootstrap captures this source of variation, while non-parametric does not.

Even though we state that, from a theoretical point of view, parametric bootstrap is the one to prefer, the results in Figure 5 and 6 show that the mean deviation of the bootstrap estimate from the point $\widehat{\boldsymbol{\kappa}}_R$ is about the same for the two bootstrap methods. For the corner points $\boldsymbol{\kappa}_R = (0, 1, 0)$ and $\boldsymbol{\kappa}_R = (1, 0, 0)$ the mean deviation is slightly higher for the parametric bootstrap procedure, indicating higher variance of the bootstrap estimate and/or more bias in the bootstrap estimate. The same analysis as for Figure 6 was also performed for markers with 10 equifrequent alleles and markers with 9 equifrequent alleles and one allele with low frequency. The mean deviation showed the same behaviour. In addition, a modified parametric bootstrap was tested. The DNA profile for one individual was kept fixed, while genotypes for the other individual were parametrically sampled, conditioned on the genotypes of the first individual. This gave about the same mean deviation as the ordinary parametric bootstrap. Further investigation is needed to conclude regarding the use of this modified parametric bootstrap method. It is not possible to determine the bootstrap procedure that gives the best approximation of the true variance of the estimate $\widehat{\boldsymbol{\kappa}}_R$, as the true variance is unknown. In principle, it would be possible to perform a simulation study to estimate the true variance. However, it is not obvious how such a study could be implemented.

The importance of correctly specified allele frequencies should be noted. Incorrect allele frequencies lead to a misspecified likelihood model. Table 3 shows coverage probabilities for the kinship coefficient when allele frequencies are misspecified. The coverage probabilities for both bootstrap procedures are affected, but the non-parametric method seems to be more severely affected.

The marginal confidence intervals for the Jacquard coefficients shown in Figure 7 and 8 show an interesting difference between the parametric and non-parametric bootstrap. In general, parametric bootstrap gives rise to confidence intervals for the coefficients that have estimate zero, unlike for the non-parametric bootstrap. If this is a feature of the particular example presented in Section 4.4 is not known, but it is a topic worth investigating further, as the results clearly differ between the bootstrap methods.

The assumption of HWE for the likelihood, is not unique for the work in this paper. Discussions of this topic is given in the literature [7, 19, 20]. However, deviation from HWE can be adjusted for in the genotype frequencies of Table 4. The bootstrap methods were affected differently by misspecified allele frequencies,

and it is therefore reasonable to speculate that the methods are affected differently if the population is out of HWE.

Standard forensic kits with say 23 markers may suffice for LR based testing of close relationships. However, more data is typically needed for estimation. Figure 4 shows a confidence region for the IBD coefficients, for a set of 32 STR markers. The region occupies quite a big area of the IBD triangle, indicating a large uncertainty in the parameter estimate. The independence assumption for the likelihood (3) may be violated when the number of markers is extended beyond 50. An interesting topic for further work is to investigate the effect of ignoring linkage on both the point estimate of the ML estimator and the corresponding confidence region.

The main challenge of the work in this paper is that we do not know the true distribution of the ML estimate, hence, we are not able to conclude which of the bootstrap methods that is more correct. However, we think it is reasonable to regard parametric bootstrap as the appropriate method for assessing the uncertainty of estimates of the relatedness coefficients.

References

[1] B. S. Weir, A. D. Anderson, and A. B. Hepler. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, 7(10):771–780, 2006.

[2] 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(3-4):223–231, 2007.

[3] M. Prinz, A. Carracedo, W. R. Mayr, N. Morling, T. J. 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.

[4] 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.

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

[6] B. G. Milligan. Maximum-likelihood estimation of relatedness. *Genetics*, 163(3):1153–1167, 2003.

[7] A. D. Anderson and B. S Weir. A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, 176(1):421–440, 2007.

[8] A. Jacquard. Genetic information given by a relative. *Biometrics*, pages 1101–1114, 1972.

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

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

[11] E. A. Thompson. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, 2013.

[12] W. G. Hill and B. S. Weir. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, 93(1):47–64, 2011.

[13] E. L Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[14] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[15] S. Ghosh and A. M. Polansky. Smoothed and iterated bootstrap confidence regions for parameter vectors. *Journal of Multivariate Analysis*, 132:171 – 182, 2014.

[16] P. Hall. On the bootstrap and likelihood-based confidence regions. *Biometrika*, 74(3):481–493, 1987.

[17] M. D. Vigeland. *Pedigree analysis in R*. Academic press, 2021.

[18] Z. Hawass, Y. Z. Gad, S. Ismail, R. Khairat, D. Fathalla, N. Hasan, A. Ahmed, H. Elleithy, M. Ball, F. Gaballah, et al. Ancestry and pathology in King Tutankhamun's family. *Jama*, 303(7):638–647, 2010.

[19] M. D. Brustad, H. K.and Vigeland and T. Egeland. Pairwise relatedness testing in the context of inbreeding: expectation and variance of the likelihood ratio. *International Journal of Legal Medicine*, 135(1):117–129, 2021.

[20] D. Kling, T. Egeland, and A. O. Tillmar. FamLink–a user friendly software for linkage calculations in family genetics. *Forensic Science International: Genetics*, 6(5):616–620, 2012.

A Table of genotype probabilities

Tab. 4: The conditional probability $P(g \mid J_i)$ of a pair of genotypes $g = (g_A, g_B)$, given a Jacquard state J_i. The symbols a, b, c and d represent different alleles, with population frequencies p_a, p_b, p_c and p_d respectively.

g	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
(aa, aa)	p_a	p_a^2	p_a^2	p_a^3	p_a^2	p_a^3	p_a^2	p_a^3	p_a^4
(aa, bb)	0	$p_a p_b$	0	$p_a p_b^2$	0	$p_a^2 p_b$	0	0	$p_a^2 p_b^2$
(aa, ab)	0	0	$p_a p_b$	$2p_a^2 p_b$	0	0	0	$p_a^2 p_b$	$2p_a^3 p_b$
(aa, bc)	0	0	0	$2p_a p_b p_c$	0	0	0	0	$2p_a^2 p_b p_c$
(ab, aa)	0	0	0	0	$p_a p_b$	$2p_a^2 p_b$	0	$p_a^2 p_b$	$2p_a^3 p_b$
(bc, aa)	0	0	0	0	0	$2p_a p_b p_c$	0	0	$2p_a^2 p_b p_c$
(ab, ab)	0	0	0	0	0	0	$2p_a p_b$	$p_a p_b (p_a + p_b)$	$4p_a^2 p_b^2$
(ab, ac)	0	0	0	0	0	0	0	$p_a p_b p_c$	$4p_a^2 p_b p_c$
(ab, cd)	0	0	0	0	0	0	0	0	$4p_a p_b p_c p_d$

ISBN: 978-82-575-1819-6
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no

110406 / ANDVORDGRAFISK.NO