Computational challenges in family genetics

Beregningsproblemer i familiegenetikk

Philosophiae Doctor (PhD) Thesis

Daniel Kling

Department of Chemistry, Biotechnology and Food Science
Faculty of Veterinary Medicine and Biosciences
Norwegian University of Life Sciences

Ås 2015

Thesis number 2015:36
ISSN 1894-6402
ISBN 978-82-575-1284-2

List of papers

I. D. Kling, J. Welander, A. Tillmar, Ø. Skare, T. Egeland and G. Holmlund, ***DNA microarray as a tool in establishing genetic relatedness - Current status and future prospects.*** Forensic Science International Genetics 6 (2012) 322-329.

II. D. Kling, T. Egeland and P. Mostad, ***Using Object Oriented Bayesian Networks to Model Linkage, Linkage Disequilibrium and Mutations between STR Markers***. PLoS One 7 (2012) e43873

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

IV. 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 (2014) 1-12

V. D. Kling, A. O. Tillmar and T. Egeland, ***Familias 3 - Extensions and new functionality***. Forensic Science International: Genetics 13 (2014) 121-127

VI. D. Kling, B. Dell'Amico and A. O. Tillmar, ***FamLinkX - A general approach to likelihood computations for X-chromosomal markers***. Forensic Science International: Genetics (2015, submitted)

Summary

There is a constant demand to determine the most probable relationship between a set of person given some genetic marker data and some hypotheses about pedigree structure. A constant stream of paternity cases is obtained at forensic laboratories around the globe and with the modernization of many underdeveloped countries the increase in a few years may be staggering. The case may be as trivial as to find out who is the true father of a child, but also more complex, as to large inbred pedigrees. In addition, cases may involve only two persons, e.g. an alleged father and a child, but also many persons, e.g. several cousins, aunts/uncles and siblings. Furthermore we may be looking at single cases, but also large scale disaster victim identification (DVI) problems. In the latter, identification through the use of DNA has risen to become the most important and reliable tool.

With the arrival of new technologies, e.g. high density SNP microarrays and next generation sequencing, more and more genetic markers become available. Although providing opportunities they also present forensic scientists with great statistical problems as independence can no longer be assumed. This high-dimensionality problem is something recurring in all fields working with genetics and the solution is in many cases reduction of dimensionality using well established methods. However, in forensic genetics, evidence in general requires a likelihood ratio to be established, weighting the genetic evidence given hypotheses against each other. Therefore the dimensionality reduction cannot generally be applied and we need other methods to handle the dependency. One approach adopted in many situations when dependence is modeled, is Markov chains. The property of such chain relies on the fact that given the value of one node, e.g. one genetic marker, the values of the subsequent nodes in the chain is independent of all previous nodes. Variants of Markov chains will be a focus in this thesis.

With the surge of increasing computational power, simulations have become a crucial tool in many fields of research. We may now study the effects of something random using complex models and investigate the outcome with little of thought on the computation time. In forensic genetics, simulations have many possible applications. For instance, in determination of relationships, we may simulate the outcome of a case and study the distribution of probabilities in order to determine the false positive/negative rates given some probability threshold. Simulations may also be used to study how the change in some parameter in our model affects the evidence value.

In summary, this thesis describes means to solve complex computational problems arising when independence between genetic markers cannot be assumed. It further considers solutions to other statistical obstacles encountered in forensic genetics such as DVI operations, simulations and models for mutations. Different approaches are discussed and evaluated. Moreover, software is presented implementing the ideas and algorithms.

Sammanfattning

Att bestämma det biologiskt mest sannolika släktskapet, baserat på genetisk data för ett antal individer, är något som ständigt intresserar människor. Ett konstant flöde av faderskapsfall tas emot och analyseras på forensiska labb runt om i världen och med den intensiva tekniska utvecklingen i U-länder kan vi bara ana en explosion av antalet ärenden de kommande åren. Det enklaste fallet är att bestämma om en man är far till ett barn, men även mer komplexa fall, där invecklade släktskap skall utredas, blir allt vanligare. Det kan vidare vara enkla isolerade fall men också stora olyckor, där flera aspekter måste tas hänsyn till. I identifieringsprocessen som följer större masskatastrofer har DNA blivit den primära och säkraste metoden att använda.

Den tekniska utvecklingen har introducerat flera nya metoder där det är möjligt att erhålla data från en stor mängd genetiska markörer billigt och på kort tid. Mer data förbättrar generellt urskiljningsförmågan, men medför dock flera statistiska problem som måste modelleras; det kanske viktigaste är beroendet mellan enskilda beräkningar. Mångdimensionalitetsproblem är ett känt fenomen inom statistik och hanteras ofta genom reduktion av antalet dimensioner medelst etablerade metoder. Dessa tillvägagångssätt kan inte med samma självklarhet användas i forensisk statistik, givet de förutsättningar som föreligger. Vi behöver andra metoder för att hantera och modellera beroendet mellan beräkningarna. Ett vanligt tillvägagångssätt är att använda så kallade Markov-kedjor. Dessa kedjor har egenskapen att givet beräkningar/värden för en nod i kedjan så är alla senare beräkningar oberoende av tidigare beräkningar. Markov-kedjor är ett centralt tema i denna avhandling.

I enighet med Moores lag utvecklas beräkningskapaciteten hos datorer exponentiellt och som en följd har tunga beräkningar och simuleringar avsevärts förenklats. Detta har i sin tur haft som konsekvens att komplicerade modeller kan studeras med hjälp av de sistnämnda utan att ägna en tanke åt kapacitetsproblem. I forensisk genetik kan vi använda simuleringar för att studera fördelningar hos olika parametrar. Till exempel kan vi erhålla en summering av förväntade bevisvärden i ett specifikt släktskapsärende under givna förutsättningar. Vi kan undersöka hur många personer vi behöver inkludera i ärendet och hur många genetiska markörer vi behöver analysera. Detta är mycket användbart då vi på förhand kan avgöra om vi har möjlighet att lösa ett ärende eller ej.

Sammanfattningsvis presenterar denna avhandling metoder och implementeringar för att lösa flera komplexa beräkningsproblem som uppkommer när kopplade genetiska markörer används. Den beskriver också lösningar på andra statistiska problem inom forensisk genetik såsom modeller för mutationer och matchningsalgoritmer vid större identifieringsarbeten samt simuleringar. Varje lösning implementeras också i fritt tillgänglig programvara för att vara ett enkelt hjälpmedel för andra forskare inom fältet.

Acknowledgements

I am indebted to my head supervisor, Thore Egeland, sharing is deep knowledge within the forensic field as well as showing me the path when I was stuck in a corner. I would further like to thank my co-supervisor, Petter Mostad, with whom I have had many fruitful discussions and who has helped my develop many of the mathematical ideas presented in this thesis.

Thanks to Andreas Tillmar and Gunilla Holmlund at the Forensic genetics laboratory in Linköping, Sweden; the former being co-author on many of the papers and providing numerous ideas and pushing me forward, the latter for introducing me to the forensic field and helping me through my Master thesis, subsequently leading to this PhD thesis.

I am very grateful for all the support from the paternity laboratory in Oslo, discussing ideas and real casework implementations. Their input has been truly valuable when developing implementations and software. A special thanks to Marguerethe Stenersen, head at the laboratory, for making this PhD possible and providing support and thoughtful comments during the thesis. Thanks also to colleagues at NMBU (Norwegian University of Life Sciences), giving me ideas and food for thought during the thesis.

I am also grateful to my family and friends for actually reading my papers and asking me questions needing a good and easy-to-grasp answer.

Also a thank goes out to all users of my software and participants of courses I held during the thesis, who all constantly gave, and still give, me feedback and helps me improve my implementations.

Last, and most importantly, thanks for the support from my love, Sara, helping me struggle through all the obstacles encountered during this thesis and in addition, final proof reading of this introduction.

Contents

1 Introduction

Since the discovery of the DNA helix by Francis Crick, James Watson and Rosalind Franklin in the early 1950s, the research on our genetic material has exploded. Even before that, work by Mendel and others provided insights into our inheritance patterns and there are still mysteries being uncovered concerning the elaborate mechanisms governing our cells. A myriad of different research fields benefit from this progression, not the least medical genetics, aiding humanity in the struggle against diseases and genetic disorders. The focus of this thesis will be on a field known as forensic genetics. The word *Forensic* is derived from Latin and means “before the forum” and relates to the times of the Roman Empire when criminal cases were presented to the public (forum). Modern use of the word is commonly connected to the investigation of any evidence in a case presented before a court of law. The following sections will introduce the readers to forensic genetics. More specifically the thesis will focus on statistical problems encountered when performing calculations on genetic relatedness.

It is fascinating how people constantly wish to find their biological relationships and establish the genetics that bonds us together. In Norway alone, the number of relationship cases approximates 2000 each year [Personal experience]. This in a population that is, in a larger context, small, only about 5 million. Without specific knowledge about the same numbers in other countries, we can, based on the global population of 7 billion, roughly estimate the number of annual paternity cases world-wide to 2 million. This is of course only a crude estimate and we know for a fact that some countries have considerably lower number of cases whereas still some countries may have higher levels.[1]

Throughout history, disputed relationships have given rise to a number of intriguing feuds. From the first book of Kings (1 Kings 3:16-18) in the Bible we learn about possibly one of the first cases of disputed maternity. To briefly recapitulate, two women are presented to the wise King Solomon, both alleging to be the mother of a child. According to the lore, no evidence is held forward favoring either of the two women. The King sees no other option but to bring forward a sword and cut the baby in two, thus leaving each mother with a part. One of the women exclaims: “Please don’t kill my son, Your Majesty, I love him very much, but give him to her. Just don’t kill him”, while the other woman replies, “Go ahead and cut him in half. Then neither of us will have the baby”. The King is

[1] The actual number is probably considerably lower, since the extent of paternity testing in some highly populated countries is substantially smaller

wise and decides not to cut the baby in halves, but proclaims the first mother to be the true mother as she was indeed willing to sacrifice her maternity to let the baby live.

A more recent example, and perhaps more relevant in the current thesis, is the infamous case of the Romanov family [1]. The last Russian tsar and his family were allegedly killed by the Russians during the revolution, but no bodies were ever found. In 1991 a family, that could possibly be the remains of the Romanovs, was found buried in Ekaterinburg, Russia. Extensive investigations were undertaken leading to several papers [1-4], where the final conclusion was that there was a high probability of the remains actually being the Romanovs. The DNA evidence suggested that all the skeletons in the grave belonged to one family and that living distant relatives of the Romanovs matched up with the Tsar and Tsarina.

Another interesting example is the search for descendants of Thomas Jefferson, the third president of the United States. He allegedly had a child (or several children), with one of his maids, Sally Hemmings (who was a slave). This is a controversy dating back to the early 19th century when suggestions were brought forward that Jefferson had fathered one or more of Hemmings' children. The arrival of DNA technology shed new light on the discussion as a perfect match for the Y chromosome (inherited unchanged through the male line), was found between descendants of Jefferson and Hemmings [5-7]. The case has not yet reached a final conclusion as the genetic evidence only points out that Jefferson or a male relative of him is likely to be the father, although other evidence does suggest paternity as well.

In addition, more recent events include the identification of victims from mass disasters. For instance, the application of DNA played a crucial part in the identification process following the 9/11 WTC terror attack [8-10] and the South Asia tsunami disaster in 2004 [11]. In the same field, large projects are undertaken to identify victims from recent wars, e.g. the First and Second World War as well as mass graves on several sites on the Western Balkan Peninsula.

The use of biological markers to determine paternity was introduced using blood groups (ABO system) in the early 1920s. If inconsistent groups were observed for the father and the child, paternity could be excluded. However, the general exclusion rate was fairly poor since the probability to exclude for some blood groups is very low. Developments led to the introduction of serological markers with higher discrimination in the 1930s and HLA markers, which were the first real genetic markers, in the 1960s with even higher discrimination. The arrival of polymerase chain reaction (PCR) in the late 1970s led to a revolution when DNA could be amplified to virtually unlimited amounts [12]. Still ongoing developments have led to the possibility of obtaining the complete genetic setup from a biological sample using next generation sequencing techniques [13-15].

The broad motivation for this thesis is the computational obstacles encountered in forensic genetics, more specifically in family genetics, see Figure 1. It is convenient to make the following division,

1. Models for population effects
2. Models for pedigrees and family structures
3. Models for observation levels effects

The distinction between the first two points is not always easy as at some point we were all related, i.e. we all belong to a common founder or seen from another perspective, a giant pedigree. From the words of famous biologist Richard Dawkins; given an individual sufficiently long ago in time, either he/she is related to all now living individuals or none [16]. Nevertheless, for our purpose, we must at some point make a decision on where to put the limitation and what to model as something random from a population and what we like to incorporate into the pedigree. We will see that this is a topic recurring throughout the thesis and examples from each of the above mentioned points will be discussed.

In order to fully explain the scope of this thesis and the papers we need to define some of the important concepts dealt with in forensic genetics. The selection herein is not complete as there are for example numerous population genetic effects that could be described. The topics are chosen such that they reflect the research conducted in the papers.

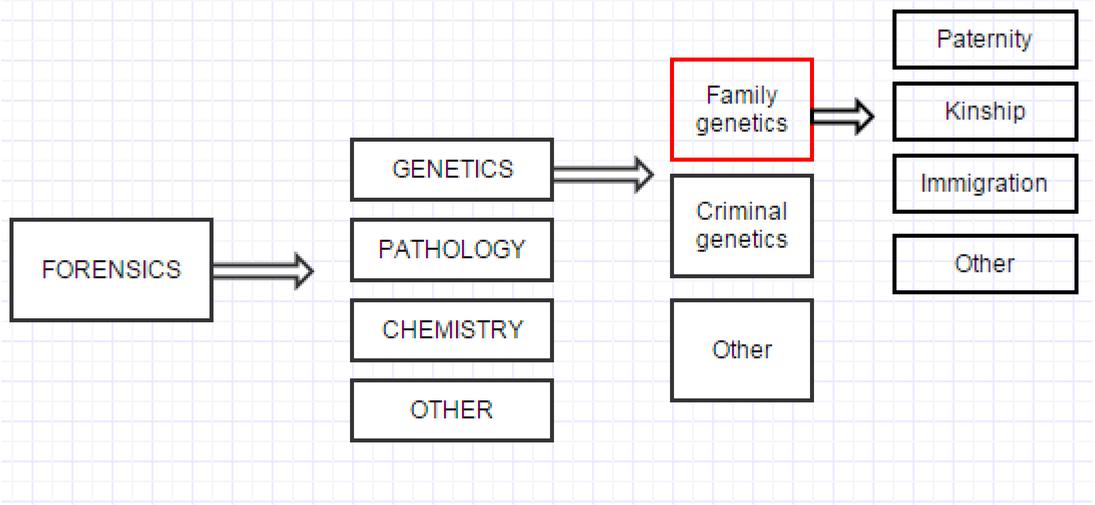

Figure 1. Flowchart illustrating the position of Family genetics in the forensic field.

1.1 Background

As mentioned in the introductory text, the analysis of our genetic code, i.e. our DNA, has provided new insights into several fields; e.g. in medical genetics to find genes associated with certain disorders, in animal genetics to establish the inbreeding and the purity of species, in evolutionary genetics to trace origins, and most importantly for this thesis, to establish the relatedness between individuals.

1.1.1 Genetic markers

The ground for the investigation of our DNA is the occurrence of genetic markers along the chromosomes [17, 18]. Genetic markers are defined as positions on the chromosomes that can be found in a majority of the population and where different variants can be observed. The degree of variation at a marker is known as its polymorphism. Consider, for instance, chromosome 1 in all individuals in the world. The first position on this chromosome may consist of an Adenine (A) base in 60% of the individuals while the remaining 40% has a Guanine (G) base on the same position. This is called a genetic marker and the specific example illustrates a single nucleotide polymorphism (SNP). In forensic genetics, "variable number of tandem repeat" (VNTR) markers are often used. More specifically, short tandem repeat (STR) markers are most commonly investigated [19]. They consist of specific genetic sequences, e.g. AAGA, occurring with a certain number of repeats. The STR markers are favorable since they are usually highly polymorphic, i.e. there are a lot of variants better known as alleles [20]. This in turns makes it unlikely that two unrelated individuals share some alleles by chance, compared to, for instance SNP markers with only two alleles. In fact, for SNPs, the probability that two unrelated individuals share at least one allele identical by state (IBS) by chance is quite high. With the example frequencies given in the beginning, this probability can be calculated as $1\text{-}2\cdot 0.6^2\cdot 0.4^2\approx 88\%$. The utility of a genetic marker in a forensic application may be addressed using population genetic parameters such as typical paternity index, observed/expected heterozygosity and polymorphic information content [21].

One downside with STR markers is their scarcity throughout the human genome while SNP:s exist in great abundance [22]. Kling et al used a microarray chip [23] where 900.000 SNP:s were genotyped in a single reaction, while the current commercially available STR multiplexes amplifies maximally 24 markers in one reaction [24]. New typing technologies, such as next generation sequencing [13], offers promising possibilities, not least sequencing of both STR markers and SNP:s, but will not be covered in this thesis. Indeed, obtaining the individual sequence of each STR allele cause an explosion of paths to explore for the biostatistical evaluations.

Genetic markers can further be divided into autosomal and gonosomal markers. The latter is also known as sex specific markers and defines the gender of an individual. For the autosomal markers, we have 22 chromosome pairs, i.e. for each genetic marker we have two variants, one on the chromosome inherited from the mother and similarly one inherited from the father. Due to chromosomal abnormalities, e.g. duplication, some individuals may have three or more variants or genes at a genetic marker. Possessing three variants, known as trisomi, is fairly uncommon, but is observed every now and then. These situations require special considerations that will not be included in this thesis. All individuals furthermore inherit one X chromosome from the mother and from the father either an additional X, specifying female gender, or Y chromosome, specifying male gender. It follows from this that Y-chromosomal markers are inherited directly between father and sons and can be used to trace paternal lineages, while the X-chromosomal markers have a more intricate inheritance pattern and is passed on between fathers and daughters while fathers and sons share no genes located on the X chromosome. The latter may be violated if other relations exists between the father and the son, e.g. through inbreeding.

1.1.2 Likelihood ratio

One of the most important statistical concepts in forensic genetics, and many other fields, is the likelihood ratio. A likelihood may be defined as

$$L(H,\phi) = P(Data \mid H,\phi)$$

where we calculate the conditional probability of observing some *Data* given hypothesis *H* and some parameters ϕ, where the latter may be implicit. In relationship testing, *H* typically refers to some hypothesis about disputed relationship, such as paternity or non-paternity. For instance, we may specify

> H_1: An alleged father is the true father of a child
>
> H_2: A random man, not related to the alleged father or the mother, is the true father of the child

To compare different hypotheses we form likelihood ratios (LR:s), e.g.

$$LR = \frac{P(Data \mid H_1 = Paternity)}{P(Data \mid H_2 = Non\ paternity)}$$

We consider an introductory example (see Figure 2) where the alleged father is homozygous[2] with alleles 12,12 and the child is heterozygous[3] with alleles 12,18 while the mother is unavailable. (Similar notation will be used throughout the thesis.)

[2] Homozygous means that an individual has inherited the same variant/allele from the mother and the father

[3] Heterozygous means that an individual has inherited different variants/alleles from the mother and the father

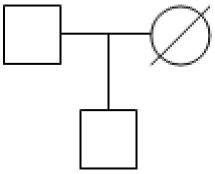

Figure 2. Pedigree describing a paternity case. Circles indicate females and squares males. Strikethrough means the genotypes for the indicated person for some reason are unavailable.

The likelihood and the corresponding ratio would then be formed as

$$LR = \frac{P(Data \mid Paternity)}{P(Data \mid Non\ paternity)} = \frac{P(G_{AF})P(G_C \mid G_{AF})}{P(G_{AF})P(G_C)} = \frac{P(12,12)p(18) \cdot 1}{P(12,12)P(12,18)} = \frac{1}{2p(12)}$$

where *P(G$_{AF}$)* and *P(G$_C$)* are the unconditional genotype probabilities of the alleged father and the child, while *P(G$_C$|G$_{AF}$)* is the conditional probability of the genotype for the child given that the alleged father is the true father. The joint probability for genotype *x,y* is denoted *P(x,y)* while the frequency of allele *x* in the population is denoted *p(x)*. We see that following simplifications, the end formula depends solely on the frequency of allele 12, i.e. the allele shared between the alleged father and the child and can be interpreted as the probability that a random man has that specific allele. The paternity case, with variations, will be used in the following sections to exemplify the various concepts discussed.

We may further combine prior information about the relationships to obtain posterior probabilities. The latter is attained using laws of conditional probabilities, in the present form known as Bayes theorem

$$P(R_j \mid Data) = \frac{P(Data \mid R_j)P(R_j)}{\sum_i P(Data \mid R_i)P(R_i)}$$

Where *P(R$_j$ | Data)* is the posterior probability for relationship *R$_j$* and *P(R$_i$)* are the prior probabilities for the different hypotheses about relatedness. In many situations we use flat priors, i.e. *P(R$_1$)=P(R$_2$)=…=P(R$_n$)=1/n*, though in large scale accidents and database searches the priors can be adjusted to reflect the large number of comparisons and thus possible false matches, see Budowle et al for a discussion [25]. How priors should be specified is a discussion in its own and will not be covered in this thesis. Bayes theorem allows multiple hypotheses to be compared in a single framework, something which is not easily provided using likelihood ratios as described above. Norgaard et al [26] as well as Buckleton et al [27] provide ideas and approaches to a likelihood ratio framework when multiple hypotheses are considered.

1.1.3 Mutations

Mutations constitute a particularly important topic in the field of forensic genetics. A mutational event is a situation bringing some change to the genome of an individual. It may occur on the somatic level, meaning that only the exposed individual will be affected, while it may also occur in the sex cells, resulting in a change that will be inherited to other generations. We are mostly interested in the latter as this could possibly spread in a population but also, and maybe more importantly in the current thesis, in a pedigree through the transmissions. There are several different causes for mutations, e.g. radiation, dysfunctional DNA repair enzymes, environmental factors. For STR markers, another mechanism for mutations is observed. The effect is commonly called DNA strand slippage error [28] and occurs during replication when the polymerase that duplicates the DNA slips, most likely due to the repeated structures of the STR markers, to produce a new variant with one (or more) repetition more or less than the original allele [20, 29]. The probability to observe a variant further away from the original allele, in terms of repeats, decreases fast. The process is illustrated in Figure 3. The slippage error is in fact quite common, compared to "normal" mutations, occurring in roughly >0.5% of all DNA replications. As a consequence, it is of paramount importance to model mutations when using STR markers in inference of relationships, not only in paternity testing but in general.

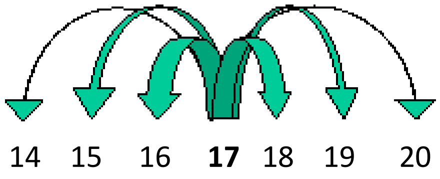

Figure 3. Illustration of the stepwise mutation model. The numbers indicate STR repeats (alleles).

Several models for mutations have been proposed, the simplest stating that it is equally probable to mutate to any other allele. A more reasonable approach is the stepwise model where we actually consider the alleles as repeats/steps [30-34]. In the basic stepwise model we define two parameters, the mutation rate μ, and the mutation range *r*. The first is the estimated overall mutation rate, i.e. the probability of observing a mutation while *r* is a parameter putting weight to different steps, i.e. how probable is one-step mutations compared to two-step mutations and so on. Mathematically, we define a mutation matrix *M*, consisting of elements $m_{i,j}$, where the diagonal elements are the probabilities of not mutating and the other $m_{i,j}$:s are the probabilities of mutating from allele *i* to allele *j*. We specify

$$m_{ij} = (1-\mu), \text{ if } i = j, \text{ i.e. the probability that an allele does not mutate.}$$

$$m_{ij} = k_i \mu r^{|i-j|}, \text{ if } i \neq j, \text{ i.e. the probability to mutate from allele } i \text{ to } j$$

The rows must sum to 1 and therefore the normalizing constants k_i are determined by the constraints $\sum_{j=1}^{N} m_{ij} = 1$.

To illustrate, consider the example where we have one marker with the set of alleles [12,13,14,15,16]. Table 1 then describes the elements of the stepwise mutation matrix *M*.

Table 1. Description of a stepwise transition model for mutations. The inner elements contain the probabilities forming the mutation matrix *M*.

Mutate to ->	**12**	**13**	**14**	**15**	**16**
12	$1-\mu$	$k_1\mu r^1$	$k_1\mu r^2$	$k_1\mu r^3$	$k_1\mu r^4$
13	$k_2\mu r^1$	$1-\mu$	$k_2\mu r^1$	$k_2\mu r^2$	$k_2\mu r^3$
14	$k_3\mu r^2$	$k_3\mu r^1$	$1-\mu$	$k_3\mu r^1$	$k_3\mu r^2$
15	$k_4\mu r^3$	$k_4\mu r^2$	$k_4\mu r^1$	$1-\mu$	$k_4\mu r^1$
16	$k_5\mu r^4$	$k_5\mu r^3$	$k_5\mu r^2$	$k_5\mu r^1$	$1-\mu$

We may calculate for instance k_1 as $k_1 = \dfrac{1}{r + r^2 + r^3 + r^4}$

An extension of the stepwise model, also accounting for microvariants or intermediate alleles, e.g. 12.3, is outlined by Kling et al and is implemented in the Familias software (described in Section 1.3.1) [35]. The model introduces a second mutation rate (α) corresponding to mutations to intermediate alleles. We extend the above notation with

$$m_{ij} = 1-(\mu+\alpha), \text{ if } i = j, \text{ i.e. the probability that an allele does not mutate.}$$

$$m_{ij} = k_i \mu r^{|i-j|}, \text{ if } i \neq j \text{ and if mutation from } i \text{ to } j \text{ is an integer step}$$

$$m_{ij} = \alpha / N_i, \text{ if } i \neq j \text{ and if mutation from } i \text{ to } j \text{ is a non-integer step}$$

Where N_i is equal to the number of non-integer mutations from allele *i*. Furthermore, for multi-generation pedigrees, allele frequencies will change slightly due to the fact that *pM≠p*, where *p* is the vector of allele frequencies at any given locus, i.e. the resulting product when multiplying the allele frequency vector with the mutation matrix is not the allele frequency vector. In other words, adding untyped parents/founders of typed persons will change the results. To counteract this, we can create a stationary matrix *S*, based on *M*, where the above mentioned criterion is fulfilled, see Dawid et al for further discussion and theory [30, 31]. One issue with the latter procedure is the fact that the matrix and its elements may change substantially, thus somewhat weakening the biological feasibility of the model. Further developments may improve the process of creating a stationary matrix, where the change, element wise, from the original mutation model is minimized.

1.1.4 Silent alleles

Silent alleles, also known as null alleles, are a subgroup of mutations where the primer binding site has a change such that no allele will be amplified by the PCR. The resulting profile is either homozygous or completely blank. As null alleles are estimated to be fairly uncommon, the latter is rarely observed, unless we consider haploid markers. In contrast, Kling et al [36] as well as Tomas et al [37] demonstrated that for the X-chromosomal markers included in the Investigator Argus X12 kit (QIAGEN), the silent allele frequency could be as high as 10% in certain populations. Nevertheless, for commercially produced kits in general, several different primers are commonly included to provide redundancy and to minimize the risk of null alleles.

The implication for the calculations is that we have to consider the possibility of a hidden allele, if an individual is genotyped as homozygous. In fact, also heterozygotes could have a silent allele given that he/she has a trisomi, but this can generally be neglected due to the rarity of such events to occur simultaneously. Consider a paternity case where the father is observed as having alleles 12,12 while the child is 18,18. The resulting LR where we consider both mutations and silent alleles would be

$$LR = \frac{P(Data \mid Paternity)}{P(Data \mid Non\ Paternity)} =$$

$$\frac{P(12,12)t(12 \to 18)\left[p(18)+p(s)\right] + P(12,s)\left[0.5t(12 \to 18)(p(18)+p(s)) + 0.5p(18)\right]}{\left[P(12,12)+P(12,s)\right]\cdot\left[P(18,18)+P(18,s)\right]} =$$

$$\frac{t(12 \to 18)p(12)\left[p(12)p(18)+p(12)p(s)+p(18)p(s)+p(s)^2\right] + p(18)p(12)p(s)}{\left[P(12,12)+P(12,s)\right]\cdot\left[P(18,18)+P(18,s)\right]}$$

where *P(12,12)* is the probability of the father's genotype and *P(12,s)* is the probability of the father having allele 12 and a silent allele, not observed in the data and with similar reasoning for *P(18,18)* and *P(18,s)* but for the child. The $t(x \to y)$ is a function describing the probability of a transition from allele *x* to *y*, and would be obtained from element m_{xy} in the mutation matrix described in Section 1.1.3. We further assume that no mutation can occur to or from a silent allele.

We see that if a mutation from allele 12 to 18 is improbable, compared to the probability of a silent allele, the formula reduces to

$$\frac{p(s)}{\left[p(12)+2p(s)\right]\cdot\left[p(18)+2p(s)\right]}$$

Fixing *p(12)=0.2* and *p(18)=0.3* we can plot the LR as a function of *p(s)*, see Figure 4.

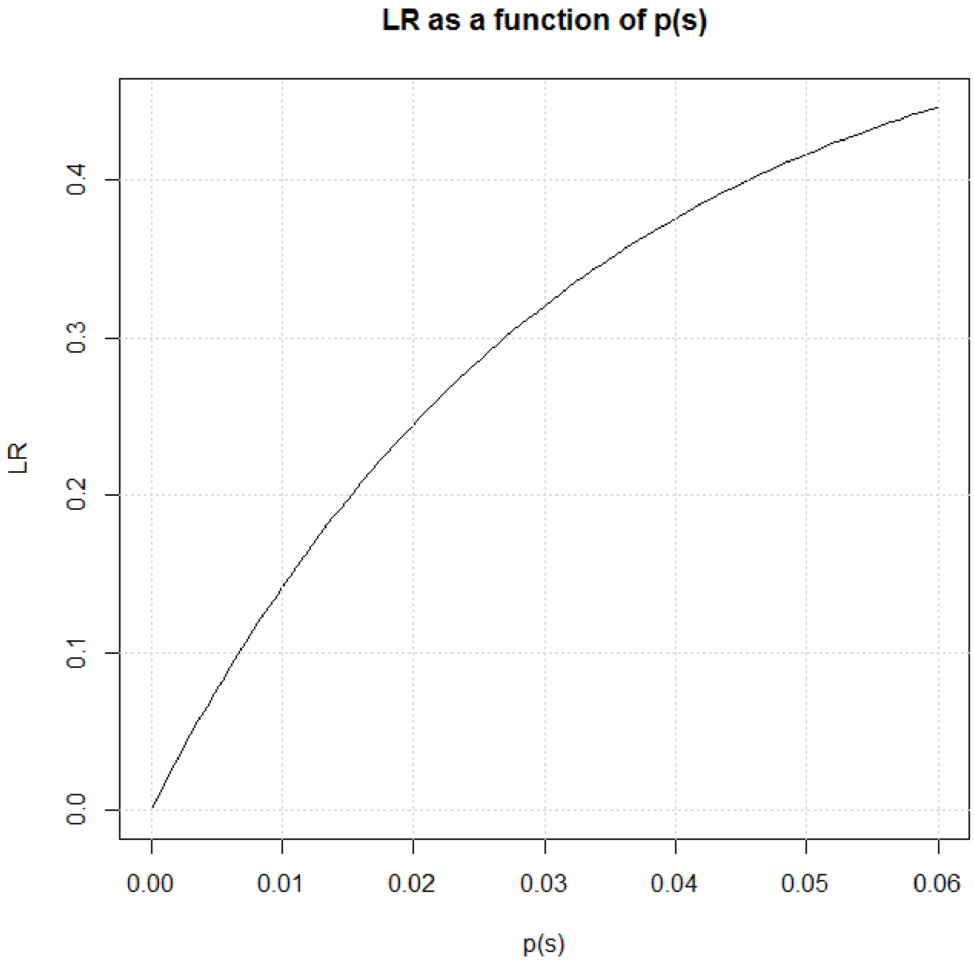

Figure 4. The LR for a paternity case with possible silent alleles. The frequency of the silent allele, p(s), is on the X-axis.

As estimating *p(s)* is usually difficult, a number of different values of *p(s)* may be considered, see Gjertson et al [38] and the homepage of NIST for some estimates [39].

1.1.5 Dropouts

When dealing with low template (LT[4]) DNA, degraded or otherwise low quality samples we may observe dropouts. A dropout is defined as an event where the PCR fails to completely amplify one or more of the alleles. For diploid markers we use the term *allelic dropout* if one of the alleles is not observed and *locus dropout* if both alleles drop out, thus resulting in a blank profile. New technologies and kits to amplify and withstand challenging samples are constantly developed, but an established framework to deal with dropouts in kinship calculations has been lacking. Several papers have proposed solutions [40-42]. A method to deal with allelic dropouts was developed by Dørum et al [43] and has been implemented in the latest version of Familias [35]. Dropouts are similar to silent alleles in that there is something hidden, not observed, that we wish to model, though the statistical implications are different. Whereas a dropout is inferred from the quality of the profile and is something random, silent alleles are non-random and will be transmitted throughout a pedigree. Consider again the paternity case in Section 1.1.2, where the alleged father is homozygous 12,12 but the child is now heterozygous 17,18. Obviously a silent allele cannot explain the data as the child would then also need to be homozygous, disregarding other observation level effects. Instead, assume we suspect a dropout in the profile of the father. The LR, where we disregard mutations, can then be formulated as

$$LR=\frac{P(Data\,|\,Paternity)}{P(Data\,|\,Non\ Paternity)}=\frac{\sum_j P(G_{AF_j})P(G^*_{AF}\,|\,G_{AF_j})P(G_C\,|\,G_{AF_j})}{P(G_C)\sum_j P(G_{AF_j})P(G^*_{AF}\,|\,G_{AF_j})}=$$

$$=\frac{P(12,17)d(1-d)0.5p(18)+P(12,18)d(1-d)0.5p(17)}{P(17,18)\left[P(12,12)(1-d^2)+P(12,x)d(1-d)\right]}=\frac{d}{p(12)(1-d)+2d}$$

where we sum over possible genotypes for the alleged father, G_{AFj}, and where d is the probability that a single allele drops out. Dropouts are assumed to occur independently so the probability that both alleles in a homozygote drops out is d^2. Furthermore, x denotes an allele different from 12. Dropout probabilities may be marker-specific, even profile specific, and may be estimated using a logistic regression model [44, 45]. The important point with the model for dropouts, described in detail in Dørum et al [43], is the conditional probabilities of observing the genotypes given the true (latent) genotypes, in this case given by the $P(G^*_{AF}|G_{AFj})$. Observe that for heterozygous genotypes we can model dropouts by stating that such an event has not occurred, obviously, with probability $(1\text{-d})^2$. In the formula above, modeling dropout for the genotype of child would cancel out as it would appear both in the numerator and the denominator. See Figure 5 for a graph of the above formula for some fixed values on *p(12)*.

[4] In the current setting meaning low concentrations of DNA, e.g. <0.5 ng/µl

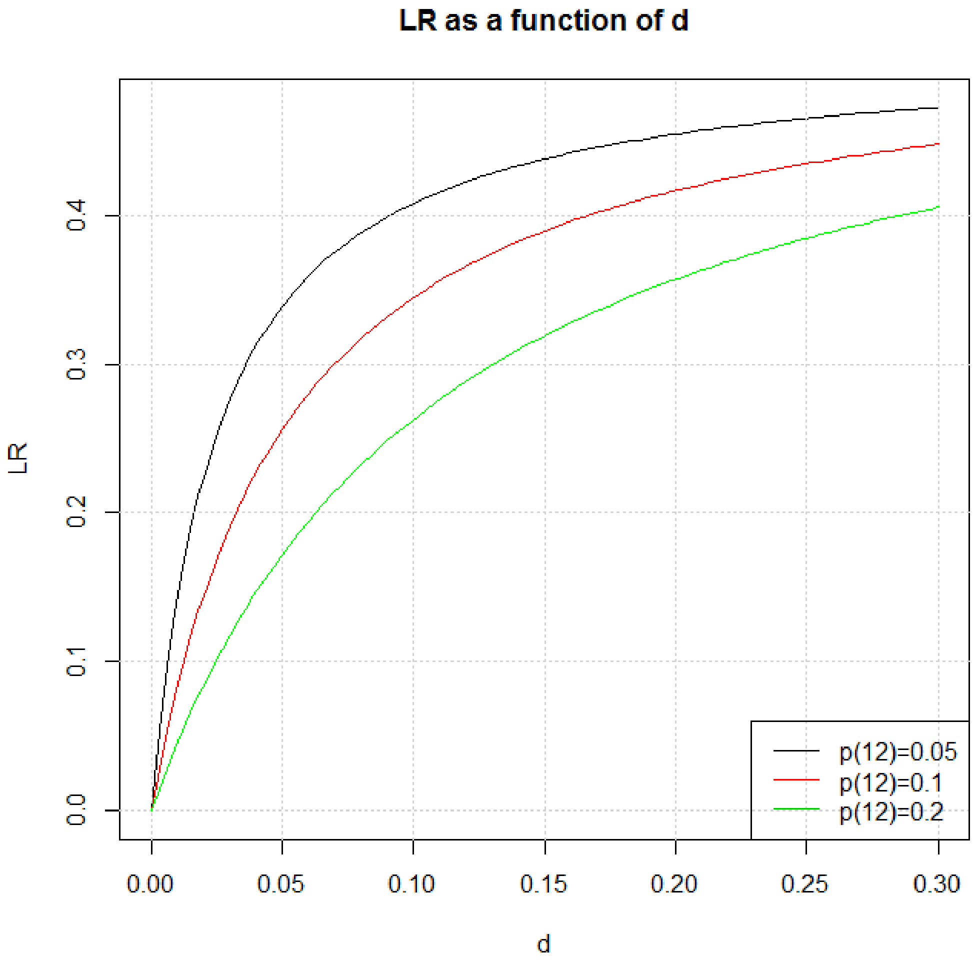

Figure 5. The LR for a paternity case with possible dropouts. The probability of dropout (d) is on the X-axis.

1.1.6 Hardy Weinberg equilibrium and subpopulation correction

From a larger perspective, there are a number of population genetic effects that could be modeled, see Balding for an overview [46]. In forensic genetics we may collectively combine several effects into one parameter known as kinship or subpopulation correction coefficient, typically denoted θ or F_{st}. To exemplify, consider some population frequency data. Due to inbreeding at the population level the data may require a correction of the allele frequencies. This is typically common in smaller isolated populations or when a general population is suspected to contain marriage between related individuals and will result in an excess of homozygotes. Other effects that may influence the allele frequencies include genetic drift, mutations and migration.

In a population where the allele frequencies obey the Hardy Weinberg equilibrium (HWE) the genotypes frequencies can be calculated as

$$\begin{array}{c} p_i p_j = p_i^2, \text{ if } i = j \\ 2p_i p_j, \text{ if } i \neq j \end{array}$$

A model for subpopulation structure was proposed by Sewall Wright in the early 1940s [47]. This was further developed and adopted in a forensic setting by Balding et al [48, 49]. As described by Balding, we may consider the procedure of calculating allele frequencies as a sampling process using a Dirichlet distribution, see formula below

$$p'_i = \frac{\theta n_i + (1-\theta) p_i}{1+(N-1)\theta} \quad (1)$$

where *p′ᵢ* is the updated frequency for allele *i*, θ is the subpopulation correction parameter, p_i is the estimated frequency of allele *i*, n_i is the total number of observations for allele *i* prior to sampling this allele and *N* is the total number of observed alleles prior to sampling this allele. For a complete derivation of the formula, see Section 5.3.2 in Balding [46]. We may use equation (1) to compute the genotype frequencies in a population where θ>0 as

$$\frac{\theta \cdot 0 + (1-\theta) p_i}{1+(0-1)\theta} \cdot \frac{\theta \cdot 1 + (1-\theta) p_i}{1+(1-1)\theta} = p_i \theta + (1-\theta) p_i^2, \text{ for homozygotes}$$

$$2\left[\frac{\theta \cdot 0 + (1-\theta) p_i}{1+(0-1)\theta} \cdot \frac{\theta \cdot 0 + (1-\theta) p_j}{1+(1-1)\theta}\right] = 2(1-\theta) p_i p_j, \text{ for heterozygotes}$$

We see that if θ=0 the formulas reduce to the same as under HWE assumptions. We further note that for homozygotes the first term $p_i\theta$ can be interpreted as the probability that the two alleles in one individual are actually identical by descent (IBD). As pointed out by Balding, when testing for HWE commonly using an exact test, we may observe deviations even though HWE can be assumed. This is a consequence of the fact that we will always have finite populations and therefore deviations will always be observed.

To better demonstrate the effect of subpopulation correction in calculation of likelihoods, we may visualize a pedigree in terms of *founders* and *non-founders*. Founders are defined as all individuals not having (defined) parents of their own, while non-founders can be defined as individuals with at least one (defined) parent. Founders, or rather the alleles of the founders, are the link between the pedigree and the population. For a pedigree with a large number of homozygous founders and a θ>0, allele frequencies will change significantly given that the founders have identical genotypes. In other words, recurring alleles for founders would be more common in an inbreed population than otherwise. We can in fact illustrate the effect of θ on a simple paternity case. Consider genetic data where the alleged father is 12,12 and the child is 12,18. We may now write (ignoring at this point complexities as mutations, silent alleles, dropouts etc.)

$$LR = \frac{P(Data \mid Paternity)}{P(Data \mid Non\ Paternity)} = \frac{P(\text{Sampling two 12:s and one 18})}{2P(\text{Sampling three 12:s and one 18})} =$$

$$= \frac{1}{2\left(\dfrac{2\theta + (1-\theta)p(12)}{1+2\theta}\right)} = \frac{1+2\theta}{2\left(2\theta + (1-\theta)p(12)\right)}$$

See Figure 6 for the effect of θ with different values of *p(12)*.

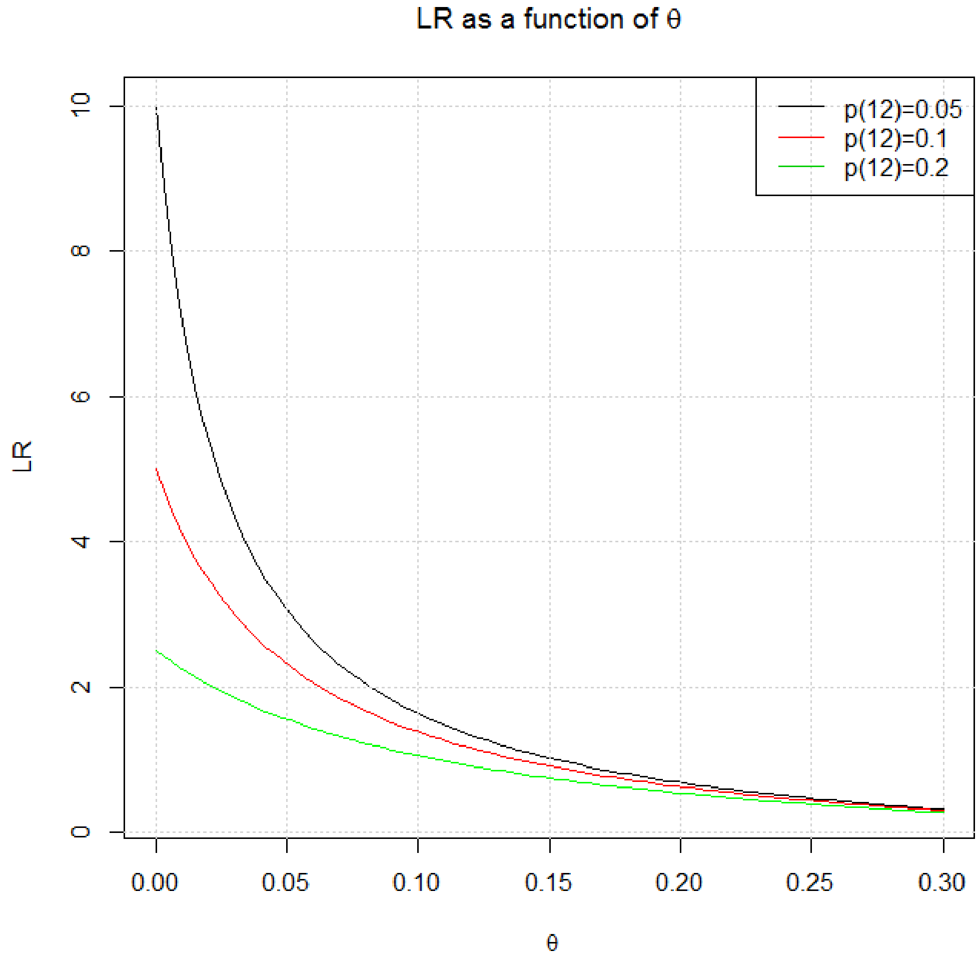

Figure 6. The LR for a paternity case with subpopulation correction. The value of θ is on the X-axis.

As an interesting detail, in a population where all individuals are full siblings the subpopulation correction parameter would be 0.25, while in a population where all individuals are 1st cousins the same value would be 0.125. This can be compared to values usually applied in statistical calculations ranging from 0.01-0.05. Details on methods for estimating the parameter can be found in Balding et al [46].

1.1.7 Inbreeding

Inbreeding is a concept that relates closely to subpopulation correction, but is handled differently in the statistical calculations. To illustrate, we may consider the relationship between two individuals in terms of identical-by-descent probabilities (IBD). For any pair wise fully outbreed relationship we may write

$$P(Data \mid R) = P(IBD = 0 \mid R)g_0 + P(IBD = 1 \mid R)g_1 + P(IBD = 2 \mid R)g_2 \quad (2)$$

where *P(IBD=x|R)=k$_x$* is the conditional probability of two persons sharing *x* alleles identical by descent (IBD probabilities) given a relationship *R*, while the set [g_0, g_1, g_2] are functions of allele frequencies depending on if 0, 1 or 2 alleles are IBD. See Table 2 for some examples of IBD probabilities for given relationships, and Hepler et al for a more comprehensive list [50]. Furthermore, using Table 2, we may deduce that g_0, g_1 and g_2 in equation (2) correspond to the probabilities of unrelated, parent-child and identical twins relationships.

Table 2. IBD probabilities for some pair wise relationships.

Relationship (R)	P(IBD=0|R)=k$_0$	P(IBD=1|R)=k$_1$	P(IBD=2|R)=k$_2$
Identical twins	0	0	1
Parent-child	0	1	0
Full siblings	0.25	0.5	0.25
Half siblings	0.5	0.5	0
Unrelated	1	0	0

To account for inbreeding we must consider an extension of equation (2) where we may actually have 0, 1,2,3 or 4 alleles IBD. To specify

$$P(Data \mid R) = \sum_{i=0}^{9} \Delta_i g_i \quad (3)$$

where the *Δ$_i$* are called the Jacquard coefficients and relates to different inheritance patterns [51]. Further, the *g$_i$*:s are still functions of allele frequencies. To illustrate, consider the example pedigree in Figure 7, illustrating two full siblings where the parents are in addition siblings of their own.

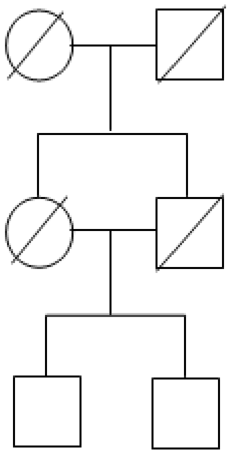

Figure 7. Illustration of an inbreed relationship where the parents of two full siblings are full siblings of their own.

Given the hypothesis depicted in Figure 7 the *Δ*$_1$, which represents the probability that the two individuals share two alleles IBD and in addition the alleles are in turn IBD to each other, is given be the events where the parents share one allele IBD (0.5) and this allele is transmitted to both siblings and where the parents share two alleles IBD (0.25) and one of these are transmitted to both siblings; $\Delta_1 = 0.5 \cdot 0.5^4 + 0.25 \cdot \left(0.5^4 + 0.5^4\right) = 0.5^4$. Similar reasoning applies for the rest of the coefficients, further details may be found in the given reference [51].

It is now fairly easy to see the distinction between inbreeding and the subpopulation correction (coancestry). Whereas the former influence the IBD patterns as illustrated above, the latter would affect the g_i in (3) by adjusting the allele frequencies. Similar to the wording in the introduction, inbreeding as discussed above deals with models within pedigrees while coancestry as discussed in Section 1.1.6, require models for population effects.

1.1.8 Linkage

Genetic linkage is the phenomenon occurring within a pedigree when alleles at different loci are inherited dependently, i.e. there is a dependent inheritance pattern. The cause of this occurrence is generally attributed to the physical proximity of loci on the same chromosome. In fact, this is a truth with some modification as linkage may actually be quite different for two loci separated by say 1000 bases on one chromosome and two loci separated by the same distance on some other chromosome, i.e. it is dependent on other things than physical distance alone. One measure of the genetic distance is centiMorgan (cM), where 1 cM is very roughly equal to 1 million bases (Mb). Even more commonly, we denote linkage in terms of recombination fraction (crossover rate), *r*, where this fraction is the probability that two loci will crossover in any given meiosis (actually the probability of any odd number of crossovers). The relation between cM and recombination fraction can be obtained from a mapping function. For instance, Haldane's mapping function specifies

$$r = \frac{1-e^{-2d/100}}{2}$$

relating recombination fraction, *r*, to the genetic distance *d*, measured in cM. The formula relies on the assumption that the pattern of recombination along a chromosome follows a Poisson process. The assumption is reasonable in calculation though interference, i.e. the occurrence of previous crossovers affecting the probability of a subsequent crossover, is not accounted for.

To obtain a measure of the linkage between two markers, we may typically analyze larger extended pedigrees where haplotypes and their inheritance as units can be traced throughout the tree. For statistical considerations, linkage only affects transitions probabilities within a pedigree, and we generally require at least two meioses to observe an effect.[5] As a consequence, random match probabilities will never be affected by linkage, unless the alternative hypothesis is for instance "My brother did it" [52]. In medical genetics, linkage is commonly used as a first step to screen for potential genes. It is a natural approach as linkage extends quite far, in theory all along the chromosome, while other means may subsequently be used to get a more exact position.

Although described for relationship estimation, see e.g. Thompson [53], the forensic genetics field has been more hesitant to using linked markers. This could be due to the fact that no user-friendly implementations have existed. In addition, linked markers introduce more parameters and require complex models. In general, they may provide crucial information in some relationship cases [54-56]. Gill et al demonstrated that linkage should be considered whenever two or more meioses separate two typed individuals in a pedigree [57]. Furthermore, Kling et al provides simulations illustrating the effect on some common relationship scenarios [58]. One scenario, which is frequently illustrated, is the example involving the relationship hypotheses

H_{UNC}: Two individuals are related as uncle/nephew

H_{HS}: The two individuals are related as half siblings

Consider two individuals P1 and P2 with genotypes 17,19 and 19,21 respectively, at a genetic marker and 14,15 and 15,17 respectively at a second marker. Using two unlinked autosomal markers we may use equation (2) and obtain LR=1 as both relationship hypotheses have the same IBD probabilities,

[5] It should be noted that this is a very crude rule

i.e. k_0, k_1 and k_2 are equal for both relationships. On the contrary, considering the same two markers to be linked we get the formula

$$\frac{P(Data\,|\,H_{HS})}{P(Data\,|\,H_{UNC})}=\frac{\begin{array}{c}0.5g_{0,1}\left[\left((1-r)^2+r^2\right)g_{0,2}+\left(2(1-r)r\right)g_{1,2}\right]+\\ 0.5g_{1,1}\left[\left(2(1-r)r\right)g_{0,2}+\left((1-r)^2+r^2\right)g_{1,2}\right]\end{array}}{\begin{array}{c}0.5g_{0,1}\left[\left((1-r)^3+r^2(1-r)3\right)g_{0,2}+\left(3(1-r)^2r+r^3\right)g_{1,2}\right]+\\ 0.5g_{1,1}\left[\left(3(1-r)^2r+r^3\right)g_{0,2}+\left((1-r)^3+3r^2(1-r)\right)g_{1,2}\right]\end{array}}$$

Where $g_{i,j}=P_j(Data|IBD=i)$ are functions of allele frequencies given that *i* alleles are IBD for locus *j*. The terms including *r* may look complicated but is understood from the fact that for half siblings we have two meioses while for uncle-nephew we have three meioses. The first term is explained by the probability that zero alleles is IBD at the second marker given zero alleles is IBD at the first marker, which can be the consequence of two recombinations or none, $r^2+(1-r)^2$. Further, evaluating the $g_{i,j}$ we see that the LR will be a function of *r*, *p(19)* and *p(15)*, i.e. the shared allele at each locus. Figure 8 illustrates the LR as a function of *r* for some fixed values on *p(19)* and *p(15)*. It is obvious that the recombination rate has an impact on the results, as different number of meiosis differs between the two hypotheses, although given the current data fairly small.

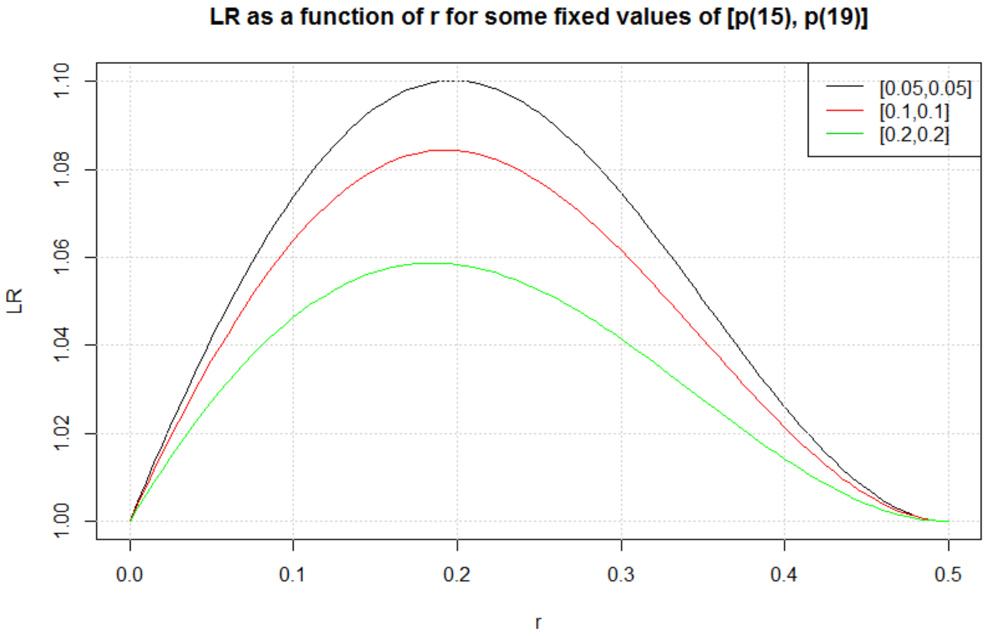

Figure 8. The LR in a case where the disputed relationships are half siblings and uncle-nephew. The recombination rate (r) is on the X-axis.

Using linked markers has, as previously indicated, generally been considered an obstacle in forensic genetics, while it can actually be turned into great advantage. As noted by Thompson [53], dependency tends to reduce the individual information contribution from each marker, but given that the alternative is to exclude linked markers from the calculations, including them is always the better option, assuming you have a model for the dependency. Their use will most probably play an even more important part in the future with the arrival of next generation sequencing technologies, inevitably leading to a greater number of markers and as a consequence dependency.

1.1.9 Linkage disequilibrium

Linkage disequilibrium (LD), also known as allelic association, is the non-random association of alleles at different loci. The concept should not be confused with genetic linkage, described in Section 1.1.8, which is the dependence between loci, although they are sometimes closely intertwined. To illustrate LD, consider two biallelic SNP markers with alleles *A,a* and *B,b*. The corresponding allele frequencies are p_A, p_a, p_B and p_b. We may now estimate the expected frequency of the combination of alleles *A* and *B*, i.e. the haplotype *[A B]*, as $p_A p_B$. Similar calculation may be conducted for the rest of the haplotypes, see Table 3.

Table 3. Expected haplotype frequencies for two biallelic SNP markers.

	A	a
B	$p_A \cdot p_B$	$p_a \cdot p_B$
b	$p_A \cdot p_b$	$p_a \cdot p_b$

In reality the haplotype frequencies may deviate from the expected, presented in Table 3, due to association between the alleles. We denote the observed haplotype frequencies with $p_{A,B}$, $p_{a,B}$, $p_{A,b}$ and $p_{a,b}$. One common measure of LD is the correlation r, defined as

$$r = \frac{p_{A,B} - p_A p_B}{\sqrt{p_A p_B p_a p_b}}$$

where *r* or the square of *r*, is a normalized parameter measuring the difference between the observed and expected haplotype frequencies.

One of the most common causes of LD is close proximity of the markers. As little recombination occurs throughout generations, the haplotypes at the markers tend to be inherited as units. This will in turn lead to a deviation from the expected haplotype frequencies in the population. In theory, LD may extend across chromosomes though normally the phenomenon is expected to occur for much shorter distances for markers located on the same chromosome. Whereas linkage, i.e. recombination fraction below 0.5, may typically be measured for any two markers located at less than 50 cM apart, LD is more common for alleles at markers located less than 1 cM apart. In addition to proximity, natural selection may be another cause of LD as possessing a specific allele may be beneficial to the survival of an individual and thus giving rise to association.

Contrary to linkage, LD does affect all calculations, even random match probabilities. We may again illustrate using a simple paternity case where we combine the two effects. Assume the alleged father is 12,14 at the first locus and 18,19 at the second locus. Similarly, the child is 12,13 at the first locus and 17,18 at the second locus. The LR (assuming no mutations or other complications except linkage and LD) may be formed as

$$LR = \frac{P(Data \mid Paternity)}{P(Data \mid Non\ Paternity)} =$$

$$= \frac{h(13,18)\left[0.5\cdot 2h(12,18)h(14,19)(1-r)+0.5\cdot h(14,18)h(12,19)r\right]}{\left(2h(12,17)h(13,18)+2h(13,17)h(12,18)\right)\left(2h(12,18)h(14,19)+2h(12,19)h(14,18)\right)}$$

where *h(x, y)* is defined as the frequency of haplotype with allele *x* at the first locus and *y* at the second locus; *r* is the recombination rate. If we assume linkage equilibrium (LE), i.e. *h(x,y)=p(x)p(y),*the formula simplifies to

$$\frac{h(13,17)\left[0.5\cdot 2h(12,18)h(14,19)(1-r)+0.5\cdot 2h(14,18)h(12,19)r\right]}{\left(2h(12,17)h(13,18)+2h(13,17)h(12,18)\right)\left(2h(12,18)h(14,19)+2h(12,19)h(14,18)\right)} =$$

$$= \frac{p(13)p(17)\left[(1-r)p(12)p(18)p(14)p(19)+rp(14)p(18)p(12)p(19)\right]}{\left(2p(12)p(17)p(13)p(18)+2p(13)p(17)p(12)p(18)\right)\left(2p(12)p(18)p(14)p(19)+2p(12)p(19)p(14)p(18)\right)} =$$

$$= \frac{p(12)p(17)p(13)p(18)p(14)p(19)}{16p(12)^2 p(17)p(13)p(18)^2 p(14)p(19)} = \frac{1}{16p(12)p(18)}$$

where the final formula does not include the recombination rate (*r*). If, however, LE cannot be assumed, we see that the weight of the different haplotypes for the father is important and thus, the value of *r* will be important to the results.

1.1.10 Simulations

Simulation is a versatile tool in virtually any scientific field. Using some stochastic model of how we think the reality works we may randomly simulate data based on this model and subsequently compare the simulated results with real data. In this way we can vary different parameters and see how they affect the outcome and also if the model may be simplified. In addition, simulations may well be the only practical way of finding summary statistics and measures of uncertainty.

In forensic genetics we can think of several situations where simulations are applicable. We may, for instance, study the distribution of likelihood ratios (LR:s) for a given relationship case. The simulations are then fairly straightforward and rely on some population frequency data to simulate founder genotypes and mutation models to simulate transitions from founder alleles to non-founder alleles, also known as the gene-dropping method. Kling et al provides a more thorough description of simulations in the software Familias [35]. In another forensic software, Forensim, simulations are used to estimate dropout probabilities [59, 60]. Simulations may further be used to solve complex models and find approximate posterior distributions. This may be particularly interesting in large node networks and Markov chains, e.g. Bayesian networks in forensics [61].

1.2 Computational methods

As mentioned in Section 1.1.2, forensic genetics commonly require a formulation of the likelihood, *P(Data|R, φ)*, where the parameters ϕ are implicit and we may write *P(Data|R)*. While it may be easy to write down the equations, algorithms to compute numerical results are generally harder to develop. There are two main such algorithms currently implemented for computations of likelihoods in relationship inference, Lander-Green and Elston-Stewart, with numerous extensions and implementations [62, 63]. For the sake of the current thesis, a brief description of the essentials is helpful, while detailed description is provided in e.g. Ziegler et al [64]. It is important to note that the complete models used in medical genetics also describe probabilities for the connection between genes and disease status, while the current description will be restricted to likelihoods for postulated relationships.

1.2.1 Elston-Stewart

In 1970, Robert Elston and John Stewart proposed an algorithm to effectively compute the likelihood for genetic marker data given some hypothesis about the relationship for the involved individuals [62], see also Ziegler et al [64] and Cannings et al [65] for further discussion. The general algorithm, without conditioning on observed data, can be formulated as

$$L(H) = \sum_{G_1} \cdots \sum_{G_N} \prod_{f} P(G_f) \prod_{\{o,p\}} P(G_o \mid G_p)$$

Where *[G_1,…,G_N]* is the set of all possible genotypes for all individuals in a pedigree *H*, *f* is the founders of the pedigree and *o* is the non-founders, defined as all individuals having parents *p* in the pedigree, and whose genotype probabilities are conditional on their parents'. The parents may be founders or non-founders and in the current setting we allow an individual to have only one (defined) parent. The likelihood, *L(H)*, only makes sense once we condition on observed data, where the set of possible genotypes *[G_1,…,G_N]*, is commonly greatly reduced. Unless we consider observational level errors, such as dropin, dropouts and genotyping errors, the observations reduce the set of possible genotypes to one for each typed individual. In addition, if we do not consider mutations, a great number of $P(G_o|G_p)$ will be zero, thus further reducing the set of possible genotypes.

In the general formula, described above, we must iterate over all possible genotypes for untyped individuals which increase exponentially, even though we condition on the observed genotype data. To effectively handle large pedigrees a peeling process is implemented. The Elston-Stewart (ES) algorithm divides the pedigree into nuclear families, where the children of a parent in a nucleus are independent of the rest of the pedigree given the parent. Conditioning on the connecting nodes, the performance time of the algorithm grows approximately linearly in terms of the number of individuals.

Consider the example in Figure 9, where the dispute concerns whether the two individuals denoted *U1* and *U2* are related to the *Child* as paternal uncles or not. The pedigree indicates two founders, the parents of *U1*, *U2* and the *Father*. The mother of the *Child*, also a founder, can be peeled away as her data is absent and not relevant using the current example. The non-founders are *U1*, *U2*, *Child* and *Father*.

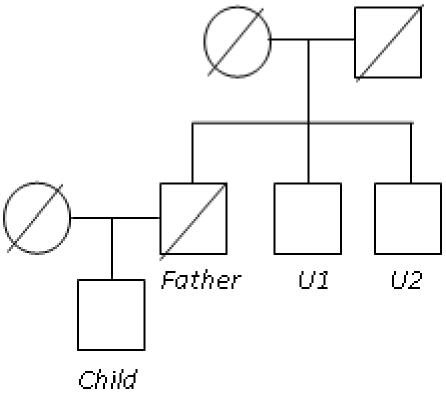

Figure 9. Illustration of a deficient paternity case where the data of two uncles (U1 and U2) are available.

The ES algorithm would typically start by calculating the likelihoods for the genotypes of the *Father* given the genotype of the *Child*. The possible genotypes for the *Father* contain all possible genotypes for the given marker. Given that we disregard mutations, the set is greatly reduced. The latter simplification, or other restrictions leading to fewer genotypes, may sometimes be necessary for extended relationships with several connecting nodes. The algorithm would continue by calculating conditional probabilities for the different genotypes of the *Father* given the uncles and their relationship as full siblings. The final step is the summation of products of the conditional probabilities and likelihoods.

The ES algorithm can generally handle large pedigrees but is restricted to outbreed relations in its original formulation. Cannings et al [65] extended the algorithm to handle inbreeding, however, inbreed relationships and loops in the pedigree causes problems in the peeling process and is generally not effectively handled. It is also apparent that mutations may cause intense computations with a large set of possible genotypes for connecting nodes. O'Connell et al [66] present details on how to combine some of the core points of the Elston-Stewart with details from the Lander-Green algorithm, described below, by implementing inheritance vectors within the nuclear families.

1.2.2 Lander-Green

Another algorithm was proposed by Eric Lander and Phillip Green in 1987 [63]. The algorithm uses hidden Markov chains to handle marker dependency (linkage). Markov chains are commonly used in statistics to model dependency problems. Given the value at a node in a Markov chain, all subsequent nodes are independent of previous nodes. For genetic markers this model is in fact an approximation as there may be interference for markers several steps away given that a recombination has occurred between two other markers. However, for practical purposes this may be used as an acceptable approximation in the calculations. The general Lander-Green algorithm for the likelihood can be formulated as

$$L(H) = \sum_{V_1} \ldots \sum_{V_N} P(V_1) \prod_{i=2}^{N} P(V_i \mid V_{i-1}) \prod_{i=1}^{N} P(G_i \mid V_i)$$

where we for a given pedigree enumerate all the possible meioses and those constitute the inheritance vectors V_i for each locus i. Furthermore, the algorithm contains the transition probabilities between states in the hidden Markov chain, $P(V_i|V_{i-1})$ as well as the probability of the genotypes given the current inheritance pattern, $P(G_i|V_i)$.

To illustrate, consider the pedigree of two full siblings (Figure 10) and data for two autosomal markers. The length of V is two as we have two markers and for each V_i we have a vector of binary indicators, v_j where $j=1,\ldots,J$ is a specific inheritance pattern. Each v_j has four elements, each element taking either the value 0 or 1. The first element represents the first meiosis in Figure 10, denoted $v_{.1}$ and the value is 0 if the paternal allele has been transmitted and 1 if the maternal allele has been transmitted from the parent. The complete set of inheritance patterns for each marker is then enumerated as, [0 0 0 0], [0 0 0 1],…,[1 1 1 1]; in total 16 different patterns. Using the marker data in Figure 10, we only have to consider the reduced set of inheritance patterns where the two individuals share one or zero alleles IBD for both markers, given that we disregard the possibility of a mutation.

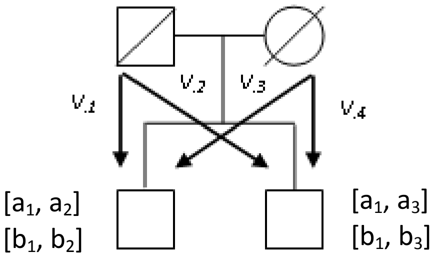

Figure 10. Pedigree describing a full sibling relationship where the parents are unavailable. The arrows indicate inheritance patterns. Genotypes are indicated with brackets for two different loci.

We may now write

$$L=\sum_{V_1}\sum_{V_2}P(V_1)P(V_2\mid V_1)\prod_{i=1}^{2}P(G_i\mid V_i)$$

where we sum over the complete inheritance space, in the current example defined by V_1 and V_2 and their components. The probability $P(V_1)$ is the prior probability for each element of V_1 and is simply $1/Length(V_1)=1/16$. We continue by calculating $P(V_2|V_1)$, i.e. the conditional probability of the current inheritance pattern in V_2 given the current inheritance pattern in V_1. This is typically a product of r and $(1-r)$ representing the total recombination probability (also known as transition probability) between two inheritance patterns. For instance, given V_1=[0 0 0 0] and V_2=[0 0 1 0] the transition probability would be $(1-r)^3r$.

The last part of the formula is given by the conditional probability of the genetic data given the current inheritance patterns at V_1 and V_2. This is typically a product of allele frequencies. Using the marker data specified in Figure 10, we may actually calculate $P(G_i|V_i)$ for the set of V_i:s at V_1 and V_2. The inheritance patterns indicating that the two siblings share one allele IBD, e.g. [0 0 0 1], $P(G_1|V_1)=p(a_1)p(a_2)p(a_3)$ and $P(G_2|V_2)=p(b_1)p(b_2)p(b_3)$ while for patterns indicating that they share zero alleles IBD, e.g. [0 0 1 1], $P(G_1|V_1)=4p(a_1)^2p(a_2)p(a_3)$ and $P(G_2|V_2)=4p(b_1)^2p(b_2)p(b_3)$.

We see that if the pedigree is large and we have a large number of meioses the inheritance vectors, V_i , grow in size and also, as a consequence, the number of possible transitions between V_i and V_{i-1} will increase substantially. Various improvements have been proposed, including reduction of the inheritance space using founder symmetries [67, 68], Fourier transforms [69] and other strategies [70]. As a general rule of thumb, the complexity increases linearly with the number of markers but exponentially with the number of persons, i.e. the number of meioses.

1.2.3 Approximate methods

Besides the two algorithms mentioned above, there are methods that implement approximate approaches instead of exact calculations. We may for instance use Monte-Carlo simulation to estimate the likelihoods [71, 72]. Approximate methods are not covered in this thesis but may be the solution to complicated cases where exact computations are not feasible and where the above mentioned algorithms fail. In theory, we may actually reduce the complexity of the calculations to scale linearly both in the number of markers and the number of individuals.

1.3 Software

There are a number of relevant software implementations that compute likelihoods for a set of individuals given some genetic marker data and hypotheses about relationships. For instance, freely available R-packages such as MasterBayes [73] and paramLink [74]. In addition several programs intended for linkage computations and calculation of LOD scores exist, e.g. Merlin, Vitesse, GeneHunter and Allegro [67, 75-77]. For forensic purposes we usually require more specialized tools, in the sense that results are easily visualized and minimal user knowledge is demanded. Moreover, these software implementations typically require a model to account for mutations, silent alleles and subpopulation structure. According to a study made by Poulsen et al [78], Familias [35, 79] was used by the greatest number of forensic laboratories, followed by DNAview [80].

Below is a list of tools, developed by the author of this thesis and relevant for some of the statistical evaluation required in forensic genetics.

1.3.1 Familias

Familias was originally developed by Egeland and Mostad in the mid 1990s [79]. Its importance has since increased and is now considered a gold standard in the forensic genetic field [78, 81]. The software calculates likelihoods for a set of persons with some genetic marker data and a number of mutually exclusive hypotheses, given population frequency data in combination with models for mutations, silent alleles and subpopulation structure.

Familias implements a variant of the Elston-Stewart algorithm, described in Section 1.2.1. The implementation works by dividing the pedigree into parts using cutsets, where each part is conditionally independent of other parts given the connecting node(s) the cutset consists of. Familias implements a number of features, such as advanced mutation models, subpopulation correction and dropouts. As pointed out by Drabek, the software has lacked some functionality, asked for among users [81]. Recent developments by Kling et al [35] have led to several extensions and new features. For instance, users may now simulate data to find false positive/negative rates as well as distributions of likelihood ratios for a given set of genetic markers and some hypotheses about relationship for a set of individuals. The simulation interface uses a gene dropping method, whereby the alleles of all founders in a pedigree are sampled from the population frequency database. The alleles for the founders are subsequently dropped down through the pedigree to all non-founders. It is in fact easy to account for several complications resulting from the need to model e.g. dropouts, subpopulation correction and mutations, by means of simulation as the complexity does not grow noticeably with the number of individuals or the structure of the pedigree. Familias now further includes a disaster victim identification (DVI) module, where users can easily compare large sets of unidentified remains with reference data from relatives of missing persons. Following larger scale mass disaster identification it is obvious that tools to determine the connection between unidentified individuals with relatives of missing persons are crucial. During the past two decades the use of DNA has increased as a primary means of identification. The first use was reported by Olaisen et al in the identification following a plane crash in Spitsbergen, Svalbard [82]. Recent examples include the 9/11 WTC terror attack, the South Asia tsunami in 2004 and the Katrina hurricane [9, 11, 41, 83-85]. Moreover, Familias now has the ability to perform blind searches where a large dataset can be scanned to find a priori unknown relationships. This is particularly useful in DVI operations where unidentified individuals may actually be related.

1.3.2 FamLink

In response to an ongoing discussion in the forensic community of how to handle linked markers [57, 86, 87], the software FamLink was developed [58]. The software uses the methods available in Merlin [75] to compute likelihoods. Briefly the software implements a variant of the Lander-Green algorithm using sparse binary trees to effectively speed up the calculations. One of the drawbacks of the software, from a forensic genetic point of view, is its lack of ability to model mutations. However, we argue that in many of the cases where linkage has an effect, mutations is not as relevant to account for. FamLink further implements a simulation tool where the user may study the impact of recombination on a given case.

1.3.3 FamLinkX

One of the most recent additions to the set of programs developed by the author of this thesis is FamLinkX. The software implements a completely new algorithm for genetic marker data on the X chromosome. In fact the algorithm, described in Kling et al [105] is general and not restricted to X-chromosomal markers. Several papers have demonstrated that the commercial X-STR kit from Qiagen includes sets of markers where linkage disequilibrium is observed [36, 88-90]. Moreover, several marker kits, developed in-house, include a number of different markers located on the X chromosome. Due to the shortness of the chromosome, the markers are often linked. Given that no existing software could provide all the necessary features, a new model was developed, simultaneously accounting for linkage, linkage disequilibrium and mutations. The model relies on similar ideas as presented by Kurbasic et al [91] but further implements models for mutations. The general utility of X-chromosomal markers is not discussed in this introduction but is rather covered elsewhere [88, 92-95].

1.4 Object Oriented Bayesian Networks

As an alternative approach to the commonly used software described above, developed as Windows based programs, Bayesian networks may prove useful [61, 96-101]. Bayesian networks rely on Markov properties where probability distributions for nodes in the network are updated as information is specified. Given specific values for a set of connecting nodes, subsequent nodes are independent of previous nodes. An introduction as well as examples in a forensic setting are given in Taroni et al [101].

A framework was developed by Kling et al [61] for some relationship scenarios accounting for linkage, linkage disequilibrium and mutations. However, as concluded by Kling et al, the complexity of these networks grow quickly and a general implementation remains to be constructed. We may use sampling methods to calculate approximate probability distributions and compute the likelihood ratio. The Bonaparte software uses Bayesian networks to compute likelihoods in relationship testing and is specifically adopted towards DVI operations [102], but does not account for linkage or linkage disequilibrium.

2 Paper summaries

The papers in this thesis cover several computational problems the modern forensic scientist is or will be faced with. We start by exploring a promising new typing method, high density SNP microarrays, where we genotype more than 900,000 genetic markers on small chips. The theme continues in the subsequent papers where we describe and explore new statistical methods.

The first paper (I) provides a broad motivation for the other papers as we demonstrate the necessity of the research conducted in the following papers (II, III, V and VI). Throughout the articles we illustrate the computational and statistical issues encountered when using closely located genetic markers. Hopefully this thesis and its papers will shed some light on possible approaches to resolve the problems and take advantage of the information inherent in such data. Whereas paper II explores a more experimental approach using object oriented Bayesian networks, papers III, IV and VI describe general implementations of methods to handle issues such as linkage and linkage disequilibrium.

We further consider other issues connected to computational problems encountered in forensic genetics, e.g. mass disaster identification problems, simulations and models for mutations. Paper VI provides new developments for the forensic software Familias [35], implementing ideas to resolve the mentioned issues.

Paper I: DNA microarray as a tool to establish genetic relatedness – Current status and future prospects

There is an increasing interest in using a greater number of genetic markers as more distant genetic relationships are investigated [23, 56]. Paper I uses data from two extended families genotyped on a high density SNP microarray chip from Affymetrix. The chip includes more than 900,000 SNP:s, more or less evenly spread throughout the entire genome. Software such as Merlin [73] is commonly used in medical genetics to study linkage in genetic disorders, but has been less used in forensic genetics. This paper is one of the first to present the application of the software on real high density marker data to calculate likelihoods for extended relationships in that setting. Previous studies have investigated the use of large sets of markers on simulated data [55], however importantly, as the present paper demonstrates there is an obvious deviation between real and constructed data. The latter is explained by the population genetic phenomena known as linkage disequilibrium, or allelic association. Whereas this can be simulated, no good algorithm to handle the implications in the statistical calculations has been proposed, though one approach is presented by Abecasis et al [103].

In summary the paper outlines the general utility of using large numbers of linked genetic markers to solve extended and complex relationships. Moreover, it motivates many future projects.

Paper II: Using Object Oriented Bayesian Networks to model linkage, linkage disequilibrium and mutations between STR markers

The second paper focuses on Object Oriented Bayesian Networks (OOBN) as a tool to model dependency between markers and alleles [61]. Bayesian networks embody the central concept of Markov chains where one node is independent of the rest of the network given the connecting nodes. The paper presents a graphical network created in the software Genie [104], a free tool to visualize networks and easily modify values thereby obtaining the posterior distribution for all other nodes. As an example, the paper uses real data from two STR markers, adopted in regular forensic casework, where dependency between the markers (linkage) and association between alleles (linkage disequilibrium) had been suggested [57, 86]. Subsequent papers demonstrated that the latter could be ignored while the former should be accounted for in statistical calculations. The paper concludes that, although easy to present, the presented model suffer some drawbacks. For example, the network experiences computational problems when calculating the exact posterior distribution for a network given some nodes when a large number of alleles is present, e.g. a typical issue with polymorphic STR markers. Solution for the mentioned difficulties are suggested though a general implementation is not presented and the final conclusion is that OOBN:s may be used for research purposes.

Paper III: FamLink – A user friendly software for linkage calculations in family genetics

As an alternative to the framework presented in Paper II, the paper adapts the functionality and algorithm presented in the software Merlin [75]. Building on the existing computational core, FamLink provides a graphical user interface aimed at forensic users with the interest of calculating likelihoods or simulating linked genetic markers [58]. The paper considers some theoretical approaches to validations and simulations demonstrating the utility of FamLink on a number of cases. The software has since its release been used by a number of laboratories.

Paper IV: A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium and mutations

The fourth paper builds on the ideas presented particularly in Paper II by presenting a general model to handle dependency between genetic markers in likelihood computations [105]. Similar to the

Lander-Green algorithm this new model relies on Markov chains to handle dependency between markers while also including a second multistep Markov chain to handle dependency between alleles across markers. In addition the model can handle data with genetic inconsistencies, i.e. mutations and is therefore specifically suited for forensic purposes. A detailed implementation of the algorithm is described for X-chromosomal marker data. X-chromosomal markers have for a period of time been of great interest in the forensic community [88, 92-95] due to their ability to provide information in several cases where autosomal markers fail. For instance, two half siblings may ask whether they are maternal or paternal half siblings; something which is undistinguishable with autosomal markers.

The paper further continues by demonstrating the utility of the implementation using simulated data as well as some real examples. In summary, the software provides means to solve cases where no previous methods or implementations appear adequate.

Paper V: Familias 3 – Extensions and new functionality

Familias is a software for calculating likelihoods for genetic marker data given some hypotheses about relatedness for a set of persons [79]. The software has long been considered a gold standard in the forensic community, but has lacked some desired functionality [81]. This paper focuses on the new version, with user requests in mind, still keeping the computational core. The new version includes the possibility to handle disaster victim identification (DVI) operations and missing person databases, where large number of unidentified remains is compared against large numbers of reference families. In addition users may now use Monte-Carlo simulations to find distributions of likelihood ratios for any given case. This is particularly interesting in case work, as laboratories may now find out if planned case data is likely to result in sufficiently strong results, i.e. a high likelihood ratio, given the number of genotyped persons. Furthermore, the paper presents a new mutation model dealing with the increasing number of microvariant alleles. In summary the new version presented in the paper provides several new features while still preserving old functionality. The new version of Familias, freely available at www.familias.no, has been developed and coded by the author of this thesis. On a mathematical note, observant readers may note that the mutation parameterization presented in Section 1.1.3 for the extended stepwise model differs from that presented in Paper VI. A small change has been made, presenting a slightly updated notation herein, to obtain a more consistent model.

Paper VI: FamLinkX – Implementation of a general model for likelihood computations for X-chromosomal marker data

The sixth paper presents a validation of the software FamLinkX. The program implements the model outlined in Paper IV for X-chromosomal marker data. The paper provides ideas to validate and confirm results when using the software to calculate likelihoods. This includes some theoretical considerations as well as simulations and a discussion on choice of parameters. Validation is in general not as straightforward as in other similar programs implementing exact computations, e.g. Familias and Merlin [35, 75, 81]. Although the calculations in FamLinkX are exact, several parameter choices can influence the results considerably. The simulations provide an idea of the general power of X-chromosomal markers in some common cases in forensic genetics.

3 Discussion

Using DNA to solve cases of disputed relationships has proven to be a truly valuable tool in many different scenarios. It may be as simple as a paternity case, where a child (or mother of a child) desires to find the true, unknown, father. More complicated cases, where distant relatives are investigated also are more and more common, definitively facilitated with the arrival of high density microarrays and next generation sequencing technologies [13, 23, 55]. It has become increasingly common to send saliva or blood samples to companies that may conduct such analyses for only 99$, e.g. *23andMe* [106] and *FamilyTreeDNA* [107]. This is actually much cheaper than what many of the current forensic laboratories charge for a common paternity case and the potential, in terms of relationship analysis, based on the raw data obtained from the mentioned companies is almost unlimited. Kling et al investigated potential limits when analyzing distant relationships with high density SNP data [23], though new methods may extend the boundaries for what relationships can be accurately established using genetics. A noteworthy point, as mentioned in Kling et al, is that 3rd cousins only share on average 0.78% of their autosomal genetic material and the same value for two unrelated individuals was found to be 0.34% in the large HAPMAP project [22].

Furthermore, DNA has in recent years been crucial in the identification process following large scale disasters, e.g. the hurricane *Katrina*, the *South Asia tsunami* and the *9/11 WTC terror attack* [9-11, 41, 83, 85]. We can in fact use information from other sources to update the prior probabilities in our Bayesian model. In combination with the genetic data we may provide a statistical statement as to the identity of an individual, something that other means of identification cannot generally do. For instance, fingerprinting and dental records generally rely on subjective opinions, even though statistical results may sometimes be produced.

Whereas much has been done since the introduction of genetic markers to solve disputes of relationships in forensic genetics, there are still a number of computational issues to be resolved; not least is to find good models for linked genetic markers. Approaches to handle linkage are well established and commonly implemented in medical genetic studies [67, 75, 76, 108, 109]. The implementations generally use the Lander-Green algorithm where markers are modeled as nodes in a Markov-Chain. Though sufficiently well adopted in medical genetics, forensic genetics usually require more extended models where mutations are also accounted for; The latter is derived from the fact that STR markers are commonly used [19, 20, 29, 110, 111]. It may also be a consequence of the fact that the Elston-Stewart algorithm is implemented in the commonly used forensic statistical software and this algorithm is not easily extended to efficiently account for a greater number of linked markers.

Within the scope of this thesis we have also developed a new mutation model, based on the step-wise transition model [29, 34]. The model has been well studied and is based on the fact that, for tandem repeat markers, mutations tend to depend on the repeat number, both in terms of the probability of mutating away from a marker and also, and perhaps most importantly, where we mutate to. We extend the model to also account for microvariants, also known as intermediate alleles, placed between distinct tandem repeats, i.e. 9.3 containing 9 repeats and three extra bases. From a mathematical point of view, the model we present is not stationary, i.e. the allele frequencies will change over time or in multi-generation pedigrees. This is a fact that is not disputed in population genetics, though in situation of calculating likelihoods this has the unwanted consequence of affecting the probabilities when introducing untyped persons in a pedigree. We are aware of developments, currently undertaken, to present a general method to create stationary mutation matrices based on any non-stationary ones.

In summary, this thesis and its papers provide research and solutions to some of the current core problems in forensic genetics,

1. Some of the problems and potentials with using high density genetic marker data are explored.
2. Means to compute likelihoods for linked autosomal markers are provided.
3. A new improved mutation model, accounting for microvariants, is described.
4. Methods to account for dropouts.
5. A new model to compute likelihoods for X-chromosomal markers, accounting for linkage, linkage disequilibrium and mutations, is described.
6. For each method, an implementation, easily and freely accessible to forensic scientists is provided.

As mentioned in the introduction, the genetic field involves a myriad of different subfields and family genetics, as discussed in this thesis, is only a small branch. However, in addition to the applications presented herein, the methods and approaches in the papers may have a wider use. For instance, some of the algorithms may be extended to medical genetics by including models about disease status. The Familias software, as introduced in Section 1.3.1, may further be adopted in the *Criminal genetics* field, see Figure 1, e.g. through *familial searching*[6].

[6] The concept familial searching has not been discussed in this thesies, briefly the concept deals with the search for relatives of an unidentified trace/stain in a database of convicted offenders.

4 What is next?

As mentioned in the discussion, there is an ongoing debate on the implementation of next generation sequencing technologies in the forensic genetics community. An argument for not introducing new markers has been the large databases existing for the current STR markers. It can easily be counter argued that the new typing technologies will most likely involve biallelic SNP markers where large databases are no longer necessary. I personally think that current typing methods will remain for a while, as they are well established and provide reliable results. At the same time, I do think the forensic community is obliged to adopt the new techniques and begin to use them more eagerly, as there will otherwise be more and more private, non-forensic, laboratories offering the services, which in theory can provide much more information about relatedness. These companies, and the research, is driven by profit and therefore dependent, non-autonomous. In my opinion, independent research is needed to achieve progress beyond mere product developments.

The thesis has provided the forensic community with description and implementations of various software. The goal has been, if not revolutionize, to provide great progression. The future certainly holds a great deal of development and maintenance of the programs, but due to shared functionality between the pieces of software, some synergy effects can be exploited.

Finally, I have, during the thesis, provided education, free-of-charge both in collaboration with the EUROFORGEN[7] initiative and on other occasions. These commitments will most probably continue following the closure of this thesis.

[7] http://www.euroforgen.eu/

5 References

[1] Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, et al., *Identification of the remains of the Romanov family by DNA analysis*, Nature Genetics, **6**:130-5 (1994)

[2] Coble MD, Loreille OM, Wadhams MJ, Edson SM, Maynard K, Meyer CE, et al., *Mystery solved: the identification of the two missing Romanov children using DNA analysis*, PLoS One, **4**:e4838 (2009)

[3] Ivanov PL, Wadhams MJ, Roby RK, Holland MM, Weedn VW, Parsons TJ, *Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II*, Nature Genetics, **12**:417-20 (1996)

[4] Stone R, *Buried, recovered, lost again? The Romanovs may never rest*, Science, **303**:753- (2004)

[5] Abbey DM, *The Thomas Jefferson paternity case*, Nature, **397**:32 (1999)

[6] Foster EA, Jobling MA, Taylor PG, Donnelly P, De Knijff P, Mieremet R, et al., *Jefferson fathered slave's last child*, Nature, **396**:27-8 (1998)

[7] Takagi M. *Thomas Jefferson and Sally Hemings: An American Controversy*: University of Virginia Press; 1999.

[8] Leclair B, Shaler R, Carmody GR, Eliason K, Hendrickson BC, Judkins T, et al., *Bioinformatics and human identification in mass fatality incidents: the world trade center disaster*, Journal of Forensic Science, **52**:806-19 (2007)

[9] Biesecker LG, Bailey-Wilson JE, Ballantyne J, Baum H, Bieber FR, Brenner C, et al., *Epidemiology. DNA identifications after the 9/11 World Trade Center attack*, Science, **310**:1122-3 (2005)

[10] Brenner CH, Weir BS, *Issues and strategies in the DNA identification of World Trade Center victims*, Theoretical Population Biology, **63**:173-8 (2003)

[11] Brenner CH, *Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities*, Forensic Science International, **157**:172-80 (2006)

[12] Bartlett JM, Stirling D. *A short history of the polymerase chain reaction*. PCR protocols: Springer; 2003. p. 3-6.

[13] Metzker ML, *Sequencing technologies - the next generation*, Nature Review Genetics, **11**:31-46 (2010)

[14] Pettersson E, Lundeberg J, Ahmadian A, *Generations of sequencing technologies*, Genomics, **93**:105-11 (2009)

[15] Shendure J, Ji H, *Next-generation DNA sequencing*, Nature biotechnology, **26**:1135-45 (2008)

[16] Dawkins R. *An Appetite for Wonder: The Making of a Scientist*: Random House; 2014.

[17] Gill P, Jeffreys AJ, Werrett DJ, *Forensic application of DNA'fingerprints'*, Nature, **318**:577-9 (1985)

[18] Jeffreys AJ, Wilson V, Thein SL, *Individual-specific'fingerprints' of human DNA*, Nature, **316**:76-9 (1985)

[19] Butler JM, *Genetics and genomics of core short tandem repeat loci used in human identity testing*, Journal of Forensic Science, **51**:253-65 (2006)

[20] Ellegren H, *Microsatellites: simple sequences with complex evolution*, Nature Review Genetics, **5**:435-45 (2004)

[21] Guo X, Elston R, *Linkage information content of polymorphic genetic markers*, Human Heredity, **49**:112-8 (1999)

[22] Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al., *A second generation human haplotype map of over 3.1 million SNPs*, Nature, **449**:851-61 (2007)

[23] Kling D, Welander J, Tillmar A, Skare Ø, Egeland T, Holmlund G, *DNA microarray as a tool in establishing genetic relatedness--Current status and future prospects*, Forensic Science International: Genetics, **6**:322-9 (2012)

[24] Oostdik K, Lenz K, Nye J, Schelling K, Yet D, Bruski S, et al., *Developmental Validation of the PowerPlex® Fusion System for Analysis of Casework and Reference Samples: A 24-locus Multiplex for New Database Standards*, Forensic Science International: Genetics, **12**:69-76 (2014)

[25] Budowle B, Ge J, Chakraborty R, Gill-King H, *Use of prior odds for missing persons identifications*, Investigative Genetics, **2**:15 (2011)

[26] Nordgaard A, Hedell R, Ansell R, *Assessment of forensic findings when alternative explanations have different likelihoods—"Blame-the-brother"-syndrome*, Science & Justice, **52**:226-36 (2012)

[27] Buckleton J, Triggs C, Champod C, *An extended likelihood ratio framework for interpreting evidence*, Science & Justice, **46**:69-78 (2006)

[28] Schlötterer C, Tautz D, *Slippage synthesis of simple sequence DNA*, Nucleic acids research, **20**:211-5 (1992)

[29] Ellegren H, *Heterogeneous mutation processes in human microsatellite DNA sequences*, Nature Genetics, **24**:400-2 (2000)

[30] Dawid AP, Mortera J, Pascali VL, *Non-fatherhood or mutation?: A probabilistic approach to parental exclusion in paternity testing*, Forensic Science International, **124**:55-61 (2001)

[31] Dawid AP, Mortera J, Pascali VL, Van Boxel D, *Probabilistic expert systems for forensic inference from genetic markers*, Scandinavian Journal of Statistics, **29**:577-95 (2002)

[32] Valdes AM, Slatkin M, Freimer N, *Allele frequencies at microsatellite loci: the stepwise mutation model revisited*, Genetics, **133**:737-49 (1993)

[33] Durrett R, Kruglyak S, *A new stochastic model of microsatellite evolution*, Journal of Applied Probability:621-31 (1999)

[34] Ota T, Kimura M, *A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population*, Genetics Res, **22**:201-4 (1973)

[35] Kling D, Tillmar AO, Egeland T, *Familias 3-Extensions and new functionality*, Forensic Science International: Genetics, **13**:121-7 (2014)

[36] Kling D, Dell'amico B, Haddeland P, Tillmar AO, *Population genetic analysis of 12 X-STRs in a Somali population sample*, Forensic Science International: Genetics, **11**:e7-8 (2014)

[37] Tomas C, Pereira V, Morling N, *Analysis of 12 X-STRs in Greenlanders, Danes and Somalis using Argus X-12*, International Journal of Legal Medicine, **126**:121-8 (2012)

[38] Gjertson DW, Brenner CH, Baur MP, Carracedo A, Guidet F, Luque JA, et al., *ISFG: Recommendations on biostatistics in paternity testing*, Forensic Science International: Genetics, **1**:223-31 (2007)

[39] Null allele estimates. http://www.cstl.nist.gov/strbase/NullAlleles.htm, Accessed: 9 September, 2014

[40] Buckleton J, Triggs C, *Dealing with allelic dropout when reporting the evidential value in DNA relatedness analysis*, Forensic Science International, **160**:134-9 (2006)

[41] Brenner CH, Weir B, *Issues and strategies in the DNA identification of World Trade Center victims*, Theoretical Population Biology, **63**:173-8 (2003)

[42] Curran J, Gill P, Bill M, *Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure*, Forensic Science International, **148**:47-53 (2005)

[43] Dørum G, Kling D, Baeza-Richer C, García-Magariños M, Sæbø S, Desmyter S, et al., *Models and implementation for relationship problems with dropout*, International Journal of Legal Medicine:1-12 (2014)

[44] Tvedebrink T, Eriksen PS, Asplund M, Mogensen HS, Morling N, *Allelic drop-out probabilities estimated by logistic regression--further considerations and practical implementation*, Forensic Science International: Genetics, **6**:263-7 (2012)

[45] Tvedebrink T, Eriksen PS, Mogensen HS, Morling N, *Estimating the probability of allelic drop-out of STR alleles in forensic genetics*, Forensic Science International: Genetics, **3**:222-6 (2009)

[46] Balding DJ. *Weight-of-evidence for Forensic DNA Profiles*: John Wiley & Sons; 2005.

[47] Wright S, *Isolation by distance*, Genetics, **28**:114 (1943)

[48] Balding DJ, Nichols RA, *DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands*, Forensic Science International, **64**:125-40 (1994)

[49] Balding DJ, Nichols RA. *A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity*. Human Identification: The Use of DNA Markers: Springer; 1995. p. 3-12.

[50] Weir BS, Anderson AD, Hepler AB, *Genetic relatedness analysis: modern data and new challenges*, Nature Review Genetics, **7**:771-80 (2006)

[51] Jacquard A, *Genetic structures of populations*, Genetics of Human Populations, (1970)

[52] Buckleton J, Triggs C, *The effect of linkage on the calculation of DNA match probabilities for siblings and half siblings*, Forensic Science International, **160**:193-9 (2006)

[53] Thompson E, Meagher T, *Genetic linkage in the estimation of pairwise relationship*, Theoretical and Applied Genetics, **97**:857-64 (1998)

[54] Egeland T, Sheehan N, *On identification problems requiring linked autosomal markers*, Forensic Science International: Genetics, **2**:219-25 (2008)

[55] Skare Ø, Sheehan N, Egeland T, *Identification of distant family relationships*, Bioinformatics, **25**:2376-82 (2009)

[56] Nothnagel M, Schmidtke J, Krawczak M, *Potentials and limits of pairwise kinship analysis using autosomal short tandem repeat loci*, International Journal of Legal Medicine, **124**:205-15 (2010)

[57] Gill P, Phillips C, McGovern C, Bright JA, Buckleton J, *An evaluation of potential allelic association between the STRs vWA and D12S391: Implications in criminal casework and applications to short pedigrees*, Forensic Science International: Genetics, **6**:477–86 (2011)

[58] Kling D, Egeland T, Tillmar AO, *FamLink - A user friendly software for linkage calculations in family genetics*, Forensic Science International: Genetics, **6**:616–20 (2012)

[59] Haned H, *Forensim: An open-source initiative for the evaluation of statistical methods in forensic genetics*, Forensic Science International: Genetics, **5**:265-8 (2011)

[60] Haned H, Egeland T, Pontier D, Pene L, Gill P, *Estimating drop-out probabilities in forensic DNA samples: a simulation approach to evaluate different models*, Forensic Science International: Genetics, **5**:525-31 (2011)

[61] Kling D, Egeland T, Mostad P, *Using Object Oriented Bayesian Networks to Model Linkage, Linkage Disequilibrium and Mutations between STR Markers*, PLoS One, **7**:e43873 (2012)

[62] Elston RC, Stewart J, *A general model for the genetic analysis of pedigree data*, Human Heredity, **21**:523-42 (1971)

[63] Lander ES, Green P, *Construction of multilocus genetic linkage maps in humans*, Proc Natl Acad Science U S A, **84**:2363-7 (1987)

[64] Ziegler A, König IR, Pahlke F. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an e-learning platform*: John Wiley & Sons; 2014.

[65] Cannings C, Thompson E, Skolnick M, *Probability functions on complex pedigrees [domesticated mammals, laboratory animals]*, Advances in Applied Probability, **10**:26-61 (1978)

[66] O'Connell JR, *Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm*, Human Heredity, **51**:226-40 (2001)

[67] Gudbjartsson DF, Jonasson K, Frigge ML, Kong A, *Allegro, a new computer program for multipoint linkage analysis*, Nature Genetics, **25**:12-3 (2000)

[68] Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES, *Parametric and nonparametric linkage analysis: a unified multipoint approach*, American Journal of Human Genetics, **58**:1347 (1996)

[69] Kruglyak L, Lander ES, *Faster multipoint linkage analysis using Fourier transforms*, Journal of Computational Biology, **5**:1-7 (1998)

[70] Idury R, Elston R, *A faster and more general hidden Markov model algorithm for multipoint likelihood calculations*, Human Heredity, **47**:197-202 (1997)

[71] George AW, Wijsman EM, Thompson EA, *MCMC multilocus lod scores: application of a new approach*, Human Heredity, **59**:98-108 (2005)

[72] Wijsman EM, Rothstein JH, Thompson EA, *Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees*, American Journal of Human Genetics, **79**:846-58 (2006)

[73] Hadfield J, Richardson D, Burke T, *Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework*, Molecular Ecology, **15**:3715-30 (2006)

[74] Egeland T, Pinto N, Vigeland MD, *A general approach to power calculation for relationship testing*, Forensic Science International: Genetics, **9**:186-90 (2014)

[75] Abecasis GR, Cherny SS, Cookson WO, Cardon LR, *Merlin--rapid analysis of dense genetic maps using sparse gene flow trees*, Nature Genetics, **30**:97-101 (2002)

[76] O'Connell JR, Weeks DE, *The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set–recoding and fuzzy inheritance*, Nature Genetics, **11**:402-8 (1995)

[77] Markianos K, Daly MJ, Kruglyak L, *Efficient multipoint linkage analysis through reduction of inheritance space*, American Journal of Human Genetics, **68**:963-77 (2001)

[78] Poulsen L, Friis SL, Hallenberg C, Simonsen BT, Morling N, *A report of the 2009-2011 paternity and relationship testing workshops of the English Speaking Working Group of the International Society For Forensic Genetics*, Forensic Science International: Genetics, (2013)

[79] Egeland T, Mostad PF, Mevåg B, Stenersen M, *Beyond traditional paternity and identification cases. Selecting the most probable pedigree*, Forensic Science International, **110**:47-59 (2000)

[80] Brenner CH, *Symbolic kinship program*, Genetics, **145**:535-42 (1997)

[81] Drabek J, *Validation of software for calculating the likelihood ratio for parentage and kinship*, Forensic Science International: Genetics, **3**:112-8 (2009)

[82] Olaisen B, Stenersen M, Mevag B, *Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster*, Nature Genetics, **15**:402-5 (1997)

[83] Dolan SM, Saraiya DS, Donkervoort S, Rogel K, Lieber C, Sozer A, *The emerging role of genetics professionals in forensic kinship DNA identification after a mass fatality: lessons learned from Hurricane Katrina volunteers*, Genetics Medicine, **11**:414-7 (2009)

[84] Prinz M, Carracedo A, Mayr WR, Morling N, Parsons TJ, Sajantila A, et al., *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**:3-12 (2007)

[85] Donkervoort S, Dolan SM, Beckwith M, Northrup TP, Sozer A, *Enhancing accurate data collection in mass fatality kinship identifications: lessons learned from Hurricane Katrina*, Forensic Science International: Genetics, **2**:354-62 (2008)

[86] O’Connor KL, Tillmar AO, *Effect of linkage between vWA and D12S391 in kinship analysis*, Forensic Science International: Genetics, **6**:840-4 (2012)

[87] O'Connor KL, Hill CR, Vallone PM, Butler JM, *Linkage disequilibrium analysis of D12S391 and vWA in U.S. population and paternity samples*, Forensic Science International: Genetics, (2010)

[88] Nothnagel M, Szibor R, Vollrath O, Augustin C, Edelmann J, Geppert M, et al., *Collaborative genetic mapping of 12 forensic short tandem repeat (STR) loci on the human X chromosome*, Forensic Science International: Genetics, **6**:778-84 (2012)

[89] Tillmar AO, *Population genetic analysis of 12 X-STRs in Swedish population*, Forensic Science International: Genetics, **6**:e80-e1 (2012)

[90] Edelmann J, Lutz-Bonengel S, Naue J, Hering S, *X-chromosomal haplotype frequencies of four linkage groups using the Investigator Argus X-12 Kit*, Forensic Science International: Genetics, **6**:e24-34 (2012)

[91] Kurbasic A, Hossjer O, *A general method for linkage disequilibrium correction for multipoint linkage and association*, Genetics Epidemiol, **32**:647-57 (2008)

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

[93] Pinto N, Silva PV, Amorim A, *A general method to assess the utility of the X-chromosomal markers in kinship testing*, Forensic Science International: Genetics, **6**:198-207 (2012)

[94] Tillmar AO, Egeland T, Lindblom B, Holmlund G, Mostad P, *Using X-chromosomal markers in relationship testing: Calculation of likelihood ratios taking both linkage and linkage disequilibrium into account*, Forensic Science International: Genetics, **5**:506–11 (2010)

[95] Szibor R, *X-chromosomal markers: past, present and future*, Forensic Science International: Genetics, **1**:93-9 (2007)

[96] Biedermann A, Taroni F, *A probabilistic approach to the joint evaluation of firearm evidence and gunshot residues*, Forensic Science International, **163**:18-33 (2006)

[97] Dawid AP, Mortera J, Vicard P, *Object-oriented Bayesian networks for complex forensic DNA profiling problems*, Forensic Science International, **169**:195-205 (2007)

[98] Gomes RR, Campos SV, Pena SD, *PedExpert: a computer program for the application of Bayesian networks to human paternity testing*, Genetics Mol Res, **8**:273-83 (2009)

[99] Hepler AB, Weir BS, *Object-oriented Bayesian networks for paternity cases with allelic dependencies*, Forensic Science International: Genetics, **2**:166-75 (2008)

[100] Koller D, Pfeffer A. *Object-oriented Bayesian networks*. Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI-97). Morgan Kaufman, San Francisco, CA.1997. p. 302-31.

[101] Taroni F, Aitken C, Garbolino P, Biedermann A. *Bayesian Networks and Probabilistic Inference in Forensic Science*. Chichester: John Wiley & Sons; 2006.

[102] Slooten K, *Validation of DNA-based identification software by computation of pedigree likelihood ratios*, Forensic Science International: Genetics, **5**:308-15 (2011)

[103] Abecasis GR, Wigginton JE, *Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers*, American Journal of Genetics, **77**:754-67 (2005)

[104] GeNie. http://genie.sis.pitt.edu/

[105] Kling D, Tillmar A, Egeland T, Mostad P, *A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations*, International Journal of Legal Medicine:1-12 (2014)

[106] 23andMe. http://www.23andme.com, Accessed: 14 August, 2014

[107] FamilyTreeDNA. www.familytreedna.com, Accessed: 14 August, 2014

[108] Keith JM, McRae A, Duffy D, Mengersen K, Visscher PM, *Calculation of IBD probabilities with dense SNP or sequence data*, Genetics Epidemiol, **32**:513-9 (2008)

[109] Leutenegger AL, Genin E, Thompson EA, Clerget-Darpoux F, *Impact of parental relationships in maximum lod score affected sib-pair method*, Genetics Epidemiol, **23**:413-25 (2002)

[110] Lynch M, *Rate, molecular spectrum, and consequences of human mutation*, Proc Natl Acad Science U S A, **107**:961-8 (2010)

[111] Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, et al., *Human genome sequence variation and the influence of gene history, mutation and recombination*, Nature Genetics, **32**:135-42 (2002)

Paper I

Forensic Science International: Genetics 6 (2012) 322–329

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

DNA microarray as a tool in establishing genetic relatedness—Current status and future prospects

Daniel Kling [a,b,d,*], Jenny Welander [c], Andreas Tillmar [a], Øivind Skare [d,e], Thore Egeland [b], Gunilla Holmlund [a,c]

[a] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Artillerigatan 12, SE-587 58, Linköping, Sweden
[b] Department for Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Norway
[c] Department of Clinical and Experimental Medicine, Faculty of Health Sciences, Linköping University, SE-581 85, Linköping, Sweden
[d] Norwegian Institute of Public Health, P.O. Box 4040, Nydalen, NO-0403 Oslo, Norway
[e] Department of Public Health and Primary Health Care, University of Bergen, P.O. Box 7804, NO-5020 Bergen, Norway

ARTICLE INFO

Article history:
Received 21 March 2011
Received in revised form 12 June 2011
Accepted 5 July 2011

Keywords:
DNA microarray
Kinship analyses
Affymetrix
Distant relatedness
Single nucleotide polymorphism
Merlin
FEST

ABSTRACT

In the past decades, microarray technology has definitely put an edge to the field of genetic research. Our aim was to determine whether single nucleotide polymorphism (SNP) microarrays could be used as a tool in establishing genetic relationships where current molecular genetic methods are not sufficient. We used the Genechip, Affymetrix GenomeWide SNP Array 6.0, which detects more than 900,000 SNP markers dispersed throughout the human genome. The intention was to find a good selection of SNP markers that could be used for statistical evaluation of relatedness in a forensic setting. We conducted pairwise comparisons in the R-package FEST as well as pedigree comparisons in Merlin. Our methods were applied on two separate families, where relationships as distant as 3rd cousins were known. In addition, a question about a possible common ancestry between the two families was tested. Relationships as distant as 2nd cousins could be readily distinguished both from unrelated and other, genetically, closer relationships. This was achieved with a selection of 5774 markers, where each pair of markers was separated by a genetic distance of at least 0.5 cM (centiMorgan). When considering 3rd cousins, and more distant relationships, the number of markers needs to be extended, consequently decreasing the genetic distance between the markers. However, inclusion of a too large number of markers presents new challenges and our results imply that the use of too dense sets of markers always yields the highest probability for the genetically closest relationship hypothesis. Simulations confirm that this is most probably caused by the fact that the computational model assumes linkage equilibrium between markers, a problem that will be further evaluated. Our results do however suggest that SNP-data derived from microarrays are well suited for kinship determination provided linkage disequilibrium is properly accounted for.

1. Introduction

In the past decades, the use of DNA has revolutionized many fields of research. It still remains the most important tool to trace genetic relationships, both in forensic casework and in clinical research. In medical research it is often crucial to accurately establish the relationships between the individuals participating in a study. In genetic association studies, unknown kinship between cases and also between controls, or even between these two groups, may give rise to false associations [1,2]. Also in linkage analysis the results can be seriously biased as a consequence of unknown relationships between pedigree founders [3].

In forensic casework, DNA can be used in crime scene investigations to find or exonerate a perpetrator. In paternity testing, the conventional problem is to determine whether a man is the biological father of a child. DNA analyses can also provide evidence to determine a disputed relationship more distant than first generation relatives. In particular, immigration cases often present genetic relationships where current forensic genetic methods do not produce sufficient evidence [4]. In forensic genetics, the choice of markers is at present mostly limited to short tandem repeats (STRs); genetically due to their high variability and their ability to provide a high power of discrimination; technically due to their suitability for multiplex PCR analyses. However, one of the disadvantages when using STR-markers is their high mutation rate. In addition the multiplex assays are often limited to 16–20 markers [5,6]. The use of single nucleotide polymorphisms (SNPs) has recently received some attention in the establishing of genetic relatedness. In the forensic field, the SNPforID Consortium has established a set of SNPs which

[*] Corresponding author at: Norwegian Institute of Public Health, Familiegenetikk, Gaustadalléen 30, NO-0027 Oslo, Norway. Tel.: +47 210 77663.
E-mail address: Daniel.Kling@fhi.no (D. Kling).

1872-4973/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2011.07.007

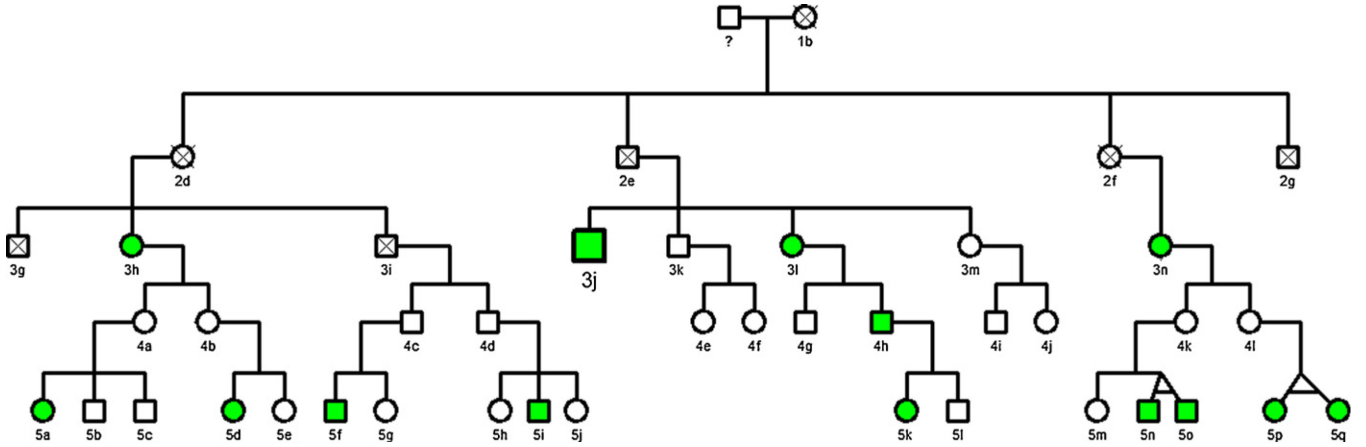

Fig. 1. Large Family. The pedigree describes a large family where relationships as distant as 3rd cousins were known. The question mark denotes an unknown paternal ancestor. Samples were drawn from the individuals marked with green.

performs sufficiently well to be used in court cases and can be multiplexed in one PCR reaction [7–10]. SNPs possess several advantages, which make them favourable when establishing complex or distant relationships. For one, they have a very low mutation rate, approximately 10^{-8} [11]. In addition, they can be analyzed in short amplicons and are generally easy to multiplex. Furthermore, there is an abundance of SNP markers to choose from in the human genome; The most recent paper from the HapMap project shows a map of 3.1 million of SNPs in the genome and the expected total number are 9–10 millions [12]. However, single SNPs provide very little genetic information, since they mostly are biallelic. The shortage of information can, however, be counteracted by analysing a larger number of markers. SNPs can be massively typed on high-density microarrays, such as the Genechips produced by Affymetrix or the HumanMap chips provided by Illumina, and have been extensively used in medical genetics [13]. A great number of markers is crucial in cases of distant relationships. The use of the standard STR markers, as well as a small set of SNP markers and a set of VNTR (Variable Number of Tandem Repeat) markers will not be enough [14]. Although easy to accomplish, the use of a larger number of markers presents challenges for the computational model used to distinguish between alternative pedigree hypotheses.

Different algorithms can be used for the purpose of calculating likelihood for a given pedigree and genotype data. They all share certain characteristics and the choice of which one to use is mainly depending on the number of markers and the number of individuals, see Gao et al. for a review [15]. One such algorithm is the Elston–Stewart algorithm [16,17], which can be described as a peeling algorithm and peels in the direction of individuals. This means that the calculation is only linear in the number of individuals. In contrast, the Lander–Green algorithm allows for a linear increase in the number of calculations to the number of markers [18]. The algorithm is implemented in the software Merlin, the main software used in this study [19]. The drawback is that both algorithms grow exponentially in one direction. In other words, the Elston–Stewart algorithm is capable of handling large pedigrees, but little genotype data, perhaps 100 markers, while the Lander–Green algorithm can handle hundreds and thousands of markers but only approximately 25 individuals in each pedigree. Besides this, the most prominent challenge, for any model, is to take genetic linkage and linkage disequilibrium (LD) properly into account. Genetic linkage has been shown, in simulation studies, to provide conclusive information in cases of relatedness [20,21]. The Lander–Green algorithm is able to take linkage into account, but assumes linkage equilibrium (LE). Therefore measures were taken to avoid the influence of LD, mainly by setting a minimum distance between the chosen markers, but also by using different sets of markers; see Supplemental Fig. S2 for a more thorough description of the selection procedure [15,19]. In addition an evaluation of possible LD for each selection of markers was carried out in PLINK [22].

In this study, we wanted to investigate if data from thousands of SNP markers could be used to resolve distant relatedness issues. For this purpose we used DNA from individuals representing different relationships known a priori and selected SNP-data derived from microarrays. We also applied our findings on a case of genealogy with a presumable half 1st cousin relationship.

2. Materials and methods

2.1. Sample data

Nineteen blood samples were collected from two families, Figs. 1 and 2, each presenting a wide selection of a priori known relationships, e.g. parent–child, grandparent–grandchild relations, full siblings, 1st cousins, 2nd cousins and 3rd cousins and uncle–niece. These known relationships were used to ascertain the validity of the statistical calculations as well as to establish which relationships could actually be determined. Finally, data from all tested individuals were used to establish whether or not the two families were related two generations back. Allele frequencies from 60 unrelated Swedish individuals were used as a reference population.

2.2. Simulations

Data were also simulated to further investigate the impact of linkage disequilibrium for different marker densities. The simula-

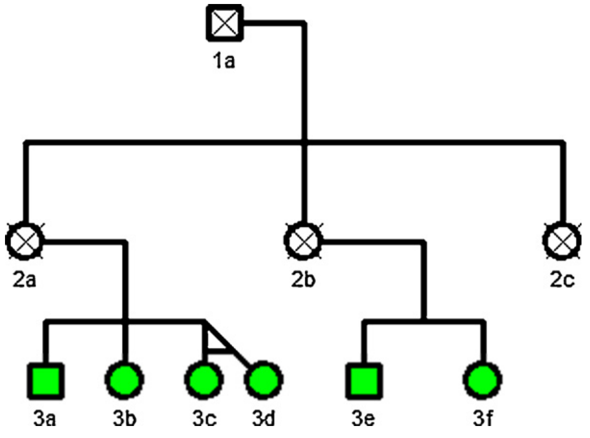

Fig. 2. Small Family. The pedigree describes a small family where relationships as distant as 1st cousins were known. Samples were drawn from the individuals marked with green.

tions were performed using FEST [20] where founder haplotypes with markers in linkage disequilibrium were drawn using the R package hapsim [23]. We used allele frequency information and LD data for chromosome 22 derived from HapMap for the CEPH (Utah residents with ancestry from northern and western Europe) population. To convert physical map distances (bp) to genetic map distances (cM), we used the Rutgers Combined Linkage-Physical Map of The Human Genome [24]. The following Bayesian approach was adopted; first, a true relation was drawn using a flat prior. Second, genotypes were simulated given the true relation: first the founder haplotypes assuming LD, then the genotypes of the descendants. Third, posterior probabilities were computed for each hypothesized family relation using Merlin. For each given marker density, these steps were repeated 5000 times, and then the posterior probabilities were averaged. Note that, if the likelihood computations were correct, the expected value of the posterior should equal the prior. This fact follows from $E[P[M = k|G]] = E[E[1_k(M)|G]] = E[1_k(M)] = P[M = k]$, where M is the family relation and G the genotype data. A bias in the averaged posterior probabilities, by not taking LD into account, would then be apparent as a deviation from the prior probabilities.

2.3. Microarray analysis

DNA was extracted as described by Lindblom and Holmlund [25]. The DNA concentration was quantified with Nanodrop (Thermo Scientific, Wilmington, DE, USA) and adjusted to 50 ng/μl prior to the microarray assay. Samples were analyzed on the Affymetrix GenomeWide SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA) according to the manufacturer's protocol.

2.4. Selection of markers

The raw data was analyzed in the software Genotyping Console version 4.0 (GTC), supplied by Affymetrix. From the original 900,000 markers, different selections of autosomal SNP markers were made. The selection criteria included minor allele frequencies (MAF), minimum distances between two neighbouring markers as well as Hardy Weinberg p-value for each marker (see Supplemental Fig. S2 for a graphical explanation of the selection procedure). In addition to the previously mentioned criteria, a subsequent evaluation of the LD between selected markers was carried out in PLINK. Two different approaches to evaluate the presence of LD were used. First computation of pairwise r^2 values between each SNP and the 100 most proximally located SNPs. For each selection of markers, the fractions of pairwise r^2 values above a "limit" (limit = 0.1; 0.2; 0.3; 0.5; 0.8) were calculated. Second, we searched for the presence of haploblocks that can be defined as a cluster of closely located SNPs in strong LD [26]. The number of haploblocks was estimated from the "haplotype block estimation" option in PLINK [22]. This estimation uses the algorithm published by Gabriel et al. [27].

A more complex selection procedure could possibly account for information content, as described by Krawczak et al. [28]. This paper describes a formula which can be used to address the issue in paternity cases. However, we consider more general pedigrees using linked markers and therefore these measures of informativity cannot be used.

2.5. Statistical calculations

Likelihoods for the hypothetical pedigree structures were obtained from the software Merlin [19]. In addition the R-package FEST, which provides a front-end user interface to Merlin, was used to perform simple pairwise comparisons between individuals [20]. FEST lets the user include certain predefined hypotheses in the analysis. There are three different simple types of pairwise relationships: (1) S–n–m – the sharing of two common ancestors n and m generations back, (2) HS–n–m, the sharing of one common ancestor n and m generations back. When $n = m$, we abbreviate to S–n and HS–n. Finally (3) PC–n denotes a parent–child relationship spaced by n generations. FEST was used due to its relative ease with which it allows the user to calculate the likelihoods for a large number of alternative hypotheses. In addition FEST provides an in-built thinning procedure for genotype data. However, FEST has some constraints. Firstly, pedigree structures with inbreed loops and marriage loops are impossible to specify in terms of simple pairwise relationships. Secondly, inclusion of genotypes from more than two individuals in each analysis is impossible, which might be necessary in distant relatedness cases.

The likelihoods, obtained from Merlin and FEST, were converted to posterior probabilities according to a Bayesian approach using flat priors. An in-house software (freely available from the corresponding author), was used to perform extensive tests in Merlin. In this study three different minor allele frequencies were tested; 0.2, 0.3 and 0.4. For each minor allele frequency, 10 separate analyses were performed based on different minimum distances between selected markers. The minimum distance was evenly spaced between 0.05 and 2 cM, yielding approximately 49,000 and 1800 markers respectively. The numbers vary slightly depending on which minor allele frequency was chosen. In addition, for each minimum distance and MAF, three separate selections, not including the same SNPs, were made in order to minimize the possible influence of linkage disequilibrium.

2.6. Genotyping errors

Genotyping errors may have an impact on the calculations [29,30]. A study was undertaken to establish the degree of genotyping errors. One control sample was typed eleven consecutive times and approximately 4000 markers, approximately 0.4% of the original 900,000, were excluded from all analyses due to overrepresentation of inconclusive results. This is an *ad-hoc* solution that requires further development for future applications, possibly by inferring an error frequency and implementing this into the statistical model. One example of a model accounting for genotyping errors is provided by Epstein et al. [31].

3. Results

3.1. Pairwise comparisons with known relationships using FEST

Using different sets of markers, pairwise relationships were shown to yield high posterior probabilities for relationships as distant as 2nd cousins (Tables 1 and 2). In Table 1, the calculated posterior probabilities are shown based on a selection of 5774 markers for six known relationships. The first row contains the true relationships; S-1 denotes full siblings, S-2 full cousins, S-3 full 2nd cousins, S-4 full 3rd cousins, while HS-1 denotes half siblings, HS-2 half cousins and PC-2 a grandfather–grandchild relationship. Table 2 shows the results where instead a selection of 12,453 markers was used to calculate likelihoods (the number of comparisons included to calculate the averaged posterior for each true relationship depends on the available data, see Supplemental Table S1). When calculating the posterior probability for a 3rd cousin relationship, see S-4 Tables 1 and 2, the highest probability achieved was 0.9991 in favour of the true hypothesis, with a selection of 12,453 markers. Although sufficient to establish the 3rd cousin relationship, comparing two unrelated individuals only yielded 0.64 in favour of the

Table 1
Posterior probabilities for each tested relationship, based on a selection of 5774 SNP markers (markers separated by at least 0.5 cM). A Bayesian approach with flat priors has been used to calculate posterior probabilities.

True relationship	S-1	S-2	S-3	S-4	PC-2	Unrelated 1	Unrelated 2
S-1	>**0.99999**	<0.00001	–	–	<0.00001	–	–
HS-1	<0.00001	<0.00001	–	–	–	<0.00001	–
S-2	<0.00001	**0.993**	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	–	0.007	–	–	–	<0.00001	–
S-3	–	<0.00001	**0.9999**	–	–	0.0035	–
PC-2	–	–	–	–	**0.9999**	–	–
S-4	–	–	–	**0.81**	–	–	0.19
Unrelated	<0.00001	<0.00001	<0.00001	0.19	<0.00001	**0.997**	**0.81**

The true relationships in the first row and corresponding probability in bold. A hyphen in a specific row means exclusion of the relationship as an alternative hypothesis. S-1 means full siblings, HS-1 half-siblings, S-2 full 1st cousins, HS-2 half 1st cousins, S-3 full 2nd cousins, PC-2 grandparent–grandchild relation and S-4 means full 3rd cousins. The same 5774 markers have been used in all comparisons. Due to the varying availability of pairwise *true* relationships (Supplemental Table S1), the number of examples included for each relationship varies; For S-1 five comparisons, S-2 ten comparisons, S-3 four comparisons, S-4 ten comparisons, PC-2 nine comparisons, Unrelated ten comparisons.

unrelated relationship, see Table 2 and the comparison denoted *Unrelated 2*. Inclusion of a large number of markers revealed to always favour the genetically closest relationship, also when unrelated was the true relationship. The threshold value, when this phenomenon starts to occur depends on which relationship is tested. As a rule of thumb, when testing relationships closer than 2nd cousins, more than 20,000 SNP markers should not be included to obtain reliable results. See Fig. 3(a)–(c) which describe an approximate threshold for three different relationships, S-2, S-3 and unrelated.

3.2. Pairwise comparisons with "unknown" relationships using FEST

The question whether the two families in Figs. 1 and 2 were related to each other was first examined with FEST. Pairwise comparisons between the individuals in the third generation, i.e. 3a/3b/3c/3d/3e/3f and 3h/3j/3l/3n, Figs. 1 and 2, were performed. The following hypotheses were included, *unrelated*, *half 1st cousins* (HS-2) and *full 1st cousins* (S-2). Table 3 shows an extraction of the results with various selections of markers. All comparisons yielded high probabilities for the two families to be unrelated.

3.3. Comparisons with "unknown" relationships using Merlin

We tested alternative hypotheses for the unknown relationship between the two families in Figs. 1 and 2, including data from all typed individuals in the third generation. All tests, independent of marker selections, revealed high posterior probability for the unrelated hypothesis (Table 4). The hypotheses tested assumed, however, that the individuals in the family in Fig. 1 were full-cousins. Separate tests also confirmed this relationship (see Supplemental Table S2 and Fig. S1).

3.4. Evaluation of linkage disequilibrium using PLINK

For each selection of markers we performed pairwise LD evaluations in PLINK. We tested for LD between markers separated by less than 100 SNPs, which roughly means comparing markers located less than 50 Mb apart, in a selection of 5774 markers. Of course this distance depends not on the number of markers but on the minimum distance chosen between two selected markers, e.g. choosing markers separated by 0.1 cM yields a distance of roughly 10 Mb. Table 5 describes the results for a selection of marker sets. Evidently, selecting markers located 0.05 cM apart, roughly 29,200 markers, yields a higher percentage of r^2 values above 0.5, while in a selection of 5800 markers, the number is considerably lower. Also, r^2 values above 0.3 are comparatively rare in the latter selection. Furthermore, Fig. 4 describes the relation between number of markers and the number of haploblocks. According to the estimation the dependence is approximately exponential, meaning that choosing more markers will yield an exponential increase in the number of haploblocks, i.e. markers in tight LD.

3.5. Simulations using FEST

Table 6 summarizes the simulation results, based on genotype data from chromosome 22, where we consider the hypotheses full cousins (S-2), half cousins (HS-2) and unrelated. We see that by reducing the distance between markers, the averaged posterior probability is shifted progressively towards full cousins, the genetically closest relationship. These results are in concordance with our experience for real data (see also Supplemental Table S3 where the same simulations have been conducted without accounting for LD).

Table 2
Posterior probabilities for each tested relationships, based on 12,453 SNP markers (markers separated by at least 0.25 cM). A Bayesian approach with flat priors has been used to calculate posterior probabilities.

True relationship	S-1	S-2	S-3	S-4	PC-2	Unrelated 1	Unrelated 2
S-1	>**0.99999**	<0.00001	–	–	<0.00001	–	–
HS-1	<0.00001	0.0002	–	–	–	<0.00001	–
S-2	<0.00001	**0.9998**	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	–	<0.00001	–	–	–	<0.00001	–
S-3	–	<0.00001	>**0.99999**	–	–	0.0017	–
PC-2	–	–	–	–	**0.99998**	–	–
S-4	–	–	–	**0.9991**	–	–	0.36
Unrelated	<0.00001	<0.00001	<0.00001	0.0009	<0.00001	**0.9983**	**0.64**

The true relationships in the first row and corresponding probability in bold. A hyphen in a specific row means exclusion of the relationship as an alternative hypothesis. S-1 means full siblings, HS-1 half-siblings, S-2 full 1st cousins, HS-2 half 1st cousins, S-3 full 2nd cousins, PC-2 grandparent–grandchild relation and S-4 means full 3rd cousins. The same 5774 markers have been used in all comparisons. Due to the varying availability of pairwise *true* relationships (Supplemental Table S1), the number of examples included for each relationship varies; For S-1 five comparisons, S-2 ten comparisons, S-3 four comparisons, S-4 ten comparisons, PC-2 nine comparisons, Unrelated ten comparisons.

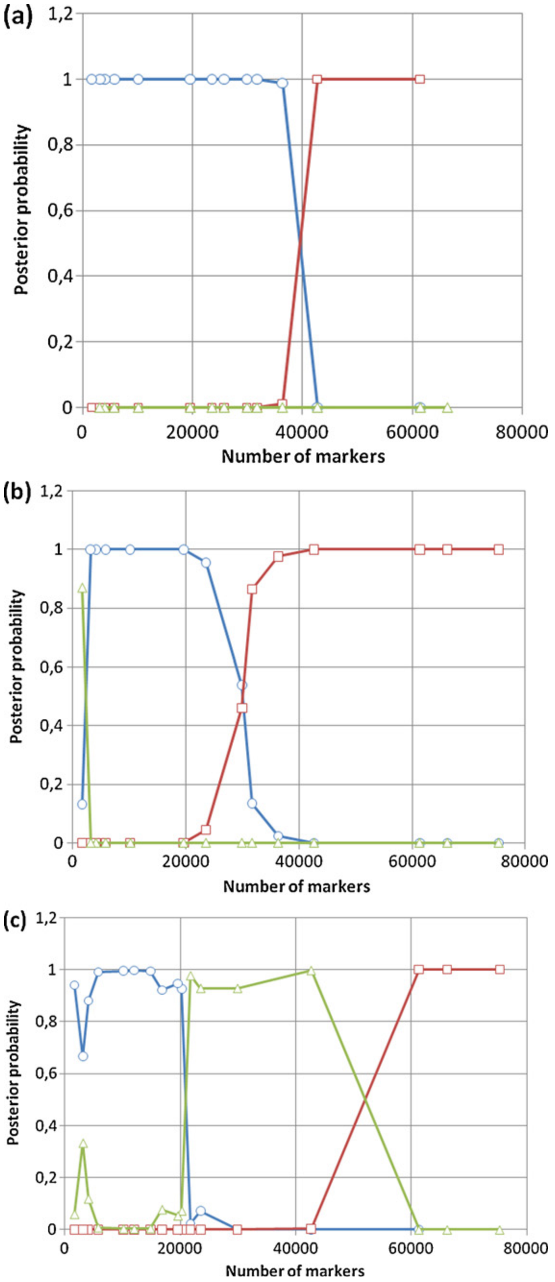

Fig. 3. (a)–(c). Graphs displaying the posterior probability for each hypothesis against the number of markers. (a) True relationship full 1st cousins (blue line) versus alternative hypotheses of full siblings (red line) and unrelated (green line). (b) True relationship full 2nd cousins (blue line) versus alternative hypotheses of full 1st cousins (red line) and unrelated (green line). (c) True relationship Unrelated (blue line) versus alternative hypotheses of full 2nd cousins (green line) and full 1st cousins (red line). The upper threshold value, when the true relationship no longer receives the highest posterior probabilities seems to be, for full 1st cousins: ~35,000 markers (markers separated by 0.05 cM), for full 2nd cousins: ~20,000 markers (markers separated by 0.1 cM), and for unrelated ~20,000 markers (markers separated by 0.1 cM).

4. Discussion

DNA has proven to be the most important tool to evaluate genetic relationships, both in forensic casework [32–34] and in medical research [1–3]. During the last decade mtDNA and gonosomal (X, Y) markers have been used to establish relatedness when lineages of maternal or paternal inheritance can be followed [35,36]. However, as soon as a line of inheritance is broken the genetic analyst loses track. Using thousands of autosomal SNP markers we showed that distant relationships could be established where the above-mentioned methods did not prevail. Although too early to draw any definite guidelines or conclusions, we believe the methods proposed in this study can be applied whenever complex family relations need to be resolved, as in for example genealogy studies.

The robustness of the tests was shown by using different sets of markers. The marker selection was based on a set of criteria that each chosen SNP-marker had to fulfil. Different minor allele frequencies did not appear to influence the results notably, though the issue was not extensively investigated. The distance between the markers did, however, show more impact on the results; especially when the number of markers exceeded 20,000, which is approximately equal to a distance of 0.1 cM between each pair of markers. It was apparent that a too dense selection of markers rendered the genetically closest relationship as the most probable, see Fig. 3(a)–(c). This phenomenon is, most likely, a consequence of linkage disequilibrium, which is also evident in Table 5 where more dense selections of markers yield a greater percentage of high r^2 values, but also, interestingly, an exponential increase in the number of haploblocks, see Fig. 4. Our simulations also further corroborates these results, see Table 6, where there obviously is a shift towards the closest relationship as more markers are included in the simulations. One of the reasons to why the results favour the closest relationship can possibly be explained by the "random" sharing of uncommon alleles. According to this admittedly speculative conjecture, a dense selection of markers amplifies the effect, as the uncommon alleles can possibly be in LD with other closely located uncommon alleles.

The Lander–Green algorithm, used to calculate the likelihoods, assumes the markers to be in LE and the likelihood computation collapses using many markers that are in LD. One reason to why the calculations fail is the large difference between the observed and the expected haplotype frequencies when dense sets of markers are used. Moreover, unrelated individuals will share certain haplotypes, as mentioned previously, due to a common ancestry, although further back, and they will appear related, i.e. false positives will arise [37,38]. In 2008 Kurbasic and Hossjer presented an extension to the Lander–Green algorithm in order to account for linkage disequilibrium [39]. They combined the Markov chain for inheritance vectors (i.e. Lander–Green) with another Lth order Markov chain that models LD structure. In this extension, the Markov chain contains information about the genotypes of the pedigree founders of L consecutive located loci. Kurbasic and Hossjer applied their method on a smaller simulation study (L = 1) and pointed out that the method is very computationally intensive unless the pedigrees are small and L is small. This limitation was also shown when the algorithm was implemented with a small number of forensically relevant STR markers [40]. Using a combination of kinship coefficient and IBS statistics, Manichaikul et al. recently presented a software, KING, which allows pairwise comparisons to be conducted on large sample material [41]. The authors claim the problem with LD is circumvented based on *large sample theory*. The KING software calculates a kinship coefficient, i.e. a rough estimate of an abstract family relationship, and not a forensically relevant probability value for a given pedigree hypothesis. We used the software on our material and the performance is comparable with our methods, for relationships closer than 3rd cousin. Using KING, 3rd cousin relationships could not be readily resolved. In addition, KING does not provide an answer to our main problem, determining the most likely pedigree.

As for Merlin/FEST, true relationships as distant as 3rd cousins could be distinguished with satisfactory posterior probabilities, using 12,453 markers. Unfortunately, inclusion of more distant relationships, e.g. 3rd cousins, as an alternative hypothesis when comparing two truly unrelated individuals, yields unsatisfactory probabilities, such as only a 64% posterior probability in favour of

Table 3
Posterior probabilities for the hypothesis of relationship between the two families, see Figs. 1 and 2, based on analyses using FEST.

Number of markers	19,518	12,453	10,144	5774	4074	3151
Comparison 1						
S-2	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	0.024	8.5e−6	0.0002	0.0001	7e−5	0.00086
Unrelated	0.975	0.99999	0.9998	0.9999	0.9999	0.999
Comparison 2						
S-2	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	0.0007	6e−6	5.1e−5	0.003	0.00085	0.006
Unrelated	0.999	0.9999	0.9999	0.997	0.999	0.994

Posterior probabilities for the included hypotheses. S-2 means full 1st cousins, HS-2 means half 1st cousins, see text for further details. Each value represents a posterior probability for a given selection of markers, see column header. Comparison 1 and 2, represents two separate tests to whether the two families in Figs. 1 and 2 are related.

Table 4
Posterior probabilities for the hypothesis of relationship between the two families, see Figs. 1 and 2, based on analyses using Merlin. A Bayesian approach with flat priors has been used.

Number of markers	19,518	12,453	10,144	5774	4074	3151
Comparison 1						
S-2	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	0.024	8.5e−6	0.0002	0.0001	7e−5	0.00086
Unrelated	0.975	0.99999	0.9998	0.9999	0.9999	0.999
Comparison 2						
S-2	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	0.0007	6e−6	5.1e−5	0.003	0.00085	0.006
Unrelated	0.999	0.9999	0.9999	0.997	0.999	0.994

The question was whether the two families were sharing a common paternal ancestor two generations back. Data was included from all individuals in the third generations of the two families. For each minor allele frequency, two different distances between two neighbouring markers has been tested, see column headings.

Table 5
Evaluation of linkage disequilibrium. The table describes the proportion of pairwise comparisons with a r^2-value above each limit. In addition the number of haploblocks in each selection has been calculated using PLINK. Limitcm stands for the minimum genetic distance between two markers in the selection.

	Proportion of pairwise-SNP with r^2 higher than r^2 limit					
	limitcm0.5	limitcm0.25	limitcm0.15	limitcm0.1	limitcm0.075	limitcm0.05
Number of markers	5865	10,227	14,869	19,420	23,263	29,277
Number of pair-wise comparisons	471,704	903,536	1,363,046	1,813,620	2,194,050	2,789,420
r^2 limit						
0.1	0.0168	0.0176	0.0568	0.0418	0.0225	0.0615
0.2	0.0020	0.0022	0.0102	0.0085	0.0053	0.0178
0.3	0.0015	0.0014	0.0073	0.0060	0.0038	0.0132
0.5	0.0014	0.0012	0.0057	0.0042	0.0026	0.0090
0.8	0.0014	0.0010	0.0044	0.0026	0.0014	0.0047
Number of haploblocks[a]	4	80	355	824	1482	2896

[a] Estimated in PLINK.

Table 6
Averaged posterior probabilities for simulated relationships. The table describes averaged posterior probabilities (with standard deviations in parentheses) from 5000 simulations of genotype data on chromosome 22. Markers are assumed to be in LD and are evenly spaced over the chromosome. Prior probabilities are equal to 1/3.

Number of markers on chr 22	Distance (cM) between markers	Number of markers if extended to all chromosomes	HS-2	S-2	Unrelated
1		45[a]	0.3333 (0.0000)	0.3335 (0.0002)	0.3332 (0.0002)
10	8.778	453	0.3333 (0.0000)	0.3339 (0.0005)	0.3327 (0.0005)
100	0.798	4535	0.3334 (0.0001)	0.3345 (0.0012)	0.3321 (0.0012)
200	0.397	9070	0.3332 (0.0001)	0.3355 (0.0013)	0.3312 (0.0013)
500	0.158	22,675	0.3337 (0.0002)	0.3445 (0.0015)	0.3218 (0.0015)
1000	0.079	45,349	0.3335 (0.0002)	0.3592 (0.0015)	0.3073 (0.0016)
1500	0.053	68,024	0.3338 (0.0003)	0.3927 (0.0015)	0.2735 (0.0016)

[a] Due to the variation in genetic length of different chromosomes, the number is not 22.

the true hypothesis, i.e. *unrelated*, see Table 2. This value is certainly not convincing in forensic genetics, nor should it be in medical genetic research. We applied our findings, based on tests using data from known relationships, on two families concerning a common paternal ancestor two generations back. The results from Merlin and FEST were unambiguous and showed that the two families did not share a common ancestor and thus, according to our findings, are unrelated.

A 2nd cousin relationship appears to be the limitation to what can be determined with current methods, or by any means presently available. It is debatable what the term unrelated really stands for [42]. The genetic material is quickly diluted as each

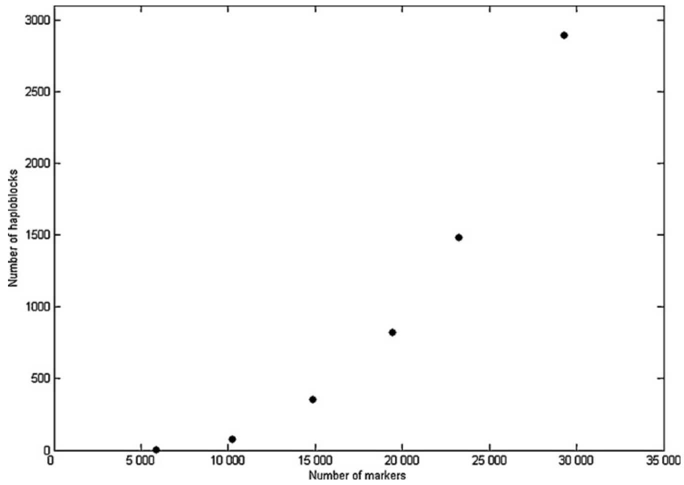

Fig. 4. Graph displaying the number of haploblocks (y-axis) versus the number of markers (x-axis). The number of haploblocks for each selection of markers has been calculated using PLINK. The graph displays an approximate exponential relationship between the number of haploblocks and the number of markers. The dot at 5865 markers corresponds to a distance of at least 0.5 cM between two markers, while the dot at 29,277 markers corresponds to a distance of at least 0.05 cM, see also Table 5.

generation passes. Perhaps the average background relatedness, shared by all individuals of the same ethnicity, lies not very far from the 3rd cousin relationship. Indeed the latest release from the HapMap project demonstrates that two unrelated individuals in the CEU population share in average 0.34% of their alleles through identity by descent (IBD) [12]. This is in fact approximately equal to the expected sharing of alleles (IBD) between two 3rd cousins. To investigate this further, more families need to be analyzed, where relationships such as half siblings, half cousins and half 2nd cousins are known. Simulation studies can be performed but they are more complicated since they raise the issue of how to model and account for linkage disequilibrium. For example, relationships can be simulated based on true haplotypes, where the issue of how to model LD in simulations is irrelevant, but haplotypes are complicated and computer demanding to infer. PHASE and IMPUTE, as well as similar available software, offer the advantage of inferring haplotypes from genotyping data, without any family or pedigree information [43,44]. We simulated relationships where instead the founder haplotypes were created using an approximate LD map from the HapMap project and the results agreed with our previous findings.

Regarding the statistical calculation, we suggest creating a new model, or modifying an existing one, which accounts for linkage disequilibrium. LD might be turned into an advantage if a proper model is developed. Moreover, other algorithms should be considered, i.e. other than Lander–Green, which is used in Merlin. Indeed, algorithms that can handle large and complex pedigrees with a large number of markers should be evaluated. For large and complex pedigrees, with thousands of markers, approximate approaches, such as Monte-Carlo Markov chain (MCMC), utilized in the software MORGAN for example, might be a good candidate. [15,45]. The existence of block-like structures, with clusters of tightly linked SNPs may also prove useful [26,46]. Merlin provides the possibility to calculate likelihoods based on specified cluster information [47]. Although theoretically promising the current implementation of the method in Merlin was, in our study, unable to handle more extended pedigrees with an average amount of clusters, i.e. 3rd cousins and 5000 clusters.

There are in addition alternative methods for the determination of the most probable relationship between individuals. One such approach is utilizing identity by state (IBS). This approach may not be optimal from a statistical point of view, but can nevertheless be useful to illustrate distant relationships [48,49].

In conclusion, genotype data from high-density SNP arrays have proved to be useful in the investigation of distant genetic relationships. In this study we solved a real case of half 1st cousinship using different selections of SNP markers. Relationships as distant as 2nd cousins could also be unambiguously resolved. However, 3rd cousins and more distant relationships revealed hard to distinguish from unrelated. Nevertheless, this task should not be insurmountable using a good computer algorithm and enough reference material to work with. Parameters such as genotyping errors and LD should be more thoroughly investigated as well as IBS approaches. Our conclusions regarding the relation between the two families (Figs. 1 and 2) are primarily based on a small number of established relationships (Tables 1 and 2) and thus further simulations and families are needed to verify our results. Even so, we are confident that our methods can be used to solve other cases of disputed distant family relationships.

Acknowledgements

Thanks to Prof. Peter Söderkvist, Department of Clinical and Experimental Medicine, Linköping University, for kindly providing the population reference material and to Master Student Iftikhar Ahmad, Department of Mathematical Science, Chalmers University, for the work on genotyping errors. We would also like to thank the two families, without whom this study would not have been possible.

Disclaimer: The study was funded by the National Board of Forensic Medicine, Sweden, Grant No. 4155225; Center of Forensic Science, Linköping University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2011.07.007.

References

[1] D.L. Newman, M. Abney, M.S. McPeek, et al., The importance of genealogy in determining genetic associations with complex traits, Am. J. Hum. Genet. 69 (2001) 1146–1148.
[2] B.F. Voight, J.K. Pritchard, Confounding from cryptic relatedness in case-control association studies, PLoS Genet. 1 (2005) e32.
[3] A.L. Leutenegger, E. Genin, E.A. Thompson, et al., Impact of parental relationships in maximum lod score affected sib-pair method, Genet. Epidemiol. 23 (2002) 413–425.
[4] A.O. Karlsson, G. Holmlund, T. Egeland, et al., DNA-testing for immigration cases: the risk of erroneous conclusions, Forensic Sci. Int. Genet. 172 (2007) 144–149.
[5] H. Ellegren, Microsatellites: simple sequences with complex evolution, Nat. Rev. Genet. 5 (2004) 435–445.
[6] J.M. Butler, Genetics and genomics of core short tandem repeat loci used in human identity testing, J. Forensic Sci. 51 (2006) 253–265.
[7] B. Budowle, A. van Daal, Forensically relevant SNP classes, Biotechniques 44 (603–608) (2008) 610.
[8] J.C. Glaubitz, O.E. Rhodes, J.A. Dewoody, Prospects for inferring pairwise relationships with single nucleotide polymorphisms, Mol. Ecol. 12 (2003) 1039–1047.
[9] J.J. Sanchez, C. Phillips, C. Borsting, et al., A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713–1724.
[10] C. Borsting, J.J. Sanchez, H.E. Hansen, et al., Performance of the SNPforID 52 SNP-plex assay in paternity testing, Forensic Sci. Int. Genet. 2 (2008) 292–300.
[11] D.E. Reich, S.F. Schaffner, M.J. Daly, et al., Human genome sequence variation and the influence of gene history, mutation and recombination, Nat. Genet. 32 (2002) 135–142.
[12] K.A. Frazer, D.G. Ballinger, D.R. Cox, et al., A second generation human haplotype map of over 3.1 million SNPs, Nature 449 (2007) 851–861.
[13] S.F. Grant, H. Hakonarson, Microarray technology and applications in the arena of genome-wide association, Clin. Chem. 54 (2008) 1116–1124.
[14] M. Nothnagel, J. Schmidtke, M. Krawczak, Potentials and limits of pairwise kinship analysis using autosomal short tandem repeat loci, Int. J. Legal Med. 124 (2010) 205–215.

[15] G. Gao, D.B. Allison, I. Hoeschele, Haplotyping methods for pedigrees, Hum. Hered. 67 (2009) 248–266.

[16] R.C. Elston, J. Stewart, A general model for the genetic analysis of pedigree data, Hum. Hered. 21 (1971) 523–542.

[17] T. Egeland, P.F. Mostad, B. Mevag, et al., Beyond traditional paternity and identification cases. Selecting the most probable pedigree, Forensic Sci. Int. Genet. 110 (2000) 47–59.

[18] E.S. Lander, P. Green, Construction of multilocus genetic linkage maps in humans, Proc. Natl. Acad. Sci. U. S. A. 84 (1987) 2363–2367.

[19] G.R. Abecasis, S.S. Cherny, W.O. Cookson, et al., Merlin—rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (2002) 97–101.

[20] O. Skare, N. Sheehan, T. Egeland, Identification of distant family relationships, Bioinformatics 25 (2009) 2376–2382.

[21] T. Egeland, N. Sheehan, On identification problems requiring linked autosomal markers, Forensic Sci. Int. Genet. 2 (2008) 219–225.

[22] S. Purcell, B. Neale, K. Todd-Brown, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (2007) 559–575.

[23] G. Montana, HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients, Bioinformatics 21 (2005) 4309–4311.

[24] Rutgers Combined Linkage-Physical Map of The Human Genome, http://compgen.rutgers.edu/RutgersMap/DownloadMap.aspx (accessed 18.05.11).

[25] B. Lindblom, G. Holmlund, Rapid DNA purification for restriction fragment length polymorphism analysis, Gene Anal. Tech. 5 (1988) 97–101.

[26] J. Ge, B. Budowle, J.V. Planz, et al., Haplotype block: a new type of forensic DNA markers, Int. J. Legal Med. 124 (2010) 353–361.

[27] S.B. Gabriel, S.F. Schaffner, H. Nguyen, et al., The structure of haplotype blocks in the human genome, Science 296 (2002) 2225–2229.

[28] M. Krawczak, Informativity assessment for biallelic single nucleotide polymorphisms, Electrophoresis 20 (1999) 1676–1681.

[29] F. Pompanon, A. Bonin, E. Bellemain, et al., Genotyping errors: causes, consequences and solutions, Nat. Rev. Genet. 6 (2005) 847–859.

[30] E. Sobel, J.C. Papp, K. Lange, Detection and integration of genotyping errors in statistical genetics, Am. J. Hum. Genet. 70 (2002) 496–508.

[31] M.P. Epstein, W.L. Duren, M. Boehnke, Improved inference of relationship for pairs of individuals, Am. J. Hum. Genet. 67 (2000) 1219–1231.

[32] D.W. Gjertson, C.H. Brenner, M.P. Baur, et al., ISFG: recommendations on biostatistics in paternity testing, Forensic Sci. Int. Genet. 1 (2007) 223–231.

[33] M. Tracey, Short tandem repeat-based identification of individuals and parents, Croat. Med. J. 42 (2001) 233–238.

[34] M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.

[35] J. Ge, A. Eisenberg, J. Yan, et al., Pedigree likelihood ratio for lineage markers, Int. J. Legal Med. 125 (2011) 519–525.

[36] R. Szibor, X-chromosomal markers: past, present and future, Forensic Sci. Int. Genet. 1 (2007) 93–99.

[37] Q. Huang, S. Shete, C.I. Amos, Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis, Am. J. Hum. Genet. 75 (2004) 1106–1112.

[38] J.M. Keith, A. McRae, D. Duffy, et al., Calculation of IBD probabilities with dense SNP or sequence data, Genet. Epidemiol. 32 (2008) 513–519.

[39] A. Kurbasic, O. Hossjer, A general method for linkage disequilibrium correction for multipoint linkage and association, Genet. Epidemiol. 32 (2008) 647–657.

[40] A.O. Tillmar, T. Egeland, B. Lindblom, et al., Using X-chromosomal markers in relationship testing: calculation of likelihood ratios taking both linkage and linkage disequilibrium into account, Forensic Sci. Int. Genet. (2010), doi:10.1016/j.fsigen.2010.11.004.

[41] A. Manichaikul, J.C. Mychaleckyj, S.S. Rich, et al., Robust relationship inference in genome-wide association studies, Bioinformatics 26 (2010) 2867–2873.

[42] B.S. Weir, A.D. Anderson, A.B. Hepler, Genetic relatedness analysis: modern data and new challenges, Nat. Rev. Genet. 7 (2006) 771–780.

[43] M. Stephens, N.J. Smith, P. Donnelly, A new statistical method for haplotype reconstruction from population data, Am. J. Hum. Genet. 68 (2001) 978–989.

[44] J. Marchini, B. Howie, S. Myers, et al., A new multipoint method for genome-wide association studies by imputation of genotypes, Nat. Genet. 39 (2007) 906–913.

[45] A.W. George, E.M. Wijsman, E.A. Thompson, MCMC multilocus lod scores: application of a new approach, Hum. Hered. 59 (2005) 98–108.

[46] K. Zhang, P. Calabrese, M. Nordborg, et al., Haplotype block structure and its applications to association studies: power and study designs, Am. J. Hum. Genet. 71 (2002) 1386–1394.

[47] G.R. Abecasis, J.E. Wigginton, Handling marker–marker linkage disequilibrium: pedigree analysis with clustered markers, Am. J. Hum. Genet. 77 (2005) 754–767.

[48] H. Miyazawa, M. Kato, T. Awata, et al., Homozygosity haplotype allows a genome-wide search for the autosomal segments shared among patients, Am. J. Hum. Genet. 80 (2007) 1090–1102.

[49] E.D. Roberson, J. Pevsner, Visualization of shared genomic regions and meiotic recombination in high-density SNP data, PLoS One 4 (2009) e6711.

Paper II

Using Object Oriented Bayesian Networks to Model Linkage, Linkage Disequilibrium and Mutations between STR Markers

Daniel Kling[1,2]*, **Thore Egeland**[1,2], **Petter Mostad**[3]

[1] Department of Family Genetics, Norwegian Institute of Public Health, Oslo, Norway, [2] Department for Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Aas, Norway, [3] Mathematical Sciences, Chalmers University of Technology and Mathematical Sciences, Gothenburg, Sweden

Abstract

In a number of applications there is a need to determine the most likely pedigree for a group of persons based on genetic markers. Adequate models are needed to reach this goal. The markers used to perform the statistical calculations can be linked and there may also be linkage disequilibrium (LD) in the population. The purpose of this paper is to present a graphical Bayesian Network framework to deal with such data. Potential LD is normally ignored and it is important to verify that the resulting calculations are not biased. Even if linkage does not influence results for regular paternity cases, it may have substantial impact on likelihood ratios involving other, more extended pedigrees. Models for LD influence likelihoods for all pedigrees to some degree and an initial estimate of the impact of ignoring LD and/or linkage is desirable, going beyond mere rules of thumb based on marker distance. Furthermore, we show how one can readily include a mutation model in the Bayesian Network; extending other programs or formulas to include such models may require considerable amounts of work and will in many case not be practical. As an example, we consider the two STR markers vWa and D12S391. We estimate probabilities for population haplotypes to account for LD using a method based on data from trios, while an estimate for the degree of linkage is taken from the literature. The results show that accounting for haplotype frequencies is unnecessary in most cases for this specific pair of markers. When doing calculations on regular paternity cases, the markers can be considered statistically independent. In more complex cases of disputed relatedness, for instance cases involving siblings or so-called deficient cases, or when small differences in the LR matter, independence should not be assumed. (The networks are freely available at http://arken.umb.no/~dakl/BayesianNetworks.)

Citation: Kling D, Egeland T, Mostad P (2012) Using Object Oriented Bayesian Networks to Model Linkage, Linkage Disequilibrium and Mutations between STR Markers. PLoS ONE 7(9): e43873. doi:10.1371/journal.pone.0043873

Editor: Frank Emmert-Streib, Queen's University Belfast, United Kingdom

Received May 25, 2012; **Accepted** July 26, 2012; **Published** September 11, 2012

Funding: The work was funded by the Norwegian Institute of Public Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: daniel.kling@fhi.no

Introduction

There are several areas of applications motivating this paper. The general problem is to determine the most likely pedigree and in this paper we discuss models to achieve this goal. It is well known that linkage analysis performed to locate disease mutations may be misguided if the pedigree is incorrectly specified as will be the case if for instance false paternities are not detected. Similarly, association analyses frequently assume that all individuals are unrelated and again deviations from this assumption may affect conclusions. In forensic cases, for instance paternity cases or identification following disasters, establishing the most likely pedigree is the main objective. Traditionally forensic applications have been based on unlinked markers in linkage equilibrium. For some applications however, these assumptions have been questioned [1,2,3] Furthermore, the conventional markers used in forensics may not have sufficient power to resolve some cases, e.g. family relationships involving more distant relations than siblings [4,5,6]. It is therefore an urgent need to consider methods and practical implementations for more general markers and this is the main objective for this paper.

The evidence is conventionally summarized by the LR (likelihood ratio) [7]. The LR is the probability of the data given one hypothesis (for instance that a specific man is the father) divided by the probability conditioned on an alternative hypotheses (for instance that some unknown man is the father). A large value of the LR results in a man being declared to be a father. In immigration cases, LR calculations can be decisive when decisions are made on granting immigration. It follows that biased LR calculations resulting from unwarranted assumptions may have serious consequences. As far as we know, methods and implementations accounting for linkage, linkage disequilibrium and mutation have not previously been presented.

A forensic example involving two short tandem repeat (STR) loci, D12S391 and vWa, will serve as a motivating case. These markers are located on chromosome 12 only 6.3 Mb apart, but the genetic distance has been estimated to be as large as 10.8 cM [3]. Following the introduction of D12S391 to the new European forensic standard set [8], questions has been raised as to whether the markers can be considered statistically independent when assessing the evidence in specific cases. In addition, studies have been performed to determine whether the physical proximity of the markers has caused linkage disequilibrium (LD) and whether this should be taken into account [1,3]. Moreover, Phillips *et al.* recently published an overview of the commercial STR kits describing several pairs of markers separated by less than 50 cM [9]. Commonly used software for likelihood ratio calculations, such as Familias [10] and DNAView [11] do not consider linkage or linkage disequilibrium in statistical calculations. Although programs exist which model linkage, they are often more complicated to use and it may be necessary to navigate a command line user-interface, e.g. Merlin [12,13]. In addition, to our knowledge, there is no complete model which simultaneously handles linkage, LD and mutations.

Object Oriented Bayesian Networks (OOBN) may provide an alternative solution with an appealing graphical interface. The object-oriented approach also provides a simple user-interface, hiding the complexities within the objects (nodes) [14]. In the model, the nodes contain sub-structures such as states, conditional probability tables and so forth. The nodes are connected to other nodes and the interplay is governed by probabilities within each node.. Several studies have already shown the advantages of using OOBN in forensic contexts [15,16,17,18,19,20]. Taroni *et al.* [15] offers a thorough introduction to the basic methodology. We used the freeware GeNIe (http://genie.sis.pitt.edu) to create the Bayesian networks. One alternative is the commercially available Hugin (http://www.hugin.com).

In this paper we model linkage, linkage disequilibrium, and mutations in a single Bayesian network (BN), freely available at http://arken.umb.no/~dakl/BayesianNetworks/. We present networks for some basic relationships, but the model can easily be extended to other pedigrees as well. In addition to previous investigations, this provides an alternative approach to the study of LD between D12S391 and vWa, but also more generally when studying pairs of linked STR markers. In contrast to other studies, which often measures the disequilibrium, or association of alleles, in terms of an r^2 value or a *p*-value depending on a sample size, our intention was to investigate the effects of LD on actual cases.

Materials and Methods

In order to model linkage disequilibrium (LD), haplotype probabilities must be estimated. A simplified model was constructed (Tillmar *et al.* [21]), based on a Dirichlet distribution, providing non-zero probability estimates also for unseen haplotypes. Specifically, a diallelic haplotype probability f_{ij} was estimated with $f_{ij}=(c_{ij}+\lambda p_i q_j)/(C+\lambda)$ where c_{ij} is the observed count of the haplotype among C unrelated individuals, p_i and q_j are the allele frequencies of the two alleles, and λ is a constant, set to 1 in the computations below. Further, to incorporate this into a Bayesian Network (BN) the haplotype probabilities were used to construct conditional allele probabilities, i.e. based on what allele is observed at the first locus we estimated the conditional probability of observing each allele at the second locus.

In order to obtain haplotype counts, we used data from regular trio paternity cases. When the parenthood is established and no mutations are present, the phase, i.e., the haplotypes can be deduced for the child using a simple algorithm. There are, however, ambiguous cases where the haplotypes cannot be determined for the child, e.g. when the parents and the child are all heterozygous for the same alleles. Out of 450 selected trios, 6 where discarded due to more than one possible haplotype configuration. As these ambiguous cases constitute only 1.3% of the total cases, it was not considered to bias the calculations enough to influence the conclusions. Notice that the phased haplotypes for the father and mother, based on the child's genotypes, are generally unknown since recombination might have occurred. Although reasonable estimates of the parents' haplotypes can be obtained, e.g. through the EM-algorithm or Gibbs sampling (PHASE by Stephens *et al.* and IMPUTE2 by Howie *et al.*, [22,23]), we found that haplotype probabilities computed this way did not differ much from those based on the children and therefore used the latter for simplicity (data not shown). Moreover, it is well known that the LD as measured by D declines with (1-recombination rate) per generation and hence,one generation will only have a minor impact on the disequilibrium.

Data

A selection of 444 unrelated Norwegian trios were used to estimate allele and haplotype probabilities at the STR loci D12S391 and vWa (using only the genotypes from the children). Table 1 describes the allele frequencies; in total 8 different alleles were observed at vWa and 16 different alleles at D12S391. To estimate haplotype probabilities, the number of observations for each haplotype was first counted (using the data from the children). In total, 100 different haplotypes were observed out of 128 possible. Haplotype probabilities were then estimated as described above. (Tables S1 and S2 provide further details on the observed haplotype frequencies and the estimated haplotype probabilities). To calculate the conditional probability of each D12S391 allele given a specific vWa allele, each column in Table S1, containing the observed haplotype probabilities, is normalized to 1. Table 2 describes the calculated conditional allele probabilities. Conditioning rather on D12S391 would of course lead to the same results.

Network

A simple Bayesian network describing a paternity case is illustrated in Fig. 1, the network is more or less self-explanatory and presents the given problem in a intuitive way. It is worth pointing out that as more parameters (i.e. recombination rates, LD

Table 1. Sample allele frequencies for STR loci vWa and D12S391, based on 444 unrelated Norwegian individuals.

	vWa	D12S391
14	0.08896	
15	0.0732	0.04392
16	0.21621	0.02252
17	0.30968	0.12387
17.3		0.01351
18	0.1982	0.19369
18.3		0.01351
19	0.10248	0.10698
19.3		0.01126
20	0.10135	0.10811
21	0.00114	0.10248
22		0.01149
23		0.09234
24		0.03829
25		0.00901
26		0.00338
27		0.00225

doi:10.1371/journal.pone.0043873.t001

Table 2. Conditional allele probabilities for the alleles at D12S391 given the allele at vWa.

	14	15	16	17	18	19	20	21
15	0.038049	0.030969	0.036497	0.043637	0.056745	0.054825	0.004392	0.021959
16	0.000282	0.030644	0.020842	0.029067	0.028376	0.011114	0.002252	0.011261
17	0.176548	0.062483	0.156082	0.112768	0.091095	0.131781	0.312387	0.061937
17.3	0.000169	0.000205	0.015614	0.018165	0.017026	0.011017	0.001351	0.006757
18	0.177421	0.199904	0.177169	0.207224	0.221433	0.154279	0.119369	0.096847
18.3	0.012669	0.015356	0.005251	0.014542	0.028325	0.000147	0.001351	0.006757
19	0.138837	0.062227	0.114544	0.09459	0.113599	0.131598	0.010698	0.053491
19.3	0.000141	0.015322	0.015602	0.018157	0.005713	0.000122	0.001126	0.005631
20	0.076351	0.168305	0.083462	0.090971	0.136204	0.142479	0.110811	0.054054
21	0.126281	0.153068	0.109339	0.098197	0.074025	0.09894	0.110248	0.051239
22	0.076437	0.122953	0.104222	0.14172	0.096694	0.109945	0.211487	0.057433
23	0.126154	0.062005	0.098924	0.083668	0.079618	0.109699	0.109234	0.546171
24	0.037979	0.061186	0.046831	0.032747	0.039764	0.022155	0.003829	0.019144
25	0.012613	0.015288	0.005228	0.010902	0.005701	0.010968	0.000901	0.004505
26	4.22E-05	5.12E-05	0.01038	1.22E-05	0.005669	3.67E-05	0.000338	0.001689
27	2.82E-05	3.41E-05	1.17E-05	0.003631	1.27E-05	0.010894	0.000225	0.001126

To account for unseen haplotypes, probabilities were estimated using a Dirichlet distribution. Each row indicates the allele at vWa, while each column indicates the allele at D12S391. The table should be interpreted as follows, for a given allele at vWa (top row), the corresponding conditional allele probabilities for D12S391 are given (column).

doi:10.1371/journal.pone.0043873.t002

and mutations), markers and more distant relationships are considered, the network grows in complexity and can become visually incomprehensible. This can be counteracted by rearranging the most relevant network nodes in a simpler way, hiding the complexity from the user. The networks created in this study use a simple naming convention, based on few abbreviations, but larger networks might require shorter node names. All networks are freely available at http://arken.umb.no/~dakl/BayesianNetworks/. In addition we provide a short user manual as well as a software to generate the networks based on your own data.

Two different scenarios were considered; a regular paternity case, Fig. 2 and a case of disputed siblingship, Fig. 3. For each network the user can vary recombination rate, decide whether to use conditional allele probabilities, based on Table 2, or allele frequencies, see Table 1. In the paternity network the user can decide whether to instantiate the mother's genotypes (trio) or to leave them unknown (duo), i.e. to use the allele frequencies. In the sibling network the hypotheses compare whether the two persons are unrelated or full siblings. (A separate network was also constructed for a halfsibling case when the siblings are known to share the same mother, see Fig. S1.) The parents' genotypes can be instantiated if available, otherwise allele frequencies will be used. The network in Fig. 1 is in principle equal to the one described by Taroni *et al.* for a paternity case [15]. The main differences lie in the existence of a *Recombination* node as well as an *LD* node. The *Recombination* node describes the probability for a cross-over to occur, i.e. the recombination rate. Also for each possible inheritance of a D12S391 allele, the *P/M* nodes transmit whether the **Paternal** or **Maternal** vWa allele have been passed on. The *LD* node is also connected to each possible inheritance of a D12S391 allele. If the *LD* node is instantiated to **Yes**, conditional allele probabilities will be used. The *Mutation* nodes contain a transition matrix. In this study a simple mutation model was used, where each transition has an equal probability of occurring, i.e. $\mu/(n-1)$, where μ is the mutation rate and n is the number of alleles. Mutation rates for each locus were obtained from a local database. The *Child Paternal Allele* (CPA) nodes are subject to the *Hypothesis* node (Either *Is Father* or *Are Siblings* depending on the network), with states **Yes** and **No**. The *Hypothesis* node will display the posterior probabilities for the given relationships. The tables for the CPA nodes are based on the Alleged father given that he is the father and the allele frequencies if he is not the father. Also, if the *LD* node is set to **Yes**, conditional allele probabilities for the D12S391 allele will be used. (Please see user manual for a more complete description.)

Results

The networks were tested on a selection of real cases where the likelihood ratio (LR), assuming marker independence, had already been calculated using the software Familias [10]. In addition an attempt was made to create a worst-case-scenario (WCS) regarding linkage disequilibrium, i.e.,selecting the haplotypes where the observed haplotype frequencies deviated maximally from the expected haplotype frequencies, see Table S1 and S2. The genotypes used in the WCS include rare alleles and as a consequence also often unobserved haplotypes. Table 3 describes the results from the likelihood ratio calculations. Each case was investigated using three different methods. The method denoted M1 in Table 3 is equivalent to the most commonly used approach in forensic laboratories, where the markers vWa and D12S391 are considered to be independent, i.e. recombination rate of 50%, and allele frequencies are utilized. In the two remaining methods, denoted M2 and M3 in Table 3, a recombination rate of 9% was used in accordance with previous studies by Budowle *et al.* [3]. In addition, the decision of whether to use conditional allele probabilities were evaluated, using in M2 allele frequencies (Table 1) and in M3 conditional allele probabilities (Table 2). Quotients between the LR values obtained using each method are

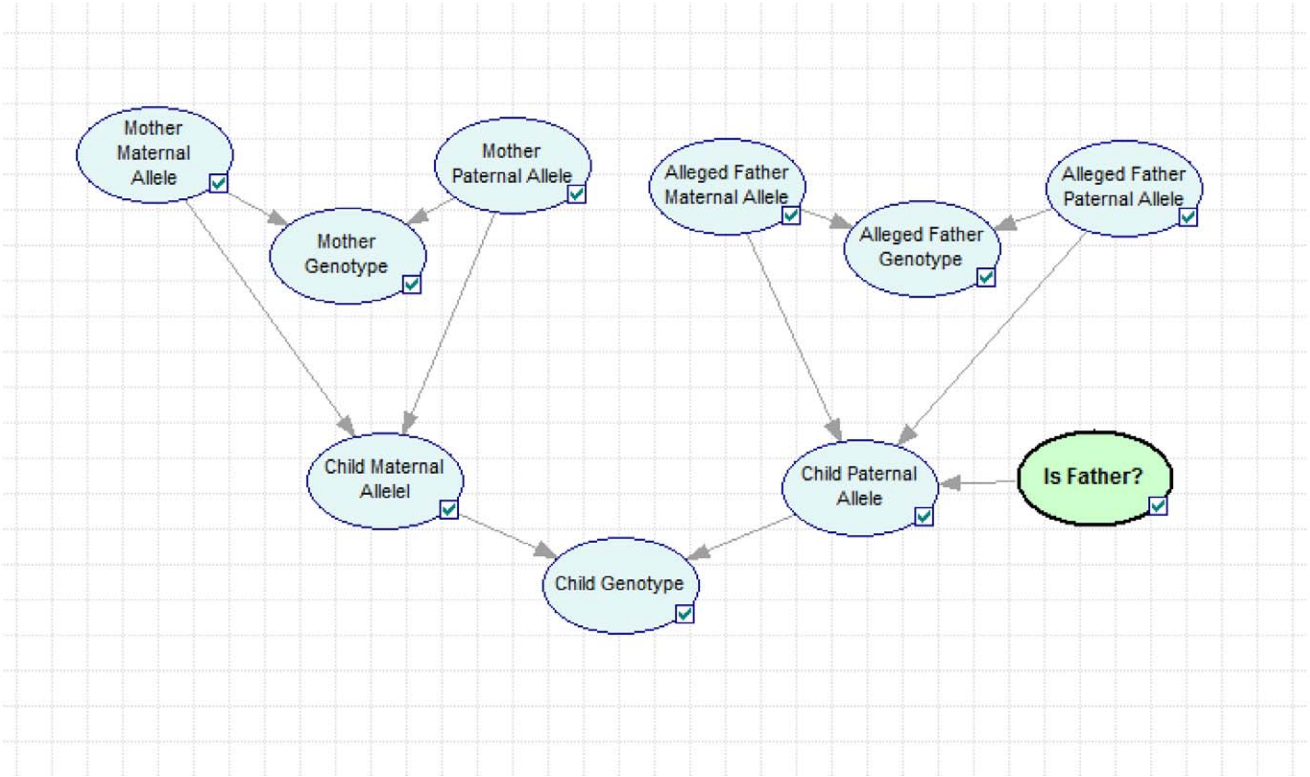

Figure 1. Bayesian network describing the basic layout for a paternity case.
doi:10.1371/journal.pone.0043873.g001

included in Table 3. (Note that M2 is not relevant in standard duo/trio cases since recombination alone does not effect the statistical calculations)

To further test the method, we also created a network where instead of using data from D12S391 and vWa we used data from two other closely located markers, D5S818 and CSF1PO (Table 4).

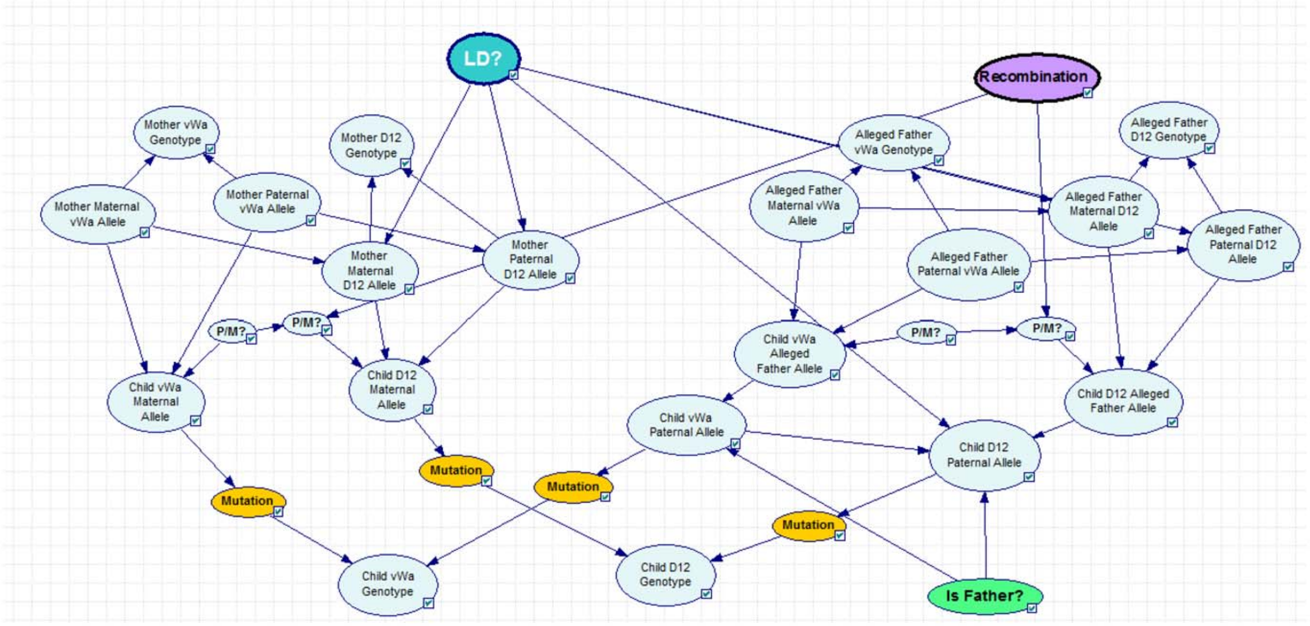
LD?
Recombination
Mother vWa Genotype
Mother D12 Genotype
Alleged Father vWa Genotype
Alleged Father D12 Genotype
Mother Maternal vWa Allele
Mother Paternal vWa Allele
Alleged Father Maternal vWa Allele
Alleged Father Maternal D12 Allele
Mother Maternal D12 Allele
Mother Paternal D12 Allele
Alleged Father Paternal vWa Allele
Alleged Father Paternal D12 Allele
P/M?
P/M?
Child vWa Alleged Father Allele
P/M?
P/M?
Child vWa Maternal Allele
Child D12 Maternal Allele
Child D12 Alleged Father Allele
Child vWa Paternal Allele
Child D12 Paternal Allele
Mutation
Mutation
Mutation
Mutation
Child vWa Genotype
Child D12 Genotype
Is Father?

Figure 2. Bayesian network describing a paternity case. The *Recombination* node contains the probability for a recombination to occur, i.e., the recombination rate. The nodes P/M tell whether the vWa paternal or maternal allele is inherited. The LD node is connected to the paternal and maternal allele nodes and decides whether or not to use conditional allele probabilities. Furthermore, the node *Is Father?* contains the different hypotheses.
doi:10.1371/journal.pone.0043873.g002

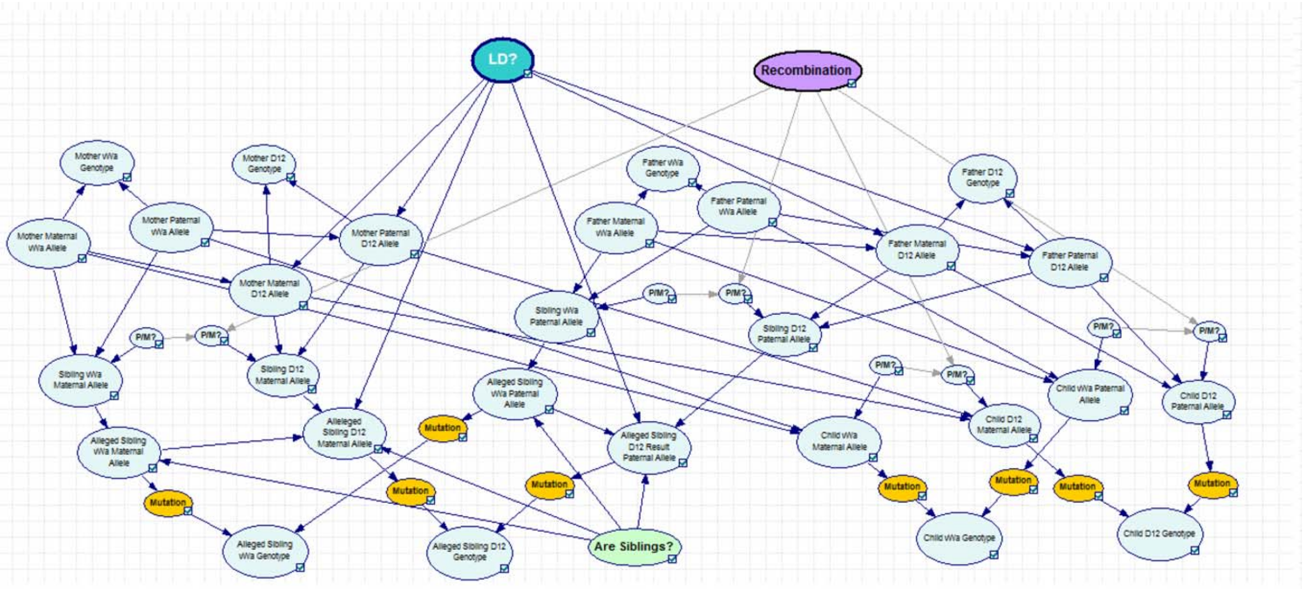
LD?
Recombination
Mother vWa Genotype
Mother D12 Genotype
Father vWa Genotype
Father D12 Genotype
Mother Paternal vWa Allele
Mother Maternal vWa Allele
Mother Paternal D12 Allele
Mother Maternal D12 Allele
Father Maternal vWa Allele
Father Paternal vWa Allele
Father Maternal D12 Allele
Father Paternal D12 Allele
Sibling vWa Paternal Allele
Sibling D12 Paternal Allele
Sibling vWa Maternal Allele
Sibling D12 Maternal Allele
P/M?
P/M?
P/M?
P/M?
P/M?
P/M?
P/M?
P/M?
Alleged Sibling vWa Paternal Allele
Alleleged Sibling D12 Maternal Allele
Alleged Sibling vWa Maternal Allele
Alleged Sibling D12 Result Paternal Allele
Child vWa Paternal Allele
Child D12 Paternal Allele
Child D12 Maternal Allele
Child vWa Maternal Allele
Mutation
Mutation
Mutation
Mutation
Mutation
Mutation
Mutation
Mutation
Alleged Sibling vWa Genotype
Alleged Sibling D12 Genotype
Are Siblings?
Child vWa Genotype
Child D12 Genotype

Figure 3. Bayesian network describing a sibling case. The nodes P/M tell whether the vWa paternal or maternal allele is inherited. The P/M nodes connected to the D12S391 allele also contains the recombination frequency. The LD node is connected to the paternal and maternal allele nodes and decides whether or not to use conditional allele probabilities. Furthermore, the node *Are Siblings?* contains the different hypotheses.
doi:10.1371/journal.pone.0043873.g003

A recombination rate of 0.3 was used, close to the value obtained using any of the mapping functions. The results reveal that, even when comparing two markers accepted to be in LE, discrepancies can be detected. Future studies should be conducted involving markers known to be in LD. Our network can of course be extended to include more linked markers in LD.

Discussion

We have demonstrated the application of Object Oriented Bayesian Networks in modeling linkage, linkage disequilibrium and mutations in cases of disputed genetic relatedness. As an example, we present data from a pair of STR markers, vWA and D12S391, recently studied with regards to possible linkage disequilibrium. Two different networks were created to investigate a selection of actual cases as well as fictional, see Worst Case Scenarios in Table 3. The small differences in calculated LRs between method M1 (not considering linkage and LD) and the *Comparison* are due to the use of slightly different allele frequency databases, where the *Comparison* LR has been calculated using a Norwegian population database utilized in routine casework. However, it is notable that the differences between the results using any of M1, M2 (10% recombination rate and LD not considered) or M3 (10% recombination rate and LD is considered) are in many cases comparable to the differences between M1 and the *Comparison* methods. Consequently, the differences between method M2 and M3, allele frequencies versus conditional allele probabilities, can perhaps be considered as merely a small bias in the estimation of allele frequencies.

Since linkage has previously been measured between vWa and D12S391, the most important concern of this paper is to evaluate the effect of using conditional allele probabilities as measured by the quotient between the LR values obtained using methods M3 and M2, see Table 3. None of the real cases display a quotient LR_{M2}/LR_{M3} larger than 2, and for most of the cases the quotient is close to 1. Also, the Worst Case Scenarios do not display quotients larger than 4. We should of course always expect some differences since no data will indicate exact linkage equilibrium (Table 4). Whereas our study has only included a small selection of real cases, we are aware that larger studies considering hundreds of cases should be conducted and also that our results, concerning possible LD between vWA and D12S391, are partly anecdotal. A recent paper by Gill *et al.* provides further evidence and discussion on the matter [2].

Haplotype frequencies are generally hard to estimate as genotype data does not normally indicate which chromosome, i.e. paternal or maternal, each allele is located on. New methods, such as mass-sequencing provide means to determine each chromosomal setup, but given current forensic casework, using STR markers, one might instead rely on the massive amount of available data from families (trios mainly) where the haplotypes from the children can, in most cases, be unambiguously determined, as long as the possibility of mutation is disregarded. In our study we used 444 phased unrelated children, i.e. 888 haplotypes, to determine the observed as well as expected haplotype frequencies. We observed 100 of a total of 128 possible haplotypes. An important consideration is if this is enough material for a reliable estimation of population haplotype frequencies? In particular, can we reliably estimate the probability of observing a haplotype that has not been observed in the database? The same dilemma exists when previously unseen or new alleles are observed in regular genotyping, but for haplotypes one may use allele frequencies to construct a reasonable guess at a probability. Our formula contains a parameter λ which loosely corresponds to the pseudo-counts often used in the estimation of population allele frequencies. Although a value for λ might be estimated for data, we have simply used $\lambda = 1$. This gives the initial estimates, constructed as products of allele frequencies, the same weight as a single haplotype observation, leading to fairly small estimates of conditional probabilities for unobserved haplotypes.

Table 3. Comparison of calculated likelihood ratios (LR) based on the genotype data from STR loci vWa and D12S391, on a selection of real cases.

Case id	M1	M2	M3	Comparison	$LR_{M1}/LR_{Comparison}$	LR_{M1}/LR_{M2}	LR_{M2}/LR_{M3}
Duos							
1	3.608	3.608	1.909	3.78	0.954	-	1.89
2	3.038	3.038	2.769	3.099	0.98	-	1.097
3	25.455	25.455	35.036	24.243	1.05	-	0.727
4	8.723	8.723	9.638	9.447	0.923	-	0.905
5	8.93	8.93	10.792	9.036	0.988	-	0.827
6	39.487	39.487	51.46	41.563	0.95	-	0.767
7	11.761	11.761	11.859	10.631	1.106	-	0.992
8	2.943	2.943	3.721	2.66	1.106	-	0.791
9	5.956	5.956	6.457	6.463	0.922	-	0.922
10	6.81	6.81	8.912	6.815	0.999	-	0.764
WCS	750.879	750.879	308.597	404	1.859	-	2.433
Trios							
11	5.567	5.567	5.055	5.239	1.063	-	1.101
12	96.809	96.809	107.696	89.208	1.085	-	0.899
13	11.626	11.626	7.026	10.834	1.073	-	1.655
14	87.652	87.652	52.191	54.74	1.601	-	1.679
15	8.32	8.32	7.772	9.498	0.876	-	1.071
16	29.479	29.479	21.491	28.919	1.019	-	1.372
17	6.214	6.214	7.347	6.624	0.938	-	0.846
18	11.234	11.234	9.624	11.628	0.966	-	1.167
19	24.483	24.483	33.811	24.8	0.987	-	0.724
20	11.635	11.635	12.358	10.827	1.075	-	0.941
WCS	2917.855	2917.855	736.46	2130	1.37	-	3.962
Siblings							
21	9.917	7.097	9.732	9.766	1.015	1.397	0.729
22	0.264	0.287	0.296	0.405	0.652	0.92	0.97
23	38.841	62.98	71.993	38.331	1.013	0.617	0.875
24	0.351	0.339	0.314	0.34	1.032	1.035	1.08
25	1.331	1.584	1.439	1.331	1	0.84	1.101
26	0.46	0.621	0.633	0.455	1.011	0.741	0.981
27	0.378	0.354	0.363	0.38	0.995	1.068	0.975
28	0.83	0.622	0.612	0.815	1.018	1.334	1.016
29	8.61	10.962	11.92	9.1278	0.943	0.785	0.92
30	13.772	19.825	19.367	13.763	1.001	0.695	1.024
WCS	200.938	298.868	134.619	115.694	1.737	0.672	2.22

Three different methods have been used, denoted M1, M2 and M3. M1: 50% recombination rate, LD not considered; M2: 10% recombination, LD not considered; M3: 10% recombination, LD taken into consideration. The column *Comparison* is the LR obtained using the software Familias with the *standard* Norwegian population database. WCS. abbreviates Worst Case Scenario and attempts to simulate a case where the likelihood ratios should differ the most due to linkage disequilibrium. The columns to the right display three relevant quotients for each case; Note that the LR calculated using M2 and the quotient LR_{M1}/LR_{M2} is only relevant in the non-paternity cases, since recombination alone will not effect the likelihoods for these cases.

doi:10.1371/journal.pone.0043873.t003

Conclusions

An imminent practical concern for forensic laboratories using closely located STR markers, such as the pair studied in this paper, is how computations should be performed with such data. One issue is whether linkage must be taken into account. Though statistical calculations in regular paternity cases is not affected by linkage and disputed paternities make up the majority of cases for most labs, we believe that in sibling cases and other more extended relationships, linkage should be taken into account. We recommend that forensic labs perform sensitivity calculations and/or simulations to investigate the effect of recombination rate, especially in kinship analyses and deficient paternity cases. The recently released software FamLink provides features to perform such analyses [24]. In addition to STR markers, our model can easily be extended to accommodate SNP data. In fact, the networks available at our repository are able to handle diallelic

Table 4. Comparison of calculated likelihood ratios (LR) based on the genotype data from STR loci D5S818 and CSF1PO, on a selection of cases.

Case id	M1	M2	M3	Comparison	$LR_{M1}/LR_{Comparison}$	LR_{M1}/LR_{M2}	LR_{M2}/LR_{M3}
Duos							
1	1.4632	1.4632	1.5058	1.4215	1.029	-	0.972
2	1.062	1.062	1.034	1.037	1.024	-	1.027
3	4.84	4.84	7.998	5.176	0.935	-	0.605
4	395.668	395.668	362.636	485.808	0.814	-	1.091
5	9.598	9.598	9.016	10.246	0.937	-	1.065
6	74.489	74.489	80.653	100.604	0.74	-	0.924
7	8.072	8.072	8.013	7.734	1.044	-	1.007
8	19.193	19.193	20.172	20.491	0.937	-	0.951
9	49.869	49.869	42.537	55.005	0.907	-	1.172
10	77.215	77.215	121.659	114.202	0.676	-	0.635
W.C.S.	1520.143	1520.143	11036.52	3656	0.416	-	0.138
Trios							
11	40.007	40.007	64.944	48.709	0.821	-	0.616
12	11.369	11.369	11.272	9.947	1.143	-	1.009
13	5.746	5.746	5.577	8.65	0.664	-	1.03
14	101.284	101.284	85.736	63.917	1.585	-	1.181
15	604.62	604.62	383.645	777.506	0.778	-	1.576
16	23.505	23.505	22.616	25.496	0.922	-	1.039
17	76.821	76.821	52.45	87.727	0.876	-	1.465
18	1838.249	1838.249	1964.408	2138.332	0.86	-	0.936
19	394.116	394.116	216.855	346.241	1.138	-	1.817
20	53.305	53.305	69.457	66.978	0.796	-	0.767
W.C.S.	139.278	139.278	709.883	138.139	1.008	-	0.196
Siblings							
21	6.218	5.808	5.02	6.742	0.922	1.071	1.157
22	0.906	0.906	0.93	0.696	1.301	1	0.974
23	4.202	3.99	3.92	3.75	1.121	1.053	1.018
24	3.632	3.343	3.499	2.856	1.272	1.086	0.955
25	0.247	0.265	0.139	0.255	0.968	0.935	1.903
26	6.407	6.407	6.165	4.441	1.443	1	1.039
27	0.158	0.177	0.171	0.154	1.022	0.892	1.037
28	0.256	0.256	0.157	0.16	1.596	1	1.636
29	0.5	0.5	0.548	0.25	2.001	1	0.912
30	0.758	0.758	0.727	0.563	1.347	1	1.043
W.C.S.	23254.65	24999	40649.41	93209.73	0.249	0.93	0.615

Three different methods have been used, denoted M1, M2 and M3. M1: 50% recombination rate and LD not considered. M2: 30% recombination and LD not considered, M3: 30% recombination and LD taken into consideration. The column *Comparison* is the LR obtained using the software Familias with the *standard* Norwegian population database. WCS abbreviates Worst Case Scenario and attempts to simulate a case where the likelihood ratios should differ the most due to linkage disequilibrium. The columns to the right display three relevant quotients for each case; Note that the LR calculated using M2 and the quotient LR_{M1}/LR_{M2} is only relevant in the non-paternity cases, since recombination alone will not effect the likelihoods for these cases.

doi:10.1371/journal.pone.0043873.t004

markers, but to process high throughput data a more automated system is needed.

The other major concern, besides recombination, is whether to use conditional allele probabilities, i.e. to account for linkage disequilibrium. All calculations are affected by the use of such probabilities, even standard paternity and match probability calculations. The effect on the marker pair vWA/D12S391 is, according to our results, reasonably small. In addition, the marker pair D5S818/CSF1PO displays equal deviation from expectation, further corroborating results in previous studies. Moreover, our implementation heavily depends on the estimates of conditional allele probabilities, which are currently fairly uncertain. We have illustrated how estimates can be generated based on data from trios, but clearly much larger datasets are needed to reduce the uncertainty. Furthermore, other models to approach the problem with unseen haplotypes should be considered.

Nevertheless, this paper demonstrates how software implementing Object Oriented Bayesian Networks can be used to assemble and code models reasonably quickly, and how these models can subsequently be used to explore complex questions about the interplay between genetic phenomena such as linkage, LD, and mutations. The models can then in fact be used for relevant computations in actual cases. We present Bayesian networks for two basic relationships, available at http://arken.umb.no/~dakl/BayesianNetworks/, which can be used as prototypes for investigations of linkage and linkage disequilibrium for pairs of closely located STR markers.

Supporting Information

Figure S1 Bayesian network describing a sibling case, where the children are known to share the same mother. The nodes P/M tell whether the vWa paternal or maternal allele is inherited. The P/M node connected to the D12S391 allele also contains the recombination frequency. The LD node is connected to the paternal and maternal allele nodes and decides whether or not to use conditional allele frequencies. Furthermore, the node *Are Siblings?* contains the different hypotheses.
(DOC)

Table S1 Observed haplotype frequencies.
(DOC)

Table S2 Expected haplotype frequencies.
(DOC)

Author Contributions

Conceived and designed the experiments: DK TE PM. Analyzed the data: DK TE PM. Wrote the paper: DK TE PM.

References

1. O'Connor KL, Hill CR, Vallone PM, Butler JM (2010) Linkage disequilibrium analysis of D12S391 and vWA in U.S. population and paternity samples. Forensic Sci Int Genet. 5:538–540
2. Gill P, Phillips C, McGovern C, Bright JA, Buckleton J (2011) An evaluation of potential allelic association between the STRs vWA and D12S391: Implications in criminal casework and applications to short pedigrees. Forensic Sci Int Genet 6:477–486
3. Budowle B, Ge J, Chakraborty R, Eisenberg AJ, Green R, et al. (2011) Population genetic analyses of the NGM STR loci. Int J Legal Med 125: 101–109.
4. Egeland T, Sheehan N (2008) On identification problems requiring linked autosomal markers. Forensic Sci Int Genet 2: 219–225.
5. Skare Ø, Sheehan N, Egeland T (2009) Identification of distant family relationships. Bioinformatics 25: 2376–2382.
6. Kling D, Welander J, Tillmar A, Skare Ø, Egeland T, et al. (2012) DNA microarray as a tool in establishing genetic relatedness–Current status and future prospects. Forensic Sci Int Genet 6: 322–329.
7. Buckleton J, Triggs CM, Walsh SJ (2004) Forensic DNA Evidence Interpretation. Bosa Roca, USA: CRC Press Inc.
8. Gill P, Fereday L, Morling N, Schneider PM (2006) New multiplexes for Europe-amendments and clarification of strategic development. Forensic Sci Int 163: 155–157.
9. Phillips C, Fernandez-Formoso L, García-Magariños M, Porras L, Tvedebrink T, et al. (2010) Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. Forensic Sci Int Genet 5:155–169
10. Egeland T, Mostad PF, Mevåg B, Stenersen M (2000) Beyond traditional paternity and identification cases. Selecting the most probable pedigree. Forensic Sci Int 110: 47–59.
11. Brenner CH (1997) Symbolic kinship program. Genetics 145: 535–542.
12. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97–101.
13. Abecasis GR, Wigginton JE (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 77: 754–767.
14. Koller D, Pfeffer A (1997) Object-oriented Bayesian networks. Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI-97); Morgan Kaufman, San Francisco, California. pp. 302–331.
15. Taroni F, Aitken C, Garbolino P, Biedermann A (2006) Bayesian Networks and Probabilistic Inference in Forensic Science. Chichester: John Wiley & Sons.
16. Hepler AB, Weir BS (2008) Object-oriented Bayesian networks for paternity cases with allelic dependencies. Forensic Sci Int Genet 2: 166–175.
17. Dawid AP, Mortera J, Vicard P (2007) Object-oriented Bayesian networks for complex forensic DNA profiling problems. Forensic Sci Int 169: 195–205.
18. Gomes RR, Campos SV, Pena SD (2009) PedExpert: a computer program for the application of Bayesian networks to human paternity testing. Genet Mol Res 8: 273–283.
19. Biedermann A, Taroni F (2006) A probabilistic approach to the joint evaluation of firearm evidence and gunshot residues. Forensic Sci Int 163: 18–33.
20. Lauritzen S, Sheehan N (2003) Graphical Models for Genetic Analyses. Statistical Science 8: 489–514.
21. Tillmar AO, Egeland T, Lindblom B, Holmlund G, Mostad P (2010) Using X-chromosomal markers in relationship testing: Calculation of likelihood ratios taking both linkage and linkage disequilibrium into account. Forensic Sci Int Genet 5:506–511
22. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978–989.
23. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.
24. Kling D, Egeland T, Tillmar AO (2012) FamLink - A user friendly software for linkage calculations in family genetics. Forensic Sci Int Genet 6:616–620

Paper III

Forensic Science International: Genetics 6 (2012) 616–620

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

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

FamLink – A user friendly software for linkage calculations in family genetics

Daniel Kling [a,b,*], Thore Egeland [b,c], Andreas O. Tillmar [d]

[a] *Norwegian Institute of Public Health, Department of Family Genetics, Oslo, Norway*
[b] *Norwegian University of Life Sciences, Department for Chemistry, Biotechnology and Food Science, Aas, Norway*
[c] *Norwegian Institute of Public Health, Department of Forensic Biology, Oslo, Norway*
[d] *National Board of Forensic Medicine, Department of Forensic Genetics and Forensic Toxicology, Linköping, Sweden*

ARTICLE INFO

Article history:
Received 19 December 2011
Received in revised form 23 January 2012
Accepted 30 January 2012

Keywords:
Linkage
Kinship
FamLink
Complex paternity

ABSTRACT

The present number of STR loci adopted in relationship testing is chiefly limited to unlinked markers, in most cases residing on different chromosomes. In order to solve more complex cases of relatedness, e.g. deficient paternities and disputed sibships, the number of core loci can be extended. The inclusion of multiple loci on the same chromosome will, however, increase the risk of possible linkage between markers. We present a new software, FamLink, freely available from http://www.FamLink.se, that can perform statistical calculations based on pedigree structures and account for linkage between pairs of markers. In addition, FamLink can simulate genotype data in order to study the effect of accounting for linkage or not. We demonstrate the importance of taking linkage properly into account using examples and real cases.

1. Introduction

For relationship testing the present number of commonly used STR loci (normally 15) is sufficient to solve the vast majority of paternity cases, trios and duos. However, there is still a need to expand the number of STR loci in order to solve more complex cases, e.g. full sibs, half sibs, etc. [1] and also to be able to solve paternity issues where a close relative is the alternative man [2].

Lately, a number of new PCR kits have been released, offering the possibility to combine DNA data from up to 40 STR loci (Supp. Data 1). For complex relationship cases, it is important to have access to as many markers as possible, given that more family members are not available, in order to reach a well founded conclusion. There are also cases which cannot be resolved based on DNA from independently segregating loci [3]. A number of the loci, included in the different PCR multiplexes, are located on the same chromosome, and in some cases the genetic distances are so small that that the linkage may have a considerable impact on calculations. Phillips et al. [4] studied pairs of closely located STR loci included in different forensic PCR multiplexes in order to estimate the genetic distance (centiMorgans, cM) between them. Among 29 syntenic loci, 14 pairs were found to have a genetic distance less than 50 cM. The closest pair of loci included in the study (SE33 and D6S1043) is separated by a genetic distance of 4.4 cM, corresponding to a recombination rate of about 4%. Moreover, when using data from the three commercial kits PowerPlex® ESI 17 System (Promega Corp.), Investigator HDplex Kit (Qiagen Inc.) and PowerPlex® 18D System (Promega Corp.) there are 9 pairs of syntenic STRs with a genetic distance less than 50 cM (Supp. Data 1).

When using DNA markers that are physically close on the same chromosome, two concepts are relevant to discuss – genetic linkage and linkage disequilibrium (LD). See Thompson [5] for a general discussion of these concepts and calculations of likelihoods on pedigrees. For the current markers of interest, LD has been reported to have negligible impact on the likelihood ratio (LR) value unless a very recent evolutionary event, such as the admixture of two previously separated populations, has occurred [6]. Genetic linkage, or *linkage*, can be described as the co-segregation of closely located loci within a family and can be measured and discussed in terms of the recombination frequency r. In criminal casework, linkage only has impact on match probability calculations when the alternative hypothesis is a close relative according to Buckleton and Triggs [7]. For relationship testing, however, linkage becomes relevant in the transition probabilities for alleles passing from founder to a child within a pedigree.

Recently, there have been discussions and concerns within the forensic community regarding the inclusion of the STR locus D12S391 as a core locus in the European Standard Set of STR loci [8,9]. The reason for these concerns is that D12S391 is located on the short arm of chromosome 12 only 6.3 megabases (Mb) from the established vWA. Studies have been performed to test the impact of ignoring linkage between D12S391 and vWA and have shown

[*] Corresponding author at: Norwegian Institute of Public Health, Department of Family Genetics, Pb 4404 Nydalen, NO-0403 Oslo, Norway. Tel.: +47 210 77663.
E-mail address: Daniel.kling@fhi.no (D. Kling).

1872-4973/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2012.01.012

that even though the median error is small, the case specific error can be considerable [6,10].

In this paper, we present a statistical freeware (FamLink) that can be used to (1) calculate case specific likelihood ratios for two (or more) hypotheses with observed DNA-data for a pair of linked DNA markers and (2) perform simulations for two or more pedigrees (hypotheses) in order to study the impact of ignoring linkage for a specified pair of linked STR markers. In addition, FamLink can analyze cases involving loops giving rise to complex pedigrees. Such loops may arise because of inbreeding or some individuals having children with different spouses, where the spouses are related (marriage loop). Moreover, several linked markers can be handled based on a Familias file [11]. For this approach, a genetic map of all markers is used. Altogether, FamLink provides relevant functionality previously not easily and freely available to the forensic community.

Below we describe the software, its main functions and also demonstrate its usefulness by applying FamLink on two different examples from routine casework. Furthermore, we make use of the simulation module to study the difference between accounting for linkage and not accounting for linkage between the two STR loci SE33 and D6S1043 for two different case examples. A mathematical derivation of a fairly general case is also included to validate the software.

2. Implementation

2.1. General description of the software and algorithm

The software provides an easy-to-use graphical user interface, which allows for linkage calculation. FamLink uses the Merlin engine [12] for numerical calculations. Merlin, as several similar programs, is based on the Lander–Green algorithm [13]. This algorithm is computationally linear in the number of markers, but not in the complexity of the pedigree. In cases where there is a large number of markers and the pedigree is complex, simulation based methods may be required and then the software MORGAN [14] is an alternative. However, we have found that fairly complex pedigrees can be handled and we have not encountered practical cases where computation time has been a serious problem.

FamLink cannot presently accommodate coancestry (theta) corrections nor mutations. The effect on LR values of coancestry is, however, minor in well-mixed populations [15,16]. When it comes to mutations, these have greatest impact on LR values for pedigrees where genetic inconsistencies are apparent. For such instances, proper modeling of the possibility of a mutation is crucial in order to avoid a likelihood of zero [17]. With the current version of FamLink, it is, however, possible to model genotyping errors whereby the true allele is recorded as a randomly chosen allele with a specified probability. In addition, FamLink allows for unseen alleles using two different methods; *Normalizing*, whereby all frequencies are normalized so the final sum is 1.0 and *Search and Substract*, whereby the new allele frequency is substracted from other alleles not used in the current case.

2.2. Theoretical considerations and validation

In this section we present a result which can be used to study the impact of linkage on the likelihood ratio and also to validate FamLink for a specific case. Consider the hypotheses:

- H_1: “Two individuals are grandparent and grandchild”
- H_2: “Two individuals are unrelated”.

Marker data is available for two markers separated by a recombination distance r. Below we show the following:

$$\frac{LR\,\text{for markers linked at distance}\,r}{LR\,\text{for unlinked markers (i.e.}\,r = 0.5)} = -2r + 2$$

if the individuals do not share any alleles.

The proof for this is based on Eq. (10) in the appendix of [3]:

$$P(data|\,ped.i) = (p_{00} + p_{11} - p_{10} - p_{01})k^{i}_{1,1}(r) + \frac{1}{2}(p_{10} + p_{01}).$$

The formula is more general than currently needed and we next explain the notation in the present context. We will use the formula for *ped. i* corresponding to hypothesis H_1 and then $k^{i}_{11}(r) + (1 - r)/2$ as explained in Chapter 4.5 of [5]. The remaining terms on the right hand side, $p_{uv}, u, v = 0, 1$ depend only on allele frequencies for the markers and can be calculated using Table 1 in [3]. Therefore the required probability is a linear function of r and may thus be written

$$f(r) = p(data|H_1) = \alpha r + \beta$$

where α and β depend only on the allele frequencies for the markers and are given as

$$\alpha = \frac{p_{10} + p_{01} - p_{00} - p_{11}}{2}$$
$$\beta = p_{10} + p_{01} - \frac{p_{00} + p_{11}}{4}$$

Observe next that

$$g(r) = \frac{LR(r)}{LR(0.5)} = \frac{LR\,\text{for markers linked at distance}\,r}{LR\,\text{for unlinked markers}} = \frac{\alpha r + \beta}{(1/2)\alpha + \beta}$$

since the likelihood for the unrelated alternative does not depend on r and therefore cancels in the above expression. By rewriting

$$g(r) = \alpha_1 r + \beta_1, \quad \text{say,}$$

it is apparent that there is a linear effect of linkage, as measured by r, on the quantity of interest, $g(r)$

It remains to show that $\alpha_1 = -2$ and $\beta_1 = 2$.

Assume the individuals do not share any alleles. Then there can be no identical by descent (IBD) sharing (mutations and genotyping errors are disregarded) and so $p_{uv} > 0$ if and only if $u = v = 0$. Then

$$\alpha = \frac{p_{10} + p_{01} - p_{00} - p_{11}}{2} = \frac{p_{00}}{2}$$
$$\beta = \frac{p_{00} + p_{11}}{2} = \frac{p_{00}}{2} = -\alpha$$
$$\alpha_1 = \frac{\alpha}{(1/2)\alpha + \beta} = \frac{\alpha}{(1/2)\alpha - \alpha} = -2$$
$$\beta_1 = \frac{\beta}{(1/2)\alpha + \beta} = \frac{-\alpha}{(1/2)\alpha - \alpha} = -2 \quad \text{and therefore}$$
$$g(r) = -2r + 2$$

Observe that above expression for $g(r)$ is valid regardless of allele frequencies as long as the individuals do not share any alleles.

For instance $g(0.1) = 1.8$. In other words, the true LR accounting for linkage is almost twice the approximation corresponding to the result obtained assuming unlinked markers. Using FamLink, we find 0.45/0.25 = 1.8 (regardless of specified allele frequencies) confirming the theoretical result. For $r = 0.2$, FamLink gives 0.4/0.25 = 1.6 which equals $g(0.2) = 1.6$ as it should.

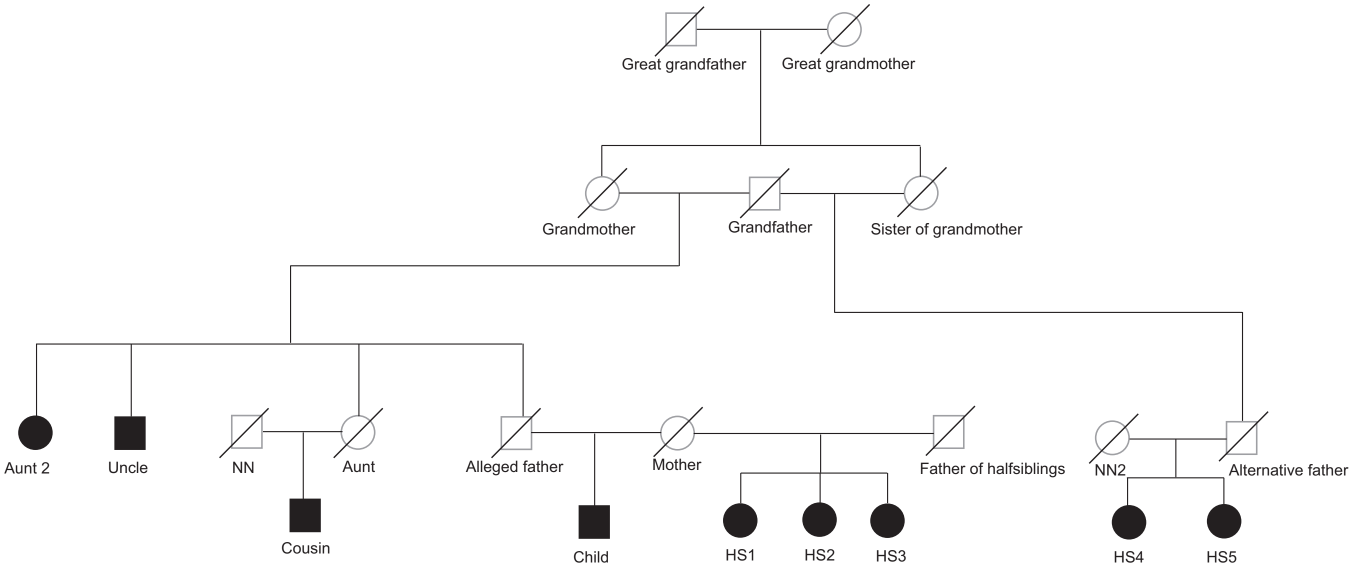

Fig. 1. Case 2 describes a complex pedigree including a marriage loop. Samples were drawn from all individuals with completely filled icons. The question put forward was whether the individual denoted *Alleged father* was the father of the individual denoted *Child*, with the second hypothesis being the individual denoted *Alternative father* as the father.

To further validate FamLink, we have written an R-function (Supp. Data 3) which extends the above examples by also considering halfsibling and avuncular relations in addition to allowing for general genotypes. The numerical results coincide as they should for a number of examples.

3. Examples

3.1. Case 1

In this case, two full siblings wanted to know if a third individual was their paternal half-sibling (Supp. Data 2). All individuals were typed for the 21 STR loci included in the AmpFlSTR® Identifiler® PCR Amplification Kit (Applied Biosystems) and PowerPlex® ESI 17 System kit (Promega Corp.). For this set of loci, there is one pair of closely located loci on chromosome 12, namely D12S391 and vWA [4]. The likelihoods for the two hypotheses "Two full siblings and one half sibling" (H_1) and "Two full siblings and one unrelated" (H_2) were first calculated by means of Familias [11] with Swedish allele frequencies ([18] and K. Montelius, personal communication) assuming the 19 loci (excluding D12S391 and vWA) to be unlinked. For the statistical interpretation of the genotype data from D12S391 and vWA we used FamLink with a recombination frequency of 0.089 [10]. This estimate may be uncertain and other values can be tried. In fact, an important purpose of FamLink is to facilitate sensitivity analyzes studying the impact of assumptions that may be questioned.

The genotypes for D12S391 and vWA can be found in Supp. Data 2. The combined LR for D12S391 and vWA was computed to 3.91 when linkage was assumed and to 1.10 when linkage was ignored. The total LR (all 21 STRs) increased from 20.0 to 71.4 when linkage was accounted for. Thus, the difference between accounting for linkage and not accounting for linkage could in this case be considered to be large, almost a four fold increase. Assuming equal priors, for the two hypotheses, the posterior probability increased from 95% in favor of the first hypothesis to almost 98.6%, when linkage was assumed. Based on this, it is obvious that reaching a threshold in terms of LR or posterior probability may well depend on whether linkage is accounted for. In other words, accounting for linkage may well directly determine how cases are reported.

3.2. Case 2

Fig. 1 depicts a rather complex relationship question, including a marriage loop located between the individuals denoted Grandmother and *Sister of grandmother*. The *Child* wanted to know if the individual denoted *Alleged father* was her biological father with the second hypothesis being that the individual denoted *Alternative father* was the father (Fig. 1). Previous results had already yielded high probabilities in favor of the *Alleged father* (or a close relative of him) being the father, with the second hypothesis including an unrelated man as the father. We used 36 STR loci included in the Investigator HDplex Kit (Qiagen Inc.), PowerPlex ESI 17 (Promega Corp.) and PowerPlex 16HS (Promega Corp) PCR multiplexes as well as an inhouse kit containing three highly polymorphic STR markers (D17S906, APOAI1 and D11S554). For these sets of loci, there are several markers located less than 50 cM from each other [4]. The likelihoods were first calculated in Familias, excluding the marriage loop, i.e. removing the *great grandparents*, and using mutation rates of zero in combination with a Norwegian frequency database for all marker sets. For the complete statistical evaluation we used the QuickAnalysis option of FamLink allowing us to include all markers and account for linkage between every pair of loci, as well as the marriage loop.

The results from Familias, excluding the marriage loop and using mutation rates of zero, yielded a combined LR of 1.3E + 08 in favor of the *Alleged father* being the father (or a close relative other than the individual denoted *Alternative father*). Using the QuickAnalysis option of FamLink, including the marriage loop and recombination between markers, a combined LR of 368577 was computed in favor of the *Alleged father* being the father. In other words, ignoring linkage and the marriage loop the LR is overestimated by factor of roughly 350. If only the marriage loop is disregarded, the LR is overestimated by a factor of 6. This clearly illustrates the importance of properly accounting for linkage and marriage loops.

3.3. Simulation cases

Two case scenarios were investigated further by simulations in order to study the impact of accounting for linkage. In the first example, the hypothesis "three full siblings" (H_1) was compared

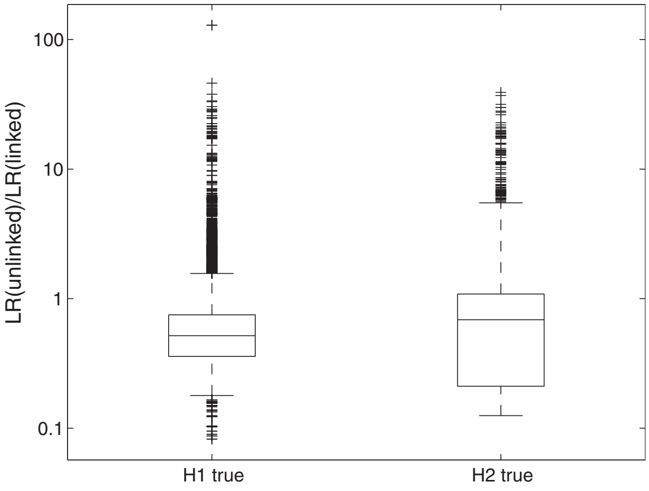

Fig. 2. Box plot of the distribution of the ratio LR(unlinked)/LR(linked) for simulated cases with the hypotheses H_1: Three full siblings, and H_2: Two full siblings and one unrelated. The data are from 10,000 simulated cases assuming H_1 and H_2, respectively, to be true.

with "two full siblings and one unrelated" (H_2) and the second case involved the hypothesis "grandparents" (H_1) versus "unrelated" (H_2) (Supp. Data 2). For both examples, the markers SE33 and D6S1043 were used with Chinese Han population frequencies [19,20], together with a recombination frequency of 0.044 [4]. The simulation was performed twice, first with 10,000 datasets where hypothesis H_1 was assumed to be true, and then 10,000 instances where hypothesis H_2 was assumed to be true.

The simulation of the impact of linkage on LR values for the case "three full sibs" versus "two full siblings and one unrelated" resulted in a general underestimation of the LR if linkage between SE33 and D6S1043 was ignored. The median of the ratio LR(unlinked)/LR(linked) was computed to 0.52 [with a 95% interval, 0.21–5.1] when the simulations were performed assuming H_1 to be true. When simulating assuming the alternative hypothesis to be true, genetic inconsistencies were found in 70% of the simulated cases. For the remaining instances the median ratio was computed to 0.67 [0.19–6.8]. Fig. 2 is a box plot of the spread of the simulated data values, showing a considerable variation of the impact of linkage on LRs for this pair of loci for these hypotheses.

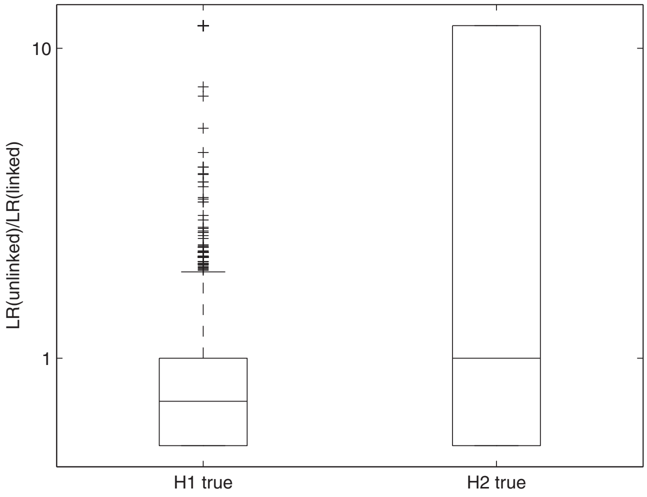

Fig. 3. Box plot of the distribution of the ratio LR(unlinked)/LR(linked) for simulated cases with the hypotheses H_1: Grandparents, and H_2: Unrelated. The data are from 10,000 simulated cases assuming H_1 and H_2, respectively, to be true.

The simulation of the second case example, "grandparents" versus "unrelated", also resulted in a general underestimation of the LR when linkage between SE33 and D6S1043 was ignored. The median of the ratio was calculated to 0.73 [0.52–2.2] when the simulations were performed assuming H_1 to be true. When simulating assuming hypothesis $H_2$to be true, genetic inconsistencies were found in 77% of the simulated cases. For the remaining instances the median ratio was computed to 1.0 [0.52–11.84]. Fig. 3 is a box plot showing the scattering of the simulated data.

4. Discussion

Recently, a number of new PCR multiplexes have been released on the market. These include a number of additional STR markers to be used in forensic relationship testing. Although more DNA markers generate more information and therefore generally makes it easier to report a complex case, it also increases the risk of having multiple markers located closely on the same chromosome. Thus linkage needs to be accounted for in the statistical evaluation when calculating the weight of evidence. Phillips et al. [4] previously studied the genetic distance between 29 syntenic STR loci included in several forensic PCR multiplexes and concluded that 14 pairs of markers were located less then 50 cM apart.

In this paper we address the issue of taking linkage properly into account in relationship testing by presenting a new statistical tool, FamLink, that calculates the likelihood ratio of observing the case specific DNA-data given two (or more) specified hypotheses of possible genetic relationships. We demonstrate, based on two real cases, simulations and a mathematical derivation, that it is important to account for linkage. This applies particularly to pedigrees beyond the standard trios/duos. In such cases, the LR may well be below, say 100,000 and accounting for linkage or at least studying sensitivity with respect to the effect of linkage, should be performed.

With this in mind, we have developed a software, freely available from http://www.FamLink.se. FamLink is a user friendly front-end to an existing framework, Merlin [12]. Merlin is widely used in medical linkage studies and provides a fast and reliable computation algorithm. Our software adopts the advantages of the computation algorithm and provides several features which simplifies the interactivity and provides the possibility to include linked markers in relationship calculations.

Although the main purpose with FamLink is focused on relationship testing, it is also possible to quantify the impact of linkage for match probability calculations including the issue of a close relative being the alternative man [7]. We refer to the manual on a more detailed description of this feature.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2012.01.012.

References

[1] M. Nothnagel, J. Schmidtke, M. Krawczak, Potentials and limits of pairwise kinship analysis using autosomal short tandem repeat loci, Int. J. Legal Med. 124 (2010) 205–215.
[2] A.O. Karlsson, G. Holmlund, T. Egeland, P. Mostad, DNA-testing for immigration cases: the risk of erroneous conclusions, Forensic Sci. Int. 172 (2007) 144–149.
[3] T. Egeland, N. Sheehan, On identification problems requiring linked autosomal markers, Forensic Sci. Int. Genet. 2 (2008) 219–225.
[4] C. Phillips, D. Ballard, P. Gill, D.S. Court, A. Carracedo, M.V. Lareu, The recombination landscape around forensic STRs: accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data, Forensic Sci. Int. Genet., doi:10.1016/j.fsigen.2011.07.012, in press.

[5] E.A. Thompson, Statistical inference from genetic data on pedigrees, in: NSF-CBMS Regional Conference Series in Probability and Statistics. vol. 6, IMS, Beachwood, OH, 2000.

[6] P. Gill, C. Phillips, C. McGovern, J.A. Bright, J. Buckleton, An evaluation of potential allelic association between the STRs vWA and D12S391: implication in criminal casework and applications to short pedigrees, Forensic Sci. Int. Genet., doi:10.1016/j.fsigen.2011.11.001, in press.

[7] J. Buckleton, C. Triggs, The effect of linkage on the calculation of DNA match probabilities for siblings and half siblings, Forensic Sci. Int. 160 (2006) 193–199.

[8] K.L. O'Connor, C.R. Hill, P.M. Vallone, J.M. Butler, Linkage disequilibrium analysis of D12S391 and vWA in U. S. population and paternity samples, Forensic Sci. Int. Genet. 5 (2011) 538–540.

[9] B. Budowle, J. Ge, R. Chakraborty, A.J. Eisenberg, R. Green, J. Mulero, R. Lagace, L. Hennessy, Population genetic analyses of the NGM STR loci, Int. J. Legal Med. 125 (2011) 101–109.

[10] K.L. O'Connor, A.O. Tillmar, Effect of linkage between vWA and D12S391 in kinship analysis, Forensic Sci. Int. Genet., under review.

[11] T. Egeland, P.F. Mostad, B. Mevag, M. Stenersen, Beyond traditional paternity and identification cases. Selecting the most probable pedigree, Forensic Sci. Int. 110 (2000) 47–59.

[12] G.R. Abecasis, S.S. Cherny, W.O. Cookson, L.R. Cardon, Merlin – rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (2002) 97–101.

[13] E.S. Lander, P. Green, Construction of multilocus genetic linkage maps in humans, Proc. Natl. Acad. Sci. U.S.A. 84 (1987) 2363–2367.

[14] E.M. Wijsman, J.H. Rothstein, E.A. Thompson, Multipoint linkage analysis with many multiallelic or dense diallelic markers: markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees, Am. J. Hum. Genet. 79 (2006) 846–858.

[15] 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, P.M. Schneider, N. Morling, ISFG: recommendations on biostatistics in paternity testing, Forensic Sci. Int. Genet. 1 (2007) 223–231.

[16] K.L. Ayres, Relatedness testing in subdivided populations, Forensic Sci. Int. 114 (2000) 107–115.

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

[18] K. Montelius, A.O. Karlsson, G. Holmlund, STR data for the AmpFlSTR Identifiler loci from Swedish population in comparison to European, as well as with non-European population, Forensic Sci. Int. Genet. 2 (2008) e49–e52.

[19] S. Huang, Y. Zhu, X. Shen, X. Le, H. Yan, Genetic variation analysis of 15 autosomal STR loci of AmpFlSTR sinofiler PCR amplification kit in Henan (central China) Han population, Leg. Med. (Tokyo) 12 (2010) 160–161.

[20] C. Liu, N. Harashima, Y. Katsuyama, M. Ota, A. Arakura, H. Fukushima, ACTBP2 gene frequency distribution and sequencing of the allelic ladder and variants in the Japanese and Chinese populations, Int. J. Legal Med. 110 (1997) 208–212.

Paper IV

Int J Legal Med
DOI 10.1007/s00414-014-1117-7

ORIGINAL ARTICLE

A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations

Daniel Kling · Andreas Tillmar · Thore Egeland · Petter Mostad

Received: 31 July 2014 / Accepted: 6 November 2014

Abstract Several applications necessitate an unbiased determination of relatedness, be it in linkage or association studies or in a forensic setting. An appropriate model to compute the joint probability of some genetic data for a set of persons given some hypothesis about the pedigree structure is then required. The increasing number of markers available through high-density SNP microarray typing and NGS technologies intensifies the demand, where using a large number of markers may lead to biased results due to strong dependencies between closely located loci, both within pedigrees (linkage) and in the population (allelic association or linkage disequilibrium (LD)). We present a new general model, based on a Markov chain for inheritance patterns and another Markov chain for founder allele patterns, the latter allowing us to account for LD. We also demonstrate a specific implementation for X chromosomal markers that allows for computation of likelihoods based on hypotheses of alleged relationships and genetic marker data. The algorithm can simultaneously account for linkage, LD, and mutations. We demonstrate its feasibility using simulated examples. The algorithm is implemented in the software FamLinkX, providing a user-friendly GUI for Windows systems (FamLinkX, as well as further usage instructions, is freely available at www.famlink.se). Our software provides the necessary means to solve cases where no previous implementation exists. In addition, the software has the possibility to perform simulations in order to further study the impact of linkage and LD on computed likelihoods for an arbitrary set of markers.

Keywords FamLinkX · Lander-Green · Likelihood computations · X chromosome · Markov-chain · Linkage disequilibrium · Linkage · Mutation

Introduction

In several applications, there is a need to determine the most probable pedigree structure given some genetic marker data for a set of persons; e.g., to get an unbiased result in a medical genetic study, for example linkage or association studies, or in a forensic setting, for example paternity

Electronic supplementary material The online version of this article (doi:10.1007/s00414-014-1117-7) contains supplementary material, which is available to authorized users.

D. Kling (✉)
Department of Family Genetics,
Norwegian Institute of Public Health, P.O. Box 4040 Nydalen,
0403, Oslo, Norway
e-mail: daniel.kling@fhi.no

D. Kling · T. Egeland · P. Mostad
Department for Chemistry, Biotechnology and Food Science,
Norwegian University of Life Sciences, Ås, Norway

A. Tillmar
Department of Forensic Genetics and Forensic Toxicology,
National Board of Forensic Medicine, Linkoping, Sweden

A. Tillmar
Department of Clinical and Experimental Medicine,
Faculty of Health Sciences, Linkoping University,
Linkoping, Sweden

T. Egeland
Department of Forensic Biology,
Norwegian Institute of Public Health, Oslo, Norway

P. Mostad
Mathematical Sciences,
Chalmers University of Technology and Mathematical Sciences,
University of Gothenburg, Gothenburg, Sweden

cases, immigration cases or in identification of victims following a disaster. Medical genetics is often concerned with thousands of markers from high-density SNP arrays and concepts such as linkage between neighboring markers is inescapable. Accounting for the possibility of a crossover in statistical computations is crucial to obtain an unbiased result. In forensic genetics, where short tandem repeats (STR) is the primary data, the use of linked markers is not as widely accepted, though their utility has been studied. Using linked markers, several symmetrical pedigrees can be distinguished due to different inheritance patterns, [7, 12, 13, 22, 27]. [8] presented an algorithm where likelihoods could be calculated based on genetic marker data and hypothesized pedigree structures. The algorithm is, in effect, a peeling algorithm calculating conditional probabilities based on cutsets, with cutsets being a set of persons separating the pedigree into independent parts. The algorithm proceeds by updating conditional probabilities as each cutset is peeled. In practice, the algorithm is feasible for complex/extended pedigrees with many individuals. Indeed, the algorithm is useful for many unlinked markers, while for linked markers, the number of iterations can grow exponentially with the number of markers, as we have to consider all possible founder haplotypes. To model many linked markers, [18] proposed another algorithm using identity-by-descent (IBD) instead. Given a pedigree, we can calculate the probability that two alleles are IBD, meaning that they originate from the same ancestral allele, see [28] for an introduction. The algorithm generalizes the concept by looking at inheritance graphs, see [2]. Given a specific pedigree, we can define the meioses, and all possible combination of meiotic outcomes determines the inheritance space. The algorithm continues by setting up each locus as a node in a hidden Markov-chain where the meiotic outcome of the next locus only depends on the previous one. This is, in reality, an approximation, since we disregard any interference caused by a crossover, but is often a good enough approach. In the absence of optimizations, the complexity grows exponentially with the number of meioses, i.e., approximately with the number of individuals.

We further consider linkage disequilibrium (LD), or allelic association, which is the non-random association of alleles at different loci. For the purpose of association studies, LD can extend across chromosomes, but for the current setting, we focus on LD between closely linked markers on the same chromosome. The degree of LD is often more apparent for STR markers (as compared to SNP markers) where more alleles at each marker yield more possible haplotypes for a set of markers; e.g., for two SNP markers we have maximally four haplotypes (given biallelic SNP:s) while for two STR markers with 10 alleles each, we have 100 possible haplotype configurations. LD is a source of bias in the statistical calculations since allele frequencies for alleles at two closely located loci may no longer be regarded as independent. As demonstrated by previous papers, LD may cause erroneous conclusions when not accounted for, see [3, 10]. [17] presented an extension of the Lander-Green algorithm modeling founder allele patterns with Markov chains. Moreover, a paper by [1] also addresses the issue by demonstrating how Merlin handles clusters of markers in LD. Both of these approaches lack the possibility to model mutations. In addition, the implementation in Merlin does not allow for recombination between markers within a cluster. As will be demonstrated, accounting for both the mentioned parameters is crucial in a forensic setting, but also often in a more general setting when establishing relatedness.

The use of the Lander-Green algorithm has gained wide acceptance in the medical genetic field, while the forensic community has been more hesitant to implement it. The latter can be explained by no current model of how to handle mutations, though the software Merlin includes the possibility to model genotyping errors, which might be sufficient in an extended pedigree with untyped founders and SNP markers. For STR markers and a forensic context where genetic inconsistencies are frequently observed, we need a better way to handle mutations. Mutation rates above 0.005 (per marker and generation) have been observed for some highly polymorphic markers and in addition, female mutation rates may vary from male mutation rates [4, 5]. Several models exist to handle transitions of alleles within a pedigree, see e.g., [6]. The most commonly accepted approach is the step-wise model where the probability of a mutation decreases with the number of transition steps. Each step here is defined as the difference between the two alleles in tandem repeats. Again, this is mostly relevant for STR markers with a distinct number of tandem repeats for each allele.

In forensic genetics, there is a growing focus on STR markers located on the X-chromosome. Previous studies demonstrate their utility in several relatedness settings, especially some cases where autosomal markers are not able to distinguish between the different alternative hypotheses [14, 20, 21, 25, 27]. Consider, for example, the hypothetical case where two sisters want to know if they share the same father or the same mother, see Fig. 1. In practicality, autosomal data cannot distinguish these two hypotheses, whereas X- chromosomal markers obviously display different inheritance patterns. Two sisters with the same father share at least one allele IBD for each X- chromosomal marker, disregarding mutations, whereas two sisters sharing only their mother, do not have the same obligate allele sharing. The point is that X- chromosomal data, be it SNP-data or STR-data, can provide crucial information in several situation where autosomal markers do not prevail. Several studies have shown that the clusters of STR markers

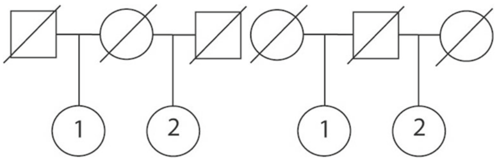

Fig. 1 Half siblings. X- chromosomal markers are useful to distinguish between the maternal (*left*) and paternal (*right*) pedigree

included in the Argus X12 kit display a high degree of LD, but also that recombination can be observed within a cluster [19, 24, 27]. In addition to constructed SNP data, we will use population data from the X12 kit to prove the necessity of our approach. This paper describes a general model to simultaneously handle linkage, linkage disequilibrium, and mutations in calculation of genetic relatedness. We demonstrate an implementation for X- chromosomal data and perform a feasibility study using some common pedigrees encountered in case work, demonstrating the utility of our implementation. In addition, we present a graphical user interface to assist in the interpretation of the statistical results.

Methods

Consider some hypotheses of postulated relationships between a set of individuals and some genetic marker data. We present a joint probability model accounting for linkage, linkage disequilibrium, and mutations. We further demonstrate a specific implementation for markers located on the X chromosome, though the model is not constrained to such data. The algorithm is similar to the algorithm of [18], but provides extensions for allelic dependencies across different loci, similar to [17]. The algorithm also implements a transition model to account for mutations, which may be different for male and female transmissions.

In mathematical notation, we want to compute the probability of observing marker data given a pedigree, a specified model, and values for the parameters in that model:

$$\Pr(d, s \mid p, r, m) \tag{1}$$

Here, the marker data is split into data s for pedigree founders and d for non-founders, and p, r, and m indicate the values of parameters for the haplotype population frequencies, the recombination frequencies, and mutation probabilities, respectively. For precise information about the notation see the Appendix.

To facilitate the specification of the model, we introduce a variable v specifying the inheritance pattern in the pedigree (whether paternal or maternal alleles are inherited at each locus) and variables g and f specifying the genotypes of typed and untyped founders, respectively. Our model specifies the probability

$$\Pr(d, s, v, f, g \mid p, r, m)$$

and our algorithm represents an efficient way to compute from this the probability in Eq. 1.

Model

We assume we have data from I different loci. In applications, it may be that we have clusters of loci where we have to take into account both LD and linkage, while we only have to take into account linkage between clusters. In its simplest form, the model uses a one-step Markov chain for the founder haplotypes, so that we assume the allele at locus i is independent of alleles at loci with indexes lower than $i-1$ given the allele at locus $i-1$. More generally, when the one-step Markov chain model is not supported by data, we use an L step Markov chain, where the allele at locus i is independent of alleles at loci with indexes lower than $i-L$ given the alleles at loci $i-L, \dots, i-1$. In all cases, we assume independence between the haplotypes of different clusters.

With this Markov assumption, we can write

$$\Pr(d, s, v, f, g \mid p, r, m) = \Pr(d_1, s_1, v_1, f_1, g_1 \mid p_1, m_1) \tag{2}$$
$$\cdot \prod_{i=2}^{I} \Pr(d_i, s_i, v_i, f_i, g_i \mid v_{i-1}, f_{i-L}, \dots, f_{i-1},$$
$$g_{i-L}, \dots, g_{i-1}, p_i, r_i, m_i).$$

Let us start with specifying the first term

$$\begin{aligned} &\Pr(d_1, s_1, v_1, f_1, g_1 \mid p_1, m_1) \\ &= \Pr(d_1 \mid v_1, f_1, g_1, m_1)\Pr(s_1 \mid g_1)\Pr(f_1 \mid p_1)\Pr(g_1 \mid p_1)\Pr(v_1), \end{aligned} \tag{3}$$

where the most involved specification is for the factor $\Pr(d_1 \mid v_1, f_1, g_1, m_1)$. Each value of v_1 specifies a sequence of transmissions from the founder alleles, whose identities are given by f_1 and g_1, to the alleles observed in d_1. This sequence of transmissions may contain branching. We get that

$$\Pr(d_1 \mid v_1, f_1, g_1, m_1) = \sum_{t_1, t_2, \dots, t_T} m_1(t_1)m_1(t_2)\dots m_1(t_T)$$

where the sum goes over all possible sequences $t_1, t_2, \dots, t_T$ of transmissions from the founder alleles to the observed alleles, and where the function m_1 indicates the probability for each transition. For the small pedigrees we consider, and in particular as we are focusing on X-chromosomal data, this sum is less formidable to handle than it might be in general. Also, simplifications can be employed where extremely unlikely combinations of multiple mutations within a single sequence of transmissions are ignored. Note that the relevant sums may be pre-computed for the transmission trees that may occur in our pedigree.

The remaining factors of Eq. 3 are easily specified: $\Pr(s_1 \mid g_1)$ is 1 if each s_{1j} is a permutation of g_{1j}, otherwise zero. The factor $\Pr(f_1 \mid p_1)$ is a product of components of p_1, as is $\Pr(g_1 \mid p_1)$. (Note that we assume Hardy-Weinberg equilibrium (HWE) in the model). Finally, $\Pr(v_1)$ should specify equal probability 2^{-N} for each of the 2^N possible values for v_1.

We continue specifying the remaining factors of Eq. 2 with

$$\begin{aligned}&\Pr(d_i, s_i, v_i, f_i, g_i \mid v_{i-1}, f_{i-L}, \ldots, f_{i-1},\\ &g_{i-L}, \ldots, g_{i-1}, p_i, r_i, m_i)\\ =\ &\Pr(d_i \mid v_i, f_i, g_i, m_i)\Pr(s_i \mid g_i)\Pr(v_i \mid v_{i-1}, r_i)\\ &\Pr(f_i \mid f_{i-L}, \ldots, f_{i-1}, p_i)\Pr(g_i \mid g_{i-L}, \ldots, g_{i-1}, p_i),\end{aligned}$$

where the factors $\Pr(d_i \mid v_i, f_i, g_i, m_i)$ and $\Pr(s_i \mid g_i)$ are specified as for $i = 1$. Each of the factors $\Pr(f_i \mid f_{i-L}, \ldots, g_{i-1}, p_i)$ and $\Pr(g_i \mid g_{i-L}, \ldots, g_{i-1}, p_i)$ are products over the components of p_i, with one factor for each haplotype. Finally, $\Pr(v_i \mid v_{i-1}, r_i) = r_i^a(1 - r_i)^b$, where a and b are the number of haplotypes where v_i and v_{i-1} indicate a recombination, respectively no recombination, between locus i and $i - 1$.

Algorithm

Our goal is to compute the probability of the data d and s, given the pedigree and the parameters p, r, and m. As mentioned, our algorithm is a version of the Lander-Green algorithm, which again is a version of the forward-backward algorithm for hidden Markov Models. In fact, for our purposes, we only need the forward part of the algorithm. Specifically, for $i = 2, \ldots, I$ and for all possible values of v_i, $f_{i-L+1}, \ldots, f_i$, and $g_{i-L+1}, \ldots, g_i$, we have the recursive formula

$$\begin{aligned}&\Pr(d_1, \ldots, d_i, s_1, \ldots, s_i, v_i, f_{i-L+1}, \ldots, f_i, g_{i-L+1}, \ldots, g_i) \quad (4)\\ =\ &\Pr(d_i \mid v_i, f_i, g_i)\Pr(s_i \mid g_i)\\ &\sum_{v_{i-1}}\sum_{f_{i-L}}\sum_{g_{i-L}}\big[\Pr(d_1, \ldots, d_{i-1}, s_1, \ldots, s_{i-1}, v_{i-1},\\ &f_{i-L}, \ldots, f_{i-1}, g_{i-L}, \ldots, g_{i-1})\Pr(f_i \mid f_{i-L}, \ldots, f_{i-1})\\ &\Pr(g_i \mid g_{i-L}, \ldots, g_{i-1})\Pr(v_i \mid v_{i-1})\big]\end{aligned}$$

where we have omitted the conditioning on the parameters p, r, and m for brevity. Our algorithm starts with computing and listing $\Pr(d_1, s_1, v_1, f_1, g_1)$ according to Eq. 3 for all values of v_1, f_1, and g_1 that makes it non-zero. Then Eq. 4 is used to compute for $i = 2, \ldots, I$ the left-hand side for all combinations of v_i, $f_{i-L+1}, \ldots, f_i$, and $g_{i-L+1}, \ldots, g_i$ making it non-zero. The final result is obtained by summing

$$\begin{aligned}&\Pr(d, s) = \sum \Pr(d_1, \ldots, d_I, s_1, \ldots, s_I, v_I,\\ &f_{I-L+1}, \ldots, f_I, g_{I-L+1}, \ldots, g_I)\end{aligned}$$

where the sum is taken over all v_I, $f_{I-L+1}, \ldots, f_I$, and $g_{I-L+1}, \ldots, g_I$. Clearly, the applicability of the algorithm depends on the number of combinations of v_i, $f_{i-L+1}, \ldots, f_i$, and $g_{i-L+1}, \ldots, g_i$ making Eq. 4 non-zero (or at least of non-negligible relative size). Note that each of $g_{i-L+1}, \ldots, g_i$ need to have components that are permutations of the corresponding data $s_{i-L+1}, \ldots, s_i$ to get a non-zero probability. However, as we are including mutations, many of the various possible values for f need to be considered for the calculations.

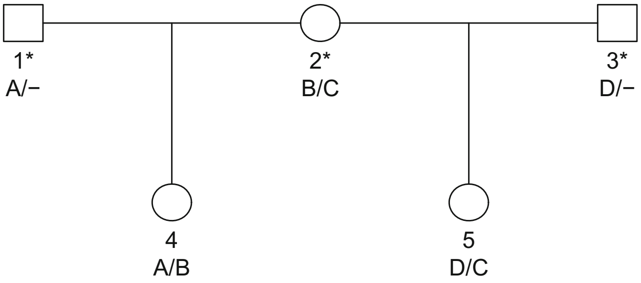

Fig. 2 Figure used to explain the notation of the likelihood calculation

Example

Let us consider an example with X- chromosomal data and a pedigree with two girls sharing the same mother but having different fathers. Initially, consider only one marker denoted i. The notation is explained with reference to Fig. 2. The founders 1, 2, and 3 are marked with '*'. Possible genotypes are indicated with the paternal allele, i.e., the one inherited from the father given to the left so that for instance individual 2 has inherited 'B' from her father and 'C' from her mother, and has passed on 'B' to her daughter 4 and 'C' to her daughter 5.

We now assume only the two girls are typed, with the data indicated in the figure. Using the notation of the Appendix, $S = 0$ and $M = 2$, and $d_i = ((A, B), (C, D))$. There are two mother-child relationships so the vector v_i has length 2, and four possible values. The value indicated in the figure is $v_i = (0, 1)$. There are no typed founders, so there is no g in the example. There are $F = 4$ founder alleles in the pedigree: the allele each girl inherits from her untyped father, and the two alleles of the untyped mother. The value of f_i indicated in the figure is $f_i = (A, B, C, D)$.

The algorithm starts by computing a table with

$$\Pr(d_1, v_1, f_1) = \Pr(d_1 \mid v_1, f_1)\Pr(v_1)\Pr(f_1)$$

for all combinations of the values of v_1 and the values of f_1 making the probability non-zero: Note that there are A_1^2 possible (phased) genotypes for the mother (remember that we include the possibility of mutations) and, given the data, two possible values for each of the two founder alleles coming from the fathers. Thus, there are, at most, $4A_1^2$ values for f_1 making the probability above non-zero, and fewer if one or both of the girls have been observed to be homozygote at locus 1. With four possible values for v_i, the table has a size at most of $16A_1^2$.

In the recursive step, the table for $\Pr(d_1, \ldots, d_i, v_i, f_{i-L+1}, \ldots, f_i)$ given the table for $\Pr(d_1, \ldots, d_{i-1}, v_{i-1}, f_{i-L}, \ldots, f_{i-1})$ is computed as follows: The value $\Pr(d_i \mid v_i, f_i)$ is computed as above for the at most $16A_i^2$ values making it non-zero, and the small table for $\Pr(v_i \mid v_{i-1})$ is computed. Then, as the values of $\Pr(f_i \mid f_{i-L}, \ldots, f_{i-1})$ are stored directly in p_i, the entries of the table for $\Pr(d_1, \ldots, d_i, v_i, f_{i-L+1}, \ldots, f_i)$, which is of length at most $4^{L+1}A_{i-L+1}^2 \ldots A_i^2$, can be computed by summation using Eq. 4.

Estimation of haplotype frequencies

Our model requires as parameters haplotype frequencies, or rather, for each $i = 1, \ldots, I$ and $j = 1, \ldots, A_i$ the probability of observing an allele of type j at locus i conditional on the alleles at loci $i - L, \ldots, i - 1$. In our implementation, we require as input a list of observed haplotypes for each cluster of loci. Estimates of haplotype frequencies based directly on haplotype counts would assign zero probability to all unobserved haplotypes, which would generate unreasonable results. Instead, and similar to [27], we use haplotype probabilities that are weighted averages between observed haplotype frequencies and haplotype frequencies estimated under the assumption of marker independence. If the count of haplotype k is c_k, we use as the haplotype probability $h_k = (c_k + \lambda x_k)/(\sum c_k + \lambda)$, where we sum over all possible haplotypes, where x_k is the product of the frequencies of the alleles of the haplotype, and λ is a positive number. We use allele frequencies based on the haplotype input, but one could also use allele frequencies from another source. Then, using a Dirichlet distribution with parameters λx_k as a prior for haplotype probabilities, h_k is the expected posterior haplotype probability after updating the prior with haplotype observations. In this paper, we use $\lambda = 1$, but one could explore more optimal ways of setting this parameter. We do not discuss how to estimate haplotype frequencies when the phase of the genetic data is unknown, since with X- chromosomal marker we can use data from males where the gametic phase is always known. For data with unknown gametic phase, frequencies can be estimated for example with the EM algorithm [23].

Models for mutations

In our setting, mutations are defined as the possibility of observing a transition from one allele to another within a pedigree. We specify a matrix M_i, containing the transition probabilities. M_i has a dimension of $A_i \times A_i$ where A_i is the number of alleles at locus i and where each row in M must sum to 1.0, such that the row indicates the initial allele and the column the resulting allele. See [6] for a summary of different models. Accounting for mutations is mostly relevant when dealing with STR marker data, where genetic inconsistencies are frequently observed. In such data, contrary to SNP marker data, alleles are determined as a distinct number of tandem repeats. For the well-accepted stepwise model, which we implement a version of, the probability of observing a transition between two alleles decreases with a specified factor with the difference in the number of repeats, such that, say, a single-step mutation can be relatively common while a four-step mutation is very improbable.

Feasibility, simplifications, approximations, and implementation

The feasibility of the algorithm depends mostly on the length of the list of probabilities that needs to be stored in each recursive step. Note that each of $g_{i-L+1}, \ldots, g_i$ need to have components that are permutations of the corresponding data $s_{i-L+1}, \ldots, s_i$ to get a non-zero probability. This means that the factor in the size of the list stemming from the variation in g becomes at most 2^{SL}. The factor coming from variation in f is at most $(A_{i-L+1} \cdots A_i)^{F-Q}2^Q$, where Q is the number of founding alleles that occur in typed persons. Finally, variation in v contributes the factor 2^N, which, for our small pedigrees, will be a small number. Clearly, the length of the list will grow very quickly with L. Unfortunately, LD seems to be a complex phenomenon which is often not well modeled with 1- or 2-step Markov chains. However, when markers are grouped into fairly small clusters, the size of these clusters limit L. When L is small and the pedigree is comparably large, a useful simplification may be based on the fact that $\Pr(d_1, \ldots, d_i, s_1, \ldots, s_i, v_i, f_{i-L+1}, \ldots, f_i, g_{i-L+1}, \ldots, g_i)$ is invariant under permutations of maternal and paternal haplotypes of founders. In other cases, one may use approximations based on excluding from the list combinations $f_{i+L+1}, \ldots, f_i$ which have very low relative probabilities. For example, many combinations may be incompatible with observed data except when assuming that several mutations have taken place, and they will thus have quite low probabilities.

The algorithm is implemented in a Windows-compatible graphical user interface FamLinkX, available at http://www.famlink.se, for a number of predefined relationships relevant to X- chromosomal marker data. In addition, the software can simulate data based on the same predefined relationships, accounting for linkage, LD, and mutations, as modeled in this paper.

Results

To demonstrate the utility of our model, we provide examples highlighting the effect of linkage/LD as well as cases where mutations are present. We compare the results from our software with results using previously published methods. All calculations were performed using three different models, M1: only linkage is accounted for and the markers are considered to be in LE [2]; M2: linkage and LD are accounted for, however no recombination is allowed for within a cluster [1]; and M3: linkage, LD and mutations are modeled as well as recombinations within a cluster are accounted for. Whenever posterior probabilities are presented, flat priors are assumed.

Simulations

We have simulated test data with an algorithm accounting for linkage, LD, and mutations (available in our software). Other simulation algorithms have similar functionality but none contains the full complexity required to test our model. The algorithm, which is similar to the "gene dropping" method, starts by defining and sampling all founder alleles, consistent with the haplotype frequencies and LD structure. Within clusters of markers, haplotype frequencies are estimated for each possible haplotype using the model described in the 'Models for mutations' section. Two haplotypes are sampled for each female founder and one haplotype for each male founder. The algorithm continues by simulating transitions from the founder alleles to all non-founders using a mutation matrix where each transition is assigned a specific probability. In addition, recombination is considered when two neighboring markers are simulated, i.e., the algorithm keeps track of the maternal/paternal chromosomes. The last step is iterated until all non-founders has been simulated. Calculations are then performed using only observed genetic data from the typed persons.

Case 1 - Demonstrating the effect of mutations and recombinations

The first example contains simulated data from the Argus X12 multiplex provided by QIAGEN. The kit divides 12 STR markers located on the X chromosome into four distinct clusters, each containing three markers. Since we are dealing with STR markers, we implement a stepwise mutation model for transmissions within the pedigree. We use haplotype frequency estimates from [26] and recombination rates from [19]. In total, 1000 simulations were performed for each relationship. For reasons that will be apparent in the M1/M2/M3 comparison below, we use an example of three full sisters versus two full sisters and a maternal half-sister. We specify,

H_1: Three full sisters
H_2: Two full sisters and one maternal half sister

We simulate data given by both hypotheses and calculate the LR using three different methods as previously described. Obviously, using the 12 STR markers we obtain sufficiently high LR values in most cases to distinguish between the two hypotheses, given that H_1 is true, see Table 1. We observe some extremes LR:s below 1, but 97.5 % of the simulations are above 28, which given equal prior probabilities yield a posterior probability above 95 %. We notice however that simulating the opposite hypothesis H_2, where we are likely to observe genetic inconsistencies when calculating likelihoods for H_1, we do observe some LRs above 1, though none above 20. It is further interesting to note that mutations and obligate recombinations within a cluster may be observed. Using highly polymorphic STR loci will also more likely result in an actual observable recombination/mutation, whereas for SNP data many of the mutations/recombinations may be hidden. In our given example, recombinations within a cluster were observed in 23 % of the cases and one or more mutations are observed in 4.3 % of the cases when H_1 was simulated. The most interesting comparison with M1 and M2 is for data simulated under H_1, where only M3 calculated non-zero likelihoods in all the 1000 cases. Even in the 957 cases where M1 provided a non-zero LR, there is a substantial bias and variation compared to M3, see Table 2, which also demonstrates that in the 723 cases where M2 provided a non-zero LR, there is less bias and variation, as expected. Figure 3 illustrates the quotients of LR:s for the cases where these can be computed. Note that the distribution of LR:s for method M2 has not been included in Fig. 3 as these values, where computable, fits closely the line $y = x$.

Case 2 - Exploring the influence of linkage disequilibrium

In the second example, we explore how linkage disequilibrium may influence the results. Consider the fairly distant relationship of two maternal cousins, see Supplementary Figure S1 depicting two individuals being maternal cousins, where LD may have a greater impact on the results. Moreover, for SNP marker even two unrelated individuals are likely to share some alleles IBS and it is interesting to see

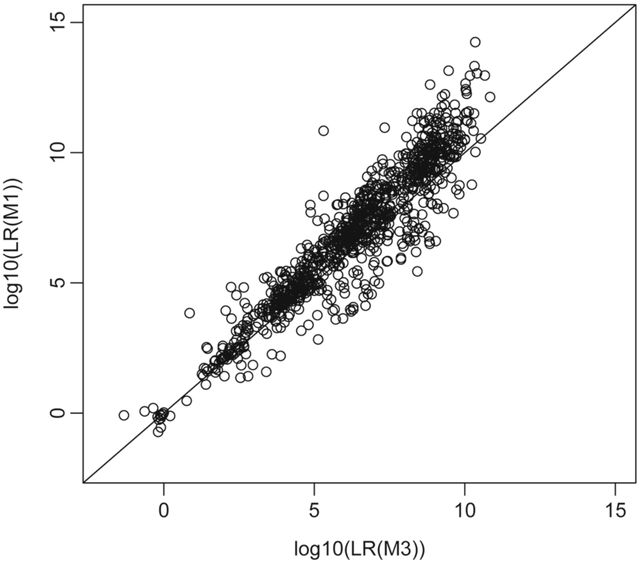

Fig. 3 Scatter plot based on 1000 simulations in Case 1. Three full sisters have been simulated using the model described in the text, alternative hypothesis is two full sisters and one maternal half sister. The plot displays LR $= P(\text{Data} \mid H_0)/P(\text{Data} \mid H_1)$, as calculated using the two models M1 and M3. If all the *dots* are projected on the *x*-axis, we obtain the distribution of LR:s calculated using M3, while if we project all *dots* on the *y*-axis we obtain the distribution of LR:s calculated using M1. If, for a specific dot, $y = x$, we have concordance between the models. (The plot actually contains data from 957 of the 1000 simulation, discarding all simulations where a mutation were observed

how this effects the IBD calculations when LD is present in the data. We specify,

H_1: Two females are maternal cousins, i.e. their mothers are full sisters.
H_2: The two females are unrelated.

The simulations are based on 100 SNP clusters, where each cluster consists of two tightly linked triallelic SNP markers. The genetic distance between the markers within the cluster is specified as 0.1 cm, while the distance between the clusters is 0.8 cm. We specify LD as illustrated in Table 3. Given the presence of LD, we expect that M1 compared to M3 and M2 will show clear bias, as the difference between these two methods relevant here is that M1 does not include LD. When we compare M2 and M3, there is very little difference, as the data was simulated with very low mutation rates and recombination rates within the clusters are not observable for either of the hypotheses. See Table 4 for a summary of the comparisons. As is evident from Figs. 4 and 5, there is an overestimation of the evidence using M1, both when the relation is simulated as true as well as when unrelated is simulated as true. In fact, as depicted in Fig. 5 using model M1 when H_2 is true yields a fairy high false positive rate of almost 10 %, while the same rate when using M3 is only 1.

Table 1 Distribution of quotients of likelihood ratios (LR:s) from 1000 simulations, using the model described in this paper (M3)

	LR (H_0)	LR (H_1)
Median	3.38E+6	0
95 % cred.	[28, 7.23E+9]	[0, 2.29E-6]
Max/min	6.939E+10 / 0.0475	17.46 / 0

We define LR $= P(\text{Data} \mid H_0)/P(\text{Data} \mid H_1)$, where H_0: three full sisters, H_1: two full sisters and one half sister

Parentheses H_0/H_1 indicate which hypothesis has been simulated. Simulating H_1, we are likely to observe genetic inconsistencies and even though we account for mutations in our model, the accumulated number of mutation needed to explain some of the data may result in an LR approximated by zero

Table 2 Distribution of quotients of likelihood ratios (LR:s) from 1000 simulations, using different methods (M1, M2 and M3) We define LR $= P(\text{Data} \mid H_0)/P(\text{Data} \mid H_1)$, where H_0: Three full sisters, H_1:Two full sisters and one half sister

	LR(M2)/LR(M3)	LR(M1)/LR(M3)
Median	1.016	4.26
95 % cred.	[0.0776, 2.4275]	[0.0121, 307]
Max/min	5.51 / 0.001	3.42E+5 / 0.00105

We are only interested in the case where H_0 is true, since simulating H_1 will likely yield genetic inconsistencies. Although our model accounts for mutation, none of the compared models do. The actual number of simulation included in the results are for LR(M2)/LR(M3) 723 and for LR(M1)/LR(M3) 957, removing cases with mutations and recombinations within a cluster

Case 3 - Further exploration of linkage disequilibrium and recombinations

In the third case, we will again use data from clusters of SNP:s. We specify,

H_1: Two females are full siblings, with data available for the mother
H_2: The two females are maternal half siblings, with data available for the mother.

Table 3 Specification of haplotype observations for two triallelic SNP markers

	A1	A2	A3	Total
B1	10	210	80	300
B2	80	35	5	120
B3	10	55	515	580
Total	100	300	600	1000

A1, A2, and A3 are the alleles for the first SNP while B1, B2, and B3 are the alleles for the second SNP

The total number of observations equals 1000

Table 4 Distribution of likelihood ratios (LR:s) from 1000 simulations, using different methods (M1, M2, and M3) We define LR $= Pr(\text{Data} \mid H_0)/Pr(\text{Data} \mid H_1)$, where H_0: Maternal cousins, H_1:Unrelated

	LR(M2)/LR(M3)		LR(M1)/LR(M3)	
	H_0	H_1	H_0	H_1
Median	1.006	1.007	144945	1.26
95 % cred.	[0.223, 1.494]	[0.84, 1.128]	[0.9, 4.9E+16]	[0.954, 14.2]
Max/min	2/0.015	1.475/0.194	3.2E+23/0.014	427/0.491

Subheading H_0 indicates H_0 is simulated as true, while H_1 indicates H_1 is simulated as true

We compare M1/M2/M3 with data from 10 clusters, where each cluster contain three markers which is more widely spaced, 0.2 cm apart, so that crossovers within clusters are likely. However, reasonably, we have lower levels of LD in this situation. For each cluster, we specify three SNP:s with alleles $[A, a]$, $[B, b]$, and $[C, c]$. We further define frequencies, $p(A) = 0.4$, $p(a) = 0.6$, $p(B|A) = 0.833$, $p(B|a) = 0.25$, $p(b|A) = 0.167$, $p(b|a) = 0.75$, $p(C|A, B) = 0.6$, $p(C|A, b) = 0.5$, $p(C|a, B) = 0.5$, $p(C|a, b) = 0.667$, $p(c|A, B) = 0.4$, $p(c|A, b) = 0.5$, $p(c|a, B) = 0.5$, and $p(c|a, b) = 0.333$. Table 5 reflects the power for this kind of data in the given case, where likelihood ratios are generally quite high even when H_1 is true. When H_2 is true, the simulated data is likely to contain several exclusions and thus result in very low LR:s (for SNP data, our model for mutations is still applicable but perhaps less relevant). Comparing M3 to M1, we observe some bias but much less than in the previous example; see Table 6 and Fig. 6. Comparing M3 to M2, however, we now observe greater problems: We can separate out the simulated cases where crossovers within the clusters have been simulated. In these cases, it is obvious that the quotient of likelihood ratios quite often will be zero, in fact, in 17 % of the simulations as a consequence of recombination within the cluster. In the remaining simulated cases where there is no crossover simulated, the differences are much smaller when M2 is compared to M3, see Table 6. As in Case 1, the distribution of LR:s for method M2 has not been included in Fig. 6 as these values, where computable, fits closely the line $y = x$.

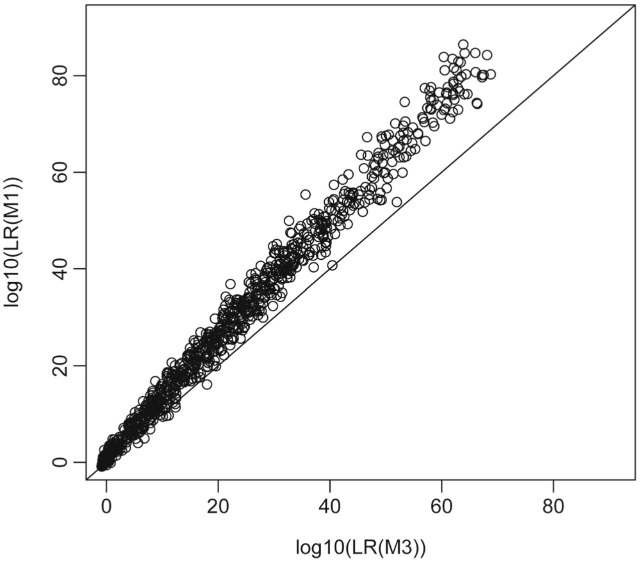

Fig. 4 Scatter plot based on 1000 simulations in Case 2. A pair of maternal cousins have been simulated using the model described in the text, alternative hypothesis is unrelated. The plot displays LR $= P(\text{Data} \mid H_0)/P(\text{Data} \mid H_1)$, as calculated using the two models M1 and M3. If all the *dots* are projected on the *x*-axis, we obtain the distribution of LR:s calculated using M3, while if we project all *dots* on the *y*-axis we obtain the distribution of LR:s calculated using M1. If, for a specific dot, $y = x$, we have concordance between the models

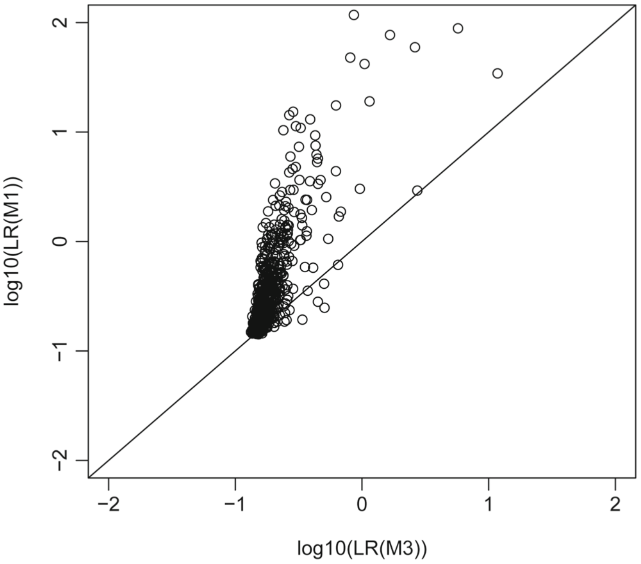

Fig. 5 Scatter plot based on 1000 simulations in Case 2. A pair unrelated pair have been simulated using the model described in the text, alternative hypothesis is maternal cousins. The plot displays LR $= P(\text{Data} \mid H_0)/P(\text{Data} \mid H_1)$, as calculated using the two models M1 and M3. If all the *dots* are projected on the *x*-axis, we obtain the distribution of LR:s calculated using M3, while if we project all dots on the *y*-axis we obtain the distribution of LR:s calculated using M1. If, for a specific dot, $y = x$, we have concordance between the models

Table 5 Distribution of quotients of likelihood ratios (LR:s) from 1000 simulations, using method M3. We define LR $= Pr(\text{Data} \mid H_0)/Pr(\text{Data} \mid H_1)$, where H_0: Full sisters (data from mother), H_1: Maternal half siblings (data from mother)

	LR
Median	1.72E+6
95 % cred.	[1959, 1.72E+8]
Max/min	1.95E+10 / 0

Case 4 - Bias in sibship test

In the last case, we explore the distribution of LR:s for SNP markers when siblingship is disputed. This may be relevant in a larger medical study where we wish to include only full siblings. We may indeed wish to study the inheritance of some disease, possibly linked to the X chromosome, where it is crucial to determine if the alleged siblings share the same father or/and mother. X- chromosomal marker data provides a crucial bit of information to resolve this. We specify the hypotheses as,

H_1: Two females are full siblings.
H_2: The two females are maternal half siblings.
H_3: The two females are paternal half siblings.
H_4: The two females are unrelated.

We use data from a set of 100 clusters of biallelic SNPs, where each cluster consisted of two tightly linked SNP markers demonstrating a high level of LD. To specify, for each cluster let the first SNP have alleles $[A, a]$ with frequencies $p(A) = 0.4$ and $p(a) = 0.6$, the second SNP have alleles $[B, b]$ with frequencies $p(B) = 0.4$ and $p(b) = 0.6$. We specify LD as conditional allele frequencies, $p(B|A) = 0.9833$, $p(B|a) = 0.025$, $p(b|A) = 0.0167$, and $p(b|a) = 0.975$, see also Supplementary Table S1.

Table 6 Distribution of likelihood ratios (LR:s) from 1000 simulations, using different methods (M1, M2, and M3) We define LR $= Pr(\text{Data} \mid H_0)/Pr(\text{Data} \mid H_1)$, where H_0: full sisters (data from mother), H_1:Maternal half siblings (data from mother)

	LR(M2)/LR(M3)	LR(M1)/LR(M3)
Median	1.0188	21.3122
95 % cred.	[1.0019, 1.0522]	[0.175, 2640]
Max/min	0.9994 / 1.2938	0.004 / 73008
Recombinations	1723	NA

We are only interested in the case where H_0 is true, since simulating H_1 will likely yield genetic inconsistencies when computing likelihoods given H_0. Row 'Recombinations' denotes the number of obligate recombinations within a cluster

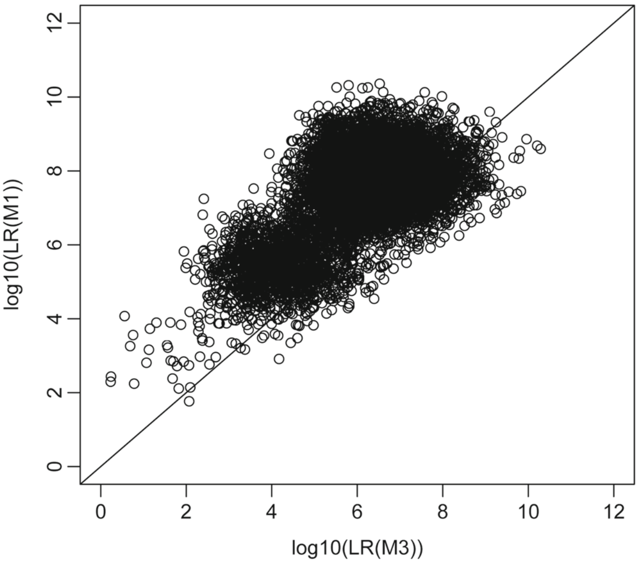

Fig. 6 Scatter plot based on 1000 simulations in Case 3. We have H_0: Full sisters (data from mother) have been simulated using the model described in the text, alternative hypothesis is H_1: maternal half sisters (data from mother). The plot displays the LR as calculated using the two models M1 and M3. If all the *dots* are projected on the *x*-axis, we obtain the distribution of LR:s calculated using M3, while if we project all *dots* on the *y*-axis, we obtain the distribution of LR:s calculated using M1. If, for a specific dot, $y = x$, we have concordance between the models

We compare results from methods M1 and M3, see Table 7, where we investigate the number of falsely classified relationships. Observe that for X- chromosomal data, the likelihood for hypotheses H_1 and H_3 may be close to zero when simulating H_2 and H_4 as multiple genetic inconsistencies is likely to be observed. With such an extreme level of allelic dependency, it is obvious that method M1, where we do not consider LD, overestimates the degree of relationship. Specially when we simulate 'unrelated', the 'Maternal half siblings' hypothesis often obtain the highest posterior probability and similarly when we simulate paternal half siblings, the full siblings hypothesis obtain the highest posterior probability. As for method M3, a majority of the falsely classified relationships are inconclusive, i.e., the probabilities are not sufficiently high to exclude the true relationship. To get further evidence as to the true relationship, we need to genotype further markers, autosomal or gonosomal, to obtain conclusive results.

Discussion

This paper describes a new probability model simultaneously handling linkage, linkage disequilibrium, and mutations, in likelihood computations using genetic marker data. An implementation of the model, aimed at X- chromosomal

Table 7 Number of correctly classified relationships from 1000 simulations, using methods M1 and M3. In the first row, each column corresponds to the true relationship while the other rows indicate the alternative hypotheses. For every column, each element contains the number of cases for each simulation where the indicated hypothesis has obtained the highest posterior probability. Results are displayed as number of cases with M1/number of cases with M3

True	Full siblings	Maternal half	Paternal half	Unrelated
Full siblings	999 / 992	0 / 0	734 / 2	0 / 0
Maternal half	1 / 0	998 / 803	0 / 1	440 / 8
Paternal half	0 / 1	0 / 0	101 / 984	0 / 0
Unrelated	0 / 0	0 / 15	0 / 0	168 / 813
Inconclusive	0 / 7	2 / 182	165 / 13	392 / 179

We specify a falsely classified relationship when the posterior probability is below 0.1 for the true relationship and correctly classified if the posterior probability is above 0.9. If the posterior probability is between 0.1 and 0.9 for the true relationship, the results are considered inconclusive

marker data, is provided in a user-friendly windows interface, FamLinkX, freely available at http://www.famlink.se. As stated in the paper, our model is not restricted to X-chromosomal data, but could fairly easily be implemented for autosomal markers. Similar to the Lander-Green algorithm, our model utilizes Markov chains to handle linkage between adjacent markers, i.e., the likelihood at a given marker is conditionally independent of any other marker given the IBD status of the previous marker. In addition, a multi-step Markov chain is used to handle linkage disequilibrium, similar to the extension presented by [17]. Whereas a single-step Markov chain is sufficiently good for linkage, LD requires a multi-step chain since allelic dependency can stretch further and a single-step chain is not a satisfactory approximation. Moreover, we implement a mutation model allowing for genetic inconsistencies in the data. The latter is crucial using STR marker data where mutations are not uncommon. This paper demonstrates the feasibility of the model whereas validation and details on the implementation is published elsewhere [manuscript in preparation].

Similar to the Lander-Green algorithm our model has limitation when it comes to the pedigree complexity, as the inheritance space grows exponentially with the number of meiosis, though to our best knowledge, this should not present any problems in any of the common pedigrees encountered in forensic genetics. For extended pedigrees with several typed persons the implementation requires optimization. We may for instance reduce the inheritance space [9, 16], e.g., by neglecting vectors with low contribution to the overall likelihood, though one must proceed carefully as many small likelihoods can yield a large combined contribution, especially when a mutation is necessary to explain the data. There are also some other considerations described by [11] and [15]. Moreoever, we assume Hardy-Weinberg equilibrium (HWE) for allele frequencies and whereas it is fairly easy to account for devations from HWE when the alleles at two loci are in LE, LD will provide computational problems where this is not as straightforward. Combining LD and subpopulation structure requires further developments of our model.

As we simulate data using our own stochastic model, it is of course given that our model will be best at differentiating between cases. However, our points are that

1. It is well documented that SNP data and, in particular, STR data do not fit the simplifications inherent in the other models discussed while they fit reasonably well in our model, so using it for simulations is reasonable.
2. Using the competing simplifications/models give a clearly different result in a considerable proportion of the cases.

In summary, our software provides the means necessary to solve extended/complex relatedness cases with clusters of X- chromosomal markers. In our implementation, we define clusters as groups of tightly linked markers also demonstrating linkage disequilibrium. It is worth noting that our model for LD is currently implemented as a two-step Markov chain. For SNP markers, it is fairly straightforward to extend this to a multi-step model spanning a greater number of steps, whereas for STR markers with more than 20 alleles, using more than two steps in the Markov model is computationally infeasible without additionalss speed-ups.

Appendix

The following section includes a more detailed description of the notation used in the paper. First, we assume locus i, $(i = 1, \ldots, I)$ has A_i possible alleles, and let p_i be a vector specifying the probabilities of a haplotype's alleles at locus i given the haplotype's alleles at lower indexes. We let $r_2, \ldots, r_I$ denote the recombination rates between the loci, which are assumed known. For a locus i, let t be a *transmission*, specifying a start allele in the parent, a resulting allele in the child, and whether the parent is a mother or a father. We then denote with $m_i(t)$ the probability that the child obtains the resulting allele, given that the parent has the start allele. This function specifies the *mutation model* at locus i. The *parameters* of our model are $p = (p_1, \ldots, p_I)$, $r = (r_2, \ldots, r_I)$, and $m = (m_1, \ldots, m_I)$.

If parents' alleles follow the population frequencies, the probabilities for a child to have various alleles are not given by the population frequencies, unless the process represented by the mutation model happens to have the population frequencies as stationary distribution. This means that adding the untyped father or mother to a person in the pedigree may change the probability results we are computing. To avoid this nuisance, we recommend that all untyped founders with only one child in the pedigree are (recursively) removed prior to computations. In our pedigree, a person may have specified no parents, only a mother, only a father, or both parents. Founders are those who have no parents in the pedigree. We also assume the pedigree does not contain untyped children with no descendants as such children cannot affect the result.

Our observed data is divided into data s for S typed founders and data d for M typed non-founders: Let s_{ij} for $i = 1, \dots, I$, $j = 1, \dots, S$ denote the observed allele or alleles of typed founder j at locus i. For males and X- chromosomal data, s_{ij} specifies only one allele, otherwise s_{ij} specifies the two observed alleles *in no particular order*. For the typed non-founders, let d_{ij} specify the similar data. We write $s_i = (s_{i1}, \dots, s_{iS})$, $s = (s_1, \dots, s_I)$, $d_i = (d_{i1}, \dots, d_{iM})$, and $d = (d_1, \dots, d_I)$.

We also need a number of ancillary variables: The inheritance pattern at locus i can be described as a vector v_i of length N, with one component for each parent-child relationship in the pedigree when the locus is autosomal, and one for each mother-child relationship for X- chromosomal loci. Each component is 0 or 1 depending on whether the paternal or maternal allele is inherited, we write $v = (v_1, \dots, v_I)$. We also need to describe the founder alleles of the pedigree: These are maternal or paternal alleles whose relevant parent is not in the pedigree. First, there are founder alleles belonging to typed founders: Let g_{ij} be the allele or alleles of typed founder j at locus i listed *with the paternal allele first*. Write $g_i = (g_{i1}, \dots, g_{iS})$ and $g = (g_1, \dots, g_I)$. For the remaining F founder *alleles*, let f_{ij} denote the $j'th$ founder allele at locus i. Finally, we write $f_i = (f_{i1}, \dots, f_{iF})$ and $f = (f_1, \dots, f_I)$.

References

1. Abecasis GR, Wigginton JE (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 77(5):754–67
2. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30(1):97–101
3. Boyles AL, Scott WK, Martin ER, Schmidt S, Li YJ, Ashley-Koch A, Bass MP, Schmidt M, Pericak-Vance MA, Speer MC, Hauser ER (2005) Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. Hum Hered 59(4):220–227
4. Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62(6):1408–1415
5. Chakraborty R, Stivers DN, Zhong Y (1996) Estimation of mutation rates from parentage exclusion data: applications to STR and VNTR loci. Mutat Res 354(1):41–48
6. Dawid AP, Mortera J, Pascali VL (2001) Non-fatherhood or mutation? A probabilistic approach to parental exclusion in paternity testing. Forensic Sci Int 124(1):55–61
7. Egeland T, Sheehan N (2008) On identification problems requiring linked autosomal markers. Forensic Sci Int Genet 2(3):219–25
8. Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21(6):523–42
9. Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nat Genet 25(1):12–13
10. Huang Q, Shete S, Amos CI (2004) Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. Am J Hum Genet 75(6):1106–1112
11. Idury R, Elston R (1997) A faster and more general hidden markov model algorithm for multipoint likelihood calculations. Hum Hered 47(4):197–202
12. Kling D, Egeland T, Tillmar AO (2012a) Famlink-a user friendly software for linkage calculations in family genetics. Forensic Sci Int: Genet 6(5):616–620
13. Kling D, Welander J, Tillmar A, Skare Ø Egeland T, Holmlund G (2012b) DNA microarray as a tool in establishing genetic relatedness—current status and future prospects. Forensic Sci Int: Genet 6(3):322–329
14. Krawczak M (2007) Kinship testing with X-chromosomal markers: mathematical and statistical issues. Forensic Sci Int: Genet 1(2):111–114
15. Kruglyak L, Lander ES (1998) Faster multipoint linkage analysis using fourier transforms. J Comput Biol 5(1):1–7
16. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58(6):1347
17. Kurbasic A, Hossjer O (2008) A general method for linkage disequilibrium correction for multipoint linkage and association. Genet Epidemiol 32(7):647–57
18. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A 84(8):2363–7
19. Nothnagel M, Szibor R, Vollrath O, Augustin C, Edelmann J, Geppert M, Alves C, Gusmao L, Vennemann M, Hou Y, Immel UD, Inturri S, Luo H, Lutz-Bonengel S, Robino C, Roewer L, Rolf B, Sanft J, Shin KJ, Sim JE, Wiegand P, Winkler C, Krawczak M, Hering S (2012) Collaborative genetic mapping of 12 forensic short tandem repeat (str) loci on the human x chromosome. Forensic Sci Int Genet 6(6):778–84
20. Pinto N, Gusmao L, Amorim A (2011) X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial. Forensic Sci Int Genet 5(1):27–32
21. Pinto N, Silva PV, Amorim A (2012) A general method to assess the utility of the x-chromosomal markers in kinship testing. Forensic Sci Int Genet 6(2):198–207
22. Skare O, Sheehan N, Egeland T (2009) Identification of distant family relationships. Bioinformatics 25(18):2376–82
23. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68(4):978–89
24. Szibor R (2007) X-chromosomal markers: past, present and future. Forensic Sci Int Genet 1(2):93–9
25. Szibor R, Krawczak M, Hering S, Edelmann J, Kuhlisch E, Krause D (2003) Use of X-linked markers for forensic purposes. Int J Legal Med 117(2):67–74

26. Tillmar AO (2012) Population genetic analysis of 12 X-STRs in Swedish population. Forensic Sci Int Genet 6(2):e80–81
27. Tillmar AO, Egeland T, Lindblom B, Holmlund G, Mostad P (2011) Using X-chromosomal markers in relationship testing: calculation of likelihood ratios taking both linkage and linkage disequilibrium into account. Forensic Sci Int Genet 5(5):506–511
28. Weir BS, Anderson AD, Hepler AB (2006) Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet 7(10):771–780

Paper V

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Familias 3 – Extensions and new functionality

Daniel Kling [a,b,*], Andreas O. Tillmar [c,d], Thore Egeland [b,e]

[a] Department of Family Genetics, Norwegian Institute of Public Health, Oslo, Norway
[b] Department for Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Aas, Norway
[c] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
[d] Department of Clinical and Experimental Medicine, Faculty of Health Sciences, Linköping University, Linköping, Sweden
[e] Department of Forensic Genetics, Norwegian Institute of Public Health, Oslo, Norway

ARTICLE INFO

Article history:
Received 14 March 2014
Received in revised form 23 June 2014
Accepted 1 July 2014

Keywords:
Familias
Paternity
Likelihood computations
Simulation
Mutation
Disaster victim identification

ABSTRACT

In relationship testing the aim is to determine the most probable pedigree structure given genetic marker data for a set of persons. *Disaster Victim Identification* (DVI) based on DNA data from presumed relatives of the missing persons can be considered to be a collection of relationship problems. Forensic calculations in investigative mode address questions like "How many markers and reference persons are needed?" Such questions can be answered by *simulations*. Mutations, deviations from Hardy–Weinberg Equilibrium (or more generally, accounting for population substructure) and silent alleles cannot be ignored when evaluating forensic evidence in case work. With the advent of new markers, so called microvariants have become more common. Previous mutation models are no longer appropriate and a new model is proposed. This paper describes methods designed to deal with *DVI* problems and a *new simulation model* to study distribution of likelihoods. There are softwares available, addressing similar problems. However, for some problems including DVI, we are not aware of freely available validated software. The Familias software has long been widely used by forensic laboratories worldwide to compute likelihoods in relationship scenarios, though previous versions have lacked desired functionality, such as the above mentioned. The extensions as well as some other novel features have been implemented in the new version, freely available at www.familias.no. The implementation and validation are briefly mentioned leaving complete details to Supplementary sections.

1. Introduction

There are several applications that require determination of genetic relatedness. The focus of this paper is to describe methods and implementations for complex relationships problems and disaster victim identification (DVI). While we have forensic applications in mind similar problems occur in a wide range of areas. The core computational problem is to calculate the likelihood of the data given competing hypotheses and from this to form the *likelihood ratio* (LR). We may further use a Bayesian approach with prior information to compute the posterior probabilities. In this paper we restrict attention to unlinked STR markers and then likelihoods are typically calculated using extensions of the Elston–Stewart (ES) algorithm [1] accommodating correction for population substructure (theta-correction), mutations and silent alleles [2]. The algorithm is in concept a peeling algorithm, where we consider subsets of a pedigree as conditionally independent given the connecting node. As a consequence, the algorithm may require long computation time, when marriage and inbreeding loops are present [3]. An implementation of the ES algorithm is provided in the software Familias [4]. The program is used by a large number of laboratories worldwide [5] when calculating likelihoods in relationship scenarios. Though previous versions of the software have included several important features, such as null/silent alleles, advanced mutation models and subpopulation correction, Familias has also lacked some desired functionality [6]. With the advent of new STR markers, micro-variant alleles (i.e., 9.3) have become more common necessitating an appropriate mutation model to handle transitions to and from such alleles. Whereas previous models are generally not designed to handle these transitions, this paper presents a new model, providing an extension of the stepwise mutation model [7,8], thereby accommodating for microvariants.

[*] Corresponding author at: Norwegian Institute of Public Health Department of Family Genetics, Gaustadalléen 30, NO-0027 Oslo, Norway. Tel.: +47 210 77663.
E-mail addresses: daniel.kling@fhi.no (D. Kling), andreas.tillmar@rmv.se (A.O. Tillmar), thore.egeland@nmbu.no (T. Egeland).

http://dx.doi.org/10.1016/j.fsigen.2014.07.004
1872-4973/© 2014 Elsevier Ireland Ltd. All rights reserved.

Monte Carlo simulation is a generic approach of relevance to virtually all areas of science. In our context, simulations can be used to get an idea of what evidential strength we will achieve for a given case. Based on simulations, one may for instance conclude that it is not worthwhile to proceed with a case unless more reference persons are genotyped or the number of genetic markers is increased. Simulation also extends the results from a point estimate of the LR to a complete description of its probability distribution. The model used for simulation is the same as the one used for likelihood and LR calculation. In other words, the simulations reflect the chosen mutation model, silent alleles and incorporate theta correction.

Disaster victim identification (DVI) applications can be considered as a potentially large collection of relationship estimation problems. Typically, LR ratios (sometimes converted to posterior probabilities) are reported and the aim is to compare large amounts of reference data, i.e., family members or personal belongings of missing persons, with unidentified remains. The underlying core computational model remains the same as in *standard* likelihood calculations. Since the early report on the successful use of DNA as a tool to identify victims of a mass disaster by Olaisen et al. [9], numerous papers have been published demonstrating its application and utility [10–15]. For the scope of this paper we consider smaller to medium sized DVI situation where the number of missing persons is typically limited to 1000.

As previously stated, the emphasis of this paper is on the new methods. Details on implementation and validation of the new software Familias 3 [hereafter called only Familias], which extends on Familias 2.0 [4], appear as supplementary material and in the manual. Some of the functionality of the new version or similar features can be found in other software [6,16,17]. However, (i) Familias is validated Drabek [6], (ii) widely used [5] (iii) freely available and (iv) the basic code is open (see http://familias.name/OpenFamilias). In addition, the implementation benefits from integrating similar problems (LR calculations, simulations and DVI feature) into one user friendly environment.

2. Methods

In relationship testing, mutually exclusive hypotheses are normally formulated. A hypothesis H corresponds to a *pedigree*, where the latter connects two or more individuals in a relationship tree. The core problem is to calculate the $P(data|H,\phi)$ where the *data* consists of alleles for different genetic markers and ϕ represents parameters needed to model e.g. mutations and subpopulation structure. The computation of the likelihood is in this paper based on the Elston–Stewart algorithm [1] and later extensions described in [18]. Briefly the algorithm peels the pedigree by calculating conditional probabilities for *cutsets*, where each *cutset* is conditionally independent given the rest of the pedigree, and can thus be effectively used on large pedigrees. The algorithm can also effectively accommodate many unlinked markers. Should we need to account for dependency between markers, other algorithms and implementations must be considered, e.g. FamLink[19] or Merlin [20]. For two different hypotheses H_1 and H_2, the likelihood ratio $\mathrm{LR} = P(data|H_1,\phi)/P(data|H_2,\phi)$ is typically calculated and reported.

The next section first describes the new mutation model, then the simulation approach and a framework to deal with DVI problems. Thereafter, some general principles related to validation are described. Finally, the implementation is briefly described deferring more complete descriptions to supplementary sections and the manual.

2.1. Mutation model

As mentioned, there is a need for a new mutation model capable of handling transitions to and from microvariants, e.g. between 9 and 9.3. Some current models treat such *microvariant mutations* (MVM) in the same way as *integer mutations* (IM) or neglect them as the mentioned transitions are considered improbable. This is biologically unreasonable and the problem has become more pronounced as MVM are more common in the latest STR kits. We provide a new stepwise mutation model accounting for MVM. The model is called the extended step wise model in the implementation.

We specify the model by letting M be the mutation matrix, with elements m_{ij}, where $i,j = 1, \ldots, N$ and where N is the number of alleles. Each element m_{ij} is the probability of a transition from allele A_i to allele A_j. The current model separates the overall mutation rate, denoted μ, into two parts, one corresponding to integer mutations, R, and one to the micro-variants α, i.e., $\mu = R + \alpha$. Biologically R is often explained by slippage error during DNA replication [8] while α is connected to insertions/deletions and point mutations. The last parameter, the mutation range r, is defined as for previous IM models; it is the value with which the probability decreases for each further step away from the original allele mutates.

Next the model is specified precisely by the transition probabilities m_{ij}. There are three different alternatives:

1. $m_{ij} = (1-(R+\alpha))$ if $i = j$, i.e. the probability that an allele does not mutate.
2. $m_{ij} = k_i(1-\alpha)r^{|i-j|}$ for integer mutations.
3. $m_{ij} = k_i\alpha/N_i$ for micro variant mutations. N_i is the number of MVM-s from allele i. The rows must sum to unity and therefore the normalizing constants k_i are determined by the constraints $\sum_{j=1}^{N} m_{ij} = 1$.

Example 1. Consider a marker containing the alleles 9, 9.3, 10, 10.3 and 15. The transition matrix M is then given by:

$$M = \begin{bmatrix} 1-(R+\alpha) & k_1\alpha/2 & (1-\alpha)k_1r^1 & k_1\alpha/2 & (1-\alpha)k_1r^6 \\ k_2\alpha/3 & 1-(R+\alpha) & k_2\alpha/3 & (1-\alpha)k_2r^1 & k_2\alpha/3 \\ (1-\alpha)k_3r^1 & k_3\alpha/2 & 1-(R+\alpha) & k_3\alpha/2 & (1-\alpha)k_1r^5 \\ k_4\alpha/3 & (1-\alpha)k_4r^1 & k_4\alpha/3 & 1-(R+\alpha) & k_4\alpha/3 \\ (1-\alpha)k_5r^6 & k_5\alpha/2 & (1-\alpha)k_5r^5 & k_5\alpha/2 & 1-(R+\alpha) \end{bmatrix}$$

In this case, k_1 is found as follows $1 = 1-(R+\alpha)+k_1\alpha/2+(1-\alpha)k_1r+k_1\alpha/2+(1-\alpha)k_1r^6 \Leftrightarrow k_1 = (R+\alpha)/(\alpha+(1-\alpha)(r+r^6))$. Similar calculation can be shown for the other k_i. Note that, the matrix M is not necessarily symmetric, meaning that the probability of observing a mutation from 9 to 9.3 is not the same as observing a mutation from 9.3 to 9. This is a consequence of the definition of M. Further note that for transitions from allele 9 for example, $N_i = 2$ as there are two MVM:s given allele 9 as starting point.

2.2. Simulation

Simulations provide means to calculate prediction intervals and investigate specific likelihood ratio thresholds for a given case. The probability of falsely including/excluding a true hypothesis with a given LR threshold can be estimated. The interface may be utilized to examine the number of genetic markers we need to obtain a sufficiently good LR, prior to deciding to accept a case, as well as providing intervals. The simulation interface accounts for all parameters in the model.

Specifically, the simulation algorithm starts by detecting all founders for a given pedigree. Founder genotypes are sampled using defined allele frequencies in combination with possible subpopulation correction, modeled by the parameter θ (sometimes denoted Fst in the literature). Furthermore, transitions within the pedigree are sampled using a transition matrix, the latter depending on the selected mutation model. Interested users may use raw data from the simulations to study observed mutation rates or the occurrence of silent alleles. Moreover, in addition to providing prediction intervals, the interface provide relevant functionality to study thresholds and false positive/negative rates, i.e., given two mutually exclusive hypotheses, H_1 and H_2, the probability $P(\mathrm{LR} \geq x|H)$ is estimated for a given threshold x and an assumed hypothesis H. Simply put, it gives the probability of obtaining a LR at least as great as a given threshold.

2.3. DVI

Since the introduction of DNA, genetic data from relatives or personal belongings of missing persons have become one of the most important and reliable means of identification [10–12,14,21]. The disaster victim identification (DVI) module in Familias is provided to assist in any operation that requires an all-against-all search. To specify, we have K number of unidentified DNA profiles and M number of reference DNA profiles. The former data set may be reduced to K' as identical DNA profiles are found through blind matching while the latter is reduced to L' if some of the M reference profiles belong to the same cluster, i.e., in this setting meaning the same reference family. We have $K \geq K'$ and $L \geq L'$. For $k = 1, \ldots, K'$ we compare each unidentified DNA profile k with the $l = 1, \ldots, L'$ reference families. In Familias we specify two sets of data; PM (Post Mortem) data – obtained from unidentified remains, where several of the remains as mentioned may originate from the same individual and AM (Ante Mortem) data, where we define missing persons. In the AM data we define reference families for each missing person, where we may have genetic data from relatives of the missing person or direct matching samples such as personal belongings. The reference family can contain arbitrary pedigree structures., Complex pedigrees, mutation models (see below) and theta correction, will typically produce longer computation times. The module calculates likelihoods for each combination of PM data and AM data. Using a Bayesian approach the likelihoods are converted to posterior probabilities including prior probabilities set by the user. The choice of priors has been debated elsewhere [22] and can be influenced in Familias by changing the size of the DVI operation. As of now, meta data is not used to adjust priors or to exclude unidentified persons based on gender. This may in some situations be appropriate as the meta data may have been incorrectly specified.

In addition to the DVI module, there is a blind search interface, allowing the user to search a set of persons for unknown relations. The feature may be used on any data set, e.g. to search for relations in a set of individuals before creating a population frequency database or as in the DVI situation to find direct matches or relations between PM samples. The blind searching is restricted to pairwise searches for a number of predefined relationships and implements a fast algorithm based on the formulas presented in Table 4 in Hepler et al. [23]. The algorithm does not account for mutation unless the parent–child relation is chosen; while theta correction is applied in all scenarios should the value be nonzero. Briefly, the formulas implemented are based on identical by descent (IBD) sharing probabilities not accounting for inbreeding. The general formula is,

$$P(data|H) = P(IBD = 0|H)g_0 + P(IBD = 1|H)g_1 + P(IBD = 2|H)g_2 \quad (1)$$

where $P(IBD = 0|H) = k_0$, $P(IBD = 1|H) = k_1$ and $P(IBD = 2|H) = k_2$ are the probabilities that two individuals share 0, 1 respectively 2 alleles identical by descent.; g_0, g_1 and g_2 are functions of allele probabilities depending only on the genotype *data*. A more general formula, also accounting for inbreeding can be derived, though its utility in the current setting is limited.

For the new direct matching feature, Familias implements a general approach. To specify, consider two profiles G_1 and G_2. Further, consider the competing hypotheses:

H_1: The profiles belong to the same person
H_2: The profiles belong to two unrelated persons

The hypotheses, and the current setting, is distinct from the more common situation where we have some trace evidence from a crime scene and a reference profile to compare with. The former being uncertain while the latter is commonly considered to be accurate.

To compute the LR we require some more definitions. We consider a *latent* genotype G_{true}, consisting of all possible genotypes for the current marker. We can now specify the LR as

$$LR = \frac{P(G_1, G_2|H_1)}{P(G_1, G_2|H_2)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} P(G_{true,i,j})P(G_1|G_{true,i,j})P(G_2|G_{true,i,j})}{P(G_1)P(G_2)} \quad (2)$$

where N is the number of alleles at the current marker and $P(G_{true,i,j})$ is the genotype probability, $p_i{}^*p_j$, for the *latent* genotype with alleles i and j. $P(G_1|G_{true,i,j})$ and $P(G_2|G_{true,i,j})$ are the transition probabilities from the *latent* genotype to the observed genotypes. To calculate the transition probabilities in the direct matching we specify three parameters, d = allelic dropout probability, c = allelic dropin probability and e = typing error probability. Here, we specify dropout as the probability of one allele not being unobserved for a heterozygous genotype (allelic dropout), dropin as the probability of an extra allele being observed for a homozygous genotype (allelic dropin) and typing error as the probability of some other laboratory error leading to an incorrect genotype [24]. Dropouts, dropins and errors are assumed to occur independently. Note that these parameters only apply to direct matching function and are not used in the kinship calculations. See Table 1 below for a list of $\mathbf{P(G_1,G_2|H_1)}$ **and** $\mathbf{P(G_1,G_2|H_2)}$ for some combinations of genotypes G_1 and G_2. (The formulas are simplified to fit, removing terms negligible in the calculations assuming $d > > c > e$; the implementation is exact, see Supplementary data 1 for a more thorough walkthrough of Eq. (2), including an example where the simplifying assumptions are omitted)

We see that if $d = c = e = 0$, the LR $[\mathbf{P(G_1,G_2|H_1)/P(G_1,G_2|H_2)}]$ in the first and fifth line of Table 1 reduces to$1/P(A,A)$ and $1/P(A,B)$, while the remaining lines simplifies to zero. Further note that if $d > >c > e$ and d is comparatively small, say below 0.1, several latent genotypes are unlikely as the transition probabilities are very small. Moreover, if subpopulation correction is nonzero the allele probabilities are not independent. The user-friendlyness of handling three parameters (d, c and e) instead of one can be discussed. Similar to Merlin [20], one may instead use a general error variable, including all the effects possibly causing an erroneous genotype.

Table 1
LRs based on the direct matching feature of Familias.

G_1	G_2	$P(G_1,G_2\|H_1)$	$P(G_1,G_2\|H_2)$
A,A	A,A	$[(1-d^2)^2(1-e)^2(1-c)^2]$P(A,A)	P(A,A)*P(A,A)
A,A	A,B	$(1-e)^2$[P(A,A)$(1-d^2)^2c$P(B) + P(A,B)$d(1-d)(1-c)(1-d^2)$]	P(A,A)*P(A,B)
A,A	B,B	$(1-e)^2[(1-c)^2d^2(1-d)^2]$P(A,B)	P(A,A)*P(B,B)
A,A	B,C	$(1-e)^2$[P(A,B)$d(1-d)c$P(C) + P(A,C)$d(1-d)c$P(B)]	P(A,A)*P(B,C)
A,B	A,B	$(1-e)^2[(1-d)^2(1-c)^2]$ P(A,B)	P(A,B)*P(A,B)
A,B	B,C	$(1-e)^2$[P(B,B)$(1-d)^2c^2$P(A)P(C)][a]	P(A,B)*P(B,C)
A,B	C,D	$(1-d^2)^2(1-c)^2$[P(A,B)e + P(C,D)e][a]	P(A,B)*P(C,D)

[a] Note that neither of the observed genotypes, G_1 or G_2, is probable as the latent genotype.

2.4. Validation

Validation can mean several things, including validation of methods and validation of the implementation. Here we focus on approaches that may be of general interest and which can be used to validate also other programs. Specific validation examples showing correct numerical results, i.e. results that can be derived by other means, typically exact formulae, appear in Supplementary data 2. (Some useful validation files are available at the Familias homepage)

2.4.1. Some useful validation formulae in simulations

The expected value of the LR assuming the denominator hypothesis H_2 to be true is 1

$$E(\mathrm{LR}|H_2) = (1-p)0 + p\frac{1}{p} = 1 \tag{3}$$

where p is the random match probability and 0 and 1/p are the two possible values for the likelihood ratio. This follows directly from the definition of the likelihood ratio and expectation as pointed out by Thompson [18]. This is true also if mutations and population substructure are modeled. Slooten and Egeland [25] presents further theoretical properties of LR:s. For instance, the identity

$$SD(\mathrm{LR}|H_2) = \sqrt{E(\mathrm{LR}|H_1) - 1} \tag{4a}$$

relating the standard deviation (SD) under H_2 to the expected value under $H_{1.}$ This last equation, however, is not valid when there are mutations or theta correction is made.

Eqs. (3) and (4) can be used to check simulations under the denominator hypothesis when p is not too small, typically for one marker. When p is small, say below 10^{-10} any reasonable number of simulations should lead to all LR-s being 0 as the probability of a random match is then negligible.

Turning to validation for simulations under the numerator hypothesis, the general formula for the expected value for all pairwise, non-inbred relationships presented in Slooten and Egeland [25] can be used

$$E(\mathrm{LR}|H_1) = \alpha L^2 + \beta L + (1-\alpha-\beta) \tag{4b}$$

where $L = \text{alleles}, \alpha = \frac{k_2^2}{2},$ and $\beta = \frac{k_1^2 + 4k_1k_2 + 2k_2^2}{4}$

As an example, note for a parent–child relation $k_1 = 1$ and $k_2 = 0$ and the expected LR is therefore $(L + 3)/4$ for one marker. This generalizes directly to n independent markers

$$E(\mathrm{LR}|H_1) = \prod_{i=1}^{n} \frac{L_i + 3}{4}$$

where L_i is the number of alleles for marker i.

We have checked the code using the above formulae for one marker at the time. To get an indication of the simulation uncertainty, several simulations can be run with different seeds.

Exact calculations are hard for general mutation models. There is, however, one exemption as explained next. Consider the hypotheses $H_{1:}$*AF is the father CH* and $H_{2:}$*AF and CH are unrelated.* The genotypes of AF and CH are denoted a/b and c/d. For instance, if both individuals are homozygote 9,9 then $a = b = c = d = 9$. A case which would need a mutation to be consistent with paternity occurs for genotypes 9,9.3 and 10,10.3 corresponding to $a = 9$, $b = 9.3$, $c = 10$ and $d = 10.3$. The likelihood ratio may be written [26]

$$\mathrm{LR} = \frac{1}{4}\frac{(m_{ac} + m_{bc})\,p_d + (m_{ad} + m_{bd})\,p_c}{p_c\,p_d} \tag{5}$$

where p denotes allele frequency. Example 2 below relies heavily on the above equation.

2.5. Implementation

The software functionality described herein is implemented in a Windows friendly software, Familias version 3.1.4 at the time of writing. See Supplementary data 2 for some validation examples. The mayor changes since Familias 2.0 is the introduction of the new mutation model, the simulation interface as well as the new DVI module. We also introduce a new blind match searching function implementing some new functionality, primarily connected to the direct matching, see previous description. The latest version of Familias is freely available at www.familias.no. Moreover, several other new features will be presented in the next releases, e.g. the possibility to model profiles with dropouts [Manuscript submitted].

3. Results

3.1. New mutation model and simulation

Example 2. In this example both simulation and the new mutation model is illustrated. Consider one marker with the mutation model and alleles as described in Section 2 of this paper. The mutation parameters are specified as:

$R = 0.005,\ r = 0.1 \text{ and } \alpha = 0.001.$

The mutation matrix M becomes

Allele	9	9.3	10	10.3	15
9	9.940e−01	2.973e−05	5.945e−03	2.973e−05	5.945e−08
9.3	1.982e−05	9.940e−01	1.982e−05	5.945e−03	1.982e−05
10	5.939e−03	2.973e−05	9.940e−01	2.973e−05	5.939e−07
10.3	1.982e−05	5.946e−03	1.982e−05	9.940e−01	1.982e−05
15	5.929e−06	2.967e−03	5.929e−05	2.967e−03	9.940e−01

For the numerical examples below, the allele frequencies for the alleles (9, 9.3, 10, 10.3, 15) are (0.05, 0.05, 0.20, 0.30, 0.40). From Eq. (5) we find, when the alleged father is 9, 9.3 and the child 10, 10.3

$$\begin{aligned}\mathrm{LR} &= \frac{1}{4}((5.94\mathrm{e}-03 + 1.98\mathrm{e}-05) * 0.2 + (2.97\mathrm{e}-05 + 5.94\mathrm{e} \\ &\quad -03) * 0.3)/(0.2 * 0.3) \\ &= 0.0124.\end{aligned}$$

which is accurately reproduced by Familias 3. Similarly, simulations closely reproduce the theoretical values. For instance, the expected value of the LR assuming AF and CH to be unrelated is 1 according to Eq. (3) and the computer output based on 10,000 simulations gives a value close to the theoretical. Furthermore, the expected LR assuming AF to be the father, $(L+3)/4 = (5+3)/5 = 2$, from Eq. (4b) is also consistent with simulations.

3.2. DVI module and blind search interface

To validate the DVI module simulated data was constructed for a number of relationships (Data available upon request). Specifically, 100 pairs of siblings, 100 pairs of grandparents/grandchildren and 100 pairs of parent/childs were generated using the simulation interface. For each pair one of the individuals was withdrawn and denoted as missing. All missing persons, in total 300, were collected into a data set of unidentified remains. The reference families were constructed according to the simulated relationship, i.e., 100 families where the reference data was from siblings, 100 families where the reference data was from grandparents and 100 families where the reference data was from a parent. An all-against-all search was performed in the DVI module, where LRs were calculated for all possible combinations of unidentified remain and reference family. In total $300 \times 300 = 90{,}000$ comparisons were done, producing a list of matches above a given threshold (in this case set as low as LR = 1). The match list indicated some false matches (i.e. false inclusions), which is most probably due to the low LR threshold. However, no false match obtained a LR higher than the *true* match. Some true matches for the missing persons obtained very low LR barely above 1.0, which was in some of the cases explained by simulated mutations (grandparents and parent) and in other cases by low number of shared alleles (sibling cases), (Data available upon request). See also Ge et al. for a discussion on choice of reference family relatives in DVI operations [27]. The point with this validation is not to investigate the match threshold but rather to demonstrate the accuracy in the calculations.

We further use constructed data to validate the blind searching function. Consider a system with alleles similar to the first example, i.e., the allele frequencies for the alleles [9, 9.3, 10, 10.3, 15] are [0.05, 0.05, 0.20, 0.30, 0.40]. For simplicity we let the mutation rate be zero, while we consider both $\theta = 0$ and $\theta = 0.01$. Consider two persons P1 and P2 with genotypes G_1 and G_2. We can now easily calculate the likelihood ratio for the predefined relationships in the blind search interface using Eq. (1). Note that the interface allows us to scale versus some other relationship rather than unrelated, but for the current calculation we use unrelated as the alternative hypothesis.

Let G_1 = 9,9 and G_2 = 9,10. For $\theta = 0.01$ we need to calculate the updated set of frequencies, [$p(9)$, $p(9|9)$, $p(9|9{,}9)$, $p(10|9{,}9{,}9)$] = [0.05, 0.0595, 0.0688, 0.194], using formulas in Balding et al. [28]. Note that this set of frequencies will change if two alleles are IBD, i.e. for IBD = 2 and IBD = 1 we need to update the frequencies as only two respectively three alleles are drawn from the population.

Table 2
LRs for some relationship hypotheses, calculated versus unrelated as alternative hypothesis, for a pair of individuals P1 and P2.

Relationship	LR ($\theta = 0$)	LR ($\theta = 0.01$)
Direct match	29.07	18.12
Siblings	5.25	3.919
Half siblings	5.5	4.169
Cousins	3.25	2.584
Parent–child	10	7.338
2nd cousins	1.5625	1.396

The likelihood for the different hypotheses of relatedness can now be calculated from Eq. (1) as, $L(Data|H) = k_0 2\,p(9)^2\,p(9)\,p(10) + k_1\,p(9)\,p(9)\,p(10)$, where k_0 and k_1 are replaced by the values according to the relationship H.

The direct match LR can be calculated according to Eq. (2), by summing over all possible genotypes for the *latent* genotype and compute the likelihood for each case according to: (We specify $d = 0.1$, $e = 0.001$ and $c = 0.001$, which are the default values in Familias)

$$
\begin{aligned}
\mathrm{LR} &= \frac{P(G_1, G_2|H_0)}{P(G_1, G_2|H_1)} = \frac{\sum_{i=1}^{A}\sum_{j=i}^{A} P(G_{true,i,j})P(G_1|G_{true,i,j})P(G_2|G_{true,i,j})}{P(G_1)P(G_2)} \\
&= \langle \text{We simplify and remove terms which is negligible in the numerator} \rangle \\
&= \frac{p(9)\,p(9|9,\theta)P(9,9|9,9)P(9,10|9,9) + 2\,p(9)\,p(9|10,\theta)P(9,9|9,10)P(9,10|9,10)}{2\,p(9)\,p(9|9,\theta)\,p(9|9,9,\theta)\,p(10|9,9,9,\theta)} \\
&= \frac{(1-e)^2[\,p(9)\,p(9|9,\theta)(1-d^2)c\,p(10) + 2\,p(9)\,p(9|10,\theta)d(1-d)(1-d)^2(1-c)]}{2\,p(9)\,p(9|9,\theta)\,p(9|9,9,\theta)\,p(10|9,9,9,\theta)}
\end{aligned}
$$

The theoretical values in coincide with the values calculated in Familias (Table 2). Also note that the Direct match obtain a high LR even though the profiles are not identical, this is due to the high values on the parameters d, c and e.

3.3. Simulations

To further corroborate output from the simulation interface we compared results on some standard forensic cases with simulations reported in Table 6 of Ge et al. [27], see Table 3. The investigated relationships are described elsewhere, op.cit., but are based on simulations on the standard 13 CODIS STR markers, in order to determine how many relatives are necessary in a given

Table 3
Distribution of log10 likelihood ratios for 10,000 simulations using three different methods.

Method	Pedigree	Mean	5percentile	1percentile
Familias3_no_mut	Both parents	10.25	8.1	7.4
Ge et al.	Both parents	10.26	8.07	7.34
Familias3_mut	Both parents	10.17	7.85	6.63
Familias3_no_mut	One parent/One child	4.08	2.47	1.90
Ge et al.	One parent/One child	4.09	2.48	1.92
Familias3_mut	One parent/One child	4.07	2.43	1.69
Familias3_no_mut	2 full sibs	5.88	2.64	1.25
Ge et al.	2 full sibs	5.88	2.65	1.34
Familias3_mut	2 full sibs	5.86	2.48	1.09
Familias3_no_mut	1 halfsib	0.92	−0.59	−1.16
Ge et al.	1 halfsib	0.91	−0.57	−1.16
Familias3_mut	1 halfsib	1.16	−0.70	−1.29
Familias3_no_mut	2 children (same parent 2)	6.97	4.31	3.43
Ge et al.	2 children (same parent 2)	6.98	4.33	3.35
Familias3_mut	2 children (same parent 2)	6.94	4.24	3.14

The methods are Familias 3 (with and without mutations considered) as well as results presented by Ge et al., the pedigrees are described elsewhere [27].

case to obtain sufficient LRs. The Familias simulation interface produces almost identical output as presented by Ge and colleagues. As a comparison we also included simulations using the extended stepwise mutation model and the results are still close to the simulations without mutations.

4. Discussion

Familias is a well-known software in the forensic community and used by a number of laboratories [5]. The software facilitates the interpretation of the evidence by computing likelihood ratios and posterior probabilities for a given set of relationship hypotheses and genetic marker data. This paper describes methods implemented in the new version (Familias 3), providing considerable extensions to previous versions [4].

A comprehensive simulation interface provides versatile functionality for studying distribution of likelihood ratios for a given case. Users may now investigate a case prior to accepting it by computing prediction intervals and decide whether decisive evidence is likely to be obtained. The authors are aware of the discussion in the forensic community on the use of case specific thresholds rather than using an general LR/Posterior probability threshold for all cases. We do not propagate for lowering the threshold only because for a given case the evidence will never reach the required value. The users should instead study the false positive/negative rates to find an appropriate limit. As presented in this paper, the algorithm can simulate arbitrary pedigree structures where the only limitation is set by the computation time.

To assist in mass disaster identifications, we have developed a DVI module, allowing users to handle small to medium scale identifications. There are several papers and online discussions following previous larger scale mass disaster incidents, e.g. the Tsunami disaster [12], the WTC terror attack [11,14] and the hurricane Katrina [29,30]. This paper includes some points on the implementation and interested users should follow the references given above for further mathematical discussions. Similar to the simulation interface, the DVI module adopts the full functionality of Familias, allowing for subpopulation frequency correction, silent alleles and mutations. The module further allows the definition of multiple alternative family hypotheses, within each family, thus permitting each reference family to have several missing persons and the user can weigh the evidence given a match based on the possibility that the unidentified person may fit in several locations in a family tree.

To further aid in the identification of unidentified remains, a blind search tool is included. As presented in this paper the tool can be used to rapidly scan data sets for unknown relations; unknown in the sense that we have no prior knowledge how the individuals in the data set are related. In addition to assist in DVI operations the search can also be performed to verify that data sets for the creation of population frequency databases do not contain related individuals. The blind search is restricted to pair wise comparisons on a number of predefined relationships implementing the formulas presented in Hepler et al. not accounting for inbreeding and mutations [23]. As the formulas are general in the sense that any non-inbreed pair wise relationship can be defined, the implementation in Familias opens up for future extensions where any non-inbred relationship between two individuals could be specified using the k_0, k_1 and k_2 parameters, see Eq. (1). Furthermore, the search also includes a newly developed direct matching function (also part of the DVI module), which incorporates dropout, dropin and typing error probabilities. The latter is probably hard to estimate but can in some situations not be neglected, and therefore equally important as the two first mentioned probabilities.

Further, to cope with the increasing polymorphism in the new STR markers, we have developed a new mutation model. The model builds on the stepwise model [7], but provides extensions for microvariants, e.g. 9.3. Microvariants are more and more common, for instance the STR marker SE33 (ACTPB2) includes several alleles with a non-integer repeat unit and even though mutation rates for transitions between non-integer alleles and integer alleles may sometimes be negligible we require an appropriate model to handle them. This transition model is not stationary. In other words, the distribution of allele frequencies will change slightly with each generation in the pedigree. A stationary version of the above model would be a welcomed extension. Such an extension should preserve the main features like the diagonal elements, i.e., the overall mutation probability. We have not yet been able to derive such a stationary model.

In summary, the software Familias has previously been proven to be a resourceful tool in calculations concerning genetic relatedness [4–6]. We believe the extensions provided in this paper will be important for many users where previous versions have lacked desired functionality. The latest version can be freely downloaded at http://www.familias.no.

Acknowledgements

The authors would like to thank the contributions from users testing and evaluating the new features of Familias. The work of the last author leading to these results was financially supported from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 285487 (EUROFORGEN-NoE).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.07.004.

References

[1] R.C. Elston, J. Stewart, A general model for the genetic analysis of pedigree data, Hum. Hered. 21 (1971) 523–542.
[2] J. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press Inc., Bosa Roca, USA, 2004.
[3] C. Cannings, E. Thompson, M. Skolnick, Probability functions on complex pedigrees [domesticated mammals, laboratory animals], Adv. Appl. Probab. 10 (1978) 26–61.
[4] T. Egeland, P.F. Mostad, B. Mevåg, et al., Beyond traditional paternity and identification cases. Selecting the most probable pedigree, Forensic Sci. Int. 110 (2000) 47–59.
[5] L. Poulsen, S.L. Friis, C. Hallenberg, et al., A report of the 2009–2011 paternity and relationship testing workshops of the English Speaking Working Group of the International Society For Forensic Genetics, Forensic Sci. Int. Genet. 9 (2013) e1–e2.
[6] J. Drabek, Validation of software for calculating the likelihood ratio for parentage and kinship, Forensic Sci. Int. Genet. 3 (2009) 112–118.
[7] T. Ota, M. Kimura, A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population, Genet. Res. 22 (1973) 201–204.
[8] H. Ellegren, Heterogeneous mutation processes in human microsatellite DNA sequences, Nat. Genet. 24 (2000) 400–402.
[9] B. Olaisen, M. Stenersen, B. Mevag, Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster, Nat. Genet. 15 (1997) 402–405.
[10] B. Leclair, C.J. Fregeau, K.L. Bowen, et al., Enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: the Swissair flight 111 disaster, J. Forensic Sci. 49 (2004) 939–953.
[11] L.G. Biesecker, J.E. Bailey-Wilson, J. Ballantyne, et al., Epidemiology. DNA identifications after the 9/11 World Trade Center attack, Science 310 (2005) 1122–1123.
[12] C.H. Brenner, Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities, Forensic Sci. Int. 157 (2006) 172–180.
[13] M. Prinz, A. Carracedo, W.R. Mayr, et al., DNA Commission of the International Society for Forensic Genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI), Forensic Sci. Int. Genet. 1 (2007) 3–12.

[14] B. Leclair, R. Shaler, G.R. Carmody, et al., Bioinformatics and human identification in mass fatality incidents: the world trade center disaster, J. Forensic Sci. 52 (2007) 806–819.

[15] L. Bradford, J. Heal, J. Anderson, et al., Disaster victim investigation recommendations from two simulated mass disaster scenarios utilized for user acceptance testing CODIS 6.0, Forensic Sci. Int. Genet. 5 (2011) 291–296.

[16] C.H. Brenner, Symbolic kinship program, Genetics 145 (1997) 535–542.

[17] K. Slooten, Validation of DNA-based identification software by computation of pedigree likelihood ratios, Forensic Sci. Int. Genet. 5 (2011) 308–315.

[18] E.A. Thompson, Statistical Inference from Genetic Data on Pedigrees, JSTOR, 2000.

[19] D. Kling, T. Egeland, A.O. Tillmar, FamLink – a user friendly software for linkage calculations in family genetics, Forensic Sci. Int. Genet. 6 (2012) 616–620.

[20] G.R. Abecasis, S.S. Cherny, W.O. Cookson, et al., Merlin – rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (2002) 97–101.

[21] C.H. Brenner, B.S. Weir, Issues and strategies in the DNA identification of World Trade Center victims, Theor. Popul. Biol. 63 (2003) 173–178.

[22] B. Budowle, J. Ge, R. Chakraborty, et al., Use of prior odds for missing persons identifications, Investig. Genet. 2 (2011) 15.

[23] B.S. Weir, A.D. Anderson, A.B. Hepler, Genetic relatedness analysis: modern data and new challenges, Nat. Rev. Genet. 7 (2006) 771–780.

[24] A. Kloosterman, M. Sjerps, A. Quak, Error rates in forensic DNA analysis: definition, numbers, impact and communication, Forensic Sci. Int. Genet. 12 (2014) 77–85.

[25] K.-J. Slooten, T. Egeland, Exclusion probabilities and likelihood ratios with applications to kinship problems, Int. J. Legal Med. (2013) 1–11.

[26] F. Ricciardi, K. Slooten, Mutation Models for DVI analysis, Forensic Sci. Int. Genet. 11 (2014) 85–95.

[27] J. Ge, B. Budowle, R. Chakraborty, Choosing relatives for DNA identification of missing persons, J. Forensic Sci. 56 (Suppl. 1) (2011) S23–S28.

[28] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.

[29] S. Donkervoort, S.M. Dolan, M. Beckwith, et al., Enhancing accurate data collection in mass fatality kinship identifications: lessons learned from Hurricane Katrina, Forensic Sci. Int. Genet. 2 (2008) 354–362.

[30] S.M. Dolan, D.S. Saraiya, S. Donkervoort, et al., The emerging role of genetics professionals in forensic kinship DNA identification after a mass fatality: lessons learned from Hurricane Katrina volunteers, Genet. Med. 11 (2009) 414–417.

Paper VI

FamLinkX – A general approach to likelihoods computations for X-chromosomal markers

Abstract

The use of genetic markers located on the X chromosome has seen a significant increase in the last years and their utility has been well studied. This paper describes the software FamLinkX, freely available at http://www.famlink.se, implementing a new algorithm for likelihood computations accounting for linkage, linkage disequilibrium and mutations. It is obvious that such software is sought for among forensic users as more and more X-chromosomal markers become available. We provide some simulated examples demonstrating the utility of the implementation as well as its application in forensic casework. Though algebraic derivations are generally unfeasible, the paper outlines some theoretical considerations and provides a discussion on the validation of the software. The focus of this paper is to compare the software to existing methods in a forensic setting, perform a validation study as well as to provide an idea of the discriminatory power for X-chromosomal markers.

Keywords

Linkage, X chromosome, FamLinkX, Linkage disequilibrium, Mutations

1. Introduction

There has been an emerging focus on X-chromosomal markers in recent years [1-6]. The most recent commercial Investigator Argus X12 kit from Qiagen divides 12 STR markers into four distinct clusters on the relatively short X chromosome [2, 7-10]. There is furthermore several in-house developed STR kits such as the Decaplex by Gusmao et al [11] as well as SNP multiplexes by Pereira et al [12, 13]. Several papers have presented approaches to handle X-chromosomal marker data in statistical calculations, though none have presented a general approach [5, 6, 14]. As the X chromosome is relatively short, and many of the kits include several markers, genetic linkage and also linkage disequilibrium (LD) is a necessary concern [1, 3]. Whereas genetic linkage is the dependence between two markers within a pedigree, i.e. two alleles at the two markers may be inherited as a unit, with increasing probability the shorter the distance is, linkage disequilibrium, also referred to as allelic association, is the non-random association of alleles at two different loci (See Thompson et al for a more thorough review [15]). The consequence is that we require an appropriate statistical model to handle these concepts in the statistical calculations.

One of the suggested solutions has been to use only one marker from each cluster, where the markers are divided into haplogroups/clusters, thus, according to propositions, removing the problem with LD. This is not an attractive approach from a statistical point of view as we may lose a lot of information in the discarded data. Another naive solution may be to disregard linkage/LD and instead use the product rule, assuming the loci to be independent. As will be obvious this can greatly overestimate the evidence and may even cause false positives. In addition, in a forensic setting we need good models to handle mutations as the mutation rates for STRs are not negligible.

Recently, we presented a joint probability model to handle genetic marker, accounting for linkage, linkage disequilibrium and mutations [16]. The model is implemented in FamLinkX for X-chromosomal markers. This paper will provide a walkthrough on some examples and prove the necessity of the software in a forensic setting. Note, the paper does not describe the general utility of X-chromosomal markers, but rather the utility of our algorithm in likelihood calculations involving linkage, linkage disequilibrium and mutations, specifically implemented for X-chromosomal data. Moreover, the paper demonstrates the application and validation of the software in forensic casework, whereas Kling et al [16] provide a more thorough general description of the algorithm.

2. Model

2.1. General description

The complete description of the model is provided in Kling et al [16]. In short, the model embodies the idea of Markov-chains to account for dependency between markers as well as dependency between alleles at different loci. A one-step Markov-chain is used to handle linkage between neighboring markers, similar to the Lander-Green algorithm [17], such that the likelihood at a given marker is independent of all other markers given the previous marker. In addition, a multi-step Markov-chain is employed to account for allelic dependency (linkage disequilibrium). The algorithm starts by defining all meiosis in a given pedigree. The combination of all possible meiotic outcomes defines the inheritance space. In addition, we define all founders and all founder alleles for the given pedigree. All possible combinations of founder alleles define the founder allele space. The algorithm proceeds by calculating the likelihood for marker 1 using the defined mutation model and the founder allele space to compute pedigree likelihoods. At marker 2 we also consider all possible paths from marker 1, such that the dependency is a sum over all possible inheritance patterns and founder allele patterns yielding non-zero likelihood at the previous marker.

2.2. The λ model

We consider an approach using a Dirichlet distribution for haplotype probabilities see Equation 1 and Tillmar et al [3].

$$H_i = \frac{c_i + \lambda p_i}{C + \lambda} \quad (1)$$

Where H_i is the updated probability for haplotype i, given c_i number of observations, p_i is the prior probability for the haplotype calculated using the expected haplotype frequencies, C is the total number of observations in the database and lambda is a parameter giving weight to the prior haplotype probabilities. We explore the impact of lambda on the likelihood ratio for a number of cases in this paper, though more studies should be undertaken to decide if the model (1) is proper and which values for lambda that should be chosen. Equation 1 allows for unobserved haplotypes to be assigned an estimated frequency, but also observed haplotype frequencies are adjusted in the model.

2.3. Theoretical example

Maternity

Algebraic derivation is generally unfeasible, but for a maternity case with a female child the formula is simple enough. We consider X-chromosomal data (two markers) and hypotheses:

H_1: An alleged mother (AM) is the true mother of the child (C)

H_2: Another woman, unrelated to the alleged mother, is the true mother of the child.

We can now calculate the likelihoods $P(Data|H_1)$ and $P(Data|H_2)$.

$$
\begin{aligned}
&P(Data \mid H_1) = P(G_{1,C} \mid G_{1,AM}) P(G_{1,AM}) P(G_{2,C} \mid G_{2,AM}, G_{1,C}) P(G_{2,AM} \mid G_{1,AM}) \\
&= \{ \text{ Switching to model notation where V is inheritence pattern and F is founder allele set } \} \\
&= \sum_{V_2} \sum_{F_2} \ \Pr(D_2 \mid V_2, F_2) \left[\sum_{V_1} \Pr(V_2 \mid V_1) \Pr(V_1) \sum_{F_1} \Pr(F_2 \mid F_1) \Pr(F_1) \Pr(D_1 \mid V_1, F_1) \right]
\end{aligned}
$$

where $G_{i,C}$ denotes the genotype data for the child at locus i. Similar reasoning applies for the $G_{i,AM}$ for the mother. Using the notation of the model, described in detail in [16], we define D_i as the genotype data at locus i, V_i as the inheritance pattern at locus i and F_i as the founder allele set at locus i. (See also Supplementary Equation file for a general description of the derivations)

Similarly

$$
\begin{aligned}
&P(Data \mid H_2) = P(G_{1,C})P(G_{1,AM})P(G_{2,C} \mid G_{1,C})P(G_{2,AM} \mid G_{1,AM}) \\
&= \{ \text{ Switching to model notation where f is founder allele set } \} \\
&= \sum_{F_1} \sum_{F_2} \ \Pr(F_2 \mid F_1)P(F_1)
\end{aligned}
$$

It is obvious from the above equations that the main difference from the Elston-Stewart algorithm [18] lies in the part where we sum over inheritance patterns for two neighboring loci. For the mentioned algorithm, we typically need to sum over all possible haplotypes, which for many consecutively linked loci grows exponentially.

Now, consider markers L1 and L2 with alleles A_1,A_2,A_3 and B_1, B_2, B_3 respectively. Allele probabilities for L1 are $p(A_1)$=0.595, $p(A_2)$=0.395, $p(A_3)$=0.01 and the unadjusted conditional allele probabilities for L2 are $p(B_1|A_1)$=585/595, $p(B_1|A_2)$=10/395, $p(B_2|A_1)$=10/595 and $p(B_2|A_2)$=385/395. For B_3 we only have $p(B_3|A_3)$=1. To estimate remaining haplotype frequencies, i.e. $p(B_1|A_3)$, $p(B_2|A_3)$, $p(B_3|A_1)$ and $p(B_3|A_2)$, we use the lambda model described in Section 2.2 with λ=0.001 and as a consequence the formerly defined conditional allele probabilities will be slightly adjusted. The markers are separated by 0.1 cM which is accepted to be approximately equal to a recombination rate of 0.001. We let the mother be heterozygous $[A_1, A_2]$ at L1 and $[B_1, B_2]$ at L2 while the child is $[A_1,A_1]$ at L1 and $[B_2, B_3]$ at L2. $P(Data|H_2)$ is calculated as a summation of the product of the probability for all possible founder allele sets for L2 given all possible founder allele sets at L1, defined by the two alleles for the mother and the two alleles for the child. $P(Data|H_1)$ is calculated in two steps. The first step calculates a table of pedigree likelihoods for L1 given all possible inheritance patterns; in this case either 0, indicating the maternal allele from the mother is passed down, or 1, indicating the paternal allele from the mother is passed down, and given all possible founder allele sets, in this case given by the possible combinations of the two alleles from the mother and one of the alleles from the child, i.e. *[A1, A2, A1]*. Each element in the table is also multiplied with the probability for the given founder allele set and the probability for the given inheritance pattern. We proceed to L2 and similarly calculate a table of pedigree likelihoods for each set of inheritance pattern and founder allele combinations. We multiply each element with a summation over all elements from the table for L1 and, for each element, computes the conditional probability for the inheritance pattern at L2 given the inheritance pattern at L1 and the conditional probability for the founder allele pattern at L2 given the founder allele pattern at L1. At last we sum all elements of the table to yield the final likelihood. To specify

$$
\begin{aligned}
&P(Data \mid H_1) = \sum_{V_2}\sum_{F_2} \Pr(D_2 \mid V_2, F_2)\left[\sum_{V_1} \Pr(V_2 \mid V_1)\Pr(V_1)\sum_{F_1}\Pr(F_2 \mid F_1)\Pr(F_1)P(D_1 \mid V_1, F_1)\right] \\
&= \ldots = p(A_2)p(A_1)^2 p(B_3 \mid A_1) \\
&\left[(1-r)\left(p(B_1 \mid A_2)p(B_2 \mid A_1)\right) + r\left(p(B_2 \mid A_2)p(B_1 \mid A_1)\right)\right] \\
&= 1.93\text{e-}13
\end{aligned}
$$

where the last step is in fact an approximation as we here disregard mutations. Furthermore

$$
\begin{aligned}
&P(Data \mid H_2) = \sum_{F_1}\sum_{F_2}\Pr(F_1)\Pr(F_2 \mid F_1) = \\
&4p(A_1)^3 p(A_2)p(B_2 \mid A_1)p(B_3 \mid A_1)\left(p(B_1 \mid A_1)p(B_2 \mid A_2) + p(B_2 \mid A_1)p(B_1 \mid A_2)\right) = 5.36\text{e-}12
\end{aligned}
$$

$$
LR = \frac{P(Data \mid H_1)}{P(Data \mid H_2)} = \frac{1.93\text{e-}013}{5.36\text{e-}012} = 0.036
$$

This value coincides with the result computed with FamLinkX, if λ is very small; see Section 2.2 for a discussion on λ. If, on the other hand, λ is larger, say above 1, the conditional probabilities will change slightly and we will obtain different likelihoods and likelihood ratio (LR) with FamLinkX. Furthermore, we note that the recombination rate is actually in the final formula and does affect the result even for a simple case such as disputed maternity.

Recombination within a cluster

We next consider the theoretical derivation on a case of two sisters with data available from the mother. We consider X-chromosomal data and hypotheses:

H_1: The two sisters are full siblings (Data from mother)

H_2: The two sisters are maternal half siblings (Data from mother)

We specify again two allele systems (L1 and L2), with alleles A_1, A_2, A_3, and B_1, B_2, B_3 respectively. We use the same allele probabilities as in the previous example and let the mother be heterozygous A_1, A_2 at L1 and B_1, B_2 at L2 while the first sister is heterozygous A_1, A_3 at L1 and B_1, B_3 and the second sister is A_1, A_3 at L1 and B_2, B_3 at L2. Again, similar to the previous example we specify

$$P(Data \mid H_1) = 1/4\, p(A_3) p(B_3 \mid A_3) p(A_1) p(A_2) \big(2p(B_1 \mid A_1) p(B_2 \mid A_2) r(1-r) + 2p(B_2 \mid A_1) p(B_1 \mid A_2) r(1-r)\big)$$
$$= 1.12\text{e-}6$$

where we disregard the possibility of mutations and where r is the recombination rate. (For details on the derivation, see Supplementary Equations). Similarly

$$P(Data \mid H_2) = 1/4\, p(A_3)^2 p(B_3 \mid A_3)^2 p(A_1) p(A_2) \big(2p(B_1 \mid A_1) p(B_2 \mid A_2) r(1-r) + 2p(B_2 \mid A_1) p(B_1 \mid A_2) r(1-r)\big)$$
$$= 1.12\text{e-}8$$

and

$$LR = \frac{P(Data \mid H_1)}{P(Data \mid H_2)} = \frac{1.12\text{e-}6}{1.12\text{e-}8} = 100$$

This value again coincides with the result computed with FamLinkX, whereas for models not accounting for recombinations within the cluster, the LR will be zero in the given case.

Moreover, should we not account for LD the LR would instead become 10,000, i.e. 100 times larger.

3. Results

We provide several examples demonstrating the utility of our software as well as illustrating the general information content in X-chromosomal markers. Unless something else is stated, we use data from the Argus X12 kit from Qiagen which divides 12 STR markers on the X chromosome into four distinct clusters, each containing three closely linked markers. We use recombination frequencies and mutation rates from Nothnagel et al [1], see Table 1.

Table 1. Recombination frequencies and mutation rates for the STR markers included in the Argus X-8 and Argus X-12 multiplexes.

Cluster	Marker	Kit	Position (cM)*	Recombination frequency	Mutation rate
Cluster 1	DXS10148	X-12	10.000	-	0.0031
	DXS10135	X-8/X-12	11.123	0.0111048	0.0041
	DXS8378	X-8/X-12	11.263	0.00139804	0.0008
Cluster2	DXS7132	X-8/X-12	321.993	0.499	0.0027
	DXS10079	X-12	322.739	0.00740462	0.0049
	DXS10074	X-8/X-12	323.637	0.00889984	0.0024
Cluster 3	DXS10103	X-12	418.626	0.425199	0.0015
	HPRTB	X-8/X-12	419.687	0.0104982	0.0018
	DXS10101	X-8/X-12	419.697	9.999E-5	0.0006
Cluster 4	DXS10146	X-12	471.329	0.321967	0.0022
	DXS10134	X-8/X-12	473.422	0.020498	0.0028
	DXS7423	X-8/X-12	473.572	0.00149775	0.0009

* Converted from recombination rates to genetic positions via Haldane’s mapping function

All calculations in the following section are, unless something else is stated, performed using five different approaches; M1: A naïve approach where all markers are considered to be independently inherited, LE is assumed, M2: Linkage between all markers is considered, LE is assumed, M3: Only the first marker from each cluster is included in the calculations, linkage is not considered, M4: Linkage and LD is accounted for through a cluster approach [19] and M5: Linkage, LD and mutations are simultaneously considered while in addition recombinations within a cluster is considered, see Table 2. The last approach (M5) is presented in Kling et al [16] and is the model implemented in FamLinkX.

Table 2. Different approaches used in this paper.

Model	All markers	Linkage	LD	Recombinations between all markers	Mutations
M1	Yes	No	No	No	No
M2	Yes	Yes	No	Yes	No
M3	No	No	No	Yes*	No
M4	Yes	Yes	Yes	No	No
M5	Yes	Yes	Yes	Yes	Yes

* Recombination is considered for the subset of markers

3.1. Validation examples

3.1.1. Definition of pedigrees

We compute the likelihood for a case containing only one marker with two alleles and compare the results to theoretical values. This is repeated for all pre-defined pedigrees such that the specification of the pedigrees, i.e. the calculation of pedigree likelihoods, can be validated. We define a system with two alleles, 13 and 14. We specify allele probabilities *p(13)*=0.2 and *p(14)*=0.8. We further specify that all typed persons are homozygous 13, 13. Given only one marker it is fairly straightforward to compute the theoretical likelihood for the data given the different hypotheses. See Table 3 where the results in FamLinkX correspond to the theoretical values in all currently defined pedigrees. See software interface for the list of corresponding pedigrees and also Supplementary Table 1 for more exhaustive information (FamLinkX save files for all pedigrees are available at http://famlink.se/fx_index.html).

Table 3. Likelihood ratios for all predefined pedigrees in FamLinkX.

Pedigree	**1**[1]	**2**[1]	**3**[2]	**4**	**5**[1]	**6**[1]	**7**[1]	**8**	**9**[3]	**10**[4]	**11**	**12**[1]	**13**[5]	**14**[1]	**15**[2]	**16**[1]	**17**[6]	**18**[7]
M5 LR	5	5	5	-	15	3	5	-	25	5	-	5	5	3	3	4	3.3	3.3
Theoretical LR	5	5	5	-	15	3	5	-	25	5	-	5	5	3	3	4	3.3	3.3
Pedigree	**19**[6]	**20**[7]	**21**[6]	**22**[6]	**23**[8]	**24**	**25**	**26**	**27**	**28**	**29**[3]	**30**[3]	**31**	**32**[1]	**33**[1]			
M5 LR	16.7	41.7	3.3	5	5	-	-	-	-	-	5	5	-	2.5	3			
Theoretical LR	16.7	41.7	3.3	5	5	-	-	-	-	-	5	5	-	2.5	3			

We consider only one marker with data as given in the text. [1]LR is scaled versus Pedigree 8, [2]Versus Pedigree 4, [3]Versus Pedigree 31. [4]Versus Pedigree 11. [5]Versus Pedigree 25. [6]Versus Pedigree 24. [7]Versus Pedigree 26. [8]Versus Pedigree 28

3.1.2. Comparison with Tillmar et al

We compute the likelihood ratio for 12 different cases using FamLinkX and compare the results to the algorithm provided in Tillmar et al [3]. Results are displayed in Table 4 and there is concordance for all cases. (DNA-profiles for each case are available in Supplementary Table 2).

Table 4. Likelihood ratios for a selection of cases. The LR:s are calculated using two different methods described in the paper.

Case	**Pedigrees**	**LR** **(FamLinkX)**	**LR** **(Tillmar et al.)**	**Comment**
C1	Full siblings vs. maternal half siblings	4101.79	4101.79	DNA data available for two female children

C2	Full siblings vs. maternal half siblings	117.245	117.245	DNA data available for two female children
C3	Full siblings vs. maternal half siblings	79.2942	79.2942	DNA data available for two female children
C4	Full siblings vs. maternal half siblings	8.45247	8.45257	DNA data available for two female children
C5	Full siblings vs. maternal half siblings	0	0	DNA data available for two female children
C6	Full siblings vs. maternal half siblings	238238	238238	DNA data available for two female children
C7	Full siblings vs. maternal half siblings	654316	654316	DNA data available for two female children and their mother
C8	Full siblings vs. maternal half siblings	89763.1	89763.1	DNA data available for two female children and their mother
C9	Full siblings vs. maternal half siblings	14536100	14536100	DNA data available for two female children and their mother
C10	Maternal half siblings vs. unrelated	242130	242130	DNA data available for two male children
C11	Maternal half siblings vs. unrelated	0.0778469	0.0778469	DNA data available for two male children
C12	Maternal half siblings vs. unrelated	0.0903502	0.0903502	DNA data available for two male children

3.1.3. On the impact of values for λ and database size

We have computed the likelihood ratio on a number of cases in order to investigate the effect of different values of λ. 15 different cases were considered, comprising five different pedigree questions (full siblings versus unrelated, full siblings versus half siblings, mother (duo) versus unrelated, father (duo) versus unrelated and father (trio) versus unrelated), with data from the Argus-X12 kit. The tested individuals were from donor families with a priori known relationships. The LRs were computed with three different values of lambda, i.e. 1, 100 and 652 (The latter being the size of the database). As indicated in Table 5 the variation using different values of lambda is in general minor, although larger discrepancies can be observed. When a greater difference does exist, the case involves an earlier unseen haplotype, necessary in order to explain the true relationship. The subject highlights the importance of a decision on how much weight the "prior" information should have, i.e. the unconditional allele probabilities. Since there is no general consensus approach whether to use large or small values of λ, we recommend users to calculate LR:s with a selection of different values on λ and report the least extreme LR (analogous to earlier recommendation on how to handle silent alleles with unknown frequencies [20]).

Table 5. Comparison of LR:s computed using λ = 1, 100 and 652.

ID	Relationship case	LR(Lambda=100)/LR(Lambda=1)	LR(Lambda=652)/LR(Lambda=1)
L1	Full siblings vs. half siblings	0.32	0.11
L2	Full siblings vs. half siblings	1.3	2.7
L3	Full siblings vs. half siblings	1.3	3.4
L4	Full siblings vs. unrelated	0.0099	0.023
L5	Full siblings vs. unrelated	0.051	0.014
L6	Full siblings vs. unrelated	0.15	0.17
L7	Paternity (duo) vs. unrelated	1.1	1.6
L8	Paternity (duo) vs. unrelated	0.7	0.37
L9	Paternity (duo) vs. unrelated	0.54	0.35
L10	Maternity (duo) vs. unrelated	0.014	0.0058
L11	Maternity (duo) vs. unrelated	0.014	0.0039
L12	Maternity (duo) vs. unrelated	3.3	9
L13	Paternity (trio) vs. unrelated	17	35
L14	Paternity (trio) vs. unrelated	0.0014	0.00011
L15	Paternity (trio) vs. unrelated	0.00001	0.00000013
L16	Mat half sibs vs. unrelated	100000	100000

Furthermore, in order to highlight the relevance of having larger population databases we performed a simulation test with the goal to demonstrate the dependence of the power to detect LD in relation to the size of the database. We used X-chromosomal data that for a large Swedish population sample displayed LD for the pair of loci in each cluster using the Argus X-8 marker kit. The haplotype data was taken from Tillmar et al [21], comprising 718 Swedish males genotyped for the eight X-STRs. For the pair of loci within each cluster, significant p-values were found when performing exact test for allelic association of the complete data set ($P<0.001$, $P=0.001$ $P<0.001$ and $P<0.001$ for the pair of markers in each of the four clusters, respectively). To calculate the power, using smaller database sizes, we performed a simulation study where we randomly picked n profiles ($n=100$, 200, 400, 600) out of the 718 profiles, and repeated this 100 times for each n. For each iteration, we calculated the p-value using Fisher's exact test and calculated the number of instances where we got significant p-values ($P<0.05$). This resulted in the power estimates presented in Table 6 below.

Table 6. Power estimates for exact test of linkage disequilibrium.

		n=100	**n=200**	**n=400**	**n=600**
Linkage Group	1	0.22	0.61	0.97	1.00
	2	0.08	0.14	0.40	0.83
	3	0.68	1.00	1.00	1.00
	4	0.11	0.33	0.86	1.00

3.1.4. Mutations

We further compare LRs calculated using FamLinkX with LRs calculated using Familias 3 [22], to verify the implementation of the model accounting for mutations. As Familias does not generally deal with X-chromosomal data we consider some special cases where the inheritance patterns are identical for autosomal marker data. Data for five different cases was set up using the same mutation model ("extended step-wise model") and mutation rate (0.001) in both

FamLinkX and Familias. We specify one artificial cluster, including two markers. Haplotype frequencies were in LE and the markers were put at a virtual distance of 200 cM in order to obtain comparable results to Familias. Identical LRs were obtained in all cases (Table 7).

Table 7. Comparison of LR:s calculated using FamLinkX and Familias 3, with a genetic inconsistency present in the data.

Case scenario	Genotype data (Marker 1; Marker 2)	LR (FamLinkX)	LR (Familias 3)
Maternity (duo) vs. unrelated - One-step inconsistency	Female: 11/12; 11/12 Child: 11/13;13/13	0.000886	0.000886028
Maternity (duo) vs. unrelated - Two-step inconsistencies	Female: 11/11; 11/11 Child: 13/13;13/13	2.066E-08	2.06612E-08
Maternal half siblings vs. mother-child and unrelated - Two-step and one-step inconsistencies	Mother: 11/11;11/11 Child1: 11/11;12/12 Child2: 11/11;13/13	0.00108982	0.00108982
Trio vs. maternity (duo) - Two-step inconsistencies	Mother: 11/11;11/11 Child: 13/13;13/13 AF: 11;11	2.06612E-08	2.06612E-08
Trio vs. maternity (duo) - Two-step and one-step inconsistencies	Mother: 11/11;11/11 Child: 13/13;13/13 AF: 12;12	6.25E-07	6.25E-07

3.2. Simulated examples

Simulations are performed using an algorithm simultaneously accounting for linkage, linkage disequilibrium and mutations. To specify, the algorithm starts by defining all founder alleles. Founder alleles are sampled using haplotype frequencies where LD structure is accounted for. The model presented in Section 2.2 is used in the frequency estimation using λ=1. The algorithm continues by simulating transitions from the founder alleles to all non-founders using a mutation matrix where each transition is assigned a specific probability. In addition, recombination is considered when two neighboring markers are simulated, i.e. the algorithm keeps track of the maternal/paternal chromosomes. Calculations are then performed using only genetic data from the typed persons and the specified approach with FamLinkX. The simulations are intended to demonstrate the necessity of our model and provide insight into the utility of X-chromosomal marker data.

3.2.1. A comparisons on different computational approaches

The first example is used to illustrate the difference between the different approaches (M1, M2, M3, M4 and M5) on a number of selected cases, see Table 8. As the table indicates the largest difference can be seen for approach M3, which is obvious as we omit all but one marker from each cluster. For several of the cases we may observe genetic inconsistencies in the data and thus methods not accounting for mutation will have a likelihood equal to zero, see intervals for all quotas in Table 8 that, besides the "one marker" approach, M1,, includes zero. It is interesting to see that for some quotas the value is close to 1 for the median, i.e. in average we can use either of the compared approaches. However, it is even more important to notice that the 95% interval is quite large, suggesting that a considerable error could be made in some cases.

Table 8. Distribution of ratios of LR for a number of forensically relevant cases. (1000 simulations have been performed for each case).

Relationship	Versus	M1/M5	M2/M5	M3/M5	M4/M5
Paternity	Unrelated	1.67 [0, 713]	1.67 [0, 713]	1.25E-5 [1.6E-7, 1.6E-2]	1.02 [0, 1.03]
Full siblings (Data mother)	Maternal half siblings (Data mother)	4.9 [0, 2.7E+5]	8.8 [0, 6.0E+5]	9.25E-7 [4.2E-3, 3.6E-9]	5.76E-2 [0, 4.95E+3]
Paternal aunt	Unrelated	0.28 [3.1E-3, 11.9]	0.55 [2.7E-3, 15.7]	5.22E-2 [2.9E-5, 9.3]	1.02 [2.8E-3, 1.2]
Paternal grandmother	Unrelated	0.42 [0, 622]	0.42 [0, 622]	1.7E-4 [6.0E-7, 0.49]	1.07 [0, 1.11]

We simulated the hypothesis in the column *Relationship* and use the hypothesis indicated in the column *Versus* as the alternative hypothesis. The table contains the medians as well as a 95% credibility interval in parenthesis. The headers indicate the compared methods.

3.2.2. Discriminatory power of X-chromosomal markers

The following example provides a range of cases where we have calculated the LR using approach M5, i.e. the model implemented in FamLinkX, with data based on the Argus X12 kit from Qiagen. Table 9 illustrates the distributions on a range of cases where X-chromosomal markers are applicable.

Table 9. Distribution of LRs for a number of forensically relevant cases. (1000 simulations have been performed for each case). We simulated the hypothesis in the column Relationship and use the hypothesis indicated in the column Versus as alternative hypothesis

Relationship	Versus	Available DNA data	Median	Max	Min	95% cred.
Paternal half	Unrelated	Two female	2.1E+05	4.9E+14	1.1E+00	[3.0E+02,3.1E+07]

sisters		children				
Paternal half sisters	Full sisters	Two female children	1.4E+01	1.6E+01	1.1E-03	[1.5E-01,1.6E+01]
Full sisters	Maternal half sisters	Two female children	1.2E+04	3.3E+10	1.3E-02	[6.8E-01,2.1E+08]
Maternal half sisters	Paternal half sisters	Two female children	1.1E+01	1.5E+01	6.1E-03	[4.2E-01,1.4E+01]
Full sisters (Data mother)	Maternal half sisters (Data mother)	Two female children and their mother	7.2E+08	3.2E+12	5.1E+03	[4.8E+05,3.2E+10]
Paternal aunt	Unrelated	One female and one female child	4.1E+01	3.8E+07	6.3E-02	[7.9E-02,2.5E+05]
Paternal grandmother	Unrelated	One female and one female child	1.5E+05	7.8E+08	2.2E-02	[2.5E+01,3.6E+07]
Maternal grandmother	Unrelated	One female and one female child	2.7E+01	2.0E+07	9.0E-02	[9.5E-02,1.0E+05]
Paternity	Unrelated	One male, one female and one female child	1.1E+07	6.3E+11	4.0E+01	[7.1E+03,8.2E+08]
Maternal half sisters	Unrelated	Two female children	3.8E+01	1.6E+08	6.4E-02	[8.3E-02,1.8E+05]

3.3. Case examples

3.3.1. Case 1 – Three full siblings

The first example is an interesting case involving three young girls. The hypotheses were presented by the client as,

H_1: All three girls are full siblings.

H_2: Other possible relationships.

The definition of H_2 is apparently not sufficient and we require narrowing the list of possible alternative hypotheses. It was assumed that the three girls were all children, with no children of their own. We used Familias [23] to generate all possible relationship hypotheses based on the three girls and two extra, untyped, persons, and based on data from 35 autosomal STRs. In total 64 pedigrees were generated, using some restrictions on the pedigree structure, see supplementary Familias file. Out of these 64, only three obtained a posterior probability above 0.001, given the autosomal marker data. The hypotheses were reduced to:

H_1: All three girls are full siblings.

H_2: Two of the girls are full siblings, the last girl being a maternal half sibling.

H_3: Two of the girls are full siblings, the last girl being a paternal half sibling.

With posterior probabilities for the autosomal data H_1:0.99773, H_2:0.001135 and H_3:0.001135 respectively.

Obviously, autosomal data, disregarding mutations, cannot distinguish between *H_2* and *H_3*. All three children were subsequently typed with the Argus X12 kit, see Supplementary Table 3.The data were analyzed in FamLinkX with the mentioned hypotheses and haplotype frequencies from a Swedish population [2]. The result indicates a strong LR in favor of *H_1*, both when compared to *H_2*, LR=4.3E+13, and *H_3*, LR=5.3E+7. Combining the autosomal results with the X-chromosomal results in the following posterior probabilities, assuming equal priors: *H_1*: >0.99999, *H_2*: <0.000001, *H_3*: <0.000001 . Thus in favor of hypothesis *H_1*. (Mitochondrial data from HV1 and HV2 further supported *H_1* and *H_2*)

3.3.2. Case 2 – Relationship testing workshop of the ESWG (ISFG) (2013)

The case involved data from two females with the hypotheses:

H_1: The alleged mother is the true mother of the child.

H_2: Another female, unrelated to the alleged mother, is the mother of the child.

We used FamLinkX to compare the results for M3/M4/M5, see Table 10. As the provided frequency data was incomplete, in the sense that not all observed haplotypes were given, the total number of observations had to be adjusted, i.e. some assumptions about the haplotypes had to be made which may affect the results. As is obvious there is a large difference, almost 100 times, accounting for LD (M4 and M5) and not (M3).

Table 10. Results from the relationship testing workshop of the ESWG (ISFG) 2013. The LR has been calculated using three different methods, described in the text.

M3	M4	M5
8.887E+6	5.758E+8	5.755E+8

See supplementary data for FamLinkX save file.

3.3.3. Case 3 – Mutation case

The last case concerns a disputed half siblingship (paternal), between two female individuals. Testing with 15 autosomal markers yielded an LR of 0.014 (H_1: Paternal half siblings, H_2: Unrelated) and additional testing of 8 X-chromosomal markers resulted in one two step inconsistency at marker DXS10134 (Individual 1: 36/38.3 and Individual 2 34/34) (See Supplementary Table 4 for complete X-chromosomal data). Taking the possibility of a mutation into account, using the implemented "Extended mutation model" with the mutation rate given in Table 1, the calculation resulted in an LR of 0.025 (H_1: Paternal half siblings, H_2: Unrelated). In total the LR was calculated to 0.00035 (or 1 in 2857), thus evidence against a paternal half siblingship between the two tested female individuals.

4. Discussion

The recent progress in forensic genetics has promoted the use of X-chromosomal markers and several papers have assessed their utility and addressed the statistical complications [1, 3, 5, 6]. We have developed a new algorithm, simultaneously handling linkage, linkage disequilibrium and mutations for such marker data [16]. We have provided examples demonstrating the necessity of a proper model and why our software could be adopted in any calculations involving linked X-chromosomal markers. With this mind, we have developed a software, FamLinkX, freely available from http://www.FamLink.se, implementing our algorithm.

Furthermore, we argue that for X-STRs, larger sample sizes should be used (compared with the size used for standard allele frequency databasing) in order to increase the power to detect possible dependence between alleles at different loci and to obtain more accurate haplotype frequency estimates. We demonstrated that using sample sizes less than 200 will most often not detect LD, even if such exists in the population. This definitely highlights the importance and relevance of the database size when testing for independence and creating haplotype databases. Our model for haplotype frequency estimation of earlier not observed haplotypes relies on the value of lambda, see Equation 1. If such haplotypes are critical for evaluation of a particular case and no data from which to estimate the lambda value exists a generous bracket of plausible values for lambda could be considered, and thus compute a corresponding range of values for the LR. The least extreme LR from such analysis could then be used in the expert report. An interesting case was recently encountered where the disputed relationship was paternal half siblings (unrelated as the alternative hypothesis) for two females, see L16 in Table 5. The genotype data indicated a common haplotype possibly derived from the same father under the half sibling hypothesis. The LR ranged from <0.01 to >100 using λ=1 and λ=650 respectively. The explanation is that the shared haplotype is extremely uncommon and the result heavily relies on the value of λ. Furthermore, under the alternative hypothesis, "Unrelated", other more common haplotype configurations are more probable for the two females and thus the LR will be higher the more ”uncommon” the shared haplotype is. As discussed above we suggest reporting the least extreme LR for such a case.

FamLinkX does not implement a model for subpopulations effects, also referred to as θ/F_{st} correction. While the model could be adopted to include correction for these effects, the theoretical considerations are more complicated. In a setting where linkage disequilibrium is not present, allele probabilities are adjusted according to the number of observations for each founder allele, i.e. there is a dependency across founder genotypes. However, when accounting for LD, where we also have dependency for genotypes at different loci, other approaches must be considered. One possible solution may be to adjust haplotype frequencies instead, though more research needs to be done on the subject.

In summary, the paper provides ideas on how to validate the software FamLinkX as well as some theoretical derivations. We acknowledge the need to better understand the estimation of haplotype frequencies and their impact on the results. The software uses a Dirichlet model, described in equation 1, which relies on a parameter λ, giving weight to unobserved haplotypes. Further work is needed to improve the understanding of how this parameter affects the outcome and if a superior model is required. Nevertheless, no other implementation, to our knowledge, is currently available providing the same features as does FamLinkX.

References

[1] Nothnagel M, Szibor R, Vollrath O, Augustin C, Edelmann J, Geppert M, et al. Collaborative genetic mapping of 12 forensic short tandem repeat (STR) loci on the human X chromosome. Forensic Science International: Genetics. 2012;6:778-84.

[2] Tillmar AO. Population genetic analysis of 12 X-STRs in Swedish population. Forensic Science International: Genetics. 2012;6:e80-e1.

[3] Tillmar AO, Egeland T, Lindblom B, Holmlund G, Mostad P. Using X-chromosomal markers in relationship testing: Calculation of likelihood ratios taking both linkage and linkage disequilibrium into account. Forensic Science International: Genetics. 2010;5:506–11.

[4] Edelmann J, Lutz-Bonengel S, Naue J, Hering S. X-chromosomal haplotype frequencies of four linkage groups using the Investigator Argus X-12 Kit. Forensic Science International: Genetics. 2012;6:e24-34.

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

[6] Pinto N, Silva PV, Amorim A. A general method to assess the utility of the X-chromosomal markers in kinship testing. Forensic Science International: Genetics. 2012;6:198-207.

[7] Kling D, Dell'amico B, Haddeland P, Tillmar AO. Population genetic analysis of 12 X-STRs in a Somali population sample. Forensic science international Genetics. 2014;11:e7-8.

[8] Ferragut J, Bentayebi K, Castro J, Ramon C, Picornell A. Genetic analysis of 12 X-chromosome STRs in Western Mediterranean populations. International Journal of Legal Medicine. 2014:1-3.

[9] Zidkova A, Capek P, Horinek A, Coufalova P. Investigator® Argus X-12 study on the Population of Czech Republic: comparison of linked and unlinked X-STRs for kinship analysis. Electrophoresis. 2014.

[10] Chen MY, Ho CW, Pu CE, Wu FC. Genetic polymorphisms of 12 X-chromosomal STR loci in Taiwanese individuals and likelihood ratio calculations applied to case studies of blood relationships. Electrophoresis. 2014.

[11] Gusmão L, Sánchez-Diz P, Alves C, Gomes I, Zarrabeitia MT, Abovich M, et al. A GEP-ISFG collaborative study on the optimization of an X-STR decaplex: data on 15 Iberian and Latin American populations. International Journal of Legal Medicine. 2009;123:227-34.

[12] Pereira R, Pereira V, Gomes I, Tomas C, Morling N, Amorim A, et al. A method for the analysis of 32 X chromosome insertion deletion polymorphisms in a single PCR. International Journal of Legal Medicine. 2012;126:97-105.

[13] Pereira V, Tomas C, Amorim A, Morling N, Gusmão L, Prata MJ. Study of 25 X-chromosome SNPs in the Portuguese. Forensic Science International: Genetics. 2011;5:336-8.

[14] Szibor R. X-chromosomal markers: past, present and future. Forensic Science International: Genetics. 2007;1:93-9.

[15] Thompson EA. Statistical inference from genetic data on pedigrees: JSTOR; 2000.

[16] Kling D, Tillmar A, Egeland T, Mostad P. A general model for likelihood computations of genetic marker data accounting for
linkage, linkage disequilibrium and mutations. International Journal of Legal Medicine. 2014 (Paper accepted).

[17] Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Science U S A. 1987;84:2363-7.

[18] Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Human Heredity. 1971;21:523-42.
[19] Abecasis GR, Wigginton JE. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. American Journal of Human Genetics. 2005;77:754-67.
[20] Gjertson DW, Brenner CH, Baur MP, Carracedo A, Guidet F, Luque JA, et al. ISFG: Recommendations on biostatistics in paternity testing. Forensic Science International: Genetics. 2007;1:223-31.
[21] Tillmar AO, Mostad P, Egeland T, Lindblom B, Holmlund G, Montelius K. Analysis of linkage and linkage disequilibrium for eight X-STR markers. Forensic Science International: Genetics. 2008;3:37-41.
[22] Kling D, Tillmar AO, Egeland T. Familias 3-Extensions and new functionality. Forensic Science International: Genetics. 2014;13:121-7.
[23] Egeland T, Mostad PF, Mevåg B, Stenersen M. Beyond traditional paternity and identification cases. Selecting the most probable pedigree. Forensic Science International. 2000;110:47-59.