Universidade do Minho
Escola de Engenharia
Departamento de Informática

Raquel Simões

Distinguishing kinships beyond identity and paternity

October 2017
Raquel Simões

Distinguishing kinships beyond identity and paternity

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Eduardo Conde-Sousa
Nádia Pinto

October 2017
ACKNOWLEDGEMENTS

First and foremost, I have to thank my research supervisors, Eduardo Conde-Sousa and Nádia Pinto. Without their assistance and dedicated involvement in every step throughout the process, this paper would have never been accomplished. I would like to thank you very much for your support and understanding over this important stage of my life.

In September 2015, I went to the University of Minho for two years of study and this time has been highly productive and enriching. I was initially very doubtful if I should really follow this path, but today, if I went back I would do exactly the same thing again.

Getting through my dissertation required more than academic support, and I have many, many people to thank for listening to and, at times, having to tolerate me over this time. Most importantly, none of this could have happened without my family and closest friends. Thank you to all!
ABSTRACT

In kinship testing powerful statistical results are usually obtained when genetic information is expected to be shared between a pair of samples, which happens in paternity and identification testing. However, there are other pedigrees where genetic information sharing is not required, such as when a pair of full-siblings or avuncular, is analyzed. Studying these pedigrees, where the sharing of genetic information is not mandatory, will be the focus of this work. We will consider several kinship problems where two (exhaustive and mutually exclusive) hypotheses will be compared, through a statistical evaluation based on the calculation of a likelihood ratio (LR) where the probabilities of genotypic configurations, assuming one or another kinship hypothesis, are compared. This analysis will allow the identification of the proportion of cases where the statistical evaluation had weak results, and those where LR favored the false hypothesis of kinship for a widely used commercial kit of genetic markers, considering simulated profiles assuming the pedigrees in question. In addition, we will compare the statistical gain of increasing the battery of analyzed markers and infer the impact of considering the genetic information given by the knowledge of the genetic profile of a relative, as the undoubted mother in the case where the hypotheses “individuals A and B are related as full-siblings” and “the individuals A and B are unrelated” are asked to be compared. Furthermore, a validation of the Familias software for two individuals will be performed for the simplest assumptions - absence of mutation and absence of silent allele - through the implementation of the algebraic formulas already established.
RESUMO

Em testes de parentesco, resultados estatísticos poderosos são geralmente obtidos quando a partilha de informação genética é esperada entre um par de amostras, o que acontece em testes de paternidade e de identificação. No entanto, existem outros pedigrees onde a partilha de informação genética não é requerida, como quando um par de irmãos ou tia(o)/sobrinha(o) é analisado. Estudar estes pedigrees, onde a partilha de informação genética não é obrigatória, será o foco deste trabalho. Consideraremos vários problemas de parentesco em que duas hipóteses (exaustivas e mutuamente exclusivas) serão comparadas, através de uma avaliação estatística com base no cálculo de razões de verossimilhança (LR, do inglês likelihood ratio) onde as probabilidades das configurações genotípicas, assumindo uma ou outra hipótese de parentesco, são comparadas. Esta análise permitirá a identificação da proporção de casos em que a avaliação estatística teve resultados fracos, e aqueles onde o LR favoreceu a hipótese falsa de parentesco para um kit comercial amplamente utilizado de marcadores genéticos, considerando perfis simulados assumindo os pedigrees em questão. Adicionalmente, será comparado o ganho estatístico de aumentar a bateria dos marcadores analisados e será também inferido o impacto de considerar a informação genética dada pelo conhecimento do perfil genético de um parente, como a mãe indubitada no caso em que as hipóteses “indivíduos A e B estão relacionadas como irmãos” e “os indivíduos A e B não estão relacionados” são comparadas. Além disso, uma validação do software Famílias para dois indivíduos será realizada para os pressupostos mais simples - ausência de mutação e inexistência de alelo silencioso - através da implementação de fórmulas algébricas já estabelecidas.
## CONTENTS

1 INTRODUCTION
   1.1 Context and Motivation 2
   1.2 Genetic inheritance 4
   1.3 Genetic Markers 9
   1.4 Mutation Models 10
      1.4.1 The equal model 10
      1.4.2 The proportional model 11
      1.4.3 The stepwise model 11
      1.4.4 The extended stepwise model 12
      1.4.5 An example illustrating how use the mutation models 12
   1.5 Statistical Evaluation 15
      1.5.1 Hypotheses 15
      1.5.2 Likelihood Ratio 16
      1.5.3 Identity-by-descent versus identity-by-state 17

2 STATE OF THE ART 19
   2.1 Exact Algebraic Expressions 19
   2.2 Software 21
   2.3 Software Validation 22

3 OBJECTIVES 24

4 MATERIAL AND METHODS 25
   4.1 Kinship problems 25
   4.2 Database 26
   4.3 Simulation of genetic profiles 27
   4.4 Validation of Familias for the simplest assumptions - absence of mutation and presence of silent allele 28
   4.5 Measurement of the impact of considering a third individual and/or a greater number of markers 28

5 RESULTS 30
   5.1 Validation of Familias for the simplest assumptions - absence of mutation and silent allele 30
   5.2 Measurement of the impact of considering a third individual and/or a greater number of markers 31
      5.2.1 Full-Siblings versus Half-Siblings 31
5.2.2 Full-Siblings versus Unrelated 37
5.2.3 Half-Siblings versus Unrelated 42
5.2.4 Avuncular versus Unrelated 48
5.2.5 First Cousins versus Unrelated 53
5.2.6 Half First Cousins versus Unrelated 58

6 CONCLUSIONS 63
6.1 Validation of Familias for the simplest assumptions - absence of mutation and silent allele 63
6.2 Measurement of the impact of considering a third individual and/or a greater number of markers 63

7 FUTURE WORK 65
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>The seven traits observed by Mendel in pea plants, from Nature Education Adapted from Pierce (2013).</td>
<td>4</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Mendel’s monohybrid crosses, from Nature Education Adapted from Pierce (2013).</td>
<td>5</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Mendel’s dihybrid cross, from Nature Education Adapted from Pierce (2013).</td>
<td>6</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Examples of uniparental and biparental pedigrees.</td>
<td>20</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Representation of the six cases under study.</td>
<td>26</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Representation of the hypotheses $H_0$: “A and B are related as full-siblings” and $H_1$: “A and B are related as half-siblings” - Figure 5(a).</td>
<td>31</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Distribution of $LR = \frac{P(G</td>
<td>FS)}{P(G</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Distribution of $LR = \frac{P(G</td>
<td>FS)}{P(G</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Distribution of $LR = \frac{P(G</td>
<td>FS)}{P(G</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Representation of the hypotheses $H_0$: “A and B are related as full-siblings” and $H_1$: “A and B are unrelated” - Figure 5(b.).</td>
<td>37</td>
</tr>
</tbody>
</table>
Figure 11 Distribution of \( LR = \frac{P(G|FS)}{P(G|Unr)} \) values (10-log scale, results for A and B simulated as FS and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (b.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line), \{A,B,D\} (dark red line) and \{A,B,E\} (green line) for 17 autosomal markers. x-axis: \( \log_{10}(LR) \); y-axis: density. 39

Figure 12 Distribution of \( LR = \frac{P(G|FS)}{P(G|Unr)} \) values (10-log scale, results for A and B simulated as FS and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (b.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line), \{A,B,D\} (dark red line) and \{A,B,E\} (green line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: \( \log_{10}(LR) \); y-axis: density. Data summarized in Table 7. 40

Figure 13 Representation of the hypotheses \( H_0 \): “A and B are related as half-siblings” and \( H_1 \):“A and B are unrelated” - Figure 5(c.). 42

Figure 14 Distribution of \( LR = \frac{P(G|HS)}{P(G|Unr)} \) values (10-log scale, results for A and B simulated as HS and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (c.) and assuming sets of analyzed individuals: \{A,B\} (blue line) for 35 STRs, \{A,B,C\} (orange line) for 27 STRs, \{A,B,D\} (dark red line) for 22 STRs and \{A,B,E\} (green line) for 22 STRs. x-axis: \( \log_{10}(LR) \); y-axis: density. 43

Figure 15 Distribution of \( LR = \frac{P(G|HS)}{P(G|Unr)} \) values (10-log scale, results for A and B simulated as HS and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (c.) and assuming sets of analyzed individuals: \{A,B\} (blue line) for 35 STRs, \{A,B,C\} (orange line) for 35 STRs, \{A,B,D\} (dark red line) for 32 STRs and \{A,B,E\} (green line) for 27 STRs. x-axis: \( \log_{10}(LR) \); y-axis: density. 45

Figure 16 Distribution of \( LR = \frac{P(G|HS)}{P(G|Unr)} \) values (10-log scale, results for A and B simulated as HS and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (c.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line), \{A,B,D\} (dark red line) and \{A,B,E\} (green line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: \( \log_{10}(LR) \); y-axis: density. Data summarized in Table 8. 46

Figure 17 Representation of the hypotheses \( H_0 \): “A and B are related as avuncular” and \( H_1 \):“A and B are unrelated” - Figure 5(d.). 48
Figure 18  Distribution of $LR = \frac{P(G|AV)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as AV and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (d.) and assuming sets of analyzed individuals: \{A,B,C\} (blue line) for 32 STRs, \{A,B,D\} (orange line) for 32 STRs and \{A,B,E\} (dark red line) for 35 STRs. x-axis: $\log_{10}(LR)$; y-axis: density.

Figure 19  Distribution of $LR = \frac{P(G|AV)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as AV and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (d.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line), \{A,B,D\} (dark red line) and \{A,B,E\} (green line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: $\log_{10}(LR)$; y-axis: density. Data summarized in Table 9.

Figure 20  Representation of the hypotheses $H_0$: “A and B are related as first cousins“ and $H_1$: “A and B are unrelated“ - Figure 5(e.).

Figure 21  Distribution of $LR = \frac{P(G|FC)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as FC and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (e.) and assuming sets of analyzed individuals: \{A,B\} (blue line) for 35 STRs, \{A,B,C\} (orange line) for 27 STRs and \{A,B,D\} (dark red line) for 17 STRs. x-axis: $\log_{10}(LR)$; y-axis: density.

Figure 22  Distribution of $LR = \frac{P(G|FC)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as FC and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (e.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line) and \{A,B,D\} (dark red line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: $\log_{10}(LR)$; y-axis: density. Data summarized in Table 10.

Figure 23  Representation of the hypotheses $H_0$: “A and B are related as half first cousins“ and $H_1$: “A and B are unrelated“ - Figure 5(f.).

Figure 24  Distribution of $LR = \frac{P(G|HC)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as HC and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (f.) and assuming sets of analyzed individuals: \{A,B\} (blue line) for 35 STRs, \{A,B,C\} (orange line) for 35 STRs and \{A,B,D\} (dark red line) for 17 STRs. x-axis: $\log_{10}(LR)$; y-axis: density.
Figure 25  Distribution of $LR = \frac{P(G|HC)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as HC and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (f.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line) and \{A,B,D\} (dark red line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: $log_{10}(LR)$; y-axis: density. Data summarized in Table 11.
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Properties of allele system.</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>Identity by descent (IBD) partitions for some pedigrees (adapted from (Weir et al., 2006)).</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>The seven distinct patterns of genotypes that result in seven algebraic expressions, that are possible for two non-inbred individuals and codominant alleles at one autosomal locus (adapted from (Weir et al., 2006)); ( P_i ) represents the frequency of the ( i ) allele and ( k_s ) the ( s )-th IBD partition associated with the pedigree linking the individuals; “Hom” being homozygous and “Het” being heterozygous.</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>The algebraic expressions for testing some biological relationships among three individuals (adapted from (Fung et al., 2006)). The joint genotype probabilities ( P(X,Y,Z) ), for all possible genetic configurations of ( X, Y ), and ( Z ).</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>The hypotheses to be compare that derive from Figure 5.</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>Maximum error per cases under study after analyzing 35 STR systems.</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>( LR^T ) results obtained for 100 000 simulated families, comparing the hypotheses of the individuals A and B being related as full-siblings (FS) or as half-siblings (HS) - see Figure 5 (a.)- assuming different sets of analyzed individuals: {A,B}, {A,B,C}, {A,B,D} and {A,B,E}.</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>( LR^T ) results obtained for 100 000 simulated families, comparing the hypotheses of the individuals A and B being related as full-siblings (FS) or as Unrelated (Unr) - see Figure 5 (b.)- assuming different sets of analyzed individuals: {A,B}, {A,B,C}, {A,B,D} and {A,B,E}.</td>
<td>41</td>
</tr>
<tr>
<td>8</td>
<td>( LR^T ) results obtained for 100 000 simulated families, comparing the hypotheses of the individuals A and B being related as half-siblings (HS) or as Unrelated (Unr) - see Figure 5 (b.)- assuming different sets of analyzed individuals: {A,B}, {A,B,C}, {A,B,D} and {A,B,E}.</td>
<td>47</td>
</tr>
<tr>
<td>9</td>
<td>( LR^T ) results obtained for 100 000 simulated families, comparing the hypotheses of the individuals A and B being related as avuncular (AV) or as Unrelated (Unr) - see Figure 5 (d.)- assuming different sets of analyzed individuals: {A,B}, {A,B,C}, {A,B,D} and {A,B,E}.</td>
<td>52</td>
</tr>
</tbody>
</table>
Table 10  \( LR^T \) results obtained for 100 000 simulated families, comparing the hypotheses of the individuals A and B being related as first cousins (FC) or as Unrelated (Unr) - see Figure 5 (e.)- assuming different sets of analyzed individuals: \{A,B\}, \{A,B,C\} and \{A,B,D\}.  57

Table 11  \( LR^T \) results obtained for 100 000 simulated families, comparing the hypotheses of the individuals A and B being related as half first cousins (HC) or as Unrelated (Unr) - see Figure 5 (f.)- assuming different sets of analyzed individuals: \{A,B\}, \{A,B,C\} and \{A,B,D\}.  62
ACRONYMS

A adenine.
AV avuncular.

C cytosine.
COTS Commercial off-the-shelf.

DNA Deoxyribonucleic acid.
DVI Disaster victim identification.

FC first cousins.
FS full-siblings.

G guanine.

HC half first cousins.
HS half-siblings.

IBD identity by descent.
IBS identity by state.
IM integer mutations.

LR likelihood ratio.

MVM microvariant mutations.

RFLP Restriction Fragment Length Polymorphism.

SNPs Single Nucleotide Polymorphisms.
SSRs Simple Sequence Repeats.
STR Short Tandem Repeat.

T thymine.

Unr unrelated.
INTRODUCTION

1.1 CONTEXT AND MOTIVATION

Forensic Genetics is the branch of genetics that makes use of the knowledge and techniques of genetics and molecular biology in helping to resolve problems and legal procedures (Jobling and Gill, 2004). Thus, the purpose of a forensic geneticist is to evaluate the likelihood of specific genetic kinships between samples and/or individuals, providing unbiased and statistically supported information (Pinto et al., 2009). Forensic Genetics based on Deoxyribonucleic acid (DNA) started in the 1980s when researchers found highly variable DNA regions. Nowadays, DNA analysis is an indispensable routine part of modern forensic casework (Jobling and Gill, 2004; Wyman and White, 1980). The sources for the biological evidence used in a forensic context are, for example, any body fluid (such as blood, semen, saliva or urine), bones, teeth, hairs, furs, muscle tissue or even touched objects.

Analyses on Forensic Genetics generally fall into one of the following frameworks:

(a.) analysis of mixtures, generally aiming identification tests, where the sample contains DNA from two or more contributors. In most cases, experts have to deal with degraded and/or contaminated samples and with a low amount of DNA. For example, the compatibility between the genetic profile of one suspect and the sample recovered from the crime scene, or in cases of sexual assaults where the material collected has genetic material from the victim and the perpetrator(s). Therefore the main objective is to carry out identification tests;

(b.) kinship analysis, in the majority of the cases aiming paternity tests, where reference samples (with good quantity/quality of DNA) are used. In criminal context, paternity tests can be computed, for example, when rapes result in pregnancy and only then is denounced/investigated (situations commonly involving intellectually disabled women). Other situations can be related with immigration cases (evaluation of the kinship existing between an individual applying for a visa and a legal immigrant already established in the country) or inheritance claims, where the alleged parent can be available for testing or
not. Another example is in identification following disasters - the so-called Disaster victim identification (DVI) problems.

However, (a.) and (b.) can intersect, that is, sometimes it is necessary to compute kinship tests considering samples which are mixtures. An example of this is when there is material of abortion, resulting from a sexual assault, where the profile of the fetus is not distinguishable from the mother (mixture), even having access to the genetic profile of hers (and of the putative father).

Genealogical and pedigree reconstruction has been a subject of high importance, not only for humans, but also for non-human populations. Applications arise, for example, in studies of inbreeding and conservation of species. Indeed, using genetic information to study relationships among human and non-human individuals is a broad topic with a large number of fields of application.

Note that, unlike other forensic areas that rely on the expert’s analysis on an intuitive basis of expertise, forensic genetics has a formal framework based on probability theory and quantification of the proof. Therefore, it is necessary to have knowledge of the genetic characterization of the population by carrying out a quantitative evaluation though population genetics studies. By way of example, if for a specific marker, a suspect has a genetic profile coincident with the one of a sample recovered from a crime scene, the evidence is valued differently if the profile is rare or common in the population. Hence, forensic genetics is a discipline of population genetics.

Within kinship analyses, very powerful statistical results are usually obtained, as the sharing of genetic information is required - in identity (or monozygosity twins) and parent-child analyses. However, there are other pedigrees where the sharing of genetic information is not required as, for example, in case of full(or half)-siblings or avuncular.

Throughout this thesis a study will be performed based on inferring genealogical relationships among those pedigrees, where the sharing of genetic information is not mandatory. Different sets of independent markers routinely used in many laboratories worldwide will be considered for the statistical analyses. The gain of increasing the battery of analyzed autosomal markers and the impact of adding an extra relative (for example, the undoubted mother), will be studied. This analysis will give us guidance on which relatives should be chosen from an initial set of candidates and the number of markers needed to be analyzed in order to maximize the chance of achieving a statistic powerful result. Moreover, we will perform a validation of the software Familias, widely used by forensic laboratories worldwide to compute likelihoods in relationship scenarios.
1.2 Genetic inheritance

Gregor Mendel (1822 – 1884), through his work on pea plants, discovered the fundamental laws of inheritance in 1865 and, because of that, he is known as the father of genetics. Our modern understanding of how traits may be inherited through generations comes from the principles proposed by this monk ((Mendel, 1866), for the English translation: (Mendel, 1901)). See box below for an overview on Mendel’s experience.

Mendel picked common garden pea plants (of the species *Pisum sativum*) for the focus of his research because they can be grown easily in large numbers, have distinct contrasting traits (for example, the color of the seed is either green or yellow) and their reproduction can be manipulated. Pea plants have both male and female reproductive organs. As a result, they can either self-pollinate (pollen could come from the same flower) or cross-pollinate with another plant (pollen could come from another plant’s flowers). So, he could easily control their fertilization by transferring pollen with a small paintbrush.

In his experiments, Mendel observed seven different characteristics in the pea plants, and each of these characteristics had two contrasting forms (see Figure 1). The characteristics included, for example, height (tall or short), pod shape (inflated or constricted), seed shape (smooth or wrinkled) or pea color (green or yellow).

![Figure 1: The seven traits observed by Mendel in pea plants, from Nature Education Adapted from Pierce (2013).](image)

In the first stage, Mendel wanted to obtain pure lineages for the selected characteristics in order to initialize his study. Once the characteristics of the pea plant were consistent generation after generation of self-fertilization (for
example, green plants had only green children and grandchildren and so forth), these parental lines of peas were considered pure-breeders.

When conducting his experiments, Mendel designated the two pure-breeding parental generations involved in a particular cross as P, and the offspring resulting from the crossing of P as F1 generation. Here, Mendel noticed that the F1 generation looked like one parent of the P generation. Upon observing the uniformity of the F1 generation, Mendel wondered whether the F1 generation could still possess the non-dominant traits of the other parent in some hidden way. To understand whether traits were hidden in the F1 generation, Mendel returned to the method of self-fertilization. Here, he created an F2 generation by letting an F1 pea plant self-fertilize.

Figure 2: Mendel’s monohybrid crosses, from Nature Education Adapted from Pierce (2013).

As Mendel suspected, the two traits are observable in the resulting F2 generation. Also, when he averaged the relative proportion of both characteristics across all F2 progeny sets, he found that one trait was consistently three times more frequent than the other (3:1). Figure 2 shows an example of Mendel’s data for the seed form (round or wrinkled), but note that Mendel studied all seven cases and in all he obtained the same conclusions. Thus, the Law of Segregation emerges, also known as Mendels’ first law, stating that every individual organism contains two “particle” for each trait and to a specific trait each “particle” came from each parent.
Mendel had thus determined what happens when two plants that are hybrid for one trait are crossed with each other, but he also wanted to determine what happens when two plants that are each hybrid for two traits are crossed.

Mendel therefore decided to examine the inheritance of two characteristics at once. So, he tested this idea of trait independence with more complex crosses. First, he generated plants that were purebred for two characteristics, such as seed color (yellow and green) and seed shape (round and wrinkled). These plants would serve now as the P generation for the experiment. In this case, Mendel crossed the plants with wrinkled and yellow seeds with plants with round and green seeds.

From his earlier monohybrid crosses, Mendel knew which traits were dominant: round and yellow. So, in the F1 generation, he expected the seeds all round and yellow from crossing these purebred varieties, and that is exactly what he observed. Mendel knew that each of the F1 progeny were dihybrids, in other words, they contained both “particles” for each characteristic. He then recurred to the self-pollination of the F1 generation of the plants, which resulted on the F2 generation, where Mendel observed that in a 9 (yellow, round) : 3 (yellow, wrinkled) : 3 (green, round) : 1 (green, wrinkled) ratio (see figure 3). Moreover, he also concluded that the proportion of each trait was still approximately 3:1 for both seed shape and seed color.
From this data, the principle of independent assortment emerged, also known as Mendel’s Second Law. According to this principle, the way in which “particles” from one specific trait separate and then recombine is unconnected to other traits. Note that, as already stated above, every individual organism contains two “particles” for each trait.

Mendel’s experiments are, indeed, supported by the current knowledge on genetic inheritance based on a parental-filial transmission of a long helical molecule of deoxyribonucleic acid, the already mentioned DNA, that comprises two helical chains containing the instructions that an organism needs to develop, live and reproduce. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). These chemical bases form units when paired (A with T, and C with G), that are called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule originating the nucleotide. These nucleotides are disposed in two long chains, forming a double helix (Watson, 1968).

The majority of human DNA is present in the nucleus of the cell - nuclear DNA - and it is packed in cell structures called chromosomes, existing 23 pairs of them. Of these, 22 pairs are similar in both sexes and are called autosomal chromosomes, which are designated by a number. Note that in each pair of chromosomes, one was paternally and other maternally inherited, keeping their individual characteristics. Moreover, each autosomal chromosome is randomly transmitted to the offspring.

In humans, the 23rd pair of chromosomes comprises the so-called heterosomal chromosomes: X and Y, and it is sex specific. A healthy female has a pair of X chromosomes, while a healthy male has one X and one Y chromosome. The transmission of these chromosomes differs from the autosomal one since it depends on the sex of the individuals. Indeed, a female transmits randomly one X-chromosome to each child she might have (male or female) such as for autosomes, while a male transmits his X chromosome to any daughter he might have and the Y chromosome to any son. Indeed, the familial information carried by the X chromosome is broken whenever there is a father-son link in a pedigree, as well as the Y information in the case of a father-daughter link.

Diploidy describes the great majority of cells in humans that contain two sets of chromosomes. On the other hand, a haploid cell is a cell that contains only one set of chromosomes in it, such as the cells of the germline that will generate gametes: eggs and sperm in females and males, respectively. Furthermore, polyploid cells and organisms are those containing
more than two paired (homologous) sets of chromosomes, the liver cells are an example of this (Duncan et al., 2010). In this work, we will assume diploidy and independent, autosomal transmission.

The position of a gene on a chromosome is designated by *locus* (plural: *loci*). An allele is one specific form of a gene, differing from other alleles on the configuration and number of bases. Note that each gene is responsible for a particular trait and the “particles” referred in Mendel’s experience are what we now know as “alleles”. The genetics term heterozygous represents having dissimilar alleles at corresponding chromosomal loci, whereas homozygous means having identical alleles at corresponding chromosomal loci. An allele that always expresses its phenotypic effect (even in heterozygosity) is said to be a dominant allele (or codominant if there are others with the same property). On the other hand, a recessive allele is an allele in which the phenotypic effect is not expressed in a heterozygote. For a better comprehension of the definitions, we will use the ABO system for blood types as an example. There are three different possible alleles: A, B, and O to determine an individual’s blood type. Of the three alleles, A and B show codominance. This means that a person possessing both A and B alleles as their genotype, has AB blood because both alleles are expressed in the phenotype. Allele O is, however, recessive. Therefore, if an individual A has blood type A (phenotype A), with no further information it is impossible to know if the genotype of the individual is AA (homozygous) or AO (heterozygous).

Another important concept to take into account is the “silent allele”, which formally designates a rare recessive allele. In fact, if a parent is apparently homozygous (i.e. if we can only distinguish one allele in his/her phenotype) and the child too, but for another different allele, the relationship is compatible with the Mendelian rules of transmission, if we assume that the parent has a silent allele in his/her genotype. As an example, suppose that the father has blood group A, but genotype AO, and the child phenotype B, but genotype BO. This genotypic configuration is explained assuming that the father transmitted the allele O to his child, that received the allele B from the mother.

In exceptional cases, it is possible for a parent to not share any allele with the child due to a genetic phenomenon known as mutation. Genetic mutation is characterized by a sudden change in the genome of somatic or germinal cells of an individual. Germline mutation, also known as hereditary mutation, corresponds to any detectable and heritable variation in the lineage of germ cells. Such mutations are susceptible to transmission to offspring (Ajf et al., 2000). For markers generally considered in forensics (Autosomal independent short tandem repeat markers - see section 1.3) mutation rate with order of magnitude $10^{-3}$ is broadly accepted. In terms of genetic variability, we must also take into account another
phenomenon - crossover or crossing over - through which two homologous chromosomes "cut" in the homologous places and exchange one of their segments, so the genes that were at one side and the other of the cut are separated and "dragged" to different gametes.

1.3 Genetic Markers

DNA fingerprinting, also known as DNA profiling, was developed in 1984 by British geneticist Alec Jeffreys while working in the Department of Genetics at the University of Leicester, UK (Jeffreys et al., 1985). Although 99.9% of human DNA sequences are the same in every person, enough of the DNA is different to distinguish one individual from another, with the exception of Identical (monozygotic) twins (Kirby, 1990). Different DNA fingerprinting methods exist, for example: Restriction Fragment Length Polymorphism (RFLP), Single Nucleotide Polymorphisms (SNPs) or, Short Tandem Repeats (STRs) (Roewer, 2013).

As mentioned above, the majority of DNA is identical between individuals. Nevertheless, there are inherited regions of DNA that may vary from individual to individual, which are known as polymorphisms. Indeed, highly polymorphic sequences are those preferably used in Forensic Genetics, since they have a greater statistical power to discriminate between individuals (Ruitberg et al., 2001; Buckleton et al., 2016).

The most commonly used molecular markers in Forensic Genetics are STRs, also known as microsatellites or Simple Sequence Repeats (SSRs). STRs are tandem repetitions of small sequence units (from 1 to 6 nucleotides), varying the number of repeats for each locus, usually from 7 to more than 30 repeat units (Saeed et al., 2015; Edwards et al., 1991).

In this thesis, STRs will be used because these are, consensually, the markers of choice used in forensics and, more importantly, STRs have some of the necessary features to distinguish one DNA sample from another or to establish relations among them, since they are polymorphic and variable across individuals, making them highly informative (Buckleton et al., 2016). Furthermore, it is possible to amplify several loci in one reaction reducing the time taken per analysis, making them genetic markers easy to work with, and still allow the study of degraded DNA samples because they have high chances of being intact after degradation (Buckleton et al., 2016). In this work, as is the standard practice in forensic genetics, we will consider the analysis of autosomal independent markers.
1.4 MUTATION MODELS

It is of obvious relevance to consider the possibility of mutations in the accomplishment of relationship inference studies and thus the probability that as an allele “a” at a marker mutates to an allele “b” (Egeland et al., 2015; Brenner, 2017; Simonsson and Mostad, 2016). In this section, we present some mutation models that are available in Familias - a free software for likelihood calculations in kinship problems based on DNA data (more details see section 2.2) and described in (Egeland et al., 2015).

The generic mutation matrix is presented below, indicating \( n \) the number of possible alleles in a specific marker and \( m_{ij} \) the probability that an allele \( i \) is transmitted as an allele \( j \):

\[
M = \begin{pmatrix}
    m_{11} & \cdots & m_{1n} \\
    \vdots & \ddots & \vdots \\
    m_{n1} & \cdots & m_{nn}
\end{pmatrix}
\]

Notice that the diagonal matrix values are those that do not represent a mutation when transferred over a generation (i.e. the parental allele is transferred to the offspring) (Simonsson and Mostad, 2016).

1.4.1 The equal model

For this model, the probability of not mutating, for each allele, is \( 1 - R \), where \( R \) is the overall mutation rate. The probability of mutating to any of the possible other alleles is the same \( \frac{R}{n-1} \), where \( n \) is the number of “possible” alleles, then the mutation matrix can be written as

\[
M = \begin{pmatrix}
    1 - R & \frac{R}{n-1} & \frac{R}{n-1} & \cdots & \frac{R}{n-1} \\
    \frac{R}{n-1} & 1 - R & \frac{R}{n-1} & \cdots & \frac{R}{n-1} \\
    \frac{R}{n-1} & \frac{R}{n-1} & 1 - R & \cdots & 1 - R \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    \frac{R}{n-1} & \frac{R}{n-1} & \frac{R}{n-1} & \cdots & 1 - R
\end{pmatrix}
\]

To use this mutation model in Familias it is necessary to specify the MutationModel parameter as “equal” (Kling et al., 2014; Simonsson and Mostad, 2016). The equal model is the simplest, although biologically unrealistic for STRs (alleles tend to mutate to neighbours).
1.4.2 The proportional model

The proportional model establishes that the probability of mutating to an allele is proportional to that allele’s frequency, leading that the more frequent alleles are considered as those more prone to mutation. The transition matrix \( M \) for this model is given by:

\[
M = \begin{pmatrix}
1 - k + kp_1 & kp_2 & \ldots & kp_n \\
k p_1 & 1 - k + kp & \ldots & kp_n \\
\vdots & \vdots & \ddots & \vdots \\
k p_1 & \ldots & \ldots & 1 - k + kp_n
\end{pmatrix}
\]

The overall mutation rate becomes \( R = k \sum_{i=1}^{n} p_i (1 - p_i) \), therefore the constant: \( k = \frac{R}{\sum_{i=1}^{n} p_i (1 - p_i)} \).

1.4.3 The stepwise model

In this model it is assumed that the list of alleles is expanded to include all "possible" alleles, and that they are listed by increasing lengths. The probability of mutation varies as a function of the difference in length between the alleles. The matrix \( M \) for this model is given by:

\[
M = \begin{pmatrix}
1 - R & k_1 r^{1-2} & k_1 r^{1-3} & \ldots & k_1 r^{1-n} \\
k_2 r^{2-1} & 1 - R & k_2 r^{2-3} & \ldots & k_2 r^{2-n} \\
k_3 r^{3-1} & k_3 r^{3-2} & 1 - R & \ldots & k_3 r^{3-n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
k_n r^{n-1} & k_n r^{n-2} & k_n r^{n-3} & \ldots & 1 - R
\end{pmatrix}
\]

Where \( r \) is a constant between 0 and 1 and this parameter is provided by the user in Familias as \textit{MutationRange} (Kling et al., 2014; Simonsson and Mostad, 2016). And the normalizing constant:

\[
k_i = \frac{R(1 - r)}{r^{i - 1} - r^{n - i}}
\]

The main limitation of the stepwise model is that, when faced with non-consensual alleles, it orders them in ascending order without analyze if they belong to the same microvariant group (i.e. towards the alleles 13, 14 and 14.2, it considers that the probability of mutation of 14 to neighbour 13 is equal to the mutation of 14 to neighbour 14.2, which is known to be not true).
1.4.4 The extended stepwise model

Mutations are divided into two types: those that add or subtract an integer number of repeats to the allele (e.g. 12 mutates to 13), and those that add or subtract some fractional amount (e.g. 13 mutates to 13.2). Consequently we have integer mutations (IM) and microvariant mutations (MVM), respectively. The rate of these two types of mutations are given separately as $\text{MutationRate}$ and $\text{MutationRate2}$ respectively, to Familias as the first is much more common ($\approx 10^{-3}$) than the second ($\approx 10^{-6}$). Note that the models described until now treat MVM in the same way as IM, which is biologically unrealistic. But the current model separates the overall mutation rate, denoted $\mu$, into two parts, one corresponding to integer mutations (IM), $R$, and one to the microvariant mutations (MVM) $\alpha$, i.e. $\mu = R + \alpha$ (Kling et al., 2014; Simonsson and Mostad, 2016). So the probability of a transition from allele $i$ to allele $j$, $m_{ij}$, has three different alternatives:

- In the case of no mutation occurring: $m_{ij} = 1 - \mu$;
- In the case of integer mutations (IM): $m_{ij} = k_i R r_{i-j}$;
- In the case of microvariant mutations (MVM): $m_{ij} = \frac{\alpha n_i}{n_j}$, where $n_i$ is the number of MVMs from allele $i$.

The extended stepwise model is the closest to biological reality, but there is still a lot of uncertainty (especially in parameters $r$ and $R$) because they are so rare.

1.4.5 An example illustrating how use the mutation models

This example is a paternity case with an alleged father with genotype (11, 12) and a child with genotype (12.1, 13). The frequency of alleles in the population is shown in the table below (Table 1). An overall mutation rate of $R = 0.005$ will be assumed.

<table>
<thead>
<tr>
<th>Alleles</th>
<th>11</th>
<th>12</th>
<th>12.1</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency - $p_i$</td>
<td>0.1</td>
<td>0.19</td>
<td>0.06</td>
<td>0.12</td>
<td>0.25</td>
<td>0.23</td>
<td>0.04</td>
<td>0.01</td>
</tr>
</tbody>
</table>

To simulate this example by Familias, the first steps are: specify the individuals involved (which in this case it will be the alleged father and child), present the sex of the individuals (i.e. male or female) and expose the pedigrees corresponding to the hypotheses under analysis:
Then, give to the software the corresponding DNA data and the frequency of the alleles per marker:

```r
person_id <- c("father","child")
sex <- c("male", "male")
ped_fatherChild <- FamiliasPedigree(id=person_id, dadid = c(NA,"father"), momid = c(NA,NA), sex = sex)
ped_Unrelated <- FamiliasPedigree(id=person_id, dadid = c(NA,NA), momid = c(NA,NA), sex = sex)
mypedigrees <- list(Un = ped_Unrelated, FC = ped_fatherChild)
```

The steps performed so far are common to all models, below is presented the code that will allow the selection of a specific mutation model as well as the specification of the associated parameters.

To apply the equal model (see subsection 1.4.1), which leads to $m_{11;12.1} = m_{11;13} = m_{12;12.1} = m_{12;13} = 0.005/7$, since the mutation rate is 0.005 (MutationRate parameter) and there are 8 possible filial alleles. Note that it is needed to select such option, specifying the respective parameter. Assuming such model and the previous case-example, the software returns the following value:

$$LR = 0.008928571$$

```r
locus <- FamiliasLocus(frequencies = freqalleles, allelenames = namealleles, name = "MarkerExample", MutationModel = "equal", MutationRate = 0.005)
myloci <- list(locus)
result = FamiliasPosterior(mypedigrees, myloci, datamatrix)
```

For the proportional model (see subsection 1.4.2), we have $m_{11;13} = m_{12;13} = kp_{13}$ and $m_{11;12.1} = m_{12;12.1} = kp_{12.1}$. The constant $k$ is equal to

$$k = \frac{R}{\sum_{i=1}^{N} p_i (1-p_i)} = \frac{0.005}{0.8188}$$
1.4. Mutation Models

From the code presented below, where we have defined the proportional mutation model (MutationModel parameter equal to “proportional”) and a mutation rate of 0.005 (MutationRate parameter), the software returns:

\[ LR = 0.006106497 \]

```r
locus <- FamiliasLocus(frequencies = freqalleles, allelenames = namealleles, name = "MarkerExample", MutationModel = "Proportional", MutationRate = 0.005)
myloci <- list(locus)
result = FamiliasPosterior(mypedigrees,myloci,datamatrix)
```

For the stepwise model (see subsection 1.4.3), considering a mutation range \( r = 0.5 \) (parameter MutationRange). The individual mutation probabilities are

\[ m_{11,13} = k_1 r^3, m_{12,13} = k_2 r^2, m_{11,12.1} = k_1 r^2, m_{12,12.1} = k_2 r. \]

The constants \( k_i \) are equal to

\[ k_1 = \frac{R(1 - r)}{r(1 - r^2)} = 0.005, k_2 = \frac{R(1 - r)}{r(2 - r - r^6)} = 0.003. \]

Once again it is required that the parameter MutationModel has the value “stepwise”, which leads to:

\[ LR = 0.006191602 \]

```r
locus <- FamiliasLocus(frequencies = freqalleles, allelenames = namealleles, name = "MarkerExample", MutationModel = "Stepwise", MutationRange = 0.5, MutationRate = 0.005)
myloci <- list(locus)
result = FamiliasPosterior(mypedigrees,myloci,datamatrix)
```

Finally, for the extended stepwise model, assuming the rate of non-integer-step mutations (MutationRate2 parameter) equal to \( 10^{-6} \) and a rate of integer-step mutations of 0.005, it is obtained:

\[ LR = 0.006199935 \]
1.5. Statistical Evaluation

1.5.1 Hypotheses

Before beginning with specific descriptions involved in kinship analysis, it is important to realize that any two individuals in a population are related in the sense that they belong to a finite population and therefore have common ancestors (Weir et al., 2006). So, for kinship analyses, some point in the past must be considered, according to which individuals are assumed to be unrelated.

For the inferential studies on genetic kinship, the approach that will be used consists of evaluating the possibility of two individuals being related by specific alternative pedigrees a priori determined (i.e. established previously to any knowledge on their genetic profiles).

As examples, the probabilities that are needed to be compared can be calculated under the assumptions: “The two genetic profiles correspond to the same donor” (prosecution hypothesis) versus “The two profiles correspond to different, genetically unrelated, donors” (defence hypothesis) in a criminal case; or “The profiles correspond to a pair of individuals genetically related as father/child;” versus “The profiles correspond to a pair of genetically unrelated individuals” in a paternity test.

Note that more than two individuals can be analyzed. So, for cases of trios, an example of the hypotheses of the profiles for the putative father, undoubted mother and offspring we have:”The profiles correspond to individuals related as mother/father/child“ versus “The profiles correspond to a trio where the alleged father is an individual genetically unrelated with the unquestioned duo mother/child”.

Examples of hypotheses that will be analyzed in this paper are: “A and B are related as avuncular” versus “A and B are genetically unrelated” when considering only a pair of individuals, and, “A, daughter of mother C, and B are related as full-siblings” versus “A, daughter of mother C, and B are genetically unrelated“ when considering three individuals in the analysis.
1.5.2 Likelihood Ratio

In this work we will consider several kinship problems where two (exhaustive and mutually exclusive) hypotheses will be compared, as referred in the section above. The method generally accepted to compute the statistical evaluation is based on the calculation of likelihood ratios (LRs), which compare conditional probabilities and expresses how many times one genotypic configuration is more likely than the other, assuming one or other hypothesis.

In order to obtain the probabilities of kinship conditioned to the genotypes of the individuals the Bayes’ Theorem is applied. Given that representing $H_1$ and $H_0$ the alternative hypotheses of kinship (which as already mentioned are independent and mutually exclusive), $G$ is the combination of the genetic types of the individuals analyzed and considering that Bayes’ Theorem is stated mathematically as

$$PosteriorOdds = PriorOdds \times LR$$

or, alternatively,

$$\frac{P(H_1 | G)}{P(H_0 | G)} = \frac{P(H_1)}{P(H_0)} \times \frac{P(G | H_1)}{P(G | H_0)}.$$

Assuming the alternative hypotheses of kinship as a priori equally likely, it results that:

$$\frac{P(H_1 | G)}{P(H_0 | G)} = \frac{0.5}{0.5} \times \frac{P(G | H_1)}{P(G | H_0)} = \frac{P(G | H_1)}{P(G | H_0)}.$$

Therefore, considering the assumptions above:

$$LR = \frac{P(G | H_1)}{P(G | H_0)}.$$

In addition, it should be noted that the standard practice is to consider independent markers. Thus, the final result is obtained by the called ‘product rule’, considering the product of the LR values obtained for each independent system. For example, using the above-mentioned $H_0$ and $H_1$ hypotheses: “the combination of genotypes belongs to a pair of individuals related to half-siblings” versus “the combination of genotypes belongs to a pair of unrelated individuals” for a set of 17 unlinked markers, we will obtain 17 partial LRs (i.e. one for each marker) that will be used to get the final LR, which in turn results from the multiplication of all the partial LRs.
1.5.3 Identity-by-descent versus identity-by-state

Note that when talking about quantification of kinships it is important to distinguish between two key concepts, identity by descent (IBD) and identity by state (IBS). IBD corresponds to a segment of DNA that was inherited from a common ancestor. On the other hand, IBS is the phenomenon where two or more individuals share similar nucleotide sequences, but not necessarily with the same ancestral origin. Note that, unless mutation occurs, IBD implies IBS, but the opposite is not true. For example, if for a given marker a pair father-son shares one, and exactly one, allele we can reasonably assume that the shared allele of the child is IBD relatively to the one of the father. Conversely, if two unrelated individuals share one, and exactly one, allele we can say that the shared alleles are IBS but not IBD, because the individuals do not have, by definition of unrelatedness, common ancestors and the similar alleles do not have the same ancestral origin.

Considering a pair of individuals, the IBD partitions between their four (autosomal) alleles are well established since 1970, through nine coefficients also known as Jacquard coefficients (Jacquard, 1974). For a pair of non-inbred individuals (i.e. individuals whose parents are unrelated) the nine IBD partitions may be reduced to three possibilities for the sharing of genetic information with the same ancestral origin: the individuals share exactly one pair of IBD alleles (with probability $k_1$), two pairs of IBD alleles (probability $k_2$), or none (probability $k_0$). Note that there are only three extreme cases of pedigrees having one and only one non null IBD partition: (a.) parent/child where exactly a pair of alleles is IBD ($k_1 = 1, k_0 = k_2 = 0$); (b.) identity or identical twins where two pairs of alleles are IBD ($k_2 = 1, k_0 = k_1 = 0$); and (c.) unrelated where none allele is IBD ($k_0 = 1, k_2 = k_1 = 0$). For all the other pedigrees the sharing of IBD alleles is possible but not mandatory.

In Table 1 we have represented the values of the IBD partitions for different pedigrees (adapted from (Weir et al., 2006)). Pedigrees are said to belong the same autosomal kinship class if they have the same Jacquard’s coefficients. As an example, half-siblings and avuncular are said to belong to the same kinship class since they have both $k_0 = k_1 = 1/2$ and $k_2 = 0$ (Pinto et al., 2010).
Table 1: IBD partitions for some pedigrees (adapted from (Weir et al., 2006)).

<table>
<thead>
<tr>
<th>Pedigrees</th>
<th>IBD partitions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k_0$</td>
</tr>
<tr>
<td>Unrelated</td>
<td>1</td>
</tr>
<tr>
<td>Identity Identical twins</td>
<td>0</td>
</tr>
<tr>
<td>Parent-child</td>
<td>0</td>
</tr>
<tr>
<td>Full-siblings</td>
<td>$1/4$</td>
</tr>
<tr>
<td>First-cousins</td>
<td>$3/4$</td>
</tr>
<tr>
<td>Half-first cousins</td>
<td>$7/8$</td>
</tr>
<tr>
<td>Avuncular</td>
<td>$1/2$</td>
</tr>
<tr>
<td>Grandparent-grandchild</td>
<td></td>
</tr>
</tbody>
</table>
STATE OF THE ART

2.1 EXACT ALGEBRAIC EXPRESSIONS

This study will be based on mathematical expressions developed for a pair of non-inbred individuals depending on the frequency of the alleles on the population and on the IBD probabilities of each kinship hypothesis. Such algebraic expressions are presented in the Table 2 (adapted from (Weir et al., 2006)), for unlinked markers and for the simplest assumptions (i.e. absence of mutation and silent allele). Very recently, exact algebraic formulas have been published (Egeland et al., 2017) for a pair of non-inbred individuals, allowing the inclusion of mutations and silent allele.

Table 2: The seven distinct patterns of genotypes that result in seven algebraic expressions, that are possible for two non-inbred individuals and codominant alleles at one autosomal locus (adapted from (Weir et al., 2006)); $P_i$ represents the frequency of the $i$ allele and $k_s$ the s-th IBD partition associated with the pedigree linking the individuals; “Hom” being homozygous and “Het” being heterozygous.

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Genotypic state</th>
<th>Probabilities for non-inbred individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_iA_i, A_i, A_i$</td>
<td>Hom/Hom</td>
<td>$k_2P_i^2 + k_1P_i^3 + k_0P_i^4$</td>
</tr>
<tr>
<td>$A_iA_i, A_j, A_j$</td>
<td>Hom/Hom</td>
<td>$k_0P_i^2P_j^2$</td>
</tr>
<tr>
<td>$A_iA_i, A_i, A_j$</td>
<td>Hom/Het</td>
<td>$k_1P_i^2P_j + 2k_0P_i^3P_j$</td>
</tr>
<tr>
<td>$A_iA_i, A_j, A_m$</td>
<td>Hom/Het</td>
<td>$2k_0P_i^2P_jP_m$</td>
</tr>
<tr>
<td>$A_iA_j, A_i, A_j$</td>
<td>Het/Het</td>
<td>$2k_2P_iP_j + k_1P_iP_j(P_i + P_j) + 4k_0P_i^2P_j^2$</td>
</tr>
<tr>
<td>$A_iA_j, A_i, A_m$</td>
<td>Het/Het</td>
<td>$k_1P_iP_jP_m + 4k_0P_i^2P_jP_m$</td>
</tr>
<tr>
<td>$A_iA_j, A_m, A_l$</td>
<td>Het/Het</td>
<td>$4k_0P_iP_jP_mP_l$</td>
</tr>
</tbody>
</table>

Moreover, in Table 3 we present the analytical formulas for testing some biological relationship among three individuals (adapted from (Fung et al., 2006)). So, in the expressions
shown in Table 3 we have all the possible joint genotypic probabilities \( P(X,Y,Z) \) assuming some biological relationship among the three individuals \( X, Y, \) and \( Z \). Individuals \( X \) and \( Z \) are considered as unrelated, \( X \) and \( Y \), as well as \( Y \) and \( Z \), are assumed as either parent-child, uniparentally related or unrelated. Note that two individuals are considered as uniparentally related if they are related either paternally or maternally (as avuncular or half-siblings) and thus they are linked through a pedigree for which \( K_2 = 0 \). On the other hand, biparental pedigree describes individuals related both maternally and paternally (for which \( K_2 \neq 0 \) - see Figure 4. We can examine the biological relationship between \( Y \) and \( Z \) when the biological relationship between \( X \) and \( Y \) are certainly known. For example, in the paternity testing case we can use the mother who will be \( X \) - when there is no doubt that this is the mother, the father will be \( Z \) - where the doubt resides, and the child will be \( Y \) (Fung et al., 2006).

These formulas can not always be used. For example, when testing the hypotheses: "The individuals are related as full-siblings" versus "The individuals are related as half-siblings", or even, "The individuals are related as full-siblings" versus "The individuals
are genetically unrelated”, it is impossible to use them, regardless of the third individual analyzed since the IBD partitions of the full-siblings pedigree, as we can see from Table 1, are \( k_0 = k_2 = 1/4 \) and \( k_1 = 1/2 \), and one of the requirements is \( k_2 = 0 \).

Table 3: The algebraic expressions for testing some biological relationships among three individuals (adapted from (Fung et al., 2006)). The joint genotype probabilities \( P(X,Y,Z) \), for all possible genetic configurations of \( X \), \( Y \), and \( Z \).

<table>
<thead>
<tr>
<th>( A_iA_j )</th>
<th>( A_iA_k )</th>
<th>( A_iA_l )</th>
<th>( A_jA_k )</th>
<th>( A_jA_l )</th>
<th>( A_kA_l )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_0^2k_0^2\frac{p_i^2p_j^2}{p_k^2}P(X)P(Y)P(Z) + 2k_0^3k_0^2\frac{p_i^2p_j^2}{p_k^2}p_lP(X) + 2k_0^2k_0^3\frac{p_i^2p_j^2}{p_k^2}p_lP(Y) + 4k_0^3k_0^2\frac{p_i^2p_j^2}{p_k^2}p_lP(Z) + 4k_0^4k_0^2\frac{p_i^2p_j^2}{p_k^2}p_l^2p_mP(Z) )</td>
<td>( k_0^2k_0^2\frac{p_i^2p_j^2}{p_k^2}P(X)P(Y)P(Z) + 2k_0^3k_0^2\frac{p_i^2p_j^2}{p_k^2}p_lP(X) + 2k_0^2k_0^3\frac{p_i^2p_j^2}{p_k^2}p_lP(Y) + 4k_0^3k_0^2\frac{p_i^2p_j^2}{p_k^2}p_lP(Z) + 4k_0^4k_0^2\frac{p_i^2p_j^2}{p_k^2}p_l^2p_mP(Z) )</td>
<td>( k_0^2k_0^2\frac{p_i^2p_j^2}{p_k^2}P(X)P(Y)P(Z) + 2k_0^3k_0^2\frac{p_i^2p_j^2}{p_k^2}p_lP(X) + 2k_0^2k_0^3\frac{p_i^2p_j^2}{p_k^2}p_lP(Y) + 4k_0^3k_0^2\frac{p_i^2p_j^2}{p_k^2}p_lP(Z) + 4k_0^4k_0^2\frac{p_i^2p_j^2}{p_k^2}p_l^2p_mP(Z) )</td>
<td>( 8k_0^3k_0^2\frac{p_i^2p_j^2}{p_k^2}p_l^2 + 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}(p_l + p_m)p_l^2 + 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}(p_l + p_m)p_l^2 + 8k_0^4k_0^2\frac{p_i^2p_j^2}{p_k^2}p_l^2p_m )</td>
<td>( 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}P(X)P(Y)P(Z) + 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}p_lP(X) + 2k_0^1k_0^1\frac{p_i^2p_j^2}{p_k^2}(p_l + p_m)p_lP(X) + 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}p_lP(Y) + 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}p_lP(Z) )</td>
<td>( 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}p_l^2 + 2k_0^1k_0^1\frac{p_i^2p_j^2}{p_k^2}(p_l + p_m)p_l^2 + 2k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}(p_l + p_m)p_l^2 + 2k_0^1k_0^1\frac{p_i^2p_j^2}{p_k^2}(p_l + p_m)p_l^2 + 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}p_lP(P) + 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}p_lP(Y) + 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}p_lP(Z) + 4k_0^3k_0^1\frac{p_i^2p_j^2}{p_k^2}p_lP(X)P(Y)P(Z) )</td>
</tr>
</tbody>
</table>

The application of formulas is prone to error, so in more complicated cases it is necessary to resort to programs designed for this purpose.

### 2.2 Software

There are several programs used to carry out the statistical calculations, such as Familias, PatPCR, B Gymn Simedic or PatCan. One of the most used worldwide is Familias (Mostad et al., 2012). To illustrate this wide usage of this software we present the statistics of the Intercomparison Program of 2015: “Analysis of DNA polymorphisms in blood stains and other biological samples”, where Familias was the tool used by 69% of the laboratories from the 58 which participated in the advanced level of kinship analyses (Antonio Amorim, 2015). This is an exercise that has been held annually since 1992, organized by The Spanish and Portuguese Speaking Working Group of the International Society for Forensic Genetics (GHEP-ISFG), with the objective of improving the standardization of methods and encouraging a meeting point to discuss the analytical strategies and different methodologies used by different laboratories.
Familias (http://www.familias.name) existed for several years exclusively as a Windows program to calculate probabilities in connection with the use of DNA data to infer family relationships, but it is now possible to make use of the Familias through R. It is a freeware and it is one of the most used softwares in the world for kinship analyses. Indeed, using the line of code "install.packages('Familias')" we can use the functionalities of Familias. It was developed by Petter Mostad, Thore Egeland and Ivar Simonsson and it represents an implementation of an interface to the core Familias functions, which are programmed in C++ (Mostad et al., 2012). Note that, how the command-line runs allows for the analysis of many cases simultaneously, and due to that we was able to lead this work.

Familias is used to compute likelihoods in cases where the DNA profiles of the individuals involved are known, but their kinship is in doubt and comparison between two alternatives is needed. Provided with alternative family constellations (or pedigrees) for the group of people involved, DNA observations for some of the individuals, and a database with allele frequencies in the relevant population, the program calculates several statistics of interest for kinship evaluation, namely the LR previously described. Besides the possibility of inclusion of any pedigree, it allows the incorporation mutation models that are even customizable and silent alleles with specified frequencies for each marker.

2.3 SOFTWARE VALIDATION

While computing the likelihood of all kinship problems linking two and some in case of three individuals (assuming independent transmission, allelic codominance and absence of mutation) can be done resorting to formulae presented in Tables 2 and 3, for more complex cases and/or assumptions, specific software is needed. Indeed, the development of calculations by hand for individual cases becomes, rapidly, a herculean task inadmissibly prone to error. On the other hand, the impact of the likelihood ratio calculating software on the quality of the expert witness report is critical as the wrong or inaccurate calculation or data interpretation, in extreme cases, may lead to false conclusions.

Software validation is the confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use are fulfilled. This means that any software, even the simpler ones, should be documented and properly updated (Drábek, 2009). Particularly in forensics, it is crucial that the computations which lead to a specific statistical result are able to be re-done at any time. Extended documentation and previous versions of the software should then be available (Gjertson et al., 2007).
Commercial off-the-shelf (COTS) software products for the LR calculation are not sufficiently validated, even though this is a well-known statistical technique (Drábek, 2009). Accredited laboratories should validate software related to forensic evidence, since it is this software that lead judges to decide on guilt or innocence or to identify a person or kinship. For these reasons, validation must be done, according to pre-established standards, having the International Society for Forensic Genetics (ISFG) published guidelines on this regard. More information about these recommendations can be found in (Coble et al., 2016).

Software, in comparison with machines or instruments, may contain failures without any warning or, during updates, apparently insignificant changes in the code may result in problems in another code location. In addition, even small programs can be complex due to the different answers that they have to give depending on the different inputs (Drábek, 2009).

Thereby, software can be likened to “black boxes” since the user can neither access to their content nor, most of the times, to alternative means to replicate the computations. Thus, it is necessary to validate the software to make sure that its use provides a correct result and not a result of an error in its implementation. One possible strategy, wherever possible, is to compare the results obtained using exact algebraic formulas and the results obtained via the software (only possible for some cases and simple assumptions) or comparing the results given by different softwares with the same assumptions.

Previously to the establishment of ISFG recommendations (Coble et al., 2016) the authors of Familias considered the software validated by the work (Drábek, 2009), which comprises seven test cases:

- Classical trio: mother-child-alleged father data with all possible allelic paternity situations inputed;
- Deficiency cases: missing mother case;
- Complicated pedigree: low-resolution genotypes for a Japanese cousin case;
- Mutation: trio profiles in cases of paternity inconsistency;
- Silent allele: for mother-child-alleged father;
- Simulation: 100 simulating repeats without null alleles and mutations were performed for sixteen pedigrees based on paternity pedigree;
- Kinship: compare two persons as full-siblings, half-siblings, and unrelated.
OBJECTIVES

In short, and broadly speaking, the objectives outlined for this thesis are:

• For a specific set of kinship problems involving two individuals, where the sharing of genetic information is not required, evaluating the statistical power of the result when: (a.) the number of independent autosomal markers analyzed varies; and/or, (b.) a third individual undoubtedly related to at least one of the two individuals under analysis is considered;

• Validating the software Familias for a set of pedigrees involving only two individuals using exact algebraic formulas and assuming independent autosomal transmission.
4 MATERIAL AND METHODS

4.1 KINSHIP PROBLEMS

The analyzed pedigrees that integrate into the kinship problems are Full-siblings (FS), Half-siblings (HS), Avuncular (AV), Unrelated (Unr), First cousins (FC) and Half first cousins (HC). The addressed kinship problems are represented in Figure 5, and the subsets of individuals to be analyzed for each kinship problem are: \{A, B\}, \{A, B, C\}, \{A, B, D\}, and \{A, B, E\}. Since, for Fig 5(e) and Fig 5(f), there is no “E” individual, only the subsets of individuals: \{A, B\}, \{A, B, C\} and \{A, B, D\} will be considered. Each of the six cases under study has two exhaustive and mutually exclusive hypotheses associated, \textit{a priori} determined, which are represented in Table 4.

Overall, in this thesis, six kinship tests will be carried out with the simulation of 100 000 profiles for each pedigree, where we have the following hypotheses to compare:

- “A and B are related as full-siblings“ \textit{versus} “A and B are related as half-siblings“;
- “A and B are related as full-siblings“ \textit{versus} “A and B are genetically unrelated“;
- “A and B are related as half-siblings“ \textit{versus} “A and B are genetically unrelated“;
- “A and B are related as avuncular“ \textit{versus} “A and B are genetically unrelated“;
- “A and B are related as first cousins“ \textit{versus} “A and B are genetically unrelated“;
- “A and B are related as half first cousins“ \textit{versus} “A and B are genetically unrelated“.
Figure 5: Representation of the six cases under study.

Table 4: The hypotheses to be compare that derive from Figure 5.

<table>
<thead>
<tr>
<th>Figure 5</th>
<th>Hypotheses</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>$H_0$</td>
<td>{A, full sister of D, daughter of mother C, and paternal granddaughter of E, is full sister of B}</td>
</tr>
<tr>
<td></td>
<td>$H_1$</td>
<td>{A, full sister of D, daughter of mother C, and paternal granddaughter of E, is maternal half-sister of B}</td>
</tr>
<tr>
<td>b.</td>
<td>$H_0$</td>
<td>{A, full sister of D, daughter of mother C, and paternal granddaughter of E, is full sister of B}</td>
</tr>
<tr>
<td></td>
<td>$H_1$</td>
<td>{A, full sister of D, daughter of mother C, and paternal granddaughter of E, is unrelated to B}</td>
</tr>
<tr>
<td>c.</td>
<td>$H_0$</td>
<td>{A, full sister of D, daughter of mother C, and paternal granddaughter of E, is full sister of B}</td>
</tr>
<tr>
<td></td>
<td>$H_1$</td>
<td>{A, full sister of D, daughter of mother C, and paternal granddaughter of E, is unrelated to B}</td>
</tr>
<tr>
<td>d.</td>
<td>$H_0$</td>
<td>{A, full sister of D, daughter of mother C, and sister-in-law of E, is aunt of B}</td>
</tr>
<tr>
<td></td>
<td>$H_1$</td>
<td>{A, full sister of D, daughter of mother C, and unrelated to E, is unrelated to B}</td>
</tr>
<tr>
<td>e.</td>
<td>$H_0$</td>
<td>{A, granddaughter of D, daughter of mother C, is first cousin of B}</td>
</tr>
<tr>
<td></td>
<td>$H_1$</td>
<td>{A, granddaughter of D, daughter of mother C, is unrelated to B}</td>
</tr>
<tr>
<td>f.</td>
<td>$H_0$</td>
<td>{A, granddaughter of D, daughter of mother C, is half first cousin of B}</td>
</tr>
<tr>
<td></td>
<td>$H_1$</td>
<td>{A, granddaughter of D, daughter of mother C, is unrelated to B}</td>
</tr>
</tbody>
</table>

4.2 DATABASE

The genotypic configurations were obtained considering a Norwegian database for 35 independent Au-STRs \(^1\) (Dupuy et al., 2013). The population frequency database used comprises the genetic information from 9586 unrelated Norwegians as well as individuals

\(^1\) Complete set of the 35 Au-STRs analyzed: D3S1358, TH01, D21S11, D18S51, PENTA_E, D5S818, D13S317, D7S80, D16S539, CSF1PO, PENTA_D, VWA, D8S1179, TPOX, FGA, D19S433, D2S1338, D10S1248, D1S1656, D22S1045, D2S4414, D12S391, SE33, D7S1517, D3S1744, D2S1360, D6S474, D4S2366, D8S1132, D5S2500, D21S2055, D10S2325, D17S906, APOAI1 and D11S554.
4.3. Simulation of genetic profiles

from three immigrant populations from East Africa, East Asia and Middle Asia in a total of 1531 individuals (Dupuy et al., 2013).

Throughout this study, 5 subsets of autosomal markers will be analyzed. These subsets were formed based on the number of autosomal markers comprised - 17, 22, 27, 32 and 35. The division was performed by order, that is, for example the set of 22 markers consists of the set of markers starting at the 1\textsuperscript{st} marker of the list: D3S1358, and ending at 22\textsuperscript{nd} marker of the list: D12S391. The complete ordered set of 35 Au-STRs is found in footnote\(^1\).

The goal of creating these 5 subsets is the understanding of the correlation between the statistical power of the result with the number of independent markers analyzed (and to obtain guidance for a future theoretical treatment of the problem).

4.3 SIMULATION OF GENETIC PROFILES

The R language (R Core Team) is used by scientists, statisticians and, more recently, by data scientists as a convenient tool for exploratory analysis of interactive data. R is provided with a lot of pre-installed packages, but still allows adding specific packages. Furthermore, it offers a wide variety of statistics and graphical techniques. Therefore, R language will be used to implement the code needed in this work.

The genetic profiles represented in Figure 5 were generated, R language was used to simulate 100,000 of each and the information about the individuals of interest \{A,B,C,D,E\} was stored. This generation of genetic profiles consisted of using a table with the frequencies of the alleles by genetic markers, mentioned in section 4.2.

As an example, the Full-Siblings pedigree was simulated resorting to the generation of a pair of unrelated individuals (the mother and the father) and taking into account the frequencies of the alleles in the population. Resorting to accumulated frequencies and a random number between 0 and 1, alleles with greater frequencies were more likely to be selected as belonging to individuals’ genotypes. Then, the genotype of each of their children was obtained assuming that each of the parents transmitting one or other allele is equally likely. For computations considering a pair of individuals, only the genotypes of the full-siblings were considered.

This work can be divided into two main studies:
• **Validation of Familias for the simplest assumptions - absence of mutation and silent allele:** only the individuals represented in Figure 5 as A and B are analyzed, not considering mutations or silent allele - detailed description in section 4.4.

• **Measurement of the impact of considering a third individual and/or a greater number of markers:** the sets of individuals - \{A,B\}, \{A,B,C\}, \{A,B,D\} and \{A,B,E\} - are analyzed taking into account mutations and silent allele with different sets of analyzed markers - detailed description in section 4.5.

### 4.4 Validation of Familias for the simplest assumptions - absence of mutation and presence of silent allele

The formulas presented in Table 2 were implemented and, additionally, code that allowed the choice of the respective formula was used, since there are seven formulas and its choice depends on whether the individuals are, for each marker, homozygous or heterozygous and on the number of shared alleles. For all the cases under study presented in Figure 5, using the 100 000 profiles generated and the formulas implemented, partial LR tables were obtained for 35 autosomal markers.

In order to make a comparison of the results obtained with the implementation of the formulas present in Table 2 and those obtained with Familias, using the same profiles and the same conditions, we proceed to the respective computations.

Additionally, to validate the results of Familias with those of the formulas already published, we will obtain the maximum errors per case study. So, to obtain the maximum error for each case under study, the two tables of the partial LRs were used, one considering the use of the implemented formulas (Table 2) and other considering the use of the Familias package, computing the difference between both and finding the maximum value.

### 4.5 Measurement of the impact of considering a third individual and/or a greater number of markers

A set of R scripts making use of Familias were created to obtain partial LR tables, and then, by the called ‘product rule’ described in subsection 1.5.2, results for the final LR value were performed considering the several sets of markers \{17, 22, 27, 32 or 35\} and a genetic profile of a third individual \{C, D or E\}, where C, D and E are assumed to undoubted relatives of A or B. The same database (Dupuy et al., 2013) will be used with the frequency of alleles per marker. The pairwise comparisons under study will also be the same (Figure...
Measurement of the impact of considering a third individual and/or a greater number of markers

5), and for each case we simulated 100,000 families. Note that we never give more than three individuals simultaneously as input to Familias, i.e. for each analysis we give to Familias the genotypes of \{A, B\}, \{A, B, C\}, \{A, B, D\} or \{A, B, E\}. So, the addition of a third individual will be studied and the statistical gain of considering greater batteries of markers will also be analyzed.

A silent allele with frequency equal to \(5 \times 10^{-3}\) is assumed and mutations were only considered in cases where Mendelian incompatibilities were verified, which otherwise would lead to \(LR = 0\) or, inversely, to a mathematical indeterminate form. Such mathematical indeterminate form occurred in multiple cases, and due to time and informatic constraints we had to consider different mutations models. No distinction was made depending on sex or age of the individuals.

In the cases where less meiosis were involved, the Extended Stepwise Mutation Model (Simonsson and Mostad, 2016) was considered with rate of integer-step mutations, rate of non-integer-step mutations and mutation range equal to \(10^{-3}\), \(10^{-6}\) and \(10^{-1}\), respectively. So, broadly speaking, one allele with \(i\) repetitive sequences was assume to mutate into one with \(j \neq i\) repetitive sequences with probability \(0.1^{2+|i-j|}\). Examples of cases where this approach has been applied are: Full-siblings versus Half-siblings when considering individuals \{A, B, D\} or \{A, B, E\} shown in Fig 5(a.) or Full-siblings versus Unrelated when considering individuals \{A, B, C\} or \{A, B, D\} shown in Fig 5(b.).

In the kinship testing: Avuncular versus Unrelated for individuals \{A, B, D\} shown in Fig 5(d.), the Equal Mutation Model (Simonsson and Mostad, 2016) was considered, due to the higher number of meiosis and informatics constraints, that prevented the use of other mutation models. Here, it was considered that an allele with \(i\) sequences mutates into one with \(j \neq i\) sequences with probability \(10^{-3}\).

The objective is to infer the gain of considering a greater number of markers and/or a specific third individual. The study will serve as a kind of guide on which relative should be chosen from a initial pool of candidates and/or to gain some insights on the number of markers to consider in order to maximize the chance of obtaining a conclusive statistical result.
RESULTS

In this chapter, an analysis of the results obtained will be presented taking into account what was referred to in the previous chapter (4) Material and Methods. Thus, also in this chapter there will be a division into two sections: (5.1) Validation of Familias for the simplest assumptions - absence of mutation and silent allele; (5.2) Measurement of the impact of considering a third individual and/or a greater number of markers. Section 5.2 will be split in subsections where each of them represents a case study.

5.1 VALIDATION OF FAMILIAS FOR THE SIMPLEST ASSUMPTIONS - ABSENCE OF MUTATION AND SILENT ALLELE

The results of the difference of the values of LR obtained (i.e. the maximum error), for 2x100 000 simulated families and after analyzing 35 Au-STRs, by the use of the algebraic formulas (Table 2) and by the Familias software are presented in Table 5, for duos - individuals \{A,B\} - without considering mutations or silent alleles, regarding all the case studies shown in Figure 5.

<table>
<thead>
<tr>
<th>Maximum Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS vs Unr</td>
</tr>
<tr>
<td>HS vs FS</td>
</tr>
<tr>
<td>HS vs Unr</td>
</tr>
<tr>
<td>AV vs Unr</td>
</tr>
<tr>
<td>FC vs Unr</td>
</tr>
<tr>
<td>HC vs Unr</td>
</tr>
</tbody>
</table>

Small differences were obtained by both methods. Maximum differences range from $10^{-15}$ (for case study “Avuncular versus Unrelated”, “First cousins versus Unrelated” and “Half first cousins versus Unrelated”) to $10^{-9}$ (for case study “Full-Siblings versus Unrelated”).
5.2 Measurement of the impact of considering a third individual and/or a greater number of markers

As previously mentioned (see subsection 1.5.2) we have statistically evaluated the results computing the likelihood ratio \( LR = \frac{P(G|H_i)}{P(G|H_j)} \), for each genotypic configuration \( G \) and kinship hypotheses \( H_i \) and \( H_j \). To facilitate the comparison of results, \( LR^T = LR \) will be presented when individuals were simulated through the relationship \( H_i \) (numerator), and \( LR^T = 1/LR \) is the value presented when individuals were simulated through the relationship \( H_j \) (denominator). Moreover, the proportion of misclassifications represents the proportion of \( LR^T \)'s favoring the alternative hypothesis (or, also known as false hypothesis) of kinship, and not the one assumed to simulate the individuals.

5.2.1 Full-Siblings versus Half-Siblings

As complete set of relatives we considered that three relatives: the mother \( C \) of \( A \), a full-sibling \( D \) of \( A \) and the parental grandmother \( E \) of \( A \), were available for testing - see Figure 6. Here we compare the likelihood of "A and B are related as full-siblings", with that of "A and B are related as half-siblings" considering a different number of markers: \( \{17, 22, 27, 32, 35\} \) and a different set of relatives: \( \{A,B\} \), \( \{A,B,C\} \),\( \{A,B,D\} \) or \( \{A,B,E\} \) in the study - see Figure 9 and Table 6.

![Figure 6: Representation of the hypotheses \( H_0 \): "A and B are related as full-siblings" and \( H_1 \): "A and B are related as half-siblings" - Figure 5(a).](image-url)

The proportion of \( LR^T \)'s favoring the alternative hypothesis, and not the one assumed to simulate the individuals, varied from \( \approx 11\% \) (for individuals \( \{A, B\} \) and 17 STRs) to 0.003\% (for individuals \( \{A, B, D\} \) and 35 STRs), both for the case where \( A \) and \( B \) were simulated as full-siblings. Indeed the proportion of misclassifications revealed to be slightly higher in the case where individuals \( A \) and \( B \) were simulated as full-siblings. As expected, the
proportion of misclassifications diminished when a third individual (see Figure 7) and an increasing number of markers were analyzed, being notable the drop of misclassifications when a full brother D of A was considered. The results obtained resorting to individuals \{A,B\} and 32 STRs is comparable with those obtained for \{A,B,C\} and \{A,B,E\} for 17 STRs. Note also that considering the full-sibling D of A and the minimal set of 17 Au-STRs, the proportion of misclassifications is \(\approx 4.4\) and \(\approx 1.5\) times lower (for A and B simulated as full-siblings and half-siblings, respectively) than considering just a pair of individuals and the maximal set of 35 markers is considered, see Figure 8.

![Figure 7: Distribution of LR = \(\frac{P(G|FS)}{P(G|HS)}\) values (10-log scale, results for A and B simulated as FS and HS in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (a.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line), \{A,B,D\} (dark red line) and \{A,B,E\} (green line) for 17 autosomal markers. x-axis: \(\log_{10}(LR)\); y-axis: density.](image)

Median \(LR^T\) varied from \(\approx 18\) (for \{A,B\} and 17 STRs) to \(3e+18\) (for \{A,B,D\} and 35 STRs), both for the case where A and B were simulated as half-siblings. The powerful results (median \(LR^T\) from to 1.85e+6, 17 STRs, to 3.05e+18, 35 STRs) obtained when individuals \{A,B,D\} were analyzed in the case where individuals A and B were simulated as half-siblings should also be noted. This situation derives from the fact that Mendelian incompatibilities under the assumption of full-sibship (only explainable assuming the occurrence of mutations) can occur. Beyond such case, when A and B were simulated as full-siblings, analyses with 17 STRs and individuals \{A,B,C\} and \{A,B,D\} revealed median
Measurement of the impact of considering a third individual and/or a greater number of markers

$LR_T$s with the same order of magnitude of those obtained for 32 and 35 STRs when only A and B were analyzed. When \{A,B,E\} were considered a median $LR_T$ similar to the one obtained for \{A,B\} and 27 Au-STRs was obtained. On the other hand, when A and B were simulated as half-siblings, median $LR_T$s with the same order of magnitude were obtained when: (a.) 17 Au-STRs and \{A,B,C\} or \{A,B,E\}, and (b.) individuals \{A,B\} and \{27,32,35\} Au-STRs were analyzed.

![Figure 8: Distribution of $LR = \frac{P(G|FS)}{P(G|HS)}$ values (10-log scale, results for A and B simulated as FS and HS in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (a.) and assuming sets of analyzed individuals: \{A,B\} (blue line) for 35 autosomal markers and \{A,B,D\} (orange line) for 17 autosomal markers. x-axis: log$_{10}$(LR); y-axis: density.](image)

The proportion of cases with weak results: $10^{-2} < LR_T < 10^2$, varied from $\approx 81\%$ (when \{A,B\} and 17 STRs were considered) to $\approx 0.16\%$ (for \{A,B,D\} and 35 STRs), when A and B were simulated as half-siblings. The results obtained for, at least, 22 STRs for both \{A,B,C\} and \{A,B,E\} were stronger than those achieved for \{A,B\} and 35 markers. For the set of individuals \{A,B,D\} the results revealed stronger with the minimal set of makers, than those achieved for \{A,B\} and the maximal number of markers.

The proportion of cases for which $LR_T > 10^4$ varied from 0.002% (for \{A,B\} and 17 STRs) to 99.36% (for \{A,B,D\} and 35 STRs), both for when A and B were simulated as half-siblings. In the case where A and B were simulated as full-siblings, the results obtained resorting to
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

\{A,B\} and 27, 32 and 35 markers broadly equated those obtained for \{A,B,E\}, \{A,B,C\} and \{A,B,D\} for 17 markers, respectively. For the case where A and B were simulated as half-siblings, the proportion obtained for \{A,B,C\} and 17 markers, equaled the one obtained for \{A,B\} and 32 STRs, and the results obtained for 17 STRs when were considered the sets of individuals \{A,B,D\} and \{A,B,E\}, overcome (largely in the first case) those obtained with \{A,B\} and 17 STRs. Additionally, the proportion of misclassifications, for which $LR < 10^{-2}$ varied from 0.16\% (for \{A,B\} available and 17 STRs) and 0\% (when \{A,B,D\} and \{27, 32, 35\} STRs were considered), both reached when individuals were simulated as full-siblings.

Therefore, for the battery of markers (and the assumptions) analyzed in this section, the likelihood of an analysis resorting only to individuals A and B, has low statistical impact. Indeed, the proportion of cases for which $LR_T < 10$ (including misclassifications) varied from \(\approx 36\%\) (for 17 STRs) to \(\approx 10\%\) (for 35 STRs). When one undoubted relative of A was analyzed along with A and B, these proportions decreased substantially: (a.) from 12\% to 1\% for \{A,B,C\}; (b.) from 4\% to 0.05\% for \{A,B,D\}; and (c.) from 13\% to 0.09\% for \{A,B,E\}. On the other hand, the results obtained for the minimal set of 17 STRs when one full-sibling D of A also revealed to be so powerful, to say the least, as those obtained for \{A,B\} and 35 markers. In the case of A and B not having the same father, the obtained results considering also D revealed to be incomparably stronger as Mendelian incompatibilities can be found. On the other hand, to reach the statistical relevance obtained through the analysis of the mother C of A and B, or of the paternal grandmother E of A, it was needed to consider, at least, 27 markers.
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

Figure 9: Distribution of $LR = \frac{P(G|FS)}{P(G|HS)}$ values (10-log scale, results for A and B simulated as FS and HS in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (a.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line), \{A,B,D\} (dark red line) and \{A,B,E\} (green line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: $\log_{10}(LR)$; y-axis: density. Data summarized in Table 6.
Table 6: $LR^T$ results obtained for 100,000 simulated families, comparing the hypotheses of the individuals A and B being related as full-siblings (FS) or as half-siblings (HS) - see Figure 5 (a) - assuming different sets of analyzed individuals: (A, B), (A, B, C), (A, B, D) and (A, B, E).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Med 1</td>
<td>1.140</td>
<td>0.910</td>
<td>0.810</td>
<td>1.253</td>
<td>0.759</td>
<td>1.277</td>
<td>0.787</td>
<td>1.263</td>
<td>0.787</td>
<td>1.263</td>
<td>0.787</td>
<td>1.263</td>
</tr>
<tr>
<td>LR &gt; 0.10</td>
<td>0.005</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>#LR &gt; 0.10</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

5.4. Measurement of the impact of considering a third individual and/or a greater number of markers.
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

5.2.2 Full-Siblings versus Unrelated

Assuming the individuals: the mother C of A, a full-sibling D of A and a grandparent E of A that were available for testing - see Figure 10, we consider the comparison of the likelihood of "A and B are related as full-siblings", with "A and B are unrelated" considering a different number of markers: \{17, 22, 27, 32, 35\} in the study - see Figure 12 and Table 7.

---

![Figure 10: Representation of the hypotheses $H_0$: “A and B are related as full-siblings" and $H_1$:“A and B are unrelated” - Figure 5(b.).](image)

In this case the proportion of misclassification varied from 1.25% (when A and B were simulated as full-siblings, for individuals \{A,B\} and 17 STRs) to 0% (for individuals \{A,B,C\} all scenarios except for \{17, 22\} STRs when A and B were simulated as unrelated, and, \{A,B,D\} for \{27, 32, 35\} STRs either A and B were simulated as full-siblings or were simulated as unrelated). When individuals \{A,B\} were analyzed, the proportion of misclassification was slightly higher in the case where A and B individuals were simulated as full-siblings, contrarily, when the set of individuals: \{A,B,C\}, \{A,B,D\} and \{A,B,E\} are analyzed the proportion of misclassifications when A and B were simulated as unrelated are slightly higher or equal. Very low values for the proportion of misclassifications are obtained when the mother C of A and B is analyzed and when the full-sibling D of A and B is analyzed (see Figure 11). Once again, observing the results, the proportion of misclassifications diminished when a third individual and an increasing number of markers were analyzed. Note that considering the genetic profile of a mother C or full-sibling D of A and B is relevant for computations, since for the minimal set of 17 Au-STRs the results are better than considering just the pair of individuals A and B and the maximal set of 35 markers.

Median $LR^T$ varied from $\approx 8e+3$ (for \{A,B\} and 17 STRs) to $\approx 2e+69$ (for \{A,B,C\} and 35 STRs), both for the case where A and B were simulated as unrelated. In general, when individuals \{A,B\} were analyzed, higher $LR^T$'s were obtained for the case where individuals
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

A and B were simulated as full-siblings, contrarily to when individuals \{A,B,C\}, \{A,B,D\} and \{A,B,E\} were considered as Mendelian incompatibilities can be found. For the different sets of markers, the order of magnitude of the median LR: (a.) Considering the set of individuals \{A,B\} varied from $10^3$ to $10^{11}$, (b.) Considering the set of individuals \{A,B,C\} varied from $10^9$ to $10^{69}$, (c.) Considering the set of individuals \{A,B,D\} varied from $10^8$ to $10^{43}$, and, (d.) Considering the set of individuals \{A,B,E\} varied from $10^5$ to $10^{15}$. Observe that in (b.) was where the greatest variation of the median $LR^T$ occurred.

The proportion of cases where $LR^T$ is between $10^{-2}$ and $10^{2}$, varied from $\approx 13\%$ (when A and B were simulated as unrelated, the set of individuals \{A,B\} and 17 STRs were considered) to 0\% (for \{A,B,C\} and \{27,32,35\} STRs, and, for \{A,B,D\} and \{32,35\} STRs). The results obtained for the set of individuals \{A,B,D\} revealed comparable with the minimal set of makers, to those achieved for \{A,B\} with the maximal number of markers. In addition, when analyzing the set of individuals \{A,B,D\} or \{A,B,E\} for at least 22 STRs or in the case of \{A,B,C\} the minimum set of markers, the results prove to be better than those obtained for \{A,B\} and its maximum set of markers - 35 STRs.

The proportion of cases for which $LR^T > 10^4$ varied from 48\% (when A and B were simulated as unrelated, for \{A,B\} and 17 STRs) to 100\% (for \{A,B,C\} and \{32,35\} STRs, and, for \{A,B,D\} and 35 STRs). The proportion obtained for \{A,B,D\} and 17 markers equaled the one obtained for \{A,B\} and 35 STRs, on the other hand, the results for \{A,B,C\} and 17 markers overcome those achieved with \{A,B\} and 35 STRs. Furthermore, the proportion of misclassifications, for which $LR^T < 10^{-2}$ varied from 0.061\% (when A and B were simulated as full-siblings, for \{A,B\} available and 17 STRs) and 0\% (when A and B were simulated as full-siblings, for \{A,B,E\} and 32 STRs were considered, and, for all sets of markers for \{A,B,C\} and \{A,B,D\}, except when A and B were simulated as unrelated and 17 STRs were considered).
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

Overall, for the case considered in this subsection the results obtained are quite strong considering only the set of individuals \{A,B\} and the minimum set of markers - 17 STRs, although it is notorious that when the genetic profile of the mother C of A the results prove to be incredibly powerful. If it is not possible to obtain genetic material from mother C a second good hypothesis is to consider the full-sibling D of A (notice that for 17 STRs the percentage of misclassifications varied from 1.245% to 0.0245% and for 22 STRs varied from 0.509% to 0.003%). As expected when analyzing a greater number of markers these proportions decrease: (a.) varied from 1.245% to 0.036% considering \{A, B\} and for 17 and 35 markers, respectively; (b.) varied from 0.0035% to 0% considering \{A,B,C\} and for 17 and 35 markers, respectively; (c.) varied from 0.0245% to 0% considering \{A,B,D\} and for 17 and 35 markers, respectively; and, (d.) varied from 0.554% to 0.0055% considering \{A,B,E\} and for 17 and 35 markers, respectively.

Figure 11: Distribution of $LR = \frac{P(G|FS)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as FS and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (b.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line), \{A,B,D\} (dark red line) and \{A,B,E\} (green line) for 17 autosomal markers. x-axis: $\log_{10}(LR)$; y-axis: density.
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

Figure 12: Distribution of $LR = \frac{P(G|FS)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as FS and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (b.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line), \{A,B,D\} (dark red line) and \{A,B,E\} (green line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: $\log_{10}(LR)$; y-axis: density. Data summarized in Table 7.
Table 7: LR results obtained for 100,000 simulated families, comparing the hypotheses of the individuals A and B being related as full-siblings (FS) or as Unrelated (Unr) - see Figure 5 (b) - assuming different sets of analyzed individuals: \{A,B\}, \{A,B,C\}, \{A,B,D\} and \{A,B,E\}.

<table>
<thead>
<tr>
<th>Classification</th>
<th>(%\text{ misclassification})</th>
<th>AB</th>
<th>ABC</th>
<th>ABD</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A,B}</td>
<td>(%\text{ misclassification})</td>
<td>0.07</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Median (\text{LR}^T)</td>
<td>1.126 (\times 10^{-6})</td>
<td>1.126 (\times 10^{-6})</td>
<td>1.126 (\times 10^{-6})</td>
<td>1.126 (\times 10^{-6})</td>
</tr>
<tr>
<td>Max (\text{LR}^T)</td>
<td>5.137 (\times 10^{-8})</td>
<td>5.137 (\times 10^{-8})</td>
<td>5.137 (\times 10^{-8})</td>
<td>5.137 (\times 10^{-8})</td>
</tr>
<tr>
<td>(%\text{ misclassification}) (\times 10^{-1})</td>
<td>0.013</td>
<td>0.007</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Median (\text{LR}^T)</td>
<td>1.370 (\times 10^{-5})</td>
<td>1.370 (\times 10^{-5})</td>
<td>1.370 (\times 10^{-5})</td>
<td>1.370 (\times 10^{-5})</td>
</tr>
<tr>
<td>Max (\text{LR}^T)</td>
<td>3.271 (\times 10^{-4})</td>
<td>3.271 (\times 10^{-4})</td>
<td>3.271 (\times 10^{-4})</td>
<td>3.271 (\times 10^{-4})</td>
</tr>
<tr>
<td>(%\text{ misclassification}) (\times 10^{-1})</td>
<td>0.007</td>
<td>0.004</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Median (\text{LR}^T)</td>
<td>2.48 (\times 10^{-1})</td>
<td>2.48 (\times 10^{-1})</td>
<td>2.48 (\times 10^{-1})</td>
<td>2.48 (\times 10^{-1})</td>
</tr>
<tr>
<td>Max (\text{LR}^T)</td>
<td>1.001 (\times 10^{2})</td>
<td>1.001 (\times 10^{2})</td>
<td>1.001 (\times 10^{2})</td>
<td>1.001 (\times 10^{2})</td>
</tr>
<tr>
<td>(%\text{ misclassification}) (\times 10^{-1})</td>
<td>0.002</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

5.2: Measurement of the impact of considering a third individual and/or a greater number of markers.
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

5.2.3 Half-Siblings versus Unrelated

The hypotheses to be compared in this subsection are: \( H_0 : \) “A, full sister of D, daughter of mother C, and paternal granddaughter of E, is paternal half-sister of B”, versus \( H_1 : \) “A, full sister of D, daughter of mother C, and paternal granddaughter of E, is unrelated of B” - see Figure 13. For each of the 100 000 simulated assuming A and B related as HS and, posteriorly as Unr, the statistic \( LR = \frac{P(G|HS)}{P(G|Unr)} \) was computed for 17, 22, 27, 32 and 35 independent markers, and considering different sets of individuals: \{A,B\}, \{A,B,C\}, \{A,B,D\} and \{A,B,E\} - see Figure 16 and Table 8.

![Figure 13: Representation of the hypotheses \( H_0 \): "A and B are related as half-siblings" and \( H_1 : \) "A and B are unrelated" - Figure 5(c).](image)

In this case the proportion of misclassifications varied from \( \approx 12\% \) (for \{A,B\} and 17 STRs) to \( \approx 0.2\% \) (for \{A,B,E\} and 35 STRs), both for the case where A and B were simulated as half-siblings. Observe that, the proportion of misclassifications when \{A,B\} and \{A,B,C\} were analyzed was slightly higher in the case where A and B individuals were simulated as half-siblings, contrarily to when the set of individuals: \{A,B,D\} and \{A,B,E\} and A and B were analyzed, these were slightly lower to the simulated as unrelated. Repeatedly, it is notorious that with the increase of the number of markers and with the consideration of a third individual the proportions of misclassifications decreased, being the drop particularly evident when the individual E was considered. In this case the proportion of misclassifications varied from \( 12\% \) to \( \approx 4\% \) for 17 markers, and, from \( 3\% \) to \( \approx 0.3\% \) for 35 markers, when analyzing \{A,B\} or \{A,B,E\}. The results obtained resorting to individuals \{A,B\} and the maximal set of 35 markers were comparable with those obtained for \{A,B,C\} and 27 STRs, or, \{A,B,D\} and \{A,B,E\} for 22 STRs (see Figure 14).
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

Figure 14: Distribution of $LR = \frac{P(G|H)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as HS and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (c.) and assuming sets of analyzed individuals: \{A,B\} (blue line) for 35 STRs, \{A,B,C\} (orange line) for 27 STRs, \{A,B,D\} (dark red line) for 22 STRs and \{A,B,E\} (green line) for 22 STRs. x-axis: $\log_{10}(LR)$; y-axis: density.

Median $LR^T$ varied from $\approx 16$ (for \{A,B\} and 17 STRs) to $\approx 5e+9$ (for \{A,B,E\} and 35 STRs), both for the case where A and B were simulated as unrelated. Overall, when individuals \{A,B\} and \{A,B,C\} were analyzed, higher $LR^T$s were obtained for the case where individuals A and B were simulated as half-siblings, contrarily to when individuals \{A,B,D\} and \{A,B,E\} were considered. The order of magnitude of the median $LR^T$: (a.) considering the set of individuals \{A,B\} varied from $10^1$ to $10^3$, (b.) considering the set of individuals \{A,B,C\} varied from $10^1$ to $10^5$, (c.) considering the set of individuals \{A,B,D\} varied from $10^2$ to $10^7$, and, (d.) considering the set of individuals \{A,B,E\} varied from $10^2$ to $10^9$. So, in (d.) was where the greatest variation of the median of $LR^T$ occurred, in which the genetic profile of the individual E was considered. Note that in the cases represented by (c.) and (d.) Mendelian incompatibilities were found.

The proportion of cases where $LR^T$ is between $10^{-2}$ and $10^2$, varied from $\approx 80\%$ (for \{A,B\} and 17 STRs) to $\approx 2\%$ (for \{A,B,E\} and 35 STRs), both for the case where A and B were simulated as unrelated. The results achieved for the set of individuals \{A,B,E\} for at least 22 STRs, or, for the set of individuals \{A,B,C\} or \{A,B,D\} for at least 27 STRs, revealed
to be stronger than those achieved for \{A,B\} with the maximal number of markers - 35 STRs.

The proportion of cases for which \( LR^T > 10^4 \) varied from 0.05\% (for \{A,B\} and 17 STRs) to 92\% (for \{A,B,E\} and 35 STRs), both for the case where A and B were simulated as unrelated. The proportion achieved for \{A,B,D\} and 22 markers equaled the one obtained for \{A,B\} and 35 STRs. Besides, the proportion of misclassifications, for which \( LR^T < 10^{-2} \) varied from 0.138\% (for \{A,B\} available and 32 STRs) and 0.011\% (for \{A,B,E\} and 35 STRs), both for the case where A and B were simulated as half-siblings.

Summarily, for the number of markers studied and, for the hypotheses treated in this subsection resorting only to individuals \{A,B\}, the results have low statistical power. Note that the proportion of cases for \( LR^T \) less than 10 ranges from approximately 40.5\% to approximately 10.3\% for 17 and 35 markers, respectively. Additionally, when a third individual were also analyzed these proportions decrease considerably: (a.) considering \{A,B,C\} varied from 24\% to 4\%; (b.) considering \{A,B,D\} varied from 18\% to 2\%; and (c.) considering \{A,B,E\} varied from 15\% to 1\%. In order to reach statistical relevance, it is not enough to analyze only the individuals \{A,B\}, so it is essential to consider the genetic profile of a third individual. Even considering a third individual the minimum number of markers is not enough, so depending on the third individual considered, a specific number of markers is also required. Thus, for \{A,B,C\} it is recommended to use 35 markers, for \{A,B,D\} at least 32 markers and, finally, for \{A,B,E\} at least 27 markers (see Figure 15). Hence, the most advantageous individual to considered for this analysis is the individual E since it is the one that seems to require the lowest number of markers to obtain powerful results.
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

When comparing the results obtained with those presented in the case analyzed previously (subsection 5.2.2. Full-siblings versus Unrelated), the achieved results in the previous subsection were, as expected, considerably more powerful. The results achieved for the proportion of misclassifications resorting to individuals \{A,B\} and 17 STRs for the case Full-siblings versus Unrelated, were comparable with those obtained for \{A,B,C\} and 35 STRs, \{A,B,D\} and 32 STRs or \{A,B,E\} and 27 STRs for the case Half-siblings versus Unrelated.
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

Figure 16: Distribution of $LR = \frac{P(G|HS)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as HS and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (c.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line), \{A,B,D\} (dark red line) and \{A,B,E\} (green line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: $\log_{10}(LR)$; y-axis: density. Data summarized in Table 8.
Table 8: \( LR^T \) results obtained for 100,000 simulated families, comparing the hypotheses of the individuals A and B being related as half-siblings (HS) or as Unrelated (Unr) - see Figure 5 (b) - assuming different sets of analyzed individuals: \{A,B\}, \{A,B,C\}, \{A,B,D\} and \{A,B,E\}.

<table>
<thead>
<tr>
<th>Results</th>
<th>{A,B}</th>
<th>{A,B,C}</th>
<th>{A,B,D}</th>
<th>{A,B,E}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min ( LR^T )</td>
<td>1.165</td>
<td>1.197</td>
<td>1.243</td>
<td>1.302</td>
</tr>
<tr>
<td>Max ( LR^T )</td>
<td>6.570</td>
<td>6.815</td>
<td>7.074</td>
<td>7.414</td>
</tr>
<tr>
<td>Min %misclassification</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Max %misclassification</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Min ( LR^T ) &gt; 10^3</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Max ( LR^T ) &gt; 10^3</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Min %misclassification &gt; 10^3</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Max %misclassification &gt; 10^3</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Min ( LR^T ) &lt; 10^-2</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Max ( LR^T ) &lt; 10^-2</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>ABC</td>
<td>4.513</td>
<td>4.578</td>
<td>4.648</td>
<td>4.724</td>
</tr>
<tr>
<td>ABD</td>
<td>4.075</td>
<td>4.143</td>
<td>4.211</td>
<td>4.288</td>
</tr>
<tr>
<td>ABE</td>
<td>4.513</td>
<td>4.578</td>
<td>4.648</td>
<td>4.724</td>
</tr>
<tr>
<td>ABCD</td>
<td>4.075</td>
<td>4.143</td>
<td>4.211</td>
<td>4.288</td>
</tr>
<tr>
<td>ABED</td>
<td>4.513</td>
<td>4.578</td>
<td>4.648</td>
<td>4.724</td>
</tr>
<tr>
<td>ABEE</td>
<td>4.075</td>
<td>4.143</td>
<td>4.211</td>
<td>4.288</td>
</tr>
<tr>
<td>ABCDE</td>
<td>4.513</td>
<td>4.578</td>
<td>4.648</td>
<td>4.724</td>
</tr>
</tbody>
</table>

52. Measurement of the impact of including a third individual and/or a greater number of markers.

47
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

5.2.4 Avuncular versus Unrelated

The hypotheses analyzed in this subsection are: $H_0$: “A, full sister of D, daughter of mother C, and sister-in-law of E, is aunt of B”, versus $H_1$: “A, full sister of D, daughter of mother C, and unrelated of E, is unrelated of B” - see Figure 17. Thus, for each of the 100 000 simulated profiles assuming A and B related as AV and, posteriorly as Unr, the statistic $LR = \frac{P(G|AV)}{P(G|Unr)}$ was computed for 17, 22, 27, 32 and 35 STRs, considering different sets of individuals: \{A,B\}, \{A,B,C\}, \{A,B,D\} and \{A,B,E\} - see Figure 19 and Table 9.

![Figure 17: Representation of the hypotheses $H_0$: "A and B are related as avuncular" and $H_1$:"A and B are unrelated" - Figure 5(d).](image)

The proportion of misclassifications varied from $\approx 11.84\%$ (for individuals \{A,B\} and 17 STRs) to $\approx 0.4\%$ (for individuals \{A,B,D\} and 35 STRS), both for the case where A and B were simulated as avuncular. When individuals \{A,B\} and \{A,B,E\} were analyzed, the proportion of misclassifications was slightly higher in the case where A and B individuals were simulated as avuncular, contrarily to when the set of individuals \{A,B,D\} were analyzed the proportion of misclassifications when A and B were simulated as unrelated are slightly higher. But when we analyze the set of individuals \{A,B,C\} we find a mixed situation, for the set of markers \{17, 22, 32\} the proportions are slightly higher when A and B are simulated as unrelated, whereas for the set of markers \{27, 35\} the values of the proportions are slightly lower when A and B are simulated as unrelated. The most powerful results among the range of individuals available for the study were obtained when considering the genetic profile of the full-sibling of A (i.e. the set of individuals \{A,B,D\}), in which whatever the number of markers used, was what obtained the lowest proportion of misclassifications. Notice that, once again, the proportion of misclassifications reduced when a third individual and an increasing number of independent markers were analyzed.
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

Median \( LR^T \) varied from \( \approx 17 \) (for \( \{A,B\} \) and 17 STRs) to \( \approx 1.5e+7 \) (for \( \{A,B,D\} \) and 35 STRs), both for the case where A and B were simulated as unrelated. Being that, when individuals \( \{A,B\} \) and \( \{A,B,E\} \) were analyzed, powerful \( LR^T \)s were obtained for the case where individuals A and B were simulated as avuncular, contrarily to when individuals \( \{A,B,C\} \) and \( \{A,B,D\} \) were considered. The values of median \( LR^T \) with the same order of magnitude were obtained for \( \{A,B\} \) and 35 markers and \( \{A,B,E\} \) and 27 markers, other case is, for \( \{A,B,C\} \) and \( \{A,B,D\} \) both for 22 markers.

The proportion of cases where \( LR^T \) is between \( 10^{-2} \) and \( 10^2 \), varied from \( \approx 79\% \) (for \( \{A,B\} \) and 17 STRs) to \( 4\% \) (for \( \{A,B,D\} \) and 35 STRs), both for the case where A and B were simulated as unrelated. The results obtained for the set of individuals \( \{A,B,D\} \) and 22 STRs or, for the set of individuals \( \{A,B,E\} \) and 27 STRs, revealed comparable to those achieved for \( \{A,B\} \) with the maximal number of markers - 35 STRs. Moreover, when analyzing the set of individuals \( \{A,B,C\} \) or \( \{A,B,D\} \) for at least 27 STRs or in the case of \( \{A,B,E\} \) for at least 32 STRs, the results prove to be more powerful than those obtained for \( \{A,B\} \) and the maximum set of markers.

The proportion of cases for which \( LR^T \) > \( 10^4 \) varied from 0.006\% (when A and B were simulated as unrelated, for \( \{A,B\} \) and 17 STRs) to 85\% (when A and B were simulated as unrelated, for \( \{A,B,D\} \) and 35 STRs). Note that the latter proportion for individuals \( \{A,B,D\} \) and 35 markers in the case of A and B are simulated as Avuncular was twice the obtained for the maximum set of markers considering only \( \{A,B\} \) individuals, whereas when A and B were simulated as unrelated it was almost four times higher, these values can be explained by the Mendelian incompatibilities found. Additionally, the proportion of misclassifications, for which \( LR^T \) < \( 10^{-2} \) varied from 0.009\% (when A and B were simulated as avuncular, for \( \{A,B,D\} \) available and 35 STRs) to 0.146\% (when A and B were simulated as avuncular, for the set of individuals \( \{A,B,E\} \) and 22 STRs).

Briefly, to reach the statistical relevance, the analysis resorting only to individuals A and B is not enough, even considering the maximal set of markers. In fact, the proportion of cases for which \( LR^T \) < 10 varied from 40\% (for 17 STRs) to 10\% (for 35 STRs). When considering the genetic profile of a third individual these proportions diminished: (a.) from 19\% to 2\% for \( \{A,B,C\} \) analyzed; (b.) from 18\% to 1\% for \( \{A,B,D\} \) analyzed; and (c.) from 24\% to 4\% for \( \{A,B,E\} \) analyzed. Thereby, for the hypotheses treated in this analysis the sets of alternatives in order to obtain a reasonable probability of reaching a statistically powerful result are: (I) \( \{A,B,C\} \) for at least 32 markers; (II) \( \{A,B,D\} \) for at least 32 markers; and, (III) \( \{A,B,E\} \) for 35 markers (see Figure 18). It was found that the alternative that obtains the best result is to consider the full-sibling of A (i.e. \( \{A,B,D\} \)) and 35 independent
marks. However, individual D is the most advantageous to consider for this analysis, although individual C also has close results.

Figure 18: Distribution of $LR = \frac{P(G|AV)}{P(G|Un)}$ values (10-log scale, results for A and B simulated as AV and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (d.) and assuming sets of analyzed individuals: \{A,B,C\} (blue line) for 32 STRs, \{A,B,D\} (orange line) for 32 STRs and \{A,B,E\} (dark red line) for 35 STRs. x-axis: $\log_{10}(LR)$; y-axis: density.
Figure 19: Distribution of $LR = \frac{P(G|AV)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as AV and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (d.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line), \{A,B,D\} (dark red line) and \{A,B,E\} (green line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: $\log_{10}(LR)$; y-axis: density. Data summarized in Table 9.
Table 9: LR results obtained for 100 000 simulated families, comparing the hypotheses of the individuals A and B as being related as avuncular (AV) or as Unrelated (Unr) - see Figure 5 (d.) - assuming different sets of analyzed individuals: {A,B}, {A,B,C}, {A,B,D} and {A,B,E}.

<table>
<thead>
<tr>
<th>ABC</th>
<th>A/B Unr</th>
<th>A/B AV</th>
<th>A/B Unr</th>
<th>A/B AV</th>
<th>A/B Unr</th>
<th>A/B AV</th>
<th>A/B Unr</th>
<th>A/B AV</th>
<th>A/B Unr</th>
<th>A/B AV</th>
<th>A/B Unr</th>
<th>A/B AV</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.54</td>
<td>5.38</td>
<td>5.78</td>
<td>5.24</td>
<td>3.71</td>
<td>4.08</td>
<td>4.04</td>
<td>3.31</td>
<td>4.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.2752105 × 10^3</td>
<td>4.986076 × 10^3</td>
<td>3.918065 × 10^3</td>
<td>2.533126 × 10^3</td>
<td>1.918959 × 10^3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.532531 × 10^3</td>
<td>1.948937 × 10^3</td>
<td>1.648932 × 10^3</td>
<td>1.269425 × 10^3</td>
<td>1.069248 × 10^3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.41</td>
<td>5.38</td>
<td>5.78</td>
<td>5.24</td>
<td>3.71</td>
<td>4.08</td>
<td>4.04</td>
<td>3.31</td>
<td>4.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.2752105 × 10^3</td>
<td>4.986076 × 10^3</td>
<td>3.918065 × 10^3</td>
<td>2.533126 × 10^3</td>
<td>1.918959 × 10^3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

... (Table continues)
5.2.5 First Cousins versus Unrelated

The results for the hypotheses: \( H_0 \): "A, granddaughter of D, daughter of mother C, is first cousin of B", versus \( H_1 \): "A, granddaughter of D, daughter of mother C, is unrelated of B" - see Figure 20 - for the different numbers of markers \{17, 22, 27, 32, 35\} and for the different sets of individuals: \{A,B\}, \{A,B,C\} or \{A,B,D\}, are presented and analyzed in this subsection - see Figure 22 and Table 10.

![Figure 20: Representation of the hypotheses \( H_0 \): "A and B are related as first cousins" and \( H_1 \): "A and B are unrelated" - Figure 5(e).](image)

The proportion of misclassifications varied from \( \approx 28.5\% \) (for individuals \{A,B\} and 17 STRs) to \( \approx 1.9\% \) (for individuals \{A,B,D\} and 35 STRS), both for the case where A and B were simulated as first cousins. Observing and comparing these proportions, when the grandmother D of A was considered and A and B were simulated as unrelated the values were slightly higher than those obtained when A and B were simulated as first cousins, which does not happened with \{A,B\} and \{A,B,C\} wherein here slightly higher values were obtained when A and B were simulated as first cousins. An improvement in the results was still observable when using a larger number of markers and with the introduction of a third individual in the analysis. The lowest values for the misclassification ratios were reached when the genetic profile of the grandmother D of A was considered, noting that even considering the minimum set of markers - 17 STRs - less values were obtained than when considering \{A,B\} and the maximum number of markers - 35 STRs. It should also be noted that when analyzing \{A,B,C\} 27 STRs the results were shown to be equivalent to those obtained for \{A,B\} 35 STRs, see Figure 21.

Median \( LRT \) varied from \( \approx 2 \) (for \{A,B\} and 17 STRs) to \( \approx 6821 \) (for \{A,B,D\} and 35 STRs), both for the case where A and B were simulated as first cousins. It was verified that the values of \( LRT \) obtained were low and do not vary much in the order of magnitude and
the higher medians were found in the case where the genetic profile of the grandmother D of A was considered, ranging from $10^1$ to $10^3$.

Figure 21: Distribution of $LR = \frac{P(G|FC)}{P(G|Un)}$ values (10-log scale, results for A and B simulated as FC and Un in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (e.) and assuming sets of analyzed individuals: \{A,B\} (blue line) for 35 STRs, \{A,B,C\} (orange line) for 27 STRs and \{A,B,D\} (dark red line) for 17 STRs. x-axis: \(\log_{10}(LR)\); y-axis: density.

The proportion of cases where $LR^T$ is between $10^{-2}$ and $10^2$, varied from 99.993% (when A and B were simulated as unrelated, for \{A,B\} and 17 STRs) to 16% (when A and B were simulated as first cousins, for \{A,B,D\} and 35 STRs). The results obtained for the set of individuals \{A,B,C\} and 27 STRs revealed comparable to those achieved for \{A,B\} with the maximal number of markers - 35 STRs. Furthermore, when analyzing the set of individuals \{A,B,C\} for at least 32 STRs or in the case of \{A,B,D\} the minimal number of markers, the results prove to be more powerful than those achieved for \{A,B\} and its maximum set of markers.

The proportion of cases for which $LR^T > 10^4$ varied from 0% (when A and B were simulated as unrelated, for \{A,B\} and \{A,B,C\} for all sets of markers) to 46% (when A and B were simulated as first cousins, for \{A,B,C\} and 35 STRs). When considering the sets of individuals: \{A,B\} and \{A,B,C\}, the maximum proportion was $\approx 4\%$ (when A and B were simulated as first cousins, for \{A,B,C\} and 35 STRs). Additionally, the proportion of
misclassifications, for which $LR_T < 10^{-2}$ varied from 0% (when A and B were simulated as first cousins, for \{A,B\} and \{17,22\} STRs) to 0.12% (when A and B were simulated as unrelated, for the set of individuals \{A,B,D\} and 17 STRs).

In short, to reach the statistical power it is crucial to analyze in addition of the genetic profile of the individuals A and B, the grandmother D of A and consider the maximal set of markers - 35 STRs. Effectively, when considering the genetic profile of a third individual and a greater number of markers the statistical power improves. Note that for the proportion of cases for which $LR_T < 10$ varied from 89% (for 17 STRs) to 58% (for 35 STRs) when considering the set \{A,B\} of individuals, from 76% to 39% for \{A,B,C\} individuals analyzed and from 33% to 7% for \{A,B,D\} individuals analyzed.
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

Figure 22: Distribution of $LR = \frac{P(G|FC)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as FC and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (e.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line) and \{A,B,D\} (dark red line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: $\log_{10}(LR)$; y-axis: density. Data summarized in Table 10.
Table 10: \( LRT \) results obtained for 100,000 simulated families, comparing the hypotheses of the individuals A and B being related as first cousins (FC) or as Unrelated (Unr) - see Figure 5 (e)- assuming different sets of analyzed individuals: \( \{A,B\}, \{A,B,C\} \) and \( \{A,B,D\} \).

<table>
<thead>
<tr>
<th></th>
<th>17 Au-STRs</th>
<th>22 Au-STRs</th>
<th>27 Au-STRs</th>
<th>32 Au-STRs</th>
<th>35 Au-STRs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A&amp;B:FC</td>
<td>A&amp;B:Unr</td>
<td>A&amp;B:FC</td>
<td>A&amp;B:Unr</td>
<td>A&amp;B:FC</td>
</tr>
<tr>
<td></td>
<td>% misclassification</td>
<td></td>
<td>% misclassification</td>
<td></td>
<td>% misclassification</td>
</tr>
<tr>
<td>Mean LRT</td>
<td>2.07959e-00</td>
<td>2.07959e-00</td>
<td>2.07959e-00</td>
<td>2.07959e-00</td>
<td>2.07959e-00</td>
</tr>
<tr>
<td>Max LRT</td>
<td>3.0145e-00</td>
<td>3.0145e-00</td>
<td>3.0145e-00</td>
<td>3.0145e-00</td>
<td>3.0145e-00</td>
</tr>
<tr>
<td>% LRT &gt; 10^6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &gt; 10^5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &gt; 10^4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &gt; 10^3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &gt; 10^2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &gt; 10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &gt; 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &lt; 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &lt; 10^-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &lt; 10^-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &lt; 10^-3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &lt; 10^-4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &lt; 10^-5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% LRT &lt; 10^-6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

5.2 Measurement of the impact of considering a third individual and/or a greater number of markers

57
5.2.6 *Half First Cousins versus Unrelated*

The results for the hypotheses: $H_0$: "A, granddaughter of D, daughter of mother C, is half first cousin of B", versus $H_1$: "A, granddaughter of D, daughter of mother C, is unrelated of B" - see Figure 23 - for the different numbers of markers $\{17, 22, 27, 32, 35\}$ and for the different sets of individuals: $\{A,B\}$, $\{A,B,C\}$ or $\{A,B,D\}$, are presented and analyzed in this subsection - see Figure 25 and Table 11.

![Figure 23: Representation of the hypotheses $H_0$: "A and B are related as half first cousins" and $H_1$: "A and B are unrelated" - Figure 5(f).](image)

The proportion of misclassifications varied from $\approx 41\%$ (for individuals $\{A,B\}$ and 17 STRs, when A and B were simulated as half first cousins) to $\approx 2.96\%$ (for individuals $\{A,B,D\}$ and 35 STRs, when A and B were simulated as unrelated). Observing and comparing these proportions, when A and B were simulated as half first cousins, the values were slightly higher than those obtained when A and B were simulated as unrelated. Note that percentages of equivalent misclassifications were reached when analyzed $\{A,B\}$ and 35 STRs or $\{A,B,C\}$ and 27 STRs. The lowest values for the misclassification ratios were achieved when the genetic profile of the grandmother D of A was considered. In addition, when analyzing the set of individuals $\{A,B,D\}$ with the lowest number of markers - 17 STRs - the results for misclassifications were approximately three times lower when compared to the set of individuals $\{A,B\}$ with the maximum number of markers - 35 STRs. An improvement was still seen in the results when using a larger number of markers and with the introduction of a third individual in the analysis.

Median $LR^T$ varied from $\approx 1$ (for $\{A,B\}$ and 17 STRs) to $\approx 2026$ (for $\{A,B,D\}$ and 35 STRs), both for the case where A and B were simulated as half first cousins. As observed in the previous case study, the values of $LR^T$ were low and do not vary much in the order
of magnitude. The higher medians are found when the set of individuals \{A,B,D\} were analyzed, ranging from $10^1$ to $10^3$.

The proportion of cases where $LR^T$ is between $10^{-2}$ and $10^2$, varied from 100\% (when A and B were simulated as unrelated, for \{A,B\} and 17 STRs) to $\approx 23\%$ (when A and B were simulated as half first cousins, for \{A,B,D\} and 35 STRs). For the set of individuals \{A,B\} and \{A,B,C\}, the proportion of cases where $LR^T$ is between $10^{-2}$ and $10^2$ ranges from 99\% to 100\%, that is, high percentages and slight variation. A greater variation was found when individuals \{A,B,D\} were analyzed, from 23\% (35 STRs and A and B simulated as unrelated) to 80\% (17 STRs and A and B simulated as half first cousins).

The proportion of cases for which $LR^T > 10^4$ varied from 0\% (for \{A,B\} and \{A,B,C\} for all sets of markers and when A and B were simulated as unrelated) to 36\% (when A and B were simulated as half first cousins, for \{A,B,D\} and 35 STRs). When considering the sets of individuals: \{A,B\} and \{A,B,C\}, the maximum proportion was $\approx 0.05\%$ (when A and B were simulated as first cousins, for \{A,B,C\} and 35 STRs). Moreover, the proportion of misclassifications, for which $LR^T < 10^{-2}$ varied from 0\% (when A and B were simulated as half first cousins, for \{A,B\} and \{A,B,C\} for all sets of markers, and when A and B were simulated as unrelated for \{A,B\} and 17 STRs) to 0.13\% (when A and B were simulated as half first cousins, for the set of individuals \{A,B,D\} and 22 STRs).

Therefore, considering the pairwise comparisons under study in this subsection, the analysis resorting only to the set of individuals \{A,B\}, even using the maximum number of autosomal markers, is not sufficient to reach statistical relevance (the ratio of misclassifications was approximately 30\%). The same happens when, in addition to individuals A and B, the genetic profile of mother C of A was analyzed (where the percentage of misclassifications for the maximum number of markers was approximately 25\%). There is no evident advantage in considering the genetic profile of the mother C of A as weak results were still being reached. The same does not apply to the analysis of grandfather D of A in which, even for the minimum number of markers there was a significant reduction in the proportion of misclassifications ($\approx 11.5\%$), see Figure 24. But note that the best possible scenario is the analysis of the set of individuals \{A,B,D\} and the maximum number of autosomal markers - 35 STRs (proportion of misclassifications $\approx 3\%$).
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

Figure 24: Distribution of \( LR = \frac{P(G|HC)}{P(G|Un)} \) values (10-log scale, results for A and B simulated as HC and Un in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (f.) and assuming sets of analyzed individuals: \{A,B\} (blue line) for 35 STRs, \{A,B,C\} (orange line) for 35 STRs and \{A,B,D\} (dark red line) for 17 STRs. x-axis: \( \log_{10}(LR) \); y-axis: density.

When comparing the results obtained with those presented in the case analyzed previously (subsection 5.2.5. First Cousins versus Unrelated), the achieved results in the previous subsection were more powerful. Although we are analyzing relationships involving cousins we get different results, which give us the information that it is fundamental to provide more detail when specifying the relationship to be analyzed.
5.2. Measurement of the impact of considering a third individual and/or a greater number of markers

Figure 25: Distribution of $LR = \frac{P(G|HC)}{P(G|Unr)}$ values (10-log scale, results for A and B simulated as HC and Unr in full line and dashed line, respectively) obtained when considering the hypotheses depicted at Figure 5 (f.) and assuming sets of analyzed individuals: \{A,B\} (blue line), \{A,B,C\} (orange line) and \{A,B,D\} (dark red line), for different numbers of markers: 17, 22, 27, 32 and 35. x-axis: $\log_{10}(LR)$; y-axis: density. Data summarized in Table 11.
Table 11: \( L^T \) results obtained for 100 000 simulated families, comparing the hypotheses of the individuals A and B being related as half first cousins (HC) or as Unrelated (Unr) - see Figure 5 (f)- assuming different sets of analyzed individuals: \( \{A,B\} \), \( \{A,B,C\} \) and \( \{A,B,D\} \).

<table>
<thead>
<tr>
<th></th>
<th>17 Au-STRs</th>
<th>22 Au-STRs</th>
<th>27 Au-STRs</th>
<th>32 Au-STRs</th>
<th>37 Au-STRs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{HC} )</td>
<td>( \text{Unr} )</td>
<td>( \text{HC} )</td>
<td>( \text{Unr} )</td>
<td>( \text{HC} )</td>
</tr>
<tr>
<td>% misclassification</td>
<td>0.067</td>
<td>0.069</td>
<td>0.066</td>
<td>0.065</td>
<td>0.064</td>
</tr>
<tr>
<td>Median ( L^T )</td>
<td>1.165±0.100</td>
<td>1.271±0.120</td>
<td>1.239±0.100</td>
<td>1.359±0.100</td>
<td>1.366±0.100</td>
</tr>
<tr>
<td>Min ( L^T )</td>
<td>0.266±0.010</td>
<td>0.513±0.010</td>
<td>0.810±0.005</td>
<td>0.936±0.005</td>
<td>0.952±0.005</td>
</tr>
<tr>
<td>Max ( L^T )</td>
<td>1.84±0.055</td>
<td>2.70±0.025</td>
<td>2.22±0.015</td>
<td>2.77±0.015</td>
<td>2.79±0.015</td>
</tr>
<tr>
<td>%( L^T &gt; 10^6 )</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>%( L^T &gt; 10^4 )</td>
<td>0.01</td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>%( L^T &gt; 10^3 )</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>%( L^T &gt; 10^2 )</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>%( \text{A&amp;B:HC} )</td>
<td>1.52±0.004</td>
<td>2.46±0.004</td>
<td>5.65±0.001</td>
<td>1.45±0.001</td>
<td>0.88±0.001</td>
</tr>
<tr>
<td>%( \text{A&amp;B:Unr} )</td>
<td>59.727</td>
<td>58.623</td>
<td>59.696</td>
<td>60.408</td>
<td>60.201</td>
</tr>
<tr>
<td>%( \text{Min} )</td>
<td>35.9±0.001</td>
<td>36.3±0.001</td>
<td>37.1±0.001</td>
<td>37.7±0.001</td>
<td>38.4±0.001</td>
</tr>
</tbody>
</table>

...Continued with similar data for other combinations of Au-STRs and different sets of individuals.
CONCLUSIONS

6.1 VALIDATION OF FAMILIAS FOR THE SIMPLEST ASSUMPTIONS - ABSENCE OF MUTATION AND SILENT ALLELE

The maximum deviation between LRs computed by using algebraic formulas (Table 2) and LRs resulting from the use of software Familias was of the order of magnitude $10^{-9}$. For the six kinship tests analyzed, these differences can be considered small, and likely resulted from rounding procedures, and it seems reasonable for us to conclude that the software must be doing the calculations correctly.

6.2 MEASUREMENT OF THE IMPACT OF CONSIDERING A THIRD INDIVIDUAL AND/OR A GREATER NUMBER OF MARKERS

Of all the six cases studied one conclusion is transversal: when the genetic profile of a third individual is analyzed, and the number of considered markers is increased, the results become statistically more powerful. The conclusions reached for each specific kinship problem studied are:

- For the kinship problem “Full-siblings versus Half-siblings”, the likelihood of an analysis resorting only to individuals A and B, has small probability of result in a high LR favouring one of the hypotheses (see Figure 9). When one undoubted relative of A was analyzed along with A and B, an improvement occurred in the results. The best option in this kinship problem seems to be using the genetic profile of one full-sibling (D) of A. Note that there were less cases erroneously classified when \{A,B,D\} and 17 STRs were analyzed, than for individuals \{A,B\} and 35 STRs. If the genetic profile of a full-sibling of A is not available, the second best option appears to be considering the genetic profile of a grandparent (E), but to have a better chance of reaching good results it is likely needed to consider at least 27 STRs. If forensic expert do not have access to the genetic profile of E, the genetic profile of the mother C of A has to be
considered. Once again, it will be essential to analyze at least 27 STRs to attain high LRs favouring one of the hypotheses in a reasonable proportion of cases.

• For the kinship problem "Full-siblings versus Unrelated", the results revealed to be strong even when analyzing only the set of individuals \{A,B\} and the minimum set of markers - 17 STRs. However, it is notorious that when analyzing the genetic profile of the mother (C) of A, the results show to be extremely powerful. The second best hypothesis is analyzing one full-sibling (D) of A. If there is no access to genetic profiles, either from the mother or from a full-sibling of A, there is still the option of analyzing the genetic profile of a grandparent (E) of A. The results achieved considering the set of individuals \{A,B,E\} are better than the ones resorting only to the set of individuals \{A,B\}.

• For the kinship problem “Half-siblings versus Unrelated”, to have a better chance of achieving statistical relevance it is likely not to be sufficient to only analyze A and B, seeming to be fundamental to add the genetic profile of a third individual and considering more than 17 markers. So, the best option seems to be considering the genetic profile of a grandparent (E) of A and a minimal set of 27 STRs. The second hypothesis is to add the genetic profile of a full-sibling (D) of A and consider a minimum set of 32 STRs. And, the last hypothesis is to consider the set of individuals \{A,B,C\} and analyze the maximal number of markers - 35 STRs.

• For the kinship problem ”Avuncular versus Unrelated”, and as happened in the previous case, it is also prudent to consider the genetic profile of a third individual (relative of A), and more than 17 markers. Thereby, prioritizing the alternatives: (I) analyze the set of individual \{A,B,D\} for a minimum set of 32 markers; (II) analyze the set of individuals \{A,B,C\} and a minimum set of 32 markers; and, (III) analyze the set of individuals \{A,B,E\} for 35 markers.

• For the kinship problem “First cousins versus Unrelated”, and based on the simulations performed, analyze the genetic profile of a grandparent (D) of A and consider the maximum number of markers - 35 STRs, seems to be the only alternative to achieve good results in a relevant proportion of cases.

• For the kinship problem “Half first cousins versus Unrelated”, as in the previous case study, consider the genetic profile of a grandparent (D) of A and analyze the maximum number of markers - 35 STRs, seems to be the alternative to obtain powerful results in a large proportion of cases. However, in this kinship test, comparing to the previous one, more cases were wrongly classified, as expected.
FUTURE WORK

In the near future, the study developed in this thesis will be submitted to an international reference journal in the area of legal medicine, hoping that it will contribute to the theoretical development of the problem, providing information of which individuals should use (i.e. order of preference) to improve the analysis of kinship problems.

For future work, as the validation performed for the simplest assumptions (absence of mutation and silent allele), a new validation should be performed taking into account the presence of mutations and/or silent allele, considering the same two individuals. Recently a paper (Egeland et al., 2017) was published, where algebraic formulas were proposed for pairwise cases considering the possibility of mutation and the presence of silent alleles, so a validation of kinship software for these assumptions can be accomplished.

To elaborate a reformulation of the formulas presented in (Fung et al., 2006), in order to allow any kinship between three individuals seems also required. From this point an adaptation of the work of Egeland (Egeland et al., 2017) to trios (i.e. with the possibility of the presence of mutations and silent alleles when analyzing three individuals) should be fulfilled, allowing the validation of new cases.

Following the knowledge acquired from the development of the work described above, can be attempted a general framework that allows the analysis of any number of individuals analyzed.


