Populations and Statistics in Forensic Genetics

Andreas Tillmar
Supervisors
Bertil Lindblom, Professor
Department of Clinical and Experimental Medicine, Faculty of Health Science,
Linköping University, Sweden

Gunilla Holmlund, Associate professor
Department of Clinical and Experimental Medicine, Faculty of Health Science,
Linköping University, Sweden

Petter Mostad, Associate professor
Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Sweden

Faculty opponent
Peter de Knijff, Professor
Department of Human and Clinical Genetics, Leiden University Medical Center, The Netherlands

Examination board
Xiao-Feng Sun, Professor
Department of Clinical and Experimental Medicine, Faculty of Health Science,
Linköping University, Sweden

Peter Söderkvist, Professor
Department of Clinical and Experimental Medicine, Faculty of Health Science,
Linköping University, Sweden

Elja Arjas, Professor
Department of Mathematics and Statistics, University of Helsinki, Finland

John Carstensen, Professor (as an alternative)
Department of Medicine and Health Sciences, Faculty of Health Science, Linköping University, Sweden

© Andreas Tillmar, 2010
Andreas Tillmar was formerly known as Andreas Karlsson (changed July 2008)

Printed by LiU-Tryck, Linköping 2010
ISBN: 978-91-7393-420-6
ISSN: 0345-0082
To Ida and Signe
Abstract

DNA has become a powerful forensic tool for solving cases such as linking a suspect to a crime scene, resolving biological relationship issues and identifying disaster victims. Traditionally, DNA investigations mainly involve two steps; the establishment of DNA profiles from biological samples and the interpretation of the evidential weight given by these DNA profiles. This thesis deals with the latter, with focus on models for assessing the weight of evidence and the study of parameters affecting these probability figures.

In order to calculate the correct representative weight of DNA evidence, prior knowledge about the DNA markers for a relevant population sample is required. Important properties that should be studied are, for example, how frequently certain DNA-variants (i.e. alleles) occur in the population, the differences in such frequencies between subpopulations, expected inheritance patterns of the DNA markers within a family and the forensic efficiency of the DNA markers in casework.

In this thesis we aimed to study important population genetic parameters that influence the weight of evidence given by a DNA-analysis, as well as models for proper consideration of such parameters when calculating the weight of evidence in relationship testing.

We have established a Swedish frequency database for mitochondrial DNA haplotypes and a haplotype frequency database for markers located on the X-chromosome. Furthermore, mtDNA haplotype frequencies were used to study the genetic variation within Sweden, and between Swedish and other European populations. No genetic substructure was found in Sweden, but strong similarities with other western European populations were observed.

Genetic properties such as linkage and linkage disequilibrium could be important when using X-chromosomal markers in relationship testing. This was true for the set of markers that we studied. In order to account for these properties, we proposed a model for how to take linkage and linkage disequilibrium into account when calculating the weight of evidence provided by X-chromosomal analysis.

Finally, we investigated the risk of erroneous decisions when using DNA investigations for family reunification. We showed that the risk is increased due to uncertainties regarding population allele frequencies, consanguinity and competing close relationship between the tested individuals. Additional infor-
mation and the use of a refined model for the alternative hypotheses reduced
the risk of making erroneous decisions.

In summary, as a result of the work on this thesis, we can use mitochondrial
DNA and X-chromosome markers in order to resolve complex relationship
investigations. Moreover, the reliability of likelihood estimates has been in-
creased by the development of models and the study of relevant parameters
affecting probability calculations.
DNA har blivit ett viktigt verktyg inom rättsväsendet för att kunna lösa frågeställningar liknande dom som att kunna koppla en misstänkt gärningsman till en brottsplats, utreda frågor rörande biologiska släktspå eller identifiera offer vid masskadoffer. Man kan dela in en DNA-utredning i två steg; dels framtagandet av DNA profilerna, dels en värdering av vilken betydelse de erhållna DNA profilerna har utifrån en given frågeställning. Mer specifikt, man vill t.ex. veta sannolikheten för att någon annan än den misstänkte har lämnat DNA på brottsplatsen, eller sannolikheten för att den utpekade mannen är barnets far jämfört med att han slumpmässigt passar in. Beräkningen av sådana sannolikheter baseras bl.a. på olika DNA-varianters (alleler) förekomst i befolkningen och hur de genetiska markörerna nedärvs från en generation till en annan.

Denna avhandling har syftat till att studera relevanta bakgrundsdata som påverkar sannolikhetsberäkningarna, samt studera matematiska modeller för att på ett korrekt sätt ta hänsyn till de studerade parametrarna vid fallspecifika sannolikhetsberäkningar. Avhandlingsarbetet fokuserar på den genetiska variationen i en svensk befolkning och på DNA undersökningar som gäller frågor rörande biologiskt släktskap mellan individer. Detta till trots så är de flesta resultaten och diskussionerna även giltiga vid användandet av DNA profilerings i brottsplatsundersökningar.


samt studerade betydelsen och inverkan av olika faktorer som kan påverka detta. Vi fokuserade på DNA-utredningar i familjeåterföreningsärenden, vilka kan vara complexa då de innefattar osäkerheter kring populationstillhörighet, skillnader i familjekonstellationer etc. Genom simuleringar visade vi att felens kan minimeras om man ökar undersökningens informationsgrad, t.ex. genom att använda fler DNA-markörer, DNA-profile från fler individer samt allelfrekvensdata från samma population. Dessutom visade vi att det går att minska risken för fel ytterligare genom att man använder sig av en förfinad metod för att kunna ta hänsyn till alternativa närbesläktade släktspår mellan de testade individerna.

I standardutredningar används DNA-markörer belägna på de s.k. autosomala kromosomer. För specialfall kan man även undersöka DNA-variationer som finns på mitokondriens (mtDNA) eller på könskromosomerna (X-kromosomen och Y-kromosomen). MtDNA ärvs på mödernet, och är speciellt användbart för utredning vid förmodat maternellt släktspår. I delarbete två undersökte vi mtDNA variationen i en svensk population i syfte att skapa en frekvensdatabas. Genom att analysera blodprover från ca 300 svenskar från sju geografiskt skilda regioner kunde vi visa att informationsgraden för användning i en svensk population är jämförbar med andra europeiska populationer. Dessutom visade vi i studien att det inte finns några signifikanta skillnader mellan mtDNA variationen i de olika svenska regionerna.


Slutligen i delarbete fyra tog vi fram en matematisk beräkningsmodell för att korrekt ta hänsyn till både koppling och kopplingsojämvikten vid sannolikhetsberäkningar i släktutsattutredningar baserade på X-kromosomdata. Vi applicerade denna beräkningsmodell i en simuleringssstudie på ett antal typfäl och visade på graden av fel om man använder en enklare beräkningsmodell där ingen hänsyn till koppling eller kopplingsojämvikt tas.

Sammanfattningvis, i och med arbetena i denna avhandling så kan vi använda mitokondriell DNA och X-kromosomala DNA-markörer för att lösa mer komplexa släktutsattutredningar. Genom framtagandet av modeller och
studie av relevanta parametrar som påverkar släktskapssannolikhsberäkningen har tillförlitigheten i de beräknande sannolikheterna kunnat ökas.
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the respective publishers.
Paper I © 2007 Elsevier, Forensic Science International
Abstract

Populärvetenskaplig sammanfattning

List of Papers

Contents

Abbreviations

| Introduction ................................................................................................................. | 17 |
| History of DNA and forensic genetics............................................................... | 17 |
| Population genetics................................................................................................ | 18 |
| Genetic polymorphisms................................................................................... | 18 |
| DNA inheritance .............................................................................................. | 20 |
| Population Genetics ......................................................................................... | 22 |
| The Swedish population and its genetic appearance................................. | 25 |
| Forensic mathematics/statistics........................................................................ | 26 |
| Framework for interpretation and presentation of evidential weight .......... | 26 |
| Paternity index calculation............................................................................... | 27 |
| Mathematical model for automatic likelihood computation for relationship testing | 29 |

| Aim of the thesis .............................................................................................. | 31 |
| Specific aims....................................................................................................... | 31 |
| Paper I ................................................................................................................ | 31 |
| Paper II............................................................................................................... | 31 |
| Paper III ............................................................................................................. | 31 |
| Paper IV ............................................................................................................. | 31 |

| Investigations .................................................................................................... | 33 |
| Paper I - DNA-testing for immigration cases: the risk of erroneous conclusions | 33 |
| Materials and methods...................................................................................... | 33 |
| Results and discussion...................................................................................... | 34 |
| Paper II - Homogeneity in mitochondrial DNA control region sequences in Swedish subpopulations | 36 |
Materials and methods ................................................................. 36
Results and discussion ............................................................... 36
Paper III - Analysis of linkage and linkage disequilibrium for eight X-STR markers ......................................................... 38
Materials and methods ................................................................. 38
Results and discussion ............................................................... 38
Paper IV - Using X-chromosomal markers in relationship testing: How to calculate likelihood ratios taking linkage and linkage disequilibrium into account ................................................................. 40
Materials and methods ................................................................. 40
Results and discussion ............................................................... 40
Concluding remarks ................................................................. 43
Future perspectives ................................................................. 45
Acknowledgements ................................................................. 47
References ................................................................. 49
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>θ</td>
<td>Theta, recombination frequency/fraction</td>
</tr>
<tr>
<td>AF</td>
<td>Alleged father</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>F_{ST}</td>
<td>Measure of population genetic subdivision</td>
</tr>
<tr>
<td>GD</td>
<td>Gene diversity</td>
</tr>
<tr>
<td>HWE</td>
<td>Hardy-Weinberg equilibrium</td>
</tr>
<tr>
<td>ISFG</td>
<td>International Society of Forensic Genetics</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>LR</td>
<td>Likelihood ratio</td>
</tr>
<tr>
<td>MEC</td>
<td>Mean exclusion chance</td>
</tr>
<tr>
<td>MtDNA</td>
<td>Mitochondrial DNA</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PD</td>
<td>Power of discrimination</td>
</tr>
<tr>
<td>PE</td>
<td>Power of exclusion</td>
</tr>
<tr>
<td>PI</td>
<td>Paternity index</td>
</tr>
<tr>
<td>PIC</td>
<td>Polymorphism informative content</td>
</tr>
<tr>
<td>PM</td>
<td>Match probability</td>
</tr>
<tr>
<td>Pr</td>
<td>Probability</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>STR</td>
<td>Short tandem repeat</td>
</tr>
<tr>
<td>TF</td>
<td>True father</td>
</tr>
</tbody>
</table>
Introduction

History of DNA and forensic genetics

When Watson & Crick discovered the structure of the DNA molecule (Watson & Crick 1953) they could probably not imagine the future usefulness of their finding. By analysing DNA, information about genetic diseases, evolution of biological life and population history can be retrieved. Nowadays, DNA is used in everyday practice for applications within different areas, such as medical genetics, the food processing industry and in forensic situations when solving crimes as well as in disputes about biological relationships.

Traditionally, the aim of forensic genetics is to provide a statement about the identity of a human being based on a biological sample by means of a DNA analysis. However, forensic genetics today covers a wider spectrum of areas, such as forensic molecular pathology (Karch, 2007), complex traits (Kayser et al., 2009; Pulker et al., 2007) and wildlife forensics (Alacs et al., 2010; Budowle et al., 2005). When it comes to human identification, the task could be to connect a suspect to the crime scene or investigate a biological relationship (Jobling et al., 2004b).

The first time DNA was used in court for a crime scene sample was in 1986 in the UK (Gill et al., 1987). The case involved the exclusion of a murder suspect using multi locus DNA-probes (Jeffreys et al., 1985). Since then, the techniques and methodologies of employing the information provided by DNA have undergone enormous improvements, making it an obvious tool for routine practice when dealing with forensic issues. Perhaps the most famous case when the use of DNA was put under pressure, and from which lessons still can be learned, was the trial of O.J. Simpson (Lee & Labriola, 2001). This trial is a good example of the importance of the complete process from handling evidential biological samples at the crime scene, via storage and the establishment of DNA profiles, to the presentation of the weight of the evidence provided by the DNA results in court. In no other trial has the DNA result been so thoroughly examined, discussed and questioned by the defence.

Within forensic genetics, a DNA investigation always has a question to answer. For example, is the donor of a crime scene sample the same individual as the suspect? and Is the alleged father (AF) the biological father of the child? When the establishment of DNA profiles is finished, they are used for interpre-
Introduction

...ation of the case specific question. Normally, three different statements can be presented for any given hypothesis tested; exclusion, inconclusive or inclusion. When no exclusion can be made, some sort of statistical evaluation has to be performed in order to estimate the strength of the evidence provided by the DNA profiles. Put simply, the majority of such cases involve consideration of the probability to see identical DNA profiles from unrelated individuals by coincidence. The statistical assessment and presentation of the DNA evidence are crucial for the acceptance of DNA as a routine tool.

The establishment of these figures is usually based on the genetic uniqueness of the information that exists in the DNA profile in the context of a relevant population. The main aim of the present thesis is to discuss issues that are important for relationship testing, but many aspects and parameters studied and discussed here are just as important for evaluating DNA evidence in criminal casework.

Two main areas must be studied in order to establish the probability of the evidential weight for a given DNA marker. First, population genetics including allele frequencies, population substructure, dependence within and between markers and others. Second, models for calculating and presenting the weight of evidence, taking the former information properly into account.

Population genetics

Genetic polymorphisms

Three different types of DNA marker, Short Tandem Repeats (STRs), Single Nucleotide Polymorphisms (SNPs) and DNA sequence data (Figure 1), represent the absolute majority of polymorphisms used in forensic genetic applications. They all have characteristics, making them especially useful for solving criminal cases and for relationship testing.

An STR marker consists of a short DNA sequence (e.g. GATA) repeated a variable number of times. These markers are widespread throughout the genome and account for approximately 3% of the total human genome (Ellegren, 2004). They have a relatively high mutation rate, which is the reason for their high degree of polymorphism. STRs are robust, easy to multiplex for PCR amplification and exhibit high polymorphisms among human populations (Butler, 2006). In other words, they have good characteristics for use in forensic applications. More than 10 alleles exist for the commonly used STRs, which generally makes a multi locus STR DNA profile unique. In the 1990s, the FBI concentrated on 13 STR markers, called CODIS loci (Budowle et al., 1999a). These and some additional markers were then adopted and commercialised by a few corporations, thus making them the standard set up of markers for use in routine practice. Recently, developments have taken place in relation to STRs with shorter amplicon sizes (i.e. miniSTRs) (Wiegand & Kleiber, 2001). These have
the advantage of increasing the probability of obtaining complete profiles for degraded DNA.

Another type of marker is the SNPs, which consist of single base polymorphisms. These are often biallelic, although there is an increasing interest in tri-allelic SNPs for forensic applications (Westen et al., 2009). SNPs have the advantage that short amplicons can be used for the PCR amplification, which is particularly important for degraded samples. Another feature is the low mutation rate, which is an advantage in relationship testing. The disadvantage, however, is that since the number of alleles per locus is limited, the information content is low. The amount of information from one STR marker is the same as from approximately four SNPs (Sobrino et al., 2005, Brenner (www.dna-view.com)). Regarding SNP multiplexes, there is no commercial forensic kit available, although some work has taken place and efforts made to develop such multiplexes for use in criminal cases and for relationship testing (Borsting et al., 2009, Philips et al., 2008).

A third alternative is to use nucleotide sequence variation, i.e. information from a DNA sequence spanning a pre-defined region. The main use of sequence data in forensic situations involves the analysis of variation on the mitochondrial DNA (mtDNA). No STRs are present on the mtDNA. Analysis of mtDNA SNPs, in addition to the sequence data, has, however, been shown to increase the total discrimination power (Coble et al., 2004).

| STR-Allele 1 | GATA GATA GATA |
| STR-Allele 2 | GATA GATA GATA GATA |
| SNP-Allele 1 | A |
| SNP-Allele 2 | C |
| DNA seq 1 | AGGACTCCGATAA |
| DNA seq 2 | ACCACGCCCATAA |

Figure 1. Illustration of alleles for a STR marker (top), SNP marker (middle) and DNA sequence variation (bottom).
DNA inheritance

In addition to the markers described above, there are different “types” of DNA with different properties in terms of their inheritance pattern as well as other important population genetic properties. The types discussed here are markers on the autosomal chromosomes, the sex-chromosomes (X-chromosome and Y-chromosome) and the mitochondrial DNA (mtDNA).

For an autosomal locus, each individual has two alleles, one inherited from the mother and one from the father. The traditional use of autosomal markers in forensic relationship testing only provides information on relationships spanning from one to a few generations (Nothnagel et al., 2010). However, technical improvements have made it possible to simultaneously study hundreds of thousands of autosomal markers, thus reducing the limitations associated with complex pedigree testing (Egeland et al., 2008; Skare et al., 2009).

Moving on to the X-chromosome, which has different inheritance pattern compared with autosomal markers. Females have two copies of the X-chromosome, while males normally only have one. A consequence of this is that X-chromosomal markers act as autosomal markers in their transmission to gametes in females and as haploid markers in males. Females inherit one X-chromosome from their mother and their father’s only X-chromosome, while males inherit their only X-chromosome from the two belonging to their mother. In relationship testing, X-chromosome analysis is particularly useful in deficiency cases. Consider, for example, a case where two sisters are tested to establish whether or not they have the same father, and where DNA profiles are only available for the sisters. In such instances, autosomal DNA markers cannot exclude paternity, since two sisters can inherit different alleles despite being full siblings. The use of X-chromosome markers can, however, exclude paternity, since two sisters would share the same paternal allele if they have the same father. There are several other types of relationship where analysis of X-chromosomal markers is superior to autosomal markers (Szibor et al., 2003; Pinto et al., 2010).

The use of the X-chromosome in forensic relationship testing usually involves STR markers. Detailed information regarding more than 50 X-STRs has been collected (www.chrx-str.org) and used in different PCR multiplexes (Becker et al., 2008; Hundertmark et al., 2008; Gomez et al., 2007; Diegoli et al., 2010). Linkage and linkage disequilibrium must typically be considered when using a combination of closely located X-chromosomal markers in relationship testing (Krawczak, 2007; Szibor, 2007) (Figure 2). These two genetic properties are further discussed below in terms of their definitions and impact on calculated likelihoods.

The Y-chromosome normally exists in one copy in males and is absent in females. It is inherited from father to son, thus all men in a paternal lineage share an identical Y-chromosome. Apart from the recombination region (~5%), mutation is the only force that leads to new variation on the Y-
Introduction

Due to this and the fact that the Y-chromosome has one-fourth of the relative population size compared with autosomal loci, the Y-chromosomal variation has been found to be fairly population specific (Hammer et al., 2003; Jobling et al., 2004a). As a result, regional population databases must be collected and studied.

Both SNPs and STRs are used as markers on the Y-chromosome. Y-SNPs can provide information about an individual's haplogroup status (Karafet et al., 2008), which can, for instance, be used for interpreting the paternal genetic geographical origin (Jobling & Tyler-Smith, 2003). For other forensic issues, analysis of Y-STRs (resulting in a haplotype) is more useful (Jobling et al., 1997). Nevertheless, it is crucial to bear in mind that the Y-chromosome haplotype is consistent for all males who share the same paternal lineage.

DNA from the mitochondrion can also be used in forensic investigations. It consists of a circular genome of ~16,600 nucleotides. Each cell has 100 to 1000 copies of its mtDNA, which makes it especially useful in forensic analyses, where the amount of DNA can often be very low. The mtDNA is inherited from mother to child (maternal) and can therefore be used to solve questions involving a potential maternal relationship. From a population point of view, mtDNA has many similarities with other haploid genomes (e.g. the Y-chromosome). Because of its haploid status, mtDNA profiles are also relatively population specific, which must be accounted for when conclusions are made (Holland & Parsons, 1999).

Figure 2. Illustration of the inheritance pattern of two X-chromosomal loci, located at a distance θ from each other, in a family consisting of a mother, father and a female child. X1a-c and X2a-c are alleles for the X-chromosomal markers 1 and 2, respectively. The value in parenthesis is the segregation probability for the inheritance of the given haplotype from the parents.
Population Genetics

Population genetics is the study of hereditable variation and its change over time and space and includes the process of mutation, selection, migration and genetic drift. By quantification of different DNA alleles and their occurrence within and between populations, information about parameters such as population structure, growth, size and age can be retrieved (Jobling et al., 2004a).

Substructure

In addition to the estimation of allele frequencies, it is also important to check for possible genetic substructures within a population and to study genetic variation among populations. The most common way of studying these differences is by means of $F_{ST}$-statistics (Wright, 1951, see also Holsinger & Weir, 2009 for a review). $F_{ST}$ has a direct relationship to the variance in allele frequencies within/among populations. Small $F_{ST}$-values correspond to small differences within/among populations, and vice versa. Variants of $F_{ST}$ exist, which in addition also take relevant evolutionary distance between alleles into account (e.g. $\Phi_{ST}$ and $R_{ST}$). For forensic purposes it is highly important to study possible substructure in the population of interest. If substructure exists, it has to be accounted for when producing the strengths of the DNA profile evidence (Balding & Nichols, 1994).

Linkage and Linkage disequilibrium

Linkage and linkage disequilibrium (LD) deal with the phenomenon characterized by the dependence that can exist between different loci and between alleles at different loci.

Linkage can be defined as the co-segregation of closely located markers within a family (Figure 2). During meiosis, the maternal and paternal chromosome homologs align and exchange segments by a phenomenon known as crossing over, or recombination. Consider, for example, two markers located on the same chromosome. If recombination occurs between the two markers, the resulting chromosome in the gametes now has a different appearance compared with its parental chromosomes. The allele combination of the two markers (i.e. haplotype) is thus changed compared with its parental constitution. The distance between two loci can be measured and discussed as the recombination frequency, $\theta$, and estimated based on data from family studies. The recombination frequency is correlated to the genetic distance between the loci (Ott, 1999).

Linkage disequilibrium on the other hand concerns dependencies between alleles at different loci, and can be defined as the non-random association of alleles in haplotypes. LD can originate from the fact that the loci are closely located, thus inherited together more often than randomly. However, it can also be due to population genetic events such as selection, founder effects and admixture (Ott, 1999). LD can be studied by comparison between observed hap-
lotype frequencies and haplotype frequencies expected under linkage equilibrium (LE).

If we have two loci and are interested in the population frequency for haplotype \(a-b\), where \(a\) is the allele at locus 1 and \(b\) is the allele at locus 2, the frequency can be estimated from

\[
f(ab) = f(a) \cdot f(b) + \Delta
\]

Where \(f(ab)\) is the frequency for the haplotype \(a-b\), and \(f(a)\) and \(f(b)\) are the allele frequency for alleles \(a\) and \(b\), respectively. If we have linkage equilibrium, then \(\Delta = 0\), i.e. no association exists between \(a\) and \(b\). However, if there is a dependency between the alleles in locus 1 and locus 2 then \(\Delta \neq 0\) and the loci are considered to be in LD.

If haplotype frequencies are to be estimated for markers in LD, they are best inferred directly from observed haplotype frequencies in the population rather than estimating \(\Delta\) for each allele combination, especially when dealing with multiallelic loci.

Validation of a frequency database

Prior to the introduction of new DNA markers into forensic casework, studies should be performed on the relevant population in order to establish allele (or haplotype) frequencies and investigate potential substructure. Furthermore, certain tests must be conducted concerning the independent segregation of alleles. Hardy-Weinberg equilibrium, HWE, (Hardy, 1908) and LD tests deal with the issue of independence of alleles within a locus and between loci, respectively. If the population is not in HWE or in LE it has to be accounted for when calculating the statistics in casework. When performing the HWE and LD tests, Fisher’s exact test is the preferable method (Fisher, 1951). However, it is important to note that the exact test has very limited power, making it difficult to draw any highly significant conclusions about the outcome of either test (Buckleton et al., 2001).

Another feature to consider is the forensic efficiency of using the DNA markers in casework involving criminal cases and relationship testing. Such estimates describe the theoretical value of using the specific markers for different forensic genetic situations and differ from case specific values. The estimation of such parameters is most often based on the number of distinctive alleles found in the population and their corresponding frequencies.

The description and mathematical formulation of a selection of useful parameters are provided below.

There are different definitions of gene diversity (GD). This parameter describes the probability that two alleles drawn at random from the population will be different.
Introduction

The unbiased estimator is given by (Nei, 1987)

\[ GD = \frac{n}{n-1} (1 - \sum_{i} p_{i}^{2}) \]

where \( n \) is the number of gene copies sampled and \( p_{i} \) is the frequency of the \( i \)th allele in the population.

The match probability (PM), is defined as the probability of a match between two unrelated individuals and is calculated as (Fisher, 1951)

\[ PM = \sum_{i} G_{i}^{2} \]

where \( G_{i} \) is the frequency of the genotype \( i \) at a given locus in the population. Thus, PM is the sum of all partial match probabilities for all genotypes. PM can also be interpreted from allele frequencies given that the population is in Hardy Weinberg equilibrium (Jones, 1972).

The power of discrimination (PD) is defined as the probability of discriminating between two unrelated individuals. Thus, correlated to PM discussed above

\[ PD = 1 - PM \]

Polymorphism Informative Content (PIC) can be interpreted as the probability that the maternal and paternal alleles of a child are deducible, or the probability of being able to deduce which allele a parent has transmitted to the child (Botstein et al, 1980; Guo & Elston, 1999). There are two instances when this cannot be deduced, namely when one parent is homozygous or when both parents and the child have the same heterozygous genotype. Thus

\[ PIC = 1 - \sum_{i=1}^{n} p_{i}^{2} - \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2p_{i}^{2}p_{j}^{2} \]

where \( p_{i} \) and \( p_{j} \) are allele frequencies.

The probability of excluding paternity (Q) is calculated from (Ohno et al., 1982)

\[ Q = \sum_{i=1}^{n} p_{i}(1-p_{i} + p_{i}^{2})(1-p_{i})^{2} + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{i}p_{j}(1-p_{i} + p_{j}^{2})(p_{i} + p_{j}) \]

Q is inferred from two factors. First the exclusion probability for a given mother/child genotype combination, which is either \((1-p_{i})^{2}\) or \((1-p_{i}p_{j})^{2}\), and second, the expected population frequency for the genotypes of the mother/child combination. \( p_{i} \) and \( p_{j} \) are the frequencies for the paternal alleles. Q is then interpreted from the sum of all mother/child genotype combinations
as described above. An alternative figure for the power of exclusion (PE) exists and is defined as (Brenner & Morris, 1990)

\[ PE = h^2 \cdot (1 - 2 \cdot h \cdot H^2) \]

where \( h \) is the proportion of heterozygous individuals and \( H \) the proportion of homozygous individuals in the population sample.

The formulas given so far are for autosomal markers. Corresponding formulas exist for X-chromosomal markers (Szibor et al., 2003), such as the mean exclusion chance (MEC) for trios including a daughter (Desmarais et al., 1998). This is equivalent to the probability of exclusion, \( Q \), with the difference that the exclusion probability for a given mother/child genotype combination is either \((1-p_i)\) or \((1-p_i-p_j)\). Thus, the mean exclusion chance when mother and child are tested is

\[
MEC_{Trio} = 1 - \sum_i p_{i}^2 + \sum_i p_{i}^4 - (\sum_i p_{i}^2)^2
\]

where \( p_i \) is the allele frequency for allele \( i \). \( p_i \) can also represent haplotype frequency, if such is considered. The mean exclusion chance for duos involving a man and a daughter \( MEC_{Duo} \) (Desmarais et al., 1998) is

\[
MEC_{Duo} = 1 - 2 \sum_i p_{i}^2 + \sum_i p_{i}^3
\]

The Swedish population and its genetic appearance

Immigration into Scandinavia did not start until around 12 000 years ago due to the ice that covered Northern Europe. Since then, immigration and population movements of various degrees, descent and directions have occurred within the present borders of Sweden. Many of the groups that immigrated originated from Western Europe and are suggested to represent a non-Indo-European population (Blankholm, 2008; Zvelebil, 2008). This, in combination with recorded demographic events over the last 1 000 years (Svanberg, 2005), may be the cause of the genetic composition of the modern Swedish population.

The Swedish population has been investigated regarding forensic autosomal STRs (Montelius et al., 2008) and forensic autosomal SNPs (Montelius et al., 2009). Both of these studies revealed high genetic diversities and information content for usage in relationship testing and criminal cases. Strong similarities with other European populations were also recorded. A sample of the Swedish population was recently compared with other European populations based on data from over 300 000 SNPs, which showed a strong correlation between the geographic location and the genetic variability for the tested populations (Lao et al., 2008).
Introduction

Regarding Y-chromosome variation, some studies have aimed at facilitating the setting-up of a Swedish reference database (Holmlund et al., 2006), while others have explored the demographic history of the Swedish male population (Karlsson et al., 2006; Lappalainen et al., 2009). These later studies confirm earlier findings of high similarity with other western European populations (Roewer et al., 2005). However, some Y-chromosome differences, albeit small, do exist within Sweden especially in the northern part of the country (Karlsson et al., 2006).

Y-STR and Y-SNP data from the Swedish population are included in YHRD, the world-wide Y-chromosome haplotype database (Willuweit & Roewer, 2007).

Due to continuous immigration to Sweden from various populations, knowledge about non-European populations is also crucial for a correct assessment of the weight of evidence (Tillmar et al., 2009).

Forensic mathematics/statistics

In order to assess the evidential weight for a DNA analysis, the numerical strength of the evidence must be calculated as well as presented to the court or client in an appropriate way.

Framework for interpretation and presentation of evidential weight

When presenting the probability or weight of the DNA findings, a logical framework is crucial in order to make the presentation clear and understandable to those who have to make decisions based on the DNA results. The design of such a framework has been debated and there is still no clear consensus within the forensic community.

The main discussion covers two (or perhaps three) different frameworks including a frequentist and a Bayesian approach (or a logistical approach, which could be extended to a full Bayesian approach). These have different properties as well as pros and cons, and several detailed publications about their usage exist (for example see Buckleton et al., 2003, chapter 2, for a review).

In brief, the frequentist approach is built around the calculation of a probability concerning one hypothesis. For example \( \Pr(E \mid H) \), which means the probability of the evidence, \( E \), when hypothesis \( H \) is true. In this case \( E \) is the DNA profile and \( H \) could be “the probability that the DNA come from an individual not related to the suspect”. If this probability is computed to be low, the hypothesis can be rejected, making an alternative hypothesis probable. The argument in favour of this approach is that it is intuitive and relatively easy to understand. However, it has been the subject of some criticism, mainly due to
the lack of logical rigour, which makes the set up of the hypothesis and its inter-
pretation extremely important.

The main characteristic of a Bayesian or logical approach is the use of a like-
lihood ratio (LR) connecting the prior odds to the resulting posterior odds, i.e
Bayes’s theorem (see formula below). The advantage of this approach is that
the LR can be connected to any other evidence, such as fingerprint, informa-
tion from eyewitnesses etc.

\[
\frac{\Pr(H_1 \mid E, I)}{\Pr(H_0 \mid E, I)} = \frac{\Pr(E \mid H_1, I)}{\Pr(E \mid H_0, I)} \frac{\Pr(H_1)}{\Pr(H_0)}
\]

posterior odds = likelihood ratio \cdot prior odds

\(H_1\) (or \(H_0\)) is commonly known as the prosecutor’s hypothesis, and \(H_0\) (or \(H_D\))
is the hypothesis for the defence. \(E\) represents the DNA profiles, and \(I\) is other
relevant background evidence. The ratio \(\frac{\Pr(E \mid H_1, I)}{\Pr(E \mid H_0, I)}\) is the LR and it is
within this formula that the strength of the DNA is quantified. The calculation
of the LR for paternity cases (i.e. Paternity Index PI) is discussed in the follow-
ing section.

Regarding the choice of framework for relationship testing, the Paternity
Testing Commission (PTC) of the International Society for Forensic
Genetics (ISFG) recently published biostatistical recommendations for probability calcu-
lation specific to genetic investigations in paternity cases (Gjertson et al., 2007).
They recommend the use of the LR (i.e. PI) principle for calculating the weight
of evidence. These recommendations cover the most basic issues, but lack in-
formation on how to deal with, for example, linked genetic markers.

Paternity index calculation

As an example, let \(H_p\) and \(H_d\) represent two mutually exclusive hypotheses for
and against paternity:

\(H_p\): The alleged father is the father of the child
\(H_d\): A random man, not related to the alleged father, is the father of the child.

The paternity index (PI) is typically defined as

\[
PI = \frac{\Pr(G_C, G_M, G_Af \mid H_p)}{\Pr(G_C, G_M, G_Af \mid H_d)}
\]

which means the probability of seeing the child’s (\(G_C\)), mother’s (\(G_M\)) and al-
leged father’s (\(G_Af\)) DNA profiles when the AF is the father, in comparison to
seeing the same DNA profiles when the AF is not the father.
We can use the third law of probability and simplify

\[
PI = \frac{\Pr(G_C, G_M, G_{AF} \mid H_p)}{\Pr(G_C, G_M, G_{AF} \mid H_d)} = \frac{\Pr(G_C \mid G_M, G_{AF}, H_p) \Pr(G_M, G_{AF} \mid H_p)}{\Pr(G_C \mid G_M, G_{AF}, H_d) \Pr(G_M, G_{AF} \mid H_d)}
\]

The probability of seeing the DNA profiles from the mother and the AF is the same, irrespective of the hypothesis. Thus, we can make a further simplification

\[
\Pr(G_M, G_{AF} \mid H_p) = \Pr(G_M, G_{AF} \mid H_d) = \Pr(G_M, G_{AF})
\]

resulting in

\[
PI = \frac{\Pr(G_C \mid G_M, G_{AF}, H_p)}{\Pr(G_C \mid G_M, G_{AF}, H_d)}
\]

We now need to calculate two probabilities. 1) The probability of the child’s genotype, given the genotypes of the mother and the AF, and given that the AF is the father (numerator), and 2) the probability of the child’s genotype, given the genotypes of the mother and the AF, and given that the AF is not the father, but that someone else is (denominator).

We start with the calculation of 1) and assume that we have data from a single locus. This probability is based on Mendelian heritage. If it is possible to determine the maternal (A_M) and paternal (A_P) alleles for the child (assuming that the mother is the true mother) the numerator can either be 1, 0.5 or 0.25, depending on the homozygous/heterozygous status of the mother and the AF. If both the mother and the AF are homozygous, the numerator is 1 (the mother and the AF cannot share any other alleles). If either the AF or the mother is heterozygous, the probability is 0.5 since there is a 50/50 chance that the child will inherit one of the alleles from a heterozygous parent. Consequently, if both the mother and the AF are heterozygous, the probability will be 0.25 (0.5 times 0.5).

If A_M and A_P are unambiguous, the denominator \(\Pr(G_C \mid G_M, G_{AF}, H_d)\) is either \(p_{A_P}\), or \(0.5p_{A_P}\), depending on the homozygous/heterozygous status of the mother. \(p_{A_P}\) is the population frequency of allele A_P and represents the probability of the child receiving the allele from a random man in the population. If A_M and A_P are ambiguous, the PI is calculated as the sum of all possible values for A_M and A_P.

As a simple example, let G_M have the genotype \([a,b]\), G_C have \([b,c]\) and G_AF have \([c,d]\).

Then

\[
\Pr(G_C \mid G_M, G_{AF}, H_p) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}
\]

28
Introduction

and

\[ \Pr(G_C \mid G_M, G_{AF}, H_d) = \frac{1}{2} \cdot p_c \]

thus

\[ PI = \frac{\Pr(G_C \mid G_M, G_{AF}, H_p)}{\Pr(G_C \mid G_M, G_{AF}, H_d)} = \frac{\frac{1}{4}}{\frac{1}{2} \cdot p_c} = \frac{1}{2 \cdot p_c} \]

In other words, as the more unusual allele \( c \) is in the population, the probability that the AF is the biological father of the child has higher evidential weight.

Decision

How does one interpret the PI-value? Bayes’s theorem is relevant in order to obtain posterior odds, from which a posterior probability can be computed. For paternity issues, the prior odds have traditionally been set to 1, leading to the following value for the posterior probability of paternity

\[ PI = \frac{\Pr(H_p \mid E)}{\Pr(H_d \mid E)} \]

hence

\[ PI = \frac{\Pr(H_p \mid E)}{1 - \Pr(H_p \mid E)} \]

resulting in

\[ \Pr(H_p \mid E) = \frac{PI}{PI + 1} \]

Hummel presented suggestions for verbal predicates based on the posterior probability (Hummel et al., 1981). It is however up to the forensic laboratory to set a limit or cut-off for inclusion based on the PI or the posterior probability (Hallenberg & Morling, 2002; Gjertson et al., 2007). A too low cut-off will increase the risk of falsely including a non-father as a true father and vice versa.

Mathematical model for automatic likelihood computation for relationship testing

While the calculation of the PI for trios and single markers are fairly simple, it rapidly becomes more complicated with the introduction of the possibility of
mutations (Dawid et al., 2002), silent alleles (Gjertson et al., 2007), population substructure (Ayres, 2000) and when treating deficiency cases (Brenner, 2006). In such situations the use of a model for automatic likelihood computations is helpful. In 1971 Elston & Stewart presented a model for the exact calculation of the likelihood of a given pedigree (Elston & Stewart, 1971). The likelihood can be described as

$$L(Ped) = \sum_{G_f} \ldots \sum_{G_m} \prod_i Pen(X_i \mid G_i) \prod_{\{founder\}} Pr(G_{\text{founder}}) \prod_{\{o,f,m\}} Pr(G_o \mid G_f, G_m),$$

The Elston-Stewart algorithm uses a recursive approach starting at the bottom of a pedigree by computing the probability for each child’s genotype conditional on the genotype of the parents. The advantage using this approach is that if the summation for the individual at the bottom is computed first, it can be attached as a factor in the calculation of the summation for his parents and thus this individual needs no further consideration. This procedure represents a peeling algorithm. The penetration (Pen) factor can be disregarded when treating non-trait loci.

The Elston-Stewart algorithm works well on large pedigrees but its computational efforts increases with the number of markers included. A need has emerged for a fast computational model for consideration of thousands of linked markers due to increased access to large datasets. Lander and Green developed the Lander-Green algorithm in 1987 (Lander & Green, 1987), which permits simultaneous consideration of thousands of loci and has a linear increase in computational efforts related to the number of markers. The Lander-Green algorithm has three main steps to consider; 1) the collection of all possible inheritance vectors in a pedigree for alleles transmitted from founder to offspring, 2) iteration over all inheritance vectors and the calculation of the probability of the marker specific observed genotypes conditioning on the inheritance vectors and finally 3) the joint probability of all marker inheritance vectors along the same chromosome (e.g. transmission probabilities). By the use of a hidden Markov model (HMM) for the final step, an efficient computational model can be obtained (see Kruglyak et al., 1996, for a more detailed description).

Practical implementation of the Lander-Green algorithm has been shown to work well in terms of taking linkage properly into account for hundreds of thousands of markers although it assumes linkage equilibrium for the population frequency estimation (Abecasis et al., 2002; Skare et al., 2009).
Aim of the thesis

The aim of this thesis was to study important population genetic parameters that influence the weight of evidence provided by a DNA-analysis as well as models for proper consideration of such parameters when calculating the weight of evidence.

Specific aims

Paper I
To analyse the risk of making erroneous conclusions in complex relationship testing and propose methods for reducing the risk of such errors.

Paper II
To establish a Swedish mitochondrial DNA frequency database, compare it in a worldwide context and study potential substructure within Sweden.

Paper III
To investigate eight X-chromosomal STR markers in a Swedish population sample concerning allele and haplotype frequencies and forensic efficiency parameters. Furthermore, to study recombination rates in Swedish and Somali families.

Paper IV
To propose a model for the computation of the likelihood ratio in relationship testing using markers on the X-chromosome that are both linked and in linkage disequilibrium.
Investigations

Paper I - DNA-testing for immigration cases: the risk of erroneous conclusions

The standard paternity case includes a child, the mother of the child and an alleged father (AF). An assessment of the weight of the DNA result can be performed, and a decision whether or not the AF can be included or excluded as the true father (TF) of the child can be made. This decision can, however, be incorrect, due to an exclusion or as an inclusion error (meaning falsely excluding the AF as TF, or falsely including the AF as TF, respectively). In this paper we studied the risk of erroneous decisions in relationship testing in immigration casework. These cases can involve uncertainties concerning appropriate allele frequencies, different degrees of consanguinity, a close relationship between the AF and TF and complex pedigrees.

Materials and methods

A simulation approach was used to study the impact of the different parameters on the computed likelihood ratio and error rates. Two mutually exclusive hypotheses are normally used in paternity testing. We introduced a five hypotheses model in order to account for the alternative of a close relationship between the TF and the AF (Figure 3).

Family data were generated, and in the standard case the individuals’ DNA-profiles were based on 15 autosomal STR markers with published allele frequencies.

When calculating the weight of evidence, expressed as posterior probabilities, we used a Bayesian framework with the standard two hypotheses and the five hypotheses model for comparison. The error rates were studied by comparing the outcome of the test with the simulated relationship using a decision rule for inclusion and exclusion.
Results and discussion

Simulation of a standard paternity case yielded an unweighted total error rate of approximately 0.8 % (for a 99.99% cut off). This might appear fairly high, but is due to the fact that we used an equal prior probability for the possibility of the alternative hypotheses, i.e. the same number of cases was simulated for hypothesis H1a as for H1b, H1c, and H1d, respectively. We demonstrated that when more information was added to the case, the error decreased, especially exclusion error (Table 1).

The use of an inappropriate allele frequency database had only a minor influence on the total error rate, but was shown to have a considerable impact on individual LR.

When dealing with cases where there is an expected risk of having a relative of the TF as the AF, it is essential to include a computational model for treating inconsistencies. When there is only a limited number of inconsistencies between the AF and the child, the question arises whether or not these are due to mutations, or are true exclusions. The recommended way of handling such cases is to include all loci in the calculation of the total LR (Gjertson et al., 2007), although some labs still use a limit of a maximum number of inconsistencies for inclusion/exclusion (Hallenberg & Morling, 2009). However, we demonstrated that it is better to use a probabilistic model, even if the interpretation is not totally correct, than not to employ one at all (Table 1).

Furthermore, we proposed and tested a five hypotheses model in order to reduce the risk of falsely including a relative of the TF as the biological father. The simulations revealed that utilisation of such a model significantly decreased the error rates, although the magnitude of the decrease was minor.
Investigations

The use of DNA analysis to clarify relationships for the purpose of family reunification is increasing, and the evaluation of the statistical methods used is important. In this paper we demonstrated that improvements are still necessary in order to reduce the risk of erroneous conclusions in immigration casework.

| 35 |
Investigations

Paper II - Homogeneity in mitochondrial DNA control region sequences in Swedish subpopulations

In forensics, mitochondrial DNA is mainly used in casework where a limited amount of nuclear DNA is present, or when a maternal relationship is questioned. In the case of haploid DNA markers, it is extremely relevant to set up and study regional frequency databases due to an increased risk of local frequency variations (Richards et al., 2000). In this study we analysed mtDNA sequence variation in a Swedish population sample in order to facilitate forensic mtDNA testing in Sweden.

Materials and methods

Blood samples from 296 Swedish individuals from seven geographically different regions were typed, together with 39 samples from a Swedish Saami population (i.e. Jokkmokk Saami), for the complete mtDNA control region (Figure 4). This hypervariable segment (e.g. HVS-I, HVS-II and HVS-III) spans over 1100 nucleotides.

Haplotype- and haplogroup frequencies were calculated and interpreted from the DNA sequence variation. The statistical evaluation involved enumeration of forensic efficiency parameters as well as comparison of the genetic variation found in the Swedish regions and between the Swedish, other European and non-European populations.

Results and discussion

Two hundred and forty seven different haplotypes were found among the typed Swedes. This represents a haplotype diversity of 0.996 and a random match probability of 0.5%, which are in the same magnitude as for other European populations (Budowle et al., 1999b). Comparing mtDNA haplogroup frequencies with corresponding frequencies for 20 world-wide populations grouped the Swedes with other western European populations. This was further confirmed when calculating pairwise $\Phi_{ST}$-values for a limited number of geographically close populations (Figure 4).

The mtDNA sequences were further analysed in order to study potential substructure within Sweden, as indicated by an earlier study of the Swedish Y-chromosomal variation (Karlsson et al., 2006). MtDNA haplotype frequencies from the eight different Swedish regions were compared and only the Saami population differed significantly from the rest. The difference found for Y-chromosomal data between the northern region, Västerbotten, and the rest of Sweden was not observed in the mtDNA data. This can most probably be explained by demographic events. However, the impact of the relatively small sample sizes should not be ignored.
Investigations

Figure 4. Descriptive statistics for the Swedish mtDNA haplotype database (Saami excluded). The values in parentheses are for the Saami population. GD is the gene diversity (or haplotype diversity), PM is the match probability and FST the frequency variation among the seven Swedish subpopulations. The FST distance for the Saami population represents the genetic distance between the Saami and the seven Swedish regions combined. The $\Phi_{ST}$ represents the genetic variability between the seven Swedish regions combined and the German, Finnish and Norwegian populations, respectively. The Swedish regions studied were as follows; Västerbotten (1), Värmland (2), Uppsala (3), Skaraborg (4), Östergötland/Jönköping (5), Gotland (6), Blekinge/Kristianstad (7) and Jokkmokk Saami (8).

When estimating the population frequency for a given mtDNA haplotype, it is crucial to have a large representative reference database. The high similarity between the Swedish and other western European populations allows the inclusion of Swedish mtDNA data in initiatives like the EMPOP (The European DNA profiling group (EDNAP) mtDNA population database) (Parson & Dür, 2007), thus providing a more accurate estimate of the rarity of a mtDNA sequence.
Investigations

Paper III - Analysis of linkage and linkage disequilibrium for eight X-STR markers

X-chromosomal markers are useful for deficiency relationship testing (Szibor et al., 2003). The X-chromosome occurs in one copy in males and in two in females, and the combined use of several X-chromosomal DNA markers requires consideration of linkage and linkage disequilibrium. Thus, prior application of X-STRs in casework testing it is important to study population frequencies, substructure and efficiency parameters but also to explore marker specific recombination rates and allelic association among the loci. In this study, we focused on eight X-STRs located in four linkage groups and their usability in relationship testing.

Materials and methods

718 males and 106 females from a Swedish population were studied and analysed for the eight X-chromosome STR markers included in the Argus X-8 kit (Biotype) (Figure 5). From these, data were retrieved for establishing haplotype frequencies, for performing the LD test and for estimating forensic efficiency parameters. Family data from 16 Swedish families (3-7 children) and 16 Somali families (2-9 children) resulting in 84 to 116 informative meioses were considered in order to estimate recombination frequencies. Furthermore, a model for estimating such recombination rates was presented.

Results and discussion

Diversity measurements and efficiency parameters revealed that the “first” linkage group (DXS10135-DXS8378) was the most informative, although only minor differences were seen among the linkage groups. The linkage disequilibrium test resulted in significant p-values for the pair of loci within each of the four linkage groups. Thus, for the Swedish population, the loci should be treated as haplotypes rather than single markers. By means of simulations we demonstrated that, when LD was disregarded as opposed to taken into account, the average difference in calculated LR was small, although in some individual cases it was considerable.

Recombination frequencies for the loci were established based on the family data (Figure 5). These indicated that the chance of recognising a recombination within each linkage group was small (< 1%) and that there is also a tendency for linkage between groups three (HPRTB-DXS10101) and four (DXS10134-DXS7423).
The eight X-STRs investigated in this work have previously been widely studied by other groups, mostly with regard to population allele frequencies. For relationship testing, however, it is also important to study genetically relevant properties such as linkage and linkage disequilibrium. In this paper, our results indicated that such features cannot be ignored when producing the evidential weight of the X-chromosomal profiles in relationship testing.
Paper IV - Using X-chromosomal markers in relationship testing: How to calculate likelihood ratios taking linkage and linkage disequilibrium into account

The findings in Paper III obliged us to study and develop a mathematical model for how to consider both linkage and linkage disequilibrium when calculating the likelihood ratio in relationship testing for X-chromosome data. Traditionally, the Elston-Stewart model (Elston & Stewart, 1971; Abecasis et al., 2002) could be used for computing the LR involved in questions about a relationship. However, this model is not efficient enough for dealing with data from multiple linked loci. On the other hand, the Lander-Green algorithm (Lander & Green, 1987) works perfectly well for thousands of linked markers, but assumes that the loci are in linkage equilibrium. Efforts have, however, been made to treat groups of loci in LD (Abecasis & Wigginton, 2005), although we found that this approach was not fully satisfactory for our purpose. Therefore, we here present a model for the complete consideration of linkage and linkage disequilibrium and study the impact of taking and not taking linkage and LD into account, by means of a simulation approach on typical pedigrees and X-chromosomal data for the markers studied in Paper III.

Materials and methods

A computational model was presented based on the Lander-Green algorithm, but extended by expanding the inheritance vectors to consider all of a haplotype’s loci.

Six different cases, representing pedigrees for which X-chromosomal analysis would be valuable, were studied in order to test the model. Three of these involved cases where DNA profiles were available for both the children and the founders (e.g. questions regarding paternity for a trio, paternity for half-sisters, and paternity for full-sisters) and three additional cases where DNA profiles were only available for the children (paternity for half-sisters, paternity for full-siblings and maternity for brothers). Simulations were performed for each of the six cases that considered the tested relationship to be true or not true. From these, LR distributions were studied together with comparisons of the calculated LR using our proposed model, with LR calculated by means of simpler models with no or only partial consideration of linkage and LD. Genotype (or haplotype) data were simulated from Swedish population frequencies for the eight STR markers studied in Paper III.

Results and discussion

The model for the likelihood calculation was adapted to the six different cases. The simulations showed that the median LR for the three cases where DNA
profiles were available for the founders was high (~10^6) in comparison with the considerably lower median LR for the cases where genotype data for the founders were not available (~10^-2-10^3). Furthermore, various degrees of positive LRs were obtained when the questioned relationship was simulated not to be true, although only in the cases where founder profiles were not available.

We then compared the LRs obtained using our proposed model with LRs from two simpler models. The difference was on average small, although somewhat larger for the model in which linkage and LD were not taken into account (Table 2). In some of the tested cases, the estimation of the rarity of a given haplotype had a strong impact on the calculated likelihood ratio. This was especially true when estimating the haplotype frequency for haplotypes earlier not seen.

In summary, we demonstrated that in order to reduce the risk of incorrect decisions, linkage and linkage disequilibrium should be properly accounted for when calculating the weight of evidence and we proposed an efficient model to accomplish this.

<table>
<thead>
<tr>
<th>Rel</th>
<th>LR</th>
<th>Log10 Median (95% cred.) (min;max)</th>
<th>Diff</th>
<th>LR (m_2)/LR(m_1) Median (95% cred.) (min;max)</th>
<th>Diff</th>
<th>LR (m_3)/LR(m_1) Median (95% cred.) (min;max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rel</td>
<td>2.0</td>
<td>2.0 [-1.4;7.3] (0.001;1418)</td>
<td>1.2</td>
<td>0.36 [-0.3;7.3] (0.0001;438)</td>
<td>0.4</td>
<td>0.2 [-0.036;9.3] (0.0001;438)</td>
</tr>
<tr>
<td>NoRel</td>
<td>-1</td>
<td>2.0 [-1;4.0] (0.002;6.3)</td>
<td>1.0</td>
<td>[-1;1.0] (0.002;6.3)</td>
<td>0.2</td>
<td>[-1;0.026;433]</td>
</tr>
</tbody>
</table>

Rel means simulation where the brothers have the same mother and noRel that the brothers were simulated to have different mothers. 10,000 simulations were performed for each situation. M_1 indicates the model in which both linkage and LD were considered, m_2 the model in which linkage but not LD was considered and m_3 the model where linkage and LD were not taken into account.
Concluding remarks

Several parameters influence the assessment of the weight of evidence in a DNA investigation, and each of them can have a considerable impact on the resulting figure. This is a common feature of the four papers included in this thesis, the aim of which was to study relevant population genetic properties and models for considering them when calculating likelihood in relationship testing.

- In Paper I, we showed how the risk of erroneous decisions in relationship testing in immigration casework was affected by parameters such as the number of markers tested, utilisation of relevant population allele frequencies, use of a probabilistic model for the treatment of single genetic inconsistencies and consideration of alternative close relationships between the alleged father and the true father. In addition, we proposed methods for reducing the risk of erroneous decisions.

- In Paper II, we set up an mtDNA haplotype frequency database for the Swedish population. We demonstrated that the mtDNA variation in the Swedish population is high and that the homogeneity among different subregions within Sweden supports a combined Swedish population frequency database. Furthermore, the resemblance of the mtDNA variation found in Sweden compared with other European populations makes it possible to enlarge the relevant reference population, thus increasing the reliability of the estimation of the rarity of a given mtDNA haplotype.

- In Paper III, we studied eight X-chromosomal markers in terms of their informativeness and usefulness in relationship testing. We found that the markers located in each of the four linkage groups were in linkage disequilibrium and that the linkage within and between the linkage groups for the Swedish population highlighted the need to consider such parameters when producing the evidential weight for X-chromosomal marker investigations. Thus, when considering the Swedish population, the commonly used product rule for employing the eight X-STRs is not valid.
Concluding remarks

- In Paper IV, we presented a model for the calculation of likelihood ratios, taking both linkage and linkage disequilibrium into account, and applied it on simulated cases based on DNA profiles with X-chromosomal data. We revealed that X-chromosomal analysis can be useful for choosing between alternative hypotheses in relationship testing. Furthermore, we showed that our proposed model for proper consideration of both linkage and linkage disequilibrium is efficient and that disregarding LD and linkage can have a considerable impact on the computed likelihood ratio.
Future perspectives

It is always difficult to predict the future, as things never turn out as we imagine. Nevertheless, I will attempt to make some intelligent guesses concerning the future of forensic genetics. If we look back at the developments in this field, we can see enormous progress over the past 25 years, much of which is thanks to the discovery of polymorphic genetic markers, thermostable polymerases as well as the invention of PCR and capillary electrophoresis, making DNA analysis faster, cheaper and more accurate. We should also not forget the valuable information provided by the HUGO-project and the assistant from the developments of fast computer power.

If we go back to the time just before the above-mentioned findings and adopt a client’s point of view, we can see that the chances of linking a suspect to the crime scene or solving relationship issues quickly and with a high degree of accuracy were rather small. Thus, the demand from clients to resolve these issues was a major factor behind the high degree of development. Once the discoveries were made, an incredible amount of dollars and scientific hours were invested in forensic genetics in order to meet the demands.

So, is it likely that we will see similar trends over the next 25 years? I do not really think so. One reason is that today, many cases can be solved with existing methods and major investments have already been made in forensic laboratories regarding instrumentation and the large national DNA-profile databases, which consist of a given number of predefined DNA markers (e.g. CODIS (www.fbi.gov/hq/lab/codis/clickmap.htm)). In the case of the national databases, which like CODIS consists of over 8,000,000 profiles, it would be very difficult to replace the existing markers with a different set up of markers.

Generally, I believe that future development will become more diversified, which means less money and time spent on each topic. With regard to the use of DNA in criminal cases, perhaps the most important areas concern methods for obtaining complete DNA profiles from difficult biological samples (Budowle et al., 2009), resolving mixtures (Homer et al., 2008) and predicting an individual’s physical appearance such as eye and skin colour (Kayser & Schneider, 2009). To this should be added the constant upgrading of analysis instruments, which will make DNA-profiling faster and of an even higher quality.

If we turn to the use of DNA in relationship testing, which is the main topic of this thesis, methods and markers have traditionally been adapted from developments in the much larger field of DNA in criminal cases. In terms of fu-
Future perspectives

ture developments, I feel that there will be a greater focus on statistical methods to handle data from an increased amount of DNA markers. However, I think that these developments will, in general, be less compared to the last 25 years. Take, for example, paternity testing, which constitutes the vast majority of all relationship testing cases. Today, such cases are mostly solved with a sufficiently high probability and certainty, thus the community will not continue to invest the same level of time and money in research. However, relationship questions still remain that cannot be resolved in a satisfactory manner with existing methods, for example, separating full-siblings from half-siblings and half-siblings from unrelated individuals (Nothnagel et al., 2010). Some attempts have been made to solve similar cases as well as more distant relationships by means of array SNP data (Skare et al., 2009, Ge et al., 2010). The similarity between these studies is the simultaneous use of a large number of markers. What becomes important in such cases are the statistical methods required for accurate estimation of the evidential value of the DNA profiles, taking properties such as linkage and linkage disequilibrium into account.

When it comes to the continuum efforts regarding the linked X-STRs studied in this thesis, there are mainly two things that remain to be investigated. One is to transform the model presented in Paper IV into user-friendly software, which can be applied to various sets of DNA marker systems and different sets of pedigrees. The other issue that requires extensive study is the estimation of haplotype frequencies. As more markers are included in a haplotype block, much larger population databases are required in order to obtain an estimate of its frequency with a low level of uncertainty.

Furthermore, we will continue to explore the genetic diversity of the Swedish population with a special focus on northern Sweden. Historical data have shown that there are regions that have been rather isolated until early 1900s, and we are interested to see if this is traceable in the genetics, thus could have an impact when producing the weight of evidence.

Another area that needs further improvement concerns the use of DNA for disaster victim identification (Prinz et al., 2007). This involves post-mortem DNA analysis, data management and methods for statistical interpretation.

Finally, one should not forget the incredible progress made with the so-called next generation sequencing methods (Metzker, 2010). It is now possible to sequence a complete human genome for under $100,000. This would allow access to a huge amount of information for predictions. However, although the methods are technically suitable for producing such large data sets, the downstream efforts for data management and statistical interference still require a great deal of work (Metzker, 2010).
Acknowledgements

Tack till!

All studies performed in this thesis were conducted at National Board of Forensic Medicine, Department of Forensic Genetics and Forensic toxicology.

Scientific work is never a one-man show and I am grateful for all help, support and encouragement I have received through the years. Especially I would like to thank:

My tutors Professor Bertil Lindblom, Associate professor Gunilla Holmlund and Associate professor Petter Mostad for sharing your deep knowledge within the field. Thanks for your guidance, encouragements and always having your doors open for discussions. Also, Bertil, especially for your deep knowledge and wide expertise within forensic genetics. Gunilla, who introduced me to the field of forensic genetics, for your positive spirit. Petter, especially for all your support on mathematical and statistical issues.

The staff in the “RG-team” Helena, Kerstin, Birgitta, Ann-Britt, Kersti, Johnny, Anna-Lena, Ulla, Irene, Margareta, Agnetha, Pernilla and Louise for all support big and small, and for making Artillerigatan 12 a fun place to work.

Johan Ahlner, the head of the Department of Forensic Genetics and Forensic Toxicology, for being an inspiring boss. My fellows in the “Jippo-gruppen” for fun moments, and all other people at RMV-Linköping who has been helpful in different ways.

Thore Egeland, Anders Götherström, Thomas Wallerström, Helena Jankovic Malmström, Michael Coble and Jukka Palo for valuable scientific collaborations and discussions, which I hope will continue.

As a researcher you need to discuss ideas big or small, and as a PhD-student it is always nice to have a “bollplank” for practical issues. I thank Jenny Velander, Daniel Kling, Gunnel Nilsson and Anna-Lena Zachrisson for inspiring discussions and practical tips.
As a scientist it is valuable to be able to think about other stuffs beside the work.

I would like to thank Magnus and Erik for always being such good friends.

Leonard, Claes, Ragnar, Sara and Evelin for the years in Uppsala.

My father Jan-Ake, mother Inga-Lill, sister Frida and brother Martin for your endless support and for believing in me during my life.

My relatives Ingvar & Eivor and grandfather Olle for always being around when help is needed.

My family in-law, Tillmar. It is always nice to have such a welcoming place to go to.

Finally, my beloved wife Ida and child Signe. Ida, for your constant support, love and caring. Thanks for your patient when I work at non-office hours. Signe, our wonderful daughter, for being constantly (almost) happy and for your help getting me up in the mornings during the writing of this thesis.

Thanks to all others who I not mentioned but has been helpful in different ways.

This work was supported by grants from National Board of Forensic Medicine and Center of Forensic Science, Linköping University.


Coble, MD, Just, RS, O’Callaghan, JE, Letmanyi, IH, Peterson, CT, Irwin, JA and Parsons, TJ (2004) Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int J Legal Med 118:137-146.

References


References


