Statistical evaluation of forensic DNA sequence profiles

Sanne E. Aalbers

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:
Bruce S. Weir, Chair
Sharon R. Browning
Scott R. Kennedy

Program Authorized to Offer Degree:
Public Health Genetics

University of Washington

Abstract

Statistical evaluation of forensic DNA sequence profiles

Sanne E. Aalbers

Chair of the Supervisory Committee:

Bruce S. Weir

Department of Biostatistics

DNA evidence has revolutionized criminal investigations and has played a crucial role in thousands of forensic cases. Nowadays, DNA typing is a mature field and overwhelmingly seen as the gold standard in forensic science. The interpretation of DNA evidence, however, is far from straightforward and challenges arise when evaluating complex profiles and assessing the statistical weight of the evidence. With the introduction of next generation sequencing (NGS) technologies a new dimension has been added to the field, providing distinct advantages over traditional DNA profiling methods in terms of captured information. As a consequence, there is a need to re-evaluate existing statistical models and underlying parameters to facilitate DNA evidence evaluations for forensic sequence profiles. This dissertation examines the statistical impact of NGS data and the perceptions of their use within the forensic science community. In Chapter 1, we provide sequence-based estimates of population genetic parameters, including population structure, relatedness, and inbreeding estimates for forensic autosomal markers. We found similar effects of sequence data on population genetic estimates as what has been seen for traditional forensic data. Resulting estimates for sequence-based data are also similar to estimates obtained by restricting to allele number designations. Although the increase in polymorphism observed in sequence data implies that matching proportions will decrease, we observed that population genetic estimates may increase or decrease, depending on where the extra sequence variation occurs. We also detected a positive relationship between inbreed-

ing and average kinship for forensic markers, highlighting the importance of incorporating appropriate population genetic parameters into forensic calculations. Since ignoring allelic dependencies has the potential of being prejudicial to the suspect, we recommend estimating such parameters using genotypic rather than allelic data. Chapter 2 demonstrates the effect of sequence data on match probabilities, a measure integral to DNA evidence evaluations. By comparing empirical matching proportions to expected values under the assumption of independence, we found that results for autosomal markers become less conservative the more markers we include. Results also showed that the violation of the independence assumption is exacerbated for sequence-based data as compared to length-based data. We discuss how to compensate for multi-locus dependencies by using an appropriate theta value, and that these values may differ between marker systems. We furthermore employed entropy measures to describe associations between markers for autosomal as well as Y-chromosome profiles, and saw that there exists a diminishing return on adding more markers beyond a certain point and that this point may be reached sooner for sequence data. We also examined the extent of independence of match probabilities across systems and observed that appropriate correction factors are required for both systems to obtain conservative estimates for multi-locus matches across different marker systems for autosomal data. Finally, Chapter 3 presents the results of a qualitative study involving sixteen U.S.-based forensic scientists working with DNA evidence. We conducted semi-structured interviews to assess if this group of professionals feels equipped to interpret and present DNA evidence for sequence data. Results highlighted eleven different themes that describe the views and needs of study participants surrounding the use of statistical models and sequence data for forensic purposes. On the topic of statistics, considerations include the need for ongoing training and education and improvement of presentation of results in court. When it comes to NGS considerations, participants saw huge potential for numerous applications, but also expressed concerns about existing practical barriers as well as potential ethical implications of sequence data for forensic applications. Overall, we believe that these perspectives, along with statistical considerations as addressed in the other chapters, need to be taken into account in our journey towards successfully implementing sequencing methods for DNA evidence evaluations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

If you would have asked me a couple of years ago, I could not have imagined I would currently be writing down the final words of my PhD dissertation. I am thankful for the opportunity to continue my education and to pursue this degree. I believe this program made me a better researcher by equipping me with a well-rounded perspective and I hope this dissertation can attest to that.

I would like to thank my committee members for their unwavering support and trust in my capabilities to successfully complete this project. I especially thank my academic mentor and committee chair, Bruce Weir, for his endless support and amazing mentorship along the way. I am so grateful for the opportunities you have given me over the years and your generosity in creating a supportive research environment throughout my stay at the UW. I greatly admire your kindness and humility, as well as your ability to provide reassurance when I needed it most. I am honored to have my name appear next to yours on the work we have done over the past years, and hope to see it soon on our book as well. I would also like to extend my gratitude and respect to the late Dr. Deb Bowen. Deb helped me during the initial stages of my qualitative analysis and gave me the confidence to carry out the part of my research I felt most insecure about. It saddens me that she did not see this work completed.

This research would not have been possible without dedication and input from numerous sources. I thank the National Institute of Justice for funding my research through their Graduate Research Fellowship. I thank the International Symposium on Human Identification for their help with recruitment for our qualitative study, and Mike Coble and Hanley Kingston for their help in preparing for the interviews. I thank all forensic practitioners who participated in this research for sharing their time and experiences with me. Thank you Alyna Khan, for agreeing to serve as my second coder on a project where I sometimes felt out of my depth. I

appreciated our collaboration during a time that can feel quite lonely and am very proud of our resulting work.

Thank you to the Institute for Public Health Genetics community and particularly Annique Atwater for answering my many questions over the years and making sure that any logistical issues were taken care of. Thank you to all the former and current students, including Nandana, Alyna, Hanley, EJ, and Diane, who made my time at the institute much more enjoyable and doable during these past couple of weird years.

Finally, I would like to thank everyone for continuing to make us feel at home in Seattle. Bruce and Beth, thank you for welcoming us the moment we arrived in the U.S. To our dear friends Sue and Lee, thank you for always being there for us; I am looking forward to our summer adventures. And most of all, thank you to my husband Lars. This dissertation is dedicated to you.

June 15, 2023

INTRODUCTION

Forensic DNA typing

Forensic DNA interpretation is currently centered on the analysis of micro-satellites, or short tandem repeats (STRs), i.e. short DNA sequences of $2-4$ base pairs that are repeated several times, often ranging from $10-20$ repeats. These repeat patterns are located in areas called loci and vary among individuals. Variants for a given locus are called alleles and they can be represented as a DNA profile obtained from biological material. Such profiles can then be compared between evidence and suspect or to existing DNA databases in an effort to implicate or exclude a person, or their relative, with a crime. These findings are usually accompanied by some expression on the strength of the evidence based on population genetic quantities and statistical models.

Initial DNA profiling technologies required relatively large amounts of DNA and were not suitable for degraded samples. The ability to increase the amount of DNA in a sample by the polymerase chain reaction (PCR) was of substantial benefit to forensic science [52]. Current STR typing methods combine the PCR process with capillary electrophoresis (CE) to gain access to the allele numbers contained in a DNA sample. During PCR amplification, primers labeled with fluorescent dye bind to DNA molecules within the region of interest. The PCR products are injected into a capillary where they travel in the direction of a positive charge. Since the time taken to pass a fixed point depends on fragment size it can be used to infer the number of repeats. The primers are detected by a sensor measuring the relative fluorescence units (RFU). Allelic ladders are used to determine allele designations, with integer values indicating the number of complete repeat motifs and additional nucleotides (i.e. incomplete repeats) separated by a decimal point [57]. The resulting DNA profile can be visualized with

an electropherogram, showing allele numbers along with their respective peak intensity in RFU.

More recent advances in DNA sequencing technologies allow us to not only reveal the number of repeats, but also the underlying base-pair structure of each repeat. Next generation sequencing (NGS) techniques therefore have the ability to reveal additional variation within the STRs, leading to an increase in discriminating power [8]. Table 1 shows some examples of length-based (LB) versus sequence-based (SB) allele callings that can be obtained through CE methods and NGS methods, respectively. While alleles of the same length will be recorded as matching using standard CE-based methods, even if they differ at the sequence level, NGS techniques are able to distinguish these so-called isoalleles [60]. An additional advantage of NGS, also called massively parallel sequencing (MPS), is that such platforms are capable of simultaneously producing data on a combination of different marker systems, including autosomal STRs, Y-chromosome STRs, X-chromosome STRs, and several single nucleotide polymorphism (SNP) marker sets, depending on the primer mix used [6].

Locus	**Allele number**	**Allele sequence**
D3S1358	15	$\text{TCTA}[\text{TCTG}]_2[\text{TCTA}]_{12}$
D3S1358	15	$\text{TCTA}[\text{TCTG}]_3[\text{TCTA}]_{11}$
D18S51	20	$[\text{AGAA}]_{20}$
D18S51	20	$[\text{AGAA}]_{16}\text{GGAA}[\text{AGAA}]_3$

Table 1: Examples of length-based (LB) versus sequence-based (SB) allele callings for two forensic markers.

Because sequence profiles allow the determination of the number of repeat units, they offer backward compatibility with CE-based STR profiles and the use of existing DNA databases [60]. This is advantageous as STR analysis has been well established in the forensic community and the Combined DNA index System (CODIS) as maintained in the U.S. currently contains over 20 million profiles[1]. However, CE-based and NGS-based methods differ in their underlying technique, such that existing models for DNA evidence evaluations need to be re-examined and adapted to facilitate sequence data.

[1]See the Federal Bureau of Investigation CODIS - NDIS Statistics for the most recent numbers: `https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics` (Accessed June 14, 2023).

NGS considerations

Most commonly, NGS techniques separate PCR products into single stranded fragments that are chemically attached to a slide [41, 6]. Instead of using primers tagged with fluorescent dye, fluorescently labeled nucleotides are added, showing a different color for each of the four different bases that can be observed via a laser. These nucleotides also have a terminator modification that allows the addition of only one base at a time and the reading of its fluorescence to determine which nucleotide was inserted. In contrast to the CE method, this technique allows for retrieving the STR motif at sequence level. NGS methods may reveal variation in the STR repeat region as well as the flanking region adjacent to the repeat motif. Flanking region SNPs (flSNPs) can be exploited by NGS techniques to increase the discrimination power of some STR loci even further [28]. Knowledge of such variants can be utilized in primer design to ensure optimal positioning during the PCR process [55].

By far the biggest player in the field of sequencing instruments is Illumina, a platform that characterizes itself through a wide range of instruments for short-read sequencing varying from low to ultra-high throughput [41]. It is also suitable for paired-end reads, i.e. the amplification of DNA template in two directions, such that a relatively long reading length can be reached. NGS techniques produce sequence data in the form of discrete read counts as opposed to continuous peak height data for CE-based methods. The number of sequence reads that are generated during a DNA sequencing reaction depend on the platform and instrument. As sequence reads are not distributed equally across a DNA target region, differences occur in the number of reads that cover a particular base. The coverage, or depth of coverage (DoC), refers to the average number of times a single base is read during a sequencing run [41]. Literature suggests a minimum of 650 to 1000 reads per locus for forensic applications [84, 40]. When using a 20 loci amplification kit, this amounts to 20 000 reads per sample to obtain an average of 1000 reads per locus. Illumina's ForenSeq DNA Signature Prep Kit together with the MiSeq FGx platform are developed specifically for such forensic genomic applications [42]. This is a form of targeted sequencing, which focuses on a subset of genes or regions of interest instead of the entire genome [8].

With the emergence of NGS profiling in forensic science, there is an increasing demand for fast and efficient software packages suitable for high-throughput analysis. Several tools, capable of translating raw data resulting from sequencing platforms into information suitable for further analysis, are now available for STR loci commonly used in forensic casework [48]. At the same time, new statistical models need to be developed and implemented to accommodate resulting sequence-based DNA profiles, with the ultimate goal to establish a probabilistic genotyping approach for NGS mixture interpretation. Over the years, several studies and projects, including international collaborative efforts, have been initiated to facilitate DNA evidence evaluations for forensic sequence profiles [31, 32, 57, 58].

DNA evidence interpretation

A DNA profile obtained from a crime scene sample can be compared to the profile of a suspect or person of interest, and it may be found that this person cannot be excluded as a contributor on the basis of DNA evidence alone. An 'inclusion' may be reported, but is practically worthless without some expression on the strength of this evidence. Instead, the evidence is evaluated in light of two or more competing hypotheses and a forensic scientist is concerned with assigning a value to the likelihood ratio [24].

When applying likelihood ratios (LRs) to DNA profiles, the evidence generally consists of the DNA types of the crime scene sample and the person of interest. Letting G_C denote the crime scene profile and G_S the suspect profile we seek the following likelihood ratio, which may also be written as

$$\mathrm{LR} = \frac{\Pr(G_C, G_S|H_1, I)}{\Pr(G_C, G_S|H_2, I)} = \frac{\Pr(G_C|G_S, H_1, I)}{\Pr(G_C|G_S, H_2, I)}, \tag{1}$$

where H_1, H_2 correspond to the competing hypotheses and I denotes the background information, which is usually dropped for simplicity. For a good quality single-source sample, the numerator of the LR evaluates to $\Pr(G_C|G_S, H_1) = 1$, assuming that H_1 reflects the hypothesis that the person of interest is the source of the crime scene sample and $G_C = G_S$, or in

words, the crime scene profile and suspect profile match. The LR then simplifies to

$$\mathrm{LR} = \frac{1}{\Pr(G_C|G_S, H_2)} \tag{2}$$

The simplest alternative hypothesis H_2 in this case is that some other person is the source of the crime scene sample. If knowledge of the suspect's DNA type does not affect our uncertainty about the offender's type when they are different people (i.e. when H_2 is true), then $\Pr(G_C|G_S, H_2) = \Pr(G_C|H_2)$ and the likelihood ratio becomes the reciprocal of the *profile probability* of profile G_C. In this case we try to answer the question what the probability is that a person randomly chosen from a population will have the DNA type observed from the crime scene sample. On the other hand, the conditional probability $\Pr(G_C|G_S, H_2)$ is the probability of seeing the profile in a randomly chosen person after we have already seen that profile in a typed person (the suspect). This is the *match probability*. Confusingly, in a forensic setting the term *random match probability* (RMP) usually refers to the unconditional probability, i.e. the profile probability. Either way, a rich theory of population genetics is available to evaluate such probabilities.

Population genetics

The profile probability is usually a little easier to think about, although difficult to answer in practice. The current CODIS core set of autosomal STR markers consists of 20 loci and it is very unlikely that a certain DNA profile will be seen in any sample of profiles. If we consider a single autosomal STR locus with 10 alleles, there are 55 different genotypes, 10 homozygote types and 45 heterozygote types, and it is quite likely that several of those will not be seen in a sample of a few hundred individuals, even though all alleles are seen. A solution to this problem is to assume that Hardy-Weinberg equilibrium (HWE) holds. This is a basic principle in population genetics, described in 1908 by English mathematician Hardy and German physician Weinberg. The Hardy-Weinberg law assumes infinitely large populations and random mating, with no disturbing forces such as selection, mutation, or migration that would change allele frequencies over time [24]. If the law holds, genotype frequencies are

products of allele frequencies π_u for alleles A_u:

$$\Pr(A_u A_u) = \pi_u^2$$
$$\Pr(A_u A_v) = 2\pi_u \pi_v$$

Instead of having one single, simple population, the human population consists of a number of subpopulations genetically different from each other. Allele and genotype proportions may in this case be estimated for each subpopulation separately, or for the total population. When allele proportions are different in the subpopulations, there will be a departure from HWE at the population level, even if there is equilibrium in each subpopulation. This phenomenon is known as the Wahlund effect and leads to departures of genotypic proportions from products of allele proportions. [5, 24]

Although the profile probability is interesting, it is the match probability that is of relevance in a forensic setting. This probability is bigger than the profile probability as the chance of seeing a particular profile increases after it has been seen already. The match probability requires calculations about pairs of profiles and can be constructed using the sampling formula [5]. If n alleles have been sampled in a subpopulation, of which m are of type A_u, then the probability that the next allele sampled is also of type A_u can be written as

$$\Pr(A_u | m \text{ of } A_u \text{ in } n) = \frac{m\theta + (1-\theta)\pi_u}{1 + (n-1)\theta},$$

where π_u is the probability an allele is of type A_u and θ is the probability of identity by descent (ibd) for two alleles drawn randomly from a population. This formula can be derived by assuming a Dirichlet distribution for the allele probabilities, or a beta distribution in case of biallelic data. Although this holds for some simple populations, it is unlikely that the Dirichlet assumption will be valid for STR loci. A step-wise mutation model would be more appropriate, instead of the infinite-alleles mutation model assumed here, and mating and migration patterns are generally more complex in practice [24]. Nevertheless, it gives a good approximation and is simple to evaluate.

With the sampling formula, we can construct the Balding-Nichols match probabilities,

which are now widely used in forensic applications [52]:

$$\begin{aligned}
\Pr(A_uA_u|A_uA_u) &= \Pr(A_u|A_uA_u)\Pr(A_u|A_uA_uA_u) \\
&= \frac{[3\theta+(1-\theta)\pi_u][2\theta+(1-\theta)\pi_u]}{(1+\theta)(1+2\theta)} \\
\Pr(A_uA_v|A_uA_v) &= \Pr(A_u|A_uA_v)\Pr(A_v|A_uA_uA_v) + \Pr(A_v|A_uA_v)\Pr(A_u|A_uA_vA_v) \\
&= \frac{2[\theta+(1-\theta)\pi_u][\theta+(1-\theta)\pi_v]}{(1+\theta)(1+2\theta)},
\end{aligned}$$

These expressions are also known as the "θ-correction" for autosomal profiles. The value of θ is specific to the type of data used. For Y-STR data, for example, the match probability of a haplotype A_u is $P(A_u|A_u) = [\theta + (1-\theta)\pi_u]$ and requires a multi-locus value of θ. In addition, SNP markers as observed with NGS data will have different mutation rates and, consequently, different θ values as well [78].

Since no two DNA profiles are truly independent, dependencies will affect match probabilities within as well as across forensic markers. Between-locus associations are known as linkage disequilibrium (LD) and loci are said to be linked if they are close together on a chromosome. Analogously, loci are said to be unlinked if they are on different chromosomes, or far apart on the same chromosome, which is generally the case for forensic (autosomal) markers [24]. Nevertheless, LD may occur in both situations, as a result of phenomena such as mutations, population structure and drift [20]. When calculating the statistical weight of DNA evidence, appropriate correction factors need to be incorporated to account for the effect of allele variations and dependencies within and between loci, as well as within and between populations.

Research motivation

Accurate representation of forensic evidence in court is crucial to avoid misinterpretations and, ultimately, to reduce the possibility of a miscarriage of justice. This not only requires sensible models that can handle the complexity associated with DNA profiles, but also an understanding of the methods used by forensic scientists who will be writing the reports and

potentially serving as an expert witness in court.

As sequence data as well as the data-generating processes are different from traditional typing methods, existing frameworks have become inadequate and need to be re-examined to facilitate the implementation of NGS methods [57, 58]. In recent years, several studies have reported NGS-based population genetic analyses and initial statistical models for phenomena relevant to NGS data [33, 73]. However, the methods applied to population genetic estimates in the most widely used forensic software tools are generally based on outdated models. In addition, we do not know if forensic scientists feel equipped to interpret and present DNA evidence obtained from sequencing methods.

In this dissertation, we focus on the impact of forensic sequence data on the evaluation of DNA profiles and the parameters underlying statistical models. Moreover, specific attention is given to the views and needs of forensic scientists concerning the (statistical) applications of sequence data for forensic purposes. The intent of this research is not only to focus on new theoretical analysis, but also to inform the forensic community in the move towards the implementation of sequencing methods for DNA evidence evaluations. This research is funded through NIJ's Graduate Research Fellowship with award number 2020-R2-CX-0040.

Dissertation outline

This dissertation contains a collection of papers examining the statistical impact of forensic sequence profiles and the perceptions of their use within the forensic science community. Since each chapter is written as a stand-alone manuscript, some overlap between certain sections can be expected. A brief summary of the chapters is given below.

Chapter 1 focuses on the estimation of population genetic parameters that are required for DNA evidence evaluations. The first part of this chapter contains preliminary work that provides estimates of population structure based on recent frameworks, comparing both length-based and sequence-based results. The second part of the chapter expands this work by providing estimates of population structure and relatedness for autosomal STR and identity-informative SNP data generated by sequencing technologies. We also discuss the effect of inbreeding on forensic calculations and why the use of genotypic-based estimates may be

preferred over allelic-based estimates. The overall goal of this chapter is to characterize the extent of dependencies within loci for sequence-based autosomal data.

Chapter 2 investigates the impact of sequence data on match probabilities for DNA evidence evaluations. The overall goal of this chapter is to characterize the extent of dependencies between loci for data generated by sequencing platforms. Here, we look at departures from independence within marker systems for autosomal STRs, Y-STRs, and identity-informative SNPs. We also look at dependencies between marker systems for the autosomal STRs and identity-informative SNPs and demonstrate how a theta correction can account for multi-locus dependencies. Finally, some attention is given to potential concerns regarding the practice of adding more markers with the intention to increase the discriminating power of a DNA profile.

Chapter 3 presents the results of a qualitative study involving sixteen U.S.-based forensic scientist working with DNA evidence. The goal of this chapter is to get an in-depth understanding of the current situation by assessing if this group of professionals feels equipped to interpret and present DNA evidence for sequence data. Here, we provide insight into the perceptions of forensic scientists in relation to statistical models, sequence data, and ethical implications for DNA evidence evaluations.

We conclude by synthesizing the contributions of our research findings and suggesting potential avenues for future work. We also provide some recommendations to be considered in our journey towards an effective implementation of statistical methods for sequence data.

Chapter 1

Population genetic parameter estimation for forensic DNA sequence data

Preliminary work

This section was originally published as Sanne E. Aalbers, Michael J. Hipp, Scott R. Kennedy, Bruce S. Weir. *Analyzing population structure for forensic STR markers in next generation sequencing data.* Forensic Science International: Genetics 49 (2020).

Abstract

Match probabilities calculated during the evaluation of DNA evidence profiles rely on appropriate values of the population structure quantity θ. NGS-based methods will enhance forensic identification and with the transformation to such methods comes the need to facilitate NGS-based population genetics analysis. If NGS data are to be used for match probabilities there needs to be a way to accommodate population structure, which requires values for θ for those data. Such estimates have not been available. This study assesses population structure for sequence-based data using a relatively new approach applied to STR data over 27 loci in five different geographic groups. Matching proportions between individuals or groups are used to obtain locus-specific θ estimates as well as estimates per geographic group and a global measure. The results demonstrate similar effects of sequencing data on θ estimates compared to what has been seen for CE-based results.

Introduction

Forensic DNA interpretation is currently centered on the analysis of short tandem repeats (STRs), relying on capillary electrophoresis (CE) to gain access to the allele types contained in a DNA sample. To evaluate such DNA evidence profiles, match probabilities can be calculated and these depend on appropriate values of the population structure quantity θ. It is common in forensic DNA evidence evaluations to use values of 1% - 5% [10].

With the introduction of next generation sequencing (NGS) more discrimination is provided through the ability of this technique to reveal variation within the STR. STR analysis has been well established in the forensic community so backward compatibility with CE-based STR profiles is needed to allow the use of existing DNA databases [57]. As long as this is the case, it is expected that NGS methods will continue to be implemented, stressing the need to facilitate NGS-based population genetics analysis.

In recent years, studies have reported population statistics demonstrating the increase in discrimination power by differentiating the nucleotide sequences of STR alleles with identical size [29, 54, 33]. Such statistics include allele frequencies, observed and expected heterozygosity, and tests for Hardy-Weinberg equilibrium and linkage disequilibrium. Freely accessible tools like STRAF [36] and Arlequin [25] provide a whole range of statistics, including F-statistics [82]. F-statistics, or more specifically values for F_{ST}, written here as θ, for NGS data, are required if sequence data are to be used for match probabilities. However, as with most published estimates of θ, F-statistics from these tools are produced using the Weir and Cockerham estimator [79] and a less restrictive estimator is recommended nowadays. This updated framework is detailed in Weir and Goudet [80] and applied to CE-based STR data in [10].

The Scientific Working Group on DNA Analysis Methods (SWGDAM) reported in an addendum from April, 2019, that "Currently, guidance does not exist regarding θ values for sequence-based data; therefore the existing NRC II guidance should be followed (NRC II 4.4a, where typically $\theta = 0.01$ for most U.S. groups or 0.03 for some isolated populations)." [60]. This paper addresses this gap.

Materials and methods

Estimation of θ

The parameter θ is needed for the Balding-Nichols [5] expressions for match probabilities. The probability an untyped person has homozygous genotype AA when a different person in the same population has been found to have the same type, for example, is

$$\Pr(AA|AA) = \frac{[3\theta + (1-\theta)\pi_A][2\theta + (1-\theta)\pi_A]}{(1+\theta)(1+2\theta)} \quad (1.1)$$

Here θ is specific to the population to which the two people belong and π_A is the probability an allele is of type A. Equation 1.1 can be used to assign a probability for an unknown perpetrator having the evidence profile AA after a suspect has been found to have that type.

The motivation for Equation 1.1 is that alleles within a population may have some dependencies because of shared evolutionary history. These dependencies will be small in populations with large population sizes and long histories, such as an African population, where mutation has had many opportunities to reduce the equality of different alleles. Allelic dependencies will be great in populations with small population sizes and few founders, such as Native American populations, where many alleles have the same ancestral allele type. The dependence of alleles in the same population is described by θ, the probability of two alleles taken randomly from a population are identical by descent (ibd), meaning that they are both copies of the same allele in an ancestral reference population. Larger θ values increase the probability of a person's genotype once that genotype has already been seen.

There are two problems with implementing Equation 1.1: neither θ nor π_A is known. It is not generally possible to specify the relevant population for a particular situation, so that population cannot be sampled to directly observe the proportion of pairs of individuals with the same genotype, or estimate matching probabilities from the equation. Instead, use is made of databases representing many populations, generally for a single continental ancestry. A single database by itself, however, does not provide information about the variance in allele frequencies among populations so that it does not indicate how different allele frequencies may be among populations.

The sample frequencies $\tilde{p}_A$ for alleles A in a large database are good estimates of the probabilities π_A for the populations represented in the database but there is a variance of $\tilde{p}_A$ around π_A. In Appendix A we show that $\text{Var}(\tilde{p}_A) = \pi_A(1-\pi_A)\theta_B$, where θ_B is an average of the probabilities of two alleles, one from each of two populations represented by the database, being ibd. This means that if $\tilde{p}_A$ is used instead of π_A in Equation 1.1, then the expression is estimating something that depends on θ_B. Buckleton et al. [10] offered a work-around to this problem, by introducing the average θ_W of population-specific θ values and using $\beta = (\theta_W - \theta_B)/(1-\theta_B)$ instead of θ in Equation 1.1 when $\tilde{p}_A$ is used instead of π_A. The parameter β is the probability of two alleles in one population are ibd, relative to the probability of alleles in different populations are ibd. There is no need to specify the ancestral reference population implicit in the definition of ibd, and there is no requirement that β is positive. It was estimates of β that were given by Buckleton et al. [10], and are given here for NGS data.

Buckleton et al. [10] adopted two sampling frameworks: global and single continental ancestry. In the second case, a set of populations with similar ancestry, such as "European", was used to estimate θ for that ancestry with the thought that it would provide guidance to a forensic analyst who wished to use allele frequencies from their own European ancestry database to estimate match probabilities with Equation 1.1. The other framework used data from all available ancestries, and that is the framework we use here as we had limited data within each ancestral group.

A formal justification for the Buckleton et al. [10] procedure for implementing Equation 1.1 is difficult to give, but a related situation is quite straightforward. The probability two alleles taken randomly from a random-mating population are both of type A is

$$\Pr(AA) \quad = \quad \theta\pi_A + (1-\theta)\pi_A^2$$

If this equation is averaged over a set of populations, then θ is replaced by its average, θ_W, over populations and π_A is not changed as this probability is assumed to hold for all the populations. If π_A is replaced by a database frequency $\tilde{p}_A$, then an unbiased estimator of the

average $\Pr(AA)$ is

$$\widehat{\Pr(AA)} \quad = \quad \beta\tilde{p}_A + (1-\beta)\tilde{p}_A^2$$

This expression follows from the expresion above for the variance of $\tilde{p}_A$ and it applies as an average for any population represented by the database, provided the database is large.

Estimation of β can be based on allele counts from a set of at least two populations, as implied in the discussion of the variation of allele frequencies $\tilde{p}$ about the probabilities π, or it can be based on genotype counts to allow for departures from Hardy-Weinberg equilibrium (HWE) in sample data. For allelic data, the estimates are written as $\hat{\beta}_{WT}$ and are functions of sample proportions $\tilde{M}$ of pairs of alleles that match, corresponding to probabilities of identity by descent θ. Starting with allele counts n_{iu} of allele A_u sampled from population i, the within-population sample matching proportion is $\tilde{M}_W^i = \sum_u n_{iu}(n_{iu}-1)/[n_i(n_i-1)]$, where $n_i = \sum_u n_{iu}$. For populations i and i', the between-population sample matching proportions are $\tilde{M}_B^{ii'} = \sum_u n_{iu}n_{i'u}/(n_i n_{i'})$. For sets of r populations, averaging over populations of within-population allele matching proportions gives $\tilde{M}^W = \sum_i \tilde{M}_W^i/r$, and the average over pairs of populations of between-population matching proportions $\tilde{M}^B = \sum_{i\neq i'} \tilde{M}_B^{ii'}/[r(r-1)]$.

Population-specific θ measures for allelic data can then be estimated relative to allele matching proportions between populations as $\hat{\beta}_{WT}^i = (\tilde{M}_W^i - \tilde{M}^B)/(1-\tilde{M}^B)$. An overall estimate for allelic data is obtained as $\hat{\beta}_{WT} = (\tilde{M}^W - \tilde{M}^B)/(1-\tilde{M}^B)$. These are locus-specific estimates, which are expected to vary among loci. The average β estimates over loci are calculated as the ratio of averages of numerators and denominators rather than the average of ratios, with the former leading to smaller variances. The reader is referred to [10, 80] for a more detailed discussion on this approach. The overall β estimates are used in Equation 1.1.

Equivalent genotypic expressions define within-population sample matching between individuals j and j' in population i as $\tilde{M}_{jj'}^i = \sum_u X_{ju}^i X_{j'u}^i/4$, where X_{ju} denotes the dosage, i.e. number of copies, of allele u for individual j. The average between-individual matching in a sample of N_i individuals from population i $\tilde{M}_S^i = \sum_{j\neq j'} \tilde{M}_{jj'}^i/[N_i(N_i-1)]$ can then be

averaged over populations to get $\tilde{M}^S = \sum_i \tilde{M}^i_S/r$. Similarly, matching between individual j from population i and individual j' from population i' $\tilde{M}^{ii'}_{jj'} = \sum_u X^i_{ju} X^{i'}_{j'u}/4$ leads to average between-population sample matching proportions $\tilde{M}^{ii'}_B = \sum_{j \neq j'} \tilde{M}^{ii'}_{jj'}/(N_i N_{i'})$. Averaging over pairs of populations yields $\tilde{M}^B = \sum_{i \neq i'} \tilde{M}^{ii'}_B/[r(r-1)]$.

Population-specific θ values for genotypic data are given by $\hat{\beta}^i_{ST} = (\tilde{M}^i_S - \tilde{M}^B)/(1 - \tilde{M}^B)$, with an overall estimate of $\hat{\beta}_{ST} = (\tilde{M}^S - \tilde{M}^B)/(1 - \tilde{M}^B)$ per locus. Taking the ratio of averages of numerator and denominator over these locus-specific estimates again yields an average β estimate. Such genotype-based estimates allow for for departures from HWE in the data, although we note that HWE is assumed in Equation 1.

Software to perform these allele-based estimates is simple to prepare as it requires only the number of copies of each allele in each population. More detail was given by Weir and Goudet [80]. There are some good packages now available, including SNPRelate [87] and hierFstat [35].

Data

DNA from 350 individuals, over five geographic groups, included in the 1000 Genomes Project Phase 3 (`http://www.1000genomes.org`) were obtained from the Coriell Institute for Medical Research (Camden, New Jersey, USA) and sequenced using Illumina's MiSeq FGx^{TM} and ForenSeq^{TM} DNA Signature Prep Kit. Genotype calls were obtained through their Universal Analysis Software (UAS) over 27 autosomal loci for both the length-based (LB) allele callings, equivalent to CE, as well as the sequence-based (SB) allele callings. The geographic groups being distinguished are: African (AFR), admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) groups.

Out of the 9,450 genotype calls (350 individuals over 27 STR markers) 7,485 are classified as heterozygotes, 1,772 as homozygotes and the remaining types contain either drop-ins or complete locus drop-outs and are excluded from further analysis. The 7,485 heterozygous types can be distinguished further into isoalleles, showing variation only in the STR repeat region, corresponding to homozygous in the CE case, and alleles of different length. In addition, some flanking region variation can be observed as the UAS incorporates a small amount of sequence variation in these regions for a subset of the markers [19].

Results

Sequence variation

Table 1.1 displays the number of individuals per geographic group and the observed number of unique alleles obtained by length compared to STR sequence after genotype calling. As expected, more variation has been observed for larger sample sizes and sequence-based allele callings as compared to length-based allele callings. Overall, 316 unique length-based alleles have been observed and the amount increases to 593 for sequence-based alleles, indicating differences in allele frequencies over the geographic groups.

		# Alleles		
Group	**Sample size**	**LB**	**SB**	**Increase**
African	88	260	408	57%
American	75	239	337	41%
East Asian	74	234	331	41%
European	58	228	318	40%
South Asian	55	225	317	41%

Table 1.1: Number of individuals per geographic group together with observed number of length-based (LB) alleles and sequence-based (SB) alleles over the set of 27 markers.

The first four columns of Table 1.2 show the number of unique alleles obtained by length compared to STR sequence per locus combined over all individuals, sorted based on the increase in number of alleles. Similarities can be seen between our results and the observations reported by Gettings et al. [29]. An overview of all unique sequences per locus is presented in Supplemental Table 1, together with their corresponding frequencies overall and per geographic group within the data set.

Locus-specific θ

We regard the data we generated from the 1000 Genomes samples as a database. The data are from five groups, the five identified continental ancestry designators. All estimates based on genotypic data are depicted graphically in Figure 1.1 and locus-specific estimates, obtained as an unweighted average over geographic groups, are displayed in Table 1.2. It can be seen that there is a considerable variation of estimates over loci and length-based versus sequence-based estimates may increase, decrease, or stay the same. The latter happens when

loci show no additional sequence variation, as is the case for locus D20S482 and TPOX.

	# Alleles			$\hat{\beta}_{ST}$	
Locus	**LB**	**SB**	**Difference**	**LB**	**SB**
D21S11	17	65	+48	0.0259	0.0383
D12S391	17	64	+47	0.0074	0.0064
D2S1338	13	50	+37	0.0148	0.0149
vWA	10	30	+20	0.0228	0.0260
D1S1656	18	34	+16	0.0174	0.0146
D3S1358	10	25	+15	0.0081	0.0399
D8S1179	11	25	+14	0.0153	0.0209
D13S317	7	19	+12	0.0404	0.0421
D9S1122	9	19	+10	0.0063	0.0130
D2S441	10	19	+9	0.0426	0.0368
D5S818	8	16	+8	0.0396	0.0250
D18S51	13	20	+7	0.0145	0.0141
FGA	22	28	+6	0.0049	0.0046
D19S433	15	19	+4	0.0091	0.0091
D6S1043	19	23	+4	0.0380	0.0377
PentaE	21	25	+4	0.0207	0.0210
CSF1PO	9	12	+3	0.0016	0.0019
D16S539	9	12	+3	0.0224	0.0213
D4S2408	6	9	+3	0.0787	0.0724
D7S820	9	11	+2	0.0199	0.0195
D10S1248	8	9	+1	0.0131	0.0133
D17S1301	8	9	+1	0.0102	0.0101
D22S1045	9	10	+1	0.0524	0.0515
PentaD	14	15	+1	0.0286	0.0288
TH01	7	8	+1	0.0642	0.0640
D20S482	9	9	0	0.0059	0.0059
TPOX	8	8	0	0.0402	0.0402
All				0.0245	0.0257

Table 1.2: Number of unique alleles obtained by length compared to sequence for $N = 350$ individuals for 27 autosomal markers, as well as locus-specific global β estimates based on length-based (LB) genotypic data and sequence-based (SB) genotypic data.

Locus D21S11 shows the highest increase in number of alleles, from 17 different LB alleles to 65 different SB alleles, as well as an increase in the β estimate from 0.0259 to 0.0383. This happens since the extra variation leads to relatively less matching between groups as compared to within groups. From a population genetics perspective, this may occur when populations or groups share the same length-based allele, but the underlying nucleotide sequences differ. If no additional sequence variation is observed within a group, within-group matching is higher for sequence-based genotypic data relative to the global group, leading to higher β estimates

for such groups.

An increase in the observed number of alleles does not necessarily translate to an increase in the β estimates, as can be seen for locus D1S1656. In this case, the extra variation leads to relatively less matching within a group as compared to between. This may happen in a situation where an allele originally unique to a group shows additional sequence variation. The heterozygosity within the group increases more than the overall heterozygosity, yielding smaller estimates.

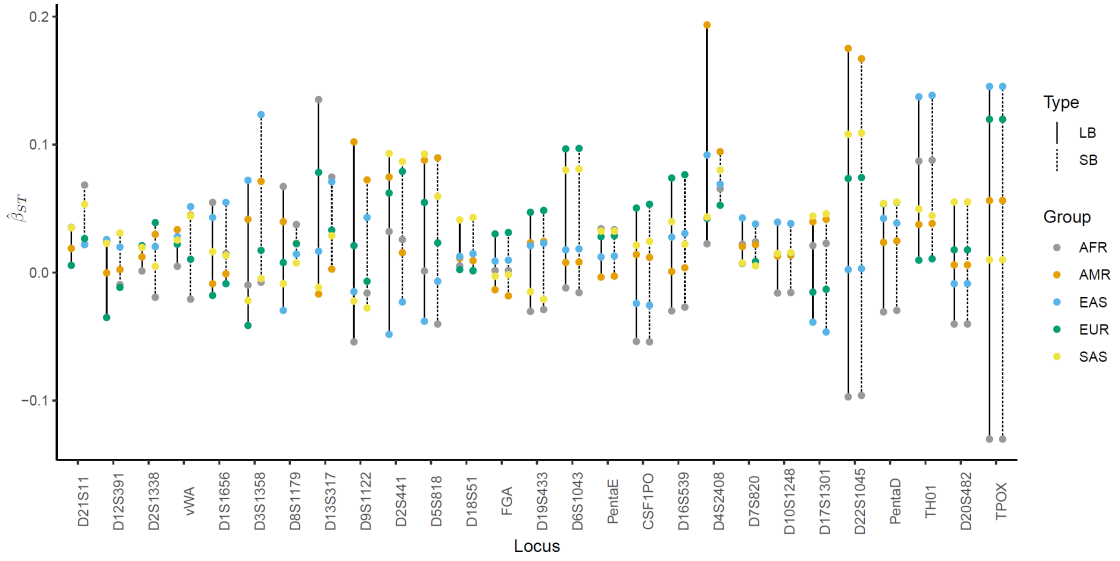

Figure 1.1: β estimates per geographic group (African (AFR), admixed American (AMR), East Asian (EAS), European (EUR), South Asian (SAS)) and locus using length-based (LB) and sequence-based (SB) genotypic data over 27 autosomal STR loci. LB estimates are connected with a solid line, while SB estimates are connected with a dotted line.

Geographic-group-specific θ

Estimated matching proportions based on length-based genotype matching yield an average within-group matching, averaged over groups and loci, of $\tilde{M}^S = 0.2165$, while the average between-group matching is $\tilde{M}^B = 0.1968$, yielding an overall estimate of $\hat{\beta}_{ST} = 0.0245$.

Group-specific estimates range from 0.0035 for the African group to 0.0347 for the American group (Table 1.3). An advantage of having population- or group-specific estimates is that the variation among the estimates reflects differences among θ values, which can be regarded as a signature of different evolutionary histories, such as age and population size. Such effects

are not possible when using the Weir & Cockerham model, as it is assumed there that the populations have equal evolutionary histories [80].

	$\hat{\beta}_{ST}$	
Group	**LB**	**SB**
African	0.0035	−0.0016
American	0.0347	0.0312
East Asian	0.0239	0.0332
European	0.0302	0.0327
South Asian	0.0302	0.0327
All	0.0245	0.0257

Table 1.3: β estimates per geographic group using length-based (LB) and sequence-based (SB) genotype counts.

Our length-based results can be compared to those in the worldwide survey by Buckleton et al. [10] and show concordance, for example, in showing the smallest values for Africa as compared to the rest of the world, as expected from our understanding of higher genetic diversity within those older populations pre-dating the migration of modern humans out of Africa. In addition, the largest values are for the American group, and the Asian and European values are lying between the African and American values.

Matching proportions based on sequence-based genotypes show somewhat lower values of $\tilde{M}^S = 0.1878$ and $\tilde{M}^B = 0.1664$ due to the increase in the number of observed types as a result of the additional variation. The global estimate is in this case $\hat{\beta}_{ST} = 0.0257$, which is an increase from the length-based estimate, albeit small. Not all geographic groups show an increase in the estimated values. For the African group, the average within-group matching proportion is now $\tilde{M}_S = 0.1651$. Relative to the between-group matching $\tilde{M}^B = 0.1664$ individuals in the African group are less similar to each other, leading to a negative β for the African group with all geographic groups as reference set. For both sequence-based as well as length-based results, the "θ-correction" has little effect when applied to the African group. The estimate for the admixed American group has decreased as well, while the Asian and European groups all show an increase in the estimate with values now larger than the American geographic group.

To check the impact of these differences 95% confidence intervals were obtained based on

bootstrapping over loci. The Balding-Nichols formulation refers to profile matching caused by evolutionary processes in previous generations. It reflects what Weir [75] referred to as "genetic sampling." The formulation also has an implicit recognition of variation of allele frequencies among replicates of these evolutionary histories: the only information generally available about these replicates is provided by multiple loci typed on the same individuals and this led Weir [75] to explain why properties of β estimates are obtained by bootstrapping over loci, rather than over individuals. This latter procedure would accommodate the "statistical sampling" of drawing individuals from the same population, but would provide no information about genetic sampling variation. Figure 1.2 demonstrates a great deal of overlap for all intervals and this also holds for the global β estimates (not shown). Overall, comparing the interval for length-based results (0.0179, 0.0315) with the sequence-based results (0.0191, 0.0329) suggests that the usual recommended value of around 3% is appropriate for DNA evaluations based on NGS data. A value of 5% is expected to yield conservative results for each system.

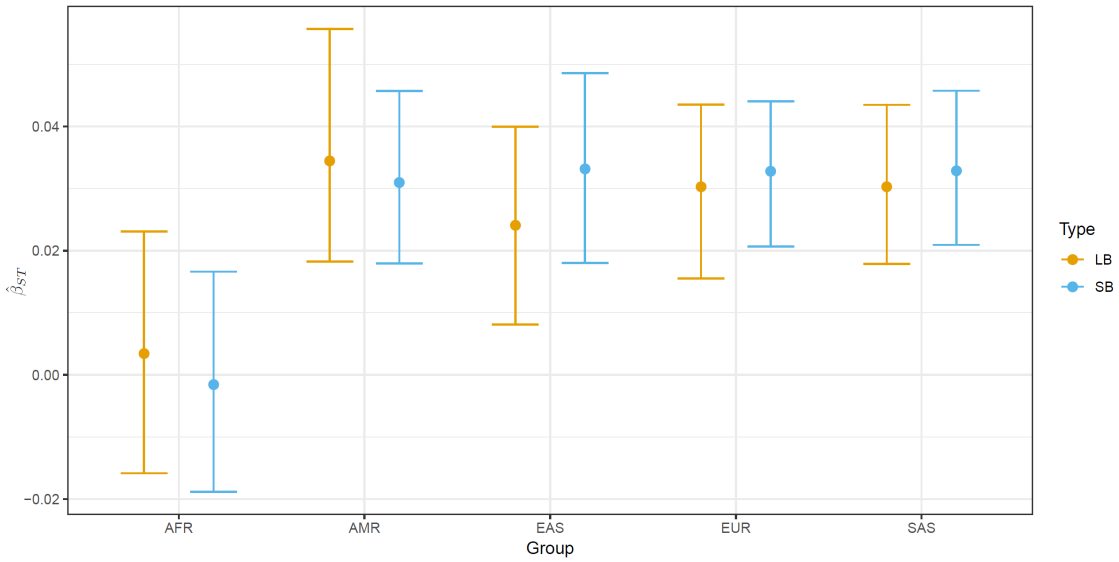

Figure 1.2: Group-specific β estimates for length-based (LB) and sequence-based (SB) genotypic data together with 95% confidence intervals obtained by bootstrapping over loci.

Discussion and conclusion

We presented here an analysis of forensic STR markers to give guidance on θ values for sequence-based data. Since we have access to genotype data, all results have been obtained using genotypic matching proportions. Allelic data may be used if there is Hardy-Weinberg equilibrium within the samples from populations, and results when applying the framework to this type of data (not shown) are almost indistinguishable from the results as presented here.

Although locus-specific estimates are interesting, we recommend averaging θ estimates over loci to reduce the bias and variance of ratio estimators. The values using all loci as shown in the last row of Table 1.2 should be used as a global estimate for θ in match probabilities. If no assumption is made on the ancestry of the true source of the DNA evidence, results may be reported for each of the different groups using group-specific estimates as displayed in Table 1.3. It is important to note that these estimates per geographic group do not reflect substructure within those groups and they are intended to be used in conjunction with global allele frequencies.

NGS-based methods will enhance forensic identification and since such data are subject to population structure, the impact on θ, a measure integral to DNA evidence evaluations, should be checked. This study gives guidance as to what values are appropriate for the population structure quantity used in match probability calculations for sequence-based data. If match probabilities are wanted for use with a single-ancestry database, a study parallel to this one would be needed with data from several populations within that ancestry group.

Although the data used in this study are limited as compared to other studies and an analysis has been performed only on a geographic basis, results for length-based data show patterns concordant with CE-based results. Availability of sequencing data is expected to increase in the upcoming years, so it is recommended to replicate this study more thoroughly. As NGS-based data better reflect the true variation among individuals, population structure estimates based on such data will be more accurate. So far, our results show similar effects of sequencing data on θ estimates as what has been seen for CE-based data.

Appendix A: Variance of sample allele frequencies

If a database has n_i alleles from population i and if x_{ijA} is 1 when allele j from population i is of type A, and is 0 otherwise:

$$\begin{aligned}
\tilde{p}_{iA} &= \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ijA} \\
\tilde{p}_{iA}^2 &= \frac{1}{n_i^2}\left(\sum_{j=1}^{n_i} x_{ijA}^2 + \sum_{\substack{j=1 \\ j\neq j'}}^{n_i}\sum_{j'=1}^{n_i} x_{ijA}x_{ij'A}\right)
\end{aligned}$$

If each population i is in Hardy-Weinberg equilibrium, then the expectation of $\tilde{p}_{iA}^2$ from Weir and Goudet [80] is

$$\begin{aligned}
\mathcal{E}(\tilde{p}_{iA}^2) &= \frac{1}{n_i^2}\left(\sum_{j=1}^{n_i}\pi_A + \sum_{\substack{j=1 \\ j\neq j'}}^{n_i}\sum_{j'=1}^{n_i}\left[\pi_A^2 + \pi_A(1-\pi_A)\theta_i\right]\right) \\
&= \pi_A^2 + \pi_A(1-\pi_A)\left(\theta_i + \frac{1-\theta_i}{n_i}\right) \\
\mathrm{Var}(\tilde{p}_{iA}) &= \pi_A(1-\pi_A)\left(\theta_i + \frac{1-\theta_i}{n_i}\right)
\end{aligned}$$

where θ_i applies to any pair of distinct alleles from population i.

For allele frequencies from two populations:

$$\begin{aligned}
\tilde{p}_{iA}\tilde{p}_{i'A} &= \left(\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ijA}\right)\left(\frac{1}{n_{i'}}\sum_{j'=1}^{n_{i'}} x_{i'j'A}\right) \\
\mathcal{E}(\tilde{p}_{iA}\tilde{p}_{i'A}) &= \frac{1}{n_i n_{i'}}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_{i'}}\left[\pi_A^2 + \pi_A(1-\pi_A)\theta_{ii'}\right] \\
&= \pi_A^2 + \pi_A(1-\pi_A)\theta_{ii'} \\
\mathrm{Cov}(\tilde{p}_{iA}, \tilde{p}_{i'A}) &= \pi_A(1-\pi_A)\theta_{ii'}
\end{aligned}$$

where $\theta_{ii'}$ applies to any pair of distinct alleles, one from population i and one from population i'.

To combine information over the r populations contributing to a database, for sample allele frequencies and for ibd measures, population i receives a weight w_i:

$$\begin{aligned}
\tilde{p}_A &= \frac{1}{(\sum_{i=1}^r w_i)} \sum_{i=1}^r w_i \tilde{p}_i \text{ overall average} \\
\theta_W &= \frac{1}{(\sum_{i=1}^r w_i)} \sum_{i=1}^r w_i \theta_i \text{ within population average} \\
\theta_B &= \frac{1}{(\sum_{i=1}^r \sum_{i'=1, i\neq i'}^r w_i w_{i'})} \sum_{\substack{i=1 \\ i\neq i'}}^r \sum_{i'=1}^r w_i w_{i'} \theta_{ii'} \text{ between population average}
\end{aligned}$$

An "unweighted" analysis sets $w_i = 1$ and it allows each population to contribute equally to ibd averages as may be important when the θ_i are different. A "weighted" analysis sets $w_i = n_i$ so that populations with more alleles in the database contribute more to average ibd measures. This is the weighting scheme used when the database allele frequencies are simply the proportions of each allele in the whole database but frequencies for the contributing populations are not available.

The variance of the sample allele frequencies, for all weighting schemes, is

$$\begin{aligned}
\text{Var}(\tilde{p}_A) &= \frac{1}{(\sum_{i=1}^r w_i)^2} \left(\sum_{i=1}^r w_i^2 \text{Var}(\tilde{p}_{iA}) + \sum_{\substack{i=1 \\ i\neq i'}}^r \sum_{i=1}^r w_i w_{i'} \text{Cov}(\tilde{p}_{iA}, \tilde{p}_{i'A}) \right) \\
&= \pi_A(1-\pi_A) \left(\theta_B + \frac{\sum_{i=1}^r w_i^2(\theta_i - \theta_B)}{(\sum_{i=1}^r w_i)^2} + \frac{\sum_{i=1}^r \frac{w_i^2}{n_i}(1-\theta_i)}{(\sum_{i=1}^r w_i)} \right)
\end{aligned}$$

The weighted and unweighted weighting schemes are the same when each population has the same number n of alleles in the database:

$$\text{Var}(\tilde{p}_A) = \pi_A(1-\pi_A) \left(\theta_B + \frac{\theta_W - \theta_B}{r} + \frac{1-\theta_W}{nr} \right)$$

but for both schemes, when the database is large and it contains alleles from many populations, $\text{Var}(\tilde{p}_A) \approx \pi_A(1-\pi_A)\theta_B$.

Appendix B: Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.fsigen.2020.102364.

Sequence-based population structure, relatedness, and inbreeding estimates for forensic autosomal markers

This section is based on Sanne E. Aalbers, Bruce S. Weir. *Sequence-based population structure, relatedness, and inbreeding estimates for forensic autosomal markers.* (2023). Manuscript in preparation.

Abstract

Population data have become available for sequence data to aid forensic investigations and prepare the forensic community in the move towards implementing NGS methods. This comes with a need for updated population genetic parameter estimates to allow DNA evidence evaluations using sequence data. Initial work has been done on a small sample and here we expand this work by providing estimates of population structure and relatedness for autosomal STR and identity-informative SNP data generated by sequencing technologies. We also discuss the effect of inbreeding on forensic calculations and discuss why the use of genotypic-based estimates may be preferred over allelic-based estimates.

Introduction

When a DNA profile obtained from a crime scene sample is found to match the profile of a suspect, we wish to express the strength of this evidence. Since autosomal profiles consist of pairs of alleles over several loci, the simplest approach to estimating a DNA profile probability is by taking the product of the individual allele probabilities and multiplying over loci. The problem with this approach is that DNA profiles are not independent, even for unrelated people. If the suspect and donor of the crime scene profile are unrelated members of a large random-mating population, the dependencies are expected to be small. Dependencies will be greater for two members of a small and/or relatively isolated population. At single loci, the allelic independence assumption can be avoided by expressions that allow for the effects of population structure.

Wright's F-statistics as described in 1951 consist of three related parameters used to

describe population structure on different levels [82]. The total inbreeding coefficient F_{IT} measures the probability of identity by descent (ibd) within individuals in a subpopulation compared to ibd between individuals in the total population. Within subpopulations, local inbreeding is referred to as F_{IS} and describes the probability of ibd within individuals in a subpopulation compared to ibd between individual but within the same subpopulation. Finally, F_{ST} measures the probability of ibd between individuals within a subpopulation relative to ibd between individuals from different subpopulations. The F-statistics satisfy the relationship $(1 - F_{IT}) = (1 - F_{ST})(1 - F_{IS})$, which demonstrates that there exists inbreeding in the total population even in the absence of local inbreeding. In this case, $F_{IS} = 0$ and $F_{ST} = F_{IT}$, which holds for large, random-mating populations.

In a forensic setting, we may be interested in allelic and genotypic frequencies in a particular subpopulation that may not be available for study, leading to uncertainty about the frequencies. The use of overall sample frequencies may be prejudicial if those frequencies are underestimating those of the relevant population. The Balding-Nichols expressions [5] allow for variation in allele frequencies between subpopulations by invoking a θ value, which measures the relatedness within subpopulations relative to the whole population and is equivalent to Wright's F_{ST}. While this approach allows for deviations from Hardy-Weinberg equilibrium (HWE) in the total population, it does assume HWE holds within subpopulations. The consequence is that the use of population-wide frequencies leads to expectations that depend on the total inbreeding F_{IT}. If the difference between F_{IT} and F_{ST} is small, then F_{IT} can be approximated by F_{ST}. We can check this assumption by verifying that $F_{IS} = 0$ if we have genotypic data available for the subpopulations [76].

In practice, the human population consists of a number of subpopulations that are genetically different from each other. Because the populations are finite, any two individuals will show some level of relatedness, meaning that there will be low levels of inbreeding. Alleles within (and between) loci will thus show dependencies and it would be of interest to quantify the extent of departure from independence. This is especially timely in light of the transformation of current approaches to the incorporation of sequence data. In this paper we focus on characterizing the extent of dependencies within loci for sequence-based autosomal data.

Methods

Data

Our data consist of a combination of samples obtained from two different studies that have been described previously [2, 33]. The first data set contains genotype calls for 350 individuals of the 1000 Genomes Project (1000GP) Phase 3 (`http://www.1000genomes.org`) over five geographic groups [2], typed for 27 autosomal short tandem repeat (aSTR) markers and 94 identity-informative single nucleotide polymorphisms (iiSNPs). The second data set includes individual-specific genotype data from the NIST 1036 data set containing samples over four U.S. populations [33] for the same set of 27 aSTRs. Both data sets were generated using a MiSeq FGx instrument with the ForenSeq DNA Signature Prep Kit. Genotype calls for STR data were restricted to the regions as reported in the sample level reports of Illumina's Universal Analysis Software.

Estimation of population structure, relatedness, and inbreeding

We obtain estimates of population structure and relatedness according to the framework outlined in Weir and Goudet [80]. In brief, genotype-based matching proportions within and between individuals are calculated as expressions of allele dosages X^i_{ju}, i.e., the number of copies of allele u for individual j in group i: $\tilde{M}^i_j = \frac{1}{2}\sum_u X^i_{ju}(X^i_{ju}-1)$, $\tilde{M}^i_{jj'} = \frac{1}{4}\sum_u X^i_{ju}X^i_{j'u}$, and $\tilde{M}^{ii'}_{jj'} = \frac{1}{4}\sum_u X^i_{ju}X^{i'}_{j'u}$. Averaging over individuals or pairs of individuals within a group i of size n_i leads to average group-specific within-individual matching proportions $\tilde{M}^i_I = \sum_j \tilde{M}^i_j/n_i$ and average between-individual matching proportions $\tilde{M}^i_S = \sum_{j\neq j'} \tilde{M}^i_{jj'}/[n_i(n_i - 1)]$. Matching proportions between groups are obtained by averaging over pairs of individuals from different groups: $\tilde{M}^{ii'}_B = \sum_{j\neq j'} \tilde{M}^{ii'}_{jj'}/(n_i n'_i)$. Averaging over a total of r groups or $r(r-1)$ pairs of groups yields overall matching proportions $\tilde{M}_I = \sum_i \tilde{M}^i_I/r$, $\tilde{M}_S = \sum_i \tilde{M}^i_S/r$, $\tilde{M}_B = \sum_{i\neq i'} \tilde{M}^{ii'}_B/[r(r-1)]$. Individual inbreeding and pairwise kinship coefficients are estimated relative to the total set:

$$\hat{\beta}^i_j = \frac{\tilde{M}^i_j - \tilde{M}_B}{1-\tilde{M}_B}, \qquad \hat{\beta}^i_{jj'} = \frac{\tilde{M}^i_{jj'} - \tilde{M}_B}{1-\tilde{M}_B}$$

Group-specific and overall inbreeding and coancestry coefficients are estimated as follows:

$$\hat{\beta}^i_{IS} = \frac{\tilde{M}^i_I - \tilde{M}^i_S}{1 - \tilde{M}^i_S}, \qquad \hat{\beta}_{IS} = \frac{\tilde{M}_I - \tilde{M}_S}{1 - \tilde{M}_S}, \qquad \hat{\beta}^i_{IT} = \frac{\tilde{M}^i_I - \tilde{M}_B}{1 - \tilde{M}_B}, \qquad \hat{\beta}_{IT} = \frac{\tilde{M}_I - \tilde{M}_B}{1 - \tilde{M}_B},$$

$$\hat{\beta}^i_{ST} = \frac{\tilde{M}^i_S - \tilde{M}_B}{1 - \tilde{M}_B}, \qquad \hat{\beta}_{ST} = \frac{\tilde{M}_S - \tilde{M}_B}{1 - \tilde{M}_B}$$

The estimates $\hat{\beta}_{IS}$, $\hat{\beta}_{IT}$, and $\hat{\beta}_{ST}$ are commonly written as F_{IS}, F_{IT} and F_{ST}, respectively, and reflect Wright's F-statistics [82].

If genotypic data are not available, equivalent expressions can be used to obtain estimates based on sample allele frequencies. Within-group matching proportions are then written as $\tilde{M}^i_W = 2n_i[\sum_u \tilde{p}^2_{iu} - 1/(2n_i)]/(2n_i - 1)$, where $\tilde{p}_{iu} = \sum_{j=1}^{n_i} X^i_{ju}/(2n_i)$ is the sample allele frequency of group i. Note that between-group matching yields the same result regardless of using genotype data or sample allele frequencies: $\tilde{M}^{ii'}_B = \sum_{j \neq j'} \tilde{M}^{ii'}_{jj'}/(n_i n_{i'}) = \sum_u \tilde{p}_{iu}\tilde{p}_{i'u}$. Estimates for F_{ST} are then written as:

$$\hat{\beta}^i_{WT} = \frac{\tilde{M}^i_W - \tilde{M}_B}{1 - \tilde{M}_B}, \qquad \hat{\beta}_{WT} = \frac{\tilde{M}_W - \tilde{M}_B}{1 - \tilde{M}_B},$$

where $\tilde{M}_W = \sum_i \tilde{M}^i_W/r$.

Matching proportions within and between each pair of groups can be used to cluster groups based on their pairwise genetic distances. For two groups i and i', the average within-group matching proportion $\tilde{M}^{ii'}_S = [\tilde{M}^i_S + \tilde{M}^{i'}_S]/2$, can be compared to the between-group matching proportion $\tilde{M}^{ii'}_B$, to obtain pairwise genetic distances for each pair of groups:

$$\hat{\beta}^{ii'} = \frac{\tilde{M}^{ii'}_S - \tilde{M}^{ii'}_B}{1 - \tilde{M}^{ii'}_B}$$

Groups can then be clustered on the basis of the genetic distances between them. Here, we use the UPGMA (unweighted pair group method with arithmetic mean) method, which is a simple algorithm that employs a bottom-up approach by clustering the closest pair of groups. At each step, the distance matrix is updated by taking the average of the remaining groups to the new cluster to construct a hierarchical tree.

Estimates are combined over markers by taking the ratio of averages of numerators and denominators rather than the average of the ratios. In other words, if locus-specific estimates $\hat{\beta}_l$ are written as N_l/D_l, the "ratio of averages" is calculated as $\sum_l w_l \hat{\beta}_l / \sum_l w_l$ by setting $w_l = D_l$. This weighted approach was advocated by Weir and Cockerham and performs better than using an unweighted approach $w_l = 1$ [80, 79]. While the variance of the estimates will be reduced by this approach, we still expect significant variation as a result of genetic as well as statistical sampling even with the use of 27 loci [81]. To accommodate this, we can generate confidence intervals for our estimates by bootstrapping over markers.

All calculations and figures in this paper were obtained through the statistical software program R (version 4.0.2).

Results

Data summary

Table 1.4 gives a summary of the total data set of $N = 1386$ individuals, with the nine different subgroups organized into five main genetic-analysis groups. The combined data set shows a total of 393 different length-based (LB) alleles over all 27 autosomal STR markers. When designating alleles by sequence, the total number increases to 849, which is a 116% increase. The most gain, when changing from length-based to sequence-based genotyping, is seen for genetic-analysis group 1, presumably due to a combination of group size and diversity of the underlying subgroups. Figure 1.3 shows a dendrogram clustering of the nine subgroups based on pairwise genetic distances using the UPGMA method. Based on these results, the subgroups have been categorized into five main genetic-analysis groups. The NIST-Caucasian (Cauc) and 1000GP-European (EUR) subgroups are most closely clustered with a genetic distance of -0.0009, followed by the NIST-African American (AfAm) and 1000GP-African (AFR) subgroups with a distance of 0.0007. The NIST-Asian (Asian) and 1000GP-East Asian (EAS) subgroups are grouped together based on a relatively short genetic distance of 0.0012. The NIST-Hispanic (Hisp) and 1000GP-admixed American (AMR) subgroups show clear similarities as well with a distance of 0.0021. Finally, as the 1000GP-South Asian (SAS) groups seems separate from other populations in terms of genetic distance, we treat this

subgroup as its own genetic-analysis group.

There are alternative ways to categorize worldwide population structure, and other methods have been used in the past [69]. The results in this paper are reported on the basis of the five genetic-analysis groups. We recognize that additional structure exists within the 1000GP data due to variation between regional groups. Since our data set is too small to address this variation, we restrict to providing estimates per subgroup in the supplementary material.

Group	**Subgroup**	n	**# Alleles - LB**	**# Alleles - SB**
1		**430**	335	628
	NIST – African American	342		
	1000GP – African	88		
2		**311**	309	519
	NIST – Hispanic	236		
	1000GP – admixed American	75		
3		**171**	267	399
	NIST – Asian	97		
	1000GP – East Asian	74		
4		**419**	299	500
	NIST – Caucasian	361		
	1000GP – European	58		
5		**55**	225	317
	1000GP – South Asian	55		
Overall		1386	393	849

Table 1.4: Number of individuals (n) per genetic-analysis group and subgroup together with observed number of length-based (LB) and sequence-based (SB) alleles over 27 autosomal STR markers. Data come from either the NIST 1036 set (NIST) or a set of 350 individuals of the 1000 Genomes Project (1000GP).

Locus-specific estimates for aSTRs

Table 1.5 shows the number of unique alleles per locus for both length-based and sequence-based allele callings, respectively, ordered by increase in the number of alleles between the two systems, together with locus-specific β estimates. Table 1.6 shows the average matching proportions per data type along with the overall β estimates. The increase in polymorphism observed in sequence data implies that matching proportions will decrease. Within-group matching decreased from 0.2124 for length-based data to 0.1843 for sequence-based data. The matching proportions between genetic-analysis groups decreased from 0.1962 to 0.1664, respectively. The effect on the β estimates is less predictable and there exists considerable

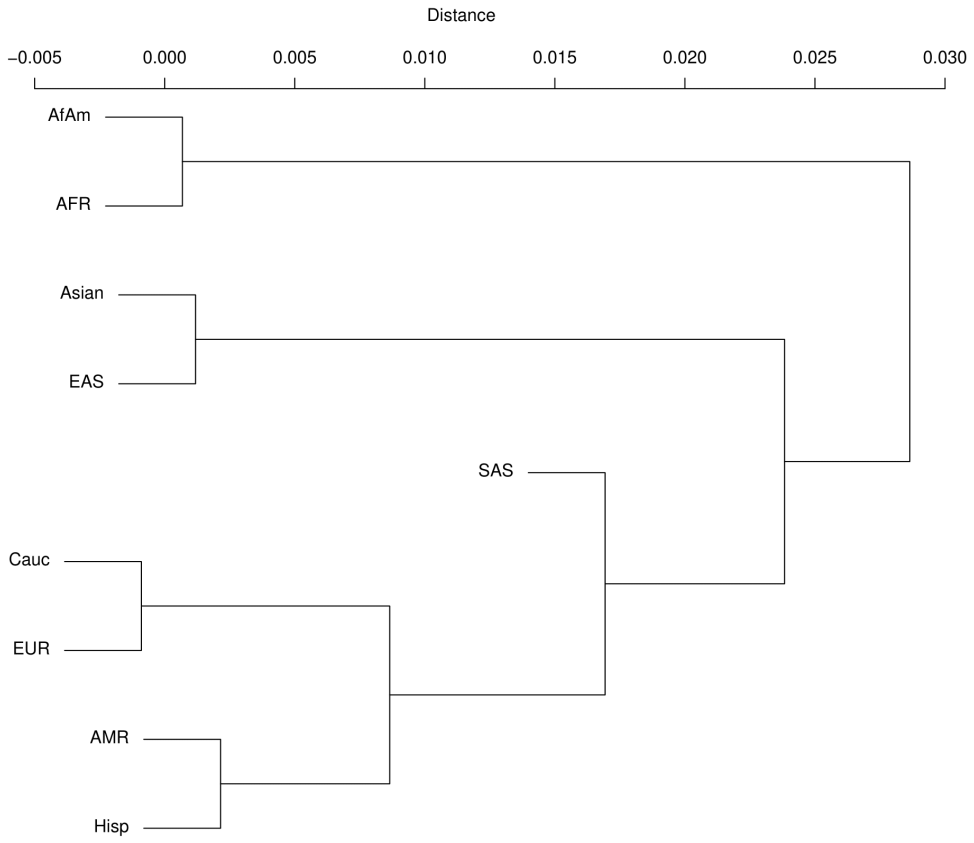

Figure 1.3: Dendrogram clustering of pairwise genetic distances between nine subgroups using the UPGMA method.

variation over loci, as also observed in Aalbers et al. [2]. Estimates may increase or decrease (or stay the same if no sequence variation is observed), depending on where the extra variation occurs (see Appendix A for a more thorough discussion of the theoretical impact of sequence data on the β estimates).

Overall, all relative inbreeding and coancestry estimates are similar between the two data types. Average matching within individuals (0.2172 and 0.1879 for length-based and sequence-based data, respectively) is higher than, but relatively close to, between individual matching within groups, which is consistent with observations from Thompson [70]. While average within-group inbreeding estimates are small, they cannot readily assumed to be zero (see also Figure 1.4). Average inbreeding relative to the whole data set is estimated around 2.6%. The coancestry estimates $\hat{\beta}_{ST}$ are estimated around 2.0−2.2%. Locus-specific estimates per subgroup are available in Supplementary Tables 1.9−1.14.

	# Alleles			$\hat{\beta}_{IS}$		$\hat{\beta}_{IT}$		$\hat{\beta}_{ST}$	
Locus	**LB**	**SB**	**Difference**	**LB**	**SB**	**LB**	**SB**	**LB**	**SB**
D21S11	27	116	89	0.0238	0.0263	0.0438	0.0518	0.0204	0.0262
D12S391	24	95	71	-0.0152	-0.0143	-0.0062	-0.0061	0.0089	0.0081
D2S1338	13	71	58	0.0095	0.0028	0.0233	0.0157	0.0140	0.0129
vWA	12	40	28	0.0111	0.0042	0.0257	0.0206	0.0147	0.0164
D3S1358	11	35	24	0.0367	0.0167	0.0433	0.0430	0.0069	0.0268
D8S1179	11	35	24	-0.0017	-0.0045	0.0111	0.0125	0.0128	0.0169
D13S317	8	31	23	0.0160	0.0151	0.0451	0.0504	0.0296	0.0358
D1S1656	18	38	20	0.0036	0.0074	0.0224	0.0247	0.0188	0.0174
FGA	28	43	15	-0.0005	-0.0003	0.0038	0.0038	0.0043	0.0041
D18S51	22	34	12	-0.0022	-0.0023	0.0127	0.0127	0.0148	0.0149
D6S1043	29	41	12	-0.0128	-0.0122	0.0075	0.0082	0.0200	0.0202
D9S1122	11	23	12	0.0074	0.0047	0.0098	0.0130	0.0024	0.0083
D5S818	9	20	11	0.0157	0.0105	0.0417	0.0280	0.0264	0.0176
D19S433	16	26	10	-0.0085	-0.0061	0.0043	0.0066	0.0127	0.0126
D2S441	15	25	10	-0.0135	-0.0067	0.0219	0.0256	0.0350	0.0320
PentaE	25	33	8	0.0166	0.0124	0.0356	0.0319	0.0194	0.0197
CSF1PO	9	15	6	-0.0024	-0.0041	0.0013	-0.0003	0.0037	0.0038
D16S539	9	14	5	0.0189	0.0185	0.0395	0.0381	0.0210	0.0201
D7S820	12	16	4	0.0046	0.0049	0.0224	0.0224	0.0179	0.0175
D17S1301	10	13	3	-0.0001	0.0007	0.0142	0.0152	0.0144	0.0145
D4S2408	7	10	3	0.0563	0.0368	0.1002	0.0899	0.0465	0.0551
D10S1248	12	14	2	-0.0139	-0.0143	-0.0008	-0.0011	0.0129	0.0130
D22S1045	11	13	2	-0.0019	-0.0021	0.0394	0.0391	0.0412	0.0410
D20S482	11	12	1	-0.0082	-0.0077	0.0012	0.0016	0.0094	0.0093
PentaD	16	17	1	0.0230	0.0223	0.0402	0.0395	0.0175	0.0176
TH01	8	9	1	0.0323	0.0336	0.0952	0.0963	0.0650	0.0648
TPOX	9	10	1	-0.0327	-0.0332	0.0049	0.0045	0.0364	0.0365
Overall	393	849	456	0.0061	0.0043	0.0261	0.0257	0.0201	0.0215

Table 1.5: Estimated values of locus-specific β's for 27 autosomal STR markers ordered by increase in the number of alleles comparing sequence-based (SB) data to length-based (LB) data.

Data type	$\tilde{M}^I$	$\tilde{M}^S$	$\tilde{M}^B$	$\hat{\beta}_{IS}$	$\hat{\beta}_{IT}$	$\hat{\beta}_{ST}$
Length-based	0.2172	0.2124	0.1962	0.0061	0.0261	0.0201
Sequence-based	0.1879	0.1843	0.1664	0.0043	0.0257	0.0215

Table 1.6: Average matching proportions within individuals, between individuals within groups, and between genetic-analysis groups, respectively, with overall estimated β values for 27 autosomal STR markers for length-based and sequence-based data.

Group-specific estimates for aSTRs

Table 1.7 shows group-specific estimates and these results are graphically displayed in Figure 1.4, along with 95% confidence intervals obtained by bootstrapping over loci. Genetic-analysis group 1 shows relatively less population structure and total inbreeding compared

to the overall set. The African America and African subgroups show the smallest values for all $\hat{\beta}_{ST}$ and $\hat{\beta}_{IT}$ (Supplementary Tables 1.9, 1.11, 1.12, 1.14). Groups 3 and 5 show the largest difference in within-individual matching relative to between-individual matching and are relatively most inbred. The East Asian subgroup shows the highest value for the local inbreeding estimates $\hat{\beta}_{IS}$, with a value of 0.0201 (95% CI: 0.0015 – 0.0408) for sequence-based data (Table 1.13). It also shows the highest total inbreeding estimate $\hat{\beta}_{IT}$ with a value of 0.0515 (95% CI: 0.0278 – 0.0781), followed by the admixed American (0.0383, 95% CI: 0.0159 – 0.0599) and South Asian (0.0385, 95% CI: 0.0148 – 0.0606) subgroups.

Supplementary Tables 1.15 and 1.16 show estimates for $\hat{\beta}_{WT}$ using allelic data for length-based and sequence-based STRs, respectively. It can be seen that estimates are close, but not identical, to the ones obtained from genotypic data. We will come back to this observation in the discussion section.

	$\hat{\beta}_{IS}$		$\hat{\beta}_{IT}$		$\hat{\beta}_{ST}$	
Group	**LB**	**SB**	**LB**	**SB**	**LB**	**SB**
Group 1	0.0002	0.0022	0.0026	-0.0009	0.0024	-0.0031
Group 2	0.0046	0.0018	0.0260	0.0229	0.0215	0.0212
Group 3	0.0151	0.0114	0.0370	0.0408	0.0222	0.0297
Group 4	-0.0018	-0.0008	0.0218	0.0259	0.0236	0.0268
Group 5	0.0126	0.0071	0.0431	0.0396	0.0309	0.0327
Overall	0.0061	0.0043	0.0261	0.0257	0.0201	0.0215

Table 1.7: Estimated values of group-specific β's averaged over 27 autosomal STR markers for length-based (LB) and sequence-based (SB) data.

Individual inbreeding and pairwise kinship estimates for aSTRs

We now briefly turn to individual-specific inbreeding estimates and pairwise coancestry estimates. Figure 1.5 shows individual-specific inbreeding $\hat{\beta}_j^i$ estimates for both data types. Estimates were obtained by using the whole data set as a reference group and we observe similar patterns as seen with the group-specific estimates. Genetic-analysis group 1 shows relatively lower inbreeding estimates compared to other groups. Overall, the distribution of inbreeding estimates for sequence data is more narrow compared to the distribution for length-based data, with lower observed variance within each genetic-analysis group.

Average pairwise coancestry estimates relative to all pairs of individuals within a group

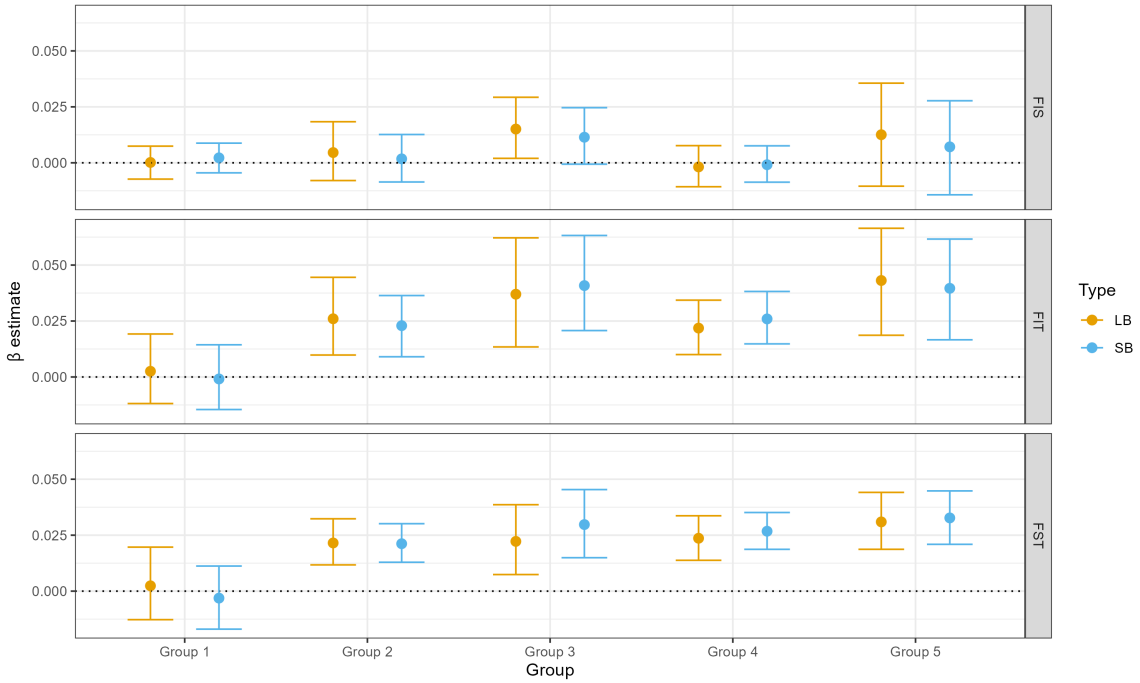

Figure 1.4: Group-specific β estimates for length-based (LB, in orange) and sequence-based (SB, in blue) data along with 95% confidence intervals obtained by bootstrapping over 27 autosomal STR loci. Estimates are given for within-group inbreeding (top), inbreeding relative to the total set (middle), and coancestry (bottom).

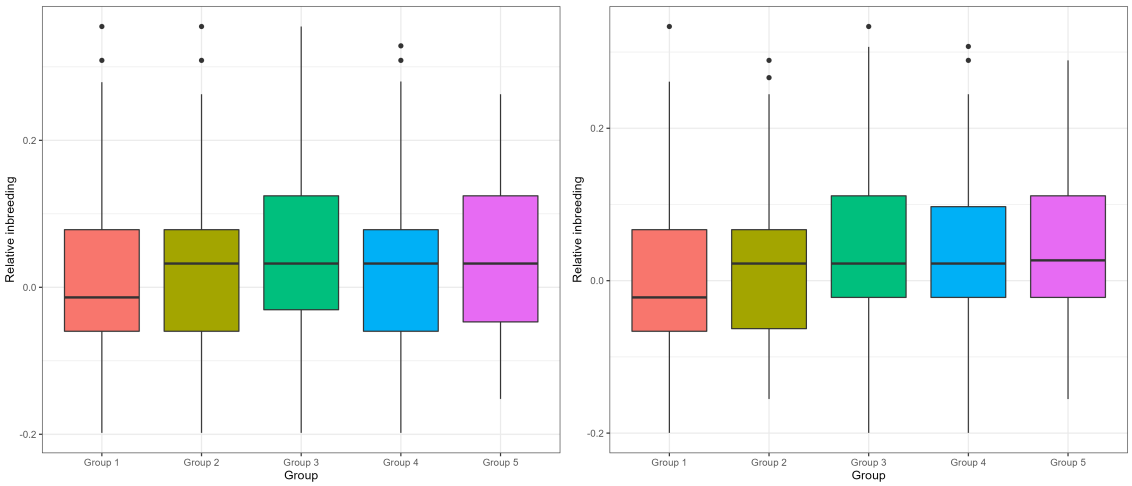

Figure 1.5: Individual-specific inbreeding estimates relative to the whole set for each of the five genetic-analysis groups for length-based data (left) and sequence-based data (right), averaged over 27 autosomal STR loci.

are zero by construct [80]. Since pedigree coancestries are non-negative, it may be of use to re-scale the estimates to get positive values. We do not see a specific need to do so as all estimates are still relative to all pairs of individuals within a group, and they do not directly translate

to pedigree values due to inherent variation of actual relatedness around pedigree relatedness [39]. Although human populations tend to avoid marriages for people more closely related than first cousins, higher general kinship levels are expected to lead to higher inbreeding levels. Figure 1.6 shows all individual-specific inbreeding estimates plotted against their average pairwise coancestry estimates for sequence-based data. A linear regression was fitted to determine the effect and strength of the relationship between the two measures within each genetic-analysis group and over all groups combined.

Since the within-individual matching proportion $\tilde{M}_j^i$ is either 0 or 1 at each locus, inbreeding estimates can only take on a limited set of values, leading to the discrete x-axis values as seen in Figure 1.6. It can be seen that group 1 (in red) shows a relatively lower degree of inbreeding as well as relatedness as compared to other groups. Although a great deal of variation is seen over the estimates, a positive significant relationship between inbreeding and average kinship is observed within each genetic-analysis group, as well as for the total set. Similar results are observed for length-based data, although a stronger correlation is observed for sequence data (adjusted R^2 values of 0.1370 and 0.1476, respectively). As such, we cannot generally assume that inbreeding or kinship values are zero. Moreover, if we estimate inbreeding coefficients based on allelic data, our estimates will be confounded by average kinship with other individuals.

Results for iiSNPs

The iiSNP analysis is restricted to the set of 350 1000GP individuals over five subgroups and does not have a differentiation between length-based and sequence-based data as observed in autosomal STRs. Locus-specific β estimates are displayed in Figure 1.7 and demonstrate a high variability of the estimates among markers. The overall average matching proportions are 0.5737 for within-individual matching, 0.5608 for within-group matching, and 0.5355 for between-group matching. As expected, observed matching is much higher compared to aSTR data due to the limited amount of variability within a SNP marker.

Group-specific and overall β estimates are given in Table 1.8 and Figure 1.8 displays these results graphically along with 95% confidence intervals obtained by bootstrapping over loci. Total inbreeding is estimated as $\hat{\beta}_{IT} = 0.0822$, within-population inbreeding is estimated as

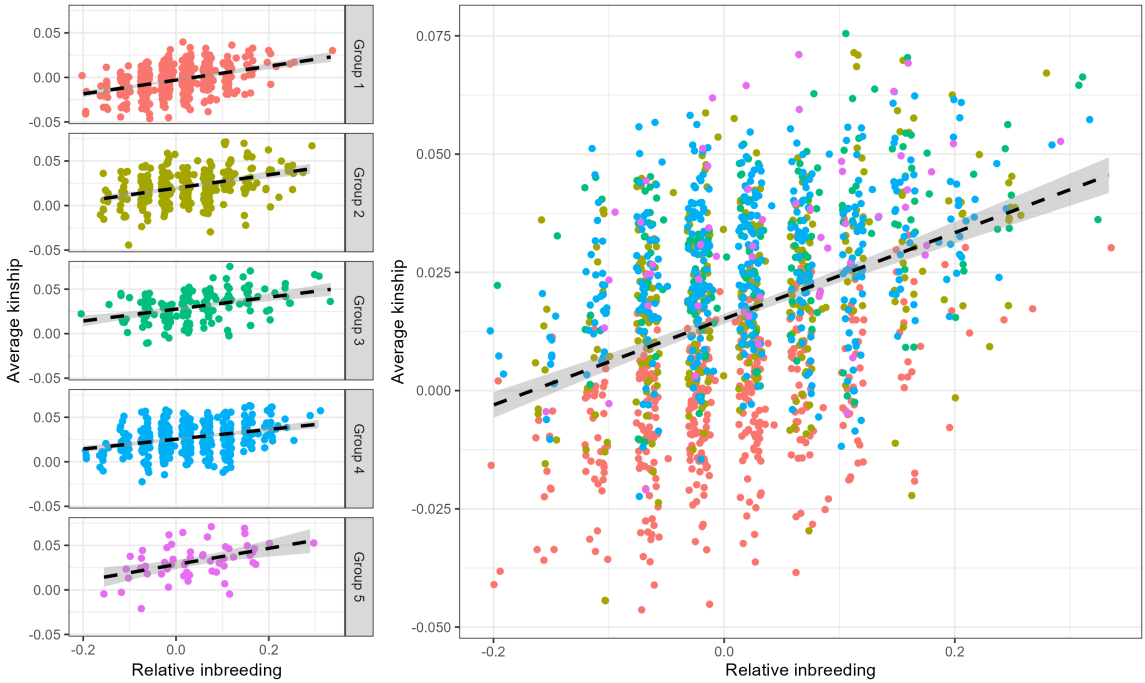

Figure 1.6: Relative inbreeding estimates versus average kinship estimates within groups (left) and for all groups combined (right), averaged over 27 autosomal STR loci using sequence data. Dashed lines indicate a linear regression fit to the data along with a confidence interval in gray.

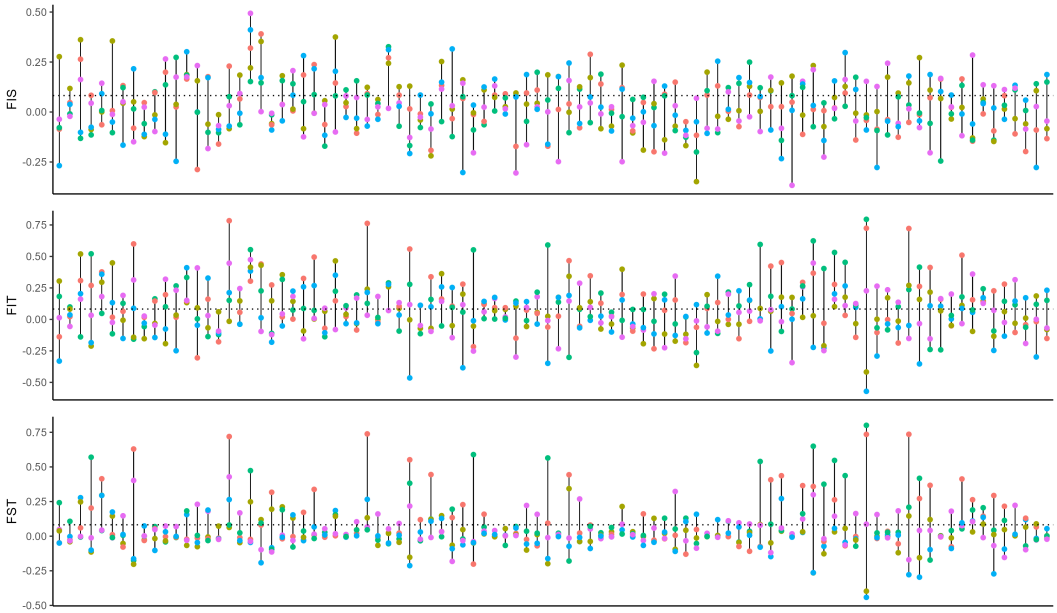

Figure 1.7: Locus-specific β estimates per 1000GP subgroup (AFR in red, AMR in yellow, EAS in green, EUR in blue, SAS in purple) for 94 iiSNP markers. Overall estimates are indicated with dashed lines.

$\hat{\beta}_{IS} = 0.0292$, and a global estimate for coancestry is given by $\hat{\beta}_{ST} = 0.0546$. Interestingly, we see different patterns for the iiSNP markers as compared to aSTR data, with the AFR subgroup now showing relatively higher coancestry and total inbreeding estimates. This observation does highlight the fact that different marker systems show different β estimates and SNP markers usually show higher coancestry estimates due to the lower mutation rates [78]. We do again detect small but non-zero β_{IS} estimates, with the AMR subgroup showing the highest within-population inbreeding estimates for iiSNP data. All locus-specific estimates per subgroup are available in Supplementary Tables 1.17−1.19. Supplementary Table 1.20 shows allelic-based estimates $\hat{\beta}_{WT}$ for comparison.

Subgroup	$\hat{\beta}_{IS}$	$\hat{\beta}_{IT}$	$\hat{\beta}_{ST}$
AFR	0.0267	0.1336	0.1098
AMR	0.0502	0.0693	0.0201
EAS	0.0276	0.1123	0.0871
EUR	0.0248	0.0306	0.0060
SAS	0.0154	0.0654	0.0507
Overall	0.0291	0.0822	0.0547

Table 1.8: Estimated values of group-specific β's averaged over 94 identity-informative SNP markers for the 1000GP data set.

Figure 1.9 shows individual-specific inbreeding estimates compared to average kinship estimates for the total iiSNP data set. We observe a significant positive correlation between the estimates and a stronger overall relationship with a slope value of 0.1935 ($p <$ 2e-16) for the iiSNP markers as compared to aSTR data. While substantial variation exists among individuals, the general increasing trend as seen for both aSTR and iiSNP data indicated that average kinship plays an important role in inbreeding estimation. Similar observations were noted in Zhang et al. [86].

Discussion

The inbreeding and coancestry estimates as presented in this paper have expected values of:

$$\mathcal{E}(\hat{\beta}_{IS}) = \frac{F_I - \theta_S}{1 - \theta_S}, \qquad \mathcal{E}(\hat{\beta}_{IT}) = \frac{F_I - \theta_B}{1 - \theta_B}, \qquad \mathcal{E}(\hat{\beta}_{ST}) = \frac{\theta_S - \theta_B}{1 - \theta_B}, \tag{1.2}$$

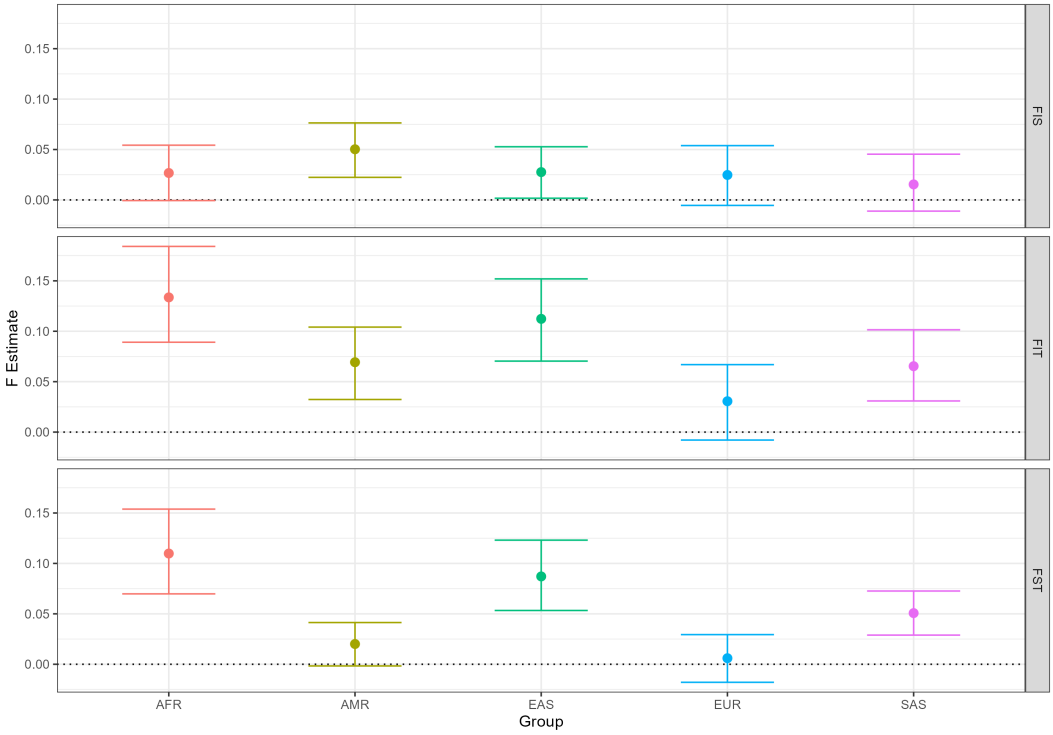

Figure 1.8: Group-specific β estimates along with 95% confidence intervals obtained by bootstrapping over 94 iiSNP markers. Estimates are given for within-group inbreeding (top), inbreeding relative to the total set (middle), and coancestry (bottom).

where F_I is the average inbreeding coefficient, θ_S is the average over all within-group coancestry coefficients, and θ_B is the average over all between-group coancestry coefficients. When genotype data are not available, allelic-based estimators may be calculated, which have expectation:

$$\mathcal{E}(\hat{\beta}_{WT}) = \frac{\theta_W - \theta_B}{1 - \theta_B},$$

where θ_W is the average probability of identity by descent for pairs of alleles within groups. Only if there is no local inbreeding, i.e., $F_{IS} = 0$, it holds that $F_I = \theta_S = \theta_W$. In other words, the estimators $\hat{\beta}_{ST}$ and $\hat{\beta}_{WT}$ give the same value if there is HWE within each group. If this assumption does not hold, the estimators will give different values and genotypic-based estimates are preferred. Also note that with genotypic data, the usual relationship $(1 - \hat{\beta}_{IT}) = (1 - \hat{\beta}_{ST})(1 - \hat{\beta}_{IS})$ is preserved and holds for all group-specific β estimates as well.

The Weir and Cockerham (W&C) estimator [79], as often used in forensic applications

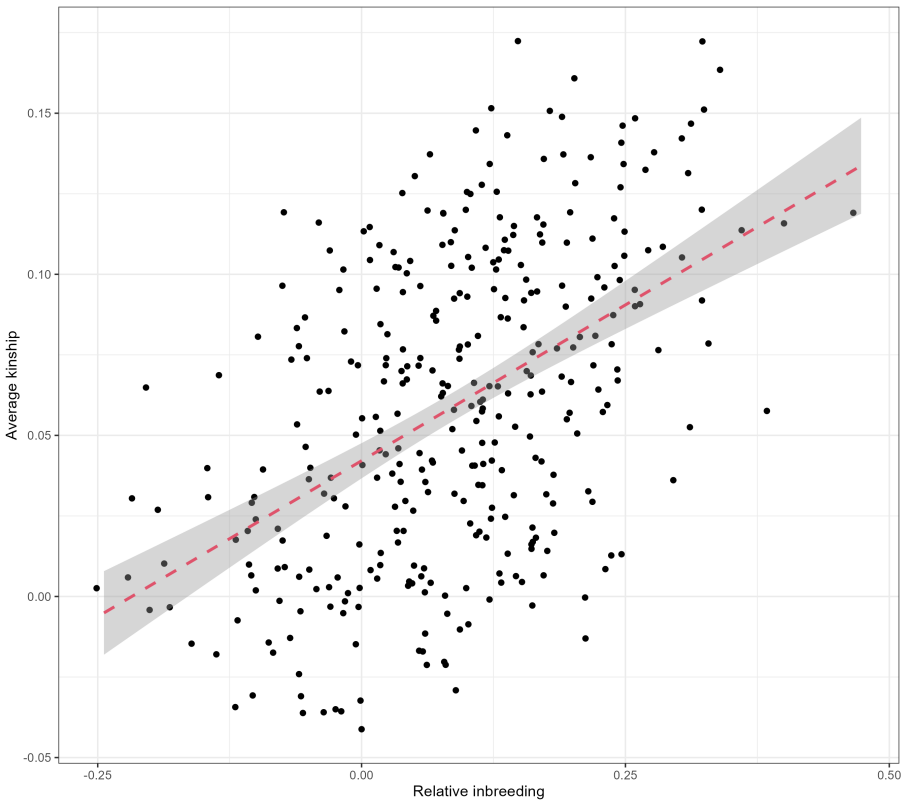

Figure 1.9: Relative inbreeding estimates versus average kinship estimates averaged over 94 iiSNP loci using the 1000GP data set. The dashed red line indicates a linear regression fit to the data along with a confidence interval in gray.

when estimating population structure, suffers from the same limitations. In addition to HWE, it assumes independent populations, i.e., $\theta_B = 0$, each having equal evolutionary history. If the W&C model is assumed, the estimators have the same expectations as shown in Equation 1.2. The estimator performs well unless coancestry values and/or sample sizes are different among populations. In the latter case, the expectation of the estimator depends on inbreeding and coancestry levels, sample sizes, and the number of populations sampled.

As demonstrated in this paper, there exists a complex, but general positive relationship between inbreeding and average kinship for forensic markers. The effect of relatedness will be greatest for small groups and small allelic frequencies, which will be important for sequence data due to their increased capability of observing rare variants. Since ignoring allelic dependencies has the potential of being prejudicial to the suspect, we recommend avoiding the independence assumption by incorporating population structure parameters. Such

parameters are best estimated using genotypic data that accurately reflect both inbreeding and relatedness levels, as allelic data is limited to providing only a composite indication of inbreeding and kinship.

Appendix A: Theoretical impact of sequence data on β estimates

When looking at sequence variation instead of just allele length, the matching proportions will generally decrease. To investigate what effect this will have on the $\hat{\beta}$ estimates, consider the fraction:

$$\frac{(\tilde{M}^S - x) - (\tilde{M}^B - y)}{1 - (\tilde{M}^B - y)} = \frac{\tilde{M}^S - \tilde{M}^B + y - x}{1 - \tilde{M}^b + y},$$

where x denotes the decrease in matching within populations and y denotes the decrease in matching between populations. Depending on x, y and the original matching proportions, the estimate may increase, decrease or stay the same. Generally, using properties of the mediant we know that the estimate will move towards the fraction $\frac{y-x}{y}$. We discuss a couple of situations below.

Situation 1 Suppose the within matching does not change, while the between matching decreases (i.e. $x = 0$, $y > 0$). Write $a = \tilde{M}^S - \tilde{M}^B$ and $b = 1 - \tilde{M}^B$. Since $\tilde{M}^S$ and $\tilde{M}^B$ are bounded by zero and one: $a \leq b$ and $b \geq 0$, such that

$$\begin{aligned} by &\geq ay \\ ab + by &\geq ab + ay \\ b(a + y) &\geq a(b + y) \\ \frac{a + y}{b + y} &\geq \frac{a}{b} \end{aligned}$$

Alternatively, this can also be graphically depicted using mediants. As the slope of $\frac{y}{y}$ is one, and the slope of $\frac{a}{b}$ will always be smaller than or equal to one, the updated β estimate will always be at least as big as the original one. From a population genetics perspective, this happens when populations show the same length-based alleles but the underlying sequences differ per population. LB-allele callings assume there is relatedness between populations,

whereas SB-allele callings show that the populations are actually different. Relative to the overall matching, the within matching increases, leading to higher estimates. In case the sequence-based matching shows no matching at all between populations, the β estimates are simply the (average of the) within matching proportions.

Situation 2 For positive estimates, if within matching decreases more than between matching, the β estimate decreases. The slope of the fraction $\frac{y-x}{y}$ will be negative as $x > y$ and since the original slope was positive, the slope of the mediant will decrease. Another explanation uses properties of fractions: the denominator of the fraction will always increase (i.e. more heterozygosity), while the numerator will decrease. The overall fraction will thus become smaller given that the numerator is positive. In terms of population genetics theory, relative to the overall matching, the within matching decreases (i.e heterozygosity within populations increases more than overall heterozygosity), yielding smaller estimates.

Situation 3 For negative estimates, if between matching decreases more than within matching, the β estimate increases. The slope of the fraction $\frac{y-x}{x}$ will be positive as $x < y$ and since the original slope was negative, the slope of the mediant will increase. Another explanation uses properties of fractions: the denominator increases, while the numerator becomes a larger negative number. The overall fraction will thus increase. In population genetics terms, relative to the overall matching the within matching was lower, but now that the overall matching decreases the within matching increases relatively, leading to increased estimates.

Situation 4 If the decrease is the same for within and between matching $x = y$ the β estimate will move towards 0. Even if both matching proportions decrease, the relative difference stays the same. As the overall heterozygosity does increase, positive estimates will decrease and negative estimates will increase (i.e. the relative heterozygosity stays the same while there is less overall matching). If the original difference in matching is offset by the decrease (i.e. $\tilde{M}^S - \tilde{M}^B = -(y - x)$) the β estimate will become 0. If there is no decrease in matching, the β estimate will not change.

Appendix B: Supplementary tables

	Genetic-analysis group						Subgroup									
Locus	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**All**	**AfAm**	**AFR**	**AMR**	**Asian**	**Cauc**	**EAS**	**EUR**	**Hisp**	**SAS**	**All**
CSF1PO	-0.0449	0.0291	-0.0217	0.0387	0.0173	0.0037	-0.0397	-0.0554	0.0128	-0.0056	0.0393	-0.0254	0.0493	0.0371	0.0201	0.0036
D10S1248	-0.0252	0.0305	0.0211	0.0273	0.0109	0.0129	-0.0289	-0.0221	0.0070	0.0058	0.0285	0.0337	0.0088	0.0349	0.0091	0.0086
D12S391	0.0162	-0.0074	0.0327	-0.0256	0.0285	0.0089	0.0116	0.0285	0.0047	0.0303	-0.0257	0.0306	-0.0300	-0.0115	0.0279	0.0074
D13S317	0.1333	-0.0235	0.0099	0.0407	-0.0123	0.0296	0.1249	0.1277	-0.0257	0.0009	0.0272	0.0078	0.0701	-0.0342	-0.0205	0.0309
D16S539	-0.0072	0.0041	0.0256	0.0504	0.0322	0.0210	0.0062	-0.0349	-0.0039	0.0351	0.0507	0.0231	0.0695	0.0095	0.0352	0.0212
D17S1301	0.0164	0.0064	-0.0138	0.0129	0.0499	0.0144	0.0138	0.0269	0.0455	-0.0036	0.0158	-0.0325	-0.0091	-0.0064	0.0499	0.0111
D18S51	-0.0010	-0.0018	0.0306	0.0017	0.0445	0.0148	-0.0037	0.0084	0.0140	0.0394	-0.0001	0.0158	0.0053	-0.0054	0.0442	0.0131
D19S433	-0.0396	0.0307	0.0255	0.0612	-0.0143	0.0127	-0.0459	-0.0336	0.0201	0.0218	0.0593	0.0180	0.0441	0.0281	-0.0180	0.0104
D1S1656	0.0348	0.0034	0.0490	-0.0084	0.0152	0.0188	0.0287	0.0520	-0.0115	0.0519	-0.0096	0.0402	-0.0208	0.0056	0.0134	0.0166
D20S482	0.0288	0.0184	-0.0550	-0.0108	0.0654	0.0094	0.0481	-0.0253	0.0203	-0.0840	-0.0128	0.0059	0.0320	0.0243	0.0689	0.0086
D21S11	0.0022	0.0116	0.0330	0.0169	0.0382	0.0204	-0.0056	0.0371	0.0207	0.0335	0.0178	0.0367	0.0072	0.0068	0.0369	0.0212
D22S1045	-0.0820	0.1154	0.0006	0.0574	0.1148	0.0412	-0.0855	-0.0929	0.1785	-0.0084	0.0496	0.0060	0.0771	0.1005	0.1116	0.0374
D2S1338	-0.0001	0.0183	0.0200	0.0138	0.0178	0.0140	-0.0014	-0.0009	0.0100	0.0214	0.0123	0.0179	0.0190	0.0193	0.0174	0.0128
D2S441	0.0211	0.0397	-0.0017	0.0111	0.1047	0.0350	0.0120	0.0415	0.0836	0.0197	-0.0023	-0.0383	0.0713	0.0235	0.1016	0.0348
D3S1358	0.0144	0.0068	0.0697	-0.0370	-0.0196	0.0069	0.0147	-0.0119	0.0398	0.0579	-0.0423	0.0702	-0.0437	-0.0072	-0.0240	0.0059
D4S2408	-0.0043	0.0887	0.0773	0.0295	0.0413	0.0465	-0.0178	0.0110	0.1842	0.0623	0.0177	0.0813	0.0309	0.0540	0.0322	0.0506
D5S818	-0.0081	0.0480	-0.0596	0.0617	0.0900	0.0264	-0.0007	0.0060	0.0923	-0.0648	0.0700	-0.0330	0.0593	0.0474	0.0970	0.0304
D6S1043	-0.0146	-0.0106	-0.0060	0.0542	0.0769	0.0200	-0.0093	-0.0117	0.0080	-0.0164	0.0527	0.0181	0.0969	-0.0066	0.0806	0.0236
D7S820	0.0347	0.0241	0.0416	-0.0135	0.0026	0.0179	0.0350	0.0142	0.0122	0.0506	-0.0198	0.0347	-0.0011	0.0243	-0.0008	0.0166
D8S1179	0.0398	0.0297	-0.0267	0.0266	-0.0055	0.0128	0.0269	0.0657	0.0381	-0.0321	0.0237	-0.0314	0.0062	0.0195	-0.0105	0.0118
D9S1122	-0.0283	0.0563	-0.0303	0.0270	-0.0127	0.0024	-0.0270	-0.0482	0.1075	-0.0517	0.0223	-0.0090	0.0266	0.0386	-0.0164	0.0047
FGA	0.0001	-0.0109	0.0196	0.0175	-0.0051	0.0043	-0.0005	-0.0016	-0.0168	0.0251	0.0155	0.0056	0.0272	-0.0100	-0.0060	0.0043
PentaD	-0.0404	0.0081	0.0485	0.0303	0.0412	0.0175	-0.0321	-0.0387	0.0161	0.0706	0.0346	0.0349	0.0465	0.0174	0.0468	0.0218
PentaE	0.0199	0.0068	0.0046	0.0311	0.0346	0.0194	0.0157	0.0361	-0.0017	-0.0028	0.0303	0.0141	0.0298	0.0083	0.0340	0.0182
TH01	0.0746	0.0312	0.1307	0.0393	0.0492	0.0650	0.0634	0.0788	0.0288	0.1198	0.0364	0.1297	0.0006	0.0197	0.0410	0.0576
TPOX	-0.1017	0.0336	0.1412	0.0869	0.0222	0.0364	-0.1138	-0.1333	0.0538	0.1140	0.0659	0.1431	0.1176	0.0104	0.0074	0.0295
vWA	0.0014	0.0229	0.0209	0.0088	0.0198	0.0147	0.0040	0.0002	0.0291	0.0299	0.0100	0.0236	0.0178	0.0227	0.0211	0.0176
All	0.0024	0.0215	0.0222	0.0236	0.0309	0.0201	0.0005	0.0025	0.0337	0.0200	0.0207	0.0230	0.0293	0.0167	0.0293	0.0195

Table 1.9: Locus-specific $\hat{\beta}_{ST}$ for each genetic-analysis group and subgroup for length-based autosomal STR data.

	Genetic-analysis group						Subgroup									
Locus	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**All**	**AfAm**	**AFR**	**AMR**	**Asian**	**Cauc**	**EAS**	**EUR**	**Hisp**	**SAS**	**All**
CSF1PO	-0.0197	-0.0231	-0.0394	-0.0125	0.0845	-0.0024	-0.0126	-0.0431	-0.0627	-0.0389	-0.0190	-0.0511	0.0327	-0.0093	0.0845	-0.0140
D10S1248	-0.0158	-0.0160	0.0181	-0.0085	-0.0471	-0.0139	-0.0212	0.0077	-0.0303	0.0320	-0.0064	0.0029	-0.0232	-0.0106	-0.0471	-0.0107
D12S391	-0.0078	-0.0050	-0.0212	-0.0108	-0.0323	-0.0152	-0.0034	-0.0221	0.0143	-0.0267	-0.0061	-0.0098	-0.0393	-0.0115	-0.0323	-0.0152
D13S317	0.0005	0.0042	0.0342	0.0059	0.0329	0.0160	0.0019	-0.0010	0.0856	0.0025	0.0294	0.0716	-0.1490	-0.0207	0.0329	0.0074
D16S539	0.0061	0.0029	0.0709	-0.0277	0.0422	0.0189	0.0090	-0.0136	-0.0075	0.0404	-0.0046	0.1079	-0.1760	0.0074	0.0422	0.0013
D17S1301	0.0036	0.0461	-0.0333	-0.0430	0.0275	-0.0001	0.0145	-0.0406	-0.0295	-0.0976	-0.0567	0.0550	0.0454	0.0691	0.0275	-0.0010
D18S51	0.0266	-0.0277	-0.0384	0.0221	0.0058	-0.0022	0.0301	0.0124	-0.0315	-0.0402	0.0241	-0.0334	0.0145	-0.0284	0.0058	-0.0050
D19S433	-0.0338	0.0041	-0.0001	-0.0082	-0.0030	-0.0085	-0.0375	-0.0182	-0.0438	0.0221	0.0042	-0.0249	-0.0790	0.0206	-0.0030	-0.0178
D1S1656	-0.0129	0.0075	0.0251	-0.0091	0.0082	0.0036	-0.0058	-0.0436	0.0114	0.0124	-0.0275	0.0427	0.1100	0.0063	0.0082	0.0131
D20S482	-0.0173	-0.0266	0.0659	0.0333	-0.1080	-0.0082	-0.0035	-0.0726	-0.0109	0.0597	0.0477	0.0686	-0.0647	-0.0330	-0.1080	-0.0115
D21S11	0.0080	-0.0132	0.0629	0.0084	0.0549	0.0238	0.0125	-0.0217	0.0434	0.0533	0.0088	0.0694	-0.0004	-0.0299	0.0549	0.0209
D22S1045	0.0411	-0.0700	0.0285	-0.0339	0.0134	-0.0019	0.0262	0.1048	-0.1029	0.0861	-0.0422	-0.0508	0.0254	-0.0714	0.0134	0.0030
D2S1338	-0.0122	0.0339	0.0130	0.0045	0.0088	0.0095	-0.0064	-0.0331	0.0555	-0.0375	0.0146	0.0791	-0.0593	0.0282	0.0088	0.0055
D2S441	-0.0320	0.0057	-0.0280	-0.0108	-0.0007	-0.0135	-0.0398	-0.0027	0.0344	-0.0810	-0.0213	0.0386	0.0579	-0.0051	-0.0007	-0.0026
D3S1358	-0.0183	0.0275	0.0295	0.0309	0.1111	0.0367	-0.0204	-0.0060	-0.0245	-0.0380	0.0442	0.1266	-0.0447	0.0411	0.1111	0.0209
D4S2408	0.0341	0.1224	0.0805	0.0570	-0.0073	0.0563	0.0232	0.0684	-0.0045	0.0655	0.0597	0.0959	0.0461	0.1507	-0.0073	0.0557
D5S818	0.0251	-0.0286	-0.0129	-0.0135	0.1154	0.0157	0.0314	-0.0045	-0.0365	-0.0183	-0.0111	-0.0063	-0.0248	-0.0304	0.1154	0.0011
D6S1043	-0.0013	-0.0111	0.0101	0.0276	-0.0936	-0.0128	-0.0086	0.0246	-0.0095	0.0560	0.0323	-0.0549	-0.0083	-0.0158	-0.0936	-0.0079
D7S820	0.0030	-0.0523	0.0535	-0.0279	0.0479	0.0046	0.0019	0.0107	-0.0583	0.1057	-0.0169	-0.0271	-0.0959	-0.0513	0.0479	-0.0098
D8S1179	-0.0013	0.0402	-0.0268	0.0253	-0.0431	-0.0017	0.0129	-0.0577	0.0625	-0.0616	0.0302	0.0190	-0.0021	0.0345	-0.0431	-0.0008
D9S1122	0.0184	-0.0032	-0.0081	-0.0329	0.0608	0.0074	-0.0095	0.1244	-0.0516	0.0268	-0.0408	-0.0593	0.0286	0.0081	0.0608	0.0112
FGA	-0.0111	0.0005	0.0134	0.0012	-0.0060	-0.0005	-0.0030	-0.0447	-0.0035	-0.0283	-0.0076	0.0733	0.0550	0.0017	-0.0060	0.0040
PentaD	0.0153	-0.0118	0.0387	-0.0324	0.1080	0.0230	0.0163	0.0069	0.0428	0.0663	-0.0399	0.0015	0.0147	-0.0329	0.1080	0.0199
PentaE	0.0032	0.0339	-0.0007	0.0051	0.0417	0.0166	0.0025	0.0036	0.0934	0.0056	0.0003	-0.0105	0.0370	0.0156	0.0417	0.0211
TH01	-0.0316	0.0476	0.0851	0.0174	0.0457	0.0323	-0.0254	-0.0556	0.1189	0.0273	0.0305	0.1598	-0.0638	0.0264	0.0457	0.0282
TPOX	0.0195	-0.0153	-0.0063	-0.0162	-0.1472	-0.0327	0.0068	0.0653	-0.1024	-0.0169	-0.0194	0.0123	0.0040	0.0085	-0.1472	-0.0197
vWA	0.0123	0.0458	0.0089	-0.0154	0.0042	0.0111	-0.0027	0.0678	0.1537	-0.0277	0.0033	0.0464	-0.1389	0.0128	0.0042	0.0131
All	0.0002	0.0046	0.0151	-0.0018	0.0126	0.0061	-0.0003	0.0007	0.0075	0.0056	0.0009	0.0263	-0.0181	0.0026	0.0126	0.0042

Table 1.10: Locus-specific $\hat{\beta}_{IS}$ for each genetic-analysis group and subgroup for length-based autosomal STR data.

	Genetic-analysis group						Subgroup									
Locus	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**All**	**AfAm**	**AFR**	**AMR**	**Asian**	**Cauc**	**EAS**	**EUR**	**Hisp**	**SAS**	**All**
CSF1PO	-0.0654	0.0068	-0.0619	0.0266	0.1004	0.0013	-0.0528	-0.1009	-0.0490	-0.0448	0.0211	-0.0778	0.0803	0.0281	0.1029	-0.0103
D10S1248	-0.0414	0.0149	0.0389	0.0190	-0.0356	-0.0008	-0.0507	-0.0142	-0.0230	0.0376	0.0223	0.0365	-0.0142	0.0247	-0.0375	-0.0021
D12S391	0.0086	-0.0124	0.0122	-0.0366	-0.0028	-0.0062	0.0082	0.0070	0.0189	0.0043	-0.0319	0.0211	-0.0705	-0.0231	-0.0034	-0.0077
D13S317	0.1337	-0.0192	0.0438	0.0464	0.0210	0.0451	0.1266	0.1268	0.0621	0.0033	0.0558	0.0788	-0.0685	-0.0556	0.0130	0.0380
D16S539	-0.0011	0.0070	0.0946	0.0241	0.0731	0.0395	0.0152	-0.0490	-0.0114	0.0741	0.0463	0.1285	-0.0942	0.0168	0.0760	0.0225
D17S1301	0.0200	0.0522	-0.0475	-0.0295	0.0760	0.0142	0.0280	-0.0126	0.0174	-0.1015	-0.0400	0.0242	0.0367	0.0631	0.0760	0.0101
D18S51	0.0256	-0.0296	-0.0066	0.0237	0.0501	0.0127	0.0265	0.0207	-0.0170	0.0008	0.0240	-0.0170	0.0198	-0.0340	0.0498	0.0082
D19S433	-0.0747	0.0347	0.0254	0.0535	-0.0174	0.0043	-0.0852	-0.0525	-0.0229	0.0434	0.0632	-0.0064	-0.0314	0.0481	-0.0211	-0.0072
D1S1656	0.0224	0.0108	0.0729	-0.0176	0.0233	0.0224	0.0230	0.0107	0.0000	0.0637	-0.0374	0.0812	0.0914	0.0118	0.0215	0.0296
D20S482	0.0121	-0.0077	0.0145	0.0228	-0.0355	0.0012	0.0447	-0.0996	0.0096	-0.0193	0.0355	0.0741	-0.0307	-0.0079	-0.0317	-0.0028
D21S11	0.0102	-0.0014	0.0938	0.0252	0.0910	0.0438	0.0070	0.0163	0.0632	0.0851	0.0265	0.1035	0.0069	-0.0229	0.0898	0.0417
D22S1045	-0.0375	0.0535	0.0291	0.0254	0.1267	0.0394	-0.0571	0.0217	0.0940	0.0784	0.0095	-0.0445	0.1005	0.0363	0.1236	0.0403
D2S1338	-0.0123	0.0516	0.0328	0.0182	0.0264	0.0233	-0.0078	-0.0341	0.0649	-0.0152	0.0266	0.0956	-0.0391	0.0469	0.0260	0.0182
D2S441	-0.0102	0.0452	-0.0298	0.0004	0.1041	0.0219	-0.0273	0.0389	0.1152	-0.0597	-0.0237	0.0018	0.1251	0.0186	0.1010	0.0322
D3S1358	-0.0036	0.0341	0.0972	-0.0049	0.0937	0.0433	-0.0054	-0.0180	0.0163	0.0221	0.0037	0.1879	-0.0903	0.0342	0.0898	0.0267
D4S2408	0.0299	0.2003	0.1515	0.0849	0.0343	0.1002	0.0059	0.0787	0.1805	0.1237	0.0763	0.1694	0.0756	0.1965	0.0252	0.1035
D5S818	0.0172	0.0207	-0.0733	0.0490	0.1950	0.0417	0.0308	0.0016	0.0592	-0.0842	0.0596	-0.0395	0.0360	0.0185	0.2012	0.0315
D6S1043	-0.0159	-0.0218	0.0042	0.0804	-0.0095	0.0075	-0.0181	0.0131	-0.0013	0.0405	0.0833	-0.0358	0.0894	-0.0225	-0.0055	0.0159
D7S820	0.0376	-0.0270	0.0929	-0.0418	0.0503	0.0224	0.0368	0.0247	-0.0454	0.1510	-0.0369	0.0086	-0.0972	-0.0257	0.0471	0.0070
D8S1179	0.0386	0.0687	-0.0542	0.0512	-0.0488	0.0111	0.0395	0.0119	0.0982	-0.0957	0.0532	-0.0119	0.0042	0.0533	-0.0540	0.0110
D9S1122	-0.0094	0.0532	-0.0386	-0.0050	0.0488	0.0098	-0.0368	0.0823	0.0614	-0.0235	-0.0176	-0.0689	0.0544	0.0464	0.0454	0.0159
FGA	-0.0110	-0.0103	0.0328	0.0187	-0.0112	0.0038	-0.0035	-0.0464	-0.0204	-0.0025	0.0081	0.0784	0.0807	-0.0083	-0.0120	0.0082
PentaD	-0.0245	-0.0036	0.0853	-0.0011	0.1447	0.0402	-0.0153	-0.0316	0.0582	0.1322	-0.0040	0.0363	0.0605	-0.0149	0.1497	0.0412
PentaE	0.0230	0.0404	0.0039	0.0360	0.0748	0.0356	0.0181	0.0396	0.0919	0.0028	0.0306	0.0038	0.0657	0.0237	0.0742	0.0389
TH01	0.0454	0.0773	0.2047	0.0561	0.0926	0.0952	0.0396	0.0276	0.1443	0.1438	0.0658	0.2688	-0.0631	0.0455	0.0848	0.0841
TPOX	-0.0802	0.0188	0.1358	0.0720	-0.1217	0.0049	-0.1063	-0.0593	-0.0432	0.0991	0.0478	0.1537	0.1211	0.0187	-0.1387	0.0103
vWA	0.0136	0.0677	0.0295	-0.0064	0.0239	0.0257	0.0013	0.0680	0.1783	0.0030	0.0132	0.0689	-0.1186	0.0352	0.0252	0.0305
All	0.0026	0.0260	0.0370	0.0218	0.0431	0.0261	0.0002	0.0033	0.0410	0.0255	0.0216	0.0487	0.0117	0.0192	0.0415	0.0236

Table 1.11: Locus-specific $\hat{\beta}_{IT}$ for each genetic-analysis group and subgroup for length-based autosomal STR data.

	Genetic-analysis group						Subgroup									
Locus	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**All**	**AfAm**	**AFR**	**AMR**	**Asian**	**Cauc**	**EAS**	**EUR**	**Hisp**	**SAS**	**All**
CSF1PO	-0.0461	0.0276	-0.0223	0.0406	0.0193	0.0038	-0.0404	-0.0564	0.0099	-0.0032	0.0417	-0.0278	0.0516	0.0365	0.0225	0.0038
D10S1248	-0.0256	0.0309	0.0208	0.0277	0.0113	0.0130	-0.0294	-0.0217	0.0074	0.0062	0.0289	0.0326	0.0092	0.0352	0.0095	0.0087
D12S391	-0.0044	-0.0066	0.0261	-0.0077	0.0333	0.0081	-0.0033	-0.0062	0.0057	0.0278	-0.0065	0.0231	-0.0081	-0.0083	0.0343	0.0065
D13S317	0.0732	-0.0030	0.0583	0.0198	0.0306	0.0358	0.0681	0.0729	0.0009	0.0489	0.0135	0.0690	0.0314	-0.0090	0.0270	0.0359
D16S539	-0.0042	0.0074	0.0288	0.0533	0.0150	0.0201	0.0076	-0.0330	-0.0021	0.0369	0.0521	0.0249	0.0712	0.0114	0.0166	0.0206
D17S1301	0.0147	0.0081	-0.0163	0.0146	0.0515	0.0145	0.0112	0.0285	0.0471	-0.0019	0.0174	-0.0405	-0.0074	-0.0048	0.0515	0.0112
D18S51	-0.0017	-0.0021	0.0325	-0.0002	0.0463	0.0149	-0.0031	0.0044	0.0126	0.0415	-0.0020	0.0179	0.0049	-0.0052	0.0462	0.0130
D19S433	-0.0378	0.0300	0.0271	0.0621	-0.0184	0.0126	-0.0456	-0.0313	0.0220	0.0206	0.0588	0.0206	0.0466	0.0252	-0.0232	0.0104
D1S1656	0.0036	0.0073	0.0628	-0.0005	0.0136	0.0174	-0.0014	0.0130	-0.0027	0.0670	-0.0023	0.0533	-0.0105	0.0077	0.0115	0.0151
D20S482	0.0292	0.0188	-0.0572	-0.0104	0.0658	0.0093	0.0485	-0.0249	0.0207	-0.0880	-0.0124	0.0063	0.0323	0.0247	0.0693	0.0085
D21S11	0.0244	0.0138	0.0116	0.0264	0.0547	0.0262	0.0150	0.0690	0.0228	0.0048	0.0257	0.0225	0.0273	0.0104	0.0541	0.0280
D22S1045	-0.0815	0.1125	0.0011	0.0579	0.1153	0.0410	-0.0847	-0.0921	0.1703	-0.0077	0.0503	0.0068	0.0777	0.0988	0.1123	0.0369
D2S1338	-0.0207	0.0307	0.0197	0.0288	0.0061	0.0129	-0.0242	-0.0209	0.0282	0.0138	0.0241	0.0189	0.0377	0.0273	0.0034	0.0120
D2S441	0.0260	0.0084	0.0025	0.0290	0.0943	0.0320	0.0223	0.0319	0.0218	0.0128	0.0182	-0.0168	0.0847	0.0026	0.0926	0.0300
D3S1358	-0.0160	0.0319	0.1072	0.0156	-0.0049	0.0268	-0.0217	-0.0117	0.0674	0.0902	0.0110	0.1198	0.0132	0.0189	-0.0087	0.0309
D4S2408	0.0363	0.0612	0.0468	0.0548	0.0763	0.0551	0.0291	0.0589	0.0881	0.0325	0.0519	0.0625	0.0459	0.0526	0.0737	0.0550
D5S818	-0.0291	0.0490	-0.0217	0.0315	0.0584	0.0176	-0.0190	-0.0356	0.0936	-0.0273	0.0377	-0.0024	0.0274	0.0435	0.0637	0.0202
D6S1043	-0.0156	-0.0104	-0.0052	0.0545	0.0776	0.0202	-0.0097	-0.0151	0.0089	-0.0155	0.0530	0.0189	0.0975	-0.0064	0.0814	0.0237
D7S820	0.0351	0.0244	0.0404	-0.0121	0.0000	0.0175	0.0349	0.0153	0.0133	0.0517	-0.0186	0.0296	0.0000	0.0239	-0.0036	0.0163
D8S1179	0.0144	0.0218	0.0052	0.0382	0.0051	0.0169	0.0071	0.0337	0.0186	-0.0011	0.0393	0.0103	0.0183	0.0208	0.0034	0.0167
D9S1122	-0.0026	0.0435	0.0293	-0.0053	-0.0232	0.0083	-0.0048	-0.0167	0.0719	0.0110	-0.0116	0.0423	-0.0076	0.0288	-0.0284	0.0094
FGA	-0.0032	-0.0123	0.0211	0.0187	-0.0036	0.0041	-0.0049	-0.0012	-0.0210	0.0265	0.0166	0.0070	0.0286	-0.0110	-0.0045	0.0040
PentaD	-0.0396	0.0088	0.0459	0.0310	0.0419	0.0176	-0.0312	-0.0378	0.0170	0.0693	0.0354	0.0309	0.0473	0.0183	0.0476	0.0219
PentaE	0.0195	0.0066	0.0054	0.0319	0.0350	0.0197	0.0153	0.0354	-0.0010	-0.0021	0.0310	0.0148	0.0305	0.0077	0.0343	0.0184
TH01	0.0757	0.0318	0.1317	0.0405	0.0444	0.0648	0.0641	0.0794	0.0294	0.1204	0.0371	0.1303	0.0013	0.0196	0.0357	0.0575
TPOX	-0.1014	0.0338	0.1414	0.0862	0.0225	0.0365	-0.1136	-0.1331	0.0539	0.1141	0.0650	0.1433	0.1177	0.0105	0.0076	0.0295
vWA	-0.0298	0.0194	0.0497	0.0095	0.0333	0.0164	-0.0248	-0.0283	0.0380	0.0719	0.0154	0.0445	0.0032	0.0201	0.0373	0.0197
All	-0.0031	0.0212	0.0297	0.0268	0.0327	0.0215	-0.0044	-0.0028	0.0300	0.0272	0.0244	0.0320	0.0315	0.0179	0.0316	0.0208

Table 1.12: Locus-specific $\hat{\beta}_{ST}$ for each genetic-analysis group and subgroup for sequence-based autosomal STR data.

	Genetic-analysis group						Subgroup									
Locus	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**All**	**AfAm**	**AFR**	**AMR**	**Asian**	**Cauc**	**EAS**	**EUR**	**Hisp**	**SAS**	**All**
CSF1PO	-0.0254	-0.0282	-0.0367	-0.0125	0.0845	-0.0041	-0.0169	-0.0543	-0.0569	-0.0389	-0.0190	-0.0460	0.0327	-0.0179	0.0845	-0.0156
D10S1248	-0.0180	-0.0160	0.0188	-0.0085	-0.0471	-0.0143	-0.0240	0.0077	-0.0303	0.0320	-0.0064	0.0044	-0.0232	-0.0106	-0.0471	-0.0109
D12S391	0.0026	-0.0113	-0.0087	-0.0023	-0.0533	-0.0143	0.0030	0.0037	0.0190	-0.0153	0.0052	0.0030	-0.0487	-0.0218	-0.0533	-0.0116
D13S317	0.0110	0.0053	0.0086	0.0105	0.0402	0.0151	0.0120	0.0111	0.0300	-0.0081	0.0214	0.0241	-0.0572	-0.0024	0.0402	0.0079
D16S539	0.0064	0.0029	0.0709	-0.0274	0.0387	0.0185	0.0095	-0.0136	-0.0075	0.0404	-0.0042	0.1079	-0.1760	0.0074	0.0387	0.0011
D17S1301	0.0035	0.0461	-0.0289	-0.0430	0.0275	0.0007	0.0143	-0.0406	-0.0295	-0.0976	-0.0567	0.0638	0.0454	0.0691	0.0275	0.0001
D18S51	0.0291	-0.0291	-0.0384	0.0203	0.0058	-0.0023	0.0315	0.0185	-0.0277	-0.0402	0.0217	-0.0334	0.0170	-0.0312	0.0058	-0.0040
D19S433	-0.0318	0.0045	0.0019	-0.0085	0.0046	-0.0061	-0.0352	-0.0179	-0.0432	0.0258	0.0037	-0.0249	-0.0790	0.0208	0.0046	-0.0161
D1S1656	0.0067	-0.0078	0.0346	-0.0082	0.0136	0.0074	0.0093	-0.0044	0.0266	0.0205	-0.0253	0.0527	0.1037	-0.0188	0.0136	0.0197
D20S482	-0.0173	-0.0266	0.0683	0.0333	-0.1080	-0.0077	-0.0035	-0.0726	-0.0109	0.0635	0.0477	0.0686	-0.0647	-0.0330	-0.1080	-0.0110
D21S11	0.0071	-0.0085	0.0326	0.0204	0.0817	0.0263	0.0039	0.0118	0.0523	0.0308	0.0306	0.0331	-0.0465	-0.0269	0.0817	0.0188
D22S1045	0.0411	-0.0707	0.0285	-0.0339	0.0134	-0.0021	0.0262	0.1048	-0.1125	0.0861	-0.0422	-0.0508	0.0254	-0.0686	0.0134	0.0022
D2S1338	-0.0110	0.0354	0.0240	0.0039	-0.0367	0.0028	-0.0108	-0.0110	0.0321	-0.0318	0.0103	0.1012	-0.0353	0.0372	-0.0367	0.0059
D2S441	-0.0235	0.0003	-0.0141	-0.0084	0.0137	-0.0067	-0.0341	0.0178	0.0607	-0.0537	-0.0207	0.0385	0.0728	-0.0194	0.0137	0.0079
D3S1358	0.0049	0.0274	-0.0320	0.0231	0.0553	0.0167	0.0120	-0.0238	0.0552	-0.0457	0.0331	-0.0107	-0.0351	0.0159	0.0553	0.0067
D4S2408	0.0336	0.0602	0.0597	0.0513	-0.0219	0.0368	0.0220	0.0720	-0.0874	0.0490	0.0511	0.0709	0.0591	0.1019	-0.0219	0.0358
D5S818	0.0331	0.0108	-0.0246	-0.0003	0.0349	0.0105	0.0365	0.0157	0.0014	-0.0462	0.0055	0.0061	-0.0315	0.0119	0.0349	0.0037
D6S1043	0.0004	-0.0105	0.0101	0.0281	-0.0936	-0.0122	-0.0074	0.0287	-0.0095	0.0560	0.0328	-0.0549	-0.0081	-0.0151	-0.0936	-0.0071
D7S820	0.0010	-0.0553	0.0561	-0.0279	0.0516	0.0049	-0.0007	0.0107	-0.0583	0.1057	-0.0169	-0.0205	-0.0959	-0.0551	0.0516	-0.0093
D8S1179	0.0080	0.0051	-0.0157	0.0063	-0.0257	-0.0045	0.0196	-0.0371	0.0415	-0.0217	0.0083	-0.0085	-0.0047	-0.0065	-0.0257	-0.0039
D9S1122	-0.0025	0.0114	-0.0291	-0.0186	0.0604	0.0047	-0.0235	0.0782	0.0097	-0.0092	-0.0302	-0.0571	0.0699	0.0120	0.0604	0.0129
FGA	-0.0062	-0.0039	0.0134	0.0014	-0.0060	-0.0003	0.0028	-0.0436	-0.0130	-0.0283	-0.0072	0.0733	0.0550	-0.0007	-0.0060	0.0034
PentaD	0.0153	-0.0118	0.0349	-0.0324	0.1080	0.0223	0.0163	0.0069	0.0428	0.0555	-0.0399	0.0064	0.0147	-0.0329	0.1080	0.0193
PentaE	0.0018	0.0348	-0.0007	0.0051	0.0212	0.0124	0.0003	0.0050	0.0934	0.0056	0.0003	-0.0105	0.0370	0.0168	0.0212	0.0189
TH01	-0.0316	0.0481	0.0851	0.0174	0.0516	0.0336	-0.0254	-0.0556	0.1189	0.0273	0.0305	0.1598	-0.0638	0.0272	0.0516	0.0289
TPOX	0.0195	-0.0153	-0.0063	-0.0190	-0.1472	-0.0332	0.0068	0.0653	-0.1024	-0.0169	-0.0226	0.0123	0.0040	0.0085	-0.1472	-0.0201
vWA	-0.0052	0.0374	0.0095	-0.0104	-0.0097	0.0042	-0.0118	0.0202	0.1362	-0.0366	0.0055	0.0543	-0.1097	0.0062	-0.0097	0.0057
All	0.0022	0.0018	0.0114	-0.0008	0.0071	0.0043	0.0014	0.0043	0.0085	0.0039	0.0013	0.0201	-0.0127	-0.0013	0.0071	0.0036

Table 1.13: Locus-specific $\hat{\beta}_{IS}$ for each genetic-analysis group and subgroup for sequence-based autosomal STR data.

	Genetic-analysis group						Subgroup									
Locus	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**All**	**AfAm**	**AFR**	**AMR**	**Asian**	**Cauc**	**EAS**	**EUR**	**Hisp**	**SAS**	**All**
CSF1PO	-0.0726	0.0001	-0.0598	0.0285	0.1022	-0.0003	-0.0580	-0.1137	-0.0465	-0.0423	0.0235	-0.0751	0.0826	0.0192	0.1051	-0.0117
D10S1248	-0.0440	0.0153	0.0392	0.0194	-0.0352	-0.0011	-0.0541	-0.0138	-0.0227	0.0380	0.0227	0.0369	-0.0138	0.0250	-0.0371	-0.0021
D12S391	-0.0017	-0.0180	0.0176	-0.0101	-0.0181	-0.0061	-0.0003	-0.0024	0.0246	0.0129	-0.0013	0.0260	-0.0572	-0.0302	-0.0171	-0.0050
D13S317	0.0835	0.0024	0.0663	0.0301	0.0696	0.0504	0.0793	0.0832	0.0309	0.0412	0.0346	0.0914	-0.0240	-0.0114	0.0662	0.0435
D16S539	0.0023	0.0103	0.0976	0.0273	0.0531	0.0381	0.0170	-0.0470	-0.0096	0.0758	0.0480	0.1301	-0.0922	0.0187	0.0546	0.0217
D17S1301	0.0182	0.0538	-0.0457	-0.0278	0.0776	0.0152	0.0253	-0.0109	0.0190	-0.0997	-0.0383	0.0258	0.0383	0.0646	0.0775	0.0113
D18S51	0.0274	-0.0313	-0.0047	0.0201	0.0519	0.0127	0.0285	0.0228	-0.0148	0.0029	0.0198	-0.0148	0.0219	-0.0366	0.0518	0.0090
D19S433	-0.0707	0.0343	0.0290	0.0541	-0.0137	0.0066	-0.0824	-0.0498	-0.0202	0.0459	0.0623	-0.0038	-0.0288	0.0454	-0.0185	-0.0055
D1S1656	0.0102	-0.0005	0.0953	-0.0087	0.0270	0.0247	0.0080	0.0086	0.0239	0.0861	-0.0277	0.1032	0.0943	-0.0109	0.0250	0.0345
D20S482	0.0125	-0.0073	0.0149	0.0232	-0.0351	0.0016	0.0451	-0.0992	0.0100	-0.0189	0.0359	0.0745	-0.0303	-0.0075	-0.0313	-0.0024
D21S11	0.0314	0.0054	0.0439	0.0463	0.1319	0.0518	0.0188	0.0800	0.0739	0.0354	0.0556	0.0549	-0.0179	-0.0163	0.1313	0.0462
D22S1045	-0.0370	0.0497	0.0296	0.0259	0.1272	0.0391	-0.0563	0.0224	0.0769	0.0791	0.0102	-0.0437	0.1012	0.0370	0.1242	0.0390
D2S1338	-0.0319	0.0650	0.0433	0.0326	-0.0304	0.0157	-0.0352	-0.0321	0.0594	-0.0176	0.0341	0.1182	0.0038	0.0635	-0.0331	0.0179
D2S441	0.0031	0.0087	-0.0115	0.0209	0.1067	0.0256	-0.0111	0.0491	0.0812	-0.0403	-0.0022	0.0223	0.1514	-0.0168	0.1051	0.0376
D3S1358	-0.0110	0.0584	0.0786	0.0383	0.0507	0.0430	-0.0094	-0.0357	0.1189	0.0486	0.0437	0.1104	-0.0215	0.0345	0.0471	0.0374
D4S2408	0.0687	0.1177	0.1037	0.1032	0.0561	0.0899	0.0504	0.1267	0.0083	0.0799	0.1004	0.1289	0.1024	0.1491	0.0534	0.0888
D5S818	0.0049	0.0593	-0.0469	0.0312	0.0913	0.0280	0.0182	-0.0193	0.0949	-0.0747	0.0430	0.0037	-0.0032	0.0549	0.0964	0.0238
D6S1043	-0.0152	-0.0210	0.0050	0.0810	-0.0088	0.0082	-0.0172	0.0140	-0.0005	0.0413	0.0841	-0.0349	0.0902	-0.0216	-0.0046	0.0168
D7S820	0.0360	-0.0296	0.0942	-0.0403	0.0517	0.0224	0.0342	0.0258	-0.0443	0.1519	-0.0358	0.0097	-0.0960	-0.0299	0.0482	0.0071
D8S1179	0.0223	0.0269	-0.0104	0.0443	-0.0205	0.0125	0.0266	-0.0021	0.0593	-0.0228	0.0473	0.0019	0.0137	0.0144	-0.0222	0.0129
D9S1122	-0.0051	0.0543	0.0011	-0.0240	0.0386	0.0130	-0.0284	0.0629	0.0809	0.0019	-0.0422	-0.0124	0.0629	0.0405	0.0337	0.0222
FGA	-0.0095	-0.0162	0.0342	0.0201	-0.0097	0.0038	-0.0021	-0.0448	-0.0343	-0.0010	0.0095	0.0798	0.0820	-0.0117	-0.0106	0.0074
PentaD	-0.0237	-0.0028	0.0792	-0.0004	0.1454	0.0395	-0.0144	-0.0307	0.0590	0.1209	-0.0031	0.0371	0.0613	-0.0140	0.1505	0.0407
PentaE	0.0212	0.0412	0.0047	0.0368	0.0554	0.0319	0.0156	0.0403	0.0925	0.0035	0.0313	0.0045	0.0664	0.0244	0.0548	0.0370
TH01	0.0465	0.0784	0.2057	0.0572	0.0937	0.0963	0.0403	0.0283	0.1449	0.1444	0.0664	0.2693	-0.0624	0.0462	0.0854	0.0847
TPOX	-0.0800	0.0190	0.1360	0.0688	-0.1215	0.0045	-0.1061	-0.0591	-0.0430	0.0992	0.0439	0.1538	0.1212	0.0188	-0.1386	0.0100
vWA	-0.0351	0.0561	0.0588	-0.0008	0.0239	0.0206	-0.0369	-0.0075	0.1690	0.0379	0.0209	0.0963	-0.1062	0.0262	0.0279	0.0253
All	-0.0009	0.0229	0.0408	0.0259	0.0396	0.0257	-0.0030	0.0015	0.0383	0.0309	0.0257	0.0515	0.0191	0.0167	0.0385	0.0244

Table 1.14: Locus-specific $\hat{\beta}_{IT}$ for each genetic-analysis group and subgroup for sequence-based autosomal STR data.

	Genetic-analysis group						Subgroup									
Locus	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**All**	**AfAm**	**AFR**	**AMR**	**Asian**	**Cauc**	**EAS**	**EUR**	**Hisp**	**SAS**	**All**
CSF1PO	-0.0449	0.0291	-0.0218	0.0386	0.0181	0.0038	-0.0397	-0.0557	0.0124	-0.0058	0.0393	-0.0257	0.0496	0.0371	0.0208	0.0036
D10S1248	-0.0252	0.0305	0.0212	0.0273	0.0105	0.0128	-0.0289	-0.0220	0.0068	0.0060	0.0285	0.0338	0.0086	0.0349	0.0087	0.0085
D12S391	0.0162	-0.0074	0.0326	-0.0256	0.0282	0.0088	0.0116	0.0284	0.0048	0.0301	-0.0257	0.0305	-0.0303	-0.0115	0.0276	0.0073
D13S317	0.1333	-0.0235	0.0100	0.0407	-0.0120	0.0297	0.1249	0.1276	-0.0251	0.0009	0.0273	0.0083	0.0688	-0.0343	-0.0202	0.0309
D16S539	-0.0072	0.0041	0.0258	0.0504	0.0326	0.0211	0.0062	-0.0350	-0.0040	0.0353	0.0507	0.0238	0.0680	0.0095	0.0356	0.0211
D17S1301	0.0164	0.0065	-0.0139	0.0129	0.0502	0.0144	0.0138	0.0267	0.0453	-0.0041	0.0157	-0.0321	-0.0087	-0.0063	0.0501	0.0112
D18S51	-0.0010	-0.0018	0.0305	0.0017	0.0446	0.0148	-0.0037	0.0085	0.0138	0.0392	-0.0001	0.0156	0.0054	-0.0054	0.0443	0.0131
D19S433	-0.0396	0.0307	0.0255	0.0612	-0.0144	0.0127	-0.0460	-0.0337	0.0198	0.0219	0.0593	0.0179	0.0434	0.0281	-0.0180	0.0103
D1S1656	0.0348	0.0034	0.0491	-0.0084	0.0153	0.0188	0.0287	0.0517	-0.0114	0.0520	-0.0096	0.0405	-0.0199	0.0056	0.0134	0.0168
D20S482	0.0288	0.0184	-0.0548	-0.0107	0.0644	0.0092	0.0481	-0.0257	0.0202	-0.0836	-0.0127	0.0065	0.0314	0.0242	0.0679	0.0085
D21S11	0.0022	0.0116	0.0332	0.0170	0.0387	0.0205	-0.0056	0.0370	0.0210	0.0338	0.0178	0.0372	0.0072	0.0067	0.0374	0.0214
D22S1045	-0.0820	0.1153	0.0007	0.0573	0.1149	0.0413	-0.0855	-0.0922	0.1780	-0.0080	0.0495	0.0057	0.0773	0.1004	0.1118	0.0374
D2S1338	-0.0001	0.0183	0.0201	0.0139	0.0179	0.0140	-0.0014	-0.0011	0.0104	0.0212	0.0123	0.0184	0.0185	0.0193	0.0175	0.0128
D2S441	0.0211	0.0397	-0.0018	0.0110	0.1047	0.0350	0.0120	0.0415	0.0839	0.0193	-0.0024	-0.0380	0.0718	0.0235	0.1016	0.0348
D3S1358	0.0144	0.0068	0.0698	-0.0369	-0.0185	0.0071	0.0147	-0.0120	0.0396	0.0577	-0.0423	0.0710	-0.0441	-0.0071	-0.0230	0.0061
D4S2408	-0.0043	0.0889	0.0775	0.0296	0.0412	0.0466	-0.0177	0.0114	0.1841	0.0626	0.0178	0.0819	0.0313	0.0543	0.0322	0.0509
D5S818	-0.0081	0.0480	-0.0596	0.0617	0.0910	0.0266	-0.0006	0.0060	0.0921	-0.0649	0.0699	-0.0331	0.0591	0.0474	0.0980	0.0304
D6S1043	-0.0146	-0.0106	-0.0059	0.0542	0.0760	0.0198	-0.0093	-0.0116	0.0080	-0.0161	0.0527	0.0177	0.0968	-0.0066	0.0797	0.0235
D7S820	0.0347	0.0240	0.0418	-0.0135	0.0031	0.0180	0.0350	0.0143	0.0118	0.0511	-0.0198	0.0345	-0.0020	0.0242	-0.0003	0.0165
D8S1179	0.0398	0.0298	-0.0268	0.0266	-0.0059	0.0127	0.0269	0.0654	0.0385	-0.0324	0.0238	-0.0313	0.0062	0.0196	-0.0109	0.0118
D9S1122	-0.0283	0.0563	-0.0303	0.0270	-0.0121	0.0025	-0.0270	-0.0474	0.1071	-0.0516	0.0222	-0.0095	0.0269	0.0386	-0.0158	0.0048
FGA	0.0001	-0.0109	0.0197	0.0175	-0.0052	0.0042	-0.0005	-0.0019	-0.0168	0.0249	0.0155	0.0061	0.0277	-0.0100	-0.0061	0.0043
PentaD	-0.0404	0.0080	0.0486	0.0302	0.0422	0.0177	-0.0321	-0.0387	0.0164	0.0709	0.0345	0.0349	0.0466	0.0174	0.0477	0.0220
PentaE	0.0199	0.0069	0.0046	0.0311	0.0350	0.0195	0.0157	0.0361	-0.0011	-0.0027	0.0303	0.0140	0.0301	0.0083	0.0343	0.0184
TH01	0.0746	0.0313	0.1309	0.0394	0.0495	0.0652	0.0634	0.0785	0.0295	0.1199	0.0365	0.1306	0.0001	0.0197	0.0414	0.0577
TPOX	-0.1017	0.0336	0.1412	0.0868	0.0209	0.0362	-0.1138	-0.1329	0.0531	0.1139	0.0659	0.1432	0.1176	0.0104	0.0061	0.0293
vWA	0.0014	0.0230	0.0209	0.0088	0.0198	0.0148	0.0040	0.0005	0.0301	0.0297	0.0100	0.0239	0.0166	0.0227	0.0212	0.0176
All	0.0024	0.0215	0.0223	0.0236	0.0310	0.0202	0.0005	0.0025	0.0338	0.0200	0.0207	0.0232	0.0291	0.0167	0.0294	0.0196

Table 1.15: Locus-specific $\hat{\beta}_{WT}$ for each genetic-analysis group and subgroup for length-based autosomal STRs using allelic data.

	Genetic-analysis group						Subgroup									
Locus	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**All**	**AfAm**	**AFR**	**AMR**	**Asian**	**Cauc**	**EAS**	**EUR**	**Hisp**	**SAS**	**All**
CSF1PO	-0.0461	0.0276	-0.0224	0.0406	0.0201	0.0039	-0.0404	-0.0567	0.0095	-0.0034	0.0417	-0.0281	0.0519	0.0365	0.0232	0.0038
D10S1248	-0.0256	0.0308	0.0209	0.0277	0.0108	0.0129	-0.0294	-0.0217	0.0072	0.0063	0.0289	0.0327	0.0090	0.0352	0.0090	0.0086
D12S391	-0.0044	-0.0066	0.0261	-0.0078	0.0329	0.0080	-0.0033	-0.0061	0.0058	0.0277	-0.0065	0.0231	-0.0085	-0.0083	0.0339	0.0064
D13S317	0.0733	-0.0029	0.0583	0.0198	0.0310	0.0359	0.0681	0.0730	0.0011	0.0488	0.0136	0.0692	0.0310	-0.0090	0.0274	0.0359
D16S539	-0.0042	0.0074	0.0290	0.0532	0.0153	0.0202	0.0076	-0.0331	-0.0021	0.0371	0.0521	0.0257	0.0698	0.0114	0.0169	0.0206
D17S1301	0.0147	0.0082	-0.0164	0.0145	0.0518	0.0146	0.0112	0.0283	0.0469	-0.0024	0.0173	-0.0401	-0.0070	-0.0046	0.0517	0.0113
D18S51	-0.0016	-0.0022	0.0323	-0.0002	0.0464	0.0149	-0.0031	0.0045	0.0124	0.0413	-0.0019	0.0177	0.0051	-0.0053	0.0463	0.0130
D19S433	-0.0378	0.0300	0.0271	0.0621	-0.0183	0.0126	-0.0457	-0.0314	0.0217	0.0207	0.0588	0.0204	0.0459	0.0252	-0.0231	0.0103
D1S1656	0.0036	0.0073	0.0629	-0.0005	0.0138	0.0174	-0.0013	0.0129	-0.0026	0.0671	-0.0024	0.0537	-0.0096	0.0077	0.0116	0.0152
D20S482	0.0292	0.0188	-0.0570	-0.0103	0.0648	0.0091	0.0485	-0.0253	0.0206	-0.0876	-0.0123	0.0069	0.0318	0.0246	0.0683	0.0084
D21S11	0.0244	0.0138	0.0117	0.0264	0.0554	0.0264	0.0150	0.0691	0.0231	0.0049	0.0258	0.0228	0.0269	0.0103	0.0548	0.0281
D22S1045	-0.0815	0.1124	0.0012	0.0578	0.1154	0.0411	-0.0847	-0.0914	0.1696	-0.0072	0.0502	0.0064	0.0779	0.0987	0.1124	0.0369
D2S1338	-0.0207	0.0307	0.0198	0.0288	0.0057	0.0129	-0.0242	-0.0210	0.0284	0.0137	0.0241	0.0196	0.0374	0.0274	0.0031	0.0120
D2S441	0.0259	0.0084	0.0025	0.0290	0.0944	0.0320	0.0223	0.0320	0.0222	0.0125	0.0181	-0.0165	0.0853	0.0025	0.0927	0.0301
D3S1358	-0.0160	0.0320	0.1071	0.0156	-0.0044	0.0269	-0.0217	-0.0118	0.0677	0.0899	0.0110	0.1198	0.0129	0.0189	-0.0082	0.0310
D4S2408	0.0364	0.0613	0.0470	0.0548	0.0761	0.0551	0.0291	0.0593	0.0875	0.0328	0.0520	0.0630	0.0464	0.0528	0.0735	0.0552
D5S818	-0.0290	0.0490	-0.0218	0.0315	0.0587	0.0177	-0.0189	-0.0355	0.0936	-0.0275	0.0377	-0.0024	0.0272	0.0435	0.0640	0.0202
D6S1043	-0.0156	-0.0104	-0.0052	0.0545	0.0767	0.0200	-0.0097	-0.0149	0.0089	-0.0152	0.0531	0.0186	0.0974	-0.0064	0.0805	0.0236
D7S820	0.0351	0.0243	0.0405	-0.0121	0.0005	0.0176	0.0349	0.0154	0.0129	0.0522	-0.0186	0.0295	-0.0008	0.0238	-0.0031	0.0162
D8S1179	0.0144	0.0218	0.0052	0.0382	0.0048	0.0169	0.0072	0.0335	0.0189	-0.0012	0.0393	0.0102	0.0183	0.0208	0.0032	0.0167
D9S1122	-0.0026	0.0435	0.0292	-0.0053	-0.0226	0.0084	-0.0048	-0.0162	0.0719	0.0109	-0.0116	0.0419	-0.0069	0.0289	-0.0278	0.0096
FGA	-0.0033	-0.0123	0.0211	0.0187	-0.0037	0.0041	-0.0049	-0.0014	-0.0211	0.0264	0.0166	0.0075	0.0291	-0.0110	-0.0046	0.0041
PentaD	-0.0396	0.0088	0.0460	0.0310	0.0429	0.0178	-0.0312	-0.0378	0.0173	0.0695	0.0354	0.0310	0.0474	0.0182	0.0485	0.0220
PentaE	0.0195	0.0067	0.0054	0.0319	0.0351	0.0197	0.0153	0.0355	-0.0004	-0.0020	0.0310	0.0147	0.0308	0.0078	0.0345	0.0186
TH01	0.0757	0.0319	0.1320	0.0405	0.0448	0.0650	0.0641	0.0791	0.0302	0.1205	0.0371	0.1312	0.0008	0.0196	0.0361	0.0576
TPOX	-0.1014	0.0338	0.1414	0.0861	0.0211	0.0362	-0.1136	-0.1327	0.0532	0.1141	0.0650	0.1433	0.1177	0.0105	0.0062	0.0293
vWA	-0.0298	0.0195	0.0498	0.0095	0.0332	0.0164	-0.0248	-0.0281	0.0389	0.0717	0.0154	0.0448	0.0022	0.0201	0.0372	0.0197
All	-0.0031	0.0212	0.0298	0.0268	0.0328	0.0215	-0.0044	-0.0028	0.0301	0.0272	0.0245	0.0322	0.0314	0.0179	0.0316	0.0208

Table 1.16: Locus-specific $\hat{\beta}_{WT}$ for each genetic-analysis group and subgroup for sequence-based autosomal STRs using allelic data.

	Subgroup					
Locus	**AFR**	**AMR**	**EAS**	**EUR**	**SAS**	**All**
rs1005533	-0.0481	0.0384	0.2418	-0.0495	0.0481	0.0462
rs10092491	-0.0261	-0.0403	0.1072	-0.0027	-0.0336	0.0009
rs1015250	0.0600	0.2475	-0.0062	0.2777	-0.0019	0.1154
rs1024116	0.2031	-0.1141	0.5700	-0.1007	-0.0119	0.1093
rs1028528	0.4145	0.2949	0.0404	0.2942	0.0436	0.2175
rs1031825	0.0106	0.1451	-0.0101	0.1744	-0.0117	0.0617
rs10488710	-0.0784	-0.0529	0.0065	0.0122	0.1471	0.0069
rs10495407	0.6293	-0.2028	-0.1751	-0.1631	0.4023	0.0981
rs1058083	-0.0309	-0.0266	0.0009	0.0738	0.0023	0.0039
rs10773760	0.0536	-0.0171	0.0692	-0.1029	0.0478	0.0101
rs10776839	-0.0035	-0.0342	-0.0413	0.0316	0.0733	0.0052
rs1109037	-0.0084	-0.0040	-0.0099	-0.0016	0.0694	0.0091
rs1294331	-0.0225	-0.0667	0.1823	0.1562	-0.0297	0.0439
rs12997453	-0.0136	-0.0758	0.0056	-0.0461	0.2303	0.0201
rs13182883	-0.0191	-0.0056	-0.0316	0.1885	0.1803	0.0625
rs13218440	-0.0151	0.0726	-0.0154	-0.0196	-0.0232	-0.0002
rs1335873	0.7193	0.0630	0.0807	0.2642	0.4283	0.3111
rs1336071	-0.0100	-0.0465	0.0241	-0.0319	0.1681	0.0208
rs1355366	-0.0233	0.2482	0.4733	-0.0453	-0.0379	0.1230
rs1357617	0.0801	0.1194	0.0943	-0.1920	-0.0976	0.0009
rs1360288	0.3171	0.1946	-0.0954	-0.0835	-0.1138	0.0438
rs1382387	-0.0188	0.2118	0.1906	-0.0075	0.0068	0.0766
rs1413212	-0.0031	0.1309	-0.0780	0.1536	-0.0309	0.0345
rs1454361	0.1719	-0.0077	0.0370	-0.0323	-0.0263	0.0285
rs1463729	0.3378	0.0090	-0.0170	0.0668	0.0142	0.0822
rs1490413	-0.0075	0.0086	0.0281	0.0092	0.0537	0.0184
rs1493232	0.0029	0.1432	0.1553	0.1850	0.0174	0.1008
rs1498553	-0.0041	-0.0063	-0.0024	-0.0067	0.0120	-0.0015
rs1523537	0.0199	0.0458	0.0454	0.0028	0.1041	0.0436
rs1528460	0.7385	0.1331	0.0461	0.2657	0.0676	0.2502
rs159606	-0.0097	-0.0664	-0.0282	0.0022	0.1605	0.0117
rs1736442	0.0256	0.0202	-0.0544	-0.0500	0.0531	-0.0011
rs1821380	0.0222	-0.0444	-0.0201	0.0067	0.0918	0.0112
rs1886510	0.5517	-0.1523	0.3822	-0.2128	0.2171	0.1572
rs1979255	0.1189	-0.0299	-0.0333	0.0300	-0.0222	0.0127
rs2040411	0.4448	0.1215	0.1235	0.1078	-0.0094	0.1576
rs2046361	0.0333	0.1481	-0.0040	0.1280	0.0321	0.0675
rs2056277	0.1342	-0.0653	0.1948	-0.0906	-0.1822	-0.0018
rs2076848	0.2277	0.0474	-0.0202	-0.0626	-0.0320	0.0321
rs2107612	-0.2017	-0.0495	0.5890	-0.0471	-0.0394	0.0502
rs2111980	0.1586	0.0163	0.0655	0.0203	0.0367	0.0595
rs214955	0.0096	-0.0043	-0.0023	0.0119	0.0405	0.0111
rs221956	0.0022	0.0546	-0.0667	0.0055	0.0034	-0.0002
rs2269355	0.0196	0.0109	0.0699	0.0587	0.0044	0.0327
rs2342747	-0.0237	-0.1001	0.0357	-0.0566	0.2217	0.0154
rs2399332	-0.0696	0.0114	0.0217	-0.0514	0.1581	0.0140
rs251934	0.0940	-0.1985	0.5647	-0.1617	-0.0102	0.0577
rs279844	0.0126	-0.0023	0.0216	-0.0027	0.0121	0.0083
rs2830795	0.4436	0.3433	-0.1794	-0.0726	-0.0137	0.1042

rs2831700	0.0222	-0.0384	-0.0390	-0.0089	0.2687	0.0409
rs2920816	0.0804	-0.0381	0.0714	-0.0890	0.0523	0.0154
rs321198	-0.0136	0.0065	0.0245	0.0013	-0.0180	0.0001
rs338882	0.0131	0.0319	0.0626	-0.0046	-0.0049	0.0196
rs354439	0.0863	0.2143	0.0156	0.0453	0.0859	0.0895
rs3780962	0.0011	0.0322	0.0172	-0.0057	0.0107	0.0111
rs430046	0.1610	-0.0023	0.0053	-0.0666	-0.0316	0.0132
rs4364205	-0.0287	0.1288	-0.0384	-0.0440	0.0585	0.0152
rs445251	-0.0187	0.0200	0.1296	0.0292	-0.0161	0.0288
rs4530059	0.0060	-0.0950	0.0520	-0.1097	0.3228	0.0352
rs4606077	-0.1300	0.0427	0.1298	0.1053	-0.0333	0.0229
rs560681	0.0483	-0.0128	-0.0530	0.1582	-0.0869	0.0108
rs576261	0.0025	-0.0036	-0.0028	-0.0013	0.0207	0.0031
rs6444724	0.0039	0.0054	-0.0071	0.1193	-0.0080	0.0227
rs6811238	-0.0029	-0.0028	0.1080	0.0219	0.1111	0.0471
rs6955448	-0.0746	-0.0452	-0.0424	0.0663	0.0965	0.0001
rs7041158	-0.1090	0.0254	0.0351	0.0064	0.0879	0.0092
rs717302	-0.0750	0.0841	0.5394	-0.0794	0.0787	0.1096
rs719366	0.4072	-0.0482	0.0865	-0.1475	-0.1185	0.0359
rs722098	0.4365	0.0359	0.0406	0.2713	0.0589	0.1687
rs722290	-0.0024	-0.0057	0.0008	-0.0070	0.0170	0.0005
rs727811	0.3650	0.0314	0.1619	0.0270	0.1241	0.1419
rs729172	0.3588	-0.2650	0.6494	-0.2636	0.2993	0.1558
rs733164	-0.0377	-0.1272	0.3751	-0.0729	-0.0185	0.0238
rs735155	0.2627	0.0309	0.5469	0.0563	0.1441	0.2082
rs737681	-0.0693	0.0560	0.4370	-0.0449	-0.0634	0.0631
rs740598	-0.0022	-0.0276	-0.0725	-0.0194	0.1633	0.0083
rs740910	0.7350	-0.3972	0.8004	-0.4415	0.0868	0.1567
rs763869	-0.0167	-0.0006	0.0237	-0.0107	0.1564	0.0304
rs8037429	0.0349	-0.0032	0.0283	0.0092	-0.0118	0.0115
rs8078417	-0.0549	-0.1196	0.0274	0.0078	0.1826	0.0087
rs826472	0.7357	0.1457	0.2091	-0.2781	-0.1702	0.1284
rs873196	0.2717	-0.1554	0.4178	-0.2960	0.0413	0.0559
rs876724	0.3672	0.1198	-0.1723	-0.0972	0.0415	0.0518
rs891700	-0.0051	0.0007	0.0029	0.0026	-0.0039	-0.0006
rs901398	-0.0763	-0.0127	0.0404	-0.0889	0.1803	0.0085
rs907100	0.4129	0.0905	0.0528	0.0963	0.0759	0.1457
rs914165	0.2632	0.0308	0.1892	0.1093	0.1053	0.1395
rs917118	0.1442	0.0873	0.2053	0.1359	-0.0104	0.1125
rs938283	0.2935	0.0116	0.0428	-0.2721	-0.0681	0.0016
rs964681	0.2159	0.0473	0.1138	-0.0932	-0.1536	0.0260
rs987640	0.0051	0.0025	0.0060	0.0138	0.2225	0.0500
rs9905977	0.1298	0.0655	-0.0704	0.1195	-0.0977	0.0293
rs993934	0.0665	0.0888	-0.0247	-0.0163	-0.0271	0.0175
rs9951171	-0.0160	0.0026	0.0015	0.0543	-0.0220	0.0041
All	0.1088	0.0206	0.0864	0.0068	0.0506	0.0546

Table 1.17: Locus-specific $\hat{\beta}_{ST}$ for each subgroup over 94 iiSNP markers using the 1000GP data set.

	Subgroup					
Locus	**AFR**	**AMR**	**EAS**	**EUR**	**SAS**	**All**
rs1005533	-0.0858	0.2768	-0.0785	-0.2682	-0.0362	-0.0418
rs10092491	0.0471	0.1177	-0.0094	0.0383	-0.0217	0.0357
rs1015250	0.2639	0.3616	-0.1318	-0.1017	0.1625	0.1078
rs1024116	0.0835	-0.0875	-0.1145	-0.0761	0.0442	-0.0267
rs1028528	-0.0638	0.0005	0.0076	0.0914	0.1440	0.0441
rs1031825	0.0066	0.3554	-0.1032	-0.0501	-0.0120	0.0325
rs10488710	0.1321	0.0442	0.1219	-0.1656	0.0510	0.0383
rs10495407	-0.0807	0.0509	0.0155	0.2161	-0.1489	0.0470
rs1058083	0.0467	-0.1232	-0.0584	-0.1116	0.0243	-0.0433
rs10773760	0.0973	-0.0145	0.1011	-0.0370	-0.0977	0.0076
rs10776839	0.1990	-0.1533	0.1361	-0.1111	0.2653	0.0646
rs1109037	0.0253	0.0383	0.2732	-0.2467	0.1747	0.0515
rs1294331	0.1650	0.1873	0.1843	0.3014	0.1756	0.1997
rs12997453	-0.2876	0.1562	-0.0012	0.0000	0.2322	0.0110
rs13182883	0.1767	-0.0606	-0.1016	0.1720	-0.1828	0.0009
rs13218440	-0.1603	-0.0133	-0.1036	-0.0867	-0.0682	-0.0877
rs1335873	0.2294	-0.0835	0.0767	-0.0704	0.0314	0.0066
rs1336071	0.0655	0.1846	-0.0638	-0.0061	0.0924	0.0547
rs1355366	0.3196	0.2206	0.1524	0.4109	0.4936	0.3455
rs1357617	0.3907	0.3528	0.1459	0.1720	0.0018	0.2020
rs1360288	-0.0638	-0.0582	-0.0139	-0.0899	-0.0062	-0.0439
rs1382387	0.0378	0.1813	0.1569	-0.0442	0.0371	0.0651
rs1413212	0.0056	0.0146	0.1412	0.0842	0.2070	0.0943
rs1454361	0.1853	-0.0835	0.0519	0.2819	-0.1242	0.0582
rs1463729	0.2371	-0.0059	0.0872	0.2161	-0.0062	0.0949
rs1490413	-0.0626	0.0565	-0.1707	-0.1163	0.0371	-0.0516
rs1493232	0.1451	0.3750	0.0808	0.2037	-0.0998	0.1339
rs1498553	0.0273	0.0465	0.1107	-0.0270	0.0804	0.0474
rs1523537	-0.1059	-0.0829	0.1561	-0.0307	0.0724	0.0001
rs1528460	0.0920	0.1236	0.0851	-0.0704	-0.0370	0.0337
rs159606	-0.0125	0.0626	0.0428	-0.0364	0.0258	0.0169
rs1736442	0.2712	0.2437	0.3265	0.3116	0.0182	0.2381
rs1821380	0.0845	0.1266	-0.0711	0.0284	0.0453	0.0428
rs1886510	0.0156	0.1294	-0.1680	-0.2076	-0.1273	-0.0710
rs1979255	-0.0095	-0.0400	-0.0775	0.0848	-0.0258	-0.0149
rs2040411	-0.1918	-0.2191	0.0403	-0.0088	-0.0851	-0.0849
rs2046361	0.1329	0.2524	-0.0480	0.1493	0.1148	0.1151
rs2056277	-0.0333	0.0146	-0.1231	0.3155	0.0314	0.0537
rs2076848	0.0675	0.1609	0.0767	-0.3025	0.1440	0.0229
rs2107612	-0.0125	-0.0039	-0.0896	0.0346	-0.2040	-0.0488
rs2111980	-0.0472	0.1107	-0.0638	0.1252	0.0243	0.0331
rs214955	0.0847	0.0733	0.0052	0.1648	0.1309	0.0912
rs221956	0.0909	0.0263	0.0708	-0.0098	0.0160	0.0394
rs2269355	-0.1719	0.0952	0.0814	0.0753	-0.3047	-0.0478
rs2342747	0.0953	0.0397	-0.0469	0.1873	-0.1636	0.0338
rs2399332	0.1099	0.0439	0.1988	0.0136	0.0258	0.0794
rs251934	-0.1710	0.1859	0.0601	-0.1610	0.0008	-0.0196
rs279844	0.0129	0.1775	0.1185	0.1769	-0.2481	0.0482
rs2830795	0.0408	-0.0013	-0.1032	0.2450	0.1576	0.0720

rs2831700	-0.0788	0.1260	0.1147	-0.0575	0.0256	0.0279
rs2920816	0.2887	0.1721	-0.0532	0.0751	0.0453	0.1055
rs321198	0.1402	-0.0835	0.1892	0.0244	-0.0093	0.0517
rs338882	0.0129	-0.0709	-0.0059	-0.0948	0.0261	-0.0266
rs354439	0.1217	0.2344	-0.0232	0.1144	-0.2486	0.0339
rs3780962	-0.0943	-0.1045	0.0630	-0.0307	-0.0682	-0.0469
rs430046	0.0478	-0.1903	0.0738	0.0012	-0.0515	-0.0262
rs4364205	-0.1987	0.0419	0.0223	-0.0683	0.1537	-0.0145
rs445251	0.0890	-0.1389	0.0808	0.1298	-0.2057	-0.0120
rs4530059	0.1493	-0.0709	-0.0929	-0.0160	0.0314	-0.0028
rs4606077	-0.0485	-0.1681	-0.1263	-0.0920	-0.1167	-0.1082
rs560681	-0.1164	-0.3481	-0.2000	-0.0488	0.0690	-0.1294
rs576261	0.0847	0.1994	0.1068	-0.1068	-0.0815	0.0411
rs6444724	0.1307	-0.0214	-0.1032	0.2541	-0.0851	0.0293
rs6811238	0.0036	-0.0363	0.1219	0.0136	0.1007	0.0375
rs6955448	-0.0731	0.0065	0.1433	0.1720	-0.0449	0.0395
rs7041158	0.0847	0.1286	0.2494	0.1475	-0.0239	0.1180
rs717302	0.0728	0.0034	0.1212	0.0773	-0.0977	0.0293
rs719366	0.0266	0.1081	0.1058	-0.0899	0.1747	0.0660
rs722098	0.0268	0.1457	-0.1419	-0.2332	-0.0815	-0.0546
rs722290	0.0492	0.1792	0.0818	0.0084	-0.3666	-0.0081
rs727811	-0.1118	-0.0155	0.1235	0.1400	0.1537	0.0672
rs729172	0.0129	0.2323	-0.0735	0.0327	0.2117	0.1104
rs733164	0.0068	-0.0729	0.0458	-0.1424	-0.2253	-0.0878
rs735155	0.0205	0.0727	-0.0338	0.1556	0.0182	0.0587
rs737681	0.0947	0.1276	0.0287	0.2972	0.1625	0.1540
rs740598	-0.1398	-0.0059	0.1739	0.1128	-0.0449	0.0238
rs740910	-0.0419	-0.0145	-0.0282	-0.0899	0.1534	-0.0063
rs763869	-0.0858	0.0022	-0.0925	-0.2774	0.1276	-0.0718
rs8037429	-0.0380	0.1745	-0.1145	-0.0459	0.2440	0.0463
rs8078417	-0.1259	0.0948	0.0759	-0.0731	-0.0568	-0.0145
rs826472	-0.0519	0.1456	0.0347	0.1799	0.0160	0.0887
rs873196	-0.0136	0.2719	-0.0052	-0.0442	-0.0779	0.0358
rs876724	0.0713	0.1097	-0.0568	0.1873	-0.2040	0.0180
rs891700	0.1645	0.0673	-0.2451	0.1024	0.1681	0.0518
rs901398	0.0236	-0.0350	-0.0094	0.0848	0.0256	0.0190
rs907100	0.1653	0.0229	0.1330	-0.0098	-0.1186	0.0293
rs914165	-0.1459	-0.1298	-0.1406	-0.0598	0.2848	-0.0339
rs917118	-0.0095	0.0648	0.0475	0.0420	0.1360	0.0592
rs938283	-0.0943	-0.1473	-0.1406	0.0202	0.1314	-0.0362
rs964681	0.0817	0.0146	0.0338	-0.0364	0.1123	0.0406
rs987640	-0.1093	-0.0333	0.1107	0.1337	0.1193	0.0406
rs9905977	-0.1976	-0.0594	0.0076	0.0593	-0.0851	-0.0537
rs993934	-0.0900	0.1066	0.1412	-0.2774	0.0274	-0.0195
rs9951171	-0.1331	-0.0835	0.1499	0.1873	-0.0454	0.0124
All	0.0269	0.0500	0.0280	0.0250	0.0154	0.0292

Table 1.18: Locus-specific $\hat{\beta}_{IS}$ for each subgroup over 94 iiSNP markers using the 1000GP data set.

	Subgroup					
Locus	**AFR**	**AMR**	**EAS**	**EUR**	**SAS**	**All**
rs1005533	-0.1380	0.3045	0.1824	-0.3310	0.0137	0.0063
rs10092491	0.0222	0.0822	0.0988	0.0357	-0.0560	0.0366
rs1015250	0.3081	0.5196	-0.1388	0.2043	0.1609	0.2108
rs1024116	0.2696	-0.2115	0.5208	-0.1845	0.0329	0.0855
rs1028528	0.3771	0.2953	0.0477	0.3587	0.1814	0.2520
rs1031825	0.0172	0.4490	-0.1144	0.1331	-0.0239	0.0922
rs10488710	0.0641	-0.0063	0.1276	-0.1514	0.1905	0.0449
rs10495407	0.5994	-0.1415	-0.1570	0.0883	0.3133	0.1405
rs1058083	0.0172	-0.1531	-0.0574	-0.0296	0.0266	-0.0393
rs10773760	0.1457	-0.0319	0.1633	-0.1437	-0.0453	0.0176
rs10776839	0.1962	-0.1927	0.1004	-0.0760	0.3192	0.0694
rs1109037	0.0170	0.0344	0.2660	-0.2486	0.2319	0.0602
rs1294331	0.1462	0.1332	0.3330	0.4105	0.1511	0.2348
rs12997453	-0.3051	0.0922	0.0044	-0.0461	0.4090	0.0309
rs13182883	0.1609	-0.0666	-0.1364	0.3281	0.0304	0.0633
rs13218440	-0.1779	0.0602	-0.1206	-0.1080	-0.0930	-0.0878
rs1335873	0.7837	-0.0152	0.1511	0.2123	0.4463	0.3156
rs1336071	0.0562	0.1467	-0.0382	-0.0382	0.2450	0.0743
rs1355366	0.3038	0.4141	0.5536	0.3842	0.4744	0.4260
rs1357617	0.4395	0.4300	0.2265	0.0131	-0.0955	0.2027
rs1360288	0.2736	0.1477	-0.1107	-0.1809	-0.1208	0.0018
rs1382387	0.0197	0.3548	0.3176	-0.0520	0.0436	0.1367
rs1413212	0.0025	0.1436	0.0742	0.2248	0.1825	0.1255
rs1454361	0.3254	-0.0918	0.0871	0.2588	-0.1539	0.0851
rs1463729	0.4949	0.0032	0.0717	0.2684	0.0081	0.1693
rs1490413	-0.0706	0.0645	-0.1377	-0.1060	0.0888	-0.0322
rs1493232	0.1476	0.4645	0.2236	0.3510	-0.0807	0.2212
rs1498553	0.0233	0.0405	0.1086	-0.0339	0.0914	0.0460
rs1523537	-0.0840	-0.0334	0.1943	-0.0279	0.1690	0.0436
rs1528460	0.7626	0.2402	0.1273	0.2140	0.0330	0.2754
rs159606	-0.0223	0.0004	0.0158	-0.0341	0.1821	0.0284
rs1736442	0.2899	0.2590	0.2899	0.2772	0.0704	0.2372
rs1821380	0.1048	0.0878	-0.0926	0.0349	0.1330	0.0536
rs1886510	0.5587	-0.0032	0.2784	-0.4647	0.1174	0.0973
rs1979255	0.1106	-0.0710	-0.1134	0.1122	-0.0486	-0.0020
rs2040411	0.3383	-0.0709	0.1588	0.0999	-0.0953	0.0862
rs2046361	0.1618	0.3631	-0.0522	0.2581	0.1431	0.1748
rs2056277	0.1054	-0.0497	0.0957	0.2535	-0.1451	0.0520
rs2076848	0.2799	0.2007	0.0580	-0.3840	0.1166	0.0542
rs2107612	-0.2167	-0.0537	0.5522	-0.0109	-0.2514	0.0039
rs2111980	0.1189	0.1252	0.0059	0.1430	0.0601	0.0906
rs214955	0.0935	0.0694	0.0029	0.1748	0.1660	0.1013
rs221956	0.0929	0.0795	0.0088	-0.0043	0.0194	0.0392
rs2269355	-0.1489	0.1050	0.1456	0.1296	-0.2989	-0.0135
rs2342747	0.0739	-0.0565	-0.0096	0.1413	0.0945	0.0487
rs2399332	0.0479	0.0548	0.2162	-0.0371	0.1797	0.0923
rs251934	-0.0610	0.0243	0.5908	-0.3487	-0.0093	0.0392
rs279844	0.0254	0.1756	0.1375	0.1747	-0.2330	0.0560
rs2830795	0.4663	0.3424	-0.3011	0.1902	0.1460	0.1688

rs2831700	-0.0548	0.0924	0.0801	-0.0669	0.2874	0.0676
rs2920816	0.3459	0.1405	0.0220	-0.0071	0.0953	0.1193
rs321198	0.1285	-0.0764	0.2091	0.0257	-0.0275	0.0519
rs338882	0.0259	-0.0367	0.0571	-0.0999	0.0213	-0.0065
rs354439	0.1975	0.3984	-0.0073	0.1545	-0.1413	0.1204
rs3780962	-0.0932	-0.0689	0.0792	-0.0366	-0.0567	-0.0352
rs430046	0.2010	-0.1931	0.0787	-0.0653	-0.0847	-0.0127
rs4364205	-0.2331	0.1653	-0.0152	-0.1154	0.2032	0.0010
rs445251	0.0719	-0.1161	0.1999	0.1551	-0.2250	0.0172
rs4530059	0.1544	-0.1726	-0.0360	-0.1275	0.3440	0.0325
rs4606077	-0.1848	-0.1182	0.0199	0.0230	-0.1539	-0.0828
rs560681	-0.0625	-0.3654	-0.2636	0.1172	-0.0119	-0.1173
rs576261	0.0870	0.1966	0.1043	-0.1082	-0.0591	0.0441
rs6444724	0.1341	-0.0160	-0.1110	0.3431	-0.0937	0.0513
rs6811238	0.0008	-0.0392	0.2168	0.0352	0.2006	0.0828
rs6955448	-0.1532	-0.0384	0.1070	0.2269	0.0560	0.0397
rs7041158	-0.0151	0.1508	0.2757	0.1529	0.0661	0.1261
rs717302	0.0032	0.0872	0.5953	0.0041	-0.0113	0.1357
rs719366	0.4230	0.0651	0.1832	-0.2506	0.0768	0.0995
rs722098	0.4516	0.1764	-0.0956	0.1014	-0.0178	0.1232
rs722290	0.0469	0.1746	0.0825	0.0015	-0.3434	-0.0076
rs727811	0.2940	0.0164	0.2654	0.1632	0.2587	0.1996
rs729172	0.3670	0.0289	0.6236	-0.2223	0.4476	0.2490
rs733164	-0.0307	-0.2093	0.4037	-0.2257	-0.2480	-0.0620
rs735155	0.2778	0.1013	0.5316	0.2031	0.1597	0.2547
rs737681	0.0319	0.1765	0.4532	0.2656	0.1094	0.2073
rs740598	-0.1424	-0.0337	0.1140	0.0957	0.1258	0.0319
rs740910	0.7239	-0.4174	0.7948	-0.5710	0.2269	0.1514
rs763869	-0.1039	0.0016	-0.0666	-0.2911	0.2641	-0.0392
rs8037429	-0.0017	0.1719	-0.0829	-0.0363	0.2351	0.0572
rs8078417	-0.1877	-0.0135	0.1012	-0.0648	0.1362	-0.0057
rs826472	0.7220	0.2701	0.2366	-0.0482	-0.1515	0.2058
rs873196	0.2618	0.1588	0.4148	-0.3533	-0.0334	0.0897
rs876724	0.4123	0.2164	-0.2389	0.1083	-0.1540	0.0688
rs891700	0.1603	0.0680	-0.2415	0.1047	0.1648	0.0513
rs901398	-0.0510	-0.0482	0.0314	0.0034	0.2013	0.0274
rs907100	0.5099	0.1113	0.1788	0.0874	-0.0337	0.1707
rs914165	0.1556	-0.0950	0.0752	0.0560	0.3601	0.1104
rs917118	0.1361	0.1464	0.2430	0.1722	0.1270	0.1650
rs938283	0.2269	-0.1339	-0.0918	-0.2463	0.0722	-0.0346
rs964681	0.2800	0.0613	0.1437	-0.1329	-0.0241	0.0656
rs987640	-0.1036	-0.0307	0.1161	0.1457	0.3153	0.0886
rs9905977	-0.0422	0.0101	-0.0623	0.1717	-0.1911	-0.0228
rs993934	-0.0175	0.1860	0.1200	-0.2982	0.0010	-0.0017
rs9951171	-0.1512	-0.0806	0.1512	0.2314	-0.0684	0.0165
All	0.1328	0.0696	0.1119	0.0316	0.0652	0.0822

Table 1.19: Locus-specific $\hat{\beta}_{IT}$ for each subgroup over 94 iiSNP markers using the 1000GP data set.

	Subgroup					
Locus	**AFR**	**AMR**	**EAS**	**EUR**	**SAS**	**All**
rs1005533	-0.0486	0.0402	0.2414	-0.0519	0.0478	0.0458
rs10092491	-0.0259	-0.0395	0.1071	-0.0024	-0.0338	0.0011
rs1015250	0.0615	0.2494	-0.0071	0.2771	-0.0004	0.1161
rs1024116	0.2034	-0.1148	0.5697	-0.1015	-0.0114	0.1091
rs1028528	0.4142	0.2949	0.0405	0.2947	0.0449	0.2178
rs1031825	0.0107	0.1472	-0.0108	0.1741	-0.0119	0.0619
rs10488710	-0.0776	-0.0525	0.0073	0.0108	0.1475	0.0071
rs10495407	0.6292	-0.2024	-0.1750	-0.1609	0.4015	0.0985
rs1058083	-0.0307	-0.0275	0.0005	0.0729	0.0025	0.0036
rs10773760	0.0541	-0.0172	0.0698	-0.1033	0.0469	0.0101
rs10776839	-0.0023	-0.0352	-0.0404	0.0306	0.0756	0.0057
rs1109037	-0.0083	-0.0037	-0.0080	-0.0037	0.0709	0.0094
rs1294331	-0.0216	-0.0652	0.1834	0.1585	-0.0280	0.0454
rs12997453	-0.0152	-0.0747	0.0056	-0.0461	0.2319	0.0203
rs13182883	-0.0181	-0.0060	-0.0323	0.1897	0.1789	0.0624
rs13218440	-0.0160	0.0725	-0.0161	-0.0203	-0.0239	-0.0008
rs1335873	0.7197	0.0625	0.0811	0.2637	0.4285	0.3111
rs1336071	-0.0096	-0.0451	0.0237	-0.0319	0.1688	0.0212
rs1355366	-0.0214	0.2493	0.4739	-0.0415	-0.0332	0.1254
rs1357617	0.0822	0.1215	0.0952	-0.1902	-0.0975	0.0022
rs1360288	0.3169	0.1942	-0.0956	-0.0844	-0.1139	0.0435
rs1382387	-0.0186	0.2128	0.1915	-0.0079	0.0071	0.0770
rs1413212	-0.0031	0.1309	-0.0770	0.1542	-0.0289	0.0352
rs1454361	0.1728	-0.0083	0.0374	-0.0297	-0.0275	0.0289
rs1463729	0.3387	0.0090	-0.0164	0.0685	0.0142	0.0828
rs1490413	-0.0079	0.0089	0.0270	0.0082	0.0541	0.0181
rs1493232	0.0037	0.1454	0.1558	0.1864	0.0165	0.1016
rs1498553	-0.0040	-0.0060	-0.0017	-0.0069	0.0127	-0.0012
rs1523537	0.0193	0.0452	0.0464	0.0025	0.1047	0.0436
rs1528460	0.7386	0.1338	0.0466	0.2653	0.0672	0.2503
rs159606	-0.0098	-0.0659	-0.0279	0.0019	0.1607	0.0118
rs1736442	0.0275	0.0228	-0.0520	-0.0471	0.0533	0.0009
rs1821380	0.0226	-0.0435	-0.0206	0.0069	0.0922	0.0115
rs1886510	0.5518	-0.1513	0.3815	-0.2150	0.2162	0.1566
rs1979255	0.1189	-0.0301	-0.0338	0.0307	-0.0225	0.0126
rs2040411	0.4442	0.1203	0.1237	0.1077	-0.0102	0.1571
rs2046361	0.0340	0.1495	-0.0043	0.1291	0.0331	0.0683
rs2056277	0.1340	-0.0652	0.1941	-0.0876	-0.1819	-0.0013
rs2076848	0.2280	0.0485	-0.0197	-0.0654	-0.0307	0.0322
rs2107612	-0.2018	-0.0496	0.5887	-0.0468	-0.0414	0.0498
rs2111980	0.1584	0.0170	0.0651	0.0214	0.0369	0.0598
rs214955	0.0101	-0.0038	-0.0023	0.0133	0.0416	0.0118
rs221956	0.0027	0.0548	-0.0662	0.0054	0.0035	0.0000
rs2269355	0.0185	0.0118	0.0704	0.0594	0.0016	0.0323
rs2342747	-0.0231	-0.0998	0.0354	-0.0549	0.2206	0.0156
rs2399332	-0.0690	0.0117	0.0230	-0.0513	0.1582	0.0145
rs251934	0.0931	-0.1970	0.5648	-0.1633	-0.0102	0.0575
rs279844	0.0127	-0.0011	0.0223	-0.0012	0.0099	0.0085
rs2830795	0.4437	0.3433	-0.1803	-0.0703	-0.0122	0.1048

rs2831700	0.0218	-0.0376	-0.0382	-0.0094	0.2688	0.0411
rs2920816	0.0820	-0.0369	0.0710	-0.0882	0.0527	0.0161
rs321198	-0.0128	0.0060	0.0258	0.0015	-0.0180	0.0005
rs338882	0.0132	0.0315	0.0626	-0.0055	-0.0046	0.0194
rs354439	0.0870	0.2155	0.0154	0.0462	0.0838	0.0896
rs3780962	0.0005	0.0316	0.0176	-0.0060	0.0101	0.0108
rs430046	0.1612	-0.0036	0.0058	-0.0666	-0.0321	0.0130
rs4364205	-0.0299	0.1290	-0.0382	-0.0447	0.0598	0.0152
rs445251	-0.0182	0.0191	0.1300	0.0303	-0.0180	0.0286
rs4530059	0.0068	-0.0955	0.0514	-0.1099	0.3230	0.0352
rs4606077	-0.1304	0.0416	0.1290	0.1046	-0.0344	0.0221
rs560681	0.0476	-0.0152	-0.0544	0.1579	-0.0862	0.0099
rs576261	0.0030	-0.0022	-0.0021	-0.0022	0.0200	0.0033
rs6444724	0.0047	0.0052	-0.0078	0.1213	-0.0088	0.0229
rs6811238	-0.0029	-0.0031	0.1088	0.0220	0.1119	0.0474
rs6955448	-0.0750	-0.0452	-0.0414	0.0677	0.0961	0.0004
rs7041158	-0.1085	0.0263	0.0368	0.0076	0.0877	0.0100
rs717302	-0.0746	0.0841	0.5398	-0.0787	0.0779	0.1097
rs719366	0.4073	-0.0474	0.0872	-0.1484	-0.1167	0.0364
rs722098	0.4366	0.0369	0.0397	0.2699	0.0582	0.1682
rs722290	-0.0021	-0.0045	0.0013	-0.0069	0.0137	0.0003
rs727811	0.3646	0.0313	0.1626	0.0282	0.1253	0.1424
rs729172	0.3588	-0.2629	0.6492	-0.2633	0.3006	0.1565
rs733164	-0.0377	-0.1277	0.3753	-0.0742	-0.0206	0.0230
rs735155	0.2628	0.0313	0.5468	0.0576	0.1442	0.2086
rs737681	-0.0687	0.0568	0.4371	-0.0422	-0.0619	0.0642
rs740598	-0.0030	-0.0276	-0.0712	-0.0184	0.1630	0.0085
rs740910	0.7349	-0.3973	0.8004	-0.4426	0.0880	0.1567
rs763869	-0.0172	-0.0006	0.0231	-0.0131	0.1574	0.0299
rs8037429	0.0347	-0.0020	0.0276	0.0088	-0.0096	0.0119
rs8078417	-0.0556	-0.1189	0.0279	0.0071	0.1822	0.0085
rs826472	0.7356	0.1468	0.2093	-0.2761	-0.1700	0.1291
rs873196	0.2717	-0.1533	0.4178	-0.2965	0.0406	0.0561
rs876724	0.3674	0.1205	-0.1728	-0.0954	0.0397	0.0519
rs891700	-0.0041	0.0012	0.0012	0.0035	-0.0024	-0.0001
rs901398	-0.0762	-0.0130	0.0403	-0.0881	0.1804	0.0087
rs907100	0.4134	0.0906	0.0537	0.0962	0.0749	0.1458
rs914165	0.2626	0.0299	0.1884	0.1088	0.1076	0.1395
rs917118	0.1442	0.0877	0.2055	0.1362	-0.0091	0.1129
rs938283	0.2931	0.0107	0.0419	-0.2718	-0.0668	0.0014
rs964681	0.2163	0.0474	0.1140	-0.0935	-0.1525	0.0263
rs987640	0.0045	0.0023	0.0068	0.0149	0.2234	0.0504
rs9905977	0.1288	0.0652	-0.0704	0.1200	-0.0986	0.0290
rs993934	0.0660	0.0895	-0.0237	-0.0187	-0.0268	0.0173
rs9951171	-0.0168	0.0020	0.0026	0.0559	-0.0224	0.0043
All	0.1090	0.0209	0.0865	0.0070	0.0507	0.0548

Table 1.20: Locus-specific $\hat{\beta}_{WT}$ for each subgroup over 94 iiSNP markers using the 1000GP data set.

Chapter 2

The impact of DNA sequence data on match probabilities for different forensic marker systems

This chapter is based on Sanne E. Aalbers, Scott R. Kennedy, Bruce S. Weir. *The impact of sequence data on match probabilities for different forensic marker systems.* (2023). Manuscript in preparation.

Abstract

In this paper we demonstrate the effect of sequence data on match probabilities, a measure integral to DNA evidence evaluations. Results show that empirical matching proportions become less conservative the more markers we include and that this problem is exacerbated with sequence-based data compared to length-based data. While a theta-correction can be invoked to compensate for multi-locus dependencies, we caution against the combination of markers across different systems due to the occurrence of dependencies even for unlinked loci.

Introduction

Forensic genetics is concerned with the matching of DNA profiles from evidence and from persons of interest. The power of DNA profiling comes from using many loci and requires the estimation of multi-locus match probabilities. For autosomal profiles, a convenient approach is to employ the product rule, which combines single-locus match probabilities by multiplying over loci. The product rule assumes that loci are independent. If this assumption is violated,

match probabilities may be underestimated, potentially leading to overstating the weight of the evidence.

Autosomal short tandem repeat (aSTR) markers are generally assumed to be independent of each other, due to their physical distance. However, between-locus dependencies occur even for unlinked loci and it has previously been shown that match probabilities for multi-locus autosomal STR profiles depart from products of single-locus probabilities [77, 46]. These departures increase with the number of loci: the more loci match, the more likely that a match occurs at the next locus as well, as also noted by Donnelly [20].

In contrast to autosomal STR markers, it is generally assumed that there is a lack of independence between markers for the Y-chromosome [4]. As such, the product rule is not employed when estimating matching proportions for multi-locus Y-STR profiles. Instead, Y-STR profiles are considered as a whole and multi-locus haplotype frequencies are used.

Whatever marker system we consider, DNA profiles are not independent since human populations are finite and any two individuals will therefore show some level of relatedness. With the introduction of sequencing techniques it would be of interest to quantify the extent of departure from independence for data generated by such technologies. In this paper we investigate the extent of dependencies between markers for sequence data over different forensic marker systems.

Methods

Data

Two data sets have been used in this study, both containing individual-level genotype data across different marker systems that have been described previously. The first data set consists of 350 individuals of the 1000 Genomes Project (1000GP) Phase 3 (`http://www.1000genomes.org`) typed for 27 autosomal STR markers, 24 Y-STR markers, and 94 identity-informative single nucleotide polymorphisms (iiSNPs) [2]. The second data set consists of the NIST 1036 data set with calls for the same set of 27 autosomal STR markers and 24 Y-STR markers [33, 63]. Both data sets were generated on a MiSeq FGx instrument with the

ForenSeq DNA Signature Prep Kit and calls were restricted to the regions as reported in the sample level reports of Illumina's Universal Analysis Software. For Y-STR data we restricted to the set of 1104 male samples showing complete profiles only.

Calculation of match probabilities

For this study we are interested in comparing observed and expected match probabilities. In Appendix A we show that expected match probabilities for autosomal data can be calculated from the following expression:

$$P_2(\theta) = \frac{1}{D}[6\theta^3 + \theta^2(1-\theta)(2+9S_2) + 2\theta(1-\theta)(2S_2+S_3) + (1-\theta^3)(2S_2^2 - S_4)],$$

where $D = (1+\theta)(1+2\theta)$ and $S_n = \sum_u \pi_u^n$. For any value of the population structure parameter θ we can calculate the expected matching proportion in a database using sample allele proportions $\tilde{p}_u$ estimated from the data to approximate population allele frequencies π_u. Multi-locus match probabilities can be estimated by taking the product of the single locus expected values.

The observed match probabilities are calculated as the sum of the number of matches for a set of markers divided by the total pairs of profiles considered. In other words, if a group of n individuals were sorted into groups of size n_g with matching genotype or haplotype, the observed match probability is $\sum_g n_g(n_g-1)/[n(n-1)]$. Genotype or haplotype counts are considered for both length-based and sequence-based STR data.

Entropy measures

An alternative approach to describe associations between markers is with the concept of entropy [14, 61]. Siegert et al. suggested the use of entropy measures to construct a set of Y-STR markers by an iterative procedure [61]. For a locus with sample frequencies $\tilde{p}_u$ for alleles A_u, the entropy is calculated as:

$$H_A = -\sum_u \tilde{p}_u \ln(\tilde{p}_u)$$

The first locus A is selected based on the largest entropy and the conditional entropy $H_{B|A} =$

$H_{AB} - H_A$ is calculated for all remaining markers B. The locus with the highest conditional entropy is added to the set and conditional entropies are updated for the remaining markers given the current selection of markers. This process can be repeated until the conditional entropy reaches zero for all markers, indicating that no additional discriminating power has been observed for the remaining markers.

Entropy measures are additive under the assumption of independence between markers. If haplotypes $A_u B_v$ have sample frequencies $\tilde{P}_{uv}$, the two-locus entropy is

$$H_{AB} = -\sum_u \sum_v \tilde{P}_{uv} \ln(\tilde{P}_{uv}) = -\sum_u \sum_v \tilde{p}_u \tilde{p}_v [\ln(\tilde{p}_u) + \ln(\tilde{p}_v)] = H_A + H_B$$

This implies that if $H_{AB} \neq H_A + H_B$ there is evidence of dependencies between markers in the sense that the multi-locus probabilities $\tilde{P}_{uv}$ are not equal to the product of the single-locus probabilities.

We employ this approach for both Y-STR data as well as genotypic-based autosomal markers.

Results

Autosomal STRs

For the autosomal STR (aSTR) markers we have a total of $1386 \cdot 1385/2 = 959\,805$ pairs of profiles, and most pairs show a match at only one locus. Table 2.1 and Figure 2.1 show the number of m-locus matches within our data set. For length-based genotypes, one pair of individuals matches at as many as 12 loci. For sequence-based data, this pair now shows a match for a total of 11 different markers, together with another pair that showed an 11-locus match for both data types. The average number of matching loci is 1.9 for length-based data and goes down to 1.5 matching loci for sequence data.

Under the assumption of independence across markers we expect to see an approximately constant decline in the number of matches with increasing number of loci. Figure 2.1 shows the decline in matching proportions on a log-scale, which more closely resembles an exponential decline and demonstrates that autosomal match probabilities for STR profiles violate the

assumption of independence.

Loci	0	1	2	3	4	5	6	7	8	9	10	11	12
LB	131 705	273 136	269 971	171 028	77 332	27 087	7453	1731	308	47	5	1	1
SB	211 779	329 961	246 895	117 802	40 148	10 565	2174	421	50	7	1	2	0

Table 2.1: Observed number of matching loci out of 27 aSTRs for $n = 1386$ individuals using length-based (LB) and sequence-based (SB) data.

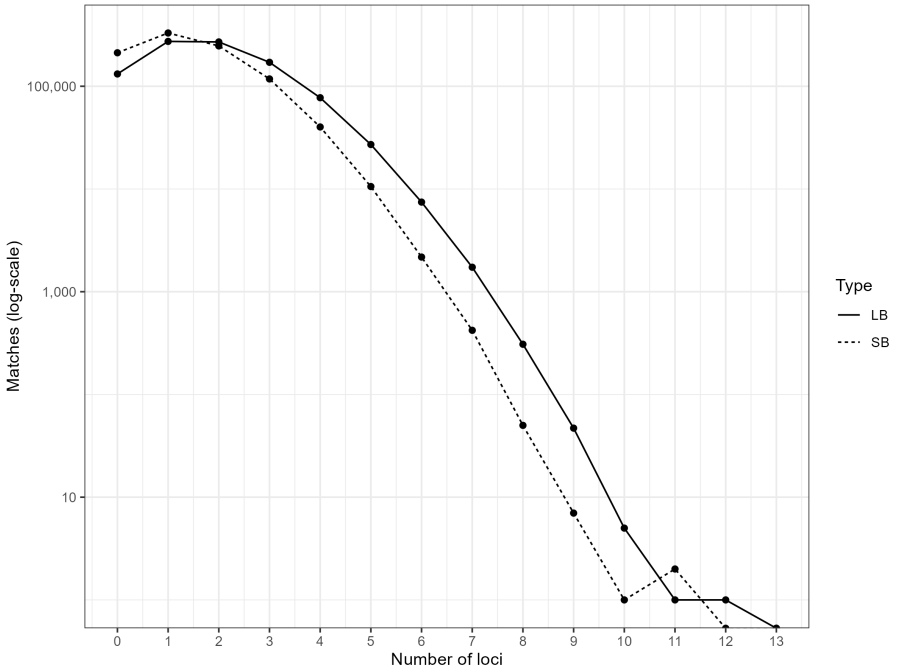

Figure 2.1: Observed number of matching loci out of 27 for pairwise comparisons of $n = 1386$ individuals for length-based (LB) and sequence-based (SB) autosomal STR data.

Table 2.2 shows single-locus match probabilities for both length-based and sequence based allele callings assuming no population structure. As expected, matching proportions are smaller for sequence data as compared to length-based data. The decrease is more dramatic for loci that show more sequence variation, with the most polymorphic markers showing a decrease in matching proportions of more than 70% (76.0% decrease for D21S11 and 72.4% decrease for D12S391). Instances where the matching proportions get underestimated are highlighted in red. Results show that sequence data more often lead to an underestimation of the observed matching proportion as compared to length-based data (13 out of 27 loci, compared to 11 out of 27 loci, respectively). Locus D5S818 is the only marker that shows an

underestimation for length-based data, but not for sequence data.

	Length-based				Sequence-based			
Locus	**Alleles**	**Obs**	**Exp**	**Entropy**	**Alleles**	**Obs**	**Exp**	**Entropy**
D21S11	27	0.0402	0.0416	3.6154	116	0.0096	0.0102	5.1469
D12S391	24	0.0280	0.0273	3.9666	95	0.0077	0.0073	5.5892
D2S1338	13	0.0210	0.0213	3.9906	71	0.0087	0.0085	5.2406
vWA	12	0.0590	0.0604	3.0269	40	0.0336	0.0343	3.9014
D3S1358	11	0.0938	0.0937	2.6100	35	0.0245	0.0233	4.0953
D8S1179	11	0.0550	0.0551	3.1861	35	0.0174	0.0175	4.4430
D13S317	8	0.0733	0.0716	2.9167	31	0.0242	0.0236	4.0189
D1S1656	18	0.0212	0.0213	4.0977	38	0.0124	0.0127	4.7512
FGA	28	0.0291	0.0293	3.8352	43	0.0282	0.0284	3.9382
D18S51	22	0.0255	0.0261	3.8882	34	0.0247	0.0252	3.9699
D6S1043	29	0.0308	0.0309	3.8698	41	0.0303	0.0304	3.9445
D9S1122	11	0.1411	0.1390	2.3205	23	0.0574	0.0556	3.3452
D5S818	9	0.1084	0.1076	2.6050	20	0.0519	0.0523	3.5170
D19S433	16	0.0539	0.0541	3.4337	26	0.0525	0.0527	3.5039
D2S441	15	0.0836	0.0842	2.8329	25	0.0575	0.0570	3.3557
PentaE	25	0.0134	0.0132	4.5549	33	0.0131	0.0129	4.5949
CSF1PO	9	0.1039	0.1049	2.5944	15	0.1030	0.1035	2.6269
D16S539	9	0.0721	0.0721	2.8593	14	0.0716	0.0716	2.8807
D7S820	12	0.0698	0.0695	2.8892	16	0.0694	0.0690	2.9055
D17S1301	10	0.1575	0.1585	2.1498	13	0.1562	0.1572	2.1745
D4S2408	7	0.0887	0.0938	2.5388	10	0.0697	0.0730	2.8155
D10S1248	12	0.0825	0.0823	2.7815	14	0.0824	0.0821	2.7882
D22S1045	11	0.0931	0.0898	2.7847	13	0.0929	0.0895	2.7931
D20S482	11	0.1351	0.1352	2.4012	12	0.1350	0.1350	2.4054
PentaD	16	0.0363	0.0358	3.6439	17	0.0361	0.0356	3.6548
TH01	8	0.0742	0.0769	2.7186	9	0.0740	0.0767	2.7264
TPOX	9	0.1373	0.1356	2.3799	10	0.1372	0.1354	2.3839

Table 2.2: Observed (Obs) and expected (Exp) match probabilities assuming no population structure and single-locus entropy per aSTR locus ordered by increase in the number of alleles comparing length-based (LB) and sequence-based (SB) data. Expected values that underestimate observations are highlighted in red.

Entropy measures can give us an idea which markers lead to the most discriminating DNA profiles and it can be seen from Tables 2.2 and 2.3 that this set does not necessarily include the loci with the most observed number of alleles. For length-based data, locus PentaE shows the highest single-locus entropy value of 4.5549. The 2-locus entropy conditional on this marker is highest for locus D2S1338, with a combined entropy of 6.9979. Note that the conditional entropy 2.4429 is smaller than the single-locus entropy 3.9906 of the selected marker D2S1338, giving further evidence of dependencies across markers. After selecting a set of five markers, consisting of PentaE, D2S1338, D12S391, D4S2408, and TH01, the

conditional entropy reaches zero for all remaining loci and a maximum combined entropy of 7.2174 is reached. In other words, all profiles for the 1386 individuals in our data set occur only once with just this set of loci. For sequence-based data the total combined entropy is 7.2218, which is reached with a selection of only three markers. Locus D12S391 shows the highest single-locus entropy in this case with a value of 5.5892. Combining this locus with markers D2S1338 and D13S317 gives a selection that leads to unique multi-locus profiles within our data set. Locus D21S11 with the most observed number of sequence-based alleles appears in neither of the sets. Match probabilities are a better indicator of loci being selected based on entropy measures as they account for allele frequencies. Locus D21S11, for example, shows most of the increase in the number of alleles in the form of singletons (44.8%), leading to relatively higher matching proportions and lower single-locus entropy compared to markers with more evenly distributed alleles among individuals. In that sense, markers D12S391 and D2S1338 are more valuable with a proportion of 23.2% and 21.1% of singletons, respectively. These markers show smaller match probabilities and relatively higher single-locus entropies as compared to locus D21S11.

	Entropy		
Locus	**Single**	**Combined**	**Conditional**
Length-based			
PentaE	4.5549	4.5549	4.5549
D2S1338	3.9906	6.9978	2.4429
D12S391	3.9666	7.2113	0.2136
D4S2408	2.5388	7.2164	0.0051
TH01	2.7186	7.2174	0.0010
D1S1656	4.0977	7.2174	0.0000
Sequence-based			
D12S391	5.5892	5.5892	5.5892
D2S1338	5.2406	7.1651	1.5759
D13S317	4.0189	7.2218	0.0567
D21S11	5.1469	7.2218	0.0000

Table 2.3: Set of markers with non-zero conditional entropy for length-based and sequence-based aSTR data.

Figure 2.2 displays multi-locus matching proportions for sequence-based data for all 2, 3, 4, and 5-locus combinations obtained by invoking the product rule. The assumption of independence works reasonably well for 2 and 3-locus matches (top), with the linear regression

fitted line closely following the identity line. Results become less conservative the more loci we include (bottom) and the slope of the linear regression fit decreases from 0.95 for 3-locus matches to 0.87 and 0.64 for 4-locus and 5-locus matches, respectively. The violation of the independence assumption is exacerbated for sequence-based data as compared to length-based data, with the latter showing a decrease in slope from 0.97, to 0.91, to 0.74 for 3, 4, and 5-locus matching, respectively.

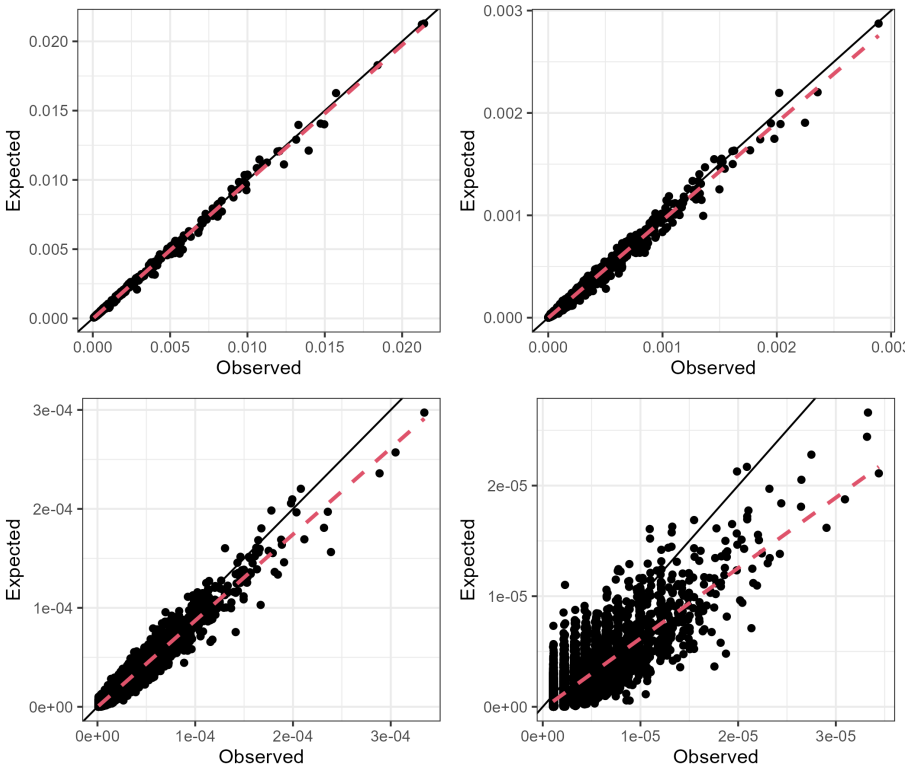

Figure 2.2: Products of 1-locus matching proportions vs. observed 2, 3, 4, and 5-locus matching proportions (from left to right, top to bottom) for sequence-based aSTR data. Solid black lines indicate the identity line and dashed red lines are a linear regression fit to the data.

To account for multi-locus dependencies, we calculated single-locus expectations for theta values in the range of $\theta \in \{0.005, 0.01, 0.03, 0.05\}$. Supplementary Table 2.8 shows the updated values for both length-based and sequence-based data for all 27 autosomal markers. Using these values, we can construct expected multi-locus match probabilities by multipying the theta-corrected single-locus expectations over a set of markers. Figure 2.3 displays 5-locus matching proportions for both length-based and sequence-based data using the different theta corrections to compensate for multi-locus dependencies. Since our matching proportions

generally decrease for sequence data as compared to length-based data, it can be seen that the data points for the former are more concentrated in the lower left corner of the plot. The proportion of conservative results (i.e., when expected values are greater than observed values) increases from 28.6% using no theta correction to 87.5% when using a theta value of $\theta = 0.03$ for length-based data. For sequence-based data these proportions are 12.3% and 64.7%, respectively. Using a theta value of $\theta = 0.05$ increases the proportions to over 90%. Table 2.4 shows the proportion of conservative estimates for up to 5-locus matches for both data types. It can again be seen that sequence-based results are less conservative than length-based results.

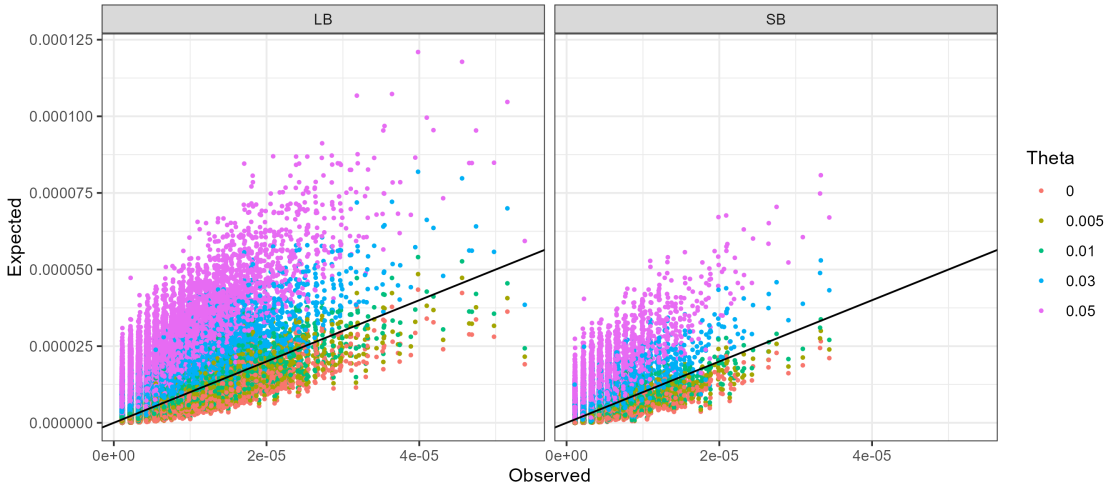

Figure 2.3: Matching proportions for 5-locus matches for different theta values for length-based (left) and sequence-based (right) aSTR data.

Loci	**Type**	$\theta = 0$	$\theta = 0.005$	$\theta = 0.01$	$\theta = 0.03$	$\theta = 0.05$
1	LB	0.593	0.963	1.000		
1	SB	0.519	0.963	1.000		
2	LB	0.499	0.946	0.989	1.000	
2	SB	0.359	0.903	0.974	1.000	
3	LB	0.440	0.857	0.979	1.000	
3	SB	0.349	0.734	0.930	1.000	
4	LB	0.443	0.666	0.842	0.998	
4	SB	0.327	0.506	0.682	0.977	
5	LB	0.286	0.390	0.505	0.875	0.984
5	SB	0.123	0.188	0.268	0.647	0.902

Table 2.4: Proportion of conservative predictions for up to 5-locus matches with different theta values for length-based (LB) and sequence-based (SB) aSTR data.

Identity-informative SNPs

For the identity-informative SNPs (iiSNPs) we resort to the 350 1000GP samples to compare $350 \cdot 349/2 = 61\,075$ pairs of individuals. All pairs match at a minimum of 16 iiSNPs out of the 94 total markers. Two pairs of individuals have as many as 58 matching loci. The average number of matching loci is 36.5 for this data set.

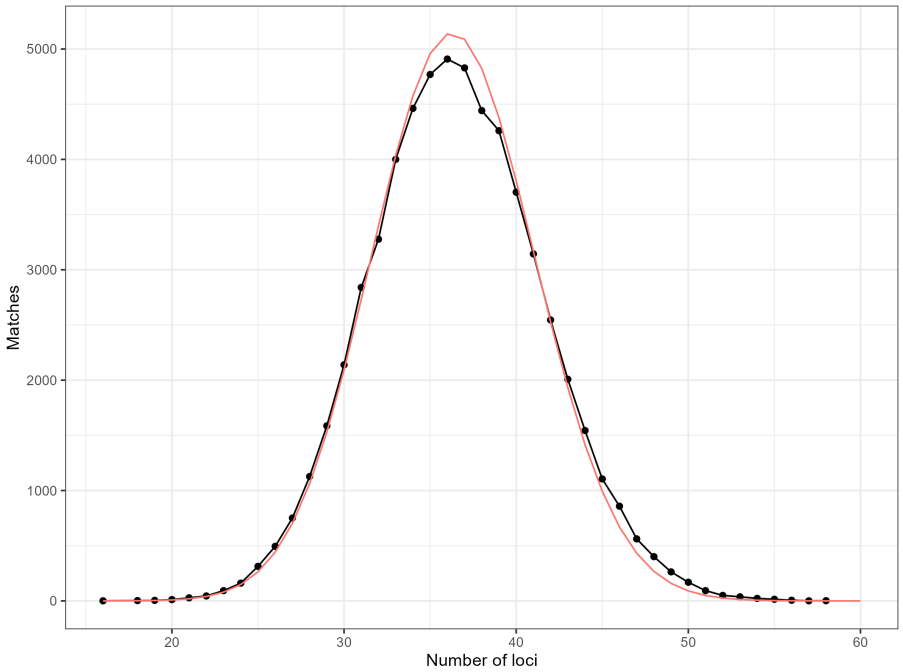

Figure 2.4: Observed and expected number of matching loci out of 94 identity-informative SNPs for $n = 350$ individuals, with expected values under a binomial distribution with median single locus match probability in red.

Figure 2.4 shows the observed number of matching loci combined over pairs. We tried to approximate the observations with a binomial distribution by assuming independence over markers and a constant match probability rate. The expected distribution of the number of matching iiSNPs per pair of individuals was obtained by estimating the probability of a match as the median single locus match probability over all 94 iiSNPs and seems to fit the data well.

Single-locus match probabilities for all iiSNPs are displayed in Supplementary Table 2.9. Assuming no population structure, a total of 16 markers underestimate the observed matching proportion. While a theta correction of 0.01 yielded conservative results for all autosomal

STR markers on a single-locus basis, for iiSNPs this required increasing the theta value to 0.10. This is in line with earlier studies showing higher theta values for SNPs due to the lower mutation rates as compared to STRs [78].

In contrast to what Figure 2.4 may suggest, we do not expect complete independence between iiSNP markers. It is difficult to examine to what extent multi-locus match probabilities depart from products of single-locus probabilities. One reason for this is the number of possible combinations. For the 27 autosomal STR markers, there are $\binom{27}{5} = 80\,730$ combinations to consider when looking at 5-locus matches. For the iiSNPs, the number of combinations increases to almost 55 million when choosing from a set of 94 markers, leading to computational difficulties. The second problem relates to limitation of the data set, as we simply do not expect to see any matches beyond a certain number of loci with just a few hundred individuals.

One workaround to deal with the number of combinations is to take samples of possible combinations for a given number of loci. For the iiSNPs, we've looked at $N = 10\,000$ random 10-, 11-, and 12-locus combinations and calculated observed and expected match probabilities within these samples. Figure 2.5 displays the results for 10-locus iiSNP combinations assuming no population structure ($\theta = 0$, in red) and when using a theta correction of $\theta = 0.10$. We observe a similar trend as seen with the autosomal STR markers where we see an underestimation of match probabilities when employing the product rule and we can compensate for multi-locus dependencies by using an appropriate theta value.

Table 2.5 shows the proportion of conservative results for different theta values and we see again that results become less conservative with increasing number of loci. Going beyond 12-locus matches proves difficult as we do not expect to see any matches when considering only a subset of the total number of possible combinations. While every pair of individuals in this data set matches at least at 16 iiSNPs, for the 12-locus matches 44% of the $N = 10\,000$ random combinations resulted in a selection of markers with no observed matches. Since expected values are always positive, results would still be conservative in case of no observed matches. Larger data sets would be required to obtain more meaningful results.

An alternative option is to look at entropy measures for iiSNPs. Table 2.6 shows the

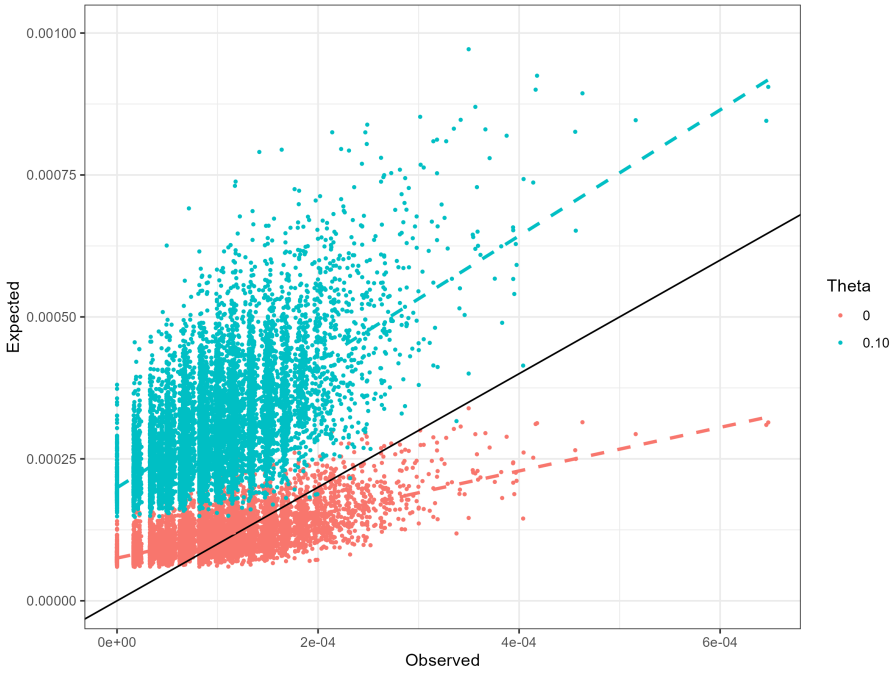

Figure 2.5: Observed and expected matching proportions for a random sample of $N = 10\,000$ 10-locus iiSNP combinations under the product rule ($\theta = 0$) and when using a theta correction of $\theta = 0.10$. The solid black line indicates the identity line and dashed lines are a linear regression fit to the data.

Loci	$\theta = 0$	$\theta = 0.01$	$\theta = 0.03$	$\theta = 0.05$	$\theta = 0.10$
10	0.685	0.759	0.885	0.964	1.000
11	0.639	0.700	0.812	0.904	0.995
12	0.590	0.645	0.758	0.850	0.978

Table 2.5: Proportion of conservative predictions for different theta values based on random samples of $N = 10\,000$ 10-, 11-, and 12-locus combinations selected from 94 iiSNPs.

set of markers with non-zero conditional entropy. The nine listed iiSNP markers result in a selection after which no additional discriminatory power is seen for our limited data set. We also note that single-locus entropy values are much smaller compared to STR markers, due to the limited value of bi-allelic markers as compared to multi-allelic markers. We observe further evidence of dependencies between markers since combined entropies are smaller than the sum of single-locus entropies.

Y-chromosome STRs

We now briefly turn to Y-STR data to assess the impact of sequence variation on this set of markers. For the Y-chromosome STR markers we consider a total of 1104 individuals and

	Entropy		
Locus	**Single**	**Combined**	**Conditional**
rs1335873	1.0965	1.0965	1.0965
rs1493232	1.0848	2.1711	1.0746
rs576261	1.0543	3.2134	1.0424
rs1015250	1.0765	4.2250	1.0115
rs214955	1.0663	5.0261	0.8011
rs917118	1.0684	5.5363	0.5102
rs1463729	1.0549	5.7482	0.2119
rs354439	1.0741	5.8186	0.0704
rs2107612	0.8886	5.8348	0.0162
rs1528460	1.0959	5.8348	0.0000

Table 2.6: Set of markers with non-zero conditional entropy for iiSNP data.

most pairs of profiles show a match at five different loci. For length-based data, the average number of matching loci is 6.81, and this decreases to 6.44 for sequence-based data. Figure 2.6 shows the number of m-locus matches out of 24 Y-STR markers within our data set. Two different pairs of individuals share the same Y-STR haplotype such that we observe a total of 1102 unique profiles in our data set.

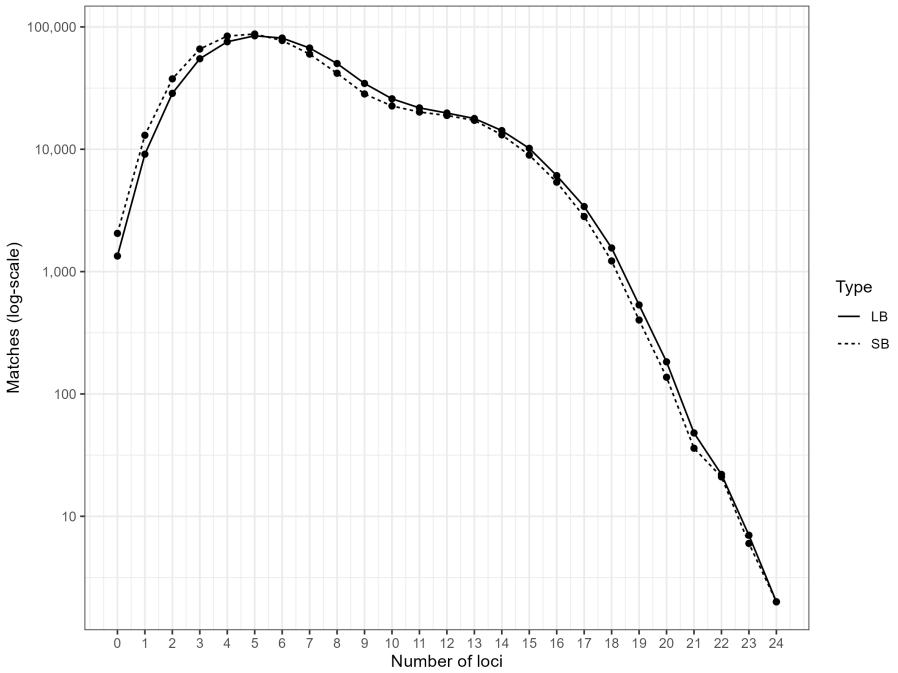

Figure 2.6: Observed number of matching loci out of 24 for pairwise comparisons of $n = 1104$ individuals for length-based (LB) and sequence-based (SB) Y-STR data.

We observe again a decline in information provided by adding additional markers after see-

ing a subset of markers due to the lack of independence between different loci. Supplementary Table 2.10 shows the entropy values for both length-based and sequence-based Y-STR data ordered by conditional entropy. For length-based data it can be seen that we do not observe additional discriminating power beyond a set of 13 loci. For sequence data, the conditional entropy reaches zero after adding 12 loci for our data set. Both sets show a similar set of markers and include the multi-copy markers DYS385 and DYF387, with the latter showing the greatest increase in sequence variation and single-locus entropy. The length-based set of non-zero conditional entropy markers includes markers DYS391 and DYS481, which are not included in the sequence-based set. The sequence-based set in turn includes marker DYS390, showing a substantial increase in sequence variation.

Match probabilities across systems

Besides characterizing the extent of independence of match probabilities within a system, we can also assess this property across systems. Figure 2.7 shows the number of matching aSTR and iiSNP markers for each pair of individuals in the 1000GP data set. Results are shown for both length-based (in red) and sequence-based (in blue) aSTR data and we observe an expected decrease in number of matching loci when introducing sequence variation. While the number of matching aSTR versus iiSNP markers varies substantially among individuals, we do observe a small but significant positive slope value of 0.053 ($p < 0.01$) and 0.088 ($p < 0.001$) for a linear regression fit to length-based and sequence-based data, respectively.

We can also compare multi-locus observed number of matches within the data set to the expected number based on the assumption of independence between systems. To account for dependencies between systems we can invoke a theta correction when calculating expected joint match probabilities. Table 2.7 shows results for both length-based and sequence-based data when calculating match probabilities for several combinations of aSTR and iiSNP markers based on $N = 10\,000$ random multi-locus combinations. To obtain conservative estimates for multi-locus matches across different marker systems, it can be seen that it is necessary to invoke a theta correction for each of the marker sets, and that these sets may come with different theta values.

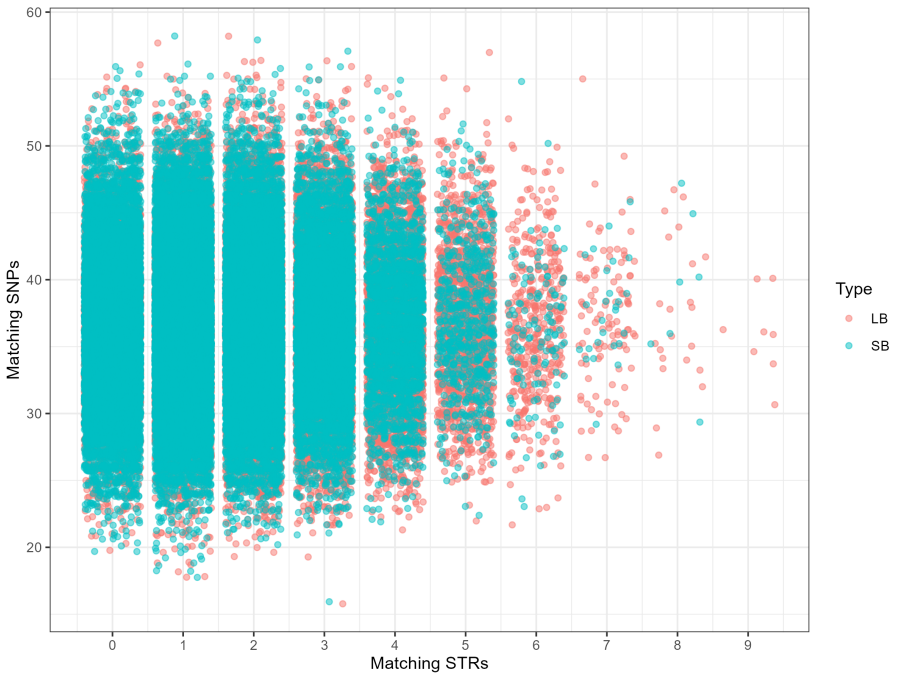

Figure 2.7: Number of matching aSTRs (out of 27) and iiSNPs (out of 94) for pairs of 350 individuals for length-based (LB) and sequence-based (SB) data.

Discussion

With the introduction of sequencing technologies, previously unknown alleles and more overall variability has been found for mainly complex and compound STRs [33]. A substantial increase in the number of alleles does not, however, automatically mean that there is also a great gain in discrimination via sequencing. Improvements will be minimal when a locus is already highly polymorphic by length, as also noted by Gettings et al. [29], or when the added variation occurs mainly in the form of singletons.

In general, match probabilities will decrease for sequence data when it leads to an increase in the number of observed alleles within one locus. For autosomal STRs, our work shows that multi-locus match probabilities may be less conservative for sequence data as compared to length-based data. We recognize that our observed values are impacted by the existence of sampling variation, and results may differ with the use of larger data sets. Nevertheless, we see similar effects of sequence data on match probabilities as observed in literature [46, 77], highlighting the need to account for multi-locus dependencies. We can compensate for multi-

# Markers			$\theta_{aSTR} = 0$,	$\theta_{aSTR} = 0.03$,	$\theta_{aSTR} = 0.03$,
STR	**SNP**	$\theta = 0$	$\theta_{iiSNP} = 0.10$	$\theta_{iiSNP} = 0$	$\theta_{iiSNP} = 0.10$
Length-based					
1	1	0.791	0.991	1.000	1.000
1	2	0.810	0.999	0.998	1.000
1	10	0.734	0.873	0.740	0.909
2	2	0.626	0.905	0.998	1.000
2	3	0.620	0.893	0.967	0.997
2	4	0.609	0.880	0.902	0.988
2	5	0.599	0.853	0.835	0.974
3	3	0.628	0.712	0.834	0.914
3	4	0.731	0.783	0.815	0.883
Sequence-based					
1	1	0.774	0.950	1.000	1.000
1	2	0.773	0.982	0.998	1.000
1	10	0.788	0.882	0.792	0.907
2	2	0.555	0.804	0.989	0.999
2	3	0.590	0.834	0.948	0.992
2	4	0.579	0.794	0.874	0.966
2	5	0.624	0.809	0.825	0.952
3	3	0.744	0.780	0.842	0.894
3	4	0.848	0.868	0.878	0.908

Table 2.7: Proportion of conservative predictions for joint match probabilities of aSTR and iiSNP data using both length-based and sequence-based data with different theta values based on $N = 10\,000$ random multi-locus combinations.

locus dependencies using a theta correction as outlined in Recommendation 4.10 from the 1996 NRC report [52]. It is common in forensic evidence evaluations to use a theta value of around 0.03 for autosomal STR markers [62, 10] and it has been shown that sequence data has similar effects on theta estimates as length-based data [2]. However, there exists a trade-off between obtaining conservative match probabilities and providing accurate estimates of the observed matching proportions. In practice, the selected theta value may depend on the forensic application. The application to a criminal trial may put more importance on being conservative in order to avoid prejudice to the defense, which comes with the need for higher theta values, while the use of lead-generating techniques may prefer optimizing accuracy.

We also showed that the existence of dependencies between markers within a system means that there is a diminishing return on adding more loci beyond a certain point and that this point may be reached sooner for sequence data. For Y-STR data, the set of non-zero conditional entropy markers consisted of 13 length-based and 12 sequence-based markers.

For autosomal STR markers, this point was reached after adding five length-based and three sequence-based markers. While highly variable loci are often selected to obtain maximally informative profiles, such markers may not necessarily be included. The data type used has an effect on the selection and we saw that the set of markers for length-based and sequence-based data may show non-overlapping loci. The order in which markers are added to the set may vary as well. These observations are in line with work from Hall, who showed entropy results for length-based Y-STR data and found variations between different data sets as well as for different regional ethnicity groups [37].

One of the advantages of sequencing technologies is the capacity to produce data on a combination of different marker systems, often including SNPs alongside STR markers. The ForenSeq DNA Signature Prep Kit reports a set of 94 identity-informative SNPs, chosen based on their low allele frequency variation among populations while remaining informative [56, 44]. It is currently unclear how the forensic community is planning to incorporate such markers but we caution against the practice of adding more markers for the sole purpose of increasing the discriminatory power while assuming independence between different marker systems. Even for bi-allelic markers, assuming independence for the iiSNPs will quickly yield extremely small expectations with increasing number of markers. Combining over all 94 loci, the expected match probability assuming no population structure is of the order 10^{-38}. This is in line with the reported numbers in Davenport et al. [17], who highlight the possibility of decreasing the match probabilities even more by including additional flanking region variation. The notion of independence or population structure is not mentioned by Davenport et al., while our results show that employing a product rule for iiSNPs may cause concern and are concordant with earlier studies investigating the independence assumption for SNP data [46].

Other studies have investigated the consequences of combining different marker systems on the weight of the evidence and showed that likelihood ratios may be overestimated if linkage is ignored [72, 85]. While the effect of linkage between STR and iiSNP pairs was reported to be minimal on average [72], our results demonstrate the occurrence of multi-locus dependencies even for unlinked loci. Some work has been done on estimating adjusted theta values for the combination of autosomal and Y chromosome match probabilities [74, 11], but

it remains to be seen if such joint theta measures will be adequate for data generated by sequencing methods.

DNA profiles show dependencies, even for unrelated individuals, and we cannot assume independence between markers within systems nor between systems. The extent of such dependencies differ between marker system, due to differences in mutation rates and inheritance patterns. As such, each system comes with their own theta value and we can incorporate these into our calculations to obtain multi-locus match probabilities. Nevertheless, it is difficult to assess how large such values should be to avoid overestimating the weight of the evidence.

Appendix A: Expected match probabilities

For autosomal data, a pair of individuals can share either 0, 1, or 2 pairs of alleles identical by state (ibs) at each locus. Individuals are said to "match" when they share 2 alleles ibs and we can calculate the expected probability by summing over the joint genotypic probabilities for homozygous and heterozygous types. The sampling formula as derived by Balding and Nichols allows us to assess such probabilities [5]. If n alleles have been sampled from a population and n_u are of type A_u, the probability of observing another allele of the same type is

$$\Pr(A_u | n_u \text{ of } A_u \text{ in } n) = \frac{n_u\theta + (1-\theta)\pi_u}{1 + (n-1)\theta},$$

where π_u is the probability an allele is of type A_u and θ is the probability of identity by descent (ibd) for two alleles drawn randomly from a population. With this formula we can construct the probability of observing a homozygous profile $\Pr(A_uA_u) = \pi_u^2 + \theta\pi_u(1-\pi_u)$ and a heterozygous profile $\Pr(A_uA_v) = 2\pi_u\pi_v(1-\theta)$. More importantly, it directly leads to the match probabilities as endorsed by the 1996 NRC report [52], which are now widely used in forensic applications:

$$\Pr(A_uA_u | A_uA_u) = \frac{[3\theta + (1-\theta)\pi_u][2\theta + (1-\theta)\pi_u]}{(1+\theta)(1+2\theta)}$$
$$\Pr(A_uA_v | A_uA_v) = \frac{2[\theta + (1-\theta)\pi_u][\theta + (1-\theta)\pi_v]}{(1+\theta)(1+2\theta)}$$

Combining results gives us the following joint genotypic probabilities:

$$\begin{aligned}
\Pr(A_uA_u, A_uA_u) &= \Pr(A_uA_u|A_uA_u)\Pr(A_uA_u) \\
&= \frac{\pi_u[\theta + (1-\theta)\pi_u][3\theta + (1-\theta)\pi_u][2\theta + (1-\theta)\pi_u]}{(1+\theta)(1+2\theta)} \\
\Pr(A_uA_v, A_uA_v) &= \Pr(A_uA_u|A_uA_u)\Pr(A_uA_v) \\
&= \frac{4\pi_u\pi_v(1-\theta)[\theta + (1-\theta)\pi_u][\theta + (1-\theta)\pi_v]}{(1+\theta)(1+2\theta)}
\end{aligned}$$

By summing over all possible alleles at a locus we get the probability $P_2(\theta)$ that two individuals share both alleles:

$$\begin{aligned}
P_2(\theta) &= \sum_u \Pr(A_uA_u, A_uA_u) + \frac{1}{2}\sum_u\sum_{v\neq u}\Pr(A_uA_v, A_uA_v) \\
&= \frac{1}{D}[6\theta^3 + \theta^2(1-\theta)(2+9S_2) + 2\theta(1-\theta)(2S_2+S_3) + (1-\theta^3)(2S_2^2 - S_4)],
\end{aligned}$$

where $D = (1+\theta)(1+2\theta)$ and $S_k = \sum_u \pi_u^k$.

Similar expressions can be obtained for the probability of a mismatch or partial match. Since the analysis on a single-locus basis is not applicable to Y-STR markers due to the absence of recombination, match probabilities for Y-STR data are based on haplotype proportions. The sampling formula can still be applied and the probability of two individuals sharing a Y-STR haplotype is $\sum_u \Pr(A_u, A_u) = \sum_u \pi_u^2(1-\theta) + \theta$, where θ is the probability of ibd for two haplotypes drawn randomly from a population and π_u is the probability that a haplotype is of type A_u.

Appendix B: Supplementary tables

	Length-based						Sequence-based					
		Expected						**Expected**				
Locus	**Observed**	$\theta = 0$	$\theta = 0.005$	$\theta = 0.01$	$\theta = 0.03$	$\theta = 0.05$	**Observed**	$\theta = 0$	$\theta = 0.005$	$\theta = 0.01$	$\theta = 0.03$	$\theta = 0.05$
CSF1PO	0.1039	0.1049	0.1075	0.1102	0.1210	0.1322	0.1030	0.1035	0.1061	0.1088	0.1196	0.1308
D10S1248	0.0825	0.0823	0.0848	0.0874	0.0978	0.1087	0.0824	0.0821	0.0846	0.0872	0.0976	0.1085
D12S391	0.0280	0.0273	0.0292	0.0311	0.0393	0.0483	0.0077	0.0073	0.0086	0.0098	0.0156	0.0224
D13S317	0.0733	0.0716	0.0742	0.0768	0.0873	0.0984	0.0242	0.0236	0.0253	0.0271	0.0348	0.0433
D16S539	0.0721	0.0721	0.0745	0.0770	0.0872	0.0979	0.0716	0.0716	0.0740	0.0765	0.0867	0.0974
D17S1301	0.1575	0.1585	0.1615	0.1645	0.1766	0.1890	0.1562	0.1572	0.1602	0.1632	0.1753	0.1877
D18S51	0.0255	0.0261	0.0279	0.0297	0.0375	0.0462	0.0247	0.0252	0.0269	0.0288	0.0365	0.0451
D19S433	0.0539	0.0541	0.0566	0.0591	0.0694	0.0802	0.0525	0.0527	0.0552	0.0577	0.0679	0.0786
D1S1656	0.0212	0.0213	0.0230	0.0247	0.0322	0.0405	0.0124	0.0127	0.0141	0.0156	0.0220	0.0294
D20S482	0.1351	0.1352	0.1384	0.1417	0.1547	0.1677	0.1350	0.1350	0.1382	0.1415	0.1545	0.1676
D21S11	0.0402	0.0416	0.0438	0.0460	0.0552	0.0650	0.0096	0.0102	0.0115	0.0129	0.0189	0.0260
D22S1045	0.0931	0.0898	0.0925	0.0952	0.1061	0.1175	0.0929	0.0895	0.0921	0.0948	0.1058	0.1172
D2S1338	0.0210	0.0213	0.0229	0.0246	0.0320	0.0402	0.0087	0.0085	0.0098	0.0110	0.0168	0.0235
D2S441	0.0836	0.0842	0.0868	0.0895	0.1004	0.1118	0.0575	0.0570	0.0595	0.0621	0.0725	0.0834
D3S1358	0.0938	0.0937	0.0962	0.0988	0.1094	0.1204	0.0245	0.0233	0.0251	0.0270	0.0348	0.0434
D4S2408	0.0887	0.0938	0.0963	0.0988	0.1091	0.1198	0.0697	0.0730	0.0753	0.0777	0.0876	0.0981
D5S818	0.1084	0.1076	0.1103	0.1130	0.1243	0.1358	0.0519	0.0523	0.0547	0.0572	0.0674	0.0781
D6S1043	0.0308	0.0309	0.0329	0.0349	0.0434	0.0526	0.0303	0.0304	0.0324	0.0344	0.0429	0.0521
D7S820	0.0698	0.0695	0.0719	0.0743	0.0844	0.0949	0.0694	0.0690	0.0714	0.0739	0.0839	0.0945
D8S1179	0.0550	0.0551	0.0574	0.0598	0.0695	0.0799	0.0174	0.0175	0.0191	0.0207	0.0279	0.0359
D9S1122	0.1411	0.1390	0.1417	0.1445	0.1559	0.1676	0.0574	0.0556	0.0580	0.0605	0.0706	0.0813
FGA	0.0291	0.0293	0.0311	0.0330	0.0412	0.0501	0.0282	0.0284	0.0302	0.0321	0.0402	0.0491
PentaD	0.0363	0.0358	0.0378	0.0399	0.0485	0.0579	0.0361	0.0356	0.0376	0.0397	0.0483	0.0577
PentaE	0.0134	0.0132	0.0146	0.0160	0.0225	0.0299	0.0131	0.0129	0.0143	0.0158	0.0222	0.0296
TH01	0.0742	0.0769	0.0793	0.0817	0.0918	0.1024	0.0740	0.0767	0.0791	0.0815	0.0916	0.1022
TPOX	0.1373	0.1356	0.1388	0.1420	0.1548	0.1677	0.1372	0.1354	0.1386	0.1418	0.1546	0.1675
vWA	0.0590	0.0604	0.0627	0.0651	0.0748	0.0852	0.0336	0.0343	0.0363	0.0384	0.0473	0.0568

Table 2.8: Observed and expected single-locus matching proportions for aSTRs using length-based and sequence-based data for different theta values. Expected values that underestimate observations are highlighted in red.

		Expected				
Locus	**Observed**	$\theta = 0$	$\theta = 0.01$	$\theta = 0.03$	$\theta = 0.05$	$\theta = 0.10$
rs1005533	0.3857	0.3887	0.3922	0.3997	0.4077	0.4295
rs10092491	0.3866	0.3950	0.3989	0.4069	0.4154	0.4381
rs1015250	0.3454	0.3796	0.3825	0.3889	0.3960	0.4163
rs1024116	0.4047	0.4144	0.4192	0.4288	0.4385	0.4635
rs1028528	0.3403	0.3778	0.3807	0.3869	0.3938	0.4137
rs1031825	0.3593	0.3785	0.3814	0.3877	0.3947	0.4147
rs10488710	0.3923	0.4020	0.4062	0.4149	0.4239	0.4475
rs10495407	0.4470	0.4525	0.4584	0.4699	0.4812	0.5087
rs1058083	0.3894	0.3819	0.3850	0.3917	0.3990	0.4197
rs10773760	0.3999	0.4051	0.4095	0.4185	0.4276	0.4516
rs10776839	0.3698	0.3847	0.3879	0.3950	0.4026	0.4237
rs1109037	0.3612	0.3767	0.3795	0.3856	0.3923	0.4120
rs1294331	0.3640	0.3920	0.3957	0.4035	0.4117	0.4340
rs12997453	0.3940	0.3961	0.4000	0.4082	0.4167	0.4396
rs13182883	0.3679	0.3814	0.3845	0.3911	0.3984	0.4190
rs13218440	0.4060	0.3850	0.3883	0.3953	0.4030	0.4241
rs1335873	0.3328	0.3755	0.3781	0.3840	0.3907	0.4101
rs1336071	0.3704	0.3851	0.3884	0.3955	0.4031	0.4243
rs1355366	0.3659	0.3915	0.3951	0.4029	0.4111	0.4333
rs1357617	0.5160	0.5079	0.5146	0.5274	0.5397	0.5685
rs1360288	0.4211	0.4221	0.4271	0.4372	0.4474	0.4730
rs1382387	0.3682	0.3906	0.3943	0.4020	0.4101	0.4322
rs1413212	0.3767	0.3945	0.3984	0.4064	0.4149	0.4375
rs1454361	0.3686	0.3856	0.3889	0.3961	0.4038	0.4250
rs1463729	0.3582	0.3853	0.3886	0.3957	0.4034	0.4246
rs1490413	0.3857	0.3767	0.3795	0.3856	0.3923	0.4120
rs1493232	0.3403	0.3772	0.3799	0.3861	0.3929	0.4127
rs1498553	0.3627	0.3751	0.3777	0.3836	0.3902	0.4095
rs1523537	0.3700	0.3769	0.3796	0.3857	0.3925	0.4122
rs1528460	0.3332	0.3752	0.3779	0.3838	0.3904	0.4098
rs159606	0.3862	0.3923	0.3960	0.4039	0.4121	0.4344
rs1736442	0.3622	0.3915	0.3952	0.4029	0.4111	0.4333
rs1821380	0.3723	0.3844	0.3876	0.3946	0.4022	0.4233
rs1886510	0.4476	0.4525	0.4584	0.4699	0.4812	0.5087
rs1979255	0.3842	0.3844	0.3876	0.3946	0.4022	0.4233
rs2040411	0.3741	0.3876	0.3911	0.3984	0.4063	0.4279
rs2046361	0.3528	0.3822	0.3853	0.3921	0.3994	0.4201
rs2056277	0.5963	0.5950	0.6018	0.6146	0.6267	0.6541
rs2076848	0.3843	0.3956	0.3995	0.4076	0.4162	0.4389
rs2107612	0.4461	0.4480	0.4537	0.4651	0.4763	0.5036
rs2111980	0.3624	0.3790	0.3820	0.3883	0.3954	0.4155
rs214955	0.3539	0.3752	0.3779	0.3838	0.3904	0.4098
rs221956	0.3864	0.3955	0.3994	0.4075	0.4160	0.4388
rs2269355	0.3796	0.3754	0.3780	0.3839	0.3906	0.4099
rs2342747	0.4005	0.4083	0.4128	0.4220	0.4313	0.4556
rs2399332	0.3766	0.3932	0.3970	0.4049	0.4132	0.4356
rs251934	0.4481	0.4514	0.4572	0.4687	0.4800	0.5074
rs279844	0.3591	0.3751	0.3777	0.3835	0.3902	0.4095
rs2830795	0.4239	0.4336	0.4390	0.4497	0.4603	0.4868

rs2831700	0.3712	0.3838	0.3870	0.3939	0.4015	0.4224
rs2920816	0.3811	0.3992	0.4033	0.4117	0.4205	0.4438
rs321198	0.3634	0.3778	0.3807	0.3869	0.3938	0.4137
rs338882	0.3760	0.3757	0.3783	0.3843	0.3909	0.4104
rs354439	0.3486	0.3750	0.3776	0.3835	0.3901	0.4094
rs3780962	0.3862	0.3770	0.3797	0.3859	0.3927	0.4124
rs430046	0.3967	0.3959	0.3999	0.4080	0.4166	0.4394
rs4364205	0.3888	0.3869	0.3903	0.3976	0.4054	0.4269
rs445251	0.3778	0.3829	0.3861	0.3930	0.4004	0.4212
rs4530059	0.4039	0.4089	0.4134	0.4227	0.4321	0.4565
rs4606077	0.4208	0.4083	0.4128	0.4220	0.4313	0.4556
rs560681	0.4308	0.4057	0.4101	0.4191	0.4284	0.4524
rs576261	0.3617	0.3764	0.3791	0.3851	0.3919	0.4115
rs6444724	0.3641	0.3764	0.3791	0.3851	0.3919	0.4115
rs6811238	0.3614	0.3774	0.3802	0.3863	0.3932	0.4130
rs6955448	0.4137	0.4189	0.4238	0.4338	0.4437	0.4691
rs7041158	0.3856	0.4029	0.4072	0.4160	0.4250	0.4487
rs717302	0.3807	0.3989	0.4030	0.4115	0.4202	0.4434
rs719366	0.4347	0.4415	0.4471	0.4582	0.4692	0.4961
rs722098	0.3509	0.3757	0.3784	0.3844	0.3910	0.4105
rs722290	0.3704	0.3750	0.3776	0.3835	0.3901	0.4094
rs727811	0.3409	0.3751	0.3778	0.3836	0.3902	0.4096
rs729172	0.4610	0.4622	0.4683	0.4801	0.4917	0.5196
rs733164	0.4401	0.4365	0.4419	0.4528	0.4636	0.4902
rs735155	0.3364	0.3757	0.3783	0.3843	0.3909	0.4104
rs737681	0.3725	0.3964	0.4004	0.4086	0.4172	0.4401
rs740598	0.3930	0.3981	0.4021	0.4105	0.4192	0.4423
rs740910	0.5370	0.5323	0.5391	0.5521	0.5645	0.5932
rs763869	0.3905	0.3792	0.3822	0.3886	0.3956	0.4158
rs8037429	0.3655	0.3771	0.3799	0.3860	0.3928	0.4126
rs8078417	0.4208	0.4213	0.4263	0.4364	0.4465	0.4720
rs826472	0.4689	0.4700	0.4762	0.4883	0.5000	0.5282
rs873196	0.4771	0.4791	0.4855	0.4978	0.5097	0.5381
rs876724	0.4216	0.4290	0.4342	0.4447	0.4552	0.4813
rs891700	0.3625	0.3750	0.3776	0.3835	0.3901	0.4094
rs901398	0.4103	0.4144	0.4192	0.4288	0.4385	0.4635
rs907100	0.3490	0.3796	0.3826	0.3890	0.3961	0.4164
rs914165	0.3563	0.3753	0.3780	0.3839	0.3905	0.4099
rs917118	0.3507	0.3794	0.3824	0.3888	0.3959	0.4161
rs938283	0.6170	0.6216	0.6282	0.6407	0.6524	0.6789
rs964681	0.4233	0.4299	0.4351	0.4457	0.4562	0.4824
rs987640	0.3606	0.3752	0.3779	0.3838	0.3904	0.4098
rs9905977	0.4085	0.4034	0.4076	0.4164	0.4255	0.4493
rs993934	0.3801	0.3821	0.3853	0.3920	0.3994	0.4201
rs9951171	0.3762	0.3787	0.3816	0.3879	0.3949	0.4149

Table 2.9: Observed and expected match probabilities per iiSNP marker for different theta values. Expected values that underestimate observations are highlighted in red.

Length-based					Sequence-based				
		Entropy					**Entropy**		
Locus	**Alleles**	Single	Combined	Conditional	**Locus**	**Alleles**	Single	Combined	Conditional
DYS385	70	3.3838	3.3838	3.3838	DYF387	220	4.4882	4.4882	4.4882
DYF387	65	3.1933	5.4420	2.0583	DYS612	36	2.4163	5.9328	1.4446
DYS612	17	2.2910	6.3538	0.9117	DYS385	99	3.4987	6.5666	0.6338
DYS570	12	1.7702	6.6837	0.3300	DYS389II	63	2.8672	6.8051	0.2385
DYS576	9	1.7834	6.8588	0.1751	DYS576	10	1.7866	6.9354	0.1303
DYS389II	10	1.5122	6.9370	0.0782	DYS570	22	1.8509	6.9744	0.0390
DYS549	7	1.2963	6.9722	0.0352	DYS390	27	1.8621	6.9886	0.0143
DYS481	14	2.0073	6.9907	0.0185	DYS549	9	1.3119	6.9949	0.0063
DYS391	7	0.8867	6.9962	0.0055	DYS505	9	1.2996	6.9987	0.0038
DYS439	7	1.2368	6.9992	0.0030	DYS439	7	1.2368	7.0017	0.0030
DYS505	9	1.2996	7.0017	0.0025	DYS19	9	1.4127	7.0029	0.0013
DYS19	8	1.4066	7.0029	0.0013	DYS522	8	1.2134	7.0042	0.0013
DYS522	7	1.2073	7.0042	0.0013	DYS389I	6	1.0228	0.0000	0.0000
DYS389I	6	1.0228	0.0000	0.0000	DYS391	8	0.8934	0.0000	0.0000
DYS390	8	1.6089	0.0000	0.0000	DYS392	10	1.1897	0.0000	0.0000
DYS392	10	1.1897	0.0000	0.0000	DYS437	14	1.1979	0.0000	0.0000
DYS437	5	1.0694	0.0000	0.0000	DYS438	9	1.3989	0.0000	0.0000
DYS438	7	1.3704	0.0000	0.0000	DYS448	29	2.0590	0.0000	0.0000
DYS448	11	1.5111	0.0000	0.0000	DYS460	8	1.0319	0.0000	0.0000
DYS460	6	1.0189	0.0000	0.0000	DYS481	23	2.1817	0.0000	0.0000
DYS533	6	1.2275	0.0000	0.0000	DYS533	8	1.2450	0.0000	0.0000
DYS635	11	1.6161	0.0000	0.0000	DYS635	28	1.9732	0.0000	0.0000
DYS643	11	1.6714	0.0000	0.0000	DYS643	12	1.6768	0.0000	0.0000
YGATAH4	6	1.0847	0.0000	0.0000	YGATAH4	8	1.0977	0.0000	0.0000

Table 2.10: Entropy measures for length-based and sequence-based Y-STR data.

Chapter 3

Perceptions of forensic scientists on statistical models, sequence data, and ethical implications for DNA evidence evaluations: a qualitative assessment

This chapter is based on Sanne E. Aalbers, Alyna T. Khan, Bruce S. Weir. *Perceptions of forensic scientists on statistical models, sequence data, and ethical implications for DNA evidence evaluations: a qualitative assessment.* Forensic Science International: Synergy 6 (2023). https://doi.org/10.1016/j.fsisyn.2023.100335.

Abstract

With the introduction of next generation sequencing (NGS) technology in the forensic field, it will be of interest to assess if forensic scientists feel equipped to interpret and present DNA evidence for sequence data. Here, we describe perceptions of sixteen U.S.-based forensic scientists on statistical models, sequence data, and ethical implications for DNA evidence evaluations.

To get an in-depth understanding of the current situation, we used a qualitative research approach with a cross-sectional study design. Semi-structured interviews ($N = 16$) were conducted with U.S. forensic scientists working with DNA evidence. Open-ended interview questions were used to explore participants' views and needs surrounding the use of statistical models and sequence data for forensic purposes. We conducted a conventional content analysis using ATLAS.ti software and employed a second coder to ensure reliability of our results.

Eleven themes emerged: 1) a statistical model that maximizes the value of the evidence is

preferred; 2) a high-level understanding of the statistical model used is generally sufficient; 3) transparency is key in minimizing the risk of creating black boxes; 4) training and education should be an ongoing effort; 5) the effectiveness of presenting results in court can be improved; 6) NGS has the potential to become revolutionary; 7) some hesitations surrounding the use of sequence data remain; 8) there is a need for a concrete plan to alleviate barriers to the implementation of sequencing techniques; 9) ethics plays a major part in the role of a forensic scientist; 10) ethical barriers for sequence data depend on the application; 11) DNA evidence has its limitations.

The results of this study give insight into the perceptions of forensic scientists regarding the use of statistical models and sequence data, providing valuable information in the move towards implementing sequencing methods for DNA evidence evaluations.

Introduction

DNA typing is a mature field and overwhelmingly seen as the gold standard in forensic evidence. The interpretation of DNA evidence, however, is far from straightforward and challenges arise when evaluating complex profiles and assessing the statistical weight of the evidence. Accurate representation of forensic evidence in court is crucial to avoid misinterpretations and, ultimately, to reduce the possibility of a miscarriage of justice. This not only requires sensible models that can handle the complexity associated with DNA profiles, but also an understanding of the methods used by forensic scientists who will be writing the reports and potentially serving as expert witnesses in court.

In general, when reporting an inclusion, admissible DNA evidence in court needs to be accompanied by a quantitative statement. The forensic scientist is often requested to provide additional meaning to these results. Although the scientific evidence is restricted to the DNA profile, the trier of fact needs to incorporate this to decide on the ultimate issue of guilt. This requires additional links between the evidence and an inference of contact with the crime scene as well as an association with the crime, while also incorporating all other relevant information available. Correct presentation of the DNA results by the forensic expert is crucial to ensure

that statements are related only to probabilities regarding the DNA evidence.

Valid probabilistic reasoning is not easy and numerous studies have been conducted showing the occurrence of fallacies manifesting within the forensic community [21, 43, 59, 71]. Over the years, mitigation strategies have been proposed to reduce the effect of bias in forensic decision-making [22, 53]. To increase our understanding of how forensic scientists feel about interpreting and presenting DNA evidence, it will be valuable to obtain direct input from this group. Such studies have the potential to illuminate the barriers faced by these professionals and can serve as guidance for the implementation of statistical models for new techniques and applications. This is especially timely in light of the transformation of current approaches to the incorporation of next generation sequencing (NGS) technology.

Early research has focused on the perceptions of sequencing technologies within the field through surveys and highlighted opinions on current use, future views, and challenges in forensics [26]. To get a more in-depth understanding of the current situation, we conducted a qualitative study involving interviews with U.S. forensic scientists working with DNA evidence. The objective of this study was to describe the views and needs of these professionals surrounding the use of statistical models and sequence data for forensic purposes.

Methods

Study design

We conducted a cross-sectional study involving forensic scientists based in the U.S. and working with DNA evidence. Semi-structured interviews were used to explore the views and needs of this group of professionals over three domains. The first two domains focused on the application of statistical models and use of sequence data in forensic DNA evidence evaluations, respectively. The final part assessed some ethical topics concerning these concepts. An interview guide was developed with open-ended questions over the three domains and subsequently refined using a key informant (see the Appendix for a detailed overview of the final guide). A mock interview was conducted before proceeding with official interviews. All study activities were reviewed and approved by the University of Washington Human Subjects

Division.

Recruitment

We reached out to the organization behind the International Symposium on Human Identification (ISHI) for recruitment purposes. ISHI is the largest symposium focusing solely on DNA forensics with about 1,000 forensic experts from around the world attending their yearly event[1]. They agreed to use their network to reach out to 90 individuals who previously attended NGS-based workshops. Initial emails sent out by the organization contained a study description and a link to an external form where people could indicate their interest in participating in the study and leave their contact information. We followed up with those individuals who expressed interest in participating with a second email asking for more background information to confirm eligibility and to set up a time for an interview. To increase response rates, a second batch of recruitment took place by reaching out to our own contacts.

Individuals were eligible if they were employed by a U.S. forensic laboratory at the time of the study and they worked with DNA evidence. Eligible candidates were invited to participate in a one-time 45-minute confidential interview over Zoom. A modest incentive in the form of a $25 gift card was offered in return for their participation, although not all participants could accept this incentive. During recruitment, we collected background information on the size of the workplace, number of years in the field, and whether or not the participant had experience with court testimony. Our goal was to recruit individuals with different work experiences to gather a range of perspectives. Recruitment took place over two months and resulted in 24 completed recruitment forms. A total of 16 individuals were successfully contacted and interviewed. Of the remaining eight individuals, six did not respond to our follow-up emails and two had to drop out due to personal circumstances.

Data collection

Data collection was performed over a period of three months (April – June 2022). Zoom interviews were scheduled at a time convenient to the interviewee. All interviews were conducted in English and digitally recorded. Verbal consent was obtained from each of the

[1] `https://www.ishinews.com/` (Accessed February 23, 2023).

participants at the beginning of the interview and included permission to record the interview. The automatically generated audio transcription files of completed recordings were manually curated to remove errors and to anonymize the data. The resulting transcripts were assigned a unique identifier to ensure confidentiality.

Data analysis

ATLAS.ti v.9 [1] was used to support coding, analysis, and data management. Transcripts were subjected to a conventional content analysis using a mixed approach of top-down and open coding until code saturation was reached [49, 38]. A subset of transcripts was independently coded and reviewed by a second coder. Coding differences were resolved through discussion. We collapsed the final codes into initial themes and translated these into underlying concepts as they emerged from the data. During the late-stage analysis, we used groundedness (total occurrence of a code) and pervasiveness (occurrence over unique transcripts) metrics to maximize our ability to identify all relevant themes.

Results

Participant characteristics

The participants of our study represented twelve different states from around the U.S.[2] Their educational background included degrees in chemistry, biology, genetics, and forensic science. Work experience with DNA evidence ranged from less than a year to over 33 years (median of 13 years). The majority of interviewees indicated using probabilistic genotyping (PG) software as part of their jobs and a handful described having been actively involved in the validation process. Except for early-career scientists, almost every participant had experience with expert testimony in a court setting. Most participants reported that their workplaces did not use sequence data at the time of the interview and thus had no practical experience working with it. Four participants reported to be in the validation process of a sequencing technique or had recently completed validation and were working on implementa-

[2]The following states are represented in this study: Arizona, California, Colorado, Florida, Louisiana, Minnesota, Nebraska, New York, North Carolina, Oklahoma, South Carolina, Washington.

tion. Participant characteristics are summarized in Table 3.1.

Characteristic – $N(\%)$	$N = 16$
Work experience	
≤ 5 years	5 (31.3%)
> 5 years	11 (68.7%)
Experience with court testimony	
No	4 (25.0%)
Yes	12 (75.0%)
Use of statistical software	
STRmix	11 (68.7%)
TrueAllele	2 (12.5%)
None or other (e.g., Popstats)	3 (18.8%)
Experience with sequence data	
No	10 (62.5%)
Yes	2 (12.5%)
In validation/implementation	4 (25.0%)

Table 3.1: Participant characteristics.

Theme identification

Nine main themes emerged describing participants' views and needs concerning the application of statistical models and sequence data for forensic DNA evidence evaluations. Two additional themes came up while exploring some ethical topics related to these concepts. The themes are graphically illustrated in Figure 3.1 over the three domains.

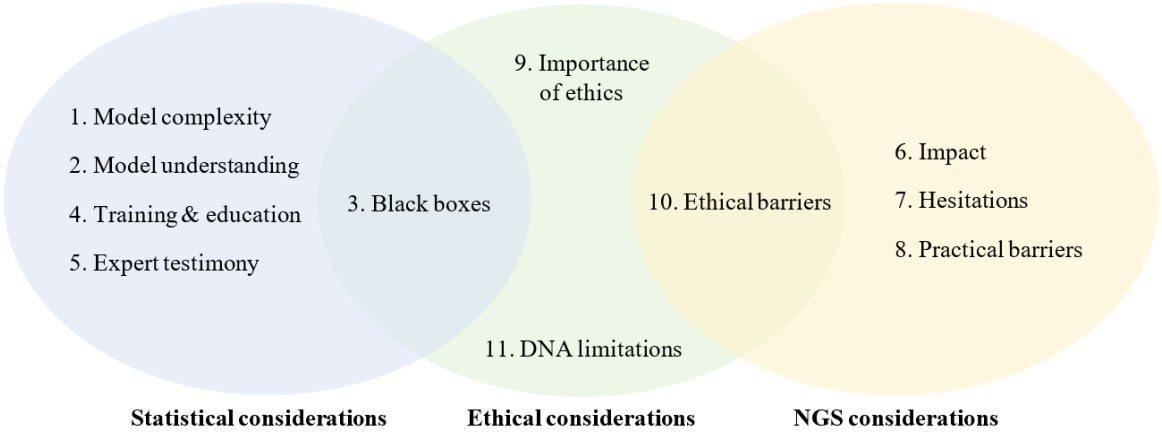

Figure 3.1: Graphical illustration of the identified themes within each of three domains and areas of overlap.

Theme 1: A statistical model that maximizes the value of the evidence is preferred.

Participants expressed their support for probabilistic genotyping software and generally preferred a statistical model that uses as much data as possible. Although there was agreement that such models can be more difficult to understand, most (81%) noted that this should not be a reason for resorting to more simplistic approaches.

> *"I wouldn't think: 'Oh it's too difficult, I wish we didn't have to do this'. If that's the best way to do it, then that's the way we do it." (P9)*

> *"[F]or me it is 100% worth it to have a more complicated model in order to be able to generate more accurate results, whether it's an inclusion or an exclusion." (P7)*

Others mentioned the need to find a middle ground between model complexity and an analyst's needs. These participants highlighted the importance of keeping in mind the question being asked, the quality of the data, and whether results can be explained well in court.

> *"I feel like it's just this middle ground that we have to find where an analyst does feel comfortable enough to talk about it and then understand what [the statistical models] are saying." (P2)*

Overall, participants valued consistency and being able to use the same software for all cases. They were looking for something they can trust and that is easy to use. Moreover, the forensic scientists included in this study wanted to make sure that they are maximizing the value of the evidence, which was considered most fair to all parties involved.

Theme 2: A high-level understanding of the statistical model used is generally sufficient.

While it may be preferred to have a full understanding of the models used during an analysis, participants overwhelmingly agreed that this is not realistic or even needed in most settings. Multiple participants used an analogy to describe their views on this topic:

> *"One analogy that I thought of is: I've taken a class on how to use a graphing calculator. I can use a lot of functions on it, but do I know the programming that went into it, to be able to do these things...?" (P8)*

> *"You have this analogy; it is so simple. I drive a car. I know how to drive the car, but do I have to know how the engine works from top to bottom? Absolutely not. I don't know how to do that, nor do I want to learn. (P12)*

We noted though that participants did feel responsible to not "just plug in data" and have at least a high-level understanding of the statistical models used. This holds especially in light of being able to adequately present results in court. Although a simplified explanation is often preferred in a court setting, participants acknowledged that this required a strong foundational knowledge of the statistical model.

> *"It's not just about I press go, and then accept everything that comes out. [...] I really need to understand how it's working in order to present to court that I have some confidence, that I trust what it's doing." (P13)*

Furthermore, the required level of proficiency may also depend on the maturity of the technology.

> *"At the beginning of the technology, as it's being introduced, yeah, you have to have a really good understanding of it to be able to present it in court for the first time. And then, as it gets more routine, I think it's less important that you have a deep dive understanding of it." (P14)*

Theme 3: Transparency is key in minimizing the risk of creating black boxes.

PG software has been criticized in the literature due to its perceived black box nature [64] and the use of sophisticated modeling techniques has been the subject of several discussions within the forensic science community [12, 50, 65, 51, 7]. Participants understood the sentiment and agreed that, from the outside, PG software can look like a "black box", especially in light of the underlying code not being readily accessible. However, this was not necessarily seen as a problem or even a relevant issue. Participants highlighted the amount of research, training, and validation that goes into setting up PG software for DNA casework. They expressed trust in the software used within their labs, provided that the developers remain transparent about their work.

> *"I don't love [the black box] part of it. I think it simplifies how much work goes into understanding how it works within your lab setting, and how much work goes into all the validation that has to happen before we start using it." (P16)*

Furthermore, a recurring thought amongst participants was that the "black box issue" was not a relevant point of concern and mainly used as a Defense tactic.

> *"[I]t's part of the game, I guess, that they're going to challenge. They want to see what they can't see and as soon as you let them see it then it's not a big issue." (P14)*

Theme 4: Training and education should be an ongoing effort.

When asked about their needs with respect to statistical modeling for DNA evidence evaluations, participants noted that they can never have enough training. While extensive training occurs in certain situations, such as for new hires and with the introduction of a new technique or software, participants expressed interest in having ongoing training opportunities as well. Training courses are seen as most useful when they provide hands-on experience and are adapted to deal with different learning needs and levels. It may also be helpful to focus on new scientific developments as participants reported struggling to keep up with relevant papers.

> *"I know people learn differently, but when it comes to [PG software] I think you need to do it to learn it. [...] I think the biggest thing is training. [...] And definitely even continued training." (P2)*

> *"I think probably the biggest challenge would be the continuing education of keeping up to date with whatever the current model is, because it is easy in school to learn what the most cutting-edge technique is, but then that only stays cutting edge for so long." (P9)*

That being said, it is important to be mindful about time constraints as forensic scientists may experience work pressure and are not always able to make education a priority.

Theme 5: The effectiveness of presenting results in court can be improved.

Participants felt a huge responsibility when it comes to effectively explaining DNA analysis results in court. One difficulty brought up was finding a balance between explaining results thoroughly but not too complicated.

> *"That's always one of the challenges, to explain our results, and one of the downsides to using [PG software] is that it is hard to explain and that a [random match probability] is much easier to explain. [...] It falls on us to explain that, as best as we can." (P1)*

> *"I'm always seeking what's the best...simplest explanation that I can provide somebody to help them understand something that they perceive as being super complicated." (P3)*

Participants also expressed frustrations in dealing with lawyers, especially in case of fallacies or a (deliberate) misdirection.

> *"No matter how we explain it, [lawyers] are going to interpret [the likelihood ratio] in a different way. They are going to interpret it as the transposed conditional most of the time. [...] That part is difficult." (P10)*

> *"I guess really the thing that's hardest with the statistics right now is just different defense experts come forward and, generally speaking, bring up something that really does not matter. [...] It makes you have to jump through a whole bunch of more hoops." (P7)*

Notably, one participant mentioned that well-intentioned efforts from the field to create awareness regarding fallacies may lead to more paranoia among forensic scientists. The result may be a rigid approach with statements being made solely for the transcript instead of focusing on the people in court. Multiple participants were proponents of offering training courses to justice officials, especially for the Defense. Others also suggested creating handouts or videos for juries.

"It's tough because at the end of the day you got to explain it to the jury. The jury won't have that background in statistics. But it will definitely be helpful if at least the lawyers will have them." (P15)

Theme 6: NGS has the potential to become revolutionary.

Participants expressed their excitement when talking about the potential of NGS techniques in a forensic setting.

"I think this is one of the most exciting things that's come about in forensic DNA probably since STRs." (P3)

"I feel it's going to be revolutionary. [...] There have been tremendous changes and I'll just say...the opportunity now is greater than any time that I've seen." (P6)

Numerous applications for sequence data were brought up, including but not limited to unidentified human remains, cold cases, investigative genetic genealogy (IGG), forensic phenotyping and ancestry inference, and mixture and low-level contributor deconvolution. Having a technique that provides access to all the data at once was seen as the main benefit. Participants thought that NGS techniques will initially be employed for lead-generating approaches, such as IGG, noting that it would be relatively easy to implement and has already proven to be successful. Most agreed that we are still a long way out until sequence data will be fully incorporated into forensic casework. Some envisioned sequencing techniques to become the standard, while more than half of the participants (63%) saw the use of these as a specialized application with short tandem repeat (STR) typing through capillary electrophoresis (CE) remaining the default approach.

"I think, in my opinion, it would just take over the whole platform. No capillary electrophoresis, it would spit out the STR, Y, and the SNP at the same time and you're good." (P15)

"If the technology gets robust enough that we can convert it, all of our STR profiles, to be compatible with that, I could see it replacing it entirely. But I think that would be a very long way out." (P8)

"I don't think it will ever outpace CE because that workflow is very straightforward and reasonably...it's still expensive but it's just a lot fewer steps and it's so much easier to analyze than [NGS] data." (P11)

Theme 7: Some hesitations surrounding the use of sequence data remain.

Despite the enthusiasm among participants about NGS techniques, there also exist concerns. Participants acknowledged that there are still a lot of uncertainties surrounding the application of sequence data and it is not always clear what the added value exactly is. It was also noted that there may be a reluctance within the community to accept new technologies, partly because of a fear of change and (public) misconceptions. Some expressed worries that this may lead to sequence data not reaching their potential.

"I have a hard time convincing myself that the extra information is worth it, given the current statistics that we get. Is ten to the 30th not high enough?" (P10)

"I think there's so much potential, and I think people have hesitated. Probably because there's certain things that weren't there, but it has kind of become like this chicken or egg thing." (P16)

Participants noted that such concerns are always an issue with the introduction of new technologies, and it takes time to get over the acceptance hump. Promotions and a push from high up are believed to be beneficial to overcoming existing hesitations.

"We are all human beings, so you do have to get over the little bit of personal fear. People don't like change. Sell me on it a little bit. Help me become part of the change." (P6)

Theme 8: There is a need for a concrete plan to alleviate barriers to the implementation of sequencing techniques.

When asked about specific barriers surrounding sequence data, participants brought up numerous practical issues. First, it was noted that the implementation of sequencing techniques requires a massive investment from a laboratory. The need for money, time, training, and staff was brought up numerous times. Moreover, participants acknowledged that the decision-making was mostly out of their hands, and it could well be that priority was given to other applications.

> *"[Laboratories] don't want to train their whole staff all over. Maybe they don't have time to do the validation work and to do research for it." (P5)*

> *"I feel the sale is there, [...] so you have the industry's interest and now it's a matter of working to reduce the barriers through training, technology, business case, all of those things." (P6)*

Second, participants experienced a lack of data and resources. This includes the need for sequence-based databases, samples to be used for validation purposes, and the need for reliable software. Specifically, participants noted that they do not want to change PG software and were hoping for an update that could accommodate sequence data if their labs decided to proceed in that direction. At the time of writing, the developers of the PG software STRmix just released a paper introducing their newest software for sequence data [15].

> *"[T]here's no autosomal sequence data that exists in a database that we can search the same way that we can search STRs." (P9)*

> *"The amount of data being generated from [NGS technology] is huge. How to deal with that is going to be another problem." (P4)*

> *"We definitely need a probabilistic genotyping tool to be able to incorporate [sequence] data. [...] I don't think I would want to go back to a different model, or to a different kind of statistical tool." (P16)*

Overall, participants expressed a need for a concrete plan. This includes guidance on preparing a laboratory for the implementation of sequencing technology to ultimately running an analysis and presenting results when applied to casework.

Theme 9: Ethics plays a major part in the role of a forensic scientist.

Participants described ethics as playing a major daily role in their jobs. Responses indicated that the topic was seen as extremely important, with a need for continuous promotion and requiring yearly training. Some participants brought up the potential of bias creeping in during an investigation and valued a work environment that promotes open discussions to minimize such risks. While a forensic scientist may strive towards objectivity, a few participants noted that they are not operating in a vacuum. There exists an added complexity in being linked to law enforcement. This may lead to misconceptions that forensic scientists work for one side only. One participant also expressed frustration in dealing with pressure coming from investigators.

> *"Everybody has conspiracy theories, like the labs are in cahoots with the prosecution, which is totally wrong. [...] I'm for the truth, for the evidence." (P13)*

> *"[Investigators] come in guns blazing and want us to start working. [...] Not that they are asking us to be unethical, [...] they just want it done now. And our stance at the laboratory has always been quality over quantity, all day long." (P12)*

Theme 10: Ethical barriers for sequence data depend on the application.

The majority of the participants (75%) indicated not seeing any potential harm for sequence data obtained from STR regions. Responses also showed that using NGS techniques for investigative leads was seen as unproblematic. Participants were more hesitant when it comes to novel sequencing techniques and data outside of the standard STR regions, including phenotyping, ancestry-related information, and single nucleotide polymorphism (SNP) data that may inadvertently reveal medical information.

> *"When you think about the traditional analysis of STRs, I don't think that [potential harms are] too much of a problem. The things that concern me are some of the newer marker types like ancestry estimation." (P11)*

That being said, many noted that law enforcement is held to high standards and that forensic laboratories are used to having tremendous checks and balances in place. Nevertheless, some

participants acknowledged that having safeguards in place never completely takes away all risks and there always exists a potential of misuse.

> *"I don't think there could be any harm in [DNA sequence] information if the information is interpreted or looked at in an appropriate way. Everybody has the ability to spin things, or make things look worse than they are, or not in the right context." (P12)*

While privacy concerns and the risk of misuse and misinterpretation should be reasons for thoroughly vetting the use of sequence data, participants said that they should not be a reason to discontinue NGS techniques. Others mentioned the existence of misconceptions from the public and how time and information would likely overcome fears.

When talking about the use of whole genome sequence (WGS) data, participants were more divided in their opinions. About a third saw this as a next step and great opportunity to get even more data, while others were more hesitant and saw no direct need for such data for forensic purposes.

> *"I think that [using WGS data] is our next step." (P3)*

> *"I don't see why we need to move to [WGS data] into the future. [...] I think it seems a bit excessive and probably an unnecessary amount of information, but how the future moves forward, I don't know." (P4)*

> *"Let's do the information we need. Let's not do it all, because we can do it all." (P2)*

Theme 11: DNA evidence has its limitations.

Interestingly, multiple participants brought up situations where DNA evidence may be of limited value. One observation was that participants experienced a shift in the court room from source level to activity level, something that has been noted in the literature as well [68, 34, 83]. Furthermore, while the increase in sensitivity of technologies may help deconvolute complex mixtures involving low contributors, participants noted that there exists a line

between pushing the limits and getting unreliable results. This may point towards concerns about an overreliance on DNA evidence, an issue that was also reported in a worldwide survey of forensic scientists [3]. This notion can also be problematic when dealing with extreme statistics that may become overwhelming.

> *"[I]n the big scheme of things: what is the difference between one in a billion and one in 10 billion?" (P5)*

> *"We've worked so hard at getting to sensitivity levels where we can detect so little DNA. I don't necessarily think it's bad, but I think that sometimes DNA is not the answer to the question. Sometimes DNA doesn't help at all." (P1)*

Discussion

Our findings describe several factors contributing to the perceptions of forensic scientists surrounding the use of statistical models and sequence data for DNA evidence evaluations. Although our study results are limited to a small set of professionals and are therefore not readily generalizable, we can frame our findings in a broader context by drawing from existing literature. With respect to statistical modeling, participants noted that forensic scientists are generally not experts in statistics and that this holds even more true for legal practitioners and jurors in a court setting. Yet, the forensic scientist is tasked with the difficult job of presenting the statistical weight of DNA evidence during expert testimony. The occurrence of themes relating to the need for training and education, and the improvement of presenting results in court, was therefore unsurprising. Similar observations are described in the literature. Eldridge suggests that it is apparent that improvements should be made when it comes to expert testimony, noting that juries do generally not interpret results as intended [23]. Unfortunately, it is less clear how such changes will look, and the forensic community is still in search of the most desirable way to present evidence, if it even exists [23]. In terms of providing handouts to juries, as suggested by one participant of our study, focusing on visual aids may be helpful [23]. Still, there will always exist factors that complicate expert testimony for DNA evidence due to the nature of the criminal justice setting. While frustrations in

dealing with such difficulties situations, as brought up by some of our participants, may be valid, it has also been noted that it is precisely the defense lawyer's job to create doubt [66].

While participants acknowledged the increased level of complexity involved in PG software, we noticed high levels of support for such models. Even though it may complicate the presentation of results as compared to resorting to more simplistic approaches, this was not seen as a reason to oppose the use of PG software. This sentiment is echoed by other key criminal justice stakeholders, as described in a recent qualitative study on probabilistic reporting in forensic science [66]. These different stakeholder groups, including judges, prosecutors, and defense attorneys, highlighted the need for forensic scientists to "accurately and impartially convey their findings" and for investments in training and education on the use of statistical models [66]. A big aspect in creating comfort surrounding the use of statistical models is having validations in place, something that was brought up by our participants but is also pointed out in the aforementioned study [66]. Validations can help alleviate concerns and show that results are accurate and trustworthy. In addition, transparency is invaluable in creating trust. This may include having access to a good support team from software developers and being aware of the workings and limitations of the model. While disclosure of the source code may also be helpful in establishing trust, opinions among different stakeholder groups seem to differ on the necessity of providing such access by default [66].

When it comes to sequence data participants saw huge potential for numerous applications, with a primary initial use case for SNP data in investigative leads. The NGS applications brought up during the interviews have been well-described in the literature [6, 13]. Participants' opinions differed on the future of forensic DNA typing, with some seeing sequencing techniques as an addition to the current workflow while others saw it replacing CE-based STR technology entirely. Whatever the situation, all participants agreed that a shift will take years, potentially ranging from about 5 to 10, or even 20 years. Several factors play a role in these beliefs. First and foremost, participants noted that practical barriers currently prohibit the implementation of sequencing techniques for routine casework. Over the years, studies have published guidelines as well as sequence-based data necessary for NGS-based DNA evidence evaluations [60, 31, 32, 2]. Despite these developments, technical barriers still remain,

and the issues noted by our participants reflect findings from other publications [26, 18, 27]. In addition to practical barriers, participants advocated for the consideration of perceptual and ethical barriers. While some participants expressed interest in wanting to be at the forefront of new developments, others would rather wait until all the kinks have been worked out. Participants also expressed concern about the (lack of) added value of sequence data. Opinions on ethical implications seem to differ based on the application. While the use of the standard forensic STR markers is seen as unproblematic among our participants, potential harm from sequence data was deemed more likely to occur for other marker systems.

Our work has both strengths and limitations. Our background in forensic statistics put us in a great position to carry out this work. However, we are aware of the possibility of our work being biased. To minimize the risk of results being influenced by personal views, we deliberately opted for a largely open coding approach to let the data guide our analysis. Our second coder was less familiar with the topic, which provided a unique perspective and maximized our ability to identify relevant themes. As with most qualitative studies, this work may suffer from self-selection bias. Specifically, our sample consists of individuals who expressed interest in NGS technology and had previously attended NGS-based workshops, which may have led to biased opinions in favor of NGS applications. Furthermore, due to the limited sample size and way of sampling, our findings are not representative of the forensic community in general. Nevertheless, we believe our data provides a rich set of perspectives on the topics of interest.

Conclusion

In this paper, we highlighted important themes concerning statistical concepts, sequence data, ethical implications, and their interactions within the field of forensic DNA evidence evaluations based on in-depth interviews with sixteen U.S. forensic scientists. We showed to what extent they felt the need to have an understanding of the statistical models used to be able to perform their work and what aspects they valued in such models and PG software. We also identified experienced barriers and needs in light of feeling better prepared to work

with statistical models, as well as to present such results in a court setting. Finally, we discussed the perceived impact of sequence data on the forensic field and revealed practical barriers and ethical considerations to take into account for such new technologies. While the perceptions of forensic scientists with respect to these topics provide valuable input, it is important to remember that they are part of a larger system. On the one hand, barriers need to be addressed from a scientific standpoint, including the creation of databases and PG software for sequence data. On the other hand, the business side requires consideration of higher-level organizational issues, including funding and management needs. We believe both perspectives need to be taken into account in our journey towards successfully implementing sequencing methods for DNA evidence evaluations.

Appendix: Interview guide

Part / Domain	Activities / Questions
I. Introduction	Provide general study background
	Describe interview process
	Obtain consent
	Start recording
	Verify participant background
	Q1. Can you tell me more about your job and what you do at work?
II. Content	
Statistical considerations	Q1. What is your personal experience working with statistical concepts and modeling in a forensic setting?
	Q2. How comfortable do you feel with concepts like the random match probability, the likelihood ratio, and probabilistic genotyping?
	Q3. To what extent do you feel you need to be proficient in statistics to perform your work?
	Q4. To what extent can or should we ask for proficiency in statistics of lawyers and judges?
	Q5. For probabilistic genotyping software, to what extent do you feel you need to understand the underlying modeling concepts?
	Q6. Do you think using the most sophisticated statistical model is always the best option?
	Q7. Are there any barriers related to this topic you are facing in your work?
	Q8. Do you feel you have adequate access to training opportunities?
	Q9. What would you need to feel better equipped to work with statistical models for DNA evidence?
	Q10. What would you need to feel better equipped to present such evidence to a jury / in court?
NGS considerations	Q1. What is your personal experience working with sequence data in a forensic setting?
	Q2. How comfortable do you currently feel working with sequence data?
	Q3. How do you think sequence data will change what forensic scientists do?
	Q4. Do you think sequencing technologies will take over STR CE typing completely?
	Q5. What would you need to feel better equipped/prepared to work with sequence data?
Ethical considerations	Q1. How do you feel about the term “black boxes”, which is sometimes used to describe probabilistic genotyping software?
	Q2. Do you think the introduction of sequence data will be beneficial to the forensic community?
	Q3. Do you foresee potential harm with the introduction of sequence data for forensic purposes?
	Q4. How do you feel about the use of whole genome sequence data for forensic purposes?
	Q5. Are there any ethical issues you are dealing /struggling with in regard to your work as a forensic scientist?
III. Conclusion	Q1. Is there anything else you would like to add?
	Stop recording
	Describe follow-up process

CONCLUSION

This research has yielded several scientific contributions to the forensic field and the study of forensic sequence data in particular. By integrating both quantitative and qualitative work, we provided a holistic view of the current state of the forensic field with respect to NGS methods. We were able to gain insights into theoretical aspects as well as some practical implications of the use of sequence data for forensic purposes.

In the first part of this dissertation, we provided previously unavailable sequence-based estimates of population genetic parameters. Such estimates are crucial for the evaluation of the statistical weight of DNA evidence profiles. While software tools are available to obtain population genetic estimates for forensic data, these tools, as well as most published estimates, use the Weir and Cockerham estimator [79]. Nowadays, the less restrictive model as outlined by Weir and Goudet [80] is recommended, which allows for variation between populations and has been gaining attention in recent years [86, 16]. To the best of our knowledge, we are the first to apply this framework to forensic sequence data. Moreover, our use of genotypic data allows for dependencies that cannot be captured with allelic data.

The second chapter builds upon existing knowledge on the lack of independence between genetic markers [77, 46, 20]. Forensic calculations often assume independence across autosomal markers and our work brings renewed attention to this problem by showing that sequence data generally results in even less conservative results compared to length-based data. While NGS technologies are capable of producing data across different marker systems simultaneously, it is currently still unclear how the forensic community is planning to incorporate such markers for DNA evidence evaluations. Our focus on the diminishing return of observing matching markers after a certain point, provides caution against the practice of adding more markers with the intent to increase the discriminatory power of DNA profiling methods.

Finally, we contributed to a more in-depth understanding of the current situation with respect to sequencing technologies within the forensic community. While early research has been conducted through the use of a survey [26], our qualitative approach provided unique insights from the forensic analyst's perspective. Overall, we hope that our consideration of both theoretical and practical implications of forensic sequence data provides valuable guidance in the move towards implementing NGS technology for DNA evidence evaluations.

Recommendations

Below we describe and summarize several recommendations based on our research findings.

- While the use of a theta correction is endorsed within the forensic community [52, 60], there is currently no formal requirement to account for existing genetic dependencies when evaluating the strength of DNA evidence. We strongly recommend the incorporation of appropriate population genetic parameter estimates in forensic calculations, as this can account for dependencies within as well as between forensic markers.

- Since available forensic software tools provide population genetic estimates based on outdated models, we urge developers to consider updating their analyses to be based on more recent frameworks. Ideally, such estimates are based on genotypic data, but we recognize the need for the support of allelic data as those are generally more readily available.

- We recommend avoiding the continual addition of forensic markers to DNA profiles, without careful consideration of the impact on DNA evidence evaluations. The intention of increasing discriminatory power and decreasing match probabilities by including more markers, while assuming independence, is misguided and has the potential of being prejudicial.

- Continuing on the previous point, care should be taken when considering markers across different systems as dependencies occur even for unlinked loci. At a minimum, we

recommended the use of joint theta measures if marker sets are going to be combined across different systems.

- Lastly, we advocate for a continued close collaboration between disciplines within the forensic science community to enhance the interpretation of DNA evidence evaluations. While collaboration efforts are common for validation studies [45, 9], we believe it is beneficial to initiate ongoing discussions during and after the implementation stage of forensic methods as well. Such an approach allows for the incorporation of different perspectives and may help alleviate existing barriers as experienced by forensic science professionals.

Future work

We conclude with providing several ideas for future research that could build upon and extend the work in this dissertation.

- It would be of interest to repeat the estimation of population genetic parameters for Y-STR data. While the notion of inbreeding does not apply to haploid data, estimates of population structure are required if Y-STR sequence data are to be used for forensic evidence evaluations. Our data set was too limited for this analysis and a much larger data set would be required to obtain useful results.

- At the time of our analysis, we did not have access to the NIST 1036 data on iiSNP markers and resorted to iiSNP data on a small sample of 350 individuals. These data have since been published [44], so it is now possible to use an extended data set to assess population genetic estimates for iiSNP data. Our initial results showed different patterns for iiSNP data as compared to STR data, as well as compared to SNP analysis not restricted to forensic markers [85]. Although this may be a consequence of differences between marker systems and number of markers used, it would be useful to check for concordance between our observations and results obtained from the NIST iiSNP data.

- Our results are based on sequence data that were restricted mostly to the repeat regions.

As such, the impact of additional flanking region variation that can be observed with NGS techniques, was not assessed in this dissertation. Due to their close proximity to the STR region, this type of variation requires the consideration of haplotype frequencies [30], and it would be of interest to characterize the extent of dependencies between these regions. It should be noted that this may require updating software features as standard reports may not automatically incorporate known flanking region variations.

- Initial work on estimating adjusted population structure values for the combination of different marker systems may be extended to the use of sequence data. We could attempt to estimate joint values and examine whether such measures will be adequate for data generated by NGS techniques. We stress, though, that this is a difficult exercise as most data sets are too limited to allow meaningful observations beyond a couple of markers.

- Our qualitative analysis could be extended by trying to address some of the limitations of this work. To account for potential biases in our data set, we might try to seek input from individuals who are more skeptical towards the implementation of NGS techniques by using a different sample population. We may also want to more closely examine differences in perspectives between individuals within different characteristic groups. For example, forensic scientists employed in public versus private, or in large versus smaller laboratories may have different experiences that could provide additional insights not captured in our analysis.

- Another potential interesting avenue for future research, is to assess the impact of sequence data on specific forensic applications. For example, familial DNA searching (FDS) techniques currently perform searches for first-order relationships only due to the limited value of traditional STR profiles. We could examine the performance of sequence data applied to this technique and assess whether such data allow consideration of more distant relatives. It remains to be seen though how the forensic field develops in terms of these applications, especially in light of the emergence of investigative genetic genealogy techniques. Another caveat is that current CODIS databases consist of CE-based DNA

profiles and are thus currently not suitable for the application of sequence-based FDS techniques.

- Further analysis could be carried out with the use of simulated data sets. There are several software packages available that aim to generate sequence-based DNA profiles and some studies have been conducted that use such data in their analyses [47, 67, 85]. A potential concern is that most tools are based on independence assumptions, while we have seen in this dissertation that dependencies occur even for unlinked markers. Future research may try to assess if such simulated profiles provide a realistic representation of real-life NGS profiles by checking whether our results are consistent with results obtained from simulated data.

In the end, we believe it is important to consider the context in which forensic sequence data are to be used. The type of application may greatly impact whether or not some of our observations and recommendations need to be taken into account. While we have argued throughout this dissertation that independence assumptions for genetic data generally do not hold, it may very well be that it works reasonably well in certain settings. There are also situations in which the addition of loci beyond the set of markers incorporated in CODIS DNA profiles does provide valuable information. Requirements are different when your purpose is to generate investigative leads, in which case we may want to focus on reducing false positive hits, or when you are dealing with highly degraded samples during the identification of human remains. We are huge proponents of the use of statistics in the forensic field and commend the community for their continued efforts to incorporate statistical genetic models to enhance DNA evidence evaluations. Nevertheless, the numbers that we attach to DNA profiles can become quite overwhelming and difficult to interpret, not just for the lay person but for any human being. Moreover, while most fields may feel comfortable with methods that work on average, this certainly does not hold in the context of a criminal trial. In this case, it becomes adamant to avoid any tendency to overestimate the weight of the evidence as this may lead to prejudice against a defendant, with potentially grave consequences.

REFERENCES

[1] ATLAS.ti Scientific Software Development GmbH [ATLAS.ti v.9 Windows]. 2022.

[2] S. E. Aalbers, M. J. Hipp, S. R. Kennedy, and B. S. Weir. Analyzing population structure for forensic STR markers in next generation sequencing data. *Forensic Sci. Int. Genet.*, 49, 2020.

[3] M. Airlie, J. Robertson, M. N. Krosch, and E. Brooks. Contemporary issues in forensic science—Worldwide survey results. *Forensic Sci. Int.*, 320, 2021.

[4] M. M. Andersen and D. J. Balding. How convincing is a matching Y-chromosome profile? *PLoS Genet.*, 13, 2017.

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

[6] D. Ballard, J. Winkler-Galicki, and J. Wesoły. Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. *Int. J. Legal Med.*, 134, 2020.

[7] A. Biedermann. Machine learning enthusiasts should stick to the facts. Response to Morrison et al. (2022). *Forensic Sci. Int. Synergy*, 4, 2022.

[8] C. Børsting and N. Morling. Next generation sequencing and its applications in forensic genetics. *Forensic Sci. Int. Genet.*, 18, 2015.

[9] J.-A. Bright, R. Richards, M. Kruijver, H. Kelly, C. McGovern, A. Magee, A. McWhorter, A. Ciecko, B. Peck, C. Baumgartner, C. Buettner, S. McWilliams, C. McKenna, C. Gallacher, B. Mallinder, D. Wright, D. Johnson, D. Catella, E. Lien, C. O'Connor, G. Duncan, J. Bundy, J. Echard, J. Lowe, J. Stewart, K. Corrado, S. Gentile, M. Kaplan, M. Hassler, N. McDonald, P. Hulme, R. H. Oefelein, S. Montpetit, M. Strong, S. Noël, S. Malsom, S. Myers, S. Welti, T. Moretti, T. McMahon, T. Grill, T. Kalafut, M. M. Greer-Ritzheimer, V. Beamer, D. Taylor,

and J. S. Buckleton. Internal validation of STRmix^{TM} – A multi laboratory response to PCAST. *Forensic Sci. Int. Genet.*, 34, 2018.

[10] J. Buckleton, J. Curran, J. Goudet, D. Taylor, A. Thiery, and B. S. Weir. Population-specific Fst values for forensic STR markers: A worldwide survey. *Forensic Sci. Int. Genet.*, 23, 2016.

[11] J. Buckleton and S. Myers. Combining autosomal and y chromosome match probabilities using coalescent theory. *Forensic Sci. Int. Genet.*, 11, 2014.

[12] J. Buckleton, B. Robertson, J. Curran, C. Berger, D. Taylor, J. Bright, T. Hicks, S. Gittelson, I. Evett, S. Pugh, G. Jackson, H. Kelly, T. Kalafut, and F. Bieber. A review of likelihood ratios in forensic science based on a critique of Stiffelman "No longer the Gold standard: Probabilistic genotyping is changing the nature of DNA evidence in criminal trials". *Forensic Sci. Int.*, 310, 2020.

[13] J. M. Butler. Recent advances in forensic biology and forensic DNA typing: INTERPOL review 2019–2022. *Forensic Sci. Int. Synergy*, 6, 2023.

[14] A. Caliebe, A. Jochens, S. Willuweit, L. Roewer, and M. Krawczak. No shortcut solution to the problem of Y-STR match probability calculation. *Forensic Sci. Int. Genet.*, 15, 2015.

[15] K. Cheng, J.-A. Bright, H. Kelly, Y.-Y. Liu, M.-H. Lin, M. Kruijver, D. Taylor, and J. Buckleton. Developmental validation of STRmix^{TM} NGS, a probabilistic genotyping tool for the interpretation of autosomal STRs from forensic profiles generated using NGS. *Forensic Sci. Int. Genet.*, 62, 2023.

[16] X. Dai, Q. Zhu, C. Wang, A. Rukeye, Z. Cao, T. Shan, Y. Wang, and J. Zhang. FST estimates of 94 populations in China based on STR markers. *Forensic Sci. Int. Genet.*, 64, 2023.

[17] L. Davenport, L. Devesse, D. Syndercombe Court, and D. Ballard. Forensic identity SNPs: characterisation of flanking region variation using massively parallel sequencing. *Forensic Sci. Int. Genet.*, 64, 2023.

[18] P. de Knijff. From next generation sequencing to now generation sequencing in forensics. *Forensic Sci. Int. Genet.*, 38, 2019.

[19] L. Devesse, D. Ballard, L. Davenport, I. Riethorst, G. Mason-Buck, and D. Syndercombe Court. Concordance of the ForenSeq^{TM} system and characterisation of sequence-specific autosomal STR alleles across two major population groups. *Forensic Sci. Int. Genet.*, 34, 2018.

[20] P. Donnelly. Nonindependence of matches at different loci in dna profiles: Quantifying the effect of close relatives on the match probability. *Heredity*, 75, 1995.

[21] I. E. Dror and G. Hampikian. Subjectivity and bias in forensic DNA mixture interpretation. *Sci. Justice*, 51, 2011.

[22] I. E. Dror, W. C. Thompson, C. A. Meissner, I. Kornfield, D. Krane, M. Saks, and M. Risinger. Context management toolbox: A linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. *J. Forensic Sci.*, 60, 2015.

[23] H. Eldridge. Juror comprehension of forensic expert testimony: A literature review and gap analysis. *Forensic Sci. Int. Synergy*, 1, 2019.

[24] I. W. Evett and B. S. Weir. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates, Inc., Sunderland, MA, 1998.

[25] L. Excoffier and H. E. L. Lischer. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resources*, 10, 2010.

[26] M. M. Foley and F. Oldoni. A global snapshot of current opinions of next-generation sequencing technologies usage in forensics. *Forensic Sci. Int. Genet.*, 63, 2023.

[27] Forensic Technology Center of Excellence. Landscape study of next generation sequencing technologies for forensic applications, 2023. Available at `https://nij.ojp.gov/library/publications/list`.

[28] S. L. Friis, A. Buchard, E. Rockenbauer, C. Børsting, and N. Morling. Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs. *Forensic Sci. Int. Genet.*, 21, 2016.

[29] K. Gettings, K. Kiesler, S. Faith, E. Montano, C. Baker, B. Young, R. Guerrieri, and P. Vallone. Sequence variation of 22 autosomal STR loci detected by next generation sequencing. *Forensic Sci. Int. Genet.*, 21, 2016.

[30] K. B. Gettings, R. A. Aponte, K. M. Kiesler, and P. M. Vallone. The next dimension in STR sequencing: Polymorphisms in flanking regions and their allelic associations. *Forensic Sci. Int. Genet.*, 5, 2015.

[31] K. B. Gettings, D. Ballard, M. Bodner, L. A. Borsuk, J. King, W. Parson, and C. Phillips. Report from the STRAND Working Group on the 2019 STR Sequence Nomenclature Meeting. *Forensic Sci. Int. Genet.*, 43, 2019.

[32] K. B. Gettings, L. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, J. L. King, W. Parson, C. Phillips, and P. M. Vallone. STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci. *Forensic Sci. Int. Genet.*, 31, 2017.

[33] K. B. Gettings, L. A. Borsuk, C. R. Steffen, K. M. Kiesler, and P. M. Vallone. Sequence-based U.S. population data for 27 autosomal STR loci. *Forensic Sci. Int. Genet.*, 37, 2018.

[34] P. Gill, T. Hicks, J. M. Butler, E. Connolly, L. Gusmão, B. Kokshoorn, N. Morling, R. A. van Oorschot, W. Parson, M. Prinz, P. M. Schneider, T. Sijen, and D. Taylor. DNA commission of the International society for forensic genetics: Assessing the value of forensic biological evidence - Guidelines highlighting the importance of propositions. Part II: Evaluation of biological traces considering activity level propositio. *Forensic Sci. Int. Genet.*, 44, 2020.

[35] J. Goudet. Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, 5, 2005.

[36] A. Gouy and M. Zieger. STRAF - A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci. Int. Genet.*, 30, 2017.

[37] T. O. Hall. The Y-Chromosome in Forensic and Public Health Genetics. 2016. Available at Seattle: University of Washington Libraries.

[38] M. M. Hennink, B. N. Kaiser, and V. C. Marconi. Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough? *Qualitative Health Research*, 27, 2017.

[39] W. G. Hill and B. S. Weir. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genet. Res.*, 93, 2011.

[40] J. Hoogenboom, K. J. van der Gaag, R. H. de Leeuw, T. Sijen, P. de Knijff, and J. F. Laros. FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise. *Forensic Sci. Int. Genet.*, 27, 2017.

[41] Verogen. An introduction to Next-Generation Sequencing Technology. 2015. Available at `https://www.illumina.com/science/technology/next-generation-sequencing.html`.

[42] Verogen. ForenSeq^{TM} Universal Analysis Software Guide. 2016. Available at `https://verogen.com/products/forenseq-dna-signature-prep-kit/`.

[43] A. M. Jeanguenat, B. Budowle, and I. E. Dror. Strengthening forensic DNA decision making through a better understanding of the influence of cognitive bias. *Sci. Justice*, 57, 2017.

[44] K. M. Kiesler, L. A. Borsuk, C. R. Steffen, K. B. Gettings, and P. M. Vallone. Population data for 94 identity SNPs in four U.S. population groups. *Forensic Sci. Int. Genet. Supplement Series*, 8, 2022.

[45] S. Köcher, P. Müller, B. Berger, M. Bodner, W. Parson, L. Roewer, S. Willuweit, and T. D. Consortium. Inter-laboratory validation study of the ForenSeq^{TM} DNA Signature Prep Kit. *Forensic Sci. Int. Genet.*, 36, 2018.

[46] C. Laurie and B. S. Weir. Dependency effects in multi-locus match probabilities. *Theoretical Population Biology*, 63, 2003.

[47] S. M. Leal, K. Yan, and B. Muller-Myhsok. SimPed: A simulation program to generate haplotype and genotype data for pedigree structures. *Hum. Hered.*, 60, 2005.

[48] Y.-Y. Liu and S. Harbison. A review of bioinformatic methods for forensic DNA analyses. *Forensic Sci. Int. Genet.*, 33, 2018.

[49] M. Maguire and B. Delahunt. Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *AISHE-J*, 8, 2017.

[50] G. S. Morrison, N. Basu, E. Enzinger, and P. Weber. The opacity myth: A response to Swofford & Champod (2022). *Forensic Sci. Int. Synergy*, 5, 2022.

[51] G. S. Morrison, D. Ramos, R. J. Ypma, N. Basu, K. de Bie, E. Enzinger, Z. Geradts, D. Meuwly, D. van der Vloed, P. Vergeer, and P. Weber. A strawman with machine learning for a brain: A response to Biedermann (2022) the strange persistence of (source) "identification" claims in forensic literature. *Forensic Sci. Int. Synergy*, 4, 2022.

[52] National Research Council. The Evaluation of Forensic DNA Evidence. The National Academies Press, Washington, DC. 1996.

[53] C. Neumann, D. Kaye, G. Jackson, V. Reyna, and A. Ranadive. Presenting Quantitative and Qualitative Information on Forensic Science Evidence in the Courtroom. *Chance*, 29, 2016.

[54] N. Novroski, F. Wendt, A. Woerner, M. Bus, M. Coble, and B. Budowle. Expanding beyond the current core STR loci: An exploration of 73 STR markers with increased diversity for enhanced DNA mixture deconvolution. *Forensic Sci. Int. Genet.*, 38, 2019.

[55] N. M. Novroski, J. L. King, J. D. Churchill, L. H. Seah, and B. Budowle. Characterization of genetic sequence variation of 58 STR loci in four major population groups. *Forensic Sci. Int. Genet.*, 25, 2016.

[56] A. J. Pakstis, W. C. Speed, J. R. Kidd, and K. K. Kidd. Candidate SNPs for a universal individual identification panel. *Human Genetics*, 121, 2007.

[57] W. Parson, D. Ballard, B. Budowle, J. M. Butler, K. B. Gettings, P. Gill, L. Gusmão, D. R. Hares, J. A. Irwin, J. L. King, P. de Knijff, N. Morling, M. Prinz, P. M. Schneider, C. Van Neste, S. Willuweit, and C. Phillips. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Sci. Int. Genet.*, 22, 2016.

[58] C. Phillips, K. B. Gettings, J. L. King, D. Ballard, M. Bodner, L. Borsuk, and W. Parson. "The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide. *Forensic Sci. Int. Genet.*, 34, 2018.

[59] M. J. Saks and J. J. Koehler. The individualization fallacy in forensic science evidence. *Vanderbilt Law Review*, 61, 2008.

[60] Scientific Working Group on DNA Analysis Methods (SWGDAM). Addendum to SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories to Address Next Generation Sequencing, 2019. Available at `https://www.swgdam.org/publications`.

[61] S. Siegert, L. Roewer, and M. Nothnagel. Shannon's equivocation for forensic Y-STR marker selection. *Forensic Sci. Int. Genet.*, 16, 2015.

[62] C. D. Steele, D. S. Court, and D. J. Balding. Worldwide FST Estimates Relative to Five Continental-Scale Populations. *Annals of Human Genetics*, 78, 2014.

[63] C. R. Steffen, T. I. Huszar, L. A. Borsuk, P. M. Vallone, and K. B. Gettings. A multi-dimensional evaluation of the 'NIST 1032' sample set across four forensic Y-STR multiplexes. *Forensic Sci. Int. Genet.*, 57, 2022.

[64] B. Stiffelman. No longer the gold standard: probabilistic genotyping is changing the nature of DNA evidence in criminal trials. *Berkeley J. Crim. L.*, 24, 2019.

[65] H. Swofford and C. Champod. Machine learning algorithms in forensic science: A response to Morrison et al. (2022). *Forensic Sci. Int. Synergy*, 5, 2022.

[66] H. Swofford and C. Champod. Probabilistic reporting and algorithms in forensic science: Stakeholder perspectives within the American criminal justice system. *Forensic Sci. Int. Synergy*, 4, 2022.

[67] R. Tao, Q. Xu, S. Wang, R. Xia, Q. Yang, A. Chen, Y. Qu, Y. Lv, S. Zhang, and C. Li. Pairwise kinship analysis of 17 pedigrees using massively parallel sequencing. *Forensic Sci. Int. Genet.*, 57, 2022.

[68] D. Taylor, B. Kokshoorn, and A. Biedermann. Evaluation of forensic genetics findings given activity level propositions: A review. *Forensic Sci. Int. Genet.*, 36, 2018.

[69] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526, 2015.

[70] E. A. Thompson. Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, 194, 2013.

[71] W. C. Thompson and E. J. Newman. Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents. *Law and Human Behavior*, 39, 2015.

[72] A. O. Tillmar and C. Phillips. Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets. *Forensic Sci. Int. Genet.*, 26, 2017.

[73] S. B. Vilsen, T. Tvedebrink, P. S. Eriksen, C. Hussing, C. Børsting, and N. Morling. Modelling allelic drop-outs in STR sequencing data generated by MPS. *Forensic Sci. Int. Genet.*, 37, 2018.

[74] B. Walsh, A. J. Redd, and M. F. Hammer. Joint match probabilities for Y chromosomal and autosomal markers. *Forensic Sci. Int.*, 174, 2008.

[75] B. S. Weir. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data.* Sinauer Associates, Inc., Sunderland, MA, 1996.

[76] B. S. Weir. The effects of inbreeding on forensic calculations. *Annu. Rev. Genet.*, 28, 1994.

[77] B. S. Weir. Matching and partially-matching DNA profiles. *J. Forensic Sci.*, 49, sep 2004.

[78] B. S. Weir, L. R. Cardon, A. D. Anderson, D. M. Nielsen, and W. G. Hill. Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, 15, 2005.

[79] B. S. Weir and C. C. Cockerham. Estimating F-Statistics for the analysis of population structure. *Evolution*, 38, 1984.

[80] B. S. Weir and J. Goudet. A unified characterization of population structure and relatedness. *Genetics*, 206, 2017.

[81] B. S. Weir and W. G. Hill. Estimating F-Statistics. *Annu. Rev. Genet.*, 36, 2002.

[82] S. Wright. The genetical structure of populations. *Ann. Eugen.*, 15, 1951.

[83] Y. J. Yang, M. Prinz, H. McKiernan, and F. Oldoni. American forensic DNA practitioners' opinion on activity level evaluative reporting. *J. Forensic Sci.*, 67, 2022.

[84] B. A. Young, J. L. King, B. Budowle, and L. Armogida. A technique for setting analytical thresholds in massively parallel sequencing-based forensic DNA analysis. *PLoS ONE*, 12, 2017.

[85] Q. Zhang, X. Wang, P. Cheng, S. Yang, W. Li, Z. Zhou, and S. Wang. Complex kinship analysis with a combination of STRs, SNPs, and indels. *Forensic Sci. Int. Genet.*, 2022.

[86] Q. S. Zhang, J. Goudet, and B. S. Weir. Rank-invariant estimation of inbreeding coefficients. *Heredity*, 128, 2021.

[87] X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28, 2012.