The Effects of True Match Rate, Imputation Accuracy, and Population Structure on a Method for Genetic Record Matching

Jacob Alfieri

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

Master of Science

University of Washington

2021

Reading Committee:

Bruce Weir, Chair

Sharon Browning

Program Authorized to Offer Degree:

Biostatistics - Public Health

University of Washington

Abstract

The Effects of True Match Rate, Imputation Accuracy, and Population Structure on a Method for Genetic Record Matching

Jacob Alfieri

Chair of the Supervisory Committee:

Professor Bruce Weir

Department of Biostatistics

As the number and size of genetic databases continues to expand, linking between them will become increasingly important, especially in forensic contexts. Edge et al. have proposed a best-in-class method for matching profiles between an STR sequenced and a SNP sequenced database. Through the use of additional genetic profiles and newly available whole genome sequencing data, we assess how this method is affected by the true match rate between the two databases, imputation accuracy, and population structure. We find that matching accuracy decreases monotonically as the true match rate decreases. Greater imputation accuracy from whole genome sequenced profiles enables substantially higher matching accuracy. Finally, accounting for population structure can modestly increase matching accuracy, but must be done carefully.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank the many people that have supported me during my graduate studies and the work on this thesis. Their encouragement and guidance has been invaluable during this most unusual year.

First, I would like to thank Bruce Weir for fostering my interest in forensic genetics and advising me as I worked through my thesis. Thank you to Sharon Browning for her technical guidance and useful comments. I would also like to thank the many other graduate students in the UW Departments of Biostatistics and Statistics for their help and camaraderie in our classes together.

I would like to thank my friends for their support. Thank you to Emma Nuelle for always reminding me to keep things in perspective. Thank you to Nina for always making me laugh.

Lastly and most importantly, I would like to thank my family. Thank you to my Mom for making sure that I took breaks from working to watch TV with her. Thank you to my Dad for always reminding me to get back to work. Thank you to Patrick, Samantha and my Grandparents for their boundless love and faith in me.

Chapter 1

INTRODUCTION

Advances in genome-sequencing and computing have reduced the cost of sequencing a single human genome from approximately $14 million in 2006 to just $1500 in 2015 [28]. This has enabled the creation of large DNA databases for a variety of purposes. The 1000 Genomes Project (1kGP) and Human Genome Diversity Project (HGDP) have been created for research. Many countries have created national forensic DNA databases including the United States, the United Kingdom and China. Private companies including Ancestry.com and 23andMe have created DNA databases for genealogy and predictive medicine. These different databases use a variety of sequencing approaches, such as whole genome sequencing at 1kGP, SNP arrays at 23andMe and short tandem repeats (STR) in the US and UK forensic databases. As these genetic databases proliferate, it will become useful to combine them in certain contexts, including situations where the individuals have been sequenced using different technologies.

Linking genetic data sets across sequencing technologies is already well established in the context of SNP arrays. Arrays produced by different companies contain different SNPs, which can make it difficult to compare them directly. However, imputation methods have been developed that enable meta-analyses across data sets from different types of SNP arrays. Many of these methods use the Li and Stephens model, which takes advantage of the natural patterns of linkage disequilibrium on the genome [10].

Combining STR data sets with SNP arrays or whole genome sequencing data sets presents additional challenges. Mutation rates at STR loci are on average several orders of magnitude

higher than for SNPs [12]. As a result, SNP-STR linkage disequilibrium is much weaker than SNP-SNP linkage disequilibrium [25]. This makes it much harder to impute STRs from SNP data, than it is to impute additional SNPs.

1.1 Forensic Databases

STRs are especially important in forensic applications. Forensic DNA samples collected from crime scenes are often degraded by environmental exposure. Many STRs have widely variable lengths, which can make it possible to analyze DNA samples despite partial degradation [23]. STR loci can have many alleles, which makes analysis of DNA samples containing mixtures of individuals possible [7]. This polymorphism also enables high discriminatory power with a small number of STRs. As a result, more than 60 countries have created forensic DNA databases using STR profiles [30].

In 1998 the FBI, in collaboration with nine states, launched the National DNA Index System (NDIS), the first national DNA database for law enforcement. As of October 2020, the NDIS included profiles from more than more than 18 million individuals who have been convicted or arrested [26]. All 50 states plus the District of Columbia and Puerto Rico have subsequently established their own DNA databases for law enforcement, referred to as State DNA Index Systems (SDIS), which contain additional profiles. The FBI maintains the Combined DNA Index System (CODIS), which allows law enforcement to search these databases.

During the creation of the NDIS, the FBI established a standard of 13 core CODIS loci, STR that were required for each profile [17]. On January 1, 2017 the core CODIS loci were expanded to 20 loci. Among other reasons, these loci were chosen because there is a low level of linkage disequilibrium between them and they provide a high level of discrimination.

While STRs are likely to remain the dominant technology in forensic applications, the use of whole genome sequencing and SNP array data is growing. In 2018 police were finally able to arrest the Golden State Killer after decades of unsuccessful investigations using a commercial genealogical SNP array database, GEDmatch [15]. In 2019, GEDmatch was

purchased by Verogen, which plans to launch a new version of the database specifically for law enforcement, using their 1.2 million customer SNP array profiles [27].

As these techniques become more widespread, linking records across databases will be important. As described, most forensic databases use STR profiles, while most other databases use SNP profiles from SNP arrays, whole genome sequencing or whole exome sequencing. In some cases, such as with the Golden State Killer, it may be possible to re-sequence the original evidence using different technologies. However, that is not always the case as original evidence may be lost or discarded over time. Degraded DNA samples or mixtures can be challenging to sequence for SNPs. Finally, re-sequencing old evidence imposes new costs.

The possibility of linking forensic STR profiles with SNP profiles also raise new concerns. When the number of CODIS core loci was expanded from 13 to 20, the additional loci were selected in part for having no known association with any medical condition [17]. If STR profiles can be linked with SNP profiles, it may be possible to determine some genetic conditions and risk factors. Individuals may not be aware that their SNP profile DNA is being used for law enforcement purposes. Customers may intend for their profiles to be used for genealogical or predictive health and not read fine print print that details uses in other contexts.

Chapter 2

GENETIC RECORD LINKING

2.1 Edge et al Method

Edge et al. proposed a method for genetic record linkage between STR profiles and SNP profiles [11]. This method relies on the Beagle software package which we will first briefly describe [5].

Beagle enables imputation of genetic markers using a reference panel of phased individuals who are typed at the desired markers [5]. Like many other genotype imputation methods, Beagle is based on the Li and Stephens framework which uses a hidden Markov model (HMM) [10] [22]. Many of the different imputation methods built using this framework provide similarly high levels of accuracy, but Beagle requires less computation time [5]. Edge et al.'s method requires a training set of individuals typed at both the SNP and STR locations in each set of records in order to perform the necessary imputation using Beagle. This is shown visually in Figure 2.1. Additionally, Edge et al. used Beagle 4.1. Since their publication, an updated version, Beagle 5.0, has been released, which achieves effectively identical results, but requires less computation time.

2.1.1 Edge et al. Score Details

In order to describe Edge et al.'s proposed method, we first define some useful quantities. We have two sets of genetic records, R and S. Let $R = \{R_1, ..., R_n\}$ be a set of n STR profiles, where each individual is unique and unrelated. Each profile contains L STR loci. Then R_{il} is the diploid STR genotype for individual i at locus l. Let $S = \{S_1, ..., S_m\}$ be a

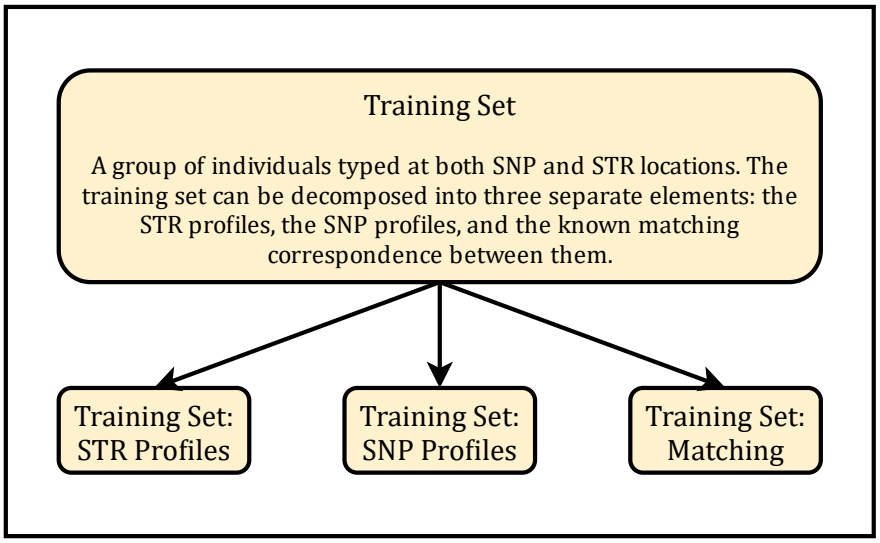


Figure 2.1: Edge et al. Training Set

Visual depiction of training set used in Edge et al. method for Beagle imputation. The color yellow corresponds to the training set. This figure primarily serves to introduce these elements that will be used in later figures describing the full Edge et. method.

set of m SNP profiles, where each individual is unique and unrelated. Then S_{jl} is the diploid window of SNPs near the STR locus l for individual j. The two sets of genetic records, R and S, together form the test set and are shown visually in Figure 2.2.

The motivation for creating a window of SNPs near each STR will be further explained later, but the general idea is that the SNPs near an STR provide the most information about the STR due to linkage disequilibrium. Effectively the SNP profiles are broken into L seperate chunks corresponding to each STR of interest. Let M be an indicator variable, where $M = 0$ indicates that two records are not from the same person while $M = 1$ indicates that two records are from the same person.

Fellegi and Sunter proposed a general framework for record matching, where match scores comparing each pair of records are log-likelihood ratios of the probabilities of observing the

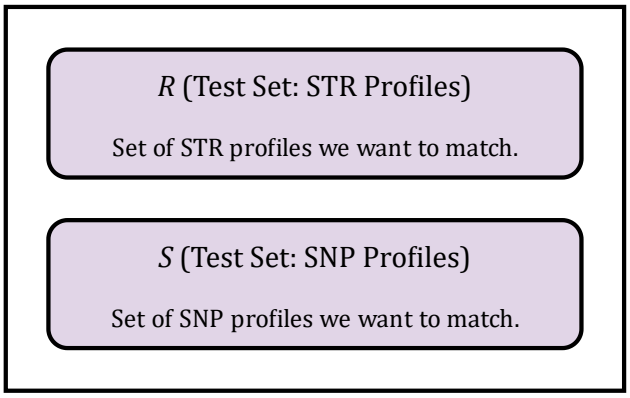


Figure 2.2: Edge et al. Testing Set

Visual depiction of testing set used in Edge et al. method for Beagle imputation. The color purple corresponds to the test set. This figure primarily serves to introduce these elements that will be used in later figures describing the full Edge et. method.

two records given they are from the same entity versus the probability of observing the two records given they are from different entities [13]. Using that framework, Edge et al. proposed a match score $\lambda(R_i, S_j)$, shown in Equation 2.1. It compares a STR profile, R_i, to a SNP profile, S_j.

$$\lambda(R_i, S_j) = \ln\left[\frac{P(R_i, S_j|M=1)}{P(R_i, S_j|M=0)}\right] \tag{2.1}$$

Edge et al. then simplify this inner quantity using the rules of conditional probability. They assume that if $M = 0$, the probabilities of R_i and S_j are independent. This results in the expression shown in Equation 2.4

$$\frac{P(R_i, S_j|M=1)}{P(R_i, S_j|M=0)} = \frac{P(R_i|S_j, M=1)}{P(R_i|S_j, M=0)}\frac{P(S_j|M=1)}{P(S_j|M=0)} \tag{2.2}$$

$$= \frac{P(R_i|S_j, M=1)}{P(R_i|M=0)}\frac{P(S_j)}{P(S_j)} \tag{2.3}$$

$$= \frac{P(R_i|S_j, M=1)}{P(R_i)} \tag{2.4}$$

Thus the match score is equivalent to Equation 2.7.

$$\lambda(R_i, S_j) = \ln\left[\frac{P(R_i, S_j|M=1)}{P(R_i, S_j|M=0)}\right] \tag{2.5}$$

$$= \ln\left[\frac{P(R_i|S_j, M=1)}{P(R_i)}\right] \tag{2.6}$$

$$= \ln[P(R_i|S_j, M=1)] - \ln[P(R_i)] \tag{2.7}$$

Edge et al. further assume that that the genotypes at each STR locus are independent. This assumption of independence for CODIS loci is common in forensic applications [14]. The CODIS loci were intentionally selected to be far apart from each other on the genome to minimize dependence. However, they are not actually independent. This assumption simplifies the calculation of the match score, because we can simply multiply the probability for each STR locus to find the probability for the entire profile. Nevertheless, if the databases contain individuals from multiple populations, these assumptions will be badly violated, which we will further discuss in Section 2.3. This is shown in Equations 2.8 and 2.9

$$P(R_i|S_j, M=1) = \prod_{l=1}^{L} P(R_{il}|S_{jl}, M=1) \tag{2.8}$$

$$P(R_i) = \prod_{l=1}^{L} P(R_{il}) \tag{2.9}$$

Let us consider what these equations represent. The expression, $P(R_{il}|S_{jl}, M=1)$, is the probability of observing the STR genotype R_{il} given the SNP genotype for a window around locus l from SNP profile S_j, assuming that they are from the same individual. The expression, $P(R_i)$, is the frequency of observing a given STR profile.

The completely simplified match score is shown in Equation

$$\lambda(R_i, S_j) = \ln[P(R_i|S_j, M = 1)] - \ln[P(R_i)] \tag{2.10}$$

$$= \ln\left[\prod_{l=1}^{L} P(R_{il}|S_{jl}, M = 1)\right] - \ln\left[\prod_{l=1}^{L} P(R_{il})\right] \tag{2.11}$$

$$= \sum_{l=1}^{L} (\ln[P(R_{il}|S_{jl}, M = 1)] - \ln[P(R_{il})]) \tag{2.12}$$

2.1.2 Edge et al. Method Steps

Now that we have defined the match score, we can describe Edge et al.'s start to finish method for record linkage.

Step 1 (also shown visually in Figure 2.3):

(a) We need a separate, non-overlapping training set of individuals with both SNP and STR profiles for imputation. This training set is phased by Beagle in preparation for imputation.

(b) Using Beagle and the phased training set, we probabilistically impute STR profiles for each SNP profile in the test set. Then the output for each SNP profile in the test set, S_j, is a vector of probabilities for each STR locus l, corresponding to the imputation probability for each potential diplotype at that locus.

(c) We calculate the proportion of each allele at each STR locus from entire training STR set.

Step 2 (shown in Figure 2.4): We calculate the match score $\lambda(R_i, S_j)$ for each pair R_i and S_j, where $i \in \{1, ..., n\}$ and $j \in \{1, ..., m\}$. This results in an $n \times m$ matrix of match scores. In order to calculate the match score for each, we estimate both $P(R_{il}|S_{jl}, M = 1)$ and $P(R_{il})$ for each locus l.

(a) We estimate $P(R_{il}|S_{jl}, M = 1)$ using the imputation probability of R_{il} given S_{jl} from Beagle in Step 1b. For example, assume the locus l is TH01 and $R_{il} = \{7, 9.3\}$. Then

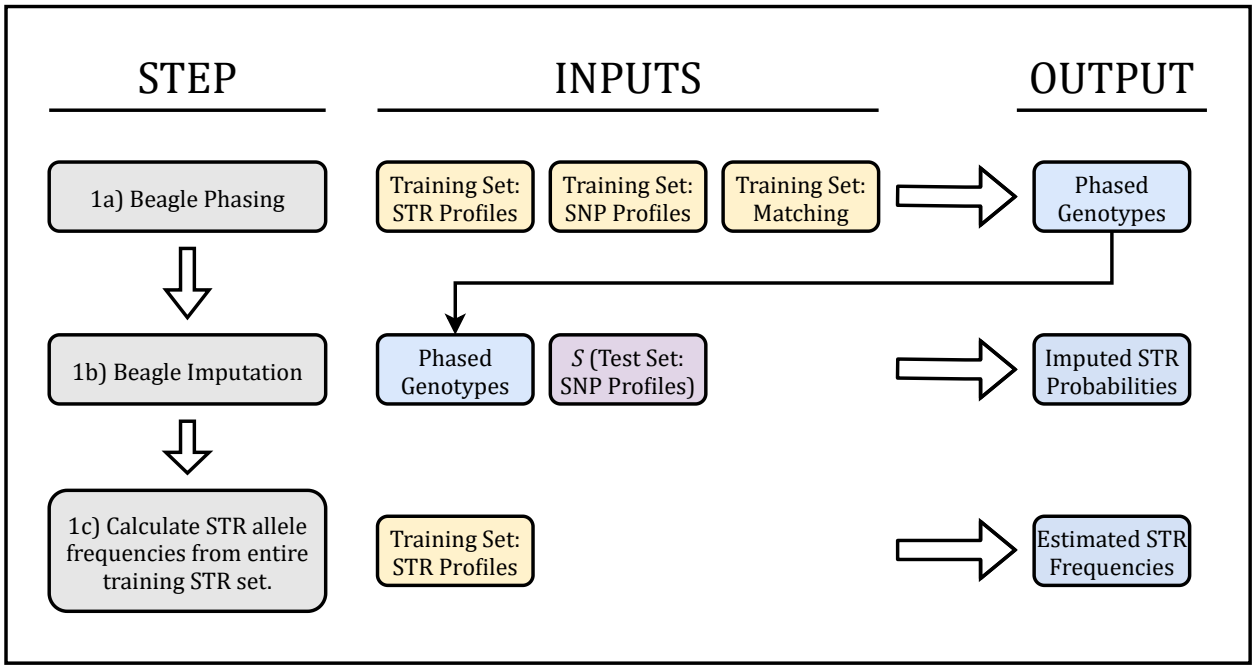


Figure 2.3: Edge et al. Method Step 1

Flowchart showing Step 1 in Edge et al.'s method. Grey boxes are the sub-steps. Blue boxes are intermediate outputs. Yellow boxes are inputs from the training set (see Figure 2.1) and purple boxes are inputs from the test set (see Figure 2.2).

$P(R_{il}|S_{jl}, M = 1)$ is the Beagle calculated imputation probability of $\{7, 9.3\}$ from the window of SNPs surrounding the TH01 locus in the S_j profile.

(b) We estimate, $P(R_{il})$, the genotype frequency, using the allele frequencies from Step 1c and assuming Hardy-Weinberg equilibrium.

Step 3 (shown in Figure 2.4): We apply the Hungarian algorithm to the score matrix to determine estimated matching assignments between the two sets of genetic records.

2.2 Testing Limitations

Edge et al. test the performance of their model using profiles from the HGDP. The HGDP contains 1064 cultured lymphoblastoid cell lines (LCL) from 51 different populations from

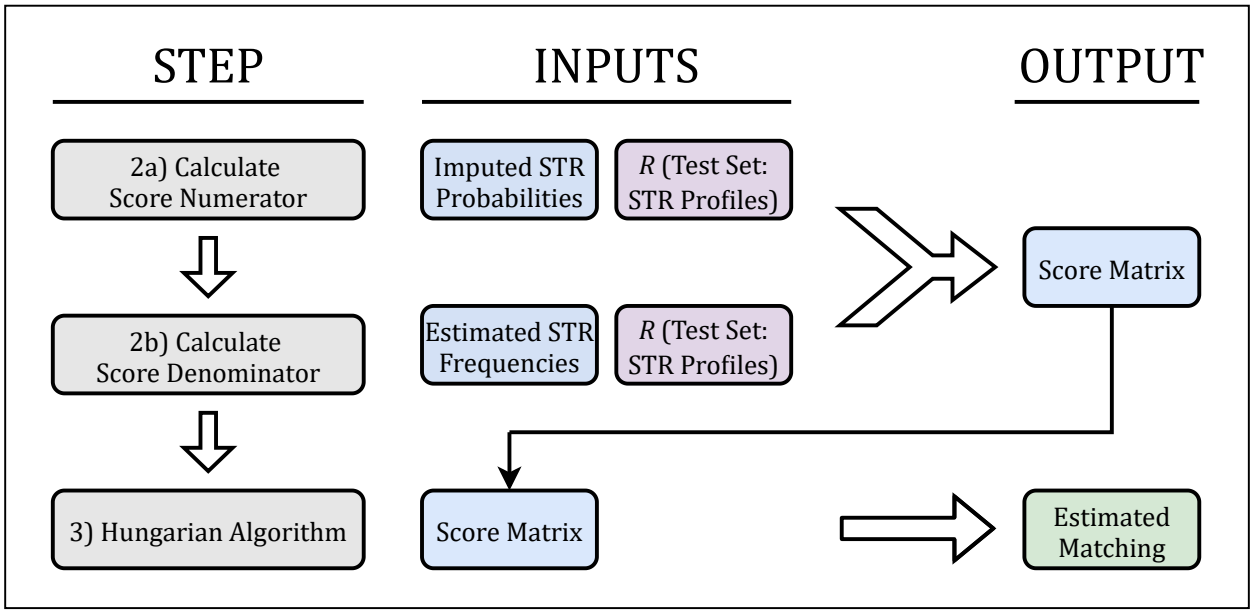


Figure 2.4: Edge et al. Method Steps 2 and 3

Flowchart showing Steps 2 and 3 in Edge et al.'s method. Grey boxes are the sub-steps. Blue boxes are intermediate outputs. Purple boxes are inputs from the test set (see Figure 2.2). The green box is the final output: estimated matching assignments between the two sets.

across the world [8]. The DNA from these LCLs are available to researchers who agree to sequence them and make their results publically available. Since the project was launched in 2002, the samples have been sequenced on a variety of technologies, including hundreds of STRs, various SNP arrays, and most recently, whole genome sequencing. The HGDP samples contain a small number of close relatives and some abnormalities, including apparent atypical and duplicate samples [24]. Since sequencing of the samples is done by independent researchers, they have used different exclusion and quality control standards. Edge et al. use data from two efforts to sequence HGDP individuals. One sequenced 938 unrelated HGDP individuals using Illumina HumanHap650K Beadchips and the other sequenced the CODIS STR loci of 1048 HGDP individuals. There are 872 individuals in the common subset of these two efforts, meaning that 872 HGDP individuals have both Illumina HumanHap650K

Beadchips sequences and CODIS STR sequences. Edge et al. use these 872 individuals to test their method [11].

They consider three different scenarios, which they designate one-to-one, one-to-many and needle-in-haystack. For the one-to-one scenario, they assume that the set of STR records and SNP records are the same size and contain identical individuals. Thus the record matching becomes an assignment problem, so they apply the Hungarian algorithm to the score matrix to determine pairings. For the one-to-many scenario, they assume that one set of records has one individual and other other a larger number, where they determine pairing by selecting the maximum score. Finally, for the needle-in-haystack, they assume that the set of STR records and SNP records contain only one true match and consider success to be perfect separation between true match and true non-match scores. Edge et al. found that their method performs well on the one-to-one and one-to-many scenarios and much worse in the needle-in-haystack scenarios.

Scenario	Median Accuracy
One-to-one	.982
One-to-many: SNP query	.913
One-to-many: STR query	.899
Needle-in-haystack	.450

Table 2.1: Median Accuracy Using 13 CODIS Loci

However, these test scenarios have limitations. The greatest accuracy was achieved in the one-to-one scenario, but the assumptions for this scenario are the least realistic. We know a priori that every profile in each set has exactly one match in the other. It is highly unlikely that any application will arise where two data sets containing the exact same set of individuals will need to be matched together. However, this is the only scenario that considers the useful possibility of multiple matches occurring in the same sets. The one-to-

many and the needle-in-haystack scenarios do not assume that the two sets are the same, which is more realistic. Yet, they do not consider situations with multiple matches.

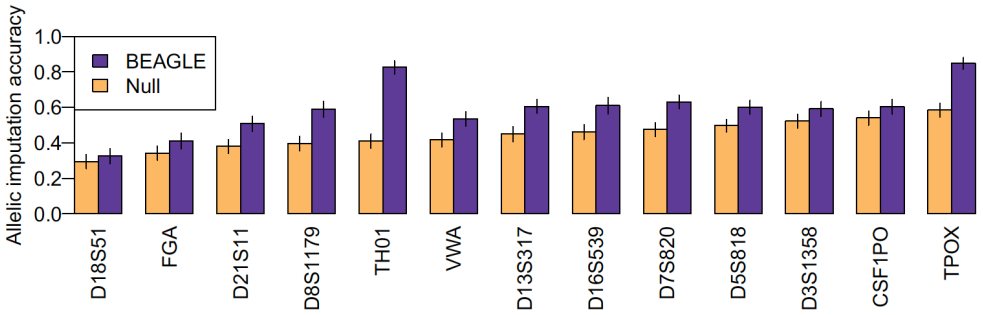


Figure 2.5: Allelic Imputation Accuracy for 13 CODIS Loci

Fraction of STR alleles correctly imputed using HGDP SNP array profiles by Edge et al. The 872 HGDP individuals were split 75% in training set and 25% in test set. Null imputation was done using the STR allele frequencies in the training set. (Note: This is Figure 1 in Edge et al.)

Additionally, Edge et al achieve a low level of accuracy in the STR imputation phase of their algorithm. This can be seen in Figure 2.5. For many of the CODIS loci, Beagle imputation performs only marginally better than null imputation using Hardy-Weinberg frequencies. This makes sense given the limitations of the data set used to test the model. Das et al. identified five factors that affect imputation accuracy including size of reference panel and density of genotyping array [10]. Edge et al.'s data contains 872 individuals, but only a subset of them are used as a reference panel for imputation, because the rest are used for testing. This results in a small sample size. The Illumina HumanHap650K Beadchips are a low density genotyping array. Additionally, Das et al. suggest that reference panels "with little genetic similarity to the target panel ... may decrease imputation accuracy." The HGDP includes individuals from 52 different populations, so an imputation reference panel will inherently have a very small number of individuals from the same population as the

individual being imputed.

2.3 Population Structure

Edge et al. do not take into account the population structure of the genetic databases. All individuals in the training and test sets are effectively considered to be part of a single combined population. This assumption may be satisfied for some applications, such as national databases where individuals are generally drawn from the same population. Edge et al. use this single population assumption to further assume that each CODIS STR is independent and that each STR is in Hardy-Weinburg equilibrium while estimating the probability of observing a given STR profile as part of calculating the match score. With population structure, these assumptions will be violated. Each population may have different allele frequencies and there may be dependencies between alleles due to shared evolutionary history [1]. This could result in poor estimates for the likelihood of observing a given STR profile, which could in turn have undesired effects on the overall matching procedure.

Chapter 3

METHODS

3.1 Effect of True Match Rate

We assessed the effect of the true match rate on the accuracy of Edge et al.'s method, by introducing additional non-matching STR profiles. The National Institute of Standards and Technology (NIST) has created a publically available database of 1036 unrelated Americans sequenced at 29 autosomal STR loci, including the 13 original CODIS loci [18]. The samples were collected anonymously from three private companies in Florida, Tennessee and Ohio. We excluded two profiles which are missing genotypes at one STR each. We chose to use profiles from NIST because it one of the only publicly available data sets including STR sequenced diplotypes.

In the testing performed by Edge et al., they achieved the highest accuracy level using 7/8 (763) of the HGDP individuals as a reference set for STR imputation and then using the remaining 1/8 (109) of the HGDP individuals as a testing set. We use this as a baseline and then consider scenarios adding 25% (259), 50% (517), 75% (776) and 100% (1034) of the NIST STR profiles to the test set.

In Figure 2.2, we visually depicted the two input sets of genetic records to be matched. Since we are performing a simulation study here, we have an additional piece of information: the true matching correspondence between the two sets. This is shown visually in Figure 3.1. In addition to the three steps described in Subsection 2.1.2 and shown visually in Figures 2.3 and 2.4, we add a fourth step for our testing: calculating the accuracy of our estimated matches, shown in Figure 3.2. Since the true matching correspondence between the two

sets is the answer the question we are attempting to solve with our matching approach, this information is only used to check the accuracy of the estimated matching assignments.

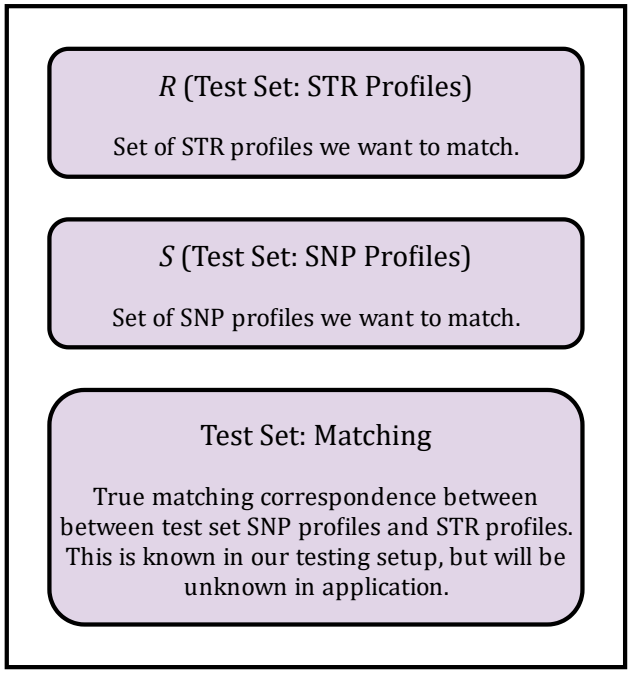


Figure 3.1: Simulation Test Set Inputs

Visual depiction of test set used in our simulation studies. The Test Set: STR Profiles, R, and Test Set: SNP Profiles, S, are analogous to normal input. The Test Set: Matching is only used to check the accuracy of our estimated assignments and not available in application.

To facilitate comparison with Edge et al., we used similar settings for testing. For the Beagle imputation stage, we used 1Mbp windows centered at each STR. The HGDP SNP sequences were mapped to the NCBI36 reference assembly. Positions for the CODIS STR loci were found using the University of California-Santa Cruz Genome Browser [19]. We use the most recent Beagle 5.0 with default settings. For each setting, we considered 100 random partitions of the HGDP samples into test and training and corresponding random selections of the NIST profiles to include in the test set.

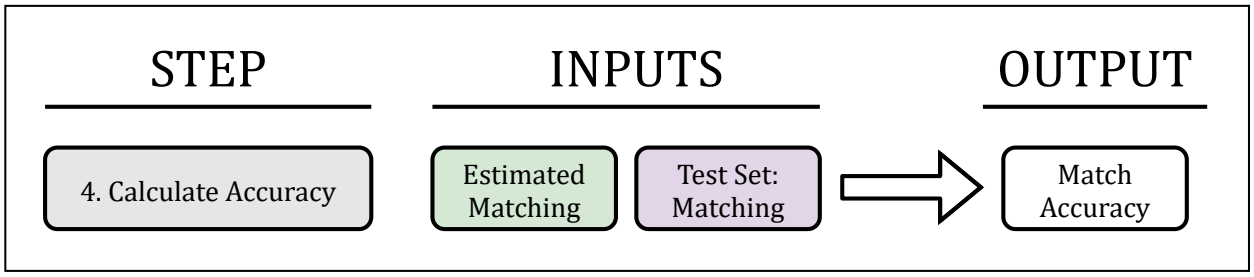


Figure 3.2: Step 4: Checking Matching Accuracy

Visual depiction of calculating matching accuracy in the same format at Edge et al. Steps 1, 2. and 3 shown above. Grey boxes are the sub-steps. Purple boxes are inputs from the test set (see Figure 3.1). The green box is the final output: estimated matching assignments between the two sets.

3.2 Effect of Imputation Accuracy

We assessed the effect of imputation accuracy on Edge et al.'s method using new data. In 2020, whole genome sequencing was performed on 929 individuals from the HGDP data set [3]. Illumina sequencing was performed with average coverage of 35x and a read length of 151bp. This gives a far greater density than the Illumina HumanHap650K Beadchips originally used by Edge et al.

Not all of the HGDP individuals sequenced using the SNP arrays were subsequently whole genome sequenced. We found the intersection of the set of unrelated individuals with whole genome sequences and typed at the CODIS loci, resulting in 865 individuals. This is 7 individuals less than the corresponding intersection for the SNP arrays. Since that comprises less than 1% of the profiles, any effects from the different sample size are likely small.

We used a similar testing procedure as the previous scenarios. We used the same number of HGDP individuals for testing, 109. We use the remaining 756 HGDP individuals as a reference set. We added the same numbers of non-matching NIST STR profiles to the test set. For imputation, we used 4Mbp windows centered at each STR. The HGDP whole

genome sequences were mapped to the GRCh38 reference assembly. Positions for the CODIS STR loci were found using the University of California-Santa Cruz Genome Browser [19].

3.3 Effect of Population Structure

We explored the effects of population structure on the calculation of the match score by comparing Edge et al.'s method with several modifications that attempt to incorporate knowledge of the population structure of the samples contained in the two databases being matched. We will focus on how that effects the denominator of the match score (Equation 2.1), simplified in Equation 2.4 as $P(R_i)$. This is also Step 1c, in the step by step breakdown. Recall, that this is the probability of observing a given STR profile, R_i. Edge et al. assume that all of the individuals in the testing and training sets are effectively sampled from a single population where the STR allele frequencies are in Hardy-Weinberg equilibrium. They set $P(R_i) = \prod_{l=1}^{L} P(R_{il})$ and use the training set STR allele frequencies to calculate $P(R_{il})$.

We modified this by calculating the STR allele frequencies by population. Let us first describe the structure of the HGDP and NIST data sets. The HGDP comprises samples from 51 separate populations, 13 of which have less than 10 individuals. It is not feasible to analyze such small samples, so instead we analyzed the profiles at the region level. The HGDP profiles are grouped into 7 regions, as shown in Table 3.1.

While on average, the HGDP contains samples from approximately 124 individuals in each region, there are less than 40 profiles sampled from both the Oceania and America regions. The NIST database contains 1034 usable profiles sampled from four populations, as shown in Table 3.2.

We will consider three modifications that take this structure into account. They are described in depth below and shown visually in Figure 3.3. For these modifications, populations refers to both the HGDP regions and NIST populations.

Region	Number of Individuals
Africa	76
America	39
Central and South Asia	198
East Asia	227
Europe	151
Middle East	155
Oceania	26

Table 3.1: Number of Individuals Sampled from Each HGDP Region

Region	Number of Individuals
African American	341
Caucasian American	361
Hispanic American	235
Asian American	97

Table 3.2: Number of Individuals Sampled from Each NIST Population

3.3.1 Modification 1

We estimate the allele frequencies for each STR, population by population, using only the samples from each population in the training set. Formally, assume that our test set contains STR profiles, collectively R, taken from T populations, designated $\{1, ..., T\}$. Then for all i, R_i is drawn from exactly one population, $t \in \{1, ..., T\}$. Each profile contains L STR loci. Let $P(R_{il}|R_i \in t)$ be the probability of observing allele R_{il} at locus l in sample t in the training set. If R_{il} is not present in sample t in the training set, we set $P(R_{il}|R_i \in t) = \frac{1}{2W_t}$, where W_t is the number of individuals in the STR test set drawn from population t. Then

Equation 3.1 shows the estimated probability of observing the profile R_i within sample t.

$$P(R_i|R_i \in t) = \prod_{l=1}^{L} P(R_{il}|R_i \in t), \quad \text{for } t \in \{1, ..., T\} \tag{3.1}$$

Notice that $P(R_{il}|R_i \in t)$ is the conditional probability of observing the profile within a specific population. We can estimate $P(R_i \in t)$ as the proportion of STR profiles in the test set drawn from population t. We then multiply this by $P(R_{il}|R_i \in t)$ to find the the unconditional probability of observing the profile, as shown in Equation 3.2.

$$P(R_i) = P(R_i|R_i \in t)P(R_i \in t) \tag{3.2}$$

Since the training set contains only HGDP profiles, we combined each NIST population with the most similar HGDP region: NIST African American with HGDP Africa, NIST Caucasian American with HGDP Europe, NIST Hispanic with HGDP America, and NIST Asian American with HGDP East Asia. HGDP America contains only individuals from Central and South America. Most Asian Americans are of East Asian decent, but a large minority are from Central Asia, especially India and Pakistan [6].

This modification has the advantage of being very similar to Edge et al.'s approach, while accounting for population structure. One disadvantage is that the NIST populations and HGDP regions may not be very similar. Additionally, the training set used to calculate the STR allele frequencies will be much smaller than in Edge et al.'s pooled method. This greatly increases the chances that a given allele will not be observed in that sample in the training set and therefore have no reference frequency. This is especially a concern for the Oceania and America regions. While our method can handle alleles not present in the sample in the training set, the default allele frequency is not very informative. To use this approach in application, every population present in the test set would have to be matched with a population present in the training set.

3.3.2 Modification 2

We estimate the allele frequencies for each STR, population by population, using only the samples from each population in the training set. We also calculate the pooled allele frequencies from the training set as Edge et al. did. Formally, assume we have samples from T populations in the training set, designated $\{1, ..., T\}$. Let all profiles in the test set not from $\{1, ..., T\}$ be designated as from T'. Then for all i, R_i is drawn from exactly one population, $t \in \{1, ..., T, T'\}$. Each STR profile, R_i, contains L STR loci.

For $t \in \{1, ..., T\}$, let $P(R_{il}|R_i \in t)$ be the probability of observing allele R_{il} at locus l in sample t in the training set. If R_{il} is not present in sample t in the training set, we set $P(R_{il}|R_i \in t) = \frac{1}{2W_t}$, where W_t is the number of individuals in the STR test set drawn from population t. The calculation of the score denominator is then the same as Modification 1 as shown in Equations 3.1 and 3.2.

For $t = T'$, let $P_t(R_{il})$ be the probability of observing allele R_{il} at locus l in the entire training set. If R_{il} is not present in the training set, we set $P(R_{il}|R_i \in t) = \frac{1}{2W}$, where W is the number STR profiles in the test set. The calculation of the score denominator is then the same as Edge et al. as shown in Equation 2.9.

This approach basically combines Edge et al.'s method and Modification 1 to take advantage of population structure to estimate the STR allele frequencies when that information is available. When no information is available, it falls back on the Edge et al. pooled training set approach. Like Modification 1, the chance that a given allele will not be observed in that sample in the training set and therefore have no reference frequency is much higher than in Edge et al.

3.3.3 Modification 3

We pool all of the STR profiles available, from both the test and training sets, and then estimate the allele frequencies for each STR, population by population. Assume that in the

pooled training and test sets, we have samples from T populations, designated $\{1, ..., T\}$. Then for all i, R_i is drawn from exactly one population, $t \in \{1, ..., T\}$. Each STR profile, R_i, contains L STR loci. Let $P_t(R_{il})$ be the probability of observing allele R_{il} of locus l in sample t in all of the STR profiles. The calculation of the score denominator then proceeds as in Modification 1, Equations 3.1 and 3.2, accounting for the proportion of profiles drawn from each population.

This modification has the advantage of using all of the STR data we have. Additionally, STRs present in a sample in the test set will never not be present in the sample in the training set.

3.3.4 Additional Comparisons

Since Modification 3 uses STR profile information from both the testing and training sets, we also considered a slight extension of Edge et. al's method, where all of the STR profiles are incorporated. We combined all 872 HGDP STR profiles with the 1034 NIST profiles and estimate $P(R_{il})$ in the match score using this combined set.

We will also consider a naive baseline, where each STR profile is assumed to be equally likely. That would mean that $P(R_i)$ is a constant for all i. This scenario is clearly not true. However, it provides insight into how well the matching procedure works in the absence of any information about STR profile frequency.

A comparison between the Edge et al. method and the modifications described above in the source of STR allele frequencies estimates is shown in Figure 3.3. A more detailed comparison of the number of STR profiles being used to estimate STR allele frequencies for each test set population is shown in Table A.1 in Appendix A.

3.3.5 Simulation Design

We compared these different methods of calculating the match score using a procedure modeled after the true match rate procedure outlined in Section 3.1. We used the same

Modif.	Source of STR allele frequencies estimates used for HGDP profiles	Source of STR allele frequencies estimates used for NIST profiles	Will any alleles be missing frequency estimates?
Edge	Pooled Training Set	Pooled Training Set	Rarely
Mod. 1	Training Set by Population	Training Set by Population	Common
Mod. 2	Training Set by Population	Pooled Training Set	Common
Mod. 3	Test & Train. Sets By Pop.	Test & Train. Sets By Pop.	Never
Edge Ext.	Pooled Test & Train. Sets	Pooled Test & Train. Sets	Never
Uniform	Uniform Probabilities	Uniform Probabilities	Never

Table 3.3: Source of STR Allele Frequency Estimates for Each Modification

split of HGDP individuals, with 7/8 in the training set and 1/8 in the test set. However, instead of randomly partitioning the entire set of HGDP individuals, we performed stratified sampling by HGDP region. Since we are testing the effects of population structure, this stratified sampling ensures that each test set has the same structure. It additionally avoids the potential for any of the regions, especially the smallest ones, to be completely missing from one of the partitions. This stratified sampling was performed 100 times. We considered the same number of added NIST profiles. We used the same Beagle settings. The same set of partitions and Beagle imputations were used for each method in order to enable direct comparisons.

In Subsection 2.1.2, we broke Edge et al.'s method into steps. In this comparison, Steps 1a and 1b were performed once. Then Step 1c was performed separately for each modification as shown in Figure 3.3. Steps 2 and 3 were then performed on that output for each of the modifications described above.

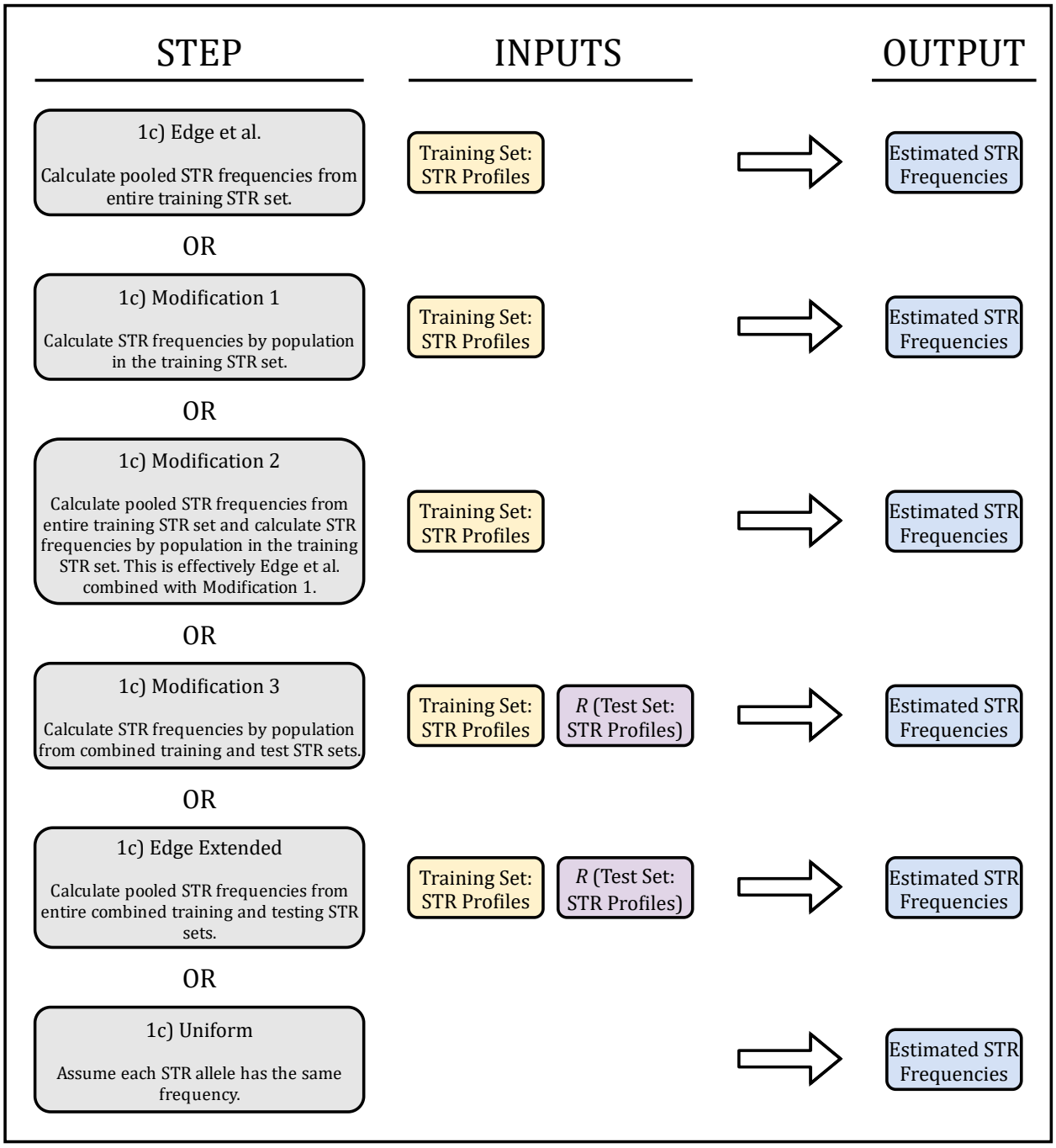


Figure 3.3: Comparison of Modifications to Step 1c

Grey boxes are the different proposed modifications. Blue boxes are intermediate outputs. Yellow boxes are inputs from the training set (see Figure 2.1) and purple boxes are inputs from the simulation test set (see Figure 3.1).

Chapter 4

RESULTS

4.1 Effect of True Match Rate

Figure 4.1 shows how the match accuracy using the SNP array HGDP profiles changes as additional non-matching NIST STR profiles are added. The trend is shown using a LOESS curve. Match accuracy is defined as the percentage of SNP profiles in the test set correctly matched to the corresponding STR profile from the same individual. As the number of non-matching NIST STR profiles added to the test set increases, the accuracy decreases, from an average of 98.6% with 0 added profiles to an average of 81.1% with 1034 added profiles.

4.2 Effect of Imputation Accuracy

We first compared the STR imputation accuracy between Beagle using SNP array profiles and WGS profiles. We measured accuracy as the percent of alleles correctly imputed at a given locus for a given individual. This is equal to 100% if both alleles are imputed correctly, 50% if 1 is, or 0% if none are. For Beagle, we considered the genotype with the highest probability as the imputed genotype. We did not further consider the imputation probability in our calculations. For both the SNP array profiles and WGS profiles, we used the respective 100 partitions of the data from the baseline scenario with 0 added NIST profiles. We calculated the average percentage of alleles correctly imputed across all individuals in all partitions for each locus.

For a baseline, we also considered STR imputation using just the Hardy-Weinberg frequencies from the training set of 100 partitions of the SNP array profiles. This is a useful baseline to consider because the match score function is effectively the log-likelihood ratio of observing

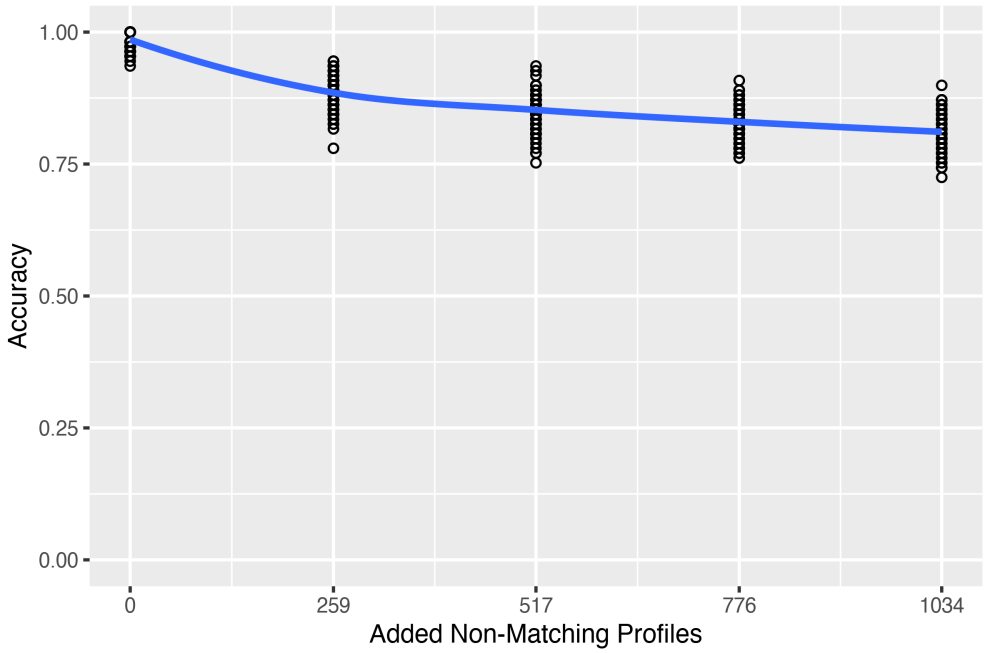


Figure 4.1: Matching Accuracy for SNP Array Scenarios

The test set always includes 109 matching SNP and STR HGDP profiles and then increasing amounts of non-matching NIST STR profiles were added. For each scenario, 100 random partitions were performed. Accuracy is the percent of SNP profiles in the test set correctly matched to the corresponding STR profile from the same individual. Trend is shown using LOESS curve.

the STR profile under the genotype probabilities from Beagle versus the Hardy-Weinberg frequencies in the training set.

The comparison of allele imputation accuracy is shown in Figure 4.2. As expected, imputation using Beagle on the WGS profiles systematically outperformed Beagle imputation using SNP arrays and the baseline Hardy-Weinberg imputation. Across all loci, Beagle using WGS profiles imputed of 81.2% of alleles correctly versus 60.5% for Beagle using SNP array profiles and 44.9% for the Hardy-Weinberg method. The difference in accuracy between Beagle using WGS profiles and Beagle using SNP array profiles is larger than the difference is accuracy

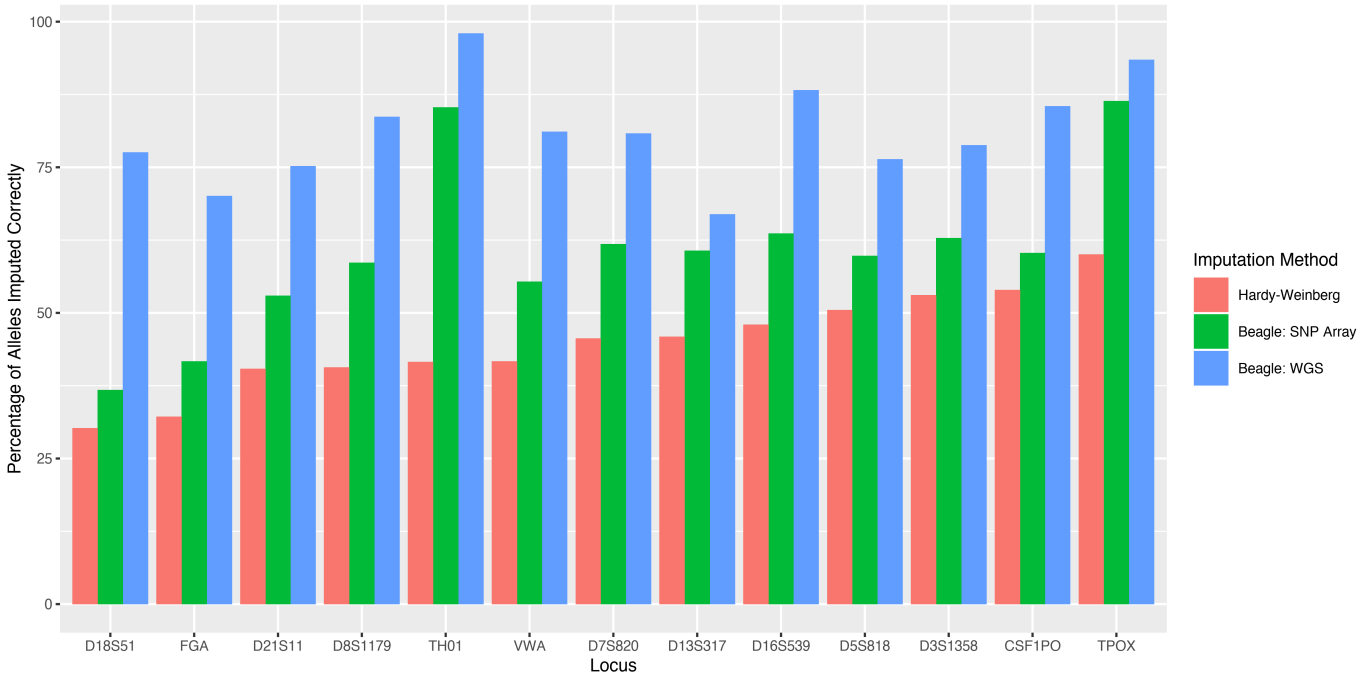


Figure 4.2: Comparison of Imputation Accuracy

Imputation accuracy is the average percentage of alleles correctly imputed at each locus over the 100 partitions for the baseline SNP array and WGS scenarios. Hardy-Weinberg imputation was done on the baseline SNP array scenarios using the training set allele frequencies.

between Beagle using SNP array profiles and the Hardy-Weinberg baseline. These improvements are even more pronounced at specific loci. For example, at the D18S51 locus Beagle using WGS profiles imputed of 77.6% of alleles correctly versus 36.8% for Beagle using SNP array profiles and 30.2% for the Hardy-Weinberg method.

The match accuracy using the WGS profiles as additional non-matching profiles are added is shown in Figure 4.3. The accuracy does not change much. When 0 non-matching profiles are added, an average match accuracy of 100% is achieved across the 100 permutations of the profiles. This decreases only to 98.3% when all 1034 non-matching profiles are added. This is a much smaller decrease than we found in the scenarios with the SNP array profiles.

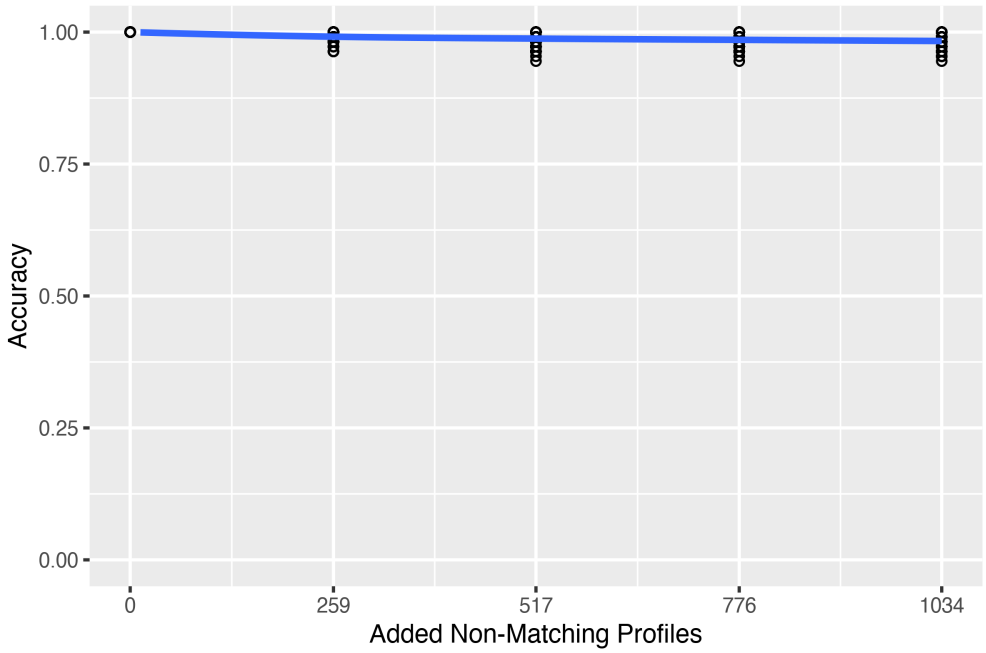


Figure 4.3: Matching Accuracy for WGS Scenarios

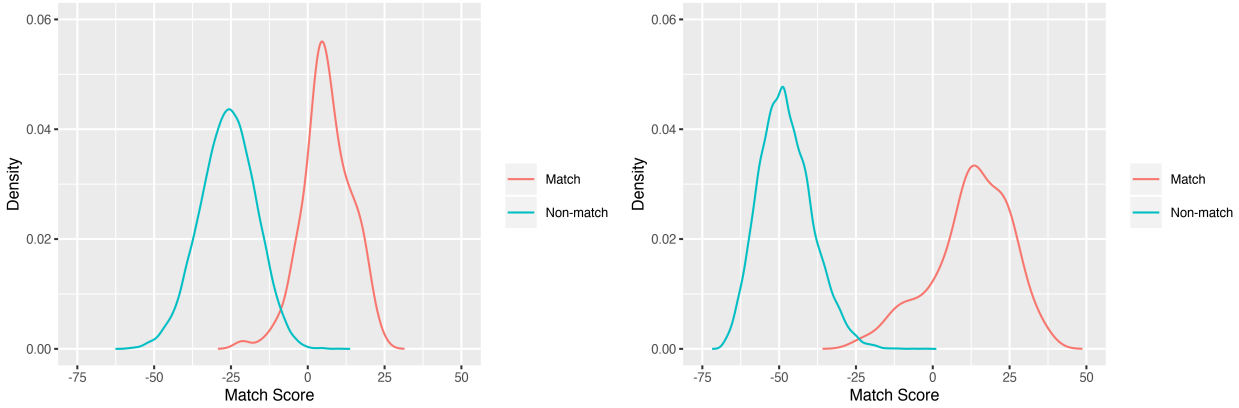
The test set always includes 109 matching SNP and STR HGDP profiles and then increasing amounts of non-matching NIST STR profiles were added. For each scenario, 100 random partitions were performed. Accuracy is the percent of SNP profiles in the test set correctly matched to the corresponding STR profile from the same individual. Trend is shown using LOESS curve.

In the easiest matching scenario, with no added non-matching profiles the results from the WGS and SNP array profiles are very close, but the accuracy gap increases as the number of added non-matching STR profiles increases.

4.3 Distribution of Match and Non-Match Scores

We also examined the distribution of the match scores from both the SNP array and WGS scenarios. We randomly selected a single permutation from each of the baseline scenarios with no added non-matching profiles. For the true matches, the match scores for the SNP

array were generally somewhat lower than for WGS scenarios, with median match scores of 5.75 and 13.73 respectively. The WGS true match scores were also more spread out, with a variance of 163.18 versus 62.06. For the non-matches, the WGS scores were much lower than the SNP array scores. The SNP array scenario had an median score of -25.87 versus -48.47 for the WGS scenario. We computed kernel density estimates for the match scores, which is shown in Figure 4.4. The estimated distributions for the WGS scores have much less overlap because the center of the distributions is further apart.



(a) SNP Array

(b) WGS

Figure 4.4: Kernel Density Estimates for Match Scores

Kernel density estimates for the distribution of match scores, estimated separately for true matches and non-matches. One baseline SNP array and one baseline WGS scenario permutation were chosen at random for estimation.

4.4 Score Modifications to Handle Population Structure

Figure 4.5 shows a comparison between Edge et al.'s original method and the five modifications described in Section 3.3. At the baseline with no added non-matching profiles, all of the methods attain the same accuracy. The differences between them become more pronounced as the number of non-matching profiles increases. Modifications 2 and 3 both achieve mod-

erately higher accuracy levels than Edge et al.'s method. In the scenarios with 1034 added profiles, Edge et al.'s method has an accuracy of 80.5% versus 88.2% for Modification 2 and 85.7% for Modification 3. Both of these modifications treat the NIST profiles separately from HGDP profiles when estimating the likelihood of observing a profile. This suggests that accounting for population structure in estimating the likelihood of observing a given STR profile can somewhat increase match accuracy.

The method using Edge et al.'s approach of combining the profiles into a single group for allele frequency estimation, but expanding to use the entire HGDP and NIST databases (Edge et al. Expanded in Figure 4.5), performs essentially identical to Edge et al.'s original approach With 1034 added profiles, the expanded version has an accuracy of 81.2% versus 80.5% for the original.

Somewhat surprisingly, the method assuming a uniform distribution of STR profiles only results in a moderate decrease in accuracy. With 1034 added profiles, the accuracy is 73.4%, which is 7.1 percentage points lower than Edge et al.'s method. This suggests that most of the discriminatory power in this matching approach comes from differences in the numerator of the log-likelihood ratio score and not the denominator that is based on the probability of observing an STR profile.

Finally, in the scenarios with 1034 added profiles Modification 1 has a significantly lower accuracy than Edge et al.'s method, 65.3% vs 80.5%. It even substantially underperforms the naive uniform method. Recall, that for Modification 1 we matched each NIST population with the most similar HGDP region. Both of the other modifications that accounted for population structure and did not match NIST and HGDP populations performed much better.

There are a couple potential explanations for the reduced accuracy. One possibility is that the NIST and HGDP regions do not correspond to the same population and therefore the STR allele frequencies from the HGDP regions in the training set are poor estimates of the

true STR allele frequencies in the NIST populations. Another possibility is that this is due to reduced sample size for STR allele frequency estimation. As shown in Table A.1 in Appendix A, Modification 1 uniformly has the smallest sample sizes for estimating allele frequencies.

Still, these results show that if population structure is going to be accounted for in the matching process, it needs to be done carefully. Attempting to account for population structure, but doing so poorly can give worse results than ignoring population structure altogether.

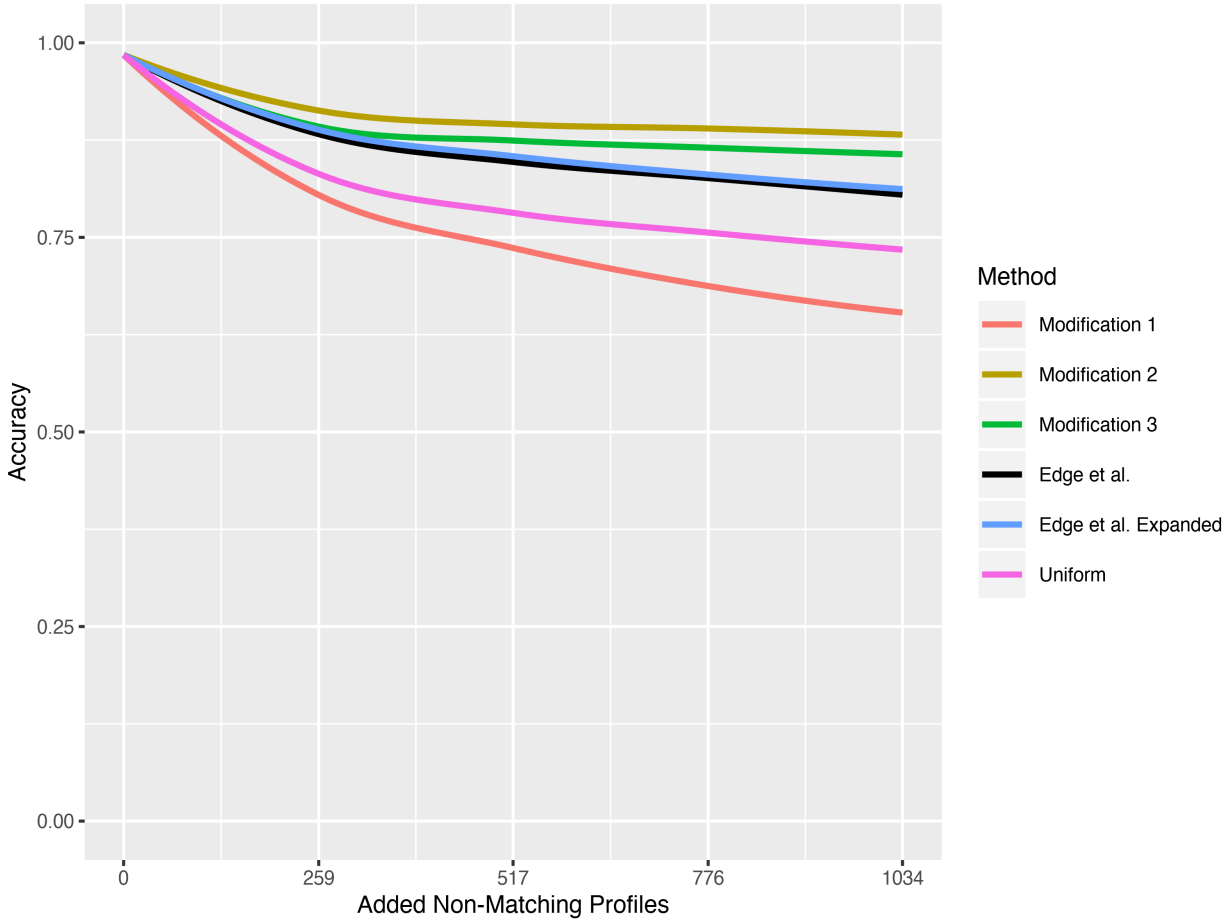


Figure 4.5: Comparison of Score Modifications

Accuracy is the percent of SNP profiles correctly matched to the corresponding STR profile. Trend is shown using LOESS curve.

Modification 1: By population using only training set frequencies, matched HGDP and NIST

Modification 2: HGDP by population and NIST from entire training set frequencies,

Modification 3: By population using training and test set frequencies, HGDP, NIST seperately

Edge et al.: Original Edge et al. method

Edge et al. Expanded: Edge et al. combined frequencies using all STR profiles

Uniform: Assume each STR profile has a uniform likelihood

Chapter 5

DISCUSSION

5.1 Conclusions

5.1.1 Effect of True Match Rate

Our results show that matching accuracy decreases monotonically as the true match rate decreases. The 17.5% decrease in accuracy between the 100% true match rate and 9.5% true match rate scenarios using the SNP array profiles is not huge. However, even the 9.5% true match rate scenarios are under favorable assumptions. It is likely that many potential applications would have very small true match rates. Additionally, we still assumed a priori that each SNP array profile did have a match in the larger set of STR profiles. Relaxing both of these assumptions would likely lead to further declines in accuracy, especially in the context of very large databases.

5.1.2 Effect of Imputation Accuracy

Our results also show that imputation accuracy makes a big difference in matching accuracy in more challenging match scenarios. Because these more challenging scenarios are more realistic, this is important to keep in mind for potential applications. We achieved greater imputation accuracy by switching from SNP array profiles to WGS profiles. In applications, it may not be possible to use WGS profiles. However, other factors, such as greater reference panel size and greater similarity to the test set have also been shown to increase imputation accuracy and could potentially be used achieve similar match accuracy gains [10].

5.1.3 Effect of Population Structure

The comparison between the different score modifications suggest that it is possible to achieve higher levels of match accuracy by incorporating known information about the population structure into the calculation of the match score. However, it is also possible to reduce the accuracy when attempting to account for population structure. Additionally, most of discriminatory power of Edge et al.'s matching method is contained in the imputation part of the score and assuming a uniform distribution of STR profile frequencies only results in a moderate reduction of accuracy. This is useful context for situations where it may be hard to estimate the STR profile frequency.

5.2 Limitations

One potential limitation of our analysis is that the NIST STR profiles come from different populations than the HGDP profiles. The NIST profiles come from 342 African Americans, 261 Caucasian Americans, 236 Hispanic Americans and 97 Asian Americans. The HGDP focused on collecting samples primarily from indigenous and isolated populations. As a result, the HGDP does not contain any profiles from the United States of America. Additionally, there are no profiles from many of the origin countries of America's largest ancestry groups including German, non-indigenous Mexican, Irish and English.

Including these additional profiles from different populations than the true matching profiles may lead us to overestimate the accuracy of this method. Many genetic databases are maintained at a national or sub-national level, including forensic databases. The individuals in these databases primarily come from a relatively small number of local populations. When matching to one of these databases, the matching process would need to distinguish between many profiles from the same population. Profiles from the same population are more likely to be similar than profiles from different populations. In our test scenario, the method needs to distinguish primarily between profiles from different populations, which may be an easier scenario because the profiles are more different from each other.

We have attempted to assess how modifications to the match score that account for population structure affect matching accuracy. Since we used the same data for all of the simulations, they all have the same population structure. As a result, our results may not generalize to different population structures or may be overfit to our data.

Additionally, Modifications 1, 2 and 3 take the population structure into account for calculating the probability of observing a given STR profile. However, this population information is not further incorporated into the calculation of the match score. This is somewhat contrived. Presumably information about the population each profile belongs to should further change the likelihood ratio that two profiles are from the same individual versus different individuals. If the population for each profile is known with 100% certainty and two profiles are from different populations the likelihood should be zero, because we already know for certain they are not a match. Our approach modifications effectively assume that population information is available for the STR profiles and not the SNP profiles. Additional work is needed to further incorporate all of this population into the matching process.

Another potential limitation of this analysis is the small sample size. Many DNA profile databases contain on the order of millions or tens of millions of DNA profiles. In our largest scenario, we consider only 1,143 individuals. We believe that our analysis provides useful insight. We are limited by the public availability of DNA profiles.

5.3 Future Work

5.3.1 Beagle Alternatives

Imputation done using Beagle is a core component of the method proposed by Edge et al. However, there are some potential alternatives for estimating STRs from whole genome sequencing data. Methods like lobSTR and HipSTR aim to sequence STRs directly from the whole genome sequencing reads, without doing imputation from a reference panel. This could have several potential advantages. They claim very high levels of accuracy: 97% from 21x coverage for lobSTR [16] and 95.2% from 30x coverage for HipSTR [29]. This is higher

than the 81.2% accuracy we achieved on the CODIS loci using Beagle on whole genome sequencing data. However, direct comparison is difficult since the testing scenarios vary significantly.

Additionally, these methods do not require a training set with both whole genome sequences and STR sequences. Since the HGDP is the only major public database with individuals sequenced with both techniques, this would be extremely helpful. Recall in the Edge et al. method, the match score is calculated as:

$$\lambda(R_i, S_j) = \ln[P(R_i|S_j, M = 1)] - \ln[P(R_i)] \tag{5.1}$$

The first part of this equation, $\ln[P(R_i|S_j, M = 1)]$, is calculated using imputed STR genotypes from Beagle using a reference panel. This step could potentially be replaced using one of these direct STR estimation methods. The second part of this equation, $\ln[P(R_i)]$, is calculated using the Hardy-Weinberg frequencies from the training set in Edge et al.'s approach or from the entire set of available STR profiles in some of the proposed modifications. This could potentially be replaced with relevant population STR frequencies. Many papers have been published which estimate STR population frequencies for various populations, including Brazil [9], Britain [20], and Thailand [4] among many others. Together, this could potentially obviate the need for a reference set of genotypes.

However, there are also some potential hurdles to using these direct STR estimation methods. They are less flexible than Beagle. Both HipSTR and lobSTR require raw read data. HipSTR only works with Illumina sequencing data [29]. lobSTR can be used on any whole genome sequencing data, but requires additional modelling steps for sequences from any method besides Illumina PCR-free sequencing [16].

Additionally, the creators of both HipSTR and lobSTR describe issues with estimating some of the CODIS loci. For both methods, it is easier to estimate shorter STRs since they are more likely to be completely spanned by a read or have a read covering part of an STR and its flanking region. However, some of the CODIS STRs are very long. For example, in the

NIST database the D21S11 locus ranges between 104 and 156 base-pairs long. For reference, the HGDP profiles were sequenced with reads lengths of 151 base-pairs [3]. For some D21S11 alleles, it is impossible that they could be spanned by these reads and for many others it is unlikely.

BIBLIOGRAPHY

[1] Sanne E. Aalbers, Michael J. Hipp, Scott R. Kennedy, and Bruce S. Weir. Analyzing population structure for forensic STR markers in next generation sequencing data. *Forensic Science International: Genetics*, 49:102364, 2020.

[2] Bridget F.B. Algee-Hewitt, Michael D. Edge, Jaehee Kim, Jun Z. Li, and Noah A. Rosenberg. Individual Identifiability Predicts Population Identifiability in Forensic Microsatellite Markers. *Current Biology*, 26:935–942, 2016.

[3] Anders Bergström, Shane A. McCarthy, Ruoyun Hui, Mohamed A. Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, Hélène Blanché, Jean-François Deleuze, Howard Cann, Swapan Mallick, David Reich, Manjinder S. Sandhu, Pontus Skoglund, Aylwyn Scally, Yali Xue, Richard Durbin, and Chris Tyler-Smith. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484), 2020.

[4] N. Boonderm, D. Suriyanratakorn, S. Sangpueng, N. Onthong, A. Nettakul, and W. Waiyawuth. Population Genetic Data of 21 STR markers in Thais of Southern Border Provinces of Thailand. *Forensic Science International: Genetics Supplement Series*, 6:e523–e525, 2017.

[5] Brian L. Browning, Ying Zhou, and Sharon R. Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[6] Abby Budiman and Neil G. Ruiz. Key facts about Asian origin groups in the U.S. *Pew Research Center*, 4 2021.

[7] John M. Butler. *Advanced Topics in Forensic DNA Typing: Methodology.* Academic Press, San Diego, 2012.

[8] Howard M. Cann, Claudia de Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, Walter F. Bodmer, Batsheva Bonne-Tamir, Anne Cambon-Thomsen, Zhu Chen, Jiayou Chu, Carlo Carcassi, Licinio Contu, Ruofu Du, Laurent Excoffier, G. B. Ferrara, Jonathan S. Friedlaender, Helena Groot, David Gurwitz, Trefor Jenkins, Rene J. Herrera, Xiaoyi Huang, Judith Kidd, Kenneth K. Kidd, Andre Langaney, Alice A. Lin, S. Qasim Mehdi, Peter Parham, Alberto Piazza, Maria Pia Pistillo, Yaping Qian, Qunfang Shu, Jiujin Xu, S. Zhu, James L. Weber, Henry T. Greely, Marcus W. Feldman, Gilles Thomas, Jean Dausset, and L. Luca

Cavalli-Sforza. A human genome diversity cell line panel. (letters). *Science*, 296:261+, Apr 2002.

[9] Vitor Rezende da Costa Aguiar, Eldamária de Vargas Wolfgramm, Frederico Scott Varella Malta, Adriana Gonçalves Bosque, Amanda de Castro Mafia, Vanessa Cristina de Oliveira Almeida, Fabiola de Andrade Caxito, Victor Cavalcanti Pardini, Alessandro Clayton Souza Ferreira, and Iúri Drumond Louro. Updated Brazilian STR allele frequency data using over 100,000 individuals: An analysis of CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, Penta D, Penta E, TH01, TPOX and vWA loci. *Forensic Science International: Genetics*, 6(4):504–509, 2012.

[10] S Das, GR Abecasis, and BL Browning. Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics*, 19:73–96, 2018.

[11] Michael D. Edge, Bridget F. B. Algee-Hewitt, Trevor J. Pemberton, Jun Z. Li, and Noah A. Rosenberg. Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *Proceedings of the National Academy of Sciences of the United States of America*, 114(22):5671–5676, 2017.

[12] Hao Fan and Jia-You Chu. A brief review of short tandem repeat mutation. *Genomics, Proteomics & Bioinformatics*, 5(1):7–14, 2007.

[13] Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[14] Jianye Ge, Arthur Eisenberg, and Bruce Budowle. Developing criteria and data to determine best options for expanding the core CODIS loci. *Investigative genetics*, 3:1–1, Jan 2012.

[15] Christi J. Guerrini, Jill O. Robinson, Devan Petersen, and Amy L. McGuire. Should police have access to genetic genealogy databases? Capturing the Golden State Killer and other criminals using a controversial new forensic technique. *PLOS Biology*, 16(10):1–9, 10 2018.

[16] Melissa Gymrek, David Golan, Saharon Rosset, and Yaniv Erlich. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, 22(6):1154–1162, Jun 2012. 22522390[pmid].

[17] Douglas R. Hares. Expanding the CODIS core loci in the United States. *Forensic Science International: Genetics*, 6(1):e52–e54, 2012.

[18] Carolyn R. Hill, David L. Duewer, Margaret C. Kline, Michael D. Coble, and John M. Butler. U.S. Population Data for 29 Autosomal STR Loci. *Forensic Science International: Genetics,*, 7:e82–e83, 2013.

[19] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle,

Alan M Zahler, and David Haussler. The Human Genome Browser at UCSC. *Genome research*, 12(6):996–1006, 2002.

[20] Urszula Krzeminska-Ahmadzai, Benjamin Buckley, Thomas Loake, Claire Nicholson, David Beesley, and Casey Randall. Population data for 23 autosomal STR loci in White British population. *Legal Medicine*, 50:101863, 2021.

[21] Jun Z. Li, Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, Gregory S. Barsh, Marcus Feldman, Luigi L. Cavalli-Sforza, and Richard M. Myers. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, 319:1100–1104, 2008.

[22] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, December 2003. 14704198[pmid].

[23] Karen Norrgard. Forensics, DNA Fingerprinting, and CODIS. *Nature Education*, 1(1):35, 2008.

[24] Noah A Rosenberg. Standardized Subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, Accounting for Atypical and Duplicated Samples and Pairs of Close Relatives. 70(6):841–847, 2006.

[25] Shubham Saini, Ileena Mitra, Nima Mousavi, Stephanie Feupe Fotsing, and Melissa Gymrek. A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nature Communications*, 9(1):4397–4397, Oct 2018. 30353011[pmid].

[26] Congressional Research Service. The Use of DNA by the Criminal Justice System and the Federal Role: Background, Current Law, and Grants. 2021.

[27] Adam Vaughan. DNA site GEDmatch sold to firm helping US police solve crime. *NewScientist*, 12 2019.

[28] Kris A. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). December 2020.

[29] Thomas Willems, Dina Zielinski, Jie Yuan, Assaf Gordon, Melissa Gymrek, and Yaniv Erlich. Genome-wide profiling of heritable and de novo STR variations. *Nature Methods*, 14:590+, June 2017. 6.

[30] Yaran Yang, Bingbing Xie, and Jiangwei Yan. Application of next-generation sequencing technology in forensic science. *Genomics, Proteomics & Bioinformatics*, 12(5):190–197, 2014. Special Issue: Translational Omics.

Appendix A

SAMPLE SIZES FOR ESTIMATING STR ALLELE FREQUENCY

	Number of STR Samples				
Test Set Population	Edge et al.	Mod. 1	Mod. 2	Mod. 3	Edge Extended
HGDP Africa	763	66	66	76	1906
HGDP America	763	34	34	39	1906
HGDP C and S Asia	763	173	173	198	1906
HGDP East Asia	763	199	199	227	1906
HGDP Europe	763	132	132	151	1906
HGDP Middle East	763	136	136	155	1906
HGDP Oceania	763	23	23	26	1906
NIST Afr American	763	69	763	341	1906
NIST Caucasian	763	132	763	361	1906
NIST Hispanic	763	34	763	235	1906
NIST Asian American	763	199	763	97	1906

Table A.1: Number of STR Profiles Used to Estimate Allele Frequencies for Each Test Set Population Group for Edge et al. Method and Each Modification