The Y-Chromosome in Forensic and Public Health Genetics

Taryn O. Hall

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Bruce Weir, Chair

Anna Mastroianni

Elizabeth Blue

Sara Goering

Program Authorized to Offer Degree:

Public Health Genetics

University of Washington

Abstract

The Y-Chromosome in Forensic and Public Health Genetics

Taryn O. Hall

Chair of the Supervisory Committee:
Bruce Weir, Professor, Department of Biostatistics

Unlike the autosomes and X-chromosome, the Y-chromosome does not recombine with a homologous partner, save obligatory recombination with the X-chromosome in the pseudoautosomal regions. Because of the lack of recombination, the male-specific region of the Y is inherited clonally from father to son. This inheritance pattern influences the way the Y-chromosome can be used forensically and in public health genetics. Chapter 1 of this dissertation details the history and use of the Y-chromosome in forensic genetics. The second half of the chapter examines the properties of a new estimator of population-specific F_{ST} on the Y-chromosome and its use in calculating the forensic match probability, as well as a comparison of the kappa method and an evolutionary model for match probability calculation. The Y-chromosome has recently been used for forensic familial searching using a commercial DNA

database. Chapter 2 presents a policy analysis of forensic familial searching in commercial DNA databases. Due to its small size and a number of analytic challenges, the Y-chromosome has largely been ignored in genetic association studies. One of the major analytic challenges is how to adjust for population stratification on the Y-chromosome, as relatedness is different between the Y and autosomes. Chapter 3 examines associations between SNPs on the Y-chromosome and obesity in Hispanic men, adjusting for autosomal and Y-chromosome principal components.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to acknowledge my wonderful dissertation committee. A huge thank you to Dr. Bruce Weir for stepping up and providing me with a new project, funding and mentorship when my first dissertation project unexpectedly fell though. Anna Mastroianni, you were instrumental in developing the ELSI chapter of this dissertation. Thank you for your mentorship – it is nice to see that someone build like me can succeed in academia. Dr. Liz Blue, I also thank you for your mentorship. And Dr. Sara Goering, thank you for stepping in and performing as my GSR.

Several other individuals have provided me technical support. Thank you to Xiuwen Zheng for providing me with the C++ code to do the bulk of my calculations for Chapter 1. My project quickly morphed into a Big Data situation and would have been totally intractable without it. And to Alex Thiery, briefly my cube mate, thank you for the pointers in R; that was the last piece I needed to be able to code proficiently in that language. Adrienne Stilp and Yatong Li both provided me with much help with the R code for Chapter 3 – thank you so much ladies.

A special thank you to Dr. Margo Bergman. In addition to being my best friend and providing some amazing caretaking to me through this protracted process, you also helped me with the last piece of thinking to complete the policy analysis for Chapter 2.

I would also like to acknowledge the amazing women I met in the course of this program: Laura Health, Nini Shridhar, Lorelei Walker, and Emmi Bane. You are all brilliant and I look forward to being your friend and colleague going forward.

Thank you to my family for always encouraging me to high achievement. Without you, I would not have gotten here.

Finally, thank you to my loving husband, Marc Teale. You have been so supportive through this process and provided me with a lot of technical help as my in-house IT consultant. I look forward to our life together and the birth of our first child in April!

INTRODUCTION

This dissertation contains a collection of papers examining the Y-chromosome in forensic and public health genetics. The first half of Chapter 1 details the history and use of the Y-chromosome in forensic genetics. The second half of Chapter 1 is an empirical study of the estimates derived from the Y-chromosome used in the forensic match calculation: F_{ST} and the Y-chromosome profile probability.

Recently, law enforcement in Idaho used a commercial DNA database, AncestryDNA, for forensic familial searching. Contrary to established forensic familial searching practice, a Y-chromosome STR profile was used exclusively to attempt to identify a possible suspect in the murder of Angie Dodge. The familial search identified Michael Usry Jr, a filmmaker from New Orleans, as a possible suspect. After a month under suspicion, additional DNA testing eliminated Usry Jr as a suspect. However, this search technique can lead to harms to innocent individuals identified. Chapter 2 presents an analysis of policy alternatives to attempt to minimize the harms to those identified through forensic familial searching of commercial DNA databases.

Finally, Chapter 3 attempts to develop a method for use of the Y-chromosome in genetic association studies. The clonal inheritance of the Y from father to son, lack of recombination, and the diminutive size of the Y make analysis challenging. One major challenge is adjusting for population structure when using Y-chromosome markers. Relatedness on the Y is different than the autosomes, but how best to estimate and adjust for this relatedness has yet to be determined. Use a dataset of Hispanic American men, Chapter 3 presents an analysis of the Y-chromosome and obesity, adjusting for population stratification using both autosomal and Y-chromosome principal components.

Chapter 1. THE Y-CHROMOSOME IN FORENSICS

1.1 HISTORY AND USE OF FORENSIC STR PROFILES

Forensic genetics (also called DNA profiling or DNA fingerprinting) is the cornerstone of modern forensic science. Genetic techniques are used forensically to match donors to crime stains, to identify missing persons and remains, and to establish parentage.

Separate genetic profiles can be constructed using standard marker sets for the autosomal (chromosomes one through 22), Y-chromosome, and mitochondrial DNA. The types of genetic marker currently used in forensic genetic practice are length polymorphisms called short tandem repeats (STRs). STRs used forensically typically have a simple or more complex tetra nucleotide repeat structure. The markers are highly polymorphic, with between 15 and more than 100 alleles observed at a locus, giving these markers, taken together, high power to discriminate between individuals.[1] In 1997, an effort was made to standardize the nomenclature for STR alleles, based on the number of repeated motifs or overall length polymorphism size.[2,3] That same year, the Federal Bureau of Investigation (FBI) announced a core of 13 autosomal STR loci required for the National DNA Index System (NDIS), a subset of the U.S. National Combined DNA Index System (CODIS).[4] These markers were chosen to be highly informative, easily amplified using PCR, and on different chromosomes or opposite ends of the same chromosome so that they are almost genetically unlinked.

Prior to the late 1990s, the use of the Y-chromosome in forensic science was limited to sex identification.[5] However, typing forensic markers on the Y-chromosome can be useful for individual identification. Because the Y-chromosome is inherited clonally from father to son, a Y-STR profile identifies males in the same patrilineage. While not individually identifying like

autosomal DNA profiling, Y-profiling has its own evidentiary value and can definitively exclude individuals.

1.1.1 *Sexual Assault*

Criminally, Y-STR profiles are particularly useful in resolving the male component of male/female DNA mixtures. In a sexual assault case, the male autosomal DNA profile can be found by isolating sperm cells from the crime stain by use of differential lysis or laser capture microdissection. When sperm are not present or unable to be isolated, inferring the male autosomal profile is unlikely, due to the large female-to-male ratio of DNA in the mixture – essentially, the female DNA saturates the PCR primers, making them unable to amplify the male DNA fraction.[6] Y-specific PCR primers are able to amplify Y-chromosome STR markers, providing an unambiguous male DNA profile[7], even when minimal male DNA and no detectable sperm are present in a sample.[8,9] Whereas autosomal markers can detect only the male component of a mixture at a male:female ratio of 1:10, the detection limit of Y-STRs is 1:2000.[8,10,11]

For a variety of reasons, sexual assaults are not always reported – and evidence gathered – right away. The longer the interval between assault and evidence collection, the less likely sperm (and the male autosomal DNA component) will be obtained from swabs.[12] However, though a full Y-STR profile will likely not be obtained, it is possible to recover a partial or nearly complete Y-STR profile 8-9 days post-coital from a vaginal swab.[13,14] Y-STR profiles can be detected following digital penetration, up to 72-hours after the event.[15]

In the case of multiple-assailant sexual offences or sexual assault following consensual intercourse, because sperm coming from different individuals are morphologically similar and gamete genomes are haploid, inferring the autosomal component of each donor is not trivial,

especially as the number of donors increases. Assuming assailants are not paternally related, Y-STR profiles can be used to determine the number of assailants.[6] Additionally, Y-STR profiling, along with laser capture microdissection and fluorescent in-situ hybridization using an X/Y probe (to capture individual, Y-containing sperm), can be used to sort the sperm cells belonging to each assailant.[16] The haploid autosomal DNA obtained from the sorted sperm cells is then used to construct a consensus autosomal profile for each assailant.

1.1.2 *Parentage/Paternity Testing*

Y-chromosome STR profiles are also useful in establishing family relationships. A male child's Y-STR profile is expected, barring mutation, to be the same as his father's. Using Y-STR typing when establishing the paternity of male children can be useful as a screening step – non-matches exclude putative fathers (with stronger evidence than autosomal markers) without the need to type autosomal markers or any DNA from the child's mother.[17,18] However, Y-STR matches do not generate as much evidence as autosomal marker matches for inclusion of paternity.[18] Y-STR profiling is also helpful to resolve parentage questions when autosomal markers do not generate a conclusive level of evidence for or against paternity.[19] This may be the case when one or both parents of the child in question (and both parents of the unavailable parent) are deceased or otherwise unavailable for testing. Assigning parentage based on putative sib-ships using only autosomal markers is often inconclusive.[20,21] Y-STR testing can help resolve relationships when two male children are tested. In the case of a deceased or unavailable father, any known paternally-related male will have the same Y-profile as the putative father and can be used to match the male child in question.[22] Here as above, the strongest evidence generated is for exclusion of non-fathers. When only the mother is unavailable for testing, combining Y-STR and autosomal STR profile matching for the putative father/son pair provides evidence of parentage

equal to or exceeding the level of evidence generated by knowing the mother's autosomal profile.[17]

Y-STR analysis may also be the method of choice when there have been many generations succeeding the paternity event in question.[23–25] Autosomal markers are essentially unlinked and recombine every generation; inferring the paternally inherited alleles over several generations without extensive pedigree information may not be possible. However, Y markers are not subject to recombination and Y-STRs can be traced from father to son over many generations and thus can be typed in descendants to provide evidence for paternity or non-paternity even hundreds of years after the paternity event in question.[25]

1.1.3 *Missing Person and Remains Identification*

The most desirable way to match samples or remains in missing person cases to the missing person is with a DNA sample left by the missing person (e.g. hair on a brush). If such a sample is not available, samples from close relatives can stand-in for the missing person. In addition to the 13 core CODIS markers, CODIS recommends attempting to collect mitochondrial DNA markers for all remains, and Y-chromosome markers for male remains, along with samples from relatives.[26] Similar to paternity testing, Y-STR profiles can be used in male missing person cases and for male remains identification by matching sample Y-profile to that of a male patrilineal relative. This is particularly helpful when a close relative is not available, as even more distantly related males in the same lineage are expected to have the same Y-profile.

1.1.4 *Familial Searching*

Y-STR profiles may be useful for familial searching – when there is no match in the database, matching a crime sample to possible relatives of the true evidence source in the forensic database

to facilitate suspect identification. Familial searching has been primarily adopted in the United Kingdom. Though not conducted federally, five states in the US conduct familial searches.[26,27] Currently, familial searching uses autosomal STR markers to find potential parent-child or full-sib relationships through partial crime stain/database matching. The UK procedure first generates lists, using number of alleles shared, for potential parent-child pairs (where samples must have one allele in common at each locus) and full-sib pairs (using a threshold of 11 alleles shared over 10 loci) and then ranks matches according to the likelihood of the profile given the degree of relatedness.[28] This process generates many false positives, therefore additional filters, often non-genetic, must be applied to eliminate false positives and reduce the number of leads to follow-up. As offenders, and consequently profiles in forensic databases, are often male, Y-STR profile can corroborate familial relationships, removing false-positive hits, as father-son and full-brothers will share the same Y-profile.[28,29]

1.1.5 *Inferring Suspect Surname*

Like Y-chromosome genetic profiles, surnames are also inherited patrilineally in many cultures; therefore, there is a correlation between surname and Y-STR profile.[30] Studies of English and Irish men have shown that many men who share a surname are also part of a Y-STR descent cluster and rarer surnames show more clustering.[31–33] While using Y-STR profile to identify surname has poor predictive power, it has been proposed as a potential screening step in suspect identification. This procedure has been used successfully to find the surname of a perpetrator in small Chinese communities by matching the Y-STR profile of the oldest male bearing the surname in the community to the Y-profile in a crime scene specimen.[34] Key individuals within a surname group were then followed up with autosomal typing, leading to suspect identification.

1.2 Determining the Weight of Evidence

If there is a match between the crime stain profile and the suspect profile, it means that the suspect is not excluded from the possibility of being the donor of the crime stain profile. Matching profiles do not unequivocally prove that the suspect is indeed the perpetrator of the crime, rather this provides evidence supporting a particular hypothesis about the crime at hand. The strength of this evidence must be quantified. In practice, this is done by calculating a likelihood ratio.[35]

We would like to answer the question: what are the odds the suspect is the source of the evidentiary profile (H_1), given the DNA match (E) and any other information (I).

$$O(H_1|E,I) = \frac{\Pr(H_1|E,I)}{\Pr(H_2|E,I)} \tag{1.1}$$

where H_2 is the alternate hypothesis that the suspect is not guilty.

We can use Bayes' Theorem to rearrange this equation:

$$O(H_1|E,I) = \frac{Pr(E|H_1,I)}{Pr(E|H_2,I)} \cdot \frac{Pr(H_1|I)}{Pr(H_2|I)}. \tag{1.2}$$

We now have the posterior odds of H_1. The last term in this equation is the prior odds of H_1, before including the DNA evidence, calculated as the probability that the suspect is the source, given the other information in the case (the prosecutor hypothesis, H_1) over the probability that someone else is the source, given the information (the defense hypothesis, H_2). The middle term is the likelihood ratio which gives the strength of the DNA evidence by calculating the probability of observing the DNA evidence, given the prosecutor hypothesis and non-DNA evidence, relative to the probability of observing the DNA evidence given the defense hypothesis and non-DNA evidence. A larger likelihood ratio favors the prosecution hypothesis.

1.2.1 *Calculating the Likelihood Ratio*

Suppose a crime scene DNA profile matches a suspect's profile. Under H_1, the prosecutor's hypothesis, the probability of the match given the suspect is the source of the evidentiary profile is: $Pr(E|H_1,I) = 1$. Calculating the probability of observing a match under the defense hypothesis is not as simple as the probability calculated under the prosecutor hypothesis. Different calculations are required if the defense hypothesis is that a relative of the suspect committed the crime or an unrelated person is the true perpetrator.

The match probability for profile A is the chance that two DNA profiles match, given we have observed profile A already, but did not, in truth, come from the same donor (Pr(A|A)). The profile probability obtained from counting the observed frequency of Y haplotypes in a database is not the same as the match probability. For both the autosomes and Y chromosome, the majority of haplotypes that could be constructed based on existing alleles at markers in populations are not observed. Observing a haplotype once in a population increases the chances we will observe the same haplotype again in the population. This is due to the inherited and shared nature of DNA. Underlying each population, is a deep, unobserved pedigree. If one person inherits an allele at a particular marker, chances are others in the population have inherited the same allele from a common ancestor. This is easy to think about in terms of the Y chromosome, where the marker profile identifies not an individual, but a patrilineage. Closely related paternal relatives – fathers and sons, brothers, uncles, paternal cousins – will all likely have the same Y chromosome profile. This deep pedigree, called population structure, must be taken into account when calculating the match probability.

Assuming the true criminal is unrelated to the suspect, the Scientific Working Group on DNA Analysis Methods (SWGDAM), an independent organization which "serves as a forum to

discuss, share, and evaluate forensic biology methods, protocols, training, and research to enhance forensic biology services as well as provide recommendations to the FBI Director on quality assurance standards for forensic DNA analysis," recommends calculating the match probability by use of an evolutionary model, constructed by applying the Balding-Nichols equation to the Y-chromosome, which employs the population structure parameter Θ and the profile probability [Pr(A)]: [36–38]

$$\Pr(A|A) = \theta + (1-\theta)\Pr(A). \quad (1.3)$$

1.2.2 *The Profile Probability, Pr(A)*

Because the CODIS markers segregate independently of one another in the genome, the frequency of any given autosomal profile, Pr(A), in a population can be calculated as the product of the population allele frequencies of the observed alleles at each marker. However, because the Y-chromosome has holandric inheritance, STR markers in the male-specific portion are linked. Linkage on the autosomes, X-chromosome and the pseudo-autosomal region of the Y-chromosome is broken up over time by recombination. There is no recombination in the male-specific region of the Y-chromosome, but mutation still acts on this region. The mutation rate of a Y-STR marker, ranging from about $5x10^{-4}$ to $1x10^{-2}$ per transmission, is large enough to break up some of the dependence between markers over time.[39,40] It has been assumed that Y-STR markers are highly dependent, but Caliebe and colleagues were the first to test this empirically. They found that was indeed a complex relationship between Y-STR markers, but that this relationship did not reveal clusters of markers based physical position or mutation rate.[41]

Since the markers and their allele frequencies are not expected to be independent, the product rule cannot be employed to calculate the profile frequency; rather, all markers must be taken together to determine a multi-locus haplotype frequency. Haplotype frequencies are also population specific – genetic markers on the Y-chromosome are more prone to genetic drift than the autosomes due to lack of recombination in the male-specific region of the chromosome, reduced number of Y-chromosomes relative to the autosomes, and gender-specific behaviors leading to further reduction in the effective number of Y-chromosomes in a population.[42]

In practice the frequency of the Y-STR profile in question can be estimated by referencing databases with large numbers of Y-STR profiles and counting the number of matching profiles within the population of interest. This provides an estimate of the profile probability for the Y-haplotype. One of these databases is the Y Chromosome Haplotype Reference Database (YHRD).[43] Following the guidelines of the International Society of Forensic Genetics (ISFG) for the publication of genetic population data, submissions to YHRD are required to consist of at least 17 Y-STR markers (including the eight-marker minimal Y-haplotype: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385)[5] for 200 or more individuals from a population.[44] As of 2015, YHRD contained more than 84,000 Y-profiles of 17 markers or greater in 572 populations.[43]

To increase Y-chromosome profile ability to distinguish between related Y-haplotypes there has been a push to add more markers to forensic Y-haplotypes. As marker sets become larger, individual haplotypes become rarer. This makes estimating Pr(A) more difficult as the sought-after haplotype is more likely be represented by one copy or missing in the database. As marker panels become larger, reference databases will also need to grow. The lack of independence between Y-chromosome markers also means that the distinguishing information

provided by one marker is not independent of other markers. After a certain number of markers are added to a profile, the information provided by subsequent markers may be redundant.

Choosing the most informative subset of a large marker panel may provide an optimization of the trade-off between discrimination capacity and genotyping and database costs, and allow for more reliable estimation of Pr(A). In order to choose such a subset, one must have a measure of the additional information provided by each marker added to the subset. One such measure, used recently in forensic genetics, is Shannon entropy.[45] Applied to genetics, it measures the residual uncertainty of allele S, given its haplotype background. For marker or marker set X, Shannon entropy, B, is calculated as:

$$B(X) = -\sum_S \, p_S \ln(p_S), \tag{1.4}$$

where p_S are the allele frequencies of each allele at marker X. The residual entropy, or information, provided by marker X after marker (or marker set) Y is known can be calculated as the conditional entropy of X given Y: $B(X|Y) = B(X,Y) - B(Y)$. Recently, Siegert and colleagues developed an algorithm to select an optimal forensic maker subset on the Y-chromosome using conditional entropy.[46] The algorithm grows a chain of makers by adding the marker with the highest entropy, conditional on all the other markers in the chain, until a threshold is met. Using this method, the authors were able to select marker sets with roughly half the markers of standard marker panels, without significant loss of discrimination capacity. Selected markers are still dependent, however, so the product rule cannot be applied to calculate Pr(A) for the reduced marker set.

1.2.3 *The Kappa Method*

Charles Brenner has been highly critical of SWGDAM's evolutionary model for calculating match probability.[47] At the heart of his criticism, Brenner doubts the accuracy of the estimate of

profile probability obtained from Y-haplotype databases. Brenner considers, non-zero Y-haplotype frequencies obtained from such databases to be overestimates of the true haplotype frequencies. Brenner also disagrees with the practice of using the sample Y-haplotype frequency as an estimate of the probability an unrelated, innocent man would carry the same haplotype. In frequentist statistics, the probability of an event is approximated as the relative frequency of that event occurring in a number of trials; as the number of trials approaches infinity, estimated and true probability converge. Brenner espouses a different philosophy of science as a follower of the philosophy of John Stuart Mill. Probability, under this philosophy, is the summary of the evidence generated by a conceptually repeatable experiment.[47] Both Mill and Brenner emphasize that probability must be based on the evidence at hand.[48] Brenner believes that the true Y-haplotype frequency may well be a fine estimate of the probability that an innocent person's haplotype matches a crime-scene haplotype, but if the sample Y-haplotype frequency is a poor estimate of the true frequency, then it is a poor estimate of probability. This is in contrast to a frequentist interpretation that the sample Y-haplotype frequency is an unbiased estimate of the true frequency.

To get around reliance on haplotype frequency, Brenner devised an alternate methodology called the "kappa method". This method focuses on addressing the problem that occurs when the crime scene haplotype is not observed in the population sample – therefore, haplotype frequency is estimated to be zero. He focuses on an observed property of Y-haplotype databases, augmented by the evidentiary profile: a large proportion of haplotypes are observed only once (they are singletons). The probability of an innocent person matching any one of these database singletons, regardless of type is identical when the database has been augmented by the crime scene haplotype. Had the evidentiary profile been observed in the database once, it would

also have this probability, so Brenner finds it quite reasonable to add the crime scene haplotype in to the database. The crime scene haplotype, is then, by definition, observed in the database and is a singleton. Then, α is the number of singletons in the database, including the crime scene haplotype. Kappa, κ, is proportion of singletons in the database. For an innocent person's haplotype (T) to match the crime scene haplotype (S0), using the database evidence, the innocent haplotype must match something in the database, must match a singleton, and finally, must match the crime scene haplotype. From these points of logic, Brenner constructs the $Pr(T = S_0)$ = Pr(Match|Singleton&Observed)*Pr(Singleton|Observed)*Pr(Observed). The Pr(Match|Singleton&Observed) is 1/ α. The Pr(Singleton|Observed) = κ. The Pr(Observed) is the probability T is not a new type. Brenner estimates the probability T is a new type as κ, so the Pr(Observed) is the complement 1- κ. Putting these together and simplifying gives:

$$Pr(T = S_0) = \frac{(1-\kappa)}{n}. \quad (1.5)$$

The probability of a match for non-singletons is calculated as (1- κ)/n multiplied by the popularity of the haplotype, i.e. the number of times it occurs in the database.

One limitation of the kappa method is that for the probability to truly reflect the match probability, the innocent person and the crime stain donor must be totally unrelated. However, cryptic relatedness is common in populations. In a sample of 1,000 Europeans after removing close relatives, 90% had a predicted 2nd to 9th cousin in the sample, most commonly 4th to 6th cousins.[49] Most people are unlikely to know a person related to this degree is a relative. Ignoring this background level of relatedness seems like too strong an assumption, particularly for Y-chromosome matching, where mutation is the only process breaking apart a haplotype within a male lineage over time.

1.2.4 *Population Structure Parameter*, Θ

To account for background relatedness, we employ the parameter Θ -- the correlation of alleles in different individuals within the same population.[50] Conceptually, Θ_i can be thought of as the probability that two alleles from population *i* are identical-by-descent (ibd).[36] When restricting this measure of coancestry between two individuals who come from the same ethnic subpopulation, Θ has been regarded as being equivalent to the parameter F_{ST}.[51]

F_{ST} was originally defined by Sewell Wright as the correlation of randomly drawn gametes from the same population, relative to the total population.[52] F_{ST} was originally estimated by Wright as the ratio of the variance in the allele frequency within a subpopulation to the variance of the allele frequency in the total population:

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2} = \frac{\sigma_S^2}{\bar{p}(1-\bar{p})}, \tag{1.6}$$

where $\bar{p}$ is the population allele frequency.

Several estimators have been proposed, employing one of three methods: method-of-moments, maximum likelihood, or Bayesian analysis.[53] F_{ST} estimates depend on the choice of estimator, the genetic markers used in analysis, and how estimates are combined across markers.[54] Bhatia and colleagues recommend using the Hudson estimator, which is the average of population-specific estimates produced by the Weir-Hill estimator.

The Weir-Hill estimator allows for the calculation of population-specific estimates of F_{ST}.[55] The moment estimator for a single population is:

$$\hat{\beta}_i = 1 - \frac{(\sum_{i=1}^{N} n_{ic}) \sum_{A=1}^{m} \frac{n_i}{n_i - 1} p_{Ai}(1-p_{Ai})}{\sum_{i=1}^{r} \sum_{A=1}^{m} [n_i (p_{Ai} - \bar{p}_A)^2 + n_{ic}(1-p_{Ai})]}, \tag{1.7}$$

where p_{Ai} is the frequency allele A in population i, $\bar{p}_A$ is the mean frequency of allele A over all populations, n_i is the sample size (in terms of alleles) in population i, r is the number of populations, m is the number of alleles at the locus, and

$$n_{ic} = \frac{n_i - n_i^2}{\sum_{i=1}^{r} n_i}. \quad (1.8)$$

A new estimator, developed by Buckelton and colleagues, formulates the population-specific F_{ST} estimate to make explicit that within-population estimates are relative to between-population pair estimates, highlighting allele sharing across populations due to migration and admixture.[56] This estimator is conceptualized as allelic matching within a population (sample homozygosity) relative to average matching between populations:

$$\hat{\beta}_{il} = \frac{(\tilde{M}_{il} - \tilde{M}_{Bl})}{(1 - \tilde{M}_{Bl})}, \quad (1.9)$$

where, for large sample sizes,

$$\tilde{M}_{il} = \sum_A \tilde{p}_{Ail}^2, \quad (1.10)$$

$$\tilde{M}_{Bl} = \sum_{i \neq i'} \frac{\sum_A \tilde{p}_{Ail} \tilde{p}_{Ai'l}}{r(1-r)}, \quad (1.11)$$

and $\tilde{p}_{Ail}$ is the frequency of allele A at locus l in population i, and r is the number of populations.

Because this estimate is relative to the total matching between populations, be they of the same continental ancestry or worldwide, $\hat{\beta}_{il}$ can be negative for populations where there is less matching of alleles within the population than between populations. This might be expected within African populations – these populations are older (on an evolutionary scale) and tend to

exhibit more genetic diversity and therefore less matching when compared to younger populations.[57,58]

1.3 A Compendium of Worldwide Surveys of Population-Specific F_{ST} for Forensic Y-STR Markers

Extending the work of Buckleton and colleagues, this compendium of surveys will examine the empirical properties of population- and locus-specific estimate of the coancestry parameter θ -- $\tilde{\beta}_{il}$ – for the Y-chromosome in three large, diverse, Y-STR datasets: the Y-Chromosome Haplotype Reference Database (YHRD),[43] Human Genome Diversity Project (HGDP),[59] and data published by Xu et al (XU).[60] Some attention will also be paid to calculation of the profile probability and match calculation, comparing Brenner's kappa method and the evolutionary model using our calculated F_{ST}.

1.3.1 *Methods: Datasets*

Because about seven percent of samples were missing data, the DYS481, DYS533, DYS549, and DYS643 loci were excluded from the HGDP dataset. Individuals not scored at any one of the loci included for analysis in each dataset were excluded. Finally, populations with fewer than ten individuals were excluded from analysis.

We used the regional ethnicities of populations assigned by the YHRD and HGDP researchers for these datasets, with one exception – one Oceanian population remained after excluding individuals and populations as detailed above; this population was re-categorized as "Asian". The populations in the XU dataset were assigned to regional ethnicity to best match the regional ethnicity assignments of the YHRD and HGDP data, using our best judgement.

The three datasets were analyzed separately to facilitate a comparison of the behavior of empirical estimates calculated for different samples. The characteristics of each dataset are reported in Table 1.1.

Table 1.1. Characteristics of the Y-STR Datasets

Region	Number of populations	Sample size Minimum	Maximum
YHRD			
Africa	10	38	509
America, Native	5	56	152
America, Admixed	10	44	403
Asia	28	30	629
Europe	76	30	2085
HGDP			
Africa	6	11	23
Asia	12	12	25
Europe	3	14	16
Middle East	3	11	25
XU			
Africa	8	12	35
America, Native	6	11	24
Asia	10	19	53
Europe	11	10	49
Middle East	4	18	40

1.3.2 *Methods: Population-specific β Estimates*

To emphasize that a population-specific estimate of θ can be calculated only relative to a total, our estimator, $\tilde{\beta}_{il}$, is written as one minus the heterozygosity within population i at locus l, divided by the average heterozygosity between all populations in the total sample,

$$\tilde{\beta}_{il} = 1 - \frac{\tilde{H}_{il}}{\tilde{H}_{Bl}}. \quad (1.12)$$

The heterozygosity within population i at locus l is in turn calculated using the sample allele frequencies for allele A at locus l,

$$\widetilde{H}_{il} = \frac{n_{il}}{n_{il}-1}\sum_A \tilde{p}_{Ail}(1-\tilde{p}_{Ail}) = \frac{n_{il}}{n_{il}-1}(1-\sum_A \tilde{p}_{Ail}^2)\,. \quad (1.13)$$

Heterozygosity between populations i and i' at locus l is one minus the sum of the products of sample allele frequencies for each pair of populations,

$$\widetilde{H}_{ii'l} = 1 - \sum_A \tilde{p}_{Ail}\,\tilde{p}_{Ai'l}. \quad (1.14)$$

The average heterozygosity is then calculated as

$$\widetilde{H}_{Bl} = \frac{1}{r(r-1)}\sum_{i\neq i'} \widetilde{H}_{ii'l}, \quad (1.15)$$

where r is the number of total populations.

Conventional F_{ST} is the average of the population specific $\tilde{\beta}_{il}$ estimates, here denoted as $\tilde{\beta}_{Wl}$, and can be calculated as one minus the average heterozygosity within populations over the average heterozygosity between populations:

$$F_{ST} = \tilde{\beta}_{Wl} = 1 - \frac{\widetilde{H}_{Wl}}{\widetilde{H}_{Bl}}, \quad (1.16)$$

where

$$\widetilde{H}_{Wl} = \frac{1}{r}\sum_{i=1}^{r} \widetilde{H}_{il}. \quad (1.17)$$

Total population can be conceptualized in different ways. To assess the impact of different conceptualizations of total population, we calculated β estimates relative to all the populations within a dataset (β_{PT}) and relative to populations within the same region (β_{PR}).

Single locus β estimates were compared using an ANOVA analysis, adjusted for locus, dataset, regional ethnicity, and their pairwise interactions, restricted to the region common between datasets (Africa, Asia, and Europe). We also examined the effect of dataset and region more specifically on locus-specific β estimates by use of a linear regression analysis, stratified by locus. Adjusted for region, we also examined the relationship of mutation rate on β estimates, stratified by dataset.

Because markers on the Y-chromosome are assumed to be linked, multi-marker haplotypes are considered a single locus. B_{il}'s were calculated using the above equations for the population in each of the three datasets for l = $\binom{n}{k}$ haplotype combinations of Y-STR loci, where n=19 (for YHRD and HGDP datasets) and n = 16 (for the XU dataset) and k = 1, 2, …, 13, in a custom C++ script. B_{Wl}s were calculated for each haplotype within regional ethnicity.

To assess the dependency between locus-specific β_{il} estimates, Pearson correlation coefficients were generated for each pair of single locus β_{il} values for all populations within a dataset and subset by regional ancestry within a dataset. Linkage disequilibrium between Y-STR loci was calculated using the 'gap' R package.[61] The linkage disequilibrium parameter ρ and β_{il} correlation matrices were compared using a Mantel test of correlation between two matrices as implemented in the 'ade4' R package, using 9,999 replications.[62,63]

1.3.3 *Methods: Marker Information and Selection*

Y-STR marker informativeness was assessed using the information measure Shannon entropy, B (Equation (1.4)). Following the algorithm developed by Siegert et al, the most informative subset of markers was chosen by growing a chain of markers according to the maximum entropy of the next added marker, conditional on the entropy of the preceding markers. Markers were added until conditional entropy could no longer be maximized (i.e. conditional entropy = 0).

We assessed the relationship between marker order of addition and published mutation rate, single locus entropy, and genetic diversity ($\frac{n}{n-1}1-\sum p_{Ail}^2$) using a linear regression, stratified by dataset and adjusted for regional ethnicity. We restricted analysis to only the three regional ethnicities common between datasets. The analysis used the rank of each variable rather than value. Genetic diversity and single locus entropy were ranked separately for each ethnicity in each dataset.

To assess dependencies between Y-STR markers, in addition to linkage disequilibrium, we calculated the shared information distance of each pair of markers as the sum of the conditional entropies of each marker divided by the joint entropy of both markers:

$$D(X,Y) = \frac{B(X|Y)+B(Y|X)}{B(X,Y)}. \quad (1.18)$$

A shared information distance of one means markers are completely independent, and zero means they are completely dependent.

1.3.4 *Methods: Comparison of Kappa and Evolutionary Matching*

We constructed haplotypes based on the 12 PowerPlex Y loci to estimate the frequency of haplotype popularity and calculate match probability estimates.[10] Match probabilities were calculated using the recommendation of the Scientific Working Group on DNA Analysis Methods (SWGDAM) using the equation $\Pr(A|A) = \theta + (1-\theta)*\Pr(A)$ where $\Pr(A)$ is the profile probability and θ is our region-specific β_W estimate, calculated relative other populations in the total sample ($\beta_{W,PT}$).[37] Match probabilities were also calculated using the kappa method using the equation $p(1-\kappa)/n$, where p is the popularity of the haplotype, κ is the frequency of singletons in the database and n is the number of unique haplotypes.[47]

1.3.5 *Results: Single Locus β_i Estimates*

Median worldwide $\beta_{il,PT}$ estimates ranged from -0.01 to 0.14 for the YHRD dataset, 0.02 to 0.18 for the HGDP dataset, and 0.04 to 0.26 for the XU dataset (Figure 1.1). A significant locus effect on $\beta_{il,PT}$ estimates remained after adjusting for regional ancestry, database, their interaction, and the interactions of region and database with locus, which were themselves highly significant (Table 1.2).

For the YHRD data, the largest median β_i values among markers were seen for DYS392, DYS393, DYS438, and DYS437. As expected, these markers also have the lowest mutation rates. The smallest median β_i estimates were seen for markers DYS576, DYS570, and DYS458. These three loci also had less variation in β_i estimates and had the highest mutation rates. In a regression analysis of $\beta_{il,PT}$ estimates, stratified by database, higher mutation rate was significantly associated with smaller β_i estimates in all three datasets (YHRD, p=2x10-16; HGDP, p=0.02; Xu et al, p=0.003).

Figure 1.2 examines $\beta_{il,PT}$ estimate distribution by regional ancestry. In a regression analysis stratified by locus, African β_i estimates were significantly larger than European estimates for markers DYS390, DYS391, DYS392, DYS437, DYS438, and DYS456 (Table 1.3). African β_i estimates were significantly smaller than European estimates for DYS393 and DYS389II.I. Asian estimates were larger than European estimates for markers DYS391, DYS437, DYS438, and DYS456 and smaller for DYS385ab, DYS389I, II, and II.I, DYS393, DYS448, DYS570 and DYS635. A database effect was observed for about half the markers.

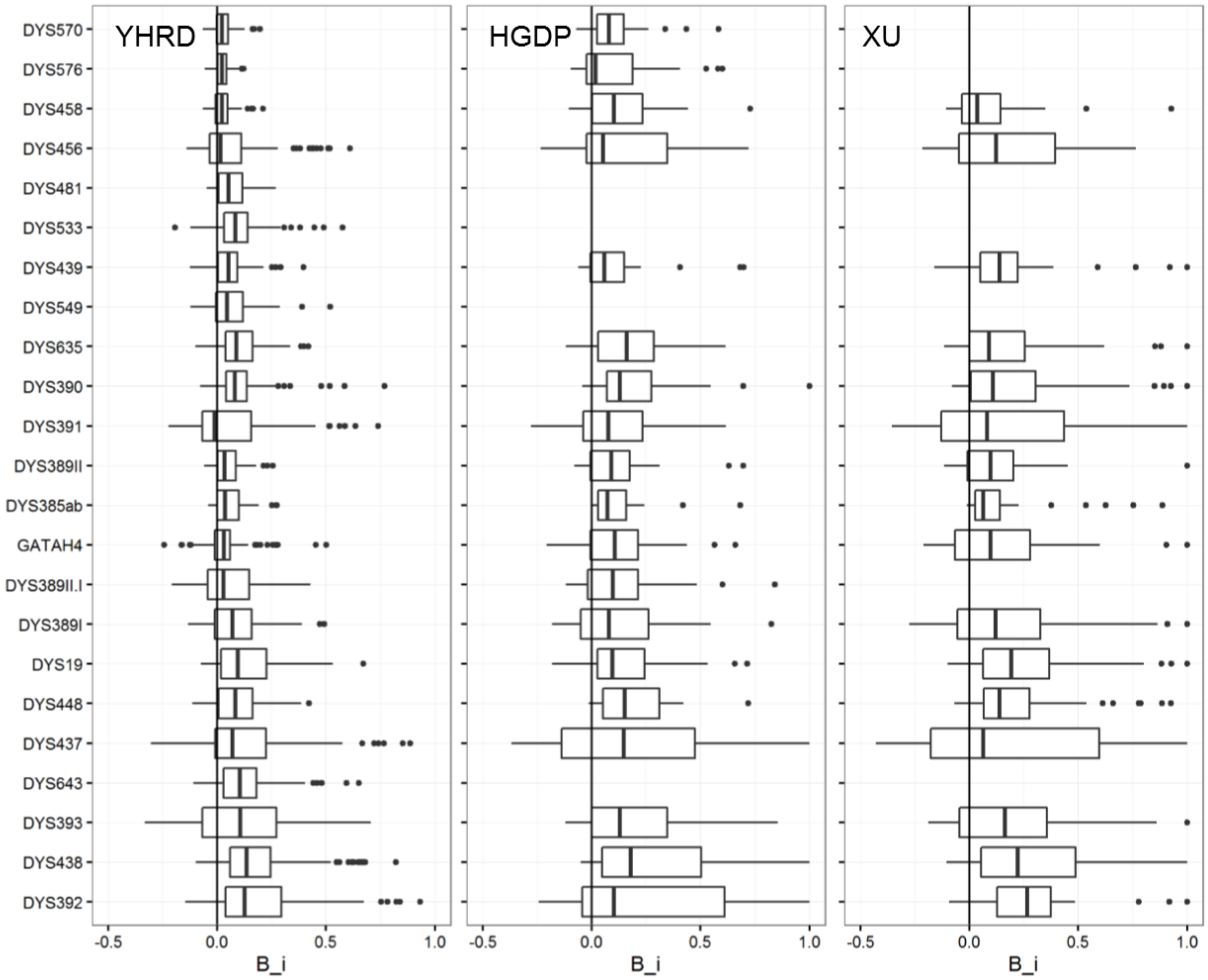

Figure 1.1. Single locus β_i's. Y-STR loci are ordered by published mutation rate.

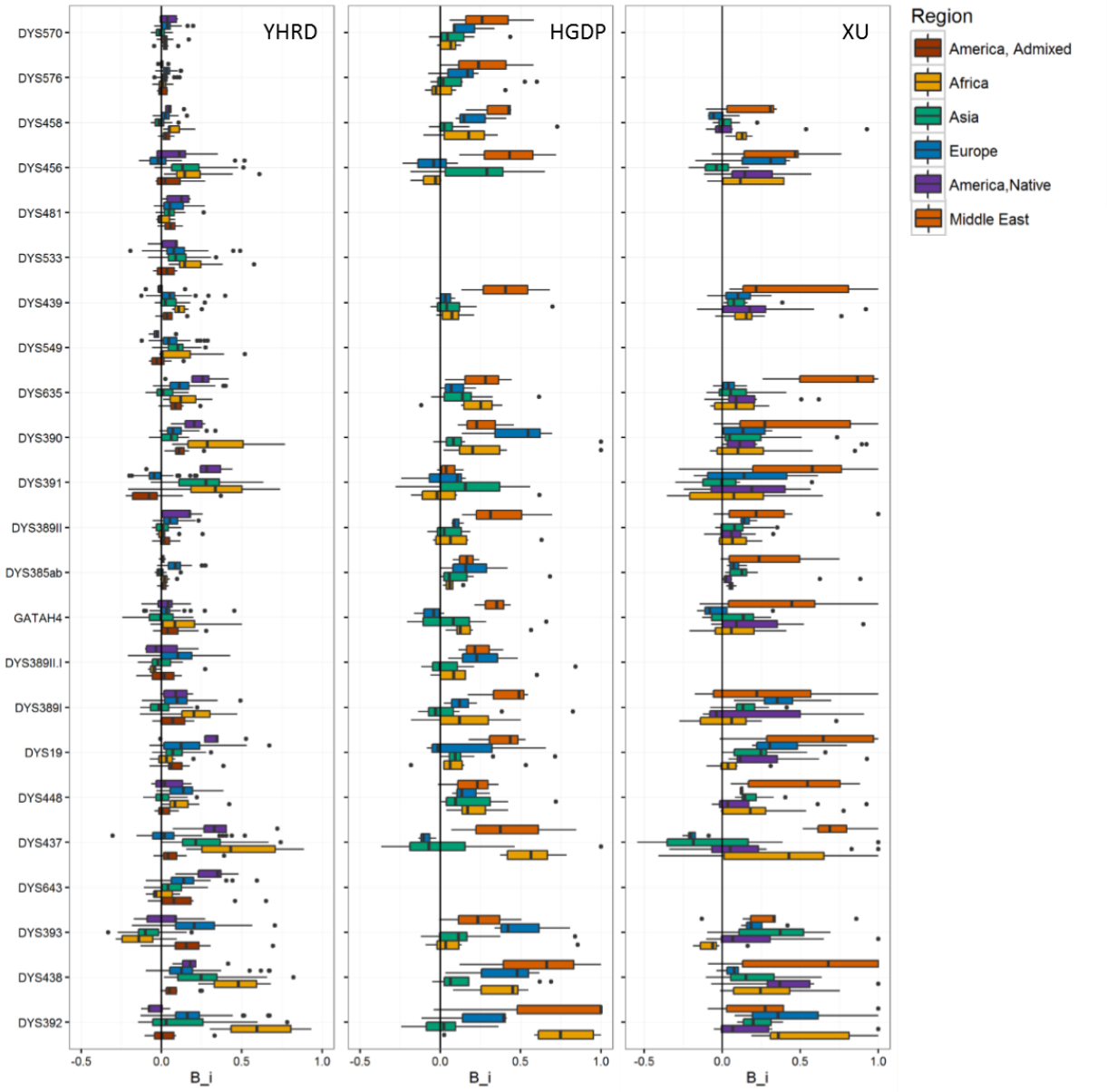

Figure 1.2. Single locus β_i's by regional ethnicity. Y-STR loci are ordered by published mutation rate.

Table 1.2. ANOVA of single locus $\beta_{il,PT}$ estimates, by Locus, Region, Database, and their pairwise interactions.

Source	Df	SS	MS	F	Pr(>F)	
Locus	22.000	9.230	0.4196	17.53	<2.00E-16	***
Region	2.000	1.990	0.9934	41.51	<2.00E-16	***
Database	2.000	1.690	0.8448	35.30	6.75E-16	***
Region*Database	4.000	1.120	0.2792	11.67	2.07E-09	***
Locus*Region	44.000	15.280	0.3473	14.51	<2.00E-16	***
Locus*Database	33.000	2.900	0.088	3.68	9.26E-12	***
Residuals	3358.000	80.370	0.0239			

Df = degrees of freedom, SS = sum of squares, MS = mean squared error, F = f-statistic

Table 1.3. Regression analysis of single locus $\beta_{il,PT}$ estimates by stratified by locus, adjusted by regional ancestry, database, and their pairwise interactions.

	Estimate	Std.Error	t	Pr(>|t|)	
DYS19					
(Intercept)	0.14841	0.01989	7.464	5.80E-12	***
HGDP	0.03765	0.10204	0.369	0.7126	
XU	0.09173	0.05592	1.64	0.103	
Africa	-0.11241	0.05832	-1.928	0.0557	.
Asia	-0.0576	0.03832	-1.503	0.1349	
HGDP*Africa	0.03355	0.13575	0.247	0.8051	
XU*Africa	-0.07116	0.09944	-0.716	0.4754	
HGDP*Asia	0.02983	0.11924	0.25	0.8028	
XU*Asia	0.0933	0.08489	1.099	0.2735	
DYS385ab					
(Intercept)	0.083195	0.012273	6.779	2.43E-10	***
HGDP	0.110314	0.06298	1.752	0.08184	.
XU	0.024737	0.034516	0.717	0.474645	
Africa	-0.060282	0.035992	-1.675	0.095982	.
Asia	-0.09107	0.023653	-3.85	0.000173	***
HGDP*Africa	-0.068143	0.083781	-0.813	0.417274	
XU*Africa	0.007931	0.061376	0.129	0.897349	
HGDP*Asia	0.033389	0.073594	0.454	0.650688	
XU*Asia	0.152667	0.052392	2.914	0.004101	**
DYS389I					
(Intercept)	0.09526	0.02106	4.524	1.21E-05	***
HGDP	0.02849	0.10806	0.264	0.7924	
XU	0.05377	0.05922	0.908	0.3653	
Africa	0.09928	0.06175	1.608	0.1099	
Asia	-0.09672	0.04058	-2.383	0.0184	*
HGDP*Africa	-0.07663	0.14375	-0.533	0.5947	
XU*Africa	-0.17037	0.10531	-1.618	0.1077	
HGDP*Asia	0.063	0.12627	0.499	0.6185	
XU*Asia	0.15181	0.08989	1.689	0.0933	.

	Estimate	Std.Error	t	Pr(>|t|)	
DYS389II					
(Intercept)	0.059298	0.010173	5.829	3.17E-08	***
HGDP	0.041466	0.052205	0.794	0.4282	
XU	0.041132	0.02861	1.438	0.1526	
Africa	-0.026026	0.029834	-0.872	0.3844	
Asia	-0.049959	0.019606	-2.548	0.0118	*
HGDP*Africa	0.066668	0.069447	0.96	0.3386	
XU*Africa	0.009063	0.050876	0.178	0.8588	
HGDP*Asia	0.014759	0.061003	0.242	0.8091	
XU*Asia	0.023028	0.043429	0.53	0.5967	
DYS389II.I					
(Intercept)	0.10686	0.01664	6.422	2.41E-09	***
HGDP	0.14611	0.08539	1.711	0.089485	.
Africa	-0.12707	0.0488	-2.604	0.010304	*
Asia	-0.11999	0.03207	-3.742	0.000275	***
HGDP*Africa	0.01696	0.11359	0.149	0.881545	
HGDP*Asia	-0.03799	0.09978	-0.381	0.704067	
DYS390					
(Intercept)	0.08806	0.0209	4.213	4.28E-05	***
HGDP	0.37101	0.10727	3.458	0.000703	***
XU	0.08941	0.05879	1.521	0.130363	
Africa	0.25632	0.0613	4.181	4.85E-05	***
Asia	-0.03619	0.04029	-0.898	0.370434	
HGDP*Africa	-0.39191	0.1427	-2.746	0.006745	**
XU*Africa	-0.22977	0.10454	-2.198	0.029452	*
HGDP*Asia	-0.25869	0.12535	-2.064	0.040726	*
XU*Asia	0.10961	0.08924	1.228	0.221237	

Table 1.3. Regression analysis of single locus $\beta_{il,PT}$ estimates, continued.

	Estimate	Std.Error	t	Pr(>|t|)	
DYS391					
(Intercept)	-0.03556	0.02154	-1.651	0.10085	
HGDP	0.04292	0.11054	0.388	0.69832	
XU	0.04517	0.06058	0.746	0.45706	
Africa	0.37066	0.06317	5.868	2.61E-08	***
Asia	0.282	0.04151	6.793	2.25E-10	***
HGDP*Africa	-0.31418	0.14704	-2.137	0.03421	*
XU*Africa	-0.30876	0.10772	-2.866	0.00473	**
HGDP*Asia	-0.15238	0.12916	-1.18	0.23994	
XU*Asia	-0.12181	0.09195	-1.325	0.18722	
DYS392					
(Intercept)	0.18566	0.02355	7.885	5.38E-13	***
HGDP	0.0443	0.12083	0.367	0.714	
XU	0.04111	0.06622	0.621	0.536	
Africa	0.42262	0.06905	6.12	7.42E-09	***
Asia	-0.0604	0.04538	-1.331	0.185	
HGDP*Africa	0.03182	0.16074	0.198	0.843	
XU*Africa	-0.1233	0.11775	-1.047	0.297	
HGDP*Asia	-0.16538	0.14119	-1.171	0.243	
XU*Asia	0.02902	0.10052	0.289	0.773	
DYS393					
(Intercept)	0.20942	0.02326	9.003	7.80E-16	***
HGDP	0.31285	0.11936	2.621	0.00965	**
XU	0.09988	0.06542	1.527	0.12886	
Africa	-0.34709	0.06821	-5.088	1.04E-06	***
Asia	-0.28965	0.04483	-6.461	1.29E-09	***
HGDP*Africa	-0.01944	0.15879	-0.122	0.90272	
XU*Africa	-0.02493	0.11632	-0.214	0.8306	
HGDP*Asia	-0.08907	0.13948	-0.639	0.52402	
XU*Asia	0.19865	0.0993	2.001	0.04719	*

	Estimate	Std.Error	t	Pr(>|t|)	
DYS437					
(Intercept)	0.03654	0.02951	1.238	0.218	
HGDP	-0.12594	0.15144	-0.832	0.407	
XU	-0.0785	0.083	-0.946	0.346	
Africa	0.44054	0.08655	5.09	1.03E-06	***
Asia	0.22898	0.05688	4.026	8.88E-05	***
HGDP*Africa	0.20847	0.20146	1.035	0.302	
XU*Africa	-0.03707	0.14759	-0.251	0.802	
HGDP*Asia	-0.05612	0.17696	-0.317	0.752	
XU*Asia	-0.01617	0.12598	-0.128	0.898	
DYS438					
(Intercept)	0.14411	0.02327	6.193	5.13E-09	***
HGDP	0.23198	0.11941	1.943	0.05387	.
XU	0.05729	0.06544	0.876	0.38266	
Africa	0.31229	0.06824	4.576	9.68E-06	***
Asia	0.11726	0.04485	2.615	0.00982	**
HGDP*Africa	-0.31964	0.15885	-2.012	0.04594	*
XU*Africa	-0.21694	0.11637	-1.864	0.06418	.
HGDP*Asia	-0.33089	0.13953	-2.371	0.01896	*
XU*Asia	0.11204	0.09933	1.128	0.26111	
DYS439					
(Intercept)	0.054247	0.015112	3.59	0.000445	***
HGDP	-0.021389	0.077549	-0.276	0.783064	
XU	0.052152	0.0425	1.227	0.221652	
Africa	0.066605	0.044317	1.503	0.134908	
Asia	-0.004951	0.029125	-0.17	0.865226	
HGDP*Africa	-0.024228	0.103161	-0.235	0.814634	
XU*Africa	0.026662	0.075574	0.353	0.72472	
HGDP*Asia	0.076938	0.090618	0.849	0.397179	
XU*Asia	0.12323	0.064511	1.91	0.057965	.

Table 1.3. Regression analysis of single locus $\beta_{il,PT}$ estimates, continued.

	Estimate	Std.Error	t	Pr(>|t|)	
DYS448					
(Intercept)	0.133968	0.016743	8.001	2.76E-13	***
HGDP	0.038383	0.085919	0.447	0.6557	
XU	0.054494	0.047087	1.157	0.24894	
Africa	-0.007029	0.0491	-0.143	0.88635	
Asia	-0.11825	0.032268	-3.665	0.00034	***
HGDP*Africa	0.041966	0.114295	0.367	0.714	
XU*Africa	0.051422	0.08373	0.614	0.54003	
HGDP*Asia	0.164531	0.100398	1.639	0.1033	
XU*Asia	0.1095	0.071474	1.532	0.12757	
DYS456					
(Intercept)	-0.009102	0.016475	-0.552	0.5814	
HGDP	-0.045706	0.084544	-0.541	0.5895	
XU	-0.017394	0.046333	-0.375	0.7079	
Africa	0.218622	0.048314	4.525	1.20E-05	***
Asia	0.188509	0.031752	5.937	1.86E-08	***
HGDP*Africa	-0.225041	0.112466	-2.001	0.0472	*
XU*Africa	-0.030095	0.08239	-0.365	0.7154	
HGDP*Asia	0.140038	0.098791	1.418	0.1584	
XU*Asia	0.025255	0.07033	0.359	0.72	
DYS458					
(Intercept)	0.027556	0.013085	2.106	0.0368	*
HGDP	0.190789	0.067147	2.841	0.0051	**
XU	0.003379	0.036799	0.092	0.927	
Africa	0.052422	0.038373	1.366	0.1739	
Asia	-0.036087	0.025218	-1.431	0.1544	
HGDP*Africa	-0.12262	0.089323	-1.373	0.1718	
XU*Africa	0.036098	0.065437	0.552	0.582	
HGDP*Asia	-0.079957	0.078462	-1.019	0.3098	
XU*Asia	0.132999	0.055858	2.381	0.0185	*

	Estimate	Std.Error	t	Pr(>|t|)	
DYS570					
(Intercept)	0.036461	0.007045	5.176	8.55E-07	***
HGDP	0.130255	0.03615	3.603	0.000449	***
Africa	-0.003082	0.020659	-0.149	0.881633	
Asia	-0.042371	0.013577	-3.121	0.002229	**
HGDP*Africa	-0.109893	0.04809	-2.285	0.023948	*
HGDP*Asia	-0.04322	0.042243	-1.023	0.308176	
DYS576					
(Intercept)	0.032803	0.009366	3.502	0.000636	***
HGDP	0.078507	0.048065	1.633	0.10485	
Africa	-0.025642	0.027468	-0.934	0.352299	
Asia	-0.013055	0.018051	-0.723	0.470883	
HGDP*Africa	-0.035831	0.063939	-0.56	0.576188	
HGDP*Asia	0.030124	0.056165	0.536	0.592648	
DYS635					
(Intercept)	0.12519	0.01394	8.978	9.09E-16	***
HGDP	-0.02968	0.07156	-0.415	0.678943	
XU	-0.04168	0.03922	-1.063	0.289498	
Africa	0.01841	0.04089	0.45	0.653125	
Asia	-0.10559	0.02688	-3.929	0.000128	***
HGDP*Africa	0.09068	0.09519	0.953	0.342299	
XU*Africa	-0.01228	0.06974	-0.176	0.860418	
HGDP*Asia	0.16499	0.08362	1.973	0.050265	.
XU*Asia	0.18596	0.05953	3.124	0.002133	**
GATAH4					
(Intercept)	0.03254	0.01646	1.977	0.0498	*
HGDP	-0.09253	0.08446	-1.096	0.275	
XU	0.05293	0.04629	1.143	0.2546	
Africa	0.08729	0.04827	1.809	0.0725	.
Asia	-0.03092	0.03172	-0.975	0.3313	
HGDP*Africa	0.16453	0.11236	1.464	0.1451	
XU*Africa	-0.09708	0.08231	-1.179	0.24	
HGDP*Asia	0.16543	0.0987	1.676	0.0957	.
XU*Asia	0.15863	0.07026	2.258	0.0254	*

1.3.6 *Results: β estimate dependencies*

Because Y-STR loci are linked, we hypothesized that there would also be substantial dependency between pairs of single-locus β_{il} estimates. Figure 1.3 through Figure 1.5 show the correlation between single-locus β_{il} estimates (upper triangle) alongside the linkage disequilibrium parameter ρ (lower triangle) for all populations in each of the datasets. Cells are shaded where the values are statistically significant.

Considering the worldwide β_{il} estimates, we observed low to moderate correlation for all three datasets (Table 1.4). About half of the correlations were statistically significant for the YHRD dataset and HGDP dataset. None of the correlations were significant in the XU dataset. Though there appears to be more correlation in the YHRD dataset, due to the larger sample size, weaker correlations are statistically significant.

Evidence for linkage disequilibrium between markers was nearly absent for all three datasets. Comparing the β_{il} correlations to ρ using a Mantel test, there appears to be a weak relationship between β_{il} dependencies and LD, though this was statistically significant only for the YHRD dataset (Table 1.5).

Table 1.4. Mean, standard deviation and range of correlation between single locus β_{il} estimates, by region, among the YHRD, HGDP, and XU datasets.

	YHRD			HGDP			XU		
Region	N pops	μ ± sd	range	N pops	μ ± sd	range	N pops	μ ± sd	range
World	129	0.18 ± 0.21	-0.39 - 0.67	24	0.40 ± 0.22	-0.35 – 0.75	39	0.44 ± 0.17	-0.02 – 0.73
Africa	10	0.29 ± 0.41	-0.83 – 0.92	6	0.19 ± 0.45	-0.94 – 0.99	8	0.15 ± 0.46	-0.78 – 0.91
Europe	76	0.22 ± 0.24	-0.38 – 0.80	3	0.37 ± 0.61	-0.63 – 0.95	11	0.10 ± 0.34	-0.61 – 0.82
Asia	28	0.12 ± 0.26	-0.49 – 0.84	12	0.48 ± 0.31	-0.99 – 0.99	11	0.67 ± 0.27	-0.32 – 0.98
America, Native	5	0.10 ± 0.51	-0.98 – 0.99	--	--	--	6	0.43 ± 0.42	-0.67 – 0.98
America, Admixed	10	0.18 ± 0.42	-0.76 – 0.91	--	--	--		--	--
Middle East	--	--	--	3	0.48 ± 0.57	-0.99 – 0.99	4	0.03 ± 0.59	-0.95 – 0.99

Table 1.5. Relationship between correlation of β_{il} estimates and linkage disequalibrium by region, among the YHRD, HGDP, and XU datasets.

	YHRD		HGDP		XU	
Region	Mantel r	p-value	Mantel r	p-value	Mantel r	p-value
World	0.20	0.01	0.14	0.17	0.25	0.09
Africa	0.30	0.01	0.21	0.04	0.47	0.002
America, Native	0.16	0.04	--	--	0.23	0.07
America, Admixed	0.22	0.03	--	--	--	--
Asia	0.10	0.08	-0.06	0.63	0.06	0.36
Europe	0.32	0.001	0.25	0.04	0.10	0.29
Middle East	--	--	-0.08	0.70	0.38	0.02

Upon stratifying by regional ancestry, we observed more LD among the European, African, and Middle Eastern populations in the HGDP dataset. Patterns of correlation remained similar to the unstratified correlation matrix (Figure 1.6 - Figure 1.19).

Overall, the mean and standard deviations of the correlations between β_{il} estimates suggest no dependence between β_{il} estimates. This may be misleading, as there are many negative correlations, some quite strong, driving the mean to zero. Negative correlations

also imply dependence between β_{il} estimates, but that the estimates are inversely related, i.e. one locus' estimates are large and the other's are small. While positively correlated β_{il} estimates represent that the two loci may contain some redundant information – estimates should be similar in magnitude, depending on correlation – negatively correlated estimates are dissimilar and should provide different information about β_{il}.

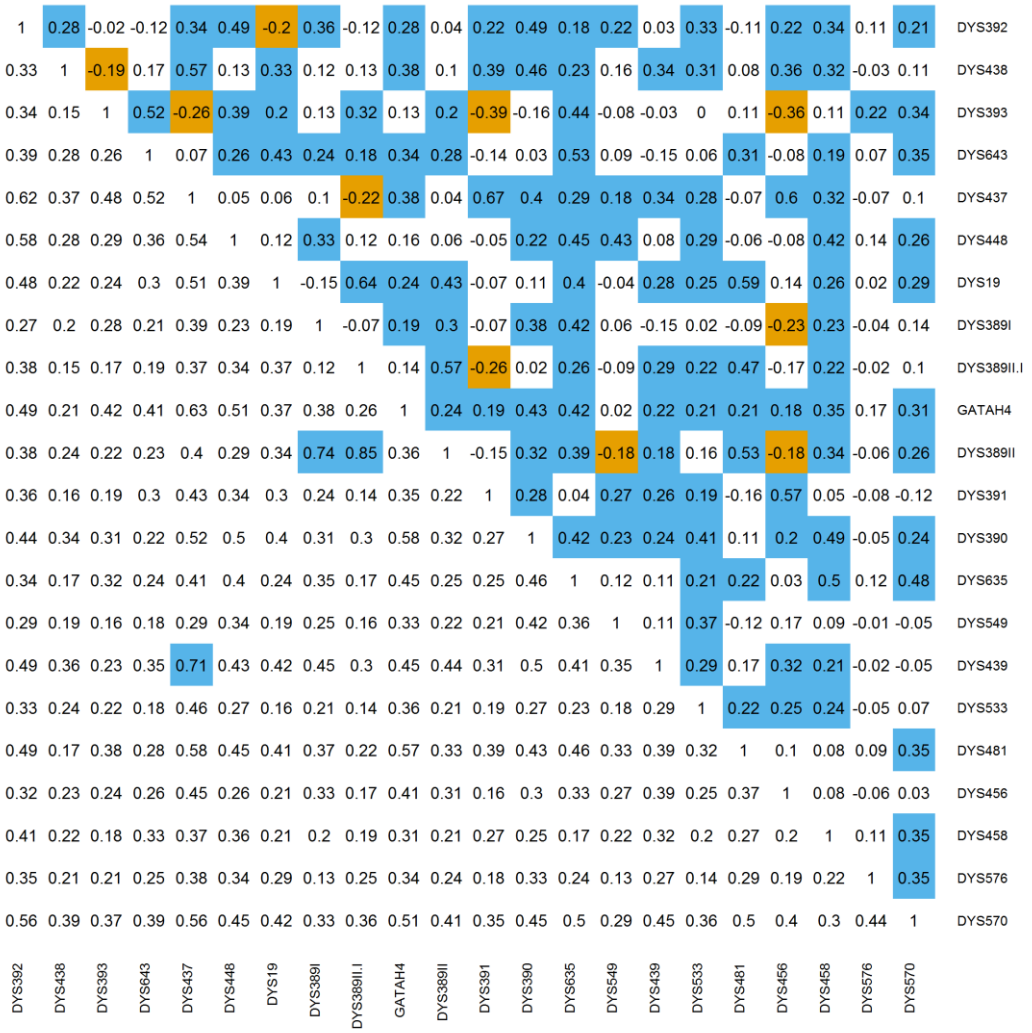

Figure 1.3. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. YHRD dataset, all populations.

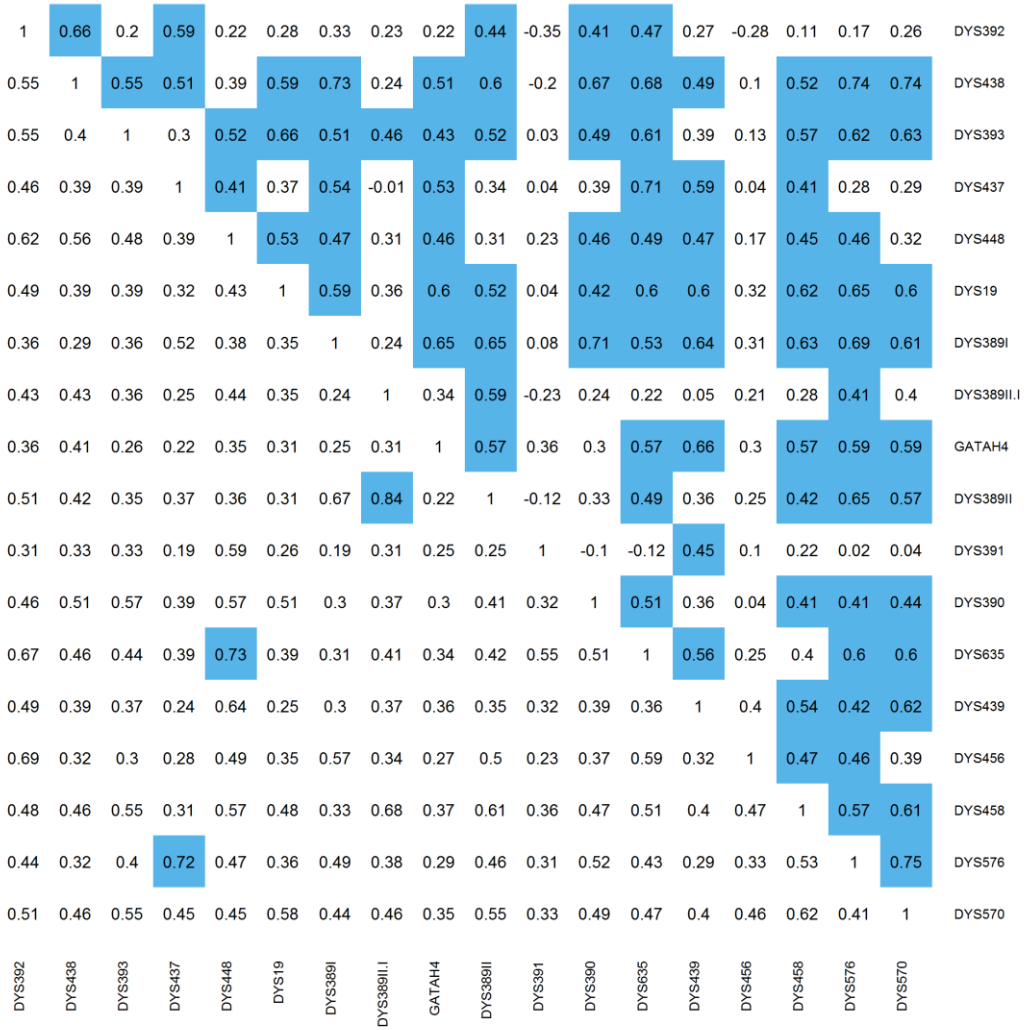

Figure 1.4. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. HGDP dataset, all populations.

1	0.45	0.01	0.27	0.54	0.38	0.52	0.34	0.4	0.2	0.58	0.28	0.34	0.26	0.46	DYS392
0.53	1	0	0.52	0.65	0.55	0.62	0.48	0.57	0.33	0.65	0.55	0.51	0.38	0.5	DYS438
0.36	0.41	1	0.16	0.29	0.43	0.28	0.41	0.17	0.21	0.41	0.25	0.18	-0.02	0.36	DYS393
0.36	0.45	0.32	1	0.56	0.4	0.32	0.37	0.35	0.55	0.49	0.65	0.32	0.35	0.53	DYS437
0.59	0.45	0.39	0.32	1	0.7	0.58	0.54	0.55	0.47	0.61	0.69	0.51	0.38	0.73	DYS448
0.62	0.49	0.32	0.39	0.48	1	0.59	0.56	0.62	0.37	0.49	0.73	0.38	0.35	0.6	DYS19
0.35	0.34	0.31	0.33	0.28	0.28	1	0.48	0.65	0.44	0.69	0.45	0.54	0.15	0.51	DYS389I
0.32	0.25	0.21	0.18	0.29	0.22	0.19	1	0.46	0.3	0.55	0.61	0.58	0.25	0.7	GATAH4
0.68	0.43	0.36	0.34	0.38	0.29	0.67	0.23	1	0.14	0.56	0.62	0.63	0.31	0.44	DYS389II
0.32	0.34	0.36	0.23	0.35	0.29	0.25	0.19	0.25	1	0.44	0.51	0.26	0.05	0.38	DYS391
0.52	0.51	0.42	0.38	0.56	0.49	0.31	0.17	0.37	0.34	1	0.54	0.63	0.39	0.59	DYS390
0.49	0.55	0.41	0.43	0.46	0.5	0.37	0.25	0.35	0.49	0.47	1	0.54	0.42	0.61	DYS635
0.39	0.36	0.3	0.25	0.31	0.36	0.31	0.25	0.35	0.31	0.44	0.27	1	0.34	0.67	DYS439
0.34	0.37	0.29	0.17	0.3	0.35	0.31	0.21	0.26	0.22	0.35	0.36	0.37	1	0.23	DYS456
0.41	0.41	0.38	0.28	0.37	0.55	0.31	0.31	0.28	0.2	0.33	0.42	0.24	0.33	1	DYS458
DYS392	DYS438	DYS393	DYS437	DYS448	DYS19	DYS389I	GATAH4	DYS389II	DYS391	DYS390	DYS635	DYS439	DYS456	DYS458	

Figure 1.5. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. XU dataset, all populations.

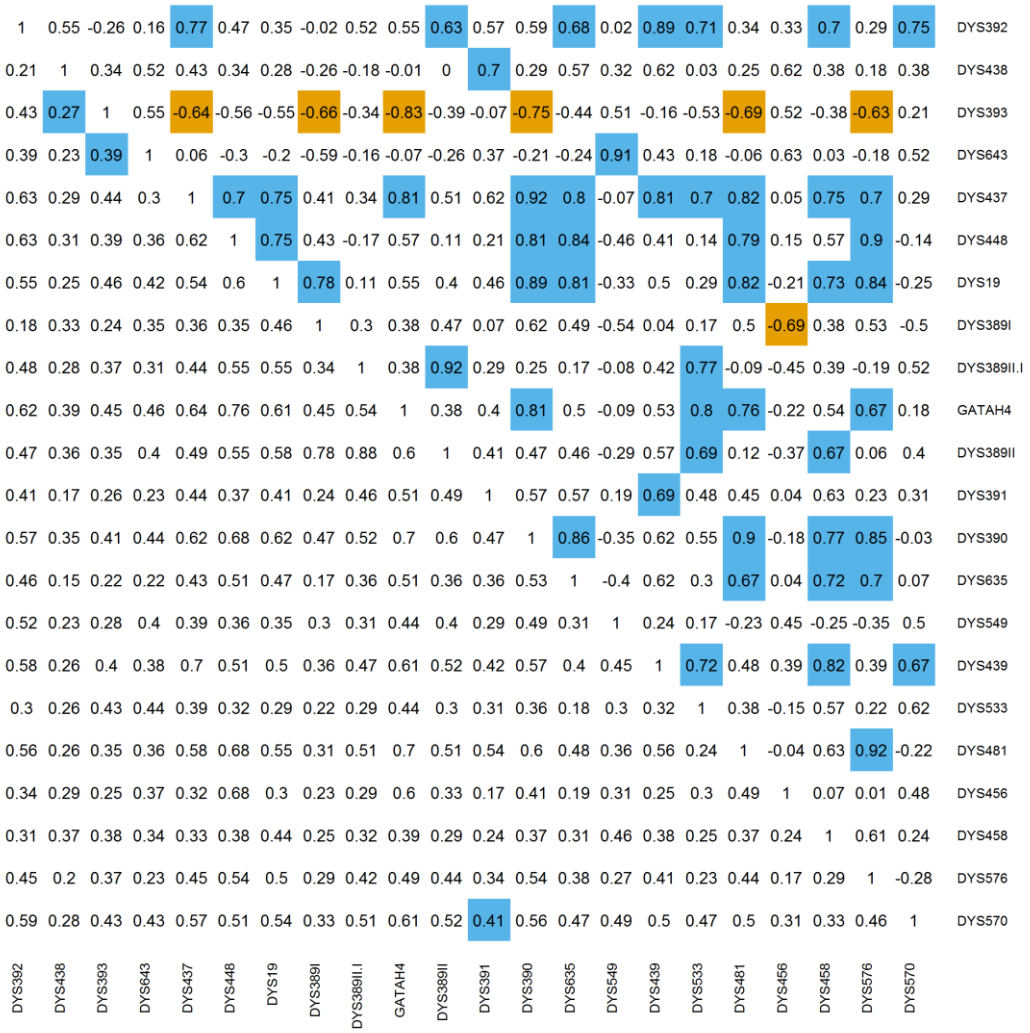

Figure 1.6. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. YHRD dataset, African populations.

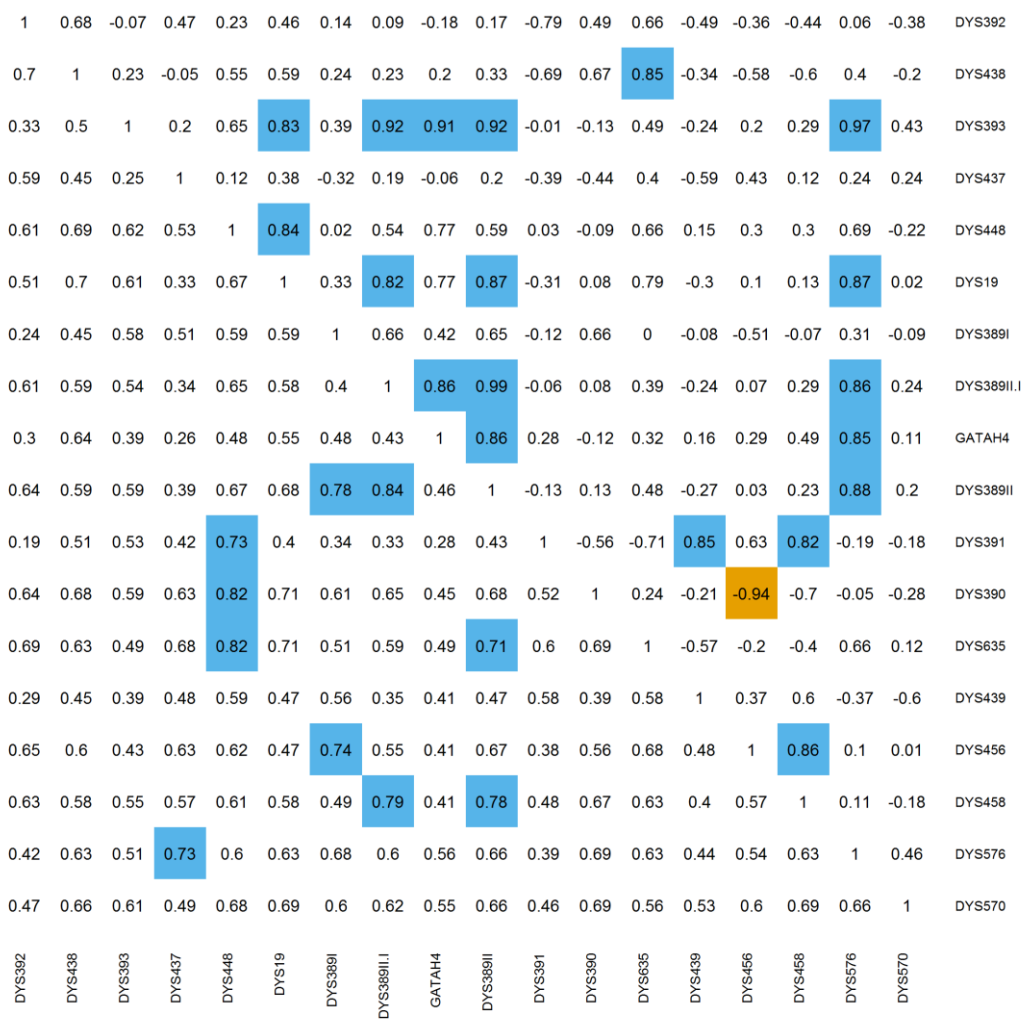

Figure 1.7. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. HGDP dataset, African populations.

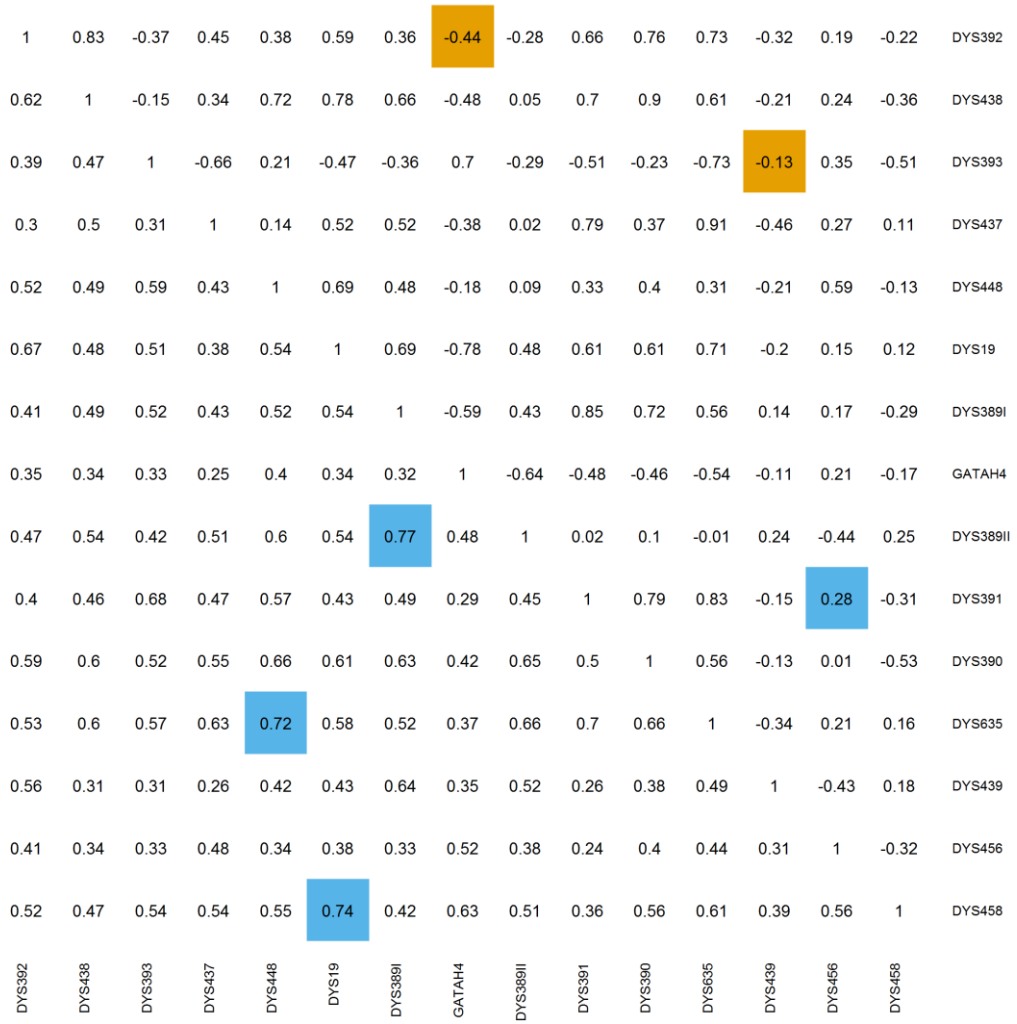

Figure 1.8. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. XU dataset, African populations.

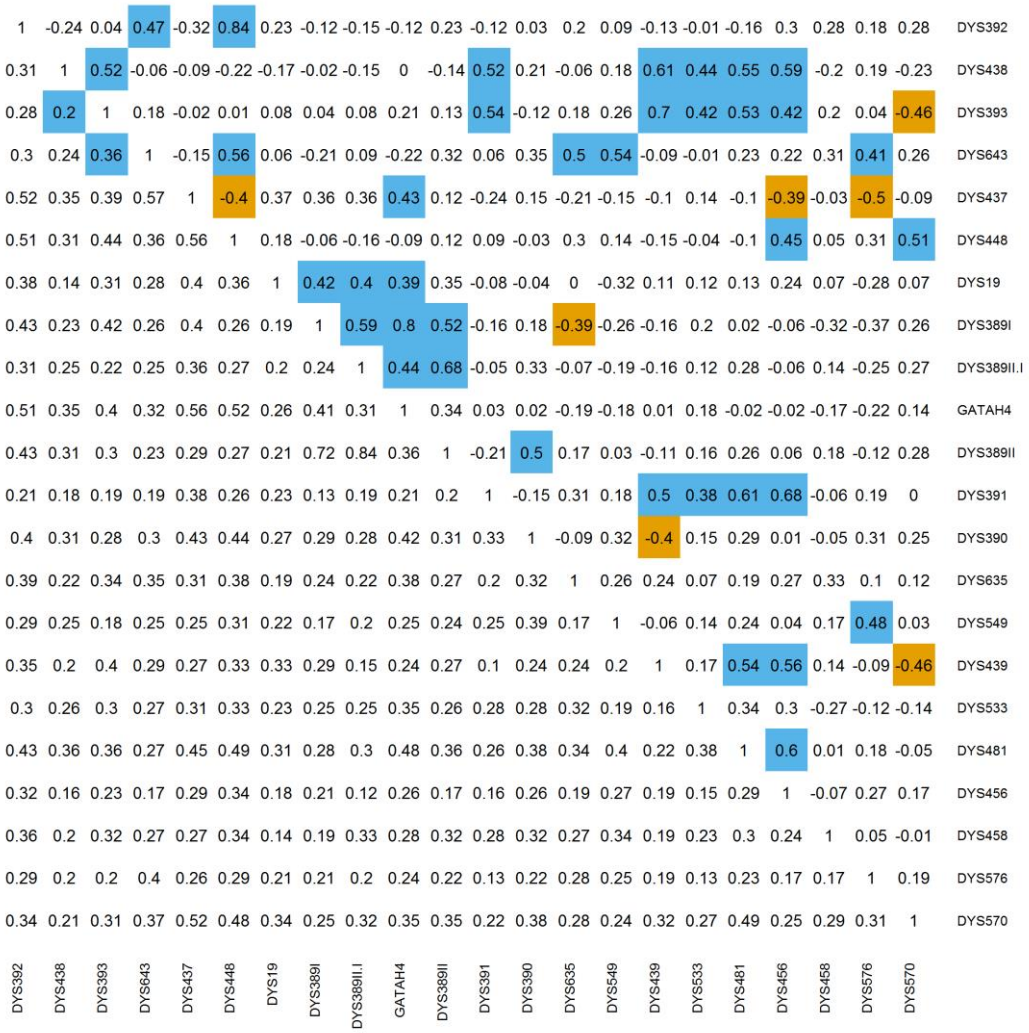

Figure 1.9. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. YHRD dataset, Asian populations.

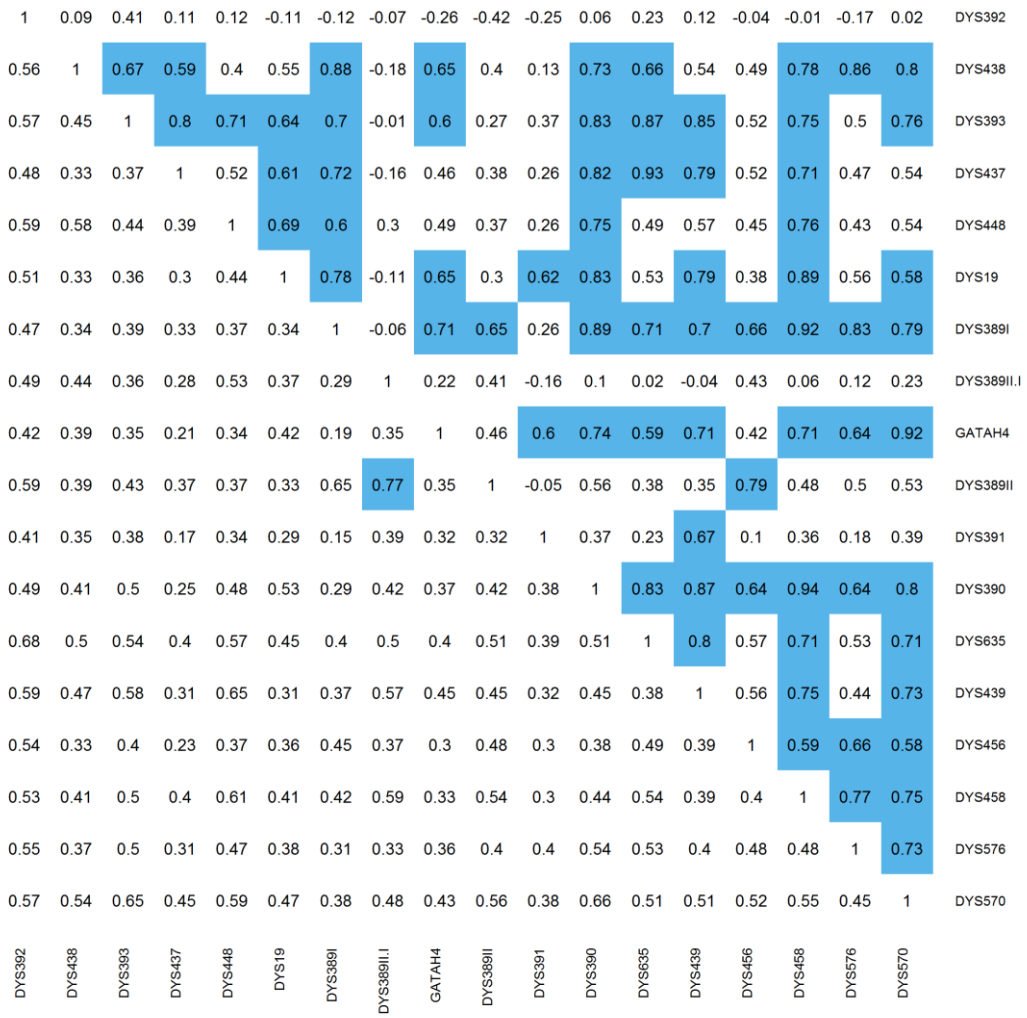

Figure 1.10. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. XU dataset, all populations.

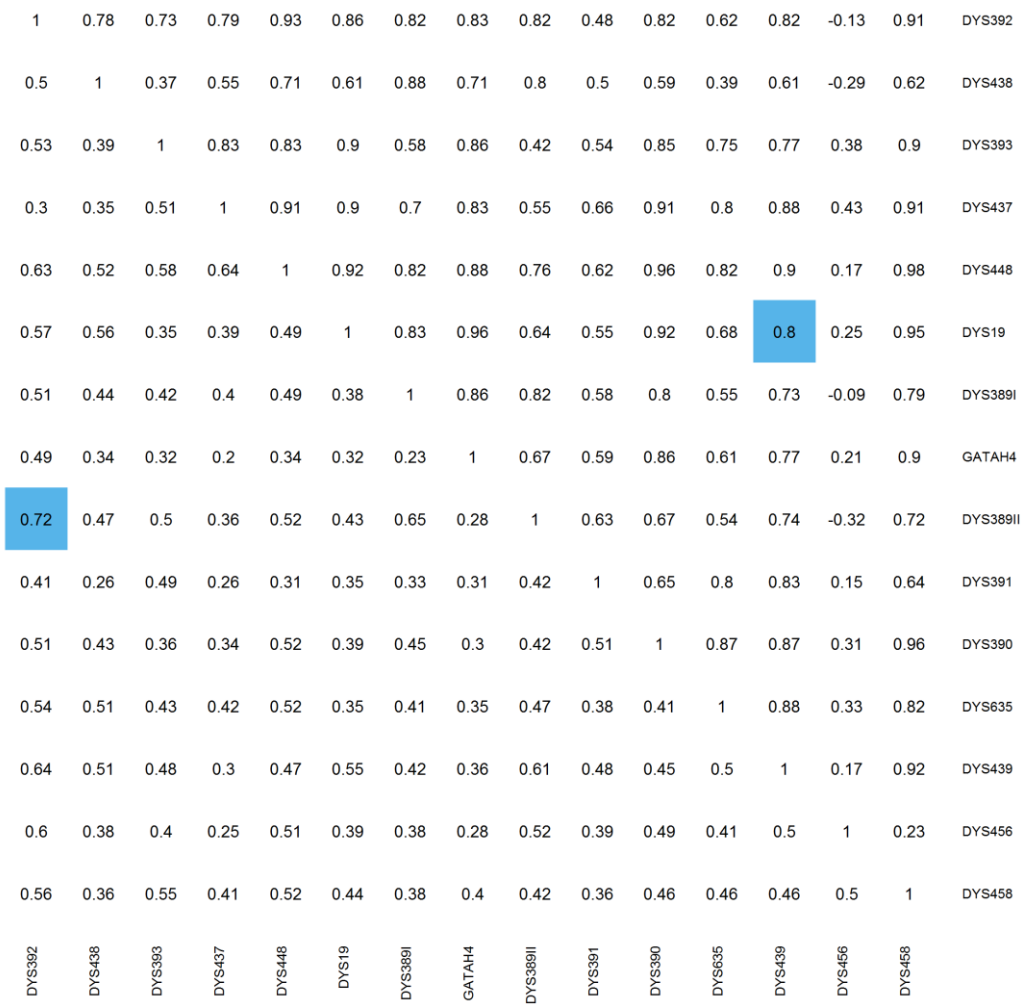

Figure 1.11. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. XU dataset, Asian populations.

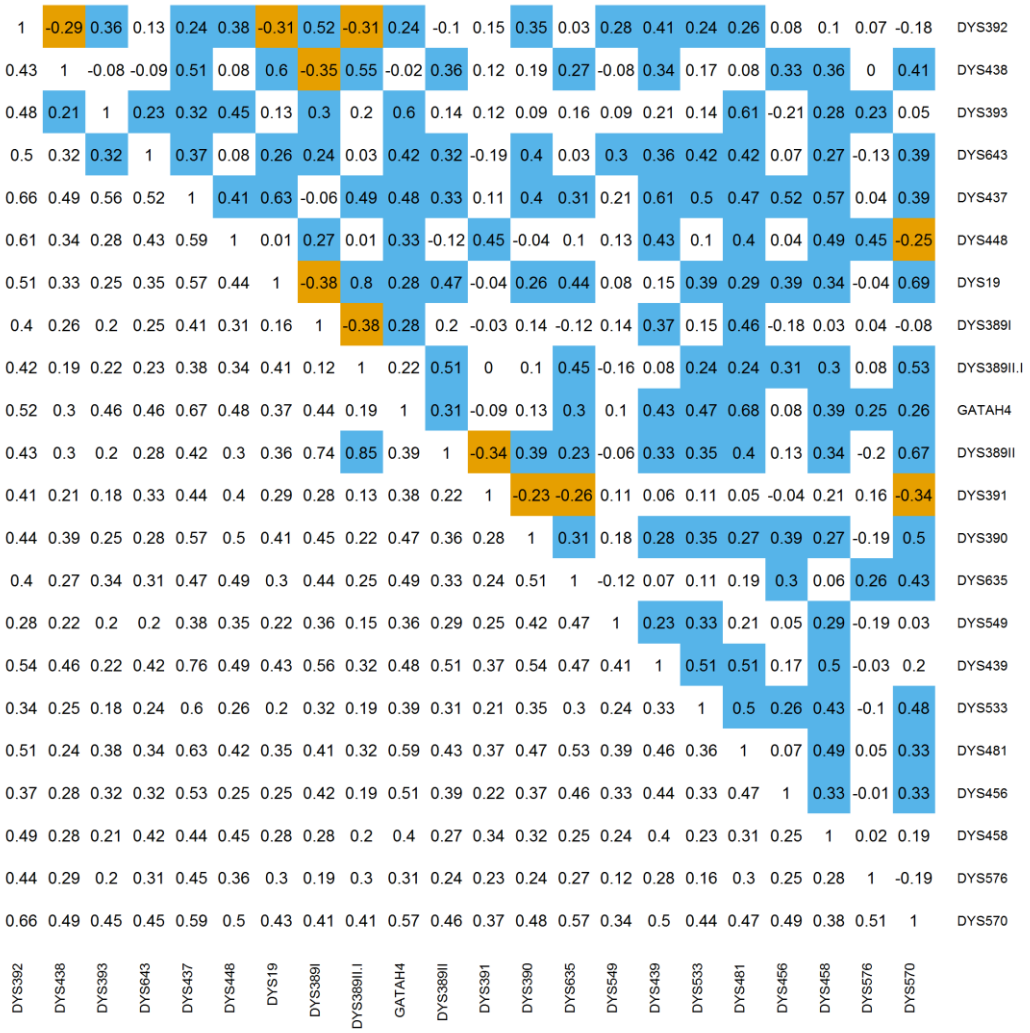

Figure 1.12. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. YHRD dataset, European populations.

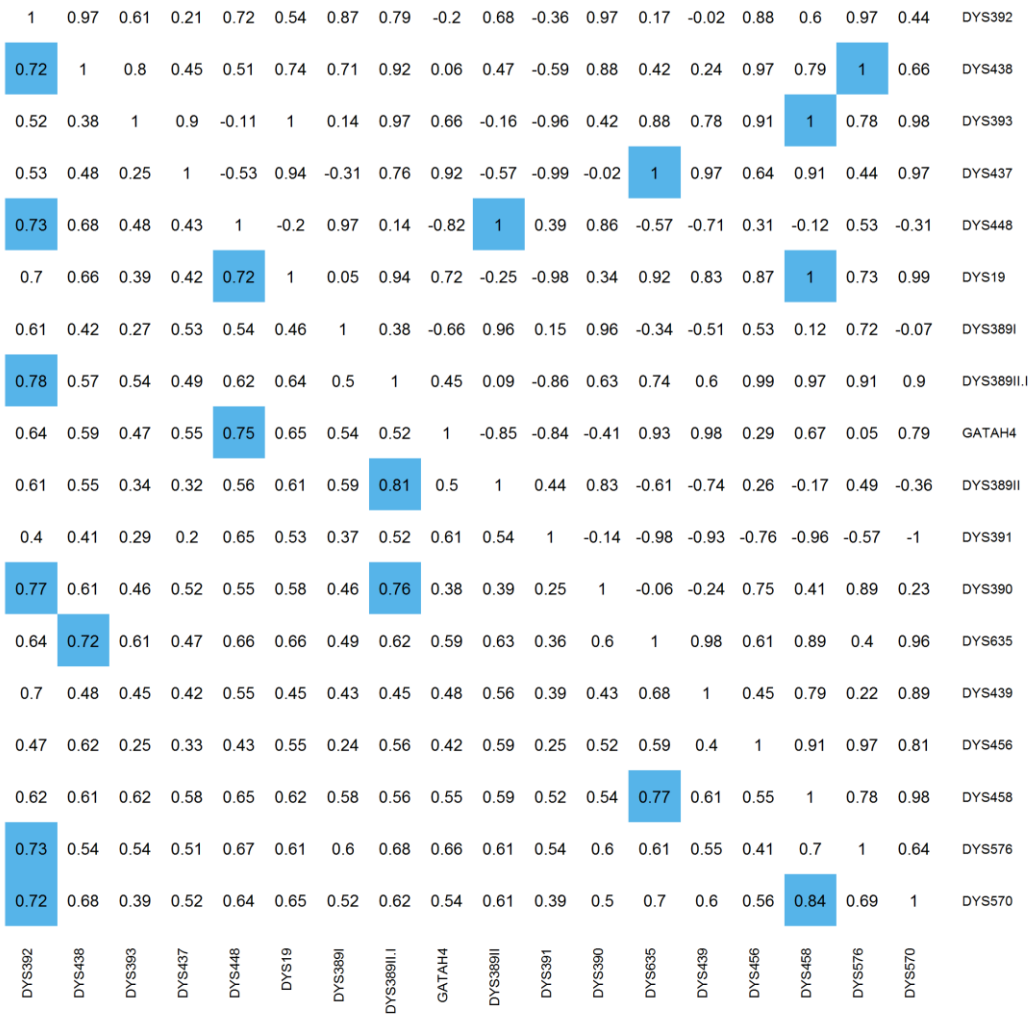

Figure 1.13. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. HGDP dataset, European populations.

1	0.78	0.73	0.79	0.93	0.86	0.82	0.83	0.82	0.48	0.82	0.62	0.82	-0.13	0.91	DYS392
0.68	1	0.37	0.55	0.71	0.61	0.88	0.71	0.8	0.5	0.59	0.39	0.61	-0.29	0.62	DYS438
0.51	0.58	1	0.83	0.83	0.9	0.58	0.86	0.42	0.54	0.85	0.75	0.77	0.38	0.9	DYS393
0.59	0.63	0.23	1	0.91	0.9	0.7	0.83	0.55	0.66	0.91	0.8	0.88	0.43	0.91	DYS437
0.64	0.53	0.31	0.39	1	0.92	0.82	0.88	0.76	0.62	0.96	0.82	0.9	0.17	0.98	DYS448
0.58	0.56	0.25	0.47	0.43	1	0.83	0.96	0.64	0.55	0.92	0.68	0.8	0.25	0.95	DYS19
0.5	0.44	0.31	0.57	0.31	0.35	1	0.86	0.82	0.58	0.8	0.55	0.73	-0.09	0.79	DYS389I
0.35	0.36	0.28	0.29	0.42	0.23	0.32	1	0.67	0.59	0.86	0.61	0.77	0.21	0.9	GATAH4
0.5	0.39	0.32	0.58	0.31	0.41	0.67	0.37	1	0.63	0.67	0.54	0.74	-0.32	0.72	DYS389II
0.44	0.5	0.41	0.37	0.56	0.27	0.2	0.29	0.24	1	0.65	0.8	0.83	0.15	0.64	DYS391
0.58	0.61	0.34	0.5	0.45	0.42	0.47	0.26	0.34	0.37	1	0.87	0.87	0.31	0.96	DYS390
0.56	0.66	0.55	0.48	0.43	0.4	0.49	0.35	0.47	0.39	0.54	1	0.88	0.33	0.82	DYS635
0.48	0.5	0.35	0.58	0.4	0.4	0.35	0.34	0.46	0.33	0.41	0.38	1	0.17	0.92	DYS439
0.48	0.59	0.42	0.36	0.4	0.41	0.46	0.35	0.43	0.24	0.52	0.61	0.43	1	0.23	DYS456
0.51	0.48	0.37	0.4	0.43	0.46	0.27	0.47	0.37	0.3	0.33	0.44	0.32	0.35	1	DYS458
DYS392	DYS438	DYS393	DYS437	DYS448	DYS19	DYS389I	GATAH4	DYS389II	DYS391	DYS390	DYS635	DYS439	DYS456	DYS458	

Figure 1.14. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. XU dataset, European populations.

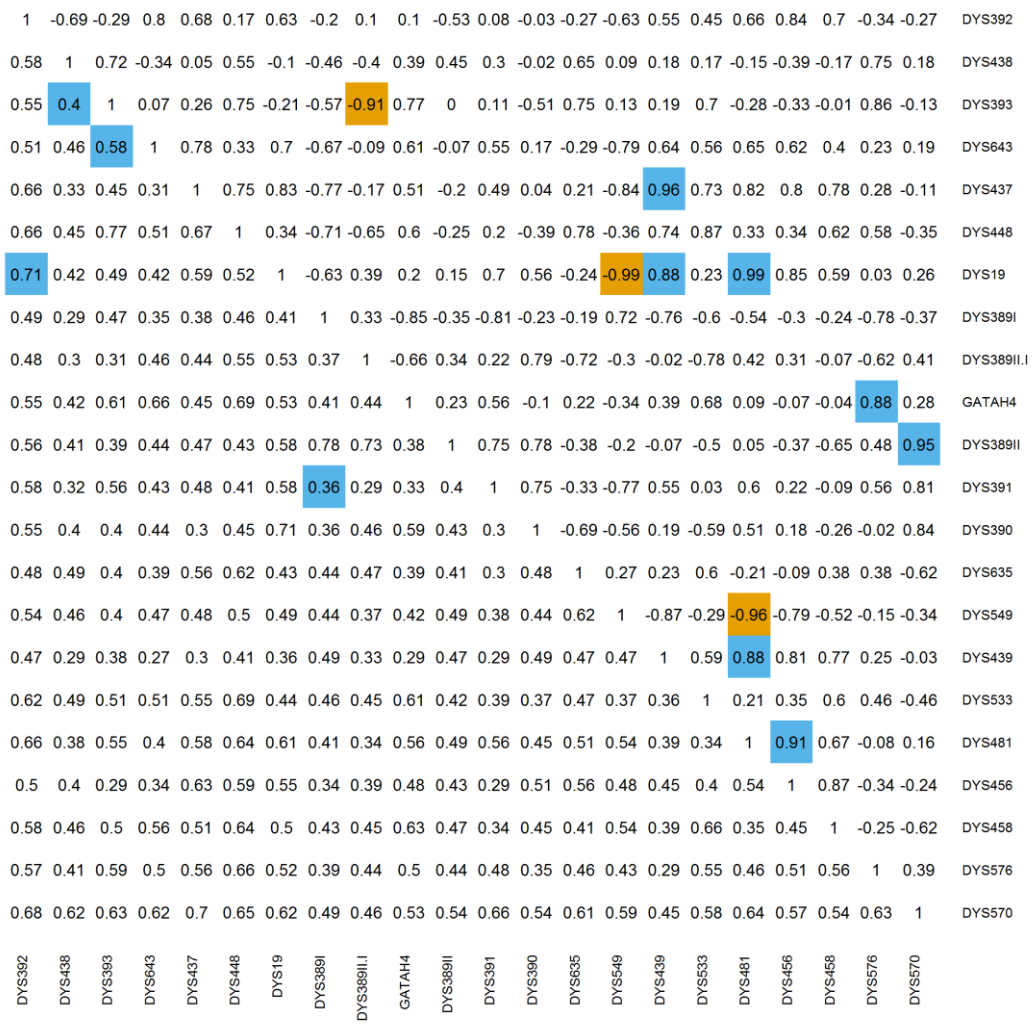

Figure 1.15. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. YHRD dataset, Native American populations.

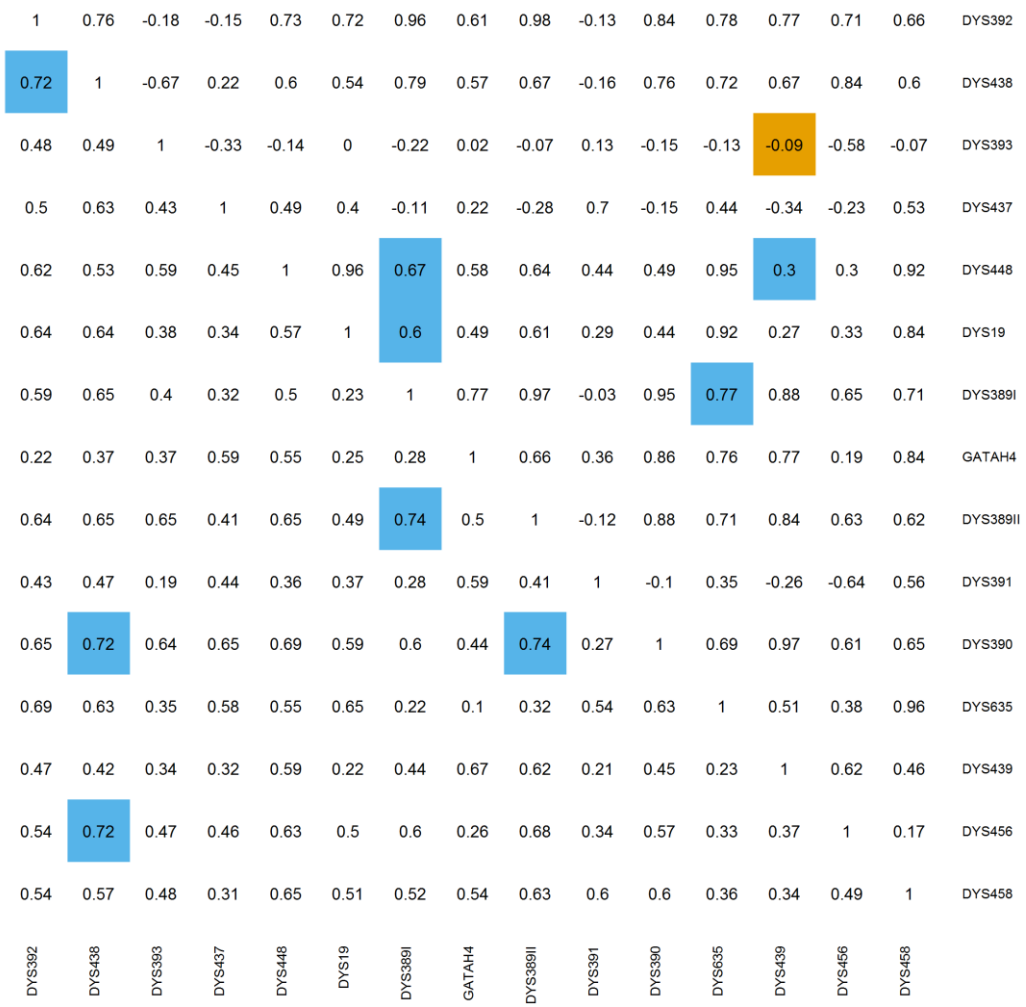

Figure 1.16. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. XU dataset, Native American populations.

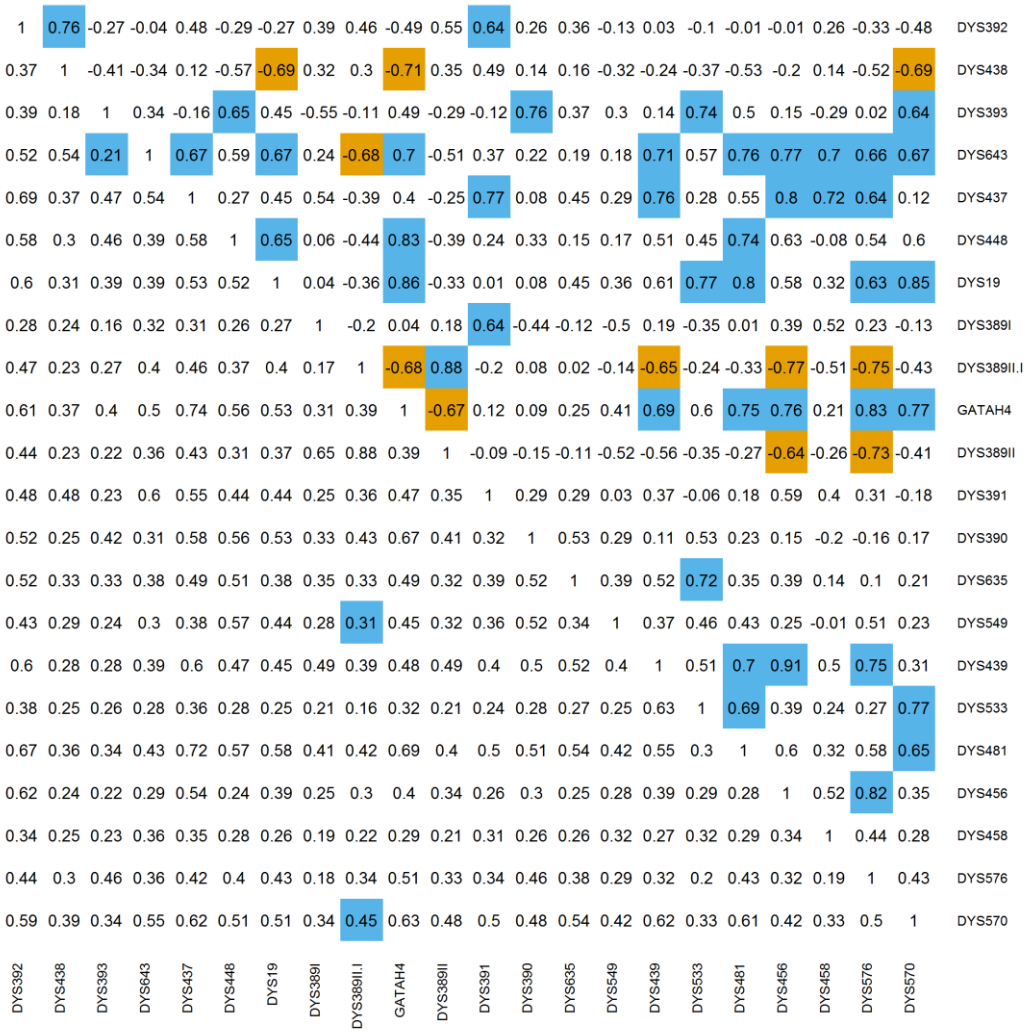

Figure 1.17. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. YHRD dataset, Admixed American populations.

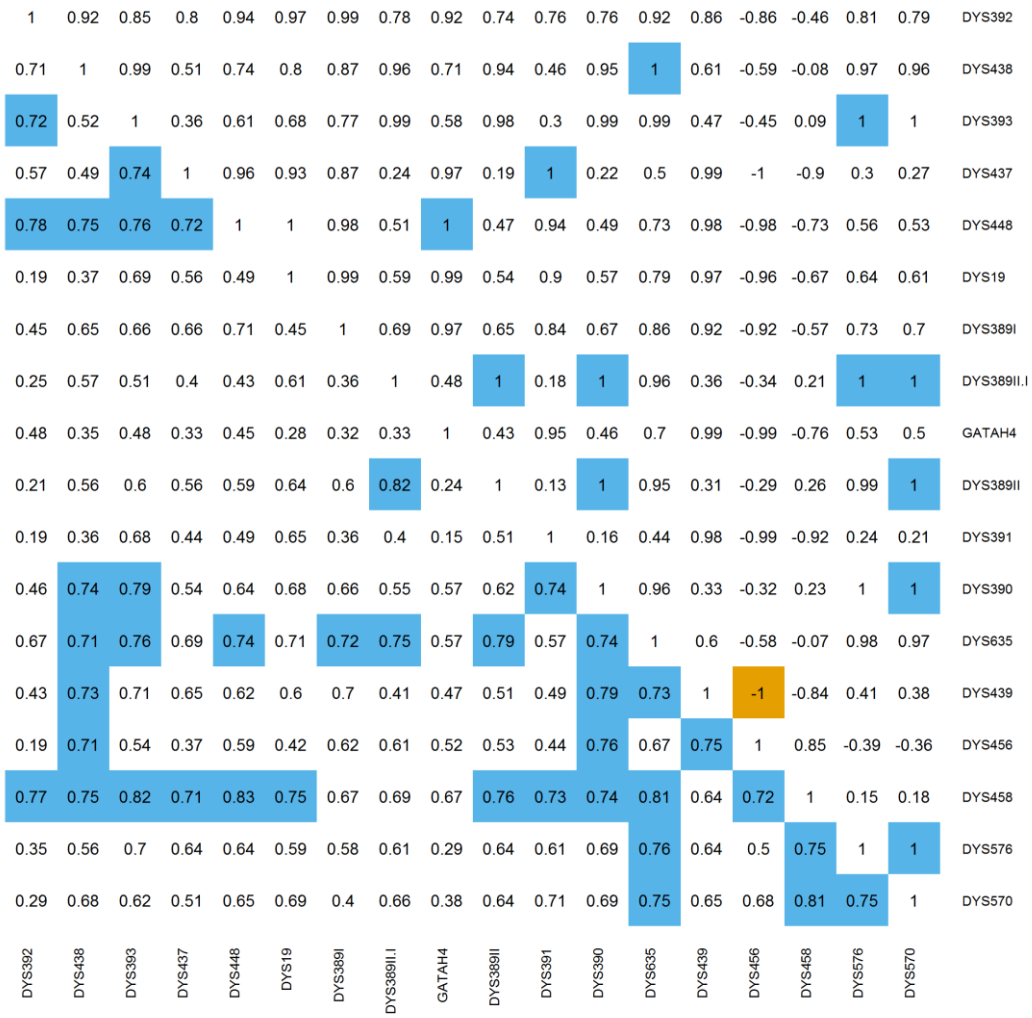

Figure 1.18. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. HGDP dataset, Middle Eastern populations.

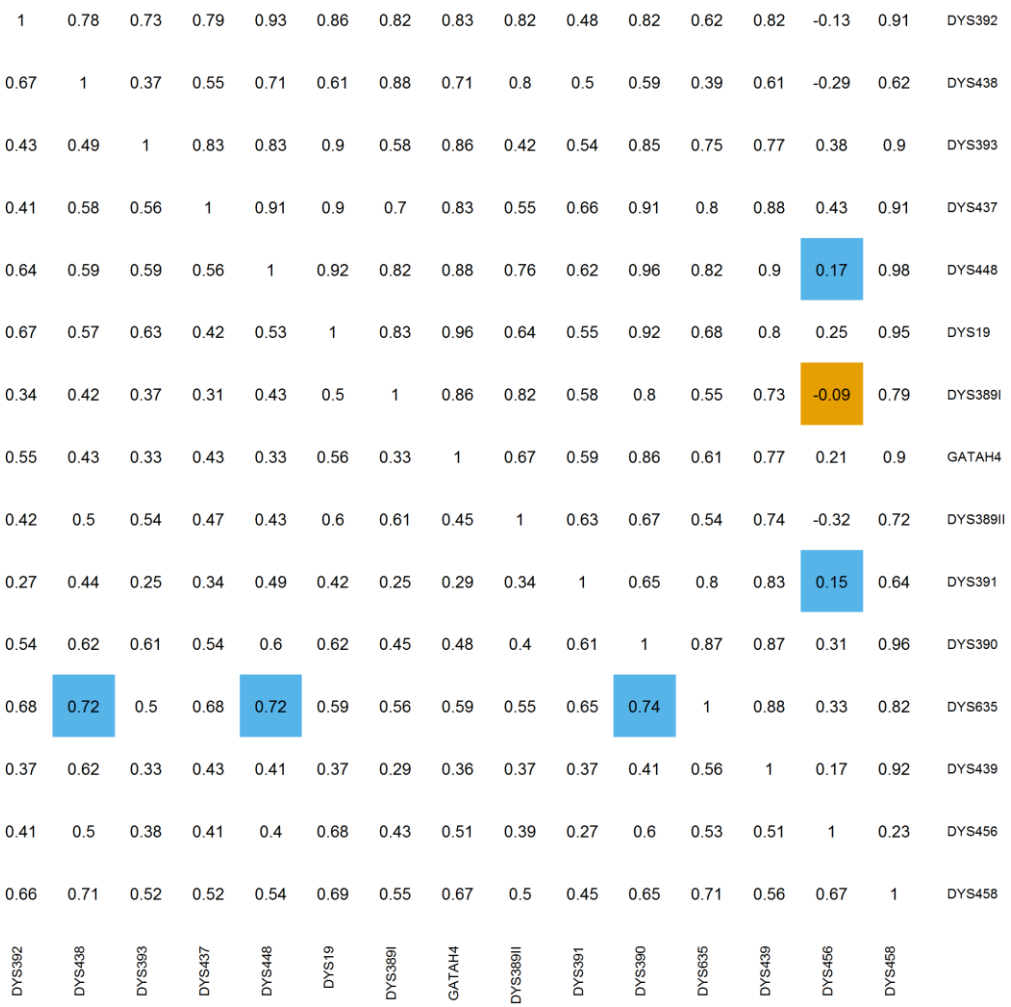

Figure 1.19. Comparison of LD (lower triangle) with single locus β_i correlations (upper triangle) by regional ethnicity and database. Statistically significant LD and positive correlations are shaded blue. Significant negative correlations are shaded orange. Y-STR loci are ordered by published mutation rate. XU dataset, Middle Eastern populations.

1.3.7 *Results: Multi-Locus β_i Estimates*

After examining the behavior of single-locus β_{il} estimates, we then calculated multi-locus β_{il} estimates considering all possible combinations of Y-haplotypes from combinations of 2 to 13 loci. Figure 1.20 shows the mean and range of the β_{Wl}'s calculated for each order of haplotype, relative to regional ancestry, for each dataset, plotted on a log scale. As the number of loci included in a haplotype increased, mean $\beta_{W,PR}$ decreased. The range of $\beta_{W,PR}$ values also became smaller. We did observe some negative $\beta_{W,PR}$ estimates. This occurred for populations in the smaller datasets when one or more loci were monomorphic for an allele. Particularly evident for the YHRD data, the decrease in mean $\beta_{W,PR}$ is not linear.

The rate of mean $\beta_{W,PR}$ decrease does seem to be related to the correlation between single-locus β_{il} estimates. Regional ancestry groups with more correlation between single-locus β_{il} estimates show smaller change in mean β_W estimates upon adding more loci to a haplotype. Across datasets, mean $\beta_{W,PR}$ estimates for Native American and Middle Eastern ancestry show little change moving from smaller to larger-order haplotypes. These populations also have more β_{il} pairs with strong, significant correlations, both positive and negative.

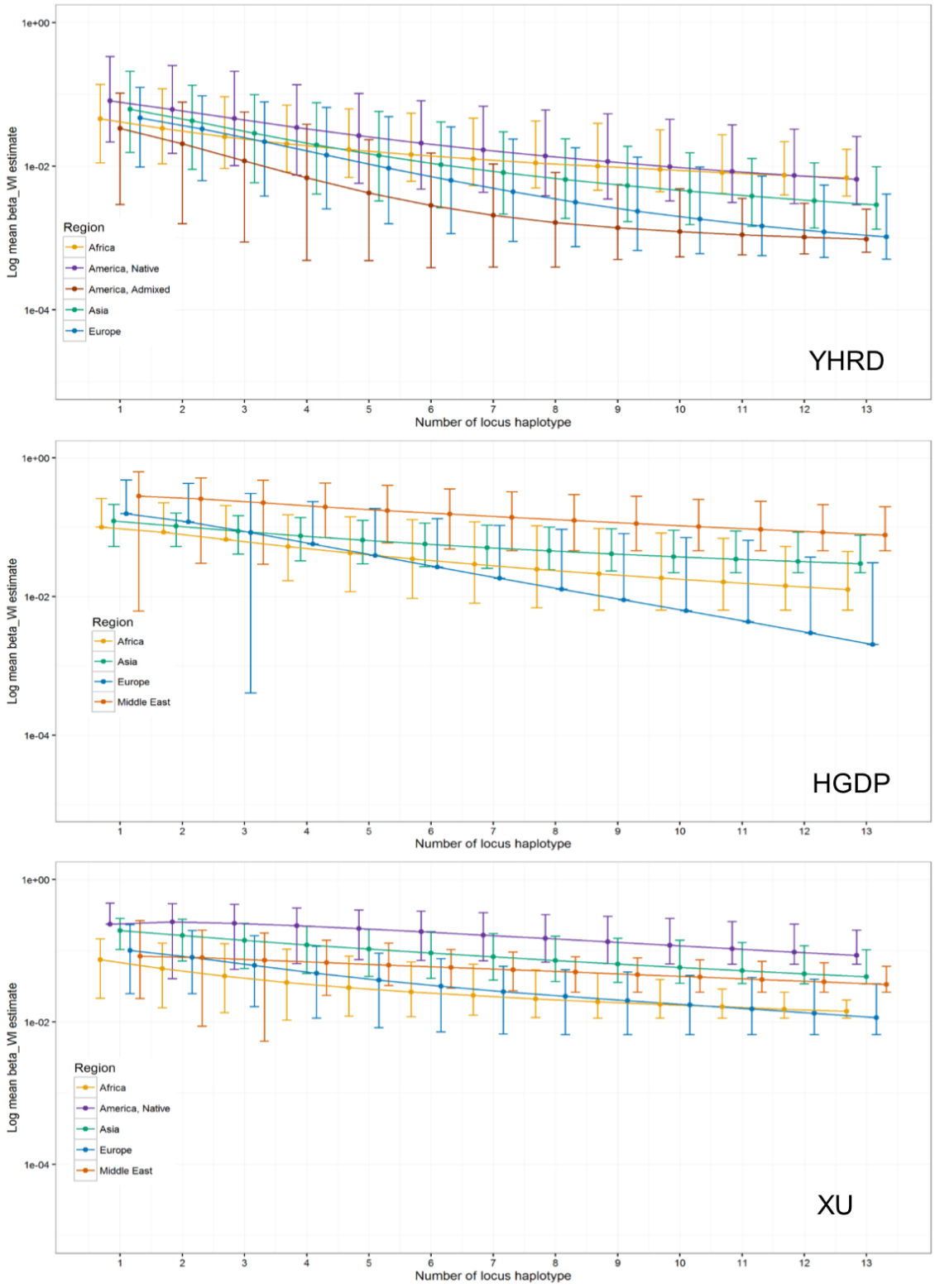

Figure 1.20. Multi-locus β_W's by regional ethnicity.

β_{il}'s can be calculated relative to the total worldwide sample of populations (β_{PT}) or relative to populations within the same regional ancestry(β_{PR}). Figure 1.21, Figure 1.22, and Figure 1.23 compare these two methods, summarized within a population, over loci (β_i, left column) or summarized within a locus, over populations (β_W, right column), for each dataset. Median $\beta_{i,PR}$'s are generally smaller than $\beta_{i,PT}$'s. As the number of loci in a haplotype increases, the median $\beta_{i,PR}$ and $\beta_{i,PT}$ estimates and the variability are nearly identical. The same patterns hold comparing the distribution of the $\beta_{W,PR}$'s and $\beta_{W,PT}$'s. In the YHRD dataset, which has the largest number of populations and individuals within populations, the median $\beta_{W,PR}$ estimates across regions are nearly identical even for single-locus estimates and are identical for five- and ten-locus estimates.

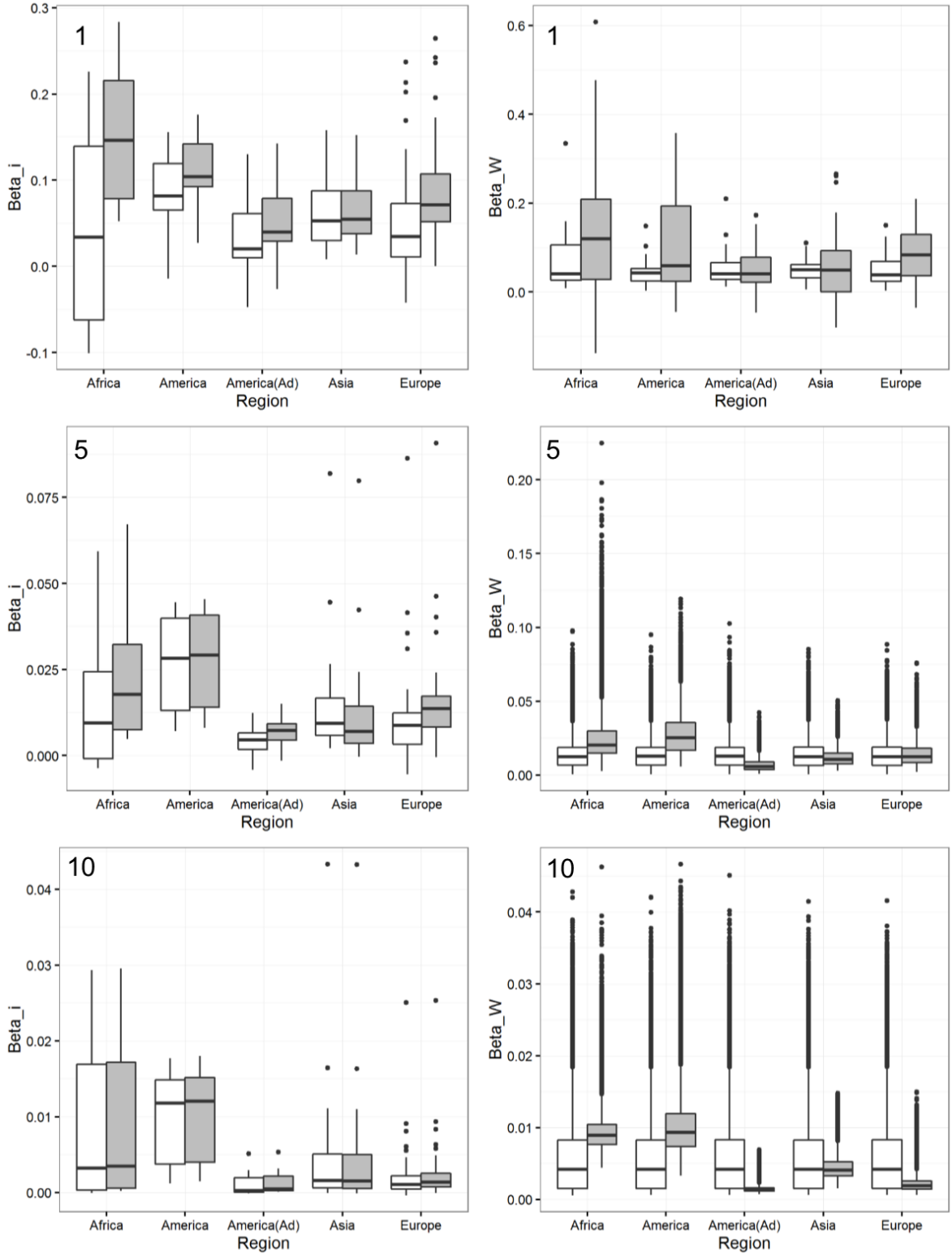

Figure 1.21 Distribution of β estimates for 1, 5, and 10 loci combination haplotypes, YHRD. Left column: β_i. Right column: β_W. White: β_{PR}. Grey: β_{PT}.

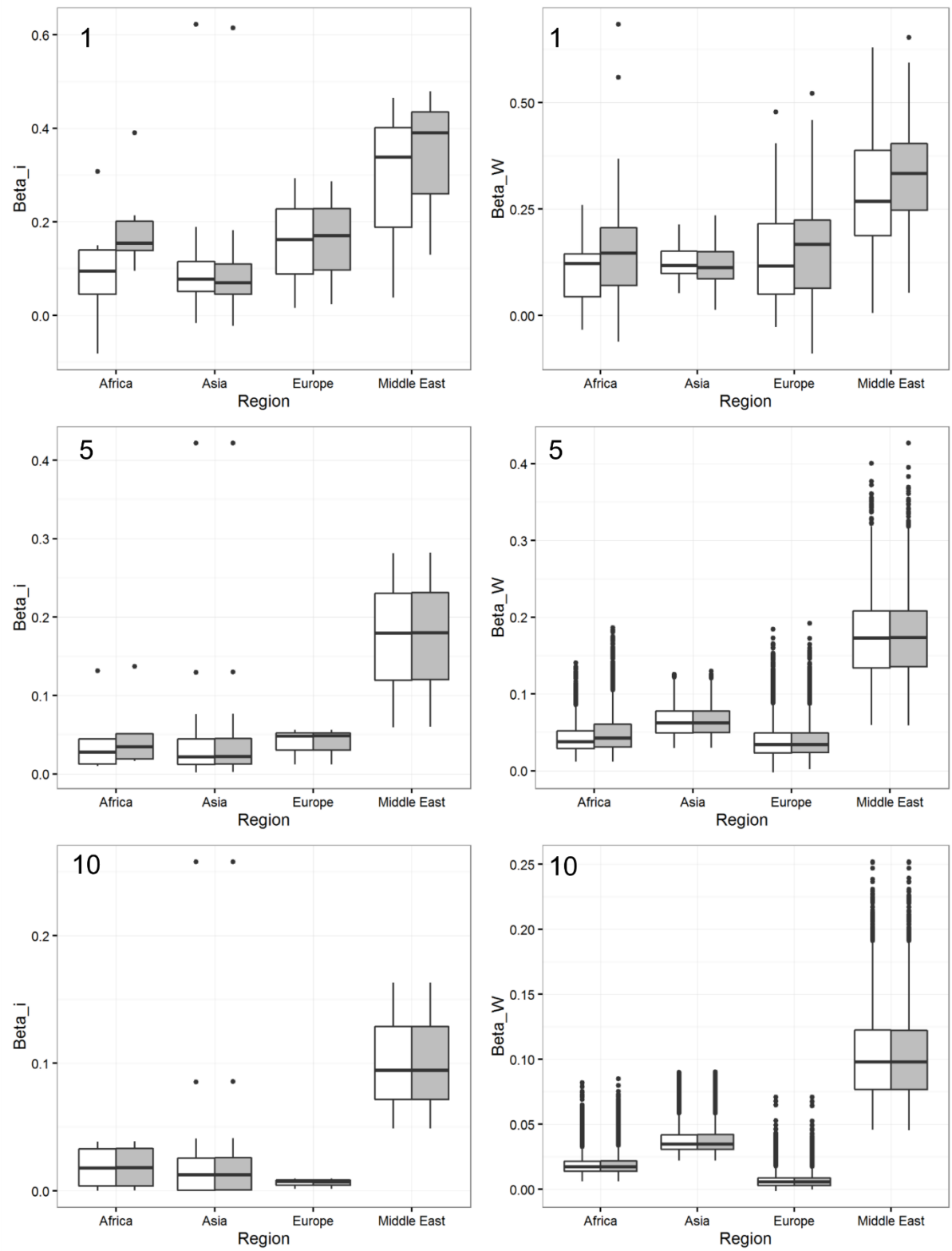

Figure 1.22 Distribution of β estimates for 1, 5, and 10 loci combination haplotypes, HGDP. Left column: β_i. Right column: β_W. White: β_{PR}. Grey: β_{PT}.

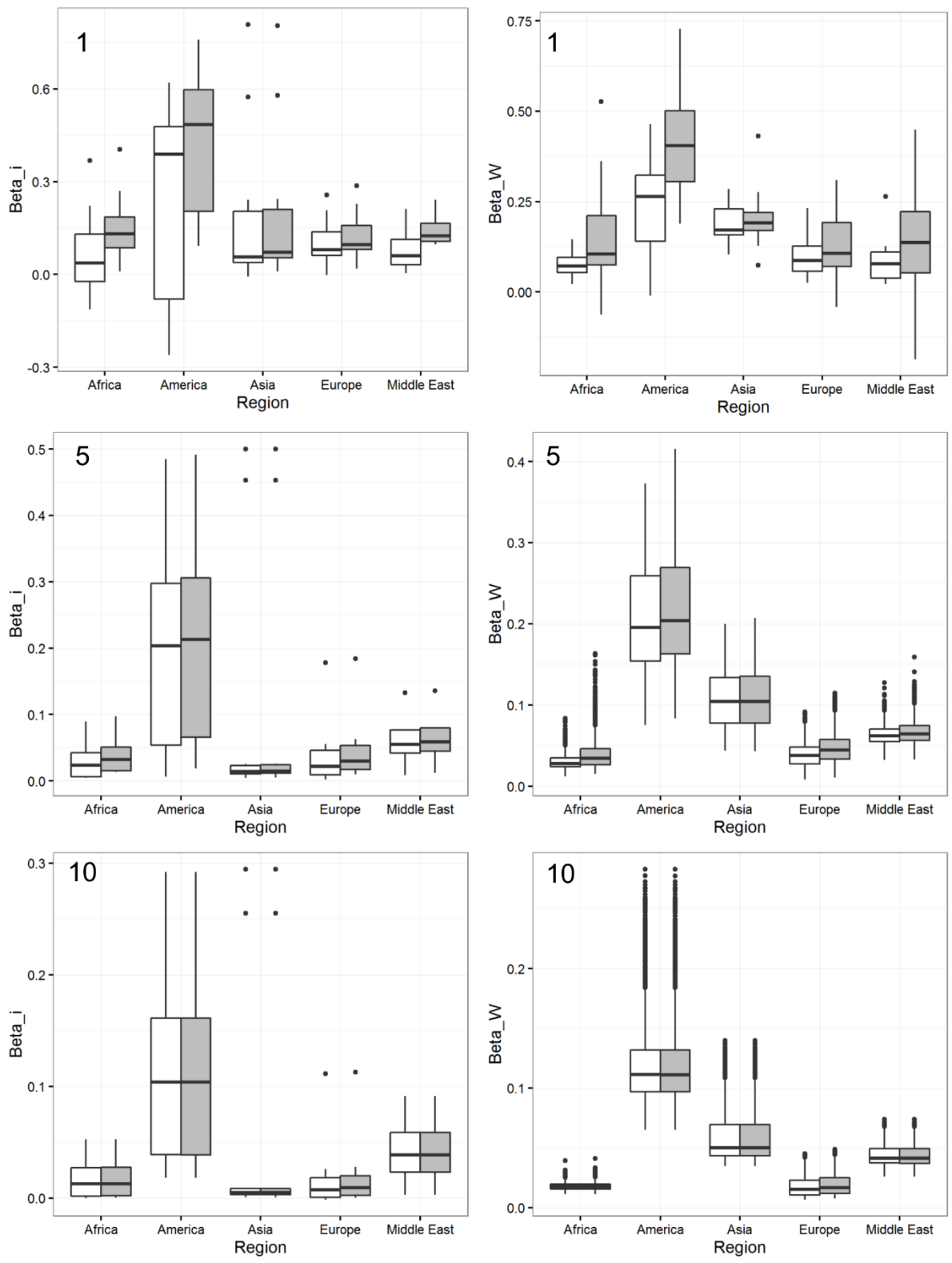

Figure 1.23 Distribution of β estimates for 1, 5, and 10 loci combination haplotypes, XU. Left column: β_i. Right column: β_W. White: β_{PR}. Grey: β_{PT}.

1.3.8 *Results: Shannon Entropy and Marker Selection*

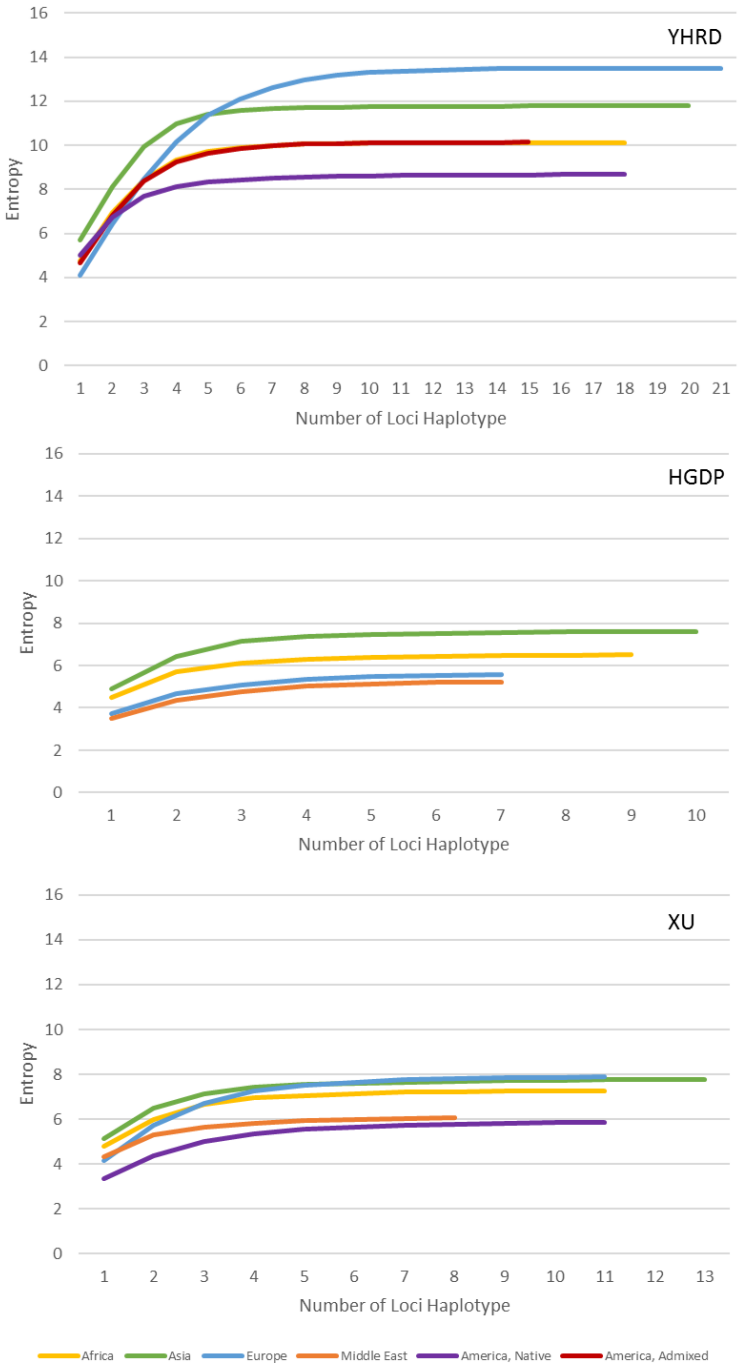

Figure 1.24 Combined entropy for each order of optimally selected haplotype, by region and dataset.

Figure 1.24 shows the combined entropy for each region within a dataset as markers were added to the haplotype chain. In each case, the information gained by successive marker addition decreased rapidly, generally after four to eight markers were added to the haplotype. Total length of the haplotype selected ranged from 7 to 21 loci.

The order that markers were added to the optimal haplotype by the selection algorithm and entropy values are shown in Table 1.6. The selected haplotypes explained 87-100% of the maximum entropy calculated with all possible markers. In all cases, marker DYS385ab was selected as the first marker in the haplotype. For marker panels that contained the high mutation rate loci, DYS570 and DYS576, one or both of these markers were added next, even though these markers had neither the next highest single locus entropy or number of alleles. The order that other makers were added varied considerably among datasets and regional ethnicity groups.

In a linear regression analysis, stratified by dataset and adjusted for region, ordered genetic diversity was independently associated with the order of marker addition in all three datasets (YHRD p=0.004, HGDP p=0.003, XU p=0.007). The association between ordered mutation rate and the order of marker rate addition was highly significant, but only in the YHRD dataset ($p=2.84 \times 10^{-11}$). Markers with larger mutation rate and genetic diversity were added earlier in the chain (Figure 1.25).

Figure 1.26 shows the relationship between matching within a population and its entropy for all haplotypes one to four loci in length, restricted to the YHRD dataset. We observed an inverse relationship between matching and entropy. When entropy was greater than 5.0, matching approached zero.

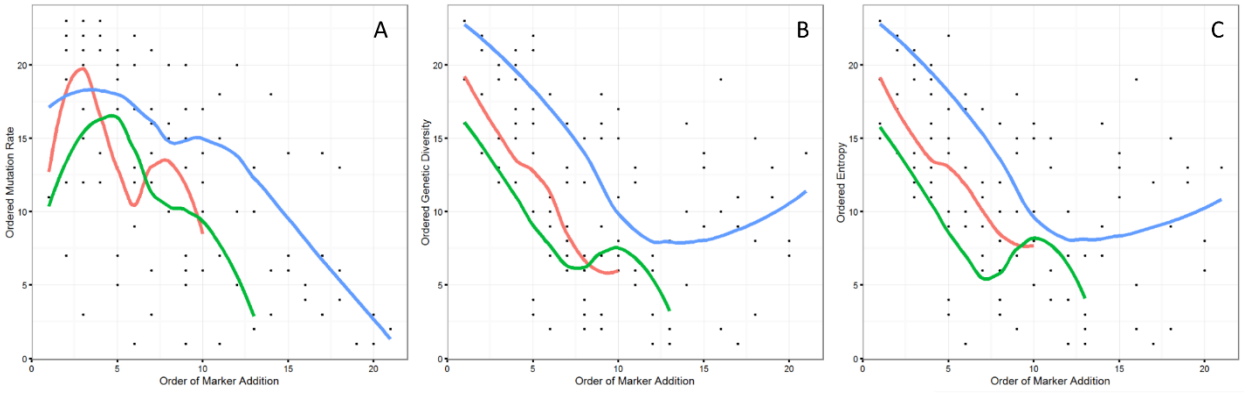

Figure 1.25. Relationship between order of marker addition to the haplotype and mutation rate (A), genetic diversity (B), and single locus entropy (C) by dataset (Blue = YHRD, Red = HGDP, Green = XU).

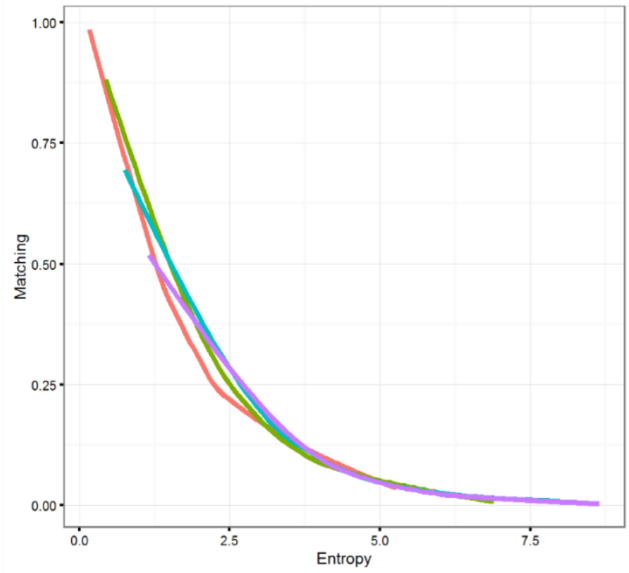

Figure 1.26. Relationship matching and entropy in the YHRD dataset by haplotype length. Red = 1 locus, Green = 2 loci, Blue = 3 loci, Purple = 4 loci.

Table 1.6. Resulting marker order from choosing optimally informative set of Y-STR markers and single locus, combined, and conditional entropy, by region and dataset.

YHRD																
Africa				Asia				Europe				America, Admixed				America, Native			
Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond
DYS385ab	4.750	4.750	4.750	DYS385ab	5.716	5.716	5.716	DYS385ab	4.100	4.100	4.100	DYS385ab	4.673	4.673	4.673	DYS385ab	4.995	4.995	4.995
DYS481	2.962	6.972	2.222	DYS570	2.769	8.115	2.399	DYS570	2.563	6.435	2.336	DYS576	2.538	6.839	2.166	DYS576	2.552	6.756	1.762
DYS570	2.554	8.447	1.474	DYS576	2.562	9.944	1.828	DYS576	2.381	8.475	2.040	DYS570	2.593	8.366	1.526	DYS570	2.508	7.691	0.934
DYS576	2.493	9.318	0.871	DYS458	2.598	10.998	1.055	DYS458	2.362	10.170	1.695	DYS458	2.321	9.239	0.873	DYS389II	1.942	8.115	0.424
DYS458	2.220	9.741	0.423	DYS481	2.860	11.406	0.408	DYS481	2.842	11.360	1.190	DYS389II	2.213	9.649	0.410	DYS439	2.059	8.330	0.215
DYS389II	2.329	9.906	0.165	DYS389II	2.319	11.582	0.176	DYS456	2.163	12.099	0.739	DYS549	1.913	9.870	0.221	DYS456	1.816	8.451	0.121
DYS549	1.719	9.999	0.093	DYS439	1.923	11.664	0.082	DYS389II	2.095	12.627	0.528	DYS456	1.973	9.995	0.125	DYS549	1.990	8.529	0.079
DYS635	2.136	10.052	0.053	DYS549	1.773	11.703	0.039	DYS549	1.792	12.964	0.337	DYS439	1.851	10.064	0.070	DYS481	2.628	8.575	0.046
DYS19	2.112	10.080	0.028	DYS635	2.465	11.728	0.024	DYS439	1.920	13.182	0.218	DYS391	1.425	10.093	0.028	DYS458	2.419	8.602	0.027
DYS439	1.637	10.104	0.024	GATAH4	1.727	11.744	0.016	DYS390	2.046	13.304	0.122	DYS19	2.011	10.109	0.016	DYS391	1.103	8.616	0.015
DYS533	1.433	10.114	0.010	DYS533	1.708	11.756	0.012	DYS635	2.001	13.372	0.068	DYS390	2.055	10.121	0.013	DYS643	1.675	8.629	0.013
DYS456	1.691	10.120	0.006	DYS456	1.775	11.765	0.009	GATAH4	1.569	13.420	0.049	DYS533	1.756	10.129	0.007	DYS393	1.603	8.642	0.013
GATAH4	1.512	10.124	0.005	DYS391	1.097	11.774	0.009	DYS391	1.279	13.454	0.033	GATAH4	1.497	10.135	0.007	GATAH4	1.730	8.651	0.009
DYS393	1.654	10.128	0.003	DYS448	2.299	11.778	0.005	DYS533	1.668	13.471	0.018	DYS448	1.978	10.137	0.002	DYS392	2.158	8.658	0.008
DYS448	1.858	10.130	0.002	DYS390	2.187	11.782	0.004	DYS19	1.837	13.484	0.013	DYS635	2.094	10.139	0.002	DYS635	1.748	8.666	0.007
DYS643	2.456	10.132	0.002	DYS437	1.212	11.786	0.003	DYS437	1.579	13.491	0.007					DYS390	1.826	8.669	0.004
DYS390	1.844	10.134	0.002	DYS19	1.974	11.788	0.002	DYS393	1.218	13.497	0.006					DYS437	1.264	8.673	0.004
DYS391	1.058	10.135	0.002	DYS643	2.267	11.790	0.002	DYS448	1.709	13.501	0.004					DYS448	2.055	8.677	0.004
				DYS392	2.124	11.791	0.001	DYS643	1.885	13.504	0.003								
				DYS393	1.754	11.791	0.001	DYS392	1.674	13.506	0.002								
								DYS438	1.908	13.508	0.002								
Max entropy		10.284				11.859				13.581				10.195				9.071	
Selected set percent of max		0.986				0.994				0.995				0.995				0.957	

HGDP															
Africa				Asia				Europe				Middle East			
Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond
DYS385ab	4.483	4.483	4.483	DYS385ab	4.919	4.919	4.919	DYS385ab	3.714	3.714	3.714	DYS385ab	3.507	3.507	3.507
DYS570	2.777	5.707	1.224	DYS576	2.532	6.423	1.504	DYS570	2.406	4.695	0.981	DYS576	2.121	4.355	0.848
DYS458	2.310	6.122	0.414	DYS389II	2.322	7.142	0.719	DYS439	1.902	5.097	0.402	DYS448	2.016	4.786	0.431
DYS390	2.064	6.291	0.169	DYS570	2.572	7.386	0.244	DYS576	2.224	5.368	0.271	DYS389II	1.751	5.051	0.265
DYS19	2.215	6.384	0.093	DYS458	2.364	7.467	0.081	DYS437	1.326	5.469	0.101	DYS19	1.372	5.125	0.074
DYS576	2.518	6.432	0.048	DYS392	2.298	7.524	0.057	DYS389II.I	1.803	5.512	0.043	DYS439	1.477	5.193	0.068
DYS393	1.691	6.460	0.028	DYS439	2.025	7.556	0.032	DYS458	1.940	5.555	0.043	DYS458	2.581	5.230	0.038
DYS635	1.973	6.488	0.028	DYS456	1.522	7.579	0.023								
DYS456	1.820	6.508	0.020	DYS437	1.332	7.601	0.022								
				DYS448	1.983	7.615	0.014								
Max entropy		6.629				7.895				5.555				5.728	
Selected set percent of max		0.982				0.965				1.000				0.913	

Table 1.6, continued. Resulting marker order from choosing optimally informative set of Y-STR markers and single locus, combined, and conditional entropy, by region and dataset

XU																
Africa				Asia				Europe				Middle East				America, Native			
Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond
DYS385ab	4.799	4.799	4.799	DYS385ab	5.120	5.120	5.120	DYS385ab	4.156	4.156	4.156	DYS385ab	4.317	4.317	4.317	DYS385ab	3.356	3.356	3.356
DYS19	2.256	5.990	1.191	DYS389II	2.380	6.473	1.353	DYS458	2.233	5.719	1.562	DYS458	2.473	5.304	0.987	DYS458	1.885	4.366	1.010
DYS393	2.012	6.661	0.671	DYS635	2.158	7.123	0.650	DYS456	2.209	6.687	0.968	DYS389II	1.709	5.637	0.332	DYS439	1.486	5.008	0.642
DYS458	2.010	6.942	0.281	DYS458	2.495	7.418	0.295	DYS389II	2.070	7.244	0.557	DYS456	1.880	5.817	0.180	DYS393	1.578	5.328	0.321
DYS391	1.246	7.062	0.120	DYS456	1.964	7.545	0.127	DYS439	1.906	7.499	0.255	DYS19	1.535	5.924	0.107	DYS391	0.789	5.541	0.213
DYS439	1.732	7.133	0.071	DYS439	1.922	7.615	0.070	DYS391	1.208	7.653	0.155	DYS391	1.079	5.983	0.059	DYS389I	1.451	5.643	0.102
DYS448	1.819	7.196	0.063	DYS393	1.832	7.657	0.042	DYS390	2.001	7.763	0.110	GATAH4	1.698	6.029	0.046	DYS437	0.433	5.717	0.074
DYS456	1.585	7.228	0.032	GATAH4	1.349	7.691	0.034	GATAH4	1.353	7.822	0.058	DYS439	1.741	6.063	0.035	DYS389II	2.269	5.775	0.058
DYS392	1.087	7.248	0.019	DYS391	1.296	7.720	0.029	DYS448	1.793	7.853	0.031					DYS438	1.091	5.830	0.055
DYS389II	2.432	7.259	0.012	DYS390	1.875	7.744	0.024	DYS635	2.127	7.871	0.018					DYS456	1.288	5.855	0.025
DYS635	2.215	7.271	0.012	DYS392	2.188	7.760	0.016	DYS19	1.752	7.881	0.011					GATAH4	1.165	5.874	0.018
				DYS437	1.078	7.769	0.009												
				DYS438	1.338	7.776	0.007												
Max entropy		7.409				8.229				8.028				6.700				6.768	
Selected set percent of max		0.981				0.945				0.982				0.905				0.868	

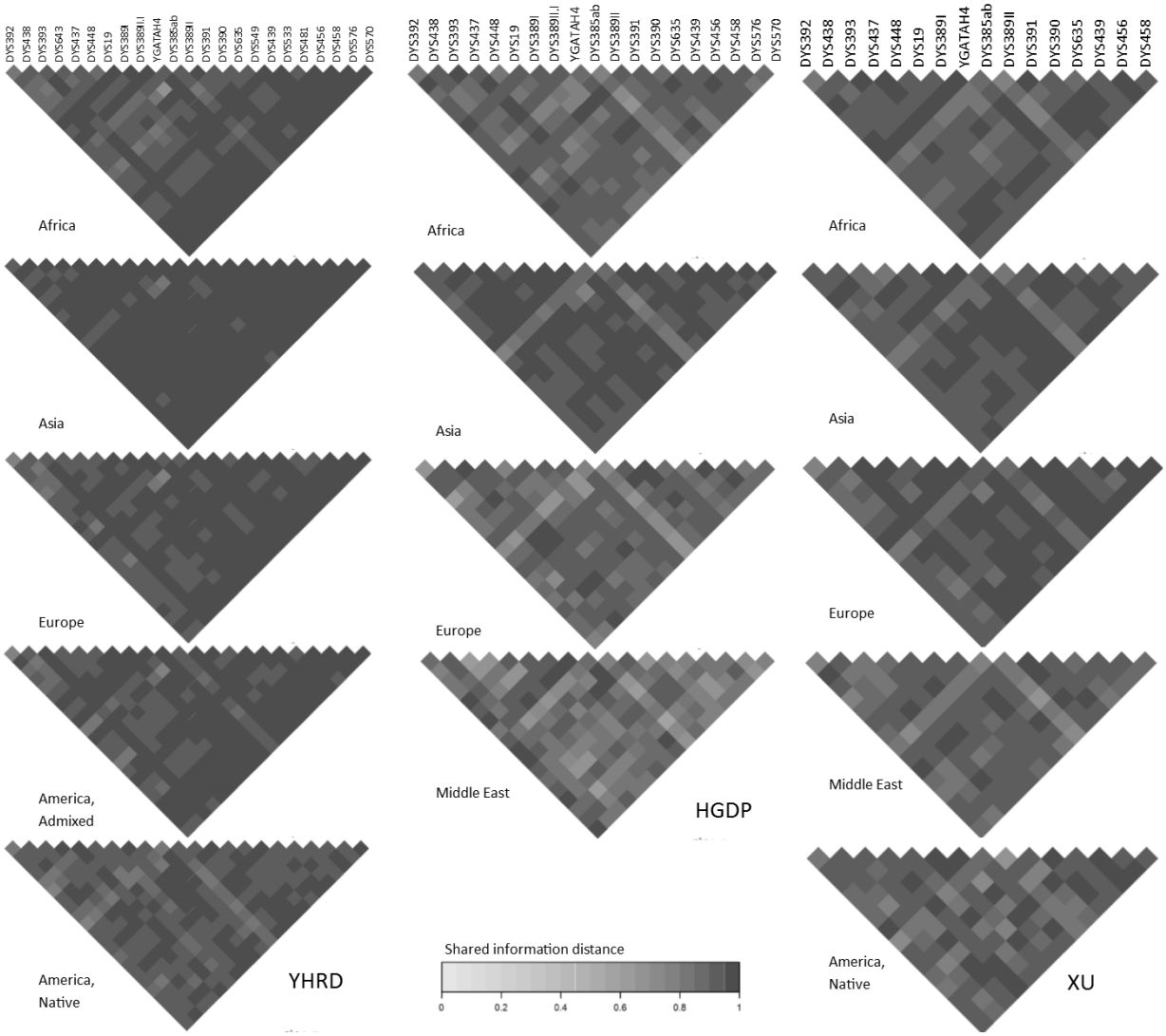

Figure 1.27 Shared information distance for two-locus haplotypes by regional ethnicity. Loci are ordered by mutation rate (left to right = low to high).

Shared information distance can be used to assess the unique information contributed by each locus in our marker panels. Figure 1.27 shows shared information distance heat maps for each region in each dataset. Darker color corresponds to greater information independence

between a pair of markers. In all groups, DYS385ab shares the most information with all other markers, as evidenced by the light gray bands running diagonally left and right from the center of each plot. Within the YHRD dataset, the markers with the highest mutation rates (DYS576, DYS570, and DYS458) also contribute the most independent information, however, this trend is not seen in the other datasets.

1.3.9 *Results: Comparison of Kappa and Evolutionary Matching*

We constructed 12-locus haplotypes, analogous to the PowerPlexY, from the markers available in each dataset to facilitate the comparison of the kappa method for Y-STR match calculation versus the evolutionary model advocated by SWGDAM. Figure 1.28 shows the frequency of haplotypes by popularity for the three datasets. Across datasets, approximately 80% of these 12-locus haplotypes were singletons, ranging from 75-93%, stratified by region. While most haplotypes were rare, some were quite common, occurring up to 50 times.

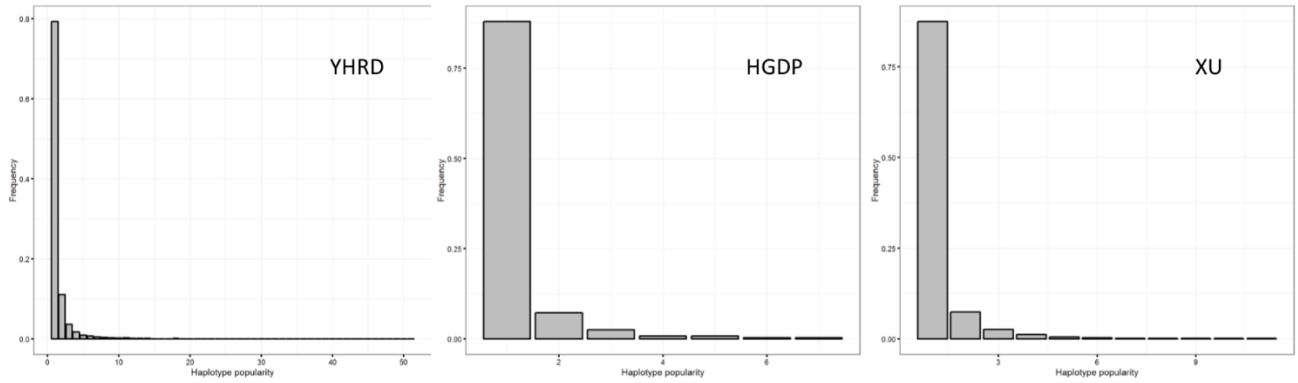

Figure 1.28 Popularity of a 12-locus haplotype based on the PowerPlexY marker system.

The likelihood ratios (LRs) of a match calculated via the kappa method were much larger – about 20 to 200 times larger -- than the match probabilities calculated by the

evolutionary model using the regional βw's calculated for this 12-locus haplotype (Table 1.7). The ranking order of the LR calculated for each region was also not the same between methods. Comparing the LRs calculated for each region in the HGDP and XU datasets (which have more similar samples sizes) under each method, regional LRs calculated using the kappa method were more disparate in magnitude, reflecting the lack of a genetic basis for this model, with the difference in LR calculated within a region varying hundreds to thousands. The LRs calculated using the evolutionary method were quite close, differing by just 4 to 15 within a region.

Table 1.7. Comparison of match calculations for a 12-locus haplotype between the kappa method and the evolutionary model.

	Region	N haplotypes	N unique haplotypes	N singleton haplotypes	kappa	LR_{kappa}	β_W	p	LR_p	q	LR_q
YHRD	America, Admixed	1206	1003	896	0.8933	11305	0.0013	0.0008	479	0.0005	555
	Africa	1283	989	843	0.8524	8691	0.0114	0.0008	82	0.0007	83
	Asia	3958	3149	2709	0.8603	28327	0.0054	0.0003	177	0.0002	180
	Europe	12523	6538	5018	0.7675	53865	0.0028	0.0001	343	0.0011	254
	America, Native	555	362	275	0.7597	2309	0.0098	0.0018	87	0.0021	84
HGDP	Africa	99	83	72	0.8675	747	0.0196	0.0101	34	0.0025	45
	Asia	238	194	171	0.8814	2007	0.0347	0.0042	26	0.0022	27
	Europe	47	44	41	0.9318	689	0.0086	0.0213	34	0.0013	102
	Middle East	53	36	30	0.8333	318	0.0894	0.0189	9	0.0281	9
XU	Africa	177	155	139	0.8968	1715	0.0208	0.0056	38	0.0010	46
	Asia	316	252	229	0.9087	3462	0.0601	0.0032	16	0.0052	15
	Europe	261	222	195	0.8784	2146	0.0200	0.0038	42	0.0010	48
	Middle East	116	83	69	0.8313	688	0.0336	0.0086	24	0.0080	24
	America, Native	109	64	48	0.7500	436	0.0976	0.0092	9	0.0156	9

LR = likelihood ratio, p = 1/N haplotypes, q = sum of squares of haplotype probabilities, LR_p = LR for a singleton haplotype, LR_q = LR for an average haplotype

1.3.10 *Discussion*

Using our estimator, we found that single-locus β estimates varied by locus, regional ethnicity, and the dataset used. Unlike the autosomal markers, we observed larger β estimates for Y-chromosome markers among African populations. β estimates were also related to mutation rate, with markers with higher mutation rates producing smaller β estimates. Higher mutation opposes the drift in Y-chromosomes, maintaining the diversity of alleles both within and between populations. The heterozygosity within and between populations is more similar, producing smaller β estimates.

We observed very little linkage disequilibrium between Y-chromosome markers. Overall, single-locus β estimates were at best moderately correlated, though in some regions a few markers showed a perfect correlation in β estimates. Correlation between β estimates was associated with LD, but only in the YHRD dataset.

Looking at multi-locus haplotypes, we observed a decrease in β estimates as the number of markers added increased. The β estimates reached their nadir when around 7 markers were included in the haplotype. The change in β estimates as more loci were added to the haplotype for populations the exhibited more correlation between single-locus estimates was smaller than populations with less correlation. Presumably, in both these situations, the addition of another locus is not contributing any new information the β calculation.

To examine the information contributed by each locus more explicitly, we used Shannon entropy. Using a locus' conditional entropy, we found the optimal order of marker addition to a maximally informative haplotype. As with the β estimates, a diminishing amount of entropy was contributed to the total entropy of the haplotype after the addition of between 4 and 8 loci. Calculating the shared information distance between loci showed a complex relationship

between Y-STR marker information. Few markers contributed independent information to a haplotype.

Our work is concordant with the work of Siegert and colleagues, who showed that the close maximum information in a marker panel can be obtained by using only the maximally informative markers -- about half the markers in the full panel.[46] This suggests that it may be prudent to reduce existing panels to their most informative constituents and focus on only adding new markers to panels that contribute independent information. However, this is complicated by the fact that marker information varies between regional ethnicities.

β_W estimates for a 12-locus haplotype varied quite considerably across datasets and regional groups. β_W was smallest for Admixed American samples, possibly due to the contribution of Y-chromosomes from multiple different ancestral groups. Coancestry in one Middle Eastern group and one Native American group was quite high for this haplotype, nearly 10%.

Likelihood ratios calculated using the kappa method were much larger than using the evolutionary model. The kappa method fails to take coancestry into account. In light of the large coancestries observed in several of our populations, with contribute substantially to the probability of a chance match, the kappa method appears anti-conservative.

Chapter 2. USE OF COMMERCIAL DNA DATABASES FOR FORENSIC FAMILIAL SEARCHING: A POLICY PERSPECTIVE

2.1 THE PROBLEM

Familial searching is a forensic technique that has been used by state law enforcement in efforts to identify the source of a crime scene DNA profile using the DNA information of a relative. Typically, it refers to searching for partial matches to a crime stain profile within a state CODIS offender database. The composition of the CODIS offender database can vary by state and may include convicted offenders, arrestees or detainees. For each individual, the database contains a 13 locus STR profile and other identifying information, such as name, date of birth, and race. A partial match between a crime stain profile and a CODIS profile indicates that the CODIS profile may belong to a first-degree relative of the perpetrator. This technique for suspect generation is accepted in the United Kingdom, but controversial in the United States.

Recently, law enforcement officials in the US have expressed interest in applying forensic familial searching to commercial DNA databases used for genetic genealogy. In March 2016, 23andMe, a direct-to-consumer genetic testing company, reported four court orders from US law enforcement officials for user data, including genetic data, from five user accounts in their database.[64–67] 23andMe successfully persuaded officials to withdraw the requests, arguing that present technical and legal challenges made it infeasible to use 23andMe's data for familial searching. In 2014, AncestryDNA, the direct-to-consumer ancestry testing service of the genealogy search website Ancestry.com, assisted law enforcement in response to a warrant request for access to its database. As discussed below, that acquiescence led to a police

investigation of Michael Usry Jr., who was ultimately eliminated as a suspect following additional DNA testing.

Prior to this use in the criminal context, the biggest potential harms for commercial DNA database users and their relatives concerned kinship, identity and health: learning family members are or are not related to them in the way they thought they were (e.g., non-paternity); learning something about their ethnicity that disagrees with their personal identity; or learning about a troublesome health risk (e.g., 23andMe users receive some health-related information). However, the prospect of forensic familial searching in those databases subjects consumers and their relatives to the risk of exposure to undue law enforcement scrutiny. With an estimated five million worldwide users of US commercial DNA databases, worldwide, and their relatives, the potential for familial searching makes this risk is relevant to a significant portion of the US population. Consumers who share their genetic information with commercial genetic databases are unaware of potential access to and uses of their data by law enforcement and other interested parties which may expose them and their relatives to tangible and intangible harms.

2.1.1 *Motivating Case*

AncestryDNA, the world's largest consumer DNA database, responded to a warrant from law enforcement officials to attempt to shed light on a cold case from Idaho Falls, Idaho. In 1998, Christopher Tapp was convicted and sentenced to life in prison for the 1996 rape and murder of 18-year-old Angie Dodge.[68] However, the only DNA sample recovered from the crime was not a match for Tapp. Law enforcement officials believe that someone else, in addition to Tapp, was involved in the crime but had no suspects. In hopes of generating a lead, Idaho law enforcement obtained a warrant granting access to AncestryDNA's genetic information and submitted the crime sample from the Dodge case to AncestryDNA for additional analysis.

AncestryDNA genotyped the sample for Y-chromosome STR markers – the same type of markers used forensically -- and searched the resulting profile against its extensive database.[69] This search revealed 41 partial Y-profile matches. The top match, with 34 of 35 alleles matching, after another warrant compelled AncestryDNA to reveal the identity of the donor, belonged to Michael Usry Sr., a resident of Mississippi.

Encouraged by the Mormon Church, Usry Sr. had donated a DNA sample and pedigree information to the Sorenson Molecular Genealogy Foundation – a non-profit, genealogy research study.[70] The database of more than 100,000 samples, though consented for research only, was acquired by AncestryDNA in 2012.[71]

Because Michael Usry Sr. was not a perfect match and otherwise did not fit the profile of the suspect, law enforcement officials looked at five generations of Usry's extended pedigree, concentrating on three paternal male relatives, including Usry's son, Michael Usry Jr., then living in New Orleans. The younger Usry was 19 years old at the time of the crime, similar to Tapp and Dodge, was in Idaho vacationing with friends coinciding with the crime, and happened to be an indie horror movie filmmaker who co-produced a film called "Murderabilia." With the combination of this circumstantial evidence and the DNA evidence, Idaho law enforcement thought they had found their suspect.

Law enforcement officials set up a meeting with Usry Jr. on the pretense of investigating a hit-and-run incident. Two officers and a federal agent escorted Usry Jr. downtown to an interrogation room and proceeded to interview him – for six hours -- not about a hit-and-run, but his travels to Idaho. Usry Jr. reported to *The New Orleans Advocate* that it only occurred to him as the federal agent requested a DNA sample with the sample kit ready that he might want to consult a lawyer, but volunteered the sample anyway.[72]

Forensic testing of the sample provided by Usry Jr. took about a month and he remained under suspicion during that time, but was never arrested. Results of the analysis conclusively excluded Usry Jr. as the perpetrator.

2.1.2 *Tangible and intangible harms*

As hinted to in the motivating case, there are a series of tangible and intangible harms that could be realized from familial searching of commercial DNA databases. At the forefront is concern for the privacy of the relative of the DNA database user.

Familial searching is at odds with a normative definition of privacy proposed by government privacy scholar Alan Westin as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."[73] The relative of an offender in a CODIS database or user in a commercial DNA database has not been compelled or chosen to expose his or her genetic information in a database, but a part of that genetic information exists in such a database nonetheless, and could be used, along with public records, for identification. Under Westin's definition, relatives could argue that their privacy rights are being violated by a search of the database even though they have not personally deposited their genetic profile into the database.

From a social and ethical perspective, violation of privacy is distressing. A person who believes a privacy right has been violated feels a loss of personal control, as well as negative impacts to self-esteem and emotional well-being.[74,75] The potential for such a harm is evident in the motivating case. Usry Sr.'s DNA, which was specifically collected for research purposes; he did not expressly authorize public sharing of that information. In public reports he expressed distress that his privacy was violated. Associated Press reporter Paul Elias quotes an interview

with Usry Sr.: “it does bother me that Sorenson sold that information after they told me it wouldn’t be shared. It does bother me that my DNA was used for this.”[65]

Legal scholar David Kaye argues persuasively that familial searching via CODIS databases would not be subject to the Fourth Amendment’s prohibition on unreasonable searches and seizures.[76] A summary of his argument is as follows. First, through interaction with the justice system, offenders have a diminished privacy interest in their personal identity. Second, a major part of a DNA search is the physical intrusion required to obtain the sample, which has already been obtained and is not present in a database search. In a familial search, the person being searched is not the relative (who may be the perpetrator), but the previous offender in the CODIS database, and this search is minimally intrusive to the offender in the database. Thus, no physical search of the relative is required. Third, the familial search itself does not reveal any private information, nor does it require that information that is discovered in the course of investigating the lead produced by the familial search is revealed to either the offender in the CODIS database or the relative being investigated.

Familial searching using commercial DNA databases is likely a reasonable search under the Fourth Amendment analogous to CODIS offender database searches. Consumers of commercial DNA databases have diminished privacy interest in their genetic information because they have shared it with a third party. The Supreme Court has concluded that Fourth Amendment protections do not apply to information exposed to third parties.[77] As with a search of a CODIS offender database, the person searched is not the relative, but the commercial DNA database user, whose privacy interest is diminished.

In addition to the negative emotions associated with feeling one’s privacy has been violated, relative identified as a suspect via a familial search may experience other psychosocial

harms. The experience of being identified as a suspect through a familial search is itself intensely stressful.[78] In addition to the stress of being under suspicion, if suspect-status becomes more widely known, a person may experience diminished reputation and may be perceived to be guilty.

Those harms can extend to the DNA database user. The person whose DNA profile is in the commercial DNA database will, by definition of familial searching, not be the person sought by police. The database inhabitant may experience guilt for exposing a relative to an ordeal akin to that felt by carriers who pass on a genetic disease to their children or "survivor's guilt" that they themselves do not have to deal with law enforcement scrutiny.[79–81]

While also a risk of participating in a commercial DNA database, the familial search may ultimately uncovers an unknown family relationship or non-paternity event.[82] This revelation may change family dynamics and increase feelings of guilt by the commercial DNA database user as the instigator of family disruption.

A more tangible harm that a person identified through a familial search of a commercial DNA database may experience are the financial costs resulting from required interaction with law enforcement. Those costs would likely include the costs of legal counsel and associated investigation as well as loss of income for leave from work to interface with the justice system.

Loss of employment is another harm of being identified as a potential suspect through a low prior probability DNA search. A DNA dragnet is a technique for DNA searching when law enforcement has no suspects, but believes the perpetrator is within a defined geographic location and exhibits certain traits. Officials will ask, for example, all men age 18-45 in a town to volunteer a DNA sample to compare against the crime stain. Each man has a low prior probability of being the perpetrator. As one example, Blair Shelton submitted a DNA sample to

Ann Arbor police as part of a DNA dragnet involving a rape in 1994. Though Shelton's DNA did not match the crime sample and he was not arrested in connection with this crime (another man, Ervin Mitchell, Jr., was ultimately convicted of the rape), Shelton was fired from his job after a detective informed his manager that he was a suspect in a rape investigation.[83]

Such accusations and unproven claims maybe particularly problematic for someone who works with vulnerable populations, such as children or the elderly, as employers are unlikely to take risks concerning client safety.

Finally, being identified as a suspect through familial searching may lead to a miscarriage of justice. While a rare occurrence, errors in forensic genetics have led to erroneous indictments, convictions, and incarcerations in the U.S. and abroad.[84–87] A fresh DNA sample from the putative suspect is required to validate a familial searching match, due to the partially matching nature of the initial relative match, so an erroneous match and conviction stemming from it are very unlikely in a familial searching scenario, whether it be by CODIS database search or otherwise.[76]

One potential scenario where a miscarriage of justice could occur in a familial search context is very similar to the Usry case. Given a 34 of 35 allele match on AncestryDNA's Y-profiling platform, it is quite likely that Usry's profile would have been a 23 of 23 match on a standard CODIS Y-profiling platform. If a Y-chromosome profile is the only perpetrator DNA available and a suspect match is identified through a familial search of this lineage marker, a confirmatory sample will also match and there will be no autosomal profile to exclude the suspect.

In addition, members of the public, including law enforcement officers and jurors, tend to give exceptional status to DNA evidence. A 2008 study found that people surveyed viewed DNA

evidence as more accurate and more persuasive than all other types of forensic evidence.[88] Studies consistently show that mock jurors are more likely to convict a suspect when presented with a scenario involving DNA evidence than when it is absent. Mock jurors were three times more likely to convict the defendant in a weak, circumstantial case if DNA evidence was presented than if there was no DNA evidence.[89] Eighty-eight percent of those surveyed in the above 2010 study were somewhat or much more likely to convict if there was DNA evidence linking a defendant to the crime.[90] Most people also view DNA evidence as infallible. Gallup polls in 2000 and 2005, and a scientific survey in 2010 showed that around 80% of those polled thought DNA evidence was either completely or very reliable.[90,91] A 2009 survey in New Zealand agreed with these results; 82% of those surveyed ". . . fully confident in the ability of DNA to provide an error-free tool for solving crime."[92]

In a study involving a database trawl scenario, where the evidence was the same, mock jurors were also 60% more likely to convict if DNA evidence was presented before conventional evidence.[93] This appears to be an example of what the philosopher Francis Bacon called the "primacy effect". In 1620, he noted, "the first conclusion colors and brings into conformity with itself all that come after." The "primacy effect" is driven by a cognitive bias called confirmation bias – a tendency to interpret new information in a way that validates a preconceived theory. Confirmation bias can come in many forms including: restricting attention to favored hypotheses, preferential treatment of evidence supporting existing beliefs, looking for only positive cases, or overweighting positive conforming instances.[94]

As it has been noted that stakeholders are more prone to confirmation bias, it is not surprising that law enforcement officials are not immune to its effects.[95] In fact, Peter Gill, a well-known forensic scientist, has coined a term related to confirmation bias and law

enforcement – The Naïve Investigator Effect – as the "tendency for DNA evidence to override any neutral or exculpatory evidence."[85] This can be seen as the exceptional status given to DNA evidence interacting with confirmation bias. News reports of the Usry case indicate this phenomenon may have been affecting the investigating officers. In light of a DNA match, the circumstantial evidence –ties to Idaho (not unexpected, given that Usry Sr. is Mormon) and made horror movies – looked much more damning. The New Orleans Advocate reported, quoting one of the investigating officers, "All of the circumstantial evidence was right. He seemed like a really good candidate."[72]

An infamous example of confirmation bias interacting with the exceptional status of DNA evidence is the Adam Scott case in the UK.[85,86] A contamination event linked Adam Scott to a rape in a town he had never been to and cell phone records indicated he was 2,000 miles away from at the time of the rape. Investigators ignored this exculpatory evidence and Scott spent three months in jail until the contamination was uncovered. Unlike the Scott case where additional evidence was exculpatory, the circumstantial evidence in the Usry case, perhaps seen in the light of confirmation bias, corroborated the hypothesis that Usry was the perpetrator. Michael Usry Jr was lucky there was an autosomal profile from the crime scene profile available to definitively rule him out as a suspect. If there hadn't been, Usry may have been involved in a miscarriage of justice.

2.2 TECHNICAL CONSIDERATIONS

In order to evaluate the problem of familial searching in commercial DNA databases, it is helpful to review how familial searching is conducted in forensic DNA databases and how commercial DNA databases differ from forensic DNA databases.

The United Kingdom, Australia, New Zealand and ten US states (Arkansas, California, Colorado, Florida, Michigan, Texas, Utah, Virginia, Wisconsin, and Wyoming) include familial searching among their DNA investigation techniques.[96] Most familial searching programs are modeled after the familial searching program within the UK's National DNA Database (NDNAD) -- is the oldest and most established in the world. It began in 2002 and, as of 2014, has been attempted in 188 violent "cold cases" and identified 41 suspects.[28] Several features of this program were designed to minimize false positives, prioritize resource allocation within the familial searching program, and protect potential suspects identified via this technique from that harms described above

First, familial searches in the UK's NDNAD are restricted to first-degree relatives. Partially matching profiles are required to have at least one shared allele at each locus for parent-child searches and 11 total matching alleles for full sibling searches. First degree relatives share a greater number of alleles identically by descent than less closely related individuals (e.g. grandparent-grandchild pairs or first cousins) so one is less likely to observe that number of alleles matching by chance than including second- or third-degree relatives in a familial search. Put more simply, restricting analyses to first-degree relatives only results in fewer false positives than including higher order relative pairs.

Second, the lists generated from the partial matching searches are then prioritized using a likelihood ratio approach called the Kinship Index (KI). This is the probability of the genetic information of two people given specified relationship type, divided by the probability of that information, given they are unrelated. KI of greater than one favors the individuals being of the specified relation and the higher the KI, the more likely they are of that relation. For autosomal markers, it is computed by comparing the probability of observing the genotypes (G) given the

probability alleles are identical-by-descent (k) for the hypothesized relationship to the probability of observing the genotypes given the hypothesis that the individuals are unrelated:

$$KI_A = \frac{\Pr(G|k_0,k_1,k_2)}{\Pr(G|k_0=1,k_1=0,k_2=0)}. \quad \text{(Equation 2.1)}$$

For a full sibling relationship, $k_0 = 0.25$, $k_1=0.50$, and $k_2=0.25$. For parent-child relationships, $k_1=1$, $k_0=k_2=0$.

Ranked lists are then filtered against non-genetic information, such as age, location and ethnicity, in the Police National Computer, which is linked to the NDNAD. These filters, particularly location, are very helpful to reduce the number of leads to follow-up and prioritize the order of investigation. Y-STR profiling and/or mitochondrial DNA profiling (which is a lineage marker inherited maternally) can then be conducted on stored samples to exclude individuals outside the putative lineage and further exclude false positives without re-contacting the NDNAD offender, protecting some of the privacy of the NDNAD offender and his or her relatives.

Lastly, in order to ensure this technique is used properly and justly, familial searches must be approved by the NDNAD Strategy Board and is monitored by the NDNAD Ethics Board. These oversight boards can help mitigate the effects of confirmation bias at the level of the investigator and help balance the needs of the investigation with the concerns of the NDNAD offender and relatives sought.

In the US, SWGDAM has made recommendations for reporting partial matches found in CODIS searches, which can be found in Appendix B (Figure B.1). To limit false-positive matches, SWGDAM recommends that the crime scene profile searched should be from a single source and that all the available CODIS loci should be typed, using a minimum of 10 loci. SWGDAM also recommends additional typing of lineage markers to exclude unrelated matches.

Because the number false-positive matches are related to database size, SWGDAM recommends first searching in smaller State CODIS databases, as opposed to the National CODIS database, and using statistical measures to determine the probative value of a partial match that account for database size. So that partial matches are not misinterpreted, SWGDAM stresses training of DNA analysts. Finally, SWGDAM calls for some quality assurance oversight from the FBI for partial match searches.

States that have robust familial searching programs are generally modeled after the UK program and also follow SWGDAM's guidelines. As an example, California's familial searching program is limited to searching for partial matches in its State CODIS database assuming a parent-child or full sibling relationship, using a single source profile of 15 CODIS markers.[97] Samples are prioritized for follow-up Y-STR typing using a KI calculation. In addition, "DNA Partial Match (Crime Scene DNA Profile to Offender) Policy" from the California Department of Justice (DOJ) requires that all other investigative leads have been exhausted before a familial search can be conducted.[98] A DOJ committee also oversees the familial searching process and requires a review of all available information concerning the case before the name of an offender will be released to law enforcement. It is unclear if the DOJ continues to oversee the investigation once the offender's name has been released.

Protecting against the potential harms of familial searching relies heavily on preventing contacting individuals identified through this technique that are false-positives. Major differences between commercial DNA databases and CODIS databases may cause the former to produce more false-positive results than the latter. These differences are the types of markers used, the degree of relatives that may be sought, and the database composition.

2.2.1 *Markers*

One of the major differences between forensic and commercial DNA databases are the type of markers used for individual identification. Genetic markers on the chromosomes, excluding the sex chromosomes, i.e. autosomal markers, are the most powerful for individual identification. CODIS uses a type of marker called a STR. Thirteen of these markers, spread across the autosomes, make up a CODIS profile, which are used to create both crime stain and suspect profiles. Genetic genealogy databases like AncestryDNA and 23andMe use single nucleotide polymorphisms (SNPs) for their autosomal markers, which can't be used to make a direct comparison with STR markers for forensic matching. While it is possible currently for law enforcement to have a sample processed on a microarray compatible with commercial databases, limited resources for DNA typing and interpretation may pose significant barriers to attempting familial searching in commercial DNA databases. However, forensic scientists are currently investigating the use of genome sequencing – typing all the base pairs in a person's genome – for forensic identification.[99] Both STR and SNP markers can be inferred from genome sequence data. Sequencing is targeted at currently used forensic markers to maintain the utility of existing CODIS databases, but depending on how this information is stored, genetic information from sequenced offenders could be compatible with both existing CODIS databases and commercial genetic databases, which may make familial searching commercial genetic databases even more appealing to law enforcement in the future.

The point of overlap currently between forensic DNA and commercial DNA platforms are STR makers on the Y-chromosome. Y-STR markers were the first DNA markers adopted by the genetic genealogy community. In the male-specific portion of the Y-chromosome, where Y-STRs lie, there is no recombination between Y and X chromosomes. As such, the Y-STR profile

is inherited clonally from father to son (save a mutation rate of 0.002 per marker per generation) and can be used to trace male lineage over many generations.[24,25,40] Surnames, a major cornerstone of genealogy research, are also inherited patrilinearly (excepting in cases of non-paternity or adoption) in many cultures and have a strong correlation with Y-STR markers.[30,31,33,100,101] There are several companies hosting Y-Chromosome Surname Projects, most notably Family Tree DNA, which hosts 9,000 projects.[102–104] Family Tree DNA lists the alleles at each Y-STR marker, surname spelling, ancestor, geographic location, and kit number of the user in a Y-Chromosome Surname Project, freely available to the public, on the website. For an example, see The Ussery Y-Chromosome Project on Family Tree DNA's site.[105]

While the patrilineal clonal inheritance of Y-STRs makes these markers quite useful for genealogy, the same property makes them less than ideal for suspect identification, since all close paternal relatives will have the same Y-STR profile. However, Y-STR profiling remains forensically useful particularly in cases where there is a male-female DNA mixture. If the female autosomal component of the mixture is greater than 10 times the male autosomal component, it is generally not possible to resolve a male autosomal DNA profile.[7,9,10] In such cases, it is often still possible to recover a Y-STR profile. The primers used to recover the profile are Y-chromosome specific and the Y is male-specific, Y-profiling is not hampered by excess female DNA. Because the Y-chromosome profile identifies a patrilineage, it has less probabitive value than an autosomal profile, which can identify an individual, and is typically more powerful to exclude individuals.[37,106] However, a Y-STR profile is virtually useless on its own to identify a close familial relationship in a database search context. While matching Y-profiles indicate relatedness, it is not possible to distinguish whether the two individuals are brothers or 3rd cousins. Identical Y-haplotypes have been observed in men whose common ancestor was alive in

the 1600s, separated by 20 father-son transmissions.[25] Because of this fact, familial search matches identified through a database trawl using Y-STR profile alone are likely to be false positives or incorrectly interpreted as close relatives when in fact the pair are more distantly related.

In the Usry case, presented above, law enforcement officials attempted to apply familial searching techniques to AncestryDNA's database using Y-STR markers alone. This approach does not follow SWGDAM's guidelines for the use and interpretation of Y-STR profiles in forensic genetics or their guidelines for the reporting of partial matches found on database searches.[37,107] It is also markedly different from the procedure of the established UK familial searching program and established US state policies, where partial matching search algorithms look first for putative relatives using autosomal markers, and then follow with additional Y-STR testing to winnow out false positives.

2.2.2 *Relatives sought*

There is a policy consensus among the four US states conducting familial searching and abroad that searches are limited to first-degree relatives. This is a practical consideration owing the type of markers used in forensic databases. There are a limited number of markers, and though each has many alleles (generally around ten), allele sharing by unrelated individuals at multiple loci is not uncommon.[108,109] More distant relatives share fewer alleles identical-by-descent and are therefore less distinguishable from unrelated individuals. As a result, familial search lists for first-cousins, for example, would be unmanageably large.

Familial searching in commercial DNA databases could expand beyond first-degree relatives due to several features of these databases. First, often multiple family members have contributed genetic information to these databases. Though commercial DNA database users

often recruit first and second degree relatives (i.e. parents, siblings, and grandparents) having more distant relatives in a commercial DNA database is common. Researchers at 23andMe found that, after excluding known close relations of users, most of its database inhabitants had at least one 2nd to 9th degree cousin in the database and the most common relationship between pairs was 4th cousins.[49] Pedigree information is also commonly linked to profiles, allowing users and potentially law enforcement to triangulate between hits between multiple users. Pedigree information can also be included in kinship calculations, improving the probative value. Finally, using genome-wide SNP markers or sequence data, the degree of relationship between two individuals can be predicted based on number and length of segments identical-by-descent.[110,111] This is also a feature of 23andMe's DNA Relatives service. As a practical and scientific matter, expanding beyond first-degree relatives will increase the number of people potentially affected by familial searching of commercial DNA databases and will increase leads, but also false positives.

2.2.3 *Database composition and prior probability*

The success of the familial searching technique relies on a relative of the perpetrator being in the database; if no relative is present, all the hits generated will be false positives. In Baysian statistical terms, the prior probability of a relative of the true perpetrator being in the database is greater than zero. Justification for the assumption that a first-degree relative of the perpetrator is among the CODIS offender database sample is that criminality may run in families such that if a person is an inhabitant of the offender database, he is more likely to have a close relative that also offends. This assertion is based on a 1996 survey of jail inmates, where 46.1% of inmates reported that a family member had ever been incarcerated, most often a brother (30.3%) or a father (17.1%).[112]

When applying familial searching techniques to a commercial DNA database, assuming that a perpetrator's close relative is among the database inhabitants may not be a valid assumption. Most genealogical records tend to be from European countries and early US Census, most genealogy cites that link these records have monthly subscription fees of $30 or more, and genetic ancestry testing cost $100 or more (Table B.1). Therefore, it follows that the inhabitants of a commercial DNA database are more likely to be of European ancestry and fairly affluent – this does not match the offender population in the US.[112] If we assume that commercial DNA database inhabitants have relatives that offend at the US population rate of 3%, the prior probability that a relative of a perpetrator is in a commercial DNA database is likely much smaller than a CODIS offender database.[113] Furthermore, the large size of commercial DNA databases compared to state CODIS offender databases does not increase the chance of finding a relative of the perpetrator. In fact, the chance of finding a parent-child match in the top familial search match decreases as database size increases, assuming there is a match to be found.[114–116]

2.3 CRITERIA FOR POLICY ANALYSIS

Having laid out the problem and the context of the issue of familial searching in commercial DNA databases, I will now describe the methods I used to conduct this policy analysis.
I utilized the framework developed by Eugene Bardach.[117] To evaluate the policy alternatives, I applied Beauchamp and Childress' "Principles of Biomedical Ethics" framework.[118] The Beauchamp and Childress ethical framework attempt to balance four principles: Respect for Autonomy, Beneficence, Non-Maleficence, and Justice.

Respect for Autonomy is "to acknowledge [an individual's] right to hold views, to make choices, and to take actions based on their personal values and beliefs." The principle of Beneficence "requires agents to provide benefits to others" and "that agents balance benefits,

risks, and costs to produce the best overall results." Non-Maleficence "imposes an obligation not to inflict harm on others." Justice is the principle that individuals are given "fair, equitable, and appropriate treatment in light of what is due or owed to persons."

Because many of the harms associated with familial searching in a commercial DNA database are intangible and not easily quantifiable, an ethical analysis makes most sense the policy alternatives. This particular ethical framework attempts to balance the needs of the individual and the needs of society, which reflects the composition of the stakeholders involved in this problem.

In order to more easily compare policy alternatives, each Principle for each stakeholder was given a magnitude based on the anticipated impact and likelihood of event occurrence with respect to the policy. This magnitude was based on a number line scale from -10 to +10.

2.3.1 *Stakeholders*

The policy analysis considered the perspectives of four stakeholders: Law enforcement; commercial DNA database service providers; commercial DNA database users; and the relatives of the commercial DNA database users identified through a familial search.

Law enforcement has an interest in apprehending criminals, both to seek justice on the behalf of victims and their families and to preserve public safety. They have a duty to use all viable investigative techniques available to attempt to catch criminals, while appropriately and fairly utilizing public resources to do so.

Commercial DNA database service providers maintain databases of genetic and sometimes pedigree information, along with providing tools and analyses to assist their users with genealogic research. The utility of these databases depends not only on the quality of service provided, but also the size of the database. Simply, the more people in the database the

more connections between individuals can be made and genealogic leads generated. Most of these providers also charge money for their services -- for initial genetic testing and potentially monthly subscription fees – and commercial DNA database profits also proportional to the number of users in their databases. Commercial DNA database service providers have an interest in maintaining a trust relationship with their users so that users will continue to keep their genetic information in the database and actively use the service. Protecting a user's private information, transparent communication about database uses, and preventing unfavorable uses of private information are all mechanisms to promote trust between service providers and their users. If the user finds forensic familial searching of their genetic data to be an unfavorable use, then it is in the best interest of the commercial DNA database provider to resist this use by law enforcement, though they can be compelled to disclose this information to comply with legal processes.

The commercial DNA database user is someone who is interested in genetics and genealogy. Many of these users are serious genealogy hobbyists who do extensive genealogy research and construct detailed family trees. Employing genetic testing and connecting with other genetic relatives is a way to validate relationships identified through paper records, like birth and death records or Census data, discover new branches on their family tree, and tap into genealogy research already conducted by a genetic relative. These users are benefited by other users providing detailed information about their relatives, so that connections can be more easily deduced and family tree branches more easily populated. Commercial DNA database users have an interest in keeping their personal information private, but not from other users so much as interested outside third parties such as law enforcement. The DNA database user is not being targeted as a suspect under a familial searching technique, by definition. They may experience some harms related to their information being used for this type of search, but the bulk of the

harms will be visited upon the relative identified as a suspect. The relative also does not realize the benefits of being connected to the commercial DNA database, as the user does, unless the relative is also interested in genealogy and this information is communicated to them through other means. While the profile in the database is unique to the user, due to the inherited nature of DNA markers, the *information* contained within the profile is not unique to the user, but rather shared within a family. So even though the relative of the DNA database user does not have a genetic profile in the database, they do have genetic information contained in the database, as well as any other personal information the user chose to input. However, the relative had no say in the creation of these database records and may not know that they exist. Where they made aware, the relative may feel that the existence of these records is a violation of his or her privacy and may damage the relationship between the user and the relative, it one exists.

2.4 POLICY OPTIONS TO PROTECT USERS AND THEIR RELATIVES

From the perspective of these four stakeholders, I analyzed one existing "Status Quo" policy developed by commercial DNA database service providers and five policy alternatives. To create a consensus "Status Quo" policy, I reviewed the Privacy Policy and Terms of Service documents of the five commercial DNA databases in the US that provide genotyping services using the autosomes, X and/or the Y chromosomes; AncestryDNA, 23andMe, Family Tree DNA, National Geographic's The Genographic Project and YSEQ. A review of these documents revealed that each privacy policy provides that personal information of users may be disclosed to law enforcement when required to do so by law or legal investigation. All except YSEQ had some mechanism for the user to remove personal and genetic information from the database, with some exceptions for the limits of internet-based technology and other activities the user has

agreed to, such as participation in research. A summary of this review can be found in the appendix. (Table B.1) From this review, the "Status Quo" policy was constructed as:

> **Status Quo Policy:** Commercial DNA database service providers do inform users that their information can be accessed by law enforcement and provide a mechanism for the user to remove his or her information.

This policy respects the autonomy of users by informing them that their data can be accessed by law enforcement and gives them the option to participate in the service. By supplying a mechanism for users to remove their data, the user has the continuing autonomy to participate or not in the future. The user benefits from having this information and both the user and the database service provider benefits from the trust relationship cultivated by transparency. This policy has a neutral impact on law enforcement – it neither hinders nor helps them utilize this information forensically above the rights given by law. The relative, however has no autonomy over their genetic or personal information being a part of the information contained in the database. While not actively harmed by this policy, there is nothing in this policy that prevents harm to the relative either. From the principle of Justice, the harms of being identified through a familial search are still disproportionately realized by the relative.

> **Policy Alternative #1:** Commercial DNA database service providers could inform users of the potential legal impact of using the service on themselves *and their relatives*.

Truly informed consent requires that users are not only told that an activity using their information may potentially occur, but also the risks and benefits of that activity. When law

enforcement can use the information in commercial DNA databases for familial searching, the risks extend to relatives of the user. Depending on the relationship between the user and his or her relatives, and the degree of relationship sought by the familial search, information the user has about his or her relatives' wishes and other characteristics, and the user's perceived likelihood of a familial search taking place, he or she may reconsider contributing a genetic profile to the database.

Policy Alternative #1 respects the user's autonomy by fully informing of the risks and benefits of participating in the service. The relationship between the user and the database provider benefits from more transparency. The database provider may be harmed by this policy if users decide not to contribute information to the database or remove their information, as database utility and profits are proportional to the number of database users. However, it is probable that few users will decide not to continue using the service upon being informed of these risks, resulting in minimal harm to the database service providers. If a user decides not to contribute or removes his or her profile, the relatives of that user are now protected from the harms of familial searching, as their information will no longer be represented in the database. As stated above, this is likely to be a rare occurrence. This policy does not give the relative of the database user any additional autonomy however, and they still bear the bulk of the harms associated with familial searching. This policy option is also fairly neutral to law enforcement, though the utility of the familial search may be lessened by users choosing to remove profiles from the database.

Policy Alternative #2: Commercial DNA database service providers could inform users when and under what circumstances their personal information has been requested by law enforcement.

In this policy scenario, the passing on of timely information helps the user to infer the risks of continuing participation in the service, allowing them to make a more informed decision to participate. However, in the context of the current search, the user is not able to remove or otherwise protect his or her information, which may lead to a feeling of a loss of autonomy. Knowing that a mechanism is in place to communicate between the database service provider and the user when the user's personal information is requested will benefit the database providers and the users trust relationship. However, knowing about the actual search or request for information may harm the user by inspiring feelings of anxiety. While a search or request for information is likely to be rare, once a user has been searched, he or she is likely to remove his or her genetic and personal information from the database and discontinue using the service, to prevent further searches. This will negatively impact the database service providers and law enforcement as in Policy Alternative #1, if users remove their information. And Policy Alternative #2 does little to respect the autonomy of relatives over their information or protect them from the harms of familial searching.

Policy Alternative #3: Commercial DNA database service providers could adopt site acceptable use guidelines following the National Genealogical Society recommendations for sharing information with others.

The National Genealogical Society (NGS) is a non-profit organization that is more than a century old with the mission to promote accurate and ethical genealogy.[119] The NGS has published as set of guidelines for how and when information should be shared with others.[120] The full list of guidelines is available in the appendix (Figure B.), but the three guidelines most relevant to the problem at hand, paraphrased, are: to inform people how family in may be used, respecting any reservations or conditions imposed about use; require evidence of consent from living people before sharing their information; using personal information, including genetic information, only in ways that have been agreed to.

This is the first policy alternative considered to expressly consider the autonomy of the relatives of the DNA database users by requiring their consent to have their information included in the database. It also attempts to address the privacy concerns of the relatives. However, the gains to Autonomy and Beneficence for the relative are limited because the relative still relies on the user to comply with the guidelines. Enforcement of these guidelines by the database service providers would be very difficult and harmful to this stakeholder if required to allocate resources to monitor user generated content. The autonomy of the user is slightly diminished by this policy because they no longer have total control over what information they include in the database. Users may also feel harmed by this policy if they can't post or access all the information about their or other users' relatives to conduct research or complete family trees. Knowledge of an ethical standard is beneficial to the users and may help preserve the relationship between a user and their relative in the event of a familial search, if the standard has been followed.

Policy Alternative #4: Law enforcement could be prohibited from searching commercial DNA databases for partial matches to users who are inferred to be the relatives of potential suspects.

Policy Alternative #4 is the first policy considered here that has a negative impact on law enforcement. By prohibiting familial searching using commercial DNA databases, law enforcement officers lose the autonomy to choose to utilize this investigative technique. This may also negatively impact the principles of Non-Maleficence or Justice if not being able to utilize this familial searching technique harms the investigation or contributes to a perpetrator remaining free. However, for the technical reasons described above, familial searching in commercial DNA databases are not likely to produce probative results. "Cold hit" database trawling within CODIS databases is only successful in identifying a perpetrator 10% of the time; adding a familial search to trawls of these databases is estimated in increase the "cold hit" success rate to 14%.[121] Law enforcement agencies and public safety in general may benefit from refraining to use this low yield technique, as the resources that would have been spent can be allocated to more impactful search techniques or programs to improve public safety.

The relative of the DNA database user is highly benefitted by this policy. By not allowing this type of familial search, all the harms associated with it are totally prevented for the relative, though this policy does not give them any additional autonomy over their personal information. The database user is also benefited by the removal of this threat to privacy. The database service providers will see this as a benefit from a public relations standpoint.

Policy Alternative #5: Law enforcement could utilize established forensic familial searching procedures when conducting familial searching within commercial DNA databases.

As discussed earlier, these procedures limit familial searching to situations when all other techniques have been exhausted, and searching for only first-degree relatives using a single source profile, autosomal markers augmented with a lineage marker profile (typically a Y-chromosome profile), and prioritizing suspects using appropriate statistical calculations and non-genetic information. These steps minimize the number of false positive leads generated by the technique. Familial searching should also be overseen by a committee or board of ethics removed from the daily investigation, who also approves the results of the familial search investigation before potential suspects are contacted. This helps to mitigate confirmation bias on the part of the investigator. Reducing false positives and requiring oversight and approval are steps that actively attempt to protect the relative identified through a familial search from the harms listed above. However, they still do not have autonomy over their personal or genetic information being present in the database. Law enforcement autonomy to conduct familial searches is slightly reduced by the inclusion of an oversight board, but not removed entirely as in Policy Alternative #4. These procedures will also limit the impact of familial searching on law enforcement resources, which is a benefit to that stakeholder and public safety. This policy has a neutral effect on commercial DNA database service providers and users compared to the Status Quo policy.

Table 2.1 summarizes the magnitude of the impacts of the policy alternatives on the Four Principles by stakeholder described above. Relatives of the DNA database users are the most vulnerable stakeholder. Policy Alternative #1 and Policy Alternative #2 can be eliminated from the list of alternatives because they do little to improve the balance of principles for the relatives of the DNA database users above the Status Quo policy. While Policy Alternative #4 is the best policy from the perspective of the relatives, it is the most negatively impactful for law enforcement. Policy Alternatives #3 and #5 better balance the Four Principle across the stakeholder groups.

Policy Alternative #3 seeks to regulate the commercial DNA database service providers, whereas Policy Alternative #5 seeks to regulate law enforcement. Both would require voluntary adoption by each commercial DNA service provider or each State's DOJ.

Adopting NGS's recommendations for information sharing -- Policy Alternative #3 – would be easy to adopt and implement by the commercial DNA database service providers and only requires adoption by five providers, but it would be difficult to compel users to comply with the NGS's recommendations. The benefits of the policy relies on user compliance, so reluctance to follow the NGS's recommendations by users would limit the impact of Policy Alternative #3.

Policy Alternative #5, to have the most uniform impact, would need to be adopted by the DOJ of each of the 50 states. However, if adopted by a State DOJ, the benefits of the policy would not necessarily be realized by residents of that state because commercial DNA databases are not state specific, unlike CODIS offender databases. This may limit the DOJ justification for adopting Policy Alternative #5. In addition, because of current technical limitation, Policy Alternative #5 effectively eliminates familial searching using commercial DNA databases.

Table 2.1. Magnitude of each policy's impact on the Four Principle by stakeholder

Policy	Stakeholder	Autonomy	Beneficence	Non-Maleficence	Justice
Status Quo	User	+5	+5	0	0
	Relative	-10	-5	0	-10
	Database Provider	0	+5	0	0
	Law Enforcement	0	0	0	0
Alternative #1	User	+10	+7	0	0
	Relative	-10	-4	0	-9
	Database Provider	0	+7	-1	0
	Law Enforcement	0	0	-1	0
Alternative #2	User	+3	+6	-8	0
	Relative	-10	-5	0	-10
	Database Provider	0	+6	-2	0
	Law Enforcement	0	0	-2	0
Alternative #3	User	+3	+7	-3	0
	Relative	+2	+2	0	-5
	Database Provider	0	0	-5	0
	Law Enforcement	0	0	0	0
Alternative #4	User	+5	+7	0	0
	Relative	-10	+10	0	0
	Database Provider	0	+7	0	0
	Law Enforcement	-10	+2	-4	-2
Alternative #5	User	+5	+5	0	0
	Relative	-10	+2	0	-2
	Database Provider	0	+5	0	0
	Law Enforcement	-4	+4	0	0

2.5 Recommendations

My recommendation is that both Policy Alternative #3 and Policy Alternative #5 be implemented. Because they regulate different entities, both policies could be adopted independently. Policy Alternative #5 is supported by science and likely to provide the most protection to relatives of database users if implemented. Policy Alternative #3 addresses some of the autonomy concerns for the relatives of database users that are not addressed by Policy Alternative #5.

Chapter 3. AN ASSOCIATION STUDY OF OBESITY AND Y-CHROMOSOME SNPS AMONG HISPANIC MEN

3.1 INTRODUCTION

A 2015 study of men with obesity found that microdeletions in the male-specific region of the Y-chromosome (MSY) were more common among men who were overweight or obesity than men of normal weight.[122] The authors hypothesized that the association between microdeletions in the MSY and excess weight may be mediated by testosterone. Testosterone and obesity have bi-directional relationship. Testosterone and other androgens inhibit the adipocyte differentiation, limiting fat deposition.[123] Without this inhibitory mechanism, men with low testosterone have an increase in body mass. At the same time, adipose tissue is very hormonally active. Aromatase, an enzyme found in adipose tissue, converts testosterone to estrogen, leading to lower androgen function in obese men. In addition, other hormones secreted by adipocytes appear to downregulate testosterone production in the Leydig cells via the hypothalamus-pituitary axis or directly, leading to a testosterone-obesity cycle.[124]

Both obesity and low testosterone are common among Hispanic men in the US. In an analysis of data from the 2011-2012 National Health and Nutrition Examination Survey (NHANES), 44% of Mexican American were obese and 27% had low testosterone (<300 ng/mL).[125] Among men of Hispanic ethnicity outside Mexico, 37% were obese and 29% had low testosterone. Both Class I (BMI between 30 and 34.9 kg/m^2) and Class II (BMI greater than or equal to 35 kg/m^2) obesity were associated with low testosterone. Mexican American men with Class I obesity were nearly 3 times and Mexican American men with Class II obesity were 6.5 times more likely to have low testosterone compared to normal weight Mexican American men

(OR [95%CI]: 2.69 [1.24-5.84], 6.51 [2.87-14.78], respectively). Men of Hispanic ethnicity outside Mexico with Class I obesity were 4 times and men with Class II obesity were nearly 16 times more likely to have low testosterone than normal weight men of the same ancestry (OR [95%CI]: 3.97 [1.37-11.56], 15.96 [4.64-54.97], respectively).

Genome-wide association studies (GWAS) typically do not include the sex chromosomes because their properties do not lend well to procedures used to analyze the autosomes. Males and females have different dosages of the X-chromosome. X-inactivation in women randomly silences one of the X-chromosome in a cell, meaning different cells may be expressing a different X. The Y-chromosome present in men only and inherited as a unit clonally from one's father, meaning all markers on the Y-chromosome are highly linked. Methods are being developed to correctly analyze the X-chromosome, but methods for Y-chromosome analysis are lacking.[126,127]

Accounting for population structure on the sex chromosomes is also problematic. Because they have different inheritance patterns compared to the autosomes, relatedness is also different on the X and Y compared to the autosomes. This is particularly important for Hispanic populations as they have a history of strong gender-biased admixture between European males and Native American females.[128] A recent association study using the X-chromosome found that incorporating the first X-chromosome principal component (PC), along with autosomal PCs, adjusted for population stratification better than using the autosomal PCs alone.[129]

Because of these challenges, previous studies of the Y-chromosome and human disease have focused on analyzing associations with the whole MSY region, using restriction fragment length polymorphism (RFLPs) and Y-chromosome haplotypes. Besides infertility, cardiovascular disease and its associated risk factors have been the most often studied in connection with the Y-

chromosome.[130–139] However, these methods are not able to determine the location of a putative causal variant.

Using single nucleotide polymorphisms (SNPs) for association mapping on the Y-chromosome may provide more information on the location of a causal variant, if the causal variant is observed directly. In order to preserve power, GWAS depends on the selection of tagSNPs – one SNP representing the variation present in one linkage disequilibrium (LD) block. LD between markers on the autosomes are broken up over time by recombination and the LD between a pair of autosomal markers is related to their physical distance. The most highly associated SNP is generally not thought to be the causal variant, but is in LD with the causal variant, which is assumed to be physically close to the tagSNP. There is no recombination on the Y-chromosome, but LD can be broken up between markers through mutation. As a result, LD on the Y-chromosome does not depend on the physical distance between two markers. Since there is no relationship between position on the Y-chromosome and LD, if it is not the causal variant itself, a highly associated tagSNP does not give any indication as to the location of the causal variant.

An alternative to selecting tagSNPs based on LD is to select them based on information content. One measure of information, Shannon entropy, applied to genetics, is a measure of the residual uncertainty of an allele given its haplotype background.[41,45] Using conditional Shannon entropy, Hampe and colleagues found that this information-based SNP selection procedure improved efficiency by 30% compared to distance or LD based selection procedures.[140]

Here we present an attempt to analyze the association between maximally-informative Y-chromosome SNPs and obesity among a Hispanic sample, accounting for population structure by incorporating both autosomal and Y-chromosome principal components.

3.2 METHODS

3.2.1 *Participants*

There are 5,122 men in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL).[141] Related participants were excluded. Because low BMI is also associated with hormonal disruption, but via alternate mechanisms, participants with a BMI of less than 18.5 were also excluded. A total of 4,332 men were included in the analysis. Participants have previously been assigned to six genetic groups. The mean and standard deviation of age and BMI were calculated overall and among each genetic group. Analysis of variance was used to determine if age and BMI were different between genetic groups.

Because the hypothesized mechanism for the relationship between obesity and Y-chromosome polymorphisms is via testosterone deficiency, we chose our case status based on a preliminary association analysis of low testosterone and BMI in the 2011-2012 NHANES dataset. Since Hispanic men with Class II obesity were at much higher risk for low testosterone, we chose this extreme phenotype as our case variable.

3.2.2 *SNPs*

A total of 1,864 SNPs in the MSY passed quality control filters. These SNPs were then passed to a Shannon entropy-based pruning algorithm to select the most informative SNPs to analyze. For marker or marker set X, Shannon entropy, H, is calculated as:

$$H(X) = -\sum_{S=1} p_S \ln(p_S), \tag{3.19}$$

where p_S are the allele frequencies of each allele at marker X. The residual entropy, or information, provided by marker X after marker (or marker set) Y is known can be calculated as the conditional entropy of X given Y: $H(X|Y) = H(X,Y) - H(Y)$. The pruning algorithm starts

by choosing the SNP with the maximum entropy value, Y. The entropy of the remaining SNPs is calculated conditional on the entropy of Y. The SNP with the maximal conditional entropy is added to marker set Y, and the process is repeated with the remaining SNPs until the conditional entropy of the remaining SNPs crosses a set threshold. Hampe and colleagues recommends a threshold of 0.1. This reduced our SNP set to 19 markers. We chose a less strict threshold of 0.001, which reduced the SNP set to 595. Further filtering to exclude SNPs with a minor allele frequency of less than 0.05 reduced the SNP set to 145 markers for use in this analysis.

3.2.3 *Regression*

Under logistic model, we regressed the 145 chosen SNPs on the Class II Obesity trait using the assocRegression command in the R package GWASTools. To best adjust for population stratification, we adjusted for combinations of the first {0, 1, 2} autosomal principal components by the first {0, 1, 2} Y-chromosome principal components. Q-Q plots were generated for these combinations. A Manhattan plot was created for the best fitting model.

3.3 RESULTS

Overall, the mean age of participants was 46.5 years and the mean BMI was 29.2 kg/m^2. Eleven percent of the men had Class II obesity. Mean age, BMI, and the proportion of men with Class II obesity different significantly by genetic group (Table 3.2).

Table 3.2. Characteristics of the HCHS/SOL participants included in analysis, by ethnicity

Genetic group	Participants **N**	Age **Mean (SD)**	BMI **Mean (SD)**	BMI >=35 **%**
All	4332	46.5 (13.1)	29.2 (5.1)	11
Central American	474	43.9 (13.1)	29.4 (5.1)	13
Cuban	930	49.9 (12.8)	28.9 (5.1)	11
Dominican	329	46.1 (14.0)	28.8 (4.6)	7
Mexican	1495	44.7 (13.6)	29.2 (4.9)	10
Puerto Rican	787	47.9 (13.6)	29.6 (6.1)	16
South American	298	46.4 (12.8)	28.6 (4.1)	6
Pr(>F) or Pr(χ^2)		<2e-16	0.02	8.78e-7

Figure 1.1 shows the effect different combinations of autosomal and Y-chromosome have on population stratification. Adjusting for zero autosomal and two Y-chromosome principal components resulted in the λ closest to one ($\lambda = 1.032$), but evidence for residual population stratification remained.

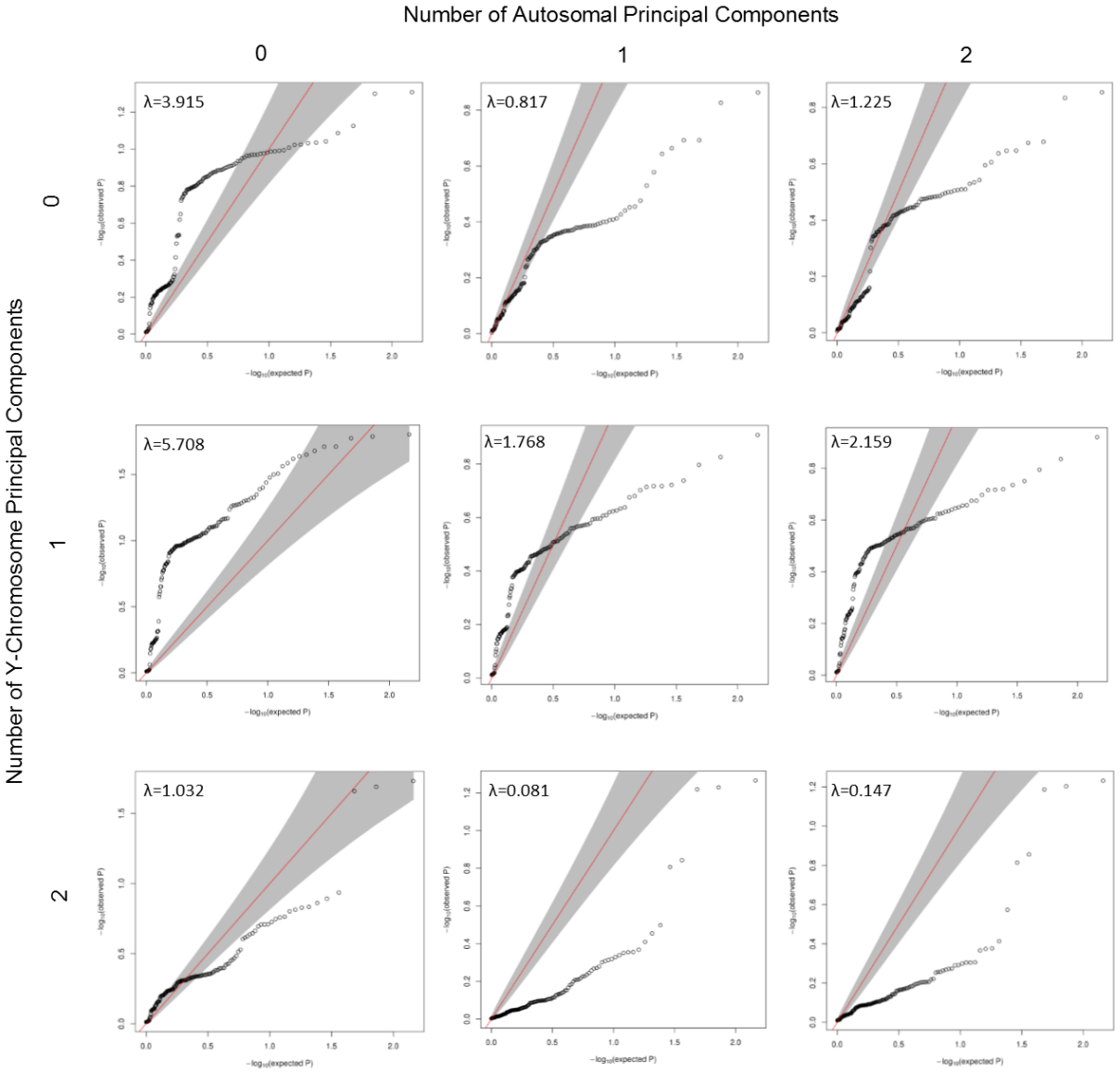

Figure 3.1. Grid of Q-Q plots of regression analysis adjusted by combinations of autosomal and Y-chromosome principal components.

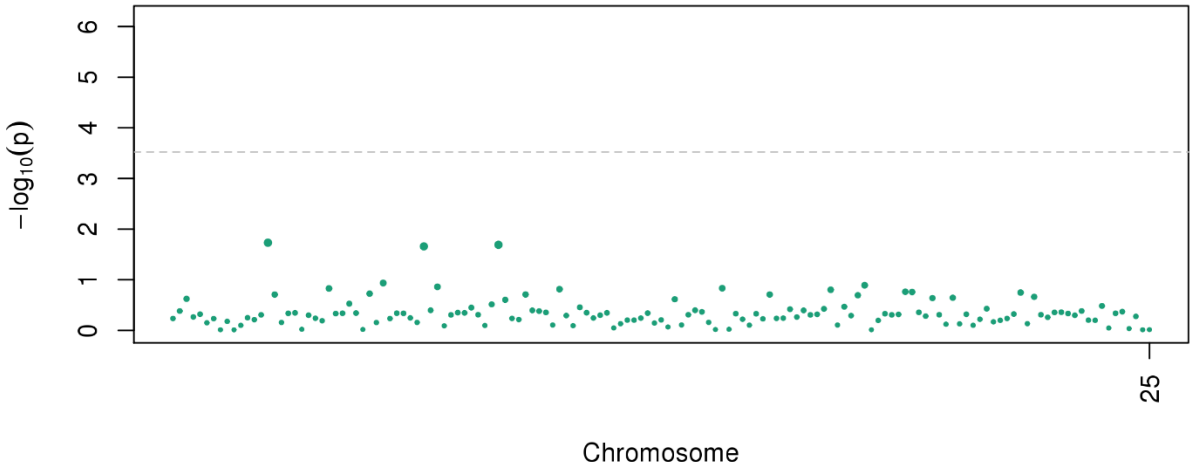

Figure 3.2. Manhattan plot of the association between Y-chromosome SNPs and Class II obesity

Based on the model with the best correction for population stratification, Figure 3.2 shows the Manhattan plot of the association between Y-SNPs and Class II obesity. The top SNPs rs17250803, rs7892876, and rs9341301 each had a p-value of 0.02 which did not meet the Bonferroni corrected significance threshold of 0.0003.

3.4 Discussion

None of the SNPs included in this study reached the Bonferroni corrected significance level of 0.0003. Even though the Y-SNPs were selected to have the least dependencies among them, the Y-chromosome SNPs used in this study are not totally independent. As such, a Bonferroni correction may be too strict. However, none of the SNPs in this study would have reached a more relaxed significance threshold either.

Further investigation of the top SNPs indicate they are likely to represent coincidental associations. SNPs rs17250803 and rs7892876 are located in or near pseudogenes. SNP rs9341301 is located in an intron about 400bp upstream from the sixth exon in the *DDX3Y* gene.

This gene is one of two genes (the other being *USP9Y*) located in the first of three azoospermia factor regions (AZFa). Microdeletions in the AZF region taken in total are a common cause of infertility in men, however microdeletions in this region are seen in fertile men as well. Of all these microdeletions, those located in the AZFa region are the rarest, because its structure makes a deletion event more difficult and deletion has a quite deleterious effect on fertility. Unlike other AZF region genes, *DDX3Y* expression in adult tissue is limited to the testes. Specifically, *DDX3Y* is expressed in highest concentration within the cytoplasm of spermatogonia, the undifferentiated male germ cells that go on to produce sperm cells. This gene appears to have a very specific function in pre-meiotic cell division for spermatogenesis. While the microdeletions associated with obesity in the Biabangard Zak et al. study were also in the AZF region, the study looked only at AZFc region microdeletions, which have a less severe impact on male fertility than alterations in AZFa region genes. It is unclear the role microdeletions in the AZF region play in testosterone levels, as most research on the AZF region focuses on male infertility caused by abnormal or absent spermatogenesis. A few studies included measures of hormone function. Among men with variant Y-chromosomes, men with Y-chromosome translocations had lower levels of testosterone, but men with microdeletions were not different from controls.[142] A study of AZFc microdeletions among men with an XXY karyotype found that they also had normal levels of testosterone, though other hormonal abnormalities were present.[143]

It is perhaps unsurprising that the best correction for population structure in the present study of Class II obesity and Y-chromosome SNPs came from adjusting for the first two Y-chromosome principal components, given the special genetic features of this chromosome. However, residual population stratification remained that could not be accounted for by adding

either more Y-chromosome PCs or autosomal PCs. The best way to proceed with association mapping on the Y-chromosome still requires some additional thought and experimentation.

BIBLIOGRAPHY

1. Butler, J. M. Genetics and genomics of core short tandem repeat loci used in human identity testing. *J. Forensic Sci.* **51,** 253–265 (2006).
2. Gill, P. *et al.* Considerations from the European DNA profiling group (EDNAP) concerning STR nomenclature. *Forensic Sci. Int.* **87,** 185–192 (1997).
3. Olaisen, B. *et al.* DNA Recommendations 1997 of the International Society for Forensic Genetics. *Vox Sang* **74,** 61–63 (1998).
4. Budowle, B. *et al.* CODIS and PCR-Based Short Tandem Repeat Loci: Law Enforcement Tools. in *The Second European Symposium on Human Identification* 73–88 (1998). at <https://www.promega.com/~/media/files/resources/conference proceedings/ishi 02/oral presentations/17.pdf>
5. Kayser, M. *et al.* Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Leg. Med* **110,** 125–133 (1997).
6. Daniels, D. L., Hall, A. M. & Ballantyne, J. SWGDAM developmental validation of a 19-locus Y-STR system for forensic casework. *J. Forensic Sci.* **49,** 668–683 (2004).
7. Prinz, M., Ishii, A., Coleman, A., Baum, H. J. & Shaler, R. C. Validation and casework application of a Y chromosome specific STR multiplex. *Forensic Sci. Int.* **120,** 177–188 (2001).
8. Prinz, M., Boll, K., Baum, H. & Shaler, B. Multiplexing of Y chromosome specific STRs and performance for mixed samples. *Forensic Sci. Int.* **85,** 209–218 (1997).
9. Sibille, I. *et al.* Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. *Forensic Sci. Int.* **125,** 212–216 (2002).
10. Krenke, B. E. *et al.* 'Validation of a male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex' [Forensic Sci. Int. 148 (1) (2005) 1-14]. *Forensic Sci. Int.* **151,** 111–124 (2005).
11. Parson, W., Niederstätter, H., Köchl, S., Steinlechner, M. & Berger, B. When autosomal short tandem repeats fail: optimized primer and reaction design for Y-chromosome short tandem repeat analysis in forensic casework. *Croat. Med. J.* **42,** 285–7 (2001).
12. Allard, J. E. The collection of data from findings in cases of sexual assault and the significance of spermatozoa on vaginal, anal and oral swabs. *Sci. Justice* **37,** 99–108 (1997).
13. Quarino, L. & Kishbaugh, J. The utility of Y-STR profiling in four-, six- and eight-day postcoital vaginal swabs. *Med. Sci. Law* **52,** 81–8 (2012).
14. Hanson, E. & Ballantyne, J. A Y-short tandem repeat specific DNA enhancement strategy to aid the analysis of late reported (>=6 days) sexual assault cases. *Med. Sci. Law* **54,** 209–218 (2014).
15. Sween, K. R., Quarino, L. a. & Kishbaugh, J. M. Detection of Male DNA in the Vaginal Cavity After Digital Penetration Using Y-Chromosome Short Tandem Repeats. *J. Forensic Nurs.* **11,** 33–40 (2015).
16. Han, J. P. *et al.* A new strategy for sperm isolation and STR typing from multi-donor sperm mixtures. *Forensic Sci. Int. Genet.* **13,** 239–246 (2014).
17. Ayadi, I., Mahfoudh-Lahiani, N., Makni, H., Ammar-Keskes, L. & Reba??, A. Combining autosomal and Y-chromosomal short tandem repeat data in paternity testing with male child: Methods and application. *J. Forensic Sci.* **52,** 1068–1072 (2007).

18. Chakraborty, R. Paternity testing with genetic markers: Are Y-linked genes more efficient than autosomal ones? *Am. J. Med. Genet.* **21,** 297–305 (1985).
19. Liu, H. M. *et al.* Y-chromosome short tandem repeats analysis to complement paternal lineage study: A single institutional experience in Taiwan. *Transfusion* **47,** 918–926 (2007).
20. Basgalupp, S. P. *et al.* Investigation of paternity with alleged father deceased or missing: Analysis of success at the end of the report. *Forensic Sci. Int. Genet.* **12,** 120–121 (2014).
21. Wenk, R. E., Traver, M. & Chiafari, F. A. Determination of sibship in any two persons. *Trans* **36,** 259– (1996).
22. Santos, F. R., Epplen, J. T. & Pena, S. D. J. Testing deficiency paternity cases with a Y-linked tetranucleotide repeat polymorphism. *EXS* **67,** (1993).
23. Rolf, B., Keil, W., Brinkmann, B., Roewer, L. & Fimmers, R. Paternity testing using Y-STR haplotypes: assigning a probability for paternity in cases of mutations. *Int J Leg. Med* **115,** 12–15 (2001).
24. Ambers, A. *et al.* Autosomal and Y-STR analysis of degraded DNA from the 120-year-old skeletal remains of Ezekiel Harper. *Forensic Sci. Int. Genet.* **9,** 33–41 (2014).
25. Greeff, J. M. & Erasmus, J. C. Appel Botha Cornelitz: The abc of a three hundred year old divorce case. *Forensic Sci. Int. Genet.* **7,** 550–554 (2013).
26. The Federal Bureau of Investigation. Frequently Asked Questions (FAQs) on the CODIS Program and the National DNA Index System. at <https://www.fbi.gov/about-us/lab/biometric-analysis/codis/codis-and-ndis-fact-sheet>
27. *DNA-Sample Collection and Biological Evidence Preservation in the Federal Jurisdiction. Federal Register*. **73,** 74932–74943 (2008).
28. Maguire, C. N., McCallum, L. A., Storey, C. & Whitaker, J. P. Familial searching: A specialist forensic DNA profiling service utilising the National DNA Database® to identify unknown offenders via their relatives - The UK experience. *Forensic Sci. Int. Genet.* **8,** 1–9 (2014).
29. Ge, J., Chakraborty, R., Eisenberg, A. & Budowle, B. Comparisons of familial DNA Database searching strategies. *J. Forensic Sci.* **56,** 1448–1456 (2011).
30. Jobling, M. A. In the name of the father: Surnames and genetics. *Trends Genet.* **17,** 353–357 (2001).
31. Sykes, B. & Irven, C. Surnames and the Y chromosome. *Am. J. Hum. Genet.* **66,** 1417–9 (2000).
32. King, T. E. & Jobling, M. A. Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. *Mol. Biol. Evol.* **26,** 1093–102 (2009).
33. McEvoy, B. & Bradley, D. G. Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. *Hum. Genet.* **119,** 212–9 (2006).
34. Huang, D. *et al.* Y-haplotype Screening of Local Patrilineages Followed by Autosomal STR Typing Can Detect Likely Perpetrators in Some Populations. *J. Forensic Sci.* **56,** 1340–1342 (2011).
35. Evett, I. W. What is the probability that this blood came from that person? A meaningful question? *J. Forensic Sci. Soc.* **23,** 35–39 (1983).
36. Balding, D. J. & Nichols, R. A. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96,** 3–12 (1995).
37. Scientific Working Group on DNA Analysis Methods. *SWGDAM Interpretation*

Guidelines for Y-Chromosome STR Typing by Forensic DNA Laboratories. (2014).
38. SWGDAM. SWGDAM Mission Statement. (2014). at <http://www.swgdam.org/>
39. Oh, Y. N. *et al.* Haplotype and mutation analysis for newly suggested Y-STRs in Korean father–son pairs. *Forensic Sci. Int. Genet.* **15,** 64–68 (2015).
40. Ballantyne, K. N. *et al.* Mutability of Y-chromosomal microsatellites: Rates, characteristics, molecular bases, and forensic implications. *Am. J. Hum. Genet.* **87,** 341–353 (2010).
41. Caliebe, A., Jochens, A., Willuweit, S., Roewer, L. & Krawczak, M. No shortcut solution to the problem of Y-STR match probability calculation. *Forensic Sci. Int. Genet.* **15,** 69–75 (2015).
42. Santos, F. R., Bianchi, N. O. & Pena, S. D. Worldwide distribution of human Y-chromosome haplotypes. *Genome Res.* **6,** 601–611 (1996).
43. Willuweit, S. & Roewer, L. Y Chromosome Haplotype Reference Database. *Forensic Sci. Int. Genet.* **15,** 43–48 (2013).
44. Carracedo, A. *et al.* Update of the guidelines for the publication of genetic population data. *Forensic Sci. Int. Genet.* **10,** 1–2 (2014).
45. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27,** 623–656 (1948).
46. Siegert, S., Roewer, L. & Nothnagel, M. Shannon's equivocation for forensic Y-STR marker selection. *Forensic Sci. Int. Genet.* **16,** 216–225 (2015).
47. Brenner, C. H. Fundamental problem of forensic mathematics - The evidential value of a rare haplotype. *Forensic Sci. Int. Genet.* **4,** 281–291 (2010).
48. Aldrich, J. Probability , Statistics & Political Economy in Mill ' s Logic. 1–43
49. Henn, B. M. *et al.* Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One* **7,** (2012).
50. Weir, B. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution (N. Y).* **38,** 1358–1370 (1984).
51. Balding, D. J. & Nichols, R. a. DNA profile match probability calculation: How to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci. Int.* **64,** 125–140 (1994).
52. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15,** 323–54 (1951).
53. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat. Rev. Genet.* **10,** 639–650 (2009).
54. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23,** 1514–1521 (2013).
55. Weir, B. S. & Hill, W. G. Estimating F-S Tatistics. *Ann. Hum. Genet.* **36,** 721–750 (2002).
56. Buckleton, J. *et al.* Population-specific FST values for forensic STR markers: A worldwide survey. *Forensic Sci. Int. Genet.* **23,** 91–100 (2016).
57. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517,** 327–332 (2015).
58. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324,** 1035–44 (2009).
59. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296,** 261–2 (2002).
60. Xu, H. *et al.* Inferring population structure and demographic history using Y-STR data from worldwide populations. *Mol. Genet. Genomics* **290,** 141–150 (2014).
61. Zhao, J. H. gap : Genetic Analysis Package. *J. Stat. Softw.* **23,** (2007).

62. Chessel, D., Dufour, A. B. & Thioulouse, J. The ade4 package - I : One-table methods. *R News* **4,** 5–10 (2004).
63. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27,** 209–20 (1967).
64. Lynch, J. How Private DNA Data Led Idaho Cops on a Wild Goose Chase and Linked an Innocent Man to a 20-year-old Murder Case. *Electronic Freedom Founadation* (2015). at <https://www.eff.org/deeplinks/2015/05/how-private-dna-data-led-idaho-cops-wild-goose-chase-and-linked-innocent-man-20>
65. Elias, P. Law enforcement investigators seek out private DNA databases. *Associated Press* (2016). at <http://www.bigstory.ap.org/article/e32f553002594ecfa5e83c160b4ba720/law-enforcement-investigators-seek-out-private-dna>
66. Black, K. & Curevac, Z. 23andPrivacy: Your Data and Law Enforcement. *23andMe Blog* (2016). at <http://blog.23andme.com/23andme-and-you/23andprivacy-your-data-law-enforcement/>
67. 23andMe. Transparency Report. (2016). at <https://www.23andme.com/transparency-report/>
68. *State v. Tapp*. *Idaho* **136,** 354 (2001).
69. Y-Chromosome Marker Details. *Sorenson Molecular Genealogy Foundation* (2012). at <https://web.archive.org/web/20120206162449/http://www.smgf.org/ychromosome/marker_details.jspx>
70. Sorenson Molecular Genealogy Foundation. (2012). at <https://web.archive.org/web/20120106054213/http://www.smgf.org/?>
71. Ancestry.com Launches new AncestryDNA Service: The Next Generation of DNA Science Poised to Enrich Family History Research. *Ancestry.com Press Releases* (2012). at <http://corporate.ancestry.com/press/press-releases/2012/05/ancestry.com-dna-launches/>
72. Mustian, J. New Orleans filmmaker cleared in cold-case murder; false positive highlights limitations of familial DNA searching. *The New Orleans Advocate* (2015). at <http://www.theneworleansadvocate.com/news/11707192-123/new-orleans-filmmaker-cleared-in>
73. Westin, A. F. *Privacy and Freedom*. (Atheneum, 1967).
74. Diener, E., Suh, E. M., Lucas, R. E. & Smith, H. L. Subjective well-being: Three decades of progress. *Psychol. Bull.* **125,** 276–302 (1999).
75. Siu, G. E., Bakeera-Kitaka, S., Kennedy, C. E., Dhabangi, A. & Kambugu, A. HIV serostatus disclosure and lived experiences of adolescents at the Transition Clinic of the Infectious Diseases Clinic in Kampala, Uganda: a qualitative study. *AIDS Care* **24,** 606–11 (2012).
76. Kaye, D. H. The Genealogy Detectives: A Constitutional Analysis of 'Familial Searching'. *Am. Crim. Law Rev.* **50,** 109–163 (2013).
77. Skopek, J. M. Reasonable Expectations of Anonymity. *Va. Law Rev.* **101,** 691–762 (2015).
78. Volbert, R. False Confessions in Police Interviews Interviewing-Error or Immanent Risk? *R. Psychiatr.* **34,** 4–10 (2016).
79. Lewis, C., Skirton, H. & Jones, R. Can we make assumptions about the psychosocial impact of living as a carrier, based on studies assessing the effects of carrier testing? *J. Genet. Couns.* **20,** 80–97 (2011).

80. Hutson, S. P., Hall, J. M. & Pack, F. L. Survivor Guilt. *Adv. Nurs. Sci.* **38,** 20–33 (2015).
81. Lehmann, A., Speight, B. S. & Kerzin-Storrar, L. Extended family impact of genetic testing: The experiences of x-linked carrier grandmothers. *J. Genet. Couns.* **20,** 365–373 (2011).
82. Haimes, E. Social and Ethical Familial Searching in Forensic Investigations : Insights from. (2006).
83. Washtenaw County Circuit Court. *Shelton v Police Dept of the City of Ann Arbor, 1995*. (1995). doi:10.1017/CBO9781107415324.004
84. Vincent, F. H. R. *Report -- inquiry int the circumstances taht lead to the conviction of Mr. Farah Abdulkadir Jama*. (2010). at <http://assets.justice.vic.gov.au/justice/resources/4cd228fd-f61d-4449-b655-ad98323c4ccc/vincentreportfinal6may2010.pdf>
85. Gill, P. *Misleading DNA Evidence. Reasons for Miscarriages of Justice*. (Academic Press, 2014).
86. Rennison, A. Report into the circumstances of a complaint received from the Greater Manchester Police on 7 March 2012 regarding DNA evidence provided by LGC Forensics. *Forensic Sci. Regul. Rep.* (2012). at <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/118941/dna-contam-report.pdf>
87. Duster, T. Explaining Differential Trust of DNA Forensic Technology : Grounded Inexplicable. *DNA fingerprinting Civ. Lib.* **Summer,** 293–300 (2006).
88. Lieberman, J. D., Carrell, C. A., Miethe, T. D. & Krauss, D. A. Gold versus platinum: Do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? *Psychol. Public Policy, Law* **14,** 27–62 (2008).
89. Tomison, A., Goodman-delahunty, J. & Hewson, L. Enhancing fairness in DNA jury trials. *Trends issues crime Crim. justice* (2010).
90. Brewer, P. R. & Ley, B. L. Media Use and Public Perceptions of DNA Evidence. *Sci. Commun.* **32,** 93–117 (2010).
91. Gallup Poll. From what you have read or heard, do you think that DNA evidence is completely reliable, very reliable, only somewhat reliable, or not reliable at all? (2000). at <http://www.gallup.com/poll/1603/Crime.aspx>
92. Curtis, C. Public Perceptions and Expectations of the Forensic Use of DNA: Results of a Preliminary Study. *Bull. Sci. Technol. Soc.* **29,** 313–324 (2009).
93. Scurich, N. & John, R. S. Mock Jurors ' Use of Error Rates in DNA Database Trawls. **37,** 424–431 (2013).
94. Nickerson, R. S. Confirmation Bias : A Ubiquitous Phenomenon in Many Guises. *Rev. Gen. Psychol.* **2,** 175–220 (1998).
95. Thurstone, L. L. *The nature of intelligence*. (Kegan Paul, Trench Trubner & Co., 1924).
96. The Federal Bureau of Investigation. Combined DNA Index System (CODIS). at <https://www.fbi.gov/services/laboratory/biometric-analysis/codis>
97. Myers, S. P. *et al.* Searching for first-degree familial relationships in California's offender DNA database: Validation of a likelihood ratio-based approach. *Forensic Sci. Int. Genet.* **5,** 493–500 (2011).
98. Anderson, G. B. & Sims, G. DNA Partial Match (Crime Scene DNA Profile to Offender) Policy. 7–9 (2008).
99. Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat*

Protoc **9,** 2586–2606 (2014).
100. Calderón, R., Hernández, C. L., Cuesta, P. & Dugoujon, J. M. Surnames and Y-chromosomal markers reveal low relationships in Southern Spain. *PLoS One* **10,** e0123098 (2015).
101. Jobling, M. A. & Tyler-Smith, C. Fathers and sons: the Y chromosome and human evolution. *Trends Genet.* **11,** (1995).
102. FamilyTreeDNA. Family Tree DNA Surname, Lineage and Geographical Projects. (2015). at <https://www.familytreedna.com/projects.aspx>
103. iGenia. Surname projects. at <https://www.igenea.com/en/surname-projects>
104. AncestryDNA. (2016). at <http://dna.ancestry.com/>
105. FamilyTreeDNA. Ussery Y-Chromosome Ussery (and Related Spellings) Surname Project. at <https://www.familytreedna.com/groups/ussery/about>
106. Roewer, L. Y chromosome STR typing in crime casework. 77–84 (2009). doi:10.1007/s12024-009-9089-5
107. Scientific Working Group on DNA Analysis Methods Ad Hoc Committee on Partial Matches. SWGDAM Recommendations to the FBI Director on the 'Interim Plan for the Release of Information in the Event of a "Partial Match" at NDIS'. *Forensic Sci. Commun.* **11,** 1–9 (2009).
108. Weir, B. S. THE RARITY OF DNA PROFILES. *Ann. Appl. Stat.* **1,** 358–370 (2007).
109. Weir, B. S. Matching and Partially-Matching DNA Profiles. *J Forensic Sci* **49,** (2004).
110. Browning, S. R. & Browning, B. L. High-resolution detection of identity by descent in unrelated individuals. *Am. J. Hum. Genet.* **86,** 526–39 (2010).
111. Li, H. *et al.* Relationship Estimation from Whole-Genome Sequence Data. *PLoS Genet.* **10,** (2014).
112. Harlow, C. W. Correctional populations in the United States, 1996. *Bur. Justice Stat. Bull.* (1999). at <http://www.bjs.gov/content/pub/pdf/cpus14.pdf>
113. Kaeble, D., Glaze, L., Tsoutis, A. & Minton, T. Correctional Populations in the United States, 2014. *Bur. Justice Stat. Bull.* (2015). doi:10.1001/jama.295.13.1549
114. Bieber, F. R., Brenner, C. H. & Lazer, D. Supporting Online Material.Finding Criminals Through DNA Testing of Their Relatives Finding Criminals Through DNA Testing of Their Relatives. *Science (80-.).* (2006). doi:10.1126/science.1122655
115. Gershaw, C. J., Schweighardt, A. J., Rourke, L. C. & Wallace, M. M. Forensic utilization of familial searches in DNA databases. *Forensic Sci. Int. Genet.* **5,** 16–20 (2011).
116. Kruijver, M., Meester, R. & Slooten, K. Optimal strategies for familial searching. *Forensic Sci. Int. Genet.* **13,** 90–103 (2014).
117. Bardach, E. *A Practical Guide for Policy Analysis: The Eightfold Path to More Effective Problem Solving*. (SAGE Publlications, 2012).
118. Beauchamp, T. L. & Childress, J. F. *Principles of Biomedical Ethics*. (Oxford University Press, 2012).
119. The National Genealogical Society. About NGS: NGS Bylaws. (2014). at <http://www.ngsgenealogy.org/cs/ngs_bylaws>
120. The National Genealogical Society. Guidelines for Sharing Information with Others. (2016). at <http://www.ngsgenealogy.org/galleries/Ref_Researching/Guidelines_SharingInfo2016.pdf>
121. Bieber, F. R., Brenner, C. H. & Lazer, D. Human genetics. Finding criminals through

DNA of their relatives. *Science* **312,** 1315–1316 (2006).
122. Biabangard Zak, A., Golalipour, M. & Hadadchi, G. High Prevalence of Y Chromosome Partial Microdeletions in Overweight Men. *Avicenna J. Med. Biotechnol.* **7,** 97–100 (2015).
123. De Maddalena, C., Vodo, S., Petroni, A. & Aloisi, A. M. Impact of testosterone on body fat composition. *J. Cell. Physiol.* **227,** 3744–3748 (2012).
124. Roumaud, P. & Martin, L. J. Roles of leptin, adiponectin and resistin in the transcriptional regulation of steroidogenic genes contributing to decreased Leydig cells function in obesity. *Horm. Mol. Biol. Clin. Investig.* **24,** 25–45 (2015).
125. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011-2012. at <http://wwwn.cdc.gov/nchs/nhanes/search/nhanes11_12.aspx>
126. Ma, C., Boehnke, M. & Lee, S. Evaluating the Calibration and Power of Three Gene-Based Association Tests of Rare Variants for the X Chromosome. *Genet. Epidemiol.* **39,** 499–508 (2015).
127. Gao, F. *et al.* XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome. *J. Hered.* **106,** 666–671 (2015).
128. Salazar-flores, J., Ferna, L. G., Mart, G., Velarde-fe, S. & Mun, F. Admixture and population structure in Mexican-Mestizos based on paternal lineages. 1–7 (2012). doi:10.1038/jhg.2012.67
129. McHugh, C. Statistical Methods for the Analysis of Autosomal and X Chromosome Genetic Data in Samples with Unknown Structure. *PhD thesis, Univ. Washingt.* (2016).
130. Bloomer, L. D. S. *et al.* Male-Specific Region of the Y Chromosome and Cardiovascular Risk: Phylogenetic Analysis and Gene Expression Studies. *Arterioscler. Thromb. Vasc. Biol.* **33,** 1722–1727 (2013).
131. Charchar, F. J. *et al.* Association of the Human Y Chromosome with Cholesterol Levels in the General Population. *Arterioscler. Thromb. Vasc. Biol.* **24,** 308–312 (2004).
132. Charchar, F. J. *et al.* Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. *Lancet (London, England)* **379,** 915–22 (2012).
133. Charchar, F. J. *et al.* The Y Chromosome Effect on Blood Pressure in Two European Populations. *Hypertension* 353–356 (2002).
134. Ellis, J. a, Stebbing, M. & Harrap, S. B. Association of the human Y chromosome with high blood pressure in the general population. *Hypertension* **36,** 731–733 (2000).
135. García, E. C. *et al.* Association between genetic variation in the Y chromosome and hypertension in myocardial infarction patients. *Am. J. Med. Genet. A* **122A,** 234–237 (2003).
136. Hiura, Y. *et al.* Effects of the Y chromosome on cardiovascular risk factors in Japanese men. *Hypertens. Res.* **31,** 1687–94 (2008).
137. Kostrzewa, G., Broda, G., Konarzewska, M., Krajewki, P. & Płoski, R. Genetic polymorphism of human Y chromosome and risk factors for cardiovascular diseases: a study in WOBASZ cohort. *PLoS One* **8,** e68155 (2013).
138. Molina, E., Clarence, E. M., Ahmady, F., Chew, G. S. & Charchar, F. J. Coronary Artery Disease: Why we should consider the Y chromosome. *Hear. Lung Circ.* 1–11 (2016). doi:10.1016/j.hlc.2015.12.100
139. Shankar, R. R. *et al.* Studies of an Association in Boys of Blood Pressure and the Y

Chromosome. *Am. J. Hypertens.* **20,** 27–31 (2007).
140. Hampe, J., Schreiber, S. & Krawczak, M. Entropy-based SNP selection for genetic association studies. *Hum. Genet.* **114,** 36–43 (2003).
141. LaVange, L. M. *et al.* Sample Design and Cohort Selection in the Hispanic Community Health Study/Study of Latinos. *Ann Epidemiol.* **20,** 642–649 (2010).
142. Dong, Y. *et al.* Impact of Chromosomal Translocations on Male Infertility, Semen Quality, Testicular Volume and Reproductive Hormone Levels. *J. Int. Med. Res.* **40,** 2274–2283 (2012).
143. Hadjkacem-Loukil, L., Ghorbel, M., Bahloul, A., Ayadi, H. & Ammar-Keskes, L. Genetic association between AZF reqion polymorphism and Klinefelter syndrome. *Reprod. Biomed. Online* **19,** 547–551 (2009).

APPENDIX A

Table A.1. Number of alleles, genetic diversity (GD), Shannon entropy (H), and average normalized entropy difference (NED) for each Y-STR locus by region for the YHRD dataset.

YHRD																			
	Africa				Asia				Europe				America, Admixed				America, Native			
Locus	N alleles	GD	H	NED	N alleles	GD	H	NED	N alleles	GD	H	NED	N alleles	GD	H	NED	N alleles	GD	H	NED
DYS19	8	0.74	2.11	0.11	8	0.69	1.97	0.03	10	0.64	1.84	0.06	8	0.70	2.01	0.09	6	0.59	1.63	0.15
DYS385ab	59	0.94	4.75	0.16	116	0.97	5.72	0.09	99	0.87	4.10	0.11	61	0.92	4.67	0.15	50	0.96	4.99	0.25
DYS389I	6	0.53	1.41	0.06	7	0.66	1.70	0.04	8	0.56	1.42	0.07	5	0.58	1.47	0.04	5	0.61	1.51	0.10
DYS389II	10	0.76	2.33	0.11	11	0.77	2.32	0.05	10	0.72	2.10	0.07	11	0.74	2.21	0.07	8	0.68	1.94	0.11
DYS389II.I	10	0.68	1.91	0.10	8	0.66	1.86	0.04	10	0.59	1.62	0.05	9	0.64	1.82	0.07	6	0.64	1.73	0.10
DYS390	7	0.59	1.84	0.15	9	0.75	2.19	0.04	9	0.72	2.05	0.07	8	0.69	2.06	0.09	7	0.64	1.83	0.09
DYS391	7	0.41	1.06	0.07	8	0.40	1.10	0.02	7	0.54	1.28	0.05	8	0.56	1.43	0.08	5	0.42	1.10	0.10
DYS392	10	0.33	1.03	0.13	10	0.74	2.12	0.06	10	0.62	1.67	0.13	8	0.70	2.04	0.14	8	0.74	2.16	0.18
DYS393	8	0.62	1.65	0.07	7	0.65	1.75	0.04	7	0.43	1.22	0.04	6	0.46	1.29	0.05	7	0.61	1.60	0.12
DYS437	6	0.39	1.15	0.12	6	0.49	1.21	0.03	9	0.64	1.58	0.11	6	0.60	1.52	0.10	3	0.55	1.26	0.08
DYS438	6	0.48	1.35	0.13	9	0.61	1.79	0.07	10	0.69	1.91	0.16	8	0.70	1.94	0.15	5	0.66	1.66	0.15
DYS439	7	0.63	1.64	0.03	8	0.69	1.92	0.03	9	0.70	1.92	0.04	7	0.67	1.85	0.04	8	0.72	2.06	0.08
DYS448	9	0.66	1.86	0.13	10	0.76	2.30	0.05	11	0.63	1.71	0.09	7	0.70	1.98	0.10	8	0.74	2.06	0.16
DYS456	8	0.58	1.69	0.04	8	0.59	1.77	0.02	12	0.74	2.16	0.06	10	0.68	1.97	0.04	8	0.65	1.82	0.09
DYS458	9	0.75	2.22	0.04	11	0.81	2.60	0.02	12	0.77	2.36	0.04	10	0.76	2.32	0.03	10	0.78	2.42	0.13
DYS481	12	0.85	2.96	0.09	17	0.83	2.86	0.03	18	0.80	2.84	0.09	15	0.81	2.82	0.10	11	0.80	2.63	0.15
DYS533	6	0.54	1.43	0.08	8	0.63	1.71	0.03	8	0.59	1.67	0.07	9	0.63	1.76	0.08	6	0.63	1.73	0.12
DYS549	7	0.64	1.72	0.08	8	0.64	1.77	0.02	10	0.65	1.79	0.04	8	0.69	1.91	0.06	6	0.71	1.99	0.11
DYS570	12	0.79	2.55	0.05	14	0.83	2.77	0.03	14	0.79	2.56	0.05	12	0.79	2.59	0.06	9	0.80	2.51	0.10
DYS576	10	0.80	2.49	0.05	15	0.80	2.56	0.02	13	0.77	2.38	0.03	10	0.80	2.54	0.05	10	0.80	2.55	0.10
DYS635	10	0.68	2.14	0.10	13	0.78	2.47	0.04	12	0.67	2.00	0.09	9	0.72	2.09	0.12	7	0.62	1.75	0.12
DYS643	9	0.77	2.46	0.13	10	0.75	2.27	0.05	11	0.64	1.88	0.10	10	0.66	2.05	0.11	8	0.57	1.67	0.13
GATAH4	6	0.58	1.51	0.05	7	0.63	1.73	0.03	9	0.60	1.57	0.04	7	0.59	1.50	0.03	5	0.67	1.73	0.14

Table A.2. Number of alleles, genetic diversity (GD), Shannon entropy (H), and average normalized entropy difference (NED) for each Y-STR locus by region for the HGDP dataset.

HGDP																
	Africa				Asia				Europe				Middle East			
Locus	N alleles	GD	H	NED	N alleles	GD	H	NED	N alleles	GD	H	NED	N alleles	GD	H	NED
DYS19	6	0.77	2.21	0.20	5	0.70	1.90	0.08	5	0.73	2.03	0.23	4	0.57	1.37	0.22
DYS385ab	30	0.95	4.48	0.33	52	0.95	4.92	0.26	20	0.90	3.71	0.41	18	0.89	3.51	0.39
DYS389I	5	0.64	1.67	0.16	5	0.66	1.66	0.07	3	0.55	1.31	0.16	3	0.54	1.25	0.27
DYS389II	8	0.75	2.35	0.20	7	0.77	2.32	0.11	5	0.74	2.00	0.21	5	0.67	1.75	0.23
DYS389II.I	5	0.69	1.81	0.19	6	0.70	1.92	0.10	5	0.68	1.80	0.23	5	0.54	1.39	0.18
DYS390	7	0.68	2.06	0.23	7	0.75	2.14	0.10	4	0.63	1.54	0.17	5	0.68	1.78	0.31
DYS391	4	0.45	1.18	0.10	4	0.46	1.10	0.07	3	0.50	1.13	0.12	3	0.52	1.15	0.21
DYS392	4	0.17	0.57	0.13	7	0.73	2.30	0.15	6	0.64	1.75	0.28	3	0.21	0.58	0.16
DYS393	4	0.65	1.69	0.15	6	0.70	1.86	0.10	3	0.36	0.95	0.13	5	0.74	1.97	0.34
DYS437	5	0.21	0.67	0.13	3	0.55	1.33	0.07	3	0.58	1.33	0.13	3	0.58	1.37	0.29
DYS438	5	0.56	1.47	0.22	5	0.66	1.74	0.10	4	0.68	1.73	0.25	4	0.27	0.81	0.22
DYS439	5	0.67	1.76	0.11	6	0.73	2.02	0.09	5	0.71	1.90	0.17	4	0.60	1.48	0.31
DYS448	7	0.67	1.91	0.22	8	0.69	1.98	0.11	5	0.74	2.01	0.25	6	0.72	2.02	0.30
DYS456	7	0.65	1.82	0.15	6	0.53	1.52	0.08	4	0.69	1.70	0.15	5	0.39	1.18	0.22
DYS458	9	0.76	2.31	0.18	10	0.76	2.36	0.10	8	0.63	1.94	0.22	9	0.81	2.58	0.37
DYS570	8	0.84	2.78	0.21	11	0.81	2.57	0.11	8	0.79	2.41	0.22	6	0.67	1.83	0.28
DYS576	7	0.82	2.52	0.18	9	0.81	2.53	0.08	6	0.76	2.22	0.19	5	0.76	2.12	0.28
DYS635	8	0.61	1.97	0.19	9	0.80	2.51	0.12	7	0.76	2.23	0.25	7	0.76	2.24	0.34
GATAH4	5	0.59	1.50	0.13	5	0.63	1.59	0.07	5	0.66	1.70	0.17	3	0.46	1.01	0.14

Table A.3. Number of alleles, genetic diversity (GD), Shannon entropy (H), and average normalized entropy difference (NED) for each Y-STR locus by region for the XU dataset.

XU																			
	Africa				Asia				Europe				Middle East				America, Native			
Locus	N alleles	GD	H	NED	N alleles	GD	H	NED	N alleles	GD	H	NED	N alleles	GD	H	NED	N alleles	GD	H	NED
DYS19	8	0.76	2.26	0.13	6	0.65	1.86	0.10	5	0.63	1.75	0.10	7	0.50	1.53	0.18	4	0.28	0.84	0.15
DYS385ab	44	0.96	4.80	0.28	61	0.95	5.12	0.25	42	0.89	4.16	0.21	30	0.94	4.32	0.34	20	0.87	3.36	0.29
DYS389I	6	0.64	1.79	0.14	5	0.66	1.65	0.11	4	0.56	1.39	0.11	4	0.44	1.21	0.13	4	0.57	1.45	0.17
DYS389II	7	0.79	2.43	0.15	8	0.78	2.38	0.11	6	0.72	2.07	0.09	5	0.67	1.71	0.14	6	0.76	2.27	0.23
DYS390	6	0.64	1.91	0.18	8	0.67	1.87	0.09	5	0.73	2.00	0.11	5	0.68	1.86	0.19	6	0.59	1.65	0.22
DYS391	5	0.46	1.25	0.10	6	0.53	1.30	0.08	4	0.52	1.21	0.07	4	0.46	1.08	0.11	5	0.24	0.79	0.10
DYS392	8	0.33	1.09	0.13	7	0.73	2.19	0.15	7	0.71	2.01	0.17	4	0.50	1.38	0.19	6	0.61	1.75	0.19
DYS393	6	0.73	2.01	0.11	5	0.69	1.83	0.12	5	0.48	1.30	0.08	5	0.51	1.27	0.15	4	0.61	1.58	0.16
DYS437	4	0.31	0.92	0.11	4	0.44	1.08	0.09	4	0.64	1.62	0.13	4	0.57	1.43	0.17	4	0.12	0.43	0.10
DYS438	4	0.52	1.41	0.15	6	0.49	1.34	0.13	5	0.68	1.83	0.19	5	0.70	1.95	0.24	5	0.35	1.09	0.22
DYS439	6	0.65	1.73	0.08	7	0.66	1.92	0.12	5	0.71	1.91	0.09	4	0.67	1.74	0.14	4	0.62	1.49	0.13
DYS448	7	0.62	1.82	0.15	7	0.71	2.05	0.15	6	0.67	1.79	0.10	4	0.68	1.72	0.20	5	0.67	1.73	0.21
DYS456	6	0.59	1.58	0.08	7	0.69	1.96	0.10	7	0.76	2.21	0.11	7	0.65	1.88	0.15	5	0.46	1.29	0.16
DYS458	8	0.70	2.01	0.13	9	0.80	2.50	0.11	7	0.75	2.23	0.08	8	0.80	2.47	0.21	6	0.69	1.89	0.15
DYS635	8	0.71	2.22	0.16	7	0.74	2.16	0.10	8	0.71		0.13	6	0.80	2.40	0.23	4	0.16	0.52	0.13
GATAH4	5	0.56	1.42	0.08	4	0.53	1.35	0.07	4	0.55	1.35	0.06	5	0.64	1.70	0.16	4	0.50	1.17	0.12

APPENDIX B

Figure B.1. SWGDAM Recommendations to the FBI Director on the "Interim Plan for the Release of Information in the Event of a 'Partial Match' at NDIS"

Recommendation 1: The crime scene sample should be from a single source.

Rationale: When more than two alleles at a locus are used for searching, the number of partial matches from unrelated individuals increases very quickly. Although the Committee did not calculate the extent of this increase, it is clear that the added number of partial matches is expected to be high when using moderate stringency CODIS searching rules. Similarly, when only one allele (perhaps an obligate, nonvictim allele) is used in moderate stringency searches, there is also a large increase in the number of unrelated individuals in a database who would have partial matches to the crime scene evidence. To help limit the number of individuals in a database who would have partial matches to a crime scene DNA profile, the Committee recommends using only single-source forensic profiles in any evaluation of partial matches. For this same reason, mixtures should be deconvoluted into single-source profiles with as many heterozygous loci as possible before a database search is done.

Recommendation 2: Local DNA Index System (LDIS) and State DNA Index System (SDIS) searches should be performed before searching at the National DNA Index System (NDIS) level.

Rationale: The number of false-positive partial matches is strongly influenced by the size of the offender database searched. Large databases will have more false-positive partial matches. Another consideration is that LDIS and SDIS databases that are more geographically limited have a higher a priori chance of containing profiles actually related to the perpetrator. Both of these factors make a partial match more valuable (i.e., have a higher chance of resulting from the comparison of two related people). As an efficient search strategy, when no full or partial matches are found at the LDIS and SDIS levels, it would be logical to search geographically nearby states before searching the NDIS level. This strategy might be impractical, but it illustrates the general principle that the smaller the database searched, the lower the probability of finding false-positive partial matches and consequently the higher the chance of detecting a true relative.

Recommendation 3: All available CODIS core loci should be used for searching.

Rationale: Even though omitting a locus from a search will increase the chance of finding a true relative that partially matches, the number of false positives increases. The result is that relaxing the partial search by omitting a locus or two is counterproductive and buries the true match to a relative in an even larger number of partial matches to nonrelatives. A minimum of 10 of the CODIS core loci is required for searching forensic DNA profiles at the National DNA Index System level.

Recommendation 4: Whenever possible, partial DNA matches that result from searching databases should have additional loci typed.

Rationale: Y-chromosome short tandem repeat (Y-STR) and mitochondrial DNA (mtDNA) can eliminate unrelated individuals, and although offenders in DNA databases are currently not typed for Y-STR and mtDNA, it is feasible to do this on a small number of candidate partial matches. This, of course, is useful only when enough forensic sample is available. Y-STR and mtDNA are lineage markers and are not highly discriminating on their own, but they eliminate pairs of unrelated people. Because most perpetrators are male, and in this discussion of partial matches we are typically looking for father-son or full-sibling pairs, Y-STR information can winnow relatives from false-positive partial DNA matches. In some cases, additional autosomal genetic loci can add probative information to a partial match.

Recommendation 5: An Expected Match Ratio (EMR) and an Expected Kinship Ratio (EKR) should be calculated for a partial match.

Rationale: The size of the database searched has a dominating effect on the probative value of a partial match. The larger the database, the greater the number of false-positive partial matches. It is possible to calculate a ratio of the probability of observing a partial match in true relatives to the probability of observing that same partial match when a database of size N is searched. The EMR and EKR can provide somewhat objective measures of the partial match value as a lead to an actual relative. Details of these calculations are given in an Appendix2 to this report.

Recommendation 6: Four individual EMRs and EKRs should be calculated on the assumption that the database searched is made of (1) African Americans, (2) Caucasians, (3) S.E. (Southeastern) Hispanics, and (4) S.W. (Southwestern) Hispanics. The partial match is considered useful only if either the EMR or the EKR satisfies the following thresholds: at least one of the four database values is greater than or equal to 1.0 and all of the others are greater than or equal to 0.1.

Rationale: Because we do not know the actual ethnic composition of offender databases, we can do a pragmatic set of calculations under an assumption that the database is 100 percent of each of the four major ethnic groups in the FBI allele-frequency databases. Although the thresholds that we agreed upon (i.e., at least one EMR greater than or equal to 1.0 and all three of the other EMRs greater than or equal to 0.1) are somewhat arbitrary, they do set useful thresholds for the partial match to identify a true relative, if one exists in the database.

Recommendation 7: In order to implement these recommendations, it is important that CODIS Administrators have training in the evaluation of partial matches and in reporting the potential value of these matches.

Rationale: Partial matches come in many varieties, and the probative value of one, if any, can be determined only by further calculations and possibly by additional analytical tests. These calculations require the use of a spreadsheet or other software that currently is not in use but should be created and distributed. Report-wording suggestions need to be developed to stress the "limited" nature of the partial match and to state explicitly the possible family relationships.

Recommendation 8: All CODIS laboratories using these recommendations will report the profiles and associated EMRs and EKRs to the FBI, which will monitor the effectiveness of this approach.

Rationale: This is an emerging issue, and we have had little actual data to evaluate. With this or any other novel approach, an assessment seems necessary. It also would be useful if laboratories using alternate methods of identifying database partial matches report that method and data to the FBI. The FBI should evaluate these data provided by the LDIS and SDIS laboratories with the intent of modifying these recommendations and/or refining the thresholds as more data are collected.

Table B.1. US commercial DNA database company characteristics and law enforcement access policies. Accessed July 29th, 2016.

	DNA markers	Number of participants in database	Raw data provided	Price in US dollars	Statement about law enforcement access	Can data be removed from database?
AncestryDNA http://dna.ancestry.com/en/legal/us/privacyStatement	Autosomal SNPs, Y-SNPs, Y-STRs	2 million+	Yes	$99	"AncestryDNA will not disclose personal information to third parties except in very limited circumstances which are set out below… (c) as may be required by law, regulatory authorities,or legal process."	Yes, but copies of some data may have been saved by relatives. Backup copies may be retained on server.
23andMe https://www.23andme.com/about/tos https://www.23andme.com/about/privacy	Autosomal SNPs, Y-SNPs	1 million+, worldwide	Yes	$199	"Under certain circumstances your information may be subject to disclosure pursuant to judicial or other government subpoenas, warrants, or orders, or in coordination with regulatory authorities. 23andMe will preserve and	Yes, but not if you participate in 23andMe research.

					disclose any and all information to law enforcement agencies or others if required to do so by law…" Profiles of those participating in 23andMe Research are afforded additional protection under Certificate of Confidentiality. "Furthermore you agree not to use the Services to: … (6) use any information received through the Services to attempt to identify other customers, to contact other customers …, or for any forensic use"	

Family Tree DNA https://www.familytreedna.com/privacy-policy	Autosomal SNPs, Y-STRs	805,143, worldwide, including transfers from other test providers	Yes	$99 for autosomal panel. $169-359 for 37-111 Y-STR panel. $39 to transfer raw autosomal data from 23andMe or AncestryDNA. $19-58 to transfer raw Y-STR data from AncestryDNA	"We may disclose your Personal Information…as may be required by law, regulatory authorities, legal process or to protect the rights or property of Gene by Gene or other Users (including outside your country of residence)"	Yes
The Genographic Project www.nationalgeographic.com/community/privacy_complete www.shop.nationalgeographic.com/html/shopping/termsConditions2	Y-SNPs	779,696, worldwide	Unclear. Can be transferred to Family Tree DNA for free.	$179.95	"We will use and share your personal information as described in this Privacy Policy, including: … in response to legal process and when we believe that doing so is required by law, may be necessary to protect any person's property, rights, or safety, or to investigate a	"If you are under 18 and a registered user of the Services, you may request removal of content or information that you have publicly posted to the Services." "It is not always possible to completely remove or delete all of your information

					potential violation of law, will help to enforce any terms of use or other legal agreement, or in the event of a corporate transaction, such as a divestiture, merger, consolidation, bankruptcy or asset sale;"	due to technical constraints, contractual, financial or legal requirements. "
YSEQ http://www.yseq.net/privacy http://www.yseq.net/conditions	Y-STRs, Y-SNPs	Not listed.	Yes	$58-85 for a panel of 16-30 Y-STR markers, $75-99 for a haplogroup specific SNP panel, $9.95-19 for an individual STR, $17.50 for an individual SNP, $89+ for a custom SNP panel	"We may disclose your Personal Information…as may be required by law, regulatory authorities, legal process or to protect the rights or property of YSEQ or other Users (including outside your country of residence)"	Personal information, including test results are YSEQ assets

NATIONAL GENEALOGICAL SOCIETY

Guidelines for Sharing Information with Others

Recommended by the National Genealogical Society

Aware that sharing information or data with others is important and that it needs continuing support and encouragement, genealogists and family historians consistently

- respect the restrictions on sharing information that arise from the rights of another as an author, originator or compiler; as a living private person; or as a party to a mutual agreement;
- observe meticulously the legal rights of copyright owners, copying or distributing any part of their works only with their permission, or to the limited extent specifically allowed under the law's "fair use" exceptions;
- identify the sources for all ideas, information, and data from others, and the form in which they were received, recognizing that the unattributed use of another's intellectual work is plagiarism;
- inform people who provide information about their families how it may be used, observing any conditions they impose and respecting any reservations they may express regarding the use of particular items;
- require evidence of consent before assuming that living people are agreeable to further sharing or publication of information about themselves;
- convey personal identifying information about living people—such as age, home address, genetic information, occupation, or activities—only in ways that those concerned have expressly agreed to;
- recognize that legal rights of privacy may limit the extent to which information from publicly available sources may be further used, disseminated, or published;
- communicate no information to others that is known to be false, or without making reasonable efforts to determine its truth, particularly information that may be derogatory; and
- are sensitive to the hurt that information discovered or conclusions reached in the course of genealogical research may bring to other persons and consider that in deciding whether to share or publish such information and conclusions.

Figure B.2. Guideline for Sharing Information with Others Recommended by the National Genealogical Society.

VITA

Taryn attended the University of Minnesota earning a Bachelor of Science degree in Genetics, Cell Biology and Development in 2005. She continued on at the University of Minnesota and completed a Master of Public Health degree in Epidemiology, with an emphasis in Genetic Epidemiology in 2007. After finishing this degree, she moved to Montana where she worked as the State's Diabetes Epidemiologist for three years. Taryn was part of a team that very successfully translated the Diabetes Prevention Program clinical trial to community settings across Montana. Their program has since been deployed in YMCAs across the country. In 2010, Taryn entered the Public Health Genetics program at the University of Washington and worked under the supervision of Karen Edwards, studying Parkinson Disease, melanoma, metabolic syndrome and gestational diabetes. After Dr. Edwards left the university, Taryn began to work with Bruce Weir on forensic and public health genetic issues on the Y-chromosome. Taryn currently teaches a hands-off data analysis course for the fledgling Public Health program at Lake Washington IT and, if she passes her exam today, will start as a senior fellow trainee in the National Library of Medicine Biomedical and Health Informatics training program at the University of Washington in September.