OPEN ACCESS Freely available online

PLoS GENETICS

Familial Identification: Population Structure and Relationship Distinguishability

Rori V. Rohlfs[1]*, Stephanie Malia Fullerton[2], Bruce S. Weir[3]

1 Department of Integrative Biology, University of California Berkeley, Berkeley, California, United States of America, **2** Department of Bioethics and Humanities, University of Washington, Seattle, Washington, United States of America, **3** Department of Biostatistics, University of Washington, Seattle, Washington, United States of America

Abstract

With the expansion of offender/arrestee DNA profile databases, genetic forensic identification has become commonplace in the United States criminal justice system. Implementation of familial searching has been proposed to extend forensic identification to family members of individuals with profiles in offender/arrestee DNA databases. In familial searching, a partial genetic profile match between a database entrant and a crime scene sample is used to implicate genetic relatives of the database entrant as potential sources of the crime scene sample. In addition to concerns regarding civil liberties, familial searching poses unanswered statistical questions. In this study, we define confidence intervals on estimated likelihood ratios for familial identification. Using these confidence intervals, we consider familial searching in a structured population. We show that relatives and unrelated individuals from population samples with lower gene diversity over the loci considered are less distinguishable. We also consider cases where the most appropriate population sample for individuals considered is unknown. We find that as a less appropriate population sample, and thus allele frequency distribution, is assumed, relatives and unrelated individuals become more difficult to distinguish. In addition, we show that relationship distinguishability increases with the number of markers considered, but decreases for more distant genetic familial relationships. All of these results indicate that caution is warranted in the application of familial searching in structured populations, such as in the United States.

Citation: Rohlfs RV, Fullerton SM, Weir BS (2012) Familial Identification: Population Structure and Relationship Distinguishability. PLoS Genet 8(2): e1002469. doi:10.1371/journal.pgen.1002469

Editor: Greg Gibson, Georgia Institute of Technology, United States of America

Received July 4, 2011; **Accepted** November 22, 2011; **Published** February 9, 2012

Funding: This work was supported in part by National Institutes of Health grants R01 GM75091, T32 GM07735, and P50 HG003374. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rrohlfs@berkeley.edu

Introduction

Forensic identification via exact genetic profile matching has become common practice in the United States [1]. In exact genetic identification, genetic markers found in a crime scene sample are genotyped and exactly matched to a suspect or database entry, suggesting that the sample originates from the matched individual. In some cases, a database search yields no exact genetic profile matches, but does reveal partial matches where some, but not all, alleles match. A partial match could result from a genetic familial relationship between the individual who left the sample and the database entrant. If the database entrant has relatives, they might be investigated to determine if any of their genetic profiles exactly match the sample.

Familial searching is now used fairly frequently in the United Kingdom and was instrumental in the identification of suspects of violent crimes for 20 cases lacking other evidence as of 2008 [2]. Its use in the United States has been more limited due to concerns regarding civil liberty infringement, racial bias, and efficacy [3–6]. However, in July 2010, familial searching was used in a highly publicized California case to identify a suspect serial killer (the "Grim Sleeper") [7–10].

Despite the increasing use of familial searching in the United States, important questions about the method remain on both social and scientific grounds. In order to understand these concerns, we must appreciate that familial searching is most useful as a database mining method in cases with no suspects. In the United States, the Combined DNA Index System (CODIS) is the Federally administered system for National DNA Index System (NDIS), the national offender/arrestee database, which includes entries from State DNA Index Systems [11]. CODIS has standardized the use of genotypes at 13 particular short tandem repeats (STRs) (the CODIS loci) in forensic identification. The CODIS loci were chosen based on several criteria including reliable multiplexed PCR amplification, availability of commercial genotyping kits, clearly distinguishable alleles, linkage equilibrium, Hardy-Weinberg equilibrium, and high polymorphism in examined population samples [12–15]. An NDIS entry contains CODIS loci genotypes and a traceable index number, without other identifying information (e.g. location, race, or ethnicity) [16]. In September 2011, NDIS included over 10 million genotype profiles and continues to grow through new cases and expanded inclusion criteria [1].

These features of the forensic testing landscape matter because, unlike exact DNA identification, a typical database search for familial matches prospectively identifies candidate suspects who, while closely genetically related to database entrants, are not in themselves in the database, provoking complex privacy concerns [4,5,9,17,18]. Additionally, social groups which both share genetic relationships and are over-represented in the database would

Author Summary

The forensic identification of criminal suspects through DNA profiling is now common in the United States. Indirect identification by familial DNA profiling is increasingly proposed to extend the utility of DNA databases. In familial searching, a DNA profile from a crime scene partially matches a database profile entry, implicating close relatives of the partial match. While the basic principles behind familial searching methods are simple and elegant, statistical confidence that a partially matched profile belongs to a true genetic relative has not been fully explored. Here, we derive relative identification likelihood ratio statistics and consider how the ability of familial searching to distinguish relatives from unrelated individuals varies over population samples and is affected by inaccurately assumed population background. We observe lower relationship distinguishability for population samples with less identifying information in the genetic loci considered. Additionally, we show that relationship distinguishability decreases with discordance between true and assumed population samples. These results indicate that, if an inappropriate genetic population group is assumed, individuals from certain marginalized groups may be disproportionately more often subject to false familial identification. Our results suggest that care is warranted in the use and interpretation of familial searching forensic techniques.

experience a disproportionate increase in genetic surveillance if familial matching were routinely implemented, further exacerbating their over-representation in these databases [6,12,17–19].

The question of relative inference has been well-studied in other contexts with varying marker types, relationships, and numbers of individuals [20–28]. Here we focus on statistical and population genetic assumptions underpinning the familial searching methodology in the forensic context. Specifically, we consider the effects of both uncertainty in allele frequency estimation and population structure. First note that allele frequency estimates calculated within socially-defined population groups (e.g. African American, European American, Latino) are used to estimate the probability of an observed partial match, assuming a particular genetic relationship. Match probabilities for some individuals may not be accurately estimated using the available categorical socially-defined population group model and sample allele frequency data, particularly individuals with genetic ancestry outside of typically studied groups or individuals whose socially-defined population group does not inform their genetic ancestry. In exact identification, the probability of observing two individuals with identical specific 13-locus genotypes is astronomically low, with the exception of monozygotic twins. With these extremely low probabilities, differences or inaccuracies in allele frequency estimates are almost inconsequential, possibly changing the probability of an observed genotype a few orders of magnitude, but unlikely to alter the conclusion of the statistical analysis [29]. However, in familial identification, the probability of observing a coincidental partial match is much higher (e.g. for a parent-offspring relationship exactly one allele is shared by descent per locus). With these higher probabilities, population genetic differences in marker informativeness and errors in allele frequency estimation can perturb match probability estimations to such a degree as to affect the interpretation outcome.

In this study, we aim to examine some of these concerns by exploring how familial searching techniques behave on populations with varying allele frequency distributions and varying accuracy of allele frequency specification. We formulate and calculate confidence intervals for familial identification likelihood ratio (LR) estimates, and investigate how well siblings and unrelated individuals can be distinguished over different population samples with varying allele frequency distributions and under accurately and inaccurately assumed allele frequency distributions. We show that population samples vary in the amount of identifying information encoded in the CODIS loci and, therefore, in relationship distinguishability, even with correctly specified allele frequencies. Since completely accurate allele frequency specification is not guaranteed and the most appropriate population sample may not be known or available, we are also interested in the systematic effects of assuming allele frequencies which are appropriate for one population, but which are not appropriate for the individuals investigated. We show that relationship distinguishability decreases with the accuracy of allele frequency estimates, potentially resulting in high rates of coincidental familial identification for some groups. These results are especially pertinent in the multiple testing context of large database searching. In addition, we explore the relationships between relationship distinguishability, the number and type of markers used for identification, the relationship considered, and the true and assumed coancestry coefficient parameter value.

Results

Likelihood ratios for relationships with confidence intervals

To determine if a partial genotype match is better explained by a genetic familial relationship or stochasticity, we used the ratio of the likelihood of the observed partial match assuming the individuals share a given genetic familial relationship, to the likelihood of the observed partial match assuming the individuals are unrelated. With the data available, this LR is the most powerful statistic to separate relatives from unrelated individuals [30]. So even though the exact methodology used by forensic agencies for familial forensic identification is not readily publicly available, our use of the LR optimistically assumes the most powerful method using the CODIS loci. In the first part of this analysis, only sibling relationships are evaluated to reduce dimensionality. Other genetic familial relationships were explored and are reported below.

Unrelated individuals were simulated in a randomly mating population by independently drawing alleles from allele frequency distributions, similarly to Bieber *et al.* [31]. Siblings were then simulated by dropping alleles through a pedigree with unrelated parents. We simulated both unrelated individuals and siblings using allele frequency distributions from five socially-defined population samples, Vietnamese, African American, European American, Latino, and Navajo. Using both unrelated individuals and siblings, we calculated the sibling relationship $\widehat{LR}$ and 95% confidence interval of that estimate, assuming allele frequencies from each population sample. We simulated siblings and unrelated individuals under each of the five allele frequency distributions and calculated $\widehat{LR}$ and 95% confidence interval of that estimate assuming each of the five allele frequency distributions 10,000 times for each pair of population samples. As a result, we have $\widehat{LR}$ with confidence intervals for sibling relationships between unrelated individuals and siblings simulated from every population sample, assuming allele frequencies from every population sample. In most of the analyses presented here, we focus specifically on the lower 95% confidence limit of $\widehat{LR}$ (LCL) to account for sampling and biological variance in allele frequency estimation and to conservatively identify relationships. We refer to the population

sample used to simulate the individuals as the true population sample, as opposed to the assumed population sample used to calculate the LR for their relationship. Figure S1 shows the $\widehat{LR}$ 95% confidence intervals for 100 simulations of unrelated individuals, where individuals were simulated based on each population sample and confidence intervals were computed assuming the allele frequency distribution of each population sample.

Note that across all of these simulations specific parameter values were chosen and kept constant, specifically, sibling relationships, the assumed coancestry coefficient (probability of two alleles being identical by descent (IBD) between two individuals not recently related) used in calculations of $\theta_a = 0.01$, confidence interval length parameterized by significance level α as $1-\alpha = 0.95$, and the use of the 13 CODIS STRs. Regardless of the values of these parameters, the relative trends across true and assumed population samples will be maintained, although the scale may vary with parameter value choice.

Distinguishing relatives and unrelated individuals. To understand the degree to which $\widehat{LR}$ distinguishes relatives and unrelated individuals, we considered the distributions of LCLs for sibling relationships on simulated siblings and unrelated individuals. Figure 1 shows the density plots of the log LCL for both siblings and unrelated individuals using different true and assumed population samples. First we consider plots along the diagonal of Figure 1 showing density curves for unrelated individuals and siblings when the true allele frequency distributions are assumed. Plots with more overlap between the sibling and unrelated pair densities indicate less ability to distinguish relatives from unrelated individuals, a feature we term distinguishability, for the assumed and true population samples. Overlap can be observed visually in both density curve overlap and the bars above the density curves which show the simulated empirical central 95% of LCL over genotypes. To quantify the differences in distinguishability between population sample pairs, $\tilde{D}_{VH}$ measures the distinctness of the distributions of LCLs for individuals who are truly unrelated and truly siblings (see Methods). Table 1 shows $\tilde{D}_{VH}$ over true and assumed population samples. When the true population sample is assumed, $\tilde{D}_{VH}$ ranges from 5.87 for the Navajo sample to 7.38 for the African American sample (Table 1).

Gene diversity and distinguishability. Differences in distinguishability between population samples are rooted in differences in the shapes of allele frequency distributions. Since alleles and individuals are simulated independently, varying distinguishability over populations cannot be due to varying consanguinity and must be attributed to varying allele frequency distributions. In the examined population samples, the shape of allele frequency distributions can vary substantially. As a dramatic, but atypical, example, Figure S2 shows the different shapes of allele frequency distributions of D3S1358 for each population sample. Generally, the Navajo sample, and to a lesser extent the Vietnamese sample, allele frequency distributions have lower variance than that of the other samples, though not typically to the extreme extent seen at D3S1358.

Intuitively, it is clear that a monomorphic locus contains no identifying information, while a locus with a unique polymorphism for every individual contains complete identifying information. Along this spectrum, a locus with a low-variance allelic type distribution is less identifying than a locus with a high-variance allele frequency distribution.

This concept of varying identifying information can be quantified as observed gene diversity (or equivalently, average expected heterozygosity) [32]

$$\tilde{H}_l = 1 - \sum_u \tilde{p}_{l,u}^2$$

where $\tilde{H}_l$ is the observed gene diversity for locus l and $\tilde{p}_{l,u}$ is the observed allele frequency of allele u at locus l. Observed gene diversity can be combined across loci as the mean of observed gene diversity at each individual locus to get average observed gene diversity $\tilde{H}$. Using this method, we calculated the average observed gene diversity of the CODIS loci as $\tilde{H} = 0.77$, 0.79, 0.78, 0.79, and 0.70 for the Vietnamese, African American, European American, Latino, and Navajo samples, respectively (Text S1).

The calculated $\tilde{H}$ values show that the CODIS loci provide varying amounts of identifying information for different population samples. As our intuition suggests, population samples with lower-variance allele frequency distributions, particularly the Navajo sample, have lower average gene diversity. Even when assuming the correct allele frequency distribution, there is significant correlation between relationship distinguishability ($\tilde{D}_{VH}$) and average gene diversity ($\tilde{H}$) across population samples, as seen in Figure 2 ($r^2 = .95$, $p = 2.7e-3$).

Information theory can provide a more direct measure of identifying information through entropy, which we calculate to quantify the number of bits required to encode an equivalent amount of information as a CODIS haplotype for each population group. We find that relationship distinguishability is even more correlated with entropy than observed gene diversity, which follows since entropy quantifies information content which directly affects distinguishability (see Text S1 and Figure S3).

Allele frequency misspecification and distinguishability. By calculating LCL under different assumed and true population sample allele frequencies, the relationship between allele frequency misspecification and relationship distinguishability can be examined. By looking at plots and values off the diagonal, Figure 1 and Table 1, it is clear that distinguishability is particularly low when the true sample is Navajo and the assumed sample is different. This indicates that unrelated Navajo individuals more often appear sibling-like when non-Navajo allele frequencies are assumed. The same is true for the Vietnamese sample, though the trend is less pronounced.

In this study, we chose not to define a single decision threshold for determining positive relative identifications since such a threshold depends on a number of factors beyond the scope of this study (e.g., the social, economic, and political cost of false positives and negatives). For a range of decision thresholds, Figure 3 shows the false positive rate and the power. To intuitively calibrate $\tilde{D}_{VH}$ by commonly-used statistics, Figure 3 plots $\tilde{D}_{VH}$ along with each set of false positive rate and power curves. False positive rate and power vary by population, with the true Navajo and assumed non-Navajo samples having particularly high false positive rates for decision thresholds shown. If a high decision threshold is chosen so that the false positive rate for true Navajo cases is comparably low as it is for other population samples, the power to identify siblings drops to levels that may render the investigation ineffective. In Figure 3 this can be visualized by choosing a point on the x-axis where the Navajo sample false positive rate is low (perhaps a decision threshold of 6) and looking up to the power to detect relationships at that threshold. A similar, but less pronounced, pattern appears with the Vietnamese data.

Low nominal false positive rates. It is notable that even when the correct allele frequencies are used, the false positive rate is lower than the $\widehat{LR}$ confidence interval significance level α. However, this is not surprising since the parameter α determines

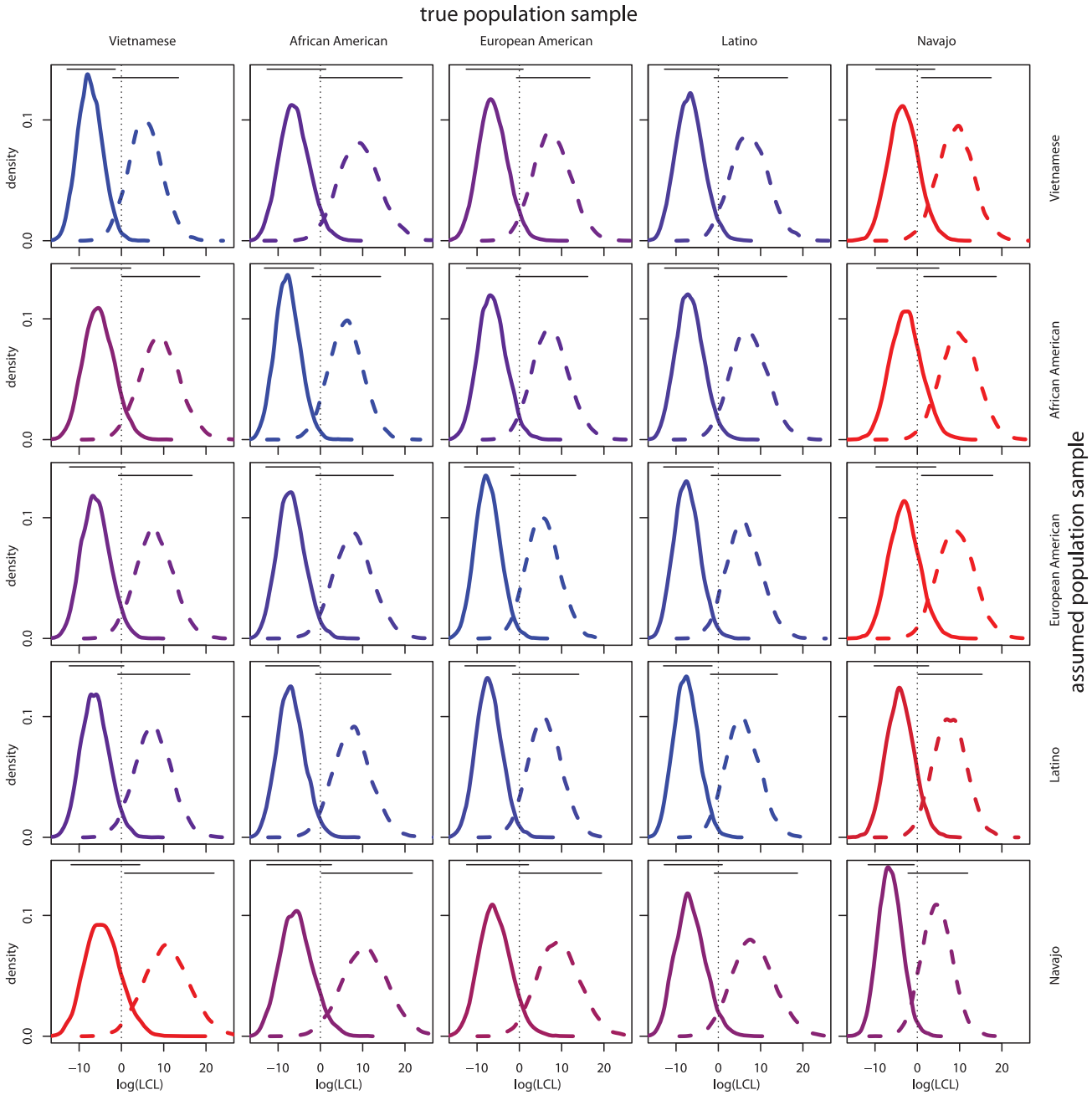

Figure 1. LCL distributions for siblings and unrelated individuals by population samples. Each individual plot shows the distribution of log LCLs for unrelated individuals (solid) and siblings (dashed). The dotted vertical lines indicate $\widehat{LR}=1$. The horizontal lines show the central 95% of observed values over genotypes. The true and assumed population samples are listed on the column and row headings, respectively. Plot coloring indicates distinguishability where red represents low $\tilde{D}_{VH}$ and blue represents high $\tilde{D}_{VH}$.
doi:10.1371/journal.pgen.1002469.g001

the width of the $\widehat{LR}$ confidence interval, not the false positive rate. The confidence interval describes uncertainty in the LR estimation due to variance in the allele frequencies. In contrast, the false positive rate is a function of the low probability that two unrelated individuals share alleles in a pattern that resembles sibling relationships, which is often lower than the unrelated $\alpha=.05$ parameter value used here. See Text S1 for more details.

$\tilde{D}_{VH}$ and $\hat{\theta}$

We observed lower distinguishability when the true and assumed allele frequency distributions differ more. The degree of difference between population sample allele frequency distributions at the CODIS alleles is quantified for every population pair using $\hat{\theta}$ (Table 2). To account for multiple alleles at multiple loci and varying sample sizes, we estimate θ with the method of Weir and Cockerham [33]. Note that $\hat{\theta}$s reported here were calculated using the only CODIS loci, as is appropriate for an analysis of forensic methods. For a thorough investigation of the population genetics of these samples, more loci would be required, producing different results than those shown here, as reported in other studies [34,35].

To explore the relationship between distinguishability and the genetic distance between true and assumed population samples, in

Table 1. $\tilde{D}_{VH}$ between population samples.

	True population sample				
	Vietnamese	**African American**	**European American**	**Latino**	**Navajo**
Vietnamese	6.94	6.35	6.19	6.51	5.00
African American	5.91	7.38	6.41	6.59	4.68
European American	6.23	6.64	6.98	6.90	4.71
Latino	6.31	6.80	6.83	7.07	5.21
Navajo	5.01	5.94	5.74	5.94	5.87

$\tilde{D}_{VH}$ between each true (columns) and assumed (rows) population sample pair calculated using the CODIS loci.
doi:10.1371/journal.pgen.1002469.t001

Figure 4, $\tilde{D}_{VH}$ is plotted against $\hat{\theta}$ for each pair of true and assumed population samples. $\tilde{D}_{VH}$ and $\hat{\theta}$ are significantly correlated ($r^2 = 0.74$, $p = 2.6e-8$), supporting the hypothesis that incorrectly assuming allele frequencies leads to low distinguishability and high false positive rates. In particular, we observe low distinguishability when Navajo, or to a lesser extent Vietnamese, is the true population sample, correlating with higher $\hat{\theta}$ with the other assumed samples.

Intuitively, when allele frequencies are misspecified, the most likely error is assuming that common alleles are more rare simply because truly common alleles are more likely to be observed than truly rare alleles. In the same way, rare alleles are assumed to be common, but by definition, rare alleles are less likely to be observed shared between individuals, so overall the misspecification of common alleles as rare dominates. When misspecifying common alleles as rare, observing the same common alleles in multiple individuals seems surprising, so a genetic relationship model is favored over a model of no relationship. That is, the probability of a partial match assuming a relationship is inflated and the probability of a partial match assuming no relationship is deflated. In this way, allele frequency misspecification results in an increase in false positive relative identifications.

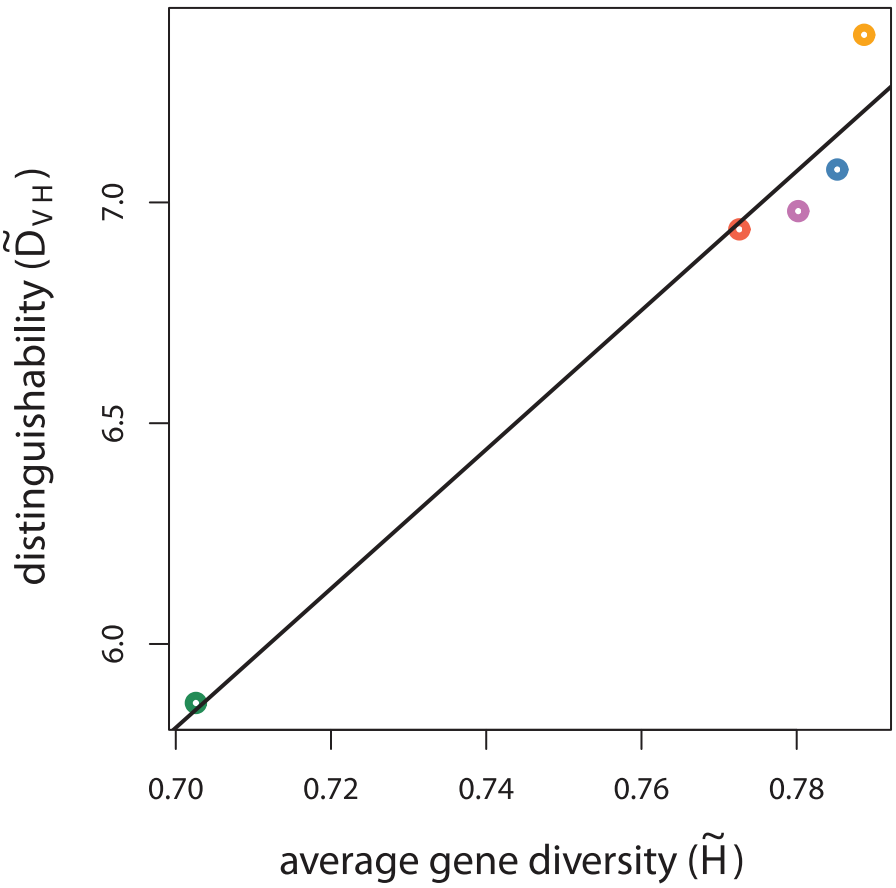

Figure 2. $\tilde{D}_{VH}$ **versus** $\tilde{H}$**.** The empirical distinguishability ($\tilde{D}_{VH}$) for siblings and unrelated individuals is plotted against average gene diversity ($\tilde{H}$) for each population sample. Points are colored according to the true population sample where red signifies Vietnamese, orange African American, purple European American, blue Latino, and green Navajo.
doi:10.1371/journal.pgen.1002469.g002

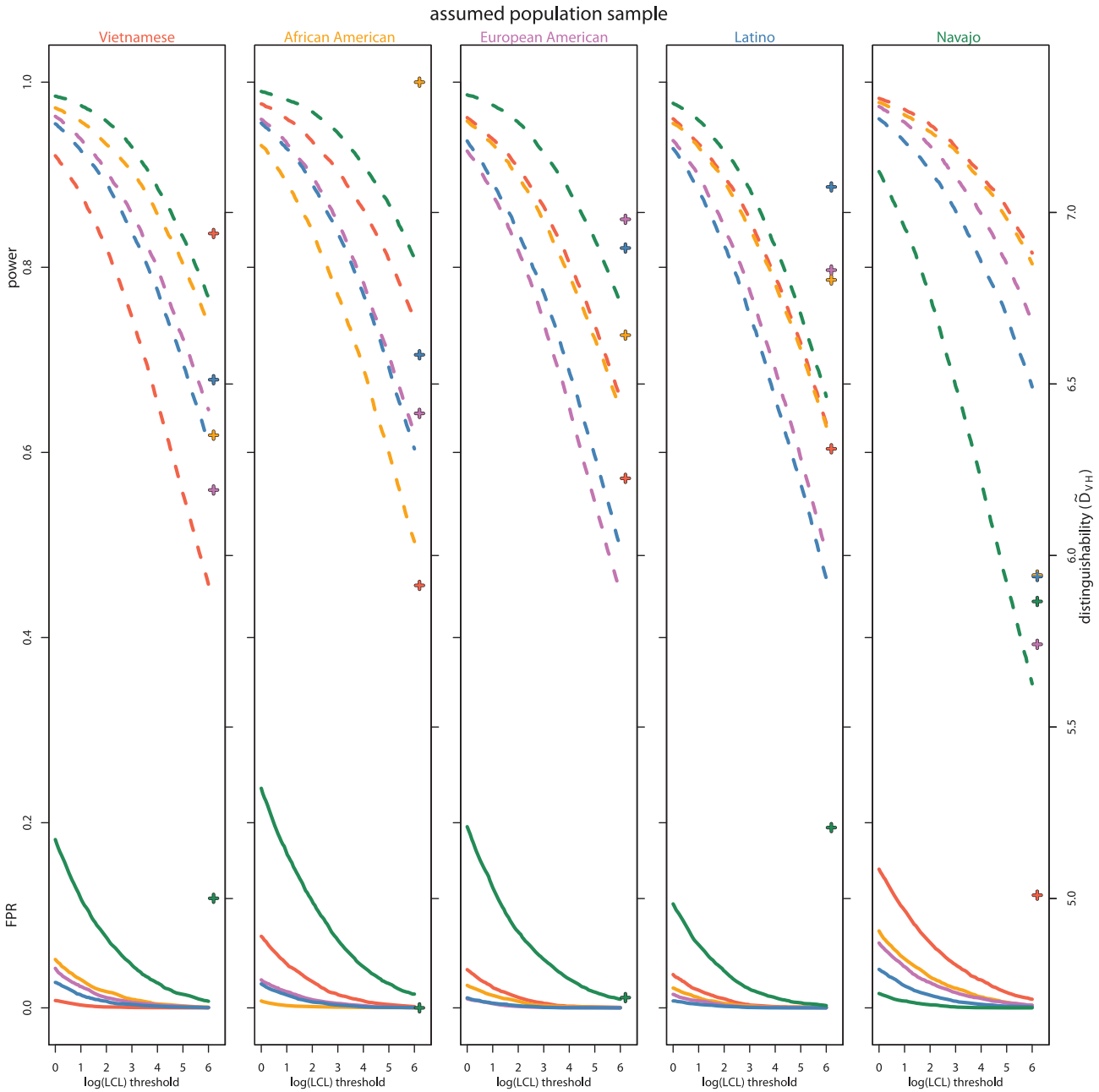

Figure 3. Power and false positive rate over thresholds and by population samples. The empirical power (dashed) and false positive rate (solid) are shown for a range of sibling versus unrelated log LCL decision thresholds. In each plot, the indicated population sample is assumed in the calculations. Within each plot, the colored curves indicate the true population sample allele frequencies used to simulate individuals. Red signifies Vietnamese, orange African American, purple European American, blue Latino, and green Navajo. Similarly color-coded crosses indicate $\tilde{D}_{VH}$ for each population sample pair.
doi:10.1371/journal.pgen.1002469.g003

Although the relationship between distinguishability and allele frequency misspecification has not yet been deeply considered in the context of genetic familial identification (but see [36]), it has been discussed in the forensic literature for exact matching and it is well-known in the linkage analysis community. For exact forensic identification using the 13 CODIS loci, discrepancies between assumed and true allele frequencies affect the computed match probabilities, but seldom change the ultimate outcome [37–40]. In linkage analysis, when inaccurate population allele frequencies are used to calculate genotype probabilities, false linkage signals between genotype and phenotype are common [41,42].

Additional populations. We have shown clear differences in average observed gene diversity of the CODIS loci and resulting differences in sibling and unrelated individual distinguishability in the five population samples considered. To ensure that these findings extend beyond the samples examined, we considered a larger dataset with a total of 32 population samples [43]. As in the five-population sample dataset, average observed gene diversity at the CODIS loci varies between samples, with particularly low

Table 2. $\hat{\theta}$ between population samples.

	African American	European American	Latino	Navajo
Vietnamese	0.038	0.026	0.021	0.068
African American		0.017	0.015	0.067
European American			0.000	0.059
Latino				0.041

$\hat{\theta}$ between each population sample as calculated using the CODIS loci. Estimates less than 0.0 are reported as 0.0.
doi:10.1371/journal.pgen.1002469.t002

values for Native American samples (Text S1). We performed a comparable analysis of average observed gene diversity versus distinguishability using ten population samples and found again that $\tilde{D}_{VH}$ is correlated with $\hat{\theta}$ over true and assumed population samples ($r^2 = .74$, $p = 2.2e-16$, Figure S4).

Distinguishability over parameters

In the analysis presented thus far, we showed how distinguishability varies over true and assumed population samples with varying allele frequency distributions. To maintain manageable dimensionality, some key parameters likely to vary in forensic analyses were kept constant. Here we explore the relationships between these parameters, particularly different genetic relationships, varying marker data, and varying the true and assumed coancestry coefficients (θ and θ_a). To focus on the relationships

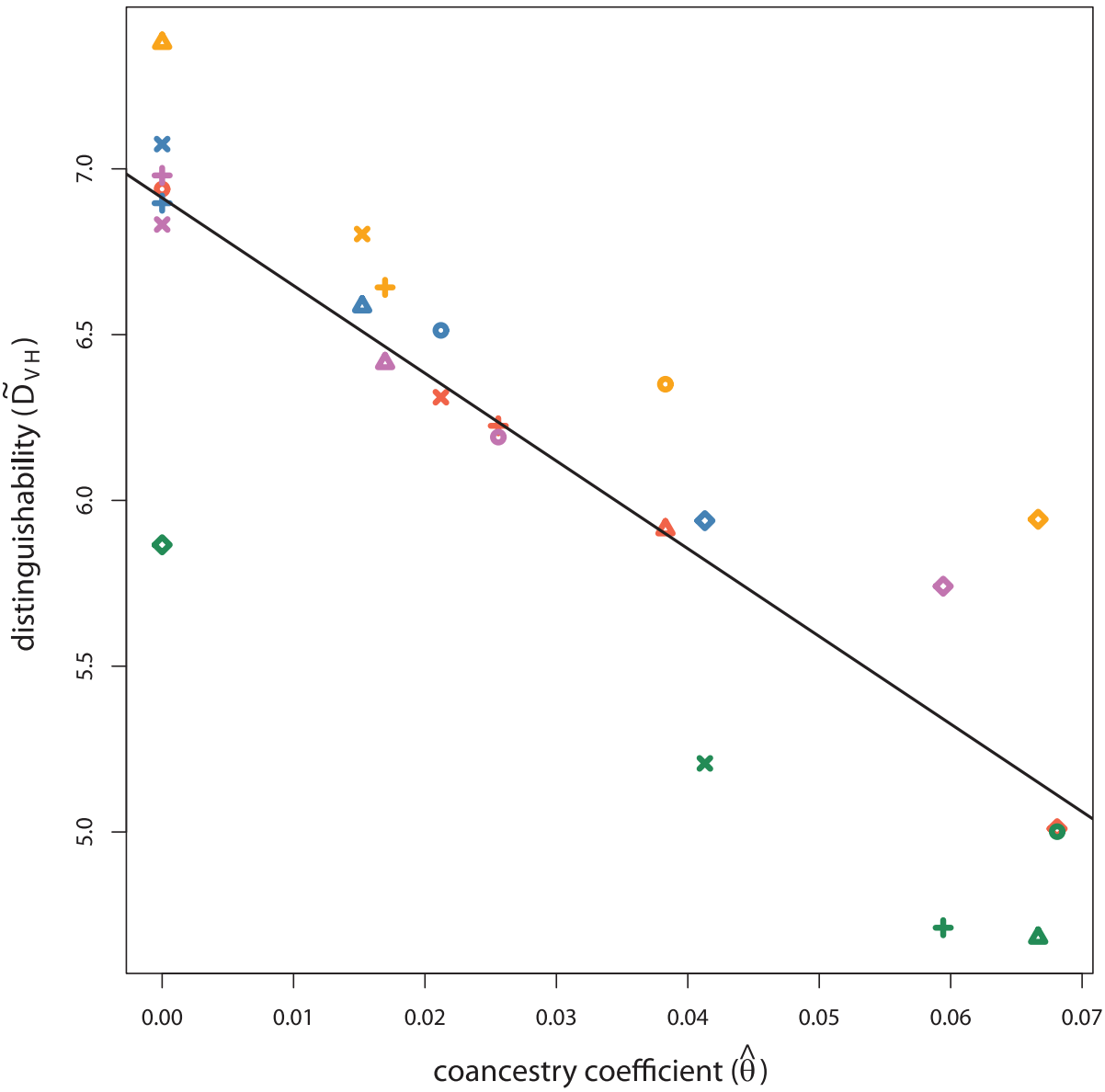

Figure 4. $\tilde{D}_{VH}$ versus $\hat{\theta}$. The empirical measure of distinguishability ($\tilde{D}_{VH}$) for siblings and unrelated individuals is plotted against $\hat{\theta}$ for each pair of true and assumed population samples. Points are colored according to the true population sample and take a shape according to the assumed population group where red and circles signify Vietnamese, orange and triangles African American, purple and plus marks European American, blue and multiplication marks Latino, and green and diamonds Navajo. $\hat{\theta}$ estimates less than 0.0 are reported as 0.0.
doi:10.1371/journal.pgen.1002469.g004

between these parameters, in these analyses the correct known allele frequencies were used.

Pairs of individuals were simulated taking into account the true coancestry coefficient, θ, using the genotype probabilities described in the Text S1, for the following genetic relationships: parent-offspring, sibling, half-sibling, first cousin, second cousin, and unrelated. Note that in contrast with the analyses presented above, here θ is used to model background relatedness. LRs were computed comparing the probabilities of the simulated data assuming the true relationship and assuming the individuals are unrelated. This analysis was repeated over varying numbers and types markers and a variety of assumed θ_a values.

Varying number and type of markers. We simulated two types of markers with equi-frequent alleles: 10-allele STRs and 2-allele SNPs. We varied the number of simulated markers over 10, 20, 30, 40, 50, and 60 STRs and 10, 50, 100, 150, 200, and 250 SNPs in independent simulations. Distinguishability between the LCL distributions of true relatives and unrelated individuals were calculated for each of these simulations (Figure S5). Distinguishability varies widely over relationships, with sibling $\tilde{D}_{VH}$ being two or three orders of magnitude higher than second cousin $\tilde{D}_{VH}$. We also see distinguishability increase with the number of markers.

For unrelated individuals, $\widehat{LR}$ for a parent-offspring relationship is often exactly 0 since unrelated individuals are unlikely to share at least one allele at each locus. As a result, the distribution of $\log(\widehat{LR})$ is not definable and distinguishability cannot be computed, so parent-offspring relationships are excluded from these results.

Varying θ and θ_a. The genetic similarlity of relatives can be quantified with the kinship coefficient, which is the probability that a pair of alleles from relatives are IBD. The kinship coefficient for parent-offspring, sibling, half-sibling, first cousin, and second cousin relationships are 0.25,0.25,0.125,0.0625, and 0.015625, respectively. Intuitively, as the kinship coefficient of the tested relationship approaches the population background relatedness (θ), it will become increasingly difficult to discern relatives from unrelated individuals.

To explore the relationship between true coancestry coefficient θ, assumed coancestry coefficient θ_a used in probability calculations, genetic similarity of relatives, and $\tilde{D}_{VH}$, we consider 15 STRs and 100 SNPs and simulated individuals with true population $\theta = 0.00,0.01,0.03,0.05,0.07$, and 0.09. We then calculated LRs using $\theta_a = 0.00,0.01,0.03,0.05,0.07$, and 0.09. For each type of marker, distinguishability decreased as θ increased and the slope of that decrease flattens as θ_a increased (Figure S6). Again, distinguishability varied over relationships where $\tilde{D}_{VH}$ for siblings was about three orders of magnitude greater than $\tilde{D}_{VH}$ for second cousins. This consistent with findings by Anderson and Weir that IBD sharing estimation accuracy increased with the number of markers considered and decreased as θ increased [44].

Discussion

The analysis presented here confirms and quantifies the intuition from population genetics that for particular loci, groups with comparatively low-variance allele frequency distributions have less identifying information encoded in genotypes. Decreased identifying information results in lower relationship distinguishability, even when the correct allele frequency estimates are used (Figure 2, Figure S2). This is abundantly apparent for the Native American samples considered in this analysis.

With a basic understanding of population genetics, it is clear that socially defined groups, like Navajo, Latino, or European American, have very different underlying population structures reflecting distinct demographic history, degrees of genetic diversity, and admixture. It is hardly surprising that a group which has undergone multiple population size reductions, like the Navajo, has a lower-variance allele frequency distribution than a group with a history of genetic diversity and social inclusion, like African Americans. This is particularly evident at the CODIS loci, which were chosen in part because of their broad allele frequency distributions in a few studied populations, without considering all relevant populations [13–15].

These population differences in allele frequency distributions are key when considering a potential source of error: inappropriately assumed allele frequency distributions. When the allele frequency distributions for an inaccurately specified population group are assumed, the probabilities of the observed data under a sibling relationship and under no close genetic relationship become less distinct, so relationship distinguishability decreases. We found that distinguishability decreases with increased distance between assumed and true allele frequency distributions, as measured through $\hat{\theta}$. Specifically, both Navajo and Vietnamese samples are more genetically distant to the other three samples considered and show decreased distinguishability when allele frequencies of one of those three samples are assumed.

The results of this analysis indicate that when a decision threshold is chosen so that the power to identify siblings is reasonably high, population samples with allele frequencies which differ from those assumed would experience disproportionately higher rates of false positive familial identification (Figure 3). This could be exacerbated by unknown population-based differences in genotyping which would distort allele frequencies, for example, population-specific mutations in PCR primer binding sites [45–51]. More extensive genotyping of genetically diverse populations may make available more appropriate allele frequency distributions. However, it is not clear how or if the most appropriate allele frequency distribution for a pair of samples can be determined. Population-based differential distinguishability will persist, regardless of additional population-specific allele frequency distributions or uniformly applied corrections. One possible correction would be increasing the value of the parameter θ, however, in Figure S6 we see that even when the true allele frequencies are assumed, increasing θ decreases distinguishability. If more genetic data were used, particularly markers on the Y chromosome or mitochondrial DNA, as are in some states but not Federally, profile informativeness could be increased to the point where allele frequency approximations made little difference in the ultimate outcome (Figure S5) [10,52]. However, additional Y chromosome and mitochondrial markers will only inform matrilinial or patrilinial relationships and any additional markers will be subject to similar population-specific variation, and will be limited by practical genotyping constraints and the need to avoid medically-associated regions. Additionally, it is not clear if more distant relationships (cousins, second cousins, etc) would be confidently identified, even with more independent genetic loci (Figure S5) [53,54]. As it is, the core 13 CODIS loci, or the minimum 10 loci recommended by the Scientific Working Group on DNA Analysis Methods Ad Hoc Committee on Partial Matches (SWGDAM), seem inadequate to implement sibling matching with low false positive rate and high power in structured populations [52,55]. More complex situations, like mixed or low-template DNA samples, require further study and may not be feasible with the 13 CODIS loci [55,56].

Motivated by the question of forensic familial searching, in this analysis we focus on distinguishing relatives with a specified relationship and unrelated individuals. In other contexts, it may be more appropriate to distinguish different kinds of relatives (e.g.

siblings and parent-offspring) or relatives with an unspecified relationship and unrelated individuals. In the former case, the ratio of LRs for the relationships of interest versus unrelated individuals reduces to the LR comparing the two specified relationships. In the later case, models allowing IBD sharing probabilities to vary can be formulated and incorporated into the LR. For example, when comparing a null model with set IBD sharing probabilities for unrelated individuals and an alternative where the likelihood of data is maximized over any IBD sharing probabilities, a LR test can be formulated which follows a χ^2_2 distribution under the null hypothesis.

This analysis considers familial identification in a forensic context, but is applicable to tests for relatedness applied in the various contexts especially when considering unlinked genetic markers as in paternity investigation, ecological surveys, and conservation biology. When more extensive genotype or sequence data are available, it is appropriate to use more sophisticated tests for relatedness considering linkage or shared haplotype length [28,57,58].

The population genetic model used in forensic identification is remarkably coarse. In direct identification, the CODIS loci provide ample data to determine identity and non-identity, even with the coarse population genetic model of a small number of discrete homogenous genetic groups corresponding to social racial groups. We have shown that under this model, new concerns arise with familial searching. However, the model itself requires some scrutiny. It is clear that human genetic population structure is complex and humans are not easily split into a small number of discrete homogenous genetic groups [59–62]. Even with carefully chosen and defined population samples, it is practically impossible to account for human genetic variation and the discrete population group model fails to account for individuals with mixed ancestry. Additionally, individuals are typically assigned to genetic population groups based on social race. While there is correlation between genetic ancestry and social race, one does not determine the other [63]. As a result, in the discrete population group model, some individuals may not be grouped with the most similar genetic group.

Forensic familial searching will most likely be implemented in the context of a large offender/arrestee database, introducing questions of multiple testing over both database entrants, and the number of genetic familial relationships considered. Because forensic methodology practice varies over jurisdictions, it is not clear how these multiple testing issues have been, or will be, addressed. However, it is reasonable to assume that familial searching will result in a list of partial database matches with $\widehat{LR}$ for genetic familial relationships. The parameter values used in the $\widehat{LR}$ calculations must be conservative to keep the number of high $\widehat{LR}$ partial matches manageably short, but the parameters also must allow enough leniency so that a true match will appear in the list considered. Ideally, parameter values used in practice should be tuned using simulations based on real genotype data representing realistic cryptic relatedness and population structure appropriate to the database and relevant population. When tuning parameters, as power increases, false positive rate will as well. Both of these values must be considered in deciding on appropriate parameter values. However across parameter values, some groups may have higher rates of false identification, as we have shown here, raising questions about the practicality of familial searching. Without access to accurate database or population information, or to a clear decision procedure practice, we refrain from making specific recommendations about parameter choice or methodology in this analysis.

Individual and population genotype information is necessary to determine the extent to which inaccurately assumed allele frequencies cause high false positive rate in familial matching in practice. For instance, in this study, we considered unrelated individuals, conforming to exactly one of five allele frequency distributions, in completely randomly mating populations. However the use of familial searching rests on the premise that relative groups are in the database and population structure is undeniably present in most databases [64]. Access to suitably secure and encrypted database information would enable analyses with an accurate portrayal of relatedness and population substructure. As recommended by Krane *et al.*, increased transparency in database makeup, search procedure, and database access are required for rigorous analyses of forensic methodology [65].

If implemented with the core CODIS loci, familial searching may result in low distinguishability and potentially high false positive rates among certain groups, especially if only African American, European American, Southeastern Latino, and Southwestern Latino allele frequency distributions are in assumed LR calculations, as recommended by SWGDAM [55]. Because some of these groups (Native Americans and some immigrant groups) are correlated with social groups already over-represented in the criminal justice system, group members would be more likely to have a relative in the database, and that relative would be more likely to have a coincidental partial match with a crime scene sample [3–6,9,17,18,66–68]. Cumulatively, members of these groups are more likely to be investigated as a familial match due to over-represention in the database, and an unusually high false positive familial identification rate.

Methods

Data

Our analysis makes use of allele frequency data for the 13 CODIS loci over different population samples socially defined by race. Note that alternate schemes to group individuals will also produce genetic differences between groups [56,63,69]. Here, we consider genetic differences between socially-determined groups which are relevant to the practice of genetic familial forensic identification. To do so, we used the allele frequencies reported by Budowle and Moretti [29] for samples from 'Vietnamese,' 'African American,' 'Caucasian,' 'Hispanic,' and 'Navajo' populations. In this manuscript, these same samples are refered to with the following labels: Vietnamese, African American, European American, Latino, and Navajo. As short hand, we refer samples derived from individuals from each sample as the sample name, for example 'the Latino sample.' The number of individuals genotyped to estimate allele frequencies for each sample varied, with $n = 213,200,150,210$, and 182 individuals sampled for Vietnamese, African American, European American, Latino, and Navajo samples, respectively.

The consent and population grouping procedures used in obtaining these data are not clear. In the time since these data were collected, dominant cultural ethics regarding informed consent process have changed considerably, motivated largely by several cases of severe misuse of samples provided by Indigenous communities [70–73]. As a result, today it is becoming less acceptable to gather data in the same way [74–78]. We use the data because of its public availability, however we look forward to working with data collected using transparent informed consent methodology.

Likelihood ratio for relationship

LRs are used to compare the probability of observed genotypes for two individuals under two different hypotheses: the individuals are unrelated (H_u) and the individuals share a specified genetic

familial relationship (H_r) [79]. The LR is defined as [79]

$$LR = \frac{P(G|H_r)}{P(G|H_u)}$$

where G is the observed pair of genotypes. When $LR<1$, the observed data are more likely for unrelated individuals and when $LR>1$, the observed data are more likely for individuals with the specified genetic relationship.

By assuming independence between all CODIS loci, LR can be broken down as

$$LR = \prod_l \frac{P(G_l|H_r)}{P(G_l|H_u)}$$

where G_l is the observed genotype for each individual at locus l.

Relationships between individuals can be described using the identical by descent (IBD) sharing probabilities k_0, k_1, and k_2, which are the probabilities that individuals with the specified relationship share 0, 1, and 2 alleles IBD, respectively [79]. For example, for a parent/offspring relationship $k_0=0$, $k_1=1$, and $k_2=0$ and for a sibling relationship $k_0=0.25$, $k_1=0.5$, and $k_2=0.25$.

Using these IBD sharing probabilities, the LR becomes

$$LR = \prod_l \frac{P(G_l|k_0,k_1,k_2)}{P(G_l|k_0=1,k_1=0,k_2=0)}$$

where the IBD sharing probabilities in the numerator are specified by the specific genetic relationship considered. The probability of the observed genotype combinations given IBD sharing probabilities depends on the specific combination of alleles observed. The probabilities of all observed genotypes, given IBD sharing probabilities, are defined in Text S1. These probabilities include a correction for expected background relatedness using the coancestry coefficient θ. In the first part of this study, we use the value of $\theta=0.01$ based on standard methodology in population genetics and as recommended by SWGDAM [55,80].

Likelihood ratio confidence intervals

The LR described above provides information about whether the observed data are more likely for unrelated or related individuals. However, the true population allele frequencies (p_i) are unknown, so LR needs to be estimated with the observed allele frequencies. Available sample allele frequencies are subject to sampling variation and variation due to demographic history [81]. Observed allele frequencies follow directly from observed genotype frequencies. Using $\hat{\boldsymbol{p}}_i$, the probability of the data is calculated under different IBD sharing schemes, so the estimate of the likelihood ratio ($\widehat{LR}$) can be computed. By considering the distribution of $\hat{\boldsymbol{p}}_i$, we can find the distribution of $\widehat{LR}$ and calculate confidence intervals on reported $\widehat{LR}$ values.

Sampling variation is inherent in allele frequency estimation since a random sample must be chosen for the estimate. By their nature, different random samples vary in their representation of specific alleles, resulting in different allele frequency estimates. Additionally, random genetic sampling exists in the historical differentiation of populations, resulting in population groups with distinct allele frequencies. Since all present-day human population groups descend from a common ancestral population, the alleles present in each present-day population group reflect a sample of the alleles from the common ancestral population.

Under evolutionary equilibrium and a simple model of demographic history, the relationship between population group allele frequencies ($\hat{\boldsymbol{p}}_i$) can be modeled using a Dirichlet distribution informed by the coancestry coefficient (θ), accounting for genetic and sampling variation in estimated allele frequencies [81,82]. With this model, we define the $\widehat{LR}$ confidence interval in order to express uncertainty conferred by allele frequency estimate.

Using the same approach as Beecham and Weir [81], we note that the total $\log(\widehat{LR})$ is the sum of the $\log(\widehat{LR}_l)$ for each locus l. The central limit theorem indicates that, for even as few as 13 independent loci, this sum will be approximately normally distributed [81]. Thus, the confidence interval for $\log(\widehat{LR})$ is [81]

$$CI = \log(\widehat{LR}) \pm z_{\alpha/2}\sqrt{\mathrm{var}\,(\log(\widehat{LR}))}$$

where $\mathrm{var}\,(\log(\widehat{LR}))$ is the variance of $\log(\widehat{LR})$ and $z_{\alpha/2}$ is the standard normal value for the given α, in this study $\alpha=.05$ and so $z_{\alpha/2}=1.96$. While the typical arbitrary value of $\alpha=.05$ is used in this study, the trends explored will be maintained with different values of α. Also note that a one-sided confidence interval can be derrived similarly with z_α. This confidence interval is in log space, so we can exponentiate the results to get the confidence interval of $\widehat{LR}$. The value of $\mathrm{var}_(\log(\widehat{LR}))$ (derived in Text S1) depends on the variances of the observed allele frequencies. These, in turn, depend on θ to accommodate evolutionary variation over populations and this is why numerical techniques such as bootstrapping cannot be used to calculate likelihood ratios, as explained by Beecham and Weir [81].

Simulating individuals

Using the data provided by Budowle and Moretti [29], individuals were simulated based on the allele frequencies reported for each of the five population samples. For the population structure analysis, individuals are simulated from a given population sample by independently drawing two alleles from the appropriate allele frequency distribution for every locus. Note that the total independence between drawn alleles implicitly creates a population with a coancestry coefficient of zero ($\theta=0$). Independently generated individuals are unrelated. Related individuals are simulated by generating unrelated individuals and randomly dropping alleles through a pedigree to achieve the desired relationship. In this way, we simulate pairs of both unrelated and related individuals from each population sample.

The total lack of population structure or cryptic relatedness ($\theta=0$) in our simulated populations causes unrelated individuals to share fewer alleles than would be expected in a real population. This contrasts with our use of the $\theta=0.01$ correction in $\widehat{LR}$ calculations, conservatively lowering our calculated $\widehat{LR}$. This is consistent with forensic applications, where a conservatively high value for θ is chosen for the anticipated populations. Specifically, $\theta=0.01$ and 0.03 have been suggested for use with populations primarily of European and Native American descent, respectively [43,83].

In the second part of this analysis, when we consider the interplay between various parameters, it is necessary to simulate unrelated individuals from a population with a given non-zero coancestry coefficient (θ). To simulate unrelated and related individuals from a population with $\theta\neq0$, random alleles are drawn

using the probabilities of two-individual genotypes, given θ and a specified relationship, as written in Text S1.

Comparative distribution analysis

We are interested in comparing LCL distributions generated with different parameters, particularly LCL distributions for truly unrelated individuals and truly related individuals. If the relationship $\widehat{LR}$ perfectly distinguished relatives and unrelated individuals, these two distributions would be totally separate. The degree of overlap between the related and unrelated distributions roughly indicates the degree of genetic similarity of relatives and unrelated individuals, and so, how well $\widehat{LR}$ distinguishes the two.

To quantify distinguishability, we use an empirical version of the measure proposed by Visscher and Hill [56]

$$\tilde{D}_{VH} = \frac{(\overline{\log(LR)}_r - \overline{\log(LR)}_u)^2}{s_r^2 + s_u^2}$$

where $\overline{\log(LR)}_r$ and $\overline{\log(LR)}_u$ are the sample means of $\log(\widehat{LR})$ for the simulations of related and unrelated individuals, respectively, and s_r^2 and s_u^2 are the sample variances of $\log(\widehat{LR})$ for the simulations of related and unrelated individuals, respectively. Note that $\tilde{D}_{VH}$ is analogous to the non-centrality parameter of the LR test statistic distribution under the alternative hypothesis. Higher $\tilde{D}_{VH}$ indicates greater LR distribution differentiation and more distinguishability, while lower $\tilde{D}_{VH}$ indicates more overlap and less distinguishability. The statistic $\tilde{D}_{VH}$ accurately describes the differentiation in LR distributions, and is particularly appealing because it describes the difference in distributions, so it does not rely on a parameterized decision procedure to discretely determine relationship status.

Supporting Information

Figure S1 Confidence intervals by population samples. Each plot shows the 100 replicates of $\widehat{LR}$ 95% confidence intervals for a sibling relationship between unrelated individuals, assuming allele frequencies based on the named population sample. Within each plot, the colored bands show the population sample allele frequencies used to simulate the unrelated individuals. Red signifies Vietnamese, orange African American, purple European American blue Latino, and green Navajo. The vertical line indicates $\widehat{LR}=1$.
(EPS)

Figure S2 Allele frequency distributions. Each plot shows the D3S1358 allele frequency distribution for each population.
(EPS)

Figure S3 $\tilde{D}_{VH}$ versus entropy. The empirical distinguishability ($\tilde{D}_{VH}$) is plotted against entropy for each population sample.
(EPS)

Figure S4 Distinguishability ($\tilde{D}_{VH}$) versus distance between true and assumed population samples ($\hat{\theta}$). The empirical distinguishability ($\tilde{D}_{VH}$) is plotted against $\hat{\theta}$ for each pair of true and assumed population samples. Points are colored according to the true population sample in the stated color scheme. $\hat{\theta}$ estimates less than 0.0 are reported as 0.0.
(EPS)

Figure S5 $\tilde{D}_{VH}$ over number of markers and relationships. $\tilde{D}_{VH}$ is shown when simulating different numbers of STRs (first column) and SNPs (second column) for a variety of relationships, as labeled.
(EPS)

Figure S6 $\tilde{D}_{VH}$ over θ, θ_a, and relationships. $\tilde{D}_{VH}$ is shown when simulating 15 STRs (first column) and 100 SNPs (second column) with different values of θ used in the simulation and θ_a used in probability calculations for a variety of relationships, as labeled.
(EPS)

Text S1 Supporting work is presented, specifically genotype probability equations, $\mathrm{var}(\log(\widehat{LR}))$ derivation, low nominal false positive rates, relationship distinguishability and entropy, and tables.
(PDF)

Acknowledgments

We are immensely grateful to the individuals whose DNA samples make up the datasets referenced in this manuscript. This work could not have been performed without that information. In addition, we thank Timothy Thornton and Kirk Lohmueller for their thoughtful discussion during the preparation of this manuscript, Nanibaa' Garrison for her insightful guidance regarding data description and use, Kim TallBear for valuable consultation on appropriate data handling, and three anonymous reviewers for productive commentary and discussion.

Author Contributions

Conceived and designed the experiments: RVR SMF BSW. Performed the experiments: RVR. Analyzed the data: RVR. Wrote the paper: RVR SMF BSW.

References

1. FBI (September 2011) CODIS-NDIS statistics. URL http://www.fbi.gov/about-us/lab/codis/ndis-statistics.
2. Naik G (23 February 2008) The gene police. The Wall Street Journal.
3. Pope S, Clayton T, Whitaker J, Lowe J, Puch-Solis R (2009) More for the same? Enhancing the investigative potential of forensic DNA databases. Forensic Science International: Genetic Supplement Series 2: 458–459.
4. Rothstein M, Talbott M (2006) The expanding use of DNA in law enforcement: What role for privacy? The Journal of Law, Medicine, and Ethics 34: 153–164.
5. Haimes E (2006) Social and ethical issues in the use of familial searching in forensic investigations: Insights from family and kinship studies. The Journal of Law, Medicine, and Ethics 34: 263–276.
6. Greely H, Riordan D, Garrison N, Mountain J (2006) Family ties: The use of DNA offender databases to catch offenders' kin. Journal of Law, Medicine, and Ethics 34: 248–262.
7. Tansey B (27 April 2008) State widens DNA scanning in cold cases: Near-match a hint offender related to person in database. San Francisco Chronicle.
8. Watkins T (7 July 2010) Police make arrest in L.A.'s 'Grim Sleeper' killings. Associated Press.
9. Miller G (2010) Familial DNA testing scores a win in serial killer case. Science 329: 262.
10. Myers S, Timken M, Piucci M, Sims G, Greenwald M, et al. (2011) Searching for first-degree familial relationships in California's offender DNA database: Validation of a likelihood ratio-based approach. Forensic Science International: Genetics 5: 493–500.
11. US Department of Justice FBI. CODIS brochure. URL http://www.fbi.gov/hq/lab/pdf/codisbrochure2.pdf.
12. Gershaw C, Schweighardt A, Rourke L, Wallace M (2011) Forensic utilization of familial searches in DNA databases. Forensic Science International: Genetics 5: 16–20.
13. Santos S, Budowle B, Smerick J, Keys K, Moretti T (1997) Portuguese population data on the six short tandem repeat loci: CSF1PO, TPOX, THO1, D3S1358, VWA and FGA. Forensic Science International 83: 229–235.
14. Gutowski S, Budowle B, Auer J, van Oorschot R (1995) Statistical analysis of an Australian population for the loci gc, HLA-DQA1, D1S80 and HUMTH01. Forensic Science International 76: 1–6.
15. Urquhart A, Kimpton C, Downes T, Gill P (1994) Variation in short tandem repeat sequences: A survey of twelve microsatellite loci for use as forensic identification markers. International Journal of Legal Medicine 107: 13–20.
16. Budowle B, Planz J, Chakraborty R, Callaghan T, Eisenberg A (2006) Clarification of statistical issues related to the operation of CODIS. In:

Proceedings of the Promega Seventeenth International Symposium on Human Identification. volume 17. pp 1–20.
17. Murphy E (2010) Relative doubt: Familial searches of DNA databases. Michigan Law Review 109: 291–349.
18. Jesudason S, Ortega M, Baruch S, Lehman J, Quevedo V, et al. (2009) California forensic DNA databases: Impacts on communities of color. Technical report, Generations Ahead.
19. Hall C (12 May 2006) Experts suggest expanding DNA database: Adding relatives could point to suspects, they say. San Francisco Chronicle.
20. Goring H, Ott J (1997) Relationship estimation in affected sib pair analysis of late-onset diseases. European Journal of Human Genetics 5: 69–77.
21. Boehnke M, NJ C (1997) Accurate inference of relationships in sib-pair linkage studies. American Journal of Human Genetics 61: 423–429.
22. O'Connell J, Weeks D (1998) Pedcheck: A program for identification of genotype incompatibilities in linkage analysis. American Journal of Human Genetics 63: 259–266.
23. Ehm M, Wagner M (1998) A test statistic to detect errors in sib-pair relationships. American Journal of Human Genetics 62: 181–188.
24. McPeek M, Sun L (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. American Journal of Human Genetics 66: 1076–1094.
25. Abecasis G, Cherny S, Cookson W, Cardon L (2001) Grr: Graphical representation of relationship errors. Bioinformatics 17: 742–743.
26. Sieberts S, Wijsman E, Thompson E (2002) Relationship inference from trios of individuals, in the presence of typing error. American Journal of Human Genetics 70: 170–180.
27. Purcell S, B N, Todd-Brown K, Thomas L, Ferreira M, et al. (2007) PLINK: A tool set for wholegenome association and population-based linkage analyses. American Journal of Human Genetics 81: 559–575.
28. Stevens E, Heckenberg G, Roberson E, Baugher J, Downey T, et al. (2011) Inference of relationships in population data using Identity-by-Descent and Identity-by-State. PLoS Genet 7: e1002287. doi:10.1371/journal.pgen.1002287.
29. Budowle B, Moretti TR (1998) Examples of STR population databases for CODIS and casework. 9th International Symposium on Human Identification 1: 64–73.
30. Neyman J, Pearson E (1933) On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character 231: 289–337.
31. Bieber F, Brenner C, Lazer D (2006) Finding criminals through DNA of their relatives. Science 312: 1315–1316.
32. Nei M (1973) Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences of the United States of America 70: 3321–3323.
33. Weir B, Cockerham C (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358–1370.
34. Jakobsson M, Scholz S, Scheet P, Gibbs J, VanLiere J, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998–1003.
35. Cavalli-Sforza L, Menozzi P, Piazza A (1994) The history and geography of human genes Princeton University Press.
36. Weir B (2007) The rarity of DNA profiles. The Annals of Applied Statistics 1: 358–370.
37. Budowle B, Giusti A, Waye J, Baechtel F, Fourney R, et al. (1991) Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. American Journal of Human Genetics 48: 841–855.
38. Green P (1992) Population genetic issues in DNA fingerprinting. American Journal of Human Genetics 50: 441–443.
39. Budowle B (1992) Reply to Green. American Journal of Human Genetics 50: 443–446.
40. Weir B (1992) Population genetics in the forensic DNA debate. Proceedings of the National Academy of Sciences of the United States of America 89: 11654–11659.
41. Ott J (1992) Strategies for characterizing highly polymorphic markers in human gene mapping. American Journal of Human Genetics 51: 283–290.
42. Knapp M, Seuchter S, Baur M (1993) The effect of misspecifying allele frequencies in incompletely typed families. Genetic Epidemiology 10: 413–418.
43. Budowle B, Shea B, Niezgoda S, Chakraborty R (2001) CODIS STR loci data from 41 sample populations. Journal of Forensic Sciences 46: 453–489.
44. Anderson A, Weir B (2007) A maximum likelihood method for estimation of pairwise relatedness in structured populations. Genetics 176: 421–440.
45. Zhai XD, Xue XQ, Mo YN, Zhao GS, Ai HW, et al. (2009) False homozygosities at CSF1PO loci revealed by discrepancies between two kits in Chinese population. International Journal of Legal Medicine 124: 457–458.
46. Heinrich M, Müller M, Rand S, Brinkmann B, Hohoff C (2004) Allelic drop-out in the STR system ACTB2 (SE33) as a result of mutations in the primer binding region. International Journal of Legal Medicine 118: 361–363.
47. Forrest S, Kupferschmid T, Hendrickson B, Judkins T, Petersen D, et al. (2004) Two rare novel polymorphisms in the D8S1179 and D13S317 markers and method to mitigate their impact on human identification. Croatian Medical Journal 45: 457–460.
48. Grgicak C, Rogers S, Mauterer C (2006) Discovery and identification of new D13S317 primer binding site mutations. Forensic Science International 157: 36–39.
49. Mizuno N, Kityama T, Fujii K, Nakahara H, Yoshida K, et al. (2008) A D19S433 primer binding site mutation and frequency in Japanese of silent allele it causes. Journal of Forensic Science 53: 1068–1073.
50. Clayton T, Hill S, Denton L, Watson S, Urquhart A (2004) Primer binding site mutations affecting the typing of STR loci contained within AMPFlSTR RSGM PlusTMkit. Forensic Science International 139: 255–259.
51. Boutrand L, Egyed B, Füredi S, Mommers N, Mertens G, et al. (2001) Variations in primer sequences are the origin of allele drop-out at loci D13S317 and CD4. International Journal of Legal Medicine 114: 295–297.
52. Lewis K (2009) Genomic Approaches to Forensic DNA Analysis. Ph.D. thesis, University of Washington.
53. Pemberton T, Wang C, Li J, Rosenberg N (2010) Inference of unexpected genetic relatedness among individuals in HapMap Phase III. The American Journal of Human Genetics 87: 457–464.
54. Epstein M, Duren W, Boehnke M (2000) Improved inference of relationship for pairs of individuals. The American Journal of Human Genetics 67: 1219–1231.
55. Scientific Working Group on DNA Analysis Methods (SWGDAM) (2009) 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 Science Communications 11: 1–12.
56. Visscher P, Hill W (2009) The limits of individual identification from sample allele frequencies: Theory and statistical analysis. PLoS Genet 5: e1000628. doi:10.1371/journal.pgen.1000628.
57. Browning B, Browning S (2011) A fast, powerful method for detecting identity by descent. The American Journal of Human Genetics 88: 173–182.
58. Moltke I, Albrechtsen A, Hansen T, Nielsen F, Nielsen R (2011) A method for detecting IBD regions simultaneously in multiple individuals – with applications to disease genetics. Genome Research 21: 1168–1180.
59. Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, et al. (2002) Genetic structure of human populations. Science 298: 2381–2385.
60. Rosenberg N, Mahajan S, Ramachandran S, Zhao C, Pritchard J, et al. (2005) Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1: e70. doi:10.1371/journal.pgen.0010070.
61. DeGiorgio M, Jakobsson M, Rosenberg N (2009) Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from africa. Proceedings of the National Academy of Sciences of the United States of America 106: 16057–16062.
62. Auton A, Bryc K, Boyko A, Lohmueller K, Novembre J, et al. (2009) Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Research 19: 795–803.
63. Lee SJ, Mountain J, Koenig B, Altman R, Brown M, et al. (2008) The ethics of characterizing difference: guiding principles on using racial categories in human genetics. Genome Biology 9: 404.
64. Mueller L (2008) Can simple population genetic models reconcile partial match frequencies observed in large forensic databases? Journal of Genetics 87: 101–108.
65. Krane D, Bahn V, Balding D, Barlow B, Cash H, et al. (2009) Time for DNA disclosure. Science 326: 1631–1632.
66. Mauer M (2009) Racial disparities in the criminal justice system. Technical report, The Sentencing Project.
67. Young T (1990) Native American crime and criminal justice require criminologists' attention. Journal of Criminal Justice Education 1: 111–116.
68. Armstrong T, Guilfoyle M, Melton A (1996) Native Americans, Crime, and Justice, Westview Press, chapter Native American delinquency: An overview of prevalence, causes, and correlates. pp 75–88.
69. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4: e10000167. doi:10.1371/journal.pgen.1000167.
70. Dalton R (2002) Tribe blasts 'exploitation' of blood samples. Nature 420: 111.
71. Wiwchar D (16 December 2004) Nuu-chah-nulth blood returns to west coast. Ha-Shilth-Sa.
72. Mello M, Wolf L (2010) The Havasupai Indian tribe case – lessons for research involving stored biologic samples. The New England Journal of Medicine 363: 204–207.
73. Asociación ANDES (May 2011) Genographic project hunts the last of the Incas. ANDES Communiqué.
74. Arbour L, Cook D (2006) DNA on loan: Issues to consider when carrying out genetic research with Aboriginal families and communities. Community Genetics 9: 153–160.
75. Goering S, Holland S, Fryer-Edwards K (2008) Transforming genetic research practices with marginalized communities: A case for responsive justice. Hastings Center Report 38: 43–53.
76. Anderson J (2009) Commentary on implications of the Genographic Project. International Journal of Cultural Property 16: 213–217.
77. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P (2009) Data sharing in genomics – re-shaping scientific practice. Nature Reviews Genetics 10: 331–335.
78. McIness R (2011) 2010 presidential address: Culture: The silent language geneticists must learn – genetic research with Indigenous populations. American Journal of Human Genetics 88: 254–261.

79. Weir B, Anderson A, Helper A (2006) Genetic relatedness analysis: modern data and new challenges. Nature Reviews Genetics 7: 771–780.
80. Holsinger K, Weir B (2009) Genetics in geographically structured populations: defining, estimating and interpreting f_{ST}. Nature Reviews Genetics 10: 639–650.
81. Beecham G, Weir B (2011) Confidence interval of the likelihood ratio associated with mixed stain DNA evidence. Journal of Forensic Sciences 56: S166–S171.
82. Curran J, Triggs C, Buckleton J, Weir B (1999) Interpreting DNA mixtures in structured populations. Journal of Forensic Sciences 44: 987–995.
83. National Research Council: Committee on DNA forensic science (1996) The evaluation of forensic DNA evidence. National Academy Press.