Report

Current Biology

Individual Identifiability Predicts Population Identifiability in Forensic Microsatellite Markers

Highlights

- Ancestry information is compared for the CODIS forensic markers and non-CODIS loci
- The CODIS markers have ancestry information comparable to random marker sets
- F_{ST} can fail to adequately represent empirically observed ancestry information
- Ancestry information is inherent in markers with high individual identifiability

Authors

Bridget F.B. Algee-Hewitt, Michael D. Edge, Jaehee Kim, Jun Z. Li, Noah A. Rosenberg

Correspondence

noahr@stanford.edu

In Brief

Algee-Hewitt et al. study the relationship between the ability of genetic marker sets to determine individual identity and the ancestry information that marker sets encode. They find that ancestry information correlates with information about identity, suggesting that forensically desirable markers generally have nontrivial ancestry information.

Algee-Hewitt et al., 2016, Current Biology *26*, 935–942
April 4, 2016
http://dx.doi.org/10.1016/j.cub.2016.01.065

CellPress

Current Biology
Report

Individual Identifiability Predicts Population Identifiability in Forensic Microsatellite Markers

Bridget F.B. Algee-Hewitt,[1,3] Michael D. Edge,[1,3] Jaehee Kim,[1] Jun Z. Li,[2] and Noah A. Rosenberg[1,*]
[1]Department of Biology, Stanford University, Stanford, CA 94305, USA
[2]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA
[3]Co-first author
*Correspondence: noahr@stanford.edu
http://dx.doi.org/10.1016/j.cub.2016.01.065

SUMMARY

Highly polymorphic genetic markers with significant potential for distinguishing individual identity are used as a standard tool in forensic testing [1, 2]. At the same time, population-genetic studies have suggested that genetically diverse markers with high individual identifiability also confer information about genetic ancestry [3–6]. The dual influence of polymorphism levels on ancestry inference and forensic desirability suggests that forensically useful marker sets with high levels of individual identifiability might also possess substantial ancestry information. We study a standard forensic marker set—the 13 CODIS loci used in the United States and elsewhere [2, 7–9]—together with 779 additional microsatellites [10], using direct population structure inference to test whether markers with substantial individual identifiability also produce considerable information about ancestry. Despite having been selected for individual identification and not for ancestry inference [11], the CODIS markers generate nontrivial model-based clustering patterns similar to those of other sets of 13 tetranucleotide microsatellites. Although the CODIS markers have relatively low values of the F_{ST} divergence statistic, their high heterozygosities produce greater ancestry inference potential than is possessed by less heterozygous marker sets. More generally, we observe that marker sets with greater individual identifiability also tend toward greater population identifiability. We conclude that population identifiability regularly follows as a byproduct of the use of highly polymorphic forensic markers. Our findings have implications for the design of new forensic marker sets and for evaluations of the extent to which individual characteristics beyond identification might be predicted from current and future forensic data.

RESULTS

We aggregated microsatellite genotypes for 978 people sampled from 53 worldwide populations, considering 792 markers in total, including new CODIS genotypes. Because each CODIS locus has a tetranucleotide repeat unit, we focused much of our analysis on 432 of the 779 non-CODIS loci identified as tetranucleotide repeats [12]. For comparison with the CODIS loci, we produced 1,000 sets of 13 loci selected randomly from among these 432.

Individual Identifiability

To verify that the CODIS loci are especially suitable in individual identification, we used two statistics that quantify the extent to which the genotype of an individual determines identity: expected heterozygosity (H) and the probability that two random unrelated diploid individuals in a panmictic population have the same genotype (M). Individual genotypes are most distinctive for high-H and low-M loci.

We compared H and M between CODIS loci and non-CODIS tetranucleotide loci. The CODIS loci have higher H (Figure 1A) and lower M (Figure 1C) than non-CODIS tetranucleotides. Furthermore, $\overline{H}$ for the 13 CODIS loci exceeds $\overline{H}$ for 999 of the 1,000 sets of 13 random non-CODIS tetranucleotides (Figure 1B). $\overline{M}$ is lower for the 13 CODIS loci than for 991 of the 1,000 random sets (Figure 1D). The greater heterozygosity and lower match probability in the CODIS loci confirm that these markers possess greater individual identifiability than do random tetranucleotide sets.

F_{ST}

We next examined the relative potential of the CODIS loci for population identifiability. We compared F_{ST} among populations for CODIS and non-CODIS tetranucleotides; a low CODIS F_{ST} would usually be taken to suggest that the high individual identifiability of these loci is accompanied by low population identifiability.

The CODIS loci have lower F_{ST} than the 432 non-CODIS tetranucleotides, though not significantly so (Figure 1E). $\overline{F_{ST}}$ for the 13 CODIS loci is relatively small in relation to the 1,000 random sets of 13 non-CODIS tetranucleotides, but it still exceeds $\overline{F_{ST}}$ for 168 of these sets (Figure 1F). Thus, although the CODIS loci have high individual identifiability, as measured

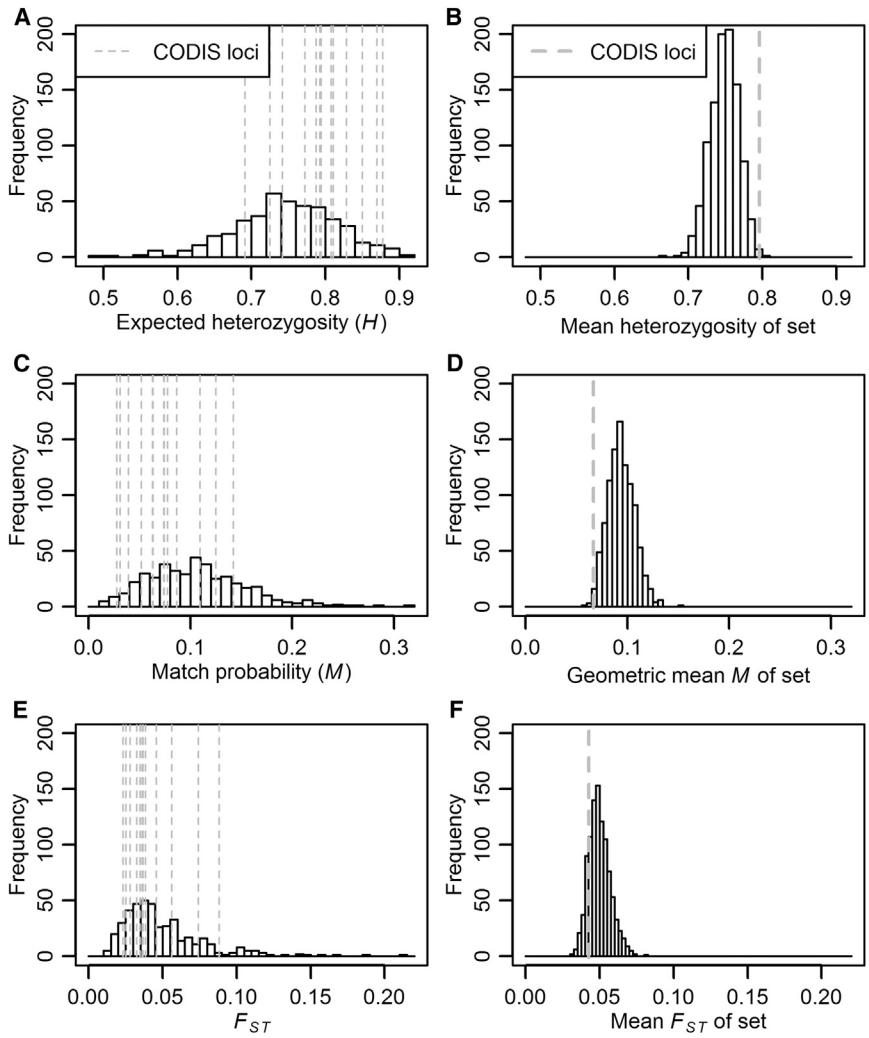

Figure 1. Population-Genetic Summary Statistics for CODIS and Non-CODIS Loci

(A–F) The 432 non-CODIS tetranucleotide markers appear in histograms, and the CODIS loci appear as dashed lines. The left side shows locus-wise results; the right side compares the single CODIS set to 1,000 sets of 13 non-CODIS tetranucleotides.

(A) Locus-wise expected heterozygosity *H*. CODIS mean 0.796; non-CODIS mean 0.747; two-sided Wilcoxon p = 0.01.

(B) $\overline{H}$.

(C) Locus-wise match probability *M* for unrelated diploid pairs. CODIS mean 0.074; non-CODIS mean 0.106; two-sided Wilcoxon p = 0.01.

(D) $\overline{M}$.

(E) Locus-wise F_{ST}. CODIS mean 0.042; non-CODIS mean 0.050; two-sided Wilcoxon p = 0.36.

(F) $\overline{F_{ST}}$.

by *H* and *M,* their level of F_{ST} genetic divergence is not unusually small.

STRUCTURE

F_{ST} only partially predicts the extent to which ancestry can be inferred from a locus set. To examine ancestry inference potential more directly, we compared the recovery by CODIS and non-CODIS marker sets of genetic clustering patterns obtained using larger sets. We compared STRUCTURE [13] solutions using 1,000 replicate runs with the CODIS loci, 100 runs with the 779 non-CODIS loci, and one run for each of the 1,000 random 13-locus non-CODIS sets. We also considered runs for each of 1,000 structure-free random null datasets with the same allele frequencies as the CODIS loci. We varied the number of clusters *K* from 2 to 6, focusing on *K* = 4, the preferred *K* for the CODIS loci.

At the continental level, the CODIS solutions indicate notable structure (Figure 2A) considerably more salient than that of null datasets (Figure 2B). This structure is, however, less apparent than that obtained from the full 779 loci (Figure 2D), which accords with the structure in past studies [10, 14]. The CODIS structure is visually similar to the pattern for 13 non-CODIS tetranucleotide markers (Figure 2C). We next considered three methods for performing this comparison quantitatively.

Clusteredness

The clusteredness statistic *B* measures the degree to which a STRUCTURE solution is clustered. *B* is greatest if estimated ancestry coefficients place each individual in exactly one cluster and smallest if each individual has equal membership in all clusters. We evaluated *B* for runs with each *K*, comparing the 1,000 CODIS replicates to the 1,000 runs using random marker sets.

Compared with random non-CODIS sets, at each *K*, the 1,000 CODIS replicates produce intermediate *B* (Table S1). At *K* = 4 clusters, they fall between the 35th and 59th percentiles of the distribution of *B* from the non-CODIS sets; the CODIS median is at the 53rd percentile (Figure 2E).

Similarity to Full-Data STRUCTURE Solutions

Because *B* measures placement into clusters but does not consider the nature of those clusters, we used a second statistic, *S*, to measure the mean similarity of a solution obtained using a small set of markers to a set of solutions obtained using a larger marker set. A large *S* indicates clustering similar to that produced with 779 non-CODIS loci. We evaluated *S* for each *K*, comparing each of the 1,000 CODIS replicates and 1,000 random sets to the *L* = 100 replicates with the full locus set.

Compared with the random tetranucleotide sets, at each *K*, the CODIS loci generate solutions with intermediate *S* (Table S1). For *K* = 4, the 1,000 CODIS replicates fall between the 41st and 56th percentiles of the *S* distribution for random sets, with the CODIS median occurring at the 52nd percentile (Figure 2F).

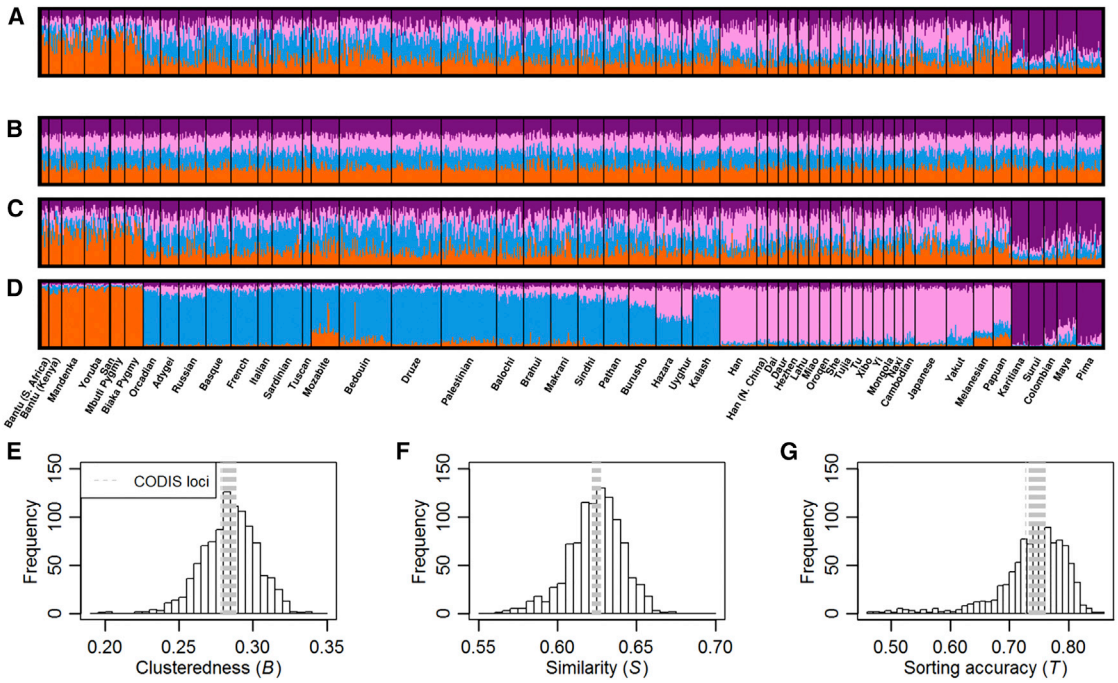

Figure 2. Properties of STRUCTURE Solutions for $K = 4$ Clusters

(A–D) STRUCTURE solutions. Each sampled individual is represented by a vertical line. Colors represent clusters, and the length of the line segment displayed in a color is proportional to the estimated membership for the associated cluster.

(A) CODIS loci. $B = 0.285$, $S = 0.626$, $T = 0.752$.

(B) 13-locus null dataset. $B = 0.084$, $S = 0.534$, $T = 0.258$.

(C) 13 random non-CODIS tetranucleotide markers. $B = 0.286$, $S = 0.625$, $T = 0.761$.

(D) 779 non-CODIS loci. $B = 0.746$, $T = 0.986$.

For (A), (B), and (C), the solution shown has the median S among 1,000 runs; the solution in (D) has the median B among 100 runs.

(E–G) Distributions of three indices describing STRUCTURE solutions. Distributions appear for 1,000 solutions using random sets of 13 non-CODIS loci (histogram) and for 1,000 solutions using the CODIS loci (dashed lines).

(E) Clusteredness B.

(F) Similarity to full data S.

(G) Sorting accuracy T.

Supporting information related to the figure appears in Tables S1, S2, and S5; a PCA analog appears in Figure S1.

Sorting Accuracy

All 100 STRUCTURE $K = 4$ replicates with 779 non-CODIS loci produced solutions in which, for each of seven geographic regions, >90% of samples from the region had their largest membership coefficient associated with the same cluster. In each solution, because regions showed the same co-clustering pattern, we defined the four groups in this pattern as "super-regions": Africa, Western Eurasia (Middle East, Europe, Central/South Asia), East Asia/Pacific (East Asia, Oceania), and the Americas.

We then defined "sorting accuracy" T, a computation that associates each of the four clusters with a super-region, evaluates for each super-region the fraction of individuals placed by STRUCTURE in the cluster associated with the super-region, and averages these fractions. T measures the extent to which the super-region of an individual is identifiable from a STRUCTURE run.

Averaging across 1,000 STRUCTURE solutions using the CODIS loci (Table S2), assignment accuracy is high for Africa (91%) and the Americas (89%), though somewhat lower for Western Eurasia (52%) and East Asia/Pacific (64%). The value of T for Table S2 is 74%, comparable to the median T across replicates, 75%. These values greatly exceed the median T of 27% for 1,000 runs with null datasets.

Compared with sets of random tetranucleotides, the CODIS loci produce $K = 4$ STRUCTURE solutions of intermediate T. The 1,000 CODIS replicates generate T values between the 33rd and 60th percentiles of the T distribution for random sets, with the CODIS median at the 52nd percentile (Figure 2G).

Individual Identifiability and Population Identifiability

By multiple measures, the CODIS loci have high individual identifiability and intermediate population identifiability. We next tested the generality of the observation that marker sets informative for individual identification possess intermediate rather than low information content for population identification. We computed Pearson correlations between the various measures, considering the 1,000 sets of 13 random tetranucleotides (Figure 3).

The measures of individual identifiability—$\overline{H}$ and $\overline{M}$—are inversely related ($r = -0.97$), as expected from the connection of individual identifiability with large H and small M. Similarly, the STRUCTURE-derived B, S, and T measures of population

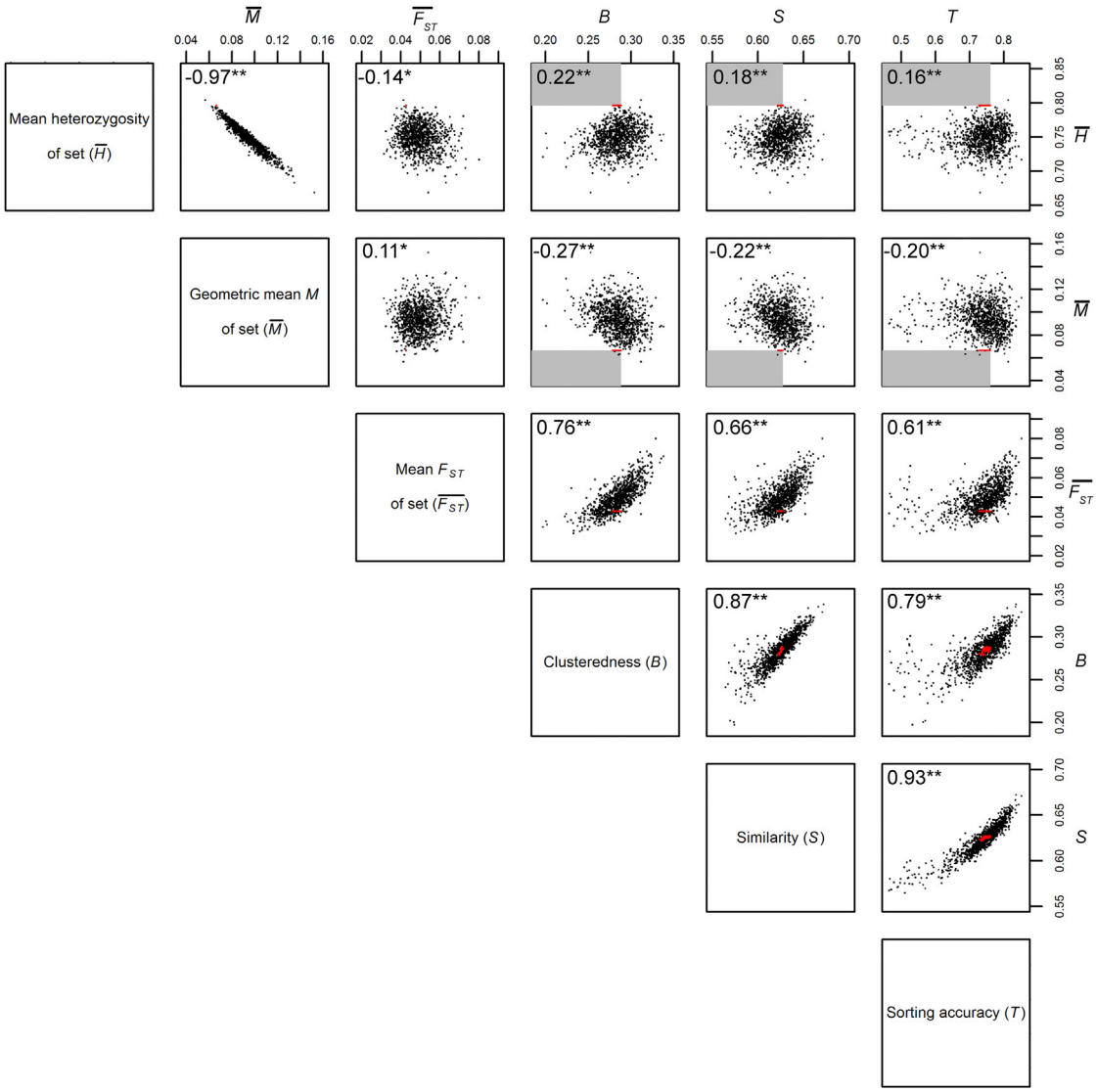

Figure 3. Individual Identifiability and Population Identifiability Statistics for 1,000 Random Non-CODIS 13-Marker Sets and 1,000 STRUCTURE Runs with the CODIS Loci

Each of 15 pairs of statistics is plotted: the 1,000 non-CODIS sets as black points and the 1,000 CODIS replicates as red points. Pearson correlation coefficients computed from the non-CODIS sets appear in the plots (*$p < 0.05$, **$p < 0.001$). For the six plots comparing $\overline{H}$ or $\overline{M}$ with B, S, or T, the gray box shows the region with individual identifiability at least as great as the CODIS markers and population identifiability at most that of the CODIS markers (as measured by the most ancestry-informative of 1,000 CODIS STRUCTURE replicates). The gray box contains 0, 0, 0, 5, 3, and 4 non-CODIS sets for the plots of ($\overline{H}$, B), ($\overline{H}$, S), ($\overline{H}$, T), ($\overline{M}$, B), ($\overline{M}$, S), and ($\overline{M}$, T), respectively. The three non-CODIS sets for ($\overline{M}$, S) are nested among the four sets for ($\overline{M}$, T), which are, in turn, nested among the five sets for ($\overline{M}$, B). Two loci are included in at least two of the five sets: D11S1986 in 4 and D12S1064 in 3. Partial-correlation adjustments for $\overline{F_{ST}}$ and analogous PCA-based correlations appear in Tables S3 and S4.

identifiability are positively correlated ($r > 0.75$). $\overline{F_{ST}}$ correlates positively with all three STRUCTURE-based measures, though less strongly ($r > 0.60$). All 15 correlations of the six statistics differ significantly from 0 ($p < 0.05$).

Relationships between information about individual identity and ancestry information differ by measure. Whereas individual identifiability correlates negatively with $\overline{F_{ST}}$, it correlates positively with ancestry information in the STRUCTURE-derived statistics B, S, and T.

We next examined the relationship between individual identifiability and STRUCTURE-based ancestry information, adjusting for $\overline{F_{ST}}$ by partial correlation. The associations between informativeness about identity with B, S, and T become more pronounced after adjustment (Table S3). This result suggests that among marker sets with similar $\overline{F_{ST}}$, sets with higher individual identifiability, measured by H and M, are associated with considerably higher population identifiability, measured by B, S, and T.

Principal Component Analysis

Having observed correlations of individual identifiability and population identifiability with STRUCTURE-based measures,

we used a second population structure inference method—principal component analysis (PCA)—to assess the generality of the relationship. We employed PCA in parallel to the STRUCTURE analysis, finding that PCA-based ancestry information measures replicate the patterns from STRUCTURE (Figure S1; Tables S2–S5).

DISCUSSION

We have examined the relationship between individual identifiability and population identifiability in microsatellite markers. The CODIS loci are more polymorphic than comparable equal-sized non-CODIS sets, confirming their relative suitability for individual identification (Figures 1B and 1D). However, although $\overline{F_{ST}}$ for the CODIS loci might suggest that they contain less ancestry information than typical tetranucleotide sets (Figure 1F), high CODIS heterozygosity enables STRUCTURE solutions just as similar to solutions from 779 microsatellites as those produced with non-CODIS tetranucleotides (Figure 2; Table S1). Moreover, non-CODIS sets with higher utility for identifying individuals possess more ancestry information via STRUCTURE-based statistics, in apparent contradiction with the negative correlation of $\overline{F_{ST}}$ and individual identifiability (Figure 3). A close relationship between individual identifiability and empirical ancestry information is further evident in partial-correlation adjustments for $\overline{F_{ST}}$ (Table S3) and in analyses using PCA in place of STRUCTURE.

Many studies have examined the use of forensic markers to produce probabilistic hypotheses about ancestry-related aspects of DNA samples [15–31], applying locus informativeness measures to propose small marker panels with potential for inferring ancestry for arbitrary samples of unknown origin or taking panels such as the CODIS set as given and evaluating their ancestry information. Our study is novel in combining elements of both types of studies; like the latter evaluative studies, we extend ancestry inferences for a standard locus set. Unlike such studies, but like some panel-design studies, we also analyze many sets to relate ancestry information for the loci of interest to that of comparable markers.

Individual Identifiability and Population Identifiability

That the CODIS loci possess similar ancestry information to non-CODIS sets is surprising given arguments from forensic genetics, which have claimed that (1) loci selected for heterozygosity and individual identification encode little ancestry information, and (2) because forensic loci are selected in this manner, they are particularly ancestry-uninformative [1, 20, 32]. In contrast with these claims, which have often relied on F_{ST} to gauge ancestry information content, by testing ancestry inference using STRUCTURE rather than relying on the more indirect F_{ST}, we found that the CODIS loci contain ancestry information comparable to tetranucleotide sets not preselected to be ancestry-informative.

More generally, in agreement with results in population genetics [3–6], our STRUCTURE analysis contradicts a perspective that loci selected for individual identification possess little ancestry information. We surmise that these problematic arguments have been grounded in an emphasis on F_{ST}, which has an underappreciated mathematical downward trend for high-heterozygosity markers [33, 34]. In particular, for such markers, F_{ST} is mathematically bounded well below 1 [33, 34], potentially masking significant potential to facilitate ancestry inference. High-heterozygosity loci can have rare alleles whose population differences contribute substantially to ancestry inference potential but only minimally to divergence statistics [3]. The effect is apparent in the detailed ancestry inference possible in high-heterozygosity African populations, despite comparatively low among-group F_{ST} [34, 35].

To illustrate that F_{ST} is mathematically constrained for loci with high heterozygosity, so that such loci might potentially be ancestry-informative even when F_{ST} is low, we plotted heterozygosity and F_{ST} for the CODIS and non-CODIS loci in the context of recent mathematical results. For each of the six super-region pairs, Figure 4 depicts the values in relation to the F_{ST} upper bound given heterozygosity.

The F_{ST} upper bound decreases monotonically for heterozygosities exceeding 0.5, strongly constraining F_{ST} for the most heterozygous loci. Many CODIS and non-CODIS loci have high enough heterozygosities that even if each of their alleles was population specific, high F_{ST} values would be unattainable. For the most heterozygous loci—those with the greatest individual identifiability—F_{ST} is necessarily small even when the loci are highly ancestry-informative. Thus, for the high-heterozygosity loci of forensic interest, low F_{ST} does not necessarily indicate low ancestry information. The decline of the F_{ST} upper bound with increasing heterozygosity explains why individual identifiability correlates negatively with $\overline{F_{ST}}$ for high-heterozygosity marker sets, even when it correlates positively with the direct empirical ancestry information measured by the STRUCTURE-derived B, S, and T.

Implications

The CODIS loci have sometimes been used to assess population-genetic questions about geographic structure [32, 36, 37]. That they partially reflect the population structure observed in larger marker sets affirms the value of this approach. Further, the existence of population structure in standard forensic markers reinforces the importance in routine forensic practice of considering population differences in allele frequencies and accounting for population structure by co-ancestry adjustment [38].

Our results contribute to discussions of the information encoded by forensic markers, about whether CODIS profiles represent "sanitized 'genetic fingerprints' that can be used to identify an individual uniquely, but do not disclose an individual's traits, disorders, or dispositions" [39], as conditions of their use intend [39–42]. Together, the markers can provide information that might be used to probabilistically characterize individual ancestry-related traits beyond a profile match. Though our study is genotypic, probabilistic connections between CODIS profiles and phenotypic traits might become accessible through partial associations that might exist between genetic ancestry, socially defined concepts of race and ethnicity, and forensic and biomedical phenotypes. Evaluation of benefits and concerns of such possibilities requires accurate data on the level of ancestry information present in the loci: incomplete, but not negligible either.

An important consideration is the finding that not only do the CODIS loci possess ancestry information, but via the link

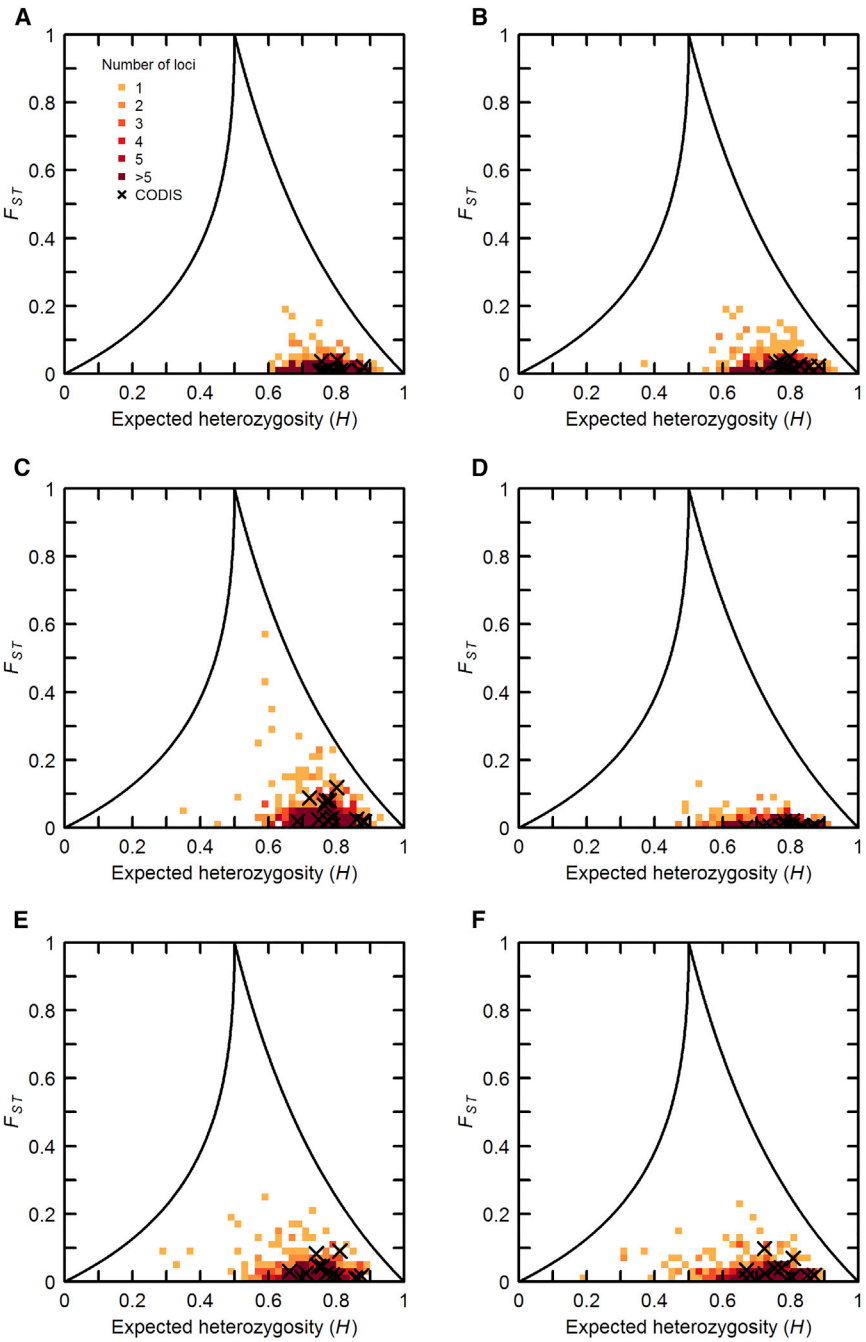

Figure 4. F_{ST} and Heterozygosity in Pairs of Super-Regions for CODIS and Non-CODIS Loci, in Relation to the Maximal F_{ST} as a Function of Heterozygosity

(A–F) In each panel, the solid curve depicts the upper bound on F_{ST} given heterozygosity, as computed in equations 31 and 32 of [34]. At each locus, allele frequencies are obtained separately for two super-regions, and the total allele frequency for use in computing heterozygosity is the average of the two frequencies. For the six panels, the pairs of super-regions are Africa, Western Eurasia (A); Africa, East Asia/Pacific (B); Africa, America (C); Western Eurasia, East Asia/Pacific (D); Western Eurasia, America (E); East Asia/Pacific, America (F). To make the mathematical bound applicable, sample allele frequencies are treated as parametric (equivalent to using infinite sample size N in the equation for H), and F_{ST} is computed using these frequencies as in equation 30 of [34]. CODIS loci are depicted individually, and non-CODIS loci are grouped in bins of size 0.02 × 0.02.

between individual identifiability and population identifiability, a goal of maximizing individual identifiability inherently conflicts with a goal of minimizing the ancestry information available from marker profiles. Indeed, the conflict is illustrated by the comparative success of the CODIS set for achieving this pair of goals: among 1,000 random sets, only 0 to 5 simultaneously contain greater individual identifiability and lower population identifiability than the CODIS loci, depending on the choice of measures (Figure 3). Nevertheless, as expected in light of the correlation between these quantities, CODIS-based ancestry information remains salient.

Recent studies using heterozygosity and F_{ST} criteria have sought to assemble globally variable marker sets with high individual identifiability and low population identifiability [43, 44]; such panels, considering single-nucleotide polymorphisms (SNPs), have achieved high individual identifiability while reducing ancestry inference potential in relation to larger random marker sets. However, although theory suggests that SNP-microsatellite heterozygosity differences will affect the relationship between individual identifiability and population identifiability [34]—with SNPs near 0.5 mean frequency in two populations less susceptible to misleading F_{ST} interpretations—in accord with our findings, ancestry inference potential in the SNP panels proposed is not eliminated; rather, PCA continues to show continental structuring [43].

Updates to forensic systems are now under consideration, capitalizing on advances made possible by new genomic data [45–48]. Irrespective of whether population identifiability is desirable, unless close attention is paid to the correlation between individual identifiability and population identifiability beyond computations of F_{ST}, future marker sets that enhance individual identifiability are likely to increase population identifiability as well.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, one figure, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2016.01.065.

AUTHOR CONTRIBUTIONS

Conceptualization, B.F.B.A.-H., M.D.E., and N.A.R.; Project Design, B.F.B.A.-H., M.D.E., J.Z.L., and N.A.R.; Data Management, B.F.B.A.-H., M.D.E., and J.Z.L.; Data Analysis, M.D.E. and J.K.; Supervision, N.A.R.; Writing, all authors.

ACKNOWLEDGMENTS

We thank two reviewers for comments and acknowledge support from National Institute of Justice grant 2014-DN-BX-K015.

Received: August 28, 2015
Revised: December 10, 2015
Accepted: January 26, 2016
Published: March 17, 2016

REFERENCES

1. Jobling, M.A., and Gill, P. (2004). Encoded evidence: DNA in forensic analysis. Nat. Rev. Genet. *5*, 739–751.
2. Butler, J.M. (2006). Genetics and genomics of core short tandem repeat loci used in human identity testing. J. Forensic Sci. *51*, 253–265.
3. Rosenberg, N.A., Burke, T., Elo, K., Feldman, M.W., Freidlin, P.J., Groenen, M.A.M., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., et al. (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics *159*, 699–713.
4. Rosenberg, N.A., Li, L.M., Ward, R., and Pritchard, J.K. (2003). Informativeness of genetic markers for inference of ancestry. Am. J. Hum. Genet. *73*, 1402–1422.
5. Liu, N., Chen, L., Wang, S., Oh, C., and Zhao, H. (2005). Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genet. *6* (*Suppl 1*), S26.
6. Haasl, R.J., and Payseur, B.A. (2011). Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. Heredity (Edinb) *106*, 158–171.
7. Budowle, B., Moretti, T.R., Niezgoda, S.J., and Brown, B.L. (1998). CODIS and PCR-based short tandem repeat loci: law enforcement tools. In Proceedings of the Second European Symposium on Human Identification (Promega Corporation), pp. 73–88.
8. Budowle, B., Shea, B., Niezgoda, S., and Chakraborty, R. (2001). CODIS STR loci data from 41 sample populations. J. Forensic Sci. *46*, 453–489.
9. Gill, P. (2002). Role of short tandem repeat DNA in forensic casework in the UK–past, present, and future perspectives. Biotechniques *32*, 366–385.
10. Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K., and Feldman, M.W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. *1*, e70.
11. Butler, J.M. (2001). Forensic DNA Typing (London: Academic Press).
12. Pemberton, T.J., Sandefur, C.I., Jakobsson, M., and Rosenberg, N.A. (2009). Sequence determinants of human microsatellite variability. BMC Genomics *10*, 612.
13. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
14. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.
15. Evett, I.W., Pinchin, R., and Buffery, C. (1992). An investigation of the feasibility of inferring ethnic origin from DNA profiles. J. Forensic Sci. Soc. *32*, 301–306.
16. Chakraborty, R., Stivers, D.N., Su, B., Zhong, Y., and Budowle, B. (1999). The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis *20*, 1682–1696.
17. Lowe, A.L., Urquhart, A., Foreman, L.A., and Evett, I.W. (2001). Inferring ethnic origin by means of an STR profile. Forensic Sci. Int. *119*, 17–22.
18. Klintschar, M., Füredi, S., Egyed, B., Reichenpfader, B., and Kleiber, M. (2003). Estimating the ethnic origin (EEO) of individuals using short tandem repeat loci of forensic relevance. Int. Congr. Ser. *1239*, 53–56.
19. Sun, G., McGarvey, S.T., Bayoumi, R., Mulligan, C.J., Barrantes, R., Raskin, S., Zhong, Y., Akey, J., Chakraborty, R., and Deka, R. (2003). Global genetic variation at nine short tandem repeat loci and implications on forensic genetics. Eur. J. Hum. Genet. *11*, 39–49.
20. Barnholtz-Sloan, J.S., Pfaff, C.L., Chakraborty, R., and Long, J.C. (2005). Informativeness of the CODIS STR loci for admixture analysis. J. Forensic Sci. *50*, 1322–1326.
21. Phillips, C., Salas, A., Sánchez, J.J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., Calaza, M., de Cal, M.C., Ballard, D., Lareu, M.V., and Carracedo, A.; SNPforID Consortium (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci. Int. Genet. *1*, 273–280.
22. Halder, I., Shriver, M., Thomas, M., Fernandez, J.R., and Frudakis, T. (2008). A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum. Mutat. *29*, 648–658.
23. Graydon, M., Cholette, F., and Ng, L.-K. (2009). Inferring ethnicity using 15 autosomal STR loci–comparisons among populations of similar and distinctly different physical traits. Forensic Sci. Int. Genet. *3*, 251–254.
24. Kidd, K.K., Speed, W.C., Pakstis, A.J., and Kidd, J.R. (2011). The search for better markers for forensic ancestry inference. In Proceedings of the 22nd International Symposium on Human Identification (Promega Corporation), pp. 1–4.
25. Pereira, L., Alshamali, F., Andreassen, R., Ballard, R., Chantratita, W., Cho, N.S., Coudray, C., Dugoujon, J.-M., Espinoza, M., González-Andrade, F., et al. (2011). PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile. Int. J. Legal Med. *125*, 629–636.
26. Phillips, C., Fernandez-Formoso, L., Garcia-Magariños, M., Porras, L., Tvedebrink, T., Amigo, J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Freire-Aradas, A., et al. (2011). Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. Forensic Sci. Int. Genet. *5*, 155–169.
27. Phillips, C., Fernandez-Formoso, L., Gelabert-Besada, M., Garcia-Magariños, M., Santos, C., Fondevila, M., Carracedo, A., and Lareu, M.V. (2013). Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing. Electrophoresis *34*, 1151–1162.
28. Kidd, K.K., Speed, W.C., Pakstis, A.J., Furtado, M.R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F.R., and Kidd, J.R. (2014). Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci. Int. Genet. *10*, 23–32.
29. Phillips, C., Gelabert-Besada, M., Fernandez-Formoso, L., García-Magariños, M., Santos, C., Fondevila, M., Ballard, D., Syndercombe Court, D., Carracedo, A., and Lareu, M.V. (2014). "New turns from old STaRs": enhancing the capabilities of forensic short tandem repeat analysis. Electrophoresis *35*, 3173–3187.
30. Phillips, C. (2015). Forensic genetic analysis of bio-geographical ancestry. Forensic Sci. Int. Genet. *18*, 49–65.
31. Phillips, C., Amigo, J., Carracedo, Á., and Lareu, M.V. (2015). Tetra-allelic SNPs: informative forensic markers compiled from public whole-genome sequence data. Forensic Sci. Int. Genet. *19*, 100–106.
32. Silva, N.M., Pereira, L., Poloni, E.S., and Currat, M. (2012). Human neutral genetic variation and forensic STR data. PLoS ONE *7*, e49666.
33. Hedrick, P.W. (2005). A standardized genetic differentiation measure. Evolution *59*, 1633–1638.

34. Jakobsson, M., Edge, M.D., and Rosenberg, N.A. (2013). The relationship between F_{ST} and the frequency of the most frequent allele. Genetics *193*, 515–528.
35. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.
36. Rowold, D.J., and Herrera, R.J. (2003). Inferring recent human phylogenies using forensic STR technology. Forensic Sci. Int. *133*, 260–265.
37. Rubi-Castellanos, R., Martínez-Cortés, G., Muñoz-Valle, J.F., González-Martín, A., Cerda-Flores, R.M., Anaya-Palafox, M., and Rangel-Villalobos, H. (2009). Pre-Hispanic Mesoamerican demography approximates the present-day ancestry of Mestizos throughout the territory of Mexico. Am. J. Phys. Anthropol. *139*, 284–294.
38. Steele, C.D., and Balding, D.J. (2014). Statistical evaluation of forensic DNA profile evidence. Annu. Rev. Stat. Appl. *1*, 361–384.
39. 73 Fed. Reg. at 74937 (2008).
40. Katsanis, S.H., and Wagner, J.K. (2013). Characterization of the standard and recommended CODIS markers. J. Forensic Sci. *58* (*Suppl 1*), S169–S172.
41. Maryland v. King, 133 S. Ct. 1958 (2013).
42. Greely, H.T., and Kaye, D.H. (2013). A brief of genetics, genomics and forensic science researchers in Maryland v. King. Jurimetrics *54*, 43–64.
43. Pakstis, A.J., Speed, W.C., Fang, R., Hyland, F.C.L., Furtado, M.R., Kidd, J.R., and Kidd, K.K. (2010). SNPs for a universal individual identification panel. Hum. Genet. *127*, 315–324.
44. Kidd, K.K., Kidd, J.R., Speed, W.C., Fang, R., Furtado, M.R., Hyland, F.C.L., and Pakstis, A.J. (2012). Expanding data and resources for forensic use of SNPs in individual identification. Forensic Sci. Int. Genet. *6*, 646–652.
45. Butler, J.M., Coble, M.D., and Vallone, P.M. (2007). STRs vs. SNPs: thoughts on the future of forensic DNA testing. Forensic Sci. Med. Pathol. *3*, 200–205.
46. Butler, J.M., and Hill, C.R. (2012). Biology and genetics of new autosomal STR loci useful for forensic DNA analysis. Forensic Sci. Rev. *24*, 15–26.
47. Ge, J., Eisenberg, A., and Budowle, B. (2012). Developing criteria and data to determine best options for expanding the core CODIS loci. Investig. Genet. *3*, 1.
48. Hares, D.R. (2012). Expanding the CODIS core loci in the United States. Forensic Sci. Int. Genet. *6*, e52–e54.

