Current Biology, Volume *26*

Supplemental Information

Individual Identifiability Predicts Population Identifiability in Forensic Microsatellite Markers

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

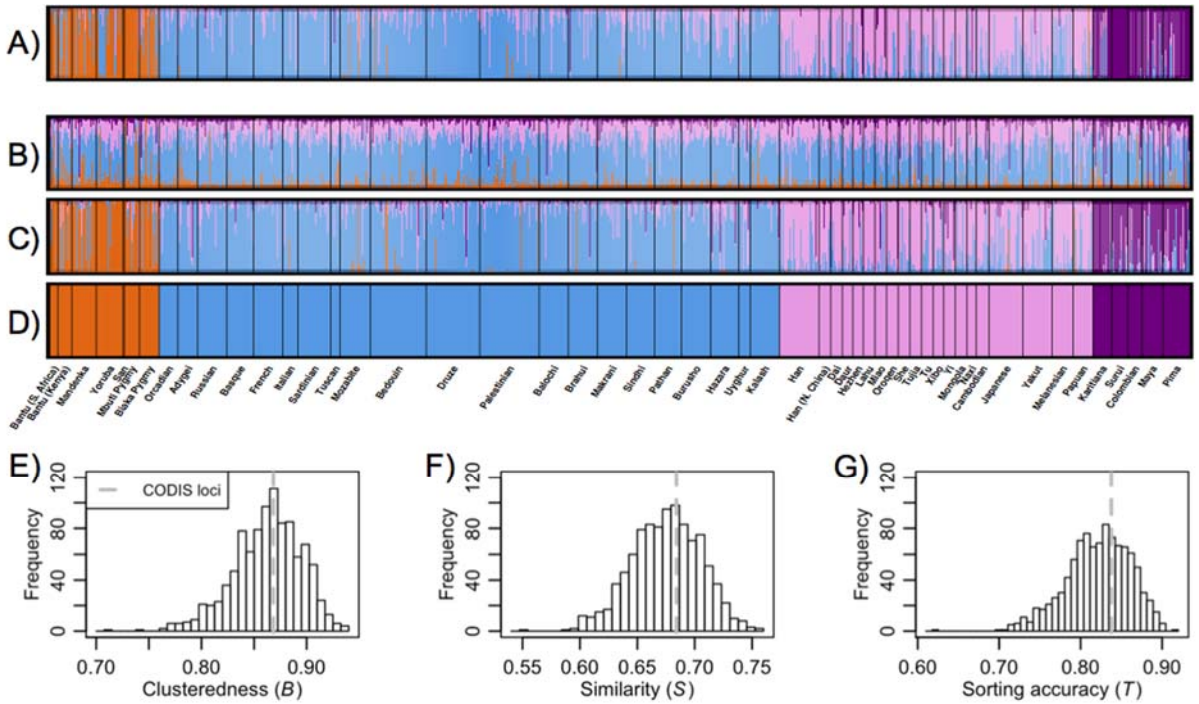

Figure S1. Properties of DAPC inferences for *K*=4 clusters. (A-D) DAPC inferences. Each sampled individual is represented by a vertical line. Colors represent clusters, and the length of the line segment displayed in a color is proportional to the estimated membership for the associated cluster. (A) CODIS loci. B_{DAPC}=0.868, S_{DAPC}=0.684, T_{DAPC}=0.838. (B) 13-locus null dataset. B_{DAPC}=0.548, S_{DAPC}=0.479, T_{DAPC}=0.374. (C) 13 random non-CODIS tetranucleotide markers. B_{DAPC}=0.867, S_{DAPC}=0.677, T_{DAPC}=0.841. (D) 779 non-CODIS loci. B_{DAPC}=1.000, S_{DAPC}=1.000, T_{DAPC}=1.000. For (B) and (C), the solution shown has the median *S* among 1000 runs. (E-G) Distributions of three indices describing DAPC inferences. The value for the CODIS loci (dashed lines) appears together with the distributions for 1000 solutions using random sets of 13 non-CODIS loci (histogram). (E) Clusteredness B_{DAPC}. (F) Similarity-to-full-data S_{DAPC}. (G) Sorting accuracy T_{DAPC}. The percentile for the CODIS loci is 54.1 for B_{DAPC}, 60.7 for S_{DAPC}, and 64.4 for T_{DAPC}. The figure design follows Figure 2.

Table S1. *B* and *S* percentiles of the median STRUCTURE replicate using the CODIS loci in relation to the distribution from runs with 1000 random sets of 13 non-CODIS loci. The *K*=4 median-*S* replicate is plotted in Figure 2A.

K	Clusteredness *B*	Similarity *S*
2	36.8	48.8
3	46.6	12.1
4	53.4	51.8
5	49.5	45.7
6	38.3	42.4

Table S2. Mean confusion matrices sorting 978 individuals into *K*=4 clusters using the CODIS loci. For the top (STRUCTURE-based) matrix, the value of *T* is 74%; the median *T* is 75% across 1000 random 13-locus sets and 27% across 1000 null datasets. For the bottom (DAPC-based) matrix, the value of T_{DAPC} is 84%; the median T_{DAPC} is 82% across 1000 random 13-locus sets and 36% across 1000 null datasets. The matrices summarize STRUCTURE and DAPC inferences of the form plotted in Figures 2A-2D and S1A-S1D.

STRUCTURE-based confusion matrix						
	Cluster 1 (orange)	Cluster 2 (blue)	Cluster 3 (pink)	Cluster 4 (violet)	Total	Sorting accuracy (%)
Africa	85.4	2.7	3.1	2.8	94	90.9
Western Eurasia	101.5	275.3	78.7	76.6	532	51.7
East Asia/Pacific	29.9	16.8	172.7	49.6	269	64.2
America	1.6	1.6	5.9	73.8	83	88.9
DAPC-based confusion matrix						
	Cluster 1 (orange)	Cluster 2 (blue)	Cluster 3 (pink)	Cluster 4 (violet)	Total	Sorting accuracy (%)
Africa	72	20	2	0	94	76.6
Western Eurasia	5	490	35	2	532	92.1
East Asia/Pacific	1	54	211	3	269	78.4
America	0	7	3	73	83	88.0

Table S3. $\overline{F_{ST}}$ –adjusted correlations between measures of individual identifiability and population identifiability for 1000 random sets of 13 non-CODIS loci. For convenience, the correlation between $\bar{H}$ and $\bar{M}$ is copied from the top (STRUCTURE) to the bottom (DAPC) part of the table. The partial correlations adjust the correlations that appear in Figure 3 and Table S4 for $\overline{F_{ST}}$.

	Diversity measures		STRUCTURE-based measures of ancestry information at *K*=4		
	$\bar{H}$	$\bar{M}$	B	S	T
$\bar{H}$		-0.97	0.51	0.36	0.21
$\bar{M}$			-0.55	-0.39	-0.23
B				0.75	0.58
S					0.89
	Diversity measures		DAPC-based measures of ancestry information at *K*=4		
	$\bar{H}$	$\bar{M}$	B_{DAPC}	S_{DAPC}	T_{DAPC}
$\bar{H}$		-0.97	0.44	0.47	0.39
$\bar{M}$			-0.45	-0.48	-0.41
B_{DAPC}				0.91	0.74
S_{DAPC}					0.88

Table S4. Correlations between measures of individual identifiability and DAPC-based measures of population identifiability for 1000 random sets of 13 non-CODIS loci. The table reports analogous correlations to those that appear in Figure 3, and for convenience, it copies from that figure the correlations between $\bar{H}$, $\bar{M}$, and $\overline{F_{ST}}$.

	Diversity measures		Variance partition	DAPC-based measures of ancestry information at K=4		
	$\bar{H}$	$\bar{M}$	$\overline{F_{ST}}$	B_{DAPC}	S_{DAPC}	T_{DAPC}
$\bar{H}$		-0.97**	-0.14*	0.28**	0.29**	0.23**
$\bar{M}$			0.11*	-0.31**	-0.31**	-0.26**
$\overline{F_{ST}}$				0.56**	0.60**	0.58**
B_{DAPC}					0.94**	0.83**
S_{DAPC}						0.92**

*p<0.05.
**p<0.001.

Table S5. Correlations between STRUCTURE-based and DAPC-based measures of population identifiability for 1000 random sets of 13 non-CODIS loci. The table provides correlations of the values that appear in the histograms in Figures 2E-2G and S1E-S1G.

	STRUCTURE-based measures of ancestry information at *K*=4			DAPC-based measures of ancestry information at *K*=4		
	B	S	T	B_{DAPC}	S_{DAPC}	T_{DAPC}
B		0.87	0.71	0.70	0.73	0.71
S			0.91	0.67	0.69	0.67
T				0.51	0.54	0.57
B_{DAPC}					0.94	0.83
S_{DAPC}						0.92

Supplemental Experimental Procedures

Data

Samples were drawn from the Human Genome Diversity Panel (HGDP-CEPH) H1048 subset [S1]. In classifying individuals as in past studies [S2-S6], the sample set included 94 people from Sub-Saharan Africa, 155 from Europe, 206 from Central and South Asia, 171 from the Middle East, 234 from East Asia, 35 from Oceania, and 83 from the Americas.

We combined 783 previously studied microsatellites [S4] with the 13 autosomal CODIS loci; CODIS genotypes were produced by Bode Technology Group (Lorton, VA) using the Promega PowerPlex 16 HS System. Four loci appeared in both marker sets; we discarded them from the data of [S4]. Locus-wise missing data ranged from 0% to 12.2% (mean 3.6% across the 792 loci). CODIS loci had no missing data.

Note that the non-CODIS loci are comparable to the CODIS loci in being highly polymorphic, widely spaced, and generally non-genic [S7]. Even the more closely spaced among them are generally pairwise-independent, producing negligible linkage disequilibrium in unstructured populations [S8].

Heterozygosity *H*

We computed locus-specific heterozygosity using the standard unbiased estimator [S9]

$$H = \frac{2N\left(1 - \sum_{j=1}^{J} p_j^2\right)}{2N - 1}, \quad (1)$$

where N is the diploid sample size, J is the number of alleles at the locus, and p_j is the frequency of allele j. Allele frequencies were estimated excluding individuals with missing data. For locus sets, we evaluated $\overline{H}$, the mean H across loci.

Match probability *M*

We calculated the locus-specific diploid random match probability as [S10,S11]

$$M = 2\left(\sum_{j=1}^{J} p_j^2\right)^2 - \sum_{j=1}^{J} p_j^4 . \quad (2)$$

For sets of loci, we computed $\overline{M}$, the geometric mean of M. The product of M values across loci directly measures a random match probability for a panel; the geometric rather than arithmetic mean reflects the contribution of a typical locus.

F_{ST}

We computed F_{ST} as the sum of among-region and among-population-within-region variance components in a three-level hierarchical partition. This partition, performed with the estimators of [S9] at each locus, considered individuals as distributed across the seven geographic regions. $\overline{F_{ST}}$ values for locus sets were obtained by averaging estimates across loci.

STRUCTURE

We conducted unsupervised model-based clustering using STRUCTURE 2.3.4 [S12,S13], employing an admixture model with correlated allele frequencies in 20,000 steps, 10,000 of which were allocated to burn-in. We set parameters α and λ to 1.

To produce "null" datasets with no ancestry information, we permuted alleles at each of the CODIS loci independently, not preserving co-occurrences of allele pairs within individuals or co-occurrences within individuals of genotypes across loci. This procedure amounts to applying Hardy-Weinberg and linkage equilibrium, holding allele frequencies constant.

K=4 was the largest K for which, in CODIS analyses, each cluster consistently had individuals for which membership in the cluster exceeded that in other clusters. K=4 was also inferred with the ΔK [S14] ($\Delta K(4)$=3.1; next was $\Delta K(3)$=1.6) and $\Pr(K=k|X)$ [S12] statistics ($\Pr(K=4|X)\approx 1$). The latter measure was computed using the median $\Pr(X|K=k)$ from 1000 runs, employing CLUMPAK [S15]. STRUCTURE solutions were plotted using DISTRUCT [S16], considering the "median" run for the purpose of plotting as the 500th smallest value among 1000.

Clusteredness *B*

Given I individuals and K clusters, STRUCTURE estimates an $I \times K$ matrix of membership coefficients. The membership for individual i in cluster k is q_{ik}. B is calculated from a STRUCTURE solution without reference to other solutions [S4]:

$$B = \frac{1}{I}\sum_{i=1}^{I}\sqrt{\frac{K}{K-1}\sum_{k=1}^{K}(q_{ik} - 1/K)^2}. \quad (3)$$

Similarity *S*

Following [S17], the similarity S measures concordance of a target STRUCTURE solution with comparison solutions at the same K. Let Q_1, Q_2, …, Q_L be membership matrices from STRUCTURE runs on the full dataset. Using G' for the similarity measure in eq. 6 of [S17], the similarity to the comparison matrices of a target matrix R of the same dimensions is

$$S(R) = \frac{1}{L}\sum_{l=1}^{L} G'(Q_l^P, R). \quad (4)$$

The superscript P indicates that columns of Q_l have been permuted in the way that maximizes similarity between Q_l and R [S17].

Sorting accuracy *T*

The sorting accuracy T measures the extent to which STRUCTURE placed individuals in clusters together with other members of the same geographic region. For each of the 100 STRUCTURE replicates at K=4 with all 779 non-CODIS loci, the cluster with the largest membership coefficient was identified for each individual. Next, we determined which of the seven geographic regions "co-clustered," where two regions co-cluster if pluralities of individuals belonging separately to the two regions sort into the same cluster. The same co-clustering pattern occurred in all 100 replicates, producing four super-regions.

For each STRUCTURE solution estimated using the CODIS markers or 13 random loci, we constructed a confusion matrix W, where $W_{i,j}$ is the number of individuals from super-region i placed into cluster j. We associated each cluster bijectively with one of the four super-regions by permuting columns of W to maximize the number of correctly classified individuals, $\sum_{i=1}^{4} W_{i,i}$. Then

$$T = \frac{1}{4}\sum_{i=1}^{4}\frac{W_{i,i}}{\sum_{j=1}^{4} W_{i,j}}. \quad (5)$$

T is the mean frequency with which individuals in a super-region sort into the cluster associated with that super-region.

Correlations

We assessed significance of correlations involving H, M, and F_{ST} by permutation, separately permuting locus-wise values of H, M, and F_{ST} appearing in the 1000 sets of 13 random markers. After applying the permutation, we recomputed $\overline{H}$, $\overline{M}$, $\overline{F_{ST}}$, and the Pearson correlations for the 1000 sets. We examined permutation distributions with 10,000 permutations. In comparisons of two among $\overline{H}$, $\overline{M}$, and $\overline{F_{ST}}$, independently permuted values were used for both statistics. Because B, S, and T are not derived from locus-level information, for significance tests involving two among B, S, and T, we instead compared correlations to a $N(0, 0.06^2)$ distribution. This distribution was chosen conservatively by noting that permutation distributions of correlations with $\overline{H}$, $\overline{M}$, and $\overline{F_{ST}}$ were roughly normal with means near 0 and smaller standard deviations, 0.048 to 0.057.

To compute partial correlations, we regressed $\overline{H}$, $\overline{M}$, B, S, and T for each set of 13 markers on $\overline{F_{ST}}$. We then evaluated Pearson correlations of residuals from these five least-squares linear regressions.

Principal component analysis

We used principal component analysis (PCA) to produce population structure estimates in a parallel manner to that used in our STRUCTURE analysis. For PCA with multiallelic markers, we placed each distinct allelic type in its own column in the input data matrix [S18]. In this matrix, each individual was represented by a row, and the entry for a row and column was the number of copies of an allelic type present in the associated individual. Missing entries were imputed as the column mean of non-missing values [S19]. Prior to running PCA, we centered and standardized columns to have mean 0 and variance 1 [S20,S21].

To facilitate comparison with STRUCTURE, we post-processed PCA output using linear discriminant analysis (LDA), as applied for population genetics in the discriminant analysis of principal components approach (DAPC [S19]). This method converts each vector of principal components (PCs) representing an individual into a vector of membership probabilities in clusters estimated by LDA from the set of individual PC vectors. Prior to implementing LDA, we performed dimensionality reduction on PC vectors, considering only the PCs associated with the largest eigenvalues required for obtaining 90% of the variance in the initial data matrix [S22]. In parallel to our STRUCTURE analysis, we applied LDA to produce membership probabilities in K=4 clusters. We implemented our PCA and LDA analyses using the ADEGENET package in R [S23,S24].

We applied DAPC to the CODIS set, the non-CODIS set, 1000 null datasets, and the same 1000 random non-CODIS datasets used in the STRUCTURE analysis. Because the procedure is deterministic—unlike in the stochastic STRUCTURE analysis—we performed only one replicate DAPC with each marker set.

The matrix of membership probabilities estimated with DAPC, containing an estimated membership probability for each individual in each cluster, is analogous to STRUCTURE-based membership estimates. We therefore computed values of B, S, and T from the DAPC solutions in the same manner as in the computations with STRUCTURE, propagating the values through analyses parallel to those performed with STRUCTURE-based B, S, and T and employing the same pipeline. To distinguish DAPC-based measures from STRUCTURE-based measures, we denote DAPC-based measures by B_{DAPC}, S_{DAPC}, and T_{DAPC}.

PCA-STRUCTURE comparisons

PCA-based ancestry information measures produce similar patterns to those observed with STRUCTURE. First, ancestry information is notable with PCA; as in the STRUCTURE analysis, PCA-based inferences are visually similar for the CODIS loci and random non-CODIS sets (Figure S1A-S1D). Measures analogous to B, S, and T, computed using PCA in place of STRUCTURE for the CODIS loci, lie near medians of their respective distributions across random sets (Figure S1E-S1G). The value of T_{DAPC} for PCA, making assignments by applying linear discriminant analysis to PCA coordinates, is 84% (Table S2). Across super-regions, assignment accuracy is greatest for Western Eurasia, the lowest-accuracy super-region for STRUCTURE.

Correlations of PCA-based B_{DAPC}, S_{DAPC}, and T_{DAPC} with $\overline{H}$, $\overline{M}$, and $\overline{F_{ST}}$ follow the patterns observed using STRUCTURE-based B, S, and T, with positive values observed between the ancestry information in PCA-based B_{DAPC}, S_{DAPC}, and T_{DAPC} and individual identifiability assessed with $\overline{H}$ and $\overline{M}$ (Table S4). As was seen with STRUCTURE-based B, S, and T, the relationship between information about identity and PCA-based B_{DAPC}, S_{DAPC}, and T_{DAPC} becomes stronger after partial-correlation adjustment for $\overline{F_{ST}}$ (Table S3). The parallel evidence of a relationship between individual and population identifiability for STRUCTURE-based and PCA-based population structure analyses is reflected in high correlations between the STRUCTURE-based B, S, and T and the analogous PCA-based B_{DAPC}, S_{DAPC}, and T_{DAPC} (Table S5).

Ancestry studies

The main text notes that among studies at the intersection of forensic genetics and genetic ancestry, some focus on use of statistical measures of marker information content to propose marker panels suitable for ancestry inference, whereas others evaluate the ancestry information for established marker sets such as the CODIS loci. An expanded list of studies in the former class, panel-design, includes [S25-S44]. An expanded list for the latter class, evaluations, includes [S35,S41,S45-S55].

Supplemental References

[S1] Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70:841-847.

[S2] Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381-2385.

[S3] Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402-1422.

[S4] Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1:660-671.

[S5] Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998-1003.

[S6] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104.

[S7] Ghebranious N, Vaske D, Yu A, Zhao C, Marth G, Weber JL (2003) STRP screening sets for the human genome at 5 cM density. BMC Genomics 4:6.

[S8] Rosenberg NA, Calabrese PP (2004) Polyploid and multilocus extensions of the Wahlund inequality. Theor Popul Biol 66:381-391.

[S9] Weir BS (1996) *Genetic Data Analysis II*. Sunderland, MA: Sinauer.

[S10] Jacquard A (1974) *The Genetic Structure of Populations*. New York: Springer.

[S11] Chakraborty R, Jin L (1993) Determination of relatedness between individuals using DNA fingerprinting. Hum Biol 65:875-895.

[S12] Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959.

[S13] Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567-1587.

[S14] Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611-2620.

[S15] Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resources 15:1179-1191.

[S16] Rosenberg NA (2004) Distruct: a program for the graphical display of population structure. Mol Ecol Notes 4:137-138.

[S17] Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801-1806.

[S18] Cavalli-Sforza LL, Menozzi P, Piazza A (1994) *The History and Geography of Human Genes*. Princeton, NJ: Princeton University Press.

[S19] Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94.

[S20] Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nature Genet 40:646-649.

[S21] François O, Currat M, Ray N, Han E, Excoffier L, Novembre J (2010) Principal component analysis under population genetic models of range expansion and admixture. Mol Biol Evol 27:1257-1268.

[S22] Jolliffe IT (2002) *Principal Component Analysis, 2nd ed*. New York: Springer-Verlag.

[S23] Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403-1405.

[S24] Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide data. Bioinformatics 27:3070-3071.

[S25] Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, Gunturi S, Ponnuswamy V, Natarajan S, Nachimuthu PK (2003) A classifier for the SNP-based inference of ancestry. J Forensic Sci 48:771-782.

[S26] Rosenberg NA (2005) Algorithms for selecting informative marker panels for population assignment. J Comput Biol 12:1183-1201.

[S27] Yang N, Li H, Criswell L, Gregersen P, Alarcon-Riquelme M, Kittles R, Shigeta R, Silva G, Patel P, Belmont J, et al. (2005) Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. Hum Genet 118:382-392.
[S28] Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M (2006) Proportioning whole-genome single-nucleotide–polymorphism diversity for the identification of geographic population structure and genetic ancestry. Am J Hum Genet 78:680-690.
[S29] Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, Drineas P (2007) PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet 3:1672-1686.
[S30] Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Álvarez-Dios J, Calaza M, Casares de Cal M, Ballard D, Lareu, MV, et al. (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet 1:273-280.
[S31] Halder I, Shriver M, Thomas M, Fernandez JR, 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.
[S32] Kersbergen P, van Duijn K, Kloosterman AD, den Dunnen JT, Kayser M, de Knijff P (2009) Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. BMC Genet 10:69.
[S33] Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, et al. (2009) Ancestry informative marker Sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat 30:69-78.
[S34] Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, et al. (2009) An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genet 10:39.
[S35] Londin ER, Keller MA, Maista C, Smith G, Mamounas LA, Zhang R, Madore SJ, Gwinn K, Corriveau RA (2010) CoAIMs: A cost-effective panel of ancestry informative markers for determining continental origins. PLoS One 5:e13443.
[S36] Paschou P, Lewis J, Javed A, Drineas P (2010) Ancestry informative markers for fine-scale individual assignment to worldwide populations. J Med Genet 47:835-847.
[S37] Ding L, Wiener H, Abebe T, Altaye M, Go RCP, Kercsmar C, Grabowski G, Martin, LJ, Khurana Hershey GK, Chakorborty R, et al. (2011) Comparison of measures of marker informativeness for ancestry and admixture mapping. BMC Genomics 12:622.
[S38] Kidd KK, Speed WC, Pakstis AJ, Kidd JR (2011) The search for better markers for forensic ancestry inference. In 22nd International Symposium on Human Identification (National Harbor, Maryland: Promega Corporation), pp. 1-4.
[S39] Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, Via M, Hidalgo-Miranda A, Contreras AV, Figueroa LU, Raska P, et al. (2012) Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genet 8:e1002554.
[S40] Nievergelt CM, Maihofer AX, Shekhtman T, Libiger O, Wang X, Kidd KK, Kidd JR (2013) Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. Investig Genet 4:13.
[S41] Phillips C, Fernandez-Formoso L, Gelabert-Besada M, Garcia-Magariños M, Santos C, Fondevila M, Carracedo Á, Lareu MV (2013) Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing. Electrophoresis 34:1151-1162.
[S42] Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, Maiers M, Middha M, Friedlaender FR, Kidd JR (2014) Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet 10:23-32.
[S43] Phillips C, Parson W, Lundsberg B, Santos C, Freire-Aradas A, Torres M, Eduardoff M, Børsting C, Johansen P, Fondevila M, et al. (2014) Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. Forensic Sci Int Genet 11:13-25.
[S44] Phillips C, Amigo J, Carracedo Á, Lareu MV (2015) Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data. Forensic Sci Int Genet 19:100-106.
[S45] Evett IW, Pinchin R, Buffery C (1992) An investigation of the feasibility of inferring ethnic origin from DNA profiles. J Forensic Sci Soc 32:301-306.
[S46] Chakraborty R, Stivers DN, Su B, Zhong Y, 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.

[S47] Lowe AL, Urquhart A, Foreman LA, Evett IW (2001) Inferring ethnic origin by means of an STR profile. Forensic Sci Int 119:17-22.
[S48] Klintschar M, Füredi S, Egyed B, Reichenpfader B, Kleiber M (2003) Estimating the ethnic origin (EEO) of individuals using short tandem repeat loci of forensic relevance. International Congress Series 1239:53-56.
[S49] Sun G, McGarvey ST, Bayoumi R, Mulligan CJ, Barrantes R, Raskin S, Zhong Y, Akey J, Chakraborty R, Deka R (2003) Global genetic variation at nine short tandem repeat loci and implications on forensic genetics. Eur J Hum Genet 11:39-49.
[S50] Barnholtz-Sloan JS, Pfaff CL, Chakraborty R, Long JC (2005) Informativeness of the CODIS STR loci for admixture analysis. J Forensic Sci 50:1322-1326.
[S51] Graydon M, Cholette F, Ng LK (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.
[S52] Pereira L, Alshamali F, Andreassen R, Ballard R, Chantratita W, Cho NS, Coudray C, Dugoujon JM, Espinoza M, Gonzalez-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.
[S53] Phillips C, Fernandez-Formoso L, Garcia-Magarinos 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.
[S54] Silva NM, Pereira L, Poloni ES, Currat M (2012) Human neutral genetic variation and forensic STR data. PLoS One 7:e49666.
[S55] Phillips C, Gelabert-Besada M, Fernandez-Formoso L, García-Magariños M, Santos C, Fondevila M, Ballard D, Court DS, Carracedo Á, Lareu MV (2014) “New turns from old STaRs”: enhancing the capabilities of forensic short tandem repeat analysis. Electrophoresis 35:3173-3187.