Supporting Information

Edge et al. 10.1073/pnas.1619944114

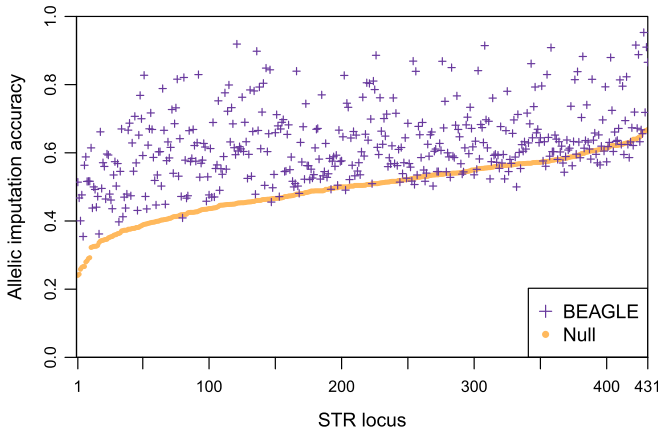

Fig. S1. Allelic imputation accuracies for 431 non-CODIS tetranucleotide STR loci. The plot considers the partition of the data represented in Fig. 1. Beagle imputation accuracy is obtained by imputing the STR genotype assigned the highest imputation probability by Beagle. Null imputation accuracy is obtained by imputing the same STR genotype for all people, irrespective of nearby SNP genotypes. Markers are sorted from left to right by null accuracy. Across all loci, the mean null accuracy is 0.497, and the mean Beagle accuracy is 0.624. Note that ref. 11 compared 432 rather than 431 non-CODIS tetranucleotides with the CODIS loci; we omitted TPO-D2S, an alias for the CODIS locus TPOX.

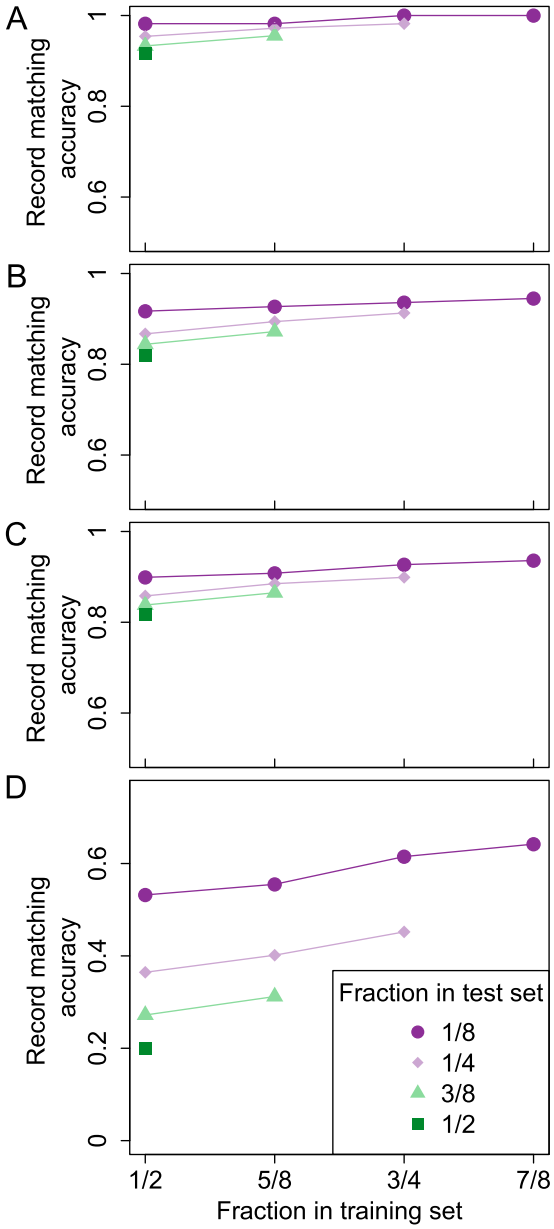

Fig. S2. The median proportion of test-set CODIS and SNP records matched correctly as a function of the sizes of the training and test sets. We divided the data into training and test sets in 1,000 ways, examining training sets of sizes 436, 545, 654, and 763—representing 50, 62.5, 75, and 87.5% of the data. For each training-set size, we used test-set sizes that were multiples of 109 (1/8 of 872), so that the sum of training-set and test-set sizes did not exceed 872. For each of 10 possible schemes for the proportions representing the training and test sets, we considered 100 random divisions of the data, using the same 100 partitions in all analyses for a given scheme. (*A*) One-to-one matching. (*B*) One-to-many matching selecting the STR profile that best matches a query SNP profile. (*C*) One-to-many matching selecting the SNP profile that best matches a query STR profile. (*D*) Needle-in-haystack matching. In *D*, the vertical axis has the same scale as in the other panels.

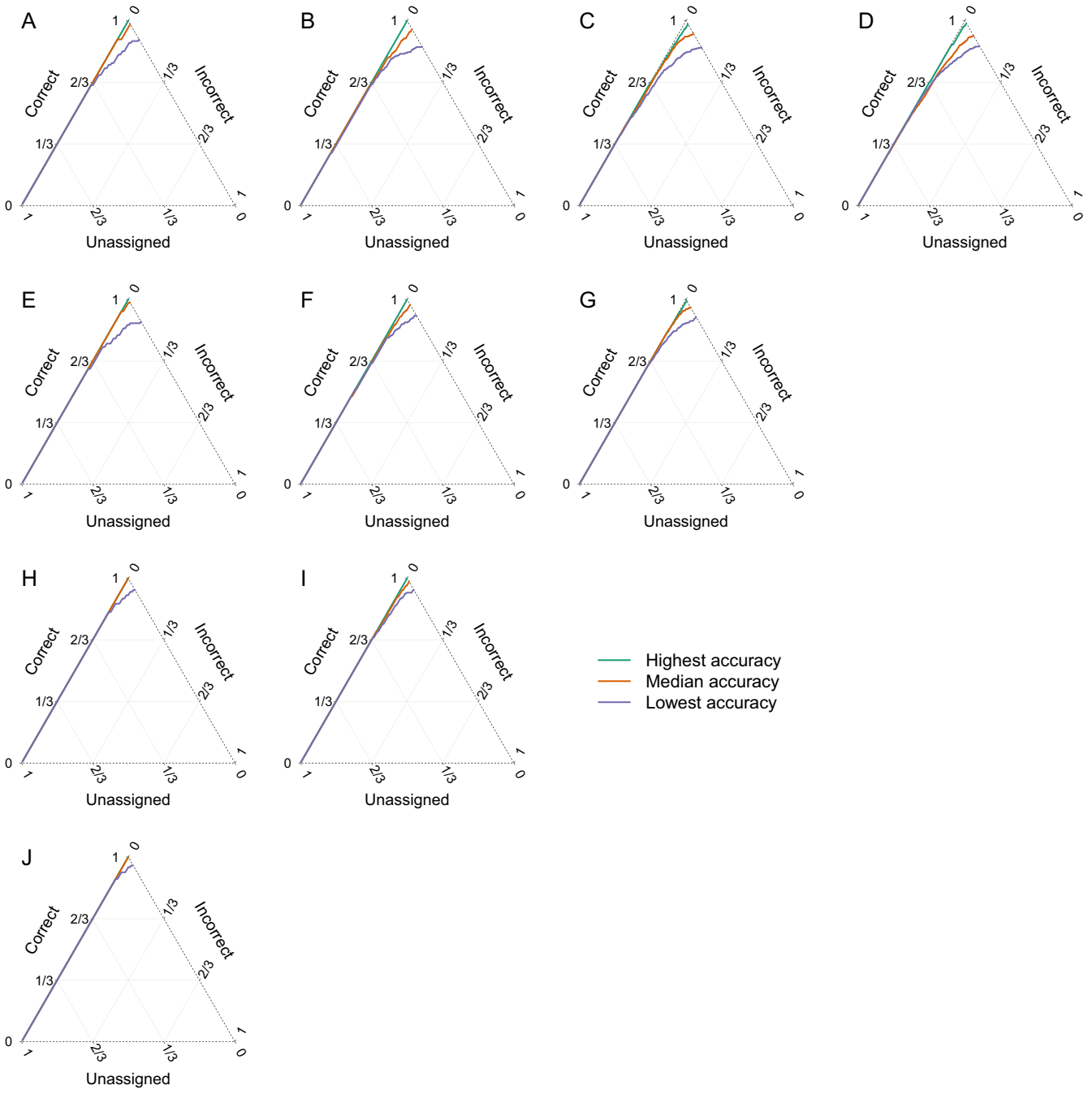

Fig. S3. Proportions of the sample unassigned, correctly assigned, and incorrectly assigned as a function of the match-score threshold under one-to-one matching using the Hungarian method. Each panel considers different proportions (training, test) of the total data (n = 872) allocated into training and test sets, with 100 allocations according to those proportions. (*A*) 1/2, 1/8. (*B*) 1/2, 1/4. (*C*) 1/2, 3/8. (*D*) 1/2, 1/2. (*E*) 5/8, 1/8. (*F*) 5/8, 1/4. (*G*) 5/8, 3/8. (*H*) 3/4, 1/8. (*I*) 3/4, 1/4. (*J*) 7/8, 1/8. The figure design follows Fig. 3.

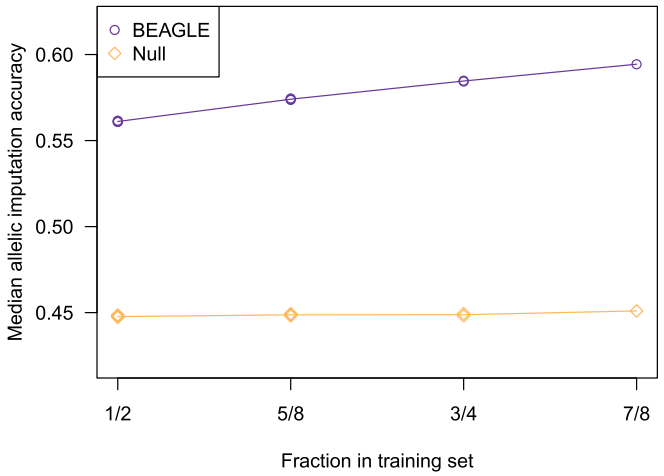

Fig. S4. The median value of the mean allelic imputation accuracy across 13 CODIS markers as a function of the size of the training set. Beagle and null imputation accuracies follow Fig. 1. The median is taken across 100 partitions into training and test sets. Imputation accuracies are plotted for all 10 schemes for the sizes of training and test sets; multiple test-set sizes produce similar values at a fixed training-set size, and they are represented by overlapping plotted points. The lines connect the median values for the test-set sizes at given training-set sizes.

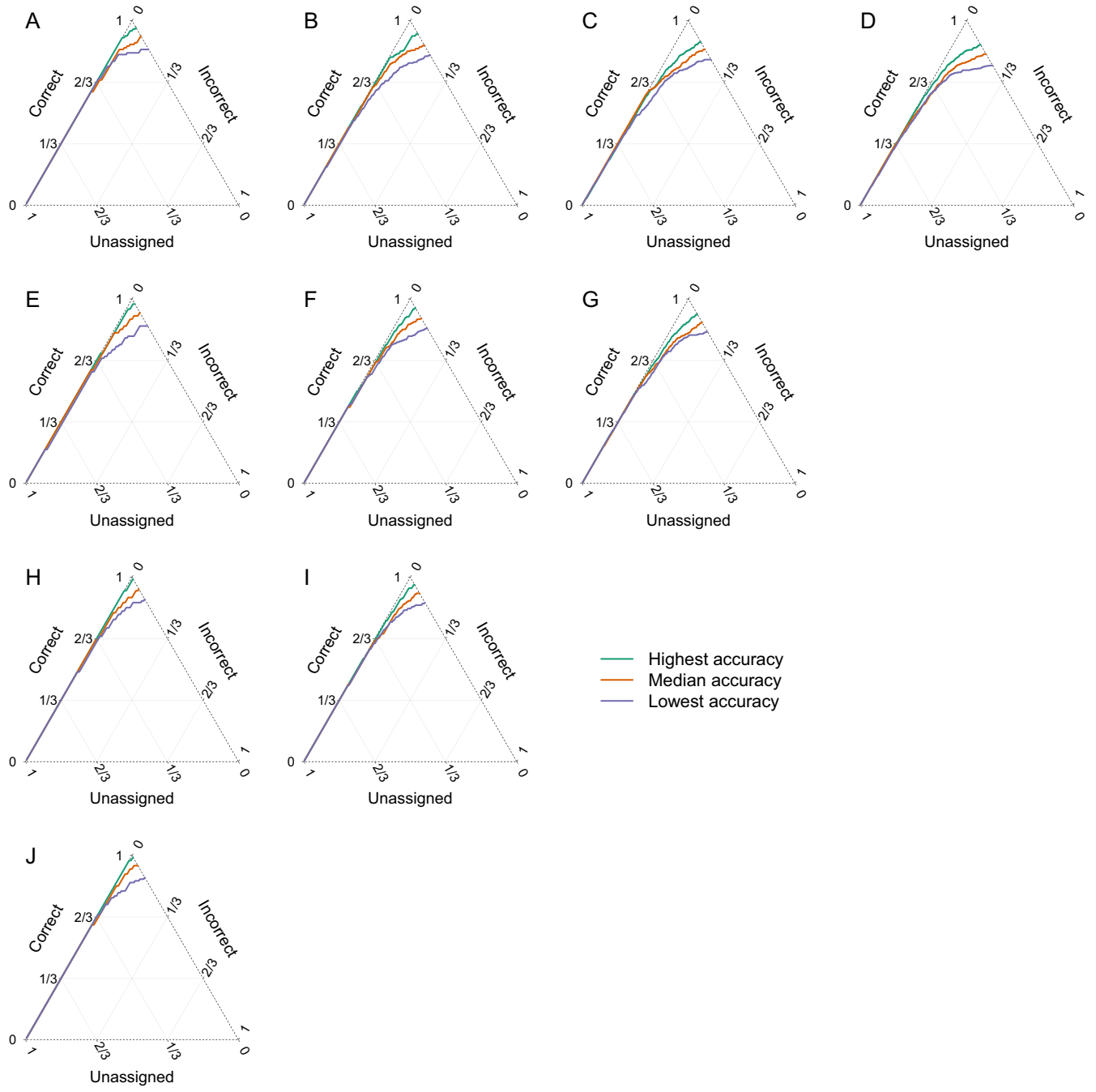

Fig. S5. Proportions of the sample unassigned, correctly assigned, and incorrectly assigned as a function of the match-score threshold under one-to-many matching that attempts to find the CODIS profile that matches a query SNP profile. Each panel considers different proportions (training, test) of the total data ($n = 872$) allocated into training and test sets, with 100 allocations according to those proportions. (*A*) 1/2, 1/8. (*B*) 1/2, 1/4. (*C*) 1/2, 3/8. (*D*) 1/2, 1/2. (*E*) 5/8, 1/8. (*F*) 5/8, 1/4. (*G*) 5/8, 3/8. (*H*) 3/4, 1/8. (*I*) 3/4, 1/4. (*J*) 7/8, 1/8. The figure design follows Fig. 3.

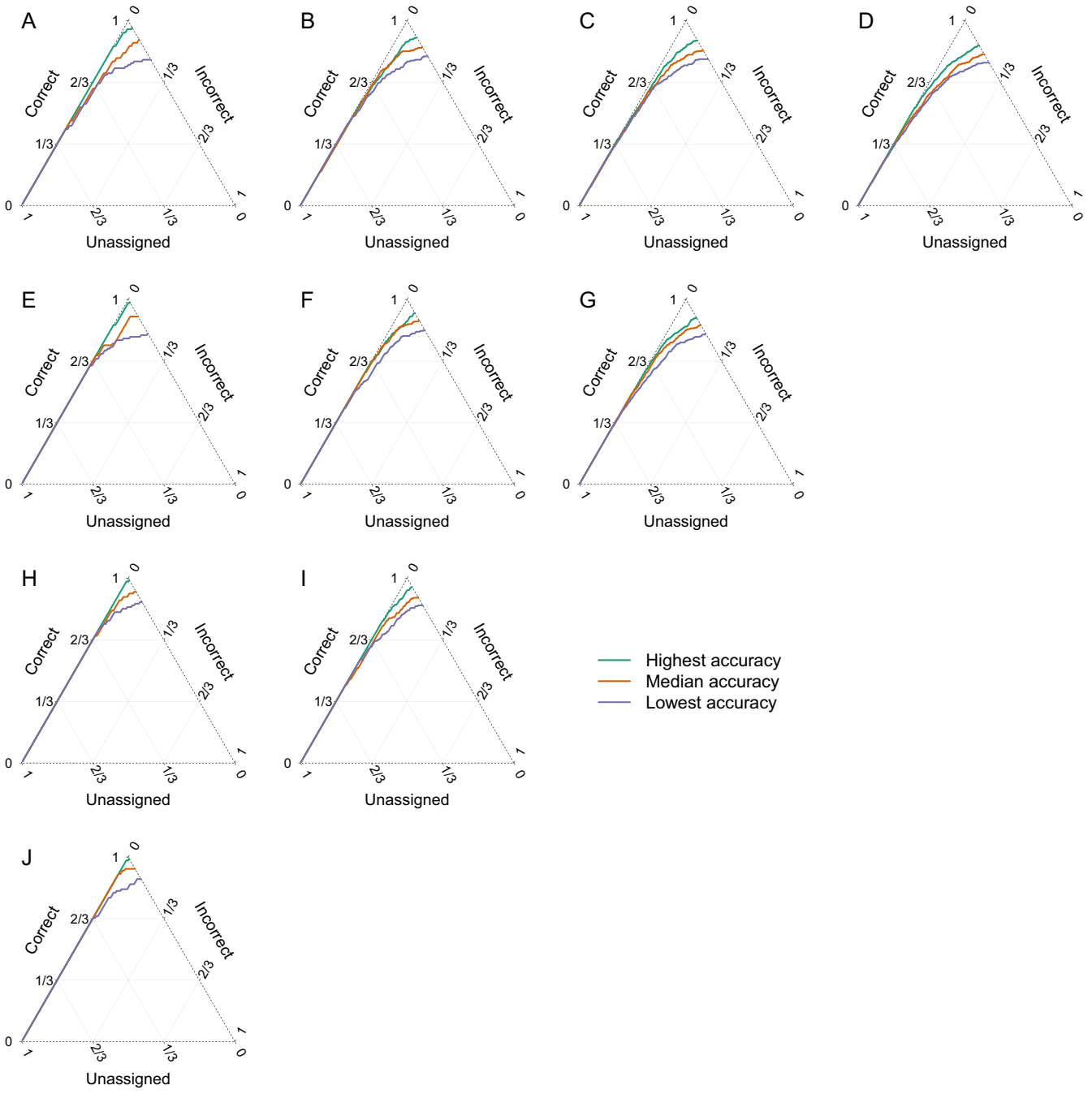

Fig. S6. Proportions of the sample unassigned, correctly assigned, and incorrectly assigned as a function of the match-score threshold under one-to-many matching that attempts to find the SNP profile that matches a query CODIS profile. Each panel considers different proportions (training, test) of the total data ($n = 872$) allocated into training and test sets, with 100 allocations according to those proportions. (*A*) 1/2, 1/8. (*B*) 1/2, 1/4. (*C*) 1/2, 3/8. (*D*) 1/2, 1/2. (*E*) 5/8, 1/8. (*F*) 5/8, 1/4. (*G*) 5/8, 3/8. (*H*) 3/4, 1/8. (*I*) 3/4, 1/4. (*J*) 7/8, 1/8. The figure design follows Fig. 3.

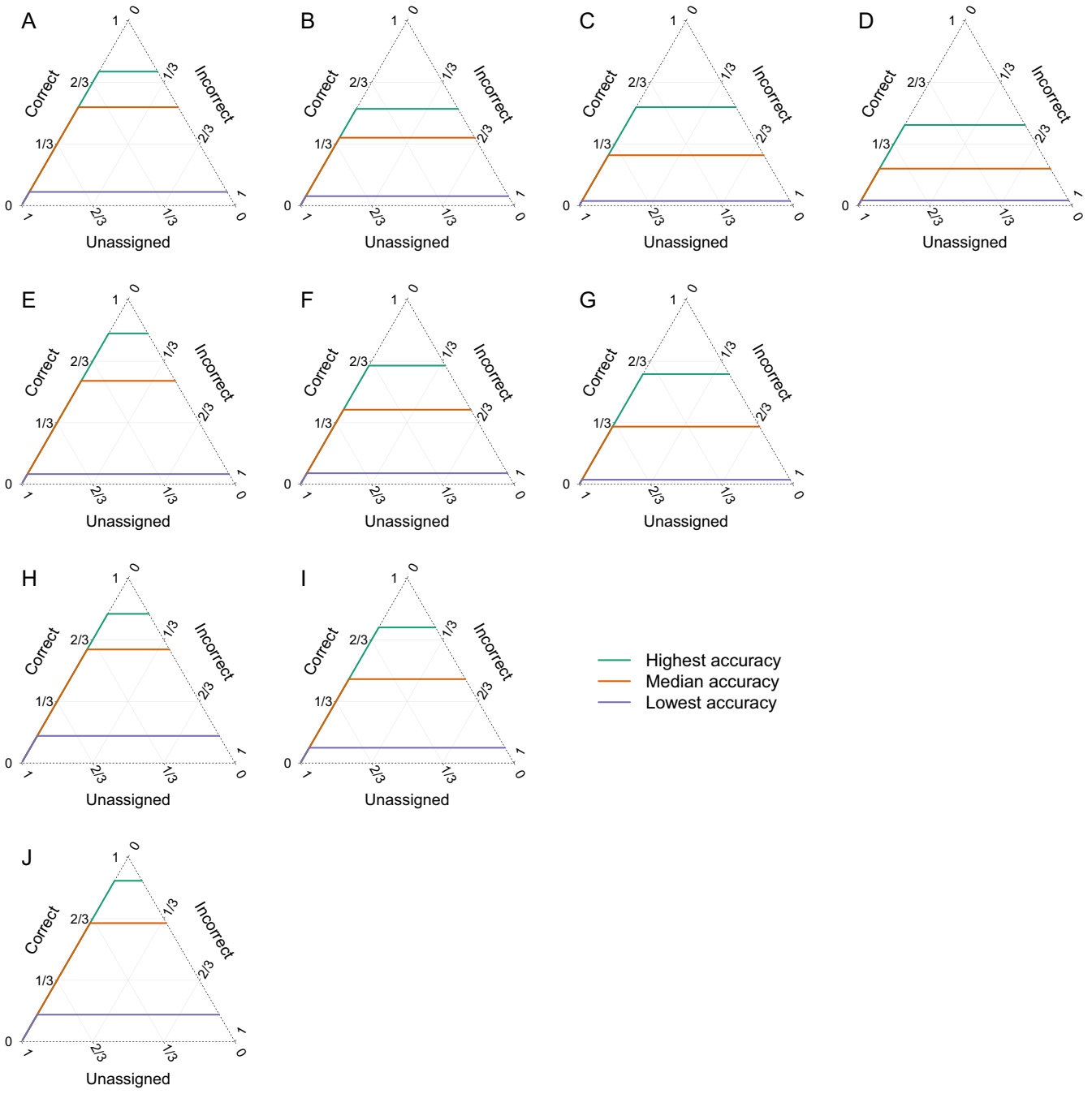

Fig. S7. Proportions of the sample unassigned, correctly assigned, and incorrectly assigned as a function of the match-score threshold under needle-in-haystack matching. Each panel considers different proportions (training, test) of the total data (n = 872) allocated into training and test sets, with 100 allocations according to those proportions. (*A*) 1/2, 1/8. (*B*) 1/2, 1/4. (*C*) 1/2, 3/8. (*D*) 1/2, 1/2. (*E*) 5/8, 1/8. (*F*) 5/8, 1/4. (*G*) 5/8, 3/8. (*H*) 3/4, 1/8. (*I*) 3/4, 1/4. (*J*) 7/8, 1/8. The figure design follows Fig. 3.

Table S1. Sample sizes by population

Continental region and population	Sample size
Sub-Saharan Africa	76
Bantu (Kenya)	11
Bantu (southern Africa)	7
Biaka Pygmy	8
Mandenka	19
Mbuti Pygmy	11
Yoruba	20
Europe	151
Adygei	17
Basque	23
French	24
Italian	12
Orcadian	15
Russian	25
Sardinian	28
Tuscan	7
Middle East	155
Bedouin	45
Druze	41
Mozabite	23
Palestinian	46
Central/South Asia	198
Balochi	24
Brahui	25
Burusho	25
Hazara	22
Kalash	23
Makrani	25
Pathan	21
Sindhi	23
Uygur	10
East Asia	227
Cambodian	10
Dai	10
Daur	9
Han	34
Han (North China)	10
Hezhen	8
Japanese	28
Lahu	8
Miao	10
Mongola	10
Naxi	7
Oroqen	9
She	10
Tu	10
Tujia	10
Xibo	9
Yakut	25
Yi	10
Oceania	26
Melanesian	9
Papuan	17
America	39
Colombian	6
Karitiana	5
Maya	14
Pima	13
Surui	1

Genotypes on 660,918 SNPs typed previously in 1,043 individuals from the Human Genome Diversity Panel (10) were submitted to quality control procedures described in ref. 39. In particular, we excluded 409 SNPs with >10% missing data among 1,043 individuals, 67 monomorphic SNPs, 696 SNPs with fewer than five alleles present in at least 1 of 52 worldwide populations, and 641 autosomal SNPs with departures from Hardy–Weinberg equilibrium. After removing these 1,813 SNPs, 659,105 SNPs remained for analysis, 642,563 of which were autosomal. We excluded relatives from the dataset of 1,043 on which SNP quality control was conducted, leaving 938 unrelated individuals (10), among whom 872 had STR data available.

Table S2. Allelic imputation accuracies and expected heterozygosities for 13 CODIS loci

Locus	Beagle imputation accuracy	Null imputation accuracy	Expected heterozygosity
D18S51	0.326	0.294	0.877
FGA	0.411	0.342	0.869
D21S11	0.509	0.381	0.850
D8S1179	0.589	0.397	0.828
TH01	0.826	0.411	0.794
VWA	0.537	0.417	0.810
D13S317	0.606	0.450	0.807
D16S539	0.610	0.461	0.787
D7S820	0.631	0.475	0.792
D5S818	0.601	0.498	0.772
D3S1358	0.592	0.523	0.742
CSF1PO	0.603	0.541	0.725
TPOX	0.849	0.585	0.691
Mean	0.591	0.444	0.796

The Beagle and null accuracies are taken from Fig. 1. Expected heterozygosities are taken from figure 1A of ref. 11. Less heterozygous loci tend to produce higher accuracies (for Beagle accuracies, Pearson $r = -0.746$, $t = -3.71$, and $p = 0.0034$; for null accuracies, $r = -0.973$, $t = -13.96$, and $p = 2.42 \times 10^{-8}$).

Table S3. Mean match scores for matching and nonmatching pairs of individuals subdivided by geographic region

		Source of STR profile						
Source of nonmatching SNP profile	Sample size	Africa	Europe	Middle East	Central/South Asia	East Asia	Oceania	America
America	15	−17.14	−19.45	−18.55	−18.22	−18.42	−13.19	−10.21
Oceania	3	−14.03	−20.83	−18.72	−18.01	−18.63	−15.83	−15.06
East Asia	59	−15.26	−18.73	−17.50	−17.92	−15.86	−16.32	−16.26
Central/South Asia	56	−14.73	−17.53	−17.04	−17.36	−18.45	−16.21	−16.78
Middle East	41	−15.49	−18.56	−18.33	−19.17	−20.67	−17.86	−19.22
Europe	30	−14.64	−15.62	−15.55	−16.14	−18.60	−14.29	−15.93
Africa	14	−19.11	−27.23	−26.04	−27.45	−28.47	−24.39	−27.00
Mean across matching profiles		0.24	8.40	8.90	9.02	9.25	1.95	13.52
Mean across matching profiles minus mean across nonmatching profile pairs from the same region		19.35	24.01	27.23	26.38	25.11	17.77	23.73

Numbers are all calculated based on the values plotted in the match-score matrix in Fig. 2*A*. Each mean is a mean of all matching or nonmatching pairs of matrix entries from a specific pair of geographic regions. Note that Oceania has a small sample size in these computations.