Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets

Michael D. Edge[a], Bridget F. B. Algee-Hewitt[a], Trevor J. Pemberton[b], Jun Z. Li[c], and Noah A. Rosenberg[a,1]

[a]Department of Biology, Stanford University, Stanford, CA 94305; [b]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, Canada R3E0J9; and [c]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109

Edited by Andrew G. Clark, Cornell University, Ithaca, NY, and approved April 10, 2017 (received for review December 6, 2016)

Combining genotypes across datasets is central in facilitating advances in genetics. Data aggregation efforts often face the challenge of record matching—the identification of dataset entries that represent the same individual. We show that records can be matched across genotype datasets that have no shared markers based on linkage disequilibrium between loci appearing in different datasets. Using two datasets for the same 872 people—one with 642,563 genome-wide SNPs and the other with 13 short tandem repeats (STRs) used in forensic applications—we find that 90–98% of forensic STR records can be connected to corresponding SNP records and vice versa. Accuracy increases to 99–100% when ∼30 STRs are used. Our method expands the potential of data aggregation, but it also suggests privacy risks intrinsic in maintenance of databases containing even small numbers of markers—including databases of forensic significance.

forensic DNA | genomic privacy | imputation | population genetics | record matching

With the increasing abundance of genetic data, the usefulness of a genetic dataset now depends in part on the possibility of productively linking it with other datasets. Thus, for example, genome-wide association study samples typed with different SNP sets are routinely combined by cross-imputation, in which markers typed only in a subset of samples are probabilistically imputed in each sample, so that all markers can be analyzed in all samples (1–3). Similarly, datasets gathered on short tandem repeat (STR) markers with different protocols can be computationally adjusted to enlarge samples for joint analysis when sets of alleles at individual markers differ between datasets (4, 5). Such efforts magnify the value of genetic datasets without requiring coordinated genotyping.

One issue that arises in combining multiple datasets is the record-matching problem: the identification of dataset entries that, although labeled differently in different datasets, represent the same underlying entity (6, 7). In a genetic context, record matching involves the identification of the same individual genome across multiple datasets when unique identifiers, such as participant names, are unavailable. This task is relatively simple when large numbers of SNPs are shared between marker sets: if records from different datasets match at enough of the shared SNPs, then they can be taken to represent the same individual.

What if no markers are shared between two genetic datasets? Can genotype records that rely on disjoint sets of markers be linked? Genetic record matching with no overlapping markers has many potential uses. Datasets could become cross-searchable even if no effort has been made to include shared markers in different marker sets. Record matching between new and old marker sets could determine whether an individual typed with a new set has appeared in earlier data, thereby facilitating deployment of new marker sets that are backward-compatible with past sets.

The presence of linkage disequilibrium (LD)—nonindependence of genotypes at distinct markers, primarily those that are proximate on the genome—can enable record matching without shared markers. As a result of LD between markers in different datasets, certain genotype pairs are more likely to co-occur, so that some potential record pairings are more likely than others. The principle applies even to different marker types not often genotyped together, such as SNPs and STRs, provided that LD exists across marker types [as is true of SNPs and STRs (8, 9)].

Relying on this principle, we devised an LD-based record-matching algorithm and evaluated its performance with nonoverlapping marker sets: one of SNPs and the other of STRs. Using 872 people from 52 populations (Table S1), we considered SNPs on a genotyping array used for population genetics and genome-wide association (10). For our STR set, we examined the Combined DNA Index System (CODIS) loci commonly used in forensic genetics (11) as well as subsets of a larger set of 432 STRs typed in the same people (12).

Our STR application enables record matching in forensic genetic contexts, where STRs are widely used. Record matching between SNP and STR panels has two additional motivations specific to forensics. First, SNP technological advances enable cost-effective genotyping of large numbers of SNPs, which could allow more precise genetic inferences than are possible with current STR panels. However, forensic testing in the United States continues to rely largely on the 13 STRs selected in the 1990s (13, 14), increasing to 20 STRs for new profiles beginning in 2017 (15), partly because millions of profiles for the 13 STRs have already been gathered in law enforcement databases (16). Reliable record matching between SNP and STR profiles could facilitate development of a backward-compatible SNP set that enables new SNP profiles to be matched against known STR profiles collected in past decades.

Second, the legality of the use of forensic genetic markers in light of US constitutional protections against unreasonable searches is based partly on a premise that these markers provide only the capacity for identification and no other information about a person (17–19). To test this premise, many investigations have examined phenotypic associations with the CODIS markers, mostly concluding that such associations are small enough to be unimportant (17, 20, 21). Record matching of CODIS and SNP data would make it possible to link a CODIS profile to a whole-genome SNP

Significance

We describe a method for identifying in distinct genetic datasets observations that represent the same person. By using correlations among genetic markers close to one another in the genome, the method can succeed even if the datasets contain no overlapping markers. We show that the method can link a dataset similar to those used in genomic studies with another dataset containing markers used for forensics. Our approach can assist in maintaining backward compatibility with databases of existing forensic genetic profiles as systems move to new marker types. At the same time, it illustrates that the privacy risks that can arise from the cross-linking of databases are inherent even for small numbers of markers.

Author contributions: M.D.E., B.F.B.A.-H., J.Z.L., and N.A.R. designed research; M.D.E., T.J.P., and N.A.R. performed research; M.D.E. and N.A.R. analyzed data; and M.D.E., B.F.B.A.-H., T.J.P., J.Z.L., and N.A.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: noahr@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1619944114/-/DCSupplemental.

profile that could enable consequential phenotypic predictions, potentially undermining the claim that the CODIS markers are phenotypically trivial. Thus, applying record matching with forensic markers is important for establishing the level of "genetic privacy" present in a forensic marker profile.

Results

We split 872 people into two disjoint subsets: a training set for learning associations between STR alleles and their surrounding SNP haplotypes and a test set for assessing record-matching accuracy. We considered 10 schemes with varying fractions of the full data allocated to training and test subsets; for each scheme, we examined 100 random assignments of people to the two subsets (100 "partitions"). We focus on a scheme with intermediate sizes for the training set (75%; $n = 654$) and the test set (25%; $n = 218$).

Imputation Accuracy. In principle, one way to link records is by genotype imputation, in which alleles of untyped loci are probabilistically predicted using genotypes at nearby typed loci (2, 3). If STR genotypes can be imputed from SNPs with perfect accuracy, then a complete set of STR genotypes can be produced from neighboring SNPs (22).

We assessed imputation accuracy at the CODIS loci using Beagle (23), imputing genotypes at each STR based on SNP genotypes within a 1-Mb window centered on the STR. First, in the training set, we used Beagle to phase the SNP genotypes together with the STR genotypes, producing a set of estimated haplotypes that included the STR alleles. Second, we imputed STR genotypes in the test set using the phased haplotypes from the training set as a reference panel.

Considering all of the CODIS markers, Beagle imputation accuracies exceed the accuracy of a null imputation method that ignores LD with nearby SNPs (Fig. 1), but they are lower than typical SNP imputation accuracies (2, 3, 24). Combining across the 13 loci and across 100 partitions into training and test sets, the null imputation method produces a mean of 11.7 of 26 alleles imputed correctly, whereas imputing with Beagle leads to a corresponding mean of 15.2. These accuracies are similar to those obtained at non-CODIS tetranucleotide STRs (Fig. S1). As has been seen previously (24), imputation accuracy is negatively correlated with measures of genetic diversity (Table S2), and the larger space of possible genotype predictions for multiallelic STRs renders their imputation accuracies lower than those observed for lower-diversity SNP loci.

Match Scores. Because imputation accuracies are not near one, records cannot be linked by simply imputing STR genotypes and identifying the record that matches the imputed genotype. It is nevertheless possible to combine imputation information across loci, producing a score that quantifies agreement between a set of STR genotypes and a set of SNP genotypes.

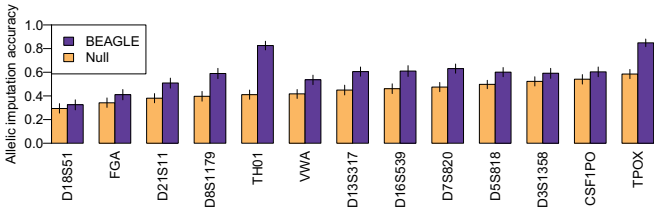

Fig. 1. Allelic imputation accuracies for 13 CODIS loci. The figure shows imputation accuracy for the partition of 872 individuals into training (75%) and test (25%) sets that yielded median (51st greatest) record-matching accuracy by the Hungarian method among 100 partitions. Beagle accuracy is obtained by imputing the STR genotype assigned the highest imputation probability by Beagle. Null accuracy is obtained by imputing the same high-frequency STR genotype in all individuals regardless of nearby SNP genotypes. Vertical lines represent 95% confidence intervals based on 10,000 bootstrap resamples of individuals from the test set. Beagle accuracies are significantly higher (Wilcoxon signed rank test, two-tailed $p < 0.05$) than null accuracies at all loci except one (D18S51; $p = 0.09$). Beagle accuracy is also higher when measuring total numbers of alleles imputed correctly in each person ($p < 2.2 \times 10^{-16}$). Beagle and null accuracies are negatively correlated with heterozygosities reported in Table S2.

We term the set of L STR genotypes carried by an individual an "STR profile," and we term the set of SNP genotypes of an individual—aggregating neighboring SNPs for all of the STR markers—a "SNP profile." R_i represents the STR profile for an individual i, with the diploid genotype at the lth locus in the profile denoted R_{il}. Similarly, S_j is the SNP profile for individual j, and S_{jl} is the set of diploid SNP genotypes in SNP profile j in the window around the lth STR.

Fellegi and Sunter (6) proposed match scores interpretable as log-likelihood ratios comparing the hypotheses that two records are drawn from the same or different people. For each possible SNP–STR profile pair, we computed the match score

$$\lambda(R_i, S_j) = \ln\left[\frac{P(R_i, S_j|M=1)}{P(R_i, S_j|M=0)}\right] = \ln\left[P(R_i|S_j, M=1)\right] - \ln[P(R_i)]. \quad [1]$$

Here, M is an indicator variable, with $M = 1$ indicating that two records are drawn from the same person (or identical twins) and with $M = 0$ indicating that they are drawn from unrelated people. The right-hand equality in Eq. **1** holds because in the ratio $P(R_i, S_j|M=1)/P(R_i, S_j|M=0) = [P(R_i|S_j, M=1)/P(R_i|S_j, M=0)]\,[P(S_j|M=1)/P(S_j|M=0)]$, the rightmost quotient is one: M affects the probability of a profile of one type, SNP or STR, only if the profile is considered jointly with a profile of the other type. The quantity $P(R_i|S_j, M=0)$ simplifies to $P(R_i|M=0)$ because R_i and S_j are independent if $M = 0$, and then to $P(R_i)$ for the same reason that the quotient $P(S_j|M=1)/P(S_j|M=0)$ reduces to 1.

STR genotypes at distinct loci are assumed to be independent in accord with the distant chromosomal locations of the CODIS loci. Consequently, the probability of observing STR profile R_i given SNP profile S_j and M is a product

$$P(R_i|S_j, M) = \prod_{l=1}^{L} P(R_{il}|S_{jl}, M). \quad [2]$$

With $M = 1$, $P(R_{il}|S_{jl}, M)$ is taken to be the imputation probability estimated by Beagle for STR genotype R_{il} given surrounding SNP genotype S_{jl}. For $M = 0$, $P(R_{il}) = P(R_{il}|S_{jl}, M=0)$ is the Hardy–Weinberg frequency of genotype R_{il} estimated using only the STR allele frequencies in the training set: the STR and SNP profiles are from different individuals, and therefore, the probability of an STR genotype is simply the genotype frequency. Thus, $\lambda(R_i, S_j)$ compares the Beagle-estimated probability of observing STR profile R_i in a person carrying SNP profile S_j with the probability of R_i in the absence of any SNP information.

For each partition into training and test sets, we computed match scores for each possible SNP–STR profile pairing in the test set (Fig. 2*A*). The method produces larger match scores when the profiles match than when they do not match ($p < 2.2 \times 10^{-16}$) (*Materials and Methods* and Fig. 2*B*). To understand the potential of the method, we used the match-score matrix to declare matches between STR and SNP profiles in four scenarios.

One-to-One Matching. We first considered the alignment of a pair of datasets on the same samples: we have n STR profiles and n SNP profiles to be matched, and it is known that each STR profile is from the same person as exactly one SNP profile. The pairing is not known and may not be trivial to determine even given an informative match-score matrix because a single STR profile might have the highest match score for multiple SNP profiles or vice versa. Given the match scores of each SNP profile with each STR profile, we conduct one-to-one matching by finding the SNP–STR profile pairing that maximizes the sum of the match scores over all paired profiles. Finding this pairing is a special case of

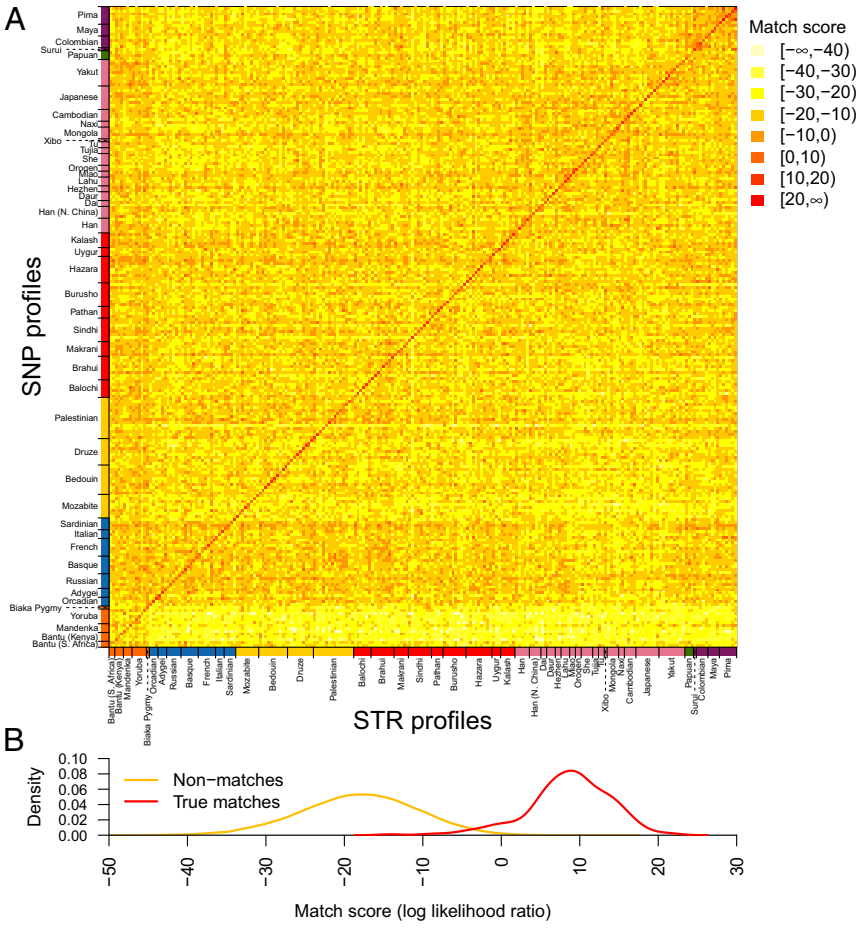

Fig. 2. Match scores of records that truly match and match scores of nonmatches. (*A*) The matrix of match scores (Eq. **1**) comparing 218 CODIS STR profiles with 218 SNP profiles for the data partition represented in Fig. 1. Each cell gives a match score for the pairing of a SNP profile with a CODIS profile. Scores pairing a given CODIS profile with each SNP profile appear in a column, and scores pairing a given SNP profile with each CODIS profile appear in a row. Darker colors represent larger values. Population memberships are colored by geographic region: Africa, orange; Europe, blue; Middle East, yellow; Central/South Asia, red; East Asia, pink; Oceania, green; Americas, purple). Of 52 populations in our dataset (Table S1), 47 appear in the test set shown. True matches are on a diagonal from the bottom left to the top right, and they tend to have higher match scores than off-diagonal nonmatches. Population structure is also visible (Table S3). For example, SNP profiles from Africans tend to have low match scores with non-Africans, and match scores of nonmatches tend to be higher when both CODIS and SNP profiles are from Native Americans. (*B*) Kernel density estimate for match scores. We applied a normal kernel with bandwidth chosen by Silverman's rule (option nrd0 in the density function in R) to the matrix entries in *A*. Nonmatches tend to have negative log-likelihood match scores, whereas true matches tend to have positive scores.

the linear sum assignment problem solvable by the "Hungarian method" (25).

Under the null hypothesis of random matching of STR and SNP profiles, the number of correct assignments among n people is distributed as the number of fixed points in a random permutation of length n. This quantity has expectation 1 and is approximately Poisson(1)-distributed (ref. 26, chap. 3, section 5). Applying the Hungarian method to the match-score matrix leads to highly accurate matches in the test set (Table 1). For 100 partitions into training and test sets, it gives a median of 214 of 218 (98.2%) correct assignments. In 18 of 100 cases, all assignments are correct. Even in the lowest-accuracy trial, 204 of 218 records are matched correctly, an extremely improbable result under random matching ($p \approx 2.8 \times 10^{-385}$).

It is possible to decrease the rate of false-positive matches by requiring that matched pairs exceed a minimum match-score threshold, leaving the remaining records unpaired. Fig. 3*A* shows the proportion of accurately assigned, inaccurately assigned, and unassigned cases as the threshold is varied for partitions with the maximum, median, and minimum numbers of correct assignments when all pairs are assigned. In the lowest-accuracy trial, 67.9% of profiles (148 of 218) can be matched accurately before a single error is made.

Figs. S2*A* and S3 display one-to-one matching results when the training-set and test-set sizes are varied. As the training-set size increases, matching accuracy increases because a larger reference haplotype set increases imputation accuracy in the test set (Fig. S4). Matching accuracy declines as the size of the test set increases because the difficulty of the matching problem increases when there are more records to be matched.

One-to-Many Matching. In some scenarios, it cannot be assumed that a one-to-one correspondence exists between profiles of one type and profiles of the other type. To examine these cases, we relax the assumption that each entry in each dataset matches exactly one entry in the other dataset. In this more challenging problem, representing the alignment of a pair of databases with partially overlapping but nonidentical samples, one STR (or SNP) profile is a "query," and we seek the SNP (or STR) profile that matches the query. Here, rather than using a matrix of match scores all at once, we consider a row or column vector, quantifying the suitability of matches of the query in one database to candidate profiles in another. This vector corresponds to either a row or a column of the matrix used for one-to-one matching; it is possible that a candidate might be identified as the best match for many queries representing different individuals.

When a SNP profile in the test set is used as a query, we choose as its match the STR profile in the test set that has the largest match score with that SNP profile. Across 100 data partitions, pairing 218 SNP profiles to their highest scoring STR match produces a median accuracy of 91.3% (199 of 218) (Table 1). The minimum accuracy is 86.2% (188 of 218), and the maximum is 95.9% (209 of 218). As was seen in one-to-one matching, matching accuracy increases with the size of the training set and declines as the size of the test set increases (Figs. S2*B* and S5).

Similarly, when an STR profile in the test set is the query, we choose as its match the SNP profile in the test set that has the largest match score with that STR profile. Pairing STR profiles to their highest-scoring SNP match produces a median accuracy of 89.9% (196 of 218) (Table 1), with a minimum of 85.3% (186 of 218) and a maximum of 95.4% (208 of 218). Accuracy increases with training-set size and declines with test-set size (Figs. S2*C* and S6).

Table 1. Record-matching accuracies for the CODIS markers

	13 CODIS markers		CODIS and 4 new markers	
Scenario	Median	Minimum, maximum	Median	Minimum, maximum
One-to-one	0.982	0.936, 1.000	1.000	0.986, 1.000
One-to-many: SNP query	0.913	0.862, 0.959	0.982	0.959, 0.995
One-to-many: STR query	0.899	0.853, 0.954	0.968	0.940, 0.991
Needle-in-haystack	0.450	0.083, 0.734	0.757	0.211, 0.899

Each entry gives the fraction of individuals matched correctly between SNP and STR profiles, with 75% of the data in the training set and 25% in the test set. The minimum, median, and maximum are taken across 100 partitions of the set of individuals into training and test sets. In columns 4 and 5, 13 CODIS STRs are augmented with 4 STRs from the 2017 CODIS addition (D2S441, D10S1248, D19S433, and D22S1045). In one-to-one matching, pairings are assigned assuming that each CODIS entry has a SNP counterpart and vice versa. In one-to-many matching, we record the proportion of times that the highest match score for a query profile (SNP or STR) arises from the profile that truly matches. In needle-in-haystack matching, we count the proportion of true matches with match scores exceeding the largest score among nonmatches.

As in the one-to-one matching case, it is possible to achieve higher confidence that proposed pairings are correct if some true matches can be missed. Fig. 3 *B* and *C* shows the proportions of accurately assigned, inaccurately assigned, and unassigned cases as the match-score threshold is varied. In the lowest-accuracy cases, when a SNP profile is the query, 41.3% of profiles (90 of 218) are assigned accurately before a single erroneous match is made, and when an STR profile is the query, 56.0% of profiles (122 of 218) are assigned accurately before an error is made.

Needle-in-Haystack Matching. An even more difficult problem arises when only one among all possible SNP–STR profile pairings is a true match. This scenario represents the case in which a database query to locate a match is performed only for one profile. Perfect accuracy is achieved if the match-score distributions for matches and nonmatches do not overlap. To evaluate accuracy in this scenario, we recorded the fraction of true matches with match scores exceeding the largest score among nonmatching profiles.

Across partitions into training and test sets, the median percentage of true matches with match scores exceeding the maximum match score among nonmatches is 45.0% (98 of 218) (Table 1). The minimum is 8.3% (18 of 218), and the maximum is 73.4% (160 of 218). As in the other cases, matching accuracy increases with increasing training-set size and declines with increasing test-set size (Figs. S2*D* and S7).

Adding STRs. Record matching proceeds by accumulating information about the agreement of a pair of records across loci. Thus, adding more loci is expected to increase record-matching accuracy. To evaluate the effect of the number of loci, we repeated our matching procedures in non-CODIS STR sets of varying size (Fig. 4). For each procedure, accuracy increases as more loci are considered. Median accuracy increases to 97.2% (212 of 218) in 20-locus panels for one-to-many matching procedures and 71.6% (156 of 218) for needle-in-haystack matching. Almost all 30-locus panels (99 of 100) produce perfect matching accuracy in one-to-one matching, and most produce accuracy above 99% in one-to-many matching (84 of 100 with query SNP profiles; 51 of 100 with query STR profiles). With 50-STR panels, in the median trial, 96.8% of true matches (211 of 218) have match scores exceeding the highest match score among unmatched pairs.

Discussion

We have shown that genetic records can potentially be linked even if they contain nonoverlapping sets of markers. Despite the small number of markers in one of our datasets—13 STRs—multimarker profiles can be matched to genome-wide SNP profiles with median accuracies in excess of 90% (Table 1). Furthermore, record-matching accuracy increases with the number of markers, and with only a few dozen markers, accuracy nears 100% (Fig. 4).

The fact that such high match accuracies are achievable despite relatively low imputation accuracies at individual STRs is perhaps surprising. In domesticated cattle, McClure et al. (22) observed that SNP haplotypes are highly predictive of the allele of an STR lying within the haplotype and that STR profiles could, therefore, be imputed with high accuracy. In humans, however, LD between STRs and surrounding SNPs is weaker, with many distinct STR alleles appearing on the same SNP haplotype in a population and with multiple SNP haplotypes possessing the same STR allele (8, 9). Nevertheless, because some LD does exist and because SNP-based imputation accuracies exceed null imputation accuracies, LD information can be accumulated across markers to permit record matching—not unlike the manner in which small differences in allele frequency across populations can be accumulated across markers to enable inference of ancestry (27, 28).

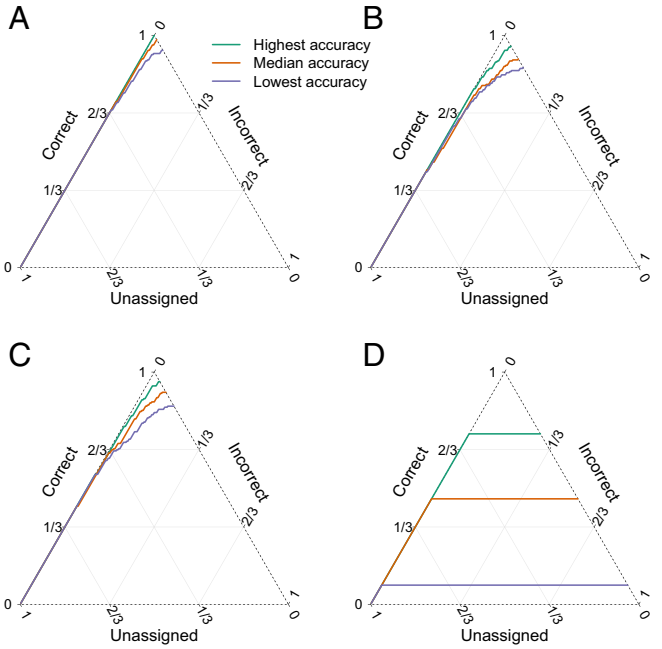

Fig. 3. The proportions of profiles unassigned, correctly assigned, and incorrectly assigned as the match-score threshold is varied. When the threshold is large, all profiles are unassigned (lower left vertex). Gradually lowering the threshold leads to assignment of all profiles, tracing a curve to the right edge. Of 100 partitions into training and test sets, the figure plots trials with maximum, median, and minimum accuracies when all possible profiles are paired. (*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 counting the proportion of true matches with match score that exceeds the maximal match score among nonmatches. In *D*, after the match-score threshold is lower than the largest match score among nonmatches, all pairings are marked incorrect.

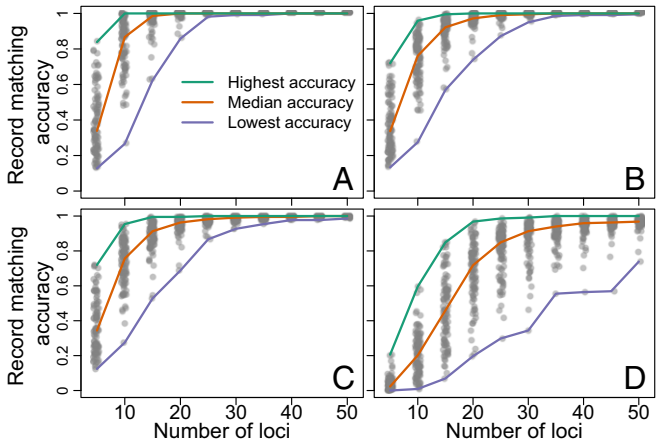

Fig. 4. Record-matching accuracy as a function of number of STRs. For each number of loci, 100 random locus sets are analyzed for the data partition in Fig. 1; results are shown horizontally jittered. (*A*) One-to-one matching. (*B*) One-to-many matching selecting the STR profile that best matches a query SNP profile. (*C*) One-to-many matching selecting the SNP profile that best matches a query STR profile. (*D*) Needle-in-haystack matching.

Imputation accuracy is negatively correlated with genetic diversity measures in the samples in which genotypes are imputed, and thus, it is often greater in low-diversity populations than in high-diversity populations (24, 29). Here, this effect accumulates across loci, and mean match scores are, therefore, greater for matches in low-heterozygosity Native Americans than in high-heterozygosity Africans (Table S3). The increase in match scores applies also to nonmatching pairs in low-diversity populations, however, because the greater similarity among individuals in such populations inflates the mean match score for these pairs (Fig. 2). Because match scores are inflated in low-diversity populations for both matches and nonmatches, differences between mean match scores for matches and nonmatches are similar across groups, so that the potential to separate matches from nonmatches need not be greatest in low-diversity groups (Table S3).

Our study adds to a growing list of record-matching scenarios in genetics. Vohr et al. (30) used SNPs to calculate likelihood ratio-based match scores, relying on a phased reference panel to assist in assigning low-coverage sequence reads to the same individual. Conceptually similar methods have also been used to identify mismatches between genotypes and expression data (31, 32). Our method augments this work by using the record-matching framework and enabling matches between markers of different types, even when imputation accuracy is far below one.

The fact that CODIS imputation accuracies are relatively low (Fig. 1) suggests that, from a SNP profile, it is unlikely that the full CODIS STR profile of an individual can be reliably imputed. However, if that profile already appears in a CODIS STR database, then a match between the SNP profile and its associated STR profile might be possible to identify. The feasibility of record matching suggests a form of backward compatibility with the CODIS STR database, in which SNP rather than STR profiles would be collected on future samples and queried against both new SNP databases and existing STR data. STRs could then be typed on such samples to validate an STR–STR match only if a strong SNP–STR match is suggested. Although matching accuracy was imperfect, that stringent match-score thresholds permitted many accurate matches before the first error was made suggests that backward compatibility by record matching and exclusion of unlikely pairings may be achievable for a substantial fraction of samples.

The utility of record matching in advancing forensic genetics depends on the degree to which it scales to typical forensic dataset sizes, numbering in the thousands or millions of profiles. We find that record-matching accuracy increases as larger training sets become available but decreases with the test-set size. An increase in the number of CODIS loci from 13 to 20 (15) increases the potential of record matching; accuracies were considerably higher in our 20-STR scenarios than in the 13-STR CODIS examples; 4 loci in the 2017 update to the CODIS markers are available in the data of ref. 12, and combining them with the 13-locus set indeed provides a substantial accuracy increase (Table 1). We expect that accuracy could potentially be increased further if STR genotypes were produced by a procedure that obtains the full DNA sequence of the STRs rather than the length of the repeated unit, thereby subdividing repeat-length alleles into finer allelic classes (33, 34); with this approach, the level of resolution at which alleles are classified as distinct could be tuned to a level that maximizes the record-matching accuracy.

Even with quite large test sets, it is plausible that some profiles could be paired with high confidence. If the match scores $\lambda(R_i, S_j)$ from Eq. **1** are viewed as likelihood ratios, then by Bayes' rule,

$$O\left[(M=1):(M=0)\middle|\lambda(R_i,S_j)\right]=O\left[(M=1):(M=0)\right]\exp\left[\lambda(R_i,S_j)\right], \quad [3]$$

where O indicates odds. Eq. **3** can be used to determine the match score necessary to obtain a specified posterior odds and thus, posterior probability of a match given the prior odds of a match. To obtain posterior odds of a match >10 (i.e., posterior probability >10/11), with prior odds 4.3×10^{-9} (1 in 235 million, the approximate adult population of the United States at the 2010 census), a match score must exceed $\ln[10/(4.3\times10^{-9})] \approx 21.6$. When we apply our method using 13 CODIS markers in the partition that leads to median one-to-one matching accuracy with 654 people in the training set, 1 of 218 true match scores in the test set exceeds this threshold, equaling 21.8. When we include four new markers of the updated CODIS panel, however, 17 of 218 (7.8%) true match scores exceed the threshold (maximum score of 31.9). Thus, with the expanded CODIS set, it is possible that a nontrivial proportion of CODIS and SNP genotypes could be matched with high confidence; furthermore, because training-set size increases matching accuracy (Fig. S2), this proportion could increase with an increase in training-set size. Our computations with Eq. **3** motivate detailed empirical evaluation in larger datasets.

The potential for record matching of SNP and CODIS STR profiles, especially with augmented CODIS panels, uncovers new risks to privacy. Some record pairings have match scores so large that they are improbable in the absence of a true match. Thus, authorized or unauthorized analysts equipped with two datasets, one with SNP genotypes and another with CODIS genotypes, could possibly identify some pairs of records that are likely to represent the same person. For people with records linked in this way, CODIS genotypes would reveal genomic SNP genotypes that could, in turn, reveal much more information than the CODIS genotypes themselves—such as precise ancestry estimates, health and identification information that accompanies SNP records, and predictions for genetically influenced phenotypes. In this sense, contrary to the view that CODIS genotypes expose no phenotypes (17, 20, 21), a CODIS profile on a person together with a SNP database—if the person is in the database—in principle may contain all of the phenotypic information that can be reliably predicted from the SNP record. Conversely, participants in biomedical research or personal genomics who have consented to share their SNP genotypes may be subject to a previously unappreciated risk: identification in a forensic STR database.

As in other situations in which data aggregation can unexpectedly reveal genetic information at the individual level (35–37), it is desirable to reevaluate the privacy of forensic STR profiles in light of the widespread availability of diverse SNP profiles to researchers and the public. Because our record-matching methods can potentially be extended beyond the detection of identical people to the detection of relatives—matching a SNP profile of an individual to an STR profile of a relative—we expect that privacy considerations will extend to this scenario as well.

Materials and Methods

Data. From the Human Genome Diversity Panel, we examined previously reported genotypes on 872 samples—the intersection of 938 unrelated samples with SNP genotypes reported (10), a subset of 1,048 samples with STR genotypes reported (12), and 978 samples with CODIS genotypes reported (11). Population information appears in Table S1. Non-CODIS STRs included 431 tetranucleotides and trinucleotide D22S1045, which is in the 2017 CODIS update (15). We obtained non-CODIS STR positions by querying University of California-Santa Cruz (UCSC) Genome Browser's BLAT (38) using the locus RefSeq sequence (table S1 of ref. 12) and build hg18; for CODIS loci, UCSC Genome Browser queries used the locus name.

Phasing and Imputation. In Beagle 4.1 (23), we set the number of iterations to 10. When phasing, we used defaults for all other parameters: maxlr = 50,000, lowmem = false, window = 50,000, overlap = 3,000, impute = true, cluster = 0.005, ne = 1 million, err = 0.0001, seed = -99,999, and modelscale = 0.8. When imputing STRs, we set gprobs = true and maxlr = 1 million, and we used a linkage map based on GRCh36 coordinates.

For each STR, windows extended 500 kb in both directions from the STR midpoint (GRCh36 coordinates corresponding to UCSC hg18). For non-CODIS loci, the number of SNPs in such windows ranged from 80 to 547, with a median of 262. For CODIS loci, the range was 164–655, with a median of 272.

Imputation Accuracy. Imputation accuracy was assessed as the number of accurately imputed alleles (24). Null imputations were made disregarding the neighboring SNP genotypes by imputing the genotype that, under Hardy–Weinberg equilibrium with the allele frequencies estimated in the training set, is predicted to lead to the highest accuracy. Denoting the alleles at a locus 1, ..., K in decreasing order of their frequencies p_1, ..., p_K, the most frequent homozygote was imputed if $p_1^2+p_2^2>2p_2$; otherwise, the most frequent heterozygote was imputed.

To verify that this condition for "null" imputations produces the highest accuracy, note that, if the most frequent homozygote is always imputed, then for each individual homozygous for allele 1 (frequency p_1^2), two alleles are imputed correctly, and for each heterozygote with allele 1 (frequency $2p_1\sum_{k=2}^{K}p_k$), one allele is imputed correctly. The expected number of correctly imputed alleles per individual is $2p_1^2+2p_1\sum_{k=2}^{K}p_k=2p_1$.

If instead, the most frequent heterozygote is imputed, then the number of alleles imputed correctly is two for individuals heterozygous for the two most frequent alleles (frequency $2p_1p_2$), one for homozygotes for allele 1 or 2 (frequency $p_1^2+p_2^2$), and one for heterozygotes with one of the two most frequent alleles and another allele that is not one of the two most frequent (frequency $2p_1\sum_{k=3}^{K}p_k+2p_2\sum_{k=3}^{K}p_k$). The expected number of correct alleles imputed per individual is, therefore, $4p_1p_2+p_1^2+p_2^2+2p_1\sum_{k=3}^{K}p_k+2p_2\sum_{k=3}^{K}p_k$ or $2p_1+2p_2-p_1^2-p_2^2$. Thus, imputing the homozygote produces a higher expected number of correctly imputed alleles than imputing the heterozygote if $2p_1>2p_1+2p_2-p_1^2-p_2^2$ or equivalently, $p_1^2+p_2^2>2p_2$.

Match Scores. To avoid likelihoods of zero in match-score computations, any diploid genotype assigned probability zero by Beagle was given probability 0.0005, one-half the lowest permissible nonzero probability in the Beagle version that we used. Probabilities were then renormalized to sum to one. Probabilities for genotypes including alleles unobserved in the training set or missing were set equal under all hypotheses about M so as not to affect match scores.

Testing Match Scores of True Matches Against Nonmatches. To account for dependencies among values in the same column or row of the match-score matrix, we fit a linear mixed model with crossed random effects using entries from the matrix in Fig. 2*A*: $Y_{ij}=\beta_0+\beta_{0i}+\beta_{0j}+\beta_1X_{ij}+\varepsilon_{ij}$. Here, Y_{ij} is the match score for the pairing of the ith STR and jth SNP profiles, β_0 is a global intercept, β_{0i} is a random intercept for scores involving the ith STR profile, β_{0j} is a corresponding intercept for the jth SNP profile, the indicator variable X_{ij} is one if Y_{ij} represents a true match ($i=j$), and ε_{ij} is a normal disturbance with expectation zero and constant variance. We used R package lmer, computing p values by Satterthwaite approximation with package lmerTest. This model was strongly preferred over models that excluded random effects for either STR or SNP profiles (Akaike Information Criterion and Bayesian Information Criterion differences >3,000). The estimate for β_1, the difference between scores for matches and nonmatches, was significant [$\hat{\beta}_1=26.8$, $\mathrm{SE}=0.43$, $t(47{,}088)=62.7$, $p<2.2\times10^{-16}$].

ACKNOWLEDGMENTS. We thank H. Greely, E. Halperin, and N. Rudin for discussions and B. Browning for assistance with Beagle. This work was supported by NIH Grant R01HG005855 and National Institute of Justice Grant 2014-DN-BX-K015.

1. de Bakker PI, et al. (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 17:R122–R128.
2. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. *Annu Rev Genomics Hum Genet* 10:387–406.
3. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11:499–511.
4. Presson AP, Sobel E, Lange K, Papp JC (2006) Merging microsatellite data. *J Comput Biol* 13:1131–1147.
5. Pemberton TJ, DeGiorgio M, Rosenberg NA (2013) Population structure in a comprehensive genomic data set on human microsatellite variation. *G3 (Bethesda)* 3:891–907.
6. Fellegi IP, Sunter AB (1969) A theory for record linkage. *J Am Stat Assoc* 64:1183–1210.
7. Winkler WE (2014) Matching and record linkage. *Wiley Interdiscip Rev Comput Stat* 6: 313–325.
8. Payseur BA, Place M, Weber JL (2008) Linkage disequilibrium between STRPs and SNPs across the human genome. *Am J Hum Genet* 82:1039–1050.
9. Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y; 1000 Genomes Project Consortium (2014) The landscape of human STR variation. *Genome Res* 24:1894–1904.
10. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
11. Algee-Hewitt BFB, Edge MD, Kim J, Li JZ, Rosenberg NA (2016) Individual identifiability predicts population identifiability in forensic microsatellite markers. *Curr Biol* 26:935–942.
12. Pemberton TJ, Sandefur CI, Jakobsson M, Rosenberg NA (2009) Sequence determinants of human microsatellite variability. *BMC Genomics* 10:612.
13. Budowle B, Shea B, Niezgoda S, Chakraborty R (2001) CODIS STR loci data from 41 sample populations. *J Forensic Sci* 46:453–489.
14. Butler JM (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. *J Forensic Sci* 51:253–265.
15. Hares DR (2015) Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci Int Genet* 17:33–34.
16. Federal Bureau of Investigation (2016) *CODIS—NDIS Statistics*. Available at https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics. Accessed June 14, 2016.
17. Katsanis SH, Wagner JK (2013) Characterization of the standard and recommended CODIS markers. *J Forensic Sci* 58:S169–S172.
18. Greely HT, Kaye DH (2013) A brief of genetics, genomics and forensic science researchers in Maryland v. King. *Jurimetrics* 54:43–64.
19. Maryland v. King, 133 S. Ct. 1958 (2013).
20. Graydon M, Cholette F, 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.
21. Lohmueller KE (2010) Graydon et al. provide no new evidence that forensic STR loci are functional. *Forensic Sci Int Genet* 4:273–274.
22. McClure M, Sonstegard T, Wiggans G, Van Tassell CP (2012) Imputation of microsatellite alleles from dense SNP genotypes for parental verification. *Front Genet* 3:140.
23. Browning BL, Browning SR (2016) Genotype imputation with millions of reference samples. *Am J Hum Genet* 98:116–126.
24. Huang L, et al. (2009) Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* 84:235–250.
25. Kuhn HW (1955) The Hungarian method for the assignment problem. *Naval Res Logistics Q* 2:83–97.
26. Riordan J (1958) *An Introduction to Combinatorial Analysis* (Wiley, New York).
27. Edwards AWF (2003) Human genetic diversity: Lewontin's fallacy. *BioEssays* 25:798–801.
28. Edge MD, Rosenberg NA (2015) Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. *Stud Hist Philos Biol Biomed Sci* 52:32–45.
29. Huang L, et al. (2011) Haplotype variation and genotype imputation in African populations. *Genet Epidemiol* 35:766–780.
30. Vohr SH, Buen Abad Najar CF, Shapiro B, Green RE (2015) A method for positive forensic identification of samples from extremely low-coverage sequence data. *BMC Genomics* 16:1034.
31. Westra HJ, et al. (2011) MixupMapper: Correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* 27:2104–2111.
32. Broman KW, et al. (2015) Identification and correction of sample mix-ups in expression genetic data: A case study. *G3 (Bethesda)* 5:2177–2186.
33. Warshauer DH, et al. (2013) STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Sci Int Genet* 7:409–417.
34. Warshauer DH, King JL, Budowle B (2015) STRait Razor v2.0: The improved STR Allele Identification Tool–Razor. *Forensic Sci Int Genet* 14:182–186.
35. Homer N, et al. (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 4:e1000167.
36. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y (2013) Identifying personal genomes by surname inference. *Science* 339:321–324.
37. Erlich Y, Narayanan A (2014) Routes for breaching and protecting genetic privacy. *Nat Rev Genet* 15:409–421.
38. Kent WJ (2002) BLAT–the BLAST-like alignment tool. *Genome Res* 12:656–664.
39. Jakobsson M, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998–1003.