PNAS RESEARCH ARTICLE | GENETICS OPEN ACCESS

Associations between forensic loci and expression levels of neighboring genes may compromise medical privacy

Mayra M. Bañuelos[a,b,d], Yuómi Jhony A. Zavaleta[c], Alennie Roldan[c], Rochelle-Jan Reyes[c], Miguel Guardado[a,1], Berenice Chavez Rojas[c,2], Thet Nyein[a,2], Ana Rodriguez Vega[c,2], Maribel Santos[c,2], Emilia Huerta-Sanchez[b,d], and Rori V. Rohlfs[c,3]

Edited by Marcus Feldman, Stanford University, Stanford, CA; received November 18, 2021; accepted August 29, 2022

A set of 20 short tandem repeats (STRs) is used by the US criminal justice system to identify suspects and to maintain a database of genetic profiles for individuals who have been previously convicted or arrested. Some of these STRs were identified in the 1990s, with a preference for markers in putative gene deserts to avoid forensic profiles revealing protected medical information. We revisit that assumption, investigating whether forensic genetic profiles reveal information about gene-expression variation or potential medical information. We find six significant correlations (false discovery rate = 0.23) between the forensic STRs and the expression levels of neighboring genes in lymphoblastoid cell lines. We explore possible mechanisms for these associations, showing evidence compatible with forensic STRs causing expression variation or being in linkage disequilibrium with a causal locus in three cases and weaker or potentially spurious associations in the other three cases. Together, these results suggest that forensic genetic loci may reveal expression levels and, perhaps, medical information.

forensic genetics | gene expression | data privacy | STRs | CODIS loci

Significance

A central assumption in forensic genetics is that the loci used for identification do not reveal any medical information. This assumption is crucial for the legal and ethical frameworks guiding how forensic samples are seized, databased, and accessed. Despite the importance of safeguarding the medical privacy of tens of millions of people with forensic genetic profiles, few studies have addressed the question since the forensic loci were established in the 1990s. Here, we show evidence calling to question this central assumption. We show significant correlations between the genotype of forensic markers and expression of neighboring genes, going on to find evidence for a molecular mechanism. These results have substantive legal and ethical implications for the treatment of forensic genetic profiles.

Forensic genetic identification in the United States is typically performed by using genotype data from 20 short tandem repeats (STRs), known as the Combined DNA Index System (CODIS) core loci. Because these markers are highly polymorphic, even just 20 loci provide an immense amount of identifying information regarding a specific individual (1). Thirteen of these CODIS core loci were established by the Federal Bureau of Investigation in 1998. These loci were selected for efficient PCR multiplexing, while maximizing identifying information and minimizing ancestry-based population differences and medically relevant information (2). In 2017, seven additional STRs were added to the CODIS core loci, selected for similar criteria, particularly no known associations with medical conditions (3).

It is important from a legal standpoint that CODIS genotypes do not reveal medical information. Laws authorizing the compulsory collection of DNA from certain persons may come into conflict with state privacy statutes or the US Constitution if medical information is embedded (4). In fact, hundreds of court cases rely on the premise that the CODIS variants are uninformative, often citing this quote relating to the DNA Analysis Backlog Elimination Act of 2000, which states that the CODIS loci "were purposely selected because they are not associated with any known physical or medical characteristics" (Letter from Robert Raben, Assistant Attorney General, to Judiciary Committee Chairman Henry Hyde) (5). Yet, some of the CODIS loci, particularly those selected before the human genome was sequenced, are very close to genes. In fact, 11 of the CODIS loci are intronic (6).

Any trait information conveyed by CODIS genotypes would raise questions regarding the medical privacy of individuals whose CODIS profiles are compelled by the government, as well as their genetic relatives. These include over 19,350,445 people arrested or convicted whose CODIS profiles are held in the US national database (7), a group that overrepresents people of color, especially Black populations (8). The historical and current treatment of arrested and convicted individuals is rife with rights unjustly curtailed, raising even more concern about a potentially lax approach to medical privacy for this population (9–11). While the access-limited, federally regulated national database is vast, it does not include all CODIS profiles held at state and local levels. For instance, in the "spit and acquit" practice by the Orange County (CA) District Attorney's Office, certain misdemeanor defendants can be offered a dismissal in exchange for a DNA sample, resulting in a local database of over 150,000 individuals (12). Other expansive local practices include using samples collected from nonsuspects for criminal investigation. A recent example is a sexual assault survivor's DNA profile later being used to connect her to a property crime (13, 14). Given the scope of individuals with CODIS profiles stored

Author contributions: R.V.R. designed research; M.M.B., Y.J.A.Z., A.R., R.-J.R., E.H.-S., and R.V.R. performed research; M.M.B., Y.J.A.Z., A.R., R.-J.R., M.G., B.C.R., T.N., A.R.V., M.S., E.H.-S., and R.V.R. analyzed data; and M.M.B., Y.J.A.Z., A.R., R.-J.R., E.H.-S., and R.V.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] Present address: Biological and Medical Informatics Graduate Program, University of California, San Francisco, CA 94158.

[2] B.C.R., T.N., A.R.V., and M.S. contributed equally to this work.

[3] To whom correspondence may be addressed. Email: rrohlfs@sfsu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2121024119/-/DCSupplemental.

Published September 27, 2022.

and/or shared, in this study, we re-examine the assumption that CODIS genotypes have no functional or medical impact.

It has long been known that variation in STR number can alter gene function and regulation, sometimes resulting in dramatic phenotypes. A classic example is the coding STR expansion in the *HD* gene, which causes increasingly severe Huntington's disease (15). Noncoding STRs have also been found to impact gene expression, resulting in trait variation. For instance, large numbers of repeats in an STR in the 5′ untranslated region of *FRAXA* impacts local methylation and gene regulation, causing Fragile X syndrome (15).

More recent studies involving genome-wide surveys have found thousands of replicable associations between STR length and gene expression level (16–18). STR length variation can impact methylation, as well as histone modifications, causing evolutionarily conserved changes in gene expression (16, 17). Some of these STR-associated expression changes were associated with clinical traits (16). Somatic STR mutations have been implicated in the development of cancer (19). One recent analysis showed that individuals with autism have significantly more de novo STR mutations (particularly in introns), as compared with their neurotypical siblings (20). This growing body of evidence suggests that STR length variation is causally responsible for a range of complex trait variations, including pathogenic conditions (21).

These results raise questions about whether the CODIS loci could impact medically relevant traits. Based on data available in 2011, a review of phenotypic associations with genetic loci concluded that there were no significant associations with the CODIS STRs (6). However, the study did report that some CODIS loci fall within predicted sites for genomic regulation, and all CODIS loci are within 1 kb of at least one genetic variant associated with a phenotype (6). Because the linkage disequilibrium (LD) surrounding the CODIS loci is strong enough to infer the genotypes of surrounding single-nucleotide polymorphisms (SNPs) (21–23), phenotype information may be inferable through the CODIS genotypes. A more recent review of literature has identified 84 significant published associations between traits and STRs for 18 of the 20 CODIS loci (25).

Here, we investigate whether genotypes at the CODIS loci could directly reveal information about a fundamental trait: the expression levels of neighboring genes. We identify CODIS loci significantly correlated with the expression of nearby genes (CODISeSTRs). We shed light on the mechanisms underlying these associations. First, we consider the possibility that the associations are caused by population structure as a confounding factor by testing for expression–genotype associations within subpopulations. With population stratification ruled out, we explore the possibility of CODISeSTRs causing expression variation by both comparing their genomic features to a panel of STRs with strong evidence of expression impact (18) and using a fine-mapping framework (CAVIAR) to identify putative causal loci (26). Finally, we investigate the hypothesis that CODISeSTRs may be in LD with a causal variant by examining their LD with putative causal sites identified by CAVIAR, as well as DNase I hypersensitivity (DHS) sites.

Results

Gene Expression and the CODIS Loci in the 1,000 Genomes Dataset. We turned to a subset of the 1,000 Genomes Project to investigate the relationship between gene expression levels and CODIS loci genotypes. STR length variation was not directly genotyped in the 1,000 Genomes Project because this dataset used short-read sequencing. Thus, additional measures were taken to genotype these loci. Saini et al. (27) imputed STR genotypes for the 1,000 Genomes data by leveraging LD between STRs and surrounding SNPs to create a publicly available haplotype reference panel. This haplotype reference panel includes 18 of the 20 core CODIS STRs currently in use (28). Genotypes for STRs D16S539 and D21S11 were unavailable because their unusually long alleles are challenging to impute from short-read data.

While these imputed STR genotypes provide a tremendous resource, their accuracy is limited. For non-European-ancestry cohorts, the accuracy of imputed STR genotypes is lower because the imputation training data consisted only of individuals with European ancestry (27). Imputation accuracy is also lower for STRs with more alleles. For example, biallelic STRs have an average concordance of 97%, while the highly polymorphic CODIS STRs have an average concordance of 70%, with values ranging from 48 to 94% (27) (Dataset S1). For context, highly polymorphic pathogenic STRs with around 70% concordance have been detected through expression-association studies (27). Another way to quantify imputation accuracy is the dosage *r*-squared (DR2), an estimate of the squared correlation between the most likely genotype and the true allele dosage. Among CODIS STRs, DR2 ranges from 10 to 92%, with an average DR2 of 63% (27) (Dataset S1). Taken together, these metrics of imputation accuracy provide varying levels of confidence across the CODIS loci. While some CODIS STRs have fairly high confidence, imputation error erodes power to detect signal for other CODIS loci genotypes; thus, we use a summary statistic, β, to describe the expected sum of alleles at a locus, given the probability assigned to each possible allele (*Materials and Methods*).

We considered gene-expression values based on transcriptome data from lymphoblastoid cell lines from 421 individuals in the 1,000 Genomes Project. The populations represented in this set are: Utah residents with northern and western European ancestry, Finnish in Finland (FIN), British in England and Scotland (GBR), Toscani in Italy (TSI), and Yoruba in Ibadan, Nigeria (YRI), each population with a sample size ranging from 89 to 95 individuals (29). We investigated a model of CODIS STRs causing or being in LD with *cis* causal loci by considering expression level variation of genes within 100 kb of the CODIS loci. Out of the 18 CODIS STRs included in the haplotype reference panel, only 14 CODIS STRs are within 100 kb of at least one gene that is expressed in the lymphoblastoid data (*SI Appendix*, Supplemental Table 1). We considered a total of 39 CODIS STR–gene pairs, as the number of expressed genes within 100 kb varied for each CODIS loci. For each CODIS STR–gene pair, we tested for correlation between CODIS loci genotypes and the expression levels of neighboring genes. Note that in this analysis, we did not correct for population structure because we are not querying the molecular causality of an STR. Instead, we are investigating informative STR-expression associations, regardless of their cause.

Of the 39 CODIS STR–gene pairs tested, 6 showed significant correlations with p values below 0.05 and a false discovery rate (FDR) of 0.23 (so the expected number of false positives is 1.4) (Fig. 1 and Dataset S1 and *SI Appendix*, Fig. 1). The strongest signal was between D3S1358 and *LARS2* (p = 1.1e-6, and coefficient of determination $R^2 = 0.059$). We see less strong correlations, although still significant, between CSF1PO and *CSF1R* (p = 0.03, $R^2 = 0.01$), between CSF1PO and *TIGD6* (p = 0.04, $R^2 = 0.009$), between D2S441 and *C1D* (p = 0.01, $R^2 = 0.014$), between D18S51 and *KDSR* (p = 0.02, $R^2 = 0.011$), and between FGA and *PLRG1* (p = 0.03, $R^2 = 0.011$) (*SI Appendix*, Supplemental Table 1). While the adjusted coefficients of determination (R^2) observed are weak, their statistical significance or

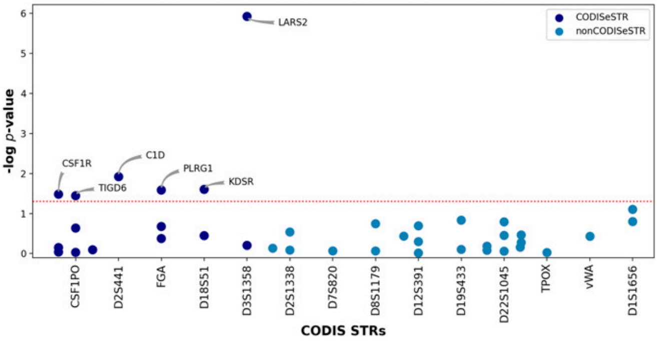

Fig. 1. Correlations between CODIS loci and the expression of neighboring genes. Associations of CODIS STR–gene pairs are shown as negative log p values. Red dotted line denotes the significant p value threshold. CODISeSTRs are shown in dark blue, and non-CODISeSTRs are shown in light blue.

marginal significance invites further investigation. The six CODIS STR–gene pairs with $p < 0.05$ represent an excess of correlation between CODIS loci and gene expression ($p = 2.9\text{e-}3$, χ^2 test). We refer to the CODIS STRs associated with gene-expression levels as CODISeSTRs.

Note that these CODISeSTRs (D3S1358, CSF1PO, D18S51, D2S441, and FGA) have imputation concordances of 0.67, 0.63, 0.51, 0.84, and 0.48, as well as a DR2 of 0.80, 0.55, 0.52, 0.68, and 0.61, respectively (Dataset S1). These measures suggest moderate to high levels of confidence in imputation accuracy.

The correlations that we observe could be explained by 1) a confounding factor like population structure and/or environmental variables in both CODISeSTR genotypes and expression levels, 2) CODISeSTRs causally impacting the expression of a neighboring gene, 3) LD between the CODISeSTR and a different causal locus that impacts expression, or 4) a spurious association in this particular dataset. However, establishing a putative mechanism for each of the observed associations may inform us about its stability and generalizability. We explore these hypotheses in the following analyses.

Exploring the Role of Population Substructure in Observed CODISeSTR Correlations. First, we investigate if the observed CODISeSTR-expression-level associations observed across the whole dataset (including CEU, GBR, FIN, TSI, and YRI) are caused by population structure as a confounding factor. We investigate this possibility by adding population membership as a covariate in the linear models of gene expression and CODISeSTR variation. For most CODISeSTR–gene pairs (D3S1358–*LARS2*, D18S51–*KDSR*, D2S441–*C1D*, and FGA–*PLRG1*), the associations remain significant with population as a covariate (Dataset S1). Thus, the associations observed are unlikely to be caused by population structure. For CSF1PO–*CSF1R* and CSF1PO–*TIGD6*, the associations are somewhat less pronounced with population as a covariate (CSF1PO–*CSF1R* p value goes from 0.03 to 0.06 and CSF1PO–*TIGD6* from 0.04 to 0.06). This is consistent with the hypothesis that CSF1PO associations may be caused in part by population stratification.

If a cumulative association is caused entirely by structure at the level of the specified subpopulations, then there would be no associations within subpopulations, and there would be differences in β and expression-level distributions across subpopulations. We further investigate population-specific associations by testing for CODISeSTR-expression-level associations within subpopulations (*SI Appendix*, Fig. 2 and Supplemental Table 2). We found six significant associations under a p value threshold of 0.05, with an FDR of 0.21 (expected number of false positives is 1.3) (Dataset S1).

For D3S1358–*LARS2*, D18S51–*KDSR*, and CSF1PO–*CSF1R*, we observe CODISeSTR-expression level associations within subpopulations. D3S1358 β is significantly correlated with *LARS2* expression in the FIN group (p = 0.0013, R^2 = 0.11), as well as showing a significant correlation in the TSI (p = 0.005, R^2 = 0.08) and the GBR (p = 0.02, R^2 = 0.06) groups. This lack of subpopulation heterogeneity ($I^2 = 0\%$) is consistent with the significant p value in the cumulative population ($p = 1.12\text{e-}6$, R^2 = 0.06) (Dataset S1 and *SI Appendix*, Fig. 3*A*). For D18S51 and *KDSR*, we see a significant association in YRI (p = 0.003, R^2 = 0.11), with nonsignificant results for all other subpopulations. However, in this case, the stronger association in YRI does result in higher subpopulation heterogeneity (I^2 = 34%) (Dataset S1 and *SI Appendix*, Fig. 3*D*). For CSF1PO–*CSF1R*, we see correlations in FIN (p = 0.04, R^2 = 0.04) and GBR (p = 0.04, R^2 = 0.04), but in different directions ($r = 0.22$ and $r = -0.24$, respectively) (Dataset S1 and *SI Appendix*, Fig. 3*B*). These results suggest that for those CODISeSTR–gene pairs, the cumulative signal is not a product of population structure, but it may be driven by stronger and/or distinct associations in some subpopulation groups.

By contrast, while the associations for D2S441–*C1D*, CSF1PO–*TIGD6*, and FGA–*PLRG1* are significant in the cumulative dataset, within subpopulations, we observed no significant or nominally significant correlations (*SI Appendix*, Supplemental Table 2). To determine whether subpopulation structure is causing both the cumulative association and lack of associations within populations, we tested for differences in the β and expression-level distributions between subpopulations (*SI Appendix*, Supplemental Table 3 and Fig. 4). We did observe some significant differences in β distributions for FGA (YRI–CEU) and CSF1PO (YRI–FIN, –GBR, and –TSI), but there were no corresponding significant differences in the expression distributions of *C1D*, *TIGD6*, or *PLRG1*. Thus, the significant associations for these CODISeSTR–gene pairs are not caused by this level of population structure as a confounding factor. Instead, it suggests either that the association is too weak to detect within subpopulations with decreased sample size and statistical power, or that the cumulative correlation is spurious.

Power Analysis for Detection of Associations between CODIS Genotypes and Expression Levels. Given the small sample sizes in the empirical data, we sought to determine the power to detect significant associations, particularly within subpopulations. Accordingly, we simulated genotypic and phenotypic data for a European-like (EUR) population for CODISeSTRs CSF1PO and D3S1358 and a Yoruban-like (YRI) population for D18S51. Because admixed populations, such as Latinx and African-Americans, are not included in the empirical data, but are overrepresented in CODIS databases (8), we also investigated the power to detect associations in an admixed American-like (AMR) population.

For CSF1PO, D3S1358, and D18S51 and for each of the populations where we found a significant association, we simulated genotypes and traits with varying phenotypic variance explained (PVE) by genetic variation (*Materials and Methods*). Using these data, we performed power analyses over varying sample sizes, from n = 20 to 200.

We found that the power to detect associations is sensitive to PVE, with higher PVEs yielding higher powers (*SI Appendix*, Fig. 5). Similarly, we observed a positive association between sample size and power. For the average sample size in the European ancestry subpopulations in the empirical data—80 individuals—and for a PVE of 0.5 in both CSF1PO and D3S1358 simulations, we estimated a power to detect associations at 0.74

(Dataset S2 and *SI Appendix*, Fig. 5). For D18S51, sampling 80 individuals from a YRI-like population yielded a power of 0.70 for a PVE of 0.5 (*SI Appendix*, Fig. 5 and Dataset S2). Similar trends were observed when sampling from the admixed AMR population. For both D3S1358 and D18S51, when sampling 80 individuals from the AMR population for a PVE of 0.5, the estimated power was 0.72, which is comparable to the powers estimated for those CODISeSTRs in the EUR and YRI populations. For CSF1PO, power in AMR was 0.66, lower than in the EUR population (*SI Appendix*, Fig. 5 and Dataset S2).

We note that PVE values reported above are high compared to most empirical data and that these power analyses may be sensitive to modeling parameter values for both genotype and phenotype simulations (30, 31). Nonetheless, these results serve as a starting point for more in-depth power analyses, show that such associations are detectable, and suggest that some associations may be slightly less detectable in admixed populations.

Comparing Genomic Features of CODISeSTRs and FMeSTRs. We go on to investigate if the CODISeSTRs resemble expression-associated STRs (eSTRs). A previous genomic analysis of a total of 1,620,030 STRs in humans identified 20,609 eSTRs with evidence of impacting gene expression in 1 of 17 tissues (18). These eSTRs were then fine-mapped and ranked by their probability of causality using the statistical framework CAVIAR (26). The eSTRs with the top 5% of probabilities of causality (1,380 unique STRs) were then characterized as fine-mapped eSTRs (FMeSTRs) to express the additional evidence for their impact on gene expression (18). Of note, 3 of the 20 CODIS STRs were previously identified as eSTRs by Fotsing et al. (18), more than expected by chance (one-tailed binomial test, $p = 0.002$). Specifically, expression associations were found for TPOX in tibial nerve tissue, for D2S1338 in heart left ventricle tissue, and for TH01 in both visceral adipose and esophagus mucosa tissues. It is unsurprising that this study in lymphoblast cell lines did not reproduce those associations. However, these potential associations raise questions about expression associations of other CODIS loci across tissues.

The genomic features of FMeSTRs were characterized, showing that they are more likely to be long, intronic, located near transcription start sites (TSSs), located near DHS sites, and to contain particular repeating units. We examine how the CODISeSTRs fit the FMeSTR profile in order to investigate the hypothesis that the CODISeSTR genotypes are directly causing changes in the expression of neighboring genes (Table 1).

In general, the CODIS loci are similar to FMeSTRs in their extreme length—the CODISeSTRs, in particular, are all in at least the 93rd percentile of lengths compared to all genomic STRs (*SI Appendix*, Fig. 6). The CODISeSTRs further resemble FMeSTRs in that four of the five are intronic (CSF1PO: *CSF1R*; D18S51:*BCL2*; D3S1358:*LARS2*; and FGA:*FGA*). Like FMeSTRs, two CODISeSTRs are unusually near to a TSS: FGA is 2.92 kb from the TSS of the gene *FGA* (92.7th genomic percentile), and CSF1PO is 4.65 kb from the TSS of *CSF1R* (88.6th genomic percentile) (*SI Appendix*, Fig. 7). Similar to FMeSTRs, which are disproportionately found near DHS sites, one CODISeSTR in particular overlaps with a DHS site observed in lymphoblasts or lymphoblast derivatives: CSF1PO (100th genomic percentile) (*SI Appendix*, Figs. 8 and 9). Finally, the repeating units of four of the five CODISeSTRs have been found to be significantly enriched among eSTRs (D3S1358, CSF1PO, D2S441, and D18S51) (18) (Table 1).

Altogether, CODISeSTRs, most particularly CSF1PO, fit the genomic profile of an FMeSTR that putatively impacts gene-expression levels. These results are consistent with a hypothesis that CSF1PO has a causal impact on *CSF1R* expression levels, without being conclusive evidence.

Identifying Local Genetic Variants to Explain Observed Variation in Gene Expression. Knowing that CODISeSTRs resemble STRs that impact expression, we attempted to determine whether the expression variation is being driven by each CODISeSTR itself or if it is due to nearby causal genetic variants (other STRs or SNPs) in LD with the CODISeSTR. To identify causal *cis* variants, we used CAVIAR, a Bayesian fine-mapping framework that leverages pairwise LD and *z* scores to identify a set of putatively causal variants (26). CAVIAR assigns a posterior probability (which we will refer to as "CAVIAR score") to each marker in the ρ-causal set (18, 32). With a ρ = 0.95, the ρ-causal set is a subset of markers that with 95% confidence contains all causal variants. Fine-mapping was performed in each subpopulation and CODISeSTR–gene combination for which we found a significant or nominally significant association (*Exploring the Role of Population Substructure in Observed CODISeSTR Correlations*). Thus, this causality analysis includes *LARS2* in the FIN, GBR,

Table 1. Genomic features of CODISeSTRs

CODISeSTR	Location relative to genes	Distance to nearest TSS, bp (genomic percentile, %)	Distance to TSS of associated gene, bp	Distance to nearest DHS site, bp (genomic percentile, %)	Distance to nearest lymph DHS site, bp (genomic percentile)	Length, bp (genomic percentile, %)	Repeating unit
D3S1358	Intronic to LARS2	31,194 (52.6)	152,15 to LARS2	1,916 (72.3)	4,651 (69.1)	63 (96.7)	$[AGAT]_n$
D2S441	Intergenic	41,143 (45.8)	41,143 to C1D	14,064 (28.1)	19,514 (39.3)	47 (93.6)	$[TGCC]_m[TTCC]_n$
CSF1PO	Intronic to CSF1R	4,649 (88.6)	4,649 to CSF1R 75,157 to TIGD6	0 (100.0)	0 (100.0)	51 (94.8)	$[AGAT]_n$
D18S51	Intronic to BCL2	36,928 (48.4)	85,535 to KDSR	13,230 (29.3)	3,714 (73.2)	71 (97.3)	$[AGAA]_n$
FGA	Intronic to FGA	2,922 (92.7)	2,922 to FGA 37,303 to PLRG1	8,065 (40)	27,933 (27.9)	87 (98.1)	$[TTTC]_m$ $TTTTTTCT[CTTT]_n$ $CTCC[TTCC]_o$

and TSI populations; *CSF1R* in the FIN and GBR populations; and *KDSR* in YRI.

These CAVIAR analyses produced scores between 0.04 and 0.60 for the putative causal variants (Dataset S3). While the relatively small sample sizes (65 to 83 individuals) mean that power may be limited for some of these analyses, scores as high as 0.60 are noteworthy.

We also used CAVIAR to estimate the most likely *n* causal variants contributing to the phenotype with a max of $n = 4$. The "putative causal set" comprises *n* variants in the ρ-causal set with the highest CAVIAR scores. While the CODISeSTRs were not tagged as putative causal variants, they do appear in most of the ρ-causal sets. For example, the highest CAVIAR score in a CODISeSTR is for D18S51 in the YRI subpopulation at 0.10.

Investigating LD between CODISeSTRs and Putative Regulatory Elements. The observed correlation between CODISeSTR genotypes and expression levels of neighboring genes could be caused by LD between a CODISeSTR and a regulatory variant that impacts gene expression. To investigate this possibility, we considered LD between CODISeSTRs and both CAVIAR-identified putative causal variants and DHS sites, indicating accessible DNA likely to contain regulatory elements.

In addition to the identification of CODISeSTR D18S51 in the ρ-causal set for *KDSR* expression in YRI, we observed high LD between CODISeSTR D3S1358 and the putative causal variants impacting *LARS2* expression levels (Fig. 2 and *SI Appendix*, Fig. 10). Two of the putative causal variants for *LARS2* expression in FIN have an LD of at least 0.61 with CODISeSTR D3S1358, while four of the putative causal variants for *LARS2* expression in GBR have an LD of 0.54 with D3S1358, and the putative causal variants with top CAVIAR scores for LARS2 expression in TSI have an LD of at least 0.68 with D3S1358. While the CAVIAR scores associated with these ρ-causal sets are modest, the convergent high LD values across subpopulations support the hypothesis that D3S1358 may be in LD with a causal locus contributing to LARS2 expression variation.

We also examined the LD between CODIS STRs and DHS sites within 100 kb (*SI Appendix*, Figs. 11 and 12). We measured the STR–DHS site LD as the maximum correlation between the STR and a SNP located in the DHS site (*SI Appendix*, Fig. 11 and Dataset S4). We observed a high DHS-site LD for some CODIS STRs (D1S1656, TH01, TPOX, and vWA) and a large number of DHS sites surrounding others (CSF1PO, D22S1045, and TH01). Note that, of these, one is a CODISeSTR, while TPOX and TH01 were previously

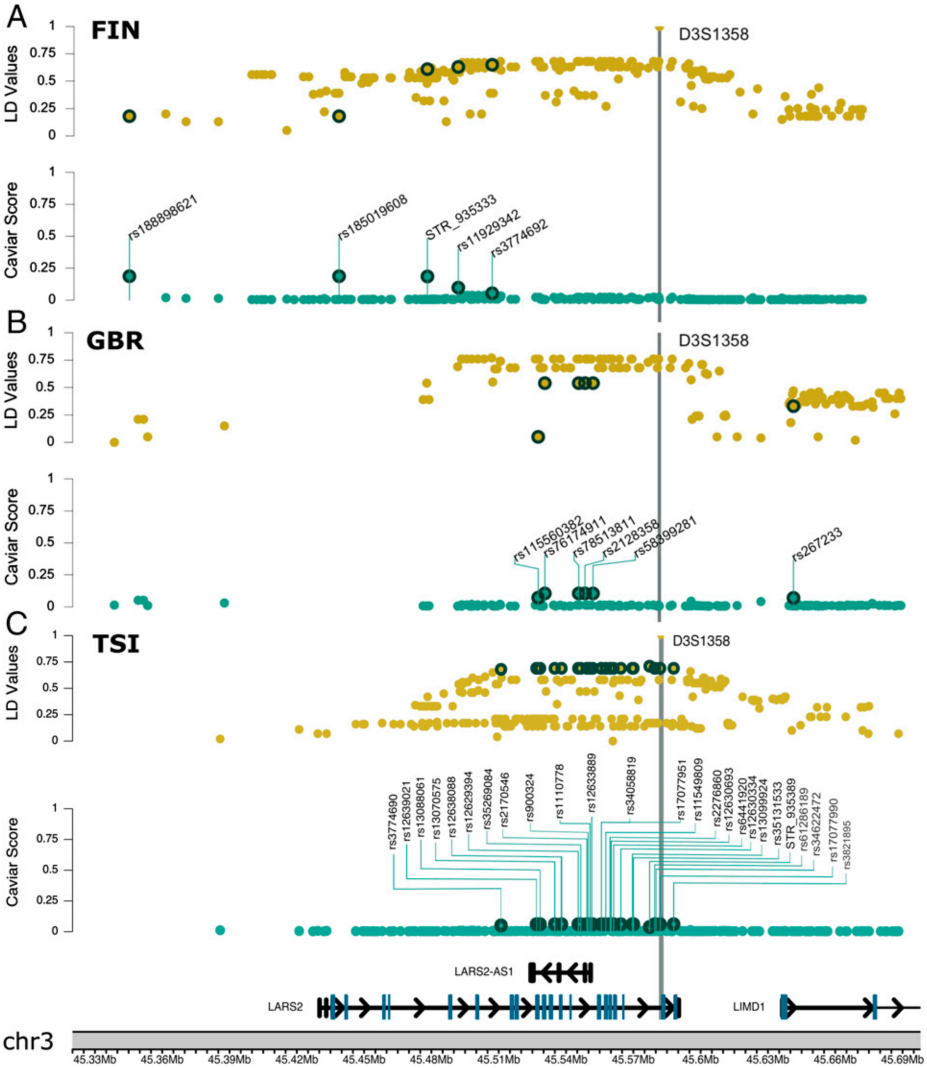

Fig. 2. *LARS2*–D3S1358 CAVIAR and local LD landscape. Local LD and CAVIAR score landscapes in a 100-kb window centered on the *LARS2* gene for the FIN subpopulation (*A*), GBR subpopulation (*B*), and TSI subpopulation (*C*). For each plot, *Upper* shows LD between the CODISeSTR D3S1358 versus each variant in the ρ causal set, and *Lower* shows CAVIAR scores for variants in the ρ causal set. Dark green circles enclose putative causal variants in both CAVIAR and LD panels. Chr, chromosome.

identified as eSTRs (18). We did not observe a general excess of LD with DHS sites for CODISeSTRs, as compared to other CODIS STRs ($p = 0.52$, two-tailed Kolmogorov–Smirnov test).

Additionally, the CODISeSTR CSF1PO overlaps with a DHS site, suggesting that variation in CSF1PO length may directly impact the action of that DHS (*SI Appendix*, Fig 12). D3S1358 is in LD ($r^2 > 0.45$) with SNPs in four DHSs detected in lymphoblasts, while D18S51 has an LD of $r^2 = 0.31$ with SNPs in one DHS (Dataset S4).

Putative Mechanisms for Observed CODISeSTR-Expression Associations. For each CODISeSTR–gene pair, we weighed the results supporting different mechanisms for the observed STR–gene-expression association. (Table 2).

Association between D3S1358 and LARS2 expression. We observed a significant negative correlation between D3S1358 allele length and the *LARS2* expression levels in our cumulative 1,000 Genomes analysis (Dataset S1). The strength of this correlation is demonstrated by the maintenance of the significant association within the smaller FIN subpopulation, as well as associations within GBR and TSI (Dataset S1). As to the mechanism for this correlation, there is weak evidence suggesting that D3S1358 resembles a causal FMeSTR (Table 2). However, there is stronger evidence that D3S1358 is in LD with both a variant that putatively impacts LARS2 expression (Dataset S3) and DHS sites active in lymphoblasts (Dataset S4). Together, these results are consistent with the hypothesis that D3S1358 may be in LD with a locus that impacts *LARS2* expression.

Association between CSF1PO and CSF1R expression. *CSF1R* expression has a significant negative correlation with the genotype of intronic CODIS locus CSF1PO (Dataset S1) in the cumulative analysis. The subpopulations FIN and GBR show a significant positive and negative correlation, respectively (Dataset S1), and there is only weak evidence that the association is partially due to population stratification (Dataset S1). CSF1PO bears a remarkable resemblance to FMeSTRs with its close proximity to *CSF1R*'s TSS and particularly with its overlap with a DHS site found in lymph cell lines, as well as its length, and repeating unit (Table 1 and *SI Appendix*, Fig. 12). These results are consistent with the hypothesis that CSF1PO could causally impact *CSF1R* expression or be in LD with a different causal locus.

Association between D18S51 and KDSR expression. D18S51 β values are significantly correlated with *KDSR* expression across all samples, as well as within the YRI subpopulation (Dataset S1 and *SI Appendix*, Table 2). We note that the distribution of β values is significantly different in YRI compared to the other subpopulations (*SI Appendix*, Table 3 and Fig. 4). Further, in the YRI subpopulation, the D18S51 was identified as the second-most-probable locus to cause *KDSR* expression variation, with a CAVIAR score of 0.10 (Dataset S3). Even if D18S51 itself is not causal, we note that its LD with a DHS site is 0.31 (Dataset S4 and *SI Appendix*, Fig. 12). Together, these results suggest that a correlation within the YRI subpopulation could be driving the correlation at the cumulative level and that the correlation likely has a molecular basis (whether causal or in LD with a causal locus).

Associations between CSF1PO and TIGD6 expression; D2S441 and C1D expression; and FGA and PLRG1 expression. For the remaining CODISeSTR–gene pairs, we see significant correlations at the cumulative population level, with no significant associations within subpopulations (Dataset S1 and *SI Appendix*, Table 2). While this might suggest that the associations are due to population structure, two factors tell a different story: 1) the maintenance of a significant association with population as a covariate (Dataset S1); and 2) the lack of consistent subpopulation differences in both β frequencies and expression levels (*SI Appendix*, Table 3 and Fig. 4). These results may be explained by either the smaller subpopulation group sample sizes reducing power to detect weak correlations, or a spurious association in the cumulative analysis. This is consistent with the other analyses showing that D2S441 and FGA weakly fit the pattern of FMeSTRs (Table 1) and have relatively low LD to local DHS sites (Dataset S4).

Discussion

CODIS loci were chosen because, at the time, researchers believed that no medical information would be revealed. However, in this study, we identified CODISeSTRs whose genotypes are correlated to the expression of neighboring genes in lymphoblast cell lines. Our results build on previous work finding expression associations with CODIS loci TPOX, TH01, and D2S1338 in other tissues (18). Specifically, we observed six significant correlations: between D3S1358 and *LARS2*, between D18S51 and *KDSR*, CSF1PO and both *CSF1R* and *TIGD6*, D2S441 and *C1D*, and FGA and *PLRG1*. We go on to investigate the putative mechanism for these correlations, finding that the associations between D3S1358–*LARS2*, D18S51–*KDSR*, and CSF1PO–*CSF1R* are likely due to a causal relationship or LD with at least one causal locus, while the other associations are weaker or possibly spurious. These results provide evidence that contravenes the assumption that CODIS genotypes convey no trait information.

Table 2. Putative mechanisms for observed CODISeSTR-expression associations

CODISeSTR–*Gene*	Association observed at subpopulation level	CODISeSTR fit to FMeSTR Profile*	CODISeSTR LD with CAVIAR causal variants†	CODISeSTR LD with DHS sites active In lymphoblasts†
CSF1PO–*CSF1R*	Yes	Strong	Low	Overlaps with DHS site
D18S51–*KDSR*	Yes	Moderate	Low	Low
D3S1358–*LARS2*	Yes	Weak	Moderate–High	Low–Moderate
CSF1PO–*TIGD6*	No	Strong	N/A	Low–Moderate
D2S441–*C1D*	No	Weak	N/A	Low
FGA–*PLRG1*	No	Weak	N/A	Low

*Strong fit is defined as satisfying most or all of the FMeSTR characteristics described in Table 1; moderate is defined as satisfying at least half; weak is defined as satisfying less than half.

†High LD is considered ≥ 0.7; moderate LD is between 0.4 and 0.69; low LD is <0.4. N/A values indicate STR–gene pairs that were not included in the CAVIAR analysis.

Medical Relevance of CODISeSTR-Associated Gene-Expression Variance. Our analysis shows that a CODIS genotype profile can be used to gain information about the expression of some genes. This raises the question: Would those expression levels reveal medical information? A number of studies have found associations between medical conditions and CODIS loci, including the CODISeSTRs (25). In addition to those findings, we consulted the medical genetics literature to begin to address this question (*SI Appendix*, Table 4 and *Text* 1–6). We discuss some of the most striking cases: *CSF1R*, *LARS2*, and *KDSR*.

CSF1R expression variation and neural and psychiatric conditions. *CSF1R*, which CSF1PO is intronic to, encodes a cytokine receptor that plays a key role in microglial regulation (33). Disruptive sequence mutations in *CSF1R* lead to a variety of brain conditions, including leukoencephalopathy (33–38), while inhibition of CSf1R protein function seems to ameliorate some neural conditions like epilepsy (39), Alzheimer's disease (40–42), and spinal cord injury recovery (43) (*SI Appendix*, *Text* 3). Further, and most relevantly for this study, variation in the expression and splicing of *CSF1R* are associated with psychiatric conditions, including depression and schizophrenia in humans (44–46). Since *CSF1R* expression is correlated with CSF1PO, the CODIS genotype may be informative about those psychiatric conditions.

LARS2 and KDSR associations with medical conditions. *LARS2* and *KDSR* function reduction or elimination have also been associated with medical conditions. *LARS2*, which contains D3S1538 in an intron, encodes a mitochondrial leucyl–transfer RNA synthetase gene (47). *LARS2* is well-established as an essential gene, as mutations that reduce or knock out its function have been associated with Perrault syndrome (48–50), MELAS syndrome (51), and other conditions (*SI Appendix*, *Text* 1). *KDSR*, which is near to D18S51, encodes an enzyme involved in synthesis of the lipid ceramide. Mutations in *KDSR* that eliminate or decrease enzyme function have been associated with a number of severe skin and platelet conditions (52–54) (*SI Appendix*, *Text* 5). These medical genetic studies provide strong evidence that *LARS2* and *KDSR* expression variation impact a number of medical conditions. The fact that dramatically reduced function leads to severe phenotypes raises the question of whether marginally lowered expression may lead to intermediate conditions. The association between CODISeSTRs and those genes' expression means that the CODIS genotype may be informative about risk of those conditions or other intermediate phenotypes.

Limitations. The associations reported here were observed in the subset of the 1,000 Genomes Project, where expression data were also available. These data are limited in a few important ways. First, expression data were only available in lymphoblastoid cell lines. This single cell type means that our analysis will miss genes that are not highly expressed, or whose expression isn't regulated by *cis*-elements, in lymphoblastoid cell lines specifically. Second, data were only available from CEU, FIN, GBR, TSI, and YRI subpopulations. As four of five of these populations are European, they do not reflect the genetic diversity of the general population of the United States, notably with regard to their lack of admixture. Correlations due to population structure as a confounding factor may be underestimated in our analysis. Further, this analysis is unable to identify correlations that are specific to subpopulations not represented here. Third, errors in the imputation of the CODIS genotypes may erode power to identify associations, particularly in non-European subpopulations, where imputation has higher error rates.

In addition, our approach to detecting associations is specifically testing for a linear relationship between STR allele length and expression levels. While this type of linear relationship is generally expected (*Materials and Methods*), there could be other nonlinear relationships present that were not detected here.

Altogether, while our analysis produced significant correlations, it is limited in scope and underpowered. This raises the question of whether stronger correlations would be identified in an analysis on a larger, more representative sample, with direct STR genotyping, using more expression data from more varied tissues.

Conclusion

Within the limitations of the publicly available data examined here, our results suggest that information on gene expression levels may be revealed by CODIS profiles. Further, some of those gene expression levels have been connected to medical phenotypes. These results join a growing body of work showing that CODIS genotypes may contain more information than purely identity. CODIS profiles have been found to provide information about the surrounding haplotype (21–23), as well as genetic ancestry (55). Together, these findings raise concerns about the medical privacy of individuals whose CODIS profiles are seized, databased, and accessed, as well as the genetic relatives of those persons.

Materials and Methods

The 1,000 Genomes Project CODIS Genotype Data. Phase 3 of the 1,000 Genomes Project sampled 2,504 individuals from 26 different populations with ancestry from Africa, East Asia, Europe, South Asia, and the Americas (56). The short-read-sequencing approach used for this dataset presents a challenge for genotyping the CODIS loci, which are highly polymorphic, often with very long alleles. We used imputed CODIS loci genotype data that were made publicly available as a haplotype reference panel (27, 57, 58). Because of the limits of this approach with particularly long alleles, genotypes for only 18 of the 20 CODIS STRs were successfully imputed (27). These 18 loci are D22S1045, TPOX, D2S441, D2S1338, vWA, D12S391, D5S818, CSF1PO, D1S1656, D10S1248, TH01, D13S317, D18S51, D19S433, D3S1358, FGA, D7S820, and D8S1179. The two CODIS STRs not included in our study are D16S539 and D21S11. We note that the very factors that make these loci difficult to impute (length and polymorphism) may make them particularly relevant for studies of phenotypic impact (18).

For our analysis of correlation between CODIS genotypes and expression levels, we created a summary statistic based on estimated allelic dosages generated by Beagle during imputation. For each individual, STR estimated allele dosages are the sum of the posterior allele probabilities for both haplotypes (59). Hence, their values range from zero to two (60).

We used the imputed STR allelic dosages to compute a normalized linear weighted genotype for each CODIS STR. We refer to this weighted average genotype as β (beta). We computed β for each individual for each CODIS STR using the following:

$$\beta = \frac{1}{2}\sum_{i=1}^{n} r_i d_i,$$

where n is the number of distinct alleles on record at the locus, r_i is the number of repeats in allele i, and d_i is its estimated allelic dosage (genotype probability). For non-CODIS STRs, β genotypes were computed by substituting r_i with the allele nucleotide length, instead of the repeat count.

The 1,000 Genomes Project Gene-Expression Data. Transcriptomes were typed from lymphoblastoid cell lines of 462 unrelated individuals from the 1,000 Genomes Project (29). The samples in this set correspond to five populations: CEU, FIN, GBR, TSI, and YRI. Transcriptomic levels were quantified with reads per kilobase of transcription per million mapped reads (RPKM).

Transcripts with zero counts in more than half the number of samples were removed (29). Full data are available at the European Bioinformatics Institute ArrayExpress portal (accession no. E-GEUV-1) (61) (see *Data, Materials, and Software Availability* for URL).

Of the 462 individuals with gene-expression data, 90 were filtered out because either β genotypes or gene-expression values were missing. Our study was performed with 372 individuals for which we have CODIS genotype data and where at least one known gene within a 100 kb window was expressed in the lymphoblastoid cell-lines data. Within-population analysis contained between 65 and 83 individuals.

Testing Associations between STR Length and Gene Expression. Using data from the University of California Santa Cruz (UCSC) Genome Browser, we identified genes within 100 kb of the CODIS markers, measuring 100 kb from the start and end of each CODIS STR genomic location. We summarized the genotypes with β values, as detailed in *The 1,000 Genomes Project CODIS Genotype Data*. We next fit linear regression models to test for Pearson correlation between the β genotypes for each CODIS STR versus the expression levels of nearby genes.

The approach implicitly assumed a linear relationship between STR alleles by length. This assumption is justified by findings that STR impact on expression level scales with allele length (15, 21, 62). For an STR that is not causal, but is in LD with a causal locus, the step-wise STR mutational process (63, 64) will lead to multiple similarly lengthed STR alleles on the causal haplotype.

We controlled for FDR using q values (65). With 6 features under a p value threshold of 0.05, we expect 1.4 of those to be a false positive (Dataset S1).

Power Analysis in European, Yoruban, and Admixed Individuals. For each CODISeSTR with a significant association within a subpopulation (CSF1PO, D18S51, and D3S1358), we performed 1,000 coalescent simulations per sample size using Msprime (66). Sample sizes ranged from 20 to 200 diploid individuals in increments of 20. We simulated a single chromosome composed of one STR surrounded by 100 kb of DNA sequence on either side. STR variation was generated by using a stepwise model of mutation with low and high repeat values following each population's empirical genotypic variance (67). The demographic parameters are those of the American Admixture model from Browning et al. (68). Under this model, admixture from African, European, and East Asian populations into admixed individuals from the Americas was introduced 12 generations ago with admixture proportions of 1/6, 2/6, and 3/6, respectively. We used the mutation rates for CSF1PO, D18S51, and D3S1358 that were estimated by Saini et al. (27).

A pool of potential causal variants was created to include SNPs with both a minor allele frequency above 0.05 and an LD of at least $r^2 = 0.20$ with the CODISeSTR. A set of four causal SNPs was then randomly sampled, without replacement, from the pool of potential causal variants.

Using the genotypes corresponding to the set of randomly sampled causal SNPs, we generated gene-expression values by employing an additive linear model. This model is represented by: $y = X\,b + \varepsilon$, where X is the set of causal SNPs genotypes, b is a vector of linear effect sizes, and ε is the residual error. Linear effect sizes b and residual errors ε were drawn from a Gaussian distribution $N \sim (0, I)$, where I represents the identity matrix. We scaled the effect sizes b and I (additive effects) such that the genotypes explained some fixed proportion of the phenotypic variation. Following PVE ranges observed in real data, we generated gene-expression phenotypes for an array of PVE values starting at 0.05 and up to 0.5 in increments of 0.05 (30, 31).

With these simulated genotype and phenotype data, we tested for associations in the same way as in the empirical analyses.

Subpopulation Heterogeneity Analysis. For each CODISeSTR with a significant association in the cumulative analysis, we performed the I^2 test for subpopulation heterogeneity using the *metaforest* library in R. We used a random-effects model assuming different effect sizes across studies and created forest plots using the *forest* function.

Characterizing Genomic Features of CODISeSTRs. We quantified several genomic features of the CODISeSTRs in order to examine how they compare to the characteristics of putatively causative expression-altering STRs (referred to as FMeSTRs) (18). Genomic STR coordinates, including those of CODIS STRs and CODISeSTRs, were gathered from a genome-wide survey of STRs (32, 69).

For context, we used the survey data to compute CODISeSTR lengths and their length percentiles as the proportion of genome-wide STRs that are at least as long. Distance, in base pairs, between STRs to the nearest gene and the nearest TSS was determined by additionally using genetic coordinates from the UCSC Genes track in the UCSC Genome Browser (70). We also used the UCSC Genes track to determine the distance of each STR relative to its associated gene(s) and the TSS(s) thereof. The repeating units for each CODISeSTR were gathered from STRbase (71, 72).

Like in the analysis of FMeSTRs (18, 32), for each CODISeSTR, we computed the distance between each STR and the nearest DHS site. DHS site cluster locations were taken from the ENCODE Regulation "DNase Clusters" track via the UCSC Genome Browser (73, 74). All distances between STRs and nearby genomic elements, except for TSSs, which are represented by the starting coordinate of the protein-coding region, were calculated to reflect the distance between the closest endpoints of the elements in question.

For the general analysis, we considered DHS site clusters annotated in at least 20 sources. We performed additional analyses focusing on DHS site clusters observed in lymphoblasts or lymphoblast derivatives. We identified 20 cell lines that are lymphoblasts or lymphoblast derivatives, specifically, Adult_CD4_Th0*, CD20+, CLL, CMK, GM06990, GM12864, GM12865, GM12878, GM12891, GM12892, GM18507, GM19238, GM19239, GM19240, HL-60, Jurkat, K562, NB4, Th1, and Th2. For lymphoblast-specific analyses, we consider DHS sites that were observed in at least 5 of the 20 lymphoblasts or lymphoblast derivatives in the dataset.

Evaluating the Potential Causality of *cis* Variants. We performed a fine-mapping analysis with CAVIAR (18, 32) to identify specific local genetic variants (either CODISeSTRs, other STRs, or SNPs) that are putatively causal of the variation in expression levels. CAVIAR employs the variants' LD structure, as well as association statistics to predict a subset of variants, the ρ causal set, in which all causal markers are said to be included with a certain probability ρ, with $\rho =$ 95% in our case. Each variant in the CAVIAR ρ credible set is then assigned a probability of being causal. We refer to this posterior probability as CAVIAR score.

We considered SNPs and STRs within 100 kb upstream and downstream of genes with significant or marginally significant CODIS locus associations at the subpopulation level: *LARS2*, *CSF1R*, and *KDSR*. SNPs that did not exhibit variation within each subpopulation group were removed. We followed the CAVIAR protocol established by Fotsing et al. (18). Next, we filtered for SNPs and STRs that hold a significant association with gene-expression level. Specifically, we tested for correlation between gene expression and either SNP or STR genotypes. For non-CODIS STRs, genotypes were considered as the nucleotide length of each allele, and β values were computed, while for CODIS STRs, we considered the number of repeats. Variants with $p > 0.05$ were excluded from further analysis. Since it is unlikely for a phenotype to be caused by one variant alone, we allowed for CAVIAR to consider up to four independent causal variants per locus by including the parameters -f 1 -c 4. We define the putative causal variants as the n number of variants with the highest CAVIAR scores, where n is CAVIAR's predicted number of putative causal variants, ranging from one to four. In the cases where a set of variants are in perfect LD with one another (and therefore have identical CAVIAR scores), the set is considered as a single prediction.

Quantifying LD. LD between STRs and SNPs was quantified as the correlation between CODIS STR β values versus the SNP genotypes (sum of alternative alleles), implicitly testing a linear relationship. This measure of LD between STRs and SNPs is similar to a haplotype-based method shown to reliably follow the expected patterns of variation when applied to phased X-chromosome haplotypes (75). LD between two STRs was quantified as the correlation between β values. For CODISeSTRs, β was based on the number of repeats, while for non-CODIS STRs, β was based on the nucleotide length. We used genotypic LD, rather than haplotypic LD, because the imputed STR estimated allele dosages lack phase information.

For analyses of LD between STRs and DHS sites, we calculated LD between CODIS loci and all SNPs within a DHS site. We considered DHS sites found within at least 5 of the 20 available lymphoblast cell lines or derivatives, as well as DHS sites found within at least 20 cell-line sources (*SI Appendix*, Figs. 9 and 12). As a summary, we considered the highest LD per the DHS site (Dataset S4). For DHS sites without SNPs in the region, LD was not computed and therefore not included in the summary.

Data, Materials, and Software Availability. Weighted average beta genotypes generated in this study, CAVIAR, regression, and power analysis scripts are publicly available at https://github.com/TalesOfDNA/CODISMarkers-Xprssn.git (76). Previously published data were used for this work. CODIS STR genotypes imputed in the 1,000 Genomes Dataset are available on the Gymrek Lab website (http://gymreklab.com/2018/03/05/snpstr_imputation.html) (58). Transcriptome data from cell lines derived from individuals participating in the 1,000 Genomes Dataset are available at ArrayExpress (https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/) (61). The genome-wide STR survey is available at GitHub (https://github.com/HipSTR-Tool/HipSTR-references/blob/master/human/hg19.hipstr_reference.bed.gz) (69). Technical details on CODIS STRs are available at STRBase (https://strbase.nist.gov/str_fact.htm) (72). DHS site locations from ENCODE are available (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegDnaseClustered/wgEncodeRegDnaseClusteredV3.bed.gz) (74).

ACKNOWLEDGMENTS. We first thank all of the research participants who shared their genomic data through the publicly available sources used for this project. This research is only possible with their generosity. We are grateful to Doc Edge, and Erin Murphy for thoughtful conversations and manuscript review that informed and improved this project. We thank Lorin Crawford, and Andy Kern for supportive conversations regarding the power analysis. We thank Camille Rey for writing suggestions that improved this manuscript. San Francisco State University (SFSU) students M.M.B., Y.J.A.Z., A.R., and M.G. were supported by the SFSU Minority Biomedical Research Support Program–Research Initiative for Scientific Enhancement (MBRS-RISE) Fellowships (R25-GM059298), SFSU Maximizing Access to Research Careers (MARC) Fellowships (T34-GM008574), Bridges Fellowships (R25-GM048972), and Genentech Foundation Fellowships. Funds through NIH-R35GM128946-01 and an Alfred P. Sloan Foundation Fellowship supported E.H.-S.; funds through Brown University Predoctoral Training Program in Biological Data Science (NIH Grant T32 GM128596) supported M.M.B. The joint sixth authors were supported through the Big Data Summer Program funded by NIH Grant 5R25MD011714-03.

Author affiliations: [a]Department of Mathematics, San Francisco State University, San Francisco, CA 94132; [b]Ecology, Evolution and Organismal Biology, Brown University, Providence, RI 02912; [c]Department of Biology, San Francisco State University, San Francisco, CA 94132; and [d]Center for Computational and Molecular Biology, Brown University, Providence, RI 02912

1. I. Evett, B. S. Weir, *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists* (Oxford University Press, Inc., Oxford, UK, 1998).
2. J. M. Butler, Genetics and genomics of core short tandem repeat loci used in human identity testing. *J. Forensic Sci.* **51**, 253–265 (2006).
3. D. R. Hares, Expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* **6**, e52–e54 (2012).
4. E. E. Murphy, *Inside the Cell* (Nation Books, 2015).
5. H.R. Rep. No. 106-900(1), at 27, 2000 U.S.C.C.A.N. 2323 (2000), 2000 WL 1420163 (letter of Robert Raben, Assistant Attorney General, to the Honorable Henry J. Hyde, Chairman, House Judiciary Committee.
6. S. H. Katsanis, J. K. Wagner, Characterization of the standard and recommended CODIS markers. *J. Forensic Sci.* **58** (suppl. 1), S169–S172 (2013).
7. Federal Bureau of Investigation, CODIS-NDIS Statistics. https://le.fbi.gov/science-and-lab-resources/biometrics-and-fingerprints/codis/codis-ndis-statistics. Accessed 13 September 2022.
8. E. E. Murphy, J. Tong, The racial composition of forensic DNA databases. *Calif. Law Rev.* **108**, 19–54 (2019).
9. R. Roth, S. L. Ainsworth, If they hand you a paper, you sign it: A call to end the sterilization of women in prison. *Hastings Womens Law J.* **26**, 3 (2015)
10. S. Bauer, My four months as a private prison guard. Mother Jones (2016). https://www.motherjones.com/politics/2016/06/cca-private-prisons-corrections-corporation-inmates-investigation-bauer/. Accessed 13 September 2022.
11. M. Chesney-Lind, M. Mauer, Eds., *Invisible Punishment: The Collateral Consequences of Mass Imprisonment* (The New Press, New York, 2003).
12. A. Roth, Spit and acquit: Prosecutors as surveillance entrepreneurs. *Calif. Law Rev.* **107**. https://www.californialawreview.org/print/spit-and-acquit-prosecutors-as-surveillance-entrepreneurs/. Accessed 16 September 2022.
13. M. Cassidy, S.F. Police crime lab ends policy allowing investigators to match victim rape exam DNA to unrelated crimes. *San Francisco Chronicle*, 23 February. https://www.sfchronicle.com/bayarea/article/S-F-police-crime-lab-ends-policy-allowing-16942375.php. Accessed 13 September 2022.
14. J. Lynch, Not just San Francisco: Police across the country are retaining and searching DNA of victims and innocent people. *Electronic Frontier Foundation* (2022). https://www.eff.org/deeplinks/2022/02/not-just-san-francisco-police-across-country-are-retaining-and-searching-dna. Accessed 17 August 2022.
15. S. M. Mirkin, Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).
16. M. Gymrek *et al.*, Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
17. J. Quilez *et al.*, Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* **44**, 3750–3762 (2016).
18. S. F. Fotsing *et al.*, The impact of short tandem repeat variation on gene expression. *Nat. Genet.* **51**, 1652–1659 (2019).
19. A. Fujimoto *et al.*, Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *Genome Res.*, 10.1101/gr.255026.119 (2020).
20. I. Mitra *et al.*, Patterns of de novo tandem repeat mutations and their role in autism. *Nature* **589**, 246–250 (2021).
21. A. J. Hannan, Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**, 286–298 (2018).
22. M. D. Edge, B. F. B. Algee-Hewitt, T. J. Pemberton, J. Z. Li, N. A. Rosenberg, Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 5671–5676 (2017).
23. J. Kim, M. D. Edge, B. F. B. Algee-Hewitt, J. Z. Li, N. A. Rosenberg, Statistical detection of relatives typed with disjoint forensic and biomedical loci. *Cell* **175**, 848–858.e6 (2018).
24. J. Kim, N. A. Rosenberg, Record-matching of STR profiles with fragmentary genomic SNP data. bioXriv [Preprint] (2022). https://www.biorxiv.org/content/10.1101/2022.09.01.505545v1.full.pdf (Accessed 16 September 2022).
25. N. Wyner, M. Barash, D. McNevin, Forensic autosomal short tandem repeats and their potential association with phenotype. *Front. Genet.* **11**, 884 (2020).
26. F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, E. Eskin, Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
27. S. Saini, I. Mitra, N. Mousavi, S. F. Fotsing, M. Gymrek, A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nat. Commun.* **9**, 4397 (2018).
28. Federal Bureau of Investigation, CODIS and NDIS fact sheet . https://www.fbi.gov/resources/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet. Accessed 17 August 2022.
29. T. Lappalainen *et al.*; Geuvadis Consortium, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
30. D. Shriner *et al.*, Phenotypic variance explained by local ancestry in admixed African Americans. *Front. Genet.* **6**, 324 (2015).
31. H. Lango Allen *et al.*, Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
32. T. Willems *et al.*, Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* **14**, 590–592 (2017).
33. T. Konno *et al.*, Haploinsufficiency of CSF-1R and clinicopathologic characterization in patients with HDLS. *Neurology* **82**, 139–148 (2014).
34. L. Guo *et al.*, Bi-allelic CSF1R mutations cause skeletal dysplasia of dysosteosclerosis-Pyle disease spectrum and degenerative encephalopathy with brain malformation. *Am. J. Hum. Genet.* **104**, 925–935 (2019).
35. A. M. Nicholson *et al.*, CSF1R mutations link POLD and HDLS as a single disease entity. *Neurology* **80**, 1033–1040 (2013).
36. F. S. Eichler *et al.*, *CSF1R* mosaicism in a family with hereditary diffuse leukoencephalopathy with spheroids. *Brain* **139**, 1666–1672 (2016).
37. R. Rademakers *et al.*, Mutations in the colony stimulating factor 1 receptor (CSF1R) gene cause hereditary diffuse leukoencephalopathy with spheroids. *Nat. Genet.* **44**, 200–205 (2011).
38. N. Oosterhof *et al.*, Homozygous mutations in CSF1R cause a pediatric-onset leukoencephalopathy and can result in congenital absence of microglia. *Am. J. Hum. Genet.* **104**, 936–947 (2019).
39. P. K. Srivastava *et al.*, A systems-level framework for drug discovery identifies Csf1R as an anti-epileptic drug target. *Nat. Commun.* **9**, 3561 (2018).
40. R. Mancuso *et al.*; NIMA Consortium, CSF1R inhibitor JNJ-40346527 attenuates microglial proliferation and neurodegeneration in P301S mice. *Brain* **142**, 3243–3264 (2019).
41. A. Olmos-Alonso *et al.*, Pharmacological targeting of CSF1R inhibits microglial proliferation and prevents the progression of Alzheimer's-like pathology. *Brain* **139**, 891–907 (2016).
42. J. Sosna *et al.*, Early long-term administration of the CSF1R inhibitor PLX3397 ablates microglia and reduces accumulation of intraneuronal amyloid, neuritic plaque deposition and pre-fibrillar oligomers in 5XFAD mouse model of Alzheimer's disease. *Mol. Neurodegener.* **13**, 11 (2018).
43. V. Bellver-Landete *et al.*, Microglia are an essential component of the neuroprotective scar that forms after spinal cord injury. *Nat. Commun.* **10**, 518 (2019).
44. J. Zhang, L. Chang, Y. Pu, K. Hashimoto, Abnormal expression of colony stimulating factor 1 receptor (CSF1R) and transcription factor PU.1 (SPI1) in the spleen from patients with major psychiatric disorders: A role of brain-spleen axis. *J. Affect. Disord.* **272**, 110–115 (2020).
45. C. Shimamoto-Mitsuyama *et al.*, Lipid pathology of the corpus callosum in schizophrenia and the potential role of abnormal gene regulatory networks with reduced microglial marker expression. *Cereb. Cortex* **31**, 448–462 (2021).
46. M. J. Gandal *et al.*; PsychENCODE Consortium, Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, eaat8127 (2018).
47. J. M. Bullard, Y. C. Cai, L. L. Spremulli, Expression and characterization of the human mitochondrial leucyl-tRNA synthetase. *Biochim. Biophys. Acta* **1490**, 245–258 (2000).
48. S. B. Pierce *et al.*, Mutations in LARS2, encoding mitochondrial leucyl-tRNA synthetase, lead to premature ovarian failure and hearing loss in Perrault syndrome. *Am. J. Hum. Genet.* **92**, 614–620 (2013).
49. G. Soldà *et al.*, First independent replication of the involvement of LARS2 in Perrault syndrome by whole-exome sequencing of an Italian family. *J. Hum. Genet.* **61**, 295–300 (2016).
50. L. A. M. Demain *et al.*, Expanding the genotypic spectrum of Perrault syndrome. *Clin. Genet.* **91**, 302–312 (2017).
51. R. Li, M.-X. Guan, Human mitochondrial leucyl-tRNA synthetase corrects mitochondrial dysfunctions due to the tRNALeu(UUR) A3243G mutation, associated with mitochondrial encephalomyopathy, lactic acidosis, and stroke-like symptoms and diabetes. *Mol. Cell. Biol.* **30**, 2147–2154 (2010).

52. L. M. Boyden *et al.*, Mutations in KDSR cause recessive progressive symmetric erythrokeratoderma. *Am. J. Hum. Genet.* **100**, 978–984 (2017).
53. T. K. Bariana *et al.*, Sphingolipid dysregulation due to lack of functional KDSR impairs proplatelet formation causing thrombocytopenia. *Haematologica* **104**, 1036–1045 (2019).
54. T. Takeichi *et al.*, Biallelic mutations in KDSR disrupt ceramide synthesis and result in a spectrum of keratinization disorders associated with thrombocytopenia. *J. Invest. Dermatol.* **137**, 2344–2353 (2017).
55. B. F. B. Algee-Hewitt, M. D. Edge, J. Kim, J. Z. Li, N. A. Rosenberg, Individual identifiability predicts population identifiability in forensic microsatellite markers. *Curr. Biol.* **26**, 935–942 (2016).
56. A. Auton *et al.*; 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
57. B. L. Browning, Y. Zhou, S. R. Browning, A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
58. S. Saini, I. Mitra, N. Mousavi, S. F. Fotsing, M. Gymrek, Data from "A reference haplotype panel for genome-wide imputation of short tandem repeats." Amazon S3 bucket. http://gymreklab.com/2018/03/05/snpstr_imputation.html. Accessed 13 September 2022.
59. B. L. Browning, S. R. Browning, Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
60. L. Yun, C. Willer, S. Sanna, G. Abecasis, Genotype imputation. *Annu. Rev. Hum. Genet.* **10**, 387–406 (2009).
61. T. Lappalainen, *et al.*, Data from "Transcriptome and genome sequencing uncovers functional variation in humans." ArrayExpress. https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/. Accessed 13 September 2022.
62. I. Tirosh, N. Barkai, K. J. Verstrepen, Promoter architecture and the evolvability of gene expression. *J. Biol.* **8**, 95 (2009).
63. G. Levinson, G. A. Gutman, High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Res.* **15**, 5323–5338 (1987).
64. H. Ellegren, Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.* **5**, 435–445 (2004).
65. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440–9445 (2003).
66. F. Baumdicker *et al.*, Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* **220**, iyab229 (2022).
67. M. Kimura, T. Ohta, Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 2868–2872 (1978).
68. S. R. Browning *et al.*, Ancestry-specific recent effective population size in the Americas. *PLoS Genet.* **14**, e1007385 (2018).
69. T. Willems, *et al.*, Data from "Genome-wide profiling of heritable and de novo STR variations." GitHub. https://github.com/HipSTR-Tool/HipSTR-references/blob/master/human/hg19.hipstr_reference.bed.gz. Accessed 13 September 2022.
70. M. Haeussler *et al.*, The UCSC genome browser database: 2019 update. *Nucleic Acids Res.* **47** (D1), D853–D858 (2019).
71. J. Butler, STRBase: Overview of STR fact sheets. https://strbase.nist.gov/str_fact.htm. Accessed 18 August 2020.
72. NIST. Data from "Overview of STR Fact Sheets." NIST. https://strbase.nist.gov/str_fact.htm. Accessed 13 September 2022.
73. K. R. Rosenbloom *et al.*, ENCODE data in the UCSC genome browser: Year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
74. Encode. Data "DHS site locations." Encode. http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegDnaseClustered/wgEncodeRegDnaseClusteredV3.bed.gz. Accessed September 13, 2022.
75. T. Willems, M. Gymrek, G. Highnam, D. Mittelman, Y. Erlich; 1000 Genomes Project Consortium, The landscape of human STR variation. *Genome Res.* **24**, 1894–1904 (2014).
76. M. M. Banuelos *et al.*, CODISMarkers-Xprssn. GitHub. https://github.com/TalesOfDNA/CODISMarkers-Xprssn. Deposited 27 February 2020.