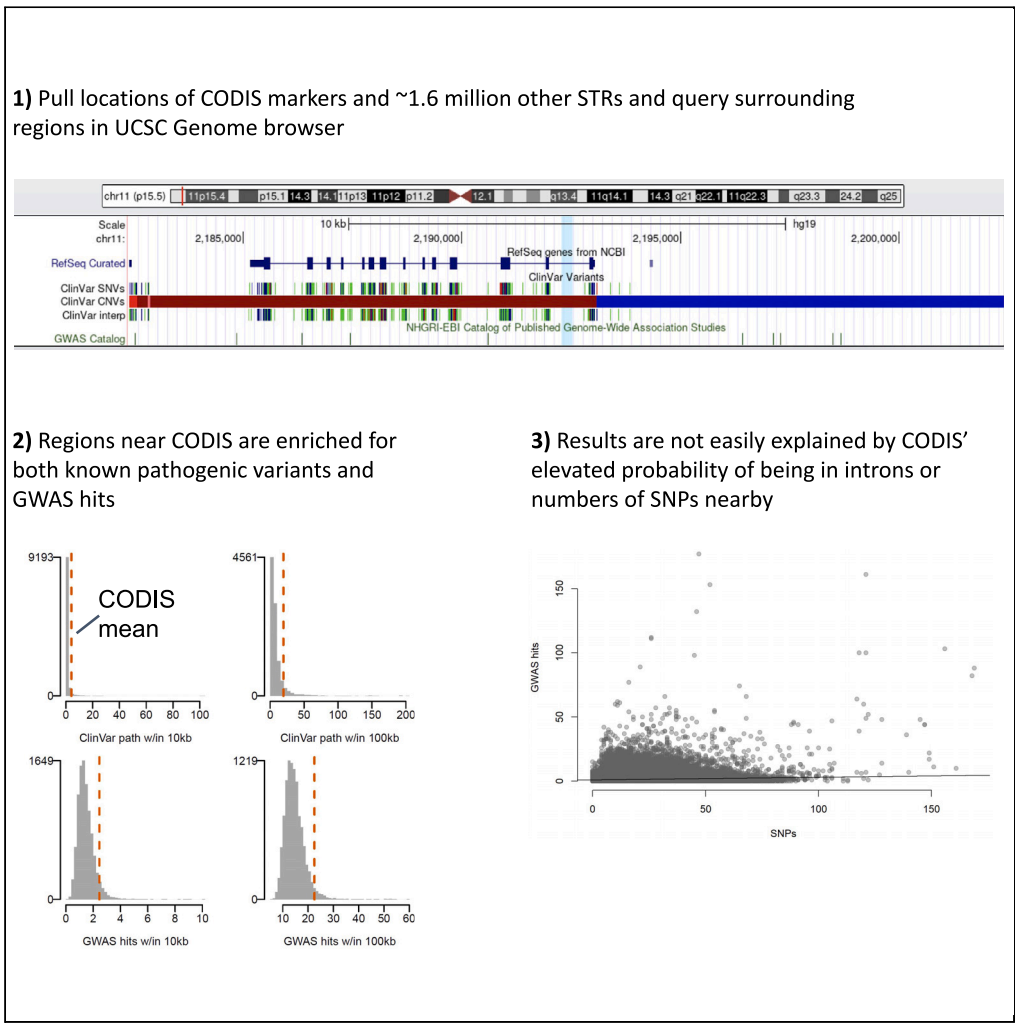
iScience

CellPress
OPEN ACCESS

Article

Microsatellites used in forensics are in regions enriched for trait-associated variants



Vivian Link, Yuómi Jhony A. Zavaleta, Rochelle-Jan Reyes, Linda Ding, Judy Wang, Rori V. Rohlfs, Michael D. Edge

rori@oregon.edu (R.V.R.)
edgem@usc.edu (M.D.E.)

Highlights

Regions near forensic loci are enriched for pathogenic variants and GWAS SNPs

The loci used in forensics are in regions enriched for DNAse I hypersensitivity sites

These regions are enriched compared with random sets of similar STRs

Link et al., iScience 26, 107992
October 20, 2023
https://doi.org/10.1016/j.isci.2023.107992

iScience

CellPress
OPEN ACCESS

Article

Microsatellites used in forensics are in regions enriched for trait-associated variants

Vivian Link,[1,4] Yuómi Jhony A. Zavaleta,[2,4] Rochelle-Jan Reyes,[2] Linda Ding,[1] Judy Wang,[1] Rori V. Rohlfs,[2,3,*] and Michael D. Edge[1,5,*]

SUMMARY

The 20 short tandem repeat (STR) loci of the combined DNA index system (CODIS) are the basis of the vast majority of forensic genetics in the United States. One argument for permissive rules about the collection of CODIS genotypes is that the CODIS loci are thought to contain little information about ancestry or traits. However, in the past 20 years, a growing field has identified hundreds of thousands of genotype-trait associations. Here, we conduct a survey of the landscape of such associations surrounding the CODIS loci as compared with non-CODIS STRs. Although this study cannot establish or quantify associations between CODIS genotypes and phenotypes, we find that the regions around the CODIS loci are enriched for both known pathogenic variants (> 90th percentile) and for trait-associated SNPs identified in genome-wide association studies (GWAS) ($\geq$ 95th percentile in 10kb and 100kb flanking regions), compared with other random sets of autosomal tetranucleotide-repeat STRs.

INTRODUCTION

DNA evidence has played a crucial role in forensic investigations for over three decades.[1–4] Beginning in the mid-1980s,[5] forensic practitioners realized that genotypes from even small numbers of genetic loci—provided that they are sufficiently heterozygous—can provide a nearly unique identifier that rules out the vast majority of people as the source of an unidentified sample. Many governments worldwide began to collect genotypes from highly variable short tandem repeat (STR, also called microsatellite) loci for the purpose of assisting forensic investigations. STR alleles differ from each other by virtue of containing different numbers of repeats of a short (generally 1–6 base pairs) motif sequence.[6] (STRs of the same length may also differ in their underlying sequence,[7] but distinct length classes are the basis for most forensic DNA work). Because many alleles are possible at each STR locus and STR mutation rates are high, STRs tend to be highly heterozygous.[8] As a result, small sets of STRs—relatively easily genotyped using technology available in the 1990s—can provide enough information to identify a person from a high-quality single-source DNA sample. Small sets of STRs remain the standard for forensic DNA practice in most countries.

In the United States (US), the Combined DNA Index System (CODIS) loci are the workhorse loci used in forensics. CODIS includes a set of 20 STR markers, 13 of which were established as the original set in the 1990s, and 7 of which were added in 2017.[9] Of the 20 CODIS STRs, 19 are tetranucleotide STRs (i.e., STRs with four-base-pair motifs), and one (D22S1045) is a trinucleotide STR. The X-linked amelogenin locus is also recorded and may be examined under more restricted circumstances. As of November 2022, CODIS genotypes from 21,791,620 people were accessible to law enforcement via the National DNA Index System (NDIS), and CODIS genotypes had been used as evidence in 622,955 investigations.[10]

The broad collection, storage, and use of CODIS genotypes is premised in part on the idea that the collection of one's CODIS genotypes entails only a minimal privacy incursion. When the CODIS set was expanded from 13 to 20 loci, an explicit goal was to avoid including loci that would allow prediction of disease.[9,11] The metaphor of a "DNA fingerprint," sometimes used to describe a person's CODIS genotypes, conveys this impression, and it has been invoked in legal decisions concerning the CODIS loci, for example the case of *Maryland v. King*, which permitted the collection of CODIS genotypes from arrestees.[12]

One piece of evidence that has been marshaled in defense of the claimed phenotypic irrelevance of the CODIS loci is that the CODIS markers themselves have not been associated with known traits. For example, ten years ago, Katsanis & Wagner[13] scoured the literature and found no record of direct associations between the CODIS loci and any known phenotypes. However, they did note that several of the CODIS loci are intragenic in genes with known phenotypic associations. It is perhaps unreasonable to expect much direct evidence of CODIS-trait associations given that STR loci are seldom tested for association with phenotype directly (but see ref.[14]), in part because

[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA
[2]Department of Biology, San Francisco State University, San Francisco, CA, USA
[3]Department of Data Science and Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA
[4]These authors contributed equally
[5]Lead contact
*Correspondence: rori@oregon.edu (R.V.R.), edgem@usc.edu (M.D.E.)
https://doi.org/10.1016/j.isci.2023.107992

STRs are less effective at "tagging" nearby causal variants than single-nucleotide polymorphisms[8,15] (SNPs). However, our knowledge of phenotypic associations has grown tremendously in the decade since Katsanis & Wagner's study, prompting a re-examination of their question, in line with calls for systematic reviews of trait information contained in CODIS loci.[16]

Here, we carry out a similar exercise to Katsanis & Wagner, searching widely used genomic databases to characterize the genomic neighborhoods of the CODIS loci. In addition to providing an update to Katsanis & Wagner's work, we extend it in four main ways. First, we examine the hundreds of thousands of known genotype-phenotype associations identified by genome-wide association study (GWAS),[17,18] particularly those loci near the CODIS loci. Second, we automate most of our procedures, facilitating replication of our work. Third, whereas Katsanis & Wagner considered only very short genomic regions around the CODIS loci (1 kilobase), we consider larger regions as well (10kb and 100kb). Though linkage disequilibrium (LD) between STRs and SNPs tends to be smaller than SNP-SNP LD, due largely to the high mutation rate of STRs, SNP-STR LD nonetheless extends over these larger regions,[8,15] making them relevant for investigation. Finally, Katsanis & Wagner considered only the 13 original CODIS loci and 11 loci suggested for inclusion, seven of which were added in 2017. Here, we consider STR loci across the genome, aggregating data (available as supplemental information) from approximately 1.6 million STRs. We focus our comparisons on 224,092 autosomal tetranucleotide-repeat STRs, as 19 of the 20 CODIS STRs have tetranucleotide repeat motifs.

RESULTS

We downloaded the locations of ~1.6 million STR regions from the hipSTR reference,[19] along with genome-wide annotations from the UCSC Genome Browser.[20] In particular, we downloaded coding gene locations from RefSeq,[21] SNP allele frequencies from HapMap[22] CEU, common SNP locations from dbSNP 153,[23] locations of phenotypically relevant variants from ClinVar,[24] trait-associated SNPs discovered in GWAS from the GWAS catalog,[25] and the locations of DNase I hypersensitivity clusters from ENCODE.[26] (For details, see STAR Methods).

We sought to describe the genomic neighborhoods of all 1.6 million STR regions identified in the hipSTR reference in terms of their density of key annotated features—in particular, of coding genes, common SNPs, trait-associated variants, and DNase I hypersensitivity sites. Before doing so, we preprocessed the feature data from UCSC as detailed in STAR Methods, with the goal of identifying manageable-sized sets of high-confidence features.

For all features and all STRs, we recorded the distance of the nearest feature to the STR midpoint, and the number of features within 1kb, 10kb, and 100kb of the STR midpoint. For the GWAS catalog, we kept track of the number of GWAS hits within each window size as well as the number of distinct associated traits after first narrowing to a set of widely studied traits.

The data processing and analysis scripts, written in R[27] version 1.4.2 and using the data.table package,[28] are available at https://github.com/edgepopgen/CODIS_proximity. The output file recording the features proximal to each STR is available as Data S1, and the file with all features recorded for each CODIS locus is available as Data S2.

Genetic neighborhoods of the CODIS loci

Table 1 shows the positions of the CODIS loci, the distance to the nearest gene, the names of genes within 100 kilobases (kb) of each locus, and the number of HapMap SNPs at minor allele frequency >1% in the CEU subset of the 1000 Genomes project within 10kb. Half of the 20 CODIS loci are intragenic, as noted previously.[13] Of the remaining 10 loci, 5 have protein-coding genes within 100kb. The CODIS locus with by far the greatest distance to the nearest protein-coding gene in RefSeq Select is D13S317, which is approximately 1.7 megabases (Mb) from the nearest gene. All CODIS loci are within 10kb of several SNPs common in people of European ancestries.

Table 2 gives information about pathogenic variants identified in ClinVar and GWAS hits within 10kb of each CODIS locus. Six of the ten intragenic CODIS loci are within 10kb of variants identified as pathogenic in ClinVar, ranging from two variants identified for CSF1PO to 25 for TH01. Sixteen of the 20 CODIS loci are within 10kb of at least one SNP identified as a GWAS hit, with TH01 again recording the most trait-associated nearby variants, with 10. TH01 is intragenic to the tyrosine hydroxylase gene *TH*, which plays an important role in synthesizing dopamine from its amino acid precursor, tyrosine.[29]

Comparisons with other autosomal tetranucleotide-repeat STRs

To place the properties of the CODIS loci in context, we compared them with the other 224,092 autosomal, tetranucleotide-repeat STRs in the hipSTR reference.[19] (Although one of the CODIS loci, D22S1045, is a trinucleotide-repeat locus, we focused our comparisons on tetranucleotide-repeat loci.) Figure 1 shows the distribution of the CODIS loci (orange) compared with non-CODIS autosomal tetranucleotide STRs (gray) with respect to their proximity to protein-coding genes, ClinVar pathogenic sites, GWAS hits, unique commonly studied traits associated with nearby GWAS hits, and DNase I hypersensitivity sites. For four of these feature categories, we show the distance to the nearest feature and the count of features within 1kb, 10kb, and 100kb. For commonly studied GWAS traits, we do not show the distance to the nearest feature. The figures suggest that the CODIS STRs are not systematically less informative about traits than non-CODIS STRs in any category, and in fact, the 10kb and 100kb windows surrounding the CODIS loci appear to harbor more trait-associated variants than average, as identified by ClinVar and the GWAS catalog.

Figure 2 shows, for the same features as in Figure 1, the mean of the CODIS loci (dashed orange line) compared with the means of 10,000 random sets of 20 tetranucleotide loci. The percentiles at which the CODIS average falls on each of these distributions, along with the distributions for transcription start site (TSS) and HapMap SNPs common in CEU, are shown in Table 3. Figure 2 and Table 3 confirm the visual

Table 1. Locations of the CODIS loci

Marker	Chr	Start position (approximate MB, hg19)	Distance to nearest protein-coding gene (0 = intragenic)	Protein-coding genes w/in 100kb, in proximity order	Common SNPs in Hapmap CEU w/in 10kb
D1S1656	1	230.9	0	*CAPN9, AGT, C1orf198, COG2*	58
TPOX	2	1.5	0	*TPO*	22
D2S441	2	68.2	29,159	*C1D*	22
D2S1338	2	218.9	11,910	*TNS1, RUFY4*	11
D3S1358	3	45.6	0	*LARS2, LIMD1*	7
FGA	4	155.5	0	*FGA, FGB, FGG, PLRG1, DCHS2*	16
D5S818	5	123.1	158,529		24
CSF1PO	5	149.5	0	*CSF1R, HMGXB3, PDGFRB, TIGD6, SLC26A2, CDX1*	36
D7S820	7	83.8	0	*SEMA3A*	19
D8S1179	8	125.9	78,404	*ZNF572*	19
D10S1248	10	131.1	172,971		39
TH01	11	2.2	0	*TH, INS, IGF2, ASCL2*	21
vWA	12	6.1	0	*VWF, ANO2*	33
D12S391	12	12.5	28,998	*MANSC1, LRP6, BORCS5*	32
D13S317	13	82.7	1,729,158		21
D16S539	16	86.4	157,803		59
D18S51	18	60.9	0	*BCL2, KDSR*	25
D19S433	19	30.4	15,972	*URI1*	22
D21S11	21	20.6	778,252		22
D22S1045	22	37.5	0	*IL2RB, TMPRSS6, C1QTNF6, SSTR3, KCTD17, RAC2*	38

impression from Figure 1. The CODIS loci, as a set, are unusually dense with nearby SNPs common in CEU, ClinVar variants marked pathogenic, and GWAS hits. For GWAS hits, the CODIS loci appear average in their number of hits within 1kb, but above the 90th percentile in the number of hits within 10kb or 100kb. At larger window sizes, the CODIS loci also appear to be in neighborhoods unusually dense in high-scoring DNase I hypersensitivity sites.

Comparing the CODIS loci with sets of random autosomal STRs irrespective of motif length from one to six (1,527,057 loci in the hipSTR reference) produces results very similar to those obtained for tetranucleotide-repeat STRs (Table S1; Figure S1).

We considered whether the unusually high number of GWAS hits and ClinVar pathogenic variants near the CODIS loci might be explained by other features of the CODIS loci. The CODIS loci are 50% intragenic (compared with 39% of non-CODIS tetranucleotide-repeat STRs), and intragenic loci might be expected to be nearer trait-associated variants than intergenic loci, assuming that genes are enriched for trait-associated variants. Further, the CODIS loci appear to be in genomic regions with unusually high numbers of SNPs common in people of European ancestry. Since such SNPs are the targets of association in GWAS studies, the high SNP density might in principle explain the high density of GWAS hits.

Table 4 shows Spearman correlations in the non-CODIS autosomal tetranucleotide STRs among intragenic status and the counts of the features in Table 3 (i.e., TSSs, genes, pathogenic variants, GWAS hits and traits, and DNase hypersensitivity sites) within 10kb. (Analogous information for 100kb windows is shown in Table S2.) Although intragenic STRs have somewhat more ClinVar pathogenic variants and GWAS hits within 10kb, the correlations between intragenic status and these features are not large (max Spearman rho = 0.22 for ClinVar pathogenic variants). Moreover, comparing the CODIS means to 10,000 random sets of non-CODIS tetranucleotide STRs matched for intragenic frequency (50%) produces a table of percentiles extremely similar to Table 3 (Table S3). The correlations between the number of nearby common SNPs and GWAS hits (or ClinVar pathogenic variants) are even smaller than those for intragenic status (Spearman's rho < 0.1), and in fact, they are mostly negative for counts within 100kb (Table S1), suggesting that the density of nearby SNPs does not explain the unusually high numbers of phenotypic associations near the CODIS loci.

DISCUSSION

We find that, in comparison with other autosomal tetranucleotide-repeat STRs, the CODIS loci are remarkably rich in nearby variants with known phenotypic associations. The most extreme example is TH01, which has the most known pathogenic variants within 10kb (25) and

Table 2. Phenotypic associations within 10kb of the CODIS loci from ClinVar and the GWAS catalog

Marker	ClinVar variants	ClinVar traits	GWAS hits	GWAS commonly studied traits
D1S1656	0		0	
TPOX	12	Deficiency of iodide peroxidase; Neurodevelopmental disorder	2	
D2S441	0		1	
D2S1338	0		1	Height
D3S1358	0		0	
FGA	22	Hepatocellular carcinoma; Congenital afibrinogenemia; Familial visceral amyloidosis,Ostertag type; Hypofibrinogenemia; Familial hypodysfibrinogenemia; Familial dysfibrinogenemia; Dysfibrinogenemia; Abnormal bleeding	4	Fibrinogen; Height; Ischemic stroke; Stroke; Venous thromboembolism
D5S818	0		3	Amyotrophic lateral sclerosis; Total body bone mineral density
CSF1PO	2	Brain abnormalities, neurodegeneration, and dysosteosclerosis	7	Aspartate aminotransferase levels; Monocyte count; Serum total protein level
D7S820	0		1	Obesity-related traits
D8S1179	0		3	Platelet count
D10S1248	0		0	
TH01	25	Permanent neonatal diabetes mellitus; not specified; Autosomal recessive DOPA responsive dystonia; Inborn genetic diseases; Dystonic disorder	10	Cystatin C levels; Height; Hematocrit; Hemoglobin; Hemoglobin concentration; Type 1 diabetes; Type 2 diabetes
vWA	17	von Willebrand disorder; von Willebrand disease type 3; Abnormality of coagulation; von Willebrand disease type 1	1	
D12S391	0		1	
D13S317	0		2	Hippocampal volume
D16S539	0		6	Appendicular lean mass; Optic cup area; Response to statin therapy
D18S51	0		2	Heel bone mineral density
D19S433	0		1	
D21S11	0		0	
D22S1045	4	Ichthyosis; Immunodeficiency 63 with lymphoproliferation and autoimmunity	4	Asthma; Eosinophil counts; Rheumatoid arthritis; Tuberculosis

also the most SNPs within 10kb implicated in GWAS studies (10). Almost 20 years ago, John Butler[30] wrote that "One core STR locus that has gotten a bad reputation over the years for supposed linkage to genetic diseases is TH01," going on to note the inconsistent nature of association evidence at the time. Although the work we report here cannot identify TH01 alleles in association with specific phenotypes, the relatively high density of variants annotated as pathogenic or identified as trait-associated in GWAS near TH01 is perhaps consistent with the reputation TH01 developed among forensic practitioners in the first decade of CODIS's use. After TH01, the loci with the most known pathogenic variants within 10kb were FGA (22) and vWA (17), and those with the most SNPs identified as trait-associated by GWAS within 10kb were CSF1PO (7) and D16S539 (6).

Although four of these five loci with the most evidence of possible trait association (all but D16S539) are intragenic, the unusual proximity of the CODIS to phenotype-associated variants is not explained by the fact that 50% of the CODIS loci are in intragenic regions (compared with 39% of non-CODIS tetranucleotide-repeat STRs). It is also not easily explained by the CODIS loci' closer proximity to SNPs with minor alleles common in people of European ancestries, since the density of such SNPs is not strongly associated with the presence of either known pathogenic variants or SNPs identified as trait-associated in GWAS.

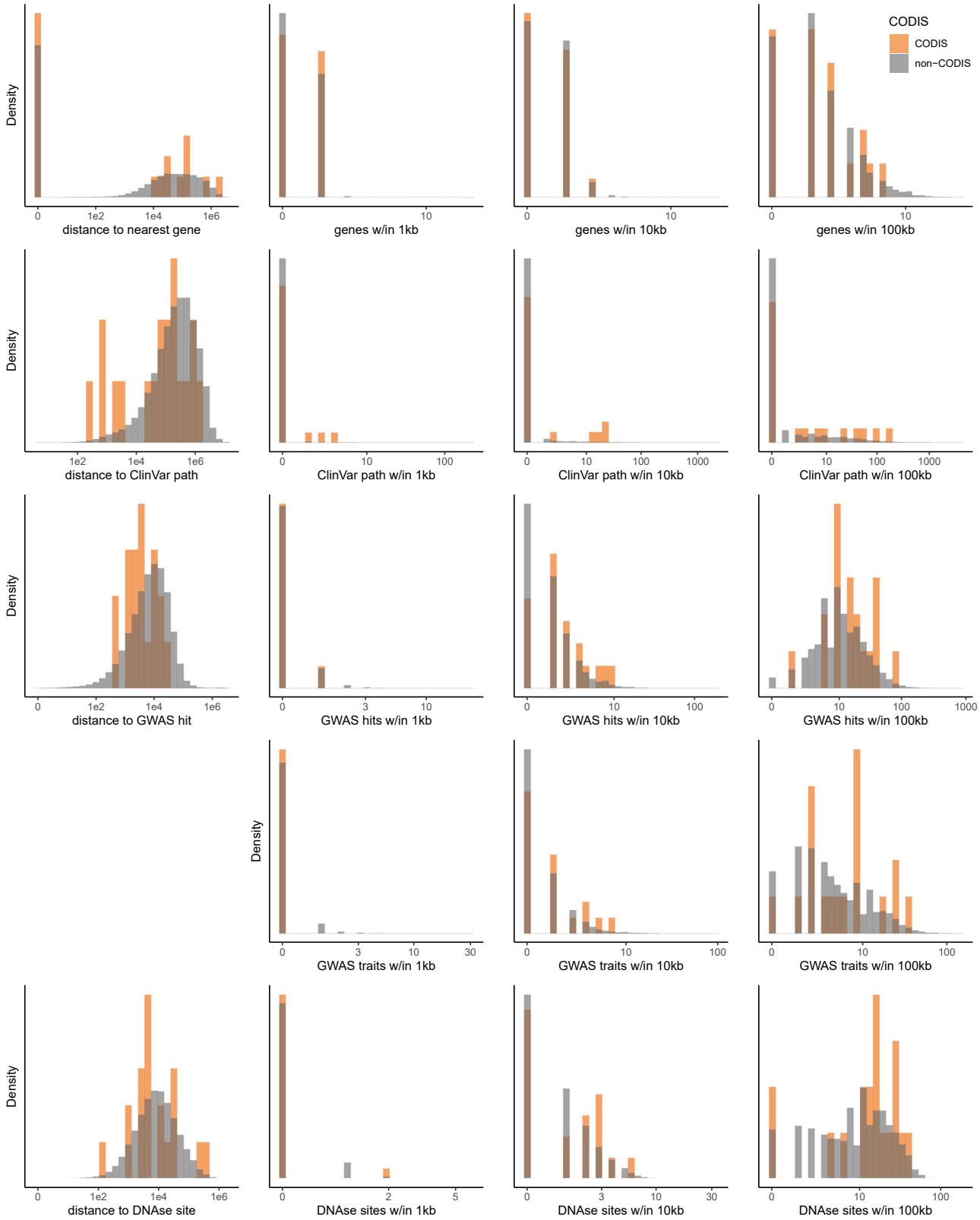

CODIS
CODIS
non−CODIS
Density
distance to nearest gene
genes w/in 1kb
genes w/in 10kb
genes w/in 100kb
distance to ClinVar path
ClinVar path w/in 1kb
ClinVar path w/in 10kb
ClinVar path w/in 100kb
distance to GWAS hit
GWAS hits w/in 1kb
GWAS hits w/in 10kb
GWAS hits w/in 100kb
GWAS traits w/in 1kb
GWAS traits w/in 10kb
GWAS traits w/in 100kb
distance to DNAse site
DNAse sites w/in 1kb
DNAse sites w/in 10kb
DNAse sites w/in 100kb

Figure 1. The values of the CODIS loci (orange histogram) compared with non-CODIS autosomal tetranucleotide-repeat STRs (gray) on variables relating to their proximity to phenotype-relevant features

The first column shows distance to the nearest feature, and the second through fourth columns show the number of features within 1kb, 10kb, and 100kb. The rows, in order, show genes included in the RefSeq Select set, variants annotated as pathogenic in ClinVar, SNPs identified as trait-associated in GWAS studies, traits included in at least 3 GWAS studies with associated variants nearby, and DNase I Hypersensitivity sites. The horizontal axes are displayed on a log scale; we added one to all values to avoid taking the logarithm of zero.

These results do not constitute direct evidence that the CODIS loci themselves are associated with any phenotypes. However, some degree of correlation (i.e., LD) is expected between STRs and SNP loci over these genomic distances.[8,15] Although the high mutation rates of STRs reduce their LD with surrounding SNPs, genetic drift continually generates LD that is slow to be removed by recombination or nullified by back mutations.[15] Direct evidence of whether the CODIS loci (or other STRs) are associated with, or causal for, phenotypes of interest is starting to appear.[6] We emphasize, however, that from the perspective of phenotype prediction, whether the CODIS loci are causal is not the central concern; any reproducible associations, even if they stem from LD with other causal loci, would still have some predictive utility.

These results add to other lines of evidence suggesting that the CODIS loci are not completely free of phenotypic or other genetic information. For example, the CODIS loci, on closer analysis, turn out to contain substantial ancestry information, despite their low values of F_{ST}.[31] Further, because the CODIS loci are correlated with—i.e. in LD with—surrounding SNP loci, it is sometimes possible to identify CODIS and genome-wide SNP genotypes as coming from the same individual, even when the sets of loci in the two datasets are disjoint.[32,33] Most recently, direct examination of the CODIS loci provides suggestive evidence that some of them are associated with gene expression levels in some tissues.[34]

More direct tests of the presence and importance of associations between CODIS (and other STR) genotypes and specific phenotypes would require high-quality STR genotypes in large samples of people, such as those typically used for GWAS. Testing for association between the mean imputed length of CODIS alleles and phenotype, as in previous work,[34] might be expected to be most informative in cases where STR length has a monotonic, causal influence on the phenotype. However, LD information is not well-preserved in these data because distinct genotypes are collapsed, leading to a loss of association signal due to LD with neighboring causal variants.

To be clear, the accuracy of phenotype predictions from the CODIS loci is not expected to be high in absolute terms for most phenotypes. The ability to predict a trait from genotype is limited by the trait's heritability,[35] and for a wide range of complex traits, the best current predictions from even genome-wide SNP data are not particularly accurate.[36] A small set of STRs will not outperform genome-wide SNPs at phenotype prediction except in rare cases. In general, whether the phenotype predictions developed directly from CODIS represent privacy incursions will depend on at least (a) the standard for how accurate prediction needs to be considered a privacy incursion, (b) the number and effect sizes of causal alleles in or near the CODIS loci, and (c) the degree to which a trait is associated with ancestry, which can be noisily reconstructed from CODIS genotypes.[31] What is clear is that the CODIS loci are not likely to be less informative about phenotypes than other, similar loci. This statement is analogous to the one made by Algee-Hewitt and colleagues,[31] who found that the CODIS loci are no less informative about ancestry than comparison loci.

It is not clear why the regions around the CODIS loci are unusually dense with phenotypic associations. The GWAS era had not yet begun at the time when the CODIS loci were selected. One possibility is simply bad luck—the original architects of the CODIS system happened to choose sites that would later be identified as near phenotype-associated sites. A second possibility is a form of historical ascertainment bias—because the CODIS loci were drawn from sets of STR loci used in linkage analysis, perhaps more is known about these regions than other regions. But this possibility does not as easily explain enrichment of nearby GWAS hits, since the design principles of SNP panels studied in GWAS make no reference to the locations of STRs used in linkage analyses.[37] Another possibility is that there is some other feature or set of features of the CODIS loci that led to their being considered favorably by the designers of CODIS and that also meant they would be near sites with trait associations, or at least sites that were liable to be discovered as trait-associated. One possible clue relevant to identifying such a feature is the enrichment of high-signal DNase I hypersensitivity sites near the CODIS loci that we observed. DNase I sites are a hallmark of accessible chromatin, and have been relied upon in searches for regulatory elements, including enhancers and promoters.[38] Chromatin accessibility may also influence the ease of PCR amplification of STRs. Because ease of genotyping by PCR was a factor in the initial selection of the CODIS loci,[30] it is possible that the CODIS loci are more likely to be near regulatory elements. Future work may consider this possibility.

In *Maryland v. King*,[12] Justice Kennedy wrote for the majority that the CODIS loci "come from noncoding parts of the DNA that do not reveal the genetic traits of the arrestee." This statement was part of the majority's argument that CODIS genotypes can be thought of as a "DNA fingerprint," a piece of information useful for identification but not informative about any of a person's traits or medical information. It followed for the majority that collection and storage of CODIS genotypes, like that of fingerprints, is an appropriate part of a routine pre-trial booking procedure. It is not obvious how much information about other traits the CODIS loci would need to convey in order to invalidate the Court's premise, nor is it yet clear how much information they actually do convey. At the same time, it appears that any attempt to choose loci for CODIS that convey unusually small amounts of information about phenotypes compared with other STRs does not seem to have been successful.

An acknowledgment that CODIS genotypes may be more revealing than previously assumed may prompt a rethinking of the patchwork of highly variable local practices governing CODIS genotype collection, storage, and access[39–41] and influence considerations regarding universal forensic DNA databases.[42] We advocate, along with Kaye,[16] that biomedical literature continue to be monitored in order to ascertain the phenotypic information accessible to a person with access to CODIS profiles.[14,34] More generally, we advocate that practices surrounding

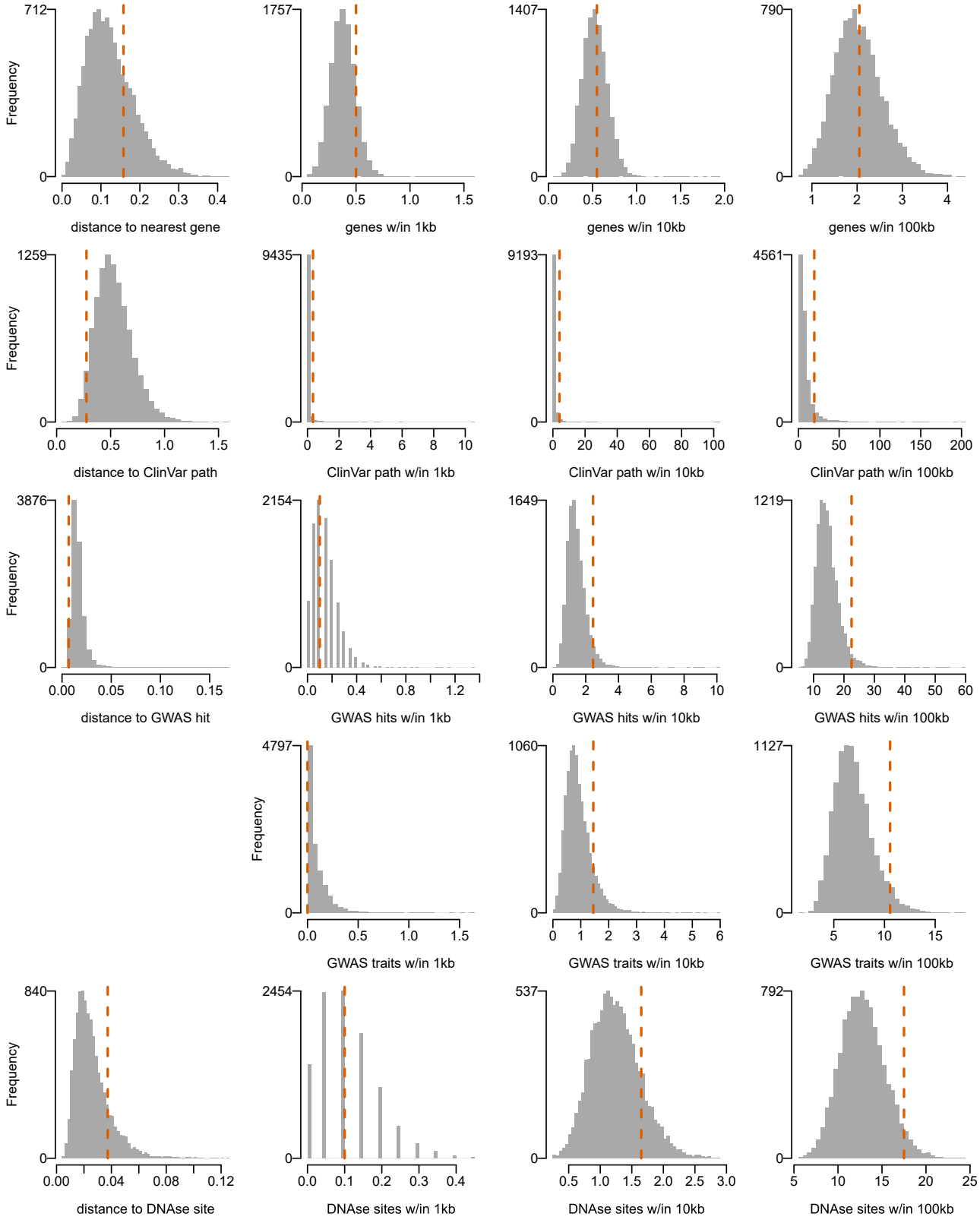


Figure 2. The mean of the 20 CODIS loci (dashed orange line) compared with random sets of 20 non-CODIS autosomal tetranucleotide-repeat loci

The variables shown are the same as in Figure 1. See also Figure S1.

Table 3. Percentiles of the CODIS loci as a set compared with 10,000 random sets of 20 tetranucleotide autosomal STRs

	Proximity to nearest[a]	w/in 1kb	w/in 10kb	w/in 100kb
RefSeq Select TSS	50.5	96.9	77.4	67.1
RefSeq Select gene	26.2	86.1	57.7	54.2
HapMap common SNPs in CEU	99.9	97.2	99.7	99.0
ClinVar pathogenic variants	96.1	97.0	97.4	92.2
GWAS hits	98.9	48.6	94.7	96.7
GWAS well-studied traits	-	22.7	87.7	95.6
DNase I Hypersensitivity sites	16.1	62.8	85.2	96.0

See also Tables S1 and S3.
[a]"Proximity" percentile is 100 minus the "distance" percentile.

CODIS profiles should be informed by a framework that considers CODIS genotypes not as isolated pieces of information but as components of a genome connected via LD produced by recombination, mutation, and our shared evolutionary history.[32,33]

Limitations of the study

This study is limited by ascertainment biases present in the various databases we considered. To take one example, the GWAS catalog is a function of the actual associations identified in GWAS, which means that associations with widely studied traits, with SNPs included in or well--imputed by genotyping arrays commonly used for GWAS, and associations that are more easily detectable in people of European ancestries are more likely to be included. Our data processing procedures, which aimed mainly to arrive at simple summaries of high-confidence features, may also have introduced additional ascertainment biases. Another limitation is that we cannot estimate the actual association between STRs and traits, merely the positions of trait-associated variants nearby.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data sources
  - Data processing
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107992.

Table 4. Spearman correlations among key measurements for non-CODIS tetranucleotide STRs (within 10kb)

	IG	SNPs	TSS	Genes	CV vars	GWAS hits	GWAS traits
Intragenic status	1						
HapMap common SNPs in CEU	−.05	1					
RefSeq Select TSS	.06	−.16	1				
RefSeq Select genes	.77	−.13	.47	1			
ClinVar pathogenic variants	.22	−.05	.16	.29	1		
GWAS hits	.09	.08	.13	.16	.10	1	
GWAS well-studied traits	.09	.01	.15	.18	.10	.80	1
DNase I Hypersensitivity sites	.06	−.05	.35	.24	.10	.21	.22

See also Table S2.

ACKNOWLEDGMENTS

MDE is funded by NIH grant R35 GM137758. Y.J.A.Z. was supported by the NIH Bridges Fellowship (R25-GM048972) and a Genentech Foundation Fellowship. We thank Andy Clark for suggesting chromatin accessibility as a hypothesis for the co-occurrence of CODIS loci and phenotype-associated variants, and we thank the anonymous reviewers for helpful comments on the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, M.D.E. and R.V.R.; Methodology, M.D.E. and R.V.R.; Software, M.D.E., Y.J.A.Z., and L.D.; Data Analysis, M.D.E., V.L., Y.J.A.Z., R-J.R., and J.W.; Writing - Original Draft, M.D.E.; Writing - Review and Editing V.L., Y.J.A.Z., R-J.R., J.W., L.D.; Visualization, M.D.E. and R.V.R.; Supervision M.D.E., R.V.R., and V.L..

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 8, 2023
Revised: August 10, 2023
Accepted: September 18, 2023
Published: September 21, 2023

REFERENCES

1. Butler, J.M. (2015). The future of forensic DNA analysis. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140252. https://doi.org/10.1098/rstb.2014.0252.
2. Jobling, M.A., and Gill, P. (2004). Encoded evidence: DNA in forensic analysis. Nat. Rev. Genet. *5*, 739–751. https://doi.org/10.1038/nrg1455.
3. Kayser, M., and de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. Nat. Rev. Genet. *12*, 179–192. https://doi.org/10.1038/nrg2952.
4. Roewer, L. (2013). DNA fingerprinting in forensics: past, present, future. Investig. Genet. *4*, 22. https://doi.org/10.1186/2041-2223-4-22.
5. Gill, P., Jeffreys, A.J., and Werrett, D.J. (1985). Forensic application of DNA 'fingerprints. Nature *318*, 577–579. https://doi.org/10.1038/318577a0.
6. Gymrek, M. (2017). A genomic view of short tandem repeats. Curr. Opin. Genet. Dev. *44*, 9–16. https://doi.org/10.1016/j.gde.2017.01.012.
7. Gettings, K.B., Aponte, R.A., Vallone, P.M., and Butler, J.M. (2015). STR allele sequence variation: Current knowledge and future issues. Forensic Sci. Int. Genet. *18*, 118–130. https://doi.org/10.1016/j.fsigen.2015.06.005.
8. Willems, T., Gymrek, M., Highnam, G., 1000 Genomes Project Consortium, Mittelman, D., and Erlich, Y. (2014). The landscape of human STR variation. Genome Res. *24*, 1894–1904. https://doi.org/10.1101/gr.177774.114.
9. Hares, D.R. (2015). Selection and implementation of expanded CODIS core loci in the United States. Forensic Sci. Int. Genet. *17*, 33–34. https://doi.org/10.1016/j.fsigen.2015.03.006.
10. FBI (2022). CODIS NDIS statistics. https://le.fbi.gov/science-and-lab-resources/biometrics-and-fingerprints/codis/codis-ndis-statistics.
11. Hares, D.R. (2012). Expanding the CODIS core loci in the United States. Forensic Sci. Int. Genet. *6*, e52–e54. https://doi.org/10.1016/j.fsigen.2011.04.012.
12. Maryland v. King. (2013). https://supreme.justia.com/cases/federal/us/569/435/#465-66.
13. Katsanis, S.H., and Wagner, J.K. (2013). Characterization of the Standard and Recommended CODIS Markers. J. Forensic Sci. *58*, S169–S172. https://doi.org/10.1111/j.1556-4029.2012.02253.x.
14. Wyner, N., Barash, M., and McNevin, D. (2020). Forensic Autosomal Short Tandem Repeats and Their Potential Association With Phenotype. Front. Genet. *11*, 884.
15. Payseur, B.A., Place, M., and Weber, J.L. (2008). Linkage Disequilibrium between STRPs and SNPs across the Human Genome. Am. J. Hum. Genet. *82*, 1039–1050. https://doi.org/10.1016/j.ajhg.2008.02.018.
16. Kaye, D.H. (2014). Open to Dispute": CODIS STR Loci as Private Medical Information (Penn State Law Res. Pap). No 23-2014.
17. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012. https://doi.org/10.1093/nar/gky1120.
18. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.
19. Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., and Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods *14*, 590–592. https://doi.org/10.1038/nmeth.4267.
20. Lee, C.M., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., Nassar, L.R., Powell, C.C., et al. (2020). UCSC Genome Browser enters 20th year. Nucleic Acids Res. *48*, D756–D761. https://doi.org/10.1093/nar/gkz1012.
21. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189.
22. International HapMap Consortium, Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The International HapMap Project. Nature *426*, 789–796. https://doi.org/10.1038/nature02168.
23. Sherry, S.T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311. https://doi.org/10.1093/nar/29.1.308.
24. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862–D868. https://doi.org/10.1093/nar/gkv1222.
25. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901. https://doi.org/10.1093/nar/gkw1133.
26. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.
27. R Core Team (2021). R A Language and Environment for Statistical Computing.
28. Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., and Parsonage, H. (2019). Package 'data. Table': Extension of 'data. Frame'.
29. Nagatsu, T., Nakashima, A., Ichinose, H., and Kobayashi, K. (2019). Human tyrosine hydroxylase in Parkinson's disease and in related disorders. J. Neural. Transm. *126*,

397–409. https://doi.org/10.1007/s00702-018-1903-3.

30. Butler, J.M. (2006). Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing. J. Forensic Sci. *51*, 253–265. https://doi.org/10.1111/j.1556-4029.2006.00046.x.
31. Algee-Hewitt, B.F.B., Edge, M.D., Kim, J., Li, J.Z., and Rosenberg, N.A. (2016). Individual Identifiability Predicts Population Identifiability in Forensic Microsatellite Markers. Curr. Biol. *26*, 935–942. https://doi.org/10.1016/j.cub.2016.01.065.
32. Edge, M.D., Algee-Hewitt, B.F.B., Pemberton, T.J., Li, J.Z., and Rosenberg, N.A. (2017). Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. Proc. Natl. Acad. Sci. USA *114*, 5671–5676. https://doi.org/10.1073/pnas.1619944114.
33. Kim, J., Edge, M.D., Algee-Hewitt, B.F.B., Li, J.Z., and Rosenberg, N.A. (2018). Statistical Detection of Relatives Typed with Disjoint Forensic and Biomedical Loci. Cell *175*, 848–858.e6. https://doi.org/10.1016/j.cell.2018.09.008.
34. Bañuelos, M.M., Zavaleta, Y.J.A., Roldan, A., Reyes, R.-J., Guardado, M., Chavez Rojas, B., Nyein, T., Rodriguez Vega, A., Santos, M., Huerta-Sanchez, E., and Rohlfs, R.V. (2022). Associations between forensic loci and expression levels of neighboring genes may compromise medical privacy. Proc. Natl. Acad. Sci. USA *119*, e2121024119. https://doi.org/10.1073/pnas.2121024119.
35. Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era — concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266. https://doi.org/10.1038/nrg2322.
36. Thompson, D.J., Wells, D., Selzam, S., Peneva, I., Moore, R., Sharp, K., Tarran, W.A., Beard, E.J., Riveros-Mckay, F., Giner-Delgado, C., et al. (2022). UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. Preprint at medRxiv. https://doi.org/10.1101/2022.06.16.22276246.
37. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. *74*, 106–120.
38. Chen, A., Chen, D., and Chen, Y. (2018). Advances of DNase-seq for mapping active gene regulatory elements across the genome in animals. Gene *667*, 83–94. https://doi.org/10.1016/j.gene.2018.05.033.
39. Joh, E.E. (2015). The Myth of Arrestee DNA Expungement. U. Pa. L. Rev. Online *164*, 51.
40. Murphy, E., and Tong, J.H. (2020). The racial composition of forensic DNA databases. Calif. Rev. *108*, 1847.
41. Roth, A. (2019). Spit and Acquit. Calif. Law Rev. *107*, 405–458.
42. Miller, S., and Smith, M. (2022). Quasi-Universal Forensic DNA Databases. Crim. Justice Ethics *41*, 238–256. https://doi.org/10.1080/0731129X.2022.2141021.
43. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591. https://doi.org/10.1038/s41588-019-0379-x.
44. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164. https://doi.org/10.1038/538161a.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
hipSTR reference (STR locations)	http://webstr.ucsd.edu/downloads, direct link https://github.com/HipSTR-Tool/HipSTR-references/blob/master/human/hg19.hipstr_reference.bed.gz	
UCSC Genome Browser	https://genome.ucsc.edu/	
Preprocessed data files	Supplemental information for this publication	
Software and algorithms		
R	R Project for Statistical Computing, https://www.r-project.org/	
Custom preprocessing and analysis scripts	github.com/edgepopgen/CODIS_proximity	

RESOURCE AVAILABILITY

Lead contact

Further questions and requests should be directed to and will be fulfilled by the lead contact, Michael "Doc" Edge (edgem@usc.edu).

Materials availability

This study did not generate new reagents.

Data and code availability

All data used in this study are publicly available as detailed in STAR Methods. Intermediate (processed) data files are available as supplementary files. All original code has been deposited at Github, as listed in the results and key resources table. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data sources

In January 2023, we downloaded the locations of ~1.6 million STR regions from the hipSTR reference (http://webstr.ucsd.edu/downloads, direct link https://github.com/HipSTR-Tool/HipSTR-references/blob/master/human/hg19.hipstr_reference.bed.gz).[19] We also downloaded a set of genome-wide annotations from the UCSC Genome Browser[20] using the DataIntegrator tool. In particular, we downloaded coding gene locations (Genes and Gene Predictions > NCBI Refseq > RefSeq All and Genes and Gene Predictions > NCBI Refseq > RefSeq Select) from RefSeq,[21] SNP allele frequencies from HapMap[22] CEU (Variation > HapMap SNPs … > HapMap SNPs CEU), common SNP locations from dbSNP 153[23] (Variation > dbSNP Archive - dbSNP 153 … > Variants), locations of phenotypically relevant variants (Phenotype and Literature > ClinVar Variants … > ClinVar SNVs) from ClinVar,[24] trait-associated SNPs discovered in GWAS (Phenotype and Literature > GWAS Catalog) from the GWAS catalog,[25] and the locations of DNase I hypersensitivity clusters (Regulation > ENCODE Regulation - DNase Clusters V3) from ENCODE.[26]

All genomic locations were expressed in hg19/GRCh37 coordinates for consistency with the hipSTR reference.

Data processing

We preprocessed the feature data downloaded from the UCSC Genome Browser in various ways. Preprocessing scripts are available at https://github.com/edgepopgen/CODIS_proximity. The output file recording the features proximal to each STR is available as Data S1, and the file with all features recorded for each CODIS locus is available as Data S2.

For coding gene locations, we used the RefSeq Select set, which contains one entry per curated coding gene (21,432 genes). We also located the transcription start site (TSS) of each gene as either the start or end coordinate of transcription, depending on whether the gene was annotated on the + (TSS = start) or - (TSS = end) strand. To identify SNPs common in people of European ancestries, heavily represented in GWAS,[43,44] we filtered to SNPs with minor allele frequencies of 1% or larger in the HapMap CEU data, reducing the number of variants from 4,029,798 to 2,705,918. We retained only ClinVar variants classified as "Pathogenic," reducing from 1,491,509 variants to 113,412.

For DNase I hypersensitivity sites, we limited to sites with the highest signal level (score 1000/1000), reducing the number of sites from 1,949,038 to 160,870.

We preprocessed the GWAS catalog in two distinct ways. The GWAS catalog contains one row per unique combination of SNP locus (rsID), study (PubMed ID), and trait, for a total of 392,271 entries. To obtain information about the number of SNPs identified as trait-associated in any GWAS, we first filtered the GWAS catalog to contain only one row per SNP locus, reducing to 183,014 rows. Thus, for counts of numbers of GWAS hits, each SNP rsID counts only once, regardless of how many studies identified it, and regardless of how many traits it was associated with. Next, we sought to identify traits with nearby GWAS associations for each STR. The trait identifiers in the GWAS catalog are not standardized, and many similar traits receive distinct names (for example "HDL cholesterol" and "HDL cholesterol levels" or "Mean corpuscular hemoglobin" and "Mean corpuscular hemoglobin concentration"). To reduce this redundancy and focus on commonly studied traits when counting the number of distinct traits near each STR, we limited to traits with associations reported in at least three distinct studies with the exact same trait name. This reduced the number of traits from 10,399 to 493, with 146,039 of the previously identified 183,014 unique GWAS-identified SNPs associated with the reduced set of traits.

For all features and all STRs, we recorded the distance of the nearest feature to the STR midpoint, and the number of features within 1kb, 10kb, and 100kb of the STR midpoint. For coding gene locations, we kept track of the distance to the nearest gene (defined as the distance to the start or end of transcription, whichever is shorter, or 0 if the STR is intragenic) and the nearest TSS separately. For the GWAS catalog, we kept track of the number of GWAS hits within each window size as well as the number of distinct associated traits (where again, distinctness merely means a non-identical character string). Because of the large size of the dbSNP common variants catalog, we recorded these locations only for the 20 CODIS loci. Additionally, for the CODIS only, we recorded the names of the traits reported as associated in ClinVar and the GWAS catalog, as well as the names of nearby protein-coding genes.

QUANTIFICATION AND STATISTICAL ANALYSIS

To assess whether the CODIS loci are unusual with respect to the features we studied, we compared the mean values of the CODIS loci on the studied features with those of many randomly selected sets of non-CODIS STRs, as described in the results.

In Tables 4 and S2, we computed Spearman correlations among the mean values of the features we studied for random sets of non-CODIS loci.