JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, 2012
doi: 10.1111/j.1556-4029.2012.02253.x
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

CRIMINALISTICS; JURISPRUDENCE

Sara H. Katsanis,[1] *M.S. and Jennifer K. Wagner,*[2] *J.D., Ph.D.*

Characterization of the Standard and Recommended CODIS Markers*

ABSTRACT: As U.S. courts grapple with constitutional challenges to DNA identification applications, judges are resting legal decisions on the fingerprint analogy, questioning whether the information from a DNA profile could, in light of scientific advances, reveal biomedically relevant information. While CODIS loci were selected largely because they lack phenotypic associations, how this criterion was assessed is unclear. To clarify their phenotypic relevance, we describe the standard and recommended CODIS markers within the context of what is known currently about the genome. We characterize the genomic regions and phenotypic associations of the 24 standard and suggested CODIS markers. None of the markers are within exons, although 12 are intragenic. No CODIS genotypes are associated with known phenotypes. This study provides clarification of the genomic significance of the key identification markers and supports—independent of the forensic scientific community—that the CODIS profiles provide identification but not sensitive or biomedically relevant information.

KEYWORDS: forensic science, genetic identity, DNA typing, forensic genetics, Combined DNA Index System, short tandem repeats

The culmination of the 1996–1997 STR Project was the selection of the 13 core CODIS markers, all highly polymorphic tetra-nucleotide short tandem repeats (STRs). In 2010, the U.S. Federal Bureau of Investigation revisited the panel composition, creating the CODIS Core Loci Working Group to consider the expansion of the core CODIS marker panel to minimize the likelihood of adventitious matches, improve international compatibility for data sharing, and improve the discriminatory power for missing persons cases and familial searching (1). This culminated in the proposal of an additional 11 STRs to be used alternatively in various identification contexts (1). Several criteria were considered in selecting markers for an expanded panel, the first of which being that they have "[n]o known association with medical conditions or defects" (1, p. e52, 2, p. 1). The primary rationale behind the emphasis on the development of panels that contain no association with biomedically relevant phenotypes is clear, as the statutory authority for CODIS itself (the DNA Identification Act of 1994 [3]; DNA Analysis Backlog Elimination Act of 2000 [4]; Justice for All Act of 2004 [5]; and the DNA Fingerprinting Act of 2005 [6]) is restricted to identification purposes. The Department of Justice has reiterated that CODIS profiles are to be "sanitized 'genetic fingerprints' that can be used to identify an individual uniquely, but do not disclose an individual's traits, disorders, or dispositions" (7). Thus, the rationale behind the criterion requires little explanation by the Working Group. On the other hand, the criterion used by the Working Group for the selection and ranking of the markers is unclear, and the literature offers little information relevant to whether (and the extent to which) any of these markers are causally related to phenotypes (1,2). Moreover, a quick review of the literature of linkage analyses and genome wide association studies (GWAS) may yield deceptive and exaggerated reports of linkage with some of these markers, because the number of reports may be simply a relic of the convenient markers' inclusion in commonly used linkage screening panel sets—such as the Marshfield linkage maps (8,9)—and, thus, the results may not be indicative of any actual causal relationship or biological function (10).

Motivated by recent court opinions (11) in which judges (frequently referencing popular science articles [12]) call the phenotypic irrelevance of the CODIS profile into question, we seek to clarify the role of phenotype in the selection criteria of markers. A myriad of important criticisms of forensic DNA analysis including the legal considerations and implications of molecular photofitting and phenotyping (13), the history and substance of criticisms aimed at forensic identification using blood grouping, HLA testing, and more recent methods, the suitability of the fingerprint analogy (which has considerable legal importance), and the appropriateness of selected CODIS markers (2) are each worthy of discussion. The focus of this technical note, however, is to address only the feasibility of creating a DNA database restricted to identity information. As the forensic community grapples with the technical and statistical benefits of the original 13 CODIS loci and the additional 11 loci for prioritization and selection, here, we examine the role of the markers and their regions as elements of the human genome. The selection of markers for identification is an ongoing process varying by population and regulatory control; as such, some markers (e.g.,

[1]Genome Ethics, Law & Policy, Duke Institute for Genome Sciences & Policy, Duke University, 304 Research Drive, Box 90141, Durham, NC 27708.

[2]Division of Translational Medicine and Human Genetics, Center for the Integration of Genetic Healthcare Technologies, Perelman School of Medicine, University of Pennsylvania, 1112 Penn Tower, 399 S. 34th St., Philadelphia, PA 19104.

*Partially funded by grant number P50HG004487-05 from the National Human Genome Research Institute (NHGRI).

Received 14 Mar. 2012; and in revised form 18 June 2012; accepted 30 June 2012.

D6S1043, D14S1434) relevant in select sub-populations are not reviewed herein.

Methods

We used the UCSC Genome Browser (build GRCh37/hg19) (14) to analyze each STR region. We conducted BLAT searches (15) of primers from each STR to locate the precise region of the repeat. We collected data on (i) phenotype and disease associations (GAD view-pack; DECIPHER-full; Online Mendelian Inheritance in Man [OMIM] AV SNPs-full; OMIM genes-full; OMIM pheno loci-full; GWAS catalog-full; RGD human quantitative trait loci [QTL]-full); (ii) genes and gene prediction tracks (UCSC genes-pack; RefSeq-dense); (iii) mRNA and EST tracks (human mRNAs-pack; spliced ESTs-pack); (iv) variation and repeats (common SNPs(132)-pack; simple repeats-pack; microsatellites-full); and (v) regulation (ENCODE Regulation-show; ENC RNA Binding-show; ORegAnno-full; Vista Enhancers-full). We used Ensembl (16) to determine intronic regions and note any reported phenotypic associations for STR genotypes. Disorders associated or linked within 1 kb of the STR were noted; chromosomal anomalies were not noted. To examine the potential relevance of the markers as noncoding genomic elements, we examined sequences for predicted enhancers and noted RNA-binding protein sites as well as predicted DNase I hypersensitivity and transcription binding sites. SNPs documented in dbSNP (build 132) overlapping and linked within 1 kb of the STR were noted. SNPs from the 1000 Genome Project were searched in SNPedia. If the STR was within a gene locus, we noted the gene name and examined the positioning with regard to the surrounding exons. Extragenic STR regions were examined to document proximity to the nearest transcript. We subsequently searched relevant genes in Database of Genotypes and Phenotypes (dbGaP) (17), OMIM (18), and GeneTests (19) to confirm genetic associations and document the availability of a genetic test for any related gene. We also examined the Marshfield linkage maps to determine which markers are used in the human genetic screening panels (20). All searches were conducted in October 2011.

Results and Discussion

Individual genotypes of the 24 STRs were not found to be associated with any documented phenotypes (note exception: DYS391 is on the Y chromosome, which if present in a DNA profile may indicate male sex). None of the 24 STRs are located within protein-coding exons (see Table 1) (see also Ref. 10). Two of the STRs (VWA and D12S391) are colocated on the same arm of chromosome 12 (12p13) within 6 Mb (21,22). Twelve are located within introns of genes, with six of those being genes with known phenotypic associations (see Table 2). Mutations in the six genes are well documented as causative of the corresponding syndromes, but no mutations have been found to be in linkage disequilibrium with any tetra-nucleotide repeat genotypes. Of the intronic STRs, two (FGA and VWA) were within 400 bp of a splice site. All STR loci were associated (within 1 kb) with at least one phenotype according to published GWAS or quantitative trait loci (QTL) studies. TH01 was associated with the most phenotypes (18 traits) ranging from alcoholism (23) and schizophrenia (24) to autosomal recessive spinocerebellar ataxia (25), while DYS391 is believed to be associated only with hairy ears (26). Such genome wide studies often span large regions of the genome; our findings demonstrate that CODIS STR loci are located within such regions, and hence potentially linked to such traits. However, association with these traits does not imply necessarily that individual CODIS marker genotypes are

TABLE 1—*Genomic characterization of CODIS markers.*

	CODIS Marker	Cytogenetic Location	Intragenic or Distance from Nearest Gene	Included in Marshfield Human Genetic Linkage Maps	Number of (#) SNPs (dbSNP Build 132) Within 1 kb
1	D18S51	18q21.33	Intron 1	Included	13
2	FGA	4q28	Intron 3		4
3	D21S11	21q21.1	>100 kb from nearest gene	Removed	7
4	D8S1179	8q24.13	>50 kb from nearest gene	Included	14
5	VWA*	12p13.31	Intron 40		27
6	D13S317	13q31.1	>100 kb from nearest gene	Included	10
7	D16S539	16q24.1	~10 kb from nearest gene	Removed	29
8	D7S820	7q21.11	Intron 1	Included	7
9	TH01	11p15.5	Intron 1		8
10	D3S1358	3p21.31	Intron 20		11
11	D5S818	5q23.2	>100 kb from nearest gene	Removed	8
12	CSF1PO	5q33.1	Intron 6		15
13	D2S1338	2q35	~20 kb from nearest gene	Included	11
14	D19S433	19q12	Intron 1	Included	10
15	D1S1656	1q42	Intron 6	Removed	22
16	D12S391*	12p13.2	~40 kb from nearest gene	Included	18
17	D2S441	2p14	~30 kb from nearest gene	Removed	13
18	D10S1248	10q26.3	~3 kb from nearest gene	Included	16
19	Penta E	15q26.2	Within uncharacterized EST; ~50 kb from nearest gene		19
20	DYS391	Yq11.21	~5 kb from nearest gene		0
21	TPOX	2p25.3	Intron 10		33
22	D22S1045	22q12.3	Intron 4	Included	19
23	SE33	6q14	psedogene, ~30 kb from nearest gene		8
24	Penta D	21q22.3	Intron 4		6

Markers are shown in their relative rank according to Hares (1).
*VWA and D12S391 are colocated on 12p13 within 6 Mb.

TABLE 2—*Reported phenotypic relevance of genomic regions of CODIS markers.*

	CODIS Marker	Gene Name	Disorder(s) Caused by Gene Mutations	Number of (#) Phenotypes Associated Within 1 kb	Predicted DNA Elements
1	D18S51	*BCL2* (B-cell CLL/lymphoma 2)	Leukemia/lymphoma, B-cell	11	ELAV1 binding site
2	FGA	*FGA* (fibrinogen alpha chain)	Congenital afibrinogenemia; hereditary renal amyloidosis; dysfibrinogenemia (alpha type)	17	PABPC1 binding site
3	D21S11	None		1	None
4	D8S1179	None		17	None
5	VWA*	*VWF* (von Willebrand factor)	Von Willebrand disease	12	ELAV1 binding site
6	D13S317	None		5	None
7	D16S539	None		8	None
8	D7S820	*SEMA3A* (sema domain, immunoglobulin domain, short basic domain, secreted (semaphorin) 3A)		8	CELF1, ELAV1 and PABPC1 binding site
9	TH01	*TH* (tyrosine hydroxylase)	Segawa syndrome, recessive	18	ELAVL1, PABPC1 and SLBP binding site
10	D3S1358	*LARS2* (leucyl-tRNA synthetase 2, mitochondria)		15	None
11	D5S818	None		5	None
12	CSF1PO	*CSF1R* (colony stimulating factor 1 receptor)	Predisposition to myeloid malignancy	15	eGFP-GATA2 transcription factor; PABPC1 binding site
13	D2S1338	None		9	None
14	D19S433	*C19orf2* (uncharacterized gene)		7	DNase I hypersensitivity site; SLBP binding site
15	D1S1656	*CAPN9* (calpain 9)		10	PABPC1 binding site
16	D12S391*	None		6	None
17	D2S441	None		6	None
18	D10S1248	None		6	DNase I hypersensitivity site
19	Penta E	EST: BG210743 (uncharacterized EST)		8	None
20	DYS391	None		1	None
21	TPOX	*TPO* (thyroid peroxidase)	Thyroid dyshormonogenesis 2A	5	PABPC1 and SLBP binding site
22	D22S1045	*IL2RB* (interleukin 2 receptor, beta)		11	None
23	SE33	None		9	None
24	Penta D	*HSF2BP* (heat shock factor 2-binding protein)		6	PABPC1 and SLBP binding site

Markers are shown in their relative rank according to Hares (1).
*VWA and D12S391 are colocated on 12p13 within 6 Mb.

predictive or causative of any particular trait. As expected, all regions were sprinkled with documented SNPs (see Table 1), with the region of TPOX having the most (33 SNPs) and the region of FGA having the fewest (four SNPs). Four SNPs (rs3829986 and rs41338945 near CSF1PO, rs34120165 near VWA, and rs28359647 near D1S1656) were among those commonly queried for the 1000 Genome Project; none of these are annotated in SNPedia. None of the STRs overlapped predicted enhancers. Ten of the STRs (CSF1PO, FGA, TH01, TPOX, VWA, D7S820, D18S51, D19S433, D1S1656, and Penta D) lay within predicted RNA-binding protein sites. Two STRs (D19S433 and D10S1248) lay within DNase I hypersensitivity sites and one (CSF1PO) lay within a transcription factor. The role of tetra-nucleotide repeats in RNA binding and DNase I hypersensitivity is unknown, although expanded tetra-nucleotide repeats may destabilize transcription factor binding sites (27). At this time, no correlation has been made between STR repeat sizes in humans and the impact on transcription factor binding. The Marshfield human genetic linkage maps include 14 of the 24 markers, with nine still in use and five identified as "cryptic duplicate markers" and removed from subsequent panels.

The current understanding of the standard and recommended CODIS panels of STR loci summarized here highlights that these markers continue to be of limited significance for assessing phenotypes. Indeed, we found no documentation of individual genotypes for the 24 STRs to be causative of any documented phenotypes either in the literature or in the interrogated databases. Several of the STRs overlay predicted sites for genomic regulation, but there is no evidence that any particular repeat genotypes are indicative of phenotype. The utility of the CODIS profile itself, even in light of the significance of various epigenetic effects and roles of noncoding RNAs, is limited to identification purposes at this time. The existence of the predicted DNA elements suggests that some STR loci may be involved in genomic regulation. However, even for CODIS marker genotypes statistically associated with biomedically relevant phenotypes, statistical association is not synonymous with positive or negative predictive value (24). While we cannot say that the standard and recommended CODIS markers are wholly absent and forever immune from any implications for potentially sensitive or medically relevant information, we can affirm that individual genotypes are not at present revealing information beyond identification (1,2,5).

References

1. Hares DR. Expanding the CODIS core loci in the United States. Forensic Sci Int Genet 2011;6(1):e52–4.
2. Ge J, Eisenberg A, Budowle B. Developing criteria and data to determine best options for expanding the core CODIS loci. Investig Genet 2012;3:1.
3. DNA Identification Act of 1994, Pub Law 103–322, 108 Stat. 1796, 2065–71.

4. DNA Analysis Backlog Elimination Act of 2000, Pub Law 106–546, 114 Stat. 2726–37.
5. Justice for All Act of 2004, Pub Law 108–405, 118 Stat. 2260.
6. DNA Fingerprinting Act of 2005, Pub Law 109–162, 119 Stat. 2960, 3085
7. 73 Fed. Reg. at 74937.
8. Weber JL. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics 1990;7(4):524–30.
9. Ghebranious N, Vaske D, Yu A, Zhao C, Marth G, Weber JL. STRP screening sets for the human genome at 5 cm density. BMC Genomics 2003;4(6):1–10.
10. Butler JM. Genetics and genomics of core short tandem repeat loci used in human identity testing. J Forensic Sci 2006;51(2):253–65.
11. *People v. Buza*, San Francisco Co. Super. Ct. SCN 207818 (First App. Dist, Ct. of App. Cal., Aug. 4, 2011) at 5 (quoting Haskell v. Brown, 677 F.Supp.2d 1187, 1190 (N.D. Cal. 2009)).
12. Gibbs WW. The unseen genome: gems among the junk. Sci Am 2003;289(5):46–53, 48.
13. Koops B-J, Schellekens M. Forensic DNA phenotyping: regulatory issues. Columbia Sci Technol Law Rev 2008;9:159–202.
14. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res 2002;12 (4):656–64, http://genome.ucsc.edu/index.html (accessed May 11, 2012).
15. Butler JM, Reeder DJ. Overview of STR fact sheets. Short Tandem Repeat DNA Internet Database. www.cstl.nist.gov/strbase/str_fact.htm (accessed May 11, 2012).
16. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, et al. Ensembl 2011. Nucleic Acids Res 2011;39(Database issue):D800–6.
17. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007;39(10):1181–6.
18. Online Mendelian Inheritance in Man (OMIM®). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, http://omim.org/ (accessed May 11, 2012).
19. GeneTests: medical genetics information resource (database online). Copyright, University of Washington, Seattle, 1993–2011, http://www.genetests.org (accessed May 11, 2012).
20. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 1998;63:861–9, http://research.marshfieldclinic.org/genetics/GeneticResearch/compMaps.asp (accessed May 11, 2012).
21. O'Connor KL, Hill CR, Vallone PM, Butler JM. Linkage disequilibrium analysis of D12S391 and vWA in U.S. population and paternity samples. Forensic Sci Int Genet 2011;5(5):538–40.
22. Gill P, Phillips C, McGovern C, Bright JA, Buckleton J. An evaluation of potential allelic association between the STRs vWA and D12S391: implications in criminal casework and applications to short pedigrees. Forensic Sci Int Genet 2012;6(4):477–86.
23. Dahmen N, Völp M, Singer P, Hiemke C, Szegedi A. Tyrosine hydroxylase val-81-Met polymorphism associated with early-onset alcoholism. Psychiatr Genet 2005;15(1):13–6.
24. Jacewicz R, Szram S, Galecki P, Berent J. Will genetic polymorphism of tetranucleotide sequences help in the diagnostics of major psychiatric disorders? Forensic Sci Int 2006;3:24–7.
25. Breedveld GJ, van Wetten B, te Raa GD, Brusse E, van Swieten JC, Oostra BA, et al. A new locus for a childhood onset, slowly progressive autosomal recessive spinocerebellar ataxia maps to chromosome 11p15. J Med Genet 2004;41:858–66.
26. Lee AC, Kamalam A, Adams SM, Jobling MA. Molecular evidence for absence of Y-linkage of the hairy ears trait. Eur J Hum Genet 2004;12:1077–9.
27. McIvor EI, Polak U, Napierala M. New insights into repeat instability: role of RNA-DNA hybrids. RNA Biol 2010;7:551–8.

Additional information and reprint requests:
Sara H. Katsanis, M.S.
Genome Ethics, Law & Policy
Duke Institute for Human Genome Sciences & Policy
Duke University
304 Research Drive, Box 90141
Durham, NC 27708
E-mail: sara.katsanis@duke.edu