Forensic Science International: Genetics 25 (2016) 175–181

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States

Tamyra R. Moretti[a,*], Lilliana I. Moreno[a], Jill B. Smerick[a], Michelle L. Pignone[a], Rosana Hizon[a], John S. Buckleton[b], Jo-Anne Bright[b], Anthony J. Onorato[a]

[a] DNA Support Unit, Federal Bureau of Investigation Laboratory, 2501 Investigation Parkway, Quantico, VA 22135, USA
[b] Institute of Environmental Science and Research, Private Bag 92021, Auckland 1025, New Zealand

ARTICLE INFO

Article history:
Received 13 October 2015
Received in revised form 28 July 2016
Accepted 30 July 2016
Available online 3 August 2016

Keywords:
STR
CODIS core loci
AmpFlSTR GlobalFiler
PowerPlex Fusion
Population genetics

ABSTRACT

Allele distributions for twenty-three autosomal short tandem repeat (STR) loci – D1S1656, D2S441, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, CSF1PO, FGA, Penta D, Penta E, SE33, TH01, TPOX and vWA – were determined in Caucasians, Southwestern Hispanics, Southeastern Hispanics, African Americans, Bahamians, Jamaicans, Trinidadians, Chamorros, Filipinos, Apaches, and Navajos. The data are included in the FBI PopStats software for calculating statistical estimates of DNA typing results and cover the expanded CODIS Core STR Loci required of U.S. laboratories that participate in the National DNA Index System (NDIS).

1. Introduction

The FBI Laboratory recently announced an expansion of the original thirteen short tandem repeat (STR) loci that have been the core of the National DNA Index System (NDIS) since 1997. The FBI requires CODIS laboratories to implement seven additional STR loci selected by the CODIS Core Loci Working Group by January 1, 2017 [1]. Collectively, these loci provide greater discrimination potential for human identification applications and enhance kinship analyses typically used in missing person inquiries. Since many of these loci are included in databases globally, the expanded STR locus set facilitates international law enforcement and counter-terrorism endeavors. Furthermore, the enhanced information content provided by the additional loci improves searching DNA profiles within large databases such as NDIS, which as of January 2016 contains over 15 million DNA profiles.

The twenty STR loci (the original set: D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, CSF1PO, FGA, TH01, TPOX and vWA; and the additional set: D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433 and D22S1045) can be simultaneously genotyped with either the AmpFlSTR® GlobalFiler® (GlobalFiler, Life Technologies, Inc., Carlsbad, CA) or PowerPlex® Fusion™ (Fusion, Promega Corporation, Madison, WI) multiplex amplification systems. These kits also enable the genotyping of SE33 and a Y indel locus (GlobalFiler), Penta D and Penta E (Fusion), and DYS391 (GlobalFiler and Fusion), as well as Amelogenin for sex determination.

The majority of forensic DNA testing laboratories in the U.S. use the FBI's PopStats software in CODIS for estimating the statistical weight of evidentiary DNA profiles. This report expands the FBI population data [2–4] in PopStats, presenting the allele frequencies for twenty-three autosomal STR loci, as determined with both the GlobalFiler and Fusion kits in African Americans, Caucasians, Southeast Hispanics, Southwestern Hispanics, Bahamians, Jamaicans, Trinidadians, Apaches, Navajos, Chamorros and Filipinos. Concordance studies demonstrate genotyping accuracy and identify instances of non-concordance due to rare kit-specific primer binding site variants that preclude allele detection. Results of population genetic analyses also presented in this report support the usage of these loci and the associated allele frequencies for estimating match statistics in human identity testing.

* Corresponding author.
E-mail address: Tamyra.Moretti@ic.fbi.gov (T.R. Moretti).

http://dx.doi.org/10.1016/j.fsigen.2016.07.022

2. Materials and methods

2.1. Sample preparation and genotyping

DNA samples that were previously typed at the original core CODIS loci and the Amelogenin locus were used in the present study. Some samples were also previously typed at the AmpliType® Polymarker, DQA1 and D1S80 loci. The source and preparation of the samples are previously described [2–8]. All procedures were conducted according to manufacturers' recommendations, except as noted. DNA samples requiring re-extraction were generated from liquid blood dried onto FTA paper and extracted either using the AutoMate Express™ DNA Extraction System with the PrepFiler Express™ DNA Extraction Kit (Life Technologies, Inc.) or the EZ1® Advanced XL with the EZ1 DNA Investigator Kit (Qiagen Sciences, Inc., Gaithersburg, MD). DNA quantities were estimated using the Quantifiler® Duo DNA Quantification Kit (Life Technologies, Inc.), and generally 0.5 or 1 ng DNA was amplified with the GlobalFiler and Fusion kits in a GeneAmp PCR System 9700 (Life Technologies, Inc.). Amplified samples were subjected to capillary electrophoresis in POP 4™ Performance Optimized Polymer (Life Technologies, Inc.) using an Applied Biosystems® 3130xl Genetic Analyzer (Life Technologies, Inc.). Genotyping was performed using GeneMapper® ID-X software version 1.4 (Life Technologies, Inc.). An off-ladder allele was sequenced using the ForenSeq™ DNA Signature Prep Kit with detection on the MiSeq FGx™ Forensic Genomics System (Illumina, San Diego, CA) and analyzed using the MFold utility within OligoAnalyzer 3.1 (Integrated DNA Technologies). Using a software tool developed in-house in Microsoft Excel, genotypes of samples that were generated using GlobalFiler and Fusion were verified against each other and published data for the same samples as generated using the AmpFlSTR Profiler Plus®, AmpFlSTR COfiler® (Life Technologies, Inc.) and/or PowerPlex 1.1 (Promega Corporation) kits [7,8].

2.2. Statistical analysis

Microsoft Excel was used to calculate allele frequencies. Arlequin version 3.5 [9] was used to perform exact test of population differentiation using 100,000 steps and to calculate genetic distance (Fst). Fisher's exact tests [10] for allelic association was undertaken with 10,000 shuffles using the software Genetic Data Analysis available from Lewis and Zaykin [11]. The truncated

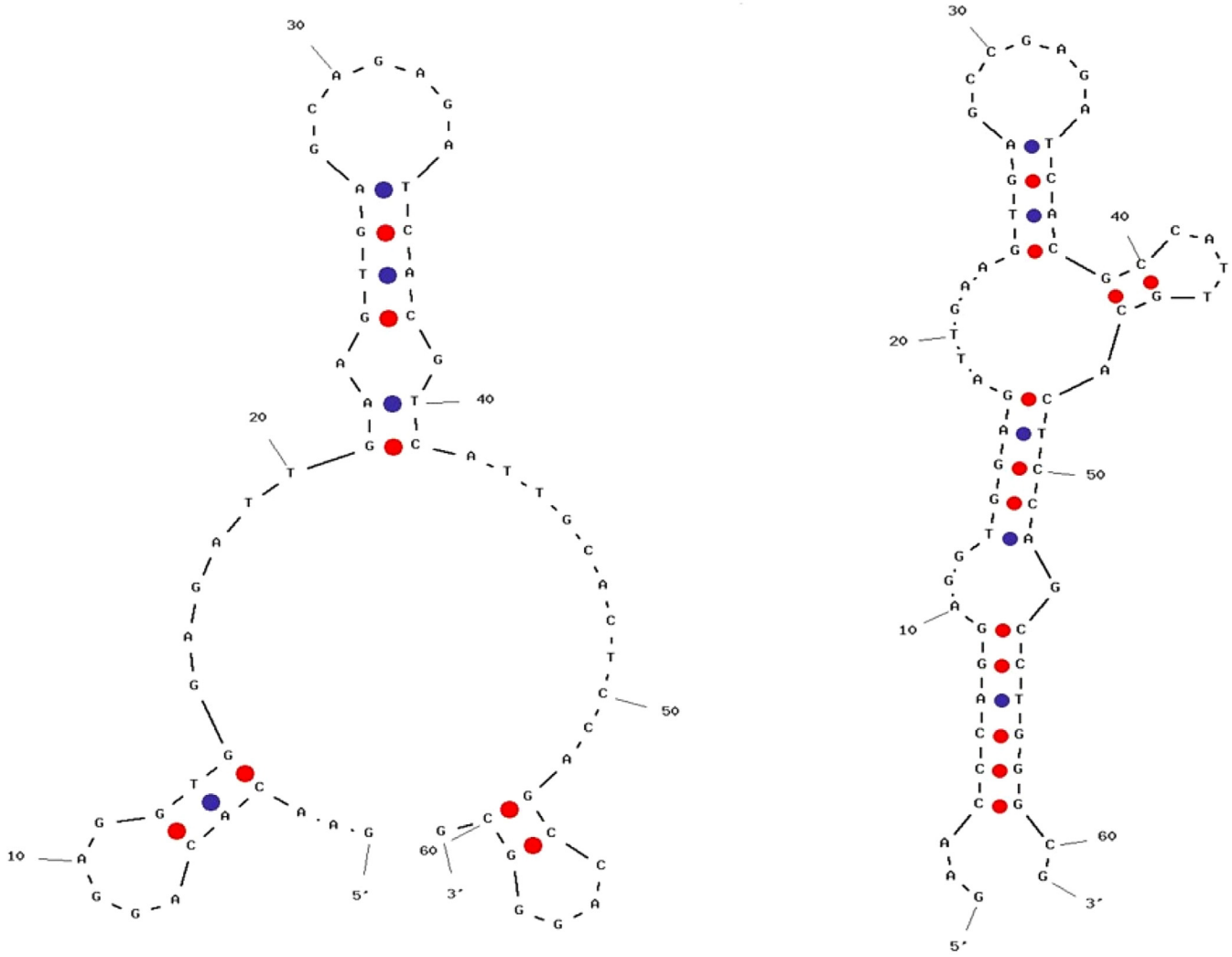


Fig. 1. Theoretical secondary structure of allele 11 at the Penta E locus. Left: allele that corresponds relative to size with the allelic ladder. Right: allele with modified mobility relative to the allelic ladder. The SNPs A/C_{24}, A/C_{49}, T/C_{59} and A/T_{75} are indicated as positions 5, 30, 40 and 56, respectively.

Table 1
Genotype non-concordances due to amplification failure of the specified alleles.

Kit	Locus	Undetected allele	Population	Kit(s) showing recovered allele
GlobalFiler	D12S391	21	Southeast Hispanic	Fusion
GlobalFiler	D12S391	23	Southwest Hispanic	Fusion
Fusion	D13S317	8	Southeast Hispanic	GlobalFiler
GlobalFiler	D13S317	13	Filipino	Fusion
Fusion	D16S539	9	Southeast Hispanic	GlobalFiler
PowerPlex 1.1	D16S539	10	Jamaican	GlobalFiler, Fusion
Fusion	D16S539	11	Filipino	GlobalFiler
PowerPlex 1.1	D16S539	12	Jamaican	GlobalFiler, Fusion
GlobalFiler	D1S1656	15	African American	Fusion
Fusion	D22S1045	14	Southwest Hispanic	GlobalFiler
Profiler Plus	FGA	22	African American	GlobalFiler, Fusion
Identifiler Plus	vWA	18	Southeast Hispanic	GlobalFiler, Fusion

product method of Fisher was used to test the hypotheses that none of the loci departed from Hardy-Weinberg and linkage equilibrium ($p > 0.05$). In addition, the data was explored graphically by producing p-p plots (observed $\sim$ expected).

3. Results

Allele frequencies and the Fisher's exact test results are available as Supplementary material and/or at http://www.fbi.gov/about-us/lab.

Seven samples (one African American, two Bahamian and four Jamaican) were identified as having an off-ladder allele at Penta E. All occurrences of this allele sized approximately 0.5 bp smaller than allele 11. Sequencing results confirm that for six of the seven samples, the allele contained 11 complete pentanucleotide repeats, and thus the correct allele call is 11. All six variant alleles each exhibited four SNPs: A/C_{24}, A/C_{49}, T/C_{59} and A/T_{75}. With the exception of A/C_{49} which resides within a loop of a theoretical stem-loop structure, these variants are presumed to disrupt the predicted base-pairing within the DNA fragment, thereby altering the secondary structure of the allele (Fig. 1) and its electrophoretic mobility [12,13]. Insufficient DNA was available for the sequencing of one such allele which, since unconfirmed, was excluded from subsequent statistical analyses.

Twelve non-concordances in which a known allele was not detected with a given amplification kit were identified during this research. Information for each non-concordance is provided in Table 1.

There are a number of loci on the same chromosome within the GlobalFiler and Fusion multiplexes. A summary of the p-value for the linked loci are provided in Table 2.

4. Discussion

Sequence variants at STR loci can have different effects on allele detection and migration, depending on the type (e.g., SNP, insertion/deletion) and position of the variant within the amplified DNA fragment [12,13]. Primer binding site variants may attenuate or impede hybridization of the amplification primer to a target sequence, resulting in a reduction in peak height or an 'apparent homozygous' typing result (i.e., a null allele at a heterozygous locus), respectively [14,15]. Some primer sequences vary among different amplification kits and, in rare instances, yield discordant typing results for the same individual's DNA typed with different kits. Consistent with kit configurations and primer differences, discordant genotypes were found in the present study when comparing Globalfiler and Fusion typing results at D1S1656, D12S391, D13S317, D16S539 and D22S1045 and when comparing PowerPlex 1.1 and Fusion results at D16S539 [16,17]. The recovery with Globalfiler of an Identifiler Plus null allele at vWA and a Profiler Plus null allele at FGA may be attributed to the incorporation of SNP-specific primers in Globalfiler [18].

The tests for departures from independence are for population level applications. There is no evidence for an effect of linkage creating disequilibrium at the population level in the physically

Table 2
p-values for linked loci for different subpopulations. Nominally significant p-values <0.05 are highlighted.

Population	vWA/D12S391	D5S818/CSF1PO	D21S11/Penta D	TPOX/D2S441
Recombination fractions	0.117	0.252	0.357	0.472
Caucasian	0.804	0.267	0.080	0.229
Southwest Hispanic	0.242	0.383	0.234	0.758
Southeast Hispanic	0.442	0.583	0.308	0.418
African American	0.425	0.135	0.703	0.707
Bahamian	0.379	0.413	0.483	0.321
Jamaican	0.163	0.041	0.564	0.928
African American/Bahamian/Jamaican	0.330	0.025	0.632	0.492
Trinidadian	0.598	0.478	0.898	0.663
Chamorro	0.243	0.927	0.442	0.542
Filipino	0.807	0.891	0.447	0.086
Apache	0.190	0.636	0.102	0.878
Navajo	0.308	0.792	0.406	0.143

The p-p plots and numbers of samples for each population group are presented in Table 3. If the hypotheses of Hardy-Weinberg and linkage equilibrium were true, then the p-values should be distributed uniformly between 0 and 1; $p \sim U[0, 1]$. As an example, the $x = y$ line in the p-p plots represents equilibrium and deviations from that line can be seen as departures from equilibrium. The 95% confidence limit is also displayed on the p-p plots as the region within the two curved lines. Evidence of departure from linkage equilibrium was observed in the Apache population. There is no evidence from departure from HW or linkage equilibrium in any of the other populations groups.

linked pairs of loci (see Table 2). There are two *p*-values below 5% out of 48 comparisons.

This is at expectation for equilibrium and consistent with the predictions of Budowle et al. [19]. We therefore recommend the multiplication of these loci for the evaluation of match probabilities for unrelated individuals and some classes of relationship. However there are some relationships where linkage is expected to have an effect [20–22].

These datasets are suitable in terms of both size and quality for the purposes of estimating DNA profile probabilities.

Table 3
p-p plot for the HWE and LE tests for each 23-locus dataset (combined GlobalFiler and Fusion). N represents the number of samples for each population group. The 95% confidence intervals generated by simulation from U[0,1] are also shown.(For interpretation of the references to colour in this table, the reader is referred to the web version of this article.)

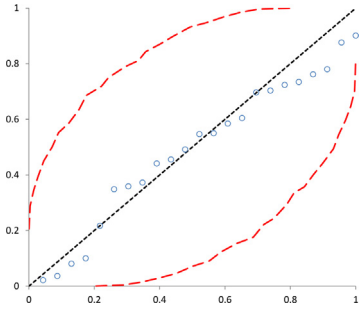
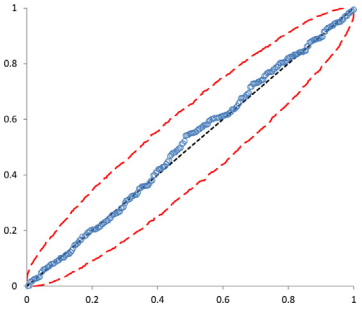
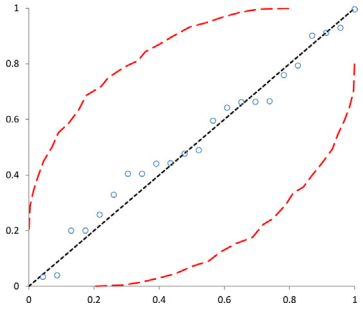
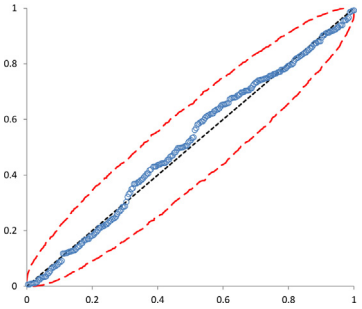
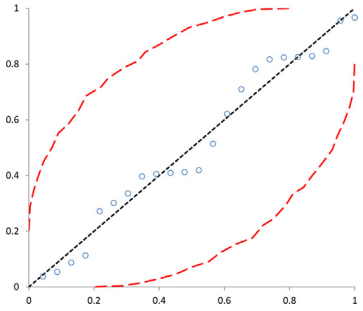
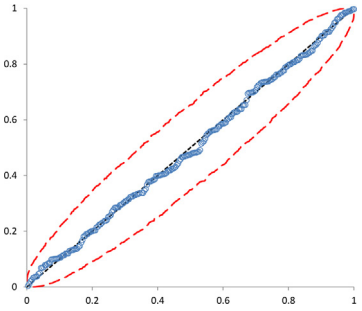
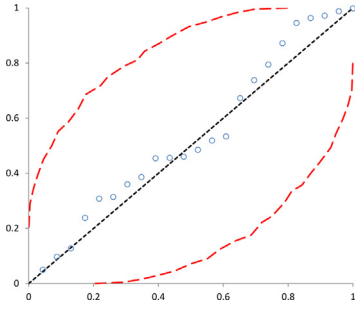
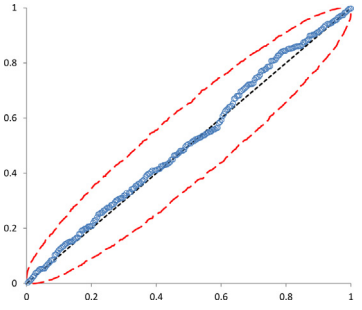
Pop	*N*	HWE	LE
Caucasian	202		
Southwest Hispanic	209		
Southeast Hispanic	263		
African American	209		

Table 3 (*Continued*)

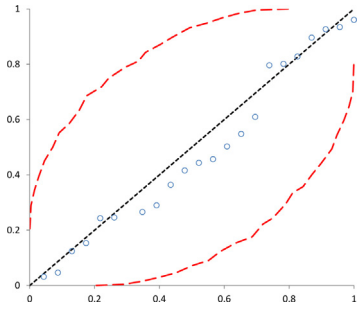
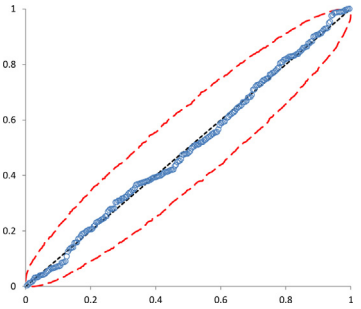
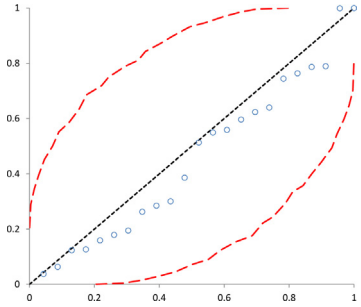
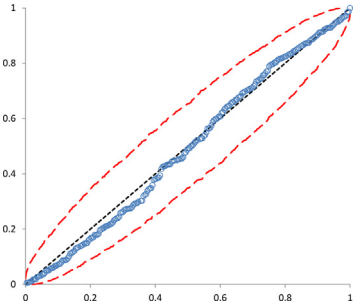
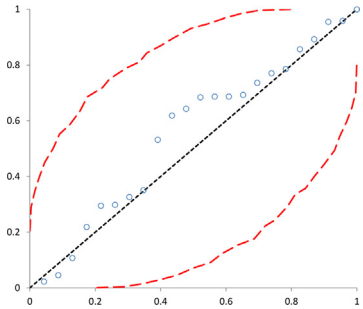
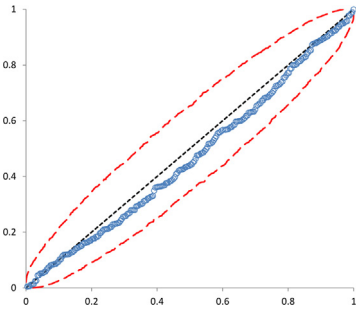
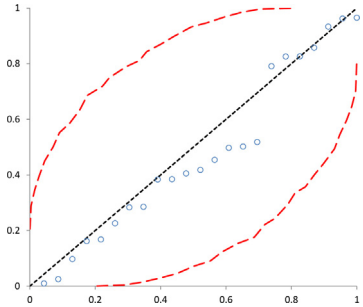
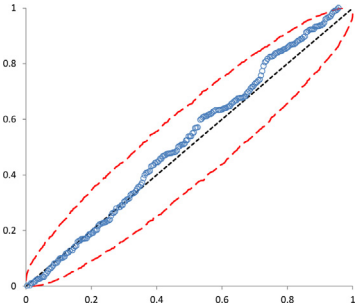
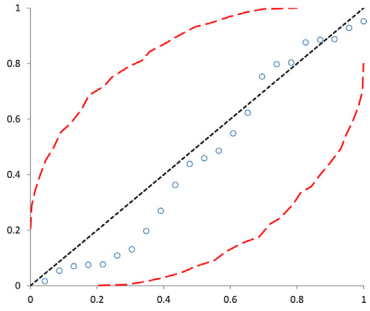
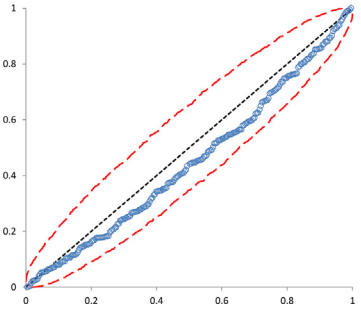
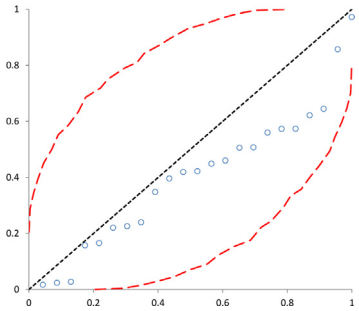
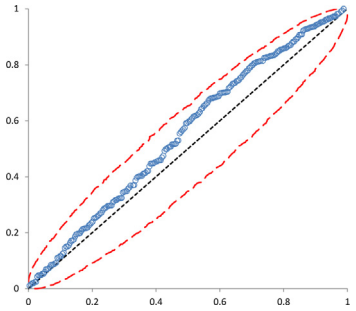
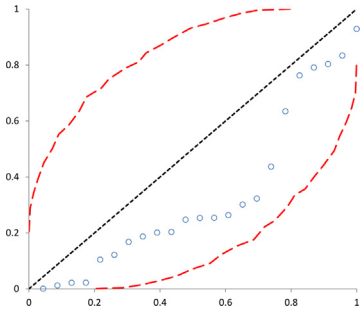
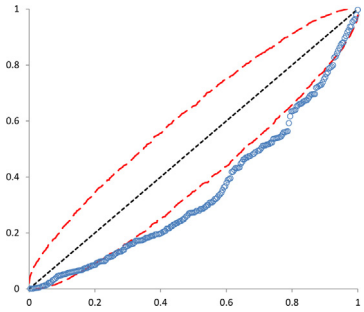
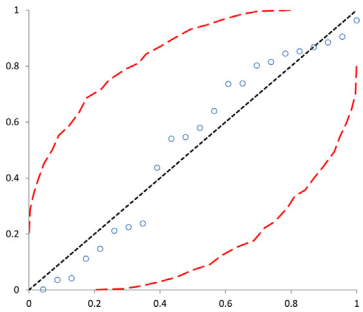
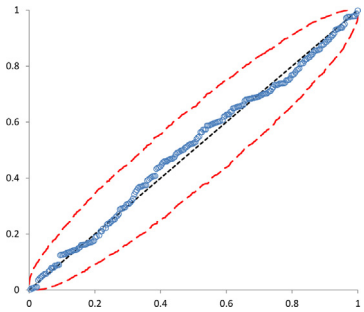
Pop	*N*	HWE	LE
Bahamian	157		
Jamaican	177		
African American, Bahamian, Jamaican	543		
Trinidadian	79		
Chamarro	95		

Table 3 (*Continued*)

Pop	*N*	HWE	LE
Filipino	91		
Apache	192		
Navajo	143		

The genetic distances between the African American, Bahamian, and Jamaican are small. The *p*-values for the hypothesis Fst = 0 are 0.5586 (African American/Jamaican), 0.1441 (African American/Bahamian) and 0.6216 (Jamaican/Bahamian), indicating that these three populations may be grouped if desired. Note that the data do not support combining other populations, including Southeast and Southwest Hispanics.

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgements

The authors gratefully acknowledge Jeremy Fletcher for the development of Excel macros and Amber Carr, Jocelyn Carlson, Jerrilyn Conway, Jade Gray, Jodi Irwin, Cayman Taylor and Leah Willis for technical review of genotyping data. We also wish to thank Thomas Callaghan, Susannah Kehl, Jodi Irwin and two anonymous reviewers for recommendations which have improved this paper.

Usage of the samples in this study was approved by the Institutional Review Board of the Federal Bureau of Investigation. In accordance with FBI IRB Docket 318-15, publication of information relative to any individual human subject is limited to aggregated allele frequencies. The authors are not authorized to publish genotype and sequence data. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer, or its products or services, by the FBI. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. government.

This work was supported in part by grant 2014-DN-BX-K028 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice or Department of Commerce. Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the

materials, instruments or equipment identified are necessarily the best available for the purpose.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2016.07.022.

References

[1] D.R. Hares, Selection and implementation of expanded CODIS core loci in the United States, Forensic Sci. Int. Genet. 17 (2015) 33–34.

[2] B. Budowle, B. Shea, S. Niezgoda, R. Chakraborty, CODIS STR loci data from 41 sample populations, J. Forensic Sci. 46 (2001) 453–489.

[3] B. Budowle, D.A. Defenbaugh, K.M. Keys, Genetic variation at nine short tandem repeat loci in Chamorros and Filipinos from Guam, Leg. Med. 2 (2000) 26–30.

[4] B. Budowle, T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh, K.M. Keys, Population data on the thirteen CODIS core short tandem repeat loci in African Americans, US, Caucasians, Hispanics, Bahamanians, Jamaicans and Trinidadians, J. Forensic Sci. 44 (1999) 1277–1286.

[5] B. Budowle, F.S. Baechtel, C.T. Comey, A.M. Giusti, L. Klevan, Simple protocols for typing forensic biological evidence: chemiluminescent detection and restriction fragment length polymorphism (RFLP) analyses and manual typing of polymerase chain reaction (PCR) amplified polymorphisms, Electrophoresis 16 (1995) 1559–1567.

[6] B. Budowle, F.S. Baechtel, Fejeran R. Polymarker, HLA-DQA1, and D1S80 allele frequency data in Chamorro and Filipino populations from Guam, J. Forensic Sci. 43 (1998) 1195–1198.

[7] B. Budowle, Genotype profiles for five population groups at the short tandem repeat loci D2S1338 and D19S433, Forensic Sci. Commun. (2001). (accessed 14.03.16) https://www2.fbi.gov/hq/lab/fsc/backissu/july2001/budowle1.htm.

[8] B. Budowle, T.R. Moretti, Genotype profiles for six population groups at the 13 CODIS short tandem repeat core loci and other PCR-based loci, Forensic Sci. Commun. (1999). Volume 1 - Number 2 (accessed 14.03.16) https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/july1999/budowle.htm.

[9] L. Excoffier, H.E.L. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (2010) 564–567.

[10] S.W. Guo, E.A. Thompson, Performing the exact test of Hardy-Weinberg proportion for multiple alleles, Biometrics 48 (1992) 361–372.

[11] P.O. Lewis, D. Zaykin, Computer program for the analysis of allelic data, Genet. Data Anal. d16c (2016) ed2001.

[12] L. Gusmão, A. Amorim, M.J. Prata, L. Pereira, M.V. Lareu, A. Carracedo, Failed PCR amplifications of MBP-STR alleles due to polymorphism in the primer annealing region, Int. J. Legal Med. 108 (1996) 313–315.

[13] C. Davis, J. Ge, J. King, N. Malik, V. Weirich, A.J. Eisenberg, B. Budowle, Variants observed for STR locus SE33: a concordance study, Forensic Sci. Int. Genet. 6 (2012) 494–497.

[14] B. Budowle, A. Masibay, S.J. Anderson, C. Barna, L. Biega, S. Brenneke, B. Brown, J. Cramer, G.A. Degroot, D. Douglas, B. Duceman, A. Eastman, R. Giles, J. Hamill, D.W. Janssen, T.D. Kupferschmid, T. Lawton, C. Lemire, B. Llewellyn, T. Morretti, J. Neves, C. Palaski, S. Schueler, J. Sgueglia, C. Sprecher, C. Tomsey, D. Yet, STR primer concordance study, Forensic Sci. Int. 124 (2001) 47–54.

[15] A.A. Westen, T. Kraaijenbrink, E.A. Robles de Medina, J. Harteveld, P. Willemse, S.B. Zuniga, K.J. van der Gaag, N.E.C. Weiler, J. Warnaar, M. Kayser, T. Sijen, P. de Knijff, Comparing six commercial autosomal STR kits in a large Dutch population sample, Forensic Sci. Int. Genet. 10 (2014) 55–63.

[16] GenePrint ® Fluorescent STR Systems Technical Manual, Revision 8/12, Promega Corporation, (2016) . (accessed 14 .03.16) https://www.promega.com/~/media/files/resources/protocols/technical%20manuals/101/geneprint%20fluorescent%20str%20systems%20protocol.pdf.

[17] PowerPlex® Fusion System Technical Manual, Revision 3/15, Promega Corporation, (2016) . (accessed 14.03.16) https://www.promega.com/~/media/files/resources/protocols/technical%20manuals/101/powerplex%20fusion%20system%20protocol.pdf.

[18] GlobalFiler™ PCR Amplification Kit User Guide, Revision D, Applied Biosystems, (2016) . (accessed 14.03.16) https://tools.thermofisher.com/content/sfs/manuals/4477604.pdf.

[19] B. Budowle, J. Ge, R. Chakraborty, A. Eisenberg, R. Green, J. Mulero, et al., Population genetic analyses of the NGM STR loci, Int. J. Legal Med. 125 (2010) 101–109.

[20] J.S. Buckleton, C.M. Triggs, The effective of linkage on the calculation of DNA match probabilities for siblings and half siblings, Forensic Sci. Int. 160 (2006) 193–199.

[21] T. Egeland, N. Sheehan, On identification problems requiring linked autosomal markers, Forensic Sci. Int. Genet. 2 (2008) 219–225.

[22] J.-A. Bright, J.M. Curran, J.S. Buckleton, Relatedness calculations for linked loci incorporating subpopulation effects, Forensic Sci. Int. Genet. 7 (2013) 380–383.