Forensic Science International: Genetics 7 (2013) 63–74

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

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

Revision of the SNP*for*ID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies

M. Fondevila [a], C. Phillips [a,*], C. Santos [a], A. Freire Aradas [a], P.M. Vallone [b], J.M. Butler [b], M.V. Lareu [a], Á. Carracedo [a]

[a] Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Spain
[b] U.S. National Institute of Standards and Technology, Biochemical Science Division, Gaithersburg, MD, USA

ARTICLE INFO

Article history:
Received 2 March 2012
Received in revised form 28 May 2012
Accepted 7 June 2012

Keywords:
SNPs
Genetic ancestry
Ancestry inference
Population genetics
Population admixture

ABSTRACT

A revision of an established 34 SNP forensic ancestry test has been made by swapping the underperforming rs727811 component SNP with the highly informative rs3827760 that shows a near-fixed East Asian specific allele. We collated SNP variability data for the revised SNP set in 66 reference populations from 1000 Genomes and HGDP-CEPH panels and used this as reference data to analyse four U.S. populations showing a range of admixture patterns. The U.S. Hispanics sample in particular displayed heterogeneous values of co-ancestry between European, Native American and African contributors, likely to reflect in part, the way this disparate group is defined using cultural as well as population genetic parameters. The genotyping of over 700 U.S. population samples also provided the opportunity to thoroughly gauge peak mobility variation and peak height ratios observed from routine use of the single base extension chemistry of the 34-plex test. Finally, the genotyping of the widely used DNA profiling Standard Reference Material samples plus other control DNAs completes the audit of the 34-plex assay to allow forensic practitioners to apply this test more readily in their own laboratories.

1. Introduction

Single nucleotide polymorphism (SNP) typing is still relatively new to the field of DNA profiling. While short amplicon approaches based around SNP typing have proved their worth in cases of severely degraded DNA analysis [1–3], the inability of SNPs to provide informative links to STR databases has hindered widespread use beyond simple identification of missing persons with reference to their surviving relatives [4]. At the current time the future of forensic SNP analysis appears to centre on two approaches where SNPs can provide important supplementary information: (a) in complex relationship tests involving deficient pedigrees, very distant relationships [5] or a single second order STR exclusion that creates ambiguous likelihood ratios [6]; and (b) helping to build an inference of the likely physical appearance of DNA donors when other information such as reliable eye witness or DNA database entries are unavailable to investigators [7,8]. In the latter category there is growing interest in establishing SNP tests able to predict hair and eye colour variation in European subjects [9–11], along with ancestry informative marker (AIM) SNP tests that equate the genotypes detected in an individual to their genetic ancestry, where in this case, genetic ancestry is a characteristic defined by broadly based continental population group SNP variability [12–14].

Five years ago we developed a 34 SNP forensic ancestry test [12] and this paper outlines an audit of the underlying multiplex assay bringing enhancements to the chemistry, swapping a single underperforming SNP rs727811 with a new, more informative replacement SNP rs3827760 and recording the genotypes of standard reference samples used as positive controls in forensic DNA laboratories [15,16]. We also outline studies of Central and South American populations where levels of admixture are higher than other parts of the world and we collate recently released genotype data from 1000 Genomes for the 34 component SNPs plus rs3827760. Finally, it is worth emphasising that in the five years since our 34-plex test was developed using Applied Biosystems (AB) SNaPshot primer extension chemistry, no other forensic SNP typing system has emerged as a viable alternative. In fact during this period two other SNP typing chemistries have been discontinued: a chip-based system used in a multiple-reaction ancestry test that typed 176 SNPs [14] (Beckman-Coulter *GenomeLab*™ *SNPstream*®) and an oligo-ligation system: AB *Genplex*®, that held much promise for a range of forensic SNP typing applications [17,18]. Since all current forensic physical characteristic and ancestry predictive tests use AB SNaPshot, more emphasis than ever should be placed on further optimisation of primer extension assay conditions in order to continue the successful application of SNPs to forensic analyses.

* Corresponding author.
E-mail address: c.phillips@mac.com (C. Phillips).

1872-4973/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2012.06.007

2. Materials and methods

2.1. Population samples

A total of 709 unrelated male samples of self-declared ancestry from the population reference collection of NIST (National Institute of Standards and Technology, Gaithersburg, MD, USA) were genotyped with the enhanced 34-plex assay reaction conditions described below. This panel is made up of representative samples from the four main U.S. population groups, comprising: 261 Caucasians, 258 African Americans, 140 Hispanics and 50 East Asians. Additionally, the six components of the recently revised NIST PCR-based DNA profiling Standard Reference Material® (SRM) 2391c [15] were typed along with the current standard forensic positive control DNAs: AB/Promega 9947a; Qiagen XY5, and; Promega 2800M, thereby complimenting the reference genotypes for the SNP*for*ID 52-plex ID-SNP assay published in 2010 [15]. NIST SRM 2391c components A, B and C are described as Caucasian, Mexican Hispanic and Melanesian in origin respectively (component D is a 3:1 mixture of A and C), while E and F do not have ancestry descriptions.

Previously generated SNP genotypes from the HGDP-CEPH Human Genome Diversity Panel (HGDP-CEPH) global diversity panel [12] were used as reference data for analyses of the NIST populations (CEPH genotype data available from the SPSmart SNP*for*ID browser at: http://spsmart.cesga.es/snpforid.php), while 1000 Genomes data was collated directly from the *ENGINES* whole genome SNP browser [19] available from: http://spsmart.cesga.es/engines.php?dataSet=engines. The HGDP-CEPH panel comprises: 105 Africans, 158 Europeans, 232 East Asians, 64 Native Americans and 28 Oceanians (South Asian and Middle East populations were excluded). The 1000 Genomes samples comprise: 246 Africans (including 61 African Americans with mixed ancestry), 380 Europeans, 286 East Asians and 181 Americans with mixed ancestry (Puerto Rican and Mexicans from Central America plus Colombians from South America). Using both panels enabled the most comprehensive geographic survey of SNP variability and *ENGINES* allows data to be collected for any SNP locus with minor allele frequencies above ~1%.

2.2. 34-plex component SNP reconfiguration

A single marker substitution was made to replace the consistently under-performing SNP rs727811 (internal code A11) with an East Asian-informative SNP rs3827760 (assigned internal code P28). Representative allele frequency distributions from 1000 Genomes and HapMap (GIH) samples comparing both SNPs are summarised in Fig. 1. Note that the SNaPshot extension primer of rs727811 interrogated the AC strand but HapMap and 1000 Genomes databases report the GT strand, while for rs3827760 SNaPshot interrogates the AG strand, as listed in the above SNP databases. The revised marker details for the complete assay are listed in Table 1. Components listed with internal codes prefixed with an 'A' are markers overlapping with the SNP*for*ID 52-plex ID-SNP multiplex [20], since they show informative allele frequency differences between populations but also serve as housekeeping markers when using both multiplexes together in an analysis. Lastly, it was necessary to design a redundancy in the extension primer of rs3827760, comprising C and T alternate bases corresponding to neighbour SNPs rs144939741 and rs121908454, 10 bp and 15 bp upstream of the target site.

2.3. Amplification and single base extension primer modifications

The PCR and single base extension (SBE) primers used in the modified assay are listed in Table 1. As well as novel primers for rs3827760, the SBE primers for component SNPs: rs2304925; rs5997008; rs2814778; rs239031; rs16891982 (internal codes P01, P02, P04, P07 and P25a respectively) have been modified to adjust peak positions for electrophoresis with AB POP-4™ capillary electrophoresis (CE) polymer. Some mobility modifying poly-CT tails were also changed to non-homologous sequences comprising all bases. POP-4™ is now in increasing use and these SBE primer rearrangements anticipate the discontinuation of AB POP-6™ CE polymer used in the original 34-plex assay development. In brief, SNPs rs2304925/rs5997008 were moved away from the fastest mobility size range where dye artefacts sometimes interfered with their peak recognition and SNPs rs2814778/rs239031/rs16891982 were size-adjusted to separate them from the co-incidental mobilities of neighbouring SNPs or closely sited non-specific peaks when using POP-4™. The rearranged SNP positions are listed in Table 2 based on average size estimates of ~750 samples with multiple runs.

2.4. Modified PCR and SBE reaction protocols

Both amplification and extension reactions were re-optimised and represent significant modifications to the originally described 34-plex assay chemistry [12]. We reduced the PCR cycling from the original 35 to 30–32 amplification cycles and from 30 to 28–30 extension cycles – each of the new cycle number ranges was modified according to the peak heights observed on identical 3130xl CE detectors in the two study laboratories, so ± two cycles adjusts for instrument sensitivity. Furthermore, cycling conditions

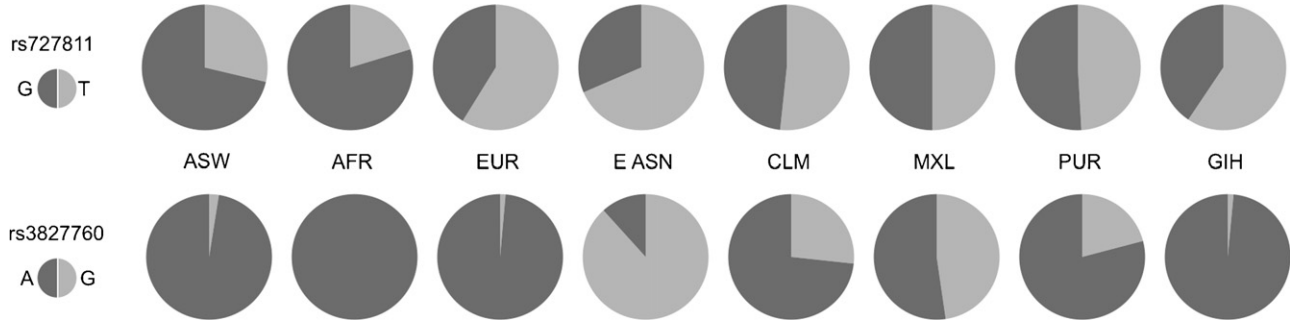

Fig. 1. Allele frequency distributions in Africans, Europeans and East Asians from 1000 Genomes and HapMap populations for the replaced rs727811 SNP and new rs3827760. Africans (AFR) comprised YRI: Yoruba in Ibadan, Nigeria and LWK: Luhya in Webuye, Kenya; Europeans (EUR) are CEU: Utah residents with N & W European ancestry from the HGDP-CEPH collection, FIN: Finnish in Finland, GBR: British in England and Scotland, IBS: Iberian populations in Spain and Toscans in Italy; East Asians (E ASN) are CHB: Han Chinese in Beijing, CHS: China, Han Chinese South and JPT: Japanese in Tokyo, Japan. Certain populations are shown separately due to levels of admixture: ASW, African ancestry in Southwest USA; CLM, Colombian in Medellín, Colombia; MXL, Mexican ancestry in Los Angeles; PUR, Puerto Rican in Puerto Rico; GIH, Gujarati Indians from Houston.

Table 1
34-plex composite SNP data.

Markers		PCR size	Strand information			Primers			Primer mix ratios	
Internal code	dbSNP rs-number		1000 G bases	Assay bases	SNaPshot direction	PCR forward primer	PCR reverse primer	SNaPshot single base extension (SBE) primer	PCR (F+R primer stock at 25 μM)	SNaPshot SBE (primer stock at 50 μM)
P02	rs5997008	90	AC	AC	F	GTCAACACTAGAGTATTTGCCCATC	ACAAACCCAAAGACTGTTCTGC	gac[aactaggtgccacgtcgtgaaagtctgac]$_2$ aactctcaCAGGATCGATTGGTTCC	1.9	1.3
P01	rs2304925	98	AC	GT	R	CCCATTAACTCATCAAAGTGGTGAT	CCCCACTCCACCGCTAAT	[aaagtctgacaactaggtgccacgtcgtg]$_2$ aaagtctgacaaCCACTCCACCGCTAAT	2.5	2.5
A07	rs917118	87	CT	AG	R	GCCCTTTAGGGTCGGTTC	GTAAGAGATGACTGAGGTCAACGAG	t[ct]$_2$TGACTGAGGTCAACGAGC	3	2.1
P03	rs1321333	80	AG	CT	R	GTCAGTAAGACGGTAACTCC	CTAACACAAGCCTAAATCCAG	AAGACGGTAACTCCATGGCTG	2.75	1.5
P04	rs2814778	102	CT	CT	F	AACCTGATGGCCCTCATTAGT	ATGGCACCGTTTGGTTCAG	agtctgacaactaggtgccacgtcgtgaaagtc tgacaactaggtgccacgtcgtgaaagtctgacat CTCATTAGTCCTTGGCTCTTA	1.5	2
A29	rs1024116	76	CT	AG	R	CCATGTGTTCTAATAAAAAGGATTGC	TGGGAAGTGAGCAAAAGTAAATACA	cttGTTCTAATAAAAAGGATTGCTCAT	1	2
P05	rs7897550	100	AG	CT	R	CGATGTGTCTTACGGAATACTAGGT	AGAGCTGACAGGCAAAAATGCTAT	t[ct]$_2$TGTGCAGGATTGAAATATAATT	2	1
A21	rs722098	80	AG	AG	F	GGAAGTACACATCTGTTGACAGTAATGA	GGGTAAAGAAATATTCAGCACATCC	agtctgacaaTGACAGTAATGAAATATCCTTG	2.3	4
P06a	rs10843344	87	CT	CT	F	TGTACAATGGTAGATGTGTGCTCAG	GATAGCTCTGGTGTTGCATTATTGT	t[ct]$_5$AGTACTTTGCCAAAGAAACTAAA	4	1.3
P08	rs12913832	99	AG	AG	F	ACGTTGGATGCGAGGCCAGTTTCATTTGAG	ACGTTGGATGAAAACAAAGAGA AGCCTCGG	[ct]$_8$CCAGTTTCATTTGAGCATTAA	3	3
P07	rs239031	70	AG	CT	R	TAGCTGTGAGATAGAAATCCTGGAC	ACTACCCTAATCTCAGCTTCCACTC	t[ct]$_8$cAATCTCAGCTTCCACTC	0.5	1
P09a	rs1978806	94	AG	CT	R	AGAGTTTGACATGATGGTGCTCTA	TCTTGTTTCTAAGCAGGAAAGTTG	t[ct]$_8$cGCAGGAAAGTTGTATTCTGATA	1.8	1
A40	rs2040411	61	AG	AG	F	TCTGGAATGCCAGTTCTTTTGT	CAGAACGCCTATGAAAACCAGT	[ct]$_7$cCTCTGTATTTTCTTACTCTAAGTGC	1	2
P10	rs773658	95	GC	CG	R	ACAAACGGAAAGTAGTATTGGACTG	AGAAGGGGCACAGCAATTTAGTA	[ct]$_{11}$cGGGAAGAATAGAGTCAATCAA	3.2	3
P11	rs10141763	82	TA	AT	R	ACAGACTTGGTTCCCTGAAGTCTA	GTAGATTGTAGGCAAGTCGTAAAGG	t[ct]$_{10}$cGTGTGAGTTGTGTGATAATCTA	3	1.1
P12	rs182549	117	CT	CT	F	AAGTACTGGGACAAAGGTGTGAG	AGAAGTCAGAATACCCCTACCCTAT	[ct]$_{16}$cAGGTGTGAGCCACCG	6.2	6.5
P13	rs1573020	81	AG	AG	R	CTATCTGCCACCTGAGAGAGTATTG	AGGTGTCAGCTTCTTCTGACCAT	[ct]$_{14}$GAGTATTGCCAGCCTGATTC	1	1.3
P14	rs896788	78	CT	CT	R	GTAATGCCTCTGTGGCCCTAT	ATTCCGTCCACATCTTCACTG	[ct]$_{17}$tctACAGTCACCAGCCAC	1.5	1.3
P15	rs2065160	66	AG	AG	F	AAGAATGGCCTCTCGATGAGTA	GATGATACCTACGCATAGTCTGTTT ACTTC	t[ct]$_{14}$GCATAGTCTGTTTACTTCATTTG	1.5	1.3
P16a	rs2572307	63	AG	AG	F	GTGTAGCTATGCCATCATTCAATC	ATCCTTAGAAGGGTGCTAAACTGAG	t[ct]$_{15}$catcATTCAATCAATAGTCATAAAC	2.25	1.1
P17	rs2303798	96	AG	CT	R	CCAGCCTGCACCACTGTC	AGAGATGTGTTCAGGAAGAGGCTA	t[ct]$_{19}$cAGAAGAGGGACGTGGG	4.3	3.5
P18	rs2065982	88	CT	AG	R	CTTGGGGCAGTCTTTAAGTCCT	AGGAAGTGGTCAGTGCCAGTAG	[ct]$_{18}$GGAAAAAAAAGTCCTCTTTGGTAT	2	1.3
P19	rs3785181	156	CT	CT	F	CTCTGTTCAGTTTCAAAGTTCTGG	TTGTGTTCAAAAATTTCAATTAGGTT	t[ct]$_{20}$AGGGCATCTTATCTTGAGC	1.25	0.75
P20	rs881929	93	GT	GT	F	AGCTACCTGGTGTCTAACTC	TTGACCCAGTGGTTCTGAGC	[ct]$_7$aaactaggtgccacgtcgtgaaagtctgacaa GCCTCTTGCCAGCTCTG	3	2.5
P21	rs1498444	106	GT	AC	R	GGCTATTACCACATTAAGAGAAACTGC	CAGCCTCTCAATGCAAATGAT	t[ct]$_{20}$cGGTTGTCAGAATATTTGCTACA	5	3
P22a	rs1426654	74	AG	CT	R	AATTCAGGAGCTGAACTGCC	TGTTCAGCCCTTGGATTGTC	[ct]$_{24}$ttCGCTGCCATGAAAGTTG	3	4
P23	rs2026721	108	CT	AG	R	GAAGACTTTTTGCAAGCACGAG	GGCAAATGCTGTAAGAATCCAT	t[ct]$_{21}$AAATGATTGACATAGTAGGCTATTG	3.2	1.75
P24	rs4540055	76	AGT[a]	ACT	R	TGTGCCTCTGATCACTTTTGAATAC	CCTAGCCAACTCCAGAGTTCAT	[ct]$_{27}$cGAAGCAGTGATCAGCAC	1.2	1.1
A52	rs1335873	110	TA	AT	R	GTGGATGATATGGTTTCTCAAGG	TTCAACAAACGTGTGATGCTCT	t[ct]$_{25}$AGGTACCTAGCTATGTACTCAGTAT	1.2	1.75
P25a	rs16891982	78	CG	GC	R	GAATAAAGTGAGGAAAACACGGAGT	GTTTCTCATCTACGAAAGAGGAGTC	[ct]$_{28}$GGTTGGATGTTGGGGCTT	1.75	2.25
P26	rs730570	92	AG	CT	R	CAGCACCCTGTAAAGTCCAG	CAGCACTCACCTGCATCTCA	t[ct]$_{27}$tCATTAATCACACAAATTTTGCAT	1.3	2.2
A13	rs1886510	86	AG	AG	F	GTCCTTGTCAATCTTTCTACCAGAG	GGATTTTCACAACAACACTTGC	[ct]$_{27}$ACAAGATTTTCACAACAACACTTGC	1.5	1
P27	rs5030240	81	CGT[a]	CGA	R	CCAAAGTGCCAGGATCACAG	TCCCTAGAAATCCTTCAGCC	[ct]$_{32}$cCACAGGAGTGAGCCACTGC	2	2
P28	rs3827760	85	AG	AG	F	GCTCAGCTCCACGTACAAC	CTGTCATGCCCCCAATCTC	tctgacaactaggtgccacgtggtgaaagtctgacaa ctaggtgccacgtcgtgaaagtctgacaactctca GGYGCCAYGTTTTCACA[b]	3	2.75
A11	rs727811	78						H_2O	104.65	5.62
								[200 μM] $(NH_3)_2SO_4$	(SBE mix only)	33.33
								Total	179.45	108.8

[a] dbSNP bases, 1000 Genomes list AC bases only for both of these tri-allelic SNPs.
[b] Y bases denote equimolar primers used with C/T at two positions (rs144939741 and rs121908454).

Table 2
SNP peak positions for electrophoresis with POP4 polymer. Std Dev: standard deviation.

dbSNP rs-number	Internal code	Amplified bases	Theoretical size	G	A	C	T	SD allele 1 size	SD allele 2 size	SD allele 3 size
rs1321333	P03	CT	22			26.89	28.63	0.05	0.06	
rs917118	A07	AG	24	28.17	30.16			0.05	0.04	
rs1024116	A29	AG	28	26.75	28.95			0.02	0.03	
rs7897550	P05	CT	28			31.44	32.87	0.03	0.08	
rs722098	A21	AG	33	33.88	36.28			0.04	0.00	
rs10843344	P06a	CT	35			37.05	38.31	0.04	0.07	
rs239031	P07	CT	36			38.90	40.45	0.08	0.05	
rs12913832	P08	AG	38	40.15	40.90			0.06	0.07	
rs2040411	A40	AG	41	43.18	44.25			0.09	0.07	
rs1978806	P09a	CT	41			44.27	45.52	0.09	0.06	
rs773658	P10	GC	45	46.25		46.73		0.01	0.07	
rs10141763	P11	TA	45		48.74		49.32	0.09	0.05	
rs182549	P12	CT	49			49.11	50.55	0.04	0.04	
rs1573020	P13	AG	49	50.27	50.98			0.06	0.05	
rs896788	P14	CT	53			52.52	53.26	0.06	0.07	
rs2065160	P15	AG	53	54.11	54.82			0.07	0.04	
rs2572307	P16a	AG	57	55.41	56.02			0.06	0.01	
rs2303798	P17	CT	57			57.14	57.73	0.03	0.02	
rs2065982	P18	AG	61	59.17	59.90			0.04	0.03	
rs3785181	P19	CT	61			60.27	61.49	0.05	0.06	
rs881929	P20	GT	64	61.83			63.86	0.05	0.07	
rs1498444	P21	AC	65		64.17	64.01		0.06	0.10	
rs1426654	P22a	CT	68			67.35	68.19	0.04	0.06	
rs2026721	P23	AG	69	68.45	69.37			0.14	0.08	
rs4540055	P24	TCA	73		72.15	71.68	72.48	0.08	0.10	0.06
rs16891982	P25a	GC	75	76.03		76.44		0.13	0.10	
rs1335873	A52	TA	77		78.16		78.31	0.09	0.13	
rs1886510	A13	AG	80	78.71	79.69			0.10	0.07	
rs730570	P26	CT	80			80.11	80.74	0.08	0.11	
rs5030240	P27	GCA	85	83.75	84.91	84.51		0.03	0.04	0.05
rs2304925	P01	GT	87	86.61			87.93	0.09	0.11	
rs5997008	P02	AC	87		88.20	87.86		0.05	0.03	
rs3827760	P28	AG	90	89.64	90.09			0.04	0.04	
rs2814778	P04	CT	90			91.04	91.93	0.07	0.04	

were changed from the original protocol by increasing the initial 95 °C denaturation from 10 to 15 min and increasing the final 65 °C extension from 6 to 15 min. Amplification reactions used a total PCR volume of 6.9 μL, comprising: 1× AB *AmpliTaq Gold* PCR buffer; 25 mM $MgCl_2$ to a final concentration of 5.9 mM; 10 mM dNTP mix to a final concentration of 0.58 mM; 3.2 μg/μL BSA to a final concentration of 0.29 μg/μL, 0.5 units of AB *AmpliTaq Gold* polymerase, 1 μL of premixed PCR primers, and; 0.75 ng of target DNA. Individual primer pair ratios have also been changed since the original assay description and are detailed in Table 1. PCR cycling comprised: 15 min at 95 °C, 30–32 cycles of 95 °C for 30 s, 60 °C for 50 s, 65 °C for 40 s, then a final extension at 65 °C for 15 min.

PCR products were cleaned up prior to extension using combined exonuclease I and Shrimp alkaline phosphatase by adding 1.3 μL *Exo-SAPit* (USB Products, Affymetrix, Santa Clara, CA, U.S.) to a 2.5 μL aliquot of PCR product. Samples were incubated at 37 °C for 45 min then enzymes were heat-inactivated at 85 °C for 15 min. SNaPshot primer extension reactions have also been modified from the original protocol to a total SBE volume of 3.0 μL combining 1 μL of cleaned-up PCR product, 1 μL of SNaPshot ready reaction mix (previously 1.25 μL), 0.5 μL of pre-mixed SBE primers, as outlined in Table 1 (previously 0.75 μL), and 0.5 μL of water, SBE cycling used: 28–30 cycles of 96 °C for 10 s, 55 °C for 5 s, 60 °C for 30 s. SBE products were cleaned-up by adding 1.3 μL of SAP to the whole 3 μL reaction volume and incubating at 37 °C for 80 min and heat-inactivated at 85 °C for 15 min.

For CE detection, SBE products were diluted 1 in 25 and 1 μL of diluted product was added to a mix of 8.9 μL AB HiDi™ formamide plus 0.3 μL AB LIZ-120 size standard. All separations were made with AB 3130xl Genetic Analyser, POP-4™ polymer and 36 cm capillary arrays. Standard SNaPshot electrophoresis conditions were used and results analysed with AB Genemapper ID-X software. The injection of 1 μL of SBE product represents a three-fold reduction in the quantity of DNA analysed compared to the previously published protocol.

2.5. Population analysis

Reference genotypes were compiled from the HGDP-CEPH SNP*for*ID listings in SPSmart for the five continental population groups of Africa, Europe, East Asia, Oceania and America. This data was used to assess the predictive value of the modified 34-plex test with emphasis on assessing the enhancement of East Asian and closely related American population classifications that can be expected from the incorporation of rs3827760. Full HGDP-CEPH genotype data for rs3827760 has been added to the 34-plex tables of the SNP*for*ID browser at: http://spsmart.cesga.es/snpforid.php?-dataSet=snpforid34, while all rs727811 data is retained.

Classification success was estimated using the verbose cross-validation option of the *Snipper* SNP classification portal (at: http://mathgene.usc.es/snipper/ – choosing the 'Thorough analysis of population data with a custom Excel file' method), which removes each sample in turn and classifies it using the modified training set '$n-1$' allele frequency estimates. Population structure analysis was performed using *Structure 2.3.3*, applying three iterations for each *K* value from 2 to 8, with a 100,000 burn-in period followed by 100,000 MCMC steps after burn-in. We applied the admixture and correlated allele frequencies models and used prior specification of the population of origin of HGDP-CEPH reference samples (i.e. applying POPFLAG = 1) before analysing the population structure of individuals taken from 1000 Genomes or NIST population panels. Analyses of the NIST panel samples combined HGDP-CEPH

reference samples with additional SNP*for*ID populations taken from SPSmart: 60 each of Mozambican; Ugandan; Danish; NW Spanish; mainland Chinese and; Taiwanese, plus 144 samples from three Colombian populations. Plots were constructed using *CLUMPP 1.1.2* and *distruct 1.1* software. After establishing the statistically optimum *K* values for the reference samples using the 34 SNPs (both original and new SNP combinations) we analysed 1000 Genomes individuals combined with the nine control DNAs in simultaneous runs by treating them as population samples of unknown ancestry (i.e. POPFLAG = 0) to test the efficacy of the 34-plex assay in characterising populations with significant admixture levels. The 709 NIST population reference samples were analysed in identical fashion, pre-empting expected high levels of admixture by also setting POPFLAG = 0. An additional set of *Structure* analyses examined just the NIST U.S. population panel samples by themselves as an alternative approach to runs combining reference samples of known un-admixed ancestry (typically HGDP-CEPH panel individuals) in order to assess how closely admixture component ratios matched those obtained with runs combining NIST and reference samples.

3. Results and discussion

3.1. Performance of the modified 34-plex SNaPshot assay: SBE product mobility variation and peak height ratios

An example 34-plex electropherogram is shown in Fig. 2. This profile was chosen due to an above-average heterozygosity of 72% with 49 peaks from a total 68, so it illustrates the improved heterozygote balance we observed in comparison to patterns obtained from the previous assay conditions (see Fig. 1 of [12]). The migration positions of SBE products not present in this example profile are marked with the SNP and allele code in grey – note three positions are shown for tri-allelic SNPs P24 and P27.

Unlike forensic STR typing, profiles generated by SNaPshot tend to give a range of non-specific peaks that are often high enough to mimic product peaks, typically in the green dye channel. Our experiences with the previous 34-plex assay profiles consistently showed that non-specific signals fall outside the established allelic peak positions and can therefore be confidently excluded as allele extension products. This makes the recording of SBE product mobility variation an important step when validating forensic SNaPshot assays. Therefore we used the genotyping of 712 NIST reference samples to optimise the reconfigured assay and, with reference to multiple runs per sample, plotting all allele mobilities and peak height ratios. Results for these extended electrophoresis analyses are summarised in Tables 2 and 3. Mean base pair size estimates for each SNP allele were made using standardised conditions of POP-4^{TM} in 36 cm capillary arrays with an AB 3130xl Genetic Analyser. The observed allele sizes with the above CE conditions are shown in Table 2. Allele size estimates were found to be very stable across multiple samples, runs, and reactions. Size values for all 34 SNPs in the modified assay gave an average standard deviation across 70 positions of ±0.06 bp and values predominantly equal to or less than 0.1 bp (highest ±0.14 bp), but for most SNPs this was much smaller. Therefore any signal peaks falling outside the highly conservative ±0.5 bp size variation window routinely applied to allelic peak positions can be securely designated as artifactual.

Heterozygote imbalance has a more disruptive effect on SNaPshot genotyping consistency than occurrence of non-specific peaks. Large disparities in reporter dye signal strength, varying

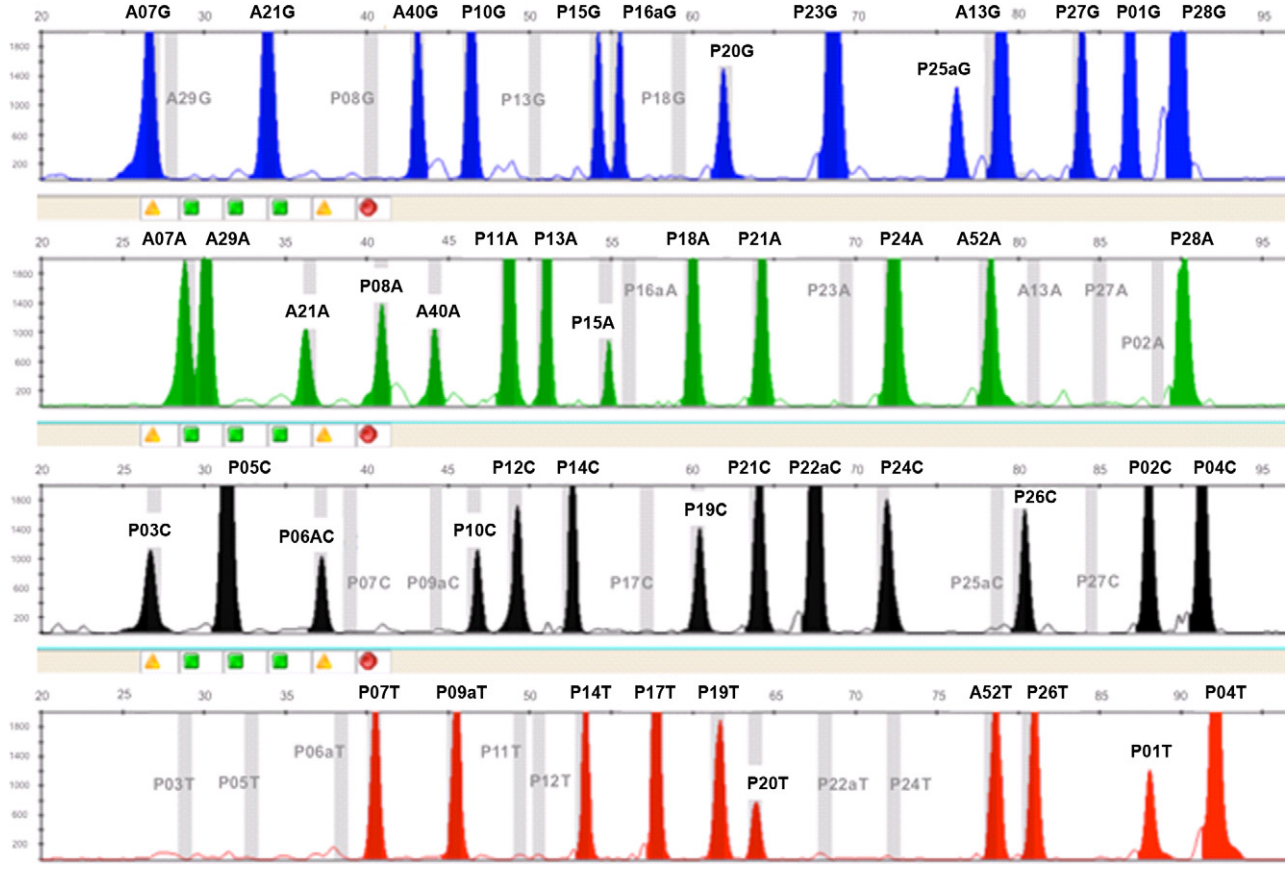

Fig. 2. Example electropherogram obtained from the revised 34-plex assay. SNP allele peaks marked with internal codes listed in Table 1.

Table 3
Mean heterozygote peak height ratios with SNPs in same order as Fig. 3.

Marker	Mean	Standard deviation	Marker	Mean	Standard deviation	Marker	Mean	Standard deviation	Marker	Mean	Standard deviation
T/C heterozygotes			G/A heterozygotes			G/T heterozygotes			A/T heterozygotes		
P24-CT	1.2092	0.1692	P27-AG	1.6372	0.1701	P20	2.0731	0.5997	P24-AT	1.7632	0.4127
P04	0.9089	0.1153	P28	5.0806	1.4789	P01	2.3225	0.0398	A52	0.9887	0.1182
P26	1.6778	0.1455	A13	1.4742	0.2727	Total GT	2.1978	0.5183	P11	1.1247	0.1829
P22a	1.2563	0.1349	P23	2.0429	0.5083				Total AT	1.2922	0.2380
P19	1.5359	0.2772	P18	1.7116	0.5266	A/C heterozygotes					
P17	2.1853	0.3817	P16	1.4132	0.3034	P27-AC	2.0240	0.3388	C/G heterozygotes		
P14	1.2118	0.1916	P15	2.3777	0.2960	P24-AC	2.1612	0.5010	P27-CG	3.5695	0.7720
P12	1.0979	0.2036	P13	1.0978	0.2317	P21	1.1858	0.1360	A25a	3.1853	0.6303
P09a	1.2072	0.1850	A40	1.5639	0.3181	P02	1.7830	0.2528	F2	3.0982	0.42424
P07	1.1672	0.1331	P08	4.8865	0.6343	Total AC	1.7885	0.3071	Total CG	3.3215	0.6089
P06a	1.3641	0.1858	A21	3.5275	1.0584						
P05	1.1904	0.3136	A07	2.3412	0.9151						
P03	1.3096	0.2822	A29	1.1750	0.0912						
Total CT	1.3324	0.2091	Total AG	2.3330	0.5234						

primer extension efficiencies between the four terminator dideoxy nucleotides and stochastic PCR effects can all impact on imbalanced signals from SBE products. While unavoidable, SNaPshot heterozygote imbalance can usually be accommodated by assigning pre-set peak height ratio (PHR) limits. Previously published guidelines [20,21] suggest that for equal amounts of dye molecules passing the detection window, recorded RFU peak heights for a G-terminating nucleotide (DR110 dye) are generally two times that of a corresponding A (DR6G) and 4 times that of corresponding C (TAMRA) or T (ROX). To explore the widest range of heterozygote peak pairs we collected RFU values from 20 heterozygotes in 80 amplified samples of European, African American, Hispanic and East Asian origin, for each of the 34 SNPs, ensuring many more heterozygotes could be analysed than normally possible in any one population group with these SNPs. The distribution of peak height ratios from these samples is shown in Fig. 3 and the mean values for each PHR range are listed in Table 3.

Amplifications were performed with three separate PCR reactions but this was not observed to have a significant effect on peak height ratios (data not shown), indicating stochastic PCR effects represent a minor source of signal variation. CT heterozygotes showed a mean T/C PHR value of 1.33 (i.e. higher T peak) with a standard deviation of 0.21 and SNP P17 a single outlier with a mean T/C = 2.19. AC and AT heterozygotes showed similar minor

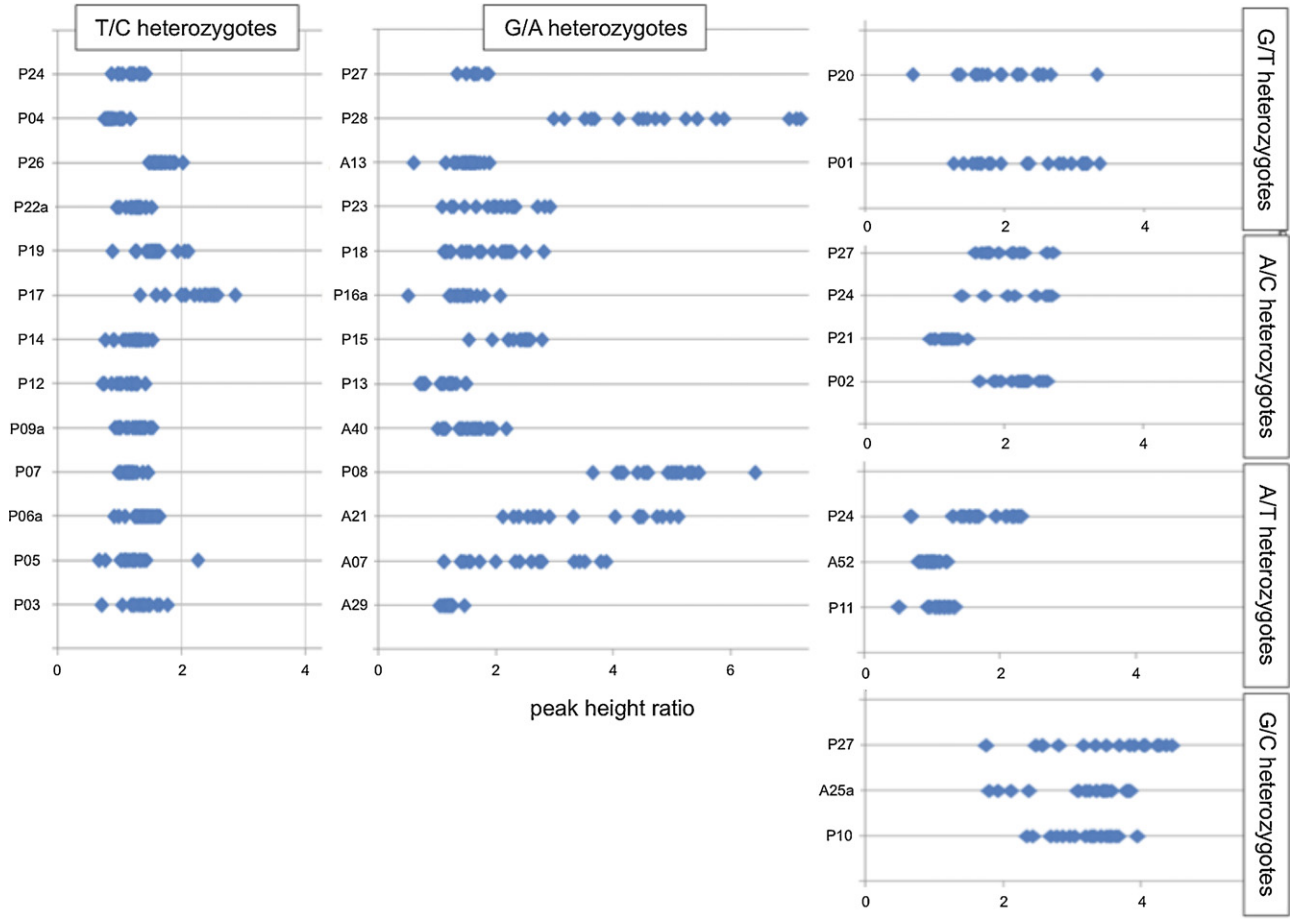

Fig. 3. Distribution of peak height ratios recorded in the heterozygotes of 80 NIST population panel samples.

variation around mean PHRs of A/C = 1.79 (±0.31) and A/T = 1.29 (±0.23) respectively. The DR110-linked G allele shows the highest deviations from predicted PHR values particularly in AG heterozygotes and it is worth noting the DR110 dye displays the highest energy emission of all the four dyes used in SNaPshot. Some AG SNPs display PHRs close to 1 with little variation, particularly SNPs A07 and A29, but these contrast with the new SNP P28 as well as A21 and P08 that have high between-run variation (up to 0.5) and PHR values greater than predicted. GT peak pairs were the most divergent from the predicted PHR of 4 being closer to a value of 2, likely due to the higher emission energy from the ROX dye of T compared to the TAMRA of C. The standard deviation from average peak height in all G-based heterozygotes was greater than 0.5: more than twice the value of the other heterozygotes and evident in the broader spread of G peak ratios in Fig. 3. Therefore average PHRs collated in our study for each SNP and summarised in Table 3, should be kept in mind when interpreting 34-plex SNaPshot profiles, especially those generated from difficult or scant DNA sources where peaks are likely to be much lower than normal and show greater stochastic variation.

3.2. Population analyses

Structure analysis of HGDP-CEPH reference samples alone using both original and revised 34-plex SNP sets are presented in Fig. 4A. The optimum *K* value corresponds to five population clusters in both cases, as shown by the *L*(*K*) mean value plots in Fig. 4B. Average ancestry membership proportions (from the HGDP-CEPH analysis) are listed in Table 4A and these show an increase in the values associated with the East Asian ancestry component using the new SNP combination, seen as a median value of 91.16% raised

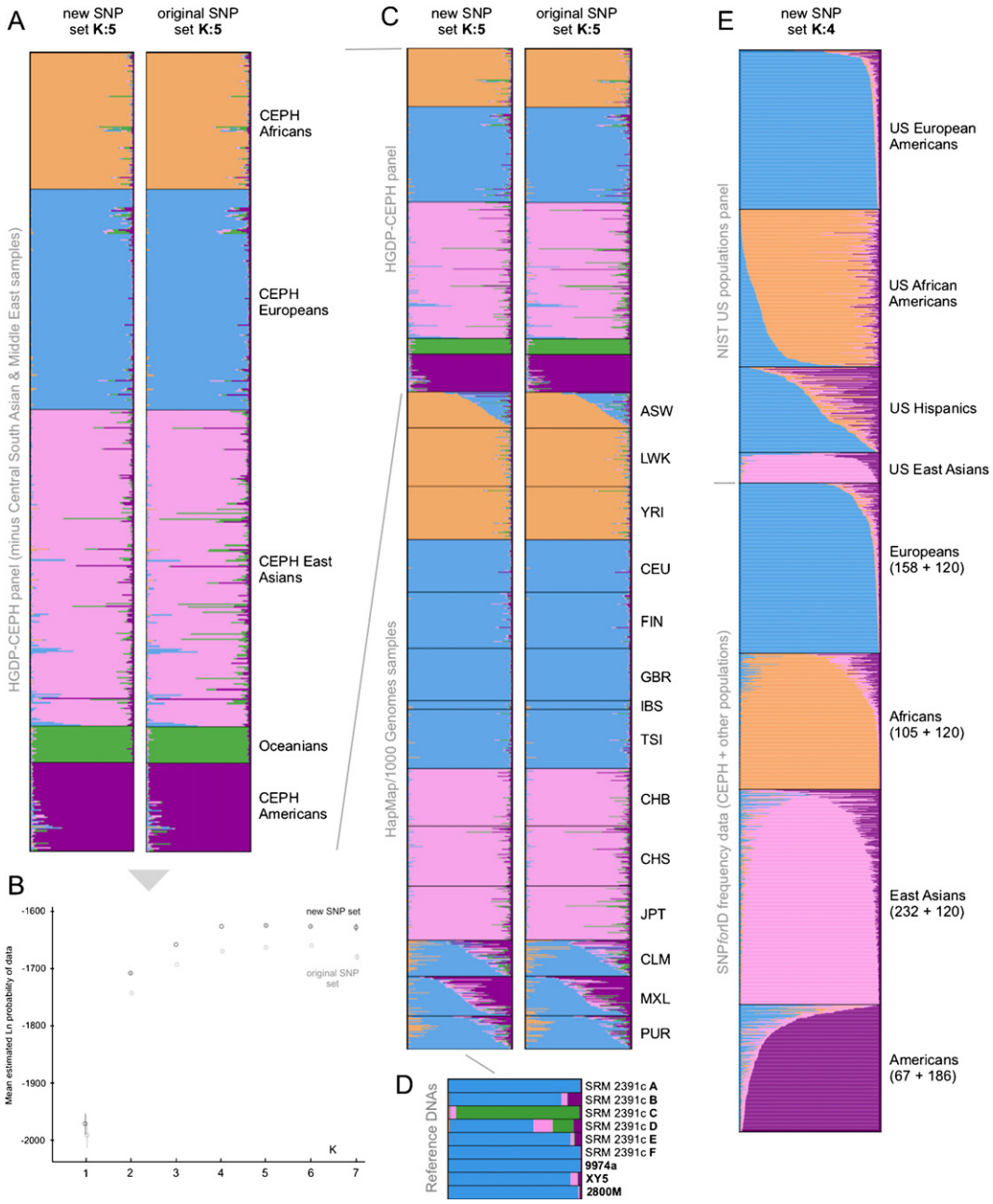

Fig. 4. Cluster plots summarising five *Structure* analyses performed on combinations of reference (HGDP-CEPH plus SNP*for*ID populations) and unknown populations (NIST SRM 2391/control DNAs, 1000 Genomes populations and NIST U.S. panel). Structure runs comprised, (A) HGDP-CEPH alone, old and new SNP sets, with the optimum *K* value of 5 indicated in by the likelihood *L*(*K*) plot of (B and C) HGDP-CEPH with unknown 1000 Genomes and SRM 2391c/control DNAs (expanded in D), (E) HGDP-CEPH plus SNP*for*ID with unknown NIST U.S. population panel and new SNPs only giving an optimum *K*:4. 1000 Genomes population descriptors as in Fig. 1.

Table 4A
Average membership proportions to *K*:5 clusters from the Structure analysis of HGDP-CEPH populations plotted in Fig. 4A. Values in bold indicate improved performance with the revised SNP combination.

Previous 34-plex (admixture model)							Revised 34-plex (admixture model)						
Given population	Proportion of membership of each pre-defined population in each of the 5 clusters					Number of individuals	Given population	Proportion of membership of each pre-defined population in each of the 5 clusters					Number of individuals
	Inferred clusters							Inferred clusters					
	AFR	EUR	EAS	OCE	NAM			AFR	EUR	E ASN	OCE	AME	
African	**0.9560**	0.0060	0.0110	0.0170	0.0100	98	African	**0.9580**	0.0067	0.0080	0.0190	0.0083	98
European	0.0060	**0.9530**	0.0120	0.0130	0.0160	158	European	0.0057	**0.9573**	0.0100	0.0130	0.0140	158
East Asian	0.0070	0.0250	**0.8764**	0.0546	0.0370	227	East Asian	0.0070	0.0230	**0.9116**	0.0270	0.0313	227
Oceanian	0.0060	0.0050	0.0123	0.9687	0.0080	26	Oceanian	0.0070	0.0060	0.0130	0.9630	0.0110	26
Native American	0.0060	0.0173	0.0190	0.0130	**0.9446**	63	Native American	0.0060	0.0173	0.0190	0.0117	**0.946**	63

Table 4B
Classification success (bold % values) from cross validation of the 5 group training set listed in supplementary Table S3. Values in bold indicate improved performance with the new SNP combination.

Previous 34-plex, 5 group training set of supplementary Table S3, worksheet 2						Revised 34-plex, 5 group training set of supplementary Table S3, worksheet 22					
	Estimation with cross-validation to compute the success ratio using the best 34 SNPs:						Estimation with cross-validation to compute the success ratio using the best 34 SNPs:				
	Africa	Europe	East Asia	Oceania	America		Africa	Europe	East Asia	Oceania	America
Population of African origin	**100%**	0.00%	0.00%	0.00%	0.00%	Population of AFRICA origin	**100%**	0.00%	0.00%	0.00%	0.00%
Population of European origin	0.00%	**98.73%**	0.63%	0.00%	0.63%	Population of EUROPE origin	0.00%	**99.37%**	0.00%	0.00%	0.63%
Population of East Asian origin	0.00%	0.00%	**92.51%**	0.88%	6.61%	Population of EAST ASIA origin	0.00%	0.00%	**94.71%**	0.44%	4.85%
Population of Oceanian origin	0.00%	0.00%	3.85%	**96.15%**	0.00%	Population of OCEANIA origin	0.00%	0.00%	0.00%	**100%**	0.00%
Population of American origin	0.00%	0.00%	0.00%	0.00%	100%	Population of AMERICA origin	0.00%	0.00%	0.00%	0.00%	100%

from a previous 87.64%. The revised SNP set creates rather cleaner *Structure* cluster plots for East Asians at $K = 5$ in HGDP-CEPH samples shown in Fig. 4A and to a lesser extent in the CHB, CHS, JPT samples of the combined HGDP-CEPH-1000 Genomes analyses described below and shown in Fig. 4C.

Analysing 1000 Genomes individuals as study samples of 'unknown' ancestry (lower half of Fig. 4C) provides results assigning all samples to their expected population clusters. Four populations are arranged in a sequence to indicate rising second admixture component proportions showing the range as a gradient of two-cluster memberships in the following 1000 Genomes populations: ASW (African Americans from Southwest U.S.); CLM (Colombians from Medellín); MXL (Mexicans from LA district, CA); and PUR (Puerto Ricans). Individual sample membership proportions in K5 clusters for both original and new 34-plex SNPs (K4 for NIST populations) estimated from the six *Structure* analyses of Fig. 4A–E, are listed in supplementary Table S1. These indicate ASW samples have a range of 5–45% European co-ancestry with African the major component and PUR; CLM; MXL show increasing levels of American co-ancestry with European the major component plus indications of a significant African third component in many Colombians and Puerto Ricans. A large number of Mexicans (MXL) have comparable proportions of European to Native American ancestry across the population sample, while many PUR, CLM and MXL show their predominant ancestry is European, not Native American.

The *Structure* analysis of forensic reference/control DNAs is presented in Fig. 4D. The patterns observed for the ancestries of SRM 2391c A, B and C components match those reported by NIST (cluster membership proportions are given in supplementary Table S2), while the other reference/control DNAs are consistent with being unadmixed European ancestries.

An initial *Structure* analysis was made for the NIST population panel alone (i.e. excluding reference populations) to measure the genetic structure discernible in this sample of the U.S. The cluster plots of the optimum $K = 4$ (revised SNP set only, and with cluster colours modified to avoid overlap with Fig. 4 plots) are shown in supplementary Fig. S1. Four genetic clusters were detected despite the heterogeneity of the sample set and a largely cultural rather than a population genetics-based definition of Hispanics. The analysis therefore detects four clusters in the NIST panel that closely correspond to the population definitions given to the samples. Furthermore most Hispanics are differentiated by showing a distinct genetic component that does not correspond to the clusters defining the other three U.S. populations.

With the observation that four clusters were detected in the NIST U.S. panel we then performed *Structure* analysis adding reference populations and the resulting *K*:4 cluster plot is shown in Fig. 4E (revised SNP set only). In this analysis reference populations were expanded to include several SNP*for*ID populations with admixture. As with the admixed populations in Fig. 4C (ASW, CLM, MXL, PUR), NIST individuals have been ordered in ascending major co-ancestry component (except descending for US African Americans). It is evident that each NIST group corresponds well with the expected population definitions. U.S. Caucasians and U.S. East Asians show largely unadmixed European and East Asian ancestry. In contrast, a large proportion of African Americans show significant levels of European co-ancestry as a major second component. U.S. Hispanics in particular show considerable heterogeneity and complex patterns of population admixture. The largest proportion of Hispanics show majority European co-ancestry, there is a wide range of Native American co-ancestry proportions and we found detectable African co-ancestry as a third component in most samples. It is evident that U.S. Hispanic individuals have a higher Native American co-ancestry compared to other U.S. population groups, which helps to explain the differentiation pattern observed in supplementary Fig. S1.

As it is difficult to characterise the overall pattern of admixture in any one population when there is a continuous range of detected second or third co-ancestry, we summarised the data from all samples analysed in the *Structure* run of Fig. 4E by compiling mean cluster membership proportions with 1st (25th percentile) and 3rd (75th) quartile ranges for the eight population groups analysed

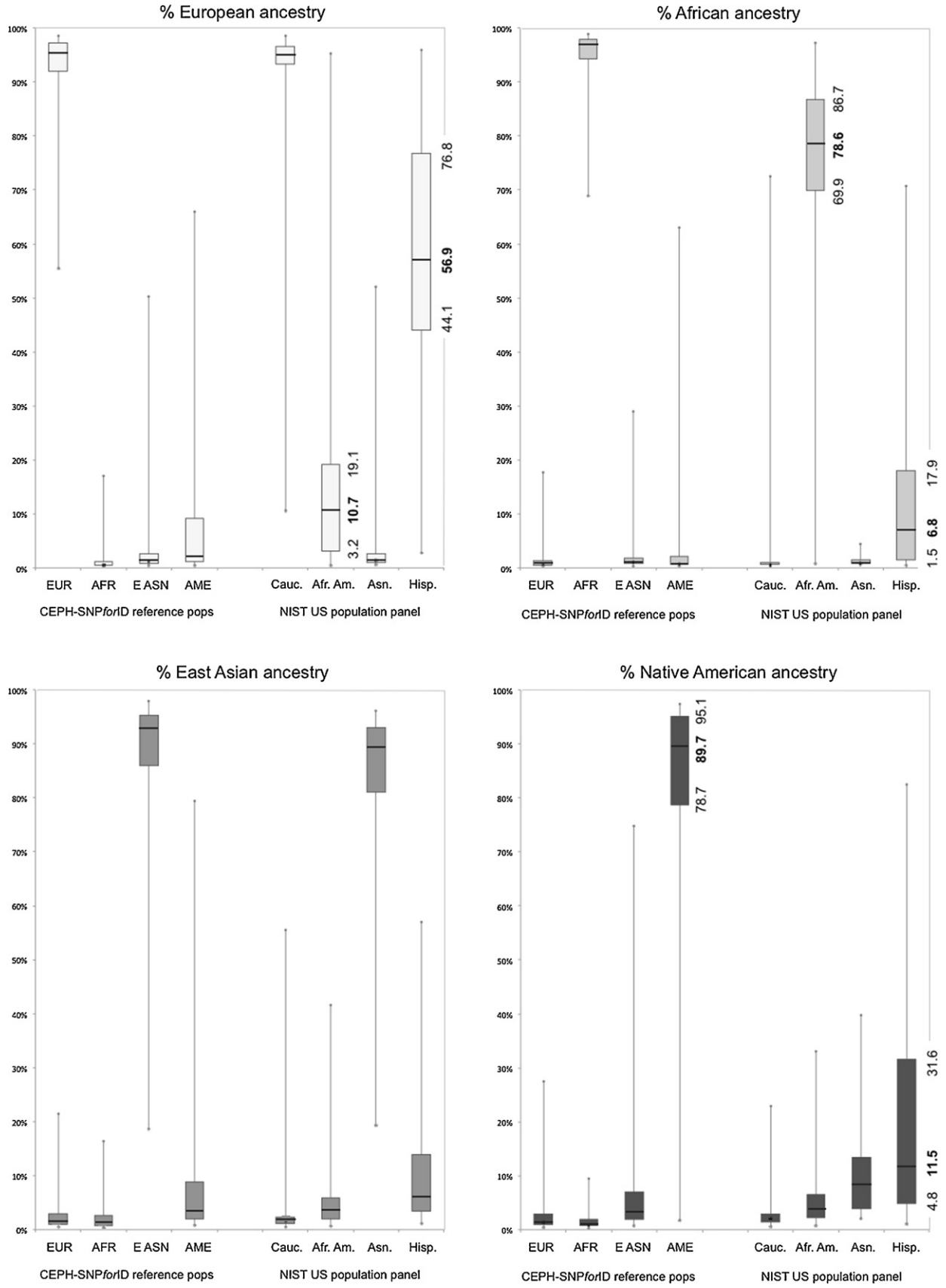

Fig. 5. Distribution of cluster membership proportions (% proportions of membership to an optimum *K*:4 clusters) shown by NIST U.S. population panel samples and HGDP-CEPH/SNP*for*ID reference populations. These values were obtained from the *Structure* run of Fig. 4E. Boxes indicate 1st and 3rd quartile ranges and contain the mean membership proportion for each population group and *K* ancestry. Whisker lines indicate the full range of values observed. Values placed next to certain selected populations show mean and quartile range values.

using the data of supplementary Tables S1. These are shown in Fig. 5 divided into four ancestry proportion plots based on the range of membership values to the four clusters. The boxes denote 1st–3rd quartile ranges each containing the mean value line with whiskers indicating the full range. The lowest means and narrowest quartile ranges are found in the reference European, African and East Asian proportions for their alternative ancestries (e.g. Europeans have minimal African/East Asian ancestry, etc.) and near identical patterns describe the NIST U.S. European and U.S. East Asian samples. NIST Hispanics show the highest mean value for European ancestry of 57% and this has the broadest quartile range, while the mean American ancestry proportion is only 11.5% but with an equally broad quartile range. In the Hispanics the observed African ancestry has a mean of 6.8% so on average this represents a quite minor third component of co-ancestry but the quartile box shows a significant range reaching 18% at the 3rd quartile, so for many NIST Hispanics African ancestry is a detectably larger component. Therefore, as a group, the NIST Hispanics can be described as showing predominantly European ancestry with comparatively minor proportions of Native American co-ancestry plus detectable but generally much lower African co-ancestry. This pattern is in sharp contrast to the HGDP-CEPH and SNP*for*ID Native American reference populations that have a mean 90% American ancestry and minimal European or African co-ancestry. NIST African Americans give a consistent pattern of African major ancestry ranging from 70 to 85% and European co-ancestry of about 5–20% in most samples.

The classification success rates from *Snipper* cross validation of the HGDP-CEPH panel samples are outlined in Table 4B, indicating enhanced performance of European, East Asian and Oceanian classifications, with 0.64%, 2.20% and 3.85% improvements in assignment success respectively. Although the 34-plex test was not designed to infer Native American or Oceanian ancestry, the enhanced differentiation of East Asians provided by rs3827760 brings improvements to the analysis of all five global groups.

3.3. Reference/control DNA profiles and AIM-SNP genotypes in seventy populations

SNP profiles for the reference/control DNA samples were compiled through independent, triplicated typing in two laboratories with full consensus in all cases and these genotypes are listed in supplementary Table S2. Fig. 4D shows the *Structure* analysis of the nine reference DNAs detecting the Oceanian ancestry of SRM 2391c-C, with ~93% membership to the Oceanian cluster identified from the HGDP-CEPH reference samples. The only other SRM 2391c component to show a significant non-European component (excluding the mixed sample D) was sample B with 10.4% membership to the Native American cluster identified from the HGDP-CEPH reference samples, this division of European and American co-ancestry broadly matches values seen in the admixed American populations of CLM, MXL and PUR as well as many NIST Hispanics. The ancestry membership proportions (K5 = five population clusters) obtained from *Structure* analysis of the nine reference DNAs are listed in supplementary Table S2.

Supplementary Table S2 also lists the likelihoods and ancestry assignments made for the reference DNAs based on the revised 34-plex SNP profiles submitted to *Snipper*. Assignments were made using a custom five group HGDP-CEPH training set (AFR-EUR-E ASN-AME-OCE). Two HGDP-CEPH based training sets are included ready to use as an uploadable Excel file in supplementary file S3, though note that one worksheet must be removed in use (revised or previous set) and that SNPs and therefore submitted profiles are ordered in the files by ascending rs-number, not SNaPshot mobility. The previous rs727811 component substitutes rs3827760 in the second training set worksheet. The Oceanian classification of SRM 2391c-C matches the *Structure* inferences described in Section 3.2 and we interpret this as an unequivocal ancestry assignment. Five reference/control DNAs (those except SRM 2391c-B to E) gave very strong European classifications of 1.E+15 to 1.E+18 times more likely European than the next closest assignment (American or East Asian). Two SRM controls, SRM 2391c-B and E gave reduced European assignments, although the non-European co-ancestry indicated by *Snipper* and measured by *Structure* as 10% and 5% American membership proportions respectively, is based on SNPs that have less power to differentiate this ancestry compared to EUR-AFR-E ASN.

Full genotype data from the 52 HGDP-CEPH populations, fourteen 1000 Genomes populations and four NIST U.S. populations for 34 component SNPs plus rs3827760 is listed in supplementary Tables S4. As the Stanford/Michigan HGDP-CEPH SNP data accessible in SPSmart (http://spsmart.cesga.es/) does not include rs3827760 we have also listed separately the HGDP-CEPH genotypes for this SNP. The 1000 Genomes genotype tables also provide allele frequency estimates for each population although it should be noted that the two tri-allelic SNPs rs4540055/rs5030240 (P24/27) are both given as binary AC substitutions by 1000 Genomes and this discrepancy is discussed in more detail in Section 3.4.

3.4. Tri-allelic AIM-SNPs rs4540055 and rs5030240

Two of the most informative components of the 34-plex assay are tri-allelic SNPs: rs4540055 and rs5030240, both having three alleles recorded at a single variable nucleotide position (A/C/T and A/C/G respectively) in all population groups, albeit at low frequencies for the third allele in some populations (rs4540055-C and rs5030240-A). We made two noteworthy observations concerning these tri-allelic SNPs: (1) their evident success in detecting three alleles in the two mixture contributors of SRM2391c D and; (2) the disparity between the allele frequency distributions obtained from our SNaPshot typing compared to the data from the next-generation deep re-sequencing analysis of 1000 Genomes as listed in supplementary Table S4.

The original motivation for identifying and characterising tri-allelic SNPs was to bring mixture detection properties into forensic SNP typing assays [22] and this approach has been more thoroughly explored in recent studies by Westen et al. [23]. Since tri-allelic SNP allele frequencies commonly show strong population differentiation it is also not surprising to see three alleles in both markers from a mixture of a European and an Oceanian donor. For instance, the rs4540055-C allele is almost eight times more frequent in Oceanians than Europeans and rs5030240-C more than eighteen times more frequent in Europeans. So while tri-allelic SNPs are much weaker than STRs at detecting multiple alleles in mixed donor profiles, the probability of finding more than two alleles is raised if donors are from different populations, potentially helping to identify the ancestry of components in simple mixtures. The three-peak patterns observed in SRM 2391c-D are shown in Fig. 6A.

The second, immediately apparent characteristic of tri-allelic SNPs seen in the allele distribution comparisons in this study, is their failure to be characterised as three allele SNPs by both the HapMap and 1000 Genomes projects. Each project has reported the two 34-plex tri-allelic SNPs as well as those additionally identified as tri-allelic SNPs by Westen [23] as binary SNPs. There appears to be problems with the genotyping of tri-allelic SNPs with next-generation whole genome sequencing approaches, and it is significant that 1000 Genomes have changed the allele calls of rs4540055 from AT to AC and rs5030240 from AG to AC between published data revisions of December 2010 and August 2011. Comparisons of the allele frequency distributions of both tri-allelic SNPs from our typing with 1000 Genomes sequencing are summarised in Fig. 6B and these appear to be quite disparate.

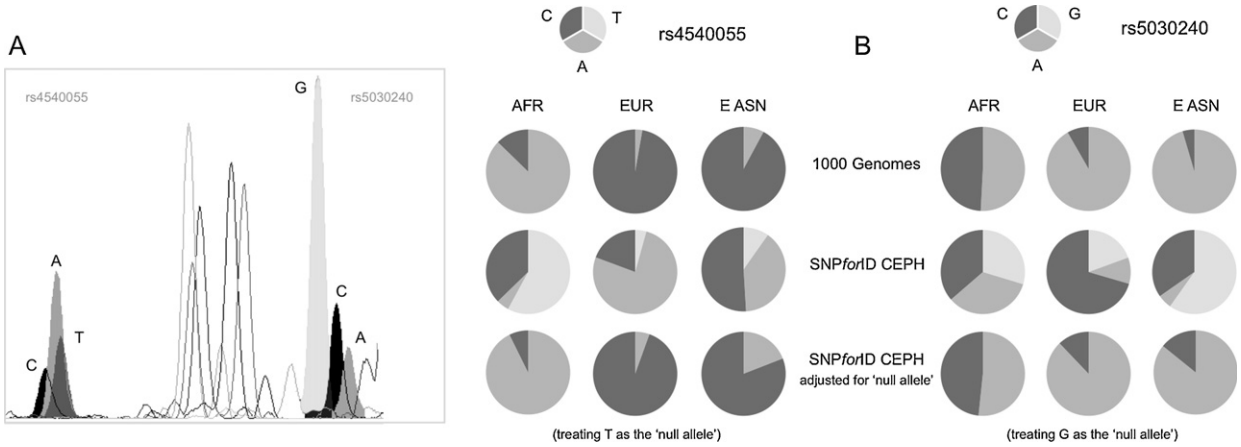

Fig. 6. (A) Observed tri-allelic SNP peak patterns in the mixed-source SRM 2391c-D sample. (B) Allele frequency estimates for both tri-allelic SNPs in 34-plex obtained from next generation sequencing by 1000 Genomes (top row) and SNaPshot typing of the HGDP-CEPH panel (middle row). The bottom row pie charts indicate the HGDP-CEPH allele frequencies found when the T allele is treated as silent (undetectable) in rs4540055 and the G allele in rs5030240, giving a close match to the estimates from 1000 Genomes and explaining the discrepant values.

We have explained these differences by treating one of the alleles as silent, for example when rs4540055-T is considered silent or undetectable, then AT appears as an AA homozygote and CT as a CC. Therefore the A frequency can be estimated as: $(AA + AT + (AC \times 0.5)/\text{total genotypes})$ and $C = (CC + CT + (AC \times 0.5)/\text{total})$. When these adjustments are made to the HGDP-CEPH panel genotypes made by SNP*for*ID SNaPshot typing, then the allele frequencies are much better matched. Fig. 6B shows these HGDP-CEPH panel genotype adjustments made by assuming silent rs4540055-T and rs5030240-G alleles. However, it is discouraging that one of the benefits of such extensive sequencing initiatives: an enhanced ability to locate and identify SNPs with more than two alternative nucleotides, appears to be lost in progressing to next generation sequencing technologies. This is further exacerbated by 1000 Genomes using multiple sequencing centres/platforms but organising analyses to concentrate on samples from one population in any one centre.

4. Concluding remarks

Swapping a single component SNP at the same time as making adjustments to the chemistry and certain mobility positions in the 34-plex assay provided relatively little disruption to routine forensic SNP typing with this ancestry analysis test. The enhancement to the differentiation of East Asians is evident amongst the seventy populations characterised and the comparisons made in this study. As East Asians had the weakest population differentiation using the original 34 SNP test, the introduction of a near-fixed SNP in this population group represents a major enhancement of the test without upsetting its forensic performance. While SNaPshot primer extension assays remain the only viable multiplexed SNP typing option available to forensic analysis it is crucial to thoroughly gauge the mobility and signal variation entailed with using this chemistry as we have done with the 34 sets of peak pair data observed with the NIST U.S. population panel.

The Hispanics that were studied as part for the NIST panel highlight the difficulties of analysing individuals with complex, potentially three-way, patterns of admixture. Not only is it difficult to make accurate assessments of co-ancestry when these are so varied within the group as a whole, but also component AIM-SNPs in small forensic multiplexes such as the 34-plex test cannot provide equally informative differentiations of all co-ancestry contributors. With this in mind we have continued development of additional AIM-SNP based ancestry tests, complimentary to the 34-plex, that improve the differentiation of Native Americans from EUR-AFR-E ASN populations. It is also likely that an optimum strategy will be the addition of ancestry-informative indels in small multiplexes with similar properties to the 34-plex SNPs. We recently published details of an AIM-indel set [24] that is certain to improve the inferences made for complex admixture patterns when combined with the current 34-plex SNPs and bring improved differentiation of Native American ancestry. Therefore forensic ancestry inferences based on typing a range of short binary markers will show improved precision while keeping multiplexes small-scale, manageable and with sufficient sensitivity for challenging forensic material. Additionally, we continue to advocate the use of tri-allelic SNPs as highly differentiated ancestry markers. Regrettably, the opportunity to build a well-validated catalog of tri-allelic SNPs from the output of 1000 Genomes is not available to forensic genetics research at this moment and consequently their number remains small.

Revisions to the 34-plex primers, chemistry and cycling conditions added since the original development of the assay have evolved gradually and therefore it is difficult to state how each of these changes has influenced the improved peak height balance we report. However the biggest advance in 34-plex electropherogram quality was obtained with the changes to primer ratios that accompanied changes to the positions of four SNPs. Specifically, the largest improvement resulted from increasing the PCR and EXT primers for rs2814778 (P04) and rs16891982 (P25a), while reducing those of rs2304925 (P01).

Lastly, at the time of writing, a new study of the statistical limitations of familial searching strategies has highlighted the problem of using inappropriate STR allele frequencies when the ancestry of the profile donor is not known [25]. Using generalised or incorrect allele frequencies can lead to a large number of adventitious matches and the search loses specificity. Therefore it appears that a simple forensic ancestry test with sufficient power to distinguish the major population groups could be applied as an informative adjunct to conventional STR profiling and help enhance the power such tests already provide. This approach would not involve database-wide SNP typing as only the profile would require an inference of ancestry to select the appropriate STR allele frequencies in each search.

Disclaimer

This work was funded in part through funding from the U.S. FBI Biometric Center of Excellence: 'Forensic DNA Typing as a Biometric Tool'. 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. Departments of Justice or Commerce. 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.

Acknowledgements

The work of MF at NIST has been supported by the Fundacion Barrie de la Maza Postgraduate Grant Program (2010). CS was awarded a PhD grant to by the Portuguese Foundation for Science and Technology (SFRH/BD/75627/2010 and co-financed by the European Social Fund (Human Potential Thematic Operational Program). AFA was supported by a María Barbeito grant from Xunta de Galicia. MVL was supported by funding from Xunta de Galicia INCITE 09 208163PR and Ministerio de Educación y Ciencia BIO2006-06178.

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.2012.06.007.

References

[1] M. Fondevila, C. Phillips, N. Naveran, L. Fernandez, M. Cerezo, A. Salas, Á. Carracedo, M.V. Lareu, Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur, Forensic Sci. Int. Genet. 2 (2008) 212–218.
[2] A. Freire-Aradas, M. Fondevila, A.-K. Kriegel, C. Phillips, P. Gill, L. Prieto, P.M. Schneider, Á. Carracedo, M.V. Lareu, A new SNP assay for identification of highly degraded human DNA, Forensic Sci. Int. Genet. 6 (2012) 341–349.
[3] C. Romanini, M.L. Catelli, A. Borosky, M. Salado-Puerto, R. Pereira, C. Phillips, M. Fondevila, A. Freire, C. Santos, Á. Carracedo, M.V. Lareu, L. Gusmao, C.M. Vullo, Typing short amplicon binary polymorphisms: supplementary SNP and Indel genetic information in the analysis of highly degraded skeletal remains, Forensic Sci. Int. Genet. 6 (2012) 469–476.
[4] P. Gill, D.J. Werrett, B. Budowle, R. Guerrieri, An assessment of whether SNPs will replace STRs in national DNA databases – joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDAM), Sci. Justice 44 (2004) 51–53.
[5] M.V. Lareu, M. García-Magariños, C. Phillips, I. Quintela, Á. Carracedo, A. Salas, Analysis of a claimed distant relationship in a deficient pedigree using high density SNP data, Forensic Sci. Int. Genet. 6 (2012) 350–353.
[6] C. Phillips, M. Fondevila, M. García-Magariños, A. Rodriguez, A. Salas, Á. Carracedo, M.V. Lareu, Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers, Forensic Sci. Int. Genet. 2 (2008) 198–204.
[7] M. Kayser, P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, Forensic Sci. Int. Genet. 3 (2009) 154–161.
[8] M. Kayser, M.P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.
[9] W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pośpiech, S. Walsh, T. Kupiec, A. Wojas-Pelc, M. Kayser, Model-based prediction of human hair color using DNA variants, Hum. Genet. 129 (2011) 443–454.
[10] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, O.M. Kayser, IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Forensic Sci. Int. Genet. 5 (2011) 170–180.
[11] J. Mengel-From, C. Børsting, J.J. Sanchez, H. Eiberg, N. Morling, Human eye colour and HERC2, OCA2 and MATP, Forensic Sci. Int. Genet. 4 (2010) 323–328.
[12] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M. Casares de Cal, D. Ballard, M.V. Lareu, Á. Carracedo, The SNPforID consortium, inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci. Int. Genet. 1 (2007) 273–280.
[13] O. Lao, P.M. Vallone, M.D. Coble, T.M. Diegoli, M. van Oven, K.J. van der Gaag, J. Pijpe, P. de Knijff, M. Kayser, Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA, Hum. Mutat. 31 (2010) E1875–E1893.
[14] I. Halder, M. Shriver, M. Thomas, J.R. Fernandez, T. Frudakis, A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications, Hum. Mutat. 29 (2008) 648–658.
[15] C. Børsting, C. Tomas, N. Morling, SNP typing of the reference materials SRM 2391b 1-10, K562, XY1, XX74, and 007 with the SNP*for*ID multiplex, Forensic Sci. Int. Genet. 5 (2011) e81–e82.
[16] M.C. Kline, C.R. Hill, J.L. Almeida, E.L.R. Butts, M.D. Coble, J.M. Butler, The latest and greatest NIST PCR-based DNA profiling standard: updates and status of Standard Reference Material® (SRM) 2391c. Profiles-in-DNA, Promega Corporation Web site: http://www.promega.com/resources/articles/profiles-in-dna/2011/the-latest-and-greatest-nist-pcr-based-dna-profiling-standard/.
[17] C. Phillips, R. Fang, D. Ballard, M. Fondevila, C. Harrison, F. Hyland, E. Musgrave-Brown, C. Proff, E. Ramos-Luis, B. Sobrino, Á. Carracedo, M.R. Furtado, D. Syndercombe Court, P.M. Schneider, SNPforID consortium, evaluation of the Genplex SNP typing system and a 49plex forensic marker panel, Forensic Sci. Int. Genet. 1 (2007) 180–185.
[18] C. Tomas, G. Axler-DiPerte, Z.M. Budimlija, C. Børsting, M.D. Coble, A.E. Decker, A. Eisenberg, R. Fang, M. Fondevila, S.F. Fredslund, S. Gonzalez, A.J. Hansen, P. Hoff-Olsen, C. Haas, et al., Autosomal SNP typing of forensic samples with the GenPlex™ HID System: results of a collaborative study, Forensic Sci. Int. Genet. 5 (2011) 369–375.
[19] J. Amigo, A. Salas, C. Phillips, ENGINES: exploring single nucleotide variation in entire human genomes, BMC Bioinformatics 12 (2011) 105.
[20] J.J. Sánchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P. Schneider, Á. Carracedo, N. Morling, A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713–1724.
[21] J.J. Sánchez, P. Endicott, Developing multiplexed SNP assays with special reference to degraded DNA templates, Nat. Protoc. 1 (2006) 1370–1378.
[22] C. Phillips, M.V. Lareu, A. Salas, Á. Carracedo, Nonbinary single-nucleotide polymorphism markers, Int. Congr. Ser. 1261 (2004) 27–29.
[23] A.A. Westen, A.S. Matai, J.F.J. Laros, H.C. Meiland, M. Jasper, W.J.F. de Leeuw, P. de Knijff, T. Sijen, Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples, Forensic Sci. Int. Genet. 3 (2009) 233–241.
[24] R. Pereira, C. Phillips, N. Pinto, C. Santos, S.E. Dos Santos, A. Amorim, Á. Carracedo, L. Gusmão, Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing, PLoS One 7 (2012) e29684.
[25] R.V. Rohlfs, S.M. Fullerton, B.S. Weir, Familial identification: population structure and relationship distinguishability, PLoS Genet. 8 (2012) e1002469.