Forensic Science International: Genetics 5 (2011) 155–169

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel

C. Phillips [a,b,*], L. Fernandez-Formoso [a], M. Garcia-Magariños [a], L. Porras [a], T. Tvedebrink [c], J. Amigo [b], M. Fondevila [a], A. Gomez-Tato [d], J. Alvarez-Dios [d], A. Freire-Aradas [a], A. Gomez-Carballa [a], A. Mosquera-Miguel [a], Á. Carracedo [a,b], M.V. Lareu [a]

[a] Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[b] Genomics Medicine Group, CIBERER, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[c] Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark
[d] Faculty of Mathematics, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain

ARTICLE INFO

Article history:
Received 10 December 2009
Received in revised form 2 February 2010
Accepted 6 February 2010

Keywords:
Short tandem repeat
STR
ESS
CEPH human genome diversity panel
Theta
Population substructure

ABSTRACT

The CEPH human genome diversity cell line panel (CEPH-HGDP) of 51 globally distributed populations was used to analyze patterns of variability in 20 core human identification STRs. The markers typed comprised the 15 STRs of Identifiler, one of the most widely used forensic STR multiplexes, plus five recently introduced European Standard Set (ESS) STRs: D1S1656, D2S441, D10S1248, D12S391 and D22S1045. From the genotypes obtained for the ESS STRs we identified rare, intermediate or off-ladder alleles that had not been previously reported for these loci. Examples of novel ESS STR alleles found were characterized by sequence analysis. This revealed extensive repeat structure variation in three ESS STRs, with D12S391 showing particularly high variability for tandem runs of AGAT and AGAC repeat units. The global geographic distribution of the CEPH panel samples gave an opportunity to study in detail the extent of substructure shown by the 20 STRs amongst populations and between their parent population groups. An assessment was made of the forensic informativeness of the new ESS STRs compared to the loci they will replace: CSF1PO, D5S818, D7S820, D13S317 and TPOX, with results showing a clear enhancement of discrimination power using multiplexes that genotype the new ESS loci. We also measured the ability of Identifiler and ESS STRs to infer the ancestry of the CEPH-HGDP samples and demonstrate that forensic STRs in large multiplexes have the potential to differentiate the major population groups but only with sufficient reliability when used with other ancestry-informative markers such as single nucleotide polymorphisms. Finally we checked for possible association by linkage between the two ESS multiplex STRs closely positioned on chromosome-12: vWA and D12S391 by examining paired genotypes from the complete CEPH data set.

1. Introduction

The widely used forensic markers of the AmpF/STR Identifiler® multiplex (Applied Biosystems: AB, Foster City, US) comprise amelogenin plus 15 autosomal STRs, that are also combined in the smaller scale AmpF/STR multiplexes of SGM Plus®, CoFiler®, Profiler Plus® and MiniFiler™. The forensic community has studied these STRs in numerous population samples but poor coverage is still evident in Oceania and to a lesser extent in the continental regions of South Asia, East Asia and the Americas [1].

The CEPH human genome diversity panel (CEPH-HGDP) contains just over a thousand individuals from African, European, North African/Middle Eastern, Central-South Asian, East Asian, Native American and Oceanian populations [2]. The identification of duplicates and first-degree relatives within the sample set reduces the CEPH panel to 971 individuals if second-degree relatives such as cousin pairs are retained [3], further reduced to 952 with the removal of second-degree relatives (termed H971 and H952 subsets respectively). We used the H971 subset to analyze the variation of the 15 STRs of Identifiler and supplemented this study with a further five loci, included in the European Standard Set or ESS STRs. The ESS loci have been recently adopted in Europe as markers able to improve on the discriminatory power and performance of five CODIS STRs in SGM Plus and Identifiler, specifically: CSF1PO, D5S818, D7S820, D13S317 and TPOX [4]. Two ESS STRs: D1S1656 and D12S391 have been chosen to improve the

* Corresponding author at: Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain. Tel.: +34 981 582 327; fax: +34 981 580 336.
E-mail address: c.phillips@mac.com (C. Phillips).

1872-4973/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2010.02.003

overall informativeness of an optimum 15 STR core set for European use and the other three: D2S441, D10S1248 and D22S1045 add enhanced genotyping success when analyzing highly degraded DNA by amplifying fragments well below current average amplicon sizes.

Until now the five new ESS markers lack extensive population studies and more importantly have not been fully characterized at the sequence level in non-European populations. Allele frequency analysis of the ESS loci in the broadly distributed CEPH populations provided the best opportunity to find and sequence rare or population specific alleles found in intermediate positions to the common repeats or outside current reference ladder size ranges in these five STRs. As the PCR primers used in the commercial kits currently being developed for the new STR combinations have not been published we made use of primer designs we originally developed when establishing two of the ESS STRs: D1S1656 and D12S391 [5,6]. This had the advantage that we could utilize previously optimized designs and PCR conditions plus comprehensive sequenced allelic ladders for the two STRs. Similarly the three short-amplicon ESS STRs were typed using the primer designs of the National Institute of Standards and Technology (NIST) where these STRs had been originally developed and characterized [7]. We combined the new ESS loci into a simple stand-alone 5-plex permitting straightforward typing of population samples to supplement existing surveys based on Identifiler or SGM Plus STRs (the full ESS multiplex comprises the 10 SGM Plus loci plus the above five). The same 5-plex assay can add further discrimination power to criminal casework samples previously typed with any of the established kits and has proved to be both robust and sensitive when analyzing challenging DNA [8].

The geographic breadth of the CEPH panel sampling gave the opportunity to explore various aspects of the new ESS loci of interest to forensic laboratories waiting to assess the usefulness of these STRs. As well as sequence analysis of rare alleles we were able to examine the following: (i) the predicted improvement in discrimination power of ESS STRs in each population group; (ii) the ancestry informativeness of all 20 STRs using a standard population reference panel with confirmed ancestral origins; (iii) the potential association by linkage between the two most closely positioned STRs in the full ESS set: vWA and D12S391, separated by 6.37 Mb on chromosome-12; and (iv) the extent of substructure amongst CEPH populations, some of which are known to exhibit very high levels of stratification.

In order to allow easy scrutiny of the extensive allele frequency data (based on more than 19,000 genotypes) generated by this study we reorganized an existing open-access SNP allele frequency browser [9] into a dedicated site named *pop.STR* [10] to accommodate the STR data. This allele frequency information is now freely available at: http://spsmart.cesga.es/popstr.php website. The help section of the *pop.STR* website provides a summary plot outlining the marker composition of each forensic multiplex from Applied Biosystems. Finally it is important to stress that intact individual STR profiles (i.e. with each genotype linked together per individual) were not retained in any format, to protect the privacy of the CEPH-HGDP donors.

2. Materials and methods

2.1. Population samples and STR genotyping

Although the CEPH-HGDP has a total of 1064 samples originating from 51 populations we used the H971 subset which excludes individuals previously identified (3) as duplicates, atypical samples (one mixed ancestry sample each from Africa and East Asia) or close relatives comprising parent-offspring and sib pairs. The H971 subset includes 19 cousin pairs removed from the other H952 subset. One widely recognized problem with the CEPH-HGDP is the disparity in population sample sizes, ranging from 6 African San of Namibia to 50 Middle East Palestinians of Central Israel. This means a proportion of CEPH populations cannot provide realistic allele frequencies given the broad range of alleles normally observed in the forensic STRs studied here. For this reason population designations were only used to identify the existence of specific STR alleles: i.e. those confined to one population alone or to analyze stratification in populations compared to their parent groups. For all other purposes populations were combined into continental-based groups which had been previously established [11] with the following composite populations, sample sizes and labels: 6 African (112 AFR), 8 European (158 EUR), 4 North African/Middle Eastern (170 ME), 9 Central-South Asian (204 S ASN), 17 East Asian (232 E ASN), 2 Oceanian (30 OCE) and 5 American (65 AME).

Identifiler STRs were genotyped with the commercial kit from AB and standard recommended amplification conditions except using 10 μl final PCR volume. The five new ESS loci were genotyped using previously optimized primers outlined in Table 1, with a single modification made to the originally reported reverse PCR primer of D22S1045 to address peak quality problems which lengthened the amplicons for this STR. The PCR reactants and cycling conditions were briefly: 1 μl DNA (0.5 ng/μl) 1 μl primer mix (ratios in Table 1), 3 μl H_2O, 5 μl Qiagen Multiplex PCR Master Mix (Cat. No. 206143), 95 °C × 15 min pre-denaturation, then 34 cycles of: 94 °C × 30 s, 58 °C × 90 s, 72 °C × 90 s, then a final elongation of 72 °C × 10 min.

Table 1
Genotyping and sequence primers used for the ESS 5-plex STR analyses. Non-specific mobility shift tails in lower case.

STR	Dye	Genotyping primers	Ratio in PCR primer mix[a]	Observed repeat numbers	Amplicon span (with tails)	Sequencing primers
D10S1248	6-FAM	TTAATGAATTGAACAAATGAGTGAG gCAACTCTGGTTGTATTGTCTTCAT	0.75 μl	7 19	78 (79) 126 (127)	CTCTGTATCCCACCCCTG AAAGCAAACCTGAGCATTAGCC
D1S1656	6-FAM	GTGTTGCTCAAGGGTCAACT ctctctctctctctccttGAGAAATAGAATCACTAGGGA	0.75 μl	8 19.3	117 (135) 164 (182)	CCATATAAGTTCAAGCCTGTGTT GAGAAATAGAATCACTAGGGA
D12S391	6-FAM	AACAGGATCAATGGATGCAT TGGCTTTTAGACCTGGACTG	0.75 μl	12 27.2	197 259	AGAGACTGTATTAGTAAGGCTTC TGGCTTTTAGACCTGGACTG
D2S441	VIC	CTGTGGCTCATCTATGAAAACTT gAAGTGGCTGTGGTGTTATGAT	0.72 μl	8 17	76 (77) 112 (113)	CTGAGCCCTAATGCACCCA gAAGTGGCTGTGGTGTTATGAT
D22S1045	NED	ATTTTCCCCGATGATAGTAGTCT CGGCACAGTGTGAGTGATCAC[b]	0.72 μl	9 19	104 134	AGCTGCTATGGGGGCTAGATT CGGCACAGTGTGAGTGATCAC

[a] Volume of 100 μM primer stocks.
[b] Sequence primer design used for genotyping PCR to reduce peak artifacts.