A genomic audit of newly-adopted autosomal STRs for forensic identification

Abstract

In preparation for the growing use of massively parallel sequencing (MPS) technology to genotype forensic STRs, a comprehensive genomic audit of 73 STRs was made in 2016 [Parson et al., Forensic Sci. Int. Genet. 22, 54-63]. The loci examined included miniSTRs that were not in widespread use, but had been incorporated into MPS kits or were under consideration for this purpose. The current study expands the genomic analysis of autosomal STRs that are not commonly used, to include the full set of developed miniSTRs and an additional 24 STRs, most of which have been recently included in several supplementary forensic multiplex kits for capillary electrophoresis. The genomic audit of these 47 newly-adopted STRs examined the linkage status of new loci on the same chromosome as established forensic STRs; analyzed world-wide population variation of the newly-adopted STRs using published data; assessed their forensic informativeness; and compiled the sequence characteristics, repeat structures and flanking regions of each STR. A further 44 autosomal STRs developed for forensic analyses but not incorporated into commercial kits, are also briefly described.

1. Introduction

A number of autosomal STR multiplex kits have recently been developed for capillary electrophoresis (CE) analysis consisting of loci complimentary to the commonly used 24 “core” forensic STRs (comprising sets: CODIS A and B plus PentaD, PentaE and D6S1043 [1]), with the supplementary STRs often termed “non-CODIS” loci. Similarly, massively parallel sequencing (MPS) technology applied to forensic STR genotyping has been able to supplement the commonly used loci with additional STRs that enhance the discriminatory power of the marker set as a whole, but also substitute for core STRs which are not reliably genotyped with MPS at this moment, such as SE33. However, many of the supplementary STRs being considered or already developed into kits, lack sufficiently detailed genomic descriptions and this has led to a number of ambiguities; notably the misidentification of the sequence details of D5S2500 [2]. Since the erroneous genomic descriptions of D5S2500/D5S2800 was first discovered, ambiguity has also been found in the published details of supplementary STRs D6S477 and D15S659, with each STR’s reported primer pairs annealing to the same genomic region [3]. As a last example of insufficient care with the description of newly-adopted STRs, there are at least two cases of the STR D2S441 being listed as D2S411 [4,5], which is a human microsatellite [6], but not the locus being analyzed. Although simple typographic errors can easily occur, they highlight the need to pay greater attention to the precise description of newly-adopted STRs that are not familiar to the majority of forensic DNA analysts. For this reason, the STRidER database has been established to ensure the forensic community has a centrally curated allele frequency database and quality control platform for autosomal STRs [7]. The study reported here has characterized the genomic details of 47 complimentary autosomal STRs that have been adopted for forensic use in the last five years. This initiative forms part of the quality control aspect of STRidER to ensure that the growing number of population studies of newly-adopted STRs are analyzing loci that have been accurately identified and characterized in terms of their genomic data.

Four kits of mainly complimentary STRs have been recently released, but solely for Chinese forensic use: AGCU NC 21+1; Microreader 23sp; Goldeneye DNA ID 22NC and STRtyper 10-G. The STRs adopted in each kit but which are not in common use, are summarized in Fig. 1, along with the other large-scale CE kits of supplementary STRs: Investigator HDplex, released by Qiagen in 2011 [8,9]. Fig. 1 shows there is minimal overlap of component markers between AGCU 21+1 and the other STR sets, but substantial overlap amongst the other sets, meaning use of the AGCU kit plus either Microreader or Goldeneye kits would provide as many as 34 STRs complimentary to the commonly used loci, respectively. Therefore, although these kits are not currently available to forensic practitioners outside of China, their component STR's discrimination power and locus characteristics are of considerable interest, since such a large number of additional STRs would greatly enhance the likelihoods obtained when testing complex pedigrees (e.g. a deficient pedigree where most members are unavailable for testing [10,11]). The interest in developing additional sets of forensic STRs in China is likely aiming to preempt a time in the future when the management of a potentially very large national DNA database must minimize the incidence of adventitious matches. It is noteworthy that the current Chinese DNA database minimum marker set has 23 STRs, which includes Penta D, Penta E and D6S1043, but not SE33; in addition to the redefined CODIS set formed by the other commonly used STRs [1]. The Applied Biosystems Huaxia Platinum STR multiplex combines all 23 STRs of the Chinese minimum marker set [12]. Furthermore, all forensically viable autosomal STRs with sufficient levels of polymorphism are worthy of detailed scrutiny since the limits of PCR multiplexing in MPS technology are unlikely to have yet been reached. For this reason, six additional STRs, not in the above CE kits but successfully analyzed with MPS technology [13], have been included in the genomic characterization of newly-adopted autosomal STRs this study outlines. The genomic audit of the 47 STRs identified to be newly-adopted for forensic use, examined the linkage status of the supplementary loci on the same chromosome as established

84 forensic STRs [14]; analyzed world-wide population variation of those STRs with
85 published data beyond Chinese populations; assessed their forensic
86 informativeness; and compiled the sequence characteristics, repeat structures
87 and flanking regions of each STR from the current human reference sequence.
88

2. Materials and methods

2.1. The STR sets

2.1.1. 23 NIST miniSTRs

The STRs D2S441, D10S1248 and D22S1045 from the 26 miniSTRs originally developed by the National Institute of Standards and Technology (NIST, Gaithersburg, USA) and described by Hill et al. [15], are now well-established CODIS STRs in widespread use. The other 23 were all successfully analyzed with MPS by Scheible et al. in 2014 [13] and 17 of these 23 loci have been incorporated into the AGCU 21-STR CE multiplex. A developmental validation that evaluated the CE genotyping performance of the AGCU kit (PCR conditions, forensic sensitivity, electrophoretic precision, stutter ratios, etc.) was made by Zhu et al., in 2015 [16]. Although no analyses of the genomic and sequence characteristics of the AGCU STRs has been made, the original development by NIST of these loci as part of the 26 MiniSTR set included a quite comprehensive study of their sequences and they remain the best described of the supplementary autosomal STRs available to use (see also [17]).

Several miniSTRs have recently been adopted in two commercial forensic MPS STR sets: the Illumina ForenSeq DNA Signature Prep Kit (D4S2408, D9S1122, D17S1301, D20S482); and the Thermo Fisher-Applied Biosystems Precision ID GlobalFiler NGS STR Panel (D1S1677, D2S1776, D3S4529, D5S2800, D6S474, D12ATA63, D14S1434, plus D4S2408 in common with the Illumina STR set). The six miniSTRs not included in any commercial CE or MPS multiplex kit to date are: D3S3053; D4S2364; D8S1115; D9S2157; D17S974; D20S1082, which are fully described in this study but not shown in Fig. 1. Scheible's MPS study reported very informative data describing sequence analyses of the common alleles of all 26 of the NIST miniSTRs [13]. Note that four further miniSTRs: D6S1027, D9S324, D10S1430 and D14S297 were successfully genotyped by

Hill et al. [15]. D6S1027 and D14S297 had very low Heterozygosities (<0.5; less informative than the best forensic SNP loci), so are not included in this study. D9S324 and D10S1430 had complex sequence variation close to their repeat regions and were not developed further. However, both are potentially informative loci and therefore included in additional autosomal STRs of potential interest but not in commercial kits, described in section 2.1.4.

2.1.2. Qiagen HDplex STRs

The population variability and forensic informativeness of the 12 STRs of the HDplex kit, of which nine are not in common use, have already been described [7,8], but these studies did not include detailed sequence data. The HDplex kit has D2S1360, D10S2325 and D21S2055 STRs unique to the set and shares five loci with Goldeneye plus a single NIST miniSTR D6S474 in common with AGCU. The D5S2500 STR in HDplex (as well as Goldeneye and Microreader sets) is a different STR to the re-named D5S2800 locus in AGCU and Thermo Fisher-Applied Biosystems Precision ID STR sets [2].

2.1.3. Goldeneye, Microreader and STRtyper STRs

The Goldeneye DNA ID 22NC set (Peoplespot Inc, China) has 16 newly-adopted STRs in a total of 22, 14 overlapping with Microreader and five with STRtyper, but has no unique STRs. The Microreader 23sp set (Suzhou Microread Genetics, Suzhou, China) has 18 newly-adopted STRs in a total of 22, with D9S925, D20S470 and D21S1270 unique to the set. STRtyper 10-G (Condon, ZhuHai, China) is the smallest supplementary STR kit with 7 newly-adopted STRs of 9 in total, and STRs D2S1772 and D18S1364 unique to the set. The study by Zhang et al., in 2013 [3], appears to have examined a prototype 17-STR set that was then adapted to create both Goldeneye and Microreader multiplexes. There are no studies using the Goldeneye STR set that describe the loci themselves, but Li et al., 2006 [18], report the validation of the Microreader STRs examining PCR

performance, sensitivity, CE signal patterns, etc. Lastly, the study by Huang et al. in 2012, sequenced alleles in several STRs in the STRtyper kit [19].

2.1.4. Autosomal STRs of potential interest but not incorporated into commercial kits

The loci of Promega's CS7 supplementary STR kit were considered insufficiently polymorphic to merit detailed analysis and have already been studied for linkage status and population variation [14,20]. Other STR sets previously published for forensic use and of potential interest have their sequence data compiled in brief detail here. These STR sets comprise: 13 closely linked non-CODIS STRs developed by Liu et al. [21]; 12 autosomal STRs developed by Phillips et al. for ancestry inference, but indicating informative levels of polymorphism for identity applications [22]; 8 STRs developed by Asamura et al. [23]; 9 STRs developed by Pinto et al. [24] and 15 STRs developed by Grubweiser et al. [25]. Including miniSTRs D9S324 and D10S1430, discounting core STRs and certain supplementary STRs in some of the above sets, plus overlap amongst the five multiplexes, there are 44 novel supplementary STRs described in these studies but not incorporated into commercial forensic kits so far.

2.2. Linkage analysis

A total of 47 syntenic (same chromosome) STR pairs were identified comparing 47 newly-adopted STRs with 24 core STRs, although this does not correspond to all newly-adopted STRs; for example, there are no newly-adopted STRs on Chromosome 16 to compare to D16S539. Genetic distances between syntenic STRs were estimated using the HapMap b36 human recombination map, as previously described [20]. Because HapMap has now been retired as a genomic database, all SNP marker positions and centiMorgan (cM) genetic distance values of the high-density b36 SNP map originally compiled by HapMap, are available from the author upon request.

2.3. Worldwide population data and forensic informativeness estimates

Although the release of four supplementary STR kits in China has led to a plethora of published studies examining ethnic Chinese populations, many of the component STRs have been subject to more broadly-based population studies of their variability. For these STRs, we made allele frequency estimates for the major population groups comprising: African, European, East Asian, Oceanian, Native American, South Asian and Middle Eastern populations, using data from published studies of the HGDP-CEPH human genome diversity project sample set; principally those of Rosenberg Lab [26] and our own STR genotyping of this population panel [7,19,20,27] where allele frequencies have been compiled in the pop.STR database (http://spsmart.cesga.es/popstr.php [28]).

Forensic informativeness metrics comprising: random match probability (RMP), power of discrimination (Dp=1−RMP), power of exclusion from paternity (PE), and typical paternity index (TPI) were calculated using Promega PowerStats Excel calculator (no longer accessible; file available upon request). To compare the forensic informativeness of each supplementary STR with the core loci, individual STR Heterozygosity values based on African, European, and East Asian allele frequencies were estimated and the population-wide average value in each case compared with 24 established forensic STRs using previous population data from the same HGDP-CEPH samples [27].

2.4. Compilation of genomic data

The sequence characteristics, repeat structures and flanking region variants were compiled for each newly-adopted STR using the 1000 Genomes database browser (http://browser.1000genomes.org/Homo_sapiens/Info/Index) for both sequence data and the annotation of repeat region or flanking region variation. The 1000 Genomes sequence database holds the Genome Reference

213 Consortium human Reference Sequence 37 (GRCh37). Matching GRCh38
214 coordinates were obtained using the Santa Cruz Genome Browser *In Silico PCR*
215 tool (http://genome.ucsc.edu/cgi-bin/hgPcr).
216
217 STR repeat region structures were defined as much as possible with cross-
218 reference to four sequence analysis studies of the bulk of the STRs (39 of 47),
219 made by Pemberton et al., 2009 [29], Huang et al., 2012 [19]; Hill et al., 2008 [15]
220 and Barral et al., 2000 [30]. For the other eight STRs without sequence data from
221 previous studies, we applied simple rules for defining the repeat region motifs
222 based on: counting from the first continuously repeated tri-, tetra- or penta-
223 nucleotide repeat, not counting sequence runs where the repeat motif was not
224 present and counting single repeat motifs within a run of repeats that differed
225 from the main motif by one nucleotide change. For example, the STR D14S608
226 has no published sequence analysis and the repeat region in the reference
227 sequence was annotated as: $[GATA]_3$ ATAGAGATAGAT $[GATA]_1$ $[GACA]_1$
228 $[GATA]_9$ = 14 repeats. All sequence annotation descriptions followed the same
229 pattern as those made for the core autosomal, X- and Y-STRs outlined in the
230 Excel supplementary file that accompanied the STR sequence nomenclature
231 guidelines published in 2016 (Supplementary File S1 of [31]). For each STR, this
232 supplementary data file details the repeat region sequence and a minimum 100
233 nucleotides each side of the repeat region from the human reference sequence
234 (the 5' to 3' forward strand), with common-variation flanking SNP and Indel loci
235 identified and GRCh37/GRCh38 coordinates provided. In cases of low
236 complexity sequence around repeat regions, e.g. poly-base tracts, additional
237 flanking region sequence was detailed. Note that the original sequence
238 annotation file of [31] (Supplementary File S1) included several miniSTRs and
239 the 'true' D5S2500 STR, as these loci had been incorporated into commercial
240 MPS kits, but their genomic data are included here.
241

3. Results

3.1. Summary STR details

Table 1A lists in the left-hand columns the largest autosomal STR multiplexes with core loci currently available from CE kit manufacturers: Applied Biosystems, Promega and Qiagen. The middle columns list the core STRs present in each of the supplementary kits from Qiagen and four Chinese manufacturers that genotype mainly newly-adopted STRs. The right-hand columns show the core STRs present in the MPS kits of manufacturers Thermo Fisher-Applied Biosystems, Illumina and Promega (the Promega PowerSeq MPS kit does not include any newly-adopted STRs). Note that Illumina and Promega MPS kits and the Applied Biosystems Huaxia Platinum CE kit genotype all 23 STRs of the Chinese minimum database set.

Table 1B gives summary data (rs-numbers when assigned by dbSNP, 9947A control DNA genotypes, repeat region start nucleotide coordinates of GRCh37), for the 47 supplementary STRs of this study: 41 incorporated into five CE kits plus two MPS kits, and 6 NIST miniSTRs successfully analyzed with MPS in the study of Scheible et al. as part of a 48 marker assay [13].

The ‘optimum’ combination of core and supplementary STRs for CE analysis - representing the maximum number of STRs and minimum number of kit-based amplification reactions - combines the Huaxia Platinum kit with Qiagen HDplex, Microreader and AGCU supplementary STR kits. This four-kit combination allows the genotyping of all 24 core loci and 39 newly-adopted STRs from four forensic assays. Such a kit combination would genotype 62 independently segregating STRs in total, as they combine D5S2500 and D5S2800 STRs - separated on Chromosome 5 by only 1643 nucleotides, requiring the discounting of data from one of them, or treatment of both loci as a single marker. For likelihood calculations in kinship testing, the best option is to introduce the Rc value

between them (Table 2) and make adjustments with dedicated the ILIR calculator [33]. The alternative of estimating haplotype frequencies composed of D5S2500-D5S2800 allele pairs would be unwieldy and phase would not be easily discerned, which means the best option is to choose D5S2500 for CE-based analyses (e.g. HDplex) and D5S2800 when applying the Thermo Fisher-Applied Biosystems Precision ID GlobalFiler NGS STR Panel.

3.2. Linkage analysis

Table 2 lists the GRCh37 genomic positions, genetic map distances and recombination fractions (Rc values) of the 47 syntenic STR pairs identified in 71 core and supplementary loci. Six STR pairs have recombination rates estimated to be less than 10% (Rc values below 0.1) and these are highlighted in gray. Ignoring D5S2500/D5S2800 and the STR loci of SE33 and D6S1043, which were already identified as closely linked [14]; STRs D2S1772-D2S441; D18S51-D18S1364; D3S4529-D3S3045 and D21S2055-Penta D are the only marker pairs closer than the previously analyzed CODIS STR pair of VWA and D12S391 [14,32]. This is a remarkable result; given that it represents the analysis of an extra 47 supplementary STRs – almost double the number of core loci in established use. It can also be argued that only D2S1772-D2S441 (AGCU kit-core) and D18S51-D18S1364 (core-STRtyper kit) would require care when analyzing related individuals in kinship tests using a complete set of kit-based STRs. Furthermore, ILIR [33] has already been proposed for the incorporation of recombination rate estimates between such closely positioned forensic loci, as a way to reduce the effect of linkage bias in likelihood calculations. Therefore, the optimum combination of four CE kits as described above, aimed at obtaining the most informative genetic data for challenging relationship test scenarios, would only require adjustment of likelihood calculations for a single marker pair amongst the 62 STRs, since the STRtyper kit is not included in this combination and only brings D18S1364 plus the linked D2S1772 as additional loci.

It should be mentioned that a study by Wu et al. in 2014 [34] of linkage data across 12 chromosomes, estimated for the AGCU component STRs using the same HapMap recombination map, obtained identical recombination rate estimates as those reported here.

3.3. Population variability and forensic informativeness metrics of newly-adopted STRs

In all, 35 newly-adopted STRs have full HGDP-CEPH population data published from which allele frequency estimates and forensic informativeness metrics can be obtained in seven worldwide population groups. Complete sets of HGDP-CEPH genotypes based on repeat-length measurements are listed in full in Supplementary File S1; and allele frequency estimates and key forensic informativeness metrics from this genotype data are listed for each population group in Supplementary File S2, along with summary allele frequency distribution plots comparing African, European and East Asian variability. A small proportion of loci show marked population differentiation of allele frequency distributions, but in most newly-adopted STRs with detailed population data, allele frequencies are generally similar. Strong repeat-length allele frequency differentiation was observed between Africans and the other two groups in D18S1364, D12ATA63 and D6S1017; with a shift in the distribution of variation in African populations towards smaller length alleles. The same trend of a shift to smaller allele size ranges was observed for East Asian populations in D9S2157 and in European populations in D21S2025, with the 19.1 allele present at 25% in this population group and providing a high frequency population-indicative allele.

To compare levels of polymorphism amongst so many novel STRs in different populations is not straightforward, so average Heterozygosity values were calculated as a single summary variation statistic per locus with which to compare newly-adopted and core STRs. Heterozygosities were averaged from African, European and East Asian values and then compared to these values

estimated from previously published data for the core STRs analyzing the same HGDP-CEPH samples. Heterozygosity is closely associated with power of discrimination / random match probability and power of exclusion / typical paternity index metrics, so can act as an efficient proxy for the expected forensic informativeness of each new STR relative to those used routinely. Furthermore, for the twelve newly-adopted STRs that have not been characterized using the HGDP-CEPH panel, published allele frequency estimates alone allow an estimation of Heterozygosity but not the other values. Allele frequency estimates from studies of the 26 NIST miniSTRs (average Heterozygosity values from European and African American samples [35,36]) were used, or when no other data was available, East Asian population frequencies from the study of 515 Han Chinese by Liu et al. [37] provided an approximate Heterozygosity value for several Microreader STRs not yet studied in other populations.

Fig. 2 places the newly-adopted STRs in order of length allele Heterozygosity value by comparing them to the core STR-average of 80% Heterozygosity shown as the midline (averaging African, European, East Asian data). Of the 29 STRs with above-average variability in length allele polymorphisms at the top of the plot, 18 are newly-adopted (black or white bars), indicating that these loci provide a highly informative option for expanding genetic data when this is needed for challenging relationship tests. Although SNPs have also been proposed as supplementary markers in such situations [10], they provide much less power per locus compared to the supplementary STRs characterized in this study. Amongst the most powerful supplementary STRs, two genotyping kits predominate: D7S1517; D10S2325; D8S1132; D21S2055 in HDplex (D10 and D21 unique to this kit), and D7S3048; D20S470; D21S1270; D22GATA19; D8S1132; D17S1290; D15S659 in Microreader (D20 and D21 unique to this kit). Therefore, the informative HDplex CE kit is widely available to supplement core STRs when necessary; and the highly informative Microreader kit would be a very useful option if it had the same commercial availability to forensic users outside of China. With the exception of D9S2157, the six NIST miniSTRs not incorporated

into CE kits of newly-adopted loci populate the lower portion of the Fig. 2 plot and have substantially lower variability than the bulk of other forensic STRs. Lastly, D5S2500 has some 2% higher average Heterozygosity than D5S2800, so allelic data from this STR would be the best choice to use in any combined CE kit analyses, but early indications show the inclusion of D5S2800 in the Thermo Fisher-Applied Biosystems Precision ID MPS kit provides a significant increase in this STR's polymorphism levels from sequence variation within the repeat region.

Finally, the emphasis on length allele measurements in this section is intended to highlight the fact that little is known yet about which of these supplementary STRs will provide large jumps in informativeness from discerning their sequence variants. However, the sequence data described in the following section will indicate some candidates for consideration for MPS analysis.

3.4. Sequence details of newly-adopted STRs

Detailed sequence data for the 47 STRs of this study are provided in Supplementary File S3. Several STRs have flanking region SNPs that would be informative in sequence analysis with MPS. More importantly for both CE and MPS-based analyses, a number of common insertion-deletion (Indel) polymorphisms were identified within, or close to repeat regions. These are: the 6-NT (nucleotide) deletion rs143705585 in D6S477 (previously described by Barral [30]), creating intermediate alleles in this STR; the common 12-NT deletion rs143108846 in D17S974; the two common 1-NT deletions of rs139318958 and rs11477214 in D20S1082; and the common 3-NT deletion rs67809736 in D21S1270. In addition, the STR D21S2055 shows a complex arrangement of 1-NT deletions embedded as a set of loci with identical frequencies inside the two repeat runs of this STR: rs200634042; rs559131724; rs201681362 (common deletions in Europeans) in the 5' repeat run, and rs150251552; rs529394107; rs549450286; rs56317735 (common deletions in all populations) in the 3' repeat run, with a common-frequency SNP pair within the

uncounted 30-NT sequence between them. Therefore, D21S2055 is a highly informative repeat-based polymorphism STR using CE (83% average Heterozygosity), that also shows a complex repeat region structure of interspersed [CTAT][CTAA][TATC] units and common sequence variants positioned within the repeat region. As such, it would merit consideration for MPS analysis, despite a longer than average repeat region (the reference sequence comprises 142 nucleotides). All the other STRs show simpler repeat region sequence patterns than D21S2055; several have multiple motifs per repeat region – likely to be informative for MPS analysis, or have single repeat motifs without significant structural variation apart from 3-NT non variable segments commonly observed between repeat runs.

The complexity of certain repeat region sequences and the number of STRs annotated in Supplementary File S3 makes it difficult to provide exhaustive descriptions here, compared to scrutiny of each individual STR sequence string in Excel. The STR repeat motif structures are summarized in Table 3 and compared to previously published sequence studies of 39/47 STRs. Of the eight STRs without sequence details for comparison, three are simple single motifs and therefore likely to be correct descriptions of sequences underlying the repeat-allele counts genotyped by CE. In the other five STRs lacking comparative CE and sequence data, D14S608 and D3S1744 may have sequence segments counted amongst the repeats. D14S608 has a 12-NT segment that could add 3 repeats to the total count and D3S1744 has 3-NT and 2-NT segments, where it is also not possible to make the comparison with CE genotyping data to match allele numbers. These five STRs require further sequence studies to confirm repeat-allele designations correctly match their sequence annotations.

In the 39 newly-adopted STRs with published sequence data, 28 match the repeat region annotations originally made in previous studies with those outlined in Supplementary File S3; compiled independently for this study from the

reference sequence (marked in bold in Table 3), although seven STR sequences were originally described for the reverse strand. Another two are a close match but with “frame-shift differences” comprising the 1-NT difference in D11S2368 and 2-NT difference in D18S364. Differences arise from reverse strand repeat region annotation (D11) or the application of different rules for the identification of the first nucleotide of the first repeat motif. Note that D6S477 counts both [TA] motifs as a single repeat in the total count of all other tetra-nucleotide units, as they are both present in the majority of alleles [30]. Note that D6S477 counts both [TA] motifs as a single repeat in the total count of all other tetra-nucleotide units, as they are both present in the majority of alleles [30].

3.5. Additional autosomal STRs of interest yet to be included in commercial forensic kits

To complete the survey of supplementary STRs developed in parallel to commercial kit STRs in the last five years, summary details of an additional 44 autosomal STRs are listed in Table 4 and Supplementary File S4. Locus details provided here are brief, but any of these novel STRs developed in five recently published studies [21-25], but not incorporated in any commercial kit to date, are of interest and would merit further studies. Detailed population data exists in the HGDP-CEPH studies of Rosenberg [26] for 18 of 44 STRs, as indicated in Supplementary File S4.

Discussion

Between the introduction of STR analysis into forensic DNA profiling in the mid-nineties and the current period of development in the field, the process of validating new candidate STRs has changed. In fact, it can even be said that the validation process has become inverted. Originally, new STRs were checked for a number of important characteristics: specific amplification (i.e. primer designs annealing to one position in the genome, free from neighboring repetitive DNA or

unstable elements such as inversions); genomic positions carefully identified; and from a set of commonly encountered alleles - corresponding to well-defined repeat allele numbers observed in CE - sequence analyses to understand the repeat structures underlying the STR. This process has ensured all of the 24 core autosomal STRs have had the benefit of careful genomic characterization before being adopted for forensic use and included in commercial kits. To a large extent, in-depth population studies for these STRs only followed after the kits were manufactured and reference allelic ladders, based on sequenced alleles, could be run alongside population samples. In the past five years, the commercial kits of supplementary STRs for forensic CE analysis that are the subject of this study have been launched and then rapidly applied to a large number of population surveys. However, there is little evidence that any of the kit's component STRs have had the necessary detailed scrutiny of their genomic characteristics to properly establish amplification specificity, precise positions (sequence coordinates rather than inexact cytogenetic locations such as '17q12'), sequence characteristics and repeat structures. The AGCU supplementary STR kit has mainly combined NIST miniSTRs that have a much more detailed level of genomic validation and it is unsurprising that this set of STRs has also fed additional autosomal loci into two forensic MPS kits with expanded multiplex capacity. Nevertheless, the ambiguity in the description of two separate STRs named D5S2500 - identified to be one locus in the AGCU CE kit and Thermo Fisher-Applied Biosystem Precision ID MPS kit; and another, different locus in the HDplex, Goldeneye and Microreader CE kits [2] - indicates all novel STRs yet to reach mainstream use must undergo comprehensive genomic characterization. This study seeks to initiate this process, while avoiding an overly-prescriptive approach to the annotation of the repeat region sequences of the 47 newly-adopted STRs and 42 novel STRs also identified to be informative forensic loci. This process needs the systematic comparison of CE and MPS data to ensure the allele annotations match each other. The sequence data of Supplementary File S3 in this study follows the same pattern as the supplementary file made for annotation of the core autosomal and X/Y STRs

[31], but it is also meant to act as a sequence annotation template onto which new sequence and variant details can be added when they become available.

It can be argued that adding between 39 and 47 supplementary STRs to the established autosomal marker sets is aiming for much higher levels of discrimination power than is strictly necessary for most forensic analyses. This study does not seek to comment on the merits of such substantial increases in STR numbers, but the fact that 39 extra STRs from three kits can almost double the marker coverage of forensic DNA tests suggests it will be worth encouraging more widespread commercial availability of the new Chinese-market CE kits for missing person identification tests alone [38]. There is the problem that adding so many extra STRs to resolve deficient paternity tests with ambiguous outcomes (e.g. a brother of the true father tested vs. a single-step mutation producing one second order exclusion), may lead to the detection of multiple mutations and therefore further ambiguity. Therefore, choice of SNPs as supplementary kinship markers rather than more STRs, can provide a better way to resolve such ambiguities [11]. Nevertheless, one very practical outcome of the detailed characterization of so many new autosomal STRs, is the consequent assessment of these marker's sequence variability and potential to be the most informative markers for MPS-based STR tests. This potentially leads to better mixed DNA de-convolution and more secure familial search matches, should the latter investigative approach become more mainstream. So there is the potential for DNA analysis regimes applied to both kinship testing and criminal casework to benefit from the widespread adoption of new STRs.

The STRidER STR database will ensure proper quality control measures are in place for all forensic loci, whether well-established or recently introduced [6], and the checking process will extend to genomic details compiled for STRs, as well as population data generated from CE size-based and MPS sequence-based genotyping. Although the first MPS STR kits have settled on marker sets that add four or eight NIST miniSTRs to the core loci, some of the other newly-adopted

STRs are certain to be worth further consideration if forensic DNA analysis continues to add new CE dye labels or expand the multiplex scales of MPS technology. When this happens, a promising set of candidate STRs have already been identified from the newly-adopted autosomal STRs described here.

Acknowledgements

The author would like to thank Mike Coble of NIST, Gaithersburg, MD, for helpful guidance on the characterization of miniSTRs, and Noah Rosenberg, Stanford University, CA, for help in identifying several human microsatellites and sharing his expertise in the genomic characteristics and population variability of many of the STRs of this study.

References

[1] D.R. Hares, Selection and implementation of expanded CODIS core loci in the United States, Forensic Sci. Int. Genet. 17 (2014) 33–34.

[2] C. Phillips, W. Parson, J. Amigo, J.L. King, M.D. Coble, C.R. Steffen, P.M. Vallone, K.B. Gettings, J.M. Butler, B. Budowle, D5S2500 is an ambiguously characterized STR: Identification and description of forensic microsatellites in the genomics age, Forensic Sci. Int. Genet. 23 (2016) 19–24.

[3] M. Whittle, More on the genomic identification of forensic STRs, Forensic Sci. Int. Genet. 25 (2016) e1–e2.

[4] S. Zhang, H. Tian, J. Wu, S. Zhao, C. Li, A new multiplex assay of 17 autosomal STRs and amelogenin for forensic application, PLoS One 8 (2013) e57471.

[5] C. Wang, X. Ma, H. Ma, S, Li, X. Hu, M. Shang, M. Wang, Y. Zhou, Y. Te, et al., Allele frequency distribution of 21 forensic autosomal STR loci of Goldeneye™ DNA ID 22NC Kit in Chinese Tibetan group, Forensic Sci. Int. Genet. 22 (2016) e21–e24.

[6] S. Takahashi, S.V. Faraone, J. Lasky-Su, M.T. Tsuang, Genome-wide scan of homogeneous subtypes of NIMH genetics initiative schizophrenia families, Psychiatry Res. 133 (2005) 111–122.

[7] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmão, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER), Forensic Sci. Int. Genet. 24 (2016) 97–102.

[8] C. Phillips, L. Fernandez-Formoso, M. Gelabert-Besada, M. García-Magariños, J. Amigo, Á. Carracedo, M.V. Lareu, Global population variability in Qiagen Investigator HDplex STRs, Forensic Sci. Int. Genet. 8 (2014) 36–43.

[9] A.A. Westen, H. Haned, L.J.W. Grol, J. Harteveld, K.J. van der Gaag, P. de Knijff, T. Sijen, Combining results of forensic STR kits: HDplex validation including allelic association and linkage testing with NGM and Identifiler loci, Int. J. Legal Med. 126 (2012) 781–789.

564 [10] M.V. Lareu, M. García-Magariños, C. Phillips, I. Quintela, Á. Carracedo, A.
565 Salas, Analysis of a claimed distant relationship in a deficient pedigree using high
566 density SNP data, Forensic Sci. Int. Genet. 6 (2012) 350–353.
567 [11] C. Phillips, M. García-Magariños, A. Salas, Á. Carracedo, M.V. Lareu, SNPs
568 as supplements in simple kinship analysis or as core markers in distant pairwise
569 relationship tests: when do SNPs add value or replace well-established and
570 powerful STR tests? Transfus. Med. Hemother. 39 (2012) 202–210.
571 [12] Z. Wang, D. Zhou, Z. Jia, L. Li, W. Wu, C. Li, Y. Hou, Developmental
572 Validation of the Huaxia Platinum System and application in 3 main ethnic groups
573 of China, Sci. Rep. 6 (2016) 31075.
574 [13] M. Scheible, O. Loreille, R. Just, J. Irwin, Short tandem repeat typing on the
575 454 platform: strategies and considerations for targeted sequencing of common
576 forensic markers, Forensic Sci. Int. Genet. 12 (2014) 107–119.
577 [14] C. Phillips, D. Ballard, P. Gill, D. Syndercombe Court, Á. Carracedo, M.V.
578 Lareu, The recombination landscape around forensic STRs: Accurate
579 measurement of genetic distances between syntenic STR pairs using HapMap
580 high density SNP data, Forensic Sci. Int. Genet. 6 (2012) 354–365.
581 [15] C.R Hill, M.C. Kline, M.D. Coble, J.M. Butler, Characterization of 26 MiniSTR
582 loci for improved analysis of degraded DNA samples, J. Forensic Sci. 53 (2008)
583 73-80.
584 [16] B. Zhu, Y. Zhang, C. Shen, W. Du, W. Liu, H. Meng, H. Wang, G. Yang, R.
585 Jin, C. Yang, J. Yan, X. Bie, Developmental validation of the AGCU 21+1 STR
586 kit: A novel multiplex assay for forensic application, Electrophoresis 36 (2015)
587 271–276.
588 [17] C. Phillips, M. Gelabert-Besada, L. Fernandez-Formoso, M. García-
589 Magariños, C. Santos, M. Fondevila, D. Ballard, D. Syndercombe Court, Á.
590 Carracedo, M.V. Lareu, “New turns from old STaRs”: Enhancing the capabilities
591 of forensic short tandem repeat analysis, Electrophoresis 35 (2014) 3173–3187.
592 [18] J. Li, H. Luo, F. Song, L. Zhang, C. Deng, Z. Yu, T. Gao, M. Liao, Y. Hou,
593 Validation of the Microreader™ 23sp ID system: A new STR 23-plex system for
594 forensic application, Forensic Sci. Int. Genet. 27 (2016) 67–73.

595 [19] Y. Huang, Yi. Qi, J. Xi, Y. Lin, L. Yang, Z. Zeng, X. Liu, L. Guo, L. Gao,
596 Analysis of genetic polymorphism of nine short tandem repeat loci in Chinese
597 Han population of Henan province, African J. Biotech. 11 (2012) 5988-5994.
598 [20] C. Phillips, S. Kind, L. Fernandez-Formoso, M. Gelabert-Besada, Á.
599 Carracedo, M.V. Lareu, Global population variability in Promega PowerPlex CS7,
600 D6S1043, and Penta B STRs, Int. J. Legal Med. 127 (2013) 901–906.
601 [21] Q.L. Liu, K.K. Huang, Y.D. Wu, H. Zhao, C.T. Li, D.J. Lu, Genetic
602 polymorphism of 13 non-CODIS STR loci in three national populations from
603 China, Electrophoresis 35 (2014) 3395–3401.
604 [22] C. Phillips, L. Fernandez-Formoso, M. Gelabert-Besada, M. García-
605 Magariños, C. Santos, M. Fondevila, Á. Carracedo, M.V. Lareu, Development of
606 a novel forensic STR multiplex for ancestry analysis and extended identity
607 testing, Electrophoresis 34 (2013) 1151–1162.
608 [23] H. Asamura, S. Fujimori, M. Ota, H. Fukushima, MiniSTR multiplex systems
609 based on non-CODIS loci for analysis of degraded DNA samples, Forensic Sci.
610 Int. 173 (2007) 7-15.
611 [24] L.M. Pinto, C.L. de Oliveira, L.L. Dos Santos, E. Tarazona-Santos, Molecular
612 characterization and population genetics of non-CODIS microsatellites used for
613 forensic applications in Brazilian populations, Forensic Sci. Int. Genet. 9 (2014)
614 e16–17.
615 [25] P. Grubwieser, B. Zimmermann, H. Niederstätter, M. Pavlic, W. Parson,
616 Validation study and population data of 15 “new” STR loci: A highly discriminating
617 set for paternity and kinship analysis, Int. Congress Series 1288 (2006) 447–449.
618 [26] Rosenberg Lab and Marshfield HGDP-CEPH STR genotype data from:
619 http://www.stanford.edu/group/rosenberglab/data/rosenbergEtAl2005/combinedm
620 icrosats-1048.stru and:
621 http://research.marshfieldclinic.org/genetics/genotypingData_Statistics/humanDiv
622 ersityPanel.asp Accessed March 2017.
623 [27] C. Phillips, L. Fernandez-Formoso, M. García-Magariños, L. Porras, T.
624 Tvedebrink, J. Amigo, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, A. Freire-
625 Aradas, A. Gomez- Carballa, A. Mosquera-Miguel, Á. Carracedo, M.V. Lareu,

626 Analysis of global variability in 15 established and 5 new European Standard Set
627 (ESS) STRs using the CEPH human genome diversity panel, Forensic Sci. Int.
628 Genet. 5 (2011) 155–169.
629 [28] J. Amigo, C. Phillips, T. Salas, L.F. Formoso, Á. Carracedo, M. Lareu,
630 pop.STR—an online population frequency browser for established and new
631 forensic STRs, Forensic Sci. Int. Genet. Suppl. Ser. 2 (2009) 361–362.
632 [29] T.J. Pemberton, C.I. Sandefur, M. Jakobsson, N.A. Rosenberg, Sequence
633 determinants of human microsatellite variability, BMC Genomics 10 (2009) 612.
634 [30] S. Barral, M.V. Lareu, A. Salas, Á. Carracedo, Sequence variation of two
635 hypervariable short tandem repeats at the D22S683 and D6S477 loci, Int. J.
636 Legal Med. 113 (2000) 146–149.
637 [31] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L.
638 Gusmão, D. R. Hares, J.A. Irwin, J.L. King, P. de Knijff, N. Morling, M. Prinz,
639 P.M. Schneider, C. Van Neste, S. Willuweit, C. Phillips, Massively parallel
640 sequencing of forensic STRs: considerations of the DNA commission of the
641 International Society for Forensic Genetics (ISFG) on minimal nomenclature
642 requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63.
643 [32] P. Gill, C. Phillips, C. McGovern, J.A. Bright, J. Buckleton, An evaluation of
644 potential allelic association between the STRs vWA and D12S391: implications in
645 criminal casework and applications to short pedigrees, Forensic Sci. Int. Genet. 6
646 (2012) 477–486.
647 [33] A. Tillmar, C. Phillips, Evaluation of the impact of genetic linkage in forensic
648 identity and relationship testing for expanded DNA marker sets, Forensic Sci. Int.
649 Genet. 26 (2017) 58–65.
650 [34] W. Wu, H. Hao, Q. Liu, X. Han, Y. Wu, J. Cheng, D. Lu, Analysis of linkage
651 and linkage disequilibrium for syntenic STRs on 12 chromosomes, Int. J. Legal
652 Med. 128 (2014) 735–739.
653 [35] C.R. Hill, M.D. Coble, and J.M. Butler, Development of 27 new MiniSTR loci
654 for improved analysis of degraded DNA samples, Poster B105 at American
655 Academy of Forensic Sciences meeting, Seattle, 2006.
656 [36] J.M. Butler, C.R. Hill, Biology and genetics of new autosomal STR loci useful

for forensic DNA analysis, Forensic Sci. Rev. 24 (2012) 15-26.
[37] Q.L. Liu, Z.X. Chen, C.G. Chen, D.J. Lu, Genetic polymorphism of 22 autosomal STR markers in a Han population of Southern China, Forensic Sci. Int. Genet. 24 (2016) e14–e16.
[38] https://www.reference.com/government-politics/many-people-missing-year-8e9fc97c68b730f4 Accessed March 2017. “According to Todd Matthews of the National Missing and Unidentified Persons System (NamUs), most cases of missing persons in the US are quickly resolved. Of the 661,000 cases in 2012, only 2,079 were still unresolved at the end of the year, and the number of cases generally decreases each year. Sometimes unidentifiable human remains are also found. According to NamUs, as many as 40,000 unidentified human remains have been found in the US”.

Figure 1. Multiplex combinations of 41 newly-adopted STRs in five forensic kits for CE analysis. A further six supplementary STRs are not shown as they have not been incorporated into a commercial forensic kit. The newly-adopted STRs in two MPS kits are indicated with gray blocks. Overlap between STRtyper and HDplex CE kits is indicated with ° and * prefixes to simplify the graphic, so the smaller circles lower left show STRs unique to these kits.

Figure 2. Comparison of average Heterozygosity values (averaged for HGDP-CEPH African, European and East Asian allele frequency data in each STR) in 47 newly-adopted STRs (black and white bars, bold and normal text) and 24 core STRs (gray bars, light text). STRs are arranged in descending order of Heterozygosity and placed to the right (above average informativeness) or left (below average) of a midline of 80% Heterozygosity, representing the average value for the core STRs. White bars denote STR allele frequency estimates based on limited population data of European Americans/African Americans or East Asian data alone (STRs with asterisks).

Table 1A Forensic kits for CE (left and center columns) and MPS (right three columns) combining core autosomal STRs.

Table 1B Summary locus details and marker combinations of supplementary STRs newly-adopted in kits for CE (center columns) and MPS (right three columns). Light gray squares indicate NIST Mini STRs, dark gray squares with bold outlines indicate STRs unique to the kit. The white dots denote the erroneous description of D2S441 as D2S411 in published studies of two kits of newly-adopted STRs

Table 2 Genomic positions, genetic map distances and recombination fractions (Rc) of 47 syntenic (same chromosome) STR pairs. Core STRs in bold, STR pairs with less than a 10% recombination rate between them (Rc<0.1) highlighted in gray.

Table 3 Repeat structure annotation of 47 recently-adopted STRs using the human Reference Sequence (Ref. Seq.) and comparison to published sequencing studies. Bold repeat structure data denotes full agreement between the annotations of this study and previous publications.

Table 4 Genomic details of 44 novel autosomal STRs, developed in published multiplexes for CE genotyping but not in commercially available kits.

Supplementary File S1. HGDP-CEPH genotypes for 35 newly-adopted STRs for which population data has been published. African, European and East Asian allele frequency estimates at the base of each STR's columns are data-checks.

Supplementary File S2. Allele frequency estimates and forensic informativeness metrics of 35 newly-adopted STRs based on the HGDP-CEPH genotypes of Supplementary File S1.

Supplementary File S3. Reference sequence strings (forward strand) of 47 newly-adopted forensic autosomal STRs. Loci are listed in genomic position order C1 to C22 with +/- 100 nucleotides of flanking sequence, or more if low complexity sequence occurs in the flanking regions.

Supplementary File S4. Locus details and simplified sequence annotations of 42 novel STRs developed in five published studies, but not incorporated in any commercial STR kit to date. 'Rosenberg?" indicates the presence of STR population data in the HGDP-CEPH studies published in [26]. 9947A indicates the 9947A control DNA genotype of each STR when known.

Figure 1

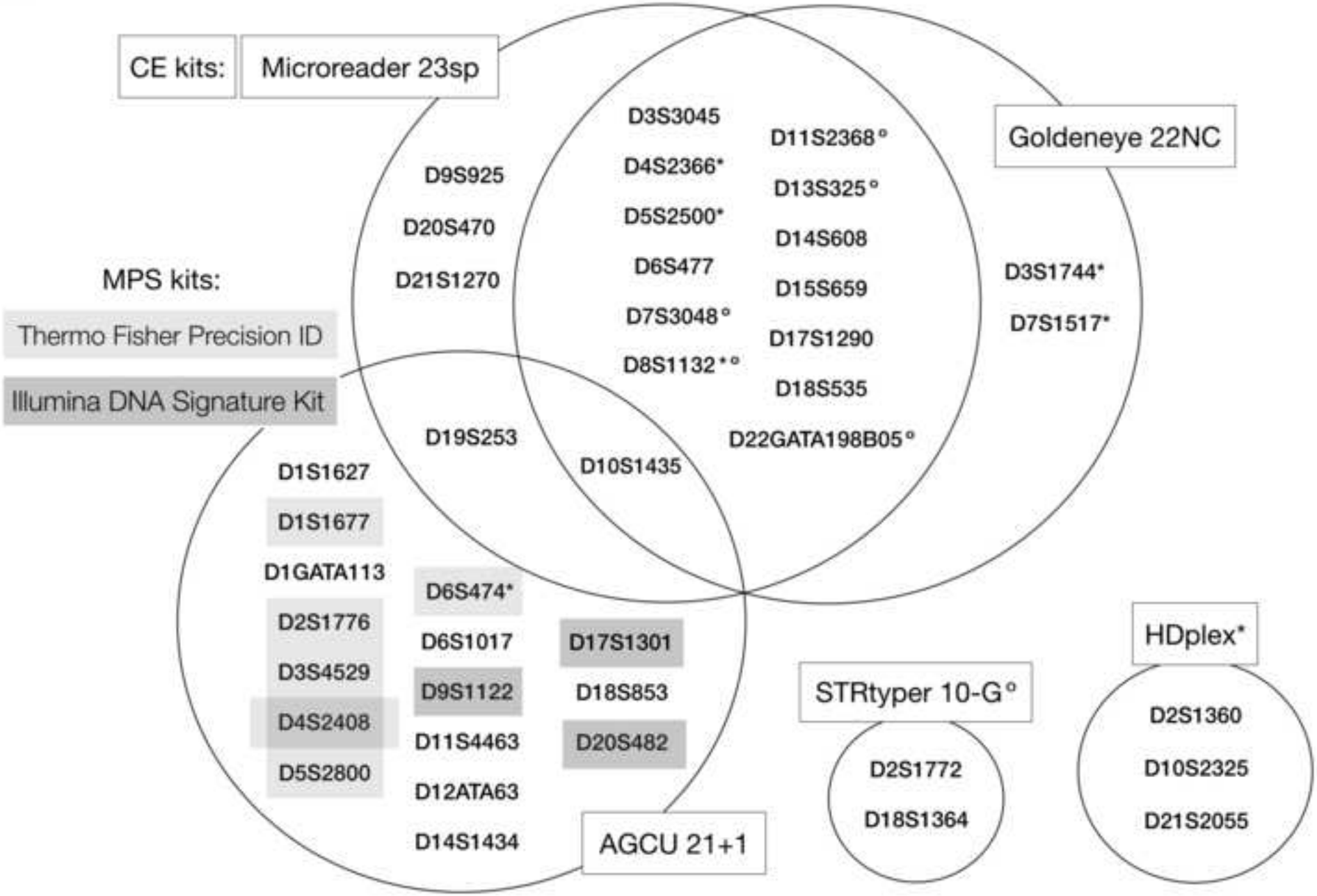
CE kits:
Microreader 23sp
Goldeneye 22NC
D3S3045
D11S2368°
D4S2366*
D13S325°
D9S925
D5S2500*
D14S608
D20S470
D6S477
D3S1744*
D21S1270
D15S659
D7S1517*
MPS kits:
D7S3048°
Thermo Fisher Precision ID
D17S1290
D8S1132**°
Illumina DNA Signature Kit
D18S535
D19S253
D22GATA198B05°
D10S1435
D1S1627
D1S1677
D1GATA113
D6S474*
D2S1776
D6S1017
D17S1301
D3S4529
D9S1122
D18S853
D4S2408
D11S4463
D20S482
D5S2800
D12ATA63
D14S1434
AGCU 21+1
STRtyper 10-G°
HDplex*
D2S1772
D2S1360
D18S1364
D10S2325
D21S2055

Figure 2

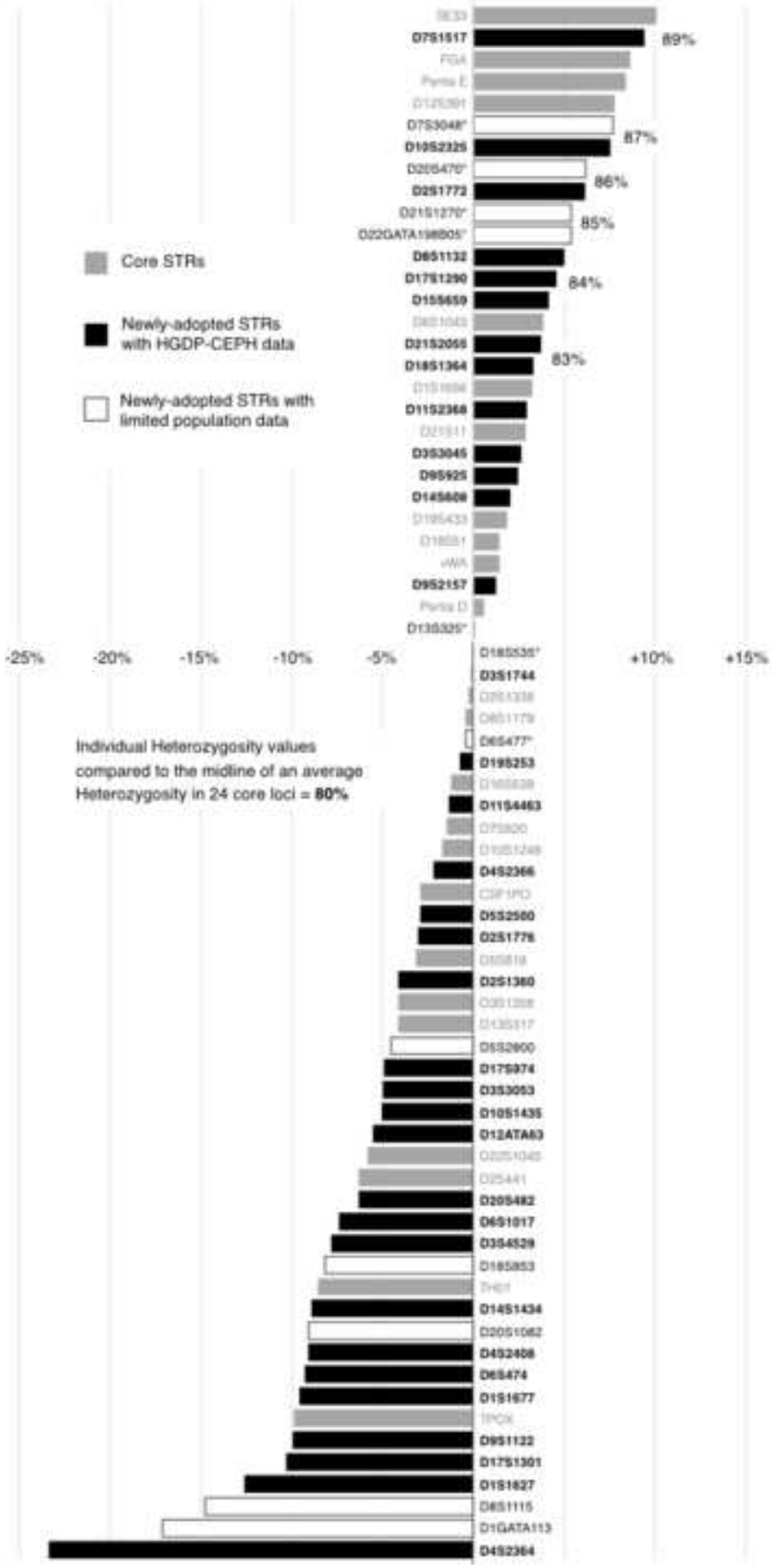
Core STRs
Newly-adopted STRs with HGDP-CEPH data
Newly-adopted STRs with limited population data
89%
87%
86%
85%
84%
83%
-25%
-20%
-15%
-10%
-5%
+10%
+15%
Individual Heterozygosity values compared to the midline of an average Heterozygosity in 24 core loci = 80%

Table 1A

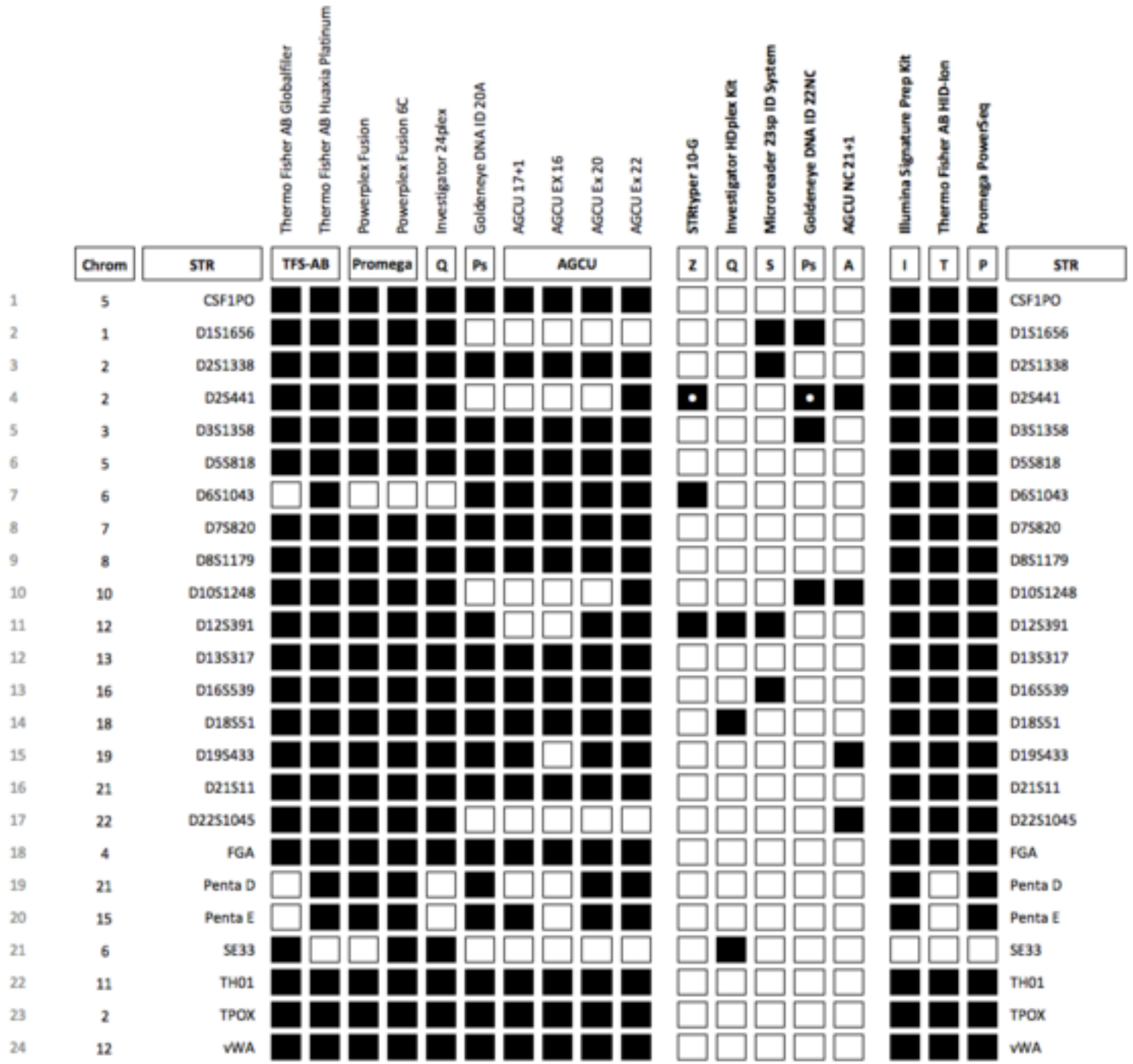
Thermo Fisher AB Globalfiler
Thermo Fisher AB Huaxia Platinum
Powerplex Fusion
Powerplex Fusion 6C
Investigator 24plex
Goldeneye DNA ID 20A
AGCU 17+1
AGCU EX 16
AGCU Ex 20
AGCU Ex 22
STRtyper 10-G
Investigator HDplex Kit
Microreader 23sp ID System
Goldeneye DNA ID 22NC
AGCU NC 21+1
Illumina Signature Prep Kit
Thermo Fisher AB HID-Ion
Promega PowerSeq
Chrom
STR
TFS-AB
Promega
Q
Ps
AGCU
Z
Q
S
Ps
A
I
T
P
STR
1 5 CSF1PO
2 1 D1S1656
3 2 D2S1338
4 2 D2S441
5 3 D3S1358
6 5 D5S818
7 6 D6S1043
8 7 D7S820
9 8 D8S1179
10 10 D10S1248
11 12 D12S391
12 13 D13S317
13 16 D16S539
14 18 D18S51
15 19 D19S433
16 21 D21S11
17 22 D22S1045
18 4 FGA
19 21 Penta D
20 15 Penta E
21 6 SE33
22 11 TH01
23 2 TPOX
24 12 vWA

Table 1B

	Chrom	STR	rs number	GRCh37 RR start NT / No.		9947A	STRtyper 10-G Z	Investigator HDplex Kit Q	Microreader 23sp ID System S	Goldeneye DNA ID 22NC Ps	AGCU NC 21+1 A	Illumina Signature Prep Kit I	Thermo Fisher AB HID-Ion T	Promega PowerSeq P	STR
1	1	D1GATA113	-	7442891	11	11,12									D1GATA113
2	1	D151627	rs112605029	106963714	13	13,14									D151627
3	1	D151677	rs112228372	163559816	15	13,14							■		D151677
4	2	D251360	rs113680434	17491986	21	23,24		■							D251360
5	2	D251772	rs111745293	67051138	25	26,28	■								D251772
6	2	D251776	rs112494037	169645403	11	10							■		D251776
9	3	D354529	-	85852634	12	13							■		D354529
8	3	D353045	rs113101511	106990014	14	9			■	■					D353045
7	3	D351744	rs113865588	147092539	18	17		■		■					D351744
10	4	D452366	rs113820309	6484841	14	11,13		■	■	■					D452366
11	4	D452408	rs113424060	31304420	9	9,10						■	■		D452408
12	5	D552500	rs111362704	58697270	11	15,16		■	■	■					D552500
13	5	D552800	-	58698958	17	14,23							■		D552800
14	6	D65477	rs113821764	6140627	18	10.2,13			■	■					D65477
15	6	D651017	-	41677269	10	9,10									D651017
17	6	D65474	rs113991233	112879153	16	13,17		■			■		■		D65474
17	7	D753048	rs112037359	21266718	19	24	■		■	■					D753048
18	7	D751517	rs112397288	123497699	22	19,25		■		■					D751517
19	8	D851132	-	107328920	20	19,21	■	■	■	■					D851132
20	9	D95925	rs112046838	18289120	17	15,16			■						D95925
21	9	D951122	rs112830978	79688742	12	12,13						■			D951122
22	10	D1051435	-	2243332	11	10,11			■	■	■				D1051435
23	10	D1052325	-	12793051	15	9,10		■							D1052325
24	11	D1152368	rs113789681	19281148	21	19,21	■		■	■					D1152368
25	11	D1154463	-	130872406	12	12,13									D1154463
26	12	D12ATA63	-	108322367	13	13							■		D12ATA63
27	13	D135325	rs113058481	43173440	19	20,21	■		■	■					D135325
28	14	D1451434	rs111914457	95308391	13	11,13							■		D1451434
29	14	D145608	rs112003987	28849469	14	7,11			■	■					D145608
30	15	D155659	rs111549851	46374109	14	14,16			■	■					D155659
31	17	D1751290	rs113538983	56331470	14	17,18			■	■					D1751290
32	17	D1751301	rs113771602	72680994	12	12						■			D1751301
33	18	D185853	-	3990629	12	11,14									D185853
34	18	D185535	rs112289696	38148827	16	13,14			■	■					D185535
35	18	D1851364	-	63400234	15	13,15	■								D1851364
36	19	D195253	-	15728295	12	7,8			■		■				D195253
37	20	D205482	rs112524392	4506338	14	14,15						■			D205482
38	20	D205470	rs112908724	17372552	10	15,16			■						D205470
39	21	D21S1270	rs113850076	31706806	15	10,14			■						D21S1270
40	21	D2152055	rs113225349	41191435	15	19.1,26		■							D2152055
41	22	D22GATA198B05	-	17650701	18	18,19	■		■	■					D22GATA198B05
42	3	D353053	rs111699340	171750965	9	9,11									D353053
43	4	D452364	rs113715572	93517373	9	9,10									D452364
44	8	D851115	rs113659271	42536591	8	9,18									D851115
45	9	D952157	rs113499999	136035669	10	7,13									D952157
46	17	D175974	rs112195386	10518745	11	7,10									D175974
47	20	D2051082	rs113175620	53865938	14	11,14									D2051082

Table 2 Genomic positions, genetic map distances and recombination fractions (Rc) of 47 syntenic (same chromosome) STR pairs. Core STRs in bold, STR pairs with less than a 10% recombination rate between them (Rc<0.1) highlighted in gray.

Chr.	STR (commonly used loci in bold)	Approximate STR position in Mb (GRCh37)	Repeat region start nucleotide coordinate (GRCh37)	HapMap proxy SNP position (GRCh37)	Proxy SNP-STR distance in nucleotides	Cumulative genetic map distance in cM	Genetic map distance in cM	Rc from Kosambi mapping function	Syntenic STR pair
1	D1GATA113	7.44	7442891	7442845	46	15.15831			
1	D1S1627	106.96	106963714	106964217	-503	131.77819	116.61988	0.490667	D1GATA113-D1S1627
1	D1S1677	163.56	163559816	163559784	32	173.11713	41.33894	0.339371	D1S1627-D1S1677
1	**D1S1656**	230.91	230905362	230905307	55	244.23489	71.11776	0.445045	D1S1677-D1S1656
2	**TPOX**	1.49	1493425	1493487	-62	1.66610			
2	D2S1360	17.49	17491856	17490232	1624	35.05805	33.39194	0.291778	TPOX-D2S1360
2	D2S1772	67.05	67051138	67051072	66	88.59125	53.53320	0.394856	D2S1360-D2S1772
2	**D2S441**	68.24	68239079	68239020	59	90.47903	1.88778	**0.018869**	**D2S1772-D2S441**
2	D2S1776	169.65	169645403	169643383	2020	177.02944	86.55041	0.469587	D2S441-D2S1776
2	**D2S1338**	218.88	218879582	218879435	147	223.48320	46.45376	0.365081	D2S1776-D2S1338
3	**D3S1358**	45.58	45582231	45582627	-396	67.17890			
3	D3S4529	85.85	85852634	85852702	-68	108.45611	41.27721	0.339038	D3S1358-D3S4529
3	D3S3045	106.99	106990014	106990619	-605	116.92246	8.46635	**0.083864**	**D3S4529-D3S3045**
3	D3S1744	147.09	147092539	147092143	396	157.24131	40.31884	0.333793	D3S3045-D3S1744
3	D3S3053	171.75	171750965	171749114	1851	177.06995	19.82864	0.188506	D3S1744-D3S3053
4	D4S2366	6.48	6484841	6484806	35	12.94670			
4	D4S2408	31.30	31304420	31305596	-1176	49.54939	36.60270	0.312160	D4S2366-D4S2408
4	D4S2364	93.52	93517373	93515918	1455	104.11410	54.56471	0.398675	D4S2408-D4S2364
4	**FGA**	155.51	155508888	155508100	788	156.81293	52.69883	0.391674	D4S2364-FGA
5	D5S2500	58.70	58697270	58697354	-84	70.32067			
5	D5S2800	58.70	58698958	58698677	281	70.32080	0.00013	**0.000001**	**D5S2500-D5S2800**
5	**D5S818**	123.11	123111250	123111652	-402	126.67284	56.35204	0.405002	D5S2800-D5S818
5	**CSF1PO**	149.46	149455887	149455757	130	154.43395	27.76112	0.252212	D5S818-CSF1PO
6	D6S477	6.14	6140627	6140845	-218	15.77241			
6	D6S1017	41.68	41677269	41677034	235	62.80381	47.03139	0.367755	D6S477-D6S1017
6	**SE33**	88.99	88986863	88986609	254	95.44921	32.64541	0.286812	D6S1017-SE33
6	**D6S1043**	92.45	92449943	92450035	-92	99.86628	4.41707	**0.044056**	**SE33-D6S1043**
6	D6S474	112.88	112879153	112879893	-740	118.66248	18.79620	0.179581	D6S1043-D6S474
7	D7S3048	21.27	21266718	21266723	-5	36.14071			
7	**D7S820**	83.79	83789542	83789257	285	100.20120	64.06049	0.428403	D7S3048-D7S820
7	D7S1517	123.50	123497699	123497068	631	132.07060	31.86940	0.281559	D7S820-D7S1517
8	D8S1115	42.54	42536591	42533383	3208	69.39870			
8	D8S1132	107.33	107328920	107330479	-1559	119.96228	50.56358	0.383144	D8S1115-D8S1132
8	**D8S1179**	125.91	125907107	125907927	-820	136.44313	16.48085	0.159088	D8S1132-D8S1179

Table 2
/Cont.

Chr.	STR (commonly used loci in bold)	Approximate STR position in Mb (GRCh37)	Repeat region start nucleotide coordinate (GRCh37)	HapMap proxy SNP position (GRCh37)	Proxy SNP-STR distance in nucleotides	Cumulative genetic map distance in cM	Genetic map distance in cM	Rc from Kosambi mapping function	Syntenic STR pair
9	D9S925	18.29	18289120	18289047	73	38.03303			
9	D9S1122	79.69	79688742	79688048	694	81.15767	43.12464	0.348770	D9S925-D9S1122
9	D9S2157	136.04	136035669	135983411	52258	156.36335	75.20567	0.452944	D9S1122-D9S2157
10	D10S1435	2.24	2243332	2243874	-542	3.40182			
10	D10S2325	12.79	12793051	12793258	-207	28.27346	24.87164	0.230048	D10S1435-D10S2325
10	**D10S1248**	131.09	131092508	131093166	-658	169.89917	141.62571	0.496547	D10S2325-D10S1248
11	**TH01**	2.19	2192318	2192166	152	4.48933			
11	D11S2368	19.28	19281148	19281171	-23	32.88891	28.39958	0.256941	TH01-D11S2368
11	D11S4463	130.87	130872406	130873262	-856	151.19289	118.30398	0.491269	D11S2368-D11S4463
12	**VWA**	6.09	6093143	6093924	-781	15.63031			
12	**D12S391**	12.45	12449954	12450501	-547	27.57129	11.94098	0.117190	VWA-D12S391
12	D12ATA63	108.32	108322367	108322352	15	126.91861	99.34732	0.481547	D12S391-D12ATA63
13	D13S325	43.17	43173440	43173444	-4	44.90825			
13	**D13S317**	82.72	82722160	82721723	437	79.83074	34.92249	0.301691	D13S325-D13S317
14	D14S1434	28.85	28849469	28850285	-816	20.49462			
14	D14S608	95.31	95308391	95308332	59	97.53227	77.03765	0.456123	D14S1434-D14S608
15	D15S659	46.37	46374109	46371620	2489	49.51748			
15	**Penta E**	97.37	97374245	97377441	-3196	124.05054	74.53306	0.451723	D15S659-Penta E
17	D17S974	10.52	10518745	10518759	-14	27.37262			
17	D17S1290	56.33	56331470	56332539	-1069	84.72292	57.35030	0.408380	D17S974-D17S1290
17	D17S1301	72.68	72680994	72680495	499	113.11145	28.38853	0.256860	D17S1290-D17S1301
18	D18S853	3.99	3990629	3990470	159	12.05264			
18	D18S535	38.15	38148827	38147797	1030	60.02540	47.97276	0.372017	D18S853-D18S535
18	**D18S51**	60.95	60948900	60949983	-1083	88.92051	28.89511	0.260570	D18S535-D18S51
18	D18S1364	63.40	63400234	63400151	83	91.21746	2.29695	0**.022953**	**D18S51-D18S1364**
19	D19S253	15.73	15728295	15728103	-321	39.27234			
19	**D19S433**	30.42	30417142	30417603	-461	51.72618	12.54384	0.122871	D19S253-D19S433
20	D20S482	4.51	4506338	4506638	-300	13.25549			
20	D20S470	17.37	17372552	17372509	43	40.68129	27.42580	0.249704	D20S482-D20S470
20	D20S1082	53.87	53865938	53865700	238	85.41915	44.73786	0.356868	D20S470-D20S1082
21	**D21S11**	20.55	20554291	20554558	-267	14.64555			
21	D21S1270	31.71	31706806	31706201	605	31.69620	17.05065	0.164191	D21S11-D21S1270
21	D21S2055	41.19	41191435	41191871	-436	49.46478	17.76858	0.170565	D21S1270-D21S2055
21	**Penta D**	45.06	45056086	45056178	-92	59.37591	9.91113	**0.097833**	**D21S2055-Penta D**
22	D22GATA198B05	17.65	17650701	17651831	-1130	7.39585			
22	**D22S1045**	37.54	37536327	37535663	664	46.21362	38.81778	0.325305	D22GATA198B05-D22S1045

Table 3 Summary of repeat region sequence annotation of 47 recently-adopted STRs using the human Reference Sequence (Ref. Seq.) and comparison to published sequencing studies. Bold repeat structure data denotes full agreement between the annotations of this study and previous publications.

No.	STR	Repeat region structure identified in Reference Sequence	Published repeat region structure	Published sequencing data compared to annotation of repeat region in this study	Orientation of Ref. Seq. to published data
1	D1GATA113	**[GATA]a**	**[GATA]**	Hill, JFS 2008 [15]	Forward
2	D1S1627	**[ATT]a**	**[ATT]**	Hill, JFS 2008	Forward
3	D1S1677	**[TTCC]a**	**[TTCC]**	Hill, JFS 2008	Forward
4	D2S1360	**[AGAT]a [AGAC]b [AGAT]c**	**[TATC][TGTC][TATC]**	Pemberton, BMC 2009 [29], with 1-NT frame shift	Reverse
5	D2S1772	[CTAT]a [(GTAT)(CTAT)]b [CTAT]c [CTGT]d [CTAT]e [CTGT]f [CTAT]g	[GATA] [CATA GATA] [CACA][GATA][CATA][GATA]	Similar to Huang, AJB 2012 [19]	Reverse
6	D2S1776	**[AGAT]a**	**[AGAT]**	Hill, JFS 2008	Forward
7	D3S4529	**[GATA]a**	**[GATA]**	Hill, JFS 2008	Forward
8	D3S3045	**[AGAT]a AT [AGAT]b**	**[AGAT] N2 [AGAT]**	Pemberton, BMC 2009	Forward
9	D3S1744	[ATAG]a ATG [ATAG]b AT [ATAG]c		No published sequence data found	Forward
10	D3S3053	**[GATA]a**	**[TATC]**	Hill, JFS 2008	Reverse
11	D4S2366	[GATA]a [GATT]b [GATA]c GAC [GATA]d	[GATA] only	More complex sequence than Pemberton, BMC 2009	Forward
12	D4S2408	**[ATCT]a**	**[ATCT]**	Hill, JFS 2008	Forward
13	D4S2364	[ATTC]a [ATCC]b [ATTC]c	[ATTC]	Similar to Hill, JFS 2008	Forward
14	D5S2500	**[CTAT]a**	**[ATAG]**	Pemberton, BMC 2009	Reverse
15	D5S2800	[GGTA]a [GACA]b [GATA]c [GATT]d	[GATA][GATT] only	More complex sequence than Hill, JFS 2008	Forward
16	D6S477	[TCTA]a [TA]b [TCTA]c [TA]d [TCTA]e [TCTG]f [TCTA]g	[TCTA] [TA] [TCTA] [TA] [TCTA]	Similar to Barral, IJLM 2000 [30]	Forward
17	D6S1017	**[GGAT]a**	**[ATCC]**	Hill, JFS 2008	Reverse
18	D6S474	[TAGA]a [GATA]b	[GATA] N3 [GATA]	Adjusted to match reported sequence in Hill, JFS 2008	Forward
19	D7S3048	**[TATC]a [TACC]b [CACC]c**	**[TATC][TACC][CACC]**	Huang, AJB 2012	Forward
20	D7S1517	[CTTT]a [GTTT]b [CTTT]c [GTTT]d [CTTT]e		No published sequence data found	Forward
21	D8S1115	**[TAA]a**	**[ATT]**	Hill, JFS 2008	Reverse
22	D8S1132	**[TCTA]a TCA [TCTA]b**	**[TCTA] TCA [TCTA]**	Huang, AJB 2012	Forward
23	D9S925	[TATA]a [TGTC]b [TATC]c [TACC]d [TATC]e		No published sequence data found	Forward
24	D9S1122	**[TAGA]a**	**[TAGA]**	Hill, JFS 2008	Forward
25	D9S2157	**[ATA]a**	**[ATA]**	Hill, JFS 2008	Forward
26	D10S1435	**[TATC]a**	**[TATC]**	Hill, JFS 2008	Forward
27	D10S2325	[ATAAG]a		No published sequence data found	Forward
28	D11S2368	[TATC]a [TGTC]b [TATC]c	[ATAG][ACAG][ATAG]	Similar to Huang, AJB 2012 (1 NT frame-shift)	Reverse
29	D11S4463	**[TATC]a**	**[TATC]**	Hill, JFS 2008	Forward
30	D12ATA63	**[TTG]a [TTA]a**	**[TAA][CAA]**	Hill, JFS 2008	Reverse
31	D13S325	**[TCTA]a**	**[AGAT]**	Huang, AJB 2012	Reverse
32	D14S1434	**[CTGT]a [CTAT]b**	**[CTGT][CTAT]**	Hill, JFS 2008	Forward
33	D14S608	[GATA]a N12 [GATA]b [GACA]c [GATA]d		No published sequence data found	Forward
34	D15S659	**[TATC]a**	**[GATA]**	Pemberton, BMC 2009	Reverse
35	D17S974	[ATAG]a ATG [ATAG]b	[CTAT]	Similar to Hill, JFS 2008 (ATG may be counted)	Reverse
36	D17S1290	**[AGAT]a N28 [ATAG]b**	**[AGAT] N28 [ATAG]**	Pemberton, BMC 2009	Forward
37	D17S1301	**[AGAT]a**	**[AGAT]**	Hill, JFS 2008	Forward
38	D18S853	**[ATA]a**	**[ATA]**	Hill, JFS 2008	Forward
39	D18S535	[AGAT]a [AGAC]b [AGAT]c GAT [AGAT]d	[GATA]	More complex sequence than Pemberton, BMC 2009	Forward
40	D18S1364	[TAGA]a [TACA]b [TAGA]c	[GATA][CATA][GATA]	Similar to Huang, AJB 2012 (2 NT frame-shift)	Forward
41	D19S253	[ATCT]a		No published sequence data found	Forward
42	D20S482	**[AGAT]a**	**[AGAT]**	Hill, JFS 2008	Forward
43	D20S470	[AGGA]a		No published sequence data found	Forward
44	D20S1082	**[ATA]a**	**[ATA]**	Hill, JFS 2008	Forward
45	D21S1270	[ATAG]a ATG [ATAG]b ATG [ATAG]c ATG [ATAG]d		No published sequence data found	Forward
46	D21S2055	[CTAT]a [CTAA]b [CTAT]c [CTA]d [CTAT]e N30 [TATC]d [TAT]e [TATC]f	[TATC] N29 [TATC]	More complex sequence than Pemberton, BMC 2009	Forward
47	D22GATA198B05	**[CTCT]a [ATCT]b [ACCT]c**	**[CTCT][ATCT][ACCT]**	Huang, AJB 2012	Forward

Table 4 Genomic details of 44 novel autosomal STRs, developed in published multiplexes for CE genotyping [15,21-25], but not in commercially available kits.

No.	STR	Multiplex	dbSNP rs-number	Repeat motif(s)	Chr.	GRCh37 nucleotide coordinates of repeat region	Repeats in Reference Sequence
1	D1S1679	Phillips AIM-STRs	rs112703460	[AAGG]a	1	1:162361962-162362021	15
2	D1S1171	Asamura	rs111787597	[AAAG]a	1	1:201917471-201917506	9
3	D2S1242	Asamura	rs112009758	[CTTT]a [CCTT]b [CTTT]c	2	2:221218009-221218072	16
4	D2S427	Phillips AIM-STRs	rs111704160	[GATT]a [GATA]b GAT [GATA]c	2	2:232206330-232206388	14
5	D3S2387	Pinto	rs113254598	[GATA]a GAT [GATA]b [GACA]c [GAGA]d [GATA]e [GACA]f [GATA]g [GACA]h [GATA]i	3	3:1036300-1036414	27
6	D3S4545	Phillips AIM-STRs	-	[CTCT]a N7 [CAGA]b [TAGA]c [CAGA]d	3	3:8585065-8585191	31
7	D3S2402	Liu linked STRs	rs113124331	[AGGA]a	3	3:58216894-58216945	13
8	D3S2452	Liu linked STRs	rs111655766	[ATCT]a ATC [ATCT]b	3	3:58698753-58698819	16
9	D3S1766	Liu linked STRs	rs111987670	[ATCT]a	3	3:58981702-58981749	12
10	D3S2406	Phillips / Pinto	rs112283702	[TATC]a [TGTC]b [CGTC]c [TGTC]d [CGTC]e [TGTC]f [CGTC]g [CATC]h	3	3:73258449-73258580	33
11	D3S4554	Liu linked STRs	rs111656608	[CTAT]a [CTGT]b [CTAT]c CAT [CTAT]d [CCAT]e [CTAT]f	3	3:82591317-82591387	17
12	D3S2388	Liu linked STRs	rs112049246	[ATCT]a	3	3:83164192-83164239	12
13	D3S1545	Asamura	rs112182395	[TATC]a	3	3:161673154-161673181	7
14	D3S3051	Liu linked STRs	-	[TCTA]a	3	3:171695743-171695786	11
15	D4S2404	Liu linked STRs	rs112609744	[TATC]a [AATC]b [TATC]c	4	4:93499868-93499915	12
16	D5S2503	Pinto	-	[ATAG]a ATG [ATAG]b	5	5:23591292-23591350	14
17	D5S1457	Phillips AIM-STRs	rs113452942	[TATC]a	5	5:41033884-41033931	12
18	D7S2201	Phillips AIM-STRs	rs111938869	[TATC]a	7	7:5630684-5630727	11
19	D8S639	Grubweiser	-	[ATAG]a ATG [ATAG]b ATG [ATAG]c	8	8:16771052-16771169	28
20	D8S306	Grubweiser	-	[GAAA]a N21 [GAAA]b [AGAA]c AAG [AAGG]d [GAAG]e [AAGG]f	8	8:60917924-60918079	33
21	D9S324	miniSTR in original study	-	[TATC]a	9	9:6585833-6585892	15
22	D9S1118	Phillips AIM-STRs	rs112557788	[TATC]a	9	9:31925368-31925418	12
23	D9S304	Grubweiser	rs112793459	[TAGA]a N19 [GATA]b [GATT]c [GATA]d [GATG]e [GATA]f	9	9:32324058-32324128	13
24	D9S938	Pinto	rs111842728	[TTCC]a N25 [TTCC]b [TTTC]c	9	9:105984102-105984198	18
25	D9S938	miniSTR in original study	rs113034989	[TATG]a [TATC]b	9	10:12695023-12695066	11
26	D10S1237	Pinto	rs111899297	[TAGA]a [TAGT]b [TAGA]c	10	10:116120245-116120324	20
27	D11S488	Grubweiser	-	[TTCC]a [TTTC]b [CTTT]c [CCTT]d [TCCT]e [TTCT]f N9 [CTTT]g	11	11:123384224-123384380	37
28	D11S1304	Phillips AIM-STRs	rs113308565	[TATC]a [TGTC]b [TATC]c	11	11:132586286-132586349	16
29	D12S297	Phillips AIM-STRs	rs113633696	[TATC]a N6 [CATC]b [TATC]c [CATC]d N3 [TATC]e [CATC]f [TATC]g	12	12:52613136-52613228	21
30	D14S1426	Phillips AIM-STRs	rs113351154	[AGAT]a	14	14:100619590-100619641	13
31	D15S822	Phillips AIM-STRs	-	[TATC]a [TCTA]b	15	15:27390743-27390810	17
32	D16S753	Pinto	rs113398559	[TCCT]a	16	16:31273566-31273601	9
33	D16S3253	Asamura / Grubweiser	-	[GATA]a	16	16:54786687-54786722	9
34	AC001348A	Liu linked STRs	-	[TTCT]a	17	17:14593456-14593503	12
35	AC001348B	Liu linked STRs	-	[ACAAT]a	17	17:14615594-14615668	15
36	D17S976	Grubweiser	-	[TAGA]a N14 [TAGA]b TGA [TAGA]c [TTGA]d	17	17:18105058-18105174	25
37	D17S975	Liu linked STRs	rs113917437	[TCCA]a TCA [TCCA]b [CCCA]c [TCCA]c TCA [TCCA]d [CCCA]e [TCCA]f TCA [TCCA]g [CCCA]h [TCCA]i	17	17:28100038-28100130	21
38	D17S1294	Liu linked STRs	rs113634282	[AAGA]a [AGAG]b [AGGA]c	17	17:28382313-28382400	22
39	D18S1270	Grubweiser	-	[TCTA]a	18	18:61392605-61392644	10
40	D20S161	Asamura	rs112764064	[ATAG]a N16 [ATAG]b	20	20:16622120-16622195	15
41	D21S1432	Phillips AIM-STRs	rs111610289	[TATC]a N14 [TATC]b [TATG]c [TATC]d	21	21:17343490-17343551	13
42	D21S1437	Pinto / Asamura / Grubweiser	rs112721088	[GAAG]a [GAGG]b [GAAG]c [GACG]d [GAAG]e [GAGG]f [GAAG]g	21	21:21646855-21646930	19
43	D22S689	Pinto	rs112683089	[TAGA]a [CAGA]a [TAGA]a [TAGG]a [CAGG]a [CAGA]a	22	22:28856596-28856687	23
44	D22S534	Pinto / Grubweiser	rs113835971	[ATAC]a	22	22:40965721-40965780	15