Forensic Science International: Genetics 25 (2016) 214–226

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research Paper

Characterization of genetic sequence variation of 58 STR loci in four major population groups

Nicole M.M. Novroski[a,*], Jonathan L. King[a], Jennifer D. Churchill[a], Lay Hong Seah[b], Bruce Budowle[a,c]

[a] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd. Fort Worth, TX 76107, USA
[b] Department of Chemistry Malaysia Kuching, Ministry of Science, Technology and Innovation (MOSTI) Malaysia
[c] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 4 August 2016
Received in revised form 15 September 2016
Accepted 27 September 2016
Available online 28 September 2016

Keywords:
STR
Sequence variation
SNP
Massively parallel sequencing
ForenSeq™
DNA signature prep kit
STRait razor
Forensic DNA

ABSTRACT

Massively parallel sequencing (MPS) can identify sequence variation within short tandem repeat (STR) alleles as well as their nominal allele lengths that traditionally have been obtained by capillary electrophoresis. Using the MiSeq FGx Forensic Genomics System (Illumina), STRait Razor, and in-house excel workbooks, genetic variation was characterized within STR repeat and flanking regions of 27 autosomal, 7 X-chromosome and 24 Y-chromosome STR markers in 777 unrelated individuals from four population groups. Seven hundred and forty six autosomal, 227 X-chromosome, and 324 Y-chromosome STR alleles were identified by sequence compared with 357 autosomal, 107 X-chromosome, and 189 Y-chromosome STR alleles that were identified by length. Within the observed sequence variation, 227 autosomal, 156 X-chromosome, and 112 Y-chromosome novel alleles were identified and described. One hundred and seventy six autosomal, 123 X-chromosome, and 93 Y-chromosome sequence variants resided within STR repeat regions, and 86 autosomal, 39 X-chromosome, and 20 Y-chromosome variants were located in STR flanking regions. Three markers, D18S51, DXS10135, and DYS385a-b had 1, 4, and 1 alleles, respectively, which contained both a novel repeat region variant and a flanking sequence variant in the same nucleotide sequence. There were 50 markers that demonstrated a relative increase in diversity with the variant sequence alleles compared with those of traditional nominal length alleles. These population data illustrate the genetic variation that exists in the commonly used STR markers in the selected population samples and provide allele frequencies for statistical calculations related to STR profiling with MPS data.

1. Introduction

The current standard methodology in forensic DNA typing relies on amplification of short tandem repeat (STR) markers by the polymerase chain reaction (PCR) and allele sizes (i.e., length-based) determined for each locus using capillary electrophoresis (CE). Massively parallel sequencing (MPS), also known as next generation sequencing (NGS), allows high throughput sequencing of STR amplicons, which can identify nominal length-based (LB) genetic variation but equally as well inter-allelic sequence (sequence-based; SB) variation [1–5]. The increased effective number of alleles per marker for some STR loci improves discrimination power, which may be invaluable in some cases of kinship analysis and for mixture de-convolution. Furthermore, allelic variation captured using MPS may be useful towards understanding of STR mutations and their rates and may contribute to evolutionary studies using STR markers.

To exploit the full power of MPS with STR typing, the underlying genetic variation needs to be described in relevant populations [6–8]. Gettings et al. [1] described a rather comprehensive characterization of allelic variation for 24 of the commonly used autosomal STR loci. While hundreds of unique sequences were identified and characterized, there likely is a great deal of genetic variation yet to be identified. Larger scale population studies are necessary for establishing allele frequencies that can be used for calculating the strength of MPS-generated DNA evidence. J.D. Churchill (personal communication; manuscript in preparation) described performance testing of the MiSeq FGx Forensic Genomics System

* Corresponding author.
E-mail address: Nicole.Novroski@live.unthsc.edu (N.M.M. Novroski).

http://dx.doi.org/10.1016/j.fsigen.2016.09.007
1872-4973/

Table 1
Summary of observed sequence variation by locus and population group.

Autosomal Loci	# of LB alleles	# of SB alleles	Difference	# of novel RR variants	# of novel FR variants
D2S1338	14	64	50	15	0
AFA	12	49	37	5	0
CAU	13	32	19	3	0
HIS	12	31	20	2	0
ASN	12	31	19	5	0
D12S391	19	79	60	18	1
AFA	17	52	35	5	0
CAU	16	56	41	8	0
HIS	14	44	30	4	1
ASN	13	41	28	6	0
D21S11	22	85	63	35	0
AFA	18	54	36	17	0
CAU	12	29	17	4	0
HIS	16	35	19	4	0
ASN	12	42	30	13	0
vWA	11	35	24	6	3
AFA	9	32	23	5	3
CAU	7	14	7	0	0
HIS	8	22	14	1	0
ASN	7	12	5	0	0
D13S317	9	26	17	0	2
AFA	8	18	10	0	1
CAU	8	16	8	0	1
HIS	9	15	6	0	0
ASN	7	16	9	0	0
D7S820	11	26	15	4	13
AFA	9	18	9	0	9
CAU	8	17	9	2	7
HIS	7	12	5	0	6
ASN	8	18	10	2	9
D8S1179	12	30	18	4	1
AFA	11	24	13	2	0
CAU	11	19	8	1	0
HIS	10	24	14	1	0
ASN	9	19	10	2	1
D16S539	8	19	11	2	8
AFA	7	13	6	0	5
CAU	7	14	7	1	5
HIS	8	14	6	1	4
ASN	7	13	6	0	5
D3S1358	10	22	12	3	0
AFA	8	18	10	2	0
CAU	8	14	6	0	0
HIS	7	18	11	1	0
ASN	6	13	7	0	0
D9S1122	8	17	9	7	1
AFA	7	14	7	4	1
CAU	7	13	6	4	0
HIS	6	12	6	3	0
ASN	8	15	7	6	0
D20S482	9	18	9	5	8
AFA	8	15	7	4	7
CAU	8	13	5	4	5
HIS	8	12	4	3	4
ASN	8	13	5	4	4
D5S818	9	18	9	0	0
AFA	8	15	7	0	0
CAU	8	13	5	0	0
HIS	8	13	5	0	0
ASN	7	12	5	0	0
D1S1656	18	34	16	11	1
AFA	14	24	10	6	1
CAU	14	24	10	5	0
HIS	14	21	7	3	0
ASN	13	17	4	0	0
D18S51	21	37	16	7	9
AFA	16	26	10	5	7
CAU	13	16	3	2	0
HIS	14	18	4	1	2
ASN	14	14	0	0	0
D2S441	14	24	10	6	2
AFA	11	19	8	4	1
CAU	8	11	3	0	1
HIS	9	14	5	0	2
ASN	9	17	8	3	2
D6S1043	20	31	11	14	1

Table 1 (*Continued*)

Autosomal Loci	# of LB alleles	# of SB alleles	Difference	# of novel RR variants	# of novel FR variants
AFA	14	19	5	7	0
CAU	13	19	6	7	0
HIS	16	19	3	4	0
ASN	13	15	2	1	1
D22S1045	11	17	6	0	7
AFA	8	13	5	0	54
CAU	8	8	0	0	0
HIS	10	11	1	0	2
ASN	7	7	0	0	0
D19S433	17	23	6	6	1
AFA	16	20	4	3	1
CAU	12	12	0	0	0
HIS	12	13	1	1	0
ASN	13	14	1	2	0
D17S1301	9	12	3	7	0
AFA	8	10	2	5	0
CAU	8	9	1	4	0
HIS	7	8	1	3	0
ASN	7	8	1	3	0
D4S2408	6	8	2	2	0
AFA	6	7	1	1	0
CAU	6	7	1	1	0
HIS	5	6	1	0	0
ASN	5	7	2	1	0
FGA	26	34	8	7	0
AFA	20	23	3	3	0
CAU	15	16	1	0	0
HIS	14	20	6	3	0
ASN	16	16	0	1	0
PentaD	16	20	4	1	4
AFA	14	16	2	0	2
CAU	12	15	3	1	3
HIS	11	11	0	0	0
ASN	11	11	0	0	0
PentaE	22	27	5	3	0
AFA	15	19	4	2	0
CAU	17	17	0	0	0
HIS	19	20	1	1	0
ASN	19	19	0	0	0
CSF1PO	9	11	2	1	0
AFA	8	9	1	0	0
CAU	9	9	0	0	0
HIS	8	8	0	0	0
ASN	8	9	1	1	0
D10S1248	9	10	1	1	0
AFA	7	7	0	0	0
CAU	7	7	0	0	0
HIS	9	9	0	0	0
ASN	7	8	1	1	0
TH01	9	10	1	0	1
AFA	7	8	1	0	1
CAU	9	9	0	0	0
HIS	6	6	0	0	0
ASN	6	6	0	0	0
TPOX	9	9	0	0	0
AFA	7	7	0	0	0
CAU	7	7	0	0	0
HIS	7	7	0	0	0
ASN	9	9	0	0	0
X-Chromosome Loci	# of LB alleles	# of SB alleles	Difference	# of novel RR variants	# of novel FR variants
DXS10135	43	107	64	79	16
AFA	37	70	33	48	14
CAU	25	52	27	31	7
HIS	30	66	36	45	10
ASN	23	42	19	27	1
DXS10074	20	46	26	23	10
AFA	18	37	19	21	4
CAU	13	20	7	4	6
HIS	15	25	10	8	5
ASN	10	12	2	2	0
DXS10103	10	23	13	12	0
AFA	7	15	8	5	0
CAU	10	17	7	6	0
HIS	7	15	8	5	0
ASN	7	11	4	2	0
DXS8378	7	13	6	2	4
AFA	6	9	3	0	3

Table 1 (*Continued*)

Autosomal Loci	# of LB alleles	# of SB alleles	Difference	# of novel RR variants	# of novel FR variants
CAU	7	9	2	2	0
HIS	5	7	2	1	1
ASN	5	5	0	0	0
DXS7132	9	16	7	4	5
AFA	7	8	1	1	0
CAU	6	7	1	0	1
HIS	9	10	1	2	1
ASN	7	12	5	1	4
HPRTB	10	14	4	2	4
AFA	9	12	3	2	2
CAU	9	9	0	0	1
HIS	6	6	0	0	0
ASN	7	8	1	0	1
DXS7423	8	9	1	1	0
AFA	7	7	0	0	0
CAU	6	6	0	0	0
HIS	6	7	1	1	0
ASN	4	4	0	0	0
Y-Chromosome Loci	# of LB alleles	# of SB alleles	Difference	# of novel RR variants	# of novel FR variants
DYS389II	9	44	35	7	2
AFA	6	12	6	4	0
CAU	7	24	17	3	2
HIS	4	11	7	0	0
ASN	7	25	18	1	0
DYF387S1	11	43	32	27	0
AFA	10	20	10	9	0
CAU	8	24	16	13	0
HIS	6	14	8	5	0
ASN	9	35	26	21	0
DYS390	7	17	10	7	1
AFA	6	7	1	1	0
CAU	6	10	4	3	1
HIS	4	5	1	0	0
ASN	5	10	5	5	0
DYS437	5	11	6	5	3
AFA	5	6	1	1	2
CAU	3	4	1	1	1
HIS	3	3	0	0	1
ASN	4	7	3	4	1
DYS635	9	19	10	10	0
AFA	8	9	1	3	0
CAU	7	13	6	5	0
HIS	5	8	3	1	0
ASN	7	11	4	8	0
DYS448	9	18	9	11	0
AFA	4	8	4	2	0
CAU	7	12	5	5	0
HIS	5	8	3	3	0
ASN	5	13	8	6	0
DYS522	5	8	3	1	2
AFA	4	4	0	0	0
CAU	5	6	1	1	0
HIS	4	4	0	0	0
ASN	5	7	2	0	2
DYS481	12	19	7	6	3
AFA	7	9	2	1	1
CAU	9	13	4	3	0
HIS	7	7	0	2	0
ASN	9	11	2	2	2
DYS612	11	17	6	7	0
AFA	9	9	0	2	0
CAU	9	9	0	1	0
HIS	7	7	0	0	0
ASN	9	14	5	5	0
DYS438	6	9	3	1	2
AFA	4	4	0	0	0
CAU	6	8	2	0	2
HIS	4	5	1	0	1
ASN	4	5	1	1	0
DYS385a-b	15	21	6	6	1
AFA	10	12	2	2	1
CAU	11	11	0	1	0
HIS	9	9	0	0	0
ASN	14	18	4	3	0
DYS533	6	8	2	1	1
AFA	5	5	0	1	0
CAU	6	6	0	0	0

Table 1 (*Continued*)

Autosomal Loci	# of LB alleles	# of SB alleles	Difference	# of novel RR variants	# of novel FR variants
HIS	5	5	0	0	0
ASN	4	5	1	0	1
DYS461	7	9	2	1	2
AFA	5	5	0	0	0
CAU	5	5	0	0	1
HIS	5	6	1	0	0
ASN	7	8	1	1	1
DYS570	9	11	2	2	0
AFA	8	8	0	0	0
CAU	7	9	2	2	0
HIS	5	5	0	0	0
ASN	8	8	0	0	0
DYS392	6	7	1	0	1
AFA	2	2	0	0	0
CAU	4	5	1	0	1
HIS	5	5	0	0	0
ASN	6	6	0	0	0
DYS460	6	7	1	0	1
AFA	3	4	1	0	1
CAU	4	4	0	0	0
HIS	3	3	0	0	0
ASN	6	6	0	0	0
DYS576	10	11	1	0	1
AFA	7	7	0	0	0
CAU	8	8	0	0	0
HIS	6	7	1	0	1
ASN	9	9	0	0	0
DYS19	5	5	0	0	0
AFA	4	4	0	0	0
CAU	5	5	0	0	0
HIS	5	5	0	0	0
ASN	4	4	0	0	0
DYS389I	6	6	0	0	0
AFA	4	4	0	0	0
CAU	4	4	0	0	0
HIS	3	3	0	0	0
ASN	6	6	0	0	0
DYS391	4	4	0	0	0
AFA	3	3	0	0	0
CAU	3	3	0	0	0
HIS	3	3	0	0	0
ASN	4	4	0	0	0
DYS439	6	6	0	1	0
AFA	4	4	0	0	0
CAU	5	5	0	1	0
HIS	5	5	0	0	0
ASN	5	5	0	0	0
DYS505	6	6	0	0	0
AFA	6	6	0	0	0
CAU	5	5	0	0	0
HIS	3	3	0	0	0
ASN	5	5	0	0	0
DYS549	6	6	0	0	0
AFA	5	5	0	0	0
CAU	6	6	0	0	0
HIS	4	4	0	0	0
ASN	5	5	0	0	0
DYS643	8	8	0	0	0
AFA	7	7	0	0	0
CAU	7	7	0	0	0
HIS	6	6	0	0	0
ASN	6	6	0	0	0
Y-GATA-H4	4	4	0	0	0
AFA	4	4	0	0	0
CAU	4	4	0	0	0
HIS	4	4	0	0	0
ASN	4	4	0	0	0

LB = length-based, SB = sequence-based, RR = repeat region, FR = flanking region.
*Counts per locus per population are characterized in Supplemental Table S2.
**Some novel allele variants contributed both to the expansion of LB and SB alleles, which is reflected by a larger number of variants when compared to the difference between SB and LB alleles.

(Illumina, San Diego, CA) and population data on the single nucleotide polymorphisms (SNPs) within the kit's genetic marker panel for the samples described herein. Because of the substantial amount of data and particularly due to the interest by the forensic DNA community, the underlying sequence variation within STR population data are described separately herein. Sequence variants that reside within flanking and repeat regions of 27 autosomal, 7 X-chromosome, and 24 Y-chromosome STR markers were identified in 777 individuals in four populations (African American, Caucasian, Hispanic, and Chinese). Using previously published sequence data [1,10–68] that described known sequence variants per locus, alleles were classified as either pre-existing (observed in the literature) or novel. The abundance of sequence-based variants within some of the commonly used STR markers demonstrates the increased genetic variation that may be exploited for human identity testing.

2. Methods and materials

2.1. Samples, extraction and quantification

Whole blood samples were obtained by venipuncture from 777 unrelated individuals from four major population groups (US Caucasian, N = 210; Hispanic, N = 198; African American, N = 200; and East Asian, i.e., Chinese, N = 169). All samples were anonymized and collected according to UNTHSC IRB-approved protocols. DNA was extracted using the Qiagen® QIAamp™ DNA Mini Kit (Qiagen, Valencia, CA) using the manufacturer's protocol and stored at −46 °C until needed [69]. The quantity of DNA was determined using the Qubit® 2.0 Fluorimeter (Thermo Fisher Scientific Inc., Waltham, NY) using the manufacturer's protocol [70].

2.2. ForenSeq™ library preparation and MiSeq sequencing

The ForenSeq™ DNA Signature Prep Kit was used to barcode and create libraries for the DNA samples, which were analyzed in batches of 32–34 samples per library, including controls, as per the manufacturer's protocols using either DNA Primer Mix A (for low template and degraded samples, a subset of the Chinese samples) or DNA Primer Mix B (African American, US Caucasian, Hispanic and the rest of the Chinese samples). Sequencing was performed on the MiSeq Desktop Sequencer using the MiSeq FGx Forensic Genomics System (Illumina) as per manufacturer's protocols, with modifications only to the Denature and Dilute step 5 of the procedure, adding 10 μl instead of 7 μl of pooled normalized library to the denatured pooled library tube for addition into the sequencing cartridge [71,72].

2.3. Sequence variation identification

Characterization of STR data and detection of inter-allelic sequence variants also were performed using a modified version of STRait Razor 2.0 and in-house Excel-based workbooks [73]. The current version of STRait Razor is both comprehensive and compatible with the STR sequences captured in this study and is available upon request (J.L. King, personal communication; manuscript in preparation). A minimum coverage threshold of 5× (selected operationally to capture more data) was used for STR allele-calling. A stutter threshold of 20% was used for initial screening. Alleles were reported using nomenclature recommended by Parson et al. [8]. The nomenclature used herein is based on practices in the literature but may change based on recommended standards for variant allele naming likely to be established by the International Society of Forensic Genetics, especially if novel variants are observed where permutations may exist for allele nomenclature. Sequence variants were described as either pre-existing or novel, based on whether or not they have been characterized in the literature, and further defined by the location of the sequence variant, as either a repeat region (RR) variant or a flanking region (FR) variant. It is important to note that the loci SE33, DYS456, DXS10148, and DXS8377 are not accessible using the ForenSeq™ UAS. Therefore no sequence data for these loci are reported.

2.4. Concordance testing

Concordance testing was performed by CE generated genotypes using the GlobalFiler® PCR Amplification Kit and AmpFLSTR® Yfiler® PCR Amplification Kit (Thermo Fisher Scientific, South San Francisco, CA) as described in Churchill et al. [36] on a subset of the population samples (n = 170 for Globalfiler; n = 59 for Yfiler). Two bioinformatics pipelines were used in this study; ForenSeq UAS and STRait Razor 2.0 [73] including in-house excel workbooks for all 777 samples. Discordance was defined as any instance in which an allele detected by one approach was not observed above the operationally defined coverage threshold by the comparison approach and/or the sequence was not the same.

2.5. Allele frequencies and population statistical analyses

Allele frequencies were determined using the counting method. Standard population genetics analyses (e.g., heterozygosity, random match probability, Hardy-Weinberg equilibrium (HWE), and linkage disequilibrium (LD) tests) were performed using Genetic Data Analysis (GDA) [74] and in-house excel workbooks. String sequence information for all characterized alleles described herein are available online at https://www.unthsc.edu/graduate-school-of-biomedical-sciences/molecular-and-medical-genetics/laboratory-faculty-and-staff/.

3. Results and discussion

The variation of all STR alleles is described both as LB and SB and follows the nomenclature recommended by Parson et al. [8] (Supplemental Table S2). Consistent with other studies [1,6,7,10], the diversity of some STR loci increased notably due to sequence variation (Table 1). A search of the peer-reviewed literature was performed to find as best is possible all known sequence variants in the 58 STR loci in the ForenSeq™ panel [1–67]. These data from 777 unrelated individuals from four major population groups were combined with variants described in the peer-reviewed literature to provide a comprehensive listing of known STR allele sequence variation (Supplemental Table S2). Novel sequences observed in the study herein are highlighted in yellow.

Four general categories of variation were observed with the SB information for the 777 population samples. First, the effective increase in allele number due to sequence variation compared with alleles characterized by repeat number (or length) alone was due predominately to internal sequence variation present within the repeat regions of some of the STR loci. Consistent with the literature, the loci D2S1338, D12S391, and D21S11 exhibit the largest contribution to increased diversity via sequence variation in the RR (Table 1) [1,6,7]. Among the four populations studied, there were 50 additional alleles detected in the D2S1338 locus (heterozygosity increase of 0.0371 ± 0.0178); 60 additional alleles found at the D12S391 locus (heterozygosity increase of 0.0699 ± 0.0134); and 63 additional alleles observed in the D21S11 locus (heterozygosity increase of 0.0697 ± 0.0220). Other loci, such as D8S1179 and D6S1043, also had a number of sequence variant alleles, and most variants observed were different from the currently published repeat motifs (Supplemental Table S2) [1,10,27,29,75].

Table 2
Summary of observed novel alleles.

(a) Loci with an increase in observed alleles with novel repeat region variation only.

Locus	# of Novel Repeat Region Variants	% Novel	Total # of Alleles
DYF387S1	27	62.79	43
DYS448	11	61.11	18
D17S1301	7	58.33	12
DYS635	10	52.63	19
DXS10103	12	52.17	23
D21S11	35	41.18	85
DYS612	7	41.18	17
D2S1338	24	37.50	64
D4S2408	2	25.00	8
FGA	7	20.59	34
DYS570	2	18.18	11
DYS439	1	16.67	6
D3S1358	3	13.64	22
PentaE	3	11.11	27
DXS7423	1	11.11	9
D10S1248	1	10.00	10
CSF1PO	1	9.09	11

(b) Loci with an increase in novel flanking region variation only.

Locus	# of Novel Flanking Variants	% Novel	Total # of Alleles
D22S1045	7	41.18	17
D5S818	7	38.89	18
DYS392	1	14.29	7
DYS460	1	14.29	7
TH01	1	10.00	10
DYS576	1	9.09	11

(c) Loci with an increase in novel repeat region and flanking region variation.

Locus	# of Novel Repeat Region Variants	# of Novel Flanking Variants	# of New Alleles	% Novel	Total # of Alleles
DXS10135[a]	79	16	91	85.05	107
DYS437	5	3	8	72.73	11
D20S482	5	8	13	72.22	18
DXS10074	23	10	33	71.74	46
D13S317	1	17	17	65.38	26
D7S820	4	13	17	65.38	26
DXS7132	4	5	9	56.25	16
D16S539	2	8	10	52.63	19
D6S1043	14	1	15	48.39	31
DYS481	6	3	9	47.37	19
D9S1122	7	1	8	47.06	17
DYS390	7	1	8	47.06	17
DXS8378	2	4	6	46.15	13
HPRTB	2	4	6	42.86	14
D18S51[a]	7	9	15	40.54	37
DYS522	1	2	3	37.50	8
D1S1656	11	1	12	35.29	34
D2S441	6	2	8	33.33	24
DYS438	1	2	3	33.33	9
DYS460pt2	1	2	3	33.33	9
D19S433	6	1	7	30.43	23
DYS385a-b[a]	6	1	6	28.57	21
vWA	6	3	9	25.71	35
PentaD	1	4	5	25.00	20
DYS533	1	1	2	25.00	8
D12S391	18	1	19	24.05	79
DYS389II	7	2	9	20.45	44
D8S1179	4	1	5	16.13	31

(d) Loci with no observed novel alleles

Locus	# of New Alleles	% Novel	Total # of Alleles
TPOX	0	0.00	9
DYS19	0	0.00	5
DYS389I	0	0.00	6
DYS391	0	0.00	4
DYS505	0	0.00	6
DYS549	0	0.00	6

[a] SNPs are present in both repeat region and flanking region in the nucleotide sequence.

Second, some loci demonstrated an effective increase in allele number only when FR sequence information was included. This category of variation includes STR loci in which the repeat regions of the alleles did not display sequence differences, but did show substantial variability (such as an increase of at least 30% of total number of alleles a metric used by Gettings et al. [1]) in the flanking regions surrounding the repeat regions of interest. For example, the loci D7S820, D13S317, and D22S1045 did not have many sequence variant alleles observed within their RRs; however within the FRs, all three loci exhibited at least a 40% increase (160%, 189%, and 55%, respectively) in the total number of alleles (Table 2a, b).

Third, several loci showed either RR variation or FR sequence variation in the SB alleles. The loci D18S51, DXS10135, and DYS385a-b are three examples in which there were alleles that contained both RR sequence and FR sequence variation (Table 1 and 2c).

Fourth, some loci such as DYS643, Y-GATA-H4 and TPOX did not display any effective increase in diversity using MPS beyond that which was observed by CE.

3.1. STR allele variation by sequencing

An increase in the number of effective alleles was observed due to variation within the RR only, the flanking sequence only, or whether the locus had alleles that had either RR or FR variants. Of the STR markers analyzed, 24 autosomal, 6 X-chromosome, and 14 Y-chromosome loci had an increase in effective alleles greater than 20% with SB alleles compared with nominal LB alleles (Fig. 1, Table 1). Only one autosomal locus (TPOX) and eight Y-STR loci (DYS19, DYS389I, DYS391, DYS439, DYS505, DYS549, DYS643, and Y-GATA-H4) did not have any effective increase in alleles using a SB analysis method. Supplemental Table S2 displays all known allele sequences for the STR loci studied herein and described in the literature [1–67] as well as highlights novel variants previously unreported for each locus in yellow.

While identifying the total sequence variation of each locus is the main goal of this study, novel alleles were described as well to demonstrate that substantial genetic variation remains to be identified. Overall, the sequence variation observed herein and those reported in the literature show similar trends [1,6,7,10,36,50] and indicate the diversity of sequence variation yet to be uncovered in larger datasets. Novel alleles substantially contribute to the number of SB alleles observed in this study (Fig. 1). Five autosomal (D17S1301, D21S11, D2S1338, D4S2408, and FGA) loci had 20% or more increase in total number of alleles due to novel RR alleles alone. The inclusion of novel FR sequence information identified another autosomal locus (D22S1045) that increased in total number of alleles by at least 20%. In addition, the combination of novel RR and FR sequence variation found 13 autosomal loci (D20S482, D13S317, D7S820, D16S539, D6S1043, D9S1122, D18S51, D1S1656, D2S441, D19S433, vWA, PentaD, and D12S391) that increased in total number of alleles by at least 20%. For the X-chromosome markers, one locus (DXS10103) had 20% or more

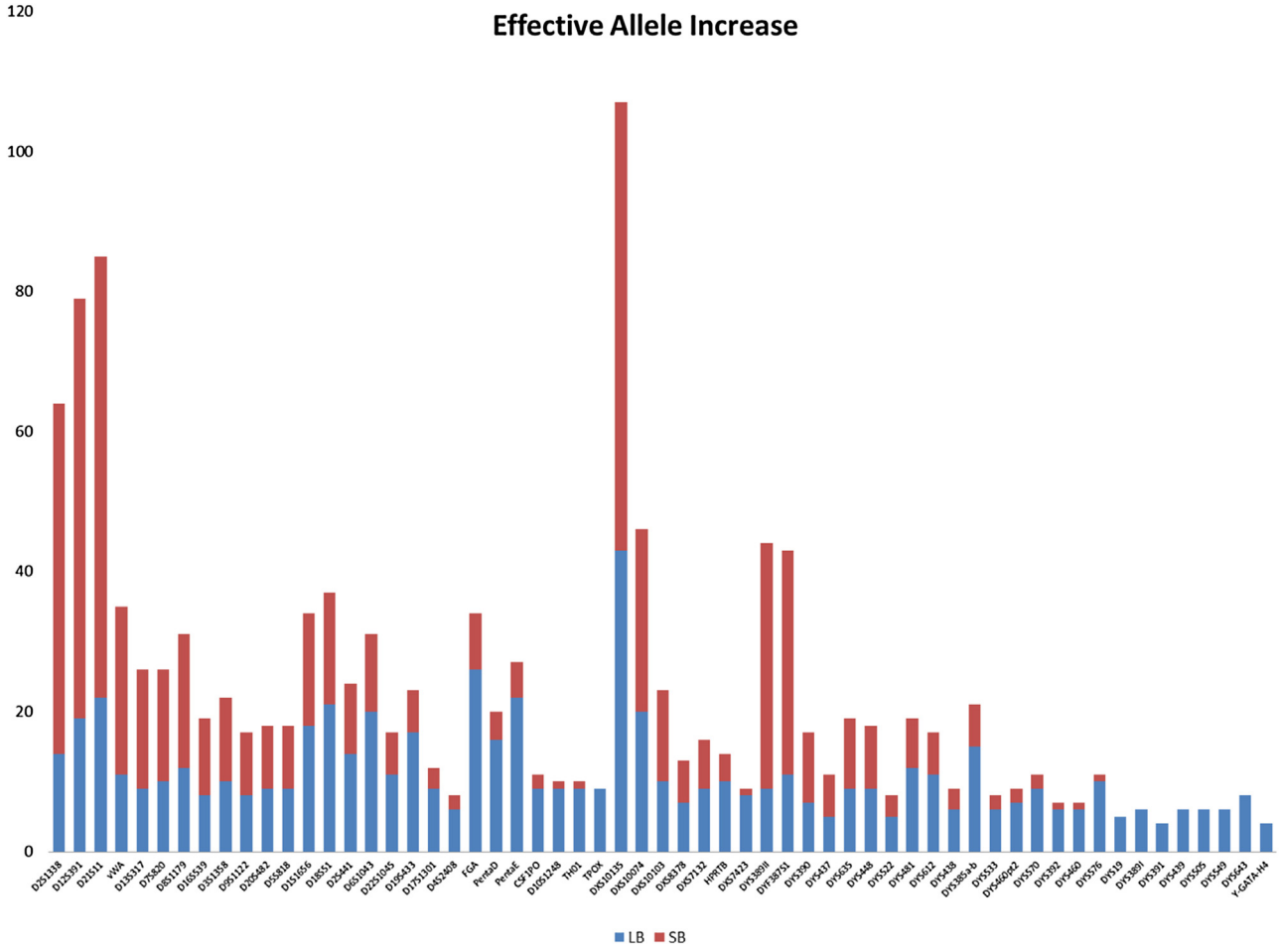

Fig. 1. The number of alleles using nominal length-based (LB; blue) as well as the effective increase in observed alleles using the sequence-based (SB; red) data from MPS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

increase in total number of alleles due to novel RR alleles alone, and five loci (DXS10135, DXS10074, DXS7132, DXS8378, and HPRTB) had 20% or more increase in total number of alleles because of novel RR and FR sequence variation. Finally, four Y-STR loci (DYF387S1, DYS448, DYS635, and DYS612) had 20% or more increase in total number of alleles due to novel RR alleles alone, and nine Y-STR loci (DYS437, DYS481, DYS390, DYS522, DYS438, DYS 461, DYS385a-b, DYS533, and DYS389II) had 20% or more increase in total number of alleles because of novel RR and FR sequence variation. Three loci, D18S51, DXS10135, and DYS385a-b had 2, 4, and 1 alleles, respectively, which had novel SB alleles with variants in both the flanking region and repeat region of each allele.

Most loci with novel RR and FR variation had good representation in all four populations contributing to the overall allelic diversity observed in the dataset (Table 1). However, some sequence variants were observed in specific populations and not observed in the other populations studied. For example, the frequency of the $[GGAA]_2$ GGAC $[GGAA]_{12}$ $[GGCA]_6$ allele 21 of the D2S1338 locus is 4% in the African American population sample, but was not observed in other populations in this study. Another example is the $[AGAT]_9$ allele 9 of the D17S1301 locus observed in the Chinese population, with an allele frequency of 3%. Interestingly, the loci DXS7423, DXS7132 and DXS8378 were notably polymorphic in the Hispanic population only, with little to no increase in diversity in the other populations studied.

Loci D9S1122, D4S2408, D20S482, D17S1301, DYS549, DYF387S1, and DXS10135 exhibited a large expansion of SB alleles observed in this study. This notable change in these loci likely is a result of limited sequence variation allele data in the literature [10,36], and only recent inclusion of such markers in forensic STR identification panels.

3.2. Sequence variation in the flanking region

The positioning of primers for the amplification of loci can impact detection of FR variation as well as successful amplification. A primer sitting on a variant site may not anneal well, and the chance of successful amplification of the particular allele may be reduced. However, instead of attempting to overcome the effects of primer binding site variants by use of degenerate primers, these FR polymorphisms can be exploited, when feasible, by MPS to increase the discrimination power of some STR loci. Therefore, knowledge of variants in the flanking region of STRs can be invaluable in future primer design to increase the diversity at a locus. FR SNPs were reported using GRCh38 reference genome coordinates (Supplemental Table S2).

In total, 30 loci demonstrated an increase in the number of alleles by inclusion of flanking sequence data residing within the amplicon sequence regions that were read. There were 145 additional alleles observed due to novel FR data that were seen at least once in one population distributed over 17 autosomal, 5 X-chromosome, and 12 Y-chromosome loci (Table 2b, c). For five loci (D22S1045, DYS392, DYS460, TH01, and DYS576 from highest to lowest increase per locus), an increase in the number of alleles was observed as a result of novel FR variation only (Table 2b). For example, the D22S1045 locus had seven novel FR sequences, increasing the total number of alleles for that locus from 10 to 17. Of note, the loci D13S317 and D7S820, which are comprised predominantly of FR sequence variants, had an effective increase in number of alleles by at least 50% per locus, with increases in overall heterozygosities of 0.0044 ± 0.008 and 0.0015 ± 0.0026, respectively. (Supplemental Tables S1–3)

The SNP in the flanking region of the D13S317 locus is important to understand as it resides at the first nucleotide position of the 3′ FR sequence. The repeat motif for the D13S317 locus is $[TATC]_n$, where the GRCh38 reference sequence 3′ flanking region begins with an 8 nucleotide "AATC AATC" pattern. However, multiple alleles exhibit an A/T variant (AATC to TATC; GRCh38 13: 82148069) (Supplemental Table S2), which may complicate the LB interpretation of the locus.

The FR sequence information described in this paper was limited to the regions designated by the primers utilized by the ForenSeq™ DNA Signature Prep Kit in the FGx system. It is possible, if not likely, that alternate primers may capture additional variation or not detect some of the variation observed in the study herein.

3.3. Interesting loci and allele designations

The DYS460 locus is unique among the loci in this data set in that there is an upstream $[TCTG]_n$ $[TCTA]_n$ repeat region (GRCh38 Y: 18,888,804-18,888,851) followed by a string of 104 nucleotides that is captured within the amplicon of the downstream RR (Supplemental Table S2). In this study, the DYS460 locus was characterized using the $[CTAT]_n$ repeat region only, beginning at GRCh38 Y: 18,888,956, consistent with the recommendations of Parson et al. [8]. For reference, the upstream motif has been included in the data analyses and is designated as DYS461 [47]. It also is important to mention that the incorrect RR motif and nucleotide sequence were reported originally by Parson et al. [8] and is not congruent with GRCh38, which currently is being addressed by Parson et al. (personal communication). As such, the correct DYS460 repeat region was aligned to the GRCh38 reference sequence using the IGV software [76] and was changed from $[TATC]_n$ to $[CTAT]_n$. The downstream flanking region commences with a G nucleotide at GRCh38 Y: 18,888,996, and if the [TATC] motif was used, each allele call would be designated as one less repeat than what truly exists in the nucleotide sequence. This example was the only discordance observed between Parson et al. [8] and the STRait Razor output generated in this analysis.

The sequence immediately proximal to the repeat region of the DYS612 locus may lead to different designations for the number of repeats comprising alleles. The motif is defined as $[CCT]_a$ $[CTT]_b$ $[TCT]_c$ $[CCT]_d$ $[TCT]_e$ [1,8,9,41,42,59]. For example, STRBase [77] lists the RR motif $[CCT]_5[CTT][TCT]_4[CCT][TCT]_{25}$ as a LB allele 36, while Parson et al. [8] recommend that the LB allele should be a 30 by not including the initial repeats $[CCT]_a$ $[CTT]_b$ [8,9,41,42].

The vWA locus is an example where Parson et al. [8] report the allele based on the forward strand whereas Gettings et al. [1] rely on the reverse strand. In Gettings et al. [1], as well as the ForenSeq™ UAS, vWA repeat sequences are designated $[TCTA]_a$ $[TCTG]_b$ $[TCTA]_c$ in which a TCCA TCTA flanking tail is not included in the repeat number designation. However, using the forward strand as per Parson et al. [8], the motif is characterized as TAGA TGGA $[TAGA]_a$ $[CAGA]_b$ $[TAGA]_c$, in which the TAGA TGGA lead sequence is not included in the repeat count. As described previously [14,15,20,21,35], there are polymorphisms in this lead/tail sequence impacting variation in allelic designations for the vWA locus (Supplemental Table S2).

3.4. Concordance analysis

Concordance refers to obtaining the same allele calls by both methods. The size of the repeat units (i.e., tri-, tetra-, penta- and hexa-nucleotides) and length of SB allele calls generally were consistent with CE-based data (Yfiler® and Globalfiler® STR kits). While not a true concordance issue, allele dropout does characterize performance. There were 12 instances of dropout for the DYS392 locus with the STRait Razor MPS data. At this time, the cause for the allele dropout is undetermined. Allele dropout or extremely imbalanced heterozygotes were observed between MPS and CE-based analysis methods for the D22S1045 locus, where

MPS was unable to detect larger alleles of a heterozygote (allele 17 or larger) in two samples (Supplemental Table S3). It should be noted that these two loci share the motif [ATT/ATA]. A similar and reproducible pattern of allele performance was observed by J.D. Churchill (personal communication; manuscript in preparation) within and between these loci.

Although not common, there were examples of MPS data that rely solely on RR variation that were discordant with operationally defined CE-based data due to indels residing in a flanking region or in instances where SNPs reside in the flanking regions immediately proximal to the repeat regions [36,78]. In total, there were only 2 instances out of 7980 comparisons (170 samples x 20 loci x 2 alleles, assuming 2 alleles for a homozygote plus 59 samples x 18 Y-STR loci x 1 allele, assuming hemizygous loci and 59 samples X 1 Y-STR locus X 2 alleles (DYS385)) where a "potential mismatch" in allele designation occurred (Supplementary Table 3). The first was an allele with repeat $[TATC]_{11}$ at the D7S820 locus where an upstream single nucleotide deletion caused the allele designation by CE to be a 10.3. If MPS analysis only targets the repeat region, this allele would be called an 11. The other example was observed at the D22S1045 locus that was designated as a 15.1 allele by CE while the true RR number is a 15. Although not part of the subsample used for concordance testing, there were another 23 alleles across all loci in the 777 individuals typed where a similar situation may occur. That is, there were indels in the flanking region that likely would affect the length of the alleles that would be designated by CE and yet would be different than solely the repeat regions detected by MPS. The alleles were observed at three autosomal (D7S820, D2S441, Penta D), two Y-STR (DYS460, DYS385) loci, and three X-STR (HPRTB, DXS10135 and DXS10074) loci. Capture of sequence data within amplicon sequence regions that were read allows for identification of the repeat number differences between MPS and CE generated data for the two instances and possibly information to readily correct MPS data and operationally defined CE data to reduce the, albeit rare, number of "apparent" discrepancies. Thus, when evaluated in total, there were no discrepancies between MPS- and CE-generated allele calls. By using the length of most, or in some instances, the entire amplicon (GRChr38 coordinates provided in Supplemental Table S2), additional sequence variation was captured that demonstrated that these few differences could be explained. By developing approaches that include the flanking regions of STRs, discordance can be reduced with legacy LB data as well as increasing discrimination power of the assays.

3.5. Population genetic analyses

SB and LB allele counts and allele frequencies were calculated for each population and each locus and are listed in Supplemental Table S1. A comprehensive characterization of total sequence variation for each locus including alignments to GRCh38 can be found in Supplemental Table S2. Observed and expected heterozygosities were determined for both LB and SB alleles and summarized by population for the autosomal and X-STRs (females only) in Supplemental Table S3. All loci were highly polymorphic as expected. In most instances, the heterozygosity increased at those loci with underlying sequence variation and is described by population in Supplemental Table S3. Tests for HWE were performed separately for SB and LB alleles for 34 loci in each of the four sample populations. For autosomal loci, the LB allele data showed detectable departures from HWE expectations ($p < 0.05$) at two STRs (D21S11 and Penta D) in the African American and three STRs (CSF1PO, D19S433, and D7S820) in the Chinese population groups. After Bonferroni correction, no loci significantly deviated from HWE expectations (Supplemental Table S3). With LB data, all X-chromosome loci (females only) met HWE expectations (Supplemental Table S3). There were three STRs (D13S317, D5S818, and D7S820) in the US Caucasian, six STRs (D13S317, D16S539, D20S482, D2S441, D5S818, and Penta D) in the African American, four STRs (CSF1PO, D16S539, D19S433 and D7S820) in the Chinese, and six STRs (D13S317, D16S539, D20S482, D4S2408, D5S818, and D7S820) in the Hispanic population that demonstrated departures from HWE expectations using SB data (i.e., $p < 0.05$). After Bonferroni correction, significant departures from HWE were observed for three STRs (D13S317, D5S818, and D7S820) in US Caucasian, two STRs (D13S317 and D16S539) in African American, one STR (D7S820) in the Chinese, and three STRs (D16S539, D20S482, and D7S820) in Hispanic population groups. The observed number of significant deviations from HWE for SB allele determinations is consistent with what is expected by chance alone ($\sim$2) but was greater than what was observed with LB data analyses. These significant departures from HWE may be due to chance, population substructure, or there may be an effect due to the different types of variation captured within the STRs. For example, the D5S818, D20S482, D16S539, D7S820, and D13S317 loci have a number of alleles due to flanking variants. The combination of FR SNPs and RR variation (i.e., slippage generated variants) may affect current HWE calculations. Future work will attempt to determine if these departures are potentially an artifact of the novel variation being measured.

Linkage disequilibrium was assessed using default parameters with the "Preserve Genotypes" setting for shuffling method in GDA to prevent within-locus disequilibrium (observed during HWE analysis) from affecting the significance of LD for the 27 autosomal and 7 X-STRs for each population group. For length based autosomal data (351 pairwise comparisons), 23 pairwise comparisons in Hispanic, 15 pairwise comparisons in US Caucasian, 20 pairwise comparisons in Chinese, and 16 pairwise comparisons in African American population groups demonstrated significant LD ($p < 0.05$). Of the 74 pairwise comparisons with detectable LD, three pairwise comparisons occurred between syntenic loci (CSF1PO and D5S818 in the African American population; D4S2408 and FGA in the Chinese population; D2S441 and TPOX in the Hispanic population). After Bonferroni correction, LD was detected for one pairwise comparison (D18S51 and D9S1122) in the Chinese population (Supplemental Table S3). With sequence based data, 23 pairwise comparisons in Hispanic, 16 pairwise comparisons in US Caucasian, 19 pairwise comparisons in Chinese, and 16 pairwise comparisons in African American population groups demonstrated significant LD ($p < 0.05$). Of the 74 pairwise comparisons with detectable LD, five pairwise comparisons occurred between syntenic loci (CSF1PO and D5S818 in the African American population; D2S1338 and TPOX, D4S2408 and FGA, D19S433 and Penta D in the Chinese population; D2S441 and TPOX in the Hispanic population). After Bonferroni correction there was no observed significant LD.

Four hundred and ninety-nine female samples were analyzed separately for LD with the combined autosomal and X-STR data (i.e., 34 total loci and 561 pairwise comparisons). Twenty-two pairwise comparisons for African American, 27 pairwise comparisons for Chinese, 27 pairwise comparisons for Caucasian, and 37 pairwise comparisons for the Hispanic population groups indicated LD for LB analysis. Twenty-four comparisons for African American, 21 pairwise comparisons for Chinese, 23 pairwise comparisons for Caucasian, and 36 pairwise comparisons for the Hispanic population groups indicated LD for SB analysis methods. After Bonferroni correction, only one pairwise comparison (TH01 and vWA) for African American and one pairwise comparison (DXS10103 and HPRTB) for Hispanic population groups showed significant LD for SB analysis, and only one pairwise comparison

(DXS10103 and HPRTB) for the Hispanic population was observed for LB analyses. Szibor [79] and Sim et al. [56] reported that DXS10103 and HPRTB are members of an X-chromosomal linkage group, and therefore, the observation of LD with these two loci may be expected.

RMPs were calculated for autosomal markers by population and are included in Supplemental Table S1. Overall, the combined RMPs were lower by at least two orders of magnitude for every population using a SB approach, highlighting the value of SB allele designations for statistical purposes. RMPs for female X-chromosome markers for each population are reported in Supplemental Table S3.

Y-Chromosome haplotypes are listed in Supplemental Table S3. Haplotype diversity calculations revealed no shared haplotypes within and across populations, using either a LB or SB approach. The haplotype diversities were the same for LB and SB allelic data and ranged from 0.9524-0.9912. These estimates do not represent the true diversity as they were calculated with very limited sample sizes. Y-haplogroup assignment was determined using World Haplogroup Predictor for each population by assuming equal prior ancestry odds (Supplemental Table S3) [80]. The predominant haplogroups associated with the haplotypes in each population were E3a (22 of 40 samples) for African American haplotypes, R1b (63 of 105 samples) for Caucasian haplotypes, R1b and Q (41 and 36 of 113 samples, respectively) for Chinese haplotypes, and R1b (10 of 21 samples) for Hispanic haplotypes.

4. Conclusions and future directions

The population data described in this study demonstrate that there is variation and substantial novel variation within RR and/or FR of a number of STR markers, whereas a few loci present little to no additional discrimination power using MPS. While the current forensically relevant STR loci were not selected based on total genetic variation, moving forward it may be worthwhile to consider inclusion of STR loci that offer additional discrimination power in the form of RR and/or FR sequence variation. In particular, consideration of loci that display similar or greater heterozygosity than existing selected STRs are particularly beneficial for mixture de-convolution. If the range of allele sizes is less for a locus with sequence variation than a comparably informative locus whose polymorphism is solely due to length variation, the former marker may be more robust during PCR. With less size difference between the alleles of a heterozygote profile, better heterozygote balance, and less amplification stochastic effects may be achieved.

Current PCR-CE genotyping methods are able to address the major portion of the current needs of the forensic community. However, higher throughput, increased resolution, and better mixture de-convolution of complex biological samples are still needed. While the data described herein are sufficient to use for calculating the strength of STR results from casework evidence, the data are still only a small representation of the variation that likely exists for these loci. Sequence data currently does not exist in DNA databases, therefore sequence-based DNA profiles will likely for the time being be used for 1:1 comparisons of suspect or victim to evidence. Continued efforts to establish population-based databases of these markers are essential for a greater understanding of STR diversity [81].

Acknowledgements

This work was supported in part by award no. 2015-DN-BX-K067, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2016.09.007.

References

[1] K.B. Gettings, R.A. Aponte, P.A. Vallone, J.M. Butler, STR allele sequence variation: current knowledge and future issues, Forensic Sci. Int. Genet. 18 (2015) 118–130.
[2] C. Børsting, N. Morling, Next generation Sequencing and its applications in forensic genetics, Forensic Sci. Int. Genet. 18 (2015) 78–89.
[3] S. Dalsgaard, E. Rockenbauer, C. Gelardi, C. Børsting, S.L. Fordyce, N. Morling, Characterization of mutations and sequence variations in complex STR loci by second generation sequencing, Forensic Sci. Int. Genet. Suppl. Ser. 4 (2013) e218–e219.
[4] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, D. Deforce Forensic, STR analysis using massively parallel sequencing, Forensic Sci. Int. Genet. 6 (2012) 810–818.
[5] D.W. Craig, J.V. Pearson, S. Szelinger, A. Sekar, M. Redman, J.J. Corneveaux, T.L. Pawlowski, T. Laub, G. Nunn, D.A. Stephan, N. Homer, M.J. Huentelman, Identification of genetic variants using barcoded multiplex sequencing, Nat. Methods (2008) 887–893.
[6] S.L. Friis, A. Buchard, E. Rockenbauer, C. Børsting, N. Morling, Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs, Forensic Sci. Int. Genet. 21 (2016) 68–75.
[7] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, R.A. Guerrieri, P.M. Vallone, Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Forensic Sci. Int. Genet. 21 (2016) 15–21.
[8] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmão, D. R. Hares, J.A. Irwin, J.L. King, K.P. de Knijff, N. Morling, M. Prinz, P.M. Schneider, C. Van Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the international society for forensic genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63.
[9] M. Klintschar, Z. Kozma, N. Al Hammadi, M. Abdull Fatah, C. Nöhammer, A study on the short tandem repeat systems HumCD4, HumTH01 and HumFIBRA in population samples from Yemen and Egypt, Int. J. Legal Med. 111 (1998) 107–109.
[10] F.R. Wendt, J.D. Churchill, N.M.M. Novroski, J.L. King, J. Ng, R.F. Oldt, K.L. McCulloh, J.A. Weise, D.G. Smith, S. Kanthaswamy, B. Budowle, Genetic analysis of the yavapai native americans from West-central Arizona using the illumina MiSeq FGx™ forensic genomics system, Forensic Sci. Int. Genet. 24 (2016) 18–23.
[11] K.J. van der Gaag, R.H. de Leeuw, J. Hoogenboom, J. Patel, D.R. Storts, J.F.J. Laros, P. de Knijff, Massively parallel sequencing of short tandem repeats —Population data and mixture analysis results for the PowerSeq™ system, forensic sci, Int. Genet. (2016) (in press).
[12] M.C. Kline, C.R. Hill, A.E. Decker, J.M. Butler, STR Sequence analysis for characterizing normal variant, and null alleles, Forensic Sci. Int. Genet. 5 (2011) 329–332.
[13] C. Allor, D.D. Einum, M. Scarpetta, Identification and characterization of variant alleles at CODIS STR loci, J.Forensic Sci. 50 (2005) 1128–1133.
[14] A. Moller, E. Meyer, B. Brinkmann, Different types of structural variation in STRs: humFES/FPS, HumVWA, and HumD21S11, Int. J. Legal Med. 106 (1994) 319–323.
[15] B. Brinkmann, A. Sajantila, H.W. Goedde, H. Matsumoto, K. Nishi, P. Wiegand, Population genetic comparisons among eight populations using allele frequency and sequence data from three microsatellite loci, Eur. J. Hum. Genet. 4 (1996) 175–182.
[16] M.D. Barber, B.J. McKeown, B.H. Parkin, Structural variation in the alleles of a short tandem repeat system at the human alpha fibrinogen locus, Int. J. Legal Med. 108 (1996) 180–185.
[17] B. Brinkmann, E. Meyer, A. Junge, Complex mutational events at the HumD21S11 locus, Hum. Genet. 98 (1996) 60–64.
[18] A.M. Lins, K.A. Micka, C.J. Sprecher, J.A. Taylor, J.W. Bacher, D. Rabbach, R.A. Bever, S. Creacy, J.W. Schumm, Development and population study of an eight-locus short tandem repeat (STR) multiplex system, J. Forensic Sci. 43 (1998) 1168–1180.
[19] R.A.L. Griffiths, M.D. Barber, P.E. Johnson, S.M. Gillbard, M.D. Haywood, C.D. Smith, J. Arnold, T. Burke, A. Urquhart, P. Gill, New reference allelic ladders to improve allelic designation in a multiplex STR system, Int. J. Legal Med. 111 (1998) 267–272.
[20] A. Kido, M. Hara, Y. Yamamoto, H. Kameyama, R. Susukida, K. Saito, A. Takada, M. Oya, Nine short tandem repeat loci analysis in aged semen stains using the AmpFLSTR Profiler Kit and description of a new vWA variant allele, Leg. Med. (Tokyo) 5 (2003) 93–96.

[21] C. Cruz, T. Ribeiro, C. Vieira-Silva, I. Lucas, R. Espinheira, H. Geada, vWA STR locus structure and variability, Int. Congr. Ser. 1261 (2004) 248–250.
[22] E.M. Dauber, W. Bär, M. Klintschar, F. Neuhuber, W. Parson, W.R. Mayr, New sequence data of allelic variants at the STR loci ACTBP2 (SE33), D21S11, FGA, vWA, CSF1PO, D2S1338, D16S539, D18S51 and D19S433 in caucasoids, Int. Congr. Ser. 1261 (2004) 191–193.
[23] P. Grubwieser, R. Muhlmann, H. Niederstatter, M. Pavlic, W. Parson, Unusual variant alleles in commonly used short tandem repeat loci, Int. J. Legal Med. 119 (2005) 164–166.
[24] M. Heinrich, H. Felske-Zech, B. Brinkmann, C. Hohoff, Characterisation of variant alleles in the STR systems D2S1338, D3S1358 and D19S433, Int. J. Legal Med. 119 (2005) 310–313.
[25] S. Hering, R. Nixdorf, J. Edelmann, C. Thiede, J. Dreβler, Further sequence data of allelic variants at the STR locus ACTBP2 (SE33): Detection of a very short off ladder allele, Int.Congr. Ser. 1288 (2006) 810–812.
[26] E.M. Dauber, G. Dorner, S. Wenda, E.M. Schwartz-Jungl, B. Glock, W. Bär, W.R. Mayr, Unusual FGA and D19S433 off-ladder alleles and other allelic variants at the STR loci D8S1132, vWA D18S51 and ACTBP2 (SE33), Forensic Sci. Int. Genet. Suppl. Ser. 1 (2008) 109–111.
[27] E.M. Dauber, E.M. Schwartz-Jungl, S. Wenda, G. Dorner, B. Glock, W.R. Mayr, Further allelic variation at the STR-loci ACTBP2 (SE33), D3S1358 D8S1132, D18S51 and D21S11, Forensic Sci. Int. Genet. Suppl. Ser. 2 (2009) 41–42.
[28] C. Phillips, L. Fernandez-Formoso, M. Garcia-Magarinos, L. Porras, T. Tvedebrink, J. Amigo, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, A. Freire-Aradas, A. Gomez-Carballa, A. Mosquera-Miguel, A. Carracedo, M.V. Lareu, Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel, Forensic Sci. Int. Genet. 5 (2011) 155–169.
[29] C. Phillips, S. Kind, L. Fernandez-Formoso, M. Gelabert-Besada, A. Carracedo, M.V. Lareu, Global population variability in promega PowerPlex CS7, D6S1043, and penta B STRs, Int. J. Legal Med. 127 (2013) 901–906.
[30] M.V. Lareu, S. Barral, A. Salas, C. Pestoni, A. Carracedo, Sequence variation of a hypervariable short tandem repeat at the D1S1656 locus, Int. J. Legal. Med. 111 (1998) 244–247.
[31] A. Morales-Valverde, S. Silva-De La Fuente, G. Nuñez-Rivas, M. Espinoza-Esquivel, Characterisation of 12 new alleles in the STR system D18S51, Forensic Sci. Int. Genet. (SS2) (2009) 43–44.
[32] M.V. Lareu, C. Pestoni, F. Barros, A. Salas, A. Carracedo, Sequence variation of a hypervariable short tandem repeat at the D12S391 locus, Gene 182 (1996) 151–153.
[33] J.A. Bright, K.E. Stevenson, M.D. Coble, C.R. Hill, J.M. Curran, J.S. Buckleton, Characterising the STR locus D6S1043 and examination of its effect on stutter rates, Forensic Sci. Int. Genet. 8 (2014) 20–23.
[34] C. Gelardi, E. Rockenbauer, S. Dalsgaard, C. Borsting, N. Morling, Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles, Forensic Sci. Int. Genet. 12 (2014) 38–41.
[35] L. Wang, X.C. Zhao, J. Ye, J.J. Liu, T. Chen, X. Bai, J. Zhang, Y. Ou, L. Hu, B.W. Jiang, F. Wang, Construction of a library of cloned short tandem repeat (STR) alleles as universal templates for allelic ladder preparation, Forensic Sci. Int. Genet. 12 (2014) 136–143.
[36] J.D. Churchill, S.E. Schmedes, J.L. King, B. Budowle, Evaluation of the Illumina® beta version ForenSeq™ DNA signature prep kit for use in genetic profiling, Forensic Sci. Int. Genet. 20 (2016) 20–29.
[37] W. Wang, T. Kishida, M. Fukuda, Y. Tamaki, The Y-27H39 polymorphism in a Japanese population, Int. J. Legal Med. 109 (1996) 157–158.
[38] Reference tables to: evaluation of Y-chromosomal STRs: a multicenter study (Kayser et al.) and Chromosome Y microsatellites: population genetic and evolutionary aspects (de Knijff et al.), Int. J. Legal Med. 110 (1997) 141–149.
[39] T. Kumoro, H. Tsutsumi, R. Mukoyama, M. Nakamura, Repeat structure of DYS389 locus, Nippon Hoigaku Zasshi, (Jpn. J. Legal Med.) 52 (4) (1998) 227–232 (Article in Japanese).
[40] Q. Ayub, A. Mohyuddin, R. Qamar, K. Mazhar, T. Zerjal, S.Q. Mehdi, C. Tyler-Smith, Identification and characterization of novel human Y-chromosomal microsatellites from sequence database information, Nucleic Acids Res. 28 (2) (2000) e8.
[41] J.M. Butler, R. Schoske, P.M. Vallone, M.C. Kline, A.J. Redd, M.F. Hammer, A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers, Forensic Sci. Int. 129 (2002) 10–24.
[42] A.J. Redd, A.B. Agellon, V.A. Kearney, V.A. Contreras, T. Karafet, H. Park, P. de Knijff, J.M. Butler, M.F. Hammer, Forensic value of 14 novel STRs on the human Y chromosome, Forensic Sci. Int. 3460 (2002) 1–15.
[43] R. Schoske, P.M. Vallone, M.C. Kline, J.W. Redman, J.M. Butler, High-throughput Y-STR typing of U.S. populations with 27 regions of the Y chromosome using two multiplex PCR assays, Forensic Sci. Int. 139 (2004) 107–121.
[44] J.M. Butler, A.E. Decker, P.M. Vallone, M.C. Kline, Allele frequencies for 27 Y-STR loci with U.S. Caucasian African American, and Hispanic samples, Forensic Sci. Int. 156 (2006) 250–260.
[45] T. Komuro, H. Tsutsumi, R. Mukoyama, M. Nakamura, Repeat structure of DYS389 locus, Nippon Hoigaku Zasshi (Jpn. J. Legal Med.) 52 (1998) 227–232.
[46] M.E. D'Amato, L. Ehrenreich, K. Cloete, M. Bejeddou, S. Davison, Characterization of the highly discriminatory loci DYS449, DYS481 DYS518, DYS612, DYS626, DYS644 and DYS710, Forensic Sci. Int. Genet. 4 (2010) 104–110.
[47] E. Bosch, A.C. Lee, F. Calafell, E. Arroyo, P. Henneman, P. de Knijff, M.A. Jobling, High resolution Y chromosome typing: 19 STRs amplified in three multiplex reactions, Forensic Sci. Int. 125 (2002) 42–51.
[48] P.S. White, O.L. Tatum, L.L. Deaven, J.L. Longmire, New, male-Specific microsatellite markers from the human Y chromosome, Genomics 57 (1999) 433–437.
[49] P.M. Schneider, S. Meuser, W. Waiyawuth, Y. Seo, C. Rittner, Tandem repeat structure of the duplicated Y-chromosomal STR locus DYS385 and frequency studies in the German and three Asian populations, Forensic Sci.Int. 97 (1998) 61–70.
[50] D.H. Warshauer, J.D. Churchill, N. Novroski, J.L. King, B. Budowle, Novel Y-chromsome short tandem repeat variants detected through the use of massively parallel sequencing, Genom. Proteom. Bioinf. 13 (2015) 250–257.
[51] J. Edelmann, S. Hering, M. Michael, R. Lessig, D. Deichsel, G. Meier-Sundhausen, L. Roewer, I. Plate, R. Szibor, 16 X-chromosome STR loci frequency data from a German population, Forensic Sci. Int. 124 (2001) 215–218.
[52] J. Edelmann, D. Deichsel, S. Hering, I. Plate, R. Szibor, Sequence variation and allele nomenclature for the X-linked STRs DXS9895, DXS8378 DXS7132, DXS6800, DXS7133, GATA172D05, DXS7423 and DXS8377, Forensic Sci. Int. 129 (2002) 99–103.
[53] M.T. Zarrabeitia, T. Amigo, C. Sañudo, M.M. de Pancorbo, J.A. Riancho, Sequence structure and population data of two X-linked markers: DXS7423 and DXS8377, Int. J. Legal Med. 116 (2002) 368–371.
[54] S. Hering, C. Augustin, J. Edelmann, M. Heidel, J. Dressler, H. Rodig, E. Kuhlisch, R. Szibor, DXS10079, DXS10074 and DXS10075 are STRs located within a 280-kb region of Xq12 and provide stable haplotypes useful for complex kinship cases, Int. J. Legal Med. 120 (2006) 337–345.
[55] I. Gomes, R. Pereira, W.R. Mayr, A. Amorim, A. Carracedo, L. Gusmão, Evaluation of DXS9902, DXS7132, DXS6809, DXS7133, and DXS7423 in humans and chimpanzees: sequence variation, repeat structure, and nomenclature, Int. J. Legal Med. 123 (2009) 403–412.
[56] J.E. Sim, H.Y. Lee, W.I. Yang, K.-J. Shin, Population genetic study of four closely-linked X-STR trios in Koreans, Mol. Biol. Rep. 37 (2010) 333–337.
[57] T.M. Diegoli, M.D. Coble, Development and characterization of two mini-X chromosomal short tandem repeat multiplexes, Forensic Sci. Int. Genet. 5 (2011) 415–421.
[58] I. Gomes, A. Brehm, L. Gusmão, P.M. Schneider, New sequence variants detected at DXS10148, DXS10074 and DXS10134 loci, Forensic Sci. Int. Genet. 20 (2016) 112–116.
[59] J.V. Planz, K.A. Sannes-Lowery, D.D. Duncan, S. Manalili, B. Budowle, R. Chakraborty, S.A. Hofstadler, T.A. Hall, Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry, Forensic Sci. Int. Genet. 6 (2012) 594–606.
[60] M.D. Barber, B.H. Parkin, Sequence analysis and allelic designation of the two short tandem repeat loci D18S51 and D8S117, Int. J. Legal Med. 109 (1996) 62–65.
[61] P. Gill, C.P. Kimpton, A. Urquhart, N.J. Oldroyd, E.S. Millican, S.K. Watson, T.J. Downes, Automated short tandem repeat (STR) analysis in forensic casework–a strategy for the future, Electrophoresis 16 (1995) 1543–1552.
[62] A. Amorim, L. Gusmao, M.J. Prata, Population and formal genetics of the STRs TPO: TH01 and VWFA31/A in north Portugal, Adv. For. Haemogenet. 6 (1996) 486–488.
[63] E. Momhinweg, C. Luckenbach, R. Fimmers, H. Ritter, D3S1358. Sequence analysis and gene frequency in a German population, Forensic Sci. Int. 95 (1998) 173–178.
[64] R. Szibor, S. Lautsch, I. Plate, K. Bender, D. Krause, Population genetic data of the STR HumD3S1358 in two regions of Germany, Int. J. Legal. Med. 111 (1998) 160–161.
[65] H.-G. Zhou, K. Sato, Y. Nishimaki, L. Fang, H. Hasekura, The HumD21S11 system of short tandem repeat DNA polymorphisms in Japanese and Chinese, Forensic Sci. Int. 86 (1997) 109–118.
[66] S.J. Walsh, S.L. Robinson, G.R. Turbett, M.P. Davies, A.N. Wilton, Characterisation of variant alleles at the HumD21S11locus implies unique Australasian genotypes and re-classification of nomenclature guidelines, Forensic Sci. Int. 135 (2003) 35–41.
[67] K.K. Kidd, A.J. Pakstis, W.C. Speed, R. Legacé, J. Chang, S. Wootton, E. Haigh, J.R. Kidd, Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics, Forensic Sci. Int. Genet. 12 (2014) 215–224.
[68] D. Becker, H. Rodig, C. Augustin, J. Edelmann, F. Götz, S. Hering, R. Szibor, W. Brabetz, Population genetic evaluation of eight X-chromosomal short tandem repeat loci using Mentype Argus X-8 PCR amplification kit, Forensic Sci. Int. Genet. 2 (2008) 69–74.
[69] Qiagen, QIAamp® DNA Mini and Blood Mini Handbook, June 2012.
[70] Thermo Fisher Scientific Qubit® 2.0 Fluorimeter User Manual 2010.
[71] Illumina® ForenSeq™ DNA Signature Prep Reference Guide, 2014.
[72] Illumina® ForenSeq™ Universal Analysis Software Guide, 2015.
[73] D.H. Warshauer, J.L. King, B. Budowle, STRait razor v2.0: the improved STR allele identification tool —Razor, Forensic Sci. Int. Genet. 14 (2015) 182–186.
[74] Genetic Data Analysis Software, (1996) . http://en.bio-soft.net/dna/gda.html.
[75] X. Zeng, J.L. King, M. Stoljarova, D.H. Warshauer, B.L. LaRue, A. Sajantila, et al., High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing, Forensic Sci. Int. Genet. 16 (2015) 38–47.
[76] Integrative Genomics Viewer (IGV) —Broad Institute, (2013) . https://www.broadinstitute.org/igv/.

[77] National Institute of Justice STRBase, www.cstl.nist.gov/strbase/.
[78] F.R. Wendt, X. Zeng, J.D. Churchill, J.L. King, B. Budowle, Analysis of short tandem repeat and single nucleotide polymorphism loci from single-source samples using a custom HaloPlex target enrichment system panel, Am. J. Forensic Med. Pathol. 37 (2016) 99–107.
[79] R. Szibor, S. X-chromosomal markers: past, present and future, Forensic Sci. Int. Genet. 1 (2007) 93–99.
[80] J. Cullen, K. Nordtvedt, World Haplogroup & Haplo-I Subclade Predictor, (2008) . http://members.bex.net/jtcullen515/haplotest.htm.
[81] 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.