Forensic Science International: Genetics 65 (2023) 102872

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Sequence-based allelic variations and frequencies for 22 autosomal STR loci in the Lebanese population

Sarah Riman [a,*], Mirna Ghemrawi [b], Lisa A. Borsuk [a], Rami Mahfouz [c], Susan Walsh [d], Peter M. Vallone [a]

[a] Applied Genetics Group, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[b] Department of Chemistry and Biochemistry and International Forensic Research Institute, Florida International University, Miami, FL 33199, USA
[c] Department of Pathology and Laboratory Medicine, American University of Beirut Medical Center, Beirut, Lebanon
[d] Department of Biology, Indiana University Purdue University Indianapolis, Indianapolis, IN 46202, USA

ARTICLE INFO

Keywords:
Lebanon
Lebanese population
Next generation sequencing
Autosomal STR loci
Sequence variations
Population *STRUCTURE*
PowerSeq 46GY

ABSTRACT

This is the first study that characterizes the sequence-based allelic variations of 22 autosomal Short Tandem Repeat (aSTR) loci in a population dataset collected from Lebanon. Genomic DNA extracts from 195 unrelated Lebanese individuals were amplified with PowerSeq 46GY System Prototype. Targeted amplicons were subjected to DNA library preparation and sequenced on the Verogen MiSeq FGx Sequencing System. Raw FASTQ data files were processed by STRait Razor v3. Sequence strings were annotated according to the considerations of the DNA Commission of the International Society for Forensic Genetics (ISFG) and tabulated herein with their respective allelic frequencies and GeneBank accession and version numbers. The sequenced Lebanese dataset resulted in 429 distinct allelic sequences as compared to the 236 alleles identified by length only. The increase in the number of alleles was observed at 18 out of 22 aSTR loci and was attributed to the sequence variations residing in both the STR repeat motifs and flanking regions. The study uncovered 25 novel aSTR allelic sequences across 12 loci for which GenBank records did not previously exist in the STRSeq BioProject, PRJNA380127. For a concordance check, the length-based allelic calls derived from the full sequences were compared to those genotyped using capillary electrophoresis (CE) methods. Population genetic parameters relevant to the evaluation of forensic DNA evidence were assessed for the sequence-based data and compared to the parameters generated from the length-based information. Using the sequence-based data, Analysis of MOlecular VAriance (AMOVA), genetic distances, and population genetic structure were evaluated for 1231 individuals sampled from the Lebanese and four U.S. populations (African American, Asian, Caucasian, and Hispanic). The results were tabulated and visualized in a population tree, multidimensional scaling scatter plots, and bar plots. This newly established sequence-based database for the Lebanese population can be beneficial for extending NGS applicability to casework or paternity testing and assessing the strength of evidence for NGS-STR profiles. The described novel sequence variants at certain loci can further help in the effort to characterize the sequence diversity of STR markers from different populations around the world.

1. Introduction

Short Tandem Repeats (STRs) are considered the most commonly used markers for individual identification, criminal casework interpretation, and kinship analysis [1]. National DNA databases are based on STR profiles and allelic calls generated via fluorescence-based capillary electrophoresis (CE) technology [2]. Typically, with CE-based methods, STR regions are amplified by polymerase chain reaction (PCR) using different dye-tagged primers. The fluorescently labeled STR amplicons are then separated on polyacrylamide gel-filled capillaries and detected based on their molecular weights/sizes [1,3]. Ultimately, this PCR-CE workflow identifies STRs only by allelic length variations present in each dye channel [4].

In recent years, there has been an interest in extending the application of Next Generation Sequencing (NGS) technologies to sequence STR loci. The major advantage offered by NGS over CE methods when

* Corresponding author.
E-mail address: sarah.riman@nist.gov (S. Riman).

https://doi.org/10.1016/j.fsigen.2023.102872
Received 22 December 2022; Received in revised form 6 April 2023; Accepted 8 April 2023
Available online 10 April 2023

genotyping STR markers is the ability to capture the full allelic sequence, allowing the (1) identification, at base-pair resolution, of the STR repeat motifs and their respective flanking regions, (2) characterization of the variation sites (single nucleotide polymorphisms (SNPs) and insertions or deletions (INDELs)) within the sequence string, and (3) resolution of isoalleles (alleles identical by length but different in sequences). Also, an added benefit of NGS method is that it allows the deduction of the length-based information associated with the obtained sequences for parallel compatibility with the CE-based data [5–7]. This gain of information has been shown to further increase the allelic diversity (i.e., range of allelic variation) and discriminating power of an STR locus, and assist in deconvoluting mixture profiles and resolving complex paternity cases, thus increasing the strength of evidence of the NGS-STR profiles [8–12].

Implementing NGS-STR genotyping workflows in forensic laboratories and taking advantage of the additional power of sequence variations in casework and DNA databanking necessitate (1) awareness of factors that influence concordance and backward compatibility of the length-based allelic calls derived from sequences with their corresponding length-based alleles obtained by the CE methods, (2) a standardized STR sequence nomenclature compatible with CE-based STR nomenclature, and (3) compilation of STR sequence patterns from across worldwide populations [6,8,13–15].

Extensive sequencing of individuals across the globe is essential for identifying novel sequences and comprehensively cataloging STR diversity and variants within and around the core repeat motifs. Several studies, using different multiplex STR panels and sequencing platforms, characterized sequence variations in major population groups (Americas [16–23], Europe [11,16,24–28], Oceania [16], East Asia [16,22,29,30], South Asia [11,16,24,31], West Africa and North East Africa [24], and Central Africa [11,16]). To date, a limited number of studies have explored patterns of sequence variation in Middle Eastern countries [16, 32–34]. To this end, DNA samples from 195 unrelated individuals sampled from the Lebanese population were amplified using a prototype version of the PowerSeq 46GY NGS STR panel (Promega Corporation, Madison, WI, USA) and subsequently sequenced on the MiSeq FGx Sequencing System (Verogen, San Diego, CA, USA). Lebanon, a Levantine country located on the eastern coast of the Mediterranean Sea, is characterized by its distinctive social structure and a heterogenous society composed of people from diverse religious and ethnic groups [35]. Several studies have genotyped STRs across the Lebanese population using CE-based methods, limiting the frequency databases to sized-based alleles only [36–39].

To the best of our knowledge, this is the first report specific to a set of Lebanese population samples that provides (1) a detailed description of the allelic sequences at each of 22 autosomal STR forensic loci with their respective observed variants in both the repeat and flanking regions, (2) frequency estimates of the sequence-based alleles, and (3) a list of sequence variants defined as novel relative to sequences curated in STRSeq BioProject, PRJNA380127 [8]. Genotype concordance was evaluated herein by comparing the length-based alleles inferred from the full sequences with the length-based allelic calls typed by CE technology in [40]. The effect of the additional resolution provided by sequencing was assessed by comparing forensic and population genetic parameters calculated from sequence-based data to those obtained using the length-based information. The genetic distance and population structure of the Lebanese individuals in relation to four U.S. populations (African American, Asian, Caucasian, and Hispanic) were also investigated.

2. Materials and methods

2.1. Lebanese population dataset and ethics statement: DNA sampling and CE-based STR typing

One hundred ninety-five genomic DNA extracts (85 males and 110 females) were shared by Indiana University-Purdue University Indianapolis. These extracts were from saliva samples collected randomly and with informed consent at the American University of Beirut from unrelated anonymized individuals of self-reported Lebanese admixed ancestry. The campus is located in Beirut, the capital of Lebanon and one of the most diverse cities in the country. The permission to share and sequence the samples for this study was in accordance with the American University of Beirut (PALM.RM.27), Indiana University-Purdue University Indianapolis Institutional Review Board (study number 1409306349), and the National Institute of Standards and Technology (NIST) Research Protections Office. The details on sample collection, methods of extraction and quantitation of genomic DNA, and amplification of the DNA extracts using the PowerPlex Fusion 6C multiplex kit (Promega) are outlined in [40].

2.2. STR typing by the NGS-based panel, PowerSeq 46GY System

Genomic DNA extracts were amplified with the PowerSeq 46GY System Prototype (Promega) [41], a PCR-based 46-locus multiplex. The kit co-amplifies 22 autosomal STR loci, 23 Y-STR loci, and Amelogenin, from which amplicons can be used to prepare NGS DNA libraries. Only the genotypes and data analysis across the 22 autosomal STR loci are reported in this study: CSF1PO, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D1S1656, D21S11, D22S1045, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, FGA, Penta D, Penta E, TH01, TPOX, vWA. Positive and no-template controls were included in each amplification plate. PCR reactions were prepared in a final volume of 25 μL and amplified at 30 cycles on GeneAmp PCR System 9700 Thermal Cycler (Thermo Fisher Scientific, Foster City, CA, USA) following the manufacturer's protocol of the Prototype as described in [42]. Amplified PCR products were then purified with AMPure XP (Beckman Coulter Inc., Brea, CA, USA) using a 3X beads-to-sample volumetric ratio. Bead washing and DNA elution were performed following the manufacturer's instructions [43]. DNA concentrations of the eluted amplicons were quantified using the Quant-iT PicoGreen dsDNA assay kit (Thermo Fisher Scientific) following the manufacturer's protocol [44].

2.3. DNA library preparation for Illumina sequencing

DNA libraries were manually prepared using Kapa Hyper Prep Kit (Roche Sequencing and Life Science|Kapa Biosystems, Roche, Wilmington, MA, USA) according to the manufacturer's procedures [45]. DNA libraries were prepared for cluster generation using the 96 CD (combinatorial dual) Indexes contained in the TruSeq DNA PCR-Free HT Library Prep Kit (Illumina, San Diego, CA, USA) [46]. Briefly, the total elution volume (~ 25 μL) of each purified amplification reaction was used as DNA input to end repair, A-tailing, and adaptor ligation according to the supplier recommended protocol [45]. Ligation products were purified using a 0.8X volume of AMPure XP Beads and DNA libraries were then eluted off the beads following the procedures in [41, 45].

2.4. Sequencing

Each indexed DNA library was quantified on the ABI 7500 Real Time PCR System (Thermo Fisher Scientific) using PowerSeq Quant MS System (Promega) following the manufacturer's technical guide [47]. Libraries were then individually normalized to an equal concentration of 4 nM and pooled by equal volume (volumetric pooling). The pooled libraries were then denatured, and diluted to 20 pM following the manufacturer's technical guide [48]. The denatured 20 pM library pools were further diluted to a final concentration of 13.5 pM and loaded onto the reagent cartridges along with a 1.5% Illumina PhiX control spike-in [42]. There was a total of three sequencing runs in which 96, 95, and 70 libraries were sequenced in run 1, 2, and 3, respectively. Paired-end

sequencing was performed on the MiSeq FGx Sequencing System (Verogen) in Research Use Only (RUO) mode using 600 cycle (2 ×300 bp reads) MiSeq v3 Reagent Kit (Illumina) [49].

2.5. Identification and assessment of sequence variations with STRait Razor

Raw FASTQ files were analyzed using STRait Razor software v3.0 with a modified version of the PowerSeqv2.31.config file (Supplementary File 1) [50]. The resulting sample files were then processed using an in-house Perl script to identify across STR markers the: (1) length-based allelic calls derived from the sequence strings, (2) flanking regions and repeat motifs of the sequence-based alleles, and (3) counts of reads (coverage) associated with each sequence variant. Various cut-off thresholds were applied on all the autosomal STR loci to capture as much data as possible and minimize manual removal of stutter artifacts and non-allelic sequences: (1) minimum depth of coverage threshold of 10X; (2) stutter ratio (SR) threshold of 20%; and (3) heterozygote sequence balance (Hb) of 20%. The SR threshold was defined as: $\left(\frac{\text{coverage of the actual stutter } \textit{sequence}}{\text{coverage of its known parental allelic } \textit{sequence}}\right) x\,100\%$. For the Hb, the two allelic sequences with the highest coverage were first identified at each heterozygous locus, and the Hb was defined as in [42]: $\left(\frac{\text{lower coverage } \textit{value}}{\text{higher coverage value}}\right) x\,100\%$. It is to be noted that the thresholds implemented in this study were chosen based on in-house experience of sequencing and analyzing large datasets of single-source samples of which the ground truth of both the length- and sequence-based genotypic data were well characterized with different CE-STR kits, NGS-STR kits, multiple sequencing platforms, and various bioinformatic pipelines [17,51,52].

The filtered sequencing results were manually reviewed for each individual sample, and sequence-derived allelic calls were compared to the CE fragments length information (discussed further in Section 2.6). Using this comparison,

- Artifacts and non-allelic sequences exceeding the 20% cut-off values were reviewed and removed from the dataset.
- Allelic sequences that did not meet the Hb threshold were investigated and retrieved only if they were present in the original unfiltered basic files.

The allelic sequences across all the loci encompassed both the flanking regions and repeat motifs. All sequence variants were reported on the forward strand of the human genome reference assembly (GRCh38) and annotated according to the "nomenclature and guidance" recommended in [13,14].

For quality control purposes, the STR sequence data from this study was inspected in STRidER (STRs for identity ENFSI Reference database) and was issued the following STRidER accession number: STR000397 [53]. SNPs were submitted to NCBI dbSNP to obtain 'rs' IDs. Sequence-based alleles were submitted to NCBI GenBank through the STRSeq BioProject [8].

2.6. Concordance assessment with CE-based STR data

Using the in-house Perl script, length-based allele calls genotyped using CE-based PowerPlex Fusion 6C multiplex (Promega) and published in [40] were compared with the length-based numerical values derived from the observed allelic sequence strings. The concordance comparison was performed across the overlapping markers between the CE- and NGS-based STR panels. The shared loci are listed in Section 2.2.

2.7. Discordance assessment: amplification and allele sizing using Sanger sequencing

A discrepancy between the CE- and NGS-based data at D16S539 for a single sample was investigated by performing Sanger sequencing as previously described in [54] using the following sequencing primers (5′→3′) (D16S539-Forward: ACTCTCAGTCCTGCCGAGGT, D16S539-Reverse: AACAGCCTACAGAGTGATTCCA) and an annealing temperature of 60 °C. Sanger sequences were analyzed with Geneious Prime v2022.1.1 software (Biomatters Ltd. Location, Auckland, New Zealand) [55].

2.8. Population genetics data analysis

Population genetics data analysis were assessed for length- and sequence-based data across each autosomal STR locus. STRAF (STR Analysis for Forensics) version 2.0.8 [56], a freely available online web application, was used to determine the following population forensic genetic parameters: STR allelic frequencies, total allele counts (N), number of different alleles (Nall), observed heterozygosity (Hobs), genetic diversity (GD) or expected heterozygosity (Hexp), polymorphism information content (PIC), match probability (PM), power of discrimination (PD), power of exclusion (PE), and typical paternity index (TPI).

Arlequin suite version 3.5.2.2 [57], a software that performs population genetics analysis, was used to test for (1) departures from Hardy-Weinberg Equilibrium (HWE) (with 1000,000 Markov chain) and (2) pairwise linkage disequilibrium (LD) tests between autosomal loci located on the same chromosome. Significant deviations from HWE (*p-values* < 0.05) were adjusted by applying a Bonferroni correction for multiple tests (comparisons).

It is to be noted that the STRAF's and Arlequin's input files containing sequence-based genotypic data were generated as following: each unique STR allelic sequence was encoded by a unique number, a special format compatible with both software.

2.9. Lebanese and four U.S. populations

The sequence-based genotypes and allele frequencies identified in the Lebanese individuals (N = 195) were compared to the sequence-based data of four U.S. populations: African American (N = 342), Asian (N = 97), Caucasian (N = 361), and Hispanic (N = 236), previously analyzed and reported in [17]. The analysis ranges of STR flanking regions are not the same across PowerSeq 46GY assay (used in this study) and ForenSeq DNA Signature Prep Kit (used in [17]) [58]. Therefore, and for the purpose of the conducted comparisons discussed in Sections 2.9.1 and 2.9.2, all the STRait Razor processed sequences obtained from the five populations were adjusted across the studied markers using the GRCh38 coordinate ranges defined in Supplementary File 2. Briefly, these ranges only reported the repeat motifs and additional nucleotides in the flanking regions adjacent to the repeat regions.

2.9.1. Multidimensional scaling, population tree, analysis of molecular variance, and population comparisons

Multidimensional scaling (MDS) plot and population tree were constructed in STRAF v.2.1.5 based on Nei's genetic distance [56]. The Analysis of MOlecular VAriance (AMOVA), locus-by-locus F_{ST} for population pairs, and pairwise F_{ST} between populations were estimated in Arlequin software version 3.5.2.2. with 10,000 permutations to obtain *p-values* for statistical significance [57].

2.9.2. Population genetic structure

Population structure analysis was determined in *STRUCTURE* v2.3.4 [59,60]. The sequence-based genotypes of the 1231 Lebanese and U.S. individuals [17] were used as the input data. *STRUCTURE* runs were first conducted using the admixture correlated allele frequencies models (i.e., unsupervised models). *STRUCTURE* was then rerun on the same

dataset using the supervised modelling assumptions (i.e., admixture correlated allele frequencies models along with the *LOCPRIOR* model). *LOCRPRIOR* is a model parameter built in *STRUCTURE* that allows the software to use prior knowledge about the origin of the studied populations [61,62].

A predefined range of populations/clusters (*K*) was set from 1 to 8. For each *K* value, *STRUCTURE* was run 20 times with a burn-in period of 100,000 followed by a 100,000 Markov Chain Monte Carlo (MCMC) replication number. The clustering runs were then processed in *CLUMPAK* (*Clustering Markov Packager Across K*) to average replicate runs with similar clustering and identify the best alignment across the range of *K* values [63,64]. The *CLUMPAK* analysis was summarized in the graphical bar plots produced by *DISTRUCT* for many *K*'s [63,65]. Herein and for each *K*, only the results of runs containing the largest number of replicates (i.e., "the major mode" [64,66]) and largest average log probability [LnP(D)] of the data are presented.

3. Results and discussion

3.1. Sequencing quality metrics and performance of the PowerSeq 46GY prototype panel

The quality metric values for each sequencing run are summarized in Supplementary Table 1. The depth of coverage (DoC) for all the true allelic sequences ranged between 131 reads observed at D2S1338 to a maximum of 60,532 reads observed at D2S441 (Supplementary Table 2). The average DoC ranged between 1290 and 8768 reads and is shown for each marker in Supplementary Table 2. The Hb ranged from 0.1 for D2S1338 to 1 for D16S539 and TPOX. The average Hb for all the markers obtained across the different sequencing runs ranged between 0.5 and 0.9.

The variability in sequence coverage (Supplementary Table 2) and extreme heterozygous imbalance attributed to preferential amplification of the shorter allelic sequences (Supplementary Table 3) have been previously reported with other studies that used the prototype version of the same NGS-based STR amplification kit [19,21]. We speculate that this sequence coverage variability could be due to the use of different prototype versions of the PowerSeq 46GY panel that was still under development while conducting this study.

3.2. Genotype concordance assessment

Discordance in length-based allelic calls between the CE-based and NGS-based data have been observed in several studies. Reasons of discordance were attributed to (1) differences in the position of PCR primers of CE-based and NGS-based STR panels leading to different amplicon sizes (e.g., occurrence of INDELS inside the CE amplicons but outside of primer regions of NGS kits) [19,30,67], (2) SNPs or INDELs observed in the sequenced flanking regions [19,22,24,28,30,67], (3) SNPs or INDELs that exist in the primer binding sites [19,24,30], and (4) differences in the sequence analysis ranges (i.e., location of anchor sites) between various software and bioinformatic pipelines used to analyze sequence-based data [28].

To check for concordance in the Lebanese dataset, the length-based alleles inferred from the sequence strings (Column D in Supplementary Table 4) were compared with the length-based alleles genotyped by CE in [40]. This comparison was conducted at the shared aSTR loci in both NGS- and CE-based STR kits (Section 2.2). Across all the 8580 comparisons (195 samples x 22 autosomal markers x 2 alleles (irrespective of zygosity)) there was only one instance of discordance in allele designation detected in one sample at D16S539, leading to an allele concordance rate of 99.98%. The CE method generated a heterozygous genotype (9,12) at D16S539 (Fig. 1A) while the sequencing data analyzed by STRait Razor v3 reported an apparent homozygous genotype (9,9) (Fig. 1B). The cause of this discrepancy was further investigated by first reanalyzing this sample using another bioinformatic pipeline, FDSTools v1.1.1 [68], that agreed with the STRait Razor v3 results (i.e., 9,9). The sample was then re-sequenced using Sanger methods and sequencing primers designed outside of the PowerSeq 46GY primer binding sites (Section 2.7). A previously characterized SNP [G>A] with an assigned rs# (rs1009712060) was detected three bases upstream of the beginning of the 5′ flanking region (i.e., near the 3′-end of the forward primer) of allele 12 (Fig. 1C). This SNP could have interfered with the binding of the primers contained in the PowerSeq 46GY panel resulting in an undetected signal of the allelic sequence "12" and to a null allele. The genotype for this specific sample at D16S539 was excluded from the allele frequency calculations and population genetics data analysis.

To conclude, there was only one discordance attributed to a null allele that was due to a SNP detected in the PCR primer binding region. The concordance rate was 99.98% and further proves the robustness of NGS applicability to sequence forensic STRs. However, examining more than one method (e.g., NGS and CE methods, different NGS kits, and various bioinformatic pipelines) would be a potential additional concordance diagnostic check.

3.3. Autosomal STR sequence variations and allelic gains

A total of 8578[1] aSTR allelic sequences, of which 429 individual sequences were detected across the 22 aSTR loci in the 195 Lebanese samples using the NGS-STR genotyping workflow. The full sequences across each STR marker are described in detail in Supplementary Table 4 with their respective GRCh38 coordinate ranges, counts, PowerSeq 46GY assay specific 5′ and 3′ flanking regions, pre-existing/novel SNPs in the flanking regions, unbracketed and bracketed repeat motifs according to [13,14], and the GenBank accession and version numbers. It is to be noted that there were no reported 3′ flanking sequences for FGA and D10S1248 loci. The nucleotides indicated in lower cases in the bracketed repeat regions of D19S433 and D21S11 were not considered in the numerical designations of the length-based alleles (Supplementary Table 4) [8,14].

Sequence level variation was detected across 18 out of the 22 aSTR loci (blue bars in Fig. 2) leading to an increase in the number of alleles observed by sequence as compared to the length-based data (magenta bars in Fig. 2). The STR markers in Fig. 2 are arranged in an ascending order of the additional allelic gain that ranged between 0 and 33 alleles (i.e., 0–220% increase in number of alleles). This observed increase in allelic diversity (a total of 82%) were attributed to variations in either the repeat motifs and/or flanking regions (Supplementary Table 4). The total number of distinct alleles observed with the length-based data was 236 and ranged from a count of 7 alleles (TH01, TPOX) to a maximum count of 18 alleles observed at Penta E. The total number of distinct sequence-based alleles were 429, showing that 193 alleles could be only distinguished by variants in either the repeat motifs and/or flanking regions of the sequences and not by length (Fig. 2 and Supplementary Table 5). D12S391, a highly polymorphic compound marker, displayed the highest number of observed allelic sequences (Nall = 48) and showed the largest increase in alleles (threefold) similar to observations reported with different sequenced populations [21,24,25,27,28,34]. Three STR loci of simple repeat motifs (D10S1248, D18S51, and Penta E) and one compound STR locus (D22S1045) showed no gains of alleles when sequenced, providing identical information and discrimination power as with the length variation data. There were other studies that (1) also showed no sequence variation at these loci (e.g., in [18,25]), or (2) observed increase in the number of alleles at all of these loci when sequenced [17,22], or (3) recorded within the same study gain or no gain in the number of alleles at one or more of these loci upon sequencing [16,20,23,24,26–28,30,31,34]; for example, there was no

[1] 8580 sequences minus the discordance genotype observed at the D16S539 locus (excluded from analysis, Section 3.2) = 8578.

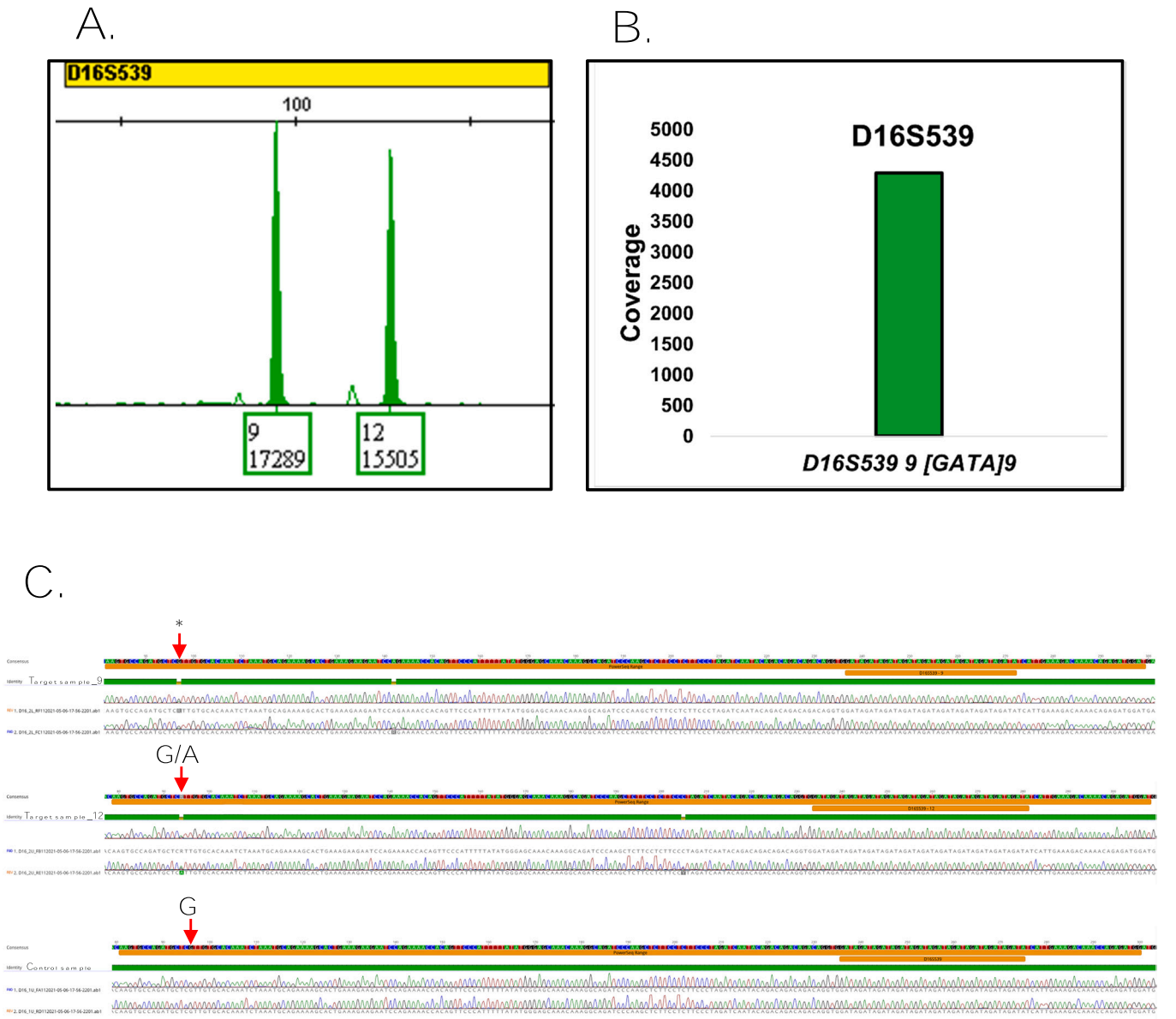

Fig. 1. Discordant genotype between PowerPlex Fusion 6C and PowerSeq 46GY systems. Instance of discordance observed in one of the Lebanese samples at D16S539 (A) vs. (B), with a (9,12) genotype typed with CE-method (A), and an apparent homozygote genotype identified with PowerSeq 46GY with a length-based allele calls of (9,9) (B). (C) Sanger sequencing results of D16S539 identified two distinct sequence variants: D16S539 9 [GATA]9 (upper sequence) and D16S539 12 [GATA]12 rs1009712060 (middle sequence). The rs1009712060 is a SNP variation (G > A) upstream the 5′ flanking region that could have resulted in the discordance shown in (B). It is to be noted that the upper Sanger sequence chromatogram displayed overlapping peaks of A and G (asterisk). Such results can be seen when sequencing a heterozygote locus where the two bands of the amplified PCR products (herein alleles 9 and 12) did not resolve and separate well on the PCR gel before cutting and sequencing the bands. In such cases, the sample might be reamplified and sequenced if enough sample is available. The lower sequence is the Sanger sequencing results of D16S539 from a positive control sample that was processed in parallel with the target sample. The control was homozygote at D16S539, and no mixed peaks were detected.

gain at D10S1248 but an increase in alleles at D18S51, Penta E, and D22S1045 in [16,24]. These various observations of which locus/loci are either showing an increase or no gain in number of alleles due to sequencing could be attributed to the total number of sampled individuals and/or population types in the dataset considered in each study.

3.4. SNPs and INDELs in STR flanking regions

Within the Lebanese dataset, thirty-four polymorphisms consisting of SNPs or INDELs were identified in regions flanking the repeat motifs of 12 aSTR loci. Herein, these polymorphisms are reported using the forward strand of the human genome reference sequence GRCh38 and are summarized in Supplementary Table 6 with hyperlinks to (1) their respective dbSNP ID Number and (2) their description using the Human Genome Variation Society (HGVS) nomenclature [69,70]. The count, frequency, and the adjacent repeat motifs for each of the described flanking polymorphisms are detailed in Supplementary Table 4. Briefly, there were five variations relative to GRCh38 at each of D7S820 (three SNPs, one deletion, and one insertion), D13S317 (four SNPs and one deletion), and vWA (five SNPs); three variations at each of D16S539 (three SNPs), D19S433 (one SNP and two deletions), Penta D (two SNPs and one deletion), and TPOX (three SNPs); two variations at each of D5S818 (two SNPs) and TH01 (two SNPs); and one variation that

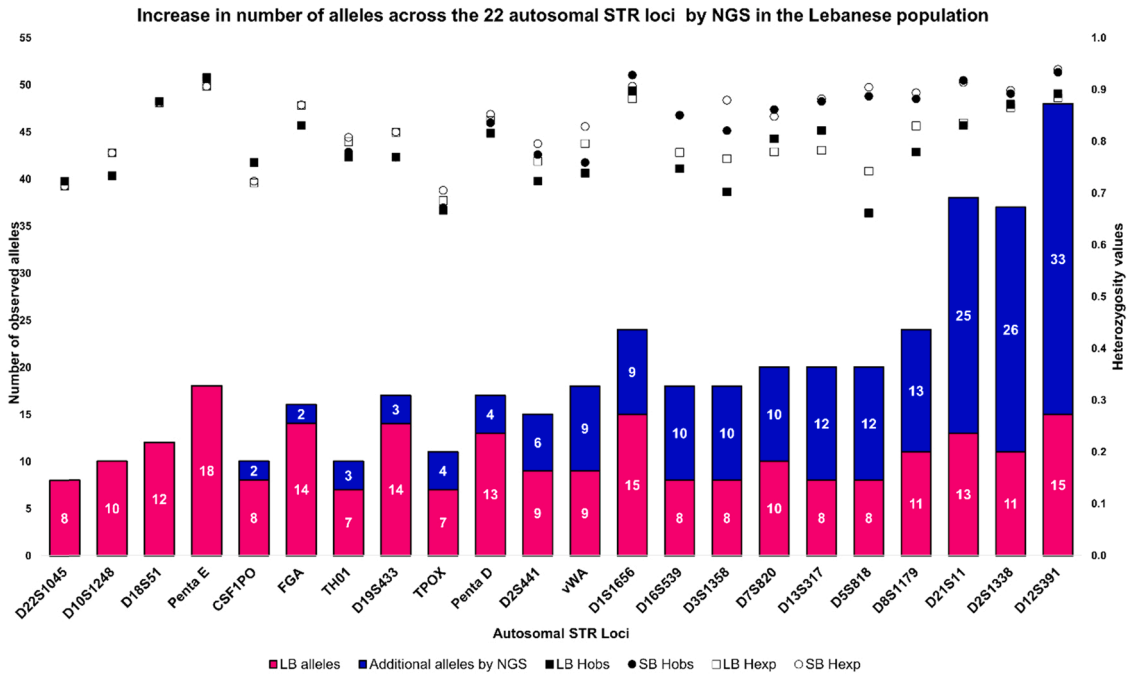

Fig. 2. Number of length-based (LB) and sequence-based (SB) alleles along with the heterozygosity values across the 22 aSTR loci in 195 Lebanese samples. Number of LB alleles (magenta bars) and the further gain in alleles due to variants in either the repeat regions and/or flanking regions (blue bars) obtained by NGS (left y-axis). Loci are sorted in ascending order of the additional information. The Hobs and Hexp values of the LB data (■ for Hobs and □ for Hexp) and SB genotypes (● for Hobs and ○ for Hexp) are plotted at each locus using the right y-axis.

consisted only of one SNP was observed at each of D1S1656, D2S1338, and D2S441. It is to be noted that D16S539 and D19S433 markers had each a newly identified SNP variant (did not previously exist in dbSNP) in their respective flanking regions. These variants were submitted to dbSNP and assigned reference SNP ID numbers: rs2143735806 residing in the 5′ flanking region of D16S539 and rs2145186901 observed in the 5′ flanking region of D19S433.

Accumulating occurrences of flanking variants as depicted in

Table 1
List of 25 aSTR sequence variants, with no previously existing records in STRSeq BioProject, identified in the Lebanese dataset.

Locus	Allele	STR sequence nomenclature	Type of novelty	Observation	Frequency	STRSeq Accession number
TPOX	9	TPOX 9 [AATG]9 rs149212737	Flanking Region	1	0.0026	OK330025.1
D2S1338	22	D2S1338 22 [GGAA]16 [GGCA]6 rs6736691	Flanking Region	3	0.0077	ON862889.1
D2S1338	24	D2S1338 24 [GGAA]2 GGAC [GGAA]14 GGCA GACA [GGCA]5	Repeat Region	1	0.0026	OK330001.1
D5S818	9	D5S818 9 [ATCT]9	Flanking Region	1	0.0026	OK330003.1
D5S818	12	D5S818 12 [ATCT]3 AGCT [ATCT]8 rs73801920 rs25768	Repeat Region	2	0.0051	OK330004.1
D5S818	13	D5S818 13 [ATCT]5 GTCT [ATCT]7	Repeat Region	1	0.0026	OK330005.1
CSF1PO	12	CSF1PO 12 [ATCT]6 ACCT [ATCT]5	Repeat Region	1	0.0026	OK329996.1
D7S820	9.1	D7S820 9.1 [TATC]9 rs7789995 rs1463708262	Flanking Region	1	0.0026	OK330006.1
D7S820	9.3	D7S820 9.3 [TATC]10 rs897512434	Flanking Region	1	0.0026	OK330007.1
D7S820	10	D7S820 10 [TATC]10 rs7786079	Flanking Region	1	0.0026	OK330008.1
D8S1179	12	D8S1179 12 TCTA TCTG [TCTA]6 TGTA [TCTA]3	Repeat Region	1	0.0026	OK330009.1
vWA	13	vWA 13 [TAGA]3 TGGA [TAGA]3 [CAGA]4 [TAGA]2 rs75219269 rs11063969 rs11063970 rs11063971	Flanking Region	1	0.0026	OK330026.1
vWA	17	vWA 17 [TAGA]11 TGGA [CAGA]4 TAGA	Repeat Region	1	0.0026	OK330028.1
D13S317	11	D13S317 11 [TATC]9 TGTC TATC	Repeat Region	1	0.0026	OK330011.1
D13S317	12	D13S317 12 [TATC]8 TGTC [TATC]3 rs9546005	Repeat Region	2	0.0051	OK330012.1
D16S539	7	D16S539 7 [GATA]7 rs1728369	Flanking and Repeat Region	1	0.0026	OL441846.1
D16S539	11	D16S539 11 [GATA]9 TATA GATA	Repeat Region	1	0.0026	OL441848.1
D16S539	12	D16S539 12 [GATA]12 rs2143735806	Flanking Region	1	0.0026	OL441849.1
D19S433	12	D19S433 12 [CCTT]11 ccta CCTT cttt CCTT rs903567912	Flanking Region	1	0.0026	OL441843.1
D19S433	13.2	D19S433 13.2 [CCTT]12 ccta CCTT cttt CCTT rs2145186901rs745607776	Flanking Region	1	0.0026	OL441844.1
D21S11	28	D21S11 28 [TCTA]6 [TCTG]5 [TCTA]3 ta [TCTA]2 tca [TCTA]2 tccata [TCTA]10	Repeat Region	2	0.0051	OK330015.1
D21S11	30	D21S11 30 [TCTA]5 [TCTG]7 [TCTA]3 ta [TCTA]3 tca [TCTA]2 tccata [TCTA]10	Repeat Region	1	0.0026	OK330018.1
D21S11	30.2	D21S11 30.2 [TCTA]5 [TCTG]5 [TCTA]3 ta [TCTA]2 tca [TCTA]2 tccata [TCTA]12 TA TCTA	Repeat Region	1	0.0026	OK330019.1
D21S11	33.1	D21S11 33.1 [TCTA]5 [TCTG]6 [TCTA]3 ta [TCTA]3 tca [TCTA]2 tccata [TCTA]10 TCA [TCTA]2 TA TCTA	Repeat Region	1	0.0026	OK330021.1
Penta D	14	Penta D 14 [AAAGA]14 rs186259515	Flanking Region	1	0.0026	OK330023.1

Supplementary Table 6 across multiple population datasets could facilitate studying the presence of associations between STR repeat alleles and their surrounding SNPs and INDELs [16,71]. For example, we noticed from this dataset that the occurrence of the rs4847015, a 3′ flanking region SNP at D1S1656, occurred adjacent to the repeat motifs of the x.3 microvariant alleles (intermediate alleles with incomplete tetranucleotide repeat), an association that has been observed with other studies [30,72].

3.5. Novel sequence variants

Twenty-five novel sequence variants identified by the PowerSeq 46GY System, with no previously existing match or records in STRSeq BioProject [8], were observed in the Lebanese samples across 12 loci. The novel sequences were issued new GenBank accession numbers and are described in Table 1 with the following associated information (1) aSTR locus at which the novel sequence(s) were observed, (2) length-based allele values determined from the STR sequence strings, (3) STR allelic sequence nomenclature in accordance with ISFG considerations [13], (4) type of novelty (differences in flanking regions or repeat motifs) relevant to STRSeq records, (5) number of observations within the considered dataset, (6) respective frequency in this study, and (7) hyperlink for the novel STRSeq accession and version numbers.

D21S11 showed the highest count of novel alleles (N = 4). Twenty-one of the novel sequence-variants were observed only once in the Lebanese dataset (Table 1). Sequencing a larger set of Lebanese individuals and/or other populations from diverse ethnic groups could confirm in the future if these singleton sequences are rare sequences and/or specific to the Lebanese population.

3.6. Population and forensic genetic statistics

3.6.1. Allelic frequency tables for length- and sequence-based data

The allelic frequencies for the sequence variants and their associated length-based alleles identified by the PowerSeq 46GY System were calculated in STRAF [56] across each autosomal STR marker and are presented in Supplementary Tables 4 and 7, respectively.

The number of alleles as well as the allele and genotype frequencies for length- and sequence-based data were then used to study population genetics and forensic parameters at each locus in the Lebanese population (Supplementary Tables 8–10). These parameters are discussed below.

3.6.2. Observed heterozygosity, expected heterozygosity, and Hardy-Weinberg equilibrium

Heterozygosity is a measure of genetic variation within a population. Hobs is the fraction of individuals that are heterozygous in the population at a given genetic marker. Hexp, also known as GD, is the probability that two randomly drawn alleles from the population will be different. Hexp depends on the increase in the number of alleles and evenness of allele distribution in the population [73]. Both Hobs and Hexp were determined for both the length-based and sequence-based data, and the values are shown in Table 2 and plotted in Fig. 2. Penta E showed the highest Hobs (> 92%) and Hexp (> 90%) values with the length-based information. When sequenced, D12S391, D1S1656, and D21S11 displayed more than 90% in Hobs and D12S391, D21S11, D1S1656, and D5S818 showed Hexp values > 90%. The gains in Hobs (0.8–34.1%) and Hexp (0.1–21.9%) due to sequencing were also studied and determined from the % difference in these values between sequence-based data and length-based information (shown in descending order in Table 2). D5S818 showed the largest gain in Hobs (~34%) and Hexp (~22%). TPOX showed the lowest gain in Hobs (0.8%) and D19S433 and FGA gained only 0.1% in Hexp after sequencing. Seven loci (CSF1PO, D10S1248, D18S51, D19S433, D22S1045, FGA and Penta E) showed identical Hobs for the length- and sequence-based data. Similarly, four loci (D10S1248, D18S51, D22S1045, and Penta E) showed no gain in Hexp at all between the length- and sequence-based alleles.

Therefore, the majority of the studied aSTRs showed an increase by sequencing in both the Hobs (15 out of 22 STR loci) as well as in the Hexp (18 out of 22 STR loci). This further indicates the effect of the additional resolution provided by the sequencing that was able to resolve isoalleles, thus leading to an increase in the number of alleles across multiple aSTR loci. However, an increase in the number of alleles did not necessarily lead to gains in either the observed or expected heterozygosity (e.g., FGA, D12S391 and D2S1338 vs. D5S818) since

Table 2
Hobs, Hexp, and HWE of the LB and SB alleles in the Lebanese population (N = 195). The gain in heterozygosity (%) at certain loci is due to the sequence-based informativeness. Loci are sorted in descending order of the increase in heterozygosity. Nall = number of the total alleles observed at each locus. A genotype for one of the samples at D16S539 has been removed due to a null allele (discussed in Section 3.2), hence a lower Nall than the rest of the loci.

Hobs					GD or Hexp				*p-values* of HWE		
aSTR Loci	Nall	LB	SB	Gain in Hobs (%)	aSTR Loci	LB	SB	Gain in Hexp (%)	aSTR Loci	LB	SB
D5S818	390	0.662	0.887	34.1	**D5S818**	0.742	0.904	21.9	**CSF1PO**	0.121	0.153
D3S1358	390	0.703	0.821	16.8	**D3S1358**	0.766	0.880	14.8	**D10S1248**	0.319	0.342
D16S539	388	0.747	0.851	13.8	**D13S317**	0.783	0.882	12.7	**D12S391**	0.009	0.030
D8S1179	390	0.779	0.882	13.2	**D21S11**	0.836	0.914	9.3	**D13S317**	0.461	0.274
D21S11	390	0.831	0.918	10.5	**D16S539**	0.778	0.851	9.3	**D16S539**	0.051	0.185
D13S317	390	0.821	0.877	6.9	**D7S820**	0.780	0.848	8.7	**D18S51**	0.822	0.804
D7S820	390	0.805	0.862	7.0	**D8S1179**	0.830	0.894	7.7	**D19S433**	0.578	0.567
D2S441	390	0.723	0.774	7.1	**D12S391**	0.884	0.939	6.2	**D1S1656**	0.910	0.672
D12S391	390	0.892	0.933	4.6	**D2S441**	0.761	0.795	4.5	**D21S11**	0.217	0.029
D1S1656	390	0.897	0.928	3.4	**D2S1338**	0.865	0.899	3.9	**D22S1045**	0.319	0.303
Penta D	390	0.815	0.836	2.5	**vWA**	0.796	0.829	4.1	**D2S1338**	0.860	0.378
vWA	390	0.738	0.759	2.8	**D1S1656**	0.883	0.907	2.7	**D2S441**	0.105	0.062
D2S1338	390	0.872	0.892	2.4	**TPOX**	0.685	0.705	2.9	**D3S1358**	0.300	0.055
TH01	390	0.769	0.779	1.3	**Penta D**	0.840	0.852	1.5	**D5S818**	0.057	0.031
TPOX	390	0.667	0.672	0.8	**TH01**	0.799	0.807	1.0	**D7S820**	0.499	0.754
CSF1PO	390	0.759	0.759	0.0	**CSF1PO**	0.720	0.723	0.4	**D8S1179**	0.023	0.011
D10S1248	390	0.733	0.733	0.0	**D19S433**	0.817	0.818	0.1	**FGA**	0.032	0.032
D18S51	390	0.877	0.877	0.0	**FGA**	0.870	0.870	0.1	**Penta D**	0.204	0.184
D19S433	390	0.769	0.769	0.0	**D10S1248**	0.778	0.778	0.0	**Penta E**	0.244	0.238
D22S1045	390	0.723	0.723	0.0	**D18S51**	0.875	0.875	0.0	**TH01**	0.805	0.534
FGA	390	0.831	0.831	0.0	**D22S1045**	0.714	0.714	0.0	**TPOX**	0.527	0.163
Penta E	390	0.923	0.923	0.0	**Penta E**	0.906	0.906	0.0	**vWA**	0.126	0.128

these alleles could be well distributed across the population (Table 2 and Fig. 2); this is in addition to the loci with no gain in alleles at all (mentioned above).

Table 2 also shows the *p-values* associated with exact tests performed to detect departure from HWE across both the length- and sequence-based 22 aSTR loci. Using a statistically significant level of 5%, the *p-values* of D12S391, D8S1179, and FGA were < 0.05 for both the length-based and sequence-based data. With sequencing data, D21S11 and D5S818 had *p-values* < 0.05. After applying Bonferroni correction (α' = 0.05/22) to HWE tests, all the *p-values* of the loci surpassed α' (*p-values* > 0.00227) and no violation of HWE was further detected for both the length-based and sequence-based data.

3.6.3. Polymorphism information content, power of discrimination, and power of exclusion

PIC is a measurement of the informativeness and polymorphism of a genetic marker [74]. PD is correlated to PM (discussed below) and defined as the probability of discriminating between two randomly selected and unrelated individuals [56], and PE is a parameter that measures the power of a genetic marker in excluding a randomly selected and unrelated individual in paternity testing [75]. PIC, PD, and PE were also evaluated for both the length- and sequence-based data. These values are presented in Supplementary Table 8. The highest PIC, PD, and PE values were observed at Penta E with the length-based information and at D12S391 with the sequence-based data. An increase in the values of PIC (0.1–28.1%), PD (0.1–8.9%), and PE (2.0–107.2%) due to sequencing was observed across 18, 17, and 15 loci, respectively. D5S818 displayed the highest gain across these parameters.

3.6.4. Match probability and typical paternity index

The PM measures the probability of a genotype match between two unrelated individuals from a population [76]. A decrease in PM was observed across 17 STR loci after sequencing and when compared to the length-based information. The combined random PM for the set of the 22 length-based and sequence-based aSTR loci were 5.7E-27 and 4.4E-31, respectively. Sequence variants lowered the cumulative PM by at least 5 orders of magnitude (Supplementary Table 9). A similar pattern of results was seen with studies using different ethnic groups and sequencing kits [20,22,27,28,34].

The TPI for a given locus is an odds ratio, measuring how many times more likely that the alleged father and not an unrelated randomly selected man from the same population is the actual father in trio cases (father-mother-child) [77]. The sequence-based data displayed an increase in TPI values across 15 out of 22 loci when compared to the length-based information. The combined TPI for the length-based and sequence-based variations across the 22 autosomal STRs were 7.5E+ 08 and 2.2E+ 11, respectively (Supplementary Table 9).

Assuming locus independence and no substructure effect, a decrease in cumulative PM and an increase in cumulative TPI should assist in mixture deconvolution and analysis of complex kinship cases [11,28,78].

3.6.5. Linkage disequilibrium

Linkage disequilibrium is the non-random association of alleles at different STR loci due to their physical proximity to one another on a chromosome [79]. The LD was evaluated for the five syntenic pairs of STR loci: TPOX-D2S441, D2S441-D2S1338, D5S818-CSF1PO, vWA-D12S391, and D21S11-Penta D. Supplementary Table 10 shows the chromosomes, physical distances, and LD *p-values* associated with each syntenic pair. Assuming a statistically significant level of 5%, all *p-values* were > 0.05, and no significant linkage disequilibrium was detected for both the length-based and sequence-based data.

3.7. Lebanese and four U.S. populations

In this section, we compared the aSTR sequencing data of the Lebanese population to the aSTR sequencing data of the U.S. populations. We chose the four U.S. populations for two main reasons: (1) the availability and accessibility of the actual sequences in our laboratory [17], and (2) the interest in studying how individuals from Lebanon, a geographically distant country, compared to U.S. individuals with diverse racial and ethnic background. We used the sequenced allele variation instead of sized alleles based on the results presented in [16]. Phillips et al. compared the genetic cluster analysis of sized and sequenced aSTR variation in the CEPH human genome diversity panel and noticed better-defined clusters and "a significant improvement in the quality of population differentiations in STRUCTURE" using sequenced-based STRs as compared to the sized-based STRs (Fig. 5 of [16]).

One caveat of the comparison conducted herein is that it should have ideally included populations that are geographically closer to Lebanon (e.g., populations from the Levant or Middle East). However, at the time of writing we did not have access to aSTR sequencing data from these countries that could have allowed us to make direct comparisons with the Lebanese population.

3.7.1. Multidimensional scaling plot and population tree

The genetic distances and relationships of 1231 individuals, sampled from the Lebanese and four U.S. populations [17], were visually examined using a MDS plot (Supplementary Fig. 1) and population tree (Supplementary Fig. 2). Based on the sequencing data of the aSTRs, the five different populations are shown to be distributed along three parts of the MDS plot. The African American and Asian American populations were distant from the rest of the populations. The studied Lebanese individuals clustered closer to the Caucasian and Hispanic individuals; highlighting a closer genetic relationship between these three different groups though distributed across different geographic locations.

Similarly, the population tree (Supplementary Fig. 2) showed the relationship among the five populations and indicated that the (1) Caucasians and Hispanics showed highest relatedness, which is expected [80], and (2) Lebanese individuals were gathered together in a major category with the Caucasians and Hispanics showing a closer relatedness with them than with the African American and Asian individuals.

3.7.2. Genetic differentiation within and among Lebanese and U.S. populations

AMOVA analysis was performed to determine the distribution of genetic variation among and within the five different populations using the sequence-based data of the aSTR markers. As shown in Supplementary Table 11, genetic differences among individuals within-population accounted for most of the variation in allele frequencies (98.42%), whereas differences among populations accounted for only 1.58% of the variance in allele frequencies; a similar observation to previous studies [81–83]. Using 10,000 permutations, the low genetic differentiation among populations was shown to be statistically significant ($F_{ST} = 0.0158$, *p-value* = 0).

Locus-by-locus F_{ST} and *p-values* were estimated between the Lebanese and each of the four U.S. populations [17] based on the sequence-based data of the aSTRs (Supplementary Table 12). After applying Bonferroni correction (α' = 0.05/88), statistically significant differences were observed at 17 loci between Lebanese and African American and 15 loci between Lebanese and Asian populations. There were significant differences at only eight and seven loci when the Lebanese population was compared with Caucasians and Hispanics, respectively.

Genetic differentiation among populations was also measured between the Lebanese and the four U.S. populations [17] using pairwise F_{ST} for all loci. The pairwise F_{ST} values for each population pair are presented in Supplementary Table 13 and ranged from 0.0051 (least differentiated) to 0.0318 (most differentiated). The permutation test indicated a statistically significant differentiation between the studied Lebanese and the four US populations (*p-value* < 0.005). The genetic

distance was highest for the pair-wise comparison between Lebanese and Asians ($F_{ST} = 0.0197$) as well as Lebanese and African American ($F_{ST} = 0.0162$) and lowest for the pairwise comparison of Lebanese and Caucasian ($F_{ST} = 0.0064$) and Lebanese and Hispanic ($F_{ST} = 0.0064$).

3.7.3. Lebanese and U.S. population genetic structure

We used the genetic clustering algorithms implemented in *STRUCTURE* [59,60] to identify population clusters in individuals sampled from Lebanon and the U.S. and to evaluate the extent of admixture in the Lebanese individuals and their similarities and/or differences to the four U.S. populations [17]. The clustering solutions of K ranging from 2 to 6 obtained using the unsupervised models (i.e., admixture model of correlated allele frequency without using prior information about the origin of populations) indicated that the Lebanese, Caucasian, and Hispanic individuals did not separate into distinct genetic clusters as did the African American individuals (shown as distinct cluster at $K = 2$) and Asians (shown at $K = 3$) (Fig. 3A).

We reevaluated the population structure using the same full dataset (N = 1231), but this time the *STRUCTURE* runs included information on the geographic sampling locations of individuals (i.e., supervised clustering approach with *LOCPRIOR* model enabled) (Fig. 3B). The *LOCPRIOR* option is recommended in cases in which the examined populations show weak population structure signals [61,62,64]. Implementing these models, as illustrated in the bar plots of Fig. 3B, yielded a different clustering pattern as compared to the runs obtained from the unsupervised models (Fig. 3A). We noticed that at $K = 4$, *STRUCTURE* generated a new cluster that exclusively corresponds to the Lebanese individuals who still share a substantial membership proportion in the cluster that corresponds mainly to the Caucasian population. It would have been interesting to examine if this additional ancestry component was also present in populations from the Levant and Middle east; however, and as mentioned above, we did not have the sequencing data from that region for comparisons.

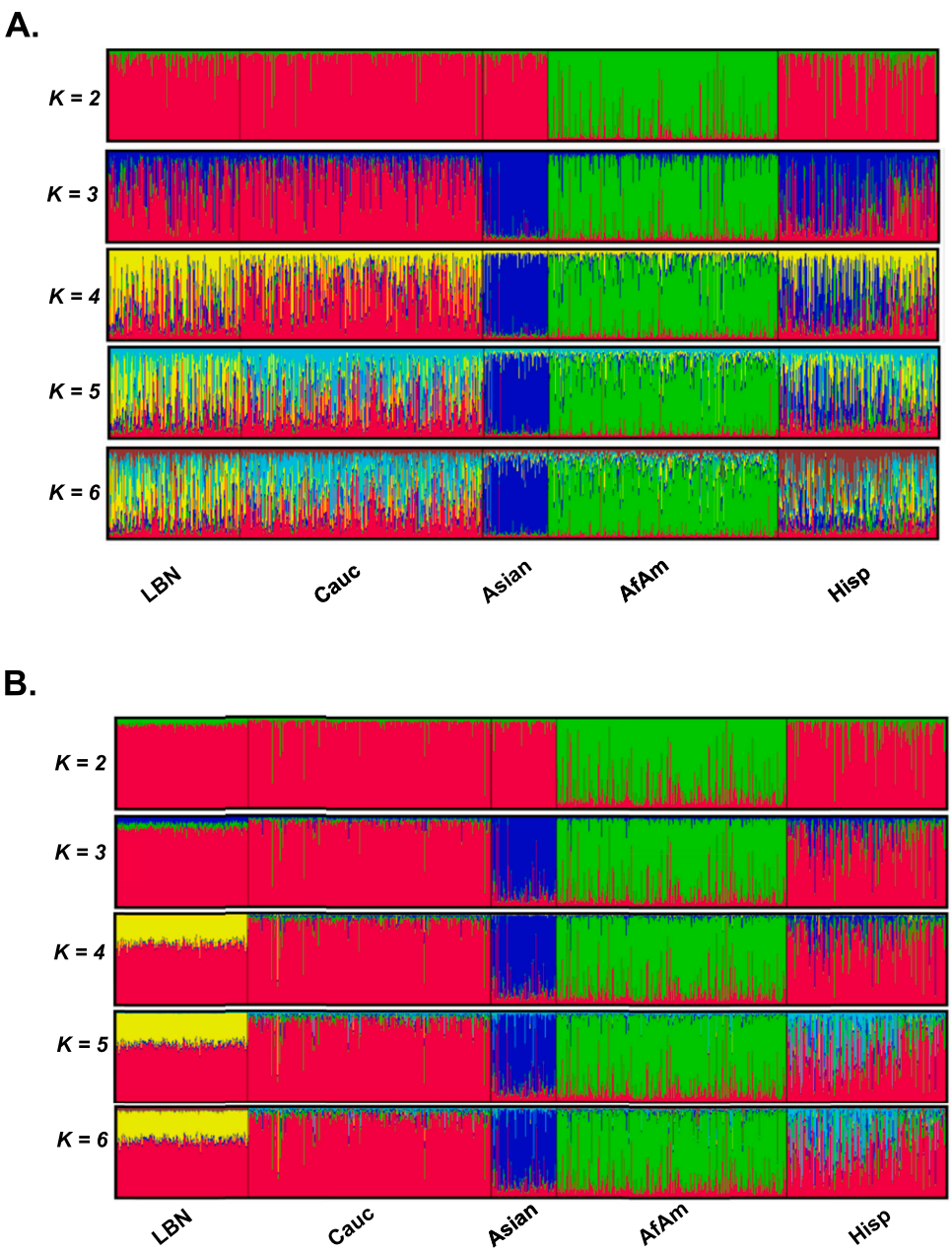

Fig. 3. Population structure. Genetic cluster analysis of Lebanese individuals (LBN, N = 195) together with the four U.S. populations: Caucasian (Cauc, N = 361), Asian (Asian, N = 97), African American (AfAm, N = 342), and Hispanic (Hisp, N = 236) were inferred using two different *STRUCTURE* models. (A) *STRUCTURE* results inferred from the unsupervised models without the use of *LOCPRIOR* approach; (B) *STRUCTURE* results using the *LOCPRIOR* model (i.e., supervised clustering). Each color represents a distinct cluster or group (K) and each of the 1231 individuals is represented by a thin K colored vertical line. The displayed color height for each line is proportional to membership of each individual to different genetic clusters/groups. Thick black vertical lines separate individuals from different populations. The studied populations are indicated on the horizontal axis and values of K are shown on the left of each bar plot. Only $K = 2$, 3, 4, 5 and 6 are shown here.

3.8. Structure analysis of the Lebanese population

We again used *STRUCTURE* [59,60] to perform cluster analysis using, this time, the sequence-based genotypes of the Lebanese individuals only (N = 195). *STRUCTURE* runs were based on the unsupervised models using the same *K* values, replicates, burn-in period, and MCMC iterations as discussed for the analysis containing the full dataset of the five populations (Section 2.9.2). As shown in the bar plots in Supplementary Fig. 3, *STRUCTURE* software did not detect population structure within the Lebanese population for this considered dataset. Individuals showed roughly a similar proportion of membership in the inferred cluster $K = 2$–6. However, that does not imply that there is no genetic distinctiveness among Lebanese individuals. Using different types of markers, increasing the number of aSTRs, combining autosomal loci with other choices of ancestry-informative markers [84–86], or sampling individuals from different geographic regions in Lebanon could have the potential to identify the presence of subclusters or resolve population structure (if exists) within the Lebanese population.

4. Conclusion

To the best of our knowledge, this is the first study that characterizes the sequence variations, establishes a sequence-based allele frequency, and analyzes population genetic parameters using sequence-based data across forensically relevant aSTR markers in samples collected from the Lebanese population (Supplementary Tables 4–10). The sequence strings are described in detail in Supplementary Table 4 in accordance with the ISFG considerations and STR Sequence Guide [13,14]. The impact of the greater resolution level offered by NGS methods was reflected in the numbers of alleles (Fig. 2 and Supplementary Table 5) and values of the forensic genetic parameters (Table 2 and Supplementary Tables 8–10) when the sequence-based data was compared to the length-based information.

The distribution of populations on the MDS scatter plot, population tree, and population *STRUCTURE* were roughly correlated with each other. Even though AMOVA indicated low genetic variations among the five populations (Supplementary Table 11), the various plots (Fig. 3 and Supplementary Figs. 1 and 2) identified genetic clusters with distinctive allele frequencies.

The Lebanese individuals (1) shared a similar pattern of allele frequency variations with the Caucasian and Hispanic (partial membership in their clusters as shown in Fig. 3A and B), and (2) as expected had a relatively distant genetic relationship with the African American and Asian individuals. These observations are based on the type and number of markers used to genotype this dataset. Different patterns or cluster signals could be obtained with a larger number of STR loci or with a combination with other ancestry informative markers (AIM-SNPs) [85, 86].

To sum up, the patterns of sequence variations identified in the Lebanese population with their associated allele frequencies will be beneficial for forensic laboratories in Lebanon, across the Middle East region, and worldwide when implementing NGS technology in human identification, mixture interpretation, or paternity testing.

Funding and disclosures

This work was supported by the NIST Special Programs Office: Forensic Genetics. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce. Certain commercial software, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose. All work has been reviewed and approved by the NIST Research Protections Office. This study was determined to be "not human subjects research" (often referred to as research not involving human subjects) as defined in U.S. Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule (45 CFR 46, Subpart A), for the Protection of Human Subjects by the NIST Human Research Protections Office and therefore not subject to oversight by the NIST Institutional Review Board. IUPUI personnel, data collection, and analyses were supported by the National Institute of Justice (NIJ) award 2014-DN-BX-K031.

CRediT authorship contribution statement

Sarah Riman: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft preparation, Writing – review & editing. **Mirna Ghemrawi**: Resources, Writing – review & editing. **Lisa A. Borsuk**: Data curation, Formal analysis, Software, Writing – review & editing. **Rami Mahfouz**: Resources, Writing – review & editing. **Susan Walsh**: Resources, Writing – review & editing. **Peter M. Vallone**: Conceptualization, Funding acquisition, Resources, Supervision, Visualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

The authors would like to thank John M. Butler, Katherine B. Gettings, and Hari Iyer at NIST for critically reading the manuscript and providing helpful comments. The authors would also like to thank Carolyn R. Steffen for Sanger sequencing one of the Lebanese samples. The authors express gratitude to Doug Storts and Spencer Hermanson at Promega for providing PowerSeq 46GY System Prototype reagents and technical guidance. Many thanks to Martin Zieger at the University of Bern, Institute of Forensic Medicine for providing helpful feedback on the MDS, population trees, and STRUCTURE plots. We thank the anonymous reviewers for their thorough revision and valuable input that greatly strengthened our manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2023.102872 and at the NIST Public Data Repository at (https://doi.org/10.18434/mds2-2867).

References

[1] J.M. Butler, Short tandem repeat typing technologies used in human identity testing, Biotechniques 43 (4) (2007) (ii-v.).
[2] P. Gill, D.J. Werrett, B. Budowle, R. Guerrieri, An assessment of whether SNPs will replace STRs in national DNA databases–joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDAM), Sci. Justice 44 (1) (2004) 51–53.
[3] J.M. Butler, Genetics and genomics of core short tandem repeat loci used in human identity testing, J. Forensic Sci. 51 (2) (2006) 253–265.
[4] K. Lazaruk, P.S. Walsh, F. Oaks, D. Gilbert, B.B. Rosenblum, S. Menchen, D. Scheibler, H.M. Wenz, C. Holt, J. Wallin, Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument, Electrophoresis 19 (1) (1998) 86–93.
[5] P. de Knijff, From next generation sequencing to now generation sequencing in forensics, Forensic Sci. Int Genet 38 (2019) 175–180.
[6] A. Alonso, P.A. Barrio, P. Müller, S. Köcher, B. Berger, P. Martin, M. Bodner, S. Willuweit, W. Parson, L. Roewer, B. Budowle, Current state-of-art of STR sequencing in forensic genetics, Electrophoresis 39 (21) (2018) 2655–2668.
[7] P.R. Haddrill, Developments in forensic DNA analysis, Emerg. Top. Life Sci. 5 (3) (2021) 381–393.
[8] K.B. Gettings, L.A. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, J. King, W. Parson, C. Phillips, P.M. Vallone, STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci. Int Genet 31 (2017) 111–117.

[9] D.S.B.S. Silva, F.R. Sawitzki, M.K. Scheible, S.F. Bailey, C.S. Alho, S.A. Faith, Paternity testing using massively parallel sequencing and the PowerSeq™ AUTO/Y system for short tandem repeat sequencing, Electrophoresis 39 (21) (2018) 2669–2673.
[10] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, D. Deforce, Forensic STR analysis using massive parallel sequencing, Forensic Sci. Int.: Genet. 6 (6) (2012) 810–818.
[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 24 (2016) 86–96.
[12] H.L. Hwa, M.Y. Wu, J.C. Lee, H.I. Yin, P.M. Hsu, S.F. Li, W.L. Hwu, C.W. Su, Analysis of nondegraded and degraded DNA mixtures of close relatives using massively parallel sequencing, Leg. Med (Tokyo) 42 (2020), 101631.
[13] 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, P. Knijff, N. Morling, M. Prinz, P.M. Schneider, C. V. 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.
[14] C. Phillips, K.B. Gettings, J.L. King, D. Ballard, M. Bodner, L. Borsuk, W. Parson, "The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide, Forensic Sci. Int Genet 34 (2018) 162–169.
[15] N.M.M. Novroski, A.E. Woerner, B. Budowle, Insertion within the flanking region of the D10S1237 locus, Forensic Sci. Int.: Genet. 35 (2018) e4–e6.
[16] C. Phillips, L. Devesse, D. Ballard, L. van Weert, M. de la Puente, S. Melis, V. Álvarez Iglesias, A. Freire-Aradas, N. Oldroyd, C. Holt, D. Syndercombe Court, Á. Carracedo, M.V. Lareu, Global patterns of STR sequence variation: Sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit, Electrophoresis 39 (21) (2018) 2708–2724.
[17] K.B. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based U.S. population data for 27 autosomal STR loci, Forensic Sci. Int Genet 37 (2018) 106–115.
[18] D. Silva, F.R. Sawitzki, M.K.R. Scheible, S.F. Bailey, C.S. Alho, S.A. Faith, Genetic analysis of Southern Brazil subjects using the PowerSeq™ AUTO/Y system for short tandem repeat sequencing, Forensic Sci. Int Genet 33 (2018) 129–135.
[19] D. Silva, M.K. Scheible, S.F. Bailey, C.L. Williams, J.S. Allwood, R.S. Just, J. Schuetter, N. Skomrock, A. Minard-Smith, N. Barker-Scoggins, C. Eichman, K. Meiklejohn, S.A. Faith, Sequence-based autosomal STR characterization in four US populations using PowerSeq™ Auto/Y system, Forensic Sci. Int Genet 48 (2020), 102311.
[20] E.K. Guevara, J.U. Palo, J.L. King, M.M. Buś, S. Guillén, B. Budowle, A. Sajantila, Autosomal STR and SNP characterization of populations from the Northeastern Peruvian Andes with the ForenSeq™ DNA Signature Prep Kit, Forensic Sci. Int Genet 52 (2021), 102487.
[21] R. Moura-Neto, J.L. King, I. Mello, V. Dias, B. Crysup, A.E. Woerner, B. Budowle, R. Silva, Evaluation of Promega PowerSeq™ Auto/Y systems prototype on an admixed sample of Rio de Janeiro, Brazil: Population data, sensitivity, stutter and mixture studies, Forensic Sci. Int Genet 53 (2021), 102516.
[22] N.M.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci. Int Genet 25 (2016) 214–226.
[23] 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.
[24] L. Devesse, L. Davenport, L. Borsuk, K. Gettings, G. Mason-Buck, P.M. Vallone, D. Syndercombe Court, D. Ballard, Classification of STR allelic variation using massively parallel sequencing and assessment of flanking region power, Forensic Sci. Int Genet 48 (2020), 102356.
[25] L. Devesse, D. Ballard, L. Davenport, I. Riethorst, G. Mason-Buck, D. Syndercombe, Court, Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups, Forensic Sci. Int Genet 34 (2018) 57–61.
[26] C. Hussing, R. Bytyci, C. Huber, N. Morling, C. Børsting, The Danish STR sequence database: duplicate typing of 363 Danes with the ForenSeq™ DNA Signature Prep Kit, Int J. Leg. Med 133 (2) (2019) 325–334.
[27] A. Delest, D. Godfrin, Y. Chantrel, A. Ulus, J. Vannier, M. Faivre, C. Hollard, F. X. Laurent, Sequenced-based French population data from 169 unrelated individuals with Verogen's ForenSeq DNA signature prep kit, Forensic Sci. Int Genet 47 (2020), 102304.
[28] P.A. Barrio, P. Martín, A. Alonso, P. Müller, M. Bodner, B. Berger, W. Parson, B. Budowle, Massively parallel sequence data of 31 autosomal STR loci from 496 Spanish individuals revealed concordance with CE-STR technology and enhanced discrimination power, Forensic Sci. Int Genet 42 (2019) 49–55.
[29] Q.X. Zhang, M. Yang, Y.J. Pan, J. Zhao, B.W. Qu, F. Cheng, Y.R. Yang, Z.P. Jiao, L. Liu, J.W. Yan, Development of a massively parallel sequencing assay for investigating sequence polymorphisms of 15 short tandem repeats in a Chinese Northern Han population, Electrophoresis 39 (21) (2018) 2725–2731.
[30] Y.L. Kwon, B.M. Kim, E.Y. Lee, K.J. Shin, Massively parallel sequencing of 25 autosomal STRs including SE33 in four population groups for forensic applications, Sci. Rep. 11 (1) (2021) 4701.
[31] H.R. Dash, K. Kaitholia, R.K. Kumawat, A.K. Singh, P. Shrivastava, G. Chaubey, S. Das, Sequence variations, flanking region mutations, and allele frequency at 31 autosomal STRs in the central Indian population by next generation sequencing (NGS), Sci. Rep. 11 (1) (2021) 23238.
[32] O. Ali Alhmoudi, R.J. Jones, G.K. Tay, H. Alsafar, S. Hadi, Population genetics data for 21 autosomal STR loci for United Arab Emirates (UAE) population using next generation multiplex STR kit, Forensic Sci. Int Genet 19 (2015) 190–191.
[33] H.M. Alsafiah, A. Iyengar, S. Hadi, W.M. Alshlash, W. Goodwin, Sequence data of six unusual alleles at SE33 and D1S1656 STR Loci, Electrophoresis 39 (19) (2018) 2471–2476.
[34] Y.M. Khubrani, P. Hallast, M.A. Jobling, J.H. Wetton, Massively parallel sequencing of autosomal STRs and identity-informative SNPs highlights consanguinity in Saudi Arabia, Forensic Sci. Int Genet 43 (2019), 102164.
[35] H. Barakat, Social and political integration in Lebanon: a case of social mosaic, Middle East J. 27 (3) (1973) 301–318.
[36] E. Chouery, M.D. Coble, K.M. Strouss, J.L. Saunier, N. Jalkh, M. Medlej-Hashim, F. Ayoub, A. Mégarbané, Population genetic data for 17 STR markers from Lebanon, Leg. Med (Tokyo) 12 (6) (2010) 324–326.
[37] A. El Andari, L. Mourad, I. Mansour, Genetic variations and population data on five supplementary STR markers in Lebanon, Ann. Hum. Genet 83 (2) (2019) 82–85.
[38] A. El Andari, H. Othman, F. Taroni, I. Mansour, Population genetic data for 23 STR markers from Lebanon, Forensic Sci. Int Genet 7 (4) (2013) e108–e113.
[39] A. El Andari, M. Khazzouh, I. Mansour, Autosomal short tandem repeat variations and population genetic data on 23 systems in seven major subpopulations from Lebanon, Int J. Leg. Med 133 (2) (2019) 433–463.
[40] M. Ghemrawi, S. Riman, N. Herrick, R. Mahfouz, P. Vallone, S. Walsh, A Genetic Investigation into a Lebanese Population: From STR's to SNP's, Chemistry Indiana University-Purdue University Indianapolis (IUPUI), 2018.
[41] Promega, PowerSeq™ 46GY System Technical Manual. TM522, August 2017.
[42] S. Riman, H. Iyer, L.A. Borsuk, P.M. Vallone, Understanding the characteristics of sequence-based single-source DNA profiles, Forensic Sci. Int Genet 44 (2020), 102192.
[43] Beckman Coulter, Instructions For Use Agencourt AMPure XP PCR Purification. B37419AB, August 2016.
[44] Thermo Fisher Scientific, Quant-iT™ PicoGreen™ dsDNA Reagent and Kit User Guide. Publication Number MAN0001931, Revision A.0, March 2022.
[45] Kapa Biosystems, KAPA Hyper Prep Kit Technical Data Sheet. August 2019.
[46] Illumina, TruSeq D.N.A. PCR-Free Library Prep Reference Guide. Document # 1000000039279 v00, October 2017.
[47] Promega, PowerSeq™ Quant MS System Technical Manual. TM511, May 2020.
[48] Illumina, MiSeq System Denature and Dilute Libraries Guide. Document # 15039740 v10, February 2019.
[49] Illumina, MiSeq System Guide. Document # 15027617 v05, August 2019.
[50] A.E. Woerner, J.L. King, B. Budowle, Fast STR allele identification with STRait Razor 3.0, Forensic Sci. Int Genet 30 (2017) 18–23.
[51] C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, US population data for 29 autosomal STR loci, Forensic Science, Int.: Genet. 7 (3) (2013) e82–e83.
[52] T.I. Huszar, K.M. Kiesler, S. Riman, L.A. Borsuk, R. Lagacé, K.B. Gettings, P.M. Vallone, Data analysis options for genotyping the Precision ID GlobalFiler NGS STR Panel v2 sequencing data from four U.S. populations, Gordon Research Conference Forensic Analysis of Human DNA, West Dover, VT, 2022.
[53] 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.
[54] 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 (4) (2011) 329–332.
[55] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, A. Drummond, Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data, Bioinformatics 28 (12) (2012) 1647–1649.
[56] A. Gouy, M. Zieger, STRAF-A convenient online tool for STR data evaluation in forensic genetics, Forensic Sci. Int Genet 30 (2017) 148–151.
[57] L. Excoffier, H.E. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (3) (2010) 564–567.
[58] K.B. Gettings, D. Ballard, M. Bodner, L.A. Borsuk, J.L. King, W. Parson, C. Phillips, Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting, Forensic Sci. Int Genet 43 (2019), 102165.
[59] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2) (2000) 945–959.
[60] D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, Genetics 164 (4) (2003) 1567–1587.
[61] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, A. Carracedo, M.V. Lareu, An overview of STRUCTURE: applications, parameter settings, and supporting software, Front Genet 4 (2013) 98.
[62] M.J. Hubisz, D. Falush, M. Stephens, J.K. Pritchard, Inferring weak population structure with the assistance of sample group information, Mol. Ecol. Resour. 9 (5) (2009) 1322–1332.
[63] Clumpak - Cluster Markov Packager Across K.). http://clumpak.tau.ac.il/.
[64] N.M. Kopelman, J. Mayzel, M. Jakobsson, N.A. Rosenberg, I. Mayrose, Clumpak: a program for identifying clustering modes and packaging population structure inferences across K, Mol. Ecol. Resour. 15 (5) (2015) 1179–1191.

[65] N.A. Rosenberg, DISTRUCT: a program for the graphical display of population structure, Mol. Ecol. Notes 4 (1) (2004) 137–138.
[66] N.M. Kopelman, L. Stone, C. Wang, D. Gefel, M.W. Feldman, J. Hillel, N. A. Rosenberg, Genomic microsatellites identify shared Jewish ancestry intermediate between Middle Eastern and European populations, BMC Genet. 10 (1) (2009) 80.
[67] T.I. Huszar, M.A. Jobling, J.H. Wetton, A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing, Forensic Sci. Int Genet 35 (2018) 97–106.
[68] J. Hoogenboom, K.J. van der Gaag, R.H. de Leeuw, T. Sijen, P. de Knijff, J.F. Laros, FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise, Forensic Sci. Int Genet 27 (2017) 27–40.
[69] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (1) (2001) 308–311.
[70] J.T. den Dunnen, Sequence Variant Descriptions: HGVS Nomenclature and Mutalyzer, Curr. Protoc. Hum. Genet 90 (2016) 7.13.1–7.13.19.
[71] M.D. Edge, B.F.B. Algee-Hewitt, T.J. Pemberton, J.Z. Li, N.A. Rosenberg, Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets, Proc. Natl. Acad. Sci. USA 114 (22) (2017) 5671–5676.
[72] K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR allele sequence variation: Current knowledge and future issues, Forensic Sci. Int Genet 18 (2015) 118–130.
[73] M. Nei, Molecular Evolutionary Genetics, Columbia University Press,, 1987.
[74] C.M.L. Serrote, L.R.S. Reiniger, K.B. Silva, S. Rabaiolli, C.M. Stefanel, Determining the Polymorphism Information Content of a molecular marker, Gene 726 (2020), 144175.
[75] S. Oliveira, A. Trindade-Filho, C. Mendes, K. Paula, F. Maia, H. Pak, G. Dalton, Power of exclusion of 18 autossomic STR loci in a Brazilian Middle–West region population sample, Int. Congr. Ser. 1288 (2006) 433–435.
[76] R.A. Fisher, 243: Standard Calculations for Evaluating a Blood-Group System, 1951.
[77] A. Gouy, M. Zieger, The STRAF Book. 2022. <⟨https://agouy.github.io/straf_book/forensic-parameters.html⟩>
[78] A. Staadig, A. Tillmar, An overall limited effect on the weight-of-evidence when taking STR DNA sequence polymorphism into account in kinship analysis, Forensic Sci. Int Genet 39 (2019) 44–49.
[79] M. Slatkin, Linkage disequilibrium—understanding the evolutionary past and mapping the medical future, Nat. Rev. Genet. 9 (6) (2008) 477–485.
[80] K. Bryc, C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds, A. Auton, M. Hammer, C.D. Bustamante, H. Ostrer, Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations, Proc. Natl. Acad. Sci. USA 107 (Suppl 2) (2010) 8954–8961.
[81] T.S. Frontanilla, G. Valle-Silva, J. Ayala, C.T. Mendes-Junior, Open-Access Worldwide Population STR Database Constructed Using High-Coverage Massively Parallel Sequencing Data Obtained from the 1000 Genomes Project, Genes 13 (12) (2022) 2205.
[82] N.A. Rosenberg, A population-genetic perspective on the similarities and differences among worldwide human populations, Hum. Biol. 83 (6) (2011) 659.
[83] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L. A. Zhivotovsky, M.W. Feldman, Genetic structure of human populations, Science 298 (5602) (2002) 2381–2385.
[84] S. Ramachandran, N.A. Rosenberg, L.A. Zhivotovsky, M.W. Feldman, Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites, Hum. Genom. 1 (2) (2004) 87–97.
[85] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet 73 (6) (2003) 1402–1422.
[86] C. Phillips, L. Fernandez-Formoso, M. Garcia-Magariños, 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 (3) (2011) 155–169.