Published in final edited form as:
Forensic Sci Int Genet. 2018 November ; 37: 106–115. doi:10.1016/j.fsigen.2018.07.013.

Sequence-based U.S. population data for 27 autosomal STR loci

Katherine Butler Gettings[a,*], **Lisa A. Borsuk**[a], **Carolyn R. Steffen**[a], **Kevin M. Kiesler**[a], **Peter M. Vallone**[a]

[a]U.S. National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD 20899, USA

Abstract

This manuscript reports Short Tandem Repeat (STR) sequence-based allele frequencies for 1,036 samples across 27 autosomal STR loci: D1S1656, TPOX, D2S441, D2S1338, D3S1358, D4S2408, FGA, D5S818, CSF1PO, D6S1043, D7S820, D8S1179, D9S1122, D10S1248, TH01, vWA, D12S391, D13S317, Penta E, D16S539, D17S1301, D18S51, D19S433, D20S482, D21S11, Penta D, and D22S1045. Sequence data was analyzed by two bioinformatic pipelines and all samples have been evaluated for concordance with alleles derived from CE-based analysis at all loci. Each reported sequence includes high-quality flanking sequence and is properly formatted according to the most recent guidance of the International Society for Forensic Genetics. In addition, GenBank accession numbers are reported for each sequence, and associated records are available in the STRSeq BioProject (https://www.ncbi.nlm.nih.gov/bioproject/380127). The D3S1358 locus demonstrates the greatest average increase in heterozygosity across populations (approximately 10 percentage points). Loci demonstrating average increase in heterozygosity from 10 to 5 percentage points include (in descending order) D9S1122, D13S317, D8S1179, D21S11, D5S818, D12S391, and D2S441. The remaining 19 loci each demonstrate less than 5 percentage point increase in average heterozygosity. Discussion includes the utility of this data in understanding traditional CE results, such as informing stutter models and understanding migration challenges, and considerations for population sampling strategies in light of the marked increase in rare alleles for several of the sequence-based STR loci. This NIST 1036 data set is expected to support the implementation of STR sequencing forensic casework by providing high-confidence sequence-based allele frequencies for the same sample set which are already the basis for population statistics in many U.S. forensic laboratories.

1. Introduction

In forensic casework, Short Tandem Repeat (STR) allele frequency data is used to calculate statistical weight when a person of interest has been included as a possible contributor of

[*]Corresponding Author's Contact Information: Katherine Butler Gettings, National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA, Phone: 301-975-6401, Fax: 301-975-8505, katherine.gettings@nist.gov.

Conflict of interest
None

Supplementary Materials
All supplementary materials are accessible via https://doi.org/10.18434/T4/1500024.

genetic material recovered from an item of evidence. This statistical weight should be derived from the same level of information which was used to include the person of interest, e.g. a *sequence-based* STR profile inclusion requires a *sequence-based* STR match statistic. Therefore, implementation of STR sequencing into forensic casework requires the existence of appropriate allele frequency databases.

Several recent publications have reported sequence-based allele frequency data for autosomal STR loci [1–6]. This manuscript reports high-confidence autosomal STR sequence-based allele frequencies for N = 1036 across 27 autosomal STR loci: D1S1656, TPOX, D2S441, D2S1338, D3S1358, D4S2408, FGA, D5S818, CSF1PO, D6S1043, D7S820, D8S1179, D9S1122, D10S1248, TH01, vWA, D12S391, D13S317, Penta E, D16S539, D17S1301, D18S51, D19S433, D20S482, D21S11, Penta D, and D22S1045. The factors which lend increased confidence to this data set include: high sequence coverage requirement; analysis with two bioinformatic pipelines; reporting of high-quality flanking sequence; comparison to CE allele calls for every sample at every locus; disallowance of dropout; and additional confirmation of all null alleles, CE discordances, isoalleles (alleles of the same length but different sequence), and sequences only observed once.

The preceding “NIST 1036” length-based data set [7, 8] reported allele frequencies for 29 autosomal STR loci, 24 of which are reported by sequence in this manuscript. The five loci reported in the length-based NIST 1036 and not reported in this manuscript are: F13A01, F13B, FESFPS, LPL, Penta C, and SE33. The sequence-based frequency data for NIST 1036 at SE33 are published separately [9]. Four additional loci are reported in this manuscript which were not present in the original data set: D4S2408, D9S1122, D17S1301, D20S482. CE data for these four loci was previously published on a subset of NIST 1036 [10], and the remaining samples were analyzed by CE in the course of this study.

As in the preceding NIST 1036 data set, this sequence-based autosomal STR data set is applicable to data derived from any current or future sequencing method, insofar as the loci and genomic coordinates interrogated overlap those reported here. Additionally, this data set serves as the foundation for STRSeq [11]; each sequence reported herein is associated with a GenBank accession number which is accessible via the STRSeq BioProject (https://www.ncbi.nlm.nih.gov/bioproject/380127), formatted according to the most recent guidance of the International Society for Forensic Genetics (ISFG) DNA commission on minimal nomenclature requirements [12] and the STR sequence working group [13].

2. Materials and Methods

Anonymous liquid blood samples with self-reported ancestries were purchased from Interstate Blood Bank (Memphis, TN) and Millennium Biotech, Inc. (Ft. Lauderdale, FL) or provided by DNA Diagnostics Center (Fairfield, OH) as buccal swabs from paternity testing samples anonymous to NIST. A total of 1,036 samples were included in this study, the same samples previously reported in [7, 8], divided among four U.S. populations: African American (AfAm, N = 342), Asian (N = 97), Caucasian (Cauc, N = 361), and Hispanic (Hisp, N = 236). Throughout this paper, N = 1036 is used to reference number of samples, whereas 2,072 is the implied number of chromosomes. All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.

See Supplementary File S1 for a full description of materials and methods. In brief, sequence data was generated using the ForenSeq DNA Signature Prep Kit on a MiSeq FGx instrument (Illumina, San Diego, CA), optimized to obtain a minimum of 30X sequence coverage per allele. Two bioinformatic analyses were performed: 1) the Illumina Universal Analysis Software (UAS)-processed .txt files were parsed off-platform with custom batch-enabled scripts and 2) the FASTQ files were analyzed with a custom pipeline based on STRait Razor 2.0 [14]. Length-based allele calls from both bioinformatic pipelines were compared to CE allele calls for all samples/loci (previously published [7, 15] and unpublished data). Any differences observed within these evaluations were investigated and arbitrated (see [8] and Supplementary File S1), with maximal information reported in the final data set.

Allele frequencies, observed heterozygosity (Hobs), expected heterozygosity (Hexp), and testing for Hardy-Weinberg Equilibrium (HWE) were all performed using STRAF [16]. Allele frequencies were cross-validated with Microsoft Excel. Linkage disequilibrium (LD) was evaluated with Arlequin [17].

3. Results

Sequencing Results

Sequencing metrics for the 42 sequencing runs reported in this study, including cluster density, percent clusters passing filter, percent phasing and pre-phasing, average coverage, standard deviation and coefficient of variance, are found in Supplementary Table S1. Coverage values obtained from the UAS-processed .txt files for all 1036 samples across all STR, YSTR and XSTR loci are presented in a heatmap-style view in Supplementary Table S2 (YSTR and XSTR coverage data included for comparative purposes). All sequence coverage values for the 27 autosomal STR loci are greater than or equal to 30X (per heterozygous and homozygous allele) except for one sample allele at the Penta D locus reported at 27X and one sample allele at the D22S1045 locus reported at 29X. Detailed conditions and troubleshooting of locus or sample specific issues are presented in Supplementary File S1.

Allele Frequencies

Supplementary Table S3 lists the sequence counts and frequencies for the 27 autosomal loci for the full set of samples (All, N = 1036) and by population group: Caucasian (Cauc, N = 361), African American (AfAm, N = 342), Hispanic (Hisp, N = 236), and Asian (N = 97). As this supplementary table is both wide and lengthy, Figure 1 illustrates the information contained in Supplementary Table S3 with three example sequences. The GRCh38 coordinates reported in this study, including hyperlinks to the chromosome reference sequences constituting this assembly, are given in Supplementary Table S4.

Three tri-allelic genotypes are omitted from the frequency data in Supplementary Table S3: one at TPOX in the African American population, one at D9S1122 in the Hispanic population, and one at Penta D in the Hispanic population (more details on these three samples are given in Supplementary File S1). These omissions result in a decrease of one from the above listed sample numbers for these populations/loci, as well as the full set.

For the range of sequence reported (see Supplementary Table S4), 16 loci contained a total of 40 flanking region polymorphisms in this data set, in descending order: eight at D13S317, five at Penta D, four each at D7S820 and D22S1045, three each at vWA and D16S539, two each at D6S1043 and D19S433, and one each at D1S1656, D2S441, D5S818, CSF1PO, D9S1122, TH01, D12S391, D18S51, and D20S482. Eight of these polymorphisms are insertion-deletions (indels) and 32 are single nucleotide polymorphisms (SNPs). Frequency data for these polymorphisms in association with repeat region alleles is found in Supplementary Table S3. Additionally, Supplementary Table S5 summarizes the 40 flanking region polymorphisms reported in this study, in leftmost (5’) orientation to GRCh38 and including hyperlinks to associated dbSNP rs numbers. Five polymorphisms were submitted to dbSNP in the course of this study. Two have been assigned rs numbers in dbSNP build 151 and three have been assigned temporary ss numbers (included in Supplementary Table S5) which will redirect to rs numbers upon release of dbSNP build 152. Interestingly, one of these submissions identifies the 13 bp deletion at the Penta D locus, which results in the 2.2 length-based allele observed at > 11 % frequency in the African American population. It is expected that this deletion has been detected in prior large sequencing studies; however, not registered in dbSNP due to alignment challenges resulting from adjacency to the repeat region.

While Supplementary Table S3 contains the full information that is needed for generating inclusion statistics, for the purpose of discussion, the data have been condensed into several *high-level* formats. Figure 2 displays the allele frequency distribution per locus, by sequence and by length in N = 1036. The first nine alleles at each locus are presented in color to facilitate comparisons within and across loci. Loci are sorted in ascending order of sequence-based frequency, such that the locus with the lowest frequency most common allele (SE33, 8.78 %) is in the upper left corner, and the locus with the highest frequency most common allele is in the lower right corner (TPOX, 46.6 %). The SE33 locus is included in this figure for comparative purposes; it is not reported in the UAS but was recovered from the FASTQ files using a modified version of the custom pipeline. Due to the considerable effort required to recover and evaluate this locus, the SE33 sequence-based frequency data for this sample set are published separately [9].

For loci which contain varying sequence motifs, Table 1 captures the motif frequency by population. The criterion used to generate this table was the presence of more than one motif at ≥ 1 % motif frequency in at least one population. Nineteen of the 27 autosomal loci met this criterion. For these 19 loci in Table 1, any motif frequencies which did not reach 1 % in at least one population were combined and reported in the table as “all other motifs”. Variable stretches in the motifs are denoted with “n”, and uncounted bases within the repeat region are denoted by lower case text. Flanking region polymorphisms as compared to GRCh38 are indicated by appending the rs number to the motif. Several examples of apparent population-specific increased frequencies are present. Three loci have motif frequencies over 20 percentage points higher in one population compared to all other populations, and these frequencies are bolded in Table 1: D3S1358 in the African American population, D4S2408 in the Asian population, and D16S539 in the Asian population. The first two examples are repeat region variant motifs, while the third is a flanking region SNP.

Population Genetics

Table 2 contains a comparison of observed heterozygosity (Hobs) by length (previously published in [8, 15]) and sequence for the 27 auSTR loci, sorted by greatest to least average gain across populations. D3S1358 demonstrates the greatest overall gain across populations, at approximately 10 percentage points average increase in heterozygosity. Loci demonstrating average increase in heterozygosity from 10 to 5 percentage points include (in descending order) D9S1122, D13S317, D8S1179, D21S11, D5S818, D12S391, and D2S441. The remaining 19 loci demonstrate less than 5 percentage points increase in average heterozygosity.

Supplementary Table S6 contains results of analysis by locus for Hexp, Hobs, and the p-values (pHW) associated with evaluation of HWE for each locus and population. Several locus/population combinations appear to significantly deviate from the frequency expectations under HWE at $\alpha = 0.05$, however after correction for multiple testing, these values are no longer significant. When comparing 23 loci with previously published length-based data for these N = 1036 samples [8], a similar number of loci appear to significantly deviate from HWE at $\alpha = 0.05$ by sequence and by length (five loci by sequence: FGA, D5S818, D12S391, D18S51, D22S1045; four loci by length [8]: D2S441, FGA, D6S1043, D22S1045). Comparing all 27 loci to published sequence-based data for similar populations, four locus/population combinations have consistently low p-values in this study and in Novroski, et al. [3]: D4S2408 and D5S818 in Hispanic samples, D13S317 in Caucasian samples, and D16S539 in African American samples.

Supplementary Table S7 contains p-values associated with LD evaluation in each population for the six syntenic pairs of loci requiring consideration: TPOX-D2S441, D2S441-D2S1338, D4S2408-FGA, D5S818-CSF1PO, vWA-D12S391, and D21S11-Penta D, and their physical distances along the chromosomes. The p-values for two pairings are significant in the Hispanic population at $\alpha = 0.05$ (D2S441-D2S1338, p-value = 0.0205 and D5S818-CSF1PO, p-value = 0.0401, bolded), but are no longer considered significant when correcting for 24 tests (six syntenic pairs across four populations, $\alpha = 0.0021$). Comparing these same pairings and populations to another study, two different locus pairings demonstrated significant evidence of LD at $\alpha = 0.05$ (Hispanic population, TPOX-D2S441, p-value = 0.0044 and Asian population, D4S2408-FGA, p-value = 0.0034) [3]; but these were also not significant after correction for multiple testing.

Supplementary Table S8 contains pairwise Fst values at all loci for the six pairings of populations in this study, sorted by highest to lowest average Fst value. This value reflects differences in allele frequency distributions and heterozygosity between populations, and is used as an indicator of the ancestry-informative value of the locus. A higher Fst indicates greater differences in frequency distributions between the populations. The maximal Fst in this study, 4.42 %, was observed for the TH01 locus when comparing the Caucasian and Asian populations. This finding is likely due to the TH01 9.3 allele which, in this study, was detected at 4 % frequency in the Asian population and 35 % frequency in the Caucasian population. It should be noted that there is negligible sequence diversity at the TH01 locus, and the 9.3 microvariant has been detectable by length since the early use of this locus as a forensic marker [18]. Therefore, our finding is only relevant in the context that no other locus gains more ancestry information by this measure than what was already present by length at TH01. The locus with the next-highest average pairwise Fst is D4S2408, and this is expected to owe largely to the sequence variant of the 9 allele, which has a substantially higher frequency in the Asian population compared to other populations in this study (previously mentioned, and bolded in Table 1). Other loci ranking relatively high in average Fst lack repeat region variation but contain flanking region polymorphisms which may be marginally ancestry-informative (D13S317-rs9546005 and D16S539-rs11642858). Overall, for the populations and range of sequence reported here, it is unlikely that these loci will be of primary use in ancestry prediction applications.

4. Discussion

Trends in STR Allelic Gains by Sequence

Some loci experience substantial gains in alleles due to variation in repeat region sequence, others demonstrate greater allelic diversity due to flanking region polymorphisms, while many of these loci do not demonstrate appreciable gains by sequence at all. Loci with an increase in repeat region alleles typically contain multiple varying subunits. For example, D3S1358, which exhibits the greatest increase in heterozygosity, is composed of a static TCTA subunit, followed by a TCTG observed from one to four times, and a final TCTA composed of eight to 17 repeats. For maximal repeat region gains, it is not only important to have multiple varying subunits, but also for that variation to be well-distributed in the population. The distribution shown in the Figure 3 histograms for D3S1358 exemplifies the gain as overlapping length-based alleles are differentiated by sequence. Other loci with similar patterns and gains include D2S1338, vWA, D12S391, and D21S11. Figure 3 also supports the motif-specific population information shown in Table 1, e.g. TCTA TCTG [TCTA]n is more common in African American population (42.4 % motif frequency) and TCTA [TCTG]2 [TCTA]n is more common in Asian population (71.1 % motif frequency).

D13S317 is an example of a locus where allelic gains derive from flanking sequence variation. In Figure 4, the D13S317 repeat region is held constant, as it is largely shown to be in this data set, while the eight flanking region variants identified in this data set are located and enumerated. The 3’ flanking region is particularly challenging to format, due to the high frequency across populations (30 % to 58 %) of the A > T SNP adjacent to the repeat, which results in the appearance of an additional TATC repeat unit (rs9546005, numbered 3 in Figure 4), and two 4 bp deletions which can lead to apparent discordance between CE- and sequence-based numerical alleles. The more common 4 bp deletion, observed 11 times in this study, had previously been submitted to dbSNP and identified as rs561167308. The less common 4 bp deletion, observed four times in this study, was submitted to dbSNP by the authors and has been identified in dbSNP build 151 as rs1442523705. Other loci demonstrating substantial gains due to flanking region polymorphisms in this data set include D5S818, D7S820, and D16S539.

While an increase in alleles is generally correlated to an increase in heterozygosity, factors such as the frequency distribution of the alleles and the heterozygosity by length will affect the gains in heterozygosity by sequence. Figure 5 displays a broad overview of the increases, or lack thereof, by locus for N = 1036. Five loci have sequence-based heterozygosity exceeding 90 % (in descending order): D12S391, D2S1338, D1S1656, D21S11 and Penta E. Interestingly, Penta E was the only locus that exceeded 90 % heterozygosity by length, and the gains by sequence are modest at this locus. While D1S1656 has a relatively low number of alleles by both length and sequence, the high heterozygosity values indicate these alleles are well-distributed in the population. This contrasts with D21S11, which has three times as many alleles by sequence as D1S1656, but has a lower heterozygosity, indicating an abundance of rare alleles. The previously discussed D13S317 and D3S1358 are adjacent in Figure 5, due to their similar sequence-based heterozygosity of nearly 86 %. While the increase in alleles was higher for D13S317, the gain in heterozygosity was higher for D3S1358. This is likely because the length-based heterozygosity for D3S1358 was lower to start; essentially, this locus has more room for improvement.

Utility of STR Sequence Data in Traditional CE Analysis

The primary purpose of this work is to facilitate technology transition; however, some interesting trends in this sequence-based data set may serve to enlighten traditional CE analysis. For example, as shown in Table 1, 81 % to 94 % of alleles by population at the vWA locus follow the repeat region pattern [TAGA]n [CAGA]3–6 TAGA. However, one sequence variant of the 14 and, more rarely, 15 allele, represents 3 % to 19 % of alleles by population. This sequence variant has an *interrupted* version of the standard pattern (interrupting tetranucleotides bolded): [TAGA]3 **TGGA** [TAGA]3 [CAGA]4 **TAGA** CAGA TAGA. This sequence variant may confound CE-based stutter models, if the models predict stutter levels based on the more common motif.

In addition, knowledge of base composition may aid understanding of migration issues in CE data. Specifically, microvariant alleles x.1 and x.3 at the D12S391 locus may be challenging to separate from neighboring integer alleles by CE [19]. Based on the allele frequency data in Supplementary Table S3, the most likely example of this is an 18.3, 19 combination. While this data does not contain sequence variants of the 18.3 allele, there are eight different sequence-based versions of the 19 allele in N = 1036, among which there are five different base composition distributions. Minor differences in molecular weight between same-length fragments typically have negligible impact on CE allele designation due to virtual allele binning in the associated software. However, the ability to resolve differences of one base at D12S391, particularly in a disproportionate mixture sample, may depend on the base composition of the alleles. Knowledge of the sequence may be useful in designing the most challenging combinations for CE validation purposes.

Population Sample Sizes for Highly Polymorphic Loci

Statistics used in forensic DNA analysis rest on the premise that a subset of the population can be used to generate accurate allele frequencies, which can then be used to assign probabilities to genotypes derived from evidence profiles. This circularity is anchored in conformity of the population data genotype frequencies to Hardy Weinberg proportions expected under random mating, and the results of HWE evaluation in this study are similar to the results of this evaluation by length for the same samples [comparison shown in Supplementary Table S7]. A further statistical question which arises when considering use of loci with very large numbers of alleles is the appropriate number of samples from which to derive allele frequencies. Intuitively, a more polymorphic system should require more samples; however, it is important to consider the history of length-based STR population sampling guidance.

The current, commonly used approach for estimating the frequency of rare variants in length-based STR analysis in U.S. forensic laboratories is the minimum allele frequency of 5/2N. This was recommended by the National Research Council in 1996 (commonly known as the NRC II, [20]) in the context of highly polymorphic VNTR loci. Page 48 states: "It is common in some statistical tests to pool very rare classes, and that is what the FBI has done by *rebinning*. If a bin in the database contains fewer than five entries, it is pooled with adjacent bins so that no bin has fewer than five. We recommend this procedure for VNTRs and for other systems in which an allele is represented fewer than five times in the database."

Further, the sample size requirements for forensic DNA typing of VNTRs was addressed in a seminal 1992 paper by Chakraborty [21], and such information may be applicable to the highly polymorphic sequence-based STR loci. Indeed, Chakraborty's description of VNTRs could equally apply to the data in this manuscript: "Based on the population genetic characteristics of the hypervariable loci, I show that the large heterozygosities at such loci necessarily imply that the expected number of alleles at each of these loci is generally quite large (often larger than 50) and that there is a predominance of rare alleles (i.e., alleles that occur in frequencies as small as 0.01) at such loci. Furthermore, the total number of alleles and the number of rare alleles are increasing functions of sample size."

VNTR allele distributions are described by Chakraborty as follows: "for most VNTR loci, even when the total number of alleles is large, the expected number of alleles having frequency p or above is generally below 10 for $p = 0.001$, 0.01, or 0.05." In our data set of 27 sequence-based STR loci, all loci are consistent this description at $p = 0.05$ (having less than 10 alleles with frequency p or above). However, only 10 loci match this description at $p = 0.01$ in the combined set of N = 1036, and this drops to five loci at $p = 0.001$. For several loci, this finding is unrelated to sequence information; more than the expected number of alleles at the given p are observable by length. There appear to be five loci for which sequencing will substantially increase the number of alleles beyond the expected range at the given p: D1S1656, D2S1338, D8S1179, D12S391, and D21S11. Of note is that the variants at these loci are primarily a product of repeat region sequence variation, rather than flanking region variation. It is possible that traditional estimates of sample numbers may not be appropriate for these loci; however, we present this information as an invitation for further analysis and discussion.

5. Conclusions

This NIST 1036 data set is expected to support the implementation of STR sequencing forensic casework by providing high-confidence sequence-based allele frequencies for the same sample set which are already the basis for population statistics in many U.S. forensic laboratories. Each sequence is properly formatted according to the most recent guidance of the ISFG. In addition, GenBank accession numbers and associated records are available for each sequence in the STRSeq BioProject (https://www.ncbi.nlm.nih.gov/bioproject/380127) [11].

As previously stated herein, the statistical weight should be derived from the same level of information which was used to include the person of interest, and this includes the extent of flanking region used for interpretation. It is more challenging for a laboratory to implement this set of sequence-based allele frequency data if the laboratory interprets a different range of sequence than is reported here. It is possible to truncate this sequence-based allele frequency data set if it represents a larger range than the laboratory's desired reporting range. However, the inverse is not true. If this sequence-based allele frequency data set is of a smaller range than the laboratory's desired reporting range, then the laboratory's reported range should be truncated to match the frequency data (or a more expansive frequency data set should be used).

The focus of this manuscript is the frequency data and use cases thereof, rather than performance of the ForenSeq DNA Signature Prep Kit, the primary assay used to generate the data. While more detailed information about optimization and performance is included in Supplementary File S1, this information is presented from a quality control perspective, applying only to this single-source data set where full CE supporting data was available. The analysis parameters described are not intended to be used as validation guidance for unknown samples.

The sequence allele frequencies reported here are limited to the 27 autosomal STR loci designated by the ForenSeq DNA Signature Prep Kit. Future publications will address the YSTR and XSTR loci included in this assay (also with full CE supporting data), as well as the identity, ancestry and phenotype SNP markers included in this assay. A concurrent publication details the NIST 1036 allele frequencies for the SE33 locus [9]. A subset of NIST 1036 were previously sequenced at 22 autosomal STR loci with the PowerSeq Auto prototype assay (Promega) [22]. Additional subsets have been sequenced with PowerSeq 46GY (Promega) and are soon-to-be sequenced with PrecisionID GlobalFiler NGS (Applied Biosystems), and these results will be included in future publications and in the STRSeq BioProject. Using the NIST 1036 set to evaluate concordance across assays, platforms, and bioinformatic pipelines continually adds the extensive forensic marker information available for these samples, and maintains their relevance in facilitating technology transition.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Jonathan King for his guidance in developing the STRaitRazor-based pipeline; John Butler for manuscript feedback; John McGuigan for feedback on allele formatting; and Chris Phillips, Jonathan King, David Ballard, Martin Bodner, and Walther Parson, for many lengthy nomenclature discussions.

Funding and Disclaimer

This work was funded in part by the Federal Bureau of Investigation (FBI) interagency agreements DJF-15-1200-0000174 and DJF-16-1200-0000246: “Forensic DNA Typing as a Biometric Tool”. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce or the Department of Justice. Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose. All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.

References

[1]. Wendt FR, King JL, Novroski NM, Churchill JD, Ng J, Oldt RF, McCulloh KL, Weise JA, Smith DG, Kanthaswamy S, Budowle B, Flanking region variation of ForenSeq DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans, Forensic Sci Int Genet 28 (2017) 146–154. [PubMed: 28273507]

[2]. Wendt FR, Churchill JD, Novroski NM, King JL, Ng J, Oldt RF, McCulloh KL, Weise JA, Smith DG, Kanthaswamy S, Budowle B, 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. [PubMed: 27243782]

[3]. Novroski NM, King JL, Churchill JD, Seah LH, Budowle B, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci Int Genet 25 (2016) 214–226. [PubMed: 27697609]

[4]. Friis SL, Buchard A, Rockenbauer E, Borsting C, Morling N, 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. [PubMed: 26722765]

[5]. Devesse L, Ballard D, Davenport L, Riethorst I, Mason-Buck G, Court DS, Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups, Forensic Science International: Genetics (2017).

[6]. van der Gaag KJ, de Leeuw RH, Hoogenboom J, Patel J, Storts DR, Laros JF, de Knijff P, 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. [PubMed: 27347657]

[7]. Hill CR, Duewer DL, Kline MC, Coble MD, Butler JM, population data for US 29 autosomal STR loci, Forensic Sci Int Genet 7(3) (2013) e82–3. [PubMed: 23317915]

[8]. Steffen CR, Coble MD, Gettings KB, Vallone PM, Corrigendum to ‘U.S. Population Data for 29 Autosomal STR Loci’ [Forensic Sci. Int. Genet. 7 (2013) e82-e83], Forensic Sci Int Genet 31 (2017) e36–e40. [PubMed: 28867528]

[9]. Borsuk L, Gettings KB, Steffen CR, Kiesler KM, Vallone PM, Sequence-based US population data for the SE33 locus, Electrophoresis (2018).

[10]. Hill CR, Butler JM, Vallone PM, A 26plex Autosomal STR Assay to Aid Human Identity Testing*, J.Forensic Sci. (2009).

[11]. Gettings KB, Borsuk LA, Ballard D, Bodner M, Budowle B, Devesse L, King J, Parson W, Phillips C, Vallone PM, STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci Int Genet 31 (2017) 111–117. [PubMed: 28888135]

[12]. Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmao L, Hares DR, Irwin JA, King JL, Knijff P, Morling N, Prinz M, Schneider PM, Neste CV, Willuweit S, Phillips C, 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. [PubMed: 26844919]

[13]. Phillips C, Gettings KB, King JL, Ballard D, Bodner M, Borsuk L, Parson W, “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. [PubMed: 29486434]

[14]. Warshauer DH, King JL, Budowle B, STRait Razor v2.0: the improved STR Allele Identification Tool--Razor, Forensic Sci Int Genet 14 (2015) 182–6. [PubMed: 25450790]

[15]. Hill CR, Kline MC, Coble MD, Butler JM, Characterization of 26 MiniSTR Loci for Improved Analysis of Degraded DNA Samples, J Forensic Sci 53(1) (2008) 73–80. [PubMed: 18005005]

[16]. Gouy A, Zieger M, STRAF-A convenient online tool for STR data evaluation in forensic genetics, Forensic Sci Int Genet 30 (2017) 148–151. [PubMed: 28743032]

[17]. Excoffier L, Lischer HEL, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Molecular Ecology Resources 10(3) (2010) 564–567. [PubMed: 21565059]

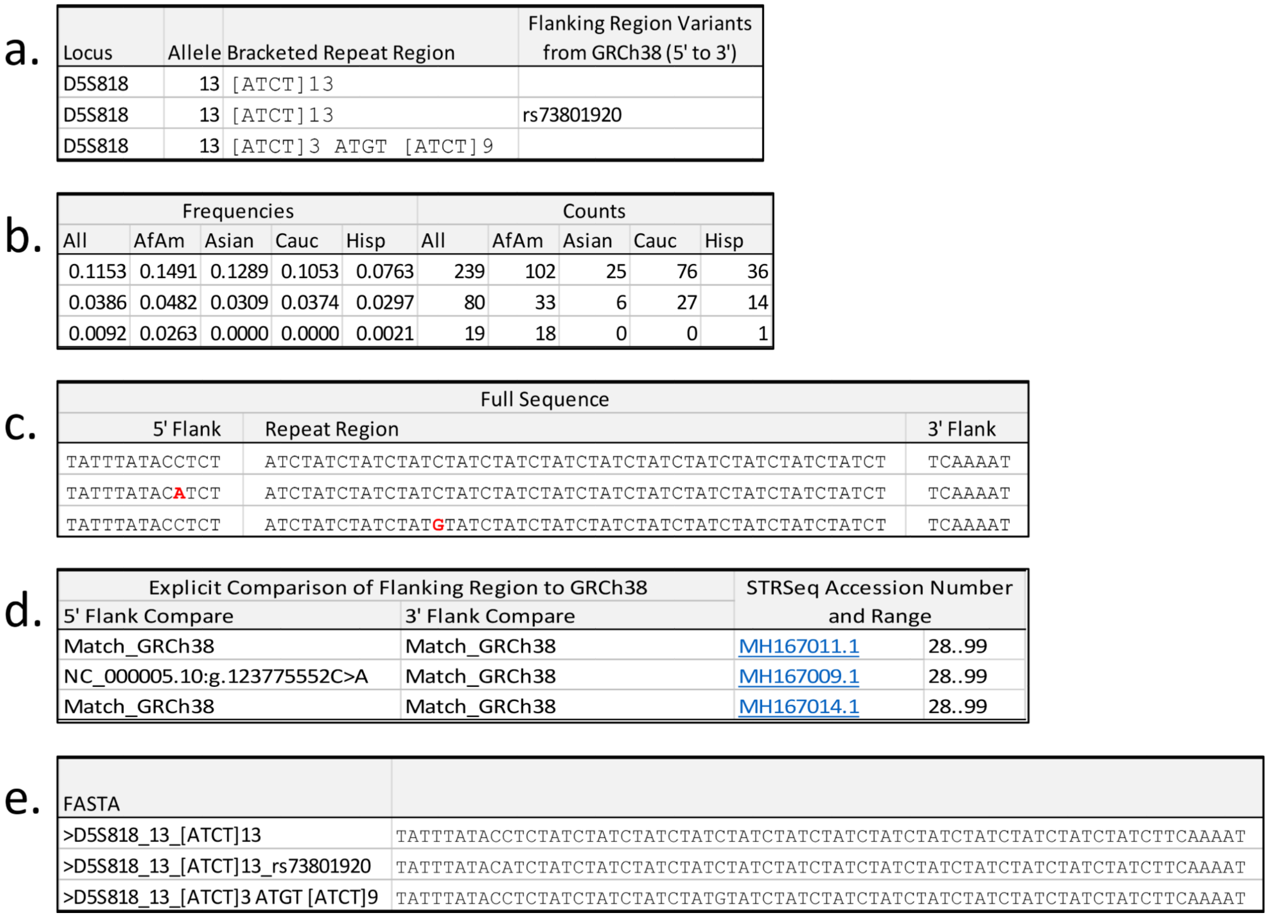
[18]. Puers C, Hammond HA, Jin L, Caskey CT, Schumm JW, Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01[AATG]n and reassignment of alleles in population analysis by using a locus-specific allelic ladder, Am.J.Hum.Genet. 53 (1993) 953–958. [PubMed: 8105685]

[19]. Dalsgaard S, Rockenbauer E, Buchard A, Mogensen HS, Frank-Hansen R, Borsting C, Morling N, Non-uniform phenotyping of D12S391 resolved by second generation sequencing, Forensic Sci Int Genet 8(1) (2014) 195–9. [PubMed: 24315608]

[20]. Council NR, The Evaluation of Forensic DNA Evidence, The National Academies Press, Washington, DC, 1996.

[21]. Chakraborty R, Sample size requirements for addressing the population genetic issues of forensic use of DNA typing, Hum Biol 64(2) (1992) 141–59. [PubMed: 1559686]

[22]. Gettings KB, Kiesler KM, Faith SA, Montano E, Baker CH, Young BA, Guerrieri RA, Vallone PM, Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Forensic Sci Int Genet 21 (2016) 15–21. [PubMed: 26701720]



a.

Locus	Allele	Bracketed Repeat Region	Flanking Region Variants from GRCh38 (5' to 3')
D5S818	13	[ATCT]13	
D5S818	13	[ATCT]13	rs73801920
D5S818	13	[ATCT]3 ATGT [ATCT]9	

b.

Frequencies					Counts				
All	AfAm	Asian	Cauc	Hisp	All	AfAm	Asian	Cauc	Hisp
0.1153	0.1491	0.1289	0.1053	0.0763	239	102	25	76	36
0.0386	0.0482	0.0309	0.0374	0.0297	80	33	6	27	14
0.0092	0.0263	0.0000	0.0000	0.0021	19	18	0	0	1

c.

Full Sequence		
5' Flank	Repeat Region	3' Flank
TATTTATACCTCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
TATTTATAC**A**TCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
TATTTATACCTCT	ATCTATCTATCTAT**G**TATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT

d.

Explicit Comparison of Flanking Region to GRCh38		STRSeq Accession Number and Range	
5' Flank Compare	3' Flank Compare		
Match_GRCh38	Match_GRCh38	MH167011.1	28..99
NC_000005.10:g.123775552C>A	Match_GRCh38	MH167009.1	28..99
Match_GRCh38	Match_GRCh38	MH167014.1	28..99

e.

FASTA	
>D5S818_13_[ATCT]13	TATTTATACCTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTTCAAAAT
>D5S818_13_[ATCT]13_rs73801920	TATTTATACATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTTCAAAAT
>D5S818_13_[ATCT]3 ATGT [ATCT]9	TATTTATACCTCTATCTATCTATCTATGTATCTATCTATCTATCTATCTATCTATCTATCTATCTTCAAAAT

Figure 1.

Example of the information provided in Supplementary Table S3: a) locus name, length-based allele, bracketed repeat region, summary of flanking region variants in 5’ to 3’ order; b) allele frequencies and counts for the full set and by population; c) reported sequence divided by 5’ flank, repeat region, 3’ flank, with important features denoted in bold red font (bold black font in printed journal); d) result of flanking region comparison to GRCh38, with differences reported explicitly relative to the chromosomal reference accession number, and the STRSeq accession number and range associated with the sequence; e) FASTA formatted sequence.

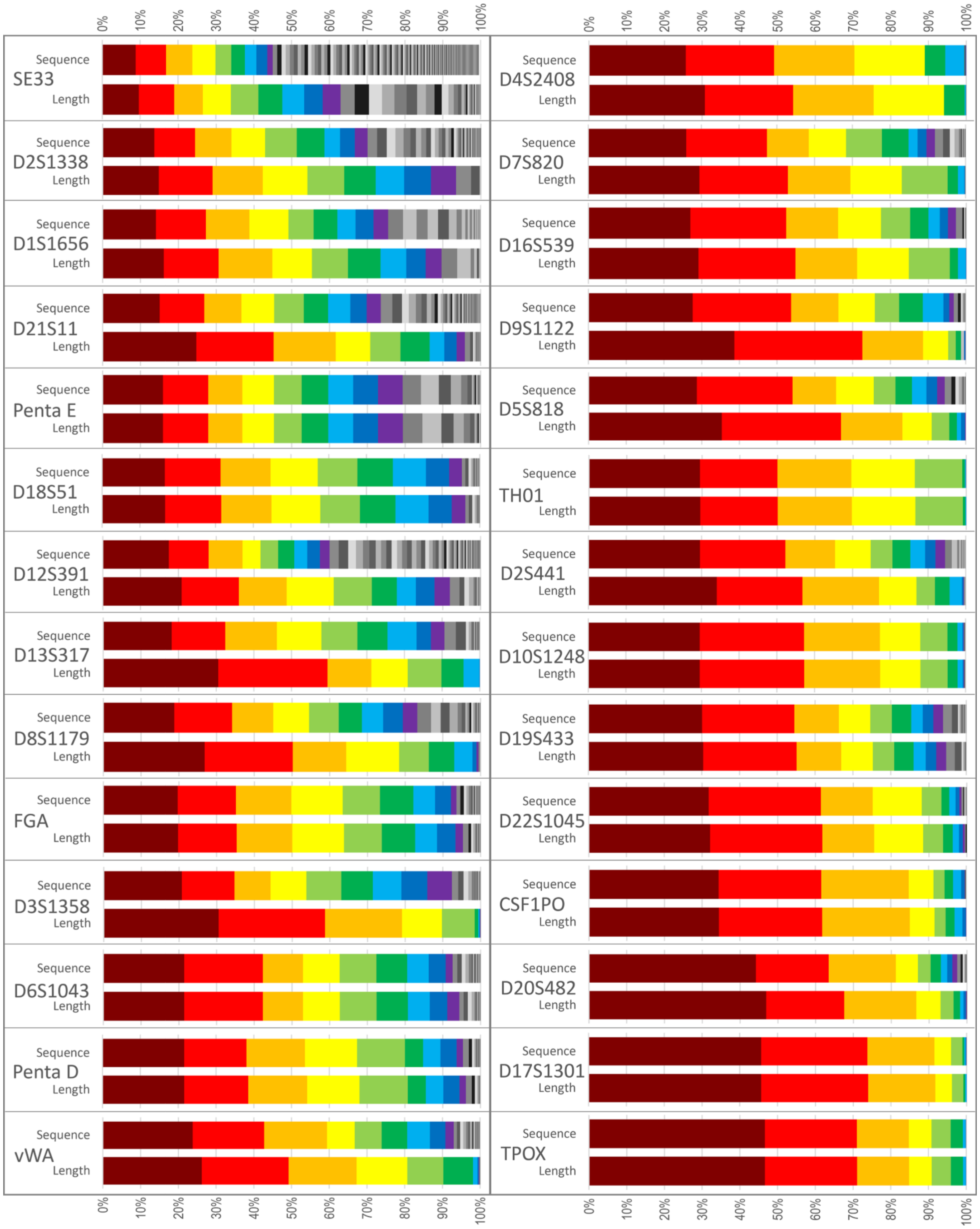


Figure 2.

Across-population allele frequency distribution per locus, by sequence and by length in N = 1036. Loci are sorted in ascending order of sequence-based frequency of the most common allele at each locus (first column top to bottom followed by second column top to bottom). The first nine alleles at each locus are colored to facilitate comparisons within and across loci, with any remaining alleles shown in grayscale. Sequence data for these samples at the SE33 locus are reported in [9].

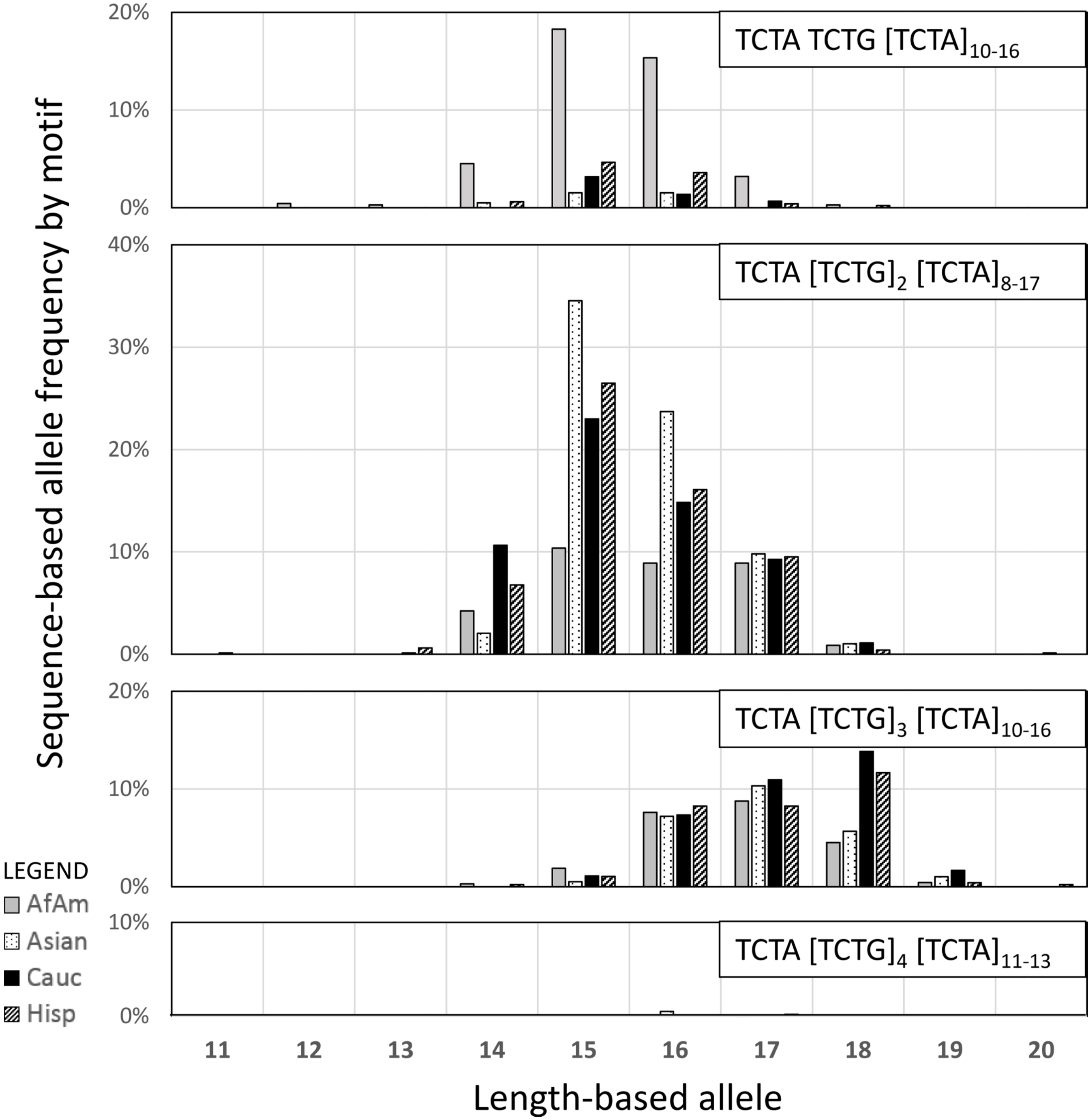


Figure 3.

D3S1358 frequency distribution among the primary motifs by length-based allele and population in N = 1036. The motif is defined as: the first subunit is fixed TCTA, the second subunit is definitive of the motif with TCTG varying from one to four, and the third subunit contains a widely varying number of TCTA repeats. For simplicity, seven additional rare motif alleles present in the data set have been excluded from this figure.

5’ Flanking Region	Repeat Region	3’ Flanking Region	AfAm	Asian	Cauc	Hisp
TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	[TATC]n	AATCAATCATCTATCTATCTTTCTGTCTGTCTTTTTGGG	69.74%	41.75%	61.77%	67.58%
..............T........................		T......................................	-	0.52%	2.35%	0.85%
...............A.......................		T......................................	0.58%	-	-	0.21%
....		T......................................	28.36%	54.12%	35.46%	29.66%
....		T...T..................................	-	3.09%	-	-
....		T.......----...........................	0.15%	-	0.14%	0.42%
....		............G..........................	-	-	-	0.42%
....		T....................----..............	1.02%	-	0.14%	0.64%
....		......................T................	-	-	0.001%	-
1 2		3 4 5 6 7 8				

Figure 4.

D13S317 frequency distribution by population of the nine flanking region motifs identified in N = 1036. The first row of 5’ and 3’ flanking sequence is consistent with GRCh38, and is the most common sequence found in this data set. Dots in subsequent rows represent bases matching the first row. Flanking polymorphisms are identified by numbers one through eight in the bottom row: 1) rs73250432 C>T, 2) rs146621667 G>A, 3) rs9546005 A>T, 4) rs202043589 A>T, 5) rs1442523705 delATCT, 6) ss2137543825 A>G, 7) rs561167308 delTCTG, and 8) rs768323113 C>T. Variation in repeat region length, combined with these flanking region polymorphisms, results in 32 sequence-based alleles at this locus. Three additional D13S317 alleles in this data set result from repeat region sequence variants, each observed once, and have been excluded from this figure.

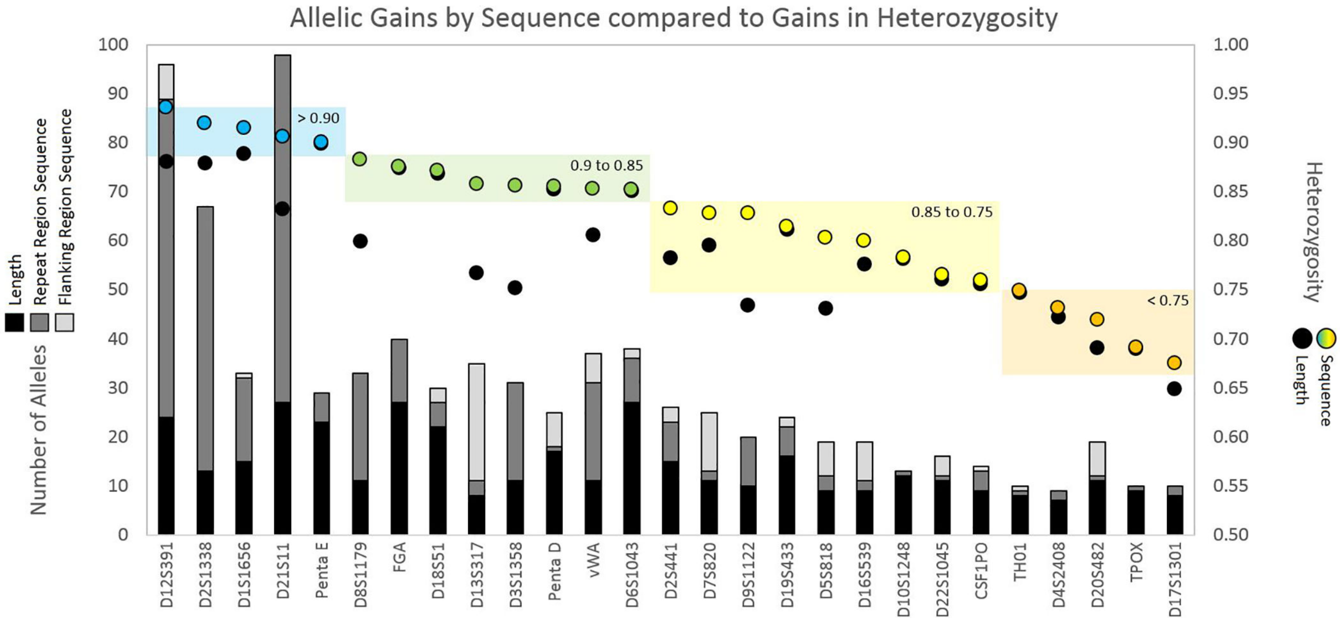


Figure 5.

Allelic gains by sequence compared to gains in heterozygosity for the 27 auSTR loci in N = 1036. Two y-axes are present: left y-axis = number of alleles, plotted as columns; right y-axis = heterozygosity, plotted as circles. Differential shading in the columns indicates number of alleles by length (black), sequence in the repeat region (dark gray), and sequence in the flanking region (light gray). Black circles represent heterozygosity by length. Colored circles represent heterozygosity by sequence, binned into ranges of heterozygosity: blue = > 0.90, green = 0.90 to 0.85, yellow = 0.85 to 0.75, and orange = < 0.75.

Table 1.

Sequence-based motifs and frequency by population in N = 1036. Motifs are represented in this table if the motif frequency exceeds 1 % in at least one population. “All other motifs” is the sum frequency of all alleles not captured by a motif shown in the table. Variable stretches are denoted with “n”. Lower case letters are not counted toward the repeat number designation. Bolded values highlight examples of frequencies over 20 % higher in one population compared to all other populations.

Locus			**Frequency of Motif by Population**			
	Allele Range	**Motif**	**AfAm**	**Asian**	**Cauc**	**Hisp**
D1S1656						
	10 to 18	CCTA [TCTA]n	46.35%	66.49%	50.69%	52.54%
	10 to 17	[TCTA]n	31.14%	17.01%	15.51%	17.58%
	14.3 to 19.3	CCTA [TCTA]n TCA [TCTA]n	21.78%	15.98%	31.99%	28.18%
	15 to 17	CTTA [TCTA]n	0.29%	-	0.97%	1.27%
		all other motifs	0.44%	0.52%	0.83%	0.42%
D2S441						
	8 to 13	[TCTA]n	53.80%	56.19%	41.55%	34.96%
	12 to 17	[TCTA]n TTTA [TCTA]2	32.02%	10.82%	32.96%	27.75%
	10 to 13	[TCTA]n TCTG TCTA	8.19%	18.04%	16.20%	27.75%
	11.3 to 14.3	[TCTA]n TCA [TCTA]n	5.26%	3.61%	6.51%	4.66%
	10 to 12	[TCTA]n rs74640515	0.58%	10.82%	2.63%	4.87%
		all other motifs	0.15%	0.52%	0.14%	-
D2S1338						
	15 to 24	[GGAA]n [GGCA]n	55.99%	51.03%	57.89%	58.05%
	19 to 27	[GGAA]2 GGAC [GGAA]n [GGCA]n	43.86%	43.30%	40.58%	38.98%
	18 to 20	[GGAA]n GAAA [GGAA]2 [GGCA]7	-	5.15%	-	2.33%
	19 and 20	[GGAA]n GGGA [GGCA]7	-	0.52%	1.52%	0.64%
		all other motifs	0.15%	-	-	-
D3S1358						
	12 to 18	TCTA TCTG [TCTA]n	**42.40%**	3.61%	5.26%	9.53%
	11 to 20	TCTA [TCTG]2 [TCTA]n	33.33%	71.13%	59.28%	59.96%
	14 to 20	TCTA [TCTG]3 [TCTA]n	23.54%	24.74%	34.90%	30.08%
		all other motifs	0.73%	-	0.55%	0.42%
D4S2408						
	7 to 13	[ATCT]n	98.39%	72.16%	97.78%	94.49%
	9	ATCT GTCT [ATCT]7	1.32%	**27.84%**	2.22%	5.51%
		all other motifs	0.29%	-	-	-
FGA						
	17 to 29	[GGAA]2 GGAG [AAAG]n AGAA AAAA [GAAA]3	89.33%	98.97%	97.51%	97.25%
	22 to 30	[GGAA]2 GGAG [AAAG]5 AAGG [AAAG]n AGAA AAAA [GAAA]3	6.29%	-	-	1.48%

Locus			Frequency of Motif by Population			
	Allele Range	Motif	AfAm	Asian	Cauc	Hisp
	16.2 to 25.2	[GGAA]2 GGAG [AAAG]n --AA AAAA [GAAA]3	3.07%	0.52%	2.22%	0.64%
		all other motifs	1.32%	0.52%	-	0.64%
D5S818						
	7 to 15	[ATCT]n	70.76%	77.32%	74.38%	76.69%
	8 to 14	[ATCT]n rs73801920	26.17%	22.68%	25.62%	23.09%
	13 to 15	[ATCT]3 ATGT [ATCT]n	3.07%	-	-	0.21%
		all other motifs	-	-	-	-
D6S1043						
	8 to 15	[ATCT]n	54.82%	61.86%	69.25%	64.41%
	15 to 22	[ATCT]n ATGT [ATCT]n	42.55%	37.63%	29.64%	27.97%
	23 to 26	[ATCT]n ATGT [ATCT]4 ATGT [ATCT]n	1.90%	-	-	-
	18.3 to 23.3	[ATCT]n ATGT [ATCT]2 ATC [ATCT]n	0.44%	-	-	6.99%
		all other motifs	0.29%	0.52%	1.10%	0.64%
D7S820						
	6 to 14	[TATC]n rs7789995	80.26%	70.62%	82.55%	86.44%
	7 to 12	[TATC]n rs7789995, rs16887642	17.40%	24.74%	5.96%	4.87%
	9 to 14	[TATC]n	2.19%	4.64%	11.36%	8.26%
		all other motifs	0.15%	-	0.14%	0.42%
D8S1179						
	8 to 16	[TCTA]n	16.67%	50.52%	46.26%	37.50%
	11 to 17	TCTA TCTG [TCTA]n	38.89%	26.29%	41.69%	42.37%
	11 to 18	[TCTA]2 TCTG [TCTA]n	41.08%	23.20%	11.36%	19.07%
	13 to 18	[TCTA]2 [TCTG]2 [TCTA]n	3.07%	-	0.14%	0.21%
		all other motifs	0.29%	-	0.55%	0.85%
D9S1122						
	9 to 17	[TAGA]n	76.61%	73.20%	57.76%	70.13%
	9 to 16	TAGA TCGA [TAGA]n	23.25%	26.29%	42.11%	29.87%
		all other motifs	-	-	-	-
vWA						
	11 to 21	[TAGA]n [CAGA]3–6 TAGA	93.71%	80.93%	91.55%	92.37%
	14 to 15	[TGGA]0,1 [TAGA]3 TGGA [TAGA]3 [CAGA]4 TAGA CAGA TAGA rs75219269	3.36%	19.07%	8.17%	7.42%
	14 to 15	[TAGA]n [CAGA]4 TAGA rs771794429, rs199970098	1.90%	-	-	0.21%
		all other motifs	1.02%	-	0.14%	-
D12S391						
	14 to 27	[AGAT]n [AGAC]n AGAT	85.82%	90.72%	66.07%	77.54%
	18 to 27	[AGAT]n [AGAC]n	10.96%	7.73%	27.70%	18.22%

Locus			**Frequency of Motif by Population**			
	Allele Range	**Motif**	**AfAm**	**Asian**	**Cauc**	**Hisp**
	17.3 to 19.3	AGAT GAT [AGAT]n [AGAC]7 AGAT	1.32%	-	4.85%	3.39%
	17.1 to 20.1	AGAT T [AGAT]n [AGAC]6 AGAT	1.75%	-	-	0.21%
		all other motifs	0.15%	1.55%	1.39%	0.64%
D13S317						
	8 to 15	[TATC]n	69.74%	41.75%	61.77%	67.58%
	9 to 15	[TATC]n rs9546005	28.36%	54.12%	35.46%	29.66%
	10 and 11, 13	[TATC]n rs9546005, rs202043589	-	3.09%	-	-
	11 and 12	[TATC]n rs73250432, rs9546005	-	0.52%	2.35%	0.85%
	9 to 12	[TATC]n rs9546005, rs561167308	1.02%	-	0.14%	0.64%
		all other motifs	0.88%	0.52%	0.28%	1.27%
D16S539						
	8 to 15	[GATA]n	77.49%	45.88%	90.72%	73.52%
	5, 9 to 14	[GATA]n rs11642858	21.20%	**54.12%**	8.86%	26.27%
	11 and 12	[GATA]n rs114697632	1.17%	-	0.28%	-
		all other motifs	0.15%	-	0.14%	0.21%
D18S51						
	9 to 24, 28	[AGAA]n	98.25%	100.00%	98.34%	98.94%
	14, 15	AGAA AGCA [AGAA]n	-	-	1.39%	0.21%
		all other motifs	1.75%	-	0.28%	0.85%
D19S433						
	9 to 17	[CCTT]n ccta CCTT cttt CCTT	73.25%	68.56%	90.86%	81.14%
	12.2 to 18.2	[CCTT]n ccta CCTT tt CCTT	25.58%	30.93%	8.31%	17.58%
		all other motifs	1.17%	0.52%	0.83%	1.27%
D20S482						
	9 to 18	[AGAT]n	95.76%	95.88%	91.41%	93.01%
	10 to 16, 19	[AGAT]n rs77560248	4.24%	3.61%	8.59%	6.99%
		all other motifs	-	0.52%	-	-
D21S11						
	26 to 39	[TCTA]n [TCTG]n [TCTA]n ta [TCTA]n tca [TCTA]2 tccata [TCTA]n	81.29%	78.35%	74.52%	70.76%
	28.2 to 34.2	[TCTA]n [TCTG]n [TCTA]3 ta [TCTA]n tca [TCTA]2 tccata [TCTA]n TA TCTA	16.52%	20.62%	25.07%	28.60%
	35, 36	[TCTA]5 [TCTG]6 [TCTA]3 ta [TCTA]3 tca [TCTA]2 tccata [TCTA]n TCA [TCTA]3 TCA [TCTA]2 TA TCTA	1.75%	-	0.14%	-
	29.3, 30.3	[TCTA]n [TCTG]n [TCTA]3 ta [TCTA]3 tca [TCTA]2 tccata [TCTA]n TCA [TCTA]n	0.15%	1.03%	-	-
		all other motifs	0.29%	-	0.28%	0.64%
Penta D						
	5 to 17	[AAAGA]n	85.96%	99.48%	99.03%	97.45%

Locus		Frequency of Motif by Population			
Allele Range	Motif	AfAm	Asian	Cauc	Hisp
2.2, 3.2	[AAAGA]n rs1190908807	12.28%	-	0.42%	1.91%
	all other motifs	1.75%	0.52%	0.55%	0.64%

Table 2.

Observed heterozygosity by length (Len) and sequence (Seq), and Increase in heterozygosity from length to sequence, for each population and combined (All). Loci are sorted in descending order of Increase for All. Increases > 0.05 are bolded. Length-based values for D9S1122, D20S482, D17S1301, and D4S2408 are derived from [15], representing a subset of N = 1036 which does not include Asian samples; NR = not reported. Length-based values for the remaining 23 STR loci are derived from [8].

	All (N=1036)			AfAm (N=342)			Cauc (N=361)			Hisp (N=236)			Asian (N=97)		
Locus	**Len**	**Seq**	**Increase**	**Len**	**Seq**	**Increase**	**Len**	**Seq**	**Increase**	**Len**	**Seq**	**Increase**	**Len**	**Seq**	**Increase**
D3S1358	0.752	0.856	**0.104**	0.763	0.886	**0.123**	0.756	0.839	**0.083**	0.733	0.847	**0.114**	0.742	0.835	**0.093**
D9S1122	0.734	0.828	**0.094**	0.753	0.839	**0.086**	0.742	0.850	**0.108**	0.686	0.783	**0.097**	NR	0.814	-
D13S317	0.767	0.858	**0.091**	0.705	0.822	**0.117**	0.762	0.864	**0.102**	0.852	0.903	**0.051**	0.804	0.856	**0.052**
D8S1179	0.799	0.883	**0.084**	0.795	0.883	**0.088**	0.784	0.864	**0.080**	0.784	0.894	**0.110**	0.907	0.928	0.021
D21S11	0.833	0.906	**0.073**	0.839	0.912	**0.073**	0.823	0.878	**0.055**	0.864	0.945	**0.081**	0.773	0.897	**0.124**
D5S818	0.731	0.803	**0.072**	0.728	0.810	**0.082**	0.706	0.789	**0.083**	0.737	0.780	0.042	0.814	0.887	**0.072**
D12S391	0.881	0.936	**0.055**	0.863	0.933	**0.070**	0.898	0.931	0.033	0.890	0.958	**0.068**	0.866	0.918	**0.052**
D2S441	0.783	0.833	**0.050**	0.789	0.827	0.038	0.787	0.823	0.036	0.754	0.843	**0.089**	0.814	0.866	**0.052**
vWA	0.806	0.853	0.047	0.816	0.889	**0.073**	0.806	0.839	0.033	0.788	0.835	0.047	0.814	0.825	0.010
D2S1338	0.879	0.920	0.041	0.901	0.965	**0.064**	0.870	0.900	0.030	0.852	0.873	0.021	0.907	0.948	0.041
D7S820	0.795	0.828	0.033	0.772	0.819	0.047	0.831	0.850	0.019	0.822	0.835	0.013	0.680	0.763	**0.082**
D20S482	0.691	0.719	0.028	0.673	0.690	0.017	0.689	0.726	0.037	0.729	0.746	0.017	NR	0.732	-
D1S1656	0.889	0.915	0.026	0.871	0.906	0.035	0.925	0.939	0.014	0.881	0.924	0.042	0.835	0.835	<0.01
D17S1301	0.649	0.675	0.026	0.626	0.658	0.032	0.717	0.701	<0.01	0.564	0.631	**0.067**	NR	0.742	-
D16S539	0.776	0.800	0.024	0.789	0.816	0.026	0.765	0.789	0.025	0.788	0.814	0.025	0.742	0.753	0.010
D4S2408	0.722	0.732	<0.01	0.752	0.766	0.014	0.709	0.726	0.017	0.691	0.686	<0.01	NR	0.742	-
D22S1045	0.761	0.765	<0.01	0.804	0.816	0.012	0.753	0.756	<0.01	0.733	0.733	<0.01	0.701	0.701	<0.01
CSF1PO	0.756	0.760	<0.01	0.789	0.795	<0.01	0.734	0.734	<0.01	0.729	0.737	<0.01	0.784	0.784	<0.01
Penta D	0.853	0.856	<0.01	0.868	0.871	<0.01	0.859	0.861	<0.01	0.868	0.868	<0.01	0.742	0.753	0.010
D18S51	0.869	0.872	<0.01	0.854	0.854	<0.01	0.856	0.861	<0.01	0.907	0.911	<0.01	0.876	0.876	<0.01
D19S433	0.812	0.815	<0.01	0.889	0.889	<0.01	0.767	0.770	<0.01	0.780	0.788	<0.01	0.784	0.784	<0.01
TH01	0.747	0.749	<0.01	0.763	0.769	<0.01	0.742	0.742	<0.01	0.758	0.758	<0.01	0.680	0.680	<0.01
TPOX	0.690	0.691	<0.01	0.760	0.760	<0.01	0.654	0.657	<0.01	0.674	0.674	<0.01	0.619	0.619	<0.01
Penta E	0.900	0.901	<0.01	0.904	0.906	<0.01	0.892	0.892	<0.01	0.898	0.898	<0.01	0.918	0.918	<0.01

	All (N=1036)			AfAm (N=342)			Cauc (N=361)			Hisp (N=236)			Asian (N=97)		
Locus	**Len**	**Seq**	**Increase**	**Len**	**Seq**	**Increase**	**Len**	**Seq**	**Increase**	**Len**	**Seq**	**Increase**	**Len**	**Seq**	**Increase**
D10S1248	0.782	0.783	<0.01	0.822	0.825	<0.01	0.765	0.765	<0.01	0.763	0.763	<0.01	0.753	0.753	<0.01
D6S1043	0.851	0.852	<0.01	0.886	0.886	<0.01	0.798	0.798	<0.01	0.890	0.894	<0.01	0.835	0.835	<0.01
FGA	0.875	0.875	<0.01	0.877	0.877	<0.01	0.867	0.867	<0.01	0.881	0.886	<0.01	0.876	0.876	<0.01