Forensic Science International: Genetics 7 (2013) 409–417

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

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

STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data

David H. Warshauer [a], David Lin [b], Kumar Hari [b], Ravi Jain [b], Carey Davis [a], Bobby LaRue [a], Jonathan L. King [a], Bruce Budowle [a,c,*]

[a] Institute of Applied Genetics, Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Boulevard, Fort Worth, TX 76107, USA
[b] cBio, Inc., 37869 Abraham Street, Fremont, CA 94536, USA
[c] Center of Excellence in Genomic Medicine (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 8 March 2013
Received in revised form 4 April 2013
Accepted 15 April 2013

Keywords:
STR
Bioinformatics
Software
Second generation sequencing
MiSeq™
GAIIx™

ABSTRACT

Recent studies have demonstrated the capability of second generation sequencing (SGS) to provide coverage of short tandem repeats (STRs) found within the human genome. However, there are relatively few bioinformatic software packages capable of detecting these markers in the raw sequence data. The extant STR-calling tools are sophisticated, but are not always applicable to the analysis of the STR loci commonly used in forensic analyses. STRait Razor is a newly developed Perl-based software tool that runs on the Linux/Unix operating system and is designed to detect forensically-relevant STR alleles in FASTQ sequence data, based on allelic length. It is capable of analyzing STR loci with repeat motifs ranging from simple to complex without the need for extensive allelic sequence data. STRait Razor is designed to interpret both single-end and paired-end data and relies on intelligent parallel processing to reduce analysis time. Users are presented with a number of customization options, including variable mismatch detection parameters, as well as the ability to easily allow for the detection of alleles at new loci. In its current state, the software detects alleles for 44 autosomal and Y-chromosome STR loci. The study described herein demonstrates that STRait Razor is capable of detecting STR alleles in data generated by multiple library preparation methods and two Illumina® sequencing instruments, with 100% concordance. The data also reveal noteworthy concepts related to the effect of different preparation chemistries and sequencing parameters on the bioinformatic detection of STR alleles.

1. Introduction

The majority of genetic analyses used in forensic casework involve the detection of short tandem repeats (STRs), which are relatively small sequences of DNA made up of repeating units of 2–6 nucleotides [1–3]. Currently, the accepted means of detecting these markers is size separation by capillary electrophoresis (CE) [4–7]. In recent years, however, second generation sequencing (SGS) technology has advanced to the point that it can be considered a viable platform for forensic DNA analyses, including STR detection [8–11]. In addition, sequencing has long been regarded as an effective means of revealing individual base variations in DNA known as single nucleotide polymorphisms (SNPs) [11–15]. While the traditional CE-based method of STR detection reveals only the length of alleles, SGS can increase resolution and detect nucleotide variation within the repeat regions and in proximal flanking regions [16]. Furthermore, the expansive genetic coverage and read length associated with current SGS methods allow for the potential capture of genetic information related to far more forensic STR markers than are possible with conventional multiplex-based CE kits. With the ability to sequence gigabases of DNA [17,18], a properly designed assay could yield STR information in a single analysis which surpasses that of all the currently available commercial CE-based kits combined, and do so on multiple samples simultaneously. However, while the extant sequencing instruments are capable of providing such extensive data, currently available software tools for identifying forensic STR alleles within the data are only beginning to address the task.

One such tool, lobSTR [8], uses an algorithm specifically designed to identify STR alleles within SGS data. First, the software analyzes a raw FASTA/FASTQ or BAM input file, detecting reads that contain an STR sequence and identifying the repeat motif. Next, lobSTR aligns the regions that flank the STR sequence to a modified reference sequence. Finally, the algorithm determines the identity of the allele(s) based on the number of detected repeat

* Corresponding author. Tel.: +1 8177352979.
E-mail address: Bruce.Budowle@unthsc.edu (B. Budowle).

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.04.005

units between the two flanking regions, applying statistical corrections to produce the most likely allelotype. While this software is certainly a refined and reasonably accurate method, it is somewhat limited in that lobSTR, by default, identifies only a single simple repeat motif. To allow the software to detect alleles at STRs that have longer, complex repeats, such as those within the D21S11 locus, for example, the user must determine the distinct simple repeats that comprise the complex motif and instruct lobSTR to identify each of these repeats individually. The resulting data must then be interpreted altogether in order to draw conclusions. This necessity makes lobSTR less applicable for the analysis of forensic STR markers, which display varying repeat motif complexity.

Recently, a method was introduced by Bornman et al. [9] that allows for the detection of STR alleles in SGS data using a different strategy. This method uses the Bowtie short read aligner [19] to align raw SGS reads to an "in silico reference," which is a user-generated FASTA file containing the full sequence of each allele at each STR locus. To reduce erroneous allele calls, reads are filtered so that only those encompassing the entire repeat region defined in the reference file are used for allelotyping. Allele calls are made using a heuristic decision model based on Fisher's exact test, and probability values are given for each allele call. This software also is effective for identifying STR alleles in sequence data, but requires prior knowledge of allelic sequence information. As a result, novel alleles or allelic variants, or those for which there are no documented sequence data, are a limitation for this system.

STRait Razor (the STR allele identification tool – Razor) is a Perl script designed for the Linux/Unix platform that identifies alleles at forensic STR loci based on the length of the repeat sequence, a method that is conceptually similar to the length-based allele detection offered by CE. This software is capable of handling repeat motifs ranging from simple to complex, and it does not require a reference composed of extensive allelic sequence data. As a result, the allele call results are consistent with those of current CE-based methods, and it is not confounded by unexpected sequence variation within repeats. In its first iteration of development, STRait Razor is capable of detecting alleles at 44 forensically-relevant STR loci, and others can be configured readily.

Fordyce et al. [10] independently developed a software tool that functions similarly to STRait Razor, isolating the repeat region of interest and performing length-based allelotyping. However, the algorithm was designed for use with the Roche® Genome Sequencer FLX™ and is only able to analyze FASTA files consisting of sequence data that contain Roche® Molecular Identifier (MID) tags. Currently this software only is able to identify alleles at 5 STR loci, compared with the 44 STR loci detected by STRait Razor. While STRait Razor was only tested on raw FASTQ files output by Illumina® instruments in this study, the software should, in theory, maintain compatibility with the raw read files generated by any second generation sequencing platform.

2. Materials and methods

To test the efficiency and accuracy of STRait Razor, an initial concordance study was performed wherein allele calls made by the software were compared with CE results. Following the University of North Texas Health Science Center Institutional Review Board approval, DNA from five Caucasian blood samples was used for this trial.

2.1. Sample preparation and CE typing

Blood samples were extracted using the Qiagen® QIAamp® DNA Mini kit (Qiagen Inc., Valencia, CA), following the manufacturer's recommendations. The quantity of extracted DNA was determined using the Applied Biosystems® Quantifiler™ Human DNA Quantification Kit (Life Technologies, Carlsbad, CA) on an Applied Biosystems® 7500 Real-Time PCR System (Life Technologies), according to the manufacturer's protocol. Amplification was performed using the reagents from the Applied Biosystems® AmpFlSTR® Identifiler® PCR Amplification Kit (Life Technologies) and the Promega® PowerPlex® 16 HS, ESI 17 Pro, and Y23 Systems (Promega Corporation, Madison, WI), on an Applied Biosystems® GeneAmp® PCR System 9700 thermal cycler (Life Technologies), according to the manufacturer's recommendations. These various kits allowed for the typing of the following STR loci: CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D5S818, D7S820, D8S1179, FGA, TH01, TPOX, vWA, PENTA D, PENTA E, D10S1248, D12S391, D1S1656, D22S1045, D2S441, SE33, DYS19, DYS385, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS481, DYS533, DYS549, DYS570, DYS576, DYS635, DYS643, and GATA H4. Capillary electrophoresis was performed on an Applied Biosystems® 3130*xl* Genetic Analyzer (Life Technologies) using POP-4™ polymer (Life Technologies) and analyzed using Applied Biosystems® GeneMapper® ID v3.2 software (Life Technologies), according to the manufacturer's recommended protocol.

2.2. Sample preparation, SGS, and STRait Razor typing

The quantity of extracted DNA from the blood sample was determined using a Qubit® 2.0 Fluorometer (Life Technologies). Library preparation prior to sequencing was performed using either the Illumina® TruSeq™ Custom Enrichment protocol (Illumina, Inc., San Diego, CA) or the Agilent Technologies HaloPlex™ Target Enrichment protocol (Agilent Technologies, Inc., Santa Clara, CA). Using the DesignStudio (Illumina, Inc.) and SureDesign (Agilent Technologies, Inc.) software, respectively, custom probes were designed to target the forensically-relevant STR loci. Paired-end sequencing was carried out on the GAIIx™ and MiSeq™ sequencing platforms (Illumina, Inc.). For these trials, the read lengths employed by these instruments were 2×146 and 2×251, respectively. Sample 1 was prepared using the HaloPlex™ chemistry and sequenced on both the GAIIx™ and MiSeq™ instruments. The sample also was prepared using the TruSeq™ chemistry, and subsequently sequenced on the MiSeq™. Sample 2 was prepared using the HaloPlex™ chemistry and sequenced on the GAIIx™. Samples 3, 4, and 5 were prepared using the TruSeq™ chemistry and sequenced on the MiSeq™. Following sequencing, the GAIIx™ output *bcl* files were demultiplexed and converted to a single FASTQ file using CASAVA v1.8.2 [20]. The MiSeq™ output was automatically converted to FASTQ format by the MiSeq™ Reporter software [21]. These FASTQ files served as the input for STRait Razor. The software was designed to detect the following forensic STR loci: CSF1PO, TPOX, D2S441, D3S1358, D5S818, D13S317, D18S51, D16S539, D7S820, D8S1179, TH01, vWA, D21S11, FGA, D2S1338, D19S433, PENTA D, PENTA E, D10S1248, D12S391, D1S1656, D22S1045, DYS389I/II, DYS390, DYS456, DYS19, DYS458, DYS437, DYS438, DYS448, GATA H4, DYS391, DYS392, DYS393, DYS439, DYS481, DYS533, DYS549, DYS570, DYS576, DYS643, DYS385, and DYS635. For these trials, allele-calling was performed using STRait Razor's default flank recognition settings (1 allowable substitution and no allowable insertions or deletions). The server used for STRait Razor analysis was a Dell™ PowerEdge™ R900 blade server, with 64 GB DDR3 ECC RAM and 4 Quad-Core Intel® Xeon® E7430 CPUs (2.13 GHz each).

2.3. STRait Razor information

The algorithm employed by STRait Razor is relatively simple (Fig. 1). First, reads containing both a leading and trailing flanking

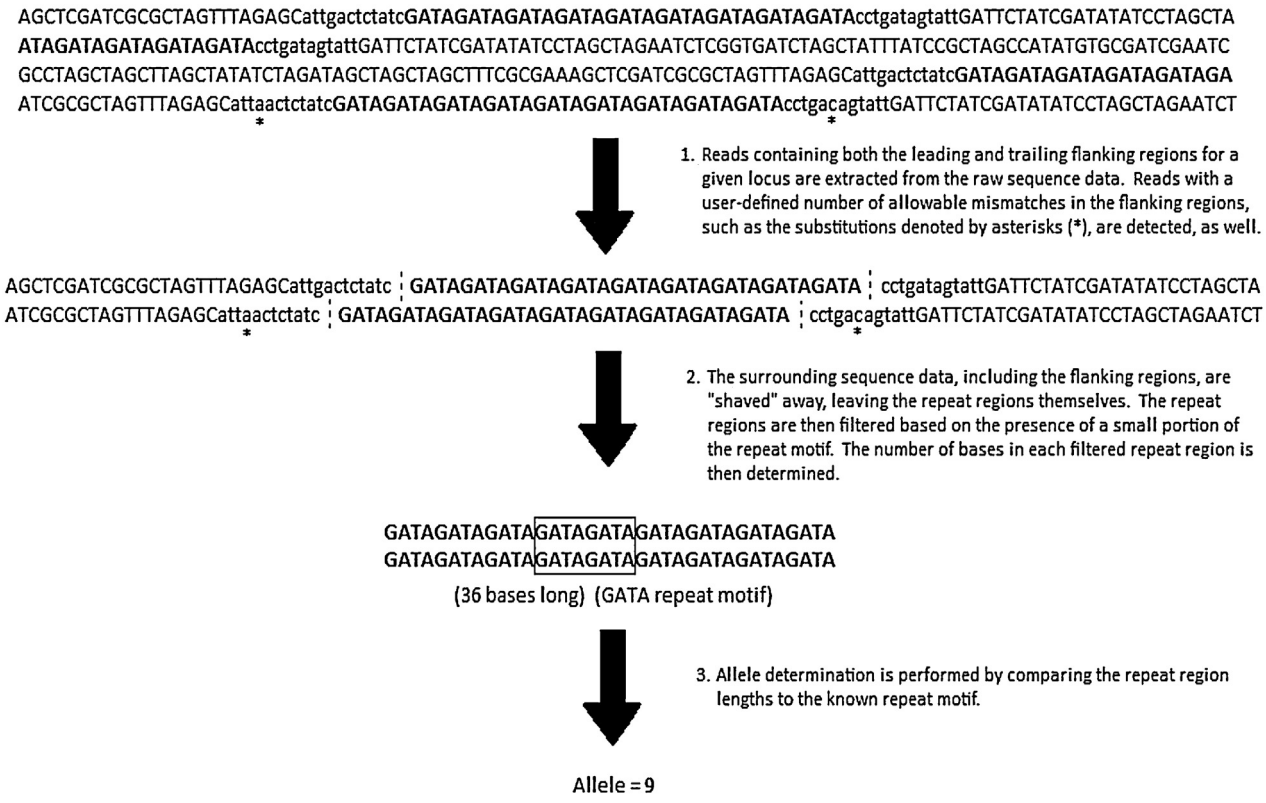

Fig. 1. STRait Razor algorithm. In this figure, each line represents an individual sequence read within the input FASTQ file. The repeat regions are shown in bold, capitalized font, while the flanking regions are shown in plain, lowercase font. Surrounding sequences are shown in plain, capitalized font.

region (Tables 1 and 2) surrounding the repeat sequence are extracted from the raw FASTQ sequence file(s) using the AGREP function, an approximate string search tool that is capable of detecting inexact matches to a provided query [22]. During this process, the software allows for customizable mismatch, insertion, and deletion penalties for each flanking region. This first step ensures that the extracted reads encompass the full repeat sequence, as partial repeat sequence data cannot be used for accurate allele-calling. With the reads containing the complete repeat sequence, the regions adjacent to the repeat sequence on each side are trimmed or "shaved" away (hence the program's name), leaving only the repeats themselves. At this point, the reads are filtered based on the presence of the repeat motif specific to the STR locus, thus discarding the majority of any irrelevant sequence data that may have been inadvertently captured. Next, the number of nucleotides in each repeat sequence is counted and compared with the expected lengths of alleles at that locus, based on the repeat motif. For example, at a STR locus with a simple tetranucleotide repeat motif, a repeat sequence consisting of 48 bases would indicate the presence of a "12" allele, while a sequence of 50 bases would indicate the presence of a "12.2" allele. Alleles can be called in this fashion regardless of intra-sequence variation or repeat motif complexity, and the length-based method of detection, to date, is concordant with CE results. The repeat region sequences of the called alleles then are sorted by length and written to a text file specific to the STR locus. With this data output, analysts can observe every nucleotide and evaluate any variations within the repeat sequences, as desired. This is not currently an automated process and requires manual inspection of the sequence data. Finally, a colon-delimited text file is generated that lists the alleles called at each STR locus, including the number of reads in which each allele was detected. These read count values can be used as a measure of abundance, similar to the RFU values in electropherograms.

STRait Razor is designed to analyze both single-end and paired-end data. Thus, the program accepts either single input files (Read 1) or dual input files (Reads 1 and 2), and recognizes STR loci in both forward and reverse complement forms. Each locus is analyzed separately, such that trimmed repeat regions always remain associated with their respective loci. The speed with which STRait Razor can provide allele calls is directly related to the size of the input file(s), as well as the depth of reads that contain the queried STR loci. However, the software utilizes PPSS, the (Distributed) Parallel Processing Shell Script [23], which allows STRait Razor to analyze one STR locus per available processor core in parallel, thus reducing the amount of time needed for analysis. Also, the user is able to choose whether STRait Razor detects only autosomal STR alleles, Y-chromosome STR alleles, or both. This option can further reduce analysis time and may be useful in cases wherein analysts wish to investigate only a subset of the loci recognized by STRait Razor, such as the typing of a known female reference sample. The information required by STRait Razor to detect STR alleles is provided in a modular format, and a workbook is provided to allow users to easily add other STR loci to the modules. Flanking regions queried by STRait Razor are each 12 bases long. These lengths allow for enough nucleotide diversity to make them fairly specific for the repeat regions of interest. Users may choose to use different flanking region sequences, as well as different flank lengths, according to their preferences. While most flanking regions used by STRait Razor are directly adjacent to the repeat regions, 6 sets (those for loci D1S1656, Penta D, DYS385, DYS393, DYS481, and DYS635) were selected at different proximities to allow for increased specificity.

Table 1
Autosomal STR loci detected by STRait Razor. Leading and trailing flanking sequences for both forward and reverse complement reads are listed. All flanking sequences are directly adjacent to the repeat region, except for those labeled with an asterisk (*), which were modified to allow for increased specificity or more efficient detection. Additional alleles can be added at the discretion of the user.

Locus	Flanking region sequences	Detectable alleles
CSF1PO	GATAGATAGATT———AGGAAGTACTTA TAAGTACTTCCT———AATCTATCTATC	5, 6, 6.3, 7, 7.3, 8, 8.3, 9, 9.1, 10, 10.1, 10.2, 10.3, 11, 11.1, 11.3, 12, 12.1, 13, 14, 15, 16
D10S1248	TATTGTCTTCAT———ACTCACTCATTT AAATGAGTGAGT———ATGAAGACAATA	8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
D12S391	AAATCCCCTCTC———ACCTATGCATCC GGATGCATAGGT———GAGAGGGGATTT	15, 16, 17, 17.3, 18, 18.3, 19, 19.3, 20, 21, 22, 23, 24, 25, 26
D13S317	AGATGATTGATT———ATGTATTTGTAA TTACAAATACAT———AATCAATCATCT	5, 6, 7, 7.1, 8, 8.1, 9, 10, 11, 11.1, 11.3, 12, 13, 13.3, 14, 14.3, 15, 16, 17
D16S539	GACAGACAGGTG———TCATTGAAAGAC GTCTTTCAATGA———CACCTGTCTGTC	4, 5, 6, 7, 8, 8.3, 9, 9.3, 10, 11, 11.3, 12, 12.1, 12.2, 13, 13.1, 13.3, 14, 14.3, 15, 16
D18S51	TCCTCTCTCTTT———GAGACAAGGTCT AGACCTTGTCTC———AAAGAGAGAGGA	7, 8, 9, 9.2, 10, 10.2, 11, 11.1, 11.2, 12, 12.2, 12.3, 13, 13.1, 13.2, 13.3, 14, 14.2, 15, 15.1, 15.2, 15.3, 16, 16.1, 16.2, 16.3, 17, 17.1, 17.2, 17.3, 18, 18.1, 18.2, 19, 19.2, 20, 20.1, 20.2, 21, 21.1, 21.2, 22, 22.1, 22.2, 23, 23.1, 23.2, 24, 24.2, 25, 26, 27, 28.1, 28.3, 39.2
D19S433	AAGATTCTGTTG———AGAGAGGTAGAA TTCTACCTCTCT———CAACAGAATCTT	5.2, 6.2, 7, 8, 9, 10, 11, 11.1, 12, 12.1, 12.2, 13, 13.1, 13.2, 13.3, 14, 14.1, 14.2, 14.3, 15, 15.2, 16, 16.2, 17, 17.2, 18, 18.2, 19, 19.2, 20
D1S1656	TAAACACACACA———CATCATACAGTT* AACTGTATGATG———TGTGTGTGTTTA*	9, 10, 11, 12, 13, 13.3, 14, 14.3, 15, 15.3, 16, 16.3, 17, 17.1, 17.3, 18, 18.3, 19, 19.3, 20, 20.3, 21
D21S11	ATAGATAGACGA———AGGCAATTCACT AGTGAATTGCCT———TCGTCTATCTAT	12, 24, 24.2, 24.3, 25, 25.1, 25.2, 25.3, 26, 26.1, 26.2, 27, 27.1, 27.2, 27.3, 28, 28.1, 28.2, 28.3, 29, 29.1, 29.2, 29.3, 30, 30.1, 30.2, 30.3, 31, 31.1, 31.2, 31.3, 32, 32.1, 32.2, 32.3, 33, 33.1, 33.2, 33.3, 34, 34.1, 34.2, 34.3, 35, 35.1, 35.2, 35.3, 36, 36.1, 36.2, 36.3, 37, 37.2, 38, 38.2, 39, 39.2, 40.2, 41.2
D22S1045	TATTTTTATAAC———GAGACTACTATC GATAGTAGTCTC———GTTATAAAAATA	8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
D2S1338	GGATTGCAGGAG———AGGCCAAGCCAT ATGGCTTGGCCT———CTCCTGCAATCC	11, 12, 13, 14, 15, 16, 17, 18, 19, 19.3, 20, 21, 22, 23, 23.2, 23.3, 24, 25, 26, 27, 28
D2S441	TCTATGAAAACT———TATCATAACACC GGTGTTATGATA———AGTTTTCATAGA	8, 9, 10, 11, 11.3, 12, 12.3, 13, 13.3, 14, 14.3, 15, 16, 17
D3S1358	AGGCTTGCATGT———ATGAGACAGGGT ACCCTGTCTCAT———ACATGCAAGCCT	8, 8.3, 9, 10, 11, 12, 13, 14, 14.3, 15, 15.1, 15.2, 15.3, 16, 16.2, 17, 17.1, 17.2, 18, 18.1, 18.2, 18.3, 19, 20
D5S818	ATTTATACCTCT———TCAAAATATTAC GTAATATTTTGA———AGAGGTATAAAT	6, 7, 8, 9, 10, 10.1, 11, 11.1, 12, 12.3, 13, 14, 15, 16, 17, 18
D7S820	GAACGAACTAAC———GACAGATTGATA TATCAATCTGTC———GTTAGTTCGTTC	5, 5.2, 6, 6.2, 6.3, 7, 7.1, 7.3, 8, 8.1, 8.2, 8.3, 9, 9.1, 9.2, 9.3, 10, 10.1, 10.3, 11, 11.1, 11.3, 12, 12.1, 12.2, 12.3, 13, 13.1, 14, 14.1, 15, 16
D8S1179	CACTGTGGGGAA———TACGAATGTACA TGTACATTCGTA———TTCCCCACAGTG	7, 8, 9, 10, 10.2, 11, 12, 12.3, 13, 14, 15, 15.1, 15.3, 16, 17, 17.1, 18, 19, 20
FGA	GAAAGGAAGAAA———CTAGCTTGTAAA TTTACAAGCTAG———TTTCTTCCTTTC	12.2, 13, 13.2, 14, 14.3, 15, 15.3, 16, 16.1, 16.2, 17, 17.1, 17.2, 18, 18.1, 18.2, 19, 19.1, 19.2, 19.3, 20, 20.1, 20.2, 20.3, 21, 21.1, 21.2, 21.3, 22, 22.1, 22.2, 22.3, 23, 23.1, 23.2, 23.3, 24, 24.1, 24.2, 24.3, 25, 25.1, 25.2, 25.3, 26, 26.1, 26.2, 26.3, 27, 27.1, 27.2, 27.3, 28, 28.1, 28.2, 29, 29.1, 29.2, 30, 30.2, 31, 31.2, 32, 32.1, 32.2, 33.1, 33.2, 34.1, 34.2, 35.2, 41.1, 41.2, 42, 42.1, 42.2, 43.1, 43.2, 44, 44.2, 44.3, 45, 45.1, 45.2, 46, 46.1, 46.2, 47, 47.2, 48, 48.2, 49, 49.1, 49.2, 50.2, 50.3, 51, 51.2
Penta D	TTTATGATTCTC———TTGAGATGGTGT* ACACCATCTCAA———GAGAATCATAAA*	1.1, 1.2, 2.2, 3.2, 4, 5, 6, 6.4, 7, 7.1, 7.4, 8, 8.1, 9, 9.1, 9.4, 10, 10.1, 10.2, 10.3, 11, 11.1, 11.2, 12, 12.1, 12.2, 12.3, 12.4, 13, 13.2, 13.3, 13.4, 14, 14.1, 14.4, 15, 15.1, 16, 17, 18
Penta E	TCCTTACAATTT———GAGACTGAGTCT AGACTCAGTCTC———AAATTGTAAGGA	5, 6, 7, 8, 9, 9.1, 9.4, 10, 10.2, 11, 11.4, 12, 12.1, 12.2, 12.3, 13, 13.2, 13.4, 14, 14.4, 15, 15.2, 15.4, 16, 16.4, 17, 17.4, 18, 18.4, 19, 19.4, 20, 20.2, 20.3, 21, 22, 23, 23.4, 24, 26
TH01	CCCTTATTTCCC———TCACCATGGAGT ACTCCATGGTGA———GGGAAATAAGGG	3, 4, 5, 5.3, 6, 6.1, 6.3, 7, 7.1, 7.3, 8, 8.3, 9, 9.1, 9.3, 10, 10.3, 11, 12, 13.3, 14
TPOX	GAACCCTCACTG———TTTGGGCAAATA TATTTACCCAAA———CAGTGAGGGTTC	4, 5, 6, 7, 7.3, 8, 9, 10, 10.1, 10.3, 11, 12, 13, 13.1, 14, 15, 16
vWA	GACTTGGATTGA———TCCATCCATCCT AGGATGGATGGA———TCAATCCAAGTC	10, 11, 12, 13, 14, 15, 15.2, 16, 16.1, 17, 18, 18.1, 18.2, 18.3, 19, 19.2, 20, 21, 22, 23, 24, 25

2.4. Allele call comparison

The alleles detected by the CE method were compared with the allele call output files generated by STRait Razor. To be considered concordant, alleles detected via CE had to be detected by STRait Razor, based on the presence of the allelic sequence data in the input FASTQ file. It should be noted that alleles not detected by STRait Razor due to lack of respective sequence data, whether the result of kit chemistry limitations or inadequate sequencing read length, were not considered discordant.

3. Results

Based on the various combinations of library preparation methods and sequencing platforms utilized, a total of 7 overall comparisons with CE data were made for the 5 samples. The assortment of multiplex-based CE kits used for these samples allowed for the comparison of STR alleles at all loci that are currently detectable with STRait Razor: 22 autosomal STR loci and 22 Y-STR loci, with DYS385 designated as a single locus. Across all trials, a total of 427 alleles were compared. The allele calls made by STRait Razor were completely concordant with the genotype results generated by the CE method. Of the 427 alleles compared, 403 alleles were detected, with 100% concordance (Tables 3 and 4). For the 24 alleles not detected by the software, a subsequent manual analysis of the FASTQ input files revealed that there were no sequence reads for these alleles in which the full repeat regions (including the surrounding flanking sequences) were present. These undetected alleles were not represented in the sequencing data because of the library preparation method (e.g., random shearing of genomic DNA) and/or the read length used, and thus could not be recognized by STRait Razor. These instances do not reflect inadequacy of the software and should not be considered evidence of discordance. Intra-repeat variation was occasionally observed in the data output by STRait Razor. For example, an examination of the

Table 2
Y-chromosome STR loci detected by STRait Razor. Leading and trailing flanking sequences for both forward and reverse complement reads are listed. All flanking sequences are directly adjacent to the repeat region, except for those labeled with an asterisk (*), where the proximities were modified to allow for increased specificity or more efficient detection. Additional alleles can be added at the discretion of the user.

Locus	Flanking region sequences	Detectable alleles
DYS19	TATATAGTGTTT———TATAGTGACACT AGTGTCACTATA———AAACACTATATA	9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
DYS385	GAGAAAGAAAGG———GGAGGACTATGT* ACATAGTCCTCC———CCTTTCTTTCTC*	7, 8, 9, 9.2, 10, 10.2, 11, 11.2, 11.3, 12, 12.1, 12.2, 12.3, 13, 13.1, 13.2, 13.3, 14, 14.2, 14.3, 15, 15.1, 15.2, 15.3, 16, 16.2, 16.3, 17, 17.1, 17.2, 17.3, 18, 18.1, 18.2, 19, 19.2, 19.3, 20, 21, 22, 23, 24, 25, 28
DYS389I	ATTATCTATGTA———TCCCTCCCTCTA TAGAGGGAGGGA———TACATAGATAAT	9, 10, 11, 12, 13, 14, 15, 16, 17
DYS389II	TCTATGTGTGTG———TCCCTCCCTCTA TAGAGGGAGGGA———CACACACATAGA	24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36
DYS390	ATATTCTATCTA———TCATCTATCTAT ATAGATAGATGA———TAGATAGAATAT	17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29
DYS391	TCTGTCTGTCTG———TCTGCCTATCTG CAGATAGGCAGA———CAGACAGACAGA	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
DYS392	TCACCATTTATT———TTACTAAGGAAT ATTCCTTAGTAA———AATAAATGGTGA	4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20
DYS393	TTGTGTCAATAC———GAGACATACCTC* GAGGTATGTCTC———GTATTGACACAA*	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
DYS437	ATGCCCATCCGG———TCATCTATCATC GATGATAGATGA———CCGGATGGGCAT	10, 11, 12, 13, 14, 15, 16, 17, 18, 19
DYS438	GTAAACAGTATA———TATTTGAAATGG CCATTTCAAATA———TATACTGTTTAC	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18
DYS439	TGATAAATAGAA———GAAAGTATAAGT ACTTATACTTTC———TTCTATTTATCA	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
DYS448	AGACATGGATAA———AGAGAGGTAAAG CTTTACCTCTCT———TTATCCATGTCT	14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24
DYS456	TGTGATAATGTA———ATTCCATTAGTT AACTAATGGAAT———TACATTATCACA	9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23
DYS458	AAGAAAAGGAAG———GGAGGGTGGGCG CGCCCACCCTCC———CTTCCTTTTCTT	11, 12, 13, 14, 15, 16, 16.2, 17, 17.2, 18, 18.2, 19, 19.2, 20, 21, 22, 23, 24
DYS481	TTCAGCATGCTG———GAGTCTTGCAAC* GTTGCAAGACTC———CAGCATGCTGAA*	16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
DYS533	TAGCTAGCTATC———ATCATCTATCAT ATGATAGATGAT———GATAGCTAGCTA	7, 8, 9, 10, 11, 12, 13, 14, 15
DYS549	GATTAGAAAGAT———GAAAAAATCTAC GTAGATTTTTTC———ATCTTTCTAATC	9, 10, 11, 12, 13, 14, 15
DYS570	CTCCAAGTTCCT———TTTTTGTAGATA TATCTACAAAAA———AGGAACTTGGAG	12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24
DYS576	CATCTCTGAATA———AAAAAGCCAAGA TCTTGGCTTTTT———TATTCAGAGATG	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22
DYS635	ATAGAATCTCTA———TCACATTTTCTT* AAGAAAATGTGA———TAGAGATTCTAT*	13, 16, 17, 18, 19, 20, 21, 21.3, 22, 23, 24, 25, 26, 27, 28, 29, 30
DYS643	AAACTACTGTGC———CTTTCTTTTTAA TTAAAAAGAAAG———GCACAGTAGTTT	7, 8, 9, 10, 11, 12, 13, 14, 15
GATA H4	TAGGTAGGTAGG———ATGGATAGATTA TAATCTATCCAT———CCTACCTACCTA	8, 9, 10, 11, 12, 13, 14, 15

sequence data output file for Sample 1 revealed that the "15" allele at locus D3S1358 consisted of both the "TCTA$(TCTG)_2(TCTA)_{12}$" variant and the "TCTA$(TCTG)_3(TCTA)_{11}$" variant, at a ratio of approximately 1:1. Conversely, the sequence data output file for Sample 2, which also displays a "15" allele at this locus, revealed that only the "TCTA$(TCTG)_2(TCTA)_{12}$" variant was present. This difference would not be detected by traditional CE methods. The time required for analysis, using a paired-end analysis method that detected both autosomal and Y chromosome STR alleles, ranged from 16 min (for dual 395 MB input files) to 285 min (for dual 7.9 GB input files) on the 16-core server used for this study.

4. Discussion

The results of this study show the efficiency and accuracy of STR allele detection with STRait Razor. They also reveal the close relationship between the software, the library preparation chemistries, and the sequencing platforms used to produce the sequence information.

Read length is arguably the most important factor (followed by coverage) that impacts the allelic detection capability of STRait Razor. The HaloPlex™ chemistry relies on enzymatic cleavage [24], which creates fragments with consistent start and end points. Depending on the length of the allele in question and the position of the repeat region within the resulting fragment(s), it is possible for sequencing reads to be produced that only partially span the repeat region and its associated flanking sequences (Fig. 2a). Since STRait Razor requires reads that contain all of this information, alleles may go undetected without complete repeat region traversal. An example of this phenomenon is locus D2S1338 in Sample 1 (HaloPlex™ preparation, GAIIx™ sequencing). For this locus, the "18" allele was called, but the "25" allele was too long to be covered completely by the sequencing read, and thus went undetected. When this same sample subsequently was sequenced on the MiSeq™ platform using a longer read length [17,18], the "25" allele was detected. In addition, if a repeat region is situated toward the beginning of a HaloPlex™ fragment, the allele is likely to be detected in one direction of a paired-end analysis. However, when the reads are sequenced from the opposite direction, the repeat region is oriented toward the end of the read and may not be completely encompassed (Fig. 2b). This situation can be seen in loci such as D7S820 and vWA in Sample

Table 3
Comparison of CE allele calls and STRait Razor results – Autosomal STRs. Alleles detected by both CE and STRait Razor analysis of SGS data are shown in bold in the columns for each sample. The numbers of reads in which an allele was detected by STRait Razor are listed in parentheses next to the respective allele. The first number in parentheses represents the abundance of the allele in Read 1 of the paired-end sequencing run, while the second number represents the abundance of the allele in Read 2. Alleles not detected by STRait Razor due to lack of relevant sequence data are denoted by "[–]".

		Detection method											
		Sample 1				Sample 2		Sample 3		Sample 4		Sample 5	
		CE	STRait Razor (TruSeq™ prep, GAIIx™)	STRait Razor (HaloPlex™ prep, GAIIx™)	STRait Razor (HaloPlex™ prep, MiSeq™)	CE	STRait Razor (HaloPlex™ prep, GAIIx™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)
S T R loci	CSF1PO	**12**	**12** (332, 315)	**12** (1135, 976)	**12** (324, 322)	**12**	**12** (2445, 4155)	**10**, **11**	**10** (89, 84), **11** (54, 75)	**12**	**12** (304, 310)	**10**, **12**	**10** (152, 189), **12** (162, 145)
	D13S317	**12**	**12** (104, 96)	**12** (639, 0)	**12** (217, 214)	**8**, **11**	**8** (2513, 4888), **11** (1995, 4377)	**11**, **12**	**11** (49, 61), **12** (50, 48)	**12**	**12** (220, 209)	**8**, **11**	**8** (101, 107), **11** (86, 76)
	D16S539	**10**, **12**	**10** (86, 70), **12** (64, 43)	**10** (886, 2896), **12** (686, 1546)	**10** (181, 135), **12** (174, 138)	**9**, **12**	**9** (1733, 7394), **12** (1165, 3209)	**8**, **13**	**8** (62, 69), **13** (51, 56)	**9**, **12**	**9** (121, 110), **12** (82, 77)	**11**, **13**	**11** (117, 98), **13** (91, 73)
	D18S51	**15**, **21**	**15** (25, 29), **21** (16, 17)	**15** (598, 3258), **21** (341, 541)	**15** (123, 123), **21** (106, 105)	**15**, **16**	**15** (957, 5711), **16** (819, 4734)	**14**, **16**	**14** (20, 16), **16** (13, 17)	**15**, **18**	**15** (34, 22), **18** (17, 17)	**12**	**12** (94, 79)
	D19S433	**12**, **14**	**12** (39, 19), **14** (23, 13)	**12** (589, 2620), **14** (542, 1543)	**12** (141, 138), **14** (126, 133)	**13**, **14**	**13** (1634, 6001), 14 (1423, 4149)	**13**, **14**	**13** (16, 5), **14** (10, 11)	**13**	**13** (44, 33)	**12**, **15**	**12** (12, 19), **15** (28, 17)
	D21S11	**29**, **30**	[–], [–]	[–], [–]	**29** (70, 17), **30** (57, 14)	**29**, **31**	[–], [–]	**29**	**29** (17, 9)	**29**	**29** (38, 25)	**28**, **29**	**28** (16, 11), **29** (16, 11)
	D2S1338	**18**, **25**	**18** (19, 18), **25** (9, 0)	**18** (80, 1), [–]	**18** (54, 52), **25** (29, 25)	**20**, **25**	**20** (137, 1), [–]	**18**	**18** (53, 53)	**17**, **24**	**17** (74, 67), **24** (35, 35)	**20**, **23**	**20** (70, 70), **23** (55, 37)
	D3S1358	**15**	**15** (109, 110)	**15** (449, 4079)	**15** (439, 395)	**15**, **16**	**15** (1046, 4579), 16 (852, 3995)	**16**, **18**	**16** (32, 29), **18** (36, 29)	**16**, **18**	**16** (71, 74), **18** (78, 66)	**14**, **17**	**14** (99, 93), **17** (72, 84)
	D5S818	**11**	**11** (76, 66)	[–]	**11** (66, 67)	**11**	[–]	**11**, **12**	**11** (20, 23), **12** (30, 26)	**11**	**11** (101, 86)	**12**, **13**	**12** (56, 62), **13** (45, 34)
	D7S820	**9**, **12**	**9** (1, 2), **12** (5, 4)	**9** (0, 1754), **12** (0, 1650)	**9** (70, 70), **12** (57, 57)	**10**	**10** (0, 7733)	**10**, **11**	**10** (4, 3), **11** (3, 4)	**9**, **13**	**9** (3, 7), **13** (8, 8)	**11**	**11** (17, 15)
	D8S1179	**13**	**13** (98, 87)	**13** (1722, 5068)	**13** (335, 220)	**13**	**13** (3941, 10527)	**10**, **12**	**10** (36, 38), **12** (24, 25)	**12**, **14**	**12** (52, 47), **14** (48, 49)	**13**, **16**	**13** (68, 78), **16** (58, 41)
	FGA	**20**, **21**	**20** (36, 25), **21** (26, 20)	**20** (770, 3819), **21** (581, 3419)	**20** (168, 129), **21** (146, 134)	**19**, **21**	**19** (1019, 5066), 21 (890, 4849)	**23**, **25**	**23** (12, 16), **25** (12, 13)	**22**, **24**	**22** (40, 30), **24** (18, 23)	**20**	**20** (84, 71)
	TH01	**9**, **9.3**	**9** (160, 146), **9.3** (132, 178)	**9** (3172, 5571), **9.3** (2893, 5493)	**9** (260, 255), **9.3** (297, 297)	**8**, **9.3**	**8** (5701, 8471), **9.3** (4963, 8350)	**7**	**7** (150, 162)	**9.3**	**9.3** (287, 292)	**9.3**	**9.3** (393, 390)
	TPOX	**8**, **9**	**8** (90, 96), **9** (99, 80)	**8** (4832, 5208), **9** (4488, 4710)	**8** (527, 479), **9** (475, 428)	**11**	**11** (11043, 15943)	**8**, **11**	**8** (35, 34), **11** (32, 29)	**8**, **12**	**8** (113, 105), **12** (84, 86)	**8**	**8** (176, 185)
	vWA	**16**, **17**	**16** (53, 37), **17** (39, 24)	**16** (299, 0), **17** (213, 0)	**16** (55, 55), **17** (36, 36)	**15**, **20**	**15** (669, 0), **20** (0, 3)	**17**	**17** (49, 56)	**14**, **17**	**14** (60, 60), **17** (57, 52)	**15**, **17**	**15** (63, 65), **17** (56, 61)
	Penta D	**10**	**10** (24, 35)	**10** (388, 0)	**10** (180, 0)	**14**, **15**	**14** (214, 0), [–]	**9**, **11**	**9** (11, 14), **11** (12, 10)	**12**, **14**	**12** (26, 23), **14** (25, 20)	**9**, **12**	**9** (23, 24), **12** (29, 30)
	Penta E	**11**, **12**	**11** (6, 7), **12** (9, 5)	**11** (98, 121), 12 (123, 104)	**11** (105, 87), **12** (107, 88)	**10**, **11**	**10** (290, 993), **11** (243, 191)	**5**, **7**	**5** (10, 7), **7** (6, 5)	**11**, **12**	**11** (9, 4), **12** (11, 6)	**7**, **14**	**7** (20, 19), **14** (8, 7)
	D10S1248	**13**, **14**	**13** (106, 94), **14** (69, 67)	**13** (672, 4537), 14 (329, 3613)	**13** (192, 199), **14** (159, 171)	**13**	**13** (3545, 17065)	**12**, **13**	**12** (81, 73), **13** (71, 69)	**13**, **14**	**13** (134, 127), **14** (152, 134)	**16**	**16** (211, 211)
	D12S391	**15**, **17**	**15** (50, 50), **17** (35, 31)	**15** (1464, 548), 17 (1101, 447)	**15** (167, 117), **17** (162, 106)	**21**	**21** (2123, 0)	**15**, **24**	**15** (37, 31), **24** (23, 19)	**17**, **21**	**17** (90, 82), **21** (84, 79)	**18**, **20**	**18** (68, 73), **20** (58, 63)
	D1S1656	**16**, **18.3**	**16** (63, 47), **18.3** (37, 52)	**16** (499, 0), **18.3** (468, 0)	**16** (59, 75), **18.3** (78, 41)	**11**, **12**	**11** (2262, 2848), **12** (1872, 2284)	**16.3**, **18.3**	**16.3** (33, 24), **18.3** (25, 29)	**15**, **17.3**	**15** (79, 75), **17.3** (67, 52)	**12**, **16**	**12** (98, 83), **16** (76, 67)
	D22S1045	**16**, **17**	**16** (10, 16), **17** (7, 16)	**16** (901, 2637), **17** (718, 1639)	**16** (107, 105), **17** (108, 104)	**15**, **16**	**15** (2042, 5133), **16** (1415, 3535)	**15**, **16**	**15** (15, 14), **16** (13, 15)	**16**	**16** (35, 38)	**11**, **16**	**11** (94, 94), **16** (40, 40)
	D2S441	**11**, **15**	**11** (72, 49), **15** (32, 37)	**11** (291, 3863), **15** (0, 3685)	**11** (124, 145), **15** (113, 167)	**11.3**, **14**	**11.3** (615, 7807), **14** (0, 7024)	**11**, **14**	**11** (74, 71), **14** (68, 64)	**10**, **11.3**	**10** (106, 100), **11.3** (149, 137)	**10**, **11**	**10** (131, 125), **11** (130, 138)
	Total alleles	38	36	34	38	37	32	40	40	37	37	38	38

Table 4
Comparison of CE allele calls and STRait Razor results – Y-Chromosome STRs. Alleles detected by both CE and STRait Razor analysis of SGS data are shown in bold in the columns for each sample. The numbers of reads in which an allele was detected by STRait Razor are listed in parentheses next to the respective allele. The first number in parentheses represents the abundance of the allele in Read 1 of the paired-end sequencing run, while the second number represents the abundance of the allele in Read 2. Alleles not detected by STRait Razor due to lack of relevant sequence data are denoted by "[–]".

		Detection method											
		Sample 1				Sample 2		Sample 3		Sample 4		Sample 5	
		CE	STRait Razor (TruSeq™ prep, GAIIx™)	STRait Razor (HaloPlex™ prep, GAIIx™)	STRait Razor (HaloPlex™ prep, MiSeq™)	CE	STRait Razor (HaloPlex™ prep, GAIIx™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)
S T R loci	DYS19	**14**	**14** (7, 11)	**14** (231, 1221)	**14** (61, 62)	**15**	**15** (6, 2341)	**14**	**14** (8, 18)	**16**	**16** (13, 13)	**14**	**14** (14, 15)
	DYS385	**11**, **13**	**11 (12**, **12)**, **13 (12**, **11)**	[–], [–]	[–], [–]	**11**, **14**	[–], [–]	**11**, **14**	**11** (5, 3), **14** (4, 5)	**11**, **14**	**11** (22, 29), 14 (8, 19)	**11**, **14**	**11** (12, 13), **14** (24, 16)
	DYS389I	**13**	**13** (102, 99)	**13** (423, 1)	**13** (26, 0)	**13**	**13** (1605, 0)	**13**	**13** (54, 47)	**13**	**13** (137, 137)	**13**	**13** (120, 103)
	DYS389II	**29**	[–]	[–]	**29** (23, 0)	**30**	[–]	**29**	**29** (9, 3)	**28**	**28** (25, 20)	**29**	**29** (13, 17)
	DYS390	**24**	**24** (14, 16)	**24** (115, 3495)	**24** (115, 126)	**25**	**25** (259, 3386)	**24**	**24** (16, 19)	**23**	**23** (37, 43)	**23**	**23** (40, 37)
	DYS391	**10**	**10** (175, 167)	**10** (952, 39)	**10** (99, 24)	**10**	**10** (3179, 1431)	**10**	**10** (73, 80)	**10**	**10** (180, 182)	**12**	**12** (166, 165)
	DYS392	**13**	**13** (3, 8)	**13** (885, 1965)	**13** (82, 78)	**11**	**11** (1466, 2850)	**13**	**13** (7, 8)	**13**	**13** (7, 8)	**13**	**13** (11, 10)
	DYS393	**13**	**13** (9, 2)	**13** (0, 360)	**13** (14, 13)	**14**	**14** (0, 1023)	**13**	**13** (2, 7)	**13**	**13** (2, 3)	**13**	**13** (10, 10)
	DYS437	**15**	**15** (85, 70)	**15** (0, 4020)	**15** (247, 238)	**14**	**14** (0, 11,064)	**15**	**15** (77, 77)	**15**	**15** (148, 133)	**14**	**14** (141, 146)
	DYS438	**12**	**12** (42, 36)	**12** (324, 285)	**12** (62, 32)	**11**	**11** (884, 871)	**12**	**12** (48, 49)	**12**	**12** (79, 68)	**12**	**12** (96, 93)
	DYS439	**12**	[–]	**12** (428, 2296)	12 (134, 78)	**10**	**10** (1789, 6072)	**12**	**12** (2, 0)	**11**	**11** (2, 2)	**13**	**13** (3, 1)
	DYS448	**19**	[–]	[–]	**19** (17, 5)	**19**	[–]	**19**	**19** (11, 3)	**19**	**19** (21, 16)	**18**	**18** (12, 11)
	DYS456	**15**	**15** (10, 13)	**15** (523, 1402)	**15** (80, 54)	**14**	**14** (2723, 6296)	**15**	**15** (13, 10)	**16**	**16** (11, 9)	**15**	**15** (22, 21)
	DYS458	**17**	**17** (10, 6)	**17** (56, 258)	**17** (31, 21)	**15**	**15** (152, 1300)	**19**	**19** (11, 9)	**17**	**17** (13, 12)	**17**	**17** (17, 24)
	DYS481	**22**	**22** (19, 21)	**22** (227, 1943)	**22** (150, 101)	**23**	**23** (663, 3830)	**25**	**25** (20, 18)	**23**	**23** (30, 43)	**22**	**22** (50, 45)
	DYS533	**12**	**12** (26, 36)	**12** (184, 0)	**12** (30, 10)	**12**	**12** (511, 0)	**12**	**12** (37, 37)	**12**	**12** (34, 37)	**14**	**14** (41, 27)
	DYS549	**13**	**13** (44, 49)	**13** (743, 0)	**13** (151, 101)	**12**	**12** (2649, 0)	**13**	**13** (39, 38)	**13**	**13** (46, 58)	**13**	**13** (60, 62)
	DYS570	**17**	**17** (44, 51)	**17** (646, 0)	**17** (76, 28)	**20**	**20** (777, 0)	**18**	**18** (64, 69)	**17**	**17** (73, 66)	**17**	**17** (145, 134)
	DYS576	**19**	**19** (45, 43)	**19** (0, 341)	**19** (3, 11)	**17**	**17** (0, 512)	**19**	**19** (34, 22)	**18**	**18** (74, 74)	**18**	**18** (65, 57)
	DYS635	**24**	**24** (10, 5)	**24** (220, 71)	**24** (18, 20)	**25**	**25** (774, 700)	**23**	**23** (18, 15)	**24**	**24** (33, 29)	**23**	**23** (30, 27)
	DYS643	**10**	**10** (43, 25)	**10** (2392, 989)	**10** (221, 111)	**10**	**10** (3314, 1407)	**11**	**11** (10, 15)	**11**	**11** (34, 30)	**10**	**10** (46, 43)
	GATA H4	**12**	**12** (23, 21)	**12** (290, 2196)	**12** (97, 93)	**12**	**12** (511, 3795)	**11**	**11** (21, 27)	**11**	**11** (34, 47)	**11**	**11** (33, 40)
	Total alleles	23	20	19	21	23	19	23	23	23	23	23	23

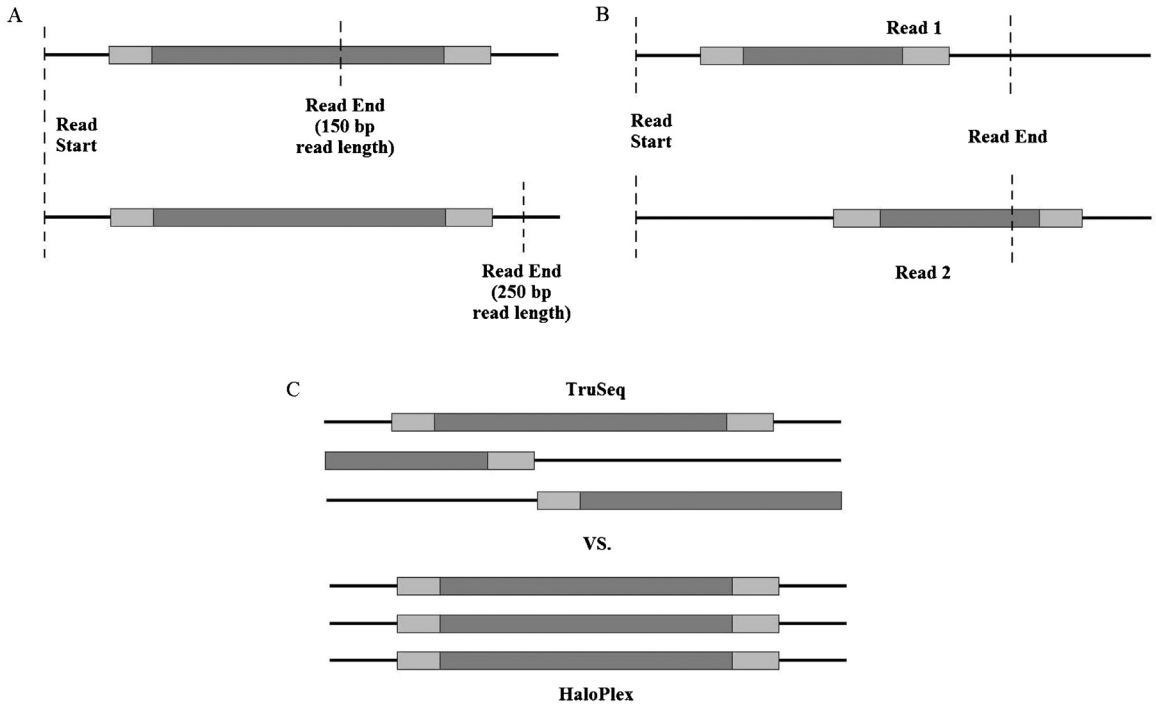

Fig. 2. Read length-related issues. In this figure, the dark gray bars represent the repeat region, while the light gray bars represent the flanking regions. The bold black lines represent surrounding sequence data. (A) Two identical fragments, such as those produced by the HaloPlex™ chemistry, are sequenced with 2 different read lengths. (B) A repeat region situated within a fragment is sequenced toward the beginning of the read in Read 1 of a paired-end sequencing run, but is sequenced toward the end of the read in Read 2 of the paired-end run. (C) A comparison of repeat region location consistency between the TruSeq™ and HaloPlex™ library preparation chemistries. This illustration is an example of a small subset of sequence reads, and does not reflect the actual proportions of reads produced by each library preparation method.

1, where the alleles are detected only in one set of paired-end reads and not the other. Some library preparation redesign may overcome the truncated repeat region reads.

The TruSeq™ chemistry [25] is less prone to these issues because the random fragmentation of DNA allows for a much more diverse positioning of repeat regions within the resulting fragments (Fig. 2c). Therefore, there is a higher likelihood of at least some reads encompassing the entire repeat region. This design explains why the majority of the alleles that were not detected by STRait Razor for Sample 1 following HaloPlex™ preparation and GAIIx™ sequencing were detected normally by the program for this sample following TruSeq™ preparation on the same instrument. Despite this advantage, the non-enzymatic random fragmentation employed by the TruSeq™ chemistry may result in lower read counts for some alleles in comparison with HaloPlex™, due to the fewer resulting fragments containing the complete repeat region. In some cases, the random fragmentation method simply may not create any fragments that contain the repeat region of interest. This limitation may explain the undetected alleles in Sample 1 at loci DYS439 and DYS448 following TruSeq™ preparation and GAIIx™ sequencing. Coverage depth differences in the results of this study, however, also may be explained by the fact that the regions targeted by the TruSeq™ kit for these trials were, by design, approximately 100 times larger than those targeted by the HaloPlex™ kit. Finally, it should be noted that in cases where the larger allele of a heterozygous pair at a given locus is not detected due to read length issues, a detected stutter allele may give the impression of a true heterozygous allele. Such issues may eventually be overcome by a combination of secondary statistical allelotyping software that takes into account the allele coverage ratios, and customization of the allelic information modules utilized by STRait Razor for the specific chemistry and instrumentation used for sequencing.

It should be noted that sequence variations (e.g., insertions or deletions) that reside outside of the flanking regions used by STRait Razor but within the primer-binding sites used by commercial STR/CE kits may result in discordant results between SGS and CE analyses. While this discordance was not observed during the course of this study, most likely due to limited sampling, it is not unique to SGS. Such discrepancies can occur due to different primer-binding site locations between various constructs of CE-based kits, as well.

The small portions of repeat sequence that are exactly matched to reads during the filtering portion of the algorithm were deliberately kept short (~3 simple repeat units or less). This exact matching step is important, as it removes irrelevant sequence data that may have been captured due to chance flanking sequence homology. However, the use of exact matching can potentially remove repeat regions with intra-repeat variation. Thus, sequences of repeats that are shorter than the smallest known allele at each locus are used, so that they may align at various points throughout the captured reads and still allow reads with intra-repeat variation to be retained. The Penta D locus, however, requires special attention, because its smallest known allele is "1.1" [26], albeit an extremely rare allele. The repeat sequence used by STRait Razor for exact match filtering at this locus is only one repeat unit long, since using a longer sequence would not allow for intra-repeat variability detection in the smaller alleles. The use of such a small portion of repeat sequence results in less thorough filtering of the reads, but this does not appear to affect the concordance of the resulting allelic data.

While the results of this study are encouraging, it should be noted that all software has its strengths and limitations. STRait Razor is simply an alternative method of approaching an increasingly important facet of SGS data analysis. The software is being offered freely so that those working with SGS and STR detection have an additional tool to facilitate their analyses. STRait

Razor is a fairly straightforward program, and the authors encourage enhancements to this software. With this phase of STRait Razor, probabilities of allele calls are not provided, nor does the program perform any stutter filtering or attempt to differentiate between heterozygosity and homozygosity. The software only reports the allele calls and related coverage at each locus; the analyst makes the final decision in an informed manner. A separate program is currently in development that reads the colon-delimited text file(s) output by STRait Razor and detects the two allele calls with the highest count values (if any), comparing them based on a user-defined "abundance ratio" to make homozygous and heterozygous allelotype calls with confidence. Presently, STRait Razor does not filter reads based on FASTQ quality information. However, the flanking region detection and small-repeat match verification performed by STRait Razor can be considered inherent filtering steps. Quality-based filtering may be incorporated into future versions of the program. In its current format, STRait Razor is designed to detect all the known alleles at each of the tested loci, according to the allelic information listed in STRBase and the Y-Chromosome Haplotype Reference Database (YHRD) [27,28]. These alleles are defined in the modules used by the software and may be modified by the user to include other rare or undocumented variants. In the future, the software may be altered to allow for the intuitive calling of alleles based on repeat length alone, without the need for allelic definitions.

5. Conclusion

In its current state of development, STRait Razor offers forensic DNA analysts a simple and effective means of detecting STR alleles in SGS data. Its ability to do so is limited primarily by the capabilities of the library preparation chemistries and sequencing platforms used to generate the raw data. The software has been shown to properly and efficiently analyze data resulting from a variety of commercially available library preparation chemistries and sequencers. The results of this study also indicate that the software is capable of identifying alleles in SGS sequence data with 100% concordance. STRait Razor provides multiple options for customization, and new loci can be added for detection at the discretion of the analyst. The program is freely available (see Supplementary Materials), and the scientific community is encouraged to make improvements to STRait Razor to increase its utility as a forensic tool.

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2013.04.005.

Acknowledgments

This project was supported in part by Award No. 2012-DN-BX-K033, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect those of the U.S. Department of Justice.

The authors would like to thank Illumina, Inc. and Agilent Technologies, Inc. for providing their technical expertise and contributing a portion of the sample preparation and sequencing chemistries used in this study.

References

[1] A. Edwards, A. Civitello, H.A. Hammond, C.T. Caskey, DNA typing and genetic mapping with trimeric and tetrameric tandem repeats, Am. J. Hum. Genet. 49 (1991) 746–756.
[2] A. Edwards, H.A. Hammond, L. Jin, C.T. Caskey, R. Chakraborty, Genetic variation at five trimeric and tetrameric repeat loci in four human population groups, Genomics 12 (1992) 241–253.
[3] H. Ellegren, Microsatellites: simple sequences with complex evolution, Nat. Rev. Genet. 5 (2004) 435–445.
[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 (1998) 86–93.
[5] 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 (2004) 51–53.
[6] P.J. Collins, L.K. Hennessy, C.S. Leibelt, R.K. Roby, D.J. Reeder, P.A. Foxall, Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: the AmpFlSTR® Identifiler® PCR Amplification Kit, J. Forensic Sci. 49 (2004) 1265–1277.
[7] K. Oostdik, J. French, D. Yet, B. Smalling, C. Nolde, P.M. Vallone, E.L. Butts, C.R. Hill, M.C. Kline, T. Rinta, A.M. Gerow, S.R. Allen, C.K. Huber, J. Teske, B. Krenke, M. Ensenberger, P. Fulmer, C. Sprecher, Developmental validation of the PowerPlex® 18D system, a rapid STR multiplex for analysis of reference samples, Forensic Sci. Int. Genet. 7 (2013) 129–135.
[8] M. Gymrek, D. Golan, S. Rosset, Y. Erlich, lobSTR: a short tandem repeat profiler for personal genomes, Genome Res. 22 (2012) 1154–1162.
[9] D.M. Bornman, M.E. Hester, J.M. Schuetter, M.D. Kasoji, A. Minard-Smith, C.A. Barden, S.C. Nelson, G.D. Godbold, C.H. Baker, B. Yang, J.E. Walther, I.E. Tornes, P.S. Yan, B. Rodriguez, R. Bundschuh, M.L. Dickens, B.A. Young, S.A. Faith, Short-read, high-throughput sequencing technology for STR genotyping, Biotechniques – Rapid Dispatches (2012) 1–6.
[10] S.L. Fordyce, M.C. Ávila-Arcos, E. Rockenbauer, C. Børsting, R. Frank-Hansen, F.T. Petersen, E. Willerslev, A.J. Hansen, N. Morling, M.T. Gilbert, High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, Biotechniques 51 (2011) 127–133.
[11] M.M. Holland, M.R. McQuillan, K.A. O'Hanlon, Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy, Croat. Med. J. 52 (2011) 299–313.
[12] R. Nielsen, J.S. Paul, A. Albrechtsen, Y.S. Song, Genotype and SNP calling from next-generation sequencing data, Nat. Rev. Genet. 12 (2011) 443–451.
[13] D.W. Craig, J.V. Pearson, S. Szelinger, A. Sekar, M. Redman, J.J. Corneveaux, T.L. Pawlowski, T. Laub, G. Nunn, D.A. Stephan, N. Homer, M.J. Huentelman, Identification of genetic variants using bar-coded multiplexed sequencing, Nat. Methods 5 (2008) 887–893.
[14] D.C. Koboldt, K. Chen, T. Wylie, D.E. Larson, M.D. McLellan, E.R. Mardis, G.M. Weinstock, R.K. Wilson, L. Ding, VarScan: variant detection in massively parallel sequencing of individual and pooled samples, Bioinformatics 25 (2009) 2283–2285.
[15] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303.
[16] E.C. Berglund, A. Kiialainen, A.C. Syvänen, Next-generation sequencing technologies and applications for human genetic history and forensics, Investig. Genet. 2 (2011) 1–15.
[17] Illumina® GAIIx™ Specifications: http://www.illumina.com/documents/products/datasheets/datasheet_genome_analyzerIIx.pdf.
[18] Illumina® MiSeq™ Specifications: http://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf.
[19] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.
[20] CASAVA v.1.8.2: http://support.illumina.com/downloads/casava_182.ilmn.
[21] MiSeq Reporter: http://support.illumina.com/sequencing/sequencing_software/miseq_reporter/downloads.ilmn.
[22] AGREP: http://laurikari.net/tre/download/.
[23] PPSS: http://code.google.com/p/ppss/.
[24] Agilent Technologies® HaloPlex™ Specifications: http://www.genomics.agilent.com/GenericB.aspx?pagetype=Custom&subpagetype=Custom&pageid=3081.
[25] Illumina TruSeq Specifications: http://www.illumina.com/Documents/%5Cproducts%5Cdatasheets%5Cdatasheet_truseq_custom_enrichment_kit.pdf.
[26] Penta D Facts Sheet (STRBase): http://www.cstl.nist.gov/strbase/str_Penta_D.htm.
[27] STRBase: http://www.cstl.nist.gov/strbase/str_fact.htm.
[28] Y-Chromosome Haplotype Reference Database: http://www.yhrd.org/Research/Loci.