Forensic Science International: Genetics 29 (2017) 21–28

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Short communication

STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems

Jonathan L. King[a,*], Frank R. Wendt[a], Jie Sun[b], Bruce Budowle[a,c]

[a] Center for Human Identification, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[b] Institute of Molecular Medicine, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[c] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 6 January 2017
Received in revised form 6 March 2017
Accepted 9 March 2017
Available online 12 March 2017

Keywords:
STRait Razor
Massively parallel sequencing
Short tandem repeat
Bioinformatics
Microhaplotypes
SNPs

ABSTRACT

STRait Razor has provided the forensic community a free-to-use, open-source tool for short tandem repeat (STR) analysis of massively parallel sequencing (MPS) data. STRait Razor v2s (SRv2s) allows users to capture physically phased haplotypes within the full amplicon of both commercial (ForenSeq) and "early access" panels (PowerSeq, Mixture ID). STRait Razor v2s may be run in batch mode to facilitate population-level analysis and is supported by all Unix distributions (including MAC OS). Data are reported in tables in string (haplotype), length-based (e.g., vWA allele 14), and International Society of Forensic Genetics (ISFG)-recommended (vWA [CE 14]-GRCh38-chr12:5983950-5984049 $(TAGA)_{10}$ $(CAGA)_3$ TAGA) formats. STRait Razor v2s currently contains a database of ~2500 unique sequences. This database is used by SRv2s to match strings to the appropriate allele in ISFG-recommended format. In addition to STRs, SRv2s has configuration files necessary to capture and report haplotypes from all marker types included in these multiplexes (e.g., SNPs, InDels, and microhaplotypes). To facilitate mixture interpretation, data may be displayed from all markers in a format similar to that of electropherograms displayed by traditional forensic software. The download package for SRv2s may be found at https://www.unthsc.edu/graduate-school-of-biomedical-sciences/molecular-and-medical-genetics/laboratory-faculty-and-staff/strait-razor.

1. Introduction

STRait Razor [1] initially was developed to capture and characterize variation in the repeat motifs of target short tandem repeat (STR) markers from next generation sequencing or, the more aptly named, massively parallel sequencing (MPS) data. Since its inception, STRait Razor has been used for length-based (LB) analysis and repeat motif variation of population data [2–4], assessment of novel MPS multiplexes [5,6], alignment-free characterization of insertion-deletion (InDel) polymorphisms [7] and other novel applications [8–10]. While STRait Razor v2 [11] improved allelic reporting and expanded target loci, the flanking regions of STR loci (and other marker types such as single nucleotide polymorphisms (SNPs) and InDels) remained largely unreported without modification to the included configuration files [12] or use of alternative software [13–16].

Variation in the flanking regions of STRs has been used to study human evolution and migration patterns [17–19] and increasing the probability of exclusion [20]. These combinations of SNP and STR (SNPSTR) loci phased within the same amplicon provide a finer granularity to better discriminate individuals. In an effort to standardize nomenclature, the International Society of Forensic Genetics (ISFG) published a set of considerations [21] regarding reporting of STR alleles, repeat and flanking regions, for comprehensive and consistent nomenclature.

The proposed ISFG nomenclature [21] is not implemented currently into commercial or third-party software. Researchers must convert their data manually which may be error prone. Traditional alignment software [22,23], while effective for genotyping SNPs, produces inconsistent results on a per read basis. For example, the Burrows-Wheeler Aligner [22], widely used for read mapping, places the insertion/deletion points in relation to the reference at different points within the repeat region when considering sequence variants and/or the end point(s) of each read. However, direct haplotype capture used by various software tools [16,24], including STRait Razor, allows users to extract phased data of flanking region variants as well as the target locus. As data

* Corresponding author.
E-mail address: jonathan.king@unthsc.edu (J.L. King).

http://dx.doi.org/10.1016/j.fsigen.2017.03.013
1872-4973/

analysis pipelines for this task are still nascent to the field of forensic genetics, bioinformatics concordance with operationally distinct methodologies are critical to ensure complete, as possible, results are obtained. Amplified products separated by capillary electrophoresis (CE) provide analysts with a comprehensive allele determination in respect to size; however, sequence characterization, thus far, has been regulated to the repeat region of the amplicons [5,6,25–28]. While effective, this approach to allele reporting has been shown to be limited in its informative value [4,14,20] and, in some cases, its backwards compatibility with CE data [2,4,14]. Therefore, characterization of the flanking region of loci is necessary to realize the full potential of MPS systems.

Bi-allelic loci (e.g., SNPs and InDels) are useful particularly in challenged samples. Kidd et al. [29] have shown the utility of combining closely linked SNPs (<200 bp) into microhaplotype loci phased within a single amplicon. However, interpretation of these small amplicon markers has been limited to the target SNP(s) of interest and have ignored potential variation in the surrounding region of the amplicon. However, variation along the entire amplicon in the form of InDels-SNPs [7], DIP-STRs [30], SNPSTRs [17–20,31], or microhaplotypes [29,32–35] may be present at some level for every forensic locus but are ignored, as yet, by first-party software. More recently, these additional data have been shown to increase the discrimination power of commercially available forensic-genomics assays from 8.54×10^{-34} to 1.31×10^{-39} for LB and sequence-based (SB) STRs and 7.66×10^{-58} to 5.49×10^{-63} when identity SNPs/microhaplotypes are included [4,36].

The two primary sequencing chemistries currently being considered for forensic applications are those of the Illumina MiSeq FGx™ Forensic Genomics System (Illumina, San Diego, CA, USA) and the Ion Torrent PGM™ and S5™ (Thermo Fisher Scientific, San Francisco, CA, USA). Each chemistry has a distinct detection method which generates unique interpretation considerations. Substitution errors (SBEs) are the primary source of error within the Illumina sequencers [37] with a relatively small proportion of insertion/deletion errors typically associated with specific sequences (sequence-specific error; SSE) [37–39]. Schirmer et al. [38] used 16S rRNA amplicons to estimate the source and distribution of sequencing error with the MiSeq 2×250 sequencing chemistry. They concluded that the accumulation of phasing and pre-phasing errors over the course of the sequencing read complicated base calling and lead to a concomitant increase in sequence errors or miscalls. There was an observed increase in sequencing error approximately at position 200–225 within the sequencing reads using MiSeq v2 2×250 sequencing chemistry. Data from the Ion Torrent, however, have a relatively high false InDel rate associated with homopolymeric stretches in conjunction with SBEs [40–43]. Bragg et al. [40] also observed a marked increase in SBE rate for both the 100 and 200 bp chemistries as the read length increased at approximately 75 and 150 bp, respectively. These sequencing errors may either complicate interpretation of the data by producing relatively high-abundance, non-target haplotypes or affect the ability of the configuration files to capture the desired haplotypes. With these limitations in mind, optimization with respect to bioinformatics of each system on a per-marker basis may be necessary to ensure application beyond single-source reference samples.

With the advent of commercial MPS multiplexes, forensic practitioners now have the capacity to interrogate a large number of markers and marker types (e.g., STRs, SNPs, InDels, and microhaplotypes) within a single analysis [4–6,27,44–48]. However, orthogonal bioinformatics solutions for phased data including microhaplotypes have been limited to traditional alignment-based software [49,50] which provide only computationally derived phasing. STRait Razor v2s (the "s" designation referring to the addition of SNP loci, SRv2s) is a freely available update that provides a direct haplotype capture approach that includes physical phasing to determine the diplotype (i.e., a specific combination of haplotypes at a particular site in an individual analogous to the genotype of alleles) at each locus (STRs, SNPs, microhaplotypes, and InDels) from current (or soon to be available) commercially available forensic multiplex assays.

2. Materials and methods

2.1. New features

STRait Razor v2s suite of tools features kit-specific locus-configuration files for the Applied Biosystems™ Precision ID GlobalFiler™ Mixture ID panel (Thermo Fisher Scientific), Illumina® ForenSeq™ DNA Signature Prep Kit (Illumina), and Promega PowerSeq™ Systems (Auto, Y, and Mito) (Promega Corporation, Madison, WI, USA). In addition to these multiplexes developed expressly for MPS, configuration files were included for forensically relevant InDels and the repeat-region configuration files for current STRs. For the purposes of this paper and supporting documentation the following definitions are visualized using IGV [51]: anchors (Fig. 1-Red Track) are defined as the substring in the amplicon that is included within the configuration file to allow the Perl script to capture reads from each FASTQ file. These substrings may, and often do, include primer sequence (Fig. 1-Blue Track); however, this configuration may be redefined by the user as preferred. The intervening space between the distal portions of the amplicon and the repeat region (Fig. 1-Purple Track) is termed

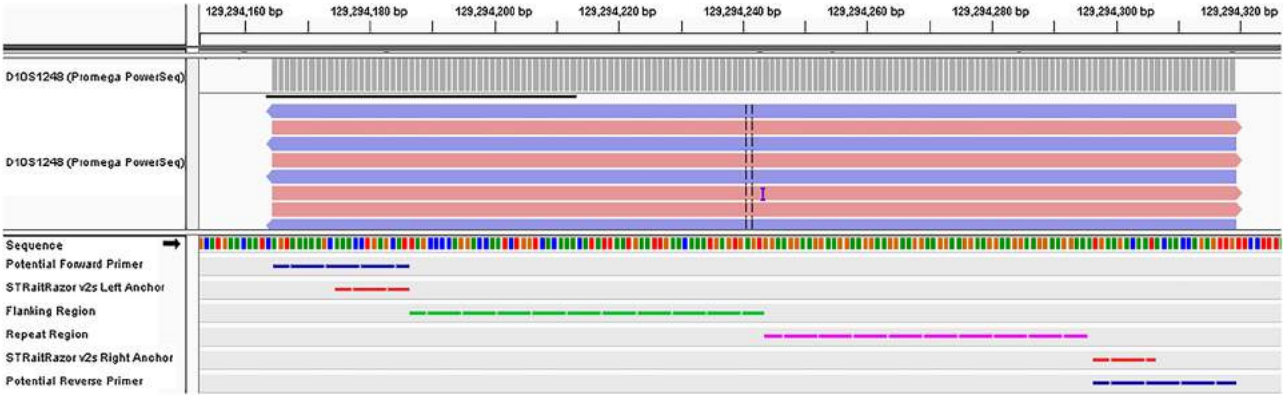

Fig. 1. Aligned reads from the Promega PowerSeq Auto System are visualized in IGV. Potential primers (blue), anchors used for SRv2s analysis (red), STR flanking region (green), and STR repeat region (purple) are displayed as tracks below the GRCh38 reference sequence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

herein as flanking region (Fig. 1-Green Track) of the locus (Fig. 1). The length of the flanking region varies with primer and anchor placement. For example, the right flank (not shown) for locus D10S1248 is a single base while the left flanking region contains 57 bases.

In addition to the substantial expansion of nearly 300 configured markers, support for novel microvariants has been included for all loci with expanded "virtual bin sets" comprising each potential base call within the allelic range rather than only observed nominal allele bins. Alongside the updates to the bins, the D5S2500 locus has been renamed to be consistent with Phillips et al. [52] and the locus D5S2500 found in the Qiagen HDplex has been added to the master configuration file. Furthermore, repeat-region configuration files have been updated for 19 loci to improve performance, to reduce noise due to high similarity flanking region sequence, to increase allelic range, and to update allele definitions for more comprehensive concordance with CE data (e.g., DYS612 previously failed to incorporate six repeats considered part of the repeat structure [53]).

STRait Razor v2s also includes a number of performance-based improvements to streamline the interface and enhance usability. First, samples now may be batched and analyzed with minimal command line knowledge. STRaitRazorBatchScript.sh is a shell script, included in the SRv2s package, which may be modified to analyze any directory path included in the script. All unzipped FASTQ files within the specified directory (folder) are placed in a queue and analyzed sequentially. Second, the Perl script has been modified to report reads now in the direction of the sequence found within the FASTQ as well as the reverse complement of each sequence. In this way, reads may be merged and reported out to a consistent strand of the reference genome (Default = Forward/Plus) in accordance with Parson et al. [21] while maintaining strand balance information. This feature is useful for assays (e.g., PowerSeq and AmpliSeq panels) that sequence both strands of the amplicon. This function may be modified within the Excel-based workbook to match each user's requirements.

STRait Razor Analysis (SRA) is a consolidation of the previous versions suite of Excel-based tools. The SRA tool allows users to place hundreds of SRv2s samples analyzed with the Perl script into a queue for analysis under user-defined conditions. The SRA tool features an AutoSave option (Default = Off; every 10th sample) that will save the workbook periodically to allow offline viewing mid-analysis for large datasets. An analytical threshold (Default = 2; i.e., singletons removed) may be defined on a per-locus basis. Additionally, strand balance and heterozygote balance may be delimited on a per-locus basis. Thus, loci with known imbalances (e.g., D22S1045) may now be properly filtered in a more dynamic fashion. STRait Razor v2s also expands the number of metrics assessed and exported from the SRA. The previous version of STRait Razor provided users with a depth of coverage for each allele and the LB allele name for backwards-compatibility to the CE. In addition to these data, SRv2s provides data tables and histograms in a manner similar to that of electropherograms displayed by traditional forensic tools such as GeneMapper and GeneMarker. These emulated electropherograms may facilitate mixture interpretation by providing a profile-level view of data rather than locus-by-locus. These data to be exported may be toggled on/off using radio buttons embedded within the workbook. In this way, users may choose the scope of the data to be output by SRv2s. These metrics currently include: allele coverage, string or haplotype sequence, LB data (e.g., vWA 14 allele), ISFG-based allele designation (e.g., vWA [CE 14]-GRCh38-chr12:5983950-5984049 $(TAGA)_{10}$ $(CAGA)_3$ TAGA), strand balance (when applicable), relative allele percentage (i.e., proportion of haplotype coverage to the locus as a whole; *Haplotype coverage/Locus coverage*), stutter percentage, heterozygote balance, sequence diversity, and locus coverage.

ISFG-based alleles are designated in accordance with Parson et al. [21] and matched against a preloaded database of, currently, ~2500 unique sequences derived from previous population studies [4,36]. However, novel sequences are expected; therefore, macros have been included to allow easy assignment of novel SNPs and/or repeat motifs. These newly assigned alleles may be added to a local database or submitted for addition to the global database provided with future STRait Razor updates. It is understood that users may possess data from earlier versions of STRait Razor or alternative software. For added flexibility, an additional worksheet has been included with the SRA that accepts string sequences from any source, including other software, and searches the database within STRait Razor for the matching ISFG-based allele. These macros align the novel sequence selected to the preloaded GRCh38 reference string of the same locus.

As MPS becomes more commonplace, the versatility of platforms is necessary to ensure long-term acceptance. The macros provided with STRait Razor v2.0 were restrictive in terms of the operating system used. In SRv2s, this limitation has been minimized by designing a MAC-friendly version of the SRA. As the MAC OS is a UNIX-based environment, STRait Razor always has been compatible with MAC OS in regards to the Perl script. However, to aid users, a walkthrough is included now as a downloadable that details installation and utilization on the MAC OS of both the SRA and the Perl script.

Novroski et al. [4] described observed and potential discordant haplotypes in a population dataset of 777 individuals across 4 major populations when comparing repeat-region-only data to CE results. To account for these positions, configuration files for the

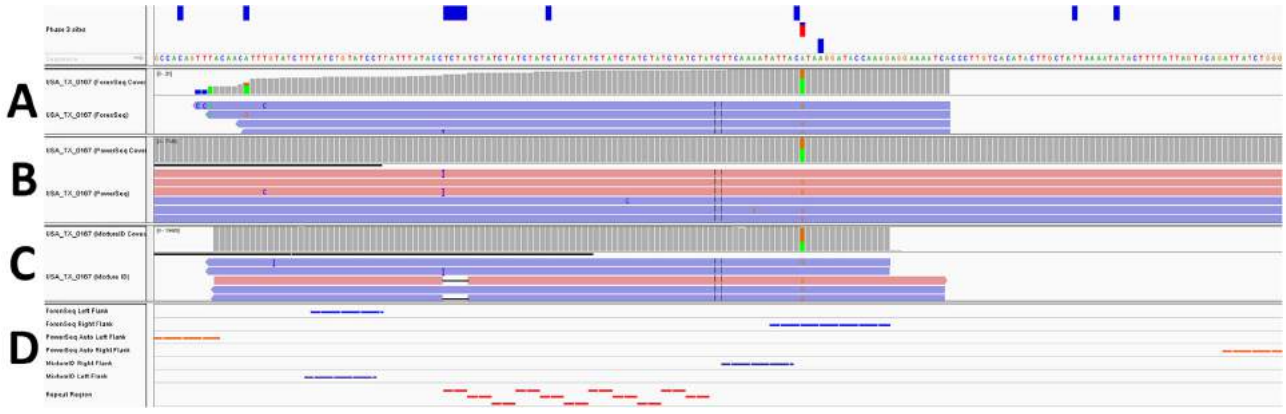

Fig. 2. IGV depiction of the aligned reads from A) ForenSeq; B) PowerSeq; C) Mixture ID. Repeat region, left, and right anchors for each kit are shown in panel D.

repeat-region data were generated with anchors adjusted for variants above a minimum allele frequency (i.e., 5/2N; MAF) or those variants present in more than one LB allele.

2.2. STRait Razor v2s configuration files

STRait Razor v2s configuration files for both commercially available and early access multiplexes were generated by aligning FASTQ files to the GRCh38 assembly of the human genome. Genomic coordinates for each locus were determined by post-processing of the mpileup output from SAMtools [23] in Microsoft Excel. Using these start/stop coordinates, operationally defined anchors were selected to capture both flanking region and repeat region data for all markers contained within the PowerSeq, Mixture ID, and ForenSeq panels. All samples were analyzed with both repeat-region only and full-region configuration to determine the coverage of both loci and alleles with each method. Optimization then was carried out to improve locus performance and reduce noise.

3. Results and discussion

Each large MPS multiplex presents unique challenges for bioinformatics development. When considering the placement of the anchors for each assay, multiple factors (e.g., size of the amplicon, presence of SSEs, sequencing platform, flanking region SNPs, etc.) should be considered. Previous studies have used the PCR primers for anchor assignment [24,54]. Anchors based on PCR primers allow for capture of the maximum amount of information from an amplicon in the form of both repeat region and flanking region SNPs. However, each MPS multiplex likely uses a different set of primers for each locus. For example, the flanking region of the locus D5S818 has a SNP (rs25768) with a frequency ranging from 0.05 in East Asians (EAS) to 0.27 in Europeans (EUR) in the five 1000 Genomes Project super populations [55,56]; however, in the ForenSeq panel, this SNP likely lies within the likely PCR primer binding site (Fig. 2). Thus, little inference may be made of the allelic state of rs25768 in this panel. Conversely, the primers for both the PowerSeq and Mixture ID panels are placed outside of this SNP. However, only the PowerSeq panel data may be phased properly. The results from the Mixture ID panel contain SSE due to a four base pair homopolymer immediately downstream (three base pairs) of the repeat region which complicates interpretation. Due to such variability in primer placement and platform chemistry, each locus configuration file required distinct optimization.

3.1. Optimization of commercial multiplexes

3.1.1. ForenSeq DNA signature prep kit

The Illumina ForenSeq kit generates ~225 amplicons with a listed amplicon size range between 61 and 467 bp. While the vast majority of loci (~90%) are smaller than 200 bp, the larger, primarily X & Y-STR, loci present a technical limitation when considering sequence quality. In fact, 20 STR loci needed to be truncated to allow successful typing of larger length alleles. Two such STRs (Penta E and DXS8378) contained, potentially, informative [frequency ranging from 0 to 0.10 in the African (AFR) super population [55,56]] SNPs within the flanking region that were removed because of larger repeat regions. In analysis of the Novroski et al. [4] data, the proportion of noise for the Penta E locus was positively correlated with the length of the amplicon making CE-based alleles ~20 or greater difficult to distinguish from non-target sequences. The final placement of anchors in this dataset allowed for minimal loss of data with a ratio of full region to repeat region when quantifying total coverage at a locus ranged from 0.96 (PentaD) to 1.00 (22 loci).

Additionally, anchors for three SNP loci (rs914165, rs891700, and rs7251928) were moved to a more proximal position excluding the problematic homopolymers. These homopolymers appear to increase the SBE rate downstream of the homopolymer. Two of the three loci have variants within the amplicon with relatively high frequency (e.g., rs750095 has a frequency range from 0.08 in AFR to 0.40 in EUR [55,56]) which were described previously by Eduardoff et al. [46]. The observed SBEs are nested between the target SNP (e.g., rs914165) and the incidental variants within the amplicon (e.g., rs750095). These SBEs presented were of relatively low abundance (~2.5% of the parent allele). However, this type of error may complicate interpretation for single-source samples with exceedingly low depth-of-coverage (DoC) or mixed samples with trace-level contributors. Thus, the anchors for these loci were placed in a more proximal position to the SBEs so that the error was at a lower abundance (~0.8%).

Anchor reassignment makes phasing of some of the linked SNPs difficult; however, the flanking region SNPs still may contain useful information for the purposes of exclusion. To capture these data, separate anchors were developed for the flanking region SNPs of these two amplicons. These SNP pairs (rs914165–rs755095; rs891700–rs12047255), though analyzed separately within SRv2s, are within the same amplicon. In some instances, it may be possible to infer the phase when quantifying the abundance of each SNP within the amplicon.

The accumulation of SBEs near the 3′ end of reads was problematic particularly when analyzing ForenSeq data due to sequencing in a unidirectional manner [57,58]. Filtering of false positive SNPs has been shown to be effective by considering strand balance or Simpson's *D* Index of evenness when interpreting potential SNPs and, likely, SBEs within a dataset [59]. Thus, the observed errors may be better filtered by considering reads in both directions during interpretation.

3.1.2. PowerSeq systems

The PowerSeq Systems provide high quality sequence data in both forward and reverse orientation for 43 STRs (22 Auto & 21 Y-chromosome), mitochondrial control region (HVI-III), and Amelogenin. PowerSeq Auto has been shown to be useful for analysis of both single source [6] and mixture samples [60]. However, only limited interrogation of the flanking region has been reported [2]. To further examine this flanking region variation, anchors were defined, and the data described by Zeng et al. [5,6] were analyzed. For the 138 FASTQ files analyzed, the ratio of full region to repeat region when quantifying total coverage at a locus ranged from 0.94 (TH01) to 1.00 (D1S1656) with 17 of the 22 STRs $\geq$0.97. Therefore, the reduction in coverage when considering the entire amplicon versus, for example, the repeat region only may be negligible with this dataset. As more population data are made available, further optimization of anchor placement is likely to occur for these loci. Despite the limited sample size, substantial increases in number of alleles were observed with three loci (D5S818, D7S820, and D16S639) consistent with the findings of Gettings et al. [2]. Three additional loci (D18S51, D8S1179, and TH01) showed nominal improvement with one additional allele each observed from this admittedly small dataset. As more population data are made available, improvements will be made to the PowerSeq Systems configuration files.

3.1.3. Precision ID GlobalFiler™ Mixture ID panel

The Mixture ID panel is currently an "early access" panel consisting of 30 STRs, 43 human identity SNPs, 36 human identity microhaplotypes, and 2 InDels. In this study, data from the ten individuals previously presented [61] were analyzed for all 111 markers included within the Mixture ID panel. As the panel has yet to be finalized, anchors for the human identity SNPs and

microhaplotypes will continue to be developed. Optimization of the flanking region configuration files for STRs typed with Ion Torrent chemistries required manual assignment of flanks for all loci. Homopolymer-associated noise in the form of false InDels varied locus-to-locus.

The sequence-related artifacts generated by the Ion Torrent systems may be visualized first by separating the haplotypes by unique sequence and then stacking those of similar length proportional to the observed abundance within a sample. For example, the amplicon for the locus D7S820 contains an 8 base adenine homopolymer 13 bp upstream of the repeat region (Fig. 3C). This homopolymer creates a number of distinct artifacts when interrogating the full amplicon (Fig. 3A) due to SSEs. These SSEs have the potential to complicate interpretation and must be removed prior to finalization of the reported regions. Selective placement of the anchors proximal to the false InDels improved

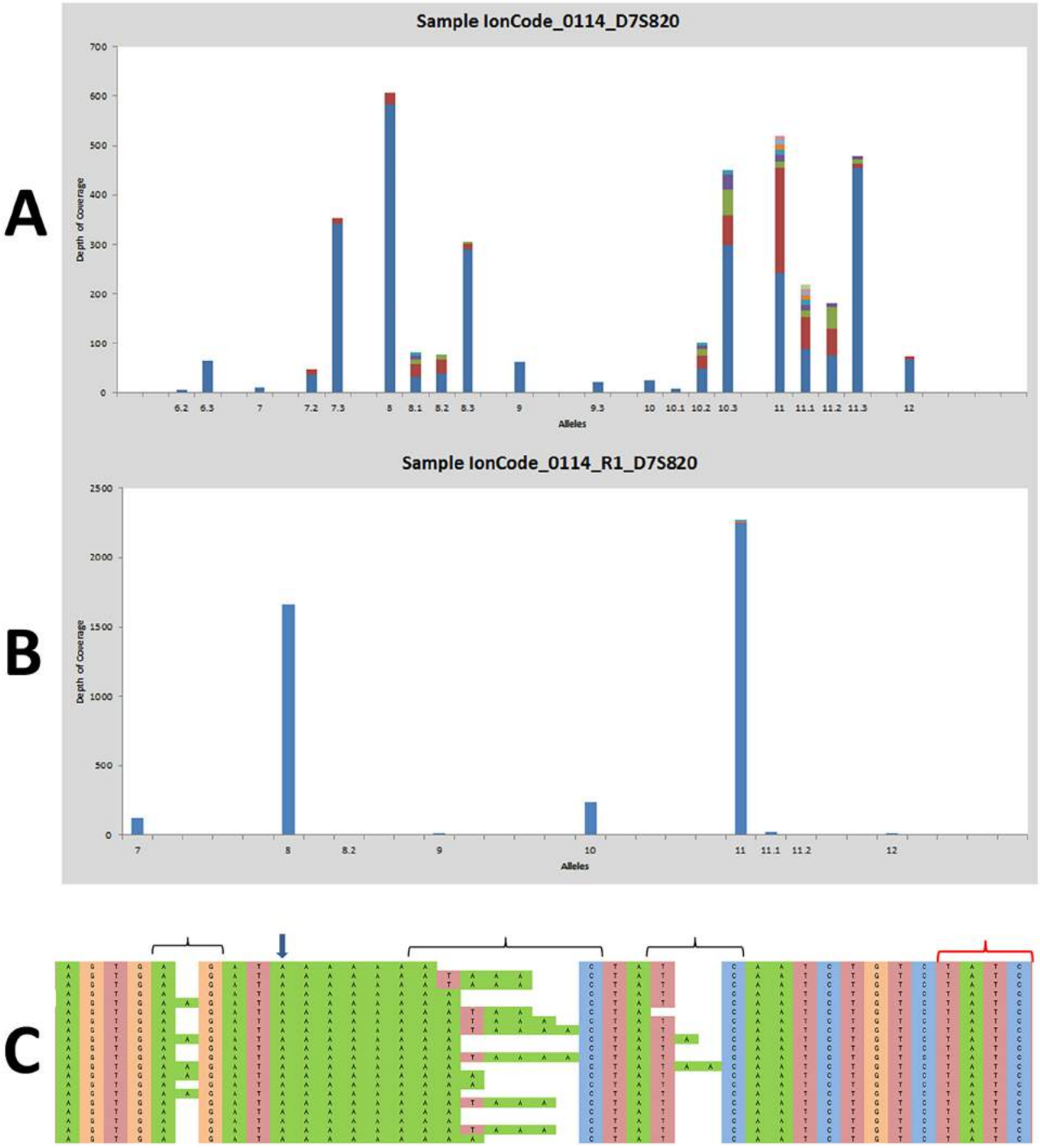

Fig. 3. Optimization of the D7S820 locus from Early Access HID Mixture ID NGS Panel. A) Depth of coverage histogram of unique sequences using distal anchors reporting the full region of the amplicon; B) Results using optimized coordinates proximal to the false INDELs generated by the Ion Torrent S5; and C) Expanded sequence view of most abundant noise (black brackets) generated surrounding the D7S820 locus. The flanking region SNP rs7789995 (blue arrow) is 13 bp upstream of the first repeat (red bracket) within the repeat region. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

interpretation of the locus by eliminating the artifact sequences and consolidating the observed reads for the locus into two primary sequences (Fig. 3B). A consequence of reducing the reportable region is the potential loss of flanking region information. The flanking region SNP rs7789995 (T→A), within the D7S820 amplicon, has a European MAF of 0.12 [55,56]. To further complicate matters, rs7789995 resides in the first position of the 8 base homopolymer and generates a 9 base homopolymer in most individuals (Fig. 3C). Thus, separate anchors were designed to capture this position. Placement of these apparent SSEs well outside the repeat region enables their excision from the reported data. When the SSEs are within or very near the repeat region, it is difficult, if not impossible, to excise the problematic region(s).

Initial testing of the locus D12S391 showed a moderate level of noise below the level of stutter (Fig. 4A). A subsequent shift of the anchors immediately adjacent to the repeat region to remove these SSEs exacerbated interpretation issues by consolidating the low-level noise into a single large artifact (Fig. 4B). Evaluation of the SSEs at the sequence level suggests interpretation of the locus may be aided by distributing the error across a larger reportable region with a more distal placement of the anchors closer to the PCR

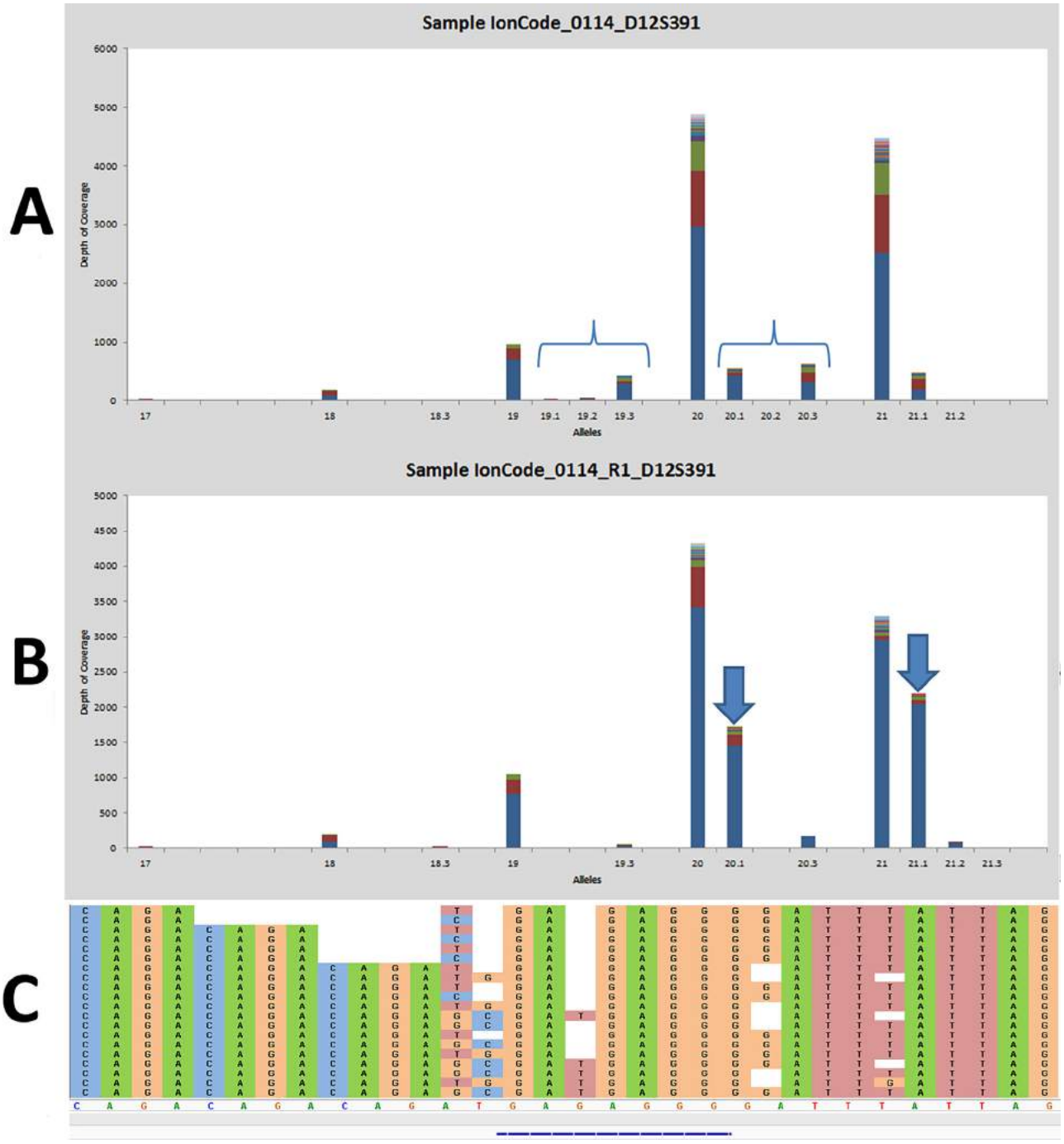

Fig. 4. Optimization of the D12S391 locus from Early Access HID Mixture ID NGS Panel. A) Depth of coverage histogram of unique sequences using distal anchors reporting the full region of the amplicon (brackets indicate the highest abundance noise); B) Results using optimized coordinates proximal to the false INDELs generated by the Ion Torrent S5 (arrows indicate noise consolidated by length); and C) Expanded sequence view of most abundant noise generated surrounding the D12S391 locus.

primers. This locus illustrates the need for additional bioinformatics solutions to better ascertain, map, and filter SSEs.

4. Conclusion

STRait Razor v2s uses a direct haplotype capture approach to extract phased data from entire amplicons of all marker types within current (or soon to be available) commercially available forensic multiplex assays. This software update provides users a database of, currently, ~2500 unique sequences matched with nomenclature in the format recommended by Parson et al. [21]. The sequences contained within this database are based on the anchors defined as reported herein; as primers define the sequences, modifications may be necessary to best capture relevant data.

This enhanced tool has been used already to detect novel variation from major population groups and will continue to be developed to meet the needs of the user community. STRait Razor v2s is freely available at https://www.unthsc.edu/graduate-school-of-biomedical-sciences/molecular-and-medical-genetics/laboratory-faculty-and-staff/strait-razor. Given the high-level of diversity observed thus far, novel variants likely will be observed. Thus, community feedback is greatly encouraged to improve the database.

Conflict of interest

The authors declare they have no conflict of interests.

Acknowledgements

This work was supported in part by award no. 2015-DN-BX-K067, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of the U.S. Department of Justice. The authors would like to thank Nicole Novroski, Jennifer Churchill, Lisa Borsuk, Lilliana Moreno, Ryan England, and Katherine Gettings for their contributions and invaluable discussions towards the development and improvement of this version of STRait Razor.

References

[1] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. LaRue, et al., STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data, Forensic Sci. Int. Genet. 7 (2013) 409–417.
[2] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, et al., Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Forensic Sci. Int. Genet. 21 (2016) 15–21.
[3] F.R. Wendt, J.D. Churchill, N.M.M. Novroski, J.L. King, J. Ng, R.F. Oldt, et al., Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx (TM) forensic genomics system, Forensic Sci. Int. Genet. 24 (2016) 18–23.
[4] N.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci. Int. Genet. 25 (2016) 214–226.
[5] X. Zeng, J.L. King, M. Stoljarova, D.H. Warshauer, B.L. LaRue, A. Sajantila, et al., High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing, Forensic Sci. Int. Genet. 16 (2015) 38–47.
[6] X. Zeng, J. King, S. Hermanson, J. Patel, D.R. Storts, B. Budowle, An evaluation of the PowerSeq Auto System: a multiplex short tandem repeat marker kit compatible with massively parallel sequencing, Forensic Sci. Int. Genet. 19 (2015) 172–179.
[7] F.R. Wendt, D.H. Warshauer, X. Zeng, J.D. Churchill, N.M. Novroski, B. Song, et al., Massively parallel sequencing of 68 insertion/deletion markers identifies novel microhaplotypes for utility in human identity testing, Forensic Sci. Int. Genet. 25 (2016) 198–209.
[8] R. England, S. Harbison, Massively parallel sequencing for the forensic scientist—sequencing archived amplified products of AmpFlSTR Identifiler and PowerPlex Y multiplex kits to capture additional information, Aust. J. Forensic Sci. (2016) 1–18.
[9] R.A. Aponte, K.B. Gettings, D.L. Duewer, M.D. Coble, P.M. Vallone, Sequence-based analysis of stutter at STR loci: characterization and utility, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e456–e458.
[10] E.H. Kim, H.Y. Lee, I.S. Yang, S.-E. Jung, W.I. Yang, K.-J. Shin, Massively parallel sequencing of 17 commonly used forensic autosomal STRs and amelogenin with small amplicons, Forensic Sci. Int. Genet. 22 (2016) 1–7.
[11] D.H. Warshauer, J.L. King, B. Budowle, STRait Razor v2.0: the improved STR allele identification tool–razor, Forensic Sci. Int. Genet. 14 (2014) 182–186.
[12] K.B. Gettings, R.A. Aponte, K.M. Kiesler, P.M. Vallone, The next dimension in STR sequencing: polymorphisms in flanking regions and their allelic associations, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e121–e123.
[13] S.Y. Anvar, K.J. van der Gaag, J.W. van der Heijden, M.H. Veltrop, R.H. Vossen, R. H. de Leeuw, et al., TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes, Bioinformatics 30 (2014) 1651–1659.
[14] S.L. Friis, A. Buchard, E. Rockenbauer, C. Børsting, N. Morling, Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs, Forensic Sci. Int. Genet. 21 (2016) 68–75.
[15] J.C.-I. Lee, B. Tseng, L.-K. Chang, A. Linacre, S.E.Q. Mapper, A DNA sequence searching tool for massively parallel sequencing data, Forensic Sci. Int. Genet. 26 (2017) 66–69.
[16] J. Hoogenboom, K.J. van der Gaag, R.H. de Leeuw, T. Sijen, P. de Knijff, J.F.J. Laros, FDSTools: a software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise, Forensic Sci. Int. Genet. 27 (2017) 27–40.
[17] J.L. Mountain, A. Knight, M. Jobin, C. Gignoux, A. Miller, A.A. Lin, et al., SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes, Genome Res. 12 (2002) 1766–1772.
[18] U. Ramakrishnan, J.L. Mountain, Precision and accuracy of divergence time estimates from STR and SNPSTR variation, Mol. Biol. Evol. 21 (2004) 1960–1971.
[19] A. Odriozola, J. Aznar, L. Valverde, S. Cardoso, M. Bravo, J. Builes, et al., SNPSTR rs59186128_D7S820 polymorphism distribution in European Caucasoid, Hispanic, and Afro-American populations, Int. J. Legal Med. 123 (2009) 527–533.
[20] H. Oberacher, F. Pitterl, G. Huber, H. Niederstätter, M. Steinlechner, W. Parson, Increased forensic efficiency of DNA fingerprints through simultaneous resolution of length and nucleotide variability by high-performance mass spectrometry, Hum. Mutat. 29 (2008) 427–432.
[21] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, et al., 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.
[22] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760.
[23] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., The Sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[24] C. Van Neste, M. Vandewoestyne, W. Van Criekinge, D. Deforce, F. Van Nieuwerburgh, My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing, Forensic Sci. Int. Genet. 9 (2014) 1–8.
[25] S.L. Fordyce, M.C. Avila-Arcos, E. Rockenbauer, C. Borsting, R. Frank-Hansen, F.T. Petersen, et al., High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, BioTechniques 51 (2011) 127–133.
[26] S.L. Fordyce, H.S. Mogensen, C. Borsting, R.E. Lagace, C.W. Chang, N. Rajagopalan, et al., Second-generation sequencing of forensic STRs using the ion torrent HID STR 10-plex and the ion PGM, Forensic Sci. Int. Genet. 14 (2015) 132–140.
[27] J.D. Churchill, S.E. Schmedes, J.L. King, B. Budowle, Evaluation of the Illumina ((R)) beta version ForenSeq DNA signature prep kit for use in genetic profiling, Forensic Sci. Int. Genet. 20 (2016) 20–29.
[28] A.D. Ambers, J.D. Churchill, J.L. King, M. Stoljarova, H. Gill-King, M. Assidi, et al., More comprehensive forensic genetic marker analyses for accurate human remains identification using massively parallel DNA sequencing, BMC Genomics 17 (2016) 750.
[29] K.K. Kidd, A.J. Pakstis, W.C. Speed, R. Lagacé, J. Chang, S. Wootton, et al., Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics, Forensic Sci. Int. Genet. 12 (2014) 215–224.
[30] V. Castella, J. Gervaix, D. Hall, DIP–STR: highly sensitive markers for the analysis of unbalanced genomic mixtures, Hum. Mutat. 34 (2013) 644–654.
[31] I. Agrafioti, M.P. Stumpf, SNPSTR: a database of compound microsatellite-SNP markers, Nucleic Acids Res. 35 (2007) D71–D75.
[32] K. Kidd, A. Pakstis, W. Speed, R. Lagace, J. Chang, S. Wootton, et al., Microhaplotype loci are a powerful new type of forensic marker, Forensic Sci. Int. Genet. Suppl. Ser. 4 (2013) e123–e124.
[33] K.K. Kidd, Proposed nomenclature for microhaplotypes, Hum. Genomics 10 (2016) 16.
[34] K.K. Kidd, W.C. Speed, Criteria for selecting microhaplotypes: mixture detection and deconvolution, Investig. Genet. 6 (2015) 1.
[35] K.K. Kidd, W.C. Speed, S. Wootton, R. Lagace, R. Langit, E. Haigh, et al., Genetic markers for massively parallel sequencing in forensics, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e677–e679.

[36] F.R. Wendt, J.L. King, N.M.M. Novroski, J.D. Churchill, J. Ng, R.F. Oldt, et al., 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.
[37] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, et al., Sequence-specific error profile of Illumina sequencers, Nucleic Acids Res. 39 (2011) e90.
[38] M. Schirmer, U.Z. Ijaz, R. D'Amore, N. Hall, W.T. Sloan, C. Quince, Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform, Nucleic Acids Res. 43 (2015) e37.
[39] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, et al., A tale of three next generation sequencing platforms: comparison of Ion Torrent Pacific Biosciences and Illumina MiSeq sequencers, BMC Genomics 13 (2012) 341.
[40] L.M. Bragg, G. Stone, M.K. Butler, P. Hugenholtz, G.W. Tyson, Shining a light on dark sequencing: characterising errors in ion torrent PGM data, PLoS Comput. Biol. 9 (2013) e1003031.
[41] W. Parson, C. Strobl, G. Huber, B. Zimmermann, S.M. Gomes, L. Souto, et al., Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM), Forensic Sci. Int. Genet. 7 (2013) 543–549.
[42] S.B. Seo, X. Zeng, J.L. King, B.L. Larue, M. Assidi, M.H. Al-Qahtani, et al., Underlying data for sequencing the mitochondrial genome with the massively parallel sequencing platform ion torrent PGM, BMC Genomics 16 (Suppl. 1) (2015) S4.
[43] J.D. Churchill, J.L. King, R. Chakraborty, B. Budowle, Effects of the Ion PGM Hi-Q sequencing chemistry on sequence data quality, Int. J. Legal Med. 130 (2016) 1169–1180.
[44] C. Hollard, C. Keyser, T. Delabarde, A. Gonzalez, C.V. Lamego, V. Zvénigorosky, et al., Case report: on the use of the HID-Ion AmpliSeq™ Ancestry Panel in a real forensic case, Int. J. Legal Med. (2016) 1–8.
[45] M. Eduardoff, T.E. Gross, C. Santos, D. de la Puente, C. Strobl, et al., Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM, Forensic Sci. Int. Genet. 23 (2016) 178–189.
[46] M. Eduardoff, C. Santos, M. de la Puente, M. Gross, C. Strobl, et al., Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM, Forensic Sci. Int. Genet. 17 (2015) 110–121.
[47] S. Elena, A. Alessandro, C. Ignazio, W. Sharon, R. Luigi, B. Andrea, Revealing the challenges of low template DNA analysis with the prototype Ion AmpliSeq Identity panel v2.3 on the PGM Sequencer, Forensic Sci. Int. Genet. 22 (2016) 25–36.
[48] A. Buchard, M.L. Kampmann, L. Poulsen, C. Børsting, N. Morling, ISO 17025 validation of a next-generation sequencing assay for relationship testing, Electrophoresis 37 (2016) 2822–2831.
[49] S.B. Seo, J.L. King, D.H. Warshauer, C.P. Davis, J. Ge, B. Budowle, Single nucleotide polymorphism typing with massively parallel sequencing for human identification, Int. J. Legal Med. 127 (2013) 1079–1086.
[50] X. Zeng, D.H. Warshauer, J.L. King, J.D. Churchill, R. Chakraborty, B. Budowle, Empirical testing of a 23-AIMs panel of SNPs for ancestry evaluations in four major US populations, Int. J. Legal Med. 130 (2016) 891–896.
[51] H. Thorvaldsdóttir, J.T. Robinson, J.P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, Brief. Bioinform. 14 (2012) 178–192.
[52] C. Phillips, W. Parson, J. Amigo, J. King, M. Coble, C. Steffen, et al., D5S2500 is an ambiguously characterized STR: identification and description of forensic microsatellites in the genomics age, Forensic Sci. Int. Genet. 23 (2016) 19–24.
[53] K.N. Ballantyne, A. Ralf, R. Aboukhalid, N.M. Achakzai, M.J. Anjos, Q. Ayub, et al., Toward male individualization with rapidly mutating Y-chromosomal short tandem repeats, Hum. Mutat. 35 (2014) 1021–1032.
[54] C. Van Neste, Y. Gansemans, D. De Coninck, D. Van Hoofstat, W. Van Criekinge, D. Deforce, et al., Forensic massively parallel sequencing data analysis tool: implementation of MyFLq as a standalone web-and IlluminaBaseSpace®-application, Forensic Sci. Int. Genet. 15 (2014) 2–7.
[55] S.T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, et al., dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (2001) 308–311.
[56] G.P. Consortium, An integrated map of genetic variation from 1,092 human genomes, Nature 491 (2012) 56–65.
[57] Illumina, MiSeq FGx™ Instrument Reference Guide (May 2015) https://support.illumina.com/downloads/miseq-fgx-instrument-reference-guide-15050524.html.
[58] Illumina, ForenSeq™ DNA Signature Prep Guide (September 2015) https://support.illumina.com/downloads/forenseq-dna-signature-prep-guide-15049528.html.
[59] A. Gonçalves da Silva, W. Barendse, J.W. Kijas, W.C. Barris, S. McWilliam, R.J. Bunch, et al., SNP discovery in nonmodel organisms: strand bias and base-substitution errors reduce conversion rates, Mol. Ecol. Resour. 15 (2015) 723–736.
[60] K.J. van der Gaag, R.H. de Leeuw, J. Hoogenboom, J. Patel, D.R. Storts, J.F. Laros, et al., 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.
[61] J.D. Churchill, B. Budowle, More and more markers use of the precision ID GlobalFiler mixture ID panel to analyze challenged and mixed samples, 27th International Symposium on Human Identification, Minneapolis, MN, 2016.