Forensic Science International: Genetics 14 (2015) 182–186

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Short Communication

STRait Razor v2.0: The improved STR Allele Identification Tool – Razor

David H. Warshauer [a], Jonathan L. King [a], Bruce Budowle [a,b,*]

[a] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Boulevard, Fort Worth, TX 76107, USA

[b] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 9 July 2014
Received in revised form 20 September 2014
Accepted 14 October 2014

Keywords:
STRait Razor
Massively parallel sequencing
STR
Bioinformatics
Software

ABSTRACT

STRait Razor (the STR Allele Identification Tool – Razor) was developed as a bioinformatic software tool to detect short tandem repeat (STR) alleles in massively parallel sequencing (MPS) raw data. The method of detection used by STRait Razor allows it to make reliable allele calls for all STR types in a manner that is similar to that of capillary electrophoresis. STRait Razor v2.0 incorporates several new features and improvements upon the original software, such as a larger default locus configuration file that increases the number of detectable loci (now including X-chromosome STRs and Amelogenin), an enhanced custom locus list generator, a novel output sorting method that highlights unique sequences for intra-repeat variation detection, and a genotyping tool that emulates traditional electropherogram data. Users also now have the option to choose whether the program detects autosomal, X-chromosome, Y-chromosome, or all STRs. Concordance testing was performed, and allele calls produced by STRait Razor v2.0 were completely consistent with those made by the original software.

1. Introduction

STRait Razor (the STR Allele Identification Tool – Razor) is a Perl-based bioinformatic software package for the Linux operating system that was created to provide an effective and reliable means of detecting short tandem repeat (STR) markers in massively parallel sequencing (MPS) data. While the current assortment of MPS platforms and methodologies are capable of sequencing STRs [1–4], there are very few options available for variant analysis of the raw data [1,5]. STRait Razor accurately analyzes STR sequence data from FASTQ files while avoiding the drawbacks associated with software tools that are unable to adequately detect complex repeat motifs or intra-repeat variation [1]. Although it was intended for forensic DNA typing, the software can be used for a wide variety of genetic research purposes. STRait Razor employs a simple yet effective method to analyze STRs. First, alleles are detected via matching of the leading and trailing flanking region(s) surrounding the repeat(s). This matching process allows for user-defined stringency parameters and results in the cleaving of all extraneous sequence data from around an STR, leaving only the repeat units. A final filtering step removes non-STR sequencing reads that may have been captured due to flanking sequence homology. Alleles are called by comparing the length of the repeat unit region with known allele lengths, and the total counts of each allele at each locus are ordered and stored in an output file for further analysis and comparison with other samples. STR analysis is performed in parallel, using a separate processor core for the detection of each locus. This straightforward process results in allele calls that are analogous to calls produced by traditional capillary electrophoresis (CE) methods, since the allele coverage counts are analogous to RFU values. In addition to size of the allele, registered as number of repeats, allelic sequence data provide an added benefit of highlighting nucleotide variation within the repeat units. Since its initial release, STRait Razor has been employed by a number of laboratories with positive results. In-house needs and resulting feedback were considered strongly to develop a number of improvements to the original software. These new features include an expanded default set of detectable STR loci (autosomal, X, and Y markers), an enhanced custom locus list configuration tool, a novel output sorting method that highlights unique sequences for each allele, and a genotyping tool that emulates traditional electropherogram data. With these improvements, STRait Razor v2.0 offers users a much wider, more flexible range of analysis options and greater ease of use.

[*] Corresponding author at: Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Boulevard, Fort Worth, TX 76107, USA. Tel.: +1 8177352979.
E-mail address: Bruce.Budowle@unthsc.edu (B. Budowle).

http://dx.doi.org/10.1016/j.fsigen.2014.10.011
1872-4973/© 2014 Elsevier Ireland Ltd. All rights reserved.

2. Materials and methods

2.1. New features

STRait Razor v2.0 includes an expanded locus configuration file which it can use to detect a wider range of forensically-relevant STR markers. Previously, the default locus definition file included 44 STR loci (22 autosomal STRs and 22 Y-chromosome STRs). An additional 42 markers have been added, for a total of 86 markers, which include: 9 new autosomal STRs (D14S1434, D17S1301, D1S1627, D2S1776, D4S2408, D5S2500, D6S1017, D6S474, and SE33), 26 new X-chromosome STRs (DXS10011, DXS10074, DXS101, DXS10101, DXS10134, DXS10135, DXS6789, DXS6795, DXS6800, DXS6801, DXS6807, DXS6809, DXS6854, DXS7132, DXS7133, DXS7423, DXS7424, DXS8377, DXS8378, DXS981, DXS9895, DXS9902, GATA165B12, GATA172D05, GATA31E08, and HPRTB), six new Y-chromosome STRs (DYS449, DYS460, DYS505, DYS518, DYS522, and DYS612), and Amelogenin. The allelic information contained in the locus configuration file was compiled using data from a variety of online databases, such as STRbase [6], ChrX-STR.org 2.0 [7], NCBI [8], and the Sorenson Molecular Genealogy Foundation database [9]. Users have the option of choosing the marker type to analyze with STRait Razor v2.0 by using the "-*typeselection*" argument (AUTO, X, Y, or ALL). This feature allows the analysis to be tailored to the specific goals of the testing and, depending on which option is selected, can reduce the time required for analysis. As with the initial version of STRait Razor, custom locus configuration files can be created for the program using the included Microsoft Excel workbook. Therefore, analysis can be performed on STR loci that are not included in the default set.

STRait Razor v2.0 includes an enhanced locus configuration workbook for the creation of custom allelic definition files. The workbook allows for the generation of a locus configuration file containing up to 10 STR loci, although the workbook can be modified to include more. Alternatively, multiple configuration files may be generated and concatenated to produce larger, more comprehensive files. The locus configuration workbook is designed to convert locus information entered by the user for any STR locus type (autosomal, X-chromosome, or Y-chromosome) to the proper format required by STRait Razor for analysis. The workbook has been redesigned for ease of use, with features such as automatic generation of reverse complement leading and trailing flanking data based on user input, and auto-conversion of lowercase entries to the proper uppercase format.

STRait Razor is designed to yield STR allele calls for each locus, as well as sequence data for all alleles so that intra-repeat variation can be detected. In its initial release, STRait Razor output sequence data for each allele in a locus-specific file, sorted by repeat region length. While useful, these data were unwieldy to interpret because sequence data from each individual read were appended to a sequence file when captured, resulting in a large amount of redundant sequence information. STRait Razor v2.0 simplifies the sequence output so that only unique sequences for each allele are displayed, and the results are sorted based on the total read count for each sequence (Fig. 1). This output results in a clear and concise sequence file for each locus that can be interpreted quickly for intra-repeat nucleotide variation.

The first version of STRait Razor enabled users to generate a substantial amount of information on both the alleles and underlying sequence variants. While this updated version of STRait Razor is considerably more efficient, the tremendous amount of data generated requires a toolset for subsequent analysis. To further facilitate data analysis, a set of Excel-based workbooks have been developed.

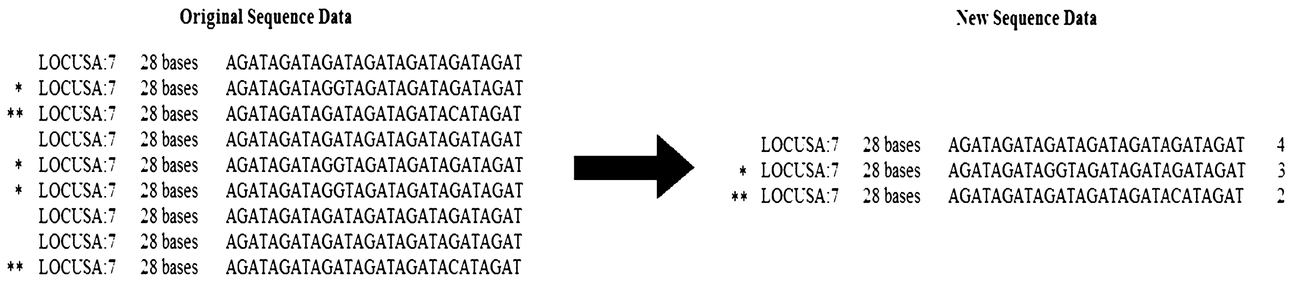
The first workbook, RazorGenotyper, converts the output files "RawSTRcallsR1" and "RawSTRcallsR2" into final genotypes. Users can set thresholds for coverage and sister allele balance to ensure that accurate genotypes are generated. Data from multiple samples can be exported and compiled via embedded macros. These exported data then can be further visualized using the second workbook.

STRait Razor Histogram Generator separates the output data of RazorGenotyper into an "allele table" of all loci. These data then are displayed as histograms showing all read variants observed (e.g., alleles, stutter, and PCR artifacts). These charts are parsed into "Autosomal," "Y," and "X" STR tabs. The autosomal tab also contains Amelogenin and is divided into "Core" loci (i.e., loci contained in either PowerPlex Fusion (Promega Corp., Madison, WI, USA) or GlobalFiler (ThermoFisher, San Francisco, CA, USA)) and "Additional Loci" (i.e., autosomal loci not found in either kit). Macros included, but not active, allow a user to uniformly change the axes of all charts to visualize locus-to-locus balance.

The final workbook included in the toolset, STRait Razor_SNP ID Tool, converts the *LOCUS.SEQUENCES* files into a table showing the top 20 sequence variants at each locus. First, the user must transfer the *LOCUS.SEQUENCES* file from each locus of interest into a single folder. Next, the user can run the included 'SeqCompile.pl' script to combine all loci into a single file. The data from this file then are pasted into the STRait Razor_SNP ID Tool. After which, data are displayed by locus showing the most relevant sequence variants. These data for all loci can be exported and compiled via embedded macros for ease of use.

2.2. Concordance testing

The accuracy and reliability of STRait Razor have been reported [1]. Therefore, concordance testing was performed to verify that STRait Razor v2.0, with its updates, provides the same allele calls as it did in its first release. The same seven sequence datasets used in the initial testing phase were re-analyzed using STRait Razor v2.0. These datasets are summarized in Table 1.



```
   LOCUSA:7  28 bases  AGATAGATAGATAGATAGATAGATAGAT
*  LOCUSA:7  28 bases  AGATAGATAGGTAGATAGATAGATAGAT
** LOCUSA:7  28 bases  AGATAGATAGATAGATAGATACATAGAT
   LOCUSA:7  28 bases  AGATAGATAGATAGATAGATAGATAGAT
*  LOCUSA:7  28 bases  AGATAGATAGGTAGATAGATAGATAGAT
*  LOCUSA:7  28 bases  AGATAGATAGGTAGATAGATAGATAGAT
   LOCUSA:7  28 bases  AGATAGATAGATAGATAGATAGATAGAT
   LOCUSA:7  28 bases  AGATAGATAGATAGATAGATAGATAGAT
** LOCUSA:7  28 bases  AGATAGATAGATAGATAGATACATAGAT
```

```
   LOCUSA:7  28 bases  AGATAGATAGATAGATAGATAGATAGAT  4
*  LOCUSA:7  28 bases  AGATAGATAGGTAGATAGATAGATAGAT  3
** LOCUSA:7  28 bases  AGATAGATAGATAGATAGATACATAGAT  2
```

Fig. 1. Sequence data sorting. The original sequence output from STRait Razor, using an example locus named "LOCUSA" (AGAT repeat unit), is shown on the left. Sequences were appended as they were captured. Sequences denoted with (*) contain an A/G SNP, while sequences denoted with (**) contain a G/C SNP. Sequence output from STRait Razor v2.0 is shown on the right. With this update, unique sequences were identified, counted, compiled, and finally sorted based on the total read count.

Table 1
Datasets used for concordance testing. The seven datasets used consisted of sequence data from 5 samples, obtained through library preparation using the TruSeq™ (Illumina, Inc., San Diego, CA) and HaloPlex™ (Agilent Technologies, Inc., Santa Clara, CA) kits and sequencing using the MiSeq™ (Illumina, Inc.) and GAIIx™ (Illumina, Inc.) platforms.

Dataset	Sample	Library preparation method	Sequencing platform
1	A	Illumina® TruSeq™ Custom Enrichment	Illumina® GAIIx™
2	A	Agilent Technologies® HaloPlex™	Illumina® GAIIx™
3	A	Agilent Technologies® HaloPlex™	Illumina® MiSeq™
4	B	Agilent Technologies® HaloPlex™	Illumina® GAIIx™
5	C	Illumina® TruSeq™ Custom Enrichment	Illumina® MiSeq™
6	D	Illumina® TruSeq™ Custom Enrichment	Illumina® MiSeq™
7	E	Illumina® TruSeq™ Custom Enrichment	Illumina® MiSeq™

The sample preparation methods and sequencing techniques used in obtaining these datasets were described previously [1]. Allele calls made by the updated software were compared to those made by the initial version of the software, and new allele calls resulting from use of the larger default locus configuration file were noted. The time required for analysis with the wider range of detectable loci also was determined. Additionally, genotyping and histogram generation were performed to validate the effectiveness of these tools.

In addition, the allele detection capability of STRait Razor v2.0 was evaluated with MiSeq™-generated FASTQ files from a set of 12 samples (6 male and 6 female) that had previously undergone library preparation using an experimental in-house Nextera™ Rapid Capture (Illumina, Inc.) assay designed to specifically target the loci included in STRait Razor v2.0's locus configuration file. These datasets were analyzed using both STRait Razor v2.0 and the original version of the software, and the genotype results were compared.

3. Results

STRait Razor v2.0 yielded identical allele calls from all 7 original sequence datasets with regard to the loci previously analyzed. The read counts generated for each allele detected by STRait Razor v2.0 were 100% concordant with those indicated by the original version of the software. The testing process also demonstrated the wider range of locus detection afforded by the new expanded locus configuration file. Across these 7 datasets, STRait Razor v2.0 detected a range of 25–41 new alleles at between 19 and 35 new autosomal, X-chromosome, and Y-chromosome loci, as well as Amelogenin (Table 2).

Table 2
Additional alleles and loci detected in the original 7 datasets. New loci identified by STRait Razor v2.0 have been categorized by their respective types. (+) Indicates that amelogenin was detected in a particular dataset, while (–) indicates that amelogenin was not detected.

Dataset	Additional alleles	Additional loci			
		Autosomal	X	Y	Amelogenin
1	26	7	9	3	+
2	25	6	10	3	–
3	34	7	17	4	–
4	27	7	10	3	–
5	35	8	16	4	+
6	40	8	21	5	+
7	41	8	21	4	+

The results for the additional 12 datasets showed a similar pattern of enhanced allele detection with STRait Razor v2.0. The original software detected between 60 and 64 alleles at 44 loci in male samples, while between 37 and 41 alleles were identified at 22 loci in female samples, given the absence of Y-chromosome STRs in these samples (Table 3). STRait Razor v2.0, however, increased the number of alleles detected to a range of 105–111 at between 82 and 84 loci in male samples. Female samples analyzed with the updated software yielded 92–97 alleles at 56 loci. Amelogenin was detected in all 12 additional samples. Comparison of the allele calls made by STRait Razor v2.0 and the original version of STRait Razor for these samples showed that they were 100% concordant, as was the case for the initial group of seven datasets.

As noted previously, allele detection is dependent on a number of factors, including sequence read length and library preparation chemistry [1]. Lack of detection of alleles at loci included in the new locus configuration file was due to these same issues and is not indicative of improper software function. Time required for analysis is directly related to the number of loci that are included in the locus configuration file. Thus, the new larger configuration file does increase STRait Razor v2.0's analysis time. In this study, the time required for analysis of all 86 markers for dual 400MB MiSeq™-generated FASTQ files on a 16-core server was approximately 29 min. When only Y-chromosome or X-chromosome STRs were analyzed, the analysis time dropped to approximately 9 min for each. The time required for analysis of only autosomal STRs was approximately 11 min. Shorter custom configuration files can be used to reduce analysis time, and the time required will vary depending on the specifics of each application and the computing platform utilized.

Genotypes were displayed, along with allele read counts, in a manner that was clear and easy to interpret. Histograms were generated for the alleles, stutter, and noise detected from these datasets that approximate electropherogram displays (Fig. 2). Given the similarity between these histograms and traditional electropherograms, this option provides a simple way to visually inspect allele calls at each detected locus. In this manner, reads resulting from stutter or noise can be interpreted visually.

Finally, the sequence data for each detected allele that were output by STRait Razor v2.0 were investigated. The unique sorting process employed by the updated software allows for quick detection of intra-repeat nucleotide variation. For example, an examination of the sequence data output file for dataset 4 revealed

Table 3
Allele and locus detection comparison for the additional 12 datasets. The alleles and loci identified by the original version of STRait Razor are listed first, followed by those detected by STRait Razor v2.0. The loci are grouped based on their respective chromosomal type.

Sample	Alleles detected	Loci detected			
		Autosomal	X	Y	Amelogenin
1	41/96	22/30	0/25	*/*	–/+
2	64/111	22/30	0/25	22/27	–/+
3	37/95	22/30	0/25	*/*	–/+
4	61/108	22/30	0/25	22/27	–/+
5	61/107	22/31	0/25	22/27	–/+
6	37/93	22/30	0/25	*/*	–/+
7	63/109	22/31	0/25	22/27	–/+
8	39/97	22/30	0/25	*/*	–/+
9	64/110	22/30	0/24	22/27	–/+
10	37/97	22/30	0/25	*/*	–/+
11	37/92	22/30	0/25	*/*	–/+
12	60/105	22/30	0/24	22/27	–/+

(*) Denotes that the sample was female, and thus did not contain any Y-chromosome STRs to detect.

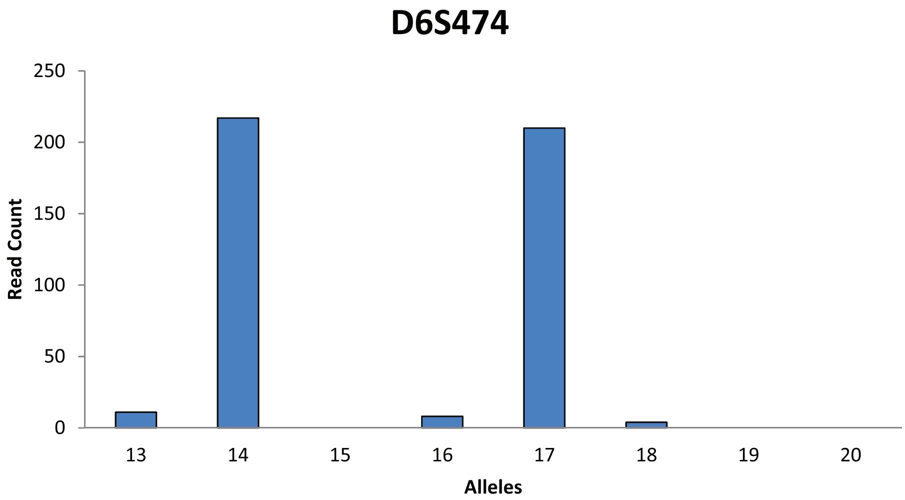


Fig. 2. Histogram generated for locus D6S474 in dataset 3 (Read 1). Histograms generated by the supplemental tool included with STRait Razor v2.0 resemble traditional electropherograms. Read counts displayed on the *Y*-axis are analogous to RFU values for peak heights. Here, the true alleles are "14" and "17." Stutter peaks "13" and "16" can be seen to the left of the major allele peaks. This profile also shows the plus stutter "18" peak to the right of the major "17" allele peak.

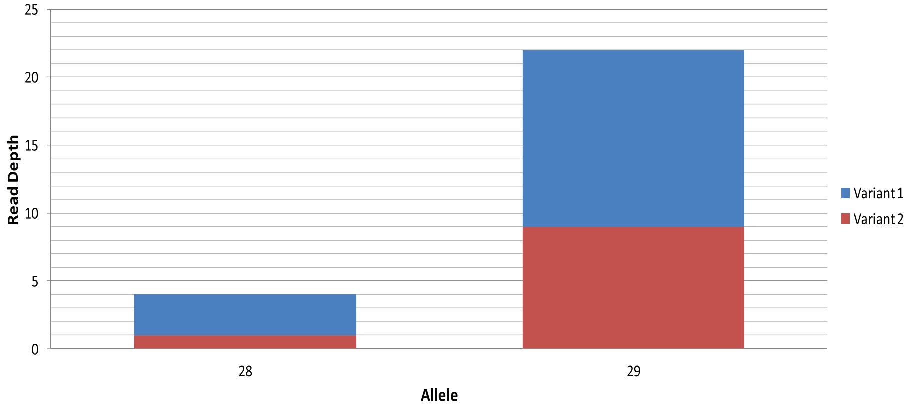


Fig. 3. Example of intra-repeat variation detected at locus D21S11 in dataset 4. This individual is homozygous for the "29" allele, Two distinct variants of this allele were detected. These variants also are seen in the stutter allele.

that the homozygote "29" allele at locus D21S11 consisted of the variants "$(TCTA)_4(TCTG)_6(TCTA)_3TA(TCTA)_3TCA(TCTA)_2TCCATA(TCTA)_{11}$" and "$(TCTA)_5(TCTG)_6(TCTA)_3TA(TCTA)_3TCA(TCTA)_2$-$TCCATA(TCTA)_{10}$", at a ratio approaching 1:1 (Fig. 3). The latter "29" variant is not listed in STRbase. It should be noted that both variants are found in the "28" stutter allele, as well. Additional sequence variation such as this can provide increased discrimination power, especially in the case of DNA mixtures, where it can facilitate deconvolution of both unique and shared alleles. However, intra-repeat variants must be characterized and allele frequencies generated before they can be used to their full potential. Mixture deconvolution using STRait Razor v2.0 will be evaluated in future studies.

4. Conclusions

STRait Razor v2.0 includes a number of improvements over the original version of this STR detection software. This update retains the reliable and accurate allele-calling capability of the initial release, with the added benefit of a much larger range of detectable loci. The ability to detect autosomal, Y-chromosome, and now X-chromosome STRs augments the usefulness of the software and provides for a wider range of potential applications. The enhanced custom locus configuration file generator and supplemental tools, such as the genotyper and histogram generator, have made STRait Razor v2.0 much more user-friendly and facilitate ease and speed of analysis. Genotypes are determined quickly and can be reviewed by the analyst readily. Aspects such as stutter can be investigated in a manner that resembles that of current electropherogram interpretation. Finally, intra-repeat nucleotide variation is presented in a much easier way to analyze. The sorting of unique sequences for each allele by the total read count allows analysts to rapidly distinguish true allelic variants from those that may be due to simple sequencing errors or background noise.

STRait Razor v2.0 is free to use and available online (http://web.unthsc.edu/info/200210/molecular_and_medical_genetics/887/research_and_development_laboratory/5). Updates and new content will be added to the website as they are developed and tested. As always, modification and improvement to the program by other users and developers, as well as constructive feedback, are encouraged.

Conflict of interest

None.

Acknowledgements

This work was supported in part by award no. 2012-DN-BXK033, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice.

References

[1] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. LaRue, J.L. King, B. Budowle, 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] M. Gymrek, D. Golan, S. Rosset, Y. Erlich, lobSTR: a short tandem repeat profiler for personal genomes, Genome Res. 22 (2012) 1154–1162.

[3] 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, Biotech. Rapid Dispatches (2012) 1–6.

[4] 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.

[5] 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.

[6] STRBase: http://www.cstl.nist.gov/strbase/str_fact.htm.

[7] ChrX-STR.org 2.0: http://www.chrx-str.org/.

[8] NCBI: http://www.ncbi.nlm.nih.gov/.

[9] Sorenson Molecular Genealogy Foundation Y Marker Details: http://www.smgf.org/ychromosome/marker_details.jspx.