Forensic Science International: Genetics 6 (2012) 810–818

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

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

Forensic STR analysis using massive parallel sequencing

Christophe Van Neste [1], Filip Van Nieuwerburgh [1], David Van Hoofstat, Dieter Deforce *

Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium

ARTICLE INFO

Article history:
Received 21 October 2011
Received in revised form 24 January 2012
Accepted 14 March 2012

Keywords:
Forensic
STR
Next generation sequencing
Second generation sequencing
Roche GS FLX
Pyrosequencing

ABSTRACT

We explore the applicability of second generation sequencing (SGS) to sequence multiplexed forensic STR amplicons, both in a single contributor sample as in multiple-person mixtures with different ratios. We compare the results of a commercial STR profiling kit (Applied Biosystems AmpFlSTR® Profiler Plus®), analyzed both with capillary electrophoresis and with Roche GS FLX sequencing. An easy to use open-source software pipeline is provided, chaining together the different steps needed to start the analysis from a GS FLX FASTA file, resulting in a FASTA file containing the called and quantified alleles present in the data. Sequencing of multiplexed STR amplicons using Roche GS FLX titanium technology is technically feasible but the technology is not ideal for this purpose. The fraction of full length reads is small and the homopolymer sequencing error rate is high. The pipeline compresses the homopolymers to a single base to avoid false results caused by these homopolymers. The qualitative and quantitative results from the SGS STR analysis pipeline are comparable to the electrophoresis method. Additionally, the SGS method provides extra information and is able to call allele subtypes based on STR sequences in a database. In mixed samples, all alleles were reported from individuals that contributed at least 10% to the mixture.

1. Introduction

1.1. Background

Short tandem repeat (STR) profiling using PCR and capillary electrophoresis is currently the most commonly used method to obtain a forensic DNA profile. Advancing knowledge in human single nucleotide polymorphism (SNP) markers, generated by projects such as the International HapMap Project [1], raised the question within the forensic DNA typing community if SNP markers have the potential to replace the currently used STR loci.

Second generation sequencing (SGS) technologies present an entirely new paradigm for DNA sequence data generation. SGS technology offers the possibility to sequence up to millions of individual DNA strands in DNA mixtures such as fragmented genomic DNA and multiplexed amplicons. Using SGS to sequence STR amplicons can combine both abovementioned approaches as STR amplicons can contain SNPs, making them even more discriminative as a genetic marker for individuals. SGS can generate individual sequences of the alleles present in an STR amplicon mixture. This way, alleles with the same length can be distinguished based on SNPs or different repeat structures.

SGS of forensic STR amplicons has other potential advantages. When separating alleles using capillary electrophoresis, the number of loci with overlapping size ranges is limited to the number of different dyes that can be used to discriminate these loci in an electropherogram. In SGS data, sequences originating from different loci can be identified based on the primer sequences. In Holland et al. [2] this was demonstrated for different markers of forensic interest. It is then possible to design an STR multiplex with overlapping amplicon sizes for all loci. This advantage is extremely useful in the design of a reduced size STR amplicon multiplex, removing the current limitation that not all amplicons can have the smallest possible size as size is used to discriminate between loci in the electropherogram.

Another difference and possible advantage of SGS compared to capillary electrophoresis, is the digital nature of the signal generated by SGS. Capillary electrophoresis produces an analog signal with peaks consisting out of a height and surface which are influenced by baseline noise due to spectral bleeding, cross-talk between capillaries and fluctuations in instrument parameters such as laser output, voltages and temperature. Theoretically, a SGS STR profile can be considered digital: the quantity of an allele is determined by the number of reads present for that allele. The quantity of an allele is described by a discrete value and can be visualized as a peak in a histogram. Both an electropherogram and a SGS digital profile can contain errors and not necessarily reflect

* Corresponding author at: Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Harelbekestraat 72, 9000 Ghent, Belgium. Tel.: +32 0 92648052; fax: +32 0 92206688.
E-mail addresses: christophe.vanneste@ugent.be (C. Van Neste), filip.vannieuwerburgh@ugent.be (F. Van Nieuwerburgh), david.vanhoofstat@ugent.be (D. Van Hoofstat), dieter.deforce@ugent.be (D. Deforce).

[1] These authors contributed equally to this work.

1872-4973/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2012.03.004

the true STR profile. An electropherogram can contain dye blobs, peaks due to spectral bleeding, baseline spikes, etc. In an SGS profile, sequencing errors can convolute the profile. In both methodologies, aberrant peaks due to PCR errors such as stutter PCR products can be present. SGS however offers more possibilities to analyze and filter aberrant peaks because the sequences of a suspected peak can be analyzed. The source of an electrophoresis signal can unfortunately not be further investigated.

Of the commercially available SGS technologies such as GS FLX from Roche, SOLiD from Applied Biosystems, and HiSeq from Illumina, the Roche GS FLX system is best suited for amplicon sequencing of PCR products from the current forensic kits as it is the only technology that can generate full length reads of amplicons of 400–500 bp in length. A disadvantage of the Roche GS FLX technology is that it generates a substantial amount of sequencing errors when homopolymers are sequenced, resulting in miscalled homopolymer sizes [3].

1.2. Experimental setup

We explored the applicability of SGS to sequence multiplexed STR amplicons, both in a single contributor sample as in multiple-person mixtures with different ratios. We compared the results of a commercial STR profiling kit (Applied Biosystems AmpFlSTR® Profiler Plus®), analyzed both with Roche GS FLX sequencing and with capillary electrophoresis. Fordyce et al. [4] recently published the results of a similar experimental set-up comparing GS FLX sequencing of 5 STR loci to capillary electrophoresis. They however only studied single contributor samples.

The PCR conditions used are those recommended by the ProfilerPlus Kit. These conditions are optimized to achieve as much sensitivity and robustness as possible in STR forensic casework samples with a limited amount of available input DNA. These conditions are certainly not ideal to generate PCR products that will be used for sequencing. The used polymerase has no high fidelity and the high number of PCR cycles adds to the creation of PCR artifacts. Because it is the purpose of our study to explore the possibility to analyze forensic casework samples using the SGS technology, the recommended parameters of the Profiler Plus kit were not changed and the number of PCR cycles was not reduced.

The fact that the amplicons generated with the Profiler Plus kit are labeled with different fluorescent dyes, might influence the efficiency of GS FLX sequencing. There is no published data available on the effect of fluorescent labels attached to the template amplicons in GS FLX amplicon sequencing. We reasoned that this should be a negligible issue because during sequencing, only one fluorescent molecule will end up per well in the picotiter plate. Furthermore Roche GS FLX technology does not make use of fluorescence to detect signals. The scope of our study is to explore the applicability of GS FLX sequencing of forensic STR reads, to compare the results with a commercial STR profiling kit showing the advantages and disadvantages of the SGS approach and to present an easy to use open-source data analysis method which handles the several pitfalls that are inherent to GS FLX sequencing of forensic STRs.

1.3. Data analysis

Roche GS FLX XLR70 Titanium amplicon sequencing generates up to 700,000 reads per run with a mean per base error rate of 1.07%. The error rate is not randomly distributed and can rise to more than 50% [3]. When considering the average length of the STR alleles, this means that on average there is at least one error per read. Current forensic STR allele calling is based on size calling of the peaks in an electropherogram, making it possible to discriminate between alleles that differ only one base pair in length. Because of the many homopolymers present in most STR amplicons, it is prone to error to discriminate between alleles based on a small difference in length using Roche GS FLX technology. When the assumption can be made that the data are originating from a single contributor sample, this problem can be solved by retaining only the 1 or 2 most prevalent sequences for each homozygous or heterozygous locus, respectively. In practice, this approach may be of limited use, since many forensic casework samples derive from multiple donors. This approach could however still be used for databanking single source samples. We show an analysis method that circumvents this problem by compressing all homopolymers to a single base, which can be used to analyze mixed samples.

Even after compressing the homopolymers, it is necessary to perform additional error correction. Reads originating from the same amplicon need to be clustered together, making it possible to find the consensus sequence. The challenge for a clustering algorithm is to cluster all reads that differs only by errors while separating reads that differ by a SNP or an indel (insertion/deletion). We performed the clustering step with the PCR noise removal tools from AmpliconNoise [5]. These tools use a Needleman–Wunsch and an expectation maximization algorithm to produce clusters with the highest likelihood of clustering the reads that differ only by errors and not by SNPs or indels. This software has been used previously to adequately perform error correction on GS FLX data [5]. The consensus sequence of such a cluster should be the sequence of the original amplicon, if all errors are successfully eliminated.

2. Materials and methods

2.1. Samples

One nanogram of DNA from 5 non-related individuals was used as PCR template to amplify D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, D21S11, FGA and vWA using the AmpFlSTR® Profiler Plus® kit (Applied Biosystems), following standard instructions in the manual, using 34 PCR cycles. The resulting PCR product of the 5 reference samples was quantified using the Quant-iT™ PicoGreen® dsDNA Kit (Invitrogen) and mixed to create three experiment samples. Table 1 shows the composition of the three experiment samples.

2.2. Electrophoresis and SGS

The PCR products of the 5 reference samples and the 3 experiment samples were analyzed by capillary electrophoresis using an Applied Biosystems 3100 Genetic Analyzer and the Applied Biosystems Genemapper software. Only the 3 experiment samples were sequenced, using standard unmodified Roche FLX protocols. Two hundred nanograms of each experiment sample were ligated with adaptors following the GS FLX Titanium General Library Preparation Method Manual, version April 2009. In this ligation step, only one fifth of the method recommended amount of adaptors was used to compensate for the lower amount of input DNA. All subsequent steps of the method were followed to create a

Table 1
Composition of the 3 experiment samples.

Individual	Single contributor (%)	Mixture 1 (%)	Mixture 2 (%)
1	100	–	0.1
2	–	40	0.5
3	–	30	1
4	–	20	5
5	–	10	93.4

single stranded DNA library. To determine the amount of library to use in emPCr amplification, an emulsion titration assay was performed. After completion of this method, the libraries were used as template in an emulsion-based clonal amplification according to the Roche GS FLX titanium series emPCR Method Manual – Lib L, version October 2009. The 3 libraries were sequenced on a picotiter plate according to the Roche GS FLX titanium Sequencing Method Manual, version October 2009. Using a rubber gasket, the picotiter plate was divided into 8 physically separated sections. Each library was sequenced in 1 section. The other 5 sections of the picotiter plate were used for several other, unrelated sequencing experiments.

2.3. Theoretical profile

A theoretical profile of the experiment samples was made based on the electropherograms of the 5 individual references samples. The theoretical profile splits one peak into stacked components when two peaks coincide because an individual is homozygous, when two or more contributing individuals have an allele in common, or when a stutter of a bigger allele coincides with the peak of a smaller allele.

The stutters for the theoretical profile are calculated based on the highest percent observed stutter reported in the Profiler Plus manual: The highest percent stutter observed for any D5S818, D13S317, or D7S820 allele was less than 8%, for any D8S1179 allele less than 9%, for any D3S1358, vWA, FGA, or D21S11 allele less than 10%, and for any D18S51 allele less than 13% [6].

In the theoretical profile, the sum of the percentages of all peaks for a locus is 100%. This makes the comparison to the SGS data easier as the total number of considered reads for a locus is 100%. When a locus, e.g. consists out of 1 allele with 10% stutter, this locus will be presented with peaks of 91% and 9%.

2.4. Data analysis

2.4.1. Custom pipeline

An open-source software pipeline was developed, chaining together the different steps needed to start the analysis from a raw GS FLX SFF file or GS FLX FASTA file, resulting in a FASTA file containing the called and quantified alleles present in the data. This pipeline is available at http://www.labfbt.ugent.be/STRbySGS. Fig. 1 shows the different steps that are sequentially performed by the pipeline. The pipeline makes use of the SeqNoiseM clustering tool from AmpliconNoise [5] and the Needle pairwise aligner from the EMBOSS package [7]. First, the homopolymers in all reads of the input FASTA file are compressed to a single base. A sequence stretch consisting of twice the same basepair is already considered a homopolymer. Next, the reads in this FASTA file are categorized for each of the STR loci and for each reading direction. We required the presence of the first 20 bases of the primer at the beginning of the read and last 10 bases of the primer at the end of the read. Additionally, we required a minimum length for the compressed read of 50 bp. This way, only full length reads are selected and small artifacts are eliminated. The used categorizing sequences for each locus are available in online supplementary data. Next, the categorized reads are clustered with the SeqNoiseM tool from the AmpliconNoise software, using standard parameters for cluster sensitivity ($s = 30.0$) and cut-off ($c = 0.01$)[5]. Next, clusters with the same consensus sequence but from an opposite reading direction are merged and the number of reads in the resulting cluster is counted. Next, clusters consisting out of one sequence and clusters smaller than a customizable percentage of the total number of reads for that locus are discarded. We chose a percentage of 1% reasoning that with this filter, it should theoretically be possible to detect a heterozygous allele of a contributor that contributes only 5% to the STR profile.

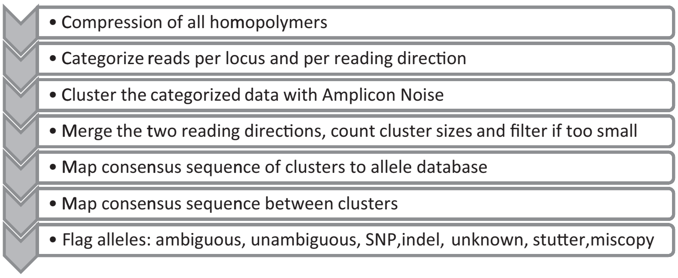
• Compression of all homopolymers
• Categorize reads per locus and per reading direction
• Cluster the categorized data with Amplicon Noise
• Merge the two reading directions, count cluster sizes and filter if too small
• Map consensus sequence of clusters to allele database
• Map consensus sequence between clusters
• Flag alleles: ambiguous, unambiguous, SNP,indel, unknown, stutter,miscopy

Fig. 1. Pipeline.

The consensus sequences of the remaining clusters are matched to an allele database using Needle from the EMBOSS package [7]. This tool uses a Needleman–Wunsch algorithm, which performs a global alignment between two sequences and returns the alignment score and characteristics. Our script queries a sequence to a database and returns the best possible global match based on the output from Needle. The allele database is composed of all regular and variant STR alleles of which the sequence could be retrieved from STRbase [8]. Because the consensus sequences of the clusters have compressed homopolymers, the homopolymers in the database were also compressed. This can create ambiguous sequences in the database: different alleles can have the same sequences after compressing the homopolymers. Based on the alignment results, the consensus sequences are flagged. The first flag can have 3 possible values: The 'N' or 'no match' flag means that no match with less than 2 gaps and less than 3 SNPs were found. The 'U' or 'unambiguous' and the 'A' or 'ambiguous' flag means a match to an ambiguous or unambiguous database sequence was found in the database with less than 2 gaps and less than 3 SNPs. A 'no match' sequence never gets a second flag. A sequence with a match can have an additional 'SNP' and an additional 'INDEL' flag when at least one SNP or one gap was allowed to make the match with the compressed STR database entry.

During the final step of the pipeline, the consensus sequence of the clusters are mapped against each other using Needle from the EMBOSS package. This allows flagging clusters as stutters or miscopy errors. When a cluster is identical to another cluster except for 4 contiguous gaps it is flagged as a stutter. When a cluster is identical to another cluster except for one gap or one SNP, both clusters are flagged as a possible sequencing error with the flag 'miscopy'. When one of the 2 'miscopy' clusters consist of reads in only one sequencing direction and the other cluster consists of reads in both reading directions, the reads of the first cluster are added to the second cluster and the remaining cluster is flagged as a binned miscopy cluster with the flag 'miscopy_bin'. Clusters with the 'stutter', 'miscopy' or 'miscopy_bin' flag need to be manually checked. Based on the percentages of the number of reads in the clusters, it is possible to asses if it is likely that these clusters are stutters, sequencing errors or true alleles. An optional script 'Isolate_Miscopy' is made available that can be used to show the difference in the original uncompressed sequences of a pair of 'miscopy' clusters. This allows a manual analysis of what may have caused the miscopy cluster.

For each reported allele, following information is reported in the resulting FASTA output: matching allele in database, is allele present in forward and/or reverse reading direction, total number of considered reads for that locus, percentage of the considered reads in the reported allele, number of gaps needed to allow a match to the database, matching allele based on the length of the cluster, abovementioned flags. The output from the pipeline for our three experimental samples is available in online supplementary data. The intermediate results of the pipeline are saved, making it, e.g. possible to look at the sequences in a specific cluster.

A genuine allele of a contributor that is not in the database, to which the alleles are matched, will either be flagged as a 'no match', 'SNP' or 'miscopy'. The proportion of the cluster can be used as an indication whether or not to consider it a genuine allele. In a single contributor sample it should be relatively simple to mark possible genuine alleles that are not yet in the database. Subsequent analysis using Sanger sequencing should be used to confirm the new allele, before it can be added to the database.

3. Results

3.1. General properties of the GS FLX dataset

Table 2 shows the total number of active wells, filtered wells, and past-filter reads in the Roche GS FLX dataset, as well as the number of reads which contain at least one primer sequence and the number of full length reads selected by the pipeline. Huse et al. [9] report that on average about 45% of the wells that contain detectable sequencing templates, produce usable reads. Comparing the total number of active wells and the total number of reads in Table 2, on average 48% of the wells produced reads. Table 3 zooms in on the number of reads which contain at least one primer sequence, showing how many reads are sequenced in the forward and in the reversed direction. These results show that longer loci are underrepresented in the dataset. Although one would expect a similar proportion of reads in the forward direction as in the reversed direction, this is not the case in our limited dataset.

3.2. Single contributor sample

Fig. 2 shows a comparison between the results obtained from the single contributor sample. In the top panel of the figure, the peak heights from the capillary electrophoresis profile are plotted. In the middle is the theoretical profile. In the theoretical profile from one individual, each locus should have either 2 peaks of 45% when the individual is heterozygous for that locus, or 1 peak of 90% when the individual is homozygous for that locus. Just before the main peaks, stutter peaks of around 5–10% should be present. The lower panel is the profile obtained from the GS FLX dataset using the pipeline output without manual corrections. The *X*-axis of the figure shows the assigned alleles. In the SGS profile, several alleles are marked with a number 1. For these alleles, the SGS method provides additional SNP, indel or uncommon allele subtype information compared to the electrophoresis method. Alleles in the SGS profile that are marked with a number 2 have an ambiguous match in the database, thus providing less information compared to the electrophoresis method. Alleles in the SGS profile that are marked with a number 3 did not match to the database with less than 2 gaps or 3 SNPs. Alleles that were flagged both as a 'no match' and as a stutter of a matched allele are presented in the figure on the stutter position. Table 4 compares the allele information between the electrophoresis and the SGS method and provides more information on the alleles marked with a number. The output file from the pipeline is available in online supplementary data.

Table 2
Number of wells and reads in the Roche GS FLX dataset.

	Mixture 1	Mixture 2	Single
Total active wells	208,069	187,237	206,779
Total filtered wells	204,924	184,494	203,456
Total number of reads	103,770	99,261	88,909
Reads with primer sequence	52,100	51,156	47,114
Full length reads	18,337	22,668	14,840

Table 3
Number of reads in forward (FW) and reversed (RV) direction: size-range of alleles according to STRbase.

	Size-range (bp)	Mixture 1		Mixture 2		Single contributor	
		FW	RV	FW	RV	FW	RV
D3S1358	97–145	6849	156	6810	154	7240	37
D5S818	130–178	667	9621	778	8666	339	8059
D7S820	253–293	133	175	147	193	85	134
D8S1179	123–175	175	11,401	547	10,472	419	11,692
D13S317	193–241	436	1357	268	2663	219	822
D18S51	264–351	1501	444	1660	274	1170	119
D21S11	186–256	6602	920	6467	867	4569	784
FGA	196–352	437	544	510	734	277	657
vWA	152–212	9766	916	8894	1052	9923	569

3.3. Mixed samples

The STR profiles and allele call results of the mixture samples are represented in Fig. 3 and Table 5 and in Fig. 4 and Table 6, respectively. Unlike Table 4 of the single contributor sample, Tables 5 and 6 for the mixture samples only contain information on the alleles that were marked with a number in the respective figures. In the theoretical profile, the relative contribution of alleles of identical length, but originating from more than one individual, is shown by stacked bars in the bar graph. In the SGS profile, bars are stacked when the pipeline result show different subtypes of alleles with the same length. The output files from the pipeline are available in online supplementary data.

4. Discussion

Roche GS FLX sequencing technology is more expensive and more laborious than the capillary electrophoresis method although multiplexing samples can reduce the per-sample cost. This limits the use of this technology to routinely analyze forensic STR casework samples. In addition, the high error rate in the produced sequences impedes data analysis. Of the commercially available SGS technologies, the Roche GS FLX system is currently the only technology that can generate full length reads from amplicons generated by the current forensic STR profiling kits. Because the Roche GS FLX system is based on pyrosequencing, it generates a substantial amount of sequences with miscalled homopolymer sizes.

To be able to make a comparison with the current gold standard capillary electrophoresis method, we decided to use a commercial multiplex for the current exercise. Current commercial STR profiling kits are designed and optimized for analysis with capillary electrophoresis and do not take advantage of the full potential of SGS. An SGS STR profiling kit should generate small amplicon sizes. The effect of reduced efficiency of PCR amplification of longer amplicons is more pronounced in SGS because the current SGS methods contain an additional PCR step. Because pyrosequencing of homopolymers results in a high error rate, a SGS STR profiling kit designed for a pyrosequencing based SGS system should generate amplicons with as few homopolymers as possible. STRs such as FGA and D18S51 can contain up to approximately 30 homopolymers. Sequencing errors in one or more of these stretches, randomly distributed over the reads, convolutes the Roche GS FLX data and easily results in mistakes during analysis of the data.

SGS poses limitations on STR amplicon design, but also opens new possibilities. In SGS data, sequences originating from different

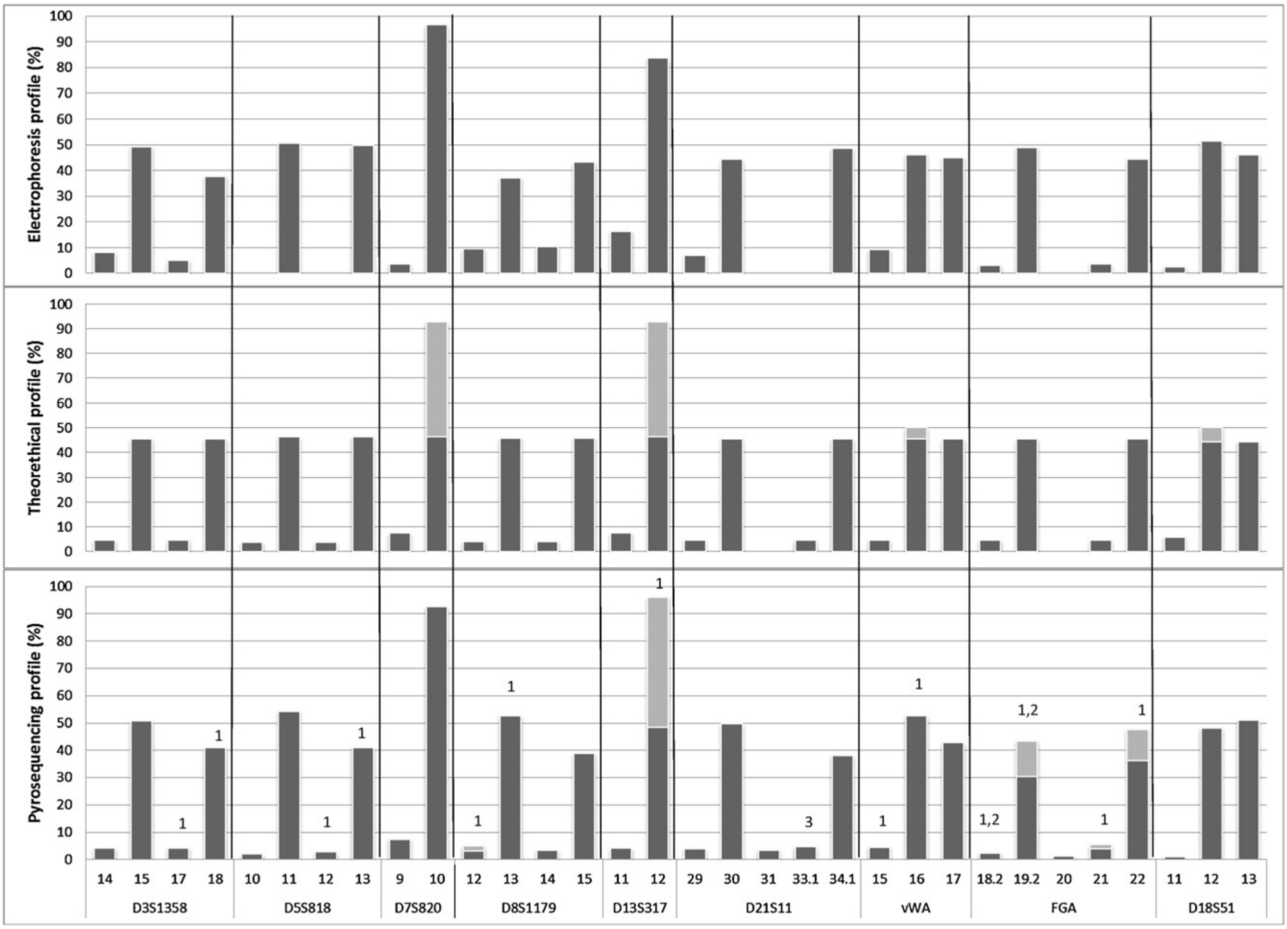

Fig. 2. STR profiles for the single contributor sample.

loci can be identified based on the primer sequence that is incorporated in the sequence, making it possible to design a STR multiplex with an unlimited number of overlapping amplicon sizes for all loci. This opens the door to design a mini-STR kit wherein more mini-STRs can be multiplexed.

Our results show a total number of reads of approximately 100,000 reads per sample. This is around the expected range of ~87,500 reads as reported by Roche [10] when sequencing an amplicon sample on 1/8th of a GS FLX titanium run. Approximately half of the reads start with a primer sequence. The other half of the reads consists of PCR artifacts, STR amplicon fragments and amplicons which start with a primer sequence that contains sequencing errors. Approximately half of the reads starting with a primer are full length. The number of reads in the forward and reversed reading direction are skewed for all loci. The reason for this is unclear. A possible cause is the fact that the template amplicons are labeled with a fluorescent dye. Steric hindrance of the dye could impede hybridization to the beads in the emPCR step or impede the sequencing by synthesis step. Unfortunately, due to time and cost limitations, we were not able to perform replicate testing for this study, so we were not able to check this hypothesis. Replicate testing in similar future studies should be a part of the experimental design where possible. There is no published data available on the effect of fluorescent labels attached to the template amplicons in GS FLX amplicon sequencing. Imbalance between the number of forward and reversed reads is not desirable. The sequence error frequency increases toward the end of reads. When an amplicon is sequenced in 2 directions, each end of the amplicon is covered by high quality sequencing data. Ideally both sequencing directions are equally represented.

We present an easy to use open-source data analysis method, which handles the several pitfalls that are inherent to GS FLX sequencing of homopolymer containing STR amplicons. In preliminary versions of our pipeline, we intuitively tried to approach the SGS data using the electrophoresis paradigm of calling an allele by its size. This approach is prone to error when using GS FLX data because this data contain too many size errors due to homopolymer sequencing errors, resulting in miscalled homopolymer sizes. This is not a problem when analyzing single contributor profiles because sequencing errors can easily be filtered. When analyzing SGS data from mixtures, it becomes impossible to distinguish small contributors from sequencing errors. Considerable effort was made trying to filter the errors from the small contributors. We, e.g., used an extra module from the AmpliconNoise software which reduces homopolymer errors using an expectation maximization algorithm on the standard flowgram files (SFF files) produced by the GS FLX sequencer. Results from the best version of the pipeline, which uses the size of clustered reads are available in online supplementary data. Results are comparable to the electrophoresis results but with several small errors that are indicated with an arrow in supplementary Fig. S1. We left this approach and chose to implement a pipeline, which reduces all homopolymers to a single base. After clustering and error correction, the homopolymer-reduced consensus sequence of the clusters is matched to a database containing the homopolymer-reduced reference sequences to call the allele. By not considering the homopolymers in the data, homopolymer sequencing errors do not influence the analysis. On the other hand information contained within the homopolymers is inherently lost. As a consequence some alleles cannot be unambiguously

Table 4
Comparison results from electrophoresis and SGS STR pipeline for single contributor sample.

	Electrophoresis (*E*)		SGS		Remarks
	Allele	Quantity (%)	Allele	Quantity (%)	
D3S1358	14	8	14	4	Same result
	15	49	15	51	Same result
	17	5	17 with SNP	4	One stutter visible from allele nr 18 with same SNP as allele nr 18
	18	38	18 with SNP	41	SNP detected in SGS profile
D5S818	10	ND	10	2	Stutter visible in SGS profile
	11	50	11	54	Same result
	12	ND	12 with SNP	3	One stutter visible from allele nr 13 with same SNP as allele nr 13
	13	50	13 with SNP	41	SNP detected in SGS profile
D7S820	9	ND	9	7	Stutter visible in SGS profile
	10	100	10	93	Same result
D8S1179	12	9	12 with SNP + 12 with SNP	2 + 3	Two stutters visible from allele nr 13 with same SNP as allele nr 13
	13	37	13 with SNP	53	SNP detected in SGS profile
	14	10	14	4	Same result
	15	43	15	39	Same result
D13S317	11	16	11	4	Same result
	12	84	12 + 12 with SNP	48 + 48	E calls homozygotic allele. SGS calls a heterozygotic allele: half of the reads cluster as the standard allele 12 and half of the reads cluster as allele 12 with a SNP
D21S11	29	7	29	4	Same result
	30	44	30″	50	Subtype 30″ detected in SGS profile
	31	ND	31	4	Unusual stutter detected in SGS profile
	33.1	ND	33.1	5	The SGS pipeline could not match the cluster to the STR database, however it was flagged as a stutter of 34.1
	34.1	49	34.1	38	Same result
vWA	15	9	15 with SNP	5	SNP detected in SGS profile
	16	46	16 with SNP	53	Same SNP in stutter
	17	45	17	43	Same result
FGA	18.2	3	18 or 18.2	2	Ambiguous call in SGS profile compared to E
	19.2	49	19 or 19.2	30 + 13	Ambiguous call in SGS profile compared to E; 30% cluster is due to out of sync sequencing
	20	ND	20	1	Unusual stutter detected in SGS profile
	21	4	21	4 + 2	4% cluster is due to out of sync sequencing
	22	44	22	36 + 12	36% cluster is due to out of sync sequencing
D18S51	11	3	11	1	Same result
	12	51	12	48	Same result
	13	46	13	51	Same result

called because different alleles can have the same sequences after compressing the homopolymers. This does not mean that homopolymer compressed GS FLX profiles contain less information and are less discriminating compared to electrophoresis profiles. The GS FLX pipeline calls allele subtypes that cannot be called by electrophoresis because they differ in sequence and not in length. Error correction of sequencing errors that are not related to homopolymers was done using the AmpliconNoise software. We chose to use the software with the default settings. The clustering parameters could be tweaked to increase the correlation between the pipeline output and the theoretically expected output. Because we analyzed only 3 samples, this could lead to parameterization that is optimized for our specific dataset, but could be sub-optimal for other datasets. Therefore this was not performed.

The called GS FLX STR alleles and their relative quantity correspond almost perfectly to the theoretical and the electrophoresis profile. In both methodologies, aberrant peaks due to PCR errors such as stutter PCR products can be present. GS FLX however offers more possibilities to analyze and flag aberrant peaks because the sequences of a suspected peak can be analyzed. The source of an electrophoresis signal can unfortunately not be further investigated. It is common practice to set a minimum threshold on the peak size in an electropherogram before the peak is considered. Our pipeline also has a similar threshold and filters clusters that are smaller than a configurable percentage of the total number of reads for that locus. Additionally, the reported alleles could be filtered based on the flagging by the pipeline, but this was not implemented in the pipeline. We chose to present our data with the filter parameter set to 1%. Empirically, we determined that a lower filter results in too many errors in the reported alleles. With the filter set to 1%, it should theoretically be possible to detect a heterozygous allele of a contributor that contributes only 5% to the STR profile. Our dataset is too limited to determine an acceptable filter for routine forensic SGS STR sequencing. A filter should not necessarily filter all aberrant peaks/clusters. In an electropherogram, stutter peaks are normally also not filtered by the minimum fluorescence intensity filter, but are simply ignored when comparing forensic profiles. Similarly, our pipeline reports and flags stutter peaks which can be ignored. The pipeline also reports another kind of aberrant peaks, which are flagged as ‘miscopy’. When a cluster is identical to another cluster except for one gap or one SNP, both clusters are flagged as a possible miscopy. The presence of 2 clusters, which differ only in one basepair is possibly caused by sequencing errors. The FGA locus is a good example showing the need for this: for each true allele, an identical allele except for one lacking thymidine just after the repeat sequence is

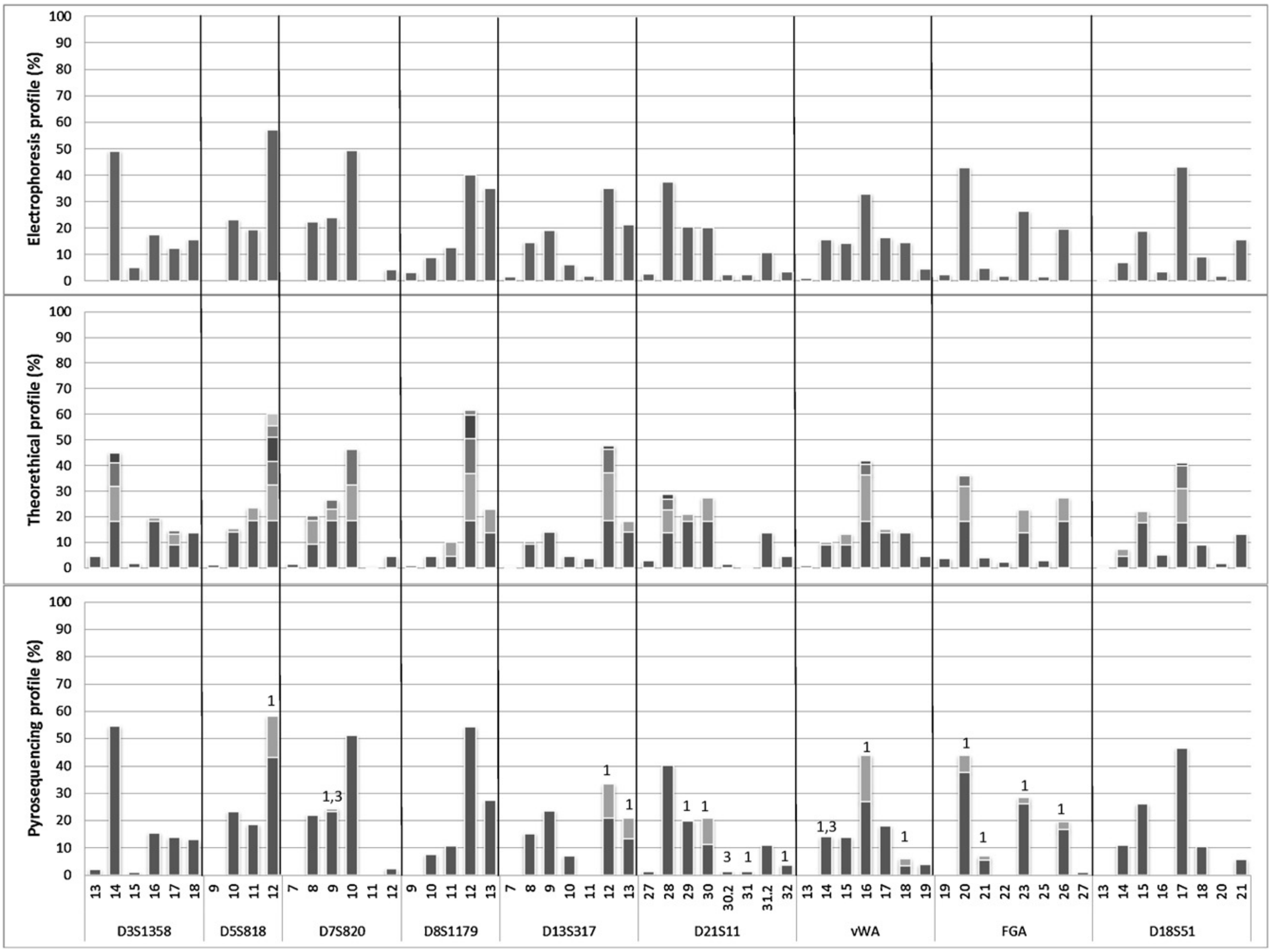

Fig. 3. STR profiles for Mixture 1.

Table 5
Comparison results from electrophoresis and SGS STR pipeline for Mixture 1.

	Electrophoresis (*E*)		SGS		Remarks
	Allele	Quantity	Allele	Quantity	
D5S818	12	57	12 + 12 with SNP	43 + 15	SNP detected in SGS profile
D7S820	9	24	9 + 9 N	23 + 1	*Extended remark:* The 23% cluster is the normal allele, the 1% cluster was unmatchable to the database but only differs by two gaps from the normal allele. Probably it is a PCR aberration that could not be filtered away
D13S317	12	35	12 + 12 with SNP	21 + 12	SNP detected in SGS profile
	13	21	13 + 13 with SNP	13 + 7	SNP detected in SGS profile
D21S11	29	20	29′ with SNP	20	Subtype detected
	30	20	30″ + 30‴	10 + 11	Two subtypes detected
	30.2	2	30.2 N	1	The SGS pipeline could not match the cluster to the STR database, however it was flagged as a stutter of 31.2
	32	4	32′	4	Subtype detected
vWA	14	16	14 N	14	*Extended remark:* The SGS pipeline could not match the cluster to the STR database. Manual sequence analysis showed it to be a variant with two SNPs compared to the normal database sequence. It appears both in mixture 1 and 2 at the percentages expected for the vWA 14 allele of contributor 4
	16	33	16 + 16 with SNP	17 + 27	SNP detected in SGS profile
	18	14	18 with SNP + 18 with SNP	3 + 2	SNP detected in SGS profile
FGA	All alleles but one have an extra cluster in the SGS profile that results from an out of sync sequencing error that could not be corrected, because there were no reads available in the reverse direction				

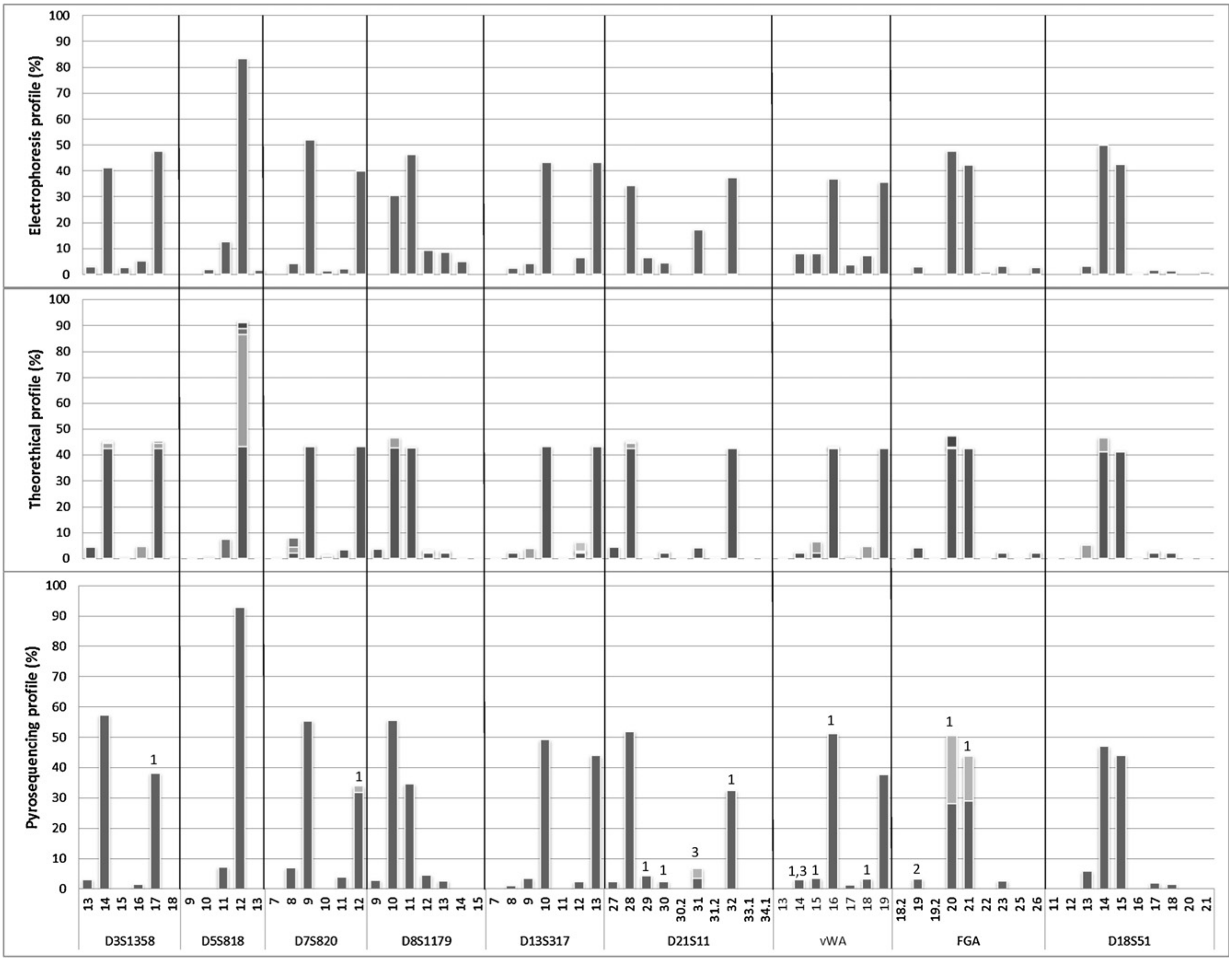

Fig. 4. STR profiles for Mixture 2.

also reported. Many of the original reads support the presence of these aberrant alleles and the clusters that contain this aberrant allele can consist of up to 30% of the reads for that locus. These reads are the result of sequencing errors caused by a long homopolymer preceding the error. As a consequence, all FGA alleles are flagged as a ‘miscopy’. Another good example is the D13S317 locus in the single contributor sample. Two alleles with equal size, but differing by one SNP, were flagged by the pipeline as miscopies. Both alleles have an almost equal number of reads in both reading directions. Therefore the most likely hypothesis is

Table 6
Comparison results from electrophoresis and SGS STR pipeline for Mixture 2.

	Electrophoresis (*E*)		SGS		Remarks
	Allele	Quantity	Allele	Quantity	
D3S1358	17	48	17 with SNP	38	SNP detected in SGS profile
D7S820	12	40	12 + 12 with SNP	32 + 2	SNP detected in SGS profile
D21S11	29	ND	29′ with SNP	4	Subtype detected
	30	6	30″	2	Subtype detected
	31	ND	31 + 31 N	3 + 3	The SGS pipeline could not match the second cluster to the STR database, however it was flagged as a stutter of 32′
	32	49	32′	32	Subtype detected
vWA	14	8	14 N	3	*Extended remark:* The SGS pipeline could not match the cluster to the STR database. Manual sequence analysis showed it to be a variant with two SNPs compared to the normal database sequence. It appears both in mixture 1 and 2 at the percentages expected for the vWA 14 allele of contributor 4
	15	8	15 with SNP	3	SNP detected in SGS profile
	16	37	16 with SNP	51	SNP detected in SGS profile
	18	7	18 with SNP	3	SNP detected in SGS profile
FGA	All alleles but one have an extra cluster in the SGS profile that results from an out of sync sequencing error that could not be corrected, because there were no reads available in the reverse direction				

that the contributor is heterozygous for this locus. However, the reported SNP could also have originated from an early PCR error leading to an equal number of reads for the two alleles. Clusters flagged as miscopies clearly need extra attention from an expert interpreting the data.

As expected, by filtering clusters smaller than 1%, the smaller contributors (0,1%, 0,5% and 1%) in mixture 2 are not detectable. It should theoretically be possible to detect a heterozygous allele of a contributor that contributes only 5% to the STR profile. All but one of the alleles of the 5% contributor to mixture 2 is reported. Only the FGA 26 allele of the 5% contributor is missing. This is probably because it was the longest allele present in that mixture for FGA. Due to the negative amplification bias toward longer alleles, the number of reads for this allele remained under the filter threshold of 1%. All alleles of the 10% contributor to mixture 1 are correctly reported by the pipeline.

5. Conclusion

Sequencing of multiplexed STR amplicons using Roche GS FLX titanium technology is technically feasible but the technology is not ideal for this purpose. The fraction of full length reads is small and the homopolymer sequencing error rate in the generated dataset is high. We present an easy to use pipeline, which compresses the homopolymers to a single base to avoid false results caused by these homopolymers. The qualitative and quantitative results from the pipeline are comparable to the results from electrophoresis. The SGS method provides extra information and is able to call allele subtypes based on STR sequences in a database. In mixed samples, all alleles were reported from individuals that contribute at least 10% to the mixture.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2012.03.004.

References

[1] D.R. Bentley, R.A. Gibbs, J.W. Belmont, P. Hardenbol, T.D. Willis, F.L. Yu, H.M. Yang, L.Y. Ch'ang, et al., The international HapMap project, Nature 426 (6968) (2003) 789–796.

[2] 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.

[3] A. Gilles, E. Meglécz, N. Pech, S. Ferreira, T. Malausa, J.F. Martin, Accuracy, quality assessment of 454 GS-FLX titanium pyrosequencing, BMC Genomics 12 (2011).

[4] S.L. Fordyce, M.C. Avila-Arcos, E. Rockenbauer, C. Borsting, R. Frank-Hansen, F.T. Petersen, E. Willerslev, A.J. Hansen, N. Morling, 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.

[5] C. Quince, A. Lanzen, R.J. Davenport, P.J. Turnbaugh, Removing noise from pyrosequenced amplicons, BMC Bioinformatics 12 (2011) 38.

[6] Applied Biosystems AmpFlSTR® Profiler Plus® PCR Amplification Kit User's Manual, 2006.

[7] P. Rice, I. Longden, A. Bleasby, EMBOSS: The European molecular biology open software suite, Trends Genet. 16 (6) (2000) 276–277.

[8] J.M. Butler, C.M. Ruitberg, D.J. Reeder, STRBase: a short tandem repeat DNA database for the human identity testing community, Nucleic Acids Res. 29 (1) (2001) 320–322.

[9] S.M. Huse, J.A. Huber, H.G. Morrison, M.L. Sogin, D.M. Welch, Accuracy and quality of massively parallel DNA pyrosequencing, Genome Biol. 8 (2007).

[10] Roche Diagnostics, 454 Sequencing Systems Portfolio Brochure, 2010.