Forensic Science International: Genetics 9 (2014) 1–8

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing☆

Christophe Van Neste [a], Mado Vandewoestyne [a], Wim Van Criekinge [b], Dieter Deforce [a,1], Filip Van Nieuwerburgh [a,1,*]

[a] Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium
[b] Biobix, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

ARTICLE INFO

Article history:
Received 25 July 2013
Received in revised form 4 October 2013
Accepted 22 October 2013

Keywords:
Illumina
MiSeq
STR
Forensic loci
MPS
NGS

ABSTRACT

Forensic scientists are currently investigating how to transition from capillary electrophoresis (CE) to massive parallel sequencing (MPS) for analysis of forensic DNA profiles. MPS offers several advantages over CE such as virtually unlimited multiplexy of loci, combining both short tandem repeat (STR) and single nucleotide polymorphism (SNP) loci, small amplicons without constraints of size separation, more discrimination power, deep mixture resolution and sample multiplexing. We present our bioinformatic framework My-Forensic-Loci-queries (MyFLq) for analysis of MPS forensic data. For allele calling, the framework uses a MySQL reference allele database with automatically determined regions of interest (ROIs) by a generic maximal flanking algorithm which makes it possible to use any STR or SNP forensic locus. Python scripts were designed to automatically make allele calls starting from raw MPS data. We also present a method to assess the usefulness and overall performance of a forensic locus with respect to MPS, as well as methods to estimate whether an unknown allele, which sequence is not present in the MySQL database, is in fact a new allele or a sequencing error. The MyFLq framework was applied to an Illumina MiSeq dataset of a forensic Illumina amplicon library, generated from multilocus STR polymerase chain reaction (PCR) on both single contributor samples and multiple person DNA mixtures. Although the multilocus PCR was not yet optimized for MPS in terms of amplicon length or locus selection, the results show excellent results for most loci. The results show a high signal-to-noise ratio, correct allele calls, and a low limit of detection for minor DNA contributors in mixed DNA samples. Technically, forensic MPS affords great promise for routine implementation in forensic genomics. The method is also applicable to adjacent disciplines such as molecular autopsy in legal medicine and in mitochondrial DNA research.

1. Introduction

Forensic DNA profiles of short tandem repeat loci are currently obtained using PCR followed by capillary electrophoresis (CE). CE separates fluorescently labeled PCR products based on their length [1]. Because of this, in order to produce unambiguous allele calls, the size ranges of STR loci with the same fluorescent tag must not overlap. This limits the number of loci that can be investigated in a single PCR and in a single capillary injection. Massive parallel sequencing (MPS) technologies, also known as second or next generation sequencing, do not rely on size separation and thus relieve the limitation on locus multiplexy [2,3]. Additionally, multiple samples can be multiplexed at the same time in a single run. MPS allows for analysis of millions of individual DNA strands (reads) in a DNA mixture, which in theory would allow for high resolution mixture analysis. Sequencing also makes it possible to detect single nucleotide polymorphisms (SNPs) and STR sequence variants in addition to the gross STR repeat number [4]. This allows analysts to tell the difference between equilength alleles in a DNA mixture. Certain mass spectrometry techniques also make it possible to differentiate equilength alleles, but complete characterization of polymorphism can only be accomplished by sequencing [5].

Forensic scientists are currently investigating how to transition from CE to MPS. Several bioinformatic tools are being developed to that end [2,6–8]. Previously, we reported that sequencing of multiplexed STR amplicons using Roche GS FLX titanium technology was technically feasible both in single contributor samples and in multiple person DNA mixtures, notwithstanding a

* Corresponding author. Tel.: +32 92648052.

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.fsigen.2013.10.012
1872-4973/

poor performance for some frequently used forensic loci [6]. For those loci, the GS FLX reads needed to be transformed by a homopolymer-compression algorithm to obtain results consistent enough for mixture analysis. In the GS FLX STR analysis, the reads needed to be clustered around a consensus sequence. The region of interest (ROI) was set using fixed flanking based on the specific PCR primers for each locus. A signal threshold was used to determine which read clusters are considered a signal and which clusters are considered noise.

In the current study we created a general bioinformatic framework, that we call My-Forensic-Loci-queries (MyFLq), for analyzing MPS-generated forensic data. It is designed to handle multiple locus types, including STR length polymorphisms and SNPs. The framework uses a MySQL reference allele database with automatically determined ROIs and python scripts to compare MPS sequences against the known allele database. We also present a method to asses the usefulness and overall performance of a forensic locus when used in an MPS analysis, and a method to estimate whether an allele which is not present in the database is in fact a newly typed allele or a PCR/sequencing error.

The MyFLq framework was used on an Illumina MiSeq dataset of an Illumina forensic amplicon library generated from STR multilocus PCR on both single contributor samples and multiple person mixtures. The multilocus PCR was not specifically optimized for MPS in terms of criteria such as locus selection or amplicon length. The results are promising, showing excellent results on most but not all loci. The raw read accuracy was high enough so that the reads did not need to be clustered around a consensus accuracy as with the GS FLX reads [6]. The frequency of homopolymer errors in the MiSeq data is insignificant compared to the GS FLX technology. Nevertheless, the homopolymer compression algorithm still proved useful to group some reads containing errors, whether from PCR or from sequencing, with those that were completely error-free in respect to a contributor's reference sequence.

The MyFLq framework has a Creative Commons open source license (CC BY-SA 3.0). The source code is available as supplementary material or for the latest version at http://MyFLq.UGent.vanneste.be. Currently we are implementing it as an Illumina BaseSpace application, which should be available by the beginning of 2014.

2. Materials and methods

2.1. Sample preparation and processing using Illumina chemistry

DNA mixtures were prepared according to Table 1 from the following purified genomic DNA sources: K562 (Promega), 9947A (Promega), 2800M (Promega) and two National Institute of Standards and Technology (NIST) standard reference materials (SRM 2391c: DNA A, DNA B). DNA concentration of each sample was measured using the Qubit® dsDNA HS Assay and Qubit Fluorimeter following manufacturer's instructions. A mixture of four source DNAs and a mixture of five source DNAs, along with 9947A and K562 single source samples, were used in multilocus PCR (Table 1).

Table 1
DNA composition of samples.

DNA standard	Sample 1 (%)	Sample 2 (%)	Sample 3 (%)	Sample 4 (%)
K562	100		0.10	
9947A		40	0.50	100
NIST SRM 2391c DNA A		30	1	
NIST SRM 2391c DNA B		20	5	
2800M		10	93.40	

Primer sequences were ordered without fluorescent tags (Integrated DNA Technologies) for loci: Amelogenin, CSF1PO, D13S317, D16S539, D18S51, D21S11, D3S1358, D5S818, D7S820, D8S1179, FGA, PentaD, PentaE, TH01, TPOX, and vWA [9]. This multiplex has yet to undergo optimization (*e.g.*, for intra- and inter-locus balance, polymerase stutter (slippage) and other artifacts) for use in MPS. Primers were used at a concentration of 2 μM each in PCR with 1 ng of DNA or DNA mixture in 1× Gold STR buffer (Promega) and 0.16 U AmpliTaq GOLD (Invitrogen). The samples were amplified using a BioRad Tetrad instrument as follows: 95 °C for 11 min, 96 °C for 1 min; 10 cycles of 94 °C for 30 s, ramp 0.5 °C/s to 60 °C and then 30 s at 60 °C, ramp 0.2 °C/s to 70 °C and then 45 s at 70 °C; 22 cycles of 90 °C for 30 s, ramp 0.5 °C/s to 60 °C and then 30 s at 60 °C, ramp 0.2 °C/s to 70 °C and then 45 s at 70 °C; hold at 60 °C for 30 min; 4 °C soak.

PCR products were quantified using Qubit® (Invitrogen) without purification. Libraries were generated by ligating TruSeq DNA adapters to the PCR products from 50 ng of unpurified PCR product (Illumina). Samples were subjected to 5 cycles of PCR and purified with SPRI (TruSeq DNA Sample Preparation Guide). The completed libraries were quantified using a qPCR assay as recommended by Illumina.

Libraries were pooled with Phi X Universal Library and a Human DNA library. Pooled libraries were denatured and diluted to 10 pM following Illumina guidelines and sequenced on a MiSeq using a MiSeq Reagent Kit v2, 500 cycles with a modified recipe. Samples were demultiplexed using the index sequences, FASTQ files were generated automatically using MiSeq Reporter (MSR).

2.2. MiSeq data analysis

2.2.1. The MyFLq framework

All necessary steps for an MPS forensic analysis were incorporated into our open-source framework. MyFLq is not yet a full application. The end results need to be statistically analyzed: probabilities of allele calls, stutter filtering, hetero- and homozygotic allele calling and visualization are not yet implemented.

The framework consists of two parts: (1) A MySQL database backend that is populated by known reference alleles, and (2) a Python frontend with functions for adding reference alleles to the reference allele database, and analysis of MPS STR data from forensic samples. The source code and documentation of all functions can be found in supplementary materials. It also lists specifically the functions used for this paper, and provides a short description for each.

2.2.2. Building the reference allele database

The allele database is ideally populated with the sequences of all known STR alleles that exist in the general population. Because each of these sequences is currently not available, the database was initially populated with the STR sequences from the DNA sources in Table 1. These reference sequences were manually inferred from the Illumina sequencing data and the STRBase allele database [10]. They are not the best representation of population alleles, but suffice for the current study. In the future, with MPS the known diversity will be better determined.

After building the reference allele database, the function processLociNames was used to determine the flanking region for each locus. The function processLociAlleles produced a table containing the ROI for each allele and its integer allele number (if STR) according to standard nomenclature [11,12].

Flanking regions are the maximum right- and left-end consensus between all alleles of a locus in the reference allele database. Fig. 1 shows a simplified example. Each allele has two primers, two flanks and the ROI. If the reference database alleles for a locus only differ by the number of STR repeats and thus no SNPs,

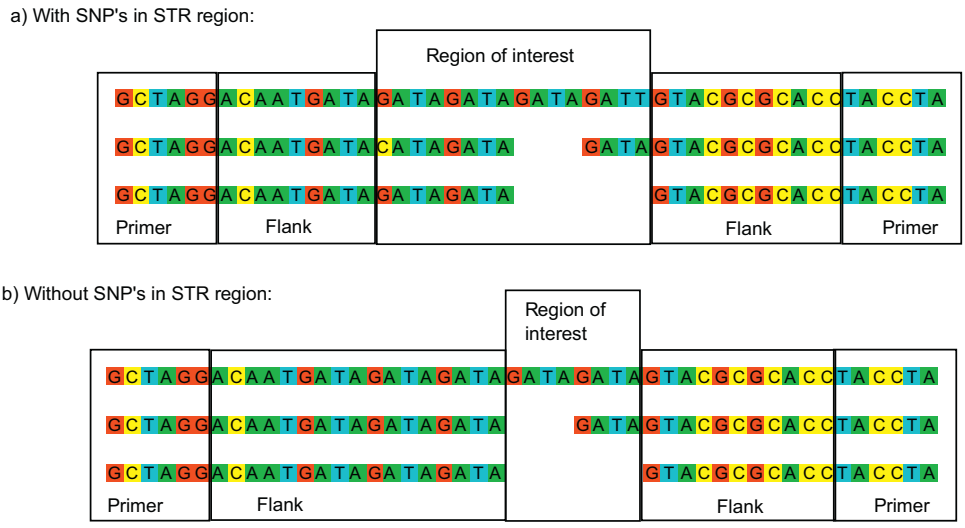

Fig. 1. Flanks and region of interest in the reference allele database. A generic locus with three alleles in the database is considered for two possible cases: with or without SNPs within the STR region.

non-consensus, or partial repeat patterns are present, one of the alleles for that locus can have a ROI of zero length, as Fig. 1b demonstrates. Because the ROI is calculated dynamically based on the sequences present in the reference allele database, the ROI of the loci can change when the reference allele database is updated.

2.2.3. STR data analysis

After building the database, all FASTQ files were analyzed with the framework software as if they would have been from unknown samples. Analysis consisted of following steps: (1) reads were assigned to a locus based on the presence of both PCR primer sequences for that locus; (2) primers and flanks were removed from the reads, leaving only the read ROI; (3) reads were grouped based on their exact sequence; (4) groups with an abundance lower than 0.5% were discarded; (5) the ROI of each group was compared to the allele database table and an allele call was made when an exact match was found; and (6) groups within a locus were compared to each other, a connection was flagged when two ROIs differed by maximum two SNPs or STR size indels.

In step (2) the determination of the flanking regions was done as follows: if a read-end matched exactly with the database flank, the read-end was removed and flagged 'clean'. Otherwise, the database flank and read-end were homopolymer-compressed (two same consecutive bases were already considered a homopolymer). When this resulted in an exact match between those sequences, the read-end was removed and flagged 'compressed-clean'. If the flank was still not removed, the read-end was flagged as 'unclean' and was removed by our flexible flanking algorithm (see below). In step (4), for each group, counts were gathered on 'clean', 'compressed-clean', and 'unclean' read-ends. Step (6) indicates how the ROIs in the results are interrelated. This helps forensic researchers to decide on the likelihood of an allele call, *e.g.* a ROI that is not in the database, has a low abundance in the results and only differs by one SNP of another high abundant ROI, is probably an error-containing ROI.

2.2.4. Flexible flanking algorithm

This algorithm always removes a flank from a read, no matter how dissimilar to the database flank. The starting hypothesis is that the read-end to remove is as long as the database flank. K-mers for the database flank with increasing length are searched around their expected index in the read. The found index is scored depending on how informative the k-mer is: more informative if longer and if closer to the flank-end. The score is calculated as the square root of the product of those two values (length of k-mer and inverse distance to database flank end). Finally, the proposed index with the highest summed score is considered to be the most likely ending flank position. Based on that position, the read-end is removed. For a more in depth explanation, documentation is provided for this algorithm in the source file in supplementary materials.

3. Results

3.1. Setting general threshold

In the analysis of MPS STR data, groups with an abundance lower than 0.5% were discarded (see step 4 of 2.2.3). This threshold was arbitrarily determined based on the results in Fig. 2. Fig. 2 shows a histogram of the abundances of all grouped identical reads for the single contributor samples. To generate this figure, the complete sequences were considered, except for the part outside of the primers.

Erroneous reads are expected to have much smaller abundances than error free reads, especially in single contributor samples. There are around 10^5 groups of identical reads with abundances smaller than 0.5%, which are in the context of the single contributor samples definitely reads with errors. The threshold for unique reads was set to 0.5%, to avoid cluttering the results with noise. By doing so, minor contributors which contribute less than 0.5% to the DNA mixture will not be detected.

3.2. General properties of the Illumina dataset

Table 2 shows the total number of MiSeq reads for each sample, the number of reads filtered based on the 0.5% abundance threshold, and the number of error free ROI after filtering. Filtered reads consist of both reads with and without errors. Error free ROI are the number of reads of which the complete ROI is identical to a reference database ROI and of which the ROI is expected to be present in the relevant sample. These reads can still have errors in the flanks around the ROI, but these flanks are not considered when matching the reads to the reference allele database. This percentage is influenced by many factors such as PCR accuracy, amplicon length, ROI and flank length, stutter, sequencing accuracy and the abundance filter cut-off.

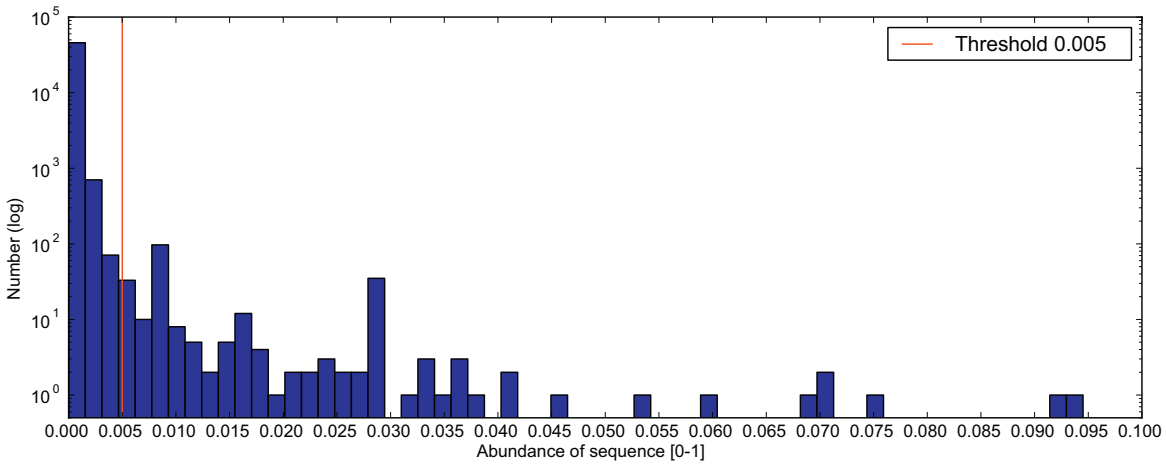

Fig. 2. Histogram of the full-sequence abundances. Groups with abundances higher than 0.1 not shown.

Table 2
General Illumina samples characteristics.

	Sample 1	Sample 2	Sample 3	Sample 4
Total MiSeq reads	246,347	1,176,806	1,261,848	961,236
Filtered reads	203,181 (82%)	981,935 (83%)	1,073,164 (85%)	771,723 (80%)
Error free ROI	173,692 (71%)	912,643 (72%)	996,466 (79%)	681,918 (71%)

3.3. STR allele calls in four and five person DNA mixtures

Figs. 3–6 show the allele calls for the different samples. In the figures, blue bars denote the theoretical abundance of the allele based on how the samples were prepared, green bars the detected read abundance, and red bars erroneous reads (including polymerase stutter). Pure samples (Figs. 3 and 6), contain, at most, two blue bars per locus. Erroneous read bars have been drawn narrower to make it possible to show all erroneous read groups. The figures are automatically generated from the MyFLq result files (in supplementary materials), with manual addition of the blue bars.

3.4. Progressive abundance threshold

Fig. 7 shows, after full analysis, for each locus, the percentage of error free sequences (*Y*-axis) for a given abundance threshold (*X*-axis). Higher abundances for erroneous sequences are less likely.

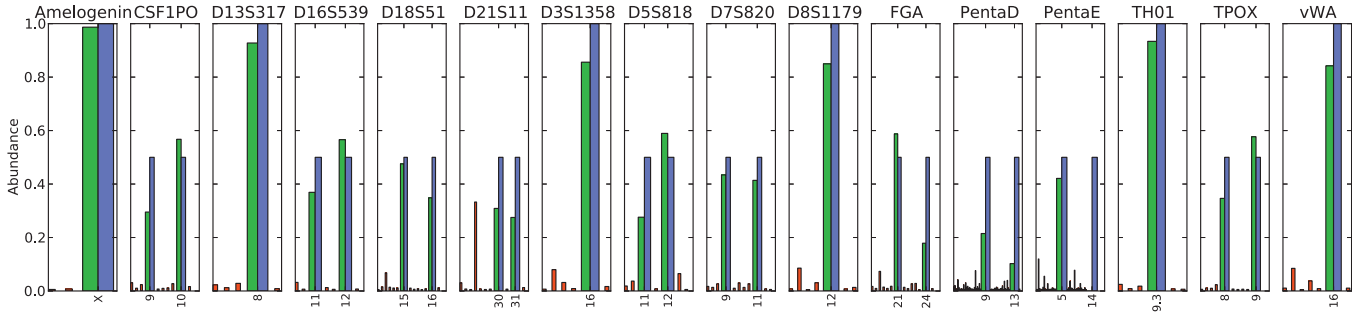

Fig. 3. Sample 1 profile. Blue bars = theoretical abundance, green bars = detected read abundance, red bars = erroneous read abundance.

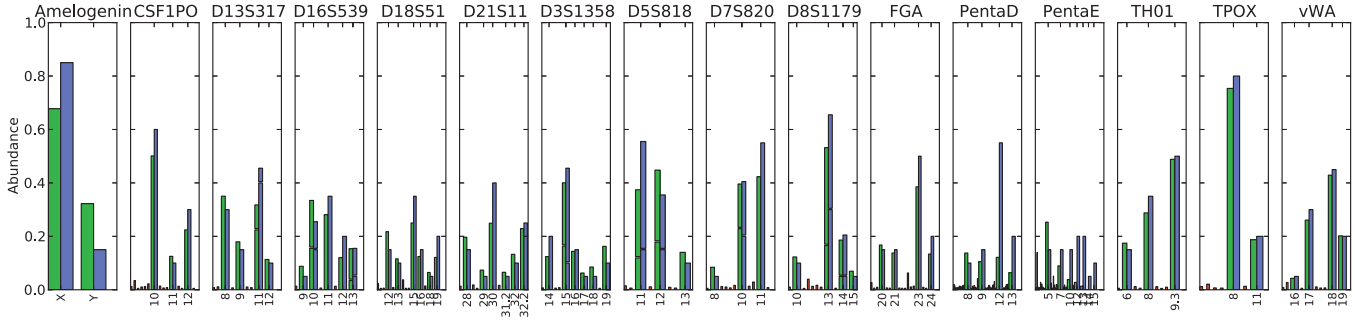

Fig. 4. Sample 2 profile. Blue bars = theoretical abundance, green bars = detected read abundance, red bars = erroneous read abundance.

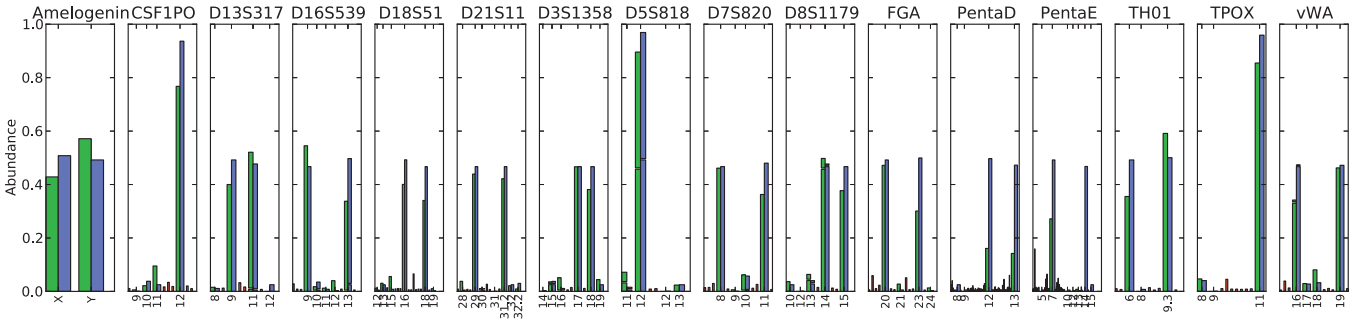

Fig. 5. Sample 3 profile. Blue bars = theoretical abundance, green bars = detected read abundance, red bars = erroneous read abundance.

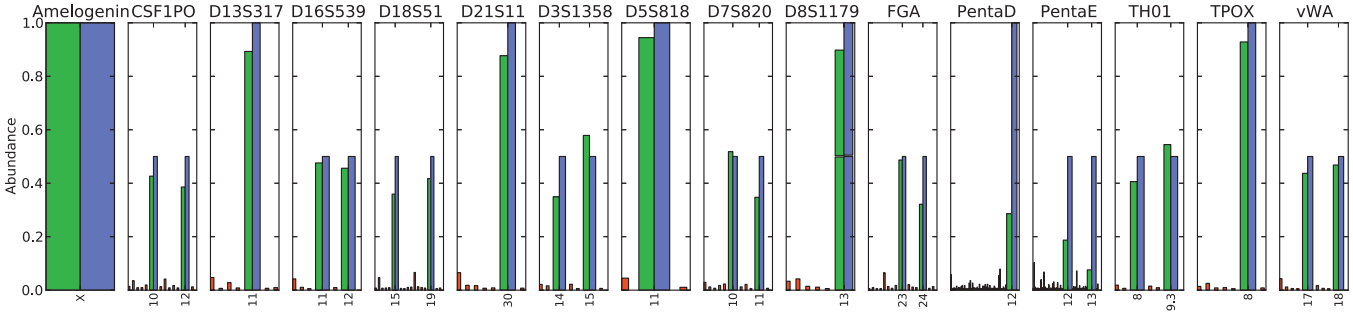

Fig. 6. Sample 4 profile. Blue bars = theoretical abundance, green bars = detected read abundance, red bars = erroneous read abundance.

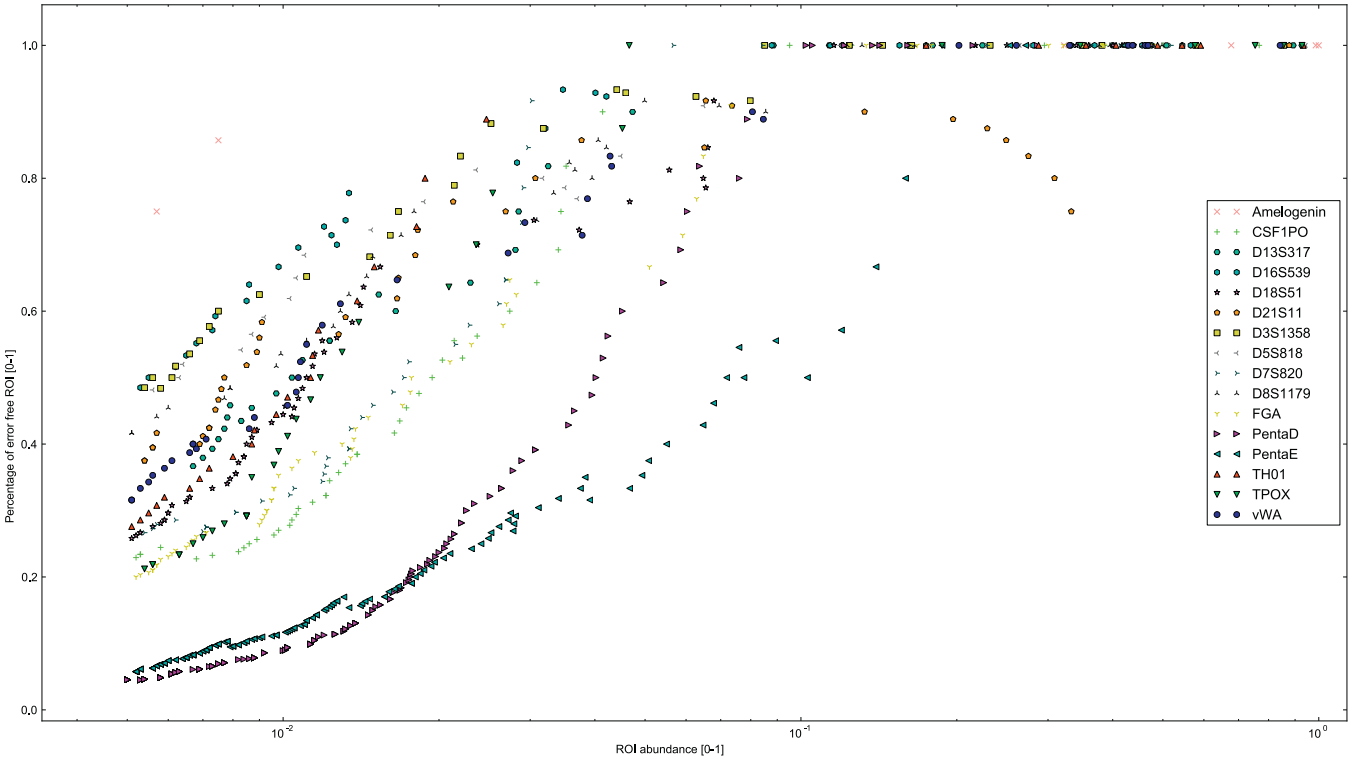

Fig. 7. Percentage of error free ROI for a given abundance threshold.

When the abundance threshold is increased, the percentage of error free sequences increases and is expected to become 100%. Loci for which less error sequences are produced, have a lower threshold for which 100% error free reads are remaining. Based on this criterion, Amelogenin and D16S539 are the best performing loci while PentaE and PentaD perform at a decreased level.

3.5. Locus quality analysis

For each locus, Table 3 shows the percentage of the theoretical abundance which is measured as MPS signal. The theoretical abundance can be calculated from the proportion of each contributor to the sample (see Table 1) and are shown as blue

Table 3
MPS signal percentage of theoretical abundance.

Locus	MPS signal/theoretical abundance (%)	Std-error (%)
Amelogenin	95	4
CSF1PO	79	5
D13S317	89	5
D16S539	88	6
D18S51	75	6
D21S11	78	5
D3S1358	86	5
D5S818	89	5
D7S820	84	6
D8S1179	95	6
FGA	73	6
PentaD	25	5
PentaE	27	7
TH01	92	5
TPOX	90	4
vWA	85	5

bars in Figs. 3–6. The measured signals are shown as green bars in the same figures. When there would be no PCR errors like stutter and no sequencing errors, the MPS signal would be 100%. The numbers in Table 3 were calculated by linear model for each locus separately but for all locus specific alleles of all samples together.

Fig. 8 shows the average percentage of flanks which are 'clean' in reads with an error-free ROI and in reads with an error-containing ROI. For all loci, except PentaD and D5S818, the proportion of clean flanks is higher in reads which also contain an error-free ROI. The proportion of clean flanks is also impacted by the length of the flanks. The negative correlation between the length of the flanks and the proportion of clean flanks is expected as longer sequences carry a higher probability of containing errors.

4. Discussion

4.1. Framework performance

The MyFLq framework was designed to incorporate all algorithms necessary for an STR MPS analysis in one open framework. It contains about a thousand lines of code and has minimal dependencies: It only requires a working python2 environment and the common python packages numpy and MySQLdb. Currently the target audience of the framework are bioinformaticians that work in the field of MPS forensic analysis. An Illumina BaseSpace application is being developed, which should be available by the beginning of 2014. This way, users will not have to interact with the framework, but will be able to analyze their STR data files with the MyFLq application.

MyFLq can also handle other types of forensic data, such as SNP or mitochondrial regions. However, it has not been extensively tested to operate with such data. From the analyzed loci in this study, only Amelogenin could be considered a SNP locus. STRait Razor is another tool that has been recently developed for forensic genomics and as the name indicates only deals with STR's [8]. In our opinion, future software packages should handle both STR and SNP loci equally well, to provide a full solution to forensic researchers.

MyFLq has currently not been designed to be computational efficient. A full analysis of a MiSeq dataset takes approximately one hour on a single CPU. The code contains many sections that could

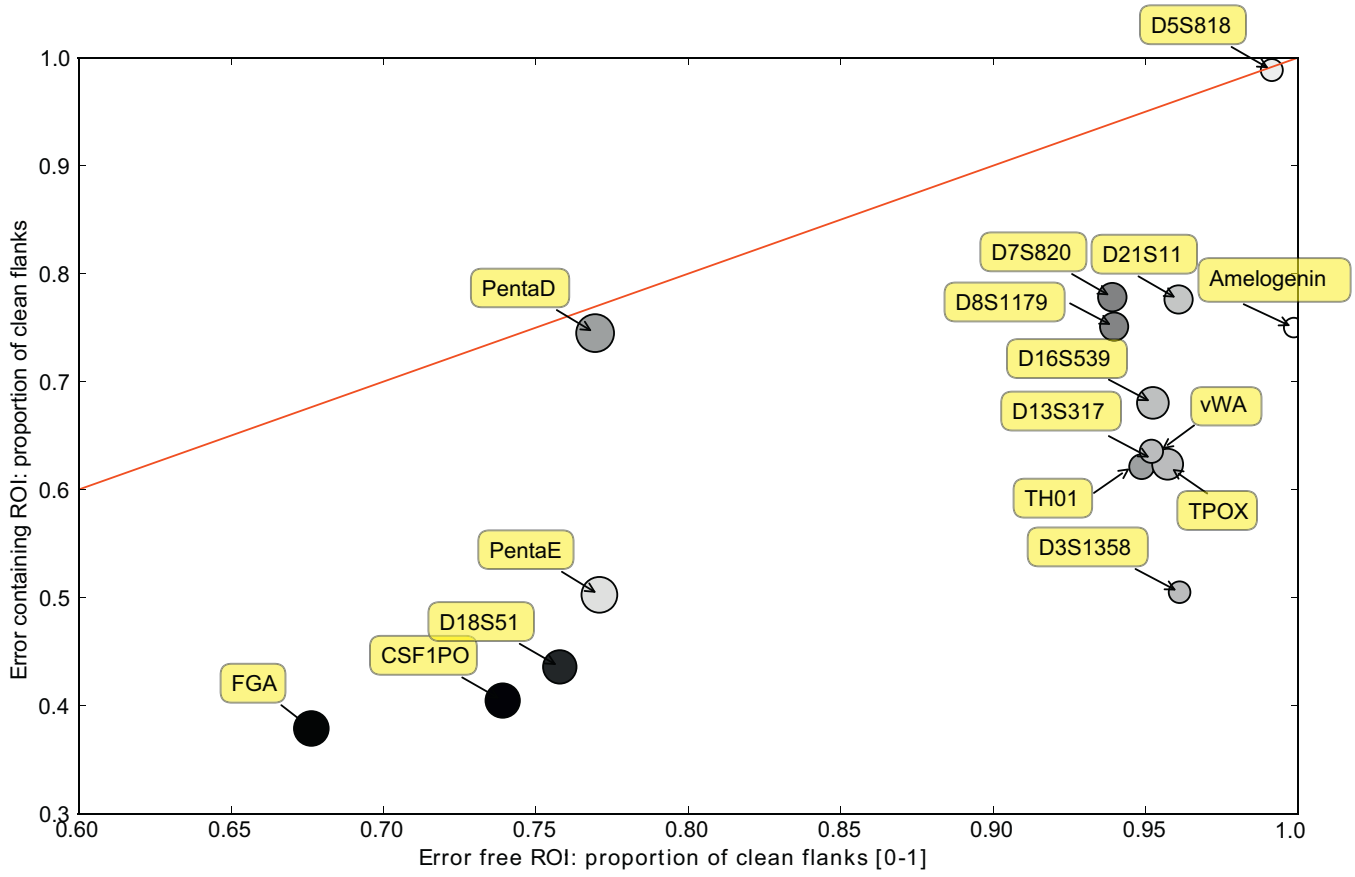

Fig. 8. Proportion of clean flanks in reads with an expected ROI and in reads with an error-containing ROI. Size of circles is proportional to average allele length of the locus, darkness of circles is proportional to flank length of the locus.

be parallelized to reduce the analysis time significantly. This will also be implemented in the BaseSpace application.

4.2. Accuracy and detection of minor contributors

In general the Illumina MiSeq reads were of high quality, with more than 70% of reads containing a complete error free ROI and constitute the MPS signal. Several factors are influencing this percentage. Stutters (around 10%) and the reads filtered by the abundance threshold (around 20%) are a big part of the other 30% of the reads. PCR and sequencing accuracy with a per base sequencing accuracy of approximately 99.5% are other main factors. Reducing the amplicon, and ROI length could further improve the MPS signal. Reads can be used without clustering them to produce a consensus sequence. This is of particular importance, since clustering should be avoided when investigating forensic mixtures. An allele from a minor contributor containing a SNP or an insertion/deletion could cluster with the sequences from the corresponding allele of a major contributor and go undetected. When the contributors are unknown, it is impossible to categorize *a priori* between alleles and amplification errors. Consequently, clustering should be avoided if possible, in order to confront the forensic investigator at least with the presence of these sequences.

Because the minimal abundance threshold during data analysis was set to 0.5%, theoretically it was still possible to detect alleles from the 1% contributor in *Sample 3*. However, only for the 5% contributor the alleles were detected consistently, showing a correct allele call for all loci except D13S317. However, the experiment was not set up to determine the allelic detection limit. Many alleles of the minor contributors are the same as one of the more abundant contributors. Future research will be needed to establish a general detection limit in mixtures for each MPS technology and PCR multiplex.

4.3. Loci performance

In capillary electrophoresis (CE), the allele specific signal is 85–90%, as polymerase stutter products usually comprise approximately 10–15% of the parent allele's signal. These stutter products were also observed in MPS data. These stutter sequences combined with sequencing errors results in an allele specific MPS signal that is slightly lower than the absolute DNA input. For most loci these values are very reasonable, as shown in Table 3, except for PentaE and PentaD. Those two loci are obvious underperformers in the current experimental setup, both with allele specific signals around 25%. Compared to the other loci, these 2 loci have long amplicons, long ROI and long flank lengths. While these factors are partially contributing to a higher proportion of reads containing errors in the ROI, they are not completely explaining the low signal. It is unclear at which point that the errors are introduced, but given the high abundance of some of the errors, they are probably introduced at the PCR step. Future research will show how useful they will be in MPS STR analyses.

Fig. 8 shows a higher proportion of clean flanks in reads with error-free ROIs. This data could be modeled to estimate the likelihood of error in ROIs of an unknown sample. Fig. 8 shows that the usefulness of this strategy will depend on the considered locus because values for 'clean', 'compressed-clean' and 'unclean' depend on locus sequence characteristics such as average length and homopolymer content.

For an MPS technology to become a valid alternative to CE it must produce at least equivalent results for the commonly used loci, and produce additional, valuable information for an expanded set of loci. Most loci already performed very well with the Illumina MiSeq, while some (PentaD, PentaE) need further optimization, such as primer re-design, in order to eclipse CE. As MPS becomes a valid alternative, our framework can be used to help identify an additional set of ideal MPS loci.

4.4. Dynamic flank calculation

Using sequencing, only the ROI needs to be considered during data analysis, *i.e.* the part of the sequence that differs between different alleles for a locus. Any part outside that region can be treated as flanking. As our framework is built to generically handle all forensic loci, the ROI and the flanks around the ROI are not predefined, but are dynamically determined. This dynamic flanking tries to maximize the flanks and to minimize the ROI for each locus based on the reference alleles present in the reference allele database. Removing flanks from the reads minimizes the impact of errors in the flanks on the analysis. This aids detecting alleles of minor DNA contributors, as a significant portion of the noise is eliminated.

Application of our method is currently limited by the size of the reference allele database, which contains a small subset of all of the possible alleles, *i.c.* only alleles present in the DNA from Table 1. To determine the practical applicability of our framework, more samples need to be analyzed to increase the allele database size. However, for future applications it would also be useful to automatically limit the database size by subsetting it. Database alleles could, for example, be selected based on the size of the sequences in a specific sample, because the size of reads already reduces the set of possible alleles to which they could be assigned. This subset of possible alleles would then serve to calculate maximum flanks *on the fly*. Smaller subsets will result in bigger flanks which in turn reduces the impact of errors.

5. Conclusion

When MPS becomes routinely used to analyze forensic DNA profiles, decisions will need to be made on how the data should be processed. We present our bioinformatic framework, MyFLq, that processes forensic MPS data prudently: without clustering, extracting maximal information with automatically determined regions of interest. The results show Illumina MiSeq is ready to analyze STR profiles. For routine implementation in forensic laboratories, a careful selection of loci, PCR multiplex optimization, and an MPS-based STR allele reference database for alignment, are needed.

Supplementary materials

MyFLq.py The main framework program.
InstallMyFLqTables.py Installs the MySQL tables necessary for using MyFLq.
results.css Needed in the same directory as the sample result files, for opening the result files in a browser. This file also contains documentation to interpret the result files.
sample1–4.xml Sample result files.

Acknowledgments

Funding was provided by the Ghent University Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks'.

The authors would also like to thank Illumina, Inc. for providing their technical expertise in sample preparation and MiSeq sequencing.

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.2013.10.012.

References

[1] D.L. Deforce, R.E. Millecamps, D.V. Hoofstat, E.G.V. den Eeckhout, Comparison of slab gel electrophoresis and capillary electrophoresis for the detection of the fluorescently labeled polymerase chain reaction products of short tandem repeat fragments, J. Chromatogr. A 806 (1) (1998) 149–155. , http://dx.doi.org/10.1016/S0021-9673(97)00394-4.

[2] S.L. Fordyce, M.C. Avila-Arcos, E. Rockenbauer, C. Borsting, R. Frank-Hansen, F.T. Petersen, E. Willerslev, A.J. Hansen, N. Morling, M.T.P. Gilbert, High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, Biotechniques 51 (2) (2011) 127+, http://dx.doi.org/10.2144/000113721.

[3] K.K. Kidd, J.R. Kidd, W.C. Speed, R. Fang, M.R. Furtado, F.C.L. Hyland, A.J. Pakstis, Expanding data and resources for forensic use of SNPs in individual identification, Forensic Sci. Int. Genet. 6 (5) (2012) 646–652, doi:10.1016/j.fsigen.2012.02.012.

[4] L.J. McIver, J.W. Fondon III, M.A. Skinner, H.R. Garner, Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments, Genomics 97 (4) (2011) 193–199. , http://dx.doi.org/10.1016/j.ygeno.2011.01.001.

[5] J.V. Planz, K.A. Sannes-Lowery, D.D. Duncan, S. Manalili, B. Budowle, R. Chakraborty, S.A. Hofstadler, T.A. Hall, Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry, Forensic Sci. Int. Genet. 6 (5) (2012) 594–606. , http://dx.doi.org/10.1016/j.fsigen.2012.02.002.

[6] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, D. Deforce, Forensic STR analysis using massive parallel sequencing, Forensic Sci. Int. Genet. 6 (6) (2012) 810–818. , http://dx.doi.org/10.1016/j.fsigen.2012.03.004.

[7] M. Gymrek, D. Golan, S. Rosset, Y. Erlich, lobSTR: a short tandem repeat profiler for personal genomes, Genome Res. 22 (6) (2012) 1154–1162. , http://dx.doi.org/10.1101/gr.135780.111.

[8] 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 (7) (2013) 409–417.

[9] A. Masibay, T. Mozer, C. Sprecher, Promega Corporation reveals primer sequences in its testing kits, J. Forensic Sci. 45 (6) (2000) 1360–1362.

[10] C. Ruitberg, D. Reeder, J. Butler, STRBase: a short tandem repeat DNA database for the human identity testing community, Nucleic Acids Res. 29 (1) (2001) 320–322. , http://dx.doi.org/10.1093/nar/29.1.320.

[11] P. Gill, C. Brenner, B. Brinkmann, B. Budowle, A. Carracedo, M. Jobling, P. de Knijff, M. Kayser, M. Krawczak, W. Mayr, N. Morling, B. Olaisen, V. Pascali, M. Prinz, L. Roewer, P. Schneider, A. Sajantila, C. Tyler-Smith, DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs, Int. J. Legal Med. 114 (6) (2001) 305–309. , http://dx.doi.org/10.1007/s004140100232.

[12] B. Olaisen, W. Bar, B. Brinkmann, B. Budowle, A. Carracedo, P. Gill, P. Lincoln, W. Mayr, S. Rand, DNA recommendations 1997 of the International Society for Forensic Genetics, Vox Sang. 74 (1) (1998) 61–63. , http://dx.doi.org/10.1046/j.1423-0410.1998.7410061.x.