Forensic Science International: Genetics 24 (2016) 86–96

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Massively parallel sequencing of short tandem repeats—Population data and mixture analysis results for the PowerSeq™ system

Kristiaan J. van der Gaag[a,b], Rick H. de Leeuw[a], Jerry Hoogenboom[b,c], Jaynish Patel[d], Douglas R. Storts[d], Jeroen F.J. Laros[c,e,f], Peter de Knijff[a,*]

[a] Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University Medical Centre, Postzone S 05 P, P.O. Box 9600, 2300 RC Leiden, The Netherlands
[b] Biological Traces, Netherlands Forensic Institute, Laan van Ypenburg 6, 2497GB The Hague, The Netherlands
[c] Department of Human Genetics, Leiden University Medical Centre, 2300 RC Leiden, The Netherlands
[d] Promega Corporation, 2800 Woods Hollow Road, Madison, WI 53711, USA
[e] Leiden Genome Technology Centre, 2300 RC Leiden, The Netherlands
[f] Netherlands Bioinformatics Centre, Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 5 April 2016
Received in revised form 25 May 2016
Accepted 29 May 2016
Available online 7 June 2016

Keywords:
Forensic science
Short tandem repeat (STR)
Next Generation Sequencing (NGS)
Massively Parallel Sequencing (MPS)
Mixture analysis
Sequence variants
Bioinformatics
STR stutter
MiSeq
Illumina
Promega
PowerSeq
TSSV
fdstools

ABSTRACT

Current forensic DNA analysis predominantly involves identification of human donors by analysis of short tandem repeats (STRs) using Capillary Electrophoresis (CE). Recent developments in Massively Parallel Sequencing (MPS) technologies offer new possibilities in analysis of STRs since they might overcome some of the limitations of CE analysis. In this study 17 STRs and Amelogenin were sequenced in high coverage using a prototype version of the Promega PowerSeq™ system for 297 population samples from the Netherlands, Nepal, Bhutan and Central African Pygmies. In addition, 45 two-person mixtures with different minor contributions down to 1% were analysed to investigate the performance of this system for mixed samples. Regarding fragment length, complete concordance between the MPS and CE-based data was found, marking the reliability of MPS PowerSeq™ system. As expected, MPS presented a broader allele range and higher power of discrimination and exclusion rate. The high coverage sequencing data were used to determine stutter characteristics for all loci and stutter ratios were compared to CE data. The separation of alleles with the same length but exhibiting different stutter ratios lowers the overall variation in stutter ratio and helps in differentiation of stutters from genuine alleles in mixed samples. All alleles of the minor contributors were detected in the sequence reads even for the 1% contributions, but analysis of mixtures below 5% without prior information of the mixture ratio is complicated by PCR and sequencing artefacts.

1. Introduction

Current forensic DNA analysis almost exclusively focuses on the identification of human sample donors using multiplex short tandem repeat (STR) genotyping with commercial kits based on polymerase chain reaction (PCR) and capillary electrophoresis (CE). Although this type of analysis has proven its value over the past decades, it is not without limitations. In CE, multiplexing of more than 5 loci in a single assay can only be achieved by using different fluorescent labels in the PCR and by using non-overlapping PCR fragment lengths for STRs with the same fluorescent label. Consequently, most commercial assays have a PCR fragment range between 80 and 500 bp [20].

When analysing degraded DNA samples, this variation in fragment length frequently results in noticeable lower, or even absent, signals for the longer PCR fragments. As a consequence, profiles of degraded DNA often have a lower discriminating power.

Another potential difficulty associated with the CE detection of STRs is the background signal arising from stutter peaks [19], caused by slippage of the polymerase in the PCR. In DNA samples from a single person, genuine alleles and stutter alleles can be easily distinguished. However, the analysis of unbalanced mixtures with low minor contributions is frequently complicated by stutter

* Corresponding author.
E-mail addresses: k.van.der.gaag@nfi.minvenj.nl (K.J. van der Gaag), r.h.de_leeuw@lumc.nl (R.H. de Leeuw), j.hoogenboom@nfi.minvenj.nl (J. Hoogenboom), Jaynish.Patel@promega.com (J. Patel), doug.storts@promega.com (D.R. Storts), j.f.j.laros@lumc.nl (J.F.J. Laros), p.de_knijff@lumc.nl (P. de Knijff).

http://dx.doi.org/10.1016/j.fsigen.2016.05.016

alleles that cannot be distinguished from genuine alleles of the minor contributors [4].

In theory, these limitations can mostly be solved by the use of massively parallel sequencing (MPS) of STR loci. STR alleles can be identified by repeat number and sequence variation and primers can be designed in such a way that PCR fragments have similar size ranges for all loci. Moreover, many more loci can be multiplexed in the same reaction because the detection is no longer based on a limited number of fluorescent labels. A few studies have indicated the potential of MPS STR genotyping [6,8,15,21]. They showed that, in addition to the variation in repeat number and repeat sequence, the repeat-flanking regions provide an additional source of variation and add to the discriminating power of the loci. However, the additional power of this new sequence variation cannot be fully used until sufficient population frequency data is available for all loci. We speculated that this additional information could help in distinguishing genuine alleles from stutter alleles although it is not likely that this problem will be completely overcome.

For this purpose, we assessed population data for 297 samples of three distinct populations (Dutch, Himalayan, and Central African Pygmies) for 17 STR loci included in a prototype version of the PowerSeq™ MPS STR assay [21]. These data were compared to the results of CE-based data from the PowerPlex® Fusion System [12]. We also present data from several series of mixed DNA samples in different ratios down to 1:99 to survey the possibilities and limits for this assay in analysis of mixed samples.

We examined the additional sequence variation of the loci, both within the STR motifs and in the flanking regions, and assessed the impact of this variation on the discriminating power of the loci. In addition, stutter ratios were studied and compared to those obtained with CE-based profiling.

2. Material and methods

2.1. Population samples

To assess the potential genetic variation, 297 DNA samples were selected from a European population (101 Dutch samples [20]), an Asian population (97 samples from Nepal and Bhutan [10]) and an African population (99 Central African Pygmy samples [9]).

2.2. Capillary electrophoresis

PCR reactions were performed according to the protocol of the PowerPlex® Fusion System [14] using 0.5 ng of DNA and 30 amplification cycles using a GeneAmp® PCR System 9700 (Life Technologies). For every reaction, 2800 M Control DNA (Promega) was included as a positive control and a water sample was included as negative control sample. CE was performed using an AB3500XL (Life Technologies) according to the PowerPlex® Fusion System protocol, data was analysed using GeneMarker® software v2.4.0 (Softgenetics).

2.3. Massively parallel sequencing

PCR reactions were performed with a prototype PowerSeq™ sequencing assay primer mix and master mix (Promega) amplifying 17 STR loci and Amelogenin. All PCRs were performed on a GeneAmp® PCR System 9700 using the following program: 96 °C for 1 min, 30 cycles of 94 °C for 10 s, 59 °C for 1 min, 72 °C for 30 s and a final extension of 60 °C for 10 min, for every reaction 2800 M Control DNA was included as a positive control and a water sample was included as negative control sample.

Illumina sequencing libraries were prepared from the PCR products by ligating barcoded adapters using the KAPA Library Preparation kit (KAPA Biosystems) without additional amplification using 2.5 μl of PCR product directly in the end repair reaction (without prior purification) in a total volume of 35 μl. The A-tailing and ligation step were performed in a total volume of 25 μl. For ligation, a 10-fold dilution of a barcoded TruSeq adapter (Illumina) was used. To confirm successful ligation of the adapters, 1 μl of library was analysed on the Qiaxcel (Qiagen) for a selection of libraries. To enable balanced pooling, sequencing libraries were quantified in duplicate by real time PCR using the KAPA SYBR® FAST qPCR kit. Quantification reactions were performed on a LightCycler® 480 (Roche) or a 7500 Real Time PCR System (Life Technologies) using a dilution series of PhiX control library (Illumina) as standard. After pooling the libraries, the final pool was quantified again using the same method to enable optimal loading of the flow cell. Sequencing was performed on the MiSeq® sequencer (Illumina) using v3 sequencing reagents according to the manufacturer's protocol with approximately 5% of PhiX control library and 14–19 pM final library concentration.

2.4. Data analysis

For the analysis of STR sequences, the use of simple alignment-based methods could lead to errors. In the analysis pipeline, the first step is the alignment of both paired-end reads that are generated by the sequencer to obtain one high quality consensus read. We used the paired-end read aligner FLASH [11] that aims for a maximum overlap of both reads when creating one consensus read (matching any two paired reads with a mismatch ratio of under 0.33 in the overlapping part). If both reads end within a repeated element, the alignment could lead to a shortened repeated element in the consensus read. To be able to recognise possible misalignment of the reads we altered FLASH version 1.2.11 (this altered version is available via https://github.com/Jerrytha-fast/FLASH-lowercase-overhang). We added an option to mark the bases that were not overlapped by both reads in small letters in the consensus read. Hereby, when all the bases of the flanking regions are in small letters (and thus the sequence reads ended within the repeated element), they can be filtered out in later analysis. When a difference occurred between the two reads, the base call with the highest quality value was used for the consensus. Analysis of the paired-end consensus reads was performed using TSSV [2] (install using: pip install tssv). A TSSV library was created based on all observed variants (Supl. File 1). In Fig. 1, the analysis of STRs using TSSV is illustrated. To further support the interpretation of STR sequencing data, we developed Stuttermark (part of the Python package fdstools, for installation use: pip install fdstools); a Python script that marks possible stutter alleles based on the sequence structure. With this software a column is added to the table of 'known alleles' from TSSV where alleles that could be derived from an *n-1*, *n-2* or *n+1* stutter of an allele (based on the complete allele sequence) in the sample are marked. Thresholds for *n-1* and *n+1* stutter ratios (*n-2* is considered as an *n-1*-1 using a squared value of the *n-1* threshold) are used to decide whether a sequence is marked as an allele or as a possible stutter. A shell script (available upon request) was written to automate all the analysis steps in parallel on a computer cluster for large sample series. An Excel sheet (available upon request) was subsequently used to summarise the results and score variants according to *a priori* defined criteria for the number of reads per variant (total and per orientation) and a minimum percentage from the reads of the highest allele for every locus.

2.5. Analysis of single source samples

In every sequencing run for the population samples (7 runs in total), a maximum of 48 barcoded samples were sequenced aiming for a coverage of at least 1000 reads for every STR allele in each

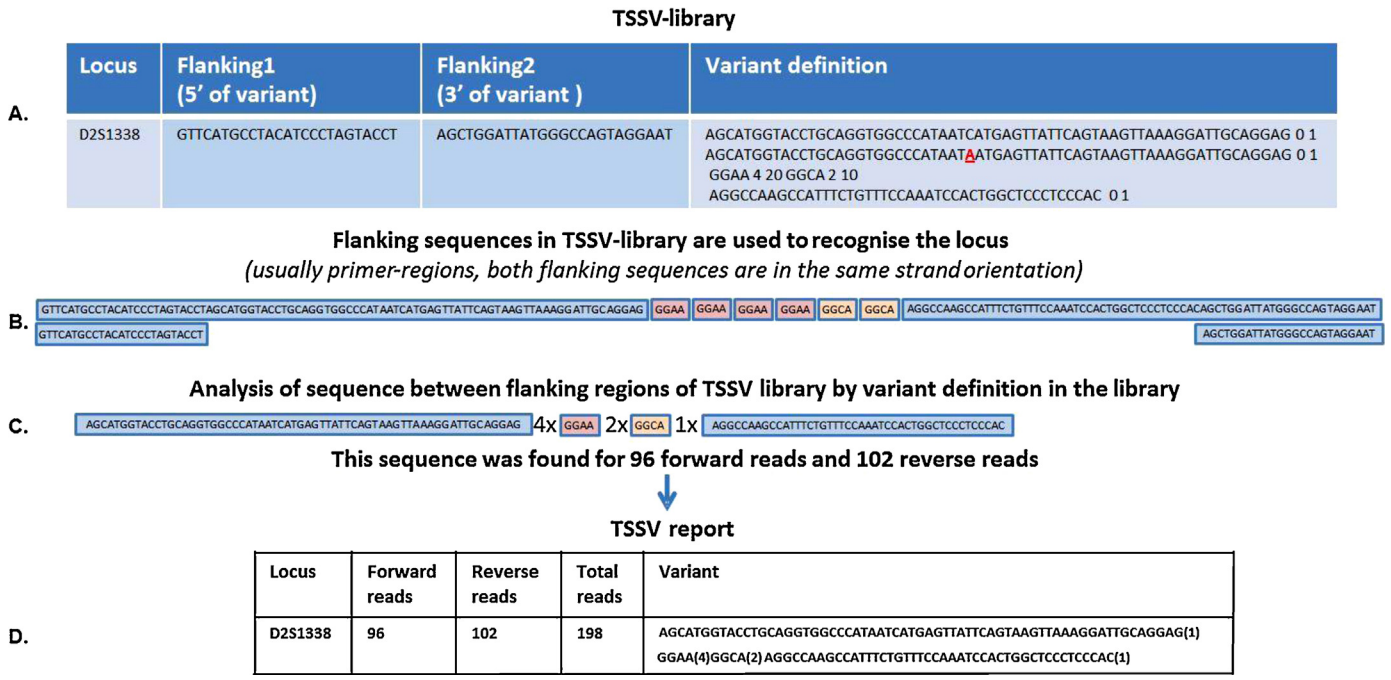


Fig. 1. An overview of the TSSV analysis strategy of short tandem repeat sequences.
(A) An example of the TSSV library entry for locus D2S1338 with from left to right the locus name, flanking 1, flanking 2 (in the same orientation as flanking 1) and the variant definition. Both flanking sequences usually represent the PCR primers. The numbers at the ends of the variant definition sequences (in this example "0 1", "0 1", "4 20", "2 10", and "0 1") indicate how often (based on current knowledge) a sequence could be repeated. (B) Both flanking sequences of the library are used to recognise which locus (in both orientations) any read represents. The observed sequence variation between the two flanking sequences will be reported by TSSV. In this example, some of the surrounding sequence of the STR is included to not only report the STR variation, but also the sequence variation in the surrounding region of the STR. (C) The sequence between the flanking regions is compared to the variant definition of the library. A sequence that complies with the variant definition is reported and summarised (by counting the separate repeated motifs) in the 'known alleles' table and a sequence that doesn't comply with the variant definition is reported in the 'new alleles' table. (D) A TSSV report summarising the displayed allele which was observed 96 times in the forward orientation and 102 times in the reverse orientation. The variant starts with AGCATGG... (not repeated), followed by GGAA (repeated 4 times), GGCA (repeated 2 times) and AGGCCAA... (not repeated). In addition to the tables, fasta files are generated containing the complete sequence reads for the known and new alleles at each locus, but also for the reads that are not recognised or in which only one of the flanking sequences of a locus is recognised. In this way, it is possible to keep track of the sequences that are not reported.

sample. After measuring concentrations of the sequencing libraries, all samples of a run were pooled in an equimolar fashion prior to sequencing. The output of TSSV was analysed with Stuttermark using two different threshold settings; first, *n-1* position stutters with ratios below 10% of the genuine allele and *n+1* position stutters below 2% of the genuine allele were marked while in the second analysis thresholds of respectively 20% and 3% were used. As a final step, a sequence read profile (see Fig. 2) was generated showing all the alleles that have met defined thresholds for read coverage (further described in the Results section). In the sequence read profile, allele names for alleles marked as stutter for both settings of Stuttermark are automatically removed. As with CE analysis, remaining alleles with an assigned allele name were inspected by a trained expert and alleles interpreted as stutter were removed. In this article, allele names are described according to the nomenclature described by van der Gaag and de Knijff [17]. In all figures, locus coordinates were removed to shorten the allele name.

2.6. Analysis of mixed samples

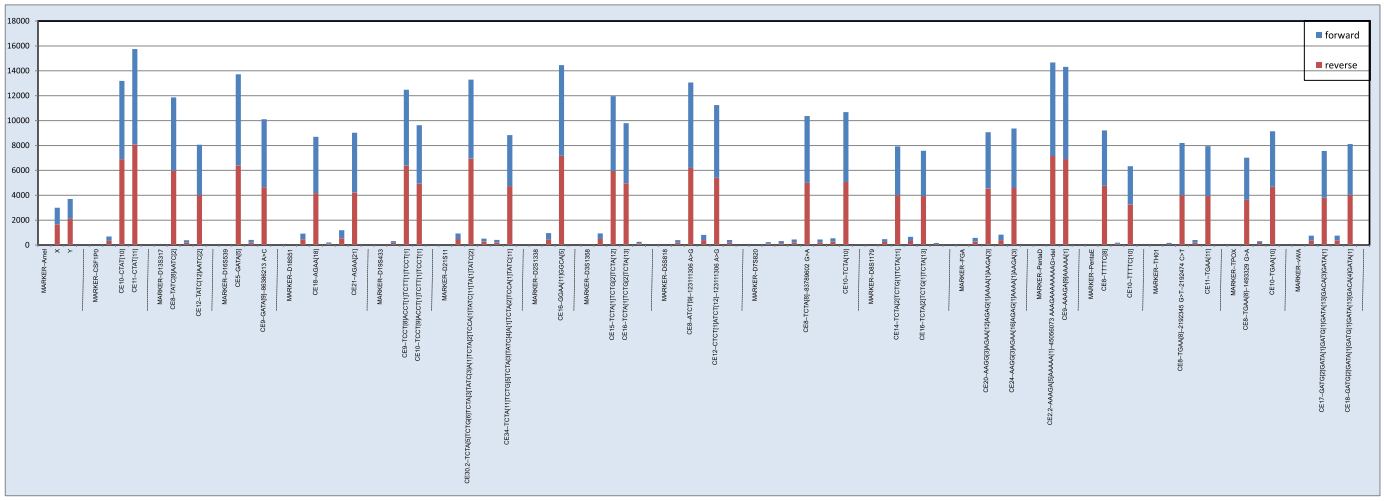
For five two-person combinations selected from the Dutch population samples, mixtures were prepared in the ratios 1:99, 5:95, 10:90, 20:80, 50:50, 80:20, 90:10, 95:5 and 99:1 by mixing the samples based on triplicate DNA quantifications acquired using the Quantifiler® Duo DNA Quantification Kit (Life Technologies).

In the PCR reaction for STR amplification, DNA input amounts were adjusted to add at least 60 pg (≈10 cells) of the minor contributor. To achieve this, total DNA input varied from 0.5 to 6 ng. Samples were sequenced in two runs and pooling ratios were calculated to achieve a minimum of 20 reads for every allele of the minor contributor in each mixture. Analysis was performed in the same way as for the single source samples, but the threshold for the percentage of reads from the highest allele of a marker was lowered depending on the mixture ratio (further discussed in results and discussion). Based on the sequence variation and allele ratios, suspected stutter peaks were marked by an expert to distinguish genuine alleles from stutter peaks.

2.7. Analysis of stutter ratios

2.7.1. Stutter analysis of CE data

The sized output trace data (containing fluorescence intensity data for every position in the electropherogram) was exported from GeneMarker® to Excel. Using peak heights, the stutter ratios at *n-1*, *n+1* and *n-2* stutter positions, were determined for every allele. Peaks that may represent overlapping stutter events (e.g. stutters in between two genuine alleles that may represent both an n-1 and an n+1 stutter) were removed. Supl. Fig. 1 illustrates which combinations of stutter peaks and alleles were used for analysis. Peaks with intensities below 30 rfu were discarded in order to avoid miscalled CE artefacts and to minimise the influence of run-to-run variation of the Genetic Analyser. For some loci, a large proportion of the peaks on stutter positions were lower than 30 rfu (because of the low stutter ratio and the limit in detection range), these peaks did not necessarily represent a zero stutter ratio and were therefore considered to miss a stutter value to avoid underestimation of the stutter ratio (resulting in a slight overestimation of low ratio stutter peaks).



B. Sample read statistics

Read-category	Read-counts	Proportion of total reads
Total passed filter reads	537665	100,0%
Matched pairs	510409	94,9%
Known alleles (including stutters)	406437	75,6%
Genuine alleles (excluding stutter)	350294	65,2%
Reads with errors in the variant region (new alleles in TSSV analysis) *(Singletons)*	103972 *(27973)*	19,3% *5,2%*
Reads representing stutters	56143	10,4%
Primer dimers	27256	5,1%

Fig. 2. An example of a PowerSeq™ MPS read profile and read statistics for all 18 loci in a single-source sample.
(A) An MPS-STR sequence read profile showing all observed alleles of a single-source reference sample with the corresponding number of forward reads (blue bars) and reverse reads (red bars) for every allele. Only the observed variants with coverage of at least 5 reads and a within locus proportion of 2% of the highest allele are displayed in this profile. (B) Read statistics of the displayed sample, all percentages are displayed as a proportion of the total passed filter reads. 94.9% of the reads of this sample were recognised for both flanking sequences (matched pairs) of a locus using TSSV. 75.6% of the total reads represented known alleles and after removing the stutter reads, 65.2% of the reads represent the genuine alleles of this sample. From the 19.3% of matched pairs that were marked as new alleles by TSSV, a large proportion (5.2% of the total reads) consisted of singletons. The remaining 5.1% of passed filter reads (not recognised as matched pairs) represented primer dimers.

2.7.2. Stutter analysis of STR sequencing data

Stutter analysis was performed for all samples for which we obtained more than 50.000 total reads (271 out of 297 samples) to avoid bias introduced by low coverage alleles. To check for possible differences in coverage between long and short alleles, the within locus allele balance was calculated for every marker. For the stutter analysis, sequence variants with coverage below 5 reads were discarded to minimise bias in the stutter ratio. For every observed sequence allele, a table was generated with 6 possible stutter sequences; the two most likely stutter sequences for the *n-1* stutter reads, the *n+1* stutter reads and the *n-2* stutter reads. The most likely stutter sequences were determined based on the length of the longest repeating element in the sequence assuming that longer repeats produce the most stutter [3]. For these 6 stutter alleles, the stutter percentage was determined by dividing the read count of the stutter allele by the read count of the genuine allele. Stutter alleles that could overlap with other alleles or stutter reads were removed taking sequence-specific differences into account as illustrated in Supl. Fig. S1.

2.8. Statistical calculations

For all STRs in the assay, the match likelihood and power of exclusion were calculated for the alleles observed in CE and MPS for all three populations using the Powerstat excel spreadsheet [13].

3. Results and discussion

To assess sequence variation in STR loci and stutter characteristics of a prototype MPS STR sequencing assay (PowerSeq™), 297 samples from three globally dispersed populations were sequenced. To avoid the influence of possible somatic cell line mutations on the analysis of stutter characteristics, we preferred to use DNA samples derived from blood over the use of cell line material from worldwide panels like HapMap or the Human Genome Diversity Panel [1,5].

In the PowerSeq™ assay, all PCRs are designed to amplify STR fragments which are around the same fragment length (shortest to longest allele: 180–310 bp, 180–280 bp excluding the exceptionally long FGA-alleles). Fig. 3 displays the fragment length distribution of the sequenced alleles in this study for all 17 STRs and Amelogenin.

3.1. Optimisation

Reliable quantification of the sequence libraries is an important step for optimal sequencing. It is used to achieve optimal balance

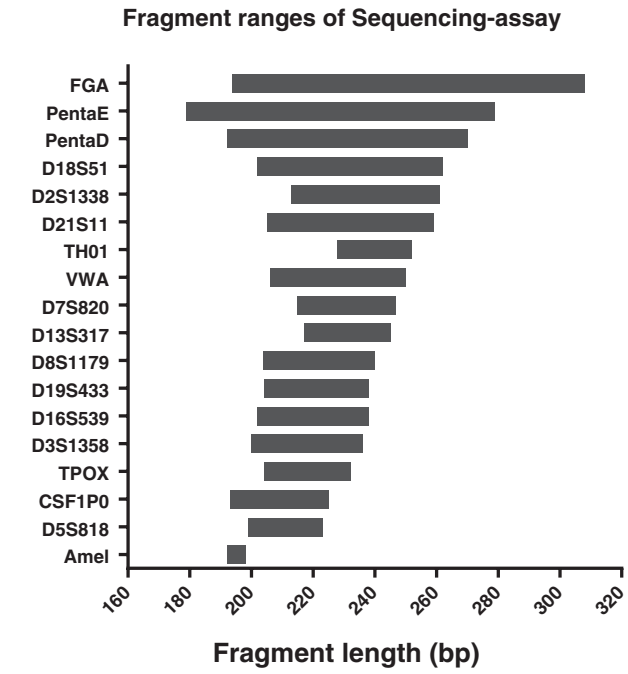


Fig. 3. Overview of fragment range for all loci in the prototype PowerSeq™ assay. The prototype MPS PowerSeq™ multiplex assay used in this study contains 17 autosomal STR loci and Amelogenin. This figure shows the PCR fragment size variation of all alleles sequenced in this study.

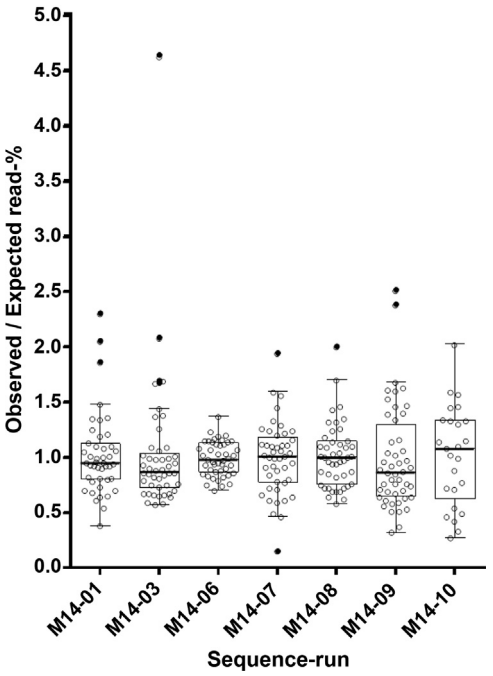


Fig. 4. Tukey boxplot of the ratio of observed versus expected read proportion of pooled samples over different sequencing runs.
Tukey boxplot showing the ratio of observed versus expected read proportion of 297 pooled samples analysed in 7 sequencing runs. The box displays the interquartile range (IQR), the line in the box displays the median and the whiskers display the range until the last sample within 1.5 IQR.

for pooling different libraries in a run and it influences the number of molecules that are loaded on the sequencer. To assess whether equimolar pooling was achieved, the observed and expected proportion of sequences were compared for all samples in the 7 sequencing runs comprising the 297 population samples (Fig. 4). The majority of libraries are represented in 0.5–2 times the expected proportion of reads in the sequencing run, which is sufficiently balanced for the current design. Thus, the quantitation method that was used (real time PCR) allows effective library pooling. Different loading concentrations were used on the MiSeq® sequencer to determine optimal cluster density on the flow cell (higher loading concentrations result in higher cluster densities). Higher cluster density results in a higher amount of unfiltered reads but decreases sequence quality (Supl. Fig. S2). We infer that a flow cell cluster density around 800–1000 K/mm^2 may be most optimal (further discussed in the section 'filtering noise from alleles').

An example of a read profile is shown in Fig. 2A. The sequence profile resembles a CE profile with the y-axis displaying the number of reads observed for every sequence variant, the labels on the x-axis display a more detailed description of the sequence for every allele. Note that the range of amplicon sizes is similar for all STRs (Fig. 3) even though the loci are displayed next to each other on the x-axis. The number of reads is directly proportional to the number of actual molecules for every allele, which is distinct from CE profiles where peak height is influenced by the intensity of emission for different fluorescent labels.

3.2. Sequence efficiency

In Fig. 2, we display the statistics of read counts and the sequencing profile for a typical sample which is prepared using the recommended input of 0.5 ng DNA in the PCR reaction for this assay. 65% of the reads represented the genuine allele sequences of the alleles, approximately 5% of the reads were occupied by stutter reads, the remaining 25% of recognised reads consisted of reads containing PCR and/or sequencing errors. The 5% of unrecognised reads consisted mostly of primer-dimers which is a well-known side effect when large multiplexes such as this 18-plex are used. Remaining primer-dimers could be minimised by purification steps involving size selection such as using a low bead-to-volume ratio for AMPure XP beads. However, we chose to use the PCR product without purification before the library preparation and we used a 2:1 bead ratio in the purification steps of the library preparation to avoid size selection which may affect the balance in sequence reads between longer and shorter STR alleles.

3.3. Filtering noise from alleles

In order to be accepted as a reliable forensic diagnostic tool, MPS results should be retrieved and stored in much more detail compared to CE data. Processing of millions of reads involves complex bioinformatics. It is for this purpose that the tools we developed to analyse MPS reads not only report genuine alleles but also facilitate storing and screening those reads that do not represent genuine STR alleles. Detailed tables of read statistics are produced and checked before allele interpretation. These tables contain read counts for new alleles and for alleles that are only recognised for either one or none of the flanking sequences of the TSSV library. In case of high read numbers for these categories, fasta files containing the complete sequences of the reads can be checked for every locus and for each category (known alleles, new alleles, reads with only the start flanking sequence recognised, reads with only the end-flanking sequence recognised and reads with no recognised flanking sequences at all) separately.

The frequency of sequencing errors varies per locus, but is also strongly influenced by the cluster density in the sequence run. A good indicator for sequence quality of a sequencing run is the balance between forward and reverse reads. Since read errors tend to be influenced by sequence content, the same error will usually not appear in both orientations [16]. For the longest alleles from

PentaD, PentaE and FGA we noted that sequencing errors may accumulate in the end of the reads. As a consequence, the flanking sequences for that strand may no longer be recognised by TSSV, which could lead to strand bias of over five-fold differences between both orientations, even when analysing paired-end consensus reads. Thus, one should not straightforwardly aim for a high cluster density to retain the highest number of reads, as this may be accompanied with strong strand bias. We observed increased rates of sequence errors and strand bias for cluster densities over 1000 K/mm^2 which is below the recommended cluster density of 1200–1500 K/mm^2. When a cluster density of 800 K/mm^2 is used, at least 1.5×10^7 Passed Filter reads are retained (all sequenced for both read orientations) which is a sufficient read number to multiplex an effective number of libraries.

Quality filtering of the data was done in the following order:

1. Paired-end consensus alignment: the two paired-end reads of each cluster are combined. In case of discrepancies, the highest quality base call is used in the consensus read for further analysis. Parts of the read that are not overlapped by both reads are marked in lower case (reads that have one of the library flanking sequences completely represented by lower case letters are later on moved to the TSSV category of reads recognised for only one of the flanking sequences).
2. Singletons are discarded during analysis using TSSV (TSSV option: '-a 2'). These reads can only be checked afterwards by restarting the analysis without this option. Discarding singletons significantly decreases the report file size and memory demand in the follow up analyses. Singletons will not meet forensic standards, but could be used to decide whether sequence coverage needs to be increased for a low coverage sample. New alleles (that do not match the variant description of the TSSV library) are reported in a separate table.
3. After performing TSSV analysis, the table of known alleles is filtered by a priori defined criteria in an Excel sheet while ensuring that the sequences, which are filtered out in these steps, can easily be retrieved and investigated. We used a minimum of 8 reads as allele coverage and a minimum of 2 reads for both sequence orientations which removed the majority of sequencing errors. These numbers may seem low, but it should be noted that we use 'allele coverage' (only including reads without errors) and not 'total coverage' (which would mean the sum of all reads for one locus and could include reads with errors). Since forensic samples often carry allele imbalance due to low amounts of template or multiple contributors to a sample, the use of total summed coverage of all alleles for a target can give a misleading sense of quality and should be avoided. The threshold of 2 reads for both sequence orientations is sufficient to remove the majority of sequence artefacts. A higher threshold could result in the loss of some (mostly longer) alleles that exhibit a strong strand-bias due to structural sequence errors. Retained alleles were interpreted before being reported.
4. In the same Excel sheet an additional criterion is a within-locus proportion (the read count of an allele divided by the read count of the highest allele of a locus) that is required for reporting an allele. This threshold is used to remove PCR errors and structural sequencing errors that may especially occur at high coverage. This value can be adjusted depending on the required detection of low percentage contributions. When the input amount of DNA in the PCR is available, it can also be used to filter out unrealistic mixture contributions (for example: for a start amount of 60 pg in the PCR it is not realistic to look for a 1% contribution since this would represent the DNA equivalent of only 0.1 cell). For single reference samples we used a threshold of 15%, for mixtures, this threshold was lowered to 1% except for mixtures with a minor contribution of 1% in which this threshold was lowered to 0.25%. All retained alleles appear as a bar in the STR sequence profile (Fig. 2).
5. Allele variants that are not represented in the TSSV library are added to the table of new alleles. This table is filtered using the same settings as used for the known alleles. When a new allele is identified as a genuine allele, it is added to the TSSV library and samples are reprocessed using the new TSSV library which will move it to the known alleles category.
6. Stuttermark is used to mark alleles that could be (partly) derived from stutter (as described in the Materials and methods

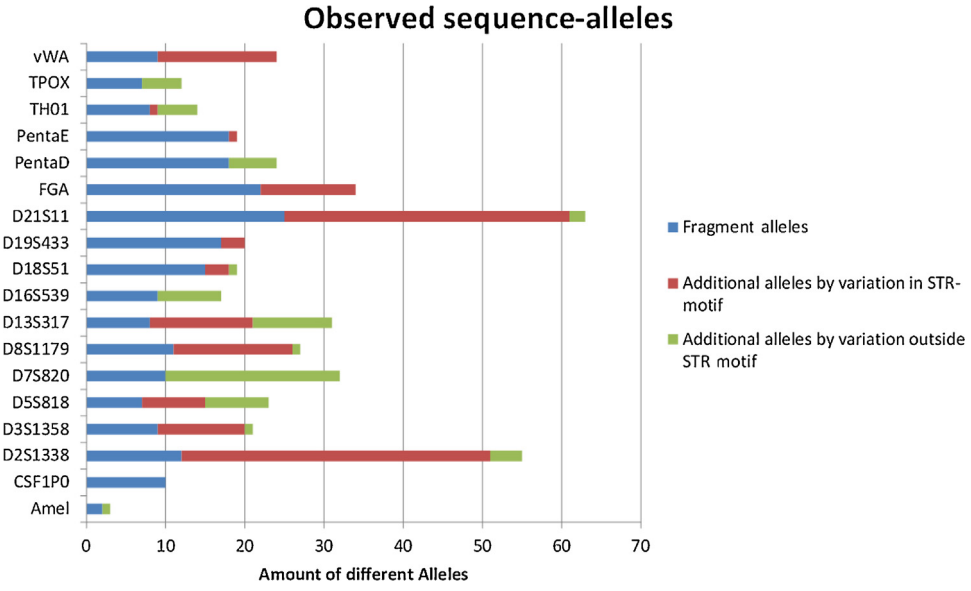


Fig. 5. STR sequence variation divided in length variation, complex STR variation and SNP variation.
The stacked bar graph displays the number of different alleles observed in sequence analysis of 297 samples divided in three categories: In blue, the number of alleles observed when performing CE. In red, the additional alleles observed by sequencing when taking into account variation within the STR motif. In green, the additional alleles when taking into account variation flanking the STR motif. When the variation flanking the STR motif is linked with variation inside the STR motif, the green portion of the bar graph doesn't display those alleles (they are included in the red portion of the bar graph).

section). When interpreting the alleles that pass the filtering steps mentioned before, alleles at a stutter position of another allele (based on the sequence) and with less reads than an *a priori* defined percentage of the reads of a genuine allele are marked as stutter.

7. Interpretation of the retained alleles is done by inspection of the markings from Stuttermark in combination with the ratio between the retained alleles and the strand balance for every allele. In this step, the label of the alleles that are marked as stutter (or any other artefact) will be removed from the STR sequence profile. However, in the sequence profile, the bar representing the removed allele will remain without a label as is common practice for CE-based profiles.

Supl. Fig. 3 shows examples of STR sequencing profiles for a single and a mixed source sample after different filtering settings to illustrate the effect of the used parameters.

3.4. Concordancy

Reliability of sequencing results was assessed for the 297 population samples by comparison of CE data from the PowerPlex® Fusion System with the sequencing data. All STR alleles from the sequencing data were in concordance with CE analysis except for two alleles from PentaD. These alleles were missed when using the 15% within locus threshold (heterozygote balance), as they had a frequency of 8% and 12% of the highest allele (Supl. Fig. S4). Since both samples are from the same population, and both alleles have the same repeat length and sequence, it is likely that this difference in read numbers is caused by a SNP under the PCR primer used in the PowerSeq™ sequencing assay as observed for rare null alleles in commercial CE-based assays [20].

3.5. Sequence variation

As was expected, MPS STR genotyping revealed substantial genetic variation in addition to the variation in repeat length that is detected using CE (Fig. 5). Supl. Fig. S5 displays the sequence of the genome reference (GRCh37/hg19) and of control sample 2800 M (which is provided with the assay). Supl. Fig. S6 displays the observed alleles for all loci and the frequencies of these alleles in the three tested populations. Since we describe our variants according to nomenclature rules [17] in which all variants are described in the forward orientation of the genome reference, the start position and orientation of some of the alleles is slightly different than the reference alleles described by Gettings et al. [7]. Based on the observed variation in this study, the analysed STRs can be divided into four classes.

1. Simple STRs: Loci that only show variation in the number of repeats without additional sequence variation. CSF1P0 is the only simple STR locus.
2. Complex STRs: Loci where the repeat motif consists of several repeating blocks with a different sequence. D19S433, FGA and PentaE are complex STRs.
3. Simple STRs with SNPs in the flanking sequence of the repeat region. D7S820, D16S539, TPOX and PentaD are simple STRs with SNPs.
4. Complex STRs containing SNPs in the flanking sequence of the repeat region: D2S1338, D3S1358, D5S818, D8S1179, D13S317, D18S51, D21S11, TH01 and vWA (interestingly, for vWA, all SNPs are associated with specific repeat region variation) are complex STRs containing SNPs in the flanking sequence.

Using CE, uniquely identified alleles comprise only 48% of the total alleles observed using sequencing in these 17 STRs for the

Table 1
Locus statistics for CE and MPS analysis of the same samples from three populations.

	Total Alleles		Heterozygous%		Match Likelihood						Power of Exclusion					
					Netherlands		Nepal + Buthan		Biaka Pygmees		Netherlands		Nepal + Buthan		Biaka Pygmees	
Marker	Fragment length	Sequence variation	Fragment length	Sequence variation	Fragment length	Sequence variation	Fragment length	Sequence variation	Fragment length	Sequence variation	Fragment length	Sequence variation	Fragment length	Sequence variation	Fragment length	Sequence variation
CSF1P0	10	10	75.8%	75.8%	0.12	0.12	0.12	0.12	0.15	0.15	0.57	0.57	0.45	0.45	0.56	0.56
D2S1338	12	55	83.6%	90.6%	0.04	0.03	0.04	0.03	0.04	0.01	0.82	0.82	0.57	0.69	0.61	0.92
D3S1358	9	21	73.8%	87.2%	0.07	0.04	0.11	0.06	0.14	0.05	0.48	0.66	0.50	0.75	0.49	0.81
D5S818	7	23	72.5%	84.9%	0.16	0.04	0.09	0.05	0.13	0.02	0.54	0.80	0.43	0.51	0.44	0.77
D7S820	10	32	81.9%	87.9%	0.07	0.03	0.08	0.03	0.09	0.03	0.64	0.70	0.48	0.63	0.79	0.94
D8S1179	11	27	80.2%	86.9%	0.07	0.04	0.06	0.03	0.09	0.03	0.61	0.76	0.65	0.75	0.56	0.69
D13S317	8	31	72.5%	85.6%	0.09	0.03	0.07	0.03	0.17	0.05	0.57	0.74	0.55	0.69	0.31	0.69
D16S539	9	17	75.5%	81.2%	0.10	0.07	0.09	0.05	0.08	0.03	0.48	0.55	0.50	0.55	0.58	0.77
D18S51	15	19	83.9%	85.2%	0.04	0.04	0.04	0.04	0.05	0.04	0.70	0.72	0.69	0.69	0.63	0.69
D19S433	17	20	83.6%	83.6%	0.09	0.08	0.06	0.06	0.03	0.03	0.57	0.57	0.67	0.67	0.77	0.77
D21S11	25	63	84.2%	88.3%	0.05	0.03	0.06	0.02	0.04	0.02	0.70	0.78	0.69	0.79	0.65	0.71
FGA	23	35	86.2%	86.2%	0.05	0.05	0.03	0.03	0.04	0.04	0.82	0.82	0.75	0.75	0.60	0.60
PentaD	18	24	83.2%	84.2%	0.06	0.05	0.06	0.06	0.04	0.04	0.62	0.66	0.57	0.59	0.79	0.79
PentaE	18	19	85.2%	85.2%	0.03	0.03	0.03	0.03	0.04	0.04	0.66	0.66	0.75	0.75	0.69	0.69
TH01	8	14	70.8%	72.5%	0.10	0.10	0.16	0.16	0.13	0.08	0.61	0.61	0.33	0.33	0.41	0.49
TPOX	7	12	65.4%	71.1%	0.20	0.20	0.21	0.19	0.13	0.05	0.38	0.38	0.31	0.33	0.39	0.67
vWA	9	24	78.5%	83.6%	0.07	0.05	0.08	0.07	0.06	0.02	0.62	0.70	0.46	0.50	0.63	0.81
Amel	2	3														
				Overall	5.4E-20	1.0E-22	3.0E-20	8.0E-23	4.7E-20	2.5E-25						

Heterozygosity, Match Likelihood and Power of Exclusion for STR CE and sequence analysis for all 17 STRs as observed in the three tested populations.

analysed set of 297 samples. However, the variation is not evenly dispersed over the loci (Table 1). Since not every available software tool for analysis of STRs capture the variation within the repeat structure and the flanking sequence [18] it is important to be aware of the information that is missed when variation outside the repeat structure is not reported. Obviously, the discriminating power of the loci is increased when all the variation on sequence level is taken into account. In Table 1 we display the match likelihood (ML) for every locus in all three populations for sequence analysis and for CE analysis in comparison. The additional sequence variation has the strongest effect on the discriminating power of D5S818 and D13S317 with an average three-fold difference in the ML over all populations between the two methods. D2S1338, D3S1358, D7S820, D8S1179, D16S539 and D21S11 exhibit more than a two-fold difference in the ML over all populations. When only taking into account the Dutch and Himalayan population, D5S818, D7S820, D13S317 and D21S11 still exhibit a greater than two-fold difference in match likelihood between length and sequence variation.

3.6. Stutter analysis

Stutter ratios were determined when the CE signal intensity or MPS read coverage was sufficient for alleles which are not influenced by stutters from other alleles. An overview of the read coverage statistics and within locus allele balance of the samples used for this analysis is shown in Suppl. Fig. 7. For each locus, dot plots were generated displaying the average stutter ratios for all STR alleles for which at least four stutter ratios could be calculated (Supl. Fig. 8). In general, stutter ratios of both methods are very similar with the exception of PentaE where stutter ratios for CE are lower than for sequence data. Some sequence alleles correspond to the same CE allele (e.g. D2S1338 allele 21). For complex STRs, the longest uninterrupted repeat stretch determines the stutter ratio [19] which is confirmed by our data as illustrated in Fig. 6. Here, detailed stutter graphs for D18S51 are shown for both methods; the dots of the alleles carrying an interrupted repeat motif (marked in red) tend to have lower stutter ratios than the uninterrupted alleles of the same length. Because of the separation of these new sequence alleles it is expected that the stutter ratio per sequence allele would show less variation than the CE stutter ratio which represents several sequence variants. To test this, the Coefficient of Variance of the stutter ratio was determined for every allele with stutter data for at least four samples (Supl. Fig. 8). Most obtained CV values are either similar or lower for sequencing stutter ratios than for CE stutter ratios. As expected, the loci for which the CV of the stutter ratio is generally lower for sequencing data than for CE are all complex STRs (especially D5S818, D8S1179, D13S317, D21S11, FGA and vWA). In addition it was noted that the CV of the stutter ratio for sequence data remains relatively stable for all alleles within the same locus (even though the stutter becomes higher for longer alleles). For CE-based stutter ratios, much more variation in CV is observed between different alleles within the same locus which is partly explained by alleles that are subdivided into different sequence alleles. For some STRs (in particular D2S1338, D3S1358, D7S820 and D18S51) the CV shows a downward trend for increasing allele length in CE data. An explanation for this decreasing CV could be that low percentage stutter peaks in a CE profile are often below the detection threshold (30 rfu in this analysis). Since a certain number (at least several thousand depending on the fluorescent label) of molecules is needed before a CE peak becomes visible, the signal intensity might not be linearly correlated with the number of molecules for alleles with low peak heights. This could contribute to an increased variation of stutter ratios.

3.7. Mixture analysis

A total of 45 two-person mixtures (from five donor combinations) were analysed with minor contributions of 1%, 5%, 10%, 20% and 50% using the PowerSeq™ sequencing assay. In every mixture, all alleles of both contributors were recovered in the sequence reads, mostly with allele ratios close to expected. Fig. 7a displays the read percentage for each allele call of the minor contributor

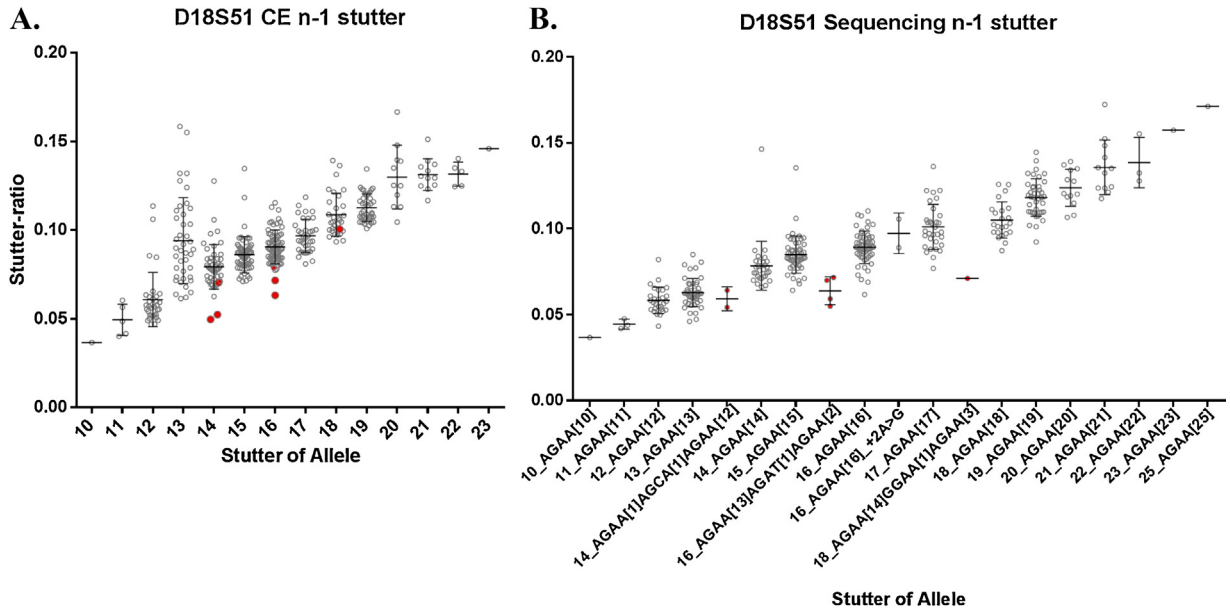
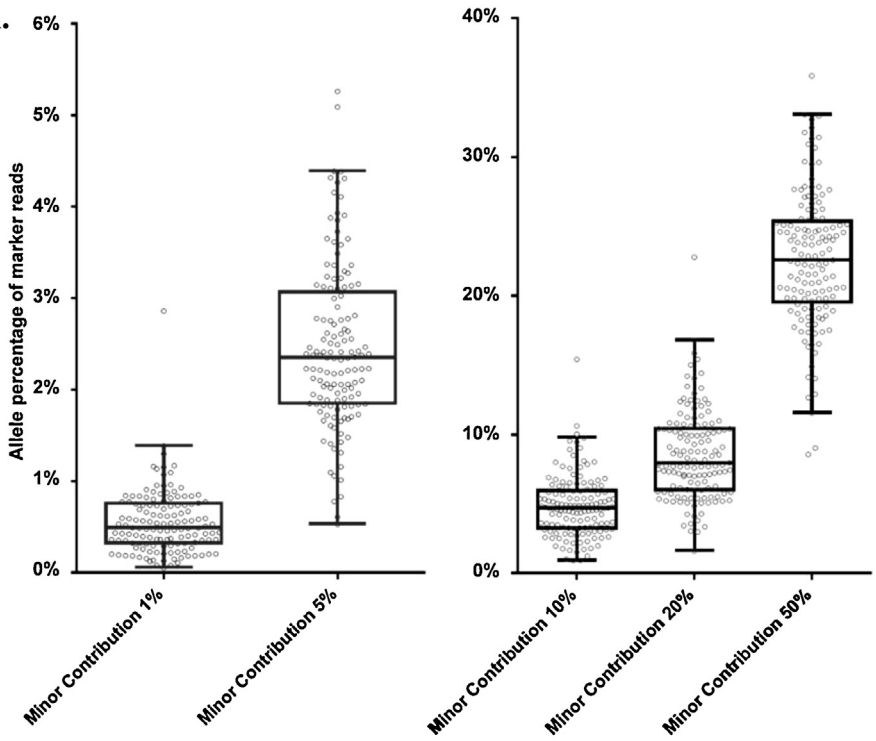


Fig. 6. Comparison of stutter ratios for locus D18S51 analysed by CE and MPS.
(A) Dot plot displaying the distribution of stutter ratios for the locus D18S51 analysed by CE using the PowerPlex® Fusion System. Every dot represents the stutter ratio of one allele in a single sample, lines display the median and whiskers display 1.5 interquartile range. Red dots represent alleles in which the sequence revealed an interrupted repeat (resulting in a shorter length of the longest repeated motif). (B) Dot plot displaying the distribution of stutter ratios for the locus D18S51 analysed by MPS using the prototype PowerSeq™ system. Red dots represent alleles in which the sequence revealed an interrupted repeat. It is apparent that the stutter ratio of the alleles carrying an interrupted repeat motif is generally lower than the alleles of the same length without interruption of the repeat motif.

A.



B.

Minor Contribution	Average calculated minor contribution	StDev	COV
1%	1,1%	0,32%	28,1%
5%	5,09%	0,28%	5,6%
10%	9,97%	0,30%	3,0%
20%	17,28%	0,26%	1,5%
50%	45,33%	0,28%	0,6%

Fig. 7. Tukey boxplot displaying the observed within locus read percentages of all minor alleles for 10 two-person mixtures for each of the five tested mixture ratios. (A) Tukey boxplots showing the within locus read percentages (read-count allele/total read-count of the locus) for all alleles of the minor contributions in two-person mixtures. For each mixture ratio (minor contributions of 1%, 5%, 10%, 20% and 50%), read percentages are displayed for 10 minor contributions. Only the alleles that are not overlapping with another allele of the mixture or with n-1 stutter of another allele are displayed. The box displays the interquartile range (IQR), the line displays the median and the whiskers display the ranges until the last sample within 1.5 IQR. Since the within locus percentages are displayed per allele, one allele of a 50:50 mixture should be represented by 25% of the reads for genuine alleles (25% for the other allele and again 25% for each allele of the other contributor). (B) Summary statistics of the average calculated minor contribution (calculated by averaging the read percentages of all alleles of the minor contribution in each mixture) for 10 minor contributors of each mixture ratio.

grouped by mixture ratio. Although there is variation, we found that the observed percentage of reads (per allele) from the total locus reads is a good indication of the ratio between two contributors in a mixture. For each of the 45 mixtures the minor contribution was estimated based on the read frequencies of the minor alleles that are not overlapping other alleles or stutter reads in the mixture (see Supl. Fig. 9 for further explanation of this procedure for a hypothetical three locus mixture profile). Fig. 7b shows the summary statistics for calculation of the minor contribution in the 10 mixtures (for the 50/50 mixtures, calculations were performed for both contributors) of each ratio. Since the total marker reads also contain reads representing stutter, the quantitative prediction of the minor contribution is expected to be slightly lower than the genuine contribution which is apparent for the mixtures with 50% and 20% minor contribution. Not surprisingly, a quantitative prediction of the minor contribution becomes less accurate (relative to the percentage of contribution) when the minor contribution decreases. It is apparent that the standard deviation is almost stable across all mixture ratios.

When analysing alleles with abundance below 5% of the highest allele of the locus, additional PCR/sequence error variants were observed for several loci which can complicate the interpretation of a DNA sample. Therefore, the analysis of minor contributions of 5% or less in a mixture without prior knowledge of the ratio between the different donors, remains difficult for some, but not all loci, using the current experimental and analysis setup for this assay. Increasing the sequencing coverage increases the read counts of these artefacts as well and will not help to distinguish them from genuine alleles.

3.8. Analysing an unknown trace

When unknown samples are analysed that could have more than one contributor, one needs to decide on the minimal allele coverage and level of minor allele detection prior to sequencing.

The minimal allele coverage of 8 reads for every allele and 2 reads for both orientations used in this study was chosen for investigative purposes to get an indication of general sequence quality. Although in most cases these thresholds were sufficient to remove artefacts, some erroneous reads can still occur due to a relatively low sequence quality that may be caused by variation in cluster density or other factors yet unknown. In addition to a minimal read coverage to guarantee sequence quality, an additional threshold can be used for the minimal percentage of reads compared to the allele with the highest read count within a locus to filter out structural sequence errors. Below 0.5%, most STRs show a high amount of additional sequence artefacts that coexist with the genuine alleles at a relatively stable ratio. However, when using a high threshold, low percentage contributions might be missed.

3.9. Recommendations

In this study, the population samples were sequenced with an average allele coverage of over 800 reads (also including the samples that were not used for stutter analysis), which is crucial for a reliable characterisation of stutter reads and structural sequence errors in this stage of the development of this new technique. We assume that, eventually, for reliable MPS-STR genotyping of a single-source reference sample (e.g. for database purposes) a much lower coverage could be sufficient. To distinguish genuine allele sequences from errors, we recommend a coverage of at least 20 reads for every allele (sequences from both ends combined) with representation in both orientations. This means that, for the current assay, 5.000 reads per sample will probably be sufficient to achieve the recommended allele coverage. For evidentiary traces, more sequences will be needed since locus balance will be influenced by low template concentrations and low contributions can only be analysed reliably using sufficient reads for the alleles of the minor contribution. For example, when we want to retain sufficient data to detect a minor contribution of 5% we need at least $(100/5) \times 5000 = 100{,}000$ reads (meaning 100,000 reads for read1 and 100,000 reads for read2) for the current assay. This assumes that the sample is of sufficient quality to retain the same locus balance as a reference sample.

4. Conclusion

The analysis of STRs by MPS using the MiSeq® provides several advantages over the routinely used CE. We observed full concordance between CE (Powerplex® Fusion) and MPS (PowerSeq™) based genotyping of STR loci among 297 individuals.

We observed substantial sequence variation within the repeat motifs of STR loci and their immediate flanking regions, in addition to the length variation of the STR-motifs. Since design of a multiplex assay for MPS is no longer limited by the number of different fluorescent labels, PCR primers can be designed to amplify all STR loci within a much more similar fragment size range. This offers advantages for degraded DNA samples and reduces some of the amplification bias due to length variation among the various PCR-templates in a single multiplex PCR reaction. In addition, the exact nature of MPS data (which is as simple as sequence-specific read counts for every allele) provides opportunities for a more standardised follow-up analysis. The study of stutter in MPS data shows that the highest stutter artefact is determined by the longest repeated element in the STR. STR stutter ratios in MPS data are generally similar to those of CE data except for many of the complex STRs since those CE alleles can be differentiated into separate MPS alleles with their own respective stutter profile. Mixture analysis down to a minor contribution of 5% is routinely feasible for most STR loci. Even sequence reads representing a minor contribution down to 1% can be recovered, although here, obviously, reads representing stutters still cause interpretation problems in the reads.

Acknowledgements

This study was supported by a grant from the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands. The authors wish to thank Titia Sijen (Netherlands Forensic Institute) for carefully reviewing the manuscript and Martin Ensenberger and Cynthia Sprecher (Promega Corporation) for their contributions to the design of the prototype PowerSeq™ System.

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.2016.05.016.

References

[1] D.M. Altshuler, R.A. Gibbs, L. Peltonen, D.M. Altshuler, R.A. Gibbs, L. Peltonen, E. Dermitzakis, S.F. Schaffner, F. Yu, L. Peltonen, E. Dermitzakis, P.E. Bonnen, D.M. Altshuler, R.A. Gibbs, P.I. de Bakker, P. Deloukas, S.B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, F. Yu, K. Chang, A. Hawes, L.R. Lewis, Y. Ren, D. Wheeler, R.A. Gibbs, D.M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J.M. Korn, K. Kristiansson, C. Lee, S.A. McCarrol, J. Nemesh, E. Dermitzakis, A. Keinan, S.B. Montgomery, S. Pollack, A.L. Price, N. Soranzo, P. E. Bonnen, R.A. Gibbs, C. Gonzaga-Jauregui, A. Keinan, A.L. Price, F. Yu, V. Anttila, W. Brodeur, M.J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, S. F. Schaffner, Q. Zhang, M.J. Ghori, R. McGinnis, W. McLaren, S. Pollack, A.L. Price, S.F. Schaffner, F. Takeuchi, S.R. Grossman, I. Shlyakhter, E.B. Hostetter, P.C. Sabeti, C.A. Adebamowo, M.W. Foster, D.R. Gordon, J. Licinio, M.C. Manca, P.A. Marshall, I. Matsuda, D. Ngare, V.O. Wang, D. Reddy, C.N. Rotimi, C.D. Royal, R.R. Sharp, C. Zeng, L.D. Brooks, J.E. McEwen, Integrating common and rare genetic variation in diverse human populations, Nature 467 (2010) 52–58.

[2] S.Y. Anvar, K.J. van der Gaag, J.W. van der Heijden, M.H. Veltrop, R.H. Vossen, R. H. de Leeuw, C. Breukel, H.P. Buermans, J.S. Verbeek, K.P. de, J.T. den Dunnen, J.F. Laros, TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes, Bioinformatics 30 (2014) 1651–1659.

[3] C. Brookes, J.A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Sci. Int. Genet. 6 (2012) 58–63.

[4] B. Budowle, A.J. Onorato, T.F. Callaghan, M.A. Della, A.M. Gross, R.A. Guerrieri, J. C. Luttman, D.L. McClure, Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework, J. Forensic Sci. 54 (2009) 810–821.

[5] H.M. Cann, T.C. de, L. Cazes, M.F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G.B. Ferrara, J.S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R.J. Herrera, X. Huang, J. Kidd, K.K. Kidd, A. Langaney, A.A. Lin, S.Q. Mehdi, P. Parham, A. Piazza, M.P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J.L. Weber, H.T. Greely, M.W. Feldman, G. Thomas, J. Dausset, L.L. Cavalli-Sforza, A human genome diversity cell line panel, Science 296 (2002) 261–262.

[6] C. Gelardi, E. Rockenbauer, S. Dalsgaard, C. Borsting, N. Morling, Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles, Forensic Sci. Int. Genet. 12 (2014) 38–41.

[7] K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR allele sequence variation: current knowledge and future issues, Forensic Sci. Int. Genet. 18 (2015) 118–130.

[8] M.C. Kline, C.R. Hill, A.E. Decker, J.M. Butler, STR sequence analysis for characterizing normal, variant, and null alleles, Forensic Sci. Int. Genet. 5 (2011) 329–332.

[9] P. de Knijff, J. Pijpe, Population genetics of an African Pygmy population, 2015, Ref. Type: Unpublished work.

[10] T. Kraaijenbrink, K.J. van der Gaag, S.B. Zuniga, Y. Xue, D.R. Carvalho-Silva, C. Tyler-Smith, M.A. Jobling, E.J. Parkin, B. Su, H. Shi, C.J. Xiao, W.R. Tang, V.K. Kashyap, R. Trivedi, T. Sitalaximi, J. Banerjee, N.M. Karma Tshering of Gaselo Tuladhar, J.R. Opgenort, G.L. van Driem, G. Barbujani, K.P. de, A linguistically informed autosomal STR survey of human populations residing in the greater Himalayan region, PLoS One 9 (2014) e91534.

[11] T. Magoc, S.L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies, Bioinformatics 27 (2011) 2957–2963.

[12] K. Oostdik, K. Lenz, J. Nye, K. Schelling, D. Yet, S. Bruski, J. Strong, C. Buchanan, J. Sutton, J. Linner, N. Frazier, H. Young, L. Matthies, A. Sage, J. Hahn, R. Wells, N. Williams, M. Price, J. Koehler, M. Staples, K.L. Swango, C. Hill, K. Oyerly, W. Duke, L. Katzilierakis, M.G. Ensenberger, J.M. Bourdeau, C.J. Sprecher, B. Krenke, D.R. Storts, Developmental validation of the PowerPlex((R)) Fusion System for

analysis of casework and reference samples: a 24-locus multiplex for new database standards, Forensic Sci. Int. Genet. 12 (2014) 69–76.
[13] Promega Corporation. Powerstats v12, 2015, Ref. Type: Unpublished work.
[14] Promega Corporation Technical Manual, PowerPlex® Fusion System, 2015, Ref. Type: Unpublished work.
[15] M. Scheible, O. Loreille, R. Just, J. Irwin, Short tandem repeat typing on the 454 platform: strategies and considerations for targeted sequencing of common forensic markers, Forensic Sci. Int. Genet. 12 (2014) 107–119.
[16] M. Schirmer, U.Z. Ijaz, R. D'Amore, N. Hall, W.T. Sloan, C. Quince, Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform, Nucleic Acids Res. 43 (6) (2015) e37.
[17] K.J. van der Gaag, P. de Knijff, Forensic nomenclature for short tandem repeats updated for sequencing, Forensic Sci. Int. Genet. Suppl. Ser. 4 (2015).
[18] D.H. Warshauer, J.L. King, B. Budowle, STRait Razor v2.0: the improved STR Allele identification Tool–Razor, Forensic Sci. Int. Genet. 14 (2015) 182–186.
[19] A.A. Westen, L.J. Grol, J. Harteveld, A.S. Matai, K.P. de, T. Sijen, Assessment of the stochastic threshold, back- and forward stutter filters and low template techniques for NGM, Forensic Sci. Int. Genet. 6 (2012) 708–715.
[20] A.A. Westen, T. Kraaijenbrink, E.A. Robles de Medina, J. Harteveld, P. Willemse, S.B. Zuniga, K.J. van der Gaag, N.E. Weiler, J. Warnaar, M. Kayser, T. Sijen, K.P. de, Comparing six commercial autosomal STR kits in a large Dutch population sample, Forensic Sci. Int. Genet. 10 (2014) 55–63.
[21] X. Zeng, J.L. King, M. Stoljarova, D.H. Warshauer, B.L. LaRue, A. Sajantila, J. Patel, D.R. Storts, B. Budowle, High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing, Forensic Sci. Int. Genet. 16 (2015) 38–47.