Development of forensic genomics research toolkits by the use of Massively Parallel Sequencing

Gaag, K.J. van der

Author: Gaag, K.J. van der
Title: Development of forensic genomics research toolkits by the use of Massively Parallel Sequencing
Issue Date: 2019-11-27

Chapter 4

Massively Parallel Sequencing of Short Tandem Repeats – Population data and mixture analysis results for the PowerSeq™ system

Forensic Science International Genetics 2016 Sep;24:86-96
Published online June 07, 2016

Kristiaan J. van der Gaag
Rick H. de Leeuw
Jerry Hoogenboom
Jaynish Patel
Douglas R. Storts
Jeroen F.J. Laros
Peter de Knijff

Supplementary material is available via https://www.fsigenetics.com

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.

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 80-500 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 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.

As a consequence of various mechanisms such as DNA recombination, replication and repair-associated processes, the spectrum of human genetic variation ranges from single nucleotide differences to large chromosomal events. Among the different types of genetic changes, repetitive DNA sequences show more polymorphism than single nucleotide variants (Conrad et al., 2010; Hinds et al., 2006; Iafrate et al., 2004; Kidd et al., 2008; Redon et al., 2006; Sebat et al., 2004; Tuzun et al., 2005), and they are important in human diseases (Conrad et al., 2010; de Cid et al., 2009; Girirajan et al., 2011; Hollox et al., 2008; McCarroll et al., 2009; Pinto et al., 2010), complex traits and evolution (Mills et al., 2011; Stephens et al., 2011; Sudmant et al., 2010). In particular, microsatellite variants, also known as short tandem repeats (STR), and their expansion/shortening have been linked to a variety of human genetic disorders (Mirkin, 2007; Pearson et al., 2005; Sutherland and Richards, 1995), and have been used in genotyping (Kimura et al., 2009; Weber and May, 1989) and forensic DNA fingerprinting studies (Kayser and de Knijff, 2011; Moretti et al., 2001).

Because of the repetitive nature of STRs and often the low level of complexity of the DNA sequences in which they occur (Treangen and Salzberg, 2012), characterization of STR variability and understanding of their functional consequences are challenging (Weischenfeldt et al., 2013). So far, sequencing-based strategies have focused on reads mapped to the reference genome and subsequent identification of discordant signatures and classification of associated STRs (Medvedev et al., 2009; Mills et al., 2011). Yet, the mainstream aligners, such as BWA (Li and Durbin, 2009) or Bowtie (Langmead and Salzberg, 2012), do not tolerate repeats or insertions and deletions (indels) as a trade-off of run time (Li and Homer, 2010). This limitation leads to ambiguities in the alignment or assembly of repeats which, in turn, can obscure the interpretation of results (Treangen and Salzberg, 2012). Moreover, the current human genome reference still remains incomplete and provides only limited information on expected and potentially uncharacterized STRs in different individuals (Alkan et al., 2011; Iafrate et al., 2004; Kidd et al., 2008; Sebat et al., 2004). Consequently, STRs are not routinely analyzed in whole-genome or whole-exome sequencing studies, despite their obvious applications and their role in human diseases, complex traits and evolution.

Here, we present a method for targeted profiling of STRs that reports a full spectrum of all observed genomic variants along with their respective abundance. Our tool, TSSV, can accurately profile and characterize STRs without the use of a complete reference genome, and therefore minimizes biases introduced during the alignment and downstream analysis. TSSV scans sequencing data for reads that fully or partially encompass loci of interest based on the detection of unique flanking sequences. Subsequently, TSSV characterizes the sequence between a pair of non-repetitive flanking regions and reports statistics on known and novel alleles for each locus of interest. We show the performance of TSSV on robust characterization of all allelic variants in a given targeted locus by its application in several case studies: forensic DNA fingerprinting of mixed samples by STR profiling, characterization of variants introduced by transcription activator-like effector nucleases (TALENs) in embryonic stem (ES) cells and detailed characterization of errors derived from a next-generation sequencing (NGS) experiment.

Material and Methods

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]).

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, 2800M 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).

Massively Parallel Sequencing

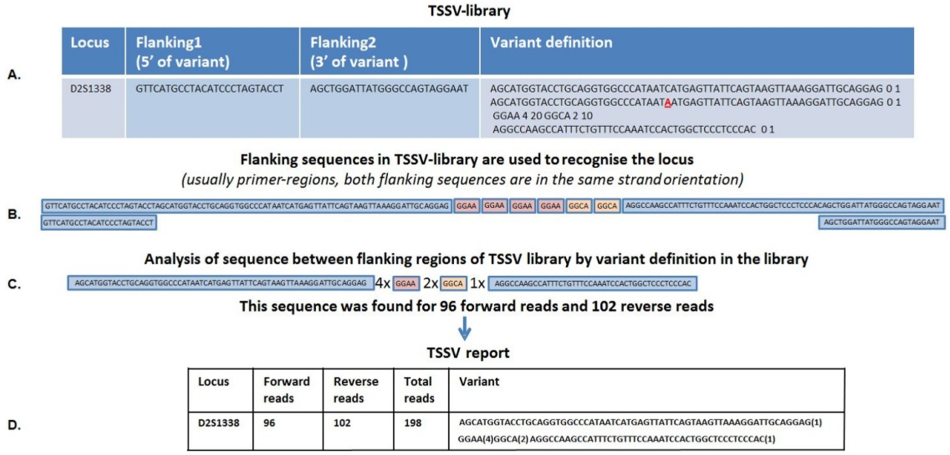
PCR reactions were performed with a prototype PowerSeq™ sequencing assay primer mix and master mix (Promega) amplifying 17 STR loci and Amelogenin. All PCRs were performed on a GeneAmp® PCR System 9700 using the following program: 96 °C for 1 min, 30 cycles of 94 °C for 10s, 59 °C for 1 min, 72 °C for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

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/Jerrythafast/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 (Sup. File 1). In Figure 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.

Figure 1. An overview of the TSSV analysis strategy of short tandem repeat sequences



TSSV-library

A.

Locus	Flanking1 (5' of variant)	Flanking2 (3' of variant)	Variant definition
D2S1338	GTTCATGCCTACATCCCTAGTACCT	AGCTGGATTATGGGCCAGTAGGAAT	AGCATGGTACCTGCAGGTGGCCCATAATCATGAGTTATTCAGTAAGTTAAAGGATTGCAGGAG 0 1 AGCATGGTACCTGCAGGTGGCCCATAATAATGAGTTATTCAGTAAGTTAAAGGATTGCAGGAG 0 1 GGAA 4 20 GGCA 2 10 AGGCCAAGCCATTTCTGTTTCCAAATCCACTGGCTCCCTCCCAC 0 1

Flanking sequences in TSSV-library are used to recognise the locus

(usually primer-regions, both flanking sequences are in the same strand orientation)

B.

GTTCATGCCTACATCCCTAGTACCTAGCATGGTACCTGCAGGTGGCCCATAATCATGAGTTATTCAGTAAGTTAAAGGATTGCAGGAG GGAA GGAA GGAA GGAA GGCA GGCA AGGCCAAGCCATTTCTGTTTCCAAATCCACTGGCTCCCTCCCACAGCTGGATTATGGGCCAGTAGGAAT

GTTCATGCCTACATCCCTAGTACCT AGCTGGATTATGGGCCAGTAGGAAT

Analysis of sequence between flanking regions of TSSV library by variant definition in the library

C.

AGCATGGTACCTGCAGGTGGCCCATAATCATGAGTTATTCAGTAAGTTAAAGGATTGCAGGAG 4x GGAA 2x GGCA 1x AGGCCAAGCCATTTCTGTTTCCAAATCCACTGGCTCCCTCCCAC

This sequence was found for 96 forward reads and 102 reverse reads

TSSV report

D.

Locus	Forward reads	Reverse reads	Total reads	Variant

D2S1338	96	102	198	AGCATGGTACCTGCAGGTGGCCCATAATCATGAGTTATTCAGTAAGTTAAAGGATTGCAGGAG(1)
				GGAA(4)GGCA(2) AGGCCAAGCCATTTCTGTTTCCAAATCCACTGGCTCCCTCCCAC(1)

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.

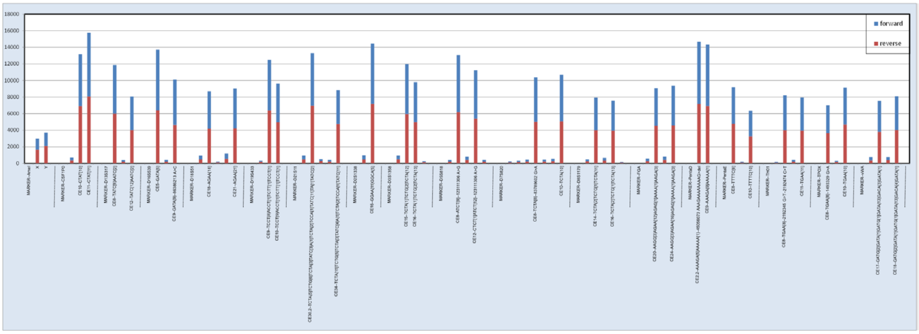
D. A TSSV report summarising the displayed allele which was observed 96 times in the forward orientation and 102 times in the reverse orientation. The variant starts with AGCATGG... (not repeated), followed by GGAA (repeated 4 times), GGCA (repeated 2 times) and AGGCCAA... (not repeated). In addition 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.

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 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 Figure 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.

Figure 2. An example of a PowerSeq^{TM} MPS read profile and read statistics for all 18 loci in a single-source sample

A. An STR sequence read profile



B. Sample read statistics

Read-category	Read-counts	Proportion of total reads
Total passed filter reads	537665	100,0%
Matched pairs Known alleles (including stutters) Genuine alleles (excluding stutter)	510409 406437 350294	94,9% 75,6% 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%

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.

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 ng – 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.

Analysis of stutter ratios

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. Sup. Figure 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).

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 Sup. Figure 1.

Statistical Calculations

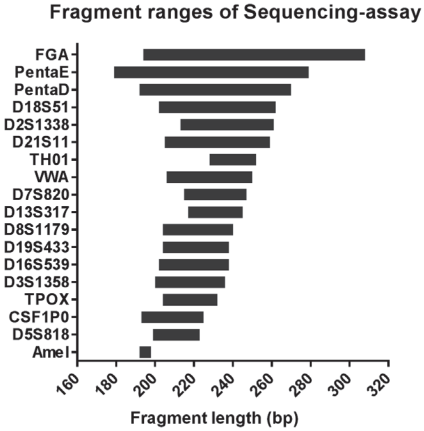
For all STRs in the assay, the match likelihood and power of exclusion were calculated for the alleles observed in CE and MPS for all three populations using the Powerstat excel spreadsheet [13].

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). Figure 3 displays the fragment length distribution of the sequenced alleles in this study for all 17 STRs and Amelogenin.

Figure 3. Overview of fragment range for all loci in the prototype PowerSeq™ assay

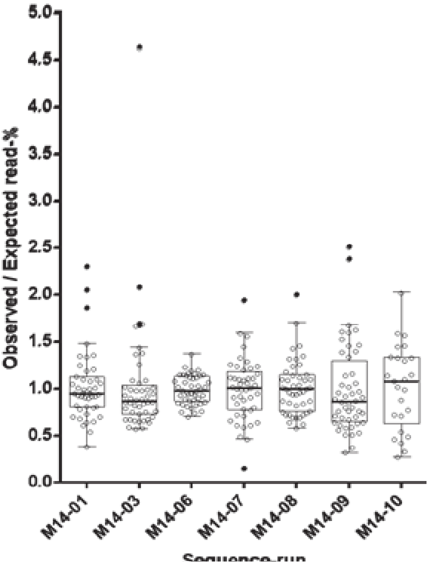


The prototype MPS PowerSeq™ multiplex assay used in this study contains 17 autosomal STR loci and Amelogenin. This figure shows the PCR fragment size variation of all alleles sequenced in this study.

Optimisation

Reliable quantification of the sequence libraries is an important step for optimal sequencing. It is used to achieve optimal balance 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 (Figure 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 (Sup. Figure 2). We infer that a flow cell cluster density around 800-1000 K/mm2 may be most optimal (further discussed in the section 'filtering noise from alleles').

Figure 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.

An example of a read profile is shown in Figure 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 (Figure 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.

Sequence efficiency

In Figure 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.

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 1.5x10^7 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 (Figure 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 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.

Sup. Figure 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.

Concordancy

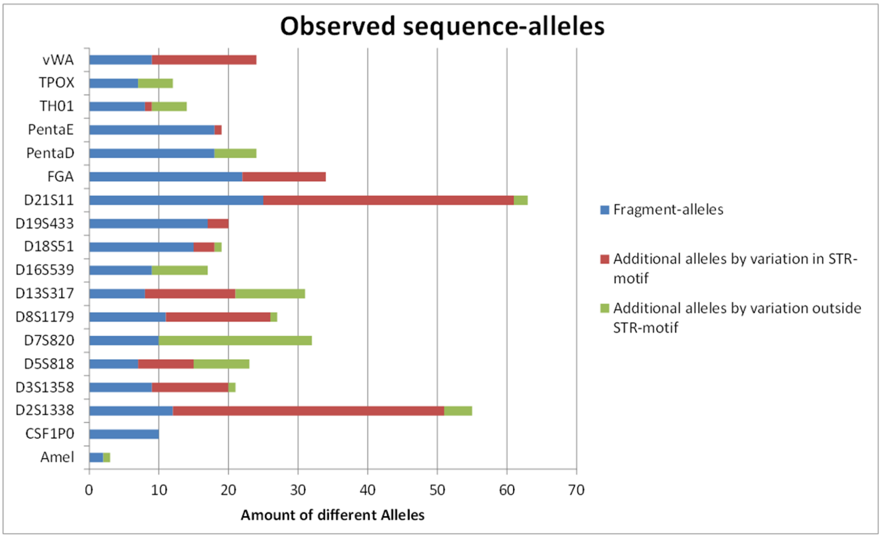
Reliability of sequencing results was assessed for the 297 population samples by comparison of CE data from the PowerPlex® Fusion System with the sequencing data. All STR alleles from the sequencing data were in concordance with CE analysis except for two alleles from PentaD. These alleles were missed when using the 15% within locus threshold (heterozygote balance), as they had a frequency of 8% and 12% of the highest allele (Sup. Figure 4). 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].

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 (Figure 5). Sup. Figure 5 displays the sequence of the genome reference (GRCh37/hg19) and of control sample 2800M (which is provided with the assay). Sup. Figure 6 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.

Figure 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).

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	5.3E-20	8.6E-23	2.2E-20	4.6E-23	4.1E-20	5.2E-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.

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

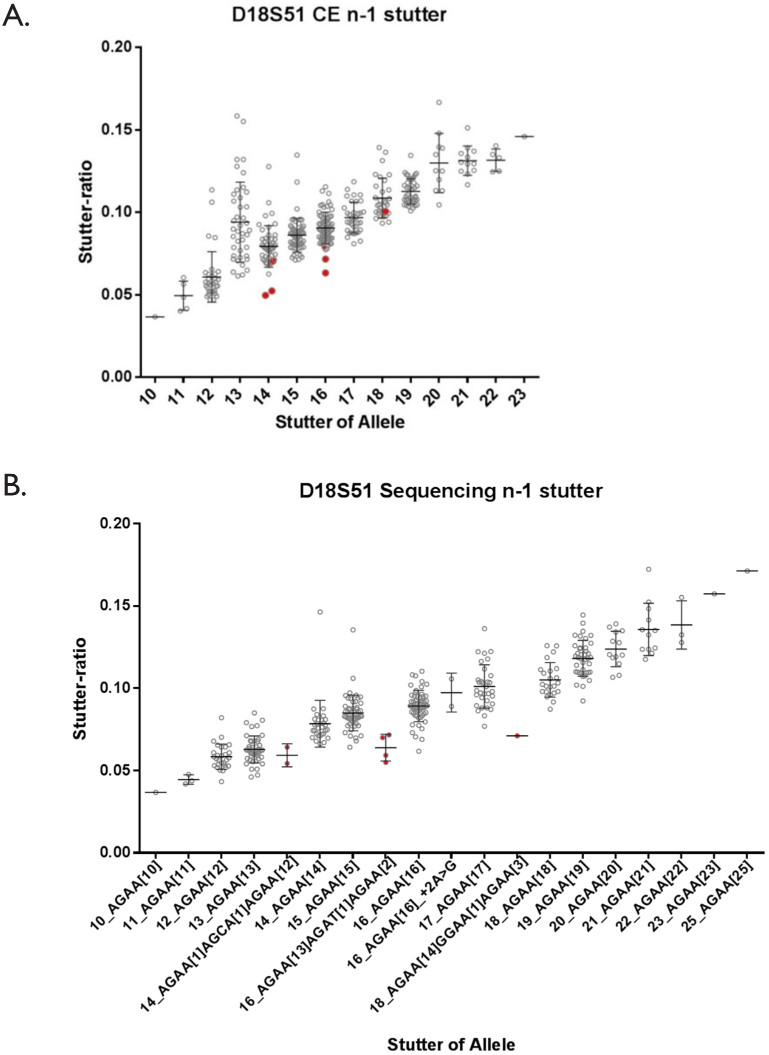
Stutter analysis

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 Supplemental Figure 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 (Sup. Figure 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 Figure 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 (Sup. Figure 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.

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. Figure 7a displays the read percentage for each allele call of the minor contributor 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 Sup. Figure 9 for further explanation of this procedure for a hypothetical three locus mixture profile). Figure 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.

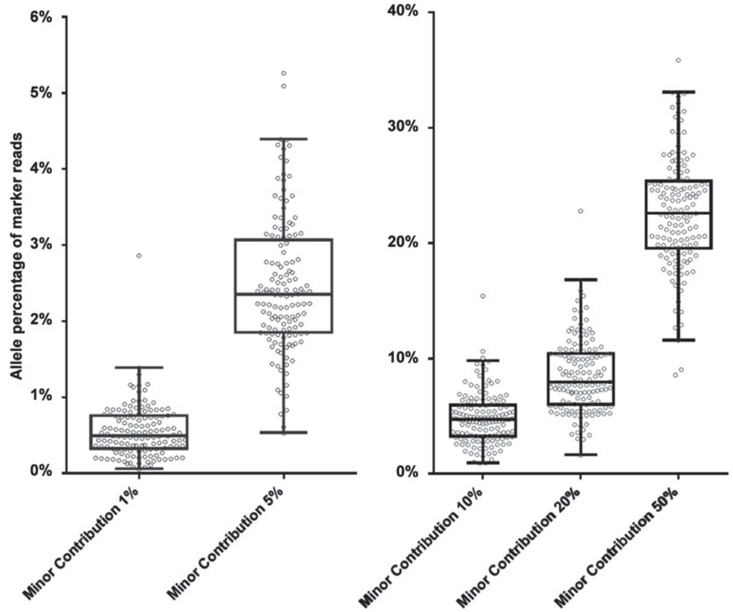
Figure 6. Comparison of stutter ratios for locus D18S51 analysed by CE and MPS



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.*

Figure 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.



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%

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.

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.

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) x 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.

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.

Reference List

1. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE,

Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJ, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. Integrating common and rare genetic variation in diverse human populations. Nature 2010; 467:52-58

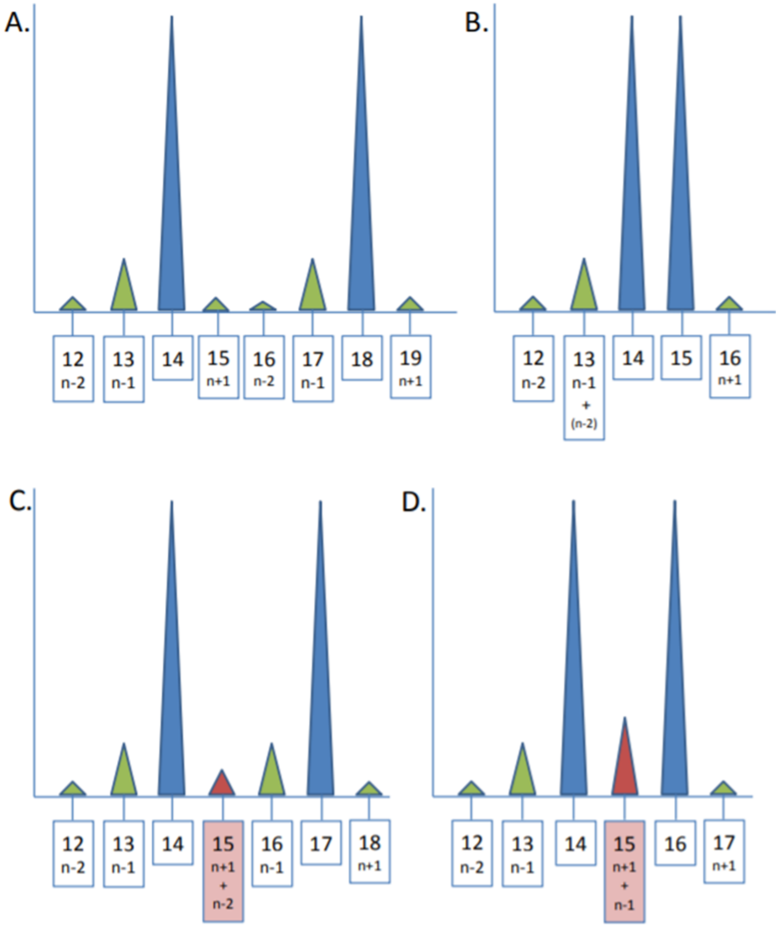
2. Anvar SY, van der Gaag KJ, van der Heijden JW, Veltrop MH, Vossen RH, de Leeuw RH, Breukel C, Buermans HP, Verbeek JS, de KP, den Dunnen JT, Laros JF. TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes. Bioinformatics. 2014; 30:1651-1659
3. Brookes C, Bright JA, Harbison S, Buckleton J. Characterising stutter in forensic STR multiplexes. Forensic Sci.Int.Genet. 2012; 6:58-63
4. Budowle B, Onorato AJ, Callaghan TF, Della MA, Gross AM, Guerrieri RA, Luttman JC, McClure DL. Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J.Forensic Sci. 2009; 54:810-821
5. Cann HM, de TC, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL. A human genome diversity cell line panel. Science 2002; 296:261-262
6. Gelardi C, Rockenbauer E, Dalsgaard S, Borsting C, Morling N. Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. Forensic Sci.Int.Genet. 2014; 12:38-41
7. Gettings KB, Aponte RA, Vallone PM, Butler JM. STR allele sequence variation: Current knowledge and future issues. Forensic Sci.Int.Genet. 2015; 18:118-130
8. Kline MC, Hill CR, Decker AE, Butler JM. STR sequence analysis for characterizing normal, variant, and null alleles. Forensic Sci.Int.Genet. 2011; 5:329-332
9. Knijff, P. and Pijpe, J. Population genetics of African Pygmies. 2015. (GENERIC). Ref Type: Unpublished Work
10. Kraaijenbrink T, van der Gaag KJ, Zuniga SB, Xue Y, Carvalho-Silva DR, Tyler-Smith C, Jobling MA, Parkin EJ, Su B, Shi H, Xiao CJ, Tang WR, Kashyap VK, Trivedi R, Sitalaximi T, Banerjee J, Karma Tshering of Gaselo, Tuladhar NM, Opgenort JR, van Driem GL, Barbujani G, de KP. A linguistically informed autosomal STR survey of human populations residing in the greater Himalayan region. PLoS.One. 2014; 9:e91534
11. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011; 27:2957-2963
12. Oostdik K, Lenz K, Nye J, Schelling K, Yet D, Bruski S, Strong J, Buchanan C, Sutton J, Linner J, Frazier N, Young H, Matthies L, Sage A, Hahn J, Wells R, Williams N, Price M, Koehler J, Staples M, Swango KL, Hill C, Oyerly K, Duke W, Katzilierakis L, Ensenberger MG, Bourdeau JM, Sprecher CJ, Krenke B, Storts DR. 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. 2014; 12:69-76
13. Promega Corporation. Powerstats v12. 2015. (GENERIC). Ref Type: Unpublished Work
14. Promega Corporation. TECHNICAL MANUAL, PowerPlex® Fusion System. 1-3-2015.

(GENERIC). Ref Type: Unpublished Work

15. Scheible M, Loreille O, Just R, Irwin J. Short tandem repeat typing on the 454 platform: strategies and considerations for targeted sequencing of common forensic markers. Forensic Sci.Int.Genet. 2014; 12:107-119
16. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015;
17. van der Gaag KJ, de Knijff P. Forensic nomenclature for short tandem repeats updated for sequencing. Forensic Science International: Genetics Supplement Series 2015; Volume 4
18. Warshauer DH, King JL, Budowle B. STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Forensic Sci.Int.Genet. 2015; 14:182-186
19. Westen AA, Grol LJ, Harteveld J, Matai AS, de KP, Sijen T. Assessment of the stochastic threshold, back- and forward stutter filters and low template techniques for NGM. Forensic Sci.Int.Genet. 2012; 6:708-715
20. Westen AA, Kraaijenbrink T, Robles de Medina EA, Harteveld J, Willemse P, Zuniga SB, van der Gaag KJ, Weiler NE, Warnaar J, Kayser M, Sijen T, de KP. Comparing six commercial autosomal STR kits in a large Dutch population sample. Forensic Sci.Int. Genet. 2014; 10:55-63
21. Zeng X, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, Patel J, Storts DR, Budowle B. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci.Int.Genet. 2015; 16:38-47

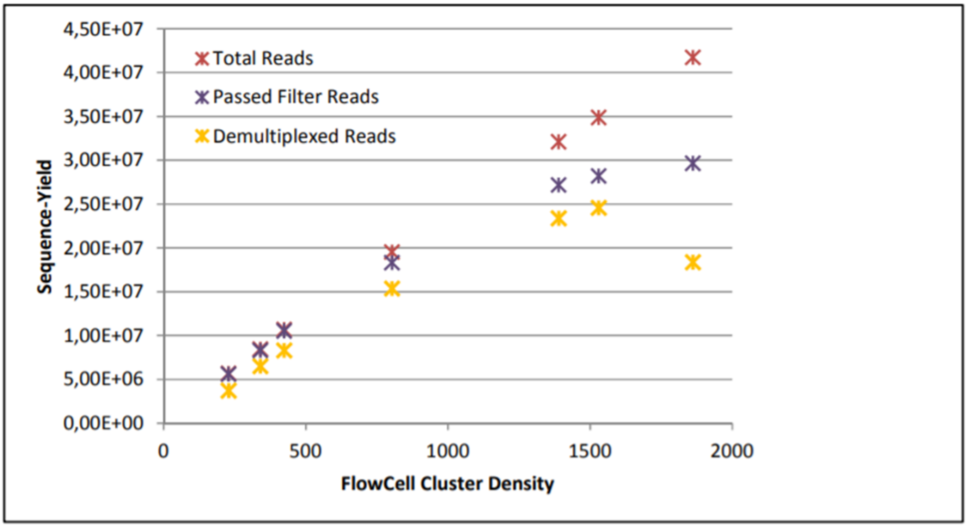
Supplementary materials

Supplemental Figure 1, four examples of allele combinations to illustrate the criteria used for inclusion of stutters for the calculation of stutter ratios



The peak profiles display which stutter peaks are included and excluded for analysis of stutter ratios. In all four examples, the blue peaks represent the genuine alleles, the green peaks represent stutters that are included in the calculation of stutter ratios and the red peaks represent stutters that are excluded for the calculation of stutter ratios. Example A.: Both alleles have no overlapping stutters, all the stutter peaks are included for calculation of stutter ratios. Example B.: Allele 14 is overlapping with the n-1 stutter of allele 15. For this allele 15, the n-1 stutter cannot be used for calculation of the stutter ratio. The n-1 stutter of allele 14 overlaps the n-2 stutter of allele 15 but is still included in the calculation of the stutter ratio for allele 14 since the contribution of the n-2 stutter to the peak height is considered to be negligible. Example C.: The n+1 stutter of allele 14 is overlapping with the n-2 stutter of allele 17. This stutter position is removed from the analysis of the stutter ratio. Example D.: The n-1 stutter of allele 16 overlaps with the n+1 stutter of allele 14 and is excluded from the analysis of the stutter ratio. For the sequencing data, the same criteria were used but sequence variation was considered resulting in fewer alleles to be excluded from the calculation of the stutter ratios.

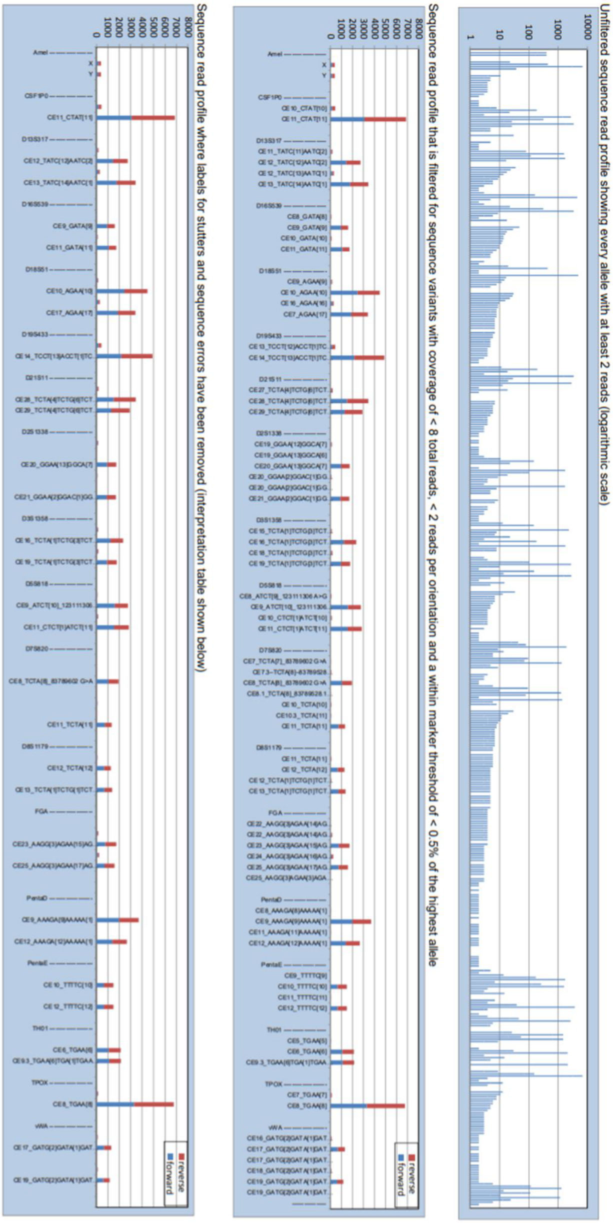
Supplemental Figure 2. Overview of sequence efficiency for MiSeq® sequencing runs with different cluster density

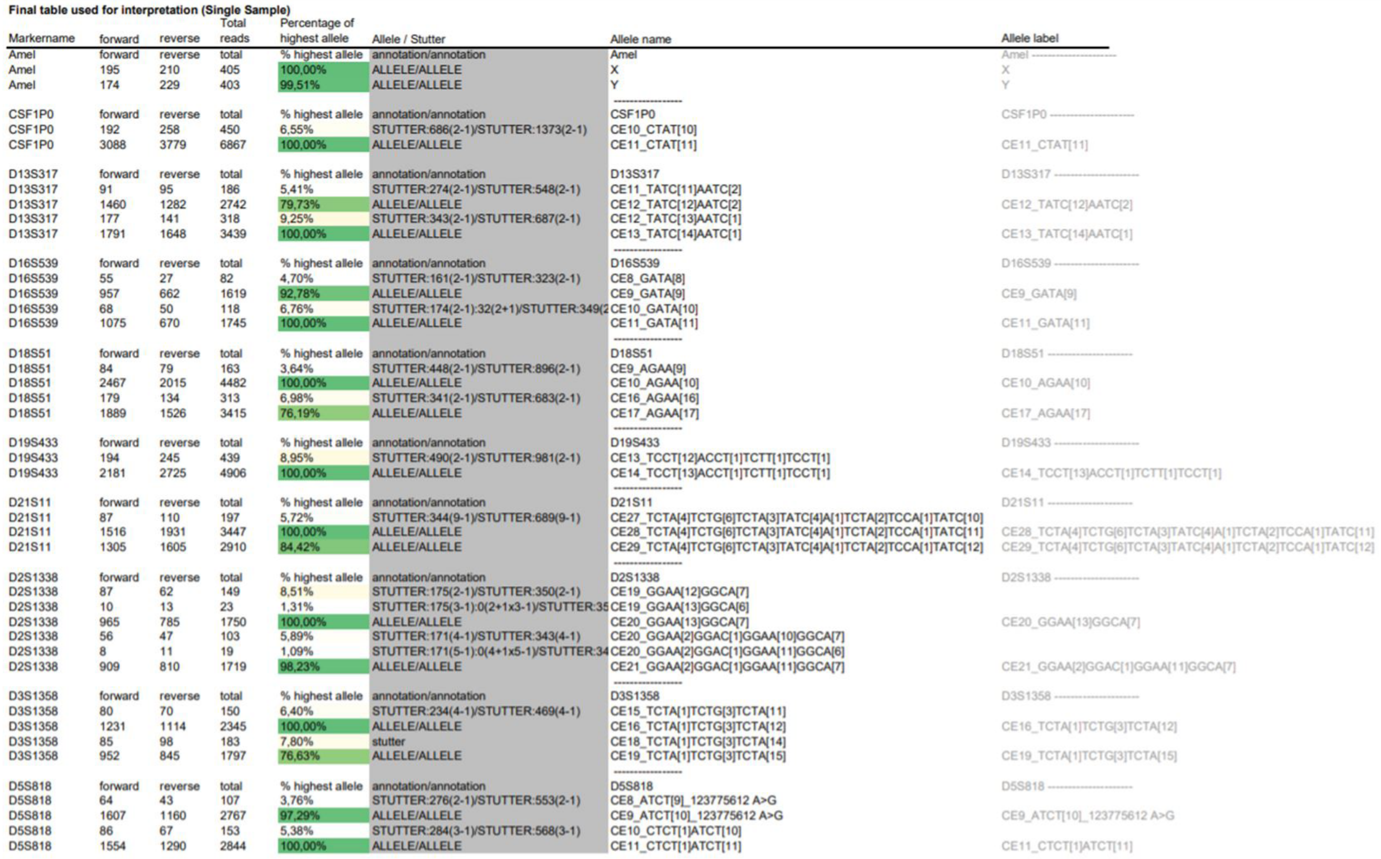


Scatterplot displaying the yield of sequence reads after different filtering steps in the analysis from signal on the MiSeq® until the reads retained in the fastq files after demultiplexing for runs with different levels of cluster density. In red, the initial number of reads are displayed before any filtering took place on the MiSeq®. In purple, remaining reads after MiSeq® quality filtering are displayed. In orange, the reads are displayed in which a barcode is recognised during de-multiplexing.

Supplemental Figure 3. Overview of analysis filtering / interpretation steps of a sequence DNA profile for a single source sample and a mixture

A.





Final table used for interpretation (Single Sample)

Markername	forward	reverse	Total reads	Percentage of highest allele	Allele / Stutter	Allele name	Allele label
Amel	forward	reverse	total	% highest allele	annotation/annotation	Amel	Amel ---------------------
Amel	195	210	405	100,00%	ALLELE/ALLELE	X	X
Amel	174	229	403	99,51%	ALLELE/ALLELE	Y	Y

CSF1P0	forward	reverse	total	% highest allele	annotation/annotation	CSF1P0	CSF1P0 ---------------------
CSF1P0	192	258	450	6,55%	STUTTER:686(2-1)/STUTTER:1373(2-1)	CE10_CTAT[10]	
CSF1P0	3088	3779	6867	100,00%	ALLELE/ALLELE	CE11_CTAT[11]	CE11_CTAT[11]
D13S317	forward	reverse	total	% highest allele	annotation/annotation	D13S317	D13S317 ---------------------
D13S317	91	95	186	5,41%	STUTTER:274(2-1)/STUTTER:548(2-1)	CE11_TATC[11]AATC[2]	
D13S317	1460	1282	2742	79,73%	ALLELE/ALLELE	CE12_TATC[12]AATC[2]	CE12_TATC[12]AATC[2]
D13S317	177	141	318	9,25%	STUTTER:343(2-1)/STUTTER:687(2-1)	CE12_TATC[13]AATC[1]	
D13S317	1791	1648	3439	100,00%	ALLELE/ALLELE	CE13_TATC[14]AATC[1]	CE13_TATC[14]AATC[1]

D16S539	forward	reverse	total	% highest allele	annotation/annotation	D16S539	D16S539 ---------------------
D16S539	55	27	82	4,70%	STUTTER:161(2-1)/STUTTER:323(2-1)	CE8_GATA[8]	
D16S539	957	662	1619	92,78%	ALLELE/ALLELE	CE9_GATA[9]	CE9_GATA[9]
D16S539	68	50	118	6,76%	STUTTER:174(2-1):32(2+1)/STUTTER:349(2	CE10_GATA[10]	
D16S539	1075	670	1745	100,00%	ALLELE/ALLELE	CE11_GATA[11]	CE11_GATA[11]

D18S51	forward	reverse	total	% highest allele	annotation/annotation	D18S51	D18S51 ---------------------
D18S51	84	79	163	3,64%	STUTTER:448(2-1)/STUTTER:896(2-1)	CE9_AGAA[9]	
D18S51	2467	2015	4482	100,00%	ALLELE/ALLELE	CE10_AGAA[10]	CE10_AGAA[10]
D18S51	179	134	313	6,98%	STUTTER:341(2-1)/STUTTER:683(2-1)	CE16_AGAA[16]	
D18S51	1889	1526	3415	76,19%	ALLELE/ALLELE	CE17_AGAA[17]	CE17_AGAA[17]

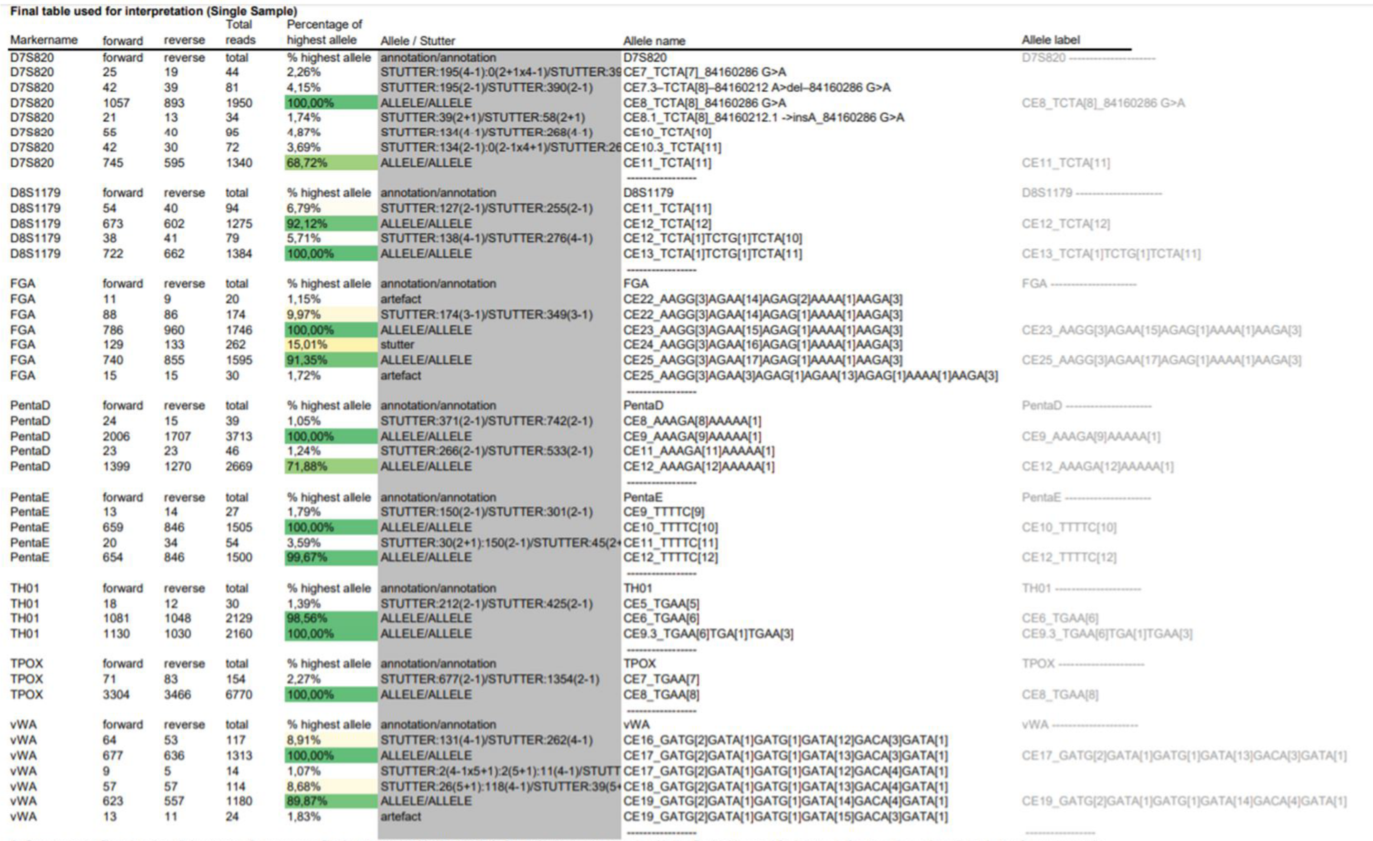
D19S433	forward	reverse	total	% highest allele	annotation/annotation	D19S433	D19S433 ---------------------
D19S433	194	245	439	8,95%	STUTTER:490(2-1)/STUTTER:981(2-1)	CE13_TCCT[12]ACCT[1]TCTT[1]TCCT[1]	
D19S433	2181	2725	4906	100,00%	ALLELE/ALLELE	CE14_TCCT[13]ACCT[1]TCTT[1]TCCT[1]	CE14_TCCT[13]ACCT[1]TCTT[1]TCCT[1]

D21S11	forward	reverse	total	% highest allele	annotation/annotation	D21S11	D21S11 ---------------------
D21S11	87	110	197	5,72%	STUTTER:344(9-1)/STUTTER:689(9-1)	CE27_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]	
D21S11	1516	1931	3447	100,00%	ALLELE/ALLELE	CE28_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]	CE28_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]
D21S11	1305	1605	2910	84,42%	ALLELE/ALLELE	CE29_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]	CE29_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]

D2S1338	forward	reverse	total	% highest allele	annotation/annotation	D2S1338	D2S1338 ---------------------
D2S1338	87	62	149	8,51%	STUTTER:175(2-1)/STUTTER:350(2-1)	CE19_GGAA[12]GGCA[7]	
D2S1338	10	13	23	1,31%	STUTTER:175(3-1):0(2+1x3-1)/STUTTER:35	CE19_GGAA[13]GGCA[6]	
D2S1338	965	785	1750	100,00%	ALLELE/ALLELE	CE20_GGAA[13]GGCA[7]	CE20_GGAA[13]GGCA[7]
D2S1338	56	47	103	5,89%	STUTTER:171(4-1)/STUTTER:343(4-1)	CE20_GGAA[2]GGAC[1]GGAA[10]GGCA[7]	
D2S1338	8	11	19	1,09%	STUTTER:171(5-1):0(4+1x5-1)/STUTTER:34	CE20_GGAA[2]GGAC[1]GGAA[11]GGCA[6]	
D2S1338	909	810	1719	98,23%	ALLELE/ALLELE	CE21_GGAA[2]GGAC[1]GGAA[11]GGCA[7]	CE21_GGAA[2]GGAC[1]GGAA[11]GGCA[7]

D3S1358	forward	reverse	total	% highest allele	annotation/annotation	D3S1358	D3S1358 ---------------------
D3S1358	80	70	150	6,40%	STUTTER:234(4-1)/STUTTER:469(4-1)	CE15_TCTA[1]TCTG[3]TCTA[11]	
D3S1358	1231	1114	2345	100,00%	ALLELE/ALLELE	CE16_TCTA[1]TCTG[3]TCTA[12]	CE16_TCTA[1]TCTG[3]TCTA[12]
D3S1358	85	98	183	7,80%	stutter	CE18_TCTA[1]TCTG[3]TCTA[14]	
D3S1358	952	845	1797	76,63%	ALLELE/ALLELE	CE19_TCTA[1]TCTG[3]TCTA[15]	CE19_TCTA[1]TCTG[3]TCTA[15]

D5S818	forward	reverse	total	% highest allele	annotation/annotation	D5S818	D5S818 ---------------------
D5S818	64	43	107	3,76%	STUTTER:276(2-1)/STUTTER:553(2-1)	CE8_ATCT[9]_123775612 A>G	
D5S818	1607	1160	2767	97,29%	ALLELE/ALLELE	CE9_ATCT[10]_123775612 A>G	CE9_ATCT[10]_123775612 A>G
D5S818	86	67	153	5,38%	STUTTER:284(3-1)/STUTTER:568(3-1)	CE10_CTCT[1]ATCT[10]	
D5S818	1554	1290	2844	100,00%	ALLELE/ALLELE	CE11_CTCT[1]ATCT[11]	CE11_CTCT[1]ATCT[11]



Final table used for interpretation (Single Sample)

Markername	forward	reverse	Total reads	Percentage of highest allele	Allele / Stutter	Allele name	Allele label
D7S820	forward	reverse	total	% highest allele	annotation/annotation	D7S820	D7S820 ---------------------
D7S820	25	19	44	2,26%	STUTTER:195(4-1):0(2+1x4-1)/STUTTER:39	CE7_TCTA[7]_84160286 G>A	
D7S820	42	39	81	4,15%	STUTTER:195(2-1)/STUTTER:390(2-1)	CE7.3–TCTA[8]–84160212 A>del–84160286 G>A	
D7S820	1057	893	1950	100,00%	ALLELE/ALLELE	CE8_TCTA[8]_84160286 G>A	CE8_TCTA[8]_84160286 G>A
D7S820	21	13	34	1,74%	STUTTER:39(2+1)/STUTTER:58(2+1)	CE8.1_TCTA[8]_84160212.1 ->insA_84160286 G>A	
D7S820	55	40	95	4,87%	STUTTER:134(4-1)/STUTTER:268(4-1)	CE10_TCTA[10]	
D7S820	42	30	72	3,69%	STUTTER:134(2-1):0(2-1x4+1)/STUTTER:26	CE10.3_TCTA[11]	
D7S820	745	595	1340	68,72%	ALLELE/ALLELE	CE11_TCTA[11]	CE11_TCTA[11]

D8S1179	forward	reverse	total	% highest allele	annotation/annotation	D8S1179	D8S1179 ---------------------
D8S1179	54	40	94	6,79%	STUTTER:127(2-1)/STUTTER:255(2-1)	CE11_TCTA[11]	
D8S1179	673	602	1275	92,12%	ALLELE/ALLELE	CE12_TCTA[12]	CE12_TCTA[12]
D8S1179	38	41	79	5,71%	STUTTER:138(4-1)/STUTTER:276(4-1)	CE12_TCTA[1]TCTG[1]TCTA[10]	
D8S1179	722	662	1384	100,00%	ALLELE/ALLELE	CE13_TCTA[1]TCTG[1]TCTA[11]	CE13_TCTA[1]TCTG[1]TCTA[11]

FGA	forward	reverse	total	% highest allele	annotation/annotation	FGA	FGA ---------------------
FGA	11	9	20	1,15%	artefact	CE22_AAGG[3]AGAA[14]AGAG[2]AAAA[1]AAGA[3]	
FGA	88	86	174	9,97%	STUTTER:174(3-1)/STUTTER:349(3-1)	CE22_AAGG[3]AGAA[14]AGAG[1]AAAA[1]AAGA[3]	
FGA	786	960	1746	100,00%	ALLELE/ALLELE	CE23_AAGG[3]AGAA[15]AGAG[1]AAAA[1]AAGA[3]	CE23_AAGG[3]AGAA[15]AGAG[1]AAAA[1]AAGA[3]
FGA	129	133	262	15,01%	stutter	CE24_AAGG[3]AGAA[16]AGAG[1]AAAA[1]AAGA[3]	
FGA	740	855	1595	91,35%	ALLELE/ALLELE	CE25_AAGG[3]AGAA[17]AGAG[1]AAAA[1]AAGA[3]	CE25_AAGG[3]AGAA[17]AGAG[1]AAAA[1]AAGA[3]
FGA	15	15	30	1,72%	artefact	CE25_AAGG[3]AGAA[3]AGAG[1]AGAA[13]AGAG[1]AAAA[1]AAGA[3]	

PentaD	forward	reverse	total	% highest allele	annotation/annotation	PentaD	PentaD ---------------------
PentaD	24	15	39	1,05%	STUTTER:371(2-1)/STUTTER:742(2-1)	CE8_AAAGA[8]AAAAA[1]	
PentaD	2006	1707	3713	100,00%	ALLELE/ALLELE	CE9_AAAGA[9]AAAAA[1]	CE9_AAAGA[9]AAAAA[1]
PentaD	23	23	46	1,24%	STUTTER:266(2-1)/STUTTER:533(2-1)	CE11_AAAGA[11]AAAAA[1]	
PentaD	1399	1270	2669	71,88%	ALLELE/ALLELE	CE12_AAAGA[12]AAAAA[1]	CE12_AAAGA[12]AAAAA[1]

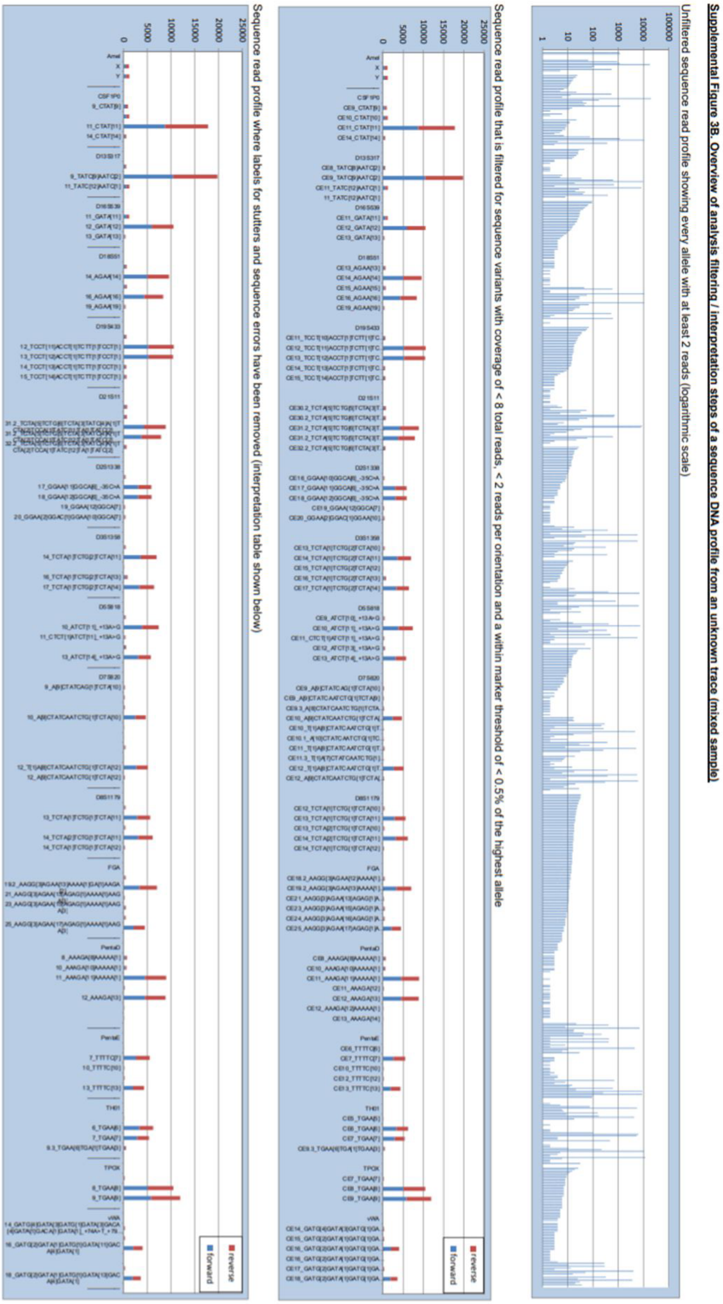
PentaE	forward	reverse	total	% highest allele	annotation/annotation	PentaE	PentaE ---------------------
PentaE	13	14	27	1,79%	STUTTER:150(2-1)/STUTTER:301(2-1)	CE9_TTTTC[9]	
PentaE	659	846	1505	100,00%	ALLELE/ALLELE	CE10_TTTTC[10]	CE10_TTTTC[10]
PentaE	20	34	54	3,59%	STUTTER:30(2+1):150(2-1)/STUTTER:45(2+	CE11_TTTTC[11]	
PentaE	654	846	1500	99,67%	ALLELE/ALLELE	CE12_TTTTC[12]	CE12_TTTTC[12]

TH01	forward	reverse	total	% highest allele	annotation/annotation	TH01	TH01 ---------------------
TH01	18	12	30	1,39%	STUTTER:212(2-1)/STUTTER:425(2-1)	CE5_TGAA[5]	
TH01	1081	1048	2129	98,56%	ALLELE/ALLELE	CE6_TGAA[6]	CE6_TGAA[6]
TH01	1130	1030	2160	100,00%	ALLELE/ALLELE	CE9.3_TGAA[6]TGA[1]TGAA[3]	CE9.3_TGAA[6]TGA[1]TGAA[3]

TPOX	forward	reverse	total	% highest allele	annotation/annotation	TPOX	TPOX ---------------------
TPOX	71	83	154	2,27%	STUTTER:677(2-1)/STUTTER:1354(2-1)	CE7_TGAA[7]	
TPOX	3304	3466	6770	100,00%	ALLELE/ALLELE	CE8_TGAA[8]	CE8_TGAA[8]
						------------------	
vWA	forward	reverse	total	% highest allele	annotation/annotation	vWA	vWA ---------------------
vWA	64	53	117	8,91%	STUTTER:131(4-1)/STUTTER:262(4-1)	CE16_GATG[2]GATA[1]GATG[1]GATA[12]GACA[3]GATA[1]	
vWA	677	636	1313	100,00%	ALLELE/ALLELE	CE17_GATG[2]GATA[1]GATG[1]GATA[13]GACA[3]GATA[1]	CE17_GATG[2]GATA[1]GATG[1]GATA[13]GACA[3]GATA[1]
vWA	9	5	14	1,07%	STUTTER:2(4-1x5+1):2(5+1):11(4-1)/STUTT	CE17_GATG[2]GATA[1]GATG[1]GATA[12]GACA[4]GATA[1]	
vWA	57	57	114	8,68%	STUTTER:26(5+1):118(4-1)/STUTTER:39(5+	CE18_GATG[2]GATA[1]GATG[1]GATA[13]GACA[4]GATA[1]	
vWA	623	557	1180	89,87%	ALLELE/ALLELE	CE19_GATG[2]GATA[1]GATG[1]GATA[14]GACA[4]GATA[1]	CE19_GATG[2]GATA[1]GATG[1]GATA[14]GACA[4]GATA[1]
vWA	13	11	24	1,83%	artefact	CE19_GATG[2]GATA[1]GATG[1]GATA[15]GACA[3]GATA[1]	
						------------------	------------------

A. Sequence profiles showing allele names after separate filtering steps used in the analysis for a single source sample with the final table used for interpretation to call genuine alleles in a reference sample.

B.

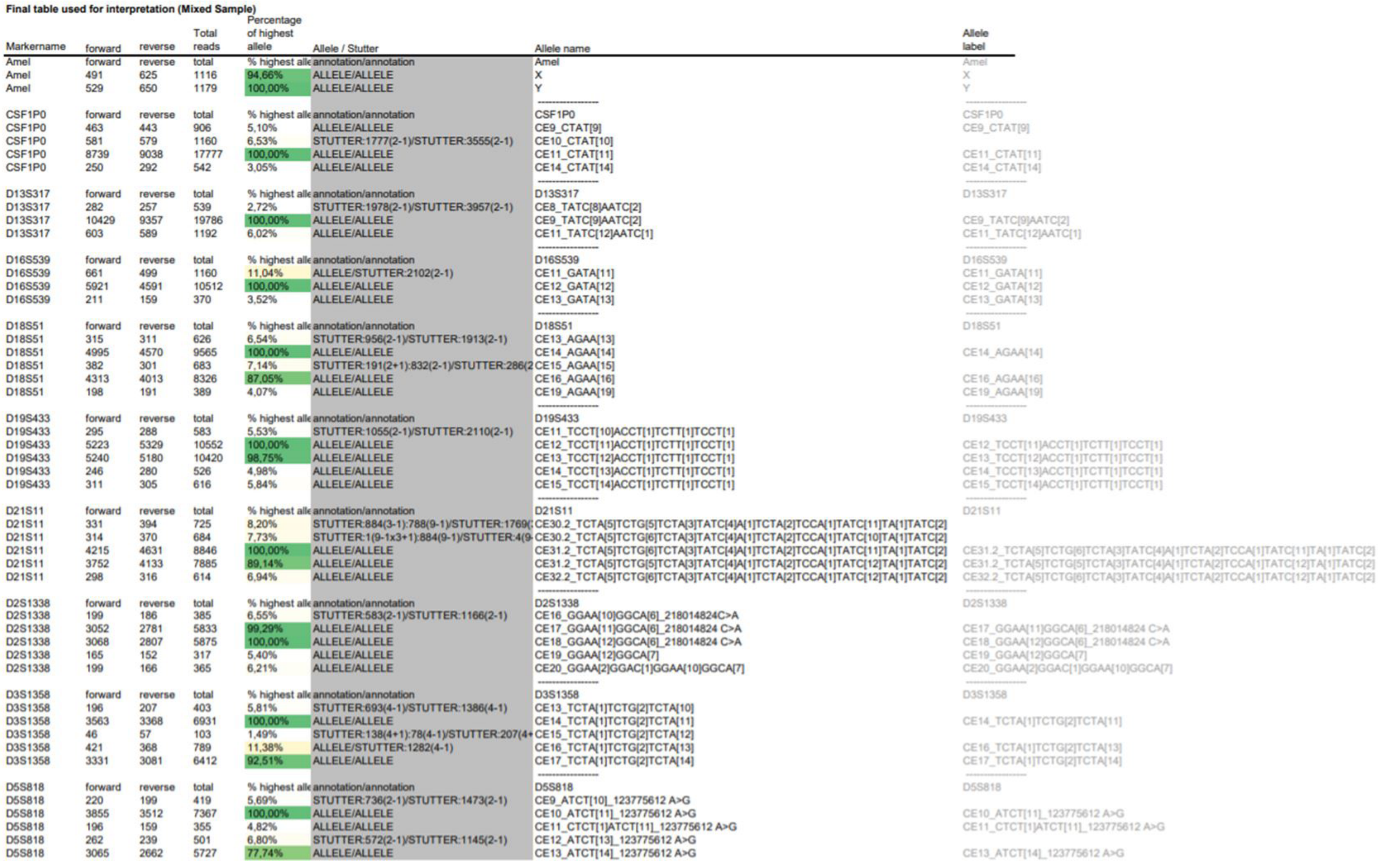


Supplemental Figure 3B. Overview of analysis filtering / interpretation steps of a sequence DNA profile from an unknown trace (mixed sample)

Unfiltered sequence read profile showing every allele with at least 2 reads (logarithmic scale)

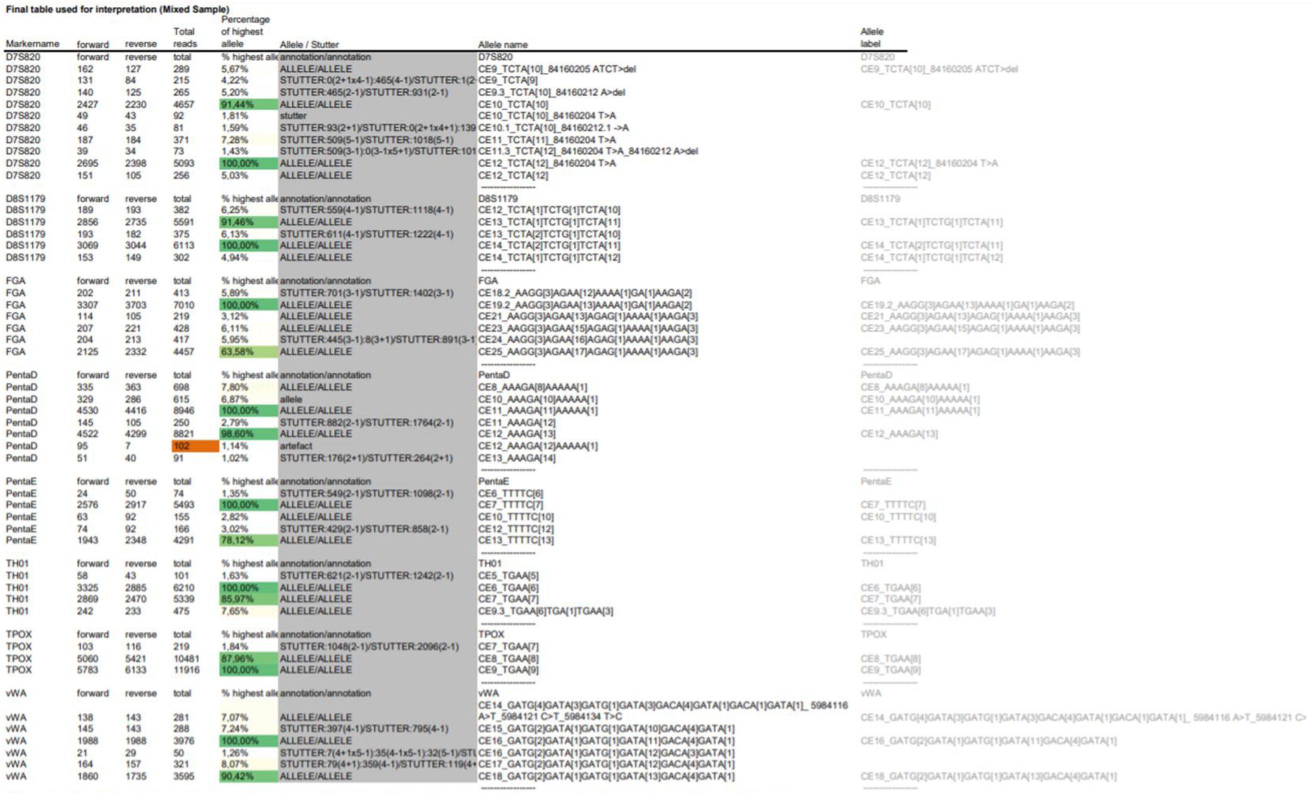
Sequence read profile that is filtered for sequence variants with coverage of < 8 total reads, < 2 reads per orientation and a within marker threshold of < 0.5% of the highest allele

Sequence read profile where labels for stutters and sequence errors have been removed (interpretation table shown below)



Final table used for interpretation (Mixed Sample)

Markername	forward	reverse	Total reads	Percentage of highest allele	Allele / Stutter	Allele name	Allele label
Amel	forward	reverse	total	% highest alle	annotation/annotation	Amel	Amel
Amel	491	625	1116	94,66%	ALLELE/ALLELE	X	X
Amel	529	650	1179	100,00%	ALLELE/ALLELE	Y	Y
						------------------	------------------
CSF1P0	forward	reverse	total	% highest alle	annotation/annotation	CSF1P0	CSF1P0
CSF1P0	463	443	906	5,10%	ALLELE/ALLELE	CE9_CTAT[9]	CE9_CTAT[9]
CSF1P0	581	579	1160	6,53%	STUTTER:1777(2-1)/STUTTER:3555(2-1)	CE10_CTAT[10]	
CSF1P0	8739	9038	17777	100,00%	ALLELE/ALLELE	CE11_CTAT[11]	CE11_CTAT[11]
CSF1P0	250	292	542	3,05%	ALLELE/ALLELE	CE14_CTAT[14]	CE14_CTAT[14]
						------------------	------------------
D13S317	forward	reverse	total	% highest alle	annotation/annotation	D13S317	D13S317
D13S317	282	257	539	2,72%	STUTTER:1978(2-1)/STUTTER:3957(2-1)	CE8_TATC[8]AATC[2]	
D13S317	10429	9357	19786	100,00%	ALLELE/ALLELE	CE9_TATC[9]AATC[2]	CE9_TATC[9]AATC[2]
D13S317	603	589	1192	6,02%	ALLELE/ALLELE	CE11_TATC[12]AATC[1]	CE11_TATC[12]AATC[1]
						------------------	------------------
D16S539	forward	reverse	total	% highest alle	annotation/annotation	D16S539	D16S539
D16S539	661	499	1160	11,04%	ALLELE/STUTTER:2102(2-1)	CE11_GATA[11]	CE11_GATA[11]
D16S539	5921	4591	10512	100,00%	ALLELE/ALLELE	CE12_GATA[12]	CE12_GATA[12]
D16S539	211	159	370	3,52%	ALLELE/ALLELE	CE13_GATA[13]	CE13_GATA[13]
						------------------	------------------
D18S51	forward	reverse	total	% highest alle	annotation/annotation	D18S51	D18S51
D18S51	315	311	626	6,54%	STUTTER:956(2-1)/STUTTER:1913(2-1)	CE13_AGAA[13]	
D18S51	4995	4570	9565	100,00%	ALLELE/ALLELE	CE14_AGAA[14]	CE14_AGAA[14]
D18S51	382	301	683	7,14%	STUTTER:191(2+1):832(2-1)/STUTTER:286(2	CE15_AGAA[15]	
D18S51	4313	4013	8326	87,05%	ALLELE/ALLELE	CE16_AGAA[16]	CE16_AGAA[16]
D18S51	198	191	389	4,07%	ALLELE/ALLELE	CE19_AGAA[19]	CE19_AGAA[19]
						------------------	------------------
D19S433	forward	reverse	total	% highest alle	annotation/annotation	D19S433	D19S433
D19S433	295	288	583	5,53%	STUTTER:1055(2-1)/STUTTER:2110(2-1)	CE11_TCCT[10]ACCT[1]TCTT[1]TCCT[1]	
D19S433	5223	5329	10552	100,00%	ALLELE/ALLELE	CE12_TCCT[11]ACCT[1]TCTT[1]TCCT[1]	CE12_TCCT[11]ACCT[1]TCTT[1]TCCT[1]
D19S433	5240	5180	10420	98,75%	ALLELE/ALLELE	CE13_TCCT[12]ACCT[1]TCTT[1]TCCT[1]	CE13_TCCT[12]ACCT[1]TCTT[1]TCCT[1]
D19S433	246	280	526	4,98%	ALLELE/ALLELE	CE14_TCCT[13]ACCT[1]TCTT[1]TCCT[1]	CE14_TCCT[13]ACCT[1]TCTT[1]TCCT[1]
D19S433	311	305	616	5,84%	ALLELE/ALLELE	CE15_TCCT[14]ACCT[1]TCTT[1]TCCT[1]	CE15_TCCT[14]ACCT[1]TCTT[1]TCCT[1]
						------------------	------------------
D21S11	forward	reverse	total	% highest alle	annotation/annotation	D21S11	D21S11
D21S11	331	394	725	8,20%	STUTTER:884(3-1):788(9-1)/STUTTER:1769(	CE30.2_TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2]	
D21S11	314	370	684	7,73%	STUTTER:1(9-1x3+1):884(9-1)/STUTTER:4(9-	CE30.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]TA[1]TATC[2]	
D21S11	4215	4631	8846	100,00%	ALLELE/ALLELE	CE31.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2]	CE31.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2]
D21S11	3752	4133	7885	89,14%	ALLELE/ALLELE	CE31.2_TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2]	CE31.2_TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2]
D21S11	298	316	614	6,94%	ALLELE/ALLELE	CE32.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2]	CE32.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2]
						------------------	------------------
D2S1338	forward	reverse	total	% highest alle	annotation/annotation	D2S1338	D2S1338
D2S1338	199	186	385	6,55%	STUTTER:583(2-1)/STUTTER:1166(2-1)	CE16_GGAA[10]GGCA[6]_218014824C>A	
D2S1338	3052	2781	5833	99,29%	ALLELE/ALLELE	CE17_GGAA[11]GGCA[6]_218014824 C>A	CE17_GGAA[11]GGCA[6]_218014824 C>A
D2S1338	3068	2807	5875	100,00%	ALLELE/ALLELE	CE18_GGAA[12]GGCA[6]_218014824 C>A	CE18_GGAA[12]GGCA[6]_218014824 C>A
D2S1338	165	152	317	5,40%	ALLELE/ALLELE	CE19_GGAA[12]GGCA[7]	CE19_GGAA[12]GGCA[7]
D2S1338	199	166	365	6,21%	ALLELE/ALLELE	CE20_GGAA[2]GGAC[1]GGAA[10]GGCA[7]	CE20_GGAA[2]GGAC[1]GGAA[10]GGCA[7]
						------------------	------------------
D3S1358	forward	reverse	total	% highest alle	annotation/annotation	D3S1358	D3S1358
D3S1358	196	207	403	5,81%	STUTTER:693(4-1)/STUTTER:1386(4-1)	CE13_TCTA[1]TCTG[2]TCTA[10]	
D3S1358	3563	3368	6931	100,00%	ALLELE/ALLELE	CE14_TCTA[1]TCTG[2]TCTA[11]	CE14_TCTA[1]TCTG[2]TCTA[11]
D3S1358	46	57	103	1,49%	STUTTER:138(4+1):78(4-1)/STUTTER:207(4+	CE15_TCTA[1]TCTG[2]TCTA[12]	
D3S1358	421	368	789	11,38%	ALLELE/STUTTER:1282(4-1)	CE16_TCTA[1]TCTG[2]TCTA[13]	CE16_TCTA[1]TCTG[2]TCTA[13]
D3S1358	3331	3081	6412	92,51%	ALLELE/ALLELE	CE17_TCTA[1]TCTG[2]TCTA[14]	CE17_TCTA[1]TCTG[2]TCTA[14]
						------------------	------------------
D5S818	forward	reverse	total	% highest alle	annotation/annotation	D5S818	D5S818
D5S818	220	199	419	5,69%	STUTTER:736(2-1)/STUTTER:1473(2-1)	CE9_ATCT[10]_123775612 A>G	
D5S818	3855	3512	7367	100,00%	ALLELE/ALLELE	CE10_ATCT[11]_123775612 A>G	CE10_ATCT[11]_123775612 A>G
D5S818	196	159	355	4,82%	ALLELE/ALLELE	CE11_CTCT[1]ATCT[11]_123775612 A>G	CE11_CTCT[1]ATCT[11]_123775612 A>G
D5S818	262	239	501	6,80%	STUTTER:572(2-1)/STUTTER:1145(2-1)	CE12_ATCT[13]_123775612 A>G	
D5S818	3065	2662	5727	77,74%	ALLELE/ALLELE	CE13_ATCT[14]_123775612 A>G	CE13_ATCT[14]_123775612 A>G



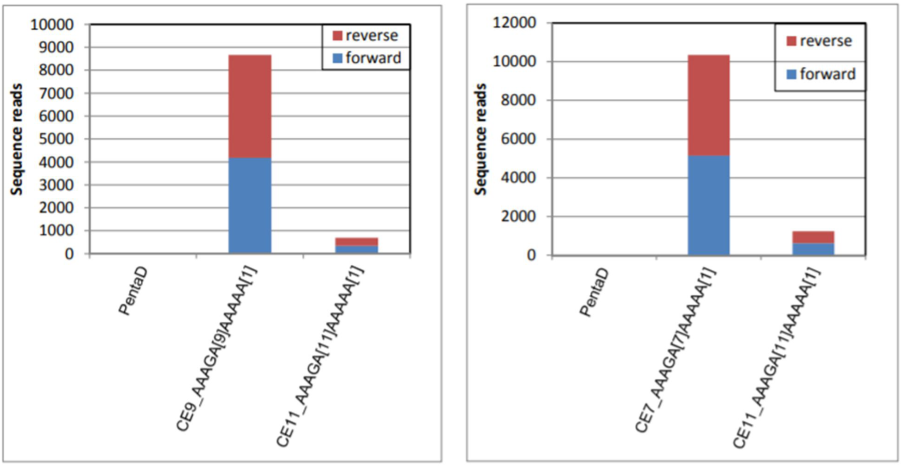
Final table used for interpretation (Mixed Sample)

Markername	forward	reverse	Total reads	Percentage of highest allele	Allele / Stutter	Allele name	Allele label
D7S820	forward	reverse	total	% highest alle	annotation/annotation	D7S820	D7S820
D7S820	162	127	289	5,67%	ALLELE/ALLELE	CE9_TCTA[10]_84160205 ATCT>del	CE9_TCTA[10]_84160205 ATCT>del
D7S820	131	84	215	4,22%	STUTTER:0(2+1x4-1):465(4-1)/STUTTER:1(2-	CE9_TCTA[9]	
D7S820	140	125	265	5,20%	STUTTER:465(2-1)/STUTTER:931(2-1)	CE9.3_TCTA[10]_84160212 A>del	
D7S820	2427	2230	4657	91,44%	ALLELE/ALLELE	CE10_TCTA[10]	CE10_TCTA[10]
D7S820	49	43	92	1,81%	stutter	CE10_TCTA[10]_84160204 T>A	
D7S820	46	35	81	1,59%	STUTTER:93(2+1)/STUTTER:0(2+1x4+1):139	CE10.1_TCTA[10]_84160212.1 ->A	
D7S820	187	184	371	7,28%	STUTTER:509(5-1)/STUTTER:1018(5-1)	CE11_TCTA[11]_84160204 T>A	
D7S820	39	34	73	1,43%	STUTTER:509(3-1):0(3-1x5+1)/STUTTER:101	CE11.3_TCTA[12]_84160204 T>A_84160212 A>del	
D7S820	2695	2398	5093	100,00%	ALLELE/ALLELE	CE12_TCTA[12]_84160204 T>A	CE12_TCTA[12]_84160204 T>A
D7S820	151	105	256	5,03%	ALLELE/ALLELE	CE12_TCTA[12]	CE12_TCTA[12]
						-----------------	-----------------
D8S1179	forward	reverse	total	% highest alle	annotation/annotation	D8S1179	D8S1179
D8S1179	189	193	382	6,25%	STUTTER:559(4-1)/STUTTER:1118(4-1)	CE12_TCTA[1]TCTG[1]TCTA[10]	
D8S1179	2856	2735	5591	91,46%	ALLELE/ALLELE	CE13_TCTA[1]TCTG[1]TCTA[11]	CE13_TCTA[1]TCTG[1]TCTA[11]
D8S1179	193	182	375	6,13%	STUTTER:611(4-1)/STUTTER:1222(4-1)	CE13_TCTA[2]TCTG[1]TCTA[10]	
D8S1179	3069	3044	6113	100,00%	ALLELE/ALLELE	CE14_TCTA[2]TCTG[1]TCTA[11]	CE14_TCTA[2]TCTG[1]TCTA[11]
D8S1179	153	149	302	4,94%	ALLELE/ALLELE	CE14_TCTA[1]TCTG[1]TCTA[12]	CE14_TCTA[1]TCTG[1]TCTA[12]
						-----------------	-----------------
FGA	forward	reverse	total	% highest alle	annotation/annotation	FGA	FGA
FGA	202	211	413	5,89%	STUTTER:701(3-1)/STUTTER:1402(3-1)	CE18.2_AAGG[3]AGAA[12]AAAA[1]GA[1]AAGA[2]	
FGA	3307	3703	7010	100,00%	ALLELE/ALLELE	CE19.2_AAGG[3]AGAA[13]AAAA[1]GA[1]AAGA[2]	CE19.2_AAGG[3]AGAA[13]AAAA[1]GA[1]AAGA[2]
FGA	114	105	219	3,12%	ALLELE/ALLELE	CE21_AAGG[3]AGAA[13]AGAG[1]AAAA[1]AAGA[3]	CE21_AAGG[3]AGAA[13]AGAG[1]AAAA[1]AAGA[3]
FGA	207	221	428	6,11%	ALLELE/ALLELE	CE23_AAGG[3]AGAA[15]AGAG[1]AAAA[1]AAGA[3]	CE23_AAGG[3]AGAA[15]AGAG[1]AAAA[1]AAGA[3]
FGA	204	213	417	5,95%	STUTTER:445(3-1):8(3+1)/STUTTER:891(3-1	CE24_AAGG[3]AGAA[16]AGAG[1]AAAA[1]AAGA[3]	
FGA	2125	2332	4457	63,58%	ALLELE/ALLELE	CE25_AAGG[3]AGAA[17]AGAG[1]AAAA[1]AAGA[3]	CE25_AAGG[3]AGAA[17]AGAG[1]AAAA[1]AAGA[3]
						-----------------	-----------------
PentaD	forward	reverse	total	% highest alle	annotation/annotation	PentaD	PentaD
PentaD	335	363	698	7,80%	ALLELE/ALLELE	CE8_AAAGA[8]AAAAA[1]	CE8_AAAGA[8]AAAAA[1]
PentaD	329	286	615	6,87%	allele	CE10_AAAGA[10]AAAAA[1]	CE10_AAAGA[10]AAAAA[1]
PentaD	4530	4416	8946	100,00%	ALLELE/ALLELE	CE11_AAAGA[11]AAAAA[1]	CE11_AAAGA[11]AAAAA[1]
PentaD	145	105	250	2,79%	STUTTER:882(2-1)/STUTTER:1764(2-1)	CE11_AAAGA[12]	
PentaD	4522	4299	8821	98,60%	ALLELE/ALLELE	CE12_AAAGA[13]	CE12_AAAGA[13]
PentaD	95	7	102	1,14%	artefact	CE12_AAAGA[12]AAAAA[1]	
PentaD	51	40	91	1,02%	STUTTER:176(2+1)/STUTTER:264(2+1)	CE13_AAAGA[14]	
						-----------------	-----------------
PentaE	forward	reverse	total	% highest alle	annotation/annotation	PentaE	PentaE
PentaE	24	50	74	1,35%	STUTTER:549(2-1)/STUTTER:1098(2-1)	CE6_TTTTC[6]	
PentaE	2576	2917	5493	100,00%	ALLELE/ALLELE	CE7_TTTTC[7]	CE7_TTTTC[7]
PentaE	63	92	155	2,82%	ALLELE/ALLELE	CE10_TTTTC[10]	CE10_TTTTC[10]
PentaE	74	92	166	3,02%	STUTTER:429(2-1)/STUTTER:858(2-1)	CE12_TTTTC[12]	
PentaE	1943	2348	4291	78,12%	ALLELE/ALLELE	CE13_TTTTC[13]	CE13_TTTTC[13]
						-----------------	-----------------
TH01	forward	reverse	total	% highest alle	annotation/annotation	TH01	TH01
TH01	58	43	101	1,63%	STUTTER:621(2-1)/STUTTER:1242(2-1)	CE5_TGAA[5]	
TH01	3325	2885	6210	100,00%	ALLELE/ALLELE	CE6_TGAA[6]	CE6_TGAA[6]
TH01	2869	2470	5339	85,97%	ALLELE/ALLELE	CE7_TGAA[7]	CE7_TGAA[7]
TH01	242	233	475	7,65%	ALLELE/ALLELE	CE9.3_TGAA[6]TGA[1]TGAA[3]	CE9.3_TGAA[6]TGA[1]TGAA[3]
						-----------------	-----------------
TPOX	forward	reverse	total	% highest alle	annotation/annotation	TPOX	TPOX
TPOX	103	116	219	1,84%	STUTTER:1048(2-1)/STUTTER:2096(2-1)	CE7_TGAA[7]	
TPOX	5060	5421	10481	87,96%	ALLELE/ALLELE	CE8_TGAA[8]	CE8_TGAA[8]
TPOX	5783	6133	11916	100,00%	ALLELE/ALLELE	CE9_TGAA[9]	CE9_TGAA[9]
						-----------------	-----------------
vWA	forward	reverse	total	% highest alle	annotation/annotation	vWA	vWA
vWA	138	143	281	7,07%	ALLELE/ALLELE	CE14_GATG[4]GATA[3]GATG[1]GATA[3]GACA[4]GATA[1]GACA[1]GATA[1]_ 5984116 A>T_5984121 C>T_5984134 T>C	CE14_GATG[4]GATA[3]GATG[1]GATA[3]GACA[4]GATA[1]GACA[1]GATA[1]_ 5984116 A>T_5984121 C>
vWA	145	143	288	7,24%	STUTTER:397(4-1)/STUTTER:795(4-1)	CE15_GATG[2]GATA[1]GATG[1]GATA[10]GACA[4]GATA[1]	
vWA	1988	1988	3976	100,00%	ALLELE/ALLELE	CE16_GATG[2]GATA[1]GATG[1]GATA[11]GACA[4]GATA[1]	CE16_GATG[2]GATA[1]GATG[1]GATA[11]GACA[4]GATA[1]
vWA	21	29	50	1,26%	STUTTER:7(4+1x5-1):35(4-1x5-1):32(5-1)/STU	CE16_GATG[2]GATA[1]GATG[1]GATA[12]GACA[3]GATA[1]	
vWA	164	157	321	8,07%	STUTTER:79(4+1):359(4-1)/STUTTER:119(4+	CE17_GATG[2]GATA[1]GATG[1]GATA[12]GACA[4]GATA[1]	
vWA	1860	1735	3595	90,42%	ALLELE/ALLELE	CE18_GATG[2]GATA[1]GATG[1]GATA[13]GACA[4]GATA[1]	CE18_GATG[2]GATA[1]GATG[1]GATA[13]GACA[4]GATA[1]
						-----------------	-----------------

B. Sequence profiles showing allele names after separate filtering steps used in the analysis for a mixed sample with the final table used for interpretation to call genuine alleles in a mixed sample.

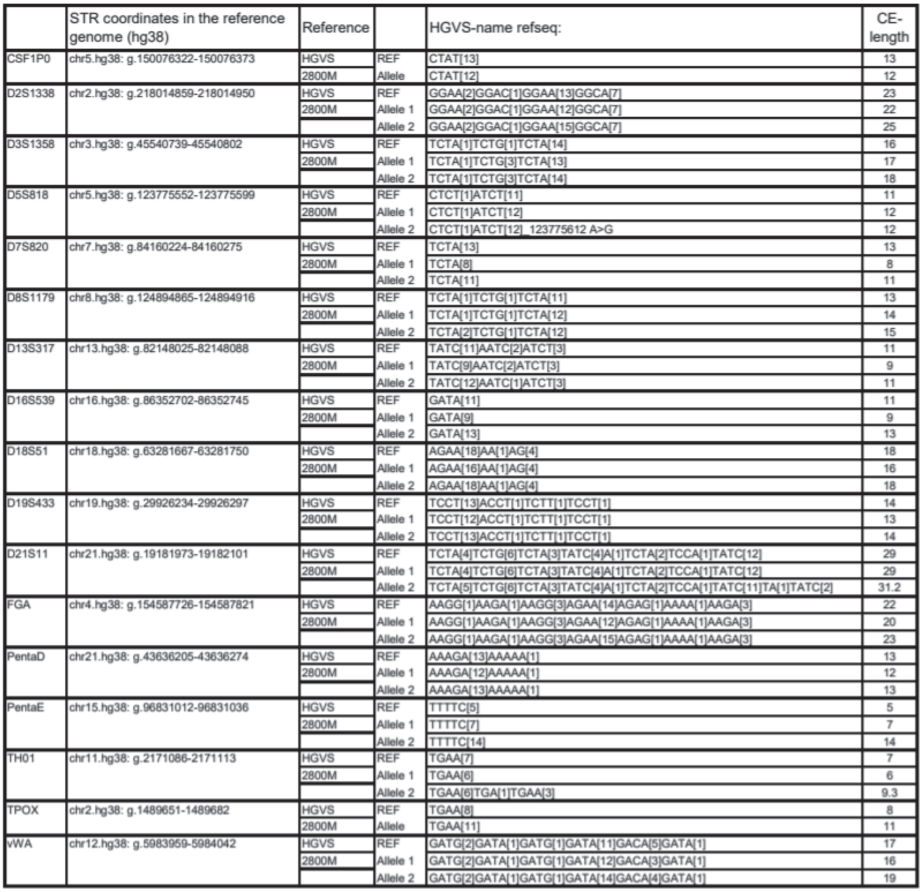
__A.__ Sequence profiles showing allele names after separate filtering steps used in the analysis for a single source sample with the final table used for interpretation to call genuine alleles in a reference sample. __B.__ Sequence profiles showing allele names after separate filtering steps used in the analysis for a mixed sample with the final table used for interpretation to call genuine alleles in a mixed sample.

Supplemental Figure 4. Sequence read profile for the two samples where PentaD showed a strong allelic imbalance



This figure displays the sequence read profile for PentaD for the two samples where a strong allelic imbalance was observed. For these samples, the number of reads for the allele with CE-length 11 was only observed for 8% and 12% of the reads of the second allele.

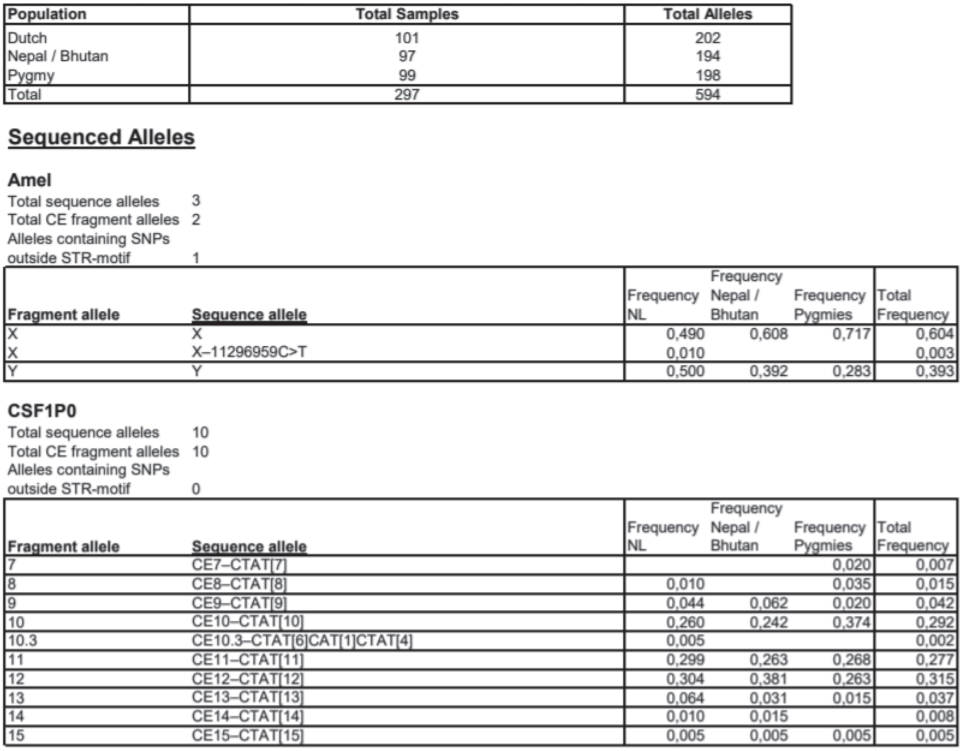
Supplemental Figure 5. STR sequence alleles for the reference genome and 2800M Control DNA



	STR coordinates in the reference genome (hg38)	Reference		HGVS-name refseq:	CE-length
CSF1P0	chr5.hg38: g.150076322-150076373	HGVS	REF	CTAT[13]	13
		2800M	Allele	CTAT[12]	12
D2S1338	chr2.hg38: g.218014859-218014950	HGVS	REF	GGAA[2]GGAC[1]GGAA[13]GGCA[7]	23
		2800M	Allele 1	GGAA[2]GGAC[1]GGAA[12]GGCA[7]	22
			Allele 2	GGAA[2]GGAC[1]GGAA[15]GGCA[7]	25
D3S1358	chr3.hg38: g.45540739-45540802	HGVS	REF	TCTA[1]TCTG[1]TCTA[14]	16
		2800M	Allele 1	TCTA[1]TCTG[3]TCTA[13]	17
			Allele 2	TCTA[1]TCTG[3]TCTA[14]	18
D5S818	chr5.hg38: g.123775552-123775599	HGVS	REF	CTCT[1]ATCT[11]	11
		2800M	Allele 1	CTCT[1]ATCT[12]	12
			Allele 2	CTCT[1]ATCT[12]_123775612 A>G	12
D7S820	chr7.hg38: g.84160224-84160275	HGVS	REF	TCTA[13]	13
		2800M	Allele 1	TCTA[8]	8
			Allele 2	TCTA[11]	11
D8S1179	chr8.hg38: g.124894865-124894916	HGVS	REF	TCTA[1]TCTG[1]TCTA[11]	13
		2800M	Allele 1	TCTA[1]TCTG[1]TCTA[12]	14
			Allele 2	TCTA[2]TCTG[1]TCTA[12]	15
D13S317	chr13.hg38: g.82148025-82148088	HGVS	REF	TATC[11]AATC[2]ATCT[3]	11
		2800M	Allele 1	TATC[9]AATC[2]ATCT[3]	9
			Allele 2	TATC[12]AATC[1]ATCT[3]	11
D16S539	chr16.hg38: g.86352702-86352745	HGVS	REF	GATA[11]	11
		2800M	Allele 1	GATA[9]	9
			Allele 2	GATA[13]	13
D18S51	chr18.hg38: g.63281667-63281750	HGVS	REF	AGAA[18]AA[1]AG[4]	18
		2800M	Allele 1	AGAA[16]AA[1]AG[4]	16
			Allele 2	AGAA[18]AA[1]AG[4]	18
D19S433	chr19.hg38: g.29926234-29926297	HGVS	REF	TCCT[13]ACCT[1]TCTT[1]TCCT[1]	14
		2800M	Allele 1	TCCT[12]ACCT[1]TCTT[1]TCCT[1]	13
			Allele 2	TCCT[13]ACCT[1]TCTT[1]TCCT[1]	14
D21S11	chr21.hg38: g.19181973-19182101	HGVS	REF	TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]	29
		2800M	Allele 1	TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]	29
			Allele 2	TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2]	31.2
FGA	chr4.hg38: g.154587726-154587821	HGVS	REF	AAGG[1]AAGA[1]AAGG[3]AGAA[14]AGAG[1]AAAA[1]AAGA[3]	22
		2800M	Allele 1	AAGG[1]AAGA[1]AAGG[3]AGAA[12]AGAG[1]AAAA[1]AAGA[3]	20
			Allele 2	AAGG[1]AAGA[1]AAGG[3]AGAA[15]AGAG[1]AAAA[1]AAGA[3]	23
PentaD	chr21.hg38: g.43636205-43636274	HGVS	REF	AAAGA[13]AAAAA[1]	13
		2800M	Allele 1	AAAGA[12]AAAAA[1]	12
			Allele 2	AAAGA[13]AAAAA[1]	13
PentaE	chr15.hg38: g.96831012-96831036	HGVS	REF	TTTTC[5]	5
		2800M	Allele 1	TTTTC[7]	7
			Allele 2	TTTTC[14]	14
TH01	chr11.hg38: g.2171086-2171113	HGVS	REF	TGAA[7]	7
		2800M	Allele 1	TGAA[6]	6
			Allele 2	TGAA[6]TGA[1]TGAA[3]	9.3
TPOX	chr2.hg38: g.1489651-1489682	HGVS	REF	TGAA[8]	8
		2800M	Allele	TGAA[11]	11
vWA	chr12.hg38: g.5983959-5984042	HGVS	REF	GATG[2]GATA[1]GATG[1]GATA[11]GACA[5]GATA[1]	17
		2800M	Allele 1	GATG[2]GATA[1]GATG[1]GATA[12]GACA[3]GATA[1]	16
			Allele 2	GATG[2]GATA[1]GATG[1]GATA[14]GACA[4]GATA[1]	19

Description of the coordinates of the STR motif for each STR and the alleles represented in the reference genome (hg38) and in 2800M Control DNA.

Supplemental Figure 6. Observed sequence variation in 297 analysed samples from three populations



Population	Total Samples	Total Alleles
Dutch	101	202
Nepal / Bhutan	97	194
Pygmy	99	198
Total	297	594

Sequenced Alleles

Amel

Total sequence alleles	3
Total CE fragment alleles	2
Alleles containing SNPs outside STR-motif	1

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
X	X	0,490	0,608	0,717	0,604
X	X–11296959C>T	0,010			0,003
Y	Y	0,500	0,392	0,283	0,393

CSF1P0

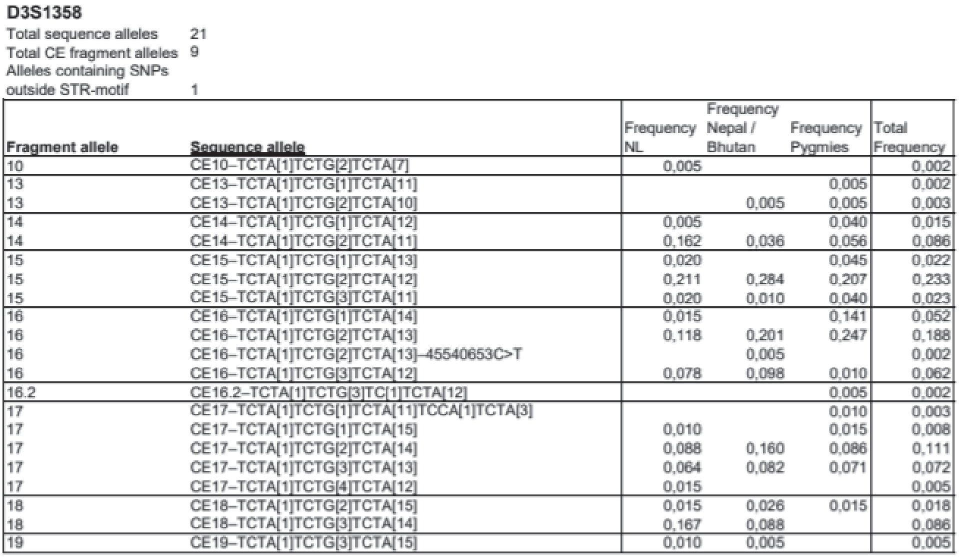
Total sequence alleles	10
Total CE fragment alleles	10
Alleles containing SNPs outside STR-motif	0

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
7	CE7–CTAT[7]			0,020	0,007
8	CE8–CTAT[8]	0,010		0,035	0,015
9	CE9–CTAT[9]	0,044	0,062	0,020	0,042
10	CE10–CTAT[10]	0,260	0,242	0,374	0,292
10.3	CE10.3–CTAT[6]CAT[1]CTAT[4]	0,005			0,002
11	CE11–CTAT[11]	0,299	0,263	0,268	0,277
12	CE12–CTAT[12]	0,304	0,381	0,263	0,315
13	CE13–CTAT[13]	0,064	0,031	0,015	0,037
14	CE14–CTAT[14]	0,010	0,015		0,008
15	CE15–CTAT[15]	0,005	0,005	0,005	0,005

D2S1338

Total sequence alleles	55
Total CE fragment alleles	12
Alleles containing SNPs outside STR-motif	14

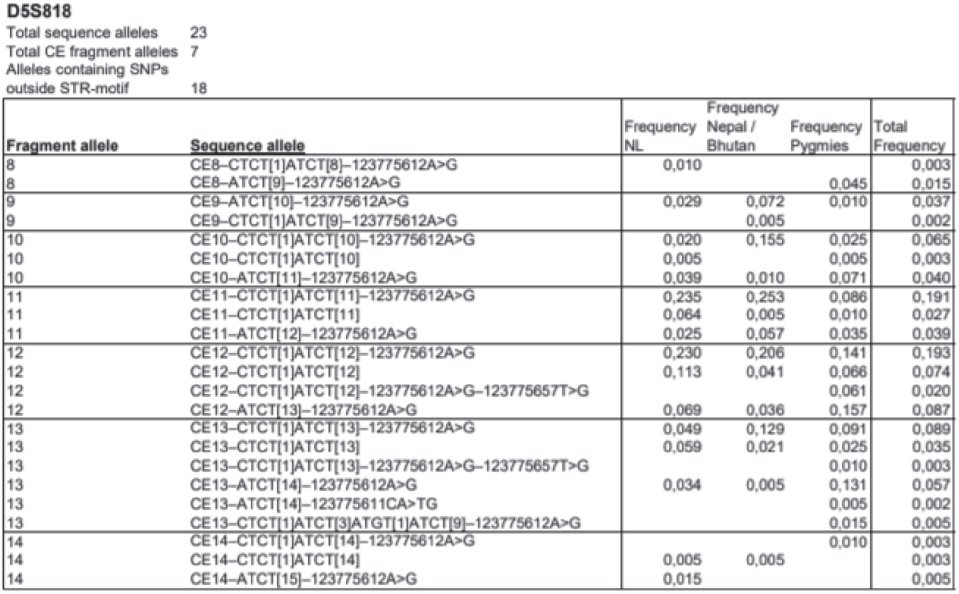
Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
14	CE14–GGAA[8]GGCA[6]–218014824C>A	0,005			0,002
16	CE16–GGAA[12]GGCA[4]–218014824C>A			0,066	0,022
16	CE16–GGAA[11]GGCA[5]			0,015	0,005
16	CE16–GGAA[10]GGCA[6]–218014824C>A	0,074			0,025
17	CE17–GGAA[12]GGCA[5]–218014824C>A			0,005	0,002
17	CE17–GGAA[11]GGCA[6]		0,005		0,002
17	CE17–GGAA[11]GGCA[6]–218014824C>A	0,230	0,041	0,015	0,097
18	CE18–GGAA[15]GGCA[3]–218014824C>A			0,005	0,002
18	CE18–GGAA[14]GGCA[4]–218014824C>A			0,010	0,003
18	CE18–GGAA[13]GGCA[5]		0,005		0,002
18	CE18–GGAA[12]GGCA[6]		0,010		0,003
18	CE18–GGAA[12]GGCA[6]–218014824C>A	0,074		0,005	0,027
18	CE18–GGAA[11]GGCA[7]	0,015	0,077	0,040	0,044
19	CE19–GGAA[15]GGCA[4]			0,005	0,002
19	CE19–GGAA[14]GGCA[5]	0,005	0,010	0,040	0,018
19	CE19–GGAA[2]GGAC[1]GGAA[10]GGCA[6]			0,005	0,002
19	CE19–GGAA[13]GGCA[6]		0,005	0,061	0,022
19	CE19–GGAA[13]GGCA[6]–218014824C>A	0,015			0,005
19	CE19–GGAA[12]GGCA[7]	0,059	0,170	0,106	0,111
19	CE19–GGAA[11]GGCA[8]			0,005	0,002
20	CE20–GGAA[17]GGCA[3]–218014824C>A			0,005	0,002
20	CE20–GGAA[16]GGCA[4]–218014824C>A			0,005	0,002
20	CE20–GGAA[14]GGCA[6]	0,020	0,010	0,071	0,034
20	CE20–GGAA[14]GGCA[6]–218014824C>A		0,005		0,002
20	CE20–GGAA[12]GGGA[1]GGCA[7]	0,015			0,005
20	CE20–GGAA[2]GGAC[1]GGAA[10]GGCA[7]	0,015	0,005		0,007
20	CE20–GGAA[13]GGCA[7]	0,088	0,108	0,056	0,084
20	CE20–GGAA[10]GAAA[1]GGAA[2]GGCA[7]		0,015		0,005
20	CE20–GGAA[12]GGCA[8]	0,005		0,005	0,003
20	CE20–GGAA[2]GGAC[1]GGAA[9]GGCA[8]		0,010		0,003
21	CE21–GGAA[17]GGCA[4]–218014824C>A			0,035	0,012
21	CE21–GGAA[2]GGAC[1]GGAA[13]GGCA[5]		0,005		0,002
21	CE21–GGAA[16]GGCA[5]			0,020	0,007
21	CE21–GGAA[2]GGAC[1]GGAA[12]GGCA[6]			0,030	0,010
21	CE21–GGAA[15]GGCA[6]			0,056	0,018
21	CE21–GGAA[2]GGAC[1]GGAA[11]GGCA[7]	0,005		0,010	0,005
21	CE21–GGAA[14]GGCA[7]	0,025	0,046	0,040	0,037
21	CE21–GGAA[13]GGCA[8]		0,010	0,030	0,013
22	CE22–GGAA[18]GGCA[4]–218014824C>A			0,005	0,002
22	CE22–GGAA[2]GGAC[1]GGAA[14]GGCA[5]			0,005	0,002
22	CE22–GGAA[2]GGAC[1]GGAA[13]GGCA[6]		0,010	0,040	0,017
22	CE22–GGAA[2]GGAC[1]GGAA[12]GGCA[7]	0,010	0,026	0,045	0,027
22	CE22–GGAA[15]GGCA[7]			0,005	0,002
22	CE22–GGAA[2]GGAC[1]GGAA[11]GGCA[8]			0,005	0,002
23	CE23–GGAA[2]GGAC[1]GGAA[15]GGCA[5]		0,010		0,003
23	CE23–GGAA[2]GGAC[1]GGAA[14]GGCA[6]	0,005	0,005		0,003
23	CE23–GGAA[2]GGAC[1]GGAA[13]GGCA[7]	0,069	0,180	0,025	0,091
23	CE23–GGAA[14]GGCA[9]			0,005	0,002
24	CE24–GGAA[2]GGAC[1]GGAA[15]GGCA[6]	0,005	0,005	0,010	0,007
24	CE24–GGAA[2]GGAC[1]GGAA[14]GGCA[7]	0,093	0,155	0,076	0,107
24	CE24–GGAA[2]GGAC[1]GGAA[13]GGCA[8]		0,005		0,002
25	CE25–GGAA[2]GGAC[1]GGAA[16]GGCA[6]	0,010	0,005		0,005
25	CE25–GGAA[2]GGAC[1]GGAA[15]GGCA[7]	0,142	0,046	0,020	0,070
25	CE25–GGAA[2]GGAC[1]GGAA[14]GGCA[8]	0,005	0,005		0,003
26	CE26–GGAA[2]GGAC[1]GGAA[16]GGCA[7]	0,015	0,005	0,010	0,010



D3S1358

Total sequence alleles	21
Total CE fragment alleles	9
Alleles containing SNPs outside STR-motif	1

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
10	CE10–TCTA[1]TCTG[2]TCTA[7]	0,005			0,002
13	CE13–TCTA[1]TCTG[1]TCTA[11]			0,005	0,002
13	CE13–TCTA[1]TCTG[2]TCTA[10]		0,005	0,005	0,003
14	CE14–TCTA[1]TCTG[1]TCTA[12]	0,005		0,040	0,015
14	CE14–TCTA[1]TCTG[2]TCTA[11]	0,162	0,036	0,056	0,086
15	CE15–TCTA[1]TCTG[1]TCTA[13]	0,020		0,045	0,022
15	CE15–TCTA[1]TCTG[2]TCTA[12]	0,211	0,284	0,207	0,233
15	CE15–TCTA[1]TCTG[3]TCTA[11]	0,020	0,010	0,040	0,023
16	CE16–TCTA[1]TCTG[1]TCTA[14]	0,015		0,141	0,052
16	CE16–TCTA[1]TCTG[2]TCTA[13]	0,118	0,201	0,247	0,188
16	CE16–TCTA[1]TCTG[2]TCTA[13]–45540653C>T		0,005		0,002
16	CE16–TCTA[1]TCTG[3]TCTA[12]	0,078	0,098	0,010	0,062
16.2	CE16.2–TCTA[1]TCTG[3]TC[1]TCTA[12]			0,005	0,002
17	CE17–TCTA[1]TCTG[1]TCTA[11]TCCA[1]TCTA[3]			0,010	0,003
17	CE17–TCTA[1]TCTG[1]TCTA[15]	0,010		0,015	0,008
17	CE17–TCTA[1]TCTG[2]TCTA[14]	0,088	0,160	0,086	0,111
17	CE17–TCTA[1]TCTG[3]TCTA[13]	0,064	0,082	0,071	0,072
17	CE17–TCTA[1]TCTG[4]TCTA[12]	0,015			0,005
18	CE18–TCTA[1]TCTG[2]TCTA[15]	0,015	0,026	0,015	0,018
18	CE18–TCTA[1]TCTG[3]TCTA[14]	0,167	0,088		0,086
19	CE19–TCTA[1]TCTG[3]TCTA[15]	0,010	0,005		0,005



D5S818

Total sequence alleles	23
Total CE fragment alleles	7
Alleles containing SNPs outside STR-motif	18

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
8	CE8–CTCT[1]ATCT[8]–123775612A>G	0,010			0,003
8	CE8–ATCT[9]–123775612A>G			0,045	0,015
9	CE9–ATCT[10]–123775612A>G	0,029	0,072	0,010	0,037
9	CE9–CTCT[1]ATCT[9]–123775612A>G		0,005		0,002
10	CE10–CTCT[1]ATCT[10]–123775612A>G	0,020	0,155	0,025	0,065
10	CE10–CTCT[1]ATCT[10]	0,005		0,005	0,003
10	CE10–ATCT[11]–123775612A>G	0,039	0,010	0,071	0,040
11	CE11–CTCT[1]ATCT[11]–123775612A>G	0,235	0,253	0,086	0,191
11	CE11–CTCT[1]ATCT[11]	0,064	0,005	0,010	0,027
11	CE11–ATCT[12]–123775612A>G	0,025	0,057	0,035	0,039
12	CE12–CTCT[1]ATCT[12]–123775612A>G	0,230	0,206	0,141	0,193
12	CE12–CTCT[1]ATCT[12]	0,113	0,041	0,066	0,074
12	CE12–CTCT[1]ATCT[12]–123775612A>G–123775657T>G			0,061	0,020
12	CE12–ATCT[13]–123775612A>G	0,069	0,036	0,157	0,087
13	CE13–CTCT[1]ATCT[13]–123775612A>G	0,049	0,129	0,091	0,089
13	CE13–CTCT[1]ATCT[13]	0,059	0,021	0,025	0,035
13	CE13–CTCT[1]ATCT[13]–123775612A>G–123775657T>G			0,010	0,003
13	CE13–ATCT[14]–123775612A>G	0,034	0,005	0,131	0,057
13	CE13–ATCT[14]–123775611CA>TG			0,005	0,002
13	CE13–CTCT[1]ATCT[3]ATGT[1]ATCT[9]–123775612A>G			0,015	0,005
14	CE14–CTCT[1]ATCT[14]–123775612A>G			0,010	0,003
14	CE14–CTCT[1]ATCT[14]	0,005	0,005		0,003
14	CE14–ATCT[15]–123775612A>G	0,015			0,005

D7S820

Total sequence alleles	32
Total CE fragment alleles	10
Alleles containing SNPs outside STR-motif	25

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
6	CE6–TCTA[6]–84160204T>A			0,005	0,002
7	CE7–TCTA[7]	0,025		0,005	0,010
7	CE7–TCTA[7]–84160204T>A		0,005		0,002
8	CE8–TCTA[8]	0,127	0,057	0,136	0,107
8	CE8–TCTA[8]–84160204T>A		0,005		0,002
8	CE8–TCTA[8]–84160204T>A–84160161A>C			0,010	0,003
8	CE8–TCTA[8]–84160204T>A–84160286G>A	0,044	0,191	0,066	0,099
9	CE9–TCTA[10]–84160219ATCT>del–84160204T>A	0,005			0,002
9	CE9–TCTA[9]	0,176	0,036	0,086	0,101
9	CE9–TCTA[9]–84160204T>A		0,010		0,003
9	CE9–TCTA[9]–84160204T>A–84160161A>C		0,005	0,025	0,010
9	CE9–TCTA[9]–84160204T>A–84160286G>A	0,015	0,072	0,056	0,047
10	CE10–TCTA[10]	0,162	0,098	0,222	0,161
10	CE10–TCTA[10]–84160204T>A	0,049	0,015		0,022
10	CE10–TCTA[10]–84160204T>A–84160161A>C			0,005	0,002
10	CE10–TCTA[10]–84160161A>C	0,010	0,031	0,035	0,025
10	CE10–TCTA[10]–84160204T>A–84160286G>A		0,005	0,020	0,008
10.3	CE10.3–TCTA[11]–84160204T>del	0,005			0,002
11	CE11–TCTA[11]	0,191	0,155	0,071	0,139
11	CE11–TCTA[11]–84160204T>A	0,015	0,041		0,018
11	CE11–TCTA[11]–84160204T>A–84160161A>C			0,010	0,003
11	CE11–TCTA[11]–84160161A>C		0,021	0,066	0,029
11	CE11–TCTA[11]–84160204T>A–84160286G>A		0,026	0,005	0,010
11	CE11–TCTA[8]CCTA[1]TCTA[2]–84160204T>A		0,010		0,003
12	CE12–TCTA[12]	0,108	0,175	0,066	0,116
12	CE12–TCTA[12]–84160204T>A	0,034	0,021		0,018
12	CE12–TCTA[12]–84160204T>A–84160161A>C			0,040	0,013
12	CE12–TCTA[12]–84160161A>C			0,030	0,010
12	CE12–TCTA[12]–84160204T>A–84160286G>A			0,035	0,012
13	CE13–TCTA[13]	0,015	0,015		0,010
13	CE13–TCTA[13]–84160204T>A	0,010	0,005		0,005
14	CE14–TCTA[14]–84160204T>A	0,010		0,005	0,005

D8S1179

Total sequence alleles	27
Total CE fragment alleles	11
Alleles containing SNPs outside STR-motif	1

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
8	CE8–TCTA[8]	0,010	0,005		0,005
9	CE9–TCTA[9]	0,010			0,003
10	CE10–TCTA[10]	0,088	0,072	0,010	0,057
11	CE11–TCTA[1]TCTG[1]TCTA[9]		0,005		0,002
11	CE11–TCTA[11]	0,083	0,036		0,040
11	CE11–TCTA[2]TCTG[1]TCTA[8]			0,030	0,010
12	CE12–TCTA[1]TCTG[1]TCTA[10]	0,010	0,046	0,005	0,020
12	CE12–TCTA[12]	0,137	0,062	0,020	0,074
12	CE12–TCTA[2]TCTG[1]TCTA[9]			0,086	0,029
12.1	CE12.1–TCTA[1]TCTG[1]TCTA[9]TCTTA[1]	0,005			0,002
13	CE13–TCTA[1]TCTG[1]TCTA[11]	0,260	0,155	0,106	0,174
13	CE13–TCTA[13]	0,064	0,062		0,042
13	CE13–TCTA[2]TCTG[1]TCTA[10]			0,071	0,023
14	CE14–TCTA[1]TCTG[1]TCTA[12]	0,142	0,139	0,172	0,151
14	CE14–TCTA[1]TCTG[1]TGTA[1]TCTA[11]	0,005			0,002
14	CE14–TCTA[14]	0,020	0,005	0,005	0,010
14	CE14–TCTA[2]TCTG[1]TCTA[11]	0,049	0,062	0,192	0,101
15	CE15–TCTA[1]TCTG[1]TCTA[13]	0,044	0,031	0,061	0,045
15	CE15–TCTA[2]TCTG[1]TCTA[12]	0,059	0,180	0,116	0,117
15	CE15–TCTA[2]TCTG[2]TCTA[11]			0,030	0,010
16	CE16–TCTA[1]TCTG[1]TCTA[14]		0,005		0,002
16	CE16–TCTA[1]TCTG[3]TCTA[12]			0,010	0,003
16	CE16–TCTA[2]TCTG[1]TCTA[13]	0,010	0,113	0,056	0,059
16	CE16–TCTA[2]TCTG[1]TCTA[13]–124894972G>A		0,010		0,003
16	CE16–TCTA[2]TCTG[2]TCTA[12]			0,010	0,003
17	CE17–TCTA[1]TCTG[3]TCTA[13]			0,005	0,002
17	CE17–TCTA[2]TCTG[1]TCTA[14]	0,005	0,010	0,015	0,010

D13S317

Total sequence alleles	31
Total CE fragment alleles	8
Alleles containing SNPs outside STR-motif	10

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
7	CE7–TATC[7]AATC[2]ATCT[3]	0,005			0,002
7	CE7–TATC[8]AATC[1]ATCT[3]		0,010		0,003
8	CE8–TATC[8]AATC[2]ATCT[3]	0,083	0,175	0,030	0,096
9	CE9–TATC[10]AATC[1]ATCT[3]	0,010	0,077		0,029
9	CE9–TATC[9]AATC[2]ATCT[3]	0,083	0,139	0,015	0,079
10	CE10–TATC[10]AATC[2]ATCT[3]	0,049	0,041	0,005	0,032
10	CE10–TATC[11]AATC[1]ATCT[3]		0,124	0,020	0,047
10	CE10–TATC[12]ATCT[3]		0,005		0,002
10	CE10–TATC[12]AATC[1]ATCT[3]–82148097GTCT>del			0,005	0,002
11	CE11–TATC[11]AATC[2]ATCT[3]	0,103	0,036	0,096	0,079
11	CE11–TATC[12]AATC[1]ATCT[3]	0,181	0,149	0,157	0,163
11	CE11–TATC[12]AATC[1]ATCT[3]–82148000C>T	0,039	0,005		0,015
11	CE11–TATC[12]AATC[1]ATCT[3]–82147972G>A			0,020	0,007
11	CE11–TATC[13]ATCT[3]		0,021		0,007
11	CE11–TATC[8]TGTC[1]TATC[3]AATC[1]ATCT[3]		0,021		0,007
12	CE12–TATC[12]AATC[2]ATCT[3]	0,167	0,026	0,308	0,168
12	CE12–TATC[13]AATC[1]ATCT[3]	0,127	0,124	0,111	0,121
12	CE12–TATC[13]AATC[1]ATCT[3]–82148001G>A			0,035	0,012
12	CE12–TATC[13]AATC[1]ATCT[3]–82148000C>T	0,005			0,002
12	CE12–TATC[13]AATC[1]ATCT[3]–82147972G>A			0,035	0,012
12	CE12–TATC[14]ATCT[3]		0,021		0,007
12	CE12–TATC[7]TATT[1]TATC[5]AATC[1]ATCT[3]			0,010	0,003
13	CE13–TATC[13]AATC[2]ATCT[3]	0,039		0,116	0,052
13	CE13–TATC[14]AATC[1]ATCT[3]	0,034	0,005	0,005	0,015
13	CE13–TATC[14]AATC[1]ATCT[3]–82148001G>A			0,005	0,002
13	CE13–TATC[14]AATC[1]ATCT[3]–82148000C>T		0,005		0,002
13	CE13–TATC[14]AATC[1]ATCT[3]–82147972G>A			0,005	0,002
13	CE13–TATC[15]ATCT[3]		0,010		0,003
13	CE13–TATC[15]AATC[1]ATCT[3]–82148097GTCT>del			0,010	0,003
14	CE14–TATC[14]AATC[2]ATCT[3]	0,074		0,010	0,029
14	CE14–TATC[15]AATC[1]ATCT[3]		0,005		0,002

D16S539

Total sequence alleles	17
Total CE fragment alleles	9
Alleles containing SNPs outside STR-motif	9

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
5	CE5–GATA[5]			0,040	0,013
8	CE8–GATA[8]	0,010	0,010	0,010	0,010
9	CE9–GATA[9]	0,186	0,278	0,162	0,208
9	CE9–GATA[9]–86352607A>C			0,096	0,032
10	CE10–GATA[10]	0,039	0,098	0,116	0,084
10	CE10–GATA[10]–86352607A>C			0,040	0,013
11	CE11–GATA[11]	0,309	0,237	0,101	0,216
11	CE11–GATA[11]–86352607A>C	0,044	0,067	0,207	0,106
11	CE11–GATA[5]GACA[1]GATA[5]–86352607A>C		0,010		0,003
11	CE11–GATA[11]–86352607A>C–86352571A>C			0,005	0,002
12	CE12–GATA[12]	0,235	0,077	0,051	0,122
12	CE12–GATA[12]–86352584G>C	0,005			0,002
12	CE12–GATA[12]–86352607A>C	0,020	0,119	0,081	0,072
13	CE13–GATA[13]	0,118	0,021	0,051	0,064
13	CE13–GATA[13]–86352607A>C	0,029	0,062	0,035	0,042
14	CE14–GATA[14]	0,005		0,005	0,003
14	CE14–GATA[14]–86352607A>C		0,021		0,007

D18S51

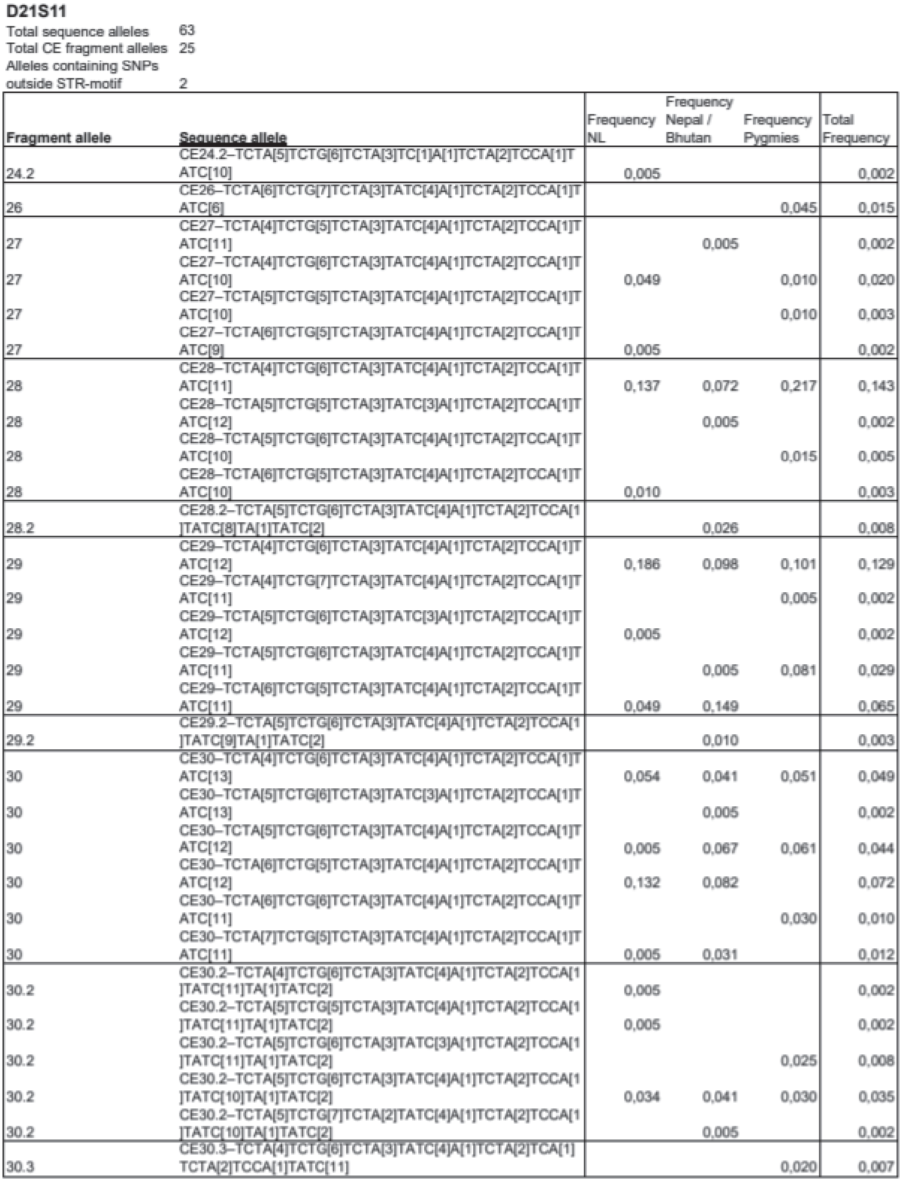
Total sequence alleles	19
Total CE fragment alleles	15
Alleles containing SNPs outside STR-motif	0

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
10	D18S51–CE10–AGAA[10]AA[1]AG[4]	0,005			0,002
11	D18S51–CE11–AGAA[11]AA[1]AG[4]	0,020	0,005		0,008
12	D18S51–CE12–AGAA[12]AA[1]AG[4]	0,167	0,036	0,005	0,070
13	D18S51–CE13–AGAA[13]AA[1]AG[4]	0,069	0,227	0,025	0,106
14	D18S51–CE14–AGAA[1]AGCA[1]AGAA[12]AA[1]AG[4]	0,015			0,005
14	D18S51–CE14–AGAA[14]AA[1]AG[4]	0,132	0,108	0,061	0,101
15	D18S51–CE15–AGAA[15]AA[1]AG[4]	0,162	0,180	0,131	0,158
16	D18S51–CE16–AGAA[13]AGAT[1]AGAA[2]AA[1]AG[4]			0,020	0,007
16	D18S51–CE16–AGAA[16]AA[1]AG[4]	0,206	0,129	0,217	0,185
16	D18S51–CE16–AGAA[16]AG[5]	0,005		0,005	0,003
17	D18S51–CE17–AGAA[17]AA[1]AG[4]	0,098	0,057	0,197	0,117
18	D18S51–CE18–AGAA[14]GGAA[1]AGAA[3]AA[1]AG[4]			0,005	0,002
18	D18S51–CE18–AGAA[18]AA[1]AG[4]	0,064	0,067	0,126	0,086
19	D18S51–CE19–AGAA[19]AA[1]AG[4]	0,025	0,103	0,141	0,089
20	D18S51–CE20–AGAA[20]AA[1]AG[4]	0,015	0,041	0,020	0,025
21	D18S51–CE21–AGAA[21]AA[1]AG[4]	0,010	0,015	0,040	0,022
22	D18S51–CE22–AGAA[22]AA[1]AG[4]	0,010	0,015		0,008
23	D18S51–CE23–AGAA[23]AA[1]AG[4]		0,010	0,005	0,005
25	D18S51–CE25–AGAA[25]AA[1]AG[4]		0,005		0,002

D19S433

Total sequence alleles	20
Total CE fragment alleles	17
Alleles containing SNPs outside STR-motif	0

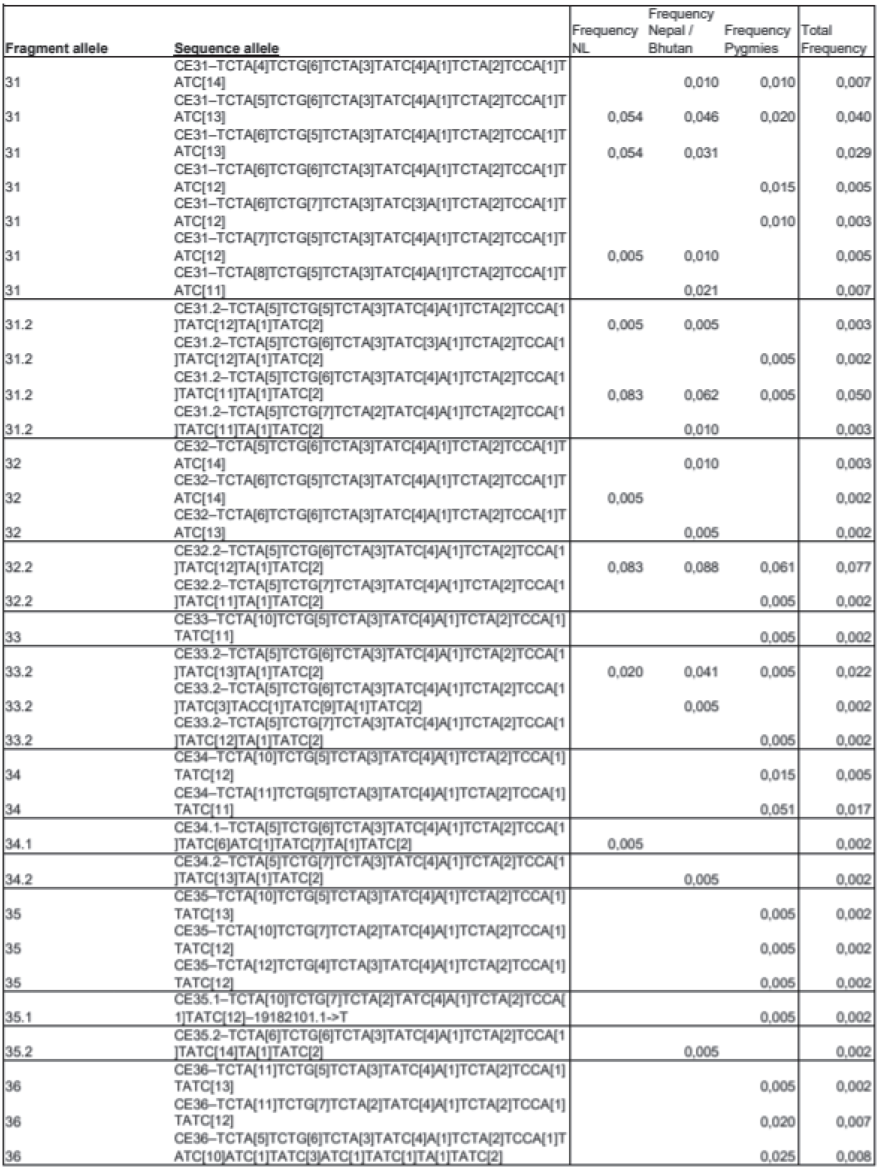
Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
9	CE9–TCCT[8]ACCT[1]TCTT[1]TCCT[1]			0,015	0,005
10	CE10–TCCT[9]ACCT[1]TCTT[1]TCCT[1]			0,177	0,059
10.2	CE10.2–TCCT[10]ACCT[1]TT[1]TCCT[1]			0,005	0,002
11	CE11–TCCT[10]ACCT[1]TCTT[1]TCCT[1]			0,045	0,015
11.2	CE11.2–TCCT[11]ACCT[1]TT[1]TCCT[1]		0,005		0,002
12	CE12–TCCT[11]ACCT[1]TCTT[1]TCCT[1]	0,054	0,041	0,081	0,059
12.2	CE12.2–TCCT[12]ACCT[1]TT[1]TCCT[1]		0,015	0,025	0,013
13	CE13–TCCT[12]ACCT[1]TCTT[1]TCCT[1]	0,260	0,242	0,167	0,223
13	CE13–TCCT[13]TCTT[1]TCCT[1]	0,005			0,002
13.2	CE13.2–TCCT[13]ACCT[1]TT[1]TCCT[1]	0,020	0,036	0,111	0,055
14	CE14–TCCT[13]ACCT[1]TCTT[1]TCCT[1]	0,328	0,237	0,136	0,235
14	CE14–TCCT[9]TCCC[1]TCCT[3]ACCT[1]TCTT[1]TCCT[1]		0,010		0,003
14	CE14–TCCT[14]TCTT[1]TCCT[1]	0,005			0,002
14.2	CE14.2–TCCT[14]ACCT[1]TT[1]TCCT[1]	0,025	0,082	0,051	0,052
15	CE15–TCCT[14]ACCT[1]TCTT[1]TCCT[1]	0,201	0,077	0,020	0,101
15.2	CE15.2–TCCT[15]ACCT[1]TT[1]TCCT[1]	0,039	0,170	0,101	0,102
16	CE16–TCCT[15]ACCT[1]TCTT[1]TCCT[1]	0,029	0,031	0,025	0,029
16.2	CE16.2–TCCT[16]ACCT[1]TT[1]TCCT[1]	0,020	0,046	0,035	0,034
17	CE17–TCCT[16]ACCT[1]TCTT[1]TCCT[1]	0,010	0,005		0,005
17.2	CE17.2–TCCT[17]ACCT[1]TT[1]TCCT[1]	0,005		0,005	0,003



D21S11

Total sequence alleles	63
Total CE fragment alleles	25
Alleles containing SNPs outside STR-motif	2

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
24.2	CE24.2–TCTA[5]TCTG[6]TCTA[3]TC[1]A[1]TCTA[2]TCCA[1]TATC[10]	0,005			0,002
26	CE26–TCTA[6]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[6]			0,045	0,015
27	CE27–TCTA[4]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]		0,005		0,002
27	CE27–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]	0,049		0,010	0,020
27	CE27–TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]			0,010	0,003
27	CE27–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[9]	0,005			0,002
28	CE28–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]	0,137	0,072	0,217	0,143
28	CE28–TCTA[5]TCTG[5]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[12]		0,005		0,002
28	CE28–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]			0,015	0,005
28	CE28–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]	0,010			0,003
28.2	CE28.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[8]TA[1]TATC[2]		0,026		0,008
29	CE29–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]	0,186	0,098	0,101	0,129
29	CE29–TCTA[4]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]			0,005	0,002
29	CE29–TCTA[5]TCTG[6]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[12]	0,005			0,002
29	CE29–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]		0,005	0,081	0,029
29	CE29–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]	0,049	0,149		0,065
29.2	CE29.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[9]TA[1]TATC[2]		0,010		0,003
30	CE30–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]	0,054	0,041	0,051	0,049
30	CE30–TCTA[5]TCTG[6]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[13]		0,005		0,002
30	CE30–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]	0,005	0,067	0,061	0,044
30	CE30–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]	0,132	0,082		0,072
30	CE30–TCTA[6]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]			0,030	0,010
30	CE30–TCTA[7]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]	0,005	0,031		0,012
30.2	CE30.2–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2]	0,005			0,002
30.2	CE30.2–TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2]	0,005			0,002
30.2	CE30.2–TCTA[5]TCTG[6]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2]			0,025	0,008
30.2	CE30.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]TA[1]TATC[2]	0,034	0,041	0,030	0,035
30.2	CE30.2–TCTA[5]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]TA[1]TATC[2]		0,005		0,002
30.3	CE30.3–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCA[1]TCTA[2]TCCA[1]TATC[11]			0,020	0,007

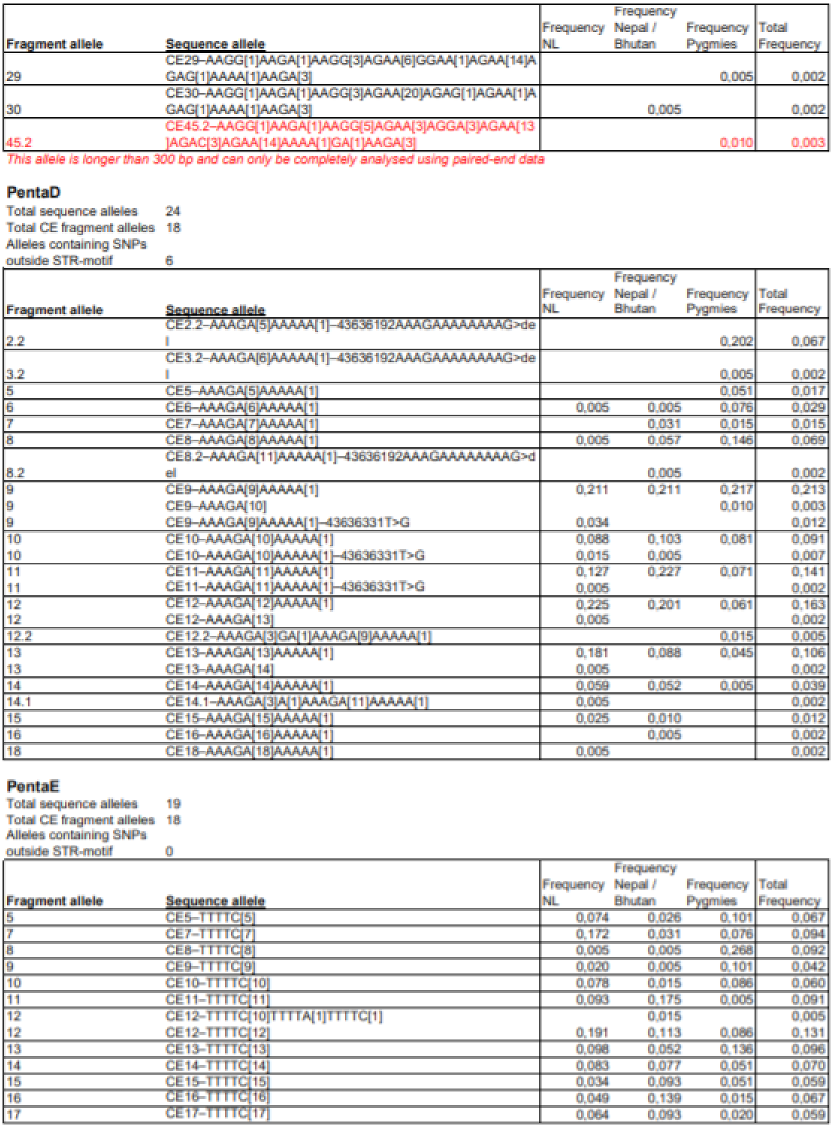


Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
31	CE31–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[14]		0,010	0,010	0,007
31	CE31–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]	0,054	0,046	0,020	0,040
31	CE31–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]	0,054	0,031		0,029
31	CE31–TCTA[6]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]			0,015	0,005
31	CE31–TCTA[6]TCTG[7]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[12]			0,010	0,003
31	CE31–TCTA[7]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]	0,005	0,010		0,005
31	CE31–TCTA[8]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]		0,021		0,007
31.2	CE31.2–TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2]	0,005	0,005		0,003
31.2	CE31.2–TCTA[5]TCTG[6]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2]			0,005	0,002
31.2	CE31.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2]	0,083	0,062	0,005	0,050
31.2	CE31.2–TCTA[5]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2]		0,010		0,003
32	CE32–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[14]		0,010		0,003
32	CE32–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[14]	0,005			0,002
32	CE32–TCTA[6]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]		0,005		0,002
32.2	CE32.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2]	0,083	0,088	0,061	0,077
32.2	CE32.2–TCTA[5]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2]			0,005	0,002
33	CE33–TCTA[10]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]			0,005	0,002
33.2	CE33.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]TA[1]TATC[2]	0,020	0,041	0,005	0,022
33.2	CE33.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[3]TACC[1]TATC[9]TA[1]TATC[2]		0,005		0,002
33.2	CE33.2–TCTA[5]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2]			0,005	0,002
34	CE34–TCTA[10]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]			0,015	0,005
34	CE34–TCTA[11]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]			0,051	0,017
34.1	CE34.1–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[6]ATC[1]TATC[7]TA[1]TATC[2]	0,005			0,002
34.2	CE34.2–TCTA[5]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]TA[1]TATC[2]		0,005		0,002
35	CE35–TCTA[10]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]			0,005	0,002
35	CE35–TCTA[10]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]			0,005	0,002
35	CE35–TCTA[12]TCTG[4]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]			0,005	0,002
35.1	CE35.1–TCTA[10]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]–19182101.1->T			0,005	0,002
35.2	CE35.2–TCTA[6]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[14]TA[1]TATC[2]		0,005		0,002
36	CE36–TCTA[11]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]			0,005	0,002
36	CE36–TCTA[11]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]			0,020	0,007
36	CE36–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]ATC[1]TATC[3]ATC[1]TATC[1]TA[1]TATC[2]			0,025	0,008

FGA

Total sequence alleles	34
Total CE fragment alleles	22
Alleles containing SNPs outside STR-motif	0

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
17	CE17–AAGG[1]AAGA[1]AAGG[3]AGAA[9]AGAG[1]AAAA[1]AAGA[3]			0,040	0,013
18	CE18–AAGG[1]AAGA[1]AAGG[3]AGAA[10]AGAG[1]AAAA[1]AAGA[3]		0,077		0,025
19	CE19–AAGG[1]AAGA[1]AAGG[3]AGAA[11]AGAG[1]AAAA[1]AAGA[3]	0,059	0,093	0,056	0,069
19.2	CE19.2–AAGG[1]AAGA[1]AAGG[3]AGAA[13]AAAA[1]GA[1]AAGA[2]	0,005		0,005	0,003
20	CE20–AAGG[1]AAGA[1]AAGG[3]AGAA[12]AGAG[1]AAAA[1]AAGA[3]	0,108	0,046	0,051	0,069
21	CE21–AAGG[1]AAGA[1]AAGG[3]AGAA[13]AGAG[1]AAAA[1]AAGA[3]	0,176	0,052	0,086	0,106
21.2	CE21.2–AAGG[1]AAGA[1]AAGG[3]AGAA[16]GA[1]AAGA[2]		0,005		0,002
21.2	CE21.2–AAGG[1]AAGA[1]AAGG[3]AGAA[15]AAAA[1]GA[1]AAGA[2]	0,005	0,010		0,005
22	CE22–AAGG[1]AAGA[1]AAGG[3]AGAA[14]AGAG[1]AAAA[1]AAGA[3]	0,196	0,103	0,288	0,196
22.1	CE22.1–AAGG[1]AAGA[1]AAGG[3]AGAA[14][A[1]AGAG[1]AAAA[1]AAGA[3]	0,005			0,002
22.2	CE22.2–AAGG[1]AAGA[1]AAGG[3]AGAA[16]AAAA[1]GA[1]AAGA[2]	0,005	0,031		0,012
22.2	CE22.2–AAGG[1]AAGA[1]AAGG[3]AGAA[14]AAAG[1]AGAA[1]AAAA[1]GA[1]AAGA[2]			0,020	0,007
23	CE23–AAGG[1]AAGA[1]AAGG[3]AGAA[15]AGAG[1]AAAA[1]AAGA[3]	0,157	0,186	0,167	0,169
23.2	CE23.2–AAGG[1]AAGA[1]AAGG[3]AGAA[17]AAAA[1]GA[1]AAGA[2]	0,005	0,005		0,003
24	CE24–AAGG[1]AAGA[1]AAGG[3]AGAA[16]AGAG[1]AAAA[1]AAGA[3]	0,137	0,180	0,126	0,148
24.2	CE24.2–AAGG[1]AAGA[1]AAGG[3]AGAA[18]AAAA[1]GA[1]AAGA[2]		0,026		0,008
24.2	CE24.2–AAGG[1]AAGA[1]AAGG[3]AGAA[16]AAAG[1]AGAA[1]AAAA[1]GA[1]AAGA[2]			0,010	0,003
25	CE25–AAGG[1]AAGA[1]AAGG[3]AGAA[2]ACAA[1]AGAA[14]AGAG[1]AAAA[1]AAGA[3]		0,005		0,002
25	CE25–AAGG[1]AAGA[1]AAGG[3]AGAA[15]AGAG[1]AGAA[1]AGAG[1]AAAA[1]AAGA[3]	0,005			0,002
25	CE25–AAGG[1]AAGA[1]AAGG[3]AGAA[17]AGAG[1]AAAA[1]AAGA[3]	0,093	0,077	0,051	0,074
25.2	CE25.2–AAGG[1]AAGA[1]AAGG[3]AGAA[19]AAAA[1]GA[1]AAGA[2]		0,005		0,002
26	CE26–AAGG[1]AAGA[1]AAGG[3]AGAA[16]AGAG[1]AGAA[1]AGAG[1]AAAA[1]AAGA[3]		0,005		0,002
26	CE26–AAGG[1]AAGA[1]AAGG[3]AGAA[6]GGAA[1]AGAA[11]AGAG[1]AAAA[1]AAGA[3]			0,005	0,002
26	CE26–AAGG[1]AAGA[1]AAGG[3]AGAA[2]ACAA[1]AGAA[15]AGAG[1]AAAA[1]AAGA[3]	0,010			0,003
26	CE26–AAGG[1]AAGA[1]AAGG[3]AGAA[18]AGAG[1]AAAA[1]AAGA[3]	0,029	0,046	0,040	0,039
26.3	CE26.3–AAGG[1]AAGA[1]AAGG[3]AGAA[10]GAA[1]AGAA[8]AGAG[1]AAAA[1]AAGA[3]		0,005		0,002
27	CE27–AAGG[1]AAGA[1]AAGG[3]AGAA[17]AGAG[1]AGAA[1]AGAG[1]AAAA[1]AAGA[3]		0,015		0,005
27	CE27–AAGG[1]AAGA[1]AAGG[3]AGAA[6]GGAA[1]AGAA[12]AGAG[1]AAAA[1]AAGA[3]			0,010	0,003
27	CE27–AAGG[1]AAGA[1]AAGG[3]AGAA[2]ACAA[1]AGAA[16]AGAG[1]AAAA[1]AAGA[3]	0,005			0,002
27	CE27–AAGG[1]AAGA[1]AAGG[3]AGAA[19]AGAG[1]AAAA[1]AAGA[3]		0,010	0,015	0,008



Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
29	CE29–AAGG[1]AAGA[1]AAGG[3]AGAA[6]GGAA[1]AGAA[14]AGAG[1]AAAA[1]AAGA[3]			0,005	0,002
30	CE30–AAGG[1]AAGA[1]AAGG[3]AGAA[20]AGAG[1]AGAA[1]AGAG[1]AAAA[1]AAGA[3]		0,005		0,002
45.2	CE45.2–AAGG[1]AAGA[1]AAGG[5]AGAA[3]AGGA[3]AGAA[13]AGAC[3]AGAA[14]AAAA[1]GA[1]AAGA[3]			0,010	0,003

This allele is longer than 300 bp and can only be completely analysed using paired-end data

PentaD

Total sequence alleles	24
Total CE fragment alleles	18
Alleles containing SNPs outside STR-motif	6

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
2.2	CE2.2–AAAGA[5]AAAAA[1]–43636192AAAGAAAAAAAAAG>del			0,202	0,067
3.2	CE3.2–AAAGA[6]AAAAA[1]–43636192AAAGAAAAAAAAAG>del			0,005	0,002
5	CE5–AAAGA[5]AAAAA[1]			0,051	0,017
6	CE6–AAAGA[6]AAAAA[1]	0,005	0,005	0,076	0,029
7	CE7–AAAGA[7]AAAAA[1]		0,031	0,015	0,015
8	CE8–AAAGA[8]AAAAA[1]	0,005	0,057	0,146	0,069
8.2	CE8.2–AAAGA[11]AAAAA[1]–43636192AAAGAAAAAAAAAG>del		0,005		0,002
9	CE9–AAAGA[9]AAAAA[1]	0,211	0,211	0,217	0,213
9	CE9–AAAGA[10]			0,010	0,003
9	CE9–AAAGA[9]AAAAA[1]–43636331T>G	0,034			0,012
10	CE10–AAAGA[10]AAAAA[1]	0,088	0,103	0,081	0,091
10	CE10–AAAGA[10]AAAAA[1]–43636331T>G	0,015	0,005		0,007
11	CE11–AAAGA[11]AAAAA[1]	0,127	0,227	0,071	0,141
11	CE11–AAAGA[11]AAAAA[1]–43636331T>G	0,005			0,002
12	CE12–AAAGA[12]AAAAA[1]	0,225	0,201	0,061	0,163
12	CE12–AAAGA[13]	0,005			0,002
12.2	CE12.2–AAAGA[3]GA[1]AAAGA[9]AAAAA[1]			0,015	0,005
13	CE13–AAAGA[13]AAAAA[1]	0,181	0,088	0,045	0,106
13	CE13–AAAGA[14]	0,005			0,002
14	CE14–AAAGA[14]AAAAA[1]	0,059	0,052	0,005	0,039
14.1	CE14.1–AAAGA[3]A[1]AAAGA[11]AAAAA[1]	0,005			0,002
15	CE15–AAAGA[15]AAAAA[1]	0,025	0,010		0,012
16	CE16–AAAGA[16]AAAAA[1]		0,005		0,002
18	CE18–AAAGA[18]AAAAA[1]	0,005			0,002

PentaE

Total sequence alleles	19
Total CE fragment alleles	18
Alleles containing SNPs outside STR-motif	0

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
5	CE5–TTTTC[5]	0,074	0,026	0,101	0,067
7	CE7–TTTTC[7]	0,172	0,031	0,076	0,094
8	CE8–TTTTC[8]	0,005	0,005	0,268	0,092
9	CE9–TTTTC[9]	0,020	0,005	0,101	0,042
10	CE10–TTTTC[10]	0,078	0,015	0,086	0,060
11	CE11–TTTTC[11]	0,093	0,175	0,005	0,091
12	CE12–TTTTC[10]TTTTA[1]TTTTC[1]		0,015		0,005
12	CE12–TTTTC[12]	0,191	0,113	0,086	0,131
13	CE13–TTTTC[13]	0,098	0,052	0,136	0,096
14	CE14–TTTTC[14]	0,083	0,077	0,051	0,070
15	CE15–TTTTC[15]	0,034	0,093	0,051	0,059
16	CE16–TTTTC[16]	0,049	0,139	0,015	0,067
17	CE17–TTTTC[17]	0,064	0,093	0,020	0,059

TH01

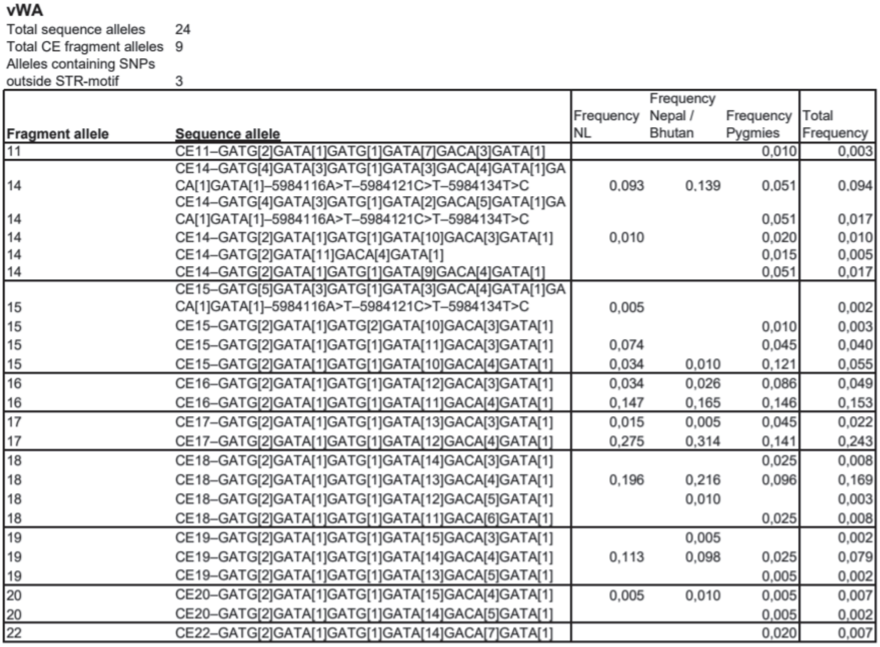
Total sequence alleles	14
Total CE fragment alleles	8
Alleles containing SNPs outside STR-motif	5

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
5	CE5–TGAA[5]	0,010			0,003
6	CE6–TGAA[6]	0,201	0,046	0,056	0,102
7	CE7–TGAA[1]TTAA[1]TGAA[5]			0,005	0,002
7	CE7–TGAA[7]	0,196	0,258	0,404	0,285
7	CE7–TGAA[7]–2171244C>T			0,025	0,008
7	CE7–TGAA[7]–2171200C>A		0,005		0,002
8	CE8–TGAA[8]–2171115G>T–2171244C>T			0,015	0,005
8	CE8–TGAA[8]	0,127	0,098	0,116	0,114
8	CE8–TGAA[8]–2171244C>T			0,101	0,034
9	CE9–TGAA[9]	0,123	0,479	0,197	0,263
9	CE9–TGAA[9]–2171244C>T			0,010	0,003
9.3	CE9.3–TGAA[6]TGA[1]TGAA[3]	0,333	0,103	0,020	0,154
10	CE10–TGAA[10]	0,010	0,010	0,035	0,018
11	CE11–TGAA[11]			0,015	0,005

TPOX

Total sequence alleles	12
Total CE fragment alleles	7
Alleles containing SNPs outside STR-motif	5

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
6	CE6–TGAA[6]			0,091	0,030
8	CE8–TGAA[8]	0,461	0,443	0,273	0,393
8	CE8–TGAA[8]–1489601G>T	0,005	0,021		0,008
8	CE8–TGAA[8]–1489557G>A			0,030	0,010
8	CE8–TGAA[8]–1489556C>T			0,121	0,040
9	CE9–TGAA[9]	0,113	0,129	0,152	0,131
9	CE9–TGAA[9]–1489544C>A			0,111	0,037
9	CE9–TGAA[9]–1489557G>A			0,015	0,005
10	CE10–TGAA[10]	0,064	0,015	0,076	0,052
11	CE11–TGAA[11]	0,333	0,371	0,121	0,275
12	CE12–TGAA[12]	0,020	0,021	0,010	0,017
13	CE13–TGAA[13]	0,005			0,002



vWA

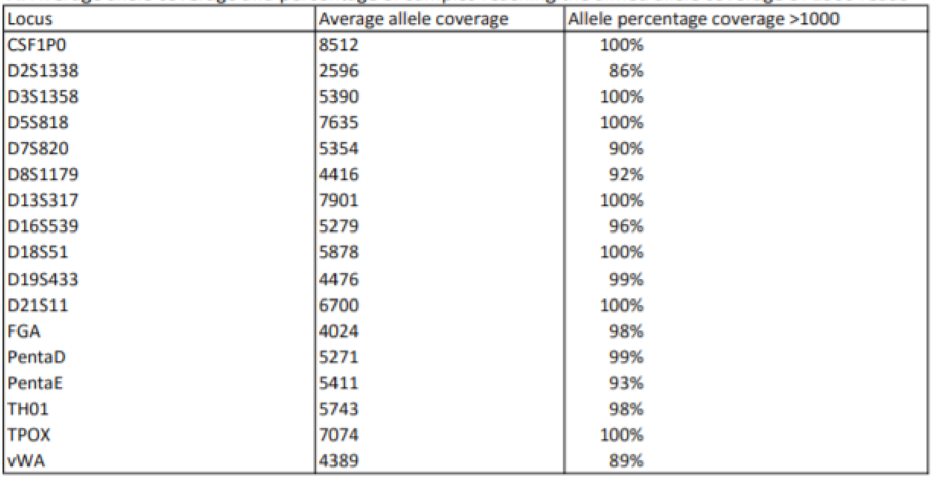
Total sequence alleles	24
Total CE fragment alleles	9
Alleles containing SNPs outside STR-motif	3

Fragment allele	Sequence allele	Frequency NL	Frequency Nepal / Bhutan	Frequency Pygmies	Total Frequency
11	CE11–GATG[2]GATA[1]GATG[1]GATA[7]GACA[3]GATA[1]			0,010	0,003
14	CE14–GATG[4]GATA[3]GATG[1]GATA[3]GACA[4]GATA[1]GACA[1]GATA[1]–5984116A>T–5984121C>T–5984134T>C	0,093	0,139	0,051	0,094
14	CE14–GATG[4]GATA[3]GATG[1]GATA[2]GACA[5]GATA[1]GACA[1]GATA[1]–5984116A>T–5984121C>T–5984134T>C			0,051	0,017
14	CE14–GATG[2]GATA[1]GATG[1]GATA[10]GACA[3]GATA[1]	0,010		0,020	0,010
14	CE14–GATG[2]GATA[11]GACA[4]GATA[1]			0,015	0,005
14	CE14–GATG[2]GATA[1]GATG[1]GATA[9]GACA[4]GATA[1]			0,051	0,017
15	CE15–GATG[5]GATA[3]GATG[1]GATA[3]GACA[4]GATA[1]GACA[1]GATA[1]–5984116A>T–5984121C>T–5984134T>C	0,005			0,002
15	CE15–GATG[2]GATA[1]GATG[2]GATA[10]GACA[3]GATA[1]			0,010	0,003
15	CE15–GATG[2]GATA[1]GATG[1]GATA[11]GACA[3]GATA[1]	0,074		0,045	0,040
15	CE15–GATG[2]GATA[1]GATG[1]GATA[10]GACA[4]GATA[1]	0,034	0,010	0,121	0,055
16	CE16–GATG[2]GATA[1]GATG[1]GATA[12]GACA[3]GATA[1]	0,034	0,026	0,086	0,049
16	CE16–GATG[2]GATA[1]GATG[1]GATA[11]GACA[4]GATA[1]	0,147	0,165	0,146	0,153
17	CE17–GATG[2]GATA[1]GATG[1]GATA[13]GACA[3]GATA[1]	0,015	0,005	0,045	0,022
17	CE17–GATG[2]GATA[1]GATG[1]GATA[12]GACA[4]GATA[1]	0,275	0,314	0,141	0,243
18	CE18–GATG[2]GATA[1]GATG[1]GATA[14]GACA[3]GATA[1]			0,025	0,008
18	CE18–GATG[2]GATA[1]GATG[1]GATA[13]GACA[4]GATA[1]	0,196	0,216	0,096	0,169
18	CE18–GATG[2]GATA[1]GATG[1]GATA[12]GACA[5]GATA[1]		0,010		0,003
18	CE18–GATG[2]GATA[1]GATG[1]GATA[11]GACA[6]GATA[1]			0,025	0,008
19	CE19–GATG[2]GATA[1]GATG[1]GATA[15]GACA[3]GATA[1]		0,005		0,002
19	CE19–GATG[2]GATA[1]GATG[1]GATA[14]GACA[4]GATA[1]	0,113	0,098	0,025	0,079
19	CE19–GATG[2]GATA[1]GATG[1]GATA[13]GACA[5]GATA[1]			0,005	0,002
20	CE20–GATG[2]GATA[1]GATG[1]GATA[15]GACA[4]GATA[1]	0,005	0,010	0,005	0,007
20	CE20–GATG[2]GATA[1]GATG[1]GATA[14]GACA[5]GATA[1]			0,005	0,002
22	CE22–GATG[2]GATA[1]GATG[1]GATA[14]GACA[7]GATA[1]			0,020	0,007

For every locus the table displays the observed sequence alleles and respective allele frequencies for the three populations tested.

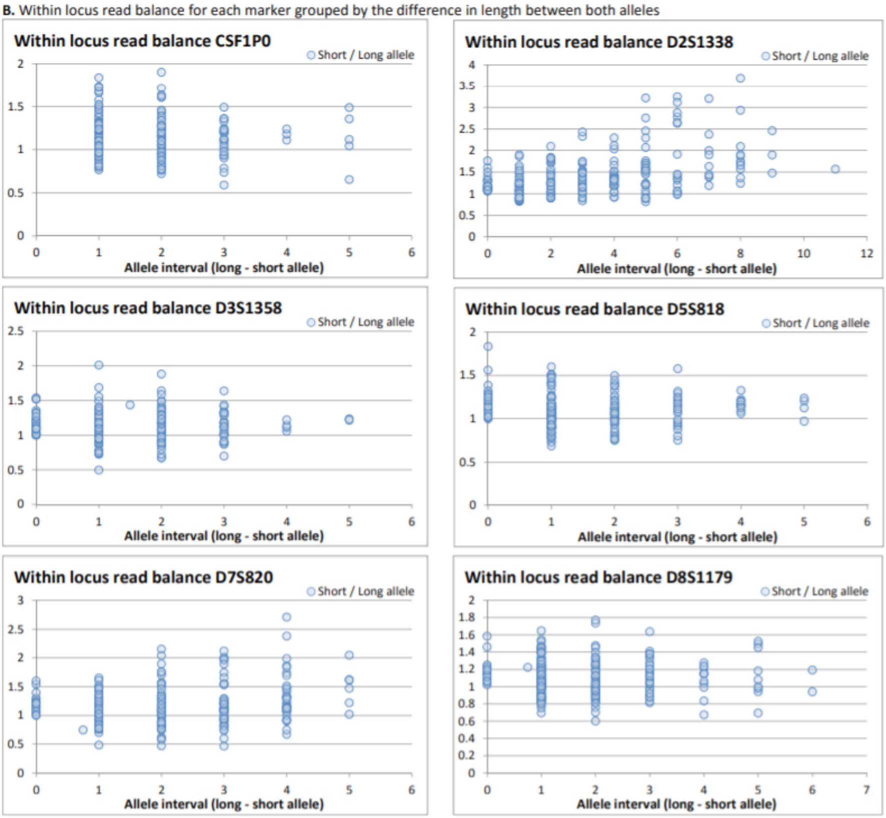
Supplemental Figure 7. Coverage and within locus allele balance for the samples used for stutter analysis

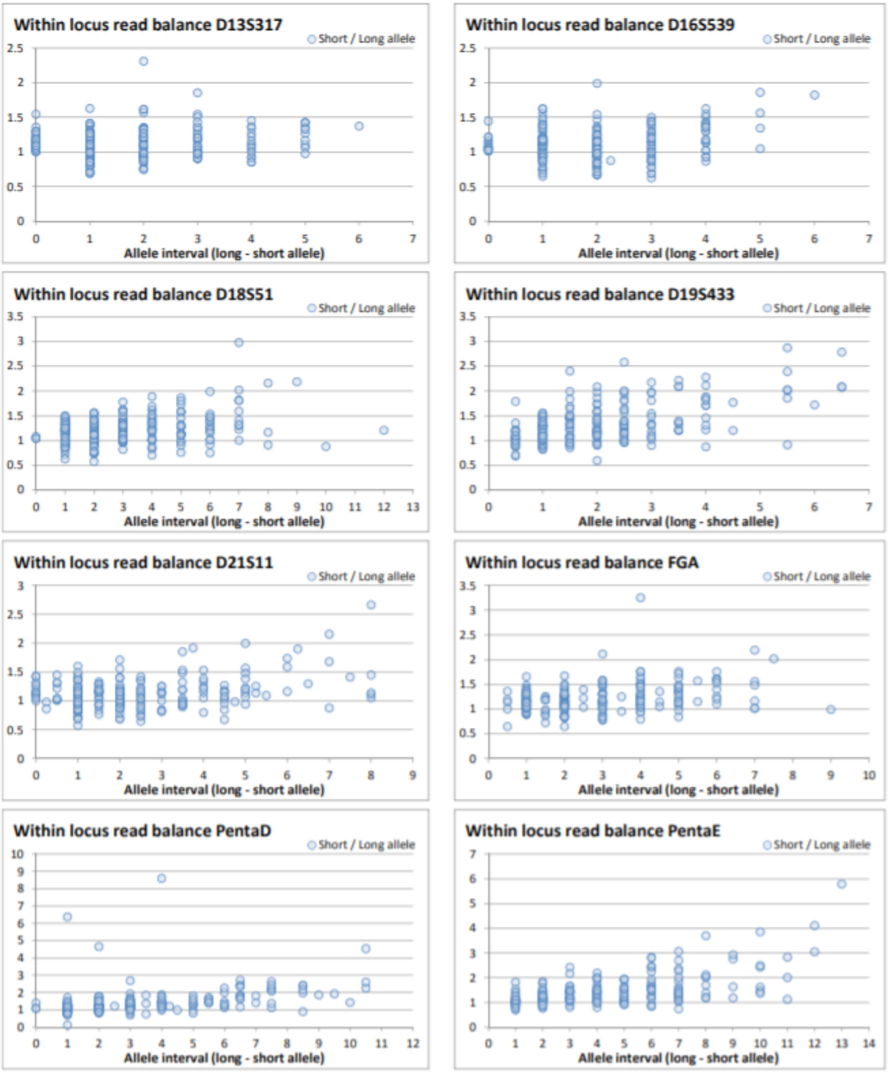
A. Average allele coverage and percentage of samples reaching the aimed allele coverage of 1000 reads

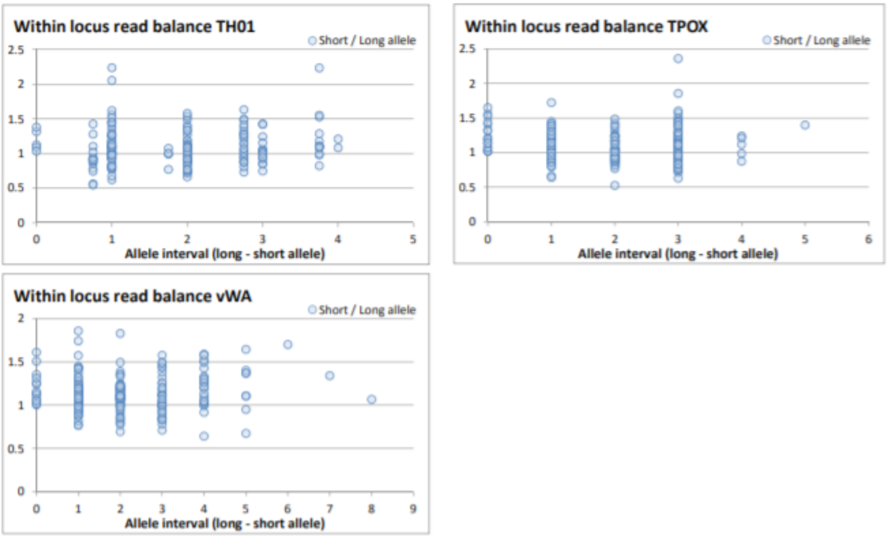


Locus	Average allele coverage	Allele percentage coverage >1000
CSF1P0	8512	100%
D2S1338	2596	86%
D3S1358	5390	100%
D5S818	7635	100%
D7S820	5354	90%
D8S1179	4416	92%
D13S317	7901	100%
D16S539	5279	96%
D18S51	5878	100%
D19S433	4476	99%
D21S11	6700	100%
FGA	4024	98%
PentaD	5271	99%
PentaE	5411	93%
TH01	5743	98%
TPOX	7074	100%
vWA	4389	89%

B. Within locus read balance for each marker grouped by the difference in length between both alleles

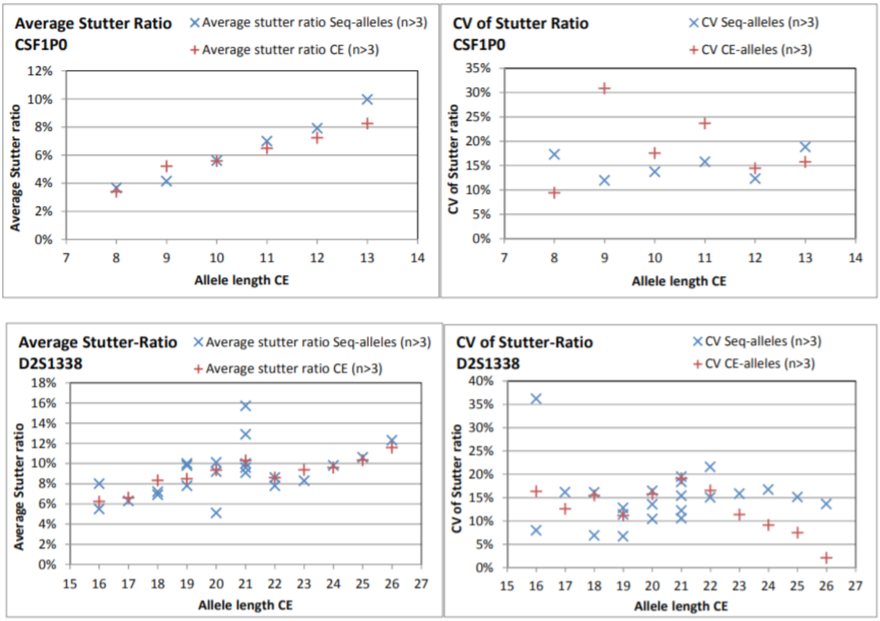


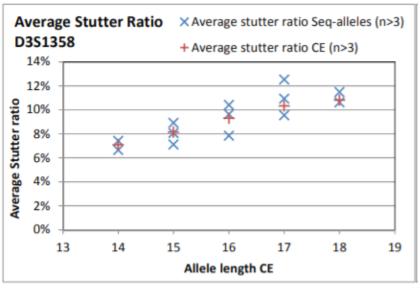

Within locus read balance D13S317
○ Short / Long allele
0 0.5 1 1.5 2 2.5
0 1 2 3 4 5 6 7
Allele interval (long - short allele)
Within locus read balance D16S539
○ Short / Long allele
0 0.5 1 1.5 2 2.5
0 1 2 3 4 5 6 7
Allele interval (long - short allele)
Within locus read balance D18S51
○ Short / Long allele
0 0.5 1 1.5 2 2.5 3 3.5
0 1 2 3 4 5 6 7 8 9 10 11 12 13
Allele interval (long - short allele)
Within locus read balance D19S433
○ Short / Long allele
0 0.5 1 1.5 2 2.5 3 3.5
0 1 2 3 4 5 6 7
Allele interval (long - short allele)
Within locus read balance D21S11
○ Short / Long allele
0 0.5 1 1.5 2 2.5 3
0 1 2 3 4 5 6 7 8 9
Allele interval (long - short allele)
Within locus read balance FGA
○ Short / Long allele
0 0.5 1 1.5 2 2.5 3 3.5
0 1 2 3 4 5 6 7 8 9 10
Allele interval (long - short allele)
Within locus read balance PentaD
○ Short / Long allele
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10 11 12
Allele interval (long - short allele)
Within locus read balance PentaE
○ Short / Long allele
0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Allele interval (long - short allele)

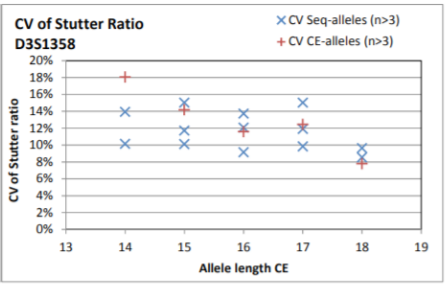


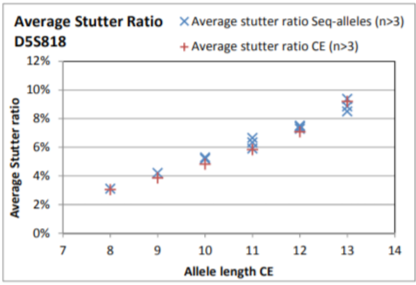
Dot plots for each locus displaying the ratio of the reads of the shortest and longest allele in each sample grouped by the STR length difference of the two alleles. In general, the shorter allele has a slightly higher number of reads than the longer allele (on average 1.2 times higher). For some loci, large length differences between the two alleles can result in stronger within marker allele inbalance.

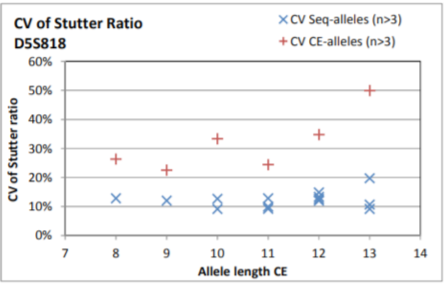
Supplemental Figure 8. Stutter characteristics for the 17 STRs of the prototype Powerseq™ system and the Powerplex® Fusion system

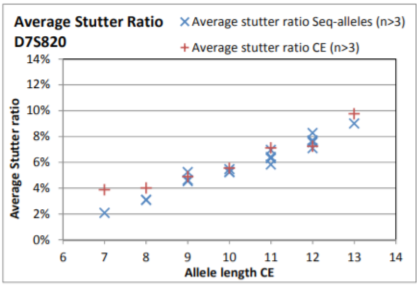


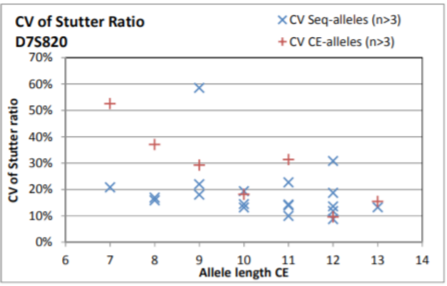

Average Stutter Ratio
D3S1358
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
14%
12%
10%
8%
6%
4%
2%
0%
13
14
15
16
17
18
19
Allele length CE

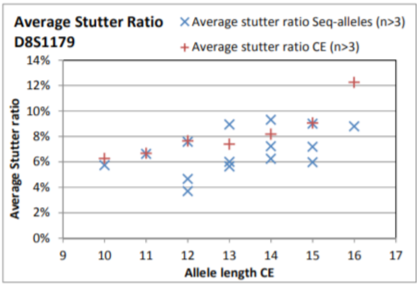

CV of Stutter Ratio
D3S1358
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
13
14
15
16
17
18
19
Allele length CE

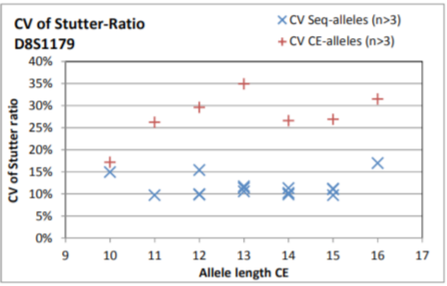

Average Stutter Ratio
D5S818
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
12%
10%
8%
6%
4%
2%
0%
7
8
9
10
11
12
13
14
Allele length CE

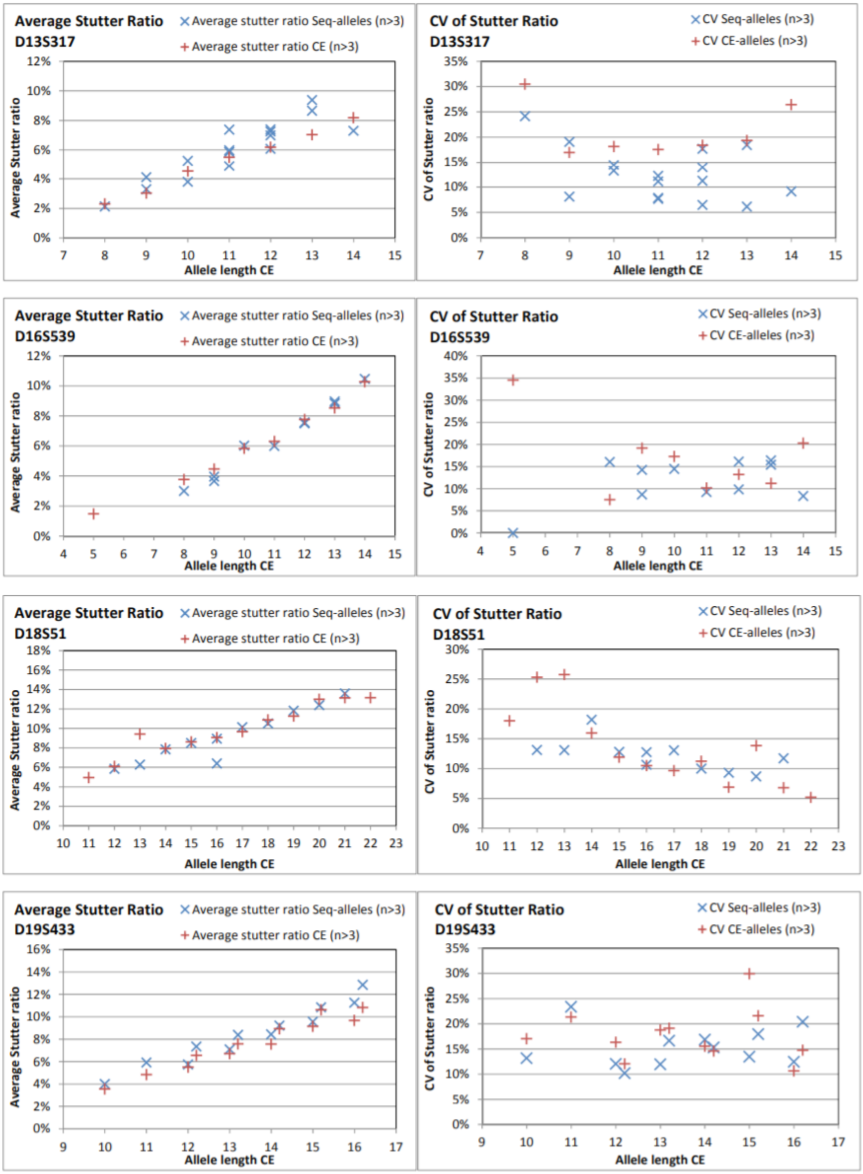

CV of Stutter Ratio
D5S818
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
60%
50%
40%
30%
20%
10%
0%
7
8
9
10
11
12
13
14
Allele length CE

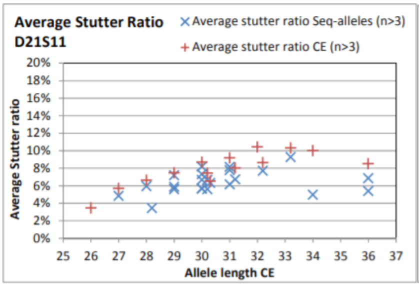

Average Stutter Ratio
D7S820
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
14%
12%
10%
8%
6%
4%
2%
0%
6
7
8
9
10
11
12
13
14
Allele length CE

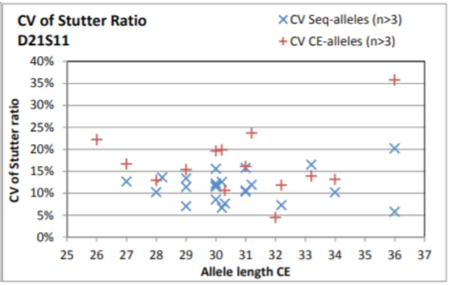

CV of Stutter Ratio
D7S820
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
70%
60%
50%
40%
30%
20%
10%
0%
6
7
8
9
10
11
12
13
14
Allele length CE

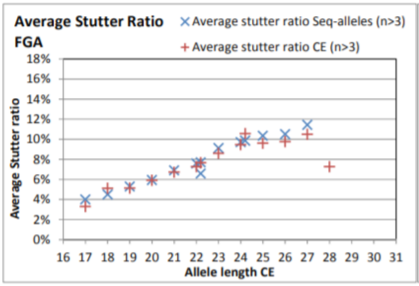

Average Stutter Ratio
D8S1179
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
14%
12%
10%
8%
6%
4%
2%
0%
9
10
11
12
13
14
15
16
17
Allele length CE

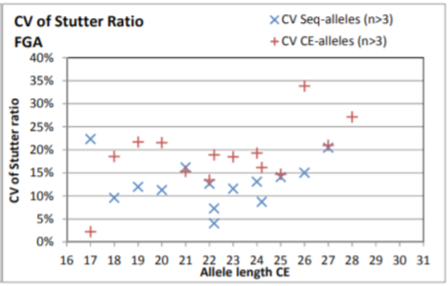

CV of Stutter-Ratio
D8S1179
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
40%
35%
30%
25%
20%
15%
10%
5%
0%
9
10
11
12
13
14
15
16
17
Allele length CE

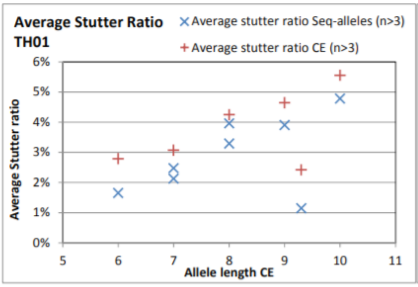

Average Stutter Ratio
D13S317
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
12%
10%
8%
6%
4%
2%
0%
7 8 9 10 11 12 13 14 15
Allele length CE
CV of Stutter Ratio
D13S317
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
35%
30%
25%
20%
15%
10%
5%
0%
7 8 9 10 11 12 13 14 15
Allele length CE
Average Stutter Ratio
D16S539
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
12%
10%
8%
6%
4%
2%
0%
4 5 6 7 8 9 10 11 12 13 14 15
Allele length CE
CV of Stutter Ratio
D16S539
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
40%
35%
30%
25%
20%
15%
10%
5%
0%
4 5 6 7 8 9 10 11 12 13 14 15
Allele length CE
Average Stutter Ratio
D18S51
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
10 11 12 13 14 15 16 17 18 19 20 21 22 23
Allele length CE
CV of Stutter Ratio
D18S51
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
30%
25%
20%
15%
10%
5%
0%
10 11 12 13 14 15 16 17 18 19 20 21 22 23
Allele length CE
Average Stutter Ratio
D19S433
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
16%
14%
12%
10%
8%
6%
4%
2%
0%
9 10 11 12 13 14 15 16 17
Allele length CE
CV of Stutter Ratio
D19S433
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
35%
30%
25%
20%
15%
10%
5%
0%
9 10 11 12 13 14 15 16 17
Allele length CE

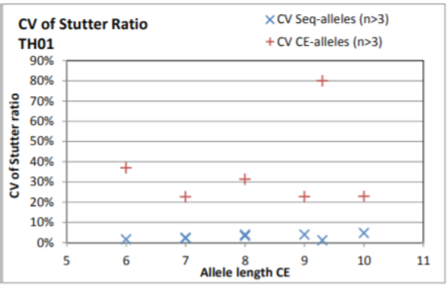

Average Stutter Ratio
D21S11
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
25 26 27 28 29 30 31 32 33 34 35 36 37
Allele length CE

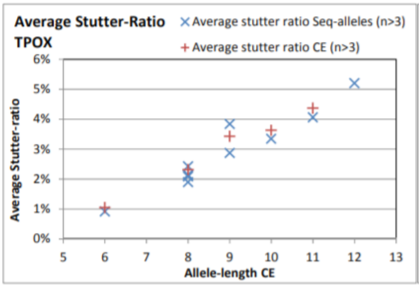

CV of Stutter Ratio
D21S11
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
40%
35%
30%
25%
20%
15%
10%
5%
0%
25 26 27 28 29 30 31 32 33 34 35 36 37
Allele length CE

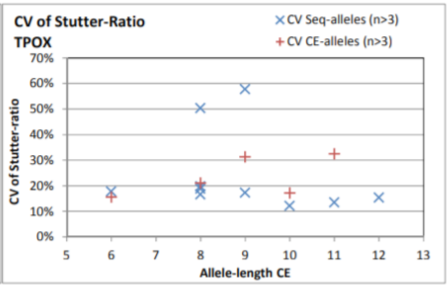

Average Stutter Ratio
FGA
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Allele length CE

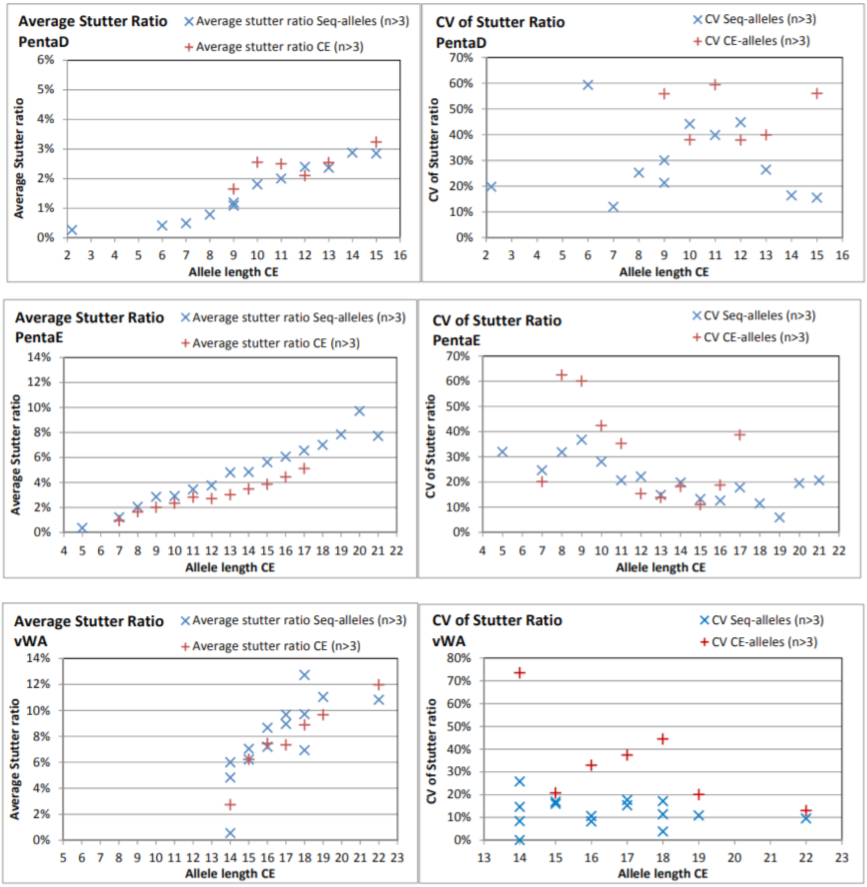

CV of Stutter Ratio
FGA
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
40%
35%
30%
25%
20%
15%
10%
5%
0%
16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Allele length CE


Average Stutter Ratio
TH01
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
6%
5%
4%
3%
2%
1%
0%
5 6 7 8 9 10 11
Allele length CE


CV of Stutter Ratio
TH01
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
5 6 7 8 9 10 11
Allele length CE


Average Stutter-Ratio
TPOX
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter-ratio
6%
5%
4%
3%
2%
1%
0%
5 6 7 8 9 10 11 12 13
Allele-length CE

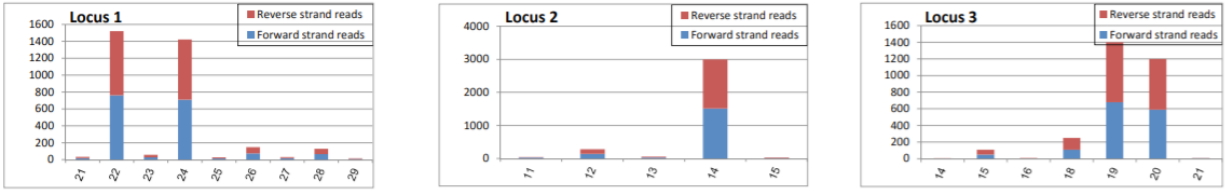

CV of Stutter-Ratio
TPOX
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter-ratio
70%
60%
50%
40%
30%
20%
10%
0%
5 6 7 8 9 10 11 12 13
Allele-length CE



Comparison of stutter characteristics for the MPS-based Powerseq^{TM} system and the CE-based Powerplex® Fusion system. For every marker, two graphs display the average stutter ratio for every allele and the Coefficient of Variance (CV) for the stutter ratio of each allele. Since for MPS-based analysis some alleles of the same CE-length are represented by distinct sequence alleles, the stutter ratios and CV of these alleles are displayed separately. It is apparent that in general, stutter ratios of the MPS-based analysis are similar to the CE-based stutter ratios except for CE-alleles that are subdivided into several sequence alleles. The subdivided sequence alleles of the same total CE-length often have different stutter ratios which results in a lower variation of stutter ratio per allele compared to the combined CE-allele. This can be observed from the generally lower values for the CV of the stutter ratio for the sequence alleles.

Supplemental Figure 9. A three locus hypothetical example illustrating our method for the calculation of the proportion of a minor contribution in a mixture

A. The sequence read profiles for three loci of a single two-person mixture



X-axis shows allele number
Y-axis shows number of sequence reads observed

X-axis shows allele number
Y-axis shows number of sequence reads observed

X-axis shows allele number
Y-axis shows number of sequence reads observed

B. Summary tables showing the read numbers of all unique alleles and stutters that passed all quality filters

Locus 1

Alleles	Forward strand reads	Reverse strand reads	All reads combined	Proportion of total locus reads	Interpretation
21	16	14	30	0.0090	stutter
22	760	740	1500	0.4478	major
23	28	32	60	0.0179	stutter
24	710	690	1400	0.4179	major
25	13	17	30	0.0090	stutter
26	74	76	150	0.0448	minor
27	15	15	30	0.0090	stutter
28	64	66	130	0.0388	minor
29	8	12	20	0.0060	stutter
Total	1688	1662	**3350**		

Locus 2

Alleles	Forward strand reads	Reverse strand reads	All reads combined	Proportion of total locus reads	Interpretation
11	22	18	40	0.0118	stutter
12	135	145	280	0.0824	minor
13	24	26	50	0.0147	stutter
14	1510	1490	3000	0.8824	major
15	14	16	30	0.0088	stutter
Total	1705	1695	**3400**		

Locus 3

Alleles	Forward strand reads	Reverse strand reads	All reads combined	Proportion of total locus reads	Interpretation
14	1	4	5	0.0017	stutter
15	50	60	110	0.0369	minor
16	4	6	10	0.0034	stutter
18	110	140	250	0.0838	stutter
19	680	720	1400	0.4690	major
20	590	610	1200	0.4020	major
21	5	5	10	0.0034	stutter
Total	1440	1545	**2985**		

Read interpretation of Locus 1
(Please note that in this example we only consider +1 or -1 stutters)

None of the possible minor alleles overlap with stutters or with major alleles

We use both minor allele (26 and 28) reads for calculation

Read interpretation of Locus 2
(Please note that in this example we only consider +1 or -1 stutters)

One minor allele (12) is visible. There is either only one (homozygous) minor allele, or the second minor allele overlaps the major. Only clear heterozygous minor alleles are used for the calculation which is not the case here

Reads from this minor allele are not used for calculation

Read interpretation of Locus 3
(Please note that in this example we only consider +1 or -1 stutters)

Because of the high read count of allele 18, this allele was interpreted as a minor allele + stutter, the stutter influences the read-count of this allele and allele 18 can therefore not be used for a reliable calculation of the proportion of minor contribution. Allele 15 remains as one of the heterozygous minor alleles and can be included in the calculation of the proportion of minor contribution.

Only allele 15 is used for the calculation

C. Calculation of the proportion of the minor contribution

Average minor allele proportion: (0,0448 + 0,0388 + 0,0369) / 3 = 0,041
Total minor contribution in profile: 0.0401 * 2 * 100(%) = 8%