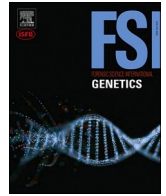Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Ultrasensitive sequencing of STR markers utilizing unique molecular identifiers and the SiMSen-Seq method

Maja Sidstedt [a,1], Arvid H. Gynnå [a,1], Kevin M. Kiesler [b], Linda Jansson [a,c], Carolyn R. Steffen [b], Joakim Håkansson [d,e,f], Gustav Johansson [g], Tobias Österlund [h,i,j], Yalda Bogestål [d], Andreas Tillmar [k], Peter Rådström [c], Anders Ståhlberg [h,i,j], Peter M. Vallone [b], Johannes Hedman [a,c,*]

[a] National Forensic Centre, Swedish Police Authority, Linköping SE-581 94, Sweden
[b] National Institute of Standards and Technology, 100 Bureau Drive, M/S 8314, Gaithersburg, MD 20899, USA
[c] Applied Microbiology, Department of Chemistry, Lund University, Lund SE-221 00, Sweden
[d] RISE Unit of Biological Function, Division Materials and Production, Box 857, Borås SE-501 15, Sweden
[e] Department of Laboratory Medicine, Institute of Biomedicine, University of Gothenburg, Gothenburg SE-405 30, Sweden
[f] Department of Chemistry and Molecular Biology, University of Gothenburg, Gothenburg SE-405 30, Sweden
[g] SIMSEN Diagnostics, Sahlgrenska Science Park, Gothenburg, Sweden
[h] Department of Laboratory Medicine, Sahlgrenska Center for Cancer Research, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Medicinaregatan 1F, Gothenburg 41390, Sweden
[i] Wallenberg Center for Molecular and Translational Medicine, University of Gothenburg, Gothenburg 41390, Sweden
[j] Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital, Gothenburg, Region Västra Götaland 41390, Sweden
[k] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping SE-587 58, Sweden

ARTICLE INFO

ABSTRACT

Massively parallel sequencing (MPS) is increasingly applied in forensic short tandem repeat (STR) analysis. The presence of stutter artefacts and other PCR or sequencing errors in the MPS-STR data partly limits the detection of low DNA amounts, *e.g.*, in complex mixtures. Unique molecular identifiers (UMIs) have been applied in several scientific fields to reduce noise in sequencing. UMIs consist of a stretch of random nucleotides, a unique barcode for each starting DNA molecule, that is incorporated in the DNA template using either ligation or PCR. The barcode is used to generate consensus reads, thus removing errors. The SiMSen-Seq (Simple, multiplexed, PCR-based barcoding of DNA for sensitive mutation detection using sequencing) method relies on PCR-based introduction of UMIs and includes a sophisticated hairpin design to reduce unspecific primer binding as well as PCR protocol adjustments to further optimize the reaction. In this study, SiMSen-Seq is applied to develop a proof-of-concept seven STR multiplex for MPS library preparation and an associated bioinformatics pipeline. Additionally, machine learning (ML) models were evaluated to further improve UMI allele calling. Overall, the seven STR multiplex resulted in complete detection and concordant alleles for 47 single-source samples at 1 ng input DNA as well as for low-template samples at 62.5 pg input DNA. For twelve challenging mixtures with minor contributions of 10 pg to 150 pg and ratios of 1–15% relative to the major donor, 99.2% of the expected alleles were detected by applying the UMIs in combination with an ML filter. The main impact of UMIs was a substantially lowered number of artefacts as well as reduced stutter ratios, which were generally below 5% of the parental allele. In conclusion, UMI-based STR sequencing opens new means for improved analysis of challenging crime scene samples including complex mixtures.

1. Introduction

Massively parallel sequencing (MPS) has revolutionized molecular biology and genomics. Numerous applications of MPS have been presented in forensic genetics, including single nucleotide polymorphism (SNP) analysis for appearance and ancestry prediction, kinship analysis and investigative genetic genealogy, as well as short tandem repeat (STR) DNA profiling [1–6]. Choosing MPS over the present state-of-the-art methodology of capillary electrophoresis enables the generation of larger multiplex assays and, in the case of STR markers, improves allele resolution through the addition of sequence information. The latter feature is especially valuable in the analysis of complex mixtures, *i.e.*, crime scene traces containing DNA from two or more individuals [7,8].

One challenge with sequencing is to distinguish true alleles from noise [7,9,10]. In theory, MPS enables the analysis of individual template molecules. In practice, the detection of rare DNA variants is limited by artefacts from library preparation or from the sequencing itself [11, 12]. Generally, these erroneous sequences make it difficult to confidently detect variants that constitute less than 1% of the template molecules for a specific marker [13,14]. STR markers have even higher error rates due to stutter artefacts formed by polymerase slippage. The most common type of stutter is n-1, *i.e.,* molecules that have lost one repeat unit compared to the true allele [15–17]. Polymerase slippage can also lead to the introduction of a repeat unit (n+1 stutter) and multiple stuttering events may occur, leading to stutters such as n-2 and n-3 [18,19].

The systematic nature and high incidence of STR stutter artefacts hinder the full exploitation of the power of MPS for the analysis of challenging casework samples. A commercial STR sequencing system has been shown to enable the detection of the minor contributor at 5% of the total amount of DNA in two-person mixtures [20,21]. However, minor alleles in n-1 stutter positions of the major donor will be filtered at substantially higher rates, preventing detection of any masked true alleles. Taking alleles in stutter positions into account, a state-of-the-art forensic STR kit enabled the complete detection of minor contributions in 1:3 mixtures, whereas one-third to two-thirds of the markers were called in 1:19 mixtures [22]. Being able to analyze and interpret even smaller contributions in mixed stains would be highly beneficial, as a substantial part of biological traces in severe crimes are mixtures [23–26].

Aside from stutter, sequencing of STRs is affected by single base errors [27]. Such errors can originate either during PCR-based library preparation [28], or through a read error in the process of sequencing [29]. A PCR error is more problematic, since a substitution (or error in general) will proliferate through the following cycles, so that a replication mistake occurring in an early cycle in a single molecule can make up a substantial portion of the final reads for a locus. An error during the sequencing process, on the other hand, will only result in a single incorrect read. In addition to stutter and single base errors, artefacts such as PCR hybrids and insertions and deletions of other sizes than an STR repeat unit also exist at lower levels [10].

A solution that aims to reduce sequencing noise, and thus enable the detection of low-abundance alleles, is the application of unique molecular identifiers (UMIs) [30–34]. UMIs are random oligonucleotide stretches of 8–18 bases. The UMIs are attached to the target DNA early in library preparation, making it possible to group all reads stemming from a specific template molecule post-sequencing [33]. The labelling is done either through enzymatic digestion and ligation or via a PCR step where the UMI is included in one of the primers [35–37]. Generally, all reads carrying the same UMI sequence are grouped into a UMI family from which a consensus read aiming to represent the true allele is generated. Theoretically, most of the errors introduced in PCR or sequencing will thus be removed.

UMI barcoding is very effective in reducing random errors, and has greatly improved the possibility of detecting rare sequence variants in medical applications [34]. During the last decade, UMIs have been applied for prenatal testing, cancer diagnostics and quantitative RNA sequencing [33,35,38]. Recently, UMIs have found their way into forensic genetics through the application of QIAGEN fragmentation and ligation methodology for sequencing of SNP [39] and STR markers [40, 41], respectively. However, the 10 ng recommended DNA input for the QIAGEN library preparation kit is often unavailable in forensic applications, although the methodology has been evaluated with lower DNA amounts [39,41].

PCR-based introduction of UMIs provides low limits of detection [35], but also comes with its own complexity: the random bases of the UMI may bind nonspecifically to the DNA template, leading to loss of primers and the risk of forming unspecific PCR products. One technology developed to handle this issue, originally intended for cancer diagnostics, is SiMSen-Seq [42]. SiMSen-Seq (Simple, multiplexed, PCR-based barcoding of DNA for sensitive mutation detection using sequencing) includes protection of the UMI in a hairpin to avoid interaction with the template during primer annealing and has been shown to enable the detection of rare SNP variants related to tumors down to 0.1% in liquid biopsies [42,43]. The use of low primer concentrations (≈ 40 nmol/L), increased PCR extension times (≈ 6 minutes), low amounts of DNA polymerase (≈ 0.1X) and low numbers of PCR cycles (≈ 3 cycles) in the first of two PCR steps are other key features of SiMSen-Seq library preparation [42–45].

Here, the SiMSen-Seq method is applied to develop a proof-of-concept seven STR marker multiplex for MPS library preparation and an associated bioinformatics pipeline. The overall aim is to determine the potential of UMIs and SiMSen-Seq to improve the performance of MPS-STR in general, with a particular focus on the detection of minor contributors in mixed crime scene traces. To this end, the assay and pipeline are used to analyze and categorize the types of errors that occur in PCR and sequencing of STR markers. Initially this is done without applying any allele calling thresholds or stutter filters. The impact of UMIs on error reduction is investigated, studying both the systematic stutter artefacts and random single base errors. Two methods for generation of consensus reads are compared with respect to the quality of the results: a naïve approach using the most common sequence and multiple sequence alignment. Additionally, machine learning (ML) models are trained to distinguish between correct and erroneous consensus reads based on UMI family information. Finally, the performance of the SiMSen-Seq STR multiplex is evaluated for analysis of mock casework samples including low-template samples and complex mixtures containing DNA of up to five persons. To the best of our knowledge, this is the first application of PCR-based UMI labelling in forensic genetics and STR profiling.

2. Materials and methods

The seven STR multiplex assay developed in this study is based on the previously described SiMSen-Seq methodology [43]. The method includes a barcoding PCR followed by adaptor PCR (Fig. 1A). One of the primers in each pair includes a stretch of 12 random nucleotides constituting the UMI or barcode. The UMI is protected during the barcoding PCR by a stem-loop structure preventing base pairing below the hairpin melting temperature of 74 °C. In the barcoding PCR, a combination of low concentrations of primers, low amounts of DNA polymerase, low numbers of PCR cycles, and increased extension times contribute to keeping errors as low as possible while tagging the copies of each original template molecule with UMIs. This enables, through bioinformatic analysis, tracking of each individual template molecule while also correcting for errors that occur both in the PCR-based library preparation and sequencing (Fig. 1B). All reads carrying the same UMI sequence are grouped into one UMI family, which is used to generate one consensus read.
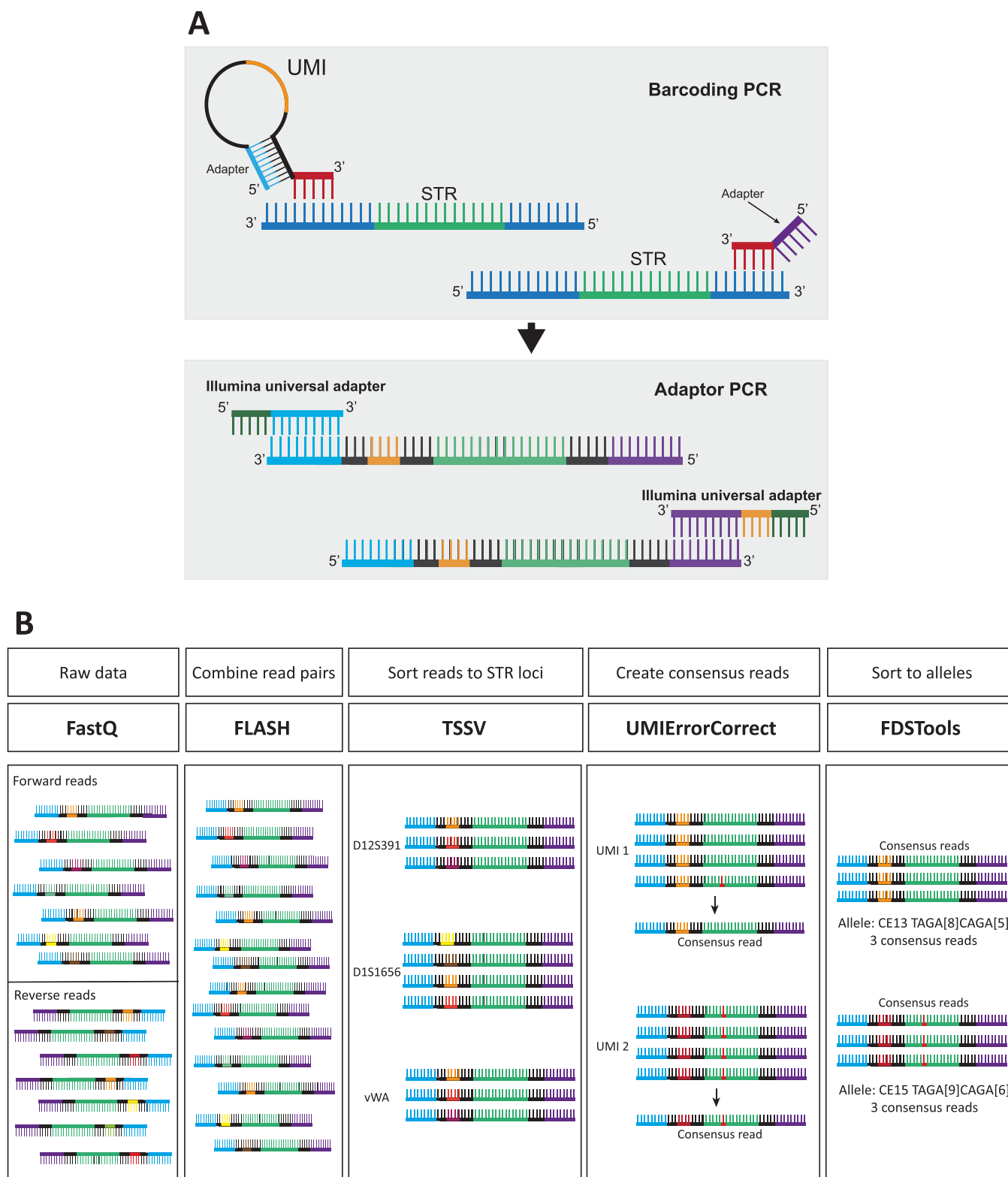
Fig. 1. Schematic illustration of (A) the barcoding and adaptor PCR and (B) the bioinformatic pipeline.

2.1. DNA samples

In this study, reference DNA samples, single-source DNA samples, and DNA mixtures were used. The DNA material used as a positive control was 2800 M Control DNA (Cat. nr. DD7101, 10 ng/μL, Promega, Madison, WI, USA), hereafter referred to as 2800 M. 2800 M is a single-source male human genomic DNA commonly used as a control in STR

analysis. NIST Standard Reference Material (SRM) 2391d components A, B and C [46] as well as 44 well-characterized NIST samples from different individuals with published or known STR profiles were used in this study [4,47]. The latter samples were quantified using an in-house digital PCR assay and subsequently diluted to 0.5 ng/μL prior to analysis. The twelve included DNA mixtures are made up of the same contributors and ratios as twelve of the mixtures in the Forensic DNA Open

Dataset [47]. The mixture proportions of the samples used are presented in Table 1. All work presented has been reviewed and approved by the NIST Human Research Protections Office (MML-16–0080).

Low-template DNA input was also investigated, applying two dilution series from two of the single-source samples above to give DNA inputs of 2 ng, 1 ng, 500 pg, 250 pg, 125 pg and 62.5 pg (each amount analyzed in duplicate).

2.2. SiMSen-Seq primers for STR markers

The development of the seven multiplex STR assay was performed as described in the SiMSen-Seq protocol [43]. The seven STR markers were chosen based on analytical performance and the added value of increased allelic information due to sequence variants as reported in the literature [4,48]. The primer sequences were obtained from published studies through STRbase [49,50] and are listed in Supplementary Table S1.

The primer pairs in the seven STR multiplex assay target the following STR loci, ordered from the smallest amplicons to the largest: D2S441 [51], D1S1656 [52], D3S1358 [53], vWA [53], D8S1179 [53], D21S11 [53] and D12S391 [52]. The complete amplicon sizes and read lengths for these markers are described in Supplementary Table S1. The theoretical maximum read lengths (excluding 28 bp for the UMI and hairpin) for analysis of 2800 M DNA are 100 bp for D2S441, 137 bp for D1S1656, 139 bp for D3S1358, 159 bp for vWA, 232 bp for D8S1179, 233 bp for D21S11 and 241 bp for D12S391.

The primers are rather long (up to 96 bp), since, in addition to the target-specific sequences, the UMI and hairpin are included in one primer and the adapter-specific sequence in the other, as described in the SiMSen-Seq protocol [43]. The UMI and hairpin sequence was added to the sequence upstream (D2S441, D8S1179, D12S391, D21S11) or downstream (D1S1656, D3S1358, vWA) of the repeat region to optimize the performance of each primer pair. Primers for the barcoding PCR were ordered from Integrated DNA Technologies (IDT, Coralville, IA, USA) as DNA Ultramer oligomers with standard desalting to achieve the best possible oligonucleotide quality [54]. Primers for the adaptor PCR were also ordered from IDT but as DNA oligomers with HPLC purification.

2.3. Library preparation

The first step in SiMSen-Seq library preparation is barcoding PCR where the following reagents and concentrations were used in a total reaction volume of 10 μL: 1X SuperFi buffer (Thermo Fisher Scientific, Waltham, MA, USA), 2.5 mmol/L MgCl$_2$ (Roche, Basel, Schweiz), 0.5 mol/L L-carnitine inner salt (Sigma-Aldrich, Burlington, MA, USA), 0.2 mmol/L dNTPs (Roche), 40 nmol/L to 100 nmol/L of each barcode primer (IDT), 0.05 μL (0.25X) Platinum SuperFi II DNA polymerase (Thermo Fisher Scientific). One ng of template DNA was added to each reaction unless otherwise noted. Cycling was performed on a ProFlex

Table 1
Description of the analyzed DNA mixtures. The mixtures have been prepared and described in detail previously [47].

Number of contributors	Ratio (%)	DNA amounts (pg)
3 P	3/3/94	30/30/940
3 P	3/3/94	120/120/3760
3 P	5/47/48	50/470/480
3 P	10/45/45	100/450/450
4 P	5/5/5/85	50/50/50/850
4 P	10/10/10/70	100/100/100/700
4 P	1/5/47/47	10/50/470/470
4 P	10/10/40/40	100/100/400/400
5 P	5/5/5/5/80	50/50/50/50/800
5 P	15/15/15/15/40	150/150/150/150/400
5 P	10/10/10/35/35	100/100/100/350/350
5 P	5/5/30/30/30	50/50/300/300/300

PCR System (Thermo Fisher Scientific) using the following settings: 98 °C for 3 min, 4 cycles of [98 °C for 10 s, 59 °C for 6 min and 72 °C for 30 s], 72 °C for 30 s and hold at 4 °C.

The second step in library preparation is adaptor PCR. The reaction mix was prepared accordingly, with a total reaction volume of 25 μL: 1X SuperFi buffer (Thermo Fisher Scientific), 2.5 mmol/L MgCl$_2$ (Roche), 0.5 mol/L L-carnitine inner salt (Sigma-Aldrich), 0.2 mmol/L dNTPs (Roche), 0.4 μmol/L of each primer (IDT), 0.5 μL (1X) Platinum SuperFi II DNA polymerase (Thermo Fisher Scientific). Additionally, 8 μL of the barcoding PCR reaction mixture was added as template. Cycling was performed on a ProFlex PCR System (Thermo Fisher Scientific) using the following settings: 98 °C for 2 min, 30 cycles of [98 °C for 10 s, 80 °C for 1 s, 72 °C for 30 s, 76 °C for 30 s] and hold at 4 °C. The ramp rate during cycling was set to 0.4 °C/s.

After the adaptor PCR, purification of the products was performed with AMPure XP Beads (Beckman Coulter, Brea, CA, USA) at a 0.8X ratio vol/vol. Before adding the beads (28 μL), the reaction volume was adjusted to 35 μL by the addition of 10 μL nuclease-free water to the adaptor PCR reaction mixture. The protocol from the manufacturer was used, and the final products were eluted in 20 μL Low EDTA TE, pH 8.0 (Quality Biological, Gaithersburg, MD, USA).

The purified libraries were analyzed with the TapeStation 4150 (Agilent, Santa Clara, CA, USA) and High Sensitivity D1000 Reagents (Product nr. 5067–5584–5587, Agilent) as a quality control before sequencing.

The DNA concentration of each library was determined using the Qubit dsDNA HS Assay Kit (Product nr. Q32851, Thermo Fisher Scientific). Thereafter, all samples were equimolarly normalized and diluted to 4 nmol/L in one pool. For the negative controls, the average concentration of the samples was applied in normalization. The pooled libraries were further diluted to 8 pmol/L with ~ 10% PhiX (Illumina, San Diego, CA, USA) spike-in. Sequencing was performed on a MiSeq FGx (Verogen, San Diego, CA, USA) using the MiSeq Reagent Kit v3, 600 cycles (Illumina) in 2×300 cycles paired-end read mode.

2.4. Bioinformatic processing and data analysis

The UMIec Forensics bioinformatic pipeline was developed to determine STR genotypes from barcoded sequencing reads (available under MIT license at https://github.com/agynna/UMIec_forensics). The pipeline builds on the UMIErrorCorrect pipeline for deduplicating barcoded sequencing reads and the FDStools suite for typing STR markers [10,55]. The pipeline uses FLASH to combine paired end-reads [56], which are assigned to a STR marker by FDStools TSSV, and sorted into UMI families by UMIErrorCorrect. By default, one mismatch is allowed in the UMI sequence.

A schematic illustration of the bioinformatic analysis outlines the main steps (Fig. 1B). Briefly, the forward and reverse FASTQ files from the MiSeq FGx instrument contain sequences with the sample-specific index denoted in the sample sheet. The first step in the analysis is to remove adapter ends using AdapterRemoval [57] and thereafter to combine the read pairs using a modified version of FLASH [56]. Next, the UMI and spacer sequences are trimmed from each read by creating a new FASTQ file with the UMI-information in the header. Here, UMIs within a Hamming distance ≤ 1 are clustered according to the "directional" method [58], to allow for sequencing errors in the UMI. Since STR markers are highly repetitive and generally do not work well with standard aligners [59–61], TSSV was used [10,61] to sort the sequences in a FASTQ file for each of the seven included STR markers. Thereafter, a conversion from FASTQ to BAM using the samtools fastq command was included to fit the input requirement in the next pipeline step, UMIErrorCorrect [55]. UMIErrorCorrect analyzes all sequences containing the same UMI and creates one consensus read per UMI. Consensus reads were generated by either taking the most common sequence with each UMI or performing a multiple sequence alignment using mafft [62]. For both of these, the consensus read generation required at least 50%

identity at each base and only UMI families with at least three reads were accepted. During explorative investigation and when the subsequent ML filter was used, a minimum of two reads per UMI family were required to accept a consensus read. As a last step in the analysis, FDSTools v. 2.0 [10] was used to determine the STR alleles in each sample using the UMI consensus reads. In parallel, each sample was also analyzed with FDSTools ignoring the UMI information for a comparison between using and not using the UMIs.

During data processing, rather than reporting the obtained sequence strings, the module STRnaming (part of FDSTools) was used to shorten the strings into interpretable brackets [63]. Here, each sequence variant is labelled according to the corresponding fragment size obtained in standard STR capillary electrophoresis analysis (*e.g.*, CE13), followed by the DNA sequence repeat structure in brackets.

The data was summarized at different levels to determine the added value of utilizing UMIs. Raw reads are the total number of reads in the FASTQ file. The term "**read**" is used for reads mapped to STRs and the term "**consensus read**" is used for data that has been constructed from the UMI families.

The n-1 stutter ratios were calculated by dividing the number of reads/consensus reads for the stutter artefact by the number of reads/consensus reads for the parental allele. Heterozygote balance was calculated by dividing the number of reads/consensus reads for the allele with the lower read number by that for the allele with the higher read number.

Data analysis and plotting was done in Python using the packages *pandas* (v 2.0.2), *scipy* (v. 1.8.1), *matplotlib* (v. 3.7.1) and *seaborn* (v. 0.12.2).

200 randomly selected UMI families with incorrect consensus reads were inspected to investigate why and which types of errors still persist after UMI correction. The families were classified according to the type of error.

Machine learning (ML) was explored to filter out unreliable UMI families. For this purpose, the single-source data was divided into training and test sets with 39 and 10 profiles, respectively. The Python packages *scikit-learn* (v. 1.1.1) and *imbalanced-learn* (v 0.9.1) were used for machine learning [64].

Thirteen features based on the family members and the consensus read were calculated for each UMI family and used as model input (Supplementary Table S2). Features related to sequence, size or allele number were excluded to decrease the risk of biasing the model towards specific alleles present in the training set. Three learning algorithms were evaluated: support vector machine (SVM), random forest (RF) and small fully connected neural networks (NN, in *scikit-learn* also known as multi-layer perceptron, MLP) [65–67]. The models were trained to predict whether a consensus read was correct, *i.e.* identical to the known allele sequence, or not.

The SVM model used the *SVC* classifier in *scikit-learn*. The numeric features were transformed by Yeo-Johnson transformation. Hyperparameters were decided by stepwise grid searches and five-fold cross validation. Separate models were trained for each marker. The final models used equal sampling from correct and incorrect UMI families, radial basis function kernel with kernel coefficient (gamma) 100, regularization (C) 0.05, and equal class weights.

The RF model used *RandomForestClassifier* in *scikit-learn*. Hyperparameters were decided by stepwise random searches and five-fold cross validation. Separate models were trained for each marker. The final models used equal sampling from correct and incorrect UMI families, 70 estimators, maximum depth of 8, maximum 3 features per split and minimum 1 sample per leaf. Isotonic probability calibration was performed per marker by *CalibratedClassifierCV*.

The NN model used *MLPClassifier* in *scikit-learn*. The numeric features were transformed by Yeo-Johnson transformation. Hyperparameters were decided by random search and five-fold cross validation. The final model used a 1:8 sampling ratio of incorrect to correct UMI families, two layers with 10 and 5 ReLU nodes, respectively, and L2 regularization

(alpha) 0.005. Isotonic probability calibration was performed by *CalibratedClassifierCV*.

After cross validation and hyperparameter selection, the whole training set was used to train the final models which were evaluated on the test set.

The NN algorithm was found to have the best performance, closely followed by RF, with the SVM being considerably worse when compared by metrics suitable for imbalanced classification problems (Supplementary Figure S1 A-C).

Initially, no allele calling thresholds or stutter filters were applied to enable the study of PCR and sequencing artefacts and the impact of UMIs and ML filter on a fundamental level. Then, the effect of varying the allele calling threshold on the detection of true alleles and artefacts was investigated.

3. Results

Four sequencing runs were performed in this study, each run including 25 samples and a negative control. The quantity and quality of all libraries was verified before sequencing using fluorometry and fragment analysis (Supplementary Table S3). The sequencing quality metrics were within the expected ranges (Supplementary Table S4). The number of raw reads and reads mapped to STR markers per sample were consistent between the four sequencing runs (Supplementary Table S5-S6, Supplementary Figure S2). Excluding negative controls, the overall mean number of reads per STR marker was 95 039 (SD = 43 613). Taking the UMI information into account, the mean number of consensus reads per STR marker was 1 143 (SD = 676) (Supplementary Figure S3). All 47 analyzed single-source samples, as well as 2800 M DNA, showed full allele fragment size and sequence concordance with published or verified values for all seven STR markers. For all analyzed samples and markers, the alleles with the highest read number/consensus read number supported the correct genotype (Supplementary Figure S4, S5).

The negative controls consistently yielded below 250 000 raw reads and a maximum of 2% of the raw reads were mapped to STR markers. On average, 18 (SD = 9) reads were mapped to each STR locus, the majority of which were not recognized as alleles. With UMIs, the negative controls did not result in any consensus reads for any of the markers due to the low numbers of obtained reads.

3.1. Generation of consensus reads from UMI families

Two methods for generation of consensus reads from each UMI family were compared to recover the original template sequences as faithfully as possible: (1) the naïve approach of taking the most common sequence among the members of each family and (2) performing a multiple sequence alignment of all family members. When applied to 47 single-source samples, the alignment method produced overall 1.6% more correct consensus reads compared to the "most common" method while being far more computationally expensive. Simultaneously, 18% more incorrect consensus reads were obtained, leading to an increased proportion of errors. Accordingly, the "most common" method was used in the remainder of the study, resulting in between 0.4% and 4.2% incorrect consensus reads, depending on STR marker (Fig. 2A).

A subset of the UMI families was manually inspected to improve the understanding of the consensus read generation. Most UMI families (more than 95% of the obtained consensus reads for each marker) supported the correct allele. These families typically consisted of a majority of reads supporting the correct sequence and an ensemble of less common variants attributed to stutter artefacts and single base errors (Fig. 2B). It is evident that a main benefit of using UMIs is that these errors are corrected by the consensus read generation.

However, about 2.3% of the UMI families still yielded consensus reads not matching the known alleles. An erroneous consensus read can be caused by either an error in the barcoding PCR or in an early cycle of
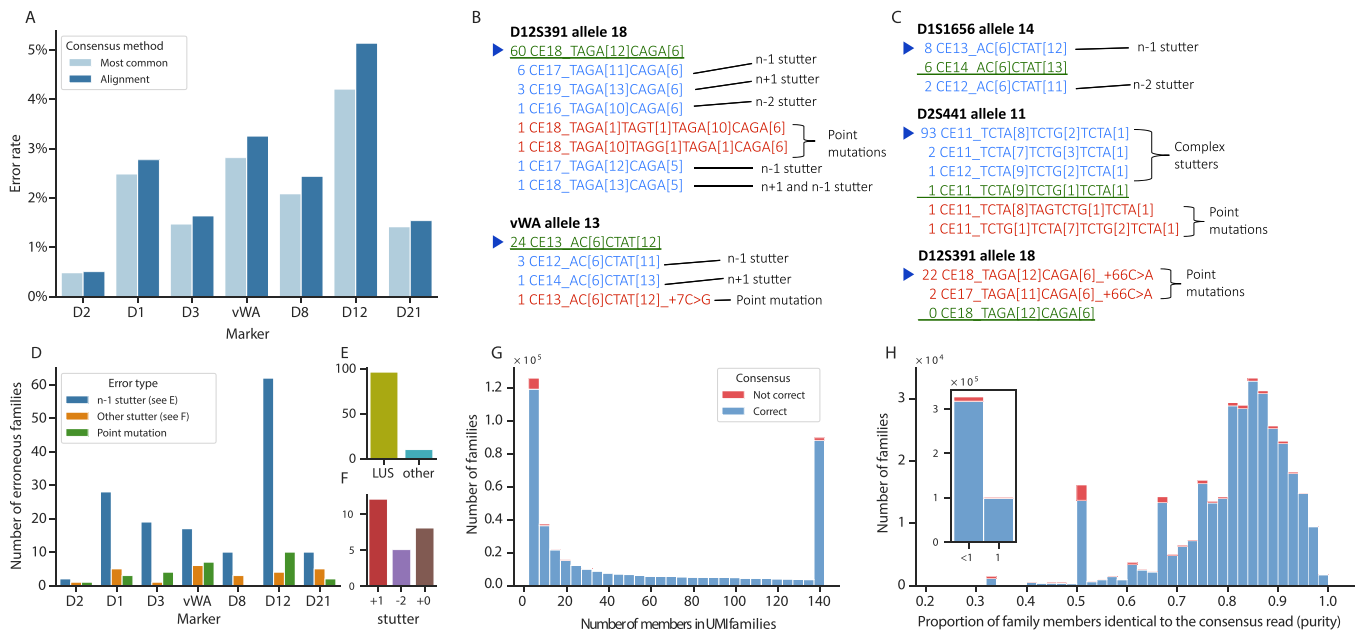
Fig. 2. Generation of consensus reads from UMI families. A. Proportion of incorrect consensus reads per marker using either "most common" or "multiple sequence alignment" (Alignment) consensus methods. B-C. Examples of UMI families with sequences and numbers of reads. Sequences are given in STRnaming format. These families gave (B) correct and (C) incorrect consensus reads, with the correct sequence underlined (green) and the proposed cause of error indicated in blue (stutters) and red (single base errors). The most common, i.e., selected, consensus read is indicated by a triangle. D. Type of errors among 200 randomly selected UMI families with incorrect consensus reads. E. Whether n-1 stutter in D occurred in the longest uninterrupted repeat sequence (LUS), or other repeat sequence, summarized for markers D8S1179, D12S391 and D21S11. F. Other types of stutter in D, summarized for all markers. G. Number of members in UMI families and whether they supported the correct (blue) or incorrect (red) allele, respectively. The rightmost bar includes families with at least 140 family members. H. Proportion of family members identical to the consensus read (purity) with correct (blue) and incorrect (red) allele, respectively. Inset shows proportion where all members are identical to consensus, main plot shows families with any deviant members. The peaks at 0.5, 0.66 and 0.75 are caused by families with 2, 3 and 4 members. See Supplementary Fig. S6 for marker-wise plots of G-H.

the adaptor PCR. In the former case, the correct template sequence is expected to be absent in the erroneous family. In the latter case, the template sequence should be present at a level close to the most common incorrect sequence, as long as the family is well sampled. Accordingly, when families with incorrect consensus reads were inspected, both families containing the correct sequence (but overwhelmed by incorrect reads) and families without the correct sequence were found (Fig. 2C). The dominant type of consensus read error was n-1 stutters (Fig. 2D). Most of these stutters occurred in the Longest Uninterrupted Stretch (LUS), but about one in ten were found in shorter repeat stretches (Fig. 2E). Single base errors, as well as n+1, n-2 and "zero" stutter (*i.e.*, the simultaneous loss and gain of repeats in different repeat stretches), were also observed (Fig. 2F).

For the dataset with 47 single-source samples, a majority of the UMI families had less than 30 member reads. However, there was a substantial long tail of large families with over 100 members (Fig. 2G). As could be intuitively expected, families with fewer members (below ten) were found to have less reliable consensus reads (Fig. 2G). About three-quarters of the families contained reads which were different from the consensus read, *i.e.*, had a purity of less than one (inset of Fig. 2H). Those where all reads were identical were predominantly families with five or fewer members, which may have been insufficiently sampled to discover any deviant members. In the families with a consensus read identical to the correct allele, it typically made up most of the member reads (Fig. 2H, Supplementary Figure S6).
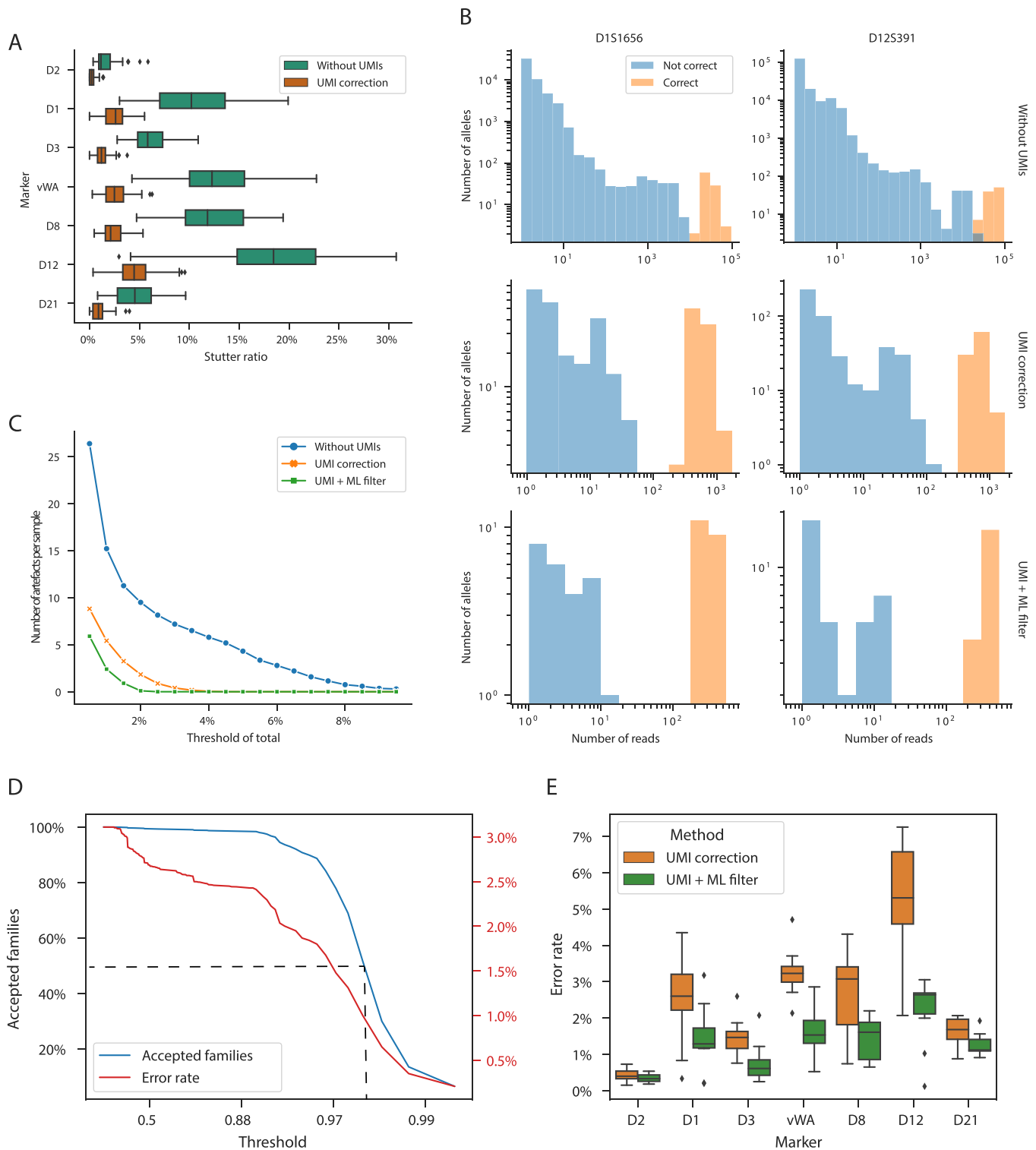
3.2. Error reduction with UMIs

The effect on error reduction by using UMIs was initially investigated without using any acceptance thresholds or stutter filters. Without applying the UMIs, between 75% (D12S391) and 95% (D2S441) of the total number of sequencing reads supported the correct genotypes for

the 47 single-source samples. Application of the UMIs led to a significantly lowered incidence of errors (P < 0.001, two-tailed paired t-test), as 95% (D12S391) to 99% (D2S441) of the total number of consensus reads coincided with the correct genotypes.

Both with and without UMIs, n-1 stutters in the LUS were the most abundant type of artefact. Before applying the UMI information, the average ratio of n-1 stutters (relative to parental allele) ranged from 1.6% (SD = 1.1%) for D2S441 to 18% (SD = 6.0%) for D12S391 (Fig. 3A, Supplementary Table S7). Focusing on the most heavily affected marker, D12S391, the n-1 stutter ratio ranged from 3.0% to 31% for the 82 observations. Using the UMI families to generate consensus reads led to drastic reductions in the n-1 stutter ratios. D2S441 and D12S391 still showed the lowest and highest incidence of n-1 stutters, respectively, at 0.3% (SD = 0.3%) and 4.6% (SD = 2.1%) of the parental allele (Fig. 3A). For D12S391, the lowest stutter ratio recorded was 0.4% and the highest was 9.6%. Overall, the seven STR markers showed four to six times lower stutter ratios when using the UMIs.

Other than the n-1 stutters, the artefacts included other stutters, single base errors and combinations thereof. Prior to applying the UMIs and without using any acceptance thresholds, each marker showed a multitude of different errors. For example, for 2800 M at 1 ng input and the marker D12S391, there were 3 820 different artefacts of which 2 693 were represented by single reads and 179 had above ten reads. The most common artefacts other than n-1 stutters were n-2 and n+1 stutter in LUS or in other repeat segments. The stutter ratios were between 0.5% and 3% of the parental allele read number. Applying UMIs, the number of detected artefacts dropped to twelve of which five were single consensus reads. The remaining n-2 and n+1 stutters were all below 0.3% of the parental allele. For D1S1656 in the same 2800 M sample, there were 1 121 different artefacts of which 677 were singletons and 35 had more than ten reads. The ratios for stutters other than n-1 were

Fig. 3. Error reduction by using UMI families to generate consensus reads. A. Ratio of n-1 stutters (relative to parental allele) without UMI information (green) and after using the UMIs to generate consensus reads (orange) for single-source samples. The boxplots show median values, the first and third quartiles and the whiskers 1.5 interquartile ranges, dots represent outliers (n=48 samples, number of stutters between 69 and 82). B. Distribution of read counts (or consensus read counts) for correct alleles (orange) and artefacts (blue) for single-source samples, determined without using the UMI information (top), with UMI consensus reads (middle) and with UMIs and ML filtration (bottom) (top, middle n=48, bottom n=10 samples). See Supplementary Fig. S9 for all markers. C. Number of artefacts per sample depending on method and allele calling threshold, expressed as a proportion of total reads for marker. No stutter filter was applied (blue, orange n=48, green n=10 samples). D. Tradeoff between error rate and number of consensus reads when applying ML filter. Proportion of accepted UMI families (left hand y scale) and proportion of incorrect consensus reads (right hand y scale) depending on chosen acceptance threshold according to the NN model, summarized for all markers. See Supplementary Fig. S1D-E for marker-separate plots (n=10 samples). E. Proportion of incorrect consensus reads with UMI correction or with UMIs and ML model, with thresholds set for each marker to accept half the number of families as the fixed thresholds. The boxplots show median values, the first and third quartiles and the whiskers 1.5 interquartile ranges, dots represent outliers (n=10 samples).

0.1–2.4% of the parental allele and below 1% of the total number of reads for the marker. Using UMIs reduced the number of artefacts to seven of which two were single consensus reads. The two remaining stutters had two to three consensus reads each, corresponding to less than 0.3% of the parental allele. Similar results were obtained for all markers and samples. Without using the UMIs, D12S391 and D21S11 showed the highest numbers of artefacts (about 3 500–5 000, Supplementary Figure S7A) while the other markers had around 1 000 artefacts each. With UMIs, the number of artefacts was reduced to five to fifteen (Supplementary Figure S7B).

Heterozygote balance, *i.e.*, the ratio between the read numbers of the two alleles of a heterozygous genotype, was significantly improved by applying UMIs (P < 0.001, two-tailed paired t-test). The improvement was most notable for the markers with the largest amplicon sizes (D8S1179: 0.89 vs 0.83, D12S391: 0.86 vs 0.80 and D21S11: 0.90 vs 0.86, see Supplementary Figure S8A, Supplementary Table S8).

In MPS analysis, acceptance thresholds are commonly used to determine if a detected sequence variant should be called or rejected. Allele calling thresholds are often set as a minimum number of reads and a proportion of the total number of reads of a marker. Here, usage of UMIs clearly improved the separation between the number of (consensus) reads observed for correct sequence variants and artefacts, respectively (Fig. 3B and Supplementary Figure S9). For example, for D12S391 the correct allele with the fewest reads had only 1.1 times more reads compared to the most abundant artefact without using UMIs (20 095 reads versus 18 316 reads). With UMIs this margin improved to a 4.7-fold difference (343 versus 73). Similar results were obtained for D1S1656; without using the UMIs the least observed correct sequence variant had two times as many reads as the most abundant artefact (14 432 versus 7 328). Applying UMIs elevated the margin to 7-fold (277 versus 41). This improved separation enables the use of robust allele calling thresholds that minimize both false positives and false negatives.

The effect of various allele calling thresholds set as proportions of the total number of reads of a marker is shown in Fig. 3C. Without UMIs, setting the acceptance threshold to 13% was necessary to remove all artefacts. Using the UMIs to form consensus reads, a threshold of 4.5% had the same effect. Without UMIs, a 4.5% allele calling threshold would result in an average of five artefacts per sample. Note that no stutter filters were applied in this example.

3.3. Applying machine learning to improve UMI error reduction

Using UMIs to generate consensus reads resulted in a substantial decrease in errors. It was hypothesized that additional information present in the UMI families, but currently not used for consensus read generation, could be utilized to further improve error reduction. Machine learning (ML) was thus applied on UMI family features to determine whether the families giving erroneous consensus reads could be excluded as unreliable. A neural network was trained to assign each family a classification score, with a higher score indicating that the consensus read is more likely to be correct (see Methods). Families with scores below a set ML score threshold are then discarded. The user can adjust the ML score threshold to achieve the desired balance between fidelity and the number of consensus reads (*i.e.*, proportion of accepted families) (Fig. 3D, Supplementary Figure 1D-E). Since the included markers have different consensus read error levels, separate ML score thresholds may be set for each marker. Here, ML score thresholds that reduce the number of consensus reads for each marker to approximately half compared to the UMI correction without ML filtration (*i.e.*, with fixed thresholds on purity and number of members) were applied for demonstration purposes. When applied to the 10 samples in the test set, this setting was found to decrease the proportion of incorrect consensus reads for all markers without losing any allele information (Fig. 3E). Other ML score thresholds may be applied. The effect of using ML score thresholds that yield similar numbers of reads as the fixed three member/50% purity thresholds is shown in the Supplementary Tables as "ML

low thresholds". As with applying UMIs in the first place, the markers with the highest error rates benefited the most from the ML filter; the largest improvement was observed for D12S391 with a reduced error rate from 5.2% (SD = 1.8%) with UMIs to 2.2% (SD = 0.9%) with UMIs and ML filter.

The ML filter gave a further decrease of the n-1 stutter ratios by a factor of two (Supplementary Figure S10), and improved the separation between true alleles and artefacts (Fig. 3B), enabling the use of even lower allele calling thresholds (Fig. 3C). For example, the calling threshold of 4.5% with UMIs could be lowered to about 2.5% with UMIs and ML filter while still eliminating all artefacts.

3.4. Analysis of low-template samples

All expected alleles were detected with full concordance to known allele values for all DNA amounts from 62.5 pg to 2 ng. Below 500 pg the amplicon quantity decreased, as seen in the fluorometric quantification and library quality control (Supplementary Table S3). The libraries were pooled equimolarly to compensate for any differences. However, for the lowest amount tested (62.5 pg DNA), most of the sequences still did not map to any of the STR markers. For 62.5 pg DNA, 14–22% of the total number of raw reads mapped to STR markers, as compared to 48–70% for 1 ng DNA (Supplementary Table S5 and Supplementary Figure S2).

The number of consensus reads per marker increased linearly with the DNA input from 62.5 pg to 1 ng (Fig. 4A, slope and r^2 per marker D2S441: 0.43, 0.983; D1S1656: 0.56, 0.988; D3S1358: 0.56, 0.986; vWA: 0.46, 0.994; D8S1179: 0.33, 0.980; D12S391: 0.62, 0.996; D21S11: 0.66, 0.998). The heterozygote balance decreased with lowered DNA input, as expected due to sampling effects and stochastic amplification, regardless of if the UMIs were used or not (Supplementary Figure S8B).

Without UMIs, the rate of erroneous reads was constant at around 16% (SD = 0.8%) over a 32-fold range of DNA input levels. Using UMIs with ML filter, the error rate was more than tenfold lower at 1.3% (SD = 0.4%) and stable between the dilutions, indicating that the UMI methodology is applicable for a wide range of template concentrations (Fig. 4B). Notably, the rate of acceptance of UMI families by the ML filter was stable from 62.5 pg to 1 ng, despite the model only being trained on 1 ng. Generation of consensus reads led to a large reduction or errors for samples with 250 pg DNA or less (Fig. 4C). For example, applying an allele calling threshold of 1% gave on average 15 artefacts per sample when studying sequencing reads, five artefacts when using the UMIs and three with UMIs and the ML filter.

3.5. Analysis of DNA mixtures

Almost all expected alleles were detected (994 of 1 002 possible correct alleles, 99.2%) when analyzing 12 three- to five-person mixtures with minor contributions of 10 pg to 150 pg DNA and ratios of 1–15% relative to the major donor, and applying UMIs and ML filter. In total, eight drop-out alleles were observed, distributed over four samples. Setting an allele calling threshold as a percentage of total reads for a locus is a compromise between the risk of false negatives (drop-outs) and false positive sequence variants (*e.g.*, stutter artefacts). The results for all the analyzed mixtures were combined and compared in terms of the relation between drop-outs and artefacts per sample when applying different allele calling thresholds (Fig. 5A). For example, when setting the threshold to 0.7% almost all expected alleles were detected. Without UMIs, there were on average 17 artefacts per sample, whereas using UMIs gave 4.4 and UMIs and ML filter 2.3. Using a threshold of 2.1% resulted in 12% drop-out alleles both with and without using the UMIs, whereas the number of artefacts was six-fold lower with UMIs. Applying the ML filter further improved the outcome, as an acceptance threshold of 2.1% showed a reduction from 1 to 0.6 artefacts per sample.

As an example, the obtained sequence variants and corresponding
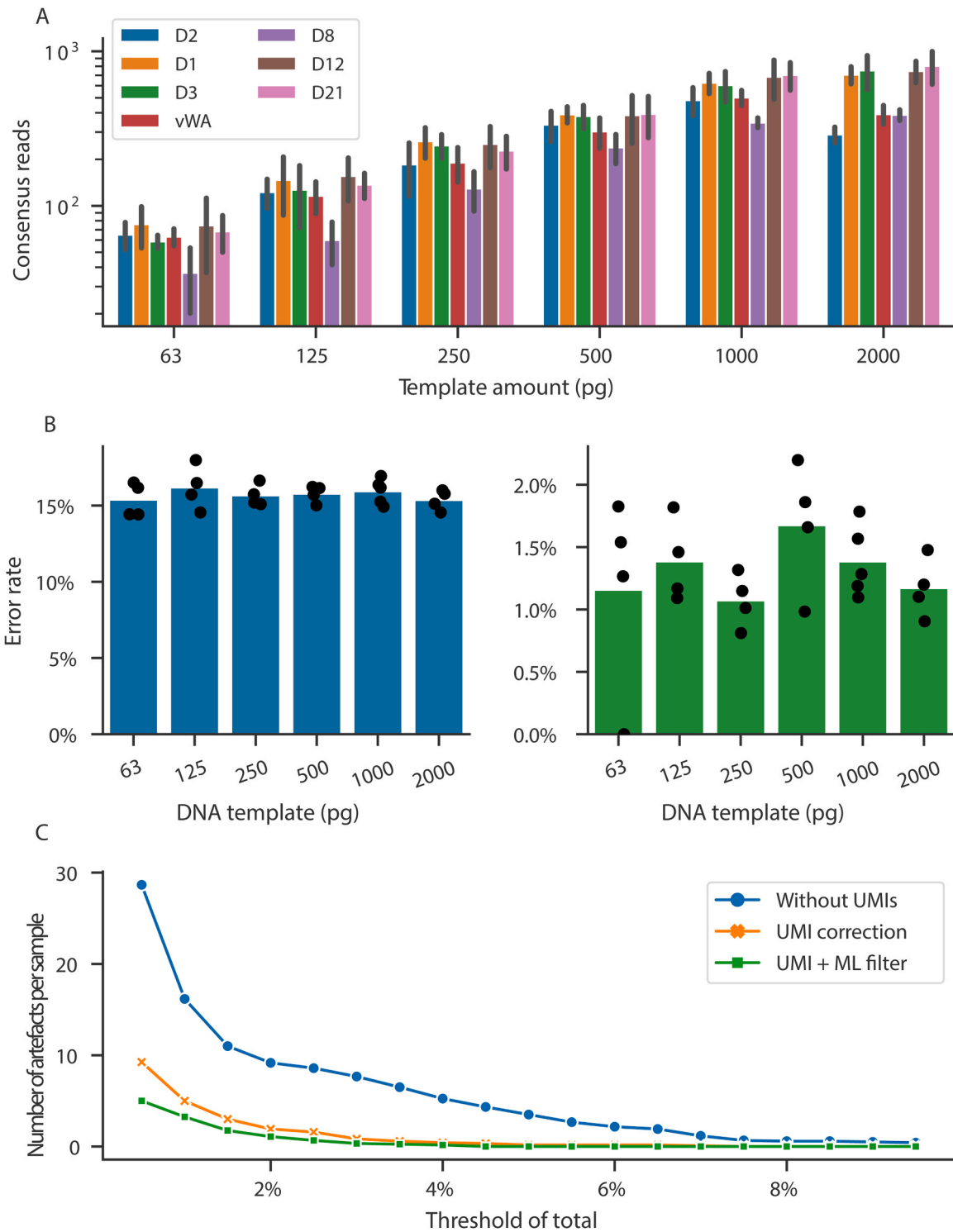
Fig. 4. Effect of UMIs on low-template DNA analysis. A. Number of consensus reads per marker obtained with UMI + ML filter. Error bar indicates SD. B. Proportion of reads for incorrect alleles without UMI correction (left) and with UMIs and ML filter (right). Dots represent each data point. C. Number of artefacts per sample with 250 pg or less template, for each allele calling threshold expressed as proportion of total reads for each marker. No stutter filter was applied. A-C: n=4 samples at each concentration.

numbers of sequencing reads (without using the UMIs) or consensus reads (with UMIs) are displayed for one of the most complex mixtures (markers D12S391 and D21S11, Fig. 5B-E). The sample is a 4-person mixture with one minor contributor constituting 1% (10 pg) of the DNA and with another minor at 5% (50 pg). Without UMIs, the D12S391 alleles for the two major contributors (P3 and P4) obtained high numbers of reads (above 15 000 per allele) and could easily be called

(Fig. 5B). However, the alleles for the two minor contributors (P1 and P2) may be difficult to call since at least three artefacts (corresponding to stutter artefacts of P3 and P4) show higher numbers of reads than the correct alleles. Stutter filters could be introduced, but this also comes with the risk of filtering true alleles. Similar results were obtained for D21S11, where the number of reads of the stutters were comparable to those of the minor contributor alleles (Fig. 5D). Applying the UMIs, the
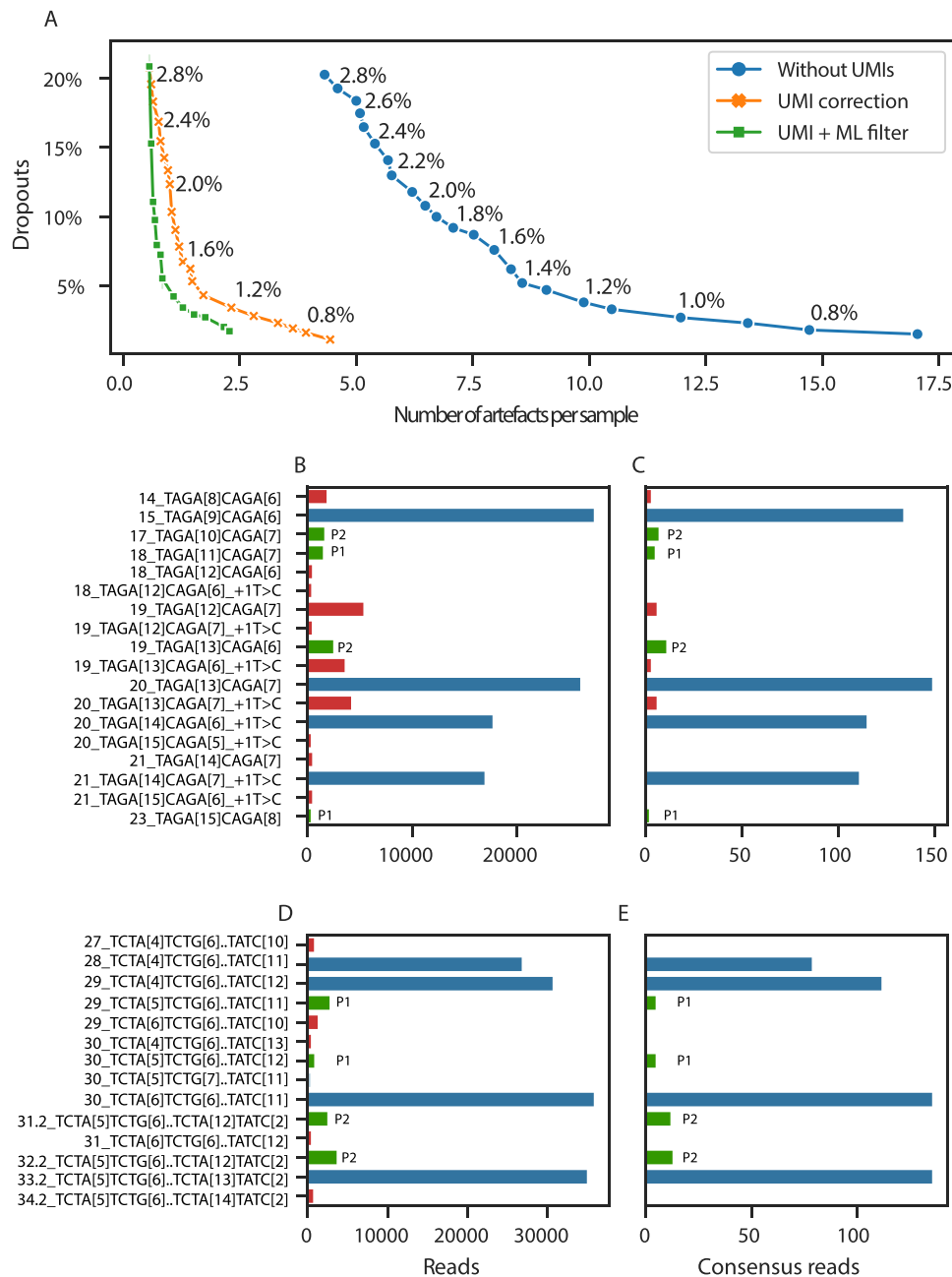
Fig. 5. Effect of UMIs on DNA mixtures. A. Average number of artefacts and drop-outs for the twelve mixtures (analyzed in duplicate) depending on method and allele calling threshold expressed as the proportion of total reads for each marker. The allele calling threshold was varied between 0.7% and 2.9% and is indicated above the data points. No stutter filter was applied. B. Alleles obtained for D12S391 without UMIs for a 4-person mixture with ratios of P4 at 47% (470 pg), P3 at 47% (470 pg), P2 at 5% (50 pg) and P1 at 1% (10 pg). Alleles with at least 10 reads and 1.2% of the highest allele are displayed. Blue indicates alleles for P3 and P4, green indicates alleles for P1 and P2 and red indicates artefacts. C. Same as B, but with UMIs and ML filter applied, and with a threshold of at least 2 consensus reads. D. Same as in B, but for the locus D21S11. E. Same as D, but with UMIs and ML filter applied, and with a threshold of at least 2 consensus reads.

number of artefacts were substantially fewer, and the stutter ratios were lower, making the evaluation of the minor contributor profiles more straightforward (Fig. 5C and E).

4. Discussion

The application of MPS in forensic STR analysis is increasingly researched and several assays have been presented [68–70]. A challenge in MPS-STR analysis is to distinguish the true alleles from artefacts such as stutter and single base errors. Here, applying UMIs and the SiMSen-Seq method made it possible to discern minute contributions in complex mixtures. Almost all expected alleles (99.2%) were detected in

mixtures with minor contributions of 10 pg to 150 pg DNA and ratios of 1–15% relative to the major donor with substantially lowered rates of artefacts. In other studies, using MPS without UMIs, about 10–80% of the minor contributor alleles were detected in mixtures where the minor constituted 5% of the total DNA amount [5,8,22,69].

In a previous study applying UMIs for STR analysis with a ligation-based method, it was demonstrated that UMIs led to improved allele assessment, but the limit of detection hindered some of the potential [41]. Here, analyzing low-template samples, 62.5 pg of DNA input gave complete STR profiles. Thus, the limit of detection when applying UMIs and SiMSen-Seq is comparable with other MPS-STR assays, which typically detect all expected alleles for 62.5 pg [22,69] or 250 pg [68,

71] depending on filtering strategies.

Studying all generated data, without applying any thresholds or stutter filters, when examining the effect of UMIs on sequencing quality enabled a deeper understanding of the impact of UMIs and SiMSen-Seq. Primarily, many artefacts showed substantially lower relative incidence and thousands of incorrect singletons were discarded. This provides the opportunity to use lower acceptance thresholds in casework. Others have demonstrated that filters and thresholds are among the most important factors for success when interpreting MPS-STR mixtures [7, 10,22,72], highlighting that the key challenge is to distinguish the true signal from noise [7]. Here, using the UMIs to generate consensus reads made it possible to apply acceptance thresholds of 0.5% of the total read number per locus with a low level of protruding errors. Note that this is without applying any stutter filters. Using such a low analytical threshold without applying the UMIs resulted in on average more than 25 artefacts per sample, which would be unacceptable when analyzing unknown traces. With UMIs, the number of artefacts dropped to nine per sample. Thus, the various thresholds and filters that are commonly applied to MPS-STR data [7,9,22] may be substantially lowered when using UMIs and SiMSen-Seq. The main artefact that remains after UMI correction is n-1 stutter. Applying a low, marker-specific stutter filter in combination with a threshold of 0.5% of the total read number per locus should lead to complete removal of stutters and single base errors while providing sensitive detection of the STR alleles.

One major benefit of applying UMIs and SiMSen-Seq was the great reduction of stutter artefacts, which were generally below 5% of the parental allele. Thus, stutter ratios were substantially lower than those found in current state-of-the-art capillary electrophoresis kits (Supplementary Figure S11) and in other MPS-STR studies [1,5,19,71]. In the latter, the stutter ratios were similar to the levels found here prior to applying the UMI information. The common strategy to handle stutters is to apply filters, but there are also more sophisticated solutions such as using models to distinguish noise from true alleles [19,73,74]. A few attempts have been made to reduce the incidence of stutters through biochemical modifications of PCR, such as the use of additives, lowered annealing temperatures, and application of alternative DNA polymerases [75]. Regarding the DNA polymerase, it has been suggested that the presence of a DNA binding domain or high-fidelity properties may reduce the incidence of stutters [16,76]. The DNA polymerase applied here, Platinum SuperFi II, has previously been found to result in lower stutter ratios compared to five other thermostable polymerases [77]. Optimized reaction conditions may be combined with UMIs to further reduce the occurrence of stutters. Through stutter reduction, it is shown that UMIs and SiMSen-Seq are powerful tools for minimizing systematic errors occurring in PCR-based library preparation. Most previous studies applying UMIs have been focused on random sequencing errors such as single base substitutions [33,34].

Using UMIs and SiMSen-Seq lead to a substantial reduction in sequencing errors which enables enhanced interpretation possibilities for complex mixtures. However, some artefacts persisted after consensus read generation (between 1% and 5% depending on STR marker). Since many of the incorrect families had similar properties, the application of an ML model on top of the UMI generation may further decrease the level of errors, as seen in previous work [40,41]. Here, the error reduction by ML filtration was most effective on the systematic errors created by stuttering. The ML filter worked well over a large range of DNA amounts and for complex mixtures. The filtration level may be chosen by the user to tune the desired balance between number of consensus reads and fidelity. In a practical application, different settings may be used for different STR markers (*e.g.*, D2S441 may not need filtration at all, while D12S391 requires stringent settings). This may be incorporated into the design of the PCR reaction. For example, tuning the primer concentrations so that better performing markers yield fewer pre-ML consensus reads while those with many errors produce larger numbers of consensus reads. The consensus reads may then be filtered until all markers are balanced both in terms of read counts and error levels.

Two sufficiently advanced ML model architectures (RF and differently sized NN models) performed similarly and had similar performance on the training and test data sets, respectively (data not shown). This suggests that adding more training data would bring little benefit and that the models are complex enough to capture the relationship, but it could still be beneficial to use more information from each family. Some consensus read generation errors appear inevitable, *e.g.*, when an error occurred in the barcoding PCR and all family members are identical but wrong. However, including more measures describing the families*, i.e.*, model input features, may improve the discrimination further in other cases. Alternatively, entirely different model designs could be used to avoid the information bottleneck represented by a limited number of numeric features. A model that takes the family members and their sequences directly as input would be exposed to all information present in the UMI family and would be able to make the most informed decision. Here, the filter is used with a threshold to either discard of accept each family. It is also possible to forward the score to a specialized allele calling method to use as a measure of reliability for each consensus read as has been demonstrated previously [41].

Another alternative is a STR-aware algorithm that generates a reliable consensus read (or discards unreliable UMI families) directly from the members. Taking sequencing quality information into account may specifically improve the quality of consensus reads from families with very few members, where sequencing errors have a larger impact. Such an algorithm could be either heuristic or based on first principles, and could possibly generate a larger number of reliable consensus sequences than the naïve "most common" method plus a post-hoc filter as in this work. We are, however, not aware of any suitable model algorithms available at the time of writing. Further, both when giving a filtering model access to all information contained in the families or if using a ML model to generate consensus reads directly, great care should be taken to prevent that the models become biased to the alleles present in the training data.

5. Conclusions

UMIs and SiMSen-Seq are promising tools for improved forensic STR profiling. The proof-of-concept seven STR multiplex yielded concordant alleles for 47 single-source samples at 1 ng as well as for low-template samples at 62.5 pg (the lowest amount tested). Minor contributions at 10 pg or 1% of the total DNA amount were detected in complex mixtures. The main impact of UMIs and SiMSen-Seq was a reduction of errors, seen as lowered numbers of artefacts and greatly reduced stutter ratios. Thus, the UMIs removed both random and systematic errors. Application of an ML model on the UMI families led to a decrease in erroneous consensus reads. Overall, the SiMSen-Seq method provided for better separation between true alleles and artefacts, which makes it possible to apply substantially lower allele calling thresholds and stutter filters compared to regular MPS analysis. Lowered thresholds and filters, in turn, may lead to improved detection of minute DNA amounts. As UMIs are relatively cheap and straightforward to incorporate into existing sequencing methods there is great potential for wide application within forensic STR sequencing, allowing better interpretation of, for example, complex mixtures.

CRediT authorship contribution statement

Joakim Håkansson: Writing – review & editing, Resources, Project administration, Funding acquisition. **Becky Steffen:** Writing – review & editing, Investigation. **Yalda Bogestal:** Writing – review & editing, Resources, Project administration, Funding acquisition. **Tobias Österlund:** Writing – review & editing, Software. **Peter Rådström:** Writing – review & editing, Methodology. **Andreas Tillmar:** Writing – review & editing, Methodology. **Peter Vallone:** Writing – review & editing, Resources, Conceptualization. **Arvid Gynnå:** Writing – original

draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Anders Ståhlberg:** Writing – review & editing, Resources, Methodology. **Maja Sidstedt:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Linda Jansson:** Writing – review & editing, Investigation. **Johannes Hedman:** Writing – original draft, Validation, Supervision, Methodology, Funding acquisition, Conceptualization. **Kevin Kiesler:** Writing – review & editing, Methodology. **Gustav Johansson:** Writing – review & editing, Methodology.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Anders Ståhlberg is co-inventor of the SiMSen-Seq technology that is patent protected (U.S. Serial No.:15/552,618). Anders Ståhlberg declares stock ownership and is a board member in Tulebovaasta, Iscaff Pharma and SiMSen Diagnostics. Gustav Johansson declares employment and stock ownership in SiMSen Diagnostics.

Conflict of interest

A.S. is co-inventor of the SiMSen-Seq technology that is patent protected (U.S. Serial No.:15/552,618). A.S. declares stock ownership and is a board member in Tulebovaasta, Iscaff Pharma and SiMSen Diagnostics. G.J. declares employment and stock ownership in SiMSen Diagnostics.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2024.103047.

References

[1] S.L. Fordyce, H.S. Mogensen, C. Børsting, R.E. Lagace, C.W. Chang, N. Rajagopalan, N. Morling, Second-generation sequencing of forensic STRs using the Ion Torrent HID STR 10-plex and the Ion PGM. Forensic Sci. Int Genet 14 (2014) 132–140.

[2] Y. Yang, B. Xie, J. Yan, Application of Next-generation Sequencing Technology in Forensic Science, Genom., Proteom. Bioinforma. 12 (2014) 190–197.

[3] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, R. A. Guerrieri, P.M. Vallone, Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Forensic Sci. Int Genet 21 (2015) 15–21.

[4] K.B. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, U.S. Population Sequence Data for 27 Autosomal STR Loci, Forensic Sci. Int. Genet. 37 (2018) 106–115.

[5] K.J. van der Gaag, R.H. de Leeuw, J. Hoogenboom, J. Patel, D.R. Storts, J.F. Laros, P. de Knijff, Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq system, Forensic Sci. Int Genet 24 (2016) 86–96.

[6] D. Ballard, J. Winkler-Galicki, J. Wesoły, Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects, Int. J. Leg. Med. 134 (2020) 1291–1303.

[7] Ø. Bleka, R. Just, J. Le, P. Gill, An examination of STR nomenclatures, filters and models for MPS mixture interpretation, Forensic Sci. Int. Genet. 48 (2020) 102319.

[8] C.C.G. Benschop, K.J. van der Gaag, J. de Vreede, A.J. Backx, R.H. de Leeuw, S. Zuñiga, J. Hoogenboom, P. de Knijff, T. Sijen, Application of a probabilistic genotyping software to MPS mixture STR data is supported by similar trends in LRs compared with CE data, Forensic Sci. Int. Genet. 52 (2021) 102489.

[9] K.J. van der Gaag, J. Hoogenboom, L. Busscher, C.C.G. Benschop, S. Zuñiga, T. Sijen, The impact of FDSTools noise correction on the analysis of data from the Forenseq™ DNA Signature Prep Kit. Forensic Sci. Int. Genet. Suppl. Ser. 7 (2019) 797–799.

[10] J. Hoogenboom, K.J. van der Gaag, R.H. de Leeuw, T. Sijen, P. de Knijff, J.F. J. Laros, FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise, Forensic Sci. Int. Genet. 27 (2017) 27–40.

[11] K. Best, T. Oakes, J.M. Heather, J. Shawe-Taylor, B. Chain, Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding, Sci. Rep. 5 (2015) 14629.

[12] M.A. Quail, I. Kozarewa, F. Smith, A. Scally, P.J. Stephens, R. Durbin, H. Swerdlow, D.J. Turner, A large genome center's improvements to the Illumina sequencing system, Nat. Methods 5 (2008) 1005–1010.

[13] E.J. Fox, K.S. Reid-Bayliss, M.J. Emond, L.A. Loeb, Accuracy of next generation sequencing platforms, Gener., Seq. Appl. 1 (2014) 1000106.

[14] D. Andersson, H. Kristiansson, M. Kubista, A. Ståhlberg, Ultrasensitive circulating tumor DNA analysis enables precision medicine: experimental workflow considerations, Expert Rev. Mol. Diagn. 21 (2021) 299–310.

[15] R. Li, R. Wu, H. Li, Y. Zhang, D. Peng, N. Wang, X. Shen, Z. Wang, H. Sun, Characterizing stutter variants in forensic STRs with massively parallel sequencing, Forensic Sci. Int. Genet. 45 (2020) 102225.

[16] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence Analysis and Characterization of Stutter Products at the Tetranucleotide Repeat Locus VWA, Nucleic Acids Res 24 (1996) 2807–2812.

[17] R.A. Aponte, K.B. Gettings, D.L. Duewer, M.D. Coble, P.M. Vallone, Sequence-based analysis of stutter at STR loci: Characterization and utility, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e456–e458.

[18] O. D'Angelo, A.C.W. Vandepoele, J. Adelman, M.A. Marciano, Assessing non-LUS stutter in DNA sequence data, Forensic Sci. Int. Genet. 59 (2022) 102706.

[19] M.M. Agudo, H. Aanes, A. Roseth, M. Albert, P. Gill, Ø. Bleka, A comprehensive characterization of MPS-STR stutter artefacts, Forensic Sci. Int. Genet. 60 (2022) 102728.

[20] A.C. Jäger, M.L. Alvarez, C.P. Davis, E. Guzmán, Y. Han, L. Way, P. Walichiewicz, D. Silva, N. Pham, G. Caves, J. Bruand, F. Schlesinger, S.J.K. Pond, J. Varlaro, K. M. Stephens, C.L. Holt, Developmental validation of the MiSeq FGx Forensic Genomics System for targeted next generation sequencing in forensic DNA casework and database laboratories, Forensic Sci. Int. Genet. 28 (2017) 52–70.

[21] S. Köcher, P. Müller, B. Berger, M. Bodner, W. Parson, L. Roewer, S. Willuweit, Inter-laboratory validation study of the ForenSeq™ DNA Signature Prep Kit, Forensic Sci. Int. Genet. 36 (2018) 77–85.

[22] K.M. Stephens, R. Barta, K. Fleming, J.C. Perez, S.-F. Wu, J. Snedecor, C.L. Holt, B. LaRue, B. Budowle, Developmental validation of the ForenSeq MainstAY kit, MiSeq FGx sequencing system and ForenSeq Universal Analysis Software, Forensic Sci. Int. Genet. 64 (2023) 102851.

[23] P. Gill, H. Haned, O. Bleka, O. Hansson, G. Dørum, T. Egeland, Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches—Twenty years of research and development, Forensic Sci. Int. Genet. 18 (2015) 100–117.

[24] M.D. Coble, J.-A. Bright, Probabilistic genotyping software: an overview, Forensic Sci. Int. Genet. 38 (2019) 219–224.

[25] H.Y. Wong, J. Tan, Z.G. Lim, R. Kwok, W. Lim, C.K.-C. Syn, DNA profiling success rates of commonly submitted crime scene items, Forensic Sci. Int. Genet. Suppl. Ser. 7 (2019) 597–599.

[26] M.N. Krosch, Variation in forensic DNA profiling success among sampled items and collection methods: a Queensland perspective, Aust. J. Forensic Sci. 53 (2021) 612–625.

[27] V. Sharma, H.Y. Chow, D. Siegel, E. Wurmbach, Qualitative and quantitative assessment of Illumina's forensic STR and SNP kits on MiSeq FGx. PLOS One 12 (2017) e0187932.

[28] C. Brandariz-Fontes, M. Camacho-Sanchez, C. Vilà, J.L. Vega-Pla, C. Rico, J. A. Leonard, Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results, Sci. Rep. 5 (2015) 8056.

[29] M. Schirmer, R. D'Amore, U.Z. Ijaz, N. Hall, C. Quince, Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data, BMC Bioinforma. 17 (2016) 125.

[30] Q. Peng, R. Vijaya Satya, M. Lewis, P. Randad, Y. Wang, Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes, BMC Genom. 16 (2015) 589.

[31] J.A. Sena, G. Galotto, N.P. Devitt, M.C. Connick, J.L. Jacobi, P.E. Umale, L. Vidali, C.J. Bell, Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis, Sci. Rep. 8 (2018) 13121.

[32] R. Kou, H. Lam, H. Duan, L. Ye, N. Jongkam, W. Chen, S. Zhang, S. Li, Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations, PLOS One 11 (2016) e0146638.

[33] T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, J. Taipale, Counting absolute numbers of molecules using unique molecular identifiers, Nat. Methods 9 (2012) 72–74.

[34] D.B. Sloan, A.K. Broz, J. Sharbrough, Z. Wu, Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods, Trends Biotechnol. 36 (2018) 729–740.

[35] I. Kinde, J. Wu, N. Papadopoulos, K.W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 9530–9535.

[36] M.L. McCloskey, R. Stöger, R.S. Hansen, C.D. Laird, Encoding PCR Products with Batch-stamps and Barcodes, Biochem. Genet. 45 (2007) 761–767.

[37] D.W. Craig, J.V. Pearson, S. Szelinger, A. Sekar, M. Redman, J.J. Corneveaux, T. L. Pawlowski, T. Laub, G. Nunn, D.A. Stephan, N. Homer, M.J. Huentelman, Identification of genetic variants using bar-coded multiplexed sequencing, Nat. Methods 5 (2008) 887–893.

[38] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, S. Linnarsson, Quantitative single-cell RNA-seq with unique molecular identifiers, Nat. Methods 11 (2014) 163–166.

[39] A. Staadig, J. Hedman, A. Tillmar, Applying Unique Molecular Indices with an Extensive All-in-One Forensic SNP Panel for Improved Genotype Accuracy and Sensitivity, Genes 14 (2023) 818.

[40] A.E. Woerner, S. Mandape, J.L. King, M. Muenzler, B. Crysup, B. Budowle, Reducing noise and stutter in short tandem repeat loci with unique molecular identifiers, Forensic Sci. Int. Genet. 51 (2021) 102459.

[41] B. Crysup, S. Mandape, J.L. King, M. Muenzler, K.B. Kapema, A.E. Woerner, Using unique molecular identifiers to improve allele calling in low-template mixtures, Forensic Sci. Int. Genet. 63 (2023) 102807.

[42] A. Ståhlberg, P.M. Krzyzanowski, J.B. Jackson, M. Egyud, L. Stein, T.E. Godfrey, Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing, Nucleic Acids Res 44 (2016) e105.

[43] A. Ståhlberg, P.M. Krzyzanowski, M. Egyud, S. Filges, L. Stein, T.E. Godfrey, Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing, Nat. Protoc. 12 (2017) 664–682.

[44] G. Johansson, D. Andersson, S. Filges, J. Li, A. Muth, T.E. Godfrey, A. Ståhlberg, Considerations and quality controls when analyzing cell-free tumor DNA, Biomol. Detect Quantif. 17 (2019) 100078.

[45] S. Filges, E. Yamada, A. Ståhlberg, T.E. Godfrey, Impact of Polymerase Fidelity on Background Error Rates in Next-Generation Sequencing with Unique Molecular Identifiers/Barcodes, Sci. Rep. 9 (2019) 3503.

[46] C.R. Steffen, E.L. Romsos, K.M. Kiesler, L.A. Borsuk, K.B. Gettings, P.M. Vallone, Make it "SNPPY" - Updates to SRM 2391d: PCR-Based DNA Profiling Standard, Forensic Sci. Int. Genet. Suppl. Ser. 8 (2022) 9–11.

[47] K. Gettings, Forensic DNA Open Dataset, Natl. Inst. Stand. Technol. (2020), https://doi.org/10.18434/M32157. Accessed 2023-04-04.

[48] L. Devesse, D. Ballard, L. Davenport, I. Riethorst, G. Mason-Buck, D. Syndercombe Court, Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups, Forensic Sci. Int. Genet. 34 (2018) 57–61.

[49] Short Tandem Repeat DNA Internet Database, STRBase (SRD-130), National Institute of Standards and Technology (NIST), https://strbase.nist.gov (Accessed 2023-04-06).

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

[51] M.D. Coble, J.M. Butler, Characterization of new miniSTR loci to aid analysis of degraded DNA, J. Forensic Sci. 50 (2005) 43–53.

[52] M.V. Lareu, S. Barral, A. Salas, C. Pestoni, A. Carracedo, Sequence variation of a hypervariable short tandem repeat at the D1S1656 locus, Int. J. Leg. Med. 111 (1998) 244–247.

[53] B.E. Krenke, A. Tereba, S.J. Anderson, E. Buel, S. Culhane, C.J. Finis, C.S. Tomsey, J.M. Zachetti, A. Masibay, D.R. Rabbach, E.A. Amiott, C.J. Sprecher, Validation of a 16-locus fluorescent multiplex system, J. Forenisc Sci. 47 (2002) 773–785.

[54] S. Filges, P. Mouhanna, A. Ståhlberg, Digital Quantification of Chemical Oligonucleotide Synthesis Errors, Clin. Chem. 67 (2021) 1384–1394.

[55] T. Österlund, S. Filges, G. Johansson, A. Ståhlberg, UMIErrorCorrect and UMIAnalyzer: Software for Consensus Read Generation, Error Correction, and Visualization Using Unique Molecular Identifiers, Clin. Chem. 68 (2022) 1425–1435.

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

[57] M. Schubert, S. Lindgreen, L. Orlando, AdapterRemoval v2: rapid adapter trimming, identification, and read merging, BMC Res. Notes 9 (2016) 88.

[58] T.S. Smith, A. Heger, I. Sudbery, UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy, Genome Res 27 (2017) 491–499.

[59] A.E. Woerner, J.L. King, B. Budowle, Fast STR allele identification with STRait Razor 3.0, Forensic Sci. Int. Genet. 30 (2017) 18–23.

[60] S.L. Friis, A. Buchard, E. Rockenbauer, C. Børsting, N. Morling, Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs, Forensic Sci. Int. Genet. 21 (2015) 68–75.

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

[62] K. Katoh, D.M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability, Mol. Biol. Evol. 30 (2013) 772–780.

[63] J. Hoogenboom, T. Sijen, K.J. van der Gaag, STRNaming: Generating simple, informative names for sequenced STR alleles in a standardised and automated manner, Forensic Sci. Int. Genet. 52 (2021) 102473.

[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucker, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[65] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32.

[66] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Adv. Large Margin Classif. 10 (1999) 61–74.

[67] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.

[68] P. Müller, C. Sell, T. Hadrys, J. Hedman, S. Bredemeyer, F.-X. Laurent, L. Roewer, S. Achtruth, M. Sidstedt, T. Sijen, M. Trimborn, N. Weiler, S. Willuweit, I. Bastisch, W. Parson, The SeqFor, S.T.R.C, Inter-laboratory study on standardized MPS libraries: evaluation of performance, concordance, and sensitivity using mixtures and degraded DNA, Int. J. Leg. Med. 134 (2020) 185–198.

[69] J. Silvery, S. Ganschow, P. Wiegand, C. Tiemann, Developmental validation of the monSTR identity panel, a forensic STR multiplex assay for massively parallel sequencing, Forensic Sci. Int. Genet. 46 (2020) 102236.

[70] D.S.B.S. Silva, M.K. Scheible, S.F. Bailey, C.L. Williams, J.S. Allwood, R.S. Just, J. Schuetter, N. Skomrock, A. Minard-Smith, N. Barker-Scoggins, C. Eichman, K. Meiklejohn, S.A. Faith, Sequence-based autosomal STR characterization in four US populations using PowerSeq™ Auto/Y system, Forensic Sci. Int. Genet. 48 (2020) 102311.

[71] C. Hussing, C. Huber, R. Bytyci, H.S. Mogensen, N. Morling, C. Børsting, Sequencing of 231 forensic genetic markers using the MiSeq FGx forensic genomics system – an evaluation of the assay and software, Forensic Sci. Res. 3 (2018) 111–123.

[72] B. Young, J.L. King, B. Budowle, L. Armogida, A technique for setting analytical thresholds in massively parallel sequencing-based forensic DNA analysis, PLOS One 12 (2017) e0178005.

[73] S.B. Vilsen, T. Tvedebrink, P.S. Eriksen, C. Hussing, C. Børsting, N. Morling, Modelling allelic drop-outs in STR sequencing data generated by MPS, Forensic Sci. Int. Genet. 37 (2018) 6–12.

[74] O. Raz, T. Biezuner, A. Spiro, S. Amir, L. Milo, A. Titelman, A. Onn, N. Chapal-Ilani, L. Tao, T. Marx, U. Feige, E. Shapiro, Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise, Nucleic Acids Res 47 (2019) 2436–2445.

[75] S.B. Seo, J. Ge, J.L. King, B. Budowle, Reduction of stutter ratios in short tandem repeat loci typing of low copy number DNA samples, Forensic Sci. Int. Genet. 8 (2014) 213–218.

[76] E. Viguera, D. Canceill, S.D. Ehrlich, In vitro replication slippage by DNA polymerases from thermophilic organisms11Edited by K. Morikawa, J. Mol. Biol. 312 (2001) 323–333.

[77] E. Yamanoi, M. Sakurada, Y. Ueno, Low stutter ratio by SuperFi polymerase, Forensic Sci. Int.: Rep. 3 (2021) 100204.