Developmental Validation of the IDseek® OmniSTR™ Global autosomal STR profiling kit.

Erik A.C. de Jong, Melanie H.J. Arts, Pieter A.M. van Oers, Joop P.G. Theelen.

NimaGen BV, Hogelandseweg 88, 6545 AB, Nijmegen, The Netherlands

Abstract

Forensic science takes advantage of population variability in autosomal Short Tandem Repeat (STR) lengths to establish human identification. The most common method for DNA profiling by STR is based on PCR, where the highly polymorphic STR regions are amplified and analysed using Capillary Electrophoresis (CE) or Massively Parallel Sequencing (MPS). MPS determines not only the repeat length, but also the repeat structure and variations in the flanking regions, making this method superior in discriminatory power compared to CE. Reverse Complement PCR (RC-PCR) is a novel, more sophisticated PCR based MPS library preparation method combining indexing and PCR amplification in a single closed-tube reaction. In this document we describe the complete developmental validation of the IDseek® OmniSTR™ kit, a RC-PCR based MPS library preparation kit. The developed IDseek® OmniSTR™ kit contains 28 autosomal STR targets, one Y-chromosomal STR and the Amelogenin gene covering all relevant STR core loci from the USA, EU, UK and Interpol.

Keywords: RC-PCR, MPS, STR, Forensic DNA, human identification, OmniSTR™, IDseek®

1. Introduction

Forensic scientists continually strive to find the perfect match for their biological evidence, for instance to identify missing persons, confirm familial relationships or link persons of interest to crime scenes. Short Tandem Repeats (STRs) are currently the golden standard to use for forensic human identification [1]. STRs are short, repeated DNA sequences of 2-6bp in length that comprise approximately 3% of the human genome. The number of repeated units these STRs contain is highly variable between individuals in the human population, therefore providing high discriminatory power in terms of human identification [2].

To analyse PCR amplified STR regions, DNA analysis techniques Capillary Electrophoresis (CE) and Massively Parallel Sequencing (MPS), also known as Next Generation Sequencing (NGS), can be used. CE is a more traditional technique to analyse STRs, where the fragment size of each allele reflects the number of repeats. The forensic field is slowly switching to MPS, which allows more information as the full base pair composition of the allele is determined, also providing information about the repeat structure and variations in the flanking regions. Obtaining insight in the full base pair composition has several advantages. Firstly, it can help predict stutter behaviour. Secondly, having the exact base pair composition can significantly improve in deconvoluting mixed samples, which can aid in solving sexual assault, for example. Lastly, a major advantage of MPS over CE for forensic STR analysis is that the amplicon size ranges do not need to differ, because of the limited number of color channels used by CE. This enables the possibility of creating a multiplex pool with the smallest possible STR amplicons, facilitating enhanced amplification of degraded DNA [3].

Reverse-Complement PCR (RC-PCR) is a novel, more sophisticated PCR technology, that enables MPS library preparation where indexing, adapter tail addition and multiplex PCR amplification occur in a single closed-tube reaction. RC-PCR based MPS library preparation has a number of highly desirable advantages for forensic science. The first major advantage is the minimization of contamination and sample swapping chance, due to the closed tube reaction: Early indexing, no need to re-amplify PCR pools and reduced hands-on time and workflow steps compared to other MPS library preparation protocols. Secondly, it allows for the analysis of samples with very low DNA input which is highly preferred in forensic investigations as the sample material can be very compromised. RC-PCR kinetics result in high sensitivity and specificity as target specific primers are synthesized during the reaction, so concentrations of primers and amplicons are more in line, reducing potential primer dimerization and off-target primer binding. Lastly, it is possible to combine different samples in the library purification due to early indexing, which saves costs on consumables as well as time needed for the purification [4][5]. The RC-PCR is a simple, sensitive, safe and robust method for cost-effective and high-quality DNA analysis.

The Federal Bureau of Investigation (FBI) created and maintains the Combined DNA Index System (CODIS), a database that contains multiple DNA databases to support criminal justice. For privacy reasons, no personal identification data, such as names or phenotypic traits are stored in the CODIS databases. Most identification methods using CODIS are based on STRs. As of January 2017, 20 STR loci have been defined as the 'CODIS core loci', creating a reference system for forensic identifications. These specific loci were originally included due to their location in non-coding regions of the genome, which precluded identification of any phenotypic information. Most of the commercially available STR profiling kits include all CODIS core loci [2].

The developed IDseek® OmniSTR™ kit contains 28 autosomal STR targets, one Y-chromosomal STR and the Amelogenin gene covering all relevant STR core loci from the USA, EU, UK and Interpol. All amplicons are designed to meet the shortest possible fragment lengths and are compatible with Illumina® MiSeq™ and Illumina® (Qiagen/Verogen) MiSeq™ FGx systems, with 2 x 10 bp Unique Dual Index reads. The IDseek® OmniSTR™ kit can be used to establish the identity of missing persons, confirm kinship and link persons to crime scenes. Gender identification can be performed in conjunction with STR typing, using the PCR product generated from the Amelogenin gene on both the X- and Y-chromosome [6].

In this document we describe the complete developmental validation of the IDseek® OmniSTR™ kit, a RC-PCR based MPS library preparation kit. The IDseek® OmniSTR™ MPS library preparation kit is evaluated on both reference samples and direct PCR samples. In addition, the sensitivity of the kit is determined, mixtures up to 1:99 (M:F) are analysed, the effect of inhibitors such as tannic acid and humic acid is determined, the repeatability and reproducibility are investigated and at last, the human specificity is studied.

2. Methods

2.1 DNA samples

Five human genomic DNA samples were used during this study consisting of 2800M (male) (Promega® Corporation) and NA12877 (male), NA12878 (female), NA24143 (female), NA24631 (male) (Coriell institute for medical research). Additional DNA samples were obtained from 12 volunteers (6 male and 6 female) who signed an informed consent form, authorizing the use of their DNA for research purposes. Buccal cells were collected using sterile OmniSwabs (Qiagen®). Swab tips were ejected in sterile tubes and stored at -20 °C while awaiting further processing. From each volunteer, 2 mL of saliva was collected and purified using the Genefix™ saliva collector and saliva DNA isolation kit (Isohelix®). DNA samples were quantified prior to PCR using the Investigator® Quantiplex kit (Qiagen®) on the Quantstudio™ 1 (Applied biosystems™) and diluted to 125 pg/µL in TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0 (Invitrogen™)) for a total DNA input of 1 ng/reaction (8 µL sample input). No template controls (NTC) were included using molecular grade water (MilliQ® IQ 7005 system, Merck®), 2800M was used as a positive control.

2.2 Reference samples, sensitivity and Direct PCR

Performance of reference samples was assessed on the 12 samples isolated from saliva using the recommended input amount of 1 ng per reaction.

Sensitivity was assessed by preparing serial dilutions of 2800M and NA12878 for total DNA input amounts of 10 ng, 1 ng, 500 pg, 250 pg, 125 pg, 62.5 pg, 31.25 pg and 15.625 pg.

Direct PCR was performed by cutting of half a tooth from the buccal swabs (approximately 1.5 mm^2) which was then placed directly in the reaction tube. Reaction volumes were adjusted to 20 µL by adding 8 µL of molecular grade water. All reactions were run in triplicate.

2.3 Human specificity

Human specificity was assessed and kindly shared by the Netherlands Forensic Institute, Biological Traces Division. A minimal DNA input of 2 ng was used for six animal species which are likely to be encountered around humans (sheep, dwarf goat, cow, pig, dog and cat).

2.4 Mixtures

Mixtures were prepared from human genomic DNA using NA12877, NA12878, NA24631 and NA24143 in varying compositions. Mixtures of two males (NA12877 and NA24143) and two females (NA12878 and NA24143) were tested in the ratios of 1:2, 1:5, 1:10, 1:20. Mixtures of a male and female sample (NA12877 and NA24143) were prepared in the ratios 1:2, 1:5, 1:10, 1:20 and 1:99 (M:F). All mixture compositions were performed in triplicate with a total DNA input of 1 ng and each contributor was also analysed as a single source reference.

2.5 PCR inhibition

Inhibitor solutions were prepared for Humic acid, Tannic acid, Hematin and Indigo carmine (Merck®) and spiked into 2800M DNA in four different concentrations each. 8 µL of the spiked samples were used for RC-PCR resulting in a total human DNA input of 1 ng and reaction concentrations of 133.33 µM, 66.67 µM, 33.33 µM and 16.67 µM per inhibitor. Inhibition by bacterial DNA was assessed by spiking reactions with a bacterial DNA mixture consisting of equal parts *E. Coli*, *P. Aeruginosa* and *S.*

Aureus for a total bacterial DNA input of 100 ng, 50 ng, 20 ng and 10 ng. Control samples were spiked with molecular grade water. All reactions were run in triplicate.

2.6 Repeatability and reproducibility

Repeatability was assessed by letting a single technician setup triplicate reactions for eight different DNA samples (2800M, NA12877, NA12878, NA24143, NA24631 and three samples obtained from volunteers). Library preparation, purification and sequencing were all performed separately.

Reproducibility was assessed by letting three different technicians setup triplicate reactions for eight different DNA samples (2800M, NA12877, NA12878, NA24143, NA24631 and three samples obtained from volunteers). Library preparation and purification was performed separately for each technician, the resulting libraries were sequenced on a single flow cell.

2.7 Library preparation

MPS libraries were generated using the IDseek® OmniSTR™ global autosomal STR profiling kit (NimaGen® BV). Reaction setup was performed in a pre-PCR environment according to the OmniSTR™ instructions for use (IFU version 1.2 NimaGen® BV) and as follows. The RC-PCR mix was prepared by combining 10 µL of the HiFi mastermix, 0.2 µL of the RC-PCR probe panel and 1.8 µL of the probe dilution buffer per sample. The required number of strips were cut from the EasySeq™ 96-well dehydrated index primer plate (NimaGen® BV) and 12 µL of the RC-PCR mix was dispensed per reaction well. 8 µL of sample or control was then added to each well before carefully and thoroughly sealing the plates with the accompanying cap strips. Reactions were vortexed briefly for a homogenously light pink colour before being centrifuged. Thermocycling was performed on the SimpliAmp™ Thermal cycler (Applied Biosystems™) using the protocol described in the IFU. Successful amplification was confirmed using the Tapestation™ 2200 (Agilent Technologies®) in a post-PCR environment. 1 µL of positive control was diluted 1:10 in molecular grade water and run on a high sensitivity D1000 ScreenTape in combination with high sensitivity D1000 sample buffer (Agilent technologies®) according to the manufacturer's instructions.

2.8 Library purification

5 µL of the reactions were then pooled and purified as described in the IFU. Pools were created based on recommendations in the IFU regarding sample type and input as follows. One pool was made containing all reference samples, positive control and NTC. Another pool was made for all direct PCR samples including NTC. Sensitivity samples were pooled based on DNA input amount not exceeding a factor of four. One pool containing the NTC, 15.625 pg, 32.25 pg and 62.5 pg input samples. One pool containing the 125 pg, 250 pg and 500 pg input samples and 2 separate pools for the 1 ng and 10 ng input samples. Three separate pools were made for each type of mixture (M:M, F:F, F:M). An additional pool was made containing the respective single source reference profiles and NTC. A separate pool was made for each type of inhibitor where each pool contained three reference samples and three samples each of the four different concentrations of inhibitor tested (15 total). NTCs were pooled separately. For each of the repeatability and reproducibility experiments a pool was made containing all reactions including NTC.

40 µL of each PCR product pool was diluted in 60 µL of molecular grade water before subsequent purification and size selection using Ampliclean™ (NimaGen® BV) utilizing a 1:1, product:bead ratio according to the IFU. Washing of the beads was performed with freshly prepared 75% (v/v) ethanol (Merck®). The initial elution was performed in 110 µL of TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0 (Invitrogen™)) from which 100 µL was transferred to a fresh tube for the second bead purification. The final elution volume for each library was 40 µL of which 35 µL was transferred for sequencing. The purified libraries were quantified using the Qubit® dsDNA HS Assay kit on the Qubit® 3 fluorometer (Invitrogen™). A qualitative check of the libraries was performed on a high sensitivity D1000 ScreenTape on the Tapestation™ 2200 (Agilent technologies®) according to the manufacturer's instructions. All libraries were stored at 4 °C until sequencing.

2.9 Sequencing

The molarity of each library was calculated based on an average fragment length of 320 bp. Libraries were then diluted to 4 nM in TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0 (Invitrogen™) and pooled based on the number of samples in each library in combination with the desired read depth. Libraries for reference samples, direct PCR, repeatability and reproducibility were calculated for a desired read depth of 30,000 reads per sample. Libraries containing mixtures, inhibitors and those with less than 1 ng DNA input were sequenced to a desired read depth of 300,000 reads per sample.

5 µL of the pooled libraries was then denatured by adding 5 µl of 0.2 N NaOH (Illumina®) and incubated for 5 minutes at room temperature. Following incubation, 10 µL of 200 mM Tris-HCL pH 7.0 (Merck®) was added to hydrolyse the NaOH, followed by 980 µL of ice cold HT1 buffer (Illumina®) resulting in a 20 pM solution. Due to the sample concentration of one library being lower than 4 nM, one run used the adapted denaturing workflow. Libraries were diluted to 2 nM and pooled based on the number of samples and desired read depth. 10 µL of the pooled libraries was then denatured by adding 10 µL of 0.1 N NaOH and incubated for 5 minutes at room temperature. Following incubation, 10 µL of 200 mM Tris-HCL pH 7.0 was added to hydrolyse the NaOH, followed by 970 µL of ice cold HT1 buffer resulting in a 20 pM solution.

240 µL of the denatured 20 pM library was then added to 280 µL of ice cold HT1 buffer and 80 µL of denatured 20 pM PhiX control (Illumina®). The total volume of 600 µL was then loaded into position 17 of a MiSeq® reagent kit v3 for 2 x 300 bp (Illumina®). Sequencing was performed on the MiSeq® system (Illumina®) via the generate fastq module in the local run manager. A samplesheet was loaded specifying the adapter sequences, a cycle count of 2 x 301 and index reads of 10 base pairs.

2.10 Data analysis

The IDseek® OmniSTR™ kit does not come with predetermined data analysis software, instead the user is given full freedom to implement the analysis tools which best suits their needs. Several options already include the OmniSTR™ kit as a preset library including open-source software such as FDSTools [7] (Netherlands Forensic Institute), STRait Razor online [8] and STRait Razor v3 [9] (The University of North Texas Health Science Centre) and commercial software solutions such as MixtureAce™ (NicheVision®).

Due to its size, larger alleles of SE33 cannot be sequenced fully from one side and therefore require merging of the two individual reads into a single long fragment for downstream analysis. The read 1 and 2 fastq files were merged using Fast Length Adjustment of Short reads (FLASh) [10] (Centre for Computational Biology, John Hopkins University). The settings that were used include a minimal overlap of 30 and a maximum overlap of 300 base pairs; maximum mismatch density was set at 0.33. Merged fastq files were then analysed using FDSTools (v2.0.4) using the pipeline setting as “case sample”. This pipeline analyses a single sample with the TSSV, BGpredict, BGmerge and BGcorrect tools. Noise and stutter correction was applied using a model (courtesy of the Netherlands Forensic Institute, Biological Traces Division) trained on approximately 300 samples.

3. Results

3.1 Sample read depth

The libraries for reference samples, direct PCR and sensitivity were assessed on several statistics such as the read distribution between samples, the distribution of reads between the various loci, heterozygous balance within the loci as well as concordance and off-target reads.

The read depth for samples was determined as the sum of all aligned reads for each locus, thus excluding any off-target reads and primer dimers. The variation in read depth was assessed for the library containing reference samples as well as that for direct PCR. The samples in each library were not normalised prior to pooling and purification. The read depth variation for the library with normalised DNA input had relatively low variance, with most samples falling between the average 28,676 ± 35%, excluding outliers. The variance for direct PCR was greater (***Figure 1.***).

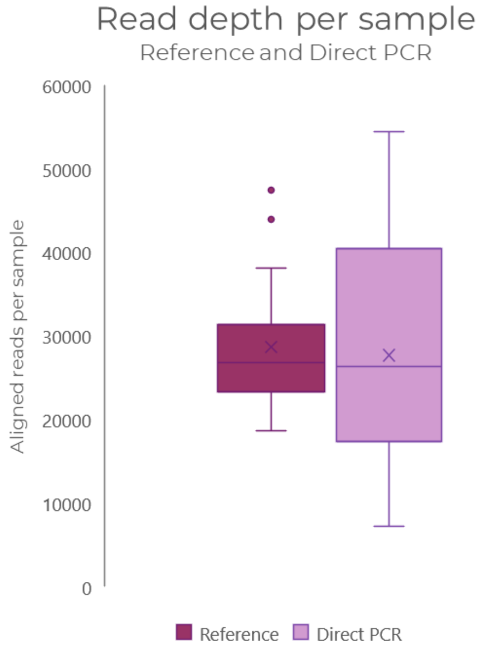

Figure 1. Variation in read depth
Read depth variation in both the reference and direct PCR libraries. PCR output in both libraries was not normalised prior to pooling and purification. PCR input quantity was normalised for the library with reference samples resulting in a more evenly distributed PCR output. PCR output varies to a far greater degree for the Direct PCR samples.

Samples for the sensitivity study were pooled based on the known DNA input ranges not exceeding a factor of 4. The library containing samples in the 500-125 pg range revealed a roughly linear correlation between DNA input and read depth (***Figure 2A.***). On average, the 500 pg samples received approximately 3.93 times more reads than those with 125 pg input. For the library containing samples in the 62.5-0 pg range the read depth difference between the average of the 62.5 pg and 15.625 pg samples was about a factor of 2.56 (***Figure 2B.***).

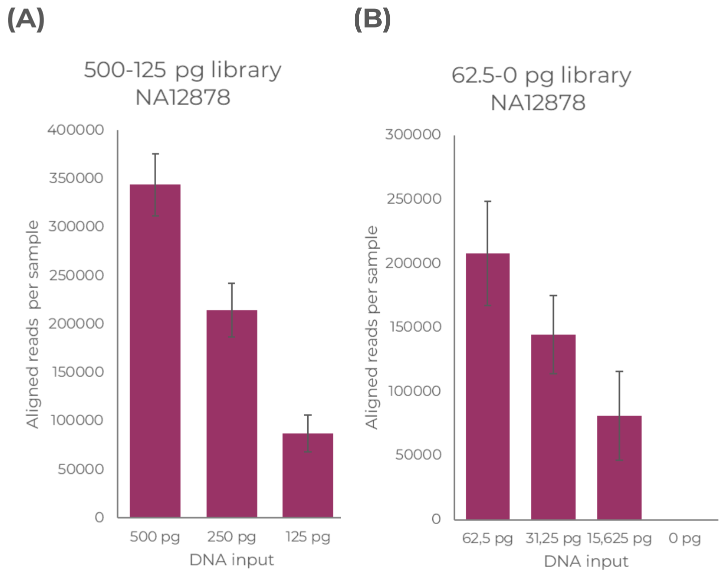

Figure 2. Correlation of read depth to PCR input
The average read depth of samples in the 500-125 pg ***(A)*** *and 62.5-0 pg* ***(B)*** *libraries. No aligned reads could be found for the NTCs. Read depth correlates roughly linear with DNA input.*

3.2 Interlocus and heterozygous balance

In addition to the distribution of reads across different samples, the distribution between the different loci within a sample was also assessed (***Figure 3.***). The read depth was determined as the total number of aligned reads for that locus.

Read distribution between the different loci is uniform at recommended DNA input with the difference in read depth between the highest and lowest marker ranging from approximately 2.4 to 5.9 times as many reads. At reduced DNA input the balance between loci read depth remains stable but gradually increases to an average of factor 7 to 14 between the highest and lowest marker. Samples that were amplified with direct PCR displayed an overall higher spread between the different loci than those with isolated DNA. The difference between the highest and lowest marker ranged from between a factor of 5.3 to 19.2 with a median factor of 9.9. Two outliers of 44.6 and 51.5 were observed for direct PCR, both due to a low read depth of SE33. The most notable differences in read depth for direct PCR could be found at seven loci. Markers SE33, PentaD, D16S539 and D2S1338 all received significantly less reads compared to those on purified DNA, on the other hand, D4S2408, D13S317 and D9S112 all showed a substantial gain.

The heterozygous balance for the reference samples and direct PCR was determined by dividing the read depth of the lowest allele by the read depth of the highest allele, where 1 signifies perfect heterozygosity (***Figure 4.***). With 97.80%, most samples had a heterozygosity balance of 0.6 or higher. 77.76% of all alleles had a balance between 0.8 and 1 regardless of input material.

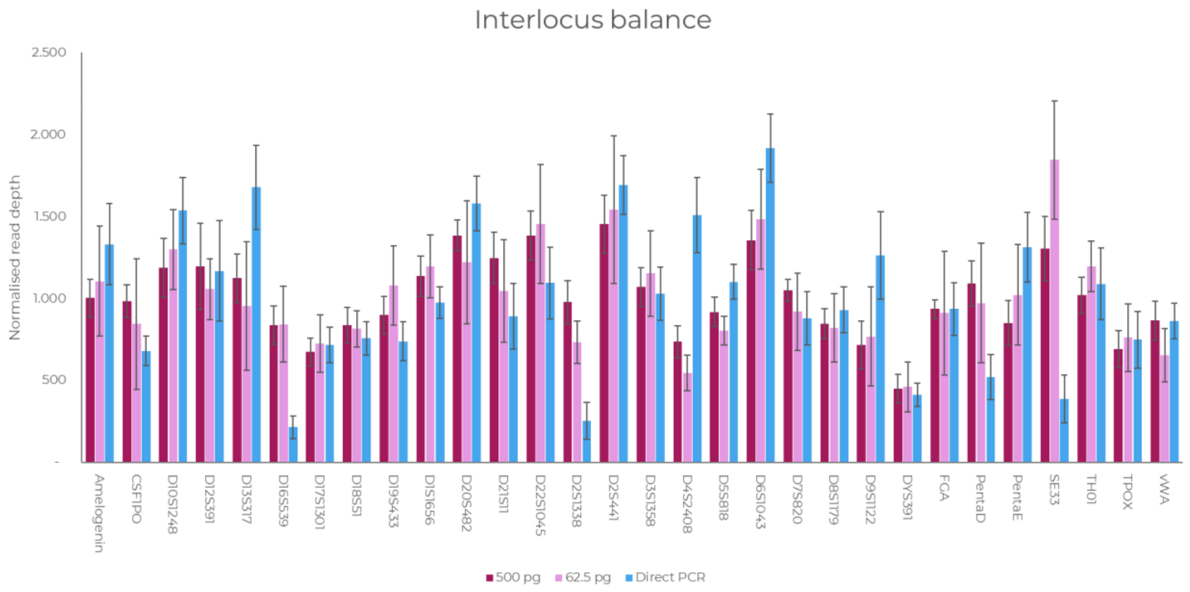

Figure 3. Interlocus balance
The average aligned read distribution between the different loci for samples with 500 pg and 62.5 pg isolated DNA input, as well as direct PCR on a piece of OmniSwab. Reads were normalised to 30.000 reads per sample.

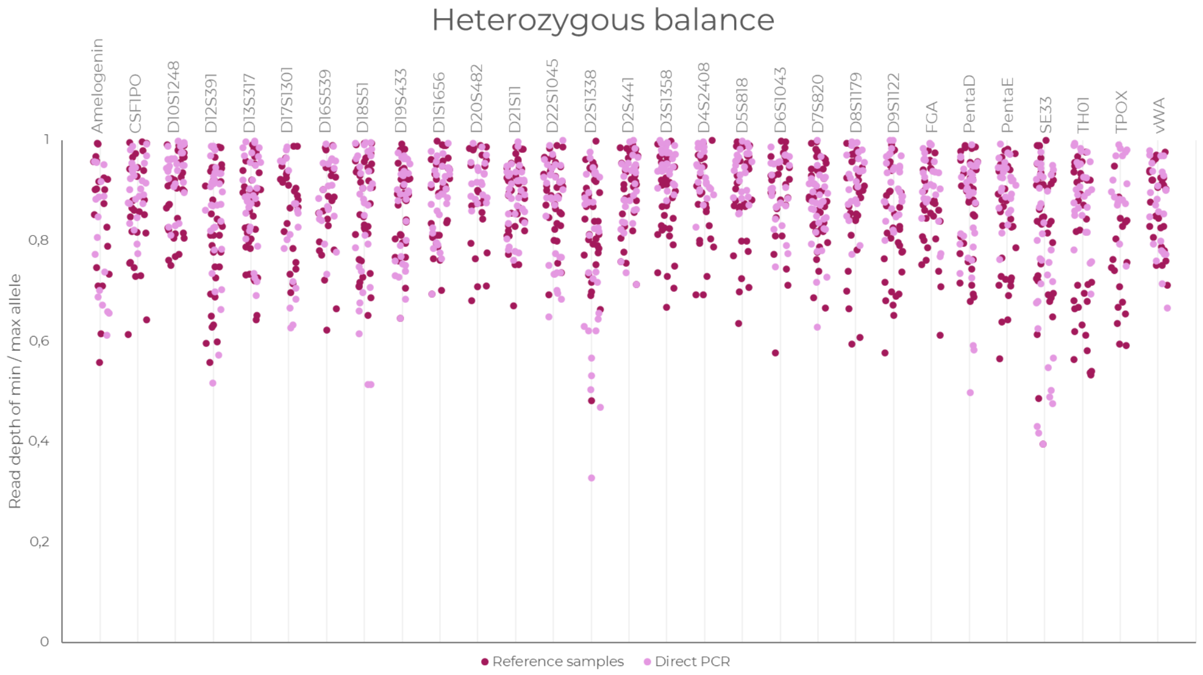

Figure 4. Heterozygous balance
Heterozygous balance was determined by dividing the allele with the least coverage by the allele with the most coverage. 77.76% of the alleles show a value of 0.8 or higher. 97.8% of all samples show a value greater than 0.6.

3.3 On-target percentage

The on-target percentage was determined for samples using both the uncombined fastq files as well as those combined using FLASh (***Figure 5.***). The on-target percentage was determined as the total number of aligned reads compared to the total number of reads in each file. The process of combining fastq files filters primer dimers due the minimal overlap requirement of 30 bp.

The overall on-target percentage for samples with recommended or higher input quantity of DNA (1-10 ng) lies between approximately 85-89%. When DNA input is lowered, the on-target rate drops gradually to an average of 78% at 125 pg. For the lower DNA input range, the percentage on-target reads drops more progressively to approximately an average of 72%, 64% and 47% at 62.5 pg, 31.25 pg and 15.625 pg respectively. When analysing the combined fastq files, the on-target percentage is significantly higher with all samples remaining above 80% on-target reads down to 31.25 pg. At 15.625 pg input the average on-target percentage remains high at approximately 75%.

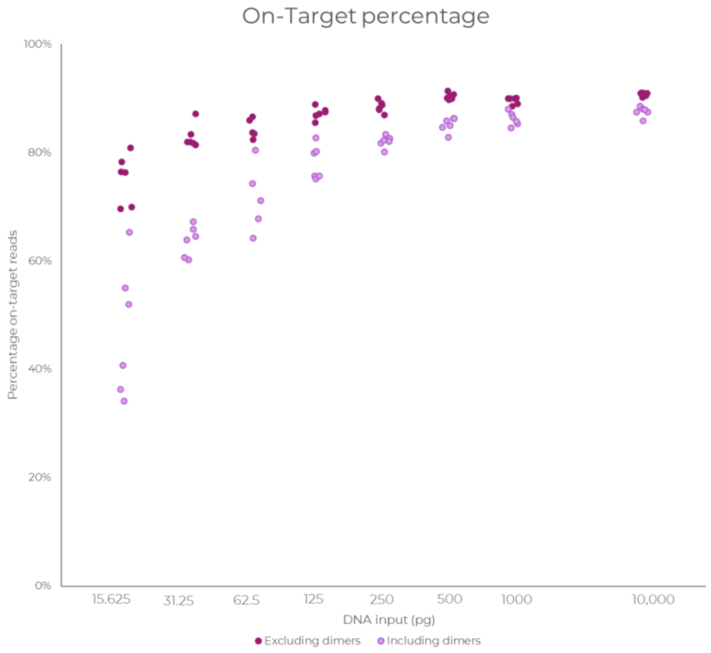

Figure 5. On-Target percentage
The on-target percentage was determined as the total number of aligned reads compared to the total number of reads in each file for both the combined fastq files (with primer dimers filtered) and overall fastq output (including primer dimers). The overall on-target percentage remains high and only gradually decreases at reduced input quantity. Only at the lowest DNA input amounts, the percentage starts to drop more significantly. Filtering fastq for dimers reveals that most off-target reads were produced by increased primer dimer formation which persist after purification and size selection.

The specificity of the assay remains high with minimal off-target amplification, even when very little template is available. The library purification is highly efficient at size selection and dimer removal for the recommended DNA input quantity. Unsurprisingly, primer-dimer formation is increased with reduced DNA input. While the purification process remains efficient, an increase in dimer carry-over is observed with progressively lower DNA input quantities.

3.4 Concordance

Concordance with known CE based alleles was assessed for samples 2800M, NA12877 and NA12878. For 2800M, the found alleles were referenced against those stated by the manufacturer [11] and found to be 100% concordant for the loci for which data was available. Samples NA12877 and NA12878 were referenced against available PowerPlex® Fusion (Promega®) data on 17 loci [12] and found to be 100% concordant.

3.5 Sensitivity

The allele calls across the various DNA quantities were assessed on concordance using allele flagging from the sample-stats tool (see chapter 2.2). The percentage of alleles found to be concordant, discordant or below the analysis threshold was calculated based on the total cumulative number of allele calls (***Figure 6.***).

For DNA input quantities down to 125 pg all samples were found to be 100% concordant. At 62.5 pg of DNA input the first instances of allelic drop-out and drop-in were observed at 1.24% and 0.62% of the alleles respectively. The observed drop-out alleles all fell short to meet the requirements to be flagged as an allele. At 31.25 pg input the number of drop-out alleles increased slightly to 3.06% while drop-ins made up 1.83% of all allele calls. Allelic drop-outs increased more significantly for the lowest DNA input quantity, with 14.07% at 15.625 pg, drop-ins remained stable at 1.83%. All observed drop-in alleles could be attributed to increased stutter.

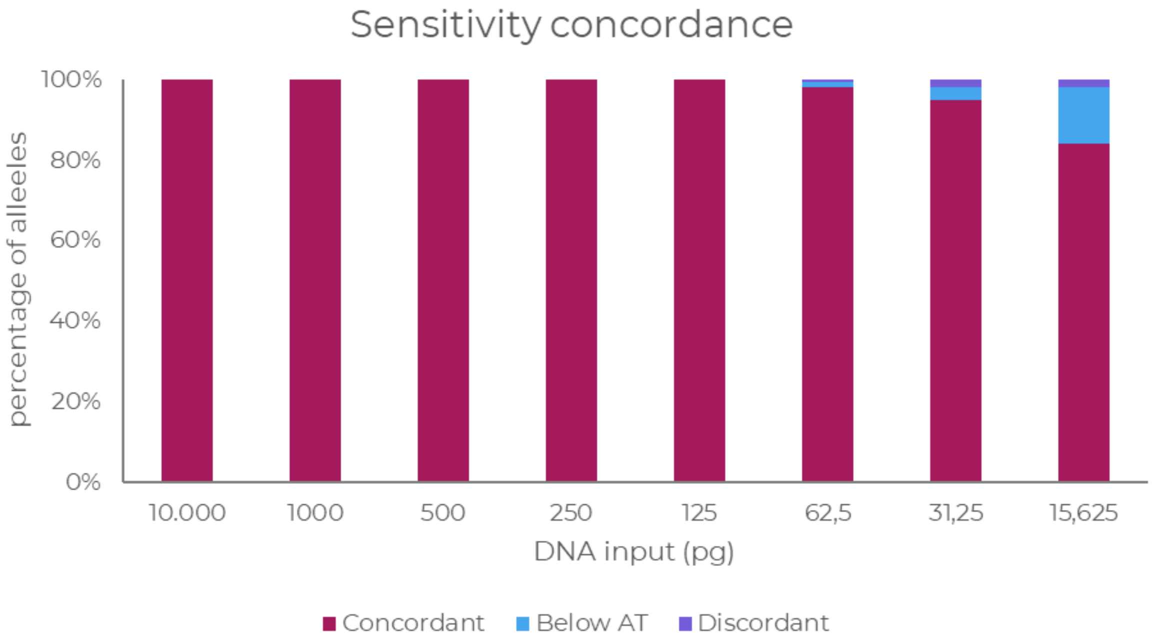

Figure 6. Sensitivity concordance

The percentage of alleles found to be either concordant, discordant or below the analysis threshold of the total number of alleles (cumulative across samples). Down to 125 pg all allele calls were found to be fully concordant. At 62.5 pg the first instances of allelic drop-in and drop-out were observed. At a DNA input of 15.625 pg, 84.1% of alle alleles were still found to be concordant with 14.07% of the alleles dropping out and 1.83% discordant alleles.

3.6 Human specificity

Out of the six species which were tested (see chapter 2.3), the cow managed to produce the fewest (22) reads. Both the dog and cat samples produced substantially more, 7298 and 9221 reads respectively. For all three species 0% of the total reads in each sample could be aligned to any marker. For the sheep and dwarf goat samples a respective 14% and 12% of the total reads in the sample could be aligned to a marker but none met the allelic threshold of 15 reads. Only the pig sample managed to produce enough aligned reads to meet the allelic threshold. 30% of the total reads from the pig samples could be aligned to Amelogenin X. However, the observed variant had a total of 11 sequence differences compared to the human reference gene (***Table 1.***).

Table 1. Human Specificity results

Out of the six species which were tested, only the pig managed to produce enough aligned reads for a variant call. However, the observed variant of Amelogenin X had substantial sequence variation compared to the human reference gene.

Species	Total reads	Aligned	Variants called
Sheep	273	14%	
Dwarf goat	92	12%	
Cow	22	0%	
Pig	113	30%	AmelogeninX: 11296898T>G 11296901T>G 11296909T>C 11296912C>T 11296914CA>AG 11296923C>G 11296927AGTG>- 11296935TGA>CAT 11296942A>T 11296949CT>TC 11296953CA>TG
Dog	7298	0%	
Cat	9221	0%	

3.7 Mixtures

Different types of mixtures were created using Coriell samples (see chapter 2.4). The single source profile of each contributor was assessed. Samples were analysed for allele calls using the visual .html interface. Allele call threshold was reached when meeting one of the following criteria (after applying noise and stutter correction): a minimum of 30 reads, a minimum of 2% of the highest allele per marker and at least 1.5% of the total reads for that marker. The percentage typed was determined as the percentage of non-overlapping alleles belonging to the minor contributor that were called. Drop-in alleles were determined as a percentage of the total number of alleles called (***Figure 7.***).

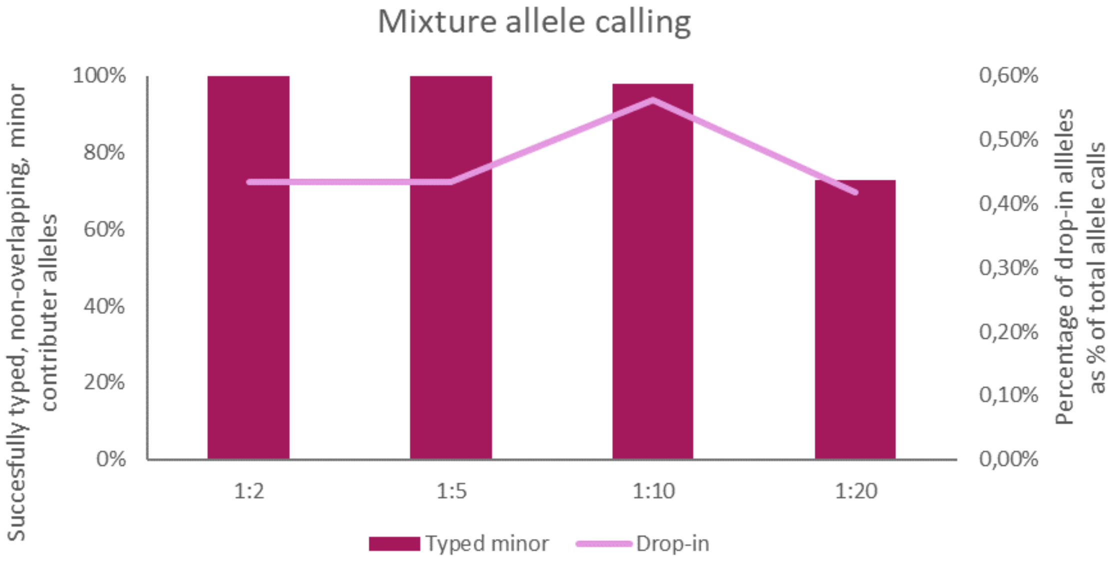

Figure 7. Mixture allele calling

For the 1:2 and 1:5 mixtures, all non-overlapping alleles from the minor contributor were called across all samples with 0.43% of the total cumulative allele calls being drop-ins. For the 1:10 ratio, 97.83% of the unshared minor contributor alleles could still be successfully typed while drop-in increased to 0.56% of the total allele calls. For the 1:20 ratio, 72.87% of the unshared minor contributor alleles could still be successfully typed while drop-ins were responsible for 0.42% of total called alleles.

The unshared alleles of the minor contributor were called for all mixture samples with the 1:2 and 1:5 ratios. For the 1:10 mixtures some drop-outs were observed but 97.83% of the unshared minor contributor alleles were still called. Drop-outs increased for the 1:20 mixtures where about 72.87% of the unshared minor contributor alleles could still be called. Drop-in alleles were observed for all mixtures and comprised 0.43% of the total number of alleles calls for both the 1:2 and 1:5 mixtures. For the 1:10 and 1:20 mixtures, drop-in was responsible for 0.56% and 0.42% of the allele calls respectively.

Drop-out alleles were mostly observed due to the allele of the minor contributor overlapping with the stutter of the major. However, in some instances complete drop-out was observed, either completely lacking any reads or due to not meeting the allelic thresholds. All observed drop-in alleles could be attributed to increased stutter. 83.33% of the Y-markers (Amelogenin-Y and DYS391) could still be detected in the 1:99 mixtures, 33.33% in sufficient quantity to meet the allelic threshold.

3.8 Inhibitor study

PCR inhibition was tested on 2800M using five different types of inhibitors in four different concentrations each (see chapter 2.5). The impact of each inhibitor was assessed by comparing allele calls, read depth per sample, interlocus- and heterozygous balance. No significant deviations in performance were observed for the samples containing bacterial DNA regardless of the amount of inhibitor used (***Figure 8.***). No allelic drop-out or drop-ins were observed for any of the samples (***Figure 9.***). The average read depth per sample was 258,534 (± 34,028) for the control samples and 303,421 (± 10,506) for the samples containing the maximum amount (100 ng) of bacterial DNA tested. Interlocus balance remained consistent with an average of 2.44 (± 0.18) fold difference in read depth between the highest and lowest marker for the control samples, and a 2.29 (± 0.11) fold difference for the samples with 100 ng of bacterial DNA. Samples containing indigo carmine also did not display any significant deviation in performance for any of the tested amounts. The average read depth for control samples was 451,003 (± 29,453) with a 2.53 (± 0.09) fold difference between the highest and lowest markers, compared to 479,083 (± 98,580) and a 2.33 (± 0.06) fold difference for the samples with 133.3 µM indigo.

Samples containing humic acid showed a gradual decline in read depth per sample with increasing concentrations of humic acid. The control samples gave an average of 638,564 (± 57,251) reads per sample while those with 133.3 µM humic acid showed an average read depth of 405,324 (± 75,705). The relative performance of most markers remained stable with the notable exception of SE33, PentaD, PentaE and FGA which all had significantly reduced read depth compared to the control samples. The overall difference in read depth between the highest and lowest marker increased from an average factor of 2.03 for the control samples to an average factor of 5.57 for those with the maximum amount of humic acid tested. No allelic drop-outs were observed in any of the samples.

The performance for samples containing hematin remained stable up until 66.7 µM whereafter the performance significant drops with 133.3 µM hematin samples. The read depth per sample dropped from an average of 478,474 (± 54,395) down to 92,645 (± 17,188). The interlocus and heterozygous balance were also significantly impacted. Additionally, SE33 and PentaD failed to produce reads above the analytic threshold (30 reads) for all alleles in all samples. In one case, FGA also failed to produce enough reads to reach the threshold for both alleles.

Tannic acid was observed to have the most significant impact of all tested inhibitors. In the presence of 16.7 µM tannic acid the average read depth dropped by almost 40% from 898,452 (± 432,845) to 536,638 (± 95,187). Twice that amount resulted in an additional 50% reduction in average read depth and at 66.7 µM the average read depth was reduced to 7% of the control samples. 133.3 µM resulted in complete sample drop-out. At 66.7 µM, the allelic drop-out percentage was 16.97%, with an additional 12.73% called with low read counts (30-100). This occurs for the longer loci such as SE33, PentaD, FGA and D16S539, D7S820 as well as the longer allele of PentaE.

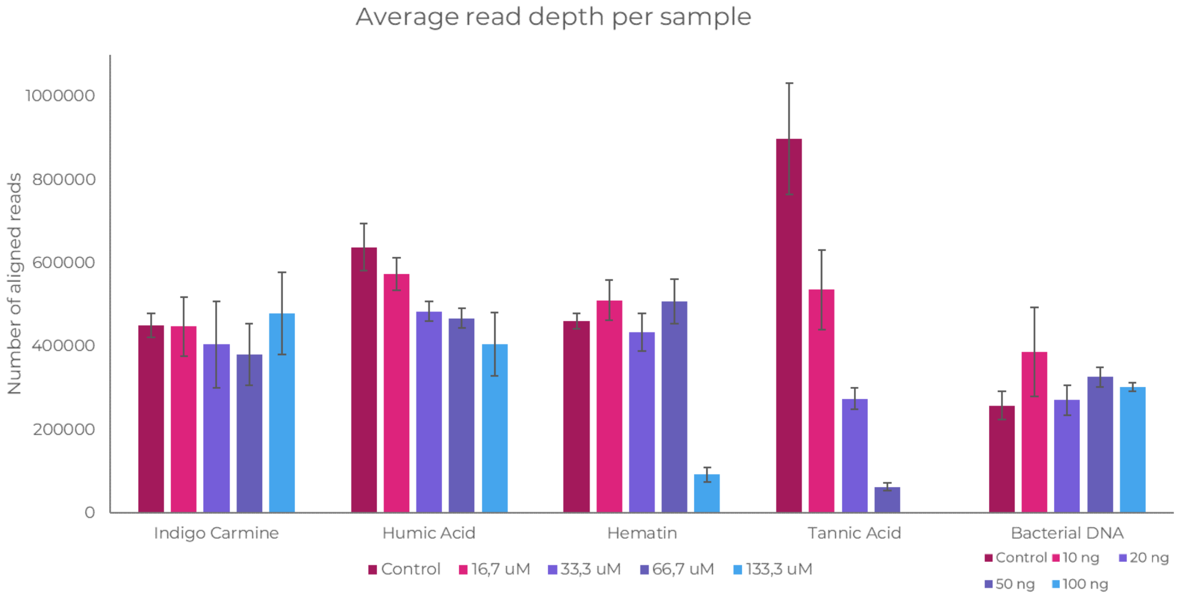

Figure 8. Read depth per sample in inhibitor libraries

The average read depth per sample in the five different inhibitor libraries. Samples were pooled without normalisation to assess the effect of each inhibitor on read depth. Both indigo carmine and bacterial DNA had no visible effect on sample read depth even at the highest concentration inhibitor tested. For humic acid, the average read depth gradually decreased for higher concentrations but still performed well at 133.3 µM. Hematin had no visible effect up to 66.7 µM but had a significant drop in performance at 133.3 µM. Tannic acid displayed the most impact on read depth even at the lowest concentration tested. The highest concentration of tannic acid resulted in complete drop-out.

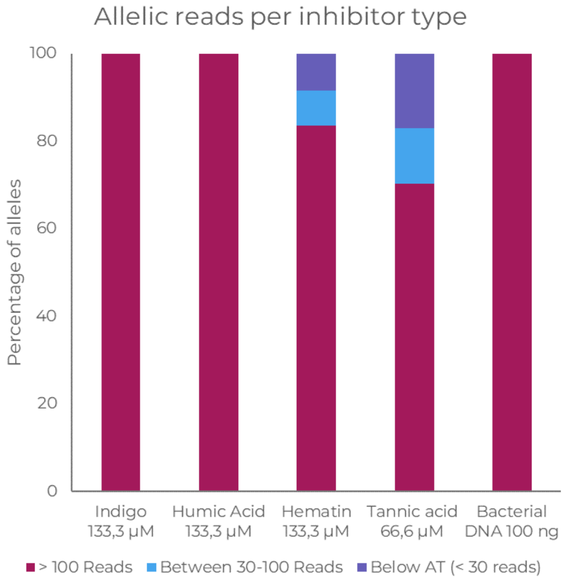

Figure 9. Allele calls per inhibitor

Percentage of (cumulative) allele calls that were called with confidence (>100 reads), with a read count between 30 and 100 reads, and those that could not be called (<30 reads). Indigo, Bacterial DNA and Humic acid displayed no issues for allele calling at the highest concentration inhibitor tested. Hematin displayed an allelic drop-out of 8.48% at 133.3 µM. An additional 7.88% could be called but at reduced read counts. Tannic acid presented the most significant impact to the assay, with a lower allelic read count of 12.73%, and an allelic drop-out of 16.97% for 66.6 µM. Tannic acid gave no results for 133.3 µM and is therefore not displayed.

3.9 Repeatability and reproducibility

Repeatability was assessed by comparing three independent library preparations and sequencing results performed by a single technician. Reproducibility was assessed by comparing the results of three different technicians (see chapter 2.6). The results of the experiments were assessed on concordant allele calls, heterozygous balance, on-target percentage and read depth per sample. Allele calling for each sample was fully concordant across all repeatability and reproducibility experiments. In addition, the heterozygous balance, read depth per sample as well as the on-target percentage fell within similar ranges for all repeatability and reproducibility experiments (***Figure 10.***).

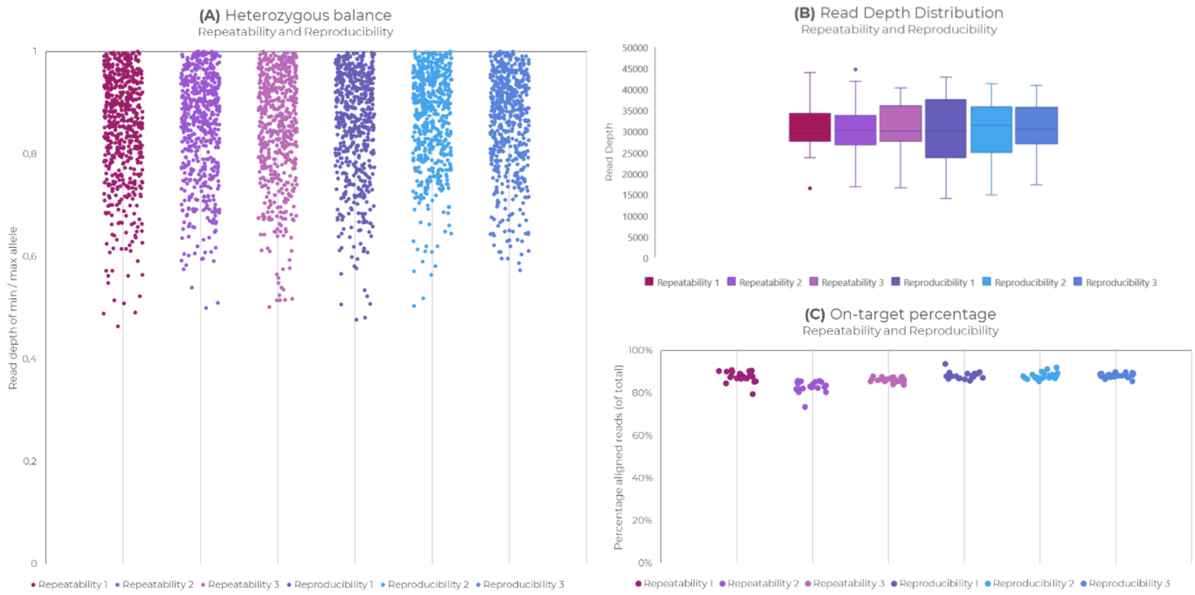

Figure 10. Repeatability and reproducibility results

*The results for the heterozygous balance **(A)**, read depth per sample **(B)** and on-target percentage **(C)** of the various repeatability and reproducibility experiments. In each panel the results from left to right: repeatability 1, repeatability 2, repeatability 3, reproducibility 1, reproducibility 2, reproducibility 3. Heterozygous values for all loci have been compacted for each experiment. All parameters fall within similar ranges indicating highly reproducible results across different library preparations and flow cells.*

4. Discussion

STRs are used in forensic science to establish human identification as their length varies from person to person. STRs are most often analysed using CE, however, MPS provides for more information creating greater discriminatory power. RC-PCR is a novel, more sophisticated PCR based MPS library preparation method combining indexing and PCR amplification in a single closed-tube reaction.

This developmental validation assessed the capabilities of the IDseek® OmniSTR™ kit, a RC-PCR based MPS library preparation kit, for use in forensic investigations. Multiple metrics were tested and assessed including general run parameters (such as on-target percentage, interlocus balance and heterozygous balance), direct PCR, sensitivity, human specificity, DNA mixtures, inhibitor tolerance, repeatability and reproducibility.

Data generated with the OmniSTR™ kit reveals great potential for the routine application of MPS in forensic casework. This RC-PCR based MPS library preparation kit offers reduced risk of sample contamination due to its single closed-tube reaction workflow, while the hands-on time of initial library preparation is similar to setting up a regular PCR. In contrast, other library preparation methods typically require two separate PCR reactions. Performing tagging and amplification of targets during an initial PCR, before enrichment with indices and adapters during the second PCR reaction [13]. When setting up the second PCR, samples are not yet provided with unique indices, and sample (or index) swapping remains a realistic concern. The RC-PCR combines the entire library preparation process within a single reaction, minimising sample handling to a single pipetting step, which is a much simpler and safer solution. Combined with the pre-dispensed index plates it greatly diminishes the potential for contamination.

Following the addition of adapters and indices, other library preparation methods would then require the individual purification of each sample before proceeding to a magnetic bead normalisation step to reduce read depth variation between the various samples [13]. The results of this validation study show that for recommended input quantities of DNA (such as reference samples), samples prepared via RC-PCR can be pooled prior to purification while maintaining an even distribution between samples (***Figure 1.***). This further reduces hands-on time and provides an easy workflow that would be well suited for generating profiles for large numbers of reference samples. Moreover, the IDseek® OmniSTR™ kit allows direct amplification of DNA without the need for extraction (***Figure 1.***), resulting in only minimally increased spread between samples as well as between the various loci. Samples with less than the recommended input amount can still receive more than sufficient coverage when following the pooling and sequencing coverage guidelines regarding DNA input ranges described in the IFU. When pooling samples with unknown and varying degrees of input, sufficient coverage should be taken for the library to compensate for increased spread between samples (***Figure 2.***). These features make the IDseek® OmniSTR™ kit very powerful and reliable when analysing both extracted and non-extracted DNA.

Samples prepared with the OmniSTR™ kit display uniform amplification, with the reads for each sample evenly distributed across loci (***Figure 3.***), with a factor difference of 2.4–5.9x at 1 ng DNA input. The reduced performance of SE33, D16S539 and D2S1338 is likely linked to fragment length as the fragment sizes of these markers (200+ bp) are amongst the highest in the assay. The majority of markers had fragment sizes between 100 to 150 bp with the remaining markers falling in the 150-180 bp range. Understandably, the performance of smaller markers and alleles was observed to be generally higher than those exceeding approximately 175 bp due to the absence of any DNA extraction. In addition, the reads are also heterozygous evenly distributed ***(Figure 4.)*** with 97.80% having a heterozygosity balance of 0.6 or higher. Combined with the high rate of 85-89% on-target reads (***Figure 5.***), the required read depth for each sample can be minimised. This enables efficient use of sequencing capacity by enabling analysis of a large number of samples on a single flow cell, leading to a reduction in sequencing costs.

The assay displays many qualities that are highly desirable in forensic testing. One of these is the high sensitivity of the IDseek® OmniSTR™ kit. Forensic samples do often not comply with the recommended DNA input guidelines. When using lower input quantities, the assay is still highly sensitive with an allelic recovery of over 84% for as little as 15.625 pg DNA input (***Figure 6.***). In addition, animals likely to be encountered in close proximity to humans (sheep, dwarf goat, cow, pig, dog and cat) are unlikely to interfere with the analysis as none of the species tested produced any usable result, expressing the high human specificity of the kit (***Table 1.***).

DNA evidence often consists of mixed DNA from multiple people. The data shows that in the 1:20 mixtures 72.87% of the minor allele could still be found, while in the 1:2 and 1:5 mixtures all non-overlapping alleles could be detected (***Figure 7.***). Interestingly, 83.33% of the Y-markers could still be detected in the 1:99 mixtures, 33.33% in sufficient quantity to meet the allelic threshold. This data support that the IDseek® OmniSTR™ kit is strong in detecting minor alleles. The ability to deconvolute mixture samples based on the full base pair composition is a major advantage of MPS compared to CE, providing greater discriminatory power.

Inhibitor tolerance was assessed for common co-extracted PCR inhibitors, including hematin, humic and tannic acid. In addition, indigo carmine and bacterial DNA were included, as both may be present in forensic samples. Complete profiles could be recovered from most samples and inhibitor concentrations, except for the highest concentrations of hematin and second highest concentration of tannic acid, which yielded only partial profiles. The highest concentration of tannic acid resulted in complete drop-out (***Figure 8.*** and ***Figure 9.***). These results prove that the IDseek® OmniSTR™ kit still has a good performance in the presence of inhibitors, unlike other MPS library preparation methods [15].

Finally, the results obtained with the IDseek® OmniSTR™ kit are highly reproducible. Results from independent technicians, library preparations and flowcells yielded concordant results for various parameters such as allele calling, heterozygous balance, read depth distribution per sample and on-target percentage (***Figure 10.***). This confirms that the IDseek® OmniSTR™ kit is very robust.

This developmental validation administrates the advantages of RC-PCR in combination with MPS over CE analysis for STRs. A global survey, with a vast majority of respondents from the USA, indicated that funding is the clear number one barrier to implementing MPS in their laboratory, followed by under-staffing in second place [14]. The IDseek® OmniSTR™ would tackle both hurdles as a cost efficient, low hands-on solution for MPS in forensics.

In conclusion, the IDseek® OmniSTR™ Global Autosomal STR Profiling Kit provides a multiplex amplicon-based MPS library preparation for sequencing 28 autosomal STR targets, one Y-chromosomal STR and the Amelogenin gene. This RC-PCR based library prep kit contains all reagents to generate Illumina compatible libraries in a simple, sensitive, robust and safe method for cost-effective and high-quality STR analysis and sex determination.

Conflict of interest

All authors are current employees of NimaGen BV and declare no other competing interests.

Acknowledgements

We would like to express our appreciation to the Biological Traces Division of Netherlands Forensic Institute, in particular Dr. Kris van der Gaag, for not only sharing their data on human specificity, but also providing the noise and stutter model used for our analysis with FDSTools.

References

[1] Bukyya JL, Tejasvi MLA, Avinash A, P CH, Talwade P, Afroz MM, Pokala A, Neela PK, Shyamilee TK, Srisha V. DNA Profiling in Forensic Science: A Review. Glob Med Genet. 2021 May 31;8(4):135-143. doi: 10.1055/s-0041-1728689. PMID: 34877570; PMCID: PMC8635824.

[2] Wyner N, Barash M, McNevin D. Forensic Autosomal Short Tandem Repeats and Their Potential Association With Phenotype. Front Genet. 2020 Aug 6;11:884. doi: 10.3389/fgene.2020.00884. PMID: 32849844; PMCID: PMC7425049.

[3] Ballard D, Winkler-Galicki J, Wesoły J. Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. Int J Legal Med. 2020 Jul;134(4):1291-1303. doi: 10.1007/s00414-020-02294-0. Epub 2020 May 25. PMID: 32451905; PMCID: PMC7295846.

[4] Kieser RE, Buś MM, King JL, van der Vliet W, Theelen J, Budowle B. Reverse Complement PCR: A novel one-step PCR system for typing highly degraded DNA for human identification. Forensic Sci Int Genet. 2020 Jan;44:102201. doi: 10.1016/j.fsigen.2019.102201. Epub 2019 Nov 6. PMID: 31786458.

[5] Bus MM, de Jong EA, King JL, van der Vliet W, Theelen J, Budowle B. Reverse complement-PCR, an innovative and effective method for multiplexing forensically relevant single nucleotide polymorphism marker systems. Biotechniques. 2021 Sep;71(3):484-489. doi: 10.2144/btn-2021-0031. Epub 2021 Aug 5. PMID: 34350776.

[6] Instructions For Use IDseek® OmniSTR™ version 1.2. Nimagen.com. 2023.

[7] Hoogenboom J, van der Gaag KJ, de Leeuw RH, Sijen T, de Knijff P, Laros JF. 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. 2017 Mar;27:27-40. doi: 10.1016/j.fsigen.2016.11.007. Epub 2016 Nov 27. PMID: 27914278.

[8] King JL, Woerner AE, Mandape SN, Kapema KB, Moura-Neto RS, Silva R, Budowle B. STRait Razor Online: An enhanced user interface to facilitate interpretation of MPS data. Forensic Sci Int Genet. 2021 May;52:102463. doi: 10.1016/j.fsigen.2021.102463. Epub 2021 Jan 13. PMID: 33493821.

[9] Woerner AE, King JL, Budowle B. Fast STR allele identification with STRait Razor 3.0. Forensic Sci Int Genet. 2017 Sep;30:18-23. doi: 10.1016/j.fsigen.2017.05.008. Epub 2017 Jun 1. PMID: 28605651.

[10] Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011 Nov 1;27(21):2957-63. doi: 10.1093/bioinformatics/btr507. Epub 2011 Sep 7. PMID: 21903629; PMCID: PMC3198573.

[11] James, R. New Control DNA for PowerPlex® Systems. [Internet] 2011. [cited: 2023, July, 17]. Available from: https://nld.promega.com/resources/profiles-in-dna/2011/new-control-dna-for-powerplex-systems/

[12] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019 Aug;37(8):907-915. doi: 10.1038/s41587-019-0201-4. Epub 2019 Aug 2. PMID: 31375807; PMCID: PMC7605509.

[13] Verogen, ForenSeq MainstAY Product Line Reference Guide, document# VD2020050 Rev. C, April 2022

[14] Foley MM, Oldoni F. A global snapshot of current opinions of next-generation sequencing technologies usage in forensics. Forensic Sci Int Genet. 2023 Mar;63:102819. doi: 10.1016/j.fsigen.2022.102819. Epub 2022 Dec 10. PMID: 36509023.

[15] Jäger AC, Alvarez ML, Davis CP, Guzmán E, Han Y, Way L, Walichiewicz P, Silva D, Pham N, Caves G, Bruand J, Schlesinger F, Pond SJK, Varlaro J, Stephens KM, Holt CL. 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. 2017 May;28:52-70. doi: 10.1016/j.fsigen.2017.01.011. Epub 2017 Jan 27. PMID: 28171784.