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

Gaag, K.J. van der

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

Chapter 7

General Discussion

Kristiaan J. van der Gaag

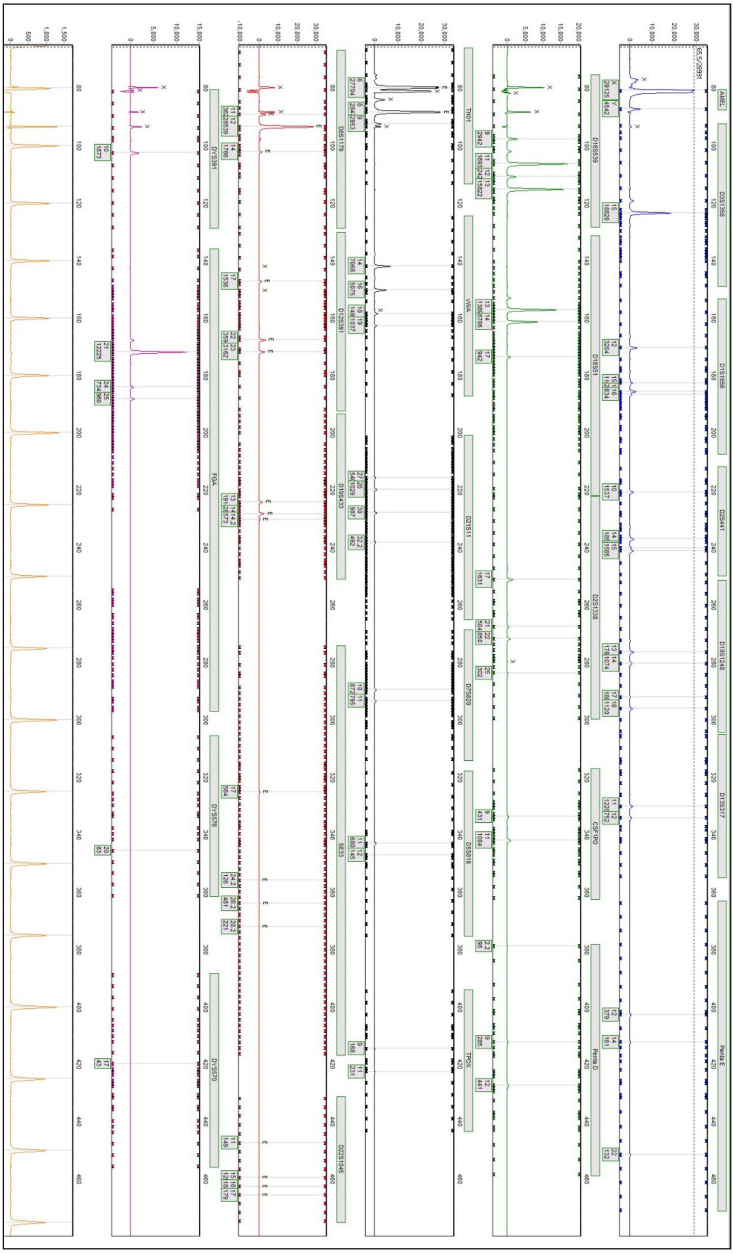
General Discussion

Short tandem repeat (STR) analysis by capillary electrophoresis (CE) has provided important investigative leads and crucial evidence in numerous forensic cases requiring human identification all over the world. While CE analysis of STRs has been the golden standard in forensic DNA evidence for over two decades this method is not without limitations. Parts of these limitations are caused by the analysis method CE and parts by the nature of the chosen DNA marker: STRs. Several of these limitations may be overcome by using Massively Parallel Sequencing (MPS) and the limits of the marker may minimised by developing new software tools for MPS data, or by selecting new markers with sufficient discriminating power.

Strengths and limits of routine CE STR-analysis

CE analysis is an easy operable, fast running and relatively cheap method; all three very good arguments for routine use in a high throughpu¬t workflow such as forensic DNA analysis. A drawback is that CE can only make use of a limited number of fluorescent dye labels (currently 5-6 labels; soon up to eight labels [21]) which means that each label needs to accommodate several STR loci of different fragment sizes to enable the multiplex range of current STR typing systems that is 16 to 27 loci. The commonly used size range of the most recent CE STR assays is around 75-450 bp (enabling 3-7 loci distributed over the range for each label). In forensic traces, DNA is often fragmented which can result in imbalance of the smaller and longer loci of a DNA profile (as shown in *figure 1*). Since the excitation spectra of the used fluorescent labels are partly overlapping and the used detectors in a CE system do not have an unlimited detection range, strong imbalance of loci can lead to either allelic drop-out for the longer loci or off-scale signals for the shorter loci. Off-scale signal can lead to bleed through signal in adjacent colour channels which complexes profile analysis as these artefact signals need to be differentiated from genuine alleles to prevent apparent allelic drop-in signals as shown in *figure 1*.

Figure 1 – Capillary Electrophoresis STR profile of a degraded sample



The figure displays a profile (Powerplex Fusion© 6C) of a severely degraded DNA sample. The signal drops as the length of alleles increases resulting in drop-out of many long alleles while the signal is off-scale for the shortest loci. The signal is so high for the short loci that bleed through signal is observed from the highest peaks to other colors.

Strengths and limits of STRs as marker for human identification

STRs have a high mutation rate. This is the very reason that these loci show a lot of variability between individuals thereby providing a strong discriminating power. The high mutation rate is most likely caused by slippage of polymerase [10] when it encounters stretches of repeated sequence motifs. Unfortunately, the process of generating a DNA profile involves an amplification step by PCR where slipping of the polymerase results in PCR stutter artefacts that can reach intensities of over 20% of the original allele (**chapter 4**). In imbalanced mixtures, minor contributions with alleles at stutter position of the major contributor can be overshadowed by PCR stutters resulting in allelic drop-out of (part of) the minor contributor.

Potential of Massively Parallel Sequencing (MPS) of STRs

Even though STRs have the drawback of stutter formation during amplification, the forensic DNA databases consist almost exclusively of STRs. It therefore makes sense to start the implementation of MPS in forensics by analysis of STRs (**chapter 4**). This can also aid in a stepwise training of experts for court, lawyers and judges to present and explain MPS data in criminal cases.

Although noise from stutter will remain as long as sample preparation for STR analysis contains an amplification step, MPS does provide some advantages over CE.

- Loci are recognised by sequence rather than by length and fluorescent label
- All amplicons can have overlapping sizes (see section 'recognition of loci by sequence rather than length and fluorescent label')
- MPS is not limited by the detection range for fluorescence signal as for CE
- Sequencing of STRs reveals additional variation that is not visible by performing CE analysis
- MPS data is more straightforward (digital) than CE data

Recognition of loci by sequence rather than length and fluorescent label

During MPS data analysis, fragments are recognised by sequence so there is no need to design consecutive size ranges for STR loci. Design of all amplicons in a multiplex in a narrow size window will improve the within-sample locus balance for degraded samples. The only remaining fragment size variation will derive from limits for successful primer design and the variation in repeat length. As long as loci can be separated by sequence, an almost indefinite number of loci can be multiplexed in one reaction. Increased numbers of loci open up new options for answering other forensic questions; this will be discussed later in '*potential new forensic markers*'.

Analysis with an almost indefinite dynamic range

Off-scale signal no longer exists in MPS data so bleed through signal can no longer complicate the interpretation for other loci. A single PCR product yields sufficient material for more than one run on the MiSeq sequencer so even in case of strong locus imbalance, the number of reads per sample can be increased in a rerun with more input so that coverage is sufficient at all loci. In principal, multiplexes do not need to be optimally balanced but for cost effectiveness, a balanced assay is opportune as it allows for a higher number of samples per run. The MPS PCR primers are much cheaper than those used in combination with CE analysis that have a fluorescent label, but sequencing costs are substantially higher than the costs of an electrophoresis run. For a routine application of MPS analysis in forensic casework, cost will be an important factor which can be achieved with a well-balanced assay for which many barcoded samples can be combined in a single sequencing run.

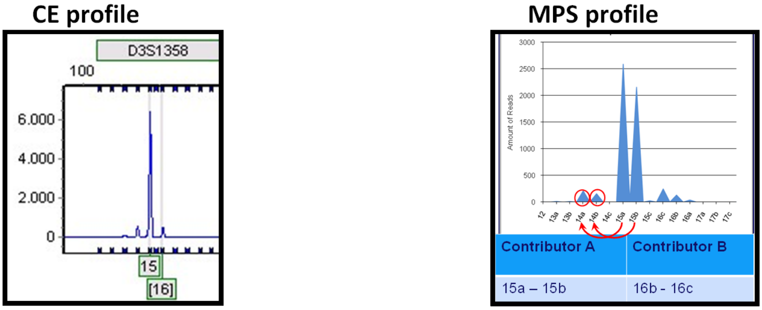
STR sequence variation in addition to length

On the sequence level, many of the STRs exhibit more variation than repeat length only (**chapter 4**). By using a reference database which includes sequence variation, the discriminating power of the same loci increases substantially. The additional information also aids in differentiating genuine alleles from PCR noise such as stutter since overlapping alleles on stutter positions may differ in sequence, which is either due to the repeat structure (in case of complex STRs that exist of more than one repeat motif) or to the presence of SNPs in the repeat or flanking regions. *Figure 2* shows part of a CE and MPS profile of the same two-person mixture with a ratio of 95:5.

As can be observed from the example of *figure 2* the additional discriminatory power gained by sequence variation can be substantial already for a single locus when comparing genotype frequencies, but will be even larger when analysing mixtures.

The sequence variation is likely to also provide additional information for the genetic biogeographic ancestry of a person. Already with CE STR data [1], it is possible to make a prediction of the biogeographic ancestry of a person, but the additional, often less variable, sequence variation will likely increase the accuracy of biogeographic ancestry prediction.

Figure 2 – Illustration of the influence of additional sequence variation for a mixed profile of D3S1358 for CE and MPS



CE profile

MPS profile

Contributor A	Contributor B
15a – 15b	16b - 16c

CE and MPS allele frequencies D3S1358

CE allele	CE allele frequency	MPS allele	MPS allele frequency
15	0,232	**15a**: CE15_TCTA[1]TCTG**[2]**TCTA**[12]**	0,227
		15b: CE15_TCTA[1]TCTG**[3]**TCTA**[11]**	0,005
16	0,226	**16a**: CE16_TCTA[1]TCTG**[2]**TCTA**[13]**	0,147
		16b: CE16_TCTA[1]TCTG**[3]**TCTA**[12]**	0,068
		16c: CE16_TCTA[1]TCTG**[1]**TCTA**[14]**	0,011

CE and MPS genotype frequencies of observed alleles D3S1358

CE genotype	CE genotype frequency	MPS genotype	MPS genotype frequency
15-15	**0,054**	15a-15a	0,0515
		15a-15b	**0,0011**
		15b-15b	0,0000
15-16	**0,0524**	15a-16b	0,0154
		15a-16c	0,0025
		15b-16b	0,0003
		15b-16c	0,0001
16-16	**0,0511**	16b-16b	0,0046
		16b-16c	**0,0007**
		16c-16c	0,0001

Match probability for the major and the minor for CE and MPS

	Major CE	Major MPS	Minor CE	Minor MPS
Possible alleles	15-15	15a-15b	15-16 or 16-16	16b-16c
Genotype frequencies	0,054	0,0011	0,0524+0,0511 = **0,1035**	**0,0007**

On top, the CE and MPS profile are illustrated for a two person mixture (5% minor). For the MPS profile, the stutters (recognised by sequence) are marked in red. The middle two tables display the (Dutch) allele and genotype frequencies for the called alleles. At the bottom, the match probability is calculated for both contributors based on the frequencies of the deconvoluted genotypes for the major and the minor for CE and MPS.

When evaluating the MPS based PowerseqTM assay more insight was obtained in the formation of stutter artefacts as, for complex STRs, the abundance of stutter artefacts of the same length depended largely of the repeat length of the longest uninterrupted stretch. Although additional stutter artefacts (and PCR hybrids which will be discussed later on) might provide a total profile which is more complex. Many of the complexity can now be much better explained than all the piled up artefacts that are visible as a single peak of the same length in a CE profile. This will be further addressed in the software part later on.

STR sequence nomenclature

Alleles generated by CE are named based on their fragment length which is assumed to correspond to a certain repeat length and this repeat length is simply used as the CE allele name. (Massively parallel) sequenced STR alleles will require a new way of describing variation. Currently available databases that describe STR sequence variation have followed the convention of describing the sequence in accordance with length variation observed using CE and using the predominant repeat motif(s) in the variable region as basis. This procedure does not always make sense when the actual sequence is regarded as shown below in two examples of sequences from allele CE12.

D13S317-CE12-TATC[12]AATC[2]ATCT[3]
and
D13S317-CE12-TATC[13]AATC[1]ATCT[3]

The observed sequence variation for this locus as displayed in Sup. Figure 6 of chapter 4 shows that the AATC motif adjacent to the predominant TATC motif is common and variable while it is originally not used in calling the CE allele length. Comparing CE allele names and descriptions of the complete sequence variation may therefore cause confusion. It should be noted that for comparison of CE data and sequence data in a casework setting, the CE allele length is the only relevant detail.

Much insight for sequence description can be obtained from already existing nomenclature used in human genome projects [28]. For clarity, and because the number of loci constantly increases (and is likely to increase further with the use of MPS in forensics), it makes sense to use general and fixed rules for describing sequence variation for the STR motifs as well as in the flanking sequences. In **chapter 3** we suggested a general way to describe sequence variation in detail in accordance with the HGVS recommendation except for a few adjustments that derive from the targeted approach (instead of whole genome) that is common in forensics. We manually applied these rules to a first set of autosomal loci for which data was present at that time (a prototype version of the Powerseq^{TM} assay). These rules should be assessed for a larger number of loci and should preferably be integrated in an automated and freely available software tool to allow a general format for exchanging sequence results.

Because of the international debate on the nomenclature of forensic STR sequencing data, the ISFG prepared a set of recommendations accommodating the points that most groups agreed on in an attempt to harmonise forensic STR sequence variation naming in literature [16]. Notably, the name should include a reference to the CE allele designation to avoid incorrect comparison with data in old cases or with existing CE DNA databases. Also it was recommended to describe all sequences in the forward orientation of the genome reference. This point opposes some existing STR sequence databases (most importantly the NIST STRBase [22]) in which some STRs follow the reverse orientation. Unfortunately, several articles published since then disregard this ISFG recommendation and use the original orientation. Until a general nomenclature consensus is accepted, the ISFG recommendation to include the complete sequence string for exchanging STR sequencing data remains even more important to avoid wrongful comparison of deviant allele calling systems.

Forensic MPS data analysis

A new type of data requires new software to handle the data. For CE data, companies that provided the analysers provided software to translate the electrophoresis data to either an STR profile (describing the number of repeats of each allele plus the observed fluorescence intensity) or a consensus sequence when applying Sanger sequencing. When needed, results could be checked manually. For MPS however, the initial output files were huge fastq files and very limited possibilities for further data processing were provided. Nowadays, the companies provide tools that are capable of doing many of the basic analyses, although for many applications and data interpretation additional third-party tools are still required.

Early users of MPS applications benefit greatly from the help of bioinformaticians. Although some basic knowledge of programming is acquired more easily than most people would expect, data analysis from MPS is too massive and complex for most biologist to handle completely on their own in an efficient way. It is therefore not surprising that a large part of the work in this thesis is done in close collaboration with bioinformaticians and focusses on development of specific software with tools for forensic MPS analysis.

Analysing STRs and investigation of STR stutter

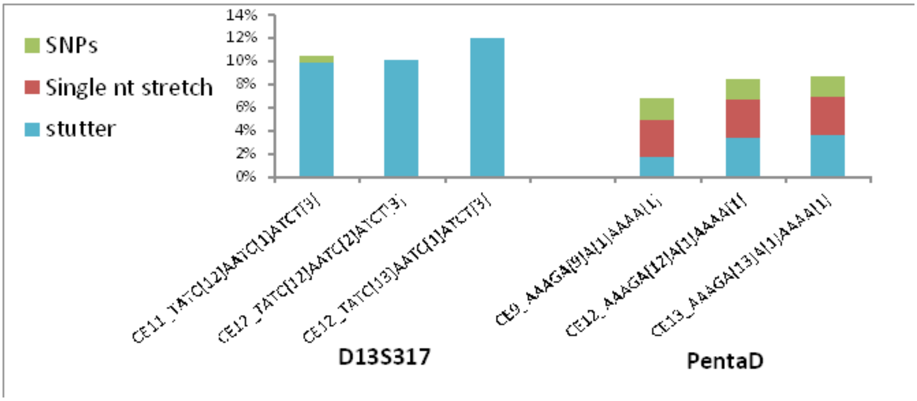
Repeating sequences such as STRs pose a challenge for most alignment algorithms [30]. Since STRs were an important target of interest, we developed the software TSSV using a new approach for analysis of fixed targets containing repeated sequence motifs by only mapping small parts in both the flanking, non-repetitive sequences and reporting all variation between those two flanks. In this way, mapping bias of repeating motifs was avoided. To achieve a readable output, repeated motifs were summarised (**chapter 2**). While TSSV was a good solution for analysing the first experiments, it was still lacking many of the functionalities needed for a forensic casework setting. Then, one needs to apply filters using quality thresholds, differentiate the genuine allele calls from artefacts, and visualise data to explain your sequence profiles in court. The output of such software could subsequently be used to perform statistical calculations on the data for further interpretation of the context and provide information for the evidentiary value/ weight of evidence in a case.

Once the evaluation of the Powerseq assay showed that the majority of observed stutter followed clear patterns in relation to the length of the longest uninterrupted stretch of repeats, a concept was developed to recognise stutter patterns based on a set of training data. This concept was included in the development of FDSTools (**chapter 5**) combined with all the needed data analysis functions intended for implementation in forensic casework. Although this concept could potentially be used for CE as well, it is not likely that it will ever perform as well as for sequencing data since the level of the most abundant stutter depends primarily on the longest uninterrupted repeated motif which varies for alleles of the same CE length.

FDSTools

While the initial concept for reduction of noise in STR sequencing data was only focussed on developing a model for STR stutter correction, one of the implemented approaches for noise correction was able to characterise not only STR stutter, but any systemic allele related noise in the set of training data. An example of this is noise on a SNP positions as shown for D13S317 and PentaD in *Figure 3*.

Figure 3 – Distribution of different types of noise for the three most common alleles of D13S317 and PentaD



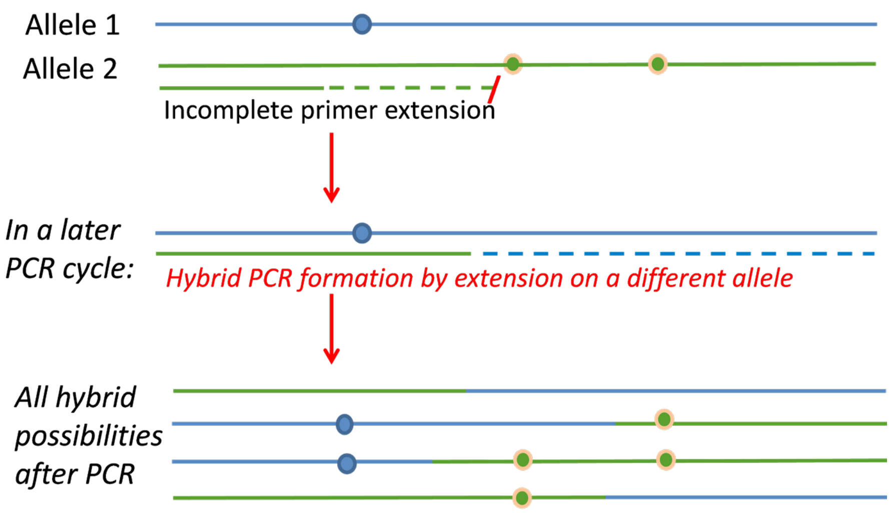
Systemically observed noise for the references in the FDSTools training set that carry the three most frequent alleles for D13S317 and PentaD is shown divided in noise caused by stutter, slippage in a single nucleotide stretch and errors on specific nt positions (noted in the figure as 'SNPs') as a percentage of the reads of the corresponding allele. As can observed, the stutter increases for the longer alleles of both loci and while D13S317 noise consists almost exlusively of stutter, PentaD shows substantial levels of noise from single nucleotide slippage but also of errors on specific nt positions.

In accordance with CE data, we see that alleles with longer uninterrupted stretches of repeats 'loose' more of their original intensity to stutter events and other PCR noise than smaller alleles (also see *Sup. figure7* of **chapter 4**). Once noise sequences can be attributed to the respective genuine alleles based on the knowledge gained in the training data, there is no reason why the noise reads can't be added to their respective genuine allele. In **chapter 5**: *table 2* we show that the within locus balance of single source samples is substantially improved when noise is not only filtered, but also added to the respective alleles. After applying the noise correction to mixtures it was shown that noise was substantially reduced and the performance of analysis of unbalanced mixtures was substantially improved. It was even shown that, after allele related noise correction, stutter is no longer the limiting factor for analysis of unbalanced mixtures. The most abundant noise that is still complicating mixture analysis is now caused by PCR hybrids.

PCR hybrids, the next generation of PCR noise

For MPS, sequence reads are generated for each molecule separately. Therefore, linkage of variants can now be monitored in detail as linked variants occur within one read. While this provides opportunities for the analysis of microhaplotypes (discussed in more detail later on) it also reveals a type of PCR noise that could not be seen by Sanger sequencing. Figure 4 shows an illustration of our theory on PCR hybrid formation (also referred to as jumping PCR). Although PCR hybrids are a new type of noise for the forensic community, they were already described in 1989 [25]) and are well known in the field of metagenomics (also known as PCR chimeras or jumping PCR artefacts) where the hybrids are visible when performing Sanger sequencing after cloning of PCR products.

Figure 4 – Illustration of the formation of PCR hybrids


Allele 1
Allele 2
Incomplete primer extension
In a later PCR cycle:
Hybrid PCR formation by extension on a different allele
All hybrid possibilities after PCR

On top, two 'parent' alleles are displayed that can form PCR hybrids during the PCR. In this example, the primer binds to allele 2 and is only partially extended. In a later cycle (displayed in the middle), the partially extended primer hybridises to allele 1 and is again extended, thereby creating a hybrid PCR product with part of allele 1 and part of allele 2. Depending on the position until where the primer is extended and the fragment / orientation where the extension starts, four different hybrids (displayed on bottom) can be derived from these two alleles.

The formation of PCR hybrids is not allele specific but depends on the combination of alleles in a sample. Hybrids are currently not recognised and corrected by FDSTools (since it corrects systemic allele dependant and not genotype dependant noise). In metagenomics, tools are available that completely filter out alleles that could, by sequence, reside from PCR hybrids [13,14]. However, in the forensic setting, many of the PCR hybrids result in a sequence that also exists as genuine alleles in the population. Therefore, discarding any possible hybrid would result in regular drop-out of genuine alleles in mixed samples. Since the level of hybrid formation seems to be sequence dependant [24], setting a common threshold for hybrid removal would be a suboptimal solution too. Therefore, further investigation of the dynamics of hybrid formation might improve analysis of low level mixture contributions and increase the possible level of automated allele calling for STR sequencing data.

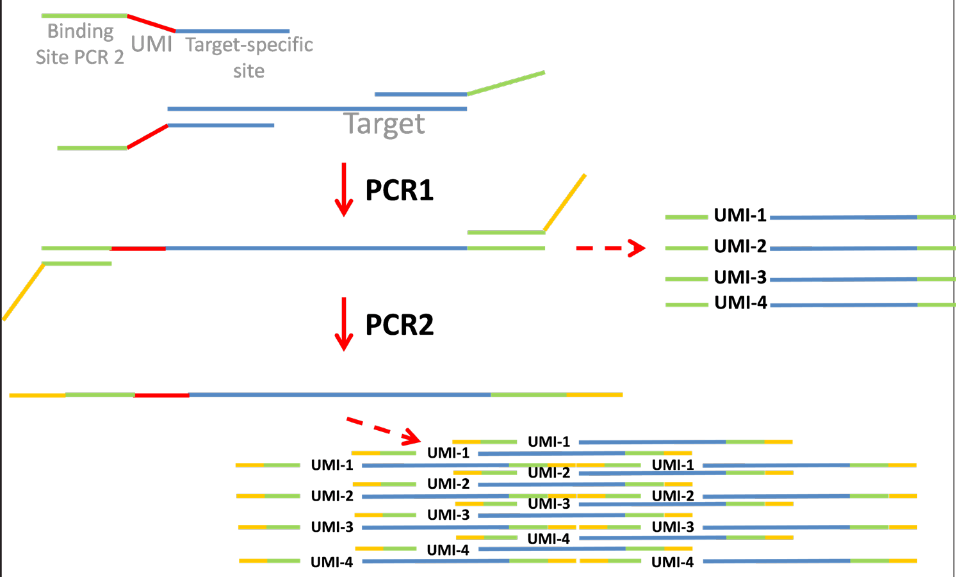
How to further reduce 'MPS' noise

The noise discussed before represents noise created during the PCR. Besides, sequencing errors occur, but this is only a small minority of the noise. Although a lot of the noise can be recognised and corrected during MPS data analysis, ideally noise is prevented to arise at all, which would mean working without PCR. With the minute amounts of cell material in forensic cases, this seems unrealistic for now. Probably it would be possible to reduce the number of PCR cycles, but this decreases inputs in the adapter-ligation step (of the libraryprep) which is shown to increase adapter dimers (as described in **chapter 4**),. Adapter dimers could be prevented by using new library prep methods such as the NEBNext Ultra II FS DNA library prep kit, but this protocol includes an additional amplification step to generate the complete sequencing adapters as required for sequencing.

Random barcodes

Another solution for reducing both, STR stutter and PCR hybrids, is the use of random barcodes (also referred to as unique molecule identifiers: UMIs) [8]. By performing a two-step nested PCR, UMIs can be included in the primers used for the first few cycles of step 1 to be amplified with the target (and remain stable) in step 2 of the PCR. In this way, by counting the number of UMIs for each variant instead of the total number of reads, any noise resulting from the PCR can be reduced to the noise that is generated in the cycles of step 1. The use of UMIs is further explained in figure 5. Although the use of UMIs seems like an excellent solution to reduce noise and authenticate genuine alleles in samples, our first tests (unpublished data) were not very successful. The inclusion of UMIs in the primers for the cycles of step 1 of the PCR resulted in highly increased primer-dimers which were amplified preferentially in step 2 of the PCR, thereby decreasing the sensitivity essential for forensic DNA analysis. So far, published methods using UMIs were all using large amounts of DNA (>20 ng).

Figure 5 – Using random barcodes to trace back PCR noise



On top, the design of a primer from PCR step 1 is displayed including a piece of random nucleotides (UMI) that are integrated in the primer. During the cycles of PCR 1 (middle part) PCR-products are generated that each contain their own unique molecule identifier (UMI). In PCR step 2, the complete molecules are amplified by targeting the tail 5' of the UMI thereby copying the UMI from PCR step 1 without generating new UMIs. During the analysis, results can be summarised by UMI rather than by read to reduce PCR noise to the the cycles performed in PCR1.

Alternative MPS methods

Target enrichment by DNA capture methods

Target enrichment is essential for forensics since most countries have legal restrictions in the features and thus genomic regions that can be analysed without informed consent (as is generally the case with crime scene investigations).

Target enrichment in forensics is commonly achieved by PCR, but an alternative approach would be to use capture methods using probes to fish out the targets of fragmented DNA and get rid of PCR bias before sequencing. However, current MPS methods still require a substantial number of input molecules. For the MiSeq, one sequencing run requires around 3 – 9 fmol (equalling 1,8 – 5,4 * 10^9 molecules) meaning that a substantial number of PCR cycles is needed to achieve sufficient material even if DNA capture methods would be 100% efficient (which they probably are not). Still, it would be worth investigating these methods since new platforms such as single molecule sequencing (also referred to as third generation sequencing) might need less material.

Single molecule sequencing

Single molecule sequencing platforms such as Pacific Biosystems and Nanopore (minion) sequencing are currently mostly known for being able to sequence long DNA fragments. While long DNA fragments are often not available in forensic DNA research, this new type of technology can provide opportunities for forensics once small amounts of material can be used and sequence read quality is increased to a level to a level such as current MPS methods. Unfortunately both of these requirements are not yet achieved for single molecule sequencing (to my knowledge). Still, developments in this field should be closely monitored by the forensic field since analysis without PCR could reduce bias, substantially increase the speed of sample preparation and might facilitate possibilities such as on-site preparation of samples at a crime scene or mobile labs.

Potential 'new' forensic markers

Although MPS increases the evidential value that can be gained by analysing STRs, it also enables the application of new types of markers for forensics. It would be possible to simply sequence high numbers of SNPs until the same discriminating power of STRs is reached (which will require >60 SNPs [37]), but an increased number of loci will substantially increase sequence demand and cost. One type of potential target that could allow forensic analysis using a limited number of loci is microhaplotypes.

Microhaplotypes

Since MPS generates sequence reads separately for each molecule, the linkage of all the observed variation within a read (also referred to as microhaplotype) can be monitored in detail. In **chapter 6** we used genome data from two large genome projects [12,29] to select potential microhaplotypes dispersed over the human genome. While this seemed a straightforward analysis, this project revealed the risk of using data from genome projects (derived from short read data) since they are not without error. Apparently, mapping of multicopy fragments sometimes results in wrongful calling of SNPs in a small region which is exactly the criteria that we used to select potential microhaplotypes. Genome data derived from long reads (the longer the better!) would probably be a much better source for selecting these loci. Unfortunately, this kind of data was not yet available at the time that this research was conducted. A large part of this part of our project focussed on selection of fragments that showed actual variation, and wet lab validation is essential. Surprisingly, many studies still publish new potential loci (including microhaplotypes). It would be beneficial for the scientific community if respected journals would demand wet lab validation for appointing new loci. Analysis of samples from families aids in authenticating genuine SNPs from the inheritance patterns (SNPs that derive from mapping errors will have deviating patterns). Using sample from globally dispersed populations minimise the chance of fixed combinations of common haplotypes and are immediately useful to assess global variation.

After discarding the majority of originally selected loci that did not show the expected variation, a final set of 16 loci remained with a discriminating power comparable to nine STRs. However, in the current format, also the microhaplotypes are not the ideal markers since, as with STRs, PCR hybrids were observed as a complicating factor for interpretation. Like as with STRs, PCR hybrids often result in sequences that also exist as genuine alleles and cannot be simply filtered out. It is surprising that PCR hybrids are still hardly mentioned in forensic MPS publications since they are likely to become the next limiting factor for mixture analysis. Despite the issue of PCR hybrids, microhaplotypes are still potentially interesting loci for forensic DNA analysis since they provide a high discriminating power for a small number of loci and also provide information for prediction of biogeographic ancestry [5].

Increased numbers of loci

Because, during the analysis, different loci are recognised by sequence, the number of loci that can be analysed in one reaction can be increased almost indefinitely for MPS as long as the PCR remains sufficiently sensitive for all the loci. This opened many new possibilities for potential forensic application. Many SNP panels have recently been developed for different purposes:

- Identification
- Prediction of geographic ancestry
- Prediction of external visual characteristics (EVCs)
- Analysis of SNPs on the RNA level

SNP panels for identification

When analysing a large number of SNPs, the same discriminating power can be reached as for the currently used sets of STRs without the burden of STR stutter [37]. Although the current forensic DNA databases consist exclusively of STRs, SNP panels for identification can be used for comparison of known references with forensic evidentiary material and can provide better means for analysing more distant kinships. The only disadvantage of these panels is the bi-allelic nature of most SNPs which is suboptimal for analysis of mixtures containing DNA of more than two contributors. Tri-allelic and tetra-allelic SNPs [18,34] could provide a potential solution to this, but the numbers of available SNPs for these types of markers in the genome are limited. When using panels of large numbers of SNPs, one can choose to use SNPs dispersed over the genome, but for kinship analysis, panels that contain dense coverage of SNPs over a genomic region can provide information about the number of recombinations that occurred which can help distinction between kinships such as cousins or grandparent – grandchildren [38].

Investigative leads

While the most commonly known application of forensic DNA analysis is human identification, DNA can also provide investigative leads. While this application is not restricted to MPS, the possibility of analysing more markers certainly facilititates more options.

SNP panels for geographic biogeographical ancestry

Biogeographical ancestry prediction has recently become a hot item in forensic DNA analysis [19]. When there are no leads available but sufficient perpetrator material is available, biogeographical ancestry information can limit the pool of potential suspects and provide investigative leads. In case of the discovery of unidentifiable human remains or body parts, information on biogeographical ancestry can also be useful to narrow down where to look for a missing person. While SNP analysis on the Y chromosome and mitochondrial DNA are important markers for predicting the biogeographical ancestry in the maternal and paternal linage [31], the use of autosomal markers for this purpose is becoming increasingly important since numbers of persons with admixed biogeographical ancestry are constantly increasing. For the use of biogeographical ancestry prediction in forensic cases it should be noted that the prediction will never be a hundred percent accurate and should be interpreted carefully. Current application in casework is extra complicated by the way of presenting the data which usually consists of principal component analysis (PCA) or structure plots [6,23]. While this way of presenting the data is a neat way of visually presenting the data, the reliability of the prediction is probably interpreted in a less biased way by using likelihood ratios [19] while strong guidance of a trained expert will remain essential for application of biogeographical ancestry prediction in a forensic case.

SNP panels for prediction of external visible characteristics

In addition to biogeographical ancestry prediction, SNP panels have been published in the last decade for prediction of EVCs such as eye colour, hair colour / structure, early onset baldness (alopecia) [15] and skin colour [26]. While the initially published traits such as eye colour and hair colour can be predicted with substantial reliability, EVCs such as alopecia and skin colour often do not reach probabilities over 80% which can easily be misinterpreted by policemen in the field who will use the information for selection of potential suspects. This and the limited number of visible traits that is reliably analysed may be why prediction of EVCs is currently hardly applied in forensic casework. Once prediction of more visual traits is possible and the reliability is increased for several of the traits (i.e. by using a high number of SNPs which is now possible using MPS) prediction of EVCs in forensic cases will probably be more informative.

Analysis of SNPs on the RNA level

Analysis of specific RNA targets has been presented in many publications for identification of the origin of the organ or body fluid of cell material in an evidentiary trace [3]. This can provide important context information for interpretation of the related DNA profile. However, when a sample is mixed, the organ or body fluid type cannot be attributed to the donors by using conventional techniques since the expression level of the different RNA targets varies (except for gender related markers). Thereby the highest RNA signal is not necessarily from the donor with the highest signal in the DNA profile. However, by typing SNPs in body fluid or tissue type related RNA loci, SNPs where the donors carry a different variant can be used to attribute the tissue type or body fluid to a donor. It should be noted however, that the number of SNPs in cell- or tissue type related loci will not be sufficient to reach the same level of discriminating power as is common for routine STR analysis. However, by combining RNA-SNP and STR results, one could simply look for cell type related SNPs of references discriminate the observed donors (deducted from the STR profile) [35].

Quantitative analysis by MPS

In the analysis of mixtures performed in this thesis (**chapter 4 and 5**) it was indicated that the dynamic nature of MPS data (numbers of sequence reads in contrast to rfus) provides a way to perform a quantitative analysis of the contributors in a mixture. While this aids the interpretation of mixtures, it also provides opportunities for other quantitative analyses such as the epigenetic marker methylation which can be measured by bisulphite sequencing [32]. By treating DNA by bisulphite, non-methylated Cytosines (Cs) are converted to Uracil (U) which is translated to a T during PCR and methylated Cs are protected from conversion. By comparing the levels of Cs and Ts the level of methylation can be quantified.

Several studies [32] have recently addressed estimation of age based on MPS analysis of methylation levels of specific CpG sites. Age is a very interesting forensic trait since age is searchable in genealogical databases. It can provide an investigative lead in case of an unknown perpetrator, assist in familial searches (search the family tree where individuals of corresponding age occur) and provide additional information for kinship analyses (i.e. a boy of 18 years old cannot be the grandson of a reference of 25 years old but could be a cousin).

Vidaki et al. [32] recently published a review suggesting numerous potential forensic applications of methylation-based epigenetic studies that we might expect in the near future, such as:

- Tissue identification; if no RNA is present, tissue specific methylation can still be assessed.
- Differentiating monozygotic twins; currently only possible by looking for de novo mutations using deep complete genome sequencing. Analysis of levels of methylation in only a few loci might provide a much cheaper and straightforward method
- Information on alcohol / drug abuse, body size and shape might be of great importance for investigative purposes.

The accuracy of quantitative analyses using MPS can probably be increased by using UMIs (as discussed before), since this will reduce bias introduced during the PCR by preferential amplification.

Combining different forensic loci

The possibility to analyse numerous markers in one analysis provides new opportunities for forensic samples with limited sample material. If a sufficiently balanced assay can be designed, in principle, autosomal, Y and X chromosomal STRs could be analysed at once in combination with SNPs on the autosomes, Y chromosome and mtDNA for identification purposes and for prediction of biogeographical ancestry and prediction of EVCs, all in one reaction. The first commercial assays have been developed already by Promega and Illumina / Verogen [20] that combine different types of loci. However, for many cases it will depend on the research question to be answered if all these loci are actually needed. Only if all loci can be combined in one assay for a price that is not substantially higher than the current routine analysis, a combined analysis will be implemented for routine work. In general, the level of implementation of MPS as routine tool or as a tool for exceptional cases will depend on the cost of the commercially available assays in the near future.

Metagenomics

Metagenomics (analysing not only human but also microbial and any other DNA) is a long existing research field with many applications in the medical field [2] but is currently only applied occasionally in forensics [9]. Once more extended (preferably curated) reference databases become available and specific analysis tools are developed for this application there is an immense potential for this type of analysis. The first studies have already indicated the possibility of attributing a sampling to a donor by analysing skin microbial sequence variation [11] although studies using considerable numbers of references and traces still need to be performed. Since the amount of microbial that we leave behind in a trace can be much higher than the amount of human material (on average, only half of the DNA in saliva is of human origin [27], forensic metagenomics can probably yield a sensitivity beyond the most sensitive forensic method ever available targeting human DNA. In addition, a lot of investigative information can probably be retrieved such as geographic location (based on databases of earth metagenomics).

Ethics related to new MPS data

In most countries, the allowed forensic investigations are bound by law. While the possibilities for e.g. prediction of EVCs is progressing rapidly, it will still take some time before the developed methods can be put into practice. In the Netherlands, additional forensic DNA analyses next to the routine STR profiles are described in detail [4]. For example, since 2007, it is allowed to predict biogeographical ancestry, in 2012, eye colour was added to the list of allowed EVCs and it took until 2017 until hair colour was allowed as well while the methods to predict hair colour were already published 2013 [33]. However, any additional EVCs need to be mentioned separately in the law which is a political process that usually takes several years. STR sequencing is allowed within the current documentation of the Dutch law since no visible traits or diseases can be deducted from the non-coding sequences surrounding the STRs. However, several foreign laws (such as in Belgium) do not allow any expansion of the region that is routinely analysed at the moment unless specifically stated in the law. This might delay the possibility to exchange sequence-based allele information between countries on short term. However, if big successes are achieved in forensic cases using this extra information, it is likely that the political system will pick this up and will work on new legislation. Perhaps this would be a topic that would benefit from European legislation rather than country-based legislation.

Application of forensic MPS tools for non-forensic purposes

The development of Forensic DNA tools largely follows on initially developed techniques for the medical field. In return, tools specifically designed for forensics often find applications in other fields. In the medical field, cell lines are now regularly authenticated using forensic assays [7,36] and FFPE samples with limited cell material can be analysed using forensic assays because of the high sensitivity. Techniques that will be applied for forensic mixture analysis might find an application in Non-invasive prenatal testing (or the other way around) and prediction of biogeographical ancestry and EVCs find a use in ancient DNA research as well as a commercial application for people who are curious to find out more about their roots.

Implementation and accreditation of MPS in a forensic setting

Before implementation of a new method in forensics, detailed validation studies need to be performed, the method needs to be ISO-accredited and experts for court need to be trained to be able to integrate results in a report and explain the data in court. While this is relatively straightforward (although still a tremendous amount of work) for established routine techniques such as CE analysis it is more challenging for a new method such as MPS. Since general scientifically accepted thresholds for allele calling are lacking, detailed testing was required to support the new interpretation guidelines. After writing detailed documentation and validation, one can apply for ISO accreditation and is visited by an expert to judge whether the presented method is fit for accreditation.

At the moment, very few forensic laboratories are using MPS for forensic casework and even less groups do so under accreditation but the number of groups that are investigating the method is increasing rapidly. At the LUMC, accreditation was achieved in September 2015 for using MPS in forensic casework as one of the first laboratories in the world and at the NFI (my current employer) we received accreditation for MPS early 2018 (still as one of the first labs to completely implement MPS for a casework setting). While there isn't any lab that is currently using MPS as the routine method in forensic DNA analysis, this might change in a few years. However, this is unlikely to happen if the costs for sample preparation and/ or sequencing do not decrease further. For application in high profile cases MPS will probably applied regularly in the coming years, especially for analysis of unbalanced mixtures and for analysis of EVCs in order to gain investigative leads to limit the pool of suspects.

References

1. Algee-Hewitt BF, Edge MD, Kim J, Li JZ, Rosenberg NA. Individual Identifiability Predicts Population Identifiability in Forensic Microsatellite Markers. Curr Biol. 2016 Apr 4;26(7):935-42.
2. Amrane S, Raoult D, Lagier JC. Metagenomics, culturomics, and the human gut microbiota. Expert Rev Anti Infect Ther. 2018 May;16(5):373-375.
3. van den Berge M, Bhoelai B, Harteveld J, Matai A, Sijen T. Advancing forensic RNA typing: On non-target secretions, a nasal mucosa marker, a differential co-extraction protocol and the sensitivity of DNA and RNA profiling. Forensic Sci Int Genet. 2016 Jan;20:119-129.
4. Besluit DNA-onderzoek in strafzaken, Nederlandse wetboek van strafvordering

5. Bulbul O, Pakstis AJ, Soundararajan U, Gurkan C, Brissenden JE, Roscoe JM, Evsanaa B, Togtokh A, Paschou P, Grigorenko EL, Gurwitz D, Wootton S, Lagace R, Chang J, Speed WC, Kidd KK. Ancestry inference of 96 population samples using microhaplotypes. Int J Legal Med. 2018 May;132(3):703-711.
6. Byun J, Han Y, Gorlov IP, Busam JA, Seldin MF, Amos CI. Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure. BMC Genomics. 2017 Oct 16;18(1):789.
7. Amanda Capes-Davis George Theodosopoulos Isobel Atkin Hans G. Drexler Arihiro Kohara Roderick A.F. MacLeod John R. Masters Yukio Nakamura Yvonne A. Reid Roger R. Reddel R. Ian Freshney. Check your cultures! A list of cross□contaminated or misidentified cell lines. IJC 2010 July;1-8
8. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. 2011 Jul;39(12):e81.
9. Clarke TH, Gomez A, Singh H, Nelson KE, Brinkac LM. Integrating the microbiome as a resource in the forensics toolkit. Forensic Sci Int Genet. 2017 Sep;30:141-147.
10. Cox R. and Mirkin S.M. Characteristic enrichment of DNA repeats in different genomes. PNAS 94 (10) 5237-5242 (1997)
11. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A. 2010 Apr 6;107(14):6477-81.
12. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014 Aug;46(8):818-25.
13. Gonzalez JM, Zimmermann J, Saiz-Jimenez C. Evaluating putative chimeric sequences from PCR-amplified products. Bioinformatics. 2005 Feb 1;21(3):333-7.
14. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methé B, DeSantis TZ; Human Microbiome Consortium, Petrosino JF, Knight R, Birren BW. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 2011 Mar;21(3):494-504.
15. Liu F, Hamer MA, Heilmann S, Herold C, Moebus S, Hofman A, Uitterlinden AG, Nöthen MM, van Duijn CM, Nijsten TE, Kayser M. Prediction of male-pattern baldness from genotypes. Eur J Hum Genet. 2016 Jun;24(6):895-902.
16. Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmão L, Hares DR, Irwin JA, King JL, Knijff P, Morling N, Prinz M, Schneider PM, Neste CV, Willuweit S, Phillips C. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Sci Int Genet. 2016 May;22:54-63.
17. Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A; SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet. 2007 Dec;1(3-4):273-80.
18. Phillips C, Amigo J, Carracedo Á, Lareu MV. Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data. Forensic Sci Int Genet. 2015 Nov;19:100-106.
19. Phillips C. Forensic genetic analysis of bio-geographical ancestry. Forensic Sci Int Genet.

2015 Sep;18:49-65.

20. Phillips C, Devesse L, Ballard D, van Weert L, de la Puente M, Melis S, Álvarez Iglesias V, Freire-Aradas A, Oldroyd N, Holt C, Syndercombe Court D, Carracedo Á, Lareu MV. Global patterns of STR sequence variation: Sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit. Electrophoresis. 2018 Nov;39(21):2708-2724.
21. Promega. Spectrum CE system brochure (www.promega.com)
22. Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res. 2001 Jan 1;29(1):320-2.
23. N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, M.W. Feldman Genetic structure of human populations Science, 298 (5602 December (20)) (2002), pp. 2381-2385
24. Shin S, Lee TK, Han JM, Park J. Regional effects on chimera formation in 454 pyrosequenced amplicons from a mock community. J Microbiol. 2014 Jul;52(7):566-73.
25. A.R. Shuldiner, A. Nirula, J. Roth. Hybrid DNA artifact from PCR of closely related target sequences. Nucleic Acids Res. 17 (11) (1989) 4409.
26. Spichenok O, Budimlija ZM, Mitchell AA, Jenny A, Kovacevic L, Marjanovic D, Caragine T, Prinz M, Wurmbach E. Prediction of eye and skin color in diverse populations using seven SNPs. Forensic Sci Int Genet. 2011 Nov;5(5):472-8.
27. Sun F, Reichenberger EJ. Saliva as a source of genomic DNA for genetic studies: review of current methods and applications. Oral Health Dent Manag. 2014 Jun;13(2):217-22.
28. Taschner, P.E., den, Dunnen, J.T. Describing structural changes by extending HGVS sequence variation nomenclature. Hum. Mutat. 2011; 32: 507–511
29. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526: 68–74
30. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011 Nov 29;13(1):36-46.
31. Underhill PA, Kivisild T. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet. 2007;41:539-64.\
32. Vidaki A, Kayser M. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. Genome Biol. 2017 Dec 21;18(1):238.
33. Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz A, Branicki W, Kayser M. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet. 2013 Jan;7(1):98-115.
34. Westen AA, Matai AS, Laros JF, Meiland HC, Jasper M, de Leeuw WJ, de Knijff P, Sijen T. Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples. Forensic Sci Int Genet. 2009 Sep;3(4):233-41.
35. Yousefi S, Abbassi-Daloii T, Kraaijenbrink T, Vermaat M, Mei H, van 't Hof P, van Iterson M, Zhernakova DV, Claringbould A, Franke L, 't Hart LM, Slieker RC, van der Heijden A, de Knijff P; BIOS consortium, 't Hoen PAC. A SNP panel for identification of DNA and RNA specimens. BMC Genomics. 2018 Jan 25;19(1):90.
36. Yu M, Selvaraj SK, Liang-Chu MM, Aghajani S, Busse M, Yuan J, Lee G, Peale F, Klijn C, Bourgon R, Kaminker JS, Neve RM. A resource for cell line authentication, annotation and quality control. Nature. 2015 Apr 16;520(7547):307-11.
37. Zhang S, Bian Y, Chen A, Zheng H, Gao Y, Hou Y, Li C. Developmental validation of a custom panel including 273 SNPs for forensic application using Ion Torrent PGM. Forensic Sci Int Genet. 2017 Mar;27:50-57

38. Fang R, Pakstis AJ, Hyland F, Wang D, Shewale J, Kidd JR, Kidd KK, Furtado MR. Multiplexed SNP detection panels for human identification. Forensic sup. series. 2009 Dec; vol2; 538-539