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

Gaag, K.J. van der

Summary thesis

This thesis focusses on the development, application and validation of new forensic methods bases on recently developed techniques referred to as Massively Parallel Sequencing (MPS, also known as Next Generation Sequencing).

In **chapter 1** the background of the currently used methods is discussed and the basics and challenges of MPS methods are discussed.

The development and application of forensic DNA research has evolved rapidly since the discovery of hypervariable DNA 'fingerprints' by Jeffreys et al. (1985) leading to the powerful tool that is currently referred to as genetic 'human identification'.

Over the past two decades, investigation of Short Tandem Repeat (STR) markers has played a major role in human identification. STRs are pieces of DNA that contain a repeated sequence of 2-6 nucleotides. The length of this repeated sequence can vary between people and by analysing a number of these STRs, a practically unique DNA profile can be generated. An example of a DNA sequence containing an STR is shown below.

Figure 1, DNA sequence containing an AGAT-repeat

ATCCGA AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT TGGACCAG

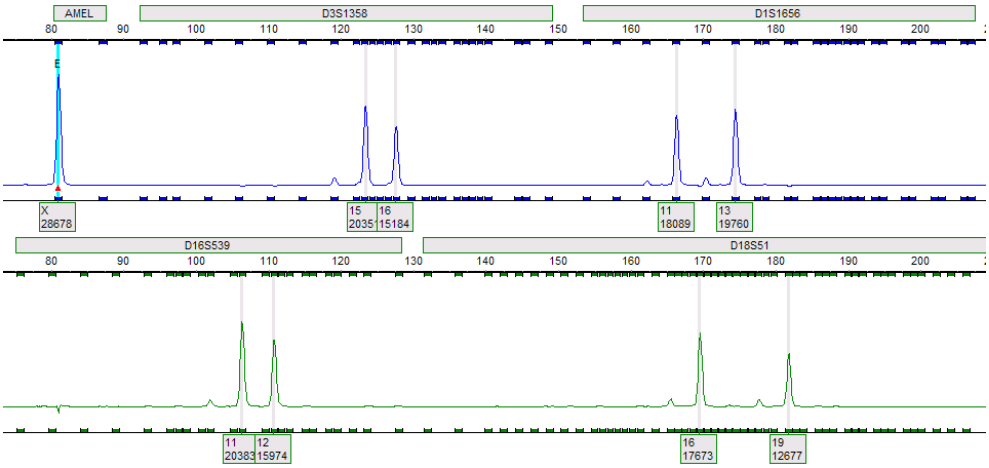
Conventional analysis of DNA profiles

Conventionally, STRs are analysed by the technique 'capillary electrophoresis' (CE) which basically means that fixed fragments containing the STRs are amplified and tagged with a (fluorescent) label and separated by length. By comparing the obtained fragment lengths to a known ladder, the number of STR repeats is deduced for each marker, resulting in a DNA profile.

While CE is a relatively easy and straightforward technique, it is not without limitations. During the amplification of STRs, so-called stutter artefacts emerge that can complicate the interpretation of a DNA profile. In addition, analysing sufficient STRs in one reaction to obtain a 'unique' DNA profile results in sub-optimal conditions for samples where DNA is degraded as is often the case for forensic samples. Figure 2 shows an example of (part of) a conventional DNA profile.

Epilogue

Figure 2, example of a DNA profile



Part of a DNA profile from a single person. The peaks are shown for five loci separated by length (in the same colour) and by fluorescent label.

New technologies: Massively Parallel Sequencing

New DNA analysis techniques have recently been developed that are capable of analysing millions of DNA molecules in parallel in a highly automated fashion. While these 'Massively Parallel Sequencing' (MPS) techniques are already being implemented in the medical and other molecular analysis fields, MPS is still relatively new to forensic DNA analysis.

In principle, MPS could overcome some of the limitations of CE analysis. However, some of these limitations are caused rather by the nature of the STR marker than by the technique that is used to analyse it. Since the type, and especially the amount, of data for MPS are very different from conventional DNA analysis techniques, development of specialised forensic software to properly handle this data is crucial for implementation of MPS in actual casework.

In this thesis we explore the applications of MPS for human forensic DNA analysis, not only focussing on the lab-related practical work, but also on the development of specialised forensic software for MPS data analysis. In addition, we survey alternative DNA markers to STRs with the goal to further expand the application of DNA analysis in forensic cases in the near future.

Massively Parallel Sequencing vs Capillary Electrophoresis

The expected potential of MPS in forensic DNA analysis relies on the following differences between CE and MPS.

- *While CE only analyses the fragment length MPS determines the exact DNA sequence*. Sequence variation (in addition to length only) increases the statistical power of each locus. In addition, sequence variation can help to differentiate stutter artefacts from genuine alleles.
- *The detection range of CE is limited while the MPS detection range is almost unlimited*. For CE, too much signal results in artefacts in a DNA profile while very low signal cannot be separated from noise. For MPS, the number of sequence reads for a sample can be increased almost indefinitely without interfering with other markers (although not without cost)
- *CE is limited in multiplex capacity while MPS can potentially analyse thousands of markers at once*. Since currently, no more than six fluorescent labels can be used, multiplexing for CE is achieved by designing the amplification in fragments of different lengths. For MPS, all markers can be designed in the same fragment range since markers are distinguished by sequence during the analysis rather than by length.

Massively Parallel Sequencing of Short Tandem Repeats

Most of the work discussed in this thesis focusses on MPS analysis of STRs. Since forensic DNA databases consist of STR data, it makes sense to start by analysing this marker using MPS so the resulting data can still be used to perform searches in the DNA databases.

Basic analysing of MPS STR data: TSSV

While STRs are the common markers for forensic DNA research they are not used very often in other fields of DNA analysis. As a result, initial software packages for MPS data analysis were performing poorly when it comes to handling repeating sequences. In **chapter 2** of this study, a new (open source) software tool, TSSV, was designed as one of the first data analysis packages that focusses on these markers. By using two anchor sequences (on either side of the STR), markers are recognised and the variation observed between these anchor sequences is summarised by counting the reads for each variant and abbreviating repeated sequences. This software also turned about to be useful in assessing sequence errors in MPS data.

Sequencing data requires sequencing nomenclature

Alleles from CE are described as allele numbers that represent the number of repeats in the allele. Since sequencing reveals additional variation on the sequence level, a new nomenclature is required that allows for a more detailed description of the allele sequence composition. In **chapter 3**, we described the first recommendations for forensic STR sequencing nomenclature attempting to reveal all relevant sequence information for reconstructing the underlying sequencing while aiming for a description as short and readable as possible.

Validating a (prototype) commercial assay for sequencing of STRs

In collaboration with the company Promega©, a prototype version of a commercial assay, designed for sequencing of STRs, was tested. With this assay, 17 STRs and Amelogenin (a sex typing marker) were amplified in one reaction and, after further preparation of the amplified product, sequenced using the Miseq™ system. **Chapter 4** describes a detailed assessment of the performance of the assay and the observed variation in each locus.

To calculate how 'unique' a DNA profile is, the frequencies of each sequence variant in the population must be determined. When these frequencies are known, a statistical calculation can be used to estimate how likely it is by chance for a person to have that exact profile. 297 samples from a European, Asian and African population were sequenced in order to assess the additional variation that is obtained compared to CE and to get an idea of the allele frequencies in these three populations. For most of the tested markers, a substantial increase of alleles was observed on the sequence level compared to the CE length alleles in the same individuals.

When analysing STRs, the amplification reaction needed to visualise the variation, results in so called stutter artefacts caused by slippage of the enzyme Polymerase. STR stutter is a complicating factor for interpretation of imbalanced mixtures since peaks of the minor contributor to the mixture cannot always be differentiated from stutter. The additional information gained by sequencing also provides inside in the occurrence of stutter. It was determined that the proportion of stutter relative to the corresponding allele is mainly determined by the length of the longest interrupted stretch of a repeat. This information presented new opportunities for interpreting mixtures when using the exact sequence of each allele in the sample.

Correcting stutter in a case sample using bioinformatics

In **chapter 5**, the bioinformatics software FDSTools is described as the first software to actually correct PCR- and sequencing artefacts in MPS data. Based on a set of training data it characterises systemic noise. This information can subsequently be used to correct the noise in an unknown sample, thereby substantially reducing the baseline for analysis.

The software also contains tools to perform quality control on the samples used in the reference data and to determine analysis thresholds after performing correction. Using these thresholds data of unknown case samples can be filtered and alleles passing the determined thresholds are called and visualised in an interactive format.

It was shown that alleles of the minor contribution in unbalanced mixtures were recovered after performing correction while they could not be differentiated from stutter peaks before performing correction.

Alternative forensic markers to STRs

While bioinformatics tools can improve analysis for STRs, it remains clear that STRs are not the ideal forensic markers for all purposes. While the high variability of STRs helps to obtain a practically unique profile from a limited number of markers, stutter artefacts and allelic length variation can increase the level of complexity of interpreting a DNA profile. In **chapter 6**, hypervariable microhaplotypes are selected from publically available genome data that contain four or more SNPs within a fragment of 70 bp. These markers can almost reach the same variability as STRs without the disadvantage of stutter artefacts and variation in length.

The study revealed that the vast majority of fragments containing this number of SNPs within a small sequence span resided from erroneous reported variation in the publically available genomes. However, a subset of 16 microhaplotypes was successfully selected and validated in the lab with a discriminating power that roughly resembles 9 STRs. In addition to the high discriminating power, data from these microhaplotypes could also be used to separate the samples from the tested European, Asian and African population suggesting that they also provide ancestry informative information.

Conclusions and future perspectives

In **chapter 7**, the overall performed studies and future perspectives are discussed. MPS is now ready to be applied in forensic casework (and is already being applied in some cases). While MPS is currently still a relatively expensive method, it is not unlikely that this might change in the near future and CE could be gradually replaced by MPS once it is used in a more automated high throughput fashion. MPS seems a promising

method for the analysis of mixtures, either by sequencing STRs or other markers, such as microhaplotypes. There are many other applications for MPS that are not (or only limited) possible using CE, mostly because the number of markers that can be analysed simultaneously is much higher for MPS than for CE. Currently, ancestry prediction and prediction of a few externally visible characteristics are already feasible, but for many other characteristics more basic knowledge needs to be acquired before the analysis can be applied in casework. Many studies are currently ongoing to acquire the basic knowledge for predicting more phenotypic characteristics.