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

Gaag, K.J. van der

Chapter 1

Introduction / Outline

Kristiaan J. van der Gaag

Introduction / Outline

Since the discovery of hypervariable DNA 'fingerprints' by Jeffreys et al. [12] in 1985, development and application of forensic DNA research has evolved rapidly. Over the past two decades, investigation of Short Tandem Repeat (STR) loci has played a major role in human identification. The implementation of short tandem repeat (STR) analysis by capillary electrophoresis in forensic casework [15] and the establishments of large STR databases [30] have provided crucial investigative leads or evidence in numerous cases.

Short Tandem Repeats and Capillary Electrophoresis

STRs are DNA loci containing repeated sequence motifs of 2-6 bp in length. An important feature of STRs is the high mutation rate caused by polymerase slippage [5] which results in a high degree of variability between individuals. By analysing a relatively small number of STRs, DNA profiles can be obtained with very high discriminatory power. Unfortunately, STR slippage also occurs in the PCR amplification step resulting in PCR artefacts known as stutter (discussed in more detail later in this thesis).

Routine analysis of STRs is a relatively straightforward process performed by size separation of fluorescently labelled PCR fragments using Capillary Electrophoresis (CE) [18]. Multiplexing of sufficient loci in one reaction for CE is achieved by labelling loci with several fluorescent labels and PCR design of different loci for separated fragment length ranges.

In addition to autosomal STR analysis, targets such as mitochondrial DNA [20], Y chromosomal markers [16], biogeographical informative markers [23] or tissue informative RNA markers [3] have also been investigated. However, until recently, application of DNA analyses in casework was performed almost exclusively by capillary electrophoresis (CE) based methods such as fragment length analysis, Sanger Sequencing [20] and SNaPshot single base extension [27].

Molecular genetics developments (non-forensic)

In the meantime, a major transition is taking place in the molecular and medical genetic field where routine analysis by Sanger sequencing is being replaced by Massively Parallel Sequencing (MPS), also known as Next Generation Sequencing (targeted or whole exome / genome) [4]. This transition is initiated either to reduce cost or to expand possibilities. The price per sequence for MPS is much lower than for Sanger when sufficient targets / samples are analysed simultaneously. Expansion of possibilities is achieved by analyses of more / larger genomic regions, or by a more quantitative analysis (e.g. bisulphite sequencing [28] and low-level mixture analysis used for NIPT

[31]).

Introduction of a new technique in the forensic field commonly happens in a later stage than developments in the medical field since methods need to be optimised for minute amounts of (often degraded) DNA. In addition, software needs to be optimised to handle mixed DNA samples and answer specific forensically relevant questions.

Why use MPS in forensics?

MPS has some features which make these techniques interesting for application in forensics.

- A sequence is generated separately for each DNA molecule, instead of a consensus sequence as is generated by Sanger sequencing.
- MPS can generate millions of sequence reads (hence, the name 'massively').
- Many different targets can be analysed simultaneously without the need of separate sequence reactions, thereby expanding multiplex possibilities, even allowing complete genome sequencing
- MPS data can be quantified by simply counting reads for every sequence variant, thereby creating 'discrete data' in contrast to CE where interpretation of the shape of peaks is an entity that is difficult to define using straightforward parameters.
- The number of reads for a single sample can be increased almost indefinitely, resulting in an unprecedented dynamic detection range.
- Small and overlapping fragment sizes can be used for all loci since, during the analysis, targets are recognised by sequence rather than by fragment size / label as for CE which will benefit the analysis of degraded DNA samples.

Several commercial companies have developed an MPS platform [28], which illustrates the large current and future market expected for these techniques, for instance in the field of medical research. The early versions of MPS platforms (2005 – 2010) were undergoing constant improvements in sequence data quality and read output but since 2011 several platforms (such as MiSeq and Ion Torrent) started to focus on a more diagnostic application where data quality is relatively stable [11,31], also opening opportunities for the forensic field.

Not every platform is suited for use in forensic research. The specifics are further discussed below in the text box 1 '**MPS platforms with forensic potential**'

MPS platforms with forensic potential

Several Massively Parallel Sequencing platforms are (or have been) available. Since forensic DNA analysis is different from that in medical genetic research (where mostly SNP genotyping is used) the forensic context has specific demands for MPS platforms. Current forensic reference databases exist of STR-profiles and the ability to sequence short tandem repeat is a prerequisite for the forensic community. To span the entire repeat region for the majority of STRs, read lengths of at least 200 nt are required which limits the choice to a few platforms. Current proven high quality MPS platforms with the ability to sequence such read lengths are the 454 pyrosequencers (454 / Roche, discontinued in 2017), Ion Torrent semiconductor sequencers (Ion Torrent / Life Technologies), Solexa / Illumina sequencers and Pacific Biosystems / Roche SMRT sequencers.

454 and Ion Torrent are both based on one-by-one addition of unlabeled nucleotides [REF Rothberg, Merriman]. After incorporation of a nucleotide, the number of incorporated nucleotides is detected via an enzymatic cascade (Pyrosequencing) or by direct measurement of a pH-influx (semiconductor sequencing). Illumina sequencers incorporate labeled blocked nucleotides and enzymatically remove the label / block to incorporate one nucleotide during each sequencing cycle [7]. Pacific Biosystems provides the only platform (at the time of this study) that performs single-molecule sequencing reaching read-lengths of many kbs [24]. This platform is not used for the projects used in this thesis but possible applications will be discussed in the general discussion. From the tested platforms in this thesis, the Illumina Miseq provided the most suitable and high-quality method (similar to the results shown by Salipante et al. [26], therefore, most data discussed in this thesis is generated using this system.

MPS data and data analysis

Since 2004, the read-output and sequence read length increased steadily (with marked increases around 2007-2008 due to introduction and further expansion of the capacity of the Illumina HiSeq systems) which resulted in a notable decrease of cost per sequence. In the same time, more computer power is required to analyse the data. Gordon Moore predicted in 1965 that the speed of computer chips doubles each year while retaining the same costs. Figure 1 shows that MPS costs per megabase dropped substantially faster than the costs for increasing computation power. As a result of this increase of data, new approaches and tools needed to be developed to analyse the data fast, efficiently and accurately.

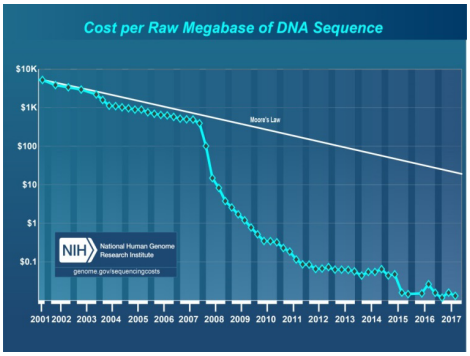

Figure 1 – cost per raw megabase of DNA sequence in combination with Moore's Law
This graph (source: National Human Genome Research Institute) displays the decrease of cost / Mbase of DNA sequence since 2001 in combination with Moore's law (development in increased computer power). It is apparent that the sequencing costs dropped substantially faster than the cost for the computer power required to perform the analysis.

Data analysis in forensics

An important aspect for forensics is robustness of the data (also known as sequencing quality, see text box 2 '**Sequencing quality**'). The choice of a robust platform, that is no longer undergoing continuous improvements, only prevents part of the challenges as low-level sequencing errors are inevitable and need to be recognised and filtered using appropriate data analysis tools. Thus, the development of software for analysis of forensic MPS data is highly important [10,19] and will be a crucial part of the work described and discussed in this thesis.

Sequencing quality

Sceptics of MPS application in diagnostics or forensics are eager to bring up the issue of sequencing errors 2]. Sequence errors occur and could compromise the interpretation of MPS data. However, setting appropriate quality thresholds that filter reads containing sequence errors will solve the issue, alike done for any other technique, such as the filtering of baseline noise and stutter peaks in CE. In general, the quality of a sequence read is highest at the beginning of the read and signal noise slowly decreases as the read progresses as can be seen from the figure below [6].

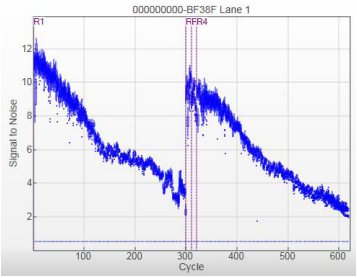

Figure 2 – Signal to noise ratio of an Illumina MiSeq® run.
The graph shows the signal to noise ratio over all data points in an Illumina MiSeq® run for nucleotide incorporation (1nt / cycle). As can be observed, the signal quality decreases as the read progresses for both, read 1 (cycle 1-300) and read 2 (cycle 321-620).

To compensate for a reducing quality in a progressing read, Paired End sequencing can be applied which refers to sequencing a molecule from both the 5'and 3' end; the low quality base calls at the end of one strand will be the high quality base calls in the beginning of the read of the opposite strand.

For every base call the signal-to-noise ratio is translated to a quality score defining the likelihood of erroneous base calling. Quality thresholds during MPS data analysis usually involve coverage (number of reads covering a specific position), base quality and mapping quality (when matching reads to a reference). When paired end sequencing is applied, the length of the overlapping part between the reads in both orientations can vary if the amplicon length exceeds the read lengths which can impact sequence quality in the non-overlapped part.

Public genome databases

The decrease in sequencing cost per nucleotide resulted first; in the (affordable) possibility to generate an (almost) complete individual genome sequence and later; by impressive collaboration efforts, to the generation of databases of complete genomes [9, 8,33]. One of these projects, the 1000 genomes project, provides a public database of variation observed in individuals of globally dispersed populations and is a powerful tool for the selection of new (globally) informative forensic markers. Since different genome projects are started for different goals and budgets for these projects are not unlimited, there is a trade-off between quality in terms of per base sequence coverage and the number of individuals included in the project. While publicly available databases can drastically decrease the amount of wet lab work that is needed to reliably select new markers, it should be taken into account that genome databases are not free of errors.

Microhaplotypes

While STRs and SNPs are well-known loci in molecular genetics, the use of microhaplotypes [14,17] is not common although they are potential forensic loci without some of the disadvantages of STRs (discussed in more detail in the discussion). Microhaplotypes are fragments that contain more than one SNP within a span of ≤200 bp. The combinations of the different alleles of the SNPs form multiple haplotypes resulting in a higher number of alleles than the SNPs separately. In particular, the fragments that contain more than two SNPs within a short sequence span can be informative for forensic identification purposes, but might also serve as markers for prediction of biogeographic origin. *These loci are discussed in more detail later on in this thesis.*

Goal of this thesis

The research described in this thesis aimed to convert the power of MPS to the forensic field. We developed a complete forensic research tool kit for human identification by DNA analysis based on MPS. We focused on all the aspects that are essential for development and optimisation of MPS assays and analysis tools with the final goal of implementation in a forensic setup. MPS enables many more possibilities besides identification, for example in providing investigative leads by analysis of human, but also non-human DNA [1]. Several of these upcoming methods, targets, technologies and applications are briefly mentioned in the General Discussion to provide a more complete picture of the expected impact of MPS on the forensic field in the near future.

Outline

The obvious application of MPS in forensics would be the sequencing of STRs as forensic DNA databases are filled with data from this marker type and MPS can have added value for the profiling of severely degraded DNA (all amplicons can be small-sized) or discrimination of mixed samples (by adding sequence variants). However, when this project started, tools for analysis of MPS STR data were absent since MPS analysis outside the forensic field focussed almost exclusively on analysis of SNPs. In **Chapter 2: 'TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes'**, the development of the software TSSV is described, which represents one of the first tools for analysis of MPS STR data. Since mapping of STRs to a reference is complex and error-prone due to the repetitive nature of these loci, we chose not to map the actual repeat region but instead map short parts of both flanking regions (usually covering part of the primer binding sites). Any variation observed in the sequence in between is reported as strings without comparison to a reference thereby avoiding any mapping bias that may occur depending on the resemblance of sequences to the reference. Repeated elements were abbreviated as a first step towards a universal nomenclature.

Since CE allele calling for STRs does not provide information about the exact sequence of an allele, nomenclature needed to be developed for describing the additional variation typed by sequencing. **In Chapter 3: 'Forensic nomenclature for short tandem repeats updated for sequencing'**; we provided the first recommendations for a nomenclature system for forensic STR sequencing data. These recommendations were largely incorporated in the ISFG recommendations for STR sequencing nomenclature which were published later that year [22].

After collaborating with Promega® in the optimisation of an STR sequencing assay for the commercial market, we performed a detailed in-house validation of a prototype version of the Powerseq™ assay in preparation of ISO17025 accreditation. The study of the performance of this assay included analysis of STR stutters, sequence variation in three distinct globally dispersed populations and analysis of mixtures is described in **Chapter 4: 'Massively parallel sequencing of short tandem repeats - Population data and mixture analysis results for the PowerSeq™ system'.**

In the study described in chapter 3 it was apparent that, not surprisingly, STR stutter remained the limiting factor for analysis of mixtures. In **Chapter 5: '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.'** we describe a new software we developed namely FDSTools. While the software TSSV described in

Chapter 2 was able to catalogue the variation observed in a sample, FDSTools can use a set of training data to characterise structural PCR and sequencing noise including STR stutter and use this information to correct noise in case samples. In this way, noise can be reduced substantially to facilitate analysis of much lower contributions in mixtures. This software was designed for application in forensic casework and includes many features for visualisation, validation and quality filtering of all types of MPS data.

Although correction of noise improved the analysis of mixtures it can be debated whether STRs are the ideal loci for the purpose of analysing complex mixtures since the remaining levels of stutter after correction will still complicate analysis of highly in unbalanced mixtures. In **Chapter 6: 'Short hypervariable microhaplotypes: A novel set of very short high discriminating power loci without stutter artefacts.'** we describe an alternative set of loci for the analysis of mixtures namely microhaplotypes which are small fragments containing several SNPs. Although these loci cannot be used for comparison with the established databases (unless it is decided to type these loci routinely), they can still be used for comparison with known references in a case. The concluding **Chapter 7** discusses the potential of MPS in comparison with the currently used method CE and the future steps needed to use the full potential of this new technique. Software development is discussed including additional options to improve MPS data analysis in order to deal with remaining PCR noise. Novel targets and applications are presented that could provide new possibilities for forensic investigations concluded by suggestions for implementation of MPS in a casework setting.

References

1. Arenas M, Pereira F, Oliveira M, Pinto N, Lopes AM, Gomes V, Carracedo A, Amorim A. Forensic genetics and genomics: Much more than just a human affair. PLoS Genet. 2017 Sep 21;13(9):e1006960.
2. Bandelt HJ, Salas A. Current next generation sequencing technology may not meet forensic standards. Forensic Sci Int Genet. 2012 Jan;6(1):143-5.
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. Biesecker LG, Biesecker BB. An approach to pediatric exome and genome sequencing. Curr Opin Pediatr. 2014 Dec;26(6):639-45.
5. Cox R. and Mirkin S.M. Characteristic enrichment of DNA repeats in different genomes. PNAS 94 (10) 5237-5242 (1997)
6. Endrullat C, Glökler J, Franke P, Frohme M. Standardization and quality management in next-generation sequencing. Appl Transl Genom. 2016 Jul 1;10:2-9.
7. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV. The challenges of sequencing by synthesis. Nat Biotechnol. 2009 Nov;27(11):1013-23.

8. 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.
9. Hayden EC. Technology: The $1,000 genome. Nature. 2014 Mar 20;507(7492):294-5.
10. 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.
11. 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.
12. Jeffreys, A.J., Wilson, V., Thein, S.L. Hypervariable ' minisatellite ' regions in human DNA. Nature 314, 67-73 (1985)
13. Jeffreys, A.J., Wilson, V., Thein, S.L. Individual-specific ' fingerprints' of human DNA. Nature 316, 76-79 (1985)
14. Jin, L., Underhill, P.A., Doctor, V., Davis, R.W., Shen, P., Cavalli-Sforza, L.L., Oefner, P.J. Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. Proc. Natl. Acad. Sci. U. S. A. 1999;96:3796–3800.
15. Jobling M.A., Gill P., Encoded evidence: DNAin forensic analysis. Nature Reviews 5: 739-751 (2004)
16. Kayser M. Forensic use of Y-chromosome DNA: a general overview. Hum Genet. 2017 May;136(5):621-635.
17. Kidd, K.K., Pakstis, A.J., Speed, W.C., Lagacé, R., Chang, J., Wootton, S., Haigh, E., Kidd, J.R. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. Forensic Sci. Int. Genet. 2014;12:215–224.
18. Kimpton C.P., Gill P., Walton A., Urquhart A., Millican E.S., Adams M. Automated DNA Profiling Employing Multiplex Amplification of Short Tandem Repeat Loci. Gen. Research 1993. 3:13-22
19. de Knijff P. From next generation sequencing to now generation sequencing in forensics. Forensic Sci Int Genet. 2019 Jan;38:175-180.
20. Melton T., Holland C., Holland M., Forensic Mitochondrial DNA: Current Practice and Future Potential. Forensic Science Review 24(2):110
21. Merriman B; Ion Torrent R&D Team, Rothberg JM. Progress in ion torrent semiconductor chip based sequencing. Electrophoresis. 2012 Dec;33(23):3397-417.
22. 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.
23. Phillips C, Santos C, Fondevila M, Carracedo Á, Lareu MV. Inference of Ancestry in Forensic Analysis I: Autosomal Ancestry-Informative Marker Sets.Methods Mol Biol. 2016;1420:233-53.
24. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):278-89.
25. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. Nat

Biotechnol. 2008 Oct;26(10):1117-24. doi: 10.1038/nbt1485.
26. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, Harkins TT, Cookson BT, Hoffman NG. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Appl Environ Microbiol. 2014 Dec;80(24):7583-91.
27. Sanchez JJ, Børsting C, Hallenberg C, Buchard A, Hernandez A, Morling N. Multiplex PCR and minisequencing of SNPs--a model with 35 Y chromosome SNPs. Forensic Sci Int. 2003 Oct 14;137(1):74-84.
28. Vidaki A, Kayser M. Recent progress, methods and perspectives in forensic epigenetics. Forensic Sci Int Genet. 2018 Nov;37:180-195.
29. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009 Apr;55(4):641-58.
30. Werret D.J. The National DNA Database. Forensic Science International vol88:33-42 (1997)
31. Willems P.J., Dierickx H., Vandenakker E., Bekedam D., Segers N., Deboulle K., Vereecken A. The first 3,000 Non-Invasive Prenatal Tests (NIPT) with the Harmony test in Belgium and the Netherlands. Facts Views Vis Obgyn. 2014;6(1):7-12.
32. 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.
33. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526: 68–74