Universiteit
Leiden
The Netherlands

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

Gaag, K.J. van der

**Citation**
Gaag, K. J. van der. (2019, November 27). *Development of forensic genomics research toolkits by the use of Massively Parallel Sequencing*. Retrieved from https://hdl.handle.net/1887/80956

| | |
|---|---|
| Version: | Publisher's Version |
| License: | Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden |
| Downloaded from: | https://hdl.handle.net/1887/80956 |

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page

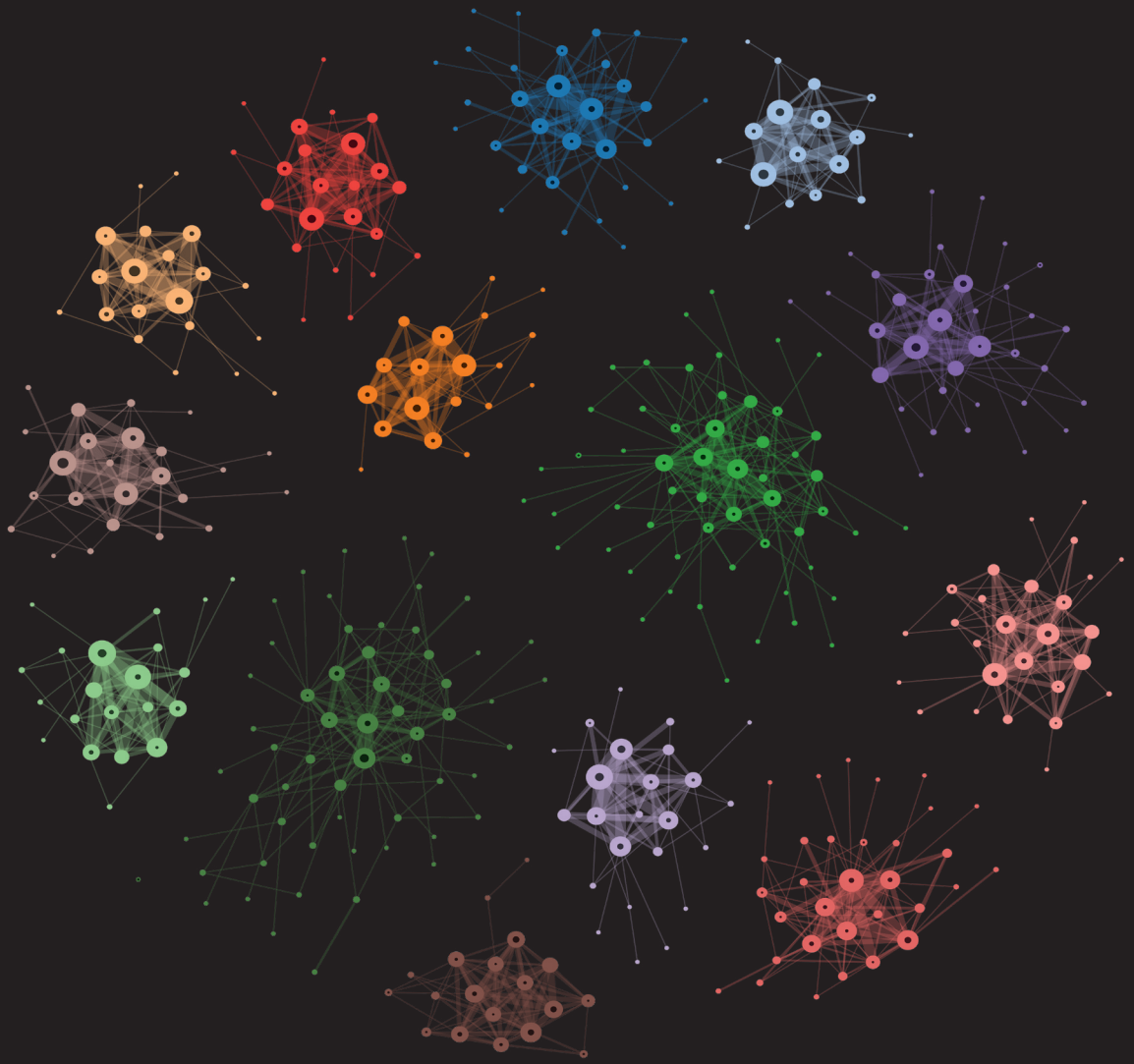
Universiteit Leiden

The handle http://hdl.handle.net/1887/80956 holds various files of this Leiden University dissertation.

**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

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

Kristiaan van der Gaag



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

Kristiaan van der Gaag

## Colophon

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

The research described in this thesis was financially supported by the Leiden University Medical Center and by a grant from the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands

| | |
|---|---|
| **Author** | Kristiaan J. van der Gaag |
| **ISBN / EAN:** | 978-94-6332-585-1 |
| **Cover** | FDSTools visualisation (samplevis) of sequencing variation from 14 STR loci in a set of 300 dutch inidividuals |
| **Back cover** | FDSTools samplestats table |
| **Layout & cover design** | Kristiaan van der Gaag and Jerry Hoogenboom |


© K.J. van der Gaag, 2019
All rights reserved. No part of this thesis may be reproduced in any form or by any means without prior written consent of the author.

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

**Proefschrift**

Ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 27 november 2019
klokke 11:15 uur

door

**Kristiaan Johannes van der Gaag**

geboren te Vlaardingen
in 1980

# Promotiecommissie

| | |
|---|---|
| Promotores | Prof. dr. P. de Knijff |
| | Prof. dr. J.T. den Dunnen |
| Copromotor | Dr. T. Sijen – *Nederlands Forensisch Instituut (NFI)* |
| Leden promotiecomissie | Prof. dr. D.J.M. Peters |
| | Prof. dr. M.H. Kayser – Erasmus MC |
| | Prof. W. Parson, PhD – Institute of Legal Medicine, Medical University of Innsbruck (GMI), Austria |

aan mijn ouders
aan Melanie
ter nagedachtenis aan Martijn van der Gaag

# Table of contents

Chapter 1: 9

Introduction / Outline

Chapter 2 21

TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes

Chapter 3 65

Forensic Nomenclature for Short Tandem Repeats updated for sequencing

Chapter 4 73

Massively Parallel Sequencing of Short Tandem Repeats – Population data and mixture analysis results for the PowerSeq™ system

Chapter 5 129

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

Chapter 6: 181

Short hypervariable microhaplotypes: A novel set of very short high discriminating power loci without stutter artefacts

Chapter 7: 221

General Discussion

Epilogue 245

List of publications 247

Summary thesis 251

Nederlandse samenvatting 257

Dankwoord 263

Curriculum Vitae 264

# 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.

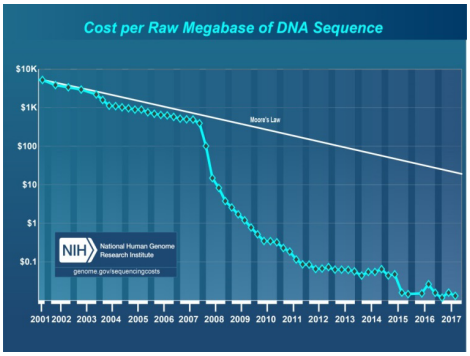



**Figure 1 – cost per raw megabase of DNA sequence in combination with Moore's Law**
*This graph (source: National Human Genome Research Institute) displays the decrease of cost / Mbase of DNA sequence since 2001 in combination with Moore's law (development in increased computer power). It is apparent 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].

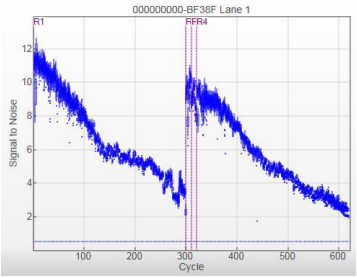



**Figure 2 – Signal to noise ratio of an Illumina MiSeq® run.**
*The graph shows the signal to noise ratio over all data points in an Illumina MiSeq® run for nucleotide incorporation (1nt / cycle). As can be observed, the signal quality decreases as the read progresses for both, read 1 (cycle 1-300)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

# Chapter 2

## TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes

Bioinformatics, Vol. 30 no. 12 2014, pages 1651–1659
Avance Access publication February 13, 2014

Seyed Yahya Anvar
Kristiaan J. van der Gaag
Jaap W. F. van der Heijden
Marcel H. A. M. Veltrop
Rolf H. A. M. Vossen
Rick H. de Leeuw
Cor Breukel
Henk P. J. Buermans
J. Sjef Verbeek
Peter de Knijff
Johan T. den Dunnen
Jeroen F. J. Laros

# Abstract

**Motivation:** Advances in sequencing technologies and computational algorithms have enabled the study of genomic variants to dissect their functional consequence. Despite this unprecedented progress, current tools fail to reliably detect and characterize more complex allelic variants, such as short tandem repeats (STRs). We developed TSSV as an efficient and sensitive tool to specifically profile all allelic variants present in targeted loci. Based on its design, requiring only two short flanking sequences, TSSV can work without the use of a complete reference sequence to reliably profile highly polymorphic, repetitive or uncharacterized regions.

**Results:** We show that TSSV can accurately determine allelic STR structures in mixtures with 10% representation of minor alleles or complex mixtures in which a single STR allele is shared. Furthermore, we show the universal utility of TSSV in two other independent studies: characterizing de novo mutations introduced by transcription activator-like effector nucleases (TALENs) and profiling the noise and systematic errors in an Ion Torrent sequencing experiment. TSSV complements the existing tools by aiding the study of highly polymorphic and complex regions and provides a high-resolution map that can be used in a wide range of applications, from personal genomics to forensic analysis and clinical diagnostics.

**Availability and implementation:** We have implemented TSSV as a Python package that can be installed through the command-line using pip install TSSV command. Its source code and documentation are available at https://pypi.python.org/pypi/tssv and http://www.lgtc.nl/tssv.

# Introduction

As a consequence of various mechanisms such as DNA recombination, replication and repair-associated processes, the spectrum of human genetic variation ranges from single nucleotide differences to large chromosomal events. Among the different types of genetic changes, repetitive DNA sequences show more polymorphism than single nucleotide variants (Conrad et al., 2010; Hinds et al., 2006; Iafrate et al., 2004; Kidd et al., 2008; Redon et al., 2006; Sebat et al., 2004; Tuzun et al., 2005), and they are important in human diseases (Conrad et al., 2010; de Cid et al., 2009; Girirajan et al., 2011; Hollox et al., 2008; McCarroll et al., 2009; Pinto et al., 2010), complex traits and evolution (Mills et al., 2011; Stephens et al., 2011; Sudmant et al., 2010). In particular, microsatellite variants, also known as short tandem repeats (STR), and their expansion/shortening have been linked to a variety of human genetic disorders (Mirkin, 2007; Pearson et al., 2005; Sutherland and Richards, 1995), and have been used in genotyping (Kimura et al., 2009; Weber and May, 1989) and forensic DNA fingerprinting studies (Kayser and de Knijff, 2011; Moretti et al., 2001).

Chapter 2

Because of the repetitive nature of STRs and often the low level of complexity of the DNA sequences in which they occur (Treangen and Salzberg, 2012), characterization of STR variability and understanding of their functional consequences are challenging (Weischenfeldt et al., 2013). So far, sequencing-based strategies have focused on reads mapped to the reference genome and subsequent identification of discordant signatures and classification of associated STRs (Medvedev et al., 2009; Mills et al., 2011). Yet, the mainstream aligners, such as BWA (Li and Durbin, 2009) or Bowtie (Langmead and Salzberg, 2012), do not tolerate repeats or insertions and deletions (indels) as a trade-off of run time (Li and Homer, 2010). This limitation leads to ambiguities in the alignment or assembly of repeats which, in turn, can obscure the interpretation of results (Treangen and Salzberg, 2012). Moreover, the current human genome reference still remains incomplete and provides only limited information on expected and potentially uncharacterized STRs in different individuals (Alkan et al., 2011; Iafrate et al., 2004; Kidd et al., 2008; Sebat et al., 2004). Consequently, STRs are not routinely analyzed in whole-genome or whole-exome sequencing studies, despite their obvious applications and their role in human diseases, complex traits and evolution.

Here, we present a method for targeted profiling of STRs that reports a full spectrum of all observed genomic variants along with their respective abundance. Our tool, TSSV, can accurately profile and characterize STRs without the use of a complete reference genome, and therefore minimizes biases introduced during the alignment and downstream analysis. TSSV scans sequencing data for reads that fully or partially encompass loci of interest based on the detection of unique flanking sequences. Subsequently, TSSV characterizes the sequence between a pair of non-repetitive flanking regions and reports statistics on known and novel alleles for each locus of interest. We show the performance of TSSV on robust characterization of all allelic variants in a given targeted locus by its application in several case studies: forensic DNA fingerprinting of mixed samples by STR profiling, characterization of variants introduced by transcription activator-like effector nucleases (TALENs) in embryonic stem (ES) cells and detailed characterization of errors derived from a next-generation sequencing (NGS) experiment.

# Material and Methods

## TSSV algorithm

The algorithm expects a FASTA file containing sequencing data and a library containing a list of loci of interest that are described by two unique sequences flanking a target locus in the form of a simple regular expression. The description of targeted loci consists of a series of triplets (i.e. CTTA 2 5), each containing a sequence followed by two integers that denote the minimum and maximum number of times the preceding sequence is expected. The notation of expected alleles is then compiled into a regular expression that is used to distinguish between known and new alleles. It is important that a library that contains a description of loci of interest according to the aforementioned instruction should be customized and provided. TSSV reports an overview of marker pair alignments and a detailed description of the identified alleles and their respective frequency per strand. TSSV also provides supporting reads of each locus of interest in separate FASTA files.

TSSV is an open source Python package that can be easily incorporated in any standard NGS pipeline. In addition, we have made the Python package fastools available at https://pypi.python.org/pypi/fastools. fastools offers a series of functions to manipulate, characterize, sanitize and convert FASTQ/FASTA files to other formats. Therefore, it can be used to convert FASTQ files to TSSV desired format (FASTA). For further information on usage and generated data see Supplementary Table S1.

## Marker alignment

Each pair of markers (unique flanking sequences) is aligned to the reads by using a semi-global pairwise alignment, a modified version of the Smith–Waterman algorithm (Smith and Waterman, 1981). The alignment matrix is initialized with penalties only for the aligned sequence and not for the reference sequence. By using this approach, we can use the alignment matrix to calculate the edit distance between the aligned sequence and all substrings of the reference sequence. Finally, TSSV uses the alignment matrix to select the rightmost alignment with a minimum edit distance. To guarantee symmetry with regard to reverse complement sequences, TSSV aligns the reverse complement of the right marker to the reverse complement of the reference sequence.

## Allele identification

Once TSSV successfully aligns a marker pair to either the forward or the reverse complement of the reference sequence, the region of interest is selected by extracting the sequence between the alignment coordinates, which is then converted to the

Chapter 2

forward orientation. The target variable sequence is then matched to the regular expression of the corresponding marker pair for classification as either a known or a new allele. In case of partial identification of markers (i.e. only the left or right marker of the pair is identified), the input sequence is flagged as having either no beginning or no end. The assessment of required runtime for TSSV to identify alleles in datasets with different sequencing depth is provided in Supplementary Figure S1. Each dataset is profiled to characterize 16 allelic STR structures. It should be noted that currently TSSV uses a single processor for the analysis.

## Annotations

Once a list of new alleles is constructed, TSSV uses a revised version of the Mutalyzer online service (Wildeman et al., 2008; https://mutalyzer.nl) to describe all observed variants compared with the reference sequence. Mutalyzer provides a description of observed variants according to the Human Genome Variation Society format for sequence variant description. This can be used to provide an overview of most frequent mutations that are observed within each locus of interest.

## Interpretation guidelines

TSSV provides the frequency in which each allelic structure is observed on plus and minus strand. Based on the experimental design, the frequencies of allelic variants and the balance between supporting reads on the plus and minus strand can aid the identification of potential sequencing biases. Moreover, based on the choice of sequencing technology, homopolymers are prone to introducing artificial allelic structures, so it is advised, when possible, to allow for a tolerance of a few base difference in the homopolymer length while describing targeted loci. The estimation of a lower boundary for the identification of variant alleles is subject to the experimental design. Thus, sequencing of control samples, if possible, can aid a more reliable analysis by ruling out potential slippage and background noise. nce a list of new alleles is constructed, TSSV uses a revised version of the Mutalyzer

## Availability

TSSV is available at http://www.lgtc.nl/tssv and https://pypi.python.org/pypi/tssv. It can also be installed through the command line: pip install tssv. All original datasets and the analysis results can be obtained from figshare (http://www.figshare.com): detection of STRs, SNPs and short indels (Anvar, 2013a), determining de novo structural variations (SVs) in TALEN-treated ES cells (Anvar, 2013b), characterization of STRs (Anvar, 2013c) and detection of systematic errors in PGM (Anvar, 2013c).

## Library preparations and sequencing

STR PCR products for sequencing were generated using the Powerplex® 16-kit from Promega (commercial assay designed and optimized for fluorescent dye-based fragment analysis of STR loci) and were purified with Ampure XP beads according to manufacturer's protocol. Library preparation was performed using the Rapid Library Preparation Kit (Roche). Emulsion PCR and sequencing were performed on the FLX Genome Sequencer (454/Roche) according to the protocol provided by the manufacturer.

PCR products for sequencing of all other samples on the Personal Genome Machine (PGM, Ion Torrent) were prepared using the Ion Plus Fragment Library Kit or amplicon fusion primers. Emulsion PCR was performed using the OneTouch (OT1, Ion Torrent). Sequencing was performed according to LifeTech protocol using the Ion PGM™ 200 Sequencing Kit. PCR reaction was done in 10 µl containing 1× FastStart High Fidelity reaction buffer (Roche), 1.8 mM MgCl2, 2% DMSO, 200 µM dNTPs, 0.5 U FastStart High Fidelity Enzyme Blend (Roche), 20 ng DNA, 300 nM universal barcoding primer, 300 nM reverse target primer and 30 nM forward target primer. After 10 min of initial denaturation at 95°C, 30 PCR cycles were performed at 20s 95°C, 30s 60°C and 40s 72°C. Primer sequences are provided in Supplementary Table S2.

## TALEN design and transfection

The TALENs -pair targeting intron 52 of the human DMD gene (hDMD) was designed using the TALEN toolbox described by Cermak et al. (2011). Next, hDMD/mdx ES cells (t Hoen et al., 2008; Veltrop et al., 2013) were transfected with the TALENs plasmids without any homologous recombination vector. ES cells were routinely cultured on murine embryonic fibroblast (MEF) feeder cells in knockout DMEM supplemented with 2 mM L-glutamine, 1 mM sodium pyruvate, non-essential amino acids, 50 units of penicillin as well as streptomycin, 1000 units of leukemia inhibitory factor and 10% fetal bovine serum (FBS Gold, all from Life Technologies Ltd). Per TALEN, total of 750 ng in 1.5 µg of DNA was used to transfect 1 000 000 hDMD/mdx ES cells using Lipofectamin 2000 (Invitrogen). DNA-Lipofectamin 2000 suspension was prepared in serum and antibiotic-free medium according to the supplier's manual Cells were incubated for 30 min in suspension with the DNA-Lipofectamine mixture and then plated in two 9 cm culture dishes coated with MEF in regular ES culture medium. ES cells were cultured for a week, and DNA was isolated from a pool of 1500 ES clones. This DNA was then prepared for sequencing using Ion Torrent PGM according to the instrument guidelines.

Chapter 2

# Results

## Characterization of STRs

We tested the performance of TSSV in characterizing known STRs from Roche/454 targeted sequencing data of 16 STR loci, amplified in a multiplex reaction. To demonstrate the added value of TSSV over mainstream aligners, we generated four sequencing libraries of which two consisted of pure individual samples and two mixtures in the ratios of 50:50 and 90:10 with comparable depth of coverage (Supplementary Table S3). A full spectrum of STR structures and their abundance was generated after a semi-global alignment of the 25 bp flanking regions adjacent to the STR structure, with tolerance of up to three mismatches (Fig. 1A). On average, 8% of reads remained uncharacterized, mostly because the sequences did not cover both flanking reference sequences or that sequences contained too many mismatches for regions that are required for identification of unique flanking reference sequences (Supplementary Table S3). The PCR product used for preparing the sequencing libraries were generated using the Powerplex 16-kit from Promega, which is an assay designed and optimized for fluorescent dye-based fragment analysis of STR loci. This resulted in a strong imbalance in sequencing yield between STR markers with different dyes in the fragment analysis (Supplementary Table S3). Thus, we restricted the analysis to the three markers with highest coverage (D3S1358, TH01 and D13S317). Frequencies of the observed alleles were interpreted to distinguish actual alleles from slippage artifacts (Supplementary Tables S4–S6).

For D3S1358 (TCTA1 TCTG1-3 TCTA12–13), TSSV robustly identified the STR structure associated with each of the samples, with >91% of reads supporting the presence of two STR alleles (Fig. 1B). In addition, TSSV could pick up a minor frequency (7.25%) for alternative STR structures, in which the DNA amplicons show false STR structures because of DNA polymerase slippage during the amplification (Ellegren, 2004; Hauge and Litt, 1993). Despite the presence of PCR amplification artifacts, the major and minor STR structures in balanced and more extreme mixtures (50:50 and 90:10, respectively) could accurately be identified by TSSV (Fig. 1B and Supplementary Table S4).

We next explored whether TSSV can correctly detect alleles of more complex cases differing based on STR length (ATCT12ATCA2 and ATCT11ATCA3) or composition (CATT9 and CATT3CAT1CATT6) as well as mixtures that shared one allelic STR structure (CATT3CAT1CATT6). Markedly, TSSV could correctly detect, characterize and quantify reads supporting all STR alleles, including mixtures with only 10% representation of the minor alleles (i.e. D3S1358 markers) and more complex

mixtures (TH01 markers) where a single STR allele is shared (Fig. 1C, Supplementary Tables S4–S6 and Supplementary Figs S2–S4). Results of the remaining STR markers are provided in supplementary materials, Supplementary Tables S7–S17.

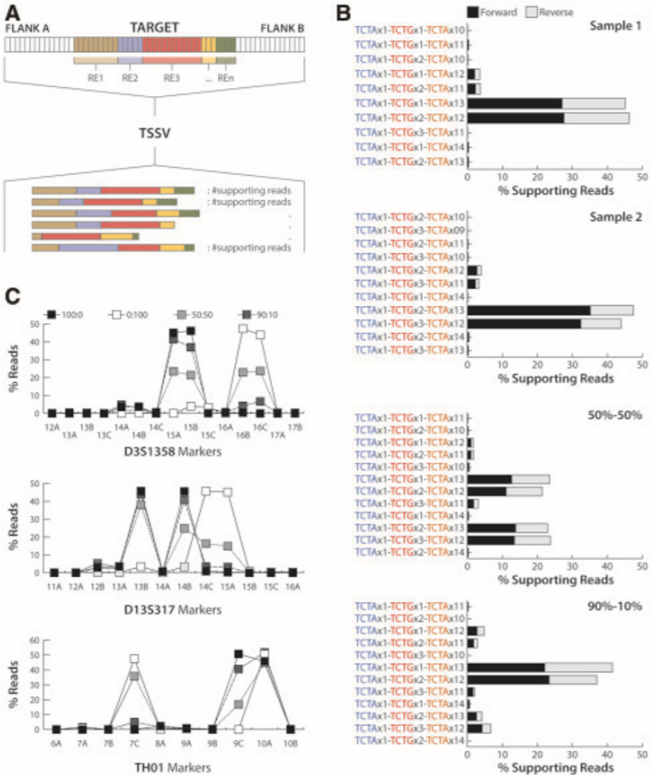



*Fig. 1. Characterization of allelic STR structures in samples and their mixtures of differing ratios (A) Schematic representation of STR structure identification and quantification. After proper alignment of two flanking sequences, TSSV performs a strand-specific classification and quantification of repetitive elements (RE) that constructs a given STR-structure. (B) The number of sequencing reads that support the presence of different allelic D3S1358 STR structures on both strands. Pure samples and their mixtures in two different ratios are presented separately. (C) The proportion of reads that support different allelic STR structures for three most abundant markers (D3S1358, D13S317 and TH01). STR markers differ in complexity based on STR length or composition as well as mixtures in which one allelic STR structure in shared.*

Chapter 2

## Determining de novo structural variations in TALEN-treated cells

TALENs have shown promising potential in site-specific genome editing (Boch, 2011; Cermak et al., 2011; Miller et al., 2011; Zhang et al., 2011). Their modular structure enables simple construction of TALENs that can specifically recognize virtually any DNA sequence of interest. On delivery of a TALENs-pair, a double strand break is introduced that is repaired by non-homologous end-joining, introducing a large variety of mutations (Supplementary Fig. S5). Because the method lacks a positive selection procedure, the applicability depends largely on its efficacy. We used TSSV to estimate the efficiency of genome editing in ES cells from a mouse model with the hDMD, stably integrated in the mouse genome (t Hoen et al., 2008), and determine the utility of an assembled TALEN pair (Supplementary Table S18) in introducing mutations within targeted intron 52 of the hDMD (Supplementary Fig. S6).

For 100.000 TALENs-transfected and non-transfected (control) ES cells, a 135 bp fragment encompassing the entire targeted locus was PCR amplified and sequenced using the Ion Torrent PGM (Supplementary Table S19). The targeted locus was covered over 450 000 times, which allows for precise detection and characterization of any variant present. From the control ES cells, we determined a background of 3.1% of reads that contain at least one mismatch, derived from sequencing errors and potential spontaneous mutations (Fig. 2A and Supplementary Table S19). In TALENs-treated ES cells, the rate of sequencing reads that contain at least one mismatch was 11.4%, almost 4-fold higher than controls (Fig. 2A). The majority of mutations introduced by TALENs pair were small insertions and deletions (75.6%; excluding duplications) (Fig. 2B), which is consistent with the expected type of variants introduced by TALENs (Cermak et al., 2011). The frequency in which individual variants occurred was specific to TALEN-treated ES cells, even for those that were observed with very low frequency (Fig. 2C and D). However, we observed a few mutations that were not specific to TALEN-treated ES cells (Fig. 2D). These were mainly duplications that arose from inaccurate detection of homopolymer stretches. Overall, TSSV results indicate significant enrichment ($P = 2.85E-09$; Kolmogorov–Smirnov test) of variants in TALEN-treated ES cells as compared with controls (Fig. 2E). Furthermore, TSSV reported a list of the most frequent variants and cleavage sites, majority of which were either exclusive to TALEN-treated ES cells or with over 3-fold higher frequency in TALEN-treated ES cells than controls (Supplementary Fig. S7). The Ion Torrent variant caller (version 3.2) did not report any variant because of the nature and frequency of variants introduced by TALENs.

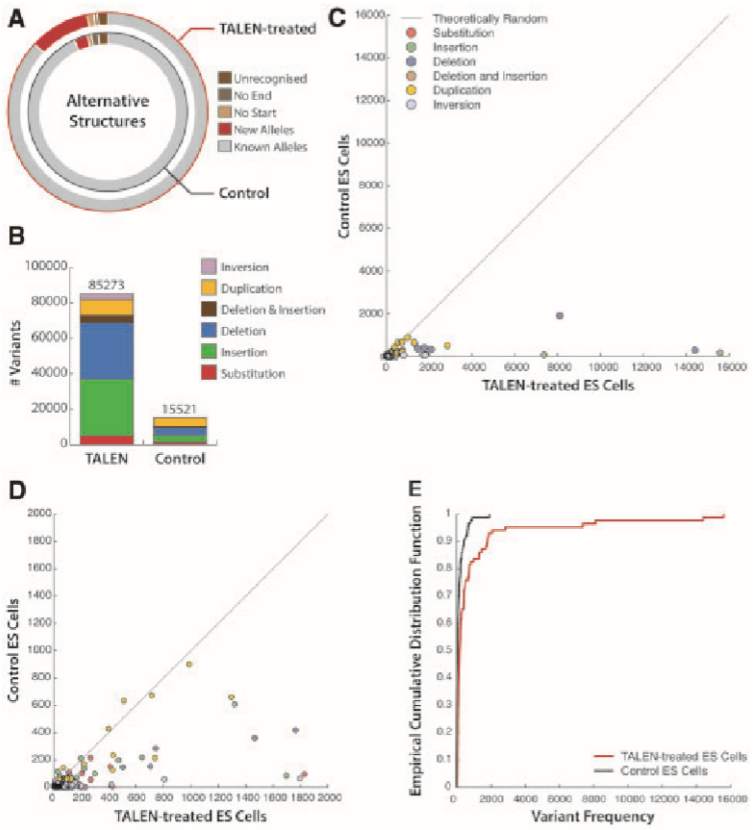



**Fig. 2.** *Variant characterization and quantification of TALEN-treated and Control ES Cells. (A) Basic statistics of the TSSV analysis. Pie charts show the proportion of sequencing reads that support the presence of new alleles in TALEN-treated (outer circle) or control (inner circle) ES Cells. No Start and No End fragments represent reads in which one of the flanking sequences was not recognized. (B) Total number of variants in TALEN-treated ES Cells and Controls, grouped by type. (C) Comparative analysis of the number of occurrences for individual variants in both samples. Data points are colored based on the type of variation. (D) Zoomed in scatter plot for variants with frequencies lower than 2000. (E) Empirical cumulative distribution of variant frequencies for TALEN-treated (red) and Control ES Cells (black). Kolmogorov–Smirnov test was performed to assess if two distributions are significantly different.*

## Detection of systematic errors in PGM Ion Torrent

During the targeted Ion Torrent resequencing of exon 19 of the DMD gene (X-chromosome) in five male patients and a female carrier, we observed a number of shared and unexplained heterozygous variants given that male patients have only one X-chromosome and DMD gene does not locate within pseudo-autosomal regions. We used TSSV to provide a high-resolution map of all sequence variants as a way to understand the origin of these artifacts (Fig. 3A). To assess the reproducibility of our findings, we performed two independent Ion Torrent PGM sequencing runs (PG090 and PG109). Two different versions of the Ion Torrent base-calling algorithm were used

Chapter 2

for PG090 (versions 2.2 and 3.0) while PG109 was only processed by version 3.0 (sequencing run was carried out after the upgrade of the Ion Torrent Suit). The three datasets enabled us to investigate potential artifacts derived from sequencing and/or different base-calling algorithms. Our first observation indicated a significant decrease in the total number of reads (average of 11.3 and 13.3% in respective to different runs and base-calling algorithms) that were recognized per individual (Supplementary Table S20). We also noticed a significant difference in the fraction of reads per dataset (44.3, 40.3 and 48.7%) that were reported as new alleles, having at least one mismatch with the reference sequence (Fig. 3B).

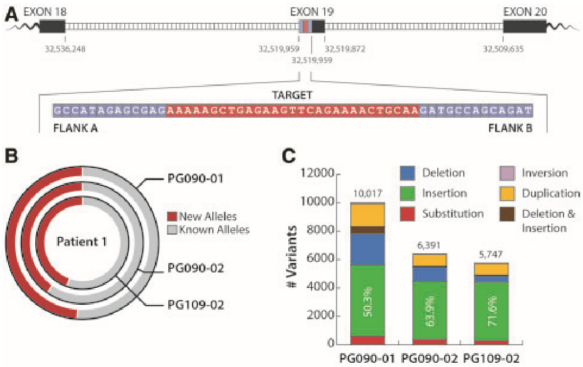



***Fig. 3**. Identification of mutations within exon 19 of DMD gene. (A) Schematic representation of the locus of interest for resequencing, the design of unique flanking sequences (blue), and the targeted region (red) to be profiled using TSSV. (B) Pie charts show the proportion of reads that support the presence of new alleles (red) in sequencing library of patient 1. Pie charts represent different sequencing runs (PG090 or PG109) or the base-calling algorithm used during the primary analysis (01 or 02). The two most outer pie charts are sequencing reads from the same PGM Ion Torrent run processed using two different versions of base-calling algorithm. The most inner pie chart represents an independent run of the same library. (C) Number of observed variants separated by variation type. Percentages show the proportion of insertion events from the total number of variants in each set.*

We observed a significant reduction of variants (36.2%) after adoption of the version 3.0 base-caller, mainly affecting the level of deletions and duplications calls (Fig. 3C). This prominent decrease (68.3 and 48.9%) arises from improvement of the algorithm in determining the length of homopolymer stretches. Notably, the majority of other variants were single nucleotide insertions (excluding duplications and indels) that remained at a comparable rate across different datasets (Fig. 3C). Next, we assessed the strand specificity of the variants based on the sequencing direction. Interestingly, while the majority of variants showed a similar frequency in both directions, the most frequent variants showed a clear imbalance between forward and reverse strand (Fig. 4A and B). The observed strand-specific bias was reproducible and was not influenced by the software version, as it was observed in all three datasets (Fig. 4C and D and Supplementary Figs S8–S11).

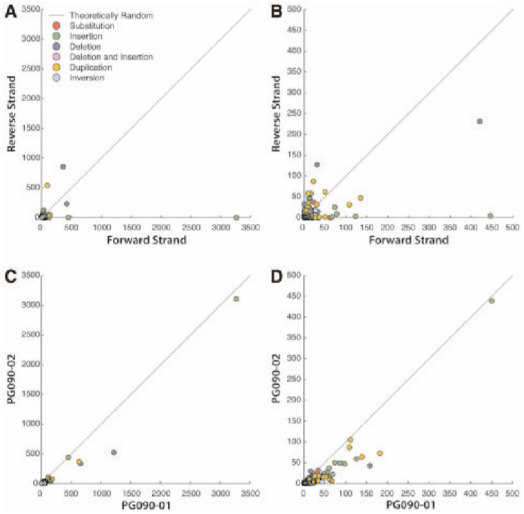



*Fig.4. Comparative analysis of observed mutations. (A) The total number of occurrences for variants to be observed on plus or minus strands is compared. Data points are colored based on the variation type. (B) Zoomed in scatter plot for variants with strand-specific frequency <500. (C) Sequencing data from the same sequencing run (PG090) are assessed for frequency of observed variants after the use of two different versions of base-calling algorithm. (D) Zoomed in scatter plot for variants with frequency <500 compared between two datasets generated using different base-caller algorithms*

To study the possible nucleotide-specific biases, we quantified the frequency of all calls that predominantly occurred on one strand. Despite slight variation, substitutions were observed on both strands at a comparable rate. However, in each dataset, the majority of substituted bases were 'A's (59.2, 61.3 and 64.9%) and 'T's (28.1, 23.0 and 20.5%) that were predominantly substituted to 'G' and 'C', respectively (Fig. 5A). Insertions were primarily observed on the forward strand (94%, on average) while 'A' remained as the most affected base (77.7%, on average) across all samples (Fig. 5B and Supplementary Figs S9–S11).

We also observed a slight enrichment of deletions and duplications on the reverse strand that were more pronounced in PG109-02 (Fig. 5C and D). Consistently, the most affected base was 'A', which was mainly the result of under- or over-calling of 'A' homopolymers. We used TSSV to report a list of most occurring variants across different samples. A single 'A' nucleotide insertion at cycle 52 was by far the most predominant variant that occurred exclusively on the forward strand (Fig. 6A). In fact, irrespective of co-occurrence of this insertion with any other variants, the new observed sequence remains strand specific (Fig. 6A). This cannot be explained from a biological standpoint and can only arise from a sequencing error. Moreover, we did not observe any variation after sequencing the same library with Sanger sequencing (Fig. 6B), ruling out the possibility of artifacts introduced by sample preparation and PCR amplification.

Chapter 2

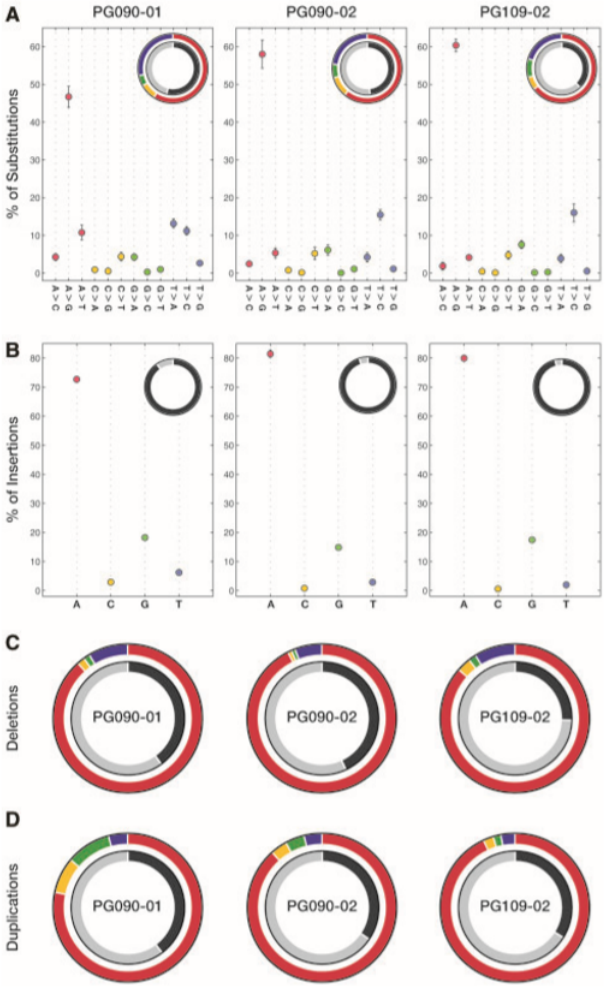



*Fig. 5. Strand-specific and nucleotide-dependent variation patterns in targeted sequencing data. (A) Breakdown of substitution type frequencies across all samples. Pie charts depict substitution events per nucleotide (outer circle) and strand (inner circle) for each dataset. The outer pie charts illustrate the proportion of substitutions based on the preferred nucleotide to which substitutions are made. A, C, G and T nucleotides are reflected in red, yellow, green and blue colors, respectively. Black and gray colors represent the plus and minus strands, respectively. (B) Breakdown of insertion frequencies per nucleotide across all samples. Pie charts represent the proportion of insertion events per nucleotide (outer circle) and strand (inner circle) for each dataset. (C) The fraction of deleted bases is shown in pie charts based on the nucleotide (outer circle) and strand (inner circle) for each dataset. (D) The amount of duplications per nucleotide (outer circle) and strand (inner circle) is presented for three datasets used in this study.*

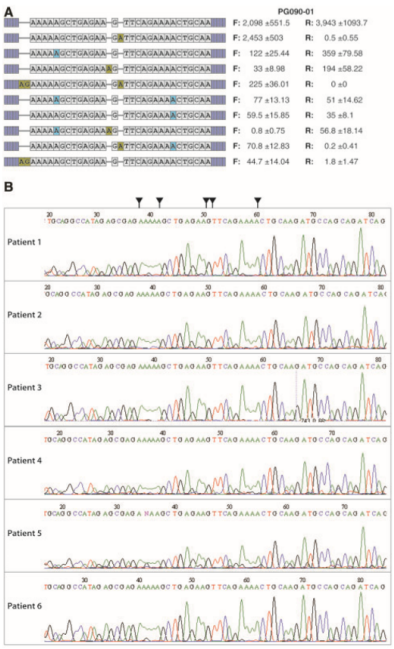



*Fig. 6. Most abundant sequences and validation by Sanger sequencing. (A) A list of most abundant sequences across different samples. The observed occurrence of each new allele is quantified per strand along with corresponding standard deviation. (B) The result of Sanger sequencing of the target locus is presented per sample. Arrows mark the location in which mutations are reported based on the PGM Ion Torrent sequencing data.*

## Comparative analysis of TSSV performance

To our knowledge, lobSTR (Gymrek et al., 2012), STRait Razor (Warshauer et al., 2013) and RepeatSeq (Highnam et al., 2013) are currently the most recent and frequently used STR profiling tools. STRait Razor has limited functionality and only provides an estimated copy number of major STR units. Therefore, we could only compare the performance of TSSV only with lobSTR and RepeatSeq. As lobSTR relies on alignment of sequencing reads to the predefined and indexed STR reference sequences, lobSTR outperformed TSSV in recognizing partial reads, containing only one of the two flanking sites required by TSSV (Supplementary Table S21). Concordantly, lobSTR accepted 1288 reads for the D3S1357 STR locus that were not reported by TSSV. However, TSSV performed significantly better on more complex STR loci (Supplementary Tables S21–S22). Across all four datasets, TSSV identified on average 2471 and 2353 reads in excess of what was recognized by lobSTR for the D13S317 and TH01 STR loci, respectively. This difference is mainly derived from increasingly problematic alignments in lobSTR that is also reflected in inaccurate estimation of STR copy number for TH01 and D13S317 markers in pure samples (Supplementary

Table S22). In addition, lobSTR does not provide information on allelic STR structure, as it only reports the copy number of the major and uninterrupted STR unit and ignores the information from other variable elements or variants outside the STR itself. Consequently, lobSTR failed to accurately detect the presence of mixed simples even in cases in which samples were mixed 50:50 (Supplementary Table S22). Although the information on strand specificity of the aligned reads is present in the alignment file, unlike TSSV, lobSTR does not provide the frequency in which each STR structure is observed. This is an important measure to detect inconsistencies and to rule out potential artifacts. RepeatSeq requires aligned data and uses predefined regions to characterize observed STR alleles. Thus, reads were mapped to the reference genome (hg19) using GS mapper, specifically designed for 454 sequencing data. RepeatSeq reported results for only one STR locus (D8S1179), despite sufficient coverage for a number of STR loci in the BAM file. After manipulating the region descriptions, we could not improve the efficiency of RepeatSeq in identifying the targeted STR loci. Thus, the result of RepeatSeq could not be used for a conclusive comparison with TSSV.

# Discussion

In the past decade, advances in sequencing technologies as well as computational analysis tools have enabled the study of genomic variations to dissect the mechanisms by which they exert their function in the case of human diseases, evolution and other complex traits. Despite this unprecedented progress, structural variations and repetitive DNA sequences (such as STRs) or coupling of de novo mutations present major obstacles for accurate and reliable allelic analysis (Alkan et al., 2011; Gymrek et al., 2012; Kidd et al., 2008; Treangen and Salzberg, 2012; Weischenfeldt et al., 2013). In particular, most computational tools are not ideal to identify STRs because of biases introduced during alignment as well as strong reliance of algorithms on coverage depth or the presence of split-reads. Here, we present a method (TSSV) that provides a high-resolution map of allele-specific genomic variants within targeted loci of interest. Our approach does not rely on the use of a complete reference sequence to reliably profile highly polymorphic sequences (such as STRs) or uncharacterized variants at a single-nucleotide resolution. However, it does require two unique flanking sequences that harbor the region of interest to identify supporting reads. We assess the performance of TSSV on profiling known allelic STR structures across pure samples from a single individual as well as mixed samples with variable abundance. Of 16 allelic STR structures that were targeted for sequencing, six STR loci were sufficiently covered so that the associated allelic STR structures could be reliably resolved. The strong imbalance between yield of STR markers is because of the assay (designed and optimized for fluorescent dye-based fragment analysis of STR loci) used for preparing the sequencing

library. We show that sensitivity of TSSV in determining allelic STR structures exceeds mixtures with only 10% representation of minor alleles and more complex mixtures in which a single STR allele is shared. The lower boundary of detecting minor allele frequencies is subject to experimental design and the complexity of the targeted locus that may result in variable rate of slippage and background noise. Our detailed analysis of three STR loci provides significant insights into forensic DNA fingerprinting of mixed samples while it confirms the feasibility of TSSV to profile causal allelic expansion of triplet, tetranucleotide or more complex repeat structures in variety of human disorders (Brook et al., 1992; Dere et al., 2004; Kremer et al., 1991; Mahadevan et al., 1992; Mirkin, 2007; Pearson et al., 2002; Verkerk et al., 1991).

Second, we sought to profile and annotate the full spectrum of de novo mutations introduced by TALENs that specifically target intron 52 of hDMD in mouse ES cells. The applicability of designed TALENs to introduce mutations in a targeted locus largely depends on its efficacy because this method lacks a selection procedure. Detected TALEN-specific editing events were almost exclusively insertions and deletions that fit the expected mutation profile of TALENs (Cermak et al., 2011). Although it has recently been reported that TALENs induce insertions at a much lower frequency than deletions (Kim et al., 2013), we have observed an extremely balanced rate of insertion and deletion events (37.26% versus 37.20%, respectively). Nevertheless, TALENs-induced deletions tend to affect more bases than insertions. We show that TSSV can resolve difficult-to-call editing events that affect the length of homopolymers based on the variant frequency in TALEN-treated ES cells versus controls. Moreover, the result of TSSV analysis of TALEN-treated and control ES cells suggests that observed de novo structural variants are predominantly caused by initiation of a double-strand break that is repaired by non-homologous end-joining mechanism and are not the result of sequencing errors. Notably, the Ion Torrent variant caller failed to identify any of the observed variants because of their complexity, and therefore does not provide any information on de novo allelic structures that were introduced.

As laboratories begin to generate deep coverage sequencing data to identify low frequent mutations (i.e. cancer genomics), the robustness and accuracy of NGS technology and library preparation methods has become vital (Costello et al., 2013). After running TSSV on a third dataset to identify potential causal mutations in samples from five DMD patients and one female carrier, we observed numerous systematic errors introduced by the Ion Torrent PGM sequencer or the base-calling algorithms. The number of sequencing reads that support the presence of a new allele was in excess of 45% while no mutation was found after Sanger sequencing of the same libraries. Moreover, the amount of allelic discordant reads were unexpected and could not be biologically explained as five out of six samples were derived from male

Chapter 2

patients who are expected to have only one copy of the X-chromosome. Across all samples, the majority of detected variations were single nucleotide insertions (~62%), excluding duplications, that were mostly the result of a single 'A' insertion (78%). Surprisingly, insertions were predominantly specific to the plus strand (94%) that can be the result of flow order in specific sequence contexts. Although the second base-caller improved the deletions and duplications rates that were derived from over- or under-calling of homopolymers, the insertion rates remained unchanged. We further observed a preference for erroneous substitution events that were more pronounced in the second base-caller. However, we were unable to identify motifs that may be associated with observed biases. We argue that the result of TSSV analysis and its ability to provide a high-resolution map of variants ever more highlights the importance of robust and vigorous assessment of downstream analysis as we generate volumes of sequencing data to identify rare mutations and in the advent of NGS in clinical diagnosis.

To demonstrate the added value of TSSV over mainstream STR profiling tools, we ran lobSTR (Gymrek et al., 2012) and RepeatSeq (Highnam et al., 2013) on four samples used for resolving allelic STR structures. Because RepeatSeq hardly reported any STR markers, the performance of TSSV could only be compared with that of lobSTR. We show that TSSV robustly and accurately resolved allelic STR structures with differing complexity. TSSV outperformed lobSTR in reporting the accurate copy number of major STR unit while it provides additional information on allelic STR structures and their strand-specific frequencies. Notably, TSSV excelled in resolving complex mixtures, whereas lobSTR failed to differentiate STR structures associated with different samples, and therefore produced unreliable and inaccurate estimations. Although lobSTR performs well on genotyping diploid samples, there is a clear need for tools to resolve mixtures with differing level of complexity and abundance.

Currently, the major limitation of TSSV is the sequencing read length because the detectable allelic structures are restricted to those that can entirely be covered by a single read. Thus, we envision that the immediate developmental outlook for TSSV can be the inference of allelic locus structure by local assembly of partial reads (reads with only one recognizable flanking region) combined with the comparative analysis of coverage of targeted loci and flanking regions. Furthermore, the promise of novel sequencing technologies (such as Pacific Biosciences RS II), and therefore significant increase in read length will aid the study of larger structural variations.

Advances in sequencing technologies and computational analysis algorithms in unraveling genetic variations from SNPs and indels to CNVs (Chen et al., 2009; DePristo et al., 2011; Goya et al., 2010; Koboldt et al., 2009; McKenna et al., 2010;

Ye et al., 2009) have facilitated the study of experimental data on an unprecedented scale to better understand the functional consequences of genetic variations. TSSV complements the existing tools by aiding the study of unknown, uncharacterized or highly polymorphic and repetitive short structural variations that can be used in a wide range of applications, from personal genomics to forensic analysis and clinical diagnostics.

# Acknowledgements

The authors thank Dr. Annemieke Aartsma-Rus for her constructive feedback and discussions. S.Y.A. and K.J.vdG. performed the analyses. S.Y.A., K.J.vdG., M.H.A.M.V., J.S.V., P.dK. and J.F.J.L. designed the study. S.Y.A., K.J.vdG., J.W.F.vdH. and J.F.J.L. were involved in developing the tool. M.H.A.M.V., R.H.A.M.V., R.H.dL., C.B. and H.P.J.B. performed the wet-lab experiments and sequencing. J.T.dD., P.dK. and J.F.JL. coordinated the study. S.Y.A. drafted the manuscript that was subsequently revised by all co-authors.

Funding: This work was partially supported by the Centre for Molecular Systems Biology (CMSB), Duchenne Parent Project (the Netherlands) and a grant from the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands.

Conflict of Interest: none declared.

# References

- Alkan,C. et al. (2011) Genome structural variation discovery and genotyping. Nat. Rev. Genet., 12, 363–376.
- Anvar,S.Y. (2013a) Allele-specific characterization of STR structures in pure and mixed forensic samples using TSSV,
- Anvar,S.Y. (2013b) Characterization of DeNovo structural variations induced by TALENs targeting hDMD in mouse ES cells using TSSV, http://dx.doi.org/10. 6084/m9.figshare.757790.
- Anvar,S.Y. (2013c) Characterizing IonTorrent PGM Error Profiles using TSSV,
- Boch,J. (2011) TALEs of genome targeting. Nat. Biotechnol., 29, 135–136.
- Brook,J.D. et al. (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 30 end of a transcript encoding a protein kinase family member. Cell, 69, 385.
- Cermak,T. et al. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res., 39, e82.
- Chen,K. et al. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic

Chapter 2

structural variation. Nat. Methods, 6, 677–681.
- Conrad,D.F. et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature, 464, 704–712.
- Costello,M. et al. (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res., 41, e67.
- de Cid,R. et al. (2009) Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat. Genet., 41, 211–215.
- DePristo,M.A. et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet., 43, 491–498.
- Dere,R. et al. (2004) Hairpin structure-forming propensity of the (CCTG.CAGG) tetranucleotide repeats contributes to the genetic instability associated with myotonic dystrophy type 2. J. Biol. Chem., 279, 41715–41726.
- Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet., 5, 435–445.
- Girirajan,S. et al. (2011) Relative burden of large CNVs on a range of neurodevelopmental phenotypes. PLoS Genet., 7, e1002334.
- Goya,R. et al. (2010) SNVMix: predicting single nucleotide variants from nextgeneration sequencing of tumors. Bioinformatics, 26, 730–736.
- Gymrek,M. et al. (2012) lobSTR: A short tandem repeat profiler for personal genomes. Genome Res., 22, 1154–1162.
- Hauge,X.Y. and Litt,M. (1993) A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. Hum. Mol. Genet., 2, 411–415.
- Highnam,G. et al. (2013) Accurate human microsatellite genotypes from highthroughput resequencing data using informed error profiles. Nucleic Acids Res., 41, e32.
- Hinds,D.A. et al. (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat. Genet., 38, 82–85.
- Hollox,E.J. et al. (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat. Genet., 40, 23–25.
- Iafrate,A.J. et al. (2004) Detection of large-scale variation in the human genome. Nat. Genet., 36, 949–951.
- Kayser,M. and de Knijff,P. (2011) Improving human forensics through advances in genetics, genomics and molecular biology. Nat. Rev. Genet., 12, 179–192.
- Kidd,J.M. et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature, 453, 56–64.
- Kim,Y. et al. (2013) TALENs and ZFNs are associated with different mutation signatures. Nat. Methods, 10, 185.
- Kimura,M. et al. (2009) Rapid variable-number tandem-repeat genotyping for Mycobacterium leprae clinical specimens. J. Clin. Microbiol., 47, 1757–1766.
- Koboldt,D.C. et al. (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics, 25, 2283–2285.
- Kremer,E.J. et al. (1991) Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science, 252, 1711–1714.
- Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2, Nat. Methods, 9, 357–359.
- Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrowswheeler transform. Bioinformatics, 25, 1754–1760.

- Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for nextgeneration sequencing. Brief. Bioinform., 11, 473–483.
- Mahadevan,M. et al. (1992) Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. Science, 255, 1253–1255.
- McCarroll,S.A. et al. (2009) Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. Nat. Genet., 41, 1341–1344.
- McKenna,A. et al. (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20, 1297–1303.
- Medvedev,P. et al. (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods, 6, S13–S20.
- Miller,J.C. et al. (2011) A TALE nuclease architecture for efficient genome editing. Nat. Biotechnol., 29, 143–148.
- Mills,R.E. et al. (2011) Mapping copy number variation by population-scale genome sequencing. Nature, 470, 59–65.
- Mirkin,S.M. (2007) Expandable DNA repeats and human disease. Nature, 447, 932–940.
- Moretti,T.R. et al. (2001) Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. J. Forensic Sci., 46, 647–660.
- Pearson,C.E. et al. (2002) Slipped-strand DNAs formed by long (CAG)*(CTG) repeats: slipped-out repeats and slip-out junctions. Nucleic Acids Res., 30, 4534–4547.
- Pearson,C.E. et al. (2005) Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet., 6, 729–742.
- Pinto,D. et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature, 466, 368–372.
- Redon,R. et al. (2006) Global variation in copy number in the human genome. Nature, 444, 444–454.
- Sebat,J. et al. (2004) Large-scale copy number polymorphism in the human genome. Science, 305, 525–528.
- Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. J. Mol. Biol., 147, 195–197.
- Stephens,P.J. et al. (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell, 144, 27–40.
- Sudmant,P.H. et al. (2010) Diversity of human copy number variation and multicopy genes. Science, 330, 641–646.
- Sutherland,G.R. and Richards,R.I. (1995) Simple tandem DNA repeats and human genetic disease. Proc. Natl Acad. Sci. USA, 92, 3636–3641.
- t Hoen,P.A.et al. (2008) Generation and characterization of transgenic mice with the full-length human DMD gene. J. Biol. Chem., 283, 5899–5907.
- Treangen,T.J.andSalzberg,S.L.(2012)Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet., 13, 36–46.
- Tuzun,E. et al. (2005) Fine-scale structural variation of the human genome. Nat. Genet., 37, 727–732.
- Veltrop,M.H.A.M. et al. (2013) Generation of embryonic stem cells and mice for duchenne research. PLoS Currents Muscular Dystrophy, 1.
- Verkerk,A.J. et al. (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell, 65, 905–914.

Chapter 2


- Warshauer,D.H. et al. (2013) STRait Razor: a length-based forensic STR allelecalling tool for use with second generation sequencing data. Forensic Sci. Int. Genet., 7, 409–417.
- Weber,J.L. and May,P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet., 44, 388–396.
- Weischenfeldt,J. et al. (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet., 14, 125–138.
- Wildeman,M. et al. (2008) Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum. Mutat., 29, 6–13.
- Ye,K. et al. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics, 25, 2865–2871.
- Zhang,F. et al. (2011) Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. Nat. Biotechnol., 29, 149–153.

# Supplementary materials

## Methodology (extended)

We developed a method to characterise short structural variations (TSSV) which is made available online. In this section, we describe the functionality and design of this program. Calibration of the algorithm and the output is done with optional command line arguments. TSSV can be installed via pip install tssv command.

**Input** – Our method expects two input files: one file containing sequencing data in FASTA format and one file containing the library description. The format of this description is shown in Table S1. The last column of the description is compiled into a regular expression. This regular expression is used to distinguish between known and unknown alleles.

**Marker Alignment** – Each pair of flanking markers is aligned to each read by using semi-global pairwise alignment, a modified version of the Smith-Waterman algorithm. In this adaptation, the alignment matrix is initialised with penalties for the aligned sequence, but not for the reference sequence. By using this approach, we can use the alignment matrix for the calculation of the edit distance between the aligned sequence and all substrings of the reference sequence. Finally, we use the alignment matrix to select the rightmost alignment with a minimum edit distance. To guarantee that this method is symmetrical with regard to the reverse complement, we align the reverse complement of the right marker to the reverse complement of the reference sequence.

**Allele Identification** – If a marker pair can be aligned to either the forward or reverse complement of the reference sequence, we can select the area of interest by extracting the sequence between the alignment coordinates and by converting it into the forward orientation. This area of interest can then be matched to the regular expression of that marker pair. Depending on the match, we either classify the area of interest as a known or new allele.

**Output** – The output of the analysis consists of an overview report that contains general statistics (such as total number of reads, number of matched pairs, number of unique newly identified alleles, etc.), an overview of the marker pair alignment, and per marker a detailed list of identified alleles (both expected and new alleles). If an output directory is selected, a folder is created to store the marker table and, per marker, a subfolder containing the new alleles and split FASTA files for supporting reads.

Chapter 2

**Table S1 -** Full description of settings and generated files

| | |
|---|---|
| **Command Line Optional Arguments** | |
| -h | Show a description of the usage |
| -m | Number of mismatches per nucleotide (default: 0.08) |
| -r | Name of the report file (default: stdout) |
| -d | Name of the output directory |
| -a | Minimum count per allele to be reported (default: 0) |
| **Library Definition** | |
| Column 1 | Name of the marker pair |
| Column 2 | Sequence of the left flanking marker |
| Column 3 | Sequence of the right flanking marker |
| Column 4 | Description of the expected alleles |
| **General Output Files per Marker** | |
| known.fa | Reads classified as known allele |
| new.fa | Reads classified as new alleles |
| unknown.fa | Reads classified as unrecognized |
| knownalleles.csv | Table of known alleles |
| newallels.csv | Table of new alleles |
| **Marker Overview** | |
| name | Name of the marker pair |
| fPaired | Number of pair matches in forward orientation |
| rPaired | Number of pair matches in reverse orientation |
| f1 | Number of left marker matches in forward orientation |
| r1 | Number of left marker matches in reverse orientation |
| f2 | Number of left marker matches in forward orientation |
| r2 | Number of left marker matches in reverse orientation |
| **Allele Overview** | |
| allele | Sequence of the area of interest |
| total | Number of times the allele was found |
| forward | Number of times the allele was found in forward orientation |
| reverse | Number of times the allele was found in reverse orientation |

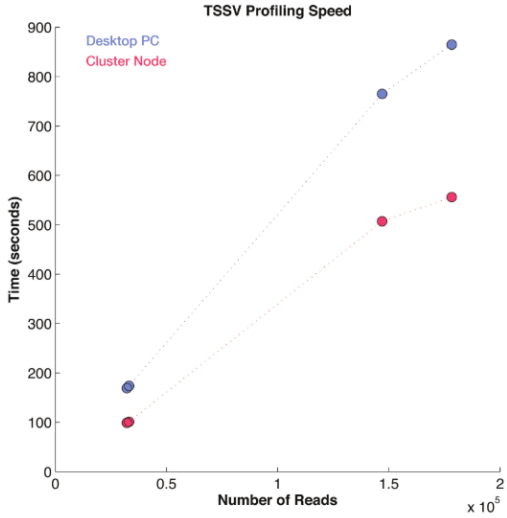



*Figure S1 – The speed of TSSV in characterizing known and novel alleles. Four datasets with different number of reads were profiled for 16 STR loci. The analysis was performed on a desktop PC (Intel Core i7 860, 2.80GHz) and a cluster node (Intel Xeon E5-2660, 2.20GHz). ost abundant sequences and validation by Sanger sequencing. (A) A list of most abundant sequences across different samples. The observed*

**Table S2 -** Description of primers used for targeting loci of interest.

| Name | Forward Primer | Reverse Primer |
|---|---|---|
| hDMD-int52 | gggaaagtgaaagagtaaccagag | ACACACCATCTAATGCTTATGAGG |
| hDMD-exn19 | ccttgtattgaattactcatc | cctaagaagattatctaaatcaactcgt |
| hDMD-exn21 | acgtgttacttactttccatact | caagttagccattttaggctt |

Chapter 2

**Table S3 -** Basic statistics of STR datasets sequenced on 454/Roche.

| | | Reference sample 1 | Reference sample 2 | Mixture 1 (50%/50%) | Mixure 2 (90%/10%) |
|---|---|---|---|---|---|
| Total sequences | | 146968 | 178282 | 33127 | 32201 |
| Sequences recognized as known alleles | | 75837 | 78228 | 16954 | 16425 |
| Sequences recognized as new alleles | | 15622 | 17887 | 4660 | 3720 |
| Sequences only recognized for the beginning | | 35154 | 106253 | 16204 | 16526 |
| Sequences only recognized for the end | | 75110 | 69957 | 9990 | 8782 |
| Unknown sequences | | 11478 | 22153 | 2108 | 2260 |
| **Sequences Recognized / locus** | | | | | |
| D13S317 | 13q31.1 | 15361 | 11473 | 3069 | 3712 |
| D21S11 | 21q21.1 | 13290 | 18477 | 2544 | 3218 |
| Amel | X / Y | 159 | 114 | 37 | 35 |
| D3S1358 | 3p21.31 | 19822 | 25590 | 5507 | 4893 |
| D16S539 | 16q24.1 | 82 | 97 | 26 | 19 |
| TPOX | 2p25.3 | 29 | 80 | 27 | 13 |
| vWA | 12p13.31 | 189 | 215 | 34 | 32 |
| D7S820 | 7q21.11 | 10998 | 7339 | 2626 | 1975 |
| FGA | 4q28 | 14 | 28 | 9 | 6 |
| TH01 | 11p15.5 | 19657 | 17204 | 4664 | 3956 |
| D8S1179 | 8q24.13 | 66 | 80 | 23 | 18 |
| CSF1P0 | 5q33.1 | 81 | 291 | 21 | 27 |
| D18S51 | 18q21.33 | 128 | 247 | 38 | 46 |
| D5S818 | 5q23.2 | 11523 | 14628 | 2983 | 2184 |
| PentaD | 21q22.3 | 25 | 36 | 6 | 5 |
| PentaE | 15q26.2 | 35 | 216 | 0 | 6 |

**Table S4 -** Sequence numbers and allele structure of D3S1358 Short Tandem Repeat.

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 | Mixture interpretation |
|---|---|---|---|---|---|---|---|---|
| **Sample 1** | 12 | 14 | 2 | 16 | TCTA(1) | TCTG(1) | TCTA(10) | |
| | 13 | 49 | 17 | 66 | TCTA(1) | TCTG(1) | TCTA(11) | |
| | 13 | 14 | 16 | 30 | TCTA(1) | TCTG(2) | TCTA(10) | |
| | 14 | 365 | 244 | 609 | TCTA(1) | TCTG(1) | TCTA(12) | |
| | 14 | 403 | 245 | 648 | TCTA(1) | TCTG(2) | TCTA(11) | |
| | 15 | 4681 | 3111 | 7792 | TCTA(1) | TCTG(1) | TCTA(13) | |
| | 15 | 4774 | 3211 | 7985 | TCTA(1) | TCTG(2) | TCTA(12) | |
| | 15 | 12 | 8 | 20 | TCTA(1) | TCTG(3) | TCTA(11) | |
| | 16 | 38 | 19 | 57 | TCTA(1) | TCTG(1) | TCTA(14) | |
| | 16 | 38 | 21 | 59 | TCTA(1) | TCTG(2) | TCTA(13) | |
| **Sample 2** | 13 | 7 | 2 | 9 | TCTA(1) | TCTG(2) | TCTA(10) | |
| | 13 | 9 | 1 | 10 | TCTA(1) | TCTG(3) | TCTA(9) | |
| | 14 | 33 | 13 | 46 | TCTA(1) | TCTG(2) | TCTA(11) | |
| | 14 | 27 | 18 | 45 | TCTA(1) | TCTG(3) | TCTA(10) | |
| | 15 | 592 | 241 | 833 | TCTA(1) | TCTG(2) | TCTA(12) | |
| | 15 | 496 | 214 | 710 | TCTA(1) | TCTG(3) | TCTA(11) | |
| | 16 | 27 | 6 | 33 | TCTA(1) | TCTG(1) | TCTA(14) | |
| | 16 | 7484 | 2602 | 10086 | TCTA(1) | TCTG(2) | TCTA(13) | |
| | 16 | 6898 | 2442 | 9340 | TCTA(1) | TCTG(3) | TCTA(12) | |
| | 17 | 70 | 29 | 99 | TCTA(1) | TCTG(2) | TCTA(14) | |
| | 17 | 35 | 25 | 60 | TCTA(1) | TCTG(3) | TCTA(13) | |
| **50:50 Mixture** | 13 | 3 | 7 | 10 | TCTA(1) | TCTG(1) | TCTA(11) | <1% of highest allele |
| | 13 | 4 | 9 | 13 | TCTA(1) | TCTG(2) | TCTA(10) | -2 stutter Al 15b |
| | 14 | 52 | 32 | 84 | TCTA(1) | TCTG(1) | TCTA(12) | -1 stutter Al 15a |
| | 14 | 51 | 38 | 89 | TCTA(1) | TCTG(2) | TCTA(11) | -1 stutter Al 16b |
| | 14 | 23 | 4 | 27 | TCTA(1) | TCTG(3) | TCTA(10) | -2 stutter Al 16c |
| | 15 | 612 | 520 | 1132 | TCTA(1) | TCTG(1) | TCTA(13) | **Allele 15a** |
| | 15 | 536 | 494 | 1030 | TCTA(1) | TCTG(2) | TCTA(12) | **Allele 15b** |
| | 15 | 88 | 63 | 151 | TCTA(1) | TCTG(3) | TCTA(11) | -1 stutter Al 16c |
| | 16 | 9 | 11 | 20 | TCTA(1) | TCTG(1) | TCTA(14) | +1 stutter Al 15a |
| | 16 | 666 | 439 | 1105 | TCTA(1) | TCTG(2) | TCTA(13) | **Allele 16b** |
| | 16 | 650 | 491 | 1141 | TCTA(1) | TCTG(3) | TCTA(12) | **Allele 16c** |
| | 17 | 14 | 1 | 15 | TCTA(1) | TCTG(2) | TCTA(14) | +1 stutter Al 16b |
| **90:10 Mixture** | 13 | 9 | 2 | 11 | TCTA(1) | TCTG(1) | TCTA(11) | <1% of highest allele |
| | 14 | 117 | 84 | 201 | TCTA(1) | TCTG(1) | TCTA(12) | -1 stutter Al 15a |
| | 14 | 76 | 45 | 121 | TCTA(1) | TCTG(2) | TCTA(11) | -1 stutter Al 15b |
| | 15 | 924 | 808 | 1732 | TCTA(1) | TCTG(1) | TCTA(13) | **Allele 15a** |
| | 15 | 978 | 566 | 1544 | TCTA(1) | TCTG(2) | TCTA(12) | **Allele 15b** |
| | 15 | 70 | 16 | 86 | TCTA(1) | TCTG(3) | TCTA(11) | -1 stutter Al 16c |
| | 16 | 23 | 3 | 26 | TCTA(1) | TCTG(1) | TCTA(14) | +1 stutter Al 15a |
| | 16 | 110 | 65 | 175 | TCTA(1) | TCTG(2) | TCTA(13) | **Allele 16b** |
| | 16 | 177 | 102 | 279 | TCTA(1) | TCTG(3) | TCTA(12) | **Allele 16c** |

**Table S5 -** Sequence numbers and allele structure of D13S317 Short Tandem Repeat.

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 | Mixture interpretation |
|---|---|---|---|---|---|---|---|---|
| **Sample 1** | 11 | 5 | 6 | 11 | ATCT(9) | ATCA(2) | | |
| | 12 | 210 | 192 | 402 | ATCT(10) | ATCA(2) | | |
| | 12 | 7 | 10 | 17 | ATCT(9) | ATCA(3) | | |
| | 13 | 3061 | 3039 | 6100 | ATCT(11) | ATCA(2) | | |
| | 13 | 249 | 222 | 471 | ATCT(10) | ATCA(3) | | |
| | 14 | 59 | 59 | 118 | ATCT(12) | ATCA(2) | | |
| | 14 | 3218 | 2864 | 6082 | ATCT(11) | ATCA(3) | | |
| | 14 | 53 | 49 | 102 | ATCT(10) | ATCA(4) | | |
| | 15 | 14 | 30 | 44 | ATCT(12) | ATCA(3) | | |
| **Sample 2** | 12 | 7 | 13 | 20 | ATCT(10) | ATCA(2) | | |
| | 13 | 182 | 160 | 342 | ATCT(11) | ATCA(2) | | |
| | 13 | 8 | 10 | 18 | ATCT(10) | ATCA(3) | | |
| | 14 | 2495 | 1982 | 4477 | ATCT(12) | ATCA(2) | | |
| | 14 | 151 | 181 | 332 | ATCT(11) | ATCA(3) | | |
| | 15 | 47 | 42 | 89 | ATCT(13) | ATCA(2) | | |
| | 15 | 2409 | 2014 | 4423 | ATCT(12) | ATCA(3) | | |
| | 15 | 22 | 29 | 51 | ATCT(11) | ATCA(4) | | |
| | 16 | 24 | 21 | 45 | ATCT(13) | ATCA(3) | | |
| **50:50 Mixture** | 12 | 35 | 38 | 73 | ATCT(10) | ATCA(2) | | -1 stutter Al 13b |
| | 13 | 525 | 467 | 992 | ATCT(11) | ATCA(2) | | **Allele 13b** |
| | 13 | 30 | 32 | 62 | ATCT(10) | ATCA(3) | | -1 stutter Al 14c |
| | 14 | 191 | 235 | 426 | ATCT(12) | ATCA(2) | | **Allele 14b** |
| | 14 | 327 | 324 | 651 | ATCT(11) | ATCA(3) | | **Allele 14c** |
| | 14 | 6 | 1 | 7 | ATCT(10) | ATCA(4) | | < 1% of highest allele |
| | 15 | 4 | 4 | 8 | ATCT(13) | ATCA(2) | | < 1% of highest allele |
| | 15 | 160 | 229 | 389 | ATCT(12) | ATCA(3) | | **Allele 15c** |
| | 15 | 2 | 4 | 6 | ATCT(11) | ATCA(4) | | +1 stutter Al14c |
| **90:10 Mixture** | 12 | 104 | 58 | 162 | ATCT(10) | ATCA(2) | | -1 stutter Al 13b |
| | 13 | 751 | 589 | 1340 | ATCT(11) | ATCA(2) | | **Allele 13b** |
| | 13 | 48 | 43 | 91 | ATCT(10) | ATCA(3) | | -1 stutter Al 14c |
| | 14 | 66 | 41 | 107 | ATCT(12) | ATCA(2) | | **Allele 14b** |
| | 14 | 691 | 546 | 1237 | ATCT(11) | ATCA(3) | | **Allele 14c** |
| | 14 | 4 | 5 | 9 | ATCT(10) | ATCA(4) | | <1% of highest allele |
| | 15 | 28 | 71 | 99 | ATCT(12) | ATCA(3) | | **Allele 15c** |

**Table S6 -** Sequence numbers and allele structure of TH01 Short Tandem Repeat.

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 | Mixture Interpretation |
|---|---|---|---|---|---|---|---|---|
| **Sample 1** | 8 | 233 | 112 | 345 | CATT(8) | | | |
| | 8.3 | 3 | 22 | 25 | CATT(4) | CAT(1) | CATT(4) | |
| | 8.3 | 91 | 34 | 125 | CATT(3) | CAT(1) | CATT(5) | |
| | 9 | 5280 | 2526 | 7806 | CATT(9) | | | |
| | 9.3 | 4626 | 2417 | 7043 | CATT(3) | CAT(1) | CATT(6) | |
| | 10 | 24 | 12 | 36 | CATT(10) | | | |
| **Sample 2** | 6 | 114 | 34 | 148 | CATT(6) | | | |
| | 6.3 | 8 | 4 | 12 | CATT(5) | CAT(1) | CATT(1) | |
| | 6.3 | 18 | 2 | 20 | CATT(4) | CAT(1) | CATT(2) | |
| | 7 | 5435 | 1412 | 6847 | CATT(7) | | | |
| | 8.3 | 70 | 24 | 94 | CATT(3) | CAT(1) | CATT(5) | |
| | 9.3 | 5789 | 1507 | 7296 | CATT(3) | CAT(1) | CATT(6) | |
| **50:50 Mixture** | 6 | 20 | 23 | 43 | CATT(6) | | | -1 stutter Al 7 |
| | 7 | 514 | 494 | 1008 | CATT(7) | | | **Allele 7** |
| | 8 | 5 | 14 | 19 | CATT(8) | | | -1 stutter Al 9 |
| | 8.3 | 9 | 4 | 13 | CATT(3) | CAT(1) | CATT(5) | -1 Stutter Al 9.3 |
| | 9 | 239 | 235 | 474 | CATT(9) | | | **Allele 9** |
| | 9.3 | 671 | 591 | 1262 | CATT(3) | CAT(1) | CATT(6) | **Allele 9.3** |
| **90:10 Mixture** | 7 | 95 | 60 | 155 | CATT(7) | | | **Allele 7** |
| | 8 | 29 | 23 | 52 | CATT(8) | | | -1 stutter Al 8 |
| | 8.3 | 0 | 11 | 11 | CATT(4) | CAT(1) | CATT(4) | <1% of highest allele |
| | 8.3 | 16 | 7 | 23 | CATT(3) | CAT(1) | CATT(5) | -1 stutter Al 9.3 |
| | 9 | 873 | 404 | 1277 | CATT(9) | | | **Allele 9** |
| | 9.3 | 1044 | 584 | 1628 | CATT(3) | CAT(1) | CATT(6) | **Allele 9.3** |

Chapter 2

**Table S7 -** Sequence numbers and allele structure of D21S11 Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 | Repeat Motif 4 | Repeat Motif 5 | Repeat Motif 6 | Repeat Motif 7 | Repeat Motif 8 | Repeat Motif 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample 1 | 31 | 1593 | 2968 | 4561 | TCTA(6) | TCTG(5) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| | 30 | 1527 | 2443 | 3970 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| | 30 | 151 | 261 | 412 | TCTA(6) | TCTG(5) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(11) |
| | 29 | 167 | 169 | 336 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(11) |
| | 30 | 28 | 102 | 130 | TCTA(5) | TCTG(5) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| | 31 | 36 | 80 | 116 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(13) |
| Sample 2 | 29 | 1182 | 4645 | 5827 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(11) |
| | 31 | 1275 | 3041 | 4316 | TCTA(5) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| | 30 | 200 | 286 | 486 | TCTA(5) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(11) |
| | 28 | 65 | 252 | 317 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(10) |
| | 30 | 130 | 158 | 288 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| 50:50 Mixture | 29 | 256 | 298 | 554 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(11) |
| | 30 | 209 | 264 | 473 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| | 31 | 156 | 186 | 342 | TCTA(6) | TCTG(5) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| | 31 | 115 | 177 | 292 | TCTA(5) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| | 30 | 17 | 15 | 32 | TCTA(5) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(11) |
| | 28 | 14 | 17 | 31 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(10) |
| 90:10 Mixture | 31 | 421 | 634 | 1055 | TCTA(6) | TCTG(5) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| | 30 | 377 | 550 | 927 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| | 29 | 40 | 58 | 98 | TCTA(4) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(11) |
| | 30 | 34 | 47 | 81 | TCTA(6) | TCTG(5) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(11) |
| | 31 | 24 | 46 | 70 | TCTA(5) | TCTG(6) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(12) |
| | 32 | 37 | 10 | 47 | TCTA(6) | TCTG(5) | TCTA(3) | TATCTA(1) | TCTA(2) | TCA(1) | TCTA(2) | TCCATA(1) | TCTA(13) |

**Table S8 -** Sequence numbers and allele structure of D16S539 Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 |
|---|---|---|---|---|---|---|---|
| Sample 1 | 12 | 12 | 57 | 69 | GATA(12) | | |
| | 11 | 1 | 1 | 2 | GATA(11) | | |
| Sample 2 | 13 | 3 | 34 | 37 | GATA(13) | | |
| | 12 | 0 | 27 | 27 | GATA(12) | | |
| 50:50 Mixture | 12 | 5 | 8 | 13 | GATA(12) | | |
| | 13 | 0 | 8 | 8 | GATA(13) | | |
| | 11 | 1 | 0 | 1 | GATA(11) | | |
| 90:10 Mixture | 12 | 7 | 8 | 15 | GATA(12) | | |
| | 11 | 0 | 1 | 1 | GATA(11) | | |

**Table S9 -** Sequence numbers and allele structure of TPOX Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 |
|---|---|---|---|---|---|---|---|
| **Sample 1** | 11 | 3 | 8 | 11 | AATG(11) | | |
| | 9 | 6 | 2 | 8 | AATG(9) | | |
| | 10 | 0 | 1 | 1 | AATG(10) | | |
| **Sample 2** | 11 | 19 | 7 | 26 | AATG(11) | | |
| | 8 | 24 | 2 | 26 | AATG(8) | | |
| | 10 | 1 | 2 | 3 | AATG(10) | | |
| **50:50 Mixture** | 11 | 1 | 2 | 3 | AATG(11) | | |
| | 10 | 1 | 1 | 2 | AATG(10) | | |
| | 9 | 2 | 0 | 2 | AATG(9) | | |
| | 8 | 1 | 0 | 1 | AATG(8) | | |
| **90:10 Mixture** | 11 | 3 | 4 | 7 | AATG(11) | | |
| | 9 | 4 | 0 | 4 | AATG(9) | | |
| | 8 | 0 | 1 | 1 | AATG(8) | | |

**Table S10 -** Sequence numbers and allele structure of vWA Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 | Repeat Motif 4 | Repeat Motif 5 |
|---|---|---|---|---|---|---|---|---|---|
| **Sample 1** | 16 | 67 | 86 | 153 | TCTA(1) | TCTG(4) | TCTA(11) | | |
| | 15 | 3 | 5 | 8 | TCTA(1) | TCTG(4) | TCTA(10) | | |
| | 16 | 0 | 5 | 5 | TCTA(1) | TCTG(5) | TCTA(9) | TCTG(1) | |
| **Sample 2** | 17 | 34 | 50 | 84 | TCTA(1) | TCTG(4) | TCTA(12) | | |
| | 16 | 39 | 42 | 81 | TCTA(1) | TCTG(4) | TCTA(11) | | |
| | 15 | 5 | 1 | 6 | TCTA(1) | TCTG(4) | TCTA(10) | | |
| | 15 | 0 | 1 | 1 | TCTA(1) | TCTG(3) | TCTA(11) | | |
| | 14 | 0 | 1 | 1 | TCTA(1) | TCTG(4) | TCTA(9) | | |
| | 17 | 1 | 0 | 1 | TCTA(1) | TCTG(4) | TCTA(10) | TCCA(1) | TCTA(1) |
| **50:50 Mixture** | 16 | 14 | 5 | 19 | TCTA(1) | TCTG(4) | TCTA(11) | | |
| | 17 | 3 | 6 | 9 | TCTA(1) | TCTG(4) | TCTA(12) | | |
| | 17 | 0 | 1 | 1 | TCTA(2) | TCTG(3) | TCTA(12) | | |
| **90:10 Mixture** | 16 | 12 | 15 | 27 | TCTA(1) | TCTG(4) | TCTA(11) | | |
| | 16 | 1 | 1 | 2 | TCTA(1) | TCTG(4) | TCTA(10) | TCCA(1) | |
| | 16 | 1 | 0 | 1 | TCTA(1) | TCTG(5) | TCTA(10) | | |

**Table S11 -** Sequence numbers and allele structure of D7S820 Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 |
|---|---|---|---|---|---|---|---|
| **Sample 1** | 7 | 3268 | 3037 | 6305 | GATA(7) | | |
| | 12 | 1535 | 1847 | 3382 | GATA(12) | | |
| | 11 | 130 | 102 | 232 | GATA(11) | | |
| | 6 | 75 | 69 | 144 | GATA(6) | | |
| **Sample 2** | 9 | 2927 | 3561 | 6488 | GATA(9) | | |
| | 8 | 140 | 170 | 310 | GATA(8) | | |
| **50:50 Mixture** | 7 | 584 | 454 | 1038 | GATA(7) | | |
| | 9 | 326 | 322 | 648 | GATA(9) | | |
| | 12 | 311 | 289 | 600 | GATA(12) | | |
| | 8 | 28 | 22 | 50 | GATA(8) | | |
| | 11 | 24 | 19 | 43 | GATA(11) | | |
| | 6 | 11 | 17 | 28 | GATA(6) | | |
| **90:10 Mixture** | 7 | 548 | 474 | 1022 | GATA(7) | | |
| | 12 | 284 | 333 | 617 | GATA(12) | | |
| | 9 | 28 | 42 | 70 | GATA(9) | | |
| | 11 | 15 | 29 | 44 | GATA(11) | | |

**Table S12 -** Sequence numbers and allele structure of FGA Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 | Repeat Motif 4 | Repeat Motif 5 | Repeat Motif 6 | Repeat Motif 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Sample 1** | 21 | 7 | 0 | 7 | TTTC(3) | TTTTTCT(1) | CTTT(13) | CTCC(1) | TTCC(2) | | |
| | 17 | 1 | 0 | 1 | TTTC(3) | TTTTTCT(1) | CTTT(9) | CTCC(1) | TTCC(2) | | |
| **Sample 2** | 20 | 5 | 0 | 5 | TTTC(3) | TTTTTCT(1) | CTTT(12) | CTCC(1) | TTCC(2) | | |
| | 24 | 5 | 0 | 5 | TTTC(3) | TTTTTCT(1) | CTTT(16) | CTCC(1) | TTCC(2) | | |
| | 19 | 1 | 0 | 1 | TTTC(3) | TTTTTCT(1) | CTTT(11) | CTCC(1) | TTCC(2) | | |
| | 21 | 1 | 0 | 1 | TTTC(3) | TTTTTCT(1) | CTTT(13) | CTCC(1) | TTCC(2) | | |
| **50:50 Mixture** | 17 | 2 | 0 | 2 | TTTC(3) | TTTTTCT(1) | CTTT(9) | CTCC(1) | TTCC(2) | | |
| | 20 | 1 | 0 | 1 | TTTC(3) | TTTTTCT(1) | CTTT(8) | CCTT(1) | CTTT(3) | CTCC(1) | TTCC(2) |
| | 20 | 1 | 0 | 1 | TTTC(3) | TTTTTCT(1) | CTTT(12) | CTCC(1) | TTCC(2) | | |
| **90:10 Mixture** | 21 | 2 | 0 | 2 | TTTC(3) | TTTTTCT(1) | CTTT(13) | CTCC(1) | TTCC(2) | | |

**Table S13 -** Sequence numbers and allele structure of D8S1179 Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 | Repeat Motif 4 |
|---|---|---|---|---|---|---|---|---|
| **Sample 1** | 10 | 12 | 13 | 25 | TCTA(10) | | | |
| | 13 | 18 | 3 | 21 | TCTA(1) | TCTG(1) | TCTA(11) | |
| | 12 | 2 | 0 | 2 | TCTA(1) | TCTG(1) | TCTA(10) | |
| **Sample 2** | 13 | 13 | 17 | 30 | TCTA(13) | | | |
| | 14 | 10 | 11 | 21 | TCTA(14) | | | |
| | 12 | 0 | 4 | 4 | TCTA(12) | | | |
| | 14 | 1 | 0 | 1 | TCTA(3) | TCTG(1) | TCTA(10) | |
| | 14 | 1 | 0 | 1 | TCTG(1) | TCTA(13) | | |
| **50:50 Mixture** | 13 | 4 | 2 | 6 | TCTA(1) | TCTG(1) | TCTA(11) | |
| | 10 | 1 | 2 | 3 | TCTA(10) | | | |
| | 14 | 0 | 2 | 2 | TCTA(14) | | | |
| | 13 | 1 | 1 | 2 | TCTA(13) | | | |
| **90:10 Mixture** | 10 | 1 | 5 | 6 | TCTA(10) | | | |
| | 13 | 1 | 3 | 4 | TCTA(1) | TCTG(1) | TCTA(11) | |
| | 14 | 1 | 0 | 1 | TCTA(14) | | | |
| | 13.1 | 0 | 1 | 1 | TCTA(1) | TCTG(1) | TCTA(10) | TCTAT(1) |

**Table S14 -** Sequence numbers and allele structure of CSF1P0 Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 |
|---|---|---|---|---|---|---|---|
| **Sample 1** | 10 | 7 | 26 | 33 | AGAT(10) | | |
| | 13 | 6 | 10 | 16 | AGAT(13) | | |
| | 12 | 1 | 1 | 2 | AGAT(12) | | |
| | 9 | 0 | 1 | 1 | AGAT(9) | | |
| | 8 | 1 | 0 | 1 | AGAT(8) | | |
| **Sample 2** | 12 | 19 | 106 | 125 | AGAT(12) | | |
| | 13 | 28 | 67 | 95 | AGAT(13) | | |
| | 11 | 1 | 4 | 5 | AGAT(11) | | |
| | 10 | 0 | 1 | 1 | AGAT(10) | | |
| **50:50 Mixture** | 13 | 1 | 7 | 8 | AGAT(13) | | |
| | 12 | 2 | 1 | 3 | AGAT(12) | | |
| | 10 | 0 | 2 | 2 | AGAT(10) | | |
| | 11 | 0 | 1 | 1 | AGAT(11) | | |
| **90:10 Mixture** | 10 | 1 | 8 | 9 | AGAT(10) | | |
| | 13 | 0 | 3 | 3 | AGAT(13) | | |
| | 8 | 0 | 1 | 1 | AGAT(8) | | |

**Table S15 -** Sequence numbers and allele structure of D18S51 Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 |
|---|---|---|---|---|---|---|---|
| **Sample 1** | 14 | 0 | 35 | 35 | AGAA(14) | | |
| | 15 | 0 | 19 | 19 | AGAA(15) | | |
| | 13 | 0 | 1 | 1 | AGAA(13) | | |
| **Sample 2** | 13 | 0 | 57 | 57 | AGAA(13) | | |
| | 14 | 0 | 35 | 35 | AGAA(14) | | |
| | 12 | 0 | 8 | 8 | AGAA(12) | | |
| | 15 | 0 | 2 | 2 | AGAA(15) | | |
| **50:50 Mixture** | 14 | 0 | 13 | 13 | AGAA(14) | | |
| | 13 | 0 | 2 | 2 | AGAA(13) | | |
| | 15 | 0 | 1 | 1 | AGAA(15) | | |
| **90:10 Mixture** | 14 | 0 | 10 | 10 | AGAA(14) | | |
| | 13 | 0 | 4 | 4 | AGAA(13) | | |
| | 15 | 0 | 3 | 3 | AGAA(15) | | |

**Table S16 -** Sequence numbers and allele structure of D5S818 Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 |
|---|---|---|---|---|---|---|---|
| **Sample 1** | 14 | 2049 | 2825 | 4874 | AGAT(11) | AGAG(1) | |
| | 17 | 1626 | 2140 | 3766 | AGAT(14) | AGAG(1) | |
| | 16 | 118 | 241 | 359 | AGAT(13) | AGAG(1) | |
| | 13 | 106 | 152 | 258 | AGAT(10) | AGAG(1) | |
| **Sample 2** | 14 | 2118 | 4512 | 6630 | AGAT(11) | AGAG(1) | |
| | 15 | 1708 | 3351 | 5059 | AGAT(12) | AGAG(1) | |
| | 13 | 114 | 261 | 375 | AGAT(10) | AGAG(1) | |
| **50:50 Mixture** | 14 | 517 | 765 | 1282 | AGAT(11) | AGAG(1) | |
| | 15 | 258 | 313 | 571 | AGAT(12) | AGAG(1) | |
| | 17 | 156 | 169 | 325 | AGAT(14) | AGAG(1) | |
| | 13 | 33 | 38 | 71 | AGAT(10) | AGAG(1) | |
| | 16 | 8 | 36 | 44 | AGAT(13) | AGAG(1) | |
| **90:10 Mixture** | 14 | 340 | 529 | 869 | AGAT(11) | AGAG(1) | |
| | 17 | 282 | 354 | 636 | AGAT(14) | AGAG(1) | |
| | 16 | 24 | 56 | 80 | AGAT(13) | AGAG(1) | |
| | 15 | 27 | 32 | 59 | AGAT(12) | AGAG(1) | |
| | 13 | 19 | 39 | 58 | AGAT(10) | AGAG(1) | |

**Table S17 -** Sequence numbers and allele structure of Penta E Short Tandem Repeat (6 most frequent alleles for every sample.)

| | Total # of repeats | Forward sequences | Reverse sequences | Total # of sequences | Repeat Motif 1 | Repeat Motif 2 | Repeat Motif 3 |
|---|---|---|---|---|---|---|---|
| Sample 1 | 17 | 0 | 6 | 6 | AAAGA(17) | | |
| | 12 | 0 | 2 | 2 | AAAGA(12) | | |
| Sample 2 | 7 | 0 | 139 | 139 | AAAGA(7) | | |
| 90:10 Mixture | 13 | 0 | 2 | 2 | AAAGA(13) | | |

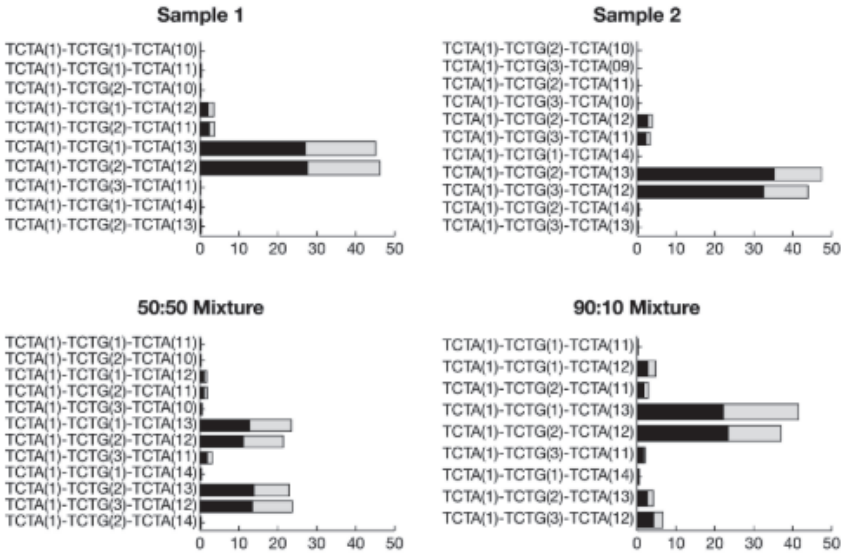



*Figure S2 – The percentage of reads supporting detected allele-structure of D3S1358 Short Tandem Repeat. Black bars depict reads supporting the forward strand and grey bars correspond to the proportion of reads supporting the reverse complement.*

Chapter 2

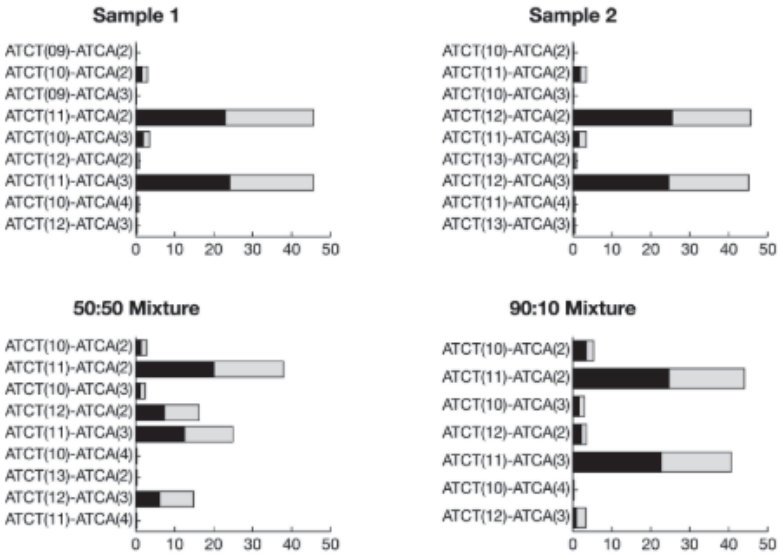



*Figure S3 – The percentage of reads supporting detected allele-structure of D13S317 Short Tandem Repeat. Black bars depict reads supporting the forward strand and grey bars correspond to the proportion of reads supporting the reverse complement.*

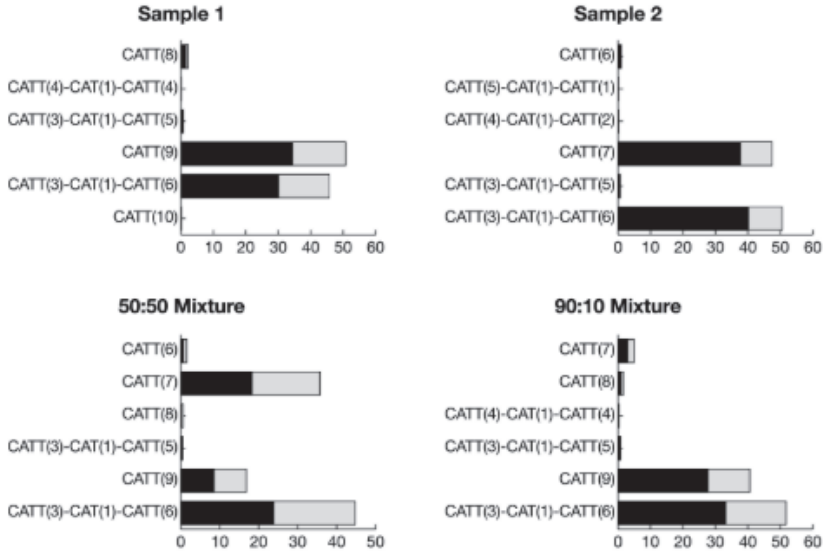



*Figure S4 – The percentage of reads supporting detected allele-structures of TH01 Short Tandem Repeat. Black bars depict reads supporting the forward strand and grey bars correspond to the proportion of reads supporting the reverse complement.*

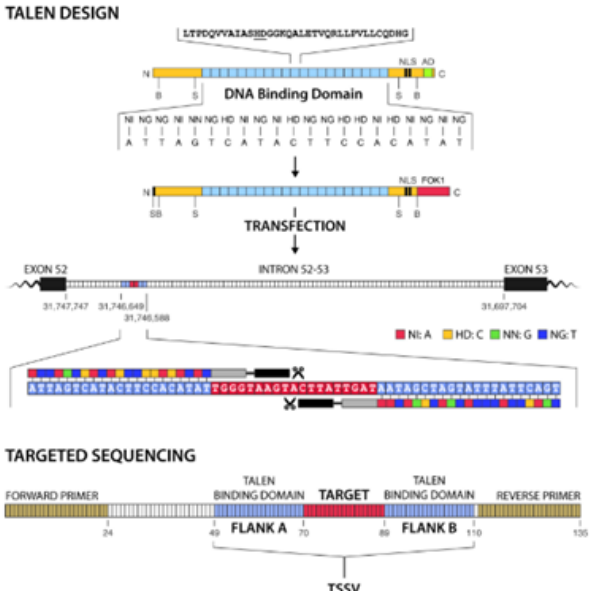



**Figure S5 -** *Schematic representation of TALEN design and targeted resequencing. TALEN-pairs were designed to specifically target intron 52 of the hDMD in mouse ES Cells. The binding sites of TALEN-pairs are 21bp long (blue). After successful transfection, TALENs initiate a double strand break within the target locus of 19bp (red). The 135bp fragment encompassing the entire targeted region was PCR-amplified, sequenced, and analysed using TSSV.*

**Table S18 –** Target sequence of the TALENS targeting intron 52 of the hDMD gene.

| | TAL1 | TAL2 |
|---|---|---|
| **Nucleotide Sequence** | ATTAGTCATACTTCCACATAT | AATAGCTAGTATTTATTCAGT |
| **RVD sequence** | NI NG NG NI NN NG HD NI NG NI HD NG NG HD HD NI HD NI NG NI NG | NI HD NG NN NI NI NG NI NI NI NG NI HD NG NI NN HD NG NI NG NG |

Chapter 2

**Table S19 –** Basic statistics on sequencing reads from TALEN-treated and control ES Cells.

**Table S19 –** Basic statistics on sequencing reads from TALEN-treated and control ES Cells.

| | # Reads | # Unique New Alleles | # Reads for New Alleles | # Reads with no Start | # Reads with no End | # Unrecognized Reads |
|---|---|---|---|---|---|---|
| TALEN-treated ES Cells | 466,633 | 3,868 | 46,337 | 4,225 | 2,861 | 7,229 |
| Control ES Cells | 423,674 | 1,458 | 12,317 | 3,614 | 5,756 | 7,414 |

**Table S20 –** Basic statistics on sequencing reads from 5 male patients and a female carrier.

| | # Reads | # Unique New Alleles | # Reads with no Start | # Reads with no End |
|---|---|---|---|---|
| **PG090-01** | | | | |
| Sample 1 | 16,346 | 1,214 | 2,304 | 673 |
| Sample 2 | 15,589 | 1,161 | 2,215 | 2,208 |
| Sample 3 | 22,397 | 1,593 | 3,533 | 3,387 |
| Sample 4 | 12,773 | 1,009 | 1,867 | 2,094 |
| Sample 5 | 14,834 | 1,093 | 2,206 | 2,581 |
| Sample 6 | 20,599 | 1,307 | 2,626 | 3,241 |
| **PG090-02** | | | | |
| Sample 1 | 14,337 | 498 | 1,536 | 620 |
| Sample 2 | 13,420 | 511 | 1,251 | 2,095 |
| Sample 3 | 18,881 | 594 | 2,004 | 3,098 |
| Sample 4 | 10,995 | 421 | 1,040 | 2,004 |
| Sample 5 | 12,769 | 429 | 1,266 | 2,483 |
| Sample 6 | 18,488 | 571 | 1,599 | 3,237 |
| **PG109-02** | | | | |
| Sample 1 | 11,612 | 447 | 1,411 | 381 |
| Sample 2 | 11,857 | 390 | 1,253 | 2,484 |
| Sample 3 | 17,386 | 510 | 1,964 | 4,018 |
| Sample 4 | 9,849 | 351 | 978 | 2,407 |
| Sample 5 | 11,545 | 365 | 1,200 | 2,949 |
| Sample 6 | 16,816 | 507 | 1,637 | 4,164 |

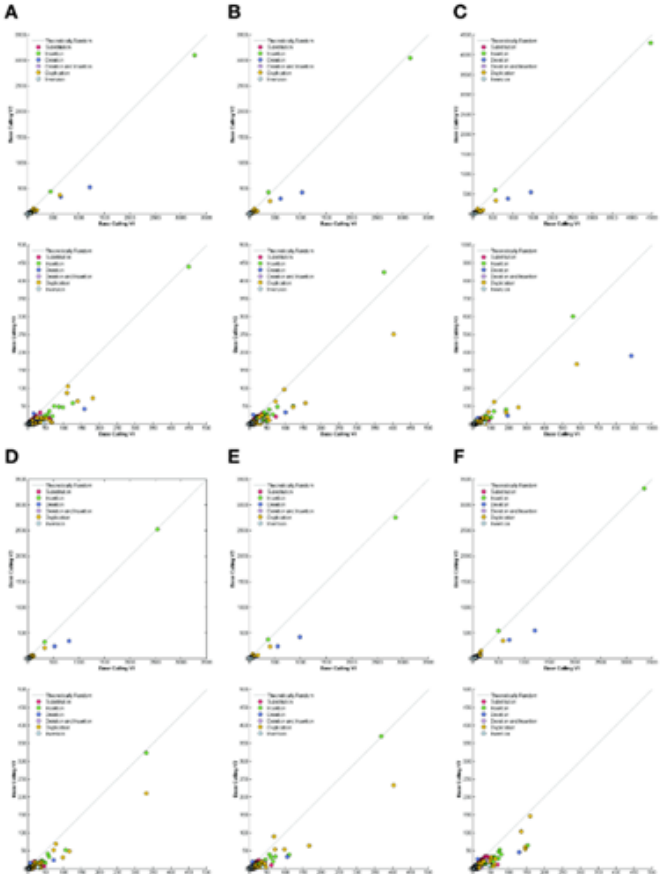



***Figure S8*** *– Variant frequency comparison between two base-calling algorithms for all samples (A, B, C, D, E, F). For each panel, the first scatter plot shows all variants and the second zooms in to variants with frequency less than 500.*

Chapter 2

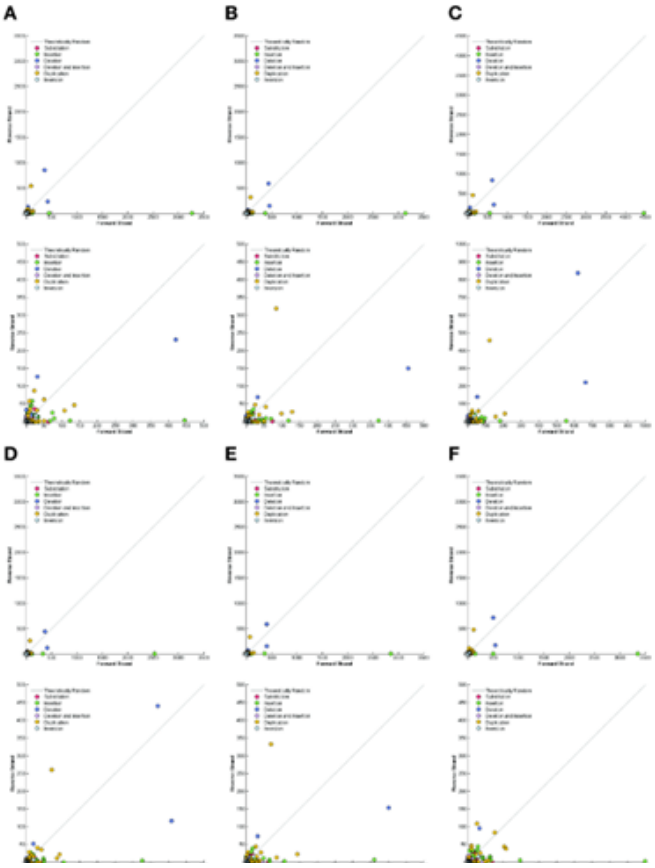



*Figure S9 – Strand specificity of observed variants in dataset PG090-01 for all samples (A, B, C, D, E, F). For each panel, the first scatter plot shows all variants and the second zooms in to variants with frequency less than 500 on each strand.*

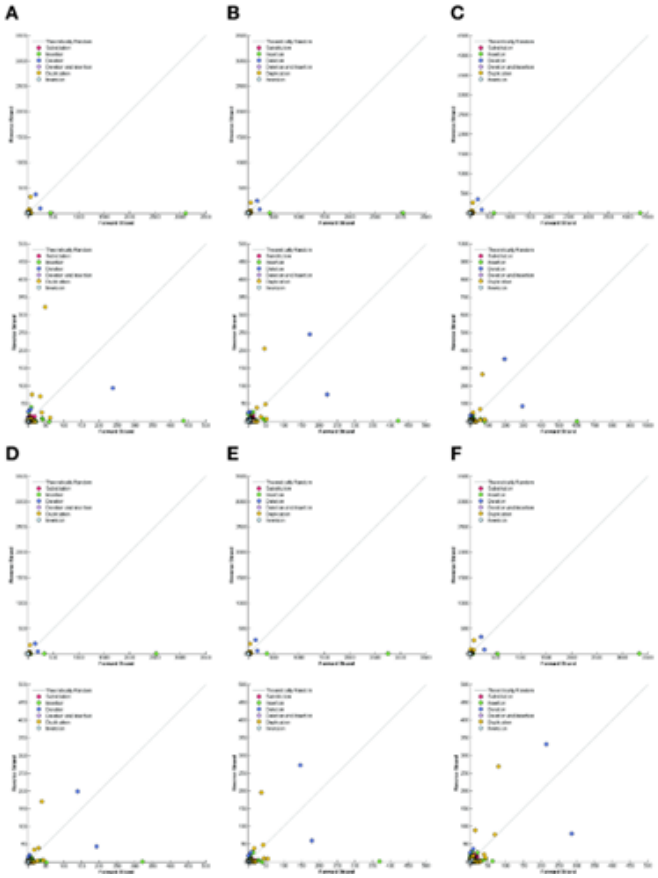



***Figure S10*** *– Strand specificity of observed variants in dataset PG090-02 for all samples (A, B, C, D, E, F). For each panel, the first scatter plot shows all variants and the second zooms in to variants with frequency less than 500 on each strand.*

Chapter 2

*Figure S11 – Strand specificity of observed variants in dataset PG109-02 for all samples (A, B, C, D, E, F). For each panel, the first scatter plot shows all variants and the second zooms in to variants with frequency less than 500 on each strand.*

**Table S21 –** General statistics on identification of informative reads for STR profiling of samples using lobSTR.

| | Sample 1 | Sample 2 | 50:50 Mixture | 90:10 Mixture |
|---|---|---|---|---|
| **Total number of reads** | 146,968 | 178,282 | 33,127 | 32,201 |
| **Number of aligned reads** | 114,122 | 127,327 | 25,339 | 24,379 |
| **Number of stitched reads** | 0 | 0 | 0 | 0 |
| **Number of single-end reads** | 114,122 | 127,327 | 25,339 | 24,379 |
| **Number of supporting end reads** | 0 | 0 | 0 | 0 |
| **Percent of partially aligned reads** | 29.03% | 37.90% | 30.42% | 31.26% |
| **Percent of reverse strand reads** | 53.07% | 50.95% | 53.34% | 51.73% |
| **Percent of non-unit allele reads** | 14.00% | 18.71% | 16.02% | 51.73% |

**Table S22 –** lobSTR performance in identification and characterization of allele specific STR structures.

| | | | | | | | Supporting Reads | | | Validation | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Official Name** | **Chr. Position** | **STR Repeat** | **Ref. Copy Number** | **lobSTR Calls *** | **Coverage** | **Total** | **Allele 1** | **Allele2** | **Copy Number** | **Structure** |
| **Sample 1** | D3S1358 | 3p21.31 | TCTA | 16 | -4,-4 | 23,661 | 20,810 | NA | NA | **Yes** | **No** |
| | D13S317 | 13q31.1 | ATCT | 11 | 0,0 | 14,536 | 13,228 | NA | NA | **Yes** | **No** |
| | TH01 | 11p15.5 | CATT | 7 | 0,11 | 19,072 | 16,828 | 9,038 | 7,790 | **No** | **No** |
| **Sample 2** | D3S1358 | 3p21.31 | TCTA | 16 | 0,0 | 21,535 | 18,668 | NA | NA | **Yes** | **No** |
| | D13S317 | 13q31.1 | ATCT | 11 | 0,5 | 10,588 | 9,994 | 5,738 | 4,256 | **No** | **No** |
| | TH01 | 11p15.5 | CATT | 7 | 11,0 | 16,802 | 15,353 | 7,973 | 7,380 | **Yes** | **No** |
| **50:50 Mixture** | D3S1358 | 3p21.31 | TCTA | 16 | -4,0 | 4,769 | 4,340 | 2,754 | 1,586 | **No** | **No** |
| | D13S317 | 13q31.1 | ATCT | 11 | 0,0 | 2,913 | 2,227 | NA | NA | **No** | **No** |
| | TH01 | 11p15.5 | CATT | 7 | 0,11 | 3,935 | 3,642 | 1,887 | 1,755 | **No** | **No** |
| **90:10 Mixture** | D3S1358 | 3p21.31 | TCTA | 16 | -4,-4 | 5,262 | 4,327 | NA | NA | **No** | **No** |
| | D13S317 | 13q31.1 | ATCT | 11 | 0,0 | 3,264 | 2,822 | NA | NA | **No** | **No** |
| | TH01 | 11p15.5 | CATT | 7 | 11,0 | 3,720 | 3,391 | 1,755 | 1,636 | **No** | **No** |

** lobSTR calls column represents the number of nucleotide differences between the called alleles and the length of the reference STR (STR lengt*

Chapter 2

# Chapter 3

## Forensic Nomenclature for Short Tandem Repeats updated for sequencing

Forensic Science International: Genetics Supplement Series 5 (2015) e542–e544
Published online 26 September 2015

Kristiaan J. van der Gaag
Peter de Knijff

# Abstract

The introduction of Massive Parallel Sequencing (MPS) techniques enables sequencing of Short Tandem Repeats (STR) as a new tool for forensic research. In addition to variation in fragment-length, MPS also reveals allelic sequence-variation in STR-fragments. This additional variation demands a new way of describing allelic variants. Here we propose a nomenclature of MPS-derived STR alleles for use in forensic research.


# Introduction

For over two decades, the analysis of Short Tandem Repeats (STRs) in forensics was routinely performed using Capillary Electrophoresis (CE). With CE, the length of a DNA fragment containing an STR is determined. STR alleles are identified by comparing unknown fragment lengths with a reference allelic ladder containing fragments with known repeat-lengths. The use of a simple number, representing the number of repeats was sufficient as nomenclature for STR allele variation. Recent developments in Massive Parallel Sequencing (MPS) technologies enable high-throughput sequencing of STRs, revealing additional sequence-variation in many of the STRs [1, 2]. A uniform nomenclature for MPS-STR alleles describing this additional variation still needs to be developed. Here, we propose a universal way of describing STR allele variation, specifically designed for use in forensic casework.

# Material and Methods

Previously suggested ways of describing STR- and other genome-variation were compared. Based on published variation [2] and in-house available data for 22 STRs (van der Gaag et al., manuscript in preparation) Human genome coordinates were identified for the genomic regions containing STR-variation, and rules were developed to describe sequence allele-variation within STRs and in repeat-flanking regions.

Chapter 3

# Results

In several studies, MPS of STRs revealed substantial sequence variation in addition to that already described in STRbase [2, 4]. For comparison of published data and, more importantly, for comparison of profiles among databases, it is important that a uniform and straight-forward nomenclature is used. For this nomenclature several aspects should be taken into consideration:

- Different MPS assays will be used to analyse the same markers, introducing variation in the genome-coordinates of the analysed fragment containing an STR (fragments of different assays will not completely overlap).
- Allele-nomenclature should be compact and readable. However, allele-description should also contain all relevant information to reconstruct the original sequence.
- A direct comparison between CE- and MPS-results should be possible.

For CE-nomenclature, ranges have been determined in the past, defining the genomic coordinates of the STRs. For some STRs (like the example of D13S317 discussed below), additional repeating elements adjacent to the defined STR turned out to vary in length resulting in a difference between the total number of variable repeat-units for sequencing and the CE allele-count. Here, we determined genomic coordinates for the region in which STR-variation has been observed for 22 commonly used autosomal STRs (Figure 1a) in the following way:

The start-position of the STR-motif represents the first possible position while retaining maximum length for the longest repeated element. Any repeated elements directly adjacent to the STR of at least three repeats long was included as part of the STR-region. If a complex repeat consists of multiple blocks, interruptions were divided into blocks of the same length where possible (D2S1338, D21S11, FGA and vWA in figure 1a).

For studies of genome variation, HGVS-nomenclature rules [5] describe almost any possible type of genomic variation including STRs. Based on the STR-variation analysed in this study we propose general rules for a straightforward forensic nomenclature that describes sequence-variation in STRs and flanking regions. We mostly follow HGVS-guidelines but some specific rules are optimised for the intended use in forensics.

Figure 1b shows an example of a hypothetical allele for marker D13S317.

**Figure 1,** determined genome-coordinates for STR-variation and example of allele-description for an STR sequence-variant of D13S317 according to nomenclature rules

| Locus | Range STR-structure in Reference-sequence (GRCh37/hg19) | CE-length | STR description refseq: |
|---|---|---|---|
| D1S1656 | chr1:g.230905351-230905426 | 16 | AC[6]CTAT[16] |
| TPOX | chr2:g.1493423-1493454 | 8 | TGAA[8] |
| D2S1338 | chr2:g.218879582-218879673 | 23 | GGAA[2]GGAC[1]GGAA[13]GGCA[7] |
| D2S441 | chr2:g.68239079-68239126 | 12 | TCTA[12] |
| D3S1358 | chr3:g.45582231-45582294 | 16 | TCTA[1]TCTG[1]TCTA[14] |
| FGA | chr4:g.155508886-155508973 | 22 | AAGG[3]AGAA[14]AGAG[1]AAAA[1]AAGA[3] |
| D5S818 | chr5:g.123111246-123111293 | 11 | CTCT[1]ATCT[11] |
| CSF1P0 | chr5:g.149455885-149455936 | 13 | CTAT[13] |
| D7S820 | chr7:g.83789540-83789591 | 13 | TCTA[13] |
| D8S1179 | chr8:g.125907107-125907158 | 13 | TCTA[1]TCTG[1]TCTA[11] |
| D10S1248 | chr10:g.131092508-131092559 | 13 | GGAA[13] |
| TH01 | chr11:g.2192316-2192343 | 7 | TGAA[7] |
| D12S391 | chr12:g.12449953-12450028 | 18 | TAGA[12]CAGA[7] |
| vWA | chr12:g.6093125-6093208 | 17 | GATG[2]GATA[1]GATG[1]GATA[11]GACA[5]GATA[1] |
| D13S317 | chr13:g.82722160-82722223 | 11 | TATC[11]AATC[2]ATCT[3] |
| PentaE | chr15:g.97374242-97374266 | 5 | TTTTC[5] |
| D16S539 | chr16:g.86386308-86386351 | 11 | GATA[11] |
| D18S51 | chr18:g.60948900-60948981 | 18 | AGAA[18]AA[1]AG[5] |
| D19S433 | chr19:g.30417141-30417204 | 14 | TCCT[13]ACCT[1]TCTT[1]TCCT[1] |
| D21S11 | chr21:g.20554291-20554419 | 29 | TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] |
| PentaD | chr21:g.45056086-45056155 | 13 | AAAGA[13]AAAAA[1] |
| D22S1045 | chr22:g.37536327-37536377 | 17 | ATT[14]ACT[1]ATT[2] |

**a) Genomic coordinates and variant-description of the reference for 22 STRs**

**b) Example of the reference and variant-sequence with description of the allele**

Reference-sequence (GRCh37:g.82.722.130-82.711.225)

tctaacgcctatctgtatttacaaatacat tatc tatc tatc tatc tatc tatc tatc tatc tatc tatc tatc aatc aatc atct atct atct tt

Variant-sequence:

TCTAACACCTATCTGTATTTACAAATACAT TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC AATC ATCT ATCT ATCT TT

Allele-description following suggested nomenclature:

D13S317-CE12–chr13-GRCh37-g.82.722.160:82.722.223–TATC[13]AATC[1]ATCT[3]–x.136G>A

Chapter 3

From left to right, allele-description contains the following 4 elements:

1. Locus-name followed by "CE" and the allele-length (described as in current CE-nomenclature).
2. Chromosome-coordinates of the STR-motif in the reference sequence (including reference genome-version), representing the position of the first, and the last base of the STR-motif in the reference genome. If coordinates of the total analysed fragment (range between the used primers) are known, this information should always be provided (as a separate table).
3. STR-motif, described in the same orientation as the reference genome. For example: TATC[13]AATC[1]ATCT[3] (TATC repeated 13 times followed by AATC repeated 1 time and ATCT repeated 3 times).
4. Variation outside the STR-region (but within the analysed fragment) is described relative to the reference by genome position. Variants are described in the following order: Genome coordinate, reference > variant. The long number of the genome coordinate can be shortened by a 'x.' followed by the last three numbers (since the total coordinates of the STR-motif are already described before).

- A SNP of G>A on position 82.722.136 can be described as x.136G>A
- For a deletion of GC on position 82.722.136-82.722.137 we only write the starting-position of the deletion followed by the variation, for example: x.136GC>del.
- An insertion of AT after the same G is described as x.136.1–>insAT ('–' before the '>' is used since this position is absent in the reference).

Although we understand the suggested use of rs-nr.'s by Gelardi et al [1] to maintain a stable allele-name over different genome-versions, this will still result in different ways of describing variants within the same table because there can always be SNPs that are not listed in dbSNP. Rs-nr.s do not provide all the information that is needed to directly translate an allele-name back to the original sequence since the exact position of the SNP will need to be retrieved from dbSNP. Comparison of results from different assays (using different primers) is complicated if it is not directly visible from the name whether a SNP is within the range of both assays or not. Thereby, we prefer the use of genome-coordinates over rs-nr's. However, it is essential that the version of the used reference genome is described.

In addition to the CE-fragment allele-name, our rules have some small deviations from the HGVS nomenclature. As in HGVS nomenclature, only the positions that differ from the reference are displayed, but they are described in the fixed order of ancestral > derived sequence. This is different from HVGS since deletions and insertions of one nt are described as delA (in our rules A>delA) and insA (in our rules x.1–>A).

This fixed order keeps the description directly translatable and is feasible for the use of computer-algorithms to automatically compare variants. To avoid accumulation of numbers, larger insertions and deletions are described in a more mtDNA-like fashion [3] displaying only the start-position of the variant followed by the ancestral > derived sequence. To combine all parts of the allele-name, an en-dash (long dash) was chosen as delimiter between the separate parts to leave an open space between the different parts and increase readability. Although we provide a method that can be used to describe variants without prior knowledge of the exact positions of the primers, we recognise that this is a suboptimal situation which limits possibilities in comparison of STR sequencing-results between different assays.

# Conclusion

Recommendations have been made for nomenclature of STRs in such a way to provide maximum information in the allele-name and help direct comparison of data from different assays and between CE- and sequencing-data.

# Acknowledgements

We thank Rick de Leeuw and Jerry Hoogenboom for their contribution in the discussion of the proposed nomenclature. This study was supported by a grant from the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands.

# Conflict of Interest Statement

The Authors declares that there is no conflict of interest.

# Reference List

1. Gelardi C, Rockenbauer E, Dalsgaard S, Borsting C, Morling N. Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. Forensic Sci.Int.Genet. 2014; 12:38-41
2. Gettings KB, Aponte RA, Vallone PM, Butler JM. STR allele sequence variation: Current knowledge and future issues. Forensic Sci.Int.Genet. 2015; 18:118-30
3. Parson W, Gusmao L, Hares DR, Irwin JA, Mayr WR, Morling N, Pokorak E, Prinz M, Salas A, Schneider PM, Parsons TJ. DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. Forensic Sci. Int.Genet. 2014; 13:134-142
4. Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res. 2001; 29:320-322
5. Taschner PE, den Dunnen JT. Describing structural changes by extending HGVS sequence variation nomenclature. Hum.Mutat. 2011; 32:507-511

Chapter 3

# Chapter 4

## Massively Parallel Sequencing of Short Tandem Repeats – Population data and mixture analysis results for the PowerSeq™ system

Forensic Science International Genetics 2016 Sep;24:86-96
Published online June 07, 2016

Kristiaan J. van der Gaag
Rick H. de Leeuw
Jerry Hoogenboom
Jaynish Patel
Douglas R. Storts
Jeroen F.J. Laros
Peter de Knijff

Supplementary material is available via https://www.fsigenetics.com

# Abstract

Current forensic DNA analysis predominantly involves identification of human donors by analysis of short tandem repeats (STRs) using Capillary Electrophoresis (CE). Recent developments in Massively Parallel Sequencing (MPS) technologies offer new possibilities in analysis of STRs since they might overcome some of the limitations of CE analysis. In this study 17 STRs and Amelogenin were sequenced in high coverage using a prototype version of the Promega PowerSeq™ system for 297 population samples from the Netherlands, Nepal, Bhutan and Central African Pygmies. In addition, 45 two-person mixtures with different minor contributions down to 1% were analysed to investigate the performance of this system for mixed samples. Regarding fragment length, complete concordance between the MPS and CE-based data was found, marking the reliability of MPS PowerSeq™ system. As expected, MPS presented a broader allele range and higher power of discrimination and exclusion rate. The high coverage sequencing data were used to determine stutter characteristics for all loci and stutter ratios were compared to CE data. The separation of alleles with the same length but exhibiting different stutter ratios lowers the overall variation in stutter ratio and helps in differentiation of stutters from genuine alleles in mixed samples. All alleles of the minor contributors were detected in the sequence reads even for the 1% contributions, but analysis of mixtures below 5% without prior information of the mixture ratio is complicated by PCR and sequencing artefacts.

# Introduction

Current forensic DNA analysis almost exclusively focuses on the identification of human sample donors using multiplex short tandem repeat (STR) genotyping with commercial kits based on polymerase chain reaction (PCR) and capillary electrophoresis (CE). Although this type of analysis has proven its value over the past decades, it is not without limitations. In CE, multiplexing of more than 5 loci in a single assay can only be achieved by using different fluorescent labels in the PCR and by using non-overlapping PCR fragment lengths for STRs with the same fluorescent label. Consequently, most commercial assays have a PCR fragment range between 80-500 bp [20].

When analysing degraded DNA samples, this variation in fragment length frequently results in noticeable lower, or even absent, signals for the longer PCR fragments. As a consequence, profiles of degraded DNA often have a lower discriminating power.

Another potential difficulty associated with the CE detection of STRs is the background signal arising from stutter peaks [19], caused by slippage of the polymerase in the PCR. In DNA samples from a single person, genuine alleles and stutter alleles can be easily distinguished. However, the analysis of unbalanced mixtures with low minor contributions is frequently complicated by stutter alleles that cannot be distinguished

Chapter 4

from genuine alleles of the minor contributors [4].

In theory, these limitations can mostly be solved by the use of massively parallel sequencing (MPS) of STR loci. STR alleles can be identified by repeat number and sequence variation and primers can be designed in such a way that PCR fragments have similar size ranges for all loci. Moreover, many more loci can be multiplexed in the same reaction because the detection is no longer based on a limited number of fluorescent labels. A few studies have indicated the potential of MPS STR genotyping [6, 8, 15, 21]. They showed that, in addition to the variation in repeat number and repeat sequence, the repeat-flanking regions provide an additional source of variation and add to the discriminating power of the loci. However, the additional power of this new sequence variation cannot be fully used until sufficient population frequency data is available for all loci. We speculated that this additional information could help in distinguishing genuine alleles from stutter alleles although it is not likely that this problem will be completely overcome.

For this purpose, we assessed population data for 297 samples of three distinct populations (Dutch, Himalayan, and Central African Pygmies) for 17 STR loci included in a prototype version of the PowerSeq™ MPS STR assay [21]. These data were compared to the results of CE-based data from the PowerPlex® Fusion System [12]. We also present data from several series of mixed DNA samples in different ratios down to 1:99 to survey the possibilities and limits for this assay in analysis of mixed samples.

We examined the additional sequence variation of the loci, both within the STR motifs and in the flanking regions, and assessed the impact of this variation on the discriminating power of the loci. In addition, stutter ratios were studied and compared to those obtained with CE-based profiling.

As a consequence of various mechanisms such as DNA recombination, replication and repair-associated processes, the spectrum of human genetic variation ranges from single nucleotide differences to large chromosomal events. Among the different types of genetic changes, repetitive DNA sequences show more polymorphism than single nucleotide variants (Conrad et al., 2010; Hinds et al., 2006; Iafrate et al., 2004; Kidd et al., 2008; Redon et al., 2006; Sebat et al., 2004; Tuzun et al., 2005), and they are important in human diseases (Conrad et al., 2010; de Cid et al., 2009; Girirajan et al., 2011; Hollox et al., 2008; McCarroll et al., 2009; Pinto et al., 2010), complex traits and evolution (Mills et al., 2011; Stephens et al., 2011; Sudmant et al., 2010). In particular, microsatellite variants, also known as short tandem repeats (STR), and their expansion/shortening have been linked to a variety of human genetic disorders (Mirkin, 2007; Pearson et al., 2005; Sutherland and Richards, 1995), and have been used in genotyping (Kimura et al., 2009; Weber and May, 1989) and forensic DNA fingerprinting studies (Kayser and de Knijff, 2011; Moretti et al., 2001).

Because of the repetitive nature of STRs and often the low level of complexity of the DNA sequences in which they occur (Treangen and Salzberg, 2012), characterization of STR variability and understanding of their functional consequences are challenging (Weischenfeldt et al., 2013). So far, sequencing-based strategies have focused on reads mapped to the reference genome and subsequent identification of discordant signatures and classification of associated STRs (Medvedev et al., 2009; Mills et al., 2011). Yet, the mainstream aligners, such as BWA (Li and Durbin, 2009) or Bowtie (Langmead and Salzberg, 2012), do not tolerate repeats or insertions and deletions (indels) as a trade-off of run time (Li and Homer, 2010). This limitation leads to ambiguities in the alignment or assembly of repeats which, in turn, can obscure the interpretation of results (Treangen and Salzberg, 2012). Moreover, the current human genome reference still remains incomplete and provides only limited information on expected and potentially uncharacterized STRs in different individuals (Alkan et al., 2011; Iafrate et al., 2004; Kidd et al., 2008; Sebat et al., 2004). Consequently, STRs are not routinely analyzed in whole-genome or whole-exome sequencing studies, despite their obvious applications and their role in human diseases, complex traits and evolution.

Here, we present a method for targeted profiling of STRs that reports a full spectrum of all observed genomic variants along with their respective abundance. Our tool, TSSV, can accurately profile and characterize STRs without the use of a complete reference genome, and therefore minimizes biases introduced during the alignment and downstream analysis. TSSV scans sequencing data for reads that fully or partially encompass loci of interest based on the detection of unique flanking sequences. Subsequently, TSSV characterizes the sequence between a pair of non-repetitive flanking regions and reports statistics on known and novel alleles for each locus of interest. We show the performance of TSSV on robust characterization of all allelic variants in a given targeted locus by its application in several case studies: forensic DNA fingerprinting of mixed samples by STR profiling, characterization of variants introduced by transcription activator-like effector nucleases (TALENs) in embryonic stem (ES) cells and detailed characterization of errors derived from a next-generation sequencing (NGS) experiment.

# Material and Methods

## Population samples

To assess the potential genetic variation, 297 DNA samples were selected from a European population (101 Dutch samples [20]), an Asian population (97 samples from Nepal and Bhutan [10]) and an African population (99 Central African Pygmy samples [9]).

## Capillary electrophoresis

PCR reactions were performed according to the protocol of the PowerPlex® Fusion System [14] using 0.5 ng of DNA and 30 amplification cycles using a GeneAmp® PCR System 9700 (Life Technologies). For every reaction, 2800M Control DNA (Promega) was included as a positive control and a water sample was included as negative control sample. CE was performed using an AB3500XL (Life Technologies) according to the PowerPlex® Fusion System protocol, data was analysed using GeneMarker® software v2.4.0 (Softgenetics).

## Massively Parallel Sequencing

PCR reactions were performed with a prototype PowerSeq™ sequencing assay primer mix and master mix (Promega) amplifying 17 STR loci and Amelogenin. All PCRs were performed on a GeneAmp® PCR System 9700 using the following program: 96 °C for 1 min, 30 cycles of 94 °C for 10s, 59 °C for 1 min, 72 °C for 30s and a final extension of 60 °C for 10 min, for every reaction 2800M Control DNA was included as a positive control and a water sample was included as negative control sample.

Illumina sequencing libraries were prepared from the PCR products by ligating barcoded adapters using the KAPA Library Preparation kit (KAPA Biosystems) without additional amplification using 2.5 µl of PCR product directly in the end repair reaction (without prior purification) in a total volume of 35 µl. The A-tailing and ligation step were performed in a total volume of 25 µl. For ligation, a 10-fold dilution of a barcoded TruSeq adapter (Illumina) was used. To confirm successful ligation of the adapters, 1 µl of library was analysed on the Qiaxcel (Qiagen) for a selection of libraries. To enable balanced pooling, sequencing libraries were quantified in duplicate by real time PCR using the KAPA SYBR® FAST qPCR kit. Quantification reactions were performed on a LightCycler® 480 (Roche) or a 7500 Real Time PCR System (Life Technologies) using a dilution series of PhiX control library (Illumina) as standard. After pooling the libraries, the final pool was quantified again using the same method to enable optimal loading of the flow cell. Sequencing was performed on the MiSeq® sequencer (Illumina) using v3 sequencing reagents according to the manufacturer's protocol with approximately 5% of PhiX control library and 14-19 pM final library concentration.

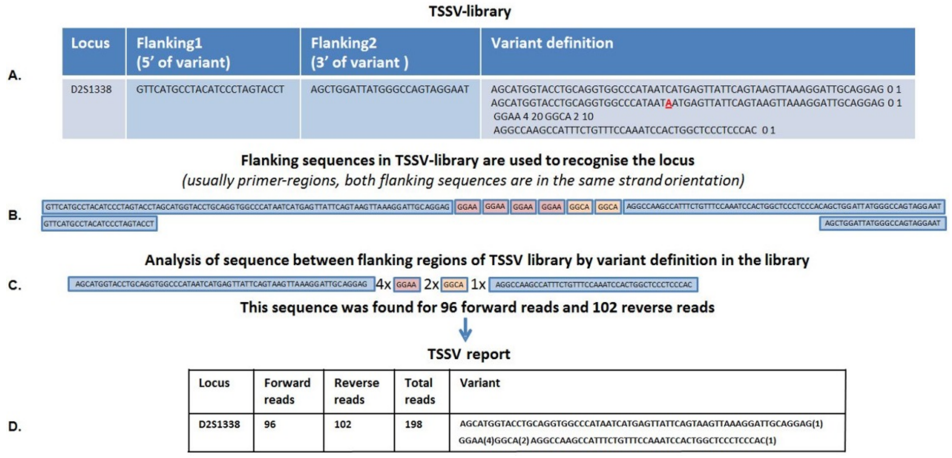
## Data analysis

For the analysis of STR sequences, the use of simple alignment-based methods could lead to errors. In the analysis pipeline, the first step is the alignment of both paired-end reads that are generated by the sequencer to obtain one high quality consensus read. We used the paired-end read aligner FLASH [11] that aims for a maximum overlap of both reads when creating one consensus read (matching any two

paired reads with a mismatch ratio of under 0.33 in the overlapping part). If both reads end within a repeated element, the alignment could lead to a shortened repeated element in the consensus read. To be able to recognise possible misalignment of the reads we altered FLASH version 1.2.11 (this altered version is available via https://github.com/Jerrythafast/FLASH-lowercase-overhang). We added an option to mark the bases that were not overlapped by both reads in small letters in the consensus read. Hereby, when all the bases of the flanking regions are in small letters (and thus the sequence reads ended within the repeated element), they can be filtered out in later analysis. When a difference occurred between the two reads, the base call with the highest quality value was used for the consensus. Analysis of the paired-end consensus reads was performed using TSSV [2] (install using: pip install tssv). A TSSV library was created based on all observed variants (Sup. File 1). In Figure 1, the analysis of STRs using TSSV is illustrated. To further support the interpretation of STR sequencing data, we developed Stuttermark (part of the Python package fdstools, for installation use: pip install fdstools); a Python script that marks possible stutter alleles based on the sequence structure. With this software a column is added to the table of 'known alleles' from TSSV where alleles that could be derived from an n-1, n-2 or n+1 stutter of an allele (based on the complete allele sequence) in the sample are marked. Thresholds for n-1 and n+1 stutter ratios (n-2 is considered as an n-1-1 using a squared value of the n-1 threshold) are used to decide whether a sequence is marked as an allele or as a possible stutter. A shell script (available upon request) was written to automate all the analysis steps in parallel on a computer cluster for large sample series. An Excel sheet (available upon request) was subsequently used to summarise the results and score variants according to a priori defined criteria for the number of reads per variant (total and per orientation) and a minimum percentage from the reads of the highest allele for every locus.

Chapter 4

**Figure 1.** An overview of the TSSV analysis strategy of short tandem repeat sequences




*A. An example of the TSSV library entry for locus D2S1338 with from left to right the locus name, flanking 1, flanking 2 (in the same orientation as flanking 1) and the variant definition. Both flanking sequences usually represent the PCR primers. The numbers at the ends of the variant definition sequences (in this example "0 1", "0 1", "4 20", "2 10", and "0 1") indicate how often (based on current knowledge) a sequence could be repeated.*

*B. Both flanking sequences of the library are used to recognise which locus (in both orientations) any read represents. The observed sequence variation between the two flanking sequences will be reported by TSSV. In this example, some of the surrounding sequence of the STR is included to not only report the STR variation, but also the sequence variation in the surrounding region of the STR.*

*C. The sequence between the flanking regions is compared to the variant definition of the library. A sequence that complies with the variant definition is reported and summarised (by counting the separate repeated motifs) in the 'known alleles' table and a sequence that doesn't comply with the variant definition is reported in the 'new alleles' table.*

*D. A TSSV report summarising the displayed allele which was observed 96 times in the forward orientation and 102 times in the reverse orientation. The variant starts with AGCATGG... (not repeated), followed by GGAA (repeated 4 times), GGCA (repeated 2 times) and AGGCCAA... (not repeated). In addition to the tables, fasta files are generated containing the complete sequence reads for the known and new alleles at each locus, but also for the reads that are not recognised or in which only one of the flanking sequences of a locus is recognised. In this way, it is possible to keep track of the sequences that are not reported.*

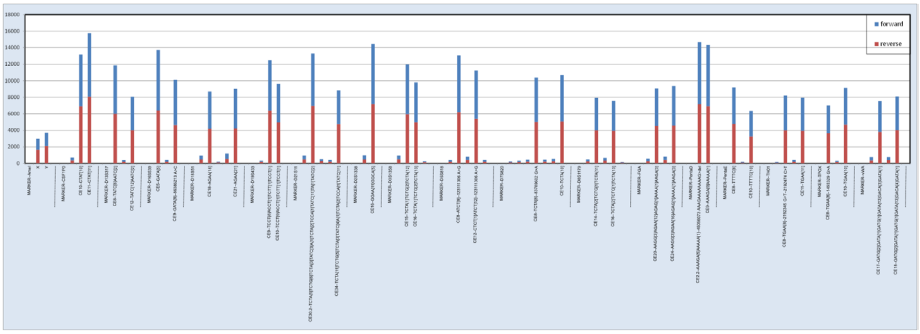
## Analysis of single source samples

In every sequencing run for the population samples (7 runs in total), a maximum of 48 barcoded samples were sequenced aiming for a coverage of at least 1000 reads for every STR allele in each sample. After measuring concentrations of the sequencing libraries, all samples of a run were pooled in an equimolar fashion prior to sequencing. The output of TSSV was analysed with Stuttermark using two different threshold settings; first, n-1 position stutters with ratios below 10% of the genuine allele and n+1 position stutters below 2% of the genuine allele were marked while in the second analysis thresholds of respectively 20% and 3% were used. As a final step, a sequence read profile (see Figure 2) was generated showing all the alleles that have met defined thresholds for read coverage (further described in the Results section). In the sequence read profile, allele names for alleles marked as stutter for both settings of Stuttermark are automatically removed. As with CE analysis, remaining alleles with an assigned

allele name were inspected by a trained expert and alleles interpreted as stutter were removed. In this article, allele names are described according to the nomenclature described by van der Gaag and de Knijff [17]. In all figures, locus coordinates were removed to shorten the allele name.

**Figure 2.** An example of a PowerSeq™ MPS read profile and read statistics for all 18 loci in a single-source sample

**A.** An STR sequence read profile




**B.** Sample read statistics

| Read-category | Read-counts | Proportion of total reads |
|---|---|---|
| Total passed filter reads | 537665 | 100,0% |
| Matched pairs | 510409 | 94,9% |
| Known alleles (including stutters) | 406437 | 75,6% |
| Genuine alleles (excluding stutter) | 350294 | 65,2% |
| Reads with errors in the variant region (new alleles in TSSV analysis) | 103972 | 19,3% |
| *(Singletons)* | *(27973)* | *5,2%* |
| Reads representing stutters | 56143 | 10,4% |
| Primer dimers | 27256 | 5,1% |

*A. An MPS-STR sequence read profile showing all observed alleles of a single-source reference sample with the corresponding number of forward reads (blue bars) and reverse reads (red bars) for every allele. Only the observed variants with coverage of at least 5 reads and a within locus proportion of 2% of the highest allele are displayed in this profile. B. Read statistics of the displayed sample, all percentages are displayed as a proportion of the total passed filter reads. 94.9% of the reads of this sample were recognised for both flanking sequences (matched pairs) of a locus using TSSV. 75.6% of the total reads represented known alleles and after removing the stutter reads, 65.2% of the reads represent the genuine alleles of this sample. From the 19.3% of matched pairs that were marked as new alleles by TSSV, a large proportion (5.2% of the total reads) consisted of singletons. The remaining 5.1% of passed filter reads (not recognised as matched pairs) represented primer dimers.*

Chapter 4

### Analysis of mixed samples

For five two-person combinations selected from the Dutch population samples, mixtures were prepared in the ratios 1:99, 5:95, 10:90, 20:80, 50:50, 80:20, 90:10, 95:5 and 99:1 by mixing the samples based on triplicate DNA quantifications acquired using the Quantifiler® Duo DNA Quantification Kit (Life Technologies).

In the PCR reaction for STR amplification, DNA input amounts were adjusted to add at least 60 pg (≈ 10 cells) of the minor contributor. To achieve this, total DNA input varied from 0.5 ng – 6 ng. Samples were sequenced in two runs and pooling ratios were calculated to achieve a minimum of 20 reads for every allele of the minor contributor in each mixture. Analysis was performed in the same way as for the single source samples, but the threshold for the percentage of reads from the highest allele of a marker was lowered depending on the mixture ratio (further discussed in results and discussion). Based on the sequence variation and allele ratios, suspected stutter peaks were marked by an expert to distinguish genuine alleles from stutter peaks.

## Analysis of stutter ratios

### Stutter analysis of CE data

The sized output trace data (containing fluorescence intensity data for every position in the electropherogram) was exported from GeneMarker® to Excel. Using peak heights, the stutter ratios at n-1, n+1 and n-2 stutter positions, were determined for every allele. Peaks that may represent overlapping stutter events (e.g. stutters in between two genuine alleles that may represent both an n-1 and an n+1 stutter) were removed. Sup. Figure 1 illustrates which combinations of stutter peaks and alleles were used for analysis. Peaks with intensities below 30 rfu were discarded in order to avoid miscalled CE artefacts and to minimise the influence of run-to-run variation of the Genetic Analyser. For some loci, a large proportion of the peaks on stutter positions were lower than 30 rfu (because of the low stutter ratio and the limit in detection range), these peaks did not necessarily represent a zero stutter ratio and were therefore considered to miss a stutter value to avoid underestimation of the stutter ratio (resulting in a slight overestimation of low ratio stutter peaks).

### Stutter analysis of STR sequencing data

Stutter analysis was performed for all samples for which we obtained more than 50.000 total reads (271 out of 297 samples) to avoid bias introduced by low coverage alleles. To check for possible differences in coverage between long and short alleles, the within locus allele balance was calculated for every marker. For the stutter analysis, sequence variants with coverage below 5 reads were discarded to minimise bias in the

stutter ratio. For every observed sequence allele, a table was generated with 6 possible stutter sequences; the two most likely stutter sequences for the n-1 stutter reads, the n+1 stutter reads and the n-2 stutter reads. The most likely stutter sequences were determined based on the length of the longest repeating element in the sequence assuming that longer repeats produce the most stutter [3]. For these 6 stutter alleles, the stutter percentage was determined by dividing the read count of the stutter allele by the read count of the genuine allele. Stutter alleles that could overlap with other alleles or stutter reads were removed taking sequence-specific differences into account as illustrated in Sup. Figure 1.

### Statistical Calculations

For all STRs in the assay, the match likelihood and power of exclusion were calculated for the alleles observed in CE and MPS for all three populations using the Powerstat excel spreadsheet [13].

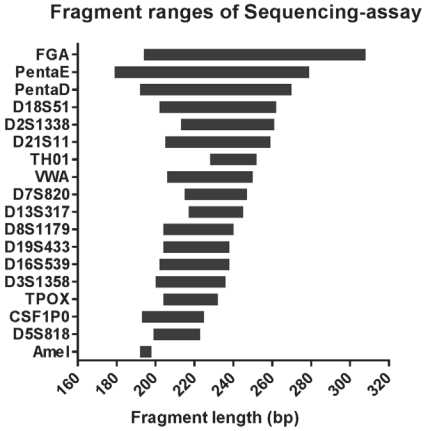
## Results and discussion

To assess sequence variation in STR loci and stutter characteristics of a prototype MPS STR sequencing assay (PowerSeq™), 297 samples from three globally dispersed populations were sequenced. To avoid the influence of possible somatic cell line mutations on the analysis of stutter characteristics, we preferred to use DNA samples derived from blood over the use of cell line material from worldwide panels like HapMap or the Human Genome Diversity Panel [1, 5].

In the PowerSeq™ assay, all PCRs are designed to amplify STR fragments which are around the same fragment length (shortest to longest allele: 180-310 bp, 180-280 bp excluding the exceptionally long FGA-alleles). Figure 3 displays the fragment length distribution of the sequenced alleles in this study for all 17 STRs and Amelogenin.

Chapter 4

**Figure 3.** Overview of fragment range for all loci in the prototype PowerSeq™ assay

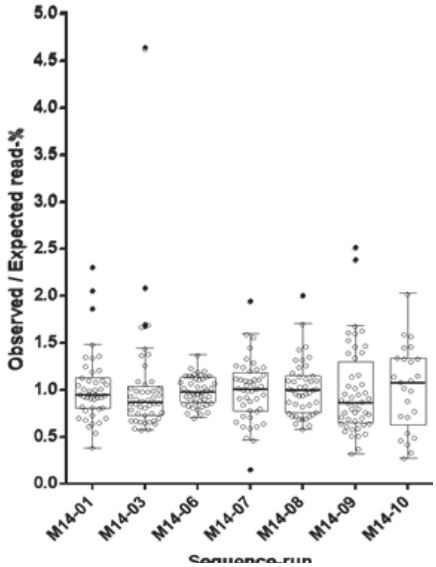



*The prototype MPS PowerSeq™ multiplex assay used in this study contains 17 autosomal STR loci and Amelogenin. This figure shows the PCR fragment size variation of all alleles sequenced in this study.*

## Optimisation

Reliable quantification of the sequence libraries is an important step for optimal sequencing. It is used to achieve optimal balance for pooling different libraries in a run and it influences the number of molecules that are loaded on the sequencer. To assess whether equimolar pooling was achieved, the observed and expected proportion of sequences were compared for all samples in the 7 sequencing runs comprising the 297 population samples (Figure 4). The majority of libraries are represented in 0.5-2 times the expected proportion of reads in the sequencing run, which is sufficiently balanced for the current design. Thus, the quantitation method that was used (real time PCR) allows effective library pooling. Different loading concentrations were used on the MiSeq® sequencer to determine optimal cluster density on the flow cell (higher loading concentrations result in higher cluster densities). Higher cluster density results in a higher amount of unfiltered reads but decreases sequence quality (Sup. Figure 2). We infer that a flow cell cluster density around 800-1000 K/mm2 may be most optimal (further discussed in the section 'filtering noise from alleles').

**Figure 4.** Tukey boxplot of the ratio of observed versus expected read proportion of pooled samples over different sequencing runs




*Tukey boxplot showing the ratio of observed versus expected read proportion of 297 pooled samples analysed in 7 sequencing runs. The box displays the interquartile range (IQR), the line in the box displays the median and the whiskers display the range until the last sample within 1.5 IQR.*

An example of a read profile is shown in Figure 2A. The sequence profile resembles a CE profile with the y-axis displaying the number of reads observed for every sequence variant, the labels on the x-axis display a more detailed description of the sequence for every allele. Note that the range of amplicon sizes is similar for all STRs (Figure 3) even though the loci are displayed next to each other on the x-axis. The number of reads is directly proportional to the number of actual molecules for every allele, which is distinct from CE profiles where peak height is influenced by the intensity of emission for different fluorescent labels.

## Sequence efficiency

In Figure 2, we display the statistics of read counts and the sequencing profile for a typical sample which is prepared using the recommended input of 0.5 ng DNA in the PCR reaction for this assay. 65% of the reads represented the genuine allele sequences of the alleles, approximately 5% of the reads were occupied by stutter reads, the remaining 25% of recognised reads consisted of reads containing PCR and / or sequencing errors. The 5% of unrecognised reads consisted mostly of primer-dimers which is a well-known side effect when large multiplexes such as this 18-plex are used. Remaining primer-dimers could be minimised by purification steps involving size selection such as using a low bead-to-volume ratio for AMPure XP beads. However,

Chapter 4

we chose to use the PCR product without purification before the library preparation and we used a 2:1 bead ratio in the purification steps of the library preparation to avoid size selection which may affect the balance in sequence reads between longer and shorter STR alleles.

## Filtering noise from alleles

In order to be accepted as a reliable forensic diagnostic tool, MPS results should be retrieved and stored in much more detail compared to CE data. Processing of millions of reads involves complex bioinformatics. It is for this purpose that the tools we developed to analyse MPS reads not only report genuine alleles but also facilitate storing and screening those reads that do not represent genuine STR alleles. Detailed tables of read statistics are produced and checked before allele interpretation. These tables contain read counts for new alleles and for alleles that are only recognised for either one or none of the flanking sequences of the TSSV library. In case of high read numbers for these categories, fasta files containing the complete sequences of the reads can be checked for every locus and for each category (known alleles, new alleles, reads with only the start flanking sequence recognised, reads with only the end-flanking sequence recognised and reads with no recognised flanking sequences at all) separately.

The frequency of sequencing errors varies per locus, but is also strongly influenced by the cluster density in the sequence run. A good indicator for sequence quality of a sequencing run is the balance between forward and reverse reads. Since read errors tend to be influenced by sequence content, the same error will usually not appear in both orientations [16]. For the longest alleles from PentaD, PentaE and FGA we noted that sequencing errors may accumulate in the end of the reads. As a consequence, the flanking sequences for that strand may no longer be recognised by TSSV, which could lead to strand bias of over five-fold differences between both orientations, even when analysing paired-end consensus reads. Thus, one should not straightforwardly aim for a high cluster density to retain the highest number of reads, as this may be accompanied with strong strand bias. We observed increased rates of sequence errors and strand bias for cluster densities over 1000 K/mm$^2$ which is below the recommended cluster density of 1200-1500 K/mm$^2$. When a cluster density of 800 K/mm$^2$ is used, at least 1.5x10$^7$ Passed Filter reads are retained (all sequenced for both read orientations) which is a sufficient read number to multiplex an effective number of libraries.

Quality filtering of the data was done in the following order:

1. Paired-end consensus alignment: the two paired-end reads of each cluster are combined. In case of discrepancies, the highest quality base call is used in the consensus read for further analysis. Parts of the read that are not overlapped

by both reads are marked in lower case (reads that have one of the library flanking sequences completely represented by lower case letters are later on moved to the TSSV category of reads recognised for only one of the flanking sequences).

2. Singletons are discarded during analysis using TSSV (TSSV option: '-a 2'). These reads can only be checked afterwards by restarting the analysis without this option. Discarding singletons significantly decreases the report file size and memory demand in the follow up analyses. Singletons will not meet forensic standards, but could be used to decide whether sequence coverage needs to be increased for a low coverage sample. New alleles (that do not match the variant description of the TSSV library) are reported in a separate table.
3. After performing TSSV analysis, the table of known alleles is filtered by a priori defined criteria in an Excel sheet while ensuring that the sequences, which are filtered out in these steps, can easily be retrieved and investigated. We used a minimum of 8 reads as allele coverage and a minimum of 2 reads for both sequence orientations which removed the majority of sequencing errors. These numbers may seem low, but it should be noted that we use 'allele coverage' (only including reads without errors) and not 'total coverage' (which would mean the sum of all reads for one locus and could include reads with errors). Since forensic samples often carry allele imbalance due to low amounts of template or multiple contributors to a sample, the use of total summed coverage of all alleles for a target can give a misleading sense of quality and should be avoided. The threshold of 2 reads for both sequence orientations is sufficient to remove the majority of sequence artefacts. A higher threshold could result in the loss of some (mostly longer) alleles that exhibit a strong strand-bias due to structural sequence errors. Retained alleles were interpreted before being reported.
4. In the same Excel sheet an additional criterion is a within-locus proportion (the read count of an allele divided by the read count of the highest allele of a locus) that is required for reporting an allele. This threshold is used to remove PCR errors and structural sequencing errors that may especially occur at high coverage. This value can be adjusted depending on the required detection of low percentage contributions. When the input amount of DNA in the PCR is available, it can also be used to filter out unrealistic mixture contributions (for example: for a start amount of 60 pg in the PCR it is not realistic to look for a 1% contribution since this would represent the DNA equivalent of only 0.1 cell). For single reference samples we used a threshold of 15%, for mixtures, this threshold was lowered to 1% except for mixtures with a minor contribution of 1% in which this threshold was lowered to 0.25%. All retained alleles appear as a bar in the STR sequence profile (Figure 2).
5. Allele variants that are not represented in the TSSV library are added to the

Chapter 4

table of new alleles. This table is filtered using the same settings as used for the known alleles. When a new allele is identified as a genuine allele, it is added to the TSSV library and samples are reprocessed using the new TSSV library which will move it to the known alleles category.

6. Stuttermark is used to mark alleles that could be (partly) derived from stutter (as described in the Materials and methods section). When interpreting the alleles that pass the filtering steps mentioned before, alleles at a stutter position of another allele (based on the sequence) and with less reads than an a priori defined percentage of the reads of a genuine allele are marked as stutter.
7. Interpretation of the retained alleles is done by inspection of the markings from Stuttermark in combination with the ratio between the retained alleles and the strand balance for every allele. In this step, the label of the alleles that are marked as stutter (or any other artefact) will be removed from the STR sequence profile. However, in the sequence profile, the bar representing the removed allele will remain without a label as is common practice for CE-based profiles.

Sup. Figure 3 shows examples of STR sequencing profiles for a single and a mixed source sample after different filtering settings to illustrate the effect of the used parameters.

## Concordancy

Reliability of sequencing results was assessed for the 297 population samples by comparison of CE data from the PowerPlex® Fusion System with the sequencing data. All STR alleles from the sequencing data were in concordance with CE analysis except for two alleles from PentaD. These alleles were missed when using the 15% within locus threshold (heterozygote balance), as they had a frequency of 8% and 12% of the highest allele (Sup. Figure 4). Since both samples are from the same population, and both alleles have the same repeat length and sequence, it is likely that this difference in read numbers is caused by a SNP under the PCR primer used in the PowerSeq™ sequencing assay as observed for rare null alleles in commercial CE-based assays [20].

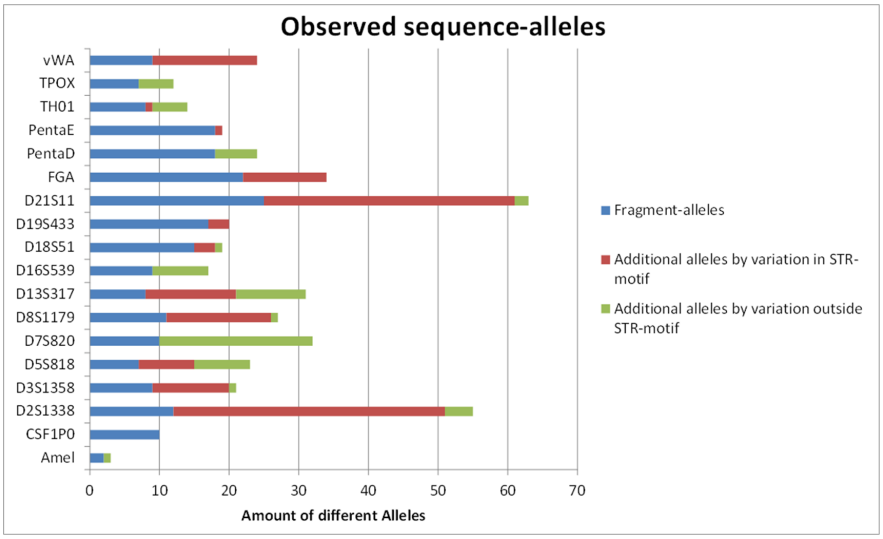
## Sequence variation

As was expected, MPS STR genotyping revealed substantial genetic variation in addition to the variation in repeat length that is detected using CE (Figure 5). Sup. Figure 5 displays the sequence of the genome reference (GRCh37/hg19) and of control sample 2800M (which is provided with the assay). Sup. Figure 6 displays the observed alleles for all loci and the frequencies of these alleles in the three tested populations. Since we describe our variants according to nomenclature rules [17] in which all variants are described in the forward orientation of the genome reference,

the start position and orientation of some of the alleles is slightly different than the reference alleles described by Gettings et al. [7]. Based on the observed variation in this study, the analysed STRs can be divided into four classes.

1. Simple STRs: Loci that only show variation in the number of repeats without additional sequence variation. CSF1P0 is the only simple STR locus.
2. Complex STRs: Loci where the repeat motif consists of several repeating blocks with a different sequence. D19S433, FGA and PentaE are complex STRs.
3. Simple STRs with SNPs in the flanking sequence of the repeat region. D7S820, D16S539, TPOX and PentaD are simple STRs with SNPs.
4. Complex STRs containing SNPs in the flanking sequence of the repeat region: D2S1338, D3S1358, D5S818, D8S1179, D13S317, D18S51, D21S11, TH01 and vWA (interestingly, for vWA, all SNPs are associated with specific repeat region variation) are complex STRs containing SNPs in the flanking sequence.

**Figure 5.** STR sequence variation divided in length variation, complex STR variation and SNP variation




*The stacked bar graph displays the number of different alleles observed in sequence analysis of 297 samples divided in three categories: In blue, the number of alleles observed when performing CE. In red, the additional alleles observed by sequencing when taking into account variation within the STR motif. In green, the additional alleles when taking into account variation flanking the STR motif. When the variation flanking the STR motif is linked with variation inside the STR motif, the green portion of the bar graph doesn't display those alleles (they are included in the red portion of the bar graph).*

Chapter 4

**Table 1.** Locus statistics for CE and MPS analysis of the same samples from three populations

| | Total Alleles | | Heterozygous % | | Match Likelihood | | | | | | Power of Exclusion | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Netherlands | | Nepal + Buthan | | Biaka Pygmees | | Netherlands | | Nepal + Buthan | | Biaka Pygmees | |
| Marker | Fragment-length | Sequence-variation | Fragment-length | Sequence-variation | Fragment-length | Sequence-variation | Fragment-length | Sequence-variation | Fragment-length | Sequence-variation | Fragment-length | Sequence-variation | Fragment-length | Sequence-variation | Fragment-length | Sequence-variation |
| CSF1PO | 10 | 10 | 75.8% | 75.8% | 0.12 | 0.12 | 0.12 | 0.12 | 0.15 | 0.15 | 0.57 | 0.57 | 0.45 | 0.45 | 0.56 | 0.56 |
| D2S1338 | 12 | 55 | 83.6% | 90.6% | 0.04 | 0.03 | 0.04 | 0.03 | 0.04 | 0.01 | 0.82 | 0.82 | 0.57 | 0.69 | 0.61 | 0.92 |
| D3S1358 | 9 | 21 | 73.8% | 87.2% | 0.07 | 0.04 | 0.11 | 0.06 | 0.14 | 0.05 | 0.48 | 0.66 | 0.50 | 0.75 | 0.49 | 0.81 |
| D5S818 | 7 | 23 | 72.5% | 84.9% | 0.16 | 0.04 | 0.09 | 0.05 | 0.13 | 0.02 | 0.54 | 0.80 | 0.43 | 0.51 | 0.44 | 0.77 |
| D7S820 | 10 | 32 | 81.9% | 87.9% | 0.07 | 0.03 | 0.08 | 0.03 | 0.09 | 0.03 | 0.64 | 0.70 | 0.48 | 0.63 | 0.79 | 0.94 |
| D8S1179 | 11 | 27 | 80.2% | 86.9% | 0.07 | 0.04 | 0.06 | 0.03 | 0.09 | 0.03 | 0.61 | 0.76 | 0.65 | 0.75 | 0.56 | 0.69 |
| D13S317 | 8 | 31 | 72.5% | 85.6% | 0.09 | 0.03 | 0.07 | 0.03 | 0.17 | 0.05 | 0.57 | 0.74 | 0.55 | 0.69 | 0.31 | 0.69 |
| D16S539 | 9 | 17 | 75.5% | 81.2% | 0.10 | 0.07 | 0.09 | 0.05 | 0.08 | 0.03 | 0.48 | 0.55 | 0.50 | 0.55 | 0.58 | 0.77 |
| D18S51 | 15 | 19 | 83.9% | 85.2% | 0.04 | 0.04 | 0.04 | 0.04 | 0.05 | 0.04 | 0.70 | 0.72 | 0.69 | 0.69 | 0.63 | 0.69 |
| D19S433 | 17 | 20 | 83.6% | 83.6% | 0.09 | 0.08 | 0.06 | 0.06 | 0.03 | 0.03 | 0.57 | 0.57 | 0.67 | 0.67 | 0.77 | 0.77 |
| D21S11 | 25 | 63 | 84.2% | 88.3% | 0.05 | 0.03 | 0.06 | 0.02 | 0.04 | 0.02 | 0.70 | 0.78 | 0.69 | 0.79 | 0.65 | 0.71 |
| FGA | 23 | 35 | 86.2% | 86.2% | 0.05 | 0.05 | 0.03 | 0.03 | 0.04 | 0.04 | 0.82 | 0.82 | 0.75 | 0.75 | 0.60 | 0.60 |
| PentaD | 18 | 24 | 83.2% | 84.2% | 0.06 | 0.05 | 0.06 | 0.06 | 0.04 | 0.04 | 0.62 | 0.66 | 0.57 | 0.59 | 0.79 | 0.79 |
| PentaE | 18 | 19 | 85.2% | 85.2% | 0.03 | 0.03 | 0.03 | 0.03 | 0.04 | 0.04 | 0.66 | 0.66 | 0.75 | 0.75 | 0.69 | 0.69 |
| TH01 | 8 | 14 | 70.8% | 72.5% | 0.10 | 0.10 | 0.16 | 0.16 | 0.13 | 0.08 | 0.61 | 0.61 | 0.33 | 0.33 | 0.41 | 0.49 |
| TPOX | 7 | 12 | 65.4% | 71.1% | 0.20 | 0.20 | 0.21 | 0.19 | 0.13 | 0.05 | 0.38 | 0.38 | 0.31 | 0.33 | 0.39 | 0.67 |
| vWA | 9 | 24 | 78.5% | 83.6% | 0.07 | 0.05 | 0.08 | 0.07 | 0.06 | 0.02 | 0.62 | 0.70 | 0.46 | 0.50 | 0.63 | 0.81 |
| Amel | 2 | 3 | Overall | | 5.3E-20 | 8.6E-23 | 2.2E-20 | 4.6E-23 | 4.1E-20 | 5.2E-25 | | | | | | |

*Heterozygosity, Match Likelihood and Power of Exclusion for STR CE and sequence analysis for all 17 STRs as observed in the three tested populations.*

Using CE, uniquely identified alleles comprise only 48% of the total alleles observed using sequencing in these 17 STRs for the analysed set of 297 samples. However, the variation is not evenly dispersed over the loci (Table 1). Since not every available software tool for analysis of STRs capture the variation within the repeat structure and the flanking sequence [18] it is important to be aware of the information that is missed when variation outside the repeat structure is not reported. Obviously, the discriminating power of the loci is increased when all the variation on sequence level is taken into account. In Table 1 we display the match likelihood (ML) for every locus in all three populations for sequence analysis and for CE analysis in comparison. The additional sequence variation has the strongest effect on the discriminating power of D5S818 and D13S317 with an average three-fold difference in the ML over all populations between the two methods. D2S1338, D3S1358, D7S820, D8S1179, D16S539 and D21S11 exhibit more than a two-fold difference in the ML over all populations. When only taking into account the Dutch and Himalayan population, D5S818, D7S820, D13S317 and D21S11 still exhibit a greater than two-fold difference in match likelihood between length and sequence variation.

## Stutter analysis

Stutter ratios were determined when the CE signal intensity or MPS read coverage was sufficient for alleles which are not influenced by stutters from other alleles. An overview of the read coverage statistics and within locus allele balance of the samples used for this analysis is shown in Supplemental Figure 7. For each locus, dot plots were generated displaying the average stutter ratios for all STR alleles for which at least four stutter ratios could be calculated (Sup. Figure 8). In general, stutter ratios of both methods are very similar with the exception of PentaE where stutter ratios for CE are lower than for sequence data. Some sequence alleles correspond to the same CE allele (e.g. D2S1338 allele 21). For complex STRs, the longest uninterrupted repeat stretch determines the stutter ratio [19] which is confirmed by our data as illustrated in Figure 6. Here, detailed stutter graphs for D18S51 are shown for both methods; the dots of the alleles carrying an interrupted repeat motif (marked in red) tend to have lower stutter ratios than the uninterrupted alleles of the same length.. Because of the separation of these new sequence alleles it is expected that the stutter ratio per sequence allele would show less variation than the CE stutter ratio which represents several sequence variants. To test this, the Coefficient of Variance of the stutter ratio was determined for every allele with stutter data for at least four samples (Sup. Figure 8). Most obtained CV values are either similar or lower for sequencing stutter ratios than for CE stutter ratios. As expected, the loci for which the CV of the stutter ratio is generally lower for sequencing data than for CE are all complex STRs (especially D5S818, D8S1179, D13S317, D21S11, FGA and vWA). In addition it was noted that

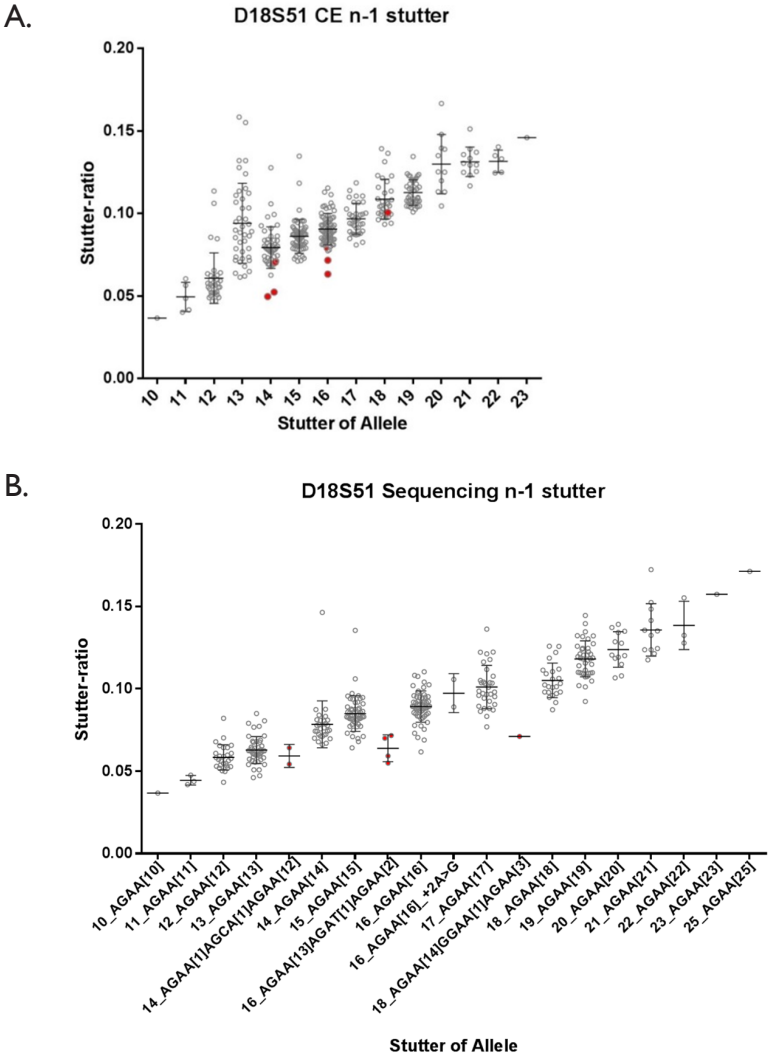
Chapter 4

the CV of the stutter ratio for sequence data remains relatively stable for all alleles within the same locus (even though the stutter becomes higher for longer alleles). For CE-based stutter ratios, much more variation in CV is observed between different alleles within the same locus which is partly explained by alleles that are subdivided into different sequence alleles. For some STRs (in particular D2S1338, D3S1358, D7S820 and D18S51) the CV shows a downward trend for increasing allele length in CE data. An explanation for this decreasing CV could be that low percentage stutter peaks in a CE profile are often below the detection threshold (30 rfu in this analysis). Since a certain number (at least several thousand depending on the fluorescent label) of molecules is needed before a CE peak becomes visible, the signal intensity might not be linearly correlated with the number of molecules for alleles with low peak heights. This could contribute to an increased variation of stutter ratios.

## Mixture analysis

A total of 45 two-person mixtures (from five donor combinations) were analysed with minor contributions of 1%, 5%, 10%, 20% and 50% using the PowerSeq™ sequencing assay. In every mixture, all alleles of both contributors were recovered in the sequence reads, mostly with allele ratios close to expected. Figure 7a displays the read percentage for each allele call of the minor contributor grouped by mixture ratio. Although there is variation, we found that the observed percentage of reads (per allele) from the total locus reads is a good indication of the ratio between two contributors in a mixture. For each of the 45 mixtures the minor contribution was estimated based on the read frequencies of the minor alleles that are not overlapping other alleles or stutter reads in the mixture (see Sup. Figure 9 for further explanation of this procedure for a hypothetical three locus mixture profile). Figure 7b shows the summary statistics for calculation of the minor contribution in the 10 mixtures (for the 50/50 mixtures, calculations were performed for both contributors) of each ratio. Since the total marker reads also contain reads representing stutter, the quantitative prediction of the minor contribution is expected to be slightly lower than the genuine contribution which is apparent for the mixtures with 50% and 20% minor contribution. Not surprisingly, a quantitative prediction of the minor contribution becomes less accurate (relative to the percentage of contribution) when the minor contribution decreases. It is apparent that the standard deviation is almost stable across all mixture ratios.

**Figure 6.** Comparison of stutter ratios for locus D18S51 analysed by CE and MPS

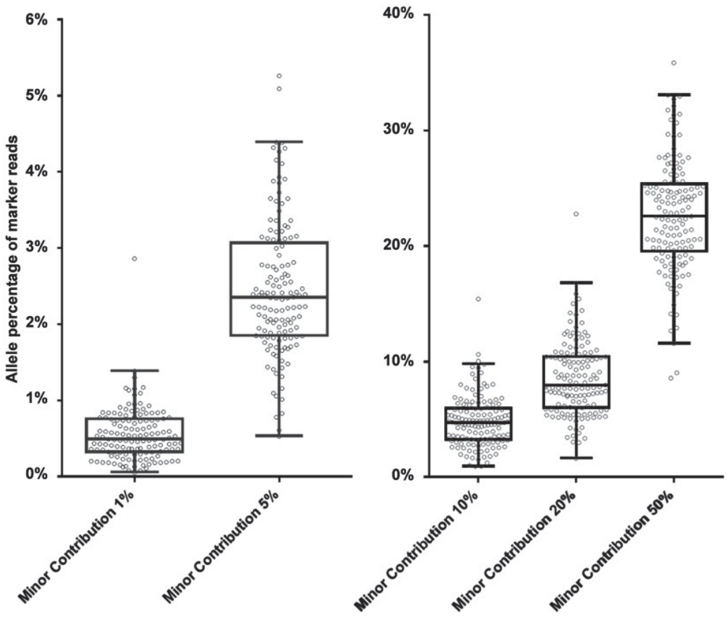



*A. Dot plot displaying the distribution of stutter ratios for the locus D18S51 analysed by CE using the PowerPlex® Fusion System. Every dot represents the stutter ratio of one allele in a single sample, lines display the median and whiskers display 1.5 interquartile range. Red dots represent alleles in which the sequence revealed an interrupted repeat (resulting in a shorter length of the longest repeated motif).* ***B.*** *Dot plot displaying the distribution of stutter ratios for the locus D18S51 analysed by MPS using the prototype PowerSeq™ system. Red dots represent alleles in which the sequence revealed an interrupted repeat. It is apparent that the stutter ratio of the alleles carrying an interrupted repeat motif is generally lower than the alleles of the same length without interruption of the repeat motif.*

Chapter 4

**Figure 7.** Tukey boxplot displaying the observed within locus read percentages of all minor alleles for 10 two-person mixtures for each of the five tested mixture ratios

A.




B.

| Minor Contribution | Average calculated minor contribution | StDev | COV |
|---|---|---|---|
| 1% | 1,1% | 0,32% | 28,1% |
| 5% | 5,09% | 0,28% | 5,6% |
| 10% | 9,97% | 0,30% | 3,0% |
| 20% | 17,28% | 0,26% | 1,5% |
| 50% | 45,33% | 0,28% | 0,6% |

*When analysing alleles with abundance below 5% of the highest allele of the locus, additional PCR/sequence error variants were observed for several loci which can complicate the interpretation of a DNA sample. Therefore, the analysis of minor contributions of 5% or less in a mixture without prior knowledge of the ratio between the different donors, remains difficult for some, but not all loci using the current experimental and analysis setup for this assay. Increasing the sequencing coverage increases the read counts of these artefacts as well and will not help to distinguish them from genuine alleles.*

## Analysing an unknown trace

When unknown samples are analysed that could have more than one contributor, one needs to decide on the minimal allele coverage and level of minor allele detection prior to sequencing. The minimal allele coverage of 8 reads for every allele and 2 reads for both orientations used in this study was chosen for investigative purposes to get an indication of general sequence quality. Although in most cases these thresholds were sufficient to remove artefacts, some erroneous reads can still occur due to a relatively low sequence quality that may be caused by variation in cluster density or other factors yet unknown. In addition to a minimal read coverage to guarantee sequence quality, an additional threshold can be used for the minimal percentage of reads compared to the allele with the highest read count within a locus to filter out structural sequence errors. Below 0.5%, most STRs show a high amount of additional sequence artefacts that coexist with the genuine alleles at a relatively stable ratio. However, when using a high threshold, low percentage contributions might be missed.

## Recommendations

In this study, the population samples were sequenced with an average allele coverage of over 800 reads (also including the samples that were not used for stutter analysis), which is crucial for a reliable characterisation of stutter reads and structural sequence errors in this stage of the development of this new technique. We assume that, eventually, for reliable MPS-STR genotyping of a single-source reference sample (e.g. for database purposes) a much lower coverage could be sufficient. To distinguish genuine allele sequences from errors, we recommend a coverage of at least 20 reads for every allele (sequences from both ends combined) with representation in both orientations. This means that, for the current assay, 5.000 reads per sample will probably be sufficient to achieve the recommended allele coverage. For evidentiary traces, more sequences will be needed since locus balance will be influenced by low template concentrations and low contributions can only be analysed reliably using sufficient reads for the alleles of the minor contribution. For example, when we want to retain sufficient data to detect a minor contribution of 5% we need at least (100/5) x 5000 = 100.000 reads (meaning 100.000 reads for read1 and 100.000 reads for read2) for the current assay. This assumes that the sample is of sufficient quality to retain the same locus balance as a reference sample.

# Conclusion

The analysis of STRs by MPS using the MiSeq® provides several advantages over the routinely used CE. We observed full concordance between CE (Powerplex® Fusion) and MPS (PowerSeq™) based genotyping of STR loci among 297 individuals.

We observed substantial sequence variation within the repeat motifs of STR loci and their immediate flanking regions, in addition to the length variation of the STR-motifs. Since design of a multiplex assay for MPS is no longer limited by the number of different fluorescent labels, PCR primers can be designed to amplify all STR loci within a much more similar fragment size range. This offers advantages for degraded DNA samples and reduces some of the amplification bias due to length variation among the various PCR-templates in a single multiplex PCR reaction. In addition, the exact nature of MPS data (which is as simple as sequence-specific read counts for every allele) provides opportunities for a more standardised follow-up analysis. The study of stutter in MPS data shows that the highest stutter artefact is determined by the longest repeated element in the STR. STR stutter ratios in MPS data are generally similar to those of CE data except for many of the complex STRs since those CE alleles can be differentiated into separate MPS alleles with their own respective stutter profile. Mixture analysis down to a minor contribution of 5% is routinely feasible for most STR loci. Even sequence reads representing a minor contribution down to 1% can be recovered, although here, obviously, reads representing stutters still cause interpretation problems in the reads.

# Acknowledgements

This study was supported by a grant from the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands. The authors wish to thank Titia Sijen (Netherlands Forensic Institute) for carefully reviewing the manuscript and Martin Ensenberger and Cynthia Sprecher (Promega Corporation) for their contributions to the design of the prototype PowerSeq™ System.

# Reference List

1. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE,

Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJ, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. Integrating common and rare genetic variation in diverse human populations. Nature 2010; 467:52-58

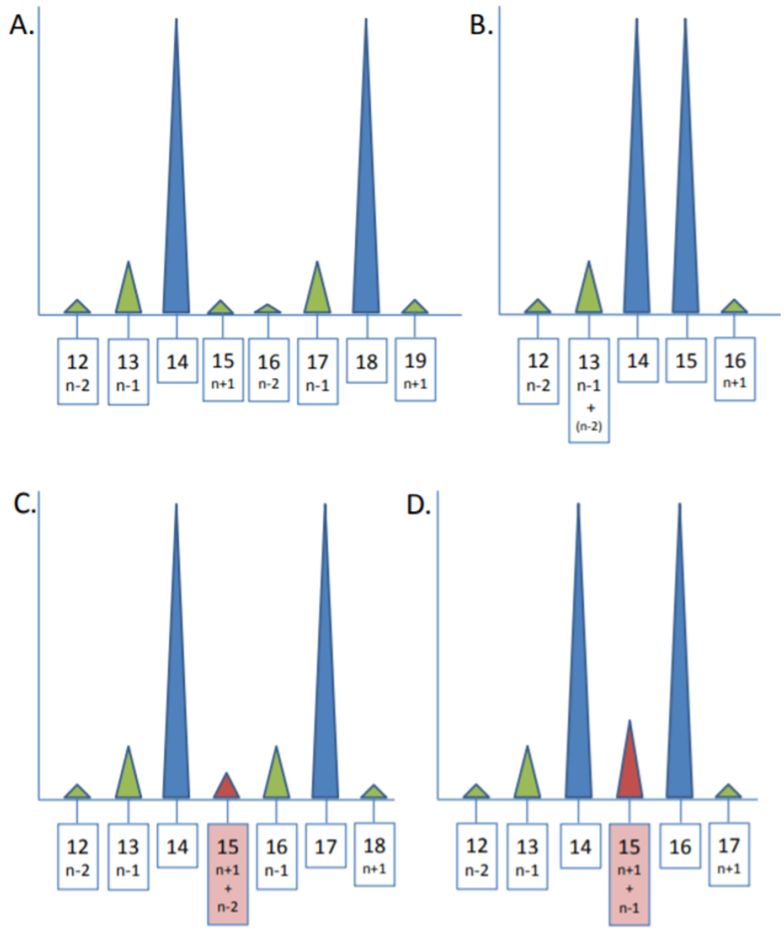
2. Anvar SY, van der Gaag KJ, van der Heijden JW, Veltrop MH, Vossen RH, de Leeuw RH, Breukel C, Buermans HP, Verbeek JS, de KP, den Dunnen JT, Laros JF. TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes. Bioinformatics. 2014; 30:1651-1659
3. Brookes C, Bright JA, Harbison S, Buckleton J. Characterising stutter in forensic STR multiplexes. Forensic Sci.Int.Genet. 2012; 6:58-63
4. Budowle B, Onorato AJ, Callaghan TF, Della MA, Gross AM, Guerrieri RA, Luttman JC, McClure DL. Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J.Forensic Sci. 2009; 54:810-821
5. Cann HM, de TC, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL. A human genome diversity cell line panel. Science 2002; 296:261-262
6. Gelardi C, Rockenbauer E, Dalsgaard S, Borsting C, Morling N. Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. Forensic Sci.Int.Genet. 2014; 12:38-41
7. Gettings KB, Aponte RA, Vallone PM, Butler JM. STR allele sequence variation: Current knowledge and future issues. Forensic Sci.Int.Genet. 2015; 18:118-130
8. Kline MC, Hill CR, Decker AE, Butler JM. STR sequence analysis for characterizing normal, variant, and null alleles. Forensic Sci.Int.Genet. 2011; 5:329-332
9. Knijff, P. and Pijpe, J. Population genetics of African Pygmies. 2015. (GENERIC). Ref Type: Unpublished Work
10. Kraaijenbrink T, van der Gaag KJ, Zuniga SB, Xue Y, Carvalho-Silva DR, Tyler-Smith C, Jobling MA, Parkin EJ, Su B, Shi H, Xiao CJ, Tang WR, Kashyap VK, Trivedi R, Sitalaximi T, Banerjee J, Karma Tshering of Gaselo, Tuladhar NM, Opgenort JR, van Driem GL, Barbujani G, de KP. A linguistically informed autosomal STR survey of human populations residing in the greater Himalayan region. PLoS.One. 2014; 9:e91534
11. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011; 27:2957-2963
12. Oostdik K, Lenz K, Nye J, Schelling K, Yet D, Bruski S, Strong J, Buchanan C, Sutton J, Linner J, Frazier N, Young H, Matthies L, Sage A, Hahn J, Wells R, Williams N, Price M, Koehler J, Staples M, Swango KL, Hill C, Oyerly K, Duke W, Katzilierakis L, Ensenberger MG, Bourdeau JM, Sprecher CJ, Krenke B, Storts DR. Developmental validation of the PowerPlex((R)) Fusion System for analysis of casework and reference samples: A 24-locus multiplex for new database standards. Forensic Sci.Int.Genet. 2014; 12:69-76
13. Promega Corporation. Powerstats v12. 2015. (GENERIC). Ref Type: Unpublished Work
14. Promega Corporation. TECHNICAL MANUAL, PowerPlex® Fusion System. 1-3-2015.

Chapter 4

(GENERIC). Ref Type: Unpublished Work
15. Scheible M, Loreille O, Just R, Irwin J. Short tandem repeat typing on the 454 platform: strategies and considerations for targeted sequencing of common forensic markers. Forensic Sci.Int.Genet. 2014; 12:107-119
16. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015;
17. van der Gaag KJ, de Knijff P. Forensic nomenclature for short tandem repeats updated for sequencing. Forensic Science International: Genetics Supplement Series 2015;Volume 4
18. Warshauer DH, King JL, Budowle B. STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Forensic Sci.Int.Genet. 2015; 14:182-186
19. Westen AA, Grol LJ, Harteveld J, Matai AS, de KP, Sijen T. Assessment of the stochastic threshold, back- and forward stutter filters and low template techniques for NGM. Forensic Sci.Int.Genet. 2012; 6:708-715
20. Westen AA, Kraaijenbrink T, Robles de Medina EA, Harteveld J, Willemse P, Zuniga SB, van der Gaag KJ, Weiler NE, Warnaar J, Kayser M, Sijen T, de KP. Comparing six commercial autosomal STR kits in a large Dutch population sample. Forensic Sci.Int. Genet. 2014; 10:55-63
21. Zeng X, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, Patel J, Storts DR, Budowle B. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci.Int.Genet. 2015; 16:38-47

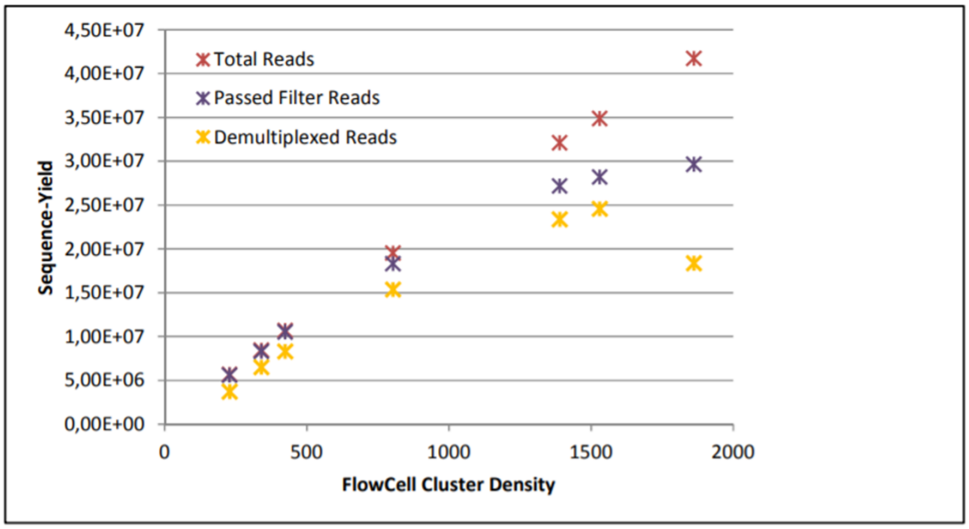
# Supplementary materials

**Supplemental Figure 1,** four examples of allele combinations to illustrate the criteria used for inclusion of stutters for the calculation of stutter ratios




*The peak profiles display which stutter peaks are included and excluded for analysis of stutter ratios. In all four examples, the blue peaks represent the genuine alleles, the green peaks represent stutters that are included in the calculation of stutter ratios and the red peaks represent stutters that are excluded for the calculation of stutter ratios. Example A.: Both alleles have no overlapping stutters, all the stutter peaks are included for calculation of stutter ratios. Example B.: Allele 14 is overlapping with the n-1 stutter of allele 15. For this allele 15, the n-1 stutter cannot be used for calculation of the stutter ratio. The n-1 stutter of allele 14 overlaps the n-2 stutter of allele 15 but is still included in the calculation of the stutter ratio for allele 14 since the contribution of the n-2 stutter to the peak height is considered to be negligible. Example C.: The n+1 stutter of allele 14 is overlapping with the n-2 stutter of allele 17. This stutter position is removed from the analysis of the stutter ratio. Example D.: The n-1 stutter of allele 16 overlaps with the n+1 stutter of allele 14 and is excluded from the analysis of the stutter ratio. For the sequencing data, the same criteria were used but sequence variation was considered resulting in fewer alleles to be excluded from the calculation of the stutter ratios.*

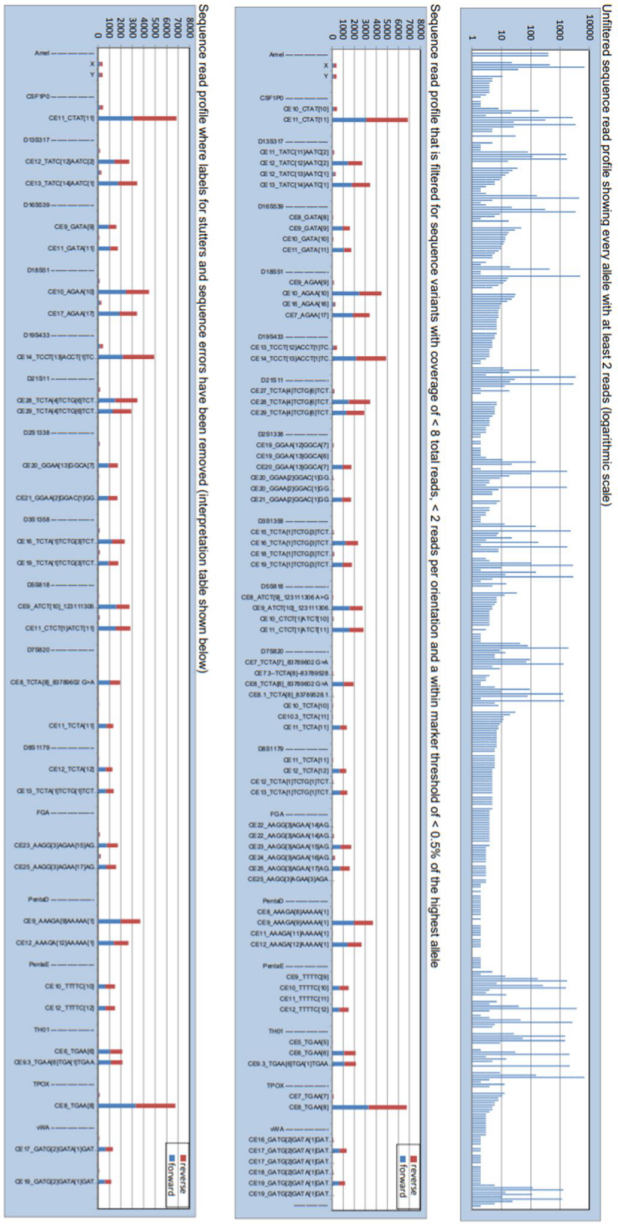
**Supplemental Figure 2.** Overview of sequence efficiency for MiSeq® sequencing runs with different cluster density




*Scatterplot displaying the yield of sequence reads after different filtering steps in the analysis from signal on the MiSeq® until the reads retained in the fastq files after demultiplexing for runs with different levels of cluster density. In red, the initial number of reads are displayed before any filtering took place on the MiSeq®. In purple, remaining reads after MiSeq® quality filtering are displayed. In orange, the reads are displayed in which a barcode is recognised during de-multiplexing.*

**Supplemental Figure 3.** Overview of analysis filtering / interpretation steps of a sequence DNA profile for a single source sample and a mixture

A.



**Final table used for interpretation (Single Sample)**

| Markername | forward | reverse | Total reads | Percentage of highest allele | Allele / Stutter | Allele name | Allele label |
|---|---|---|---|---|---|---|---|
| Amel | forward | reverse | total | % highest allele | annotation/annotation | Amel | Amel ---------- |
| Amel | 195 | 210 | 405 | 100,00% | ALLELE/ALLELE | X | X |
| Amel | 174 | 229 | 403 | 99,51% | ALLELE/ALLELE | Y | Y |
| | | | | | | ---------- | |
| CSF1P0 | forward | reverse | total | % highest allele | annotation/annotation | CSF1P0 | CSF1P0 ---------- |
| CSF1P0 | 192 | 258 | 450 | 6,55% | STUTTER:686(2-1)/STUTTER:1373(2-1) | CE10_CTAT[10] | |
| CSF1P0 | 3088 | 3779 | 6867 | 100,00% | ALLELE/ALLELE | CE11_CTAT[11] | CE11_CTAT[11] |
| | | | | | | | |
| D13S317 | forward | reverse | total | % highest allele | annotation/annotation | D13S317 | D13S317 ---------- |
| D13S317 | 91 | 95 | 186 | 5,41% | STUTTER:274(2-1)/STUTTER:548(2-1) | CE11_TATC[11]AATC[2] | |
| D13S317 | 1460 | 1282 | 2742 | 79,73% | ALLELE/ALLELE | CE12_TATC[12]AATC[2] | CE12_TATC[12]AATC[2] |
| D13S317 | 177 | 141 | 318 | 9,25% | STUTTER:343(2-1)/STUTTER:687(2-1) | CE12_TATC[13]AATC[1] | |
| D13S317 | 1791 | 1648 | 3439 | 100,00% | ALLELE/ALLELE | CE13_TATC[14]AATC[1] | CE13_TATC[14]AATC[1] |
| | | | | | | ---------- | |
| D16S539 | forward | reverse | total | % highest allele | annotation/annotation | D16S539 | D16S539 ---------- |
| D16S539 | 55 | 27 | 82 | 4,70% | STUTTER:161(2-1)/STUTTER:323(2-1) | CE8_GATA[8] | |
| D16S539 | 957 | 662 | 1619 | 92,78% | ALLELE/ALLELE | CE9_GATA[9] | CE9_GATA[9] |
| D16S539 | 68 | 50 | 118 | 6,76% | STUTTER:174(2-1):32(2+1)/STUTTER:349(2 | CE10_GATA[10] | |
| D16S539 | 1075 | 670 | 1745 | 100,00% | ALLELE/ALLELE | CE11_GATA[11] | CE11_GATA[11] |
| | | | | | | ---------- | |
| D18S51 | forward | reverse | total | % highest allele | annotation/annotation | D18S51 | D18S51 ---------- |
| D18S51 | 84 | 79 | 163 | 3,64% | STUTTER:448(2-1)/STUTTER:896(2-1) | CE9_AGAA[9] | |
| D18S51 | 2467 | 2015 | 4482 | 100,00% | ALLELE/ALLELE | CE10_AGAA[10] | CE10_AGAA[10] |
| D18S51 | 179 | 134 | 313 | 6,98% | STUTTER:341(2-1)/STUTTER:683(2-1) | CE16_AGAA[16] | |
| D18S51 | 1889 | 1526 | 3415 | 76,19% | ALLELE/ALLELE | CE17_AGAA[17] | CE17_AGAA[17] |
| | | | | | | ---------- | |
| D19S433 | forward | reverse | total | % highest allele | annotation/annotation | D19S433 | D19S433 ---------- |
| D19S433 | 194 | 245 | 439 | 8,95% | STUTTER:490(2-1)/STUTTER:981(2-1) | CE13_TCCT[12]ACCT[1]TCTT[1]TCCT[1] | |
| D19S433 | 2181 | 2725 | 4906 | 100,00% | ALLELE/ALLELE | CE14_TCCT[13]ACCT[1]TCTT[1]TCCT[1] | CE14_TCCT[13]ACCT[1]TCTT[1]TCCT[1] |
| | | | | | | ---------- | |
| D21S11 | forward | reverse | total | % highest allele | annotation/annotation | D21S11 | D21S11 ---------- |
| D21S11 | 87 | 110 | 197 | 5,72% | STUTTER:344(9-1)/STUTTER:689(9-1) | CE27_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10] | |
| D21S11 | 1516 | 1931 | 3447 | 100,00% | ALLELE/ALLELE | CE28_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | CE28_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] |
| D21S11 | 1305 | 1605 | 2910 | 84,42% | ALLELE/ALLELE | CE29_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | CE29_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] |
| | | | | | | ---------- | |
| D2S1338 | forward | reverse | total | % highest allele | annotation/annotation | D2S1338 | D2S1338 ---------- |
| D2S1338 | 87 | 62 | 149 | 8,51% | STUTTER:175(2-1)/STUTTER:350(2-1) | CE19_GGAA[12]GGCA[7] | |
| D2S1338 | 10 | 13 | 23 | 1,31% | STUTTER:175(3-1):0(2+1x3-1)/STUTTER:35 | CE19_GGAA[13]GGCA[6] | |
| D2S1338 | 965 | 785 | 1750 | 100,00% | ALLELE/ALLELE | CE20_GGAA[13]GGCA[7] | CE20_GGAA[13]GGCA[7] |
| D2S1338 | 56 | 47 | 103 | 5,89% | STUTTER:171(4-1)/STUTTER:343(4-1) | CE20_GGAA[2]GGAC[1]GGAA[10]GGCA[7] | |
| D2S1338 | 8 | 11 | 19 | 1,09% | STUTTER:171(5-1):0(4+1x5-1)/STUTTER:34 | CE20_GGAA[2]GGAC[1]GGAA[11]GGCA[6] | |
| D2S1338 | 909 | 810 | 1719 | 98,23% | ALLELE/ALLELE | CE21_GGAA[2]GGAC[1]GGAA[11]GGCA[7] | CE21_GGAA[2]GGAC[1]GGAA[11]GGCA[7] |
| | | | | | | ---------- | |
| D3S1358 | forward | reverse | total | % highest allele | annotation/annotation | D3S1358 | D3S1358 ---------- |
| D3S1358 | 80 | 70 | 150 | 6,40% | STUTTER:234(4-1)/STUTTER:469(4-1) | CE15_TCTA[1]TCTG[3]TCTA[11] | |
| D3S1358 | 1231 | 1114 | 2345 | 100,00% | ALLELE/ALLELE | CE16_TCTA[1]TCTG[3]TCTA[12] | CE16_TCTA[1]TCTG[3]TCTA[12] |
| D3S1358 | 85 | 98 | 183 | 7,80% | stutter | CE18_TCTA[1]TCTG[3]TCTA[14] | |
| D3S1358 | 952 | 845 | 1797 | 76,63% | ALLELE/ALLELE | CE19_TCTA[1]TCTG[3]TCTA[15] | CE19_TCTA[1]TCTG[3]TCTA[15] |
| | | | | | | ---------- | |
| D5S818 | forward | reverse | total | % highest allele | annotation/annotation | D5S818 | D5S818 ---------- |
| D5S818 | 64 | 43 | 107 | 3,76% | STUTTER:276(2-1)/STUTTER:553(2-1) | CE8_ATCT[9]_123775612 A>G | |
| D5S818 | 1607 | 1160 | 2767 | 97,29% | ALLELE/ALLELE | CE9_ATCT[10]_123775612 A>G | CE9_ATCT[10]_123775612 A>G |
| D5S818 | 86 | 67 | 153 | 5,38% | STUTTER:284(3-1)/STUTTER:568(3-1) | CE10_CTCT[1]ATCT[10] | |
| D5S818 | 1554 | 1290 | 2844 | 100,00% | ALLELE/ALLELE | CE11_CTCT[1]ATCT[11] | CE11_CTCT[1]ATCT[11] |



Chapter 4

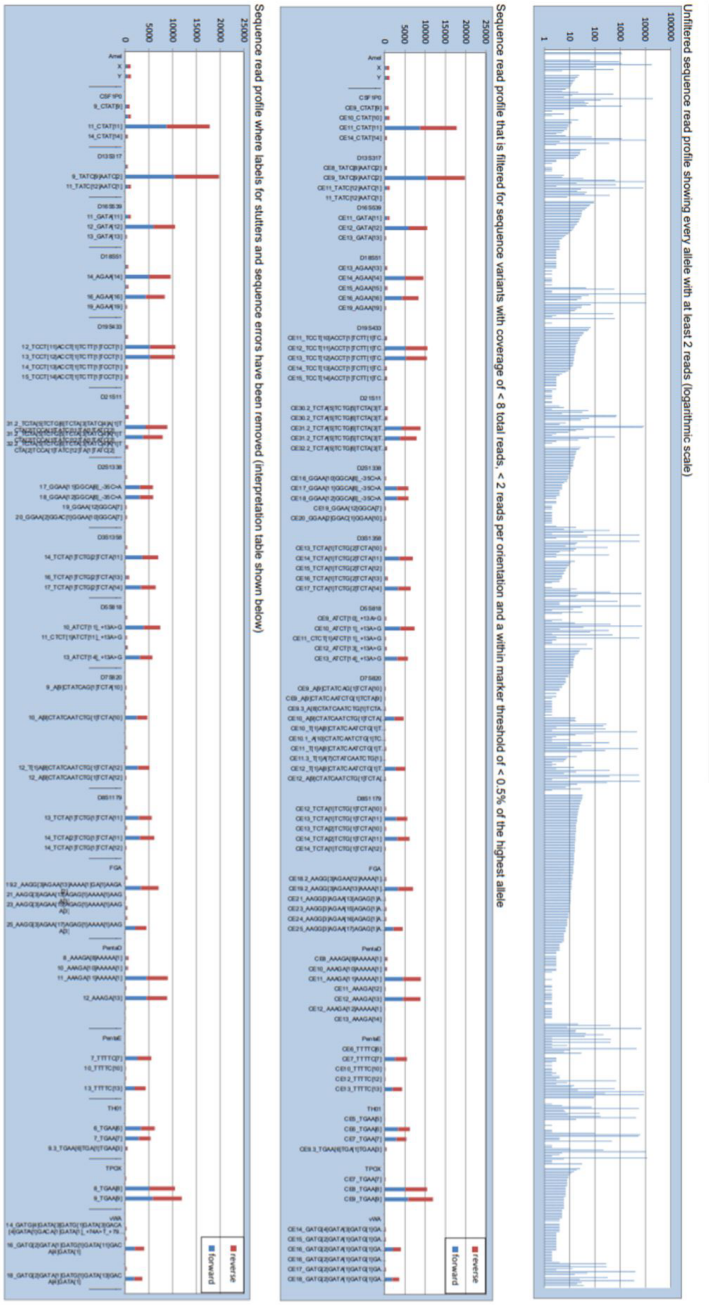
**Final table used for interpretation (Single Sample)**

| Markername | forward | reverse | Total reads | Percentage of highest allele | Allele / Stutter | Allele name | Allele label |
|---|---|---|---|---|---|---|---|
| D7S820 | forward | reverse | total | % highest allele | annotation/annotation | D7S820 | D7S820 ---------- |
| D7S820 | 25 | 19 | 44 | 2,26% | STUTTER:195(4-1):0(2+1x4-1)/STUTTER:39 | CE7_TCTA[7]_84160286 G>A | |
| D7S820 | 42 | 39 | 81 | 4,15% | STUTTER:195(2-1)/STUTTER:390(2-1) | CE7.3–TCTA[8]–84160212 A>del–84160286 G>A | |
| D7S820 | 1057 | 893 | 1950 | 100,00% | ALLELE/ALLELE | CE8_TCTA[8]_84160286 G>A | CE8_TCTA[8]_84160286 G>A |
| D7S820 | 21 | 13 | 34 | 1,74% | STUTTER:39(2+1)/STUTTER:58(2+1) | CE8.1_TCTA[8]_84160212.1 ->insA_84160286 G>A | |
| D7S820 | 55 | 40 | 95 | 4,87% | STUTTER:134(4-1)/STUTTER:268(4-1) | CE10_TCTA[10] | |
| D7S820 | 42 | 30 | 72 | 3,69% | STUTTER:134(2-1):0(2-1x4+1)/STUTTER:26 | CE10.3_TCTA[11] | |
| D7S820 | 745 | 595 | 1340 | 68,72% | ALLELE/ALLELE | CE11_TCTA[11] | CE11_TCTA[11] |
| | | | | | | ---------- | |
| D8S1179 | forward | reverse | total | % highest allele | annotation/annotation | D8S1179 | D8S1179 ---------- |
| D8S1179 | 54 | 40 | 94 | 6,79% | STUTTER:127(2-1)/STUTTER:255(2-1) | CE11_TCTA[11] | |
| D8S1179 | 673 | 602 | 1275 | 92,12% | ALLELE/ALLELE | CE12_TCTA[12] | CE12_TCTA[12] |
| D8S1179 | 38 | 41 | 79 | 5,71% | STUTTER:138(4-1)/STUTTER:276(4-1) | CE12_TCTA[1]TCTG[1]TCTA[10] | |
| D8S1179 | 722 | 662 | 1384 | 100,00% | ALLELE/ALLELE | CE13_TCTA[1]TCTG[1]TCTA[11] | CE13_TCTA[1]TCTG[1]TCTA[11] |
| | | | | | | ---------- | |
| FGA | forward | reverse | total | % highest allele | annotation/annotation | FGA | FGA ---------- |
| FGA | 11 | 9 | 20 | 1,15% | artefact | CE22_AAGG[3]AGAA[14]AGAG[2]AAAA[1]AAGA[3] | |
| FGA | 88 | 86 | 174 | 9,97% | STUTTER:174(3-1)/STUTTER:349(3-1) | CE22_AAGG[3]AGAA[14]AGAG[1]AAAA[1]AAGA[3] | |
| FGA | 786 | 960 | 1746 | 100,00% | ALLELE/ALLELE | CE23_AAGG[3]AGAA[15]AGAG[1]AAAA[1]AAGA[3] | CE23_AAGG[3]AGAA[15]AGAG[1]AAAA[1]AAGA[3] |
| FGA | 129 | 133 | 262 | 15,01% | stutter | CE24_AAGG[3]AGAA[16]AGAG[1]AAAA[1]AAGA[3] | |
| FGA | 740 | 855 | 1595 | 91,35% | ALLELE/ALLELE | CE25_AAGG[3]AGAA[17]AGAG[1]AAAA[1]AAGA[3] | CE25_AAGG[3]AGAA[17]AGAG[1]AAAA[1]AAGA[3] |
| FGA | 15 | 15 | 30 | 1,72% | artefact | CE25_AAGG[3]AGAA[3]AGAG[1]AGAA[13]AGAG[1]AAAA[1]AAGA[3] | |
| | | | | | | ---------- | |
| PentaD | forward | reverse | total | % highest allele | annotation/annotation | PentaD | PentaD ---------- |
| PentaD | 24 | 15 | 39 | 1,05% | STUTTER:371(2-1)/STUTTER:742(2-1) | CE8_AAAGA[8]AAAAA[1] | |
| PentaD | 2006 | 1707 | 3713 | 100,00% | ALLELE/ALLELE | CE9_AAAGA[9]AAAAA[1] | CE9_AAAGA[9]AAAAA[1] |
| PentaD | 23 | 23 | 46 | 1,24% | STUTTER:266(2-1)/STUTTER:533(2-1) | CE11_AAAGA[11]AAAAA[1] | |
| PentaD | 1399 | 1270 | 2669 | 71,88% | ALLELE/ALLELE | CE12_AAAGA[12]AAAAA[1] | CE12_AAAGA[12]AAAAA[1] |
| | | | | | | ---------- | |
| PentaE | forward | reverse | total | % highest allele | annotation/annotation | PentaE | PentaE ---------- |
| PentaE | 13 | 14 | 27 | 1,79% | STUTTER:150(2-1)/STUTTER:301(2-1) | CE9_TTTTC[9] | |
| PentaE | 659 | 846 | 1505 | 100,00% | ALLELE/ALLELE | CE10_TTTTC[10] | CE10_TTTTC[10] |
| PentaE | 20 | 34 | 54 | 3,59% | STUTTER:30(2+1):150(2-1)/STUTTER:45(2+ | CE11_TTTTC[11] | |
| PentaE | 654 | 846 | 1500 | 99,67% | ALLELE/ALLELE | CE12_TTTTC[12] | CE12_TTTTC[12] |
| | | | | | | ---------- | |
| TH01 | forward | reverse | total | % highest allele | annotation/annotation | TH01 | TH01 ---------- |
| TH01 | 18 | 12 | 30 | 1,39% | STUTTER:212(2-1)/STUTTER:425(2-1) | CE5_TGAA[5] | |
| TH01 | 1081 | 1048 | 2129 | 98,56% | ALLELE/ALLELE | CE6_TGAA[6] | CE6_TGAA[6] |
| TH01 | 1130 | 1030 | 2160 | 100,00% | ALLELE/ALLELE | CE9.3_TGAA[6]TGA[1]TGAA[3] | CE9.3_TGAA[6]TGA[1]TGAA[3] |
| | | | | | | ---------- | |
| TPOX | forward | reverse | total | % highest allele | annotation/annotation | TPOX | TPOX ---------- |
| TPOX | 71 | 83 | 154 | 2,27% | STUTTER:677(2-1)/STUTTER:1354(2-1) | CE7_TGAA[7] | |
| TPOX | 3304 | 3466 | 6770 | 100,00% | ALLELE/ALLELE | CE8_TGAA[8] | CE8_TGAA[8] |
| | | | | | | ---------- | |
| vWA | forward | reverse | total | % highest allele | annotation/annotation | vWA | vWA ---------- |
| vWA | 64 | 53 | 117 | 8,91% | STUTTER:131(4-1)/STUTTER:262(4-1) | CE16_GATG[2]GATA[1]GATG[1]GATA[12]GACA[3]GATA[1] | |
| vWA | 677 | 636 | 1313 | 100,00% | ALLELE/ALLELE | CE17_GATG[2]GATA[1]GATG[1]GATA[13]GACA[3]GATA[1] | CE17_GATG[2]GATA[1]GATG[1]GATA[13]GACA[3]GATA[1] |
| vWA | 9 | 5 | 14 | 1,07% | STUTTER:2(4-1x5+1):2(5+1):11(4-1)/STUTT | CE17_GATG[2]GATA[1]GATG[1]GATA[12]GACA[4]GATA[1] | |
| vWA | 57 | 57 | 114 | 8,68% | STUTTER:26(5+1):118(4-1)/STUTTER:39(5+ | CE18_GATG[2]GATA[1]GATG[1]GATA[13]GACA[4]GATA[1] | |
| vWA | 623 | 557 | 1180 | 89,87% | ALLELE/ALLELE | CE19_GATG[2]GATA[1]GATG[1]GATA[14]GACA[4]GATA[1] | CE19_GATG[2]GATA[1]GATG[1]GATA[14]GACA[4]GATA[1] |
| vWA | 13 | 11 | 24 | 1,83% | artefact | CE19_GATG[2]GATA[1]GATG[1]GATA[15]GACA[3]GATA[1] | |
| | | | | | | ---------- | ---------- |

A. Sequence profiles showing allele names after separate filtering steps used in the analysis for a single source sample with the final table used for interpretation to call genuine alleles in a reference sample.

Chapter 4

B.




**Supplemental Figure 3B. Overview of analysis filtering / interpretation steps of a sequence DNA profile from an unknown trace (mixed sample)**

**Final table used for interpretation (Mixed Sample)**

| Markername | forward | reverse | Total reads | Percentage of highest allele | Allele / Stutter | Allele name | Allele label |
|---|---|---|---|---|---|---|---|
| Amel | forward | reverse | total | % highest alle | annotation/annotation | Amel | Amel |
| Amel | 491 | 625 | 1116 | 94,66% | ALLELE/ALLELE | X | X |
| Amel | 529 | 650 | 1179 | 100,00% | ALLELE/ALLELE | Y | Y |
| | | | | | | --- | --- |
| CSF1P0 | forward | reverse | total | % highest alle | annotation/annotation | CSF1P0 | CSF1P0 |
| CSF1P0 | 463 | 443 | 906 | 5,10% | ALLELE/ALLELE | CE9_CTAT[9] | CE9_CTAT[9] |
| CSF1P0 | 581 | 579 | 1160 | 6,53% | STUTTER:1777(2-1)/STUTTER:3555(2-1) | CE10_CTAT[10] | |
| CSF1P0 | 8739 | 9038 | 17777 | 100,00% | ALLELE/ALLELE | CE11_CTAT[11] | CE11_CTAT[11] |
| CSF1P0 | 250 | 292 | 542 | 3,05% | ALLELE/ALLELE | CE14_CTAT[14] | CE14_CTAT[14] |
| | | | | | | --- | --- |
| D13S317 | forward | reverse | total | % highest alle | annotation/annotation | D13S317 | D13S317 |
| D13S317 | 282 | 257 | 539 | 2,72% | STUTTER:1978(2-1)/STUTTER:3957(2-1) | CE8_TATC[8]AATC[2] | |
| D13S317 | 10429 | 9357 | 19786 | 100,00% | ALLELE/ALLELE | CE9_TATC[9]AATC[2] | CE9_TATC[9]AATC[2] |
| D13S317 | 603 | 589 | 1192 | 6,02% | ALLELE/ALLELE | CE11_TATC[12]AATC[1] | CE11_TATC[12]AATC[1] |
| | | | | | | --- | --- |
| D16S539 | forward | reverse | total | % highest alle | annotation/annotation | D16S539 | D16S539 |
| D16S539 | 661 | 499 | 1160 | 11,04% | ALLELE/STUTTER:2102(2-1) | CE11_GATA[11] | CE11_GATA[11] |
| D16S539 | 5921 | 4591 | 10512 | 100,00% | ALLELE/ALLELE | CE12_GATA[12] | CE12_GATA[12] |
| D16S539 | 211 | 159 | 370 | 3,52% | ALLELE/ALLELE | CE13_GATA[13] | CE13_GATA[13] |
| | | | | | | --- | --- |
| D18S51 | forward | reverse | total | % highest alle | annotation/annotation | D18S51 | D18S51 |
| D18S51 | 315 | 311 | 626 | 6,54% | STUTTER:956(2-1)/STUTTER:1913(2-1) | CE13_AGAA[13] | |
| D18S51 | 4995 | 4570 | 9565 | 100,00% | ALLELE/ALLELE | CE14_AGAA[14] | CE14_AGAA[14] |
| D18S51 | 382 | 301 | 683 | 7,14% | STUTTER:191(2+1):832(2-1)/STUTTER:286(2 | CE15_AGAA[15] | |
| D18S51 | 4313 | 4013 | 8326 | 87,05% | ALLELE/ALLELE | CE16_AGAA[16] | CE16_AGAA[16] |
| D18S51 | 198 | 191 | 389 | 4,07% | ALLELE/ALLELE | CE19_AGAA[19] | CE19_AGAA[19] |
| | | | | | | --- | --- |
| D19S433 | forward | reverse | total | % highest alle | annotation/annotation | D19S433 | D19S433 |
| D19S433 | 295 | 288 | 583 | 5,53% | STUTTER:1055(2-1)/STUTTER:2110(2-1) | CE11_TCCT[10]ACCT[1]TCTT[1]TCCT[1] | |
| D19S433 | 5223 | 5329 | 10552 | 100,00% | ALLELE/ALLELE | CE12_TCCT[11]ACCT[1]TCTT[1]TCCT[1] | CE12_TCCT[11]ACCT[1]TCTT[1]TCCT[1] |
| D19S433 | 5240 | 5180 | 10420 | 98,75% | ALLELE/ALLELE | CE13_TCCT[12]ACCT[1]TCTT[1]TCCT[1] | CE13_TCCT[12]ACCT[1]TCTT[1]TCCT[1] |
| D19S433 | 246 | 280 | 526 | 4,98% | ALLELE/ALLELE | CE14_TCCT[13]ACCT[1]TCTT[1]TCCT[1] | CE14_TCCT[13]ACCT[1]TCTT[1]TCCT[1] |
| D19S433 | 311 | 305 | 616 | 5,84% | ALLELE/ALLELE | CE15_TCCT[14]ACCT[1]TCTT[1]TCCT[1] | CE15_TCCT[14]ACCT[1]TCTT[1]TCCT[1] |
| | | | | | | --- | --- |
| D21S11 | forward | reverse | total | % highest alle | annotation/annotation | D21S11 | D21S11 |
| D21S11 | 331 | 394 | 725 | 8,20% | STUTTER:884(3-1):788(9-1)/STUTTER:1769( | CE30.2_TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2] | |
| D21S11 | 314 | 370 | 684 | 7,73% | STUTTER:1(9-1x3+1):884(9-1)/STUTTER:4(9- | CE30.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]TA[1]TATC[2] | |
| D21S11 | 4215 | 4631 | 8846 | 100,00% | ALLELE/ALLELE | CE31.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2] | CE31.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2] |
| D21S11 | 3752 | 4133 | 7885 | 89,14% | ALLELE/ALLELE | CE31.2_TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2] | CE31.2_TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2] |
| D21S11 | 298 | 316 | 614 | 6,94% | ALLELE/ALLELE | CE32.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2] | CE32.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2] |
| | | | | | | --- | --- |
| D2S1338 | forward | reverse | total | % highest alle | annotation/annotation | D2S1338 | D2S1338 |
| D2S1338 | 199 | 186 | 385 | 6,55% | STUTTER:583(2-1)/STUTTER:1166(2-1) | CE16_GGAA[10]GGCA[6]_218014824C>A | |
| D2S1338 | 3052 | 2781 | 5833 | 99,29% | ALLELE/ALLELE | CE17_GGAA[11]GGCA[6]_218014824 C>A | CE17_GGAA[11]GGCA[6]_218014824 C>A |
| D2S1338 | 3068 | 2807 | 5875 | 100,00% | ALLELE/ALLELE | CE18_GGAA[12]GGCA[6]_218014824 C>A | CE18_GGAA[12]GGCA[6]_218014824 C>A |
| D2S1338 | 165 | 152 | 317 | 5,40% | ALLELE/ALLELE | CE19_GGAA[12]GGCA[7] | CE19_GGAA[12]GGCA[7] |
| D2S1338 | 199 | 166 | 365 | 6,21% | ALLELE/ALLELE | CE20_GGAA[2]GGAC[1]GGAA[10]GGCA[7] | CE20_GGAA[2]GGAC[1]GGAA[10]GGCA[7] |
| | | | | | | --- | --- |
| D3S1358 | forward | reverse | total | % highest alle | annotation/annotation | D3S1358 | D3S1358 |
| D3S1358 | 196 | 207 | 403 | 5,81% | STUTTER:693(4-1)/STUTTER:1386(4-1) | CE13_TCTA[1]TCTG[2]TCTA[10] | |
| D3S1358 | 3563 | 3368 | 6931 | 100,00% | ALLELE/ALLELE | CE14_TCTA[1]TCTG[2]TCTA[11] | CE14_TCTA[1]TCTG[2]TCTA[11] |
| D3S1358 | 46 | 57 | 103 | 1,49% | STUTTER:138(4+1):78(4-1)/STUTTER:207(4+ | CE15_TCTA[1]TCTG[2]TCTA[12] | |
| D3S1358 | 421 | 368 | 789 | 11,38% | ALLELE/STUTTER:1282(4-1) | CE16_TCTA[1]TCTG[2]TCTA[13] | CE16_TCTA[1]TCTG[2]TCTA[13] |
| D3S1358 | 3331 | 3081 | 6412 | 92,51% | ALLELE/ALLELE | CE17_TCTA[1]TCTG[2]TCTA[14] | CE17_TCTA[1]TCTG[2]TCTA[14] |
| | | | | | | --- | --- |
| D5S818 | forward | reverse | total | % highest alle | annotation/annotation | D5S818 | D5S818 |
| D5S818 | 220 | 199 | 419 | 5,69% | STUTTER:736(2-1)/STUTTER:1473(2-1) | CE9_ATCT[10]_123775612 A>G | |
| D5S818 | 3855 | 3512 | 7367 | 100,00% | ALLELE/ALLELE | CE10_ATCT[11]_123775612 A>G | CE10_ATCT[11]_123775612 A>G |
| D5S818 | 196 | 159 | 355 | 4,82% | ALLELE/ALLELE | CE11_CTCT[1]ATCT[11]_123775612 A>G | CE11_CTCT[1]ATCT[11]_123775612 A>G |
| D5S818 | 262 | 239 | 501 | 6,80% | STUTTER:572(2-1)/STUTTER:1145(2-1) | CE12_ATCT[13]_123775612 A>G | |
| D5S818 | 3065 | 2662 | 5727 | 77,74% | ALLELE/ALLELE | CE13_ATCT[14]_123775612 A>G | CE13_ATCT[14]_123775612 A>G |

Chapter 4

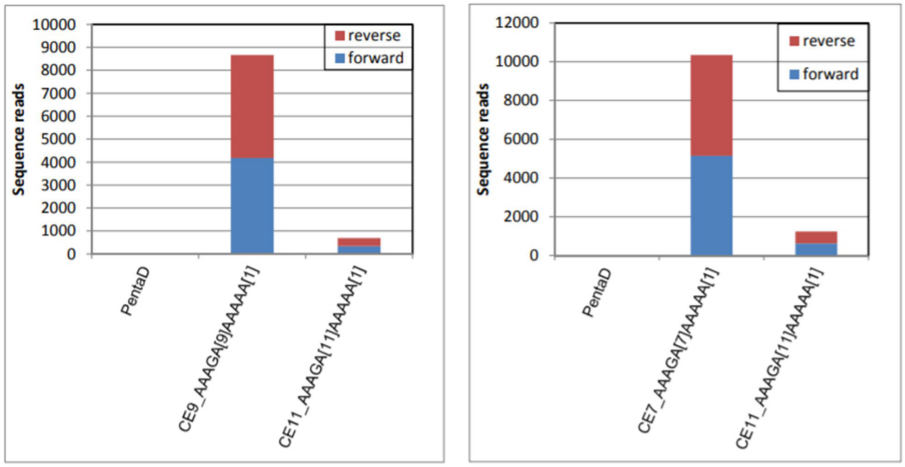
**Final table used for interpretation (Mixed Sample)**

| Markername | forward | reverse | Total reads | Percentage of highest allele | Allele / Stutter annotation/annotation | Allele name | Allele label |
|---|---|---|---|---|---|---|---|
| D7S820 | forward | reverse | total | % highest allele | annotation/annotation | D7S820 | Allele label |
| D7S820 | 127 | 127 | 254 | 5.67% | ALLELE/ALLELE | CE9_TCTA[10]_84160205 ATCT>del | D7S820 |
| D7S820 | 152 | 92 | 244 | 5.44% | ALLELE/ALLELE | CE9_TCTA[10]_84160205 ATCT>del | CE9_TCTA[10]_84160205 ATCT>del |
| D7S820 | 140 | 125 | 265 | 5.91% | STUTTER 465(2-1)/STUTTER 501(2-1) | CE9.3_TCTA[10]_84160212 A>del | |
| D7S820 | 2427 | 2230 | 4657 | 91.44% | ALLELE/ALLELE | CE10_TCTA[10]_84160204 T>A | CE10_TCTA[10] |
| D7S820 | 49 | 43 | 92 | 1.81% | stutter | CE10_TCTA[10]_84160204 T>A | |
| D7S820 | 46 | 35 | 81 | 1.59% | STUTTER 930(2+1)/STUTTER 1002(2+1) | CE10.1_TCTA[10]_84160212 -->A | |
| D7S820 | 187 | 184 | 371 | 7.28% | STUTTER 508(5-1)/STUTTER 1018(5-1) | CE11_TCTA[11]_84160204 T>A | |
| D7S820 | 34 | 39 | 73 | 1.43% | STUTTER 1018(5-1)/STUTTER 1129(5-1) | CE11.1_TCTA[11]_84160204 T>A_84160212 A>del | |
| D7S820 | 2805 | 2368 | 5093 | 100.00% | ALLELE/ALLELE | CE12_TCTA[12]_84160204 T>A_84160212 A>del | |
| D7S820 | 151 | 105 | 256 | 5.03% | ALLELE/ALLELE | CE12_TCTA[12] | CE12_TCTA[12] |
| D8S1179 | forward | reverse | total | % highest allele | annotation/annotation | D8S1179 | D8S1179 |
| D8S1179 | 189 | 193 | 382 | 6.25% | STUTTER 559(4+1)/STUTTER 1118(4-1) | CE12_TCTA[1]TCTG[1]TCTA[10] | |
| D8S1179 | 3556 | 3581 | 7137 | 116.78% | ALLELE/ALLELE | CE13_TCTA[1]TCTG[1]TCTA[11] | CE13_TCTA[1]TCTG[1]TCTA[11] |
| D8S1179 | 1823 | 1735 | 3558 | 58.19% | STUTTER 1197(4-1)/STUTTER 1222(4-1) | CE14_TCTA[2]TCTG[1]TCTA[11] | |
| D8S1179 | 3069 | 3044 | 6113 | 100.00% | ALLELE/ALLELE | CE14_TCTA[2]TCTG[1]TCTA[12] | CE14_TCTA[1]TCTG[1]TCTA[12] |
| D8S1179 | 153 | 149 | 302 | 4.94% | ALLELE/ALLELE | CE14_TCTA[1]TCTG[1]TCTA[12] | |
| FGA | forward | reverse | total | % highest allele | annotation/annotation | FGA | FGA |
| FGA | 202 | 211 | 413 | 5.89% | STUTTER 701(3-1)/STUTTER 1402(3-1) | CE18.2_AAGG[3]AGAA[1]AAAA[1]GA[1]AAGA[2] | |
| FGA | 3507 | 3503 | 7010 | 100.00% | ALLELE/ALLELE | CE19_AAGG[3]AGAA[1]AAAA[1]GA[1]AAGA[2] | CE19.2_AAGG[3]AGAA[1]AAAA[1]GA[1]AAGA[2] |
| FGA | 114 | 105 | 219 | 3.12% | ALLELE/ALLELE | CE21_AAGG[3]AGAA[1]AAAA[1]AAGA[3] | CE21_AAGG[3]AGAA[1]AAAA[1]AAGA[3] |
| FGA | 207 | 221 | 428 | 6.11% | ALLELE/ALLELE | CE23_AAGG[3]AGAA[1]AAAA[1]AAGA[3] | CE23_AAGG[3]AGAA[1]AAAA[1]AAGA[3] |
| FGA | 204 | 213 | 417 | 5.95% | STUTTER 446(3-1)/STUTTER 891(3-1) | CE24_AAGG[3]AGAA[1]6]AAGG[1]AAAA[1]AAGA[3] | |
| FGA | 2125 | 2332 | 4457 | 63.58% | ALLELE/ALLELE | CE25_AAGG[3]AGAA[1]7]AGAG[1]AAAA[1]AAGA[3] | CE25_AAGG[3]AGAA[1]7]AGAG[1]AAAA[1]AAGA[3] |
| PentaD | forward | reverse | total | % highest allele | annotation/annotation | PentaD | PentaD |
| PentaD | 335 | 363 | 698 | 7.80% | ALLELE/ALLELE | CE8_AAAGA[8]AAAAA[1] | CE8_AAAGA[8]AAAAA[1] |
| PentaD | 329 | 286 | 615 | 6.87% | allele | CE10_AAAGA[10]AAAAA[1] | CE10_AAAGA[10]AAAAA[1] |
| PentaD | 4530 | 4416 | 8946 | 100.00% | ALLELE/ALLELE | CE11_AAAGA[11]AAAAA[1] | CE11_AAAGA[11]AAAAA[1] |
| PentaD | 145 | 105 | 250 | 2.79% | STUTTER 882(2-1)/STUTTER 1764(2-1) | CE11_AAAGA[12] | |
| PentaD | 4522 | 4299 | 8821 | 98.60% | ALLELE/ALLELE | CE12_AAAGA[13] | CE12_AAAGA[13] |
| PentaD | 65 | 65 | 130 | 1.14% | artifact | CE12_AAAGA[12]AAAAA[1] | |
| PentaD | 51 | 40 | 91 | 1.02% | STUTTER 176(2+1)/STUTTER 264(2+1) | CE13_AAAGA[14] | |
| PentaE | forward | reverse | total | % highest allele | annotation/annotation | PentaE | PentaE |
| PentaE | 24 | 74 | 50 | 1.35% | STUTTER 540(2-1)/STUTTER 1098(2-1) | CE5_TTTTC[5] | |
| PentaE | 2576 | 2917 | 5493 | 100.00% | ALLELE/ALLELE | CE7_TTTTC[7] | CE7_TTTTC[7] |
| PentaE | 63 | 92 | 155 | 2.82% | ALLELE/ALLELE | CE10_TTTTC[10] | CE10_TTTTC[10] |
| PentaE | 72 | 92 | 164 | 2.98% | STUTTER 1026(2-1)/STUTTER 858(2-1) | CE12_TTTTC[12] | |
| PentaE | 1943 | 2348 | 4291 | 78.12% | ALLELE/ALLELE | CE13_TTTTC[13] | CE13_TTTTC[13] |
| TH01 | forward | reverse | total | % highest allele | annotation/annotation | TH01 | TH01 |
| TH01 | 58 | 43 | 101 | 1.63% | STUTTER 621(2-1)/STUTTER 1242(2-1) | CE5_TCAA[5] | |
| TH01 | 3325 | 2885 | 6210 | 100.00% | ALLELE/ALLELE | CE6_TCAA[6] | CE6_TCAA[6] |
| TH01 | 2969 | 2570 | 5539 | 89.20% | ALLELE/ALLELE | CE7_TCAA[7] | CE7_TCAA[7] |
| TH01 | 242 | 233 | 475 | 7.65% | ALLELE/ALLELE | CE9.3_TCAA[9]TGA[1]TGAA[3] | CE9.3_TCAA[9]TGA[1]TGAA[3] |
| TPOX | forward | reverse | total | % highest allele | annotation/annotation | TPOX | TPOX |
| TPOX | 103 | 116 | 219 | 1.84% | STUTTER 1048(2-1)/STUTTER 2096(2-1) | CE7_TGAA[7] | |
| TPOX | 5060 | 5421 | 10481 | 87.99% | ALLELE/ALLELE | CE8_TGAA[8] | CE8_TGAA[8] |
| TPOX | 5783 | 6133 | 11916 | 100.00% | ALLELE/ALLELE | CE9_TGAA[9] | CE9_TGAA[9] |
| vWA | forward | reverse | total | % highest allele | annotation/annotation | vWA | vWA |
| vWA | 138 | 143 | 281 | 7.07% | ALLELE/ALLELE | CE14_GATG[4]GATA[3]GATG[1]GATA[3]GACA[9]GATA[1]GACA[1]GATA[1]_5984116 A>T | |
| vWA | 145 | 143 | 288 | 7.24% | STUTTER 397(4-1)/STUTTER 793(4-1) | CE15_GATG[2]GATA[1]GATG[1]GATA[1]9]GACA[4]GATA[1] | |
| vWA | 1988 | 1988 | 3976 | 100.00% | ALLELE/ALLELE | CE16_GATG[2]GATA[1]GATG[1]GATA[1]3]GACA[4]GATA[1] | CE16_GATG[2]GATA[1]GATG[1]GATA[1]3]GACA[4]GATA[1] |
| vWA | 21 | 19 | 40 | 1.01% | STUTTER 1194(5+1)/STUTTER 1194(5-1) | CE17_GATG[2]GATA[1]GATG[1]GATA[1]2]GACA[4]GATA[1] | |
| vWA | 164 | 157 | 321 | 8.07% | STUTTER 794(4+1)/STUTTER 1190(4-1) | CE18_GATG[2]GATA[1]GATG[1]GATA[1]3]GACA[4]GATA[1] | CE18_GATG[2]GATA[1]GATG[1]GATA[1]3]GACA[4]GATA[1] |
| vWA | 1760 | 1735 | 3495 | 88.44% | ALLELE/ALLELE | | |

B. Sequence profiles showing allele names after separate filtering steps used in the analysis for a mixed sample with the final table used for interpretation to call genuine alleles in a mixed sample.

*A. Sequence profiles showing allele names after separate filtering steps used in the analysis for a single source sample with the final table used for interpretation to call genuine alleles in a reference sample. B. Sequence profiles showing allele names after separate filtering steps used in the analysis for a mixed sample with the final table used for interpretation to call genuine alleles in a mixed sample.*

**Supplemental Figure 4.** Sequence read profile for the two samples where PentaD showed a strong allelic imbalance




*This figure displays the sequence read profile for PentaD for the two samples where a strong allelic imbalance was observed. For these samples, the number of reads for the allele with CE-length 11 was only observed for 8% and 12% of the reads of the second allele.*

**Supplemental Figure 5.** STR sequence alleles for the reference genome and 2800M Control DNA

| | STR coordinates in the reference genome (hg38) | Reference | | HGVS-name refseq: | CE-length |
|---|---|---|---|---|---|
| CSF1P0 | chr5.hg38: g.150076322-150076373 | HGVS | REF | CTAT[13] | 13 |
| | | 2800M | Allele | CTAT[12] | 12 |
| D2S1338 | chr2.hg38: g.218014859-218014950 | HGVS | REF | GGAA[2]GGAC[1]GGAA[13]GGCA[7] | 23 |
| | | 2800M | Allele 1 | GGAA[2]GGAC[1]GGAA[12]GGCA[7] | 22 |
| | | | Allele 2 | GGAA[2]GGAC[1]GGAA[15]GGCA[7] | 25 |
| D3S1358 | chr3.hg38: g.45540739-45540802 | HGVS | REF | TCTA[1]TCTG[1]TCTA[14] | 16 |
| | | 2800M | Allele 1 | TCTA[1]TCTG[3]TCTA[13] | 17 |
| | | | Allele 2 | TCTA[1]TCTG[3]TCTA[14] | 18 |
| D5S818 | chr5.hg38: g.123775552-123775599 | HGVS | REF | CTCT[1]ATCT[11] | 11 |
| | | 2800M | Allele 1 | CTCT[1]ATCT[12] | 12 |
| | | | Allele 2 | CTCT[1]ATCT[12]_123775612 A>G | 12 |
| D7S820 | chr7.hg38: g.84160224-84160275 | HGVS | REF | TCTA[13] | 13 |
| | | 2800M | Allele 1 | TCTA[8] | 8 |
| | | | Allele 2 | TCTA[11] | 11 |
| D8S1179 | chr8.hg38: g.124894865-124894916 | HGVS | REF | TCTA[1]TCTG[1]TCTA[11] | 13 |
| | | 2800M | Allele 1 | TCTA[1]TCTG[1]TCTA[12] | 14 |
| | | | Allele 2 | TCTA[2]TCTG[1]TCTA[12] | 15 |
| D13S317 | chr13.hg38: g.82148025-82148088 | HGVS | REF | TATC[11]AATC[2]ATCT[3] | 11 |
| | | 2800M | Allele 1 | TATC[9]AATC[2]ATCT[3] | 9 |
| | | | Allele 2 | TATC[12]AATC[1]ATCT[3] | 11 |
| D16S539 | chr16.hg38: g.86352702-86352745 | HGVS | REF | GATA[11] | 11 |
| | | 2800M | Allele 1 | GATA[9] | 9 |
| | | | Allele 2 | GATA[13] | 13 |
| D18S51 | chr18.hg38: g.63281667-63281750 | HGVS | REF | AGAA[18]AA[1]AG[4] | 18 |
| | | 2800M | Allele 1 | AGAA[16]AA[1]AG[4] | 16 |
| | | | Allele 2 | AGAA[18]AA[1]AG[4] | 18 |
| D19S433 | chr19.hg38: g.29926234-29926297 | HGVS | REF | TCCT[13]ACCT[1]TCTT[1]TCCT[1] | 14 |
| | | 2800M | Allele 1 | TCCT[12]ACCT[1]TCTT[1]TCCT[1] | 13 |
| | | | Allele 2 | TCCT[13]ACCT[1]TCTT[1]TCCT[1] | 14 |
| D21S11 | chr21.hg38: g.19181973-19182101 | HGVS | REF | TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | 29 |
| | | 2800M | Allele 1 | TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | 29 |
| | | | Allele 2 | TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2] | 31.2 |
| FGA | chr4.hg38: g.154587726-154587821 | HGVS | REF | AAGG[1]AAGA[1]AAGG[3]AGAA[14]AGAG[1]AAAA[1]AAGA[3] | 22 |
| | | 2800M | Allele 1 | AAGG[1]AAGA[1]AAGG[3]AGAA[12]AGAG[1]AAAA[1]AAGA[3] | 20 |
| | | | Allele 2 | AAGG[1]AAGA[1]AAGG[3]AGAA[15]AGAG[1]AAAA[1]AAGA[3] | 23 |
| PentaD | chr21.hg38: g.43636205-43636274 | HGVS | REF | AAAGA[13]AAAAA[1] | 13 |
| | | 2800M | Allele 1 | AAAGA[12]AAAAA[1] | 12 |
| | | | Allele 2 | AAAGA[13]AAAAA[1] | 13 |
| PentaE | chr15.hg38: g.96831012-96831036 | HGVS | REF | TTTTC[5] | 5 |
| | | 2800M | Allele 1 | TTTTC[7] | 7 |
| | | | Allele 2 | TTTTC[14] | 14 |
| TH01 | chr11.hg38: g.2171086-2171113 | HGVS | REF | TGAA[7] | 7 |
| | | 2800M | Allele 1 | TGAA[6] | 6 |
| | | | Allele 2 | TGAA[6]TGA[1]TGAA[3] | 9.3 |
| TPOX | chr2.hg38: g.1489651-1489682 | HGVS | REF | TGAA[8] | 8 |
| | | 2800M | Allele | TGAA[11] | 11 |
| vWA | chr12.hg38: g.5983959-5984042 | HGVS | REF | GATG[2]GATA[1]GATG[1]GATA[11]GACA[5]GATA[1] | 17 |
| | | 2800M | Allele 1 | GATG[2]GATA[1]GATG[1]GATA[12]GACA[3]GATA[1] | 16 |
| | | | Allele 2 | GATG[2]GATA[1]GATG[1]GATA[14]GACA[4]GATA[1] | 19 |

*Description of the coordinates of the STR motif for each STR and the alleles represented in the reference genome (hg38) and in 2800M Control DNA.*

**Supplemental Figure 6.** Observed sequence variation in 297 analysed samples from three populations

| Population | Total Samples | Total Alleles |
|---|---|---|
| Dutch | 101 | 202 |
| Nepal / Bhutan | 97 | 194 |
| Pygmy | 99 | 198 |
| Total | 297 | 594 |

## Sequenced Alleles

### Amel

Total sequence alleles 3
Total CE fragment alleles 2
Alleles containing SNPs outside STR-motif 1

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| X | X | 0,490 | 0,608 | 0,717 | 0,604 |
| X | X–11296959C>T | 0,010 | | | 0,003 |
| Y | Y | 0,500 | 0,392 | 0,283 | 0,393 |

### CSF1P0

Total sequence alleles 10
Total CE fragment alleles 10
Alleles containing SNPs outside STR-motif 0

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 7 | CE7–CTAT[7] | | | 0,020 | 0,007 |
| 8 | CE8–CTAT[8] | 0,010 | | 0,035 | 0,015 |
| 9 | CE9–CTAT[9] | 0,044 | 0,062 | 0,020 | 0,042 |
| 10 | CE10–CTAT[10] | 0,260 | 0,242 | 0,374 | 0,292 |
| 10.3 | CE10.3–CTAT[6]CAT[1]CTAT[4] | 0,005 | | | 0,002 |
| 11 | CE11–CTAT[11] | 0,299 | 0,263 | 0,268 | 0,277 |
| 12 | CE12–CTAT[12] | 0,304 | 0,381 | 0,263 | 0,315 |
| 13 | CE13–CTAT[13] | 0,064 | 0,031 | 0,015 | 0,037 |
| 14 | CE14–CTAT[14] | 0,010 | 0,015 | | 0,008 |
| 15 | CE15–CTAT[15] | 0,005 | 0,005 | 0,005 | 0,005 |

Chapter 4

**D2S1338**

Total sequence alleles 55
Total CE fragment alleles 12
Alleles containing SNPs outside STR-motif 14

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 14 | CE14–GGAA[8]GGCA[6]–218014824C>A | 0,005 | | | 0,002 |
| 16 | CE16–GGAA[12]GGCA[4]–218014824C>A | | | 0,066 | 0,022 |
| 16 | CE16–GGAA[11]GGCA[5] | | | 0,015 | 0,005 |
| 16 | CE16–GGAA[10]GGCA[6]–218014824C>A | 0,074 | | | 0,025 |
| 17 | CE17–GGAA[12]GGCA[5]–218014824C>A | | | 0,005 | 0,002 |
| 17 | CE17–GGAA[11]GGCA[6] | | 0,005 | | 0,002 |
| 17 | CE17–GGAA[11]GGCA[6]–218014824C>A | 0,230 | 0,041 | 0,015 | 0,097 |
| 18 | CE18–GGAA[15]GGCA[3]–218014824C>A | | | 0,005 | 0,002 |
| 18 | CE18–GGAA[14]GGCA[4]–218014824C>A | | | 0,010 | 0,003 |
| 18 | CE18–GGAA[13]GGCA[5] | | 0,005 | | 0,002 |
| 18 | CE18–GGAA[12]GGCA[6] | | 0,010 | | 0,003 |
| 18 | CE18–GGAA[12]GGCA[6]–218014824C>A | 0,074 | | 0,005 | 0,027 |
| 18 | CE18–GGAA[11]GGCA[7] | 0,015 | 0,077 | 0,040 | 0,044 |
| 19 | CE19–GGAA[15]GGCA[4] | | | 0,005 | 0,002 |
| 19 | CE19–GGAA[14]GGCA[5] | 0,005 | 0,010 | 0,040 | 0,018 |
| 19 | CE19–GGAA[2]GGAC[1]GGAA[10]GGCA[6] | | | 0,005 | 0,002 |
| 19 | CE19–GGAA[13]GGCA[6] | | 0,005 | 0,061 | 0,022 |
| 19 | CE19–GGAA[13]GGCA[6]–218014824C>A | 0,015 | | | 0,005 |
| 19 | CE19–GGAA[12]GGCA[7] | 0,059 | 0,170 | 0,106 | 0,111 |
| 19 | CE19–GGAA[11]GGCA[8] | | | 0,005 | 0,002 |
| 20 | CE20–GGAA[17]GGCA[3]–218014824C>A | | | 0,005 | 0,002 |
| 20 | CE20–GGAA[16]GGCA[4]–218014824C>A | | | 0,005 | 0,002 |
| 20 | CE20–GGAA[14]GGCA[6] | 0,020 | 0,010 | 0,071 | 0,034 |
| 20 | CE20–GGAA[14]GGCA[6]–218014824C>A | | 0,005 | | 0,002 |
| 20 | CE20–GGAA[12]GGGA[1]GGCA[7] | 0,015 | | | 0,005 |
| 20 | CE20–GGAA[2]GGAC[1]GGAA[10]GGCA[7] | 0,015 | 0,005 | | 0,007 |
| 20 | CE20–GGAA[13]GGCA[7] | 0,088 | 0,108 | 0,056 | 0,084 |
| 20 | CE20–GGAA[10]GAAA[1]GGAA[2]GGCA[7] | | 0,015 | | 0,005 |
| 20 | CE20–GGAA[12]GGCA[8] | 0,005 | | 0,005 | 0,003 |
| 20 | CE20–GGAA[2]GGAC[1]GGAA[9]GGCA[8] | | 0,010 | | 0,003 |
| 21 | CE21–GGAA[17]GGCA[4]–218014824C>A | | | 0,035 | 0,012 |
| 21 | CE21–GGAA[2]GGAC[1]GGAA[13]GGCA[5] | | 0,005 | | 0,002 |
| 21 | CE21–GGAA[16]GGCA[5] | | | 0,020 | 0,007 |
| 21 | CE21–GGAA[2]GGAC[1]GGAA[12]GGCA[6] | | | 0,030 | 0,010 |
| 21 | CE21–GGAA[15]GGCA[6] | | | 0,056 | 0,018 |
| 21 | CE21–GGAA[2]GGAC[1]GGAA[11]GGCA[7] | 0,005 | | 0,010 | 0,005 |
| 21 | CE21–GGAA[14]GGCA[7] | 0,025 | 0,046 | 0,040 | 0,037 |
| 21 | CE21–GGAA[13]GGCA[8] | | 0,010 | 0,030 | 0,013 |
| 22 | CE22–GGAA[18]GGCA[4]–218014824C>A | | | 0,005 | 0,002 |
| 22 | CE22–GGAA[2]GGAC[1]GGAA[14]GGCA[5] | | | 0,005 | 0,002 |
| 22 | CE22–GGAA[2]GGAC[1]GGAA[13]GGCA[6] | | 0,010 | 0,040 | 0,017 |
| 22 | CE22–GGAA[2]GGAC[1]GGAA[12]GGCA[7] | 0,010 | 0,026 | 0,045 | 0,027 |
| 22 | CE22–GGAA[15]GGCA[7] | | | 0,005 | 0,002 |
| 22 | CE22–GGAA[2]GGAC[1]GGAA[11]GGCA[8] | | | 0,005 | 0,002 |
| 23 | CE23–GGAA[2]GGAC[1]GGAA[15]GGCA[5] | | 0,010 | | 0,003 |
| 23 | CE23–GGAA[2]GGAC[1]GGAA[14]GGCA[6] | 0,005 | 0,005 | | 0,003 |
| 23 | CE23–GGAA[2]GGAC[1]GGAA[13]GGCA[7] | 0,069 | 0,180 | 0,025 | 0,091 |
| 23 | CE23–GGAA[14]GGCA[9] | | | 0,005 | 0,002 |
| 24 | CE24–GGAA[2]GGAC[1]GGAA[15]GGCA[6] | 0,005 | 0,005 | 0,010 | 0,007 |
| 24 | CE24–GGAA[2]GGAC[1]GGAA[14]GGCA[7] | 0,093 | 0,155 | 0,076 | 0,107 |
| 24 | CE24–GGAA[2]GGAC[1]GGAA[13]GGCA[8] | | 0,005 | | 0,002 |
| 25 | CE25–GGAA[2]GGAC[1]GGAA[16]GGCA[6] | 0,010 | 0,005 | | 0,005 |
| 25 | CE25–GGAA[2]GGAC[1]GGAA[15]GGCA[7] | 0,142 | 0,046 | 0,020 | 0,070 |
| 25 | CE25–GGAA[2]GGAC[1]GGAA[14]GGCA[8] | 0,005 | 0,005 | | 0,003 |
| 26 | CE26–GGAA[2]GGAC[1]GGAA[16]GGCA[7] | 0,015 | 0,005 | 0,010 | 0,010 |

**D3S1358**

Total sequence alleles 21
Total CE fragment alleles 9
Alleles containing SNPs outside STR-motif 1

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 10 | CE10–TCTA[1]TCTG[2]TCTA[7] | 0,005 | | | 0,002 |
| 13 | CE13–TCTA[1]TCTG[1]TCTA[11] | | | 0,005 | 0,002 |
| 13 | CE13–TCTA[1]TCTG[2]TCTA[10] | | 0,005 | 0,005 | 0,003 |
| 14 | CE14–TCTA[1]TCTG[1]TCTA[12] | 0,005 | | 0,040 | 0,015 |
| 14 | CE14–TCTA[1]TCTG[2]TCTA[11] | 0,162 | 0,036 | 0,056 | 0,086 |
| 15 | CE15–TCTA[1]TCTG[1]TCTA[13] | 0,020 | | 0,045 | 0,022 |
| 15 | CE15–TCTA[1]TCTG[2]TCTA[12] | 0,211 | 0,284 | 0,207 | 0,233 |
| 15 | CE15–TCTA[1]TCTG[3]TCTA[11] | 0,020 | 0,010 | 0,040 | 0,023 |
| 16 | CE16–TCTA[1]TCTG[1]TCTA[14] | 0,015 | | 0,141 | 0,052 |
| 16 | CE16–TCTA[1]TCTG[2]TCTA[13] | 0,118 | 0,201 | 0,247 | 0,188 |
| 16 | CE16–TCTA[1]TCTG[2]TCTA[13]–45540653C>T | | 0,005 | | 0,002 |
| 16 | CE16–TCTA[1]TCTG[3]TCTA[12] | 0,078 | 0,098 | 0,010 | 0,062 |
| 16.2 | CE16.2–TCTA[1]TCTG[3]TC[1]TCTA[12] | | | 0,005 | 0,002 |
| 17 | CE17–TCTA[1]TCTG[1]TCTA[11]TCCA[1]TCTA[3] | | | 0,010 | 0,003 |
| 17 | CE17–TCTA[1]TCTG[1]TCTA[15] | 0,010 | | 0,015 | 0,008 |
| 17 | CE17–TCTA[1]TCTG[2]TCTA[14] | 0,088 | 0,160 | 0,086 | 0,111 |
| 17 | CE17–TCTA[1]TCTG[3]TCTA[13] | 0,064 | 0,082 | 0,071 | 0,072 |
| 17 | CE17–TCTA[1]TCTG[4]TCTA[12] | 0,015 | | | 0,005 |
| 18 | CE18–TCTA[1]TCTG[2]TCTA[15] | 0,015 | 0,026 | 0,015 | 0,018 |
| 18 | CE18–TCTA[1]TCTG[3]TCTA[14] | 0,167 | 0,088 | | 0,086 |
| 19 | CE19–TCTA[1]TCTG[3]TCTA[15] | 0,010 | 0,005 | | 0,005 |

**D5S818**

Total sequence alleles 23
Total CE fragment alleles 7
Alleles containing SNPs outside STR-motif 18

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 8 | CE8–CTCT[1]ATCT[8]–123775612A>G | 0,010 | | | 0,003 |
| 8 | CE8–ATCT[9]–123775612A>G | | | 0,045 | 0,015 |
| 9 | CE9–ATCT[10]–123775612A>G | 0,029 | 0,072 | 0,010 | 0,037 |
| 9 | CE9–CTCT[1]ATCT[9]–123775612A>G | | 0,005 | | 0,002 |
| 10 | CE10–CTCT[1]ATCT[10]–123775612A>G | 0,020 | 0,155 | 0,025 | 0,065 |
| 10 | CE10–CTCT[1]ATCT[10] | 0,005 | | 0,005 | 0,003 |
| 10 | CE10–ATCT[11]–123775612A>G | 0,039 | 0,010 | 0,071 | 0,040 |
| 11 | CE11–CTCT[1]ATCT[11]–123775612A>G | 0,235 | 0,253 | 0,086 | 0,191 |
| 11 | CE11–CTCT[1]ATCT[11] | 0,064 | 0,005 | 0,010 | 0,027 |
| 11 | CE11–ATCT[12]–123775612A>G | 0,025 | 0,057 | 0,035 | 0,039 |
| 12 | CE12–CTCT[1]ATCT[12]–123775612A>G | 0,230 | 0,206 | 0,141 | 0,193 |
| 12 | CE12–CTCT[1]ATCT[12] | 0,113 | 0,041 | 0,066 | 0,074 |
| 12 | CE12–CTCT[1]ATCT[12]–123775612A>G–123775657T>G | | | 0,061 | 0,020 |
| 12 | CE12–ATCT[13]–123775612A>G | 0,069 | 0,036 | 0,157 | 0,087 |
| 13 | CE13–CTCT[1]ATCT[13]–123775612A>G | 0,049 | 0,129 | 0,091 | 0,089 |
| 13 | CE13–CTCT[1]ATCT[13] | 0,059 | 0,021 | 0,025 | 0,035 |
| 13 | CE13–CTCT[1]ATCT[13]–123775612A>G–123775657T>G | | | 0,010 | 0,003 |
| 13 | CE13–ATCT[14]–123775612A>G | 0,034 | 0,005 | 0,131 | 0,057 |
| 13 | CE13–ATCT[14]–123775611CA>TG | | | 0,005 | 0,002 |
| 13 | CE13–CTCT[1]ATCT[3]ATGT[1]ATCT[9]–123775612A>G | | | 0,015 | 0,005 |
| 14 | CE14–CTCT[1]ATCT[14]–123775612A>G | | | 0,010 | 0,003 |
| 14 | CE14–CTCT[1]ATCT[14] | 0,005 | 0,005 | | 0,003 |
| 14 | CE14–ATCT[15]–123775612A>G | 0,015 | | | 0,005 |

Chapter 4

**D7S820**

Total sequence alleles 32
Total CE fragment alleles 10
Alleles containing SNPs outside STR-motif 25

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 6 | CE6–TCTA[6]–84160204T>A | | | 0,005 | 0,002 |
| 7 | CE7–TCTA[7] | 0,025 | | 0,005 | 0,010 |
| 7 | CE7–TCTA[7]–84160204T>A | | 0,005 | | 0,002 |
| 8 | CE8–TCTA[8] | 0,127 | 0,057 | 0,136 | 0,107 |
| 8 | CE8–TCTA[8]–84160204T>A | | 0,005 | | 0,002 |
| 8 | CE8–TCTA[8]–84160204T>A–84160161A>C | | | 0,010 | 0,003 |
| 8 | CE8–TCTA[8]–84160204T>A–84160286G>A | 0,044 | 0,191 | 0,066 | 0,099 |
| 9 | CE9–TCTA[10]–84160219ATCT>del–84160204T>A | 0,005 | | | 0,002 |
| 9 | CE9–TCTA[9] | 0,176 | 0,036 | 0,086 | 0,101 |
| 9 | CE9–TCTA[9]–84160204T>A | | 0,010 | | 0,003 |
| 9 | CE9–TCTA[9]–84160204T>A–84160161A>C | | 0,005 | 0,025 | 0,010 |
| 9 | CE9–TCTA[9]–84160204T>A–84160286G>A | 0,015 | 0,072 | 0,056 | 0,047 |
| 10 | CE10–TCTA[10] | 0,162 | 0,098 | 0,222 | 0,161 |
| 10 | CE10–TCTA[10]–84160204T>A | 0,049 | 0,015 | | 0,022 |
| 10 | CE10–TCTA[10]–84160204T>A–84160161A>C | | | 0,005 | 0,002 |
| 10 | CE10–TCTA[10]–84160161A>C | 0,010 | 0,031 | 0,035 | 0,025 |
| 10 | CE10–TCTA[10]–84160204T>A–84160286G>A | | 0,005 | 0,020 | 0,008 |
| 10.3 | CE10.3–TCTA[11]–84160204T>del | 0,005 | | | 0,002 |
| 11 | CE11–TCTA[11] | 0,191 | 0,155 | 0,071 | 0,139 |
| 11 | CE11–TCTA[11]–84160204T>A | 0,015 | 0,041 | | 0,018 |
| 11 | CE11–TCTA[11]–84160204T>A–84160161A>C | | | 0,010 | 0,003 |
| 11 | CE11–TCTA[11]–84160161A>C | | 0,021 | 0,066 | 0,029 |
| 11 | CE11–TCTA[11]–84160204T>A–84160286G>A | | 0,026 | 0,005 | 0,010 |
| 11 | CE11–TCTA[8]CCTA[1]TCTA[2]–84160204T>A | | 0,010 | | 0,003 |
| 12 | CE12–TCTA[12] | 0,108 | 0,175 | 0,066 | 0,116 |
| 12 | CE12–TCTA[12]–84160204T>A | 0,034 | 0,021 | | 0,018 |
| 12 | CE12–TCTA[12]–84160204T>A–84160161A>C | | | 0,040 | 0,013 |
| 12 | CE12–TCTA[12]–84160161A>C | | | 0,030 | 0,010 |
| 12 | CE12–TCTA[12]–84160204T>A–84160286G>A | | | 0,035 | 0,012 |
| 13 | CE13–TCTA[13] | 0,015 | 0,015 | | 0,010 |
| 13 | CE13–TCTA[13]–84160204T>A | 0,010 | 0,005 | | 0,005 |
| 14 | CE14–TCTA[14]–84160204T>A | 0,010 | | 0,005 | 0,005 |

**D8S1179**

Total sequence alleles 27
Total CE fragment alleles 11
Alleles containing SNPs outside STR-motif 1

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 8 | CE8–TCTA[8] | 0,010 | 0,005 | | 0,005 |
| 9 | CE9–TCTA[9] | 0,010 | | | 0,003 |
| 10 | CE10–TCTA[10] | 0,088 | 0,072 | 0,010 | 0,057 |
| 11 | CE11–TCTA[1]TCTG[1]TCTA[9] | | 0,005 | | 0,002 |
| 11 | CE11–TCTA[11] | 0,083 | 0,036 | | 0,040 |
| 11 | CE11–TCTA[2]TCTG[1]TCTA[8] | | | 0,030 | 0,010 |
| 12 | CE12–TCTA[1]TCTG[1]TCTA[10] | 0,010 | 0,046 | 0,005 | 0,020 |
| 12 | CE12–TCTA[12] | 0,137 | 0,062 | 0,020 | 0,074 |
| 12 | CE12–TCTA[2]TCTG[1]TCTA[9] | | | 0,086 | 0,029 |
| 12.1 | CE12.1–TCTA[1]TCTG[1]TCTA[9]TCTTA[1] | 0,005 | | | 0,002 |
| 13 | CE13–TCTA[1]TCTG[1]TCTA[11] | 0,260 | 0,155 | 0,106 | 0,174 |
| 13 | CE13–TCTA[13] | 0,064 | 0,062 | | 0,042 |
| 13 | CE13–TCTA[2]TCTG[1]TCTA[10] | | | 0,071 | 0,023 |
| 14 | CE14–TCTA[1]TCTG[1]TCTA[12] | 0,142 | 0,139 | 0,172 | 0,151 |
| 14 | CE14–TCTA[1]TCTG[1]TGTA[1]TCTA[11] | 0,005 | | | 0,002 |
| 14 | CE14–TCTA[14] | 0,020 | 0,005 | 0,005 | 0,010 |
| 14 | CE14–TCTA[2]TCTG[1]TCTA[11] | 0,049 | 0,062 | 0,192 | 0,101 |
| 15 | CE15–TCTA[1]TCTG[1]TCTA[13] | 0,044 | 0,031 | 0,061 | 0,045 |
| 15 | CE15–TCTA[2]TCTG[1]TCTA[12] | 0,059 | 0,180 | 0,116 | 0,117 |
| 15 | CE15–TCTA[2]TCTG[2]TCTA[11] | | | 0,030 | 0,010 |
| 16 | CE16–TCTA[1]TCTG[1]TCTA[14] | | 0,005 | | 0,002 |
| 16 | CE16–TCTA[1]TCTG[3]TCTA[12] | | | 0,010 | 0,003 |
| 16 | CE16–TCTA[2]TCTG[1]TCTA[13] | 0,010 | 0,113 | 0,056 | 0,059 |
| 16 | CE16–TCTA[2]TCTG[1]TCTA[13]–124894972G>A | | 0,010 | | 0,003 |
| 16 | CE16–TCTA[2]TCTG[2]TCTA[12] | | | 0,010 | 0,003 |
| 17 | CE17–TCTA[1]TCTG[3]TCTA[13] | | | 0,005 | 0,002 |
| 17 | CE17–TCTA[2]TCTG[1]TCTA[14] | 0,005 | 0,010 | 0,015 | 0,010 |

**D13S317**

Total sequence alleles 31
Total CE fragment alleles 8
Alleles containing SNPs outside STR-motif 10

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 7 | CE7–TATC[7]AATC[2]ATCT[3] | 0,005 | | | 0,002 |
| 7 | CE7–TATC[8]AATC[1]ATCT[3] | | 0,010 | | 0,003 |
| 8 | CE8–TATC[8]AATC[2]ATCT[3] | 0,083 | 0,175 | 0,030 | 0,096 |
| 9 | CE9–TATC[10]AATC[1]ATCT[3] | 0,010 | 0,077 | | 0,029 |
| 9 | CE9–TATC[9]AATC[2]ATCT[3] | 0,083 | 0,139 | 0,015 | 0,079 |
| 10 | CE10–TATC[10]AATC[2]ATCT[3] | 0,049 | 0,041 | 0,005 | 0,032 |
| 10 | CE10–TATC[11]AATC[1]ATCT[3] | | 0,124 | 0,020 | 0,047 |
| 10 | CE10–TATC[12]ATCT[3] | | 0,005 | | 0,002 |
| 10 | CE10–TATC[12]AATC[1]ATCT[3]–82148097GTCT>del | | | 0,005 | 0,002 |
| 11 | CE11–TATC[11]AATC[2]ATCT[3] | 0,103 | 0,036 | 0,096 | 0,079 |
| 11 | CE11–TATC[12]AATC[1]ATCT[3] | 0,181 | 0,149 | 0,157 | 0,163 |
| 11 | CE11–TATC[12]AATC[1]ATCT[3]–82148000C>T | 0,039 | 0,005 | | 0,015 |
| 11 | CE11–TATC[12]AATC[1]ATCT[3]–82147972G>A | | | 0,020 | 0,007 |
| 11 | CE11–TATC[13]ATCT[3] | | 0,021 | | 0,007 |
| 11 | CE11–TATC[8]TGTC[1]TATC[3]AATC[1]ATCT[3] | | 0,021 | | 0,007 |
| 12 | CE12–TATC[12]AATC[2]ATCT[3] | 0,167 | 0,026 | 0,308 | 0,168 |
| 12 | CE12–TATC[13]AATC[1]ATCT[3] | 0,127 | 0,124 | 0,111 | 0,121 |
| 12 | CE12–TATC[13]AATC[1]ATCT[3]–82148001G>A | | | 0,035 | 0,012 |
| 12 | CE12–TATC[13]AATC[1]ATCT[3]–82148000C>T | 0,005 | | | 0,002 |
| 12 | CE12–TATC[13]AATC[1]ATCT[3]–82147972G>A | | | 0,035 | 0,012 |
| 12 | CE12–TATC[14]ATCT[3] | | 0,021 | | 0,007 |
| 12 | CE12–TATC[7]TATT[1]TATC[5]AATC[1]ATCT[3] | | | 0,010 | 0,003 |
| 13 | CE13–TATC[13]AATC[2]ATCT[3] | 0,039 | | 0,116 | 0,052 |
| 13 | CE13–TATC[14]AATC[1]ATCT[3] | 0,034 | 0,005 | 0,005 | 0,015 |
| 13 | CE13–TATC[14]AATC[1]ATCT[3]–82148001G>A | | | 0,005 | 0,002 |
| 13 | CE13–TATC[14]AATC[1]ATCT[3]–82148000C>T | | 0,005 | | 0,002 |
| 13 | CE13–TATC[14]AATC[1]ATCT[3]–82147972G>A | | | 0,005 | 0,002 |
| 13 | CE13–TATC[15]ATCT[3] | | 0,010 | | 0,003 |
| 13 | CE13–TATC[15]AATC[1]ATCT[3]–82148097GTCT>del | | | 0,010 | 0,003 |
| 14 | CE14–TATC[14]AATC[2]ATCT[3] | 0,074 | | 0,010 | 0,029 |
| 14 | CE14–TATC[15]AATC[1]ATCT[3] | | 0,005 | | 0,002 |

**D16S539**

Total sequence alleles 17
Total CE fragment alleles 9
Alleles containing SNPs outside STR-motif 9

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 5 | CE5–GATA[5] | | | 0,040 | 0,013 |
| 8 | CE8–GATA[8] | 0,010 | 0,010 | 0,010 | 0,010 |
| 9 | CE9–GATA[9] | 0,186 | 0,278 | 0,162 | 0,208 |
| 9 | CE9–GATA[9]–86352607A>C | | | 0,096 | 0,032 |
| 10 | CE10–GATA[10] | 0,039 | 0,098 | 0,116 | 0,084 |
| 10 | CE10–GATA[10]–86352607A>C | | | 0,040 | 0,013 |
| 11 | CE11–GATA[11] | 0,309 | 0,237 | 0,101 | 0,216 |
| 11 | CE11–GATA[11]–86352607A>C | 0,044 | 0,067 | 0,207 | 0,106 |
| 11 | CE11–GATA[5]GACA[1]GATA[5]–86352607A>C | | 0,010 | | 0,003 |
| 11 | CE11–GATA[11]–86352607A>C–86352571A>C | | | 0,005 | 0,002 |
| 12 | CE12–GATA[12] | 0,235 | 0,077 | 0,051 | 0,122 |
| 12 | CE12–GATA[12]–86352584G>C | 0,005 | | | 0,002 |
| 12 | CE12–GATA[12]–86352607A>C | 0,020 | 0,119 | 0,081 | 0,072 |
| 13 | CE13–GATA[13] | 0,118 | 0,021 | 0,051 | 0,064 |
| 13 | CE13–GATA[13]–86352607A>C | 0,029 | 0,062 | 0,035 | 0,042 |
| 14 | CE14–GATA[14] | 0,005 | | 0,005 | 0,003 |
| 14 | CE14–GATA[14]–86352607A>C | | 0,021 | | 0,007 |

**D18S51**

Total sequence alleles 19
Total CE fragment alleles 15
Alleles containing SNPs
outside STR-motif 0

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 10 | D18S51–CE10–AGAA[10]AA[1]AG[4] | 0,005 | | | 0,002 |
| 11 | D18S51–CE11–AGAA[11]AA[1]AG[4] | 0,020 | 0,005 | | 0,008 |
| 12 | D18S51–CE12–AGAA[12]AA[1]AG[4] | 0,167 | 0,036 | 0,005 | 0,070 |
| 13 | D18S51–CE13–AGAA[13]AA[1]AG[4] | 0,069 | 0,227 | 0,025 | 0,106 |
| 14 | D18S51–CE14–AGAA[1]AGCA[1]AGAA[12]AA[1]AG[4] | 0,015 | | | 0,005 |
| 14 | D18S51–CE14–AGAA[14]AA[1]AG[4] | 0,132 | 0,108 | 0,061 | 0,101 |
| 15 | D18S51–CE15–AGAA[15]AA[1]AG[4] | 0,162 | 0,180 | 0,131 | 0,158 |
| 16 | D18S51–CE16–AGAA[13]AGAT[1]AGAA[2]AA[1]AG[4] | | | 0,020 | 0,007 |
| 16 | D18S51–CE16–AGAA[16]AA[1]AG[4] | 0,206 | 0,129 | 0,217 | 0,185 |
| 16 | D18S51–CE16–AGAA[16]AG[5] | 0,005 | | 0,005 | 0,003 |
| 17 | D18S51–CE17–AGAA[17]AA[1]AG[4] | 0,098 | 0,057 | 0,197 | 0,117 |
| 18 | D18S51–CE18–AGAA[14]GGAA[1]AGAA[3]AA[1]AG[4] | | | 0,005 | 0,002 |
| 18 | D18S51–CE18–AGAA[18]AA[1]AG[4] | 0,064 | 0,067 | 0,126 | 0,086 |
| 19 | D18S51–CE19–AGAA[19]AA[1]AG[4] | 0,025 | 0,103 | 0,141 | 0,089 |
| 20 | D18S51–CE20–AGAA[20]AA[1]AG[4] | 0,015 | 0,041 | 0,020 | 0,025 |
| 21 | D18S51–CE21–AGAA[21]AA[1]AG[4] | 0,010 | 0,015 | 0,040 | 0,022 |
| 22 | D18S51–CE22–AGAA[22]AA[1]AG[4] | 0,010 | 0,015 | | 0,008 |
| 23 | D18S51–CE23–AGAA[23]AA[1]AG[4] | | 0,010 | 0,005 | 0,005 |
| 25 | D18S51–CE25–AGAA[25]AA[1]AG[4] | | 0,005 | | 0,002 |

**D19S433**

Total sequence alleles 20
Total CE fragment alleles 17
Alleles containing SNPs
outside STR-motif 0

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 9 | CE9–TCCT[8]ACCT[1]TCTT[1]TCCT[1] | | | 0,015 | 0,005 |
| 10 | CE10–TCCT[9]ACCT[1]TCTT[1]TCCT[1] | | | 0,177 | 0,059 |
| 10.2 | CE10.2–TCCT[10]ACCT[1]TT[1]TCCT[1] | | | 0,005 | 0,002 |
| 11 | CE11–TCCT[10]ACCT[1]TCTT[1]TCCT[1] | | | 0,045 | 0,015 |
| 11.2 | CE11.2–TCCT[11]ACCT[1]TT[1]TCCT[1] | | 0,005 | | 0,002 |
| 12 | CE12–TCCT[11]ACCT[1]TCTT[1]TCCT[1] | 0,054 | 0,041 | 0,081 | 0,059 |
| 12.2 | CE12.2–TCCT[12]ACCT[1]TT[1]TCCT[1] | | 0,015 | 0,025 | 0,013 |
| 13 | CE13–TCCT[12]ACCT[1]TCTT[1]TCCT[1] | 0,260 | 0,242 | 0,167 | 0,223 |
| 13 | CE13–TCCT[13]TCTT[1]TCCT[1] | 0,005 | | | 0,002 |
| 13.2 | CE13.2–TCCT[13]ACCT[1]TT[1]TCCT[1] | 0,020 | 0,036 | 0,111 | 0,055 |
| 14 | CE14–TCCT[13]ACCT[1]TCTT[1]TCCT[1] | 0,328 | 0,237 | 0,136 | 0,235 |
| 14 | CE14–TCCT[9]TCCC[1]TCCT[3]ACCT[1]TCTT[1]TCCT[1] | | 0,010 | | 0,003 |
| 14 | CE14–TCCT[14]TCTT[1]TCCT[1] | 0,005 | | | 0,002 |
| 14.2 | CE14.2–TCCT[14]ACCT[1]TT[1]TCCT[1] | 0,025 | 0,082 | 0,051 | 0,052 |
| 15 | CE15–TCCT[14]ACCT[1]TCTT[1]TCCT[1] | 0,201 | 0,077 | 0,020 | 0,101 |
| 15.2 | CE15.2–TCCT[15]ACCT[1]TT[1]TCCT[1] | 0,039 | 0,170 | 0,101 | 0,102 |
| 16 | CE16–TCCT[15]ACCT[1]TCTT[1]TCCT[1] | 0,029 | 0,031 | 0,025 | 0,029 |
| 16.2 | CE16.2–TCCT[16]ACCT[1]TT[1]TCCT[1] | 0,020 | 0,046 | 0,035 | 0,034 |
| 17 | CE17–TCCT[16]ACCT[1]TCTT[1]TCCT[1] | 0,010 | 0,005 | | 0,005 |
| 17.2 | CE17.2–TCCT[17]ACCT[1]TT[1]TCCT[1] | 0,005 | | 0,005 | 0,003 |

**D21S11**

Total sequence alleles 63
Total CE fragment alleles 25
Alleles containing SNPs outside STR-motif 2

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 24.2 | CE24.2–TCTA[5]TCTG[6]TCTA[3]TC[1]A[1]TCTA[2]TCCA[1]TATC[10] | 0,005 | | | 0,002 |
| 26 | CE26–TCTA[6]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[6] | | | 0,045 | 0,015 |
| 27 | CE27–TCTA[4]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | | 0,005 | | 0,002 |
| 27 | CE27–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10] | 0,049 | | 0,010 | 0,020 |
| 27 | CE27–TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10] | | | 0,010 | 0,003 |
| 27 | CE27–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[9] | 0,005 | | | 0,002 |
| 28 | CE28–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | 0,137 | 0,072 | 0,217 | 0,143 |
| 28 | CE28–TCTA[5]TCTG[5]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[12] | | 0,005 | | 0,002 |
| 28 | CE28–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10] | | | 0,015 | 0,005 |
| 28 | CE28–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10] | 0,010 | | | 0,003 |
| 28.2 | CE28.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[8]TA[1]TATC[2] | | 0,026 | | 0,008 |
| 29 | CE29–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | 0,186 | 0,098 | 0,101 | 0,129 |
| 29 | CE29–TCTA[4]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | | | 0,005 | 0,002 |
| 29 | CE29–TCTA[5]TCTG[6]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[12] | 0,005 | | | 0,002 |
| 29 | CE29–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | | 0,005 | 0,081 | 0,029 |
| 29 | CE29–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | 0,049 | 0,149 | | 0,065 |
| 29.2 | CE29.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[9]TA[1]TATC[2] | | 0,010 | | 0,003 |
| 30 | CE30–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13] | 0,054 | 0,041 | 0,051 | 0,049 |
| 30 | CE30–TCTA[5]TCTG[6]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[13] | | 0,005 | | 0,002 |
| 30 | CE30–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | 0,005 | 0,067 | 0,061 | 0,044 |
| 30 | CE30–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | 0,132 | 0,082 | | 0,072 |
| 30 | CE30–TCTA[6]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | | | 0,030 | 0,010 |
| 30 | CE30–TCTA[7]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | 0,005 | 0,031 | | 0,012 |
| 30.2 | CE30.2–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2] | 0,005 | | | 0,002 |
| 30.2 | CE30.2–TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2] | 0,005 | | | 0,002 |
| 30.2 | CE30.2–TCTA[5]TCTG[6]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2] | | | 0,025 | 0,008 |
| 30.2 | CE30.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]TA[1]TATC[2] | 0,034 | 0,041 | 0,030 | 0,035 |
| 30.2 | CE30.2–TCTA[5]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]TA[1]TATC[2] | | 0,005 | | 0,002 |
| 30.3 | CE30.3–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCA[1]TCTA[2]TCCA[1]TATC[11] | | | 0,020 | 0,007 |

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 31 | CE31–TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[14] | | 0,010 | 0,010 | 0,007 |
| 31 | CE31–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13] | 0,054 | 0,046 | 0,020 | 0,040 |
| 31 | CE31–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13] | 0,054 | 0,031 | | 0,029 |
| 31 | CE31–TCTA[6]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | | | 0,015 | 0,005 |
| 31 | CE31–TCTA[6]TCTG[7]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[12] | | | 0,010 | 0,003 |
| 31 | CE31–TCTA[7]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | 0,005 | 0,010 | | 0,005 |
| 31 | CE31–TCTA[8]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | | 0,021 | | 0,007 |
| 31.2 | CE31.2–TCTA[5]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2] | 0,005 | 0,005 | | 0,003 |
| 31.2 | CE31.2–TCTA[5]TCTG[6]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2] | | | 0,005 | 0,002 |
| 31.2 | CE31.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2] | 0,083 | 0,062 | 0,005 | 0,050 |
| 31.2 | CE31.2–TCTA[5]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2] | | 0,010 | | 0,003 |
| 32 | CE32–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[14] | | 0,010 | | 0,003 |
| 32 | CE32–TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[14] | 0,005 | | | 0,002 |
| 32 | CE32–TCTA[6]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13] | | 0,005 | | 0,002 |
| 32.2 | CE32.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2] | 0,083 | 0,088 | 0,061 | 0,077 |
| 32.2 | CE32.2–TCTA[5]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]TA[1]TATC[2] | | | 0,005 | 0,002 |
| 33 | CE33–TCTA[10]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | | | 0,005 | 0,002 |
| 33.2 | CE33.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]TA[1]TATC[2] | 0,020 | 0,041 | 0,005 | 0,022 |
| 33.2 | CE33.2–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[3]TACC[1]TATC[9]TA[1]TATC[2] | | 0,005 | | 0,002 |
| 33.2 | CE33.2–TCTA[5]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]TA[1]TATC[2] | | | 0,005 | 0,002 |
| 34 | CE34–TCTA[10]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | | | 0,015 | 0,005 |
| 34 | CE34–TCTA[11]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11] | | | 0,051 | 0,017 |
| 34.1 | CE34.1–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[6]ATC[1]TATC[7]TA[1]TATC[2] | 0,005 | | | 0,002 |
| 34.2 | CE34.2–TCTA[5]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]TA[1]TATC[2] | | 0,005 | | 0,002 |
| 35 | CE35–TCTA[10]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13] | | | 0,005 | 0,002 |
| 35 | CE35–TCTA[10]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | | | 0,005 | 0,002 |
| 35 | CE35–TCTA[12]TCTG[4]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | | | 0,005 | 0,002 |
| 35.1 | CE35.1–TCTA[10]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]–19182101.1->T | | | 0,005 | 0,002 |
| 35.2 | CE35.2–TCTA[6]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[14]TA[1]TATC[2] | | 0,005 | | 0,002 |
| 36 | CE36–TCTA[11]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13] | | | 0,005 | 0,002 |
| 36 | CE36–TCTA[11]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12] | | | 0,020 | 0,007 |
| 36 | CE36–TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[10]ATC[1]TATC[3]ATC[1]TATC[1]TA[1]TATC[2] | | | 0,025 | 0,008 |

**FGA**

Total sequence alleles 34
Total CE fragment alleles 22
Alleles containing SNPs outside STR-motif 0

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 17 | CE17–AAGG[1]AAGA[1]AAGG[3]AGAA[9]AGAG[1]AAAA[1]AAGA[3] | | | 0,040 | 0,013 |
| 18 | CE18–AAGG[1]AAGA[1]AAGG[3]AGAA[10]AGAG[1]AAAA[1]AAGA[3] | | 0,077 | | 0,025 |
| 19 | CE19–AAGG[1]AAGA[1]AAGG[3]AGAA[11]AGAG[1]AAAA[1]AAGA[3] | 0,059 | 0,093 | 0,056 | 0,069 |
| 19.2 | CE19.2–AAGG[1]AAGA[1]AAGG[3]AGAA[13]AAAA[1]GA[1]AAGA[2] | 0,005 | | 0,005 | 0,003 |
| 20 | CE20–AAGG[1]AAGA[1]AAGG[3]AGAA[12]AGAG[1]AAAA[1]AAGA[3] | 0,108 | 0,046 | 0,051 | 0,069 |
| 21 | CE21–AAGG[1]AAGA[1]AAGG[3]AGAA[13]AGAG[1]AAAA[1]AAGA[3] | 0,176 | 0,052 | 0,086 | 0,106 |
| 21.2 | CE21.2–AAGG[1]AAGA[1]AAGG[3]AGAA[16]GA[1]AAGA[2] | | 0,005 | | 0,002 |
| 21.2 | CE21.2–AAGG[1]AAGA[1]AAGG[3]AGAA[15]AAAA[1]GA[1]AAGA[2] | 0,005 | 0,010 | | 0,005 |
| 22 | CE22–AAGG[1]AAGA[1]AAGG[3]AGAA[14]AGAG[1]AAAA[1]AAGA[3] | 0,196 | 0,103 | 0,288 | 0,196 |
| 22.1 | CE22.1–AAGG[1]AAGA[1]AAGG[3]AGAA[14][A[1]AGAG[1]AAAA[1]AAGA[3] | 0,005 | | | 0,002 |
| 22.2 | CE22.2–AAGG[1]AAGA[1]AAGG[3]AGAA[16]AAAA[1]GA[1]AAGA[2] | 0,005 | 0,031 | | 0,012 |
| 22.2 | CE22.2–AAGG[1]AAGA[1]AAGG[3]AGAA[14]AAAG[1]AGAA[1]AAAA[1]GA[1]AAGA[2] | | | 0,020 | 0,007 |
| 23 | CE23–AAGG[1]AAGA[1]AAGG[3]AGAA[15]AGAG[1]AAAA[1]AAGA[3] | 0,157 | 0,186 | 0,167 | 0,169 |
| 23.2 | CE23.2–AAGG[1]AAGA[1]AAGG[3]AGAA[17]AAAA[1]GA[1]AAGA[2] | 0,005 | 0,005 | | 0,003 |
| 24 | CE24–AAGG[1]AAGA[1]AAGG[3]AGAA[16]AGAG[1]AAAA[1]AAGA[3] | 0,137 | 0,180 | 0,126 | 0,148 |
| 24.2 | CE24.2–AAGG[1]AAGA[1]AAGG[3]AGAA[18]AAAA[1]GA[1]AAGA[2] | | 0,026 | | 0,008 |
| 24.2 | CE24.2–AAGG[1]AAGA[1]AAGG[3]AGAA[16]AAAG[1]AGAA[1]AAAA[1]GA[1]AAGA[2] | | | 0,010 | 0,003 |
| 25 | CE25–AAGG[1]AAGA[1]AAGG[3]AGAA[2]ACAA[1]AGAA[14]AGAG[1]AAAA[1]AAGA[3] | | 0,005 | | 0,002 |
| 25 | CE25–AAGG[1]AAGA[1]AAGG[3]AGAA[15]AGAG[1]AGAA[1]AGAG[1]AAAA[1]AAGA[3] | 0,005 | | | 0,002 |
| 25 | CE25–AAGG[1]AAGA[1]AAGG[3]AGAA[17]AGAG[1]AAAA[1]AAGA[3] | 0,093 | 0,077 | 0,051 | 0,074 |
| 25.2 | CE25.2–AAGG[1]AAGA[1]AAGG[3]AGAA[19]AAAA[1]GA[1]AAGA[2] | | 0,005 | | 0,002 |
| 26 | CE26–AAGG[1]AAGA[1]AAGG[3]AGAA[16]AGAG[1]AGAA[1]AGAG[1]AAAA[1]AAGA[3] | | 0,005 | | 0,002 |
| 26 | CE26–AAGG[1]AAGA[1]AAGG[3]AGAA[6]GGAA[1]AGAA[11]AGAG[1]AAAA[1]AAGA[3] | | | 0,005 | 0,002 |
| 26 | CE26–AAGG[1]AAGA[1]AAGG[3]AGAA[2]ACAA[1]AGAA[15]AGAG[1]AAAA[1]AAGA[3] | 0,010 | | | 0,003 |
| 26 | CE26–AAGG[1]AAGA[1]AAGG[3]AGAA[18]AGAG[1]AAAA[1]AAGA[3] | 0,029 | 0,046 | 0,040 | 0,039 |
| 26.3 | CE26.3–AAGG[1]AAGA[1]AAGG[3]AGAA[10]GAA[1]AGAA[8]AGAG[1]AAAA[1]AAGA[3] | | 0,005 | | 0,002 |
| 27 | CE27–AAGG[1]AAGA[1]AAGG[3]AGAA[17]AGAG[1]AGAA[1]AGAG[1]AAAA[1]AAGA[3] | | 0,015 | | 0,005 |
| 27 | CE27–AAGG[1]AAGA[1]AAGG[3]AGAA[6]GGAA[1]AGAA[12]AGAG[1]AAAA[1]AAGA[3] | | | 0,010 | 0,003 |
| 27 | CE27–AAGG[1]AAGA[1]AAGG[3]AGAA[2]ACAA[1]AGAA[16]AGAG[1]AAAA[1]AAGA[3] | 0,005 | | | 0,002 |
| 27 | CE27–AAGG[1]AAGA[1]AAGG[3]AGAA[19]AGAG[1]AAAA[1]AAGA[3] | | 0,010 | 0,015 | 0,008 |

Chapter 4

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 29 | CE29–AAGG[1]AAGA[1]AAGG[3]AGAA[6]GGAA[1]AGAA[14]AGAG[1]AAAA[1]AAGA[3] | | | 0,005 | 0,002 |
| 30 | CE30–AAGG[1]AAGA[1]AAGG[3]AGAA[20]AGAG[1]AGAA[1]AGAG[1]AAAA[1]AAGA[3] | | 0,005 | | 0,002 |
| 45.2 | CE45.2–AAGG[1]AAGA[1]AAGG[5]AGAA[3]AGGA[3]AGAA[13]AGAC[3]AGAA[14]AAAA[1]GA[1]AAGA[3] | | | 0,010 | 0,003 |

*This allele is longer than 300 bp and can only be completely analysed using paired-end data*

**PentaD**

Total sequence alleles 24
Total CE fragment alleles 18
Alleles containing SNPs outside STR-motif 6

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 2.2 | CE2.2–AAAGA[5]AAAAA[1]–43636192AAAGAAAAAAAAG>del | | | 0,202 | 0,067 |
| 3.2 | CE3.2–AAAGA[6]AAAAA[1]–43636192AAAGAAAAAAAAG>del | | | 0,005 | 0,002 |
| 5 | CE5–AAAGA[5]AAAAA[1] | | | 0,051 | 0,017 |
| 6 | CE6–AAAGA[6]AAAAA[1] | 0,005 | 0,005 | 0,076 | 0,029 |
| 7 | CE7–AAAGA[7]AAAAA[1] | | 0,031 | 0,015 | 0,015 |
| 8 | CE8–AAAGA[8]AAAAA[1] | 0,005 | 0,057 | 0,146 | 0,069 |
| 8.2 | CE8.2–AAAGA[11]AAAAA[1]–43636192AAAGAAAAAAAAG>del | | 0,005 | | 0,002 |
| 9 | CE9–AAAGA[9]AAAAA[1] | 0,211 | 0,211 | 0,217 | 0,213 |
| 9 | CE9–AAAGA[10] | | | 0,010 | 0,003 |
| 9 | CE9–AAAGA[9]AAAAA[1]–43636331T>G | 0,034 | | | 0,012 |
| 10 | CE10–AAAGA[10]AAAAA[1] | 0,088 | 0,103 | 0,081 | 0,091 |
| 10 | CE10–AAAGA[10]AAAAA[1]–43636331T>G | 0,015 | 0,005 | | 0,007 |
| 11 | CE11–AAAGA[11]AAAAA[1] | 0,127 | 0,227 | 0,071 | 0,141 |
| 11 | CE11–AAAGA[11]AAAAA[1]–43636331T>G | 0,005 | | | 0,002 |
| 12 | CE12–AAAGA[12]AAAAA[1] | 0,225 | 0,201 | 0,061 | 0,163 |
| 12 | CE12–AAAGA[13] | 0,005 | | | 0,002 |
| 12.2 | CE12.2–AAAGA[3]GA[1]AAAGA[9]AAAAA[1] | | | 0,015 | 0,005 |
| 13 | CE13–AAAGA[13]AAAAA[1] | 0,181 | 0,088 | 0,045 | 0,106 |
| 13 | CE13–AAAGA[14] | 0,005 | | | 0,002 |
| 14 | CE14–AAAGA[14]AAAAA[1] | 0,059 | 0,052 | 0,005 | 0,039 |
| 14.1 | CE14.1–AAAGA[3]A[1]AAAGA[11]AAAAA[1] | 0,005 | | | 0,002 |
| 15 | CE15–AAAGA[15]AAAAA[1] | 0,025 | 0,010 | | 0,012 |
| 16 | CE16–AAAGA[16]AAAAA[1] | | 0,005 | | 0,002 |
| 18 | CE18–AAAGA[18]AAAAA[1] | 0,005 | | | 0,002 |

**PentaE**

Total sequence alleles 19
Total CE fragment alleles 18
Alleles containing SNPs outside STR-motif 0

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 5 | CE5–TTTTC[5] | 0,074 | 0,026 | 0,101 | 0,067 |
| 7 | CE7–TTTTC[7] | 0,172 | 0,031 | 0,076 | 0,094 |
| 8 | CE8–TTTTC[8] | 0,005 | 0,005 | 0,268 | 0,092 |
| 9 | CE9–TTTTC[9] | 0,020 | 0,005 | 0,101 | 0,042 |
| 10 | CE10–TTTTC[10] | 0,078 | 0,015 | 0,086 | 0,060 |
| 11 | CE11–TTTTC[11] | 0,093 | 0,175 | 0,005 | 0,091 |
| 12 | CE12–TTTTC[10]TTTTA[1]TTTTC[1] | | 0,015 | | 0,005 |
| 12 | CE12–TTTTC[12] | 0,191 | 0,113 | 0,086 | 0,131 |
| 13 | CE13–TTTTC[13] | 0,098 | 0,052 | 0,136 | 0,096 |
| 14 | CE14–TTTTC[14] | 0,083 | 0,077 | 0,051 | 0,070 |
| 15 | CE15–TTTTC[15] | 0,034 | 0,093 | 0,051 | 0,059 |
| 16 | CE16–TTTTC[16] | 0,049 | 0,139 | 0,015 | 0,067 |
| 17 | CE17–TTTTC[17] | 0,064 | 0,093 | 0,020 | 0,059 |

**TH01**

Total sequence alleles 14
Total CE fragment alleles 8
Alleles containing SNPs outside STR-motif 5

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 5 | CE5–TGAA[5] | 0,010 | | | 0,003 |
| 6 | CE6–TGAA[6] | 0,201 | 0,046 | 0,056 | 0,102 |
| 7 | CE7–TGAA[1]TTAA[1]TGAA[5] | | | 0,005 | 0,002 |
| 7 | CE7–TGAA[7] | 0,196 | 0,258 | 0,404 | 0,285 |
| 7 | CE7–TGAA[7]–2171244C>T | | | 0,025 | 0,008 |
| 7 | CE7–TGAA[7]–2171200C>A | | 0,005 | | 0,002 |
| 8 | CE8–TGAA[8]–2171115G>T–2171244C>T | | | 0,015 | 0,005 |
| 8 | CE8–TGAA[8] | 0,127 | 0,098 | 0,116 | 0,114 |
| 8 | CE8–TGAA[8]–2171244C>T | | | 0,101 | 0,034 |
| 9 | CE9–TGAA[9] | 0,123 | 0,479 | 0,197 | 0,263 |
| 9 | CE9–TGAA[9]–2171244C>T | | | 0,010 | 0,003 |
| 9.3 | CE9.3–TGAA[6]TGA[1]TGAA[3] | 0,333 | 0,103 | 0,020 | 0,154 |
| 10 | CE10–TGAA[10] | 0,010 | 0,010 | 0,035 | 0,018 |
| 11 | CE11–TGAA[11] | | | 0,015 | 0,005 |

**TPOX**

Total sequence alleles 12
Total CE fragment alleles 7
Alleles containing SNPs outside STR-motif 5

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 6 | CE6–TGAA[6] | | | 0,091 | 0,030 |
| 8 | CE8–TGAA[8] | 0,461 | 0,443 | 0,273 | 0,393 |
| 8 | CE8–TGAA[8]–1489601G>T | 0,005 | 0,021 | | 0,008 |
| 8 | CE8–TGAA[8]–1489557G>A | | | 0,030 | 0,010 |
| 8 | CE8–TGAA[8]–1489556C>T | | | 0,121 | 0,040 |
| 9 | CE9–TGAA[9] | 0,113 | 0,129 | 0,152 | 0,131 |
| 9 | CE9–TGAA[9]–1489544C>A | | | 0,111 | 0,037 |
| 9 | CE9–TGAA[9]–1489557G>A | | | 0,015 | 0,005 |
| 10 | CE10–TGAA[10] | 0,064 | 0,015 | 0,076 | 0,052 |
| 11 | CE11–TGAA[11] | 0,333 | 0,371 | 0,121 | 0,275 |
| 12 | CE12–TGAA[12] | 0,020 | 0,021 | 0,010 | 0,017 |
| 13 | CE13–TGAA[13] | 0,005 | | | 0,002 |

Chapter 4

**vWA**

Total sequence alleles 24
Total CE fragment alleles 9
Alleles containing SNPs outside STR-motif 3

| Fragment allele | Sequence allele | Frequency NL | Frequency Nepal / Bhutan | Frequency Pygmies | Total Frequency |
|---|---|---|---|---|---|
| 11 | CE11–GATG[2]GATA[1]GATG[1]GATA[7]GACA[3]GATA[1] | | | 0,010 | 0,003 |
| 14 | CE14–GATG[4]GATA[3]GATG[1]GATA[3]GACA[4]GATA[1]GACA[1]GATA[1]–5984116A>T–5984121C>T–5984134T>C | 0,093 | 0,139 | 0,051 | 0,094 |
| 14 | CE14–GATG[4]GATA[3]GATG[1]GATA[2]GACA[5]GATA[1]GACA[1]GATA[1]–5984116A>T–5984121C>T–5984134T>C | | | 0,051 | 0,017 |
| 14 | CE14–GATG[2]GATA[1]GATG[1]GATA[10]GACA[3]GATA[1] | 0,010 | | 0,020 | 0,010 |
| 14 | CE14–GATG[2]GATA[11]GACA[4]GATA[1] | | | 0,015 | 0,005 |
| 14 | CE14–GATG[2]GATA[1]GATG[1]GATA[9]GACA[4]GATA[1] | | | 0,051 | 0,017 |
| 15 | CE15–GATG[5]GATA[3]GATG[1]GATA[3]GACA[4]GATA[1]GACA[1]GATA[1]–5984116A>T–5984121C>T–5984134T>C | 0,005 | | | 0,002 |
| 15 | CE15–GATG[2]GATA[1]GATG[2]GATA[10]GACA[3]GATA[1] | | | 0,010 | 0,003 |
| 15 | CE15–GATG[2]GATA[1]GATG[1]GATA[11]GACA[3]GATA[1] | 0,074 | | 0,045 | 0,040 |
| 15 | CE15–GATG[2]GATA[1]GATG[1]GATA[10]GACA[4]GATA[1] | 0,034 | 0,010 | 0,121 | 0,055 |
| 16 | CE16–GATG[2]GATA[1]GATG[1]GATA[12]GACA[3]GATA[1] | 0,034 | 0,026 | 0,086 | 0,049 |
| 16 | CE16–GATG[2]GATA[1]GATG[1]GATA[11]GACA[4]GATA[1] | 0,147 | 0,165 | 0,146 | 0,153 |
| 17 | CE17–GATG[2]GATA[1]GATG[1]GATA[13]GACA[3]GATA[1] | 0,015 | 0,005 | 0,045 | 0,022 |
| 17 | CE17–GATG[2]GATA[1]GATG[1]GATA[12]GACA[4]GATA[1] | 0,275 | 0,314 | 0,141 | 0,243 |
| 18 | CE18–GATG[2]GATA[1]GATG[1]GATA[14]GACA[3]GATA[1] | | | 0,025 | 0,008 |
| 18 | CE18–GATG[2]GATA[1]GATG[1]GATA[13]GACA[4]GATA[1] | 0,196 | 0,216 | 0,096 | 0,169 |
| 18 | CE18–GATG[2]GATA[1]GATG[1]GATA[12]GACA[5]GATA[1] | | 0,010 | | 0,003 |
| 18 | CE18–GATG[2]GATA[1]GATG[1]GATA[11]GACA[6]GATA[1] | | | 0,025 | 0,008 |
| 19 | CE19–GATG[2]GATA[1]GATG[1]GATA[15]GACA[3]GATA[1] | | 0,005 | | 0,002 |
| 19 | CE19–GATG[2]GATA[1]GATG[1]GATA[14]GACA[4]GATA[1] | 0,113 | 0,098 | 0,025 | 0,079 |
| 19 | CE19–GATG[2]GATA[1]GATG[1]GATA[13]GACA[5]GATA[1] | | | 0,005 | 0,002 |
| 20 | CE20–GATG[2]GATA[1]GATG[1]GATA[15]GACA[4]GATA[1] | 0,005 | 0,010 | 0,005 | 0,007 |
| 20 | CE20–GATG[2]GATA[1]GATG[1]GATA[14]GACA[5]GATA[1] | | | 0,005 | 0,002 |
| 22 | CE22–GATG[2]GATA[1]GATG[1]GATA[14]GACA[7]GATA[1] | | | 0,020 | 0,007 |

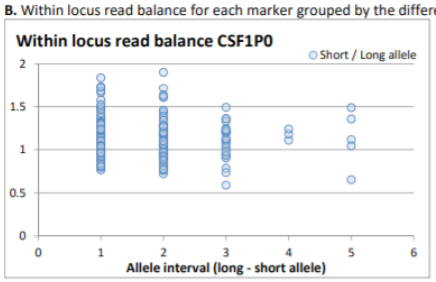
*For every locus the table displays the observed sequence alleles and respective allele frequencies for the three populations tested.*

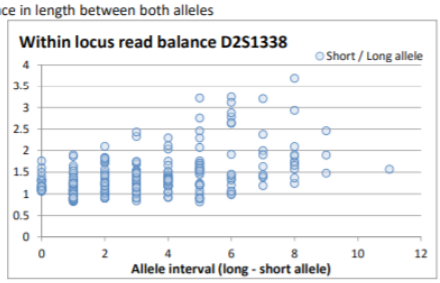
**Supplemental Figure 7.** Coverage and within locus allele balance for the samples used for stutter analysis

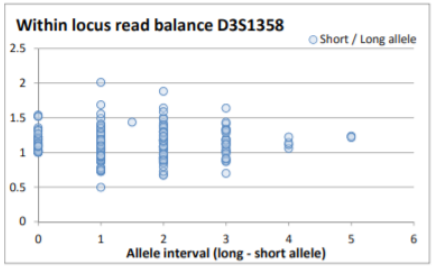
**A.** Average allele coverage and percentage of samples reaching the aimed allele coverage of 1000 reads

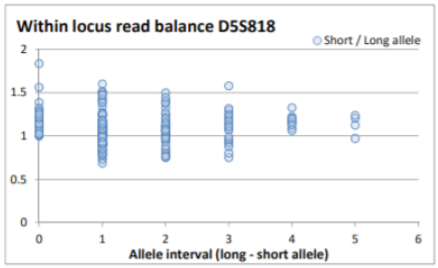
| Locus | Average allele coverage | Allele percentage coverage >1000 |
|---|---|---|
| CSF1P0 | 8512 | 100% |
| D2S1338 | 2596 | 86% |
| D3S1358 | 5390 | 100% |
| D5S818 | 7635 | 100% |
| D7S820 | 5354 | 90% |
| D8S1179 | 4416 | 92% |
| D13S317 | 7901 | 100% |
| D16S539 | 5279 | 96% |
| D18S51 | 5878 | 100% |
| D19S433 | 4476 | 99% |
| D21S11 | 6700 | 100% |
| FGA | 4024 | 98% |
| PentaD | 5271 | 99% |
| PentaE | 5411 | 93% |
| TH01 | 5743 | 98% |
| TPOX | 7074 | 100% |
| vWA | 4389 | 89% |

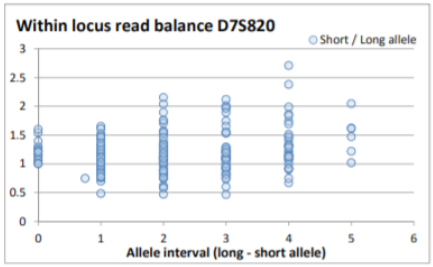
**B.** Within locus read balance for each marker grouped by the difference in length between both alleles

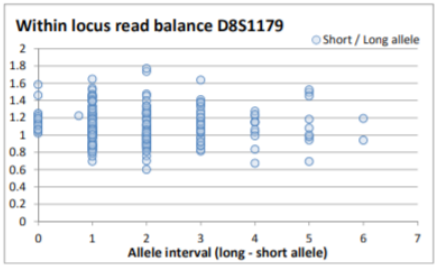


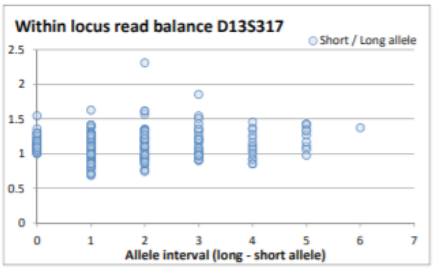



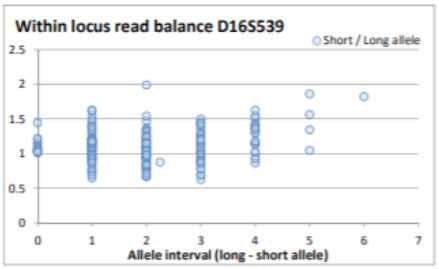



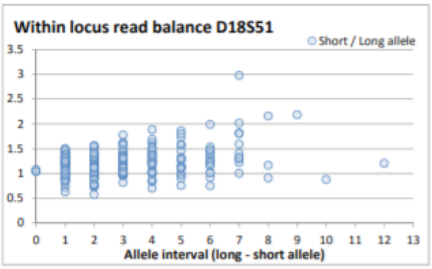



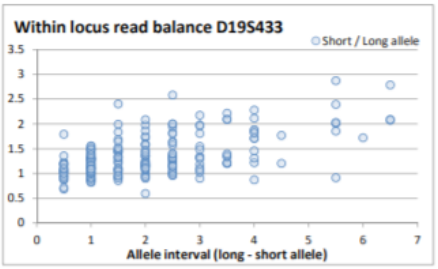



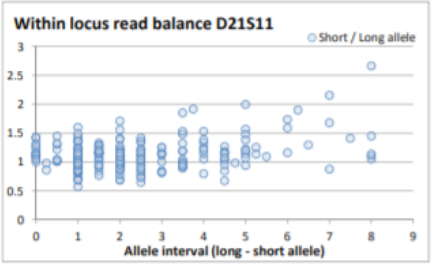
Chapter 4

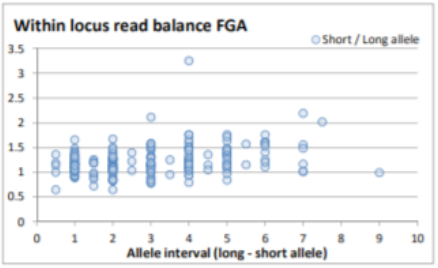

Within locus read balance D13S317
Short / Long allele
Allele interval (long - short allele)

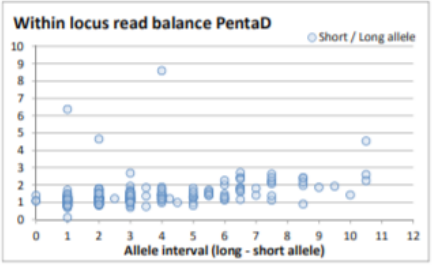

Within locus read balance D16S539
Short / Long allele
Allele interval (long - short allele)

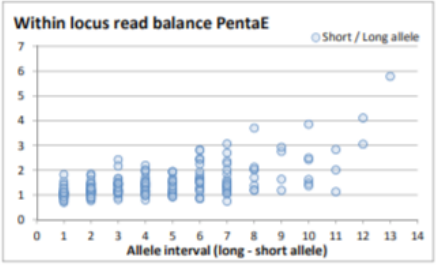

Within locus read balance D18S51
Short / Long allele
Allele interval (long - short allele)

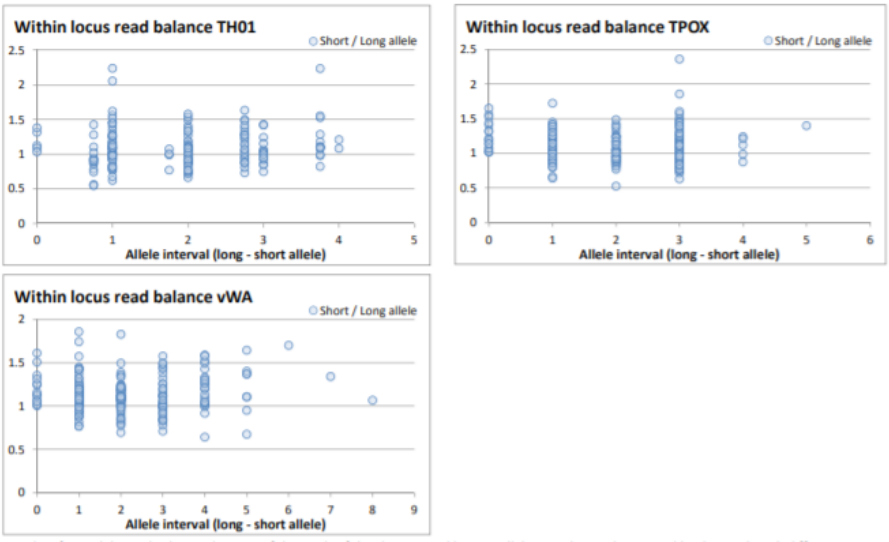

Within locus read balance D19S433
Short / Long allele
Allele interval (long - short allele)


Within locus read balance D21S11
Short / Long allele
Allele interval (long - short allele)


Within locus read balance FGA
Short / Long allele
Allele interval (long - short allele)

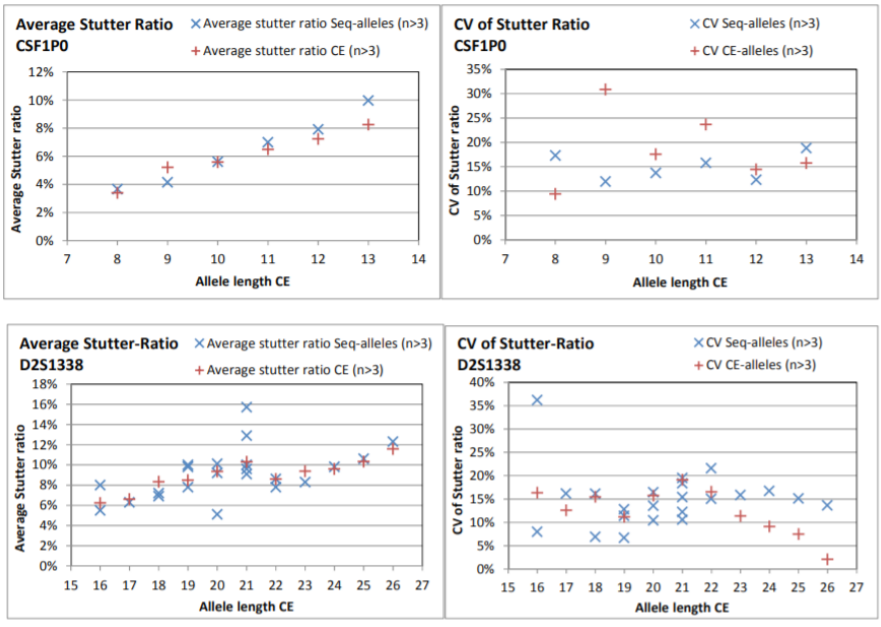

Within locus read balance PentaD
Short / Long allele
Allele interval (long - short allele)


Within locus read balance PentaE
Short / Long allele
Allele interval (long - short allele)

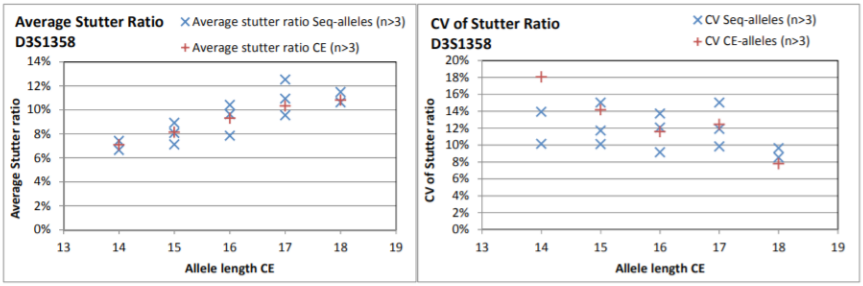


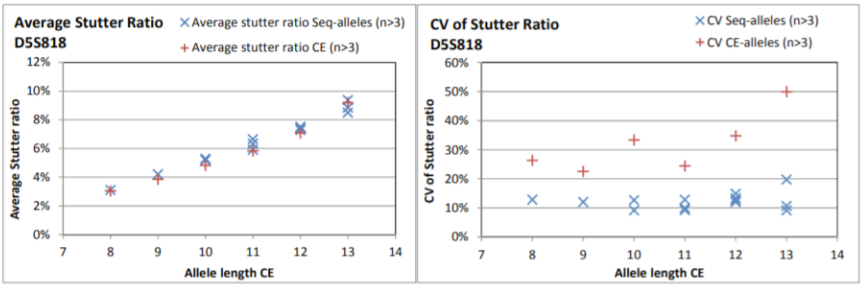

*Dot plots for each locus displaying the ratio of the reads of the shortest and longest allele in each sample grouped by the STR length difference of the two alleles. In general, the shorter allele has a slightly higher number of reads than the longer allele (on average 1.2 times higher). For some loci, large length differences between the two alleles can result in stronger within marker allele inbalance.*

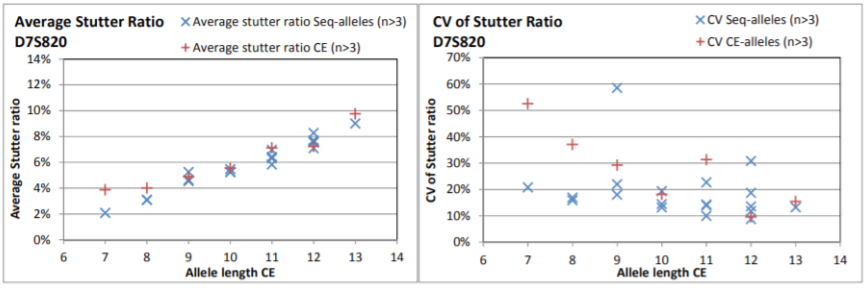
**Supplemental Figure 8.** Stutter characteristics for the 17 STRs of the prototype Powerseq™ system and the Powerplex® Fusion system

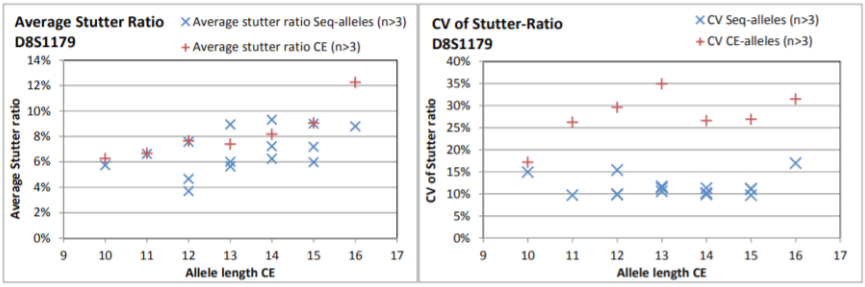


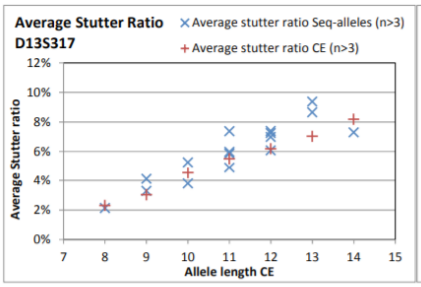
Chapter 4

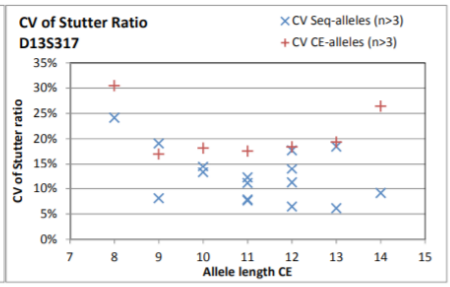

Average Stutter Ratio
D3S1358
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE
CV of Stutter Ratio
D3S1358
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE

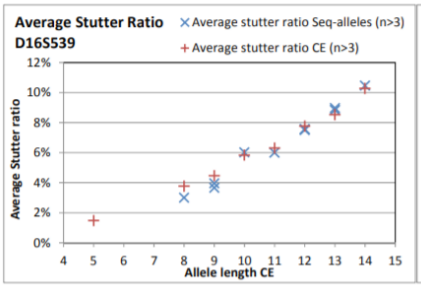

Average Stutter Ratio
D5S818
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE
CV of Stutter Ratio
D5S818
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE

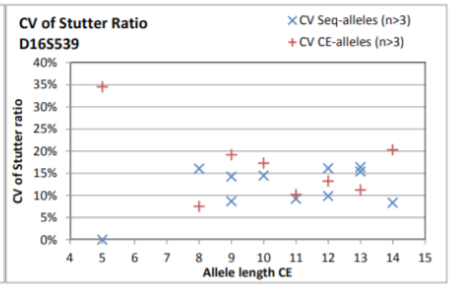

Average Stutter Ratio
D7S820
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE
CV of Stutter Ratio
D7S820
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE

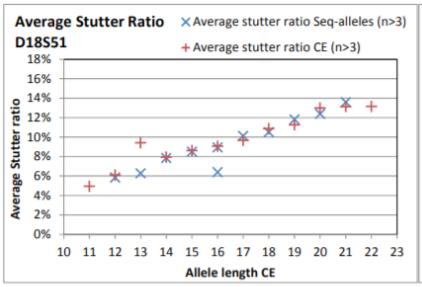

Average Stutter Ratio
D8S1179
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE
CV of Stutter-Ratio
D8S1179
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE

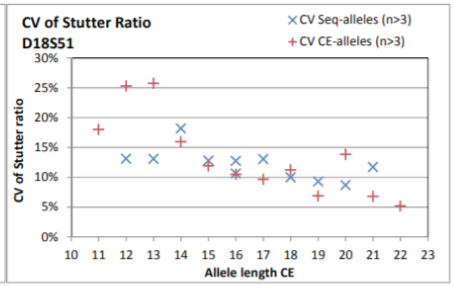

Average Stutter Ratio
D13S317
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE

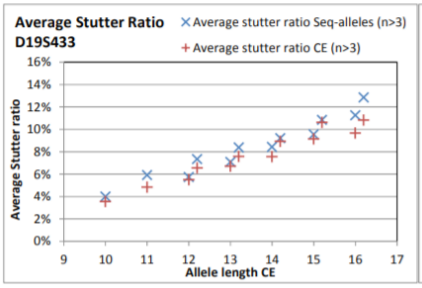

CV of Stutter Ratio
D13S317
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE

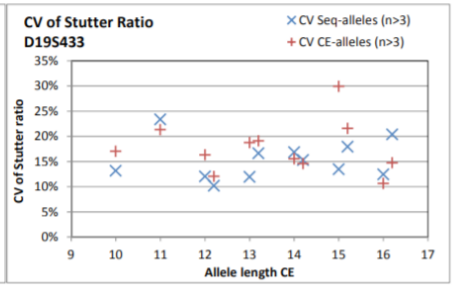

Average Stutter Ratio
D16S539
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE


CV of Stutter Ratio
D16S539
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE

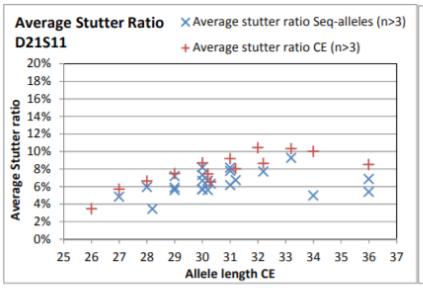

Average Stutter Ratio
D18S51
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE

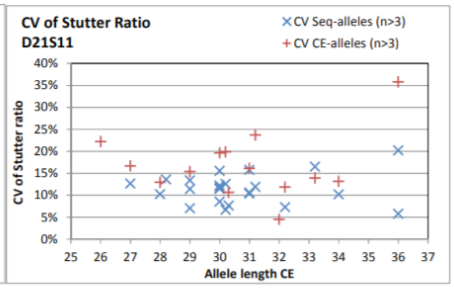

CV of Stutter Ratio
D18S51
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE

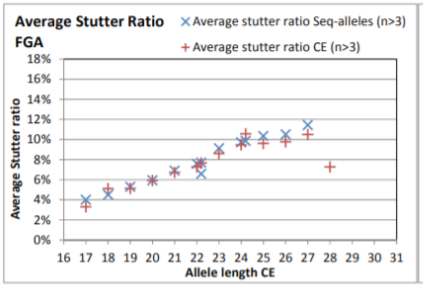

Average Stutter Ratio
D19S433
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE

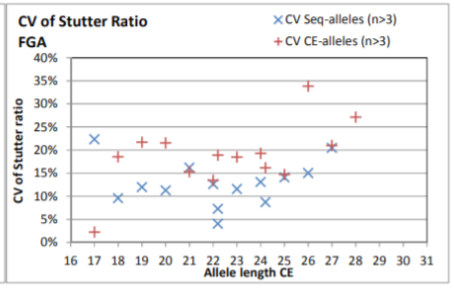

CV of Stutter Ratio
D19S433
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE

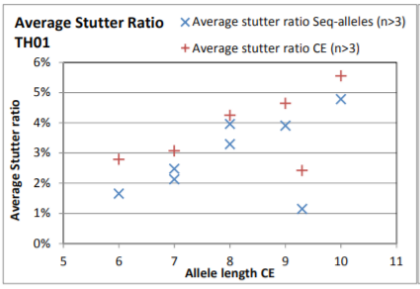
Chapter 4

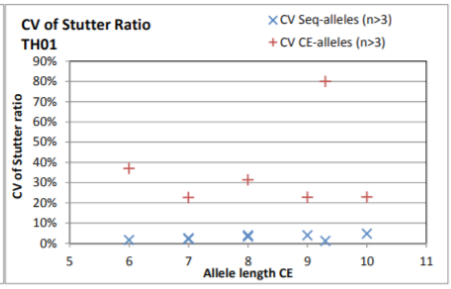

Average Stutter Ratio
D21S11
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE

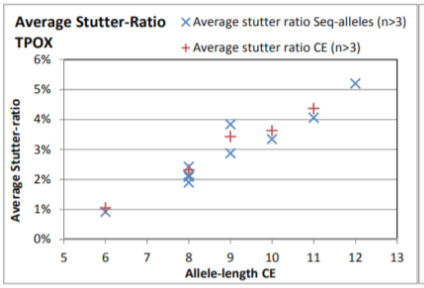

CV of Stutter Ratio
D21S11
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE

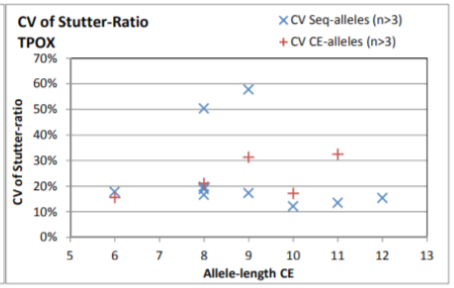

Average Stutter Ratio
FGA
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE

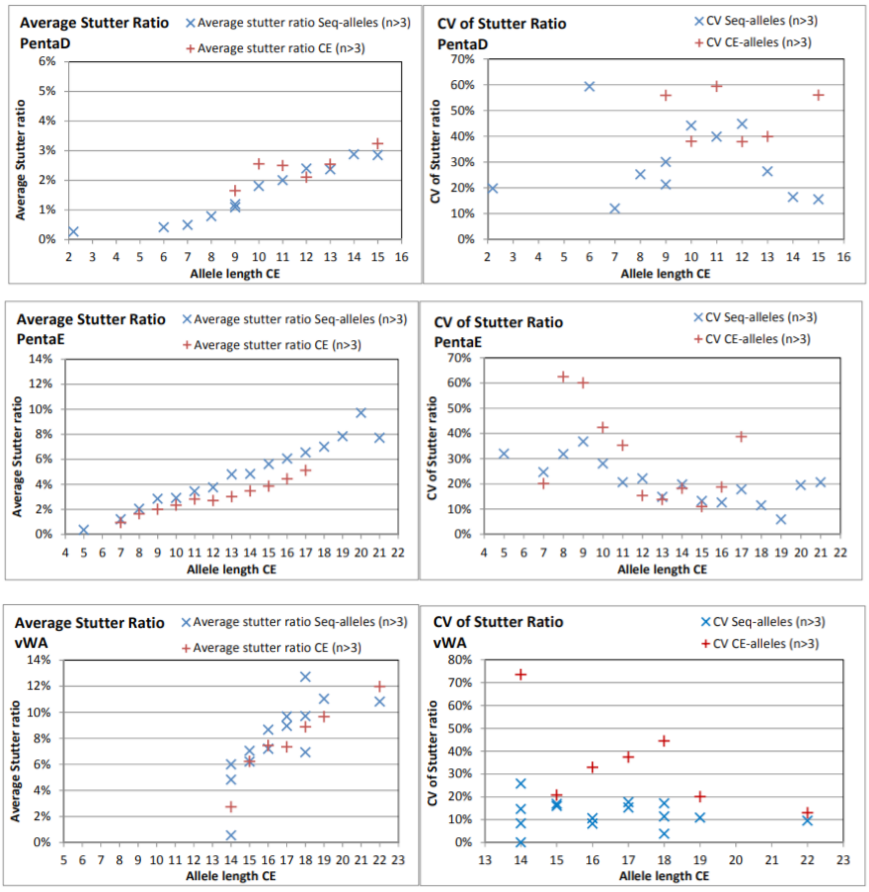

CV of Stutter Ratio
FGA
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE


Average Stutter Ratio
TH01
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter ratio
Allele length CE


CV of Stutter Ratio
TH01
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter ratio
Allele length CE


Average Stutter-Ratio
TPOX
Average stutter ratio Seq-alleles (n>3)
Average stutter ratio CE (n>3)
Average Stutter-ratio
Allele-length CE


CV of Stutter-Ratio
TPOX
CV Seq-alleles (n>3)
CV CE-alleles (n>3)
CV of Stutter-ratio
Allele-length CE




*Comparison of stutter characteristics for the MPS-based Powerseq™ system and the CE-based Powerplex® Fusion system. For every marker, two graphs display the average stutter ratio for every allele and the Coefficient of Variance (CV) for the stutter ratio of each allele. Since for MPS-based analysis some alleles of the same CE-length are represented by distinct sequence alleles, the stutter ratios and CV of these alleles are displayed separately. It is apparent that in general, stutter ratios of the MPS-based analysis are similar to the CE-based stutter ratios except for CE-alleles that are subdivided into several sequence alleles. The subdivided sequence alleles of the same total CE-length often have different stutter ratios which results in a lower variation of stutter ratio per allele compared to the combined CE-allele. This can be observed from the generally lower values for the CV of the stutter ratio for the sequence alleles.*

Chapter 4

**Supplemental Figure 9.** A three locus hypothetical example illustrating our method for the calculation of the proportion of a minor contribution in a mixture

**A.** The sequence read profiles for three loci of a single two-person mixture




**B.** Summary tables showing the read numbers of all unique alleles and stutters that passed all quality filters

**Locus 1**

| Alleles | Forward strand reads | Reverse strand reads | All reads combined | Proportion of total locus | Interpretation |
|---|---|---|---|---|---|
| 21 | 16 | 14 | 30 | 0.0090 | stutter |
| 22 | 760 | 740 | 1500 | 0.4478 | major |
| 23 | 28 | 32 | 60 | 0.0179 | stutter |
| 24 | 710 | 690 | 1400 | 0.4179 | major |
| 25 | 13 | 17 | 30 | 0.0090 | stutter |
| 26 | 74 | 76 | 150 | 0.0448 | minor |
| 27 | 15 | 15 | 30 | 0.0090 | stutter |
| 28 | 64 | 66 | 130 | 0.0388 | minor |
| 29 | 8 | 12 | 20 | 0.0060 | stutter |
| **Total** | 1688 | 1662 | **3350** | | |

**Read interpretation of Locus 1**
(Please note that in this example we only consider +1 or -1 stutters)

None of the possible minor alleles overlap with stutters or with major alleles

*We use both minor allele (26 and 28) reads for calculation*

**Locus 2**

| Alleles | Forward strand reads | Reverse strand reads | All reads combined | Proportion of total locus | Interpretation |
|---|---|---|---|---|---|
| 11 | 22 | 18 | 40 | 0.0118 | stutter |
| 12 | 135 | 145 | 280 | 0.0824 | minor |
| 13 | 24 | 26 | 50 | 0.0147 | stutter |
| 14 | 1510 | 1490 | 3000 | 0.8824 | major |
| 15 | 14 | 16 | 30 | 0.0088 | stutter |
| **Total** | 1705 | 1695 | **3400** | | |

**Read interpretation of Locus 2**
(Please note that in this example we only consider +1 or -1 stutters)

One minor allele (12) is visible. There is either only one (homozygous) minor allele, or the second minor allele overlaps the major. Only clear heterozygous minor alleles are used for the calculation which is not the case here

*Reads from this minor allele are not used for calculation*

**Locus 3**

| Alleles | Forward strand reads | Reverse strand reads | All reads combined | Proportion of total locus | Interpretation |
|---|---|---|---|---|---|
| 14 | 1 | 4 | 5 | 0.0017 | stutter |
| 15 | 50 | 60 | 110 | 0.0369 | minor |
| 16 | 4 | 6 | 10 | 0.0034 | stutter |
| 18 | 110 | 140 | 250 | 0.0838 | stutter |
| 19 | 680 | 720 | 1400 | 0.4690 | major |
| 20 | 590 | 610 | 1200 | 0.4020 | major |
| 21 | 5 | 5 | 10 | 0.0034 | stutter |
| **Total** | 1440 | 1545 | **2985** | | |

**Read interpretation of Locus 3**
(Please note that in this example we only consider +1 or -1 stutters)

Because of the high read count of allele 18, this allele was interpreted as a minor allele + stutter. the stutter influences the read-count of this allele and allele 18 can therefore not be used for a reliable calculation of the proportion of minor contribution. Allele 15 remains as one of the heterozygous minor alleles and can be included in the calculation of the proportion of minor contribution.

*Only allele 15 is used for the calculation*

**C.** Calculation of the proportion of the minor contribution

Average minor allele proportion: (0.0448 + 0.0388 + 0.0369) / 3 = 0,041
Total minor contribution in profile: 0.0401 * 2 * 100(%) = 8%

# 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

Forensic Science Intenational Genetics 2017 Mar;27:27-40
Published online November 26, 2016

Jerry Hoogenboom
Kristiaan J van der Gaag
Rick H de Leeuw
Titia Sijen
Peter de Knijff
Jeroen F.J. Laros

Supplementary material is available via https://www.fsigenetics.com

# Abstract

Massively parallel sequencing (MPS) is on the advent of a broad scale application in forensic research and casework. The improved capabilities to analyse evidentiary traces representing unbalanced mixtures is often mentioned as one of the major advantages of this technique. However, most of the available software packages that analyse forensic short tandem repeat (STR) sequencing data are not well suited for high throughput analysis of such mixed traces. The largest challenge is the presence of stutter artefacts in STR amplifications, which are not readily discerned from minor contributions. FDSTools is an open-source software solution developed for this purpose. The level of stutter formation is influenced by various aspects of the sequence, such as the length of the longest uninterrupted stretch occurring in an STR. When MPS is used, STRs are evaluated as sequence variants that each have particular stutter characteristics which can be precisely determined. FDSTools uses a database of reference samples to determine stutter and other systemic PCR or sequencing artefacts for each individual allele. In addition, stutter models are created for each repeating element in order to predict stutter artefacts for alleles that are not included in the reference set. This information is subsequently used to recognise and compensate for the noise in a sequence profile. The result is a better representation of the true composition of a sample. Using Promega Powerseq™ Auto System data from 450 reference samples and 31 two-person mixtures, we show that the FDSTools correction module decreases stutter ratios above 20% to below 3%. Consequently, much lower levels of contributions in the mixed traces are detected. FDSTools contains modules to visualise the data in an interactive format allowing users to filter data with their own preferred thresholds.

# Introduction

Analysis of Short Tandem Repeats (STRs) has been a successful forensic tool in the past two decades. The comparison of STR profiles from forensic DNA evidentiary traces with reference samples and DNA databases has provided essential information in many forensic cases. [1] Standard practice is to use Capillary Electrophoresis (CE) to analyse STR length variation. In recent years, Massively Parallel Sequencing (MPS) was introduced as a new method to analyse STRs and other forensic DNA markers [2,3]. MPS enables the simultaneous detection of both length and sequence variation of STRs, which increases the discriminatory value substantially [4,5,6]. The output of CE consists of peaks reflecting fluorescent signal intensities with their own respective shapes and peak heights. The output of MPS data analysis consists simply of read counts of the observed sequences. Both methods can suffer from the occurrence of PCR artefacts such as STR stutters [7]. This especially complicates the analysis of STR profiles coming from multiple contributors, which is common in forensic evidentiary traces. [8] The

level of stutter formation depends on a number of distinct aspects of the sequence, including the A/T content of the repeat unit and the number of consecutive repeat units occurring in an STR [9]. Since any specific STR length identified by CE can consist of multiple different sequences, these CE-identified length variants show a larger variation in measured stutter percentage than individual sequences analysed through MPS. This decreased variation in stutter percentage for MPS STR data may aid in the interpretation of mixtures [2], as it allows for a better prediction of stutter behaviour, which can be used to filter the data for stutter products. Existing software packages for the analysis of STR sequencing data [10,11,12] do not support extensive filtering and correction of systemic PCR and/or sequencing errors and therefore seem less suited for analysis of mixed DNA samples. This prompted us to develop a software package that harbours the following features: 1) characterisation and correction of noise in the sequencing data caused by PCR stutter or other systemic PCR and/or sequencing errors; 2) visualisation of sequencing data as comprehensive profiles; 3) filtering of data in graphs and tables with user definable thresholds and 4) open-source accessibility. Forensic DNA Sequencing Tools (FDSTools) is available via the Python Package Index (either by manual installation or by using the command 'pip install fdstools'). We assess the performance of FDSTools on 31 two-person mixtures genotyped via the Promega Powerseq™ Auto System for which we first generated a reference dataset of 450 samples.

# Material and Methods

## Sample preparation

PCR products and sequencing libraries were prepared as described previously [2] using a prototype Promega Powerseq™ Auto System containing 23 STRs and amelogenin. A set of 450 Dutch samples [13] and 31 two-person mixtures were amplified and sequenced. The mixtures consisted of three combinations of two donors selected randomly from a pool of unrelated individuals, which were mixed in different ratios. The minor components in the mixtures contributed 0.5% (six mixtures), 1% (six mixtures), 5% (four mixtures), 10% (six mixtures), 20% (six mixtures) and 50% (three mixtures).

Since the mixtures were used to test the performance of the software and also to determine analysis thresholds that are fit for purpose, we balanced the influence of varying DNA inputs in the PCR and increased drop-out due to low DNA input. This was achieved by the use of a minimum of the minor component of 60 pg in the 0.5%, 1% and 5% mixtures, resulting in a total DNA input of 12 ng, 6 ng and 1.2 ng, respectively (60 pg resulted in less than 20% drop-out in the validation of Powerplex 6C [14]). The same total DNA input of 1.2 ng was used for the 5%, 10%, 20% and

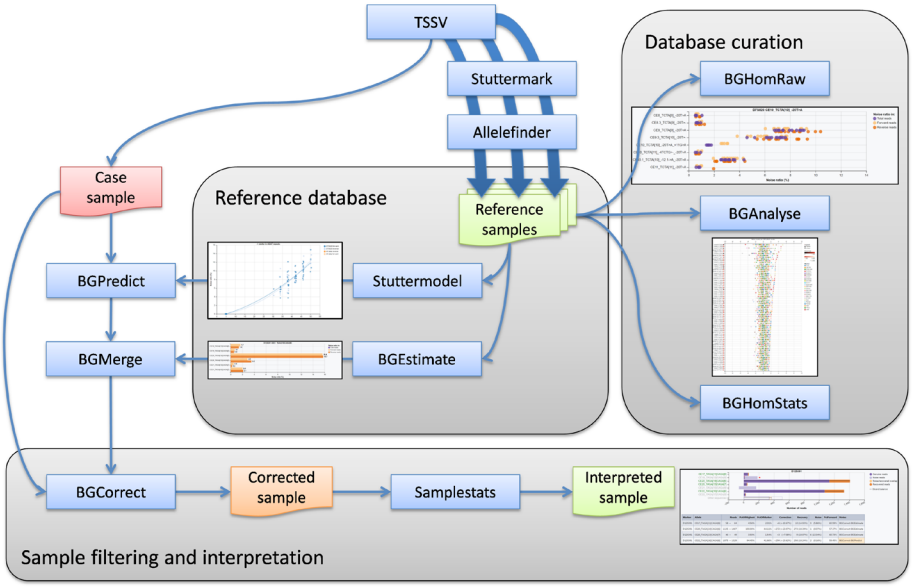
50% mixtures (resulting in 120 pg and 240 pg of the minor components in the 10% and 20% mixtures, respectively). The DNA input was 0.5 ng for single donor samples.

The genotypes of the donors used in the mixtures were known, which enables the identification of drop-in and drop-out allele calls. Paired-end sequencing data of all amplicons was generated using the MiSeq® Sequencer (Illumina).

## Initial data processing

In Figure 1, the main tools of the FDSTools package and their role in the data analysis pipeline are displayed. The tools can be split into three functional groups: tools for reference database creation, tools for reference database curation (data quality assessment) and tools for case sample filtering and data interpretation. In addition, the package contains initial data processing tools such as TSSV [10] that are common to reference database samples and case samples.

**Figure 1.** Flow chart of the analysis process, showing the main tools of FDSTools




*Flow chart showing the main tools (blue rectangles) of the FDSTools package and their roles in the data analysis pipeline. The output of each tool can be visualised using the Vis tool (not shown).*

## Paired-end read merging

Using paired-end sequencing, forward and reverse strand molecules of each amplicon were sequenced from both ends. The first ~300 nucleotides from either end were obtained. These read pairs were merged into a consensus read by aligning the read pair such that the largest possible overlap is obtained while allowing for up to 33% mismatches in the overlapped region. Most amplicons were about 300 base pairs in length and provided fully complementary read pairs.

With STR amplicons that are longer than 300 bp, a problem may occur when both reads end in the middle of the STR structure and the pair may be merged into a truncated STR sequence. A modified version of FLASH 1.2.11 [15] (available via github.com/Jerrythafast/FLASH-lowercase-overhang) was used to mark the bases that were not in the overlapped region in lower case in the consensus read. This enables detection of truncated STR sequences in downstream analysis.

## Linking reads to loci and alleles

The merged reads are linked to specific loci and alleles by the TSSV tool, which is a wrapper around a simplified version of the TSSV [10] program called TSSV-Lite. TSSV links reads to loci by scanning the reads for the sequences flanking the STR loci used. The flanking sequences of each locus, that usually represent the most 5' nucleotides of the primers, are provided to FDSTools in a library file, together with various other details about the loci used. Supplementary File 1 represents the library file used in this study. The file contains a description for the contents of each section.

Each read is scanned for these flanking sequences by computing alignments. In this study, the flanking sequences were 18 nucleotides in length and two substitutions (or two inserted or deleted bases) per flank were allowed in the alignment. Reads are categorised as 'unrecognised' if no flanking sequence is found. Furthermore, both flanking sequences are required to have at least one upper case letter, which ensures that overlapped reads that are potentially truncated are categorised as 'unrecognised' as well. Reads in which only one flanking sequence is found with at least one upper case letter get linked to a locus but flagged as 'no start' or 'no end' depending on whether the left or right flank is missing, respectively (optionally, these reads can be written to separate fasta or fastq files).

The main output of TSSV is a text file with tab-separated values. The file contains one line for every unique sequence of each locus. The columns include the name of the locus, the sequence, and the number of reads carrying this particular sequence. Read counts are given separately for the forward and reverse strand.

TSSV includes additional options for filtering sequences that are seen too few times and sequences with a length outside a given range (e.g., primer-dimers). This range can be specified separately for each locus. Furthermore, filtered sequences can be aggregated into a single 'other sequences' category for each locus. In this project, only singletons (i.e., sequences with only one read) were aggregated to the 'other sequences' category.

## Building a reference database

One function of FDSTools is the building of a reference database. Such a database can be used to obtain estimates of recurring allele-specific systemic noise. Here, 'noise' refers to the complete collection of sequences observed in a sample, except the sample's true allelic sequences. Noise includes any artefact deriving from the PCR as well as the sequencing (such as PCR stutter or single-nucleotide errors). Additionally, based on the reference data a statistical model can be derived that aims to predict stutter ratios for alleles not present in the reference set.

The creation of a reference database involves various tools included in the FDSTools package, which will be discussed in the next sections. In addition to these separate tools, FDSTools offers the Pipeline tool, which conveniently integrates the entire data analysis pipeline. Users are advised to use Pipeline as it removes the complexity of having to run several separate tools and to combine their output. Pipeline takes a simple configuration file containing the analysis parameters and automatically runs the appropriate tools.

Building a reference database is a two-phase process. In the first phase, the reference samples are analysed in a global manner to identify their alleles and reject those samples in which the alleles are not readily identified. In the second phase, the systemic noise of each of these alleles is analysed in detail.

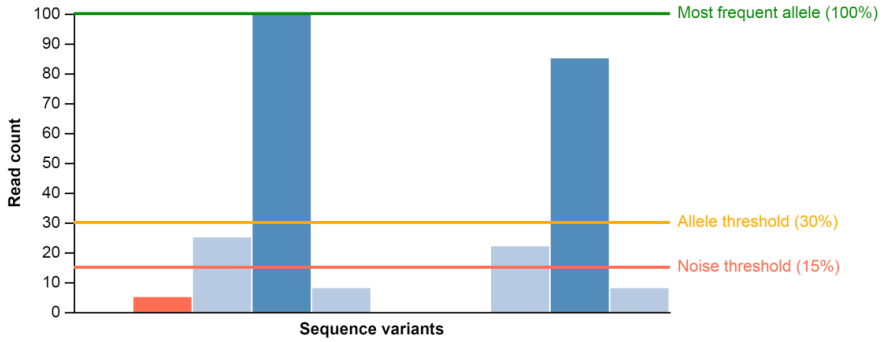
## Allele calling for reference samples

Determining the alleles of single donor reference samples is a fairly straightforward process because these generally represent the one or two most abundant sequences for any locus. FDSTools includes Allelefinder to call alleles this way. It is applied after Stuttermark, which is described below. A number of thresholds are used to guard against including alleles of potential low-level contaminations, which are outlined in Figure 2. For heterozygous loci, a second allele is only called if it passes the allele threshold, which is defined in this project as 30% of the read count of the most frequent allele at the same locus. As we expect no stutter above 30% [2], this threshold separates the alleles from noise. No alleles are called at a locus when additional sequences occur that have a read count below the allele threshold but above the noise threshold (which is

Chapter 5

defined as 15% of the most frequent allele in this project) or if a third sequence passes the allele threshold. If more than two loci in the same sample fail to give a result for these reasons, the overall quality of the sample is considered too poor to report any alleles. Additionally, Allelefinder can be configured to call at most one allele at haploid loci.

**Figure 2.** Thresholds used by Allelefinder to call alleles in reference samples




*Sequence variants with a read count above the allele threshold are called as alleles. The four lighter-shaded bars represent stutter variants (as recognised by Stuttermark), which are ignored by Allelefinder.*

The three potential pitfalls are 1) PCR stutter artefacts that exceed the noise threshold; 2) strong read count imbalance for heterozygous alleles, which may be the result of e.g., primer-site sequence variants and 3) autosomal trisomy, which is rare. To deal with the problem of stutter, each sample was analysed with Stuttermark [2] before calling alleles. With Stuttermark, sequences that are in a stutter position of another sequence while having a read count below a user-supplied percentage with respect to the other sequence are marked as 'stutter'. Sequences that have a read count that is too high to be explained by stutter alone will not be marked as 'stutter', as they may coincide with a genuine allele. The thresholds used here were 30% for −1 stutter (loss of a repeat unit) and 10% for +1 stutter (gain of a repeat unit). For −2 stutter products, a 30% threshold of the −1 stutter product is used. Sequences that are marked as 'stutter' are completely ignored by Allelefinder.

Allelefinder produces the list of alleles and a report detailing for which samples and loci allele calling is rejected and for which reasons.

## Estimating average allele-specific systemic noise

For each allele, a profile of recurring systemic noise, including PCR stutter products as well as any other 'side products', can be generated based on the reference data.

Noise profiles are always computed separately for forward and reverse reads, because strand bias may exist in the sequencing technology used. Profiles are also computed separately for each locus, under the assumption that noise production is not influenced by alleles of other loci. The level of noise is expressed as the number of noise reads as a percentage of the number of reads of the parent allele. In the context of PCR stutter analysis, this quantity is often referred to as the 'stutter ratio', despite the representation as a percentage of the parent allele. We use the generalised term 'noise ratio' (also represented as a percentage of the parent allele) to account for all other systemic noise as well.

$$\text{Noise ratio} = \frac{\text{Noise reads}}{\text{Allele reads}} \times 100\%$$

In homozygous samples, the noise ratio can be calculated by dividing the number of reads of a non-allelic sequence by the number of reads of the allele. Allele-specific noise profiles are readily computed from homozygous samples carrying this allele by scaling the read counts in each sample such that the parent allele is 100 and averaging the noise ratios for each noise sequence. These per-allele noise statistics and other statistics, such as the standard deviations of the noise ratios can be obtained using BGHomStats. In heterozygous samples the extraction of noise sequences is more complex, because it has to be determined which proportions each allele contributed to the observed noise sequences. We assume that noise in heterozygous samples corresponds to the sum of the noise profiles of the two alleles, after the application of a scaling correction to account for differences in the amount of each allele amplified. This is needed as even for heterozygous allele pairs, PCR efficiency may vary due to primer binding site sequence variation or STR length. [16] To extract noise from heterozygous reference samples an iterative approach was taken and implemented in the BGEstimate tool in FDSTools.

In essence, the algorithm, which is discussed in more detail in Supplementary Text 1, seeks a non-negative least squares solution to the matrix equation $A\,P = C$. In this equation, C is an $N \times M$ matrix of constants derived from the read counts in the reference samples, A is an $N \times N$ matrix summarising the allele balance in the samples, and P is an $N \times M$ matrix containing the estimated profiles of systemic noise. N is the number of unique genuine alleles among the reference samples and thus also the number of profiles produced and M is the total number of unique sequences observed.

Matrix C is computed once at the start of the algorithm. Each row in C corresponds to one allele and contains the sum of the read counts of all samples that have that particular allele, after scaling the allele to 100 reads for homozygous samples and 50 reads for heterozygotes. The noise profiles in P are initialised with the assumption that no systemic noise is present, i.e., all elements are set to 0, except for the elements that correspond to the actual alleles, which are set to 100.

Chapter 5

The algorithm then proceeds by repeatedly re-estimating the allele balance matrix A while reading cross-contributions between the alleles from the current profiles P and subsequently re-estimating P by finding a non-negative least squares optimal solution to A P = C. The values thus obtained in P are the average noise ratios of all observed systemic noise for all alleles (i.e., each row in P contains the noise profile of one allele).

To avoid noise from one allele being incorporated in the noise profile of another allele, a minimum of three different heterozygous genotypes per allele was used in this study. A threshold can be set for the minimal read count of noise to consider and the minimal percentage (we used 80%) of reference samples with the same allele which should contain the same noise before it is included in the noise profile. Each of these parameters can be set using various options of the 'fdstools bgestimate' command.

Relating the amount of stutter to repeat length

With the methods outlined above, profiles of systemic noise were obtained for each allele present in the reference set. However, one would also like to be able to filter and correct the noise originating from alleles that are not (yet) included in the reference set, as case samples may be encountered that contain alleles for which no reference sample was available. For this purpose, we developed a method to predict the sequence and corresponding amount of PCR stutter artefacts that would be produced for any allele of a given locus. Note that this method does not predict noise other than noise resulting from STR stutter or single nucleotide stretches.

Previous studies have shown that the amount of stutter is strongly correlated with the length of the repeated sequence [17] and even more so with the number of consecutive repeat units [2,18]. The FDSTools tool Stuttermodel seeks to fit polynomial functions to the repeat length and stutter ratio in homozygous reference samples. Stuttermodel scans each of the alleles for all positions where a particular repeat unit (e.g., the sequence 'AGAT') is repeated and records the length of this repeat, as the number of nucleotides, including incomplete repeats at the beginning or end of the repeated stretch. For each sample with this allele, the number of noise reads that lack exactly one repeat is counted. Reads that combine the loss of one repeat with one or more other differences (e.g., substitutions, or stutter in another stretch of repeats in the same allele) are included in this count. The counts thus obtained are used to compute the noise ratios of individual stutter sites and a polynomial function is fitted to quantify the relationship between the length of the repeat and the stutter ratio.

This analysis is repeated for each unique repeat unit of a length between one and a configurable maximum number of nucleotides (inclusive), treating cyclically equivalent units (e.g., 'ATAG' and 'AGAT') and their respective reverse complements (e.g., 'CTAT' and 'ATCT') synonymously. The amount of +1 stutter, −2 stutter etc. is analysed the same way.

Because different loci behave different in stutter formation, a separate function is fitted for each locus. Additionally, a polynomial function is fitted to all data at once,

which is used to predict stutter in alleles of loci for which insufficient reference data was available to fit a locus-specific function. Separate functions are fitted for the forward and reverse strands.

For each fitted function, Stuttermodel also determines the lower bound of the repeat length for which the function gives meaningful results. This lower bound is defined as the lowest repeat length for which the function produces a nonnegative result and the function is non-decreasing. Below this threshold, and in any other points where the function value would be negative, the function value is set to zero.

The quality of fit is assessed by computing the coefficient of determination,

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where

$$\hat{y}_i = \begin{cases} f_i & f_i \geq 0 \\ 0 & f_i < 0 \end{cases}$$

with $y_i$ the noise ratios of the reference samples, $\bar{y}$ the mean, $f_i$ the polynomial function's estimate of the noise ratio of sample $i$, and $\hat{y}_i$ the modified function value. The $R^2$ score will be close to one when the function is a good fit and lower otherwise.

Stuttermodel supports fitting polynomial functions of any degree. To prevent overfitting while still allowing a non-linear relationship, second-degree polynomials (with a minimum $R^2$ score) were used. In cases where the fit for one strand has an $R^2$ score above the threshold while the fit for the other strand scores below the threshold, both fits are rejected to prevent unintended introduction of strand bias by filtering stutter on only one strand.

## Curating the reference database

To make sure all reference samples were of good quality and all alleles were called correctly, they were put through the same analysis pipeline as case samples, thereby performing noise filtering and correction on the reference samples. It is important to note that these reference samples were previously genotyped by us in great detail using CE [13]. The remaining amounts of noise in each sample were assessed using BGAnalyse (described below) to identify potentially unsuitable reference samples that still passed the thresholds of Allelefinder. Any sample with a notably higher amount of remaining background was manually removed from the set of reference samples to prevent pollution of the noise profiles.

BGAnalyse was developed and employed to analyse the remaining noise after

Chapter 5

correction. For each locus and each sample, this tool calculates the least frequent (this can be a negative value because of over-correction), most frequent, and total noise as a percentage of the number of reads of the highest allele at each locus. These results are subsequently visualised to easily identify potentially problematic samples. In the visualisation, samples can be sorted by any of the calculated values or by coverage (total number of reads). Samples were subjected to manual inspection and any sample that exhibited non-stutter products with corrected read counts above 4% of the most frequent allele or above 2% of the total reads was rejected.

## Analysing case samples

The analysis of mock case samples was performed in a three-step process which is described in the following sections.

1. A prediction was made for the amount of stutter for each sequence in the sample, using the fitted polynomial functions obtained from running Stuttermodel on the reference samples. These predictions are used to extend the allele-specific noise profiles obtained from running BGEstimate on the reference samples.
2. The extracted noise profiles are used to filter and correct the noise in the case sample.
3. Alleles are called and the sample is subjected to manual interpretation.

Similar to the creation of a reference database, analysing case samples involves multiple tools discussed in the following sections. Pipeline offers a convenient way to automatically analyse a case sample with all tools discussed.

## Predicting stutter amounts for unknown alleles

Because case samples may contain alleles that are not present in the reference samples, noise profiles for these alleles need to be predicted. FDSTools includes the BGPredict tool, which uses a previously created Stuttermodel file to predict the amounts of stutter artefacts for alleles not present in the reference data. BGPredict finds all sequences in the analysed case sample in which a particular repeat unit is repeated. The expected amount of stutter in this repeat is then computed using the corresponding fitted polynomial function from the Stuttermodel file. All possible combinations of stutter are taken into consideration when the frequencies of each stutter artefact are computed. The noise profiles created in this way are used to extend the noise profiles in the previously created BGEstimate file (a tool called BGMerge is included in FDSTools for this purpose).

## Noise filtering and correction in case samples

To be able to filter systemic noise in case samples, one first needs to determine which alleles are likely present in the sample. To this end, the algorithm of BGEstimate is essentially reversed, i.e., the goal is now to solve for **a** in $\mathbf{a}\,\mathbf{P} = \mathbf{c}$, where c is a row vector with the sample's read counts for the $M$ sequences in the noise profiles and **a** is a row vector with the estimated amount of each of the $N$ profiles in the sample. **P** is the $N \times M$ matrix of noise profiles obtained from BGEstimate, extended with the predictions obtained from BGPredict. Solving for **a** is done in a non-negative least squares sense as before, giving estimated allele contributions that best fit the various sequences – alleles as well as noise – present in the sample.

Background-corrected read counts can then be computed by first subtracting the scaled profiles from the sample's read counts

$$\mathbf{d} \leftarrow \mathbf{c} - \mathbf{a}\,\mathbf{P}$$

and then adding the total size of each profile to the corresponding allele, i.e.,

$$\mathbf{d}_n \leftarrow \mathbf{d}_n + \mathbf{a}_n \sum_{m=1}^{M} \mathbf{P}_{n,m}, \quad \forall n \in [1 \ldots M]$$

Note that **d** may have negative elements if the sample contains a lower amount of a certain sequence than was predicted by the profiles of its dominant alleles.

FDSTools offers *BGCorrect* to filter and correct background noise following the procedure outlined above. Given a sample data file (obtained from *TSSV* for example) and a file containing noise profiles, *BGCorrect* produces a copy of the sample data with additional columns giving the amounts of each sequence attributed to noise and the amounts of each sequence that would be recovered by noise correction (i.e., adding the noise to the originating allele). These values are given separately for the forward and reverse strand. Although the method by which *BGCorrect* computes them results in non-integer values, it was decided not to round these numbers to avoid unnecessary loss of precision. If necessary, these numbers can be rounded to integer values, thereby easing the interpretation as 'read counts' when presented in a graph or table in a report.

## Allele calling for case samples

The naïve method of calling alleles that *Allelefinder* uses is not appropriate for case samples, since these may contain alleles of multiple contributors in different quantities. Therefore, calling alleles in case samples is done by computing various statistics based on the information of the detected sequences and subsequently setting interpretation thresholds on these statistics. For this, *Samplestats* was developed, which operates on

Chapter 5

and adds various columns to the output of *BGCorrect*. *Samplestats* automatically marks sequences as 'allele' using the thresholds outlined in Table 1.

Alleles can also be called while visualising the sample data, hence, FDSTools includes the *Samplevis* visualisation. By means of the interactive graphical user interface of *Samplevis*, the same set of thresholds as depicted in Table 1 are available to filter the visible sequences and to automatically call alleles. Thresholds can be specified separately for the graphs and for the tables. While the table displays the called alleles, less conservative settings may be used for the filtering of the corresponding graph to ensure visibility of alleles just below the allele-calling threshold. The results of changing the thresholds are immediately visible. Clicking a sequence in any of the graphs toggles its 'allele' status. This allows the user to manually add alleles to and remove alleles from the profile. A note is added to manually added alleles, stating that the allele is 'User-added'. Similarly, if the user removes any alleles, the allele remains visible but a 'User-removed' note is added. In this way it remains easy to trace back exactly which alleles meet the thresholds and which ones were manually added and removed.

*Samplestats* can also be used to filter sequences using the same types of thresholds (albeit with more stringent threshold values than used for allele calling, as potential alleles should not be filtered out) and (optionally) aggregate the filtered sequences per locus to a single line categorised 'other sequences'.

**Table 1.** Interpretation thresholds for case samples in Samplestats and Samplevis.

| Threshold | Description | Allele calling default | Filtering default |
|---|---|---|---|
| **Total reads** | Minimum number of reads per allele. Non-systemic (and thus unfilterable) sequence errors occur sporadically. This threshold ensures that a minimal amount of amplified product is present to support the allele call. | 30 | 5 |
| **Reads per strand** | Minimum number of reads per allele for both strands. This threshold can be used to exclude low template sequences with strong strand bias. | 1 | 0 |
| **Percentage of most frequent** | The number of reads as a percentage of the number of reads of the most frequent allele at the locus. This threshold sets a limit to the mixture proportions that can be analysed in mixed samples or to the allele balance in samples with a single contributor. | 2% | 0.5% |
| **Percentage of locus** | Each allele contributes at least this percentage to the total number of reads of the locus. With this threshold, a minimum contribution percentage can be enforced. | 1.5% | 0% |
| **Percentage correction** | This percentage derives from the number of reads after noise correction minus the number of reads before correction, which is divided by the number of reads before correction. Consequently, the percentage correction is negative if noise correction resulted in a reduction of the read count of a sequence. Therefore, with this threshold set to 0%, any sequence representing noise will not be called as an allele. To be able to detect alleles of minor contributors that coincide with noise products for the major contributor's alleles, the 'percentage recovery' threshold described below is allowed to overrule this threshold. | 0% | 0% |
| **Percentage recovery** | The number of reads added by noise correction as a percentage of the total number of reads after noise correction. After noise correction, at least this percentage of reads must have originated from corrected noise. The rationale behind this threshold is that only allelic sequences will have substantial amounts of recovered reads. When an allele of a minor contribution coincides with the stutter of an allele of the major contributor, noise will be extracted and added to the major contributor's parent allele resulting in a negative percentage correction. Yet, since the minor contributor's contribution to the reads also results in noise products that are corrected, the allele will receive recovered reads and a percentage recovery >0%. To allow the calling of alleles for which no noise profile exists (or no noise was detected) in the reference database the threshold is set at 0% by default. | 0% | 0% |

*Sequences that meet either the 'Percentage correction' or 'Percentage recovery' threshold (or both) as well as all the other thresholds will be marked as 'allele'. These threshold values are evaluated after noise correction. The 'Allele calling default' column lists the default threshold values for calling alleles. The 'Filtering default' column lists the default values used for filtering displayed sequences in Samplevis graphs.*

Chapter 5

### Visualisation

For visualisation of the data, FDSTools makes use of the JavaScript graphing library Vega [19]. Vega graphs can be embedded on a web page, exposing a JavaScript programming interface that allows for updating the graphs based on the user's interaction with the web page. Vega can also run on Node.js, which allows it to be included in automated analysis pipelines to generate (static) image files.

FDSTools comes with Vega graph specifications and accompanying interactive web pages (HTML files) to visualise the output of each tool. The Vis tool can be used to obtain self-contained HTML files containing visualisations of various types of data files generated by the other tools. For example, Samplevis visualises a sample data file as a sequence profile and Profilevis visualises background noise profiles obtained from BGEstimate or BGPredict. A description of each visualisation can be found in Supplementary Table 1. When viewed in a web browser, the web page provides additional controls that allow the user to filter the data, switch between linear and logarithmic scales, or select different subsets of the data to visualise. The default values for the settings on the web page can be set when the HTML file is generated by the Vis tool.

The web pages also offer the option to save the displayed graphs as a Scalable Vector Graphics (SVG) or rasterised Portable Network Graphics (PNG) image, so that they can be imported into documents. Alternatively, the Vis tool can supply a raw Vega graph specification file (either with or without embedded data), which can then be used by Vega to generate SVG or PNG images directly on the command line.

## Results and discussion

We developed FDSTools, a software package containing a suite of tools that can be used for the analysis of forensic MPS data. With these tools, FDSTools provides detailed insight in the quality of a sample and the noise profile of a certain allele (or sequence variant). In Supplementary Table 1, an overview of all tools currently available in the package is provided, of which a selection was described in more details in Section 2.

To enhance the analysis of mixed samples, FDSTools identifies, extracts and corrects for PCR or sequencing noise such as stutter from a reference database with the aim to discern low mixture proportions. Different STR amplification assays and different amplification protocols could result in different noise. It is therefore important to base the database for noise correction on references generated by a method that is representable for the casework samples to be analysed.

Note that it is not possible to correct all noise completely as the level of noise shows variation between samples.

## Reference database

Our reference samples were sequenced with an average coverage of 65,000 reads and a mode of about 45,000 reads. For the present study, a minimum coverage of 6,000 reads per sample was required, which relates to an average of 250 reads per locus as 24 loci were co-amplified. For heterozygous loci, less than 250 reads per locus is not sufficient to quantify low amounts of noise accurately.

## Reference sample curation

Since the reference database is used to filter and correct noise in case samples, it is essential that the reference samples contain no contaminants and reference alleles are called correctly. Although all other steps can be performed automatically by FDSTools, a manual curation of samples in the reference database is needed. BGAnalyse was developed to facilitate this process by visualising potential outliers.

Allelefinder automatically rejected two out of the initial 450 samples which were clearly contaminated and three samples that had too low coverage to detect alleles reliably. Manual inspection of samples with a notably higher amount of remaining noise after correction in BGAnalyse resulted in the rejection of an additional 16 samples. Reasons for rejection were low-level contamination, low coverage and low sequencing quality. The interactive BGAnalyse visualisations displaying the remaining noise for the reference samples are available in Supplementary File 2a (before database curation) and 2b (after curation). For the majority of samples, the highest remaining noise variant in the complete profile did not exceed 3% of the number of reads of the highest allele at the locus while without correction STR stutters can represent over 20%. For the remaining 429 samples, no drop-in or drop-out was observed when calling alleles using Allelefinder with the settings described in Section 2.3.1.

## Extending noise profiles for noise correction

As described in Section 2.4.1, case samples may contain alleles which are not present in the reference database. In such cases, FDSTools resorts to noise prediction instead of noise estimation. A column in the output file of *BGCorrect* marks if correction has been performed using data obtained from *BGEstimate* (if the allele was available in the reference database) or by using *BGPredict* (if not available in the reference database).

From the results from *Stuttermodel* it becomes evident that for simple STRs consisting of a single repeating element or for long stretches of a specific repeating element within a complex STR, only few reference samples are needed to reliably fit a stutter model. However, when complex repeats consist of several repeating elements of which

some show little length variation, correction using the stutter model is suboptimal as exemplified by the predictions for D12S391. This STR locus consists of two repeat units; an AGAT repeat stretch of highly variable length and an ACAG repeat that is repeated 6 to 8 times for most individuals. Since *Stuttermodel* predicts the amount of stutter based on the repeat length, at least four different repeat lengths need to be available in homozygous reference samples to obtain a reliable fit. However, the set of reference samples used in this study only contained homozygotes with 6 to 8 repeats of ACAG, which is not sufficiently variable to obtain a reliable fit. Consequently, ACAG is omitted from the stutter model for D12S391, even though this repeat stutters up to 9% for the longer repeats (8 repeat units, data not shown). When *BGEstimate* does not obtain a background noise profile, *BGPredict* will not correct stutter in this repeat and thus stutters will remain present. As a last resort, *BGPredict* offers the possibility to use a stutter model based on data from all loci that have the same repeat unit sequence if no locus-specific fit is available. Supplementary Figure 1 displays the stutter model obtained from the set of 429 reference samples, including the individual observations on which the model was based.

Combining *BGEstimate* and *BGPredict* (by using *BGMerge*) instead of using *BGPredict* alone is expected to reduce the noise remaining after correction, as the combined correction also corrects for noise other than stutters . This is confirmed when we determine the percentage of remaining noise (the reads representing remaining noise as a percentage of the reads for the most frequent allele at the locus) and plot the highest percentage and various percentiles ($90^{th}$, $95^{th}$ and $99^{th}$) (Supplementary Figure 2a–b). The percentiles illustrate how often samples exhibit outlying noise sequence variants and when the 99th percentile is regarded, *BGPredict* alone retains on average 2.6% noise and the combined correction 2.4%. Also, the combined correction results in less overcorrected variants.

Thus, *BGPredict* can be used without *BGEstimate* with a slightly reduced accuracy in correction. Note that *BGEstimate* should not be used without *BGPredict* since alleles not included in the reference database will not be corrected, which can result in a combination of corrected and uncorrected alleles and remaining noise for the uncorrected alleles.

## Reference database size and coverage

To test the effect of the sample size and type from which the reference database is built, we used the complete curated reference database of 429 samples and a random selection of 100 samples (both with combined *BGEstimate* and *BGPredict* correction, which was found to be slightly better as described in Section 3.1.2). Supplementary

Figures 2c–d display an overview of the most frequent and the total remaining noise at each locus after correction. The different percentiles of the reference samples are given to illustrate how often samples exhibit outlying noise sequence variants.

When comparing the results for the complete database with the results for the subset of 100 samples, the difference in remaining noise seems surprisingly small (Supplementary Figure 2c–d). However, with a smaller database, less alleles will fit the criteria to create a *BGEstimate* noise profile and more alleles rely on noise prediction by *BGPredict*. Indeed, for the reference set of 429 samples, only 3.5% of the alleles are corrected using *BGPredict*. This percentage increases to 10.2% when the correction is based on the subset of 100 samples.

In a larger reference database more alleles will be observed. Supplementary Figure 3 displays the alleles observed in the reference databases of 429 and 100 samples. To fit the criteria to create a *BGEstimate* noise profile, alleles need to be present as a homozygous genotype or be available as part of shared genotypes with at least three other alleles that must also fit these criteria. For the stutter model, only the homozygous genotypes are used. In the larger 429 database, more alleles fit these criteria than in the smaller 100 sample set database.

To examine the effect of read coverage of the reference samples on noise profile analysis, we generated two subsets comprising samples with high or low coverage, which is specified as a total read count between 82,000 and 350,000 or 8,000 and 44,000 respectively. The high coverage set comprised 71 samples; the low coverage set 70. We noticed that in the low-coverage noise profiles, strand bias can occur especially for the low-percentage noise that is due to single-strand drop-out of this noise. This is illustrated by the *BGEstimate* noise profiles for the CE10_TCTA[10]_-20T>A allele for locus D7S820 in Supplementary Figure 4, in which forward and reverse reads are in good or reasonable balance for all seven noise sequences in the high coverage sample set while good balance is only seen for the two main noise sequences in the low coverage set.

Since the most abundant noise after correction in a sample is usually in the range of 0.5–3% (for STR analysis), we recommend a coverage of at least 1,000 reads per locus (which relates to a 24,000 total read coverage for our 24 loci amplification kit) for the samples of the reference database to obtain the most accurate noise estimates.

## Infrequent alleles

Depending on the composition of the reference database, occasionally alleles will be encountered that are not included in the database. *BGPredict* can predict the noise from stutter or other repeating elements but correction of other types of noise (like low level SNPs caused by sequence errors) is not possible for these infrequent alleles.

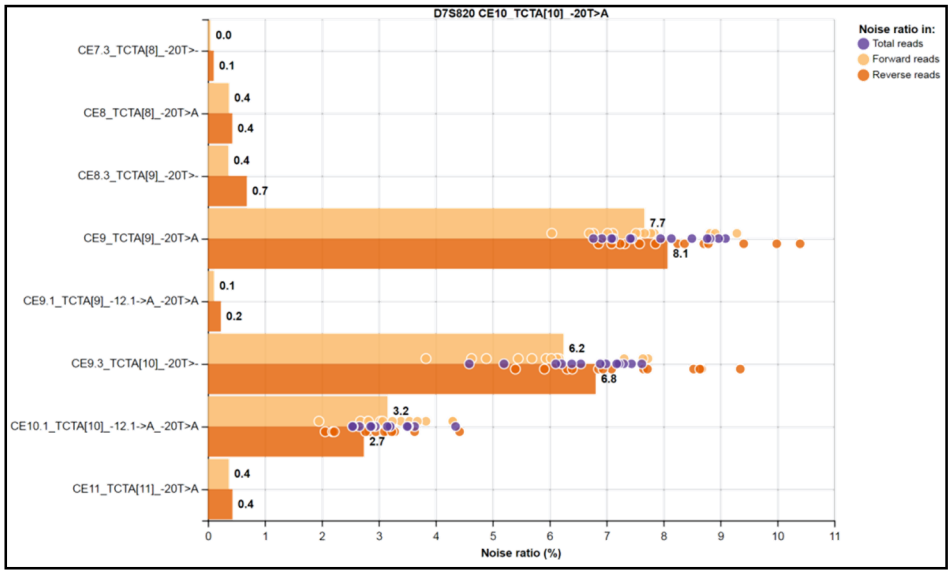
We therefore recommend to obtain *BGEstimate* noise profiles for as many alleles as possible, while retaining good quality of these noise profiles. Several filtering criteria can be applied, such as the minimum number of different heterozygous genotypes per allele, the minimum number of samples per allele and the minimum number of homozygous samples per allele. The effect of increasing the stringency on the filtering criteria on the number of retrieved *BGEstimate* noise profiles for our 429 reference set is shown in Supplementary Table 2. The settings selected for use in this study are: at least two samples per allele (which ensures noise is not based on a single sample as that could be an outlier) that present at least three different heterozygous or at least one homozygote genotype (i.e., the samples can be three different heterozygotes or two homozygotes or one homozygote and one heterozygote).

When an allele at a heterozygous locus fails the criteria, the complete locus carrying this allele cannot be used for establishment of a noise profile since the noise cannot be attributed to any of the two alleles. Thus, for both alleles at a heterozygous locus no noise profile is extracted.

## Accuracy of noise reference database and stutter model

To verify the accuracy of the noise profiles obtained through *BGEstimate* and *BGPredict*, it can be useful to compare the average noise ratios with the noise ratios observed in individual homozygous samples. The noise ratios of all noise in all homozygous reference samples can easily be collected using the *BGHomRaw* tool. These data points can be plotted on top of a noise profile to inspect the consistency and variation in the noise ratios of various types of noise for each allele. In Figure 3, the noise profile of the most frequent allele of D7S820 (CE10_TCTA[10]_-20T>A ) is displayed, which has foremost a –1 stutter (CE10_TCTA[9]_-20T>A) in addition to a –1 nt slippage product at the A-stretch (CE9.3_TCTA[10]_-20T>-). The individual observations for the homozygous samples coincide nicely with the estimated noise profile ratios.

**Figure 3.** Noise profile of D7S820 allele CE10_TCTA[10]_-20T>A



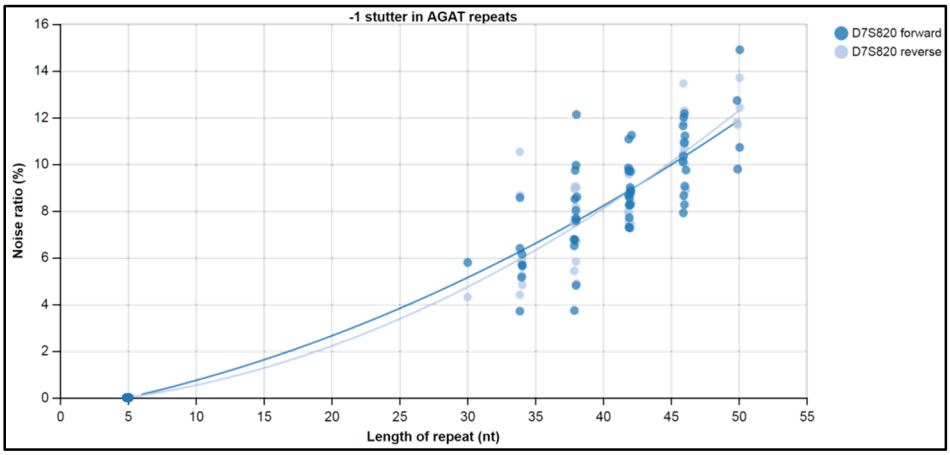

*The noise ratio is shown for each systemic noise sequence observed with a noise ratio of 0.1% or higher. Individual observations in homozygous samples (above 0.5%) are displayed as circles. As expected, the most frequently observed noise sequence is the −1 stutter, but since the allele contains a single-nucleotide stretch of 9 A nucleotides, a considerable portion of the noise consists of sequences with slippage at this A-stretch (or a combination of the two).*

Similarly, it is useful to compare the functions fitted by *Stuttermodel* to the data points to which they were fitted. *Stuttermodel* includes an option to write the raw data points to a separate output file, which can be visualised together with the fitted model as shown in Figure 4 for D7S820. This example shows that the homozygous calls and the Stuttermodel estimation follow the same trend and that there is no discrepancy between forward and reverse reads. The same holds for the A-stretch (data not shown).

In the stutter model, fits with an R2 score below 0.75 were rejected. Although this may seem a very low R2 score, we obtained better results by including more fits than by excluding them, which would result in the inability of the stutter model to be used to filter and correct stutter for the respective repeat units at all.

Chapter 5

**Figure 4.** Stutter model for the –1 stutter of D7S820




*On the x-axis the length of the repeat is displayed (in nucleotides) and on the y-axis the –1 stutter noise ratio (as percentage of reads of the parent allele) is displayed. Each homozygous reference sample is displayed as a dot and the lines display the fitted functions used for calculating the expected stutter of each allele.*

## Sample analysis

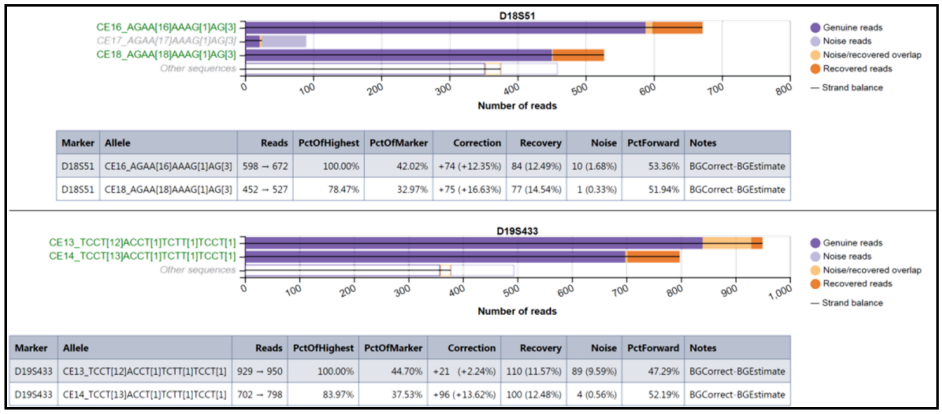
### Allele calling, interpretation and visualisation

When a reference database has been created, one can proceed with the analysis of samples. FDSTools analyses sequencing data, calls alleles and interprets the data by correction for noise as inferred from the reference database. Results can be represented as a graphical sequence profile output and as an interactive profile report.

In Figure 5, an example of a sequence profile of two loci of a single-source sample is displayed (generated by the command 'fdstools vis sample'). A sequence profile displays the read counts before and after correction and visualises the effects of noise filtering and noise correction. A more detailed explanation of the interpretation of a sequence profile can be found in Supplementary Figure 5.

The interactive sequence profile reports provide separate filtering options for the graphs and tables displayed (see Section 2.4.3). In the graphs, all alleles that are hidden by the filtering options are (optionally) aggregated as a separate bar (displaying the cumulative numbers of reads) with the label 'other sequences'. In addition, we aggregate all singleton reads into 'other sequences' already in the first step of the analysis (using 'fdstools tssv --minimum 2 --aggregate-filtered') which has the additional benefits of speeding up subsequent analysis and decreasing data storage demand.

**Figure 5.** Sequence profile of a single-source sample




| Marker | Allele | Reads | PctOfHighest | PctOfMarker | Correction | Recovery | Noise | PctForward | Notes |
|---|---|---|---|---|---|---|---|---|---|
| D18S51 | CE16_AGAA[16]AAAG[1]AG[3] | 598 → 672 | 100.00% | 42.02% | +74 (+12.35%) | 84 (12.49%) | 10 (1.68%) | 53.36% | BGCorrect-BGEstimate |
| D18S51 | CE18_AGAA[18]AAAG[1]AG[3] | 452 → 527 | 78.47% | 32.97% | +75 (+16.63%) | 77 (14.54%) | 1 (0.33%) | 51.94% | BGCorrect-BGEstimate |

| Marker | Allele | Reads | PctOfHighest | PctOfMarker | Correction | Recovery | Noise | PctForward | Notes |
|---|---|---|---|---|---|---|---|---|---|
| D19S433 | CE13_TCCT[12]ACCT[1]TCTT[1]TCCT[1] | 929 → 950 | 100.00% | 44.70% | +21 (+2.24%) | 110 (11.57%) | 89 (9.59%) | 47.29% | BGCorrect-BGEstimate |
| D19S433 | CE14_TCCT[13]ACCT[1]TCTT[1]TCCT[1] | 702 → 798 | 83.97% | 37.53% | +96 (+13.62%) | 100 (12.48%) | 4 (0.56%) | 52.19% | BGCorrect-BGEstimate |

*Sequence profile of loci D18S51 and D19S433 of a single-source sample. A sequence profile displays the read count before correction (in purple bars) and shows the effects of noise filtering (light purple for the reads that are removed) and noise correction (with the noise reads added to the parent alleles in dark orange). When performing correction, it is possible that an allele gains reads because the noise reads originating from this allele are added, but loses reads at the same time since the noise of another allele in the profile includes reads of this allele. This overlapping part of added and removed reads is marked separately in light orange. This means that the original read count of an allele before correction is the combination of the purple and the light orange bar. The lines in the bars indicate the strand balance; the line is drawn near the top of the bar if the majority of reads of a sequence is on the forward strand, near the bottom of the bar if the majority of reads is on the reverse strand, and in the middle of the bar in the absence of strand bias. Sequences displayed in green in the graphs are the alleles that the software infers to be genuine alleles in the sample. These are also displayed in the table.*

## Improving heterozygote balance through noise correction

The amplification of long STR alleles in the PCR is generally less efficient than shorter alleles and, in addition, long STR alleles suffer from a higher degree of stutter resulting in reduced heterozygote balance between the two alleles. [2] Since FDSTools determines which 'noise reads' are derived from which parent alleles, these reads can (optionally) be added to the read counts of the parent alleles, which theoretically will improve the heterozygote balance. When we examine the heterozygote allele balance in the 429 single-source reference samples, an improved heterozygote balance is indeed observed when the stutter reads are added to the read counts of the parent alleles (Table 2). Heterozygote balance was determined per locus by dividing the read counts for the less frequent alleles by those for the more frequent alleles, and taking the average of all 429 samples.

## Table 2. Heterozygote balance for original, filtered and corrected datasets

| Dataset \ locus | Amel | CSF1PO | D10S1248 | D12S391 | D13S317 | D16S539 | D18S51 | D19S433 | D1S1656 | D21S11 | D22S1045 | D2S1338 | D2S441 | D3S1358 | D5S818 | D7S820 | D8S1179 | FGA | PentaD | PentaE | TH01 | TPOX | vWA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Uncorrected data | 0.83 | 0.88 | 0.82 | 0.79 | 0.88 | 0.87 | 0.85 | 0.84 | 0.88 | 0.89 | 0.85 | 0.77 | 0.90 | 0.88 | 0.90 | 0.88 | 0.89 | 0.85 | 0.87 | 0.78 | 0.88 | 0.88 | 0.85 |
| Filtered data without noise reads added to allele read count | 0.83 | 0.90 | 0.87 | 0.80 | 0.89 | 0.89 | 0.87 | 0.88 | 0.89 | 0.90 | 0.88 | 0.78 | 0.90 | 0.90 | 0.91 | 0.89 | 0.90 | 0.87 | 0.87 | 0.78 | 0.88 | 0.89 | 0.88 |
| Corrected data with noise reads added to allele read count | 0.83 | 0.91 | 0.89 | 0.84 | 0.90 | 0.91 | 0.89 | 0.90 | 0.91 | 0.91 | 0.93 | 0.80 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.89 | 0.88 | 0.81 | 0.89 | 0.90 | 0.90 |

*The read counts for the less frequent alleles are divided by those for the more frequent alleles, and the average for all 429 single-source reference samples is taken.*

## Mixture analysis

For the analysis mixtures, noise correction may assist in identifying the alleles of a low minor contributor. We used 31 two-person mixtures with minor contributions of 50%, 20%, 10%, 5%, 1% and 0.5% to assess this expectation.

We varied the 'percentage of locus' threshold (Table 1) for calling alleles, which sets a limit to the mixture proportion. When no noise correction was applied the threshold was varied between 5.0% and 1.5%; when noise correction was applied, a lower threshold could be used, varying between 3.0% to 0.5%. We compared the various methods by calculating percentage missed alleles (a.k.a. drop-out) and the number of erroneous allele calls (a.k.a. drop-in). The percentage drop-out was calculated by dividing the number of donor alleles not called by the total number of possible alleles (homozygous and shared alleles are counted as one, Amel is included), and the percentage was averaged for the mixtures with the same mixture ratio. Drop-in is presented in the average number occurring in profiles with the same mixture ratio. In Table 3, the results of these analyses are displayed and it is obvious that without correction more drop-in alleles occur that mostly represent stutters. Consequently, the threshold for calling an allele can be lower when correction is applied, as less stutters remain in the corrected profile that can be wrongfully called as an allele. As expected, the percentage of drop-out depends largely on the 'percentage of locus' threshold for allele calling (and hardly on the application of noise correction); drop-out is more frequent with a higher (more stringent) threshold. When the threshold for corrected data is decreased below 1.5%, the number of drop-ins rapidly increases for all ratios. Not surprisingly, the drop-out percentage for the mixtures with 1% and 0.5% is very high when using a threshold that is higher than the minor component. Therefore, the data from the 5% and 10% minor contribution was used to determine the optimal threshold for allele calling.

In Figure 6, we show the relation between the 'percentage of locus' allele-calling threshold, drop-out and drop-in for the mixtures with a 5% or 10% minor contribution. In the used dataset, a threshold of 1.5% appears to be the most effective for calling

genuine alleles in mixtures with minimal erroneous calling of remaining noise in the mixtures. With mixture ratios down to 10%, no drop-out and only minimal drop-in is observed with this threshold, whereas with contributions smaller than 10% an optimal balance between drop-out and drop-in is achieved (Table 3). When investigating the noise that is erroneously called using this threshold it is apparent that the drop-in alleles are rarely resulting from stutter but almost exclusively consist of PCR hybrids [20]. In Table 3b, the percentage of drop-out when using the 1.5% allele-calling threshold is categorised and illustrates that drop-out alleles consist mostly of heterozygous minor alleles. Most of these drop-out alleles represent minor contributions on stutter positions (where stutter ratio of the major contributor was lower than the average observed in the set of reference samples, thereby causing over-correction) and long alleles that suffer from heterozygote imbalance. In Figure 6 the trends from Table 3 are confirmed: drop-out is hardly and drop-in is largely affected by the use of noise correction. Thus, calling of genuine alleles is not negatively influenced by noise correction.

Note that the number of drop-ins may be reduced further by applying additional thresholds from Table 1, but the effects of varying additional threshold values were not studied in depth.

**Table 3.** Average number of drop-in alleles and average drop-out percentage per sample for different 'percentage of locus' allele-calling thresholds

**a)** Summary of drop-in and drop-out rates for various allele-calling thresholds

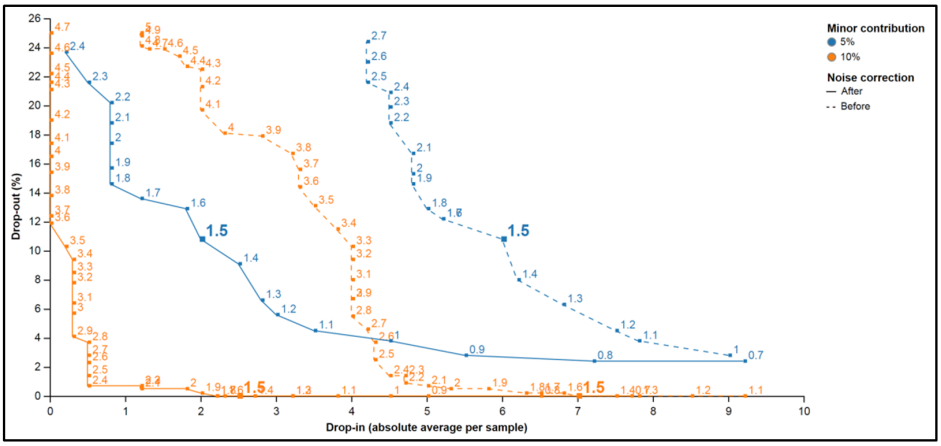
| Minor contribution | 50%: 600 pg | 20%: 240 pg | 10%: 120 pg | 5%: 60 pg | 1%: 60 pg | 0.5%: 60 pg |
|---|---|---|---|---|---|---|
| Analysis method and threshold | # drop-in / % drop-out | # drop-in / % drop-out | # drop-in / % drop-out | # drop-in / % drop-out | # drop-in / % drop-out | # drop-in / % drop-out |
| Without correction, ≥ 5.0% | 0.7 / 0.0% | 0.8 / 2.1% | 1.2 / 25.0% | 1.0 / 33.1% | 1.7 / 39.9% | 0.8 / 40.8% |
| Without correction, ≥ 2.5% | 4.3 / 0.0% | 3.8 / 0.0% | 4.3 / 2.5% | 4.2 / 21.6% | 3.5 / 38.3% | 2.3 / 40.4% |
| Without correction, ≥ 2.0% | 5.3 / 0.0% | 5.3 / 0.0% | 5.3 / 0.5% | 4.8 / 15.3% | 4.2 / 38.1% | 2.8 / 40.1% |
| Without correction, ≥ 1.5% | 7.7 / 0.0% | 7.7 / 0.0% | 7.0 / 0.0% | 6.0 / 10.8% | 6.0 / 37.4% | 4.7 / 39.7% |
| With correction, ≥ 3.0% | 0.0 / 0.0% | 0.3 / 0.0% | 0.3 / 5.7% | 0.0 / 30.3% | 0.7 / 41.1% | 0.3 / 41.1% |
| With correction, ≥ 2.5% | 0.3 / 0.0% | 0.3 / 0.0% | 0.5 / 1.4% | 0.0 / 25.1% | 0.7 / 40.6% | 0.3 / 41.1% |
| With correction, ≥ 2.0% | 0.7 / 0.0% | 1.2 / 0.0% | 1.8 / 0.5% | 0.8 / 17.4% | 0.8 / 40.6% | 1.2 / 40.8% |
| **With correction, ≥ 1.5%** | 1.7 / 0.0% | 2.3 / 0.0% | 2.5 / 0.0% | 2.0 / 10.8% | 2.7 / 39.9% | 2.3 / 40.8% |
| With correction, ≥ 1.0% | 4.0 / 0.0% | 5.2 / 0.0% | 4.5 / 0.0% | 4.5 / 3.8% | 5.7 / 37.8% | 3.2 / 40.6% |
| With correction, ≥ 0.5% | 16.3 / 0.0% | 17.3 / 0.0% | 14.0 / 0.0% | 15.5 / 2.4% | 14.0 / 30.3% | 11.0 / 37.4% |

**b)** Categorised drop-out rates when using 1.5% allele-calling threshold (with correction)

| Minor contribution | 20%: 240 pg | 10%: 120 pg | 5%: 60 pg | 1%: 60 pg | 0.5%: 60 pg |
|---|---|---|---|---|---|
| Alleles unique to the minor (homozygous) | 0.0% | 0.0% | 0.0% | 91.3% | 100.0% |
| Alleles unique to the minor (heterozygous) | 0.0% | 0.0% | 31.6% | 98.1% | 99.4% |
| Alleles unique to minor | 0.0% | 0.0% | 27.0% | 97.2% | 99.4% |
| All alleles of the minor | 0.0% | 0.0% | 18.5% | 67.7% | 69.3% |
| All alleles of the major | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |

Chapter 5

**Figure 6.** Average number of drop-in and percentage of drop-out per sample for different 'percentage of locus' allele-calling thresholds




*Effect of different 'percentage of locus' allele-calling thresholds on the drop-in and drop-out rates. The numbers next to the points display allele-calling thresholds. The position of each point illustrates the number of drop-ins and percentage of drop-out for the corresponding threshold in mixtures with a ratio of 90:10 (orange) and 95:5 (blue). Points connected by dashed lines correspond to results obtained without noise correction, points connected by solid lines correspond to results obtained after noise correction. The 1.5% 'percentage of locus' allele-calling threshold that appears most optimal is indicated in bold.*

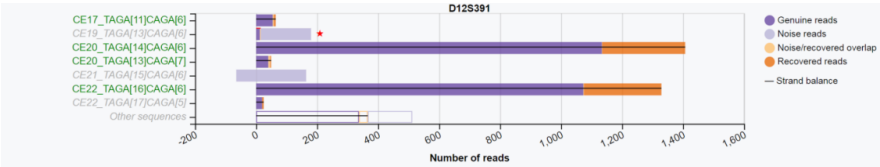
In Figure 7a–b, the effect of noise correction on allele calling is shown for a highly unbalanced mixture (95:5 mixture ratio) in which the alleles of the minor contributor have a similar or lower read count than the stutter products of the alleles of the major contributor. Without noise correction the four most frequent sequence variants are the major contributor's alleles and the corresponding –1 stutters and interpretation of the less frequent sequence variants becomes intractable; after noise correction, the stutter products and other PCR artefacts are filtered out and four sequence variants meet the 'percentage of locus' allele-calling threshold of 1.5%, which correspond to the four alleles of the two heterozygous donors. Also, the alleles of both the major and minor contributor have gained recovered reads.

**Figure 7.** Interpretation of a mixed sequence profile before and after correction

**a)** Sequence profile of locus D12S391 of a mixed sample with a ratio of 95:5, without noise filtering and correction

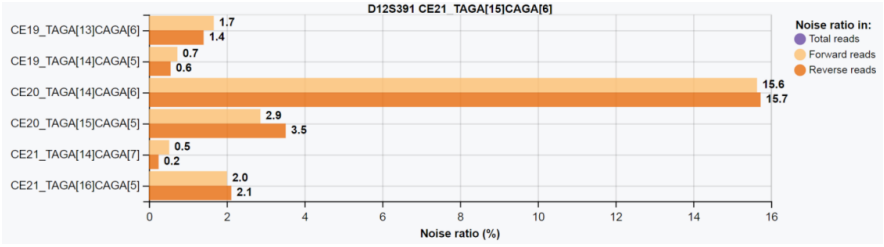
| Marker | Allele | Reads | PctOfHighest | PctOfMarker | Correction | Recovery | Noise | PctForward | Notes |
|---|---|---|---|---|---|---|---|---|---|
| D12S391 | CE17_TAGA[11]CAGA[6] | 58 | 5.11% | 1.82% | +0 (+0.00%) | 0 (0.00%) | 0 (0.00%) | 62.07% | |
| D12S391 | CE19_TAGA[13]CAGA[6] | 180 | 15.86% | 5.64% | +0 (+0.00%) | 0 (0.00%) | 0 (0.00%) | 59.44% | |
| D12S391 | CE20_TAGA[14]CAGA[6] | 1135 | 100.00% | 35.58% | +0 (+0.00%) | 0 (0.00%) | 0 (0.00%) | 59.21% | |
| D12S391 | CE21_TAGA[15]CAGA[6] | 164 | 14.45% | 5.14% | +0 (+0.00%) | 0 (0.00%) | 0 (0.00%) | 56.10% | |
| D12S391 | CE22_TAGA[16]CAGA[6] | 1075 | 94.71% | 33.70% | +0 (+0.00%) | 0 (0.00%) | 0 (0.00%) | 61.40% | |

**b)** Sequence profile of locus D12S391 of a mixed sample with a ratio of 95:5, with noise filtering and correction applied




| Marker | Allele | Reads | PctOfHighest | PctOfMarker | Correction | Recovery | Noise | PctForward | Notes |
|---|---|---|---|---|---|---|---|---|---|
| D12S391 | CE17_TAGA[11]CAGA[6] | 58 → 64 | 4.56% | 2.01% | +6 (+10.67%) | 10 (14.93%) | 3 (5.86%) | 60.59% | BGCorrect·BGEstimate |
| D12S391 | CE20_TAGA[14]CAGA[6] | 1135 → 1407 | 100.00% | 44.11% | +272 (+23.97%) | 273 (19.39%) | 1 (0.07%) | 57.37% | BGCorrect·BGEstimate |
| D12S391 | CE20_TAGA[13]CAGA[7] | 46 → 49 | 3.50% | 1.54% | +3 (+7.08%) | 9 (19.07%) | 6 (13.34%) | 60.73% | BGCorrect·BGEstimate |
| D12S391 | CE22_TAGA[16]CAGA[6] | 1075 → 1329 | 94.45% | 41.66% | +254 (+23.62%) | 256 (19.24%) | 2 (0.16%) | 59.45% | BGCorrect·BGPredict |

**c)** Noise profile of D12S391 allele CE21_TAGA[15]CAGA[6]




*Sequence profile of locus D12S391 of a mixed sample with a ratio of 95:5, A without and B with applying noise filtering and correction. The table displays all sequence variants with at least 1.5% of the reads of the locus (the 'percentage of locus' allele-calling threshold used), which are also marked in green in the graph. A note in the table in panel B warns that no noise profile was available for the major CE22_TAGA[16]CAGA[6] allele and a stutter prediction has been used instead. An additional variant CE22_TAGA[17]CAGA[5], which is derived from the major CE22 allele, remains visible in the sequence profile (although not marked green as it does not meet the 1.5% threshold). C Noise profile of a similar allele, showing a non-stutter PCR artefact with a noise ratio of about 2%.*

Chapter 5

This example also displays a pitfall of the interpretation of a mixed DNA profile where the major contributor has an infrequent allele for which no *BGEstimate* noise profile is available. The noise correction of allele CE22_TAGA[16]CAGA[6] is only based on the stutter model (using BGPredict), which fails to correct for the CE22_TAGA[17]CAGA[5] PCR artefact. Looking at the noise profile of the most resembling allele in the reference database, CE21_TAGA[15]CAGA[6], we find a similar PCR artefact CE21_TAGA[16]CAGA[5] that represents a C to T substitution at the first CAGA repeat unit , with a noise ratio of about 2% (Figure 7c). This suggests that the CE22_TAGA[17]CAGA[5] artefact would be properly corrected if a noise profile for the CE22_TAGA[16]CAGA[6] allele would be available. Thus, additional inspection of the applied method of correction (*BGPredict* or *BGEstimate*) may be useful when infrequent alleles occur.

### Analysis time and computer demand

To indicate the required time and computer memory demand, five samples with different numbers of reads (15,169–318,403 total read pairs) were analysed and the time and peak memory usage for each separate tool was registered (Supplementary Figure 6). With the used 2.0 GHz processor, the analysis time is mostly consumed by TSSV (≈75% of the total analysis time) and the complete analysis only takes up to 13:30 minutes for a sample with 318,403 reads. BGCorrect shows the highest peak memory usage but does not exceed 200 MB for the largest sample (of the five tested samples). Both the required time and memory increase more or less linearly when the read count of the analysed samples is increased.

## Conclusions

We developed FDSTools for the analysis of forensic MPS data. FDSTools can determine systemic PCR and/or sequencing noise from the data of reference samples, build a database from this data and use it to correct for systemic noise in case samples. The software is also able to predict the noise caused by stutter for alleles not included in the reference database and uses this information in the correction of case samples.

With automatic threshold-based allele calling, noise correction reduces the occurrence of drop-in and drop-out substantially and improves the balance between alleles of a heterozygote pair. This decreases the detection limits of minor contributions in mixtures. STR stutter variants are no longer the most frequent remaining noise as PCR hybrid artefacts now generally exceed the corrected read counts of stutters.

Although reliable noise correction can already be obtained from a database of 100 samples, a larger database is preferred as a larger number of alleles can be corrected by

the use of a complete noise profile instead of relying on noise predictions based on the stutter model. This will also reduce manual inspection of retained non-stutter noise for infrequent alleles. When building the database, it is important to use an amplification kit representative for the kit used for the samples. Although not extensively tested, we anticipate that noise prediction will be less precise when less DNA is used and more stochastic PCR effects occur. Also, more strand bias will occur during the massively parallel sequencing. These effects are intrinsic to low-level DNA typing and probabilistic genotyping software have been developed that accommodate drop-in and dropout during profile interpretation [21,22,23,24,25,26,27,28]. Such software are not yet straightforwardly able to deal with MPS data, but the necessary adaptations are feasible and include nomenclature for sequence variants, allele frequencies databases and read counts replacing peak heights in continuous models (not required for semi-continuous models) . In CE-based analysis, PCR replicates are often used to reduce profiling uncertainty [7]; replicates can be entered in probabilistic genotyping software or used to prepare a consensus profile [7,8]. A future version of FDSTools will feature a consensus-based analysis method alike those used with CE data [8]. Besides, export options for DNA database systems such as CODIS will be added.

FDSTools has been validated following recommendations for software validation [29,30] and is already implemented in the ISO17025 certified environment of the LUMC for forensic casework. The validation for use and performance of the software in casework was a separate study which was not based on the data described in this manuscript. By providing tools to evaluate the performance of noise correction in reference samples FDSTools facilitates the determination of analysis thresholds that are fit for purpose.

The application of FDSTools is not limited to the analysis of STRs. FDSTools has already been applied successfully to the analysis of multiplex assays of SNP fragments (manuscript in preparation) and complete mtDNA data (Weiler et al., submitted). Note that the minimum number of required reference samples for loci other than STRs will depend on the amount of variation observed in these loci.

# Acknowledgements

This study was supported by a grant from the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands. The authors wish to thank Arie Koppelaar (Netherlands Forensic Institute) for providing help and expertise in setting up the computer cluster for the analyses. We also thank Martin Ensenberger, Cynthia Sprecher and Douglas R. Storts (Promega Corporation) for their contributions to designing and providing the PowerSeq™ Auto System.

# References


1. M.A. Jobling, P. Gill. Encoded evidence: DNA in forensic analysis. Nat. Rev. Genet. 5 (10) (2004) 739–751.
2. K.J. van der Gaag, R.H. de Leeuw, J. Hoogenboom, J. Patel, D.R. Storts, J.F.J. Laros, P. de Knijff. Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq™ System. Forensic Sci. Int. Genet. 24 (2016) 86-96.
3. K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, R.A. Guerrieri, P.M. Vallone. Sequence variation of 22 autosomal STR loci detected by next generation sequencing. Forensic Sci. Int. Genet. 21 (2016) 15–21.
4. C. Gelardi, E. Rockenbauer, S. Dalsgaard, C. Børsting, N. Morling. Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. Forensic Sci. Int. Genet. 12 (2014) 38–41.
5. M. Scheible, O. Loreille, R. Just, J. Irwin. Short tandem repeat typing on the 454 platform: Strategies and considerations for targeted sequencing of common forensic markers. Forensic Sci. Int. Genet. 12 (2014) 104–119.
6. X. Zeng, J.L. King, M. Stoljarova, D.H. Warshauer, B.L. LaRue, A. Sajantila, J. Patel, D.R. Storts, B. Budowle. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci. Int. Genet. 16 (2015) 38–47.
7. C.C. Benschop, C.P. van der Beek, H.C. Meiland, A.G. van Gorp, A.A. Westen, T. Sijen. Low template STR typing: effect of replicate number and consensus method on genotyping reliability and DNA database search results. Forensic Sci. Int. Genet. 5 (2011) 316–328.
8. C.C. Benschop, T. Sijen. LoCIM-tool: An expert's assistant for inferring the major contributor's alleles in mixed consensus DNA profiles. Forensic Sci. Int. Genet. 11 (2014) 154–165.
9. C. Brookes, J.A. Bright, S. Harbison, J. Buckleton. Characterising stutter in forensic STR multiplexes. Forensic Sci. Int. Genet. 6 (2012) 58–63.
10. S.Y. Anvar, K.J. van der Gaag, J.W. van der Heijden, M.H. Veltrop, R.H. Vossen, R.H. de Leeuw, C. Breukel, H.P. Buermans, J.S. Verbeek, P. de Knijff, J.T. den Dunnen, J.F.J. Laros. TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes. Bioinformatics 30 (2014) 1651–1659.
11. D.H. Warshauer, J.L. King, B. Budowle. STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Forensic Sci. Int. Genet. 14 (2015) 182–186.
12. S.L. Friis, A. Buchard, E. Rockenbauer, C. Børsting, N. Morling. Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs. Forensic Sci. Int. Genet., 21 (2016) 68–75.
13. A.A. Westen, T. Kraaijenbrink, E.A. Robles de Medina, J. Harteveld, P. Willemse, S.B. Zuniga, K.J. van der Gaag, N.E. Weiler, J. Warnaar, M. Kayser, T. Sijen, P. de Knijff. Comparing six commercial autosomal STR kits in a large Dutch population sample. Forensic Sci. Int. Genet. 10 (2014) 55–63.
14. M.G. Ensenberger, K.A. Lenz, L.K. Matthies, G.M. Hadinoto, J.E. Schienman, A.J. Przech, M.W. Morganti, D.T. Renstrom, V.M. Baker, K.M. Gawrys, M. Hoogendoorn, C.R. Steffen, P. Martín, A. Alonso, H.R. Olson, C.J. Sprecher, D.R. Storts. Developmental validation of the PowerPlex® Fusion 6C System. Forensic Sci. Int. Genet. 21 (2016) 234–144.
15. T. Magoc, S.L. Salzberg. FLASH: fast length adjustment of short reads to improve genome

assemblies. Bioinformatics 27 (2011) 2957–2963.
16. B.C. Hendrickson, B. Leclair, S. Forrest, J. Ryan, B.E. Ward, D. Petersen, T.D. Kupferschmid, T. Scholl. Accurate STR allele designations at the FGA and vWA loci despite site polymorphisms. J. Forensic Sci. 49 (2) (2004) 250–254.
17. D. Shinde, Y. Lai, F. Sun, N. Arnheim. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. Nucleic Acids Res. 31 (2003) 974–980.
18. M. Klintschar, P. Wiegand. Polymerase slippage in relation to the uniformity of tetrameric repeat stretches. Forensic Sci. Int. 135 (2) (2003) 163–166.
19. A. Satyanarayan, R. Russell, J. Hoffswell, J. Heer. Reactive Vega: A Streaming Dataflow Architecture for Declarative Interactive Visualization. IEEE Trans. Vis. Comput. Graphics. 22 (2016) 659–668.
20. A.R. Shuldiner, A. Nirula, J. Roth. Hybrid DNA artifact from PCR of closely related target sequences. Nucleic Acids Res. 17 (11) (1989) 4409.
21. H. Haned, P. Gill. Analysis of complex DNA mixtures using the Forensim package. Forensic Sci. Int. Genet. Supp. Series 3 (2011) e79–e80.
22. H. Swaminathan, A. Garg, C.M. Grgicak, M. Medard, D.S. Lun. CEESIt: A computational tool for the interpretation of STR mixtures. Forensic Sci. Int. Genet. (2016) 149–160.
23. D.J. Balding. Evaluation of mixed-source, low-template DNA profiles in forensic science. Proc. Natl. Acad. Sci. U. S. A. 110 (30) (2013) 12241–12246.
24. D. Taylor, J.A. Bright, J. Buckleton. The interpretation of single source and mixed DNA profiles. Forensic Sci. Int. Genet. 7 (2013) 516–528.
25. R.G. Cowell, T. Graversen, S.L. Lauritzen, J. Mortera. Analysis of forensic DNA mixtures with artefacts. Appl. Stat. 64 (1) (2015) 1–32.
26. O. Bleka, G. Storvik, P. Gill. EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. Forensic Sci. Int. Genet. 21 (2016) 35–44.
27. M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B.W. Duceman. Validating TrueAllele® DNA Mixture Interpretation. J. Forensic Sci. 56 (2011) 1430–1447.
28. R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I.W. Evett, J. Curran, D. Balding. Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelelic dropout and stutters. Forensic Sci. Int. Genet. 7 (2013) 555–563.
29. H. Haned, P. Gill, K. Lohmueller, K. Inman, N. Rudin. Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations. Sci. Justice 56 (2) (2016) 104–108.
30. E.J. Rykiel. Testing ecological models: the meaning of validation. Ecol. Model. 90 (1996) 229–244.

Chapter 5

# Supplementary materials

**Supplementary Text 1 -** Allele-centric systemic noise estimation (BGEstimate)

The BGEstimate tool of FDSTools does the computation of a background noise profile for each allele found among a set of reference samples. Computing background noise profiles from a set of homozygous samples is straightforward, whereas for heterozygous samples this becomes more complicated as the alleles of one sample in general appear in different amounts, thereby systematically contributing to a different amount of the same background noise sequence.

Algorithm 1, which enables the computation of these background noise profiles from heterozygous samples, is implemented in BGEstimate. In testing, the best results were obtained if for each allele at least one homozygous sample or at least three different heterozygous samples were available. With default settings, BGEstimate will ensure these conditions are met before executing Algorithm 1.

In essence Algorithm 1 seeks a non-negative least squares solution to the matrix equation $AP = C$. In this equation, C is an N×M matrix of constants derived from the observed read counts in the reference samples (see Figure 1), A is an N×N matrix in which the estimated allele balance in the samples is summarised and P is an N×M matrix containing the estimated profiles of systemic noise. N is the number of unique alleles among the observed reference samples and therefore also the number of profiles produced and M is the number of unique sequences observed. It is possible to include additional sequences beyond the N alleles of the samples if this is deemed appropriate. With default settings, BGEstimate will include all sequences that appear in at least 80% of the samples with any particular allele, since these sequences are probably the result of systemic noise. In any case, the first N columns in P and C correspond to the N alleles of the samples and the order of the rows and columns is the same (i.e., row n and column n in both P and C correspond to the same sequence).

The input of Algorithm 1 consists of a K×M matrix S which contains the observed number of reads of each of the M sequences in each of the K samples. The genotype of each sample is provided as a set of indices $g_k \forall g_k \leq N$.

Any element $P_{n,m}$ of P can be interpreted as the amount of sequence m that is observed, on average, for every 100 reads of sequence n. Therefore, the algorithm initialises P to a diagonal N×M matrix with the elements on its major diagonal set to 100. The number 100 was chosen for practical reasons since it directly results in noise ratios expressed as percentages of the actual allele.

**Algorithm 1** Systemic noise profile estimation.

**Require:** $K \times M$ matrix $\mathbf{S}$ containing the $K$ samples in the rows
**Require:** number of alleles $N \leq M$, corresponding to the first $N$ columns of $\mathbf{S}$
**Require:** list $g$ of sets $g_k$: the genotypes (indices of alleles $\leq N$) of samples $\mathbf{S}_k$
**Return:** $N \times M$ matrix $\mathbf{P}$ containing the profiles

**function** EstimateBackgroundProfiles($\mathbf{S}, g, N$)
  *Initialisation:*
  $\mathbf{P}_{n,m} \leftarrow 0, \quad 1 \leq n \leq N, \quad 1 \leq m \leq M$
  $\mathbf{P}_{n,n} \leftarrow 100, \quad 1 \leq n \leq N$ ▷ Set actual allele to 100.
  $\mathbf{C}_{n,m} \leftarrow 0, \quad 1 \leq n \leq N, \quad 1 \leq m \leq M$
  **for** $k \leftarrow 1$ **to** $K$ **do**
    $\mathbf{C}_{i,:} \leftarrow \mathbf{C}_{i,:} + \mathbf{S}_{k,:} \times \frac{100}{|g_k|} / \mathbf{S}_{k,i}, \quad \forall i \in g_k$ ▷ Compute separately for each $i$.
  **end for**
  *Optimisation:*
  **repeat**
    *Estimate allele balance:*
    $\mathbf{A}_{i,j} \leftarrow 0, \quad \forall i, j \in [1 \dots N]$
    **for** $k \leftarrow 1$ **to** $K$ **do**
      $\mathbf{Q}_{i,j} \leftarrow \mathbf{P}_{g_{k,i}, g_{k,j}}, \quad \forall i, j \in [1 \dots |g_k|]$
      $\mathbf{r}_i \leftarrow \mathbf{S}_{k, g_{k,i}}, \quad \forall i \in [1 \dots |g_k|]$
      $\mathbf{B} \leftarrow \left( \frac{100}{|g_k|} / \mathbf{r}^T \right) \times \textsc{nnls}(\mathbf{Q}^T, \mathbf{r}^T)^T$
      $\mathbf{A}_{g_{k,i}, g_{k,j}} \leftarrow \mathbf{A}_{g_{k,i}, g_{k,j}} + \mathbf{B}_{i,j}, \quad \forall i, j \in [1 \dots |g_k|]$
    **end for**
    *Update profiles:*
    $\mathbf{E} \leftarrow \mathbf{A}^T \mathbf{A}$
    $\mathbf{F} \leftarrow \mathbf{A}^T \mathbf{C}$
    **repeat**
      **for** $n \leftarrow 1$ **to** $N$ **do**
        $\mathbf{P}_{n,:} \leftarrow (\mathbf{F}_{n,:} - \mathbf{E}_{n,\setminus n} \mathbf{P}_{\setminus n,:}) / \mathbf{E}_{n,n}$
        $\mathbf{P}_{n,:} \leftarrow \textsc{max}(\mathbf{P}_{n,:}, 0)$ ▷ Set negative elements to 0.
        $\mathbf{P}_{n,n} \leftarrow 100$ ▷ Keep actual allele at 100.
      **end for**
    **until** stop condition is met
  **until** stop condition is met
  **return** $\mathbf{P}$
**end function**

Chapter 5

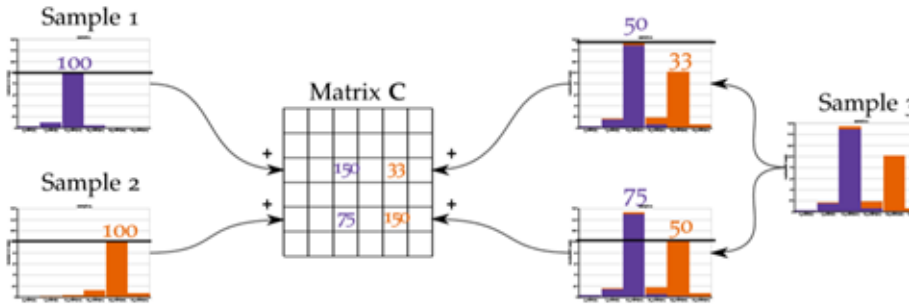



*Figure 1. Construction of matrix C in Algorithm 1 by summing scaled read counts of the samples that share the same alleles. Left: Samples 1 and 2 are homozygous for the third and fifth allele respectively. The read counts of both samples are scaled such that their true allele is equal to 100, after which they are added to the third and fifth row of matrix C respectively. Right: Sample 3 is heterozygous, having both the third and fifth allele. Therefore, it is added to both the third and fifth row of matrix C, with its read counts scaled such that the third or fifth allele is equal to 50 respectively.*

Similarly, the elements Cn,m of C can be interpreted as the total amount of sequence m that is observed in all samples with allele n. Line 6 in Algorithm 1 initialises C. For homozygous samples, it scales the read counts of each sample Sk such that its allele Sk,i,i ∈ gk is equal to 100 and then adds the scaled counts to row Ci of C. Heterozygous samples are treated likewise twice — once for each of their alleles — except that the allele is scaled to 50 instead of 100 to compensate for the fact that the sample is added to two rows in C, as compared to just one for homozygous samples.[1]

Each row Ci of C thus contains the sum of the read counts of all samples that have allele i, with the read counts of each sample scaled such that allele Sk,i,i ∈ gk becomes 100/|gk|.[2] After matrices P and C are initialised, the algorithm enters its main loop wherein it alternately estimates the allele balance in the samples (matrix A) and refines the least squares fit of the profiles P. The main loop is exited and the profiles are returned when the sum of the squared errors,

$$\sum_{n=1}^{N}\sum_{m=1}^{M}(\mathbf{D}_{n,m})^2, \quad \mathbf{D}=\mathbf{C}-\mathbf{AP}$$

is reduced by less than 0.01% in one iteration. This stopping condition is generally met within 20 iterations.

*1 One may also say that the samples are added once for each allele, adding them to the same row twice*
*2 Interestingly, because Algorithm 1 scales read counts to 100 divided by the number of alleles in the sample, it handles samples with more than two alleles without problems. Since Algorithm 1 makes no assumptions about the number of alleles each sample can have, it is possible to use mixed samples to compute systemic noise profiles as well. This has not been tested, however.*

Estimation of the allele balance is done for every sample in isolation. At line 11, the elements in P that correspond to cross-contributions between the alleles $i,j \in [1...|g_k|]$ of sample k are extracted. Similarly, the corresponding elements from Sk are extracted at line 12. For heterozygous samples, this expands to (shortening $g_{k,i}$ to i for brevity):

$$\mathbf{Q} \leftarrow \begin{bmatrix} \mathbf{P}_{i,i} & \mathbf{P}_{i,j} \\ \mathbf{P}_{j,i} & \mathbf{P}_{j,j} \end{bmatrix} = \begin{bmatrix} 100 & \mathbf{P}_{i,j} \\ \mathbf{P}_{j,i} & 100 \end{bmatrix}$$

$$\mathbf{r} \leftarrow \begin{bmatrix} \mathbf{S}_{k,i} & \mathbf{S}_{k,j} \end{bmatrix}$$

At line 13, a non-negative least squares algorithm is employed to estimate the allele balance within the sample. The nnls function can be any algorithm that solves JK = L for K subject to K≥0 in the least squares sense, e.g., [1]. Line 13 of Algorithm 1 uses this function to solve bQ = r for b (by solving QTbT = rT), which gives an estimation of the proportions in which the alleles are present in the sample. The resulting row vector b is left-multiplied by a column vector with the same scaling factors as previously calculated at line 6. The result is, for heterozygotes, a 2×2 matrix B.[3] Finally, at line 14, the elements of B are added to their corresponding elements of A.

With the allele balance estimates all added to A, the second step in the main loop of Algorithm 1 is to update the profiles P such that they form a non-negative least squares solution to the equation AP = C subject to the additional requirement that the elements on the diagonal of P must be 100. Lines 16–24 implement nnls(A, C) with this additional requirement enforced on line 22. Indeed, with the omission of line 22, lines 16–24 are a general purpose implementation of the nnls function. This implementation is based on [1].

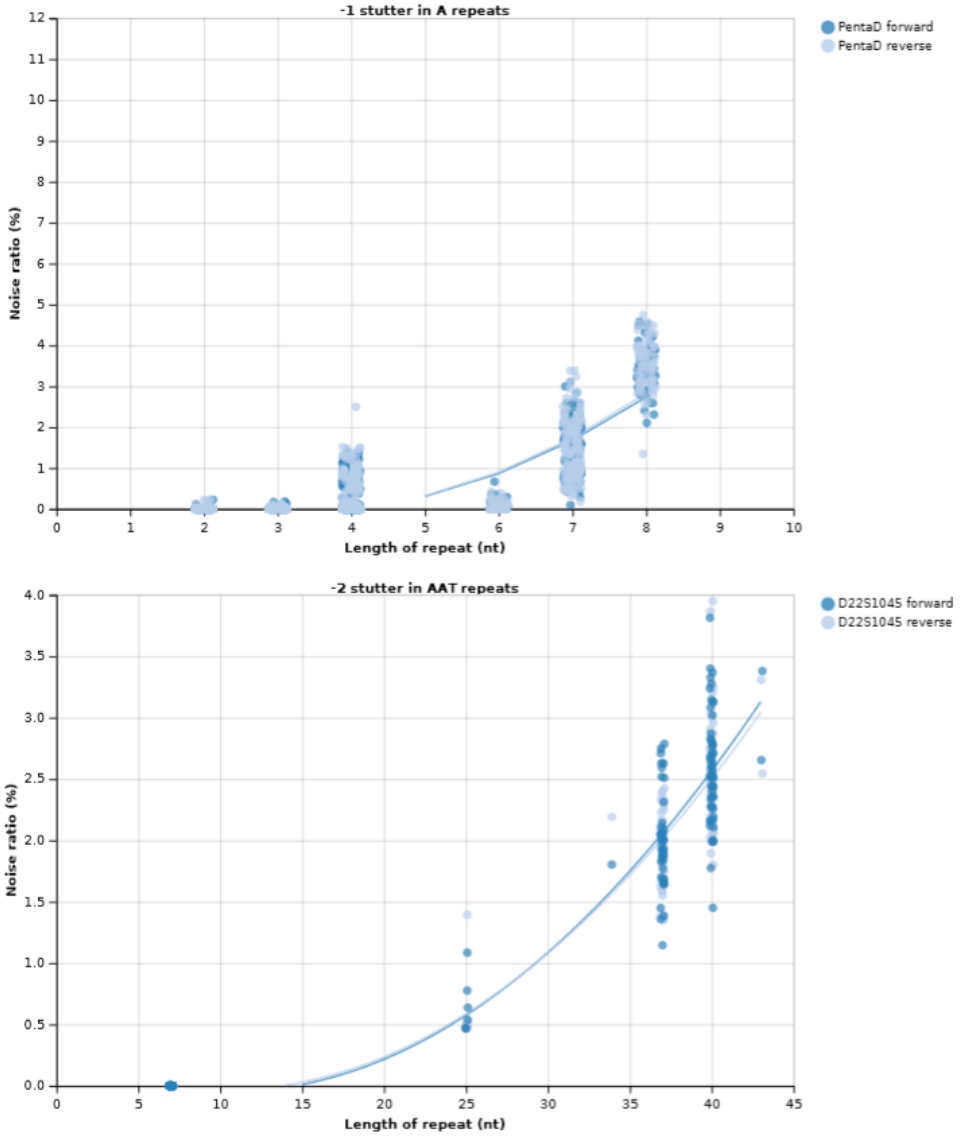
Separate profiles for the numbers of forward and reverse reads can be constructed by doubling the number of columns in P and C, where the left half corresponds to the forward stand and the right half to the reverse strand. This ensures that the same allele balance matrix A is used for both strands.

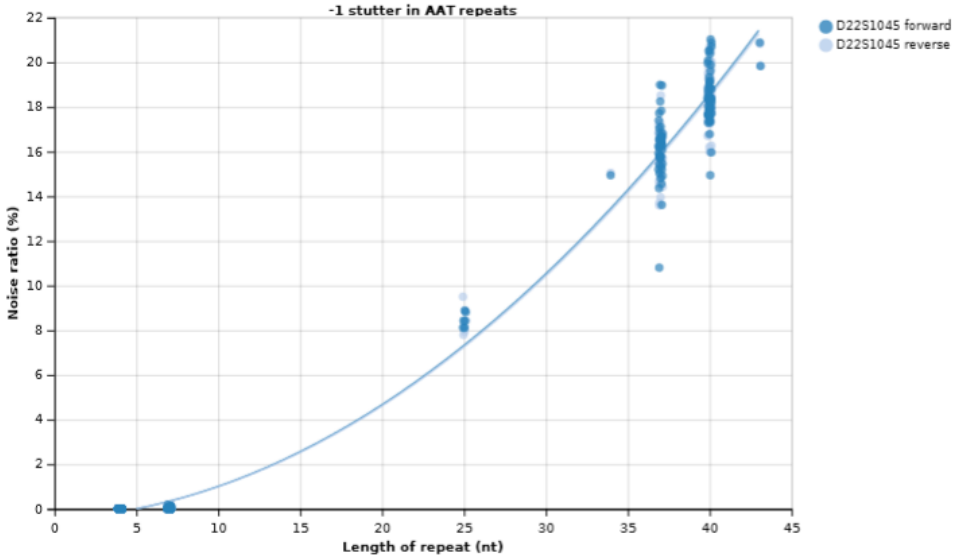
## References

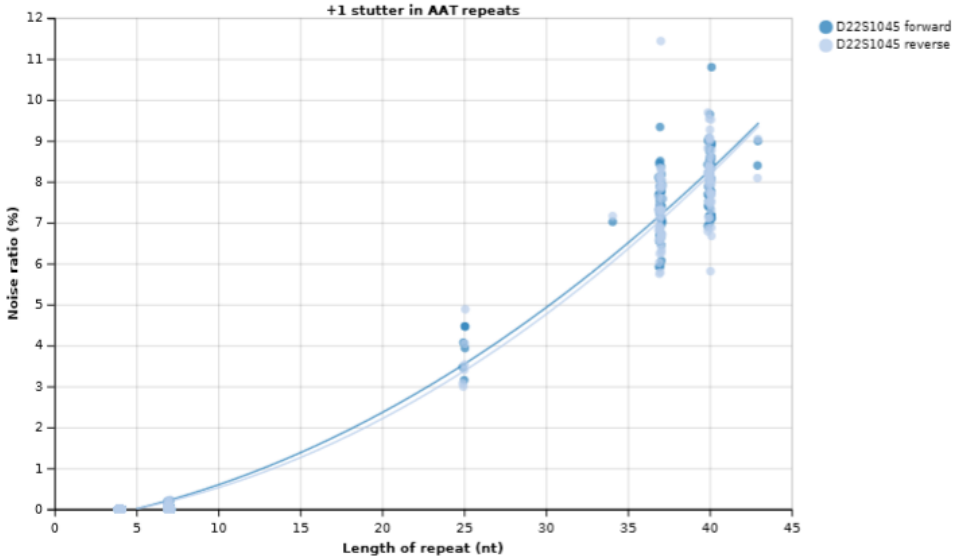
1. M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In: Independent Component Analysis and Signal Separation. Springer, 2009, pp. 540–547.

Chapter 5

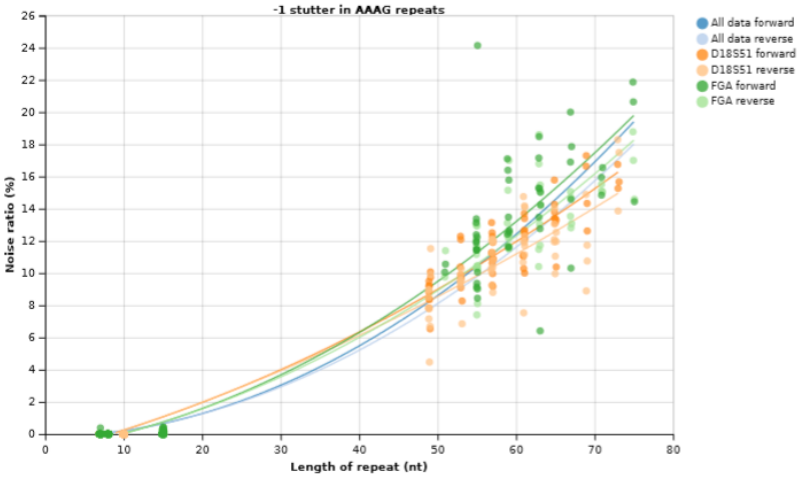
**Supplementary Figure 1 –** Stutter model based on a reference database of 429 samples

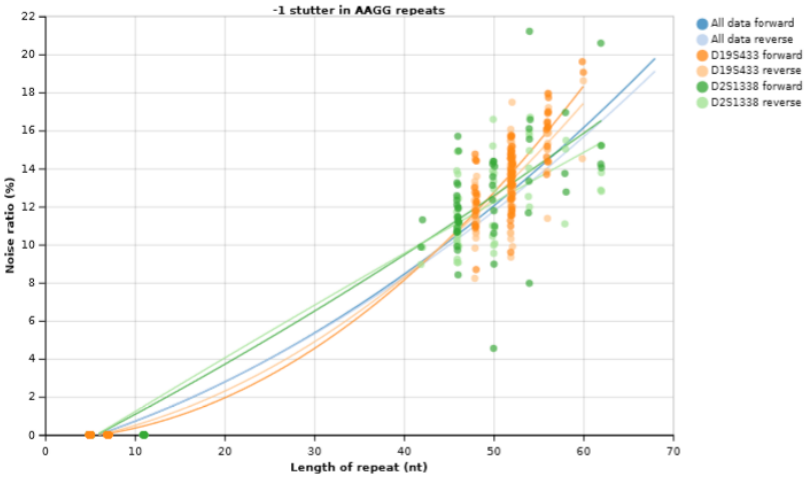


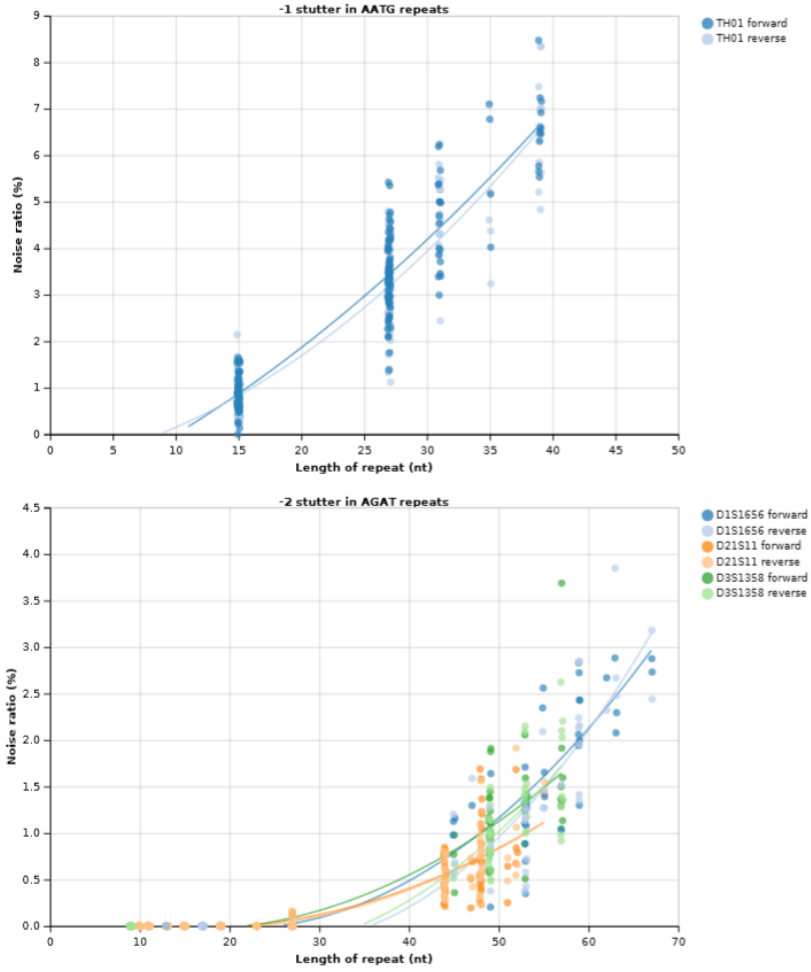

-1 stutter in AAT repeats
Noise ratio (%)
Length of repeat (nt)
D22S1045 forward
D22S1045 reverse


+1 stutter in AAT repeats
Noise ratio (%)
Length of repeat (nt)
D22S1045 forward
D22S1045 reverse

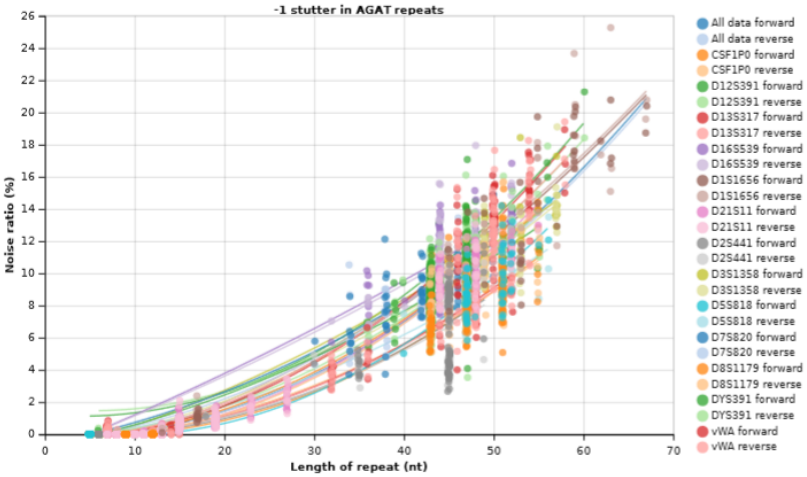
Chapter 5

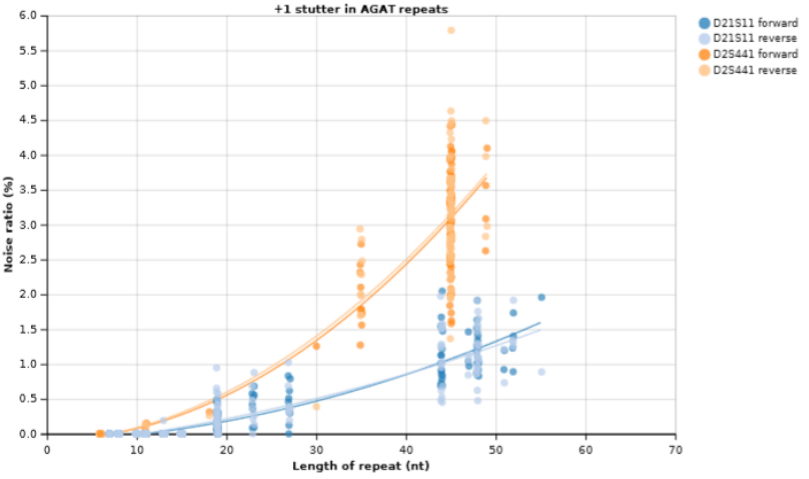

-1 stutter in AAAG repeats
Noise ratio (%)
Length of repeat (nt)
All data forward
All data reverse
D18S51 forward
D18S51 reverse
FGA forward
FGA reverse

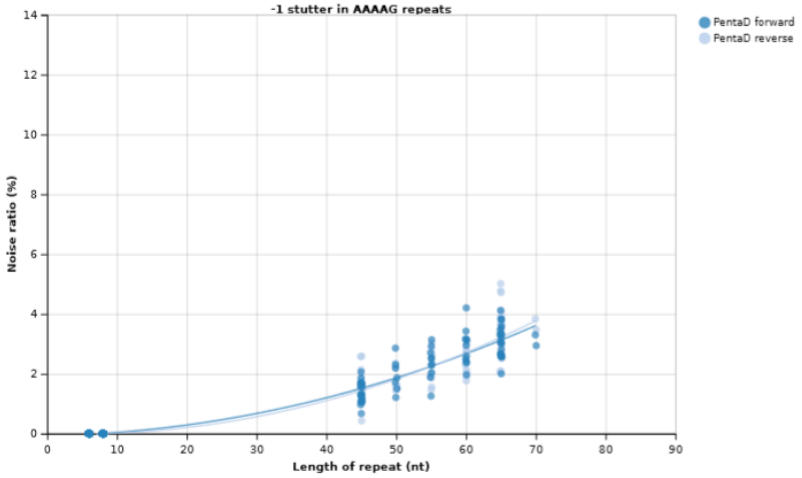

-1 stutter in AAGG repeats
Noise ratio (%)
Length of repeat (nt)
All data forward
All data reverse
D19S433 forward
D19S433 reverse
D2S1338 forward
D2S1338 reverse


-1 stutter in AATG repeats
Noise ratio (%)
Length of repeat (nt)
TH01 forward
TH01 reverse
-2 stutter in AGAT repeats
Noise ratio (%)
Length of repeat (nt)
D1S1656 forward
D1S1656 reverse
D21S11 forward
D21S11 reverse
D3S1358 forward
D3S1358 reverse

Chapter 5

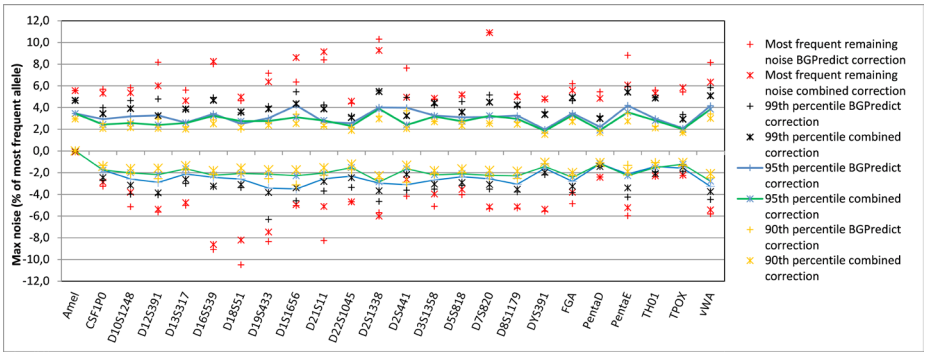

-1 stutter in AGAT repeats
Noise ratio (%)
Length of repeat (nt)
All data forward
All data reverse
CSF1P0 forward
CSF1P0 reverse
D12S391 forward
D12S391 reverse
D13S317 forward
D13S317 reverse
D16S539 forward
D16S539 reverse
D1S1656 forward
D1S1656 reverse
D21S11 forward
D21S11 reverse
D2S441 forward
D2S441 reverse
D3S1358 forward
D3S1358 reverse
D5S818 forward
D5S818 reverse
D7S820 forward
D7S820 reverse
D8S1179 forward
D8S1179 reverse
DYS391 forward
DYS391 reverse
vWA forward
vWA reverse


+1 stutter in AGAT repeats
Noise ratio (%)
Length of repeat (nt)
D21S11 forward
D21S11 reverse
D2S441 forward
D2S441 reverse



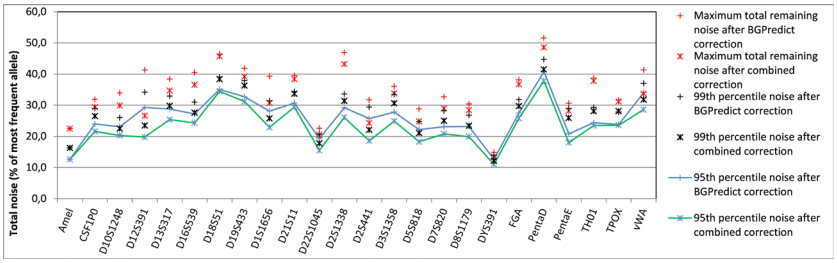

**Supplementary Figure 2** – Comparison of remaining noise using different correction settings

a) Most frequent noise variant for each locus after correction in 429 reference samples, BGPredict correction vs combined BGEstimate and BGPredict correction

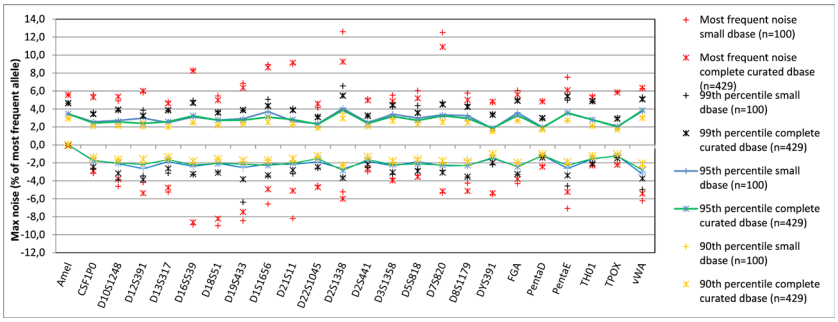


Chapter 5

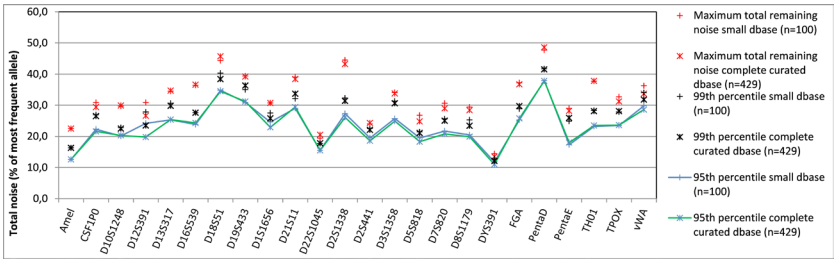
b) Total remaining noise for each locus after correction in 429 reference samples, BGPredict correction vs combined BGEstimate and BGPredict correction




c) Most frequent noise variant for each locus after combined correction in 429 reference samples, correction based on 100 vs 429 reference samples




d) Total remaining noise for each locus after correction in 429 reference samples, correction based on 100 vs 429 reference samples




*a) The dotplot displays the most frequently observed noise that remained after correction in any of the 429 analysed reference samples when performing the correction using BGPredict only (based on the stutter model) or a combination of BGEstimate and BGPredict. In addition, the 99th and the 95th percentile are plotted to illustrate the variation in remaining noise.*
*b) The dotplot displays the highest total observed noise (cumulative percentage of the reads of the most frequent allele) that remained after correction in any of the 429 analysed reference samples when performing the correction using BGPredict only (based on the stutter model) or a combination of BGEstimate and BGPredict. In addition, the 99th and the 95th percentile are plotted to illustrate the variation in remaining noise.*
*c) The dotplot displays the most frequently observed noise that remained after correction in any of the 429 analysed reference samples when performing the correction based on all 429 samples or only 100 randomly selected samples from this database. In addition, the 99th, 95th, and 90th percentile are plotted to illustrate the variation in remaining noise.*
*d) The dotplot displays the highest total observed noise (cumulative percentage of the reads of the most frequent allele) that remained after correction in any of the 429 analysed reference samples when performing the correction based on all 429 samples or only 100 randomly selected samples from this database. In addition, the 99th and 95th percentile are plotted to illustrate the variation in remaining noise.*

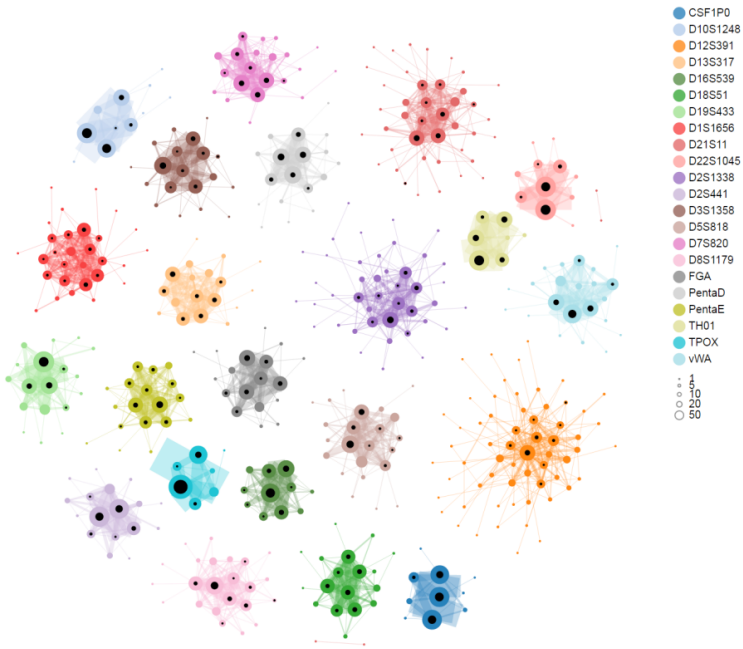
**Supplementary Figure 3 - Visualisation of the frequency of each allele and the frequency of co-occurence with other alleles as heterozygous genotypes**

The graphs depict every allele among the reference samples as a circle. The size of the circle corresponds to the number of samples with that particular allele. A black inner circle depicts the number of homozygotes.

The circles of two alleles are connected by a line whenever samples exist that have a combination of the two connected alleles. The thickness of the line corresponds to the number of heterozygotes with that particular combination of alleles.

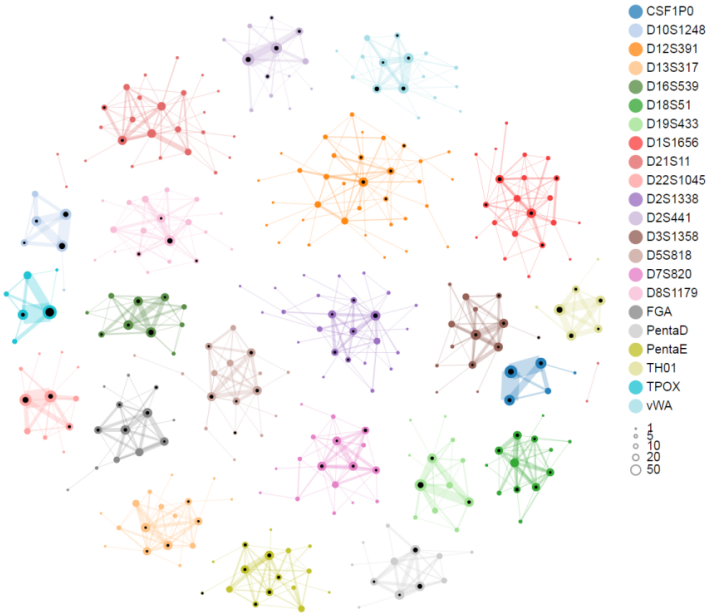
To fit the criteria to create a BGEstimate noise profile, alleles need to be present as a homozygous genotype (displayed as a black circle) or be connected with at least three other alleles that must also fit these criteria.
These figures were generated using the command 'fdstools vis allele'.

**a)** Allele visualisation of 429 reference samples

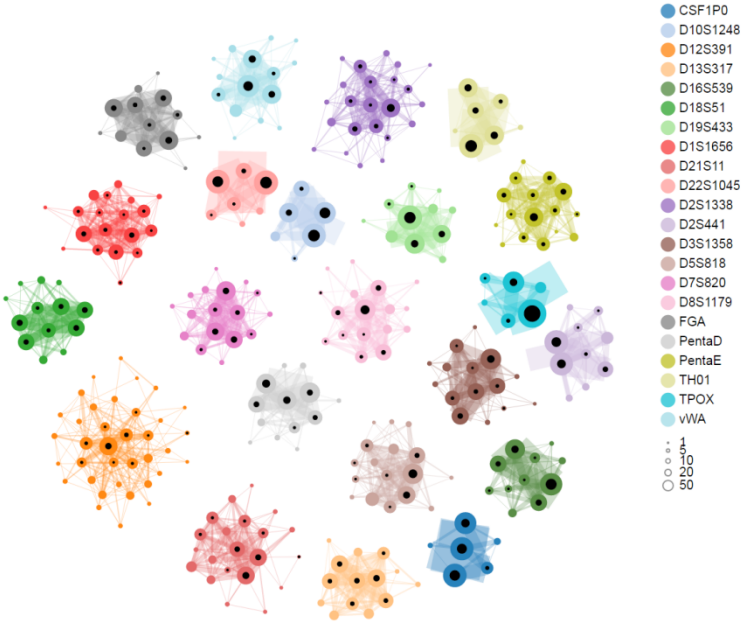


Chapter 5

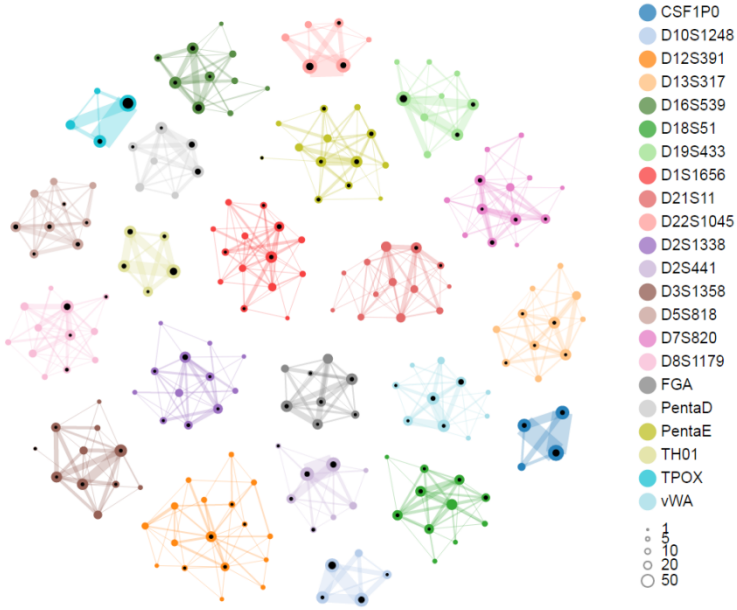
**b)** Allele visualisation of 100 reference samples




**c)** Allele visualisation of 429 reference samples after removing alleles that do not meet thresholds for determining a reliable noise profile



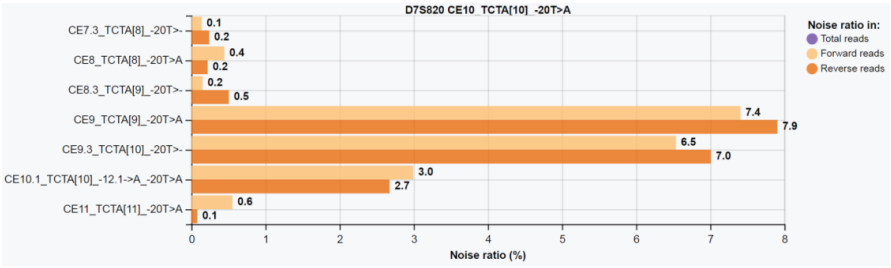
**d)** Allele visualisation of 100 reference samples after removing alleles that do not meet thresholds for determining a reliable noise profile



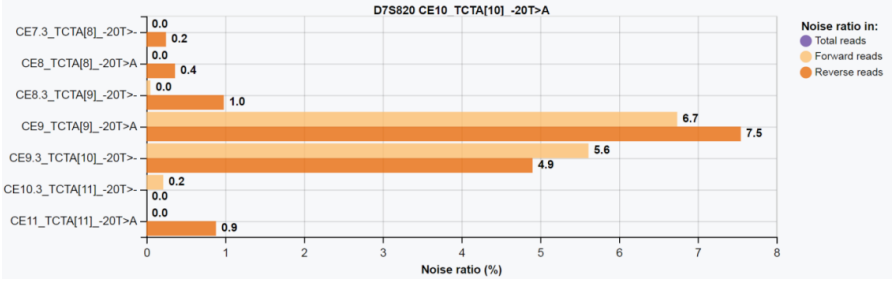
Chapter 5

**Supplementary Figure 4** – Noise profiles of D7S820 allele CE10_TCTA[10]_-20T>A estimated from high and low-coverage samples

**a)** Noise profile of D7S820 allele CE10_TCTA[10]_-20T>A estimated from high-coverage samples




**b)** Noise profile of D7S820 allele CE10_TCTA[10]_-20T>A estimated from low-coverage samples



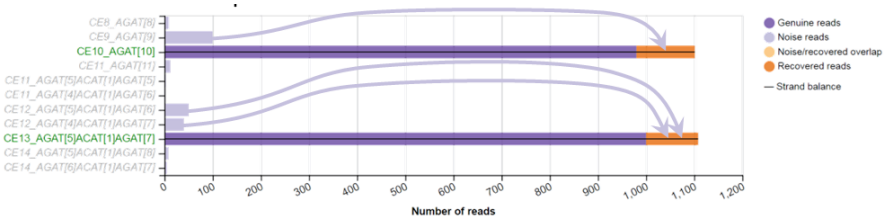

*Noise profiles created with BGEstimate based on a selection of a) 71 high-coverage samples (82,000–350,000 total reads) and b) 70 low-coverage samples (8,000–44,000 total reads). The noise ratio is shown for each systemic noise sequence observed. It is clear that for the low-percentage noise in the low-coverage noise profile, more strand bias is introduced due to single-strand drop-out of this noise caused by insufficient coverage of the reference samples.*

## Supplementary Figure 5a – Explanation of a sequence profile for raw, filtered and corrected data

### Raw data

### Noise reads filtered out

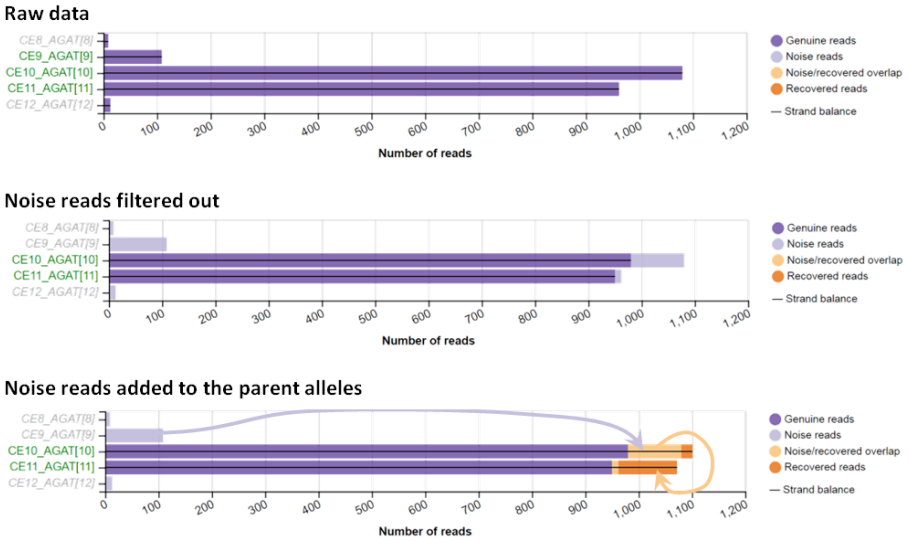
### Noise reads added to the parent alleles




*Sequence profiles for three different stages of the analysis for a simple reference sample without overlap between stutter and genuine alleles.*

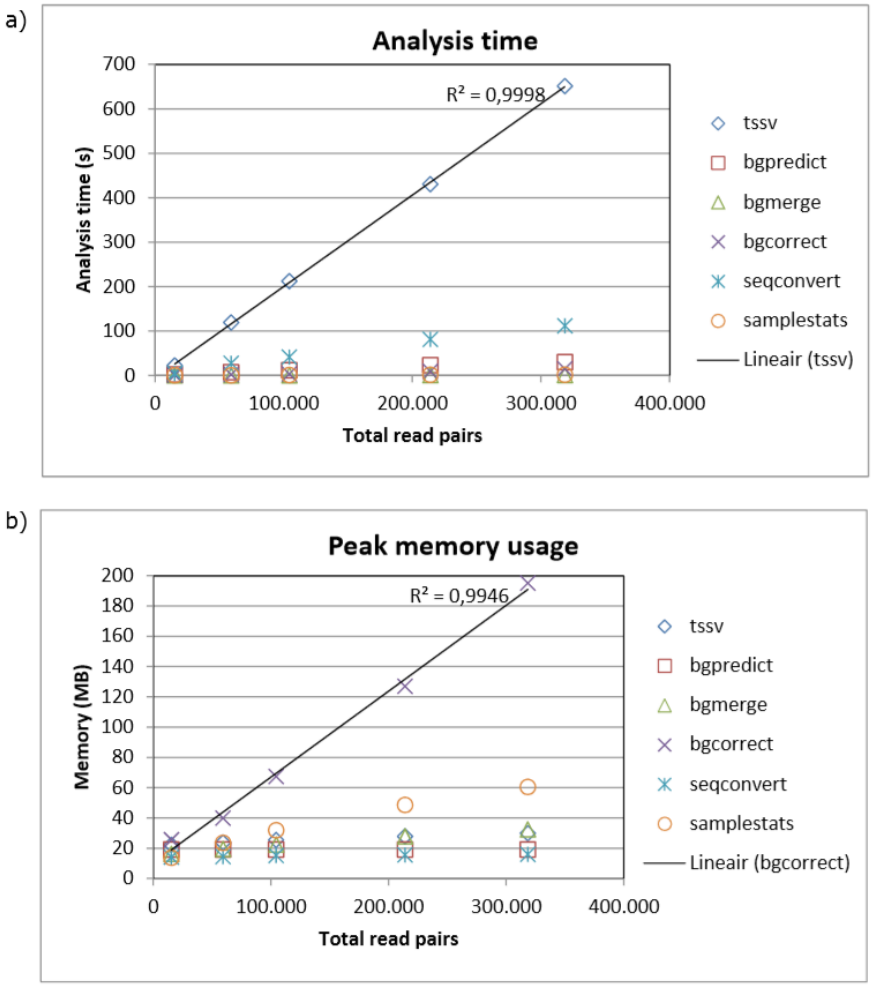
***On top,*** *raw read counts are displayed for each observed variant.*
***In the middle****, noise reads are filtered out (based on the observed reproducible noise for each allele in the reference database). Filtered noise reads are displayed in light purple.*
***At the bottom****, filtered reads are added to the parent allele (as determined by the noise profiles) as recovered reads marked in dark orange.The lines in the bars indicate the strand balance; the line is drawn near the top of the bar if the majority of reads of a sequence is on the forward strand, near the bottom of the bar if the majority of reads is on the reverse strand, and in the middle of the bar in the absence of strand bias. Sequences displayed in green in the graphs are the alleles that the software infers to be genuine alleles in the sample, based on a threshold of 1.5% of the total number of reads of the locus.*

**Supplementary Figure 5b –** Explanation of a sequence profile for raw, filtered and corrected data




*Sequence profiles for three different stages of the analysis for a reference sample with overlap between stutter and genuine alleles.*
***On top**, raw read counts are displayed for each observed variant.*
***In the middle**, noise reads are filtered out (based on the observed reproducible noise for each allele in the reference database). Filtered noise reads are displayed in light purple. Note that part of allele CE10 is filtered out as noise from allele CE11.*
***At the bottom**, filtered reads are added to the parent allele (as determined by the noise profiles) as recovered reads marked in dark orange. For allele CE10, part of the reads are removed as noise, but some reads are recovered as well. The overlap of this filtered noise and recovered reads is marked in light orange. The lines in the bars indicate the strand balance; the line is drawn near the top of the bar if the majority of reads of a sequence is on the forward strand, near the bottom of the bar if the majority of reads is on the reverse strand, and in the middle of the bar in the absence of strand bias. Sequences displayed in green in the graphs are the alleles that the software infers to be genuine alleles in the sample, based on a threshold of 1.5% of the total number of reads of the locus.*

**Supplementary Figure 6 –** Required time and memory for the analysis of case samples




*The dot plots display the registered time (a) and the peak memory usage (b) of analysis for five samples for each tool of the standard casework analysis pipeline. The analysis was performed using a single core of an Intel(R) Xeon(R) E5-2620 processor at 2.00 GHz. The analysis time is mostly consumed by TSSV and the highest memory demand is measured for the tool BGCorrect . Both the analysis time and memory increase more or less linearly when the coverage of a sample is increased.*

**Supplementary Table 1 –** Currently available tools and visualisations in FDSTools

| | |
|---|---|
| **General** | |
| Pipeline | Automatically run complete, predefined analysis pipelines using the other tools in the package. Recommended starting point for new users. |
| TSSV | Link raw reads in a FastA or FastQ file to markers and count the number of reads for each unique sequence. Wrapper around the TSSV-Lite program. |
| Vis | Create a data visualisation web page or Vega graph specification. |
| Seqconvert | Convert between raw sequences, TSSV-style sequences (shortened notation of STRs), and allele names. |
| BGMerge | Merge multiple files containing background noise profiles. Used to extend the output of BGEstimate with output of BGPredict. |
| Library | Create an empty FDSTools library file. |
| Libconvert | Convert between TSSV and FDSTools library file formats. Primarily useful for users migrating from the older, standalone TSSV programme to FDSTools. |
| **Reference sample analysis** | |
| Stuttermark | Mark potential stutter products by assuming a fixed maximum percentage of stutter product with respect to the parent sequence. Used to mask stutter products for Allelefinder. |
| Allelefinder | Find true alleles in reference samples and detect possible contaminations. |
| BGHomRaw | Compute noise ratios for all noise detected in homozygous reference samples. |
| BGHomStats | Compute allele-centric statistics for background noise in homozygous reference samples (min, max, mean, sample variance). |
| BGEstimate | Estimate allele-centric background noise profiles (means) from reference samples. |
| Stuttermodel | Train a stutter prediction model using homozygous reference samples. |
| BGAnalyse | Find contaminated or otherwise unsuitable reference samples. |
| **Case sample analysis** | |
| BGPredict | Predict background profiles of new alleles based on a model of stutter occurrence obtained from Stuttermodel. |
| BGCorrect | Match background noise profiles to samples to filter and correct for systemic noise. |
| FindNewAlleles | Mark all sequences that are not in a list of known (e.g., previously encountered) allelic sequences. |
| Samplestats | Compute various statistics for each sequence in a given sample data file. Can also call alleles and filter sequences using these statistics. |

| **Visualisations** | |
|---|---|
| Samplevis | Visualise and interpret sample data files. The web page version of Samplevis offers many features to help interpreting the sample, such as automatically marking sequences that match given criteria as 'allele'. |
| Profilevis | Visualise background noise profiles obtained with BGEstimate, BGHomStats, and BGPredict. |
| BGRawvis | Visualise raw background noise data obtained with BGHomRaw. |
| Stuttermodelvis | Visualise models of stutter obtained from Stuttermodel. |
| Allelevis | Visualise the allele list obtained from Allelefinder as a graph in which the nodes correspond to alleles and the edges correspond to heterozygous samples combining the connected alleles. |
| BGAnalysevis | Visualise remaining background noise levels as obtained from BGAnalyse. |

**Supplementary Table 2 –** Effects of criteria for admission of alleles to noise profile estimation

Number of alleles for which a BGEstimate noise profile can be obtained in our 429 sample reference set when applying different criteria. These comprise the minimum number of different heterozygous genotypes per allele, minimum number of samples per allele and minimum number of homozygous samples per allele. When a criterion is varied, the other criteria are kept at the minimum value possible which is at least 1 heterozygous genotype per allele, 1 sample per allele and 0 homozygous samples per allele. The criterion of the minimum number of different heterozygotes per allele does not apply if the allele is present in at least one homozygote.

When the settings are more stringent, BGEstimate noise profiles are obtained for fewer alleles. The results for the settings selected for this study are indicated in the rightmost column labelled 'used settings' and represent at least 3 different heterozygous genotypes per allele or, if a homozygote is available, at least 2 samples per allele.
Note that three different alleles have been detected for the gender locus Amel, which is due to the detection of 2 sequence variants for the X allele.

| Criteria | Heterozygous genotypes per allele (or one homozygote available) | | | | | Samples per allele (homozygous or heterozygous) | | | | | Homozygous samples per allele (situation for *Stuttermodel*) | | | | | Used settings |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1+ | 2+ | **3+** | 4+ | 5+ | 1+ | **2+** | 3+ | 4+ | 5+ | 1+ | 2+ | 3+ | 4+ | 5+ | |
| Locus | Alleles | | | | | | | | | | | | | | | |
| Amel | 3 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 1 | 1 |
| CSF1P0 | 10 | 8 | 6 | 5 | 4 | 10 | 8 | 6 | 6 | 6 | 4 | 4 | 4 | 4 | 3 | 6 |
| D10S1248 | 11 | 7 | 6 | 6 | 6 | 11 | 7 | 6 | 6 | 6 | 5 | 4 | 4 | 4 | 3 | 6 |
| D12S391 | 61 | 44 | 36 | 29 | 27 | 61 | 45 | 36 | 31 | 28 | 10 | 5 | 3 | 2 | 2 | 36 |
| D13S317 | 18 | 14 | 14 | 14 | 13 | 18 | 14 | 14 | 14 | 14 | 7 | 7 | 7 | 6 | 6 | 14 |
| D16S539 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 8 | 4 | 4 | 4 | 4 | 11 |
| D18S51 | 18 | 15 | 13 | 13 | 12 | 18 | 15 | 14 | 14 | 14 | 7 | 7 | 7 | 6 | 5 | 13 |
| D19S433 | 18 | 15 | 11 | 10 | 10 | 18 | 16 | 14 | 13 | 10 | 5 | 4 | 3 | 3 | 3 | 11 |
| D1S1656 | 29 | 22 | 19 | 17 | 17 | 29 | 22 | 20 | 18 | 17 | 13 | 7 | 4 | 3 | 2 | 19 |
| D21S11 | 45 | 28 | 21 | 18 | 15 | 45 | 29 | 22 | 18 | 17 | 9 | 7 | 3 | 3 | 3 | 21 |
| D22S1045 | 11 | 9 | 7 | 7 | 6 | 11 | 11 | 8 | 8 | 8 | 5 | 4 | 3 | 3 | 3 | 7 |
| D2S1338 | 42 | 28 | 23 | 22 | 17 | 42 | 28 | 24 | 22 | 22 | 10 | 8 | 5 | 4 | 3 | 23 |
| D2S441 | 17 | 13 | 10 | 10 | 9 | 17 | 14 | 12 | 10 | 10 | 6 | 4 | 4 | 3 | 3 | 10 |
| D3S1358 | 19 | 16 | 15 | 14 | 11 | 19 | 16 | 15 | 14 | 13 | 8 | 5 | 5 | 3 | 3 | 15 |
| D5S818 | 23 | 18 | 17 | 14 | 13 | 23 | 18 | 17 | 16 | 15 | 8 | 5 | 4 | 3 | 3 | 17 |
| D7S820 | 21 | 19 | 17 | 16 | 15 | 21 | 19 | 17 | 16 | 16 | 8 | 6 | 5 | 5 | 4 | 17 |
| D8S1179 | 23 | 19 | 17 | 14 | 14 | 23 | 19 | 18 | 18 | 16 | 9 | 5 | 5 | 4 | 4 | 17 |
| DYS391 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 4 | 4 | 6 | 5 | 5 | 4 | 4 | 5 |
| FGA | 18 | 15 | 13 | 10 | 9 | 18 | 15 | 14 | 11 | 9 | 7 | 7 | 6 | 4 | 4 | 13 |
| PentaD | 18 | 14 | 13 | 10 | 9 | 18 | 14 | 13 | 13 | 11 | 6 | 6 | 5 | 5 | 5 | 13 |
| PentaE | 18 | 16 | 16 | 15 | 14 | 18 | 16 | 16 | 15 | 14 | 11 | 9 | 8 | 7 | 4 | 16 |
| TH01 | 7 | 7 | 7 | 7 | 5 | 7 | 7 | 7 | 7 | 7 | 5 | 5 | 5 | 5 | 4 | 7 |
| TPOX | 7 | 6 | 6 | 6 | 6 | 7 | 6 | 6 | 6 | 6 | 4 | 3 | 3 | 3 | 2 | 6 |
| vWA | 22 | 16 | 15 | 11 | 9 | 22 | 16 | 15 | 14 | 12 | 5 | 5 | 5 | 3 | 3 | 15 |

# Chapter 6

## Short hypervariable microhaplotypes: A novel set of very short high discriminating power loci without stutter artefacts

Forensic Science International Genetics 2018 Jul;35:169-175
Published online May 22, 2018

Kristiaan J. van der Gaag
Rick H. de Leeuw
Jeroen F.J. Laros
Johan T. den Dunnen
Peter de Knijff

Supplementary material is available via https://www.fsigenetics.com

# Abstract

Since two decades, short tandem repeats (STRs) are the preferred markers for human identification, routinely analysed by fragment length analysis. Here we present a novel set of short hypervariable autosomal microhaplotypes (MH) that have four or more SNPs in a span of less than 70 nucleotides (nt). These MHs display a discriminating power approaching that of STRs and provide a powerful alternative for the analysis of forensic samples that are problematic when the STR fragment size range exceeds the integrity range of severely degraded DNA or when multiple donors contribute to an evidentiary stain and STR stutter artefacts complicate profile interpretation. MH typing was developed using the power of massively parallel sequencing (MPS) enabling new powerful, fast and efficient SNP-based approaches. MH candidates were obtained from queries in data of the 1000 Genomes, and Genome of the Netherlands (GoNL) projects. Wet-lab analysis of 276 globally dispersed samples and 97 samples of nine large CEPH families assisted locus selection and corroboration of informative value. We infer that MHs represent an alternative marker type with good discriminating power per locus (allowing the use of a limited number of loci), small amplicon sizes and absence of stutter artefacts that can be especially helpful when unbalanced mixed samples are submitted for human identification.

# Introduction

Short Tandem Repeats (STRs) have been the preferred marker for human identification for over two decades. Although the high degree of variation at STR-loci [1] provides useful discriminatory power for forensic and paternity cases, STRs are not the ideal marker type when degraded or mixed samples are involved. The interpretation of samples that have multiple contributors (and especially those with unequal contributions) can be complicated by the effects of slippage of DNA polymerases at the repeat stretches, resulting in stutter peaks that reside foremost at the n-1 position (representing products of one repeat unit less than the original allele length) [2]. Also, STR fragments with higher repeat numbers can be too long to allow amplification in severely degraded DNA samples [3]. The ideal forensic marker has a high degree of variation per fragment, allows for the design of small amplicons and is devoid of the production of stutter artefacts. In 1999, Jin et al. [4] published such a marker: a hypervariable fragment close to the MX1 gene on chromosome 21 containing several single nucleotide polymorphisms (SNPs) within a stretch of 100 nucleotides that proved to be informative in population genetics. However, at that time, the full power of such loci could not be exploited since routine analysis performed by Sanger sequencing only provides consensus information for each position without revealing how the variants of different SNPs within a fragment are connected (as a

Chapter 6

microhaplotype).

The development of massive parallel sequencing (MPS) platforms has provided promising new possibilities, especially for marker types that reveal their discriminatory value upon sequencing analysis. For STRs, MPS reveals substantial sequence variation in addition to repeat length, thereby increasing the discriminatory power of STRs compared to conventional fragment analysis [5,6]. However, even with MPS, the complication of stutter formation in the interpretation of complex mixtures remains. MPS also allows for the analysis of large panels of SNPs when severely degraded DNA is involved [7,8]. Recently, microhaplotypes (MH) or fragments with two to four SNPs, within a 200 nucleotide (nt) stretch, have been described [9] as an alternative for STR typing of mixtures. Note that both SNPs and MHs do not allow for searches in DNA databases that are generally built from STR data and that relevant reference samples need to be available. Here we examine a new set of short hypervariable haplotypes, consisting of four or more SNPs contained in genomic fragments of less than 70 nt. We indicate that these MHs represent a discriminating power close to that of STR loci and facilitate mixture analysis without the hindrance of stutter. The data of the 1000 genome [10] and the GoNL projects [11] were used to identify potentially useful MHs. To confirm the genetic variation of these loci, data from 276 individuals of three globally distinct populations and 97 DNA samples from nine large families were analysed using MPS. Variant data of the most promising MHs was made publically available via the Leiden Open (source) Variation Database (LOVD) [12,13].

# Material and Methods

## Marker selection

We screened the Variant Call Format (VCF) files of the African samples of the 1000 Genome and all of the GoNL project samples (Dutch selected for European ancestry) for genome fragments spanning 100 nt containing six SNPs with Minor Allele Frequencies (MAFs) in the relevant population ≥ 0.1. To select a subset of fragments for wetlab confirmation from the total set (which was >100,000 fragments), filtering was performed using the following criteria:

- At least four out of six SNPs need to occur in both the 1000 Genomes and the GoNL projects, we do this for both validation purposes, but also to confirm that the variants are present in samples from different ancestry;
- All six SNPs should be within 70 nt to maximise possibilities for small amplicon design (the number of fragments from the 100 nt interval search allowed us to further reduce the fragment size);
- At least five of the six SNPs should not share the same MAF to maximise the number of possible haplotypes (many identical frequencies suggests perfect

linkage and lack of variation);
- One of the SNPs should have a MAF of at least 0.4 to avoid overrepresentation of one haplotype.
- The highest and the lowest MAF of the SNPs within a fragment should have a difference of at least 0.2 to maximise variation in frequencies between haplotypes.
- The genomic distance to the nearest fragment should be at least 100,000 nt.

For the remaining fragments, the MH sequence spanning all SNPs plus 60 nucleotides up- and downstream was checked for homology in the genome using BLAST. Fragments with multiple hits (both within one chromosome and on different chromosomes) were discarded. For the remaining fragments, primer design was performed using primer3 v4.0.0 allowing a Tm of 57-63 °C, a primer length of 18-27 nt and amplicon sizes of 80-120 bp. Fragments containing repeating elements (repeated four or more times) or single nucleotide stretches over 8 nt were discarded. After primer design, the complete amplicon was checked again for homology using BLAST to achieve the final set of fragments for wet-lab testing. The set was completed by designing an amplicon representing the most variable part of the fragment described in Jin et al. [4] which includes seven of the nine SNPs excluding the last SNP of the 248 bp fragment and the SNP in the additional 227 bp fragment.

## Microhaplotype selection by monoplex PCR and Ion PGM analysis

To confirm the sequence variation for the selected candidates, 92 MHs were sequenced in 15 samples using the Ion PGM™ System according to the manufacturer's procedures. Five Dutch, three Bhutanese, two Ghanese, two Pygmy and three Amerindian samples from the HGDP CEPH-panel [14] were amplified in monoplex reactions. PCRs were performed using a 10µl reaction containing PCR buffer (Life Technologies), 3mM MgCl2, 0.2µM dNTPs, primer concentrations of 0.1-0.8µM, 0.6 units Amplitaq Gold (Life Technologies) and 1.5ng DNA. PCR specificity was checked using the Qiaxcel system (Qiagen) according to the manufacturer's procedures and MHs for which additional bands were visible in eight or more samples were discarded. All monoplex PCR products of the same sample were pooled and adapters were ligated to the amplicon pool using the IonXpress library preparation kit according to the manufacturer's procedures (Ion Torrent / Thermo Fisher). Sequencing was performed using the PGM™ System according to the manufacturer's procedures (Thermo Fisher) and data analysis was performed using FDSTools [5].

## MH confirmation by multiplex PCR and MiSeq analysis

A multiplex PCR was designed (amplicon sizes 87-126bp including primers) to examine the most informative 16 MHs in more detail. To test for global variation, 99 samples from the Netherlands [15], 87 Asian samples of the Han Chinese and Japan HapMap panel [16], and 90 African samples of the Luhya (Kenya), Yoruba (Kenya / Nigeria) and Maasai (Kenya) HapMap panel were analysed. To confirm stable transmission of the variants, nine CEPH families (family 12, 66, 1328, 1347, 13281, 13291, 13292, 13293 and 13294; 97 samples in total) were analysed. Multiplex PCR was performed using a total volume of 12.5µl containing PCR buffer (Life Technologies), 4mM MgCl2, 0.4µM dNTP, primer concentrations of 0.03-0.35µM, 2.5 units Amplitaq Gold (Life Technologies) and 1.5ng DNA. Adapters were ligated using the KAPA HTP Library Preparation Kit for Illumina® platforms according to the manufacturer's procedures (KAPA Biosystems / Roche) and sequencing was performed using the MiSeq® Sequencer according to the manufacturer's procedures (Illumina, v3 chemistry). Data analysis was performed using FDSTools [5]. Data for all observed sequence variants of the final set of MHs was submitted to LOVD (http://databases.lovd.nl/DNA_profiles/) [12,13].

## Statistical analysis

All statistic calculations were performed on haplotype data (not separately for each SNP). Population statistics were calculated for all populations (Chinese / Japanese and Kenyan / Nigerian were respectively grouped together) using Powerstats [17] and Genalex [18]. The power to detect mixtures (chance to observe a third allele for at least one locus) was calculated as described by Phillips et al. [19] by adding an extra sheet to the Powerstats Excel sheet (file available upon request). An Excel sheet was used to check for correct transmission of variants in the CEPH families. Neighbour joining networks were drawn for the 16 MHs of the final multiplex using Network 5 and Network Publisher [20] using the homologous sequence of a Chimpanzee as out-group. Recombination rates were retrieved for all MHs from the HapMap recombination maps [21] and the average number of meioses for recombination to occur within the fragment was calculated considering the fragment lengths. To test the potential of these fragments to inform about geographic ancestry, STRUCTURE [22] was run 100 times with a K-value of 2, 3 and 4. CLUMPAK [23] was used to combine and visualise the data of the repeated runs.

# Results

## MH candidate selection

A search in the VCF files for genomic intervals of 100 nt containing at least six SNPs with a MAF of at least 0.1 resulted in 14,890 potential MHs in the African samples of the 1000 Genomes project and 105,129 MH candidates in the GoNL dataset. An overview of the number of remaining fragments for each chromosome after applying several filtering criteria is shown in Table 1. After checking the remaining 410 fragments for homologous regions in the genome and the possibility for PCR design, 92 fragments dispersed over the genome remained and amplicons were prepared for wet-lab testing.

**Table 1 -** Numbers of remaining short hypervariable microhaplotypes after applying several filtering criteria for selection of potentially informative fragments

| | | | Chromosome | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Selection Criteria | Total | % of total* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| At least four SNPs in 1000 Genome and GoNL selection | 10464 | 10.0% | 181 | 332 | 277 | 291 | 133 | 5858 | 338 | 173 | 183 | 185 |
| Clusters <70bp | 5612 | 5.3% | 89 | 155 | 120 | 128 | 64 | 3435 | 175 | 73 | 91 | 105 |
| At least five SNPs with different MAF within the fragment | 4726 | 4.5% | 80 | 139 | 89 | 102 | 57 | 2925 | 146 | 58 | 75 | 87 |
| Max MAF in fragment at least 0.4 | 3910 | 3.7% | 61 | 108 | 67 | 81 | 41 | 2402 | 118 | 57 | 69 | 77 |
| Max within fragment freq-distance > 0.2 | 2882 | 2.7% | 50 | 71 | 46 | 61 | 31 | 1863 | 84 | 42 | 47 | 51 |
| Distance between two fragments > 100,000 nt | 410 | 0.4% | 26 | 16 | 21 | 28 | 13 | 32 | 23 | 20 | 16 | 20 |
| Succesful PCR-design | 92 | | 4 | 2 | 6 | 4 | 1 | 6 | 5 | 5 | 4 | 5 |

| | Chromosome | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Selection Criteria | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| At least four SNPs in 1000 Genome and GoNL selection | 185 | 223 | 210 | 410 | 103 | 227 | 244 | 115 | 255 | 282 | 115 | 144 |
| Clusters <70bp | 96 | 124 | 112 | 171 | 33 | 103 | 125 | 62 | 120 | 113 | 53 | 65 |
| At least five SNPs with different MAF within the fragment | 87 | 109 | 91 | 125 | 26 | 82 | 98 | 57 | 96 | 102 | 43 | 52 |
| Max MAF in fragment at least 0.4 | 70 | 97 | 83 | 113 | 25 | 63 | 96 | 47 | 75 | 82 | 39 | 39 |
| Max within fragment freq-distance > 0.2 | 48 | 57 | 53 | 66 | 17 | 50 | 63 | 33 | 41 | 53 | 29 | 26 |
| Distance between two fragments > 100,000 nt | 24 | 17 | 17 | 11 | 11 | 22 | 17 | 13 | 26 | 14 | 11 | 12 |
| Succesful PCR-design | 6 | 2 | 4 | 4 | 5 | 6 | 5 | 2 | 6 | 4 | 2 | 4 |

**percentages are calculated as proportion of the GoNL candidate fragments*

## MH candidate testing

From the 92 MH candidates amplified for the first set of 15 samples, 83 MHs passed the selection criterion regarding PCR specificity (no differently sized amplification products in at least eight of the 15 samples) and were subjected to sequence analysis.

The sequence variation observed for these 83 candidates was generally very low: in most cases only a single haplotype was observed. In addition, several fragments showed more than two alleles in the same sample reflecting multiple genomic copies of different sequence. Using these results as a selection criterion, 29 fragments remained with more than two haplotypes in 15 samples and no indication of fragment amplification from homologous loci based on the available data.

## Performance of MH set

A multiplex PCR was designed and 23 of the 29 fragments were successfully amplified and sequenced in 276 population samples and 97 CEPH family samples. Three of the 23 fragments revealed multiple amplification products suggesting more than one genomic location. Four fragments showed insufficient sequence variation. Thus, 16 fragments remained for which the genome positions and primer sequences are displayed in Sup. Table 1a. Microhaplotpes were named according to the suggested names by Kidd et al. [24].

The observed number of variable SNP-positions within a MH varied from four to 22 and the number of unique haplotypes varied from 4-26 as displayed in Table 2.

A sequence alignment of the observed haplotypes of each MH is displayed in Sup. Figure 1 and Sup. Table 2 displays the allele frequencies in each of the three tested populations.

Networks were drawn from the population samples for each MH to visualise the SNP-distance between the separate haplotypes and the observed number of haplotypes for each population. An example of the network of mh07PK-38311 is displayed in Figure 1. For this figure, an illustration of the fragment was included connecting the position of the SNPs with the branches in the network. Sup. Figure 2 displays the networks for each of the 16 MHs, statistics of the Chi-Square tests for Hardy-Weinberg Equilibrium are displayed in Sup. Table 3.

The observed degree of variation is different for each MH. 13 of the 16 loci result in a simple network with either no or one reticulation. The MH with the most haplotypes (mh17PK-86511) has a slightly more complex structure with a few low-frequency haplotypes that result in reticulations. MHs mh11PK-62906 and mh14PK-72639 result in complex web-like structure. mh11PK-62906 is the only fragment located in a region with a substantially elevated recombination rate. For the tested allele transfers of the selected 16 MHs in the CEPH families (144 allele transfer events for each locus in total), no inconsistent haplotype inheritance was observed. As an example, Sup. Figure 3 displays the joined family tree for CEPH family 1328, 13281, 13291, 13292, 13293 and 13294 (50 individuals in total) with the corresponding genotypes and read counts of each haplotype for mh16PK-83544.

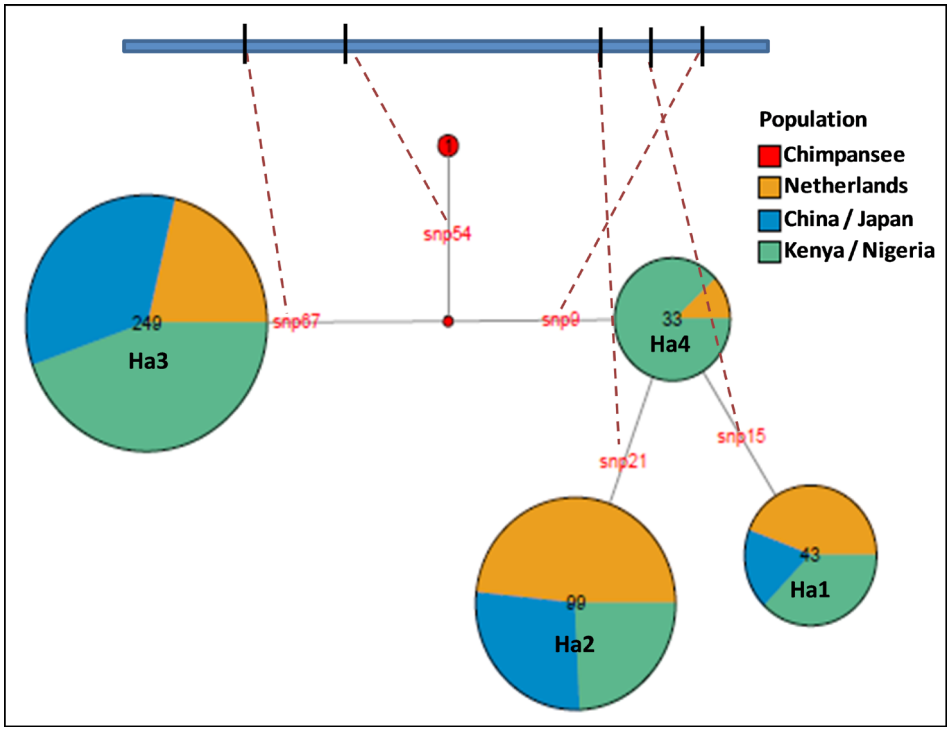
**Table 2 -** Overview of the observed variation for each Microhaplotype

| Locus | Unique observed haplotypes* | Number of observed SNP-positions in MH |
|---|---|---|
| mh06PK-24844 | 9 | 10 |
| mh06PK-25713 | 6 | 6** |
| mh07PK-38311 | 4 | 5 |
| mh08PK-46625 | 5 | 4 |
| mh10PK-62104 | 5 | 7** |
| mh11PK-62906 | 19 | 7 |
| mh11PK-63643 | 7 | 7 |
| mh14PK-72639 | 15 | 9** |
| mh15PK-75170 | 12 | 13** |
| mh16PK-83362 | 7 | 8 |
| mh16PK-83483 | 8 | 9 |
| mh16PK-83544 | 5 | 6** |
| mh17PK-86511 | 26 | 22** |
| mh18PK-87558 | 4 | 6 |
| mh22PK-104638 | 11 | 12 |
| mh21PK-MX1s | 5 | 4 |

**Each haplotype is defined as a unique observed combination of the SNP-variants within a fragment (in the tested human samples).*
*** For mh06PK-25713, mh10PK-62104, mh14PK-72639, mh15PK-75170, mh16PK-83544 and mh17PK-86511, one of the SNPs is tri-allelic. For each fragment, the number of observed unique haplotypes is displayed and the number of SNP-positions in the fragment from which these haplotypes are comprised.*

Chapter 6

**Figure 1 –** Illustration of the fragment of mh07PK-38311 and the corresponding network of the haplotypes




*On top, the fragment of mh07PK-38311 is displayed with the observed SNP positions indicated by vertical lines. Below, the network displays the distribution of each haplotype over the different tested populations and the SNP-distance between each haplotype. The circles are sized by the number of haplotypes observed in each population with colours representing the haplotypes of each analysed population. Each branch of the network is connected to the corresponding SNP in the fragment by a dotted line.*

Forensic and paternity statistics are summarised for each tested population in Sup. Table 4. The random match probability (RMP) of the total set of 16 MHs is $9.2\times10^{-13}$ for the African population, $4.4\times10^{-11}$ for the Dutch population and $1.0\times10^{-9}$ for the Asian population. In comparison, Table 3 displays the RMP for several panels of different kind of loci and Table 4 displays the power to detect a mixture (PMD) for the MHs and the tri-allelic [25] and tetra-allelic SNPs [19].

**Table 3** - Overview of the Random Match Probability for different panels of forensic loci

| Panel | number of loci | Type of loci | RMP** | Based on population | Source |
|---|---|---|---|---|---|
| Short hypervariable microhaplotypes | 16 | Micro haplotypes | $4.4\times10^{-11}$ | NL | This study |
| | | | $1.0\times10^{-9}$ | China/Japan | |
| | | | $9.2\times10^{-13}$ | Kenya / Nigeria | |
| SGM Plus® Kit | 10 | STRs | $7.9\times10^{-14}$ | African American | ThermoFisher |
| | | | $3.0\times10^{-13}$ | US Caucasian | |
| NGM™ | 15 | STRs | $1.6\times10^{-19}$ | US Hispanic | ThermoFisher |
| | | | $4.6\times10^{-20}$ | African American | |
| | | | $2.2\times10^{-19}$ | US Caucasian | |
| NGM™ <= 200 bp* | 9* | STRs | $3.1\times10^{-12}$ | US Hispanic | ThermoFisher |
| | | | $8.8\times10^{-13}$ | African American | |
| | | | $2.6\times10^{-12}$ | US Caucasian | |
| Powerplex Fusion | | STRs | $1.6\times10^{-28}$ | African American | Promega |
| | | | $2.4\times10^{-27}$ | US Caucasian | |
| | | | $2.1\times10^{-27}$ | US Hispanic | |
| | | | $1.4\times10^{-25}$ | US Asians | |
| SNPforID | 52 | SNPs | $5.0\times10^{-21}$ | European | Sanchez et al. (2006) |
| | | | $1.1\times10^{-19}$ | Somali | |
| | | | $5.0\times10^{-19}$ | Asian | |
| tri-allelic SNPs | 13 | SNPs (tri-allelic) | $3.2\times10^{-6}$ | Dutch | Westen et al. |
| | | | $4.4\times10^{-7}$ | Dutch Antilles | (2009) |
| tetra-allelic SNPs* | 24 | SNPs (tetra-allelic) | $1.5\times10^{-12}$ | European | Phillips et al. (2015) |
| | | | $5.2\times10^{-10}$ | East Asian | |
| | | | $2.0\times10^{-15}$ | African | |
| Kidd MicroHaps | 31 | Micro haplotypes | $1\times10^{-13}$ - $4\times10^{-21}$ | Various global populations | Kidd et al. (2014) |

** the boundary of 200 nt is within the range of some loci, on average 9 loci determine the RMP*
*** abbreviations: RMP = Random Match Probability*

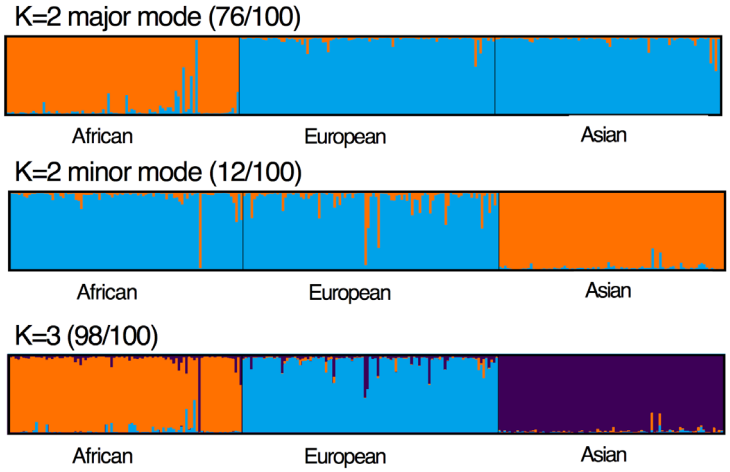
**Table 4 -** Overview of the Power of Mixture Detection for different panels of forensic loci

| Panel | number of loci | Type of loci | PMD* | Based on population | Source |
|---|---|---|---|---|---|
| Short hypervariable microhaplotypes | 16 | Micro haplotypes | 0.9989 | NL | This study |
| | | | 0.9947 | China/Japan | |
| | | | 0.9999 | Kenya / Nigeria | |
| tri-allelic SNPs | 10 | SNPs (tri-allelic) | 0.7490 | Dutch | Westen et al. (2009) |
| | | | 0.9471 | Dutch Antilles | |
| tetra-allelic SNPs* | 24 | SNPs (tetra-allelic) | 0.9939 | European | Phillips et al. (2015) |
| | | | 0.9260 | East Asian | |
| | | | 0.9999 | African | |

** abbreviations: PMD = Power of Mixture Detection*

To test the power of the 16 MHs to differentiate populations of different ancestry, 100 Structure runs were performed using two to four groups (K=2, K=3 and K=4, Figure 2). A major cluster (76 of the 100 runs) and a minor cluster (12 of the 100) was obtained for K=2 separating the African or the Asian samples respectively from the other two populations. For K=3, 98 of the 100 runs resulted in an almost complete separation of all three populations involved. For K=4, a major and a minor fourth cluster were obtained resulting in a poor differentiation of either the African or the Dutch population (data not shown).

**Figure 2** - STRUCTURE / CLUMPAK population differentiation for 16 MHs in the tested populations




*The figure displays the CLUMPAK results of 100 STRUCTURE runs, every bar displays one individual. On top, the major mode is displayed of STRUCTURE runs with K=2 derived from 76/100 repeated analyses where the African samples are mostly differentiated from Europe and Asia. In the middle, the minor K=2 mode is displayed derived from 12/100 repeated analyses where most Asian samples are differentiated from Africa and Europe. At the bottom, the results for K=3 are displayed derived from 98/100 repeated analyses where most of the samples of the three continents are properly differentiated.*

# Discussion

Detection of degraded DNA and of minor contributions in mixed samples is often complicated when conventional forensic STR typing is applied. Due to the large range of amplicon sizes for some loci and the occurrence of stutter products, it can be difficult to generate reliable and reproducible STR profiles. It would therefore be ideal to use a marker type of small amplicon sizes with a discriminating power equivalent to STRs but without the burden of stutter artefacts.

We selected hypervariable micro haplotype loci with at least six SNPs within a range of 100 nt from genomic reference data of a European and African populations and tested the final set on additional populations (including Asian samples) in order to provide a set of markers which is likely to be informative in the majority of global populations. Since the data available to us consisted merely of SNP-frequencies and did not contain any information about haplotype frequencies of the combined SNPs within a fragment, we used variation of SNP allele frequencies within each fragment as a means to maximise haplotype variation.

BLAST results of the first selection of 410 fragments exposed that many (≈25%) of the hypervariable fragments contained homologous regions in the genome, which

Chapter 6

suggested that part of the variation in the databases might have resulted from something else than actual SNP variation. After discarding those fragments, sequencing of the remaining fragments still revealed much less variation than we observed in the data of the two genome projects. Since most of such reference data is derived from alignment of short reads (for these projects reads of mostly ≤150 nt) to a reference sequence, there are two likely issues that could cause discrepancy in the estimated frequencies for these hypervariable fragments:

1. Homologous fragments may map to the same position, falsely suggesting a heterozygous genotype.
2. Fragments with many SNPs in a short range may exceed the number of allowed mismatches for mapping reads to the reference during analysis, meaning that only the reads that overlap part of the SNPs and haplotypes that are most similar to the reference sequence will be mapped to the correct location.

In combination with relatively low coverage, these two issues can result in erroneous variant calling for separate SNP positions within one (heterozygous) sample. An extensive wet-lab confirmation of new possibly hypervariable loci is therefore essential. Testing of samples from globally dispersed populations will not only give information about discriminating power in different populations, but also increase the chance to find different heterozygous allele combinations that can help to identify possible co-amplified homologous regions. Testing of samples from large families will confirm correct inheritance of the haplotypes and assist the internal validation of genotyping results.

Although many of the initial candidate loci were rejected, a final set of 16 MHs remained with expected inheritance of the haplotypes in the tested families and a high degree of variation in the population samples. With a varying number of haplotypes for each MH (2-19) and corresponding haplotype frequencies, the discriminating power is not as strong as STRs but the set of 16 loci still reaches strong random match probabilities (RMP) of: $1.0 \times 10^{-9}$ in the Asian population, $4.4 \times 10^{-11}$ in the Dutch population and $9.2 \times 10^{-13}$ in the African population. For identification purposes, our set of loci proved to be more informative than other alternative non-STR loci as can be observed from Table 3. Since the populations tested for the different loci are not exactly the same, a direct comparison of the RMP should be interpreted carefully. Notwithstanding, the discriminating power of the 16 short hypervariable MHs roughly resembles that of nine STRs [26], 25 tri-allelic SNPs [25], 21 tetra-allelic SNPs [19] or 23 of the earlier described MHs [9].

An important advantage of the use of MHs for mixture analysis is the number of hap-lotypes that is observed for several loci. The statistical power (likelihood ratio) of matching a person with a two-person mixture is substantially increased when more

than two alleles are present for a specific locus. The power to detect a third allele for a two-person mixture in at least one of the 16 loci ranges from 0.995 in the Asian population to 0.9989 in the Dutch population and even 0.99992 in African population. For detecting additional contributors in mixtures, the assay outperforms the published sets of tetra-allelic and tri-allelic SNPs (Table 4). For the 130 MHs of Kidd et al [27], the average PMD is estimated based on the top 28 loci for different numbers of loci divided in ranges of effective number of alleles (3-4, 4-5 and >5). These 28 loci together reach a PDM of 0.9999999875 from which 16 loci contain all SNPs within a 150 nt span. However, only three these 28 loci contain all SNPs within a 100 nt span as is the case for the loci described in this paper. The two sets together could complete an even more optimal set of loci for mixture detection.

## Observed variation

Reticulations in a neighbour joining network can be caused by either recombination or by recurrent mutations. The only fragment located in a region with exceptionally high recombination rate is mh11PK-62906, but considering the small fragment length, recombination would only be expected to occur within the fragment once every $5.5 \times 10^4$ meioses. This might suggest that the web-like networks of mh11PK-62906 and mh14PK-72639 (and in lower extent mh17PK-86511) are more likely to be explained by mutation hotspots concentrated on a few specific positions rather than by recombination. Indeed, in none of the tested allele transfers of the CEPH families (144 allele transfer events for each locus in total), recombination has occurred in such a way that the allele inheritance of any of the loci was impacted. When using these loci for paternity cases, it should be considered that mh11PK-62906 and mh14PK-72639 are more likely to display mutations than an average fragment.

For the network of mh06PK-25713, a fairly even distribution of the haplotype frequencies was observed for all populations but for most of the loci, several haplotypes vary substantially in frequency between the tested populations. This suggests that the MHs provide ancestry information although the design and selection of the loci was not intended for this purpose. STRUCTURE analysis indeed showed that the three analysed populations are differentiated almost completely based on the data of these 16 MHs. Data from a larger set of samples with a more global representation would be needed to test the full potential of these MHs as ancestry informative markers. From the 16 MHs that remained after all selection criteria, several of the loci failed Hardy-Weinberg equilibrium test since the frequency of some homozygous genotypes (usually with low frequency) is higher than would be expected.

None of the fragments is located in gene regions, so strong natural selection is not expected for these fragments. An explanation for this could be that some samples in the tested populations are somewhat genetically distinct from the rest of the population,

which is not unlikely since we grouped samples of two Asian populations and of three African populations in order to achieve comparable sample sizes. It also cannot be excluded that some fragments could have occasional SNPs under the primer binding sites although we did not observe any discrepancy of inheritance in the nine CEPH families.

### Sequence data analysis

It should be noted that not every software for sequence data analysis is capable to analyse single-fragment haplotype data. When using an analysis software that maps the complete sequences to a reference, results are often summarised by SNP instead of haplotypes. In this study we used FDSTools [5] since variant frequencies in the data are always reported for the complete sequence between two flanks instead of a summary for each position.

## Conclusions

A new set of short hypervariable microhaplotypes were selected as potential loci for application in forensic DNA analysis. For 16 MFs, confirmation of the variation and inheritance was performed by analysing 276 samples of three globally dispersed populations and 97 samples of nine large families. MHs provide an alternative type of loci for cases where STR stutter or degradation of DNA limits or complicates the analysis. Since the discriminating power of the selected hypervariable MHs is larger than other published non-STR loci, they provide a practical and financially advantageous method with a relatively small number of loci. For the purpose of increased discriminating power and ancestry informative information, a combination of these loci with (part of) the loci of Kidd et al [21] could provide an even more powerful tool.

The selection of short hypervariable MHs from genomic reference data is complicated since the generally short read length of reference data is not ideal to resolve the exact variation in short range hypervariable fragments. Since the read length of most MPS platforms is increasing, future reference data will most likely be better suited for selection and analysis of additional MHs.

National forensic DNA databases currently consist of STR data. Although it is not expected that all database samples will be typed for new loci in the near future, loci such as MHs could provide a powerful tool in cases where reference samples are available for comparison.

## Competing interests

The authors declare no conflict of interest.

## Authors' contributions

KJG wrote the manuscript. KJG, JFJL,RHL and PK performed the selection of loci. KJG and RHL performed the practical work and performed analyses. JD uploaded the data to the LOVD. PK and JD supervised the project and provided funding and equipment.

## Acknowledgements

This study was supported by a grant from the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands. The authors wish to thank Titia Sijen and Jerry Hoogenboom (Netherlands Forensic Institute) for carefully reviewing the manuscript.

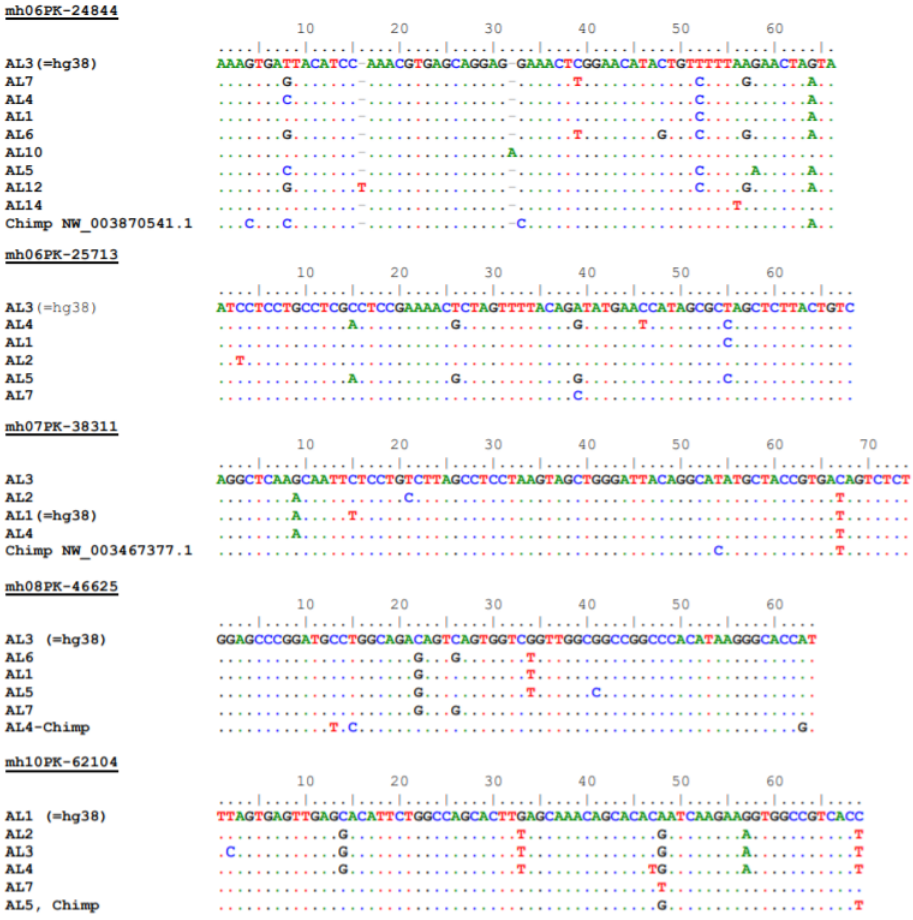
## References

1. K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler. STR allele sequence variation: current knowledge and future issues. Forensic Sci. Int. Gent. 18(2015) 118-130.
2. A.A. Westen, L.J. Grol, J. Harteveld, A.S. Matai, de Knijff .P., T. Sijen. Assessment of the stochastic threshold, back- and forward stutter filters and low template techniques for NGM. Forensic Sci. Int. Genet. 6 (2012) 708-715.
3. Reza Alaeddini, Simon J. Walsh, Ali Abbas. Forensic implications of genetic analyses from degraded DNA—A review. (Forensic Science International: Genetics 4 (2010) 148–157).
4. Jin L, Underhill PA, Doctor V, Davis RW, Shen P, Cavalli-Sforza LL, Oefner PJ. Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. Proc Natl Acad Sci U S A. 1999 Mar 30;96(7):3796-800.
5. 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. 2016 Nov 27;27:27-40.
6. van der Gaag KJ, de Leeuw RH, Hoogenboom J, Patel J, Storts DR, Laros JF, de Knijff P. Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq™ system. Forensic Sci Int Genet. 2016 Sep;24:86-96.
7. Elena S, Alessandro A, Ignazio C, Sharon W, Luigi R, Andrea B. Revealing the challenges of low template DNA analysis with the prototype Ion AmpliSeq™ Identity panel v2.3 on the PGM™ Sequencer. Forensic Sci Int Genet. 2016 May;22:25-36.
8. Grandell I, Samara R, Tillmar AO. A SNP panel for identity and kinship testing using massive parallel sequencing. Int J Legal Med. 2016 Jul;130(4):905-14.
9. Kidd KK, Pakstis AJ, Speed WC, Lagacé R, Chang J, Wootton S, Haigh E, Kidd JR. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. Forensic Sci Int Genet. 2014 Sep;12:215-24.
10. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

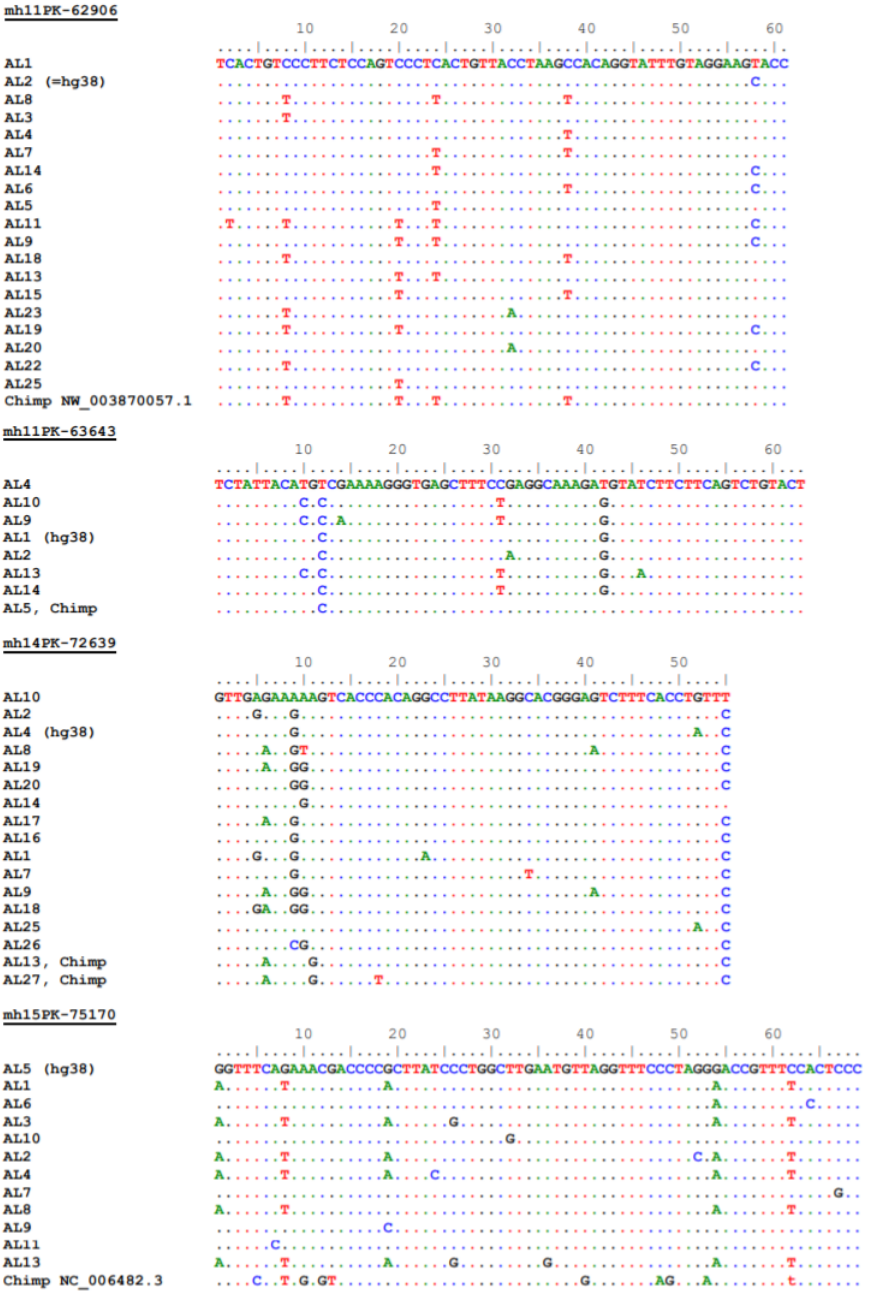
11. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 46:818-25.
12. Fokkema IF, den Dunnen JT, Taschner PE. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat. 2005 Aug;26(2):63-8.
13. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat. 2011 May;32(5):557-63. doi: 10.1002/humu.21438. Epub 2011 Feb 22.
14. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, et al. A human genome diversity cell line panel. Science. 2002 Apr 12;296(5566):261-2.
15. A.A. Westen, T. Kraaijenbrink, E.A. Robles de Medina, J. Harteveld, P. Willemse, S.B. Zuniga, K.J. van der Gaag, N.E. Weiler, J. Warnaar, M. Kayser, T. Sijen, P. de Knijff. Comparing six commercial autosomal STR kits in a large Dutch population sample. Forensic Sci. Int. Genet. 10 (2014) 55–63.
16. The International HapMap Consortium, Integrating ethics and science in the International HapMap Project. Nat Rev Genet. 2004 Jun;5(6):467-75.
17. Promega Corporation, Powerstats v1.2. Unpublished work.
18. Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics. 2012 Oct 1;28(19):2537-9.
19. 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;19:100-6.
20. Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999 Jan;16(1):37-48.
21. International HapMap Consortium et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct 18;449(7164):851-61.
22. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. Science. 2002 Dec 20;298(5602):2381-5.
23. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour. 2015 Sep;15(5):1179-91.
24. Kidd KK. Proposed nomenclature for microhaplotypes. Hum Genomics. 2016 Jun17;10(1):16.
25. 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.
26. ThermoFisher. cms_073986 - The AmpFlSTR® NGM™ PCR Amplification Kit: The Perfect Union of Data Quality and Data Sharing. Forensic News ThermoFisher.
27. Kidd KK, Speed WC, Pakstis AJ, Podini DS, Lagacé R, Chang J, Wootton S, Haigh E, Soundararajan U. Evaluating 130 microhaplotypes across a global set of 83 populations. Forensic Sci Int Genet. 2017 Jul;29:29-37.

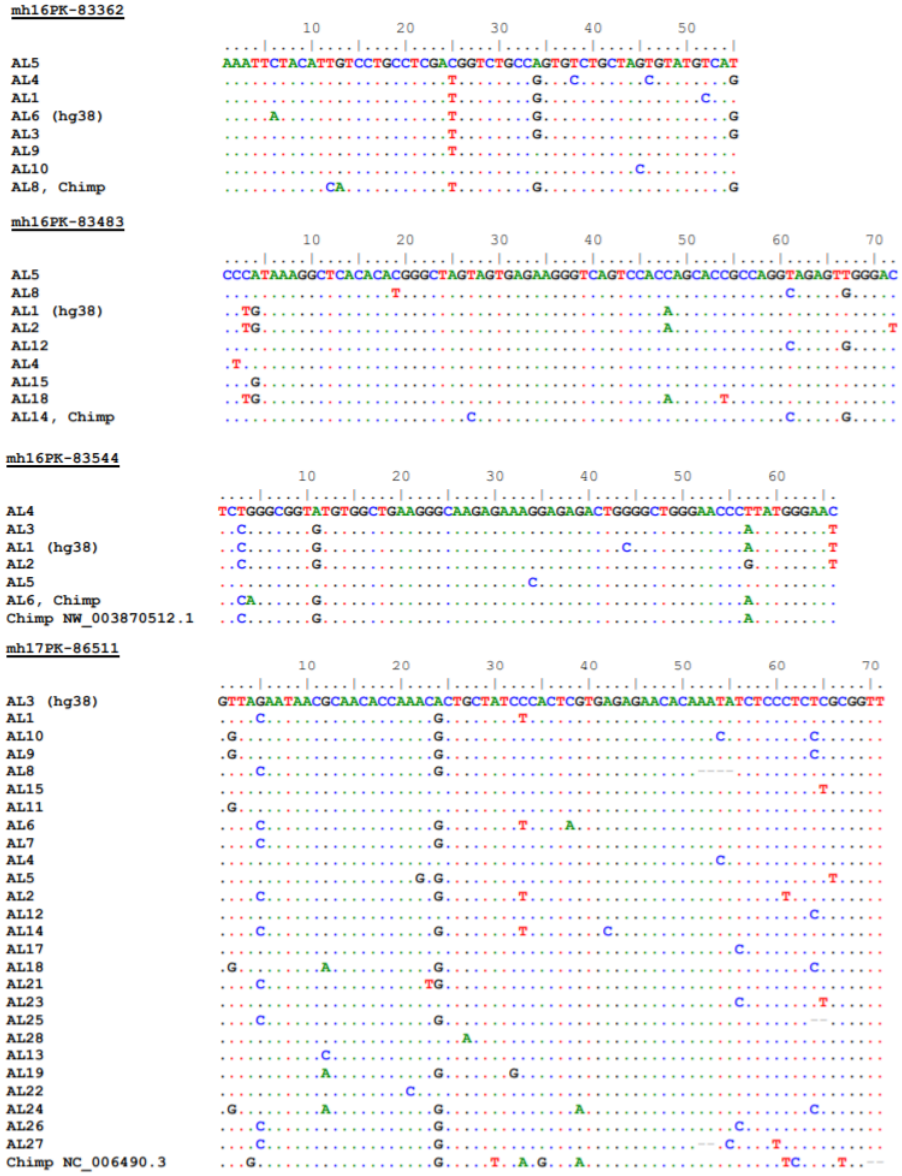
# Supplementary materials

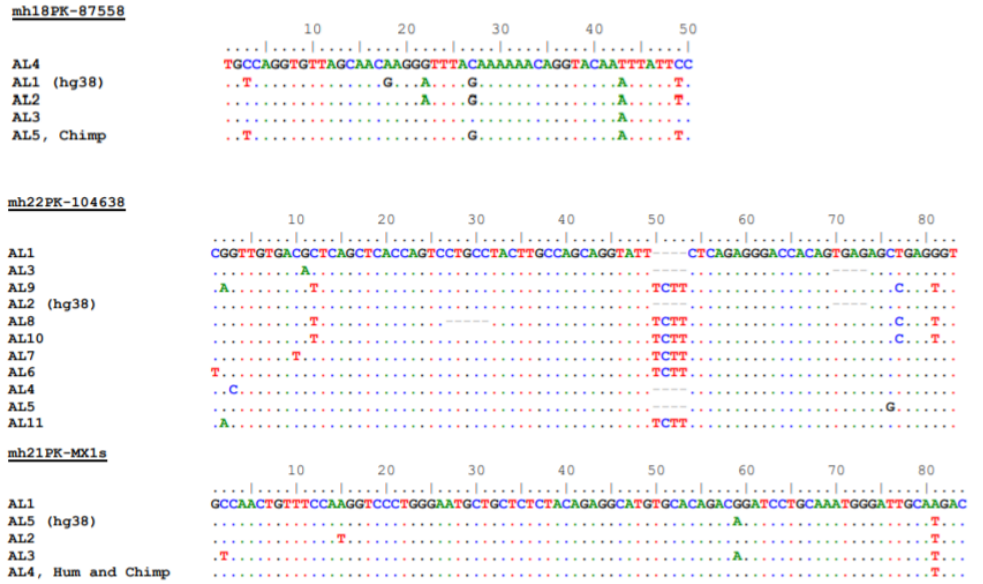
**Supplementary Figure 1** – Alignments of the observed microhaplotype variation



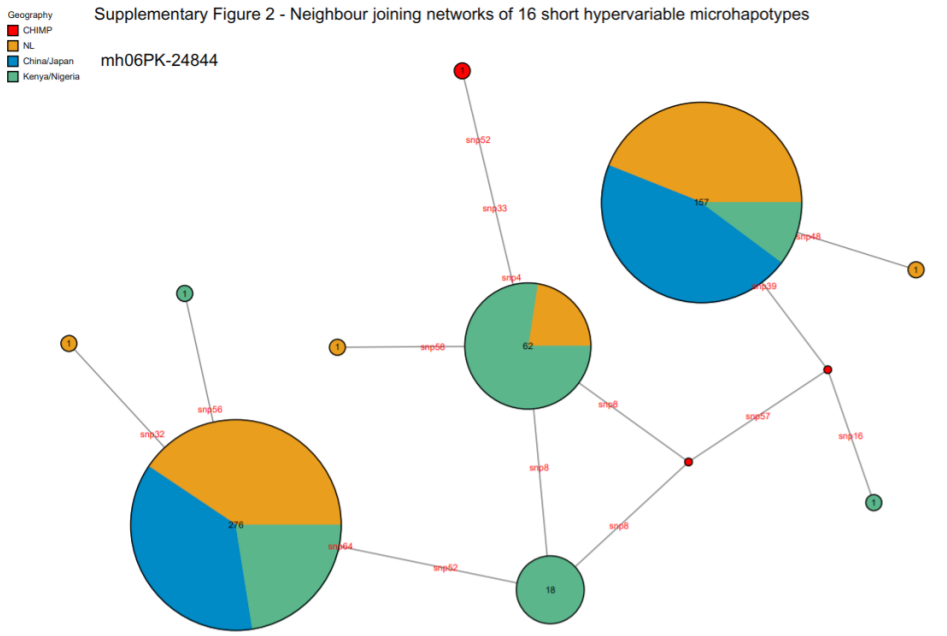
Chapter 6

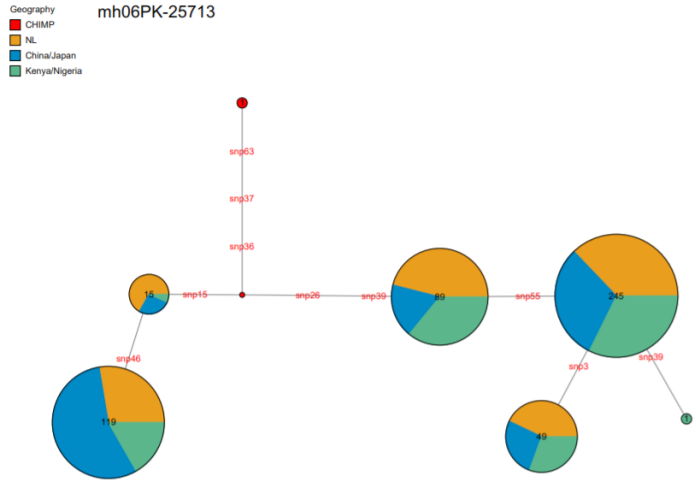

mh11PK-62906
AL1
AL2 (=hg38)
AL8
AL3
AL4
AL7
AL14
AL6
AL5
AL11
AL9
AL18
AL13
AL15
AL23
AL19
AL20
AL22
AL25
Chimp NW_003870057.1
mh11PK-63643
AL4
AL10
AL9
AL1 (hg38)
AL2
AL13
AL14
AL5, Chimp
mh14PK-72639
AL10
AL2
AL4 (hg38)
AL8
AL19
AL20
AL14
AL17
AL16
AL1
AL7
AL9
AL18
AL25
AL26
AL13, Chimp
AL27, Chimp
mh15PK-75170
AL5 (hg38)
AL1
AL6
AL3
AL10
AL2
AL4
AL7
AL8
AL9
AL11
AL13
Chimp NC_006482.3

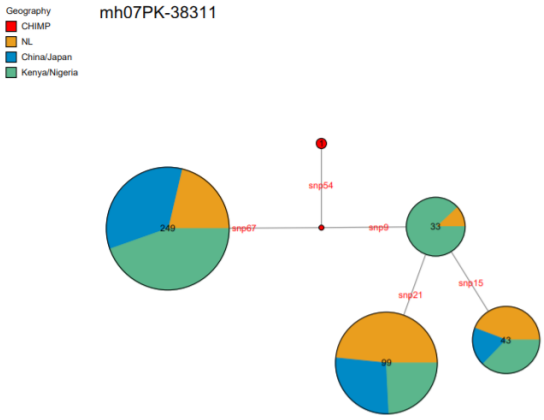

mh16PK-83362
AL5
AL4
AL1
AL6 (hg38)
AL3
AL9
AL10
AL8, Chimp
mh16PK-83483
AL5
AL8
AL1 (hg38)
AL2
AL12
AL4
AL15
AL18
AL14, Chimp
mh16PK-83544
AL4
AL3
AL1 (hg38)
AL2
AL5
AL6, Chimp
Chimp NW_003870512.1
mh17PK-86511
AL3 (hg38)
AL1
AL10
AL9
AL8
AL15
AL11
AL6
AL7
AL4
AL5
AL2
AL12
AL14
AL17
AL18
AL21
AL23
AL25
AL28
AL13
AL19
AL22
AL24
AL26
AL27
Chimp NC_006490.3



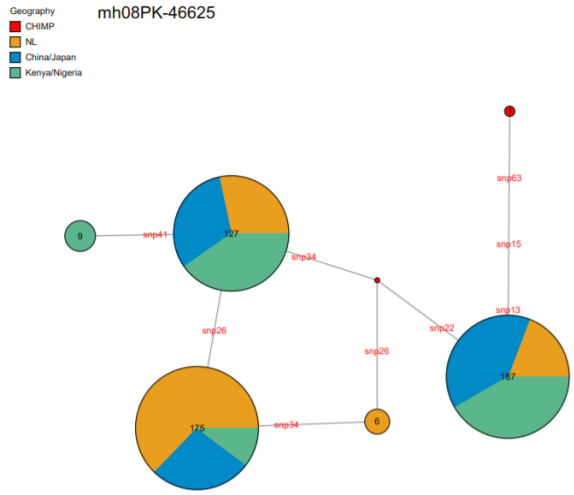

*For each MH, the alignment of all observed variants is displayed sorted (top to bottom) by the overall frequency (of all tested populations combined). When chimpanzee sequence data is available, the sequence is displayed at the bottom of the alignment. The haplotype with the sequence of GRCh38 is marked as (hg38).*

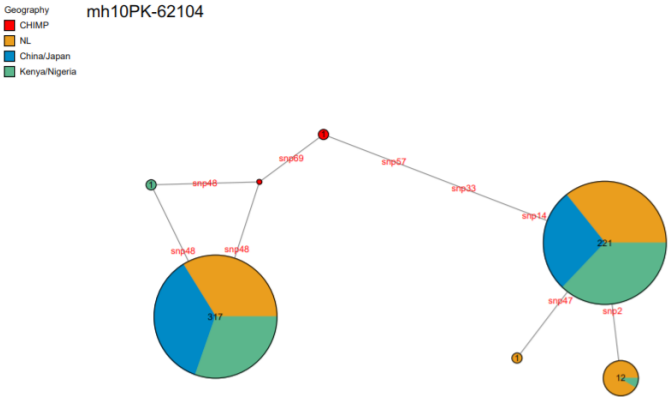


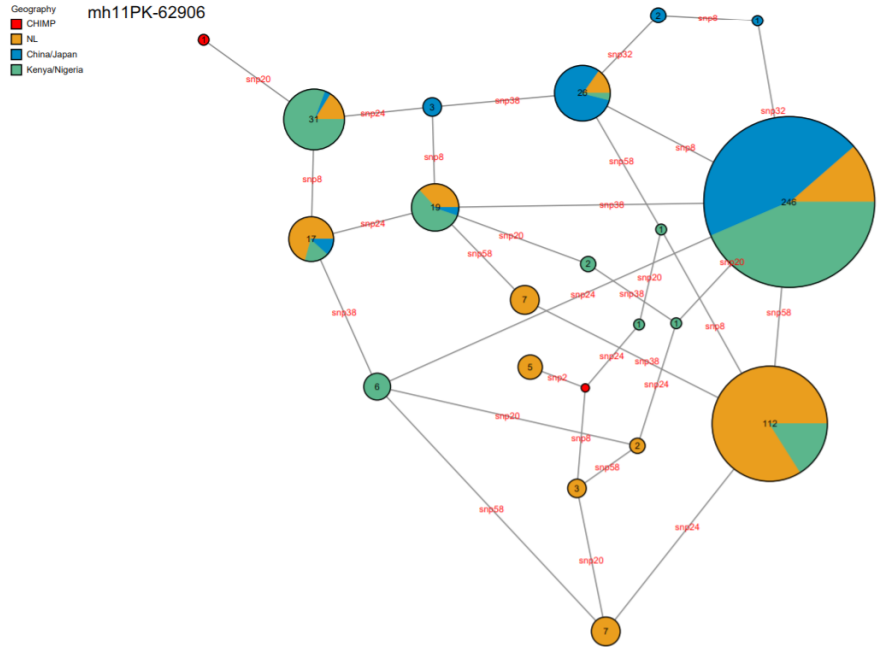

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh06PK-25713
snp63
snp37
snp36
snp15
snp26
snp39
snp55
snp46
snp3
snp39
16
89
245
119
49
1

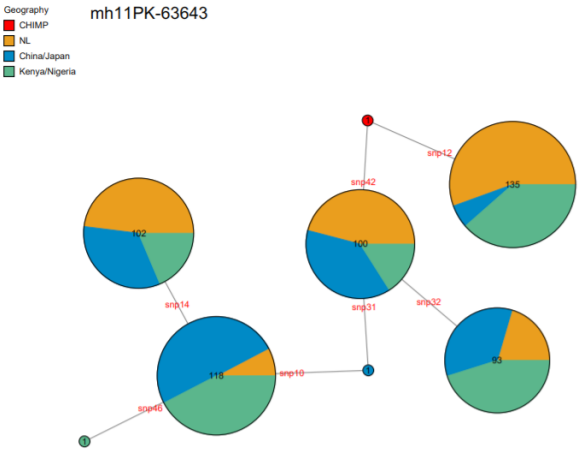

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh07PK-38311
snp54
snp67
snp9
snp21
snp15
249
33
99
43

Chapter 6

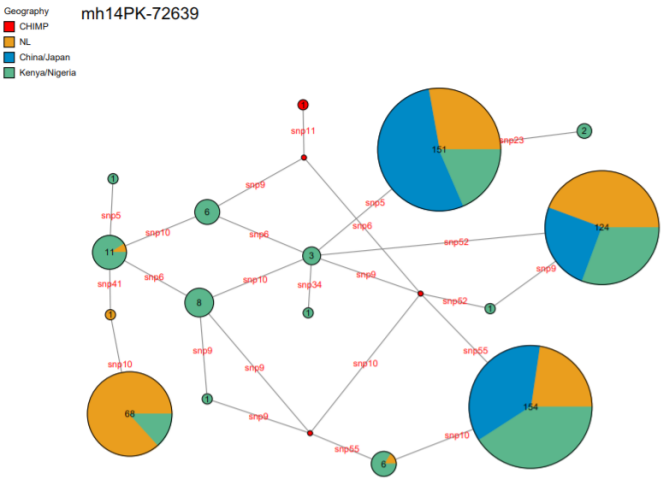

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh08PK-46625
1
snp63
9
snp41
127
snp34
snp15
snp13
snp26
snp22
snp26
187
175
snp34
6

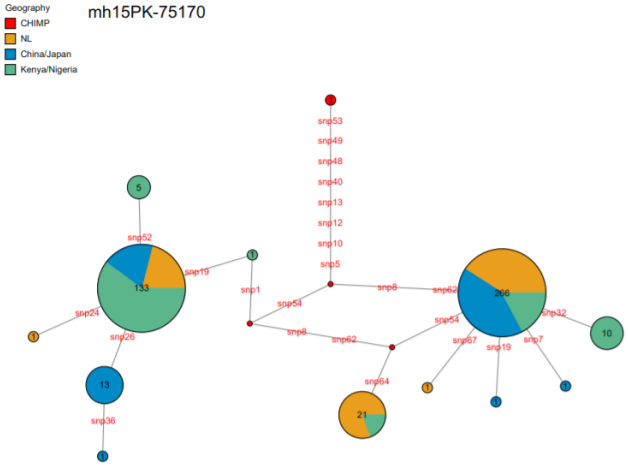

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh10PK-62104
1
snp69
snp57
1
snp48
snp33
snp14
221
snp48
snp48
317
snp47
snp2
1
12

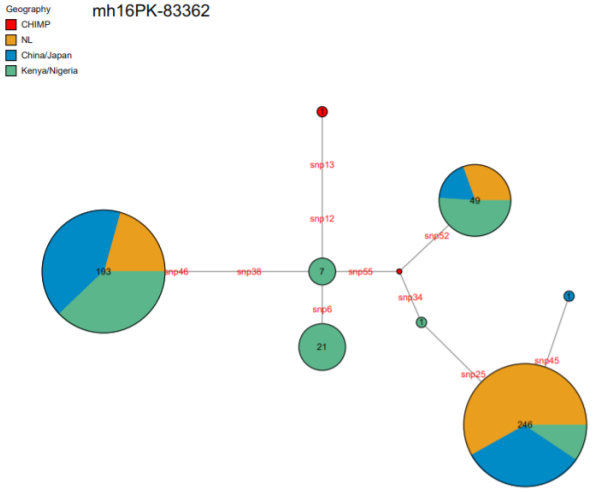

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh11PK-62906
snp20
snp24
snp38
snp32
snp8
snp58
snp2
snp12
snp42
snp14
snp31
snp10
snp46
31
3
26
246
19
17
7
6
5
2
112
135
102
100
118
93

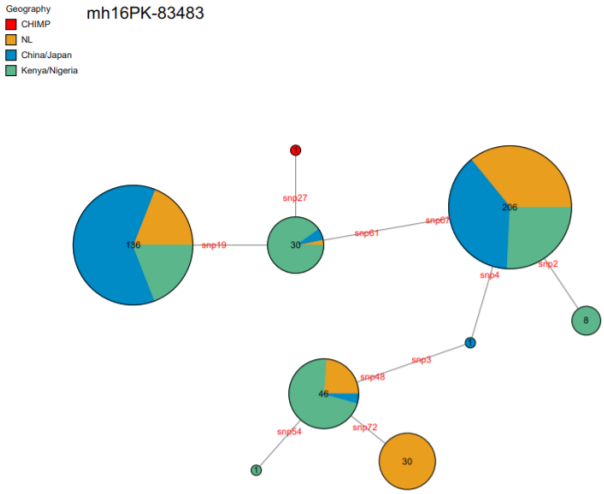

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh11PK-63643

Chapter 6

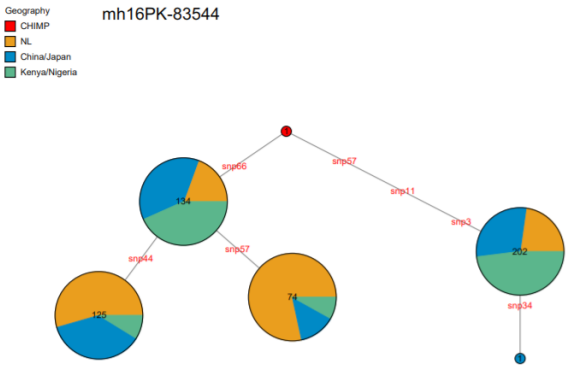

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh14PK-72639
151
124
68
154
snp11
snp23
snp9
snp5
snp10
snp6
snp52
snp41
snp34
snp62
snp55

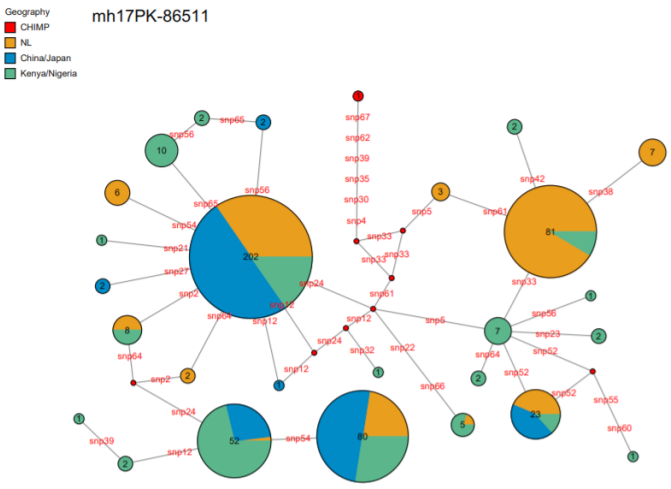

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh15PK-75170
snp53
snp49
snp48
snp40
snp13
snp12
snp10
snp5
snp52
snp19
133
snp24
snp26
13
snp36
snp1
snp54
snp8
snp62
266
snp32
10
snp67
snp7
snp64
21

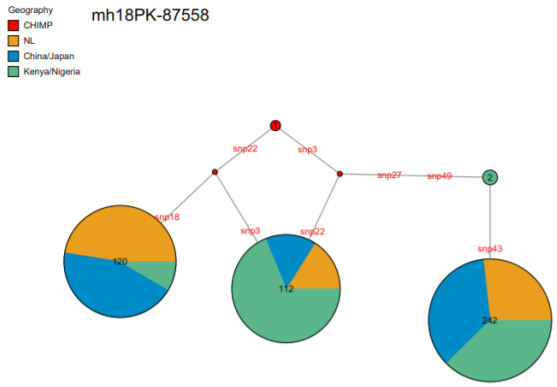

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh16PK-83362
snp13
snp12
snp52
snp46
snp38
snp55
snp34
snp6
snp23
snp45
193
7
49
21
248

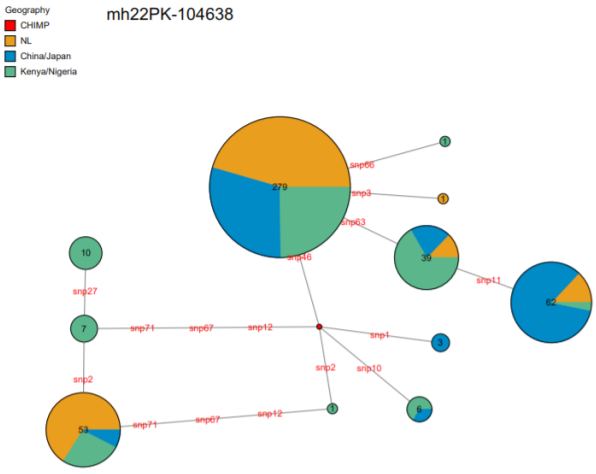

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh16PK-83483
snp27
snp19
snp61
snp67
snp4
snp2
snp3
snp48
snp54
snp72
136
30
206
8
46
30

Chapter 6

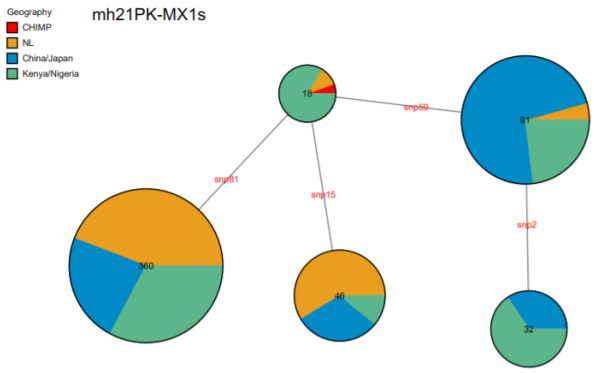

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh16PK-83544
snp66
snp57
snp11
snp3
snp44
snp67
snp34
134
125
74
202


Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh17PK-86511
202
81
80
52
74
23


Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh18PK-87558
snp22
snp3
snp27
snp49
snp18
snp3
snp22
snp43
120
112
242

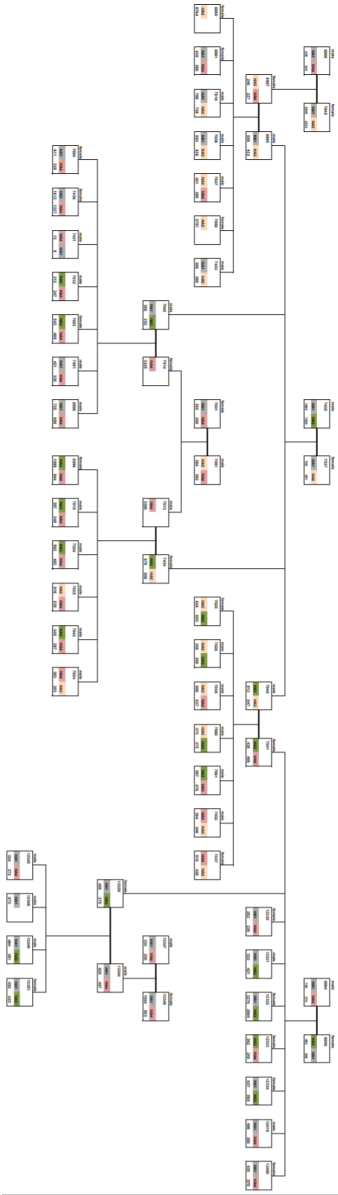

Geography
CHIMP
NL
China/Japan
Kenya/Nigeria
mh22PK-104638
snp66
snp3
snp63
snp46
snp11
snp27
snp71
snp67
snp12
snp1
snp2
snp10
snp2
snp71
snp67
snp12
279
39
62
10
7
53

Chapter 6




*Neighbour joining networks are displayed for each of the final 16 MHs. The total number of haplotypes in each circle is displayed and the size of each coloured circle displays which number of alleles of the specific haplotype was present in each population. Every number on a branch displays a SNP that separates one haplotype from another.*

**Supplementary Figure 3** – Joined family tree of six CEPH families displaying the observed haplotypes and read counts for mh16PK-83544 for each individual



*The joined family tree is displayed of CEPH family 1328, 13281, 13291, 13292, 13293 and 13294 (50 individuals in total) with the observed haplotypes for mh16PK-83544. In the tree, different haplotypes are marked in different colours with the read numbers of each haplotype displayed below. Parents are connected in the tree with double lines and to their respective offspring with single lines.*

**Supplementary Table 1** - Primer details, genome locations and recombination rate for the 16 selected microhaplotypes

Supplementary Table 1a - Primer details of the 16 selected short hypervariable microhaplotypes

| Fragment name | Forward primer | Reverse primer | Amplicon size (including primers) |
|---|---|---|---|
| mh06PK-24844 | GTGAACCCAGCAAAAGGAAG | GACTTTAAAGCTGCAAACTCTAGTGA | 110 |
| mh06PK-25713 | CAAACTCCTGGTCTCACAyG | GGGAGGACGTGTTGAAGArA | 108 |
| mh07PK-38311 | CCAGCTGGTCTTGAACTCCT | CCTGAGTCAAATTAAATTACAyAAAT | 120 |
| mh08PK-46625 | GTCCCGGCTGGTGGAG | CTGCCTAGACAsGGTGAGC | 99 |
| mh10PK-62104 | GGCTTTCAGGGTGGTCATT | CCAAATAAGGAACTTTGTGGAAA | 118 |
| mh11PK-62906 | GCTAATTCCTCCTmGCTCCT | CAGAAGGTGTTGGCrTCAT | 100 |
| mh11PK-63643 | CTTGGATTCTGCCTCCACAT | GTTGGAACTGGTCCTTGTGAA | 104 |
| mh14PK-72639 | CGAAGCGAGCACGTTG | CGGTTCGGTCACCGTAAGTA | 87 |
| mh15PK-75170 | TCCCTGGCTTkAAAGTGC | GTTGAGGGGAGGAGGCAG | 114 |
| mh16PK-83362 | CTCCTTTGACTGTCCCGACT | TGAAGAGAGAGCAGAAGAACACA | 98 |
| mh16PK-83483 | CCAGAGGGAGAGGAGATGC | CTGTTTTTATTTAACCCCTTCTGG | 115 |
| mh16PK-83544 | AGGGGTGTGTTCTGGAGATG | TGGCCTGGACACTCAGC | 103 |
| mh17PK-86511 | TCACCAACAGCAAACAGCA | TGAACGTGTAGAACGCTGGA | 110 |
| mh18PK-87558 | GTAGCAGCCTAGCCAAGAGC | CCCAGAAACTACCCATAGGTTACTA | 114 |
| mh21PK-MX1s | GGGAGCAAGCACCTTACAGT | GGACTTAACCTTGAAATGGAAA | 126 |
| mh22PK-104638 | GGAGTGCTCAGTGACATTGG | AAGTGGGTACTGGTGCAGGT | 116 |

**Supplementary Table 1b** - Genome locations, distance to the closest microhaplotypes of Kidd et al. 2017 and recombination rate (from HapMap) for the 16 selected microhaplotypes

| Fragment name | Chromosome location (GRCh38) | Upstream MH Kidd et al. 2017 | Distance to MH (nt) | Downstream MH Kidd et al. 2017 | Distance to MH (nt) | recombination rate (cM/Mb) | average nr of meioses / recombination |
|---|---|---|---|---|---|---|---|
| mh06PK-24844 | chr6:13861366-13861475 | mh06KK-101 | 2,1E+06 | mh06KK-026 | 3,2E+07 | 0,50 | 1,8E+06 |
| mh06PK-25713 | chr6:31196928-31197035 | mh06KK-101 | 1,9E+07 | mh06KK-026 | 1,5E+07 | 0,13 | 7,1E+06 |
| mh07PK-38311 | chr7:52677423-52677542 | mh07KK-031 | 2,3E+07 | mh07KK-082 | 2,9E+07 | 3,30 | 2,5E+05 |
| mh08PK-46625 | chr8:1194316-1194414 | | | mh08KK-032 | 2,2E+08 | 0,93 | 1,1E+06 |
| mh10PK-62104 | chr10:127392544-127392661 | mh10KK-084 | 2,0E+07 | mh10KK-087 | 6,1E+06 | 0,42 | 2,0E+06 |
| mh11PK-62906 | chr11:247959-248058 | | | mh11KK-090 | 4,8E+06 | 18,25 | 5,5E+04 |
| mh11PK-63643 | chr11:34415786-34415889 | mh11KK-040 | 2,5E+07 | mh11KK-039 | 1,7E+07 | 0,02 | 6,0E+07 |
| mh14PK-72639 | chr14:32203258-32203344 | mh14KK-101 | 1,2E+07 | mh14KK-068 | 1,0E+08 | 0,29 | 3,9E+06 |
| mh15PK-75170 | chr15: 24802296-24802409 | | | mh15KK-104 | 5,4E+07 | 1,98 | 4,4E+05 |
| mh16PK-83362 | chr16:77999218-77999315 | mh16KK-302 | 2,8E+07 | mh16KK-096 | 4,0E+07 | 0,49 | 2,1E+06 |
| mh16PK-83483 | chr16:84516809-84516923 | mh16KK-302 | 3,4E+07 | mh16KK-096 | 3,4E+07 | 0,51 | 1,7E+06 |
| mh16PK-83544 | chr16:85934053-85934155 | mh16KK-302 | 3,6E+07 | mh16KK-096 | 3,3E+07 | 1,21 | 8,0E+05 |
| mh17PK-86511 | chr17:58631674-58631783 | mh17KK-055 | 7,0E+06 | mh17KK-052 | 5,2E+06 | 0,00 | 1,5E+09 |
| mh18PK-87558 | chr18:1960521-1960634 | | | mh18KK-293 | 4,9E+06 | 0,53 | 1,7E+06 |
| mh21PK-MX1s | chr21:41464724-41464849 | mh21KK-324 | 5,2E+06 | mh21KK-316 | 5,3E+07 | 0,90 | 8,8E+05 |
| mh22PK-104638 | chr22:44857863-44857978 | mh22KK-060 | 2,2E+07 | mh22KK-064 | 1,7E+06 | 1,12 | 7,7E+05 |

*Primer details, genome locations and recombination rate of the MHs distance to the closest microhaplotypes of Kidd et al 2017.*

**Supplementary Table 2** - Allele frequencies of MH haplotypes for the analysed populations

| Locus | | NL | Asia | Africa |
|---|---|---|---|---|
| **mh06PK-24844** | **Haplotype \ n** | 99 | 87 | 73 |
| | 1 | 0,000 | 0,000 | 0,123 |
| | 3 | 0,566 | 0,586 | 0,425 |
| | 4 | 0,071 | 0,000 | 0,329 |
| | 5 | 0,005 | 0,000 | 0,000 |
| | 6 | 0,005 | 0,000 | 0,000 |
| | 7 | 0,348 | 0,414 | 0,110 |
| | 10 | 0,005 | 0,000 | 0,000 |
| | 12 | 0,000 | 0,000 | 0,007 |
| | 14 | 0,000 | 0,000 | 0,007 |
| **mh06PK-25713** | **Haplotype \ n** | 98 | 87 | 74 |
| | 1 | 0,209 | 0,092 | 0,216 |
| | 2 | 0,107 | 0,075 | 0,101 |
| | 3 | 0,464 | 0,431 | 0,534 |
| | 4 | 0,168 | 0,379 | 0,135 |
| | 5 | 0,051 | 0,023 | 0,007 |
| | 7 | 0,000 | 0,000 | 0,007 |
| **mh07PK-38311** | **Haplotype \ n** | 62 | 60 | 90 |
| | 1 | 0,153 | 0,067 | 0,089 |
| | 2 | 0,387 | 0,225 | 0,133 |
| | 3 | 0,427 | 0,708 | 0,617 |
| | 4 | 0,032 | 0,000 | 0,161 |
| **mh08PK-46625** | **Haplotype \ n** | 94 | 80 | 78 |
| | 1 | 0,191 | 0,250 | 0,327 |
| | 3 | 0,191 | 0,456 | 0,500 |
| | 5 | 0,000 | 0,000 | 0,058 |
| | 6 | 0,585 | 0,294 | 0,115 |
| | 7 | 0,032 | 0,000 | 0,000 |
| **mh10PK-62104** | **Haplotype \ n** | 99 | 87 | 90 |
| | 1 | 0,540 | 0,655 | 0,533 |
| | 2 | 0,399 | 0,345 | 0,456 |
| | 3 | 0,056 | 0,000 | 0,006 |
| | 4 | 0,005 | 0,000 | 0,000 |
| | 7 | 0,000 | 0,000 | 0,006 |

| mh11PK-62906 | Haplotype \ n | 87 | 71 | 88 |
|---|---|---|---|---|
| | 1 | 0,161 | 0,782 | 0,608 |
| | 2 | 0,540 | 0,000 | 0,102 |
| | 3 | 0,023 | 0,148 | 0,006 |
| | 4 | 0,040 | 0,007 | 0,063 |
| | 5 | 0,000 | 0,000 | 0,034 |
| | 6 | 0,040 | 0,000 | 0,000 |
| | 7 | 0,069 | 0,014 | 0,017 |
| | 8 | 0,029 | 0,007 | 0,142 |
| | 9 | 0,017 | 0,000 | 0,000 |
| | 11 | 0,029 | 0,000 | 0,000 |
| | 13 | 0,011 | 0,000 | 0,000 |
| | 14 | 0,040 | 0,000 | 0,000 |
| | 15 | 0,000 | 0,000 | 0,011 |
| | 18 | 0,000 | 0,021 | 0,000 |
| | 19 | 0,000 | 0,000 | 0,006 |
| | 20 | 0,000 | 0,007 | 0,000 |
| | 22 | 0,000 | 0,000 | 0,006 |
| | 23 | 0,000 | 0,014 | 0,000 |
| | 25 | 0,000 | 0,000 | 0,006 |
| mh11PK-63643 | Haplotype \ n | 99 | 86 | 90 |
| | 1 | 0,232 | 0,221 | 0,089 |
| | 2 | 0,096 | 0,186 | 0,233 |
| | 4 | 0,379 | 0,047 | 0,289 |
| | 9 | 0,247 | 0,198 | 0,106 |
| | 10 | 0,045 | 0,343 | 0,278 |
| | 13 | 0,000 | 0,000 | 0,006 |
| | 14 | 0,000 | 0,006 | 0,000 |

Chapter 6

| mh14PK-72639 | Haplotype \ n | 97 | 84 | 88 |
|---|---|---|---|---|
| | 1 | 0,000 | 0,000 | 0,011 |
| | 2 | 0,216 | 0,482 | 0,159 |
| | 4 | 0,284 | 0,185 | 0,216 |
| | 7 | 0,000 | 0,000 | 0,006 |
| | 8 | 0,304 | 0,000 | 0,051 |
| | 9 | 0,005 | 0,000 | 0,000 |
| | 10 | 0,180 | 0,333 | 0,358 |
| | 14 | 0,005 | 0,000 | 0,028 |
| | 16 | 0,000 | 0,000 | 0,017 |
| | 17 | 0,000 | 0,000 | 0,034 |
| | 18 | 0,000 | 0,000 | 0,006 |
| | 19 | 0,005 | 0,000 | 0,057 |
| | 20 | 0,000 | 0,000 | 0,045 |
| | 25 | 0,000 | 0,000 | 0,006 |
| | 26 | 0,000 | 0,000 | 0,006 |
| mh15PK-75170 | Haplotype \ n | 78 | 76 | 73 |
| | 1 | 0,179 | 0,164 | 0,548 |
| | 2 | 0,000 | 0,000 | 0,034 |
| | 3 | 0,000 | 0,086 | 0,000 |
| | 4 | 0,006 | 0,000 | 0,000 |
| | 5 | 0,699 | 0,730 | 0,315 |
| | 6 | 0,109 | 0,000 | 0,027 |
| | 7 | 0,006 | 0,000 | 0,000 |
| | 8 | 0,000 | 0,000 | 0,007 |
| | 9 | 0,000 | 0,007 | 0,000 |
| | 10 | 0,000 | 0,000 | 0,068 |
| | 11 | 0,000 | 0,007 | 0,000 |
| | 13 | 0,000 | 0,007 | 0,000 |
| mh16PK-83362 | Haplotype \ n | 99 | 85 | 75 |
| | 1 | 0,076 | 0,053 | 0,167 |
| | 3 | 0,000 | 0,000 | 0,047 |
| | 4 | 0,202 | 0,471 | 0,487 |
| | 5 | 0,722 | 0,471 | 0,153 |
| | 6 | 0,000 | 0,000 | 0,140 |
| | 9 | 0,000 | 0,000 | 0,007 |
| | 10 | 0,000 | 0,006 | 0,000 |

| mh16PK-83483 | Haplotype \ n | 71 | 84 | 74 |
|---|---|---|---|---|
| | 1 | 0,077 | 0,012 | 0,223 |
| | 2 | 0,211 | 0,000 | 0,000 |
| | 4 | 0,000 | 0,000 | 0,054 |
| | 5 | 0,521 | 0,470 | 0,358 |
| | 8 | 0,183 | 0,500 | 0,176 |
| | 12 | 0,007 | 0,012 | 0,182 |
| | 15 | 0,000 | 0,006 | 0,000 |
| | 18 | 0,000 | 0,000 | 0,007 |
| **mh16PK-83544** | **Haplotype \ n** | 99 | 83 | 86 |
| | 1 | 0,343 | 0,277 | 0,064 |
| | 2 | 0,293 | 0,060 | 0,035 |
| | 3 | 0,131 | 0,301 | 0,337 |
| | 4 | 0,232 | 0,355 | 0,564 |
| | 5 | 0,000 | 0,006 | 0,000 |

Chapter 6

| mh17PK-86511 | Haplotype \ n | 98 | 85 | 70 |
|---|---|---|---|---|
| | 1 | 0,378 | 0,000 | 0,050 |
| | 2 | 0,015 | 0,000 | 0,000 |
| | 3 | 0,357 | 0,594 | 0,221 |
| | 4 | 0,031 | 0,000 | 0,000 |
| | 5 | 0,005 | 0,000 | 0,029 |
| | 6 | 0,036 | 0,000 | 0,000 |
| | 7 | 0,000 | 0,000 | 0,050 |
| | 8 | 0,051 | 0,059 | 0,021 |
| | 9 | 0,005 | 0,082 | 0,264 |
| | 10 | 0,092 | 0,235 | 0,157 |
| | 11 | 0,020 | 0,000 | 0,029 |
| | 12 | 0,010 | 0,000 | 0,000 |
| | 13 | 0,000 | 0,006 | 0,000 |
| | 14 | 0,000 | 0,000 | 0,014 |
| | 15 | 0,000 | 0,000 | 0,071 |
| | 17 | 0,000 | 0,012 | 0,000 |
| | 18 | 0,000 | 0,000 | 0,014 |
| | 19 | 0,000 | 0,000 | 0,007 |
| | 21 | 0,000 | 0,000 | 0,014 |
| | 22 | 0,000 | 0,000 | 0,007 |
| | 23 | 0,000 | 0,000 | 0,014 |
| | 24 | 0,000 | 0,000 | 0,007 |
| | 25 | 0,000 | 0,000 | 0,014 |
| | 26 | 0,000 | 0,000 | 0,007 |
| | 27 | 0,000 | 0,000 | 0,007 |
| | 28 | 0,000 | 0,012 | 0,000 |
| **mh18PK-87558** | **Haplotype \ n** | 70 | 78 | 90 |
| | 1 | 0,407 | 0,340 | 0,056 |
| | 2 | 0,129 | 0,109 | 0,428 |
| | 3 | 0,000 | 0,000 | 0,011 |
| | 4 | 0,464 | 0,551 | 0,506 |

| mh22PK-104638 | Haplotype \ n | 88 | 76 | 67 |
|---|---|---|---|---|
| | 1 | 0,722 | 0,546 | 0,515 |
| | 2 | 0,028 | 0,053 | 0,194 |
| | 3 | 0,045 | 0,342 | 0,015 |
| | 4 | 0,006 | 0,000 | 0,000 |
| | 5 | 0,000 | 0,000 | 0,007 |
| | 6 | 0,000 | 0,020 | 0,000 |
| | 7 | 0,000 | 0,013 | 0,030 |
| | 8 | 0,000 | 0,000 | 0,075 |
| | 9 | 0,199 | 0,026 | 0,104 |
| | 10 | 0,000 | 0,000 | 0,052 |
| | 11 | 0,000 | 0,000 | 0,007 |
| **mh21PK-MX1s** | **Haplotype \ n** | 96 | 87 | 90 |
| | 1 | 0,828 | 0,477 | 0,656 |
| | 2 | 0,141 | 0,080 | 0,028 |
| | 3 | 0,000 | 0,063 | 0,117 |
| | 4 | 0,010 | 0,000 | 0,083 |
| | 5 | 0,021 | 0,379 | 0,117 |

Chapter 6

# Chapter 7

## General Discussion

Kristiaan J. van der Gaag

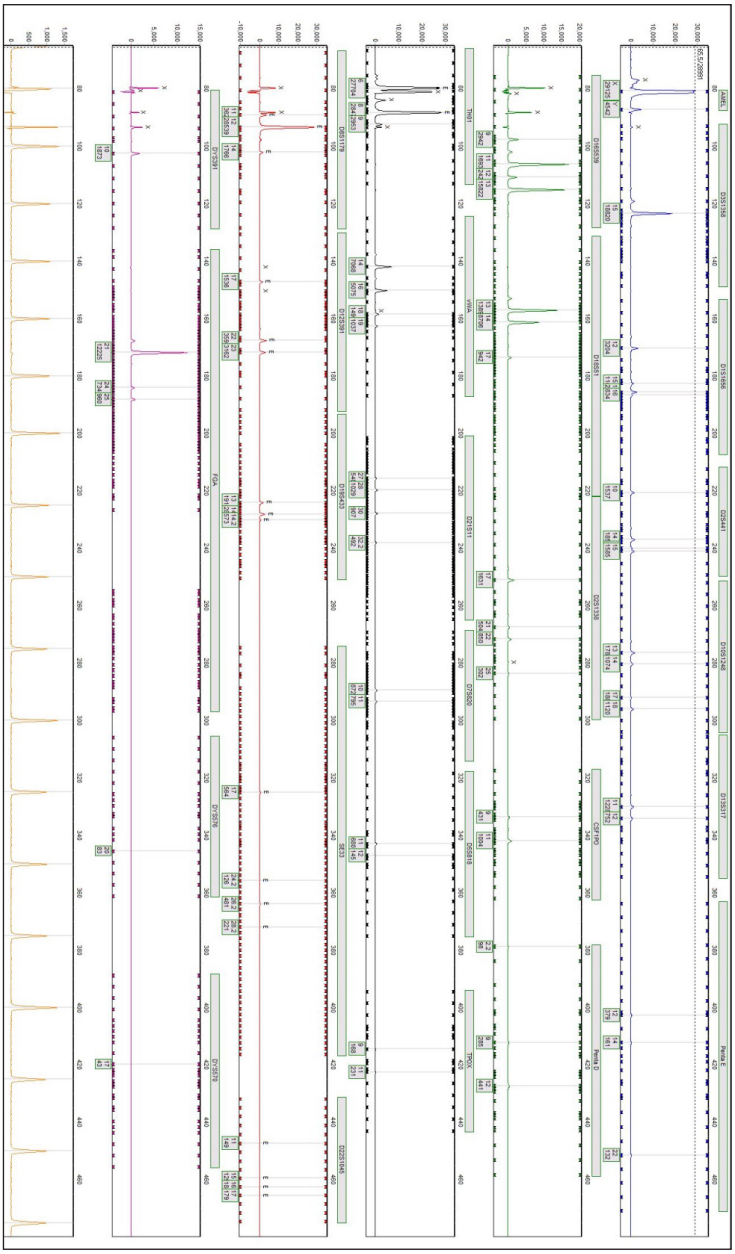
# General Discussion

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

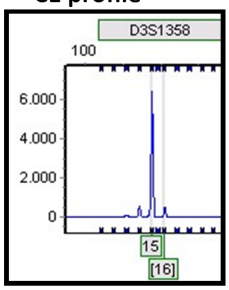
On the sequence level, many of the STRs exhibit more variation than repeat length only (**chapter 4**). By using a reference database which includes sequence variation, the discriminating power of the same loci increases substantially. The additional information also aids in differentiating genuine alleles 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.

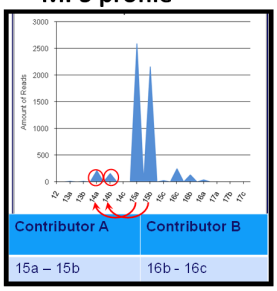
The sequence variation is likely to also provide additional information for the genetic biogeographic ancestry of a person. Already with CE STR data [1], it is possible to make a prediction of the biogeographic ancestry of a person, but the additional, often less variable, sequence variation will likely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




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™ 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

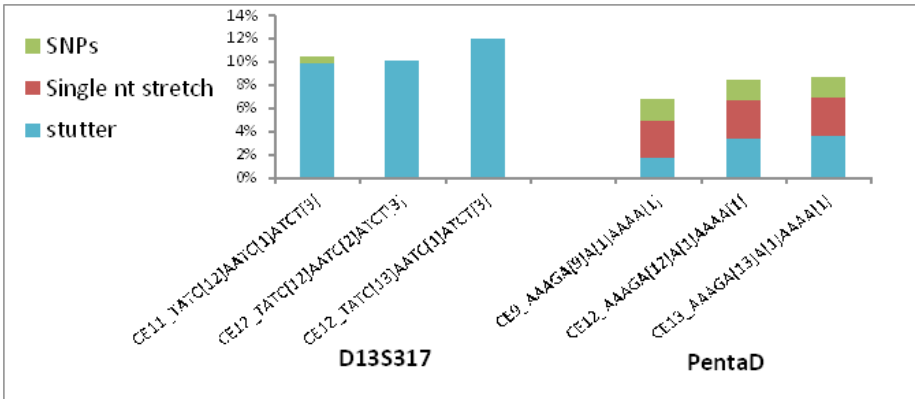
Repeating sequences such as STRs pose a challenge for most alignment algorithms [30]. Since STRs were an important target of interest, we developed the software TSSV using a new approach for analysis of fixed targets containing repeated sequence motifs by only mapping small parts in both the flanking, 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.*

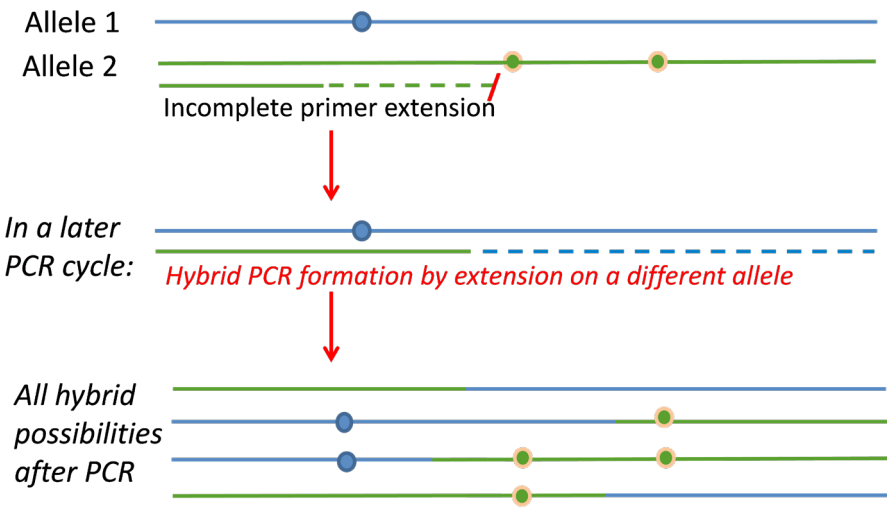
In accordance with CE data, we see that alleles with longer uninterrupted stretches of repeats 'loose' more of their original intensity to stutter events and other PCR noise than smaller alleles (also see *Sup. figure7* of **chapter 4**). Once noise sequences can be attributed to the respective genuine 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.

Chapter 7

## 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




*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

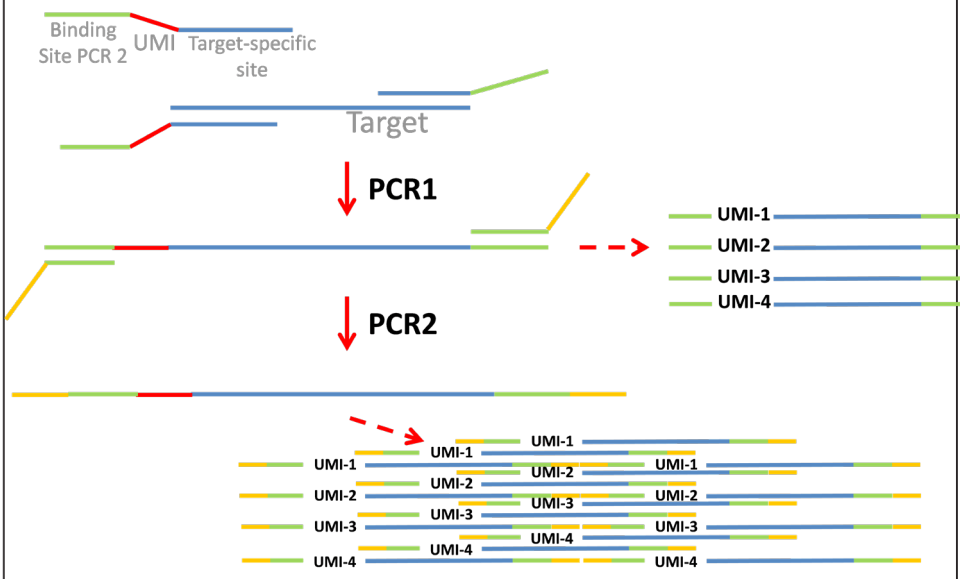
The noise discussed before represents noise created during the PCR. Besides, sequencing errors occur, but this is only a small minority of the noise. Although a lot of the noise can be recognised and corrected during MPS data analysis, ideally noise is prevented to arise at all, which would mean working 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).

Chapter 7

**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

Chapter 7

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

# Epilogue

List of publications
Summary thesis
Nederlandse samenvatting
Dankwoord
Curriculum Vitae

# List of publications

## This thesis was based on the following publications:

- S Y Anvar, K J van der Gaag, J W F van der Heijden, M H A M Veltrop, R H A M Vossen, R H de Leeuw, C Breukel, H P J Buermans, J S Verbeek, P de Knijff, J T den Dunnen, J F J Laros. TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes. Bioinformatics 02/2014
- van der Gaag K J, de Knijff P. Forensic nomenclature for short tandem repeats updated for sequencing. Forensic Sci Int: Gen Supplement Series Volume 5, December 2015, Pages e542-e544
- van der Gaag KJ, de Leeuw RH, Hoogenboom J, Patel J, Storts DR, Laros JFJ, de Knijff P. Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq™ system. Forensic Sci Int Genet. 2016 Sep;24:86-96. doi: 10.1016/j.fsigen.2016.05.016.
- 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.
- van der Gaag KJ, de Leeuw RH, Laros JFJ, den Dunnen JT, de Knijff P. Short hypervariable microhaplotypes: A novel set of very short high discriminating power loci without stutter artefacts. Forensic Sci Int Genet. 2018

## Publications besides this thesis

- Breslin K, Wills B, Ralf A, Ventayol Garcia M, Kukla-Bartoszek M, Pospiech E, Freire-Aradas A, Xavier C, Ingold S, de La Puente M, van der Gaag KJ, Herrick N, Haas C, Parson W, Phillips C, Sijen T, Branicki W, Walsh S, Kayser M. HIrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms. Forensic Sci Int Genet. 2019 Aug 26;43:102152.
- Altena E, Smeding R, van der Gaag KJ, Larmuseau MHD, Decorte R, Lao O, Kayser M, Kraaijenbrink T, de Knijff P. The Dutch Y-chromosomal landscape. Eur J Hum Genet. 2019
- Pośpiech E, Chen Y, Kukla-Bartoszek M, Breslin K, Aliferi A, Andersen JD, Ballard D, Chaitanya L, Freire-Aradas A, van der Gaag KJ, Girón-Santamaría L, Gross TE, Gysi M, Huber G, Mosquera-Miguel A, Muralidharan C, Skowron M, Carracedo Á, Haas C, Morling N, Parson W, Phillips C, Schneider PM, Sijen T, Syndercombe-Court D, Vennemann M, Wu S, Xu S, Jin L, Wang S, Zhu G, Martin NG, Medland SE, Branicki W, Walsh S, Liu F, Kayser M; EUROFORGEN-NoE Consortium. Towards broadening Forensic DNA Phenotyping beyond pigmentation: Improving the prediction of head hair shape from DNA. Forensic Sci Int Genet.

2018 Nov;37:241-251.
- Ingold S, Dørum G, Hanson E, Berti A, Branicki W, Brito P, Elsmore P, Gettings KB, Giangasparo F, Gross TE, Hansen S, Hanssen EN, Kampmann ML, Kayser M, Laurent FX, Morling N, Mosquera-Miguel A, Parson W, Phillips C, Porto MJ, Pośpiech E, Roeder AD, Schneider PM, Schulze Johann K, Steffen CR, Syndercombe-Court D, Trautmann M, van den Berge M, van der Gaag KJ, Vannier J, Verdoliva V, Vidaki A, Xavier C, Ballantyne J, Haas C. Body fluid identification using a targeted mRNA massively parallel sequencing approach - results of a EUROFORGEN/EDNAP collaborative exercise. Forensic Sci Int Genet. 2018 May;34:105-115.
- Van der Gaag K J, Hoogenboom J, Sijen T. Validation and implementation of MPS mtDNA control region analysis for forensic casework: Determination of C-stretch lengths by the FDSTools noise correction feature. Forensic Sci Int: Gen Supplement Series Volume 6, Pages e558–e559.
- Casals F, Anglada R, Bonet N, Rasal R, van der Gaag KJ, Hoogenboom J, Solé-Morata N, Comas D, Calafell F. Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations. Forensic Sci Int Genet. 2017 Sep;30:66-70. doi: 10.1016/j.fsigen.2017.06.006.
- Weiler NE, Baca K, Ballard D, Balsa F, Bogus M, Børsting C, Brisighelli F, ☐ervenáková J, Chaitanya L, Coble M, Decroyer V, Desmyter S, van der Gaag KJ, Gettings K, Haas C, Heinrich J, João Porto M, Kal AJ, Kayser M, Kúdelová A, Morling N, Mosquera-Miguel A, Noel F, Parson W, Pereira V, Phillips C, Schneider PM, Syndercombe Court D, Turanska M, Vidaki A, Woliński P, Zatkalíková L, Sijen T. A collaborative EDNAP exercise on SNaPshot™-based mtDNA control region typing. Forensic Sci Int Genet. 2017 Jan;26:77-84.
- de Haan HG, van Hylckama Vlieg A, van der Gaag KJ, de Knijff P, Rosendaal FR. Male-specific risk of first and recurrent venous thrombosis: a phylogenetic analysis of the Y chromosome. J Thromb Haemost. 2016 Oct;14(10):1971-1977.
- Bertola LD, Jongbloed H, van der Gaag KJ, de Knijff P, Yamaguchi N, Hooghiemstra H, Bauer H, Henschel P, White PA, Driscoll CA, Tende T, Ottosson U, Saidu Y, Vrieling K, de Iongh HH. Phylogeographic Patterns in Africa and High Resolution Delineation of Genetic Clades in the Lion (Panthera leo). Sci Rep. 2016 Aug 4;6:30807.
- Westen AA, Kraaijenbrink T, Clarisse L, Grol LJ, Willemse P, Zuniga SB, Robles de Medina EA, Schouten R, van der Gaag KJ, Weiler NE, Kal AJ, Kayser M, Sijen T, de Knijff P. Analysis of 36 Y-STR marker units including a concordance study among 2085 Dutch males. Forensic Sci Int Genet. 2015 Jan;14:174-81
- Wielstra B, Arntzen JW, van der Gaag KJ, Pabijan M, Babik W. Data concatenation, Bayesian concordance and coalescent-based analyses of the species tree for the rapid radiation of Triturus newts. PLoS One. 2014 Oct 22;9(10)
- Ballantyne KN et al. Toward male individualization with rapidly mutating y-chromosomal short tandem repeats. Hum Mutat. 2014 Aug;35(8):1021-32
- József Geml, Barbara Gravendeel, Kristiaan J van der Gaag, Manon Neilen, Youri

Lammers, Niels Raes, Tatiana A Semenova, Peter de Knijff, Machiel E Noordeloos. The Contribution of DNA Metabarcoding to Fungal Conservation: Diversity Assessment, Habitat Partitioning and Mapping Red-Listed Fungi in Protected Coastal Salix repens Communities in the Netherlands. PLoS ONE 06/2014; 9(6)

- Antoinette A Westen, Thirsa Kraaijenbrink, Elizaveta A Robles de Medina, Joyce Harteveld, Patricia Willemse, Sofia B Zuniga, Kristiaan J van der Gaag, Natalie E C Weiler, Jeroen Warnaar, Manfred Kayser, Titia Sijen, Peter de Knijff. Comparing six commercial autosomal STR kits in a large Dutch population sample. Forensic Science International: Genetics 05/2014; 10C:55-63
- Thirsa Kraaijenbrink, Kristiaan J van der Gaag, Sofia B Zuniga, Yali Xue, Denise R Carvalho-Silva, Chris Tyler-Smith, Mark A Jobling, Emma J Parkin, Bing Su, Hong Shi, V K Kashyap, R Trivedi, T Sitalaximi, Jheelam Banerjee, Karma Tshering Of Gaselô, Nirmal M Tuladhar, Jean-Robert M L Opgenort, George L van Driem, Guido Barbujani, Peter de Knijff. A Linguistically Informed Autosomal STR Survey of Human Populations Residing in the Greater Himalayan Region. PLoS ONE 01/2014; 9(3)
- Frido Welker, Elza Duijm, Kristiaan J. van der Gaag, Bas van Geel, Peter de Knijff, Jacqueline van Leeuwen, Dick Mol, Johannes van der Plicht, Niels Raes, Jelle Reumer, Barbara Gravendeel. Analysis of coprolites from the extinct mountain goat Myotragus balearicus. Quaternary Research. 01/2014
- Jeroen Pijpe, Alex de Voogt, Mannis van Oven, Peter Henneman, Kristiaan J van der Gaag, Manfred Kayser, Peter de Knijff. Indian Ocean crossroads: human genetic origin and population structure in the Maldives. American Journal of Physical Anthropology 03/2013
- Hernando Sanchez-Faddeev, Jeroen Pijpe, Tom van der Hulle, Hans J Meij, Kristiaan J van der Gaag, P Eline Slagboom, Rudi G J Westendorp, Peter de Knijff. The influence of clan structure on the genetic variation in a single Ghanaian village. European journal of human genetics: EJHG 02/2013
- Antoinette A Westen, Kristiaan J van der Gaag, Peter de Knijff, Titia Sijen. Improved analysis of long STR amplicons from degraded single source and mixed DNA. Deutsche Zeitschrift für die Gesamte Gerichtliche Medizin 01/2013
- Iris Kindt, Roeland Huijgen, Marieke Boekel, Kristiaan J van der Gaag, Joep C Defesche, John J P Kastelein, Peter de Knijff. Quality assessment of the genetic test for familial hypercholesterolemia in the Netherlands. Cholesterol 01/2013; 2013:531658.
- Hernando Sanchez-Faddeev, Jeroen Pijpe, David van Bodegom, Tom van der Hulle, Kristiaan J van der Gaag, Ulrika K Eriksson, Thomas Spear, Rudi G J Westendorp, Peter de Knijff. Ancestral stories of Ghanaian Bimoba reflect millennia-old genetic lineages. PLoS ONE 01/2013; 8(6):e65690.
- Antoinette A Westen, Hinda Haned, Laurens J W Grol, Joyce Harteveld, Kristiaan J van der Gaag, Peter de Knijff, Titia Sijen. Combining results of forensic STR kits: HDplex validation including allelic association and linkage testing with NGM and Identifiler loci. Deutsche Zeitschrift für die Gesamte Gerichtliche Medizin

07/2012; 126(5):781-9.
- Luba M Pardo, Giovanna Piras, Rosanna Asproni, Kristiaan J van der Gaag, Attilio Gabbas, Andres Ruiz-Linares, Peter de Knijff, Maria Monne, Patrizia Rizzu, Peter Heutink. Dissecting the genetic make-up of North-East Sardinia using a large set of haploid and autosomal markers. European journal of human genetics: EJHG 02/2012; 20(9):956-64.
- Oscar Lao, Peter M Vallone, Michael D Coble, Toni M Diegoli, Mannis van Oven, Kristiaan J van der Gaag, Jeroen Pijpe, Peter de Knijff, Manfred Kayser. Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. Human Mutation 09/2010; 31(12):E1875-93.
- Richard J L F Lemmers, Patrick J van der Vliet, Kristiaan J van der Gaag, Sofia Zuniga, Rune R Frants, Peter de Knijff, Silvère M van der Maarel. Worldwide population analysis of the 4q and 10q subtelomeres identifies only four discrete interchromosomal sequence transfers in human evolution. The American Journal of Human Genetics 03/2010; 86(3):364-77.
- Daniel Corach, Oscar Lao, Cecilia Bobillo, Kristiaan van Der Gaag, Sofia Zuniga, Mark Vermeulen, Kate van Duijn, Miriam Goedbloed, Peter M Vallone, Walther Parson, Peter de Knijff, Manfred Kayser. Inferring continental ancestry of argentineans from Autosomal, Y-chromosomal and mitochondrial DNA. Annals of Human Genetics 01/2010; 74(1):65-76.
- Mark Vermeulen, Andreas Wollstein, Kristiaan van der Gaag, Oscar Lao, Yali Xue, Qiuju Wang, Lutz Roewer, Hans Knoblauch, Chris Tyler-Smith, Peter de Knijff, Manfred Kayser. Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. Forensic Science International: Genetics 10/2009; 3(4):205-13.
- Wentao Shi, Qasim Ayub, Mark Vermeulen, Rong-guang Shao, Sofia Zuniga, Kristiaan van der Gaag, Peter de Knijff, Manfred Kayser, Yali Xue, Chris Tyler-Smith. A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. Molecular Biology and Evolution 10/2009; 27(2):385-93.
- Richard J L F Lemmers, Marielle Wohlgemuth, Kristiaan J van der Gaag, Patrick J van der Vliet, Corrie M M van Teijlingen, Peter de Knijff, George W Padberg, Rune R Frants, Silvere M van der Maarel. Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. The American Journal of Human Genetics 12/2007; 81(5):884-94.

# 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

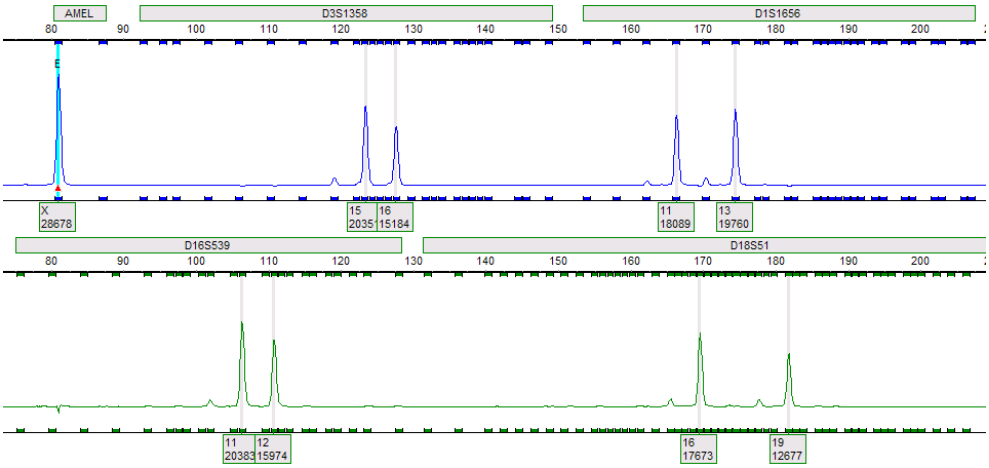
ATCCGA AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT TGGACCAG

## 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.

## 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.

# Nederlandse samenvatting

Dit proefschrift richt zich op de ontwikkeling, toepassing en validatie van nieuwe forensische methoden gebruik makend van nieuwe beschikbare technieken bekend als 'Massively Parallel Sequencing' (afgekort MPS, ook bekend onder de naam 'Next Generation Sequencing').

In **hoofdstuk 1** wordt de achtergrond beschreven van MPS zelf en de verschillende methoden die verder in het proefschrift worden besproken.

De ontwikkeling en toepassing van forensisch DNA onderzoek heeft zich snel ontwikkeld sinds de ontdekking van zeer variabele 'DNA vingerafdrukken' door Jeffreys et al. (1985).

De laatste ca. twintig jaar heeft de analyse van 'Short Tandem Repeats' (STRs) een grote rol gespeeld in identificatie m.b.v. DNA onderzoek. STRs zijn stukjes DNA die een herhaalde sequentie bevatten van 2-6 nucleotiden. Het aantal herhalingen van deze sequentie kan variëren tussen personen en door het analyseren van meerdere van deze STRs kan een vrijwel uniek DNA profiel worden vastgesteld. Een voorbeeld van een stukje DNA sequentie met daarin een STR wordt hieronder weergegeven.

**Figuur 1,** DNA sequentie die een AGAT-herhaling bevat

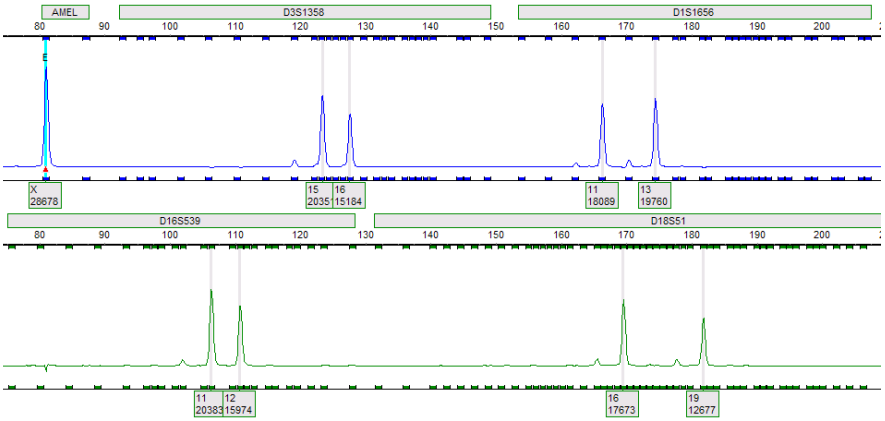
ATCCGA AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT TGGACCAG

## Conventionele DNA profielen

Capillaire elektroforese (CE) is de conventionele methode voor het analyseren van STRs. Voor analyse d.m.v. CE worden fragmenten die een STR bevatten eerst vermeerderd en gelabeld m.b.v. de zogenaamde polymerase chain reaction (PCR). Vervolgens worden de fragmenten op lengte van elkaar gescheiden en vergeleken met een ladder van fragmenten met een bekende lengte om zo het aantal herhalingen van elke STR vast te stellen in een DNA profiel.

Hoewel CE analyse van STRs een vrij eenvoudige techniek is, heeft het ook zijn limieten. Tijdens de amplificatie van STR-fragmenten ontstaan zgn. stutter-artefacten die vooral de interpretatie van een ongebalanceerd gemengd DNA profiel ingewikkelder kunnen maken. Daarnaast zijn er, om een voldoende uniek DNA profiel te verkrijgen, suboptimale condities nodig voor de analyse van gedegradeerd DNA. Figuur 2 geeft een voorbeeld van een (deel van een) DNA profiel.

**Figuur 2,** Voorbeeld van een DNA profiel




*Deel van een DNA profiel van een (enkele) persoon. De pieken zijn weergegeven voor vijf markers die gescheiden zijn bij lengte (in hetzelfde kleurkanaal) of fluorescent label.*

## Nieuwe technologieën: Massively Parallel Sequencing

Recent zijn er nieuwe DNA analyse technieken ontwikkeld die parallel de sequentie kunnen bepalen van miljoenen DNA moleculen. Terwijl deze MPS-technieken al toegepast worden in het medische DNA onderzoek en andere moleculaire onderzoeksgebieden zijn ze nog relatief nieuw in het forensische werkveld.

In principe zou MPS een aantal beperkingen van CE moeten kunnen voorkomen. Enkele beperkingen in CE STR analyse worden echter veroorzaakt door het type DNA marker en niet door de gebruikte techniek. Aangezien het type data, maar vooral de hoeveelheid data voor MPS anders is dan bij CE, is het cruciaal dat er gespecialiseerde forensische software ontwikkeld wordt om goed met deze data om te gaan en MPS toepassing in zaakwerk te implementeren.

In dit proefschrift wordt de toepassing van MPS in humaan forensisch DNA onderzoek verkend en wordt niet enkel het lab-gerelateerde praktische aspect onderzocht, maar ook de ontwikkeling van gespecialiseerde forensische software voor analyse van MPS data. Daarnaast verkennen we nieuwe niet-STR DNA markers als alternatief voor de huidige analyse om de toepassing van DNA analyse in de nabije toekomst nog verder uit te breiden.

## Massively Parallel Sequencing vs Capillaire Elektroforese

De verwachte perspectieven van MPS in forensisch DNA onderzoek zijn gebaseerd op de volgende verschillen tussen CE en MPS.

- *Terwijl CE enkel fragmentlengte analyseert, typeert MPS de exacte sequentie.*

Sequentie variatie (bovenop lengtevariatie) verhoogt de statistische kracht van een STR. Daarnaast kan sequentie-variatie helpen om stutter artefacten en echte allelen van elkaar te onderscheiden.

- *De detectierange van CE is beperkt terwijl MPS een vrijwel ongelimiteerde detectierange heeft.* Voor CE zorgt te hoog signaal voor extra artefacten (in andere kleuren) terwijl erg laag signaal niet van ruis onderscheiden kan worden. Voor MPS kan het aantal uitgelezen sequenties vrijwel oneindig worden verhoogd zonder dat er artefacten ontstaan bij andere STRs (al stijgen de kosten mee met het aantal sequenties).
- *De capaciteit van CE om meerdere markers tegelijk te analyseren is beperkt terwijl MPS in principe duizenden markers tegelijk kan analyseren.* Aangezien er momenteel niet meer dan zes fluorescente labels gebruikt kunnen worden met CE, wordt het gewenste aantal markers in een reactie verkregen door de lengte van verschillende markers niet te laten overlappen. Voor MPS kunnen alle markers in dezelfde fragment-range ontworpen worden omdat ze onderscheiden worden op sequentie in plaats van kleur en lengte.

## Massively Parallel Sequencing van Short Tandem Repeats

Het meeste werk dat beschreven wordt in dit proefschrift richt zich op MPS analyse van STRs. Aangezien forensische DNA databanken zijn opgebouwd met STR data is het logisch om te beginnen met de analyse van deze zelfde markers zodat er nog altijd zoekingen uitgevoerd kunnen worden in de DNA databank.

### Basale analyse van MPS STR data: TSSV

Terwijl STRs de gebruikelijke markers zijn voor forensisch DNA onderzoek, worden ze niet heel veel gebruikt in andere DNA onderzoeksgebieden. De initieel beschikbare software pakketten voor MPS data analyse konden dan ook slecht om gaan met herhalende sequenties. In **hoofdstuk 2** van dit proefschrift wordt de (open-source) software TSSV beschreven. Dit was een van de eerste software pakketten vooral gericht op STR analyse. Door het gebruik van twee anker sequenties (aan beiden kanten van de STR) worden markers herkend en vervolgens wordt de waargenomen variatie tussen deze anker sequenties samengevat en worden herhaalde sequenties afgekort. Naast de analyse van STRs bleek deze software ook goed toepasbaar voor de analyse van errors in sequentie-data.

### Voor sequentie data is ook sequentie nomenclatuur nodig

Allelen van CE worden benoemd als een nummer wat het aantal herhalingen in

de STR weergeeft. Aangezien het bepalen van de exacte sequentie extra variatie aan toont, is er een nomenclatuur nodig die ook deze variatie kan beschrijven. In **hoofdstuk 3** beschrijven we de eerste aanbevelingen voor een forensische STR sequentie nomenclatuur met het doel om alle relevante sequentie informatie weer te geven in een zo kort en leesbaar mogelijke naam.

### Validatie van een (prototype) commerciële methode voor STR sequentie-analyse

In samenwerking met het bedrijf Promega© werd een prototype van een commerciële methode, ontwikkeld voor sequentie-analyse van STRs, getest. Met deze methode werd het DNA van 17 STRs en een marker voor geslachtstypering vermenigvuldigd in een reactie en werd, na verdere voorbereiding, de sequentie bepaald met behulp van de Miseq™ sequencer. Hoofdstuk 4 is een gedetailleerde beschrijving van de analyse van de prestaties van deze methode en de waargenomen sequentie-variatie van elke STR.

Om te berekenen hoe uniek een DNA profiel is, moeten eerst de frequenties vastgesteld worden van elke sequentie-variant in de populatie. Wanneer deze frequenties bekend zijn kan er een statistische kansberekening gebruikt worden om in te schatten hoe groot de kans is dat een willekeurig persoon een bepaald DNA profiel heeft. De STR-sequenties van 297 DNA-monsters van personen uit een Europese, Aziatische en Afrikaanse populatie werden bepaald. Zo werd inzicht verkregen in de extra sequentie-variatie t.o.v. de CE lengte-variatie van de STRs en de frequenties van deze sequentie-varianten. Voor de meeste STRs werden substantieel meer sequentie-varianten waargenomen dan CE lengte varianten.

Bij de analyse van STRs resulteert de PCR in het vormen van zogenaamde stutter-artefacten die veroorzaakt worden door een 'slippend' enzym (Polymerase). STR stutter bemoeilijkt de interpretatie van ongebalanceerde mengsels aangezien pieken van de lager bijdragende donor niet altijd onderscheiden kunnen worden van stutter. De extra sequentie-informatie levert meer inzicht in deze stutter. Het bleek dat de hoeveelheid stutter bepaald wordt door de lengte van de langste ononderbroken herhaling in de STR. Deze informatie gaf nieuwe mogelijkheden voor de interpretatie van gemengde profielen op basis van sequentie-data.

## Correctie van stutter in een monster m.b.v. bioinformatica tools

In **hoofdstuk 5** is de bioinformatica software FDSTools beschreven als de eerste software die werkelijk PCR- en sequentie-artefacten kan corrigeren in MPS data. Op basis van een training dataset wordt systematische ruis vastgesteld en deze informatie wordt vervolgens gebruikt om in een onbekend monster de ruis te corrigeren zodat

de basislijn van de data substantieel wordt gereduceerd.

De software bevat ook tools om kwaliteitscontrole uit te voeren op de referentie data en om analysegrenzen vast te stellen na het uitvoeren van de correctie. Met behulp van deze vastgestelde grenzen kan data van onbekende monsters worden gefilterd en wordt de data weergegeven in een interactieve format.

Analyse van ongebalanceerde mengsels toonde aan dat allelen van de laag bijdragende donor vaker werden teruggevonden na correctie terwijl deze zonder correctie niet herkend konden worden.

## Alternatieve forensische markers voor STRs

Ook met de verbeterde analyses m.b.v. bioinformatica tools blijft het duidelijk dat STRs niet voor alle toepassingen ideale forensische markers zijn. Terwijl de hoge mate van variatie helpt om een zo uniek mogelijk profiel te verkrijgen met een beperkt aantal markers zorgen stutter artefacten en lengte-variatie voor beperkingen in de interpretatie van mengsels en de analyse van afgebroken DNA. In **hoofdstuk 6** zijn microhaplotypes geselecteerd uit publiek beschikbare genoom data die vier of meer SNPs (DNA verschillen op een enkele positie) bevatten in een fragment van 70 baseparen. Deze markers kunnen bijna net zo variabel zijn als STRs zonder het nadeel van stutter artefacten en variatie in lengte.

De studie toonde aan dat het merendeel van deze geselecteerde fragmenten voortkwamen uit verkeerd gerapporteerde variatie in de publiek beschikbare genoom data maar er werd toch een subset van 16 microhaplotypes geselecteerd met een discriminerend vermogen dat ongeveer overeen komt met negen STRs. Naast het hoge discriminerende vermogen lijkt de data ook informatie te geven over de biogeografische afkomst van een persoon. Het was namelijk mogelijk om aan de hand van de data de drie geteste populaties vrijwel compleet van elkaar te scheiden.

## Conclusies en toekomstperspectieven

In **hoofdstuk 7** wordt het geheel van de uitgevoerde studies besproken in combinatie met de toekomstperspectieven. MPS is nu klaar om toegepast te worden in forensisch zaken en wordt ook al in enkele zaken toegepast. Terwijl MPS nu nog een relatief dure methode is, is het niet onmogelijk dat dit in de nabije toekomst verandert en dat CE geleidelijk vervangen kan worden door MPS wanneer het in een gestroomlijnde geautomatiseerde manier toegepast kan worden. MPS lijkt een veelbelovende tool voor de analyse van mengsels, die mogelijk is door het sequencen van STRs, maar ook door het sequencen van andere markers zoals microhaplotypes. Er zijn veel andere toepassingen voor MPS die niet (of beperkt) mogelijk zijn met CE, vooral door het hogere aantal markers dat in een keer geanalyseerd kan worden. Momenteel

zijn er al toepassingen voor voorspelling van biogeografische herkomst en een aantal uiterlijke kenmerken op basis van SNP. Voor een meer gedetailleerde voorspelling van biogeografische herkomst, maar ook voor veel andere uiterlijke karakteristieken moet nog meer basale kennis verzameld worden voor het toegepast kan worden in zaken. Bij verschillende instituten worden momenteel studies uitgevoerd om deze basale kennis uit te breiden.

## Dankwoord

Deze thesis was er niet geweest zonder support van vele mensen. Peter, al toen ik nog als analist werkzaam was, heb ik de basis mogen leggen om verder te ontwikkelen tijdens mijn promotie. Er was altijd ruimte om out-of-the-box te denken en nieuwe dingen te proberen. Door samenwerking met het LGTC, via Johan en Yavuz en collega's kon altijd gebruik gemaakt worden van de nieuwste technologieën. Rick, het was fijn om te sparren en los te gaan met nieuwe ideeën. Zonder jouw hulp bij het labwerk en het uitwerken van analyses hadden we nooit zo ver kunnen komen. Dit geldt ook voor Jeroen en boven al voor Jerry. Jullie hebben me kennis laten maken met de wereld van de bioinformatica. Eveline, je bracht een hoop fijne humor en nuchterheid, jouw support, maar ook die van Sofia, Rene, Thirsa, Risha, Ron, Patricia, Monique, Denise en Laura zorgde voor een inspirerende omgeving bij het FLDO.

De inspiratie om dit project te starten kwam zeker niet alleen van werk. Zonder Martijn als grote broer en voorbeeld op het gebied van studie en wetenschap was ik nooit zo ver gekomen, graag had ik dit traject met jou gedeeld… Gelukkig is de steun van mijn ouders, mijn zus Simone en Geert (en natuurlijk Pepijn) er altijd onvoorwaardelijk geweest.

Het vertrouwen om het promotieproject te starten is ondersteund door meerdere mensen. Reinout, maar ook zeker Loredana hebben me geïnspireerd en overtuigd om te gaan promoveren. Verder heeft de afwisseling van onderzoek en muziek me altijd geholpen om met enthousiasme door te gaan. Marco, Dave, Sven, Marijn, Gil en Rob, jullie hebben me geholpen de juiste balans te houden.

Bij het NFI mocht ik het laatste 'stukje' van de promotie afmaken. Team Research heeft me welkom ontvangen en gesteund om alles af te ronden. Titia, momentum houden en reële doelen zetten voor jezelf, niemand kon me dit beter leren dan jij. Maar ook Corina, Margreet, Francisca, wederom Jerry, Sofia, Natalie, Josja, Patrick, Jeroen en alle fijne stagiaires hebben, samen met mijn coaching team Peter en Ate, mogen meegenieten van alle geduld die er nodig was om zo ver te komen. Mijn beste vrienden Gideon, Marco en Reinout en in het bijzonder natuurlijk Melanie hebben me gesteund en me zo sterk geholpen tot deze afronding. Mel, het is heerlijk om hier nu met jou, maar ook met Owen, Gwen en Jill van te kunnen genieten.

## Curriculum Vitae

Kristiaan (Johannes) van der Gaag was born on the 5th of June 1980 in Vlaardingen, the Netherlands. He completed primary school in Vlaardingen and graduated from high school in 1997 in Vlaardingen (HAVO, CSG Aquamarijn). In 1997 Kris started a Bachelor of Biomedical Science in Rotterdam (Hogeschool Rotterdam). In 2002 he started an internship at the Forensic Laboratory for DNA Research (FLDO, LUMC, Leiden the Netherlands) focussing on the validation of Y chromosome SNP typing by Pyrosequencing in order to obtain information on the geographic ancestry of a donor of a forensic trace and later on using MALDITOF (Matrix-Assisted Laser Desorption / Ionisation Time-Of-Flight mass spectrometry) as an alternative method for forensic SNP-typing. He started a position as laboratory technician at the same department in 2002 and received his Bachelor degree in 2003. During his period as technician, he performed laboratory and analysis work on STRs and SNPs in many population genetic projects, developed many SNP genotyping assays (mostly SNaPshot assays for Y-chromosome and mtDNA) and later on also performed forensic casework.

When the Forensic Genomics Consortium Netherlands (FGCN) was founded in 2009, Kris started to focus more on the application of Massively Parallel Sequencing and the corresponding analysis and software development, initially using the 454-sequencer and the PGM Ion Torrent, but later mostly using the Miseq platform namely focussing on forensics. In addition, many projects were conducted collaborating with partners from the medical and evolutionary genetic field. As part of the FGCN consortium (Forensic Genomics Consortium Netherlands), he started a PhD position at the FLDO in 2010 focussing on development of forensic genomics toolkits by the use of Massively Parallel Sequencing (as described in this thesis). Since 2015, Kristiaan is employed at the Netherlands Forensic Institute as a scientist in the Research team of the Division Biological Traces implementing MPS applications in casework where he will continue to conduct research after completing his PhD.