Library construction optimization and analysis of Short Tandem Repeats by a simple, PCR-based DNA barcoding method

Master Thesis Project

By: Markus André Soma

June 2020

Division: Applied Microbiology

Supervisors: Johannes Hedman & Linda Jansson

External supervisor: Maja Sidstedt

Examiner: Peter Rådström

Abstract

DNA profiling is evolving in the forensics community towards introducing massively parallel sequencing (MPS) as a complement to capillary electrophoresis (CE). Obstacles remain however before this technology can become routine in forensic casework, such as the development of more efficient bioinformatics solutions and recommendations concerning data analyses. Additionally, there is a need to develop MPS methods that are even more sensitive than current commercialized systems. One promising candidate is SiMSen-Seq, a method that incorporates unique molecular identifiers (UMI's), also known as barcodes, into PCR library preparation, allowing for the reduction of background noise (artefacts) in data analyses. In the development of SiMSen-Seq towards its use in forensics, the library preparation protocol, consisting of two distinct PCRs (PCR1 and PCR2), were in this project further optimized for the efficient amplification of short tandem repeats (STRs). By applying Bioanalyzer 2100, the results showed that the type of DNA polymerase and the barcode primer concentration had the greatest effect in maximizing specific products and minimizing nonspecific products. SuperFi and Immolase, two promising DNA polymerases that resulted in efficient STR amplification, were further evaluated for use in library preparation by MPS using MiSeq, and the results showed that although the use of SuperFi in PCR1 and Immolase in PCR2 resulted in the most STR products, it generated the highest amount of artefacts, complicating data interpretation. Instead, utilizing SuperFi, a proofreading enzyme, in PCR2 of library preparation, decreased the amount of generated artefacts. Based on these results, it is therefore recommended that further tests are performed with SuperFi in both PCR1&2 of library preparation. Testing other DNA polymerase combinations featuring proofreading abilities may also provide valuable data. However, before SiMSen-Seq can be implemented in the analyses of real crime scene samples, additional evaluation using low DNA concentrations and more complex DNA samples, including inhibitors, is required. Further customization of the bioinformatic data workflow is also necessary to streamline the work process.

Preface

This 20-week master project (January-June 2020) was conducted within the forensics group at the division of Applied Microbiology at Lund University and in collaboration with the Swedish National Forensic Centre (NFC). Being part of a larger project known as ULTRA-UDI (financed by VINNOVA), this study was aimed at further optimizing a method that in the future may be used to analyze crime scene samples. This gave me the opportunity to learn more and work within a field that I have found very interesting for many years. The study trips to Linköping and Gothenburg were truly highlights, and I am very thankful for being invited.

I would like to thank my examiner Peter Rådström for his time and constructive feedback. I would also like to thank Anders Ståhlberg, Gustav Johansson, and Tobias Österlund at Gothenburg University, and Adam Staadig and Andreas Tillmar at the Swedish National Board of Forensic Medicine, for their help in this project. I am forever grateful for the time, patience, kindness, and expertise my supervisors Linda Jansson, Johannes Hedman, and Maja Sidstedt gave me. Thank you for making me feel so welcome and part of the team. I would also like to thank everyone at the division of Applied Microbiology for making these 20 weeks the best time I have had at Lund University.

Popular science summary

Title: Forensic DNA analyses of complex crime scene samples can be improved using an optimized method

Optimization of the method led to increased production of DNA that is necessary to correctly identify and distinguish one person from another. DNA errors were reduced, and interestingly, changing one factor of the method had the power to further reduce DNA errors, simplifying interpretation of a DNA sample.

The Swedish National Forensic Centre (NFC) handles thousands of crime scene samples every year, many of which are highly complex or contain very little evidence in the form of DNA, complicating the work for forensic scientists. In this project, the focus is on the optimization of a method that may one day lead to improved analysis of crime scene DNA evidence, such that in the future, it will be easier to connect perpetrators with crimes.

When a method is being adapted from its use in one field to another, in this case from cancer diagnostics to forensic analyses, optimization under controlled conditions (i.e. in the laboratory) is an important part of the development process. By changing different factors of the method, it can be fine-tuned towards providing better quality data, making analyses easier. Although these factors can be difficult and time-consuming to discover, it is an important job towards implementing a new method in a laboratory handling real samples.

One key factor determined in this project was the type of DNA polymerase used when running the method. The DNA polymerase can be thought of as a machine that is necessary in building more DNA from the very small amounts of DNA that is found at the crime scene (e.g. in blood or saliva), enabling identification of who left the stain. An additional important factor determined in this optimization was the primer concentration. The primer specifies what part of the DNA that should be built, such that the identity of an individual can be correctly determined and distinguished from another person's DNA (as the DNA differs between individuals). The essence of the method is that something that can be thought of as a barcode will be added to all of the newly built DNA, which allows for a smoother process in identifying the person(s) who committed the crime.

After testing many different types of DNA polymerases, as well as various primer concentrations and other factors of the method, one specific DNA polymerase at a set primer concentration led to the production of less DNA errors, making it easier to identify the correct person(s) in the sample. DNA errors often occur when building new DNA fragments, and some DNA polymerases produce more and others less, which was shown in this study. It is therefore important to use this specific DNA polymerase further or test other low error polymerases when running this method. Errors in the produced DNA could also be reduced when taking advantage of the barcodes, simplifying data interpretation. It was then also possible to successfully distinguish two different persons from each other based on differences in their DNA. Although no real crime samples were analyzed in this project, the results look very promising.

Towards the future development of this method, more testing should be done with the goal to further reduce the amount of errors produced in the DNA when using the method. Additional customization of the analyses workflow when handling computer software is also required to further advance the method in reaching its full potential, such that it may one day be used to analyze real crime scene samples.

List of abbreviations

CE	Capillary electrophoresis
DNA	Deoxyribonucleic acid
LUS	Longest uninterrupted stretch
MPS	Massively parallel sequencing
NGS	Next generation sequencing
PCR	Polymerase chain reaction
SiMSen-Seq	Simple, Multiplexed, PCR-based barcoding of DNA for Sensitive mutation detection using Sequencing
STRs	Short tandem repeats
UMI	Unique molecular identifier

Table of contents

1 Introduction

In recent years, there has been an increasing interest among forensic genetic institutes towards developing technology for DNA profiling by massively parallel sequencing of short tandem repeats (STR-MPS) (Alonso et al., 2017). However, the transition from using capillary electrophoresis (CE), today's golden standard of forensic analyses (Alonso et al., 2018), to STR-MPS, requires the development and implementation of bioinformatics solutions for efficient data analyses (Borsting and Morling, 2015).

Among the STR-MPS multiplex systems commercially available today is the Forenseq™ DNA Signature Prep Kit. Although a highly robust, reliable, and reproducible method according to validation studies (Jäger et al., 2017), the analyses of a minor contributor in a mixed DNA sample is limited down to 5% of the major contributor (Alonso et al., 2018). This limitation is due to the background noise caused commonly by polymerase errors during PCR amplification (Fox et al., 2014), such as stutters (Walsh et al., 1996, Hauge and Litt, 1993). One approach to reduce background noise that has significantly improved sensitivity in fields such as cancer research is by incorporating unique molecular identifiers (UMI's) into PCR library preparation (Ståhlberg et al., 2016). Often referred to as barcodes, UMI's allow for the removal of error-induced sequence variants, including stutters, that can otherwise complicate NGS analyses (Filges et al., 2019).

This project, performed within the forensics group at the division of Applied Microbiology at Lund University, and in collaboration with the Swedish National Forensic Centre (NFC), aims at further optimizing a library construction protocol, central to the method first described by (Ståhlberg et al., 2016) as "Simple, Multiplexed, PCR-based barcoding of DNA for Sensitive mutation detection using Sequencing (SiMSen-Seq)". By combining an optimized library construction protocol for efficient barcoding of STRs with a compatible analysis toolkit (Ståhlberg et al., 2017), SiMSen-Seq has the potential to provide a more sensitive method for DNA profiling by STR-MPS than current NGS technologies, allowing a theoretical profile detection from a mixed sample down to ~ 0.1% total DNA (Ståhlberg et al., 2016).

1.1 Scope

The scope of this project involves experimental testing of further optimizing the SiMSen-Seq library preparation protocol, plus data analyses using bioinformatics software. Data analyses involves interpretation of gel-images, electropherograms, and next generation sequencing data using bioinformatics software.

1.2 Aim

The main goals of this project are to:

- Optimize the SiMSen-Seq protocol for PCR amplification of STRs.
- Develop a data analysis workflow for STRs provided by SiMSen-Seq data.
- Compare the effect of using different polymerase combinations in SiMSen-Seq library preparation.
- Demonstrate proof-of-concept for applying SiMSen-Seq in STR-MPS by a 7-plex.
- Distinguish a minority DNA profile from a mixed DNA sample using the developed SiMSen-Seq protocol and analysis workflow.

2 Background

The theoretical background for the applied methods and data analyses are presented in this section.

2.1 PCR

First described in the 1980's (Mullis and Faloona, 1987), the polymerase chain reaction (PCR) revolutionized molecular biology and has since become a routine procedure in many fields. Capable of amplifying minute amounts of DNA, PCR can allow for the detection and analysis of genetic material from sources such as humans, bacteria, and viruses (Garibyan and Avashia, 2013).

2.1.1 Principles of PCR

Running a PCR requires four essential regents: template DNA, primers, nucleotides, and a DNA polymerase. The template DNA is the genetic material that is targeted for amplification, and the primers specify which region of the DNA template that is to be amplified. The nucleotides adenine, thymine, cytosine, and guanine are needed to produce complementary DNA strands from the template DNA, which is catalyzed by the enzyme DNA polymerase (Garibyan and Avashia, 2013).

Combined in a tube in balanced concentrations and placed in a thermocycler machine, the four reagents described above will allow for running a successful PCR that occurs in three distinct steps. The first step, known as denaturation, requires high temperatures to separate the DNA double helix to its two complementary single strands. The next step, known as annealing, occurs at a lower temperature, and allows for the binding of primers to their specific site on the single-stranded DNA. Once the primers have attached, the third step commences, known as extension, in which the temperature is raised once again, allowing the DNA polymerase to catalyze the addition of complementary nucleotides to a growing DNA strand. These three steps can then be repeated for additional cycles, elegantly doubling the desired DNA fragment after each cycle (Garibyan and Avashia, 2013).

2.1.2 DNA polymerase

The DNA polymerase enzyme comes in many forms, although for use in PCR, the thermostable DNA polymerase isolated from *Thermus aquaticus* (Chien et al., 1976), referred to as *Taq* polymerase, is among the most known and commonplace (Saiki et al., 1988). Nowadays recombinantly produced, *Taq* polymerase is commercialized for example as AmpliTaq DNA polymerase (Ishino and Ishino, 2014). In the field of forensics, AmpliTaq Gold, a derivative of *Taq* polymerase, has been used in kits (AMPFlSTR® SGM Plus™) (Cotton et al., 2000) worldwide for amplifying DNA retrieved from crime scenes (Hedman et al., 2009).

Polymerases differ from each other in their abilities. One such ability is proofreading, also known as 3'-5' exonuclease activity. Enzymes with proofreading mechanisms are known for having lower error rates during PCR, translating to fewer artefacts and less background noise in data analyses (Ishino and Ishino, 2014, Filges et al., 2019). Polymerases that have low error rates are also commonly referred to as high-fidelity polymerases. The fidelity of an enzyme is often compared to that of *Taq* (Filges et al., 2019). For instance, Platinum SuperFi II, one of the polymerases used in this project, has a fidelity of over 300x to that of *Taq* (Invitrogen, 2019). Additional abilities include thermostability and processivity. Thermostability is especially important, as DNA polymerases that lack sufficient thermostability cannot be used in PCR applications due to the high temperatures that the reaction requires (Ishino and Ishino, 2014). Processivity on the other hand is the measure of how efficiently the DNA polymerase can continuously synthesize the new DNA strand without dissociating (Zhuang and Ai, 2010), a feature that is important when copying longer DNA templates (Wang et al., 2004).

Interestingly, certain polymerases are also more resistant to PCR inhibitors (Abu Al-Soud and Râdström, 1998, Hedman et al., 2009), which is of great importance in forensics, as contaminants such as humic substances (Sidstedt et al., 2015), commonly found in soil, are known to complicate analyses of crime scene samples (Sidstedt et al., 2020). Choosing a polymerase with inhibitor resistance in mind can therefore add a valuable ability to PCR library preparation.

2.1.3 PCR optimization

There are numerous parameters that can be adjusted and fine-tuned when running a PCR that may lead to higher reaction specificity and efficiency. Primer design, magnesium concentration, type of polymerase, thermal cycle modifications, and additives, are all key troubleshooting factors that can be targeted when performing PCR optimization (Lorenz, 2012).

Among the most effective parameters to change however during optimization is the magnesium concentration and the annealing temperature. Varying the final magnesium concentration will affect the reaction specificity. The general trend is that as the magnesium concentration is raised, more wanted products are formed, but at a cost of generating more nonspecific products. This is due to the stabilizing role magnesium has on DNA, as concentrations set too high will stabilize the binding of primers to the DNA template, including nonspecific binding. Lowering the concentration will have the opposite effect, decreasing the amount of products formed, although increasing specificity, resulting in less nonspecific products. Additionally, as the DNA polymerase requires magnesium as a cofactor, setting the concentration too low will in fact prevent amplification (Lorenz, 2012). Changes to the annealing temperature will also affect specificity. Annealing temperatures set too low will stimulate the formation of nonspecific products, whereas annealing temperatures set too high will significantly reduce the yield of wanted products (Rychlik et al., 1990). A relationship can then be drawn for these two parameters, as increasing the magnesium concentration is equivalent to lowering the annealing temperature, leading to more nonspecific products, whereas decreasing the magnesium concentration is equivalent to increasing the annealing temperature, leading to less nonspecific products.

2.1.4 DNA profiling

Analyses of STRs have long been the standard in forensics casework for the identification of perpetrators, a technique known as DNA profiling (Gill, 2002). Today performed using CE (Alonso et al., 2018), the technique relies on that a person can be distinguished from another based on statistical probability. The AMPFlSTR® SGM Plus™ multiplex kit for example, released in 1999 to forensic institutes, had a probability of randomly matching two DNA profiles of 1 in 10^9 (Cotton et al., 2000, Gill, 2002). The probability is linked to which STR markers (Cotton et al., 2000) and how many STR markers that are included in the analyses, as increasing the number of STRs studied will also reduce the probability of accidentally matching two unrelated individuals (Gill, 2002).

2.2 Massively parallel sequencing

Massively parallel sequencing (MPS), commonly referred to as next generation sequencing (NGS) (Borsting and Morling, 2015), is nowadays a widespread technology with applications in fields such as biotechnology and medical diagnostics (Bruijns et al., 2018). Since the introduction of Sanger sequencing in 1977 (Sanger et al., 1977), the costs of sequencing have been significantly reduced and its speed multiplied through the use of MPS (Borsting and Morling, 2015).

Capable of offering superior sensitivity, there is a growing interest among forensic institutes (Alonso et al., 2017) to further develop and implement MPS in casework, replacing current CE methods (Alonso et al., 2018). One such MPS system available today is the Forenseq™ DNA Signature Prep Kit, commercialized by Illumina (Jäger et al., 2017). This system is however limited to identifying individuals represented below 5% of the total DNA (Alonso et al., 2018), a shortcoming due to the background noise generated from PCR and sequencing errors (Fox et al., 2014). For instance, stutters (Walsh et al., 1996) are common artefacts that arise during PCR and are caused by the insertion or deletion of an STR motif by the DNA polymerase (Brookes et al., 2012, Hauge and Litt, 1993). Other commonly encountered errors during MPS analyses are base-pair substitutions (Schirmer et al., 2015, de Knijff, 2019). Clear recommendations regarding analytical thresholds (de Knijff, 2019), as well as more sensitive methods are therefore needed in STR-MPS analyses that can decrease the amount of background noise and allow for the identification of individuals represented at very low DNA concentrations (<1%) (Ståhlberg et al., 2017, Alonso et al., 2018).

2.2.1 SiMSen-Seq

SiMSen-Seq was originally developed for cancer diagnostics in the detection of rare variant alleles represented down to 0.1% (Ståhlberg et al., 2016). Capable of reducing background noise (Ståhlberg et al., 2017), SiMSen-Seq sparked interest for its use in other fields, including forensics, where it may allow for the detection of individuals represented at very low DNA concentrations (~ 0.1%).

Central to SiMSen-Seq is the library preparation, involving two distinct rounds of PCR: PCR1 and PCR2. In PCR1, the DNA region of interest is amplified in a total of 3 thermal cycles. The primers used in PCR1 are specifically designed to contain a hairpin structure that protects the barcode sequence as well as an adapter sequence. The hairpin also functions to minimize nonspecific binding and formation of primer-dimers. In PCR2, Illumina sequencing adaptor primers that are complementary to the adapter sequences linked during PCR1, are used to further amplify the barcoded products in additional thermal cycles. The generated products will then consist of three different parts: 1) an adaptor sequence complementary to Illumina adapters on the flow cell, 2) a barcode sequence that is unique for each original DNA molecule, and 3) the sequence of interest (Ståhlberg et al., 2017).

2.2.2 Bioinformatics data analysis

With the advent of MPS comes big data, and bioinformatics software solutions are required to efficiently store, handle, and analyze the millions of sequences that can be generated from a single run (Greene et al., 2014). Such bioinformatic pipelines are in continuous development, ranging from software provided by different MPS manufacturers to open source toolkits freely available online or for download (Liu and Harbison, 2018). ToaSTR (Ganschow et al., 2018) and FDSTools (Hoogenboom et al., 2017) are among such open source solutions that are developed for forensics, and are in this project used, together with UMIErrorCorrect (unpublished at time of writing), to discover and evaluate the effect of barcoding by SiMSen-Seq (Ståhlberg et al., 2016) in STR-MPS.

3 Materials and methods

In this section, the materials and methods used in this project are presented.

3.1 Materials

A list of the consumables and reagents, and equipment used in this project along with their respective manufacturer are shown in Table 1 and Table 3, respectively. Shown in Table 2 are the DNA polymerases and buffers used and their respective manufacturer. Fidelity compared to *Taq* is included and whether the DNA polymerase has 3’-5’ exonuclease activity.

Table 1: List of consumables and reagents used in this project.

Consumables and reagents	**Manufacturer**
Agilent High Sensitivity DNA kit	Agilent Technologies, Inc. (Santa Clara, U.S.A.)
AMPure XP	Beckman Coulter (Brea, U.S.A.)
Barcode primers (100 µM)	Integrated DNA Technologies (Coralville, U.S.A.)
Index primers (100 µM)	Integrated DNA Technologies (Coralville, U.S.A.)
KAPA Library Quantification Kit, Universal qPCR mix	Roche Diagnostics (Basel, Switzerland)
L-carnitine inner salt (≥98%)	Sigma-Aldrich (St. Louis, U.S.A.)
$MgCl_2$ stock solution (25 mM)	Roche Diagnostics (Basel, Switzerland)
MiSeq FGx Reagent Kit	Verogen (San Diego, U.S.A.)
PCR nucleotide mix (10 mM)	Roche Diagnostics (Basel, Switzerland)
PhiX Control v3	Illumina (San Diego, U.S.A.)
DEPC – Treated Water Nuclease Free (0.2 µm filtered)	Ambion (Austin, U.S.A.)
Tris-EDTA pH 8.0 Solution	Medicago AB (Uppsala, Sweden)
Protease *Streptomyces griseus* type XIV	Sigma-Aldrich (St. Louis, U.S.A.)
2800M Control DNA (10 ng/µL)	Promega (Madison, U.S.A.)

Table 2: List of polymerases and buffers used in this project. Whether the DNA polymerase has proofreading ability (3'-5' exonuclease activity) is included and the fidelity compared to Taq is shown (data retrieved from manufacturer websites and user manuals). Data not found is represented by a hyphen (-).

DNA polymerase	**Buffer**	**Manufacturer**	**3'-5' exonuclease activity**	**Fidelity vs. *Taq***
AccuStart™ Taq HiFi (5 U/µL)	HiFi PCR Buffer 10x	Quanta Biosciences Inc. ™ (Beverly, U.S.A.)	Yes	6x (Quantabio, n.d.)
AmpliTaq Gold® (5 U/µL)	AmpFlSTR® PCR Reaction Mix	Applied Biosystems (Foster City, U.S.A.)	No	1x (Applied Biosystems, 2014)
IMMOLASE™ (5 U/µL)	ImmoBuffer (10X)	Bioline (London, U.K.)	No	-
KAPA HiFi HotStart (1 U/µL)	KAPA HiFi Fidelity Buffer 5x	KAPA Biosystems (Wilmington, U.S.A.)	Yes	100x (KAPA Biosystems, 2019)
KOD Xtreme™ Hot Start (1 U/µL)	2x Xtreme Buffer	Novagen® (Darmstadt, Germany)	Yes	10x (Novagen, n.d.)
Phusion™ Hot Start II (2 U/µL	5X Phusion HF Buffer	Thermo Fisher Scientific (Waltham, U.S.A.)	Yes	50x (Thermo Fisher Scientific, 2018)
Platinum™ SuperFi II (2 U/µL)	5X SuperFi II Buffer	Invitrogen™ (Carlsbad, U.S.A.)	Yes	>300x (Invitrogen, 2019)
PowerPlex ESX	5X Master Mix	Promega (Madison, U.S.A.)	-	-
Q5® Hot Start High-Fidelity (2 U/µL)	Q5® Reaction Buffer	New England Biolabs Inc. (Ipswich, U.S.A.)	Yes	~280x (New England Biolabs, n.d.)
TEMPase Hot Start (5 U/µL)	10x Ammonium Buffer	Ampliqon (Odense, Denmark)	No	-

Table 3: List of equipment used in this project.

Equipment	**Manufacturer**
Centrifuge 5424	Eppendorf (Hamburg, Germany)
CFX96 Touch™ Real-Time PCR Detection System	Bio-Rad Laboratories, Inc. (Hercules, U.S.A.)
GeneAmp™ PCR System 9700	Applied Biosystems (Foster City, U.S.A.)
LightCycler® 8-Tube Strips (clear)	Roche Diagnostics (Basel, Switzerland)
Microcentrifuge MiniStar silverline	VWR (Radnor, U.S.A.)
MiSeq FGx System	Illumina (San Diego, U.S.A.)
Optical Flat 8-Cap Strips for 0.2 mL tube strips/plates	Bio-Rad Laboratories, Inc. (Hercules, U.S.A.)
PCR Strips without Caps, low profile, white	Bio-Rad Laboratories, Inc. (Hercules, U.S.A.)
UVC/T-M-AR, DNA/RNA UV-cleaner box	Biosan (Riga, Latvia)
Vortex-Genie 2	Scientific Industries, Inc. (New York, U.S.A.)
2100 Bioanalyzer	Agilent Technologies, Inc. (Santa Clara, U.S.A.)

3.2 Methods

The SiMSen-Seq method, thoroughly described in (Ståhlberg et al., 2017), is the foundation for the methods presented in this section. Since work in adapting SiMSen-Seq towards forensic use had been done prior to the start of this project, an initial protocol was already available including specific STR barcode primers. From the initial protocol, an optimized protocol was developed after an optimization process involving the alteration of various PCR factors, including the type of DNA polymerase, barcode primer concentration, magnesium concentration, and thermal cycle programs.

The library construction method of SiMSen-Seq is in focus for this project, with the aim of optimizing the protocol such as to generate more specific and less nonspecific products during PCR amplification. Library construction involves two distinct rounds of PCR: PCR1 and PCR2. During PCR1, the STR markers of interest are amplified in 3 thermal cycles using STR specific barcoded primers. After the last thermal cycle in PCR1, a mixture consisting of *Streptomyces griseus* protease (Sigma-Aldrich) is added, which inactivates the DNA polymerase and also dilutes the reaction by a factor of 1:3. The products of PCR1 are then added to a new PCR master mix, containing a new set of reagents and primers that are specific for the amplified barcoded products generated in PCR1. In PCR2, Illumina index adapters are added to the barcoded STR fragments and are amplified for an additional 33 thermal cycles.

All optimization work applying Bioanalyzer used 2800M control DNA. The samples used for MPS were generated using 2800M control DNA and the DNA from two different individuals (mixed samples). Each reaction performed used a total of 20 ng DNA.

The amplicons of interest, their theoretical lengths, and their observed lengths as determined by single-plex using Bioanalyzer 2100 are shown below in Table 4. The expected size in base-pairs when using SiMSen-Seq in single-plex may vary ±10% as reported by (Ståhlberg et al., 2017), however the highest variability seen for the 7 STR markers was ±2% (n=3). Described further below is the initial protocol available at project start, followed by the optimized protocol developed for both PCR1 and PCR2.

Table 4: The 7 STR markers used for multiplex PCR and their theoretical and observed product lengths for analysis of 2800M control DNA. The observed product lengths may vary ±2% as seen from single-plex replicates between different analyses of the same STR marker (n=3).

STR marker	2800M alleles	Theoretical product length (bp)	Observed product length (bp)
D2S441	10, 14	234/250	244/261
D1S1656	12, 13	283/287	300
D3S1358	17, 18	285/289	297
vWA	16, 19	297/309	307/320
D12S391	18, 23	371/391	377/396
D21S11	29, 31.2	373/383	384/394
D8S1179	14, 15	378/382	400

3.2.1 PCR1 in library preparation

The master mix protocols, initial and optimized, used for PCR1 of library preparation are shown in Table 5 and Table 6, respectively. In a UV-cleaner box (Biosan) the reagents listed were combined in a microcentrifuge tube and vortexed using a Vortex-Genie 2 (Scientific Industries). Next, 10 µL master mix was distributed to each PCR tube (Bio-Rad Laboratories, Inc.), and amplified using GeneAmp™ PCR System 9700 (Applied Biosystems) with the program shown in Table 7, which was the same for both the initial and the optimized protocol, except for the initial denaturation time, which was 10 min for Immolase and 2 min for SuperFi. Increasing the initial denaturation time for SuperFi to 10 min had no observable effect. However, according to the SuperFi user guide (Invitrogen, 2019), a 30 sec denaturation at 98°C is sufficient, although this was not tested with the optimized protocol.

Table 5: Master mix content for the initial protocol for PCR1. For one 10 µL reaction, the following reagents were mixed to the final concentration shown.

Reagent	Final concentration/**amount**
DEPC – Treated Water	-
ImmoBuffer (10X)	1 X
$MgCl_2$ (25 mM)	2.5 mM
L-carnitine inner salt (2 M)	0.5 M
dNTP (2 mM)	0.2 mM
Barcode primer mix (0.2 µM)*	40 nM
2800M Control DNA (10 ng/µL)	**20 ng**
IMMOLASE™ DNA Polymerase (5 U/µL)	**0.1 U**

*D2, D1, D3, vWA, D12, D21, D8

Table 6: Master mix content for the optimized protocol for PCR1. For one 10 µL reaction, the reagents below were mixed to the concentrations shown.

Reagent	Final concentration/**amount**
DEPC – Treated Water	-
SuperFi II Buffer (5X)	1 X
$MgCl_2$ (25 mM)	1 mM
L-carnitine (2.5 M)	0.5 M
dNTP (10 mM)	0.2 mM
Barcode primer mix (0.2/0.3 µM)*	40/60 nM**
2800M Control DNA (10 ng/µL)	**20 ng**
Platinum™ SuperFi II DNA Polymerase (2 U/µL)	**0.1 U**

* D2, D1, D3, vWA, D12, D21, D8, **All primers 60 nM except vWA at 40 nM.

Table 7: PCR1 thermal cycle program. The program consists of initial denaturation (10 min for Immolase and 2 min for SuperFi at 98°C), denaturation, annealing, extension, enzyme inactivation, protease inactivation, and hold. Only three cycles are run in PCR1. As soon as the enzyme inactivation step starts, 20 µL of protease solution is added to inactivate the polymerase.

Initial denaturation of Immolase / **SuperFi**	Denaturation	Annealing	Extension	Enzyme inactivation	Protease inactivation	Hold
98°C / **98°C**	98°C	58°C	72°C	65°C*	95°C	4°C
10 min / **2 min**	10 sec	6 min	30 sec	15 min	15 min	∞
	3 cycles					

*Immediate addition of 20 µL protease solution (10 µL 667x griseus to 1 mL TE-buffer) to each well.

3.2.2 PCR2 in library preparation

The initial and optimized PCR2 master mix protocols are shown in Table 8, the only difference being the volume of PCR1 reaction products added to the PCR2 mix (5 µL and 10 µL for the initial and optimized protocols, respectively). In a UV-cleaner box (Biosan), the reagents listed were combined in a microcentrifuge tube, except for the PCR1 reaction products which were added immediately before PCR2 amplification start. A volume of 15 µL from the master mix was then distributed to each PCR tube (Bio-Rad Laboratories, Inc.). Thereafter, the PCR1 reaction products were added, to a final volume of 20 µL and 25 µL for the initial and optimized protocol, respectively. The products were then amplified using CFX96 Touch™ Real-Time PCR Detection System (Bio-Rad Laboratories, Inc.). The PCR2 thermal cycle program used was the same for both protocols and is shown in Table 9.

Table 8: Master mix for initial and optimized protocol of PCR2. The reagents were combined to a final volume of 20 µL and 25 µL per reaction for the initial and optimized protocol, respectively.

Reagent	Final concentration/**amount**
DEPC – Treated Water	-
ImmoBuffer (10 X)	1 X
$MgCl_2$ (25 mM)	2.5 mM
L-carnitine inner salt (2 M)	0.5 M
dNTP (2 mM)	0.2 mM
Index primer F (10 µM)	0.4 µM
Index primer R (10 µM)	0.4 µM
IMMOLASE™ DNA Polymerase (5 U/µL)	**1 U**
PCR 1 reaction products	**5 µL / 10 µL***

*In the optimized protocol, the volume of added PCR1 products was doubled, from 5 µL to 10 µL.

Table 9: PCR2 thermal cycle program. The program includes initial denaturation, denaturation, two annealing steps, extension, and hold. The ramping speed is set at 0.2°C/s for 33 cycles.

Initial denaturation	Denaturation	Annealing	Annealing	Extension	Hold
98°C	98°C	80°C	72°C	76°C	4°C
10 min	10 sec	1 min	30 sec	30 sec	∞
	33 cycles, ramping 0.2°C/s				

3.2.3 Optimization of library preparation

Initial optimization testing was done in single-plex, varying different PCR parameters listed in Table 10. Further experiments, such as polymerase testing, were performed in 7-plex, aiming at demonstrating proof-of-concept for using SiMSen-Seq in STR-MPS. The different polymerases tested during library optimization are presented in Table 2. As library preparation involves two distinct PCR's (PCR1 and PCR2), one complete library preparation requires the use of two polymerases, either the same polymerase in both PCR1 and PCR2 or two different polymerases. Several polymerases were therefore tested for PCR1 and PCR2 in 48 combinations (presented in Table 11). PowerPlex ESX Master Mix was included as a reference, as it has been used in kits worldwide for the amplification of STRs in forensic samples (Ruitberg et al., 2001).

Other parameters of PCR that were tested during the optimization process included different reagent concentrations and variations in the thermal cycle program as shown in Table 10. These parameters were mostly tested in single-plex experiments.

Table 10: Optimization of PCR reagent concentrations and thermal cycle program. Magnesium concentrations between 1-4 mM were tested, and barcode primer concentrations at 40, 60 and 80 nM. Annealing temperatures between 48 and 62 degrees were tested in PCR1, including two different annealing and extension times at 30 and 90 seconds for PCR2.

PCR reagent	**Concentration**
$MgCl_2$ (mM)	1, 1.5, 2, 2.5, 3.5, 4
Barcode primer mix (nM)	40, 60, 80
Thermal cycle program	**Value**
PCR1 Annealing temperature (°C)	48, 56, 58, 60, 62
PCR2 Annealing time (s)	30, 90
PCR2 Extension time (s)	30, 90

3.2.4 Analysis using capillary electrophoresis

For the separation and quantitation of amplified DNA products, a 2100 Bioanalyzer (Agilent Technologies, Inc.) was used. The High Sensitivity DNA chips have on-chip electrophoresis for separating and quantifying DNA fragments having a length from 50 to 7,000 base-pairs at concentrations between 5-500 pg/µL. Fragments are separated by electrophoretic separation and detected and quantified by a fluorometer using fluorescent dye (Gromadski et al., 2016).

After PCR2, the DNA samples were diluted in TE buffer (10X/20X) and thereafter loaded into wells according to Agilent's protocol (Agilent Technologies, 2017). The procedure is described below. All pipetting was done by reverse pipetting, except when adding the ladder.

The gel-dye mix, marker, and ladder tubes (reagents part of kit components) were equilibrated to room temperature for at least 30 minutes. A new High Sensitivity DNA Chip was then placed in the chip priming station and 9 µL gel-dye mix was added to the well labeled **G**. The plunger was then positioned at 1 mL and the chip priming station was carefully closed. The plunger was thereafter pushed firmly and gently down, and a timer was started at exactly 60 s. After 60 s, the plunger was released and untouched for an additional 5 s. The plunger was then carefully pulled back up to the 1 mL position, and the chip priming station was opened. 9 µL gel-dye mix was then added to the three other wells marked G. 5 µL marker was then added to all wells (except the four wells marked G containing gel-dye mix), including 1 µL ladder mix to the labeled ladder well. 1 µL of the diluted amplified products (10/20X in TBE) were then added to the 11 sample wells. The chip was vortexed at 2400 rpm for 1 min, and thereafter placed in the Bioanalyzer 2100 and the run was started.

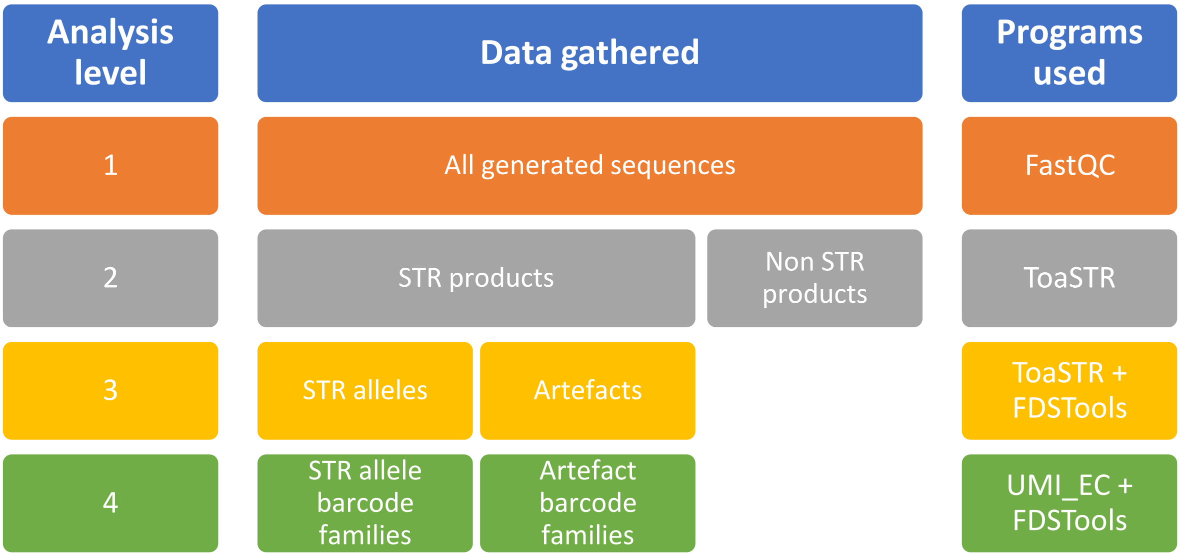
3.2.5 Analysis using MiSeq and bioinformatics software

Sequencing was performed at NFC in Linköping. A summary of the procedure is detailed below.

The DNA polymerase combinations (PCR1/PCR2) used for library preparation were SuperFi/Immolase, Immolase/SuperFi, SuperFi/SuperFi, and Immolase/Immolase, all amplified in duplicates, along with a negative control, following the optimized library preparation protocol previously described (only parameter that was varied was the polymerase). Mixed DNA samples from two individuals at three different ratios (1:1, 3:1, and 9:1) were also amplified in duplicates using the optimized protocol and the combination SuperFi/Immolase. All samples had a total DNA concentration of 20 ng. A negative control was also included using the combination SuperFi/Immolase, giving a total of 15 samples.

After library preparation by PCR1 and PCR2, the samples were cleaned using Solid Phase Reversible Immobilization (SPRI) beads to remove short, unwanted amplicons such as primer dimers. Before sequencing, the samples were first quantified by quantitative PCR (qPCR) using KAPA Library Quantification Kit and then diluted to achieve equal concentrations for all samples in the sequencing. Samples were then pooled and denatured such as to generate single-stranded DNA allowing binding to adapters on the flow-cell. The pooled DNA samples were then loaded into a cassette containing the necessary reagents for amplification, and the sequencing was initiated. Upon completion of the sequencing, the MiSeq FGx instrument performed initial data analysis, sorting the different 15 samples based on sample specific index sequences linked during PCR2 (Seidlitz et al., 2019).

The NGS data analyses for this project can be sorted into four levels, showed in Figure 1, where each level increase represents a deeper data analysis of the former. The raw data output from the MiSeq analysis is first processed to FastQ-files, which is the input file for all programs used: FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), ToaSTR (https://www.toastr.de/), FDSTools (https://www.fdstools.nl/), and UMIErrorCorrect (unpublished at time of writing). Running data through the UMIErrorCorrect pipeline and FDSTools program required programming knowledge and was therefore performed externally (Gothenburg University and NFC) and the result files from these analyses were used for further data analyses running Excel.



Analysis level	Data gathered			Programs used
1	All generated sequences			FastQC
2	STR products		Non STR products	ToaSTR
3	STR alleles	Artefacts		ToaSTR + FDSTools
4	STR allele barcode families	Artefact barcode families		UMI_EC + FDSTools

Figure 1: Overview of STR-MPS data analyses workflow. Each analysis level shows the information gathered along with respective programs used. At analysis level 1, all generated sequences in each FastQ file is obtained using the program FastQC. Out of these sequences, at analysis level 2, only the ones containing STRs are identified as STR products with ToaSTR, whilst the rest are classified as non STR products. At level 3, the alleles and artefacts for each STR product were examined using an additional program called FDSTools. The alleles and artefacts were then grouped into barcode families using UMIErrorCorrect and evaluated using FDSTools at analysis level 4.

4 Results

The results presented in this section include data analyses from the optimization process of further adapting SiMSen-Seq for forensic STR profiling. The effect of using different polymerase combinations for PCR1 and PCR2 in library preparation is shown, and how the use of barcodes can simplify forensic data analyses through the removal of artefacts such as stutters.

During the optimization process, the effects of altering magnesium concentration, annealing temperature and time, and extension temperature and time, were mostly determined by single-plex experiments. These parameters remained unchanged from the initial to the optimized protocol as they appeared optimal for the process and/or did not have a significant effect in generating more correct products and less nonspecific products (data not shown). The factors which had the greatest effect however were the DNA polymerase and the barcode primer concentrations, tested in 7-plex.

4.1 Optimization of library preparation applying Bioanalyzer

Throughout the library optimization process, a range of different DNA polymerases were tested in PCR1/PCR2 combinations as shown in Table 11 below. The use of different polymerase combinations for SiMSen-Seq library preparation showed a great effect on the generation of specific and nonspecific products. Products having a length in the size range of 50-239 base-pairs are considered artefacts, as all correct products should be in the range of 240-405 base-pairs (see Table 4). The number of detected STR markers as seen in the electropherogram for each respective combination is also included, as not all products formed within the correct product region were STR alleles. 10 out of the total 48 tested polymerase combinations enabled detection of 6-7 out of the maximum 7 included STR markers (marked in green). Among these, SuperFi in PCR1 and Immolase in PCR2 (further combinations represented in text as PCR1 polymerase/PCR2 polymerase) had the highest product concentration measured at 25,795 ± 311 pg/µL, followed by SuperFi/Phusion at 10,119 pg/µL, both using a barcode primer concentration of 60 nM for all markers except vWA at 40 nM. PowerPlex ESX Master Mix was included as a reference, and it gave mostly 7/7 detected markers. However, this DNA polymerase was not used in further analyses as it was part of a pre-made master mix that included unspecified reagent concentrations. Product concentrations for PowerPlex ESX Master Mix combinations could also not be quantified (represented as NA) due to an error in the lower and upper marker when running these samples on the Bioanalyzer.

Table 11: Different PCR1/PCR2 polymerase combinations tested during optimization, showing the concentrations of amplicons for two distinct size regions and the number of detected markers in the electropherogram. Amplicons having a length between 50-239 base-pairs are unwanted products (artefacts). All 7 STR markers of interest are contained within the size range of 240-405 base-pairs, however, unwanted products are also present for certain combinations. Therefore, the number of detected markers out of the maximum 7 is presented for each combination, visually determined from each respective electropherogram. The color indicates the level of visually detected markers: red: 0-2 markers, yellow: 3-5 markers, and green: 6-7 markers. Standard barcode primer concentrations are 40 nM for all markers. NA: Data not available due to Bioanalyzer error in lower/upper marker.

PCR1	**PCR2**	**Artefacts region 50-239 bp (pg/µL)**		**Correct product region 240-405 bp (pg/µL)**		**Number of detected markers (max 7)**	
AccuStart	AccuStart	11975	16791	3024	4125	3/7	3/7
	AmpliTaq Gold	9012	7246	696	242	0/7	0/7
	Immolase	18093	21804	12524	15550	2/7	2/7
	KAPA	27746	22126	5761	4875	0/7	0/7
	Phusion	28195	19470*	18654	18472*	2/7	5/7*
	PowerPlex	NA		NA		7/7	
	Q5	21158	24397	12232	14585	0/7	0/7
	SuperFi	370161	24793*	57216	18097*	2/7	3/7*
	Tempase	24088	25970*	14971	9812*	1/7	2/7*
AmpliTaq Gold	AccuStart	9138		631		3/7	
	AmpliTaq Gold	4734	2677	247	0	0/7	0/7
	Immolase	8890	22091	7509	5067	7/7	2/7
	KAPA	20078	28139	2852	5650	0/7	0/7
	PowerPlex	NA		NA		7/7	
	Q5	26256	28360	9083	9251	0/7	0/7
	Tempase	25612		4236		0/7	
Immolase	AccuStart	14211	16377*	3394	3616*	3/7	3/7*
	Immolase	8667 ± 5075	8988**	6218 ± 2373	8387**	7/7	7/7**
	KAPA	19403		3967		0/7	
	Phusion	17159	12607*	12969	8674*	2/7	2/7*
	PowerPlex	NA		NA		0/7	
	Q5	21130		10121		0/7	
	SuperFi	16377	10756**	9818	7321**	3/7	3/7**
KAPA	Immolase	7007	15919	6832	16888	0/7	0/7
	KAPA	23644		9484		0/7	
	Q5	12957		11908		0/7	
Q5	Immolase	21182	19409	16102	15089	0/7	0/7
	KAPA	28357	25738	7469	6926	0/7	0/7
	Q5	15814	22195	10926	15074	1/7	1/7
KOD	Immolase	4512	3451	6927	6245	3/7	
Phusion	Immolase	8443	3107	9182	4001	7/7	7/7
	Phusion	17269	23349*	19067	13771*	1/7	3/7*
	PowerPlex	NA		NA		7/7	
	SuperFi	17059	20046*	21542	21314*	0/7	0/7*
	Tempase	13802	22292*	14983	12540*	3/7	3/7*
SuperFi	Immolase	11826 ± 3059	5634 ± 323**	9356 ± 1871	25795 ± 311**	7/7	7/7**
	Phusion	9476	6062*	10119	9194*	6/7	6/7*
	SuperFi	8907	9220**	5836	10738**	4/7	5/7**
	Tempase	19985	22087*	769	7080*	3/7	5/7*
Tempase	AccuStart	13273		2997		0/7	
	AmpliTaq Gold	7373	4983	624	3	1/7	0/7
	Immolase	5821	11814	3119	7171	7/7	7/7

PCR1	PCR2	Artefacts region 50-239 bp (pg/µL)		Correct product region 240-405 bp (pg/µL)		Number of detected markers (max 7)	
	KAPA	24114		3583		0/7	
	Phusion	10077	4456*	7485	6642*	5/7	7/7*
	PowerPlex	NA		NA		0/7	
	Q5	22860		8865		0/7	
	SuperFi	16056	15607*	9992	10150*	3/7	3/7*
	Tempase	24396	27894	13724	6636	0/7	4/7

*All barcode primer concentrations at 80 nM except vWA at 40 nM, **All barcode primer concentrations at 60 nM except vWA at 40 nM.

In SiMSen-Seq, barcoding takes place during the three thermal cycles of PCR1. Increasing the barcode primer concentration from 40 nM to 60 nM for all primers except vWA had a large positive effect in terms of generating more specific products and less nonspecific products for certain DNA polymerase combinations, especially for SuperFi/Immolase, as shown in Figure 2. The concentration of products formed within the defined region of 240-405 base-pairs increased from 9,356 ± 1871 pg/µL to 25,795 ± 311 pg/µL after elevating the primer concentration.

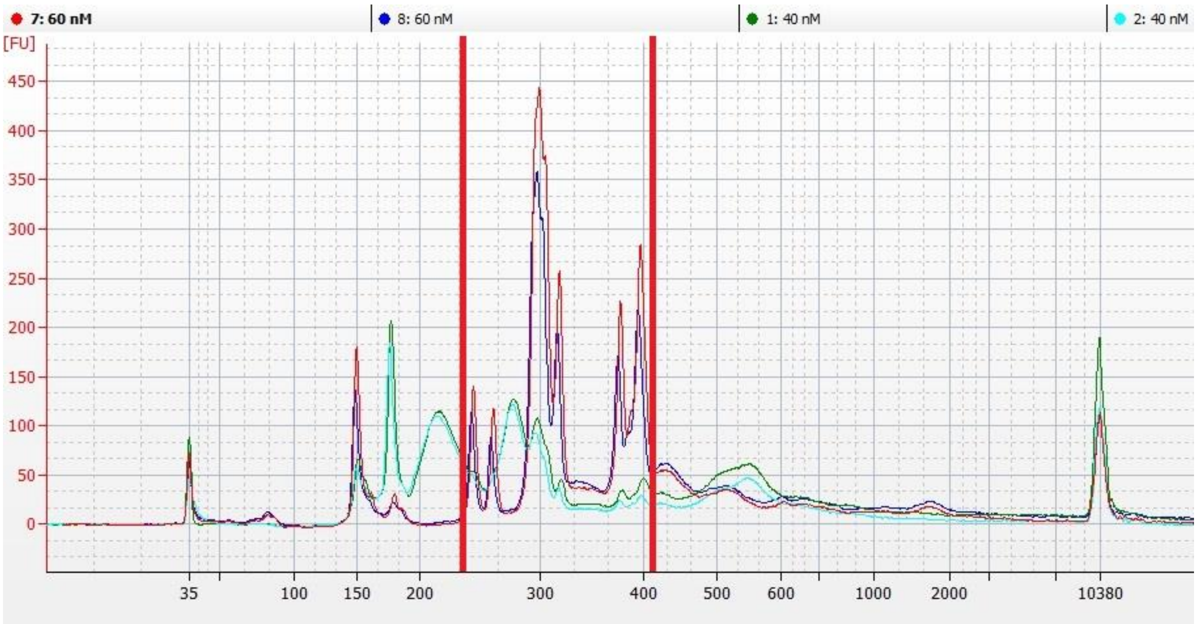


Figure 2: Combination SuperFi/Immolase at 40 nM (green and cyan) and 60 nM (red and blue) barcode primer concentration compared in duplicates. The products recorded between the two red lines are between the size range of 240-405 base-pairs, which is the region that includes all 7 STR markers.

In the optimized protocol, SuperFi was applied in PCR1 and Immolase in PCR2 of SiMSen-Seq library preparation. This polymerase combination along with an increased primer concentration in PCR1 resulted in the generation of more specific products, as shown in Table 12 and Figure 3 below.

The concentration of products formed within the desired base pair range increased more than fourfold with the optimized protocol, from 6,218 ± 2373 (Immolase/Immolase) to 25,795 ±311 pg/µL (SuperFi/Immolase).

Table 12: Comparison of initial and optimized protocol showing parameters changed, recorded concentration of desired products within the size range of 240-405 base pairs, and artefacts between 50-239 base pairs. The magnesium concentration and thermal cycle parameters remained unchanged from the initial to the optimized protocol, whereas the polymerase combination and primer concentration was altered. Data values were retrieved from duplicate samples (n=2).

	Initial protocol	**Optimized protocol**
Polymerase combination	Immolase/Immolase	SuperFi/Immolase
Primer concentration in PCR1	40 nM all markers	60 nM all markers except 40 nM for vWA
Magnesium concentration	2.5 mM	2.5 mM
Annealing temperature in PCR1 / Annealing and extension time in PCR2	58°C / 30 s	58°C / 30 s
Amount of product in artefact region (50-239 base-pairs)	8,667 ± 5075 pg/µL	5,634 ± 323 pg/µL
Amount of product in STR marker region (240-405 base-pairs)	6,218 ± 2373 pg/µL	25,795 ± 311 pg/µL

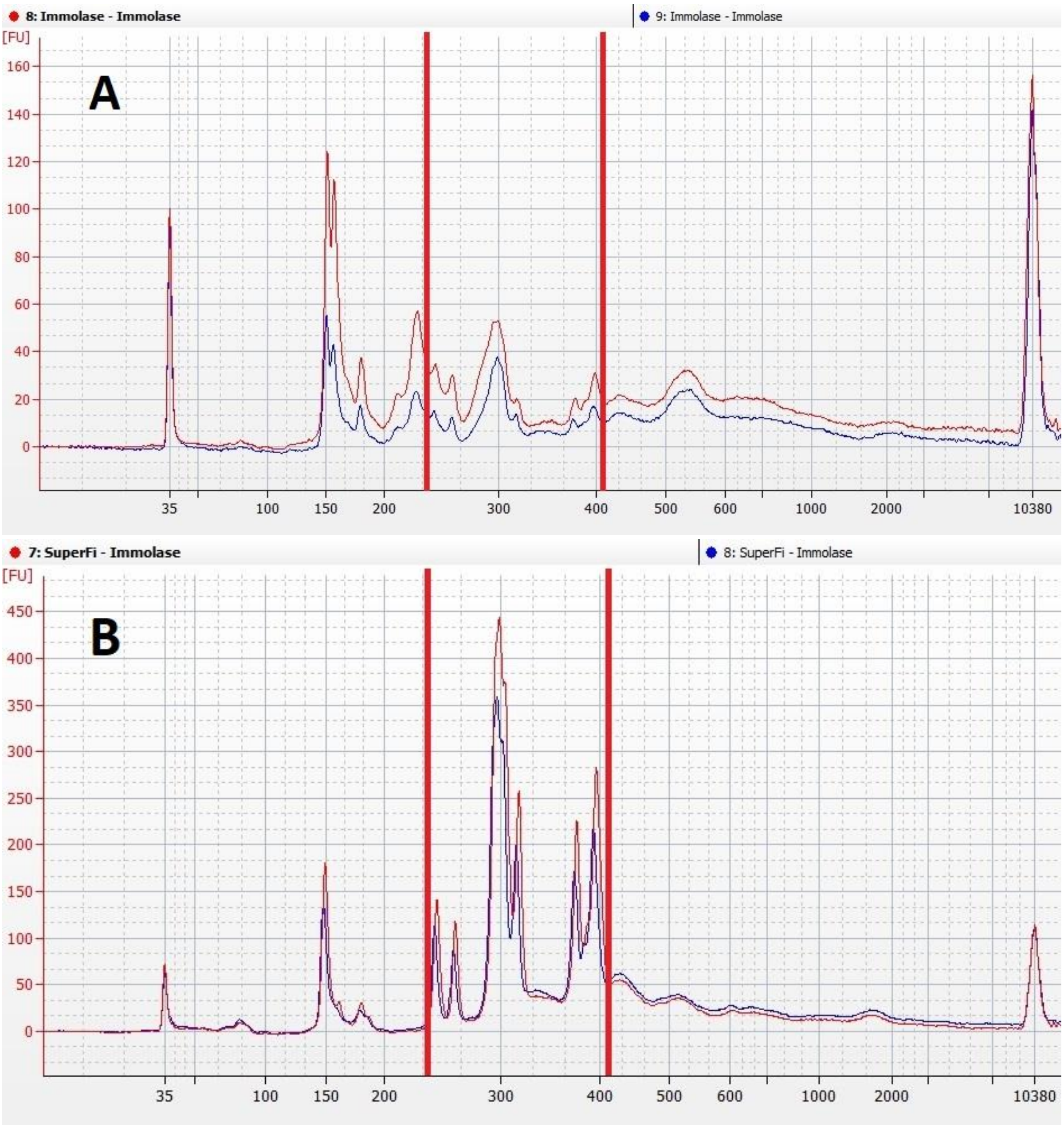


Figure 3: Initial protocol (A) and optimized protocol (B) measured in duplicates, with an amplicon size range between 240-405 marked within red lines. A) Amplified products (20X diluted) of 2800M control DNA using the DNA polymerase combination Immolase/Immolase with 40 nM barcode primer concentration. B) Amplified products (10X diluted) of 2800M control DNA using combination SuperFi/Immolase with 60 nM primer concentration for all markers except vWA at 40 nM .

4.2 Evaluation of DNA polymerase combinations by sequencing

Based on the most promising DNA polymerase combinations, library preparation with SuperFi and Immolase in different combinations was evaluated by sequencing. At analysis level 2 (see Figure 1), the number of STR product reads and non STR product reads from the total reads were evaluated. STR product reads are reads that are identified as STRs, whereas non STR product reads are reads not identified as STRs by the web application tool ToaSTR. After library preparation using the four DNA polymerase combinations, the products were sequenced by MiSeq in duplicates, with an overview of the data presented in Figure 4. The use of SuperFi in PCR1 as opposed to Immolase greatly affected the amount of non-STR products generated, as SuperFi/Immolase generated an average of 0.3 million non STR product reads, whereas Immolase/Immolase resulted in an average of nearly 1.2 million reads of non STR products. The highest number of STR product reads was generated using SuperFi/Immolase (average 1.04 million reads), followed by SuperFi/SuperFi (average 0.77 million reads).

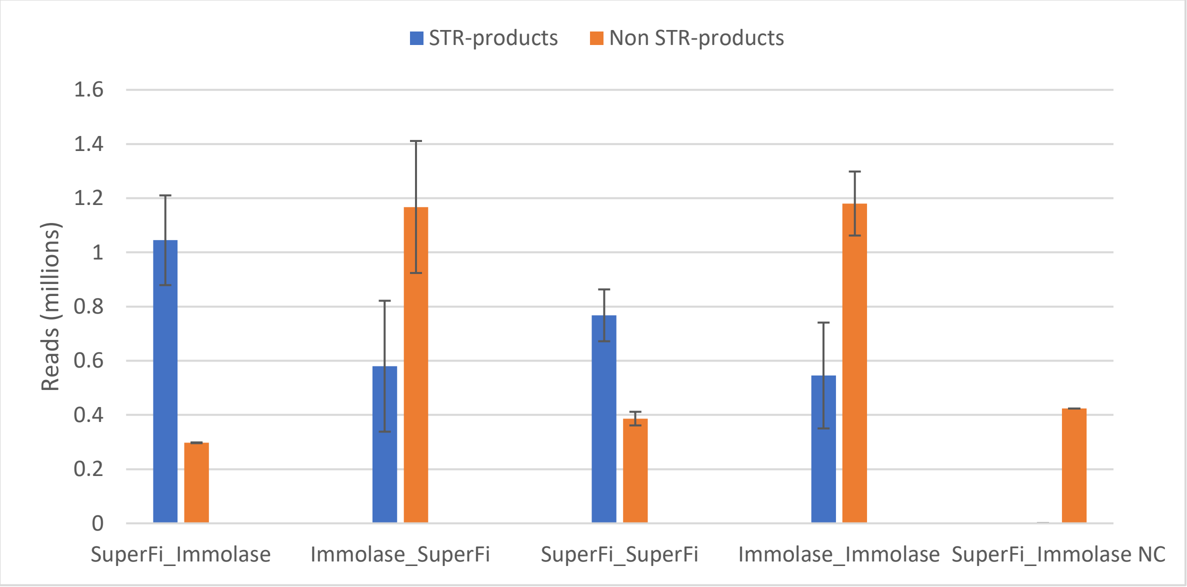


Figure 4: Number of reads identified as STR products (blue) and non STR products (orange) in millions for the four DNA polymerase combinations, including a negative control of SuperFi/Immolase. The data was determined using the bioinformatics tools FastQC and ToaSTR, with an analytical threshold of 100 reads for identifying STR product reads. Average values are presented. n=2.

The different combinations of SuperFi and Immolase for PCR1 and PCR2 in library preparation also gave substantial differences in the quality of NGS data when evaluating STR alleles and artefacts at analysis level 3 (see Figure 1). STR alleles represent the number of reads that are identified as true alleles for each respective STR marker (for 2800M alleles, see Table 4), whereas reads identified as artefacts are stutters and other PCR or sequencing errors. Combinations containing SuperFi in PCR2 as opposed to Immolase showed overall less artefacts (Figure 5). SuperFi in PCR2 also resulted in the lowest percentage of artefacts generated per STR marker. Interestingly, there was also variability among the different STR markers, as vWA and D12 consistently showed a higher degree of generated artefacts. The DNA polymerase combination of SuperFi/Immolase resulted in the highest number of reads that were specific for the correct alleles (average 770 000), followed closely by SuperFi/SuperFi (average 670 000). However, the amount of artefacts generated was higher for SuperFi/Immolase (average 280 000) compared to SuperFi/SuperFi (average 93 000).

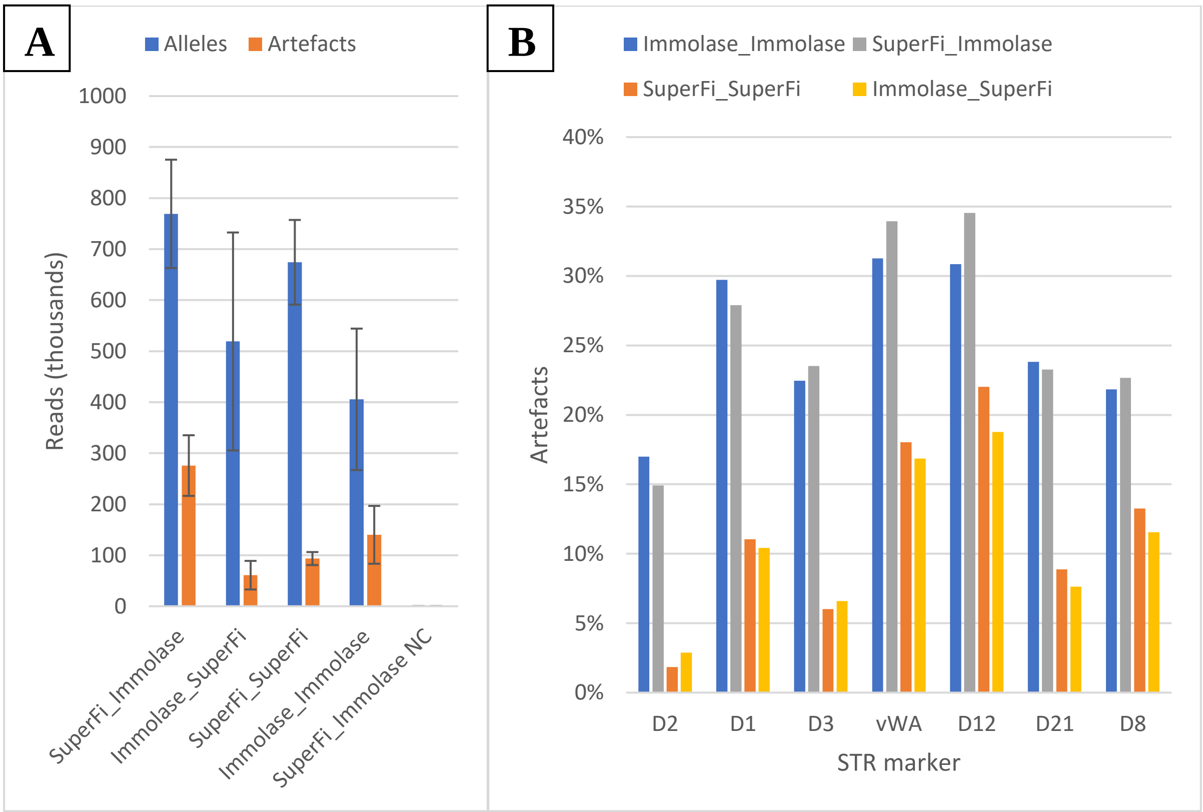


Figure 5: Levels of STR alleles and artefacts generated by the different DNA polymerase combinations. A) Average number of reads for correct STR alleles and artefacts. B) The fraction of artefacts (stutters and other errors) generated per STR marker out of the total number of reads per marker. The data was compiled using ToaSTR with an analytical threshold of 100 reads for allele identification. n=2.

4.3 Evaluation of barcoding using SiMSen-Seq

One of the main features of SiMSen-Seq is that each forward primer used in PCR1 contains a unique sequence serving as a barcode, also called a unique molecular identifier (UMI). The PCR products generated in PCR1 will therefore have a unique barcode, and the data can be filtered and sorted into so-called "barcode families", which in this case means that each unique barcode family must contain at least three identical sequence reads with the same barcode. This grouping process was done at analysis level 4 (see Figure 1) using a program known as "UMIErrorCorrect" (unpublished at time of writing), and an overview of the results are presented in Figure 6 below, compiled using FDSTools. The use of SuperFi in PCR2 resulted in a decrease in the number of artefact barcode families and a reduced amount of correct allele barcode families.

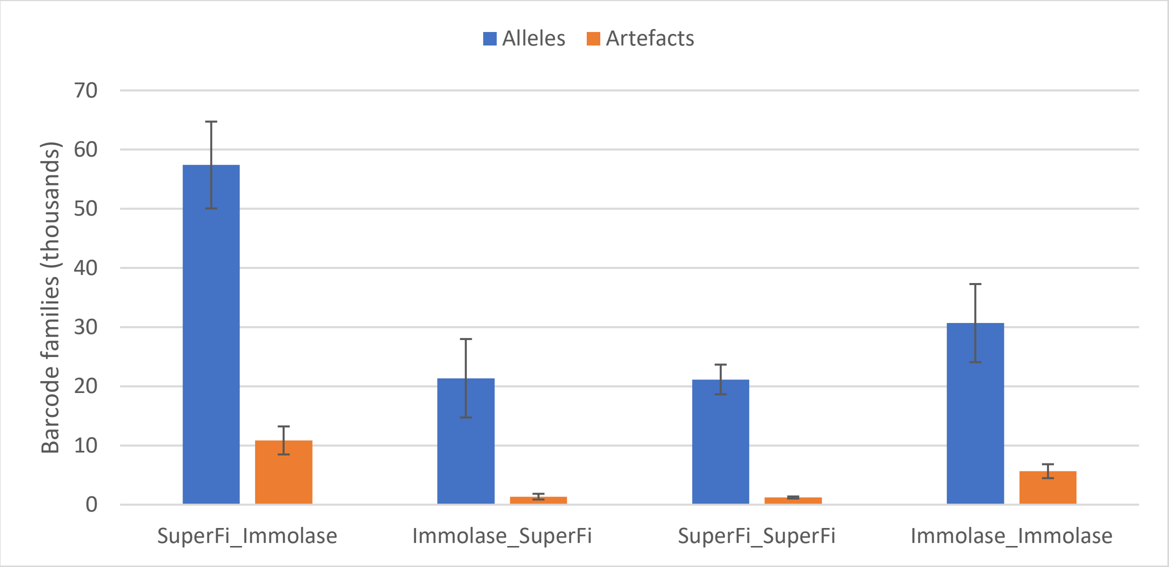


Figure 6: Overview of allele barcode families and artefact barcode families after UMIErrorCorrect shown in thousands. The program FDSTools was used to compile the data. An analytical threshold of 1 read was set for identifying allele and artefact barcode families. Each barcode family represents a minimum of three identically barcoded reads. n=2.

The use of barcodes in SiMSen-Seq allows NGS data to be filtered, where erroneous STR allele sequence variants (e.g. stutters caused by the DNA polymerase) as well as barcoded sequences that are underrepresented (< 3 reads with identical barcode) are removed at analysis level 4 (see Figure 1). An example is shown below in Figure 7, which demonstrates the removal of artefacts for STR marker D1 (2800M alleles 12; 13) using the DNA polymerase combination SuperFi/SuperFi. In A) the total number of different allele variants detected above the analytical threshold was 7, but after running the same data through UMIErrorCorrect in B) only the two correct alleles remained, removing the 5 artefacts previously present. For instance, one of the correct alleles is CE12_AC[6]CTAT[11], which has the stutter product seen in A) as CE11_AC[6]CTAT[10], that is later removed by UMIErrorCorrect in B). This filtration is further detailed in Table 13 which compares the effect of using SuperFi in PCR1 and PCR2, to that of using Immolase in both PCR steps, before and after using UMIErrorCorrect. As seen, the use of SuperFi in both PCR1&2 greatly improves the effect of artefact filtering for most STR markers, where in five of the markers (D2, D1, D3, D21, and D8) only the two correct alleles remain after UMIErrorCorrect, applying a threshold of 50 barcode families. The high analytical threshold was set to allow for a more simplified comparison between the different DNA polymerase combinations. Several identical sequence variants were common between the polymerase combinations, however, in varying quantities, and elevating the analytical threshold to 50 barcode families was sufficient for removing such artefacts generated by SuperFi/SuperFi, whereas these often remained for Immolase/Immolase after UMIErrorCorrect (see Appendix 1). In Appendix 1, different stutters (-2,-1,+1, and +2) of the correct STR alleles are shown, determined at a different analytical threshold (0.5% of highest allele per STR marker and a minimum of 5 barcode families). The -1 and +1 stutters appear more frequently than the -2 and +2 stutters as seen in Appendix 1.

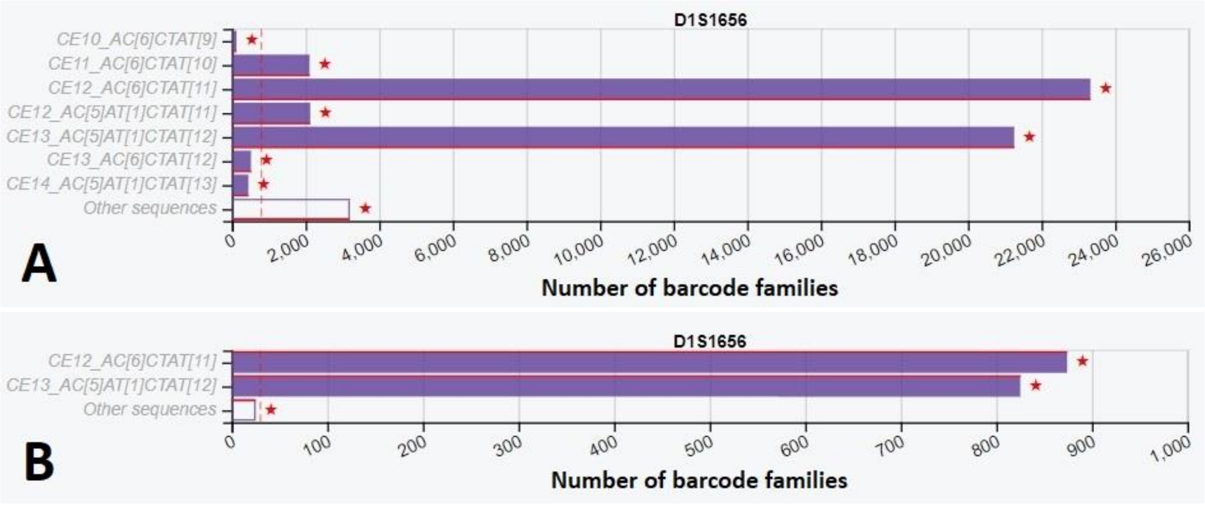


Figure 7: Identified alleles for STR marker D1 before (A) and after (B) running data through UMIErrorCorrect, with an analytical threshold of 50 reads before UMIErrorCorrect and 50 barcode families after UMIErrorCorrect. A) A total of 7 alleles was detected above the set threshold, whereas in B) only the two correct alleles remain after running the same data through UMIErrorCorrect. The combination SuperFi/SuperFi was used and the image was retrieved from the software FDSTools.

Table 13: Comparison of generated artefacts before and after using UMIErrorCorrect (UMI_EC) for the combinations SuperFi/SuperFi and Immolase/Immolase. The data was compiled using FDSTools and the analytical threshold was set at 50 reads before and 50 barcode families after UMIErrorCorrect. The number of detected artefacts above the analytical threshold is shown. n=2.

	Average number of artefacts per STR marker						
Polymerase combination	**D2**	**D1**	**D3**	**vWA**	**D12**	**D21**	**D8**
SuperFi/SuperFi Before UMI_EC	3	5	4.5	11.5	13	7	6
SuperFi/SuperFi After UMI_EC	0	0	0	1	3.5	0	0
Immolase/Immolase Before UMI_EC	6	13	11	19.5	19	19	10
Immolase/Immolase After UMI_EC	1	2	3	1	1	1	2

To demonstrate the use of barcoding in SiMSen-Seq for complex DNA samples, the analysis of mixtures with DNA from two individuals at a ratio of 9:1 is presented in Figure 8 below, using the DNA polymerase combination SuperFi/Immolase. As expected, there is a higher percentage of barcode families for person 1 compared to the more diluted person 2 DNA sample. The level of artefact barcode families for STR markers D2, D12, and D8, is however high, nearly reaching the level of true allele barcode families for person 2. The five STR markers D2, D1, D12, D21, and D8, were chosen for analyses as both individuals possess different alleles of these.

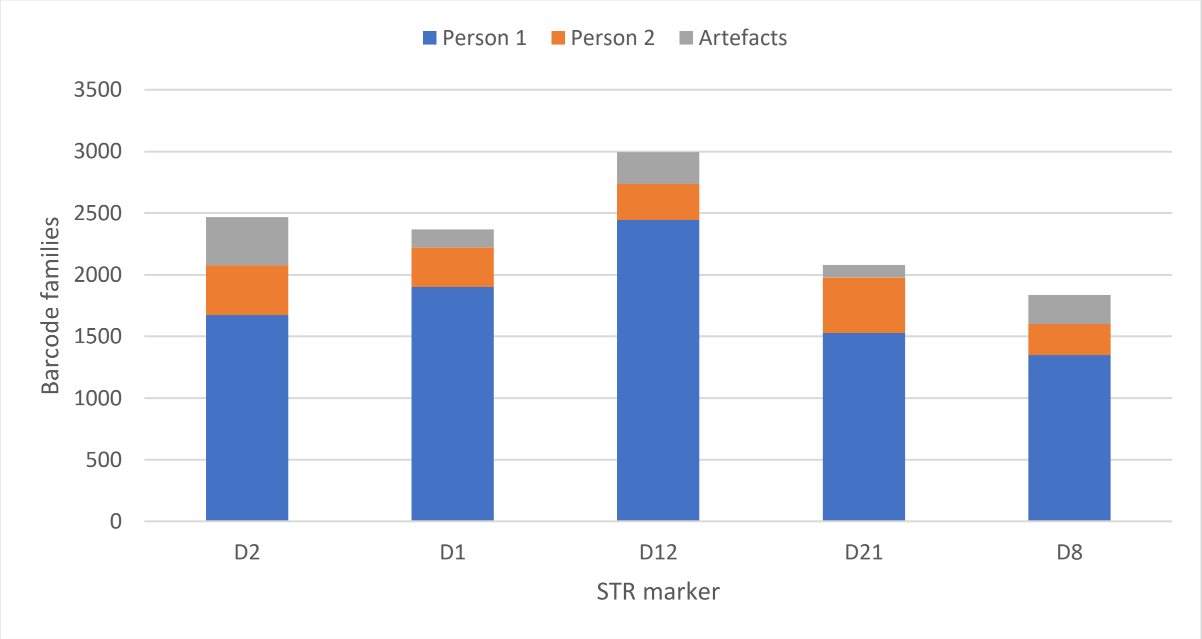


Figure 8: Barcode family count of a mixed sample containing the DNA from 2 individuals, person 1 and person 2, after filtration through UMIErrorCorrect using an analytical threshold of 1 barcode family. Before amplification by PCR, the concentration of DNA used from person 1 and person 2 was at a 9:1 ratio. Represented in blue is the number of correct allele barcode families for person 1 and in orange the correct allele barcode families for person 2. The artefacts for the 5 markers, including stutters and PCR/sequencing errors, is shown in grey.

5 Discussion

The focus of this project was to optimize the SiMSen-Seq method, originally developed for cancer diagnostics (Ståhlberg et al., 2016), towards its use in forensics for STR-MPS. The protocol already in development for this purpose was further altered with the goal of generating more specific products and less nonspecific products, evaluated by applying Bioanalyzer 2100. Throughout this optimization process, from testing different magnesium concentrations to thermal cycle programs, the results showed that the DNA polymerase and the primer concentration had the greatest effect towards maximizing correct products and minimizing nonspecific products during library preparation. Two DNA polymerases, SuperFi and Immolase, both of which performed well in generating specific STRs were further evaluated using MiSeq, allowing a deeper analysis of the sequencing data from the four possible PCR1 and PCR2 enzyme combinations used in SiMSen-Seq library preparation.

5.1 Polymerase effect on SiMSen-Seq library preparation

A range of different polymerases were tested in various combinations for evaluating the effect of generating specific STR products during SiMSen-Seq library preparation. It was clear that using different polymerases had a large impact on generating correct products. Out of the 48 combinations tested, only 10 allowed for the detection of most or all STR markers (6-7 of 7).

The combination SuperFi/Immolase, with 60 nM barcode primer concentration for all markers except vWA at 40 nM, generated the highest concentration of products within the correct product region. Elevating the primer concentration for this combination led to amplification of more specific products within the range of 240-405 base pairs, with an almost threefold increase. This is also a fourfold increase of products compared to the initial protocol where Immolase were applied in both PCR1 and PCR2. Although increased primer concentration was tested for other polymerase combinations, the largest positive effect was seen for SuperFi/Immolase.

For the optimization of the protocol, results were evaluated using Bioanalyzer which has its limitations in that only fragment length and concentration is obtained. The NGS data provided a deeper look into how the use of two enzymes, SuperFi and Immolase, in different combinations during library preparation, affected the generation of STR product artefacts, such as stutters, which are common PCR artefacts generated from STR amplification (Walsh et al., 1996). Although the DNA polymerase combination SuperFi/Immolase resulted in the highest number of STR product reads, it suffered from more STR-product artefacts. The fact that the choice of polymerase used in MPS library preparation influences sequence quality has also been documented previously (Brandariz-Fontes et al., 2015).

Interestingly, combinations having SuperFi in PCR2 of library preparation, although generating less STR product reads, resulted in higher quality reads due to a lower percentage of stutters and other artefacts, as seen for all 7 STR markers. It has been reported that certain polymerases can decrease stutter formation, and it has been theorized that possible mechanisms may include enzyme processivity, proofreading, or an increased binding of the polymerase to the DNA by a binding domain such as Sso7d protein, decreasing rate of dissociation (Fazekas et al., 2010). From the findings in this project however, it is evident that using SuperFi, an enzyme with proofreading ability, as opposed to Immolase, a non-proofreading enzyme, generates fewer stutter products, although more research is needed to explain why this is the case.

5.2 Barcoding effect on data analyses using SiMSen-Seq

One of the key features of SiMSen-Seq is that NGS data can be filtered and sorted based on barcodes, allowing the compression of relevant data and the removal of noise (Ståhlberg et al., 2016). For example, the combination Immolase/SuperFi had a total read count of roughly 1.7 million, which after filtration through UMIErrorCorrect, was sorted and compressed to just over 22 thousand barcode families. One example of how barcoding combined with UMIErrorCorrect can filtrate NGS data and remove artefacts is shown in Figure 7. Before UMIErrorCorrect, a total of 7 alleles (5 artefacts) were detected above the analytical threshold, whereas after UMIErrorCorrect, all 5 artefacts were removed from analyses leaving only the correct two D1 STR 2800M alleles. This filtration was further compared for combinations SuperFi/SuperFi and Immolase/Immolase, which showed that SuperFi/SuperFi generated less STR artefacts before UMIErrorCorrect, and the software could remove all artefacts for 5 out of the 7 STR markers. This was not true however for Immolase/Immolase, where one or more artefacts consistently remained for all STR markers after UMIErrorCorrect. Common artefacts such as stutter products arise from the insertion or deletion of one repeat motif due to polymerase errors (Brookes et al., 2012, Hauge and Litt, 1993), and its occurrence has been linked to the longest uninterrupted stretch (LUS) of an allele, meaning that the longer a motif is consecutively repeated it will have a higher chance of stutter formation (Vilsen et al., 2018, Walsh et al., 1996). Such stutter products of the LUS are also seen for most of the 7 STRs sequenced in this project, further detailed in Appendix 1. Notably, the amount of artefacts generated was higher for DNA polymerase combinations utilizing the non-proofreading enzyme Immolase in PCR2, which is also the library preparation step that includes the most thermal cycles (33 compared to 3 in PCR1). Using SuperFi in PCR2, a proofreading enzyme with high-fidelity (>300x better fidelity compared to *Taq* (Invitrogen, 2019)), results in less artefacts, suggesting that proofreading and high-fidelity are favorable qualities to include for STR-MPS by SiMSen-Seq.

The use of SiMSen-Seq for mixed identity samples (Figure 8) showed that two individuals, person 1 and person 2, represented at different DNA ratios (9:1), could both be identified and separated from the background noise (artefacts). However, for STR markers D2, D12, and D8, the amount of artefact barcode families reached the same or nearly the same amount as correct allele barcode families for person 2 (10% of the total DNA). Among these STR markers, D12 also resulted in the highest sequence read artefacts percentage when comparing the four combinations of SuperFi and Immolase (Figure 5). The sequence read artefact percentage for STR marker D2 was however low, and a correlation could therefore not be drawn between these two cases.

Mixed identity samples in forensic casework can be tricky to interpret, especially using CE, as sequence variants having the same length cannot be discriminated from each other. MPS however enables these sequence variants to be distinguished, also revealing different errors that can arise from PCR or sequencing, such as stutters and base-pair substitutions (de Knijff, 2019). Interpretation issues can then arise when it is not known how many individuals contributed to the DNA sample and at what ratios, as true alleles could be misinterpreted as artefacts. Although MPS provides more information than CE does, there is still work to be done before transitioning to STR-MPS in routine practice, such as developing more efficient data analyses solutions (Borsting and Morling, 2015) and setting recommendations on analytical thresholds for calling true STR alleles (de Knijff, 2019), both of which were issues faced during this project.

5.3 Ideas on the further development of SiMSen-Seq towards forensics

The use of SiMSen-Seq in forensics, based on these results, is very promising. All correct alleles for the 7 STR markers amplified from 2800M control DNA were identified using MPS and distinguishing two individuals in a mixed identity sample was possible. However, for future analyses of mixed identity samples using SiMSen-Seq, it is suggested to further evaluate a DNA polymerase combination utilizing SuperFi in both PCR1 and PCR2 in MPS experiments, as SuperFi resulted in the generation of less STR artefacts such as stutters and may therefore translate to a more sensitive analysis. It would also be interesting to evaluate other proofreading and high-fidelity DNA polymerases in PCR2, such as Phusion. It is also necessary to study the effect of using low DNA concentrations as well as analyzing more complex samples, e.g. 3-4 person mixtures. Additionally, testing with additives and adding contaminants (e.g. hemoglobin and humic substances) to evaluate system tolerance is also a future step, as forensic samples often contain such inhibitory substances that can cause PCR complications (Sidstedt et al., 2020). Further customization of the bioinformatic data analysis workflow is also required to streamline the analysis process and use the barcoding to full extent, such that in the future, the SiMSen-Seq method can be used for analysis of real crime scene samples.

6 Conclusions

This project can be divided into two main parts, 1) the optimization of the SiMSen-Seq library preparation for efficient amplification of STRs, and 2) massively parallel sequencing and bioinformatic data analyses using the optimized 7 plex STR assay. The results showed that the type of DNA polymerase and the barcode primer concentration had the greatest impact on STR product formation. Although using the DNA polymerase combination SuperFi/Immolase resulted in the highest concentration of desired products out of all tested combinations, as well as the most STR product reads of out the four sequenced combinations, it suffered from high amounts of STR sequence artefacts, complicating DNA profile interpretation in a mixed sample.

The use of barcodes allowed for a reduction of artefacts generated from DNA polymerase errors, simplifying data interpretation. However, for future testing of mixed samples by SiMSen-Seq, it would be interesting to further evaluate SuperFi in both PCR1 and PCR2 of library preparation, and to test other proofreading DNA polymerases, as this could lead to higher quality sequences with fewer STR artefacts such as stutters. Additionally, more testing is required in studying the effect of low DNA concentrations and PCR inhibitors, as well as further customization of the bioinformatic workflow, towards the development of a more sensitive STR-MPS method.

7 References

ABU AL-SOUD, W. & RÂDSTRÖM, P. 1998. Capacity of nine thermostable DNA polymerases to mediate DNA amplification in the presence of PCR-inhibiting samples. *Applied and Environmental Microbiology,* 64**,** 3748-3753.

AGILENT TECHNOLOGIES. 2017. *Agilent High Sensitivity DNA Kit Quick Start Guide* [Online]. Available: https://www.agilent.com/cs/library/usermanuals/public/G2938-90322_HighSensitivityDNAKit_QSG.pdf [Accessed 18/05/2020].

ALONSO, A., BARRIO, P. A., MULLER, P., KOCHER, S., BERGER, B., MARTIN, P., BODNER, M., WILLUWEIT, S., PARSON, W., ROEWER, L. & BUDOWLE, B. 2018. Current state-of-art of STR sequencing in forensic genetics. *Electrophoresis,* 39**,** 2655-2668.

ALONSO, A., MÜLLER, P., ROEWER, L., WILLUWEIT, S., BUDOWLE, B. & PARSON, W. 2017. European survey on forensic applications of massively parallel sequencing. *Forensic Science International: Genetics,* 29**,** e23-e25.

APPLIED BIOSYSTEMS. 2014. *AmpliTaq Gold® DNA Polymerase Protocol* [Online]. Available: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2Famplitaq_gold_dna_polymerase_man.pdf&title=QW1wbGlUYXEgR29sZCBETkEgUG9seW1lcmFzZQ== [Accessed 20/05/2020].

BORSTING, C. & MORLING, N. 2015. Next generation sequencing and its applications in forensic genetics. *Forensic Science International: Genetics,* 18**,** 78-89.

BRANDARIZ-FONTES, C., CAMACHO-SANCHEZ, M., VILÀ, C., VEGA-PLA, J. L., RICO, C. & LEONARD, J. A. 2015. Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. *Scientific Reports,* 5**,** 8056.

BROOKES, C., BRIGHT, J.-A., HARBISON, S. & BUCKLETON, J. 2012. Characterising stutter in forensic STR multiplexes. *Forensic Science International: Genetics,* 6**,** 58-63.

BRUIJNS, B., TIGGELAAR, R. & GARDENIERS, H. 2018. Massively parallel sequencing techniques for forensics: A review. *Electrophoresis*, 39**,** 2642-2654.

CHIEN, A., EDGAR, D. B. & TRELA, J. M. 1976. Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *Journal of Bacteriology,* 127**,** 1550-1557.

COTTON, E. A., ALLSOP, R. F., GUEST, J. L., FRAZIER, R. R. E., KOUMI, P., CALLOW, I. P., SEAGER, A. & SPARKES, R. L. 2000. Validation of the AMPFlSTR® SGM Plus™ system for use in forensic casework. *Forensic Science International,* 112**,** 151-161.

DE KNIJFF, P. 2019. From next generation sequencing to now generation sequencing in forensics. *Forensic Science International: Genetics,* 38**,** 175-180.

FAZEKAS, A. J., STEEVES, R. & NEWMASTER, S. G. 2010. Improving sequencing quality from PCR products containing long mononucleotide repeats. *BioTechniques,* 48**,** 277-285.

FILGES, S., YAMADA, E., STÅHLBERG, A. & GODFREY, T. E. 2019. Impact of Polymerase Fidelity on Background Error Rates in Next-Generation Sequencing with Unique Molecular Identifiers/Barcodes. *Scientific Reports,* 9**,** 3503-3503.

FOX, E. J., REID-BAYLISS, K. S., EMOND, M. J. & LOEB, L. A. 2014. Accuracy of Next Generation Sequencing Platforms. *Next Generation, Sequencing & Applications,* 1.

GANSCHOW, S., SILVERY, J., KALINOWSKI, J. & TIEMANN, C. 2018. toaSTR: A web application for forensic STR genotyping by massively parallel sequencing. *Forensic Science International: Genetics,* 37**,** 21-28.

GARIBYAN, L. & AVASHIA, N. 2013. Research Techniques Made Simple: Polymerase Chain Reaction (PCR). *The Journal of Investigative Dermatology,* 133**,** 1-4.

GILL, P. 2002. Role of Short Tandem Repeat DNA in Forensic Casework in the UK—Past, Present, and Future Perspectives. *BioTechniques,* 32**,** 366-385.

GREENE, C. S., TAN, J., UNG, M., MOORE, J. H. & CHENG, C. 2014. Big Data Bioinformatics. *Journal of Cellular Physiology,* 229**,** 1896-1900.

GROMADSKI, K., SALOWSKY, R. & GLUECK, S. 2016. Improving sample quality for target enrichment and next-gen sequencing with the Agilent High Sensitivity DNA Kit and the Agilent SureSelect Target Enrichment Platform. Waldbronn, Germany: Agilent Technologies, Inc.

HAUGE, X. Y. & LITT, M. 1993. A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Human Molecular Genetics,* 2**,** 411-415.

HEDMAN, J., NORDGAARD, A., RASMUSSON, B., ANSELL, R. & RÅDSTRÖM, P. 2009. Improved forensic DNA analysis through the use of alternative DNA polymerases and statistical modeling of DNA profiles. *BioTechniques,* 47**,** 951-958.

HOOGENBOOM, J., VAN DER GAAG, K. J., DE LEEUW, R. H., SIJEN, T., DE KNIJFF, P. & LAROS, J. F. J. 2017. 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 International: Genetics,* 27**,** 27-40.

INVITROGEN. 2019. *Platinum™ SuperFi™ II DNA Polymerase User Guide* [Online]. Available: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0018859_Platinum_SuperFi_II_DNA_Pol_UG.pdf [Accessed 20/05/2020].

ISHINO, S. & ISHINO, Y. 2014. DNA polymerases as useful reagents for biotechnology - the history of developmental research in the field. *Frontiers in Microbiology,* 5**,** 465.

JÄGER, A. C., ALVAREZ, M. L., DAVIS, C. P., GUZMÁN, E., HAN, Y., WAY, L., WALICHIEWICZ, P., SILVA, D., PHAM, N., CAVES, G., BRUAND, J., SCHLESINGER, F., POND, S. J. K., VARLARO, J., STEPHENS, K. M. & HOLT, C. L. 2017. Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. *Forensic Science International: Genetics,* 28**,** 52-70.

KAPA BIOSYSTEMS. 2019. *KAPA HiFi HotStart ReadyMix PCR Kit Technical Data Sheet* [Online]. Available: https://rochesequencingstore.com/wp-content/uploads/2017/10/KAPA-HiFi-HotStart-ReadyMix-PCR-Kit_KR0370-%E2%80%93-v10.19.pdf [Accessed 20/05/2020].

LIU, Y.-Y. & HARBISON, S. 2018. A review of bioinformatic methods for forensic DNA analyses. *Forensic Science International: Genetics,* 33**,** 117-128.

LORENZ, T. C. 2012. Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *Journal of Visualized Experiments : JoVE***,** e3998.

MULLIS, K. B. & FALOONA, F. A. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in Enzymology,* 155**,** 335-350.

NEW ENGLAND BIOLABS. n.d. *Q5® Hot Start High-Fidelity DNA Polymerase* [Online]. Available: https://international.neb.com/products/m0493-q5-hot-start-high-fidelity-dna-polymerase#Product%20Information [Accessed 20/05/20].

NOVAGEN. n.d. *KOD Xtreme™ Hot Start DNA Polymerase User Protocol* [Online]. Available: https://www.merckmillipore.com/SE/en/product/KOD-Xtreme-Hot-Start-DNA-Polymerase,EMD_BIO-71975#anchor_USP [Accessed 20/05/2020].

QUANTABIO. n.d. *AccuStart™ Taq DNA Polymerase HiFi* [Online]. Available: https://www.quantabio.com/media/contenttype/IFU-061.1_REV_02_95085_AccuStart_Taq_DNA_polymerase_HiFi.pdf [Accessed 20/05/2020].

RUITBERG, C. M., REEDER, D. J. & BUTLER, J. M. 2001. STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Research,* 29**,** 320-322.

RYCHLIK, W., SPENCER, W. J. & RHOADS, R. E. 1990. Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Research,* 18**,** 6409-6412.

SAIKI, R. K., GELFAND, D. H., STOFFEL, S., SCHARF, S. J., HIGUCHI, R., HORN, G. T., MULLIS, K. B. & ERLICH, H. A. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science,* 239**,** 487-491.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 5463-5467.

SCHIRMER, M., IJAZ, U. Z., D'AMORE, R., HALL, N., SLOAN, W. T. & QUINCE, C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research,* 43**,** e37-e37.

SEIDLITZ, H., HEDMAN, J. & SIDSTEDT, M. 2019. Nationellt Forensiskt Centrum, Notat: Biologisektionen 2019:24.

SIDSTEDT, M., JANSSON, L., NILSSON, E., NOPPA, L., FORSMAN, M., RÅDSTRÖM, P. & HEDMAN, J. 2015. Humic substances cause fluorescence inhibition in real-time polymerase chain reaction. *Analytical Biochemistry,* 487**,** 30-37.

SIDSTEDT, M., RÅDSTRÖM, P. & HEDMAN, J. 2020. PCR inhibition in qPCR, dPCR and MPS—mechanisms and solutions. *Analytical and Bioanalytical Chemistry,* 412**,** 2009-2023.

STÅHLBERG, A., KRZYZANOWSKI, P. M., EGYUD, M., FILGES, S., STEIN, L. & GODFREY, T. E. 2017. Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. *Nature Protocols,* 12**,** 664-682.

STÅHLBERG, A., KRZYZANOWSKI, P. M., JACKSON, J. B., EGYUD, M., STEIN, L. & GODFREY, T. E. 2016. Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. *Nucleic Acids Research,* 44**,** e105-e105.

THERMO FISHER SCIENTIFIC. 2018. *Phusion Hot Start II High-Fidelity DNA Polymerase Product Information* [Online]. Available: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0012401_Phusion_HotStartII_DNAPolymerase_500U_UG.pdf [Accessed 20/05/2020].

VILSEN, S. B., TVEDEBRINK, T., ERIKSEN, P. S., BØSTING, C., HUSSING, C., MOGENSEN, H. S. & MORLING, N. 2018. Stutter analysis of complex STR MPS data. *Forensic Science International: Genetics,* 35**,** 107-112.

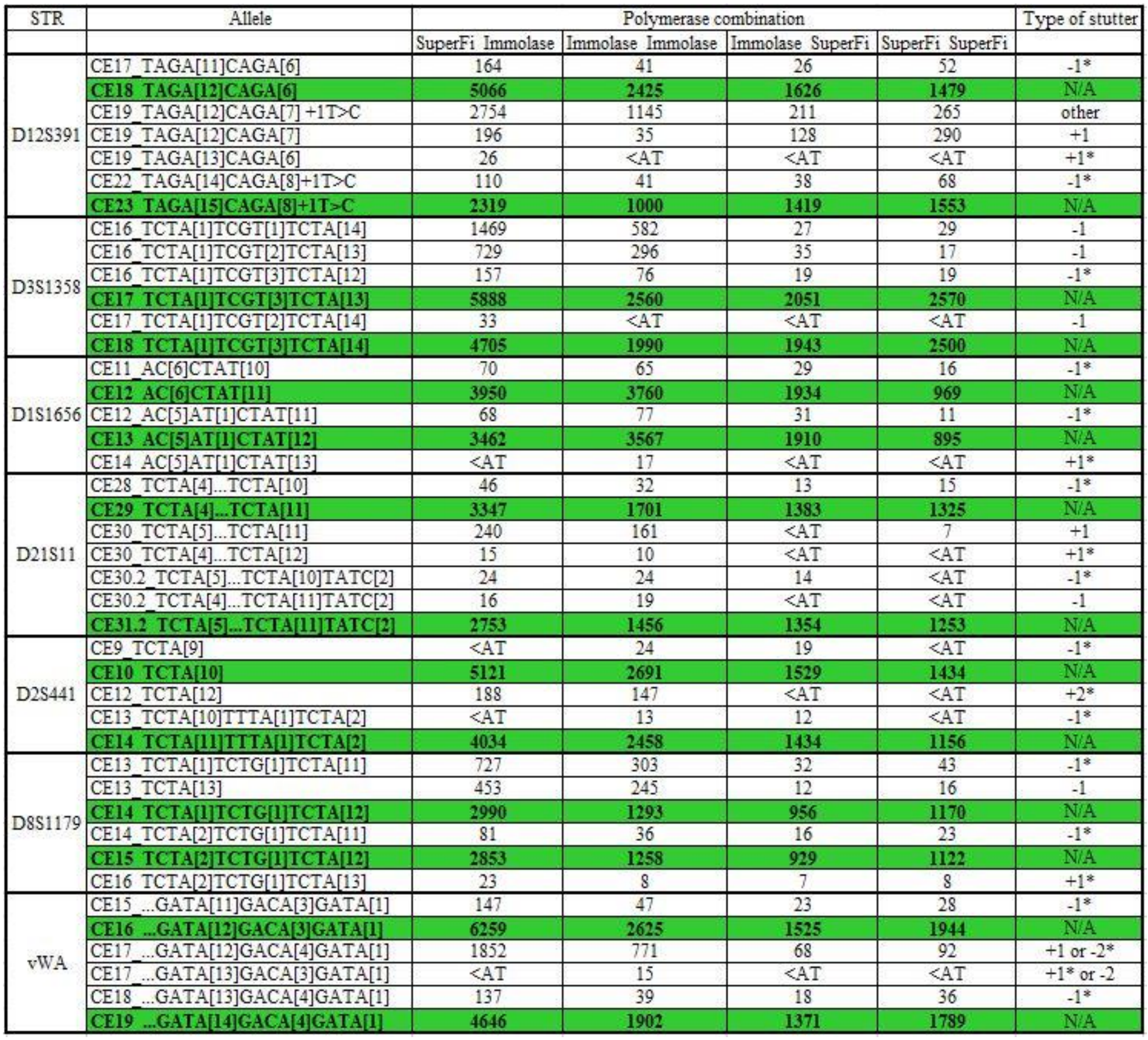
WALSH, P. S., FILDES, N. J. & REYNOLDS, R. 1996. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. *Nucleic Acids Research,* 24**,** 2807-2812.

WANG, Y., PROSEN, D. E., MEI, L., SULLIVAN, J. C., FINNEY, M. & VANDER HORN, P. B. 2004. A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro. *Nucleic Acids Research,* 32**,** 1197-1207.

ZHUANG, Z. & AI, Y. 2010. Processivity factor of DNA polymerase and its expanding role in normal and translesion DNA synthesis. *Biochimica et Biophysica Acta,* 1804**,** 1081-1093.

Appendix 1

Data comparing the generation of correct alleles (highlighted in green) and artefacts for the 7 amplified STRs (D12, D3, D1, D21, D2, D8, and vWA) by the four different polymerase combinations SuperFi/Immolase, Immolase/Immolase, Immolase/SuperFi, and SuperFi/SuperFi after UMIErrorCorrect. In the column: "Type of stutter", -2, -1, +1, and +2 repeat stutters of the correct STR alleles are shown. Stutters classified as "other" could not be categorized. Correct alleles are shown as N/A, meaning not applicable. The analytical threshold is 0.5% of the highest allele per STR marker and a minimum of 5 barcode families. Sequence variants below the analytical threshold are denoted as <AT.



STR	Allele	Polymerase combination				Type of stutter
		SuperFi Immolase	Immolase Immolase	Immolase SuperFi	SuperFi SuperFi	
D12S391	CE17_TAGA[11]CAGA[6]	164	41	26	52	-1*
	CE18 TAGA[12]CAGA[6]	**5066**	**2425**	**1626**	**1479**	N/A
	CE19_TAGA[12]CAGA[7] +1T>C	2754	1145	211	265	other
	CE19_TAGA[12]CAGA[7]	196	35	128	290	+1
	CE19_TAGA[13]CAGA[6]	26	<AT	<AT	<AT	+1*
	CE22_TAGA[14]CAGA[8]+1T>C	110	41	38	68	-1*
	CE23 TAGA[15]CAGA[8]+1T>C	**2319**	**1000**	**1419**	**1553**	N/A
D3S1358	CE16_TCTA[1]TCGT[1]TCTA[14]	1469	582	27	29	-1
	CE16_TCTA[1]TCGT[2]TCTA[13]	729	296	35	17	-1
	CE16_TCTA[1]TCGT[3]TCTA[12]	157	76	19	19	-1*
	CE17 TCTA[1]TCGT[3]TCTA[13]	**5888**	**2560**	**2051**	**2570**	N/A
	CE17_TCTA[1]TCGT[2]TCTA[14]	33	<AT	<AT	<AT	-1
	CE18 TCTA[1]TCGT[3]TCTA[14]	**4705**	**1990**	**1943**	**2500**	N/A
D1S1656	CE11_AC[6]CTAT[10]	70	65	29	16	-1*
	CE12 AC[6]CTAT[11]	**3950**	**3760**	**1934**	**969**	N/A
	CE12_AC[5]AT[1]CTAT[11]	68	77	31	11	-1*
	CE13 AC[5]AT[1]CTAT[12]	**3462**	**3567**	**1910**	**895**	N/A
	CE14_AC[5]AT[1]CTAT[13]	<AT	17	<AT	<AT	+1*
D21S11	CE28_TCTA[4]...TCTA[10]	46	32	13	15	-1*
	CE29 TCTA[4]...TCTA[11]	**3347**	**1701**	**1383**	**1325**	N/A
	CE30_TCTA[5]...TCTA[11]	240	161	<AT	7	+1
	CE30_TCTA[4]...TCTA[12]	15	10	<AT	<AT	+1*
	CE30.2_TCTA[5]...TCTA[10]TATC[2]	24	24	14	<AT	-1*
	CE30.2_TCTA[4]...TCTA[11]TATC[2]	16	19	<AT	<AT	-1
	CE31.2 TCTA[5]...TCTA[11]TATC[2]	**2753**	**1456**	**1354**	**1253**	N/A
D2S441	CE9_TCTA[9]	<AT	24	19	<AT	-1*
	CE10 TCTA[10]	**5121**	**2691**	**1529**	**1434**	N/A
	CE12_TCTA[12]	188	147	<AT	<AT	+2*
	CE13_TCTA[10]TTTA[1]TCTA[2]	<AT	13	12	<AT	-1*
	CE14 TCTA[11]TTTA[1]TCTA[2]	**4034**	**2458**	**1434**	**1156**	N/A
D8S1179	CE13_TCTA[1]TCTG[1]TCTA[11]	727	303	32	43	-1*
	CE13_TCTA[13]	453	245	12	16	-1
	CE14 TCTA[1]TCTG[1]TCTA[12]	**2990**	**1293**	**956**	**1170**	N/A
	CE14_TCTA[2]TCTG[1]TCTA[11]	81	36	16	23	-1*
	CE15 TCTA[2]TCTG[1]TCTA[12]	**2853**	**1258**	**929**	**1122**	N/A
	CE16_TCTA[2]TCTG[1]TCTA[13]	23	8	7	8	+1*
vWA	CE15_...GATA[11]GACA[3]GATA[1]	147	47	23	28	-1*
	CE16 ...GATA[12]GACA[3]GATA[1]	**6259**	**2625**	**1525**	**1944**	N/A
	CE17_...GATA[12]GACA[4]GATA[1]	1852	771	68	92	+1 or -2*
	CE17_...GATA[13]GACA[3]GATA[1]	<AT	15	<AT	<AT	+1* or -2
	CE18_...GATA[13]GACA[4]GATA[1]	137	39	18	36	-1*
	CE19 ...GATA[14]GACA[4]GATA[1]	**4646**	**1902**	**1371**	**1789**	N/A

*LUS stutter