An Evaluation of Massively Parallel Sequencing for Forensic Applications

Matthew Phipps

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy at
the University of Central Lancashire.

November 2021

STUDENT DECLARATION FORM

Type of Award: **Doctor of Philosophy**

School: **Forensic and Applied Sciences**

1. Concurrent registration for two or more academic awards

I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution.

2. Material submitted for another award

I declare that no material contained in the thesis has been used in any other submission for an academic award and is solely my own work.

3. Collaboration

Where a candidate's research programme is part of a collaborative project, the thesis must indicate in addition clearly the candidate's individual contribution and the extent of the collaboration. Please state below:

Not part of a collaborative project

4. Use of a Proof-reader

No proof-reading service was used in the compilation of this thesis.

Signature of Candidate: 

Print name: **Matthew Phipps**

Table of Contents

I. List of Figures

II. List of Tables

III. Abstract

Massively parallel sequencing, or MPS, also known as next generation sequencing, is a technique that has gained much recent attention in forensic literature. This work explores whether MPS offers practical benefits to forensic laboratories beyond those which can be achieved with existing capillary electrophoresis (CE) based methods. Specifically, this work has focussed on the use of commercially available Ion Torrent semi-conductor MPS sequencing and the kits and chemistries available for that platform.

The sensitivity, resistance to inhibition, performance on non-probative casework samples, and ability to accurately predict ancestry with MPS have all been examined. Further work has explored the ability of MPS methods to detect mixtures and to add useful information to kinship cases. Lastly, an evaluation of the analysis parameters that would be necessary for practical implementation of MPS in the forensic laboratory has been performed.

Results show a noticeable increase in sensitivity and performance with degraded DNA for MPS, something that was also shown to be a practical advantage in the analysis of the non-probative casework samples. Performance of inhibited samples and mixtures with MPS was less good, but this does not hinder the net gain in overall practical terms, with a conclusion that MPS methods do not replace CE, but offer much as a complement to CE methods for certain sample types.

Similar conclusions were drawn from the kinship analysis, with some kinship scenarios, such as sibling and grandparent cases, having their match statistics usefully increased by the addition of MPS data to the CE data, while other case types, such as cousins and great-grandparents, were not able to be resolved with either CE or the MPS panel used in this work. Results of the ancestry prediction by MPS showed promise, with 45 of 64 samples having a predicted ancestry that matched the donor's self-declared ancestry.

Investigation of the analysis parameters that are necessary for the implementation of MPS in practice uncovered the significant result that these parameters can vary based on the specifics of the run being performed, particularly with regard to the chip type used and the number of samples analysed. Carefully managed though, this does not affect the conclusion that MPS offers much of promise to practicing forensic DNA laboratories when used on certain case and sample types in conjunction with current CE-based technologies.

Chapter 1:

Introduction and Aims

1. Introduction and Aims

1.1. Introduction

The advantage that DNA analysis can offer forensic practitioners is well established (Goodwin 2015). Current methods of forensic DNA analysis focus almost entirely on capillary electrophoresis (CE) based methods of short tandem repeat (STR) analysis (Butler 2012, Butler 2015). The limitations of these methods are also well known however (Butler 2015 [2]) and include:

- Limited ability to retrieve useful information from degraded or inhibited samples
- Inability to resolve complex mixed samples
- Interference in analysis of low-level samples from stutter, dye-blobs and other artefacts inherent in CE STR systems
- The inability to make progress in a case where there is no suspect sample and no match to a DNA database

In recent years, a new method of sequencing, massively parallel sequencing (MPS), has promised to make progress on these limitations, and to offer information in cases that was not previously possible (Yang *et al.* 2014).

MPS, also known as next generation sequencing (NGS), is a method of DNA sequencing that has become well established in clinical and genetic research applications in recent years (Rehm 2013, Beadling *et al.* 2013, Millat *et al.* 2014, Tsongalis *et al.* 2014). To a large degree, it has replaced capillary electrophoresis (CE) methods in these applications, in part due to the much larger amount of sequencing data that MPS can produce.

More recently, forensic practitioners have begun to investigate whether the advantages that MPS can offer to other fields can be translated to forensic practice. To date there have been many preliminary publications on this topic (Van Neste 2012, Parson *et al.* 2013, Fordyce *et al.* 2015 Gettings, Kiesler *et al.* 2015), but little practical implementation of MPS technology in forensic casework. This is in part due to the fact that many practical aspects of how MPS will be used by forensic practitioners remain unclear.

For example, several aspects of commonly used MPS protocols have not been fully translated to forensic application from the clinical / research applications mentioned above. This includes knowledge of which MPS assays would offer benefit compared to existing CE-based methods, and lack of clarity on aspects of how data analysis should be performed. These questions need to be addressed for MPS to be successfully used in a forensic context and are examined in this work

1.1.1. Current Forensic DNA Practice

Modern forensic DNA laboratories can carry out a wide range of analyses. This covers many different sample types, which can include samples that come from blood, saliva, skin cells (often called 'touch' DNA), and semen – all deposited on items that range from swabs and drinking vessels to items of clothing, weapons, household objects, and many more (Shewale, 2014). These analyses can also cover multiple different applications, such as crime scene analysis of difficult or degraded samples, databasing of reference samples which contain relatively large amounts of DNA, kinship analysis which requires special consideration of the potential relatedness of sample donors and the corresponding specialised statistics, and disaster victim identification (DVI), which can combine the difficulty of crime scene analysis, with the statistical challenges of kinship analysis (Shewale, 2014).

In 2018 / 2019 over 37,000 samples were added to the UK National DNA Database from crime scenes of this nature, with burglary (47%) and vehicle crime (15%) making the largest proportions of these (National DNA Database Strategy Board Biennial Report 2018-2020). In addition, approximately 30,000 to 50,000 kinship samples, mostly for paternity analysis, are analysed every year in the UK by a variety of different laboratories (Thermo Fisher Scientific, personal communication).

All of these sample types and applications are typically processed today using STR analysis on CE-based technology (Butler 2012, Butler 2015). This technology is well understood, but also has limitations. These limitations include a limited ability to retrieve useful information from degraded or inhibited samples. These samples, especially degraded samples, can make up a large proportion of samples that are encountered in crime scenes. MPS technology promises to aid in the analysis of degraded samples by targeting smaller sections of DNA than the current CE methods, thus allowing profiles to be more readily obtained from damaged, degraded samples (Gettings *et al.* 2015).

A further limitation of current CE analysis in forensic practice is the inability to resolve complex mixed samples (Butler, 2015). This is a common occurrence in forensic analysis with many crime scene samples being reported as being mixed source (Petersen, 2001), making resolution of the mixture into its constituent profiles an important factor in current forensic analysis. MPS technology has been reported to aid interpretation of mixed profiles by revealing the full sequence of STR profiles, which provides extra information that may allow greater resolution of mixtures, and by making new types of analysis such as microhaplotype analysis possible. This is a new type of marker, discussed here in Section 1.1.4.4, which promises to aid in mixture resolution due to its lack of stutter, a factor which

can complicate analysis of mixed STR profiles and prevent full resolution of complex mixed samples.

Another proposed advantage of MPS compared to CE analysis for forensic cases is the lack of dye-blobs and pull-up artefacts that are inherent to CE STR systems. Dye-blobs are excess noise in CE-based profiles caused by unassociated dye-labelled primer forming artefact peaks which hinder analysis of the 'true' peaks formed by the DNA fragments of interest (Butler, 2015). Pull-up is another type of artefact inherent in CE systems, again caused by the nature of the dyes used in the system. In this case, pull-up results from the spectral overlap of the fluorescence emitted by the different dyes that are used in modern PCR multiplexes, and results in artefact peaks in one dye-channel as a result of true signal from another (Butler, 2015). Again, these false artefact peaks hinder analysis of the true result in the profile. MPS promises to remove these artefacts as, in many cases, MPS technology is not based on fluorescent dye technology and the dyes that cause these artefacts in CE analysis are simply not present in the MPS assay.

In many forensic cases there is no suspect sample and no match to a DNA database, which means that even if a clear profile is derived from a crime stain sample, little progress can be made in the case (Kayser, 2015). Here again, MPS has the potential to offer improvements over current CE-based methods, with the expanded amount of data that MPS can process offering the possibility to examine many loci in an assay and in doing so assess the ancestry or phenotype of the sample donor. This may allow an investigative team to make progress in a forensic case with knowledge of the offender's appearance, even without a reference sample to compare the crime stain to.

Lastly, another limitation of CE-based technology encountered by forensic laboratories is the number of loci that can be analysed simultaneously in kinship analysis. Determination of relatedness of samples can require large numbers of loci to be used, something which can be difficult in the limited space available in a CE-based assay, which is typically limited to around 20 to 30 STR markers (Butler, 2015). MPS analysis promises to be able to significantly increase the number of loci that can be analysed in a single assay, something that could be of significant benefit to forensic kinship analysis (Li *et al*. 2017).

This work examines many of these aspects of the use of MPS in forensic practice and attempts to evaluate its utility in aiding forensic practitioners. Specifically, this work focuses on the use of semiconductor MPS using Ion Torrent technology and evaluates the overall performance of this technology compared to CE in crime scene analysis. In addition, this work also investigates the analysis parameters and thresholds that would need to be used

for this type of analysis and evaluates the utility of MPS in forensic evaluation of sample donor ancestry and kinship. Different types of MPS are explored below.

1.1.2. Introduction to MPS

Several methods for performing 'next generation' sequencing (that is, sequencing methods developed after the wide adoption of capillary electrophoresis based Sanger sequencing) exist and although several have been experimented with, to date, none have been widely adopted in routine forensic practice.

Before sequencing by any method, the DNA to be analysed must first be prepared. In some MPS methods this can involve simply sequencing all of the DNA in the starting material. This first requires the DNA to be split, or fragmented, into pieces of appropriate length for sequencing. This can be done either enzymatically or through physical shearing methods, such as sonication (Poptsova *et al.* 2014). This method of breaking DNA in pieces and sequencing the pieces is known as 'shotgun' sequencing (Børsting and Morling 2015) (Figure 1). The drawback of shotgun sequencing for forensic application is that large amounts of input DNA are required (often micrograms, whereas forensic applications routinely deal with DNA samples measured in picograms). Another drawback is that shotgun sequencing is non-selective, i.e. all of the DNA present in the sample is sequenced, which is often not useful to the forensic analyst. The vast majority of DNA in the human genome (estimated at about 99.5%) is the same between different individuals, which makes sequencing of these shared stretches irrelevant to the identification of individual samples (Levy *et al.* 2007).

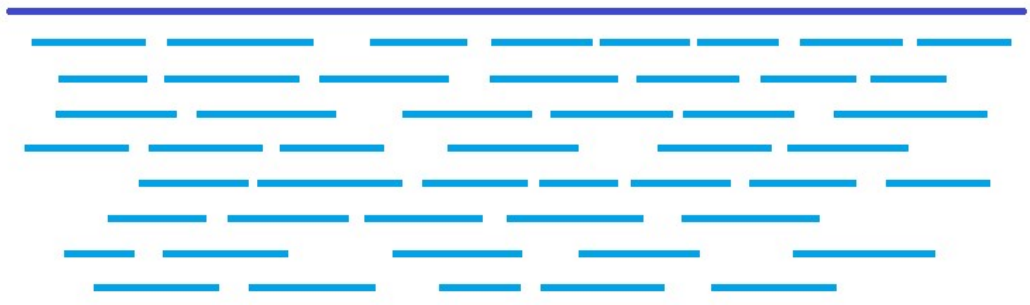


Figure 1: A representation of Shotgun sequencing. The dark blue bar represents the reference stretch of DNA. The light blue bards represent the fragments that are generated in the fragmentation of the DNA. Similarly (but not identically) sized pieces are produced that randomly cover the entire reference sequence.

For these reasons, methods of DNA preparation other than shotgun sequencing are typically preferred in forensic analysis. Specifically, a method is used which both selectively targets the regions of interest in the sample, and allows for the use of much lower levels of input DNA than would be possible with shotgun sequencing. This is called 'targeted' sequencing (or often targeted resequencing, the point being that the DNA region being analysed must have previously been sequenced and characterised for this analysis to be possible). In targeted sequencing an initial step is performed that either amplifies the regions of interest in the DNA by PCR, or uses probes to selectively capture the regions of interest (Figure 2). The PCR method is the most sensitive and can require less than 10 ng of DNA, while probe-based methods typically require 50-500 ng of input DNA (Børsting and Morling 2015).

Figure 2: A representation of targeted (re)sequencing. The dark blue bar represents the reference stretch of DNA. The light blue bars represent the fragments generated that cover the areas of interest in the reference. This is done via PCR or probe capture. Not all areas of the reference are sequenced.

Once the fragments of DNA to be sequenced have been prepared, either by shotgun or targeted methods, the next step is to use these fragments to prepare what is known as a library. A library is a collection of fragments that are ready for sequencing via the addition of known stretches of DNA to the unknown fragment that is to be sequenced (Figure 3). These known stretches can be oligonucleotide adapters that are attached to the ends of the fragments enzymatically, or known stretches of DNA that are attached via PCR. In the case of PCR, the PCR primers are tagged with additional sequence adjacent to the 'core' primer sequence that is complementary to the template DNA and drives the PCR reaction (Figure 4).

Figure 3: A simplified view of a library fragment. The blue stretch of sequence represents the unknown sequence that will be analysed in the coming sequencing run. The red stretches are known sequence that have been attached to facilitate the coming steps.

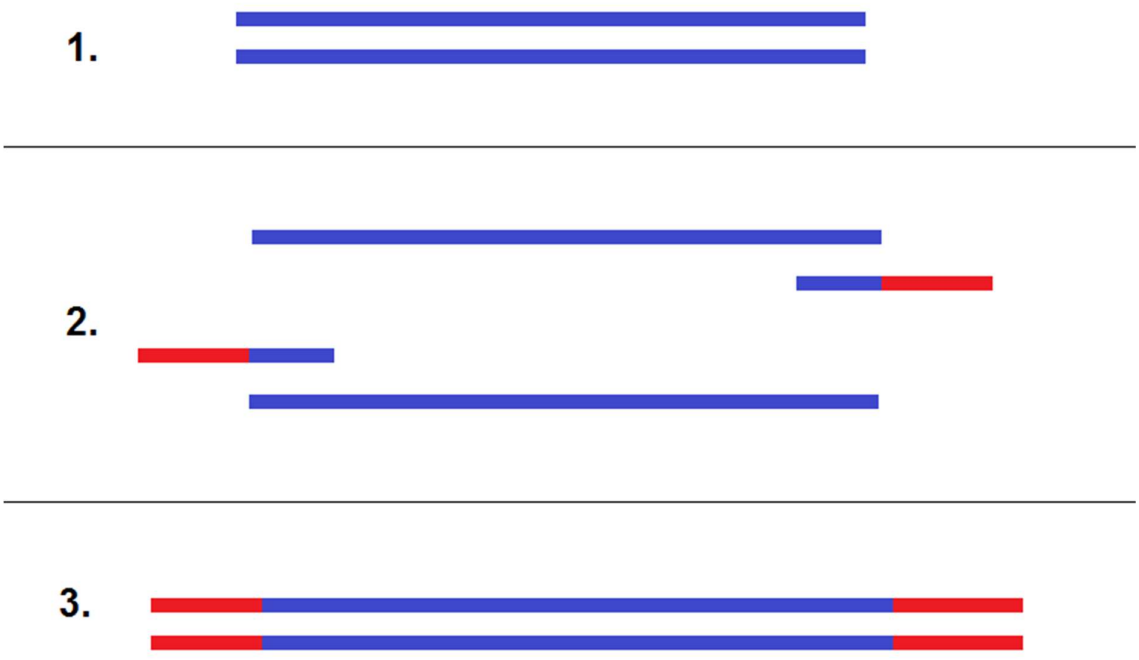


Figure 4: A representation of how known stretches of DNA can be attached to the fragments to be sequenced via PCR. Section 1 of the image shows the double stranded DNA fragment to be sequenced. In Section 2, the DNA is denatured and primers with sequence complementary to the template (blue) and the desired extra sequence (red) are added. The result of such a PCR is shown in Section 3 – DNA fragments with the red extra sequence at each end.

The extra sequence that is added to the fragments to be sequenced, hereafter referred to as ‘adapter sequence’, serves different functions depending on the specific type of sequencing that is used to analyse the library. One feature that is common to many systems is called barcoding (also known as indexing). This is where different samples each have slightly different adapter sequence attached to them. These differing adapters allow the fragments originating from each sample to be told apart in the sequencing results, thus allowing several different samples to be pooled together and analysed in one sequencing reaction. This allows for much more efficiency in analysing multiple samples simultaneously (Figure 5).

Other features of adapter sequences differ according to the specific sequencing technology that is used to sequence the library. Adapter sequence can be used to bind the library fragments to other components of the sequencing reaction, or to act as a calibration of the sequencing result that is obtained from the unknown portion of the fragment. Discussion of the details of this follows in the next sections that describe various methods of massively parallel sequencing.

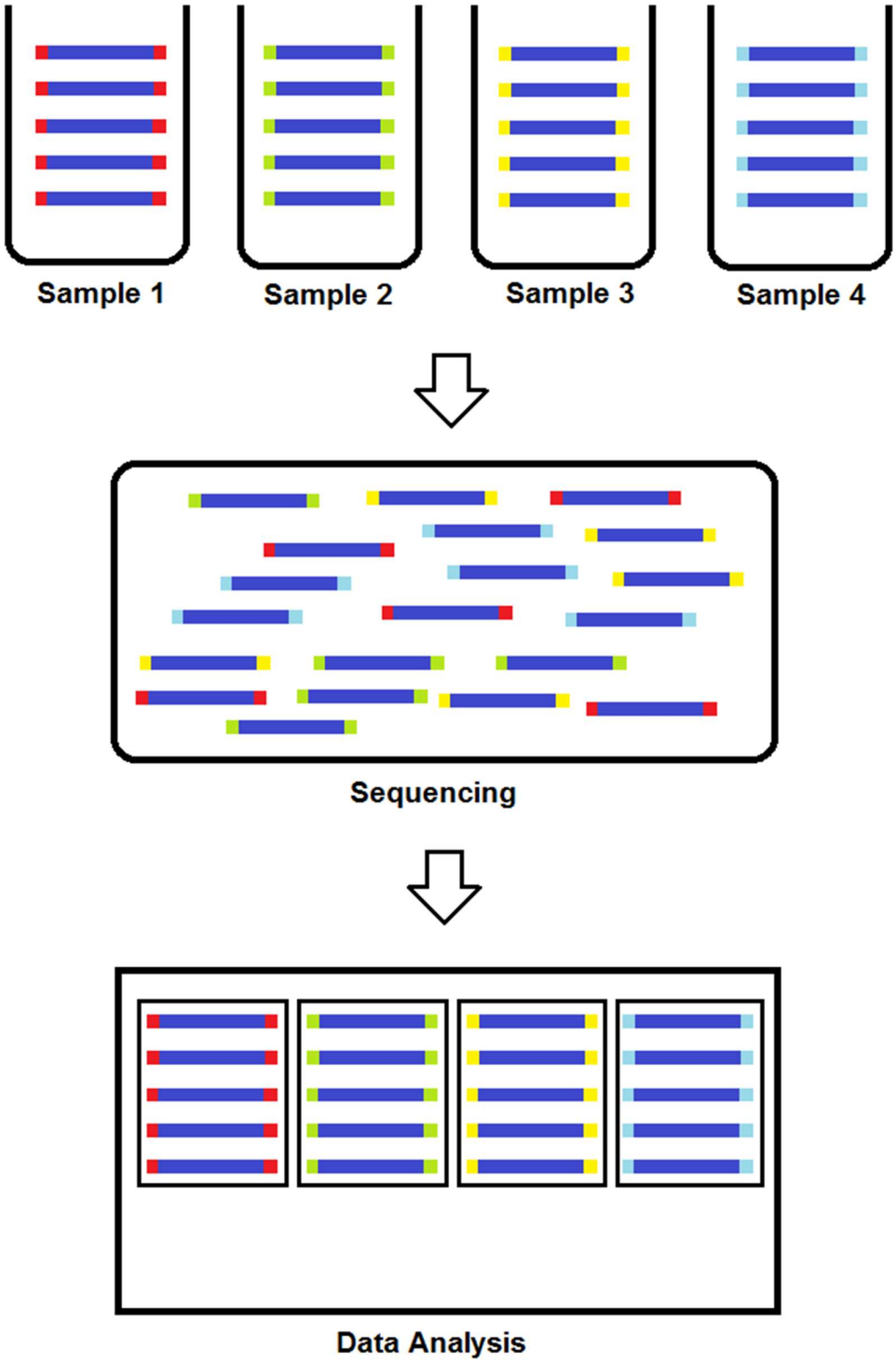


Figure 5: A representation of how barcoding allows multiple samples to be analysed in one MPS run. Each blue bar represents a stretch of unknown DNA to be sequenced. The differently coloured bars at the end represent the barcode sequences. A different barcode is added to each sample at the start of the process. All fragments are then combined and sequenced together. The differing sequence in the barcodes allows the data analysis software to tell which fragment belongs to which sample in the analysis.

1.1.3. Types of MPS

1.1.3.1. Pyrosequencing

One of the earliest methods of next generation sequencing was pyrosequencing, which was first proposed in 1996. Pyrosequencing performs sequencing by the sequential addition of nucleotides to the DNA sequence of interest, with the resulting incorporation reaction tied to an enzymatic cascade. This results in the emission of the fluorescent light which is detected by the instrument (Ronaghi *et al.* 1996). Pyrosequencing was made commercially available in instruments from 454 Life Sciences, but was never routinely adopted in forensic applications due to its need for relatively high levels of DNA input (Divine *et al.* 2010). Some sources have published forensic applications using pyrosequencing, but only in circumstances where the level of input DNA is not a factor, such as investigation of a sequence variant anomaly (Dalsgaard *et al.* 2014), analysis of mitochondrial DNA from reference samples (Mikkelsen *et al.* 2014), or investigation of the ability to use methylation sites for age prediction (Fleckhaus and Schneider, 2020 and Zbiec-Piekarska *et al.* 2018). Others have noted that the error rate of pyrosequencing is also unacceptably high for routine use in forensic analysis (Van Neste *et al.* 2012).

1.1.3.2. Sequencing by Ligation

This method of massively parallel sequencing involves the ligation of fluorescently labelled probes to a primer. Sequencing by ligation was developed by Life Technologies and was first made commercially available in 2006. Life Technologies' implementation of sequencing by ligation is named SOLiD sequencing, which stands for Sequencing by Oligonucleotide Ligation and Detection. In SOLiD sequencing a library is prepared from the sample to be sequenced and the resulting fragments are attached to beads. The reaction is setup such that only one fragment attaches to each bead, with many copies of that fragment being made over the surface of the bead. The attachment of the DNA to the beads is facilitated by the adapter sequences that are used in the library preparation – these adapters are known as the P1 adapter and are complementary to adapters that are pre-attached to the beads. The resulting beads are then bound to a glass slide for analysis. In the sequencing reaction, primers are added that are complementary to the P1 adapter contained in every fragment. After this, a set of four two-base probes compete for ligation to the primer. The probes are fluorescently labelled, allowing them to be detected once attached to the primer. Once a probe is bound and detected it is cleaved off and the process is repeated to analyse the length of the unknown sequence (Figure 6). The process is then repeated four times with

starting primers that are progressively one, two, three and four bases shorter than the initial primer. In this way, all bases in the unknown fragment are sequenced several times, resulting in what amounts to a built-in check on the accuracy of the sequence (Mardis, 2008). As a result, SOLiD sequencing has very high accuracy rates and does not suffer from the same issues that pyrosequencing does in this regard. On the other hand, SOLiD sequencing has several drawbacks which mean that it has not been adopted for forensic applications. These include short read lengths (only up to about 50 bp), very long runs times (>1 week), and the same requirement seen with pyrosequencing for relatively large amounts of input DNA (Børsting and Morling 2015).

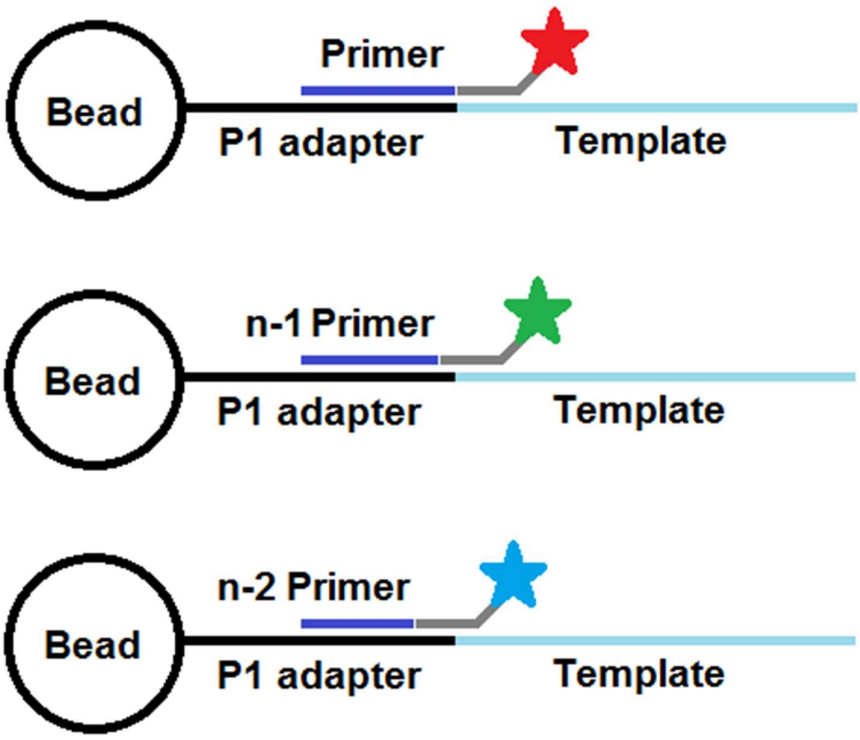


Figure 6: A representation of SOLiD sequencing. Library fragments are attached to beads via the P1 adapter. The sequencing primer then attaches to the adapter, allowing ligation of a probe, which is fluorescently labelled (see top image). The process is then repeated with progressively shorter primers (see second and third pictures with n-1 and n-2 primers) to allow full sequencing of the template strand to occur.

1.1.3.3. Single-molecule sequencing

Single molecule sequencing is another relatively new type of sequencing that may be considered alongside the other MPS methods discussed here. Single molecule sequencing

differs from the other methods discussed in that rather than a cluster of clonally amplified DNA fragments being analysed, only the original DNA fragments themselves are sequenced. In this sense, single-molecule methods are not true MPS, but another type of sequencing. As a result single-molecule methods are sometimes referred to as third-generation sequencing (with most of the other methods discussed here being second generation, or just 'next' generation sequencing). Despite these issues of nomenclature, as a 'new' type of sequencing that may be considered for use in future forensic application, single molecule sequencing is described here.

There are currently two major different types of single molecule sequencing, the PacBio platform, offered by Pacific Biosciences, and the MinION platform offered by Oxford Nanopore Technologies.

On the PacBio platform, the DNA to be sequenced is bound to polymerase and immobilised at the bottom of a well that is only nanometres in diameter and which only allows one DNA / polymerase complex to enter. Once in place, the template DNA is sequenced by use of four differently labelled fluorescent nucleotides. The fluorescent label is linked to the end phosphate in the nucleotide and is released when the next nucleotide is incorporated. A laser is used to excite the attached fluorescent label and a camera monitors the emitted fluorescence. In this way, the DNA is sequenced in real time at a rate of about five nucleotides per second. As such, the run is very fast (under one hour), and can sequence very long stretches of DNA (up to about 15,000 bp) (Eid *et al.* 2009; Børsting and Morling 2015).

The MinION platform sequences DNA based on passing the DNA to be sequenced through a very small opening (a nanopore) in a synthetic polymer membrane under the influence of an electric current. As the DNA passes through the nanopore, the passage of other ions is blocked and the decrease in current can be monitored to determine the sequence of the DNA. Like the PacBio, the MinION platform also allows very long read lengths (up to 70,000 bp) and fast run times (up to 450 bp per second) (Haque *et al.* 2013; Børsting and Morling 2015).

Both single molecule sequencing methods have received little attention to date from forensic analysts however. One barrier to adoption is the high error rate on both technologies compared to other sequencing methods (Ren *et al.* 2021). Another barrier to adoption is again the requirement for large amounts of input DNA, 250 ng or more (Børsting and Morling 2015). That said, some forensic researchers have investigated the use of the MinION platform for use in sequencing forensically relevant SNP markers. Cornelis and colleagues (Cornelis *et al.* 2019), concluded that some loci are problematic for nanopore sequencing,

and that if these loci are avoided then forensic analysis with nanopore sequencing is 'technically feasible.' Ren *et al.* (2021) also used the MinION system to sequence commonly used forensic markers and found that STR typing was 'notably error prone', but that 13 autosomal STRs, 12 Y-STRs and 4 X-STRs could be found that showed high consistency between the MinION system and 'conventional' NGS analysis. The same group also reported over 99.9% accuracy in generating SNP genotypes with MinION.

1.1.3.4. Sequencing by Synthesis

A significant method of massively parallel sequencing for forensic purposes is sequencing-by-synthesis. This method of sequencing was invented by researchers at the Pasteur Institute in Paris and published by them in 1994 (Canard and Sarfati, 1994). The method was further developed by researchers at Cambridge University and commercialised by them in a company called Solexa. Solexa was acquired by Illumina in 2007, who continue to offer this sequencing technology today.

In sequencing by synthesis, a library is prepared and adapter sequences are added to the fragments to be sequenced via the tagged PCR method shown in Figure 4. As the initial step in the library preparation is PCR based, this allows for much lower levels of input DNA to be used compared to some of the methods discussed previously. The adapter sequences can include barcodes (known as 'indices' in the Illumina workflow) which allow multiple samples to be pooled together in one sequencing run.

After library preparation, the library fragments are hybridised to a slide. The slide contains hundreds of thousands of oligonucleotides which are complementary to adapter sequences used in the library preparation. In this way, the fragments are attached to the chip. Once attached, 'clusters' of the same sequences are produced via bridge PCR (Figure 7).

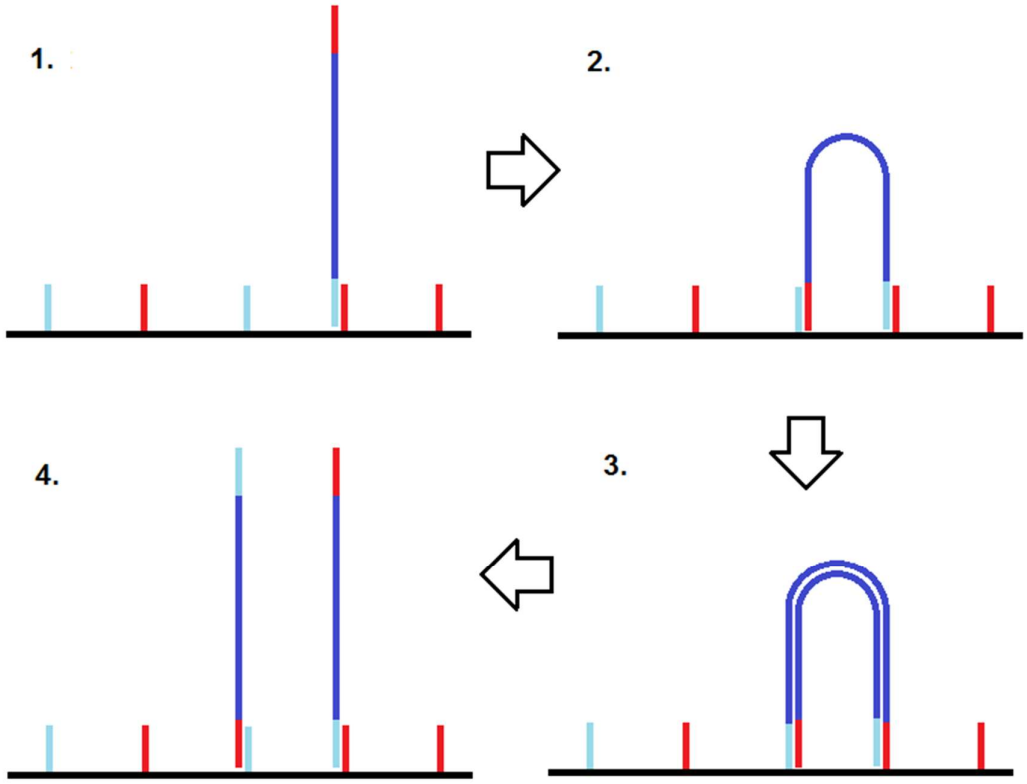


Figure 7: An illustration of the bridge PCR process used to generate clusters in sequencing by synthesis. 1. The library fragment is bound to the slide covered with oligonucleotides complementary to the library adapter sequences. 2. The library fragment forms a bridge shape. 3. PCR generates a complementary strand. 4. The result after multiple PCR cycles is a cluster of sequences attached to the slide which share the same sequence, some in the forward orientation, some reverse.

After the clusters are generated, they are then sequenced. This is done by the addition of fluorescently labelled nucleotides. These nucleotides are reversibly blocked in the 3' position, which prevents the incorporation of more than one nucleotide at a time. As such, all four nucleotides can be added simultaneously to the reaction and the nucleotide which is incorporated to each cluster can be detected. The 3' block is then reversed and the addition of nucleotides is repeated to sequence the DNA of interest (Børsting and Morling 2015).

Sequencing-by-synthesis was made commercially available by Illumina, and is the technology behind several different DNA sequencing instruments from that company. One of these instruments, the MiSeq FGx, is targeted for forensic applications. This instrument was released by Illumina in 2015, and has been the focus of much attention from the forensic community (Churchill *et al.* 2016). Since the release of the MiSeq FGx in 2015, the forensic division of Illumina has been separated into a distinct company known as Verogen. It is now

Verogen that offers the MiSeq FGx and the associated forensic chemistry for sale, rather than Illumina, although in some sources the names of the two companies are still used interchangeably. The limitations of the previously mentioned methods of MPS (accuracy, need for large amounts of input DNA, and speed of processing) are all improved in sequencing by synthesis technology, especially as implemented in the MiSeq FGx instrument.

Multiple forensic researchers have reported the use of the Illumina / Verogen technology for most commonly used forensic applications, including short tandem repeats (STRs) (Hussing *et al.* 2015, Gettings *et al.* 2016, Just *et al.* 2017, Guo *et al.* 2017, Churchill *et al.* 2017, Zeng, King, Stoljarova *et al.* 2015, Kim *et al.* 2017, Casals *et al.* 2017, Jäger *et al.* 2017, Köcher *et al.* 2018, Devesse *et al.* 2018, Hussing *et al.* 2018, Hollard *et al.* 2019, and Silvery *et al.* 2020), mitochondrial analysis, (McElhoe *et al.* 2014, Parson *et al.* 2015, Young Lee *et al.* 2016, King *et al.* 2014, Peck *et al.* 2016, Marshall *et al.* 2017, Lin *et al.* 2017, Woerner *et al.* 2018, Huszar *et al.* 2019, Wood *et al.* 2019, and Brandhagen *et al.* 2020), single nucleotide polymorphisms (SNPs), (Hussing *et al.* 2015, Casals *et al.* 2017, Jäger *et al.* 2017, Churchill *et al.* 2017, Guo *et al.* 2017, Ramani *et al.* 2017, Wendt *et al.* 2017, Köcher *et al.* 2018, Devesse *et al.* 2018, Hussing *et al.* 2018 and Hollard *et al.* 2019), and less commonly used applications such as methylation (Naue *et al.* 2017, Vidaki *et al.* 2017, Naue *et al.* 2018, and Aliferi *et al.* 2018). Details of these applications are examined in the next section on the forensic applications of MPS.

Sequencing by synthesis as implemented on the Illumina / Verogen MiSeq FGx is a significant method of MPS for the forensic community. Despite this, it has not yet achieved widespread adoption in routine forensic casework, which is still dominated by capillary electrophoresis (CE) methods (Butler 2012, Butler 2015). This is in part because MPS technology of any kind is still relatively new and many details of the implementation of this technology are still to be decided. Further to this, for it to be widely adopted, MPS must not just be robust in itself, it must also demonstrate the ability to perform tasks and solve cases that cannot be solved by the established CE methods. Exploration of these possibilities is a major theme of this work. For this work however, focus was made on the semi-conductor sequencing methods discussed in the next section rather than sequencing by synthesis methods. Many of the questions of optimisation of MPS for forensic purpose apply to both technologies equally however.

1.1.3.5. Semiconductor sequencing

This research focussed on the use of semiconductor MPS sequencing in forensic analysis. Semiconductor sequencing was first described in a paper in *Nature* in 2011 (Rothberg *et al.* 2011) and is made commercially available by Thermo Fisher Scientific under their Ion Torrent brand. The first instrument promoted for forensic use by Thermo Fisher Scientific was the Personal Genome Machine (PGM). This has been followed by the newer S5 instrument. Both instruments operate under similar principles however.

Similar to the methods described in previous sections, semiconductor sequencing again begins with preparation of a library of DNA fragments to be sequenced. Although it is possible to make the fragments for Ion Torrent sequencing via enzymatic or physical fragmentation methods, in practice for forensic purposes PCR based methods of making the fragments are preferred. As discussed previously, this allows two highly desirable features for forensic analysis – the ability to selectively target the desired loci in the sample, as well as allowing large amounts of DNA to be made from small amounts of starting material.

After the fragments are generated by PCR, adapter sequences are attached to each end of the fragments. This can include a barcode sequence that allows multiple samples to be pooled together in one run. The adapters are not attached to the fragments via PCR as in the Illumina forensic method, but by enzymatic ligation. Because the adapter sequences that are ligated are blunt-ended, this means that they can attach to either end of a given DNA fragment.

Because of the large number of fragments in the library, this means that for a given sequence some fragments with that sequence will have the adapters attach in one orientation, some in the other orientation. As a result, the sequence in question will be sequenced in both directions depending on at which end the adapters attached to each individual fragment (Figure 8).

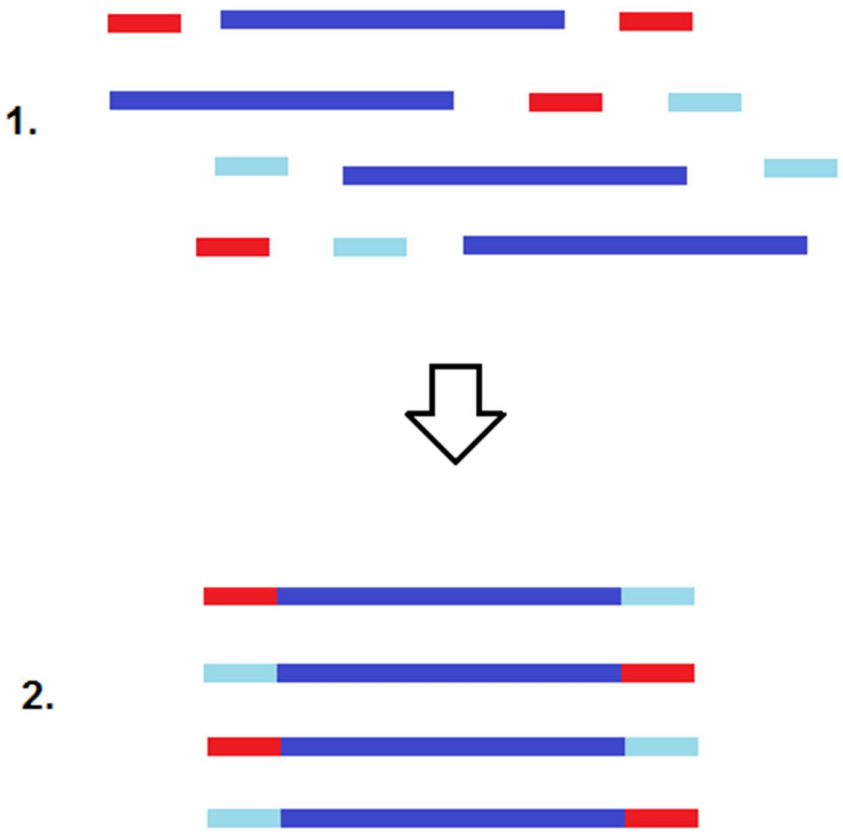


Figure 8: An illustration of the adapters used in Ion Torrent sequencing. 1. The adapters (red and light blue) are added to the fragments to be sequenced (dark blue). 2. The adapters are ligated enzymatically to the fragments. Because the adapters are blunt ended, they can attach to the fragment at either end. Sequencing always occurs in the same direction relative to the adapters (for example, from red to light blue), this means that a given sequence will be sometimes be sequenced forward, sometime reverse depending on the orientation of the adapters.

Once the library has been prepared, the library fragments are then attached to beads. This is done via an emulsion PCR. The beads, also known as Ion Sphere Particles (ISPs), have primers embedded in them that are complementary to one of the adapter sequences used in the library preparation (the P1 adapter), which allows the library fragments to attach during the PCR.

The emulsion property of the PCR is achieved by mixing the water-soluble PCR components (template DNA, dNTPs, polymerase, etc) with oil. This mixture is then either vigorously mixed or passed through a filter to create an emulsion. The resulting emulsion consists of millions of small pockets of water-based solution, which contain the PCR reagents

suspended in a background of oil. PCR cycling is then conducted within this emulsion. The purpose of creating the emulsion is that the small pockets of water-based solution (also known as reactors) act as tiny individual PCR tubes, with each reactor holding a self-contained PCR reaction. This is done so that, on average, one DNA sequence only is attached to one ISP – if the emulsion was not made, many different sequences would attach to each ISP during PCR, which would result in indecipherable data on the sequencer.

Once the emulsion PCR is done, the emulsion is broken, the water-based product is recovered, and a clean-up procedure is performed to remove any ISP that did not attach to library fragments. The resulting ISPs, known as templated ISPs (i.e. ISPs with DNA attached) are then loaded into a semi-conducting chip along with sequencing primer and polymerase. It is on this chip that the sequencing occurs and from which this form of sequencing gets its name.

The chip in question has a surface that contains millions of picolitre sized wells, which are sized to fit one templated ISP only. When the templated ISPs are added to the chip, each ISP goes into one well of the chip. From here, the sequencing reaction occurs. The instrument flows four unmodified nucleotides over the chip in turn. When the correct nucleotide is offered to a given well, it is incorporated into the strand being built by the polymerase. The natural result of the chemistry of this incorporation is that protons are given off into the solution of the well. It is these extra protons, in essence a pH change in the well, that is detected by the instrument and translated by the software into DNA sequence (Rothberg *et al.* 2011) (Figure 9).

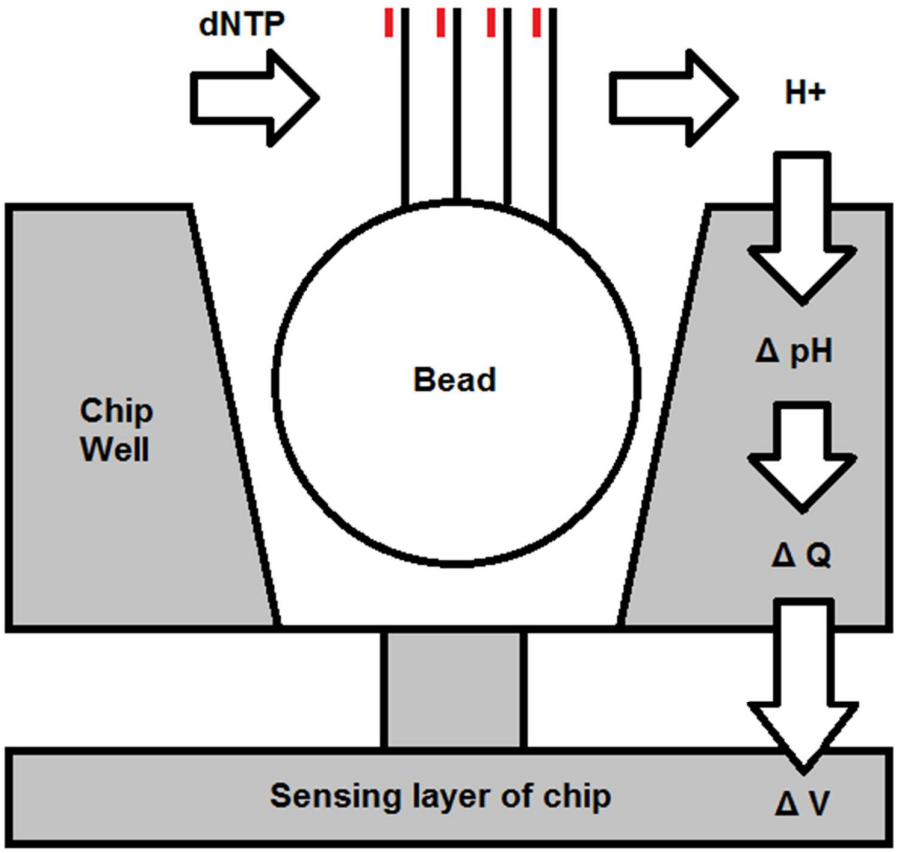


Figure 9. An illustration of how sequencing is performed on the Ion Torrent chip. DNA is attached to the bead (or ISP), which fit into the wells on the surface of the chip (one well shown here in cross-section). Each of the four dNTPs are offered in turn to the chip. When the correct base is offered, it is incorporated into the DNA on the bead by the polymerase. A sequencing primer (shown is red) is used to start the reaction, and is complementary to one of the adapters used in library preparation. When the dNTP is incorporated, the natural chemistry of the phosphate backbone locking into place results in protons being released. This is equivalent to a pH change in the well. This is in turn equivalent to a change in the electrical resistance of the solution in the well (Q), which is detected by the electrical components of the chip as a change in voltage (V). In this way, the sequence of the DNA attached to the bead is determined.

Semi-conductor sequencing has received significant attention from the forensic community and all commonly used forensic applications have been demonstrated on the Ion Torrent platform. This includes STRs (Zubakov *et al.* 2015, Bottino *et al.* 2015, Churchill *et al.* 2015, Zhao *et al.* 2015, Fordyce *et al.* 2015, Guo *et al.* 2016, Zhao *et al.* 2016, Wang *et al.* 2017, Li *et al.* 2017, Wang *et al.* 2018, Müller *et al.* 2018, Tao *et al.* 2019, Zeng *et al.* 2019, and Barrio *et al.* 2019), mitochondrial analysis (Parson *et al.* 2015, Bodner *et al.* 2015, Zhou *et al.* 2016, Juras *et al.* 2017, Pereira *et al.* 2018, Wai *et al.* 2018, Ma *et al.* 2018, Woerner *et*

al. 2018, Strobl *et al.* 2018, Avila *et al.* 2019, Strobl *et al.* 2019, Garcia *et al.* 2020, Sturk-Andreaggi *et al.* 2020, and Wang *et al.* 2020), and SNPs (Børsting *et al.* 2014, Eduardoff *et al.* 2015, Churchill *et al.* 2015, Ralf *et al.* 2015, Gettings, Kiesler *et al.* 2015, Buchard *et al.* 2016, Pilli *et al.* 2016, Themudo *et al.* 2016, Elena *et al.* 2016, Guo *et al.* 2016 [2], Bleka *et al.* 2017, Meiklejohn and Robertson 2017, van der Heijden *et al.* 2017, Santangelo *et al.* 2017, Pereira *et al.* 2017, Juras *et al.* 2017, Tasker *et al.* 2017, Garcia *et al.* 2017, Garcia *et al.* 2017 [2], Wang *et al.* 2018, Nakanishi *et al.* 2018, Mo *et al.* 2018. Kukla-Bartoszek *et al.* 2018, Liu *et al.* 2018, Phillips *et al.* 2019, Ralf *et al.* 2019, Breslin *et al.* 2019, Simayijiang *et al.* 2019, Avila *et al.* 2019, Wang *et al.* 2019, Mogensen *et al.* 2020, and Petrovick *et al.* 2020).

Given the amount of attention received from the forensic community, it seems that that Ion Torrent semi-conductor sequencing and Illumina sequencing by synthesis are by some distance the two dominant methods of MPS being considered for forensic applications. The other methods covered in this section (pyrosequencing, sequencing by ligation, and single molecule sequencing) have serious limitations which mean that very few, if any, forensic applications have been applied to these technologies.

Despite this, as noted previously, neither Ion Torrent nor Illumina sequencing have become routine in every day forensic practice, which is still dominated by CE based analysis. This is in part because much is still untested in the application of any MPS technology to forensic practice. Investigation of these aspects was a focus of this work.

1.1.4. Forensic Applications of MPS

As touched upon in the previous section, many different forensic DNA applications are possible on MPS systems, including mitochondrial DNA, STR, and single nucleotide polymorphism (SNP) analysis. SNP analysis can have many purposes and be, for example, identity SNPs, ancestry informative SNPs or phenotyping SNPs. These applications have received attention in Parson *et al.* 2013, Fordyce *et al.* 2015, Gettings, Kiesler *et al.* 2015, Churchill *et al.* 2016. The choice of which of these applications to use in forensic case depends on the goal of the case, with, for example, identity informative SNPs providing extra discrimination in difficult kinship cases, sequencing of STRs providing extra information that may help to resolve difficult mixtures, and mitochondrial analysis assisting in cases where nuclear DNA is degraded to the point that it cannot be usefully analysed.

1.1.4.1. Single Nucleotide Polymorphisms (SNPs)

Single nucleotide polymorphisms, or SNPs, are single bases in the genome that are known to vary between individuals. It is well known that the vast majority (approximately 99.5%), of the human genome is identical between different individuals, but given that the human genome is approximately 3 billion bases in size, this leaves many millions of bases that will differ between individuals (Levy *et al.* 2007). The majority of this variation between individuals takes the form of SNPs, with there being an estimated 11 million SNP sites in the human genome (Shen *et al.* 2013) Many of these SNPs have been linked to function of the body, with the discovery of previously unknown connections between SNP genotypes and manifestation of disease being an important application of DNA sequencing in medical research (Shen *et al.* 2013). For forensic purposes however, both SNPs that can be linked to body function (or phenotype), and SNPs that have no apparent bodily function can be useful. SNPs that have no role in bodily function can be highly forensically useful if they can be used to tell one individual from another with a high degree of probability, while SNPs that can be linked to body form or function can be useful to the forensic analyst if used to give information on an unknown suspect's physical characteristics such as ethnicity, hair colour, eye colour, etc.

1.1.4.1.1 Identification SNPs

Identification SNPs, sometimes called ID SNPs, or iiSNPs (where ii stands for identity informative), are SNPs chosen for analysis in a forensic context to tell one individual from another with a high degree of probability. These SNPs do not typically give useful information on the outwardly visible traits of a sample donor. This is often because they are found in the non-coding region of the genome, where lack of evolutionary pressure has allowed more genetic diversity in the population, resulting in a SNP that tends to vary from one person to another, and is thus useful for discriminating individuals. For example, a bi-allelic SNP marker that has allele frequencies of 50% for both of its alleles has higher discrimination power in a forensic context than a bi-allelic SNP linked to a disease state where one of the alleles is much more commonly observed than the other. Although it is possible for forensically useful identification SNPs to originate in the coding regions of the genome, it is much less common than their being found in non-coding areas (Budowle and van Daal, 2008).

SNPs have long been considered for use in forensic identification applications, and their analysis in this field predates the development of MPS. Due to the limitations of previously

available analytical technologies however, such as capillary electrophoresis (CE), SNP analysis was a niche forensic application that was used much less frequently than other methods of analysis such as short tandem repeat (STR) analysis (Budowle and van Daal, 2008).This is because relatively few SNP loci can be analysed in one reaction with CE technology – at most about 30 or 40 SNPs can fit into one analysis. Because one SNP locus can only have at most four alleles (each of the four DNA bases), and in practice is often only bi-allelic, (compared to STR loci which can have twenty or more commonly observed alleles) forensic CE analysis has focussed much more on STR analysis, as the approximately 25 STR loci that can fit into a CE analysis give much higher discrimination between individuals than the previously mentioned 30 to 40 SNP loci.

With the advent of MPS technology, the limit of 30 to 40 SNPs that can fit into a CE multiplex is no longer a factor. Modern MPS systems can generate many megabases of sequence data in one run, easily sufficient to genotype many hundreds or even thousands of SNPs with high coverage. Coverage in this context refers to the number of times that a given locus has been genotyped in a single MPS run. Typically it is required to genotype each locus several times over to be certain of the genotype that the sample in question contains.

Traditionally, SNPs have not been extensively used in forensic analysis in part due to limitations mentioned above in the number of loci that can be analysed with CE based methods, but also due to limitations of how mixed samples can be interpreted (Budowle and van Daal, 2008). This is because for a bi-allelic SNP locus there are only three possible genotypes (if two SNP alleles are A and B, then possible genotypes are AA, AB, and BB). This means that a mixed profile will almost always feature shared alleles and can only be reliably detected if an individual with one of the homozygote genotypes is mixed with an individual with the heterozygote genotype if the contributors are evenly balanced. Other balanced combinations will not clearly be mixed profiles at all, something that makes interpretation of these profiles very difficult. Despite this, recent advances, especially in software development, have shown that mixed SNP profiles “are not the impossibility sometimes thought.” (Kidd *et al.* 2012).

Because of this increased ability of MPS to examine large SNP multiplexes, many in the forensic community have attempted to design panels of identity SNPs that can be used with MPS in forensic practice. There are two major commercially available identity SNP panels, one for each of the dominant MPS platforms described in the previous section. These are the Precision ID Identity panel, from Thermo Fisher Scientific, which is designed for their Ion Torrent semi-conductor sequencing systems, and the ForenSeq DNA Signature kit from Illumina / Verogen, designed for their MiSeq FGx system.

The Precision ID Identity panel contains 124 SNP loci in total, with 90 of these being autosomal identity SNPs and 34 identity Y-chromosome SNPs. The ForenSeq DNA Signature kit contains 94 autosomal identity SNPs, as well as multiple other forensic markers such as STRs, that will be discussed in the coming sections.

Both of the SNP multiplexes share a large number of SNPs with each other, with the bulk of the markers having been first published by either Kenneth Kidd (Kidd *et al.* 2006) or through what was known as the SNPforID project, a consortium of laboratories who proposed a 52 marker set of SNPs for identification purposes in 2007 (Musgave-Brown *et al.* 2007).

The Thermo Fisher and Illumina panels offer the majority of both of these SNPs sets together and in doing so, allow high discrimination of individuals, with random match probabilities in the order of 10^{-35} (Gettings *et al.* 2015).

These panels have been examined by multiple sources. For example Børsting *et al.* 2014, Gettings, Kieser *et al.* 2015, Churchill *et al.* 2015, Elana *et al.* 2016, Tasker *et al.* 2017, Liu *et al.* 2018, and Avila *et al.* 2019, all examined the Thermo Fisher Identity SNP panel, concluded that the panel shows promise in forensic practice with high discrimination values, good performance with degraded DNA, and good sensitivity being demonstrated.

Equally, Jäger *et al.* 2017, Kocher *et al.* 2018, Hussing *et al.* 2018, and Hollard *et al.* 2019 all examined the identity SNP component of the ForenSeq DNA Signature kit and similarly concluded that this panel show promise in forensic practice.

Despite this, work remains to be done in evaluating the panels on a wide variety of real forensic samples – testing their performance in the presence of inhibiting substances, for example, and in determining appropriate analysis thresholds. Little research has been done to date in these areas and was a focus of this work.

1.1.4.1.2 Ancestry inference

Ancestry SNPs, sometimes referred to as aiSNPs, where ai stands for ancestry informative, or AIMs (ancestry informative markers), are SNPs that have been chosen for analysis due to their ability to predict the likely ancestry of the sample donor. This is because these SNPs have allele frequencies that have been shown to differ significantly between different ethnic groups around the world (Budowle and van Daal, 2008). For example, a given bi-allelic SNP with alleles A and B, might have allele frequencies of 80% A and 20% B in African populations, but the reverse, 20% A and 80% B, in Europe. This means that if a person is

observed with genotype AA, it is more likely that they are of African than European ancestry. Of course, with analysis of only one SNP this is far from conclusive, but if many independent SNPs are analysed in parallel it is possible for a more accurate estimate of ancestry to be determined (Phillips, 2015).

Similar to the points made in the previous section on Identity SNPs, analysis of Ancestry SNPs is not new, but it is with the advent of MPS that it has become practical to analyse large enough SNP sets to allow this type of analysis to be done.

The first panel of Ancestry SNPs targeted specifically for forensic use was the SNPstream system from a now defunct company called DNAprint. This was 178 SNPs that were analysed in multiple PCR reactions, something that can now easily be done in one MPS reaction, and an illustration of how the greater throughput MPS has facilitated this type of work. Details of the SNPstream system were published in 2008 (Halder *et al*, 2008) and used the 'δ' metric for selecting the markers used. This is an estimate of the genetic distance between populations for a marker and for populations A and B is equal to the absolute difference in allele frequencies between A and B (i.e. $p_A - p_B$). This was first proposed by Shriver in a 1997 publication (Shriver *et al.* 1997).

Two further forensic Ancestry SNP panels were then proposed, the first a 34-plex developed by the SNPforID consortium (a group of collaborating European forensic laboratories) (Phillips *et al.* 2007) and a 47-plex proposed by a group from the Netherlands (Kersbergen *et al.* 2009). Both panels were designed to be used in CE based applications and are able to discriminate samples at a continental level. Despite receiving attention from researchers, neither panel was widely adopted in forensic casework.

With the increased availability of MPS for forensic analysis, the possibility of larger, more accurate Ancestry SNP panels has been opened. In 2009 a 128 SNP panel was published by Kosoy *et al.* (Kosoy *et al.* 2009), which focusses on African, European and Native American discrimination. This panel is often referred to as the Seldin panel, after Michael Seldin, the last-named author in the Kosoy publication. A 55 SNP panel that extended the 128 in the Seldin panel was then proposed by Kenneth Kidd in 2011 (Kidd *et al.* 2011). These two panels form the basis of the commercially available Ancestry SNP panel offered by Thermo Fisher Scientific for their Ion Torrent semi-conductor sequencing systems: the Precision ID Ancestry panel. A further commercially available panel is the ForenSeq DNA Signature kit from Illumina / Verogen, designed for their MiSeq FGx system. This panel was previously referred to in the section on Identity SNPs. It contains several types of forensic markers: the Identity SNPs named previously, Ancestry SNPs, and other markers that will be discussed in coming sections.

The Precision ID Ancestry panel contains 165 SNP loci in total, with 123 of these from the Seldin panel and 55 from the Kidd panel (13 SNPs appear in both the Kidd and Seldin panels). The ForenSeq DNA Signature kit contains 56 Ancestry SNPs, based on the Kidd panel.

Both of these commercial offerings have received much attention from forensic researchers, who have demonstrated the potential that this type of analysis holds in forensic applications. Publications have found that the Precision ID panel (Garcia *et al.* 2017; Pereira *et al.* 2017; Santangelo *et al.* 2017; Tasker *et al.* 2017, Nakanishi *et al.* 2018, and Simayijiang *et al.* 2019) and the ForenSeq panel (Ramani *et al.* 2017; Churchill *et al.* 2017; and Xavier and Parson 2017, Hussing *et al.* 2018) perform well in distinguishing samples based on their ancestry, although, like the Identity SNPs, less work has been done in real-world evaluation of how these panels perform in real forensic scenarios.

Beyond these panels, further work has been done by researchers to examine the extent to which ancestry can be determined by SNP panels, with other panels being proposed and examined to attempt more detailed discrimination than the continental level discrimination discussed above. Christopher Phillips and colleagues, for example have published two studies proposing marker sets that attempt to provide greater global ancestry resolution, the 'Global AIMSNP set' (Phillips *et al.* 2014) and greater resolution of Asia-Pacific populations, the 'MAPlex' multiplex (Phillips *et al.* 2019). Other examples cover East Asian populations (Li *et al.* 2016) and in the Pacific (Santos *et al.* 2016). This will be an on-going field of research, although none to date have resulted in commercially available panels.

1.1.4.1.3 Phenotyping

Another area where SNPs can be analysed by MPS in a forensic context is for the purposes of phenotyping, that is, to determine externally visible characteristics of sample donors. This can be related to the previous topic of determination of ancestry, as people from different continents will tend to have certain physical traits that are characteristic of those areas – skin, eye and hair colour, for example. Those studying phenotyping for forensic purposes have focussed on finding SNPs that directly characterise those (and other) traits however, rather than focussing on the biogeographic ancestry as in the previous section.

The first forensic focussed eye colour prediction system was published in 2011 by Walsh *et al.* (Walsh *et al.* 2011) and is known as IrisPlex. This is a six SNP multiplex which has been shown to predict eye colour as either brown, blue or 'intermediate' with an accuracy of

approximately 84%. In 2013 a group mostly comprised of the same researchers expanded the IrisPlex to HIrisPlex, a 24-plex for predicting hair and eye colour, which comprises all six of the original IrisPlex SNPs, and which has been reported to give an average hair colour prediction accuracy of 73% (Walsh *et al.* 2013).

The HIrisPlex system was then further expanded from 24 to 36 SNPs and renamed HIrisPlex-S, in a new panel that added skin colour to the attributes that the panel can distinguish. It had previously been noted in 2013 that some of the genes involved in quantitative skin colour variation had been identified (Jacobs *et al.* 2013), in principle allowing this analysis to be done. A 2017 publication by several of the same group who later published the HIrisPlex-S panel defined the SNP set that would later become HIrisPlex-S (Walsh *et al.* 2017). In their publication on the new HIrisPlex-S panel (Chaitanya *et al.* 2018) the authors categorise skin colour into five categories: very pale, pale, intermediate, dark, and dark-black, and state that the panel will provide a probability score for each category for a given genotype. If the probability is 0.9 or over, then the category in question is predicted. If less than this, it will be affected by the category with the second highest score and may represent a mix or intermediate state of the two. This represents progress over an earlier study, (Maronas *et al.* 2014) which identified 29 SNPs that were most correlated with skin colour out of 59 SNPs that had been previously identified as being associated with skin, eye and hair colour. These SNPs allowed separation of most white skin donors from other donors, but those with black and intermediate skin overlapped considerably.

Related to the prediction of hair colour via the HIrisPlex panel noted above, in 2018 many of the same authors involved in the initial publication of the panel published a paper that explored the phenomenon of age-dependant hair darkening, where a child's hair darkens considerably from early childhood into adolescence and adulthood (Kukla-Bartoszek *et al.* 2018). Using the same HIrisPlex markers the authors found that in two-thirds of cases, HIrisPlex recognised blond hair colour from early childhood but not the darker hair colour seen at advanced childhood, resulting in an incorrect prediction. On the other hand, in one-third of cases, HIrisPlex predicted the darker hair seen at an advanced age, even though the child was blond in early childhood. This illustrates the complexity of the prediction of hair phenotypes, and indicates, as the authors conclude, that more work remains to be done in this area.

The only commercially available MPS panel to feature phenotyping markers is the ForenSeq DNA Signature kit from Illumina / Verogen. In addition to the Ancestry and Identity SNPs discussed in the previous sections, this panel also includes 22 phenotyping SNPs, which are informative of the sample donor's eye and hair colour. These are based on the HIrisPlex

markers. As discussed in previous sections, the Illumina / Verogen panel has been the subject of several forensic publications, but these have generally focussed on the other markers in the panel rather than on the performance of the phenotyping markers specifically.

Researchers have also attempted to characterise other externally visible characteristics through DNA analysis, but these studies have not advanced as far as those for eye, hair, and skin colour. Aside from these attributes, other groups have investigated height, hair shape, face shape, presence of freckles, and age as potential features that may be characterised via DNA analysis. One study of height was able to provide 75% accuracy in distinguishing between 'extremely tall' and 'normal' donors (Liu *et al.* 2014), and a group of many of the same researchers performed a later study on the same topic which improved this to 79% with an expanded SNP set (Liu *et al.* 2019). No studies however have attempted to provide a quantitative prediction of height. As such, it is clear that there is still much to be discovered in the genetic determination of height. It is suspected that if true characterisation of height is possible it will likely involve analysis of many thousands of SNPs (Kayser, 2015).

Hair shape, that is the characterisation of hair as being straight, curly, wavy, etc, has been investigated by some researchers, with a 2018 study (Pośpiech *et al.* 2018) reporting success of 66% in Europeans and 79% in non-Europeans in classifying hair as being either straight, curly, or wavy with a set of 32 SNPs, an improvement from an earlier study which performed similar classification with a set of 14 SNPs (Liu, Chen *et al.* 2018). Work continues in this area, with the publication by Pośpiech and colleagues concluding that: "this study demonstrates that the search for more hair shape associated DNA variants and the investigation of their predictive value in independent samples needs to continue." (Pośpiech *et al.* 2018).

Face shape prediction would be of significant appeal to forensic investigators, but is a difficult challenge. The large number of factors involved, both genetic and environmental, mean that true face shape prediction is some way off. As stated in a 2015 review article: "we are just at the beginning of understanding which genes determine normal facial variation, and it will likely be a long way (and wait) until enough predictive DNA markers are available for practical (forensic DNA phenotyping) of the face." (Kayser, 2015).

Some researchers have investigated the possibility of determining the presence of freckles in a sample donor via their genotype, with the first publication on this topic (Hernando *et al.* 2018) reporting an approximate 74% success rate in detecting the presence or absence of freckles with eight SNP loci, located in four genes that have been shown to be associated with freckling. Another group expanded on this work with a 19 locus SNP set that categorised samples as being non-freckled, heavily freckled or intermediate with success

rates of approximately 75%, 79% and 66% respectively (Kukla-Bartoszek *et al.* 2019). Work again continues in this area, the authors of the 2019 study noting that further improvement in resolution could be obtained by sequencing the entirety of the MC1R gene, one of those that has been shown to influence this phenotype.

Lastly, age prediction from DNA is an area that shows some promise, but not in the area of SNP analysis. This is understandable as, generally speaking, an individual's SNP genotype does not alter as they age, and so a genetic indicator of age must look beyond the SNP genotyping methods described previously. Some methods that have been investigated here include counting of age-dependent accumulation of mitochondrial DNA deletions and telomere shortening, but these have been concluded to be inappropriate for forensic use (Meissner and Ritz-Timme, 2010). The most promising method of forensic age estimation is via measurement of DNA methylation, which is discussed in Section 1.1.4.5.

1.1.4.2. Short Tandem Repeats (STRs)

Short tandem repeats, also known as STRs or microsatellites, are the most well-established forensic DNA marker, with profiles comprised of these markers forming the vast majority of forensic DNA databases and investigations globally (Butler 2012). STRs are regions in the genome that have a short sequence of bases repeated a number of times in a row. The number of these repeats in a given STR locus can vary from person to person and counting the number of repeats at each locus forms the basis of STR profiling.

When forensic DNA profiling was becoming established, STR profiles became the preferred method of profiling over SNPs and other types of markers in part due to the discrimination power that they offer. One STR locus will typically have at least five and often ten or more commonly observed alleles, in contrast to SNPs, which can have at most four alleles, and in practice often only have two. As such, one STR locus is significantly more discriminating than one SNP locus, something that is important when trying to fit loci into the relatively small read region offered by CE profiling (Budowle and Van Daal 2008). Due to this, and also due to their amenability to amplification by PCR and their relatively short length (compared to methods such as restriction fragment length polymorphisms or RFLPs), STRs became the dominant method of forensic DNA profiling from the 1990s to the present day.

With the advent of MPS methods, as described in the previous sections, this consideration of fitting loci into a relatively limited CE profile is lessened as MPS is capable of easily analysing many hundreds of SNPs in one run. Despite this, it remains desirable to analyse

STR loci, as opposed to other types of marker, via MPS for several reasons. Firstly, to maintain compatibility with existing CE base systems and databases: a new crime profile can only be matched to an older database reference sample (or vice versa) if the same markers have been used in the profiling methods for each sample. Secondly, with MPS it is possible to reduce the fragment size of STR loci further than is used for CE, allowing increased performance with degraded DNA. MPS is not size based separation, so there is no need to 'space out' the markers as there is in a CE based profile. Each marker can be made as small as possible to allow the repeat region to be measured, without issues of 'fitting' the loci into the dye channels available to the CE instrument (Figure 10).

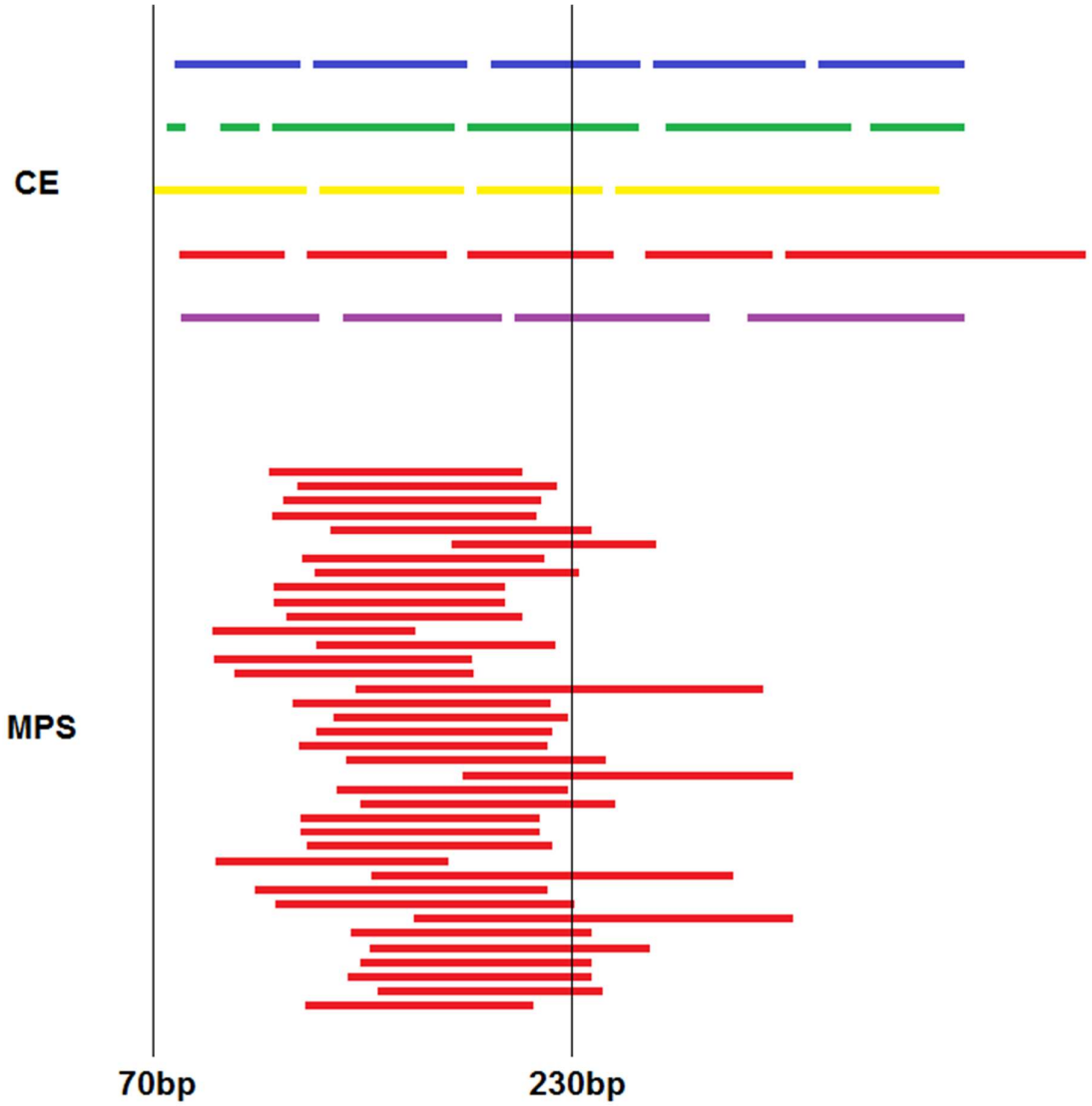


Figure 10: Schematic layouts of a typical CE-based STR multiplex (top) and MPS-based STR multiplex (bottom). CE technology requires that every locus within a dye channel (represented by the different colours) be spaced separately from the others, requiring detection up to approximately 400 bp to sequence 20 or more loci. With MPS technology, there is no equivalent requirement to space out the loci. This results in a multiplex that is capable of returning more information for a degraded sample.

The third reason for analysing STRs via MPS is that this analysis gives the full sequence of the STR loci, as opposed to CE, which simply measures the size of the fragments made in PCR and from that infers the number of STR repeats that must be present. With MPS, the STR sequence at a locus is directly measured. This results in an increase of the number of alleles detected, due to differences within the sequences and enhances the discrimination power of the individual loci. This point has been demonstrated by Gettings *et al.* (Gettings *et al.* 2016) who sequenced 183 samples from three populations with both CE and MPS methods and found a significant increase in the number of alleles seen in the MPS method for some loci.

Table 1: A list of loci examined by Gettings *et al.* (2016) and the number of alleles detected at each locus by both CE (length) and MPS (sequence) based methods. Loci are ordered by the increase in alleles seen with MPS methods.

Locus	Number of Alleles obtained by length	Number of Alleles obtained by sequence	Difference
D12S391	17	53	+36
D2S1338	12	40	+28
D21S11	19	46	+27
D8S1179	10	22	+12
D3S1358	8	19	+11
vWA	8	19	+11
D1S1656	14	23	+9
D2S441	9	14	+5
PentaE	16	19	+3
D18S51	18	21	+3
FGA	16	19	+3
D19S433	14	16	+2
CSF1PO	8	10	+2
D5S818	9	11	+2
D10S1248	9	10	+1
PentaD	14	14	0
D22S1045	11	11	0
D13S317	8	8	0
D7S820	7	7	0
D16S539	7	7	0
TPOX	7	7	0
TH01	6	6	0

A practical example of the benefits of this type of sequencing was published by Dalsgaard *et al.* (Dalsgaard *et al.* 2014) where inconsistent CE migration of D12S391 alleles was shown to be due to sequence variation between allele types that were of the same length but different sequence and thus appeared the same to CE, but were seen as distinct with MPS.

Connected, however, with this benefit of increased information from the sequence of STRs are issues of nomenclature. Several sources have noted that when attempting to describe the alleles obtained from MPS analysis of STRs, there are many complexities in establishing a nomenclature convention that is consistent, accurate, and compatible with profiles already obtained with CE profiling (Gelardi *et al.* 2014; Gettings, Aponte *et al.* 2015; Gettings *et al.* 2017, Phillips *et al.* 2018, Young *et al.* 2019). These include issues of which strand to use (until now commonly used STRs have sometimes been described using the forward and sometimes the reverse strand); how to define the repeat region relative to the flanking region, something that is not trivial if there is variation close to the edge of the repeat region; which reference sequence to use; how to describe variation in the STR flanking region; and how to maintain compatibility with length-based (i.e. CE) profiling. There is no fully agreed upon method for STR nomenclature of this type, but a step towards a standardised system was taken in 2016 when the International Society of Forensic Genetics (ISFG) published a paper with their recommendations on the topic (Parson *et al.* 2016). This included eight recommendations (or 'considerations') that attempt to provide a futureproof framework for the analysis and reporting of sequence based STR alleles.

Further steps towards an agreed upon method for categorisation of MPS STRs have been taken by the STRAND working group, which has received the endorsement of ISFG to report on this topic. The name STRAND is an acronym of Short Tandem Repeat: Align, Name, Define. A STRAND meeting was held in London in 2019 with representation from academia, practicing forensic laboratories, and industry with the aim to "present and discuss ideas, encourage mutual awareness, identify differences in approaches, opposing aspects, and opportunities for parallelization (in the nomenclature of forensic STRs)." The outcome of the STRAND meeting was published in 2019 (Gettings *et al.* 2019) and although a step towards agreement on how to define forensic MPS STRs, the STRAND report also notes that approaches in this field are still under development.

There are currently three manufacturers who provide commercially available forensic STR sequencing kits: Thermo Fisher Scientific for their Ion Torrent systems, Illumina / Verogen for their MiSeq FGx system, and Promega who do not manufacture MPS detection systems, but have released chemistry designed to be used with the Illumina / Verogen systems.

Thermo Fisher Scientific offer the Precision ID GlobalFiler NGS STR Panel v2, which contains 31 autosomal STR markers plus the Amelogenin sex determining marker and three other Y-chromosome markers designed as gender-confirmation markers (SRY, DYS391, and rs2032678, also known as the 'Y-indel' from the GlobalFiler CE STR kit). This kit, and previous versions of it that were released as prototype versions, have been the subject of several published studies (Fordyce *et al.* 2015, Bottino *et al.* 2015, Zhou *et al.* 2016, Guo *et al.* 2016, Vilsen *et al.* 2017, Wang *et al.* 2017, Li *et al.* 2017, Wang *et al.* 2018, Müller *et al.* 2018, and Tao *et al.* 2019) which have concluded that the panel displays promise for use in forensic analysis.

The ForenSeq DNA Signature kit from Verogen / Illumina, in addition to the SNP markers discussed previously, also includes 27 autosomal STR markers, 24 Y-STR markers and 7 X-STR markers. This kit has also been the subject of several publications (Zeng, King, Stoljarova *et al.* 2015, Guo *et al.* 2017, Just *et al.* 2017, Wendt *et al.* 2017, Jäger *et al.* 2017, Hussing *et al.* 2018, Hollard *et al.* 2019) which have again concluded that the system shows promise in forensic analysis

The commercially available PowerSeq Auto Y system from Promega is designed to run on Verogen / Illumina instruments and contains 22 autosomal STR markers, 23 Y-STR markers, and an Amelogenin sex test. It was preceded by the PowerSeq Auto/Mito/Y system which included an additional 10 amplicons covering the mitochondrial control region. These panels have also been examined in published papers, (Zeng, King, Hermanson *et al.* 2015, van der Gaag *et al.* 2016, Gettings *et al.* 2016, SIlva *et al.* 2018, Montano *et al.* 2018, Huszar *et al.* 2018, and Young *et al.* 2019) also with the conclusion that the results are suitable for forensic use.

As such, all of the commercially available MPS STR panels have been evaluated in the literature, with a general conclusion being formed that MPS measurement of STRs is a promising field of work. Results have been found to be generally concordant with CE technology and sensitivity and performance with degraded samples found to be fitting. Despite this, much work remains to be done in determining the real-world performance of this technology, both in terms of differing sample types, and in establishing analysis thresholds and methods that will be suitable for forensic use.

1.1.4.3. Mitochondrial Sequencing

Mitochondrial analysis has long been a significant area of attention for forensic analysis, due to the possibility of recovering mitochondrial DNA (mtDNA) from highly degraded samples. The large number of mitochondria per cell (in the order of hundreds to thousands (Just *et al.* 2015)), compared to the single cell nucleus, along with the protective double-membrane of the mitochondria mean that useful mtDNA profiles can be obtained from samples that are so degraded that no nuclear DNA remains. As such, a significant area of interest in the forensic application of MPS methods has been mitochondrial analysis. The appeal of MPS in this application is largely due to the significantly increased throughput of sequencing that MPS offers compared to CE based sequencing. CE-based systems can sequence at most in the order of 100,000 bases per run (Karger and Guttman, 2009), while all of the forensically orientated MPS systems mentioned previously can easy sequence many tens or even hundreds of megabases in a single run (Børsting and Morling, 2015). This means that the relatively laborious CE-bases sequencing of mitochondria, which often only sequenced the control region of the mitochondria for reasons of efficiency, can potentially be replaced by MPS methods that can easily sequence the entire 16,569 bp mitochondrial genome in a single run.

Work of this nature has been the focus of multiple publications in the forensic literature. One early source in 2012 cautioned that MPS methods were not yet sufficiently reliable to allow accurate forensic mtDNA profiling (Bandelt and Salas 2012), but since then several further publications have reported high quality results from MPS application of mtDNA methods, using both the Verogen / Illumina and Ion Torrent systems described previously (Parson *et al.* 2013; McElhoe *et al.* 2014, Parson *et al.* 2015; Lin *et al.* 2017; Peck *et al.* 2016; Zhou *et al.* 2016, Marshall *et al.* 2017, Strobl *et al.* 2018, Pereira *et al.* 2018, Woerner *et al.* 2018, Wai *et al.* 2018, Wood *et al.* 2019, Wang *et al.* 2020, and Brandhagen *et al.* 2020). These publications have used a variety of methods to prepare the DNA for sequencing on these platforms. This includes both targeted PCR methods, and fragmentation and capture methods. Most have reported high concordance of MPS mtDNA results to CE results, across a wide range of sensitivities. Peck *et al.* for example reported 99.9996% concordance was seen between the MPS data and the CE data for the full mitochondrial genome, with the only discrepancies involving low level point heteroplasmies (Peck *et al.* 2016).

It is of note that of the commercial providers, for some time, Thermo Fisher Scientific was the only to provide a dedicated mitochondrial panel, again under their Precision ID brand. This panel uses a PCR based approach to amplify either the entire mitochondrial genome (Precision ID mtDNA Whole Genome Panel) or a subset of the first panel (Precision ID

mtDNA Control Region Panel). Where mitochondrial analysis has been performed on the Verogen / Illumina platform, this has mostly been done without a forensic specific panel with custom PCR or fragmentation methods. Verogen have announced release of a forensic specific mitochondrial profiling kit, but at the current time, no publications have been released that use this kit.

As noted in the previous section, Promega's prototype PowerSeq Auto/Mito/Y panel also offered a PCR based method for analysing mtDNA on the Illumina system, although there have been few publications that focus on the mtDNA section of this panel this to date. Otherwise, Promega also offer the PowerSeq™ CRM Nested System for mtDNA analysis, which focuses on the mtDNA control region. Again, this is a panel that has been the focus of relatively few publications, although it was studied by Brandhagen and colleagues, who used it as the basis of their validation of MPS mtDNA profiling at the FBI laboratory (Brandhagen *et al.* 2020). In their paper on this the authors looked at the reproducibility, accuracy, efficacy, sensitivity and reliability of the panel, and conclude that the panel is suitable for use in the forensic lab and "will set the stage for the transition of additional NGS assays to casework over the coming years."

1.1.4.4. Microhaplotypes

A relatively new type of marker that has been considered for forensic use with the advent of MPS is the microhaplotype. This is a marker which is a small collection of SNPs that are so close together on the genome that rather than being treated as independent SNP loci, they must be considered together as a small haplotype, hence the term 'microhaplotype'. Just like with STR loci however, multiple microhaplotype loci can then be treated separately and independently combined together to form a profile with high discrimination power. This approach aims to combine the best features of SNPs and STRs, without several of the drawbacks. These advantages include the lower mutation rate and lack of stutter of SNPs, as well as the high discrimination power of STRs (Kidd *et al.* 2014). Discrimination power is largely a function of the number of possible alleles at a locus. A single bi-allelic SNP only has two possible alleles, while an STR locus can easily have ten or more (see Table 1). A microhaplotype that is comprised of (for example) four biallelic SNPs in a row can in theory have $2^4 = 16$ alleles. In practice, not all of these sixteen alleles may be observed in a given population, but the potential for greater discrimination is clear.

Microhaplotypes were first reported in forensic literature by Kenneth Kidd (Kidd *et al.* 2013) and were subsequently initially referred to in a small number of publications, again mostly by

Kenneth Kidd and his collaborators (Kidd *et al.* 2014; Kidd *et al.* 2017; and Wendt *et al.* 2016).

Since then, more forensic work on microhaplotypes has been performed however, with at least seven new studies being released in the last two to three years. These have explored the use of microhaplotypes from a number of forensic perspectives, with Chen and colleagues publishing two papers that looked at the ability of a 25 microhaplotype panel to resolve mixtures, and concluded that it could be 'greatly helpful' in individual identification from mixtures (Chen *et al.* 2018 and Chen *et al.* 2019). Zhu and colleagues reported a 13 locus microhaplotype panel that can be used for individual identification and ancestry inference (Zhu *et al.* 2019). Van der Gaag and colleagues published a set of 16 microhaplotype loci that were demonstrated by them to provide high discrimination power without the drawback of stutter, something that affects STR analysis of all types, whether done by MPS or CE, and has been noted as an advantage of forensic microhaplotype analysis (van der Gaag *et al.* 2018). Turchi *et al.* (2019) examined a large set of 87 microhaplotype loci and again concluded that these markers show high promise in mixture analysis. Sun and colleagues examined 30 microhaplotype markers and explored their utility in various kinship cases. They concluded that the markers show promise in this area, especially where there are mutated alleles in the STR result or cases involving close relatives (Sun *et al.* 2020). Lastly, de la Puente and colleagues proposed a 118 locus microhaplotype set, and concluded that this panel showed promise in the analysis of degraded DNA (de la Puente *et al.* 2020).

All of the above microhaplotype sequencing was done on either or both of the Verogen / Illumina or Ion Torrent platform, using custom chemistry to explore the new markers. It remains the fact that there are no commercially available microhaplotype MPS kits, perhaps a reflection of the fact that there is no widely agreed upon set of microhaplotype markers for this type of analysis. This remains an interesting field for future developments.

1.1.4.5. Methylation

Another active field of investigation in forensic DNA research is that of DNA methylation. This is an application of epigenetics, which is the study of alterations in gene function by mechanisms other than change in the DNA sequence itself (Vidaki *et al.* 2013). Put another way, epigenetics is an additional layer of information, on top of the genetic code, which controls how the genetic information is used. The genetic code provides a framework for RNA and protein structure, the epigenetic layer controls packaging of DNA and gene

regulation (Kader and Ghai, 2015). This epigenetic layer includes a wide number of factors, such as non-coding RNA, chromatin looping, nucleosomal remodelling, histone modification, and methylation (Vidaki *et al.* 2013). These epigenetic patterns are retained during cell division in the same way that the DNA sequence is also copied and retained in cell division, but unlike DNA sequence however, epigenetic factors can change over an individual's lifetime (Bird, 2002) and have been shown to change in response to environmental factors such as diet and smoking (Rando and Verstrepen, 2007). It is for this reason that epigenetic factors have drawn the attention of forensic researchers. In principle, it may be possible to determine aspects of forensic samples from epigenetic analysis that are not able to be determined with traditional DNA sequencing. This includes the discrimination of monozygotic twins, determination of the parental origin of alleles (maternal and paternal alleles can have different epigenetic markers), determination of the cause of death, age estimation, and identification of different body fluids (Vidaki *et al.* 2013).

Of the epigenetic markers, the one that has drawn the most attention from forensic researchers is DNA methylation. This is where methyl (CH_3) groups are added to the DNA molecule at certain bases, which changes the activity of the DNA without changing the sequence. This is a natural occurrence that plays a key role in the regulation of genes, typically in suppressing gene expression in the methylated areas. Of the four DNA bases cytosine and adenine can be methylated, with cytosine methylation being the most common. The exact size and location of methylated sites is known to change as a person ages and in certain disease states, and as such, analysis of methylation sites promises to help forensic analysts in determining the factors named above such as the age or cause of death of an unknown sample donor.

There are many ways in which methylation sites can be analysed, including mass spectrometry, melt-curve analysis, and enzymatic assays which, for example, can use the ability of restriction enzymes to differentially cleave methylated and unmethylated DNA sites. Relevant to the current work however is bisulphite sequencing, which is a method of methylation analysis that can be performed on MPS platforms, thus taking advantage of the sensitivity and high throughput of MPS sequencing discussed earlier. Bisulphite sequencing relies on an initial treatment of the DNA to be analysed with bisulphite, which converts cytosine to uracil, but leaves a methylated cytosine (5-methylcytosine) unaffected. As a result, the bisulphite treatment (also known as bisulphite conversion) results in changes to the DNA sequence which can then be detected on the MPS platform (Figure 11).

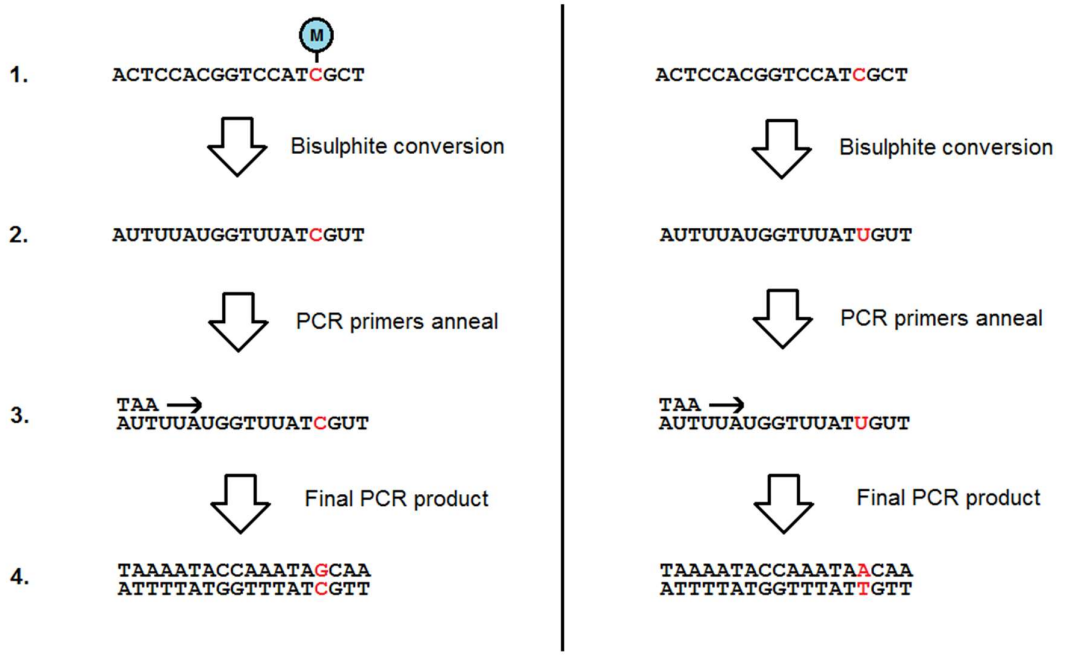


Figure 11: An illustration of how a bisulphite conversion allows methylation sites to be detected by sequencing. The left side of the figure shows a sequence with a methylated cytosine, the right shows the same sequence without the methylation. In step 2, the bisulphite treatment has turned all unmethylated cytosines into uracil. The methylated cytosine remains cytosine. The resulting product can then be amplified by PCR. The uracil bases in the template are converted to thymine via the PCR. This results in a final PCR product that will differ in sequence at the site of the original methylation.

Forensic analysis of methylation products via MPS is still a new area of investigation, with other non-MPS methods being more widely reported in literature. Despite this, some work has been done in the area. Vidaki *et al.* reported that with next generation sequencing methods on the Illumina MiSeq they could predict the age of 46 donor samples aged from 15 to 90 years with a mean average error of 7.45 years. This is worse than results achieved with other technologies (mean average error approximately 4 years), but the authors state that they believe that the MPS technology can be optimised to achieve even better results in future (Vidaki *et al.* 2017). Another 2017 publication reported a mean absolute deviation of 3.21 years on the prediction of age of a 104-sample set with true ages from 18 to 69 (Naue *et al.* 2017). A subsequent 2018 study by many of the same authors expanded this study to look at the same markers across different tissue types in deceased individuals, and concluded that an age related effect could still be seen (Naue *et al.* 2018). Another 2018 study by a different group gave an error of 4.1 years on prediction of age of 33 samples with true ages from 11-93 (Aliferi *et al.* 2018). This study also examined the sensitivity of the assay, and was able to achieve accurate results down to 10ng of input DNA – this is more input DNA than is typically required for forensic DNA methods due to the need for the

bisulphite conversion, something that makes testing of the sensitivity of methylation assays particularly important for forensic use.

Lastly, some researchers have used MPS to investigate the possibility of determination of body tissue type via methylation, with Bartling and colleagues using an Illumina MiSeq to examine ten methylation sites, and were able to identify 15 out of 16 samples of semen, saliva, skin and blood (Bartling *et al.* 2014). Further to this, Forat and colleagues again used an Illumina MiSeq to examine nine methylation markers, drawn from 150 candidate loci, and were able to distinguish venous blood, menstrual blood, saliva, vaginal fluid and sperm samples (Forat *et al.* 2016). This area continues to be an area of active forensic research.

1.1.4.6. Kinship

A practical application of the types of MPS analysis that have been described in the previous sections is kinship analysis, that is analysis of cases where the samples are members of the same extended family. While many cases of this nature can be satisfactorily resolved with a relatively small number of markers run on CE-based technology, for more ambiguous or distantly related cases, common practice is to extend the set of markers analysed in the hopes of finding more evidence that will strengthen the conclusions that can be drawn (Pinto *et al.* 2013). This desire to run more markers in certain kinship cases has led investigators to explore whether the enhanced throughput of MPS methods could be an efficient way to run large numbers of markers in complex or ambiguous kinship cases. Also of interest is the increased discrimination that can be offered with sequencing of STR markers, rather than size based CE analysis. This has initially been explored by Li *et al.* (Li *et al.* 2016), who used Thermo Fisher Scientific's Ion Torrent platform to conclude that MPS analysis of STR makers can provide more useful information in paternity cases, and Ma *et al.* (Ma *et al.* 2016) who used the Illumina MiSeq FGx to conclude that in paternity cases with apparent mismatches in CE STR loci, MPS can offer extra insights into whether these mismatches are exclusions or mutations.

In 2018, Mo and colleagues proposed a 472 custom SNP set where, using Ion Torrent technology, they demonstrated can 'sufficiently distinguish' second-degree relatives, that is avuncular, grandparent-grandchild and half siblings, from unrelated individuals (Mo *et al.* 2018).

Li and colleagues explored use of the Verogen / Illumina ForenSeq DNA Signature kit with kinship applications in their 2019 paper (Li *et al.* 2019), which found that the sequence

based information in the STR markers in that kit could boost kinship statistics, compared to the length based alleles that CE would provide from the same markers, and also that the SNP loci in the kit are of use when there are mutations in the STRs or when a relative is an alleged parent. They conclude that the kit can resolve paternity and full sibling cases and 'most' second-degree relationships, while more markers are needed for cousin cases.

A custom 1245 SNP set was proposed for kinship by Wu and colleagues in a 2019 paper (Wu *et al.* 2019), which concluded that the results for seven test paternity case with this panel were in agreement with conventional STR results, with stronger statistics for the match. Further, Wu found that the panel was suitable for cases with degraded DNA, one of the previously discussed benefits of SNPs for forensic analysis.

Lastly, as mentioned in the section on microhaplotypes, Sun and colleagues investigated the use of microhaplotype markers for kinship analysis in their 2020 paper (Sun *et al.* 2020), concluding that the microhaplotype markers show promise in this area, especially where there are mutated alleles in the STR result or cases involving close relatives

More research into the benefits of analysing different types of marker by MPS for kinship applications would be an interesting further avenue of research and was examined in this work.

Another application of MPS technology to the analysis of closely related individuals has been seen in the attempt to differentiate monozygotic (or 'identical') twins. This is a particularly difficult task given that theoretically monozygotic twins are genetically identical (Weber-Lehmann *et al.* 2014), and it is only by close examination of epigenetic factors or the small number of random mutations that occur in the genome that this differentiation is possible. Certainly with a 'traditional' forensic profile of CE based STRs, the likelihood of finding distinguishing characteristics of monozygotic twins is extremely small. This area has been examined with MPS methods however, where again, the large throughput of MPS has enabled large scale sequencing of the genomes of twins, allowing characteristic mutations to be seen that can tell the twins apart. This was published in 2014 by Weber-Lehmann *et al.* (Weber-Lehmann *et al.* 2014), who used Illumina technology to sequence around 600 megabases of a pair of twins (283 and 292 megabases for each twin respectively) and in that data found five SNPs that differed between the two. This result could be confirmed by analysis of those SNPs in a child of one of the twins, thus demonstrating that they were a real difference in the genome of the two 'identical' twins. Further studies in 2019 and 2020 confirmed this ability of MPS to distinguish monozygotic twins (Fang *et al.* 2019, and Yuan *et al.* 2020), with Fang and colleagues making use of MPS profiling of microRNA to distinguish four pairs of twins, they found that of the on average 158 microRNAs detected in each

individual, 14% of these differed between twins. Yuan and colleagues used whole-genome sequencing and confirmatory allele specific PCR to distinguish a single set of monozygotic twins.

1.2. Aims

The primary aim of this work was to establish whether MPS offers practical benefits in forensic casework that cannot be achieved with CE based methods. As such, evaluation was performed of multiple forensic massively parallel sequencing applications, with comparison in each case to the performance of an equivalent CE-based method. This included:

- Evaluation of the sensitivity of MPS
- Evaluation of the performance of MPS in the presence of inhibited DNA
- Evaluation of the performance of MPS on casework type samples
- Evaluation of the utility of MPS in detecting and resolving mixtures
- Evaluation of the utility of MPS in cases involving close relatives

Further work evaluated the performance of MPS in determining the ancestry of sample donors, an application that has received increased attention with the advent of MPS methods, given the ability of MPS to analyse numbers of loci that would be impractical with CE methods. As such, this has included:

- Evaluation of the ability of MPS to determine sample donor ancestry
- Exploration of alternative data analysis methods to determine sample donor ancestry

Lastly, aspects of MPS that are crucial to successful validation and implementation in forensic laboratories were examined, specifically:

- Evaluation of the concordance of MPS
- Evaluation of the reproducibility of MPS
- Exploration of methods to establish analytical thresholds for MPS and the factors that affect them.

Chapter 2:

Materials and Methods

2. Materials and methods

2.1. Overview

The following methods and materials were used to investigate the utility of MPS for forensic practice. These methods can be divided into three broad categories: sample preparation methods, which involve the collection of samples and the extraction and quantitation of DNA, and are commonly used in forensic DNA analysis regardless of the subsequent sequencing or detection method that is employed; capillary electrophoresis methods that cover the 'traditional' method of forensic DNA analysis, as routinely employed by most forensic DNA laboratories; and Massively Parallel Sequencing methods, the techniques that were examined here in comparison to the CE based techniques.

2.2. Sample Preparation

2.2.1. Sample Collection

Samples for the kinship component of this work (Chapter 6) were collected from nine donors using Copan Nucleic Cards and Floq swabs (Thermo Fisher Scientific, USA). Donors swabbed their inner cheeks with the swab and then transferred the collected buccal cells to the cards. These were left to dry at room temperature before processing.

For the ancestry samples studied in this work (Chapter 5), these were a collection of 64 donor samples provided by Thermo Fisher Scientific. Samples were provided as buccal scrapes on Floq swabs (Thermo Fisher Scientific, USA) and were then further processed as part of this work with the methods described below.

Where non-probative crime stain samples were used in this work, these were collected from a variety of simulated crime samples provided by Thermo Fisher Scientific. A precise description of each sample is given in Section 3.5, Table 35.

All donor samples used in this work, both for kinship and ancestry analysis, were collected with the informed consent of the donors in question and with ethical approval gained from the University of Central Lancashire in October 2017.

2.2.2. DNA extraction with PrepFiler Express BTA

All samples were extracted on the Automate Express instrument with PrepFiler Express BTA chemistry (Thermo Fisher Scientific, USA) as per the manufacturer's recommendations. Where Copan Nucleic cards were extracted with this method, two 1.2 mm diameter punches were added to the extraction lysis. Where Floq swabs were extracted with this method, the entire swab head was added to the extraction lysis. Sample lysate was incubated in LySep columns (Thermo Fisher Scientific, USA) at 56°C with 750 rpm agitation on an Eppendorf Thermal-shaker for 40 minutes, prior to purification on the Automate Express with the 'PrepFiler Express BTA' protocol. The final elution volume of the extracted samples was 50 µL.

2.2.3. DNA Quantification with Quantifiler Trio

All DNA extracts were quantitated on the 7500 Real-time PCR system using Quantifiler Trio chemistry and HID Real-Time PCR Analysis software v1.2 (all Thermo Fisher Scientific, USA) as per the manufacturer's recommendations. All plates run on this system used a five point duplicated standard curve with 50, 5, 0.5, 0.05 and 0.005 ng/µL of kit control DNA respectively.

2.3. Capillary Electrophoresis

2.3.1. PCR Amplification with GlobalFiler

Samples analysed with CE-based methods were first amplified with the GlobalFiler PCR amplification kit on the Veriti 96-Well Thermal Cycler (both Thermo Fisher Scientific, USA) as per the manufacturer's recommendations. This involved a 25 µL total PCR reaction volume, with a maximum of 15 µL of DNA extract in each reaction. All extracts amplified had been quantitated with the Quantifiler Trio kit described in the previous section. Where the quantitation result (specifically, the result of the Small Autosomal target in the Quantifiler Trio kit) indicated a relatively strong sample, the extract was diluted so that 1 ng total of DNA was added to the PCR reaction. Where samples were weaker than this, the maximum of 15 µL of neat extract was added to the PCR. Dilutions of DNA extract for PCR were made in low TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) (Teknova, USA). The following PCR protocol was used for all reactions:

Table 2: PCR parameters used in amplification of samples with the GlobalFiler PCR amplification kit.

Stage	Temperature	Time	
Initial incubation	95°C	60 seconds	
Denature	94°C	10 seconds	} These steps repeated for 29 total PCR cycles
Anneal/extend	59°C	90 seconds	
Final extension	60°C	10 minutes	
Final hold	4°C	Indefinite	

PCR negative controls (i.e. samples with 15 µL of low TE buffer and no DNA input to the PCR) were run with all PCR batches processed. These showed no significant spurious DNA peaks in any batches.

The GlobalFiler PCR amplification kit contains 24 STR loci. A full list of these loci is given in the Appendix (Chapter 9).

2.3.2. Capillary Electrophoresis with 3500 Genetic Analyzer

PCR products resulting from the GlobalFiler PCR were analysed with the 3500 Genetic Analyzer running Data Collection v3.1 (Thermo Fisher Scientific, USA). POP-4 polymer and a 36 cm capillary array were used, as validated by the manufacturer for forensic applications. Samples were run with the default instrument protocol for GlobalFiler analysis, named HID36_POP4. This used injection parameters of 1.2 kV and 15 seconds. Samples were prepared in 96-well optical plates, with each well of the plate containing 9.5 µL of HiDi formamide and 0.5 µL of internal size standard (GeneScan-600 LIZ v2.0), (both Thermo Fisher Scientific, USA), with 1 µL of sample, control or allelic ladder added.

2.3.3. Data Analysis with GeneMapper IDX v1.5

Data generated on the 3500 Genetic Analyzer was genotyped with GeneMapper ID-X software v1.5 (Thermo Fisher Scientific, USA). For this analysis, AmpFLSTR_Panels_v5X and AmpFLSTR_Bins_v5X were used, along with a peak amplitude threshold (PAT) of 50 rfu. All other analysis settings were as per the manufacturer's recommendations.

2.4. Massively Parallel Sequencing

2.4.1. Library building with Ion Chef

Samples that underwent MPS based analysis were extracted and quantified with the same methods used for the CE-based workflow as described in Sections 2.2.2 and 2.2.3. Libraries were then built from the DNA extracts with the Ion Chef instrument (Thermo Fisher Scientific, USA) as per the manufacturer's recommendations. Torrent Suite v5.2.2 was installed on the Ion Chef for this work.

Five types of libraries were made in this work: a mixed forensic panel (containing STR and microhaplotype loci), an STR panel, an Ancestry SNP panel, an Identity SNP panel, and a Mitochondrial panel.

The full names of these panels were as follows: (All Thermo Fisher Scientific, USA)

- Precision ID GlobalFiler Mixture ID Panel
- Precision ID GlobalFiler NGS STR Panel v2
- Precision ID Ancestry Panel
- Precision ID Identity Panel
- Precision ID mtDNA Whole Genome Panel

2.4.1.1. Mixed Forensic Panel

The Precision ID GlobalFiler Mixture ID Panel is not yet commercially available and was used on early access release from Thermo Fisher Scientific, USA. The early access release of the kit was first available in 2016. This panel contains 68 loci (32 STR and 36 microhaplotype) and the amplicons in the panel range in size from 57 to 275 bp with an average size of 151 bp. A full list of the loci is given in the Appendix (Table 69 and Table 70).

2.4.1.2. STR Panel

The Precision ID GlobalFiler STR NGS Panel v2 was released in 2018. The earlier 'v1' version of the panel was released for the older PGM instrument from Thermo Fisher Scientific. The v2 panel used here is a new version of that same panel, optimised for the S5 sequencing instrument used in this work. This panel contains 36 loci and the amplicons in the panel range in size from 54 to 156 bp with an average size of 86 bp. A full list of the loci is given in the Appendix (Table 71).

2.4.1.3. Ancestry SNP Panel

The Precision ID Ancestry Panel was released in 2014 and contains 165 SNP loci. The amplicons in the panel range in size from 34 to 155 bp with an average size of 78 bp. A full list of the loci is given in the Appendix (Table 72).

2.4.1.4. Identity SNP Panel

The Precision ID Identity Panel was release in 2014 and contains 124 SNP loci. The amplicons in the panel range in size from 33 to 192 bp with an average size of 87 bp. A full list of these loci is given in the Appendix (Table 73).

2.4.1.5. Whole Mitochondrial Genome Panel

The Precision ID mtDNA Whole Genome Panel was released in 2015 and contains two primer pools, each containing 81 primer pairs. These primers generate PCR amplicons that average 163 bp in size and cover the entire 16,569 bp human mitochondrial genome in a tiled fashion. The amplicons overlap each other by an average of 11 bp. There are two primer pools, each containing alternate amplicons spaced around the genome, so that the primers do not interfere with each other during the PCR that initially amplifies the amplicons. This is shown in the following diagram (Figure 12).

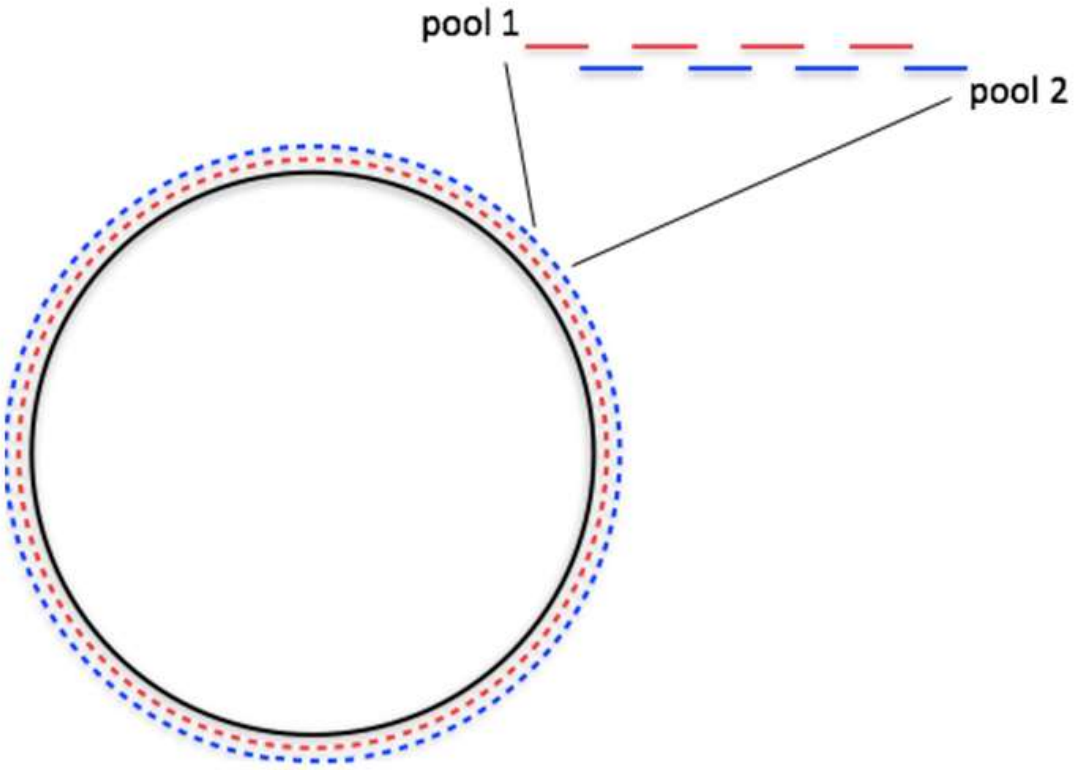


Figure 12: Schematic diagram showing how the tiled amplicons in the Precision ID mtDNA Whole Genome Panel are spaced around the mitochondrial genome (represented by the black circle). Primers for alternate amplicons are in each of the two pools so that there is no interference between the primers in the PCR.

Libraries were prepared for the four autosomal panels described above (i.e. for all except the mitochondrial panel) using 1 ng of DNA extract per library, as measured by the quantitation result (specifically, the result of the Small Autosomal target in the Quantifiler Trio kit). Where this indicated a relatively strong sample, the extract was diluted so that 1 ng total of DNA was added to the library reaction. Where samples were weaker than this, the maximum of 15 µL of neat extract was added to the reaction. Dilutions of DNA extract were made in low TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) (Teknova, USA). For the mitochondrial panel, the same procedure was followed except with a target input of 0.1ng of DNA, instead of 1ng.

The following tables contains the parameters that were used on the Ion Chef instrument for library building for each of the five panels. All are the manufacturer's recommended parameters.

Table 3: Instrument parameters used in Ion Chef library building runs.

Step	'Mixed' and STR panels	Ancestry and ID panels	mtDNA Panel
Number of primer pools	1	1	2
Target amplification cycles	24	27	27
Anneal and extension time	4 minutes	4 minutes	4 minutes

2.4.2. Templating with Ion Chef

The libraries made in Section 2.4.1 were then templated on the Ion Chef instrument. Before templating, libraries were quantified with the Ion Library TaqMan Quantitation Kit (Thermo Fisher Scientific, USA), and input library to the template step normalised to 30 pM, as per the user guide. Templating was done using the Ion S5 Precision ID Chef & Sequencing Kit (Thermo Fisher Scientific, USA) with either 520 or 530 model sequencing chips.

2.4.3. Sequencing with S5xl

The templated chips produced in Section 2.4.2 were then sequenced with the S5xl as per the manufacturer's instructions using the Ion S5 Precision ID Chef & Sequencing Kit.

To test for any possibility of unexpected contaminating DNA in the S5xl workflow, one full blank chip (with low TE buffer used in place of library input to the templating step) was sequenced. This showed no spurious reads. Negative control libraries (a library prepared with low TE buffer used in place of DNA extract) were run as part of the sensitivity study described in Section 3.2. Details of this are described in that section. Generally, no reads that could be attributed to contamination were observed in the MPS runs performed in this work.

2.4.4. Data Analysis with Torrent Suite v5.2.2 and Converge v2.1

Primary analysis of sequencing data was performed with Torrent Suite v5.2.2 (Thermo Fisher Scientific, USA). This software controls the Ion Chef and S5xl sequencer and converts the raw data generated in the sequencing run to sequence data. The result of this analysis is the BAM file. BAM files in this work were aligned to the hg19 human genome

reference using the default TMAP alignment software that is built into Torrent Suite. All primary analysis parameters used the default parameters recommended by the manufacturer.

Secondary analysis is the conversion of the BAM file to a genotype. For SNP analysis this was done using the HID_SNP_Genotyper v5.2.2 plugin to Torrent Suite v5.2.2 (Thermo Fisher Scientific, USA). Some of the parameters used for this software must be set by the user. These include:

- Minimum coverage: the minimum total number of reads that must be present at a locus for an allele to be called.
- Minimum allele frequency: the minimum proportion of the total number of reads at a locus that must match a given allele for that allele to be called.
- Minimum coverage either strand: the minimum number of either forward or reverse reads that must be present at a locus for an allele to be called.
- Maximum strand bias: a measure of the strand bias at a locus, that is, the number of forward compared to reverse reads. Measured as the maximum of the number of forward and reverse reads, divided by the sum of the number of forward and reverse reads. So a perfectly balanced locus (same number of forward and reverse reads) has a strand bias of 0.5 and a completely imbalanced locus (all forward or all reverse reads) has strand bias of 1.

For the majority of the work performed here, the manufacturer's default recommended values were used. These values, along with any variations from the defaults that were used, are described in the relevant chapters.

For STR, microhaplotype, and mitochondrial work, secondary analysis was performed using Converge v2.1 software (Thermo Fisher Scientific, USA). Some of the parameters used for this software must also be set by the user. These include the minimum coverage and strand bias metrics, as listed above, as well as the following parameters:

For mitochondrial analysis:

- Minimum number of reads to call: the minimum number of reads required to be present to call a variant.

- Minimum coverage to mark region: the minimum number of reads required to be present at any region in the data. Below this number the region is marked for the user's attention as an area of low coverage.
- Minimum coverage % of average: the minimum coverage required to be present as a percentage of the overall mean coverage. Below this number the region is marked for the user's attention as an area of low coverage.
- Threshold for confirmed/likely/possible call: Minimum frequencies for annotating a possible variant as confirmed/likely/possible, expressed as the percentage of the variants seen out of the total number of reads at the base in question.

For STR analysis:

- Target / hotspot file: manufacturer provided files that define the properties of the loci in the kit being analysed.
- STR flank length: the length of the flank area of the STR locus in question in bases.
- STR flank tolerance: how many mismatches are allowed in matching the STR flank region to the reference genome, to account for possible sequence variation in the flank region of the sample being analysed.
- STR analytical threshold: the number of reads required to call an allele at a locus, expressed as a percentage of the total reads at the locus.
- STR stochastic threshold: the number of reads required to call an allele at a locus, and flag it in the software as 'above stochastic threshold'. Intended for use to identify single alleles at loci that are homozygous, as opposed to single alleles where other alleles have dropped out due to low DNA input.
- STR stutter ratio: the percentage of reads below which a potential allele will not be called as an allele if it sits in a 'stutter position' relative to a larger allele (typically four bases smaller than the large allele). Expressed as the number reads in the small 'allele' as a percentage of reads in the large allele.

As above, for the majority of the work performed here, the manufacturer's default recommended values were used. These values, along with any variations from the defaults that were used, are described in the relevant chapters.

2.5. Statistics and Kinship Analysis with Familias v3.2.3

Where random match probabilities are reported for the profiles observed, these were calculated using the Hardy-Weinberg formulae of p^2 and 2pq with no inbreeding correction. Allele frequencies were those supplied by the manufacturers of the secondary analysis software in question: GeneMapper ID-X v1.5 for CE data and HID_SNP_Genotyper v5.2.2 for SNP data (both Thermo Fisher Scientific, USA). Specifically, the allele frequencies used are the 'ABGlobalFilerPopulationDatabase' from GeneMapper ID-X v1.5 (data from Thermo Fisher Scientific) and the '1000 genomes' data set from HID_SNP_Genotyper v5.2.2 (data from 1000 Genomes, www.internationalgenome.org).

Kinship analysis was performed with Familias v3.2.3 (Kling *et al.* 2014). Allele frequencies used for this analysis were the same as referenced above for the random match probabilities. The specific configuration for this analysis is described in Chapter 6.

Chapter 3:

Evaluation of the Performance of Massively Parallel Sequencing

3. Evaluation of the Performance of Massively Parallel Sequencing

3.1. Introduction

The first section of this work tested the general performance in forensic DNA analysis of massively parallel sequencing compared to capillary-based methods. This performance testing had several components to it, as 'performance' can be broadly defined and is not focussed on only one aspect of how an analysis method behaves. As such, in this work the performance of the MPS and CE methods was assessed by examining: the sensitivity achieved with small amounts of input DNA, the performance in the presence of inhibited DNA, the performance on mixed samples, and the performance on non-probative 'casework style' samples. These are all areas of interest to forensic DNA analysts, and have been the topic of multiple past evaluations of forensic DNA technology, although few, if any, of these previous tests have directly compared MPS to CE using these criteria.

For each of the four types of comparison named above, MPS methods were compared to CE methods on the same set of samples, prepared and extracted with the methods described in Chapter 2. A variety of MPS methods were examined, covering Identity SNP, mitochondrial, microhaplotype and STR applications, as discussed in Chapter 1 and described in Chapter 2.

3.2. Sensitivity

3.2.1. Methods – Sample set up

The sensitivity of two MPS assays were assessed in this work, the Precision ID Identity Panel and the Precision ID mtDNA Whole Genome Panel. These were run on a dilution series of control DNA, with the same dilution series also processed with a CE-based assay, the GlobalFiler PCR amplification kit. The results of the MPS and CE assays were compared to assess the utility of the MPS against the CE methods.

The following dilution series of DNA was prepared using 0.1 ng/µL Control DNA 007 (Thermo Fisher Scientific, USA). Dilutions were made in low TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) (Teknova, USA):

Table 4: Control DNA dilutions tested for sensitivity with MPS and CE-based assays. The samples are a two-fold dilution series, so that the DNA concentration halves with each successive sample. The total DNA in 15 µL is shown as this is the volume of the reaction for both the MPS and CE assays tested.

Sample number	DNA concentration (pg/µL)	DNA in 15 µL reaction (pg)
1	66.7	1000
2	33.3	500
3	16.7	250
4	8.3	125
5	4.2	62.5
6	2.1	31.3
7	1.0	15.6
8	0.5	7.8
9	0.3	3.9
10	0.1	2.0
Negative	0	0

3.2.2. Methods – Precision ID Identity Panel

This dilution series was then processed in duplicate with the Precision ID Identity Panel as described in Section 2.4.1.4. Parameters of the HID_SNP_Genotyper v5.2.2 analysis used to interpret the results of this data were as follows (all manufacturer's default values):

Table 5: Analysis parameters used in HID_SNP_Genotyper v5.2.2 analysis for Precision ID Identity Panel sensitivity analysis. All are the manufacturer's recommended default parameters.

Parameter	Value used
Minimum allele frequency	0.1
Minimum coverage	6
Minimum coverage either strand	0
Maximum strand bias	1
Trim reads	true

3.2.3. Sensitivity Results – Precision ID Identity Panel

Results of the analysis showed high coverage for the stronger samples (average locus coverage of 10,121 for the 1000 pg sample), with easily above threshold coverage being observed for even the weakest samples (average locus coverage of 823 for the 2 pg sample). Full profiles were observed down to the 125 pg sample, with the 62.5 pg samples being the first that showed dropout (i.e. a minimum locus coverage of zero). This is shown in the following table:

Table 6: Coverage results for Precision ID Identity Panel sensitivity analysis. Sample numbers and DNA input correspond to those shown in Table 4. Coverage results were averaged across two duplicates for each sample, with the maximum, minimum, mean and standard deviation of coverage for the samples are shown. Minimum coverage of zero indicates that at least one locus dropped out in that sample.

Sample Number	DNA in 15 µL reaction (pg)	Maximum coverage	Minimum coverage	Mean coverage	Std Dev of coverage
1	1000	30719	646	10121	6414
2	500	31490	431	10460	6566
3	250	32216	817	10924	6495
4	125	31098	601	10742	6558
5	62.5	32561	0	9519	6741
6	31.3	27568	0	6590	5445
7	15.6	20500	0	3743	4086
8	7.8	12051	0	1986	2323
9	3.9	12047	0	1841	2384
10	2.0	9654	0	823	1497
Neg	0	1248	0	8	-

The results were then assessed as to how much of the complete profile was present (100% in the case of a full profile with all alleles present, lower percentages for partial profiles). Random match probabilities for the observed profiles were then calculated as described in Section 2.5. All profiles showed a random match probability that would be of use in a practical case, with even the weakest sample (2.0 pg) having a match probability in the order of 10^{-18}. This is shown in the following table:

Table 7: Genotype results for Precision ID Identity Panel sensitivity analysis. Sample numbers and DNA input correspond to those shown in Table 4. There are 124 loci in the panel, 90 autosomal and 34 Y-loci, so 214 alleles represents a full profile. Results were averaged across two duplicates for each sample. The Random Match Probability (RMP) is that calculated by the HID_SNP_Genotyper v5.2.2 software used in the analysis, with allele frequencies as described in Section 2.5

Sample Number	DNA in 15 µL reaction (pg)	Alleles observed	% of profile observed	RMP of profile observed
1	1000	214	100	3.39×10^{-38}
2	500	214	100	3.39×10^{-38}
3	250	214	100	3.39×10^{-38}
4	125	214	100	3.39×10^{-38}
5	62.5	213.5	99.8	2.03×10^{-38}
6	31.3	211	98.6	1.99×10^{-38}
7	15.6	190	88.8	2.47×10^{-37}
8	7.8	162.5	75.9	3.00×10^{-33}
9	3.9	131	61.2	1.04×10^{-33}
10	2.0	79.5	37.1	1.05×10^{-18}
Neg	0	0.5	0.2	-

The results for the Precision ID Identity Panel sensitivity samples were then assessed and the profiles categorised as to the coverage that was observed for each allele. The strongest profiles (1000 pg to 31.3 pg) had no alleles under 500 coverage. Even the weakest samples, despite having some dropped out alleles (coverage = 0), had no above threshold alleles below 50 coverage, with all but one allele in the test being above 100 coverage. This is shown in the table below:

Table 8: Banded allele coverage results for Precision ID Identity Panel sensitivity analysis. Sample numbers correspond to those shown in Table 4. Columns show the number of alleles in the sample that fell in each coverage band. For example, the number of alleles with coverage of 101 to 500 are shown in the 101-500 column. Coverage of zero is equivalent to allele dropout and is included in the 0-6 column. All rows sum to 214, the number of alleles in a full profile. Results were averaged across two duplicates for each sample. Analysis was performed with an analytical threshold (also known as minimum coverage) of 6, as seen in Table 5.

Sample No.	Banded allele coverage								Total
	0 - 6	7 - 20	21 - 50	51 - 100	101 - 500	501 - 1000	1001 - 5000	5000+	
1	0	0	0	0	2	4	93	115	214
2	0	0	0	0	2	2	91	119	214
3	0	0	0	0	2	2	75	135	214
4	0	0	0	0	0	6	86	122	214
5	0	0	0	0	2	4	105	103	214
6	0	0	0	0	4	13	137	60	214
7	2	0	0	2	33	29	121	27	214
8	14	0	0	4	57	50	88	1	214
9	46	0	2	0	35	43	86	2	214
10	89	0	0	10	39	34	42	0	214

Results were then plotted on a chart of coverage against the 124 loci in the panel. The relative strength of each of the ten samples can be observed, with all samples having a six-fold or more range from the highest to lowest covered loci. Samples 6 to 10, where some dropout was observed, can be seen towards the bottom of the chart, with the dropped-out loci on the right with coverage at zero.

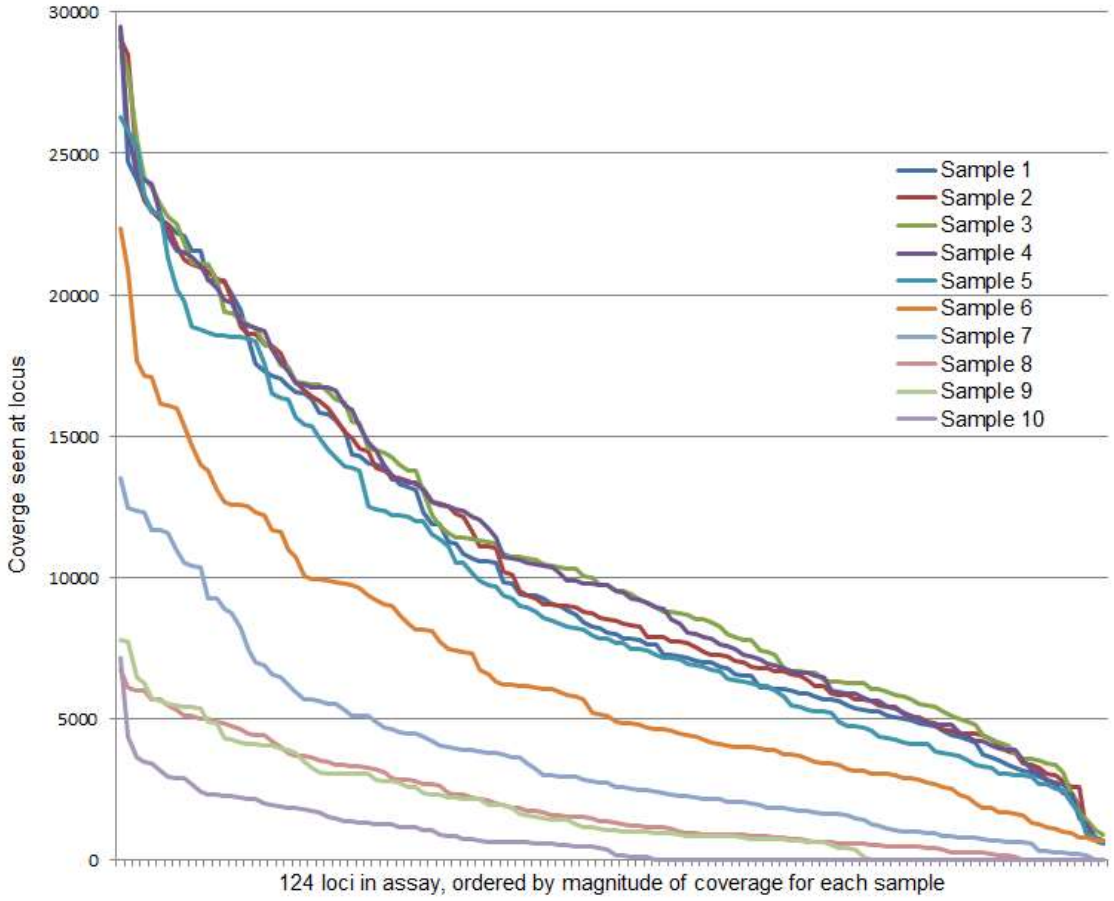


Figure 13: Chart of coverage observed at each locus in Precision ID Identity Panel sensitivity analysis, for each of the ten samples analysed. Sample numbers correspond to those shown in Table 4.The horizontal axis displays each of the 124 loci in the multiplex, sorted by magnitude of coverage for each sample. The vertical axis measures the coverage seen at each locus for the ten samples, averaged across two duplicates.

Next, results were plotted with coverage against the 124 loci, this time averaged across the entire experiment. A similar distribution of coverage as in the previous chart can be seen, this time with the detail of each locus visible. Three loci: rs4141886, rs1109037, and M479 were noticeably lower than the other loci across the experiment.

Loci in the chart are coloured blue for autosomal loci and red for Y-chromosome loci. It can also be observed that the Y-chromosome loci generally tended to be in the bottom half of the range of loci when ordered by coverage, likely due to the haploid nature of these loci.

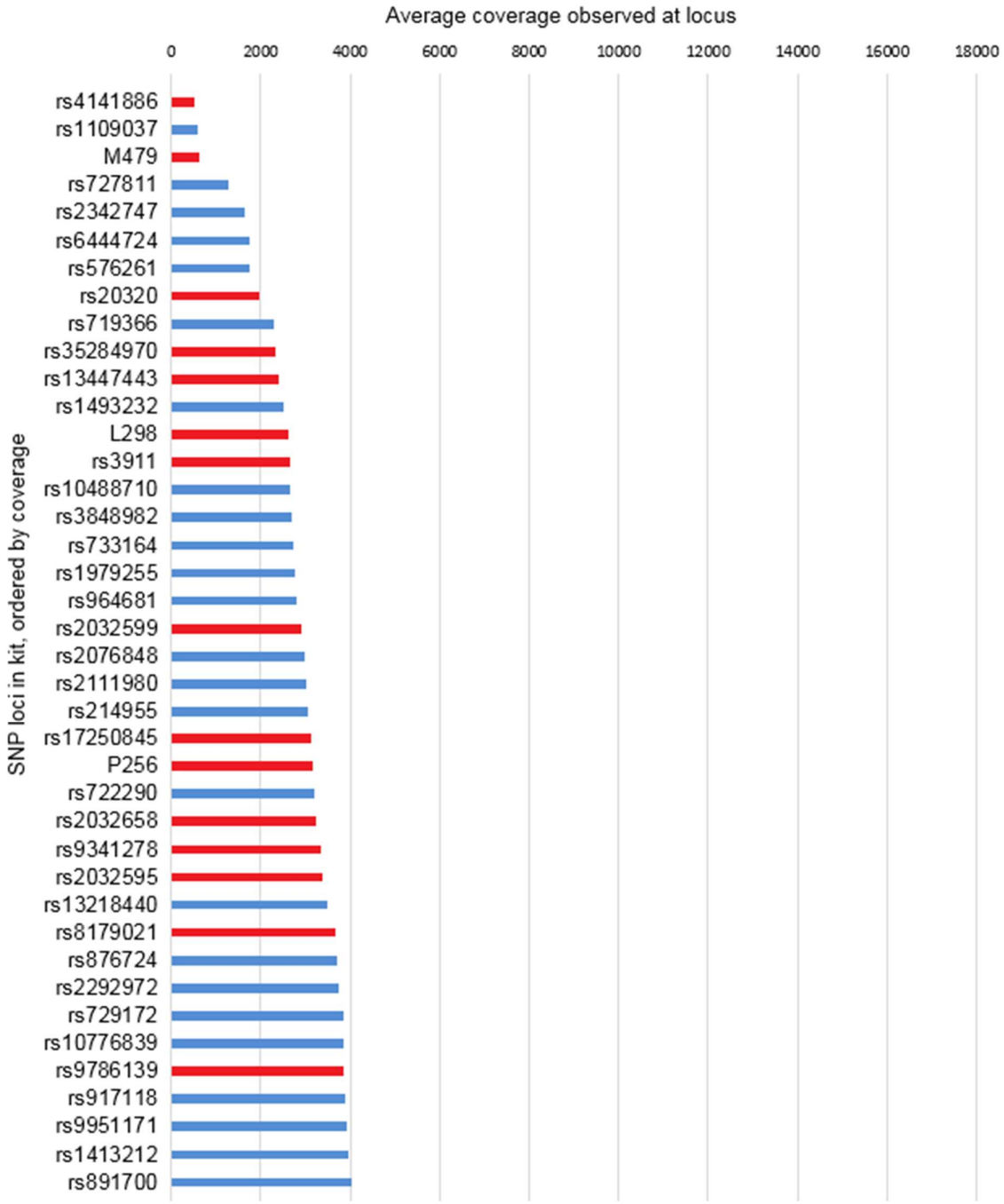


Figure 14: Chart of coverage observed at each locus in Precision ID Identity Panel sensitivity analysis, averaged across all samples in sensitivity study. The horizontal axis shows the coverage seen at each locus. The vertical axis shows the loci in the kit. Autosomal loci are shown in blue, Y-chromosome loci in red. Data is split across Figure 14, Figure 15 and Figure 16 to show all 124 loci in the kit. Each of these three figures has the same horizontal scale.

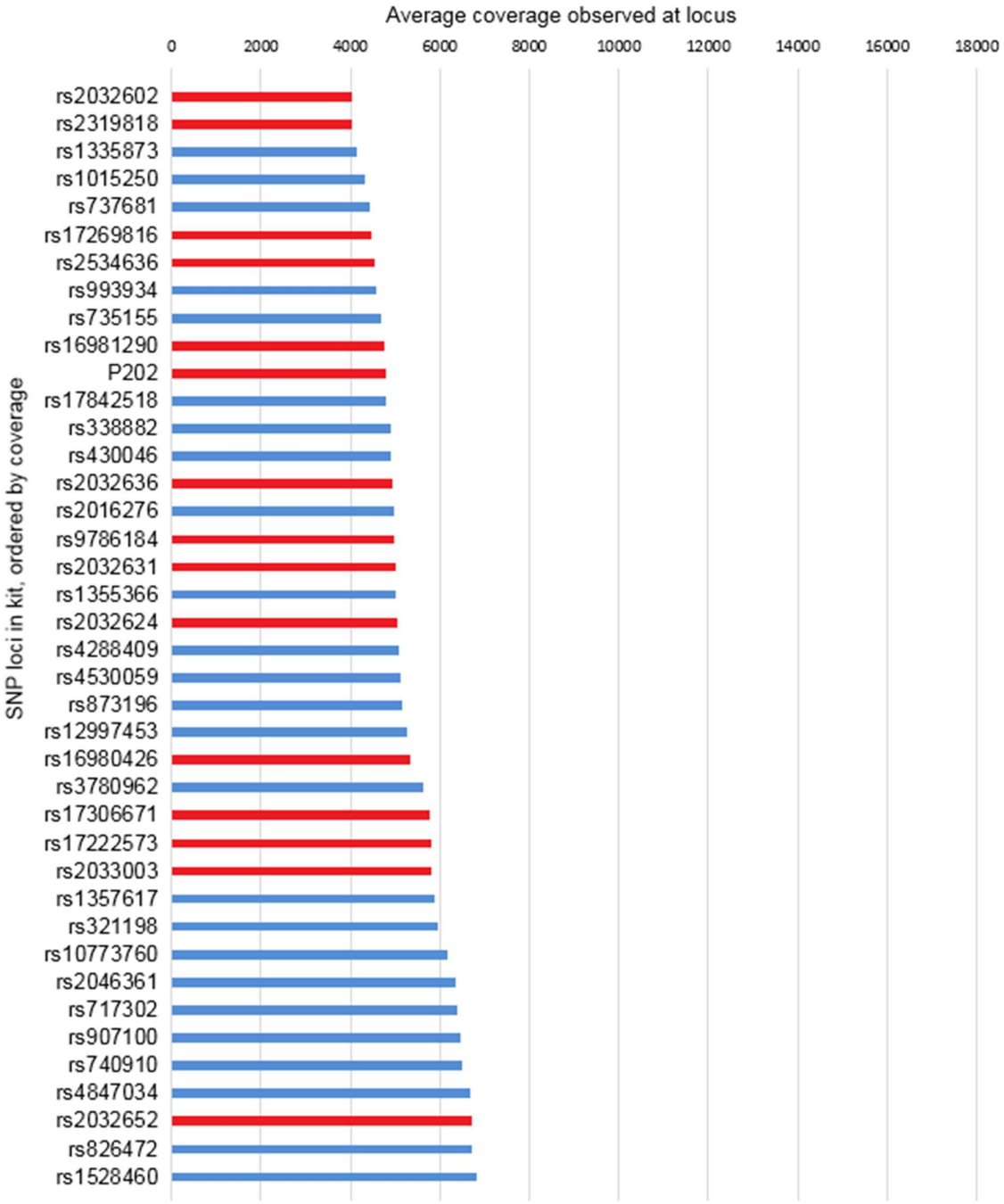


Figure 15: Chart of coverage observed at each locus in Precision ID Identity Panel sensitivity analysis, averaged across all samples in sensitivity study. The horizontal axis shows the coverage seen at each locus. The vertical axis shows the loci in the kit. Autosomal loci are shown in blue, Y-chromosome loci in red. Data is split across Figure 14, Figure 15 and Figure 16 to show all 124 loci in the kit. Each of these three figures has the same horizontal scale.

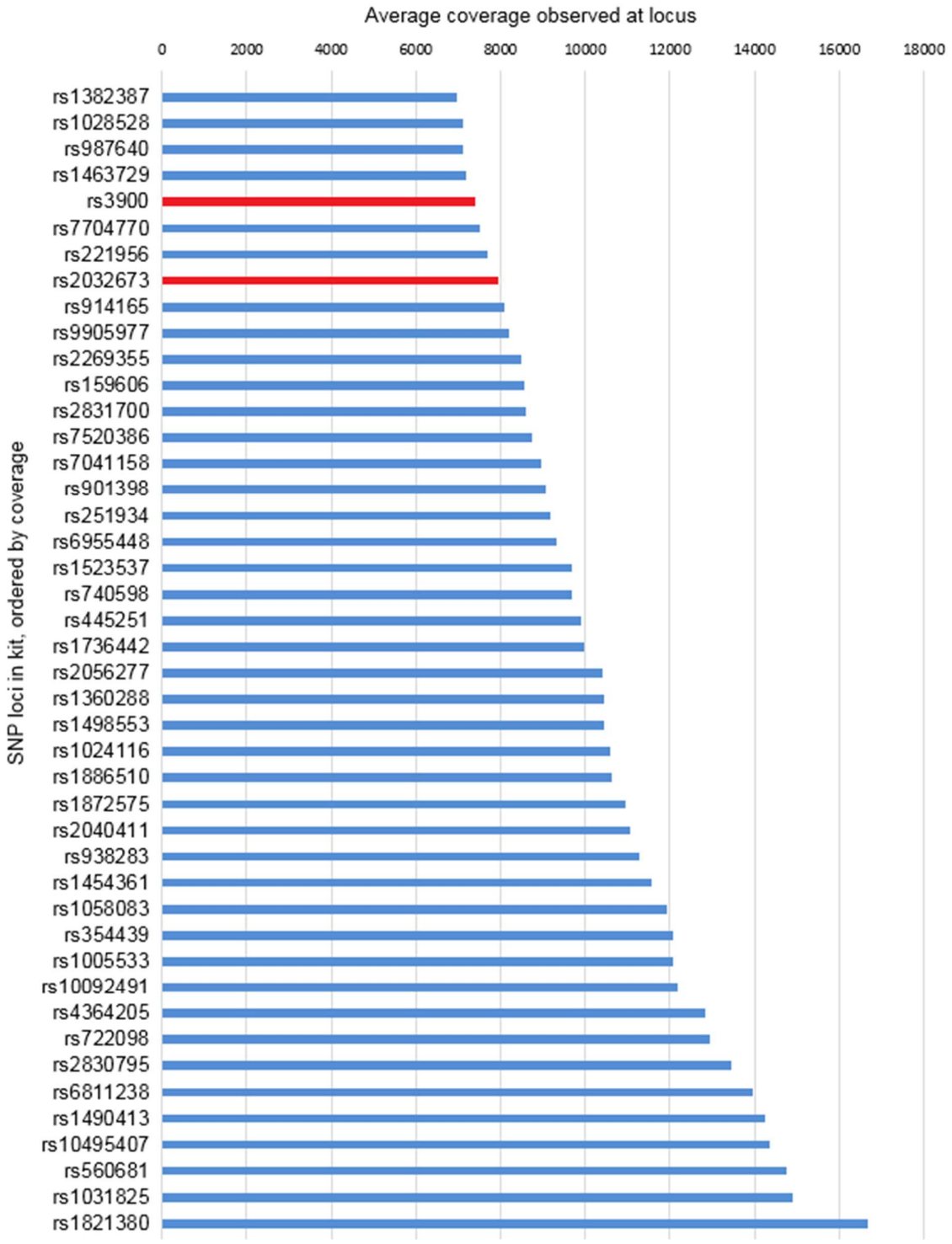


Figure 16: Chart of coverage observed at each locus in Precision ID Identity Panel sensitivity analysis, averaged across all samples in sensitivity study. The horizontal axis shows the coverage seen at each locus. The vertical axis shows the loci in the kit. Autosomal loci are shown in blue, Y-chromosome loci in red. Data is split across Figure 14, Figure 15 and Figure 16 to show all 124 loci in the kit. Each of these three figures has the same horizontal scale.

3.2.4. Methods – Precision ID mtDNA Whole Genome Panel

Samples 3 to 10 of the dilution series were then processed in duplicate with the Precision ID mtDNA Whole Genome Panel as described in Section 2.4.1.5. Samples 1 and 2 were not processed as they were too strong – 1000 pg and 500 pg of genomic DNA is considered by the Precision ID mtDNA Whole Genome Panel user guide to be approximately equivalent to 100 ng and 50 ng of mtDNA, and as such it was not useful to analyse these samples in a sensitivity study.

Parameters of the Converge v2.1 mitochondrial analysis were as follows (all manufacturer's default values):

Table 9: Analysis parameters used in Converge v2.1 mitochondrial analysis for Precision ID mtDNA Whole Genome Panel sensitivity analysis. All are the manufacturer's recommended default parameters.

Parameter	Value used
Minimum coverage	20
Minimum number of reads to call	20
Minimum coverage to mark region	20
Threshold for certain call	96%
Threshold for likely call	90%
Threshold for possible call	80%
Minimum coverage % of average	10%
Strand bias threshold	0.9

3.2.5. Sensitivity Results – Precision ID mtDNA Whole Genome Panel

Results of this analysis were examined as to the maximum, minimum, mean, median and standard deviation coverage shown. All samples showed complete coverage, with no dropped out positions observed – the lowest base in the entire experiment was covered 13 times in the sequencing. Interestingly this was in the strongest sample (sample 3 with 250 pg input). All samples showed a similar minimum of coverage (ranging from 13 to 55), but the strongest samples showed larger maximum coverage – the highest being sample 5 (62.5 pg of input DNA) with a maximum of 37,122. It is of note that the strongest two samples (samples 3 and 4, 250 pg and 125 pg of input respectively) had lower maximum coverage than sample 5. These results are shown in the following table:

Table 10: Coverage results for Precision ID mtDNA Whole Genome Panel sensitivity analysis. Sample numbers and DNA input correspond to those shown in Table 4. Coverage results were averaged across two duplicates for each sample, with the maximum, minimum, mean, median and standard deviation of all coverage across the mtDNA genome shown

Sample Number	DNA in 15 µL (pg)	Max. coverage	Min. coverage	Mean coverage	Median coverage	Std Dev coverage	%CV
3	250	9482	13	2136	1869	1371	0.64
4	125	33566	43	6690	5917	4243	0.63
5	62.5	37122	55	7476	6584	4993	0.67
6	31.3	24062	31	5453	5079	3503	0.64
7	15.6	20511	48	5199	4904	3063	0.59
8	7.8	11128	27	2901	2752	1676	0.58
9	3.9	12098	34	3064	2899	1813	0.59
10	2.0	12203	36	3275	2786	2162	0.66

Result were then examined with similar coverage metrics as in the previous table, but this time only for the variants that were seen in the sequence, rather than at every base as previous. A similar pattern was seen, with sample 5 (62.5 pg) having the highest maximum coverage. All variants expected in the profile were observed in every sample in the experiment, 36 variants from the revised Cambridge Reference Sequence (rCRS) in total.

Table 11: Genotype results for Precision ID mtDNA Whole Genome Panel sensitivity analysis. Sample numbers and DNA input correspond to those shown in Table 4. Coverage results were averaged across two duplicates for each sample, with the maximum, minimum, mean and standard deviation of coverage of the variants detected shown. Note that there are 36 variants in the control DNA used, this represents a full profile

Sample Number	DNA in 15 µL (pg)	Variants detected	% of variants detected	Max. variant coverage	Min. variant coverage	Mean variant coverage	Std. dev. variant coverage
3	250	36	100%	5342	295	2162	1199
4	125	36	100%	18351	127	6125	4051
5	62.5	36	100%	20948	871	7080	4756
6	31.3	36	100%	14811	130	5247	3709
7	15.6	36	100%	11280	88	4886	2720
8	7.8	36	100%	5835	325	2705	1411
9	3.9	36	100%	5935	48	2840	1479
10	2.0	36	100%	3202	5	1424	949

The results from the previous two tables were then plotted on a chart of coverage against sample number, where the pattern of maximum coverage in sample 5 (62.5 pg) can be clearly seen both for total coverage and for variant (allele) coverage.

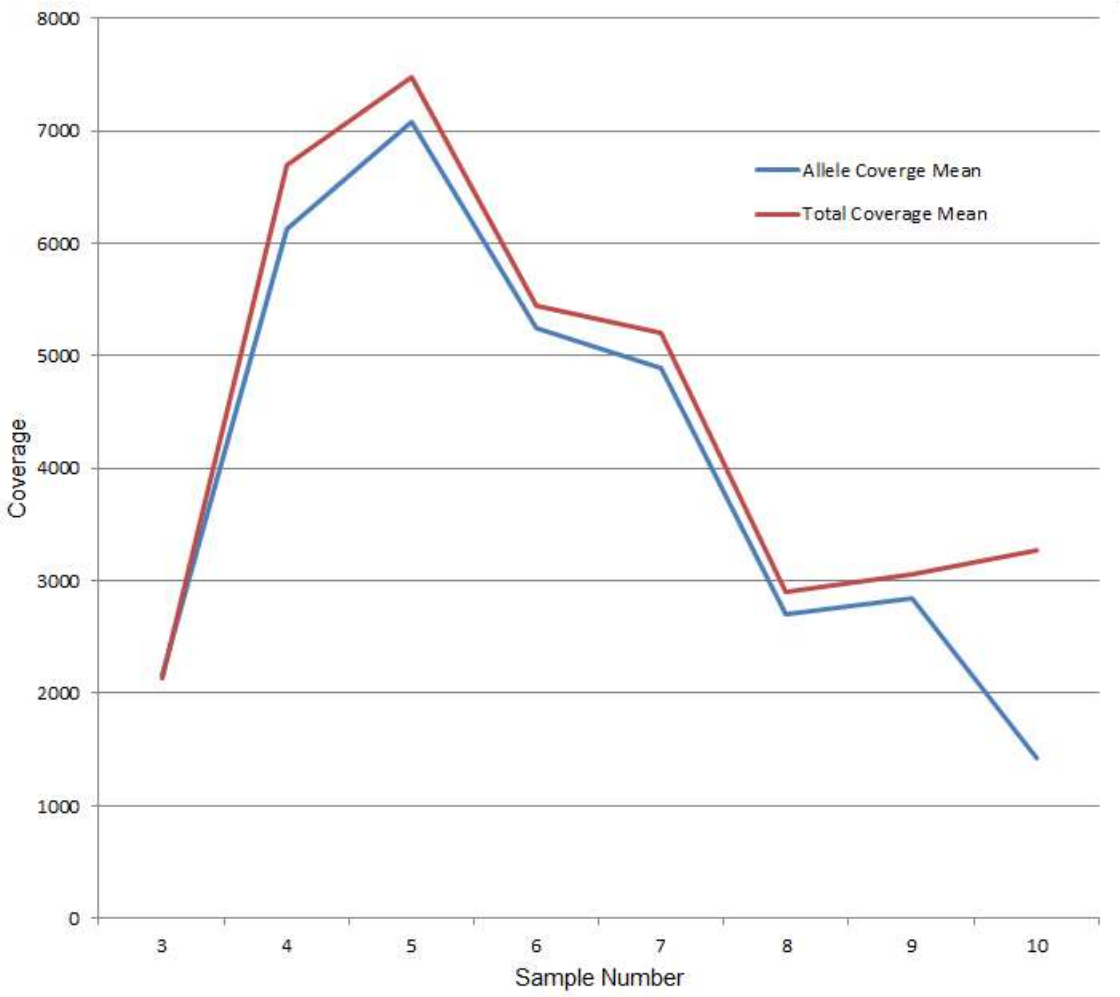


Figure 17: Chart of mean coverage observed in each sample in Precision ID mtDNA Whole Genome Panel sensitivity analysis. Sample numbers correspond to those shown in Table 4. Results are averaged across two duplicates for each sample. The red line denotes the mean of all coverage observed in the sample, while the blue line denotes mean coverage at the variants.

Coverage for the samples tested with the Precision ID mtDNA Whole Genome Panel was optimal at around the amount of DNA in sample 4 to sample 7 (Figure 17). After this, with decreasing amounts of DNA the coverage became increasingly lower, but coverage was also significantly lower for the higher amount of input DNA in sample 3. This was found to be due to the effect of amplicons in the strongest library unexpectedly joining together to form so-called ‘super-amplicons’, which are sequenced correctly, but due to the random nature of

their joining together, do not align to the reference genome, and are thus discarded by the software. As a result, samples with more than the recommended amount of input DNA can have lower coverage. This effect is visible in the Read Length Histogram generated by the software for the MPS run. This is a chart of read length against frequency, and from which the distribution of the length of the reads in the run can be visualised. Here, the super-amplicons are clearly visible at the right of the Read Length Histogram for the sample with 250 pg of input DNA, but are not present in the Read Length Histogram for the sample with 31.3 pg of input DNA (Figure 18 and Figure 19).

The way in which the super amplicons do not align to the reference genome can be seen below in Figure 20. This is a chart of all reads in the run, again arranged by read length and colour coded as to whether they have or have not aligned to the reference. All reads of approximately 150 bases and over, i.e. the super amplicons, do not align to the reference.

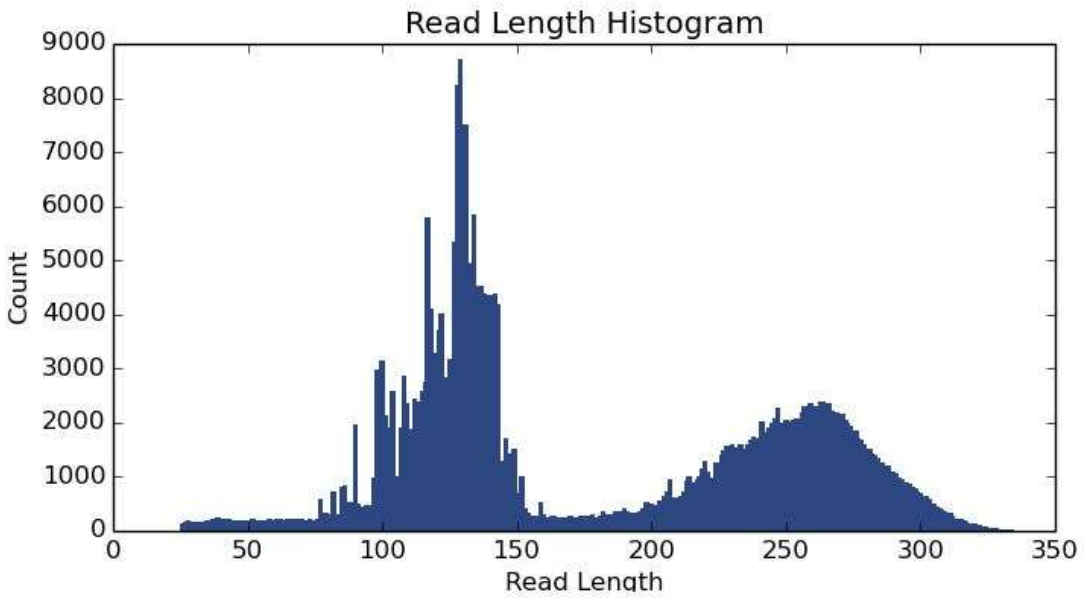


Figure 18: Read length histogram from MPS run of Precision ID mtDNA Whole Genome Panel for a sample with more than the manufacturer's recommended amount of input DNA (Sample 3 from Table 4). The jagged bars around 100 to 150bp are the expected mtDNA amplicons. The hump around 200 bp to 300 bp is comprised of super-amplicons formed due to the excess of input.

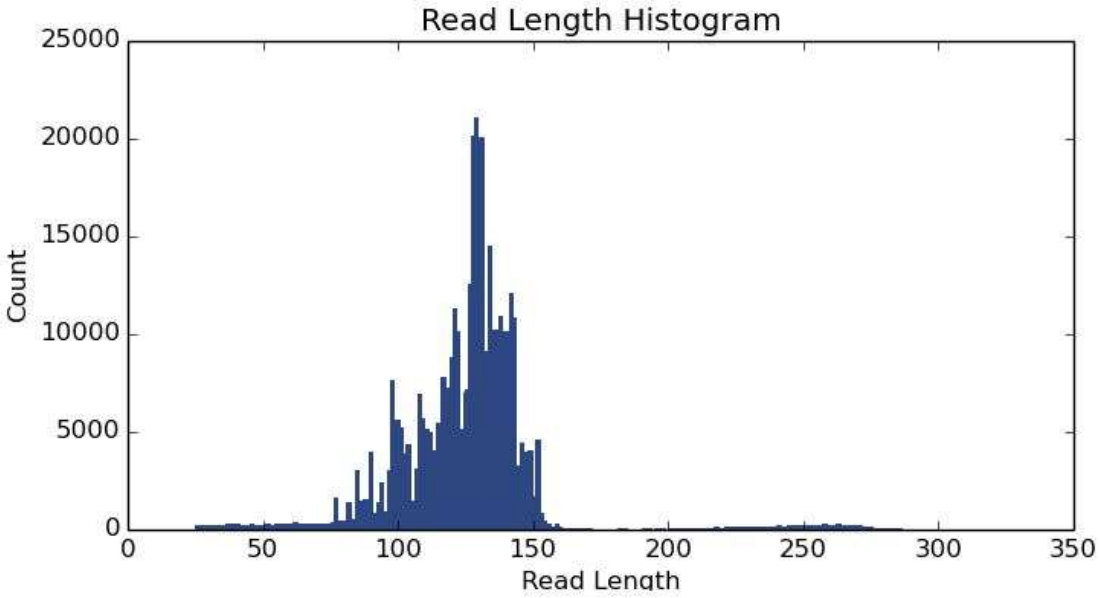


Figure 19: Read length histogram from MPS run of Precision ID mtDNA Whole Genome Panel for a sample with the manufacturer's recommended amount of input DNA (Sample 6 from Table 4). Note that the reads from 200 bp to 300 bp as seen in Figure 18 are missing from this histogram and only the true (100 bp to 150 bp) reads remain. This results in higher total coverage for the sample.

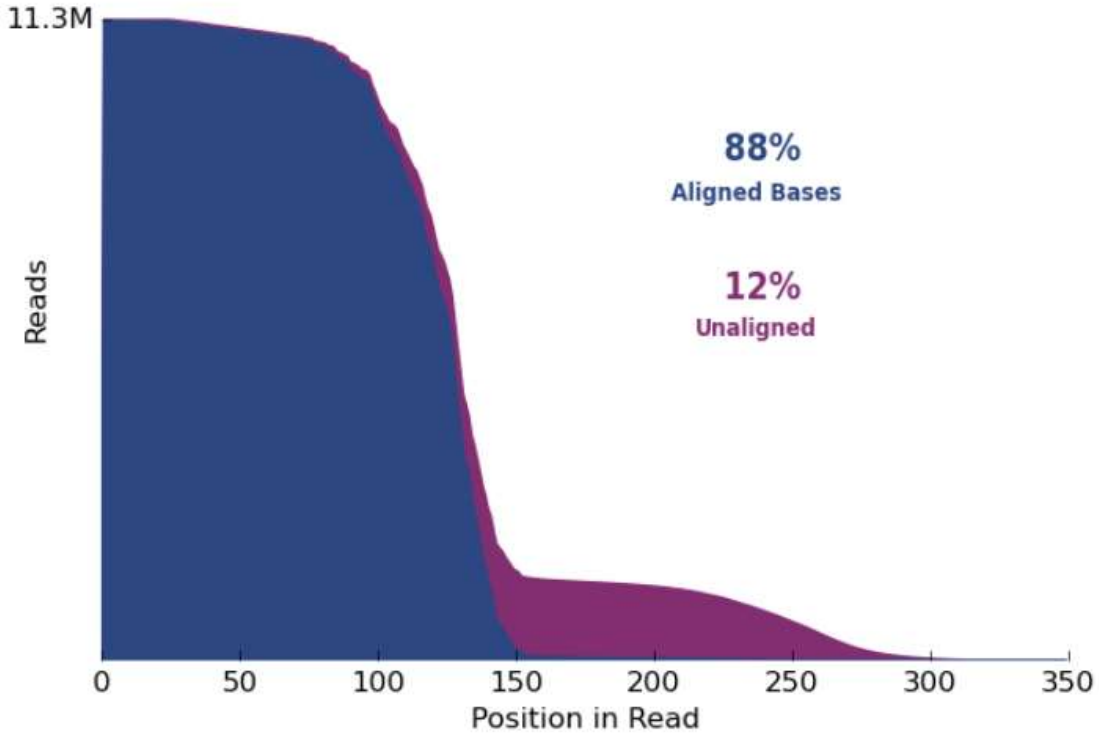


Figure 20: Alignment chart from MPS run of Precision ID mtDNA Whole Genome Panel. Alignment of all reads in the run is shown. Aligned reads are in blue, unaligned reads in purple. The large majority of reads over 150 bp do not align – these are the super amplicons shown above.

To further explore the nature of these super amplicons, the results from Sample 3 (the same sample shown in Figure 18 above) were further examined. All reads over 200 bp were extracted from the result file for this sample and stored in a new file. Any 'soft-clipped' bases in this new file were then extracted and put into a second new file. Soft-clipping is a bioinformatic technique where bases at the end of a read that do not map to the reference are taken out of the alignment, allowing the start of the read that does map to remain (Schröder *et al*. 2014). The non-aligning bases can either be removed entirely ('hard-clipped') or marked but kept in the file ('soft-clipped'). The default algorithm used by the TMAP alignment software in Torrent Suite v5.2.2 (see Section 2.4.4) soft-clips any bases that do not align at the end of a read. Examination of these soft-clipped bases was chosen as a method to reveal more about the nature of the super amplicons seen here.

The file containing the soft-clipped bases was realigned to the rCRS reference used in the original run. The file contained 148,667 total reads, of which 148,260 reads (99.73%) aligned to rCRS. In the original sample file with all reads present only 146 reads out of 365,954 total reads did not align at all. These results indicate that the non-aligning bases in the soft-clipped file were the ends of reads where front part of the reads aligned correctly, just that the second half did not align in that same position. Further, it shows that the second non-aligning half of these reads did align elsewhere on the reference, just not in the same position as the first half. As such, this analysis supports the theory that the super amplicons observed here are two 'good' amplicons, joined together at ligation due to the excess DNA in the sample. These super amplicons are sequenced correctly, but are unable to be fully aligned due to their not fitting correctly in any one place on the reference genome.

The frequency of the variants observed in the mitochondrial results was also analysed. The frequency of a variant is the number of variant reads detected as a percentage of the total of reads at that position. Low frequency values for a given variant could indicate the possible presence of heteroplasmy in the sample, something that must be accounted for in practical forensic mitochondrial analysis. Figure 21 shows variant frequencies for Sample 5 (62.5 pg input of DNA, this was the sample with the highest average coverage in the study) and Figure 22 shows variant frequencies for Sample 10 (2 pg of input of DNA, the sample with the lowest average coverage).

Figure 23 then shows the Sample 5 sequence data for the three variants with the lowest variant frequency in Figure 21 and Figure 22. These three variants were the same in both Sample 5 and Sample 10.

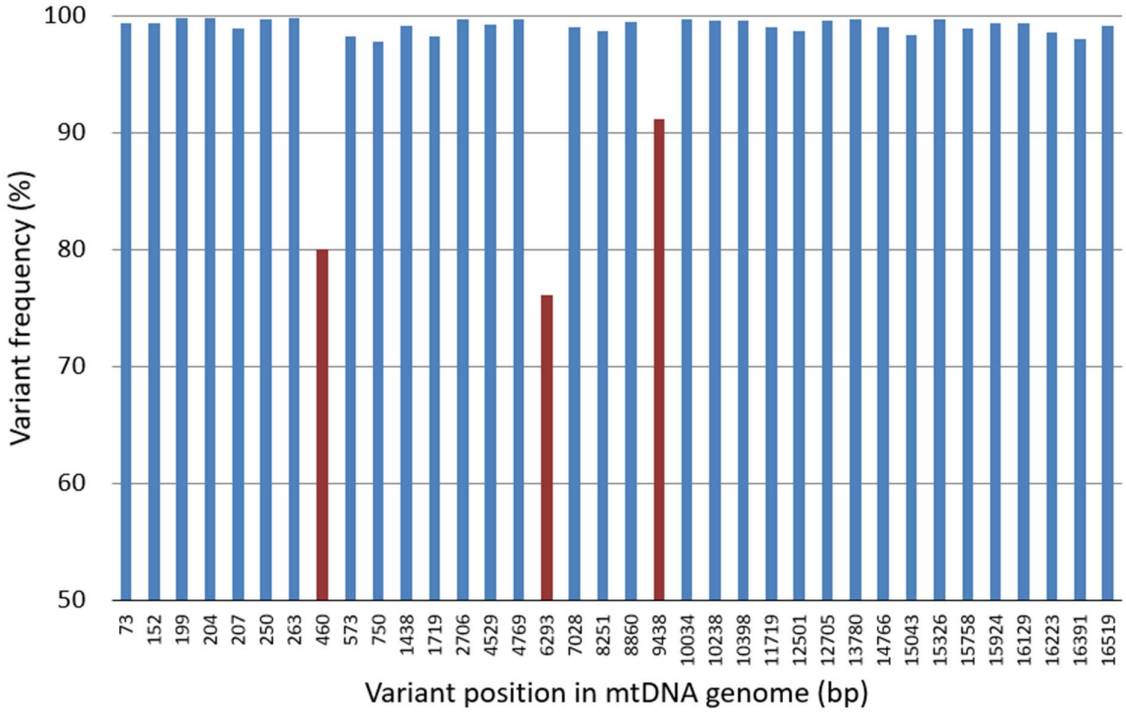


Figure 21: Variant frequency of all variants observed in Precision ID mtDNA Whole Genome Panel for Sample 5 (62.5 pg of input DNA, see Table 4). Variants below 95% frequency are highlighted in red.

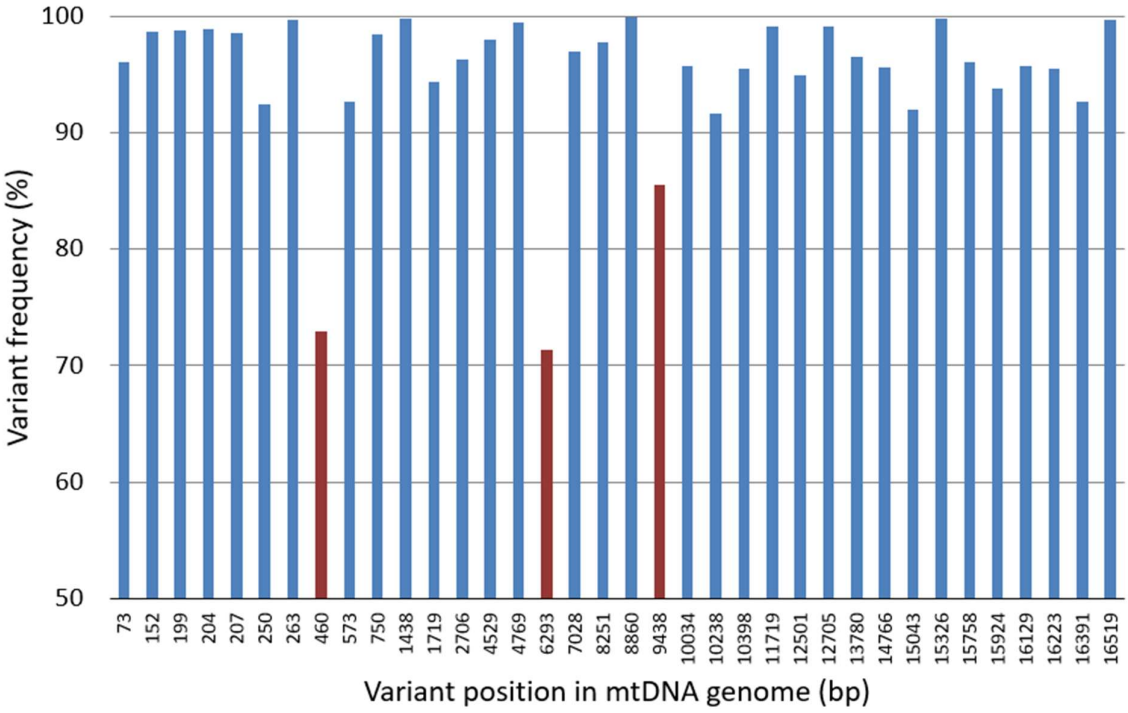


Figure 22: Variant frequency of all variants observed in Precision ID mtDNA Whole Genome Panel for Sample 10 (2 pg of input DNA, see Table 4). Variants highlighted red in Figure 21 also highlighted here in red.

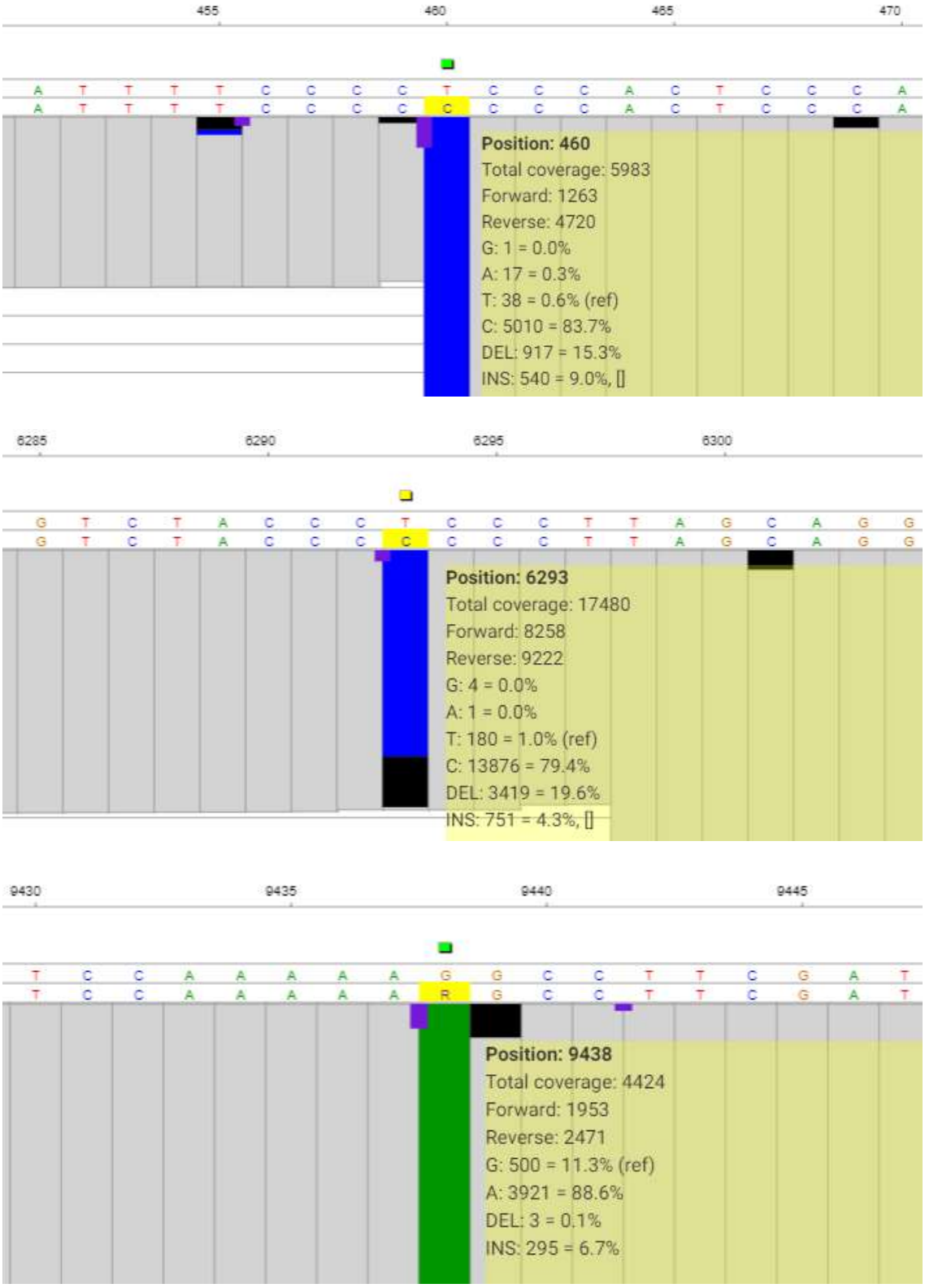


Figure 23: Sequence data from Converge v2.1 for Precision ID mtDNA Whole Genome Panel analysis of Sample 5 (62.5 pg of DNA input). Variant shown where variant frequency was under 90% (see Figure 21). From top to bottom: 460C, 6293C and 9438A variants.

3.2.6. Methods – Capillary Electrophoresis

The same dilution series was then processed in duplicate with the GlobalFiler PCR amplification kit, on the 3500xl Genetic Analyzer as described in Section 2.3.

3.2.7. Sensitivity Results – Capillary Electrophoresis

Results of the CE analysis were then analysed for the maximum, minimum and mean peak heights observed. Results showed a distinct trend from strongest to weakest, with the strongest sample (sample 1, 1000 pg input) having the highest maximum and mean peak height, and with all samples from 7 to 10 (15.6 pg to 2.0 pg) showing at least some locus drop out (i.e. minimum peak height equal to zero). This is shown in the following table:

Table 12: Peak height results for GlobalFiler CE-based sensitivity analysis. Sample numbers and DNA input correspond to those shown in Table 4. Results were averaged across two duplicates for each sample, with maximum, minimum, mean and standard deviation peak height shown. Minimum peak height of zero indicates that that at least one locus dropped out in that sample.

Sample Number	DNA in 15 µL reaction (pg)	Maximum peak height (rfu)	Minimum peak height (rfu)	Mean peak height (rfu)	Std. dev. peak height (rfu)
1	1000	7331	1706	3942	1064
2	500	4110	1837	2877	626
3	250	1787	539	1013	339
4	125	1116	138	546	216
5	62.5	597	81	217	116
6	31.3	331	32	143	60
7	15.6	327	0	97	65
8	7.8	156	0	73	36
9	3.9	98	0	65	21
10	2.0	40	0	40	0
Neg	0	0	0	0	0

As previously with the MPS results, the CE results were then assessed as to how much of the complete profile was present (100% in the case of a full profile with all loci present, lower percentages for partial profiles). Random match probabilities for the observed profiles were then calculated as described in Section 3.5. Samples 1 through 7 showed random match

probabilities that would likely be of use in a practical case, with match probabilities of at most 10^{-16} for all samples. Samples 8 to 10 (7.8 pg to 2.0 pg) however showed match probabilities that would likely not be acceptable in many forensic cases, with the weakest sample having a match probability of only 0.58. This is shown in the following table:

Table 13: Genotype results for GlobalFiler CE-based sensitivity analysis. Sample numbers and DNA input correspond to those shown in Table 4. Note that there are a maximum of 46 alleles in the control DNA used, this represents a full profile. Results were averaged across two duplicates for each sample. The Random Match Probability was calculated with method and allele frequencies as described in Section 2.5

Sample Number	DNA in 15µL reaction (pg)	Total alleles observed	% of profile observed	Random Match probability of profile observed
1	1000	46	100	2.03×10^{-30}
2	500	46	100	2.03×10^{-30}
3	250	46	100	2.03×10^{-30}
4	125	46	100	2.03×10^{-30}
5	62.5	44	95.7	5.61×10^{-29}
6	31.3	43	93.5	5.74×10^{-28}
7	15.6	28	60.9	1.99×10^{-16}
8	7.8	13	28.3	1.50×10^{-5}
9	3.9	8	17.4	1.55×10^{-4}
10	2.0	1	2.2	0.58
Neg	0	0	0	-

Results were then plotted on a chart of peak height against the 23 loci in the panel. The relative strength of each of the ten samples can be observed, with the strongest samples having an approximate two to three-fold difference from the strongest to weakest loci. Samples 7 to 10, where some dropout was observed, can be seen towards the bottom of the chart, with the dropped-out loci on the right with peak height at zero.

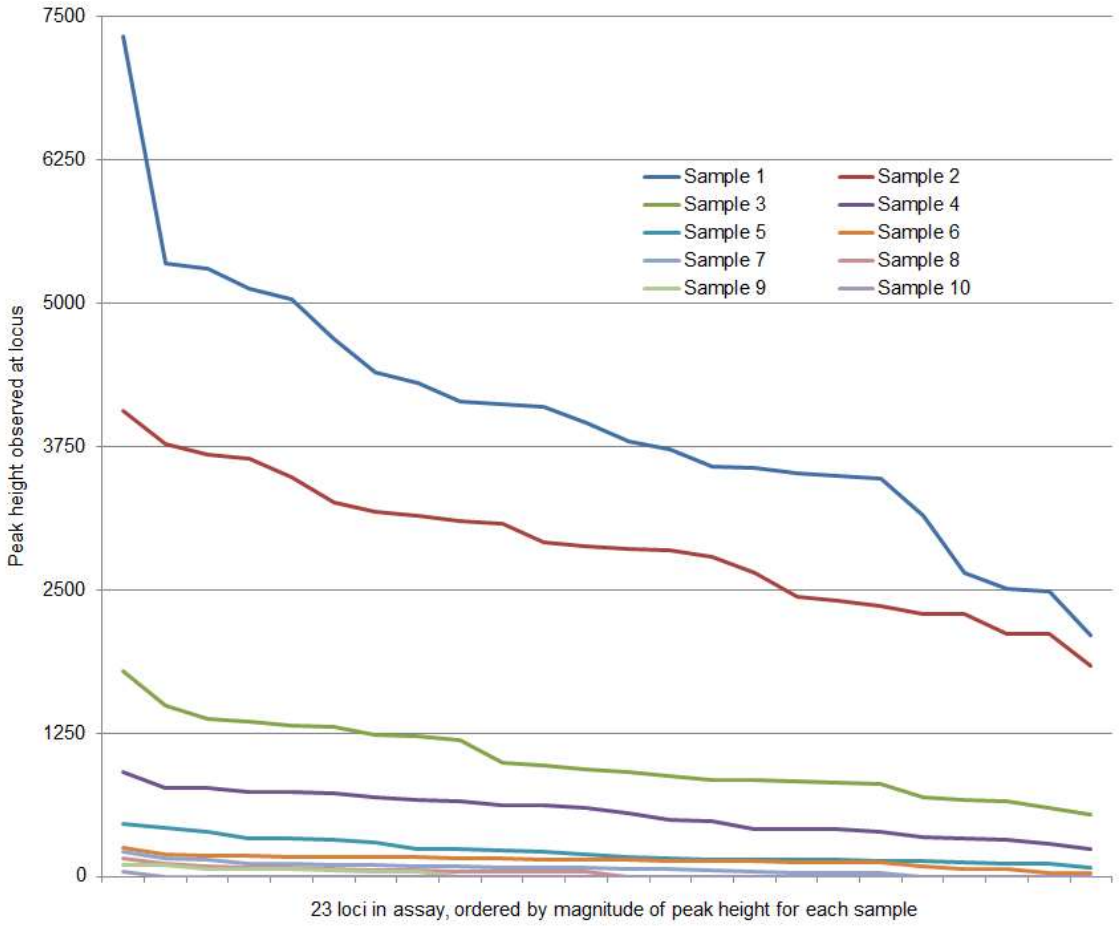


Figure 24: Chart of peak height observed at each locus in GlobalFiler CE-based sensitivity analysis for each of the ten samples analysed. Sample numbers correspond to those shown in Table 4. The horizontal axis displays each of the 23 loci in the multiplex, sorted by magnitude of peak height for each individual sample. The vertical axis measures the peak height seen at each locus for the ten samples, averaged across two duplicates.

3.2.8. Sensitivity Results – Overall

The results seen in the MPS and CE based sections of the sensitivity test were then directly compared to each other by comparing the random match probabilities achieved by MPS and CE for the same sample sets (Table 7 and Table 13). This comparison is shown in the following figure, where it can be seen that the MPS assay gave a stronger (more discriminating) random match probability for every sample.

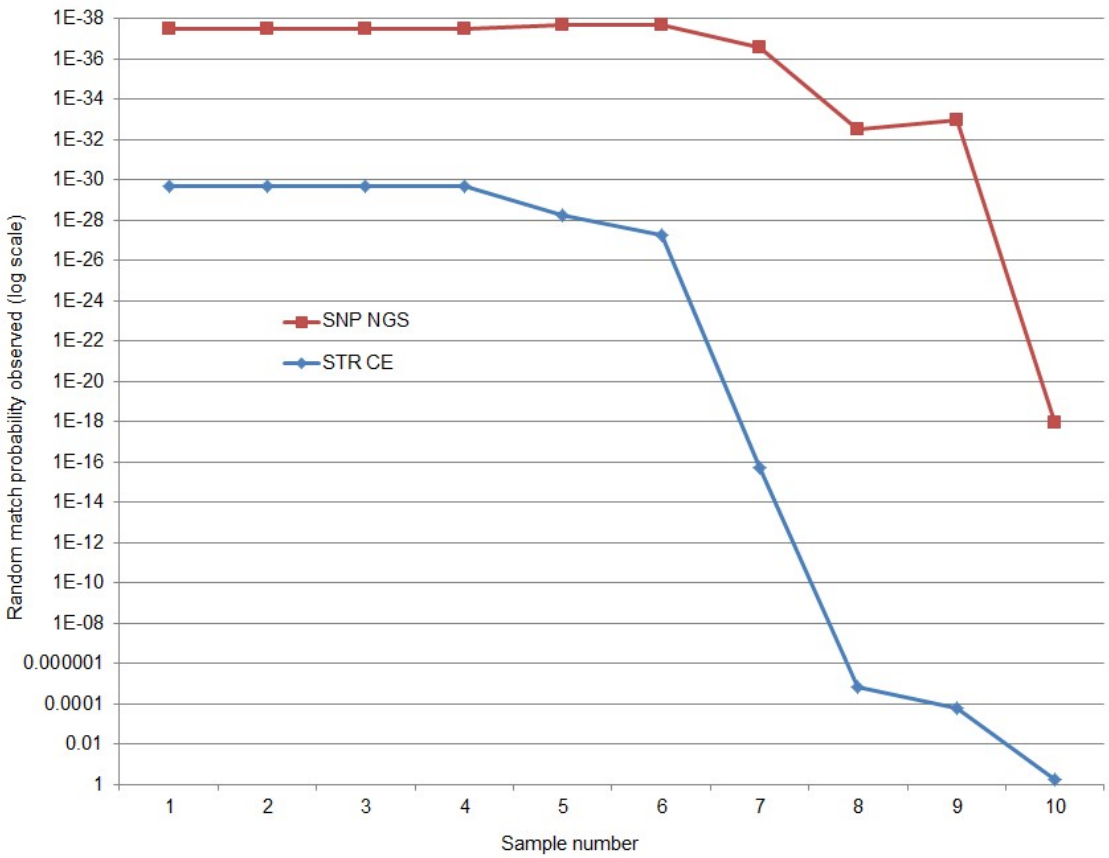


Figure 25: Comparison of random match probabilities achieved for each of the sensitivity samples with GlobalFiler CE-based analysis and Precision ID Identity Panel (SNP) analysis. Sample numbers correspond to those shown in Table 4. Random match probabilities shown are those taken from Table 7 and Table 13.

3.3. Inhibition

3.3.1. Methods – Sample set up

To investigate the resistance of MPS methods to inhibitors found in samples, a common issue in forensic analysis, a series of samples were processed that consisted of an inhibiting substance artificially added to control DNA. Eight replicates of control DNA were first prepared using 0.1 ng/µL Control DNA 007 (Thermo Fisher Scientific, USA). 1 ng total of DNA was added to each replicate. Three of these replicates were retained as positive controls. To the other five, varying amounts of 700 ng/µL humic acid were added as an inhibitor. Each of the samples was then made up to 30 µL total volume with molecular biology grade water. The 30 µL in each sample was then split in two, with 15 µL being processed in each of the MPS and CE methods described below. The sample preparation is detailed in the following table:

Table 14: Sample preparation of humic acid / Control DNA samples for inhibition analysis. The total humic acid in 15 µL is shown as this is the volume of extract added to the reaction for both the MPS and CE assays tested

Sample number	Humic acid added to 30 µL sample (µL)	Total Humic acid each 15 µL reaction (ng)
1	3	1050
2	3	1050
3	6	2100
4	6	2100
5	9	3150
6 (Control)	0	0
7 (Control)	0	0
8 (Control)	0	0

3.3.2. Methods – Precision ID Ancestry Panel

These samples were then processed with the Precision ID Ancestry Panel as described in Section 2.4.1.3. Parameters of the HID_SNP_Genotyper v5.2.2 analysis were as follows (all manufacturer's default values):

Table 15: Parameters used in HID_SNP_Genotyper v5.2.2 analysis for Precision ID Ancestry Panel inhibition analysis. All are the manufacturer's recommended default parameters.

Parameter	Value used
Minimum allele frequency	0.1
Minimum coverage	6
Minimum coverage either strand	0
Maximum strand bias	1
Trim reads	true

3.3.3. Inhibition Results – Precision ID Ancestry Panel

Results of this analysis were examined for the maximum, minimum and mean locus coverage achieved for each sample. Full profiles were observed for all control samples. All samples with humic acid added showed dropout (i.e. minimum locus coverage equal to zero), with samples 3, 4 and 5 being particularly poor, with maximum coverage of only 5, 19

and 4 respectively. This represents a profile with effectively no result. No samples showed partially dropped out loci – i.e. true heterozygous loci that appear homozygous due to the dropout of one allele only. Results are in the following table:

Table 16: Coverage results for Precision ID Ancestry Panel inhibition analysis. Sample numbers and humic acid input correspond to those shown in Table 14. Coverage results of the maximum, minimum, mean and standard deviation coverage for loci within the samples are shown. Minimum coverage of zero indicates that at least one locus dropped out in that sample.

Sample Number	Total Humic acid in 15 µL (ng)	Maximum locus coverage	Minimum locus coverage	Mean locus coverage	Std. dev. locus coverage
1	1050	40557	0	1873	5643
2	1050	2551	0	51	242
3	2100	5	0	0	1
4	2100	19	0	0	2
5	3150	4	0	0	1
6 (Control)	0	25126	463	12793	6642
7 (Control)	0	29673	533	12839	7120
8 (Control)	0	22666	445	10744	6204

The same results were then analysed as to the number of loci that were present in the final profile, with total number of loci and percentage of the 165 loci in the panel shown. All control samples showed a full profile, while for the humic acid samples, at most 45% of a complete profile was seen. This is shown in the following table:

Table 17: Genotype results for Precision ID Ancestry Panel inhibition analysis. Sample numbers and humic acid input correspond to those shown in Table 14. Note that there are 165 loci in the panel, so this is the maximum number of genotypes that can be achieved and represents a full profile

Sample Number	Total Humic acid in 15 µL (ng)	Locus genotypes achieved	% of profile observed
1	1050	75	45%
2	1050	32	19%
3	2100	0	0%
4	2100	2	1%
5	3150	0	0%
6 (Control)	0	165	100%
7 (Control)	0	165	100%
8 (Control)	0	165	100%

The results were then plotted in a chart of coverage against the 165 loci in the panel. The three control samples can be clearly seen at the top of the chart, with no dropout, even at the lowest covered markers in the panel. The humic acid samples however (samples 1 to 5) can be seen to almost immediately descend to the horizontal axis, indicating the dropped-out loci with coverage of zero.

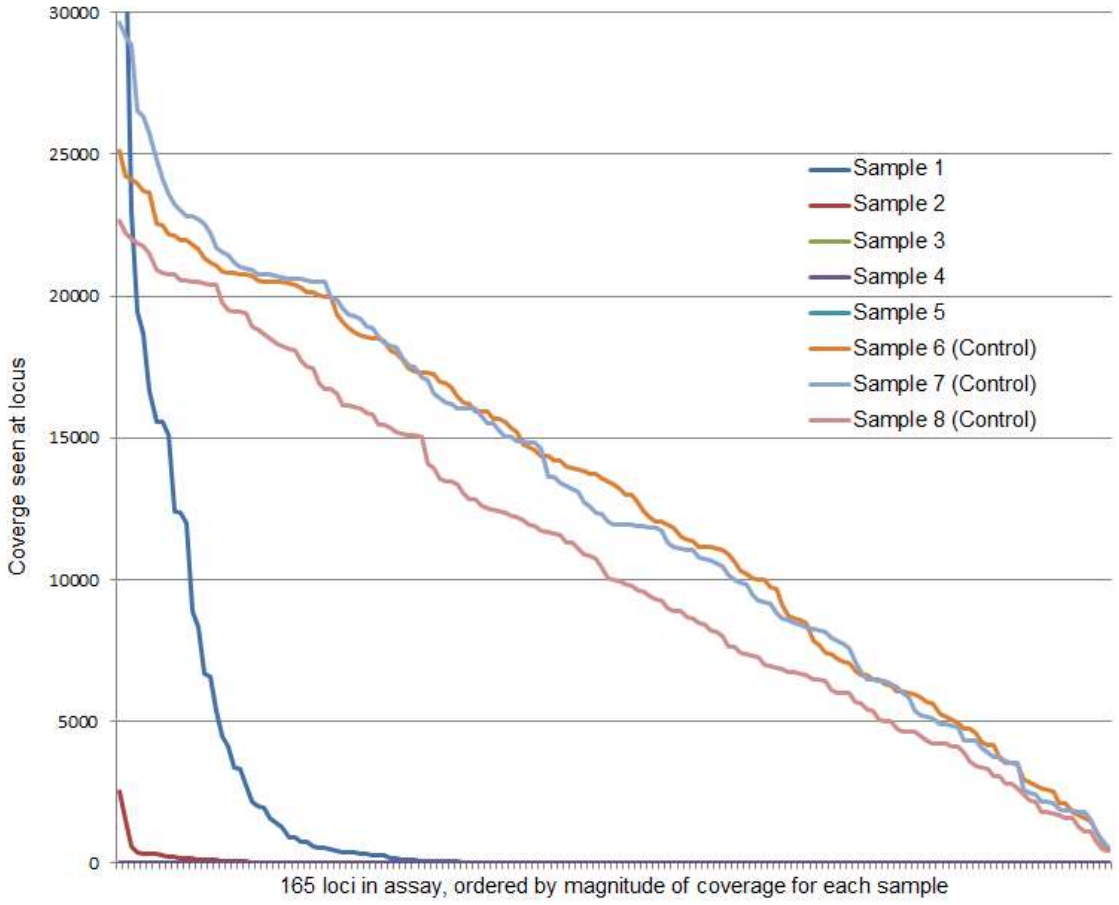


Figure 26: Chart of coverage observed at each locus in Precision ID Ancestry Panel inhibition analysis for each of the samples analysed. Sample numbers correspond to those shown in Table 14. Note that the horizontal axis displays each of the 165 loci in the multiplex, sorted by magnitude of coverage for each individual sample. The vertical axis measures the coverage seen at each locus for the samples.

3.3.4. Methods – Capillary Electrophoresis

The same extracts were then processed in duplicate with the GlobalFiler PCR amplification kit, on the 3500xl Genetic Analyzer as described in Section 2.3.

3.3.5. Inhibition Results – Capillary electrophoresis

Results of this CE analysis were analysed in a similar way to the MPS data, with maximum, minimum and mean peak height examined for each sample. Full profiles were observed for every sample, both control and humic acid, with similar peak height metrics observed in each case. This is shown in the following table:

Table 18: Peak height results for GlobalFiler CE-based inhibition analysis. Sample numbers and humic acid input correspond to those shown in Table 14. Maximum, minimum, mean and standard deviation peak height are shown for each sample.

Sample Number	Total Humic acid in 15 µL (ng)	Maximum peak height (rfu)	Minimum peak height (rfu)	Mean peak height (rfu)	Std. dev. peak height (rfu)
1	1050	7855	1823	3579	1463
2	1050	11182	2860	5122	2062
3	2100	6000	1339	2831	1179
4	2100	6810	1974	3610	1345
5	3150	8484	1900	4319	1704
6 (Control)	0	11150	2170	4690	1929
7 (Control)	0	13433	3173	5887	2520
8 (Control)	0	10900	2259	4998	2105

The results were then analysed in the same way as the MPS data, with the number of alleles observed in total and as a percentage of the number of alleles expected in a full profile. Full profiles were observed for all samples.

Table 19: Genotype results for GlobalFiler CE-based inhibition analysis. Sample numbers and humic acid input correspond to those shown in Table 14. Note that there are a maximum of 46 alleles in the control DNA used, this represents a full profile.

Sample Number	Total Humic acid in 15 µL (ng)	Total alleles observed	% of profile observed
1	1050	46	100
2	1050	46	100
3	2100	46	100
4	2100	46	100
5	3150	46	100
6 (Control)	0	46	100
7 (Control)	0	46	100
8 (Control)	0	46	100

The results were then plotted in a chart of coverage against the 23 loci in the assay. All samples in the experiment, control and humic acid, show a similar pattern of a full profile with no dropout, trending downwards from the loci with highest peaks on the left to lowest on the right:

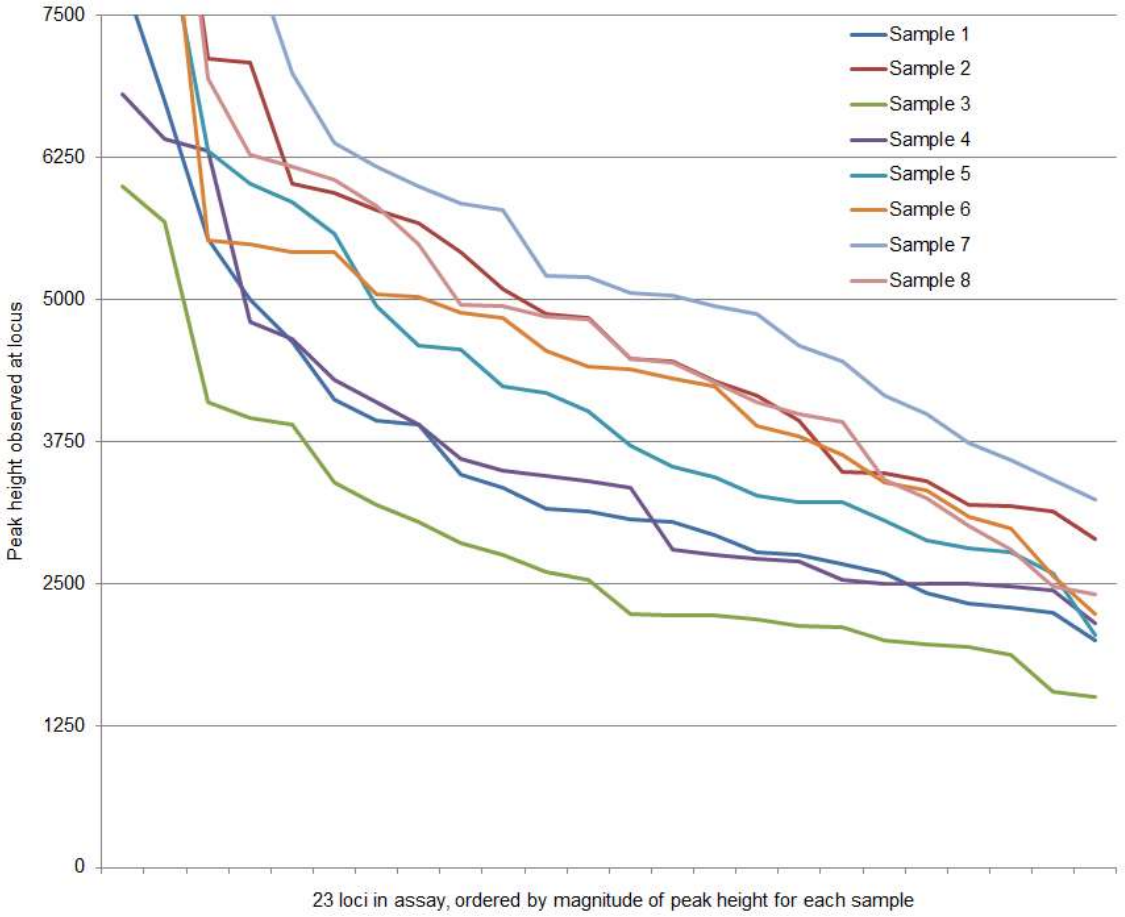


Figure 27: Chart of peak height observed at each locus in GlobalFiler CE-based inhibition analysis for each of the samples analysed. Sample numbers correspond to those shown in Table 14. Note that the horizontal axis displays each of the 23 loci in the multiplex, sorted by magnitude of coverage for each individual sample. The vertical axis measures the peak height seen at each locus for the sample.

3.3.6. Inhibition Results - Overall

The results achieved in the inhibition analysis for MPS and CE-based methods were then compared to each other through comparison of the '% of profile observed' metric, displayed for MPS and CE results in Table 17 and Table 19 respectively. This comparison is shown in the following table:

Table 20: Comparison of Precision ID Ancestry Panel MPS analysis and GlobalFiler CE-based analysis for inhibition samples. Sample numbers and humic acid input correspond to those shown in Table 14.

Sample Number	Total Humic acid in 15 µL (ng)	% of profile observed with MPS	% of profile observed with CE
1	1050	45%	100
2	1050	19%	100
3	2100	0%	100
4	2100	1%	100
5	3150	0%	100
6 (Control)	0	100%	100
7 (Control)	0	100%	100
8 (Control)	0	100%	100

3.4. Mixtures

3.4.1. Methods – Sample set up

To investigate the performance of MPS with mixed source samples, again a common issue in forensic analysis, a series of samples were processed consisting of two control DNA types mixed together in known ratios. The control DNAs used were Control DNA 007 (Thermo Fisher Scientific, USA) and Control DNA 9947A (Thermo Fisher Scientific, USA). These were diluted and combined together in different ratios as detailed in the following table:

Table 21: Sample preparation of Control DNA samples for mixture analysis. 007 and 9947A are the two control DNA types used in the analysis. A constant amount of 9947A was added to most samples, with varying amounts of 007 then mixed into the sample. The total amount of each DNA, and the resulting ratio of the two is shown.

Sample number	Amount of 9947A DNA added (ng)	Amount of 007 DNA added (ng)	Ratio of 9947A to 007
1	1	0	Neat 9947A
2	1	0.01	100 : 1
3	1	0.02	50 : 1
4	1	0.05	20 : 1
5	1	0.1	10 : 1
6	1	0.2	5 : 1
7	1	1	1 : 1
8	0	1	Neat 007

3.4.2. Methods – Precision ID Mixture ID Panel

These samples were then processed with the Precision ID Mixture ID Panel as described in Section 2.4.1.1. Parameters of the Converge v2.1 analysis were as follows (all manufacturer's default values):

Table 22: Parameters used in Converge v2.1 analysis for Precision ID Mixture ID Panel mixture analysis. All are the manufacturer's recommended default parameters.

Parameter	Value used
Target file	Globalfiler_MixtureID_targets_v1.0
Hotspot file	Globalfiler_MixtureID_hotspots_v1.0
Microhaplotype Min Coverage	0.02
STR flank length	15
STR flank tolerance	2
STR analytical threshold	0.02
STR stochastic threshold	0.05
STR stutter ratio	0.2

3.4.3. Mixture Results – Precision ID Mixture ID Panel

Results for the MPS data were then analysed as to the number of STR alleles observed in the profiles from each of the two DNA sources. Due to the genotypes of the two DNA sources in question, some alleles could be attributed to one or other of the profiles, but some alleles could not as they were common to both genotypes. Twenty-five STR alleles were shared between the two control genotypes, with there being 26 unique alleles in the 9947A genotype and 23 unique alleles in the 007 genotype. Of note was the D8S1179 in the 9947A genotype, which had a heterozygote genotype consisting of two forms of the 13 repeat allele due to sequence variation. One of these allele variants was shared with the 007 genotype, the other was unique to the 9947A profile. These genotypes are detailed in the following table:

Table 23: STR genotypes of the control DNA used for Precision ID Mixture ID Panel mixture analysis. Alleles shaded grey are those shared between the two genotypes. Alleles in blue are those in the 007 (minor) genotype that sit in a minus-4 stutter position of an allele in 9947A.The allele shaded green at D8S1179 indicates sequence variation – two forms of the 13 allele were detected in the 9947A genotype. One of these variants is shared with the 007 genotype.

Locus	9947A Control Genotype		007 Control Genotype	
	Allele 1	Allele 2	Allele 1	Allele 2
AMEL	X		X	Y
CSF1PO	10	12	11	12
D10S1248	13	15	12	15
D12ATA63	13		13	17
D12S391	18	20	18	19
D13S317	11		11	
D14S1434	11	13	11	14
D16S539	11	12	9	10
D18S51	15	19	12	15
D19S433	14	15	14	15
D1S1656	18.3		13	16
D1S1677	13	14	13	
D21S11	30		28	31
D22S1045	11	14	11	16
D2S1338	19	23	20	23
D2S1776	10		8	10
D2S441	10	14	14	15
D3S1358	14	15	15	16
D3S4529	13		13	
D4S2408	9	10	10	11
D5S2800	14	23	17	18
D5S818	11		11	
D6S1043	12	18	12	14
D6S474	14	18	14	
D7S820	10	11	7	12
D8S1179	13	13	12	13
FGA	23	24	24	26
TH01	8	9.3	7	9.3
TPOX	8		8	
vWA	17	18	14	16
DYS391			11	

All 25 shared alleles and all but one of the expected 26 alleles from the stronger of the two contributors (9947A) were detected in every mixed sample. Seven of the expected 007 alleles also fell into a ‘stutter position’ (i.e. four bases smaller) relative to a 9947A allele at the same locus, and so were removed from the analysis. This is because it was not possible to determine whether a peak present in this position was due to signal from the 007 minor component or due to stutter from the 9947A allele. These ‘stutter’ peaks are highlighted in Table 23.

At least three alleles from the weaker 007 contributor were detected in every sample, with, as could be expected, more 007 alleles being seen with more of this DNA being present in the mixture. Results were as follows:

Table 24: STR allele counts for Precision ID Mixture ID Panel mixture analysis. Sample numbers and DNA ratio correspond to those shown in Table 21. The number of STR alleles observed in the profile that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source.

Sample Number	Ratio of 9947A to 007 in sample	Number of 9947A STR alleles observed	Number of 007 STR alleles observed	Number of shared STR alleles observed
1	Neat 9947A	26	-	25
2	100 : 1	26	3	25
3	50 : 1	26	7	25
4	20 : 1	26	11	25
5	10 : 1	26	18	25
6	5 : 1	26	19	25
7	1 : 1	25	23	25
8	Neat 007	-	23	25

Results were then analysed as to the mean coverage of the observed STR alleles. This was relatively constant across the eight samples for the shared and 9947A alleles, and increased with increasing amounts of 007 being added to the mixture. Results are shown in Table 25 and Figure 28:

Table 25: Mean STR coverage results for Precision ID Mixture ID Panel mixture analysis. Sample numbers and DNA ratio correspond to those shown in Table 21. The mean coverage of STR alleles observed in the resulting profile that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source.

Sample Number	Ratio of 9947A to 007 in sample	Mean STR coverage of 9947A alleles	Mean STR coverage of 007 alleles	Mean STR coverage of shared alleles
1	Neat 9947A	1803	-	-
2	100 : 1	1673	50	2085
3	50 : 1	1658	104	1945
4	20 : 1	1610	157	2102
5	10 : 1	1723	176	2281
6	5 : 1	1835	327	2590
7	1 : 1	1011	1185	2653
8	Neat 007	-	2161	-

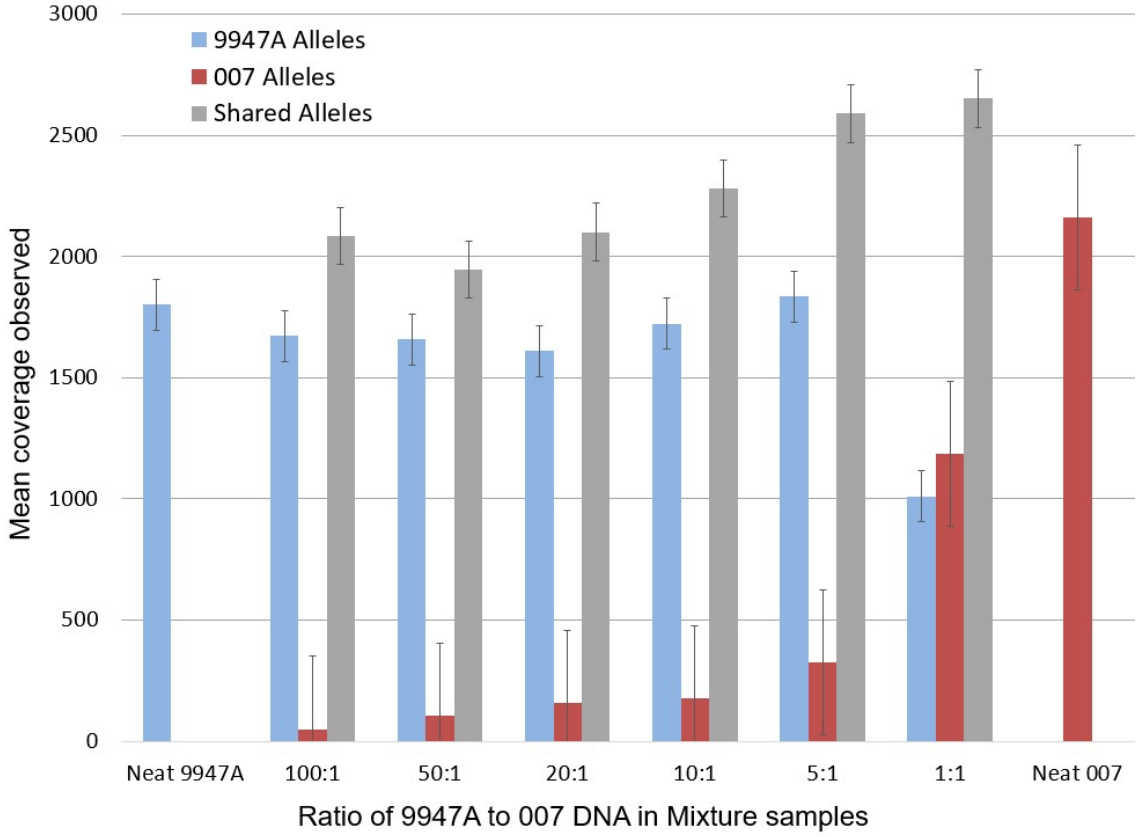


Figure 28: Mean STR coverage results for Precision ID Mixture ID Panel mixture analysis. The mean coverage of STR alleles observed that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source.

Results were then analysed as to the maximum coverage of the observed STR alleles. Again, this was relatively constant across the eight samples for the shared and 9947A alleles, as could be expected, and increased with increasing amounts of 007 being added to the mixture. Results were as follows:

Table 26: Maximum STR coverage results for Precision ID Mixture ID Panel mixture analysis. The maximum coverage of STR alleles observed in the resulting profile that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source.

Sample Number	Ratio of 9947A to 007 in sample	Max STR coverage of 9947A alleles	Max STR coverage of 007 alleles	Max STR coverage of shared alleles
1	Neat 9947A	5654	-	-
2	100 : 1	3675	84	6295
3	50 : 1	3428	158	4739
4	20 : 1	3204	251	5179
5	10 : 1	3353	297	6588
6	5 : 1	4009	685	6543
7	1 : 1	2175	2254	5246
8	Neat 007	-	5943	-

Results for the MPS data were then analysed as to the number of microhaplotype alleles observed in the profiles. Result are presented in the same way as for the STR result, with the number of 9947A, 007, and shared alleles shown. The microhaplotype genotypes for the two control DNA samples, with the shared alleles highlighted, are shown in the following table. Twenty-nine alleles were shared between the two genotypes, with there being 36 unique alleles in the 9947A genotype and 30 unique alleles in the 007 genotype.

Table 27: Microhaplotype genotypes of the control DNA used for Precision ID Mixture ID Panel mixture analysis. Alleles shaded grey are those shared between the two genotypes (Table continues on next page).

Locus	9947A Control Genotype		007 Control Genotype	
	Allele 1	Allele 2	Allele 1	Allele 2
mh01KK-001	CG	TG	CA	
mh01KK-002	AA	GG	AA	GG
mh01KK-106	CAGA	CAGG	CAGG	TAGG
mh01KK-205	CCAG	TCAG	TTAG	TTGG
mh02KK-134	TCG	TTG	CCG	TCG
mh02KK-136	GTA	TTC	GTC	TTC

Locus	9947A Control Genotype		007 Control Genotype	
	Allele 1	Allele 2	Allele 1	Allele 2
mh03KK-006	AA	AG	AA	
mh04KK-017	GCA	GTA	GCA	
mh05KK-062	AA		TA	
mh05KK-170	CAGG	CGGG	CAGA	CGGA
mh09KK-152	AGCA	GTCG	AGCA	GTCG
mh09KK-153	CAA	TAA	TAA	TAC
mh09KK-157	ACCT	GCCT	ACCT	GCCC
mh10KK-169	ACTG	GCTG	ACCG	GCTG
mh11KK-180	ACTC		ACCG	ACTC
mh11KK-187	CCCG	GCGG	GCGG	
mh11KK-191	CGAT	TGAT	TAAC	
mh12KK-046	GA		GG	TA
mh12KK-202	AACT	CATT	AATC	
mh13KK-213	CCA	CCG	CCA	CCG
mh13KK-217	AGCA	AGCG	AATG	AGCG
mh13KK-218	CTTT	TTCT	CTCT	TTTT
mh15KK-067	GT	TT	TC	TT
mh15KK-104	TAG		TAA	TAG
mh16KK-049	AAA	ACG	ACG	
mh16KK-255	ACTG	GACA	GACA	
mh16KK-302	ACTT	GCTC	ACTT	
mh17KK-272	TCCT		CCCT	TTCC
mh18KK-293	AGAA		AGAA	
mh19KK-299	CGTA	TGAA	CATG	CGTA
mh19KK-301	GAAC	GGAT	GAAC	GGAT
mh21KK-316	GCGC		GCGC	
mh21KK-320	AACA	GGCG	GACA	GACG
mh21KK-324	CTAA	TCAG	CCTA	TCTG
mh22KK-061	AAA	GAA	AAA	GGG
mh22KK-069	AG	GT	GT	

Stutter was not a factor in the microhaplotype analysis, as one of the reported advantages of microhaplotypes is that, unlike STR loci, they are not susceptible to stutter (see Section 1.1.4.4). As in the STR analysis, all shared alleles and all of the expected alleles from the stronger of the two contributors (9947A) were detected in every sample. For the weaker 007 contributor, some alleles were detected in every sample, with more 007 alleles being seen with more of this DNA present in the mixture. Results were as follows:

Table 28: Microhaplotype allele counts for Precision ID Mixture ID Panel mixture analysis. Sample numbers and DNA ratio correspond to those shown in Table 21. The number of MH alleles observed in the resulting profile that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source (MH = microhaplotype).

Sample Number	Ratio of 9947A to 007 in sample	Number of 9947A MH alleles observed	Number of 007 MH alleles observed	Number of shared MH alleles observed
1	Neat 9947A	36	-	29
2	100 : 1	36	2	29
3	50 : 1	36	8	29
4	20 : 1	36	21	29
5	10 : 1	36	30	29
6	5 : 1	36	30	29
7	1 : 1	36	30	29
8	Neat 007	-	30	29

Results were then analysed as to the mean coverage of the observed microhaplotype alleles. This was relatively constant across the eight samples for the shared and 9947A alleles, and increased with increasing amounts of 007 being added to the mixture. Results are shown in Table 29 and Figure 29:

Table 29: Mean microhaplotype coverage results for Precision ID Mixture ID Panel mixture analysis. Sample numbers and DNA ratio correspond to those shown in Table 21. The mean coverage of MH alleles observed in the resulting profile that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source (MH = microhaplotype).

Sample Number	Ratio of 9947A to 007 in sample	Mean MH coverage of 9947A alleles	Mean MH coverage of 007 alleles	Mean MH coverage of shared alleles
1	Neat 9947A	1533	-	-
2	100 : 1	1623	58	1950
3	50 : 1	1504	81	1830
4	20 : 1	1634	126	2065
5	10 : 1	1691	210	2237
6	5 : 1	1719	314	2439
7	1 : 1	1828	1273	2674
8	Neat 007	-	1717	-

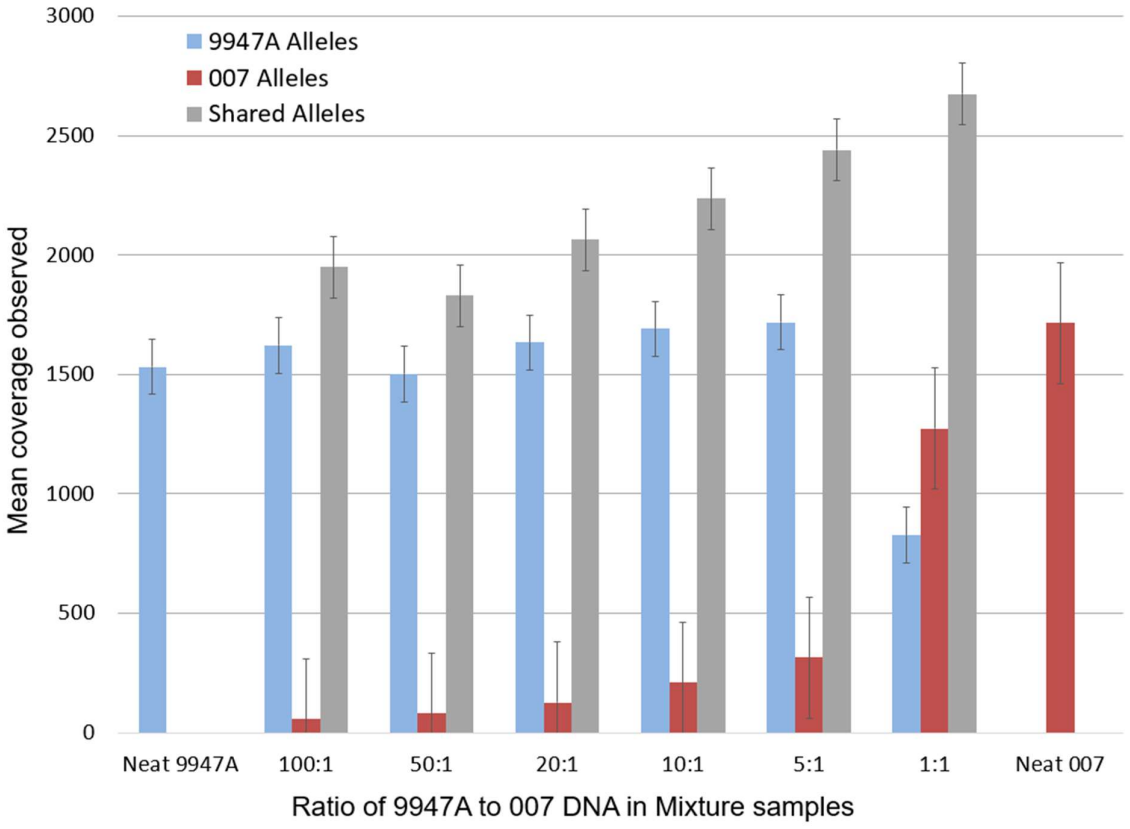


Figure 29: Mean microhaplotype coverage results for Precision ID Mixture ID Panel mixture analysis. The mean coverage of microhaplotype alleles observed that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source.

Results were then analysed as to the maximum coverage of the observed microhaplotype alleles. Again, this was relatively constant across the eight samples for the shared and 9947A alleles and increased with increasing amounts of 007 being added to the mixture. Results were as follows:

Table 30: Maximum microhaplotype coverage results for Precision ID Mixture ID Panel mixture analysis. Sample numbers and DNA ratio correspond to those shown in Table 21. The maximum coverage of MH alleles observed in the resulting profile that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source (MH = microhaplotype).

Sample Number	Ratio of 9947A to 007 in sample	Max MH coverage of 9947A alleles	Max MH coverage of 007 alleles	Max MH coverage of shared alleles
1	Neat 9947A	3888	-	-
2	100 : 1	3662	61	4050
3	50 : 1	3380	117	4519
4	20 : 1	4794	270	4301
5	10 : 1	4693	574	5191
6	5 : 1	4407	617	5448
7	1 : 1	2348	3608	5970
8	Neat 007	-	5166	-

3.4.4. Methods – Capillary Electrophoresis

The same extracts were then processed in duplicate with the GlobalFiler PCR amplification kit, on 3500xl Genetic Analyzer as described in the Section 2.3.

3.4.5. Mixture Results – Capillary Electrophoresis

Results for the CE data were then analysed in the same way as the MPS data, with the number of alleles observed in the profiles from each of the two DNA sources being tallied. Seventeen STR alleles were shared between the two control genotypes, with there being 20 unique alleles in the 9947A genotype and 19 unique alleles in the 007 genotype. These are shown in the following table:

Table 31: STR genotypes of the control DNA used for GlobalFiler PCR amplification kit mixture analysis. Alleles shaded grey are those shared between the two genotypes. Alleles in blue are those in the 007 (minor) genotype that sit in a minus-4 stutter position of an allele in 9947A.

Locus	9947A Control Genotype		007 Control Genotype	
	Allele 1	Allele 2	Allele 1	Allele 2
AMEL	X		X	Y
CSF1PO	10	12	11	12
D10S1248	13	15	12	15
D12S391	18	20	18	19
D13S317	11		11	
D16S539	11	12	9	10
D18S51	15	19	12	15
D19S433	14	15	14	15
D1S1656	18.3		13	16
D21S11	30		28	31
D22S1045	11	14	11	16
D2S1338	19	23	20	23
D2S441	10	14	14	15
D3S1358	14	15	15	16
D5S818	11		11	
D7S820	10	11	7	12
D8S1179	13		12	13
FGA	23	24	24	26
SE33	19	29.2	17	25.2
TH01	8	9.3	7	9.3
TPOX	8		8	
vWA	17	18	14	16
Y-indel			2	
DYS391			11	

All shared alleles and all of the expected alleles from the stronger of the two contributors (9947A) were detected in every sample. As with the MPS STR data, stutter was taken into consideration. As was the case in the CE data, seven of the expected 007 alleles also fell into a ‘stutter position’ relative to a 9947A allele at the same locus, and so were removed from the analysis.

This time, at least 14 of the possible 19 alleles from the weaker 007 contributor were detected in every sample, with, as could be expected, more 007 alleles being seen with more of this DNA being present in the mixture. Results were as follows:

Table 32: Allele counts for GlobalFiler CE-based mixture analysis. Sample numbers and DNA ratio correspond to those shown in Table 21. The number of alleles observed in the resulting profile that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source.

Sample Number	Ratio of 9947A to 007 in sample	Number of 9947A alleles observed	Number of 007 alleles observed	Number of shared alleles observed
1	Neat 9947A	20	-	17
2	100 : 1	20	14	17
3	50 : 1	20	14	17
4	20 : 1	20	18	17
5	10 : 1	20	19	17
6	5 : 1	20	19	17
7	1 : 1	20	19	17
8	Neat 007	-	19	17

Results were then analysed as to the mean peak height of the observed STR alleles. As with the MPS data, this was relatively constant across the eight samples for the shared and 9947A alleles, and increased with increasing amounts of 007 being added to the mixture. Results are shown in Table 33 and Figure 30:

Table 33: Mean peak heights for GlobalFiler CE-based mixture analysis. Sample numbers and DNA ratio correspond to those shown in Table 21. The mean peak height of alleles observed in the resulting profile that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source.

Sample Number	Ratio of 9947A to 007 in sample	Mean peak height of 9947A alleles	Mean peak height of 007 alleles	Mean peak height of shared alleles
1	Neat 9947A	4844	-	-
2	100 : 1	6761	165	9779
3	50 : 1	4696	140	7078
4	20 : 1	6351	424	8342
5	10 : 1	4912	560	7213
6	5 : 1	5240	1335	8733
7	1 : 1	3038	3289	8383
8	Neat 007	-	9315	-

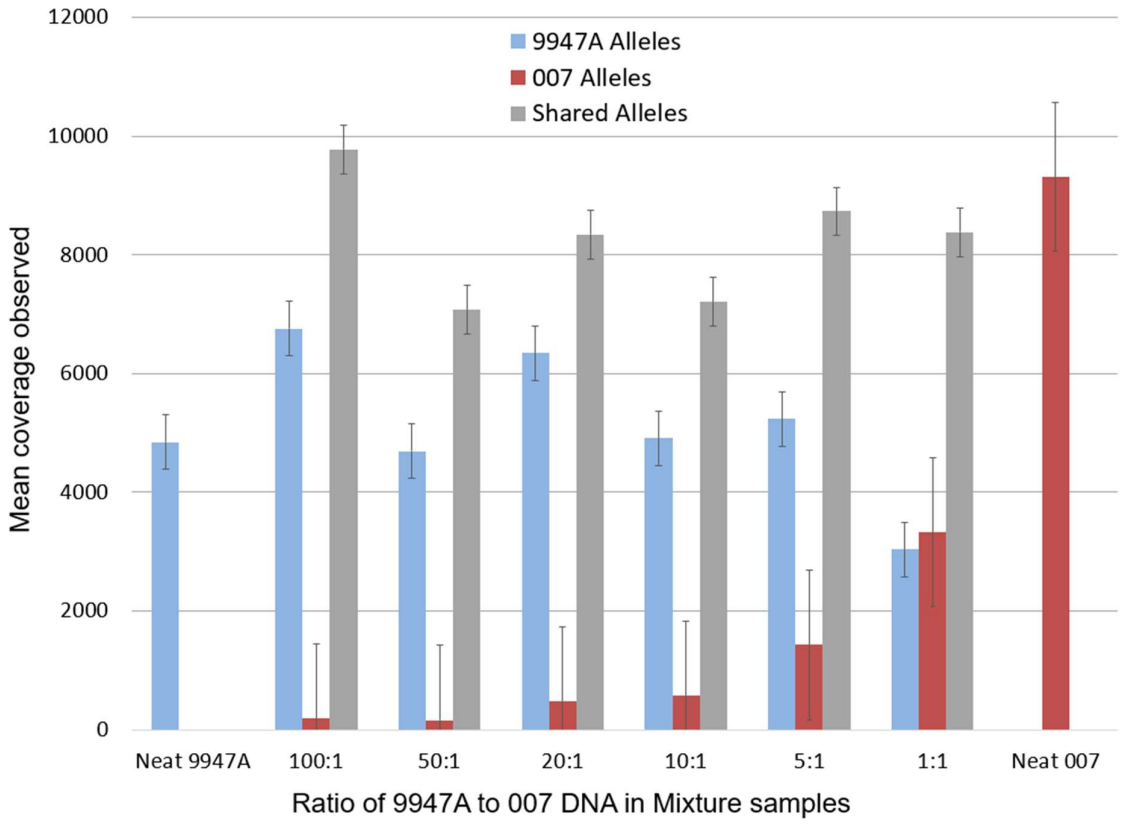


Figure 30: Mean peak heights for GlobalFiler CE-based mixture analysis. The mean peak height of alleles observed that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source.

Results were then analysed as to the maximum peak height of the observed alleles. Again, this was relatively constant across the eight samples for the shared and 9947A alleles, as could be expected, and increased with increasing amounts of 007 being added to the mixture. Results were as follows:

Table 34: Maximum peak heights for GlobalFiler CE-based mixture analysis. Sample numbers and DNA ratio correspond to those shown in Table 21. The maximum peak height of alleles observed in the resulting profile that clearly belong to either of the 007 or 9947A source DNA are shown. Shared alleles are those shared by the genotypes of both 007 and 9947A and in a mixed sample, cannot be attributed to either source.

Sample Number	Ratio of 9947A to 007 in sample	Max peak height of 9947A alleles	Max peak height of 007 alleles	Max peak height of shared alleles
1	Neat 9947A	10402	-	-
2	100 : 1	17087	353	22611
3	50 : 1	12127	380	15610
4	20 : 1	15312	1021	19696
5	10 : 1	10576	1170	16788
6	5 : 1	13714	2422	19735
7	1 : 1	7791	5042	17378
8	Neat 007	-	14315	-

The data obtained from the MPS STR profiles was then plotted in a chart of coverage against locus, averaged across all profiles in the experiment. This showed that the D22S1045 locus was a noticeable outlier, with significantly lower coverage than other loci.

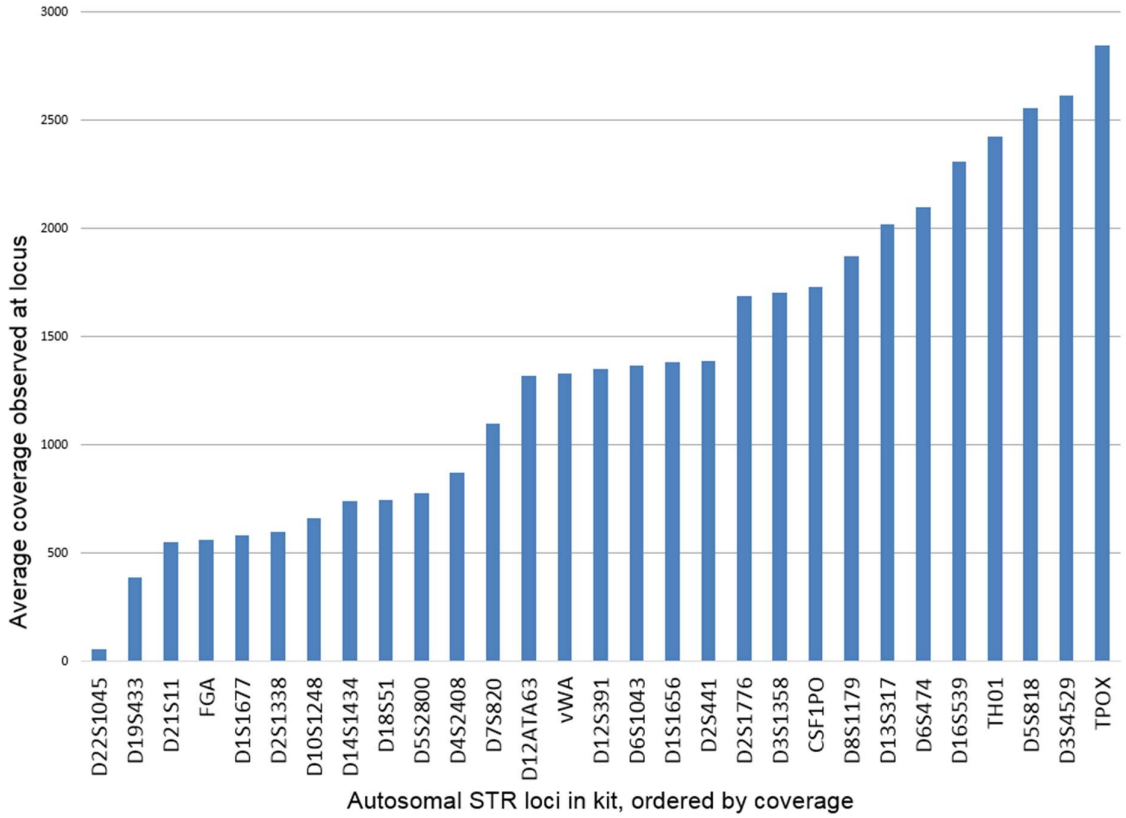


Figure 31: Chart of coverage observed at each STR locus for Precision ID Mixture ID Panel mixture analysis, averaged across all samples in mixture study. The vertical axis shows the coverage seen at each locus. The horizontal axis shows the loci in the kit.

A similar chart of coverage against loci was made for the data obtained from the MPS microhaplotype profiles, again averaged across all profiles in the experiment. In this data set, the mh13KK-217 locus could be seen to give significantly lower coverage than other loci.

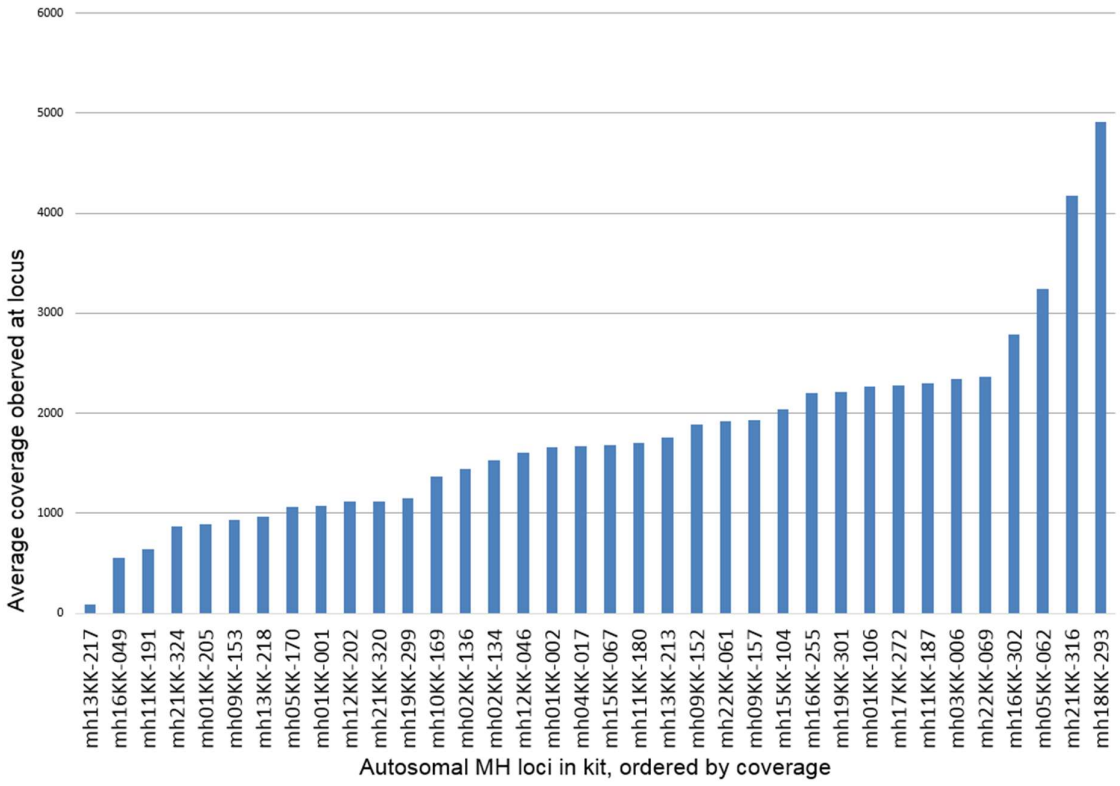


Figure 32: Chart of coverage observed at each microhaplotype locus for Precision ID Mixture ID Panel mixture analysis, averaged across all samples in mixture study. The vertical axis shows the coverage seen at each locus. The horizontal axis shows the loci in the kit.

The CE data in the mixture experiment was then plotted in a chart of coverage against each locus, again averaged across all profiles in the experiment. Unlike the MPS data, no locus stands out as significantly lower than the others.

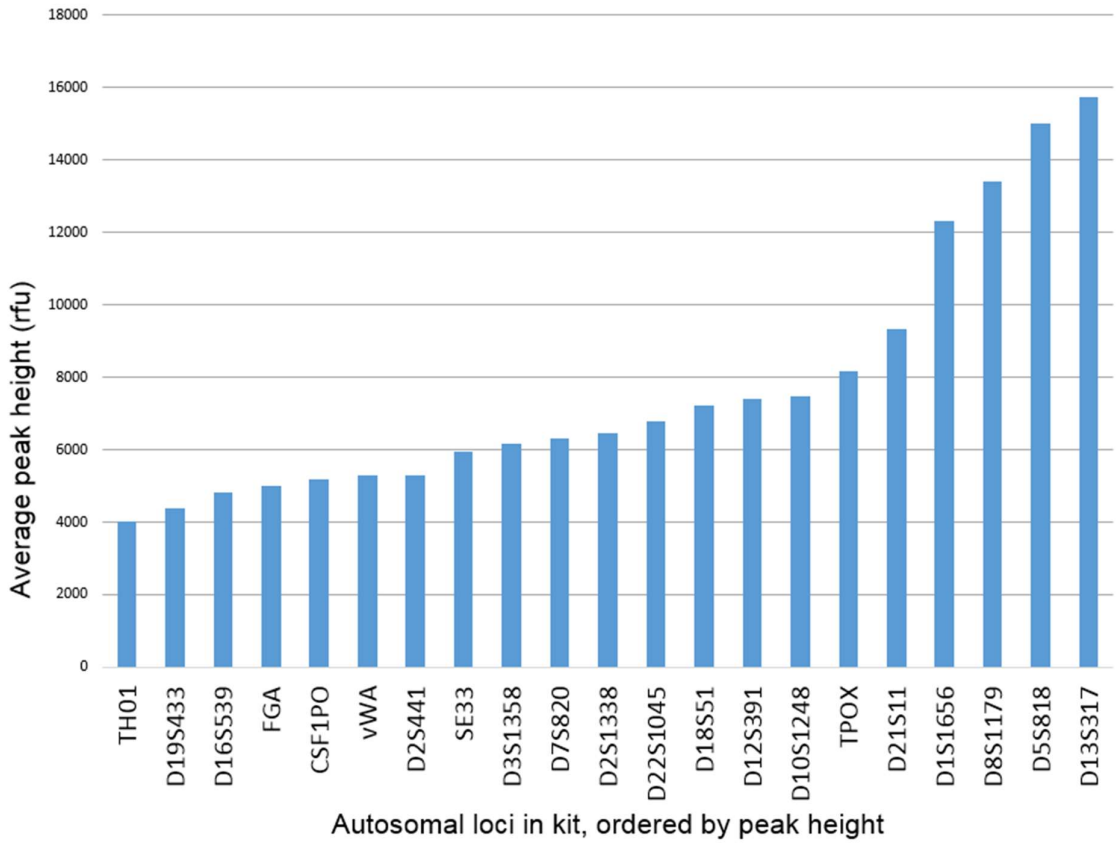


Figure 33: Chart of peak height observed at each locus for GlobalFiler CE-based mixture analysis, averaged across all samples in mixture study. The vertical axis shows the peak height seen at each locus. The horizontal axis shows the loci in the kit.

For comparison to the MPS STR data, which was gained from the Precision ID Mixture ID kit (a prototype panel not commercially released by the manufacturer), a graph is shown of the manufacturer's data from the commercially available Precision ID GlobalFiler STR NGS panel v2. This does not show the same outlying poor performance of the D22S1045 locus, as in the Precision ID Mixture ID data shown in Figure 31.

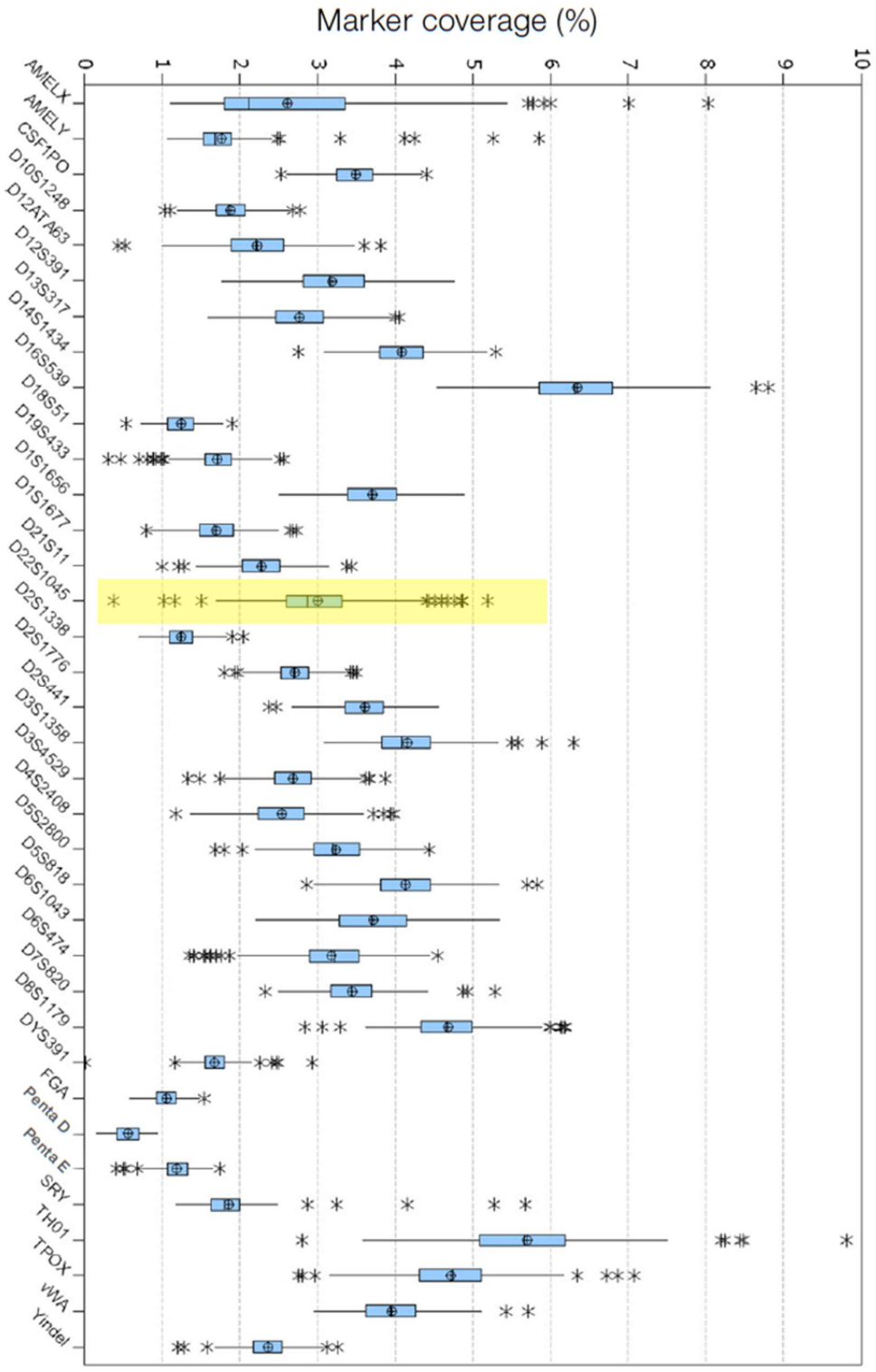


Figure 34: Manufacturer's chart of coverage observed at each STR locus for the Precision ID GlobalFiler NGS STR v2 Panel. Data shows coverage per locus plotted as a percentage of total reads across 8x 1ng samples averaged across all samples in mixture study. The horizontal axis shows the coverage seen at each locus as a percentage. The vertical axis shows the loci in the kit. The D22S1045 locus, which was the lowest covered marker in this analysis (see Figure 31) is highlighted in yellow. (Taken from Thermo Fisher Scientific Application Note: *Get more information from challenging samples with next-generation sequencing of short tandem repeats*, 2019).

3.5. Non-probative samples

3.5.1. Methods – Sample set up

To investigate the real-world value of MPS compared to CE methods, a series of eleven non-probative, 'casework style' samples were processed representing routine samples received in forensic laboratories. These were all extracted and quantified with PrepFiler Express BTA and Quantifiler Trio as described in Section 2.2.2 and Section 2.2.3. The Internal Positive Control (IPC) in the kit was analysed as an indication of possible inhibition in the samples. This value was close to the expected 28 cycles for all except sample 9, indicating some possible inhibition in that sample. The degradation index was calculated for each sample. This is the ratio of the small autosomal quantitation result to the large autosomal quantitation result, and gives an indication of how degraded the source DNA is. A range of degradation indices were observed across the experiment with several results of '1', indicating no degradation, but also with results of 26 and 55 for other samples, indicating moderate to high degradation. Full results were as follows:

Table 35: Sample source and quantitation results for non-probative analysis. Source refers to the sample type aland substrate that was used, these came from a variety of 'casework style' sources. Quantifiler Trio quantitation results for each sample are shown, these are for both the small and large autosomal target in that kit. The degradation index is the ratio of the small autosomal target result to the large autosomal target result, rounded to the nearest integer. The IPC (Internal Positive Control) result is also shown. This is an indication of possible inhibition in the sample.

Sample No.	Source	Quantitation result (ng/µL)			Degradation Index
		Small autosomal	Large autosomal	IPC	
1	Fabric cutting	0.014	0.014	28.1	1
2	Reference buccal	0.314	0.091	27.7	3
3	Tissue FFPE	0.012	0.000	27.8	55
4	Cigarette butt	0.074	0.005	27.7	15
5	Envelope flap	0.003	0.003	27.6	1
6	Bone fragment	0.002	0.000	27.9	26
7	Touch swab	0.104	0.160	28.1	1
8	Touch swab	0.036	0.049	28.1	1
9	Blood stain	25.076	37.452	29.7	1
10	Touch swab	0.017	0.007	28.0	2
11	Saliva stain	2.807	1.069	27.8	3

3.5.2. Methods – Precision ID Mixture ID Panel

These extracts were then processed with the Precision ID GlobalFiler Mixture ID Panel as described in Section 2.4.1.1. Parameters of the Converge v2.1 analysis were as follows (all manufacturer's default values):

Table 36: Parameters used in Converge v2.1 analysis for Precision ID Mixture ID Panel non-probative analysis. All are the manufacturer's recommended default parameters.

Parameter	Value used
Target file	Globalfiler_MixtureID_targets_v1.0
Hotspot file	Globalfiler_MixtureID_hotspots_v1.0
Microhaplotype Min Coverage	0.02
STR flank length	15
STR flank tolerance	2
STR analytical threshold	0.02
STR stochastic threshold	0.05
STR stutter ratio	0.2

3.5.3. Non-probative Results – Precision ID Mixture ID Panel

Results were analysed as to the maximum, minimum and mean coverage for the STR and microhaplotype results in the panel. Results were varied across the experiment, likely due to the varied nature of the samples analysed, but with all samples giving at least some high coverage loci and useable results. Results were as follows:

Table 37: Coverage results for Precision ID Mixture ID Panel non-probative analysis. Sample numbers correspond to those shown in Table 35. Maximum, minimum and mean coverage for loci within each sample is shown (MH = microhaplotype).

Sample Number	Max. STR locus coverage	Min. STR locus coverage	Mean STR locus coverage	Max. MH locus coverage	Min. MH locus coverage	Mean MH locus coverage
1	14598	897	7000	9689	587	4817
2	41116	88	8776	13100	109	2955
3	1682	134	702	1629	99	606
4	16300	498	5951	3816	10	854
5	14110	513	5840	14141	368	4180
6	3671	15	1373	1267	58	636
7	13446	188	6031	5241	351	3308
8	16140	185	6282	6512	384	2824
9	15360	115	6555	5667	367	3639
10	7082	134	2706	1974	38	634
11	14832	335	5402	4528	73	1800

The MPS results obtained were analysed as to the number of STR and microhaplotype alleles observed. This was calculated both as an absolute number, and as a percentage of a theoretical full profile for the panel. All samples showed at least a ‘useful partial’ profile, with several showing clear evidence of a mixture, that is a number of alleles over 100%, or the number that would be in a complete single source profile. Results were are follows:

Table 38: Genotype results for Precision ID Mixture ID Panel non-probative analysis. Sample numbers correspond to those shown in Table 35. Number of alleles recovered for STR and MH analysis of each sample is shown, along with a short description of the profile obtained. 61 alleles represent a hypothetical full STR profile, 72 alleles represent a hypothetical full MH profile. Percentage shown is the observed number of alleles as a percentage of these numbers (MH = microhaplotype).

Sample No.	Total STR alleles obs.	% of full STR profile obs.	Total MH alleles obs.	% of full MH profile obs.	Result description
1	91	149%	97	135%	Clear mixture at most loci
2	60	98%	68	94%	Clear full profile
3	16	26%	19	26%	Useful partial profile
4	73	120%	52	72%	Full profile with clear signs of small minor
5	86	141%	76	106%	Clear mixture at most loci
6	29	48%	6	8%	Useful partial profile
7	110	180%	91	126%	Clear mixture almost all loci
8	94	154%	99	138%	Full profile with clear signs of small minor
9	91	149%	88	122%	Full profile with clear signs of small minor
10	80	131%	64	89%	Full profile with clear signs of small minor
11	59	97%	65	90%	Clear full profile

3.5.4. Methods – Capillary Electrophoresis

The same extracts were then processed in duplicate with the GlobalFiler PCR amplification kit, on 3500xl Genetic Analyzer as described in Section 2.3.

3.5.5. Non-probative Results – Capillary Electrophoresis

The CE results for this experiment were analysed in a similar way as for the MPS result, with the maximum, minimum and mean calculated for the STR peak heights observed. All samples showed at least some above threshold peaks. Results were as follows:

Table 39: Peak height results for GlobalFiler CE-based non-probative analysis. Sample numbers correspond to those shown in Table 35. Maximum, minimum, mean and standard deviation of peak heights for loci within each sample is shown.

Sample Number	Maximum peak height (rfu)	Minimum peak height (rfu)	Mean peak height (rfu)	Std Dev peak height (rfu)
1	749	59	268	133
2	5783	58	1561	1663
3	251	56	109	95
4	4076	88	1320	1250
5	194	27	96	44
6	127	58	92	16
7	6829	339	2292	1403
8	11438	239	2339	1884
9	7757	61	1955	1188
10	613	57	255	164
11	8391	192	1917	1762

In the same way as the MPS results, the CE results were also analysed as to the number of alleles observed. This was calculated both as an absolute number, and as a percentage of a theoretical full profile for the panel. As with MPS, several samples showed clear evidence of a mixture. Some samples showed clear loss of loci due to degradation (samples 2, 4, 10 and 11) two samples showed only a few isolated peaks, with no useful STR profile being present (samples 3 and 6). Results were as follows:

Table 40: Genotype results for GlobalFiler CE-based non-probative analysis. Sample numbers correspond to those shown in Table 35. Number of alleles recovered for CE analysis of each sample is shown, along with a short description of the profile obtained. 46 alleles represent a hypothetical full STR profile. Percentage shown is the observed number of alleles as a percentage of these numbers.

Sample Number	Total alleles observed	% of profile observed	Result description
1	62	135%	Full profile with clear signs of small minor
2	41	89%	Mostly full profile, signs of degradation
3	4	9%	Non-useful partial profile
4	29	63%	Partial profile, strong signs of degradation
5	34	74%	Low level partial profile. Some indication of mix
6	7	15%	Non-useful partial profile
7	71	154%	Clear mixture. Major plus strong minor
8	67	146%	Full profile with clear signs of small minor
9	50	109%	Full profile with clear signs of small minor
10	19	41%	Partial profile, strong signs of degradation
11	46	100%	Mostly full profile, signs of degradation

3.5.6. Non-probative results – Overall

The results achieved in the non-probative analysis for MPS and CE-based methods were then compared to each other through comparison of the ‘result description’, displayed for MPS and CE results in Table 38 and Table 40 respectively. This is shown in the following table:

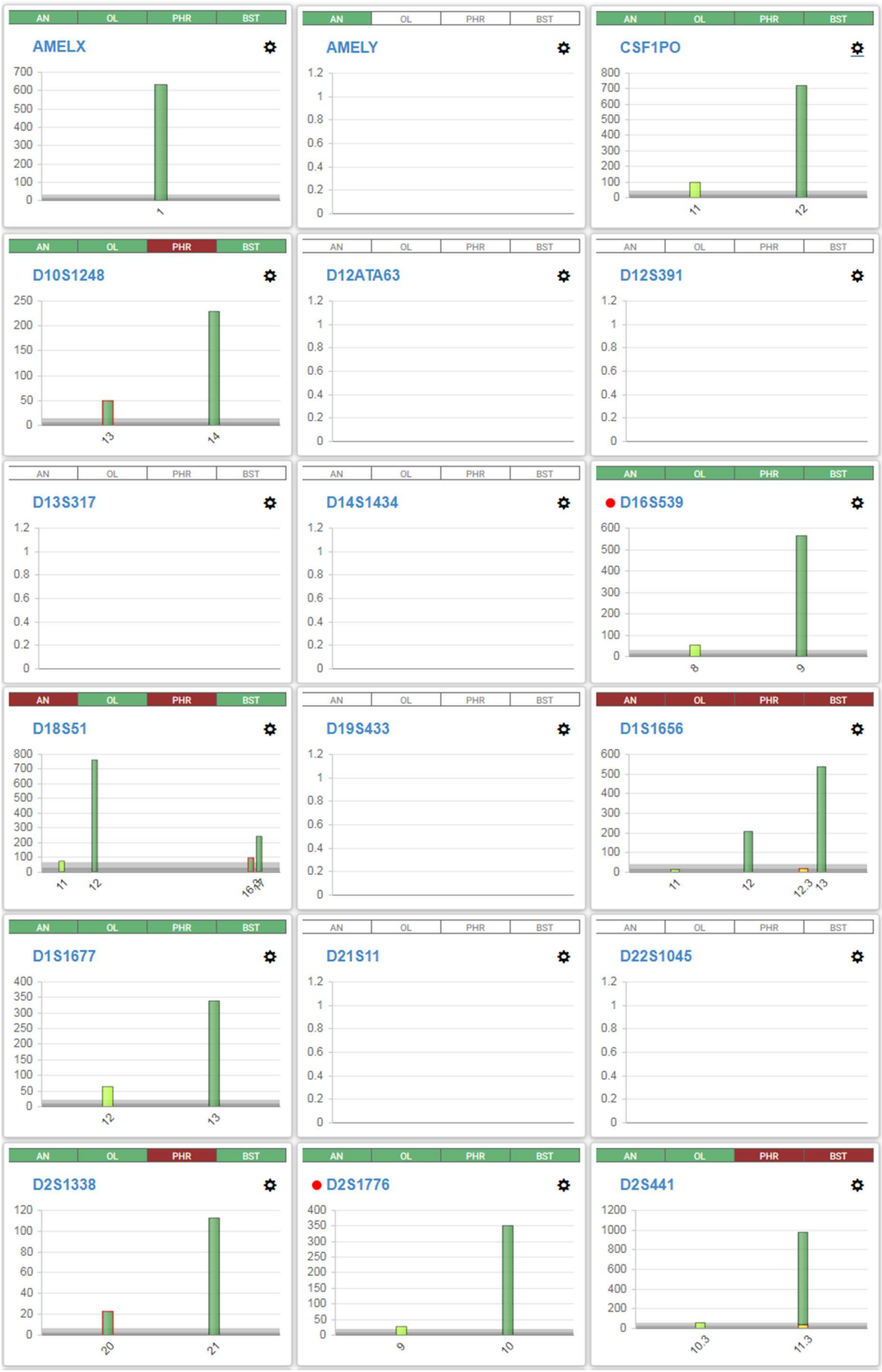
Table 41: Comparison of results for MPS-based and CE-based non-probative analysis. Sample numbers and degradation index correspond to those shown in Table 35. Results descriptions are those in Table 38 (MPS) and Table 40 (CE). Results where MPS analysis was clearly superior to CE are shaded in red. Results where MPS analysis was marginally superior to CE are shaded in blue.

Sample Number	Degradation Index	Result description for MPS Analysis	Result description for CE Analysis
1	1	Clear mixture at most loci	Full profile with clear signs of small minor
2	3	Clear full profile	Mostly full profile, signs of degradation
3	55	Useful partial profile	Non-useful partial profile
4	15	Full profile with clear signs of small minor	Partial profile, strong signs of degradation
5	1	Clear mixture at most loci	Low level partial profile. Some indication of mix
6	26	Useful partial profile	Non-useful partial profile
7	1	Clear mixture almost all loci	Clear mixture. Major plus strong minor
8	1	Full profile with clear signs of small minor	Full profile with clear signs of small minor
9	1	Full profile with clear signs of small minor	Full profile with clear signs of small minor
10	2	Full profile with clear signs of small minor	Partial profile, strong signs of degradation
11	3	Clear full profile	Mostly full profile, signs of degradation

An example of a result highlighted in the above table as a ‘clearly superior’ result with MPS is shown in the figures below (Figure 35 and Figure 36). This is sample 3, which had a ‘non-useful partial profile’ with CE and a ‘useful partial profile’ with MPS.

Figure 35 shows the MPS result (split over two pages to fit all locus results). While many loci give no result, twelve loci do give a result, consistent with an STR profile that could be used for investigative purposes.

Figure 36 shows the CE electropherogram for the same sample, this time mostly displaying no result at each locus, with four loci showing a single above threshold peak. This is an STR profile that typically would not be able to be used for investigative purposes.


AN OL PHR BST
AMELX
AMELY
CSF1PO
D10S1248
D12ATA63
D12S391
D13S317
D14S1434
D16S539
D18S51
D19S433
D1S1656
D1S1677
D21S11
D22S1045
D2S1338
D2S1776
D2S441

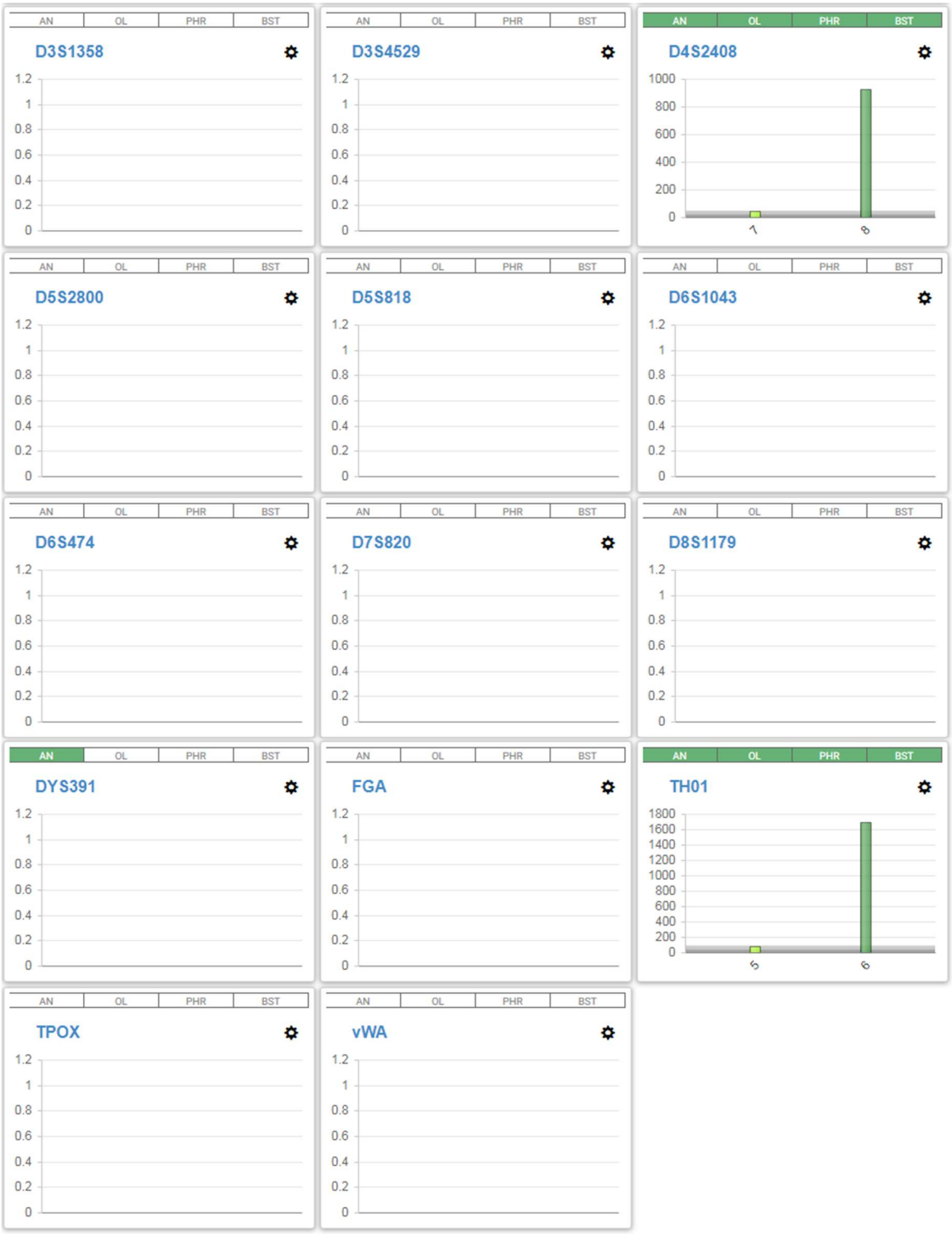


Figure 35: Example of MPS profile obtained in non-probative analysis. Sample 3 is shown, degradation index is 55 (Refer to Table 35 for sample details). Compare to Figure 36 for CE result of same sample Note that the profile picture is split over two pages to show all loci, see previous page also. Called alleles are shown in dark green, stutter peaks that were filtered from the analysis by the software are in light green. Stutter was filtered at 20% the height of the main allele, as per manufacturer's recommendation. Twelve loci give a result, described in Table 41 as a 'useful partial profile'.

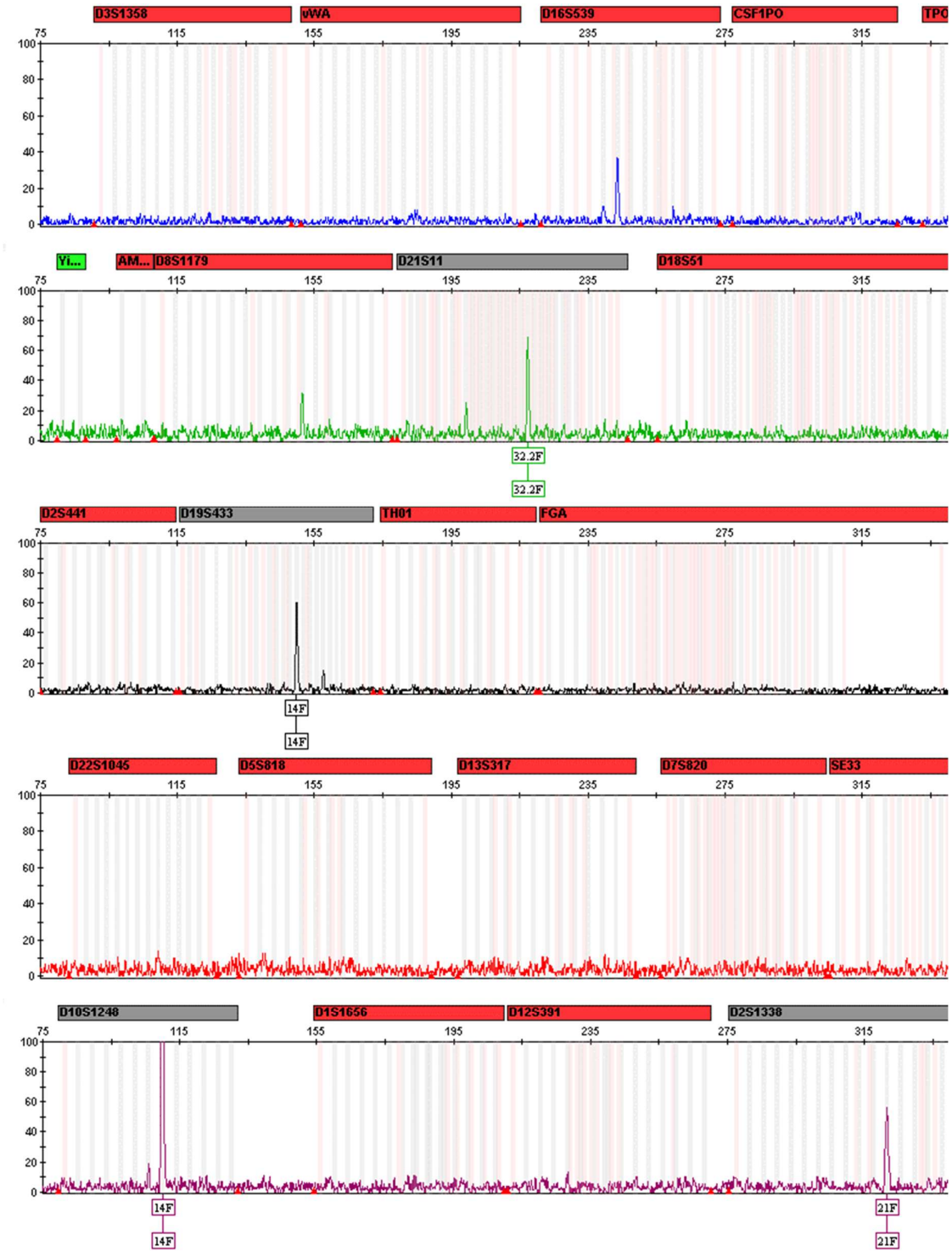


Figure 36: Example of CE profile obtained in non-probative analysis. Sample 3 is shown, degradation index is 55 (Refer to Table 35 for sample details). Compare to Figure 35 for MPS result of same sample. Four loci give an above threshold result (AT = 50rfu), described in Table 41 as a 'non-useful partial profile'.

3.6. Discussion

The results gained in this chapter have shown that MPS offers better sensitivity compared to CE-based methods. For the same set of control DNA dilutions, full profiles were achieved for the strongest four samples with both the MPS SNP assay and CE (Table 7 and Table 13). After this, alleles started to drop out as the amount of DNA decreased. It is of note however that the rate of decline in quality of the profiles was much slower for the MPS SNP method. This was particularly noticeable for samples 8, 9, and 10 (7.8 pg, 3.9 pg, and 2.0 pg of DNA respectively). For the CE methods, these samples gave only very limited partial profiles, with random match probabilities from approximately 10^{-5} to 0.5 (Table 13). In a practical forensic case, these profiles would typically be discarded as not offering sufficient information to identify the donor. For the same samples with the MPS method however, better partial profiles were obtained, with random match probabilities from approximately 10^{-33} to 10^{-18} (Table 7). This is a level that would be highly useful in a practical forensic case. In the sensitivity study for all samples the MPS method offered a higher discrimination power than the CE method (Figure 25).

This result of the MPS SNP assay giving high sensitivity is something that has also been reported by several sources, with Børsting *et al.* 2014, Guo *et al.* 2016 Hussing *et al.* 2018, and Hollard *et al.* 2019, all reporting success in obtaining SNP profiles from amounts of DNA comparable to those used in this study. None of these studies directly compared the MPS profiles to CE STR profiles in terms of discrimination power, however.

It should be noted that the conclusion of MPS-based SNP being more sensitive than CE-based analysis appears to be independent of the analysis threshold that is chosen in the MPS-based analysis. This is an important factor to consider, as the sensitivity of an analysis method can be heightened though lowering of the analytical threshold (AT) that is used to interpret the data. This lowering of the threshold at which alleles are called however can come at the expense of increased risk of false positive allele calls, and interference from artefacts in the analysis. As such, AT must be carefully chosen by forensic laboratories as a validated balance of the sensitivity and accuracy that can be achieved with the method. This issue of which AT to use is an important one here, as given that MPS is new method for forensic analysis, it is unclear precisely what AT should be used for this type of analysis. This subject of appropriate choice of AT is explored in more detail in section 4.2. For the current analysis however, the manufacturer's recommended AT of 6 was used (Table 5). AT is described there with the term used in the software of 'Minimum coverage', i.e. what minimum coverage is required to call an allele. This relatively low AT had little effect on the genotypes that were called however, with no alleles in the entire analysis being below 20

reads and only two alleles falling in the 21 to 50 reads band (Table 8). This indicates that the AT for the analysis could have been raised to as much as 50 or 100 without lowering the sensitivity of the result achieved, and as such, the result observed here is due to the inherent sensitivity of the MPS assay, not due to use of an artificially low AT.

Related to this consideration of analytical threshold is the subject of negative controls. Library negative controls were run as part of the sensitivity study with the MPS SNP assay and the CE STR assay. No spurious alleles were detected in the CE analysis. A total of one spurious allele was observed in the MPS SNP assay for the negative control (Table 6). This allele was not replicated in the duplicate negative control run in the same batch, which indicates that it was likely drop-in to the sample, rather than an inherent contamination, something also evidenced by the fact no reads that could not be attributed to the expected sample DNA were observed in the remainder of the MPS runs here. The result of the 'chip negative control' where a full chip was run with TE buffer in place of the usual library input (see Section 2.4.3) also indicates that the system is not inherently prone to contamination and that the results seen in this study are unlikely to be affected by contamination. Unlike in CE analysis, Ion Torrent MPS analysis negative control samples are pooled with other samples and run on the same chip, which means that the presence of negative controls in a batch can affect the aimed for concentration of the library pool entering templating. As such, the optimisation of this concentration, whilst allowing for the effect of negative controls in the batch could be an interesting avenue for future work.

Further to the MPS SNP section of the sensitivity analysis, the mitochondrial MPS assay shows even higher sensitivity for the MPS method. Full profiles (i.e. all variants) were observed for all samples in the dilution series (Table 11). This is likely more due to the inherent sensitivity of mitochondrial analysis than due to the sensitivity of MPS, as there are typically at least one hundred times more copies of the mtDNA genome per cell than the nuclear DNA. It is of note that although all variants were detected each time for all samples, the overall coverage of the mitochondrial assay was maximised for samples 4 and 5 (Figure 17). Coverage for sample 3, despite it containing more DNA, was about three-fold less than samples 4 and 5. This is understandable when the manufacturer's recommendations for the mitochondrial assay are taken into account (Section 2.4.1.5), these recommend the addition of only 0.1 ng of gDNA to the assay. This is based on there being approximately 100-fold more mtDNA in a cell than gDNA, which makes 0.1 ng of gDNA equal to approximately 10 ng of mtDNA. This recommended amount of DNA is close to what was in samples 4 and 5, so the fact that these were the best performing samples confirms the manufacturer's recommended input is optimal. In contrast, Sample 3 contained 0.25 ng of gDNA, 2.5x the recommended amount of DNA. This results in lessened coverage as reads are lost due to

the formation of ‘super-amplicons’ i.e. large amplicons formed of two or more ‘good’ amplicons randomly joining together. These amplicons are sequenced correctly, but as was shown in Section 3.2.5, cannot be aligned to the reference sequence in any one position and so are discarded by the analysis software, thus resulting in lowered overall coverage. These super-amplicons can be seen in the read histogram of the sequencing reports for the mitochondrial MPS runs (Figure 18 and Figure 19) and explain why the amount of input to the assay should be optimised.

Further examination of the mitochondrial results showed that there was no heteroplasmy evident in the control sample used in this work. Figure 21 and Figure 22 show the variant frequencies observed for the 36 variants detected in the mitochondrial sample, with Figure 21 showing the result for Sample 5 (62.5 pg of DNA input) and Figure 22 showing the result for Sample 5 (2 pg of DNA input). Figure 21, which was for the sample with the highest average coverage in the analysis, shows every variant detected with a frequency of 98% or more, with the exception of three variants: 460C, 6293C and 9438A. The sequence data for these three variants is shown in Figure 23. This figure shows that these three variants have lowered frequency (ranging from 76% to 91%) not due to heteroplasmy, but due to the homopolymer sequence in the mitochondrial genome at the areas of the variants. At these areas, while the majority of reads correctly call the variant in question, some reads detect an insertion or deletion due to errors in detecting the length of the homopolymer stretch. This results in the lowered variant frequency seen in the result. Figure 22, which shows the same data for Sample 10, the sample with the lowest DNA input and coverage in the study, shows the same pattern, with the same three variants having lower variant frequency than the others. It is of note however that the other variants in Figure 22 show more variability and range from 91% to 99%, in contrast to the 98% to 99% seen in Figure 21. This is likely the result of stochastic effects due the lowered DNA input to this sample. This variability of variant frequency with varied DNA input is something that would need to be carefully measured and validated if the MPS mitochondrial assay used here was implemented in forensic casework.

Despite there being no evidence of heteroplasmy in this work, it remains possible that there was heteroplasmy in the control sample used, just that it was in areas of the mitochondrial genome that correspond to primer binding sites for the Precision ID mtDNA Whole Genome assay used here. This is accounted for to some degree in the assay design, which has amplicons that overlap each other by on average 11 bp, but is still a risk with any tiled amplicon PCR-based assay such as this. One possible approach to address this was published by Huszar *et al.* (2019) who describe a method of bioinformatically selecting reads that span primer binding sites, thus filtering out reads that, by design, end with sequence

consistent with primer sequence, rather than the true underlying sequence of the sample. This method is called OREO, which stands for Overarching Read Enrichment Option. This work by Huszar *et al.* was performed with Promega mtDNA chemistry, rather than the Precision ID mtDNA Whole Genome assay used here, and further exploration of bioinformatic techniques like this could be an interesting avenue for future work with this kit.

The results gained in this work on the sensitivity of the MPS mitochondrial assay are in line with other who have studied the same area, with Pereira *et al.* (2018) reporting successful sequencing with the Precision ID mtDNA Whole Genome Panel (the same kit as used in this study) of samples with 6.25 pg of DNA, as measured by autosomal real-time PCR quantification. Pereira *et al.* also note in their analysis that “DNA input may have been lowered even further.” Others who have evaluated the same chemistry have not done a specific sensitivity study with a controlled input of DNA, (Strobl *et al.* 2018 and Woerner *et al.* 2018) but have examined case work style samples or samples with degraded or generally low levels of DNA, and have equally concluded that the chemistry shows high sensitivity. Lastly, Brandhagen and colleagues (Brandhagen *et al.* 2020) showed that this high sensitivity with mitochondrial analysis is not unique to Precision ID chemistry used here and in the above studies, in their analysis of the Promega PowerSeq CRM Nested kit, which also showed high sensitivity, with full mitochondrial profiles being consistently obtained down to 5000 mtDNA copies, as measured by an in-house mtDNA specific quantification.

Despite the high sensitivity of mitochondrial analysis, the downside of this analysis in a forensic setting is that even if a full mtDNA profile (or haplotype) is gained, as was achieved for every sample in this work, this offers much less discrimination power than a full SNP or STR profile. If DNA quantity is extremely limited however, mitochondrial analysis can be a very useful tool, and this work shows that analysing mtDNA via MPS is a viable technique. In practical terms it is certainly much easier to generate a full genome mitochondrial profile with MPS than with CE-based Sanger sequencing methods, and this point alone may be enough for laboratories to choose MPS methods over CE for this type of analysis.

The results from the inhibition analysis (Section 3.3) showed however that the MPS method performed much less well than the CE based method. For the given set of artificially inhibited samples, The CE method was entirely unaffected by the inhibition, with full profiles being obtained for all samples, both control and inhibited. This was definitely not the case for the MPS SNP assay however, with all inhibited samples showing significant loss of alleles, with all but the smallest amount of inhibition showing near complete loss of the profile (two total alleles observed across three samples). It seems clear that the library chemistry used in the MPS SNP assay is significantly less robust to inhibition than the CE-based PCR kit

chemistry. This is understandable given that CE-based STR chemistry has been used in the forensic community for around 20 years and the PCR chemistry has been significantly optimised to be robust to inhibition. The MPS chemistry used to date is much newer in a forensic context and has largely been directly inherited from non-forensic applications, where the need to be resistant to inhibition is much less.

Little work has been published on the result of inhibition on MPS analysis, something that has been pointed out in the few studies that do touch on it. Tao *et al.* (2019) note that “few validation studies have been carried out to determine the effects of PCR inhibitors on the NGS STR panels” and recommend that such studies should be performed in future. They briefly note in their work that urea has an inhibiting effect on MPS profiling, with full profiles only achieved with samples with less than 1000 ng/µL of urea. No direct comparison to CE technology was made however.

Zeng *et al.* (2018) touched on the topic of inhibitors in MPS profiling by studying the effectiveness of common forms of forensic DNA extraction on MPS analysis, particularly with regard to whether the extraction methods would remove inhibitors. Their conclusion was that DNA IQ (Promega), DNA Investigator (Qiagen), and PrepFiler BTA (Thermo Fisher Scientific) are all capable of removing inhibitors effectively for analysis with the ForenSeq DNA Signature Prep Kit (Verogen / Illumina) and Precision ID chemistry (Thermo Fisher Scientific). This matches the results of this work, where PrepFiler BTA was used for extraction of samples, and generally speaking, no inhibitory effect was seen other than in the samples in the inhibition study, where inhibitor was intentionally added to the reaction. The single exception to this was the quantitation IPC result for sample 9, which showed a slight inhibitory effect on the IPC (Table 35). This was not sufficient to have any noticeable effect on the resulting profile however, with full profiles being gained with both CE and MPS methods. The result of Zeng *et al,* in combination with the present work, makes it clear that those intending to perform forensic MPS analysis should be sure to use a high-quality extraction method to remove inhibitors before proceeding to sequencing.

Lastly, the one publication that has looked at the effect of inhibition on MPS analysis in detail is Sidstedt *et al.* (2019), who examined the effect of haematin and humic acid on analysis with the ForenSeq DNA Signature Prep kit. Their results matched the present work, with poor performance being shown by the MPS method in the presence of both inhibitors, something that was not seen when the equivalent samples were run with a CE-based assay, in this case the Powerplex Fusion kit (Promega). They concluded that the CE-STR kit was able to handle approximately 200x more inhibitor than the MPS assay.

As such, if the results of Sidstedt *et al.* and the present work are considered together, it seems clear that performance in the presence of inhibitors is something that affects both the Verogen / Illumina chemistry used by Sidstedt *et al.* and the Thermo Fisher Scientific chemistry used in this study. For successful implementation of these methods in forensic practice, it seems clear that attention should be paid by the manufacturers to make the resistance to inhibition of forensic MPS chemistry more comparable to that of the CE-based technology.

Like the results of the inhibition study, the results from the mixture analysis (Section 3.4) did not favour MPS. In Table 32 it can be seen that alleles from the lower level contributor to the mixture (the 007 control DNA) were observed down to the 100:1 level in the CE-analysis, and at that point 14 of the 19 unique alleles belonging to 007 were observed (73.7%). In the equivalent STR MPS processed 100:1 sample however, (Table 24) only 3 out of 23 007 alleles were observed (13.0%). On the surface, this indicates a worse ability of MPS to detect the lower level contributor to a mixture, something that is often of importance to a forensic mixed sample, where the lower of the two contributors to a mixture can be the suspect of interest in the case. This pattern was repeated in the microhaplotype MPS analysis, where in the 100:1 mixture only two out of 30 alleles uniquely belonging to the lower level contributor were detected (Table 28).

As such, the results show that when the same mixture of control DNA is analysed independently by these CE and MPS assays, the MPS assays are less well able to detect the alleles of the minor contributor to the mixture. It may be that the performance of these systems in detecting minor contributor alleles in a 1:100 mixture does not translate well to a ‘real world’ (i.e., non-control DNA) mixture, as in this study the genotypes of the control DNAs used were known, and so in detecting the minor alleles there was no ambiguity over whether a small peak in the profile could be a true minor peak or due to an artefact such as stutter. This ambiguity could hinder the analysis of a ‘real’ mixture with both MPS and CE such that detection of the minor contributor in the mixture ratios studied here was not possible.

That said though, given that this study directly compared the performance of the same DNA sample with MPS and CE, the results show the difference between MPS and CE, even if the ratios may not translate to real world mixtures. These results are also broadly in line with others who have looked at the performance of MPS mixtures. Silvery *et al.* (2020), for example studied the performance of artificially created control DNA mixtures with a custom STR panel, and found that 94% of minor contributor alleles could be detected in a 1:49 mixture. This was the weakest mixture that they studied, in terms of the strength of the minor

contributor. If the results here for the 1:50 in this work are compared, fewer alleles were observed, with 32 of a possible 55 minor contributor alleles were observed, or 58% (Table 24). Possibly the stutter thresholds used in the two analyses made a difference in this result, with Silvery *et al* reporting use of a relatively low 5% stutter threshold for their analysis, while in this work, peaks in stutter positions were removed entirely, due to the risk of peaks in this position as a result of stutter being mistaken for minor contributor peaks.

Hussing *et al.* (2018) also analysed similar MPS mixtures, at 1:25, 1:50, 1:100 and 1:1000 and other ratios. They presented result differently to the Silvery *et al.* study, but showed that the 1:1000 and 1:100 mixtures were not reliably detected as mixtures, often triggering the 'single source indicator' flag of the software that they used. Despite this, some evidence that the samples were mixtures was present, even if the full minor profile could not be reliably determined.

Kocher *et al.* (2018) performed an inter-laboratory study of mixture performance with the Verogen / Illumina ForenSeq DNA Signature Prep Kit and again looked at mixtures with ratios of 1:100, 1:500 and 1:1000. They concluded that almost no minor alleles were detected at the 1:500 and 1:1000 levels. They also, like this study, compared the result of the MPS mixture to the CE result, and concluded that "MPS did detect slightly less (minor contributor) alleles than CE", especially at the samples with weaker levels of minor DNA input.

Lastly, Tao *et al.* (2019) examined mixtures with the Thermo Fisher Scientific Precision ID GlobalFiler™ NGS STR Panel v2, a panel that was used elsewhere in this work, and an STR panel that is broadly similar to the STR component of the 'Mixture ID' panel used here. In mixtures of a 1:49 ratio, Tao and colleagues found that 34% to 42% of minor alleles were recovered, less than the 60% seen in this study, and significantly less than the 94% reported in the Silvery *et al.* study. The authors suggest that more library PCR cycles may have been beneficial in aiding that analysis.

As such, the mixture results seen in this study are broadly in line with other published studies, but this is also the first to examine the use of the Precision ID Mixture ID Panel in conjunction with comparison of the resulting profiles to CE analysis. Any slight differences in the result seen may be explained by experimental variation, or possibly by the choice of analytical threshold in the analysis.

This factor of analysis settings could have significant impact on the ability of these systems to detect mixtures. In the MPS section of the mixture study, a 'relative' analytical threshold of 2% was used (Table 22), whereas in the CE an 'absolute' analytical threshold of 50 rfu was

used (Section 2.3.3). This means that for CE, the cut-off to define a low level allele was 50 rfu regardless of the height of the tallest signal at that locus, whereas for MPS, the cut-off to define a low level allele was set at 2% of the height of the alleles signal at each locus. These settings are all as recommended by the manufacturer of the system, with the relative analytical threshold for the MPS assay being widely recommended for MPS analysis by a number of sources, for example in Hollard *et al.* (2019), who published developmental validation of the ForenSeq DNA Signature kit, while absolute thresholds are standard in CE, to the point that in the software used for this analysis (GeneMapper ID-X v1.5, Thermo Fisher Scientific, USA) it is not possible to set a relative analytical threshold, only an absolute one. This means that, by definition as currently implemented, the MPS assay is not able to detect minor contributor alleles less than 2% of the height of the major contributor in a mixed sample, something that may in part explain the performance of the MPS assays in this study.

The choice of relative rather than absolute thresholds for the MPS assays by the manufacturer may be due to the relatively poor locus-to-locus balance of the MPS assay compared to the longer-established CE assay. This can be seen in Figure 31, Figure 32, and Figure 33, where the inter-locus balance of the MPS STR and microhaplotype assays is seen to be much more variable than that of the CE STR assay. When performance of the assay varies between loci, relative analysis thresholds may be better able to filter out artefacts than absolute thresholds, which risk being set either too high to show the alleles at low performing loci, or too low to filter artefacts at high performing loci. It is likely for these reasons that the manufacturer of the MPS assays in this study has chosen to recommend a relative threshold for their use.

Having made this point, it is of note that the poor locus balance of the MPS assays in this study is mostly due to the underperformance of a small number of loci relative to the others in the kit. Especially, D22S1045 in the STR MPS assay (Figure 31) and mh13KK-217 in the microhaplotype MPS assay (Figure 32). It can be noted that in a newer version of an STR MPS assay from the same manufacturer, the Precision ID GlobalFiler™ NGS STR Panel v2 (Thermo Fisher Scientific, USA), the inter locus balance is improved, with D22S1045 no longer an outlier in performance to the rest of the kit. Figure 34 shows an example of this data from the manufacturer, with the D22S1045 locus highlighted in yellow. This indicates that inter-locus balance is something that may be improved by development from the manufacturer, likely in improved versions of primers for the loci in question, and opens the possibility of the use of absolute analysis thresholds, which may allow more sensitive mixture detection, in future versions of the assay.

In reviewing the results of the mixture study, a consistent pattern was also seen in the signal strength section component of the study (coverage in the case of the MPS runs, peak height in the case of CE data), with the signal strength for the minor contributor to the mixture trending downwards with lesser amounts of the minor contributor DNA, while the signal strength of the major contributor remained consistently high, all as could be expected (Table 25, Table 26, Table 29, and Table 30). As a result, it may be that the reported theoretical strengths of MPS in aiding mixture analysis (Silvery *et al.* 2020) are not as clear-cut as thought.

The three aspects of MPS performance described above, sensitivity, inhibition tolerance, and mixture detection are important factors in the evaluation of MPS technology, but in this study all were based on the use of artificial control type samples. The section in this work on non-probative samples (Section 3.5) attempted to expand on this theoretical basis and evaluate the performance of MPS on 'real world' samples. The results of this analysis showed a clear benefit in some samples for using the MPS method. For example, based on CE analysis, samples 3 and 6 (degradation indices of 55 and 26 respectively) gave a result that would not be used in a practical forensic case as the resulting partial profiles contained very few alleles. See Table 41: samples 3 and 6 are both described as 'non-useful partial profile'. Also see Figure 36, the CE electropherogram for sample 3 is shown, with only four above threshold peaks present. The MPS result for both samples 3 and 6 however showed over 30 alleles and each resulted in a useful partial profile that could be used in a forensic case. This can be seen in Table 41, where the result for both samples 3 and 6 is 'useful partial profile' and in Figure 35, where the MPS STR result for sample 3 is shown, with results at 12 loci. Given the amount of DNA in these extracts indicated by the quantitation result (0.012 ng/µL and 0.002 ng/µL for the small autosomal target in samples 3 and 6 respectively) this appears to be a practical demonstration of the enhanced sensitivity of MPS that is discussed above.

Other non-probative samples also showed benefit from MPS processing. Samples 2, 4, 10 and 11, for example, all showed signs of degradation in the CE result that were not apparent in the MPS result. These samples all showed indications of slight degradation in the quantitation result (degradation indices of 3, 15, 2 and 3 respectively), and this improvement of the MPS profiling result over CE is a demonstration of the benefit of the shorter fragment sizes that MPS analyses compared to CE. On the other hand, for some samples, for example samples 1, 5, 7, 8 and 9, while MPS did not give a worse result than CE, the extra alleles detected by MPS did not confer any particular benefit to the overall result for the sample (Table 41). Samples that were shown to be complex mixtures by CE were still shown as complex mixtures in the MPS analysis, and as shown in the previous discussion on

mixtures, there may be no particular benefit in detecting mixed contributors with MPS compared to CE.

So in this sense, the results of the non-probative analysis confirm the results of the previous analysis of sensitivity in that the increased sensitivity of MPS, as demonstrated on control DNA samples, is also seen in increased sensitivity of result on the non-probative samples. This is also the case in mixture detection, where there was no particular benefit to MPS analysis of control DNA samples, and equally, no particular benefit observed in MPS processing of the non-probative mixed samples. In this non-probative study there appeared to be no significantly inhibited samples, which showed generally equal or better profiling results for the MPS assays compared to the CE (Table 41). As the results in Section 3.3 showed, even the smallest amounts of inhibitor can seriously compromise the ability of the MPS assays to produce profiles, and as such the presence of these inhibitors can be assumed to be largely absent from the samples used here. This would typically be the case given the high quality DNA extraction and purification used on the samples (see Section 2.2.2). This confirms the work of Zeng *et al.* (2018), mentioned previously, who found that the PrepFiler BTA chemistry, also used in this study, was effective in removing inhibitors in samples for MPS analysis.

In this sense, the present work confirms the work of several sources that have showed the theoretical benefits of MPS methods in analysing degraded and low level DNA (Gettings *et al.* 2015; Zeng *et al.* 2015; Fordyce *et al.* 2015; Meiklejohn *et al.* 2015), but extends it to show that these benefits apply in real world samples, not just in a control DNA model. Some later studies have looked at non-probative style samples, as was studied here, and compared MPS methods with CE, such as Wang and colleagues (Wang, Chen *et al.* 2018), who found that of thirteen casework samples studied, five gave full results and seven showed dropout with CE, but twelve gave full profiles with MPS. No further analysis of the quality or usability of the samples was done however, which were all single source. Other studies have examined casework style non-probative samples, such as Muller *et al.* (2018), and like here, concluded that MPS shows good sensitivity and high recovery of alleles, but no direct comparison with CE results was made.

In summary, this work has shown that MPS offers significant sensitivity gains over CE analysis for forensic samples, but less clear benefits for analysis of mixed samples, and with present tests, significantly worse results from inhibited samples. As such, it is likely that when these factors are combined on real forensic samples, some samples will benefit from this type of analysis, while others may not. It may be the case that in practice a forensic lab

should choose to use both CE and MPS in parallel, with different samples used with different profiling methods depending on the specifics of the sample or case in question.

Chapter 4:

Evaluation of Analysis Metrics in Massively Parallel Sequencing

4. Evaluation of Analysis Metrics in Massively Parallel Sequencing

4.1. Introduction

The next chapter of this work examined analysis metrics that will need to be considered by forensic laboratories who wish to use MPS techniques in practical casework. These are metrics that are well established in the CE-methods commonly used by forensic laboratories today, and cover aspects of analysis that are fundamental to any analytical method. An example is signal-to-noise ratio, or in other words, how an analyst can be certain that the signal they see from their data is 'true' signal caused by the material they are testing for (in this case, human alleles in the DNA under examination) and not 'noise', that is, signal caused by reasons other than the DNA of interest, such as background electrical noise, signal from unincorporated primers, etc. The signal-to-noise cut-off, i.e. the point above which the analyst can be sure that signal they see from their system is true signal, is typically termed the analytical threshold or AT. A defined AT is used by every forensic laboratory that runs CE-based analysis to measure the output of their systems, and for successful use of MPS analysis, it would seem that laboratories must define AT in a similar way for MPS. The exact way in which AT should be defined for MPS and the factors that influence it are much less well known for MPS than for CE however. An exploration of these factors was the topic of this work.

Also investigated in this chapter was the topic of heterozygous balance, another threshold that is well established in CE-based forensic analysis. This is the ratio of signal that the two alleles give at a heterozygous locus, a ratio that ideally would be 1:1, but in practice can vary depending on factors such as the size of the alleles in question or the efficiency of the PCR that was used to generate the DNA fragments in the assay. The heterozygous balance of an assay is of particular importance in the analysis of mixed samples, because when samples that share alleles are mixed together, knowledge of the expected heterozygous balance of the assay in detecting single source samples is crucial in determining what alleles belong to which of the contributors to the mixture. The expected heterozygous balance of MPS systems was explored in this work.

Lastly, in this chapter, the metrics of reproducibility and concordance of MPS analysis were explored. These are topics that are again a feature of typical validations of CE-based forensic methods, and as such are also important in the implementation of MPS methods. Again though, to date, much less work has been done on these topics for MPS than for CE-based methods. Reproducibility is the expectation that measuring the same sample multiple times with the same method will give the same result, both in terms of the genotype achieved, but also in terms of the analysis metrics, such as AT discussed above. For these

metrics to be of use in forensic analysis, the metrics must be able to be established in validation runs of the system in question, and then be able to be applied to subsequent runs of the same system with the confidence that the metric definition is applicable to the new data. As such, understanding of the reproducibility of an MPS method is vital for its successful practical implementation. The reproducibility of MPS analysis was explored in this chapter. Concordance was also explored here. This is the expectation that analysis of a given forensic marker for a given sample will provide the same result regardless of the type of analysis used to type the marker. As the majority of forensic DNA analysis at present is done with STR markers on CE-based systems, this work compares the STR genotype achieved with a CE-based STR assay, to an MPS-based STR assay that has multiple markers in common. This, in principle, should provide the same genotype for the same samples. This is not necessarily always the case however (Parson *et al.* 2016), something that has implications for comparison of samples generated with MPS to older samples, possibly those stored on a DNA database, that were generated with CE-methods.

4.2. Analysis thresholds

To investigate various performance metrics of MPS to do with ensuring the quality of analysis, specifically: background noise, analytical thresholds, and heterozygous balance, a variety of samples were processed with either varying concentrations of Control DNA 007 (Thermo Fisher Scientific, USA) or as multiple replicates of 1 ng Control DNA 007. The samples were processed with the Precision ID Identity panel and the Precision ID Ancestry panel as described in Section 2.4.1.3.

Parameters of the HID_SNP_Genotyper v5.2.2 for secondary analysis (i.e. genotyping) were as follows (all manufacturer's default values):

Table 42: Parameters used in HID_SNP_Genotyper v5.2.2 for Precision ID Ancestry Panel and Precision ID Identity panel analysis. All are the manufacturer's recommended default parameters.

Parameter	Value used
Minimum allele frequency	0.1
Minimum coverage	6
Minimum coverage either strand	0
Maximum strand bias	1
Trim reads	true

Among 16 replicates of 1 ng control DNA, all samples gave full, clear, single source genotypes consistent with those expected from the control DNA used. The resulting data was then analysed to calculate the average background noise for each sample and each run, where background noise is the number of reads sequenced which show a base other than that expected from the genotype at the locus in question. For example, in Table 43 below, results from the first eight loci of one sample in this analysis are shown. For each locus, the software measures the number of A, C, G and T bases detected at the SNP base in question. The larger numbers in each row represent the known genotype of the sample in question at that locus. The numbers in the other columns represent background noise. In this analysis, the magnitude of the background noise was examined to attempt to determine the analytical threshold, which represents the point at which true signal, caused by the genotype of the sample being analysed, can be distinguished from the background noise.

In this analysis, the analytical threshold was calculated as the mean noise observed, plus ten times the standard deviation of the noise (Butler, 2015). Heterozygous balance is the balance in signal between the two alleles of a heterozygous genotype, and was calculated at each heterozygous locus as the coverage of the smaller of the two alleles divided by the coverage of the larger allele (i.e. as: minimum coverage at locus / maximum coverage at locus).

Table 43: Example of raw data examined for background noise in this section. Each row represents the result for one locus in one sample, only eight example rows are shown – each profile comprises 124 loci and so 124 rows. The genotype of the sample in question is known and is given in the 'Genotype' column. Total coverage for the locus is shown in the 'Coverage' column. This coverage is then broken down by each of the four bases in the 'Base detected at SNP' columns. The bases corresponding to the genotype represent the true signal of the result. The other bases represent background noise.

Sample No.	Locus	Genotype	Coverage	Base detected at SNP			
				A	C	G	T
1	rs1490413	AG	1472	698	0	774	0
1	rs7520386	AG	1618	1316	0	301	1
1	rs4847034	AA	2675	2669	0	1	5
1	rs560681	AA	3333	3332	0	1	0
1	rs10495407	AG	3393	1838	0	1553	2
1	rs891700	AA	2116	2109	0	4	3
1	rs1413212	CC	2030	1	2019	3	7
1	rs876724	CC	1289	1	1288	0	0

The samples described above were then run in different combinations on different sequencing chips so that the resulting analytical thresholds could be calculated and compared. Firstly, two runs were done using the same S5 sequencing chip (the 530 chip) each with exactly the same run conditions, other than that one chip contained seven of the Precision ID Ancestry samples and the other chip contained fifteen of these samples. Every sample contained 1 ng of input control DNA. Results were as follows:

Table 44: Analysis metrics for 1 ng control samples analysed with the Precision ID Ancestry panel on a 530 sequencing chip. The middle column shows the metrics calculated when seven 1 ng samples were run on the chip. The right column shows the metrics calculated when fifteen 1 ng samples were run on the chip. Analytical threshold and average coverage are rounded to the nearest integer. Mean and s.d. noise are rounded to one decimal place. Heterozygous balance is rounded to two decimal places.

Metric	Run with 7x 1 ng samples on 530 chip	Run with 15x 1 ng samples on 530 chip
Mean of background noise	10.9	4.5
Standard deviation of noise	49.8	29.0
Analytical threshold	509	295
Average sample coverage	11607	4209
Average heterozygous balance	0.76	0.78

Next, the same set of sixteen 1 ng input Precision ID Identity samples were run on two sequencing runs that were identical to each other, except that one used the ‘520’ model of S5 sequencing chip, while the other used the ‘530’ model of S5 sequencing chip. These two chip types are offered by the manufacturer (Thermo Fisher Scientific, USA) as being of differing ‘capacities’, i.e. as being capable of generating differing numbers of sequencing reads. The user can choose the chip type that they prefer depending on the number of reads they need to generate for the run in question. The 520 chip is specified by the manufacturer as providing four to six million reads per run, while the 530 chip is specified as providing fifteen to twenty million reads per run. There is also the practical consideration that the 530 chip has a higher cost to purchase than the 520 chip.

Running the same sample set on both of these chip types gave the following results:

Table 45: Analysis metrics for the same number of 1 ng control samples analysed with the Precision ID Identity panels on runs with two different sequencing chips – the 520 and 530 sequencing chip. The middle columns show the analysis metrics calculated when sixteen 1 ng samples were run on a 520 chip, the right column shows the analysis metrics calculated when sixteen 1 ng samples were run on a 530 chip. Analytical threshold and average coverage are rounded to the nearest integer. Mean and s.d. noise are rounded to one decimal place. Heterozygous balance is rounded to two decimal places

Metric	Run with 16x 1 ng samples on 520 chip	Run with 16x 1 ng samples on 530 chip
Mean of background noise	2.5	7.9
Standard deviation of noise	5.5	19.6
Analytical threshold	57	204
Average sample coverage	2711	8220
Average heterozygous balance	0.88	0.88

Next, the effect of differing amounts of DNA in the samples, and the effect of different MPS panels was examined. A sample set was prepared with the same control DNA described above, but this time with a dilution series of 2 ng, 1 ng, 750 pg, and 500 pg of input DNA, each duplicated to give eight samples in total. These eight samples were then used as input into two different MPS assays – the Precision ID Identity panel and the Precision ID Ancestry panels. This gave two different sets of eight libraries. One set from each of the two panels, but each with the same DNA input. These libraries were then each run on both the 520 and 530 chip described above, to give four sequencing runs in total. Analysis metrics were then calculated for each of the four runs, which gave the following results:

Table 46: Overall analysis metrics for the same sample set run four times with two different sequencing chips – the 520 and 530 sequencing chip, and with two different panels – the Precision ID Identity panel and the Precision ID Ancestry panel. The sample set run consisted of four dilutions of control DNA (2 ng, 1 ng, 750 pg and 500 pg) run in duplicate. Analytical threshold and average coverage are rounded to the nearest integer. Mean and s.d. noise are rounded to one decimal place. Heterozygous balance is rounded to two decimal places.

Metric	Run with 8x dilution series samples			
	Precision ID Identity Panel		Precision ID Ancestry Panel	
	520 chip	530 chip	520 chip	530 chip
Mean of background noise	5.6	13.8	3.5	15.3
Standard deviation of noise	11.8	26.5	8.1	72.2
Analytical threshold	124	279	85	737
Average sample coverage	5237	16660	3121	10544
Average hetero’ balance	0.86	0.86	0.84	0.85

The same data set was then analysed to examine the effect that the differing amount of DNA in the samples had on the coverage and noise metrics. These results were as follows:

Table 47: Per sample analysis metrics for the same sample set run four times with two different sequencing chips – the 520 and 530 sequencing chip, and with two different panels – the Precision ID Identity panel and the Precision ID Ancestry panel. The sample set run consisted of four dilutions of control DNA (2 ng, 1 ng, 750 pg and 500 pg) run in duplicate. Average coverage is rounded to the nearest integer. Mean and s.d. noise are rounded to one decimal place.

Panel	Chip type	Sample input (ng)	Average sample coverage	Mean of background noise	Standard deviation of noise
Precision ID Identity Panel	520	0.5	4507	4.7	7.4
		0.75	5052	5.0	7.5
		1	5375	5.8	8.2
		2	6014	6.1	9.0
	530	0.5	14670	11.3	16.0
		0.75	16231	13.1	18.5
		1	17054	13.7	18.2
		2	18683	15.3	20.1
Precision ID Ancestry Panel	520	0.5	2992	3.5	6.1
		0.75	2418	2.8	5.7
		1	3096	3.3	5.6
		2	3978	3.7	4.9
	530	0.5	10157	16.3	58.1
		0.75	8327	12.3	44.0
		1	10393	13.8	48.5
		2	13297	14.7	39.4

4.3. Reproducibility

To investigate the reproducibility of MPS analysis, specifically with regard to the sample coverage, 32 reference samples were processed with 1 ng of extracted DNA in each sample. The samples were processed with the Precision ID GlobalFiler STR NGS Panel v2 as described in Section 2.4.1.2. All samples gave full, clear, single source genotypes. Parameters of the Converge v2.1 analysis were as follows (all manufacturer's default values):

Table 48: Parameters used in Converge v2.1 analysis for Precision ID GlobalFiler STR NGS Panel v2 reproducibility analysis. All are the manufacturer's recommended default parameters.

Parameter	Value used
Target file	Precision_ID_GlobalFiler_NGS_STR_Panel_Targets_v1.1
Hotspot file	Precision_ID_GlobalFiler_NGS_STR_Panel_Hotspot_v1.1
STR flank length	Per locus setting
STR flank tolerance	2
STR analytical threshold	0.02
STR stochastic threshold	0.05
STR stutter ratio	0.2

The 32 sample libraries obtained in this analysis were then run twice on two separate sequencing chip runs. The overall quality of the two runs, and the coverage obtained for each sample was then compared between the two sequencing runs. The overall quality of the two sequencing runs were as follows:

Table 49: Overall sequencing metrics for two runs in reproducibility study. Each run contained the same 32 identical 1 ng libraries on the chip.

Parameter	Run 1	Run 2
Chip loading (%)	47%	52%
Enrichment (%)	99%	99%
Polyclonal (%)	35%	32%
Low Quality (%)	53%	48%
Total reads sequenced	5,136,915	6,442,696
Total bases sequenced	641 Mb	803 Mb
Mean Raw Accuracy	96.6%	96.6%
Test fragment reads	21,815	21,702
Test fragment reads 50AQ17 (%)	99%	99%

Coverage metrics for each of the 32 samples on the two runs were then compiled, with a consistent 23% to 27% difference in coverage between the two runs being seen for every sample. Results were as follows:

Table 50: Sample-by-sample coverage metrics achieved for the same 32 sample libraries run twice with identical conditions on two sequencing runs in reproducibility study. Total coverage for each sample on each of the two runs is shown, as is the % difference in Coverage between the two runs (Table continues on next page).

Sample No.	Total sample coverage		Difference between runs (%)
	Run 1	Run 2	
1	111,330	140,341	26%
2	19,679	24,586	25%
3	86,077	108,972	27%
4	90,825	112,139	23%
5	140,734	178,076	27%
6	112,322	140,115	25%
7	100,691	126,145	25%
8	103,800	131,329	27%
9	100,909	125,656	25%
10	97,028	121,214	25%
11	142,407	176,528	24%
12	136,796	169,679	24%

Sample No.	Total sample coverage		Difference between runs (%)
	Run 1	Run 2	
13	156,967	196,979	25%
14	149,547	187,113	25%
15	141,125	176,462	25%
16	149,880	188,825	26%
17	157,535	198,555	26%
18	130,903	163,971	25%
19	133,178	164,301	23%
20	52,770	66,300	26%
21	160,689	199,467	24%
22	150,230	188,150	25%
23	147,577	185,045	25%
24	146,478	184,744	26%
25	140,842	176,716	25%
26	78,456	99,051	26%
27	67,546	83,943	24%
28	70,765	87,935	24%
29	59,044	74,254	26%
30	80,108	100,115	25%
31	121,348	152,412	26%
32	88,239	109,838	24%
Total Coverage	3,625,825	4,538,956	24%

This data was then plotted in a chart of sample number against coverage. The same sample-to-sample pattern can be seen for the two runs, with Run 2 being consistently approximately 25% higher than Run 1:

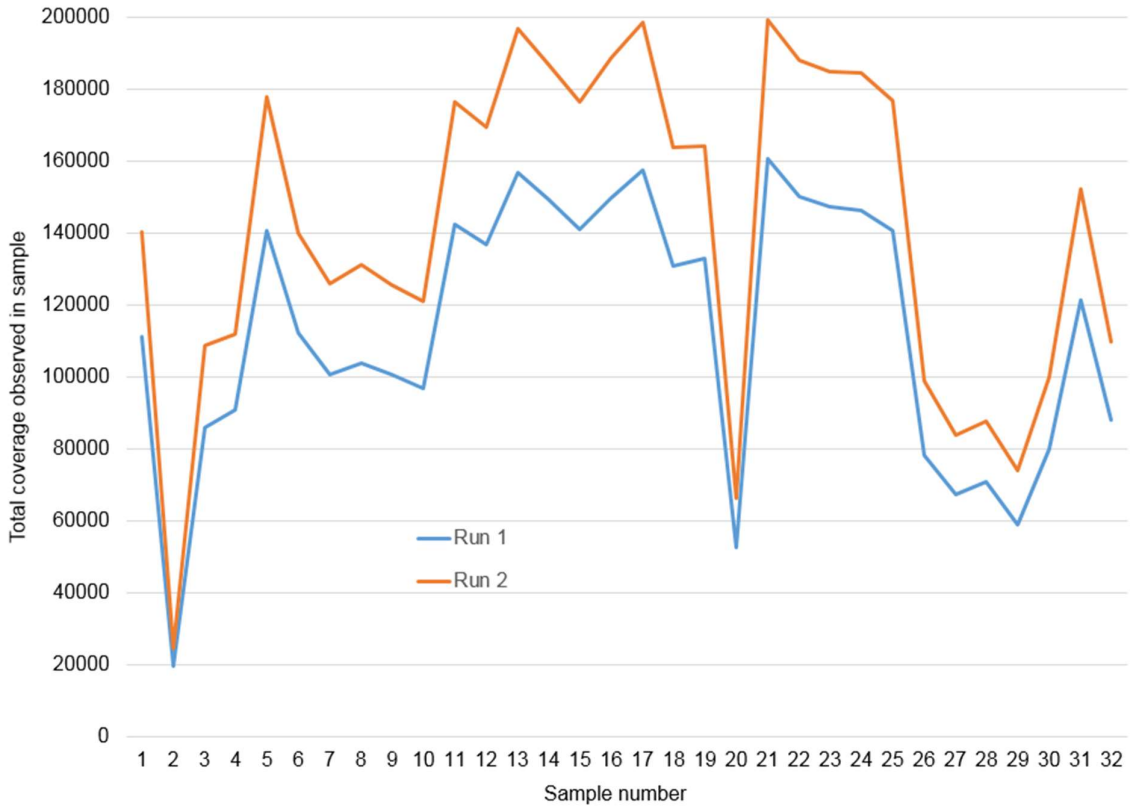


Figure 37: Chart of sample-by-sample coverage metrics achieved for the same 32 sample libraries run twice with identical conditions on two sequencing runs in in reproducibility study. The 32 samples are shown on the horizontal axis. Total coverage for each sample is plotted on the vertical axis

4.4. Concordance

To investigate the concordance of MPS analysis to CE analysis for STR, i.e. whether the same sample can be expected to give the same STR genotype with both MPS and CE methods, 25 reference samples were processed both an MPS and CE STR assay. 1 ng of extracted DNA was used to process each sample. The samples were processed with the Precision ID GlobalFiler STR NGS Panel v2 (an MPS assay) and with the GlobalFiler PCR amplification kit (a CE assay), as described in Section 2.4.1.2 and Section 2.3 respectively. All samples gave full, clear, single source genotypes with both MPS and CE methods. Parameters of the Converge v2.1 analysis were as follows (all manufacturer's default values):

Table 51: Parameters used in Converge v2.1 analysis for Precision ID GlobalFiler STR NGS Panel v2 concordance analysis. All are the manufacturer's recommended default parameters.

Parameter	Value used
Target file	Precision_ID_GlobalFiler_NGS_STR_Panel_Targets_v1.1
Hotspot file	Precision_ID_GlobalFiler_NGS_STR_Panel_Hotspot_v1.1
STR flank length	Per locus setting
STR flank tolerance	2
STR analytical threshold	Per locus setting
STR stochastic threshold	Per locus setting
STR stutter ratio	0.2

Note that the parameters used in this analysis differ slightly from those used in previous sections, in that the STR analytical and stochastic thresholds are a 'per locus setting' rather than 2% and 5% respectively for every locus as was used previously (Table 36 and Table 48). This is due to the release by the manufacturer after the previous work had been completed of new recommended settings for these parameters. Specifically, the settings used here were the 'v2.1.0' version of Converge settings for this panel (full name: Precision_ID_GlobalFiler_NGS_STR_Panel_AnalysisParams_v2.1.0.json).

Results were analysed by comparing the CE-generated genotypes to the MPS genotypes for each sample. For the MPS result only the number of STR repeats, rather than the specific sequence, was considered, this being the result that is able to be directly compared to the CE result.

Results were as follows. First, an example full result for one sample is shown. As can be seen, the CE result matches the MPS result at every locus.

Table 52: Example result from concordance analysis. Profiles for MPS and CE for the same sample are shown. Only loci common to both CE and MPS assays are shown (some loci are in one assay and not the other – refer to Table 68 and Table 71 in Appendix for details of kit contents). Results with no entry in the 'Allele 2' column are homozygote loci.

Sample No.	Locus	CE Result		MPS Result	
		Allele 1	Allele 2	Allele 1	Allele 2
1	CSF1PO	12	13	12	13
	D10S1248	14	15	14	15
	D12S391	16	23	16	23
	D13S317	12	12	12	12
	D16S539	9	11	9	11
	D18S51	12	14	12	14
	D19S433	14	-	14	-
	D1S1656	14	17.3	14	17.3
	D21S11	29	32.2	29	32.2
	D22S1045	11	14	11	14
	D2S1338	18	25	18	25
	D2S441	10	-	10	-
	D3S1358	15	16	15	16
	D5S818	10	12	10	12
	D7S820	9	12	9	12
	D8S1179	13	14	13	14
	FGA	21	24	21	24
	TH01	9.3	-	9.3	-
	TPOX	8	11	8	11
	vWA	16	17	16	17

All 25 samples in the concordance experiment were analysed in the same way as shown in Table 52 by comparing the MPS to CE result at all common loci for all samples. The CE and MPS result matched exactly at every locus for every sample, with the exception of the results shown below in Table 53.

Table 53: Differences seen between MPS and CE result in Concordance analysis. All rows show a homozygous result in CE (Allele 2 entry is empty), but a heterozygous result for MPS, where the length-based name of both of the heterozygous alleles is the same however. Rows marked 'Iso. Het.' are examples of Isometric heterozygotes, as discussed below. Rows marked 'SNP' are examples of flanking region SNPs, also discussed below.

Sample No.	Locus	CE Result		MPS Result		Conclusion
		Allele 1	Allele 2	Allele 1	Allele 2	
10	D21S11	29	-	29	29	Iso. Het.
12	D3S1358	16	-	16	16	Iso. Het.
13	D8S1179	13	-	13	13	Iso. Het.
16	D5S818	11	-	11	11	SNP
17	D8S1179	12	-	12	12	Iso. Het.
19	D12S391	20	-	20	20	Iso. Het.
22	D5S818	12	-	12	12	SNP
23	D3S1358	16	-	16	16	Iso. Het.
24	vWA	15	-	15	15	Iso. Het.
25	D3S1358	16	-	16	16	Iso. Het.

The ten results in Table 53 where the CE and MPS result differed were all investigated in greater detail. All ten results were at loci detected as homozygous in CE analysis. Eight of these results were found to be examples of STR sequence variation that is not visible to the CE analysis, also known as isometric heterozygotes. This is where the genotype is a heterozygote consisting of two alleles that share the same size, which is the same in CE analysis, but have differing sequence and so can be distinguished by MPS analysis. An example of one of the observed isometric heterozygotes is given below.

Allele	Status	Coverage	Sequence
28	STUTTER	74	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10
29	ABOVE_ST	546	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10
29	ABOVE_ST	558	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11

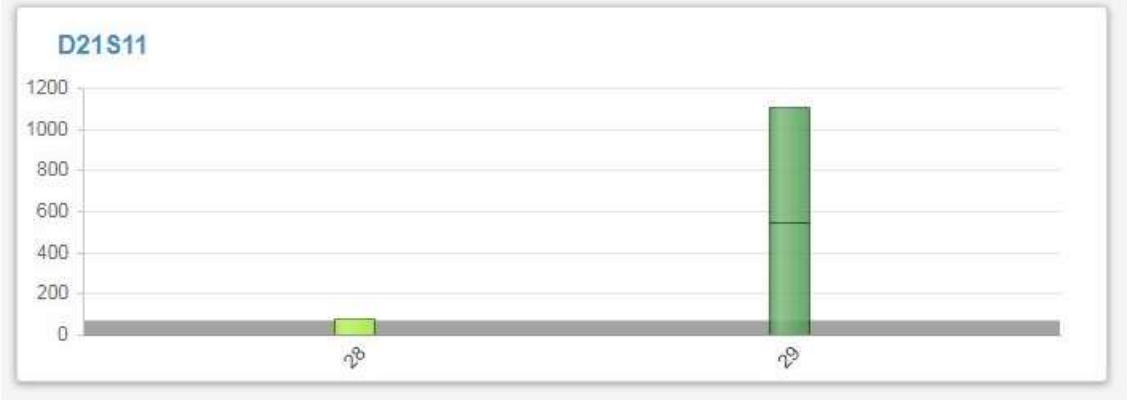


Figure 38: The MPS result for sample 10 at D21S11 (see row 1 of Table 53). An isometric heterozygote is observed in this result – two forms of the '29' allele that can be distinguished by different repeat region sequences

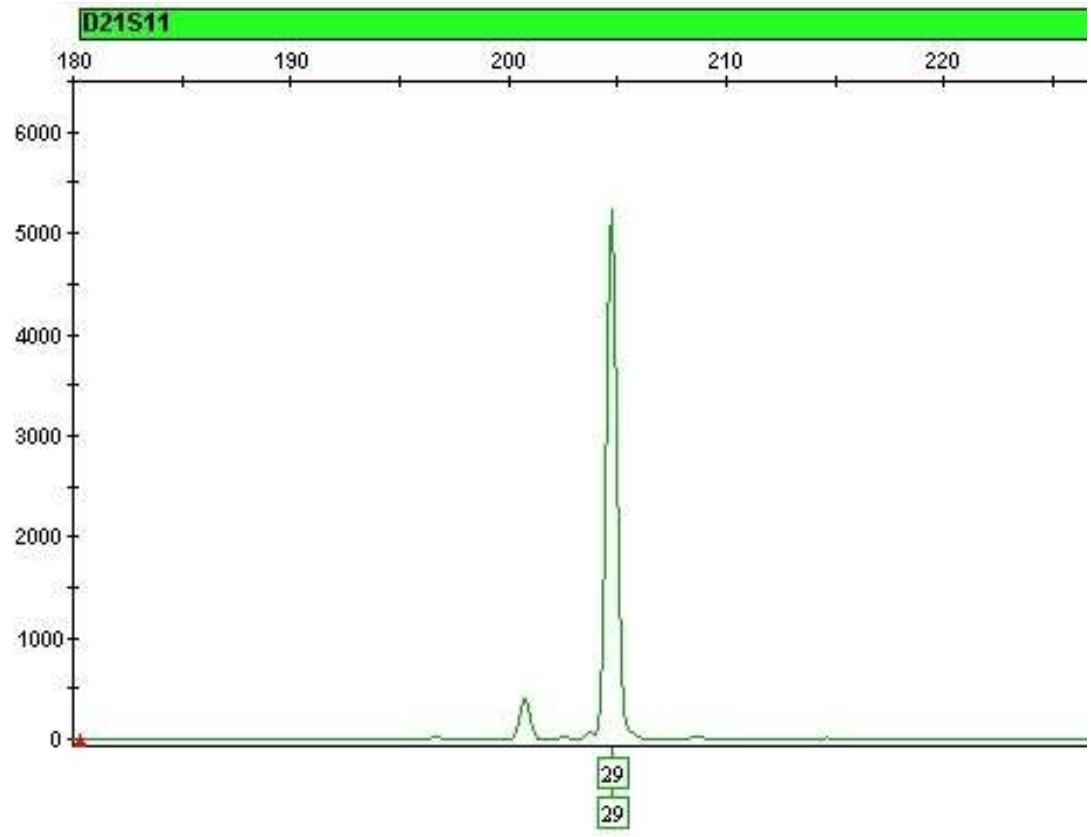
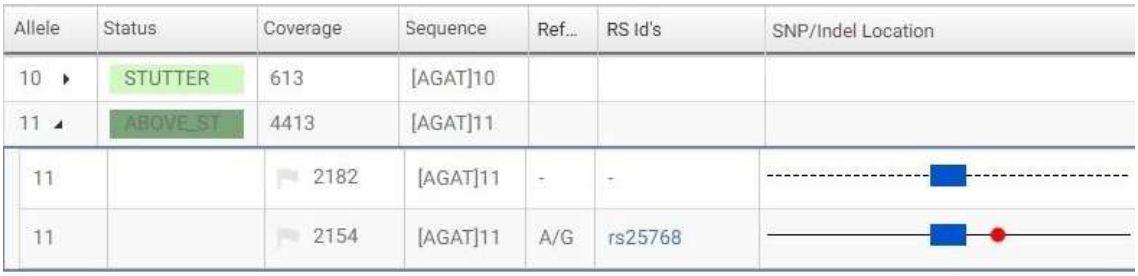


Figure 39: The CE result for sample 10 at D21S11 (see row 1 of Table 53). A 29 homozygote is observed. CE is not able to detect the sequence variation that is seen in the MPS result in Figure 38.

Two of the ten results recorded in Table 53 were found upon investigation to be due to detection of SNPs in the flanking region of the locus for the MPS result. This is another manner in which MPS can detect sequence variation that is not seen by CE, this time due to variation in the flanking region of the STR, not in the repeat region itself. The result is the same as for the isometric heterozygotes shown above however in that two true alleles are detected by MPS, where only one allele is visible to CE. An example is shown below:



Allele	Status	Coverage	Sequence	Ref...	RS Id's	SNP/Indel Location
10	STUTTER	613	[AGAT]10			
11	ABOVE_ST	4413	[AGAT]11			
11		2182	[AGAT]11	-	-	
11		2154	[AGAT]11	A/G	rs25768	

Figure 40: An example of a 'SNP' based difference between CE and MPS in concordance analysis. The MPS result for the D5S818 locus in sample 16 is shown (see row 4 of Table 53). Two forms of the 11 allele are observed by the software. One allele where an A/G SNP is observed in the flank region of the STR (2154 reads) and one allele where the SNP is not observed – i.e. the sequence at the same base matches the reference (2182 reads).

Figure 40 shows that an 11 allele was detected in the MPS result for the D5S818 of sample 16, with total coverage of 4413. In addition to the STR repeat region, the software also looks at the flanking region of the locus for variation in the observed sequence compared to the reference sequence. In this case, 2154 reads of the 4413 total at that locus showed an A to G SNP in the right flank of the locus, a SNP that has previously been characterised and given the name rs25768 (see 'RS Id's' column in Figure 40). 2182 reads were also observed with the 11 allele in the repeat region but with no A to G SNP in the same position – in these reads the same base was an A, matching the reference genome. As such, two distinct alleles were detected by MPS in the D5S818 genotype for this sample – one allele with the SNP in the flanking region and one allele without the SNP. Both CE to MPS differences marked 'SNP' in Table 53 (samples 16 and 22) followed the same pattern. As such it can be concluded that all of the apparent differences found between MPS and CE in this analysis can be explained by fundamental differences between the detection of CE and MPS, and that no unexplained discordances were observed.

4.5. Discussion

The most significant result obtained in the analysis of analytical threshold in this work was the finding that analytical threshold can vary significantly based on the specific nature of the run and samples in question. Both the number of samples used in a run, and the chip type used in the run can significantly alter the appropriate threshold to use based on analysis of the signal-to-noise that the run produces. This can be seen first in Table 44, where the same run type was performed twice (using 530 chips in each case) but with seven and fifteen samples on the two runs respectively. This resulted in an analytical threshold of 509 for the run with seven samples, but an analytical threshold of 295 for the run with fifteen samples. Equally, Table 45 shows the result where the same set of 16 samples was analysed on two different runs, one run with a 520 sequencing chip, which is specified by the manufacturer to produce four to six million total sequencing reads, and one run with a 530 sequencing chip, which is specified to produce fifteen to twenty million reads. In this case the analytical threshold was 57 for the run with the 520 chip and 204 for the run with the 530 chip.

This result fits with the specification for reads just mentioned in that the approximately three to four-fold increase in sequencing reads with the 530 chip, also resulted in a 3.6 fold increase in the analytical threshold. This also correlates with the average sample coverage seen in the two runs, which was 2711 and 8220 for the 520 chip and 530 chip runs respectively – again an approximate threefold increase (Table 45).

The above results were obtained with samples that were all of the same strength – reference samples with the 1 ng of DNA input recommended by the manufacturer. Further work examined the impact on the analytical threshold of samples of different input amounts together on the same chip, and also looked at the impact that the specific MPS assay used had on the result. The results of this are seen in Table 46, where four runs of the same dilution series of eight samples were performed. These runs were again done using both the 520 and 530 sequencing chips, as above, and also with two different MPS assays – the Precision ID Identity panel and the Precision ID Ancestry panel, as described in Sections 2.4.1.3 and 2.4.1.4. The results obtained were similar to those for the 1 ng samples, in that the calculated analytical threshold was different in each of the four runs, indicating that the analytical threshold can also vary depending on the sequencing run types and assays used for samples of different input amounts. The analytical thresholds seen in this section ranged from 85 for the Precision ID Ancestry panel on a 520 chip, to 737 for the same assay on a 530 chip, a more than eight-fold difference.

The result for the dilution series of the Precision ID Ancestry panel on a 530 chip (Table 46) can then also be compared to the previously discussed result with the 1 ng samples, where

seven 1 ng samples were run, again with the Precision ID Ancestry panel on a 530 chip (Table 44). In this way, the effect of using entirely 1 ng samples versus the dilution series for the same type of run can be seen, albeit with seven samples in one run and eight in the other. In this case, the average sample coverage was similar with the seven 1 ng samples giving average coverage of 11607, while the dilution series gave slightly lower average coverage of 10544. This small difference in average coverage may be explained entirely by there being only seven samples in the 1 ng set and eight in the dilution set – the fixed amount of reads on the chips are shared between fewer samples in the 1 ng set, resulting in slightly higher average coverage. Despite this, the analytical thresholds produced for the two runs however were more widely different to each other, with the 1 ng set of samples giving an analytical threshold of 509 and the dilution series giving an analytical threshold of 737 – an approximate 1.5x difference.

Given that the average signal strength was about the same in the two runs, this 1.5x difference in threshold can be explained by the differing amount of noise in the two runs. In the 1 ng sample set run mean noise was 10.9 reads with standard deviation of 49.8 (Table 44), while in the dilution run, mean noise was 15.3 with standard deviation of 72.2 (Table 46). This difference was further examined in Table 47 where the noise and coverage metrics for the individual samples in the dilution series can be seen. Here it is shown that while, as could be expected, coverage for the samples generally increases with increasing DNA input, so too does noise, with the stronger samples in the dilution series showing increased noise for the increased DNA input. There is no other obvious reason for this variation in 'noisiness' between the two runs, and both runs had general sequencing metrics in line with expectations for the type of run. It may be that the 2 ng input samples in the dilution samples produced more noise by nature of being above the manufacturer's recommended 1 ng of DNA input – increased signal results in increased noise. Equally, it can be noted that the 1 ng sample in the dilution series produced higher noise than the full run of 1 ng samples with the same panel and chip types. This was 3.3 for the 520 chip and 13.8 for the 530 chip for the 1ng in the dilution series (Table 47), compared to 2.5 for the 520 chip and 10.9 for the 530 chip for the full 1ng set of samples (Table 44 and Table 45). It may be that more noise was produced on the chip with the dilution series because this chip contained samples with differing amounts of DNA, rather than having every sample at the manufacturer's recommended input, as was the case on the 1ng input run. This would be an important factor to consider in a practical forensic validation of this technology, as practicing labs routinely analyse samples with vastly differing amounts of DNA present.

Possible reasons for the noise in this analysis include misincorporation of bases, either in the initial library PCR or in the sequencing reaction itself – it is known the polymerases used

in these steps have error rates, which can be in the order of one error per million bases amplified (McInerney *et al*. 2014). Given that several hundred million bases were sequenced in each of the sequencing runs studied here, it follows that there could be several hundred misincorporation errors in the final results – a source of the noise readings shown in Table 43. It is further possible that this type of error could vary per locus analysed. It is known that certain areas of the genome are more difficult to sequence than others – repeat motifs and homopolymer stretches, for example, are more difficult to sequence accurately than 'random' sequence stretches with no readily apparent pattern (Goldfeder *et al*. 2016). As such, certain loci in the assays used for forensic application could have sequence motifs which mean that the level of noise varies per locus. This then raises the possibility that locus specific thresholds may be appropriate for routine forensic MPS analysis, something that has typically not been the case with CE-based analysis. Measurement of this effect would require a large data set containing many measurements for each of the loci in the assay and could be an interesting avenue for further work in this area.

Another possible explanation for the noise observed in this analysis is low-level contamination, most likely occurring during the sequencing run. As described in Section 1.1.3.5, the Ion Torrent sequencing method used here detects the next base in the unknown sequence by sequentially flowing each of four dNTPs over the sequencing chip and detecting which are incorporated into the sequencing reaction. Between these so-called 'flows' the instrument will wash away the previous flow with a neutral buffer solution, cleaning the chip for the next flow. In practice, given that there are tens of millions of wells on the sequencing chip however, it seems likely that in at least some of these wells the previous flow is imperfectly washed away and some of the dNTP remains for the next flow. This would result in an error in the base detected in that well for the next flow. The magnitude of this type of error is unknown, but it seems likely that at least some of the noise seen in this type of experiment is due to this mechanism.

It is further possible that the magnitude of the errors noted here could vary from instrument to instrument, in the same way that it is known that peak heights, and the corresponding appropriate thresholds, can vary from instrument to instrument in CE analysis (Shewale *et al*. 2012). This is due to minor differences in factors such as the reagent quality, electrical resistance, and thermal properties of different instruments, which add up to a difference in the observed output of the system. Again, this would be an interesting area for further analysis, something which to date has not been covered in the forensic literature, perhaps for the practical reason that few forensic institutes today have the multiple MPS instruments that would be needed to conduct such work.

As such, it is clear that variation in the noise of an MPS run can be observed and that this may be due to factors such as misincorporation of bases, low-level contamination, or simply due to random fluctuation. This variation in noise needs to be accounted for by forensic laboratories that use MPS in setting and using their analysis thresholds.

Another source of variation that was seen in the result shown in Table 46, where the same dilution series was run four times in total with two different chip sizes and two different assays, was the assay itself. The two assays used were the SNP-based Precision ID Ancestry panel and Precision ID Identity panel. Although, as noted, the analytical threshold varied between the four runs, no clear pattern based on the assay could be seen, with the Precision ID Ancestry panel having a higher threshold than the Precision ID Identity panel for the 530 chip run, but a lower threshold for the 520 chip. One pattern that did emerge was that the Precision ID Identity panel did have clearly higher average coverage for both the 520 and 530-chip runs, something that could perhaps be expected due to the smaller number of loci in this panel compared to the Precision ID Ancestry panel (124 SNP loci compared to 165 SNP loci – see Section 2.4.1.3 and 2.4.1.4 for details). It is clear that different MPS assays will have different properties in this regard and so must be individually validated when implemented in forensic practice.

The result that analytical threshold can vary based on the number of samples in the run, the type of chip used in the MPS run, and the specific MPS assay used is an important one, and raises some issues that will be new to forensic analysts who have only used CE methods in the past. In particular, the need to define different analysis settings based on the number of samples in the batch is entirely foreign to CE analysis, where every sample is self-contained in the run and one sample does not affect another. As such, in CE the same analytical threshold can be used however many samples are analysed simultaneously. In the MPS systems evaluated here however, libraries that have been prepared separately are pooled together on the same chip for sequencing, with a fixed amount of chip 'real estate' available to the sample pool. If this finite space on the chip is taken up by more or fewer samples, it has been shown here that this will impact the analytical threshold that must be used to define the existence of alleles in the resulting profiles.

This possibility of varying analytical thresholds in MPS depending on specific run conditions has been noted in forensic literature, with Peter de Knijff noting the following in a 2019 review paper (de Knijff, 2019):

"(With MPS) one also has to set an analytical threshold, in this case the number of reads with an identical sequence structure. It strongly depends on the experimental design. If one has pooled many different DNA samples for database purposes into a single MPS run, one

expects less reads per sample and per locus (in case of a multiplex STR design), compared to a run with only a few case samples pooled."

That said, de Knijff does not specifically measure the size of this effect, and also notes that "clear and concise guidelines (for this) are not yet available". It is of note that there are very few publications that address this issue. One that does is by Young and colleagues, (Young *et al.* 2017), who examine setting analytical threshold for STR data on the Verogen / Illumina MiSeq FGx Forensic Genomics System. They recommend a system of defining AT based on subtraction of the minimum observed noise from the maximum observed noise, all multiplied by a scaling constant, but do not test this on runs with varying numbers of samples or read throughput, as described here. Another recent paper that looks at the issue is that of Riman *et al.* (2020) who describe several possible methods for setting AT, and discuss the issues involved, but again do not deeply examine the issue of differing number of samples in a run, or make any definitive conclusion on what method of setting AT should be used.

It seems clear that this issue of appropriate setting of AT depending on run conditions is something that will have to be carefully considered by all laboratories validating MPS methods, and may result in the labs validating methods that allow for fixed numbers of samples being processed at one time.

Subsequent to this result, it can be noted however that the heterozygous balance measured in all of this analysis was not affected by the factors mentioned above, such as the number of samples in the run, the strength of the samples, or the chip type. In Table 44, where the results comparing the effect of running seven or fifteen 1 ng samples on two otherwise identical runs are shown, the heterozygous balance observed in both runs was near identical – 0.76 on the seven sample run and 0.78 on the fifteen sample run. Equally, in Table 45 the heterozygous balance for two runs of the same sixteen 1 ng samples on a 520 and a 530 chip were 0.88 in both runs. Lastly, for the four runs shown in Table 46, where eight dilution samples were run with the Precision ID Ancestry panel and the Precision ID Identity panel on 520 and 530 chips, the heterozygous balance across the four runs were 0.84, 0.85, 0.86 and 0.86. As such it can be concluded that the number of samples in the run, the strength of the samples, or the chip type do not affect this metric.

This is intuitive given that heterozygous balance is largely a function of the PCR earlier in the sequencing workflow, the PCR that makes the DNA fragments that go on to be sequenced. Factors that affect heterozygous balance are typically factors that affect this PCR, such as changes in the amount of input DNA, changes to the thermal cycling or reaction volume of the PCR, or the design of the primers in the kit being used. Once these DNA fragments are made in the PCR, they will go on to be sequenced as they stand. Even if the total number of

fragments that are sequenced from any one sample may alter depending on the number of samples in the run, as we have seen earlier in this section, the ratio of the two alleles at a heterozygous locus should be unchanged regardless of how many samples are run. As such, it seems that heterozygous balance is an analysis metric that forensic labs employing MPS can continue to set and monitor in the same way as they may have done in the past with CE.

The next section of this work examined the reproducibility of an MPS run. This is something of interest to forensic analysts and a requirement of the validation process, especially given the factors discussed above on analysis thresholds. If thresholds are to be experimentally determined and applied, it is necessary for the reproducibility of the system to be well understood so that a threshold defined on one run can be properly applied to other runs. This work ran two identical sequencing runs of the same sample library set so that the two results could be compared. Both runs contained 32 1 ng reference samples processed with the Precision ID GlobalFiler NGS STR kit v2. Table 49 shows the high-level sequencing metrics for the two runs. Both runs were of good quality and met the specifications for the type of sequencing run made by the manufacturer (Thermo Fisher Scientific, USA) in that the number of test fragment reads, the quality of the test fragments reads (50AQ17) and the mean raw accuracy of the runs all met expectation and were near identical to each other. Despite this however, the total number of reads sequenced for the two runs were significantly different to each other in that 5.13 million reads were produced on the first run, and 6.44 million reads were produced on the second – a 25% increase from Run 1 to Run 2. Both numbers meet the manufacturer's specification for the type of run, but still represent a large difference between two runs that were identical in their preparation and run consecutively on the same instrument with the same set of reagents.

Table 50 and Figure 37 show the sample-by-sample breakdown of how this total coverage for the chip was spread across the 32 samples on each chip, and it can be seen that in both cases, the total reads for the run are spread relatively evenly across all samples, with the 25% increase in the total reads in Run 2 being reflected in a 23% to 27% increase in reads for each of samples in Run 2 compared to the same sample in Run 1. Figure 37 shows that there were some sample specific effects in the two runs – sample 2 and sample 20 for example have a notably lower number of reads than the other samples in the run, but because this pattern of lower reads in some samples is exactly repeated across the two runs, always with a 23-27% increase in reads from Run 1 to Run 2, this can be concluded to be due to variability in the sample libraries, not a sample-specific variability in the sequencing. Instead, rather than a sample related effect, it has been demonstrated in this experiment that the overall coverage of an MPS sequencing run can vary by at least 25%,

something that should be taken into account when laboratories are validating these systems and designing analysis parameters for routine use. Future work could extend the work here by doing more runs with identical parameters to build a larger data set and further characterise the variation in coverage that can be seen with MPS methods.

It is of note in this discussion of reproducibility, that several publications have addressed the topic of reproducibility of forensic MPS methods, but mostly from the perspective of genotyping – i.e. does a given sample produce the same genotype if profiled multiple times, without considering the effect of any variation in the total number of reads that the sequencing run has produced. Examples of this are seen in Hussing *et al.* (2018), Kocher *et al.* (2018), and Wang, Chen *et al.* (2018). No forensic publications have been found which address the issue of reproducibility of sequencing metrics and as such, this area seems a good candidate for future study.

Lastly in this chapter, analysis was done on the concordance of MPS methods, with a run of MPS STR samples compared to a run of the same samples with a CE-based STR assay. Specifically, the samples were processed with both the Precision ID GlobalFiler STR NGS Panel v2 (an MPS assay) and with the GlobalFiler PCR amplification kit (a CE assay), both as described in Chapter 2. The genotyping results of the two runs were then compared at all twenty STR loci that appear in both assays (see the Appendix for full details of all the loci in these assays). As can be seen in Table 53, the concordance of the MPS and CE runs was high with only ten differences being found between the profiles in the 500 loci compared (20 common loci x 25 samples tested). This meant that 490 out of 500 of the loci tested matched exactly between the MPS and CE result (98%).

As can also be seen in Table 53, the ten differences noted above were not discrepancies between MPS and CE, but rather sequence variation that was invisible to the CE assay, consisting of eight so-called isometric heterozygotes, where the genotype in question is a heterozygote consisting of two alleles that share the same size. This means that these two alleles are indistinguishable by CE methods, resulting in an apparent homozygous genotype in that analysis, but are able to be correctly distinguished by MPS as heterozygous alleles. Figure 38 and Figure 39 show an example of one of these isometric heterozygotes, all of which were similar to that in the figures. As such, these differences between the MPS and CE result were not indicative of an inconsistency in either system, but an illustration of how MPS can potentially be used in STR analysis to gain more information on a sample than is possible with CE.

The remaining two differences of the total ten noted above and in Table 53 were marked in the analysis as ‘SNP’. This meant the difference was due to detection of SNPs in the

flanking region of the locus for the MPS result. This is another example where, similar to the above case of isometric heterozygotes, variation in the flanking region of the STR resulted in two distinct alleles with MPS, alleles that were indistinguishable by CE. This is shown in Figure 40. The software indicates a SNP in the flanking region of the 11 allele, with 2154 of the 4413 total reads for the allele showing the SNP. The remaining 2182 reads at the locus represent the 11 allele without the SNP.

As such, because the ten differences described above are expected features of MPS analysis of STRs, not inconsistencies, the concordance seen in this work is 500 out of 500 loci tested, or 100%. This compares favourably to other studies in this area. One study by Wang, Chen, and colleagues (Wang, Chen *et al.* 2018) reported a concordance of 97.86% (183 out of 187 loci) using an early-access version of the Precision ID STR chemistry used here. Barrio and colleagues performed a large study in this area, finding that 5078 out of 5083 loci, or 99.9%, were concordant (Barrio *et al.* 2019). This study used the same chemistry as used here, and the percentage locus concordance seen is very similar and perhaps shows the improvement from the early access version of the chemistry used by Wang, Chen *et al.* to the 'release' version used here and by Barrio *et al.* Devesse and colleagues examined the concordance of another MPS STR chemistry, the ForenSeq DNA Signature Prep Kit, and also reported it to be highly concordant, reporting a concordance of 16451 out of 16453 alleles (99.98%) (Devesse *et al.* 2018). This is the number of concordant alleles found, not loci as in the other publications, but still demonstrates a high level of concordance of the system.

It the current work, it is of note that the latest analysis parameters published by the manufacturer were used in this analysis (Table 51), settings that were aimed at reducing the number of unexpected artefacts that were seen in analysis with previous versions of the settings. This is described in a technical note released by the manufacturer in 2019 (*Performance of the Precision ID GlobalFiler NGS STR Panel v2: Artifacts, Thresholds and Chip Loading*, available from www.thermofisher.com). To date, no other studies are known that have taken advantage of these updated settings. This further illustrates one of the central points of this chapter – the importance of carefully defining analysis parameters and thresholds in MPS analysis.

In summary, this chapter has shown that the analytical threshold that should be applied to any analytical method in the forensic DNA laboratory can vary significantly in MPS analysis due to factors such as the number of samples in a run, the strength of the samples in that run, the panel which is used in the analysis, the specific analysis parameters that are used, and fluctuation in the run-to-run performance of the instrument in question. This results in

issues of measuring appropriate analytical threshold that will be unfamiliar to analysts used only to working with CE-based analysis, where the samples in a given instrument run do not affect each other in the way that can happen in MPS analysis. These issues of defining the appropriate analytical threshold are not insurmountable however, but do require careful validation through rigorous testing of the same conditions under which real samples would be run – replicating the same number of samples and sample strengths in the validation as would be used in casework, for example. By careful consideration of these factors and thorough validation of the appropriate analysis metrics, it will be possible for forensic laboratories to take advantage of the benefits offered by MPS analysis.

Chapter 5:

Evaluation of the Precision ID Ancestry Panel for Ancestry Prediction

5. Evaluation of the Precision ID Ancestry Panel for Ancestry Prediction

5.1. Introduction

The next chapter of this work examined the ability of MPS methods to infer the ancestry of a sample donor. The ability of MPS to analyse large numbers of markers simultaneously has made ancestry analysis practical for forensic laboratories. This is because in the past, analysis was limited in the number of markers that could be successfully analysed in a CE-based reaction. With the higher capacity of MPS, it is now possible to analyse hundreds, or even thousands, of markers in one reaction, and in doing so, it is theoretically possible to genotype many ancestry indicative markers, which individually may only give a small amount of evidence of the ancestry of the sample, but when taken together can provide accurate inference of ancestry. This work examined one of the commercial solutions for forensic ancestry analysis, the Precision ID Ancestry panel (Thermo Fisher Scientific, USA), and evaluates its ability to accurately predict the ancestry of the sample donor for a diverse panel of 64 samples where the self-declared ancestry of the sample donor was known in each case.

The Precision ID Ancestry panel contains 165 SNP loci in total, with 123 of these comprising the so-called 'Seldin' panel, a set of SNPs that were proposed in 2009 for ancestry inference by Michael Seldin and colleagues (Kosoy *et al.* 2009), and 55 comprising the 'Kidd' panel, which was proposed by Kenneth Kidd in 2011 (Kidd *et al.* 2011). 13 SNPs appear in both the Kidd and Seldin panels, which is how the 123 SNP and 55 SNP sub-panels combine for a total of 165 SNPs. The manufacturer's recommendation is for all of the SNPs in the panel (so for both Seldin and Kidd sub-panels) to be analysed simultaneously to provide a combined estimate of ancestry. In this work, this combined analysis is examined, as is the power of each of the Seldin and Kidd panels to infer ancestry individually.

5.2. Genotyping of Ancestry samples

To investigate the performance of MPS in ancestry prediction, 64 samples were processed that were collected from donor individuals of varying ethnicities. The self-reported ethnicity of the donors was collected and is shown in the following results tables. Samples were processed with the Precision ID Ancestry panel as described in Section 2.4.1.3. Parameters of the HID_SNP_Genotyper v5.2.2 for secondary analysis (i.e. genotyping) were as follows (all manufacturer's default values):

Table 54: Parameters used in HID_SNP_Genotyper v5.2.2 for Precision ID Ancestry Panel ancestry analysis (secondary analysis). All are the manufacturer's recommended default parameters.

Parameter	Value used
Minimum allele frequency	0.1
Minimum coverage	6
Minimum coverage either strand	0
Maximum strand bias	1
Trim reads	true

All samples gave full, clear, single-source genotypes.

5.3. Ancestry prediction using manufacturer recommendation

Predicted ethnicities from the genotypes were then generated with the HID_SNP_Genotyper v5.2.2 (tertiary analysis). Parameters for this were as follows (all manufacturer's default values):

Table 55: Parameters used in HID_SNP_Genotyper v5.2.2 for Precision ID Ancestry Panel ancestry analysis (tertiary analysis). All are the manufacturer's recommended default parameters.

Parameter	Value used
Target file	PrecisionID_AncestryPanel_targets_v1.0
Hotspot file	PrecisionID_AncestryPanel_hotspots_v1.0
Algorithm	Admixture Prediction - AISNPs

The ancestry prediction algorithm in HID_SNP_Genotyper v5.2.2 works by taking allele frequencies for the SNPs in the panel stored within the software from seven major ethnic groups around the world (frequencies obtained from 1000 genomes (www.internationalgenome.org) and ALFRED (alfred.med.yale.edu). These seven ethnic groups are:

- Native American (referred to as 'America' in the software)
- East Asia
- Oceania
- Africa

- Europe
- South West Asia
- South Asia

The software then considers the genotype for which it has been asked to make a prediction of ethnicity. The software first simulates all possible admixture combinations of at most four of the above seven ethnic groups, in increments of 5%. In doing so, it makes a simulated set of allele frequencies for each combination that it considers. The software then chooses which of the simulated admixture combinations best describes the observed genotype. It does this by selecting the admixture combinations that gives the highest random match probability for the genotype in question. This admixture combination is then presented to the user as the most likely ethnicity of the sample donor.

Results for this analysis for the 64 samples are shown in the following table. The self-declared ethnicity of each sample donor is also shown. Where declared ethnicities are shown with two entries separated by a "/", this means that the donor declared themselves to have one parent of each of the ethnicities shown. The 'Software prediction' columns of the table show the result of the MPS software ancestry prediction described above. Each of the seven possible ethnic groups are shown, each scored from 0% to 100% in the combination that the software simulation concluded would best represent the ethnicity of the donor in question. The result column of the table records then either a 'Match', 'No match', or 'Unclear' result gained by comparing the donor's declared ethnicity to the software's predicted result.

Table 56: Results for Precision ID Ancestry panel ancestry analysis. Am = America, EA = East Asia, Oc = Oceania, Af = Africa, Eu = Europe, SA = South Asia, SWA = South West Asia. Predictions are % of each ethnic group the software predicts are admixed in the donor's ethnic background. Each row sums to 100%. Cells left blank represent 0%. Where declared ethnicities are shown with two entries separated by a "/", this means that the donor declared themselves to have one parent of each of the ethnicities shown. The result column records either a 'Match', 'No match', or 'Unclear' result of comparing the donor's declared ethnicity to the SNP result (Table continues over next two pages).

No.	Self-declared donor ethnicity	Software prediction							Result
		Am	EA	Oc	Af	Eu	SWA	SA	
1	Afghan	5	15			45		35	Unclear
2	Afghan	10				35	35	20	Unclear
3	Afghan					25	65	10	Unclear
4	Albanian					55	45		Match
5	Angolan				100				Match

No.	Self-declared donor ethnicity	Software prediction							Result
		Am	EA	Oc	Af	Eu	SWA	SA	
6	Asian / Black			15	40	25		20	Unclear
7	Bangladeshi		30		5			65	Match
8	Black British				100				Match
9	Bosnian					80	20		Match
10	British Asian					55	35	10	Match
11	Canadian Caribbean			25	65	10			Match
12	Chinese	15	85						Match
13	Chinese		100						Match
14	Chinese		100						Match
15	Chinese Malay / French		40			25		35	Match
16	Cornish / Burmese		10	5		70		15	Unclear
17	Dutch / Vietnamese		40			60			Match
18	Egyptian						100		Match
19	Ethiopian				40		50	10	Match
20	Ethiopian				50		50		Match
21	French / British					70	30		Unclear
22	Ghanaian				100				Match
23	Greek			70		30			No Match
24	Greek Cypriot			5		70	20	5	Match
25	Indian	5	25			30		40	No Match
26	Iran / UK					85	15		Match
27	Iraqi					30	70		Match
28	Iraqi			5			80	15	Match
29	Irish					90	10		Match
30	Irish			5		95			Match
31	Irish					100			Match
32	Irish / Black British				55	45			Match
33	Irish / Chinese		45			55			Match
34	Irish / Thai		35	5		40		20	Match
35	Italian / Filipino		40	15	5	40			Match
36	Kenyan				55		45		Match

No.	Self-declared donor ethnicity	Software prediction							Result
		Am	EA	Oc	Af	Eu	SWA	SA	
37	Kuwaiti		10			30	60		Match
38	Mixed Race British				10	50	35	5	Match
39	Mongolian	5	85	10					Match
40	Morocco / Korea		40			15	45		Match
41	Nigerian				100				Match
42	Nigerian				100				Match
43	Nigerian				100				Match
44	Pakistani							100	Match
45	Pakistani					20	45	35	Match
46	Polish					100			Match
47	Somalian				50		50		Match
48	Somalian				55		45		Match
49	Somalian		5		60		35		Match
50	South Italian					50	50		No Match
51	Spanish					70	30		Unclear
52	Sri Lankan	5			10			85	Match
53	Tanzanian				90		5	5	Match
54	Ugandan				100				Match
55	Ugandan				100				Match
56	White Brit. / Black Brit.				50	50			Match
57	White British					100			Match
58	White British	5				75	20		Unclear
59	White British					95	5		Match
60	White British					100			Match
61	White British					100			Match
62	White British					100			Match
63	White British					100			Match
64	White British					100			Match

5.4. Ancestry prediction using custom parameters

Predicted ethnicities from the same genotypes were then generated, again using HID_SNP_Genotyper v5.2.2, but this time using parameters that differ from those provided by the manufacturer. The algorithm used to generate the ethnicities was the same as Section 5.3, but in this analysis customised grouping of SNP sets were used. These involved analysing the 'Kidd' and 'Seldin' SNP sets (discussed in Section 1.1.4.1.2 and detailed in Table 72) individually. These SNP sets contain 123 SNPs in the case of the Seldin panel and 55 SNPs in the case of the Kidd panel, and under the manufacturer's recommendation, are analysed together as one large panel in the Precision ID Ancestry assay. In this work, customised parameter files were created which allowed the two panels to be considered independently for the same genotypes tested in the previous section.

Results for this analysis for the 64 samples are presented in the following table, where three results are shown for each of the 64 samples, following the same pattern seen in Table 56. The result marked 'C' represents the 'Combined' result, i.e. the result for the entire 165 SNP panel analysed together, as shown in Table 56, while the results marked 'K' and 'S' represent the results for the Kidd and Seldin sub-panels respectively.

Table 57: Results for Precision ID Ancestry panel ancestry analysis, showing results for each sample for the whole combined panel (C), the Kidd panel SNPs only (K) and the Seldin panel SNPs only (S). Am = America, EA = East Asia, Oc = Oceania, Af = Africa, Eu = Europe, SA = South Asia, SWA = South West Asia. Predictions are % of each ethnic group the software predicts are admixed in the donor's ethnic background. Each row sums to 100%. Cells left blank represent 0%. Where declared ethnicities are shown with two entries separated by a "/", this means that the donor declared themselves to have one parent of each of the ethnicities shown. The result column records either a 'Match', 'No match', or 'Unclear' result of comparing donor's declared ethnicity to the SNP result (Table continues over next six pages).

No.	Self-declared donor ethnicity	SNP Set	Software prediction							Result
			Am	EA	Oc	Af	Eu	SWA	SA	
1	Afghan	C	5	15			45		35	Unclear
		K		15					80	Unclear
		S	10		15		75			Unclear
2	Afghan	C	10				35	35	20	Unclear
		K					60	5	30	Unclear
		S	5				20	15	60	Unclear
3	Afghan	C					25	65	10	Unclear
		K					20	80		Unclear
		S					35	45	20	Unclear

No.	Self-declared donor ethnicity	SNP Set	Software prediction							Result
			Am	EA	Oc	Af	Eu	SWA	SA	
4	Albanian	C					55	45		Match
		K					70	30		Match
		S					45	55		Match
5	Angolan	C				100				Match
		K				100				Match
		S				100				Match
6	Asian / Black	C			15	40	25		20	Unclear
		K				40	35	25		Unclear
		S			25	30		35	10	Unclear
7	Bangladeshi	C		30		5			65	Match
		K		25	10				65	Match
		S	10	20					70	Match
8	Black British	C				100				Match
		K				100				Match
		S				100				Match
9	Bosnian	C					80	20		Match
		K					60	35	5	Match
		S					90	10		Match
10	British Asian	C					55	35	10	Match
		K					70	30		Match
		S		10			55	35		Match
11	Canadian Caribbean	C			25	65	10			Match
		K			15	70	15			Match
		S			30	60		10		Match
12	Chinese	C	15	85						Match
		K	10	90						Match
		S	20	80						Match
13	Chinese	C		100						Match
		K		100						Match
		S		95		5				Match

No.	Self-declared donor ethnicity	SNP Set	Software prediction							Result
			Am	EA	Oc	Af	Eu	SWA	SA	
14	Chinese	C		100						Match
		K		100						Match
		S	5	95						Match
15	Chinese Malay / French	C		40			25		35	Match
		K		40			20		30	Match
		S		40			25		35	Match
16	Cornish / Burmese	C		10	5		70		15	Unclear
		K		10			80		10	Unclear
		S		5	15		60	20		Unclear
17	Dutch / Vietnamese	C		40			60			Match
		K		45	10		45			Match
		S	5	25			70			Match
18	Egyptian	C						100		Match
		K						50	45	Unclear
		S					5	95		Match
19	Ethiopian	C				40		50	10	Match
		K				25		75		Match
		S				40		55		Match
20	Ethiopian	C				50		50		Match
		K				30		60	10	Match
		S				55		45		Match
21	French / British	C					70	30		Unclear
		K					60	40		Unclear
		S					75	25		Unclear
22	Ghanaian	C				100				Match
		K				100				Match
		S				100				Match
23	Greek	C			70		30			No Match
		K	75				25			No Match
		S			100					No Match

No.	Self-declared donor ethnicity	SNP Set	Software prediction							Result
			Am	EA	Oc	Af	Eu	SWA	SA	
24	Greek Cypriot	C			5		70	20	5	Match
		K					70		30	Match
		S			5		65	30		Match
25	Indian	C	5	25			30		40	No Match
		K		20		15	35		30	No Match
		S	5	15			40		40	No Match
26	Iran / UK	C					85	15		Match
		K					80	20		Match
		S					100	0		No Match
27	Iraqi	C					30	70		Match
		K					45	55		Match
		S					10	80	10	Match
28	Iraqi	C			5			80	15	Match
		K	10					90		Match
		S					15	65	20	Match
29	Irish	C					90	10		Match
		K					100			Match
		S					80	20		Unclear
30	Irish	C			5		95			Match
		K			5	5	90			Match
		S					95		5	Match
31	Irish	C					100			Match
		K					100			Match
		S			5		95			Match
32	Irish / Black British	C				55	45			Match
		K				55	40	5		Match
		S				50	40	10		Match
33	Irish / Chinese	C		45			55			Match
		K	5	50			45			Match
		S		40			60			Match

No.	Self-declared donor ethnicity	SNP Set	Software prediction							Result EA
			Am	EA	Oc	Af	Eu	SWA	SA	
34	Irish / Thai	C		35	5		40		20	Match
		K		50			45	5		Match
		S		15	10		25		50	No Match
35	Italian / Filipino	C		40	15	5	40			Match
		K		40	15	5	40			Match
		S		35	20		25	20		Unclear
36	Kenyan	C				55		45		Match
		K				45		55		Match
		S				60		40		Match
37	Kuwaiti	C		10			30	60		Match
		K		15			10	75		Match
		S		10			35	55		Match
38	Mixed Race British	C				10	50	35	5	Match
		K				5	20	60	15	Match
		S			10	10	60	20		Match
39	Mongolian	C	5	85	10					Match
		K	10	80	10					Match
		S	5	90	5					Match
40	Morocco / Korea	C		40			15	45		Match
		K		45			20	35		Match
		S		30		5		55	10	Match
41	Nigerian	C				100				Match
		K				100				Match
		S				100				Match
42	Nigerian	C				100				Match
		K				100				Match
		S				100				Match
43	Nigerian	C				100				Match
		K				100				Match
		S				100				Match

No.	Self-declared donor ethnicity	SNP Set	Software prediction							Result
			Am	EA	Oc	Af	Eu	SWA	SA	
44	Pakistani	C							100	Match
		K							100	Match
		S							100	Match
45	Pakistani	C					20	45	35	Match
		K					40	50	10	Match
		S					5	55	40	Match
46	Polish	C					100			Match
		K					100			Match
		S					100			Match
47	Somalian	C				50		50		Match
		K				40		60		Match
		S				50		50		Match
48	Somalian	C				55		45		Match
		K				40		30	30	Match
		S				50		50		Match
49	Somalian	C		5		60		35		Match
		K			5	65		30		Match
		S		5		60		35		Match
50	South Italian	C					50	50		No Match
		K		5			45	50		No Match
		S					30	70		No Match
51	Spanish	C					70	30		Unclear
		K	10				30	45	15	No Match
		S					100			Match
52	Sri Lankan	C	5			10			85	Match
		K				5			95	Match
		S	10			10			80	Match
53	Tanzanian	C				90		5	5	Match
		K				90		5	5	Match
		S				95			5	Match

No.	Self-declared donor ethnicity	SNP Set	Software prediction							Result
			Am	EA	Oc	Af	Eu	SWA	SA	
54	Ugandan	C				100				Match
		K				90		10		Match
		S				95		5		Match
55	Ugandan	C				100				Match
		K				100				Match
		S				100				Match
56	White / Black British	C				50	50			Match
		K				45	55			Match
		S				45	30	25		No Match
57	White British	C					100			Match
		K				5	95			Match
		S					100			Match
58	White British	C	5				75	20		Unclear
		K	5				95			Match
		S	5		10		60	25		Unclear
59	White British	C					95	5		Match
		K					90	10		Match
		S					100			Match
60	White British	C					100			Match
		K					100			Match
		S					100			Match
61	White British	C					100			Match
		K					95	5		Match
		S					100			Match
62	White British	C					100			Match
		K					100			Match
		S				5	95			Match
63	White British	C					100			Match
		K					100			Match
		S					100			Match
64	White British	C					100			Match
		K					100			Match
		S					100			Match

The results from Table 57 are then summarised in the following table, which counts the number of 'Match', 'No Match, and 'Unclear' results seen in the analysis for each of the Combined panel, Kidd panel and Seldin panel.

Table 58: Summary of results seen in Ancestry analysis. Total number of samples with a result of 'Match', 'No Match' and 'Unclear' by customer parameter and original ('Combined panel') analysis in Table 57 are shown.

Panel used in Analysis	No. of 'Match' results	No. of 'Unclear' results	No. of 'No match' results
Combined panel	53	8	3
Kidd panel	53	7	4
Seldin panel	49	9	6

The consistency of the result for each of the 64 samples across the three panels was then examined and is summarised in the table below.

Table 59: Summary of results seen in Ancestry analysis. The number of samples that had the a result of 'Match' for all three panels in Table 57 are shown, as are the number of samples that had a result of 'Unclear' or 'No match' across all three panels. Also shown are the number of 'Inconsistent' samples – these are samples in Table 57 that had differing results across the three panels tested.

Result	Samples
No. of samples 'Match' for all three panels	47
No. of samples 'Unclear' for all three panels	6
No. of samples 'No Match' for all three panels	3
No. of samples with inconsistent result across the three panels	8
Total Samples	64

Next, the eight samples shown in Table 59 as having an inconsistent result across the three panels tested – i.e. where all three sub-panel results were not identical to each other, are shown in the table below.

Table 60: The eight samples with inconsistent results across the three Ancestry panels tested, as shown in Table 57 and summarised in Table 59. The result for each sample for each of the three panels tested ('Combined', 'Seldin' and 'Kidd') is shown, as is the sample donor's self-declared ethnicity.

Sample No.	Ethnicity	SNP Set	Result
18	Egyptian	Combined	Match
		Kidd	Unclear
		Seldin	Match
26	Iran / UK	Combined	Match
		Kidd	Match
		Seldin	No Match
29	Irish	Combined	Match
		Kidd	Match
		Seldin	Unclear
34	Irish / Thai	Combined	Match
		Kidd	Match
		Seldin	No Match
35	Italian / Filipino	Combined	Match
		Kidd	Match
		Seldin	Unclear
51	Spanish	Combined	Unclear
		Kidd	No Match
		Seldin	Match
56	White / Black British	Combined	Match
		Kidd	Match
		Seldin	No Match
58	White British	Combined	Unclear
		Kidd	Match
		Seldin	Unclear

Given that the results shown so far were evaluated with knowledge of the declared ethnicity of the donors, it is possible that there was an element of unconscious bias in this evaluation. To examine whether there was any bias in the determination of the results to date as a 'Match' or 'No Match', the same results were subjected to a blind evaluation by another party. In this analysis, an experienced independent analyst reviewed the 'Software prediction' columns of Table 57, without seeing the 'Self-declared donor ethnicity' column

and in doing so, categorised the software prediction for each of the samples as belonging to one of the seven ethnic categories described in Section 5.3 (America, East Asia, Oceania, Africa, Europe, South West Asia and South Asia), or as a mixture of two of these categories. Independently of this analysis, each of the donor self-declared ethnicities were also categorised in the same way. Results were as follows:

Table 61: 'Blinded' results of Ancestry analysis. Samples number and 'Self-declared donor ethnicity' correspond to entries in Table 57. 'Self-declared category' is the assignation of self-declared donor ethnicity to the seven ethnic categories in Section 5.3. 'Blind categorisation' is the independent assignation of the software prediction in Table 57 to the same seven categories. 'Blinded Result' indicates whether columns 3 and 4 match each other. 'Original Result' is the result from the 'unblinded' analysis in Table 56. Samples where the 'Blinded Result' and 'Original Result' differ are highlighted in blue (Table continues over next three pages).

No.	Self-declared donor ethnicity	Self-declared category	Blind categorisation	Blinded Result	Original Result
1	Afghan	South West Asia	European / South Asian	No Match	Unclear
2	Afghan	South West Asia	South West Asia	Match	Unclear
3	Afghan	South West Asia	South West Asia	Match	Unclear
4	Albanian	European	South West Asia	No Match	Match
5	Angolan	African	African	Match	Match
6	Asian / Black	South Asia / African	Admixed African	Match	Unclear
7	Bangladeshi	South Asia	South Asia	Match	Match
8	Black British	African	African	Match	Match
9	Bosnian	European	European	Match	Match
10	British Asian	South Asia	European / South West Asian	No Match	Match
11	Canadian Caribbean	African	Admixed African	Partial Match	Match
12	Chinese	East Asia	East Asia	Match	Match
13	Chinese	East Asia	East Asia	Match	Match

No.	Self-declared donor ethnicity	Self-declared category	Blind categorisation	Blinded Result	Original Result
14	Chinese	East Asia	East Asia	Match	Match
15	Chinese Malay / French	East Asia / European	South Asia	No Match	Match
16	Cornish / Burmese	East Asia / European	Admixed European	Partial Match	Unclear
17	Dutch / Vietnamese	East Asia / European	East Asia / European	Match	Match
18	Egyptian	South West Asia	South West Asia	Match	Match
19	Ethiopian	African	Admixed African	Partial Match	Match
20	Ethiopian	African	Admixed African	Partial Match	Match
21	French / British	European	European / South West Asian	Partial Match	Unclear
22	Ghanaian	African	African	Match	Match
23	Greek	European	Unclear	No Match	No Match
24	Greek Cypriot	European	European	Match	Match
25	Indian	South Asia	South Asia	Match	No Match
26	Iran / UK	South West Asia / European	European	Partial Match	Match
27	Iraqi	South West Asia	South West Asia	Match	Match
28	Iraqi	South West Asia	South West Asia	Match	Match
29	Irish	European	European	Match	Match
30	Irish	European	European	Match	Match
31	Irish	European	European	Match	Match
32	Irish / Black British	European / African	European / African	Match	Match

No.	Self-declared donor ethnicity	Self-declared category	Blind categorisation	Blinded Result	Original Result
33	Irish / Chinese	European / East Asia	European / East Asia	Match	Match
34	Irish / Thai	European / East Asia	European / East Asia	Match	Match
35	Italian / Filipino	European / East Asia	European / East Asia	Match	Match
36	Kenyan	African	African / South West Asian	Partial Match	Match
37	Kuwaiti	South West Asia	South West Asia	Match	Match
38	Mixed Race British	European / African	South West Asia	No Match	Match
39	Mongolian	East Asia	East Asia	Match	Match
40	Morocco / Korea	South West Asia / East Asia	South West Asia	Partial Match	Match
41	Nigerian	African	African	Match	Match
42	Nigerian	African	African	Match	Match
43	Nigerian	African	African	Match	Match
44	Pakistani	South Asia	South Asia	Match	Match
45	Pakistani	South Asia	South West Asia	No Match	Match
46	Polish	European	European	Match	Match
47	Somalian	African	African / South West Asian	Partial Match	Match
48	Somalian	African	African / South West Asian	Partial Match	Match
49	Somalian	African	African / South West Asian	Partial Match	Match
50	South Italian	European	European / South West Asian	Partial Match	No Match
51	Spanish	European	European	Match	Unclear

No.	Self-declared donor ethnicity	Self-declared category	Blind categorisation	Blinded Result	Original Result
52	Sri Lankan	South Asia	South Asia	Match	Match
53	Tanzanian	African	African	Match	Match
54	Ugandan	African	African	Match	Match
55	Ugandan	African	African	Match	Match
56	White / Black British	European / African	European / African	Match	Match
57	White British	European	European	Match	Match
58	White British	European	European	Match	Unclear
59	White British	European	European	Match	Match
60	White British	European	European	Match	Match
61	White British	European	European	Match	Match
62	White British	European	European	Match	Match
63	White British	European	European	Match	Match
64	White British	European	European	Match	Match

A comparison of the ‘unblinded’ and ‘blinded’ results in Table 61 was then made. Results were as follows:

Table 62: Comparison of ‘blinded’ and ‘unblinded’ results for Ancestry analysis. Total number of samples with a result of ‘Match’, ‘No Match’ and ‘Unclear / Partial Match’ result in Table 57 and Table 61 are shown.

Panel used in Analysis	No. of ‘Match’ results	No. of ‘Unclear’ or ‘Partial’ results	No. of ‘No match’ results
Combined panel	53	8	3
Blinded analysis	45	12	7

5.5. Principal Component Analysis

As a comparison to the results generated with the manufacturer’s software for the Precision ID Ancestry panel in Sections 5.3 and 5.4, a principal component analysis (PCA) for the same data set of 64 samples was performed. This was performed using PAST v3.21 (Hammer *et al*. 2001). Results were as follows (Figure 41).

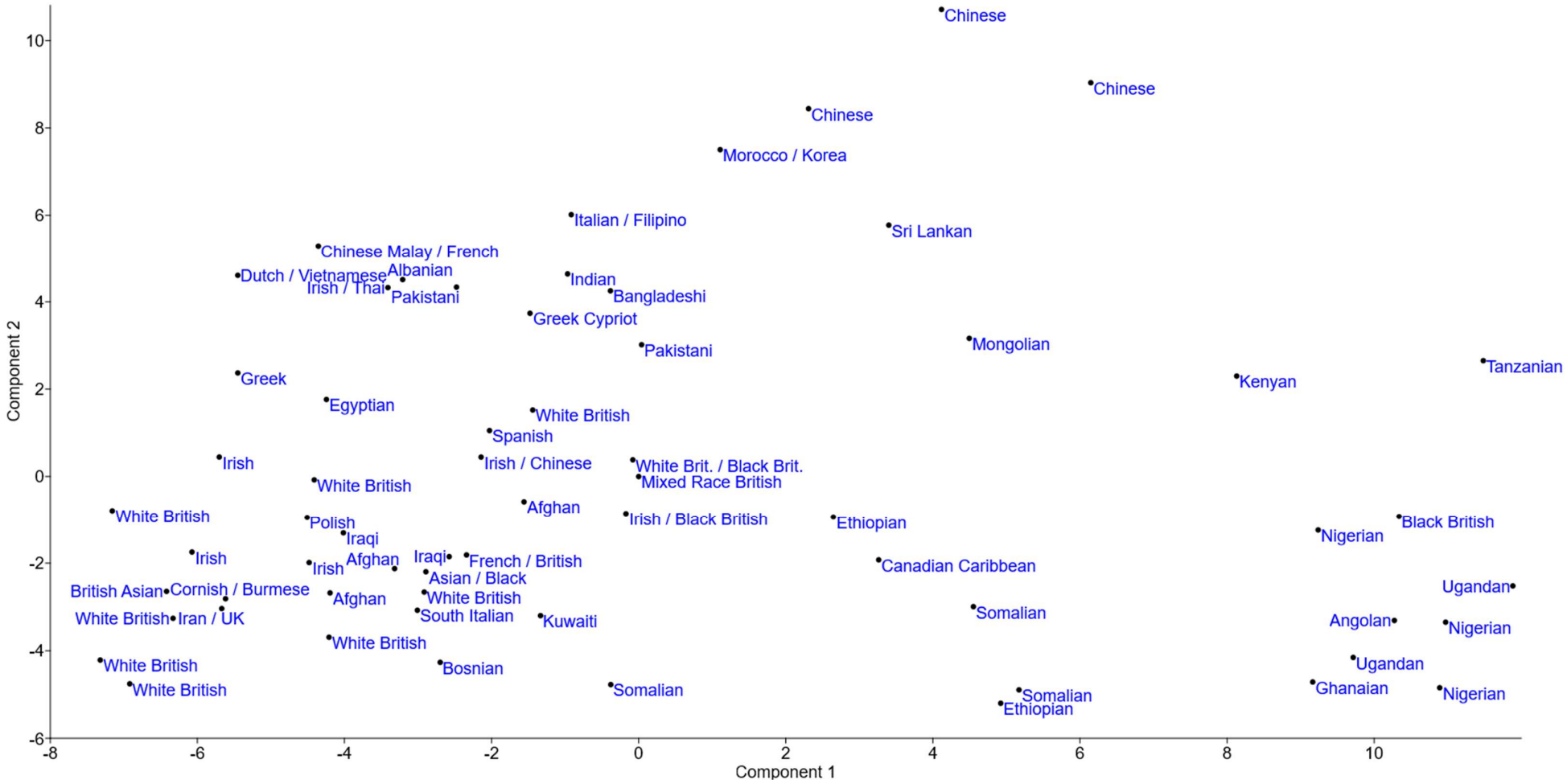


Figure 41: Principal Component Analysis (PCA) chart for the 64 ancestry samples. The sample donor's self-declared ancestry is shown at each point on the chart.

5.6. Discussion

The results of this study showed that the MPS assay tested, the Precision ID Ancestry panel, has a good ability to predict the self-declared ancestry of a range of sample donors. The results with the system using the built-in manufacturer recommended analysis of all loci in the panel, i.e. both the Kidd and Seldin loci together, referred to in the results as the 'Combined' panel, showed that 53 of the 64 donors could have their ancestry accurately predicted by this system (Table 56). Eight of the 64 were scored as 'Unclear' with this system, with three being 'No Match'. The three samples scored as 'No Match' had declared ancestry of 'Greek', 'Indian' and 'South Italian' (see samples 23, 25 and 50 in Table 56). No particular pattern is evident to the origins of these three samples, and it is difficult to conclude anything specific on results of three isolated samples, especially without any knowledge of the donor's family history except a single self-declared summary of their ancestry. The eight 'Unclear' results in the same analysis comprised three 'Afghan' samples (all the Afghan samples that were tested), one 'Asian / Black' sample, one 'Cornish / Burmese' sample, one 'French / British' sample, one 'Spanish' and one 'White British' sample (see samples 1, 2, 3, 6, 16, 21, 50, and 58 in Table 56). Again it is hard to conclude significant amounts from isolated samples, but if a pattern can be seen it is that the system struggled to identify donors of Afghan ancestry, and also that the majority of the rest of the 'Unclear' samples were donors of mixed heritage, where it could perhaps be expected that it is harder for the system to accurately infer ancestry. Despite this however, the system did give a 'Match' result for nine other samples of declared mixed ancestry in the same analysis, (see samples 15, 17, 26, 32, 33, 34, 35, 40, and 56 in Table 56), meaning that more often than not, the system could accurately detect the ancestry of mixed ancestry samples.

When the analysis was expanded to include the custom method of looking at the 'Kidd' and 'Seldin' panels in isolation, in addition to the 'Combined' approach already discussed, the results were not improved with either of the individual panels (Table 58). In the case of the 'Seldin' panel taken in isolation, the results on the same sample set were worse than with the 'Combined' panel, with only 49 instead of 53 samples resulting in 'Match', and nine and six 'Unclear' and 'No Match' results respectively, instead of eight and three with the 'Combined' panel. The 'Kidd' panel taken in isolation performed approximately the same as the 'Combined' panel, with the same number of profiles scored 'Match' – 53, and with seven 'Unclear' and four 'No Match' results, instead of eight and three respectively from the 'Combined' panel. Table 60 shows the results for the eight samples in the analysis that had an inconsistent result across the three panels, again with there being no particular pattern to

the source of these eight donors compared to the wider sample set, but again perhaps of note that four of the eight ‘inconsistent’ donors were of self-declared mixed ancestry.

The samples noted above in the ‘Combined’ analysis that were scored as ‘No Match’ were also all scored as ‘No Match’ by the individual systems (see samples 23, 25, and 50 in Table 57). It is of note also that the three Afghan samples, which were all scored ‘Unclear’ by the combined panel, were also all scored ‘Unclear’ in each analysis by the individual panels, indicating that the loci in the assay tested struggle to detect the ancestry from donors of this region and that investigation of additional SNP characteristic of this region may be needed to allow this prediction. On the other hand, in this work, donors that were self-declared as ‘Chinese’, ‘White British’, ‘Irish’, ‘Ugandan and ‘Nigerian’ were all accurately detected in multiple samples, indicating that the loci in the Precision ID Ancestry panel can accurately determine these ancestry types and may be of use to forensic investigators for this purpose.

It is of note however in this analysis, that the number of samples that were found to be matches of the declared donor ancestry to the software prediction was lowered when the analysis was blinded, dropping from 53 out 64 matches to 45 out of 64 (Table 62). This indicates that the bias of knowing the declared donor ancestry can play a part in evaluating the software prediction and is a better indicator of how this analysis would perform in a real forensic case, where, of course, the ethnicity of the offender is not known in advance. Despite this, the general conclusions reached above still apply, with the blinded analysis again correctly determining the ancestry of the multiple ‘Chinese’, ‘White British’, ‘Irish’, ‘Ugandan and ‘Nigerian’ samples. The blinded analysis was similarly less accurate with admixed samples, with the ‘Asian / Black’, ‘Cornish / Burmese’ and ‘French / British’ samples again not being accurately identified (Table 61).

Where the blinded analysis proved less accurate than the ‘unblinded’ analysis was in cases where the country of origin is generally dissimilar to other parts of the same continental area. For example, the blinded analysis did less well in identifying the six ‘Somalian’, ‘Ethiopian’ and ‘Kenyan’ in the analysis (Table 61). These are all countries in Africa, but all six donors with these ethnicities did not give an unambiguous ‘African’ result but tended to show an African result mixed with other, mostly Asian, ethnicities. This can be seen as a consistent result with prior knowledge of the donor ethnicity, but as has been shown, cannot be predicted when this is not known. This indicates that the so-called ‘African’ results in this analysis are more accurately ‘West African’, (as was shown by the clear results from the Ugandan and Nigerian samples), and do not predict the East African ethnicities as well.

This illustrates the importance of careful choice of allele frequency databases in analysis of this type, and highlights an area in which this analysis could potentially be improved with

larger databases with more representative frequencies. The databases used in this analysis were the default databases provided with the HID_SNP_Genotyper v5.2.2 software used here (see Section 2.5), which contain 'African' allele frequencies, but which are largely comprised of populations in West, rather than East, Africa. Expansion of the available databases for these SNP sets to include East African and other underrepresented groups could improve the results seen in this analysis in future.

The analysis in this study compares well to other groups who have researched this area, with 45 of the 64 samples tested matching the declared ancestry of the donor, or 70%. Although this number is less than many have come to expect in forensic DNA analysis, which when dealing with random match probabilities of STR profiles typically produces match statistics in the magnitude of billions, it is in line with figures reported by others who have attempted to use DNA as a tool for predicting sample donor ancestry, for example Ramani *et al.* (2017) who reported success rates of 80-94% although with a slightly different SNP group and a less diverse range of sample ethnicities than the ones used here.

Other groups have used the same Precision ID Ancestry chemistry as used in this study, but often on narrower ranges of ancestry as was examine in this study. Pereira *et al.* (2017) used the Precision ID Ancestry chemistry and had found that it was able to correctly categorise the samples that they tested as being either Somali or Danish. A third group of samples tested by Pereira *et al*, who had ancestry in Greenland, were not correctly categorised, the authors speculate, due to the lack of Greenlander data in the allele frequencies used. Despite using the same chemistry as used here, the Precision ID Ancestry panel, the authors of the Pereira *et al.* study used different analysis methods to infer ancestry from the data including STRUCTURE, a method of clustering genotypes into distinct populations based on shared locus genotypes (Pritchard *et al.* 2000), and principal component analysis (PCA), a mathematical method for grouping observations based on shared correlated variables (Jolliffe and Cadima, 2016). Despite this, the overall result for the different analysis methods used were broadly the same for the samples analysed.

STRUCTURE and PCA, as methods of analysis in this field, were also used by Anantharaman *et al*. (2020) to attempt to differentiate 484 individuals of Asian ancestry. In this work the authors found that STRUCTURE and PCA worked equally well to distinguish five population groups within the set of 484 samples. Karamizadeh *et al.* (2013) evaluated the PCA method itself, and conclude that while PCA has the advantage of reducing complexity in large data sets and low sensitivity to noise, it can be difficult to evaluate and requires comprehensive training data sets to allow data to be correctly interpreted, a conclusion that could also be applied to STRUCTURE. In this work PCA was performed as a

comparison to the results gained from the manufacturer's recommended form of analysis with the HID_SNP_Genotyper_v5.2.2. This is shown in Figure 41. Here it can be seen that while some clustering of similar samples can be seen, for example the West African samples in the lower right, the White British and Irish samples in the lower left, and the Chinese samples in the top middle, it would be difficult, if not impossible, to accurately predict the ancestry of an unknown sample based solely on its placement on this chart. In this sense, this is an illustration of the limitation noted above by Karamizadeh *et al.* regarding training data – with 64 samples of diverse origin in this analysis, at most eight samples in the set being from any one ethnicity, and in most cases only a single example of each ancestry, the training data needed to define distinct clusters is not present.

Wang, He *et al.* (2018) also used the Precision ID Ancestry panel to analyse samples with ancestry from Tibetan-Burmese minority populations in China, and had success in categorising nine of the sixteen samples into the correct region, with the remaining seven samples being placed in neighbouring regions. Samples from wider populations were not tested however.

Nakanishi *et al.* (2018) reported the ability of the Precision ID Ancestry panel to distinguish different Japanese sub-populations, specifically 'mainland' Japanese and Okinawa Japanese. They concluded that the panel was not able to differentiate these two populations of 'East Asians', something that matches the results seen here in that only broad continental ancestry categories were studied, not narrower sub-populations. The author of the Nakanishi *et al.* paper report that distinction of the mainland and Okinawa sub-populations is possible however, based on analysis of a much larger set of 140,387 SNPs by another group (Yamaguchi-Kabata *et al.* 2008). As such they state that an MPS panel that can make this distinction should be possible, just with a wider range of SNPs than is available in the Precision ID Ancestry panel.

Simayijiang and colleagues (Simayijiang *et al.* 2019) used the Precision ID Ancestry panel to look at Uyghur and Kazakh populations in China. In a similar way to the Nakanishi *et al.* paper just mentioned, the panel was not able to distinguish these populations with the allele frequencies provided by the manufacturer, but by adding a custom set of allele frequencies to the analysis, they were able to correctly distinguish 42 out of 49 test individuals.

Another study with the Precision ID Ancestry panel used a broader range of sample ancestries than the previously mentioned studies, and as in the present work, analysed a set of samples with self-declared ancestries from around the world (Al-Asfi *et al.* 2018). These samples were comprised of 36 samples from single populations, and 14 from multiple populations (i.e. they were from donors of mixed ancestry). The authors did not score the

results of any of these samples as being a ‘Match’ or ‘No match’ as in this work, but broadly achieved similar results, with their conclusion being that the panel could accurately determine the continent of ancestry for East Asian, African, European and South Asian individuals, but also that the panel was less accurate for individuals of mixed ancestry. Similar results to this were also seen in a publication by Jin *et al.* (2018), who also studied a large collection of samples with global self-declared ancestry. Jin *et al.* also reported success in distinguishing samples with one ancestral population. 644 of the samples analysed were scored as ‘concordant’, with one ‘not concordant’ and four ‘uninformative’, giving a 99.2% success rate. For samples of declared mixed ancestry however, the results were less good with 15 ‘concordant’, nine ‘not concordant,’ and nine ‘uninformative’, for a 45.5% success rate.

Lastly, a recent study by Morgensen *et al.* (2020) also studied a large set of profiles with the Precision ID Ancestry panel loci, this time 3606 profiles obtained from reference population data sets. They compared two methods of analysing these profiles, the first a similar method to that studied here, which resulted in 78.1% of profiles being correctly assigned ancestries. The second analytical method they studied, freeware software called GenoGrapher (Tvedebrink *et al.* 2018), resulted in improved performance, with 83.6% correct assignment of ancestry, both results comparable to the 82% found in this study.

As a result, it has been found in this work that the Precision ID Ancestry panel is capable of detecting the ancestry of unknown forensic samples with an approximate 70% success rate. As also found by others who have investigated this area, this may be improved in future with improved allele frequencies for the analysis, from a wider set of populations; alternative analysis methods for inferring the ancestry of the samples; or a larger SNP panel that includes more ancestry informative markers. In the meantime however, as it stands, Precision ID Ancestry panel is a useful tool in predicting ancestry and as an intelligence tool in guiding investigation of a case, rather than as evidence that is used in isolation to convict a suspect, may find a place in the techniques used by practicing forensic DNA laboratories.

Chapter 6:

Evaluation of Massively Parallel Sequencing for Kinship Analysis

6. Evaluation of Massively Parallel Sequencing for Kinship Analysis

6.1. Introduction

The next chapter of this work explored the ability of MPS to provide additional useful information to the forensic analyst in examination of kinship cases, that is, cases where there are questions of if and how sample donors are related to each other. As discussed in Section 1.1.4.6, MPS promises to improve this type of analysis by allowing more markers to be analysed from a single sample than was previously possible. These extra markers could potentially allow samples from relatives to be determined as such with a much greater degree of statistical certainty than has been previously possible, due to the genotypic information provided by the extra markers. That said, however, for many types of kinship cases, current CE-based methods can already give a result in a kinship case of 99.9999% or more certainty that the samples in question have the claimed relationship. This statistic is known as the probability of relatedness. As such, it is unclear how much extra benefit MPS would add to these cases, especially given than MPS analysis involves a greater outlay of time and money than CE-based analysis. On the other hand, some kinship cases can achieve much less clear results with current CE-based methods, and a much lower probability of relatedness than that stated above. It is these cases, often involving more distantly related individuals than the first category, that could benefit from MPS analysis.

It is the aim of this work to examine multiple kinship case scenarios, and having performed CE-based and MPS-based analysis on the samples in each case, determine whether the case would benefit from use of CE-based analysis, MPS-based analysis, or a combination of the two.

The cases examined were derived from donors of known relationship, and involved lab-based analysis of the samples with the CE-based GlobalFiler PCR amplification kit and the MPS-based Precision ID Identity panel (both Thermo Fisher Scientific, USA). The statistical strength of these results in confirming the known relationships in each case were then compared as a means of determining which analysis method, CE or MPS, was superior in each case.

6.2. Genotyping of Kinship samples

To investigate the performance of MPS in kinship analysis, nine samples were processed that were collected from a single extended family. The relationships between the donors of the samples are shown in the following diagram.

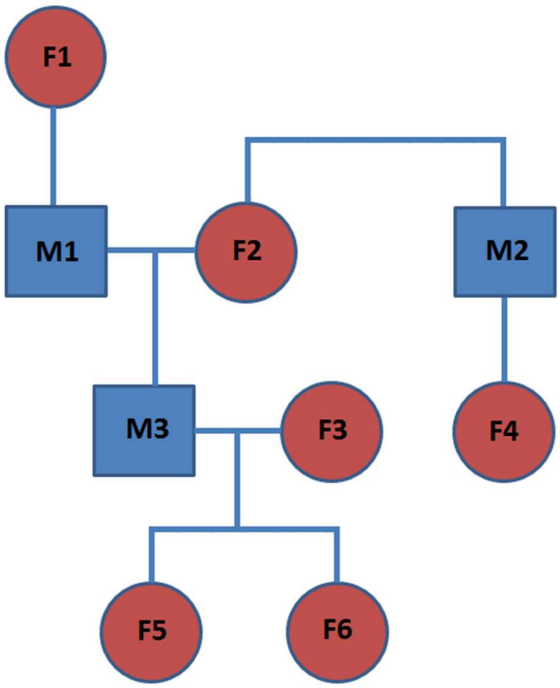


Figure 42: Pedigree diagram of the nine samples analysed for kinship. Blue squares represent males and red circles represent females. Horizontal lines directly connecting two shapes represent a male and female who are the parents of the offspring indicated by the vertical lines running downwards from the couple in question. F2 and M2 are siblings, with the parents of this pair not shown as not included in the testing.

Samples were processed with the CE-based GlobalFiler PCR amplification kit and with the MPS-based Precision ID Identity panel as described in Section 2.3 and Section 2.4.1.4 respectively. Parameters of the HID_SNP_Genotyper v5.2.2 for secondary analysis (i.e. genotyping) were as follows (all manufacturer's default values):

Table 63: Parameters used in HID_SNP_Genotyper v5.2.2 for Precision ID Identity Panel kinship analysis. All are the manufacturer's recommended default parameters.

Parameter	Value used
Minimum allele frequency	0.1
Minimum coverage	6
Minimum coverage either strand	0
Maximum strand bias	1
Trim reads	true

All samples gave full, clear, single source genotypes with both the CE and MPS based assays. All markers in all genotypes showed alleles as per the expected inheritance shown in Figure 42. Mutations were not observed at any markers.

6.3. Kinship analysis

The profiles obtained in the previous section were analysed in multiple different kinship scenarios to explore the utility of the MPS assay in comparison to the conventional CE-based assay. Familias v3.2.3 was the software package used for the analysis, as described in the Section 2.5. These kinship scenarios analysed were as described in the following tables:

Table 64: The first 22 relationship types tested in kinship analysis. All relationships tested with the stated (true) relationship as the test hypothesis, with the alternative hypothesis being that the donors were unrelated. Sample names are consistent with those used in Figure 42. Sample names starting with 'F' are females, those starting 'M' are males.

No.	Relationship type tested	Samples used
1	Siblings	F5, F6
2	Siblings	M2, F2
3	Duo Paternity	M1, M3
4	Duo Paternity	M2, F4
5	Duo Paternity	M3, F5
6	Duo Paternity	M3, F6
7	Duo Paternity	F1, M1
8	Trio Paternity	M3, F3, F5
9	Trio Paternity	M3, F3, F6
10	Trio Paternity	M1, F2, M3
11	Duo Grandparent	M3, F1
12	Duo Grandparent	M1, F5
13	Duo Grandparent	M1, F6
14	Duo Grandparent	F2, F5
15	Duo Grandparent	F2, F6
16	Trio Grandparent	M1, F2, F5
17	Trio Grandparent	M1, F2, F6
18	Great-grandparent	F1, F5
19	Great-grandparent	F1, F6
20	Cousin	M3, F4
21	Uncle / Nephew	M2, M3
22	Aunt / Niece	F2, F4

Analysis of the cases was performed by comparison of two hypotheses. For the twenty-two cases listed above, the first hypothesis was that the relationship was the one proposed in the table above (and in the case of this work, known to be the true relationship). The second hypothesis was that the two individuals were unrelated.

A further thirteen cases were constructed from the same group of samples as follows, making thirty-five test cases in total.

Table 65: The second 13 relationship types tested in kinship analysis. Sample names are consistent with those used in Figure 42. Sample names starting with 'F' are females, those starting 'M' are males.

No.	True relationship	Alternative hypothesis	Samples used
23	Mother / Son	Aunt / Nephew	F1, M2
24	Mother / Son	Aunt / Nephew	F2, M3
25	Father / Son	Uncle / Nephew	M1, M3
26	Mother / Daughter	Aunt / Niece	F3, F5
27	Mother / Daughter	Aunt / Niece	F3, F6
28	Father / Daughter	Uncle / Niece	M3, F5
29	Father / Daughter	Uncle / Niece	M3, F6
30	Father / Daughter	Uncle / Niece	M2, F4
31	Aunt / Niece	Mother / Daughter	F2, F4
32	Uncle / Nephew	Father / Son	M2, M3
33	Siblings	Cousins	F2, M2
34	Siblings	Cousins	F5, F6
35	Cousins	Siblings	M3, F4

For this second group of cases, the test hypothesis was the true relationship shown in the table. The alternative hypothesis is also shown in the table and for the cases where the test hypothesis was a parent/child pair, the alternative hypothesis was that the two samples were an avuncular or materteral pair, and vice versa. This is a common practical kinship case that forensic labs encounter, where it is disputed, often for immigration purposes, whether an adult and child pair are parent and child or aunt/uncle and niece/nephew. For cases where the test hypothesis was that the two sample donors were siblings, the alternative hypothesis is that they were cousins, and vice versa. Again, this is a common case type that forensic labs encounter.

The results for all thirty-five cases shown in Table 64 and Table 65 are presented below as the ratio of the probability of these hypotheses, known as the likelihood ratio or relationship index (and in the case of paternity analysis, commonly called the paternity index). Results are also shown as a probability of relatedness, expressed as a percentage. This is calculated from the relationship index with prior odds of 0.5 for each hypothesis.

Results of this analysis were calculated for both the MPS and CE profiles individually, and then also for the MPS and CE profiles combined. Result are shown in the following table:

Table 66: Results of kinship analysis. Relationships tested are as described in Figure 42, Table 64 and Table 65. For each relationship tested, two figures are shown for each of the CE, MPS, and CE + MPS result. The top figure is the relationship index (rounded to nearest integer, unless <5 when one decimal place is shown). The bottom figure is the probability of relatedness. This is shown to two decimal places after the last repeated '9'. Results are capped at 99.999999% and any result greater than this shown as >99.999999% (Table continues on the next two pages).

No.	Test (true) Hypothesis	Alternative Hypothesis	Result with CE	Result with MPS	Result with CE + MPS
1	Sibling	Unrelated	38,869	7,154,819	2.781×10^{11}
			99.9974%	99.999986%	>99.999999%
2	Sibling	Unrelated	4,203,854	2,799,805	1.177×10^{13}
			99.999976%	99.999964%	>99.999999%
3	Duo Paternity	Unrelated	14,398,675	19,304	2.859×10^{11}
			99.999993%	99.9948%	>99.999999%
4	Duo Paternity	Unrelated	240,747,801	8,758,770	2.109×10^{15}
			>99.999999%	99.999989%	>99.999999%
5	Duo Paternity	Unrelated	201,998	761,006	1.569×10^{11}
			99.99950%	99.99986%	>99.999999%
6	Duo Paternity	Unrelated	1,339,078	127,560,369	1.755×10^{14}
			99.999925%	>99.999999%	>99.999999%
7	Duo Paternity	Unrelated	32,251,201	365,320	1.178×10^{13}
			99.999997%	99.99973%	>99.999999%
8	Trio Paternity	Unrelated	2,558,072,896	22,555,094	7.451×10^{16}
			>99.999999%	99.999996%	>99.999999%
9	Trio Paternity	Unrelated	3.471×10^{10}	1.237×10^{10}	4.293×10^{20}
			>99.999999%	>99.999999%	>99.999999%
10	Trio Paternity	Unrelated	5,886,718,690	51,756,249	3.047×10^{17}
			>99.999999%	99.999998%	>99.999999%
11	Duo Grandparent	Unrelated	47	7	332
			97.95%	87.42%	99.70%
12	Duo Grandparent	Unrelated	3.8	0.2	0.9
			79.25%	19.28%	47.72%

No.	Test (true) Hypothesis	Alternative Hypothesis	Result with CE	Result with MPS	Result with CE + MPS
13	Duo Grandparent	Unrelated	12	3.0	37
			92.42%	75.05%	97.34%
14	Duo Grandparent	Unrelated	29	145	4249
			96.69%	99.32%	99.98%
15	Duo Grandparent	Unrelated	6	47	287
			85.98%	97.72%	99.65%
16	Trio Grandparent	Unrelated	1186	101	120,233
			99.915%	99.02%	99.99916%
17	Trio Grandparent	Unrelated	1091	335	365,922
			99.908%	99.70%	99.99973%
18	Great Grandparent	Unrelated	1.1	0.9	0.9
			51.37%	46.88%	48.25%
19	Great Grandparent	Unrelated	2.5	1.8	4.4
			71.22%	63.96%	81.45%
20	Cousin	Unrelated	0.6	5.0	3.0
			35.52%	83.31%	73.33%
21	Uncle / Nephew	Unrelated	3.0	16	57
			77.13%	94.37%	98.26%
22	Aunt / Niece	Unrelated	68	709	48,093
			98.55%	99.86%	99.9979%
23	Mother / Son	Aunt / Nephew	824	373	307,497
			99.87%	99.73%	99.99967%
24	Mother / Son	Aunt / Nephew	596	1093	639,654
			99.83%	99.909%	99.99984%
25	Father / Son	Uncle / Nephew	452	47	27,027
			99.78%	97.92%	99.9963%
26	Mother / Daughter	Aunt / Niece	365	321	116,588
			99.73%	99.69%	99.99914%
27	Mother / Daughter	Aunt / Niece	446	2217	982,515
			99.78%	99.955%	99.99989%

No.	Test (true) Hypothesis	Alternative Hypothesis	Result with CE	Result with MPS	Result with CE + MPS
28	Father / Daughter	Uncle / Niece	114	258	29,273
			99.13%	99.61%	99.9966%
29	Father / Daughter	Uncle / Niece	257	2261	577,369
			99.61%	99.956%	99.99983%
30	Father / Daughter	Uncle / Niece	356	854	301,827
			99.72%	99.88%	99.99967%
31	Aunt / Niece	Mother / Daughter	7,255,823	2.416×10^{11}	1.786×10^{18}
			99.999986%	>99.999999%	>99.999999%
32	Uncle / Nephew	Father / Son	1.807×10^{15}	1.391×10^{17}	2.544×10^{32}
			>99.999999%	>99.999999%	>99.999999%
33	Siblings	Cousins	6382	29,759	188,265,607
			99.984%	99.9966%	>99.999999%
34	Siblings	Cousins	70	29,753	1,775,859
			98.60%	99.9966%	99.999944%
35	Cousins	Siblings	2948	2760	9,748,063
			99.966%	99.964%	99.999990%

It is important to consider the statistical independence of the loci used in kinship analysis. As such, the chromosomal distance of each of the loci used, both STR and SNP was examined. This is shown in the following table:

Table 67: The list of loci used in kinship analysis, ordered by position on chromosome, with the distance between loci shown. Position on chromosome is as per the hg19 human genome. Loci with distance from the neighbouring locus under 5 million bp are highlighted. Where the highlighted distance is between two SNP loci, this is in red. Where the highlighted distance is between a SNP and STR locus, this is in blue (Table continues on next three pages.)

Locus Name	Locus Type	Chromosome	Position on chromosome (bp)	Distance from previous locus (bp)
rs1490413	NGS SNP	1	4,367,323	-
rs7520386	NGS SNP	1	14,155,402	9,788,079
rs4847034	NGS SNP	1	105,717,631	91,562,229
rs560681	NGS SNP	1	160,786,670	55,069,039
D1S1656	CE STR	1	230,905,362	70,118,692
rs10495407	NGS SNP	1	238,439,308	7,533,946
rs891700	NGS SNP	1	239,881,926	1,442,618
rs1413212	NGS SNP	1	242,806,797	2,924,871
rs876724	NGS SNP	2	114,974	-
TPOX	CE STR	2	1,493,425	1,378,451
rs1109037	NGS SNP	2	10,085,722	8,592,297
D2S441	CE STR	2	68,239,079	58,153,357
rs993934	NGS SNP	2	124,109,213	55,870,134
rs12997453	NGS SNP	2	182,413,259	58,304,046
D2S1338	CE STR	2	218,879,582	36,466,323
rs907100	NGS SNP	2	239,563,579	20,683,997
rs1357617	NGS SNP	3	961,782	-
rs4364205	NGS SNP	3	32,417,644	31,455,862
D3S1358	CE STR	3	45,582,231	13,164,587
rs1872575	NGS SNP	3	113,804,979	68,222,748
rs1355366	NGS SNP	3	190,806,108	77,001,129
rs6444724	NGS SNP	3	193,207,380	2,401,272
rs2046361	NGS SNP	4	10,969,059	-
FGA	CE STR	4	155,508,888	144,539,829
rs6811238	NGS SNP	4	169,663,615	14,154,727
rs1979255	NGS SNP	4	190,318,080	20,654,465
rs717302	NGS SNP	5	2,879,395	-
rs159606	NGS SNP	5	17,374,898	14,495,503

Locus Name	Locus Type	Chromosome	Position on chromosome (bp)	Distance from previous locus (bp)
D5S818	CE STR	5	123,111,250	105,736,352
CSF1PO	CE STR	5	149,455,887	26,344,637
rs7704770	NGS SNP	5	159,487,953	10,032,066
rs251934	NGS SNP	5	174,778,678	15,290,725
rs338882	NGS SNP	5	178,690,725	3,912,047
rs13218440	NGS SNP	6	12,059,954	-
SE33	CE STR	6	88,986,988	76,927,034
rs214955	NGS SNP	6	152,697,706	63,710,718
rs727811	NGS SNP	6	165,045,334	12,347,628
rs6955448	NGS SNP	7	4,310,365	-
rs917118	NGS SNP	7	4,457,003	146,638
D7S820	CE STR	7	83,789,542	79,332,539
rs321198	NGS SNP	7	137,029,838	53,240,296
rs737681	NGS SNP	7	155,990,813	18,960,975
rs10092491	NGS SNP	8	28,411,072	-
D8S1179	CE STR	8	125,907,107	97,496,035
rs4288409	NGS SNP	8	136,839,229	10,932,122
rs2056277	NGS SNP	8	139,399,116	2,559,887
rs1015250	NGS SNP	9	1,823,774	-
rs7041158	NGS SNP	9	27,985,938	26,162,164
rs1463729	NGS SNP	9	126,881,448	98,895,510
rs1360288	NGS SNP	9	128,968,063	2,086,615
rs10776839	NGS SNP	9	137,417,308	8,449,245
rs826472	NGS SNP	10	2,406,631	-
rs735155	NGS SNP	10	3,374,178	967,547
rs3780962	NGS SNP	10	17,193,346	13,819,168
rs740598	NGS SNP	10	118,506,899	101,313,553
D10S1248	CE STR	10	131,092,508	12,585,609
rs964681	NGS SNP	10	132,698,419	1,605,911
TH01	CE STR	11	2,192,319	-
rs1498553	NGS SNP	11	5,709,028	3,516,709

Locus Name	Locus Type	Chromosome	Position on chromosome (bp)	Distance from previous locus (bp)
rs901398	NGS SNP	11	11,096,221	5,387,193
rs10488710	NGS SNP	11	115,207,176	104,110,955
rs2076848	NGS SNP	11	134,667,546	19,460,370
vWA	CE STR	12	6,093,143	-
rs2269355	NGS SNP	12	6,945,914	852,771
D12S391	CE STR	12	12,449,954	5,504,040
rs2111980	NGS SNP	12	106,328,254	93,878,300
rs10773760	NGS SNP	12	130,761,696	24,433,442
rs1335873	NGS SNP	13	20,901,724	-
rs1886510	NGS SNP	13	22,374,700	1,472,976
D13S317	CE STR	13	82,722,160	60,347,460
rs1058083	NGS SNP	13	100,038,233	17,316,073
rs354439	NGS SNP	13	106,938,411	6,900,178
rs1454361	NGS SNP	14	25,850,832	-
rs722290	NGS SNP	14	53,216,723	27,365,891
rs873196	NGS SNP	14	98,845,531	45,628,808
rs4530059	NGS SNP	14	104,769,149	5,923,618
rs2016276	NGS SNP	15	24,571,796	-
rs1821380	NGS SNP	15	39,313,402	14,741,606
rs1528460	NGS SNP	15	55,210,705	15,897,303
rs729172	NGS SNP	16	5,606,197	-
rs2342747	NGS SNP	16	5,868,700	262,503
rs430046	NGS SNP	16	78,017,051	72,148,351
rs1382387	NGS SNP	16	80,106,361	2,089,310
D16S539	CE STR	16	86,386,308	6,279,947
rs9905977	NGS SNP	17	2,919,393	-
rs740910	NGS SNP	17	5,706,623	2,787,230
rs938283	NGS SNP	17	77,468,498	71,761,875
rs2292972	NGS SNP	17	80,765,788	3,297,290
rs1493232	NGS SNP	18	1,127,986	-
rs9951171	NGS SNP	18	9,749,879	8,621,893

Locus Name	Locus Type	Chromosome	Position on chromosome (bp)	Distance from previous locus (bp)
rs1736442	NGS SNP	18	55,225,777	45,475,898
D18S51	CE STR	18	60,948,900	5,723,123
rs1024116	NGS SNP	18	75,432,386	14,483,486
rs719366	NGS SNP	19	28,463,337	-
D19S433	CE STR	19	30,417,142	1,953,805
rs576261	NGS SNP	19	39,559,807	9,142,665
rs1031825	NGS SNP	20	4,447,483	-
rs445251	NGS SNP	20	15,124,933	10,677,450
rs1005533	NGS SNP	20	39,487,110	24,362,177
rs1523537	NGS SNP	20	51,296,162	11,809,052
rs722098	NGS SNP	21	16,685,598	-
D21S11	CE STR	21	20,554,291	3,868,693
rs2830795	NGS SNP	21	28,608,163	8,053,872
rs2831700	NGS SNP	21	29,679,687	1,071,524
rs914165	NGS SNP	21	42,415,929	12,736,242
rs221956	NGS SNP	21	43,606,997	1,191,068
rs733164	NGS SNP	22	27,816,784	-
rs987640	NGS SNP	22	33,559,508	5,742,724
D22S1045	CE STR	22	37,536,327	3,976,819
rs2040411	NGS SNP	22	47,836,412	10,300,085
rs1028528	NGS SNP	22	48,362,290	525,878

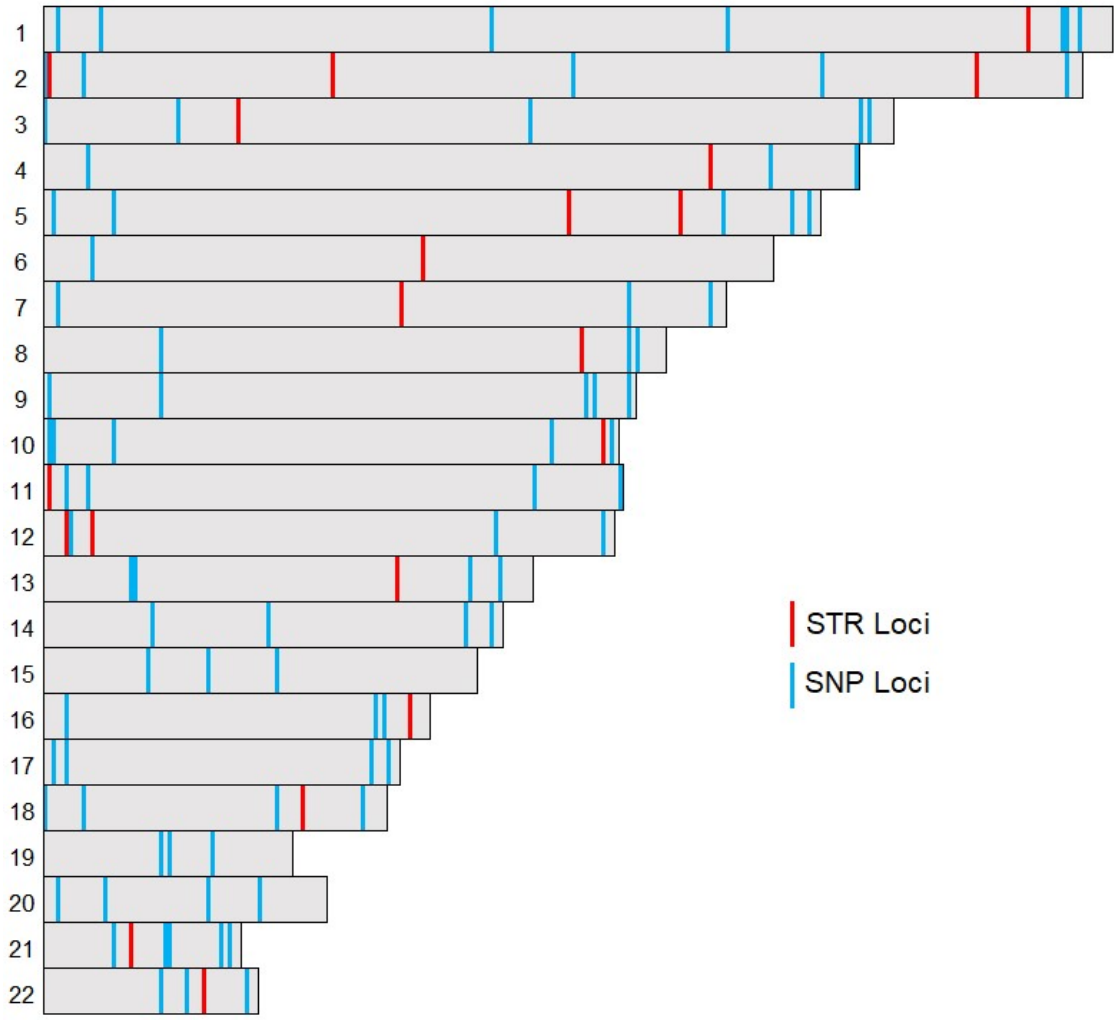


Figure 43: A visual representation of the spacing across the genome of the STR and SNP loci used in kinship analysis. Numbers on the left of the figure represent the 22 autosomal chromosomes. Red stripes are the position of STR loci. Blue stripes are the position of SNP loci.

6.4. Discussion

The results of this study show some value in the addition of MPS methods to the capillary based methods routinely adopted today by forensic laboratories. This is not the case in 'simple' paternity cases however. As can be seen in Table 66, in rows three to ten, all paternity scenarios tested showed a 99.999% or higher result with only CE profiling, making the addition of extra discrimination power with MPS unnecessary. This was especially the case in the Trio paternity cases, where the genotype of the mother is known, which then gives more information on which alleles in the child genotype must have been inherited from the true father. For all Trio paternity cases tested here, CE analysis gave a >99.999999% probability of relatedness, rendering any extra discrimination from MPS entirely

unnecessary. In paternity cases such as these, it could be expected that forensic laboratories will remain with CE-based methods for the foreseeable future due to the extra expense and time needed for MPS analysis.

For other case types however, MPS showed that it can be of value in kinship analysis. For the two sibling relationship cases that were tested, although CE gave a probability of relatedness of 99.99976% for one case (Table 66, row two), for the other (Table 66, row one) the CE-based probability of relatedness was only 99.9974%, or a relatedness index of 38,869. Although still strong evidence in favour of the proposed relationship, this figure is not as high as might be desired in a practical case. This lower figure for one of the sibling cases is due just to chance, specifically due to the number of alleles that the two siblings share. True siblings do not necessarily share any alleles with each other, but in practice will typically share many more alleles than two unrelated individuals (Butler, 2015). The exact number that are shared however, can impact the relatedness index observed in a specific case, as was seen in this work with the slightly lowered statistic seen in one of the sibling cases. That said, it is of note that for both sibling cases, the addition of MPS boosted the probability of relatedness to over 99.999999% and the relationship index to 2.781×10^{11} in one case and 1.177×10^{13} in the other. For the case that started with a relationship index of 38,869 from CE alone, this represents a significant increase that would be of interest in the reporting of a practical case, and shows a practical benefit in the addition of the MPS method.

For the more distantly related scenarios that were tested in this work, MPS methods showed some benefit. In the grandparent cases, especially in the ‘Trio grandparent’ cases, MPS was of benefit to the analysis and increased a result that would likely not be acceptable with CE alone. Trio grandparent cases are those where the genotype of one grandparent is known and it is the genotype of the other alleged grandparent, the partner of the donor of known genotype, that is tested against the child. In this work, two such cases were tested (Table 66, rows 16 and 17). In both cases CE alone gave a relationship index of approximately one thousand (1186 and 1091 respectively), which resulted in probabilities of relatedness of 99.915% and 99.908%. These statistics are on the border of what would be considered acceptable in a practical case. In both cases, the addition of the MPS data increased both of these statistics to over 99.999%, with relationship indices of 120,233 and 365,922 respectively. This demonstrates a clear benefit in the addition of the MPS method, with it boosting both cases from borderline inconclusive to definitely useable.

The grandparent cases where the second grandparent genotype was not known, where the test is simply a determination of whether the alleged grandparent in isolation could be the

grandparent of the child, known as ‘Duo grandparent’ cases, were less conclusive. In these five cases, the strength of the statistics gained were all much weaker than for the Trio grandparent cases (Table 66, rows 11 to 15). The statistics for the CE analysis here gave relationship indices ranging from 3.8 to 47, which translate to probabilities of relatedness of 79.25% to 97.95%. In all five cases, these would not be reportable as conclusive evidence of the relationship in a practical case. In all but one of the cases however, the addition of the MPS data to the CE data increased the strength of the statistic. In three of the five cases, this increase potentially resulted in a strengthening of the statistic that would make the case reportable in a practical scenario. These are rows 11, 14 and 15 in Table 66. In row 14 in particular, the probability of relatedness from CE was 96.69%, which was increased to 99.98% with the addition of MPS. Row 11 saw an increase from 97.95% to 99.70%, and row 15 an increase from 85.98% to 99.65%. In all three of these cases, a result that was definitely not reportable was increased to one on the border of reportable with the addition of the MPS data. As to whether the exact statistic achieved would be useable in a given case would depend of the rules of the laboratory and jurisdiction in question, but either way, in this work, the addition of MPS data to these three Duo grandparent cases definitely improved the statistic gained from the case.

In one Duo grandparent case however, the MPS data actually weakened the statistic (Table 66, row 12), with the combined result of the CE and MPS giving a probability of relatedness of 47.72%. This result is due to chance, similar to the sibling case noted above, and is due to the number of alleles that the grandparent and grandchild share – as with siblings, grandparents do not necessarily share any alleles with their grandchildren, it just that on average, they can be expected to share a greater number than two unrelated individuals. The exact number of shared alleles in any given case however can affect the statistic that is observed. This is particularly prone to happen with the MPS assay used here, the Precision ID Identity panel, which, as described in Section 2.4.1.4, contains exclusively bi-allelic SNP markers. This means that there are only two possible alleles and three possible genotypes at each marker in this assay. When the relationship index for a case is calculated, depending on the specific genotypes observed, the relationship index of a locus is often equal to 0.5 divided by the frequency of the allele that is shared between the two genotypes being tested. For a bi-allelic locus there are only two possible alleles at the locus, which means that one of these alleles will have a frequency greater than or equal to 0.5 and the other less than or equal to 0.5. If by chance the frequency of the shared allele in a case is greater than 0.5, this means that the relationship index for the locus is 0.5 / >0.5, which gives a result less than one. The relationship indices calculated for each locus in the assay are then multiplied together to give the final relationship index for the assay as a whole. If the relationship index

for a given locus is less than one however, this will have the effect of lowering the overall relationship index, even though the result of that locus represents a true relationship with a shared allele. As such, although this work has shown that the Precision ID Identity panel is helpful in many relationship testing scenarios, it is likely that an improved kit that features SNP markers with three or four alleles per locus, could be even more effective in this type of analysis. This is a point in favour of the use of microhaplotypes for kinship analysis, which, as discussed in Section 1.1.4.4, are loci consisting of multiple SNPs that are so close together in the genome that they must be analysed together as a small haplotype (hence the name, 'micro' haplotype) and depending on the number of SNPs in the microhaplotype, these loci will have multiple possible different alleles. This point was demonstrated by Sun *et al.* (2020) who showed that a panel of 30 microhaplotype markers could be used effectively to distinguish avuncular and grandparent-grandchild cases.

For the most distantly related cases tested in this work, the cousin, great-grandparent avuncular and materteral tests, the results were similar to those in the Duo grandparent tests. In one case, that of the materteral test (Table 66, row 22), the addition of MPS analysis raised the CE statistic of 68 for the relationship index and 98.55% for the probability of relatedness to 48,093 and 99.9979% respectively, converting a result that would not be reportable to one that likely would be. In the cousin, great-grandparent, and avuncular tests however (Table 66, rows 18 to 21), although the addition of the MPS assay improved the statistics seen, the result was still too low to be reported however, with a probability of relatedness of 73.33% and 98.26% for the cousin and avuncular tests respectively.

In the cases examined where the alternative hypothesis was not that the sample donors were unrelated, rather that they were an aunt / niece pair rather than parent and child, for example, (Table 66, rows 23 to 35), many of these cases showed value in adding MPS to the CE data for the case. For cases 23 to 30, where the cases in question tested the true hypothesis of a parent/child pair against an aunt/uncle relationship, all cases showed a CE STR result of around 99% exactly, again on the borderline of what would be acceptable in a practical case. For all of these cases the addition of MPS to these statistics increased the result to 99.99% or more, showing a definite practical advantage in the addition of this data.

For the reverse cases, where the question was again whether the donors concerned were parent / child or aunt / uncle and niece / nephew, but this time the true relationship was the aunt / uncle rather than the parent relationship (Table 66, rows 31 to 32), the addition of the MPS data was less necessary as in these cases the CE STR data alone gave a result of 99.9999% or greater. This is likely due to the fact that those in an uncle/nephew relationship will typically share fewer alleles than parent and child, whereas there is obligate sharing

between parent and child in the absence of a de novo mutation. Although of course in a practical case it would not be known ahead of time which was the true relationship in the case, it could be possible to pursue a strategy of performing CE STR analysis first, and only adding MPS analysis if the resulting statistic was not as strong as desired for reporting.

Lastly, for the cases that compared the possibility of two donors being siblings or cousins, (Table 66, rows 33 to 35) there was again value added by the addition of the MPS data, both for the cases where sibling was the true relationship and for the case where cousin was the true relationship. In all three of these cases a CE STR statistic of 98.60% to 99.984% was increased to 99.9999% or more with the addition of the MPS loci. This shows a definite benefit to the combined analysis in this type of case.

As such, this work, which has specifically investigated the utility of the Precision ID Identity panel as a complement to CE STRs in kinship cases, has confirmed results published by others who have more generally explored the area of whether SNPs can aid in kinship analysis. A publication by Phillips and colleagues in 2012 (Phillips *et al.* 2012) explored this issue outside of the MPS area, by simply investigating if SNPs could in principle aid kinship cases by being added to the existing CE methods. Phillips *et al.* concluded that this definitely was the case, with even second-cousin relationships being able to be identified with as few as 7000 SNPs. For most applications, the authors recommend a 'medium-scale' multiplex of 256 to 1000 markers as being suitable for most 'challenging kinship analyses.' This then is a larger multiplex than the Precision ID Identity panel used here, which has 90 autosomal SNPs (the 34 Y-chromosome SNPs in the panel were not used in this work). As such, the results achieved in this work, where the panel can assist some types of case (such as sibling cases, most grandparent cases) but not others (e.g. cousins), make sense in the context of Phillips *et al*'s recommendation. It also makes sense relative to the work of others in the same field, such as Mo *et al.* (2018) and Wu *et al.* (2019) who found that 472 and 1245 marker SNP panels respectively were appropriate for use in distinguishing second-degree relationships, such as cousins.

In all kinship analysis it is important to demonstrate that the loci used to gain the kinship statistics are independent of each other, i.e. that the genotype a person has at one locus does not allow prediction of the genotype they will have at another locus due to the proximity of the two loci on the genome. As such, in real-world kinship analysis all loci that are used must either be on separate chromosomes or be sufficiently distant on the same chromosome such that recombination ensures that they are inherited separately (O'Connor and Tillmar, 2012). In this work, no literature was found that has demonstrated the independence of the CE STR and MPS panels used in this work, the GlobalFiler PCR amplification kit and the

Precision ID Ancestry kit respectively, despite their promotion for this type of combined analysis by the manufacturer. To explore the independence of these loci, the distance of each locus across the hg19 human genome was recorded and is shown in Table 67. This table notes the position of each locus in hg19 and also the distances between the loci on the same chromosome. The distances between the loci can also be visualised in Figure 43. Twenty-three locus pairs were found in the 112 total loci (22 STR and 90 SNP) that were less than five million bases apart from each other, with the shortest distance between two loci being the 146,638 bases between the two SNP loci rs6955448 and rs917118 on chromosome seven. It is of note that of these 23 locus pairs separated by less than five million bases, only six of the pairs are where an STR locus is near to a SNP locus, or in other words, a consequence of the CE and MPS-based panels being combined together. The smallest such STR to SNP distance is the 852,771 base distance between rs2269355 and vWA on chromosome 12. It is these six locus pairs that would need particular attention to test whether they perform independently of each other in kinship analysis.

Although position on the chromosome as examined here is a simple indication of whether loci can be expected to be independent of each other, true genetic independence of loci depends on the frequency at which recombination events occur across the genome, which varies from location to location. To measure this effect for a specific set of loci, a large set of genotypes for the loci in question is needed, from individuals who are not related to each other. A data set of this nature was not available for this work, and as such full demonstration of the independence of the loci used here was not possible. Another approach to this problem is described in Phillips *et al*. (2012) where high density SNPs from the HapMap project – a human haplotype map of over 3.1 million SNPs (Frazer *et al.* 2007) – was examined to infer the independence of 29 syntenic STR pairs. This approach was also used by Alsafiah *et al*. (2019) in their work evaluating the SureID 23comp Human Identification Kit, which adds several new STR loci to the set routinely used in commercially available CE-based STR kits. This would be an interesting avenue for future work and essential if these panels are to be used in practical forensic cases. Despite this, this does not affect the conclusion of this work as a proof of concept analysis of whether results from a commercially available MPS panel can be usefully added to a commonly used CE STR panel.

In summary, this work has shown that MPS offers promise for forensic laboratories that practice relationship testing. This is as a complement to, rather than as a replacement of, CE testing. In many cases, such as simple paternity cases, MPS is not necessary and CE will give a result that is entirely acceptable in itself, without the need for the extra expense and time of MPS analysis. In some cases however, such as sibling, grandparent and aunt/uncle

cases where the alternative hypothesis is that the donors are unrelated, or in cases where it is a question of determining which of a parent / pibling or sibling / cousin relationship is true, it is possible that after a CE result that is not entirely conclusive, an analyst may increase the result to something that is useable by using an MPS assay. The Precision ID Identity panel tested here has some promise in this area, as the 90 autosomal SNPs it provides have been shown to add useful information to some types of case, such as sibling or grandparent cases, and as an off-the-shelf ready to use kit, it may appeal to practicing forensic labs. Equally, due to the panels lack of success in adding further useful information to other case types, such as cousin cases, forensic labs may also prefer to use larger SNP panels for this purpose when adding MPS to the tools that they use in investigating kinship cases.

Chapter 7:
Conclusion

7. Conclusion

This work has addressed the question of whether massively parallel sequencing, or MPS, a method that has received much recent attention in forensic literature, actually confers benefits that are not able to be achieved with the CE-based methods that are currently typically used in forensic DNA laboratories.

The first phase of the work examined the sensitivity of MPS in comparison to CE methods. This was done by comparing the performance of the Precision ID Identity panel, an MPS assay consisting of 90 autosomal SNP markers, to the GlobalFiler PCR kit, a CE-based assay consisting of 24 STR markers. These two assays were run on the same set of low-level DNA samples to test their sensitivity. The results showed the MPS assay to have excellent sensitivity, with the Precision ID Identity panel giving better discrimination power in the genotype achieved for every sample tested (Figure 25). This result was added to in a further sensitivity test that ran the same sample set with the Precision ID mtDNA Whole genome panel, an MPS assay that targets mitochondrial DNA. The sensitivity of this test was even better than the first test, with no dropouts noted in any sample tested, down to 2 pg of input DNA, as measured by a nuclear DNA quantification (Table 11). This result is likely due to the inherent sensitivity of mitochondrial DNA analysis, a function of the large number of mtDNA copes per cell compared to the single cell nucleus, but still illustrates the practical utility of MPS given how much easier it is for a laboratory sequence mtDNA with MPS compared to previously used Sanger CE techniques.

A further interesting result from the mitochondrial work showed the importance of optimising the input of DNA into the Precision ID mtDNA Whole genome panel. These results showed that adding too much DNA could have the effect of reducing the number of sequencing reads that were achieved due to the formation of unwanted ‘super-amplicons’ that do not align to the reference sequence (See Figure 17, Figure 18, and Figure 19). Avoiding these, and properly optimising the input DNA to the panel would be an important step in the practical forensic implementation of this panel

The next section of this work examined the performance of MPS in the presence of inhibitors. Here, like others have found in studying other MPS assays, the Precision ID Ancestry Panel showed significantly worse inhibition response than the GlobalFiler CE assay. Amounts of humic acid that showed almost no effect in the CE analysis (Table 19) would almost entirely knock out the MPS assay (Table 17). This result illustrates the significant forensic specific development that has gone into current CE-based forensic PCR chemistry, something that has not yet been done on forensic MPS chemistry, which to date has largely been inherited from non-forensic applications where inhibitors are not as much of

an issue as in the forensic laboratory. This is an area that would benefit from the attention of the MPS manufacturers, although there is also an argument to be made that modern forensic extraction methods do such a good job of removing inhibitors from samples, that the inhibition response of the downstream PCR chemistry is not quite as important a factor as it once was.

Other results then examined the performance of MPS with mixed samples, this time using the prototype Precision ID Mixture panel, an as yet unreleased MPS panel from Thermo Fisher Scientific, which combines STR and microhaplotype markers in one panel, and is specifically designed for mixture analysis. Standardised samples consisting of two control DNAs mixed together in varying proportions were analysed with both this panel and the CE-based GlobalFiler kit. The results here did not favour MPS, with the CE method being capable of detecting 15 out of 26 alleles (57.7%) of the minor contributor in the weakest mixture tested (Table 32), with MPS being capable of detecting only 4 of the 30 alleles in the same mixture (13.3%) (Table 24). Improvements in the locus-to-locus balance of the MPS assay, along with related improvements and optimisation to the thresholds used in the analysis may aid this in future – at present a 2% relative analytical threshold is recommended by the manufacturer, which by definition limits the ability to detect mixtures below a theoretical 1:50 ratio – but at present this work has shown that current CE methods are at least as good at detecting mixtures as MPS methods.

The last component of Chapter 3 examined the utility of MPS on ‘real world’ or non-probative casework style samples. Given that the previous work found advantages for MPS in sensitivity of analysis, yet disadvantages in use with inhibited or mixed samples, as perhaps could be expected, the non-probative samples showed success in MPS with some samples, but not with others. Two samples in particular of the eleven in total that were tested showed a markedly better result with MPS, where a CE profile that was effectively ‘no result’ was turned into a useful partial profile in both cases – a practical illustration of the enhanced sensitivity of MPS analysis noted earlier (see Table 41, Figure 35, and Figure 36).

As a result of this testing, it seems that MPS does have a place in the forensic laboratory for testing of low level, potentially degraded samples that fail to achieve a good result with CE methods. These samples would need to be well extracted and purified at the pre-sequencing steps to ensure that no inhibitors are present, and if mixtures are present, then although the MPS method can still be used, it is unlikely to reveal any further information than the equivalent CE method would. As such, based on these results, it seems that MPS has a place in the forensic laboratory as a complement to, rather than as a replacement of, CE technology.

The next chapter of this work expanded upon the points already touched on regarding analysis thresholds and explored the variables that affect the definition of analytical thresholds in MPS. Results showed that the appropriate analytical threshold for an MPS run can vary significantly based on both the number of samples in the run, and the run type that is performed. This has significant implications for practicing forensic laboratories, who up until now have been able to define a single set of analysis parameter for a CE run regardless of how many samples are in it, but for MPS are faced with either having to put the same number of samples into every run, or have a different set of analysis parameters depending on the specifics of the run in question.

Related to this is the result found in the next section of work, which looked at the reproducibility of an MPS run, with identical runs of the Precision ID GlobalFiler NGS STR kit v2 being performed and the results compared. This showed a consistent 23% to 27% increase in sequencing reads on every sample. Both runs were within the specification of the manufacturer for overall run performance, and it seems that the variability seen between the two runs is an inherent part of the system. This has implications on the threshold discussion above, as this varying number of reads could affect the definition of analytical threshold, which as has been shown, can vary based on the number of reads achieved in the run based on the chip type used. More work is needed in this area to fully characterise the magnitude of the variability of run performance, and its resulting impact on analysis thresholds, on a larger data set.

The final section of Chapter 4 then explored the concordance of an MPS run, with the same Precision ID GlobalFiler NGS STR v2 kit being studied as in the reproducibility section. Results here showed the MPS assay to be generally very concordant with the CE method tested, the GlobalFiler PCR kit, with only ten out of 500 loci tested not exactly matching between CE and MPS result (see Table 53). All of these ten differences were not discordance however, but were examples of sequence variation being detected by MPS that cannot be seen by CE. Eight of the ten differences were isometric repeat variant heterozygotes, where the two alleles of an STR heterozygote are the same size, so cannot be distinguished by CE, only by analysis of the sequence in MPS. The remaining two differences were examples of where MPS detected multiple alleles due to SNPs in the STR flanking region. As such, this work found 500 out of 500, or 100%, concordance between STR profiles generated with MPS and CE based methods.

Chapter 5 of this work went on to examine the performance of the Precision ID Ancestry panel, an MPS based SNP assay, in predicting the ancestry of a set of 64 sample donors with self-declared ancestries from all around the world. The panel performed well in

identifying the continental ancestry of the 64 donors, with 45 of the donors having a predicted ancestry that matched their self-declared ancestry (Table 62). Twelve results were unclear and seven were ‘No match’, with there being no apparent pattern to the ancestries of those that did not match. Further investigation was then performed into how the two constituent panels of the Precision ID Ancestry panel, the ‘Kidd’ and ‘Seldin’ SNP sets performed in predicting the ancestry for the same sample set. The result showed that the Seldin panel taken in isolation was slightly worse than the combined effect of both panels together, while the Kidd panel performed about the same as the combined panel (Table 58). Overall, no advantage in splitting the panel could be seen, so it is advised for anyone using the Precision ID Ancestry panel for ancestry prediction to use it in its entirety. Further, the results of this study showed that the panel has good ability to predict the continental ancestry of a range of unknown donors, results that are in line with what others have reported for similar analysis. As such the prediction of ancestry via MPS SNP analysis could prove to be a useful tool in forensic investigation.

In the last section of this work, another forensic application of MPS was examined. This was kinship analysis, specifically the method of adding MPS SNP data to the CE STR results typically used in the analysis of this type of case. To investigate this, a set of samples from an extended family with known relationships was analysed, with profiles for each member of the family being generated by both the CE-based GlobalFiler kit and the MPS-based Precision ID Identity kit. Thirty-five different kinship scenarios were then tested, with the statistical strength of the CE data in isolation, the MPS data in isolation and the CE and MPS data combined being compared (Table 66). It was found that, as perhaps could be expected, for simple paternity scenarios, the CE data was adequate in giving a very strong match statistic on its own and there was no particular benefit in adding the MPS data to the analysis. Equally, for the more distant relationships tested, such as cousins and great-grandparents, both the CE and MPS data, even when combined together, failed to give a match statistic that would be useful in a practical case. In the middle of this range however, some cases were found where the CE result could be usefully boosted by the addition of the MPS data, as was the case in the sibling analysis, several of the grandparent analyses, and in cases where the chance of the putative relationship was compared to another relationship type, such as comparing a sibling to cousin relationship. As such, this work shows that the Precision ID Identity panel has potential as a useful tool for labs performing kinship analysis as a source of extra information in certain case types. With expansion of the SNPs used in this type of analysis, this can only improve in future and provide further benefits from MPS to this field.

Overall, this work points to a clear benefit of MPS over CE based analysis for certain forensic sample and case types, whilst also balancing this against the fact that these benefits may not currently apply to all scenarios that a practicing forensic laboratory encounters. Further barriers to adoption of MPS by practicing laboratories include simply the cost of the analysis: despite recent advances in MPS technology, MPS reagents, in particular, remain an order of magnitude more expensive than the equivalent CE-based reagents. It is possible that these prices will come down in future as the technology advances, but for now this remains a practical obstacle to routine implementation for many laboratories.

Another barrier to adoption is the significant amount of validation that would be required to adopt these MPS methods, as was discussed in Chapter 4. As the field develops however, and as the requirements for MPS validation become better understood, this burden should become lessened however as best practices for implementation of MPS become shared across the forensic community. For example, better characterisation of expected levels of run-to-run and sample-to-sample variation in an MPS run would be of great benefit in further defining what is needed in a practical MPS validation of analysis thresholds.

As such, it seems certain that CE will still have a place in the forensic laboratory for the foreseeable future. Despite this, it has been shown here that MPS is a method that can offer much to the forensic DNA laboratory, and adoption of MPS as a complement to existing CE technology would let a forensic lab significantly expand beyond what is possible with CE analysis only. This would offer significant advances in the detection of low level or degraded casework samples, in the prediction of sample ancestry, and in the analysis of cases involving related individuals. Future advances in the field, such as increased tolerance of MPS PCR assays to inhibition, expanded marker sets to allow for greater kinship resolution, and discovery of new SNP sets to allow greater ancestry discrimination and phenotype determination of presently little examined traits such as height and face shape would only increase this utility of MPS to the forensic laboratory.

This work has also shown that implementation of MPS methods into forensic practice must be done with care, with particular attention being paid to the analysis thresholds used. Development and management of these settings may need to be done in ways that are different to how CE analysis has been performed in the past. Despite this however, it is clear that MPS methods have a place in the forensic laboratory today and, if the field of MPS continues to evolve at its present fast rate, potentially much more to offer in future.

Chapter 8:
References

8. References

AL-ASFI M, MCNEVIN D, MEHTA B, POWER D, GAHAN ME, DANIEL R. 2018. Assessment of the Precision ID Ancestry panel. International Journal of Legal Medicine. 132 pp 1581-1594.

ALIFERI A, BALLARD D, GALLIDABINO MD, THURTLE H, BARRON L, SYNDERCOMBE COURT D. 2018. DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. Forensic Science International: Genetics. 37 pp 215-226.

ALSAFIAH HM, ALJANABI AA, HADI S, ALTURAYEIF SS, GOODWIN W. 2019. An evaluation of the SureID 23comp Human Identification Kit for kinship testing. Scientific Reports. 14 pp 16859.

ANANTHARAMAN R, SHUE BH, TAN SZ, WONG Y, SYN C. 2020. Differentiation of Asian population samples using the Illumina ForenSeq kit. Forensic Science International: Genetics. 48 pp 102318

AVILA E, FELKL AB, GRAEBIN P, NUNES CP, ALHO CS. 2019. Forensic characterization of Brazilian regional populations through massive parallel sequencing of 124 SNPs included in HID ion Ampliseq Identity Panel. Forensic Science International: Genetics. 40 pp 74-84.

AVILA E, GRAEBIN P, CHEMALE G, FREITAS J, KAHMANN A, ALHO CS. 2019. Full mtDNA genome sequencing of Brazilian admixed populations: A forensic-focused evaluation of a MPS application as an alternative to Sanger sequencing methods. Forensic Science International: Genetics. 42 pp 154-164.

BANDELT HJ, SALAS A. 2012. Current Next Generation Sequencing technology may not meet forensic standards. Forensic Science International: Genetics. 6 pp 143-145.

BARRIO PA, MARTÍN P, ALONSO A, MÜLLER P, BODNER M, BERGER B, PARSON W, BUDOWLE B; DNASEQEX CONSORTIUM. 2019. Massively parallel sequence data of 31 autosomal STR loci from 496 Spanish individuals revealed concordance with CE-STR technology and enhanced discrimination power. Forensic Science International: Genetics. 42 pp 49-55.

BARTLING CM, HESTER ME, BARTZ J, HEIZER E JR, FAITH SA. 2014. Next-generation sequencing approach to epigenetic-based tissue source attribution. Electrophoresis. 35 pp 3096-3101

BEADLING C, NEFF TL, HEINRICH MC, RHODES K, THORNTON M, LEAMON J, ANDERSEN M, CORLESS CL. 2013. Combining highly multiplexed PCR with semiconductor-based sequencing for rapid cancer genotyping, Journal of Molecular Diagnostics. 15 pp 171-176.

BIRD A. 2002. DNA methylation patterns and epigenetic memory. Genes and Development. 16 pp 6-21.

BLEKA Ø, EDUARDOFF M, SANTOS C, PHILLIPS C, PARSON W, GILL P. 2017. Open source software EuroForMix can be used to analyse complex SNP mixtures. Forensic Science International: Genetics. 31 pp 105-110.

BODNER M, IUVARO A, STROBL C, NAGL S, HUBER G, PELOTTI S, PETTENER D, LUISELLI D, PARSON W. 2015. Helena, the hidden beauty: Resolving the most common West Eurasian mtDNA control region haplotype by massively parallel sequencing an Italian population sample. Forensic Science International: Genetics. 15 pp 21-26.

BØRSTING C, FORDYCE SL, OLOFSSON J, MOGENSEN HS, MORLING N. 2014. Evaluation of the Ion Torrent HID SNP 169-plex: A SNP typing assay developed for human identification by second generation sequencing. Forensic Science International: Genetics. 12 pp 144-54.

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

BOTTINO CG, CHANG CW, WOOTTON S, RAJAGOPALAN N, LANGIT R, LAGACE R, SILVA R, MOURA-NETO RS. 2015. STR genotyping using ion torrent PGM and STR 24-plex system: Performance and data interpretation. Forensic Science International: Genetics Supplement Series. 5 pp 325–326.

BRANDHAGEN MD, JUST RS, IRWIN JA. 2020. Validation of NGS for mitochondrial DNA casework at the FBI Laboratory. Forensic Science International: Genetics. 44 pp 102151.

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. 2019. HIrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms. Forensic Science International: Genetics. 43 pp 102152.

BUCHARD A, KAMPMANN ML, POULSEN L, BØRSTING C, MORLING N. 2016. ISO 17025 validation of a next-generation sequencing assay for relationship testing. Electrophoresis. 37 pp 2822-2831.

BUDOWLE B, VAN DAAL A. 2008. Forensically relevant SNP classes. Biotechniques. 44 pp 603-608.

BUTLER JM 2012. Advanced topics in forensic DNA typing: methodology. Elsevier Academic Press. San Diego, CA, USA.

BUTLER JM 2015. Advanced topics in forensic DNA typing: interpretation. Elsevier Academic Press. San Diego, CA, USA.

BUTLER JM 2015 (2). The future of forensic DNA analysis. Philosophical Transactions of the Royal Society B: Biological Sciences. 370 pp 1674.

CANARD B, SARFATI RS. 1994. DNA polymerase fluorescent substrates with reversible 3′-tags. Gene. 148.1 pp 1-6.

CASALS F, ANGLADA R, BONET N, RASAL R, VAN DER GAAG KJ, HOOGENBOOM J, SOLE-MORATA N, COMAS D, CALAFELL F. 2017. Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations. Forensic Science International: Genetics. 30 pp 66-70.

CHAITANYA L, BRESLIN K, ZUÑIGA S, WIRKEN L, POŚPIECH E, KUKLA-BARTOSZEK M, SIJEN T, KNIJFF P, LIU F, BRANICKI W, KAYSER M, WALSH S. 2018. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Science International: Genetics. 35 pp 123-135.

CHEN P, DENG C, LI Z, PU Y, YANG J, YU Y, LI K, LI D, LIANG W, ZHANG L, CHEN F. 2019. A microhaplotypes panel for massively parallel sequencing analysis of DNA mixtures. Forensic Science International: Genetics. 40 pp 140-149.

CHEN P, YIN C, LI Z, PU Y, YU Y, ZHAO P, CHEN D, LIANG W, ZHANG L, CHEN F. 2018. Evaluation of the Microhaplotypes panel for DNA mixture analyses. Forensic Science International: Genetics. 35 pp 149-155.

CHURCHILL JD, CHANG J, GE J, RAJAGOPALAN N, WOOTTON SC, CHANG CW, LAGACÉ R, LIAO W, KING JL, BUDOWLE B. 2015. Blind study evaluation illustrates utility of the Ion PGM™ system for use in human identity DNA typing. Croatian Medical Journal. 56(3) pp 218-229.

CHURCHILL JD, NOVROSKI NMM, KING JL, SEAH LH, BUDOWLE B. 2017. Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System. Forensic Science International: Genetics. 30 pp 81-92.

CHURCHILL JD, SCHMEDES SE, KING JL, BUDOWLE B. 2016. Evaluation of the Illumina Beta Version ForenSeq DNA Signature Prep Kit for use in genetic profiling. Forensic Science International: Genetics. 20 pp 20-29.

CORNELIS S, GANSEMANS Y, VANDER PLAETSEN AS, WEYMAERE J, WILLEMS S, DEFORCE D, VAN NIEUWERBURGH F. 2019. Forensic tri-allelic SNP genotyping using nanopore sequencing. Forensic Science International: Genetics. 38 pp 204-210.

DALSGAARD S, ROCKENBAUER E, BUCHARD A, MOGENSEN HS, FRANK-HANSEN R, BØRSTING C, MORLING N 2014. Non-uniform phenotyping of D12S391 resolved by second generation sequencing. Forensic Science International: Genetics. 8 pp1 95–199.

DE KNIJFF P. 2019. From next generation sequencing to now generation sequencing in forensics. Forensic Science International: Genetics. 40 pp 182-191

DE LA PUENTE M, PHILLIPS C, XAVIER C, AMIGO J, CARRACEDO A, PARSON W, LAREU MV. 2020. Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems. Forensic Science International: Genetics. 45 pp 102213.

DEVESSE L, BALLARD D, DAVENPORT L, RIETHORST I, MASON-BUCK G, SYNDERCOMBE COURT D. 2018. Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups. Forensic Science International: Genetics 34. pp 57-61.

DIVNE A.M, EDLUND H, ALLEN M. 2010. Forensic analysis of autosomal STR markers using Pyrosequencing, Forensic Science International: Genetics. 4 pp 122-129.

EDUARDOFF M, SANTOS C, DE LA PUENTE M, GROSS TE, FONDEVILA M, STROBL C, SOBRINO B, BALLARD D, SCHNEIDER PM, CARRACEDO Á, LAREU MV, PARSON W, PHILLIPS C. 2015. Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM. Forensic Science International: Genetics. 17 pp 110-121.

EID J, FEHR A, GRAY J, LUONG K, LYLE J, OTTO G, PELUSO P, RANK D, BAYBAYAN P, BETTMAN B, BIBILLO A, BJORNSON K, CHAUDHURI B, CHRISTIANS F, CICERO R, CLARK S, DALAL R, DEWINTER A, DIXON J, FOQUET M, GAERTNER A, HARDENBOL P, HEINER C, HESTER K, HOLDEN D, KEARNS G, KONG X, KUSE R, LACROIX Y, LIN S, LUNDQUIST P, MA C, MARKS P, MAXHAM M, MURPHY D, PARK I, PHAM T, PHILLIPS M, ROY J, SEBRA R, SHEN G, SORENSON J, TOMANEY A, TRAVERS K, TRULSON M, VIECELI J, WEGENER J, WU D, YANG A, ZACCARIN D, ZHAO P, ZHONG F, KORLACH J, TURNER S. 2009. Real-time DNA sequencing from single polymerase molecules. Science. 323 pp 133-138.

ELENA S, ALESSANDRO A, IGNAZIO C, SHARON W, LUIGI R, ANDREA B. 2016. Revealing the challenges of low template DNA analysis with the prototype Ion AmpliSeq™ Identity panel v2.3 on the PGM™ Sequencer. Forensic Science International: Genetics. 22 pp 25-36.

FANG C, ZHAO J, LIU X, ZHANG J, CAO Y, YANG Y, YU C, ZHANG X, QIAN J, LIU W, WU H, YAN J. 2019. MicroRNA profile analysis for discrimination of monozygotic twins using massively parallel sequencing and real-time PCR. Forensic Science International: Genetics 38. pp 23-31.

FLECKHAUS J, SCHNEIDER PM. 2020. Novel multiplex strategy for DNA methylation-based age prediction from small amounts of DNA via Pyrosequencing. Forensic Science International Genetics. 44 pp 102189.

FORAT S, HUETTEL B, REINHARDT R, FIMMERS R, HAIDL G, DENSCHLAG D, OLEK K. 2016. Methylation Markers for the Identification of Body Fluids and Tissues from Forensic Trace Evidence. PLOS One. 11. e0147973

FORDYCE SL, MOGENSEN HS, BØRSTING C, LAGACÉ RE, CHANG CW, RAJAGOPALAN N, MORLING N. 2015. Second-generation sequencing of forensic STRs using the Ion Torrent™ HID STR 10-plex and the Ion PGM. Forensic Science International: Genetics. 14 pp 132-40.

FRAZER KA, BALLINGER DG, COX DR, HINDS DA, STUVE LL, GIBBS RA, BELMONT JW, BOUDREAU A, HARDENBOL P, LEAL SM, PASTERNAK S, WHEELER DA, WILLIS TD, YU F, YANG H, ZENG C, GAO Y, HU H, HU W, LI C, LIN W, LIU S, PAN H, TANG X, WANG J, WANG W, YU J, ZHANG B, ZHANG Q, ZHAO H, ZHAO H, ZHOU J, GABRIEL SB, BARRY R, BLUMENSTIEL B, CAMARGO A, DEFELICE M, FAGGART M, GOYETTE M, GUPTA S, MOORE J, NGUYEN H, ONOFRIO RC, PARKIN M, ROY J, STAHL E, WINCHESTER E, ZIAUGRA L, ALTSHULER D, SHEN Y, YAO Z, HUANG W, CHU X, HE Y, JIN L, LIU Y, SHEN Y, SUN W, WANG H, WANG Y, WANG Y, XIONG X, XU L, WAYE MM,

TSUI SK, XUE H, WONG JT, GALVER LM, FAN JB, GUNDERSON K, MURRAY SS, OLIPHANT AR, CHEE MS, MONTPETIT A, CHAGNON F, FERRETTI V, LEBOEUF M, OLIVIER JF, PHILLIPS MS, ROUMY S, SALLÉE C, VERNER A, HUDSON TJ, KWOK PY, CAI D, KOBOLDT DC, MILLER RD, PAWLIKOWSKA L, TAILLON-MILLER P, XIAO M, TSUI LC, MAK W, SONG YQ, TAM PK, NAKAMURA Y, KAWAGUCHI T, KITAMOTO T, MORIZONO T, NAGASHIMA A, OHNISHI Y, SEKINE A, TANAKA T, TSUNODA T, DELOUKAS P, BIRD CP, DELGADO M, DERMITZAKIS ET, GWILLIAM R, HUNT S, MORRISON J, POWELL D, STRANGER BE, WHITTAKER P, BENTLEY DR, DALY MJ, DE BAKKER PI, BARRETT J, CHRETIEN YR, MALLER J, MCCARROLL S, PATTERSON N, PE'ER I, PRICE A, PURCELL S, RICHTER DJ, SABETI P, SAXENA R, SCHAFFNER SF, SHAM PC, VARILLY P, ALTSHULER D, STEIN LD, KRISHNAN L, SMITH AV, TELLO-RUIZ MK, THORISSON GA, CHAKRAVARTI A, CHEN PE, CUTLER DJ, KASHUK CS, LIN S, ABECASIS GR, GUAN W, LI Y, MUNRO HM, QIN ZS, THOMAS DJ, MCVEAN G, AUTON A, BOTTOLO L, CARDIN N, EYHERAMENDY S, FREEMAN C, MARCHINI J, MYERS S, SPENCER C, STEPHENS M, DONNELLY P, CARDON LR, CLARKE G, EVANS DM, MORRIS AP, WEIR BS, TSUNODA T, MULLIKIN JC, SHERRY ST, FEOLO M, SKOL A, ZHANG H, ZENG C, ZHAO H, MATSUDA I, FUKUSHIMA Y, MACER DR, SUDA E, ROTIMI CN, ADEBAMOWO CA, AJAYI I, ANIAGWU T, MARSHALL PA, NKWODIMMAH C, ROYAL CD, LEPPERT MF, DIXON M, PEIFFER A, QIU R, KENT A, KATO K, NIIKAWA N, ADEWOLE IF, KNOPPERS BM, FOSTER MW, CLAYTON EW, WATKIN J, GIBBS RA, BELMONT JW, MUZNY D, NAZARETH L, SODERGREN E, WEINSTOCK GM, WHEELER DA, YAKUB I, GABRIEL SB, ONOFRIO RC, RICHTER DJ, ZIAUGRA L, BIRREN BW, DALY MJ, ALTSHULER D, WILSON RK, FULTON LL, ROGERS J, BURTON J, CARTER NP, CLEE CM, GRIFFITHS M, JONES MC, MCLAY K, PLUMB RW, ROSS MT, SIMS SK, WILLEY DL, CHEN Z, HAN H, KANG L, GODBOUT M, WALLENBURG JC, L'ARCHEVÊQUE P, BELLEMARE G, SAEKI K, WANG H, AN D, FU H, LI Q, WANG Z, WANG R, HOLDEN AL, BROOKS LD, MCEWEN JE, GUYER MS, WANG VO, PETERSON JL, SHI M, SPIEGEL J, SUNG LM, ZACHARIA LF, COLLINS FS, KENNEDY K, JAMIESON R, STEWART J. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature. 449 pp 851-61.

GARCIA O, ALONSO S, HUBER N, BODNER M, PARSON W. 2020. Forensically relevant phylogeographic evaluation of mitogenome variation in the Basque Country. Forensic Science International: Genetics. 46 pp 102260.

GARCIA O, AJURIAGERRA JA, ALDAY A, ALONSO S, PEREZ JA, SOTO A, URIARTE I, YURREBASO I. 2017. Frequencies of the Precision ID Ancestry Panel markers in Basques using the Ion Torrent PGM platform. Forensic Science International: Genetics. 31 pp e1-e4.

GARCIA O, SOTO A, YURREBASO I. 2017 (2). Allele frequencies and other forensic parameters of the HID-Ion AmpliSeq Identity Panel markers in Basques using the Ion Torrent PGM platform. Forensic Science International: Genetics. 28 pp 8-10.

GELARDI C, ROCKENBAUER E, DALSGAARD S, BØRSTING C, MORLING N. 2014. Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. Forensic Science International: Genetics. 12 pp 38-41.

GETTINGS KB, APONTE RA, VALLONE PM, BUTLER JM. 2015. STR allele sequence variation: Current knowledge and future issues. Forensic Science International: Genetics. 18 pp 118-130.

GETTINGS KB, BALLARD D2, BODNER M, BORSUK LA, KING JL, PARSON W, PHILLIPS C. Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting. Forensic Science International: Genetics. 43 pp 102165.

GETTINGS KB, BORSUK LA, BALLARD D, BODNER M, BUDOWLE B, DEVESSE L, KING J, PARSON W, PHILLIPS C, VALLONE PM. 2017. STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci. Forensic Science International: Genetics. 31 pp 111-117.

GETTINGS KB, KIESLER KM, FAITH SA, MONTANO E, BAKER CH, YOUNG BA, GUERRIERI RA, VALLONE PM. 2016. Sequence variation of 22 autosomal STR loci detected by next generation sequencing. Forensic Science International: Genetics. 21 pp 5–21.

GETTINGS KB, KIESLER KM, VALLONE PM. 2015. Performance of a next generation sequencing SNP assay on degraded DNA. Forensic Science International: Genetics. 19 pp 1-9.

GOODWIN W. 2015. DNA profiling: The first 30 years. Science and Justice. 55 pp 375–376.

GOLDFEDER RL, PRIEST JR, ZOOK JM, GROVE ME, WAGGOTT D, WHEELER MT, SALIT M, ASHLEY EA. 2016. Medical implications of technical accuracy in genome sequencing. Genome Medicine. 8.24 pp 1-12.

GUO F, YU J, ZHANG L, LI J. 2017. Massively parallel sequencing of forensic STRs and SNPs using the Illumina ForenSeq DNA Signature Prep Kit on the MiSeq FGx Forensic Genomics System. Forensic Science International: Genetics. 31 pp 135-148.

GUO F, ZHOU Y, LIU F, YU J, SONG H, SHEN H, ZHAO B, JIA F, HOU G, JIANG X. 2016. Evaluation of the Early Access STR Kit v1 on the Ion Torrent PGM™ platform. Forensic Science International: Genetics 23. pp 111-120.

GUO F, ZHOU Y, SONG H, ZHAO J, SHEN H, ZHAO B, LIU F, JIANG X. 2016 [2]. Next generation sequencing of SNPs using the HID-Ion AmpliSeq™ Identity Panel on the Ion Torrent PGM™ platform. Forensic Science International: Genetics. 25 pp 73-84

HALDER I, SHRIVER M, THOMAS M, FERNANDEZ JR, FRUDAKIS T. 2008. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Human Mutation. 29 pp 648-658.

HAMMER Ø, HARPER DAT, RYAN PD. 2001 Paleontological statistics software package for education and data analysis. Palaeontologia Electronica. 4 pp 9.

HAQUE F, LI J, WU HC, LIANG XJ, GUO P.2013. Solid-State and Biological Nanopore for Real-Time Sensing of Single Chemical and Sequencing of DNA. Nano Today 8. Pp 56-74.

HERNANDO B, IBAÑEZ MV, DESERIO-CUESTA JA, SORIA-NAVARRO R, VILAR-SASTRE I, MARTINEZ-CADENAS C. 2018. Genetic determinants of freckle occurrence in

the Spanish population: Towards ephelides prediction from human DNA samples. Forensic Science International: Genetics. 33 pp 38-47.

HOLLARD C, AUSSET L, CHANTREL Y, JULLIEN S, CLOT M, FAIVRE M, SUZANNE É, PÈNE L, LAURENT FX. 2019. Automation and developmental validation of the ForenSeq™ DNA Signature Preparation kit for high-throughput analysis in forensic laboratories. Forensic Science International: Genetics. 40 pp 37-45.

HUSSING C, BØRSTING C, MOGENSEN HS, MORLING N. 2015. Testing of the Illumina ForenSeq kit. Forensic Science International: Genetics Supplement Series 5 pp 449–450.

HUSSING C, HUBER C, BYTYCI R, MOGENSEN HS, MORLING N, BØRSTING C. 2018. Sequencing of 231 forensic genetic markers using the MiSeq FGx™ forensic genomics system - an evaluation of the assay and software. Forensic Sciences Research. 3(2) pp 111-123.

HUSZAR TI, JOBLING MA, WETTON JH. 2018. A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing. Forensic Science International: Genetics. 35 pp 97-106.

HUSZAR TI, WETTON JH, JOBLING MA. 2019. Mitigating the effects of reference sequence bias in single-multiplex massively parallel sequencing of the mitochondrial DNA control region. Forensic Science International: Genetics. 40 pp 9-17.

JACOBS LC, WOLLSTEIN A, LAO O, HOFMAN A, KLAVER CC, UITTERLINDEN AG, NIJSTEN T, KAYSER M, LIU F. 2013. Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans. Human Genetics. 132 pp 147-158

JÄGER AC, ALVAREZ ML, DAVIS CP, GUZMAN 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. 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 pp 52–70.

JIN S, CHASE M, HENRY M, ALDERSON G, MORROW JM, MALIK S, BALLARD D, MCGRORY J, FERNANDOPULLE N, MILLMAN J, LAIRD J. 2018. Implementing a biogeographic ancestry inference service for forensic casework. Electrophoresis. 39 pp 2757-2765

JOLLIFFE IT, CADIMA J. 2016. Principal component analysis: a review and recent developments. Philosophical Transactions. Series A 374 pp 20150202.

JURAS A, CHYLEŃSKI M, KRENZ-NIEDBAŁA M, MALMSTRÖM H, EHLER E, POSPIESZNY Ł, ŁUKASIK S, BEDNARCZYK J, PIONTEK J, JAKOBSSON M, DABERT M. 2017. Investigating kinship of Neolithic post-LBK human remains from Krusza Zamkowa, Poland using ancient DNA. Forensic Science International: Genetics. 26 pp 30-39.

JUST RS, IRWIN JA, PARSON W. 2015. Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. Forensic Science International: Genetics. 18 pp 131-139.

JUST RS, MORENO LI, SMERICK JB, IRWIN JA. 2017. Performance and concordance of the ForenSeq system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens. Forensic Science International: Genetics. 28 pp 1–9.

KADER F, GHAI M. 2015. DNA methylation and application in forensic sciences. Forensic Science International. 249 pp 255-265.

KARAMIZADEH S, SBDULLAH SM, MANAF1 AA, ZAMANI1 M, HOOMAN A. 2013. An Overview of Principal Component Analysis. Journal of Signal and Information Processing. 4 pp 173-175

KARGER BL, GUTTMAN A. 2009. DNA Sequencing by Capillary Electrophoresis. Electrophoresis. 30 pp 196-202.

KAYSER M. 2015. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Science International: Genetics. 18 pp 33-48.

KERSBERGEN P, VAN DUIJN K, KLOOSTERMAN AD, DEN DUNNEN JT, KAYSER M, DE KNIJFF P. 2009. Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. BMC Genetics. 10 pp 69.

KIDD JR, FRIEDLAENDER FR, SPEED WC, PAKSTIS AJ, DE LA VEGA FM, KIDD KK. 2011. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investigative Genetics. 2 pp 1.

KIDD KK, KIDD JR, SPEED WC, FANG R, FURTADO MR, HYLAND FC, PAKSTIS AJ. 2012. Expanding data and resources for forensic use of SNPs in individual identification. Forensic Science International: Genetics. 6 pp 646-652.

KIDD KK, PAKSTIS AJ, SPEED WC, GRIGORENKO EL, KAJUNA SL, KAROMA NJ, KUNGULILO S, KIM JJ, LU RB, ODUNSI A, OKONOFUA F, PARNAS J, SCHULZ LO, ZHUKOVA OV, KIDD JR. 2006. Developing a SNP panel for forensic identification of individuals. Forensic Science International. 164 pp 20-32.

KIDD KK, PAKSTIS AJ, SPEED WC, LAGACÉ R, CHANG J, WOOTTON S, HAIGH E, KIDD JR. 2014. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. Forensic Science International Genetics. 12 pp 215-224.

KIDD KK, PAKSTIS AJ, SPEED WC, LAGACE R, CHANG J, WOOTTON S, IHUEGBU N. 2013. Microhaplotype loci are a powerful new type of forensic marker. Forensic Science International: Genetics Supplement Series. 4 pp 123-124.

KIDD KK, SPEED WC, PAKSTIS AJ, PODINI DS, LAGACÉ R, CHANG J, WOOTTON S, HAIGH E, SOUNDARARAJAN U. 2017. Evaluating 130 microhaplotypes across a global set of 83 populations. Forensic Science International: Genetics. 29 pp 29-37.

KIM EH, LEE HY, KWON SY, LEE EY, YANG WI, SHINN KJ. 2017. Sequence-based diversity of 23 autosomal STR loci in Koreans investigated using an in-house massively parallel sequencing panel. Forensic Science International: Genetics. 30 pp 134-140.

KING JL, LARUE BL, NOVROSKI NM, STOLJAROVA M, SEO SB, ZENG X, WARSHAUER DH, DAVIS CP, PARSON W, SAJANTILA A, BUDOWLE B. 2014. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Science International: Genetics. 12 pp 128-135.

KLING, D, TILLMAR, AO, EGELAND, T. 2014. Familias 3-Extensions and new functionality. Forensic Science International: Genetics. 13 pp 121-127.

KÖCHER S, MÜLLER P, BERGER B, BODNER M, PARSON W, ROEWER L, WILLUWEIT S, DNASEQEX CONSORTIUM. 2018. Inter-laboratory validation study of the ForenSeq™ DNA Signature Prep Kit. Forensic Science International: Genetics. 36 pp 77-85.

KOSOY R, NASSIR R, TIAN C, WHITE PA, BUTLER LM, SILVA G, KITTLES R, ALARCON-RIQUELME ME, GREGERSEN PK, BELMONT JW, DE LA VEGA FM, SELDIN MF. 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Human Mutation 2009. 30 pp 69-78.

KUKLA-BARTOSZEK M, POŚPIECH E, SPÓLNICKA M, KARŁOWSKA-PIK J, STRAPAGIEL D, ŻĄDZIŃSKA E, ROSSET I, SOBALSKA-KWAPIS M, SŁOMKA M, WALSH S, KAYSER M, SITEK A, BRANICKI W. 2018. Investigating the impact of age-depended hair colour darkening during childhood on DNA-based hair colour prediction with the HIrisPlex system. Forensic Science International: Genetics. 36 pp 26-33.

KUKLA-BARTOSZEK M, POŚPIECH E, WOŹNIAK A, BOROŃ M, KARŁOWSKA-PIK J, TEISSEYRE P, ZUBAŃSKA M, BRONIKOWSKA A, GRZYBOWSKI T, PŁOSKI R, SPÓLNICKA M, BRANICKI W. 2019. DNA-based predictive models for the presence of freckles. Forensic Science International: Genetics. 42 pp 252-259.

LEVY S, SUTTON G, NG PC, FEUK L, HALPERN AL, WALENZ BP, AXELROD N, HUANG J, KIRKNESS EF, DENISOV G, LIN Y, MACDONALD JR, PANG AW, SHAGO M, STOCKWELL TB, TSIAMOURI A, BAFNA V, BANSAL V, KRAVITZ SA, BUSAM DA, BEESON KY, MCINTOSH TC, REMINGTON KA, ABRIL JF, GILL J, BORMAN J, ROGERS YH, FRAZIER ME, SCHERER SW, STRAUSBERG RL, VENTER JC. 2007. The diploid genome sequence of an individual human. PLoS Biology. 5 pp 254.

LI CX, PAKSTIS AJ, JIANG L, WEI YL, SUN QF, WU H, BULBUL O, WANG P, KANG LL, KIDD JR, KIDD KK. 2016. A panel of 74 AISNPs: Improved ancestry inference within Eastern Asia. Forensic Science International: Genetics. 23 pp 101-110.

LI H, ZHAO X, MA K, CAO Y, ZHOU H, PING Y, SHAO C, XIE J, LIU W. 2017. Applying massively parallel sequencing to paternity testing on the Ion Torrent Personal Genome Machine. Forensic Science International: Genetics. 31 pp 155-159.

LI R, LI H, PENG D, HAO B, WANG Z, HUANG E, WU R, SUN H. 2019. Improved pairwise kinship analysis using massively parallel sequencing. Forensic Science International: Genetics. 38 pp 77-85

LIN CY, TSAI LC, HSIEH HM, HUANG CH, YU YJ, TSENG B, LINACRE A, LEE JCI. 2017. Investigation of length heteroplasmy in mitochondrial DNA control region by massively parallel sequencing. Forensic Science International: Genetics. 30 pp 127-133.

LIU F, CHEN Y,, ZHU G, HYSI PG, WU S, ADHIKARI K, BRESLIN K, POSPIECH E, HAMER MA, PENG F,, MURALIDHARAN C, ACUNA-ALONZO V, CANIZALES-QUINTEROS S, BEDOYA G, GALLO C, POLETTI G, ROTHHAMMER F, BORTOLINI MC, GONZALEZ-JOSE R, ZENG C, XU S, JIN L,, UITTERLINDEN AG, IKRAM MA, VAN DUIJN CM, NIJSTEN T, WALSH S, BRANICKI W, WANG S, RUIZ-LINARES A, SPECTOR TD, MARTIN NG, MEDLAND SE, KAYSER M. 2018. Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair. Human Molecular Genetics. 27 pp 559-575.

LIU F, HENDRIKS AE, RALF A, BOOT AM, BENYI E, SÄVENDAHL L, OOSTRA BA, VAN DUIJN C, HOFMAN A, RIVADENEIRA F, UITTERLINDEN AG, DROP SL, KAYSER M. 2014. Common DNA variants predict tall stature in Europeans. Human Genetics. 133 pp 587-597.

LIU F, ZHONG K, JING X, UITTERLINDEN AG, HENDRIKS AEJ, DROP SLS, KAYSER M. 2019. Update on the predictability of tall stature from DNA markers in Europeans. Forensic Science International: Genetics. 42 pp 8-13.

LIU J, WANG Z, HE G, ZHAO X, WANG M, LUO T, LI C, HOU Y. 2018. Massively parallel sequencing of 124 SNPs included in the precision ID identity panel in three East Asian minority ethnicities. Forensic Science International: Genetics. 35 pp 141-148.

MA Y, KUANG JZ, NIE TG, ZHU W, YANG Z. 2016. Next generation sequencing: Improved resolution for paternal/maternal duos analysis. Forensic Science International: Genetics. 24 pp 83-85.

MA K, ZHAO X, LI H, CAO Y, LI W, OUYANG J, XIE L, LIU W. 2018. Massive parallel sequencing of mitochondrial DNA genomes from mother-child pairs using the Ion Torrent Personal Genome Machine (PGM). Forensic Science International: Genetics. 32 pp 88-93.

MARDIS ER. 2008, Next-generation DNA sequencing methods, Annual Review of Genomics and Human Genetics. 9 pp 387-402.

MAROÑAS O, PHILLIPS C2, SÖCHTIG J, GOMEZ-TATO A, CRUZ R, ALVAREZ-DIOS J, DE CAL MC, RUIZ Y, FONDEVILA M, CARRACEDO Á, LAREU MV1. 2014. Development of a forensic skin colour predictive test. Forensic Science International: Genetics. 13 pp 34-44.

MARSHALL C, STURK-ANDREAGGI K, DANIELS-HIGGENBOTHAM J, OLIVER RS, BARRITT-ROSS S, MCMAHON TP. 2017. Performance evaluation of a mitogenome capture and Illumina sequencing protocol using non-probative, case-type skeletal samples: Implications for the use of a positive control in a next-generation sequencing procedure. Forensic Science International: Genetics. 31 pp 198-206.

MCELHOE JA, HOLLAND MM, MAKOVA KD, SU MS, PAUL IM, BAKER CH, FAITH SA, YOUNG B. 2014 Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. Forensic Science International: Genetics. 13 pp 20-29.

MCINERNEY P, ADAMS P, HADI MZ. 2014. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. Molecular Biology International 287430.

MEIKLEJOHN KA, ROBERTSON JM. 2017. Evaluation of the Precision ID Identity Panel for the Ion Torrent PGM sequencer. Forensic Science International: Genetics. 31 pp 48-56.

MEISSNER C, RITZ-TIMME S. 2010. Molecular pathology and age estimation. Forensic Science International. 203 pp 34-43.

MIKKELSEN M, FRANK-HANSEN R, HANSEN AJ, MORLING N 2014 Massively parallel pyrosequencing of the mitochondrial genome with the 454 methodology in forensic genetics. Forensic Science International: Genetics. 12 pp 30–37.

MILLAT G, CHANAVAT V, ROUSSON R. 2014. Evaluation of a New High-Throughput Next-Generation Sequencing Method Based on a Custom AmpliSeq Library and Ion Torrent PGM Sequencing for the Rapid Detection of Genetic Variations in Long QT Syndrome. Molecular Diagnosis & Therapy. 18 pp 533-539.

MO SK, REN ZL, YANG YR, LIU YC, ZHANG JJ, WU HJ, LI Z, BO XC, WANG SQ, YAN JW, NI M. 2018. A 472-SNP panel for pairwise kinship testing of second-degree relatives. Forensic Science International: Genetics. 34 pp 178-185.

MOGENSEN HS, TVEDEBRINK T, BØRSTING C, PEREIRA V, MORLING N. 2020. Ancestry prediction efficiency of the software GenoGeographer using a z-score method and the ancestry informative markers in the Precision ID Ancestry Panel. Forensic Science International: Genetics. 44 pp 102154.

MONTANO EA, BUSH JM, GARVER AM, LARIJANI MM, WIECHMAN SM, BAKER CH, WILSON MR, GUERRIERI RA, BENZINGER EA, GEHRES DN, DICKENS ML. 2018. Optimization of the Promega PowerSeq™ Auto/Y system for efficient integration within a forensic DNA laboratory. Forensic Science International: Genetics. 32 pp 26-32.

MÜLLER P, ALONSO A, BARRIO PA, BERGER B, BODNER M, MARTIN P, PARSON W, DNASEQEX CONSORTIUM. 2018. Systematic evaluation of the early access Applied Biosystems Precision ID Globalfiler Mixture ID and Globalfiler NGS STR panels for the Ion S5 system. Forensic Science International: Genetics. 36 pp 95-103.

MUSGRAVE-BROWN E, BALLARD D, BALOGH K, BENDER K, BERGER B, BOGUS M, BØRSTING C, BRION M, FONDEVILA M, HARRISON C, OGUZTURUN C, PARSON W, PHILLIPS C, PROFF C, RAMOS-LUIS E, SANCHEZ JJ, SÁNCHEZ DIZ P, SOBRINO REY B, STRADMANN-BELLINGHAUSEN B, THACKER C, CARRACEDO A, MORLING N, SCHEITHAUER R, SCHNEIDER PM, SYNDERCOMBE COURT D. 2007. Forensic validation of the SNPforID 52-plex assay. Forensic Science International: Genetics. 1 pp 186–190.

NAKANISHI H, PEREIRA V, BØRSTING C, YAMAMOTO T, TVEDEBRINK T, HARA M, TAKADA A, SAITO K, MORLING N. 2018. Analysis of mainland Japanese and Okinawan Japanese populations using the precision ID Ancestry Panel. Forensic Science International: Genetics. 33 pp 106-109.

NATIONAL DNA DATABASE STRATEGY BOARD BIENNIAL REPORT 2018 – 2020. 2020. United Kingdom Home Office. Published September 2020 at gov.uk

NAUE J, HOEFSLOOT HCJ, MOOK ORF, RIJLAARSDAM-HOEKSTRA L, VAN DER ZWALM MCH, HENNEMAN P, KLOOSTERMAN AD, VERSCHURE PJ. 2017. Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression. Forensic Science International: Genetics. 31 pp 19-28.

NAUE J, SÄNGER T, HOEFSLOOT HCJ, LUTZ-BONENGEL S, KLOOSTERMAN AD, VERSCHURE PJ. 2018. Proof of concept study of age-dependent DNA methylation markers across different tissues by massive parallel sequencing. Forensic Science International: Genetics. 36 pp 152-159.

O'CONNOR KL, TILLMAR AO. 2012. Effect of linkage between vWA and D12S391 in kinship analysis. Forensic Science International: Genetics. 6 pp 840-844.

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. 2016. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Science International: Genetics. 22 pp 54-63.

PARSON W, HUBER G, MORENO L, MADEL MB, BRANDHAGEN MD, NAGL S, XAVIER C, EDUARDOFF M, CALLAGHAN TC, IRWIN JA. 2015. Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples. Forensic Science International: Genetics. 15 pp 8–15.

PARSON W, STROBL C, HUBER G, ZIMMERMANN B, GOMES SM, SOUTO L, FENDT L, DELPORT R, LANGIT R, WOOTTON S, LAGACÉ R, IRWIN J. 2013. Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Science International: Genetics. 7 pp 543-9.

PECK MA, BRANDHAGEN MD, MARSHALL C, DIEGOLI TM, IRWIN JA, STURK-ANDREAGGI K. 2016. Concordance and reproducibility of a next generation mtGenome sequencing method for high-quality samples using the Illumina MiSeq. Forensic Science International: Genetics. 24 pp 103-111.

PEREIRA V, LONGOBARDI A, BØRSTING C. 2018. Sequencing of mitochondrial genomes using the Precision ID mtDNA Whole Genome Panel. Electrophoresis. 39 pp 2766-2775.

PEREIRA V, MOGENSEN HS, BØRSTING C, MORLING N. 2017. Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers. Forensic Science International: Genetics. 28 pp 138-145.

PETERSEN J, MOHAMMAD AA. 2001 Clinical and Forensic Applications of Capillary Electrophoresis. Humana Press Inc. Totowa, NJ, USA.

PETROVICK MS, BOETTCHER T, FREMONT-SMITH P, PERAGALLO C, RICKE DO, WATKINS J, SCHWOEBEL E. 2020. Analysis of complex DNA mixtures using massively parallel sequencing of SNPs with low minor allele frequencies. Forensic Science International: Genetics 46 pp 102234.

PHILLIPS C. 2015. Forensic genetic analysis of bio-geographical ancestry. Forensic Science International: Genetics. 18 pp 49-65.

PHILLIPS C, BALLARD D, GILL P, COURT DS, CARRACEDO A, LAREU MV. 2012. The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data. Forensic Science International: Genetics. 6 pp 354-365.

PHILLIPS C, GARCÍA-MAGARIÑOS M, SALAS A, CARRACEDO A, LAREU MV. 2012. SNPs as Supplements in Simple Kinship Analysis or as Core Markers in Distant Pairwise Relationship Tests: When Do SNPs Add Value or Replace Well-Established and Powerful STR Tests? Transfusion Medicine and Hemotherapy. 39 pp 202-210.

PHILLIPS C, GETTINGS KB, KING JL, BALLARD D, BODNER M, BORSUK L, PARSON W. 2018. "The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide. Forensic Science International: Genetics. 34 pp 162-169.

PHILLIPS C, MCNEVIN D, KIDD KK, LAGACÉ R, WOOTTON S, DE LA PUENTE M, FREIRE-ARADAS A, MOSQUERA-MIGUEL A, EDUARDOFF M, GROSS T, DAGOSTINO L, POWER D, OLSON S, HASHIYADA M, OZ C, PARSON W, SCHNEIDER PM, LAREU MV, DANIEL R. 2019. MAPlex - A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. Forensic Science International: Genetics. 42 pp 213-226.

PHILLIPS C, PARSON W, LUNDSBERG B, SANTOS C, FREIRE-ARADAS A, TORRES M, EDUARDOFF M, BØRSTING C, JOHANSEN P, FONDEVILA M, MORLING N, SCHNEIDER P, CARRACEDO A, LAREU MV. 2014. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. Forensic Science International: Genetics. 11 pp 13-25.

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. 2007. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Science International: Genetics. 1 pp 273-280.

PILLI E, AGOSTINO A, VERGANI D, SALATA E, CIUNA I, BERTI A, CARAMELLI D, LAMBIASE S. 2016. Human identification by lice: A Next Generation Sequencing challenge. Forensic Science International. 266 pp 71-78.

PINTO N, MAGALHÃES M, CONDE-SOUSA E, GOMES C, PEREIRA R, ALVES C, GUSMÃO L, AMORIM A. 2013. Assessing paternities with inconclusive STR results: The suitability of bi-allelic markers. Forensic Science International: Genetics. 7 pp 16-21.

POPTSOVA MS, IL'ICHEVA IA, NECHIPURENKO DY, PANCHENKO LA, KHODIKOV MV, OPARINA NY, POLOZOV RV, NECHIPURENKO YD, GROKHOVSKY SL. 2014. Non-random DNA fragmentation in next-generation sequencing. Scientific Reports 4 pp 4532.

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. 2018. Towards broadening Forensic DNA Phenotyping beyond pigmentation: Improving the prediction of head hair shape from DNA. Forensic Science International: Genetics 37 pp 241-251.

PRITCHARD JK, STEPHENS M, DONNELLY P. 2000. Inference of population structure using multilocus genotype data. Genetics 155 pp 945-959.

RALF A, VAN OVEN M, MONTIEL GONZÁLEZ D, DE KNIJFF P, VAN DER BEEK K, WOOTTON S, LAGACÉ R, KAYSER M. 2019. Forensic Y-SNP analysis beyond SNaPshot: High-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing. Forensic Science International: Genetics 41 pp 93-106.

RALF A, VAN OVEN M, ZHONG K, KAYSER M. 2015. Simultaneous analysis of hundreds of Y-chromosomal SNPs for high-resolution paternal lineage classification using targeted semiconductor sequencing. Human Mutation 36 pp 151-159.

RAMANI A, WONG Y, TAN SZ, SHUE BH, SYN C. 2017. Ancestry prediction in Singapore population samples using the Illumina ForenSeq Kit. Forensic Science International Genetics 31 pp 171-179.

RANDO O, VERSTREPEN K. 2017. Timescales of genetic and epigenetic inheritance. Cell. 128 pp 655-668.

REHM H.L. 2013. Disease-targeted sequencing: a cornerstone in the clinic, Nature Reviews Genetics, 14 pp 295-300.

REN ZL, ZHANG JR, ZHANG XM, LIU X, LIN YF, BAI H, WANG MC, CHENG F, LIU JD, LI P, KONG L, BO XC, WANG SQ, NI M, YAN JW. 2021. Forensic nanopore sequencing of STRs and SNPs using Verogen's ForenSeq DNA Signature Prep Kit and MinION. International Journal of Legal Medicine 135 pp 1685-1693

RIMAN S, IYER H, BORSUK LA, VALLONE PM. 2020. Understanding the characteristics of sequence-based single-source DNA profiles. Forensic Science International: Genetics 44 pp 102192.

RONAGHI M, KARAMOHAMED S, PETTERSSON B, UHLEN M, NYREN P. 1996. Real-time DNA sequencing using detection of pyrophosphate release, Analytical Biochemistry, 242 pp 84-89.

ROTHBERG JM, HINZ W, REARICK TM, SCHULTZ J, MILESKI W, DAVEY M, LEAMON JH, JOHNSON K, MILGREW MJ, EDWARDS M, HOON J, SIMONS JF, MARRAN D, MYERS JW, DAVIDSON JF, BRANTING A, NOBILE JR, PUC BP, LIGHT D, CLARK TA, HUBER M, BRANCIFORTE JT, STONER IB, CAWLEY SE, LYONS M, FU Y, HOMER N, SEDOVA M, MIAO X, REED B, SABINA J, FEIERSTEIN E, SCHORN M, ALANJARY M, DIMALANTA E, DRESSMAN D, KASINSKAS R, SOKOLSKY T, FIDANZA JA, NAMSARAEV E, MCKERNAN KJ, WILLIAMS A, ROTH GT, BUSTILLO J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 20.475(7356) pp 348-352.

SANTANGELO R, GONZALEZ-ANDRADE F, BØRSTING C, TORRONI A, PEREIRA V, MORLING N. 2017. Analysis of ancestry informative markers in three main ethnic groups from Ecuador supports a trihybrid origin of Ecuadorians. Forensic Science International: Genetics. 31 pp 29-33.

SANTOS C, PHILLIPS C, FONDEVILA M, DANIEL R, VAN OORSCHOT RAH, BURCHARD EG, SCHANFIELD MS, SOUTO L, UACYISRAEL J, VIA M, CARRACEDO Á, LAREU MV. 2016. Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region. Forensic Science International: Genetics. 20 pp 71-80.

SEO SB, KING JL, WARSHAUER DH, DAVIS CP, GE J, BUDOWLE B. 2013. Single nucleotide polymorphism typing with massively parallel sequencing for human identification. International Journal of Legal Medicine 127 pp 1079-1086.

SHEN H, LI J, ZHANG J, XU C, JIANG Y, WU Z, ZHAO F, LIAO L, CHEN J, LIN Y, TIAN Q, PAPASIAN CJ, DENG HW. 2013. Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty-four Caucasians. PLoS ONE. 8(4). e59494.

SHEWALE JG. 2014. Forensic DNA Analysis: Current Practice and Emerging Technologies. CRC Press. Boca Raton, FL, USA

SHEWALE JG, QI L, CALANDRO LM. 2012. Principles, Practice, and Evolution of Capillary Electrophoresis as a Tool for Forensic DNA Analysis. Forensic Science Review 24 pp 79-100.

SHRIVER MD, SMITH MW, JIN L, MARCINI A, AKEY JM, DEKA R, FERRELL RE. 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. American Journal of Human Genetics. 60 pp 957-964.

SCHRÖDER J, HSU A, BOYLE SE, MACINTYRE G, CMERO M, TOTHILL RW, JOHNSTONE RW, SHACKLETON M, PAPENFUSS AT. 2014. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. Bioinformatics. 30 pp 1064-1072.

SIDSTEDT M, STEFFEN CR, KIESLER KM, VALLONE PM, RÅDSTRÖM P, HEDMAN J. 2019. The impact of common PCR inhibitors on forensic MPS analysis. Forensic Science International: Genetics. 40 pp 182-191.

SILVA DSBS, SAWITZKI FR, SCHEIBLE MKR, BAILEY SF, ALHO CS, FAITH SA. 2018. Genetic analysis of Southern Brazil subjects using the PowerSeq™ AUTO/Y system for short tandem repeat sequencing. Forensic Science International: Genetics 33 pp 129-135.

SILVERY J, GANSCHOW S, WIEGAND P, TIEMANN C. 2020. Developmental validation of the monSTR identity panel, a forensic STR multiplex assay for massively parallel sequencing. Forensic Science International: Genetics. 46 pp 102236.

SIMAYIJIANG H, BØRSTING C, TVEDEBRINK T, MORLING N. 2019. Analysis of Uyghur and Kazakh populations using the Precision ID Ancestry Panel. Forensic Science International: Genetics 43 pp 102144.

STROBL C, CHURCHILL CIHLAR J, LAGACÉ R, WOOTTON S, ROTH C, HUBER N, SCHNALLER L, ZIMMERMANN B, HUBER G, LAY HONG S, MOURA-NETO R, SILVA R, ALSHAMALI F, SOUTO L, ANSLINGER K, EGYED B, JANKOVA-AJANOVSKA R, CASAS-VARGAS A, USAQUÉN W, SILVA D, BARLETTA-CARRILLO C, TINEO DH, VULLO C, WÜRZNER R, XAVIER C, GUSMÃO L, NIEDERSTÄTTER H, BODNER M, BUDOWLE B, PARSON W. 2019. Evaluation of mitogenome sequence concordance, heteroplasmy detection, and haplogrouping in a worldwide lineage study using the Precision ID mtDNA Whole Genome Panel. Forensic Science International: Genetics 42 pp 244-251.

STROBL C, EDUARDOFF M, BUS MM, ALLEN M, PARSON W. 2018. Evaluation of the precision ID whole MtDNA genome panel for forensic analyses. Forensic Science International: Genetics 35 pp 21-25.

STURK-ANDREAGGI K, PARSON W, ALLEN M, MARSHALL C. 2020. Impact of the sequencing method on the detection and interpretation of mitochondrial DNA length heteroplasmy. Forensic Science International: Genetics 44 pp 102205.

SUN S, LIU Y, LI J, YANG Z, WEN D, LIANG W, YAN Y, YU H, CAI J, ZHA L. 2020. Development and application of a nonbinary SNP-based microhaplotype panel for paternity testing involving close relatives. Forensic Science International: Genetics. 46 pp 102255.

TASKER E, LARUE B, BEHEREC C, GANGITANO D, HUGHES-STAMM S. 2017. Analysis of DNA from post-blast pipe bomb fragments for identification and determination of ancestry. Forensic Science International: Genetics. 28 pp 195-202.

THEMUDO GE, MOGENSEN HS, BØRSTING C, MORLING N. 2016. Frequencies of HID-ion ampliseq ancestry panel markers among Greenlanders. Forensic Science International: Genetics 24 pp 60-64.

THERMO FISHER SCIENTIFIC TECHNICAL NOTE: Performance of the Precision ID GlobalFiler NGS STR Panel v2: Artifacts, Thresholds and Chip Loading. Available at www.thermofisher.com

TAO R, QI W, CHEN C, ZHANG J, YANG Z, SONG W, ZHANG S, LI C. 2019. Pilot study for forensic evaluations of the Precision ID GlobalFiler™ NGS STR Panel v2 with the Ion S5™ system. Forensic Science International: Genetics 43 pp 102147.

TSONGALIS GJ, PETERSON JD, DE ABREU FB, TUNKEY CD, GALLAGHER TL, STRAUSBAUGH LD, WELLS WA, AMOS CI. 2014. Routine use of the Ion Torrent AmpliSeq Cancer Hotspot Panel for identification of clinically actionable somatic mutations. Clinical Chemistry and Laboratory Medicine, 52 pp 707-14.

TURCHI C, MELCHIONDA F, PESARESI M, TAGLIABRACCI A. 2019. Evaluation of a microhaplotypes panel for forensic genetics using massive parallel sequencing technology. Forensic Science International: Genetics. 41 pp 120-127.

TVEDEBRINK T, ERIKSEN PS, MOGENSEN HS, MORLING N. 2018. Weight of the evidence of genetic investigations of ancestry informative markers. Theoretical Population Biology 120 pp 1-10.

VAN DER GAAG KJ, DE LEEUW RH, HOOGENBOOM J, PATEL J, STORTS DR, LAROS JFJ, DE KNIJFF P. 2016. Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq™ system. Forensic Science International: Genetics 24 pp 86-96.

VAN DER GAAG KJ, DE LEEUW RH, LAROS JFJ, DEN DUNNEN JT, DE KNIJFF P. 2019. Short hypervariable microhaplotypes: A novel set of very short high discriminating power loci without stutter artefacts. Forensic Science International: Genetics. 35 pp 169-175.

VAN DER HEIJDEN S, DE OLIVEIRA SJ, KAMPMANN ML, BØRSTING C, MORLING N. 2017. Comparison of manual and automated AmpliSeq workflows in the typing of a Somali population with the Precision ID Identity Panel. Forensic Science International: Genetics 31 pp 118-125.

VAN NESTE C, VAN NIEUWERBURGH F, VAN HOOFSTAT D, DEFORCE D. 2012. Forensic STR analysis using massive parallel sequencing. Forensic Science International: Genetics 6 pp 810–818.

VIDAKI A, BALLARD D, ALIFERI A, MILLER TH, BARRON LP, SYNDERCOMBE COURT D. 2017 DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Science International: Genetics. 28 pp 225-236.

VIDAKI A, DANIEL B, SYNDERCOMBE COURT D. 2013. Forensic DNA methylation profiling—Potential opportunities and challenges. Forensic Science International: Genetics 7 pp 499-507.

VILSEN SB, TVEDEBRINK T, MOGENSEN HS, MORLING N. 2017. Statistical modelling of Ion PGM HID STR 10-plex MPS data. Forensic Science International: Genetics 28 pp 82-89.

WAI KT, BARASH M, GUNN P. 2018. Performance of the Early Access AmpliSeq™ Mitochondrial Panel with degraded DNA samples using the Ion Torrent™ platform. Electrophoresis 39 pp 2776-2784.

WALSH S, LIU F, BALLANTYNE KN, VAN OVEN M, LAO O, KAYSER M. 2011. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Science International: Genetics. 5 pp 170-180.

WALSH S, CHAITANYA L, BRESLIN K, MURALIDHARAN C, BRONIKOWSKA A, POSPIECH E,, KOLLER J, KOVATSI L, WOLLSTEIN A, BRANICKI W, LIU F, KAYSER M. 2017. Global skin colour prediction from DNA. Human Genetics 136 pp 847-863.

WALSH S, LIU F, WOLLSTEIN A, KOVATSI L, RALF A, KOSINIAK-KAMYSZ A, BRANICKI W, KAYSER M. 2013. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Science International: Genetics 7 pp 98-115.

WANG L, CHEN M, WU B, LIU YC, ZHANG GF, JIANG L, XU XL, ZHAO XC, JI AQ, YE J. 2018. Massively Parallel Sequencing of Forensic STRs Using the Ion Chef™ and the Ion S5™ XL Systems. Journal of Forensic Science 63 pp 1692-1703.

WANG M, WANG Z, HE G, LIU J, WANG S, QIAN X, LANG M, LI J, XIE M, LI C, HOU Y. 2019. Developmental validation of a custom panel including 165 Y-SNPs for Chinese Y-

chromosomal haplogroups dissection using the Ion S5 XL system. Forensic Science International: Genetics 38 pp 70-76.

WANG M, WANG Z, HE G, WANG S, ZOU X, LIU J, WANG F, YE Z, HOU Y. 2020. Whole mitochondrial genome analysis of highland Tibetan ethnicity using massively parallel sequencing. Forensic Science International: Genetics 44 pp 102197.

WANG Z, HE G, LUO T, ZHAO X, LIU J, WANG M, ZHOU D, CHEN X, LI C, HOU Y. 2018. Massively parallel sequencing of 165 ancestry informative SNPs in two Chinese Tibetan-Burmese minority ethnicities. Forensic Science International: Genetics 34 pp 141-147.

WANG Z, ZHOU D, WANG H, JIA Z, LIU J, QIAN X, LI C, HOU Y. 2017. Massively parallel sequencing of 32 forensic markers using the Precision ID GlobalFiler NGS STR Panel and the Ion PGM System. Forensic Science International: Genetics 31 pp 126-134.

WEBER-LEHMANN J, SCHILLING E, GRADL G, RICHTER DC, WIEHLER J, ROLF B. 2014. Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. Forensic Science International: Genetics 9 pp 42-46.

WENDT FR, KING JL, NOVROSKI NM, CHURCHILL JD, NG J, OLDT RF, MCCULLOH KL, WEISE JA, SMITH DG, KANTHASWAMY S, BUDOWLE B. 2017. Flanking region variation of ForenSeq DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans. Forensic Science International: Genetics 28 pp 146-154.

WENDT FR, WARSHAUER DH, ZENG X, CHURCHILL JD, NOVROSKI NMM, SONG B, KING JL, LARUE BL, BUDOWLE B. 2016. Massively parallel sequencing of 68 insertion/deletion markers identifies novel microhaplotypes for utility in human identity testing. Forensic Science International: Genetics 25 pp 198-209.

WOERNER AE, AMBERS A, WENDT FR, KING JL, MOURA-NETO RS, SILVA R, BUDOWLE B. 2018. Evaluation of the precision ID mtDNA whole genome panel on two massively parallel sequencing systems. Forensic Science International: Genetics 36 pp 213-224.

WOOD MR, STURK-ANDREAGGI K, RING JD, HUBER N, BODNER M, CRAWFORD MH, PARSON W, MARSHALL C. 2019. Resolving mitochondrial haplogroups B2 and B4 with next-generation mitogenome sequencing to distinguish Native American from Asian haplotypes. Forensic Science International: Genetics 43 pp 102143.

WU L, CHU X, ZHENG J, XIAO C, ZHANG Z, HUANG G, LI D, ZHAN J, HUANG D, HU P, XIONG B. 2019. Targeted capture and sequencing of 1245 SNPs for forensic applications. Forensic Science International: Genetics 42 pp 227-234

XAVIER C, PARSON W. 2017. Evaluation of the Illumina ForenSeq DNA Signature Prep Kit – MPS forensic application for the MiSeq FGx benchtop sequencer. Forensic Science International: Genetics. 28 pp 188-194.

YAMAGUCHI-KABATA Y, NAKAZONO K, TAKAHASHI A, SAITO S, HOSONO N, KUBO M, NAKAMURA Y, KAMATANI N. 2008. Japanese population structure, based on SNP

genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. American Journal of Human Genetics 83 pp 445-456.

YANG Y, BINGBING, X, JIANGWEI, Y, 2014. Application of Next-generation Sequencing Technology in Forensic Science. Genomics Proteomics Bioinformatics 12(5). pp 190–197.

YOUNG B, FARIS T, ARMOGIDA L. 2019. A nomenclature for sequence-based forensic DNA analysis. Forensic Science International: Genetics 42 pp 14-20.

YOUNG BA, GETTINGS KB, MCCORD B, VALLONE PM. 2019. Estimating number of contributors in massively parallel sequencing data of STR loci. Forensic Science International: Genetics 38 pp 15-22.

YOUNG B, KING JL, BUDOWLE B,, ARMOGIDA L. 2017. A technique for setting analytical thresholds in massively parallel sequencing-based forensic DNA analysis. PLOS One 2017 12 e0178005.

YOUNG LEE E, YOUNG LEE H, YOON OH S, JUNG SE, SEOK YANG I, LEEYH, ICK YANG W, SHIN KJ. 2016. Massively parallel sequencing of the entire control region and targeted coding region SNPs of degraded mtDNA using a simplified library preparation method. Forensic Science International: Genetics 22 pp 37–43.

YUAN L, CHEN X, LIU Z, LIU Q, SONG A, BAO G, WEI G, ZHANG S, LU J, WU Y. 2020. Identification of the perpetrator among identical twins using next-generation sequencing technology: A case report. Forensic Science International: Genetics 44 pp 102167.

ZBIEC-PIEKARSKA R, SPÓLNICKA M, KUPIEC T, MAKOWSKA Z, PARYS-PROSZEK A, KRZYSZTOF K, ELLIOTT K, PŁOSKI R, BRANICKI W. 2018. Use of Methylation Markers for Age Estimation of an Unknown Individual Based on Biological Traces. Qiagen Technical note. Available from www.qiagen.com.

ZENG X, ELWICK K, MAYES C, TAKAHASHI M, KING JL, GANGITANO D, BUDOWLE B, HUGHES-STAMM S. 2019. Assessment of impact of DNA extraction methods on analysis of human remain samples on massively parallel sequencing success. International Journal of Legal Medicine 133 pp 51-58.

ZENG X, KING J, HERMANSON S, PATEL J, STORTS DR, BUDOWLE B. 2015. An evaluation of the PowerSeq™ Auto System: A multiplex short tandem repeat marker kit compatible with massively parallel sequencing. Forensic Science International: Genetics 19 pp 172-179.

ZENG X, KING JL, STOLJAROVA M, WARSHAUER DH, LARUE BL, SAJANTILA A, PATEL J, STORTS DR, BUDOWLE B. 2015. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Science International: Genetics 16 pp 38–47.

ZHAO X, LI H, WANG Z, MA K, CAO Y, LIU W. 2016. Massively parallel sequencing of 10 autosomal STRs in Chinese using the ion torrent personal genome machine (PGM). Forensic Science International: Genetics 25 pp 34-38.

ZHAO X, MA K, LI H, CAO Y, LIU W, ZHOU H, PING Y. 2015. Multiplex Y-STRs analysis using the ion torrent personal genome machine (PGM). Forensic Science International: Genetics 19 pp 192–196

ZHOU Y, GUO F, YU J, LIU F, ZHAO J, SHEN H, ZHAO B, JIA F, SUN Z, SONG H, JIANG X. 2016. Strategies for complete mitochondrial genome sequencing on Ion Torrent PGM platform in forensic sciences. Forensic Science International: Genetics 22 pp 11-21.

ZHU J, LV M, ZHOU N, CHEN D, JIANG Y, WANG L, HE W, PENG D, LI Z, QU S, WANG Y, WANG H, LUO H, AN G, LIANG W, ZHANG L. 2019. Genotyping polymorphic microhaplotype markers through the Illumina® MiSeq platform for forensics. Forensic Science International: Genetics. 39 pp 1-7.

ZUBAKOV D, KOKMEIJER I, RALF A, RAJAGOPALAN N, CALANDRO L, WOOTTON S, LANGIT R, CHANG C, LAGACE R, KAYSER M. 2015. Towards simultaneous individual and tissue identification: A proof-of-principle study on parallel sequencing of STRs, Amelogenin, and mRNAs with the Ion Torrent PGM. Forensic Science International: Genetics 17 pp 122-128.

Chapter 9:

Appendix

9. Appendix

This appendix contains full locus details of the kits referred to in Chapter 2:

- The GlobalFiler PCR amplification kit (CE)
- The Precision ID GlobalFiler Mixture ID Panel (MPS)
- The Precision ID GlobalFiler STR NGS Panel v2 (MPS)
- The Precision ID GlobalFiler Ancestry Panel (MPS)
- The Precision ID GlobalFiler Identity Panel (MPS)

Table 68: The loci in the GlobalFiler PCR Amplification kit.

Locus name	Locus type	Repeat Structure	Chromosome
Amelogenin	Sex determination	N/A	X / Y
CSF1PO	Simple STR	AGAT	5
D1S1656	Compound STR	TAGA	1
D2S441	Compound STR	TCTA/TCAA	2
D2S1338	Compound STR	TGCC/TTCC	2
D3S1358	Compound STR	TCTA/TCTG	3
D5S818	Simple STR	AGAT	5
D7S820	Simple STR	GATA	7
D8S1179	Compound STR	TCTA/TCTG	8
D10S1248	Simple STR	GGAA	10
D12S391	Compound STR	AGAT/AGAC	12
D13S317	Simple STR	TATC	13
D16S539	Simple STR	GATA	16
D18S51	Simple STR	AGAA	18
D19S433	Compound STR	AAGG/TAGG	19
D21S11	Complex STR	TCTA/TCTG	21
D22S1045	Simple STR	ATT	22
DYS391	Simple STR	TCTA	Y
FGA	Compound STR	CTTT/TTCC	4
SE33	Complex STR	AAAG	6
TH01	Simple STR	TCAT	11
TPOX	Simple STR	AATG	2
vWA	Compound STR	TCTA/TCTG	12
Y-Indel	Sex determination	N/A	Y

Table 69: STR loci in the Precision ID GlobalFiler Mixture ID Panel. This panel contains STR and microhaplotype markers. See Table 70 below for a list of the microhaplotype loci

STR Loci			
Locus name	**Locus type**	**Repeat Structure**	**Chromosome**
AMEL-X	Sex determination	N/A	X
AMEL-Y	Sex determination	N/A	Y
CSF1PO	Simple STR	AGAT	5
D1S1656	Compound STR	TAGA	1
D2S441	Compound STR	TCTA/TCAA	2
D2S1338	Compound STR	TGCC/TTCC	2
D3S1358	Compound STR	TCTA/TCTG	3
D5S818	Simple STR	AGAT	5
D7S820	Simple STR	GATA	7
D8S1179	Compound STR	TCTA/TCTG	8
D10S1248	Simple STR	GGAA	10
D12S391	Compound STR	AGAT/AGAC	12
D13S317	Simple STR	TATC	13
D16S539	Simple STR	GATA	16
D18S51	Simple STR	AGAA	18
D19S433	Compound STR	AAGG/TAGG	19
D21S11	Complex STR	TCTA/TCTG	21
D22S1045	Simple STR	ATT	22
DYS391	Simple STR	TCTA	Y
FGA	Compound STR	CTTT/TTCC	4
TH01	Simple STR	TCAT	11
TPOX	Simple STR	AATG	2
vWA	Compound STR	TCTA/TCTG	12
D1S1677	Simple STR	TTCC	1
D2S1776	Simple STR	AGAT	2
D3S4529	Simple STR	ATCT	3
D4S2408	Simple STR	ATCT	4
D5S2800	Compound STR	GATA/GATT	5
D6S474	Complex STR	GATA/GACA	6
D6S1043	Compound STR	AGAT/AGAC	6

STR Loci			
Locus name	**Locus type**	**Repeat Structure**	**Chromosome**
D12ATA63	Compound STR	TAA/CAA	12
D14S1434	Complex STR	CTGT/CTAT	14

Table 70: Microhaplotype loci in the Precision ID GlobalFiler Mixture ID Panel. This panel contains STR and microhaplotype markers. See Table 69 above for a list of the STR loci.

Microhaplotype Loci		
Locus name	**SNPs that comprise locus**	**Chromosome**
mh01KK-001	rs4648344, rs6663840	1
mh01KK-106	rs12123330, rs16840876, rs56212601, rs4468133	1
mh01KK-205	rs11810587, rs1336130, rs1533623, rs1533622	1
mh01KK-002	rs4528199, rs6604596	1
mh02KK-134	rs3101043, rs3111398, rs72623112	2
mh02KK-136	rs6714835, rs6756898, rs12617010	2
mh03KK-006	rs1919550, rs9873644	3
mh04KK-017	rs4699748, rs2584461, rs1442492	4
mh05KK-170	rs74865590, rs438055, rs370672, rs6555108	5
mh05KK-062	rs870348, rs870347	5
mh09KK-152	rs10867949, rs4282648, rs10780576, rs7046769	9
mh09KK-153	rs10125791, rs2987741, rs7047561	9
mh09KK-157	rs606141, rs8193001, rs56256724, rs633153	9
mh10KK-169	rs10796164, rs10796165, rs17154765, rs10796166	10
mh11KK-180	rs12802112, rs28631755, rs7112918, rs4752777	11
mh11KK-187	rs493442, rs17137917, rs551850, rs17137926	11
mh11KK-191	rs12421109, rs12289401, rs12420819, rs770566	11
mh12KK-202	rs10506052, rs4931233, rs10506053, rs4931234	12
mh12KK-046	rs1503767, rs11068953	12
mh13KK-213	rs8181845, rs679482, rs9510616	13
mh13KK-217	rs7320507, rs9562648, rs9562649, rs2765614	13
mh13KK-218	rs1927847, rs9536429, rs7492234, rs9536430	13
mh15KK-067	rs701463, rs701464	15

Microhaplotype Loci		
Locus name	**SNPs that comprise locus**	**Chromosome**
mh15KK-104	rs11631544, rs10152453, rs80047978	15
mh16KK-049	rs9937467, rs17670098, rs17670111	16
mh16KK-302	rs1395579, rs1395580, rs1395582, rs9939248	16
mh16KK-255	rs16956011, rs3934955, rs3934956, rs4073828	16
mh17KK-272	rs2934897, rs7207239, rs16955257, rs7212184	17
mh18KK-293	rs621320, rs621340, rs678179, rs621766	18
mh19KK-299	rs4932999, rs4932769, rs2361019, rs2860462	19
mh19KK-301	rs10408594, rs11084040, rs10408037, rs8104441	19
mh21KK-316	rs961302, rs17002090, rs961301, rs2830208	21
mh21KK-320	rs2838081, rs2838082, rs78902658, rs2838083	21
mh21KK-324	rs6518223, rs2838868, rs7279250, rs8133697	21
mh22KK-069	rs8137373, rs2235845	22
mh22KK-061	rs763040, rs5764924, rs763041	22

Table 71: The loci in the Precision ID GlobalFiler STR NGS Panel v2.

Locus name	Locus type	Repeat Structure	Chromosome
AMEL-X	Sex determination	N/A	X
AMEL-Y	Sex determination	N/A	Y
CSF1PO	Simple STR	AGAT	5
D1S1656	Compound STR	TAGA	1
D2S441	Compound STR	TCTA/TCAA	2
D2S1338	Compound STR	TGCC/TTCC	2
D3S1358	Compound STR	TCTA/TCTG	3
D5S818	Simple STR	AGAT	5
D7S820	Simple STR	GATA	7
D8S1179	Compound STR	TCTA/TCTG	8
D10S1248	Simple STR	GGAA	10
D12S391	Compound STR	AGAT/AGAC	12
D13S317	Simple STR	TATC	13

Locus name	Locus type	Repeat Structure	Chromosome
D16S539	Simple STR	GATA	16
D18S51	Simple STR	AGAA	18
D19S433	Compound STR	AAGG/TAGG	19
D21S11	Complex STR	TCTA/TCTG	21
D22S1045	Simple STR	ATT	22
DYS391	Simple STR	TCTA	Y
FGA	Compound STR	CTTT/TTCC	4
TH01	Simple STR	TCAT	11
TPOX	Simple STR	AATG	2
vWA	Compound STR	TCTA/TCTG	12
D1S1677	Simple STR	TTCC	1
D2S1776	Simple STR	AGAT	2
D3S4529	Simple STR	ATCT	3
D4S2408	Simple STR	ATCT	4
D5S2800	Compound STR	GATA/GATT	5
D6S474	Complex STR	GATA/GACA	6
D6S1043	Compound STR	AGAT/AGAC	6
D12ATA63	Compound STR	TAA/CAA	12
D14S1434	Complex STR	CTGT/CTAT	14
PENTA D	Simple STR	AAAGA	21
PENTA E	Simple STR	AAAGA	15
Y-Indel	Sex determination	N/A	Y
SRY	Sex determination	N/A	Y

Table 72: The loci in the Precision ID Ancestry Panel. The 'Source' column indicates whether the SNP in question was first proposed in the 'Seldin' panel (Kosoy *et al.* 2009) or the 'Kidd' panel (Kidd *et al.* 2011). More detail of these panels is in Section 1.1.4.1.2. Note that 13 SNPs in the Precision ID Ancestry Panel appear in both the Kidd and Seldin panels.

Locus name	Source	Chromosome
rs10007810	Seldin	4
rs10108270	Seldin	8
rs10236187	Seldin	7
rs1040045	Seldin	6

Locus name	Source	Chromosome
rs1040404	Seldin	1
rs10496971	Seldin	2
rs10497191	Kidd	2
rs10511828	Seldin	9
rs10512572	Seldin	17
rs10513300	Seldin	9
rs1079597	Kidd	11
rs10839880	Seldin	11
rs10954737	Seldin	7
rs11227699	Seldin	11
rs11652805	Kidd and Seldin	17
rs12130799	Seldin	1
rs1229984	Kidd	4
rs12439433	Kidd and Seldin	15
rs12498138	Kidd	3
rs12544346	Seldin	8
rs12629908	Seldin	3
rs12657828	Seldin	5
rs12913832	Kidd	15
rs1296819	Seldin	22
rs1325502	Seldin	1
rs13400937	Seldin	2
rs1369093	Seldin	4
rs1407434	Seldin	1
rs1426654	Kidd	15
rs1462906	Kidd	8
rs1471939	Seldin	8
rs1500127	Seldin	5
rs1503767	Seldin	12
rs1513056	Seldin	12
rs1513181	Seldin	3
rs1569175	Seldin	2
rs1572018	Kidd	13

Locus name	Source	Chromosome
rs16891982	Kidd	5
rs174570	Kidd	11
rs1760921	Seldin	14
rs17642714	Kidd	17
rs1800414	Kidd	15
rs1834619	Kidd	2
rs1837606	Seldin	11
rs1871428	Seldin	6
rs1871534	Kidd	8
rs1876482	Kidd	2
rs1879488	Seldin	17
rs192655	Kidd and Seldin	6
rs1950993	Seldin	14
rs2001907	Seldin	8
rs200354	Kidd and Seldin	14
rs2024566	Kidd	22
rs2030763	Seldin	3
rs2033111	Seldin	17
rs2042762	Kidd	18
rs2070586	Seldin	12
rs2073821	Seldin	9
rs2125345	Seldin	17
rs214678	Seldin	12
rs2166624	Kidd	13
rs2196051	Kidd	8
rs2238151	Kidd	12
rs2306040	Seldin	9
rs2330442	Seldin	7
rs2357442	Seldin	14
rs2416791	Seldin	12
rs2504853	Seldin	6
rs2532060	Seldin	19
rs2593595	Kidd	17

Locus name	Source	Chromosome
rs260690	Kidd and Seldin	2
rs2627037	Seldin	2
rs2702414	Seldin	4
rs2814778	Kidd	1
rs2835370	Seldin	21
rs2899826	Seldin	15
rs2946788	Seldin	11
rs2966849	Seldin	16
rs2986742	Seldin	1
rs310644	Kidd	20
rs3118378	Seldin	1
rs316598	Seldin	5
rs316873	Seldin	1
rs32314	Seldin	7
rs37369	Seldin	5
rs3737576	Kidd and Seldin	1
rs3745099	Seldin	19
rs3784230	Seldin	14
rs3793451	Seldin	9
rs3793791	Seldin	10
rs3811801	Kidd	4
rs3814134	Kidd	9
rs3823159	Kidd	6
rs3827760	Kidd	2
rs385194	Seldin	4
rs3907047	Seldin	20
rs3916235	Kidd	18
rs3943253	Seldin	8
rs4411548	Kidd	17
rs4458655	Seldin	6
rs4463276	Seldin	6
rs4471745	Kidd	17
rs459920	Kidd	16

Locus name	Source	Chromosome
rs4666200	Seldin	2
rs4670767	Seldin	2
rs4717865	Seldin	7
rs4746136	Seldin	10
rs4781011	Seldin	16
rs4798812	Seldin	18
rs4821004	Seldin	22
rs4833103	Kidd	4
rs4880436	Seldin	10
rs4891825	Kidd and Seldin	18
rs4908343	Seldin	1
rs4918664	Kidd	10
rs4918842	Seldin	10
rs4951629	Seldin	1
rs4955316	Seldin	3
rs4984913	Seldin	16
rs5768007	Seldin	22
rs6104567	Seldin	20
rs6422347	Seldin	5
rs6451722	Seldin	5
rs6464211	Seldin	7
rs647325	Seldin	1
rs6541030	Seldin	1
rs6548616	Seldin	3
rs6556352	Seldin	5
rs671	Kidd	12
rs6754311	Kidd	2
rs6990312	Kidd	8
rs705308	Seldin	7
rs7226659	Kidd	18
rs7238445	Seldin	18
rs7251928	Kidd	19
rs731257	Seldin	7

Locus name	Source	Chromosome
rs7326934	Kidd	13
rs734873	Seldin	3
rs735480	Kidd	15
rs7421394	Seldin	2
rs7554936	Kidd and Seldin	1
rs7657799	Kidd and Seldin	4
rs7722456	Kidd	5
rs772262	Seldin	12
rs7745461	Seldin	6
rs7803075	Seldin	7
rs7844723	Seldin	8
rs798443	Kidd and Seldin	2
rs7997709	Kidd and Seldin	13
rs8021730	Seldin	14
rs8035124	Seldin	15
rs8113143	Seldin	19
rs818386	Seldin	16
rs870347	Kidd and Seldin	5
rs874299	Seldin	18
rs881728	Seldin	18
rs917115	Kidd	7
rs9291090	Seldin	4
rs9319336	Seldin	13
rs946918	Seldin	14
rs948028	Seldin	11
rs9522149	Kidd and Seldin	13
rs9530435	Seldin	13
rs9809104	Seldin	3
rs9845457	Seldin	3

Table 73: The loci in the Precision ID Identity Panel. The 'Source' column indicates whether the SNP in question was first proposed in the 'Kidd' panel (Kidd *et al.* 2006) or in the panel developed by the SNPforID project (Musgave-Brown *et al.* 2007). SNPs marked 'Thermo' are Y-chromosome SNPs first proposed by Thermo Fisher Scientific in this panel. More detail of these panels is in Section 1.1.4.1.1 . Note that four SNPs in the Precision ID Identity Panel appear in both the Kidd and SNPforID panels.

Locus name	Source	Chromosome
rs1005533	SNPforID	20
rs10092491	Kidd	8
rs1015250	SNPforID	9
rs1024116	SNPforID	18
rs1028528	SNPforID	22
rs1031825	SNPforID	20
rs10488710	Kidd	11
rs10495407	SNPforID	1
rs1058083	Kidd	13
rs10773760	Kidd	12
rs10776839	Kidd	9
rs1109037	Kidd	2
rs12997453	Kidd	2
rs13218440	Kidd	6
rs1335873	SNPforID	13
rs1355366	SNPforID	3
rs1357617	SNPforID	3
rs1360288	SNPforID	9
rs1382387	SNPforID	16
rs1413212	SNPforID	1
rs1454361	SNPforID	14
rs1463729	SNPforID	9
rs1490413	Kidd and SNPforID	1
rs1493232	SNPforID	18
rs1498553	Kidd	11
rs1523537	Kidd	20
rs1528460	SNPforID	15
rs159606	Kidd	5
rs1736442	Kidd	18

Locus name	Source	Chromosome
rs1821380	Kidd	15
rs1872575	Kidd	3
rs1886510	SNPforID	13
rs1979255	SNPforID	4
rs2016276	SNPforID	15
rs2040411	SNPforID	22
rs2046361	Kidd and SNPforID	4
rs2056277	SNPforID	8
rs2076848	SNPforID	11
rs2111980	SNPforID	12
rs214955	Kidd	6
rs221956	Kidd	21
rs2269355	Kidd	12
rs2292972	Kidd	17
rs2342747	Kidd	16
rs251934	SNPforID	5
rs2830795	SNPforID	21
rs2831700	SNPforID	21
rs321198	Kidd	7
rs338882	Kidd	5
rs354439	SNPforID	13
rs369616152	SNPforID	13
rs372157627	SNPforID	13
rs372687543	SNPforID	13
rs3780962	Kidd	10
rs4288409	Kidd	8
rs430046	Kidd	16
rs4364205	Kidd	3
rs445251	Kidd	20
rs4530059	Kidd	14
rs4847034	Kidd	1
rs560681	Kidd	1

Locus name	Source	Chromosome
rs576261	Kidd	19
rs6444724	Kidd	3
rs6811238	Kidd	4
rs6955448	Kidd	7
rs7041158	Kidd	9
rs717302	SNPforID	5
rs719366	SNPforID	19
rs722098	SNPforID	21
rs722290	Kidd	14
rs727811	SNPforID	6
rs729172	SNPforID	16
rs733164	SNPforID	22
rs735155	SNPforID	10
rs737681	SNPforID	7
rs740598	Kidd	10
rs740910	SNPforID	17
rs7520386	Kidd	1
rs7704770	Kidd	5
rs826472	SNPforID	10
rs873196	SNPforID	14
rs876724	SNPforID	2
rs891700	Kidd and SNPforID	1
rs901398	Kidd and SNPforID	11
rs907100	SNPforID	2
rs914165	SNPforID	21
rs917118	SNPforID	7
rs938283	SNPforID	17
rs964681	SNPforID	10
rs987640	Kidd	22
rs9905977	Kidd	17
rs993934	Kidd	2
rs9951171	Kidd	18

Locus name	Source	Chromosome
rs35284970	Thermo	Y
rs2032599	Thermo	Y
rs2032602	Thermo	Y
rs2534636	Thermo	Y
rs2032631	Thermo	Y
rs2032652	Thermo	Y
rs2033003	Thermo	Y
rs2319818	Thermo	Y
rs3848982	Thermo	Y
rs4141886	Thermo	Y
rs9786139	Thermo	Y
rs16980426	Thermo	Y
rs17222573	Thermo	Y
rs17250845	Thermo	Y
rs17269816	Thermo	Y
rs17306671	Thermo	Y
rs17842518	Thermo	Y
P256	Thermo	Y
rs2032624	Thermo	Y
rs2032636	Thermo	Y
rs8179021	Thermo	Y
rs13447443	Thermo	Y
rs2032673	Thermo	Y
rs9786184	Thermo	Y
rs16981290	Thermo	Y
rs3900	Thermo	Y
rs3911	Thermo	Y
rs2032595	Thermo	Y
rs2032658	Thermo	Y
rs9341278	Thermo	Y
rs20320	Thermo	Y