



The author(s) shown below used Federal funding provided by the U.S. Department of Justice to prepare the following resource:

**Document Title: Development of Reference Sample DNA
 Profiling for Databases Using Next
 Generation Sequencing Technologies**

Author(s): Bruce Budowle

Document Number: 251814

Date Received: July 2018

Award Number: 2012-DN-BX-K033

This resource has not been published by the U.S. Department of Justice. This resource is being made publically available through the Office of Justice Programs' National Criminal Justice Reference Service.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Federal Agency: Office of Justice Programs, National Institute of Justice, Department of Justice

Federal Grant: Award Number 2012-DN-BX-K033

Project Title: Development of Reference Sample DNA Profiling for Databases Using Next Generation Sequencing Technologies

Final Report

PI: Bruce Budowle, University of North Texas Health Science Center at Fort Worth, Institute of Applied Genetics, Department of Molecular and Medical Genetics, Fort Worth, TX 76107, tel: 817-735-2979; email: bruce.budowle@unthsc.edu

Submission Date: February 09, 2015

DUNS Number:


EIN:

Recipient Organization: University of North Texas Health Science Center at Fort Worth, Fort Worth, TX 76107

Institutional Profile Number:

Project grant Period: 10/01/2012-12/31/2014

Signature:



Bruce Budowle, Ph.D.

Project Abstract

Massively parallel sequencing (MPS) technologies provide DNA sequencing data with unprecedented capacity and speed at a reduced cost. These features make the technology desirable for generating DNA profiles that may be uploaded into forensic offender, arrestee, and family reference database files. Because of the exquisitely high throughput, a large battery of genetic markers can be analyzed simultaneously. Indeed, different classes of markers may be analyzed simultaneously and a number of samples can be analyzed at the same time. The studies herein demonstrate autosomal short tandem repeats (STRs), Y-chromosome STRs, X-chromosome STRs, and human identity SNPs can be typed simultaneously. A final panel was designed. A total of 88 STRs (31 autosomal, 26 X-chromosome, 31 Y-chromosome) and 229 autosomal identity SNPs were tabulated including details regarding chromosomal positioning, target selection (Full Region), probe density requirements (due to the alignment-specific requirements of STRs, density of these markers was increased to ‘ADJACENT’) and marker information. Marker data then were uploaded to Design Studio v1.5 and probes were generated under the default conditions (with hg19 for probe reference). Technology advancements using the Nextera Rapid Capture system (Illumina, San Diego, CA) enable typing with 50 ng of input DNA. Probes were designed using Design Studio (Illumina). In addition, instead of sequencing only the hypervariable regions of the mitochondrial genome (mtGenome), the entire molecule can be sequenced with the benefits of increased power of discrimination and more accurate haplogroup assignment. Two MPS platforms were used: the MiSeq (Illumina, San Diego, CA) and the Personal Genome Machine (PGM) (Ion Torrent, ThermoFisher, South San Francisco, CA). Both yielded reliable results. The general procedure, common to all MPS methods was: extraction of DNA, library preparation, sequencing, and data analysis. The amounts of input DNA ranged from 1 ng to 500-1000 ng. The larger amounts of input DNA do not limit the use of MPS for typing reference samples. The lower amounts of input DNA point to the possibility of achieving the same sensitivity of detection of current methods of DNA analysis. A variety of library preparations were tested and all provided results for the various genetic marker systems; each library method had benefits and limitations. Data analysis was demanding and required a number of software tools, including those provided by commercial manufacturers, freeware, and in-house built tools. The software employed (depending on platform and marker) were on-board software (i.e., Real-TimeAnalysis and MiSeq Reporter), Binary Alignment/Map (BAM), Variant Call Format (VCF) v4.1 files, Ion Torrent Software Suite (v 4.0.2) using the plug-in variant caller (v 4.0), Genome Analysis Toolkit (GATK), Integrative Genomics Viewer (IGV), and Haplogrep. Two additional tools that were developed for these studies were STRait Razor and mitoSAVE. The former enables STR allele calls for MPS data that are back compatible with standard STR allele calls. In addition, intra-repeat variation is identified. The latter facilitates haplotype alignment selection to ensure proper nomenclature that meets forensic standards. All goals of the project were met: large multiplex systems were developed (or obtained) and tested for typing reference samples; STRs and SNPs could be typed simultaneously (SNPs also were can be typed in their own multiplexes); and whole mtGenomes could be sequenced with relative ease and at a relatively low cost.

Table of Contents

	Page
Abstract.....	2
Table of Contents.....	3
Executive Summary.....	4
I. Introduction.....	11
a. Project Goals.....	13
II. Materials and Methods.....	14
a. mtGenome Analysis by MPS – the MiSeq Protocol.....	14
b. mitoSAVE: Mitochondrial sequence analysis of variants in Excel.....	18
c. PGM mtGenome Sequencing Protocol and Concordance Testing.....	23
III. Whole Genome mtDNA Sequencing Results and Discussion.....	23
a. MiSeq mtGenome Results.....	24
b. PGM Sequencing Results.....	37
c. mitoSAVE Results.....	40
IV. STR and SNP Panels- Materials and Methods.....	42
a. Selected Markers (Proof of Concept Panel).....	42
V. Software for STR Typing – STRait Razor.....	45
a. Methods for STRait Razor and Data Analysis of STRs from the Large Multiplex Panel.....	48
VI. STR Typing Results and Discussion.....	50
VII. STRait Razor v 2.0.....	55
VIII. Full Panel STR Results from STRait Razor v2.0 Analyses.....	60
IX. SNP Typing with Large Marker Panels by MPS Results and Discussion.....	61
X. PGM SNP Panel - Ion AmpliSeq™ HID SNP panel (v1) Methods.....	72
XI. SNP Panel Assessment Results and Discussion.....	73
XII. Updated Marker Panel.....	78
XIII. Library Preparation Summary.....	92
XIV. Library Preparation Materials and Methods.....	94
XV. TruSeq™ Forensic Amplicon Results and Discussion.....	95
XVI. Final Concluding Remarks.....	97
XVII. References.....	98
XVIII. Dissemination of Research Findings.....	108
XIX. Participating Scientists and Project Collaborations.....	111

Executive Summary

Over the past 25-30 years robust and reliable DNA typing technologies for human identity testing have been implemented (see 1-3 and references within). The technologies enable analyses of minute quantities of DNA and provide a resolving power such that in many cases the number of potential contributors of an evidence sample can be reduced to only a few individuals or a single source. The demands of generating DNA profiles for a national DNA database have fostered developments in automation and robust molecular assays. One particular challenge is the selection of markers that should be used routinely (or for that matter, special case scenarios) by forensic laboratories. There are differences of opinions on how to proceed on core marker selection (10). Additionally, use of core markers, while useful for formalizing a common set for data exchange, inadvertently can limit progress or stifle innovation that may serve well specialized needs of the forensic community. However, these issues can be rendered moot with the advent of massively parallel sequencing (MPS).

The MPS technologies provide DNA sequencing data with unprecedented capacity and speed at a reduced cost that can meet the requirements for uploading DNA profiles into forensic offender, arrestee, and family reference database files. Because of the exquisitely high throughput, a large battery of genetic markers can be analyzed simultaneously. It is entirely possible that all forensically-relevant identified autosomal STRs, such as the 24 STR loci selected by Hares (9) and beyond, a set of Y STRs and X STRs, and human identity SNPs can be typed simultaneously. Moreover, with the high throughput capacity afforded by MPS, many different samples, distinguished by barcoding, may be sequenced simultaneously. SNPs are desirable because they may be typed in degraded samples where STR typing fails; they have a much lower mutation rate; and depending on the SNPs they can provide novel lead information, such as bioancestry and phenotype (3).

Mitochondrial DNA (mtDNA) typing is used by various disciplines (12, 13, 26-35). The higher copy number of mitochondrial genomes (mtGenomes) per cell compared with the nuclear genome makes typing of mtDNA particularly useful for forensic human and species-identity testing, where samples typically are of low quality and contain minute or undetectable amounts of nuclear DNA. With Sanger sequencing (11) the ~16,569 base mtGenome is not feasibly sequenced in a practical manner. Thus, most forensic laboratories focus only on the control region of the mtGenome and, more specifically, hypervariable regions I and II (HVI and HVII) for database construction, database queries, and direct and indirect sample comparisons. However, MPS may make it practical to sequence the entire mtGenome in a rapid and facile manner.

With an increased number of markers and lineage markers, indirect searches can be performed. Familial searching would be highly successful and provide an increased number of investigative leads. The sheer number of markers (and the inclusion of lineage based markers, i.e., Y STRs, Y SNPs, and mtDNA) will provide more robust associations and reduce substantially candidate lists. It is likely that as more kinship associations result in solving crimes there will be motivation to further exploit familial searching with MPS profiling.

With its economies of scale, MPS can provide a system such that reference samples can be typed economically for a large battery of identity markers and the mtGenome. The primary goals of this work were to develop and evaluate MPS systems that can type reference samples for 1) the entire mtGenome; and 2) a large battery of autosomal, Y-chromosome, and X-chromosome STRs and human identity SNPs in a single multiplex analysis.

The overall results support that sequencing of the entire mtGenome from reference samples is feasible by MPS and, in fact, is easier and more cost effective than Sanger sequencing. In addition, haplogroup assignment, useful for quality control and evolutionary studies, is more accurate with sequence data from the entire mtGenome than from sequence data solely from HVI and HVII. The basic methodology is: 1) extraction of DNA from reference; 2) amplification of the entire mtGenome by long PCR by generating two ~8kb long amplicons; 3) library preparation which modifies the amplified DNA so it can be sequenced; 4) sequencing by MPS; 5) raw data analysis; 6) interpretation of results; and 7) population statistical analyses. There were two different MPS platform systems employed for mtGenome sequencing: the MiSeq (Illumina, San Diego, CA) and the Personal Genome Machine (PGM) (Ion Torrent, ThermoFisher, South San Francisco, CA) and reliable results were obtained with both systems. While the general protocols for MPS with each system are similar the chemistries of each are different. For the MiSeq the Nextera XT DNA Sample Preparation Kit (Illumina) was used for library preparation because it requires only 1ng template DNA, can be performed in a relatively short time frame, and multiple samples can be prepared simultaneously. This library preparation protocol exploits “Tagmentation” (53) which combines transposase activity to fragment the DNA into short fragments and adapter ligation in one reaction. For the PGM the amplicons were enzymatically fragmented using Ion Shear™ Plus Reagents (ThermoFisher) and Ion adapters and barcodes were ligated to the fragmented amplicons using the Ion Plus Fragment Library and Ion Xpress™ Barcode Adapters Kits (ThermoFisher). The clonal amplification of DNA fragments on the MiSeq employs bridge amplification, while the PGM uses emulsion PCR. Sequencing strategies are different with the former using terminator chemistry and fluorescent detection and the latter employing detection by a pH change using a semiconductor of an elicited proton during synthesis. Although different, concordant results by orthogonal testing further supported reliability of mtGenome sequencing results.

Data analyses relied on a combination of software tools, including those provided by commercial manufacturers, freeware, and in-house built tools. The software employed were on-board software (i.e., Real-Time Analysis and MiSeq Reporter), Binary Alignment/Map (BAM), Variant Call Format (VCF) v4.1 files, Ion Torrent Software Suite (v 4.0.2) using the plug-in variant caller (v 4.0), Genome Analysis Toolkit (GATK), Integrative Genomics Viewer (IGV), and Haplogrep. The sequenced regions were aligned to the revised Cambridge Reference Sequence (rCRS). Each nucleotide position (np) was interrogated and variations from the reference were annotated by base difference (e.g., 73G). The VCF files were analyzed subsequently using mitoSAVE (57).

mitoSAVE was built in-house to facilitate haplotype alignments to derive reliable and accurate nomenclature. There was a data analysis bottleneck in the process of extracting the information necessary to call mtDNA variants properly. Beyond relatively simple parsing bottlenecks,

“clerical errors” due to alternate alignments of the same string require attention (68). Variants are reported based on alignment software and require some post-processing to comply with current forensic standards. To facilitate semi-automated mtDNA variant designations, mitoSAVE, an Excel-based workbook, evaluates and converts haplotypes to a standardized forensic format. Once familiar with the workflow an analyst using mitoSAVE generated a haplotype from a VCF file in less than one minute per sample. Thus, the automated variant reassignment and haplotype generation allowed for a much faster processing time and higher throughput concomitant with increased sample sequencing of MPS systems. Because accurate haplotypes are reliant on quality sequence data, users can set thresholds, review variants, and generate haplotypes in a more consistent manner than current MPS-related software allow. This level of control promotes accurate haplotype nomenclature and allows consistent haplotypes to be generated by different users.

With the MiSeq system, up to 72 samples could be analyzed simultaneously, with 96 samples being a possibility. Depth of coverage at each mtDNA position was consistent among all samples sequenced. Strand balance was met at the majority of mtDNA sites. **Strand bias is when coverage is notably different between the forward and reverse strand of a targeted sequence.** While some strand bias was observed, it generally was limited to areas of low coverage and did not diminish the ability to assign variant calls. Also, it was possible, due to a high depth of interrogation, to type length and point heteroplasmies. This MiSeq methodology offers a substantial improvement in throughput compared with current Sanger sequencing and (with other similar MPS procedures) given the robustness of sequence results should be considered the method of choice based on quality, cost, and information generating whole mtGenome data. Because of the high throughput of this protocol, whole genome sequences were generated for 283 individuals. It would have been impossible to generate the same amount of data with a Sanger sequencing protocol (routinely used in forensic laboratories) during this phase of the project. The entire sequence data for all individuals were published in King et al (69).

As expected, polymorphism density was clustered heavily in the HVI and HVII regions. However, 74.7% of all variants observed resided outside of the HVI and HVII regions. An increase in random match probability (RMP) was observed from HVI/HVII data of 2.42%, 3.12%, and 3.33% (Table 6) in African American, Caucasian, and Southwest Hispanic populations, with RMPs based on mtGenome sequences of 1.31%, 1.20%, and 0.98%, respectively. Similar patterns were observed for all sample populations with genetic diversity (GD). The GD was 0.987, 0.981, and 0.975 for HVI/HVII compared with 0.998, 1.000, and 0.999 using the mtGenome data for African Americans, Caucasians and Southwest Hispanics, respectively. These findings illustrated the untapped potential of the coding region for discriminatory power. Given the ease of generating sequence data and concomitant high quality results, MPS sequencing of the entire mtGenome should be considered as a viable approach to supplement power of discrimination, when warranted.

Whole mtGenome sequencing also was performed on the PGM to determine its feasibility, accuracy, and reliability. In this study, 24 samples were sequenced, in which 23 were in common with samples sequenced by the MiSeq system. The depth of coverage pattern was similar among all 24 samples and strand bias was limited to a small subset of sites. False deletions can occur with the PGM chemistry and must be understood to acquire reliable data. False deletions may be

due to the limitations of the sequencing chemistry at short to long homopolymer regions and have been observed previously (73). They were measured as a ratio and only 156 positions out of 16,569 positions displayed a false deletion of greater than 0.15 in one or more individuals. These false deletions were associated largely with homopolymers (155/156). All 1237 (SNP) variants (across the 23 mtGenomes) were concordant between the PGM and MiSeq data, excluding the number of Cs in homopolymers around np 310 and 16189 regions. These regions are well known sites for heteroplasmic length variants and typically are not used in forensic identifications (63).

Concordance testing (as above), when feasible, provides information on reliability of results. With current technologies such testing would be a time-consuming, arduous task which is impractical and cost prohibitive. However, previously-generated Sanger sequencing data for HVI and HVII were available for a subset of samples (n=8). All MPS data were concordant at all positions with Sanger sequencing. These data included point and length heteroplasmy, the latter of which was previously difficult to interpret given the nature of Sanger sequencing (and not considered for concordance). Whole genome sequence data for the 9947A cell line were compared and were completely concordant; however some heteroplasmy was observed at some sites consistent with the findings of (71). A 7-sample exchange with Walther Parson's laboratory (Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria) produced concordant results. His data were generated on the PGM MPS system and further supported reliability of both systems (i.e. PGM and MiSeq) by orthogonal testing. Two of the exchanged samples each had an example of point heteroplasmy which was concordant between the MPS methods. Interestingly the quantitative assessment between laboratories and different MPS systems was very similar. For position 195 Y (T/C) in one sample and position 234 R (A/G) in another sample the relative contributions of the heteroplasmic variants between laboratories were 0.71C/0.29T vs 0.67C/0.33T and 0.51A/0.49G vs 0.54A/0.46G, respectively. These concordant results support the reliability of sequence results obtained by the methodologies described herein.

To the best of our knowledge, this was the first reported study of a relatively large number of mtGenomes that have been sequenced in a high-throughput fashion using the Illumina MiSeq system. This study permitted an evaluation of the performance of PGM for mtGenome sequencing and data generated from the PGM were demonstrated to be highly reliable. Therefore, this project demonstrates that whole mtGenome sequencing by MPS is feasible and practical and (with the analytical tools available and developed) is sufficiently robust for use by the forensic community for typing reference samples.

An initial panel of forensically-relevant genetic markers (STRs and SNPs) was selected from the literature (81-92) and from existing commercial STR kits. The markers were a collection of autosomal, X-chromosome and Y-chromosome STRs and human identity and bioancestry SNPs (the bioancestry SNPs will not be part of the final identity panel as they are not as well suited for typical identity testing which require markers of high heterozygosity and low F_{st} ; they were used solely to increase the number of markers for demonstration purposes of high throughput). While the current MPS instruments are capable of providing extensive data, available software tools were limited for identifying forensic STR alleles. Without suitable software for STR analysis the process was tedious and time consuming and comparison of results with current capabilities was difficult. Therefore, this project required the development of STR typing software for MPS data. A novel STR typing software was developed in-house and named STRait Razor (the STR allele

identification tool - Razor). The software is a Linux-based Perl script that identifies alleles at STR loci based on the length of the repeat sequence. This software is capable of handling repeat motifs ranging from simple to complex, and it does not require a reference composed of extensive allelic sequence data. As a result, the allele call results are consistent with those of current CE-based methods, and it is not confounded by unexpected sequence variation within repeats. In its first iteration STRait Razor could identify alleles at 44 forensically-relevant STR loci, and other loci could be configured readily. The details of STRait Razor were described in Warshauer et al (94). This software facilitated data analysis from the initial large panel of STRs.

The results provide support for 1) the functionality and accuracy of STRait Razor for calling STR alleles; and 2) the reliability of STR typing by MPS. The alleles detected by CE methodology were compared with the allele call output files generated by STRait Razor from MPS analyses and were completely concordant. In addition, the STR data comparison revealed the relationship between software, library preparation chemistries, and sequencing platforms used to produce the sequence information. Read length is an important factor (followed by coverage) that impacts STR locus and allele detection of MPS.

The HaloPlex and TruSeq chemistries both provided reliable STR results; however, they have different features that impact STR detection. The haloplex chemistry for library preparation relies on enzymatic cleavage and therefore a benefit is fragments with consistent start and end points are created. The limitation of such a method is that the cleavage sites are based on the restriction endonucleases employed. Depending on the length of the allele in question and the position of the repeat region within the resulting fragment(s), it is possible for sequence reads to be produced that partially span the repeat region. If the sequencing start point of the fragment is too distant from the repeat region, the read may not extend through the entire repeat region of an allele. Also, a very few loci may not be captured. Given that many more STRs can be typed by MPS, one may have to balance forfeiting a couple of current “core” STR loci with the benefit of read depth. The loss of loci can be more than compensated by sheer number of additional markers. The overall practicality of design should be a criterion for long term functionality. In contrast, the TruSeq chemistry is less prone to HaloPlex specific cleavage site issues because DNA is fragmented randomly for a much more varied positioning of repeat regions within the resulting fragments. Therefore, there is a greater likelihood of at least some reads encompassing the entire repeat region at “problematic” loci. The majority of the alleles that were not detected when prepared using HaloPlex were detected with the TruSeq preparation. Despite this beneficial feature, non-enzymatic random fragmentation employed by the TruSeq chemistry resulted in lower read counts for some alleles in comparison with HaloPlex. The random fragmentation method simply may not generate as many fragments that contain the complete repeat region of interest.

Since its initial release, STRait Razor has been employed by a number of laboratories with positive results. In-house needs and resulting feedback were considered strongly to enhance the original software. New features (v2.0) include an expanded default set of detectable STR loci (autosomal, X, and Y markers) that covers all the STR loci in the proposed panel, an enhanced custom locus list configuration tool, a novel output sorting method that highlights unique sequences for each allele, and a genotyping tool that emulates traditional electropherogram data.

With these improvements, STRait Razor v2.0 offers users a much wider, more flexible range of analysis options and greater ease of use.

MPS provides a platform for more comprehensive coverage of genetic markers. There were 379 SNPs in the initial panel. Only 1-4 SNPs per sample failed to yield a result. The SNP rs938283 did not yield a result in any sample and SNP rs9845457 yielded a result in only about half of the samples. Out of all the SNPs, 328 (86.5%) were heterozygous in one or more of the samples. Allele coverage ratios (ACRs) were calculated by dividing the coverage of one allele by the total coverage at that locus (e.g. 450X/970X=46%; 50% indicating equal coverage). The ACR for heterozygous types ranged from 0.460 to 1.00, of which only 8 SNPs displayed an average ACR <0.60. The average depth of coverage per SNP that yielded a result ranged from 6.5X to 564X with only 13 (3.4%) of the SNPs displaying an average depth of coverage <50X.

These data supported that typing reference samples with a large battery of markers is feasible. However, one cannot confirm that all SNPs were typed correctly without an orthogonal approach. Fortunately, a subset of these SNPs (i.e., 95 SNPs) could be compared with the Ion AmpliSeq™ HID SNP panel (v1) and some inference on typing accuracy was obtained. The Ion AmpliSeq™ HID SNP panel (v1), a primer pool of 103 autosomal SNPs and 33 Y-SNPs, was evaluated using the Ion 314™ Chip on the Ion PGM Sequencer with four DNA samples. Genotypes at all SNP loci in the panel were obtained for all samples. Of the 103 autosomal SNPs in the Ion AmpliSeq™ HID SNP panel there were 95 SNPs in common with our in-house panel. All SNP typing results were concordant for the SNPs in common between the two systems, except for SNP rs1029047. This SNP is flanked by homopolymeric stretches, and the SNP states are the same as the homopolymer regions (TTT(T/A)AAAAAAAAA). *A priori* this SNP was suspected of posing a potential typing problem because of the continuum of flanking homopolymers. Operationally, signals generated from homopolymers with the PGM system are not entirely linear, and typing this SNP was problematic with the PGM chemistry.

Overall, the PGM chemistry with its Ion AmpliSeq™ HID SNP panel and the in-house panel with its supporting Illumina system were quite successful in typing SNPs. The data supported that a viable panel of identity SNPs (separately or in concert with STRs) can be analyzed successfully by MPS.

Based on the results described above, a final multiplex STR and SNP identification panel was designed with the Nextera Rapid Capture system (Illumina). Technology advancements suggested that data capture was feasible at a substantially lower quantity of template DNA of 50 ng using the Nextera Rapid Capture system compared with 500 ng to 1 µg with the Illumina® TruSeq™ Custom Enrichment protocol. Probes for the Nextera Rapid Capture Custom Enrichment Kit were designed using Design Studio (Illumina), a freely-available software. A total of 88 STRs (31 autosomal, 26 X-chromosome, 31 Y-chromosome) and 229 autosomal identity SNPs were tabulated including details regarding chromosomal positioning, target selection (Full Region), probe density requirements (due to the alignment-specific requirements of STRs, density of these markers was increased to ‘ADJACENT’) and marker information. Marker data then were uploaded to Design Studio v1.5 and probes were generated under the default conditions (with hg19 for probe reference).

Lastly, throughout this project different enrichment/library preparation methods were considered and tested. Four library preparation strategies have been used, two for the Illumina system, HaloPlex, and a PCR-based one for the PGM/SNP panel and mtGenome sequencing. All library preparations were suitable for the intended purpose. However, some require more template DNA; some are more labor intensive; and some may not be compatible with a very few markers. The amount of initial template was not considered a limitation as the methods herein were developed for reference sample typing. The Illumina® TruSeq™ Custom Enrichment protocol, based on a capture strategy, can target a large number of target sites simultaneously. This library preparation protocol was selected initially because PCR amplification for target enrichment was not required. Therefore, a challenging PCR multiplex primer design would not have to be accomplished and errors due to the PCR would not impact sequencing results. However, it is a laborious method. HaloPlex also is a capture-based approach. It has benefits of known start and stop points for the DNA being sequenced, higher coverage, and high sample throughput. But a small number of loci may not be compatible with the restriction enzyme cocktail that is used. For the mtGenome sequencing and Ion AmpliSeq™ HID SNP panel PCR enrichment was employed. The features that make PCR enrichment approaches desirable are: 1) only 1ng template DNA is required; 2) the process can be performed in a relatively short time frame; and 3) multiple samples can be prepared simultaneously.

In this project, another approach, the TruSeq™ ChIP protocol (Illumina), was modified to enable library preparation of forensically-relevant SNP-containing amplicons. This protocol, known as TruSeq™ Forensic Amplicon, was used to detect a battery of 160 human identification SNPs (HIDs) and AIMs in a set of 12 reference samples. SNP genotypes were obtained for all 160 SNPs in 11 of the 12 samples analyzed. In one sample, only one SNP was not called due to low coverage. Sequence coverage and heterozygote allele balance were comparable with other PCR enrichment systems. This method appears to be less labor-intensive than alternative techniques. Additionally, the TruSeq™ Forensic Amplicon library preparation method is highly sensitive; 0.5 ng input DNA was used in this study. In conjunction with a properly designed multiplex PCR, this preparation method is capable of producing sequencing results with relatively even allele balance at heterozygous loci. The results of this proof-of-concept preparation method suggested that this novel use of the original TruSeq™ ChIP protocol could support forensic marker typing by MPS.

In conclusion, the goals of the project were met. Large multiplex systems were developed (or obtained) and tested for typing reference samples. STRs and SNPs could be typed simultaneously. SNPs also were can be typed in their own multiplex. Whole mtGenomes could be sequenced with relative ease. The data support that reliable results can be obtained. To facilitate analyses software was developed. STRait Razor (v1.0 and v2.0) for STR typing and mitoSAVE for haplotype alignment/nomenclature have been created and are freely available. The protocols described within the final report and published in the scientific literature should enable novel users to perform MPS in their respective laboratories.

I. Introduction

Over the past 25-30 years various robust and reliable DNA typing technologies for human identity testing have been implemented (see 1-3 and references within). The technologies enable analyses of minute quantities of DNA and provide a resolving power such that in many cases the number of potential contributors of an evidence sample can be reduced to only a few individuals, if not only one source. The success of DNA typing has led to further applications; one notable use has been developing investigative leads. The potential of DNA typing for developing investigative leads and for solving future crimes came to fruition with the development of DNA databases. Many countries have established DNA databanks that contain DNA profiles from convicted offenders, arrestees and forensic samples from unsolved cases (4, 5). These databases are designed to associate DNA profiles from individuals with those derived from forensic samples or to identify missing persons. The U.S. databank - COmbined DNA Index System (CODIS) - houses more than 10,971,392 offender profiles, 1,892,952 arrestee profiles and 559,705 forensic profiles as of May 2014 and is relied upon routinely for helping to develop meaningful investigative leads. Because of their success, these DNA databases continue to increase in size and may generate additional information other than solely direct matching of DNA profiles for investigative leads, such as that which can be obtained by familial searching (6-8).

The demands of generating, entering, and maintaining DNA profiles in a national DNA database have fostered developments in automation and robust molecular assays. The number of reference samples from convicted felons, arrestees, detainees, and missing persons continues to increase and there is no indication of the demand subsiding. To meet the needs of forensic DNA typing and its infrastructure it is incumbent on forensic scientists to be vigilant and embrace new technologies that will benefit the process, as well as society, by being able to analyze ever increasing numbers of reference samples, to address more challenging samples, to continue to exonerate the innocent, to enhance abilities to solve crime and to identify missing persons.

One particular challenge is the selection of markers that should be used routinely (or for that matter, special case scenarios) by forensic laboratories. To be able to share and compare DNA results a core set of short tandem repeat (STR), or microsatellite, loci was selected sixteen years ago (4,5). Recently, Hares (9), representing the FBI, recommended that the core 13 STR loci for CODIS should be changed and augmented. The FBI advocated 20 STR loci (24 total if a second panel of four additional STRs is considered) to serve as the new CODIS core markers. Ge et al (8) suggested that there were additional factors and applications beyond that which Hares (9) relied upon for selecting a core set of markers. This alternate viewpoint was that the loci selected should be driven by the demands of casework, i.e., loci should be selected based on performance with degraded and inhibited samples or that the markers selected might have been more versatile to enable a variety of search strategies. Thus, there are differences of opinions on how to proceed on core marker selection (10). Additionally, use of core markers, while useful for formalizing a common set for data exchange, inadvertently can limit progress or stifle innovation for alternate markers that may serve well specialized needs of the forensic community. However, these discussions on a fixed core set of loci and unintentional stymied growth of novel marker sets can

be rendered moot with the advent of massively parallel sequencing (MPS), also termed next generation sequencing.

The MPS technologies provide DNA sequencing data with unprecedented capacity and speed at a reduced cost. Sequencing for the past few decades has primarily been performed by Sanger sequencing (11). While Sanger sequencing is robust and used particularly in forensics for mitochondrial DNA (mtDNA) sequencing (12, 13) and some SNP-based assays (14, 15), it is labor intensive, has a relatively low throughput, and is costly on a per nucleotide basis. In contrast, MPS technologies sequence DNA in a highly parallel fashion with high coverage, which result in high throughput of specified targets with potentially low error. In fact, whole human genomes have been sequenced with costs dropping dramatically to \$1000 or less. Typically, a Sanger-sequenced mtDNA targeted region for forensic applications provides a 1X coverage for each strand or it can be considered 2X if the complementarity of the two strands is used for confirmation and accuracy. In contrast, MPS technology can provide 100s to 1000s fold coverage for the same target region and not even begin to exploit the full throughput of the systems (16-24). These different coverage features between technologies should not be construed as a sensitivity of detection difference, but as a molecule interrogation difference. The developments of MPS, in recent years, have made the technology sufficiently robust for typing reference samples that can meet the requirements for uploading DNA profiles into forensic offender, arrestee, and family reference database files. Because of the exquisitely high throughput, a large battery of genetic markers can be analyzed simultaneously, far exceeding the current capacity of 15-27 STRs of a fluorescent multiplex/capillary electrophoresis (CE) system. It is entirely possible that all forensically-relevant identified autosomal STRs, such as the 24 STR loci selected by Hares (9) and beyond, a set of Y STRs and X STRs, and human identity SNPs (comprising hundreds of markers) can be typed simultaneously. Moreover, with the high throughput capacity afforded by MPS, many different samples which can be distinguished by barcoding may be sequenced simultaneously. In theory, hundreds to thousands of barcodes could be synthesized, but currently 12 to 384 different reference samples could be coded at one time (25). SNPs are desirable because they may be typed in degraded samples where STR typing fails, they have a much lower mutation rate, and depending on the SNPs they can provide novel lead information, such as bioancestry and phenotype (3).

Mitochondrial DNA (mtDNA) typing is used by various disciplines, such as forensic genetics (12,13, 26-28), medical genetics (29-31), genealogy and evolutionary anthropology (32-35). The higher copy number of mitochondrial genomes (mtGenomes) per cell compared with the nuclear genome makes typing of mtDNA particularly useful for forensic human and species-identity testing, and ancient DNA analyses, where samples typically are of low quality and contain minute or undetectable amounts of nuclear DNA. With Sanger sequencing (11) the ~16,569 base mtGenome is not feasibly sequenced in a practical manner in an application-oriented laboratory. Thus, most forensic laboratories focus only on the control region (CR) of the mtGenome and, more specifically, hypervariable regions I and II (HVI and HVII) for database construction, database queries, and direct and indirect sample comparisons.

Current mtDNA databases allow for haplotype searching (36-41) as well as variant-specific queries (40,42). To date, forensic databases contain limited, if any, coding region data. mtGenome data provide greater discriminatory power and allow resolution of common

HVI/HVII haplotypes (43-46). Though not routinely performed in forensic casework, haplogroup assignments allow analysts a measure of data quality control (47,48). Haplogroup assignments can be performed manually using Phylotree (34) or with haplogroup-assignment software (37,39,49-51). Regardless, the accuracy of a haplogroup assignment is reliant on the genetic data used (e.g., CR vs. mtGenome). MPS could make it feasible to sequence the entire mtGenome, thus increasing discrimination power and haplogroup assignment accuracy.

The inclusion of a more comprehensive set of markers for reference samples will overlap all current databases and foster investigations. Thus, all STR and mtDNA legacy data in forensic databases can be compared with MPS data. With MPS generated reference sample data, extant genetic marker data from evidence samples can be compared among the majority (if not all) of the reference DNA profiles in databases worldwide that contain a more limited set of marker sets.

With an increased number of markers and lineage markers that can be included in the set, indirect searches can be performed. Familial searching would be highly successful and provide an increased number of investigative leads. The sheer number of markers (and the inclusion of lineage based markers, i.e., Y STRs, Y SNPs, and mtDNA) will provide more robust associations and substantially reduce candidate lists. It is likely that as more kinship associations result in solving crimes there will be motivation to further exploit familial searching with MPS profiling.

With its economies of scale, MPS can provide a system such that reference samples can be typed economically for a large battery of identity markers and the whole genome of mtDNA. The latter shall be sequenced separately due to its much higher copy number. Eventually, if commercialized, MPS systems could provide a notable cost benefit compared with current costs for typing a modicum of autosomal STRs. The primary goals of this work were to develop and evaluate MPS systems that can type reference samples for 1) a large battery of autosomal, Y chromosome, and X chromosome STRs and human identity SNPs in a single multiplex analysis; and 2) the entire genome of mtDNA.

1. Project Goals

1. Select and assess strategies for amplifying and enriching mtDNA amid the background of nuclear DNA to prepare for sequencing. The approach herein employs long PCR to generate two 8 kb amplicons of the mtDNA genome;
2. Select and finalize a candidate panel of STRs (autosomal, Y chromosome and X chromosome) and SNPs suitable for human identity testing based on those forensic markers used worldwide, the scientific literature, and previous work supported by NIJ;
3. Select, develop, and/or evaluate library generation strategies to facilitate sample preparation. The strategy should be commensurate with the MPS platform system and if at all possible reduce the substantial labor associated with generating libraries;

4. Based on the outcomes from goals 1-3, test capability and determine the limitations of the designed MPS assays on the available platforms for profiling individuals for the specified forensically-relevant genetic marker systems.

The research described herein is divided into two sections. The first section describes the methods and results for mtGenome sequencing. The second section addresses the work of nuclear markers, both STRs and SNPs.

II. Materials and Methods

1. mtGenome Analysis by MPS – the MiSeq Protocol

In this section, the methodology, output results, overall performance, and findings on mtGenome sequencing are presented. To facilitate describing the results some discussion is inserted where warranted, as opposed to only in the Conclusion of the section. The overall results support that sequencing of the entire mitochondrial genome from reference samples is feasible by MPS and, in fact, is easier and more cost effective than Sanger sequencing. Lastly, haplogroup assignment is more accurate with sequence data from the entire ntGenome than from sequence data solely from HVI and HVII, which in turn can improve quality control for forensic analyses and better elucidate evolutionary studies.

DNA Extraction

Whole blood samples were collected by venipuncture according to protocols approved by the University of North Texas Health Science Center's Institutional Review Board. DNA was extracted from these samples using the QIAamp® DNA Blood Mini Kit (QIAGEN, Hilden, Germany) according to the manufacturer's recommendations. The quantity of DNA obtained from extraction was determined using the Qubit dsDNA BR Quantification Kit and a Qubit spectrofluorometer (ThermoFisher, South San Francisco, CA). Samples were normalized to 0.1ng/μL of DNA with molecular grade water and stored at either 4°C or -20°C.

Long PCR Amplification of Whole MtGenome DNA

Amplification of the entire mtGenome was performed by long PCR in two separate PCRs. The primers for each reaction were described previously by Gunnarsdóttir et al. (52) and are listed in Table 1. The TaKaRa LA PCR Kit (TaKaRa Bio; Otsu, Shiga, Japan) was used for long-range PCR amplification. The long PCR master mix is shown in Table 2.

Table 1. Long PCR Primers

F1: 5'- ggc atc tac ggc tca aca tt -3'

R1: 5'- ttg gct ctc ctt gca aag tt -3'

F2: 5'- tat ccg cca tcc cat aca tt -3'

R2: 5'- gtg gcc ttg gta tgt gct tt -3'

Table 2. Long PCR Components for mtDNA amplification

6.25 μ L of Nuclease-free Water
2.5 μ L of 10X TaKaRa LA Buffer
4.0 μ L of 2.5mM dNTPs
1 μ L each of 10 μ M forward and reverse primers
0.25 μ L of 5U/ μ L TaKaRa LA Taq, 5 Units/ μ L
10 μ L of template DNA

The total template DNA (i.e., based on genomic DNA measurement) was 1.0 ng per reaction. Amplification was performed on a GeneAmp 9700 thermocycler (ThermoFisher) using the following thermocycling parameters: an initial temperature of 94°C for one minute; followed by thirty-five cycles 98°C for ten seconds, 60 °C for two minutes, and 68°C for ten minutes. After cycling there was a final extension step of 72°C for ten minutes. The amplified product was maintained at 4°C.

Amplicon Pooling

Two separate PCRs were performed to achieve amplification of the mtGenome. Therefore, the two amplicons were combined back into one sample. First, the quantity of the amplicon products was determined using the Qubit dsDNA BR kit ((ThermoFisher) and then the quantities were normalized to 0.2 ng/ μ L. Second, the size of amplicons was verified to be ~8.3 and 8.6 kb using the Agilent High Sensitivity DNA Kit and Bioanalyzer 2100 (Agilent Technologies; Santa Clara, CA). This quality check ensures that the amplicons were the correct length before proceeding. Then the two amplicons per sample were pooled in equimolar amounts (quantity determined using the Qubit) to a final volume of 5 μ L. Once a pooled set of amplicons was generated, the samples were ready for library preparation.

Library Preparation

The Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, CA) was selected for library preparation because it requires only 1ng template DNA, can be performed in a relatively short time frame, and multiple samples can be prepared simultaneously. This library preparation protocol exploits “Tagmentation” (53) which combines transposase activity to fragment the DNA and adapter ligation in one reaction. For tagmentation, 10 μ L of Tagmentation DNA Buffer (TD), 5 μ L of the Amplicon Tagmentation Mix (ATM), and 5 μ L pooled template were mixed via pipetting up and down five times. The resulting tagmentation mix then was centrifuged at 280 x g for one minute at room temperature. The samples then were placed for five minutes onto a thermocycler preheated to 55°C. Following this step, 5 μ L of the Neutralize Tagmentation Buffer (NT) were added immediately to each sample. Then, the whole volume was mixed by pipetting up and down five times, followed by centrifugation, and allowed to incubate at room temperature for five minutes.

The Nextera XT system employs a dual index system (i.e., barcoding both ends of a fragment) to enable identification of different samples that are multiplexed for sequencing. Indices are 6 bases long and anneal to a 5' overhang to the tags added during the tagmentation reaction and then are ligated, in theory, to all fragments of target DNA. These indices are unique in their sequence and allow for differentiation of 96 samples (by combinations of eight separate 500 series indices with

twelve separate 700 series indices). The library PCR reagents were added directly to the sample tube from the tagmentation reaction at the following quantities: 5µL each of the sample specific 500 series Index and 700 series Index (which was then mixed by pipetting up and down five times), and then 15µL of Nextera PCR Master Mix (NPM) were added which was in turn mixed by pipetting up and down three times. This mixture was centrifuged at 280 x g for one minute at room temperature. Subsequently, the reaction was amplified with the following conditions: Two successive single cycle steps of 72°C for three minutes and 95°C for 30 seconds followed by twelve cycles of 95°C for ten seconds, 55°C 30 seconds, and 72°C for 30 seconds. A final extension step of 72°C for five minutes was performed. The plate can either be removed immediately, or maintained on the thermocycler overnight at 10°C.

In the same sample well or tube, 50 µL of Agencourt AMPure XP beads (Beckman-Coulter Indianapolis, Indiana, USA) were added to each sample well or tube and mixed by gently pipetting up and down ten times. Samples then were allowed to stand at room temperature for five minutes. Following this step, the samples were placed on a magnet for two minutes, and then 90 µL of the supernatant were removed and discarded. The beads, due to association with the magnet, remained in the tubes/wells. The beads then were washed two times with 200 µL of freshly prepared 80% ethanol. The bead preparations were allowed to air dry for 15 minutes at room temperature, and then were resuspended in 52.5µL of the Nextera XT Resuspension Buffer (RSB). The Buffer/bead mixture was mixed ten times by pipetting up and down, and the slurry was allowed to stand at room temperature for two minutes. The samples were placed on a magnet for two minutes. Once the beads were pulled out of suspension by the magnet, 50 µL of supernatant were transferred to a new tube/plate for further processing. Care was taken not to disturb the beads during this step to minimize bead carryover.

Following the PCR clean-up, the libraries were quantified using the Qubit dsDNA BR kit, and evaluated for fragment size using the High Sensitivity D1K ScreenTape and Tape Station 2200 (Agilent Technologies). Based on Illumina’s technical note for Cluster optimization (54) and the resultant size and quantity data, libraries of each sample to be multiplexed for sequencing were normalized to 2 nM and pooled in an equimolar fashion into a single tube for a final volume of 600 µl. The pooled libraries were mixed briefly by pulse vortexing and briefly subjected to centrifugation. This pooled sample (14 µL) and 2 µL of 2nM PhiX DNA (diluted in RSB) were combined in a new tube, briefly vortexed, and briefly subjected to centrifugation. To denature the DNA 10 µL of this resultant pool were added to 10µL of freshly prepared 0.1 N sodium hydroxide, briefly vortexed, briefly subjected to centrifugation, and allowed to stand at room temperature for five minutes. Subsequently, 980 µL of pre-chilled HT1 buffer were added to bring the total library concentration to 20 pM. A final dilution step was performed where 600 µL of the 20 pM library were combined with an additional 400 µL of cold HT1 Buffer for a final concentration of 12 pM. The basic steps and approximate time required are summarized in Table 3.

Table 3. Summary of Library and Sequencing Steps with Approximate Time Requirements

Steps	Time Required
Long PCR	~7 hours
Amplicon Pooling	~1 hour

Tagmentation	~1.5 hours
Library PCR	~1 hour
PCR Cleanup	~1.5-2 hours
Library Pooling	~1.5 hours
Sequencing	~39 hours

Sequencing and Raw Data Processing

The MiSeq (Illumina) re-sequencing protocol for small genome sequencing was followed according to the manufacturer's recommendations. Sequencing entailed: thawing of the reagent cartridge, cleaning and insertion of a new flow cell, the addition of sample to the sample well in the reagent cartridge, and the proper insertion of the reagent cartridge, waste reservoir, and accompanying reagent buffer reservoir. Sequencing reactions were carried out using the MiSeq v2 (2 x 250 bp and 2 x 150 bp) chemistries (Illumina). Sequencing proceeded on a MiSeq platform in an automated fashion for ~39 hours. On-board software (i.e., Real-Time Analysis and MiSeq Reporter) converted raw data to Binary Alignment/Map (BAM) and Variant Call Format (VCF) v4.1 files using Genome Analysis Toolkit (GATK) (55). The sequenced regions were aligned to the revised Cambridge Reference Sequence (rCRS) (56). Each nucleotide position (np) was interrogated and variations from the reference were annotated by base difference (e.g., 73G). These VCF files were analyzed subsequently using mitoSAVE (57).

Data Analysis

Software for data analysis and the flow of use of software (i.e., pipeline) are listed in Table 4.

Table 4. Summary of Data Analysis Steps in Order of Processing

FASTQ file generated from Miseq raw data
 Generate SAI file (BWA)
 Generate SAM (BWA)
 Convert to BAM (SAM Tools)
 Sort BAM File (SAM Tools)
 Index BAM File (SAM Tools)
 Variant Calling (GATK)
 In-house Work Book Analyses Tailored to Task

VCF files were compared initially with BAM files in Integrative Genomics Viewer (IGV) (58) to ensure that all variants have been called according to conventions established in the forensic community (5). Following this step, the VCF files were converted via in-house software (i.e., mitoSAVE (57)) to a format that was amenable to Haplogrep (34,62) for genome analyses. Haplogrep analyzes mtDNA sequence data in a phylogenetic manner. Variants not known to be associated with a haplogroup (local private mutations), not previously observed in the database (global private mutations), or variants expected, but not observed, for each haplotype were verified by manually viewing BAM files in IGV. Random match probability (RMP) and Genetic

diversity (GD) were calculated according to methods described by Stoneking *et al.* (59) and Tajima (60), respectively.

2. mitoSAVE: Mitochondrial sequence analysis of variants in Excel A Variant Caller File (VCF) Conversion Tool

The current mtDNA data analysis pipeline consists of taking FASTQ file format reads, performing alignment of reads against the rCRS using BWA (61) and subsequent calling variants that differ from the rCRS using GATK (Table 4). Various pipelines use a similar approach and many use these same software within their pipelines. GATK generates a VCF file for each sample that consists of a row of data for each nucleotide position aligned. The output describes position, rCRS allele, alternate allele (if applicable), quality score, and two information strings.

Regardless of the region of interest, inconsistency in haplotype assignment persists across all disciplines. Attempts to standardize nomenclature have been met with varying success (63-67). Sequence data are not reported currently in string format, but rather are listed as variants from the rCRS (56). This manner of nomenclature creates a shorthand haplotype that facilitates communication, can be stored in various databases and queried as needed. Guidelines have been produced for consistent interpretation of sequence data by applying a rule-based “least number of differences” approach to sequence analysis while at the same time, known patterns of polymorphisms based on previously-described phylogenetic structure frame possible alignments (63,65,67,68). Haplotypes generated using these guidelines should align with sequences in forensic databases such as EMPOP (36) and Phylotree (34).

One advantage of phylogenetic evaluation of mtDNA sequence data in identity testing has been its use as a means of quality control of the data (47,48). By evaluating haplogroup assignments, an analyst may identify potential errors *a posteriori*. These assignments may be done manually or using software applications (49,51,62) with debatable success (48). Software applications, such as HaploGrep, have proven to be successful and allow haplogroup generation for thousands of samples at a time making it ideal for analysis of large population data sets.

However, there is a data analysis bottleneck in the process of extracting the information necessary to call mtDNA variants properly. The information string contains critical information about each nucleotide position (e.g. allelic depth of coverage, genotype, phred-scaled genotype likelihood, etc.) and is configurable in BWA. However, the information is in string format delimited with colons and thus difficult to analyze in a time effective manner.

Beyond relatively simple parsing bottlenecks, clerical errors can create alternate alignments of the same string and require attention (68). Variants are reported based on alignment software and require some post-processing to comply with current forensic standards. Designated variants often require realignment in areas of length heteroplasmy (i.e., homopolymeric stretches) and areas of repeats (e.g., HVIII AC stretch; np 8272-8289) to allow more accurate and consistent haplotype nomenclature. To facilitate semi-automated mtDNA variant designations, mitoSAVE was developed for haplotype evaluation and conversion to a standardized forensic format.

mitoSAVE is an Excel-based workbook that provides users a tool to analyze mtGenome VCF files in a semi-automated fashion in an expeditious manner.

Data Collection

Samples used for development and evaluation of mitoSAVE were from VCF files obtained from 325 indexed samples (283 different individuals) of whole mtDNA genomes described in King et al. (69). Reads were aligned to the rCRS with BWA, and VCF files were generated using GATK with no downsampling.

VCF Format

GATK allows for multiple options when creating a VCF file. Such data were annotated in the column labeled FORMAT directly preceding the GENOTYPE column and were listed for each position annotated in the VCF (either all positions or variants with respect to the rCRS depending on user preference) in the column labeled GENOTYPE. Data were colon-delimited and easily parsed. mitoSAVE uses the following genotype information for analysis: genotype, allelic depth, read depth, and genotype quality (Figure 1). The sub-fields are required in this order for proper parsing and subsequent data analysis.

Figure 1. User interface overview.

B	C	D	U	V	W	Z	AA	AB	AC	AD	AE	AH	AJ	CPCQ	CR	CS
1																
2				Converted VCF Calls												
4	Quality Thresh (GATK)		Review	Chrom:Co	Manual Call	Coverage	Ref	Ref Cov	Alt	Alt Cov			Variants Observed			
6	70			chrM.73	73G	1416	A	21	G	1393						
7				chrM.146	146C	1719	T	1	C	1713			73G			Generate Sample Report
8	Heteroplasmy Thresh			chrM.152	152C	1678	T	18	C	1660			146C			
9	0.18			chrM.263	263G	654	A	2	G	650			152C			
10				chrM.302	309.1CC#	371	A	179	ACC,AC	122,66			263G			
11	Coverage Thresh			chrM.309		381							309.1CC#			
12	40			chrM.310		375							315.1C			
13				chrM.310	315.1C	383	T	158	TC	198			523del_524del			
14	Coverage Screen			chrM.469		1356							750G			
15	200			chrM.513	523del_524del	2187	GCA	818	GCCCA,G	0,6			1438G			
16				chrM.750	750G	4944	A	4	G	4938			1811G			
17	Sample ID			chrM.1438	1438G	4903	A	11	G	4883			2217T			
18	19502			chrM.1811	1811G	4313	A	50	G	4257			2706G			
19				chrM.2217	2217T	4126	C	72	T	4022			3480G			
20	Shortcuts for Review			chrM.2706	2706G	4356	A	62	G	4217			3995G			
21				chrM.3106		4493							4769G			
22	A-Accept			chrM.3480	3480G	1972	A	133	G	1825			5231A			
23	R-Remove			chrM.3995	3995G	4994	A	46	G	4935			7028T			
24	S-Show Seq			chrM.4769	4769G	3925	A	7	G	3910			8860G			
25	L-Mark LHP			chrM.5231	5231A	3950	G	53	A	3878			9055A			
26				chrM.7028	7028T	4623	C	63	T	4546			9698C			
27				chrM.8860	8860G	4798	A	12	G	4775			9716C			
28	Target			chrM.9055	9055A	3356	G	51	A	3281			9972G			
29	Whole Genome			chrM.9698	9698C	2063	T	34	C	2024			10550G			
30				chrM.9716	9716C	2054	T	30	C	2020			11084G			
31				chrM.9972	9972G	1907	A	19	G	1886			11299C			
32				chrM.10550	10550G	2658	A	17	G	2631			11467G			
33	Compile Variant List			chrM.11084	11084G	2525	A	29	G	2485			11719A			
34				chrM.11299	11299C	3402	T	38	C	3360			11869A			
35				chrM.11467	11467G	3611	A	21	G	3581			12308G			
36				chrM.11719	11719A	3716	G	44	A	3634			12372A			
37				chrM.11869	11869A	3690	C	62	A	3576			14037G			
38				chrM.12308	12308G	2685	A	35	G	2643			14167T			
39				chrM.12372	12372A	2505	G	28	A	2475			14502C			
40	Haplogrep Export			chrM.14037	14037G	2951	A	53	G	2890			14766T			
41				chrM.14167	14167T	2958	C	48	T	2905			14798C			
42				chrM.14502	14502C	2536	T	29	C	2496			15326G			
43				chrM.14766	14766T	2654	C	68	T	2570			16224C			
44				chrM.14798	14798C	2711	T	37	C	2667			16311C			
45				chrM.15326	15326G	3568	A	13	G	3544			16342C			
46				chrM.16224	16224C	4393	T	48	C	4342			16519C			
47				chrM.16311	16311C	4646	T	38	C	4605						
48				chrM.16342	16342C	4431	T	37	C	4386						
49				chrM.16519	16519C	1554	T	2	C	1551						
50																
51																
52																

Data Interpretation

The VCF files were pasted into mitoSAVE. Each mtDNA np was defined using a combination of phylogenetic and parsimony rule-based “least number of differences” approaches. Each read was aligned initially with the rCRS. Since there were times that multiple slightly-different alignments were possible for certain regions (i.e., homopolymeric stretches) (65,67), the alignment was called parsimoniously initially. Next, a correction based on phylogenetically-established variants was applied to some nps to maintain established descriptions of known patterns of polymorphism. A sheet (‘Watchlist’) within mitoSAVE allowed for positions to be predefined according to well-defined structures or scenarios and can be configured by the user. This ‘Watchlist’ currently encompasses the variants observed in our dataset and should not be

used in place of haplogrouping-software. However, this functionality enables mitoSAVE to be modified based on the data observed and processed by users and will be especially useful as experience with sequencing of whole mtDNA genomes reaches a level that is consistent with that of current forensic sequence analysis of the mitochondrial genome non-coding region.

The data then were transferred automatically to a new tab for final review (Figure 2). In this tab, the user can select the target area for which the final haplotype (based on extant data) is generated (i.e., HVI/HVII, mtGenome, or portion thereof). Thresholds and allowances were set for data interpretation. Quality scores, heteroplasmy level, and depth of coverage are all customizable thresholds for visualizing data. For the purposes of this study, the following criteria were used: a quality threshold of 70; a heteroplasmy threshold of 0.18; and a coverage threshold of 40X (all values arbitrarily chosen for this study). Conditional formatting also was applied to the coverage columns. Reference alleles at a depth greater than or equal to the coverage threshold were highlighted for quick review of potentially-missed multiple SNP states at a position (i.e., point heteroplasmy). An overall coverage view of each variant is available for highlighting potential variants in low-coverage areas.

Figure 2. Review process. A) Sample variants are reported out for review; B) Review shortcuts are placed into the review column next to the corresponding nucleotide position.

A.

Review	Chrom:Co	Manual Call	Coverage	Ref	Ref Cov	Alt	Alt Cov
	chrM.146	146C	1719	T	1	C	1713
	chrM.152	152C	1678	T	18	C	1660
	chrM.263	263G	654	A	2	G	650
	chrM.302	309.1CC_Review	371	A	179	ACC,AC	122,66
	chrM.310	315.1C_Review	383	T	158	TC	198
	chrM.469	469Y_Review	1356	C	939	T	416
	chrM.513	523del_524del_REVIEW 62% READS MISSING	2187	GCA	818	GCCCA,G	0,6
	chrM.750	750G	4944	A	4	G	4938
	chrM.1438	1438G	4903	A	11	G	4883
	chrM.1811	1811G	4313	A	50	G	4257
	chrM.2217	2217T	4126	C	72	T	4022
	chrM.2706	2706G	4356	A	62	G	4217
	chrM.3106		4493				

B.

Review	Chrom:Co	Manual Call	Coverage	Ref	Ref Cov	Alt	Alt Cov	
	chrM.146	146C	1719	T	1	C	1713	
	chrM.152	152C	1678	T	18	C	1660	
	chrM.263	263G	654	A	2	G	650	
✓ +	chrM.302	309.1CC#	371	A	179	ACC,AC	122,66	chrM.508_A
✓	chrM.310	315.1C	383	T	158	TC	198	chrM.511_C
✗	chrM.469		1356					chrM.512_A
?	chrM.513	523del_524del_REVIEW 62% READS MISSING	2187	GCA	818	GCCCA,G	0,6	chrM.513_G
	chrM.750	750G	4944	A	4	G	4938	chrM.514_C
	chrM.1438	1438G	4903	A	11	G	4883	chrM.515_A
	chrM.1811	1811G	4313	A	50	G	4257	chrM.516_C
	chrM.2217	2217T	4126	C	72	T	4022	chrM.517_A
	chrM.2706	2706G	4356	A	62	G	4217	chrM.518_C
	chrM.3106		4493					

Once these values were set, the user reviewed the variants that meet filter criteria. Shortcuts currently available allowed the user to accept or reject ambiguous calls, mark variants for future reference (e.g., length heteroplasmy), and view five bases upstream and downstream of the variant to assist in quick resolution of ambiguous calls. Using this shortcut, the user was able to quickly review surrounding reference sequence in Excel without opening a sequence viewer. This feature facilitated elimination of reads or portions of reads that were inconsistent with the rCRS and may have been attributed to alignment or sequencing noise. The final haplotype then was compiled for review. A report of each sample can be saved containing the parameters used for analysis and the final haplotype. Haplotypes can be exported to a separate tab using an Excel macro that currently allows up to 350 individual haplotypes to be compiled for evaluation by phylogenetic-based haplogroup assignment using internet-based software packages (e.g., HaploGrep (27)).

Upload into HaploGrep

Haplotype strings generated by mitoSAVE were saved in a text file with the extension .hsd (e.g., *SampleFile.hsd*). This text file contained several columns for sample identification, targeted sequence ranges (e.g., 1-16569, or 16024-16365 and 73-340) defining the mtGenome, or HVI and HVII, respectively, and expected haplogroup (or blank if undetermined). The remaining columns contained the haplotype, with variants separated by tabs. To facilitate file generation, a small accompaniment file was available for .hsd generation. The .hsd file was uploaded to HaploGrep for haplogroup assignment and phylogenetic-based variant-call checking.

Ease of Use Testing

mitoSAVE is designed for application with sequence viewers by users with some mtDNA typing experience. To evaluate the usability of mitoSAVE, three novice users of the workbook with backgrounds in mtDNA analysis were given a brief (~5 minute) tutorial on use of the program with alignment files for resolving ambiguous variant calls.

3. PGM mtGenome Sequencing Protocol and Concordance Testing

Long PCR and Library Preparation

DNA from 24 samples (23 that had been sequenced with the MiSeq protocol) was amplified by long PCR (52). The PCR included SequalPrep™ 10× Reaction Buffer (ThermoFisher), SequalPrep™ 10× Enhancer B (ThermoFisher), SequalPrep™ long polymerase (5U/μl) (ThermoFisher), DMSO (ThermoFisher), primer sets (ThermoFisher), DNase-free water, and 5 ng of total genomic DNA according to the manufacturer's protocol. The amplification conditions were 2 min at 94 °C for polymerase activation, 30 cycles of 10 s at 94 °C for denaturation, 30 s at 60 °C for annealing, 8 min at 68 °C for extension; followed by a final extension of 5 min at 72 °C. The two amplicons were pooled in equimolar amounts (quantity determined using the Qubit). The PCR amplicons were enzymatically fragmented using Ion Shear™ Plus Reagents (ThermoFisher) Ion adapters and barcodes were ligated to the fragmented amplicons using the Ion Plus Fragment Library and Ion Xpress™ Barcode Adapters Kits (ThermoFisher). The library was size-selected at 315 bp with the Pippin Prep™ instrument (Sage Science, Beverly, MA).

Template Preparation

A diluted library (26 pM) was used to generate template positive Ion Sphere™ Particles (ISPs) containing clonally amplified DNA. Emulsion PCR was conducted using the OneTouch™ 200 Template Kit v2 DL with the Ion OneTouch™ DL configuration (ThermoFisher), template-positive ISPs were enriched with the Ion OneTouch™ ES (ThermoFisher), and quality of template-positive ISPs was assessed by using the Ion Sphere™ Quality Control Kit (ThermoFisher) on the Qubit® 2.0 Fluorometer, following the recommended protocol.

Sequencing and Data Analysis

Libraries were sequenced on the Ion 314™ Chip with the Ion PGM™ 200 Sequencing Kit (ThermoFisher) following the recommended protocol (70). Six barcoded samples were sequenced per 314 Chip. All PGM sequences were analyzed with the Ion Torrent Software Suite (v 4.0.2) using the plug-in variant caller (v 4.0). The VCF output of the variant caller was presented in tabular format, as a list of differences to the rCRS. BAM files were visualized with IGV. Whole mtGenome sequence data were compared with mtDNA sequences previously analyzed on the MiSeq (69).

III. Whole Genome mtDNA Sequencing Results and Discussion

The overall data show that mtGenome sequencing can be performed accurately and reliably on two different MPS platforms (i.e., MiSeq and PGM) based on the observation that two different platforms and chemistries provided the same results. Long PCR worked effectively for template enrichment and therefore DNA from reference samples, which typically is high in quantity and quality, is readily sequenced. The protocols herein for library preparation and sequencing are relatively efficient and should be able to be transferred to an operational laboratory. Compared with Sanger sequencing, a large number of samples can be prepared simultaneously (see results

below). The cost of obtaining whole genome data is far less than that of sequencing just regions HVI and HVII by Sanger sequencing (see results below).

1. MiSeq mtGenome Results

In a single run 72 to 96 samples could be analyzed simultaneously. Therefore, this MiSeq methodology offers a substantial improvement in throughput compared with current Sanger sequencing and (with other similar MPS procedures) should be considered the method of choice based on quality, cost, and information generating whole mtGenome data. Depending on the number of samples multiplexed (~72-96) reagent costs for mtGenome sequencing ranged from \$50-\$70 per sample.

Samples were multiplexed in an increasing series of 6, 12, 24, 48, and 96 to determine the number of whole genomes that could be sequenced with sufficient coverage. The cell line 9947A, which has been sequenced for mtDNA, was common to all series and correctly sequenced. Sufficient coverage was obtained for all multiplex series except the 96 sample multiplex. At the 96 sample level some regions for a few samples had too low coverage to yield complete sequence data. Therefore, the multiplex was set at approximately 72 samples (a few more samples could be added if desired), which performed well for obtaining full coverage of the whole mtDNA genome for all samples. A multiplex of 96 samples may still be feasible, but would require normalization and flow cell density to be near optimum conditions.

Given the throughput of this protocol, whole genome sequences were generated for 283 individuals. It would be impossible to have generated the same amount of data with a Sanger sequencing protocol (routinely used in forensic laboratories) during this phase of the project. The data were processed and are shown in Tables 5-7. All data were run through the Haplogrep program and sites to check were verified manually. Heteroplasmy determination was set arbitrarily at 0.18 for the study. Because this study is for reference samples, the focus primarily was on the predominant type and not on the lowest level possible to detect heteroplasmy. The entire sequence data for all individuals were published in King et al (69).

Table 5. Haplotype Diversity by Population

	HV1 and HV2 (16024-16365;73-340)			Whole mtGenome		
	AFA	CAU	HIS	AFA	CAU	HIS
N	89	74	115	89	74	115
Unique Haplogroups	56	62	57	71	71	70
Unique Haplotypes	79	68	99	87	74	111

AFA = African American; CAU – Caucasian; HIS = southwestern Hispanic

Table 6. Summary Population Statistics

Population	N	HV1 and HV2 (16024-16365;73-340)		Whole mtGenome	
		RMP	GD	RMP	GD
AFA	89	2.34%	0.988	1.28%	0.998
CAU	74	3.76%	0.976	1.35%	1.000
HIS	115	3.23%	0.976	1.14%	0.997

RMP = Random match Probability

GD = Gene Diversity

AFA = African American; CAU – Caucasian; HIS = southwestern Hispanic

Table 7. Haplogroup Assignment Based on Whole Genome mtDNA Data for 278 Individuals

<u>Haplogroup</u>	<u># AFA Individuals with Haplogroup</u>	<u>Haplogroup</u>	<u># CAU Individuals with Haplogroup</u>	<u>Haplogroup</u>	<u># HIS Individuals with Haplogroup</u>
A2d1	1	A2	1	A2	3
B4a1a1a2	1	B2	1	A2+64	10
H5a1p	1	C1b1	1	A2+64+!16111	2
L0a1b1a	1	H1	1	A2ae	3
L1b1a	1	H13a1a1a	1	A2d1	3
L1b1a+!1629 3	1	H16c	1	A2d1a	2
L1b1a15	2	H1a	1	A2g	1
L1b1a3	2	H1a1	1	A2h1	5
L1b1a9	1	H1a3b	1	A2j1	1
L1b2a	1	H1bk	1	A2p	1
L1c1b	2	H1c1a	1	A2q	2
L1c1c	1	H1c7	1	A2r	2
L1c2b1a	1	H1f	1	A2v	1
L1c2b1b	1	H1h1	1	A2w	2
L1c2b1b1	2	H1n1a	1	B2	8
L1c3a1a	1	H3	1	B2a	1
L1c3b1a	1	H3af	1	B2b	2
L1c3b2	1	H3v+16093	1	B2c1	3
L2a1+143	1	H4a1a1a1a1	1	B2c2a	1
L2a1a	1	H52	1	B2f	1

L2a1a1	1	H56a1	1	B2g1	3
L2a1a2a1	2	H5a1	1	C1b	2
L2a1a3	1	H5a1g1	1	C1b1	1
L2a1c	1	H5a1j	1	C1b10	1
L2a1c3	2	H5b1	1	C1b11	1
L2a1c4a1	2	H5e1	1	C1b3	1
L2a1e	1	H6a1a	1	C1b5	1
L2a1e1	2	H6c	1	C1b5b	1
L2a1f	3	H76	1	C1b7a	1
L2a11a	1	H7c1	1	C1b9a	2
L2a1m1	1	H7h	1	C1c	2
L2b1a3	1	H86	1	C1c1a	2
L2b1b	1	HV17	1	C1c1b	1
L2b2	1	I1a1d	1	C1c2	1
L2b2a	1	I2	1	C1c4	1
L2c	2	J1b1a1a	1	C1d1c1	3
L2c1	1	J1b1a1c	1	D1c	1
L2c2	2	J1c	1	D1d	1
L2d+16129	1	J1c2	1	D1f	1
L2e1	1	J1c2c1	1	D1i	2
L3b1a	1	J1c3	1	H1ag	1
L3b1a+!1612 4	1	J1c3a1	1	H1ag1	1
L3b1a1	1	J1c4c	1	H1ba	3
L3b1a1a	1	J1c5d	1	H1g1	1
L3b1a4	1	J2a1a1a2	2	H1j1	1
L3b1a7	2	K1a11	1	H2+152	1
L3b2	2	K1a1b2b	1	H30a	1
L3d2a	1	K1b1a1	1	H3aa	1
L3d3a1a	1	K2b	1	H5p	1
L3e1e	1	K2b1a1	1	H82	1
L3e1f	1	T1a1	1	I5a2	1
L3e2a	2	T1a1c	1	J1c3	1
L3e2a1a	1	T2a1a	1	K2a8	1
L3e2a1b	1	T2b3	1	L1c3a1b	1
L3e2b	3	T2c1d1a	2	L2a1+143+16189	1
L3e2b+152	1	T2f3	1	L2a1+16189 (16192)	1
L3e2b1a1	1	U2e1b1	1	L2c	1
L3e2b3	1	U3a2	1	L2c2b	1

L3e3b	1	U4a	1	L3d1b1	1
L3e3b2	1	U4a2	1	L3e2a1a	1
L3f1b	1	U4c1a	1	T2a1a	1
L3f1b1	1	U4d2	1	T2b	1
L3f1b1a	2	U5a1a1	1	T2b+152	1
L3f1b1a1	1	U5a1a1d	2	U4a	1
L3f1b3	1	U5a1b1	1	U5a1b	1
L3f1b4	1	U5a1c	1	U5a1b1c2	1
L3f1b4a	1	U5a2a1+152	1	U5b1b1b	1
L3k1	1	U5b1b1a	1	U5b2a1a	1
M1a1	1	U5b1f	1	W3a1c	1
M23	1	U5b2c2b	1	U5b2b3a	1
U4c1a	1	V2	1		
Unique	Count	Unique	Count	Unique	Count
71	89	71	74	70	115

AFA = African American; CAU – Caucasian; HIS = southwestern Hispanic

From 283 mtGenome population samples (African American, $n=87$; Caucasian, $n=83$; Southwest Hispanic, $n=113$) 11,607 variants, defined in relation to the rCRS, were observed. These variants were distributed across 1,353 nucleotide positions throughout the mtGenome (Figure 3). Of these 1,353 positions, more than one variant type was observed at 55 base positions among all samples sequenced. A total of 722, 220, and 96 of the 11,607 variants were observed in one, two and three samples respectively, and three variants (263G, 4769G, and 15326G) were observed in all samples, which is a reflection of the reference used. The remaining variants were observed in between 4 and 282 samples with 1,302 variants (approximately 92.3% of the 1,411 total unique variants) being observed in 20 or less of all samples sequenced.

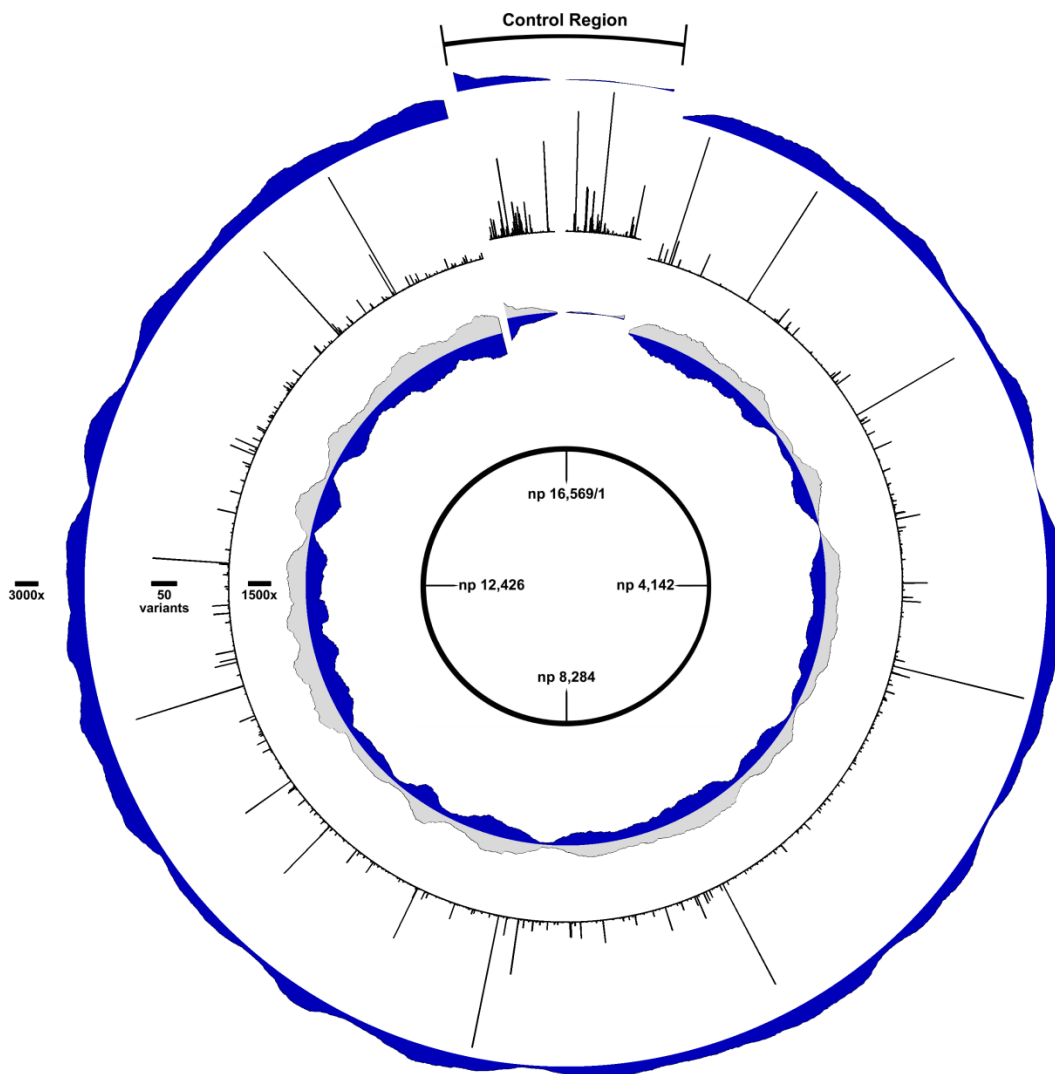


Figure 3. A concentric Circos plot of the mtGenome representing mean coverage (outer circle; $n=24$), variants observed per nucleotide position (middle circle; $n=283$), and mean coverage differentiated by reverse (dark) or forward (light) strand (inner circle; $n=24$). The rose diagram in the center is included for nucleotide position orientation and scale bars are included to the left of the individual plots to approximate values. The control region is offset slightly for orientation. The disproportionately-low coverage observed in HVII is likely an artifact of alignment to a linear reference.

As expected, polymorphism density was clustered heavily in the HVI and HVII regions. Out of all observed variants, 2,938 of the variants (25.3%) were observed in these two regions which comprise only 3.7% of the mtGenome. However, 8,669 of the variants (74.7% of all variants observed) resided outside of the HVI and HVII regions. The distribution of variants is inflated artificially, however, by high frequency variants (because of the artifact of using a reference for allele calling). A total of 15 variants, 4 of which reside in HVI/HVII, appeared in more than half the samples and account for 3,638 (31.3%) of all variants observed. These high-frequency reference-alignment artifacts are unavoidable and do not change the observed distribution of variants. These findings illustrated the untapped potential of the coding region for discriminatory power and more effective haplogroup assignment (see below).

The task of generating concordance mtGenome data using Sanger sequencing for such a large dataset is a time-consuming, arduous task which is impractical. However, previously-generated Sanger sequencing data for HVI and HVII were available for a subset of samples (n=8). All MPS data were concordant at all positions with Sanger sequencing. These data included point and length heteroplasmy, the latter of which was previously difficult to interpret given the nature of Sanger sequencing (and not considered for concordance). Whole genome sequence data for the 9947A cell line were compared and were completely concordant; however some heteroplasmy was observed at some sites consistent with the findings of (71). Lastly, a 7-sample exchange with Walther Parson's laboratory (Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria) produced concordant results. His data were generated on the PGM MPS system and further supported reliability of both systems (i.e. PGM and MiSeq) by orthogonal testing. Two of the exchanged samples each had an example of point heteroplasmy which was concordant between the MPS methods. Interestingly the quantitative assessment between laboratories and different MPS systems was very similar. For position 195 Y (T/C) in one sample and position 234 R (A/G) in another sample the relative contributions of the heteroplasmic variants between laboratories were 0.71C/0.29T vs 0.67C/0.33T and 0.51A/0.49G vs 0.54A/0.26G, respectively. These data further support the concordance and reliability of the methodology described herein.

Figure 4 shows the mean coverage of 24 representative samples across the mtGenome (from np 1 to np 16569). Although the template was generated with two approximately 8kb amplicons, the coverage does vary across the genome but consistently among samples. Therefore, the variation across the genome was likely due to post PCR effects during library preparation and/or sequencing and likely will persist with the current protocol. Nonetheless, quality results can be obtained. The lower coverage areas will be the threshold performance determinants for obtaining full sequence data.

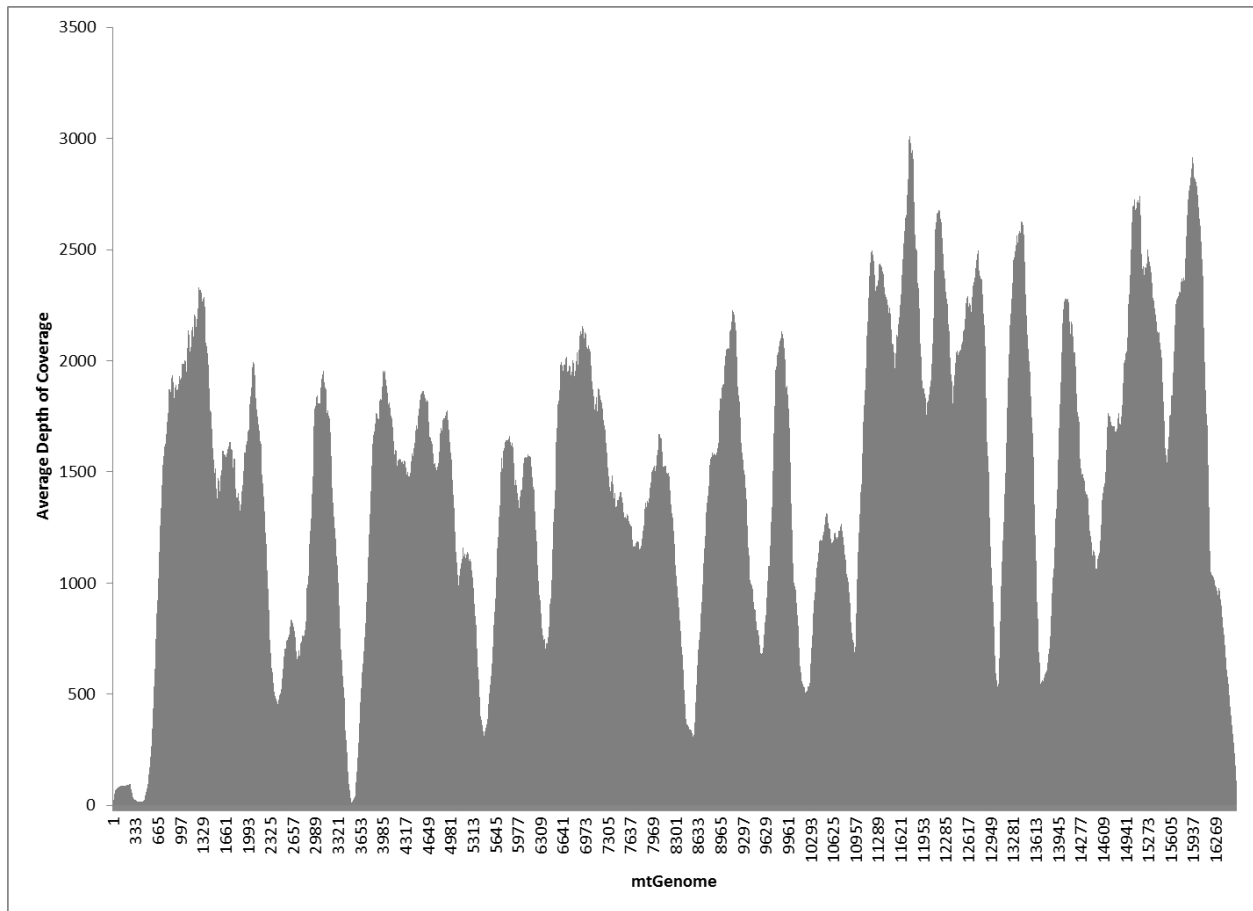


Figure 4. A histogram illustrating mean depth of coverage per nucleotide position of the mtGenome ($n=24$).

Although coverage varies across the genome, there were few areas where strand bias was observed. Figure 5 displays the ratio of coverage between the forward and reverse strands at each nucleotide position (lower coverage/higher coverage) and indicated that most ratios were greater than 60%. However, a few sites had low strand coverage ratios. In Figure 5 the Y axis is the number of positions and the X axis is the strand bias ratio (x100). The data showed that the majority of positions had relatively little strand bias. There were 209 positions out of 16570 positions where the strand bias ratio was ≤ 0.30 (Table 8). These low ratios represent a very small percentage of the total sites and are localized to specific regions of the genome. While strand bias does not necessarily indicate lower quality data, balanced strand representation does provide a high degree of confidence that a correct base call was made. Validation studies will be needed to determine whether strand bias has any effect on allele call accuracy.

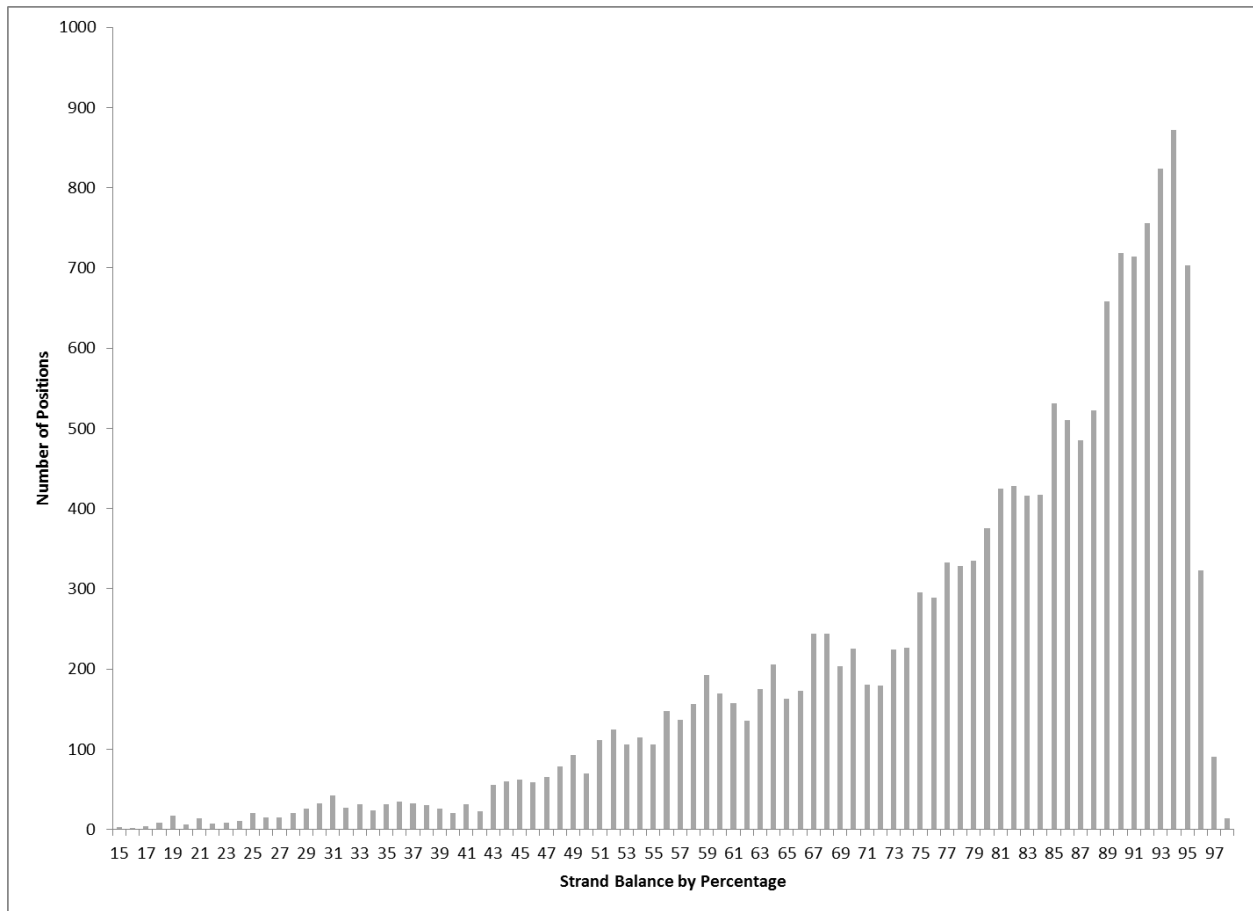


Figure 5. Strand bias histogram displaying the distribution of strand balance across all nucleotide positions of the mtGenome for an arbitrary subset of samples ($n=24$).

Table 8. mtDNA nucleotide positions with average strand bias (based on coverage) ratios ≤ 0.300

np	Avg SR	SD	np	Avg SR	SD	np	Avg SR	SD
1	0.158	0.294	3475	0.215	0.097	3543	0.284	0.158
6	0.282	0.295	3476	0.224	0.101	3553	0.288	0.166
7	0.294	0.295	3477	0.211	0.097	3580	0.29	0.13
8	0.219	0.204	3478	0.204	0.1	3581	0.276	0.153
511	0.299	0.182	3479	0.197	0.102	3582	0.262	0.149
514	0.293	0.184	3480	0.198	0.101	3583	0.247	0.144
515	0.29	0.178	3481	0.197	0.103	3584	0.248	0.121
516	0.285	0.181	3482	0.197	0.103	3585	0.236	0.113
517	0.284	0.172	3483	0.196	0.105	3586	0.22	0.109
518	0.271	0.175	3484	0.18	0.087	3587	0.214	0.107
519	0.284	0.178	3485	0.186	0.095	3588	0.199	0.111
521	0.26	0.177	3486	0.195	0.109	3589	0.185	0.107
523	0.269	0.177	3487	0.197	0.109	3590	0.186	0.106
524	0.29	0.181	3488	0.206	0.117	3591	0.19	0.096
525	0.264	0.153	3489	0.215	0.14	3592	0.207	0.121
526	0.267	0.146	3490	0.221	0.131	3593	0.21	0.108
527	0.258	0.146	3491	0.226	0.143	3594	0.199	0.06
528	0.258	0.137	3492	0.245	0.16	3595	0.216	0.059
529	0.242	0.134	3493	0.252	0.163	3596	0.221	0.059
530	0.262	0.133	3494	0.245	0.159	3597	0.218	0.06
531	0.276	0.133	3495	0.259	0.166	3598	0.217	0.059
532	0.282	0.126	3496	0.263	0.17	3599	0.219	0.061
533	0.293	0.131	3497	0.254	0.183	3600	0.224	0.057
537	0.297	0.138	3498	0.271	0.177	3601	0.237	0.052
538	0.298	0.135	3499	0.279	0.185	3602	0.236	0.055
539	0.29	0.13	3500	0.286	0.232	3603	0.24	0.051
540	0.287	0.129	3502	0.294	0.242	3604	0.238	0.044

541	0.293	0.123	3503	0.254	0.222	3605	0.245	0.048
542	0.279	0.122	3504	0.27	0.211	3606	0.253	0.063
543	0.285	0.142	3508	0.298	0.286	3607	0.247	0.062
544	0.287	0.129	3510	0.295	0.277	3608	0.255	0.059
545	0.275	0.133	3511	0.273	0.302	3609	0.256	0.064
546	0.254	0.124	3512	0.256	0.262	3610	0.259	0.066
547	0.251	0.119	3513	0.247	0.267	3611	0.263	0.066
548	0.238	0.121	3514	0.227	0.272	3612	0.259	0.064
549	0.24	0.116	3515	0.183	0.223	3613	0.257	0.057
550	0.244	0.094	3516	0.193	0.212	3614	0.255	0.062
551	0.262	0.1	3517	0.177	0.203	3615	0.268	0.067
552	0.258	0.104	3518	0.167	0.164	3616	0.275	0.065
553	0.255	0.101	3519	0.172	0.135	3617	0.279	0.063
554	0.255	0.1	3520	0.153	0.126	3618	0.276	0.061
555	0.264	0.11	3521	0.18	0.13	3619	0.289	0.065
556	0.268	0.106	3522	0.185	0.152	3620	0.289	0.065
557	0.28	0.095	3523	0.199	0.149	3621	0.287	0.064
558	0.295	0.091	3524	0.171	0.154	3622	0.295	0.066
559	0.3	0.095	3525	0.201	0.146	3623	0.294	0.069
572	0.291	0.098	3526	0.188	0.142	3624	0.295	0.063
573	0.287	0.099	3527	0.201	0.144	8591	0.299	0.121
574	0.276	0.101	3528	0.195	0.137	8592	0.293	0.12
575	0.286	0.09	3529	0.172	0.152	16563	0.277	0.167
576	0.293	0.087	3530	0.196	0.152	16564	0.242	0.161
577	0.296	0.079	3531	0.193	0.14	16565	0.231	0.188
2464	0.293	0.085	3532	0.226	0.155	16566	0.216	0.171
2465	0.293	0.102	3533	0.197	0.143	16567	0.209	0.168
2466	0.3	0.109	3534	0.213	0.149	16568	0.199	0.181
3468	0.296	0.102	3535	0.216	0.169	16569	0.161	0.16
3472	0.294	0.171	3536	0.23	0.169	16570	0.158	0.15

3473	0.265	0.119	3537	0.269	0.23
3474	0.25	0.101	3538	0.262	0.23
			3539	0.289	0.223

np= nucleotide position; SR= strand ratio; SD = standard deviation

Forensic Population Statistic Parameters

The haplotype and haplogroup diversity of HVI/HVII were compared with that of the mtGenome. The variant calls from HVI/HVII resulted in 38 fewer unique haplogroups (55, 70, 56) for African Americans, Caucasians, and Southwest Hispanics, respectively), and 30 fewer unique haplotypes (76, 77, 96 for African Americans, Caucasians, and Southwest Hispanics, respectively) than when whole mtGenome sequence data were assessed (Table 5).

Population genetics parameters of mtDNA sample sets are reliant partially on the size of the database. Generating a total of 283 mtGenome sequences in a relatively short time was impressive compared with Sanger sequencing capabilities. However, it was a relatively small number for assessing mean RMP and GD. The increase in RMP can be appreciated better by comparison of HVI/HVII sequences and mtGenome sequences from the same sets of individuals. The RMPs for HVI/HVII data were 2.42%, 3.12%, and 3.33% (Table 6) in African American, Caucasian, and Southwest Hispanic populations, respectively. In contrast, the RMPs based on mtGenome sequences were 1.31%, 1.20%, and 0.98%, respectively. This difference was significant ($p=0.0036$; paired, two-tailed Student's T-test).

A similar pattern held with GD. The GD was 0.987, 0.981, and 0.975 for HVI/HVII compared to 0.998, 1.000, and 0.999 using the mtGenome data for African Americans, Caucasians and Southwest Hispanics, respectively. The increase in GD was significant ($p=0.0063$; paired, two-tailed Student's T-test). As the database increases in size, it is expected that the RMP will decrease and GD will increase. Interestingly, both the RMP and GD for the coding region alone yielded equivalent RMP and GD to the mtGenome reinforcing that there is more variation residing in the coding region compared with the control region of the mtGenome. Given the ease of generating sequence data and concomitant high quality results, MPS sequencing of the entire mtGenome should be considered as a viable approach to supplement power of discrimination, when warranted.

Database querying of sequence data typically is done using a haplotype defined by differences from the rCRS rather than a "string search" (i.e., using the entire sequence). Thus, variant reports (VCF v4.1 files) were converted into concise haplotypes using mitoSAVE (2). Haplotypes were exported from mitoSAVE in *.hsd* or *.txt* file format for upload to HaploGrep (8). For this study, the highest ranking haplogroup was relied upon with no assumed haplogroup status prior to assignment. As part of the quality assessment of haplotype data, haplogroup assignment was performed to discern established haplogroup specific mutations from yet undescribed "private" mutations in the respective haplogroup backgrounds. The latter were subjected to additional quality checks to confirm their authenticity. From all 283 individual samples, 14 different clades were represented with 208 distinct haplogroups and 279 unique haplotypes. By population, there were 70, 79, and 70 distinct haplogroups and 85, 83, and 111 distinct haplotypes for African Americans, Caucasians, and Southwest Hispanics, respectively. The haplotypes can be found in King et al (69).

The haplogroup assignments were consistent with the declared population affinity of the individual (Table 7). As expected, haplogroup assignment was better determined with whole genome data than solely regions HVI and HVII (Table 9). Nine samples (3.2% of all samples

sequenced) changed clades (e.g., G→L) between limited HVI/HVII data and that of the mtGenome data (Table 9).

Table 9. Samples in which the clade^a was reassigned based on mtGenome vs. HVI/HVII sequence data.

HVI/HVII		mtGenome	
HaploGroup Assignment	Quality %	HaploGroup Assignment	Quality %
N11a	80.3	L2a1c3	93.1
M73'79	95.1	L3b1a+!16124	95.1
G3	89.3	L3b1a7	97.2
N2	95.2	L3e1f	95.1
HV0	93.9	V2	95.4
R0+16189	87.7	H4a1a1a1a1	97.8
M33c	83.6	A2+64	90.3
D4e1	82.8	A2+64	91.8
P5	95.9	H32	92.5

a. As assigned by HaploGrep (8) and Phylotree (9)

b. As labeled in EMPOP (13)

In fact, six of nine samples changed macrohaplogroups (i.e., L, M, or N). Further analysis of the HVI/HVII haplogroup assignments indicated a variation in top-ranked haplogroups independent of stated quality. Quality scores for these nine samples ranged from 80.3 to 95.9 using HVI/HVII data. In fact, four of nine samples had quality scores greater than 90.0 (rank equivalent 0.900). These observations can be explained by the fact that HaploGrep's assignment is based on signature mutations indicated on the branches of Phylotree only, while other mutations present in the corresponding mtGenomes were not taken into consideration. The ranked haplogroups displayed by HaploGrep for these samples varied widely. Sample USA_TX_0257, for example, listed the following as the top three possible haplogroups: P5 (rank-0.959), U5b2a1a (rank-0.914), H32 (rank-0.886). Conversely, the mtGenome haplotype analyzed with HaploGrep listed H2+152 (rank-0.925), H (rank-0.917), and H32 (rank-0.916) as the top three possible haplogroups. These observations demonstrated that haplogroup assignment can be misleading with limited sequence data from HVI/HVII regions. While such effects would not necessarily have an impact on profile comparisons for forensic applications, they could have a negative impact on population and evolutionary studies.

Conclusions of MiSeq mtGenome Protocol

To the best of our knowledge, this was the first reported study of a relatively large number (compared with Sanger sequencing throughput) of mtGenomes that have been sequenced in a high-throughput fashion using the Illumina MiSeq system. Subsequently, another study by McElhoe et al (72) reported mtGenome sequencing on the same platform but with fewer samples sequenced. The throughput level of the Nextera XT DNA Sample Preparation Kit was tested and found to be exceedingly high. While some strand bias was observed, it generally was limited to

areas of low coverage and did not diminish the ability to assign variant calls. Also, it was possible, due to a high depth of interrogation, to type length and point heteroplasmies.

By sequencing the entire mtGenome versus only HVI/HVII, the additional variant calls significantly improved the discrimination power of haplotypes. An overall improvement in the resolution of haplogroup assignments was observed compared with only the control region to the mtGenome, while haplogroup assignment was ambiguous for HVI/HVII segments of those mtGenome sequences that were not represented by control region polymorphisms in Phylotree.

Software tools can facilitate analyses and are imperative for technology transfer. The time of analysis was greatly reduced compared with Sanger sequencing, as data analyses become more automated. It should be noted that with Sanger sequencing, it would be very demanding to sequence just the control region of 283 samples, let alone whole genomes in a 1-12 month time frame. The 1-12 month frame herein was not reflective of the actual time as development time was included. Indeed, one individual can sequence 72-96 samples from DNA to VCF file in approximately 4.5 days. Subsequently, haplotypes can be compiled in a matter of hours using mitoSAVE (see below). This MPS approach will facilitate generation of whole mtGenome population data to support human evolution, forensic, and medical studies. The overall conclusion is that this protocol allows for reliable typing of whole mtDNA genomes from reference samples in a high throughput and relatively low cost manner.

2. PGM Sequencing Results

Parson et al. (73) demonstrated that sequence results with the PGM were highly concordant with those obtained with Sanger sequencing. Whole mtGenome sequencing was performed herein on the PGM to determine its feasibility, accuracy, and reliability. An ancillary benefit of developing this protocol was that the generated mtGenome data could be used for additional concordance testing. It is difficult to validate whole mtGenome sequencing by MPS with Sanger-based sequencing systems. The throughput of the latter is so much lower than MPS that only a small region of the mtGenome can be assessed by both approaches within a reasonable time and cost. An orthogonal MPS technology is a better approach to effectively compare whole mtGenome results for concordance and hence obtain supporting data on the reliability of MPS. Results were compared with sequence data from the MiSeq protocol.

In this study, 6 samples were multiplexed and sequenced at one time on a 314 chip (10 megabase throughput). The average throughout of 4 chips was 84 Mb (± 17), and the average total reads was 448,129 ($\pm 78,773$). Sufficient coverage was obtained to reliably determine the sequence for the entire mtGenome of six pooled libraries. In all, 24 samples were sequenced successfully on 4 chips.

The depth of coverage pattern was similar among all 24 samples. As with the MiSeq data balanced coverage across nucleotide positions did not occur. However, coverage was consistently low at certain positions and high at other positions across the mtGenome. Coverage of, for example, one sample (no. 8) ranged from approximately 25X to 2815X. This range in coverage might be attributed to homopolymeric stretches as these areas may be difficult to

sequence due to chemistry-related limitations (74) and/or may be filtered out due to low quality. To elucidate coverage variation, areas of relatively high ($\geq 810X$) and low coverage ($\leq 500X$) were analyzed. There were 17 regions with relatively high coverage and 18 regions with low coverage. Areas of low coverage had substantially more C homopolymers (defined as two or more C's in a row) than high coverage areas. Interestingly, all regions with C homopolymers interrupted by another base (e.g., CnTCn) displayed relatively low coverage. However, long homopolymers (≥ 4 or more C's) alone did not explain all the reduction in coverage.

In theory, both strands of a DNA duplex should be sequenced equally. For all 24 samples, two-thirds of the positions of the genome had strand ratios that were greater than 0.5. A few sites had more extreme strand bias. For example, in one sample (no. 8), out of a total number of 69 reads at np 300, 7 forward direction reads were aligned, while 62 reversed direction reads were aligned; the average strand bias at this position was 0.08. Across the 16,568 nucleotide positions surveyed, only 1045 positions showed an average ratio less than or equal 0.1. While strand bias does not necessarily indicate lower quality data for base calling, balanced strand representation does provide a higher degree of confidence that a correct base call was made. In circumstances where in one strand direction there may be an indication of a deletion and in the other strand there is no indication (due to chemistry and/or software), this site might be deemed inconclusive (but this interpretation will depend on further validation studies). Special attention should be given to high strand bias sites and deletions (see below).

Parson et al. (73) reported some reads had false deletions in PGM-generated mtDNA sequence data. These deletions could not be verified with Sanger sequencing. In the study herein, a number of positions ($n=1391$) showed some level of false deletions. These false deletions were measured as a ratio ($DR = \text{deletion reads} / \text{total reads}$). In the 16,568 mtDNA nucleotide positions, 156 positions displayed a false deletion of greater than 0.15 in one or more individuals. These false deletions were associated largely with homopolymers (155/156) with a single guanine residue showing a DR of 0.18 in one sample (sample no. 17). DRs were observed up to 0.84, although very few positions across the 24 samples had this high ratio. The np 11635 had the highest average DR (0.69). In this position, 23 samples showed DRs greater than 0.58 except for one sample (no. 23) with a DR of 0.18. After reviewing the BAM file of this sample in IGV, a variant was observed at the nearby site A11654G. Several positions showed a similar pattern with a variant unique within the dataset that might indicate an association with a reduction in false deletions. However, the sample size is too small for any inferences and further study is needed to determine if such SNP variants could somehow be associated with a reduction of DR.

In some specific regions with 2 consecutive guanine residues (GG), false deletions were observed in PGM sequence results (e.g., nps 6957, 7077 and 12629). In fact, two of the six highest positions in terms of DR and 16/156 positions with high DR showed this GG pattern. However, this pattern alone does not account for all false deletions observed. Across the mtGenome, there were 296 GG homopolymers of which only 16 were associated with substantial false deletions. These observations suggested that homopolymers were not the sole cause of this phenomenon, and it likely may be sequence specific. No discernable sequence pattern was observed for these false deletions. The frequency of sequence errors has been a subject of other studies. Nakamura et al. (75) showed that sequence specific errors occur in Illumina Genome

Analyzer II data, and that these errors were triggered by inverted repeats and GGC motifs. Meacham et al. (76) developed a statistically principled framework and reported that the most common sequence context error is associated with the GGT motif. Furthermore, Allhoff et al. (77) analyzed errors on three different Illumina platforms (GAIIx, MiSeq, HiSeq2000), confirmed previously known error-causing sequence contexts and reported new specific ones. A similar scenario may be occurring with a GG motif described herein for PGM data.

For the 24 samples analyzed, 31-98 SNP variants were observed (each annotated as a difference from the rCRS) per sample. Of the 24 samples, 23 samples had been sequenced previously on the MiSeq platform (69). All 1237 (SNP) variants (across the 23 mtGenomes) were concordant between the PGM and MiSeq data, excluding the number of Cs in homopolymers around np 310 and 16189 regions. These regions are well known sites for heteroplasmic length variants and typically are not used in forensic identifications (63). Parson et al. (73) reported similar findings in which they described that approximately two-thirds of the different bases (compared with Sanger sequencing data) were observed in or around homopolymeric sequences stretches.

There were two sites worth noting that presented apparent differences between PGM and MiSeq sequence data. One site was the dinucleotide CA insertion at the np 514-524 region. For example, a CACA (83.3%) insertion was predominant in one sample (no. 6) with PGM sequence data; however there were other insertions (CA, 8.3%; and CACACA, 8.3%) also present at much lower representation. This region had low coverage and some reads were not sequenced fully. In contrast, data from the MiSeq showed overwhelmingly CACA reads (95.7%), a relatively small portion of CA (4.3%; less than observed in the PGM data), and no CACACA reads. Based on this comparison, there is no way currently to indicate whether the minor (CA)_n types and their proportions are real, and therefore the lower representation CA variants were considered inconclusive.

Another site was a 9-bp deletion of ACCCCCTCT at np 8280-8288 (also known as CCCCCTCTA at np 8281-8289) (21). The 9-bp deletion was confirmed easily from PGM data. In the PGM workflow, sequence data were aligned with TMAP (78) and variants called using the variant caller v4.0. The MiSeq workflow employed BWA to align reads and GATK to call variants. This difference in workflows between the two MPS platforms created a “perceived” difference in insertion/deletion calling because of alignment strategies. The underlying data were the same, but the outputs yielded different nomenclature. To demonstrate this workflow-dependent difference FASTQ files generated by the PGM were aligned and called using BWA/GATK and similar alignment problems were observed. Software dependent alignment illustrated the importance of validating bioinformatics workflows in haplotype nomenclature for reliability and consistency among laboratories.

Conclusions for PGM mtGenome Protocol

This study permitted an evaluation of the performance of PGM for mtGenome sequencing and highlights performance in general that may need to be addressed for the application of methodology in forensic genetics (or for that matter any discipline that may seek to sequence mtDNA). mtDNA sequence data generated from the PGM were analyzed and demonstrated to be

highly reliable. Sequence data generated on the PGM and the MiSeq systems were highly concordant except for the number of Cs in homopolymers around np 310 and 16189 regions, which are not used currently for forensic identifications generated using Sanger methods (63). Depth of coverage variation and strand bias were identified but did not impact reliability of variant calls. In addition, multiplexing of samples was demonstrated which can improve throughput and reduce overall cost per sample analyzed.

Overall, the results of this study supported that whole mtGenome sequence data with high accuracy can be obtained using the PGM platform. The study demonstrated the importance of validation studies to better understand the system(s) used, to highlight potential limitations in specific target regions, and to identify robust and/or inconclusive sequences to refine diagnostic interpretations.

3. mitoSAVE Results

Manual translation of a VCF file is an arduous task that creates a bottleneck in data analysis and introduces potential for user error. Additionally, the treatment of insertions in homopolymeric stretches by currently employed commercially-available alignment software does not necessarily follow well-established forensic conventions (63,67,68). In addition, forensic standards dictate that certain sites and variants are anchored in position, e.g., 310T, and that alignments be adjusted at these fixed positions (65). No current MPS alignment software meets these criteria. Thus, the overall time needed to process a single VCF file manually and have it concordant with alignment and nomenclature standards can exceed 10-15 minutes for an experienced user. This time frame will increase as new alignment scenarios arise when analyzing the entire mtGenome and can be quite burdensome to meet the throughput of MPS in which 100s to 1000s of whole mtGenomes can be readily generated.

mitoSAVE was developed to facilitate alignments choices. Three novice users were given a copy of mitoSAVE, BAM and VCF files for a small subset ($n = 6$) of sample data and tasked to generate haplotypes for the samples. IGV was made available to view reads and confirm variant calls. Processing time ranged from 110 s to 200 s averaging ~150 s per sample. Once familiar with the workflow a haplotype was generated from a VCF file in less than one minute per sample. Thus, the automated variant reassignment and haplotype generation allowed for a much faster processing time and higher throughput concomitant with increased sample sequencing of MPS systems.

In addition to a substantial reduction in processing time, mitoSAVE accounted for known phylogenetic variants in its haplotype generation. A 'Watchlist' maintained locally within mitoSAVE defines alignment issues known to occur with traditional alignment software and with specific MPS-offered tools provided by commercial manufacturers. For instance, hypervariable region III (HVIII) contains an AC repeat of ten bases from np 515-525 (in the rCRS). This region is prone to both insertions and deletions. Forensically-naive alignment tools tend to place both the insertion and deletion of these AC repeats starting at np 513-514 while forensic convention places them at 523-524. Thus, haplotypes would need to be corrected manually for this alignment and likely reviewed in a sequence viewer. mitoSAVE, however, reassigns this

dinucleotide repeat polymorphism to the proper position. When mitoSAVE detects a two base pair deletion at np 513, it automatically reassigns it to the 3' end of the repeat (i.e., 513del, 514del→523del, 524del).

Another example is np 249del which is a relatively common variant that GATK aligns to np 248. This position is called accurately by the onboard rule-based formulas which correct for the alignment software. However, when a haplotype contains a transition at np 247 (G247A) and a deletion at np 249, mitoSAVE initially calls the deletion at np 247 according to rule-based formulas. This alignment, however, differs from the phylogenetically-established 249del by two differences from the rCRS. In this situation, though, mitoSAVE automatically reassigns the deletion to the phylogenetically-established position (np 249) and correctly calls the allelic state at np 247 (247A).

A third example is the intergenic region between tRNA lysine and cytochrome oxidase II that contains a 9-bp repeat at np 8272-8289. A deletion of one of these repeats has been described in a subset of Asians and Native Americans (79,80). The resulting alignment of reads ending near this repeat creates noise with some alignment software. mitoSAVE, however, labels this polymorphism as a possible 9-bp deletion which is resolved and confirmed by viewing the associated BAM file.

mitoSAVE currently has a number of such positions listed in its 'Watchlist' allowing for reassignment of variants based on previous data sets and conventions preferred by the user. The list of positions in the 'Watchlist' can be increased and configured with experience of processing VCF files. Not every scenario can be accounted with the current list as alignment issues are based on VCF files processed from a relatively-small sampling of mtGenomes ($n = 278$). As more samples are sequenced more situations will be discovered. Such positions can be added easily to the list. The 'Watchlist' allows users to maintain defined positions to generate haplotypes.

mitoSAVE offers control over variant inclusion/exclusion based on data quality and coverage. Because accurate haplotypes are reliant on quality sequence data, users can set thresholds, review variants, and generate haplotypes all in a more consistent manner than current MPS-related software allow. Thresholds for coverage allowances and degree of heteroplasmy may be set by users. These thresholds may not yet be defined well and can be defined better as more mtDNA sequence data are accumulated. mitoSAVE allows each user to set these levels as deemed appropriate based on extant data and internal studies. This level of control promotes accurate haplotype nomenclature and allows consistent haplotypes to be generated by different users.

An optional macro has been included within mitoSAVE that, when selected, emails the VCF being processed to the UNTHSC for troubleshooting and continued improvement of the overall tool for global use. With this in mind, submission is encouraged of novel variants, upgrade suggestions, and troubleshooting questions. Finally, mitoSAVE in its present iteration is specific for human mtDNA analysis; however, it could be configured to using sequence data from any small genome or genome region and is limited only by the number of rows in Excel (e.g., Excel 2010 ~1.5 Mb). mitoSAVE is publicly available

(http://web.unthsc.edu/info/200210/molecular_and_medical_genetics/887/research_and_development_laboratory/4).

Short mtDNA Amplicon Sequencing

The success of the above protocols for mtGenome sequencing demonstrated that long PCR was more than sufficient for the enrichment phase of the analysis of reference samples. Therefore, there was no need to focus on amplifying shorter regions of the mtGenome for reference samples. However, transitioning mtGenome sequencing to analysis of challenged evidentiary samples will require development of short amplicon that span the genome.

IV. STR and SNP Panels- Materials and Methods

1. Selected Markers (Proof of Concept Panel)

In this section, the methodology, output results, overall performance, and findings for nuclear markers (i.e., STRs and SNPs) and MPS are presented. To facilitate flow some discussion is inserted in various section where warranted, as opposed to only in the Conclusion of the section. Different enrichment strategies (which impact library preparation) were employed and all were able to accommodate genetic marker typing, although each one had advantages and limitations when compared with each other. These findings demonstrate that different marker types and a large battery of markers can be sequenced simultaneously by MPS.

A panel of forensically-relevant genetic markers was selected from the literature (81-92) and from existing commercial STR kits. The markers were a collection of autosomal, X chromosome and Y chromosome STRs and human identity and bioancestry SNPs (the bioancestry SNPs will not be part of the final identity panel as they are not best suited for standard human identity testing; they were used solely to increase the number of markers for demonstration purposes of high throughput). The markers and their chromosomal locations are listed in Tables 10-12 and 15. This set was an initial test set for preliminary analysis and the final panel will be determined based on the general review of the output data.

Table 10. Autosomal STRs in the test panel.

Autosomal STR	Chromosome	Location
D1S1656	1	230,905,305-230,905,457
D2S441	2	68,238,998-68,239,157
D2S1338	2	218,879,515-218,879,706
D3S1358	3	45,582,205-45,582,335
FGA	4	155,508,848-155,509,043
D5S818	5	123,111,198-123,111,332
CSF1PO	5	149,455,735-149,456,053
D7S820	7	83,789,441-83,789,683
D8S1179	8	125,907,080-125,907,260
D10S1248	10	131,092,482-131,092,583
TH01	11	2,192,214-2,192,381

D12S391	12	12,449,930-12,450,154
vWA	12	6,093,104-6,093,254
D13S317	13	82,722,056-82,722,247
Penta E	15	97,374,212-97,374,590
D16S539	16	86,386,124-86,386,411
D18S51	18	60,948,814-60,949,143
D19S433	19	30,417,027-30,417,232
D21S11	21	20,554,259-20,554,481
TPOX	2	1,493,393-1,493,662
SE33	6	88,986,820-88,987,106
Penta D	21	45,055,996-45,056,424
D22S1045	22	37,536,303-37,536,407
D6S474	6	112,879,106-112,879,267
D1S1627	1	106,963,665-106,963,777
D6S1017	6	41,677,196-41,677,354
D4S2408	4	31,304,234-31,304,509
D17S1301	17	72,680,935-72,681,088
D14S1434	14	95,308,357-95,308,578
D2S1776	2	169,645,211-169,645,507
D5S2500	5	58,697,193-58,697,344

Table 11. X-chromosome STRs in the test panel.

X-STR	Chromosome	location
DXS8378	X	9,370,226-9,370,429
DXS7132	X	64,655,336-64,655,623
DXS6800	X	78,680,410-78,680,603
DXS6801	X	92,511,172-92,511,301
DXS6809	X	94,938,153-94,938,411
DXS6789	X	95,449,414-95,449,554
DXS7424	X	100,618,816-100,618,983
DXS101	X	101,413,016-101,413,242
GATA172D05	X	113,174,984-113,175,103
HPRTB	X	133,615,405-133,615,691
DXS8377	X	149,566,471-149,566,716
DXS10135	X	9,306,118-9,306,616
DXS10074	X	66,976,953-66,977,449
DXS10101	X	133,654,443-133,654,698
DXS10134	X	149,649,916-149,650,436
DXS7423	X	149,710,903-149,711,089

DXS10011	X	151,188,026-151,188,418
GATA31E08	X	140,234,255-140,234,502
DXS9895	X	7,377,107-7,377,253
DXS981	X	68,197,359-68,197,545
DXS7133	X	109,041,543-109,041,664
DXS6807	X	4,743,382-4,743,648
DXS6795	X	23,244,500-23,244,783
GATA165B12	X	120,877,968-120,878,096
DXS6854	X	128,688,898-128,689,006
DXS9902	X	15,323,616-15,323,787

Table 12. Y-chromosome STRs in the test panel.

Y-STR	Chromosome	Location
DYS456	Y	4,270,942-4,271,090
DYS389I/II	Y	14,612,070-14,612,436
DYS390	Y	17,274,884-17,275,099
DYS458	Y	7,867,840-7,867,983
DYS19	Y	9,521,878-9,522,129
DYS385a/b	Y	20,842,336-20,842,716
DYS385a/b	Y	20,801,456-20,801,824
DYS393	Y	3,131,128-3,131,247
DYS391	Y	14,102,766-14,102,872
DYS439	Y	14,515,188-14,515,408
DYS635	Y	14,379,517-14,379,692
DYS392	Y	22,633,847-22,634,156
Y GATA H4	Y	18,743,528-18,743,664
DYS437	Y	14,466,964-14,467,156
DYS438	Y	14,937,785-14,938,104
DYS448	Y	24,364,964-24,365,273
DYS576	Y	7,053,302-7,053,492
DYS481	Y	8,426,347-8,426,474
DYS549	Y	21,520,078-21,520,317
DYS533	Y	18,393,105-18,393,318
DYS570	Y	6,861,115-6,861,370
DYS643	Y	17,425,985-17,426,129
DYS460	Y	21,050,792-21,050,902
DYS612	Y	15,752,549-15,752,752
DYS449	Y	8,217,985-8,218,232
DYS522	Y	7,415,373-7,415,724

DYS505	Y	3,640,750-3,640,923
DYS627	Y	8,649,930-8,650,266
Amelogenin Y	Y	6,736,679-6,736,894

The markers in Tables 10-12 were combined into one panel with 379 SNPs (see below) and thus represent to date the largest number of forensically-relevant markers to reside in a single multiplex MPS system.

The STR methods and results are described primarily in the STRait Razor Section. SNPs will be discussed separately, although the data for both marker types were generated simultaneously. For evaluation of STR typing, success was based on obtaining a result and then comparing the results with that of standard CE typing. One limitation of this study was that there were only 22 autosomal and 18 Y STRs that had been typed previously using commercial STR kits - Applied Biosystems® AmpFlSTR® Identifiler® PCR Amplification Kit and the Promega® PowerPlex® 16 HS, ESI 17 Pro, and Y23 Systems (Promega Corporation, Madison, WI). However, there were no typing discrepancies between the MPS and CE-based systems for the common STRs which lends support to the reliability of STR typing by MPS. Depending on the sample preparation method, some loci (e.g., D21S11 and DYS393) did not yield results. The failure to type the former locus was likely due to the single-end sequencing format (with GAIIX data) which did not completely cover the large allele lengths for the D21S11 locus. These results were similar to those described by Bornman et al (93). Pair-end reads should and do overcome this limitation (as observed with MiSeq data). The failure to detect the DYS393 locus may be a probe design issue and can be overcome with further work on design (in the final panel). Overall the initial results demonstrated that a large panel of STRs can be typed by MPS, and the results will enable STR analysis for many more markers than can be analyzed simultaneously by CE. However, initial data analyses were difficult and in depth review of results initially were not practical. Therefore, STR typing software was developed to facilitate analysis (94) and now is part of the pipeline that will be used for assessing STR data that facilitate final panel analyses..

V. Software for STR Typing

STRait Razor (STR allele identification tool – Razor)

While the current MPS instruments are capable of providing extensive data, available software tools were limited for identifying forensic STR alleles and calling them with the same nomenclature that has been used for data generated using CE-based systems. Without suitable software for STR analysis the process was tedious and time consuming and comparison of results with current capabilities was difficult. Therefore, this project required STR typing software for MPS data.

One existing software tool, lobSTR (95), uses an algorithm specifically designed to identify STR alleles within MPS data. First, this software analyzes a raw FASTA/FASTQ or BAM input file, detecting reads that contain a STR sequence and the identifying repeat motif. Next, lobSTR

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

aligns the regions that flank the STR sequence to a modified reference sequence. Lastly, the allele(s) are identified based on the number of detected repeat units between the two flanking regions, applying statistical corrections to produce the most likely allele type. This software is limited for compatibility with legacy STR data in that lobSTR identifies only a single simple repeat motif. To allow the software to detect alleles at STRs that have longer, complex or compound repeats, such as those within the D21S11 locus, the user must determine the distinct simple repeats that comprise the complex motif and instruct lobSTR to identify each of these repeats individually. The resulting data must be interpreted altogether to draw conclusions. Thus lobSTR is less applicable for the analysis of a number of forensically-relevant STRs due to their varying repeat motif complexity.

Another approach was introduced by Bornman et al (93) that allows for the detection of STR alleles in MPS data using a different strategy. This method uses the Bowtie short read aligner (96) to align raw MPS reads to an "*in silico* reference," which is a user-generated FASTA file containing the full sequence of each allele at each STR locus. To reduce erroneous allele calls, reads are filtered so that only those encompassing the entire repeat region defined in the reference file are used for allele typing. Allele calls are made using a heuristic decision model based on Fisher's Exact Test, and probability values are given for each allele call. This software is effective for identifying STR alleles in sequence data, but requires substantial, prior knowledge of allelic sequence variation. As a result, novel alleles or allelic variants, or those for which there are no documented sequence data, may be missed.

For the study herein, a novel STR typing software was developed named STRait Razor (the STR allele identification tool - Razor) (94). The purpose was to facilitate STR allele calling that is compatible with CE-based STR typing nomenclature so that evaluation and validation can be effectively assessed. The software is a Linux-based Perl script that identifies alleles at STR loci based on the length of the repeat sequence, a method that is conceptually similar to the length-based allele detection offered by CE. This software is capable of handling repeat motifs ranging from simple to complex, and it does not require a reference composed of extensive allelic sequence data. As a result, the allele call results are consistent with those of current CE-based methods, and it is not confounded by unexpected sequence variation within repeats. In its first iteration STRait Razor could identify alleles at 44 forensically-relevant STR loci, and other loci could be configured readily.

The details of STRait Razor were described in Warshauer et al (94). The software can be downloaded from the publication site. Briefly the software identifies only reads containing both a leading and trailing flanking region surrounding the repeat sequence of a designated locus (or loci) and these sequences are extracted from the raw FASTQ sequence file(s) using the AGREP function (97). This first step ensures that the extracted reads encompass the full repeat sequence, as partial repeat sequence data cannot be used for accurate allele-calling. With the reads containing the complete repeat sequence, the regions adjacent to the repeat sequence on each side are trimmed away, leaving only the repeat regions (a future version will capture variation that resides in the flanking regions). Then, the reads are filtered based on the presence of the repeat motif specific to the STR locus; thus the majority of irrelevant sequence data that may have been inadvertently captured are discarded. Next, the number of nucleotides in each repeat

sequence is counted and compared with the expected lengths of alleles at that locus, based on the repeat motif (Figure 6).



Figure 6. STRait Razor algorithm. The repeat region is shown in bold, capitalized font, while the flanking regions are shown in plain, lowercase font. Surrounding sequences are shown in plain, capitalized font.

For example, at a STR locus with a tetranucleotide repeat motif, a repeat sequence consisting of 48 bases would indicate the presence of a "12" allele, while a sequence of 50 bases would indicate the presence of a "12.2" allele. Alleles can be called in this fashion regardless of intra-sequence variation or repeat motif complexity, and the length-based method of detection, to date, has been concordant with CE results. The repeat region sequences of the called alleles then are sorted by length and written to a text file specific to the STR locus. With this data output, analysts can observe every nucleotide and evaluate any variations within the repeat sequences, as desired. Finally, a colon-delimited text file is generated that lists the alleles called at each STR locus, including the number of reads in which each allele was detected. These read count values can be used as a measure of abundance, similar to the manner RFU values in electropherograms are used for quantitative assessments.

STRait Razor is designed to analyze both single-end and paired-end data. Thus, the program accepts either single input files (single-end reads) or dual input files (paired-end reads), and recognizes STR loci in both forward and reverse complement forms. The speed with which STRait Razor can provide allele calls is directly related to the size of the input file(s), as well as the depth of reads that contain the queried STR loci. The software utilizes PPSS, the (Distributed) Parallel Processing Shell Script (98), which allows STRait Razor to analyze one

STR locus per available processor core in parallel, thus reducing the amount of time needed for analysis. Also, the user is able to choose whether STRait Razor detects only autosomal STR alleles, Y-chromosome STR alleles, or both. This option can further reduce analysis time and may be useful in cases wherein analysts wish to investigate only a subset of the loci recognized by STRait Razor, such as the typing of a known female reference sample would not require Y STR analyses. The information required by STRait Razor to detect STR alleles is provided in a modular format, and a workbook is provided to allow users to easily add other STR loci to the modules. Flanking regions queried by STRait Razor are each 12 bases long. These lengths allow for sufficient nucleotide diversity for specificity of target loci. Users may choose to use different flanking region sequences, as well as different flanking lengths. While most flanking regions used by STRait Razor are directly adjacent to the repeat regions, 6 sets (those for loci D1S1656, Penta D, DYS385, DYS393, DYS481, and DYS635) were selected at different proximities to allow for increased specificity.

It should be noted that Fordyce et al. (99) independently developed a software tool that functions similar to that of STRait Razor, isolating the repeat region of interest and performing length-based allele typing. However, the algorithm was designed for use with the Roche® Genome Sequencer FLX™ and is only able to analyze FASTA files consisting of sequence data that contain Roche® Molecular Identifier (MID) tags. FASTQ files tend to be the MPS format of choice. In addition, their software only is able to identify alleles at 5 STR loci, compared with the 44 STR loci detected by STRait Razor. More recently, Van Este et al (100), subsequent to the development of STRait Razor, produced another effective STR typing software for MPS data. Therefore, there now are two viable methods for calling STR alleles from MPS data.

While STRait Razor was tested only on raw FASTQ files output by the Illumina® instruments, the software should, in theory, maintain compatibility with the raw read files generated by the Ion Torrent PGM System as well. To test the efficiency and accuracy of STRait Razor, a concordance study was performed wherein allele calls made by the software were compared with CE results. This study allowed verification of the performance of STRait Razor and also permitted evaluation of STR data generated by MPS.

1. Methods for STRait Razor and Data Analysis of STRs from the Large Multiplex Panel

Samples

DNA was extracted using the Qiagen® QIAamp® DNA Mini kit, following the manufacturer's recommendations. The quantity of extracted DNA was determined using the Applied Biosystems® Quantifiler™ Human DNA Quantification Kit on an Applied Biosystems® 7500 Real-Time PCR System, according to the manufacturer's protocol. The quantity of extracted DNA from some samples also was determined using a Qubit® 2.0 Fluorometer.

CE-based Typing

Amplification was performed using the reagents from the Applied Biosystems® AmpFlSTR® Identifiler® PCR Amplification Kit and the Promega® PowerPlex® 16 HS, ESI 17 Pro, and Y23

Systems (Promega Corporation, Madison, WI), on an Applied Biosystems® GeneAmp® PCR System 9700 thermal cycler, according to the manufacturer's recommendations. These various kits allowed for the typing of the following STR loci: CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D5S818, D7S820, D8S1179, FGA, TH01, TPOX, vWA, PENTA D, PENTA E, D10S1248, D12S391, D1S1656, D22S1045, D2S441, SE33, DYS19, DYS385, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS481, DYS533, DYS549, DYS570, DYS576, DYS635, DYS643, and GATA H4. CE was performed on an Applied Biosystems® 3130xl Genetic Analyzer (ThermoFisher) using POP-4™ polymer (ThermoFisher) and analyzed using Applied Biosystems® GeneMapper® ID v3.2 software (ThermoFisher), according to the manufacturer's recommended protocol.

Sample Preparation for Large Marker Panel

Library preparation was performed using the Illumina® TruSeq™ Custom Enrichment protocol (Illumina). Custom probes were designed to target the panel loci using the DesignStudio software (Illumina). The TruSeq library chemistry was selected initially because no PCR amplification is required. Therefore, PCR generated errors were not a primary consideration for potential artifacts. PCR errors were not a large concern initially as STR typing was assessed for length variation solely (most PCR errors would be substitutions). However, sequence variation will add another dimension that will facilitate STR profile interpretation. As one proceeds forward and develops STR typing on more challenged samples, a PCR enrichment may be required to attain desired levels of sensitivity. There are two limitations with the TruSeq chemistry: 1) it requires a good amount of template DNA (50 ng -1 ug); and 2) it is a laborious method. Since this initial panel is designed for reference sample typing, the amount of template DNA was not so limiting and thus was not an issue for using the TruSeq library preparation method. However, for casework applications (not a focus of this project), where template DNA is limited, an alternate library preparation method will be necessary. After capture, single-end sequencing (1x-146) was carried out on the GAIIx™ and paired-end sequencing (2x-151) was performed on MiSeq™ sequencing platform. In order to call genotypes, an alignment program (BWA) with a virtual allelic ladder and some internal scripts as well as a preliminary program from Illumina was used to parse out the data. STRait Razor was developed to facilitate allele calling of STR alleles from MPS data.

Library Preparation and Sequencing

Library preparation prior to sequencing was performed using either the Illumina® TruSeq™ Custom Enrichment protocol or the Agilent Technologies HaloPlex™ Target Enrichment protocol (Agilent Technologies). Using the DesignStudio (Illumina, Inc.) and SureDesign (Agilent Technologies, Inc.) software, respectively, custom probes were designed to target the forensically-relevant STR loci (SNPs also were included in the DesignStudio panel, see below). Paired-end sequencing was carried out on the GAIIx™ (Illumina) and MiSeq sequencing platforms. For these trials, the read lengths employed by these instruments were 2x146 and 2x251, respectively. Sample 1 was prepared using the HaloPlex chemistry and sequenced on both the GAIIx and MiSeq instruments. The sample also was prepared using the TruSeq

chemistry, and subsequently sequenced on the MiSeq. Sample 2 was prepared using the HaloPlex chemistry and sequenced on the GAIIx. Samples 3, 4, and 5 were prepared using the TruSeq chemistry and sequenced on the MiSeq. Following sequencing, the GAIIx output *bcl* files were demultiplexed and converted to a single FASTQ file using CASAVA v1.8.2 (101). The MiSeq output was automatically converted to FASTQ format by the MiSeq Reporter software (102). These FASTQ files served as the input for STRait Razor. The software was designed to detect the following forensic STR loci: CSF1PO, TPOX, D2S441, D3S1358, D5S818, D13S317, D18S51, D16S539, D7S820, D8S1179, TH01, vWA, D21S11, FGA, D2S1338, D19S433, PENTA D, PENTA E, D10S1248, D12S391, D1S1656, D22S1045, DYS389I/II, DYS390, DYS456, DYS19, DYS458, DYS437, DYS438, DYS448, GATA H4, DYS391, DYS392, DYS393, DYS439, DYS481, DYS533, DYS549, DYS570, DYS576, DYS643, DYS385, and DYS635. For these trials, allele-calling was performed using STRait Razor's default flank recognition settings (1 allowable substitution and no allowable insertions or deletions). The server used for STRait Razor analysis was a Dell™ PowerEdge™ R900 blade server, with 64 GB DDR3 ECC RAM and 4 Quad-Core Intel® Xeon® E7430 CPUs (2.13 GHz each).

VI. STR Typing Results and Discussion

The results of this portion of the study provide support for 1) the reliability of STR typing by MPS; and 2) the functionality and accuracy of STRait Razor for calling STR alleles. The alleles detected by CE methodology were compared with the allele call output files generated by STRait Razor from MPS analyses. To be considered concordant, alleles detected via CE had to be detected by STRait Razor, based on the presence of the allelic sequence data in the input FASTQ file. At this time, loci analyzed with MPS but without concordant data could not be verified as correctly called. They only could be scored as being typed. Based on the various combinations of library preparation methods and the two sequencing platforms, 7 comparisons with CE data were made for 5 different samples resulting in a total of 427 alleles being compared. The allele calls made by STRait Razor were completely concordant with the genotype results generated by the CE method. Of the 427 alleles compared, 403 alleles were detected, with 100% concordance. For the 24 alleles not detected by the software, manual analysis of the FASTQ input files revealed that there were no sequence reads for these alleles in which the full repeat regions (including the surrounding flanking sequences) were present. These undetected alleles were not represented in the sequencing data most likely because of the library preparation method (e.g., random shearing of genomic DNA) and/or the read length used, and thus could not be recognized. These instances were not considered evidence of discordance. The time required for analysis, using a paired-end analysis method that detected both autosomal and Y chromosome STR alleles, ranged from 16 minutes (for dual 395 MB input files) to 285 minutes (for dual 7.9 GB input files) on the 16-core server. The results of this study demonstrated the efficiency and accuracy of STR allele detection by MPS and analysis with STRait Razor. In addition, the output text file revealed underlying sequence variants within repeat and flanking areas. These variants go undetected by CE typing (103). In fact, a variant was observed at the 15 allele of the D3S1358 locus in one sample (Table 13). Therefore, underlying sequence variation is additional information that could increase the discrimination power of STR typing. While with a full profile an increase in discrimination power may not have a practical impact for reference sample characterization, such information could be invaluable for partial profiles (e.g., with degraded samples) or in mixtures. For

mixtures, SNP variants in the repeat region could assist in mixture deconvolution of a stutter peak with that of true alleles from a minor contributor. Overall, these concordant data support the accuracy of STR allele typing by MPS and the functionality of STRait Razor.

Library Preparation Evaluation

The STR data comparison revealed the relationship between software, library preparation chemistries, and sequencing platforms used to produce the sequence information. Read length is an important factor (followed by coverage) that impacts STR locus and allele detection of MPS. A summary of results are displayed Tables 13 and 14.

	Detection Method											
	Sample 1				Sample 2		Sample 3		Sample 4		Sample 5	
	CE	STRait Razor (TruSeq™ prep, GAIlix™)	STRait Razor (HaloPlex™ prep, GAIlix™)	STRait Razor (HaloPlex™ prep, MiSeq™)	CE	STRait Razor (HaloPlex™ prep, GAIlix™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)
CSF1PO	12	12 (332, 315)	12 (1135, 976)	12 (324, 322)	12	12 (2445, 4155)	10, 11	10 (89, 84), 11 (54, 75)	12	12 (304, 310)	10, 12	10 (152, 189), 12 (162, 145)
D13S317	12	12 (104, 96)	12 (639, 0)	12 (217, 214)	8, 11	8 (2513, 4888), 11 (1995, 4377)	11, 12	11 (49, 61), 12 (50, 48)	12	12 (220, 209)	8, 11	8 (101, 107), 11 (86, 76)
D16S539	10, 12	10 (86, 70), 12 (64, 43)	10 (886, 2896), 12 (686, 1546)	10 (181, 135), 12 (174, 138)	9, 12	9 (1733, 7394), 12 (1165, 3209)	8, 13	8 (62, 69), 13 (51, 56)	9, 12	9 (121, 110), 12 (82, 77)	11, 13	11 (117, 98), 13 (91, 73)
D18S51	15, 21	15 (25, 29), 21 (16, 17)	15 (598, 3258), 21 (341, 541)	15 (123, 123), 21 (106, 105)	15, 16	15 (957, 5711), 16 (819, 4734)	14, 16	14 (20, 16), 16 (13, 17)	15, 18	15 (34, 22), 18 (17, 17)	12	12 (94, 79)
D19S433	12, 14	12 (39, 19), 14 (23, 13)	12 (589, 2620), 14 (542, 1543)	12 (141, 138), 14 (126, 133)	13, 14	13 (1634, 6001), 14 (1423, 4149)	13, 14	13 (16, 5), 14 (10, 11)	13	13 (44, 33)	12, 15	12 (12, 19), 15 (28, 17)
D21S11	29, 30	[-], [-]	[-], [-]	29 (70, 17), 30 (57, 14)	29, 31	[-], [-]	29	29 (17, 9)	29	29 (38, 25)	28, 29	28 (16, 11), 29 (16, 11)
D2S1338	18, 25	18 (19, 18), 25 (9, 0)	18 (80, 1), [-]	18 (54, 52), 25 (29, 25)	20, 25	20 (137, 1), [-]	18	18 (53, 53)	17, 24	17 (74, 67), 24 (35, 35)	20, 23	20 (70, 70), 23 (55, 37)
D3S1358	15	15 (109, 110)	15 (449, 4079)	15 (439, 395)	15, 16	15 (1046, 4579), 16 (852, 3995)	16, 18	16 (32, 29), 18 (36, 29)	16, 18	16 (71, 74), 18 (78, 66)	14, 17	14 (99, 93), 17 (72, 84)
D5S818	11	11 (76, 66)	[-]	11 (66, 67)	11	[-]	11, 12	11 (20, 23), 12 (30, 26)	11	11 (101, 86)	12, 13	12 (56, 62), 13 (45, 34)
D7S820	9, 12	9 (1, 2), 12 (5, 4)	9 (0, 1754), 12 (0, 1650)	9 (70, 70), 12 (57, 57)	10	10 (0, 7733)	10, 11	10 (4, 3), 11 (3, 4)	9, 13	9 (3, 7), 13 (8, 8)	11	11 (17, 15)
D8S1179	13	13 (98, 87)	13 (1722, 5068)	13 (335, 220)	13	13 (3941, 10527)	10, 12	10 (36, 38), 12 (24, 25)	12, 14	12 (52, 47), 14 (48, 49)	13, 16	13 (68, 78), 16 (58, 41)
FGA	20, 21	20 (36, 25), 21 (26, 20)	20 (770, 3819), 21 (581, 3419)	20 (168, 129), 21 (146, 134)	19, 21	19 (1019, 5066), 21 (890, 4849)	23, 25	23 (12, 16), 25 (12, 13)	22, 24	22 (40, 30), 24 (18, 23)	20	20 (84, 71)
TH01	9, 9.3	9 (160, 146), 9.3 (132, 178)	9 (3172, 5571), 9.3 (2893, 5493)	9 (260, 255), 9.3 (297, 297)	8, 9.3	8 (5701, 8471), 9.3 (4963, 8350)	7	7 (150, 162)	9.3	9.3 (287, 292)	9.3	9.3 (393, 390)
TPOX	8, 9	8 (90, 96), 9 (99, 80)	8 (4832, 5208), 9 (4488, 4710)	8 (527, 479), 9 (475, 428)	11	11 (11043, 15943)	8, 11	8 (35, 34), 11 (32, 29)	8, 12	8 (113, 105), 12 (84, 86)	8	8 (176, 185)
WVA	16, 17	16 (53, 37), 17 (39, 24)	16 (299, 0), 17 (213, 0)	16 (55, 55), 17 (36, 36)	15, 20	15 (669, 0), 20 (0, 3)	17	17 (49, 56)	14, 17	14 (60, 60), 17 (57, 52)	15, 17	15 (63, 65), 17 (56, 61)
Penta D	10	10 (24, 35)	10 (388, 0)	10 (180, 0)	14, 15	14 (214, 0), [-]	9, 11	9 (11, 14), 11 (12, 10)	12, 14	12 (26, 23), 14 (25, 20)	9, 12	9 (23, 24), 12 (29, 30)
Penta E	11, 12	11 (6, 7), 12 (9, 5)	11 (98, 121), 12 (123, 104)	11 (105, 87), 12 (107, 88)	10, 11	10 (290, 993), 11 (243, 191)	5, 7	5 (10, 7), 7 (6, 5)	11, 12	11 (9, 4), 12 (11, 6)	7, 14	7 (20, 19), 14 (8, 7)
D10S1248	13, 14	13 (106, 94), 14 (69, 67)	13 (672, 4537), 14 (329, 3613)	13 (192, 199), 14 (159, 171)	13	13 (3545, 17065)	12, 13	12 (81, 73), 13 (71, 69)	13, 14	13 (134, 127), 14 (152, 134)	16	16 (211, 211)
D12S391	15, 17	15 (50, 50), 17 (35, 31)	15 (1464, 548), 17 (1101, 447)	15 (167, 117), 17 (162, 106)	21	21 (2123, 0)	15, 24	15 (37, 31), 24 (23, 19)	17, 21	17 (90, 82), 21 (84, 79)	18, 20	18 (68, 73), 20 (58, 63)
D1S1656	16, 18.3	16 (63, 47), 18.3 (37, 52)	16 (499, 0), 18.3 (468, 0)	16 (59, 75), 18.3 (78, 41)	11, 12	11 (2262, 2848), 12 (1872, 2284)	16.3, 18.3	16.3 (33, 24), 18.3 (25, 29)	15, 17.3	15 (79, 75), 17.3 (67, 52)	12, 16	12 (98, 83), 16 (76, 67)
D22S1045	16, 17	16 (10, 16), 17 (7, 16)	16 (901, 2637), 17 (718, 1639)	16 (107, 105), 17 (108, 104)	15, 16	15 (2042, 5133), 16 (1415, 3535)	15, 16	15 (15, 14), 16 (13, 15)	16	16 (35, 38)	11, 16	11 (94, 94), 16 (40, 40)
D2S441	11, 15	11 (72, 49), 15 (32, 37)	11 (291, 3863), 15 (0, 3685)	11 (124, 145), 15 (113, 167)	11.3, 14	11.3 (615, 7807), 14 (0, 7024)	11, 14	11 (74, 71), 14 (68, 64)	10, 11.3	10 (106, 100), 11.3 (149, 137)	10, 11	10 (131, 125), 11 (130, 138)
Total Alleles	38	36	34	38	37	32	40	40	37	37	38	38

Table 13. Comparison of CE allele calls and STRait Razor results – Autosomal STRs. Alleles detected by both CE and STRait Razor analysis of SGS data are shown in bold in the columns for each sample. The numbers of reads in which an allele was detected by STRait Razor are listed in parentheses next to the respective allele. The first number in parentheses represents the abundance of the allele in Read 1 of the paired-end sequencing run, while the second number represents the abundance of the allele in Read 2. Alleles not detected by STRait Razor due to lack of relevant sequence data are denoted by “[-].”

	Detection Method											
	Sample 1				Sample 2		Sample 3		Sample 4		Sample 5	
	CE	STRait Razor (TruSeq™ prep, GAIlix™)	STRait Razor (HaloPlex™ prep, GAIlix™)	STRait Razor (HaloPlex™ prep, MiSeq™)	CE	STRait Razor (HaloPlex™ prep, GAIlix™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)
DYS19	14	14 (7, 11)	14 (231, 1221)	14 (61, 62)	15	15 (6, 2341)	14	14 (8, 18)	16	16 (13, 13)	14	14 (14, 15)
DYS385	11, 13	11 (12, 12), 13 (12, 11)	[-], [-]	[-], [-]	11, 14	[-], [-]	11, 14	11 (5, 3), 14 (4, 5)	11, 14	11 (22, 29), 14 (8, 19)	11, 14	11 (12, 13), 14 (24, 16)
DYS389I	13	13 (102, 99)	13 (423, 1)	13 (26, 0)	13	13 (1605, 0)	13	13 (54, 47)	13	13 (137, 137)	13	13 (120, 103)
DYS389II	29	[-]	[-]	29 (23, 0)	30	[-]	29	29 (9, 3)	28	28 (25, 20)	29	29 (13, 17)
DYS390	24	24 (14, 16)	24 (115, 3495)	24 (115, 126)	25	25 (259, 3386)	24	24 (16, 19)	23	23 (37, 43)	23	23 (40, 37)
DYS391	10	10 (175, 167)	10 (952, 39)	10 (99, 24)	10	10 (3179, 1431)	10	10 (73, 80)	10	10 (180, 182)	12	12 (166, 165)
DYS392	13	13 (3, 8)	13 (885, 1965)	13 (82, 78)	11	11 (1466, 2850)	13	13 (7, 8)	13	13 (7, 8)	13	13 (11, 10)
DYS393	13	13 (9, 2)	13 (0, 360)	13 (14, 13)	14	14 (0, 1023)	13	13 (2, 7)	13	13 (2, 3)	13	13 (10, 10)
DYS437	15	15 (85, 70)	15 (0, 4020)	15 (247, 238)	14	14 (0, 11064)	15	15 (77, 77)	15	15 (148, 133)	14	14 (141, 146)
DYS438	12	12 (42, 36)	12 (324, 285)	12 (62, 32)	11	11 (884, 871)	12	12 (48, 49)	12	12 (79, 68)	12	12 (96, 93)
DYS439	12	[-]	12 (428, 2296)	12 (134, 78)	10	10 (1789, 6072)	12	12 (2, 0)	11	11 (2, 2)	13	13 (3, 1)
DYS448	19	[-]	[-]	19 (17, 5)	19	[-]	19	19 (11, 3)	19	19 (21, 16)	18	18 (12, 11)
DYS456	15	15 (10, 13)	15 (523, 1402)	15 (80, 54)	14	14 (2723, 6296)	15	15 (13, 10)	16	16 (11, 9)	15	15 (22, 21)
DYS458	17	17 (10, 6)	17 (56, 258)	17 (31, 21)	15	15 (152, 1300)	19	19 (11, 9)	17	17 (13, 12)	17	17 (17, 24)
DYS481	22	22 (19, 21)	22 (227, 1943)	22 (150, 101)	23	23 (663, 3830)	25	25 (20, 18)	23	23 (30, 43)	22	22 (50, 45)
DYS533	12	12 (26, 36)	12 (184, 0)	12 (30, 10)	12	12 (511, 0)	12	12 (37, 37)	12	12 (34, 37)	14	14 (41, 27)
DYS549	13	13 (44, 49)	13 (743, 0)	13 (151, 101)	12	12 (2649, 0)	13	13 (39, 38)	13	13 (46, 58)	13	13 (60, 62)
DYS570	17	17 (44, 51)	17 (646, 0)	17 (76, 28)	20	20 (777, 0)	18	18 (64, 69)	17	17 (73, 66)	17	17 (145, 134)
DYS576	19	19 (45, 43)	19 (0, 341)	19 (3, 11)	17	17 (0, 512)	19	19 (34, 22)	18	18 (74, 74)	18	18 (65, 57)
DYS635	24	24 (10, 5)	24 (220, 71)	24 (18, 20)	25	25 (774, 700)	23	23 (18, 15)	24	24 (33, 29)	23	23 (30, 27)
DYS643	10	10 (43, 25)	10 (2392, 989)	10 (221, 111)	10	10 (3314, 1407)	11	11 (10, 15)	11	11 (34, 30)	10	10 (46, 43)
GATA H4	12	12 (23, 21)	12 (290, 2196)	12 (97, 93)	12	12 (511, 3795)	11	11 (21, 27)	11	11 (34, 47)	11	11 (33, 40)
Total Alleles	23	20	19	21	23	19	23	23	23	23	23	23

Table 14. Comparison of CE allele calls and STRait Razor results – Y-Chromosome STRs. Alleles detected by both CE and STRait Razor analysis of SGS data are shown in bold in the columns for each sample. The numbers of reads in which an allele was detected by STRait Razor are listed in parentheses next to the respective allele. The first number in parentheses represents the abundance of the allele in Read 1 of the paired-end sequencing run, while the second number represents the abundance of the allele in Read 2. Alleles not detected by STRait Razor due to lack of relevant sequence data are denoted by “[-].”

The HaloPlex chemistry for library preparation relies on enzymatic cleavage (104). The benefit of specific site digestion is creation of fragments with consistent start and end points. This feature of HaloPlex makes it a good candidate for further evaluation of a library preparation method. The limitation of such a method is that the cleavage sites are based on the restriction endonucleases employed. Depending on the length of the allele in question and the position of the repeat region within the resulting fragment(s), it is possible for sequence reads to be produced that partially span the repeat region. If the sequencing start point of the fragment is too distant from the repeat region, the read may not extend through the entire repeat region of an allele. The cleavage positions cannot be changed without substituting enzymes. A change in enzymes may impact other loci that were typeable with the current restriction enzyme cocktail. However, HaloPlex is a system that can be automated and provides high sample throughput, which could be desirable for a databasing laboratory. It requires ~200ng of DNA. In addition, having defined starting points allows for kit development with defined locus success parameters (assuming good quality DNA). Given that many more STRs can be typed by MPS, one may consider giving up a couple of current “core” STR loci as the loss can be more than compensated by sheer number of additional markers. The overall practicality of design should be a criterion for long term functionality.

If a repeat region is situated toward the beginning of a HaloPlex fragment, the allele is likely to be detected in one direction of a paired-end analysis. However, when the reads are sequenced from the opposite direction and the repeat region is oriented toward the end of the read, the region may not be completely encompassed. This situation will be dependent on read length and was observed in loci such as D7S820 and vWA, where the alleles were detected only in one set of paired-end reads and not the other. However, correct calls were made. Some library preparation redesign may overcome the truncated repeat region reads. An example of an allele not detected due to incomplete repeat region traversal is at locus D2S1338 in one sample (HaloPlex preparation, GAIIX sequencing). For this locus, the “18” allele was called, but the “25” allele was too long to be covered completely by the allowable sequencing read. When this same sample was sequenced on the MiSeq platform using a longer read length, the “25” allele was detected.

The TruSeq chemistry (105) is less prone to HaloPlex specific cleavage site issues because DNA is fragmented randomly for a much more varied positioning of repeat regions within the resulting fragments. Therefore, there is a greater likelihood of at least some reads encompassing the entire repeat region. This design supports the finding of the majority of the alleles following TruSeq preparation that were not detected when prepared using HaloPlex. Despite this beneficial feature, non-enzymatic random fragmentation employed by the TruSeq chemistry resulted in lower read counts for some alleles in comparison with HaloPlex, due to the fewer resulting fragments containing the complete repeat region. The random fragmentation method simply may not generate as many fragments that contain the complete repeat region of interest (i.e., resulting in lower coverage). This limitation may explain the undetected alleles in a sample at loci DYS439 and DYS448 following TruSeq preparation and GAIIX sequencing. Coverage depth differences in the results of this study (data not shown for the preliminary panel), however, also may be

explained by the fact that the regions targeted by the TruSeq kit for these trials were, by design, approximately 100 times larger than those targeted by the HaloPlex kit.

Because a number of STR loci (other than those described above) in the panel did not have concordance data with another system, correct typing could not be demonstrated for all loci. Only success in obtaining results could be recorded. Typing results were greater with the MiSeq runs compared with those of the GAIIX because of the different read lengths 2x251 and 2x146, respectively. If a read length is too short, the sequence may not traverse the repeat region and no definitive repeat count can be obtained. Therefore, the final designed panel will be sequenced solely on the MiSeq with a longer chemistry 2x251. In addition, at the onset of the large panel development, tools were not in place to analyze the output data beyond simple typing results. Subsequent to the first iteration, workbooks and STRait razor were developed to obtain more information on performance. The final panel will have a full data analysis on success of typing, concordance testing, depth of coverage, allele coverage ratios, and strand bias.

VII. STRait Razor v 2.0

Since its initial release, STRait Razor has been employed by a number of laboratories with positive results. In-house needs and resulting feedback were considered strongly to enhance the original software. These new features include an expanded default set of detectable STR loci (autosomal, X, and Y markers) that covers all the STR loci in the proposed panel, an enhanced custom locus list configuration tool, a novel output sorting method that highlights unique sequences for each allele, and a genotyping tool that emulates traditional electropherogram data. With these improvements, STRait Razor v2.0 offers users a much wider, more flexible range of analysis options and greater ease of use.

New Features

STRait Razor v2.0 includes an expanded locus configuration file which it can use to detect a wider range of forensically-relevant STR markers. Previously, the default locus definition file included 44 STR loci (22 autosomal STRs and 22 Y-chromosome STRs). An additional 42 markers have been added, for a total of 86 markers, which include: 9 new autosomal STRs (D14S1434, D17S1301, D1S1627, D2S1776, D4S2408, D5S2500, D6S1017, D6S474, and SE33), 26 new X-chromosome STRs (DXS10011, DXS10074, DXS101, DXS10101, DXS10134, DXS10135, DXS6789, DXS6795, DXS6800, DXS6801, DXS6807, DXS6809, DXS6854, DXS7132, DXS7133, DXS7423, DXS7424, DXS8377, DXS8378, DXS981, DXS9895, DXS9902, GATA165B12, GATA172D05, GATA31E08, and HPRTB), 6 new Y-chromosome STRs (DYS449, DYS460, DYS505, DYS518, DYS522, and DYS612), and Amelogenin. The allelic information contained in the locus configuration file was compiled using data from a variety of online databases, such as STRbase (106), ChrX-STR.org 2.0 (107), NCBI (108), and the Sorenson Molecular Genealogy Foundation database (109). Users have the option of choosing the marker type to analyze with STRait Razor v2.0 by using the “-*typeselection*” argument (AUTO, X, Y, or ALL). This feature allows the analysis to be tailored to the specific goals of the testing and, depending on which option is selected, can reduce the

time required for analysis. As with the initial version of STRait Razor, custom locus configuration files can be created for the program using the included Microsoft Excel workbook. Therefore, analysis can be performed on STR loci that are not included in the default set.

STRAit Razor v2.0 includes an enhanced locus configuration workbook for the creation of custom allelic definition files. The workbook allows for the generation of a locus configuration file containing up to 10 STR loci, although the workbook can be modified to include more. Alternatively, multiple configuration files may be generated and concatenated to produce larger, more comprehensive files. The locus configuration workbook is designed to convert locus information entered by the user for any STR locus type (autosomal, X-chromosome, or Y-chromosome) to the proper format required by STRait Razor v2.0 for analysis. The workbook has been redesigned for ease of use, with features such as automatic generation of reverse complement leading and trailing flanking data based on user input, and auto-conversion of lowercase entries to the proper uppercase format.

STRAit Razor v2.0 is designed to yield STR allele calls for each locus, as well as sequence data for all alleles so that intra-repeat variation can be detected. In its initial release, STRait Razor output sequence data for each allele in a locus-specific file, sorted by repeat region length. While useful, these data were unwieldy to interpret because sequence data from each individual read were appended to a sequence file when captured, resulting in a large amount of redundant sequence information. STRait Razor v2.0 simplifies the sequence output so that only unique sequences for each allele are displayed, and the results are sorted based on the total read count for each sequence (Figure 7). This output results in a clear and concise sequence file for each locus that can be interpreted quickly for intra-repeat nucleotide variation.

Original Sequence Data			New Sequence Data		
LOCUSA:7	28 bases	AGATAGATAGATAGATAGATAGATAGAT	LOCUSA:7	28 bases	AGATAGATAGATAGATAGATAGATAGAT 4
* LOCUSA:7	28 bases	AGATAGATAGGTTAGATAGATAGATAGAT	* LOCUSA:7	28 bases	AGATAGATAGGTTAGATAGATAGATAGAT 3
** LOCUSA:7	28 bases	AGATAGATAGATAGATAGATACATAGAT	** LOCUSA:7	28 bases	AGATAGATAGATAGATAGATACATAGAT 2
LOCUSA:7	28 bases	AGATAGATAGATAGATAGATAGATAGAT			
* LOCUSA:7	28 bases	AGATAGATAGGTTAGATAGATAGATAGAT			
* LOCUSA:7	28 bases	AGATAGATAGGTTAGATAGATAGATAGAT			
LOCUSA:7	28 bases	AGATAGATAGATAGATAGATAGATAGAT			
LOCUSA:7	28 bases	AGATAGATAGATAGATAGATAGATAGAT			
** LOCUSA:7	28 bases	AGATAGATAGATAGATAGATACATAGAT			

Figure 7. Sequence Data Sorting. The original sequence output from STRait Razor, using an example locus named “LOCUSA” (AGAT repeat unit), is shown on the left. Sequences were appended as they were captured. Sequences denoted with (*) contain an A/G SNP, while sequences denoted with (**) contain a G/C SNP. Sequence output from STRait Razor v2.0 is shown on the right. With this update, unique sequences were identified, counted, compiled, and finally sorted based on the total read count.

The first version of STRait Razor enabled users to generate a substantial amount of information on both the alleles and underlying sequence variants. While this updated version of STRait Razor

is considerably more efficient, the tremendous amount of data generated requires a toolset for subsequent analysis. To further facilitate data analysis, a set of Excel-based workbooks have been developed.

The first workbook, RazorGenotyper, converts the output files "RawSTRcallsR1" and "RawSTRcallsR2" into final genotypes. Users can set thresholds for coverage and sister allele balance to ensure that accurate genotypes are generated. Data from multiple samples can be exported and compiled via embedded macros. These exported data then can be further visualized using the second workbook.

STRait Razor Histogram Generator separates the output data of RazorGenotyper into an "allele table" of all loci. These data then are displayed as histograms showing all read variants observed (e.g., alleles, stutter, and PCR artifacts). These charts are parsed into "Autosomal," "Y," and "X" STR tabs. The autosomal tab also contains Amelogenin and is divided into "Core" loci (i.e., loci contained in either PowerPlex Fusion (Promega Corp.) or GlobalFiler (ThermoFisher)) and "Additional Loci" (i.e., autosomal loci not found in either kit). Macros included, but not active, allow a user to uniformly change the axes of all charts to visualize locus-to-locus balance.

The final workbook included in the toolset, STRait Razor_SNP ID Tool, converts the *LOCUS.SEQUENCES* files into a table showing the top 20 sequence variants at each locus. First, the user must transfer the *LOCUS.SEQUENCES* file from each locus of interest into a single folder. Next, the user can run the included 'SeqCompile.pl' script to combine all loci into a single file. The data from this file then are pasted into the STRait Razor_SNP ID Tool. Finally, data are displayed by locus showing the most relevant sequence variants. These data for all loci can be exported and compiled via embedded macros for ease of use.

Concordance Testing STRait Razor v2.0

The accuracy and reliability of STRait Razor have been reported (94). Therefore, concordance testing was performed to verify that STRait Razor v2.0, with its updates, provided the same allele calls as it did in its first release. The same 7 sequence datasets used in the initial testing phase were re-analyzed using STRait Razor v2.0.

Allele calls made by the updated software were compared to those made by the initial version of the software, and new allele calls resulting from use of the larger default locus configuration file were noted. The time required for analysis with the wider range of detectable loci also was determined. Additionally, genotyping and histogram generation were performed to validate the effectiveness of these tools.

Results of Testing STRait Razor v2.0

STRait Razor v2.0 yielded identical allele calls from all 7 sequence datasets with regard to the loci previously analyzed. The read counts generated for each allele detected by STRait Razor v2.0 were 100% concordant with those indicated by the original version of the software. The testing process also demonstrated the wider range of locus detection afforded by the new

expanded locus configuration file. For dataset 1, a total of 26 additional alleles were detected at 20 new loci (7 autosomal, 9 X-chromosome, 3 Y-chromosome, and Amelogenin). Dataset 2 yielded 25 additional alleles that were detected at 19 new loci (6 autosomal, 10 X-chromosome, and 3 Y-chromosome). For dataset 3, a total of 34 additional alleles were detected at 28 new loci (7 autosomal, 17 X-chromosome, and 4 Y-chromosome). For dataset 4, a total of 27 additional alleles were detected at 20 new loci (7 autosomal, 10 X-chromosome, and 3 Y-chromosome). Dataset 5 yielded 35 additional alleles at 29 new loci (8 autosomal, 16 X-chromosome, 4 Y-chromosome, and Amelogenin), while a total of 40 additional alleles at 35 new loci (8 autosomal, 21 X-chromosome, 5 Y-chromosome, and Amelogenin) were detected in dataset 6. For dataset 7, 41 additional alleles at 34 new loci (8 autosomal, 21 X-chromosome, 4 Y-chromosome, and Amelogenin) were found. For more detail on these findings see Warshauer et al (110). As noted previously, allele detection is dependent on a number of factors, including sequence read length and library preparation chemistry. Lack of detection of alleles at loci included in the new locus configuration file was due to these same issues and was not indicative of improper software function. Time required for analysis was directly related to the number of loci that were included in the locus configuration file. Thus, the new larger configuration file did increase STRait Razor v2.0's analysis time. In this study, the time required for analysis of all 86 markers for dual 400MB MiSeq™-generated FASTQ files on a 16-core server was approximately 29 minutes. When only Y-chromosome or X-chromosome STRs were analyzed, the analysis time dropped to approximately 9 minutes for each. The time required for analysis of only autosomal STRs was approximately 11 minutes. Shorter custom configuration files can be used to reduce analysis time, and the time required will vary depending on the specifics of each application and the computing platform utilized.

Genotypes were displayed, along with allele read counts, in a manner that was clear and easy to interpret. Histograms were generated for the alleles, stutter, and noise detected from these datasets that approximated electropherogram displays (Figure 8). Given the similarity between these histograms and traditional electropherograms, this option provided a simple way to visually inspect allele calls at each detected locus. In this manner, reads resulting from stutter or noise could be interpreted visually.

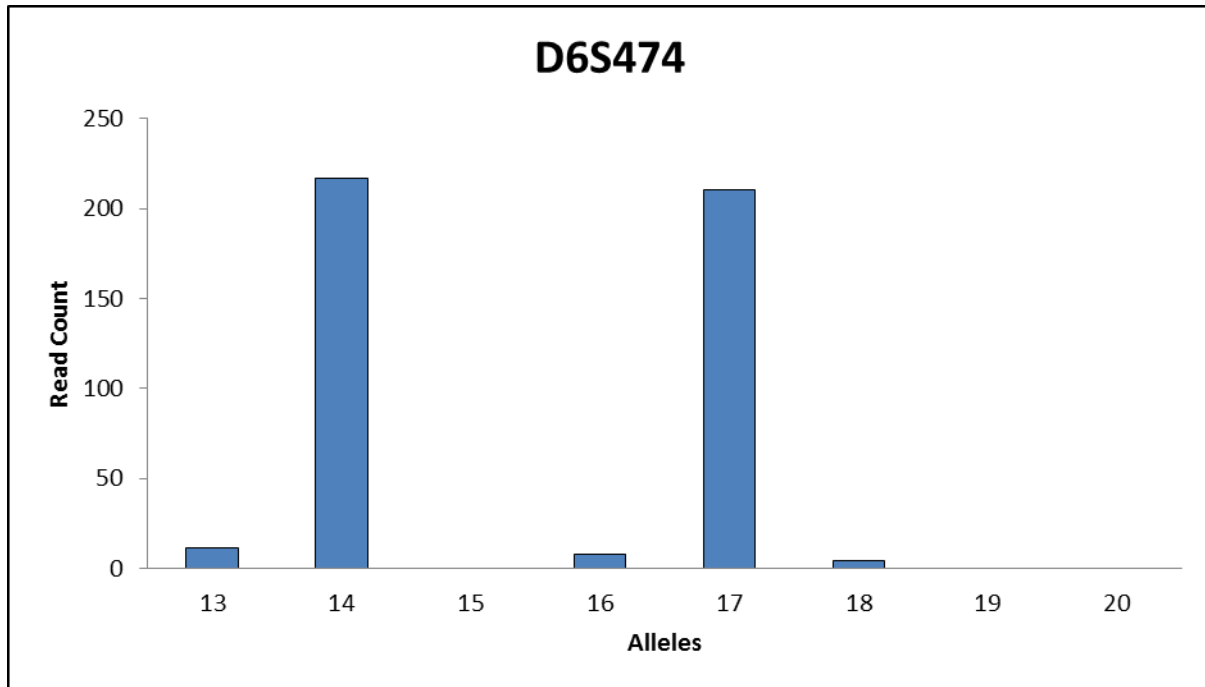


Figure 8. Histogram generated for locus D6S474 in dataset 3 (Read 1). Histogram generated by the supplemental tool included with STRait Razor v2.0 resembles traditional electropherograms. Read counts displayed on the Y-axis are analogous to RFU values for peak heights. Here, the true alleles are “14” and “17.” Stutter peaks “13” and “16” can be seen to the left of the major allele peaks. This profile also shows the plus stutter “18” peak to the right of the major “17” allele peak.

Finally, the sequence data for each detected allele that was output by STRait Razor v2.0 were investigated. The unique sorting process employed by the updated software allowed for quick detection of intra-repeat nucleotide variation. For example, an examination of the sequence data output file for dataset 4 revealed that the homozygote “29” allele at locus D21S11 consisted of the variants “(TCTA)₄(TCTG)₆(TCTA)₃TA(TCTA)₃TCA(TCTA)₂TCCATA (TCTA)₁₁” and “(TCTA)₅(TCTG)₆(TCTA)₃TA(TCTA)₃TCA(TCTA)₂TCCATA(TCTA)₁₀”, at a ratio approaching 1:1. The latter “29” variant is not listed in STRbase. Additional sequence variation can provide increased discriminatory power.

STRait Razor v2.0 retains the reliable and accurate allele-calling capability of the initial release, with the added benefit of a much larger range of detectable loci. The ability to detect autosomal, Y-chromosome, and now X-chromosome STRs augments the usefulness of the software and provides for a wider range of potential applications. The enhanced custom locus configuration file generator and supplemental tools, such as the genotyper and histogram generator, have made STRait Razor v2.0 much more user-friendly and facilitate ease and speed of analysis. Genotypes were determined quickly and could be reviewed by the analyst readily. Aspects such as stutter can be investigated in a manner that resembles that of current electropherogram interpretation. Finally, intra-repeat nucleotide variation was presented in a much easier way to analyze. The

sorting of unique sequences for each allele by the total read count allowed the user to distinguish true allelic variants from those that may be due to simple sequencing errors or background noise.

STRait Razor v2.0 is free to use and available online (http://web.unthsc.edu/info/200210/molecular_and_medical_genetics/887/research_and_development_laboratory/5). Updates and new content will be added to the website as they are developed and tested.

VIII. Full Panel STR Results from STRait Razor v2.0 Analyses

With the enhancements of STRait Razor v2.0, it was possible to analyze the full STR panel. Because of differences in probe design between TruSeq and HaloPlex, start and stop points for sequencing, and read length (as described in the initial STR evaluation above), coverage was not compared among the methods. The criterion for the initial evaluation was the number of loci that provided typeable results. These results would guide the decision for the final panel construct. It should be noted that successfully typed is being used instead of correctly typed as the additional loci typed here were not typed with a previously established method. The number of successfully typed loci (Table 15) varied by method. The shorter reads on the GAIIx instrument yielded the fewest successfully typed loci. The longer reads afforded with the MiSeq provided the highest number of successfully typed loci. As expected, and previously reported (93,94), longer read lengths are necessary to capture longer STR alleles. Therefore, the final panel will be run only on the MiSeq instrument. Although the coverage was higher with the HaloPlex approach (data not shown), the number of typeable loci was slightly higher in 2 out of three samples with a 10X minimum coverage (arbitrarily set) for calling alleles (which was not significant with the TruSeq approach). However, interpretation of allele calls with HaloPlex data is far more complicated because of the start and stop position points where read one and/or read two may not sequence an allele in one or both read directions. In order to exploit fully HaloPlex generated data, software will be needed to facilitate allele calling. Therefore, based on these results the final panel will be based on Nextera Rapid Capture Custom Enrichment Kit and the MiSeq instrument with read lengths of 250 bases (see below).

Table 15. Number of STR loci (n=85 STRs) that provided a result

Sample	Sample Preparation	Instrument	Loci Typeable ^a	Above Threshold ^b
1	HaloPlex	GAIIx	60	58
1	TruSeq	GAIIx	62	58
2	HaloPlex	GAIIx	60	58
1	HaloPlex	MiSeq	74	71
3	TruSeq	MiSeq	82	69
4	TruSeq	MiSeq	82	74
5	TruSeq	MiSeq	82	74

a. Coverage of $\geq 1X$

b. Coverage of $\geq 10X$ (arbitrarily set)

IX. SNP Typing with Large Marker Panels by MPS Results and Discussion

Currently, STRs are the primary genetic markers used for forensic analyses because of their high discrimination power and relatively short amplicon size. However, some evidence samples are highly degraded and may not be characterized well with the current battery of STRs. Although commercial mini-STR typing kits enable generation of amplicons ranging from approximately 70-280 bp, some degraded samples still may produce partial or no STR profiles. In addition, STRs have relatively high mutation rates, which at times limit their use in kinship analyses. In contrast, SNP typing may be applied successfully to degraded samples that are not amenable to STR typing. SNPs amplicons (if PCR enrichment is employed) can be designed to be smaller than 150 bp and, in theory, as short as 50-60 bp in length. In addition, SNPs have orders of magnitude lower mutation rates than that of STRs. Even with these desirable features, there has been resistance in embracing alternate markers beyond STRs because of the large number of STR profiles in CODIS and other national DNA databases. However, MPS enables multiplexing a much larger number of markers and can combine STR and SNP typing into a single analysis. Thus, the legacy data would not be in jeopardy with advancements in technology and both markers types can be accommodated for the overall betterment of providing enhanced forensic analyses toolkits.

There have been a number of reports on SNP typing methods describing high discrimination power. For example, the SNPforID group developed a multiplex assay with 52 autosomal SNPs with a mean match probability of at least 5.0×10^{-19} in nine different populations (85). Pakstis et al (83) reported on a panel of 45 unlinked SNPs providing matching probabilities of less than 1.0×10^{-15} in 44 populations. Considering that the match probability for the 13 CODIS core STR loci is approximately 2.4×10^{-15} in, for example, the US Caucasian population (8), these SNP panels provide discrimination power comparable to that of the STR core loci. Various approaches have been used to analyze SNPs, such as single base extension, chip-based microarrays, allele-specific hybridization assays and mass spectrometry (83, 111-114). Each of these methods has some limitations; the most notable is not being able to type a large battery of SNPs in a single analysis.

MPS provides a platform for more comprehensive coverage of genetic markers. There were 379 SNPs (Table 16).

Table 16. The 379 SNPs in the panel with location and genotypes obtained by MPS.

		1	2	3	4	5	047	412
chr20.60058231	rs1000322	A/A	A/G	A/A	A/G	G/G	A/A	A/G
chr19.16449517	rs1000329	T/T	T/T	T/T	C/T	C/C	T/T	T/T
chr4.41554364	rs10007810	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr20.23530035	rs1003204	C/T	T/T	C/T	T/T	C/T	T/T	T/T
chr21.36446597	rs1003473	G/C	G/C	G/G	C/C	G/G	G/G	G/C
chr17.41691526	rs1004357	A/A	A/A	A/A	A/A	A/A	A/A	G/A
chr20.39487110	rs1005533	G/A	G/A	G/G	A/A	G/A	A/A	A/A
chr6.168321659	rs1008457	T/T	C/T	C/T	C/C	C/C	C/T	T/T
chr9.17602496	rs1008730	G/G	G/G	A/G	G/G	G/G	G/G	A/A

chr8.28411072	rs10092491	C/C	T/C	C/C	T/C	T/T	T/C	T/T
chr8.4190793	rs10108270	C/A	C/A	C/C	C/A	C/C	C/C	C/C
chr14.36170607	rs10141763	T/T	T/T	T/T	T/T	T/T	T/T	T/T
chr9.1823774	rs1015250	G/C	G/G	G/C	G/G	G/C	G/C	G/C
chr18.24363110	rs1017415	A/A	A/G	G/G	A/G	A/A	A/G	G/G
chr7.13894276	rs1019029	A/A	A/G	A/A	A/G	A/A	G/G	A/G
chr2.60012802	rs1019264	G/G	G/A	G/A	G/G	G/A	G/A	G/A
chr2.33186261	rs1020636	A/G	A/G	A/G	A/G	A/A	A/A	A/A
chr11.103842542	rs1021290	T/T	C/T	T/T	C/T	C/C	C/C	C/T
chr7.139447377	rs10236187	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr18.75432386	rs1024116	T/T	C/T	T/T	C/T	C/C	T/T	T/T
chr5.164680288	rs1024997	C/T	T/T	T/T	C/T	C/T	C/C	C/T
chr17.46510697	rs1027895	A/A	A/A	A/G	A/G	A/A	A/A	A/A
chr6.65003904	rs1028484	T/T	T/T	T/T	C/T	C/T	T/T	C/C
chr22.48362290	rs1028528	A/G	A/A	A/G	A/A	A/A	A/G	A/A
chr6.1135939	rs1029047	A/A	T/T	T/T	T/T	T/T	T/A	T/T
chr20.4447483	rs1031825	A/C	C/C	C/C	C/C	C/C	A/A	C/C
chr6.4747159	rs1040045	G/A	G/A	G/A	G/A	A/A	A/A	G/A
chr1.168159890	rs1040404	G/G	G/G	G/G	G/A	G/G	G/G	G/G
chr11.115207176	rs10488710	G/G	C/G	C/G	G/G	C/C	G/G	C/G
chr1.238439308	rs10495407	G/G	A/A	G/G	G/A	G/A	G/G	G/A
chr2.145769943	rs10496971	T/T	T/T	T/T	T/T	T/T	T/T	T/T
chr11.5099393	rs10500617	T/A	T/A	T/T	T/A	T/T	T/A	T/A
chr3.2208832	rs10510228	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr9.28628500	rs10511828	T/T	T/T	T/T	T/T	T/T	T/T	T/T
chr17.69512099	rs10512572	G/G	G/G	G/G	G/G	A/A	G/G	G/A
chr9.120130206	rs10513300	T/T	T/T	T/T	T/T	T/C	T/T	T/T
chr13.100038233	rs1058083	A/G	A/G	A/G	A/G	A/G	G/G	G/G
chr7.15518610	rs1072292	G/G	A/G	A/G	A/G	A/G	G/G	G/G
chr6.88366602	rs1075665	A/G	A/A	A/A	A/G	A/G	A/A	G/G
chr12.130761696	rs10773760	A/A	G/G	A/G	G/G	A/A	A/G	A/A
chr9.137417308	rs10776839	G/T	G/T	G/T	T/T	G/T	G/T	T/T
chr2.154933789	rs1079861	A/G	A/G	A/G	A/G	A/A	A/G	A/G
chr11.7850316	rs10839880	C/C	T/T	T/T	T/T	T/T	C/T	T/T
chr12.29369871	rs10843344	C/C	C/T	C/T	C/T	C/C	C/T	C/C
chr7.83533047	rs10954737	T/T	T/T	T/T	T/C	T/T	T/T	T/T
chr18.34124950	rs1105459	A/A	G/G	A/A	A/G	A/G	A/G	A/G
chr13.100380429	rs1105576	T/T	T/C	T/T	T/C	T/T	T/T	T/C
chr20.42703547	rs1108943	C/C	C/C	C/C	C/C	T/C	T/C	T/C
chr2.10085722	rs1109037	G/A	G/A	G/G	G/A	G/A	G/A	A/A
chr11.66898492	rs11227699	G/G	G/A	G/G	G/G	G/G	G/G	G/G
chr17.62987151	rs11652805	T/T	C/T	T/T	T/T	C/T	T/T	T/T

chr1.55663372	rs12130799	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/G
chr7.103243492	rs123714	T/C	T/C	C/C	C/C	C/C	C/C	C/C	C/C
chr15.36220035	rs12439433	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr20.16241416	rs12480506	A/G	A/A	A/G	A/G	A/A	A/G	A/A	A/A
chr8.86424616	rs12544346	A/A	G/G	A/A	G/G	A/A	G/A	A/A	A/A
chr3.120522716	rs12629908	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr5.79085726	rs12657828	A/A	A/G	A/G	A/G	A/G	G/G	A/A	A/A
chr15.28365618	rs12913832	A/G	A/G	A/A	G/G	A/G	G/G	A/G	A/G
chr1.233448413	rs1294331	C/T	C/C	C/T	C/T	C/C	C/C	C/C	C/C
chr22.18076546	rs1296819	C/C	C/C	C/A	C/A	C/A	C/A	C/C	C/C
chr2.182413259	rs12997453	A/A	A/G	A/G	A/G	A/A	G/G	A/G	A/G
chr4.76425896	rs13134862	G/A	G/A	G/G	G/G	G/A	G/A	G/A	G/A
chr5.136633338	rs13182883	G/A	G/G	G/G	G/A	G/G	A/A	G/G	G/G
chr20.38849642	rs1321333	A/G	A/G	A/A	A/G	A/G	G/G	A/G	A/G
chr6.12059954	rs13218440	G/A	G/G	G/G	G/A	G/G	G/A	G/G	G/G
chr1.42360270	rs1325502	G/A	G/G	G/G	G/A	G/A	G/G	G/G	G/G
chr9.93436252	rs1331494	C/C	C/C	G/C	C/C	G/C	G/C	C/C	C/C
chr13.20901724	rs1335873	A/A	T/T	A/A	T/A	T/A	A/A	T/A	T/A
chr6.94537255	rs1336071	T/T	C/C	T/C	T/C	C/C	T/C	T/T	T/T
chr2.79864923	rs13400937	T/T	T/G	T/G	T/G	T/G	T/G	T/T	T/T
chr3.190806108	rs1355366	T/T	T/C	T/T	T/C	T/C	T/C	T/T	T/T
chr3.961782	rs1357617	T/T	T/T	A/T	A/A	T/T	T/T	A/T	A/T
chr6.123894978	rs1358856	A/A	A/A	C/C	C/A	A/A	A/A	A/A	A/A
chr9.128968063	rs1360288	C/C	C/C	T/T	T/T	C/C	C/C	C/C	C/C
chr4.73245191	rs1369093	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
chr16.80106361	rs1382387	C/A	C/A	A/A	C/A	A/A	C/A	A/A	A/A
chr22.43579708	rs138952	C/C	C/C	C/C	C/C	C/C	T/C	T/C	T/C
chr1.186149032	rs1407434	G/G	G/G	G/A	G/G	G/A	G/A	G/G	G/G
chr9.12672320	rs1408801	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr10.97172595	rs1410059	C/C	T/T	C/C	T/C	T/T	T/C	T/C	T/C
chr1.242806797	rs1413212	C/C	C/C	C/C	T/C	C/C	C/C	C/C	C/C
chr15.48426484	rs1426654	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr14.25850832	rs1454361	A/A	A/A	T/T	T/T	T/T	T/A	T/A	T/A
chr9.126881448	rs1463729	C/T	T/T	C/T	C/C	C/T	T/T	C/T	C/T
chr17.43984399	rs1467966	C/C	C/C	C/C		C/C	C/C	C/C	C/C
chr8.28941305	rs1471939	C/T	T/T	C/T	C/T	T/T	T/T	C/T	C/T
chr6.120560694	rs1478829	T/A	A/A	T/T	T/A	T/A	T/A	T/T	T/T
chr3.43484669	rs1482650	A/A	A/A	A/A	A/A	A/A	A/T	A/T	A/T
chr1.4367323	rs1490413	G/A	G/G	G/G	G/A	A/A	G/A	G/G	G/G
chr18.1127986	rs1493232	A/A	A/A	C/C	A/A	A/A	C/A	C/A	C/A
chr3.168645035	rs1498444	T/T	G/G	T/T	G/G	T/G	T/G	T/G	T/G
chr11.5709028	rs1498553	C/C	C/C	T/T	C/T	C/C	C/T	C/T	C/T

chr5.165739982	rs1500127	C/C	C/C	C/C	C/C	C/C	C/C	C/T
chr5.169436953	rs1501643	T/A	A/A	T/A	A/A	A/A	A/A	A/A
chr12.118889488	rs1503767	T/T	T/T	T/G	T/T	T/T	T/T	T/G
chr12.17407792	rs1513056	G/A	G/G	G/G	G/G	G/G	G/G	G/G
chr3.188574996	rs1513181	G/G	G/G	G/A	G/G	G/G	G/G	G/A
chr20.51296162	rs1523537	T/C	T/C	T/C	T/C	C/C	T/T	T/T
chr15.55210705	rs1528460	C/T	C/T	T/T	C/T	C/T	C/C	C/T
chr8.91823568	rs1542931	C/G	C/G	C/G	C/G	C/C	C/G	G/G
chr4.157489906	rs1554472	A/A	A/A	G/A	G/G	G/A	G/G	G/A
chr2.201021954	rs1569175	C/C	C/C	C/C	C/C	C/C	C/C	C/C
chr1.36768200	rs1573020	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr5.52811560	rs1593055	T/T	T/T	T/A	T/A	T/T	T/A	A/A
chr5.17374898	rs159606	A/G	A/G	G/G	G/G	A/A	G/G	A/G
chr5.33951693	rs16891982	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr18.55225777	rs1736442	C/C	C/C	T/C	T/C	T/C	T/T	C/C
chr11.61672235	rs174473	T/T	T/T	T/T		T/T	T/T	T/T
chr14.20818131	rs1760921	T/T	T/T	T/T	T/C	T/T	T/T	T/T
chr10.119362760	rs181619	G/G	T/T	G/G	G/G	G/G	G/G	G/T
chr15.39313402	rs1821380	C/G	C/G	C/C	C/C	C/C	C/G	C/G
chr2.136616754	rs182549	C/T	C/C	T/T	C/T	T/T	C/T	T/T
chr11.15838137	rs1837606	T/T	T/T	T/T	T/C	T/C	T/T	T/T
chr21.18565025	rs18579	G/A	G/A	G/A	G/G	G/G	A/A	G/A
chr7.24516433	rs1858958	G/G	G/G	G/G	G/G	G/C	G/G	G/G
chr6.168665760	rs1871428	G/A	G/G	G/G	G/A	G/A	G/A	A/A
chr3.113804979	rs1872575	G/A	G/A	G/A	A/A	G/A	G/A	G/A
chr17.1401613	rs1879488	C/C	C/C	C/C	C/C	A/A	C/C	C/C
chr13.22374700	rs1886510	G/G	G/A	A/A	A/A	G/A	G/A	G/A
chr6.90518278	rs192655	A/A	A/A	G/A	A/A	A/A	G/A	A/A
chr14.58238687	rs1950993	G/G	G/T	G/T	G/G	T/T	G/T	G/G
chr10.34755348	rs1978806	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr4.190318080	rs1979255	C/G	C/C	G/G	C/G	C/G	C/C	G/G
chr6.14012999	rs1997680	T/T	A/T	A/T	A/T	T/T	T/T	T/T
chr8.140241181	rs2001907	C/C	C/C	C/C	C/C	C/C	C/C	C/C
chr14.99375321	rs200354	G/T	G/G	G/G	G/G	G/G	G/G	G/T
chr17.39150443	rs2010209	G/G	G/A	G/G	G/A	A/A	G/A	G/A
chr1.166899807	rs2013526	T/T	C/C	T/C	C/C	T/C	T/C	T/T
chr2.53037869	rs2015632	C/T	T/T	T/T	T/T	T/T	T/T	T/T
chr15.24571796	rs2016276	T/C	T/T	T/T	T/C	T/C	T/T	T/C
chr13.70300514	rs2018205	C/T	C/C	C/C	C/C	T/T	C/T	C/T
chr4.159181963	rs2026721	C/C	C/C	C/C	C/C	C/C	C/C	C/C
chr3.179964727	rs2030763	G/A	G/G	G/G	G/A	G/G	G/A	G/G
chr17.53788280	rs2033111	A/A	A/G	A/G	A/G	A/A	A/A	A/A

chr22.47836412	rs2040411	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr4.10969059	rs2046361	T/A	T/A	T/T	T/A	T/A	T/T	T/A
chr8.139399116	rs2056277	C/C	C/T	C/C	C/C	C/T	C/C	C/C
chr1.204790977	rs2065160	A/A	A/A	G/G	A/A	A/A	A/G	A/A
chr13.34864240	rs2065982	T/T	T/T	T/T	T/T	T/T	T/C	T/T
chr12.109277720	rs2070586	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr22.23802171	rs2073383	T/C	C/C	C/C	C/C	T/T	T/T	T/C
chr9.135933122	rs2073821	C/C	C/C	C/C	C/C	C/C	C/C	C/C
chr11.134667546	rs2076848	A/A	A/A	A/T	A/T	A/T	A/T	T/T
chr12.888320	rs2107612	A/A	A/A	A/A	A/A	G/A	G/A	A/A
chr12.106328254	rs2111980	C/C	T/C	T/C	T/C	T/C	C/C	C/C
chr17.73782191	rs2125345	T/T	T/T	T/C	T/T	C/C	T/C	T/C
chr12.47676950	rs214678	T/T	T/T	T/T	T/C	T/T	T/C	T/T
chr6.152697706	rs214955	C/T	C/C	C/T	C/T	C/C	C/C	T/T
chr21.43606997	rs221956	C/C	C/C	T/C	T/C	T/C	C/C	T/C
chr12.6909442	rs2255301	C/C	T/C	T/T	T/C	C/C	C/C	T/C
chr12.6945914	rs2269355	C/C	C/C	C/G	C/G	C/G	C/C	G/G
chr16.19272908	rs2269793	T/T	G/G	T/T	T/T	T/T	T/T	T/T
chr9.14747133	rs2270529	T/T	T/C	T/T	T/T	C/C	T/C	T/T
chr6.148761456	rs2272998	G/C	G/G	G/G	C/C	G/C	G/G	G/C
chr17.80526139	rs2291395	A/A	A/G	G/G	G/G	A/A	A/G	G/G
chr19.42410331	rs2303798	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr17.75551667	rs2304925	A/A	A/A	A/C	A/A	A/C	A/A	A/A
chr9.93641199	rs2306040	T/T	T/T	T/C	T/T	T/C	T/T	T/T
chr3.79427470	rs2311046	T/T	A/A	A/T	A/T	A/T	T/T	A/T
chr7.42380071	rs2330442	A/G	A/A	A/A	A/A	A/G	A/A	A/G
chr16.5868700	rs2342747	A/G	A/A	A/G	G/G	G/G	A/G	G/G
chr14.52607967	rs2357442	A/C	C/C	A/A	A/A	A/A	A/A	A/A
chr5.8293937	rs2388618	T/T	T/A	T/T	T/T	T/A	T/T	T/T
chr21.17710424	rs239031	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr6.51611470	rs2397060	T/T	T/T	T/C	T/T	T/T	T/T	T/T
chr3.110301126	rs2399332	G/G	G/G	T/G	G/G	T/G	T/G	T/T
chr12.11701488	rs2416791	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr6.127463376	rs2503107	C/A	C/C	C/A	C/A	A/A	C/C	C/A
chr6.12535111	rs2504853	C/C	T/C	C/C	T/T	T/C	T/C	T/T
chr5.174778678	rs251934	A/A	A/G	A/A	A/A	G/G	A/G	A/G
chr19.55614923	rs2532060	T/C	T/T	T/C	T/T	T/C	T/C	T/C
chr20.23017082	rs2567608	T/T	T/T	T/T	T/C	T/T	T/T	T/C
chr21.25672460	rs2572307	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr2.109579738	rs260690	A/A	C/A	A/A	A/A	A/A	A/A	A/A
chr2.179606538	rs2627037	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr4.179399523	rs2702414	G/A	G/G	G/A	G/G	G/G	G/A	G/G

chr4.46329655	rs279844	A/T	A/A	A/T	A/T	A/T	A/T	A/T
chr6.55155704	rs2811231	A/A	C/A	C/A	C/C	A/A	C/C	A/A
chr1.159174683	rs2814778	T/T	T/T	T/T	T/T	T/T	T/C	T/T
chr21.28608163	rs2830795	A/A	A/G	A/G	A/A	A/A	A/A	A/A
chr21.29679687	rs2831700	G/G	A/A	A/A	G/G	G/G	A/G	A/G
chr21.33582722	rs2833736	G/G	G/A	G/A	G/G	A/A	A/A	G/A
chr21.37885625	rs2835370	T/T	T/T	T/T	C/C	T/T	T/T	T/T
chr15.74734500	rs2899826	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr12.40863052	rs2920816	A/G	G/G	A/G	A/G	A/G	A/A	A/A
chr11.24010530	rs2946788	G/G	G/T	G/G	G/T	G/T	T/T	G/T
chr16.85183682	rs2966849	G/G	A/A	G/G	G/G	G/G	G/G	G/G
chr1.6550376	rs2986742	T/T	T/T	T/T	T/T	T/T	T/C	T/T
chr2.204838091	rs3096741	A/G	A/G	A/G	A/G	A/A	A/G	G/G
chr1.68849687	rs3118378	A/G	A/A	A/A	A/A	G/G	A/G	A/G
chr5.169735920	rs315791	A/C	C/C	A/A	A/A	A/A	A/A	C/C
chr5.2364626	rs316598	T/T	T/T	T/T	T/T	T/T	T/C	T/C
chr1.242342504	rs316873	C/C	C/C	C/C	C/T	C/C	C/C	C/C
chr7.137029838	rs321198	T/T	T/C	C/C	C/C	T/C	C/C	T/T
chr7.32179124	rs32314	T/T	T/T	T/T	T/T	C/T	T/T	C/T
chr5.178690725	rs338882	G/A	G/A	G/A	G/A	G/G	G/A	A/A
chr13.106938411	rs354439	A/T	A/T	A/T	T/T	A/T	A/T	A/A
chr5.35037115	rs37369	C/C	C/C	C/C	C/C	C/C	C/C	C/C
chr1.101709563	rs3737576	T/T	T/T	T/T	T/T	T/C	T/T	T/T
chr17.80739859	rs3744163	G/G	G/C	G/C	G/G	G/C	G/G	G/G
chr19.52901905	rs3745099	A/A	G/A	A/A	A/A	A/A	A/A	G/A
chr10.17193346	rs3780962	G/G	G/G	A/G	A/G	A/G	A/G	A/G
chr14.105679055	rs3784230	A/G	A/A	A/G	A/G	A/A	A/G	A/G
chr16.90105333	rs3785181	C/C	C/C	C/C	C/C	C/T	C/C	C/C
chr9.71659280	rs3793451	C/C	C/C	C/C	C/C	C/C	C/C	T/T
chr10.50841704	rs3793791	T/T	T/T	T/T	T/T	T/T	T/T	T/T
chr4.85309078	rs385194	A/G	G/G	G/G	G/G	A/A	A/G	A/A
chr20.54000914	rs3907047	T/T	T/T	T/T	T/T	T/C	T/T	T/T
chr8.13359500	rs3943253	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr3.9152374	rs420426	C/C	C/C	C/T	C/T	T/T	C/T	T/T
chr8.136839229	rs4288409	C/C	C/C	A/C	C/C	C/C	A/C	A/C
chr16.78017051	rs430046	C/T	T/T	C/C	T/T	C/T	C/T	C/T
chr3.32417644	rs4364205	T/G	G/G	T/G	T/G	T/G	T/G	T/T
chr20.15124933	rs445251	C/C	G/G	G/C	C/C	G/G	G/G	G/C
chr6.163221792	rs4458655	T/C	T/T	T/T	T/C	T/T	T/T	T/T
chr6.145055331	rs4463276	G/G	G/G	G/A	G/A	G/A	A/A	G/G
chr14.104769149	rs4530059	G/A	G/A	G/A	G/A	A/A	G/A	G/A
chr4.38803255	rs4540055	A/T	A/A	A/A	A/T	A/T	A/A	A/A

chr8.144656754	rs4606077	C/C	T/C	C/C	C/C	C/C	T/C	C/C
chr21.28023370	rs464663	C/C	C/C	T/C	C/C	T/C	C/C	C/C
chr2.29538411	rs4666200	A/A	A/A	G/A	A/A	A/A	G/A	G/A
chr2.37941396	rs4670767	G/G	G/T	G/G	G/G	G/G	G/G	G/G
chr7.73454199	rs4717865	G/G	G/G	G/A	G/G	G/G	G/G	G/G
chr10.75300994	rs4746136	G/A	G/G	G/G	G/G	G/G	G/G	G/G
chr16.10975311	rs4781011	G/G	G/G	G/G	T/G	T/G	T/G	G/G
chr17.6811529	rs4796362	G/A	G/A	G/G	G/G	G/A	G/G	G/A
chr18.9420504	rs4798812	G/A	G/G	G/G	G/A	G/A	G/A	G/A
chr18.19651982	rs4800105	C/C	C/C	C/C	C/C	C/C	C/T	T/T
chr22.32366359	rs4821004	C/T	C/T	T/T	T/T	T/T	T/T	C/C
chr1.105717631	rs4847034	A/G	A/G	G/G	A/G	G/G	G/G	G/G
chr10.134650103	rs4880436	C/C	C/C	T/T	C/C	C/C	C/C	C/T
chr18.67867663	rs4891825	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr1.27931698	rs4908343	A/A	G/A	A/A	A/A	G/G	A/A	A/A
chr10.115316812	rs4918842	T/T	T/T	T/T	T/T	T/T	T/C	T/C
chr1.212786883	rs4951629	T/T	T/C	T/C	T/T	T/T	T/T	T/T
chr3.30415612	rs4955316	T/G	T/T	T/G	T/T	T/T	T/T	T/T
chr16.740466	rs4984913	A/G	G/G	A/G	A/G	A/A	A/G	A/G
chr1.38182164	rs502776	A/A	T/A	T/T	T/A	T/A	T/A	T/T
chr11.32424389	rs5030240	A,G	C/C	C/C	C/G	C/C	A,G	C/C
chr18.47371014	rs521861	G/G	G/G	C/C	C/G	C/C	C/G	C/G
chr1.160786670	rs560681	A/G	G/G	A/A	A/G	A/A	A/A	A/G
chr22.19920646	rs5746846	C/C	C/G	C/C	C/C	C/G	C/G	G/G
chr19.39559807	rs576261	C/C	C/C	A/C	A/C	A/C	A/A	C/C
chr22.48207872	rs5768007	C/C	C/T	C/C	C/T	C/C	C/T	C/T
chr10.113627886	rs585070	T/C	T/C	T/C	T/C	C/C	C/C	T/C
chr11.122195989	rs590162	T/T	C/T	C/C	C/T	T/T	C/T	C/T
chr18.4237534	rs595601	T/T	T/T	A/T	A/A	A/A	A/T	A/T
chr22.26350103	rs5997008	C/C	C/C	C/C	C/C	C/C	C/C	C/A
chr20.10195433	rs6104567	T/T	T/G	T/T	T/G	T/T	T/T	T/T
chr11.35541878	rs627119	A/G	A/G	A/A	A/G	A/A	A/A	A/A
chr5.177863083	rs6422347	T/T	T/T	T/T	T/C	T/T	T/T	T/T
chr3.193207380	rs6444724	T/T	T/C	T/C	T/C	T/C	T/T	T/T
chr5.43711378	rs6451722	G/G	G/A	G/G	G/G	G/G	G/A	G/G
chr7.151873853	rs6464211	C/T	C/C	C/C	C/T	C/T	C/T	C/T
chr8.117122598	rs6469629	G/G	A/A	A/G	A/A	A/G	G/G	A/A
chr1.18170886	rs647325	A/A	A/A	G/G	A/A	A/A	A/A	A/G
chr1.12608178	rs6541030	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr3.79399575	rs6548616	T/C	T/C	T/C	C/C	T/T	T/T	T/T
chr5.155471714	rs6556352	C/T	C/C	C/T	C/C	T/T	C/C	C/T
chr11.105912984	rs6591147	T/C	T/C	C/C	C/C	C/C	T/C	T/C

chr4.169663615	rs6811238	G/G	G/G	T/G	G/G	T/G	G/G	G/G
chr7.4310365	rs6955448	C/T	C/C	C/C	C/C	C/C	C/T	C/T
chr1.54195018	rs702490	A/G	A/G	A/G	A/G	A/A	A/G	G/G
chr9.27985938	rs7041158	C/C	C/C	T/T	C/C	C/T	C/C	C/T
chr7.97695363	rs705308	C/C	A/A	A/A	C/C	C/A	C/A	A/A
chr4.182192291	rs716360	A/A	G/A	G/A	A/A	A/A	G/A	A/A
chr6.39882750	rs716856	A/G	G/G	G/G	A/G	A/G	A/G	A/G
chr5.2879395	rs717302	G/G	G/G	A/A	G/A	G/A	A/A	G/A
chr15.54523909	rs719211	A/G	G/G	A/A	G/G	A/A	A/G	A/G
chr19.28463337	rs719366	G/A	G/A	A/A	A/A	G/A	G/G	G/G
chr16.7520254	rs7205345	C/C	C/G	C/G	G/G	C/C	C/G	C/C
chr21.16685598	rs722098	A/G	A/A	A/A	A/A	A/G	A/G	A/G
chr14.53216723	rs722290	G/G	C/C	C/C	G/C	G/G	G/G	G/G
chr18.22739001	rs7229946	G/G	G/G	G/A	G/A	G/G	A/A	G/G
chr18.49781544	rs7238445	G/A	G/G	G/G	G/G	G/G	G/A	G/G
chr17.31918109	rs727206	G/A	A/A	G/A	G/A	G/A	G/G	G/A
chr6.165045334	rs727811	T/T	G/G	T/T	G/G	G/T	T/T	G/T
chr16.5606197	rs729172	G/T	G/G	G/T	G/T	G/T	G/T	G/G
chr13.109415188	rs729549	C/T	T/T	T/T	C/T	C/T	T/T	C/C
chr11.19977718	rs729999	G/G	G/G	G/G	G/G	G/G	A/G	G/G
chr12.102149981	rs730013	A/G	A/G	G/G	A/G	A/G	G/G	G/G
chr13.40101740	rs730249	T/T	C/C	C/T	C/T	C/T	T/T	C/T
chr6.3350185	rs730488	T/T	T/T	T/T	T/T	T/T	T/C	T/T
chr14.101142890	rs730570	A/A	A/A	A/A	G/A	A/A	G/A	A/A
chr5.118058631	rs730907	A/A	A/G	A/G	A/G	G/G	A/G	A/G
chr7.12669251	rs731257	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr1.34155501	rs732889	G/G	G/A	G/G	G/A	A/A	A/A	G/A
chr5.132655625	rs733023	G/G	A/G	G/G	G/G	A/G	G/G	G/G
chr22.27816784	rs733164	G/G	G/G	G/A	G/G	G/G	G/A	G/A
chr2.53828410	rs734295	T/C	C/C	T/T	T/C	T/C	T/C	C/C
chr14.84668023	rs734656	A/A	A/A	A/A	G/A	G/A	G/A	G/A
chr1.14996654	rs734664	A/A	A/G	A/A	A/G	A/A	A/G	A/G
chr8.6388247	rs734701	G/A	G/A	A/A	G/A	A/A	G/A	A/A
chr3.147750355	rs734873	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr10.3374178	rs735155	C/C	C/C	T/T	C/C	C/T	C/T	T/T
chr22.35948435	rs736210	C/C	C/T	T/T	C/C	C/T	C/C	C/T
chr8.287398	rs737168	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr7.155990813	rs737681	C/C	T/C	C/C	T/T	T/C	C/C	T/T
chr22.37119800	rs738518	T/T	C/T	T/T	C/T	T/T	T/T	T/T
chr22.43172267	rs738532	C/T	C/T	C/T	C/C	T/T	C/T	C/C
chr10.118506899	rs740598	A/A	A/A	G/A	A/A	G/A	A/A	A/A
chr17.5706623	rs740910	G/G	A/A	G/G	A/G	A/A	A/G	A/A

chr2.14756349	rs7421394	A/G	A/A	A/G	A/G	G/G	A/G	A/A
chr20.25053105	rs743018	G/G	G/G	G/A	G/G	G/G	G/G	G/A
chr14.68053124	rs749270	T/T	T/A	T/T	T/T	T/T	T/T	T/T
chr1.14155402	rs7520386	A/A	G/A	G/A	A/A	G/A	G/A	G/G
chr1.151122489	rs7554936	C/T	C/T	C/C	T/T	T/T	C/T	C/C
chr8.1375610	rs763869	A/A	G/A	A/A	G/G	A/A	G/A	A/A
chr6.124142944	rs765533	A/A	A/G	A/A	G/G	A/A	A/A	A/G
chr4.105375423	rs7657799	T/T	T/T	T/T	T/T	T/T	T/T	T/T
chr5.159487953	rs7704770	G/A	G/A	G/A	G/G	A/A	G/A	G/G
chr12.56163734	rs772262	G/G	G/G	G/A	G/G	G/G	G/A	G/G
chr2.7833821	rs772436	C/C	C/T	T/T	C/T	T/T	C/T	T/T
chr12.56603834	rs773658	C/C	C/C	C/C	C/C	C/C	C/C	C/C
chr6.21911616	rs7745461	G/G	G/G	A/G	G/G	A/G	A/A	G/G
chr7.130742066	rs7803075	G/G	A/G	A/A	A/G	A/A	A/G	G/G
chr8.122908503	rs7844723	C/T	C/T	T/T	C/T	C/T	T/T	C/T
chr10.17064992	rs7897550	G/G	G/A	G/G	G/A	G/G	G/G	G/A
chr2.7968275	rs798443	G/A	A/A	A/A	A/A	G/A	A/A	G/A
chr13.34847737	rs7997709	T/T	T/T	T/T	T/T	T/T	C/T	T/T
chr14.67886781	rs8021730	G/G	G/G	G/G	G/G	G/G	G/G	G/G
chr15.92105708	rs8035124	A/A	A/A	A/A	A/C	A/C	A/C	A/C
chr15.53616909	rs8037429	C/T	C/T	T/T	T/T	C/T	C/T	T/T
chr17.41341984	rs8070085	A/G	A/G	A/G	A/A	A/A	A/A	A/A
chr17.80461935	rs8078417	C/T	C/T	C/T	C/C	C/C	C/C	C/C
chr19.33652247	rs8113143	C/A	C/C	C/C	C/C	C/A	C/C	C/C
chr16.65406708	rs818386	C/C	T/C	C/C	C/C	T/C	C/C	T/C
chr10.2406631	rs826472	C/C	C/C	T/C	T/C	T/C	T/C	T/C
chr17.78877735	rs868432	G/G	G/A	G/G	G/A	G/G	G/G	G/A
chr5.6845035	rs870347	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr14.98845531	rs873196	T/T	T/T	C/C	C/C	C/T	T/T	C/T
chr19.1175396	rs873289	A/A	A/A	A/G	A/G	G/G	G/G	G/G
chr18.75056284	rs874299	T/T	C/C	T/C	C/C	T/T	T/C	T/C
chr2.114974	rs876724	C/C	C/C	C/T	C/T	C/C	T/T	C/T
chr15.61076591	rs877228	A/G	A/A	A/G	G/G	A/A	G/G	A/G
chr5.153861047	rs880083	C/T	C/T	C/T	C/T	C/T	C/C	C/T
chr18.59333108	rs881728	C/C	C/C	C/C	C/C	C/C	C/C	C/C
chr16.31079371	rs881929	G/T	G/G	G/T	G/T	G/T	G/G	G/G
chr19.30585036	rs887754	C/C	C/C	C/C	C/T	C/T	C/C	C/C
chr1.239881926	rs891700	A/G	A/G	A/G	A/G	A/G	A/G	A/G
chr8.103550014	rs892503	C/T	C/T	T/T	C/T	C/T	C/T	T/T
chr2.121350385	rs896499	C/C	T/C	T/T	T/C	C/C	T/C	T/C
chr2.7149155	rs896788	C/C	C/C	C/C	C/T	C/C	C/C	C/T
chr11.11096221	rs901398	T/T	C/T	T/T	C/T	C/C	C/T	T/T

chr2.239563579	rs907100	G/G	G/C	G/C	G/C	G/G	C/C	C/C
chr14.55125716	rs911621	T/C	T/C	T/C	C/C	T/T	T/C	T/C
chr21.42415929	rs914165	G/G	G/G	G/G	G/G	G/A	G/A	G/G
chr6.167030062	rs916388	G/A	G/G	G/G	G/G	G/A	G/A	G/G
chr7.4457003	rs917118	C/T	T/T	C/C	C/T	C/T	C/T	C/C
chr17.55150205	rs917927	G/A	G/G	G/A	G/A	A/A	G/G	A/A
chr8.57562039	rs919023	T/C	T/T	T/T	T/T	T/C	T/C	T/T
chr11.132091073	rs921269	C/C	C/T	C/C	T/T	C/C	C/T	C/T
chr10.82771574	rs922992	A/G	A/A	A/A	A/G	A/A	A/G	A/G
chr1.110680114	rs924181	G/G	G/A	G/A	G/A	G/A	G/A	G/A
chr6.119798030	rs924397	C/T	C/C	C/T	C/C	C/T	C/C	C/C
chr6.15010230	rs927628	C/T	C/T	C/T	C/T	C/C	C/T	C/T
chr4.5390637	rs9291090	A/A	A/A	A/A	A/A	A/A	A/A	A/A
chr13.27624356	rs9319336	T/T	T/T	T/T	T/T	T/T	T/T	T/T
chr17.77468498	rs938283							
chr14.83472868	rs946918	G/G	G/T	G/G	T/T	G/T	G/T	G/T
chr11.120644447	rs948028	A/A	A/A	A/A	A/A	A/C	A/A	A/A
chr13.111827167	rs9522149	T/C	C/C	T/C	C/C	C/C	C/C	T/C
chr13.75993887	rs9530435	T/C	C/C	C/C	T/C	C/C	T/C	C/C
chr13.84456735	rs9546538	T/C	T/C	T/T	T/T	T/T	T/C	T/T
chr12.30268737	rs959566	T/T	T/C	T/C	T/C	C/C	T/C	T/C
chr10.132698419	rs964681	T/C	C/C	T/C	C/C	T/C	T/C	T/C
chr3.39146429	rs9809104	T/T	T/T	T/T	T/T	T/C	T/T	T/C
chr3.135914476	rs9845457		A/A	A/A			G/G	A/A
chr18.29311034	rs985492	G/A	G/A	G/G	G/G	G/A	A/A	G/A
chr3.59488340	rs9866013	T/C	T/C	T/C	T/T	T/C	T/C	T/T
chr22.33559508	rs987640	T/A	T/T	T/A	T/A	T/A	T/A	T/A
chr17.2919393	rs9905977	A/G	G/G	G/G	G/G	A/G	A/G	G/G
chr2.124109213	rs993934	A/A	A/A	A/A	A/A	A/G	A/G	A/A
chr18.9749879	rs9951171	G/A	A/A	G/A	G/A	G/G	G/G	A/A
chr7.51964745	rs997556	C/C	T/C	T/T	T/T	T/C	T/T	T/T
chr1.184182392	rs997568	A/G	A/A	A/G	G/G	G/G	G/G	A/A
chr10.27919931	rs997750	G/G	A/A	A/A	A/A	A/G	G/G	G/G
chr15.23000272	rs999842	G/G	A/A	A/A	A/A	G/G	A/G	A/G
Total SNPs Called		377	378	378	375	377	378	378

Seven samples, run on the GAIIX chemistry, were analyzed for these SNPs (the shorter read length should not compromise typing success). SNP typing was highly successful. Only 1-4 SNPs per sample failed to yield a result. The SNP rs938283 did not yield a result in any sample and SNP rs9845457 yielded a result in only about half of the samples.

Out of all the SNPs, 328 were heterozygous in one or more of the samples. The ACR for heterozygous types ranged from 0.460 to 1.00 (Figure 9), of which only 8 SNPs displayed an average ACR <0.60. The average depth of coverage per SNP that yielded a result ranged from 6.5X to 564X (Figure 10) with only 13 of the SNPs displaying an average depth of coverage <50X. These data supported that typing reference samples with a large battery of markers is feasible. However, one cannot confirm that all SNPs were typed correctly without an orthogonal approach. A subset of these SNPs (i.e., 95 SNPs), however, could be compared with a panel of SNPs (i.e., Ion AmpliSeq™ HID SNP panel v1) and some inference on typing accuracy can be obtained. The success in typing is described below.

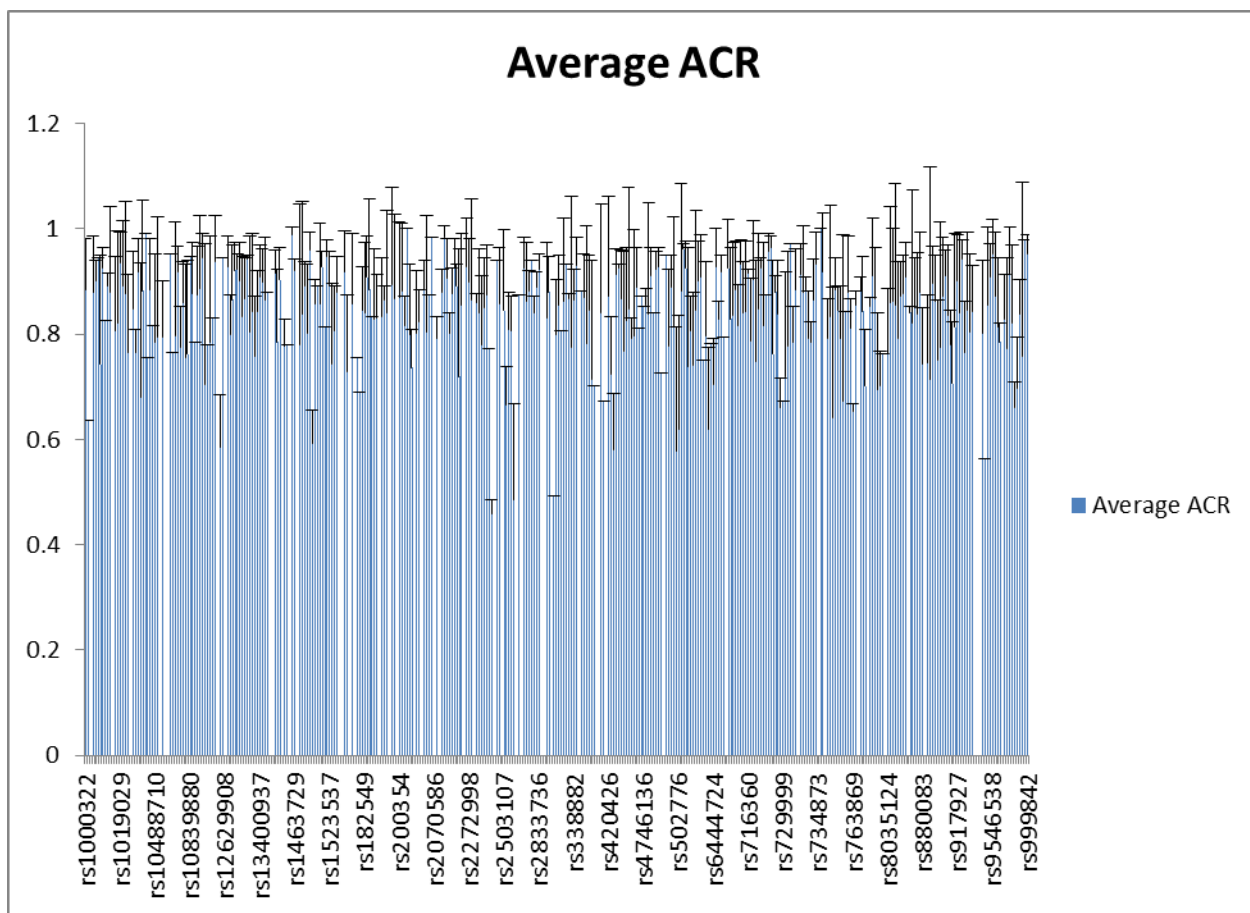


Figure 9. Average heterozygote coverage ratios for the 328 SNPs. The bars represent standard deviations.

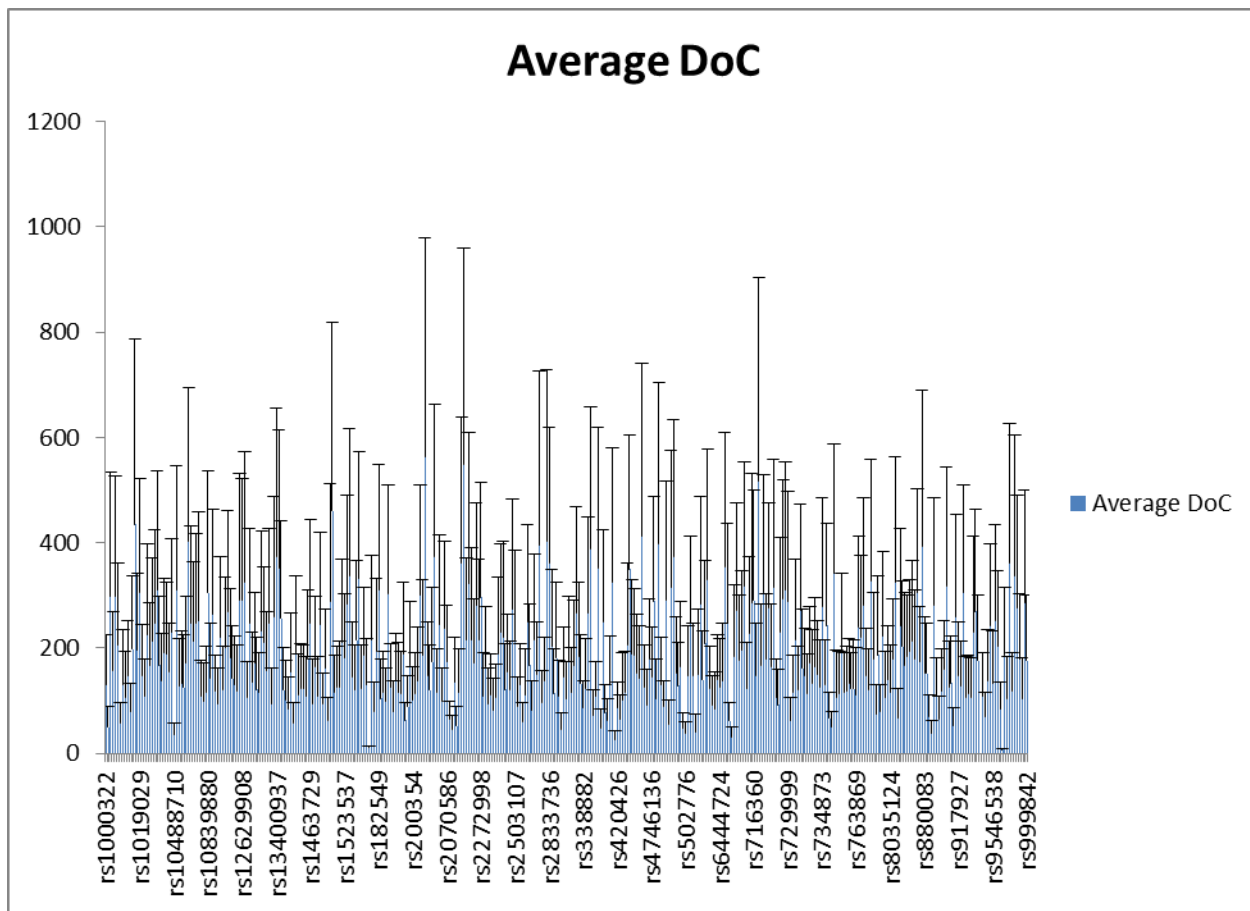


Figure 10. Average depth of coverage across the 379 SNPs.

X. PGM SNP Panel - Ion AmpliSeq™ HID SNP panel (v1) Methods

The Ion AmpliSeq™ HID SNP panel (v1), a primer pool of 103 autosomal SNPs and 33 Y-SNPs, was evaluated using the Ion 314™ Chip on the Ion PGM Sequencer with four DNA samples. The study focused on the sequencing of DNA at three different initial target quantities and related interpretation issues. Overall, the data supported that genotyping a large battery of SNPs is feasible with the PGM MPS and data were highly concordant with Illumina-based data from our initial in-house panel.

Sample Preparation

DNA was extracted from whole blood of four volunteers (one female, three males) with informed consent. The QIAamp DNA Blood Mini Kit was used for DNA extraction. The quantity of extracted DNA was estimated using the Qubit® dsDNA BR Assay Kit on a Qubit® 2.0 Fluorometer and the Quantifiler® Human DNA Quantification Kit on an ABI Prism® 7500 Sequence Detection System.

Human Identification SNP Primer Pool

The 2X Ion AmpliSeq™ HID SNP panel primer pool (panel: HID_SNP_v0.1) (ThermoFisher) was used for this study. This panel was designed to amplify 103 autosomal SNPs and 33 Y-SNPs. Information on the primer pool is described on Ion Community (<http://ioncommunity.lifetechnologies.com/community/applications/hid/snps>).

Library Preparation

This library preparation is different than that of the TruSeq and HaloPlex methods in that the fragments were generated by PCR. Thus, similar to that of HaloPlex the fragments have defined starting points for sequencing. The fragment size and starting points for sequencing are fixed (based on the primers used for PCR of the SNP). However, in contrast to HaloPlex the starting points can be designed to be sufficiently close to the target by positioning of the PCR primers. PCR also has the desirable feature of increasing the sensitivity of detection and in turn reducing the amount of template DNA required for analysis. To amplify the targeted 136 SNPs, 10 ng, 1 ng and 100 pg of genomic DNA were used for each of the four samples. PCRs were prepared using the Ion AmpliSeq™ Library Kit 2.0 and 2X Ion AmpliSeq™ HID SNP panel containing the pool of PCR primers on a GeneAmp® PCR System 9700 following the manufacturer's recommended protocols (115). The PCR conditions for 10 ng of template DNA were 2 min at 99°C for polymerase activation and 18 cycles of 15 sec at 99°C for denaturation and 4 min at 60°C for annealing/extension. For 1 ng of template DNA, the amplification cycles were increased to 22 cycles. For 100 pg of template DNA, the amplification cycles were carried out at 26 and 28 cycles and conducted in duplicate. The resulting amplicons were treated with FuPa Reagent (ThermoFisher) to partially digest primers. Amplicons then were ligated to Ion P1 and Ion Xpress™ Barcode adapters (ThermoFisher) and purified using Agencourt® AMPure® XP Reagent (Beckman Coulter, Brea, CA). Barcoded libraries were assessed by quantitative PCR with the Ion Library Quantitation Kit (ThermoFisher) following the recommended protocol (116) and diluted to ~20 pM. Equal volumes of the four diluted libraries were combined.

Template Preparation

The diluted library (20 µl) was used to generate template-positive Ion Sphere™ Particles (ISPs) containing clonally amplified DNA. Emulsion PCR was conducted by using the OneTouch™ 200 Template Kit v2 DL with the Ion OneTouch™ DL configuration (ThermoFisher) following the recommended protocol (117). Template-positive ISPs were enriched with the Ion OneTouch™ ES (ThermoFisher). Quality of template-positive ISPs was assessed by using the Ion Sphere™ Quality Control Kit (ThermoFisher) on the Qubit® 2.0 Fluorometer.

Sequencing and Data Analysis

The PGM was the MPS instrument used in this analysis. Libraries were sequenced on the Ion 314™ Chip with the Ion PGM™ 200 Sequencing Kit (Life Technologies) following the recommended protocol (118). The plugin “HID SNP Genotyper” on the Ion Torrent server and IGV were used for data analysis. The reference genome was Hg19.

XI. SNP Panel Assessment Results and Discussion

Full details of the findings were published by Seo et al (119). Results showed that at 10 ng of template DNA, there was consistently high coverage with little variation between samples.

Genotypes at all SNP loci were obtained for all samples. Genotypes of Y-SNPs were not detected in the female sample. However, variation in coverage was observed among the SNPs. Each SNP generally showed similarly high or low coverage across the samples. The lowest coverage was at rs2072422 (a Y-SNP) at 5-9X in the three male DNA samples. The consistent (among individuals) variation in coverage may be due to primer design and PCR amplification efficiency and may be adjusted with modified primers and/or PCR conditions.

The average coverage of autosomal SNPs with 1 ng of template DNA was comparable to the 10 ng samples for the female and lower for the male DNAs. The average coverage of Y-SNPs for 1 ng samples was lower than 10 ng male samples. Most SNP genotypes were detected. There were 2-6 SNPs and 1-5 SNPs not detected in sample nos. 3, 4 and 4275 using 26 and 28 PCR cycles, respectively. For 26 and 28 PCR cycles, an average of 5.5 SNPs and 3 SNPs showed heterozygote imbalance of <20%, respectively. Most SNP genotypes were detected correctly at 1 ng of template DNA. One SNP in one sample demonstrated extreme heterozygote imbalance on allele coverage. At rs13218440 in sample no.4, the true genotype was AG, and allele coverage of A and G was 452X and 30X, respectively. The allele coverage ratio was calculated by dividing the coverage of one allele (showing lower coverage) by the coverage of the other allele (showing higher coverage) (ex. $30/452=6.6\%$ at rs13218440 in sample no.4). For the average coverage of autosomal and Y-SNPs with 100 pg of DNA for 26 and 28 cycles, overall, observed allele coverage was higher for 26 PCR cycles than for 28 cycles. However, this coverage difference did not appear to be directly related to the number of complete locus drop-out and heterozygote imbalance events. For more details the reader should see Seo et al (119).

No detectable pattern of heterozygote imbalance was observed across the SNP loci (although the sample size may be too small at this time to identify any patterns), other than those clearly low-performing loci with low coverage at 10 ng of DNA. This phenomenon was more severe in the results from 100 pg of template DNA than those from 1 ng of template DNA. Most discordant genotypes were due to heterozygote imbalance, resulting in changes from heterozygous genotypes to apparent homozygous genotypes. However, one SNP showed a homozygous genotype that changed to a heterozygous genotype. At rs576261 in sample no. 1, the genotype was designated as AC (assumed true type: CC). The number of reads of A and C was 108 and 210, respectively. A possible reason for this observation might be contamination (i.e., allele drop-in), as might be expected with an assay with high sensitivity.

The observed average allele coverage ratio with 10 ng, 1 ng, 100 pg (26 cycles) and 100 pg (28 cycles) of template DNA was $89.6\pm 11.3\%$, $70.7\pm 18.3\%$, $60.4\pm 21.1\%$ and $63.2\pm 21.6\%$, respectively. This balance remained relatively similar among samples at each template amount. In tests with 10 ng of template DNA, SNPs displaying imbalanced allelic coverage ratios (< 60%) were in sample no.1 at rs1029047 and rs4530059 which showed allele coverage ratios of 54.6% and 33.7%, respectively; in sample no.3 at rs4530059 and rs576261 which showed allele coverage ratios of 31.2% and 57.4%, respectively; in sample no.4 at rs4530059 which showed allele coverage ratio of 32.2%; and in sample no. 4275 at rs576261 which showed an allele coverage ratio of 52.7%. Since two of these imbalanced SNPs were seen in multiple samples, heterozygote allelic imbalance may be attributed to a primer mismatch.

Concordant SNP Results by Orthogonal MPS Testing

As with STR and mtDNA typing, the high throughput of MPS technology makes it difficult to verify typing results with standard CE-based methods, as the latter method does not have sufficient throughput. Concordance typing is more efficient for determining correct typing results using two MPS systems that use different chemistries (i.e., orthogonal testing). Of the 103 autosomal SNPs in the Ion AmpliSeq™ HID SNP panel (Table 17) there were 95 SNPs in common with our in-house panel (described above). The SNPs rs10495407, rs10768550, rs901398, rs2175957, rs4789798, rs689512, rs2292972 and rs9606186 were not included in the in-house panel.

Table 17. 103 autosomal SNPs in Ion AmpliSeq™ HID SNP panel v1

Chromosome	Position	Target ID	Chromosome	Position	Target ID
chr1	4367323	rs1490413	chr11	11096221	rs901398
chr1	14155402	rs7520386	chr11	105912984	rs6591147
chr1	160786670	rs560681	chr11	122195989	rs590162
chr1	238439308	rs10495407	chr12	888320	rs2107612
chr1	239881926	rs891700	chr12	6909442	rs2255301
chr1	242806797	rs1413212	chr12	6945914	rs2269355
chr2	114974	rs876724	chr12	106328254	rs2111980
chr2	182413259	rs12997453	chr12	130761696	rs10773760
chr3	961782	rs1357617	chr13	22374700	rs1886510
chr3	59488340	rs9866013	chr13	84456735	rs9546538
chr3	113804979	rs1872575	chr13	100038233	rs1058083
chr3	190806108	rs1355366	chr13	106938411	rs354439
chr3	193207380	rs6444724	chr14	25850832	rs1454361
chr4	76425896	rs13134862	chr14	98845531	rs873196
chr4	157489906	rs1554472	chr14	104769149	rs4530059
chr4	169663615	rs6811238	chr15	39313402	rs1821380
chr4	190318080	rs1979255	chr16	5606197	rs729172
chr5	2879395	rs717302	chr16	5868700	rs2342747
chr5	17374898	rs159606	chr16	78017051	rs430046
chr5	136633338	rs13182883	chr16	80106361	rs1382387
chr5	159487953	rs7704770	chr17	41286822	rs2175957
chr5	174778678	rs251934	chr17	41341984	rs8070085
chr5	178690725	rs338882	chr17	41691526	rs1004357
chr6	1135939	rs1029047	chr17	80526139	rs2291395
chr6	12059954	rs13218440	chr17	80531643	rs4789798
chr6	55155704	rs2811231	chr17	80715702	rs689512
chr6	120560694	rs1478829	chr17	80739859	rs3744163
chr6	123894978	rs1358856	chr17	80765788	rs2292972

chr6	148761456	rs2272998	chr18	1127986	rs1493232
chr6	152697706	rs214955	chr18	9749879	rs9951171
chr6	165045334	rs727811	chr18	22739001	rs7229946
chr7	4310365	rs6955448	chr18	29311034	rs985492
chr7	4457003	rs917118	chr18	47371014	rs521861
chr7	13894276	rs1019029	chr18	55225777	rs1736442
chr7	137029838	rs321198	chr18	75432386	rs1024116
chr7	155990813	rs737681	chr19	28463337	rs719366
chr8	28411072	rs10092491	chr19	39559807	rs576261
chr8	136839229	rs4288409	chr20	16241416	rs12480506
chr8	139399116	rs2056277	chr20	23017082	rs2567608
chr8	144656754	rs4606077	chr20	39487110	rs1005533
chr9	14747133	rs2270529	chr20	51296162	rs1523537
chr9	27985938	rs7041158	chr21	16685598	rs722098
chr9	126881448	rs1463729	chr21	28023370	rs464663
chr9	137417308	rs10776839	chr21	33582722	rs2833736
chr10	3374178	rs735155	chr21	42415929	rs914165
chr10	17193346	rs3780962	chr22	19920359	rs9606186
chr10	97172595	rs1410059	chr22	23802171	rs2073383
chr10	118506899	rs740598	chr22	27816784	rs733164
chr10	132698419	rs964681	chr22	33559508	rs987640
chr11	5098714	rs10768550	chr22	47836412	rs2040411
chr11	5099393	rs10500617	chr22	48362290	rs1028528
chr11	5709028	rs1498553			

All SNP typing results using 10 ng of template DNA were concordant for the SNPs in common between the two platforms, except for the SNP rs1029047. This SNP is flanked by homopolymeric stretches, and the SNP states are the same as the homopolymer regions (TTT(T/A)AAAAAAAAA). *A priori* this SNP was suspected of posing a potential typing problem because of the continuum of flanking homopolymers. Based on the chemistry and detection system of the PGM, the intensity of the electronic signal due to pH change increases proportionally with the number of incorporated bases added (16). In theory, a homopolymer with 10 residues should produce twice the signal of homopolymer with 5 residues. However, operationally, signals generated from homopolymers with the PGM system are not entirely linear (120), and each locus with adjacent homopolymers needs to be evaluated and tested. For example, in sample no.1, the locus appeared to be heterozygous; a mixture of T and A was observed with IGV. However, 44.4% of bases showed a quality score of ≤ 10 at the locus when bases with a quality score of ≥ 4 were counted. IGV aligned bases to the reference genome are based on a 3' end alignment strategy. Therefore, misalignment could have occurred at bases at the 5' end of the homopolymers when homopolymer length was not correctly determined, e.g., the alignment at the first T position at the T stretch (TTTT). The first T was incorrectly

designated as a deletion in 22.6% of the reads when bases with a quality score of ≥ 4 were counted; in these reads, the T bases were shifted to other T positions. This observation indicated a high probability that the SNP genotype was incorrect with the PGM data. The overall data supported that the true genotype of rs1029047 for sample no. 1 is AA. The in-house panel yielded an AA type for this SNP.

The rs1029047 SNP was examined in the other samples. In sample nos. 3 and 4, 99.0% of the bases were detected as T. The TT genotype was correctly called and was concordant with the in-house-generated results. However, A deletion and A calls with low quality scores were still observed in the homopolymeric A stretch. The insertion of A, AA or TA between A and T stretches also was observed. In sample no. 4275, the portion of reads calling T and A was 67.0% and 33.0%, respectively, and the TT genotype was determined using the HID SNP Genotyper plugin. When the TTT(T/A)AAAAAAAAA region flanking of SNP rs1029047 was examined using IGV, 14.9% of the first T in the T stretch was incorrectly designated as a deletion due to a shift of bases when bases with BPQ of ≥ 4 were counted. After correcting the alignment problem, 46.0% of bases showed A and 54.0% of bases showed T at the locus. This observation indicated that the true SNP type was an A/T heterozygote. It also indicated that even with flanking homopolymers it may be possible to overcome incorrect calls with software that uses a specifically designed algorithm for alignment.

Table 18 lists those SNPs within the Ion AmpliSeq™ HID SNP panel that were proximal to homopolymers of 3 bases or more. Only SNP rs1029047 has such an extremely long homopolymer immediately flanking the site. All others demonstrated no genotyping errors, i.e., complete concordance among the two MPS platforms. As an example, a TT genotype (TTT(C/T), T stretch), was determined with a T called in 94.1% of the reads at SNP rs430046 (sample no. 4). Immediate flanking SNPs that differ from the known allelic state of a SNP could anchor alignments, further reducing error in allele calls. Consider, for example, SNP rs10092491, where a G residue lays immediately 5' to the C/T SNP. Even if an incorrect estimation of the number of homopolymer bases were to arise, the alignment could anchor on the G residue and reduce the chance of mistyping. Although no typing errors were observed with 94 of the 95 SNPs on the PGM™, it would be beneficial to review sequences around all SNPs for potential homopolymer and alignment issues. The SNPs in the Ion AmpliSeq™ HID SNP that were not in common with our in-house panel were reviewed for adjacent homopolymers and none were observed. The data supported that the calls were correctly obtained.

Table 18. SNPs within the Ion AmpliSeq™ HID SNP panel that were adjacent to homopolymers with ≥ 3 bases

SNP position	Flanking regions of each SNP
rs10092491	AATTCCAGATAGAGCTAAAACCTGAAG[C/T]TTTCCTTATAGAGATTTATCCTAGT
rs1029047	AAAAGTAAGAATTCAAGATGGTATTT[A/T]AAAAAAAAACCTCATATCTTTTTTC
rs12997453	AGATACAGGTTATCTGTATTACATTG[A/G]GTTTTTACCTACCTTTCTTGACAT
rs1357617	TTTGACTTCCCAAGCTGAATTTGGGG[A/T]GCTTGGTCATGTTTCTTATCAGCTA
rs1493232	CTATTCTCTTTTTGGGTGCTAGGCC[A/C]CAAAATAAACAGGCCTACAATAAA
rs1872575	TCAACTAAAAGAATTAGTCTAGAAGT[C/T]TTAAAGGTCACAGTTCAATTCTCTC

rs2811231	CATACCATGTATTCTTGTAGGAGATT[A/C/G/T]TTTCATGCTTATCACTGATCAACTT
rs3744163	GCAGAGAAACCTACCCTGGGGAGCCC[C/G]GGCTGCGTGGCACCCTGCCCTCTG
rs430046	TGATGTAAAAGCTTGGGAGGTGATTT[C/T]TGAGGGTAGGTGCTGGGTTTAATGG
rs4606077	AGTGTGGGATCTGACTCCCCACAGCC[C/T]ACCCAAAGCCGGGGAATCCTCACT
rs521861	CTCTTGAGTACATGGTTGACATTTGG[C/G]CATTTTATAGGTCCAGCAGATGGCT
rs576261	TCCGTGTACCACCTTCTCTGTACCA[A/C]CCCTGGCCTCACAACCTCTCTCCTTT
rs727811	TCTCTTACCGGAACCTCAACGACTTA[A/C]AATCATCTGCATCTCCAGCAATCT
rs733164	CACCAACAGGCCATCCCACTTGAAA[A/G]TTTGCCTGACATTCCTGAGCCGGGC
rs873196	CTGCATTCAAATCCCAAGTGCTGCCC[C/T]TTGTAATGTGAACATGCCTGATTGA

Overall, the PGM chemistry with its Ion AmpliSeq™ HID SNP panel and the in-house panel with its supporting Illumina system were quite successful in typing SNPs. The data supported that a viable panel of identity SNPs (separately or in concert with STRs) can be analyzed successfully by MPS.

XII. Updated Marker Panel

Based on the results described above, it was decided to continue with design of a multiplex STR and SNP identification panel based on the Nextera Rapid Capture system. Over the course of this project improvements occurred in target capture methodology. Moreover, longer reads (i.e., ~250) for STR typing were necessary. These features drove the design of the final panel. Technology advancements suggested that data capture was feasible at a substantially lower quantity of template DNA (i.e., 50 ng of genomic DNA) using the Nextera Rapid Capture system (Illumina) compared with 500 ng to 1 µg with the Illumina® TruSeq™ Custom Enrichment protocol). Probes (80 bases in length) for the Nextera Rapid Capture Custom Enrichment Kit were designed using Design Studio (Illumina), a freely-available software. First, 88 STRs (31 autosomal, 26 X-chromosome, 31 Y-chromosome) and 229 autosomal identity SNPs were tabulated including details regarding chromosomal positioning, target selection (Full Region), probe density requirements (due to the alignment-specific requirements of STRs, density of these markers was increased to ‘ADJACENT’) and marker information. Marker data then were uploaded to Design Studio v1.5 and probes were generated under the default conditions (with hg19 for probe reference). The multiplex panel of STRs and SNPs was finalized and probes were designed and ordered for testing.

The 88 STRs are: CSF1PO, D10S1248, D12S391, D13S317, D14S1434, D16S539, D17S1301, D18S51, D19S433, D1S1627, D1S1656, D21S11, D22S1045, D2S1338, D2S1776, D2S441, D3S1358, D4S2408, D5S2500, D5S818, D6S1017, D6S474, D7S820, D8S1179, FGA, PENTAD, PENTAE, SE33, TH01, TPOX, vWA, DXS10011, DXS10074, DXS101, DXS10101, DXS10134, DXS10135, DXS6789, DXS6795, DXS6800, DXS6801, DXS6807, DXS6809, DXS6854, DXS7132, DXS7133, DXS7423, DXS7424, DXS8377, DXS8378, DXS981, DXS9895, DXS9902, GATA165B12, GATA172D05, GATA31E08, HPRTB, DYF387S1A, DYF387S1B, DYS19, DYS385A, DYS385B, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS449, DYS456, DYS458, DYS460, DYS481, DYS505, DYS518, DYS522, DYS533, DYS549, DYS570, DYS576, DYS612, DYS627, DYS635, DYS643, and GATAH4 and Amelogenin.

The autosomal identity SNPs are: rs1000322, rs1000329, rs1003204, rs1003473, rs1004357, rs1005533, rs1008457, rs1008730, rs10092491, rs1015250, rs1017415, rs1019029, rs1019264, rs1020636, rs1021290, rs1024997, rs1027895, rs1028484, rs1028528, rs1029047, rs1031825, rs10488710, rs10495407, rs10500617, rs1058083, rs1072292, rs1075665, rs10768550, rs10773760, rs10776839, rs1079861, rs1105459, rs1105576, rs1108943, rs1109037, rs123714, rs12480506, rs1294331, rs12997453, rs13134862, rs13182883, rs13218440, rs1331494, rs1336071, rs1355366, rs1357617, rs1358856, rs1360288, rs1382387, rs138952, rs1410059, rs1413212, rs1454361, rs1463729, rs1467966, rs1478829, rs1482650, rs1490413, rs1493232, rs1498553, rs1501643, rs1523537, rs1528460, rs1542931, rs1554472, rs1593055, rs159606, rs1736442, rs174473, rs181619, rs1821380, rs18579, rs1858958, rs1872575, rs1979255, rs1997680, rs2010209, rs2013526, rs2015632, rs2016276, rs2018205, rs2046361, rs2056277, rs2073383, rs2076848, rs2107612, rs2111980, rs214955, rs2175957, rs221956, rs2255301, rs2269355, rs2270529, rs2272998, rs2291395, rs2292972, rs2311046, rs2342747, rs2388618, rs2399332, rs2503107, rs251934, rs2567608, rs279844, rs2811231, rs2830795, rs2831700, rs2833736, rs2920816, rs315791, rs321198, rs338882, rs354439, rs3744163, rs3780962, rs420426, rs4288409, rs430046, rs4364205, rs445251, rs4530059, rs4606077, rs464663, rs4789798, rs4796362, rs4847034, rs502776, rs521861, rs560681, rs5746846, rs576261, rs585070, rs590162, rs595601, rs627119, rs6444724, rs6591147, rs6811238, rs689512, rs6955448, rs702490, rs7041158, rs716360, rs716856, rs717302, rs719211, rs719366, rs7205345, rs722290, rs7229946, rs727206, rs729172, rs729549, rs729999, rs730013, rs730249, rs730488, rs730907, rs732889, rs733023, rs733164, rs734295, rs734656, rs734664, rs734701, rs735155, rs736210, rs737168, rs737681, rs738518, rs738532, rs740598, rs740910, rs743018, rs749270, rs7520386, rs763869, rs765533, rs7704770, rs772436, rs8037429, rs8070085, rs8078417, rs826472, rs868432, rs873196, rs873289, rs876724, rs877228, rs880083, rs887754, rs891700, rs892503, rs896499, rs901398, rs907100, rs911621, rs914165, rs916388, rs917927, rs919023, rs921269, rs922992, rs924181, rs924397, rs927628, rs938283, rs9546538, rs959566, rs9606186, rs964681, rs985492, rs9866013, rs987640, rs9905977, rs993934, rs9951171, rs997556, rs997568, rs997750, rs999842, P256, rs1024116, rs1335873, rs1886510, rs2040411, rs722098, rs727811, and rs917118.

Library Preparation and Sequencing

DNA samples (n=94) were collected following the University of North Texas Health Science Center IRB approval. The samples were obtained from 16 African American females, 13 African American males, 20 Caucasian females, 12 Caucasian males, 17 Hispanic females, and 16 Hispanic males. These 94 samples were prepared for sequencing using the Nextera® Rapid Capture Custom Enrichment (Illumina, Inc.) protocol. The quantity of DNA for each sample was determined using the Qubit® platform, according to the manufacturer's protocol. After normalization to 10 ng/μL with 10mM Tris-HCl (pH 8.5), the quantity of DNA was determined again and normalized to 5 ng/μL, to ensure an accurate dilution. 10 μL of each sample (50 ng total DNA per sample) were then subjected to tagmentation, where the DNA is simultaneously fragmented and tagged with adapters, in an Eppendorf Thermomixer® (Eppendorf AG, Hamburg, Germany) at 58° C for 10 minutes. Following tagmentation, the samples were washed with two 80% ethanol washes, and the resulting fragment sizes were verified using an Agilent

Technologies 2200 TapeStation™ (Agilent Technologies, Inc.). The samples then were indexed and amplified in an Eppendorf Mastercycler® Pro S (Eppendorf AG) thermal cycler, using the following PCR parameters: 72° C for 3 minutes, 98° C for 30 seconds, 10 cycles of 98° C for 10 seconds, 60° C for 30 seconds, and 72° C for 30 seconds, 72° C for 5 seconds and a final hold at 10° C. With indexing each sequencing run would consist of 11-12 samples multiplexed together. Amplification was followed by two 80% ethanol washes, and amplification success was verified using the Agilent 2200 TapeStation™. At this point, the samples were pooled, and the first hybridization of the custom oligonucleotide probes was performed in an Eppendorf Mastercycler® Pro S thermal cycler, using the following parameters: 95° C for 10 minutes, 18 cycles of 1 minute incubations, starting at 94° C and decreasing 2° C per cycle, and final hold at 58° C for 3 hours. The first streptavidin magnetic bead capture reaction was performed by mixing the beads with the hybridized samples at 1200 rpm for 5 minutes and incubating them at room temperature for 25 minutes. The samples then underwent 2 heated washes at 50° C for 30 minutes, and the DNA targets were eluted. The second hybridization reaction was performed in an Eppendorf Mastercycler® Pro S thermal cycler, using the same parameters, except that the final hold at 50° C was extended to a minimum of 14.5 hours. The second hybridization cleanup and elution reactions were performed as described above. The samples then were washed twice with 80% ethanol. A final amplification was performed in an Eppendorf Mastercycler® Pro S thermal cycler, using the following parameters: 98° C for 30 seconds, 12 cycles of 98° C for 10 seconds, 60° C for 30 seconds, and 72° C for 30 seconds, 72° C for 5 minutes, and a final hold at 10° C. Following amplification, the samples were washed twice with 80% ethanol. Library validation was performed on the Agilent 2100 Bioanalyzer™. Quantification was performed using the Qubit® platform, according to the manufacturer's protocol. The pooled libraries were normalized to 2 nM and then diluted to 12 pM for paired-end sequencing (2x250 bp reads) on the MiSeq.

STR data analysis was performed by using STRait Razor v2 to process the FASTQ files generated by the MiSeq Reporter software. The included Razor Genotyper workbook was used to produce genotypic information, depth of coverage values, and heterozygote balance statistics from the STRait Razor output. SNP analysis was performed by processing the BAM files output by MiSeq Reporter with the GATK. Genotypes, depth of coverage information, and heterozygote balance values for these markers were calculated using the resulting VCFs.

Results and Discussion

These panel probes were used to analyze 94 different individuals to assess the general performance of the large capture-based multiplex. A tremendous amount of data is generated with these analyses. Therefore, to present the performance information summary charts were generated on depth of coverage and, where appropriate, heterozygote balance (termed here also as allele coverage ratio). For STRs the data were separated into autosomal, Y chromosome, X chromosome male, and Y chromosome female (Figures 11-16). For SNPs the data were

separated into autosomal and Y chromosome (Figures 17-19). The overall performance of depth of coverage and heterozygote is similar to that of commercial PCR-based MPS kits. For all marker systems the depth of coverage ranged from some low signal loci to high signal loci. These extremes are a small subset of the total markers, and the majority are well-balanced (i.e., within 2 SD of the mean; calculations not shown). With a CE-based approach having such a wide range of signal in a multiplex would not be feasible because the largest signal loci typically would be blown out. However, the dynamic range with MPS is much greater as the signal (and concomitant noise) from one marker does not directly affect the signal at another marker. So the range of coverage seen in our capture panel (and commercial PCR-based MPS kits) can be accommodated easily. The limitation on such a wide range of coverage is sample throughput. Detection of lower performing markers will drive the number of samples that can be run to ensure that routine typing will not result in unreasonable amount of allele and locus dropout. Future studies could improve the balance by increasing probe density for the lower signal markers from ADJACENT to OVERLAPPING and the highest signal markers from ADJACENT to INTERMEDIATE or STANDARD. A more balance system will increase sample multiplexing capability (which cannot be predicted until a new probe panel is produced).

For the autosomal STRs, the D14S1434 locus accounted for 80% of the total locus dropout (8 out of 10 total dropouts in Reads 1 and 2 combined). For the X-STRs in females, the GATA165B12 locus accounted for 43.8% of the total dropout (7 out of 16 total dropouts in Reads 1 and 2 combined). The next most prevalent locus dropout was observed at the DXS6809 locus, which accounted for 25% of the total dropout (4 out of 16 total dropouts in Reads 1 and 2 combined). For the X-STRs in males, the DXS6809 locus had the highest dropout with 22% of the total dropout (20 out of 91 total dropouts in Reads 1 and 2 combined), and the DXS10134 locus was second, accounting for 16.5% of the total dropout (15 out of 91 total dropouts in Reads 1 and 2 combined). The DXS101 locus accounted for 12.1% of the total dropout (11 out of 91 total dropouts in Reads 1 and 2 combined), while the DXS6789 locus accounted for 11% of the total dropout (10 out of 91 total dropouts in Reads 1 and 2 combined).

For the Y-STRs, only 1 locus stood was low performer based on locus dropout, i.e., the DYS448 locus, which accounted for 73.5% of the total dropout (25 out of 34 total dropouts in Reads 1 and 2 combined).

Most of the low-performing autosomal SNP loci only dropped out in 1 or 2 samples (more likely due to the overall low signal in these samples). The two autosomal SNPs with the highest dropout rates were rs502776, which accounted for 17.5% of the total dropout (10 out of 57 total dropouts), and rs1406945, which accounted for 7% of the total dropout (4 out of 57 total dropouts). As for the Y-SNPs, rs16980360 and rs34486382 were the only ones that dropped out in more than one sample, each accounting for 25% of the total dropout (2 out of 8 total dropouts each).

Allele coverage ratios (or heterozygote balance) were quite good for the vast majority of loci with autosomal STRs performing slightly better than X chromosome STRs. The SNPs were well-balanced as would be expected. Heterozygote balance with the capture panel is similar to that of CE-based systems and other commercial PCR-based MPS kits (data not shown).

Issues that arose point to the need of information curation, to be cognizant of limitations of the various components of a system, and to appreciate how limitations can impact assessment of the performance of a panel. Although not displayed as a low performer in the Figures above, the STR locus D5S2500 initially suffered from an apparent high degree of dropout. In fact this one locus accounted for 94.5% of the total dropout (154 out of 163 total dropouts) for the autosomal STR loci in the 94 samples. The cause of the dropout was not due to the probe design per se. Indeed, the probes for the D5S2500 locus actually performed quite well. The coordinates for the D5S2500 locus were based on consistent data from a number of sources (121-124). These same coordinates were used to design the primers for the D5S2500 locus in the Qiagen Investigator HDplex (125). However, the flanking regions for this locus that were used in STRait Razor were derived from a different source, i.e., Hill et al (126,127). There is discordance between the coordinates used to design the probes in the panel and the coordinates described by Hill et al (126,127). This discordance was supported by a difference in reported genotype for the 9947A cell line. The Hill et al result was 14,23 and the other groups reported 15,16 (128,129). There are two different STRs being identified by these sources. To address the false dropouts in our study STRait Razor was reconfigured to identify flanking regions for the coordinates used in the panel, and then there was no evidence of dropout for the D5S2500 locus. At this time, it is not clear which of the two sites is the correct D5S2500 locus. However, such discrepancies do point out how false conclusions can occur about the performance of a marker(s) and during developmental stages of methods some review by alternate means may be warranted. In addition, one should be aware of limitations of STR calling software, STRait Razor included. Reads for any marker will only be identified if they are configured in the software. Therefore, as was done in our study, all apparent STR dropouts should be reviewed manually for developmental work.

Other examples of apparent locus dropout that were not due to the chemistry of the system or a sample with overall low signal were at the loci GATA172D05, DXS981, and DYS518. For the GATA172D05 locus the dropout was due a STRait Razor configuration file. The allelic definitions lacked 10 bases in the offset value and thus some alleles were not detected. 10 bases were added to the offset value, which eliminated dropout at this locus. For the DXS981 locus STRait Razor was configured correctly. However, the configured flanking regions were set unnecessarily far apart and therefore a number of reads that did span the whole region between the flanks were missed. The distance between the flanking regions was shortened which completely eliminated dropout at this locus. For the DYS518 locus, nomenclature/repeat structure was based on old repeat motif data. The definitions for this locus were changed to comport with an alternate nomenclature and dropout was reduced.

The data herein support that a capture-based approach can produce robust data for typing reference samples. A large set of markers and different types of markers can be typed simultaneously; thus the potential for gaining substantially more data in a single analysis is demonstrated. A very few loci were low performers and their signals likely could be increased, if desired, by increasing probe density in the design phase. The main difference between maintaining the panel in its current form and creating a more balanced depth of coverage panel (if possible) would be sample throughput (using barcoding). Fewer samples can be analyzed simultaneously with a less-balanced panel. One motivation for using a capture-based assay was that the vagaries of PCR would not impact the results. However, the data indicate that the

performance and artifacts observed with a PCR enrichment method persist with our capture-based approach. There is some locus-to-locus coverage variation; stutter does occur (data not shown) mostly due to an amplification stage prior to sequencing; heterozygote balance is similar; and low level noise exists (that can be filtered out). The artifacts of locus-to-locus signal difference, stutter and noise are not new to DNA typing and can be managed in a similar fashion as they are with CE-based systems.

Although not part of this study, but worth considering for future studies, is that a probe-based capture system may be better suited for typing degraded samples than a PCR enrichment approach. Primers define the size of a PCR amplicon. If DNA is degraded, such that the fragments are too small to generate amplicons, no PCR product will be generated. However, a probe capture system is not as limited due to the size of the fragments of DNA in a sample. Indeed, the probe design could be increased readily for ADJACENT to OVERLAPPING to enhance capture with challenged samples. Two groups – Carpenter et al (130) and Templeton et al (131) - have described a novel capture procedure that enriches highly degraded endogenous ancient genomic DNA. To make the current probe-based system in our study practical for analyzing challenged samples the amount of input DNA will need to be substantially reduced from the current amount of 50 ng. However, during the course of this project the input DNA was reduced initially from 500 ng to 50 ng, which is an order of magnitude change in template requirements. With the rapid advancements in MPS technologies and chemistries, it is anticipated that the amount of input DNA required for capture based approaches will continue to decrease.

Lastly, there are commercial PCR-based kits coming to market. These kits show great promise (data not shown). However, they require substantial effort and resource to produce similar data with what is observed with the capture-based large panel described herein. Our data indicate that development of the capture panel was much easier and required resources than the PCR-based systems. The design is simple and did not require substantial modification with the probe panel.

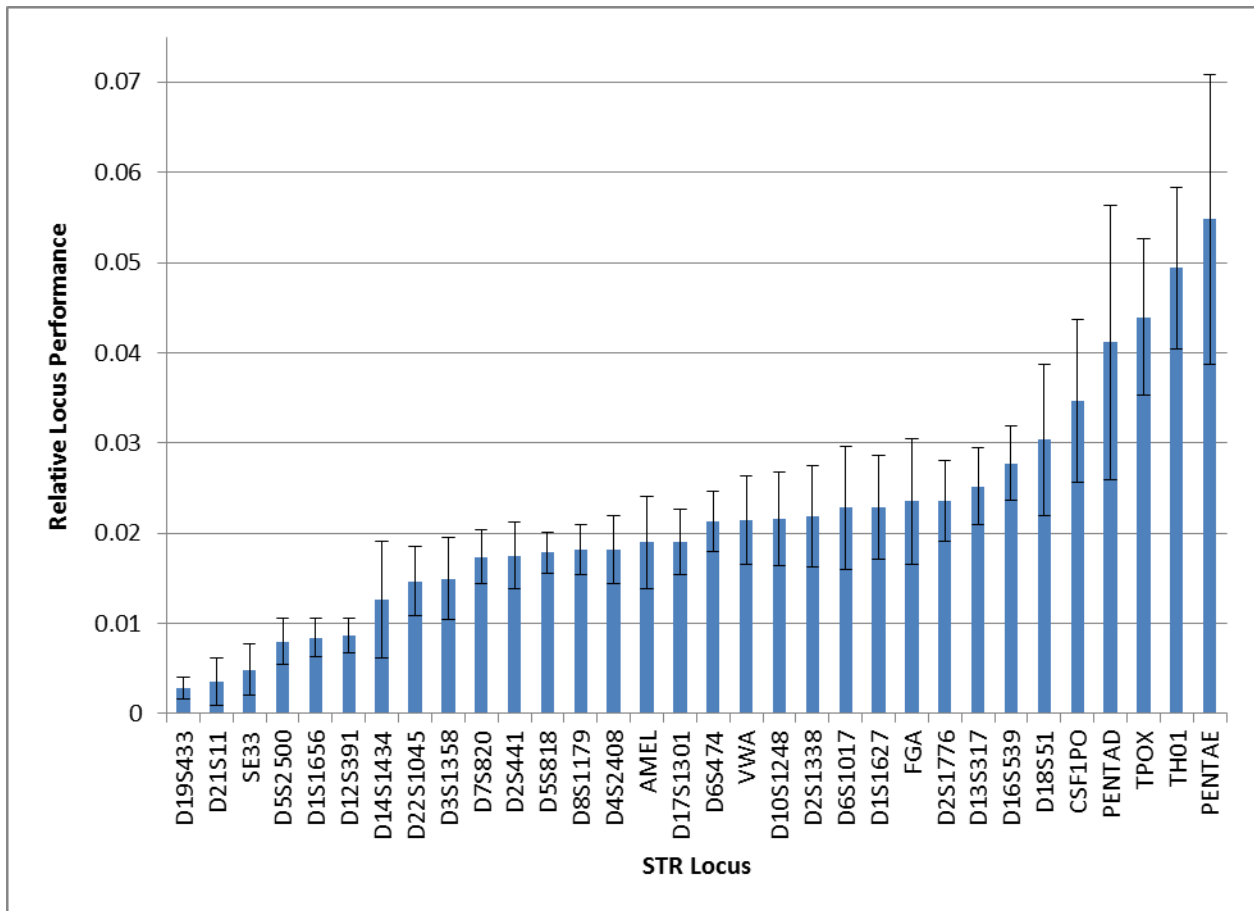


Figure 11. Relative depth of coverage for autosomal STRs. Calculated by coverage at the locus divided by total coverage across all autosomal STR loci.

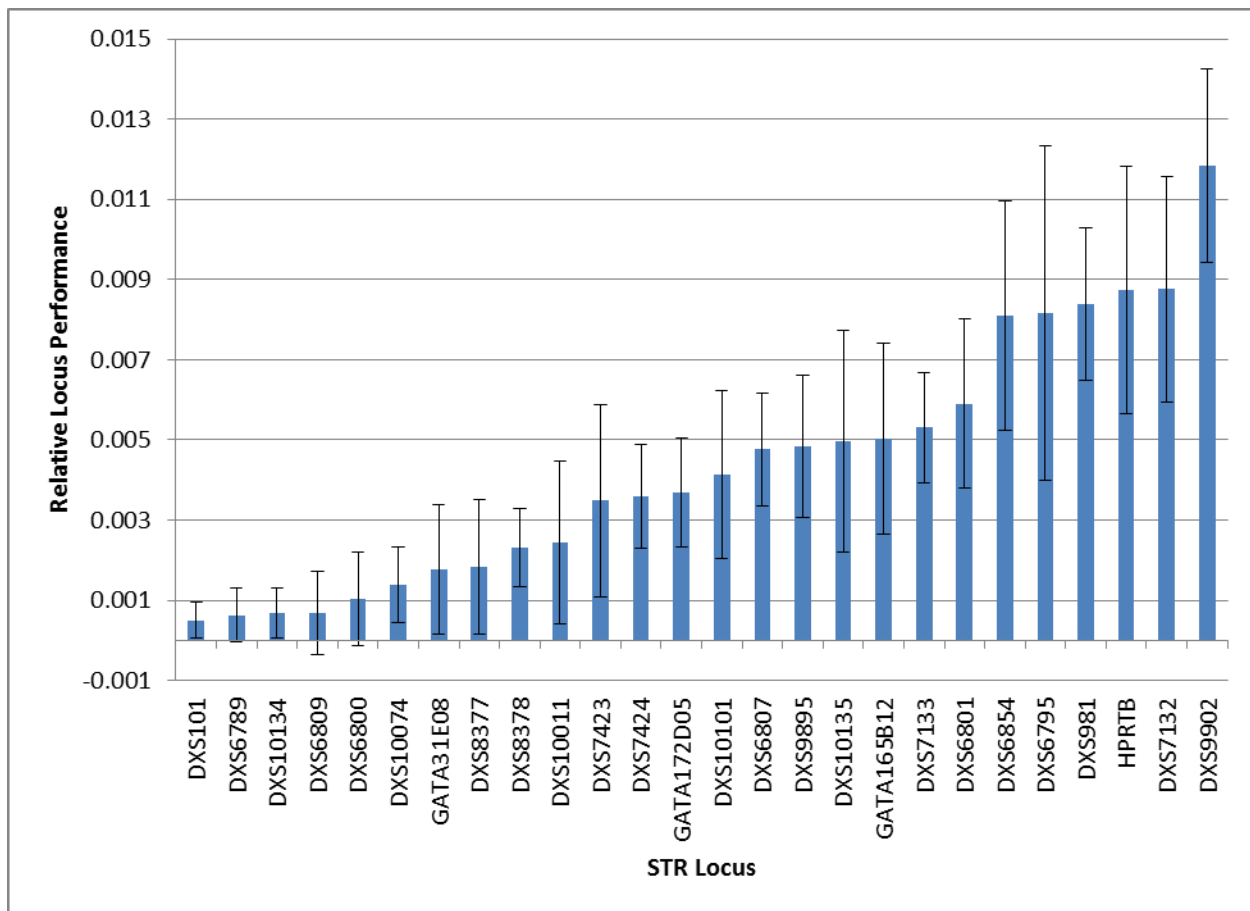


Figure 12. Relative depth of coverage for X chromosome STRs for male individuals. Calculated by coverage at the locus divided by total coverage across all X chromosome STR loci.

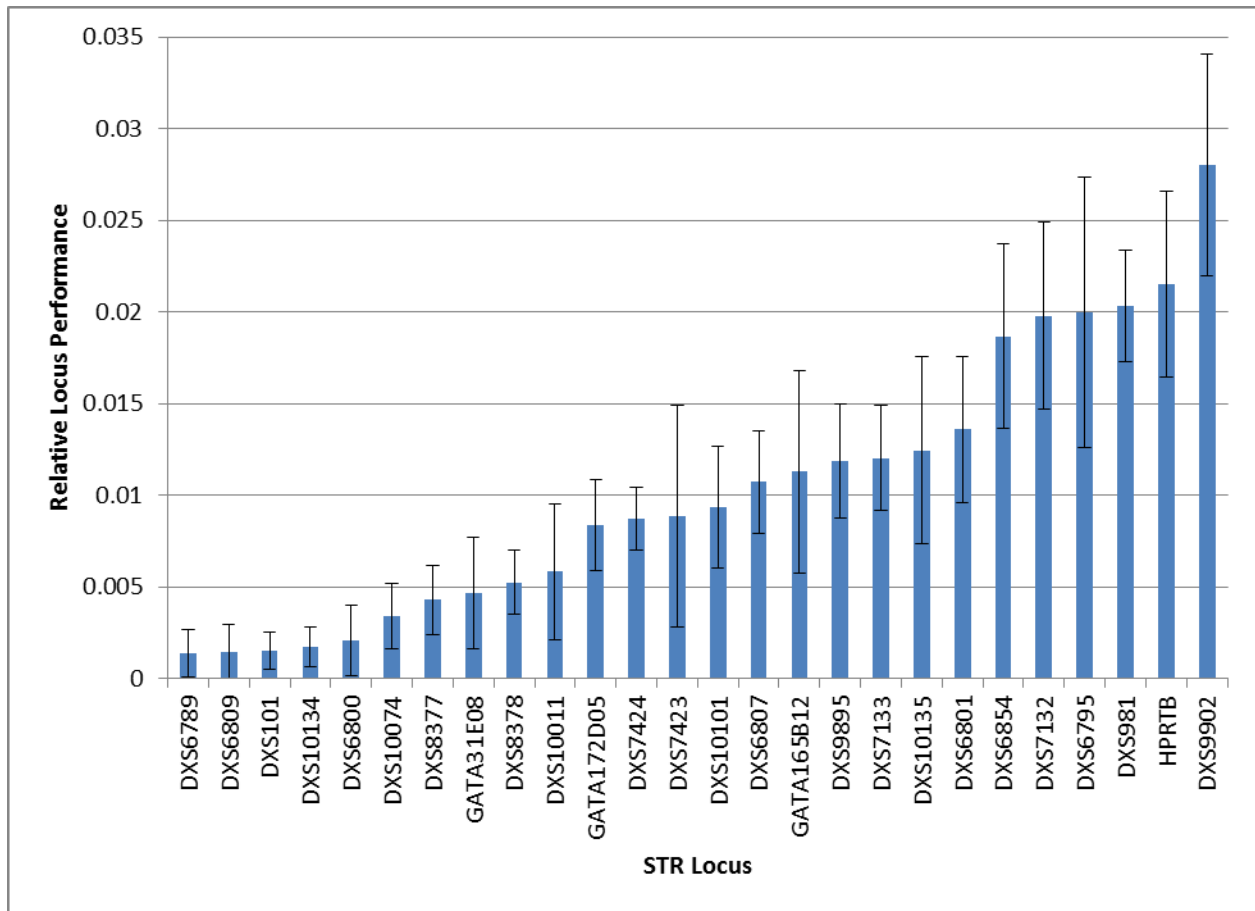


Figure 13. Relative depth of coverage for X chromosome STRs for female individuals. Calculated by coverage at the locus divided by total coverage across all X chromosome STR loci.

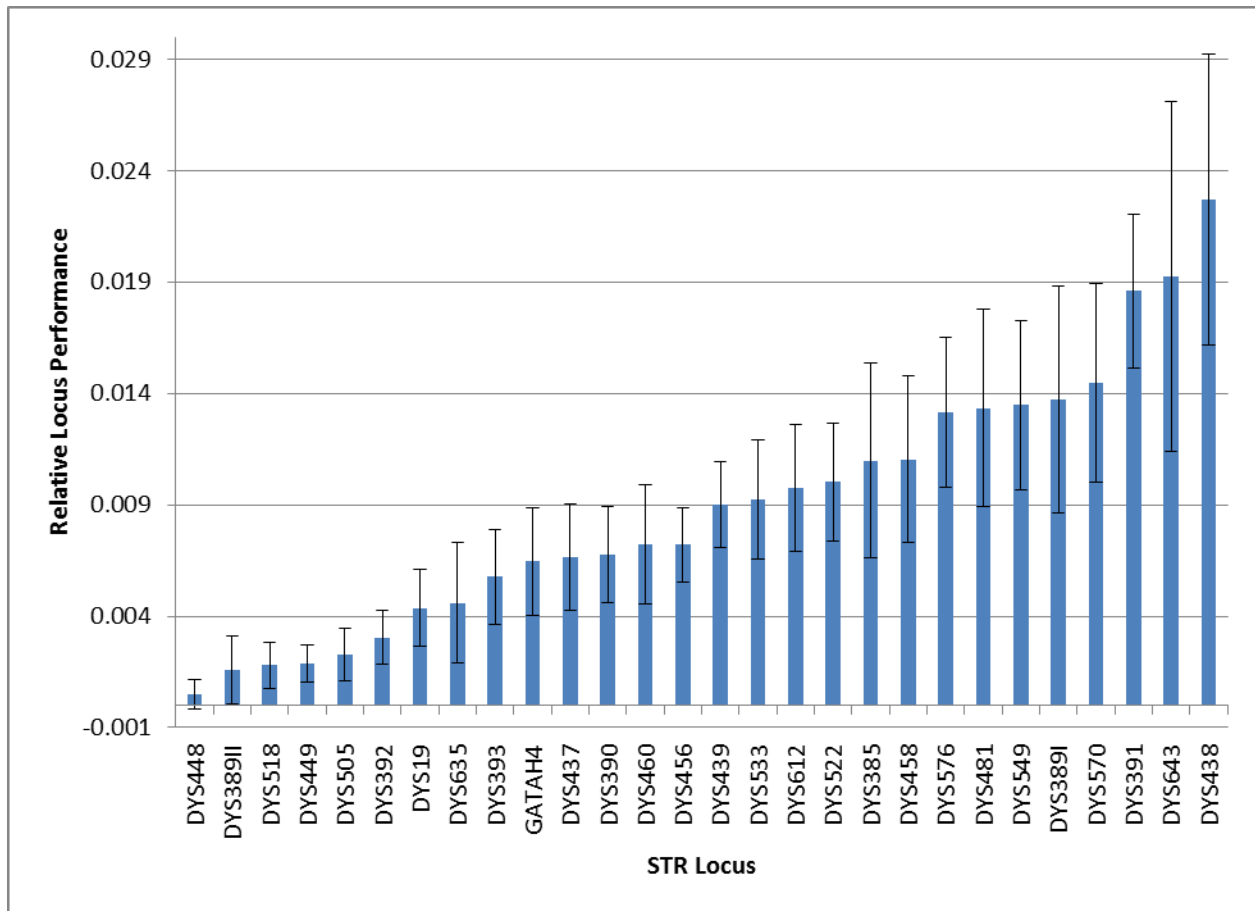


Figure 14. Relative depth of coverage for Y chromosome STRs. Calculated by coverage at the locus divided by total coverage across all Y chromosome STR loci.

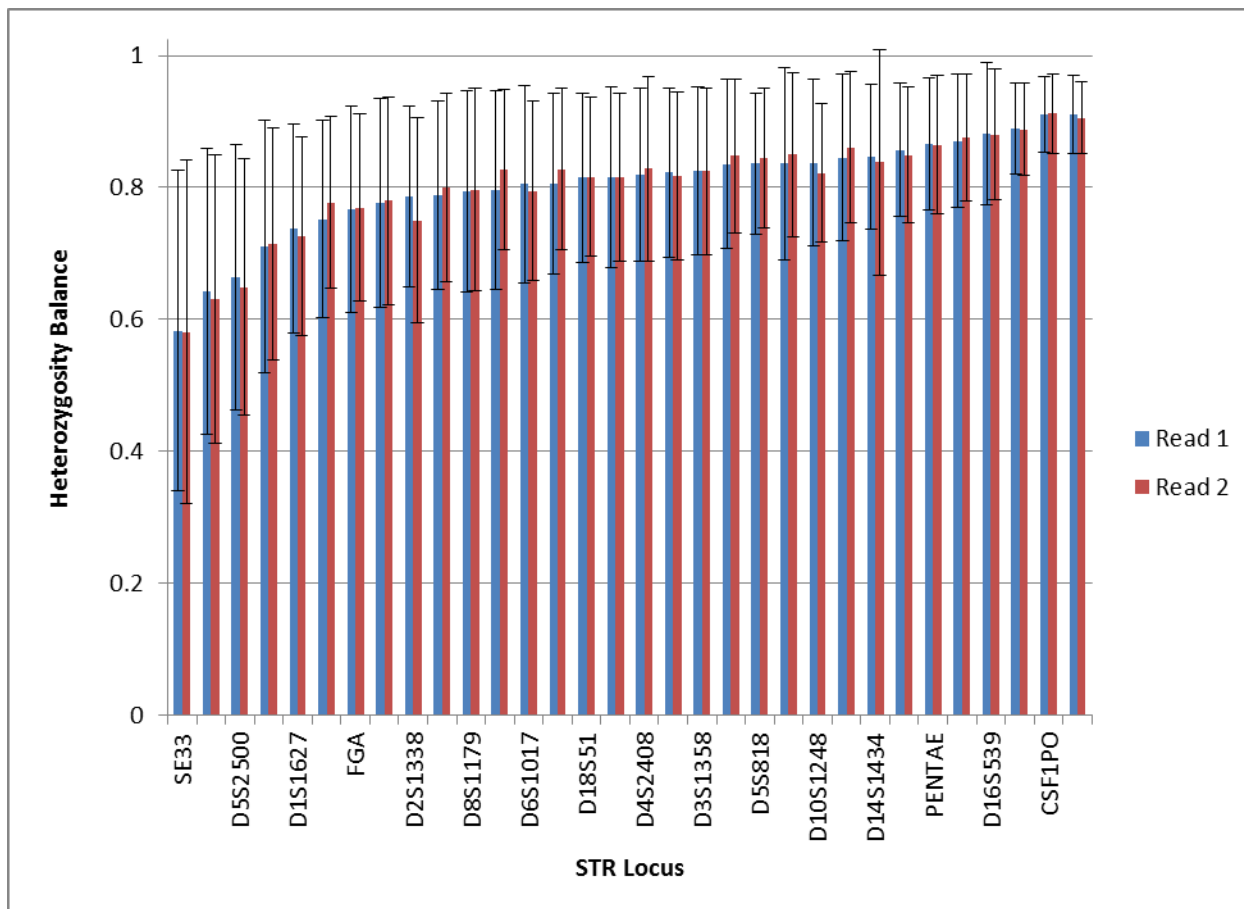


Figure 15. Heterozygote balance (or allele coverage ratio) for autosomal STRs. Calculated by the allele with lower coverage divided by the allele with higher coverage at a locus.

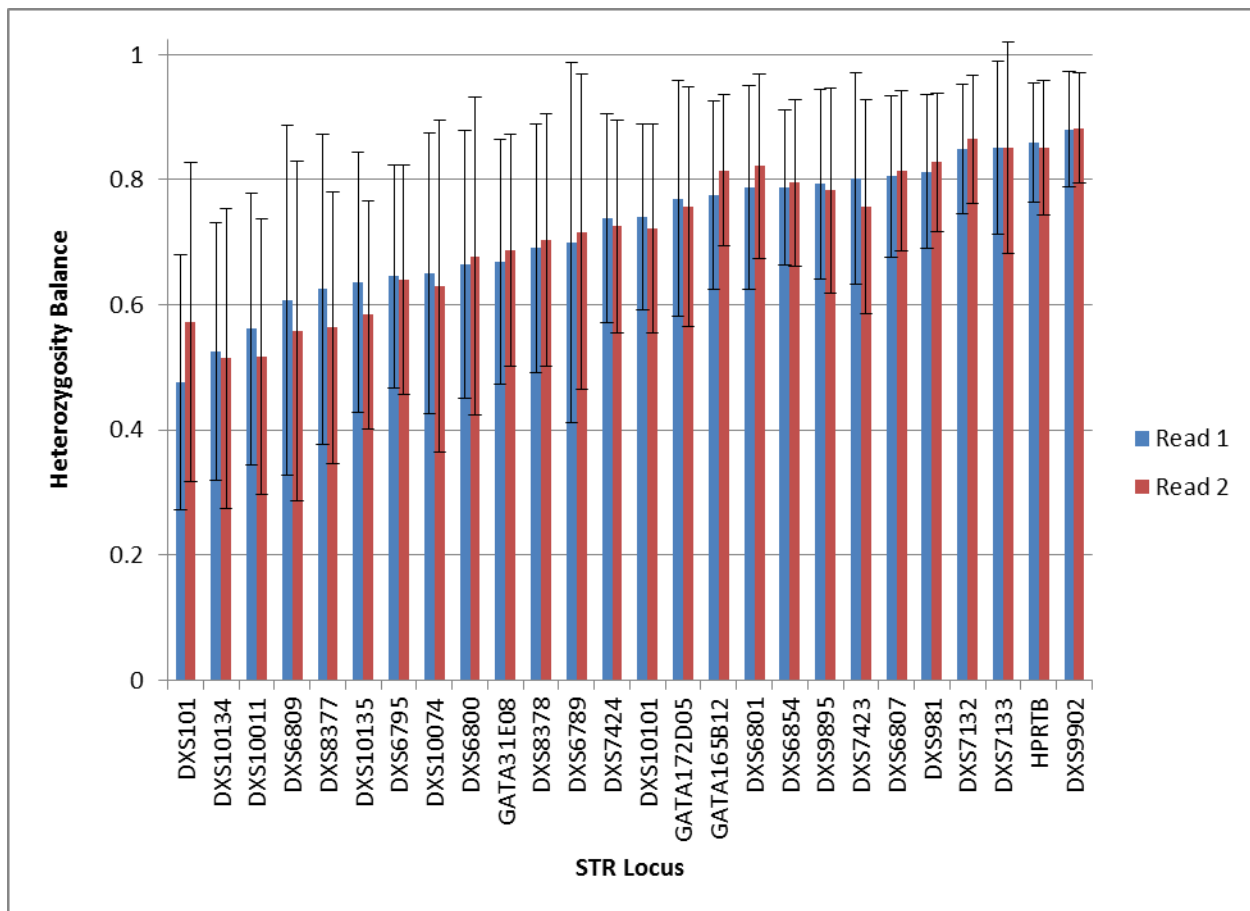


Figure 16. Heterozygote balance (or allele coverage ratio) for X chromosome STRs (only possible to calculate with female individuals). Calculated by the allele with lower coverage divided by the allele with higher coverage at a locus.

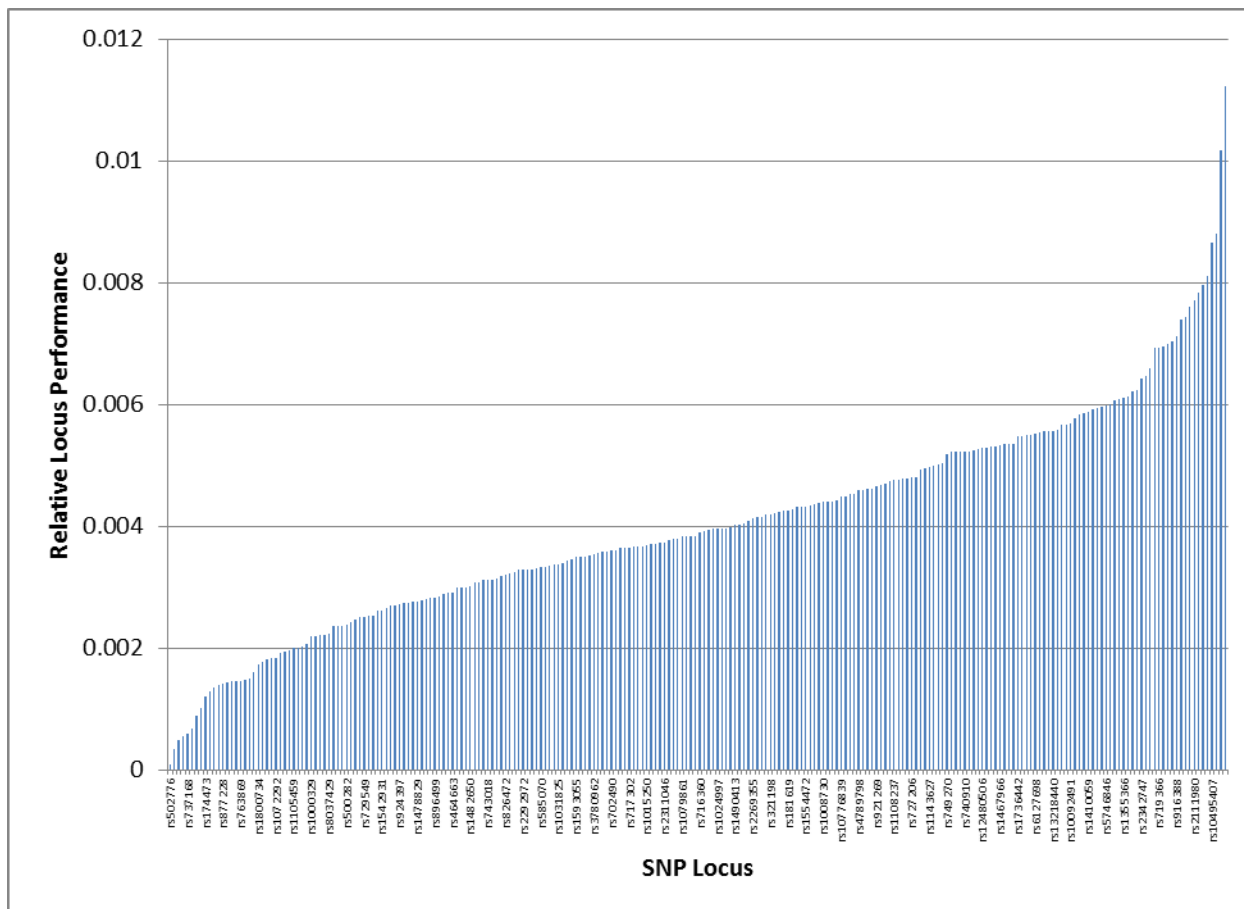


Figure 17. Relative depth of coverage for autosomal SNPs. Calculated by coverage at the locus divided by total coverage across all autosomal SNP loci.

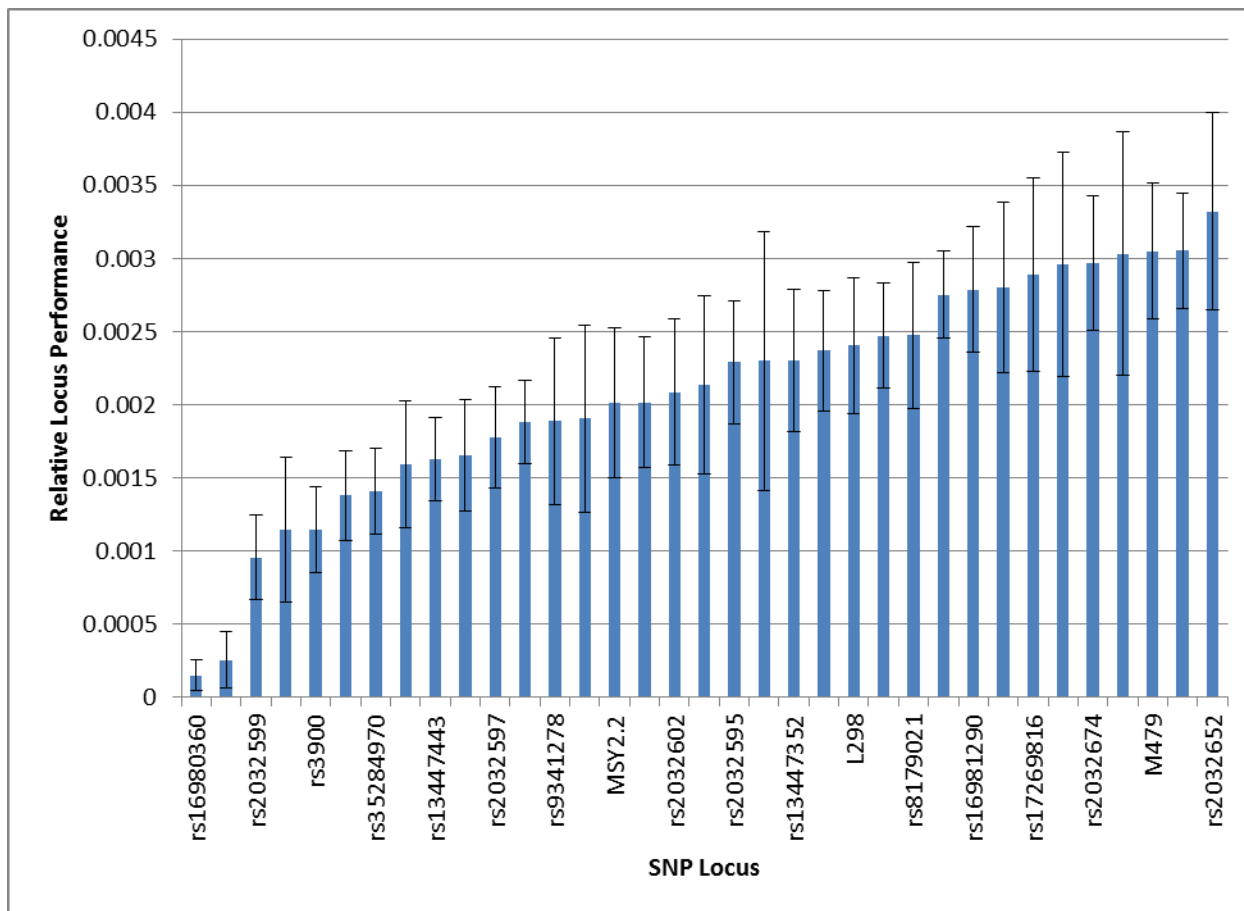


Figure 18. Relative depth of coverage for Y chromosome SNPs. Calculated by coverage at the locus divided by total coverage across all Y chromosome SNP loci.

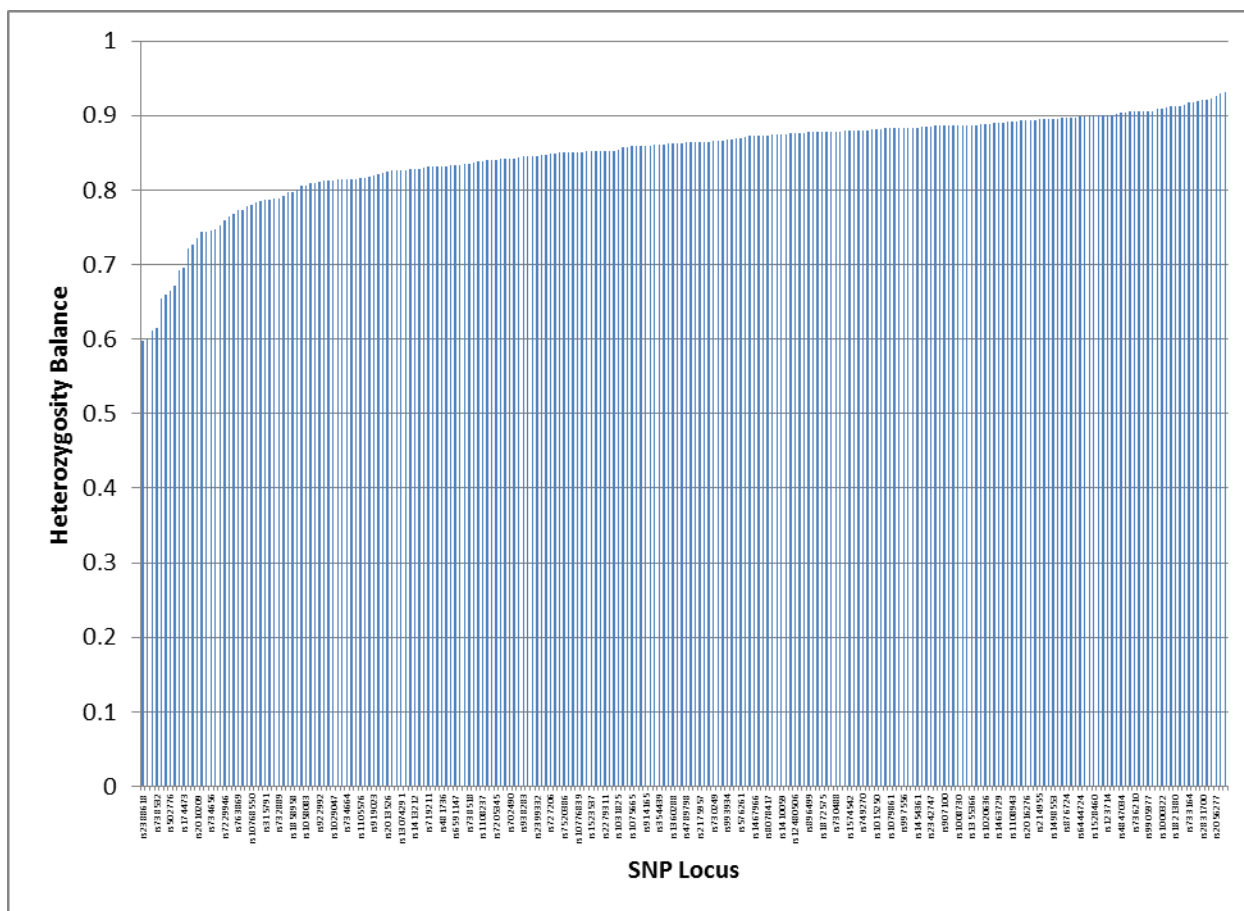


Figure 19. Heterozygote balance (or allele coverage ratio) for autosomal SNPs. Calculated by the allele with lower coverage divided by the allele with higher coverage at a locus.

XIII. Library Preparation Summary

Throughout this project different enrichment/library preparation methods were considered and tested. Four library preparations have been used, two for the Illumina system, HaloPlex, and a PCR-based one for the PGM/SNP panel and mtGenome sequencing. All library preparations described above (in various studies) were suitable for the intended purpose. However, some require more template DNA; some are more labor intensive; and some may not be able to work with a very few markers. The results and successes were described above so here they only will be summarized.

The Illumina® TruSeq™ Custom Enrichment protocol, based on a capture strategy, can target a large number of target sites simultaneously. This library preparation protocol was selected initially because PCR amplification for target enrichment was not required. Therefore, a challenging PCR multiplex primer design would not have to be accomplished and errors due to the PCR would not impact sequencing results with this library preparation methodology. There are two limitations with the TruSeq chemistry: 1) it requires a good amount of template DNA

(~50-500ng), and it is a laborious method. Since the initial panel was tested for reference sample typing, the amount of template DNA was not particularly limiting.

HaloPlex also is a capture-based approach, which requires a relatively large amount of DNA around 200 ng, also not limiting for reference sample typing. It has benefits of known start and stop points for the DNA being sequenced, higher coverage, and high sample throughput. But a small number of loci may not be compatible with the restriction enzyme cocktail that is used. If a new cocktail is developed it is likely that a few different loci would suffer. However, the loss of a couple of “core” loci can be more than compensated by the many more markers that can be contained in a MPS multiplex.

For the mtGenome sequencing protocol PCR enrichment was performed. The two amplicons are slightly larger than 8 kb in length and must be fragmented to a requisite length prior to attaching adapters. The Nextera XT DNA Sample Preparation Kit (Illumina) protocol is based on tagmentation (53). Tagmentation combines transposase activity to fragment the DNA and primers and adapter addition in one reaction. The features that make this method desirable are: 1) it requires only 1ng template DNA; 2) it can be performed in a relatively short time frame; and 3) multiple samples can be prepared simultaneously.

Libraries of mtGenomes for sequencing on the PGM were prepared in a similar fashion to the tagmentation approach in that fragmentation of the long amplicons is required (and all of the amplification product is used). The ~8 kb long PCR amplicons were enzymatically fragmented using Ion Shear™ Plus Reagents (ThermoFisher). Ion adapters and barcodes were ligated to the fragmented amplicons using the Ion Plus Fragment Library and Ion Xpress™ Barcode Adapters Kits (ThermoFisher).

The majority of published reports on MPS for potential forensic applications rely on (and for the foreseeable future will rely on) PCR for enrichment. Tagmentation, for example, requires relatively long templates. However, most large multiplex PCR panels will generate short amplicons (for forensic utility) and tagmentation or shearing will not be required. The AmpliSeq panel approach (115) is based on short amplicons serving as the input for library preparation. For the MiSeq a similar approach was needed.

A library preparation method based on chromatin immunoprecipitation sequencing (ChIP-Seq) protocol was investigated. With this technology genomic DNA is cross-linked with chromatin and enriched before being subjected to MPS (132). Traditionally, it has been used to investigate the distribution, abundance, and characteristics of DNA-bound protein targets across a genome of interest. The TruSeq™ ChIP sample preparation kit (Illumina) provides a simple workflow that allows the preparation of chromatin-bound DNA for sequencing via the attachment of TruSeq™ adapters.

In this study, the TruSeq™ ChIP protocol was modified to enable library preparation of forensically-relevant SNP-containing amplicons. This modified protocol, known as TruSeq™ Forensic Amplicon, was used to detect a battery of 160 human identification SNPs (HIDs) and AIMs in a set of 12 reference samples. The resulting data were analyzed for both sequence coverage and heterozygote allele balance.

XIV. Library Preparation Materials and Methods

The TruSeq™ Forensic Amplicon library preparation protocol recommends an amplified DNA input volume of 50 µL, at a concentration of 20-2000 pg/µL (i.e., 1-100 ng total input DNA). Amplified products generated from each PCR (0.5 ng of template DNA was amplified) were normalized in a 96-well plate at a volume of 50 µL at 0.5 ng/µL, or 25 ng of amplified DNA.

The TruSeq™ Forensic Amplicon library preparation process is similar to that of TruSeq ChIP, except that it uses PCR amplicons as starting material, rather than chromatin-bound DNA. The process starts with end repair, where the 5' ends of the amplicons were made blunt and phosphorylated during a 30-minute incubation at 30° C in an Applied Biosystems® GeneAmp® PCR System 9700 thermal cycler. Next, the samples were washed using AMPure XP beads and 80% ethanol. The blunt ends then were adenylated, which prevented them from ligating to each other during adapter ligation. Adenylation was performed in the thermal cycler using the following parameters: 37° C for 30 minutes, 70° C for 5 minutes, and a final hold at 4° C. Following adenylation, adapter ligation was performed, wherein TruSeq™ indexed adapters were bound to the adenylated 3' ends of the amplicons. Adapters were bound to each sample with a unique index sequence for multiplexed sequencing. Adapter ligation required a 10-minute incubation at 30° C, followed by washing using AMPure XP beads and 80% ethanol. For enrichment of adapter-bound amplicons, PCR was carried out using primers designed to amplify only those amplicons with adapters bound to them. The enrichment PCR parameters were: 98° C for 30 seconds, 18 cycles of 98° C for 10 seconds, 60° C for 30 seconds, and 72° C for 30 seconds, a final extension at 72° C for 5 minutes, and a final hold at 4° C. PCR products were washed with AMPure XP beads and 80% ethanol.

Following library preparation, the adapter-ligated amplicons were quantified using the Qubit® platform, according to the manufacturer's protocol. The samples were normalized to a concentration of 10 nM with 10 mM tris-HCl buffer at pH 8.5 with 0.1% Tween 20. Five µL of each sample were used to pool samples, for a total 10 nM sample pool of 120 µL.

MiSeq Sequencing and Data Analysis

Ten µL of the 10 nM sample pool were combined with 40 µL of 10 mM tris-HCl buffer at pH 8.5 with 0.1% Tween 20, for a resultant concentration of 2 nM. The concentration of the pooled sample was brought down to 12 pM using chilled HT1 buffer. Paired-end sequencing was performed on the MiSeq™ with a read length of 120 bases.

MiSeq Reporter was used to produce VCF files for each sample which identified each SNP detected during sequencing. Since MiSeq Reporter limits, by default, sequence coverage values for SNPs to 5,000X, a separate method of variant-calling was required to ascertain the coverage at each locus of interest so that conclusions could be drawn with regard to the depth of sequencing and heterozygote balance afforded by the TruSeq Forensic Amplicon library preparation method. To this end, BAM files produced by the MiSeq were subjected to variant-calling without downsampling using GATK.

XV. TruSeq™ Forensic Amplicon Results and Discussion

SNP genotypes were obtained for all 160 SNPs in 11 of the 12 samples analyzed. In sample 9, one SNP (rs10776839) was not called due to low coverage. Whole genome sequencing (WGS)-based SNP calls were obtained from the Complete Genomics FTP site (133) for concordance testing of these samples. The allele calls derived from the data produced by the TruSeq™ Forensic Amplicon library preparation method displayed high concordance (96.23% to 98.74%) across all 12 samples. Discordance between the WGS-derived SNP calls and the trial calls was observed at a total of 9 out of the 160 SNPs (rs1029047, rs1058083, rs10776839, rs10954737, rs12997453, rs182549, rs2399332, rs430046, and rs907100). These SNP loci appear to be discordance "hotspots", as all but one of the loci showed discordance in at least 4 of the samples tested (Table 19). The vast majority of the discordance (all but 3 of the total 53 discordant calls, across all samples) consisted of differences between heterozygous and homozygous allele SNP calls. There are a number of reasons why this discordance may have occurred. First, nucleotide variation within the primer binding site may have resulted in a failure to amplify one of the alleles at a given locus, which could explain a homozygous SNP call at a truly heterozygous locus. For most of the discordant results, the differences were consistent at the locus among the samples. For example, at rs1029047 all discordant samples by amplicon sequencing were A/T and by WGS were A. Conversely, at rs907100 all samples were G by amplicon sequencing and C/G by WGS. Second, it is possible that primer mismatching may result in chimeric products that confound the alignment process. Sequence analysis of discordant loci revealed a noteworthy portion of mismatched nucleotides surrounding the polymorphic position. Partial homology of primer sequences with other regions of the genome may have caused primers to partially anneal to regions other than intended and being extended before being denatured. If these fragments re-anneal to their proper primer binding sites and continue extension, sequence data would be generated that might pass the alignment software's stringency thresholds and yield a discordant call. Other explanations include factors such as multiplex inefficiency, low coverage leading to skewed SNP calls, and simple alignment errors. Alternatively, some errors may reside with the WGS-generated data. At this time the discrepancies cannot be resolved. Regardless, the high concordance supports that the library preparation method is an effective process.

Table 19. SNP discordance.

	1	2	3	4	5	6	7	8	9	10	11	12
rs1029047	A/T:A	A/T:A		A/T:A		A/T:A				A/T:A	A/T:A	
rs1058083					A/G:G		A/G:G	A/G:G		A/G:G		
rs10776839	G:G/T			G:G/T	G:G/T		G:G/T	G:G/T		G:G/T	G:G/T	G:T
rs10954737	T:C/T	T:C/T	T:C/T					T:C/T	T:C/T	T:C/T		
rs12997453		G:A/G										
rs182549	C/T:T	C/T:T	C/T:T	C/T:T					C/T:T	C/T:T		
rs2399332	G/T:G	G/T:G					G/T:G	G/T:G		G/T:G	G/T:G	
rs430046	C:C/T	C:C/T		C:C/T	C:C/T	C:T	C:C/T	C:C/T	C:C/T		C:C/T	C:T
rs907100						G:C/G	G:C/G	G:C/G	G:C/G	G:C/G		G:C/G

Discordance between the SNP calls generated in this study and those obtained through whole genome sequencing are shown. Discordance is shown in the following format "study call: WGS call".

Overall, good heterozygote balance was achieved with the multiplex PCR and TruSeq Forensic Amplicon library preparation method. On a per sample basis, between 91.9% and 100% of the heterozygous loci showed allelic balance ratios of 1:2 (50% balance, arbitrarily set) or better. Figure 11 shows the heterozygote allele balance for the SNP panel. In some cases, allelic imbalance was associated with low coverage, but other factors, such as those noted above, may explain imbalance in heterozygous loci with higher coverage values.

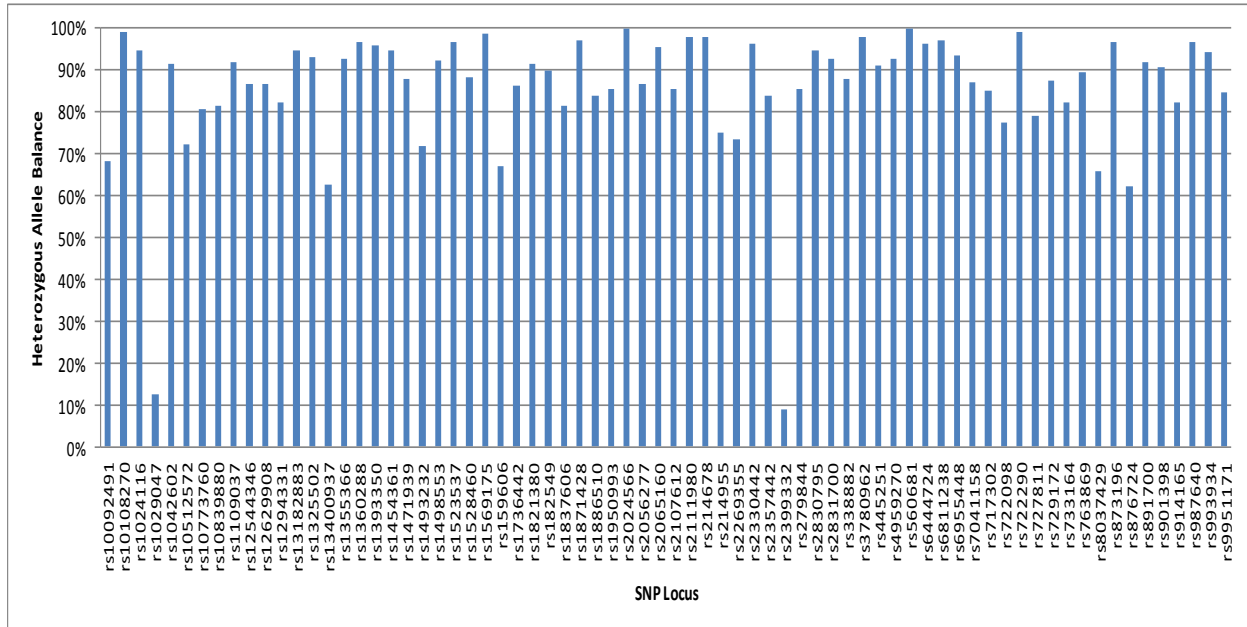


Figure 11. Heterozygote allele balance for a representative sample no. 2). Allele balance at heterozygote loci, expressed as a percentage, is shown for one sample. A value of 100% denotes a perfect 1:1 balance of alleles. In this sample, only 2 loci (rs1029047 and rs2399332) display an allele balance value of less than 50%.

The average sequencing coverage per locus across all 12 samples ranged from 142X to 46,908X, and coverage was relatively consistent among samples at each locus (Figure 12). The wide range of coverage is most likely due to differences in amplification efficiency of the multiplex PCR.

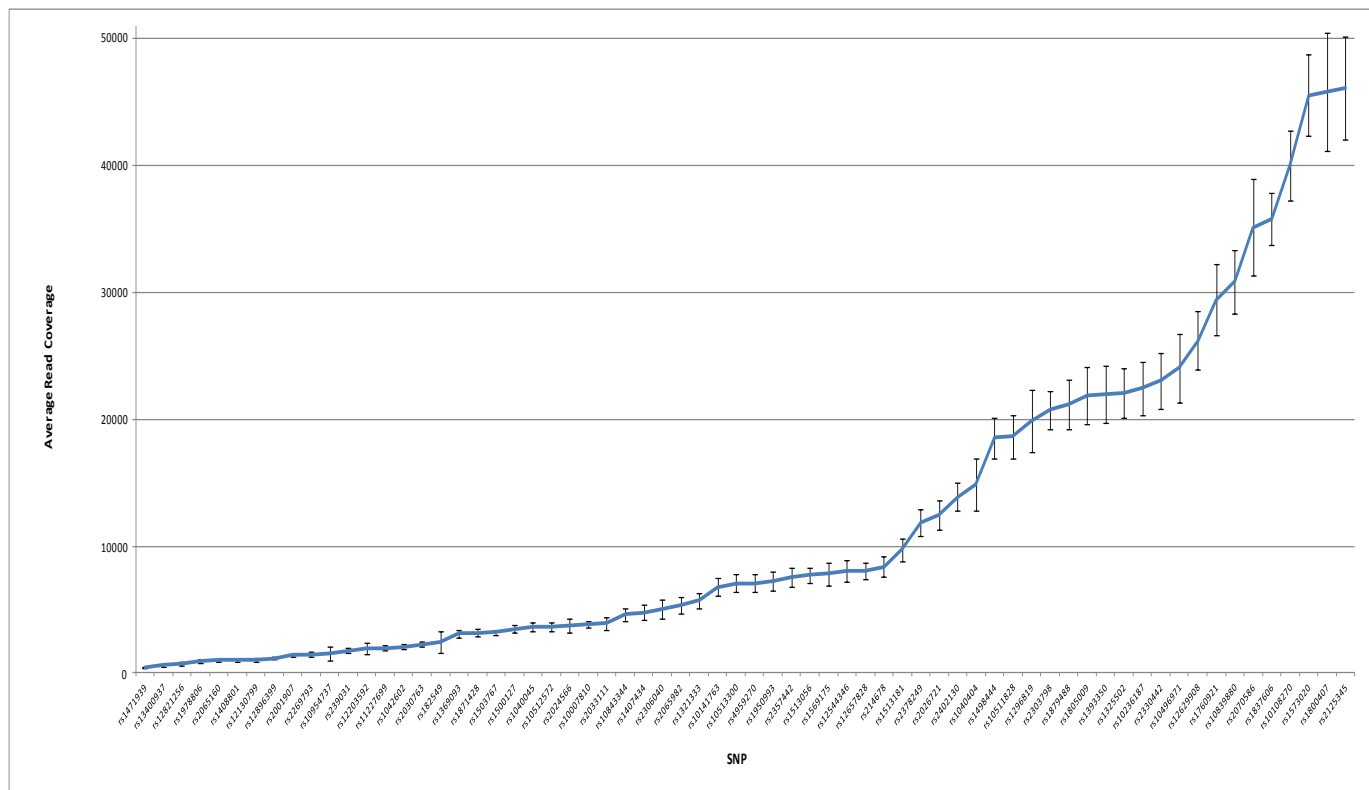


Figure 12. Average sequence coverage for AIM SNP loci. The average depth of coverage across all samples for each AIM SNP locus is shown. Bars represent the standard deviation.

Conclusions on Library Preparation Method

The results of this proof-of-concept study indicate that the TruSeq™ Forensic Amplicon library preparation protocol can be another effective method of preparing amplified nuclear DNA for MPS. This method appears to be less labor-intensive than alternative techniques. Unlike the TruSeq™ Custom Amplicon workflow, TruSeq™ Forensic Amplicon workflow does not require the use of custom-designed oligonucleotide probes for library preparation. Additionally, the TruSeq™ Forensic Amplicon library preparation method is highly sensitive, with a relatively low input DNA requirement (25 ng of amplified DNA, from a starting quantity of 0.5 ng were used in this study, as opposed to the recommended 50-500 ng of input DNA recommended for the TruSeq™ Enrichment protocol). In conjunction with a properly designed multiplex PCR, this preparation method is capable of producing sequencing results with relatively even allele balance at heterozygous loci. The results of this proof-of-concept preparation method suggested that this novel use of the original TruSeq™ ChIP protocol could support forensic marker typing by MPS.

XVI. Final Concluding Remarks

All goals of the project were met. Large multiplex systems were developed (and also obtained) and tested for typing reference samples. STRs 9autosomal, X chromosome and Y chromosome) and identity SNPs could be typed simultaneously. SNPs also were typed in their own multiplex.

Whole mtGenomes could be sequenced with relative ease. The data support that reliable results can be obtained. To facilitate analyses software was developed. STRait Razor (v1.0 and v2.0) for STR typing and mitoSAVE for haplotype alignment/nomenclature have been created and are freely available. The protocols described within the final report and published in the scientific literature should enable novel users to perform MPS in their respective laboratories.

XVII. References

1. Budowle, B., Eisenberg, A.J. 2007. Forensic Genetics. In: Emery and Rimoin's Principles and Practice of Medical Genetics, fifth edition, Vol. 1, Rimoin, D.L., Connor, J.M., Pyeritz, R.E., and Korf, B.R. (eds.), Elsevier, Philadelphia, pp.501-517.
2. Budowle, B., Planz, J.V., Campbell, R., Eisenberg, A.J. 2005. Molecular diagnostic applications in forensic science. In: Molecular Diagnostics, Patrinos, G. and Ansorge, W., (eds.), Elsevier, Amsterdam, pp. 267-280.
3. Budowle, B., van Daal, A. 2008. Forensically relevant SNP classes. *Biotechniques* 44:603-610.
4. Budowle, B., Moretti, T.R., Niezgoda, S.J., Brown, B.L. 1998. CODIS and PCR-based short tandem repeat loci: Law enforcement tools. In: Second European Symposium on Human Identification 1998, Promega Corporation, Madison, Wisconsin, pp. 73-88.
5. Martin, P.D., Schmitter, H., Schneider, P.M. 2001. A brief history of the formation of DNA databases in forensic science within Europe. *Forens. Sci. Int.* 119(2):225-231.
6. Budowle, B. 2010. Familial searching: extending the investigative lead potential of DNA typing. *Profiles in DNA* 13(2), 2010, Available at: www.promega.com/profiles/1302/1302_07.html.
7. Ge, J., Chakraborty, R., Eisenberg, A., Budowle, B. 2011. Comparisons of the familial DNA database searching policies. *J. Forens. Sci.* 56(6):1448-1456.
8. Ge, J., Eisenberg, A., Budowle, B. 2012. Developing criteria and data to determine best options for expanding the core CODIS loci. *BMC Investigative Genetics* 3:1.
9. Hares, D.R. 2011. Expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 6 (1):e52-54
10. Ge, J., Sun, H., Li, H., Liu, C., Yan, J., Budowle, B. 2014. Future directions of forensic DNA databases. *Croatian Med. J.* 55:163-166.
11. Sanger, F., Nicklen, S., Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74(12):5463-5467.

12. Wilson, M.R., Polanskey, D., Butler, J., DiZinno, J.A., Replogle, J., Budowle, B. 1995. Extraction, PCR amplification, and sequencing of mitochondrial DNA from human hair shafts. *BioTechniques* 18:662-669.
13. Wilson, M.R., DiZinno, J.A., Polanskey, D., Replogle, J., Budowle, B. 1995. Validation of mitochondrial DNA sequencing for forensic casework analysis. *Int. Journal Leg. Med.* 108:68-74.
14. Budowle, B. 2004. SNP typing strategies. *Forens. Sci. Int.* 146(suppl):S139-S142.
15. Budowle, B., Planz, J., Campbell, R., Eisenberg, A. 2004. SNPs and microarray technology in forensic genetics: development and application to mitochondrial DNA. *Forens. Sci. Rev.* 16:22-36.
16. Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W. et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475(7356):348-352.
17. Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermoud, J.J., Mayer, P., Kawashima, E. 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* 28(20):E87.
18. Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, H. et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18:630-634.
19. Brenner, S., Williams, S.R., Vermaas, E.H., Storck, T., Moon, K., et al. 2000. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. USA* 97:1665-1670.
20. Holt, K.E., Parkhill, J., Mazzoni, C.J., Roumagnac, P., Weill, F.X., et al. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat. Genet.* 40:987-993.
21. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.
22. Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H., Turner, D.J. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* 5:1005-1010.
23. Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., Sonstegard, T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5:247-252.

24. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872-876.
25. Knapp, M., Stiller, M., Meyer, M. 2012. Generating barcoded libraries for multiplex high throughput sequencing. *Methods Mol. Biol.* 840:155–170.
26. Gill, P., Ivanov, P.L., Kimpton, C., Piercy, R., Benson, N., et al. 1994. Identification of the remains of the Romanov family by DNA analysis. *Nat. Genet.* 6:130-135.
27. Holland, M.M., Parsons, T.J. 1999. Mitochondrial DNA sequence analysis-validation and use for forensic casework. *Forensic Sci Rev.* 11:21-50.
28. Palo, J.U., Hedman, M., Soderholm, N., Sajantila, A. 2007. Repatriation and identification of Finnish World War II soldiers. *Croat Med J.* 48:528-535.
29. Bannwarth, S., Procaccio, V., Lebre, A.S., Jardel, C., Chaussenot, A., et al. 2013. Prevalence of rare mitochondrial DNA mutations in mitochondrial disorders. *J Med Genet.* 50:704-714.
30. Nunnari, J., Suomalainen, A. 2012. Mitochondria: in sickness and in health. *Cell.* 148:1145-1159.
31. Wallace, D.C., Chalkia, D. 2013. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harbor Persp. Biol.* 5:a021220.
32. Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., et al. 2004. Ethiopian Mitochondrial DNA Heritage: Tracking Gene Flow Across and Around the Gate of Tears. *Amer. J. Hum. Genet.* 75:752-770.
33. Richards, M., Macaulay, V., Torroni, A., Bandelt, H.-J. 2002. In search of geographical patterns in European mitochondrial DNA. *Amer. J. Hum. Genet.* 71:1168-1174.
34. van Oven, M., Kayser, M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* 30:E386-E94.
35. Sajantila, A., Salem, A.-H., Savolainen, P., Bauer, K., Gierig, C., Pääbo, S. 1996. Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc. Natl. Acad. Sci.* 93:12035-12039.
36. Parson, W., Dür, A. 2007. EMPOP—A forensic mtDNA database. *Forensic Sci. Int. Genet.* 1:88-92.
37. Lee, H.Y., Song, I., Ha, E., Cho, S.-B., Yang, W.I., Shin, K.-J. 2008. mtDNAMAN: a Web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences. *BMC Bioinformatics* 9:483.

38. Ingman, M., Gyllensten, U. 2006. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Res.* 34:D749-D751.
39. Fan, L., Yao, Y.-G. 2013. An update to MitoTool: Using a new scoring system for faster mtDNA haplogroup determination. *Mitochondrion* 13:360-363.
40. Attimonelli, M., Accetturo, M., Santamaria, M., Lascaro, D., Scioscia, G., et al. 2005. HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research. *BMC Bioinformatics* 6:S4.
41. Kohl, J., Paulsen, I., Laubach, T., Radtke, A., von Haeseler, A. 2006. HvrBase++: a phylogenetic database for primate species. *Nucleic Acids Res.* 34:D700-D704.
42. Brandon, M.C., Lott, M.T., Nguyen, K.C., Spolim, S., Navathe, S.B., et al. 2005. MITOMAP: a human mitochondrial genome database—2004 update. *Nucleic Acids Res.* 33:D611-D613.
43. Levin, B.C., Cheng, H., Reeder, D.J. 1999. A Human Mitochondrial DNA Standard Reference Material for Quality Control in Forensic Identification, Medical Diagnosis, and Mutation Detection. *Genomics* 55:135-146.
44. Coble, M., Just, R., O’Callaghan, J., Letmanyi, I., Peterson, C., et al. 2004. Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. *Int. J. Leg. Med.* 118:137-146.
45. Brandstätter, A., Parsons, T.J., Parson, W. 2003. Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups. *Int. J. Leg. Med.* 117:291-298.
46. Parsons, T.J., Coble, M.D. 2001. Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. *Croat. Med. J.* 42:304-309.
47. Bandelt, H.-J., Lahermo, P., Richards, M., Macaulay, V. 2001. Detecting errors in mtDNA data by phylogenetic analysis. *Int. J. Leg. Med.* 115:64-69.
48. Bandelt, H.-J., van Oven, M., Salas, A. 2012. Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics. *Int. J. Leg. Med.* 126:901-916.
49. Röck, A.W., Dür, A., van Oven, M., Parson, W. 2013. Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA). *Forensic Sci. Int. Genet.* 7(6):601-609.

50. Kloss-Brandstätter, A., Pacher, D., Schönherr, S., Weissensteiner, H., Binna, R., et al. 2011. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* 32:25-32.
51. Vianello, D., Sevini, F., Castellani, G., Laura, L., Capri, M., Franceschi, C. 2013. HAPLOFIND: A New Method for High-Throughput mtDNA Haplogroup Assignment. *Hum. Mutat.* 34 (9):1189-1194.
52. Gunnarsdóttir, E.D., Li, M., Bauchet, M., Finstermeier, K., Stoneking, M. 2011. High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res.* 21:1-11.
53. Syed, F., Gruenwald, H., Caruccio, N. 2009. Next-generation sequencing library preparation: Simultaneous fragmentation and tagging using in vitro transposition. *Nat. Methods*. Available: <http://www.nature.com/nmeth/journal/v6/n11/full/nmeth.f.272.html>.
54. Illumina. 2013. Nextera® library validation and cluster density optimization.
55. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297-1303.
56. Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowers, R.N., Turnbull, D.M., Howell, N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* 23:147.
57. King, J.L., Sajantila, A., Budowle, B. 2014. mitoSAVE: mitochondrial sequencing analysis of variants in excel. *Forens. Sci. Genet. Int.* 12:122-125.
58. Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*.
59. Stoneking, M., Hedgecock, D., Higuchi, R.G., Vigilant, L., Erlich, H.A. 1991. Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *Amer. J. Hum. Genet.* 48:370-382.
60. Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 123:585-595.
61. Li, H., Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics* 25:1754–1760.

62. Kloss-Brandstätter, A., Pacher, D., Schönherr, S., Weissensteiner, H., Binna, R., Specht, G., Kronenberg, F. 2011. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* 32 (1):25-32.
63. Scientific Working Group on DNA Analysis Methods (SWGDM). 2013. Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories, http://swgdam.org/SWGDAM%20mtDNA_Interpretation_Guidelines_AP-PROVED_073013.pdf.
64. Wilson, M.R., Allard, M.W., Monson, K., Miller, K.W.P., Budowle B. 2002. Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region, *Forensic Sci. Int.* 129:35–42.
65. Budowle, B., Polansky, D., Fisher, C.L., Den Hartog B.K., Kepler, R.B., Elling, J.W. 2010. Automated alignment and nomenclature for consistent treatment of polymorphisms in the human mitochondrial DNA control region, *J. Forensic Sci.* 55:1190–1195.
66. Tully, G., Bar, W., Brinkmann, B., Carracedo, A., Gill, P., et al. 2001. Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles, *Forensic Sci. Int.* 124:83–91.
67. Bandelt, H.-J., Parson, W. 2008. Consistent treatment of length variants in the human mtDNA control region: a reappraisal, *Int. J. Leg. Med.* 122:11–21.
68. Parson, W., Brandstatter, A., Alonso, A., Brandt, N., Brinkmann, B., et al. 2004. The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation, results and perspectives, *Forensic Sci. Int.* 139:215–226.
69. King, J.L., LaRue, B.L., Novroski, N., Stoljarova, M., Seo, S.B., Zeng, X., Warshauer, D., Davis, C., Parson, W., Sajantila, A., Budowle, B. 2014. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. *Forens. Sci. Int. Genet.* 12:128-135.
70. Ion PGM™ Sequencing 200 Kit v2 [<http://ioncommunity.lifetechnologies.com/docs/DOC-6775>]
71. Mikkelsen, M., Hansen, R.F., Hansen, A.J., Morling, N. 2014, Massively parallel pyrosequencing 454 methodology of the mitochondrial genome in forensic genetics. *Forensic Sci Int Gent.* (in press).
72. McElhoe, J.A., Holland, M.M., Makova, K.D., Su, M.S.-W. Paul, I.M., Baker, C.H., Faith, S.A., Young, B. 2014. Development and assessment of an optimized next-generation DNA

sequencing approach for the mtgenome using the Illumina MiSeq. *Forens. Sci. Int. Genet.* (in press).

73. Parson, W., Strobl, C., Huber, G., Zimmermann, B., Gomes, S.M., et al. 2013. Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). *Forens. Sci. Int. Genet.* 7:543-549.

74. McElroy, K., Thomas, T., Luciani, F. 2014. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatics solutions. *Microb. Inform. Exp.* 4:1.

75. Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., et al. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* 39:e90.

76. Meacham, F., Boffelli, D., Dhahbi, J., Martin, D.I., Singer, M., Pachter, L. 2011. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* 12:451.

77. Allhoff, M., Schönhuth, A., Martin, M., Costa, I.G., Rahmann, S., Marschall, T. 2013. Discovering motifs that induce sequencing errors. *BMC Bioinformatics* 14(Suppl 5):S1.

78. Part II: Overview of Torrent Sequencing and Alignment (v4.x)
[<http://ioncommunity.lifetechnologies.com/docs/DOC-3266>]

79. Lorenz, J.G., Smith, D.G. 1994. Distribution of the 9-bp mitochondrial DNA region V deletion among North American Indians, *Hum. Biol.* 66:777–788.

80. Watkins, W., Bamshad, M., Dixon, M., Bhaskara Rao, B., Naidu, J., et al. 1999. Multiple origins of the mtDNA 9-bp deletion in populations of South India, *Amer. J. Phys. Anthropol.* 109:147–158.

81. Dixon, L.A., Murray, C.M., Archer, E.J., Dobbins, A.E., Koumi, P., Gill, P. 2005. Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. *Forensic Sci. Int.* 154:62-77.

82. Kidd, J.R., Friedlaender, F.R., Speed, W.C., Pakstis, A.J., De La Vega, F.M., Kidd, K.K. 2011. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *BMC Investigative Genetics* 2(1):1.

83. Pakstis, A.J., Speed, W.C., Fang, R., Hyland, F.C.L., Furtado, M.R., Kidd, J.R., Kidd, K.K. 2010. SNPs for a universal individual identification panel. *Human Genetics* 127:315-324.

84. Pakstis, A.J., Speed, W.C., Kidd, J.R., Kidd, K.K. 2007. Candidate SNPs for a Universal Individual Identification Panel. *Human Genetics* 121:305-317.

85. Sanchez, J.J., Phillips, C., Borsting, C., Balogh, K., Bogus, M., et al. 2006. A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis*. 27:1713-1724.
86. Vallone, P.M., Decker, A.E., Butler, J.M. 2005. Allele frequencies for 70 autosomal SNP loci with U.S. Caucasian, African-American, and Hispanic samples. *Forensic Sci. Int.* 149:279-286.
87. Kosoy, R., Nassir, R., Tian, C., White, P.A., Butler, L.M., et al. 2010. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutation* 30:69-78.
88. Phillips, C., Salas, A., Sánchez, J.J., Fondevila, M., Gómez-Tato, A., et al. 2007. The SNPforID Consortium Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci. Int. Genet.* 1:273-280.
89. Becker, D., Rodig, H., Augustin, C., Edelmann, J., Götz, F., Hering, S., Szibor, R., Brabetz, W. 2008. Population genetic evaluation of eight X-chromosomal short tandem repeat loci using Mentype Argus X-8 PCR amplification kit. *Forensic Sci. Int. Genet.* 2(1):69-74.
90. Diegoli, T.M., Coble, M.D. 2011. Development and characterization of two mini-X chromosomal short tandem repeat multiplexes. *Forensic Sci. Int. Genet.* 5(5):415-421.
91. Gomes, I., Prinz, M., Pereira, R., Meyers, C., Mikulasovich, R.S., Amorim, A., Carracedo, A., Gusmão, L. 2007. Genetic analysis of three US population groups using an X-chromosomal STR decaplex. *Int. J. Leg. Med.* 121(3):198-203.
92. Nothnagel, M., Szibor, R., Vollrath, O., Augustin, C., Edelmann, J., et al. 2012. Collaborative genetic mapping of 12 forensic short tandem repeat (STR) loci on the human X chromosome. *Forensic Sci. Int. Genet.* 6(6):778-784.
93. Bornman, D.M., Hester, M.E., Schuetter, J.M., Kasoji, M.D., et al. 2012. Short-read, high-throughput sequencing technology for STR genotyping. *Biotechniques Biotech Rapid Dispatches* 0:1-6.
94. Warshauer, D.H., Lin, D., Hari, K., Jain, R., Davis, C., LaRue, B., King, J., Budowle, B. 2013. STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forens. Sci. Int. Genet.* 7:409-417.
95. Gymrek, M., Golan, D., Rosset, S., Erlich, Y. 2012. lobSTR: a short tandem repeat profiler for personal genomes, *Genome Res.* 22:1154-1162.
96. Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

97. AGREP: <http://laurikari.net/tre/download/>
98. PPSS: <http://code.google.com/p/ppss/>
99. Fordyce, S.L., Ávila-Arcos, M.C., Rockenbauer, E., Børsting, C., Frank-Hansen, R., et al. 2011. High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform. *Biotechniques* 51:127-133.
100. Van Neste, C., Vandewoestyne, M., Van Criekinge, W., Deforce, D., Van Nieuwerburgh, F. 2014. My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing. *Forensic Sci. Int. Genet.* 9:1-8.
101. CASAVA v.1.8.2: http://support.illumina.com/downloads/casava_182.ilmn
102. MiSeq Reporter: http://support.illumina.com/sequencing/sequencing_software/miseq_reporter/downloads.ilmn.
103. Planz, J.V., Sannes-Lowery, K.A., Duncan, D.D., Manalili, S., Budowle, B., Chakraborty, R., Hofstadler, S.A., Hall, T.A. 2012. Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry. *Forens. Sci. Int. Genet.* 6(5):594-606.
104. Agilent Technologies® HaloPlex™ Specifications: <http://www.genomics.agilent.com/GenericB.aspx?pagetype=Custom&subpagetype=Custom&pageid=3081>.
105. Illumina TruSeq Specifications: http://www.illumina.com/Documents/%5Cproducts%5Cdatasheets%5Cdatasheet_truseq_custom_enrichment_kit.pdf.
106. STRBase: http://www.cstl.nist.gov/strbase/str_fact.htm.
107. ChrX-STR.org 2.0: <http://www.chrx-str.org/>.
108. NCBI: <http://www.ncbi.nlm.nih.gov/>.
109. Sorenson Molecular Genealogy Foundation Y Marker Details: http://www.smgf.org/ychromosome/marker_details.jsp.
110. Warshauer, D.H., King, J.L., and Budowle, B. 2015. STRait Razor v2.0: the improved STR allele identification tool – razor. *Forensic Sci. Int. Genet.* 14:182–186.
111. Tomas, C., Axler-DiPerte, G., Budimlija, Z.M., Børsting, C., Coble, M.D., et al. 2011. Autosomal SNP typing of forensic samples with the GenPlex™ HID System: results of a collaborative study. *Forensic Sci. Int. Genet.* 5:369-375.

112. Børsting, C., Sanchez M J.J., Morling, N. 2005. SNP typing on the NanoChip electronic microarray. *Methods Mol. Biol.* 297:155-168.
113. Mengel-Jørgensen, J., Sanchez, J.J., Børsting, C., Kirpekar, F., Morling, N. 2005. Typing of multiple single-nucleotide polymorphisms using ribonuclease cleavage of DNA/RNA chimeric single-base extension primers and detection by MALDI-TOF mass spectrometry. *Anal. Chem.* 77:5229-5235.
114. Budowle, B., Planz, J., Campbell, R., Eisenberg, A. 2004. SNPs and microarray technology in forensic genetics: development and application to mitochondrial DNA. *Forens. Sci. Rev.* 16:22-36.
115. Life Technologies. 2012. Ion AmpliSeq™ Library Preparation User Guide. Life Technologies, Foster City (CA).
116. Life Technologies. 2011. Ion Library Quantitation Kit User Guide. Life Technologies, Foster City (CA).
117. Life Technologies. 2011. Ion OneTouch™ 200 Template Kit v2 DL. Life Technologies, Foster City (CA).
118. Life Technologies. 2011. Ion PGM™ 200 Sequencing Kit. Life Technologies, Foster City (CA).
119. Seo, S.B., King, J., Warshauer, D., Davis, C., Ge, J., Budowle, B. 2013. Single nucleotide polymorphism typing with massively parallel sequencing for human identification. *Int. J. Leg. Med.* 127(6):1079-1086.
120. Voelkerding, K.V., Dames, S.A., Durtschi, J.D. 2009. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* 55:641-658.
121. Marshfield Clinic;
<http://research.marshfieldclinic.org/genetics/GeneticResearch/sets/Set10PrimerSequences.htm>);
122. NCBI; <http://www.ncbi.nlm.nih.gov/nuccore/G08468>
123. Mentype Chimera;
(http://www.biotype.de/fileadmin/user/MANUALS/190213_Manual_MentypeChimera_CE.pdf),

124. Thiede, C., Bornhauser, M., Ehninger, G. 2004. Evaluation of STR informativity for chimerism testing – comparative analysis of 27 STR systems in 203 matched related donor recipient pairs. *Leukemia* 18: 248–254.
125. Qiagen Investigator HDplex Kit;
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0CC8QFjAC&url=http%3A%2F%2Fwww.qiagen.com%2Fresources%2Fdownload.aspx%3Fid%3D7d1661bd-a47b-4b19-a882-357a61b48c64%26lang%3Den&ei=f9tjVI7GJoiYyASKI4GQDw&usg=AFQjCNEXCa2eHD1yg5bzJl0AbkyB9k_4gg&sig2=uJ7RG8yki7pvbp_w4aVFkw&bvm=bv.79189006,d.aWw.
- 126 Hill, C.R., Kline, M.C., Coble, M.D., Butler, J.M. 2008. Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples. *J. Forens. Sci.* 53(1):73-80.
127. Hill, C.R, Butler, J.M., Vallone, P.M. 2009. A 26plex autosomal STR assay to aid human identity testing. *J Forens. Sci.*54(5):1008-1015.
128. STRbase; http://www.cstl.nist.gov/biotech/strbase/miniSTR/miniSTR_NC_loci_types.htm
129. Becker D, Bender K, Edelmann J, Götz F, Henke L, Hering S, Hohoff C, Hoppe K, Klintschar M, Muche M, Rolf B, Szibor R, Weirich V, Jung M, Brabetz W. 2007. New alleles and mutational events at 14 STR loci from different German populations. *Forens. Sci. Int. Genet.* 1(3-4):232-237.
130. Carpenter, M. L., Buenrostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., Sikora, M., et al. 2013. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Amer. J. Hum. Genet.* 93(5):852-864.
131. Templeton, J. E., Brotherton, P. M., Llamas, B., Soubrier, J., Haak, W., Cooper, A., and Austin, J. J. 2013. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *Investigative Genetics*, 4(1):1-13.
132. Park, P.J. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10:669-680.
133. Complete Genomics FTP site: <ftp://ftp2.completegenomics.com>.

XVIII. Dissemination of Research Findings

Presentations

1. Budowle, B.: Genomics and technologies for the next generation forensic laboratory. Joint Conference of HGM 2013 and 21st International Congress of Genetics, Singapore, 2013.
2. Budowle, B. and Kroupa, K.: Update on the next generation of DNA typing technologies. American Society of Crime Laboratory Directors 40th Anniversary Meeting, Durham, North Carolina, 2013.
3. Budowle, B. and Eisenberg, A.: A view of the new DNA technologies and their potential impact on the next generation forensic laboratory. American Society of Crime Laboratory Directors 40th Anniversary Meeting, Durham, North Carolina, 2013.
4. Budowle, B.: The genetic technology revolution. 3rd Annual Life Sciences Symposium – Improving the Quality of Life Through DNA Technology, Ft. Worth, TX, 2013.
5. Budowle, B.: Comprehensive sequencing to address forensic needs. Technology Transition Workshop: a DNA Revolution – next generation technologies, Forensic Technology Center of excellence, National Institute of Justice, Ft. Worth, TX, 2013.
6. Budowle, B.: Genomics and technologies for the next generation forensic laboratory. Joint Conference of HGM 2013 and 21st International Congress of Genetics, Singapore, 2013.
7. Budowle, B., Warshauer, D.H., Seo, S.B., King, J.L., Davis, C., and LaRue, B.: Next generation sequencing provides comprehensive multiplex capabilities, 25th Congress of the International society of Forensic Genetics, Melbourne, Australia, 2013.
8. Zeng, X., Seo, S.B., LaRue, B., King, J., and Budowle, B.: Whole mitochondrial genome typing on Ion Torrent™ PGM™ platform, 24th International Symposium on Human Identification, Atlanta, GA, 2013.
9. Seo, S.B., King, J., Warshauer, D., Ge, J., and Budowle, B.: Large panels of SNPs for human identity typing are feasible with current generation sequencing (CGS) technology, 24th International Symposium on Human Identification, Atlanta, GA, 2013.
10. Warshauer, D.H., Lin, D., Hari, K., Jain, R., Davis, C., LaRue, B., King, J.L., and Budowle, B.: STRait Razor: a bioinformatic tool for length-based STR allele-calling in massively parallel sequencing data, 24th International Symposium on Human Identification, Atlanta, GA, 2013.
11. Budowle, B.: Massively parallel sequencing and forensic identity testing, Fifth Annual Prescription for Criminal Justice Forensics, The ABA Criminal Justice Section and the Louis Stein Center for Law & Ethics, Fordham University, New York, New York, 2014.

12. Budowle, B.: Technologies of the future have arrived and communicating with the legal community, 2014 International Symposium on Forensic DNA in Law, Seoul, Korea, 2014.
13. Budowle, B.: Principles and chemistries of next generation sequencing technologies, American Academy of Forensic Sciences, Seattle, WA, 2014.
14. Budowle, B.: Massively parallel sequencing and forensic identity testing, Fifth Annual Prescription for Criminal Justice Forensics, The ABA Criminal Justice Section and the Louis Stein Center for Law & Ethics, Fordham University, New York, New York, 2014.
15. Warshauer, D.H., King, J.L., and Budowle, B.: STRait Razor v2.0: the improved STR Allele Identification Tool – Razor, 25th International Symposium on Human Identification, Phoenix, AZ, 2014.
16. King, J.L., LaRue, B.L., Novroski, N.M., Stoljarova, M., Seo, S.B., Zeng, X., Warshauer, D.H., Davis, C.P., Parson, W., Sajantila, A., and Budowle, B.: The use of Massively Parallel Sequencing (MPS) to accurately and rapidly sequence the mtGenome of 283 individuals from 3 North American populations, 25th International Symposium on Human Identification, Phoenix, AZ, 2014.

Publications

1. Warshauer, D.H.; Lin, D., Hari, K., Jain, R., Davis, C., LaRue, B., King, J., and Budowle, B.: STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forens. Sci. Int. Genet.* 7:409–417, 2013.
2. Seo, S.B., King, J., Warshauer, D., Davis, C., Ge, J., and Budowle, B.: Single nucleotide polymorphism typing with massively parallel sequencing for human identification. *Int. J. Leg. Med.* 127(6):1079-1086, 2014.
3. Budowle, B., Warshauer, D.H., Seo, S.B., King, J.L., Davis, C., and LaRue, B.: Massively parallel sequencing provides comprehensive multiplex capabilities. *Forensic Sci. Int.: Genetics Supplement Series* 4:e334–e335, 2013.
4. King, J.L., Sajantila, A., and Budowle, B.: mitoSAVE: mitochondrial sequencing analysis of variants in excel. *Forens. Sci. Genet. Int.* 12:122-125, 2014.
5. King, J.L., LaRue, B.L., Novroski, N., Stoljarova, M., Seo, S.B., Zeng, X., Warshauer, D., Davis, C., Parson, W., Sajantila, A., and Budowle, B.: High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. *Forens. Sci. Int. Genet.* 12:128-135, 2014.

6. Warshauer, D.H., Davis, C.P., Holt, C., King, J.L., and Budowle, B.: Massively parallel sequencing of forensically-relevant single nucleotide polymorphisms using TruSeq™ Forensic Amplicon. *Int. J. Leg. Med.* (in press).
7. Warshauer, D.H., King, J.L., and Budowle, B.: STRait Razor v2.0: the improved STR allele identification tool – razor. *Forens. Sci. Int. Genet.* 14:182–186, 2015.
8. Seo, S.B., Zeng, X., King, J.L., Larue, B.L., Assidi, M., Al-Qahtani, M.H., Sajantila, A., and Budowle, B.: Underlying data for sequencing the mitochondrial genome with the massively parallel sequencing platform Ion Torrent™ PGM™. *BMC Genomics* (in press).

XIX. Participating Scientists and Collaborations

UNTHSC Scientists

Principal Investigator, Bruce Budowle, Ph.D.
Research Associate, Bobby LaRue, Ph.D.
Laboratory Manager, Jonathan King, M.Sc.
Graduate Student Assistant, David Warshauer, M.Sc.
Graduate Student Xiangpei Zeng, M.D.
Visiting Scientist, Seung Bum Seo, Ph.D.
Student Intern, Monika Stoljarova, B.S.

Collaborators

cBio, Inc. (Fremont, CA) Scientists

David Lin
Ravi Jain
Kumar Hari

Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria

Walther Parson