OPTIMIZATION OF MARKER SETS AND TOOLS FOR PHENOTYPE, ANCESTRY, AND IDENTITY USING GENETICS AND PROTEOMICS

by

Bailey Wills

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Forensic and Investigative Science

Indianapolis, Indiana

August 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Susan Walsh, Chair
Department of Biology
Dr. Christine Picard
Department of Biology
Dr. Dave Skalnik
Department of Biology

Approved by:

Dr. John Goodpastor
Head of the Graduate Program

Dedicated to my parents. Without your constant love and support, none of this would have been possible

ACKNOWLEDGMENTS

I would like to acknowledge and thank everyone who has made an impact on my life and helped me achieve all of my goals throughout my academic career. First, I would like to begin by thanking Dr. Susan Walsh, my PI and mentor. I am beyond grateful for the opportunities I have had thanks to you. When you allowed me to join the lab way back in undergrad, I didn't expect to be earning my Master's degree three years later. Your love for science sparked a passion in me that has allowed me to grow and evolve as a student. You have helped shape who I am as a scientist and have taught me to always strive to be better. I feel extremely grateful to have worked under you for three years and will never forget this experience or the lessons you taught me.

Next, I would like to thank Dr. Christine Picard and Dr. Dave Skalnik for being a valuable part of my advisory committee. I truly value the input and advice you have provided to my thesis. You are both amazing scientists and I feel privileged to have had you sit on my committee.

To Krystal Breslin, my personal therapist, mentor, and friend- I have no idea what I would have done without you and all of our late-night phone calls. Thank you for always being there to listen, help, and push me when I needed it most. You truly have been one of the biggest support systems for me throughout this entire process and I would have never made it to where I am today without you.

I would also like to acknowledge and thank each and every member of the Walsh Lab: Stephanie, Charanya, Noah, Ryan, Mirna, Racquel, Case, Clare, Paige, Lina, Lydia, Emma, Sarah, Morgan,Wesli, and Katherine. You have all made an impact on my life and I'm thankful for all of the amazing relationships I have made. To Ryan and Noah, I cannot thank you two enough for taking me under your amazing bioinformatician wings and teaching me the ropes of computers. Your patience and dedication to helping me learn was truly remarkable and I'm certain I could not

have done this without you two. To Stephanie and Racquel, thank you for always putting a smile on my face and reminding me to roll with the punches. Without your daily encouragement, this would not have been possible. Your friendships will always be precious to me.

Finally, to my family- Mom and Dad, I love you both so much! Your constant support and encouragement has helped me every step of the way. There are not enough words to thank you for all of sacrifices you have made for me to achieve my goals. I hope I have made you proud. To my sister, Blaire- thank you for always allowing me to complain about my day and reminding me to look on the bright side of things. You're my best friend and I'm so thankful to have had you by my side through this process. Finally, to my brother, Owen- thank you for always being a source of comedic relief in my life. You never fail to make me smile. I love you all and know I could not have done this without your love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Wills, Bailey, M. MS
Institution: Purdue University
Degree Received: August 2019
Title: Optimization of Marker Sets and Tools for Phenotype, Ancestry, and Identity Using Genetics and Proteomics
Committee Chair: Susan Walsh

In the forensic science community, there is a vast need for tools to help assist investigations when standard DNA profiling methods are uninformative. Methods such as Forensic DNA Phenotyping (FDP) and proteomics aims to help this problem and provide aid in investigations when other methods have been exhausted. FDP is useful by providing physical appearance information, while proteomics allows for the examination of difficult samples, such as hair, to infer human identity and ancestry. To create a "biological eye witness" or develop informative probability of identity match statistics through proteomically inferred genetic profiles, it is necessary to constantly strive to improve these methods.

Currently, two developmentally validated FDP prediction assays, 'HIrisPlex' and 'HIrisplex-S', are used on the capillary electrophoresis to develop a phenotypic prediction for eye, hair, and skin color based on 41 variants. Although highly useful, these assays are limited in their ability when used on the CE due to a 25 variant per assay cap. To overcome these limitations and expand the capacities of FDP, we successfully designed and validated a massive parallel sequencing (MPS) assay for use on both the ThermoFisher Scientific Ion Torrent and Illumina MiSeq systems that incorporates all HIrisPlex-S variants into one sensitive assay. With the migration of this assay to an MPS platform, we were able to create a semi-automated pipeline to extract SNP-specific sequencing data that can then be easily uploaded to the freely accessible online phenotypic prediction tool (found at https://hirisplex.erasmusmc.nl) and a mixture

deconvolution tool with built-in read count thresholds. Based on sequencing reads counts, this tool can be used to assist in the separation of difficult two-person mixture samples and outline the confidence in each genotype call.

In addition to FDP, proteomic methods, specifically in hair protein analysis, opens doors and possibilities for forensic investigations when standard DNA profiling methods come up short. Here, we analyzed 233 genetically variant peptides (GVPs) within hair-associated proteins and genes for 66 individuals. We assessed the proteomic methods ability to accurately infer and detect genotypes at each of the 233 SNPs and generated statistics for the probability of identity (PID). Of these markers, 32 passed all quality control and population genetics criteria and displayed an average PID of 3.58×10^{-4}. A population genetics assessment was also conducted to identify any SNP that could be used to infer ancestry and/or identity. Providing this information is valuable for the future use of this set of markers for human identification in forensic science settings.

CHAPTER 1. INTRODUCTION

1.1 Purpose and Objectives

The current gold standard of forensic DNA profiling is centered around short tandem repeats (STRs), however when this method fails to identify a possible contributor at a crime scene through DNA reference profile comparison or database search, alternative methods must be used to further the investigation. Forensic DNA Phenotyping (FDP), or the prediction of externally visible characteristics, extends current laboratory analyses and works to help identify an unknown perpetrator from trace amounts of DNA [1]. FDP is useful because of the technology's ability to strengthen or disprove eye witness statements narrow down the suspect list through the ranking of the most probable appearance characteristics [2-5], and assign pigmentation to skeletal remains [6]. Another alternative method that can be used to further an investigation when standard STR profiling is not sufficient is through reverse proteomic methods of hair proteins. Proteomics brings the ability to analyze sample types that may not be suitable for standard forensic analyses (ie.hair shaft) [7] and shows potential for the generation of ancestry inference and probability of identity match statistics.

Phenotypic prediction tools such as IrisPlex [2, 8], HIrisplex [3, 9], and HIrislex-S [4, 5] are beneficial in forensic casework, but currently these assays are limited in their scope and potential as they are run on capillary electrophoresis (CE) systems. Typically, CE systems are capped at approximately 25 DNA variants due to size and ability to adequately space fragments, in addition to the SNaPshot (ThermoFisher Scientific) chemistry's reagent limitations. Due to the complex genetic nature of physical appearance traits, with each new trait would come a limitation of number of variants per assay. To advance the field of FDP, it is therefore vital to move to a platform that has the ability to analyze more DNA variants in a singular assay instead of generating additional

multiplex assays. Doing this on the CE would increase time, cost, and personnel needed to run the assays in order to expand phenotypic predictions. This limits the application of Forensic DNA Phenotyping and the progression of this intelligence approach with current technologies. Based upon this assessment, it is evident that a transition from CE to Massive Parallel Sequencing (MPS) technologies on benchtop sequencers (Illumina MiSeq, or ThermoFisher Scientific Ion Torrent) is the next logical and advantageous step for the future of Forensic DNA Phenotyping.

Massive Parallel Sequencing (MPS), also referred to as Next-Generation Sequencing (NGS), follows some of the same basic principles of Capillary Electrophoresis (CE) by generation of fragment information. However, MPS can overcome common limitations set forth by capillary electrophoresis by increasing the multiplex capacity, requiring less input DNA, and generates output data that is far more informative [10]. Instruments like Illumina's MiSeq and ThermoFisher's Ion Torrent allow for numerous fragments of the genome to be sequenced simultaneously, resulting in a significantly lower cost than the previous methods of individually sequencing genomes [11] or Sangers singular fragments, while generating more data. The ability of this technique to produce results from hundreds to thousands of variants in a single run is an ideal tool for forensic researchers as it allows the capacity to expand the number of variants required for further developing FDP (in addition to ancestry estimation) without variant and trait limitations, to assist law enforcement investigations. In essence, the application of MPS technology to FDP assay development will revolutionize the DNA intelligence field for "biological witness" generation.

In forensic cases with samples containing low quantity, degraded, or non-nuclear DNA, common STR-typing and FDP methods may not be useful in producing genetic information that will aid in an investigation. A common type of sample found at crime scenes is hair due to its

ability to remain intact under a wide range of environmental conditions and situations. However, if a hair does not contain a root, and therefore lacks nuclear DNA, other methods of examination will be necessary for the identification of the individual who left the sample. Current forensic methods for the analysis of hair shaft samples are limited to the use of a microscopic examination or mitochondrial DNA (mtDNA) analysis, each with their own unique set of limitations [12, 13]. While the microscopic examination of hair provides vital information about the ancestral characteristics, structure, pigment, and size, it is not objective enough to match an individual to the sample with a high degree of certainty, thus creating a limitation on the evidentiary value of the examination in a court of law [14]. The most prevalent approach to the examination of a hair shaft is through mtDNA sequencing [15]. While this method does provide biogeographic information and limited identification through familial analysis, it requires careful analyst handling and is susceptible to environmental factors [13]. Mitochondrial DNA sequencing is also less discriminating than STR-typing, due to the nature of inheritance through the maternal lineage only. All maternally related individuals will inherit the same mitochondrial genome therefore limiting the capacity for identification. Due to these limitations in analysis methods, diving into other sources of forensic applications, such as proteomics, might be useful for biogeographical knowledge as well as human identification.

Due to its unique characteristics, proteomics can be used to differentiate protein-containing samples based on amino acid sequence changes [16]. Unlike DNA, which is the same in every cell of the organism, the proteome of each organ and/or tissue type is individual to the abundance and identity of its proteins [12]. Thus, the ability to identify protein sequence information can be used to distinguish genetically distinct individuals. In hair proteomics, technologies like mass spectrometry-based shotgun proteomic sequencing can be used to detect single amino acid

polymorphisms within a peptide. These single amino acid polymorphisms can then be used to infer the genotypes of non-synonymous single nucleotide polymorphism (nsSNP) alleles and provide a genetic profile for a hair sample of a particular individual [16].

The first goal of this research was to design, validate, and assess the performance of the HIrisPlex-S system on the Illumina MiSeq, create tools to aid in interpretation of raw sequence data, and propose deconvolution steps to separate two-person mixture profiles. This manuscript has been submitted to Forensic Science International: Genetics, and can be seen in Chapter 2 of this thesis. The final goal of this research was to assess proteomic data for 99 individuals at 233 different SNPs. Genotypes were inferred from the proteomic analysis data at each SNP, genotyping accuracy calculations were computed, and these SNPs were genetically evaluated with population data to generate a list of identity and ancestry informative markers.

1.2 Biology of Pigment Formation

To better understand the workings of pigmentation models, it is first necessary to provide some background information on the production of pigmentation through the melanin biochemical pathway. Melanin is a light absorbing biopolymer and is the source of human pigmentation. It is found in the melanocytes within the ocular (eye), follicular (hair), and epidermal (skin), causing the visible color in each [17]. The production of this pigmentation follows an intricate pathway with the outcome creating one of two types of melanin: Eumelanin or Pheomelanin. The production of eumelanin is responsible for darker colors, such as brownish and black pigment. Pheomelanin is responsible for lighter colors, like red-yellow pigment. The complex pigmentation pathway involves numerous different genes that code for many proteins such as receptors, transporters, and transcription factors [18, 19]. Genetically speaking, one of two pathways occur: 1) MC1R is stimulated by an agonist called alpha-melanocyte-stimulating hormone (α-MSH) and

triggers the production of eumelanin or 2) the stimulation and binding of an antagonist called Agouti-Signaling-Protein (ASIP) will occur and cause a shift towards pheomelanin instead [18, 19]. Ultimately, the visible spectrum of most pigmentation we see in hair, eye, and skin color is simply the ratio of pheomelanin to eumelanin.

1.3 Hair Structure Formation

The development of melanocytes follows a complex pathway. First, vacuoles arise and bud from the endoplasmic reticulum, forming premelanosomes. These premelanosomes then take in structural proteins and enzymes such as *TYR* and Keratin-Associated Proteins (KAP) [20]. From here, melanin synthesis begins resulting in melanosomes [19]. At this point, eumelanosomes only continue on with the binding of *TYRP1* and *DCT* enzymes. Pheomelonosomes become transferred to the keratinocytes in an area of already formed melanocytes. This mode of transportation to the keratinocytes is still unknown, but hypothesized to be through phagocytosis [18].

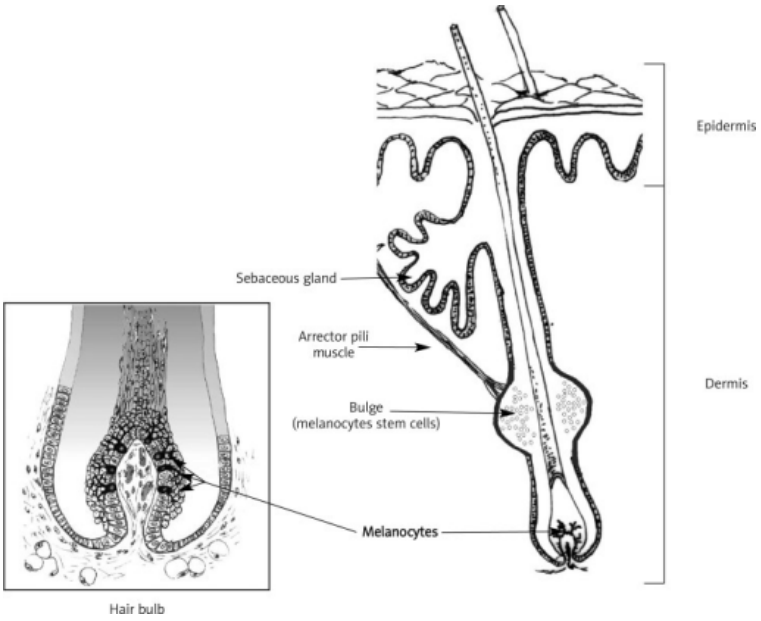


Figure 1 Distribution of Melanocytes in Hair Follicle

As seen in Figure 1, the melanocytes reside in the proximal bulb of each hair and in the sebaceous gland. The melanocytes transfer the melanosomes to keratinocytes, which move up the shaft of the hair, and provide the visible pigment in the hair [21]. The melanocytes in the hair die at the end of the hair cycle, which lasts approximately 3-8 years and is composed of four different phases: Anagen, Catagen, Telogen, and Exogen. During the anagen phase, the hair is in a state of active growth. Melanogenesis of the hair only takes place during this phase. The catagen phase is the transition phase of the hair in which the follicle begins to separate from the dermal papilla. Melanogenesis does not occur during this phase. The telogen phase is the resting phase for the hair. This is where the follicle completely separates from the dermal papilla and melanogenesis is completely absent. Finally, the hair goes through exogen phase, which is when it will shed naturally. Melanogenesis does not occur during the exogen phase [19, 22, 23].

Though pigmentation is the most noticeable hair-feature difference between individuals, the morphology of the hair is also very discriminating. In determining whether an individual has straight or curly hair, the instructions for the shape of the hair shaft begins in the hair follicle. Some of the major structural components of hair shaft fiber include hair keratin genes, keratin-associated proteins (KAP), and trichohyalin (*TCHH*) proteins [20, 24-26]. The *TCHH* gene is expressed in the developing inner root sheath of the hair follicle and provides instruction for the production of a protein called trichohyalin [25]. Trichohyalin binds to keratin intermediate filaments (KIF) to create cross-links that provide mechanical strength and structure to the hair shaft. Polymorphisms within the *TCHH* gene help to determine the shape and structure of the hair [26]. KAPs are located in the hair matrix around the keratin intermediate filaments and are coded by numerous multigene families that each consist of a single exon and no introns. They are one of the major components of the formation of the hair shaft through disulfide bond cross-linking with KIFs [20]. Studying

these major genes and proteins can provide vital information in the understanding of hair structure formation and therefore lead to improvements in phenotypic prediction models for hair type and proteomic methods for protein detection.

1.4 The HIrisPlex-S Assay for Eye, Hair, and Skin Color Prediction for DNA

Previous methods in Forensic DNA Phenotyping include the HIrisplex system [3], a novel and fully validated Forensic DNA Phenotyping tool that was released in 2014 by Walsh *et al*. This tool allowed for the prediction of the externally visible characteristics of eye and hair color. Of the 24 SNPS present in the HIrisPlex model, 6 of these SNPs were incorporated from a previously published model; the IrisPlex system [2] for the prediction of eye color. To expand on the amount of phenotypic knowledge gained from phenotyping, in 2017 Walsh *et al*. published the HIrisPlex-S system [4] which has drastically improved the hair and eye color prediction while also adding on the additional trait of skin pigmentation. To develop this, 77 SNPs were assessed from 37 genetic loci in 2025 globally dispersed individuals [5]. From this dataset, a novel prediction model was developed for a 5-scale system of skin color prediction based upon 36 SNPs from 16 genes. These five skin color categories received Area Under the Receiver Operating Curve (AUC) values of 0.83 for Very Pale, 0.76 for Pale, 0.78 for Intermediate, 0.98 for Dark, and 0.99 for Dark-Black skin color. This model was then combined with the previously established HIrisPlex system, creating the HIrisPlex-S system, a novel Forensic DNA Phenotyping tool for the prediction of eye, hair, and skin color [4, 5].

Input data for the IrisPlex and HIrisPlex, is generated with one multiplex genotyping assay while the HIrisPlex-S input data is generated with two. Both multiplex assays, HIrisPlex and HIrisPlex-S, have successfully undergone forensic developmental validation testing. Through this validation, it was determined that both assays are capable of generating full genotypic profiles

from a minimum DNA input concentration of 63 pg. In totality, the HIrisPlex-S system is composed of 41 SNPs (Figure 2): 24 SNPs in the first assay (HIrisPlex) and 17 SNPs in the second assay. Genotype data of these 41 DNA variants can then be uploaded to the online web tool found at https://HIrisPlex.erasmusmc.nl/ to generate individual prediction probabilities for 3 eye color, 4 hair color, and 5 skin color categories.

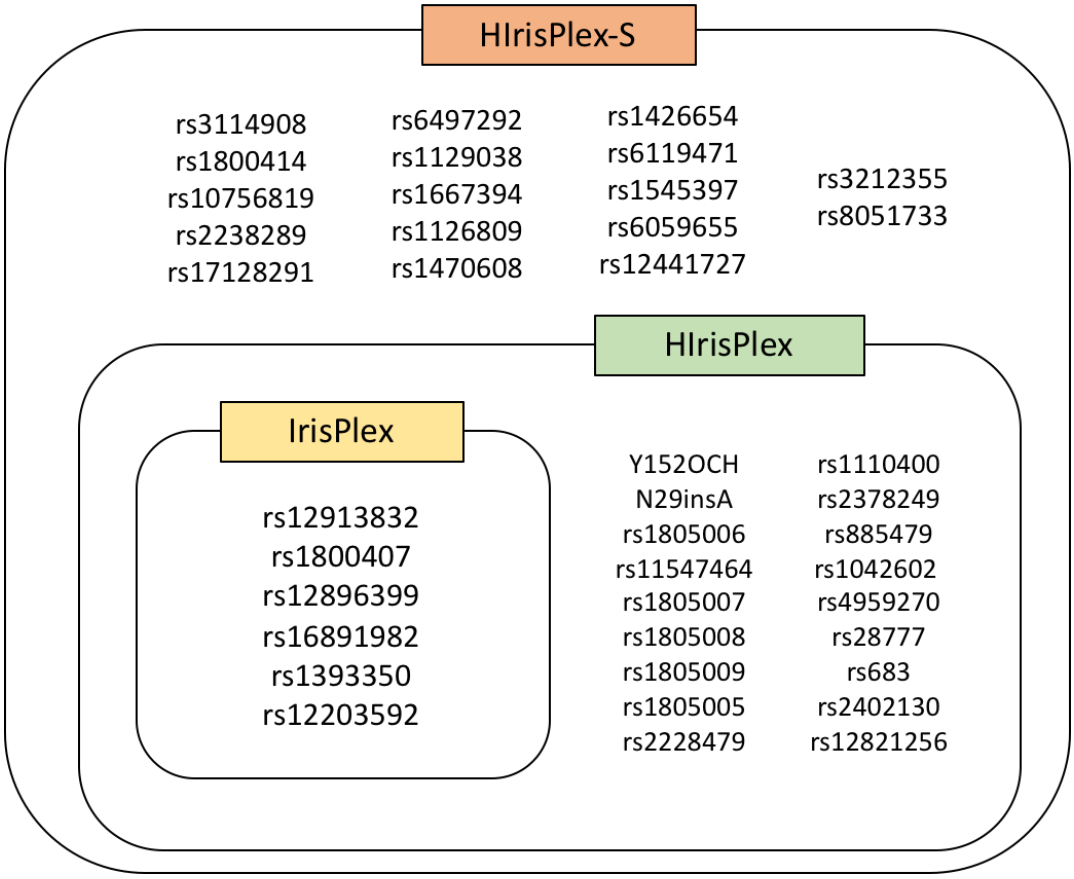


Figure 2 A summary of a total of 41 variants covered in the three prediction models: IrisPlex eye (6 variants), HIrisPlex (24 variants) and HIrisPlex-S (41 variants)

1.5 Massive Parallel Sequencing

Massive Parallel Sequencing (MPS) is an amplicon-based sequencing method. It can effectively address all of the short comings put forth by capillary electrophoresis by increasing the multiplex capacity, increasing the amount of sequencing data output, and decreasing the amount of input DNA required [10]. For massive parallel sequencing, libraries are generated after fragments, or amplicons, are isolated and then specific adapters are added for important use in downstream sequencing. These libraries are generated so that clonal amplification can take place on either a bead or flow cell. This allows for hundreds of thousands of copies or 'clusters' to be generated at once for sequencing [10]. There are multiple methods and platforms for massive parallel sequencing: semi-conductor sequencing (ThermoFisher Scientific IonTorrent™), sequencing by synthesis (Illumina™), pyrosequencing (Roche 454™), or sequencing by ligation (ThermoFisher Scientific SOLiD ™) [27-33]. The research conducted in this thesis will be focusing on the implementation of Forensic DNA Phenotyping on the Illumina Miseq FGx ™ (Illumina, San Diego, CA, USA) and the ThermoFisher Scientific Ion Torrent ™ (ThermoFisher Scientific, Waltham, MA, USA).

The concept behind how Illumina MiSeq FGx ™ MPS system functions holds several similarities to that of the CE in that a deoxyribonucleotide triphosphate (dNTP) is incorporated into the DNA strand through cycles of sequencing by synthesis that is facilitated by DNA polymerase. After this initial amplification, the PCR primers are tagged with individual barcodes for identification of the sample in downstream analyses. With more combinations of these individualizing barcodes, comes a greater ability to run a higher number of samples in one experiment. Once on the MiSeq, the pooled library containing the tagged fragments of DNA hybridizes to a flow cell through bridge PCR and amplifies, creating thousands of copies of the

sequences of interest. This is done simultaneously, creating “clusters” of millions of copies of DNA in parallel along the entire flow cell. The fluorescently labeled nucleotides incorporated within these “clusters” are imaged while being added through a process called “sequencing by synthesis”. This allows for identification by the machine after excitation from a laser generating sequence results that are then used for genome alignment and analysis [34].

In contrast, Ion Torrent differs in the chemical methodology behind the machine’s ability to sequence the fragments of interest. The Ion Torrent utilizes the simple concept of a proton being released when a nucleotide is incorporated by the polymerase into the actual DNA molecule rather than the “sequencing by synthesis” method that the Illumina Miseq FGx utilizes. This sequencing is detected through a pH change in the surrounding region due to that release of the hydrogen. The Ion Torrent utilizes a semiconductor sequencing chip that incorporates hundreds of thousands of copies of the DNA. When the nucleotides are released to the chips those that are similar to the DNA molecule are incorporated and hydrogen ions are released into the solution resulting in pH changes in the corresponding well. This change in pH results in detection by the ion sensor which then converts that chemically obtained information to actual digital information. The voltage change occurs based on the type of nucleotides, for example if two bases are identical the voltage will be doubled that of a single nucleotide [35].

1.6 Proteomics

The proteome is defined as the complete set of proteins that are expressed or modified by an organism [12], biological system, or cell. Thus, proteomics is the study of proteomes and their functions. Similar to the genome, a proteome and protein expression is unique to each human being. Therefore, studying the specific sequences of proteins in a given sample can provide important information about the DNA sequence, and thus the expression in the individual [12]. Despite their

similarities, the study of the proteome is far more complex than that of the genome. Each individual has one nuclear genome; however, this singular genome contains genes that can produce several different protein variations due to alternative splicing and other modification events. This genetic variation can come in the form of single amino acid polymorphisms (SAAP) within a genetically variant peptide (GVP) [7, 13]. Technologies like mass spectrometry-based shotgun proteomic sequencing, which utilizes microcapillary columns to separate peptides by hydrophobicity and charge, can be used to detect these SAAPs within a peptide. These SAAPs can then be used to infer the genotypes of non-synonymous single nucleotide polymorphism (nsSNP) alleles regardless of the presence of a DNA template in the sample [16]. The genotypes of these separated nsSNPs can be combined to form a profile of genetic variation for an individual and potentially be used to acquire identifying and biogeographic information [7, 13].

Current studies in forensic proteomics focus on the detection of peptides in hair. Hair is composed primarily of keratin, which is a coiled-coil protein with a high degree of intermolecular disulphide and isopeptide covalent bonds [13]. These bonds are responsible for the robustness and physical flexibility of the hair. However, despite the strong and stable properties of the hair, it is a poor source of nuclear DNA due to apoptosis of the keratinocytes during hair shaft biogenesis and the natural weathering of hair throughout a lifetime. Alternatively, the hair does not lack in protein content with more than 300 proteins already being detected in the hair proteome [36], providing solid grounds to assess the usefulness of protein analyses in forensic and bioarcheological domains.

In a study by Lee *et al*., 343 hair shaft proteins were identified using two-dimensional liquid chromatography mass spectrometry. Of these 343 detected proteins, many were keratin or keratin associated proteins and in high abundance [36]. Another study tested hair shaft proteomes of four different ethnic groups and found significant variation in the abundance of specific keratins

between individuals within each ethnic group. The variation was smaller between each ethnic group and centered on keratin-associated proteins. A study by Milan *et al*. analyzed the differences in protein abundance and genetically variant peptides in hair throughout different parts of the body. They discovered that the protein levels vary as a function of genetic origin and the genetically variant peptides are more dependent on the individual [16]. In another study by Parker *et al*., researchers have explored the ability to identify an individual from single amino acid polymorphisms (SAAPs) in hair proteins. A total of 596 SNP genotypes were accurately imputed from these SAAPs, allowing for population (European) statistics to be computed for the generated allelic profiles. By using the product rule and known allelic frequencies in the population, a power of discrimination value of up to 1 in 12,500 was calculated, showing the ability to apply match probabilities to a hair sample [13]. These studies give insight on the usefulness of proteomic analysis of hair samples and the potential for growth in forensic capacities.

1.7 References

1. Kayser, M.J.F.S.I.G., *Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes.* 2015. **18**: p. 33-48.
2. Walsh, S., et al., *Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence.* 2011. **5**(5): p. 464-471.
3. Walsh, S., et al., *Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage.* 2014. **9**: p. 150-161.
4. Chaitanya, L., et al., *The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation.* 2018. **35**: p. 123-135.
5. Walsh, S., et al., *Global skin colour prediction from DNA.* 2017. **136**(7): p. 847-863.
6. Draus-Barini, J., et al., *Bona fide colour: DNA prediction of human eye and hair colour from ancient and contemporary skeletal remains.* 2013. **4**(1): p. 3.
7. Mason, K.E., et al., *Development of a Protein-based Human Identification Capability from a Single Hair.* 2019.
8. Walsh, S., et al., *IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information.* 2011. **5**(3): p. 170-180.
9. Walsh, S., et al., *The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA.* 2013. **7**(1): p. 98-115.
10. Børsting, C. and N.J.F.S.I.G. Morling, *Next generation sequencing and its applications in forensic genetics.* 2015. **18**: p. 78-89.
11. Liu, L., et al., *Comparison of next-generation sequencing systems.* 2012. **2012**.
12. Merkley, E.D., et al., *Applications and challenges of forensic proteomics.* 2019.
13. Parker, G.J., et al., *Demonstration of protein-based human identification using the hair shaft proteome.* 2016. **11**(9): p. e0160653.
14. Laatsch, C.N., et al., *Human hair shaft proteomic profiling: individual differences, site specificity and cuticle analysis.* 2014. **2**: p. e506.
15. Higuchi, R., et al., *DNA typing from single hairs.* 1988. **332**(6164): p. 543.
16. Milan, J.A., et al., *Comparison of protein expression levels and proteomically-inferred genotypes using human hair from different body sites.* 2019.
17. Nordlund, J.J., et al., *The pigmentary system: physiology and pathophysiology*. 1998: Oxford University Press New York.
18. Scherer, D. and R.J.M.R.R.i.M.R. Kumar, *Genetics of pigmentation in skin cancer—a review.* 2010. **705**(2): p. 141-153.
19. Cichorek, M., et al., *Skin melanocytes: biology and development.* 2013. **30**(1): p. 30.
20. Shimomura, Y. and M. Ito. *Human hair keratin-associated proteins*. in *Journal of Investigative Dermatology Symposium Proceedings*. 2005. Elsevier.
21. Slominski, A., et al., *Hair follicle pigmentation.* 2005. **124**(1): p. 13-21.
22. Alonso, L. and E.J.J.o.c.s. Fuchs, *The hair cycle.* 2006. **119**(3): p. 391-393.
23. Tobin, D.J., et al. *The fate of hair follicle melanocytes during the hair growth cycle*. in *Journal of Investigative Dermatology Symposium Proceedings*. 1999. Elsevier.
24. Adav, S.S., et al., *Studies on the proteome of human hair-identification of histones and deamidated keratins.* 2018. **8**(1): p. 1599.
25. Medland, S.E., et al., *Common variants in the trichohyalin gene are associated with straight hair in Europeans.* 2009. **85**(5): p. 750-755.
26. Mlitz, V., et al., *Trichohyalin-like proteins have evolutionarily conserved roles in the morphogenesis of skin appendages.* 2014. **134**(11): p. 2685-2692.

27. Malausa, T., et al., *High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries.* 2011. **11**(4): p. 638-644.
28. Quail, M.A., et al., *A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.* 2012. **13**(1): p. 341.
29. Shin, S., et al., *Validation and optimization of the Ion Torrent S5 XL sequencer and Oncomine workflow for BRCA1 and BRCA2 genetic testing.* 2017. **8**(21): p. 34858.
30. McElhoe, J.A., et al., *Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq.* 2014. **13**: p. 20-29.
31. Jäger, A.C., et al., *Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories.* 2017. **28**: p. 52-70.
32. Schmidt, D., et al., *ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions.* 2009. **48**(3): p. 240-248.
33. Ondov, B.D., et al., *Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications.* 2008. **24**(23): p. 2776-2777.
34. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* 2008. **456**(7218): p. 53.
35. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* 2011. **475**(7356): p. 348.
36. Lee, Y.J., et al., *Proteome analysis of human hair shaft: from protein identification to posttranslational modification.* 2006. **5**(5): p. 789-800.

CHAPTER 2. HIRISPLEX-S SYSTEM FOR EYE, HAIR, AND SKIN COLOR PREDICTION FROM DNA: MASSIVELY PARALLEL SEQUENCING SOLUTIONS FOR TWO COMMON FORENSICALLY USED PLATFORMS

Krystal Breslin[1$], Bailey Wills[1$], Arwin Ralf[2], Marina Ventayol Garcia[3], Magdalena Kukla-Bartoszek[4], Ewelina Pośpiech[5], Ana Freire-Aradas[6], Catarina Xavier[7], Sabrina Ingold[8], Maria de La Puente[6,7], Kristiaan J. van der Gaag[3], Noah Herrick[1], Cordula Haas[8], Walther Parson[7,9], Christopher Phillips[6], Titia Sijen[3], Wojciech Branicki[5,10], Susan Walsh[1,#,*] and Manfred Kayser[2,#,*]
Submitted to FSI: Genetics.

Bailey Wills was responsible for optimizing the HPS-MPS assay, preparing the samples, running the samples at the U.S. site, compiling and analyzing the data from the sequencing run, creating tables, figures, and supplemental material, and writing, editing, and reviewing the manuscript draft. Bailey Wills also helped to create and assess the genotyping pipeline and the two-person mixture deconvolution tool discussed in the manuscript.

Krystal Breslin was responsible for designing the HPS-MPS assay, preparing the samples, analyzing the sequencing run data, and writing, editing, and reviewing the manuscript draft.

All other authors were involved in concordance testing for this manuscript.

[1] Department of Biology, Indiana University Purdue University Indianapolis (IUPUI), Indiana, U.S.A.
[2] Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands
[3] Department of Human Biological Traces, Netherlands Forensic Institute, The Hague, the Netherlands
[4] Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Kraków, Poland
[5] Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland
[6] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain
[7] Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria
[8] Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland
[9] Forensic Science Program, The Pennsylvania State University, University Park, PA, U.S.A.
[10] Central Forensic Laboratory of the Police, Warsaw, Poland

$,# these authors, respectively, contributed equally to this work

* Corresponding authors: SW: phone +1-317-274-0593, e-mail walshsus@iupui.edu, or MK: phone +31-10-7038073, e-mail m.kayser@erasmusmc.nl

2.1 Abstract

Forensic DNA Phenotyping (FDP) provides the ability to predict externally visible characteristics from minute amounts of crime scene DNA, which can help find unknown perpetrators who are typically unidentifiable via conventional forensic DNA profiling. Fundamental human genetics research has led to a better understanding of the specific DNA variants responsible for physical appearance characteristics, particularly eye, hair, and skin color. Recently, we introduced the HIrisPlex-S system for the simultaneous prediction of eye, hair, and skin color based on 41 DNA variants generated from two forensically validated SNaPshot multiplex assays using capillary electrophoresis (CE). Here we introduce massively parallel sequencing (MPS) solutions for the HIrisPlex-S (HPS) system on two MPS platforms commonly used in forensics, Ion Torrent and MiSeq, that cover all 41 DNA variants in a single assay, respectively. Additionally we present the forensic developmental validation of the two HPS-MPS assays. The Ion Torrent MPS assay based on Ion AmpliSeq technology illustrated successful generation of full HIrisPlex-S genotypic profiles from 100 pg of input control DNA, while the MiSeq MPS assay based on an in-house design yielded complete profiles from 250 pg of input DNA. Assessing simulated forensic casework samples such as saliva, hair (bulb), blood, semen, and low quantity touch DNA, as well as artificially degraded DNA samples, concordance testing, and samples from numerous species, all illustrated the ability of both versions of the HIrisPlex-S MPS assay to produce results that motivate forensic applications. By also providing an integrated bioinformatics analysis pipeline, MPS data can now be analyzed and a file generated for upload to the publically accessible HIrisPlex online webtool (https://hirisplex.erasmusmc.nl). In addition, we updated the website to accept VCF input data for those with genome sequence data. We thus provide a user-friendly and semi-automated MPS workflow from DNA sample to individual eye, hair, and skin color prediction probabilities. Furthermore, we present a 2-person mixture separation

tool that not only assesses genotype reliability with regards genotyping confidence, but also provides the most fitting mixture scenario for both minor and major contributor, including profile separation. We envision this MPS implementation of the HIrisPlex-S system for eye, hair and skin color prediction from DNA as a starting point for further expanding MPS-based forensic DNA phenotyping. This may include the future addition of SNPs predictive for more externally visible characteristics, as well as SNPs for bio-geographic ancestry inference, provided the statistical framework for DNA prediction of these traits is in place.

2.2 Introduction

The standard approach to forensic DNA profiling for human identification purposes uses short tandem repeats (STRs); however, when this method fails to identify a known suspect to be the contributor of a crime scene trace, alternative methods must be available to further the forensic investigation. Forensic DNA Phenotyping (FDP) utilizes fundamental genetics knowledge in order to provide information on an unknown crime scene donor. Typically, FDP involves appearance traits, but inferring bio-geographic ancestry from DNA [1, 2] and estimating a person's age from DNA [3, 4], are also considered under the wider umbrella of FDP. FDP represents an innovative investigation approach to forensic DNA application due to its ability to potentially strengthen or challenge eye-witness statements and provide biological witness information in cases without human eye-witnesses [5]. It is useful to narrow down suspect lists that are extensive in cases without known suspects [6, 7]. It has also been used to infer appearance predictions from skeletal remains [8], including those in the field of ancient DNA [9], making it useful in missing person identification cases that lack knowledge on the putative identity or on possible relatives.

Recently, Walsh et al. [10] considerably improved knowledge on categorical skin color prediction from DNA, which has subsequently been used to extend the previously established

IrisPlex system for eye color and HIrisPlex system for hair color prediction from DNA [11-13] by developing the HIrisPlex-S system for eye, hair and skin color prediction from DNA [14]. Based upon current available statistical models, eye color prediction using the IrisPlex model [13] achieves prediction accuracies expressed as Area Under the Receiver Operating Curve (AUC) of 0.94 for blue, 0.74 for intermediate, and 0.95 for brown eye color. Hair color prediction using the HIrisPlex model [11, 12] achieves an AUC performance metric of 0.93 for red, 0.81 for blond, 0.74 for brown, and 0.86 for black, and skin color prediction using the HIrisPlex-S model [10] achieves an AUC performance metric of 0.83 for Very Pale, 0.76 for Pale, 0.78 for Intermediate, 0.98 for Dark, and 0.99 for Dark to Black (based on full data model performance as measured March 2019 on https://hirisplex.erasmusmc.nl). In order to generate genetic data for input into the IrisPlex, HIrisPlex and HIrisPlex-S models, one (IrisPlex and HIrisPlex) or two (HIrisPlex-S) multiplex genotyping assays were previously established and successfully underwent forensic developmental validation testing [11, 13, 14]. The HIrisPlex-S system includes 41 DNA variants; 24 variants targeted with the HIrisPlex assay [11, 12] and 17 additional variants targeted with a second assay [14]. Both multiplex assays are capable of generating full genotypic profiles from a minimum DNA input of 63 pg [11, 14]. Genotype data of these 41 DNA variants can then be uploaded to the easy-to-use web tool found at https://HIrisPlex.erasmusmc.nl/ to generate individual prediction probabilities for 3 eye color, 4 hair color, and 5 skin color categories [14].

While the previously developed IrisPlex, HIrisPlex and HIrisPlex-S multiplex assays have been demonstrated to be sensitive, robust and reliable, and able to cope with the specific requirements of forensic DNA analysis in dealing with low quantity and quality DNA [10-15], the underlying single base primer extension (SNaPshot) and capillary electrophoresis (CE) technologies have limitations. Due to the chemistry and fragment sizing used, the multiplex

capacity of a SNaPshot genotyping assay is typically limited to approximately 25 DNA variants per single assay. As a consequence, if more DNA variants need to be analyzed for a specific forensic purpose, for example, the 41 from the HIrisPlex-S system, more multiplex assays, such as two in the case of HIrisPlex-S, have to be developed, validated and finally applied. This leads to the consumption of additional evidence DNA that in some cases may not be available, also given that FDP is typically performed subsequently to DNA-consuming conventional forensic STR profiling. Moreover, running multiple assays increases the time, cost, and efforts needed. In addition, deconvolution or separation of DNA mixture profiles generated by several contributors is challenging, if not impossible, with SNP assays using SNaPshot and CE, because of the semi-quantitative nature of these technologies. Lastly, the limited multiplex capacity of the SNaPshot-CE approach means it cannot easily be expanded to include new DNA variants from developments in the field for additional appearance traits for which statistical prediction models have already been developed, such as hair structure and hair loss [16-19] for example, as well as ancestry-informative SNPs. Therefore it is apparent that a transition towards targeted massively parallel sequencing (MPS) solutions is required for FDP purposes in order to take advantage of the increasing knowledge that improved appearance (and ancestry) genetics provides.

Targeted MPS technologies are characterized by a dramatic increase in multiplexing capacity relative to all DNA technologies previously used in forensic DNA analysis including SNaPshot-CE. The technological transition from CE to MPS has started in the forensic field, mostly for STRs, but also for SNP sequencing (e.g. using ThermoFisher Scientific Precision ID Identity Panel [20], and ThermoFisher Scientific Precision ID whole mtDNA genome [21], respectively. In addition, the multi-purpose ForenSeq™ DNA Signature Prep [22], developed by Illumina (now Verogen) includes the HIrisPlex DNA markers. Moreover, non-commercial developments have

demonstrated that several hundreds of SNPs can be simultaneously analyzed via single targeted MPS assays, as demonstrated for Y-SNPs [7] and the entire mitogenome [23]. Here we describe the development and forensic validation of MPS solutions for the HIrisPlex-S (HPS) system for the two MPS platforms most commonly used in forensic genetics; Ion Torrent (ThermoFisher Scientific) and MiSeq (Illumina). Parallel assessments were made because of the differing performance metrics, as well as the different underlying sequencing principles and methodologies in each MPS platforms; sequencing by synthesis for MiSeq and semi-conductor sequencing for Ion Torrent. Although there is the potential to run identical MPS assays on both platforms, for the present study, the applied MiSeq assay design reflected an in-house alternative to the commercial AmpliSeq assay design used for Ion Torrent.

2.3 Materials and Methods

Study samples were collected in compliance with Indiana University IRB#1409306349 and included informed consent for all individuals. Test samples were made up of single and multiple source samples, including simulated casework (saliva, blood, semen, hair (including bulb), vaginal swabs and touched items) and non-human samples. Supplementary Table 1 (Appendix A) describes the 96 samples used for this forensic developmental validation, their DNA input concentration for sequencing, and eye, hair and skin color phenotypes. DNA was extracted using an in-house salting out protocol (unpublished). Sample DNA concentrations were determined by qPCR via InnoQuant Human DNA Quantification and Degradation Assessment Kit [24] and/or the Quantifiler Trio DNA Quantification kit (ThermoFisher Scientific (TFS), Waltham, MA, USA) and plated at the US site. Results from the 96 samples were used to generate data to test the performance of two MPS assays, one in-house designed for the Illumina MiSeq platform by the US site, and a separate Ion AmpliSeq assay designed by the Rotterdam side together with

Jagiellonion University and TFS for use on the Ion Torrent platform. Both MPS assays underwent forensic developmental validation testing using these samples. In addition, all genotypes generated through amplicon sequencing were also confirmed through CE-SBE genotyping using previously published and developmentally validated HIrisPlex and HIrisPlex-S SNaPshot assays[11, 14].

HIrisPlex-S Assay design for Massive Parallel Sequencing using MiSeq (HPS-MPS-MiSeq)

The custom protocol used to generate the assay design for the Illumina MiSeq Sequencer was based on an in-house design using modifications of the protocol published by Bronner *et al.* [25]. Each of the primer pairs were designed to isolate between 100 to 300bp around the variant of interest using a proposed optimal primer pair from the free web-based design tool Primer3Plus [26]. These primers also included specific adapter sequences, therefore allowing the fragments or amplicons to adhere to the lawn found on the Illumina MiSeq flow cell. The selection of the primers used in this design was checked by using the program Bisearch [27] to ensure that specific unique amplicons were generated. Lastly, the program AutoDimer [28] was used to check for primer-dimers and/or primer to primer interactions (including potential interactions with adapter sequences) within the multiplex. The hg19 position of the 41 variants used in the HIrisPlex-S system, including the primer pair designs with incorporated adapter sequences for the Illumina MiSeq protocol, named HPS-MPS-MiSeq, can be found in Table 1 below.

Table 1 Information on the 41 DNA variants used in the HirisPlex-S system, including the primer pair designs with incorporated adapter sequences used for the HPS-MPS-MiSeq protocol, and their concentration

SNP	Gene	Chromosome	Position	Ref Allele	Alt Allele	Amplicon	Forward Primer	Reverse Primer	Product Size with adapters (bp)	Input Concentration (µM)
rs796296176	MC1R	16	89985753	A insertion	-	MC1R Amplicon 1	TCGTCGGCAGCGTCAGATGTGTATAAGAG ACAGGCAGGGATCCCAGAGAAGAC	GTCTCGTGGGCTCGGAGATGTGTATA AGAGACAGTCAGAGATGGACACCTCCAG	184	0.7
rs11547464	MC1R	16	89986091	G	A	MC1R Amplicon 2	TCGTCGGCAG CGTCAGATGTGTATA AGAGACAGCTGGTGAGCTTGGTGGAGA	GTCTCGTGGGCTCGGAGATGTGTATA AGAGACAGTCCAGCAGGAGGATGACG	225	0.7
rs885479	MC1R	16	89986154	G	A					
rs1805007	MC1R	16	89986117	C	T					
rs1805008	MC1R	16	89986144	C	T					
rs201326893	MC1R	16	89986122	C	A					
rs1110400	MC1R	16	89986130	T	C					
rs2228479	MC1R	16	89985940	G	A	MC1R Amplicon 3	TCGTCGGCAGCGTCAGATGTGTATAAGAG ACAGGTCCAGCCTCTGCTTCCTG	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGAGCGTGCTGAAGACGACAC	214	0.7
rs1805005	MC1R	16	89985844	G	T					
rs1805006	MC1R	16	89985918	C	A					
rs1805009	MC1R	16	89986546	G	C	MC1R Amplicon 4	TCGTCGGCAGCGTCAGATGTGTATAAGA GACAGCAAGAACTTCAACCTCTTTCTCG	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGCACCTCCTTGAGCGTCCTG	173	0.5
rs28777	SLC45A2	5	33958959	C	A	SLC45A2 Amplicon 1	TCGTCGGCAGCGTCAGATGTGTATA AGAGACAGCTTTCAAAAGGCTTCCACTCA	GTCTCGTGGGCTCGGAGATGTGTATAAGAG ACAGTCTTTGATGTCCCCTTCGAT	195	0.6

Table 1 continued

rs16891982	SLC45A2	5	33951693	C	G	SLC45A2 Amplicon 2	TCGTCGGCAGCGTCAGATGTGTATA AGAGACAGTCCAAGTTGTGCTAGACCAGA	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGCGAAAGAGGAGTCGAGGTTG	195	0.4
rs12821256	KITLG	12	89328335	T	C	KITLG Amplicon	TCGTCGGCAGCGTCAGAT GTGTATAAGAGACAGATGCC CAAAGGATAAGGAAT	GTCTCGTGGGCTCGGAGATGTGTATA AGAGACAGGGAGCCAAGGGCATGTTACT	185	0.6
rs4959270	EXOC2	6	457748	C	A	EXOC2 Amplicon	TCGTCGGCAGCGTCAGATGTGTA TAAGAGACAGTGAGAAATCTACCCCCACGA	GTCTCGTGGGCTCGGAGATGTGTATA AGAGACAGGTGTTCTTACCCCCTGTGGA	207	0.4
rs12203592	IRF4	6	396321	C	T	IRF4 Amplicon	TCGTCGGCAGCGTCAGATGTGTAT AAGAGACAGAGGGCAGCTGATCTCTTCAG	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGGCTTCGTCATATGGCTAAACCT	193	0.5
rs1042602	TYR	11	88911696	C	A	TYR Amplicon	TCGTCGGCAGCGTCAGATGTGTATA AGAGACAGCAACACCCATGTTTAACGACA	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGGCTTCATGGGCAAAATCAAT	191	0.55
rs1800407	OCA2	15	28230318	C	T	OCA2 Amplicon 1	TCGTCGGCAGCGTCAGATGTGTATA AGAGACAGAAGGCTGCCTCTGTTCTACG	GTCTCGTGGGCTCGGAGATGTGTATAAGA GACAGCGATGAGACAGAGCATGATGA	191	0.35
rs2402130	SLC24A4	14	92801203	G	A	SLC24A4 Amplicon	TCGTCGGCAGCGTCAGATGTGTATAA GAGACAGACCTGTCTCACAGTGCTGCT	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGTTCACCTCGATGACGATGAT	217	0.35
rs12913832	HERC2	15	28365618	A	G	HERC2 Amplicon 1	TCGTCGGCAGCGTCAGATGTGTATAA GAGACAGTGTTCTTCATGGCTCTCTGTG	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGGGCCCCTGATGATGATAGC	163	0.45
rs2378249	PIGU	20	33218090	G	A	PIGU Amplicon	TCGTCGGCAGCGTCAGATGTGTATAA GAGACAGCGCATAACCCATCCCTCTAA	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGCATTGCTTTTCAGCCCACAC	203	0.35
rs12896399	SLC24A4	14	92773663	G	T	SLC24A4 Amplicon	TCGTCGGCAGCGTCAGATGTGTATA AGAGACAGCTGGCGATCCAATTCTTTGT	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGGACCCTGTGTGAGACCCAGT	192	0.4
rs1393350	TYR	11	89011046	G	A	TYR Amplicon	TCGTCGGCAGCGTCAGATGTGTATAA GAGACAGTTTCTTTATCCCCCTGATGC	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGGGGAAGGTGAATGATAACACG	191	0.6
rs683	TYRP1	9	12709305	C	A	TYRP1 Amplicon	TCGTCGGCAGCGTCAGATGTGTATAA GAGACAGCACAAAACCACCTGGTTGAA	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGTCCCAGCTTTGAAAAGTATGC	194	0.8

Table 1 continued

rs3114908	ANKRD11	16	89383725	T	C	ANKRD11 Amplicon	TCGTCGGCAGCGTCAGATGTGTATAAG AGACAGCAGAACACAGCCACACCCTA	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGACAGGAATGGCAGCTTTGAG	166	0.2
rs1800414	OCA2	15	28197037	T	C	OCA2 Amplicon 2	TCGTCGGCAGCGTCAGATGTGTATAAG AGACAGGCTGCAGGAGTCAGAAGGTT	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGGGGACAAACGAATTGAGGAA	212	0.65
rs10756819	BNC2	9	16858084	G	A	BNC2 Amplicon	TCGTCGGCAGCGTCAGATGTGTATA AGAGACAGGACCAGTTATTTTGGGTTTGGA	GTCTCGTGGGCTCGGAGATGTGTATAAGA GACAGCGTCATGACTAGAAAAACACCAA	143	0.4
rs2238289	HERC2	15	28453215	A	G	HERC2 Amplicon 2	TCGTCGGCAGCGTCAGATGTGTAT AAGAGACAGGGAACATGAAGATTTCCCAGT	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGCTGATTCAGGTCTGCTGTCACT	179	0.25
rs17128291	SLC24A4	14	92882826	A	G	SLC24A4 Amplicon	TCGTCGGCAGCGTCAGATGTGTAT AAGAGACAGCCAGCACTGCCAAAATAACA	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGCTCTTTGGACCCATCACCTC	196	0.4
rs6497292	HERC2	15	28496195	A	G	HERC2 Amplicon 3	TCGTCGGCAGCGTCAGATGTGTATA AGAGACAGTCTGCTGTAGAACCAATGTCC	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGGAATTGCACCTGTAGCTCCAT	217	0.4
rs1129038	HERC2	15	28356859	G	A	HERC2 Amplicon 4	TCGTCGGCAGCGTCAGATGTGTATA AGAGACAGATGTCGACTCCTTTGCTTCG	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGACACCAGGCAGCCTACAGTC	204	0.4
rs1667394	HERC2	15	28530182	C	T	HERC2 Amplicon 5	TCGTCGGCAGCGTCAGATGTGTATAAG AGACAGCAGCTGTAGAGAGAGACTTTGAGG	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGCAGCAATTCAAAACGTGCAT	184	0.4
rs1126809	MC1R	16	89017961	G	A	MC1R Amplicon 5	TCGTCGGCAGCGTCAGATGTGTATAAGAG ACAGTGTTTCTTAGTCTGAATAACCTTTTCC	GTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGGGTGCATTGGCTTCTGGATA	167	0.4
rs1470608	OCA2	15	28288121	G	T	OCA2 Amplicon 3	TCGTCGGCAGCGTCAGATGTGTATAAGAG ACAGTTTCTTGTGTTAACTGTCCTTACAAA	GTCTCGTGGGCTCGGAGATGTGTATAAGA GACAGGGAAAATATGTTAGGGTTGATGG	212	0.8
rs1426654	SLC24A5	15	48426484	A	G	SLC24A5 Amplicon	TCGTCGGCAGCGTCAGATGTGTATAAG AGACAGTTCAGCCCTTGGATTGTCTC	GTCTCGTGGGCTCGGAGATGTGTATAAGAG ACAGTGAGTAAGCAAGAAGTATAAGGAGCA	190	0.8

Table 1 continued

rs6119471	ASIP	20	32785212	C	G	ASIP Amplicon	TCGTCGGCAGCGTCAGATGTGTATAAGAGAC AGAAAAGAAGTAGCTGTACTAGACGGGAT	GTCTCGTGGGCTCGGAGATGTGTATAAGAG ACAGAACCCGAAGGAAGAGTGAAAA	130	0.25
rs1545397	OCA2	15	28187772	A	T	OCA2 Amplicon 4	TCGTCGGCAGCGTCAGATGTGTATAAGAGAC AGAAAGTGTTCTGGAATTGGATACTGACAA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGAAATGGAGATATAGAATTCACACAACA	188	0.8
rs6059655	RALY	20	32665748	A	G	RALY Amplicon	TCGTCGGCAGCGTCAGATGTGTATAAGAGAC AGGTGAGGAAATCGAGGCTCAG	GTCTCGTGGGCTCGGAGATGTGTATA AGAGACAGAGGAGAAAGCTGCAGATCCA	179	0.45
rs12441727	OCA2	15	28271775	G	A	OCA2 Amplicon 5	TCGTCGGCAGCGTCAGATGTGTATAA GAGACAGGGGAAGAGACAGCTCCATGT	GTCTCGTGGGCTCGGAGATGTGTATA AGAGACAGACAATCCTGGGAGGTACACG	204	0.35
rs3212355	MC1R	16	89984378	C	T	MC1R Amplicon 6	TCGTCGGCAGCGTCAGATGTGTATAAG AGACAGTTCCACCCTTCAGCACAGA	GTCTCGTGGGCTCGGAGATGTGTATA AGAGACAGCATCAAAGGCAGACCTCTCG	211	0.8
rs8051733	DEF8	16	90024206	A	G	DEF8 Amplicon	TCGTCGGCAGCGTCAGATGTGTATAAGA GACAGAGGCGGTGGTCTCTCTCTC	GTCTCGTGGGCTCGGAGATGTGTATAA GAGACAGTTGCAACAGGAGGGTCTAGG	191	0.3

Due to the temperature range needed for the incorporation of multiple primers in this multiplex, a touchdown PCR program was applied for the first PCR of the assay, using an Eppendorf Mastercycler Nexus SX1 and cycles: 1) 94 °C for 10 minutes, 2) 14 cycles of 94 °C for 20 seconds and 64 °C (with temperature decreases of -0.6 °C per each additional cycle) for 1 minute each (touchdown range of 64 °C – 55.6 °C), 3) 20 cycles of 94 °C for 20 seconds and 57 °C for 1 minute, and 68 °C for 30 seconds, 4) 72 °C for 3 minutes, 5) hold at 10ºC. PCR amplification was performed in a single multiplex PCR assay with a 10 µL total volume containing 1 µL genomic DNA (varying concentrations), primers (see Table 1), 1X PCR gold buffer (Applied Biosystems), 2.5 mM $MgCl_2$ (Applied Biosystems), 220 µM of each dNTP (TFS) and 2 U AmpliTaq Gold DNA polymerase (Applied Biosystems). Bead clean-up was then performed using a ratio of 9 µL of AmPure XP (Beckman Coulter, Indianapolis, IN, USA) beads to 5 µL PCR product. After mixing thoroughly, the samples were incubated for 5 minutes to allow for binding of the beads to the DNA, then the samples were placed on a magnetic stand for 5 minutes. All but 5 µL of the supernatant was removed and discarded while on the stand and then washed with 200 µL of 70% Ethanol. The ethanol was removed in the same fashion and then the wash was repeated after 30s. The samples were air dried for 2-5 minutes, re-suspended in 20 µL of purified water, and mixed thoroughly. After a 2-minute incubation, the samples were placed on a magnetic stand for 1 minute, and then transferred to a new plate.

The second round of PCR amplification was then performed to add index sequences to each sample as a unique identifier in order to de-multiplex (separate) each individual's FASTQ files after sequencing. For each well 5 µL of KAPA master mix (KAPA Biosystems, Wilmington, MA), 1 µL of each Nextera index (both forward and reverse to total 2 µL), 2 µL of H_20, and 1 µL of DNA were added to each well. The samples were placed on the thermocycler with the following

protocol: 1) 98 °C for 2 minutes, 2) 12 cycles of 98 °C for 30 seconds and 72 °C for 30 seconds, 3) 72 °C for 5 minutes, 4) hold at 15ºC. Another bead clean up, as described above, followed this indexing reaction.

To successfully sequence the 96 samples in one sequencing run, the products were pooled, diluted and quantified as follows to complete the library preparation. 5 µL of each sample was pooled and then quantified using the Qubit Fluorometer (TFS) following the standard manufacturer's guidelines. An in-house dilution calculator was used then to calculate an accurate dilution to a 2nM overall library concentration. Denaturing the library used 5 µL of 0.2 N NaOH to 5 µL of the 2nM library. Tubes were spun down, then incubated for 5 minutes at room temperature. The library was diluted to 10pM with 990 µL of Hybridization Buffer (Illumina, San Diego, CA) as provided with the Illumina Nextera XT Version 2 Kit (Illumina, San Diego, CA) and further diluted to 8pM using 480 µL of library and 120 µL of Hybridization Buffer with pulse vortexing. For optimal sequencing results, a PhiX control was spiked in at 20% to standardize the run. Preparation of the controls was as follows: 5 µL of the 4nM PhiX library was added to 5 µL of 0.2 N NaOH. The control was then vortexed, spun down, and incubated for 5 minutes. A further dilution was performed using 10 µL of PhiX Library and 990 µL of Hybridization Buffer to a final concentration of 20pM PhiX library. The final dilution to 12.5 pM was then performed using 375 µL of the previously diluted PhiX library and 225 µL of Hybridization Buffer. The last step before adding the samples to the MiSeq cartridge was to spike in the 20% PhiX control to the custom library (120 µL PhiX and 480 µL of the library). 600 µL of the combined library and control was then loaded into the MiSeq cartridge. The Illumina MiSeq v2 Reagent Kit cartridge was then run on 'Research Use Only' Mode through the Nextera XT sequencing. The MiSeq Reporter software (Illumina, San Diego, CA) de-multiplexes the samples by utilizing the uploaded sample sheet to

assign sample names through correlation with the appropriate indices. Sequences are then exported as two paired-end FASTQ files, titled with their respective sample name for use in downstream analyses.

HIrisPlex-S Assay design for Massive Parallel Sequencing using Ion Torrent (HPS-MPS-ION)

The Ampliseq primers were designed and tested for good performance by TFS, Erasmus MC Rotterdam and Jagiellonian University. Ion AmpliSeq™ Library Kit 2.0 chemistry (TFS) was used following the manufacturer's guidelines and using Ion Xpress™ Barcode Adapters (TFS). Twenty cycles of amplification as well as the incubation steps were performed on a Veriti™ 96-Well Thermal Cycler (TFS), the libraries were quantified using the TaqMan™ Library Quantitation Kit (TFS) on a CFX96 Touch™ Real-Time PCR machine (Bio-Rad, Hercules, CA, USA), then normalized and pooled accordingly. Template preparation was performed using the Ion PGM™ Hi-Q™ View OT2 Kit (TFS) following the manufacturer's guidelines. Sequencing of 48 samples per chip was performed on Ion 318™ Chip Kit v2 BC (TFS) using Ion PGM™ Hi-Q™ View Sequencing Kit (TFS) following the manufacturer's guidelines. The Ion Personal Genome Machine™ (PGM™) System (TFS) was used for simultaneous sequencing of all Ion Torrent applications of this study (apart from Site 4, which used the Ion S5 system and 530 chip). Torrent Suite version 5.2.2 was used for initial data processing and base calling, the resulting FASTQ files were exported and used for downstream pipeline analysis.

Sensitivity & sequence coverage

The sensitivity of both MPS assays was evaluated to determine the minimum input needed to obtain a complete 41-SNP HPS profile. Two commercial control DNA samples, 9947A

(OriGene, Rockville, MD, USA) and 9948 (OriGene, Rockville, MD, USA), were used to prepare serial dilutions to concentrations of 5pg, 10pg, 25pg, 50pg, 100pg, 250pg, 500pg, and 1ng. For the MiSeq assessment, each concentration was performed in duplicate for both controls. These high quality control samples were used to assess each HPS amplicons' accuracy and sequencing coverage at differing concentrations for each assay design and were used to set thresholds for genotype calling used in the Threshold & Mixture Tool (Supplementary Table 2 [Appendix B]). For HPS-MPS-MiSeq calls, these threshold values were calculated from two control samples run in duplicate at 100pg and 50pg for a total of 4 samples at each concentration (see Supplementary Table 5 [Appendix E] for more details). For HPS-MPS-ION, these threshold values were calculated from two control samples run at 100pg and 50pg for a total of 2 samples at each concentration. Percent sequencing error of the controls was calculated as the number of incorrect calls at that variant site within the amplicon as determined by sequencing quality and the BCFtools mpileup and call algorithm. The allele depth DP4 classification was assessed from the VCF file and is defined as the number of: 1) forward ref alleles; 2) reverse ref; 3) forward alt; 4) reverse alt alleles, used in variant calling at a site. For example, if an expected AA genotype within an individual displayed sequence reads of an allele other than A at that site (or no A allele at all), this was used to calculate that sites % error. In addition, an assessment of the genotype calls (homozygote and heterozygote) and coverage of each HPS variant site with a 500 pg DNA input from multiple individuals (n=8), generated by the HPS-MPS pipeline, was also evaluated, including standard deviation of the mean.

Simulated casework, stability testing and mixture assessment

For the simulated casework samples, samples were manufactured with dried and UV degraded blood, dried and UV degraded saliva, wet saliva, touch DNA, hair, vaginal swab, and vaginal swab mixture with semen (see Supplementary Table 1 [Appendix A]). These samples were extracted with the salting out method and quantified using the Quantifiler Trio DNA Quantification kit (TFS) to assess quantity and quality of the samples prior to library preparation.

DNA from one individual measured at 500pg DNA was then exposed to UV light for time intervals of 0, 5, 10, and 20 minutes using the CL-1000 Ultraviolet Crosslinker (Ultra-Violet Products Ltd, Upland, CA, USA) at a strength of 50 J/cm2 in order to test the robustness of each assay to analyze Degraded DNA.

Two person mixtures were tested in ratios of 1:1, 1:2, 1:5, and 1:10 in duplicate. To ensure a mixture of DNA variants were present in the sample, there were two sets of 2-person mixtures (number of individuals = 4) were set up to contribute to the sample mixtures that had differing eye, hair, and skin colors (see Supplementary Table 1 [Appendix A] for more details). The 2-person mixture deconvolution tool (see Supplementary Table 2 [Appendix B]) was designed using a 2-person ratio calculation (Minor:Major ratio out of 1 e.g. a 1:1 ratio is ½ and was input as 0.5) in addition to using knowledge of heterozygote read count ranges as observed from the 500 pg variant coverage samples from 8 individuals (performed in duplicate). The calculator displays read counts (+/- sd) for all 2-person mixture scenarios using the allele read counts input from the HPS-MPS pipeline for that particular mixture sample. A guide on how to use the tool is outlined in Supplementary Material 1 (Appendix K).

Species specificity and concordance testing

Species specificity testing is necessary in order to determine the possible contributors of a biological sample, as crime scenes can be prone to contamination from non-human sources. Therefore, each assay was tested for human specificity against samples of cat, primate, dog, mouse, and pig DNA at 1ng input. All non-human samples were extracted through an in-house extraction method, apart from the chimp sample obtained from a collaborator (Dr. Brenda Bradley - George Washington University).

Concordance testing ensures that the two assays perform consistently among different laboratories and personnel with varying experiences. To do this, a concordance plate (sample n=16 subset) was generated from the 96-sample set used in this study by the US (Illumina MiSeq platform) and Netherlands Erasmus MC (Ion Torrent platform) sites. This concordance plate was sent to five external European collaborators. Information on the samples and the sites instrumentation can be found in Supplementary Table 3 (Appendix C). Users were asked to indicate if the sample was a mixture or a single source. If a single source was indicated, users were also asked to provide a final predicted profile.

Genotype calling and webtool upload

For consistency, a pipeline was designed so that both platforms were assessed using the same algorithms to generate the 41 genotype calls needed for prediction model input to the web-based HIrisPlex-S prediction tool. See Figure 3 (Figure 1 in manuscript) for pipeline overview. Raw data was aligned to the hg19 human reference sequences for all amplicons using the mem algorithm within BWA (http://bio-bwa.sourceforge.net/) [29]. The sequence alignment/map (SAM) file was converted and sorted using SAMtools into a BAM file [30] and read groups added via Picard Tools (https://github.com/broadinstitute/picard). Variant calling was performed by

BCFtools (http://github.com/samtools/bcftools) using mpileup (set to a depth read of 8000), call (using the multi-allelic caller for all sites –m -M) and query commands for SNP extraction. For more information of how BCFtools multiallelic caller performs during genotype calling, please see the manual found at https://samtools.github.io/bcftools/. Finally, the Java applet VarScan [31] was used to detect the presence or absence of the INDEL rs796296176 (variant 1 of HIrisPlex). The R program [32] was used to generate the upload file required for usage on the HIrisPlex webtool site. The pipeline can generate HPS-MPS results for up to 96 samples at a time; however, this script is customizable to include more samples if desired. In addition, the environment needed to run this pipeline has also been packaged into a Docker (https://www.docker.com) container image, which can be accessed via the Docker Hub under suswalsh/hpsmps. All information regarding the pipeline and a guide on how to use it can be found in Supplementary Material 2 (Appendix L).

2.4 Results and Discussion

MPS assay design and analysis pipeline

Two versions of the HIrisPlex-S MPS-based lab tool for the two MPS platforms commonly used in forensic genetics were designed and assessed in this study, HPS-MPS-MiSeq and HPS-MPS-ION, to target the 41 DNA variants included in the HIrisPlex-S system via 34 amplicons. Care was taken to design the amplicons to be as short as possible to optimize analysis of low quality DNA, commonly encountered in forensic DNA testing. For HPS-MPS-MiSeq, the size of the 34 amplicons ranged between 130 and 225 bp in length and the average length across amplicons was 124 bp. For HPS-MPS-ION, the insert size ranged between 44 and 113 bp and the average insert size across amplicons was 71 bp.

The HPS-MPS analysis pipeline was designed to be user-friendly and semi-automated to ease the entire process from DNA sample to the sample donors eye, hair, and skin color prediction probabilities, estimated via the HIrisPlex webtool (https://HIrisPlex.erasmusmc.nl/). In its current version, the analysis pipeline can run 96 samples at a time and simply requires the sample name and raw FASTQ sequence files generated from any sequencer, as per the instructions found in Supplementary Material 2 (Appendix L), but customization towards more samples is possible by the user. As part of the HPS-MPS analysis pipeline, the sequence files are aligned to the human reference sequence hg19 (obtained from ftp://hgdownload.soe.ucsc.edu/goldenPath/hg19) and genotypes are extracted at the 41 DNA variant sites using a location text file. The process and tools used are illustrated in Figure 3 (Figure 1 in manuscript) below.

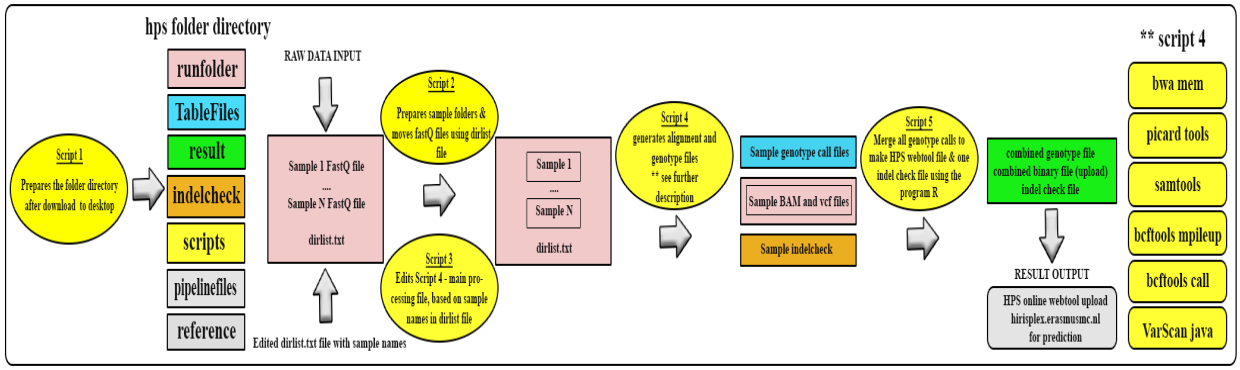


Figure 3 (Figure 1 in manuscript) Illustrative example of the HIrisPlex-S MPS pipeline used to assess and call genotypic information from raw HPS-MPS sequencing data and file generation for online webtool input using an automated set of scripts and programs

A more thorough guide is provided in Supplementary Material 2 (Appendix L) that explains the entire process including the computer set up needed to run these analyses. This pipeline can be run on all computer platforms; it is primarily designed (scripts) for use on a linux platform, however due to the use of the Docker container and its internal linux environment, it can be used on any platform (Mac OS, and Windows PC). An organized folder system is created for each sample to easily find sorted bam and vcf files. A table file folder contains all result files with

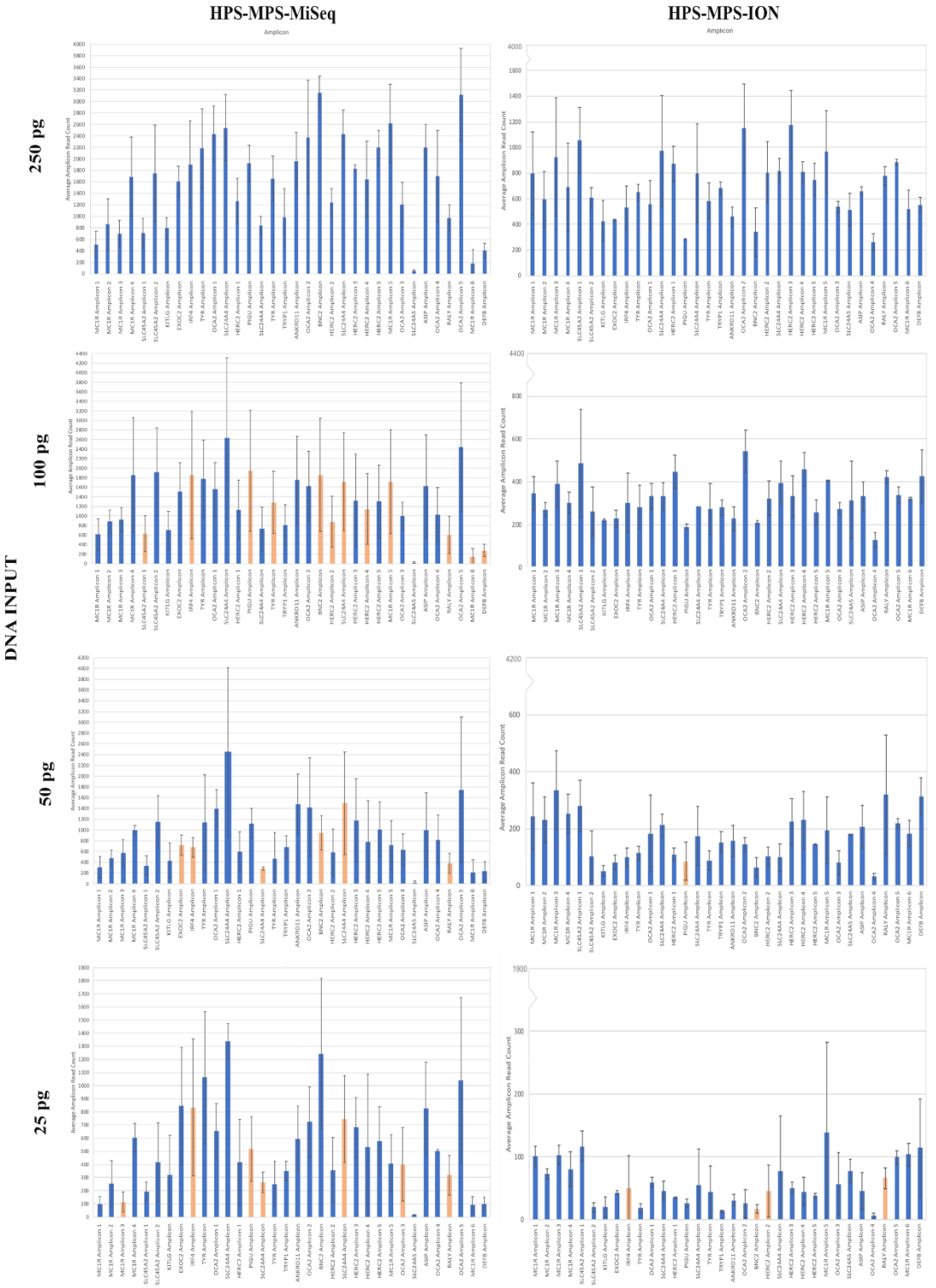
r genotype calls and read counts and finally a result folder contains all 96 samples in one upload file for use on the webtool prediction site. Notably, the HPS-MPS analysis pipeline is designed to work with any raw sequencing data including the HPS markers, not only the data generated from these targeted MPS assays. Thus, the pipeline can work with other targeted MPS data as well as with whole genome or capture sequencing data (once the HPS variant's region is covered), as it flips strands into the correct orientation for all 41 DNA variants for upload to the HIrisPlex-S webtool. For all assessments discussed below, both MPS assays utilized the same HPS-MPS analysis pipeline to generate the genotype calls and read counts.

Sensitivity testing and coverage consistency

In order to test the sensitivity of the two HPS-MPS assays on their respective MPS platforms, control DNA samples 9947A and 9948 were sequenced at DNA input concentrations of 5pg, 10pg, 25pg, 50pg, 100pg, 250pg, 500pg, and 1ng. For HPS-MPS-MiSeq, complete HPS profiles were observed for the 500pg and 250pg samples and for 3 of the 4 100pg samples (Figure 4 [Figure 2 in manuscript]). In the 100pg 9947A sample showing incomplete profiling, 12 amplicons were affected i.e., rs12203592 (*IRF4* amplicon), rs2378249 (*PIGU* amplicon), rs1393350 (*TYR* amplicon), rs10756819 (*BNC2* amplicon), rs2238289 (*HERC2* amplicon 2), rs1129038 (*HERC2* amplicon 4), rs17128291 (*SLC24A4* amplicon), rs1126809 (*MC1R* amplicon 5), rs3212355 (*MC1R* amplicon 6), rs1426654 (*SLC24A5* amplicon), rs6059655 (*RALY* amplicon), and rs8051733 (*DEF8* amplicon). At 50 pg input, drop out was seen at less loci, which included rs17128291 (*SLC24A4* amplicon) and rs6059655 (*RALY* amplicon) for sample 9947A and rs4959270 (*EXOC2* amplicon), rs12203592 (*IRF4* amplicon), rs12821256 (*KITLG* amplicon), and rs10756819 (*BNC2* amplicon) for sample 9948. Therefore based on these results, the sensitivity

threshold for the HPS-MPS-MiSeq assay is set to 250 pg. Further sensitivity testing using more DNA samples in varying concentrations shall be carried-out to clarify if drop-outs are consistently observed DNA inputs of 100 pg and below with this assay and platform, or not.

The same DNA samples in the same dilutions were tested with the HPS-MPS-ION assay, albeit not in duplicate. As seen in Figure 4 (Figure 2 in manuscript) below, complete HPS profiles were observed at 100pg DNA input in all samples tested. Drop-out started to occur at 50 pg DNA input, which affected one amplicon with one HPS DNA variant (rs683 in the *TYRP1* amplicon). At 25 pg input DNA, more drop out occurred at rs12203592 (*IRF4* amplicon) and rs2238289 (*HERC2* amplicon 2) for sample 9948, and rs2238289 (*HERC2* amplicon 2) and rs6059655 (*RALY* amplicon) for sample 9947A.



Control	9948	9947A	2800M
rs312262906	CC	CC	CC
rs11547464	GG	GG	GG
rs885479	GG	GG	GG
rs1805008	CC	CC	CC
rs1805005	GG	GG	GG
rs1805006	CC	CC	CC
rs1805007	CC	CC	CC
rs1805009	GG	GG	GG
rs201326893	CC	CC	CC
rs2228479	GG	GG	GG
rs1110400	TT	TT	TT
rs28777	AA	AA	AA
rs16891982	GG	GG	GG
rs12821256	CT	TT	TC
rs4959270	CA	CA	CA
rs12203592	CT	CT	CC
rs1042602	CC	CC	CA
rs1800407	CC	CC	CC
rs2402130	AA	AA	AA
rs12913832	GG	GG	AG
rs2378249	AA	GA	AA
rs12896399	GT	TT	GG
rs1393350	GG	AA	GG
rs683	AA	AA	CA
rs3114908	CC	CC	CT
rs1800414	TT	TT	TT
rs10756819	GA	AA	AA
rs2238289	AA	AA	AA
rs17128291	AG	AG	AA
rs6497292	AA	AA	AA
rs1129038	TT	TT	CT
rs1667394	TT	TT	CT
rs1126809	GG	AA	GG
rs1470608	GT	GG	GG
rs1426654	AA	AA	AA
rs6119471	CC	CC	CC
rs1545397	AA	AA	AA
rs6059655	GG	AG	GG
rs12441727	GA	GG	GG
rs3212355	CC	CC	CC
rs8051733	AA	AG	AG

Figure 4 (Figure 2 in manuscript) Sensitivity testing of both the HIrisPlex-S MPS tool with the HPS-MPS-MiSeq and the HPS-MPS-ION assays using control DNA samples 9947A and 9948 shown for the 34 amplicons used to sequence the 41 HIrisPlex-S DNA variants. Blue bars indicate correct calls in all samples analyzed at that DNA concentration, orange bars indicate an incorrect call was made in one sample at that concentration.

Supplementary Table 4 (Appendix D) provides a further breakdown of amplicon drop-out and sequencing error per HPS SNP variant per assay and platform-based on the two control DNA samples 9948 and 9947A at concentrations ranging from 250 pg – 25 pg DNA input. As an example, approx. 50% error indicates at least one sample had complete dropout for that DNA variant, closer to 100% indicates drop out for all samples at that site, and lastly approx. 25% would indicate at least one allele from that variant had dropped out for that sample. Apart from drop out of certain alleles at DNA input levels below the identified sensitivity threshold, percent error was broadly consistent between both assays and platforms. However, the HPS-MPS-ION assay had lower sequencing error per DNA variant than the HPS-MPS-MiSeq assay did e.g. 0.07% HPS-MPS-ION and 0.32% HPS-MPS-MiSeq at 250pg DNA input after pipeline application.

Overall, as seen in Figure 4 (Figure 2 in manuscript), the HPS-MPS-ION achieved more evenly distributed sequencing coverage across the amplicons and across DNA input concentrations compared to HPS-MPS-MiSeq analyses. However, the HPS-MPS-MiSeq assay displayed considerably higher read coverages (up to 3 times the reads at some amplicons) than the HPS-MPS-ION, where some amplicons had less than 100 reads, even at 250 pg DNA input. Figure 4 (Figure 2 in manuscript) also includes the genotype profiles consistently generated by both HPS-MPS assays (as well as the HPS SBE-CE assays) of control DNA samples 9947A, 9948, and 2800M (Promega, Madison, WI). 2800M was not assessed in this sensitivity study.

One of the reasons for the differences in performance observed with the two HPS-MPS assays may be due to the unequal number of DNA samples included in the respective singular sequencing runs. For this validation testing, 96 samples were sequenced from one cartridge for HPS-MPS-MiSeq, while for HPS-MPS-ION they were sequenced with two chips each running up to 48 samples in parallel. Reducing the number of samples in the MiSeq run may increase the

sensitivity and thus the likelihood of recovering a full genotypic profile at a lower DNA input amount than the sensitivity threshold obtained here indicates.

In an effort to measure the occurrence of drop out in a low input DNA sample and to provide a threshold for users of both HPS MPS assays, read counts were also assessed for all samples used in the sensitivity testing on whether the genotype was called correctly or if drop out occurred at that locus. Supplementary Table 5 (Appendix E) provides a confidence read count value for all 41 HPS sites for both MPS assays based on information from this validation. These confidence values reported average read counts as well as minimum read count values of the 100 pg DNA input samples (HPS-MPS-MiSeq N=4, HPS-MPS-ION N=2) used in sensitivity testing in which genotypes were accurately called by the genotyping pipeline. Supplementary Table 5 (Appendix E) also provides a recommended read count genotype confidence threshold average and minimum read count as a threshold set to ensure genotypes were called correctly between the 50 to 100 pg DNA input levels (based on 50 pg input DNA sensitivity samples HPS-MPS-MiSeq N=4, HPS-MPS-ION N=2). However, this is less than the obtained sensitivity threshold of both assays (which represents the minimal DNA input to achieve a result at all 41 DNA variants), so caution is recommended with these read count thresholds. If the read count threshold is not passed at the 50 pg minimum read count threshold, it is advised not to trust this genotype call and it should be reported as NA for upload to the prediction webtool. In order to semi-automate this process of threshold passes in a user-friendly manner, these read counts have been incorporated into the Threshold and Mixture Tool found in Supplementary Table 2 (Appendix B). A singular sample read count genotype confidence threshold for calls is not possible to determine for both HPS-MPS assays due to differing performance of the primers during the entire sequencing process and it is recommended to follow the thresholds per DNA variant instead as shown in these tables for both

workflows. Additional runs of 100 and 50 pg DNA input control and non-control samples may help further refine these read count threshold indicators, and these values can therefore be edited in the tool if the user wanted to define a more stringent threshold level. A guide on how to use this tool can be found in Supplementary Material 1 (Appendix K).

To assess coverage consistency of read counts for homozygote and heterozygote alleles for both HPS-MPS assays, several pre-selected individuals (N=8) with varying phenotype and genotype profiles, were analyzed in duplicate for a total DNA input of 500 pg per sample. Average read counts per allele were assessed for homozygote and heterozygote genotype calls using the HPS-MPS analysis pipeline and can be seen in Figure 5 (Figure 3 in manuscript) below.

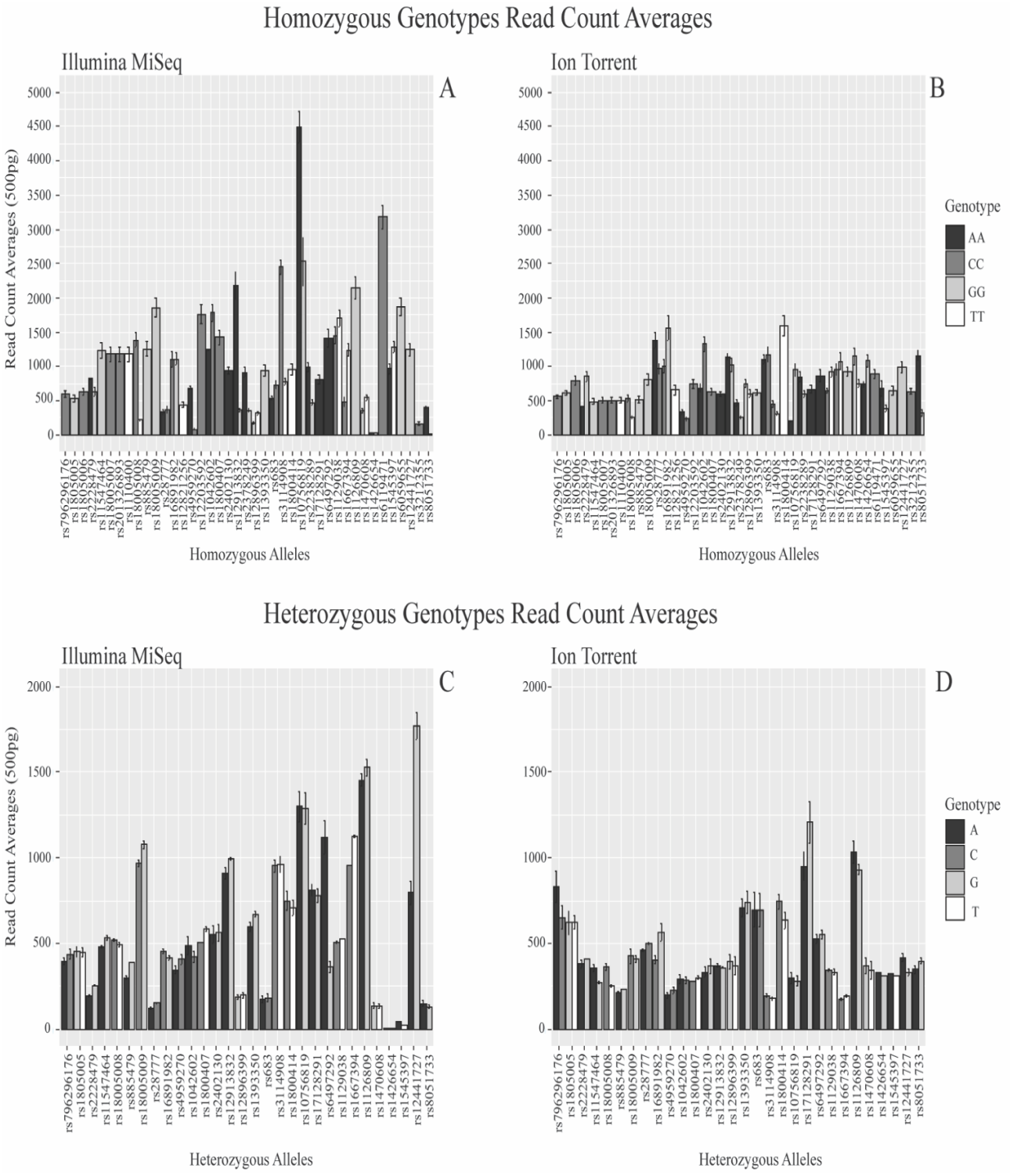


Figure 5 (Figure 3 in manuscript) Homozygote and heterozygote average peak heights from HIrisPlex-S MPS analysis with the HPS-MPS-MiSeq and HPS-MPS-ION assays

Overall, the HPS-MPS-MiSeq assay displayed slightly higher overall read counts for both homozygote and heterozygote genotypes per variant compared to HPS-MPS-ION (HPS-MPS-MiSeq: average 1039 reads homozygous, average 570 reads heterozygous calls, HPS-MPS-ION: average 750 reads homozygous, average 444 read counts heterozygous calls per variant). However, HPS-MPS-ION showed a more balanced profile with read counts more evenly distributed across the different amplicons. Notably, for HPS-MPS-MiSeq, DNA variants rs1426654 (*SLC24A5* amplicon) and rs1545397 (*OCA2* amplicon) displayed a much lower read count compared to the other DNA variants with less than 100 reads on average at 500 pg DNA input. The range in read counts at 500 pg DNA input for HPS-MPS-MiSeq was 14-4490 homozygous read counts, and 2-1771 heterozygous read counts using genotypes from a total of 16 profiles (8 samples in duplicate). For the HPS-MPS-ION this was 199-1590 homozygous read counts, and 176-1208 heterozygous read counts. Additional runs and further optimization of sample input for pooled library preparation, PCR cycle number, and/or primer input concentrations of the low coverage loci may correct the imbalances in amplicon amplification observed here at this 500 pg DNA input level.

Simulated casework

Nine simulated (mock) casework samples from six different individuals were made in duplicate from blood, semen, saliva, hair, and touch DNA sources and analyzed with the HPS-MPS-MiSeq and HPS-MPS-ION assays. The performance of both HPS-MPS assays, on the eight single source casework samples, including their concentration, are shown in Supplementary Table 6 (Appendix F). Both assays performed well analyzing samples with DNA inputs over 100 pg; in agreement with results from the sensitivity testing. The mock casework samples from saliva, dried and degraded saliva, semen, vaginal swab, hair, and dried blood samples, with DNA input ranging from 121 to 6890 pg, generated complete and correct 41-SNP HPS profiles (in comparison to

reference profiles generated by CE typing) with both MPS assays. Incorrect results due to allele drop out/in, were only seen for the two touch DNA samples analyzed, likely because of low level input DNA (~1 pg and 18 pg, respectively). The touch DNA samples prepared for this validation set were obtained from physical fingerprints swabbed from a glass slide immediately (fresh), and after 24 hours exposure on a bench top. The freshly prepared touch DNA sample (~ 1 pg DNA) showed incorrect results at 16 (39%) of the 41 HPS SNPs with the HPS-MPS-MiSeq assay and 19 (46%) SNPs with the HPS-MPS-ION assay, hence both assays had similar problems with this sample. The aged touch DNA sample (~ 18 pg DNA) revealed incorrect results for 6 (15%) of the 41 HPS SNPs with the HPS-MPS-MiSeq and for one SNP (2%) with the HPS-MPS-ION. The performance difference of both assays between fresh and aged touch DNA is likely explained by differences in DNA input amount collected from the fingerprint swab, rather than the time between touch and trace collection. Notably, both touch-DNA samples had input amounts well below the sensitivity threshold established for both assays, in contrast to all other mock casework samples used that were near or above the sensitivity threshold. The HPS-MPS-MiSeq assay displayed accurate genotypes (based on HP & HPS CE typing comparison) in the range between 100 pg and 250 pg DNA input for these mock case samples. This finding lends support to the idea that the singular duplicate sample in the sensitivity testing that showed dropouts at 100 pg input DNA may represent an outlier, and that the true sensitivity of the HPS-MPS-MiSeq assay may be closer to 100 pg (therefore matching the sensitivity of HPS-MPS-ION) rather than the 250 pg indicated by sensitivity testing. Overall, both HPS MPS assays were able to generate complete and accurate HIrisPlex-S results from all types of simulated casework scenarios tested, except from touched object samples with minute input DNA amounts that were well below the estimated sensitivity threshold of the assays.

Mixture sample testing and deconvolution tool

Mixture deconvolution is an active area of research [33] and several commercial tools [34, 35] have been developed to assist in mixture interpretation of forensic STR profiles. However, mixture separation tools for SNPs are currently lacking. As previously stated, it is very difficult to separate out mixtures when using CE-based SNP genotyping methods. Next to the increased multiplex capacity, this provided the other motivation to develop MPS-based SNP assays, which allows sequencing of the nucleotides surrounding the DNA variant and provides read count information per allele. Counting sequencing reads allows a quantitative assessment with advantages for mixture deconvolution, whereas peak height estimation using relative florescence units (RFU) generated from CE-based analysis is semi-quantitative. Although there are several criteria to detect a possible mixture, in particular, unusual read balances, at present there are no guidelines that can indicate a mixture using autosomal SNP data generated from MPS methods. Therefore, in order to test the mixture performance of both HPS-MPS assays in conjunction with the HPS-MPS analysis pipeline, a mixture calculator tool was designed to assist with 2-person mixture deconvolution designed specifically for the MPS platform and MPS assay used. This tool can be found in Supplementary Table 2 (Appendix B). The mixture tool works on the basis of the minimum read count thresholds as described in Supplementary Table 5 (Appendix E), and a ratio input calculation that separates read counts according to a Major: Minor ratio, within a 2-person mixture, all based upon the premise of an STR profile being available prior to the use of these FDP tools (i.e following common casework practice). By designing the mixture deconvolution tool around the input of a Major: Minor profile from STR data, the knowledge gleaned from heterozygote read counts per variant, and read counts for several 2-person mixture scenarios are generated within the tool for the user to decide which scenario their sample most closely resembles.

Also built into the tool is a range at which heterozygotes are called using read count information from the 500 pg input, as described above section 3.2. For example, not all heterozygote alleles are sequenced in a 50:50 ratio for single source samples; with some loci displaying a higher read count for a specific allele at a particular locus (see Supplementary Table 7 [Appendix G] for more details). Although this is a rather simple tool, it provides the basis for future tools to be more automated by using this process/guide as a starting point. A caveat to this tool in its current version is that not all heterozygotes were present in the available dataset. Therefore some HPS DNA variants such as rs1805006 *MC1R*, rs1805007 *MC1R*, rs201326893 *MC1R*, rs1110400 *MC1R*, rs12821256 *KITLG*, rs12203592 *IRF4*, rs2378249 *PIGU*, rs2238289 *HERC2*, rs6119471 *ASIP*, rs6059655 *RALY*, and rs3212355 *MC1R* do not have their heterozygote read count information incorporated into this tool at present. To overcome this data deficit, a conservative 45:55 standard deviation range is currently applied for these HPS DNA variants. Reference and alternate read counts at each site are compared to the various scenarios presented in the tool to determine the genotype profiles for the major and minor contributors to the sample and a ranking of the best scenario is generated with a value and a color code from green to red. The more green together with the lower the number, the more probable the scenario.

To assess the performance of this tool, mixture samples were sequenced with both HPS-MPS assays at mixture ratios of 1:1, 1:2, 1:5, and 1:10 (x2) for two separate sets of individuals (2 sets of 2 individual mixtures), to give 10 mixture types that were run in duplicate, total N = 20 per each MPS assay) with varying phenotypes and genotypes (see Supplementary Table 1 [Appendix A] for more specific details). A human evaluator was tasked with using the tool to infer the profiles of the contributors on a variant-by-variant basis. Other than knowledge on the ratios for each of the test mixtures, the human evaluator did not have the genotypes of the two individuals used in

the mixtures to compare, until the end of their assessment of the separated genotypes. All mixture contributors were quantified as being above the sensitivity thresholds for DNA input, therefore the chance of dropout was not accounted for in this assessment. Most scenarios (i.e. both major and minor profiles homozygous for reference allele, or major homozygote and minor heterozygote) and therefore the genotypes of the two individuals, could be separated by utilizing this tool. As the assessment was done on a variant-by-variant basis, these results are presented in Supplementary Table 8 (Appendix H). Overall, 28 of the 41 HPS SNPs could be fully separated into two individual profiles in 100% of samples across the 40 samples analyzed with both MPS assays. In the case of the other 13 variants, there were three DNA variants that resulted in inconsistent mixture separations (more than 20 errors or over half of the samples tested) leading to incorrect genotypes per person at rs1805005 *MC1R*, rs4959270 *EXOC2*, and rs2402130 *SLC24A4*. Incorrect genotype calls here signifies that he most probable scenario did not always reflect the actual prepared DNA scenario for these variants. This could be due to the fact that i) the pre-made sample did not actually reflect the true ratio for that variant (i.e. sample DNA was not exactly 1:10 with regards DNA input) or ii) that several scenarios may overlap when taking standard deviations of read count into account. The standard deviation for this tool was calculated based on the allelic imbalance observed per variant in its heterozygous state where there can be 5-15% read count variation in allele sequencing coverage (i.e. genotype GA called with 100 sequence depth, G allele called in the sequence 40 times, A allele called 60 times make it a 40:60 ratio, so for a 50:50 ratio for that heterozygote, a 10% read count deviation applies). Caution should be taken with these SNPs when utilizing this tool to access 2-person mixture profiles for input into the HIrisPlex-S prediction webtool. The final 10 variants showed a lower level of error in approximately 10-19 cases or 25-50% of total samples tested. Lastly, it is highly recommended to use the mixture tool as a guide

but still perform a manual check on how close the second and third scenarios are as some read counts may fall between two scenarios due to read count standard deviation. Additional genotyping of more individuals at these erroneous sites may provide a clearer heterozygote read count range to help refine the standard deviation generated for these sites for future developments of this deconvolution tool. It was also worth noting that the mixture samples prepared in 1:1 ratios were unable to be called using the tool alone. We recommend that in situations in which genotype read counts vastly differ from any given prediction scenario, or when ratio information still does not provide assistance, that a genome viewer such as Integrative Genomics Viewer (IGV) [36] is used to align and visualize the physical DNA strands to help assist in the resolution of the mixture data.

As an examination of an additional type of scenario that could be encountered during mixture interpretation of a sample, we tested the performance of the mixture tool without mixture ratio knowledge (i.e. no STR profile information to show minor:major ratio), using a simulated casework sample from mixed semen and vaginal material. The vaginal swab of unknown DNA concentration was dipped into semen aliquot of unknown concentration, and this sample was extracted for DNA and run through the HPS-MPS pipeline and mixture tool designed for both HPS-MPS assays. In order to successfully process this sample, the minor contributor ratio was adjusted to see if an appropriate scenario read count could be matched. Deconvolution of this mixture was possible for both HPS-MPS assays without prior ratio information once a 0.4 minor contributor ratio was input (1:2.5 ratio). It is worth noting that human examiner interpretation is still needed when making the final genotype decision, especially with the troublesome variants noted above. However, the use of this tool greatly aided mixture deconvolution on a variant-by-variant basis. In some scenarios it may not be possible to split the profile and therefore genotype options (i.e. report minor profile as being GA or GG with major being GG or GA) if separation is

not easily possible at that variant. To provide a simple visualization of how to assess a sample in terms of source (single or 2-person mixture) and read count threshold (clean calls or potential for allele drop out), a flowchart (Figure 6 [Figure 4 in manuscript]) has been designed that indicates what tools and tables to use to better understand how to deal with an unknown sample using both sequencing assays and systems. A more in-depth guide can also be found in Supplementary Material 1 (Appendix K).

Specificity and degradation testing

Five animal species were tested with the HPS-MPS-MiSeq and the HPS-MPS-ION assays. Samples included cat, dog, pig, mouse, and chimp (at DNA inputs of 1ng). The number of sequencing reads generated for the five species with both assays is shown in Supplementary Table 9 (Appendix I). Using the HPS-MPS-MiSeq assay, 31 (76 %) of the 41 DNA variants revealed sequencing reads in the cat, 34 (83 %) in pig, 40 (98 %) in the mouse, and 2 (1 %) in the dog, whereas the chimp produced a genotype profile of 39 (95 %) HIrisPlex DNA (as would be expected). With the HPS-MPS-ION assay, 21 (51 %) DNA variants yielded sequencing results in the cat, 20 (49 %) in the pig, 31 (76 %) in the mouse, 28 (68 %) in the dog, the chimp produced a full profile of all 41 HIrisPlex DNA variants. Amplification of particular HPS DNA variants can be explained by conserved genomic regions in these species; however, in such cases, the read counts obtained from non-human samples were typically much lower (> 100 times lower at some sites) and more fragmented (as seen using IGV software) than sequences typically generated from a human DNA sample at 1ng DNA input. On average, read count comparisons of the species to mean read count of the 500 pg human DNA input samples (from the coverage consistency section above) shows that the species gave a genotype call range from 49% - 98% (discounting chimp comparisons, which showed which showed much lower read counts of about 100 times lower).

Overall, the non-human species average read count was much lower than that expected for a 1 ng human DNA input if it were human DNA input. For HPS-MPS-MiSeq, the cat gave 2 times less in read count sample average, while the pig and the mouse gave 4 times less. For HPS-MPS-ION, the cat gave 100 times less in read count sample average, while the pig (10 times), the mouse (30 times) and the dog gave 8 times less. These observations coupled with the partial profiles generated may serve as a tool to help distinguish human and non-human samples when evaluating an unknown crime scene DNA sample. However, since FDP would be typically performed on crime scene DNA samples after STR profiling, human DNA is already detected in each case.

To prepare samples that would test the effect of DNA degradation on the performance of the in-house designed HPS-MPS-MiSeq assay, aliquots of a single source 500 pg DNA input sample was subjected to Ultraviolet (UV) radiation for 0 seconds, 5 minutes, 10 minutes, and 20 minutes. Even after 10 minutes of UV light exposure, a complete 41-SNP HPS profile was achieved with an average coverage of 2040 reads. After 20 minutes of UV light exposure 5 SNPs: rs28777 *SLC45A2*, rs4959270 *EXOC2*, rs12896399 *SLC24A4*, rs1426654 *SLC24A5,* and rs3212355 *MC1R* displayed drop out due to suspected degradation. These results indicate the robustness of HPS-MPS-MiSeq assay to cope with environmental degradation. Degradation information with read counts is given in Supplementary Table 10 (Appendix J). For HPS-MPS-ION, degradation testing was not performed on these artificially degraded DNA samples. However, preliminary evidence of this assay's ability to deal with naturally degraded DNA comes from the analysis of a series of DNA samples extracted from bones that spent approximately 1 to 78 years in soil, where HPS-MPS-ION efficiency was found to be comparable to that of the GlobalFiler PCR amplification kit on the same samples Although, it should be noted that the maximal DNA amount used to analyze STRs was 15 µl whereas only 6 µl was used for HPS-MPS-ION which

made a significant difference in the weak samples. Full HPS profiles were obtained from as little as 50 pg of DNA with 200 reads coverage threshold. However, performance of three SNPs: rs1545397 and rs1470608 in *OCA2* and rs10756819 in *BNC2* was slightly weaker compared to other markers (W. Branicki, personal communication).

Concordance testing

In all, five partner laboratories, with varying MPS experience and complementary to US and Rotterdam, were involved in the concordance testing of the two HPS-MPS assays, 3 for HPS-MPS-ION and 2 for HPS-MPS-MiSeq. During the initial phase of the concordance testing it became evident that there was a need for guidelines for HPS-MPS data interpretation with regards read count thresholds and data assessment for single source versus mixture interpretation. Therefore, such interpretation guidelines were designed to assist the HPS-MPS assay users in genotype calling and mixture separation using the output from the HPS-MPS analysis pipeline. Concordance testers were asked to generate data on unknown samples (N=16) that ranged in concentration from 6 pg to 25.4 ng DNA input, thus including samples that were below the DNA input thresholds established in the sensitivity testing of these two assays (see Supplementary Table 3 [Appendix C] for more details). They were also tasked with running the raw FastQ sequence files, output by the sequencers, through the HPS-MPS analysis pipeline to generate the necessary genotype calls and read count information files (for more information on the pipeline and what is generated please see Supplementary Material 1 [Appendix K]) using their respective machines/assays. Lastly, concordance testers were asked to use the threshold and mixture tool they were provided with (Supplementary Table 2 [Appendix B]) to generate each samples genotype interpretation. Figure 6 (Figure 4 in manuscript) below provides an outline of how best to deal

with single/mixture sources, and was the guideline given to the concordance sites. Concordance testers used this approach for each of their sample result files, and summarized their interpretation results in a single file where it was compared with the main US and Rotterdam development laboratory results.

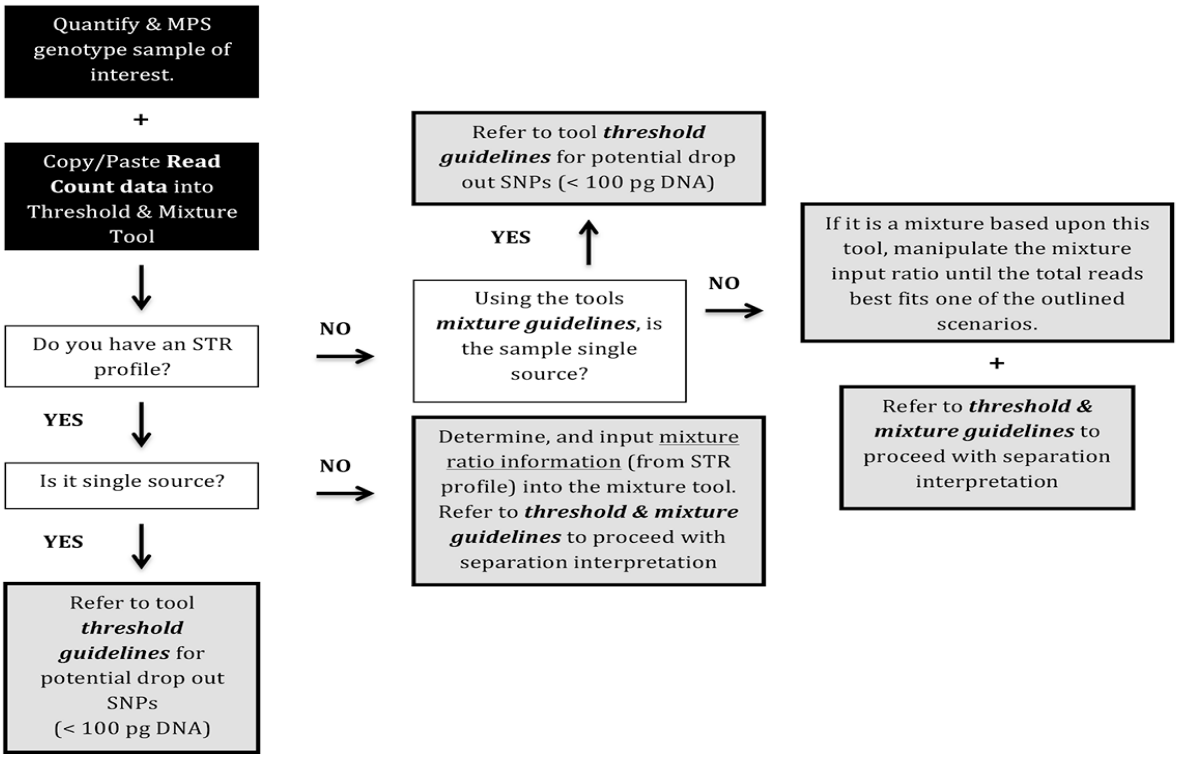


Figure 6 (Figure 4 in manuscript) Interpretation flowchart for the HIrisPlex-S MPS pipeline

The results of the concordance study can be found in Supplementary Table 3 (Appendix C), where the source type and concentration of the sample are given, together with each site's correct interpretation calls (number and %) for that sample (Supplementary Table 2 [Appendix B]). This table displays % agreement between the assay developing site and the respective concordance sites (US for the HPS-MPS-MiSeq assay versus the two MiSeq concordance sites, and Rotterdam for the HPS-MPS-ION assay versus the three Ion Torrent concordance sites). The

criteria used to report the final genotype per variant according to the minimum thresholds (per assay/machine) and final genotype calls scenario (including those used in mixture scenario separations) are described in Supplementary Table 2 (Appendix B). The top portion of the result table reflects only the correct (number and %) genotype calls generated by the scenarios (single vs mixed) but does not take into account the minimum threshold needed to call a variants genotype with confidence. As can be seen from the top portion of the table, HPS-MPS-MiSeq generated the HPS SNP genotypes for each locus, ranging in genotype success from 49% - 100% (> 100 pg DNA input average genotype success is 84%) compared to 56% -100% (> 100 pg DNA input average genotype success is 92%) for HPS-MPS-ION. The bottom portion of the table shows that HPS-MPS-MiSeq did not perform as well as HPS-MPS-ION when factoring in the minimum read count threshold (which can also be found in Supplementary Table 5 (Appendix E) under the < 50 pg level DNA input column). This threshold is required to pass the genotyping call confidence criteria and rules of interpretation proposed by this study and as described in Supplementary Material 2 (Appendix L) guide. The HPS-MPS-MiSeq concordance result interpretation assessment ranged from 12%-85% result agreement (>100 pg DNA input average interpretation success is 33%) with the reference data obtained at the US assay developing site, while for HPS-MPS-ION it ranged from 61% - 100% result agreement (> 100 pg DNA input average interpretation success is 88%) with the reference data generated at the Rotterdam assay developing site. Note that, '0%' for some samples indicates that no data passed the threshold minimum read count for that sample at that laboratory during the interpretation assessment.

Overall, the HPS-MPS-ION assay performed well in this concordance testing in both assessments with on average 89% agreement with the reference data. The HPS-MPS-MiSeq assay underperformed with on average reference data agreement from both assessments only being on

average 58%. Due to the in-house design and the multiple steps needed in the library preparation of the MiSeq assay, it is possible that primer degradation occurred (especially with regards the small indexing primers needed for an integral step in the process of the MiSeq library preparation) that affected the significant decrease in read counts generated at the MiSeq concordance sites. Overall, average read counts for the same 100 pg DNA input sample (9947A standard control) run by the reference US site (which has vast experience in running this particular in-house design) were approximately double (1185) the average read counts of the two concordance sites (859 and 577 respectively). This lends support to the possibility of HPS-MPS-MiSeq primer and/or sample degradation during material shipment to the concordance testers.

2.5 Conclusions

This study introduced and forensically validated MPS assays for the HIrisPlex-S system for eye, hair, and skin color prediction from DNA for the two MPS platforms commonly used in forensic genetics. We demonstrate that both HPS-MPS assays perform reasonably well on the respective MPS platforms they were developed for. The better performance of the HPS-MPS-ION assay may be explained by the use of the commercial AmpliSeq design compared to an in-house design of the HPS-MPS-MiSeq assay, which may be overcome in the future by applying the AmpliSeq design to MiSeq. Although, both HPS-MPS assays appeared less sensitive than the previously reported two SNaPshot assays of the HIrisPlex-S system, due to the fact that the MPS assays simultaneously analyze all 41 DNA variants, less total DNA is needed. Moreover, the HPS-MPS assays provide advantages in mixture interpretation compared to the previous SNaPshot assays. The semi-automated HPS-MPS analysis pipeline and the HPS-MPS mixture analysis tool introduced here together with the HPS-MPS assays will benefit future application of HPS-MPS analysis. We envision the MPS implementation of the HIrisPlex-S system for eye, hair and skin

color prediction from DNA described here, to be the starting point for expanding MPS-based forensic DNA phenotyping. This expansion is expected to include the addition of SNPs predictive for more externally visible traits, as well as SNPs suitable for bio-geographic ancestry inference, provided such predictive SNPs are identified and suitable statistical prediction models are developed.

2.6 Acknowledgements

We thank all participants of this study. The US site was supported in part by the US National Institute of Justice (NIJ) under grant number 2014-DN-BX-K031, and the US Department of Defense (DOD) DURIP-66843LSRIP-2015. The funding organizations did not have any influence on the design, conduct or conclusions of the study.

2.8 References

[1] K.K. Kidd, W.C. Speed, A.J. Pakstis, M.R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F.R. Friedlaender, J.R. Kidd. Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci. Int. Genet. 10 (2014) 23-32.

[2] F. Oldoni, K.K. Kidd, D. Podini. Microhaplotypes in forensic genetics. Forensic Sci. Int. Genet. 38 (2019) 54-69.

[3] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, A. Parys-Proszek, Ż. Makowska, A. Pałeczka, K. Kucharczyk, R. Płoski, W. Branicki. Development of a forensically useful age prediction method based on DNA methylation analysis. Forensic Sci. Int. Genet. 17 (2015) 173-179.

[4] A. Aliferi, D. Ballard, M.D. Gallidabino, H. Thurtle, L. Barron, D. Syndercombe Court. DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. Forensic Sci. Int. Genet. 37 (2018) 215-226.

[5] M. Kayser. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci. Int. Genet. 18 (2015) 33-48.

[6] M. Kayser, P. de Knijff. Improving human forensics through advances in genetics, genomics and molecular biology. Nature Reviews Genetics 12 (2011) 179.

[7] A. Ralf, M. van Oven, K. Zhong, M. Kayser. Simultaneous Analysis of Hundreds of Y-Chromosomal SNPs for High-Resolution Paternal Lineage Classification using Targeted Semiconductor Sequencing. Human Mutation 36(1) (2015) 151-159.

[8] J. Draus-Barini, S. Walsh, E. Pośpiech, T. Kupiec, H. Głąb, W. Branicki, M. Kayser. Bona fide colour: DNA prediction of human eye and hair colour from ancient and contemporary skeletal remains. Investigative Genetics 4(1) (2013) 3.

[9] T.E. King, G.G. Fortes, P. Balaresque, M.G. Thomas, D. Balding, P.M. Delser, R. Neumann, W. Parson, M. Knapp, S. Walsh, L. Tonasso, J. Holt, M. Kayser, J. Appleby, P. Forster, D. Ekserdjian, M. Hofreiter, K. Schürer. Identification of the remains of King Richard III. Nature Communications 5 (2014) 5631.

[10] S. Walsh, L. Chaitanya, K. Breslin, C. Muralidharan, A. Bronikowska, E. Pospiech, J. Koller, L. Kovatsi, A. Wollstein, W. Branicki, F. Liu, M. Kayser. Global skin colour prediction from DNA. Human genetics 136(7) (2017) 847-863.

[11] S. Walsh, L. Chaitanya, L. Clarisse, L. Wirken, J. Draus-Barini, L. Kovatsi, H. Maeda, T. Ishikawa, T. Sijen, P. de Knijff, W. Branicki, F. Liu, M. Kayser. Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. Forensic Sci. Int. Genet. 9 (2014) 150-161.

[12] S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, M. Kayser. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci. Int. Genet. 7(1) (2013) 98-115.

[13] S. Walsh, A. Lindenbergh, S.B. Zuniga, T. Sijen, P. de Knijff, M. Kayser, K.N. Ballantyne. Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence. Forensic Sci. Int. Genet. 5(5) (2011) 464-471.

[14] L. Chaitanya, K. Breslin, S. Zuñiga, L. Wirken, E. Pośpiech, M. Kukla-Bartoszek, T. Sijen, P.d. Knijff, F. Liu, W. Branicki, M. Kayser, S. Walsh. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Sci. Int. Genet. 35 (2018) 123-135.

[15] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser. IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci. Int. Genet. 5(3) (2011) 170-180.

[16] E. Pośpiech, Y. Chen, M. Kukla-Bartoszek, K. Breslin, A. Aliferi, J.D. Andersen, D. Ballard, L. Chaitanya, A. Freire-Aradas, K.J. van der Gaag, L. Girón-Santamaría, T.E. Gross, M. Gysi, G. Huber, A. Mosquera-Miguel, C. Muralidharan, M. Skowron, Á. Carracedo, C. Haas, N. Morling, W. Parson, C. Phillips, P.M. Schneider, T. Sijen, D. Syndercombe-Court, M. Vennemann, S. Wu, S. Xu, L. Jin, S. Wang, G. Zhu, N.G. Martin, S.E. Medland, W. Branicki, S. Walsh, F. Liu, M. Kayser. Towards broadening Forensic DNA Phenotyping beyond pigmentation: Improving the prediction of head hair shape from DNA. Forensic Sci. Int. Genet. 37 (2018) 241-251.

[17] F. Liu, Y. Chen, G. Zhu, P.G. Hysi, S. Wu, K. Adhikari, K. Breslin, E. Pospiech, M.A. Hamer, F. Peng, C. Muralidharan, V. Acuna-Alonzo, S. Canizales-Quinteros, G. Bedoya, C. Gallo, G. Poletti, F. Rothhammer, M.C. Bortolini, R. Gonzalez-Jose, C. Zeng, S. Xu, L. Jin, A.G. Uitterlinden, M.A. Ikram, C.M. van Duijn, T. Nijsten, S. Walsh, W. Branicki, S. Wang, A. Ruiz-Linares, T.D. Spector, N.G. Martin, S.E. Medland, M. Kayser. Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair. Human molecular genetics 27(3) (2018) 559-575.

[18] M. Marcińska, E. Pośpiech, S. Abidi, J.D. Andersen, M. van den Berge, Á. Carracedo, M. Eduardoff, A. Marczakiewicz-Lustig, N. Morling, T. Sijen, M. Skowron, J. Söchtig, D. Syndercombe-Court, N. Weiler, E.-N.C. The, P.M. Schneider, D. Ballard, C. Børsting, W. Parson, C. Phillips, W. Branicki. Evaluation of DNA Variants Associated with Androgenetic Alopecia and Their Potential to Predict Male Pattern Baldness. PLOS ONE 10(5) (2015) e0127852.

[19] F. Liu, M.A. Hamer, S. Heilmann, C. Herold, S. Moebus, A. Hofman, A.G. Uitterlinden, M.M. Nöthen, C.M. van Duijn, T.E. Nijsten, M. Kayser. Prediction of male-pattern baldness from genotypes. European journal of human genetics : EJHG 24(6) (2016) 895-902.

[20] K.A. Meiklejohn, J.M. Robertson. Evaluation of the Precision ID Identity Panel for the Ion Torrent™ PGM™ sequencer. Forensic Sci. Int. Genet. 31 (2017) 48-56.

[21] C. Strobl, M. Eduardoff, M.M. Bus, M. Allen, W. Parson. Evaluation of the precision ID whole MtDNA genome panel for forensic analyses. Forensic Sci. Int. Genet. 35 (2018) 21-25.

[22] C. Xavier, W. Parson. Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit – MPS forensic application for the MiSeq FGx™ benchtop sequencer. Forensic Sci. Int. Genet. 28 (2017) 188-194.

[23] L. Chaitanya, A. Ralf, M. van Oven, T. Kupiec, J. Chang, R. Lagacé, M. Kayser. Simultaneous Whole Mitochondrial Genome Sequencing with Short Overlapping Amplicons Suitable for Degraded DNA Using the Ion Torrent Personal Genome Machine. Human Mutation 36(12) (2015) 1236-1247.

[24] G.M. Pineda, A.H. Montgomery, R. Thompson, B. Indest, M. Carroll, S.K. Sinha. Development and validation of InnoQuant™, a sensitive human DNA quantitation and degradation assessment method for forensic samples using high copy number mobile elements Alu and SVA. Forensic Sci. Int. Genet. 13 (2014) 224-235.

[25] I.F. Bronner, M.A. Quail, D.J. Turner, H. Swerdlow. Improved Protocols for Illumina Sequencing. Current Protocols in Human Genetics 79(1) (2013) 18.2.1-18.2.42.

[26] A. Untergasser, H. Nijveen, X. Rao, T. Bisseling, R. Geurts, J.A.M. Leunissen. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Research 35(suppl_2) (2007) W71-W74.

[27] T. Arányi, G.E. Tusnády. BiSearch: ePCR tool for native or bisulfite-treated genomic template. Methods Mol Biol 402 (2007) 385-402.
[28] P.M. Vallone, J.M. Butler. AutoDimer: a screening tool for primer-dimer and hairpin structures. BioTechniques 37(2) (2004) 226-231.
[29] H. Li, R. Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14) (2009) 1754-1760.
[30] H. Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21) (2011) 2987-2993.
[31] D.C. Koboldt, Q. Zhang, D.E. Larson, D. Shen, M.D. McLellan, L. Lin, C.A. Miller, E.R. Mardis, L. Ding, R.K. Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research 22(3) (2012) 568-576.
[32] The R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013
[33] D. Taylor, J.-A. Bright, J. Buckleton. The interpretation of single source and mixed DNA profiles. Forensic Sci. Int. Genet. 7(5) (2013) 516-528.
[34] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B.W. Duceman. Validating TrueAllele® DNA Mixture Interpretation. Journal of Forensic Sciences 56(6) (2011) 1430-1447.
[35] J.A. Bright, D. Taylor, C. McGovern, S. Cooper, L. Russell, D. Abarno, J. Buckleton. Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. Forensic Sci. Int. Genet. 23 (2016) 226-239.
[36] H. Thorvaldsdóttir, J.T. Robinson, J.P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14(2) (2012) 178-192.

CHAPTER 3. PROTEOMIC ANALYSIS OF HAIR

3.1 Introduction

Proteomics incorporates biological and chemical processes into one effective technology for the identification and quantitation of proteins present in a sample [1, 2]. These detected proteins have the potential to provide DNA sequence information (with particular variant genotypes) obtained through the detection of amino acid changes [3]. These amino acid changes can ultimately be used to identify an individual via comparable profiling, i.e. by matching the unknown biological sample to either a known individual/suspect or potential database match [4]. For this particular research, we were interested in detecting proteins obtained from hair samples due to the ease at which hair may be shed and therefore available at a crime scene. To note, a hair bulb is not required for this laboratory approach, approximately one inch of the hair shaft will generate enough protein material to see potential amino acid changes. Due to the presence of core proteins found within hair follicles [5] (mainly the Keratin class of structural proteins), we focused on a set of 233 candidate variants commonly found in hair follicles. These variants are termed Genetically Variant Peptides (GVPs) and they have the potential to be classified as a type of identity marker, if assessed using quality control metrics such as independence and Hardy Weinberg Equilibria, and a suitable population frequency database were available. Current concepts and methods in forensic science are centered around the ability to identify humans from DNA. However, when DNA extracted from a sample is too degraded or too low in concentration, proteomic methods such as the one described above may be the solution for identification in these difficult scenarios. Due to this, it was necessary to assess the 233 candidate GVPs using population genetics approaches to identify a robust set of SNPs that may be best suited for identification purposes from hair shaft material.

3.2 Materials and Methods

Hair Sample Collection and Subset Selection Criteria

Individuals that contributed towards this study (providing saliva and hair samples amongst other questionnaire data) were collected in compliance with IRB#1409306349 and included informed consent for all participating individuals. Each individual was assigned a randomized study ID number and 5ml of saliva was collected. Each participant also filled out a questionnaire containing numerous questions about phenotypic characteristics (ie. hair type: straight, wavy, curly) and familial heritage along with having three hairs removed from the back of their head and placed in a tube containing RNA*later*^{TM} solution (Invitrogen^{TM}, Carlsbad, CA).

Saliva samples from these individuals were sequenced at the University of Chicago Genomics Center using the Muti-Ethnic Genotyping Array (MEGA) by Illumina (Illumina Inc.). Approximately 1.7 million SNPs were identified for each sample and computationally processed in GenomeStudio Software (Illumina Inc.). The SNPs were then exported into PLINK [6] format for quality control (QC) procedures to be performed. The QC procedures include filtering for minor allele frequency, SNP genotyping efficiency, individual genotyping efficiency, relatedness, sex discrepancies, and Hardy-Weinberg Equilibrium. Individuals or samples not passing these QC protocols and thresholds were excluded from the dataset. After all the stringent QC metrics were performed, the sample dataset consisted of approximately 3,300 individuals, termed the Walsh Lab database, and were represented by approximately 330,000 variants for the next analyses.

The final set of hair samples chosen (N = 99) from the 3,300 set of individuals were selected for proteomic analysis based on a principle component analysis (PCA) [7] to ensure there was a broad distribution of populations provided for analyses. For this PCA plot, individuals were plotted along with 2,504 individuals from phase 3 of the 1,000 Genomes Project [8], and 940 individuals from Human Genome Diversity Project (HGDP) [9] which represented a

total of 78 populations. The 1,000 Genomes Project and HGDP contains individuals with known population information and were therefore used as a reference set in the PCA plot to view genetic distances between population groups, and therefore between individuals, allowing for the visualization of the population distribution of these datasets. This also allowed us to view important within population variation information on the individuals chosen for proteomic analysis (ie. Northern vs. Southern Europeans).

Hair Sample Preparation for Proteomic Analysis

After the 99 individuals were selected from the PCA plot, hair samples were prepared and sent to collaborators at the University of California Davis. For this, one hair from each sample collection tube was removed using sterile tweezers, and the root of the hair was detached (with sterile medical scissors) and placed back into the original sample tube containing RNA*later*^{TM} solution (Invitrogen^{TM}, Carlsbad, CA). The remaining hair shaft was placed into a new 1.5 mL microcentrifuge tube, labeled with the corresponding sample number, and approximately 1000 μl of deionized water was added to completely cover and submerge the hair. Each hair was between 2-4 cm in length to ensure a successful amount of protein would be detected during proteomic analysis. The proteomic analysis of the 99 hair shaft samples was completed in collaboration with faculty and students from Glendon Parker's laboratory at the University of California Davis according to methods discussed in research published by Milan *et al* [3]. Please see this paper for more specific details on sample processing.

Population Genetics Analysis on 233 Candidate GVP Set

Population genetics statistics were generated using data from the five main regions of the 1000 Genomes Project: Europe, Africa, America, South Asia, and East Asia for each of the candidate variants. All population genetics statistics (i.e. HWE, LD, FREQ) were generated using PLINK [6]. PLINK is an open-source whole genome association analysis toolset that utilizes computational command line to efficiently compute large-scale analyses of genotype/phenotype data. To generate the minor allele frequency values for each SNP for each region, the --freq command was used. Hardy-Weinberg equilibrium values for each region were generated using the --hardy command. Finally, linkage disequilibrium r^2 values for the entire SNP set were generated using the --r2 command. In addition, to identify potential ancestry informative markers (AIMs) within the total set of 233 variants, Fst values were calculated in a pairwise fashion for the five main regions using the --fst command [6].

Assessment of Proteomic Analysis Method for Inferred Genotyping

The hairs were digested and analyzed using methods developed by collaborators at the University of California Davis. 233 SNPs across 127 genes were chosen as potential candidate markers for their possible effect on hair structure. Out of the 99 total hair samples sent to the University of California Davis, only 66 have been fully analyzed due to the extensive time it takes to process the samples. Our collaborators intend to send results for the remaining 33 samples when they are finished being processed. A description of the 99 total samples used for proteomic analysis can be seen in Appendix M.

For the proteomic results for the 66 analyzed hair samples, GVPs that were detected were assigned a value of "1" at each specific SNP allele for each gene. Non-synonymous SNP alleles,

and therefore genotypes, were inferred for each sample from this GVP detection notation. For example, if a GVP was detected at rs1234 allele A and rs1234 allele G, the inferred genotype for the individual would be heterozygous for GA at rs1234. However, if a GVP was only detected at rs1234 allele G, then the inferred genotype would be homozygous GG instead.

To assess the accuracy of the Liquid Chromatography tandem Mass Spectrometry (LC/MS-MS) proteomic analysis method, it was necessary to compare the inferred genotypes to the actual genotypes of each variant for the subset of individuals used. Genotypic data for the 66-sample set was extracted from SNPs included in the MEGA array. The total set of 233 SNPs used in the proteomic analysis was crossed referenced with the SNPs included in the MEGA array and a total of 114 SNPs overlapped. The 114 SNPs were then assessed based on a GVP detection threshold in order to be determined useful in measuring the accuracy of the laboratory proteomics method and therefore generating correct genotype information. For this threshold, SNPs had to show GVP detection rates of at least 90%. Therefore, if a GVP was not detected in >10% of samples, the SNP was not included in the assessment. Based on of these detection thresholds, a final set of 32 SNPs were used for further accuracy analyses.

Hair Structure Correlation Assessment

In addition to rating the GVPs success from a proteomics perspective for identity and/or ancestry inference, due to our phenotyping research, we also attempted to check if one of these GVP variants may be a potential candidate SNP for hair structure prediction. To do this, it was necessary to assess the correlation and its significance between an individual's questionnaire-based hair phenotype (straight, wavy, and curly) and genotype for these 114 variants with reliable genotypes (from the MEGA SNP array). For the first correlation assessment, the previously

mentioned proteomic sample set of 66 individuals and their genotypic data was used. For the second correlation assessment, a European-only dataset from the Walsh lab database was used for greater power. The assessments directly focused on the association between hair phenotype and genotype while controlling for the population/ancestry, sex, and age of the individuals.

Hair structure phenotype information for all samples involved in both correlation assessments was obtained through the previously mentioned questionnaire filled out by all participants. During the correlation assessments, the phenotypes were coded on a continuous scale of 1, 2, and 3, representing straight, wavy and curly, respectively. After this assessment, a binary correlation was then performed for each hair structure type, by re-coding each phenotype as a 1 or 0, depending on the phenotype in question. For example, during the correlation assessment for curly hair, all individuals with curly hair phenotype were re-coded as "1", while the remaining individuals with wavy or straight hair were re-coded as "0." This pattern was followed for wavy and straight hair correlation assessments.

Genotypic data was re-coded in both correlation assessments as 0, 1, and 2 (ex. homozygote for allele A, heterozygote, homozygote for allele B) per SNP. Age and sex of each sample in both correlation assessments was also extracted from questionnaire data received during the Walsh lab study. Age of each individual did not need recoding, as it is already numerical, however; the sex of each individual was recoded as 1 for male and 2 for female for both correlation assessments. Ancestry correction was used in the first correlation assessment to group samples into five main populations: African, European, Asian, Admixed, and Middle Eastern where they were recoded as 1, 2, 3, 4, and 5 for each individual respectively. For the second correlation, European individuals were selected from the 3,300 database set, leaving 1,821 European individuals to assess.

Each correlation assessment was performed using the R program [10] and the command "pcor.test" from the ppcor package [11]. The r^2 correlation coefficient and its significance (p-value) of the phenotype: genotype correlation was generated and output per SNP. The covariates for this analysis were age, sex, and ancestry. A Bonferroni correction value was also generated using the R command "p.adjust" and "BH" was used as the method.

3.3 Results and Discussion

Assessment of Proteomic Method for Inferring Genotypes from GVP Information

A result file was sent from our collaborators at UC Davis. It included information on gene and SNP name and their corresponding SAAP and peptide sequences. They coded the data based on the detection of the GVP according to their method metrics. As previously stated, their proteomic method and protocol can be seen in research publish by Milan *et al.* [3]. Upon receiving these proteomic analysis results, it was necessary to evaluate the accuracy of the GVP detection and genotype inference method by comparing to genotypes that were generated from the MEGA SNP array (N=114 overlapping). The current proteomic method was unable to detect GVPs for some of the SNPs. Therefore, in order to calculate the accuracy of the method, it was important to identify the SNPs with the highest rate of detection first. Any SNPs that GVPs were not detected in >10% of the sample set (N=66) were eliminated from the list. This list of SNPs was then separated based on SNPs with known genotypic information and SNPs without this genotype confirmation. Finally, the remaining markers were assessed for population heterozygosity and statistical independence by computing Hardy-Weinberg equilibrium and linkage disequilibrium (LD) statistics (discussed in the next section). The final set of markers passed these metrics (in HWE, not in LD) and was composed of 21 SNPs with genotypic confirmed data and 11 SNPs without genotypic confirmed data. The final set of markers and their GVP missingness (% of GVPs

not detected) and genotype accuracy calculations can be seen in Table 2 below. The set of 21 SNPs will now be referred to as GVP21 and the complete set of all proteomically detected markers regardless of genotypic data (21 SNPs plus 11 SNPs) will be referred to as GVPComplete. Genotype frequencies for GVPComplete for the 66 sample dataset can be seen in Appendix N to give a general idea of overall genotype distributions.

Table 2 Genotype Accuracy and GVP Missingness for GVPComplete. The SNPs highlighted in orange are included in GVP21. The remaining SNPs are the 11 SNPs without genotypic confirmed data.

CHROM	POS	SNP	REF	ALT	% GVP Missingness	Genotype % Accuracy
1	153431406	rs41265164	G	A	2%	72%
1	153520203	rs116208483	G	C	0%	100%
1	153520954	rs62624468	C	T	8%	89%
1	201289487	rs61818256	C	T	8%	89%
6	74014637	rs28763966	C	A	0%	100%
6	7581001	rs28763967	C	T	0%	100%
12	52788945	rs1791634	C	G	0%	98%
12	53069014	rs17678945	C	A	10%	84%
14	113975768	rs10148371	G	A	0%	100%
14	55609418	rs11125	A	T	0%	100%
17	38859509	rs7213256	C	T	2%	86%
17	39116603	rs17843021	G	A	3%	83%
17	39116728	rs142154718	C	T	0%	100%
17	39183254	rs62623375	C	T	5%	92%
17	39593768	rs2604953	G	T	9%	77%
17	39633354	rs138303882	G	A	0%	97%
17	39635733	rs743686	A	G	0%	92%
17	39913771	rs41283425	C	T	0%	92%
18	28605818	rs79011243	C	A	3%	92%
21	31744310	rs9636845	A	T	2%	86%
21	32253513	rs71321355	C	T	8%	75%
17	39913771	rs143043662	C	T	0%	—
17	39619115	rs2071563	G	A	0%	—
17	39620641	rs146792525	C	T	0%	—
12	52788928	rs2658658	G	A	9%	—
12	52788945	rs1732263	C	G	3%	—
12	52713088	rs2857663	G	A	0%	—
17	39183313	rs148449559	G	C	3%	—
17	39156084	rs9897046	T	C	6%	—
17	39334241	rs62067292	G	C	3%	—
17	39323971	rs428371	G	A	5%	—
21	46117792	rs34302939	G	A	8%	—

GVP21 was used to compute genotyping accuracy statistics due to the ability to compare the inferred genotypes from the detected GVPs to the actual genotypes (SNP array data) of the 66 samples. Genotyping accuracy varied per SNP and ranged from 72% - 100%, with an average genotyping accuracy of 91%. The genotyping accuracy was lower (72% - 77%) for the SNPs that displayed higher levels of heterozygosity within the sample set (rs41265164, rs71321355, rs2604953) and higher (100%) for the SNPs that were lower in heterozygosity (rs116208483, rs28763966, rs28763967, rs10148371, rs11125, rs142154718). Therefore, although most maintained an average genotyping accuracy, the distribution of these accuracies suggests that the current method of GVP detection may be missing out on important genotypic data. The true accuracy of the inferred genotyping proteomic method was unable to be calculated for the set of 11 additional SNPs in GVPComplete because genotypic data was unavailable. It is important to perform a genotyping accuracy assessment in the future to confidently include them into the optimal GVPComplete identity set.

Probability of Identity (PID) Using an Optimal Set of Hair GVPs – GVP21 & GVPComplete

In order to use the product rule to generate a match probability statistic for a sample, it is vital that the variant pass both HWE and LD assessments. Variants should be in HWE and not in LD with any other variant in the identity set. All SNPs in GVP21 passed Hardy-Weinberg equilibrium tests, however; two of the SNPs were in linkage disequilibrium with other SNPs in the set (rs17843021 and rs62623375). Based on this, only rs17843021 and rs62623375 were used in the calculations, while the SNPs that were in LD with them were not. The 11 additional SNPs in GVPComplete passed Hardy-Weinberg equilibrium tests, however; three of the SNPs were in linkage disequilibrium with other SNPs in the set (rs9897046, rs428371, and rs34302939). Once

again, this led to the inclusion of rs9897046, rs428371, and rs34302939 in calculations and the exclusion of their complementary linked SNPs. A summary of these metrics for GVPComplete can be found in Table 3 below.

Table 3 Hardy-Weinberg Equilibrium and Linkage Disequilibrium Assessments for GVPComplete

					Hardy-Weinberg Equilibrium					Linkage Disequilibrium
CHROM	**POS**	**SNP**	**REF**	**ALT**	**AFR**	**EUR**	**AMER**	**SASIA**	**EASIA**	**Global R2**
1	153431406	rs41265164	G	A	0.2032	0.3855	0.2383	1	1	
1	153520203	rs116208483	G	C	1	1	1	1	1	
1	153520954	rs62624468	C	T	0.2257	1	1	1	1	
1	201289487	rs61818256	C	T	1	1	1	1	1	
6	74014637	rs28763966	C	A	0.2496	1	1	1	1	
6	7581001	rs28763967	C	T	1	0.1141	1	1	1	
12	52788945	rs1791634	C	G	0.1928	0.1455	0.3123	0.2744	0.7601	
12	53069014	rs17678945	C	A	0.794	0.372	0.6466	0.2375	1	
14	113975768	rs10148371	G	A	0.6436	1	1	0.1982	1	
14	55609418	rs11125	A	T	1	1	1	1	1	
17	38859509	rs7213256	C	T	1	1	1	1	1	
17	39116603	rs17843021	G	A	0.3533	1	1	1	1	rs17843023 (0.382633)
17	39116728	rs142154718	C	T	1	0.7577	0.4618	1	1	
17	39183254	rs62623375	C	T	1	0.869	0.07901	1	1	rs150218495 (0.220935)
17	39593768	rs2604953	G	T	0.876	1	1	0.1279	0.1366	
17	39633354	rs138303882	G	A	1	1	1	1	1	
17	39635733	rs743686	A	G	0.04065	1	0.1874	0.3118	1	
17	39913771	rs41283425	C	T	0.2556	1	0.2907	0.02936	0.191	
18	28605818	rs79011243	C	A	1	1	1	1	1	
21	31744310	rs9636845	A	T	0.4559	1	0.027	1	1	
21	32253513	rs71321355	C	T	1	0.4996	0.09218	1	1	
17	39913771	rs143043662	C	T	0.4076	0.3864	0.8761	0.3555	1	
17	39619115	rs2071563	G	A	0.6334	0.5838	0.2662	0.5399	0.7647	
17	39620641	rs146792525	C	T	1	1	1	1	1	
12	52788928	rs2658658	G	A	1	1	1	1	1	
12	52788945	rs1732263	C	G	0.4842	0.8575	0.1576	0.2331	0.7637	
12	52713088	rs2857663	G	A	1	1	1	1	1	

Table 3 continued

17	39183313	rs148449559	G	C	1	0.3927	0.314	0.1449	1	
17	39156084	rs9897046	T	C	0.4636	1	1	1	1	rs138200823 (0.242294), rs62623375 (0.256221)
17	39334241	rs62067292	G	C	1	1	0.7421	0.08258	0.3706	
17	39323971	rs428371	G	A	0.667	0.8889	1	0.7526	0.6399	rs366700 (0.993514), rs385055 (0.830107), rs73983172 (0.233243), rs9902235 (0.222918), rs2191379 (0.214576)
21	46117792	rs34302939	G	A	0.002152	0.3573	0.1714	0.6986	1	

The probability of identity (PID) was calculated for the marker sets GVP21 and GVPComplete in order to apply power to the current proteomic detection method. The probability of identity is the probability that two individuals selected at random will have an identical profile using the same set of variants. It is useful in SNP panel studies to determine the minimum number of SNPs needed for identity calling. It is easily explained as: the lower the probability of identity, the more variable the DNA markers. The PID of a locus is calculated as the sum of squares of the genotype frequencies for that SNP. The product of these values for a set of loci then provides a PID for the total set of markers used. The full tables of calculations can be found in Appendix O. The PID was calculated for each population (Africa, America, South Asia, East Asia, and Europe) and can be seen in Table 4 below. Each population should be treated independently in identity calculations to accurately represent the true power of the SNP set.

Table 4 Probability of Identity for all populations on GVP21 and GVPComplete

	Probability of Identity (PID)				
SNP Set	Africa (n=661)	America (n=347)	East Asia (n=504)	South Asia (n=489)	Europe (n=503)
GVP21	1.14E-03	4.96E-03	1.68E-02	8.46E-03	1.34E-02
GVPComplete	2.49E-06	1.29E-04	1.23E-03	1.08E-04	3.21E-04

The average probability of identity was 8.97×10^{-3} for GVP21 and 3.58×10^{-4} for GVPComplete. The probability of identity for GVPComplete will hold more power after the 11 additional SNPs have been genotyped and checked for method accuracy. These probabilities give insight on the usefulness of these two marker sets for identification, but also proves that the current laboratory method could use improvement with regards to the detection of these GVPs. With better SNP detection in the proteomic analysis comes more power in the probability of identity calculation, which is crucial in forensic casework. In current forensic methods for human identification, STR kits like Identifiler (ThermoFisher Scientific) and Globalfiler (ThermoFisher Scientific) exhibit high probabilities of identity at 6.18×10^{-19} and 7.73×10^{-28}. These kits are forensically stable due to their high polymorphic characteristics across all populations and set an exceptional standard for other human identification methods.

Consequently, it is apparent that additional extensive research in proteomics is still required to obtain a stable set of forensically applicable markers exuding a strong probability of identity. Refining the proteomic analysis method will be the first catalyst for improving this statistic, but still may not reach the high standard of current forensic STR kits. If that is the case, researchers may find success in identifying additional hair proteins, and therefore SNPs, to proteomically re-asses.

Identifying Potential AIMs and Generating Metrics for the Full GVP Candidate List

A population genetics assessment was conducted on the full set of 233 SNPs to identify potential ancestry informative markers (AIMs). The full set of markers, as seen in Appendix P, was also assessed for their ability to be used as an identity marker, if the proteomics method were to improve in the future. Each candidate GVP was analyzed using genotype data from the 1000 Genomes project and assessed based on Fst values, heterozygosity, minor allele frequency, linkage disequilibrium, and Hardy-Weinberg equilibrium. The full list of SNPs and their allele frequencies for African, American, South Asian, East Asian, and European populations can be found in Appendix Q**.** These were obtained through population data from the 1000 Genomes Project and were generated using PLINK [6] commands. For the potential future use of these markers in forensic applications, it was necessary to generate an allelic frequency table. In order to apply statistical power to the rarity of a genetic profile, statistics such as random match probabilities or likelihood ratios are calculated. These statistics, and therefore the identification of an individual, rely on this population frequency information. If this set of markers were to be used for forensic human identification in the future, these allelic frequencies may be used to compute the match probability statistic. However, it is advisable to obtain more population datasets and individuals to improve the accuracy of this calculation as the 1000 Genomes project dataset only consists of 26 global populations and within region (i.e. within Europe) frequencies do not truly represent the countries in this region.

Analyses of population genetics structure have shown that continental population groups can be identified by examining differences in allele frequencies and measuring the degree of this differentiation by computing Fst values. Large Fst values correspond to large absolute allelic frequency differences. Therefore, calculating these values in a pairwise fashion gives insight on

markers that are distinguishable across populations. These distinguishable differences lead to the development of ancestry informative markers, which can be highly beneficial in forensic settings.

Allelic frequencies and Fst measurements were evaluated for the total set of 233 SNPs using data from the 1000 Genomes Project. The markers that had allelic frequencies near fixation in specific populations and exhibited high Fst values (>0.2) across all populations in the pairwise analysis were identified as potential ancestry informative markers and can be seen in Table 5. A scatterplot of these markers and their minor allele frequencies can be seen in Figure 7.

Table 5 Minor allele frequencies and Fst values for AIMs for all populations

		Africa	**America**	**East Asia**	**Europe**	**South Asia**
SNP	**Pairwise Fst**	**Minor Allele Frequency**	**Minor Allele Frequency**	**Minor Allele Frequency**	**Minor Allele Frequency**	**Minor Allele Frequency**
rs2037912	0.428-0.626	0.03026	0.471	0.652	0.532	0.486
rs9891361	0.641-0.703	0.8298	0.138	0.095	0.099	0.089
rs3120655	0.291-0.376	0.407	0.033	0	0.001	0
rs385055	0.348-0.430	0.466	0.033	0	0.002	0
rs214803	0.301-0.441	0.59	0.167	0.075	0.162	0.157
rs2634041	0.339-0.398	0.259	0.269	0.7589	0.283	0.308
rs6761276	0.282-0.392	0.421	0.418	0.8204	0.389	0.332

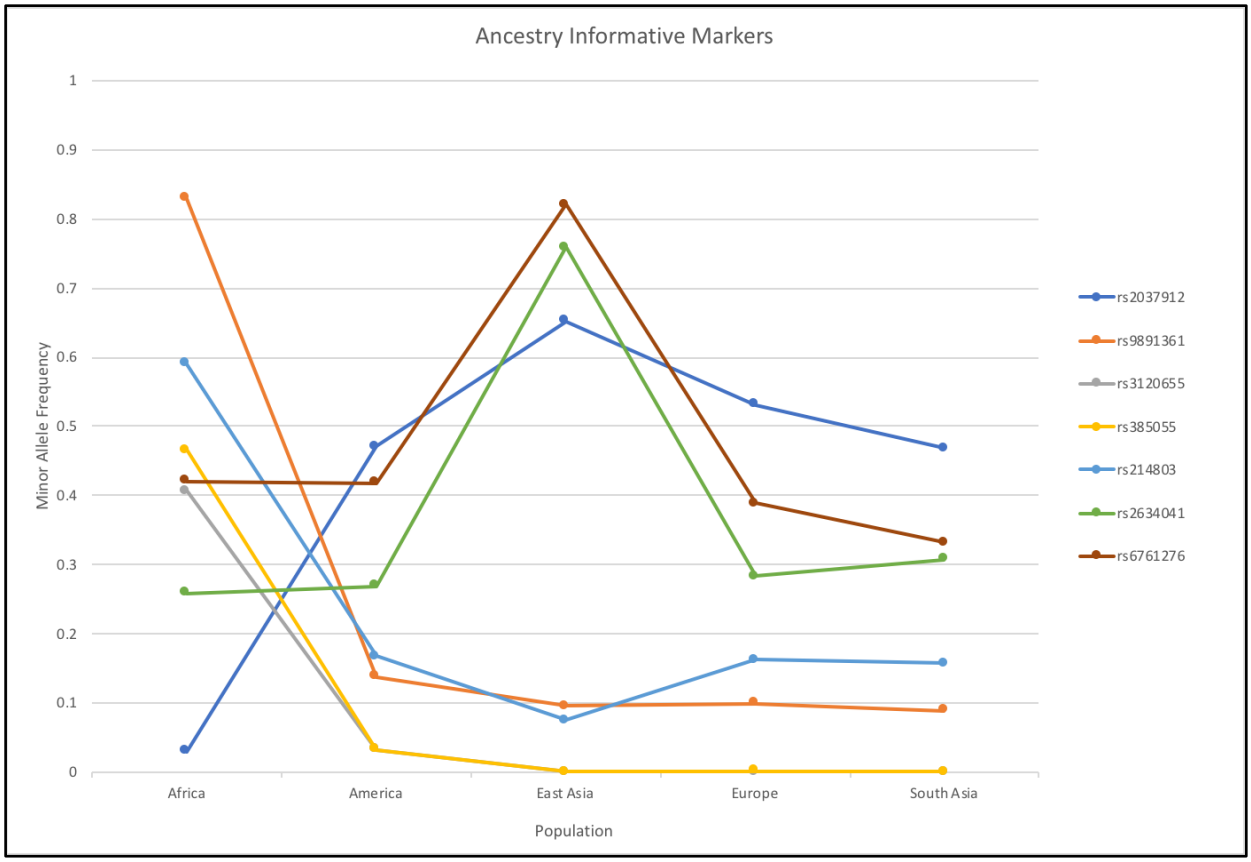


Figure 7 Ancestry Informative Markers within 233 SNP Set

Based on this analysis, 7 markers were determined to help infer ancestry for two populations: Africa versus non-African, and East Asia versus non East-Asian. For the African population, 5 SNPs were identified as potential AIMs: rs2037912, rs9891361, rs3120655, rs385055, and rs214803. These markers are highlighted in green in the table found in Appendix P. The marker with the highest pairwise Fst values, or most genetic differentiation between the other populations, was rs9891361 with values ranging from 0.64-0.70. The most informative marker based on allelic frequencies was rs3120655. This SNP displayed allele frequencies near fixation (0) for all populations other than Africa, with an African minor allele frequency of 0.407. This indicates that there is a high level of heterozygosity in the African population for this SNP and suggests that if the allele is then present in a sample, the probability of it coming from an American,

East Asian, South Asian, or European is very low. For the East Asian population, 2 SNPs were identified as potential AIMs: rs2634041 and rs6761276. Although these markers were not as informative as the African markers, they both showed high levels of Fst with ranges from 0.28-0.39, indicating that there is a high level of genetic differentiation with the other populations.

This knowledge could be useful in a forensic setting when calculating match probability statistics for a casework sample. If a hair sample were found at a crime scene and this proteomic analysis was used to generate a genetic profile, this set of AIMs could be used to infer the population to which the perpetrator belongs. If a specific population is identified first, the allelic frequencies for that specific population can be used in an RMP or likelihood ratio calculation, providing a more powerful and specific statistic.

However, if these 7 AIMs do not provide assistance in first identifying the ancestral origin of the individual, other SNPs in the 233 marker set could still be used for the individualization of the genetic profile. After removing the 7 AIMs, the 226 remaining SNPs were assessed for their suitability as an identity marker via linkage disequilibrium and Hardy-Weinberg equilibrium testing. The primary goal in human identification is to reliably distinguish unrelated individuals from one another, therefore SNPs that do not pass these tests, or are not independently inherited, should not be used in human identification. SNPs that did not pass these tests can be seen in Appendix P denoted with an "X". SNPs that displayed LD with any other SNP in the set were completely removed as well as SNPs that did not pass HWE (<0.0002). The remaining 124 SNPs that passed the population genetics assessment for LD and HWE can then be used for identification. Using these markers in a forensic setting would require all populations and their allelic frequencies to be used and reported for RMP calculations. As previously stated, these SNPs and their allele frequencies per population can be seen in Appendix Q.

Hair Structure Correlation Assessment

A Pearson's correlation test was performed in R to assess the association between genotype and hair structure phenotype for the 114 SNPs that were included in both the proteomic analysis and the MEGA array. During each correlation assessment, the ancestry, age, and sex of each individual was controlled for by assigning them as test covariates.

For the first correlation, genotype and phenotype data was assessed for the 66 samples used in the proteomic analysis. After running the correlation, there seemed to be a significant positive correlation (p-value <0.05, passing Bonferroni correction, and R^2 correlation values >0.4) for a small number of SNPs, which specifically corresponded to the curly hair phenotype. However, these significant correlation values were later hypothesized to be due to a high rate of African American or Admixed individuals with curly hair within the small dataset of 66 individuals. Out of 66 individuals, 21 self-reported their hair type as curly. Of the 21 individuals with curly hair, 20 were African American or Admixed. The effect of the ancestral origin of the individual on the hair phenotype was too strong, making it unable to be completely controlled for during the correlation test.

Therefore, the second correlation assessment was conducted to test the above hypothesis and to prove or disprove the significance of the SNPs found in the first correlation. To do this assessment, it was crucial to use a bigger sample size and isolate individuals based on their populations, specifically if they were European or non-European. By removing individuals with non-European ancestral origins from the dataset, there will be less of a population stratification effect. If a significance value or correlation remained after testing on a European-only sample set, then the SNP could be considered as associated with the curly hair phenotype. It was also important to increase the power of the phenotype in the second correlation test. To do this, the phenotypes

of straight, wavy, and curly were re-coded as 0 and 1 (depending on the phenotype of interest) instead of 1, 2, and 3. By doing this, the correlation test only has two phenotypic variables to assess (curly vs. non-curly) instead of 3 variables (straight, curly, wavy).

1,821 European individuals from the Walsh lab database were used to evaluate the candidate GVPs for their association with hair structure. Upon initial examination, one SNP (rs4796697) showed significant correlation with the binary association analyses of curly versus non-curly in a European cohort with a p-value of 0.004298. However, after further analysis, it was discovered that this was likely a false positive significance value due to a low minor allele frequency for this SNP. It is also important to note, that using questionnaire-based phenotypes can lead to increased background noise and false positives, so additional research on this phenotype (i.e. reclassification and verification using imagery or microscopy) may provide a more accurate correlation assessment. It would also be recommended to correct for hair color in future analyses, as pigment has an effect on hair structure.

Future Directions

Many significant insights were generated from this research, but there are still many areas for improvement and advancement. From this assessment, we have developed a set of 32 SNPs that could be used for assistance in identification in current forensic casework. We have also provided all necessary population genetics statistics for the total set of 233 SNPs for future use in forensics.

One of the first future directions of this research will be to finish genotyping the remaining 119 SNPs that were not included in the MEGA SNP array. With this genotyping information, correct GVP genotyping accuracy calculations can be performed for any SNP (out of the 233 set)

that is detected by the proteomic method now and in the future. It would be valuable to be able to confidently add the additional 11 variants as part of the GVPComplete set for increased power in PID calculations. Another main goal for the future of this research would be to improve and optimize the proteomic method. Out of the 233 SNPs, only 32 were detected at a trustworthy rate (> 90%). By optimizing the method, the sensitivity may increase, allowing for the detection of more GVPs (for which all the population genetics assessments and population allele frequencies have already been generated for future use), and therefore the inference of additional variant genotypes to use for ancestry/identity inference purposes. During this assessment, it was also apparent that there was a lack of criteria or thresholds for the detection of SNPs and the trustworthiness of their GVP inferred genotypes going from the laboratory method to the analyses method. If genotypic data were not available, one would not be able to know if the inference was an informative GVP or not. This research has certainly attempted to put quality control metrics to the GVP to variant genotype conversion. Additional research could also be conducted to further explore the potential predictive hair structure variants found in this assessment.

3.4 References

1. Laatsch, C.N., et al., *Human hair shaft proteomic profiling: individual differences, site specificity and cuticle analysis.* 2014. **2**: p. e506.
2. Lee, Y.J., et al., *Proteome analysis of human hair shaft: from protein identification to posttranslational modification.* 2006. **5**(5): p. 789-800.
3. Milan, J.A., et al., *Comparison of protein expression levels and proteomically-inferred genotypes using human hair from different body sites.* 2019.
4. Parker, G.J., et al., *Demonstration of protein-based human identification using the hair shaft proteome.* 2016. **11**(9): p. e0160653.
5. Shimomura, Y. and M. Ito. *Human hair keratin-associated proteins*. in *Journal of Investigative Dermatology Symposium Proceedings*. 2005. Elsevier.
6. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* 2007. **81**(3): p. 559-575.
7. Hotelling, H.J.J.o.e.p., *Analysis of a complex of statistical variables into principal components.* 1933. **24**(6): p. 417.
8. Consortium, G.P., A. Auton, and L.J.N. Brooks, *A global reference for human genetic variation.* 2015. **526**(7571): p. 68-74.
9. University, S., *Human Genome Diversity Project*.
10. Team, R.C., *A language and environment for statistical computing.* 2013, R Foundation for Statistical Computing: Vienna, Australia.
11. Kim, S., *ppcor: Partial and Semi-Partial (Part) Correlation. R package version 1.1*. 2015.

CHAPTER 4. CONCLUSIONS

4.1 HIrisPlex-S system for eye, hair, and skin color prediction from DNA: massively parallel sequencing solutions for two common forensically used platforms

The first goal of this research was to assist in the development and validation of the HIrisPlex-S assay on the Illumina MiSeq and Thermo Fisher Ion Torrent sequencing platforms. The sequencing alignment and genotype extraction process was simplified with the development of the HPS-MPS pipeline. This semi-automated pipeline aligns sequences to the human genome (hg19) and extracts the 41 SNPs of interest for HIrisPlex-S. Through this forensic validation, the assay demonstrated sensitivity down to 250pg and was able to produce accurate genotypic profiles for most simulated casework DNA samples, including saliva (fresh and degraded), semen, vaginal secretion, hair, and dried blood. The only simulated casework samples that did not produce full and accurate genotypic profiles were the touch DNA samples, which both expectedly displayed very low input DNA concentrations. Based on the mixture studies, a mixture deconvolution and minimum read count threshold tool was developed. This tool is intended to aid users in the separation of mixtures and establish confidence in the genotypes called for the major and minor contributors within the sample. Ultimately, the results of this forensic validation, including the creation of both the mixture separation tool and HPS-MPS pipeline, demonstrate the robust ability of the HIrisPlex-S Massive Parallel genotyping system to produce successful results in a variety of Forensic DNA Phenotyping scenarios. Such advancements in forensic DNA analysis allow for future expansions of Forensic DNA Phenotyping on massive parallel sequencing platforms to include additional predictive phenotypic SNPs, bio-geographic ancestry informative markers, and to ultimately produce an all-in-one highly prediction assay.

4.2 Proteomic Analysis of Hair

The second goal of this research was to evaluate a proteomic analysis method for 99 hair samples and optimize a marker set for future use in proteomics. Hair sample GVP data (n = 66) for 233 candidate markers was analyzed for overall detection and genotyping accuracy. Genotyping accuracy calculations were computed for SNPs that were also included in the Illumina MEGA SNP array, which totaled to 114 SNPs. After analyzing the GVP data based on detection thresholds, the total number of proteomically detected SNP GVPs was 32 (GVPComplete): 21 with genotypic data from SNPs in the MEGA array (GVP21) and 11 without. Genotype accuracy calculations were computed for GVP21 and displayed an average of 91%. The probability of identity, or the probability that two individuals selected at random will have an identical genotype, was calculated for GVP21 and GVPComplete. The average probability of identity was 8.97×10^{-3} for GVP21 and 3.58×10^{-4} for GVPComplete. With the current proteomic method, these 32 SNPs (GVPComplete) could be used in forensic casework to assistance in human identification.

A thorough population genetics assessment was also conducted on the set of 233 candidate SNPs to assess their ability to be used as identity markers and identify potential ancestry informative markers (AIMs). Based on the population genetics tests (HWE, LD, minor allele frequency and heterozygosity, Fst), 7 AIMs were identified to help ancestry inference for two populations: East Asia and Africa. Future research should be conducted to investigate the usefulness of these AIMs for inferring ancestry in other datasets. Also based on the population genetics tests, 124 SNPs were identified as having the potential for use in human identification. The allele frequencies of these SNPs in each population are required to compute forensic statistics, specifically in random match probability calculations, to apply statistical power to the genetic profile. If proteomic methods improve in the future, collaborators will have the necessary population genetics information to forensically assess the newly detected SNPs.

APPENDIX A

Supplementary Table 1 Full description of the samples used for the developmental validation of the HPS-MPS assays, including sample type, concentration and phenotype of the individuals used

Sample	Source	Concentration	Assessment	Phenotype		
				Eye	Hair	Skin
1	Single	1 ng	Sensitivity: 9947A			
2	Single	500 pg	Sensitivity: 9947A			
3	Single	250 pg	Sensitivity: 9947A			
4	Single	100 pg	Sensitivity: 9947A			
5	Single	50 pg	Sensitivity: 9947A			
6	Single	25 pg	Sensitivity: 9947A			
7	Single	10 pg	Sensitivity: 9947A			
8	Single	5 pg	Sensitivity: 9947A			
9	Single	1 ng	Sensitivity: 9947A			
10	Single	500 pg	Sensitivity: 9947A			
11	Single	250 pg	Sensitivity: 9947A			
12	Single	100 pg	Sensitivity: 9947A			
13	Single	50 pg	Sensitivity: 9947A			
14	Single	25 pg	Sensitivity: 9947A			
15	Single	10 pg	Sensitivity: 9947A			
16	Single	5 pg	Sensitivity: 9947A			
17	Single	1 ng	Sensitivity: 9948			
18	Single	500 pg	Sensitivity: 9948			
19	Single	250 pg	Sensitivity: 9948			
20	Single	100 pg	Sensitivity: 9948			
21	Single	50 pg	Sensitivity: 9948			
22	Single	25 pg	Sensitivity: 9948			
23	Single	10 pg	Sensitivity: 9948			
24	Single	5 pg	Sensitivity: 9948			
25	Single	1 ng	Sensitivity: 9948			
26	Single	500 pg	Sensitivity: 9948			
27	Single	250 pg	Sensitivity: 9948			
28	Single	100 pg	Sensitivity: 9948			
29	Single	50 pg	Sensitivity: 9948			
30	Single	25 pg	Sensitivity: 9948			
31	Single	10 pg	Sensitivity: 9948			
32	Single	5 pg	Sensitivity: 9948			
33	Single	500 pg	Stability: 0 min UV	Blue/Blue Gray	Light Brown	Pale
34	Single	500 pg	Stability: 5 min UV	Blue/Blue Gray	Light Brown	Pale
35	Single	500 pg	Stability: 10 min UV	Blue/Blue Gray	Light Brown	Pale
36	Single	500 pg	Stability: 20 min UV	Blue/Blue Gray	Light Brown	Pale
37	Single	0.126 ng	Simulated Casework: Dried Degraded Blood	Blue	Blond	Pale
38	Single	27.9 ng	Simulated Casework: Semen	Blue	Blond-Light Brown	Intermediate
39	Single	4.59 ng	Simulated Casework: Wet Saliva	Dark Brown	Dark Brown-Black	Dark
40	Single	0.425 ng	Simulated Casework: Dried & Degraded Saliva	Dark Brown	Dark Brown-Black	Dark
41	Single	< 1 pg	Simulated Casework: Touch DNA Fresh	Hazel/Green	Brown	Intermediate
42	Single	0.121 ng	Simulated Casework: Hair	Brown	Light Brown-Brown	Intermediate
43	Single	0.018 ng	Simulated Casework: Touch DNA Degraded	Hazel/Green	Brown	Intermediate
44	Single	6.89 ng	Simulated Casework: Vaginal Swab	Blue	Auburn	Pale
45	Mixture	29.4 ng	Simulated Casework: Vaginal Swab/Semen	Refer to Samples 38 and 44		
46	Single	500 pg	Species: Dog			
47	Single	500 pg	Species: Cat			
48	Single	500 pg	Species: Primate			
49	Single	500 pg	Species: Pig			
50	Single	500 pg	Species: Mouse			
51	Single	0.126 ng	Simulated Casework: Dried Degraded Blood	Blue	Blond	Pale
52	Single	27.9 ng	Simulated Casework: Semen	Blue	Blond-Light Brown	Intermediate
53	Single	4.59 ng	Simulated Casework: Wet Saliva	Dark Brown	Dark Brown-Black	Dark
54	Single	0.425 ng	Simulated Casework: Dried & Degraded Saliva	Dark Brown	Dark Brown-Black	Dark
55	Single	< 1 pg	Simulated Casework: Touch DNA Fresh	Hazel/Green	Brown	Intermediate
56	Single	0.121 ng	Simulated Casework: Hair	Brown	Light Brown-Brown	Intermediate
57	Single	0.018 ng	Simulated Casework: Touch DNA Degraded	Hazel/Green	Brown	Intermediate
58	Single	6.89 ng	Simulated Casework: Vaginal Swab	Blue	Auburn	Pale
59	Mixture	29.4 ng	Simulated Casework: Vaginal Swab/Semen	Refer to Samples 38 and 44		
60	Single	500 pg	Control: 2800M			
61	Single	500 pg	Reproducibility	Dark Brown	Black	Intermediate
62	Single	500 pg	Reproducibility	Blue/Green Yellow	Light Red/Strawberry Blond	Very Pale
63	Single	500 pg	Reproducibility	Blue/Green Yellow	Dark Brown	Intermediate
64	Single	500 pg	Reproducibility	Blue/Green Yellow	Red Brown/Auburn	Pale
65	Single	500 pg	Reproducibility	Blue/Blue Gray	Blond	Pale
66	Single	500 pg	Reproducibility	Dark Brown	Black	Dark
67	Single	500 pg	Reproducibility	Dark Brown	Dark Brown	Dark
68	Single	500 pg	Reproducibility	Blue/Green Yellow	Light Red/Strawberry Blond	Pale
69	Single	500 pg	Reproducibility	Dark Brown	Black	Intermediate
70	Single	500 pg	Reproducibility	Blue/Green Yellow	Light Red/Strawberry Blond	Very Pale
71	Single	500 pg	Reproducibility	Blue/Green Yellow	Dark Brown	Intermediate
72	Single	500 pg	Reproducibility	Blue/Green Yellow	Red Brown/Auburn	Pale
73	Single	500 pg	Reproducibility	Blue/Blue Gray	Blond	Pale
74	Single	500 pg	Reproducibility	Dark Brown	Black	Dark
75	Single	500 pg	Reproducibility	Dark Brown	Dark Brown	Dark
76	Single	500 pg	Reproducibility	Blue/Green Yellow	Light Red/Strawberry Blond	Pale
77	Mixture	1 ng:1 ng	Mixtures	Refer to Samples 62 and 63		
78	Mixture	1 ng: 500 pg	Mixtures	Refer to Samples 62 and 63		
79	Mixture	100 pg: 500 pg	Mixtures	Refer to Samples 62 and 63		
80	Mixture	1 ng: 100 pg	Mixtures	Refer to Samples 62 and 63		
81	Mixture	5 ng: 500 pg	Mixtures	Refer to Samples 62 and 63		
82	Mixture	1 ng:1 ng	Mixtures	Refer to Samples 70 and 71		
83	Mixture	1 ng: 500 pg	Mixtures	Refer to Samples 70 and 71		
84	Mixture	100 pg: 500 pg	Mixtures	Refer to Samples 70 and 71		
85	Mixture	1 ng: 100 pg	Mixtures	Refer to Samples 70 and 71		
86	Mixture	5 ng: 500 pg	Mixtures	Refer to Samples 70 and 71		
87	Mixture	1 ng:1 ng	Mixtures	Refer to samples 65 and 67		
88	Mixture	1 ng: 500 pg	Mixtures	Refer to samples 65 and 67		
89	Mixture	100 pg: 500 pg	Mixtures	Refer to samples 65 and 67		
90	Mixture	1 ng: 100 pg	Mixtures	Refer to samples 65 and 67		
91	Mixture	5 ng: 500 pg	Mixtures	Refer to samples 65 and 67		
92	Mixture	1 ng:1 ng	Mixtures	Refer to samples 73 and 75		
93	Mixture	1 ng: 500 pg	Mixtures	Refer to samples 73 and 75		
94	Mixture	100 pg: 500 pg	Mixtures	Refer to samples 73 and 75		
95	Mixture	1 ng: 100 pg	Mixtures	Refer to samples 73 and 75		
96	Mixture	5 ng: 500 pg	Mixtures	Refer to samples 73 and 75		

APPENDIX B

Supplementary Table 2 Two-person mixture deconvolution tool

FOR MIXTURE PLEASE ENTER
Minor Input Ratio 0.09091 (e.g. if 1:1 mixture ratio, put in 0.5, if 1:5 ratio, put in 0.166)
if 1:10, put in 0.090909

ter sample concentration inp — 100 pg input

Further breakdown to show how results are generated

Scenario	1	2	3	4	5	6	7
Major	alt	ref	het	ref	alt	het	het
Minor	ref	alt	het	het	het	ref	alt

CHROM	POS	SNP	REF	ALT	ref.forward	ref.reverse	alt.forward	alt.reverse	REF DP	ALT DP	Coverage Threshold	Major	Minor	Both Homozygous REF	Both Homozygous ALT	Mixture Scenario Conclusion REF	Mixture Scenario Conclusion ALT
chr16	89985753	rs79629617	C	-	1961	18	0	0	1979	0	Passed	C/C	C/C	Y	N	4	4
chr16	89986091	rs11547464	G	A	3601	32	1	1	3633	2	Passed	G/G	G/G	Y	N	4	4
chr16	89986154	rs885479	G	A	3604	33	1	0	3637	1	Passed	G/G	G/G	Y	N	4	4
chr16	89986144	rs1805008	C	T	3601	33	0	0	3634	0	Passed	C/C	C/C	Y	N	4	4
chr16	89985844	rs1805005	G	T	1122	10	1036	11	1132	1047	Passed	G/T	G/T	N	N	3	3
chr16	89985918	rs1805006	C	A	2153	20	3	0	2173	3	Passed	C/C	C/C	Y	N	4	4
chr16	89986117	rs1805007	C	T	3603	32	2	0	3635	2	Passed	C/C	C/C	Y	N	4	4
chr16	89986546	rs1805009	G	C	3146	8	1379	0	3154	1379	Passed	G/C	G/G	N	N	6	6
chr16	89986122	rs20132689	C	A	3604	33	1	0	3637	1	Passed	C/C	C/C	Y	N	4	4
chr16	89985940	rs2228479	G	A	2135	37	2	1	2172	3	Passed	G/G	G/G	Y	N	4	4
chr16	89986130	rs1110400	T	C	3602	33	0	0	3635	0	Passed	T/T	T/T	Y	N	4	4
chr5	33958959	rs28777	C	A	125	1	258	1	126	259	Caution	C/A	A/A	N	N	7	7
chr5	33951693	rs16891982	C	G	594	3	893	3	597	896	Caution	C/G	G/G	N	N	7	7
chr12	89328335	rs12821256	T	C	692	3	0	0	695	0	Caution	T/T	T/T	Y	N	4	4
chr6	457748	rs4959270	C	A	648	6	554	6	654	560	Caution	C/A	C/C	N	N	6	6
chr6	396321	rs12203592	C	T	5286	4	4	0	5290	4	Passed	C/C	C/C	Y	N	4	4
chr11	88911696	rs1042602	C	A	1209	6	589	1	1215	590	Passed	C/A	C/C	N	N	6	6
chr15	28230318	rs1800407	C	T	2823	116	0	0	2939	0	Passed	C/C	C/C	Y	N	4	4
chr14	92801203	rs2402130	G	A	963	8	2477	13	971	2490	Passed	G/A	A/A	N	N	7	7
chr15	28365618	rs12913832	A	G	2658	0	2590	0	2658	2590	Passed	A/G	A/G	N	N	3	3
chr20	33218090	rs2378249	G	A	247	1	762	6	248	768	Caution	A/A	G/G	N	N	1	1
chr14	92773663	rs12896399	G	T	260	3	370	4	263	374	Caution	G/T	T/T	N	N	7	7
chr11	89011046	rs1393350	G	A	1113	9	1	0	1122	1	Caution	G/G	G/G	Y	N	4	4
chr9	12709305	rs683	C	A	442	3	478	4	445	482	Passed	C/A	C/A	N	N	3	3
chr16	89383725	rs3114908	T	C	1479	5	2531	6	1484	2537	Passed	T/C	C/C	N	N	7	7
chr15	28197037	rs1800414	T	C	1794	11	1	0	1805	1	Passed	T/T	T/T	Y	N	4	4
chr9	16858084	rs10756819	G	A	2245	45	1655	39	2290	1694	Passed	G/A	G/G	N	N	6	6
chr15	28453215	rs2238289	A	G	755	12	704	18	767	722	Passed	A/G	A/G	N	N	3	3
chr14	92882826	rs17128291	A	G	1807	66	2	0	1873	2	Passed	A/A	A/A	Y	N	4	4
chr15	28496195	rs6497292	A	G	2253	33	1	0	2286	1	Passed	A/A	A/A	Y	N	4	4
chr15	28356859	rs1129038	C	T	1197	3	1469	2	1200	1471	Passed	C/T	T/T	N	N	7	7
chr15	28530182	rs1667394	C	T	816	38	667	34	854	701	Passed	C/T	C/C	N	N	6	6
chr11	89017961	rs1126809	G	A	1839	7	2	0	1846	2	Caution	G/G	G/G	Y	N	4	4
chr15	28288121	rs1470608	G	T	522	3	250	3	525	253	Caution	G/T	G/G	N	N	6	6
chr15	48426484	rs1426654	A	G	26	1	9	0	27	9	Passed	A/A	G/G	N	N	2	2
chr20	32785212	rs6119471	C	G	2582	7	0	0	2589	0	Passed	C/C	C/C	Y	N	4	4
chr15	28187772	rs1545397	A	T	820	33	163	6	853	169	Caution	A/A	T/T	N	N	2	2
chr20	32665748	rs6059655	A	G	1	0	3300	15	1	3315	Passed	G/G	G/G	N	Y	5	5
chr15	28271775	rs12441727	G	A	1076	13	2015	19	1089	2034	Passed	G/A	A/A	N	N	7	7
chr16	89984378	rs3212355	C	T	985	5	1	0	990	1	Passed	C/C	C/C	Y	N	4	4
chr16	90024206	rs8051733	A	G	602	1	20	0	603	20	Passed	A/A	A/G	N	N	4	4

SNP	Ref difference to scenario 1	2	3	4	5	6	7	Alt difference to scenario 1	2	3	4	5	6	7
rs79629617	1799	180	990	90	1889	900	1079	1799	180	990	90	1889	900	1079
rs11547464	3303	328	1816	163	3468	1650	1981	3303	328	1816	163	3468	1650	1981
rs885479	3306	330	1818	164	3472	1653	1983	3306	330	1818	164	3472	1653	1983
rs1805008	3304	330	1817	165	3469	1652	1982	3304	330	1817	165	3469	1652	1982
rs1805005	934	849	43	948	1033	57	142	934	849	43	948	1033	57	142
rs1805006	1975	195	1085	96	2074	986	1184	1975	195	1085	96	2074	986	1184
rs1805007	3304	329	1817	163	3470	1651	1982	3304	329	1817	163	3470	1651	1982
rs1805009	2742	967	888	1173	2948	681	1094	2742	967	888	1173	2948	681	1094
rs20132689	3306	330	1818	164	3472	1653	1983	3306	330	1818	164	3472	1653	1983
rs2228479	1974	195	1085	96	2073	986	1183	1974	195	1085	96	2073	986	1183
rs1110400	3305	330	1818	165	3470	1652	1983	3305	330	1818	165	3470	1652	1983
rs28777	91	224	67	242	109	84	49	91	224	67	242	109	84	49
rs16891982	461	760	150	828	529	217	82	461	760	150	828	529	217	82
rs12821256	632	63	348	32	663	316	379	632	63	348	32	663	316	379
rs4959270	544	450	47	505	599	8	102	544	450	47	505	599	8	102
rs12203592	4809	477	2643	237	5049	2402	2884	4809	477	2643	237	5049	2402	2884
rs1042602	1051	426	313	508	1133	230	395	1051	426	313	508	1133	230	395
rs1800407	2672	267	1470	134	2805	1336	1603	2672	267	1470	134	2805	1336	1603
rs2402130	656	2175	760	2333	814	917	602	656	2175	760	2333	814	917	602
rs12913832	2181	2113	34	2351	2419	205	273	2181	2113	34	2351	2419	205	273
rs2378249	156	676	260	722	202	306	214	156	676	260	722	202	306	214
rs12896399	205	316	56	345	234	84	27	205	316	56	345	234	84	27
rs1393350	1020	101	561	50	1071	509	612	1020	101	561	50	1071	509	612
rs683	361	398	19	440	403	61	24	361	398	19	440	403	61	24
rs3114908	1118	2171	527	2354	1301	709	344	1118	2171	527	2354	1301	709	344
rs1800414	1641	163	902	81	1723	820	984	1641	163	902	81	1723	820	984
rs10756819	1928	1332	298	1513	2109	117	479	1928	1332	298	1513	2109	117	479
rs2238289	632	587	23	654	699	45	90	632	587	23	654	699	45	90
rs17128291	1703	168	936	83	1788	850	1021	1703	168	936	83	1788	850	1021
rs6497292	2078	207	1143	103	2182	1039	1246	2078	207	1143	103	2182	1039	1246
rs1129038	957	1228	136	1350	1079	257	14	957	1228	136	1350	1079	257	14
rs1667394	713	560	77	630	783	6	147	713	560	77	630	783	6	147
rs1126809	1678	166	922	82	1762	838	1006	1678	166	922	82	1762	838	1006
rs1470608	454	182	136	218	490	101	171	454	182	136	218	490	101	171
rs1426654	24	6	9	7	25	7	11	24	6	9	7	25	7	11
rs6119471	2354	235	1295	118	2471	1177	1412	2354	235	1295	118	2471	1177	1412
rs1545397	760	76	342	123	807	296	388	760	76	342	123	807	296	388
rs6059655	300	3014	1657	3164	150	1808	1506	300	3014	1657	3164	150	1808	1506
rs12441727	805	1750	473	1892	947	614	331	805	1750	473	1892	947	614	331
rs3212355	900	89	495	44	945	449	540	900	89	495	44	945	449	540
rs8051733	546	37	292	8	575	263	320	546	37	292	8	575	263	320

The greener, the better scenario it is
The more red, the further apart from the scenario

Supplementary Table 2 Continued

MIXTURE SCENARIO 1			MIXTURE SCENARIO 2			MIXTURE SCENARIO 3			MIXTURE SCENARIO 4			MIXTURE SCENARIO 5			MIXTURE SCENARIO 6			MIXTURE SCENARIO 7			SINGLE SOURCE		
MAJOR HOMO(alt) and OPP MINOR (ref)			OPP MAJOR (ref) and MINOR HOMO(alt)			BOTH MAJOR and MINOR HETERO			HOMO MAJOR (ref) and MINOR HETERO			HOMO MAJOR (alt allele) and MINOR HETERO			MAJOR HETERO and HOMO MINOR (ref)			MAJOR HETERO and HOMO MINOR (alt allele)					
REF	REF+/-SD	ALT	REF	REF+/-SD	ALT	REF	REF+/-SD	ALT	REF	REF+/-SD	ALT+/-SD	REF	REF+/-SD	ALT	REF	REF+/-SD	ALT	REF	REF+/-SD	ALT	REF	REF+/-SD	ALT+/-SD
179.9089	8.995446	1799.091	89.99455	179.9089	8.995446	989.5	49.475	989.5	49.475	1889.046	94.45228	89.95446	4.497723	1889.046	94.45228	53.97272	899.5455	44.97728	1079.454	53.97272	1979	98.95	0
330.4542	16.52271	3304.546	165.2273	330.4542	16.52271	1817.5	90.875	1817.5	90.875	3469.773	173.4886	165.2271	8.261355	3469.773	173.4886	1982.727	173.4886	1652.273	82.61364	1982.727	173.4886	3633	181.65	2	0.1
330.7269	33.07269	3307.273	330.7273	330.7273	33.07269	1819	181.9	1819	181.9	3472.637	347.2637	165.3635	16.53635	3472.637	347.2637	1984.363	347.2637	1653.637	165.3637	1984.363	347.2637	3637	363.7	1	0.1
330.3633	16.51817	3303.637	165.1818	330.3633	16.51817	1817	90.85	1817	90.85	3468.818	173.4409	165.1817	8.259083	3468.818	173.4409	1982.182	173.4409	1651.818	82.59092	1982.182	173.4409	3634	181.7	0	0
198.0907	9.904536	1980.909	99.04546	198.0907	9.904536	1089.5	54.475	1089.5	54.475	2079.955	103.9977	99.04536	4.952268	2079.955	103.9977	1188.545	103.9977	990.4546	49.52273	1188.545	103.9977	1132	56.6	1047	52.35
197.818	9.890899	1978.182	98.9091	197.818	9.890899	1088	54.4	1088	54.4	2077.091	103.8546	98.90899	4.94545	2077.091	103.8546	1186.909	103.8546	989.091	49.45455	1186.909	103.8546	2173	108.65	3	0.15
330.636	16.5318	3306.364	165.3182	330.636	16.5318	1818.5	90.925	1818.5	90.925	3471.682	173.5841	165.318	8.265901	3471.682	173.5841	1983.818	173.5841	1653.182	82.6591	1983.818	173.5841	3635	181.75	2	0.1
412.0905	20.60452	4120.91	206.0455	412.0905	20.60452	2266.5	113.325	2266.5	113.325	4326.955	216.3477	206.0452	10.30226	4326.955	216.3477	2476.545	216.3477	2060.455	103.0227	2476.545	216.3477	3154	157.7	1379	68.95
330.7269	16.53635	3307.273	165.3637	330.7269	16.53635	1819	90.95	1819	90.95	3472.637	173.6318	165.3635	8.268174	3472.637	173.6318	1984.363	173.6318	1653.637	82.68183	1984.363	173.6318	3637	181.85	1	0.05
197.7271	19.77271	1977.273	197.7273	197.7271	19.77271	1087.5	108.75	1087.5	108.75	2076.136	207.6136	98.86354	9.886354	2076.136	207.6136	1186.364	207.6136	988.6365	98.86365	1186.364	207.6136	2172	217.2	3	0.3
330.4542	16.52271	3304.546	165.2273	330.4542	16.52271	1817.5	90.875	1817.5	90.875	3469.773	173.4886	165.2271	8.261355	3469.773	173.4886	1982.727	173.4886	1652.273	82.61364	1652.273	82.61364	3635	181.75	0	0
34.99997	1.749998	350	17.5	34.99997	1.749998	192.5	9.625	192.5	9.625	367.5	18.375	17.49998	0.874999	367.5	18.375	210	18.375	175	8.750001	210	18.375	126	6.3	259	12.95
135.7271	6.786357	1357.273	67.86364	135.7271	6.786357	746.5	37.325	746.5	37.325	1425.136	71.25682	67.86357	3.393178	1425.136	71.25682	814.3636	71.25682	678.6364	33.93182	814.3636	71.25682	597	29.85	896	44.8
63.18176	3.159088	631.8182	31.59091	63.18176	3.159088	347.5	17.375	347.5	17.375	663.4091	33.17046	31.59088	1.579544	663.4091	33.17046	379.0909	33.17046	315.9091	15.79546	379.0909	33.17046	695	34.75	0	0
110.3635	5.518176	1103.636	55.18182	110.3635	5.518176	607	30.35	607	30.35	1158.818	57.94091	55.18176	2.759088	1158.818	57.94091	662.1818	57.94091	551.8182	27.59091	662.1818	57.94091	654	98.1	560	84
481.2722	24.06361	4812.728	240.6364	481.2722	24.06361	2647	132.35	2647	132.35	5053.364	252.6682	240.6361	12.03181	5053.364	252.6682	2887.636	252.6682	2406.364	120.3182	2887.636	252.6682	5290	264.5	4	0.2
164.0907	8.204537	1640.909	82.04546	164.0907	8.204537	902.5	45.125	902.5	45.125	1722.955	86.14773	82.04537	4.102269	1722.955	86.14773	984.5454	86.14773	820.4546	41.02273	984.5454	86.14773	1215	121.5	590	59
267.1816	13.35908	2671.818	133.5909	267.1816	13.35908	1469.5	73.475	1469.5	73.475	2805.409	140.2705	133.5908	6.679539	2805.409	140.2705	1603.091	140.2705	1335.909	66.79546	1603.091	140.2705	2939	146.95	0	0
314.636	15.7318	3146.364	157.3182	314.636	15.7318	1730.5	86.525	1730.5	86.525	3303.682	165.1841	157.318	7.865901	3303.682	165.1841	1887.818	165.1841	1573.182	78.6591	1573.182	78.6591	1887.818	165.1841	971	97.1	2490	249
477.0904	47.70904	4770.91	477.091	477.0904	47.70904	2624	262.4	2624	262.4	5009.455	500.9455	238.5452	23.85452	5009.455	500.9455	2862.545	500.9455	2385.455	238.5455	2862.545	500.9455	2658	265.8	2590	2.4
92.36354	4.618177	923.6365	46.18182	92.36354	4.618177	508	25.4	508	25.4	969.8182	48.49091	46.18177	2.309089	969.8182	48.49091	554.1818	48.49091	461.8182	23.09091	554.1818	48.49091	248	12.4	768	38.4
57.90903	2.895452	579.091	28.95455	57.90903	2.895452	318.5	15.925	318.5	15.925	608.0455	30.40227	28.95452	1.447726	608.0455	30.40227	347.4545	30.40227	289.5455	14.47727	347.4545	30.40227	263	26.3	374	37.4
102.0908	5.10454	1020.909	51.04546	102.0908	5.10454	561.5	28.075	561.5	28.075	1071.955	53.59773	51.0454	2.55227	1071.955	53.59773	612.5454	53.59773	510.4546	25.52273	612.5454	53.59773	1122	112.2	1	0.1
84.27264	4.213632	842.7274	42.13637	84.27264	4.213632	463.5	23.175	463.5	23.175	884.8637	44.24318	42.13632	2.106816	884.8637	44.24318	505.6363	44.24318	421.3637	21.06818	505.6363	44.24318	445	44.5	482	48.2
170.4544	8.522719	1704.546	85.22728	170.4544	8.522719	937.5	46.875	937.5	46.875	1789.773	89.48864	85.22719	4.261359	1789.773	89.48864	1022.727	89.48864	852.2728	42.61364	1022.727	89.48864	1873	93.65	2	0.1
207.9089	10.39544	2079.091	103.9546	207.9089	10.39544	1143.5	57.175	1143.5	57.175	2183.046	109.1523	103.9544	5.197722	2183.046	109.1523	1247.454	109.1523	1039.546	51.97728	1247.454	109.1523	2286	114.3	1	0.05
242.8179	12.1409	2428.182	121.4091	242.8179	12.1409	1335.5	66.775	1335.5	66.775	2549.591	127.4796	121.409	6.070448	2549.591	127.4796	1456.909	127.4796	1214.091	60.70455	1456.909	127.4796	1200	60	1471	73.55
141.3635	7.068175	1413.637	70.68183	141.3635	7.068175	777.5	38.875	777.5	38.875	1484.318	74.21591	70.68175	3.534087	1484.318	74.21591	848.1817	74.21591	706.8183	35.34091	848.1817	74.21591	854	42.7	701	35.05
167.9998	8.399992	1680	84.00001	167.9998	8.399992	924	46.2	924	46.2	1764	88.2	83.99992	4.199996	1764	88.2	1008	88.2	840.0001	42	1008	88.2	1846	92.3	2	0.1
70.7272	3.53636	707.2728	35.36364	70.7272	3.53636	389	19.45	389	19.45	742.6364	37.13182	35.3636	1.768118	742.6364	37.13182	424.3636	37.13182	353.6364	17.68182	424.3636	37.13182	525	26.25	253	12.65
3.272724	0.490909	32.72728	4.909091	3.272724	0.490909	18	2.7	18	2.7	34.36364	5.154546	1.636362	0.245454	34.36364	5.154546	19.63636	5.154546	16.36364	2.454546	16.36364	2.454546	27	4.05	9	1.35
235.3634	11.76817	2353.637	117.6818	235.3634	11.76817	1294.5	64.725	1294.5	64.725	2471.318	123.5659	117.6817	5.884085	2471.318	123.5659	1412.182	123.5659	1176.818	58.84091	1412.182	123.5659	2589	129.45	0	0
92.909	9.2909	929.091	92.9091	92.909	9.2909	511	51.1	511	51.1	975.5455	97.55455	46.4545	4.64545	975.5455	97.55455	557.4545	97.55455	464.5455	46.45455	557.4545	97.55455	853	85.3	169	16.9
301.4542	15.07271	3014.546	150.7273	301.4542	15.07271	1658	82.9	1658	82.9	3165.273	158.2636	150.7271	7.536356	3165.273	158.2636	1808.727	158.2636	1507.273	75.36364	1808.727	158.2636	1	0.05	3315	165.75
283.9088	14.19544	2839.091	141.9546	283.9088	14.19544	1561.5	78.075	1561.5	78.075	2981.046	149.0523	141.9544	7.09772	2981.046	149.0523	1703.454	149.0523	1419.546	70.97728	1703.454	149.0523	1089	54.45	2034	101.7
90.09082	4.504541	900.9092	45.04546	90.09082	4.504541	495.5	24.775	495.5	24.775	945.9546	47.29773	45.04541	2.25227	945.9546	47.29773	540.5454	47.29773	450.4546	22.52273	540.5454	47.29773	990	49.5	1	0.05
56.63631	2.831815	566.3637	28.31818	56.63631	2.831815	311.5	15.575	311.5	15.575	594.6818	29.73409	28.31815	1.415908	594.6818	29.73409	339.8182	29.73409	283.1818	14.15909	339.8182	29.73409	603	30.15	20	1

APPENDIX C

Supplementary Table 3 Concordance results of the HPS-MPS assay testing for both MiSeq and Ion Torrent systems. Boxes in grey indicate assessment performed with read count thresholds in place for interpretation of genotype calls as to no grey indicating the same assessment but without read count thresholds in place for the respective system/assay

Concordance Samples			HPS-MPS-MiSeq						HPS-MPS-ION					
			Site 2			Site 3			Site 4 (Ion S5 - 530 chip)			Site 5		Site 6
			Genotypes using tools WITHOUT threshold						Genotypes using tools WITHOUT threshold					
Sample Name	Source Type	Concentration	Agree?	?/41	Agree?	?/41			Agree?	?/41	Agree?	?/41	Agree?	?/41
A1	Single	148 pg	98%	40	85%	35			100%	41	100%	41	98%	40
C1	Single	110 pg	85%	35	90%	37			100%	41	100%	41	100%	41
D1	Single	4474 pg	100%	41	93%	38			100%	41	100%	41	100%	41
H1	Single	135 pg	83%	34	90%	37			100%	41	100%	41	63%	26
B2	Single	302 pg	88%	36	78%	32			100%	41	98%	40	98%	40
D2	Single	152 pg	98%	40	83%	34			100%	41	100%	41	100%	41
E2	Mixture	2913 pg (10:1 ratio)	83%	34	76%	31			95%	39	85%	35	93%	38
B3	Single	105 pg	90%	37	88%	36			95%	39	100%	41	98%	40
C3	Single	20890 pg	95%	39	93%	38			100%	41	100%	41	100%	41
D3	Single	37 pg	85%	35	59%	24			98%	40	90%	37	100%	41
E3	Single	6 pg	54%	22	78%	32			68%	28	76%	31	76%	31
C4	Mixture	25400 pg	80%	33	66%	27			90%	37	88%	36	93%	38
D4	Single	32 pg	95%	39	56%	23			98%	40	100%	41	100%	41
F4	Single	3420 ng	98%	40	90%	37			100%	41	100%	41	100%	41
G4	Mixture	972 pg (1:1 ratio)	73%	30	66%	27			78%	32	71%	29	78%	32
H4	Mixture	873 pg (1:2 ratio)	54%	22	49%	20			61%	25	56%	23	56%	23
Sample Name	Source Type	Concentration	Genotypes using tools WITH read count thresholds						Genotypes using tools WITH read count thresholds					
A1	Single	148 pg	46%	19	66%	27			98%	40	100%	41	0%	0
C1	Single	110 pg	56%	23	44%	18			98%	40	100%	41	100%	41
D1	Single	4474 pg	41%	17	34%	14			100%	41	100%	41	100%	41
H1	Single	135 pg	59%	24	41%	17			98%	40	100%	41	0%	0
B2	Single	302 pg	39%	16	17%	7			98%	40	100%	41	100%	41
D2	Single	152 pg	54%	22	29%	12			100%	41	100%	41	100%	41
E2	Mixture	2913 pg (10:1 ratio)	37%	15	22%	9			90%	37	93%	38	93%	38
B3	Single	105 pg	27%	11	20%	8			98%	40	100%	41	100%	41
C3	Single	20890 pg	34%	14	34%	14			100%	41	100%	41	100%	41
D3	Single	37 pg	17%	7	12%	5			98%	40	95%	39	63%	26
E3	Single	6 pg	85%	35	78%	32			56%	23	61%	25	68%	28
C4	Mixture	25400 pg	29%	12	22%	9			88%	36	85%	35	88%	36
D4	Single	32 pg	24%	10	12%	5			95%	39	98%	40	100%	41
F4	Single	3420 ng	27%	11	20%	8			98%	40	100%	41	0%	0
G4	Mixture	972 pg (1:1 ratio)	17%	7	17%	7			68%	28	98%	40	100%	41
H4	Mixture	873 pg (1:2 ratio)	17%	7	12%	5			73%	30	73%	30	95%	39

APPENDIX D

Supplementary Table 4 Information on sample drop-out and % sequencing error (using counts of incorrect allele divided by total read count of both alleles) observed using the HPS-MPS pipeline per HPS variant per assay/platform using control DNA from 250 pg DNA down to 25 pg DNA input

HPS-MPS-MiSeq									
Average Genotype Standard Error									
9947A						9948			
Homozygous SNP	250pg	100pg	50pg	25pg		Homozygous SNP	250pg	100pg	50pg	25pg
rs796296176	0.13%	0.00%	0.00%	0.33%		rs796296176	0.00%	0.00%	0.00%	0.00%
rs11547464	0.07%	0.65%	0.20%	0.11%		rs11547464	0.11%	0.16%	0.00%	0.13%
rs885479	0.00%	0.00%	0.20%	0.00%		rs885479	0.00%	0.03%	0.00%	0.00%
rs1805008	0.07%	0.00%	0.00%	0.00%		rs1805008	0.06%	0.03%	0.06%	0.00%
rs1805005	0.00%	0.00%	0.07%	50.50%		rs1805005	0.11%	0.00%	0.10%	0.00%
rs1805006	0.00%	0.00%	0.00%	50.00%		rs1805006	0.17%	0.00%	0.00%	0.00%
rs1805007	0.00%	0.00%	0.11%	0.00%		rs1805007	0.08%	0.07%	0.00%	0.00%
rs1805009	0.15%	0.14%	0.19%	0.28%		rs1805009	0.12%	3.71%	0.08%	0.12%
rs201326893	0.00%	0.18%	0.11%	0.11%		rs201326893	0.00%	0.08%	0.00%	0.00%
rs2228479	0.17%	0.00%	0.00%	50.00%		rs2228479	0.12%	0.03%	0.10%	0.00%
rs1110400	0.00%	0.00%	0.11%	0.00%		rs1110400	0.12%	0.07%	0.00%	0.00%
rs28777	0.00%	50.00%	0.00%	0.00%		rs28777	0.00%	0.00%	0.00%	0.00%
rs16891982	0.00%	0.00%	0.00%	0.00%		rs16891982	0.00%	0.00%	0.06%	0.00%
rs12821256	0.00%	0.07%	0.00%	0.75%		rs12821256	-	-	-	-
rs4959270	-	-	-	-		rs4959270	-	-	-	-
rs12203592	-	-	-	-		rs12203592	-	-	-	-
rs1042602	0.10%	0.00%	0.00%	0.00%		rs1042602	0.02%	0.00%	0.00%	0.02%
rs1800407	0.02%	0.00%	0.00%	0.00%		rs1800407	0.11%	0.11%	0.09%	0.19%
rs2402130	0.06%	0.03%	0.04%	0.04%		rs2402130	1.31%	0.00%	0.05%	0.06%
rs12913832	0.00%	0.42%	0.14%	0.10%		rs12913832	3.10%	0.04%	0.65%	0.12%
rs2378249	-	-	-	-		rs2378249	0.02%	0.00%	0.00%	0.00%
rs12896399	0.00%	0.00%	0.00%	9.24%		rs12896399	-	-	-	-
rs1393350	0.00%	50.00%	0.00%	0.00%		rs1393350	0.15%	0.05%	0.33%	0.00%
rs683	0.00%	0.00%	0.00%	0.00%		rs683	0.00%	0.03%	0.00%	0.25%
rs3114908	1.84%	0.00%	1.98%	0.07%		rs3114908	2.29%	0.04%	0.03%	0.19%
rs1800414	0.05%	0.00%	0.13%	0.00%		rs1800414	0.04%	0.02%	0.12%	0.09%
rs10756819	0.00%	48.02%	9.26%	10.48%		rs10756819	-	-	-	-
rs2238289	0.05%	50.05%	0.00%	0.00%		rs2238289	0.71%	0.00%	0.33%	0.00%
rs17128291	-	-	-	-		rs17128291	-	-	-	-
rs6497292	0.00%	0.07%	0.06%	0.00%		rs6497292	0.03%	0.04%	0.10%	0.05%
rs1129038	0.03%	50.00%	0.00%	0.00%		rs1129038	0.14%	0.00%	0.20%	0.00%
rs1667394	0.14%	0.03%	0.00%	0.10%		rs1667394	0.06%	0.07%	0.00%	0.46%
rs1126809	2.78%	26.34%	3.87%	2.05%		rs1126809	4.00%	0.09%	9.43%	0.00%
rs1470608	0.00%	0.00%	0.00%	0.00%		rs1470608	-	-	-	-
rs1426654	0.00%	50.00%	0.00%	0.00%		rs1426654	0.00%	0.00%	0.00%	0.00%
rs6119471	1.42%	0.00%	0.00%	0.00%		rs6119471	0.00%	0.00%	0.00%	0.08%
rs1545397	0.20%	0.07%	0.09%	0.25%		rs1545397	0.64%	0.08%	0.11%	0.00%
rs6059655	-	-	-	-		rs6059655	0.06%	0.00%	0.00%	0.55%
rs12441727	0.02%	0.00%	0.09%	0.29%		rs12441727	-	-	-	-
rs3212355	0.00%	50.00%	0.00%	0.00%		rs3212355	0.00%	0.00%	0.00%	0.00%
rs8051733	-	-	-	-		rs8051733	0.60%	2.31%	0.00%	0.00%
Average Error	0.21%	10.75%	0.48%	4.99%		Average Error	0.43%	0.21%	0.36%	0.07%

Supplementary Table 4 Continued

HPS-MPS-ION									
Average Genotype Standard Error									
9947A					9948				
Homozygous SNP	250pg	100pg	50pg	25pg	Homozygous SNP	250pg	100pg	50pg	25pg
rs796296176	0.00%	0.00%	0.00%	0.00%	rs796296176	0.00%	0.00%	0.00%	0.00%
rs11547464	0.00%	0.00%	0.00%	0.00%	rs11547464	0.00%	0.00%	0.00%	0.00%
rs885479	0.00%	0.40%	0.00%	0.00%	rs885479	0.00%	0.00%	0.00%	0.00%
rs1805008	0.00%	0.00%	0.00%	0.00%	rs1805008	0.00%	0.00%	0.00%	0.00%
rs1805005	0.00%	0.00%	0.00%	0.00%	rs1805005	0.00%	0.00%	0.00%	0.00%
rs1805006	0.00%	0.00%	0.00%	0.00%	rs1805006	0.00%	0.00%	0.00%	0.00%
rs1805007	0.00%	0.00%	0.34%	0.00%	rs1805007	0.00%	0.00%	0.00%	0.00%
rs1805009	0.11%	0.00%	0.00%	0.00%	rs1805009	0.00%	0.00%	0.00%	0.00%
rs201326893	0.00%	0.00%	0.00%	0.00%	rs201326893	0.00%	0.00%	0.00%	0.00%
rs2228479	0.00%	0.00%	0.00%	0.00%	rs2228479	0.00%	0.00%	0.00%	0.81%
rs1110400	0.13%	0.00%	0.00%	0.00%	rs1110400	0.00%	0.00%	0.00%	0.00%
rs28777	0.00%	0.00%	0.00%	0.00%	rs28777	0.08%	0.00%	0.00%	1.96%
rs16891982	0.00%	3.24%	0.00%	0.00%	rs16891982	0.00%	0.00%	0.60%	0.00%
rs12821256	0.19%	0.00%	0.00%	0.00%	rs12821256	-	-	-	-
rs4959270	-	-	-	-	rs4959270	-	-	-	-
rs12203592	-	-	-	-	rs12203592	-	-	-	-
rs1042602	0.00%	0.28%	0.00%	4.17%	rs1042602	0.00%	0.47%	0.00%	0.00%
rs1800407	0.00%	0.00%	0.00%	0.00%	rs1800407	0.00%	0.00%	0.36%	0.00%
rs2402130	0.16%	0.00%	0.00%	0.00%	rs2402130	0.15%	0.00%	0.00%	5.26%
rs12913832	0.00%	0.00%	0.00%	0.00%	rs12913832	0.13%	0.00%	0.00%	0.00%
rs2378249	-	-	-	-	rs2378249	0.00%	0.00%	0.00%	0.00%
rs12896399	0.19%	0.00%	0.00%	0.00%	rs12896399	-	-	-	-
rs1393350	0.00%	0.00%	0.00%	0.00%	rs1393350	0.20%	0.00%	0.00%	0.00%
rs683	0.00%	0.00%	0.00%	0.00%	rs683	0.00%	0.00%	12.70%	0.00%
rs3114908	0.00%	0.00%	0.00%	0.00%	rs3114908	0.96%	0.74%	0.00%	3.85%
rs1800414	0.14%	0.16%	0.00%	0.00%	rs1800414	0.00%	0.00%	0.00%	0.00%
rs10756819	0.00%	0.00%	0.00%	0.00%	rs10756819	-	-	-	-
rs2238289	0.10%	0.00%	1.23%	94.12%	rs2238289	0.31%	0.00%	0.80%	29.33%
rs17128291	-	-	-	-	rs17128291	-	-	-	-
rs6497292	0.44%	0.00%	0.59%	0.00%	rs6497292	0.00%	0.00%	0.00%	0.00%
rs1129038	0.35%	0.19%	0.00%	0.00%	rs1129038	0.00%	0.00%	0.00%	6.45%
rs1667394	0.00%	0.00%	0.00%	0.00%	rs1667394	0.15%	0.00%	0.00%	0.00%
rs1126809	0.17%	0.97%	0.89%	0.00%	rs1126809	0.14%	0.25%	0.00%	0.00%
rs1470608	0.20%	0.00%	0.00%	0.00%	rs1470608	-	-	-	-
rs1426654	0.33%	0.00%	0.00%	0.00%	rs1426654	0.00%	0.23%	0.00%	0.00%
rs6119471	0.15%	0.00%	0.00%	0.00%	rs6119471	0.00%	0.00%	0.00%	0.00%
rs1545397	0.00%	0.00%	0.00%	0.00%	rs1545397	0.00%	0.00%	0.00%	0.00%
rs6059655	-	-	-	-	rs6059655	0.00%	0.25%	0.00%	0.00%
rs12441727	0.00%	0.00%	0.00%	0.00%	rs12441727	-	-	-	-
rs3212355	0.00%	0.00%	0.00%	0.00%	rs3212355	0.00%	0.00%	0.00%	0.00%
rs8051733	-	-	-	-	rs8051733	0.00%	0.00%	0.37%	0.00%
Average Error	0.08%	0.15%	0.09%	2.81%	Average Error	0.06%	0.06%	0.45%	1.44%

APPENDIX E

Supplementary Table 5 Read count thresholds per variant for both HPS-MPS-MiSeq and HPS-MPS-ION assays for correct genotype calls at 100pg and 50pg DNA input

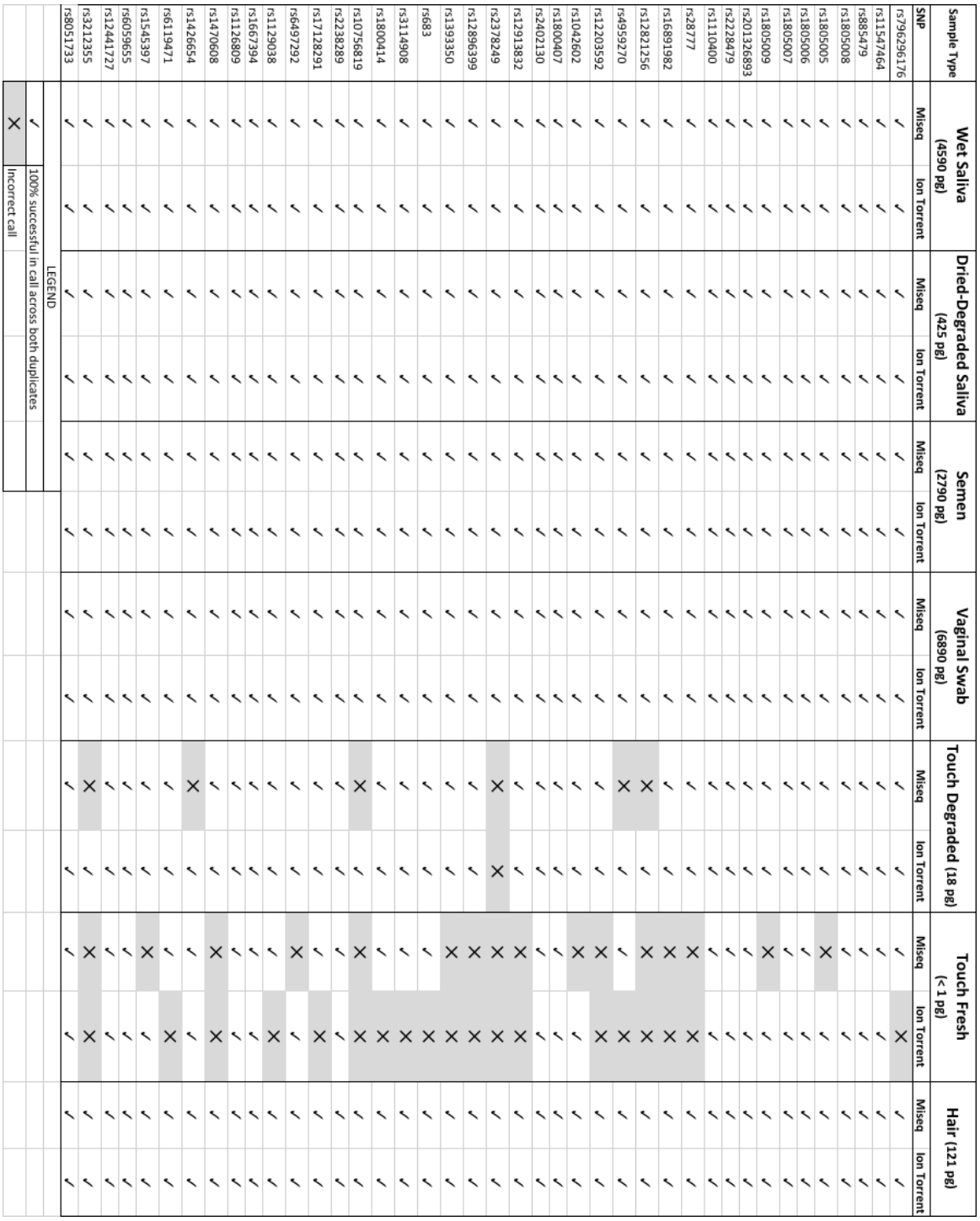
HPS-MPS-MiSeq Workflow				
Mixture or Single Source using the tool?				
If mixture, can you deconvolute with mixture tool?				
If single source, what is the concentration?				
	If >100pg DNA input*		If < 100 pg: use 50pg read count info*	
SNP	Average (n=4)	Confidence Range	Average (n=4)	Confidence Range
rs312262906	814	>479	310	>96
rs11547464	1160	>558	483	>248
rs885479	1161	>560	483	>249
rs1805008	1161	>560	483	>250
rs1805005	1244	>567	584	>201
rs1805006	1241	>565	584	>201
rs1805007	1161	>560	483	>250
rs1805009	2365	>1656	997	>754
rs201326893	1161	>558	483	>250
rs2228479	1241	>566	583	>201
rs1110400	1159	>559	483	>250
rs28777	850	>582	337	>207
rs16891982	2573	>2060	1159	>763
rs12821256	934	>698	434	>166
rs4959270	2036	>1267	729	>590
rs12203592	2313	>1294	685	>439
rs1042602	2261	>1534	1152	>424
rs1800407	2166	>1240	1452	>724
rs2402130	3550	>2920	2464	>896
rs12913832	1474	>1277	598	>325
rs2378249	2616	>1977	1128	>698
rs12896399	979	>806	294	>305
rs1393350	1724	>1427	469	>71
rs683	1080	>648	694	>293
rs3114908	2344	>882	1488	>607
rs1800414	2171	>1439	1429	>615
rs10756819	2488	>1942	998	>1392
rs2238289	1187	>948	590	>86
rs17128291	2356	>1666	1549	>813
rs6497292	1780	>1270	1194	>515
rs1129038	1534	>767	780	>230
rs1667394	1797	>1442	1054	>611
rs1126809	2298	>1897	720	>256
rs1470608	1336	>813	641	>369
rs1426654	28	>24	29	>9
rs6119471	2059	>1395	1002	>510
rs1545397	1441	>1308	872	>375
rs6059655	807	>662	380	>316
rs12441727	3286	>2399	1766	>654
rs3212355	191	>40	211	>13
rs8051733	367	>227	232	>82

Supplementary Table 5 Continued

HPS-MPS-ION Workflow				
Mixture or Single Source using the tool?				
If mixture, can you deconvolute with mixture tool?				
If single source, what is the concentration?				
	If >100pg DNA input*		If < 100 pg: use 50pg read count info*	
SNP	Average	Confidence Range	Average	Confidence Range
rs312262906	346	>291	243	>159
rs11547464	272	>249	232	>172
rs885479	269	>246	227	>172
rs1805008	271	>249	230	>172
rs1805005	385	>308	330	>233
rs1805006	392	>318	339	>238
rs1805007	271	>249	231	>173
rs1805009	301	>266	253	>205
rs201326893	270	>249	231	>172
rs2228479	393	>320	336	>236
rs1110400	270	>248	231	>172
rs28777	485	>307	280	>216
rs16891982	261	>179	104	>41
rs12821256	220	>215	51	>38
rs4959270	228	>201	81	>62
rs12203592	303	>206	104	>87
rs1042602	282	>210	113	>96
rs1800407	332	>288	182	>87
rs2402130	332	>286	212	>184
rs12913832	446	>391	109	>94
rs2378249	190	>182	87	>41
rs12896399	284	>284	174	>101
rs1393350	274	>189	88	>64
rs683	280	>254	145	>110
rs3114908	227	>187	157	>118
rs1800414	543	>473	145	>127
rs10756819	210	>203	63	>39
rs2238289	323	>266	102	>80
rs17128291	394	>323	99	>64
rs6497292	332	>263	225	>168
rs1129038	457	>401	230	>159
rs1667394	257	>214	146	>145
rs1126809	406	>404	194	>111
rs1470608	273	>252	80	>50
rs1426654	315	>187	181	>180
rs6119471	333	>284	206	>154
rs1545397	129	>104	32	>25
rs6059655	423	>402	320	>171
rs12441727	338	>311	218	>206
rs3212355	321	>316	183	>150
rs8051733	426	>338	315	>268

APPENDIX F

Supplementary Table 6 The performance of both HPS-MPS-MiSeq and HPS-MPS-ION assays using the HPS-MPS pipeline on simulated casework. A tick indicates a correct call was made for the variant for that sample. An X indicates an incorrect genotype call was made or there was drop out for that variant in that sample



APPENDIX G

Supplementary Table 7 Observed heterozygote call ratios per variant (i.e. a GA genotype with 100 read counts of allele G and 100 read counts of allele A would give a 50:50 heterozygote call ratio) using the HPS-MPS-MiSeq or the HPS-MPS-ION assay, and HPS-MPS pipeline. Only variants with a heterozygote genotype found in the 8 individuals used is this assessment are represented

			HPS-MPS-MiSeq		HPS-MPS-ION	
SNP	REF	ALT	REF %	ALT %	REF %	ALT %
rs796296176	C	A	52	48	44	56
rs11547464	G	A	53	47	42	58
rs885479	G	A	57	43	52	48
rs1805008	C	T	51	49	59	41
rs1805005	G	T	50	50	47	53
rs1805006	C	A	-	-	-	-
rs1805007	C	T	-	-	-	-
rs1805009	G	C	53	47	52	48
rs201326893	C	A	-	-	-	-
rs2228479	G	A	56	44	53	47
rs1110400	T	C	-	-	-	-
rs28777	C	A	47	53	52	48
rs16891982	C	G	48	52	44	56
rs12821256	T	C	-	-	-	-
rs4959270	C	A	53	47	52	48
rs12203592	C	T	-	-	-	-
rs1042602	C	A	50	50	52	48
rs1800407	C	T	46	54	48	52
rs2402130	G	A	50	50	50	50
rs12913832	A	G	46	54	51	49
rs2378249	G	A	-	-	-	-
rs12896399	G	T	48	52	53	47
rs1393350	G	A	53	47	50	50
rs683	C	A	51	49	50	50
rs3114908	T	C	50	50	47	53
rs1800414	T	C	50	50	45	55
rs10756819	G	A	51	49	48	52
rs2238289	A	G	-	-	-	-
rs17128291	A	G	51	49	45	55
rs6497292	A	G	48	52	49	51
rs1129038	C	T	49	51	52	48
rs1667394	C	T	46	54	48	52
rs1126809	G	A	51	49	49	51
rs1470608	G	T	52	48	58	42
rs1426654	A	G	71	29	48	52
rs6119471	C	G	-	-	-	-
rs1545397	A	T	61	39	54	46
rs6059655	A	G	-	-	-	-
rs12441727	G	A	52	48	45	55
rs3212355	C	T	-	-	-	-
rs8051733	A	G	52	48	46	54

APPENDIX H

Supplementary Table 8 Mixture separation performance per variant on simulated 2-person mixture ratios using the Threshold & Mixture Deconvolution Tool (Supplementary Table 2). A tick represents a correct call for all samples. A yellow triangle represents caution at this variant as there were correct and incorrect genotype calls generated. A red X indicates incorrect genotypes for all samples and duplicates at that ratio

Mixture Assessment Using the Threshold & Mixture Deconvolution Tool (Both Platforms)					
SNP	1:10 (0.5ng : 5ng) ratio	1:10 (0.1ng : 1ng) ratio	1:5 (0.1ng : 0.5 ng) ratio	1:2 (0.5ng : 1ng) ratio	1:1 (1ng : 1ng) ratio
rs312262906	✓	▲	▲	▲	▲
rs11547464	✓	✓	✓	✓	✓
rs885479	✓	✓	✓	✓	✓
rs1805008	✓	✓	✓	✓	✓
rs1805005*	✗	▲	▲	✓	▲
rs1805006	✓	✓	✓	✓	✓
rs1805007	✓	✓	✓	✓	✓
rs1805009	✓	✓	✓	✓	✓
rs201326893	✓	✓	✓	✓	✓
rs2228479	✓	✓	✓	✓	✓
rs1110400	✓	✓	✓	✓	✓
rs28777	✓	✓	✓	✓	✓
rs16891982	✓	✓	✓	✓	✓
rs12821256	✓	✓	✓	✓	✓
rs4959270*	✓	✗	✗	▲	▲
rs12203592	✓	✓	✓	✓	✓
rs1042602	✓	✓	✓	✓	✓
rs1800407	✓	✓	✓	✓	✓
rs2402130*	✓	✗	✗	▲	▲
rs12913832	✓	✓	✓	✓	✓
rs2378249	✓	✓	✓	✓	✓
rs12896399	✓	▲	✗	✓	▲
rs1393350	✓	✓	▲	✓	▲
rs683	✓	▲	▲	✓	▲
rs3114908	✓	▲	▲	✓	▲
rs1800414	✓	✓	✓	✓	✓
rs10756819	✓	▲	✓	▲	▲
rs2238289	✓	✓	✓	✓	✓
rs17128291	✓	✓	✓	✓	▲
rs6497292	✓	✓	✓	✓	✓
rs1129038	✓	✓	✓	✓	✓
rs1667394	✓	✓	✓	✓	✓
rs1126809	✓	✓	▲	▲	▲
rs1470608	✓	✓	✓	✓	✓
rs1426654	✓	✓	✓	✓	✓
rs6119471	✓	✓	✓	✓	✓
rs1545397	✓	✓	✓	✓	✓
rs6059655	✓	✓	✓	✓	✓
rs12441727	✓	✓	✓	✓	✓
rs3212355	✓	✓	✓	✓	✓
rs8051733	✓	✗	✓	✓	▲
Total n = 40 samples					
✓	Correct genotypes called for both major and minor for all samples and duplicates				
▲	Be cautious: incorrect and correct genotypes called for samples and their duplicates				
✗	Incorrect genotypes called for all samples and duplicates at this DNA input ratio				

APPENDIX I

Supplementary Table 9 Results of the performance of the HPS-MPS-MiSeq and HPS-MPS-ION assays and the HPS-MPS pipeline on several non-human DNA samples, cat, dog, mouse, pig and primate. An X indicates no read counts were observed

	Cat		Dog		Mouse		Pig		Chimp	
SNP	Miseq	ION	Miseq	ION	Miseq	ION	Miseq	ION	Miseq	ION
rs796296176	1	X	X	4	144	X	X	X	156	1366
rs11547464	1780	X	X	34	2	37	332	42	12	501
rs885479	1781	X	X	31	2	36	333	40	12	505
rs1805008	1782	X	X	34	2	36	333	38	11	506
rs1805005	177	9	X	1	1	34	1	5	216	1515
rs1805006	177	9	X	1	1	36	X	6	216	1552
rs1805007	1781	X	X	34	2	37	333	43	12	506
rs1805009	461	12	X	8	1261	2	3719	290	549	1482
rs201326893	1783	X	X	34	2	37	333	41	12	506
rs2228479	178	9	X	1	1	36	X	6	214	1551
rs1110400	1777	X	X	33	2	37	331	41	12	501
rs28777	1	1	1	1	299	42	X	X	570	2045
rs16891982	X	1	X	10	416	X	3	12	2341	768
rs12821256	269	X	X	X	1	6	224	X	1036	554
rs4959270	1226	X	X	X	251	X	411	1	1696	520
rs12203592	391	23	7999	X	5	44	11	1	2705	630
rs1042602	37	1	X	1424	75	36	135	487	2678	726
rs1800407	545	1	X	3	337	23	3	12	2702	42
rs2402130	3	X	X	3	686	1	221	X	2387	2458
rs12913832	X	X	X	1	1854	14	942	4	4599	1110
rs2378249	127	X	X	X	408	1	2	X	2143	456
rs12896399	193	X	X	X	83	4	1	X	388	867
rs1393350	294	X	X	6	263	1	2	X	X	5
rs683	X	49	X	X	1	X	X	X	1190	695
rs3114908	1	X	X	4	234	X	131	X	3719	817
rs1800414	1	2	X	X	98	4	2	X	2972	1446
rs10756819	42	1	X	7	1004	1	519	X	4835	620
rs2238289	161	1	X	4	179	3	54	X	136	583
rs17128291	X	X	X	2	139	1	81	X	5263	1758
rs6497292		X	X	1	3	2	5	1	2561	434
rs1129038	X	1	X	1	481	X	3	X	846	1375
rs1667394	1	3	X	2	81	X	6	X	2798	794
rs1126809	2	1	X	X	1034	7	132	1	3409	1544
rs1470608	X	X	X	X	4	X	X	1	7	706
rs1426654	X	1	X	481	X	X	1	7	54	1228
rs6119471	406	2	X	6	92	21	43	X	3518	1352
rs1545397	118	X	X	X	211	3	4	X	1764	102
rs6059655	X	X	X	X	470	X	7	X	1533	1368
rs12441727	422	3	X	1	97	2	405	X	2635	888
rs3212355	X	1	X	X	2	2	X	X	65	1008
rs8051733	X	1	X	X	80	1	51	X	X	1249

APPENDIX J

Supplementary Table 10 Results of the performance of the HPS-MPS-MiSeq and HPS-MPS-ION assays and the HPS-MPS pipeline on UV degraded samples. A grey box indicates no read counts were observed. Degradation was assessed by utilizing 500pg DNA exposed to UV light for time intervals of 0, 5, 10, and 20 min, using CL-1000 Ultraviolet Crosslinker (Ultra-Violet Products Ltd) at a strength of 50 J/cm2. An X indicates no read counts were observed

SNP	UV Degraded Sample Coverage			
Timepoints	0 minutes	5 minutes	10 minutes	20 minute
	Miseq	Miseq	Miseq	Miseq
rs796296176	579	1957	2146	48
rs11547464	1058	4529	5001	2
rs885479	1060	4513	4992	2
rs1805008	1057	4530	5019	2
rs1805005	618	1774	1565	1
rs1805006	628	1774	1566	1
rs1805007	1060	4531	5003	2
rs1805009	2406	3758	3973	645
rs201326893	1060	4532	5006	2
rs2228479	618	1753	1548	1
rs1110400	1061	4530	5015	2
rs28777	413	490	388	
rs16891982	1199	1375	598	6
rs12821256	481	388	202	1
rs4959270	905	953	834	
rs12203592	2525	1733	741	2
rs1042602	1519	1344	1190	1
rs1800407	1631	2935	2989	279
rs2402130	1562	2418	1820	2
rs12913832	2972	3976	3881	245
rs2378249	1318	1172	1009	77
rs12896399	431	712	327	
rs1393350	1241	1393	1250	1
rs683	523	392	203	1
rs3114908	2531	3095	2990	2
rs1800414	1596	918	1120	6
rs10756819	3312	3607	2546	87
rs2238289	1172	1040	937	64
rs17128291	1379	823	575	2
rs6497292	1585	3642	3376	41
rs1129038	1952	2998	2136	2
rs1667394	1412	1180	782	1
rs1126809	2638	2374	1017	5
rs1470608	497	321	121	1
rs1426654	33	21	21	
rs6119471	2987	4379	4302	608
rs1545397	998	889	673	46
rs6059655	2480	4128	3514	1
rs12441727	1870	2118	2152	126
rs3212355	289	478	603	
rs8051733	426	419	549	18
Average Coverage	**1343.46**	**2192.49**	**2040.98**	**64.81**

APPENDIX K

SUPPLEMENTARY MATERIAL 1

Guide to using the threshold & mixture deconvolution tool (*two-person mixture only)

Please consult the flowchart found in Figure 6 (Figure 4 in manuscript) for the order in which to use this tool.

1. Inputting data into the tool.

As part of the HPS-MPS pipeline, every sample has its own folder of generated data found in the runfolder (~/Desktop/hps/runfolder). However for a summary of the result of each sample in one folder, a .csv file is created and this can be found in the TableFiles folder of the hps folder. Please go to this location ~/Desktop/hps/TableFiles/SAMPLE_NAME.csv

Open this .csv file with excel to view read counts of the HPS variants. Your file will look like this.

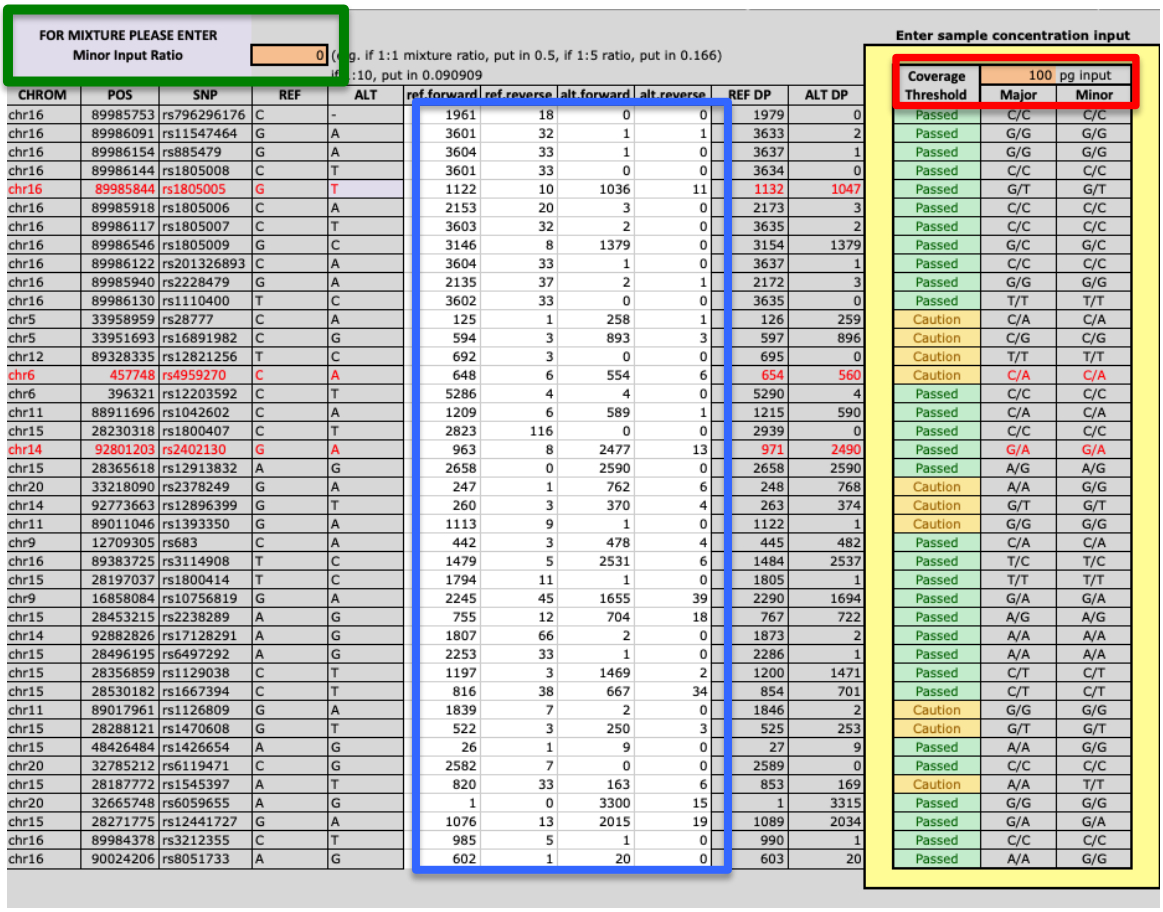
	A	B	C	D	E	F	G	H	I	J	K
1	CHROM	POS	SAMPLE	REF	ALT	GT	ref.forward	ref.reverse	alt.forward	alt.reverse	
2	chr16	89985753	9947A-500p-	C	.	C/C	112	1	0	0	
3	chr16	89986091	9947A-500p-	G	.	G/G	212	3	0	0	
4	chr16	89986154	9947A-500p-	G	.	G/G	212	3	0	0	
5	chr16	89986144	9947A-500p-	C	.	C/C	212	3	0	0	
6	chr16	89985844	9947A-500p-	G	.	G/G	186	1	1	0	
7	chr16	89985918	9947A-500p-	C	.	C/C	187	1	0	0	
8	chr16	89986117	9947A-500p-	C	.	C/C	212	3	0	0	
9	chr16	89986546	9947A-500p-	G	.	G/G	1359	7	0	0	
10	chr16	89986122	9947A-500p-	C	.	C/C	211	3	1	0	
11	chr16	89985940	9947A-500p-	G	.	G/G	186	2	0	0	
12	chr16	89986130	9947A-500p-	T	.	T/T	212	3	0	0	
13	chr5	33958959	9947A-500p-	C	A	A/A	0	0	546	2	
14	chr5	33951693	9947A-500p-	C	G	G/G	0	0	1494	5	
15	chr12	89328335	9947A-500p-	T	.	T/T	605	3	0	0	
16	chr6	457748	9947A-500p-	C	A	C/A	513	9	589	8	
17	chr6	396321	9947A-500p-	C	T	C/T	779	6	773	8	
18	chr11	88911696	9947A-500p-	C	.	C/C	2100	8	0	1	
19	chr15	28230318	9947A-500p-	C	.	C/C	2151	93	10	0	
20	chr14	92801203	9947A-500p-	G	A	A/A	2	0	2105	25	
21	chr15	28365618	9947A-500p-	A	G	G/G	31	0	1289	4	
22	chr20	33218090	9947A-500p-	G	A	G/A	576	5	771	7	
23	chr14	92773663	9947A-500p-	G	T	T/T	0	0	719	2	
24	chr11	89011046	9947A-500p-	G	A	A/A	0	0	1064	11	
25	chr9	12709305	9947A-500p-	C	A	A/A	0	0	867	11	
26	chr16	89383725	9947A-500p-	T	C	C/C	12	0	1708	4	
27	chr15	28197037	9947A-500p-	T	.	T/T	1634	7	2	0	
28	chr9	16858084	9947A-500p-	G	A	A/A	123	4	2675	48	
29	chr15	28453215	9947A-500p-	A	.	A/A	1103	13	0	0	
30	chr14	92882826	9947A-500p-	A	G	A/G	1142	37	558	17	
31	chr15	28496195	9947A-500p-	A	.	A/A	1362	12	1	0	
32	chr15	28356859	9947A-500p-	C	T	T/T	0	0	1195	3	
33	chr15	28530182	9947A-500p-	C	T	T/T	0	0	1326	52	
34	chr11	89017961	9947A-500p-	G	A	A/A	87	1	1722	4	
35	chr15	28288121	9947A-500p-	G	.	G/G	742	4	1	0	
36	chr15	48426484	9947A-500p-	A	.	A/A	27	3	0	0	
37	chr20	32785212	9947A-500p-	C	.	C/C	2079	5	43	0	
38	chr15	28187772	9947A-500p-	A	.	A/A	951	57	9	0	
39	chr20	32665748	9947A-500p-	A	G	A/G	227	0	355	3	
40	chr15	28271775	9947A-500p-	G	.	G/G	1849	23	1	0	
41	chr16	89984378	9947A-500p-	C	.	C/C	18	0	0	0	
42	chr16	90024206	9947A-500p-	A	G	A/G	164	2	215	0	

Copy and paste the items in the red box from the above image (in your sample csv file) into your Threshold & Mixture Tool below (to the highlighted blue box in the below image).

Be sure to use the MiSeq version if you used the HPS-MPS-MiSeq assay.

Be sure to use the ION version if you used the HPS-MPS-ION assay.

2. Threshold Guidelines



FOR MIXTURE PLEASE ENTER
Minor Input Ratio: 0 (e.g. if 1:1 mixture ratio, put in 0.5, if 1:5 ratio, put in 0.166)
if 1:10, put in 0.090909

Enter sample concentration input

CHROM	POS	SNP	REF	ALT	ref forward	ref reverse	alt forward	alt reverse	REF DP	ALT DP	Coverage Threshold	100 pg input Major	Minor
chr16	89985753	rs796296176	C	-	1961	18	0	0	1979	0	Passed	C/C	C/C
chr16	89986091	rs11547464	G	A	3601	32	1	1	3633	2	Passed	G/G	G/G
chr16	89986154	rs885479	G	A	3604	33	1	0	3637	1	Passed	G/G	G/G
chr16	89986144	rs1805008	C	T	3601	33	0	0	3634	0	Passed	C/C	C/C
chr16	89985844	rs1805005	G	T	1122	10	1036	11	1132	1047	Passed	G/T	G/T
chr16	89985918	rs1805006	C	A	2153	20	3	0	2173	3	Passed	C/C	C/C
chr16	89986117	rs1805007	C	T	3603	32	2	0	3635	2	Passed	C/C	C/C
chr16	89986546	rs1805009	G	C	3146	8	1379	0	3154	1379	Passed	G/C	G/C
chr16	89986122	rs201326893	C	A	3604	33	1	0	3637	1	Passed	C/C	C/C
chr16	89985940	rs2228479	G	A	2135	37	2	1	2172	3	Passed	G/G	G/G
chr16	89986130	rs1110400	T	C	3602	33	0	0	3635	0	Passed	T/T	T/T
chr5	33958959	rs28777	C	A	125	1	258	1	126	259	Caution	C/A	C/A
chr5	33951693	rs16891982	C	G	594	3	893	3	597	896	Caution	C/G	C/G
chr12	89328335	rs12821256	T	C	692	3	0	0	695	0	Caution	T/T	T/T
chr6	457748	rs4959270	C	A	648	6	554	6	654	560	Caution	C/A	C/A
chr6	396321	rs12203592	C	T	5286	4	4	0	5290	4	Passed	C/C	C/C
chr11	88911696	rs1042602	C	A	1209	6	589	1	1215	590	Passed	C/A	C/A
chr15	28230318	rs1800407	C	T	2823	116	0	0	2939	0	Passed	C/C	C/C
chr14	92801203	rs2402130	G	A	963	8	2477	13	971	2490	Passed	G/A	G/A
chr15	28365618	rs12913832	A	G	2658	0	2590	0	2658	2590	Passed	A/G	A/G
chr20	33218090	rs2378249	G	A	247	1	762	6	248	768	Caution	A/A	G/G
chr14	92773663	rs12896399	G	T	260	3	370	4	263	374	Caution	G/T	G/T
chr11	89011046	rs1393350	G	A	1113	9	1	0	1122	1	Caution	G/G	G/G
chr9	12709305	rs683	C	A	442	3	478	4	445	482	Passed	C/A	C/A
chr16	89383725	rs3114908	T	C	1479	5	2531	6	1484	2537	Passed	T/C	T/C
chr15	28197037	rs1800414	T	C	1794	11	1	0	1805	1	Passed	T/T	T/T
chr9	16858084	rs10756819	G	A	2245	45	1655	39	2290	1694	Passed	G/A	G/A
chr15	28453215	rs2238289	A	G	755	12	704	18	767	722	Passed	A/G	A/G
chr14	92882826	rs17128291	A	G	1807	66	2	0	1873	2	Passed	A/A	A/A
chr15	28496195	rs6497292	A	G	2253	33	1	0	2286	1	Passed	A/A	A/A
chr15	28356859	rs1129038	C	T	1197	3	1469	2	1200	1471	Passed	C/T	C/T
chr15	28530182	rs1667394	C	T	816	38	667	34	854	701	Passed	C/T	C/T
chr11	89017961	rs1126809	G	A	1839	7	2	0	1846	2	Caution	G/G	G/G
chr15	28288121	rs1470608	G	T	522	3	250	3	525	253	Caution	G/T	G/T
chr15	48426484	rs1426654	A	G	26	1	9	0	27	9	Passed	A/A	G/G
chr20	32785212	rs6119471	C	G	2582	7	0	0	2589	0	Passed	C/C	C/C
chr15	28187772	rs1545397	A	T	820	33	163	6	853	169	Caution	A/A	T/T
chr20	32665748	rs6059655	A	G	1	0	3300	15	1	3315	Passed	G/G	G/G
chr15	28271775	rs12441727	G	A	1076	13	2015	19	1089	2034	Passed	G/A	G/A
chr16	89984378	rs3212355	C	T	985	5	1	0	990	1	Passed	C/C	C/C
chr16	90024206	rs8051733	A	G	602	1	20	0	603	20	Passed	A/A	G/G

As soon as you paste the read counts of your file into the tool, you will see green/orange cells in color in the yellow region of the file. *For this assessment, please make sure the Minor Input Ratio highlighted in a green box above is set to 0 as you are not assessing mixtures at this point.*

There is a two-step process to assess the confidence of genotypes using the threshold guidelines once you have pasted in the read counts at the 100 pg input level as noted above in the red box. If your variant passed the threshold set for your sequencer at the 100 pg level, it will highlight the variant in green and call it “Passed”. This genotype may be trusted as reaching the acceptable levels based on 100 pg sensitivity testing of the system and that variant does not need further assessment.

If the variant is highlighted in orange with “Caution”, this variant must go through the second check of threshold. For this, you must edit the input level to 50 pg in the red box above.

When you edit to 50, you will see colors change in the boxes. If variant passed all of the 50 pg DNA level sensitivity thresholds for that sequencer, then the variant will stay orange with “Caution”. (To note, you will also see your green cell colors go to orange, this makes sense as they too have already passed the 50 pg level input). Below you can see the very same example as above but using the 50 pg level:

Enter sample concentration input

Coverage Threshold	50 pg input	
	Major	Minor
Caution	C/C	C/C
Caution	G/G	G/G
Caution	G/G	G/G
Caution	C/C	C/C
Caution	G/T	G/T
Caution	C/C	C/C
Caution	C/C	C/C
Caution	G/C	G/C
Caution	C/C	C/C
Caution	G/G	G/G
Caution	T/T	T/T
Caution	C/A	C/A
Caution	C/G	C/G
Caution	T/T	T/T
Caution	C/A	C/A
Caution	C/C	C/C
Caution	C/A	C/A
Caution	C/C	C/C
Caution	G/A	G/A
Caution	A/G	A/G
Caution	A/A	G/G
Caution	G/T	G/T
Caution	G/G	G/G
Caution	C/A	C/A
Caution	T/C	T/C
Caution	T/T	T/T
Caution	G/A	G/A
Caution	A/G	A/G
Caution	A/A	A/A
Caution	A/A	A/A
Caution	C/T	C/T
Caution	C/T	C/T
Caution	G/G	G/G
Caution	G/T	G/T
Caution	A/A	G/G
Caution	C/C	C/C
Caution	A/A	T/T
Caution	G/G	G/G
Caution	G/A	G/A
Caution	C/C	C/C
Caution	A/A	G/G

To summarize that example, all variants apart from rs28777, rs16891982, rs12821256, rs4959270, rs2378249, rs12896399, rs1393350, rs1126809, rs1470608, rs1545397 passed with a green confidence call. Those 10 variants passed with an orange caution call. This means that the user should use caution during interpretation as the variant did not pass the 100 pg read count level but did pass the 50 pg DNA input sensitivity level as measured in this study.

If the variant did not pass the 50 pg threshold filter, the cell will turn to red and say “Failed” as can be seen in the below image. It is not recommended to proceed with this variant using this tool and the variant should be inserted as NA for the prediction of that sample. It is at the users discretion if they go against this recommendation. This tool is merely a guide towards threshold and mixture rule instigation.

Enter sample concentration input

Coverage Threshold	50 pg input	
	Major	Minor
Caution	C/C	C/C
Caution	G/G	G/G
Caution	G/G	G/G
Caution	C/C	C/C
Caution	T/T	G/G
Caution	C/C	C/C
Caution	C/C	C/C
Caution	G/G	G/C
Caution	C/C	C/C
Caution	G/G	G/G
Caution	T/T	T/T
Caution	A/A	C/A
Caution	G/G	C/C
Caution	T/T	T/T
Caution	A/A	C/C
Caution	C/C	C/C
Caution	C/C	C/A
Caution	C/C	C/C
Caution	A/A	G/A
Caution	G/G	A/A
Caution	A/A	G/A
Caution	T/T	G/G
Caution	G/G	G/G
Caution	A/A	C/C
Caution	C/C	T/C
Caution	T/T	T/T
Caution	A/A	G/G
Caution	G/G	A/A
Caution	A/A	A/A
Caution	A/A	A/A
Caution	T/T	C/C
Caution	T/T	C/C
Caution	G/G	G/G
Caution	G/G	G/T
Caution	A/A	A/G
Caution	C/C	C/C
Failed	T/T	A/T
Caution	G/G	G/G
Caution	A/A	G/A
Caution	C/C	C/C
Caution	A/A	A/G

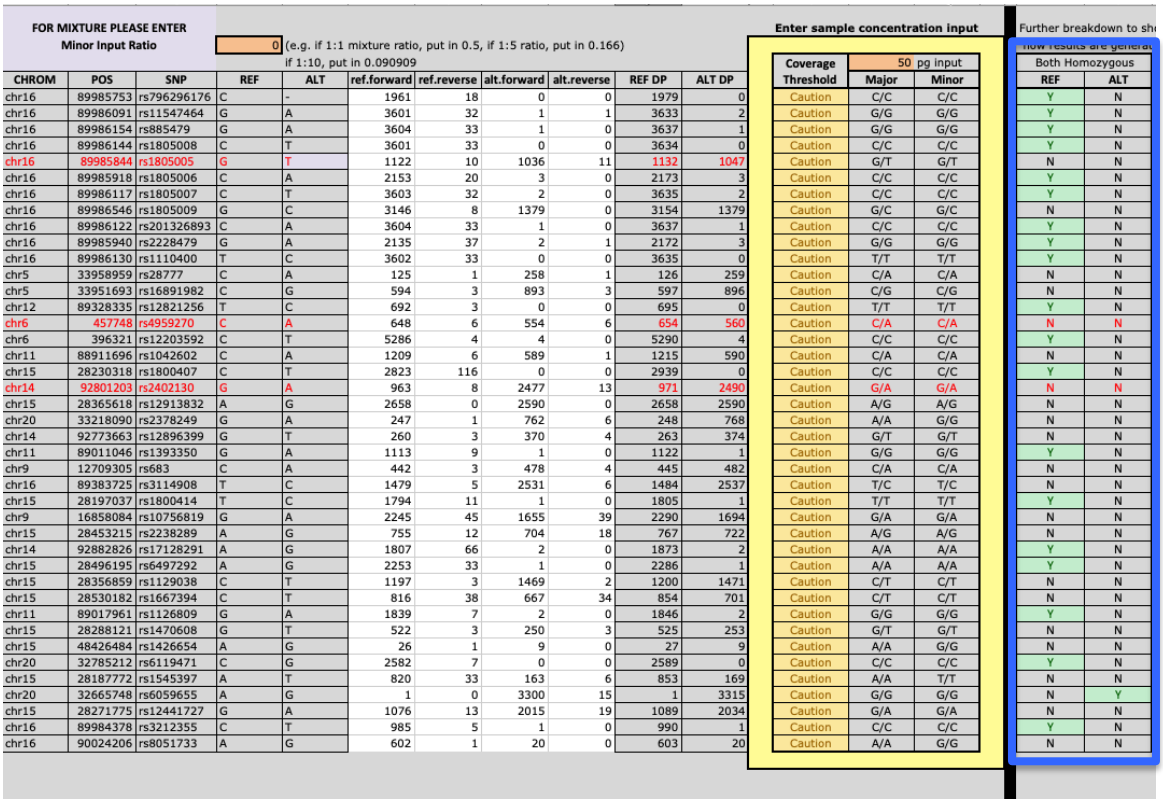
Important to note, 100 pg and 50 pg threshold level counts are set based upon the read count confidence range column found in Supplementary Table 5.

3. *Mixture Guidelines*

Once you have confirmed the variants that can be utilized for the mixture separation of the tool (Green (Passed) & Orange (Caution)), you may proceed with mixture separation guidelines if suggested by Figure 6 (Figure 4 in manuscript) flowchart.

How to check if your sample may or may not be a single source (if no STR profile available) and steps towards genotyping calling in either single/2-pereson mixture scenarios

1. Check for **sample homozygosity** by looking at the table highlighted in blue below.



FOR MIXTURE PLEASE ENTER Minor Input Ratio: 0 (e.g. if 1:1 mixture ratio, put in 0.5, if 1:5 ratio, put in 0.166; if 1:10, put in 0.090909)

Enter sample concentration input: Coverage Threshold = Caution; 50 pg input

Further breakdown to sh… how results are generat…

CHROM	POS	SNP	REF	ALT	ref.forward	ref.reverse	alt.forward	alt.reverse	REF DP	ALT DP	Major	Minor	Both Homozygous REF	Both Homozygous ALT
chr16	89985753	rs796296176	C	-	1961	18	0	0	1979	0	C/C	C/C	Y	N
chr16	89986091	rs11547464	G	A	3601	32	1	1	3633	2	G/G	G/G	Y	N
chr16	89986154	rs885479	G	A	3604	33	1	0	3637	1	G/G	G/G	Y	N
chr16	89986144	rs1805008	C	T	3601	33	0	0	3634	0	C/C	C/C	Y	N
chr16	89985844	rs1805005	G	T	1122	10	1036	11	1132	1047	G/T	G/T	N	N
chr16	89985918	rs1805006	C	A	2153	20	3	0	2173	3	C/C	C/C	Y	N
chr16	89986117	rs1805007	C	T	3603	32	2	0	3635	2	C/C	C/C	Y	N
chr16	89986546	rs1805009	G	C	3146	8	1379	0	3154	1379	G/C	G/C	N	N
chr16	89986122	rs201326893	C	A	3604	33	1	0	3637	1	C/C	C/C	Y	N
chr16	89985940	rs2228479	G	A	2135	37	2	1	2172	3	G/G	G/G	Y	N
chr16	89986130	rs1110400	T	C	3602	33	0	0	3635	0	T/T	T/T	Y	N
chr5	33958959	rs28777	C	A	125	1	258	1	126	259	C/A	C/A	N	N
chr5	33951693	rs16891982	C	G	594	3	893	3	597	896	C/G	C/G	N	N
chr12	89328335	rs12821256	T	C	692	3	0	0	695	0	T/T	T/T	Y	N
chr6	457748	rs4959270	C	A	648	6	554	6	654	560	C/A	C/A	N	N
chr6	396321	rs12203592	C	T	5286	4	4	0	5290	4	C/C	C/C	Y	N
chr11	88911696	rs1042602	C	A	1209	6	589	1	1215	590	C/A	C/A	N	N
chr15	28230318	rs1800407	C	T	2823	116	0	0	2939	0	C/C	C/C	Y	N
chr14	92801203	rs2402130	G	A	963	8	2477	13	971	2490	G/A	G/A	N	N
chr15	28365618	rs12913832	A	G	2658	0	2590	0	2658	2590	A/G	A/G	N	N
chr20	33218090	rs2378249	G	A	247	1	762	6	248	768	A/A	G/G	N	N
chr14	92773663	rs12896399	G	T	260	3	370	4	263	374	G/T	G/T	N	N
chr11	89011046	rs1393350	G	A	1113	9	1	0	1122	1	G/G	G/G	Y	N
chr9	12709305	rs683	C	A	442	3	478	4	445	482	C/A	C/A	N	N
chr16	89383725	rs3114908	T	C	1479	5	2531	6	1484	2537	T/C	T/C	N	N
chr15	28197037	rs1800414	T	C	1794	11	1	0	1805	1	T/T	T/T	Y	N
chr9	16858084	rs10756819	G	A	2245	45	1655	39	2290	1694	G/A	G/A	N	N
chr15	28453215	rs2238289	A	G	755	12	704	18	767	722	A/G	A/G	N	N
chr14	92882826	rs17128291	A	G	1807	66	2	0	1873	2	A/A	A/A	Y	N
chr15	28496195	rs6497292	A	G	2253	33	1	0	2286	1	A/A	A/A	Y	N
chr15	28356859	rs1129038	C	T	1197	3	1469	2	1200	1471	C/T	C/T	N	N
chr15	28530182	rs1667394	C	T	816	38	667	34	854	701	C/T	C/T	N	N
chr11	89017961	rs1126809	G	A	1839	7	2	0	1846	2	G/G	G/G	Y	N
chr15	28288121	rs1470608	G	T	522	3	250	3	525	253	G/T	G/T	N	N
chr15	48426484	rs1426654	A	G	26	1	9	0	27	9	A/A	G/G	N	N
chr20	32785212	rs6119471	C	G	2582	7	0	0	2589	0	C/C	C/C	Y	N
chr15	28187772	rs1545397	A	T	820	33	163	6	853	169	A/A	T/T	N	N
chr20	32665748	rs6059655	A	G	1	0	3300	15	1	3315	G/G	G/G	N	Y
chr15	28271775	rs12441727	G	A	1076	13	2015	19	1089	2034	G/A	G/A	N	N
chr16	89984378	rs3212355	C	T	985	5	1	0	990	1	C/C	C/C	Y	N
chr16	90024206	rs8051733	A	G	602	1	20	0	603	20	A/A	G/G	N	N

EXAMPLE 1

For variant 40: rs3212355 (second last variant), the homozygous REF allele is highlighted as being Homozygous (Y under REF allele, N under ALT allele) in the blue highlighted box. That means that in this sample (if single source or major profile), the individual is a C/C genotype due to all or majority ref allele being present versus the alt allele. The definition of majority is calculated by only 2% of ALT allele in comparison to REF allowed present to still call a C or Ref allele a homozygous genotype call. If over 2% ALT allele is present it will affect homozygous call for majority REF allele. If this is a 2-person mixture sample, it also means that both individuals, major and minor, are C/C at this variant site.

This genotype result is indicated in the genotype calls in the highlighted yellow area.

EXAMPLE 2

For variant 38: rs6059655 (fourth last variant), the homozygous ALT allele is highlighted as being Homozygous, that means that in this sample (if single source or major profile), the individual is a G/G genotype due to nearly all (majority as defined above but for ALT allele present) ALT read counts are G allele calls. If this is a 2-person mixture sample, it also means that both individuals, major and minor, are G/G at this variant site.

This genotype result is also indicated in the genotype calls in the highlighted yellow area.

By first doing this step, you will have generated the majority of genotypes for the single or mixture samples.

EXAMPLE 3

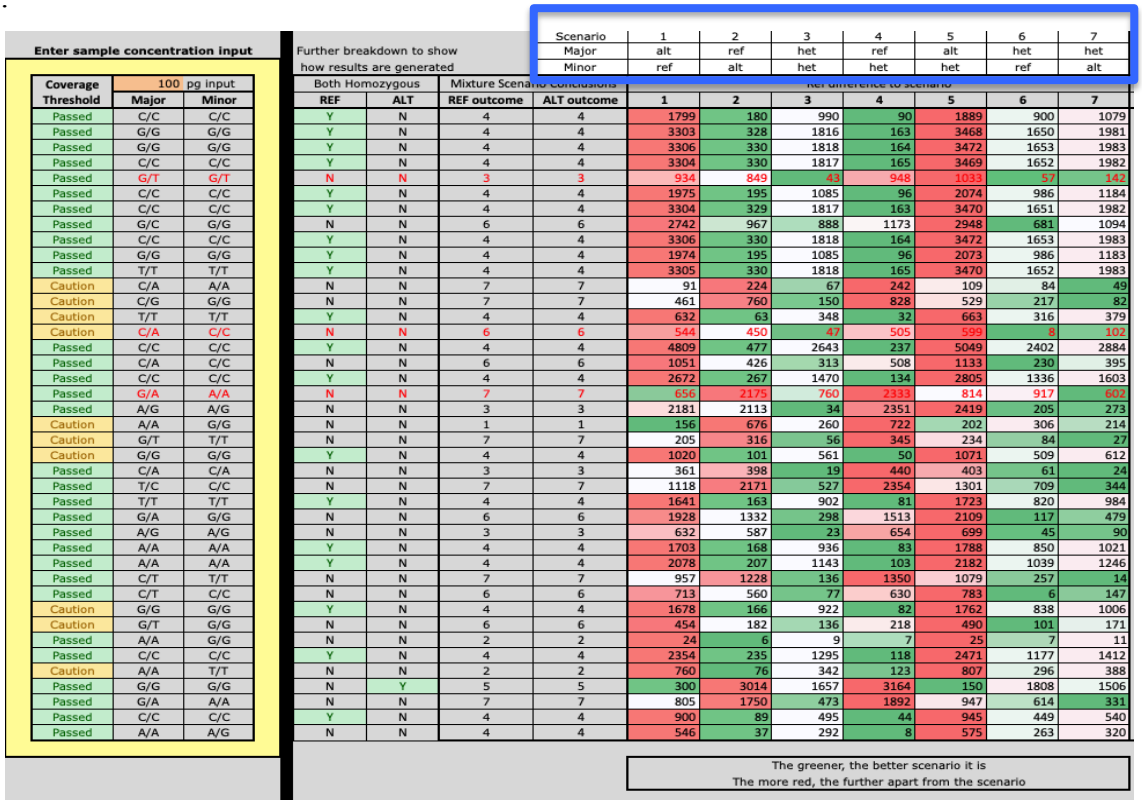
If both homozygous REF and ALT are N, you have two options. Either you have a single source profile and the Major genotype is called heterozygous in the highlighted yellow area, or there is a

mixture possible as the ratios are outside the boundaries of the heterozygote balance for that variant. This is an indication that you may not have a single source sample.

If you suspect or know (due to STR profile) you have a 2-person mixture. The next step is to establish the Major:Minor ratio of the sample. This can be obtained from STR profile data where one can utilize peak height ratios to give an estimate, i.e. 1:10 minor to major ratio. If this is not available, the user can manipulate the Major: Minor until the read counts best "match" the scenarios provided in cells BO:CP of this tool. The user should start with the 1:10 input (value 0.090909) and work their way up to 1:1 ratio, which is 0.5 input.

2. How to perform genotyping calls on non-homozygous calls when you know the minor:major ratio due to STR profile being available in a 2-person mixture

The most important part of the scenario guidelines is that the **minor ratio has to be filled in** and should be as accurate an estimate as possible. This value affects how the scenarios are built from the read count inputs. An explanation of the scenario types can be seen highlighted in a blue box below.



Enter sample concentration input

Coverage Threshold	100 pg input Major	Minor
Passed	C/C	C/C
Passed	G/G	G/G
Passed	G/G	G/G
Passed	C/C	C/C
Passed	G/T	G/T
Passed	C/C	C/C
Passed	C/C	C/C
Passed	G/C	G/G
Passed	C/C	C/C
Passed	G/G	G/G
Passed	T/T	T/T
Caution	C/A	A/A
Caution	C/G	G/G
Caution	T/T	T/T
Caution	C/A	C/C
Passed	C/C	C/C
Passed	C/A	C/C
Passed	C/C	C/C
Passed	G/A	A/A
Passed	A/G	A/G
Caution	A/A	G/G
Caution	G/T	T/T
Caution	G/G	G/G
Passed	C/A	C/A
Passed	T/C	C/C
Passed	T/T	T/T
Passed	G/A	G/G
Passed	A/G	A/G
Passed	A/A	A/A
Passed	A/A	A/A
Passed	C/T	T/T
Passed	C/T	C/C
Caution	G/G	G/G
Caution	G/T	G/G
Passed	A/A	G/G
Passed	C/C	C/C
Caution	A/A	T/T
Passed	G/G	G/G
Passed	G/A	A/A
Passed	C/C	C/C
Passed	A/A	A/G

Further breakdown to show how results are generated

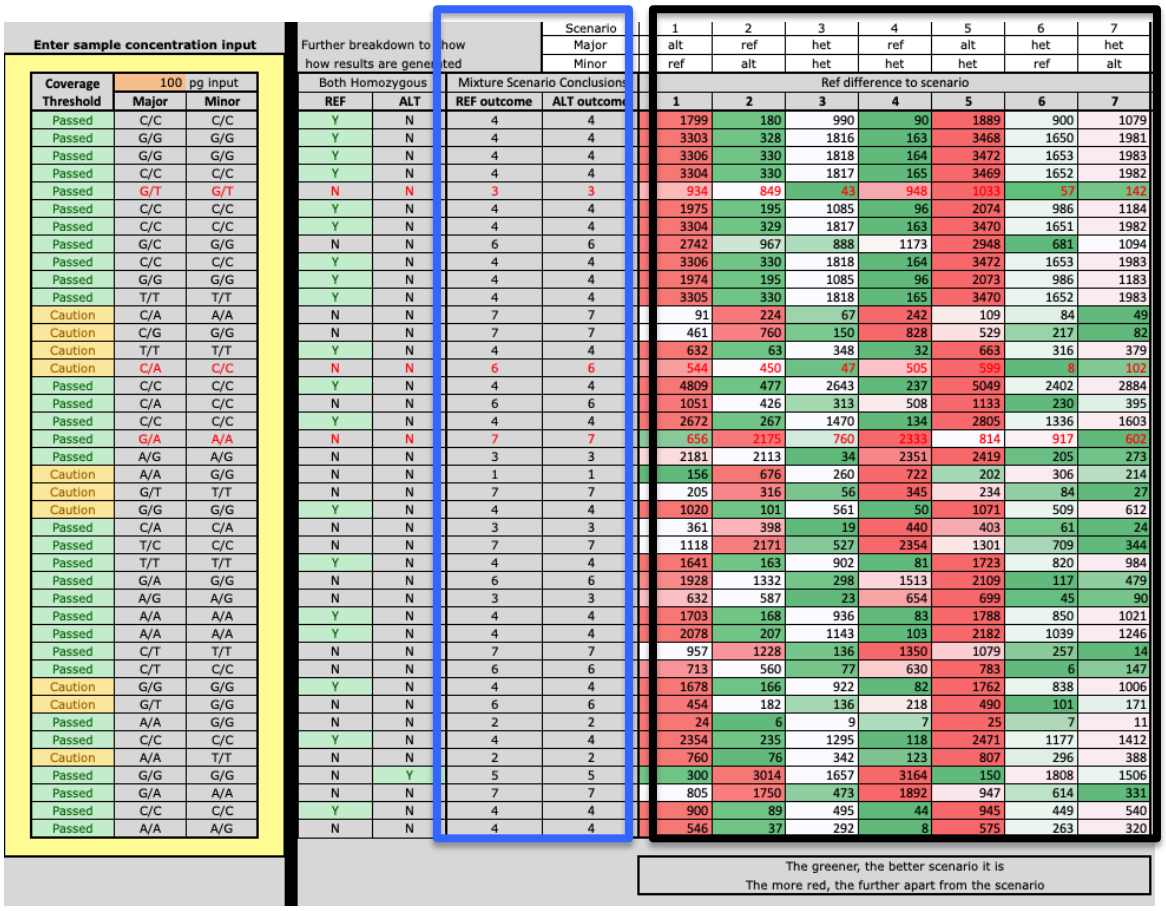
Scenario	1	2	3	4	5	6	7
Major	alt	ref	het	ref	alt	het	het
Minor	ref	alt	het	het	het	ref	alt

Both Homozygous REF	ALT	Mixture Scenario conclusions REF outcome	ALT outcome	1	2	3	4	5	6	7
Y	N	4	4	1799	180	990	90	1889	900	1079
Y	N	4	4	3303	328	1816	163	3468	1650	1981
Y	N	4	4	3306	330	1818	164	3472	1653	1983
Y	N	4	4	3304	330	1817	165	3469	1652	1982
N	N	3	3	934	849	43	948	1033	57	142
Y	N	4	4	1975	195	1085	96	2074	986	1184
Y	N	4	4	3304	329	1817	163	3470	1651	1982
N	N	6	6	2742	967	888	1173	2948	681	1094
Y	N	4	4	3306	330	1818	164	3472	1653	1983
Y	N	4	4	1974	195	1085	96	2073	986	1183
Y	N	4	4	3305	330	1818	165	3470	1652	1983
N	N	7	7	91	224	67	242	109	84	49
N	N	7	7	461	760	150	828	529	217	82
Y	N	4	4	632	63	348	32	663	316	379
N	N	6	6	544	450	47	505	599	8	102
Y	N	4	4	4809	477	2643	237	5049	2402	2884
N	N	6	6	1051	426	313	508	1133	230	395
Y	N	4	4	2672	267	1470	134	2805	1336	1603
N	N	7	7	656	2175	760	2333	814	917	602
N	N	3	3	2181	2113	34	2351	2419	205	273
N	N	1	1	156	676	260	722	202	306	214
N	N	7	7	205	316	56	345	234	84	27
Y	N	4	4	1020	101	561	50	1071	509	612
N	N	3	3	361	398	19	440	403	61	24
N	N	7	7	1118	2171	527	2354	1301	709	344
Y	N	4	4	1641	163	902	81	1723	820	984
N	N	6	6	1928	1332	298	1513	2109	117	479
N	N	3	3	632	587	23	654	699	45	90
Y	N	4	4	1703	168	936	83	1788	850	1021
Y	N	4	4	2078	207	1143	103	2182	1039	1246
N	N	7	7	957	1228	136	1350	1079	257	14
N	N	6	6	713	560	77	630	783	6	147
Y	N	4	4	1678	166	922	82	1762	838	1006
N	N	6	6	454	182	136	218	490	101	171
N	N	2	2	24	6	9	7	25	7	11
Y	N	4	4	2354	235	1295	118	2471	1177	1412
N	N	2	2	760	76	342	123	807	296	388
N	Y	5	5	300	3014	1657	3164	150	1808	1506
N	N	7	7	805	1750	473	1892	947	614	331
Y	N	4	4	900	89	495	44	945	449	540
N	N	4	4	546	37	292	8	575	263	320

The greener, the better scenario it is
The more red, the further apart from the scenario

Both the Major and Minor profile are given a row in the blue box. If ref is mentioned then that contributor is homozygous ref allele. If alt is mentioned, then that contributor is homozygous alt allele. If het is mentioned then that contributor is heterozygous at that site.

There is a ranking of the best scenario based on read count input and minor profile input. This is reflected in the blue box below under mixture scenario conclusions.



Enter sample concentration input

Coverage Threshold	100 pg input Major	Minor	Both Homozygous REF	ALT	Mixture Scenario Conclusions REF outcome	ALT outcome	Scenario	1	2	3	4	5	6	7
							Major	alt	ref	het	ref	alt	het	het
							Minor	ref	alt	het	het	het	ref	alt
								Ref difference to scenario						
								1	2	3	4	5	6	7
Passed	C/C	C/C	Y	N	4	4		1799	180	990	90	1889	900	1079
Passed	G/G	G/G	Y	N	4	4		3303	328	1816	163	3468	1650	1981
Passed	G/G	G/G	Y	N	4	4		3306	330	1818	164	3472	1653	1983
Passed	C/C	C/C	Y	N	4	4		3304	330	1817	165	3469	1652	1982
Passed	G/T	G/T	N	N	3	3		934	849	43	948	1033	57	142
Passed	C/C	C/C	Y	N	4	4		1975	195	1085	96	2074	986	1184
Passed	C/C	C/C	Y	N	4	4		3304	329	1817	163	3470	1651	1982
Passed	G/C	G/G	N	N	6	6		2742	967	888	1173	2948	681	1094
Passed	C/C	C/C	Y	N	4	4		3306	330	1818	164	3472	1653	1983
Passed	G/G	G/G	Y	N	4	4		1974	195	1085	96	2073	986	1183
Passed	T/T	T/T	Y	N	4	4		3305	330	1818	165	3470	1652	1983
Caution	C/A	A/A	N	N	7	7		91	224	67	242	109	84	49
Caution	C/G	G/G	N	N	7	7		461	760	150	828	529	217	82
Caution	T/T	T/T	Y	N	4	4		632	63	348	32	663	316	379
Caution	C/A	C/C	N	N	6	6		544	450	47	505	595	8	102
Passed	C/C	C/C	Y	N	4	4		4809	477	2643	237	5049	2402	2884
Passed	C/A	C/C	N	N	6	6		1051	426	313	508	1133	230	395
Passed	C/C	C/C	Y	N	4	4		2672	267	1470	134	2805	1336	1603
Passed	G/A	A/A	N	N	7	7		656	2175	760	2333	814	917	602
Passed	A/G	A/G	N	N	3	3		2181	2113	34	2351	2419	205	273
Caution	A/A	G/G	N	N	1	1		156	676	260	722	202	306	214
Caution	G/T	T/T	N	N	7	7		205	316	56	345	234	84	27
Caution	G/G	G/G	Y	N	4	4		1020	101	561	50	1071	509	612
Passed	C/A	C/A	N	N	3	3		361	398	19	440	403	61	24
Passed	T/C	C/C	N	N	7	7		1118	2171	527	2354	1301	709	344
Passed	T/T	T/T	Y	N	4	4		1641	163	902	81	1723	820	984
Passed	G/A	G/G	N	N	6	6		1928	1332	298	1513	2109	117	479
Passed	A/G	A/G	N	N	3	3		632	587	23	654	699	45	90
Passed	A/A	A/A	Y	N	4	4		1703	168	936	83	1788	850	1021
Passed	A/A	A/A	Y	N	4	4		2078	207	1143	103	2182	1039	1246
Passed	C/T	T/T	N	N	7	7		957	1228	136	1350	1079	257	14
Passed	C/T	C/C	N	N	6	6		713	560	77	630	783	6	147
Caution	G/G	G/G	Y	N	4	4		1678	166	922	82	1762	838	1006
Caution	G/T	G/G	N	N	6	6		454	182	136	218	490	101	171
Passed	A/A	G/G	N	N	2	2		24	6	9	7	25	7	11
Passed	C/C	C/C	Y	N	4	4		2354	235	1295	118	2471	1177	1412
Caution	A/A	T/T	N	N	2	2		760	76	342	123	807	296	388
Passed	G/G	G/G	N	Y	5	5		300	3014	1657	3164	150	1808	1506
Passed	G/A	A/A	N	N	7	7		805	1750	473	1892	947	614	331
Passed	C/C	C/C	Y	N	4	4		900	89	495	44	945	449	540
Passed	A/A	A/G	N	N	4	4		546	37	292	8	575	263	320

The greener, the better scenario it is
The more red, the further apart from the scenario

For the section highlighted in the black box above, the more green the scenario, with a number closer to 0, is the most optimal scenario in this case (IF and only IF the homozygous check is N for both REF and ALT.)

APPENDIX L

SUPPLEMENTARY MATERIAL 2

Guide to using the HPS-MPS pipeline

STEP 1: Setting up Docker

This pipeline can be used on any platform once the docker container (virtual machine) has been installed in your system. For steps on how to install docker, go to the below link.

https://www.docker.com/get-started

or

https://docs.docker.com/install/overview/

You will see install options for Mac, Windows and Linux.

Mac – https://docs.docker.com/docker-for-mac/install/

Windows – https://docs.docker.com/docker-for-windows/install/

Linux (Ubuntu) - https://docs.docker.com/install/linux/docker-ce/ubuntu/#install-docker-ce

There are particular systems requirements for these installs. If you are having trouble installing the Windows software, please go to this link to install the docker toolbox (supports more Windows operating systems).

https://docs.docker.com/toolbox/toolbox_install_windows/

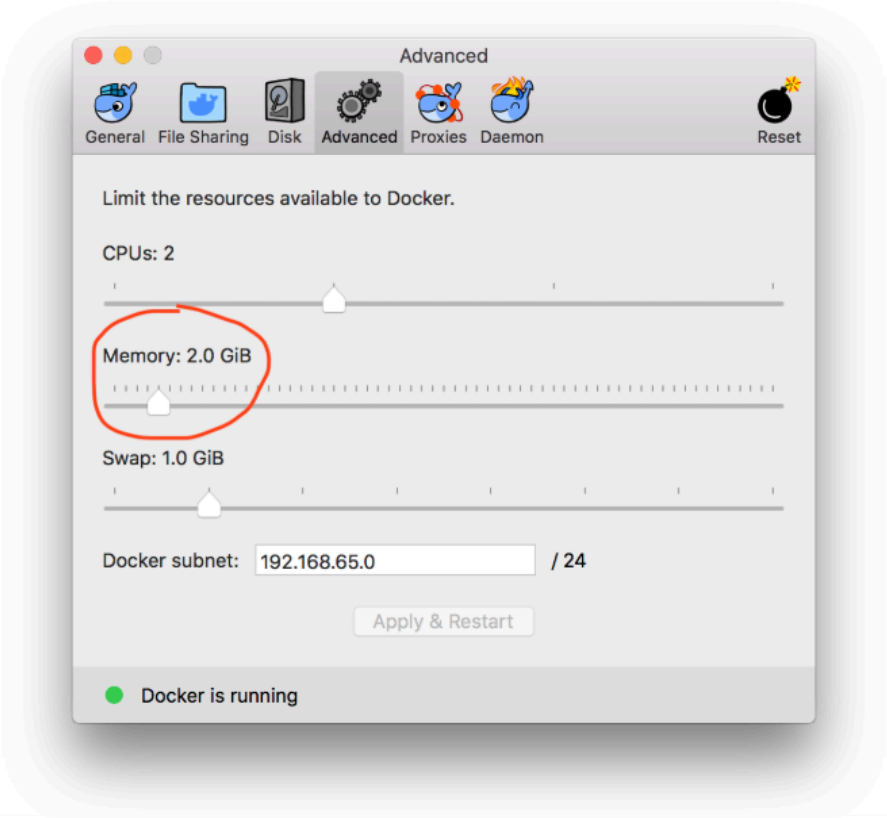
All links show you how to install and how to check you have installed correctly but doing a dummy run of

Docker run hello-world

Once you have docker installed and up and running. You will need to adjust the memory dedicated to the virtual machine so that the processes required to successfully run the hps-mps pipeline will be supported in the environment. This can be easily done using the user interface as seen below.

That `2GB` limit you see is the total memory of the VM in which docker runs.

If you are using docker-for-windows or docker-for-mac you can easily increase it from the Whale icon in the task bar, then go to Preferences -> Advanced:



Please increase the memory to at least 16gb if possible. Increasing CPU's will also shorten run time but that is up to the user.

If you have installed using the Windows toolbox option. You will need to run the following

Change default vm settings

If the default Virtual Box VM does not provide enough resources to give a good experience, we recommend you create a new VM with at least 2 CPUs and 16GB of memory.

- Double click the **Docker Quickstart icon from your desktop** and then run the following commands in that terminal.
- Remove the default vm

```
docker-machine rm default
```

- Re-create the default vm
    - Choose the number of cpus with **--virtualbox-cpu-count**. For this example we'll use two.
    - Choose the amount of RAM: **--virtualbox-memory**. This is also based on the host hardware. However, choose at least 16GB If you can.
    - Choose the amount of disk space: **--virtualbox-disk-size**. It is recommended that this be at least 50GB since building generates a lot* of output. In this example we'll choose 50GB.
    - Create vm with new settings

```
docker-machine create -d virtualbox --virtualbox-cpu-count=2 --virtualbox-memory=16384 --virtualbox-disk-size=50000 default
```

- Restart docker

```
docker-machine stop exit
```

Then open a new Docker Quickstart Terminal.

STEP 2: Setting up the hps-mps pipeline environment

run the following command in your docker terminal

```
docker run -it -v ~/Desktop:/Desktop suswalsh/hpsmps
```

This will set up your environment and give it access to your desktop. Minimize this window unstill needed in step 4.

STEP 3: Downloading folders needed for hps-mps pipeline

1. Download the hps.zip file from here using your internet browser

https://iu.box.com/shared/static/xs2omjmujowpxa4r2gi2kn8waiqrzbog.zip

2. Unzip the file using whatever unzipping software you have on your computer.

3. **IT IS VERY IMPORTANT** to place the 'hps' folder on the **desktop** of your computer (make sure it is the folder itself and not the zipped folder. Its location should be ~/Desktop/hps/ and not ~/Desktop/hps/hps

STEP 4: Preparing for your sequence files and running the pipeline

1. Open the docker terminal window again, and run the following commands in this terminal window

```
chmod +x /Desktop/hps/scripts/*.sh
/Desktop/hps/scripts/1-prepare.sh
```

If you get no errors at this stage, you have installed docker and environment correctly. If you have an error, please check to make sure the hps folder is on the desktop of your computer in the correct location (see above).

2. Minimize this terminal window and go back to your desktop hps folder
3. Now place your sample fastq files (any sequencer) into the hps/runfolder on your desktop, make sure they are **unzipped** (do not have .gz) before placing in hps/runfolder or immediately after placing in folder. **Do not proceed with the pipeline with .gz files.**
4. Edit the dirlist.txt file that is in that hps/runfolder (please use software that can save the file as a unix file – BBedit for Mac OSX or NotePad++ for Windows) to include your sample names (please use names up to the first _ in the filename) as follows

sample1-of10-of-100_S1_L001_R1_001.fastq would be titled sample1-of10-of-100, there should be a sample name in each line as follows

sample1-of10-of-100

sample2-of10-of-100

etc.

Using your sample sheet from your MiSeq or Ion Torrent run is a good way to keep track of your sample names going into the pipeline. **Currently the pipeline is set to running from 1-96 samples at once. If you include more than 96 samples, only the first 96 samples in the dirlist.csv file will run.

5. Go back to your terminal window that you left open and paste in the following commands one at a time (paste, and hit return, etc.)

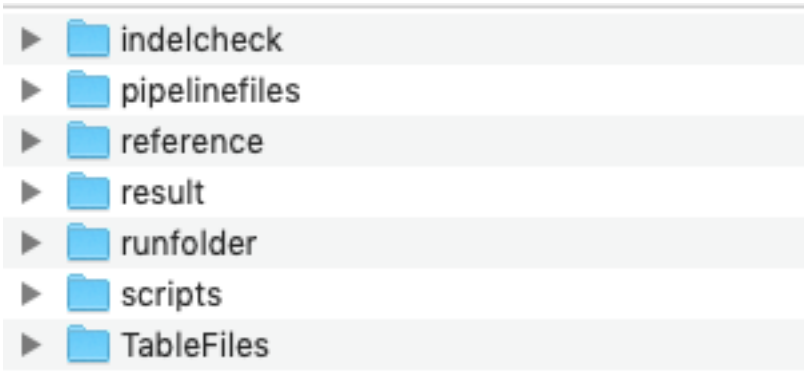
/Desktop/hps/scripts/2-organisefiles.sh

/Desktop/hps/scripts/3-sampleinput.sh

/Desktop/hps/scripts/4-generatedata.sh

/Desktop/hps/scripts/5-makeonlinefile.sh

6. All results can be found in the hps folder - with the final result folder including the upload files for the prediction web tool. A description of every folder (within the hps folder) and what it contains after the pipeline has been run is described below.



POST –RUN OUTPUT

Indelcheck folder contains every samples check for the Indel variant of HPS (variant 1 in HP: rs796296176). There is only one indel in the HPS variants. Each sample is checked for the presence of this indel and a csv file is generated either if it is present or absent.

Pipelinefiles folder contain all the necessary files for using the pipeline. There are no results generated in this folder. Do not delete any files or adjust files in this folder unless you are experienced in doing so.

Reference folder contains the human reference files needed for alignment; in this case Hg19 is being used. Do not delete any files or adjust files in this folder unless you are experienced in doing so.

Result folder contains all the main result files of the run all summed up into single files. It includes the following files:

The indelcheck file is a complete list of all samples that had the variant 1 indel. IT IS VERY IMPORTANT TO CHECK THIS FILE as it influences the online and onlineupload file, if a sample shows the presence of the indel, then you must edit both the online and onlineupload file for that individual to show their correct genotype i.e.

If indel present in sample 123, go to sample 123 in the online.csv file and the onlineupload.csv file and change the genotype.

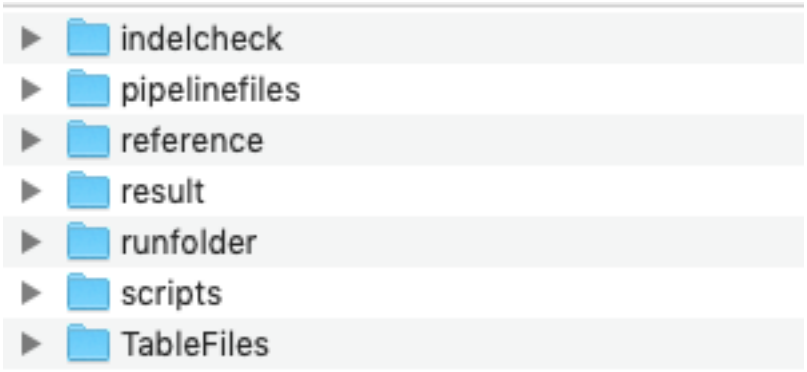
With no indel that variant would say CC (online file) or 0 (onlineupload file) for that sample, the user will have to edit this to say CA (if hetero) or AA (if homo insertion). Also for the onlineupload file which is used for prediction it should be changed from 0, to 1 or 2.

It is also best practice to edit the file found below in the TableFiles folder with this new genotype information (due to indelcheck). How to do this is described below.

The online.csv file is one file with every genotype result for all (up to 96 samples) that you ran.

The onlineupload.csv file is the online webtool file that you use to generate prediction probabilities for eye, hair and skin color using the website https://hirisplex.erasmusmc.nl/

This file needs no further editing (if the pipeline gave no errors and no indels were reported in the indelcheck file) and can simply be uploaded directly.



runfolder contains every file generated for each sample in each of their respective sample folder names.

I.e if you name a sample Sample123, there will be a folder called Sample123 in the runfolder. In here you will find:

Sample123.sam

Sample123.bam

Sample123sorted.bam

Sample123sorted.bam.bai

Sample123sorted.vcf

Sample123.recode.vcf

Scripts folder contains all the necessary scripts needed to run the pipeline. Do not delete any files or adjust files in this folder unless you are experienced in doing so.

TableFiles folder generates every samples .csv file which contains variant ref and alt alleles, genotype calls and read counts generated from the hps-mps pipeline using the programs as described in Figure 3 (Figure 1 manuscript) and the manuscript materials and methods. IF YOU

HAVE AN INDEL IN A SAMPLE FILE, you must edit the samples genotype and read count for that variant 1 indel. See example on how below. Lets call the sample Sample123

chr16	89985750	C	=+A	*/+A:1080:583:496:45.93%:4.3379E-180	Pass:0.9858870967741935:582:1:489:7:2.0375E-2	0	1	0	0	*/+A:1080:583:496:45.93%:4.3379E-180

This shows that this sample had an indel found at our variant 1 site. The most important part are the number of reads for the ref * versus the +A at that site. Here it shows

COV = 1080

READS1 583: READS2 496.

Cons.Cov.Reads1.Reads2.Freq.P.value
*/+A:1080:583:496:45.93%:4.3379E-180

The user **MUST** edit the Sample123.csv file found in the TableFiles folder as being Ref.forward of 583 and Alt.forward of 496 for that variant with ref being C and alt being +A

***** THESE FILES ARE USED AS INPUT THAT MUST PASS THE THRESHOLD AND MIXTURE TOOL RULES FOR SAMPLE INTERPRETATION**

****In order to run the pipeline from scratch on new samples, it is recommended to re-download the hps folder and go from STEP 4 of this guide.*

TROUBLESHOOTING GUIDES:

If you experience any errors, some of the likely causes are listed below.

1. You did not enter the correct sample name into the dirlist file (or the file must be saved as a unix line breaks file)

Some of the fastq files contain no data (0 bytes) and hence that sample will not contain information in any of its generated files. If this happens you must delete the 'empty' .csv files in the **TableFiles** folder before running script 5 or it will give an error when making the final result folder files – *hence the online upload file will not be made*

APPENDIX M

Description of Samples used in Proteomic Analysis

Sample	Proteomic Results Obtained	Population	Sex	Age	Hair Color	Hair Type
1		Southern European	Male	12.02	Light Brown	Light Waves
2			Female	18.97	Light Brown	Light Waves
3	✓		Male	20.33	Light Brown	Light Waves
4	✓		Female	19.5	Light Brown	Light Waves
5	✓		Female	18.73	Dark Brown	Light Waves
6			Female	18.08	Blond	Very Straight
7	✓		Female	19.57	Blond	Very Straight
8			Female	19.07	Light Brown	Light Waves
9	✓		Male	20.33	Dark Brown	Light Waves
10			Female	20.46	Blond	Light Waves
11	✓		Male	18.86	Pure Red	Light Waves
12	✓		Female	21.18	Light Brown	Very Straight
13			Male	20.91	Dark Brown	Light Waves
14	✓		Male	48.17	Blond	Light Waves
15	✓	Northern Europe	Female	22.67	Light Brown	Light Waves
16			Male	49.92	Dark Brown	Very Straight
17	✓		Female	17.99	Blond	Very Straight
18			Male	30.88	Dark Brown	Very Straight
19	✓		Male	23	Dark Brown	Very Straight
20	✓		Male	19.48	Light Brown	Very Straight
21	✓		Male	26.22	Light Brown	Very Straight
22	✓		Female	19.58	Blond	Very Straight
23			Female	18.72	Light Red/Strawberry Blond	Light Waves
24	✓		Female	18.68	Black	Very Straight
25	✓		Male	18.79	Light Brown	Very Straight
26	✓		Female	21.08	Light Brown	Very Straight
27			Male	21.33	Blond	Very Straight
28	✓		Male	19.75	Light Brown	Very Straight
29			Female	53.25	Red Brown/Auburn	Light Waves
30	✓	Western Europe	Male	19.02	Blond	Light Waves
31	✓		Female	18.75	Red Brown/Auburn	Light Waves
32	✓		Male	19.53	Dark Brown	Light Waves
33			Male	19.45	Dark Brown	Very Straight
34	✓		Female	17.89	Light Brown	Light Waves
35	✓		Male	18.21	Blond	Very Straight
36			Female	18.44	Blond	Very Straight
37			Female	19.75	Red Brown/Auburn	Very Curly/Rings of Curl
38	✓		Female	19	Light Brown	Light Waves
39			Male	26.83	Light Brown	Light Waves
40	✓		Female	19.29	Dark Brown	Light Waves
41	✓		Male	39.92	Light Brown	Very Straight
42	✓		Female	21.42	Light Brown	Light Waves
43	✓		Male	18.16	Blond	Light Waves
44			Female	18.85	Blond	Very Straight
45	✓		Female	18.68	Red Brown/Auburn	Light Waves
46	✓		Female	27.17	Pure Red	Light Waves
47	✓		Female	20.89	Pure Red	Very Curly/Rings of Curl

48	✓	Middle Eastern	Female	26.58	Dark Brown	Very Curly/Rings of Curl
49	✓		Female	20	Dark Brown	Light Waves
50	✓		Female	18.92	Dark Brown	Very Straight
51			Male	28.45	Dark Brown	Very Curly/Ring of Curl
52	✓		Female	18.75	Blond	Light Waves
53			Female	18.57	Light Brown	Very Curly/Ring of Curl
54			Female	26.38	Light Brown	Very Straight
55			Male	51.25	Dark Brown	Very Straight
56	✓		Male	20.48	Dark Brown	Light Waves
57	✓		Male	19.74	Dark Brown	Light Waves
58			Male	20.68	Dark Brown	Light Waves
59	✓		Female	18.35	Light Brown	Light Waves
60	✓	African	Male	19.19	Black	Very Curly/Ring of Curl
61	✓		Male	18.88	Black	Very Curly/Ring of Curl
62	✓		Female	18.14	Black	Very Curly/Ring of Curl
63	✓		Female	24	Black	Very Curly/Ring of Curl
64	✓		Female	26.25	Black	Very Curly/Ring of Curl
65	✓		Male	23	Black	Very Curly/Ring of Curl
66			Female	18.74	Black	Light Waves
67	✓		Female	25	Black	Very Straight
68			Female	20.42	Black	Very Curly/Ring of Curl
69	✓		Male	30.08	Black	Very Curly/Ring of Curl
70			Female	21.92	Black	Very Curly/Ring of Curl
71	✓		Male	26.25	Black	Very Curly/Ring of Curl
72	✓		Male	29	Dark Brown	Very Curly/Ring of Curl
73			Female	18.4	Black	Very Curly/Ring of Curl
74			Female	18.2	Dark Brown	Light Waves
75	✓		Female	18.46	Black	Very Curly/Ring of Curl
76	✓		Female	18.98	Black	Very Curly/Ring of Curl
77			Male	18.83	Black	Very Curly/Ring of Curl
78	✓	Admixed	Female	20	Dark Brown	Very Curly/Ring of Curl
79			Female	19.02	Dark Brown	Very Curly/Ring of Curl
80	✓		Female	19.89	Black	Very Curly/Ring of Curl
81	✓		Female	21.46	Black	Very Curly/Ring of Curl
82			Female	18.91	Dark Brown	Very Curly/Ring of Curl
83			Female	18.37	Dark Brown	Very Curly/Ring of Curl
84	✓		Female	19.13	Light Brown	Light Waves
85	✓		Female	19.83	Dark Brown	Very Curly/Ring of Curl
86	✓		Female	29.67	Dark Brown	Light Waves
87	✓		Male	19.75	Light Brown	Very Curly/Ring of Curl
88	✓		Female	22.24	Blond	Very Curly/Ring of Curl
89	✓		Male	18.56	Dark Brown	Very Curly/Ring of Curl
90	✓		Male	19.73	Dark Brown	Very Straight
91	✓		Male	21.3	Dark Brown	Very Curly/Ring of Curl
92	✓	Asian	Male	18.21	Black	Very Straight
93			Female	20.53	Black	Light Waves
94			Female	18	Black	Very Straight
95	✓		Male	18.17	Black	Very Straight
96	✓		Male	21.25	Black	Very Straight
97	✓		Male	24.5	Black	Light Waves
98			Male	20.58	Black	Very Straight
99	✓		Male	18.27	Black	Very Straight

APPENDIX N

Genotype Frequencies of GVPComplete

GVPComplete Genotype Frequencies (N=66)						
CHROM	POS	SNP	REF	ALT	Genotype Frequency	
1	153431406	rs41265164	G	A	GG	0.72727273
					GA	0.21212121
					AA	0.06060606
1	153520203	rs116208483	G	C	GG	1
					GC	0
					CC	0
1	153520954	rs62624468	C	T	CC	0.95454545
					CT	0.03030303
					TT	0
1	201289487	rs61818256	C	T	CC	0.93939394
					CT	0.06060606
					TT	0
6	74014637	rs28763966	C	A	CC	0.89393939
					CA	0.10606061
					AA	0
6	7581001	rs28763967	C	T	CC	0.96969697
					CT	0.03030303
					TT	0
12	52788945	rs1791634	C	G	CC	0.98484848
					GC	0.01515152
					GG	0
12	53069014	rs17678945	C	A	CC	0.95454545
					CA	0.04545455
					AA	0
14	113975768	rs10148371	G	A	GG	1
					GA	0
					AA	0
14	55609418	rs11125	A	T	AA	0.95454545
					AT	0.04545455
					TT	0
17	38859509	rs7213256	C	T	CC	0.65151515
					CT	0.28787879
					TT	0.06060606
17	39116603	rs17843021	G	A	GG	0.78787879
					GA	0.1969697
					AA	0.01515152
17	39116728	rs142154718	C	T	CC	1
					CT	0
					TT	0
17	39183254	rs62623375	C	T	CC	0.75757576
					CT	0.22727273
					TT	0.01515152
17	39593768	rs2604953	G	T	TT	0.71212121
					GT	0.24242424
					GG	0.04545455
17	39633354	rs138303882	G	A	GG	1
					GA	0
					AA	0
17	39635733	rs743686	A	G	GG	0.42622951
					GA	0.32786885
					AA	0.24590164
17	39913771	rs41283425	C	T	CC	0.91935484
					CT	0.06451613
					TT	0
18	28605818	rs79011243	C	A	CC	0.95454545
					CA	0.04545455
					AA	0
21	31744310	rs9636845	A	T	AA	0.87878788
					AT	0.10606061
					TT	0.01515152
21	32253513	rs71321355	C	T	CC	0.74242424
					CT	0.1969697
					TT	0.06060606
17	39913771	rs143043662	C	T	CC	0.98412698
					CT	0.01587302
					TT	0
					Not Detected	0
17	39619115	rs2071563	G	A	GG	0.57142857
					GA	0.42857143
					AA	0
					Not Detected	0
17	39620641	rs146792525	C	T	CC	1
					CT	0
					TT	0
					Not Detected	0
12	52788928	rs2658658	G	A	GG	0.73015873
					GA	0.12698413
					AA	0.04761905
					Not Detected	0.0952381
12	52788945	rs1732263	C	G	CC	0.82539683
					GC	0.12698413
					GG	0.01587302
					Not Detected	0.03174603
12	52713088	rs2857663	G	A	GG	0.88888889
					GA	0.11111111
					AA	0
					Not Detected	0
17	39183313	rs148449559	G	C	GG	0.96825397
					GC	
					CC	
					Not Detected	0.03174603
17	39156084	rs9897046	T	C	TT	0.93650794
					CT	
					CC	
					Not Detected	0.06349206
17	39334241	rs62067292	G	C	GG	0.96825397
					GC	
					CC	
					Not Detected	0.03174603
17	39323971	rs428371	G	A	GG	0.95238095
					GA	
					AA	
					Not Detected	0.04761905
21	46117792	rs34302939	G	A	GG	0.65079365
					GA	0.26984127
					AA	0
					Not Detected	0.07936508

APPENDIX O

Probability of Identity Calculations for Five Populations on GVP21 and GVPComplete SNP Sets

GVP21			p	q							
SNP	Allele 1	Allele 2	Allele 1 Freq	Allele 2 Freq	p2	(p2)^2	2pq	(2pq)^2	q2	(q2)^2	SUM of Squares
rs41265164	A	G	0.4228	0.5772	0.17875984	0.03195508	0.48808032	0.2382224	0.33315984	0.110995479	0.381172958
rs116208483	C	G	0	1	0	0	0	0	1	1	1
rs62624468	T	C	0.06884	0.93116	0.00473895	2.2458E-05	0.12820211	0.01643578	0.86705895	0.751791215	0.768249453
rs61818256	T	C	0	1	0	0	0	0	1	1	1
rs28763966	A	C	0.1876	0.8124	0.03519376	0.0012386	0.30481248	0.09291065	0.65999376	0.435591763	0.529741012
rs28763967	T	C	0.0007564	0.9992436	5.7214E-07	3.2735E-13	0.00151166	2.2851E-06	0.99848777	0.996977831	0.996980116
rs1791634	G	C	0.0348	0.9652	0.00121104	1.4666E-06	0.06717792	0.00451287	0.93161104	0.86789913	0.872413469
rs17678945	A	C	0	1	0	0	0	0	1	1	1
rs10148371	A	G	0.1195	0.8805	0.01428025	0.00020393	0.2104395	0.04428478	0.77528025	0.601059466	0.645548175
rs11125	T	A	0.0007564	0.9992436	5.7214E-07	3.2735E-13	0.00151166	2.2851E-06	0.99848777	0.996977831	0.996980116
rs7213256	C	T	0.4592	0.5408	0.21086464	0.0444639	0.49667072	0.2466818	0.29246464	0.085535566	0.376681266
rs17843021	A	G	0.1452	0.8548	0.02108304	0.00044449	0.24823392	0.06162008	0.73068304	0.533897705	0.595962279
rs142154718	T	C	0.01437	0.98563	0.0002065	4.2641E-08	0.02832701	0.00080242	0.9714665	0.943747155	0.944549617
rs62623375	T	C	0.1891	0.8109	0.03575881	0.00127869	0.30668238	0.09405408	0.65755881	0.432383589	0.527716363
rs2604953	G	T	0.2511	0.7489	0.06305121	0.00397546	0.37609758	0.14144939	0.56085121	0.31455408	0.459978925
rs138303882	A	G	0.004539	0.995461	2.0603E-05	4.2446E-10	0.00903679	8.1664E-05	0.9909426	0.981967241	0.982048906
rs743686	A	G	0.2648	0.7352	0.07011904	0.00491668	0.38936192	0.1516027	0.54051904	0.292160833	0.448680217
rs41283425	T	C	0.002269	0.997731	5.1484E-06	2.6506E-11	0.0045277	2.05E-05	0.99546715	0.990954843	0.990975344
rs79011243	A	C	0.08775	0.91225	0.00770006	5.9291E-05	0.16009988	0.02563197	0.83220006	0.692556944	0.718248205
rs9636845	T	A	0.004539	0.995461	2.0603E-05	4.2446E-10	0.00903679	8.1664E-05	0.9909426	0.981967241	0.982048906
rs71321355	T	C	0.04917	0.95083	0.00241769	5.8452E-06	0.09350462	0.00874311	0.90407769	0.817356468	0.826105427
								Probability of ID for GVP21 for AFR pop			1.14E-03

11 Non-Genotyped SNPs			p	q							
SNP	Allele 1	Allele 2	Allele 1 Freq	Allele 2 Freq	p2	(p2)^2	2pq	(2pq)^2	q2	(q2)^2	SUM of Squares
rs2857663	A	G	0.1589	0.8411	0.02524921	0.00063752	0.26730158	0.07145013	0.70744921	0.500484385	0.572572042
rs2658658	A	G	0.3374	0.6626	0.11383876	0.01295926	0.44712248	0.19991851	0.43903876	0.192755033	0.405632808
rs1732263	G	C	0.2133	0.7867	0.04549689	0.00206997	0.33560622	0.11263153	0.61889689	0.38303336	0.497734862
rs9897046	T	C	0.3888	0.6112	0.15116544	0.02285099	0.47526912	0.22588074	0.37356544	0.139551138	0.388282865
rs148449559	C	G	0.02194	0.97806	0.00048136	2.3171E-07	0.04291727	0.00184189	0.95660136	0.915086169	0.916928293
rs428371	A	G	0.3759	0.6241	0.14130081	0.01996592	0.46919838	0.22014712	0.38950081	0.151710881	0.39182392
rs62067292	C	G	0.002269	0.997731	5.1484E-06	2.6506E-11	0.0045277	2.05E-05	0.99546715	0.990954843	0.990975344
rs2071563	A	G	0.4259	0.5741	0.18139081	0.03290263	0.48901838	0.23913898	0.32959081	0.108630102	0.380671704
rs146792525	T	C	0.01967	0.98033	0.00038691	1.497E-07	0.03856618	0.00148735	0.96104691	0.923611161	0.925098661
rs143043662	T	C	0.001513	0.998487	2.2892E-06	5.2403E-12	0.00302142	9.129E-06	0.99697629	0.993961721	0.99397085
rs34302939	G	A	0.3767	0.6233	0.14190289	0.02013643	0.46959422	0.22051873	0.38850289	0.150934496	0.391589657
						Probability of ID for AFR pop with 11 Non-Geno SNPs					2.19E-03
						Probability of ID for AFR pop with GVPComplete SNPs					2.49107E-06

GVP21			p	q							
SNP	Allele 1	Allele 2	Allele 1 Freq	Allele 2 Freq	p2	(p2)^2	2pq	(2pq)^2	q2	(q2)^2	SUM of Squares
rs41265164	A	G	0.1037	0.8963	0.01075369	0.000115642	0.18589262	0.03455607	0.80335369	0.64537715	0.680048859
rs116208483	C	G	0.007205	0.992795	5.1912E-05	2.69486E-09	0.01430618	0.00020467	0.98564191	0.97148998	0.971694648
rs62624468	T	C	0.02161	0.97839	0.00046699	2.18082E-07	0.04228602	0.00178811	0.95724699	0.9163218	0.918110129
rs61818256	T	C	0.01729	0.98271	0.00029894	8.93676E-08	0.03398211	0.00115478	0.96571894	0.93261308	0.933767952
rs28763966	A	C	0.01009	0.98991	0.00010181	1.03649E-08	0.01997638	0.00039906	0.97992181	0.96024675	0.960645816
rs28763967	T	C	0.001441	0.998559	2.0765E-06	4.31177E-12	0.00287785	8.282E-06	0.99712008	0.99424845	0.994256729
rs1791634	G	C	0	1	0	0	0	0	1	1	1
rs17678945	A	C	0.01009	0.98991	0.00010181	1.03649E-08	0.01997638	0.00039906	0.97992181	0.96024675	0.960645816
rs10148371	A	G	0.008646	0.991354	7.4753E-05	5.58806E-09	0.01714249	0.00029387	0.98278275	0.96586194	0.966155811
rs11125	T	A	0.07925	0.92075	0.00628056	3.94455E-05	0.14593888	0.02129816	0.84778056	0.71873188	0.740069483
rs7213256	T	C	0.1902	0.8098	0.03617604	0.001308706	0.30804792	0.09489352	0.65577604	0.43004221	0.526244442
rs17843021	A	G	0.05476	0.94524	0.00299866	8.99195E-06	0.10352268	0.01071695	0.89347866	0.79830411	0.80903005
rs142154718	T	C	0	1	0	0	0	0	1	1	1
rs62623375	T	C	0.1167	0.8833	0.01361889	0.000185474	0.20616222	0.04250286	0.78021889	0.60874152	0.651429851
rs2604953	G	T	0.219	0.781	0.047961	0.002300258	0.342078	0.11701736	0.609961	0.37205242	0.491370037
rs138303882	A	G	0.001441	0.998559	2.0765E-06	4.31177E-12	0.00287785	8.282E-06	0.99712008	0.99424845	0.994256729
rs743686	A	G	0.4409	0.5591	0.19439281	0.037788565	0.49301438	0.24306318	0.31259281	0.09771426	0.378566008
rs41283425	T	C	0.03314	0.96686	0.00109826	1.20617E-06	0.06408348	0.00410669	0.93481826	0.87388518	0.877993077
rs79011243	A	C	0.005764	0.994236	3.3224E-05	1.10381E-09	0.01146155	0.00013137	0.98850522	0.97714258	0.977273946
rs9636845	T	A	0.2032	0.7968	0.04129024	0.001704884	0.32381952	0.10485908	0.63489024	0.40308562	0.509649582
rs71321355	T	C	0.1455	0.8545	0.02117025	0.000448179	0.2486595	0.06183155	0.73017025	0.53314859	0.59542832
							Probability of ID for GVP21 for AMER pop				4.96E-03

11 Non-Genotyped SNPs			p	q							
SNP	Allele 1	Allele 2	Allele 1 Freq	Allele 2 Freq	p2	(p2)^2	2pq	(2pq)^2	q2	(q2)^2	SUM of Squares
rs2857663	A	G	0.05764	0.94236	0.00332237	1.10381E-05	0.10863526	0.01180162	0.88804237	0.78861925	0.800431908
rs2658658	A	G	0.3746	0.6254	0.14032516	0.019691151	0.46854968	0.2195388	0.39112516	0.15297889	0.392208844
rs1732263	G	C	0.04179	0.95821	0.0017464	3.04993E-06	0.08008719	0.00641396	0.9181664	0.84302955	0.849446554
rs9897046	C	T	0.2061	0.7939	0.04247721	0.001804313	0.32724558	0.10708967	0.63027721	0.39724936	0.506143344
rs148449559	C	G	0	1	0	0	0	0	1	1	1
rs428371	A	G	0.02882	0.97118	0.00083059	6.89884E-07	0.05597882	0.00313363	0.94319059	0.88960849	0.892742811
rs62067292	C	G	0.01729	0.98271	0.00029894	8.93676E-08	0.03398211	0.00115478	0.96571894	0.93261308	0.933767952
rs2071563	A	G	0.402	0.598	0.161604	0.026115853	0.480792	0.23116095	0.357604	0.12788062	0.385157421
rs146792525	T	C	0.001441	0.998559	2.0765E-06	4.31177E-12	0.00287785	8.282E-06	0.99712008	0.99424845	0.994256729
rs143043662	T	C	0.002882	0.997118	8.3059E-06	6.89884E-11	0.00574739	3.3032E-05	0.99424431	0.98852174	0.988554772
rs34302939	A	G	0.1383	0.8617	0.01912689	0.000365838	0.23834622	0.05680892	0.74252689	0.55134618	0.608520941
						Probability of ID for AMER pop with 11 Non-Geno SNPs					2.59E-02
						Probability of ID for AMER pop with GVPComplete SNPs					1.29E-04

GVP21			p	q							
SNP	Allele 1	Allele 2	Allele 1 Freq	Allele 2 Freq	p2	(p2)^2	2pq	(2pq)^2	q2	(q2)^2	SUM of Squares
rs41265164	A	G	0.06339	0.93661	0.00401829	1.61467E-05	0.11874342	0.0141	0.87723829	0.76954702	0.783663167
rs116208483	C	G	0	1	0	0	0	0	1	1	1
rs62624468	T	C	0.003067	0.996933	9.4065E-06	8.8482E-11	0.00611519	3.7396E-05	0.99387541	0.98778832	0.987825719
rs61818256	T	C	0.003067	0.996933	9.4065E-06	8.8482E-11	0.00611519	3.7396E-05	0.99387541	0.98778832	0.987825719
rs28763966	A	C	0	1	0	0	0	0	1	1	1
rs28763967	T	C	0	1	0	0	0	0	1	1	1
rs1791634	G	C	0	1	0	0	0	0	1	1	1
rs17678945	A	C	0.003067	0.996933	9.4065E-06	8.8482E-11	0.00611519	3.7396E-05	0.99387541	0.98778832	0.987825719
rs10148371	A	G	0	1	0	0	0	0	1	1	1
rs11125	T	A	0.09202	0.90798	0.00846768	7.17016E-05	0.16710464	0.02792396	0.82442768	0.679681	0.707676662
rs7213256	T	C	0.2106	0.7894	0.04435236	0.001967132	0.33249528	0.11055311	0.62315236	0.38831886	0.500839107
rs17843021	A	G	0.08282	0.91718	0.00685915	4.7048E-05	0.1519217	0.0230802	0.84121915	0.70764966	0.730776912
rs142154718	T	C	0	1	0	0	0	0	1	1	1
rs62623375	T	C	0.02454	0.97546	0.00060221	3.62659E-07	0.04787558	0.00229207	0.95152221	0.90539452	0.907686953
rs2604953	G	T	0.1421	0.8579	0.02019241	0.000407733	0.24381518	0.05944584	0.73599241	0.54168483	0.601538403
rs138303882	A	G	0	1	0	0	0	0	1	1	1
rs743686	A	G	0.3476	0.6524	0.12082576	0.014598864	0.45354848	0.20570622	0.42562576	0.18115729	0.401462376
rs41283425	T	C	0.03783	0.96217	0.00143111	2.04807E-06	0.07279778	0.00529952	0.92577111	0.85705215	0.862353711
rs79011243	A	C	0.001022	0.998978	1.0445E-06	1.09095E-12	0.00204191	4.1694E-06	0.99795704	0.99591826	0.995922432
rs9636845	T	A	0.1738	0.8262	0.03020644	0.000912429	0.28718712	0.08247644	0.68260644	0.46595155	0.549340423
rs71321355	T	C	0.3119	0.6881	0.09728161	0.009463712	0.42923678	0.18424421	0.47348161	0.22418484	0.41789276
							Probability of ID for GVP21 for SASIA pop				8.46E-03

11 Non-Genotyped SNPs			p	q							
SNP	Allele 1	Allele 2	Allele 1 Freq	Allele 2 Freq	p2	(p2)^2	2pq	(2pq)^2	q2	(q2)^2	SUM of Squares
rs2857663	A	G	0.2096	0.7904	0.04393216	0.001930035	0.33133568	0.10978333	0.62473216	0.39029027	0.502003639
rs2658658	A	G	0.4622	0.5378	0.21362884	0.045637281	0.49714232	0.24715049	0.28922884	0.08365332	0.37644109
rs1732263	G	C	0.07771	0.92229	0.00603884	3.64676E-05	0.14334231	0.02054702	0.85061884	0.72355242	0.744135904
rs9897046	C	T	0.1943	0.8057	0.03775249	0.001425251	0.31309502	0.09802849	0.64915249	0.42139896	0.520852697
rs148449559	C	G	0	1	0	0	0	0	1	1	1
rs428371	A	G	0	1	0	0	0	0	1	1	1
rs62067292	C	G	0.009202	0.990798	8.4677E-05	7.17016E-09	0.01823465	0.0003325	0.98168068	0.96369695	0.964029461
rs2071563	A	G	0.2444	0.7556	0.05973136	0.003567835	0.36933728	0.13641003	0.57093136	0.32596262	0.46594048
rs146792525	T	C	0	1	0	0	0	0	1	1	1
rs143043662	T	C	0.002045	0.997955	4.182E-06	1.74893E-11	0.00408164	1.666E-05	0.99591418	0.99184506	0.991861718
rs34302939	A	G	0.3722	0.6278	0.13853284	0.019191348	0.46733432	0.21840137	0.39413284	0.1553407	0.39293341
						Probability of ID for SASIA pop with 11 Non-Geno SNPs					1.28E-02
						Probability of ID for SASIA pop with GVPComplete SNPs					1.08E-04

GVP21			p	q							
SNP	Allele 1	Allele 2	Allele 1 Freq	Allele 2 Freq	p2	(p2)^2	2pq	(2pq)^2	q2	(q2)^2	SUM of Squares
rs41265164	A	G	0.01488	0.98512	0.00022141	4.90243E-08	0.02931717	0.0008595	0.97046141	0.94179536	0.942654902
rs116208483	C	G	0	1	0	0	0	0	1	1	1
rs62624468	T	C	0	1	0	0	0	0	1	1	1
rs61818256	T	C	0	1	0	0	0	0	1	1	1
rs28763966	A	C	0	1	0	0	0	0	1	1	1
rs28763967	T	C	0	1	0	0	0	0	1	1	1
rs1791634	G	C	0	1	0	0	0	0	1	1	1
rs17678945	A	C	0	1	0	0	0	0	1	1	1
rs10148371	A	G	0	1	0	0	0	0	1	1	1
rs11125	T	A	0.005952	0.994048	3.5426E-05	1.25502E-09	0.01183315	0.00014002	0.98813143	0.97640372	0.97654374
rs7213256	T	C	0.03274	0.96726	0.00107191	1.14899E-06	0.06333618	0.00401147	0.93559191	0.87533222	0.879344839
rs17843021	A	G	0.1399	0.8601	0.01957201	0.000383064	0.24065598	0.0579153	0.73977201	0.54726263	0.605560991
rs142154718	T	C	0	1	0	0	0	0	1	1	1
rs62623375	T	C	0.1627	0.8373	0.02647129	0.000700729	0.27245742	0.07423305	0.70107129	0.49150095	0.566434729
rs2604953	G	T	0.2252	0.7748	0.05071504	0.002572015	0.34896992	0.12178001	0.60031504	0.36037815	0.484730168
rs138303882	A	G	0	1	0	0	0	0	1	1	1
rs743686	G	A	0.3333	0.6667	0.11108889	0.012340741	0.44442222	0.19751111	0.44448889	0.19757037	0.407422224
rs41283425	T	C	0	1	0	0	0	0	1	1	1
rs79011243	A	C	0	1	0	0	0	0	1	1	1
rs9636845	T	A	0.4663	0.5337	0.21743569	0.047278279	0.49772862	0.24773378	0.28483569	0.08113137	0.376143429
rs71321355	T	C	0.05258	0.94742	0.00276466	7.64333E-06	0.09963069	0.00992627	0.89760466	0.80569412	0.815628036
							Probability of ID for GVP21 for EASIA pop				1.68E-02

11 Non-Genotyped SNPs			p	q							
SNP	Allele 1	Allele 2	Allele 1 Freq	Allele 2 Freq	p2	(p2)^2	2pq	(2pq)^2	q2	(q2)^2	SUM of Squares
rs2857663	A	G	0.08036	0.91964	0.00645773	4.17023E-05	0.14780454	0.02184618	0.84573773	0.71527231	0.737160192
rs2658658	A	G	0.06448	0.93552	0.00415767	1.72862E-05	0.12064466	0.01455513	0.87519767	0.76597096	0.780543382
rs1732263	G	C	0	1	0	0	0	0	1	1	1
rs9897046	C	T	0.2312	0.7688	0.05345344	0.00285727	0.35549312	0.12637536	0.59105344	0.34934417	0.478576798
rs148449559	C	G	0	1	0	0	0	0	1	1	1
rs428371	A	G	0	1	0	0	0	0	1	1	1
rs62067292	C	G	0	1	0	0	0	0	1	1	1
rs2071563	A	G	0.3363	0.6637	0.11309769	0.012791087	0.44640462	0.19927708	0.44049769	0.19403821	0.406106387
rs146792525	T	C	0	1	0	0	0	0	1	1	1
rs143043662	T	C	0	1	0	0	0	0	1	1	1
rs34302939	A	G	0.1151	0.8849	0.01324801	0.00017551	0.20370398	0.04149531	0.78304801	0.61316419	0.654835007
						Probability of ID for EASIA pop with 11 Non-Geno SNPs					7.32E-02
						Probability of ID for EASIA pop with GVPComplete SNPs					1.23E-03

GVP21			p	q							
SNP	Allele 1	Allele 2	Allele 1 Freq	Allele 2 Freq	p2	(p2)^2	2pq	(2pq)^2	q2	(q2)^2	SUM of Squares
rs41265164	A	G	0.05169	0.94831	0.00267186	7.13882E-06	0.09803629	0.00961111	0.89929186	0.80872584	0.818344095
rs116208483	C	G	0.00994	0.99006	9.8804E-05	9.76215E-09	0.01968239	0.0003874	0.9802188	0.9608289	0.961216309
rs62624468	T	C	0.01789	0.98211	0.00032005	1.02433E-07	0.0351399	0.00123481	0.96454005	0.93033751	0.931572427
rs61818256	T	C	0.01292	0.98708	0.00016693	2.78644E-08	0.02550615	0.00065056	0.97432693	0.94931296	0.949963551
rs28763966	A	C	0.000994	0.999006	9.8804E-07	9.76215E-13	0.00198602	3.9443E-06	0.99801299	0.99602992	0.996033869
rs28763967	T	C	0.0159	0.9841	0.00025281	6.39129E-08	0.03129438	0.00097934	0.96845281	0.93790085	0.938880247
rs1791634	G	C	0	1	0	0	0	0	1	1	1
rs17678945	A	C	0.0169	0.9831	0.00028561	8.15731E-08	0.03322878	0.00110415	0.96648561	0.93409443	0.935198668
rs10148371	A	G	0.000994	0.999006	9.8804E-07	9.76215E-13	0.00198602	3.9443E-06	0.99801299	0.99602992	0.996033869
rs11125	T	A	0.07853	0.92147	0.00616696	3.80314E-05	0.14472608	0.02094564	0.84910696	0.72098263	0.7419663
rs7213256	T	C	0.161	0.839	0.025921	0.000671898	0.270158	0.07298534	0.703921	0.49550477	0.569162017
rs17843021	A	G	0.1322	0.8678	0.01747684	0.00030544	0.22944632	0.05264561	0.75307684	0.56712473	0.620075781
rs142154718	T	C	0	1	0	0	0	0	1	1	1
rs62623375	T	C	0.0169	0.9831	0.00028561	8.15731E-08	0.03322878	0.00110415	0.96648561	0.93409443	0.935198668
rs2604953	G	T	0.05368	0.94632	0.00288154	8.30329E-06	0.10159692	0.01032193	0.89552154	0.80195883	0.812289069
rs138303882	A	G	0	1	0	0	0	0	1	1	1
rs743686	A	G	0.4602	0.5398	0.21178404	0.04485248	0.49683192	0.24684196	0.29138404	0.08490466	0.376599095
rs41283425	T	C	0.05666	0.94334	0.00321036	1.03064E-05	0.10689929	0.01142746	0.88989036	0.79190484	0.803342609
rs79011243	A	C	0.000994	0.999006	9.8804E-07	9.76215E-13	0.00198602	3.9443E-06	0.99801299	0.99602992	0.996033869
rs9636845	T	A	0.08847	0.91153	0.00782694	6.1261E-05	0.16128612	0.02601321	0.83088694	0.69037311	0.716447581
rs71321355	T	C	0.1978	0.8022	0.03912484	0.001530753	0.31735032	0.10071123	0.64352484	0.41412422	0.516366198
							Probability of ID for GVP21 for EUR pop				1.34E-02

11 Non-Genotyped SNPs			p	q							
SNP	Allele 1	Allele 2	Allele 1 Freq	Allele 2 Freq	p2	(p2)^2	2pq	(2pq)^2	q2	(q2)^2	SUM of Squares
rs2857663	A	G	0.1034	0.8966	0.01069156	0.000114309	0.18541688	0.03437942	0.80389156	0.64624164	0.680735369
rs2658658	A	G	0.4702	0.5298	0.22108804	0.048879921	0.49822392	0.24822707	0.28068804	0.07878578	0.375892772
rs1732263	G	C	0.03877	0.96123	0.00150311	2.25935E-06	0.07453377	0.00555528	0.92396311	0.85370783	0.859265377
rs9897046	C	T	0.1223	0.8777	0.01495729	0.000223721	0.21468542	0.04608983	0.77035729	0.59345035	0.639763904
rs148449559	C	G	0	1	0	0	0	0	1	1	1
rs428371	A	G	0.001988	0.998012	3.9521E-06	1.56194E-11	0.0039681	1.5746E-05	0.99602795	0.99207168	0.992087427
rs62067292	C	G	0.07157	0.92843	0.00512226	2.62376E-05	0.13289547	0.01766121	0.86198226	0.74301343	0.760700869
rs2071563	A	G	0.4245	0.5755	0.18020025	0.03247213	0.4885995	0.23872947	0.33120025	0.10969361	0.380895207
rs146792525	T	C	0	1	0	0	0	0	1	1	1
rs143043662	T	C	0.006958	0.993042	4.8414E-05	2.34389E-09	0.01381917	0.00019097	0.98613241	0.97245714	0.972648109
rs34302939	A	G	0.1392	0.8608	0.01937664	0.000375454	0.23964672	0.05743055	0.74097664	0.54904638	0.606852386
						Probability of ID for EUR pop with 11 Non-Geno SNPs					2.39E-02
						Probability of ID for EUR pop with GVPComplete SNPs					3.21E-04

APPENDIX P

Population Genetics Assessment of 233 Candidate Marker Set

CHROM	POS	SNP	HWE	LD
1	59042181	rs14008		
1	152084419	rs201015975		×
1	152085505	rs2515663		
1	152190954	rs6659183		×
1	152276875	rs116505293		
1	152283862	rs58001094		
1	152285188	rs3120655		
1	152733301	rs4329520		
1	153431406	rs3014837		
1	153520203	rs41265164		×
1	153520280	rs116208483		
1	153520954	rs36022742		×
1	159828613	rs62624468		
1	161495463	rs142859332		
1	201282334	rs34626929		
1	201288955	rs1626370		
1	201289487	rs10920171		
1	201294910	rs61818256		
1	201458071	rs3738283		
2	28761981	rs6753929		
2	85622059	rs6886		
2	113832312	rs6761276		
2	216190020	rs2372536		
2	227886773	rs3752895		
2	233537125	rs11550699		
2	233897414	rs2233384		
2	233897503	rs2233385		
2	233899057	rs2233390		
2	233899126	rs2233391		
3	13395579	rs2280084		
3	38052725	rs933135		
3	38167095	rs2229528		
3	111828423	rs340167		
3	194080916	rs13070515		×
3	194080983	rs13060627		×
5	73981270	rs820878		
5	73992881	rs10805890		
5	74014637	rs77499935		
6	7581001	rs28763966		
6	7581032	rs28763967		
6	7581636	rs6929069		
6	26108282	rs198844		
6	31170227	rs73728294		
6	31683157	rs3749952		
6	31777946	rs2075800		
7	23300345	rs35363287		
8	17796382	rs412750		

8	37728017	rs7817179		
8	144992390	rs62642465		
8	144998444	rs201070741		
8	145001588	rs11136334		
9	139890130	rs2292923		
11	244141	rs7128029		
11	62288978	rs11828907		
11	67352689	rs1695		
11	67353579	rs1138272		
11	93797619	rs1945783		
11	103029516	rs688906		
11	120008468	rs1670195		
11	124502114	rs76570671		
12	31945000	rs7301923		
12	52631313	rs6580870		
12	52681460	rs4761786		
12	52681925	rs6580873		
12	52685096	rs2071588		×
12	52685213	rs79897879		
12	52698717	rs56677856		
12	52699525	rs139895699		×
12	52708420	rs2857671		×
12	52709855	rs140635030		×
12	52710721	rs2852464		×
12	52713088	rs2857663		×
12	52760957	rs61630004		×
12	52774235	rs951773		×
12	52788928	rs2658658		
12	52788945	rs1732263		×
12	52795099	rs1791634		
12	52825839	rs2232393		
12	52827608	rs2232387		
12	52881544	rs11540301		
12	52966428	rs11170177		
12	52981442	rs11170183		
12	52994990	rs116941214		
12	53044267	rs638043		
12	53045626	rs2634041		×
12	53069014	rs14024		×
12	53070174	rs17678945		
12	53085089	rs636127		
12	53186088	rs3887954		
12	53344142	rs59979366		
12	112241766	rs671		
13	113975768	rs9577230		
14	55609418	rs10148371		
14	55611839	rs11125		
15	60653205	rs17845226		
16	4933939	rs2037912		
16	4934286	rs143676756		
17	4856580	rs238239		

17	28576076	rs1050565		
17	38121993	rs3894194		×
17	38131187	rs56030650		×
17	38859509	rs114431517		
17	39114962	rs7213256		×
17	39116603	rs17843023		×
17	39116728	rs17843021		×
17	39122852	rs142154718		
17	39134528	rs16968862		×
17	39135207	rs150812789		×
17	39137104	rs9908389		×
17	39137297	rs721957		×
17	39137387	rs2010027		×
17	39139370	rs9908304		×
17	39140203	rs140634473		
17	39155969	rs3813050		×
17	39156027	rs3829598		×
17	39156084	rs9897046		×
17	39183171	rs146063347		
17	39183254	rs138758776		×
17	39183304	rs62623375		×
17	39183313	rs148449559		
17	39190830	rs62622849		×
17	39197609	rs138200823		×
17	39197615	rs150218495	×	×
17	39216256	rs36006291		
17	39240504	rs11655310		×
17	39240511	rs383835		×
17	39253819	rs201814486		×
17	39254318	rs138296121		×
17	39274491	rs113376601		×
17	39274518	rs9897031		×
17	39296553	rs73983172		×
17	39296715	rs73983173		
17	39305760	rs427961	×	
17	39305956	rs1497383		×
17	39306004	rs238829		×
17	39316482	rs366700		×
17	39316618	rs75030409		
17	39316841	rs444509		×
17	39316870	rs385055		×
17	39323971	rs428371		×
17	39324280	rs55690617	×	×
17	39334133	rs389784		×
17	39334241	rs62067292		
17	39340707	rs398825		×
17	39382950	rs144662088		
17	39383073	rs9902235		×
17	39383351	rs146532415		×
17	39405992	rs192691497		

17	39406409	rs2191379		×
17	39411711	rs150962386		×
17	39421781	rs12938374		×
17	39422063	rs576405629	×	×
17	39422065	rs537301040	×	×
17	39431954	rs188966259		
17	39432017	rs4890107	×	×
17	39464487	rs2074285		×
17	39503163	rs12937519		×
17	39521142	rs34771886		
17	39521468	rs114488848		
17	39521751	rs143499346		
17	39535305	rs61740668		
17	39535388	rs2071599	×	×
17	39535672	rs112570296		
17	39535859	rs2239710		×
17	39551763	rs112544857		
17	39553547	rs6503627		
17	39579112	rs16966811		×
17	39580285	rs201439644		
17	39580559	rs9916475		×
17	39580562	rs9916484		×
17	39580660	rs9916724		×
17	39580739	rs9910204		×
17	39593768	rs897416		
17	39616430	rs2604953		×
17	39619094	rs2604955		×
17	39619115	rs2071563		×
17	39619186	rs2604956		
17	39619193	rs11078993		×
17	39619283	rs16966929		
17	39620565	rs72830046		×
17	39620641	rs146792525		
17	39622068	rs2071561		×
17	39622385	rs2071560		×
17	39623363	rs3744786		×
17	39633349	rs2071601		×
17	39633354	rs12451652		×
17	39635733	rs138303882		
17	39637244	rs743686		×
17	39643340	rs11657323		
17	39643646	rs2301354		×
17	39643860	rs9904102		
17	39643934	rs75790652		×
17	39644900	rs9675246		×
17	39645761	rs8082683		×
17	39659194	rs4796697		×
17	39659913	rs9891361		×
17	39674641	rs1050784		
17	39724592	rs77688767		
17	39739524	rs59780231		

17	39767338	rs112891689		
17	39913771	rs143043662		
17	39925713	rs41283425		
17	76164779	rs142608913		
18	21140367	rs80358251		
18	28605818	rs35296997		
18	43675064	rs79011243		
18	61160287	rs2289519		
18	61170782	rs1455555		
18	61323228	rs61748838		
19	2456877	rs3764582		
19	14590279	rs45458894		
19	15582863	rs34440547		
20	2290333	rs214803		
20	2297790	rs214814		
20	2321105	rs214830		
20	32685225	rs17856024		
21	31744127	rs877346		
21	31744310	rs3804010		
21	32253513	rs9636845		
21	32253629	rs71321355		
21	45978090	rs233252		×
21	45999653	rs464391		×
21	46000445	rs5017208		
21	46011468	rs465279		
21	46032087	rs111668637		
21	46032094	rs411254		×
21	46047857	rs9980129		×
21	46057806	rs4818950		×
21	46057950	rs7280841		
21	46117792	rs34302939		×
21	46117823	rs61745911		×
22	44079728	rs6006438		

APPENDIX Q

Allelic Frequency Table for 233 Candidate Marker Set

			AFR				AMER				EASIA				EUR				SASIA			
CHROM	POS	SNP	Allele 1	Allele 2	Freq. Allele 1	Freq. Allele 2	Allele 1	Allele 2	Freq. Allele 1	Freq. Allele 2	Allele 1	Allele 2	Freq. Allele 1	Freq. Allele 2	Allele 1	Allele 2	Freq. Allele 1	Freq. Allele 2	Allele 1	Allele 2	Freq. Allele 1	Freq. Allele 2
1	59042181	rs14008	T	G	0.3449	0.6551	T	G	0.05331	0.94669	T	G	0.1379	0.8621	T	G	0.0497	0.9503	T	G	0.1074	0.8926
1	152084419	rs201015975	T	A	0.03631	0.96369	T	A	0.001441	0.998559	T	A	0.001984	0.998016	T	A	0	1	T	A	0	1
1	152085505	rs2515663	A	C	0.03858	0.96142	A	C	0.004323	0.995677	A	C	0	1	A	C	0	1	A	C	0	1
1	152190954	rs6659183	T	C	0.04009	0.95991	T	C	0.002882	0.997118	T	C	0	1	T	C	0	1	T	C	0	1
1	152276875	rs116505293	T	C	0.02496	0.97504	T	C	0	1	T	C	0.0009921	0.9990079	T	C	0	1	T	C	0	1
1	152283862	rs58001094	G	C	0.2194	0.7806	C	G	0.4611	0.5389	G	C	0.3423	0.6577	C	G	0.1759	0.8241	C	G	0.4939	0.5061
1	152285188	rs3120655	A	G	0.407	0.593	A	G	0.03314	0.96686	A	G	0	1	A	G	0.000994	0.999006	A	G	0	1
1	152733301	rs4329520	T	A	0.4213	0.5787	A	T	0.4914	0.5086	A	T	0.4177	0.5823	T	A	0.4841	0.5159	A	T	0.4182	0.5818
1	153431406	rs3014837	C	G	0.0121	0.9879	C	G	0.01441	0.98559	C	G	0	1	C	G	0.04274	0.95726	C	G	0.01431	0.98569
1	153520203	rs41265164	A	G	0.4228	0.5772	A	G	0.1037	0.8963	A	G	0.01488	0.98512	A	G	0.05169	0.94831	A	G	0.06339	0.93661
1	153520280	rs116208483	C	G	0	1	C	G	0.007205	0.992795	C	G	0	1	C	G	0.00994	0.99006	C	G	0	1
1	153520954	rs36022742	T	C	0.3896	0.6104	T	C	0.09942	0.90058	T	C	0.01488	0.98512	T	C	0.05169	0.94831	T	C	0.06339	0.93661
1	159828613	rs62624468	T	C	0.06884	0.93116	T	C	0.02161	0.97839	T	C	0	1	T	C	0.01789	0.98211	T	C	0.003067	0.996933
1	161495463	rs142859332	A	G	0	1	A	G	0.005764	0.994236	A	G	0	1	A	G	0.008946	0.991054	A	G	0	1
1	201282334	rs34626929	A	G	0.09304	0.90696	A	G	0.001441	0.998559	A	G	0	1	A	G	0.000994	0.999006	A	G	0	1
1	201288955	rs1626370	A	G	0.09228	0.90772	A	G	0.1859	0.8141	A	G	0.2302	0.7698	A	G	0.1988	0.8012	A	G	0.1074	0.8926
1	201289487	rs110920171	T	C	0.02194	0.97806	T	C	0	1	T	C	0	1	T	C	0	1	T	C	0	1
1	201294910	rs61818256	T	C	0	1	T	C	0.01729	0.98271	T	C	0	1	T	C	0.01292	0.98708	T	C	0.003067	0.996933
1	201458071	rs3738283	A	T	0	1	A	T	0	1	A	T	0.02679	0.97321	A	T	0	1	A	T	0.02658	0.97342
2	28761981	rs6753929	C	G	0.09228	0.90772	C	G	0.304	0.696	C	G	0.2232	0.7768	C	G	0.2425	0.7575	C	G	0.2014	0.7986
2	85622059	rs6886	T	C	0.2511	0.7489	C	T	0.3977	0.6023	T	C	0.496	0.504	T	C	0.3449	0.6551	T	C	0.4918	0.5082
2	113832312	rs6761276	T	C	0.4206	0.5794	T	C	0.4179	0.5821	C	T	0.1796	0.8204	T	C	0.3887	0.6113	T	C	0.3323	0.6677
2	216190020	rs2372536	G	C	0.06203	0.93797	G	C	0.3084	0.6916	G	C	0.2927	0.7073	G	C	0.3161	0.6839	G	C	0.4928	0.5072
2	227886773	rs3752895	A	G	0.3109	0.6891	G	A	0.4841	0.5159	G	A	0.4058	0.5942	G	A	0.4334	0.5666	A	G	0.4448	0.5552
2	233537125	rs11550699	G	A	0.1483	0.8517	G	A	0.464	0.536	G	A	0.4891	0.5109	G	A	0.3469	0.6531	G	A	0.3834	0.6166
2	233897414	rs2233384	A	C	0.03933	0.96067	A	C	0.02738	0.97262	A	C	0.1111	0.8889	A	C	0.000994	0.999006	A	C	0.05521	0.94479
2	233897503	rs2233385	A	G	0	1	A	G	0.03026	0.96974	A	G	0.08333	0.91667	A	G	0.002982	0.997018	A	G	0.006135	0.993865
2	233899057	rs2233390	A	G	0.02496	0.97504	A	G	0.001441	0.998559	A	G	0	1	A	G	0	1	A	G	0	1
2	233899126	rs2233391	A	C	0.1626	0.8374	A	C	0.3026	0.6974	A	C	0.02976	0.97024	C	A	0.498	0.502	A	C	0.2065	0.7935
3	13395579	rs2280084	C	A	0.1029	0.8971	A	C	0.4452	0.5548	A	C	0.3998	0.6002	C	A	0.4533	0.5467	A	C	0.4417	0.5583
3	38052725	rs933135	T	C	0	1	T	C	0.03026	0.96974	T	C	0.1825	0.8175	T	C	0.01491	0.98509	T	C	0.09816	0.90184
3	38167095	rs2229528	G	A	0.002269	0.997731	G	A	0.01441	0.98559	G	A	0	1	G	A	0.04274	0.95726	G	A	0.00409	0.99591
3	111828423	rs340167	G	A	0.08548	0.91452	G	A	0.004323	0.995677	G	A	0	1	G	A	0	1	G	A	0	1
3	194080916	rs13070515	A	G	0.09304	0.90696	A	G	0.1052	0.8948	A	G	0.06845	0.93155	A	G	0.2107	0.7893	A	G	0.1104	0.8896
3	194080983	rs13060627	T	C	0.1929	0.8071	T	C	0.1412	0.8588	T	C	0.06845	0.93155	T	C	0.2406	0.7594	T	C	0.1278	0.8722
5	73981270	rs820878	T	C	0.002269	0.997731	T	C	0.01729	0.98271	T	C	0	1	T	C	0.04076	0.95924	T	C	0.03783	0.96217
5	73992881	rs10805890	G	A	0.009077	0.990923	G	A	0.1988	0.8012	G	A	0.2609	0.7391	G	A	0.1461	0.8539	G	A	0.08487	0.91513
5	74014637	rs77499935	G	A	0.04085	0.95915	G	A	0.002882	0.997118	G	A	0	1	G	A	0	1	G	A	0	1
6	7581001	rs28763966	A	C	0.1876	0.8124	A	C	0.01009	0.98991	A	C	0	1	A	C	0.000994	0.999006	A	C	0	1
6	7581032	rs28763967	T	C	0.0007564	0.9992436	T	C	0.001441	0.998559	T	C	0	1	T	C	0.0159	0.9841	T	C	0	1
6	7581636	rs6929069	A	G	0.379	0.621	A	G	0.1859	0.8141	A	G	0.1944	0.8056	A	G	0.1352	0.8648	A	G	0.2474	0.7526
6	26108282	rs198844	G	C	0.3653	0.6347	G	C	0.4257	0.5743	G	C	0.3026	0.6974	C	G	0.4719	0.5281	G	C	0.4499	0.5501
6	31170227	rs73728294	G	A	0.1036	0.8964	G	A	0.1009	0.8991	G	A	0.04167	0.95833	G	A	0.09344	0.90656	G	A	0.1288	0.8712
6	31683157	rs3749952	G	T	0.1649	0.8351	G	T	0.08357	0.91643	G	T	0.1438	0.8562	G	T	0.02982	0.97018	G	T	0.06442	0.93558
6	31777946	rs2075800	T	C	0.02042	0.97958	T	C	0.3372	0.6628	T	C	0.3681	0.6319	T	C	0.3489	0.6511	T	C	0.3016	0.6984
7	23300345	rs35363287	T	C	0.06808	0.93192	T	C	0.007205	0.992795	T	C	0	1	T	C	0.003976	0.996024	T	C	0.001022	0.998978
8	17796382	rs412750	A	G	0.1906	0.8094	A	G	0.3674	0.6326	A	G	0.38	0.62	A	G	0.2366	0.7634	A	G	0.4315	0.5685
8	37728017	rs7817179	A	G	0.1634	0.8366	A	G	0.219	0.781	A	G	0.0248	0.9752	A	G	0.2584	0.7416	A	G	0.3078	0.6922
8	144992390	rs62642465	G	C	0.1377	0.8623	G	C	0.008646	0.991354	G	C	0	1	G	C	0.000994	0.999006	G	C	0	1
8	144998444	rs201070743	T	C	0	1	T	C	0.001441	0.998559	T	C	0.03571	0.96429	T	C	0.001988	0.998012	T	C	0.005112	0.994888
8	145001588	rs11136334	T	C	0.02799	0.97201	T	C	0.2954	0.7046	T	C	0.1429	0.8571	T	C	0.4274	0.5726	T	C	0.3476	0.6524
9	139890130	rs2292923	T	G	0	1	T	G	0	1	T	G	0.03472	0.96528	T	G	0	1	T	G	0.001022	0.998978
11	244141	rs7128029	G	A	0.1082	0.8918	G	A	0.17	0.83	G	A	0.1964	0.8036	G	A	0.2694	0.7306	G	A	0.1994	0.8006
11	62288978	rs11828907	C	T	0.04387	0.95613	C	T	0.004323	0.995677	C	T	0	1	C	T	0.000994	0.999006	C	T	0	1
11	67352689	rs1695	G	A	0.4803	0.5197	G	A	0.4755	0.5245	G	A	0.1786	0.8214	G	A	0.331	0.669	G	A	0.2945	0.7055
11	67353579	rs1138272	T	C	0.005295	0.994705	T	C	0.02882	0.97118	T	C	0.0009921	0.9990079	T	C	0.07058	0.92942	T	C	0.07055	0.92945
11	93797619	rs1945783	A	G	0.3275	0.6725	A	G	0.02161	0.97839	A	G	0	1	A	G	0.001988	0.998012	A	G	0	1
11	103029516	rs688906	G	A	0.4372	0.5628	A	G	0.2536	0.7464	A	G	0.2173	0.7827	A	G	0.2654	0.7346	A	G	0.2607	0.7393
11	120008468	rs1670195	C	G	0	1	C	G	0.008646	0.991354	C	G	0	1	C	G	0.01093	0.98907	C	G	0	1
11	124502114	rs76570671	T	C	0.06959	0.93041	T	C	0.005764	0.994236	T	C	0.09226	0.90774	T	C	0	1	T	C	0.002045	0.997955
12	31945000	rs7301923	T	C	0.05068	0.94932	T	C	0.005764	0.994236	T	C	0	1	T	C	0	1	T	C	0	1
12	52631313	rs6580870	A	G	0.08018	0.91982	A	G	0.1614	0.8386	A	G	0.09722	0.90278	A	G	0.1809	0.8191	A	G	0.1462	0.8538
12	52681460	rs4761786	A	G	0	1	A	G	0.004323	0.995677	A	G	0	1	A	G	0.02087	0.97913	A	G	0.002045	0.997955
12	52681925	rs6580873	A	C	0.009863	0.990137	A	C	0.07925	0.92075	A	C	0.003968	0.996032	A	C	0.1481	0.8519	A	C	0.09918	0.90082
12	52685096	rs2071588	G	C	0.0643	0.9357	G	C	0.3199	0.6801	G	C	0.4216	0.5784	G	C	0.328	0.672	G	C	0.3415	0.6585
12	52685213	rs79897879	G	T	0.2284	0.7716	G	T	0.02161	0.97839	G	T	0.1141	0.8859	G	T	0.002982	0.997018	G	T	0.06748	0.93252
12	52698717	rs56677856	G	A	0.06127	0.93873	G	A	0.005764	0.994236	G	A	0	1	G	A	0.000994	0.999006	G	A	0	1
12	52699525	rs139895699	T	C	0.0416	0.9584	T	C	0.004323	0.995677	T	C	0	1	T	C	0	1	T	C	0	1
12	52708420	rs2857671	G	A	0.0348	0.9652	G	A	0.4092	0.5908	G	A	0.3224	0.6776	G	A	0.3588	0.6412	G	A	0.3119	0.6881
12	52709855	rs140635030	T	C	0.04463	0.95537	T	C	0.004323	0.995677	T	C	0	1	T	C	0	1	T	C	0	1
12	52710721	rs2852464	C	G	0.4713	0.5287	C	G	0.2997	0.7003	C	G	0.1885	0.8115	C	G	0.3877	0.6123	C	G	0.3967	0.6033
12	52713088	rs2857663	A	G	0.1589	0.8411	A	G	0.05764	0.94236	A	G	0.08036	0.91964	A	G	0.1034	0.8966	A	G	0.2096	0.7904
12	52760957	rs61630004	T	C	0.01815	0.98185	T	C	0.03458	0.96542	T	C	0	1	T	C	0.04274	0.95726	T	C	0.07464	0.92536
12	52774235	rs951773	A	G	0.05371	0.94629	A	G	0.1455	0.8545	A	G	0.254	0.746	A	G	0.2654	0.7346	A	G	0.1779	0.8221
12	52788928	rs2658658	A	G	0.3374	0.6626	A	G	0.3746	0.6254	A	G	0.06448	0.93552	A	G	0.4702	0.5298	A	G	0.4622	0.5378
12	52788945	rs1732263	G	C	0.2133	0.7867	G	C	0.04179	0.95821	G	C	0	1	G	C	0.03877	0.96123	G	C	0.07771	0.92229
12	52795099	rs1791634	G	C	0.0348	0.9652	G	C	0	1	G	C	0	1	G	C	0	1	G	C	0	1
12	52825839	rs2232393	C	T	0.02874	0.97126	C	T	0.001441	0.998559	C	T	0	1	C	T	0	1	C	T	0	1
12	52827608	rs2232387	T	C	0.1891	0.8109	T	C	0.2089	0.7911	T	C	0.0754	0.9246	T	C	0.1213	0.8787	T	C	0.1268	0.8732
12	52881544	rs11540301	C	G	0.298	0.702	C	G	0.07925	0.92075	C	G	0.05556	0.94444	C	G	0.01292	0.98708	C	G	0.03579	0.96421
12	52966428	rs11170177	C	G	0.005295	0.994705	C	G	0.02594	0.97406	C	G	0.3175	0.6825	C	G	0.05169	0.94831	C	G	0.09407	0.90593
12	52981442	rs11170183	A	C	0.09682	0.90318	A	C	0.1527	0.8473	A	C	0.06548	0.93452	A	C	0.2416	0.7584	A	C	0.3569	0.6431
12	52994990	rs116941214	T	C	0	1	T	C	0	1	T	C	0.01687	0.98313	T	C	0	1	T	C	0	1
12	53044267	rs638043	T	C	0.02345	0.97655	T	C	0.1009	0.8991	T	C	0.02579	0.97421	T	C	0.1571	0.8429	T	C	0.07771	0.92229
12	53045626	rs2634041	C	T	0.2595	0.7405	C	T	0.2695	0.7305	T	C	0.2411	0.7589	C	T	0.2833	0.7167	C	T	0.3078	0.6922
12	53069014	rs14024	C	T	0.02194	0.97806	C	T	0.2478	0.7522	T	C	0.4216	0.5784	C	T	0.331	0.669	C	T	0.2689	0.7311
12	53070174	rs17678945	A	C	0	1	A	C	0.01009	0.98991	A	C	0	1	A	C	0.0169	0.9831	A	C	0.003067	0.996933
12	53085089	rs636127	T	C	0.2065	0.7935	T	C	0.0389	0.9611	T	C	0	1	T	C	0.02684	0.97316	T	C	0.007157	0.992843
12	53186088	rs3887954	C	G	0.3775	0.6225	C	G	0.4539	0.5461	C	G	0.1369	0.8631	G	C	0.328	0.672	C	G	0.2699	0.7301
12	53344142	rs59979366	G	A	0.005295	0.994705	G	A	0	1	G	A	0	1	G	A	0	1	G	A	0	1
12	112241766	rs671	A	G	0.001513	0.998487	A	G	0.002882	0.997118	A	G	0.1736	0.8264	A	G	0	1	A	G	0	1
13	113975768	rs9577230	C	T	0.1074	0.8926	C	T	0.0562	0.9438	C	T	0.1369	0.8631	C	T	0.000994	0.999006	C	T	0.003067	0.996933
14	55609418	rs10148371	A	G	0.1195	0.8805	A	G	0.008646	0.991354	A	G	0	1	A	G	0.000994	0.999006	A	G	0	1
14	55611839	rs11125	T	A	0.0007564	0.9992436	T	A	0.07925	0.92075	T	A	0.005952	0.994048	T	A	0.07853	0.92147	T	A	0.09202	0.90798
15	60653205	rs17845226	A	C	0.007564	0.992436	A	C	0.0951	0.9049	A	C	0	1	A	C	0.1252	0.8748	A	C	0.04192	0.95808
16	4933939	rs2037912	C	G	0.03026	0.96974	C	G	0.4712	0.5288	G	C	0.3482	0.6518	G	C	0.4682	0.5318	C	G	0.4683	0.5317

16	4934286	rs143676756	T	C	0.0007564	0.9992436	T	C	0.001441	0.998559	T	C	0.001984	0.998016	T	C	0.001988	0.998012	T	C	0	1
17	4856580	rs238239	C	T	0.1445	0.8555	C	T	0.4481	0.5519	C	T	0.1548	0.8452	T	C	0.4563	0.5437	C	T	0.3241	0.6759
17	28576076	rs1050565	C	T	0.1989	0.8011	C	T	0.3746	0.6254	C	T	0.1736	0.8264	C	T	0.329	0.671	C	T	0.2474	0.7526
17	38121993	rs3894194	A	G	0.2587	0.7413	G	A	0.4957	0.5043	G	A	0.4603	0.5397	A	G	0.4463	0.5537	A	G	0.4836	0.5164
17	38131187	rs56030650	A	C	0.3132	0.6868	A	C	0.3847	0.6153	A	C	0.4742	0.5258	A	C	0.4573	0.5427	A	C	0.3047	0.6953
17	38859509	rs114431517	C	T	0.02496	0.97504	C	T	0.001441	0.998559	C	T	0	1	C	T	0	1	C	T	0	1
17	39114962	rs7213256	C	T	0.4592	0.5408	T	C	0.1902	0.8098	T	C	0.03274	0.96726	T	C	0.161	0.839	T	C	0.2106	0.7894
17	39116603	rs17843023	T	G	0.0885	0.9115	T	G	0.01009	0.98991	T	G	0.08433	0.91567	T	G	0.01193	0.98807	T	G	0.01022	0.98978
17	39116728	rs17843021	A	G	0.1452	0.8548	A	G	0.05476	0.94524	A	G	0.1399	0.8601	A	G	0.1322	0.8678	A	G	0.08282	0.91718
17	39122852	rs142154718	T	C	0.01437	0.98563	T	C	0	1	T	C	0	1	T	C	0	1	T	C	0	1
17	39134528	rs16968862	A	G	0.2451	0.7549	A	G	0.1037	0.8963	A	G	0.03175	0.96825	A	G	0.1292	0.8708	A	G	0.1892	0.8108
17	39135207	rs150812788	G	A	0.06278	0.93722	G	A	0.317	0.683	G	A	0.251	0.749	G	A	0.2425	0.7575	G	A	0.2883	0.7117
17	39137104	rs9908389	A	G	0.3147	0.6853	G	A	0.2378	0.7622	G	A	0.1716	0.8284	G	A	0.2684	0.7316	G	A	0.2996	0.7004
17	39137297	rs721957	C	T	0.09531	0.90469	C	T	0.3256	0.6744	C	T	0.4157	0.5843	C	T	0.4642	0.5358	C	T	0.3016	0.6984
17	39137387	rs2010027	C	T	0.3147	0.6853	T	C	0.2378	0.7622	T	C	0.1716	0.8284	T	C	0.2684	0.7316	T	C	0.2996	0.7004
17	39139370	rs9908304	G	A	0.3139	0.6861	A	G	0.2378	0.7622	A	G	0.1716	0.8284	A	G	0.2684	0.7316	A	G	0.3006	0.6994
17	39140203	rs140634473	T	C	0.0007564	0.9992436	T	C	0.005764	0.994236	T	C	0	1	T	C	0.00994	0.990066	T	C	0.001022	0.998978
17	39155969	rs3813050	A	G	0.3896	0.6104	G	A	0.2061	0.7939	G	A	0.2302	0.7698	G	A	0.1203	0.8797	G	A	0.1943	0.8057
17	39156027	rs3829598	A	G	0.02648	0.97352	A	G	0.2709	0.7291	A	G	0.2629	0.7371	A	G	0.169	0.831	A	G	0.2188	0.7812
17	39156084	rs9897046	T	C	0.3888	0.6112	C	T	0.2061	0.7939	C	T	0.2312	0.7688	C	T	0.1223	0.8777	C	T	0.1943	0.8057
17	39183171	rs146063347	T	G	0.01059	0.98941	T	G	0.001441	0.998559	T	G	0	1	T	G	0	1	T	G	0	1
17	39183254	rs113758776	C	T	0.1891	0.8109	C	T	0.1167	0.8833	C	T	0.1627	0.8373	C	T	0.0159	0.9841	C	T	0.02454	0.97546
17	39183304	rs62623375	T	C	0.1891	0.8109	T	C	0.1167	0.8833	T	C	0.1627	0.8373	T	C	0.0169	0.9831	T	C	0.02454	0.97546
17	39183313	rs148449559	C	G	0.02194	0.97806	C	G	0	1	C	G	0	1	C	G	0	1	C	G	0	1
17	39190830	rs62622849	T	C	0.1891	0.8109	T	C	0.1167	0.8833	T	C	0.1627	0.8373	T	C	0.0159	0.9841	T	C	0.02454	0.97546
17	39197609	rs138200823	A	C	0.1838	0.8162	A	C	0.1124	0.8876	A	C	0.1409	0.8591	A	C	0.0169	0.9831	A	C	0.0184	0.9816
17	39197615	rs150218495	C	G	0.001513	0.998487	C	G	0.05331	0.94669	C	G	0.08631	0.91369	C	G	0.003976	0.996024	C	G	0.007157	0.992843
17	39216256	rs36006291	G	C	0.05522	0.94478	G	C	0.008646	0.991354	G	C	0	1	G	C	0	1	G	C	0	1
17	39240504	rs11655310	A	G	0.1127	0.8873	A	G	0.3703	0.6297	G	A	0.3938	0.6062	A	G	0.3847	0.6153	G	A	0.499	0.501
17	39240511	rs383835	A	T	0.3555	0.6445	A	T	0.3919	0.6081	T	A	0.3879	0.6121	A	T	0.3976	0.6024	A	T	0.4969	0.5031
17	39253819	rs201814486	C	G	0.1778	0.8222	C	G	0.02305	0.97695	C	G	0.009921	0.990079	C	G	0.000994	0.999006	C	G	0.002045	0.997955
17	39254318	rs138296121	T	C	0.1392	0.8608	T	C	0.1138	0.8862	T	C	0.1508	0.8492	T	C	0.0159	0.9841	T	C	0.02249	0.97751
17	39274491	rs113376601	T	C	0.1815	0.8185	T	C	0.121	0.879	T	C	0.1508	0.8492	T	C	0.0169	0.9831	T	C	0.02147	0.97853
17	39274518	rs9897031	C	T	0.03782	0.96218	C	T	0.3314	0.6686	T	C	0.4306	0.5694	C	T	0.2863	0.7137	C	T	0.4315	0.5685
17	39296553	rs73983172	A	G	0.1914	0.8086	A	G	0.01729	0.98271	A	G	0	1	A	G	0.000994	0.999006	A	G	0	1
17	39296715	rs73983173	G	C	0.01891	0.98109	G	C	0	1	G	C	0	1	G	C	0	1	G	C	0	1
17	39305760	rs427961	T	C	0.1256	0.8744	T	C	0.008646	0.991354	T	C	0.0009921	0.9990079	T	C	0.000994	0.999006	T	C	0.01022	0.98978
17	39305956	rs1497383	A	G	0.3699	0.6301	G	A	0.4856	0.5144	A	G	0.3532	0.6468	A	G	0.4861	0.5139	G	A	0.4703	0.5297
17	39306004	rs238829	G	A	0.1377	0.8623	G	A	0.01009	0.98991	G	A	0	1	G	A	0	1	G	A	0	1
17	39316482	rs366700	G	C	0.379	0.621	G	C	0.02882	0.97118	G	C	0	1	G	C	0.001988	0.998012	G	C	0	1
17	39316618	rs75030409	C	T	0.05068	0.94932	C	T	0.004323	0.995677	C	T	0	1	C	T	0	1	C	T	0	1
17	39316841	rs444509	T	A	0.3782	0.6218	T	A	0.02882	0.97118	T	A	0	1	T	A	0.001988	0.998012	T	A	0	1
17	39316870	rs385055	C	T	0.466	0.534	C	T	0.03314	0.96686	C	T	0	1	C	T	0.001988	0.998012	C	T	0	1
17	39323971	rs428371	A	G	0.3759	0.6241	A	G	0.02882	0.97118	A	G	0	1	A	G	0.001988	0.998012	A	G	0	1
17	39324280	rs55690617	G	A	0.2905	0.7095	G	A	0.02305	0.97695	G	A	0	1	G	A	0.001988	0.998012	G	A	0	1
17	39334133	rs389784	T	C	0.06051	0.93949	T	C	0.02594	0.97406	T	C	0.05159	0.94841	T	C	0.06759	0.93241	T	C	0.02249	0.97751
17	39334241	rs62067292	C	G	0.002269	0.997731	C	G	0.01729	0.98271	C	G	0	1	C	G	0.07157	0.92843	C	G	0.009202	0.990798
17	39340707	rs398825	C	T	0.06051	0.93949	C	T	0.02594	0.97406	C	T	0.05159	0.94841	C	T	0.06759	0.93241	C	T	0.02249	0.97751
17	39382950	rs144662088	A	G	0.02421	0.97579	A	G	0	1	A	G	0	1	A	G	0	1	A	G	0	1
17	39383073	rs9902235	G	C	0.3956	0.6044	G	C	0.1671	0.8329	G	C	0.3681	0.6319	G	C	0.2545	0.7455	G	C	0.1953	0.8047
17	39383351	rs146532415	C	T	0.04236	0.95764	C	T	0.002882	0.997118	C	T	0	1	C	T	0	1	C	T	0	1
17	39405992	rs192691497	T	C	0	1	T	C	0	1	T	C	0.01687	0.98313	T	C	0	1	T	C	0	1
17	39406409	rs2191379	C	A	0.4947	0.5053	C	A	0.1988	0.8012	C	A	0.369	0.631	C	A	0.2664	0.7336	C	A	0.1953	0.8047
17	39411711	rs150962388	G	C	0.1566	0.8434	G	C	0.01585	0.98415	G	C	0	1	G	C	0.000994	0.999006	G	C	0	1
17	39421781	rs12938374	G	A	0.497	0.503	G	A	0.1988	0.8012	A	G	0.37	0.63	A	G	0.2664	0.7336	A	G	0.1953	0.8047
17	39422063	rs576405629	G	A	0.1392	0.8608	G	A	0.4841	0.5159	G	A	0.2897	0.7103	G	A	0.4066	0.5934	G	A	0.4039	0.5961
17	39422065	rs537301040	C	T	0.1392	0.8608	C	T	0.4841	0.5159	C	T	0.2897	0.7103	C	T	0.4056	0.5944	C	T	0.4039	0.5961
17	39431954	rs188966259	G	C	0.001513	0.998487	G	C	0	1	G	C	0	1	G	C	0	1	G	C	0	1
17	39432017	rs4890107	T	C	0.2375	0.7625	C	T	0.4107	0.5893	T	C	0.38	0.62	C	T	0.4891	0.5109	C	T	0.4039	0.5961
17	39464487	rs2074285	G	C	0.3865	0.6135	C	G	0.2435	0.7565	C	G	0.3968	0.6032	C	G	0.2992	0.7008	C	G	0.2229	0.7771
17	39503163	rs12937519	A	G	0.3918	0.6082	A	G	0.1902	0.8098	A	G	0.37	0.63	A	G	0.2644	0.7356	A	G	0.1922	0.8078
17	39521142	rs134771886	T	C	0	1	T	C	0.007205	0.992795	T	C	0	1	T	C	0.02187	0.97813	T	C	0	1
17	39521468	rs114488848	G	C	0.008321	0.991679	G	C	0	1	G	C	0	1	G	C	0	1	G	C	0	1
17	39521751	rs143499346	G	C	0	1	G	C	0	1	G	C	0.0377	0.9623	G	C	0	1	G	C	0	1
17	39535305	rs61740668	A	G	0.0007564	0.9992436	A	G	0.008646	0.991354	A	G	0.005952	0.994048	A	G	0.006958	0.993042	A	G	0.003067	0.996933
17	39535388	rs2071599	C	T	0.4297	0.5703	C	T	0.08501	0.91499	C	T	0.07738	0.92262	C	T	0.1024	0.8976	C	T	0.04908	0.95092
17	39535672	rs112570298	A	G	0.002269	0.997731	A	G	0.01009	0.98991	A	G	0	1	A	G	0.01193	0.98807	A	G	0	1
17	39535859	rs2239710	A	G	0.03177	0.96823	A	G	0.3256	0.6744	A	G	0.4127	0.5873	A	G	0.3549	0.6451	A	G	0.2434	0.7566
17	39551763	rs112544857	A	G	0.002269	0.997731	A	G	0.02882	0.97118	A	G	0.02183	0.97817	A	G	0.02386	0.97614	A	G	0.02658	0.97342
17	39553547	rs78503627	A	G	0.03707	0.96293	A	G	0.03314	0.96686	A	G	0.05159	0.94841	A	G	0.07853	0.92147	A	G	0.02147	0.97853
17	39579112	rs16966811	A	G	0.03782	0.96218	A	G	0.1268	0.8732	A	G	0.3224	0.6776	A	G	0.1759	0.8241	A	G	0.2331	0.7669
17	39580285	rs201439644	C	T	0.0007564	0.9992436	C	T	0	1	C	T	0	1	C	T	0	1	C	T	0	1
17	39580559	rs9916475	A	T	0.1846	0.8154	A	T	0.3415	0.6585	A	T	0.4067	0.5933	A	T	0.4592	0.5408	T	A	0.3793	0.6207
17	39580562	rs9916484	C	T	0.1846	0.8154	C	T	0.3415	0.6585	C	T	0.4067	0.5933	C	T	0.4592	0.5408	T	C	0.3793	0.6207
17	39580660	rs9916724	C	T	0.1846	0.8154	C	T	0.3415	0.6585	C	T	0.4067	0.5933	C	T	0.4592	0.5408	T	C	0.3793	0.6207
17	39580739	rs9910204	A	C	0.1505	0.8495	A	C	0.2147	0.7853	A	C	0.08433	0.91567	A	C	0.2833	0.7167	A	C	0.3875	0.6125
17	39593768	rs897416	A	G	0.06127	0.93873	A	G	0.07637	0.92363	A	G	0.0009921	0.9990079	A	G	0.1163	0.8837	A	G	0.0818	0.9182
17	39616430	rs2604953	G	T	0.2511	0.7489	G	T	0.219	0.781	G	T	0.2252	0.7748	G	T	0.05368	0.94632	G	T	0.1421	0.8579
17	39619094	rs2604955	T	C	0.3691	0.6309	T	C	0.4294	0.5706	T	C	0.3185	0.6815	T	C	0.327	0.673	T	C	0.498	0.502
17	39619115	rs2071563	A	G	0.4259	0.5741	A	G	0.402	0.598	A	G	0.3363	0.6637	A	G	0.4245	0.5755	A	G	0.2444	0.7556
17	39619186	rs2604956	C	G	0.04085	0.95915	C	G	0.03746	0.96254	C	G	0.04762	0.95238	C	G	0.07654	0.92346	C	G	0.02761	0.97239
17	39619193	rs11078993	T	C	0.05295	0.94705	T	C	0.1931	0.8069	T	C	0.2073	0.7927	T	C	0.04672	0.95328	T	C	0.1401	0.8599
17	39619283	rs16966929	A	G	0.01967	0.98033	A	G	0	1	A	G	0	1	A	G	0	1	A	G	0	1
17	39620565	rs72830046	T	C	0.04312	0.95688	T	C	0.196	0.804	T	C	0.1141	0.8859	T	C	0.2724	0.7276	T	C	0.3241	0.6759
17	39620641	rs146792529	T	C	0.01967	0.98033	T	C	0.001441	0.998559	T	C	0	1	T	C	0	1	T	C	0	1
17	39622068	rs2071561	T	G	0.2413	0.7587	G	T	0.4813	0.5187	G	T	0.3631	0.6369	T	G	0.4761	0.5239	G	T	0.2822	0.7178
17	39622385	rs2071560	G	A	0.05825	0.94175	G	A	0.05187	0.94813	G	A	0.09623	0.90377	G	A	0.000994	0.999006	G	A	0.01227	0.98773
17	39623363	rs3744786	C	T	0.1672	0.8328	C	T	0.3963	0.6037	C	T	0.3413	0.6587	C	T	0.3032	0.6968	C	T	0.4959	0.5041
17	39633349	rs2071601	G	C	0.3646	0.6354	G	C	0.4568	0.5432	G	C	0.3462	0.6538	G	C	0.4612	0.5388	G	C	0.3476	0.6524
17	39633354	rs112451652	T	C	0.07413	0.92587	T	C	0.1225	0.8775	T	C	0.2956	0.7044	T	C	0.172	0.828	T	C	0.2219	0.7781
17	39635733	rs138303882	A	G	0.004539	0.995461	A	G	0.001441	0.998559	A	G	0	1	A	G	0	1	A	G	0	1
17	39637244	rs743686	A	G	0.2648	0.7352	A	G	0.4409	0.5591	G	A	0.3333	0.6667	A	G	0.4602	0.5398	A	G	0.3476	0.6524
17	39643340	rs11657323	G	T	0.1755	0.8245	G	T	0.1239	0.8761	G	T	0.003968	0.996032	G	T	0.2316	0.7684	G	T	0.1605	0.8395
17	39643646	rs2301354	A	G	0.4024	0.5976	G	A	0.4553	0.5447	A	G	0.3323	0.6677	G	A	0.4602	0.5398	G	A	0.3476	0.6524
17	39643860	rs9904102	A	G	0.01664	0.98336	A	G	0	1	A	G	0.0009921	0.9990079	A	G	0	1	A	G	0.001022	0.998978
17	39643934	rs75790652	C	G	0.0007564	0.9992436	C	G	0.04323	0.95677	C	G	0.07837	0.92163	C	G	0.000994	0.999006	C	G	0.005112	0.994888
17	39644900	rs9675246	T	C	0.1112	0.8888	T	C	0.001441	0.998559	T	C	0	1	T	C	0.000994	0.999006	T	C	0	1
17	39645761	rs80082683	A	G	0.1135	0.8865	A	G	0	1	A	G	0	1	A	G	0.000994	0.999006	A	G	0	1
17	39659194	rs4796697	T	C	0.3283	0.6717	T	C	0.02738	0.97262	T	C	0.0009921	0.9990079	T	C	0.00497	0.99503	T	C	0.02249	0.97751
17	39659913	rs9891361	A	G	0.1702	0.8298	G	A	0.1383	0.8617	G	A	0.09524	0.90476	G	A	0.0994	0.9006	G	A	0.08896	0.91104
17	39674641	rs1050784	T	C	0.407	0.593	T	C	0.2133	0.7867	T	C	0.08532	0.91468	T	C	0.3519	0.6481	T	C	0.2168	0.7832
17	39724592	rs77688767	G	A	0	1	G	A	0	1	G	A	0.01885	0.98115	G	A	0	1	G	A	0	1
17	39739524	rs59780231	T	C	0.001513	0.998487	T	C	0.08934	0.91066	T	C	0.06349	0.93651	T	C	0.007952	0.992048	T	C	0.003067	0.996933
17	39767338	rs112891685	C	T	0.01513	0.98487	C	T	0.002882	0.997118	C	T	0	1	C	T	0	1	C	T	0	1
17	39913771	rs143043662	T	C	0.001513	0.998487	T	C	0.002882	0.997118	T	C	0	1	T	C	0.006958	0.993042	T	C	0.002045	0.997955
17	39925713	rs41283425	T	C	0.002269	0.997731	T	C	0.03314	0.96686	T	C	0	1	T	C	0.05666	0.94334	T	C	0.03783	0.96217
17	76164779	rs142608913	T	G	0.1694	0.8306	T	G	0.01873	0.98127	T	G	0	1	T	G	0	1	T	G	0	1
18	21140367	rs80358251	A	G	0.0007564	0.9992436	A	G	0.004323	0.995677	A	G	0	1	A	G	0.0169	0.9831	A	G	0.002045	0.997955
18	28605818	rs35296997	G	T	0.03026	0.96974	G	T	0.002882	0.997118	G	T	0	1	G	T	0	1	G	T	0	1
18	43675064	rs79011243	A	C	0.08775	0.91225	A	C	0.005764	0.994236	A	C	0	1	A	C	0.000994	0.999006	A	C	0.001022	0.998978
18	61160287	rs2289519	T	C	0.1278	0.8722	T	C	0.3818	0.6182	T	C	0.378	0.622	T	C	0.3091	0.6909	T	C	0.3057	0.6943
18	61170782	rs1455555	G	A	0.1884	0.8116	G	A	0.4092	0.5908	G	A	0.4871	0.5129	G	A	0.4662	0.5338	G	A	0.3415	0.6585
18	61323228	rs61748838	G	C	0.205	0.795	G	C	0.01441	0.98559	G	C	0	1	G	C	0.003976	0.996024	G	C	0.001022	0.998978
19	2456877	rs3764582	T	C	0.01286	0.98714	T	C	0.005764	0.994236	T	C	0.09623	0.90377	T	C	0.02286	0.97714	T	C	0.01329	0.98671
19	14590279	rs45458894	C	G	0.1067	0.8933	C	G	0.1138	0.8862	C	G	0.00496	0.99504	C	G	0.08052	0.91948	C	G	0.06544	0.93456
19	15582863	rs34440547	T	C	0.1581	0.8419	T	C	0.2406	0.7594	T	C	0.2966	0.7034	T	C	0.2227	0.7773	T	C	0.2311	0.7689
20	2290333	rs214803	A	C	0.41	0.59	C	A	0.1671	0.8329	C	A	0.0754	0.9246	C	A	0.162	0.838	C	A	0.1575	0.8425
20	2297790	rs214814	A	G	0.2126	0.7874	A	G	0.07781	0.92219	A	G	0.06845	0.93155	A	G	0.1103	0.8897	A	G	0.04601	0.95399
20	2321105	rs214830	G	C	0.07337	0.92663	G	C	0.268	0.732	G	C	0.1359	0.8641	G	C	0.3439	0.6561	G	C	0.1871	0.8129
20	32685225	rs117856024	A	C	0.02799	0.97201	A	C	0.004323	0.995677	A	C	0	1	A	C	0	1	A	C	0	1
21	31744127	rs877346	T	A	0.1679	0.8321	T	A	0.3473	0.6527	T	A	0.2688	0.7312	T	A	0.4483	0.5517	T	A	0.4172	0.5828
21	31744310	rs13804010	C	G	0.2943	0.7057	C	G	0.09041	0.90959	C	G	0.06151	0.93849	C	G	0.1223	0.8777	C	G	0.1155	0.8845
21	32253513	rs9636845	T	A	0.004539	0.995461	T	A	0.2032	0.7968	T	A	0.4663	0.5337	T	A	0.08847	0.91153	T	A	0.1738	0.8262
21	32253629	rs71321355	T	C	0.04917	0.95083	T	C	0.1455	0.8545	T	C	0.05258	0.94742	T	C	0.1978	0.8022	T	C	0.3119	0.6881
21	45978090	rs233252	C	T	0.1036	0.8964	C	T	0.1527	0.8473	C	T	0.2351	0.7649	C	T	0.04374	0.95626	C	T	0.135	0.865
21	45999653	rs464391	G	C	0.09002	0.90998	G	C	0.05187	0.94813	G	C	0.1964	0.8036	G	C	0.03777	0.96223	G	C	0.1237	0.8763
21	46000445	rs5017208	G	C	0.1142	0.8858	G	C	0.01297	0.98703	G	C	0	1	G	C	0	1	G	C	0	1
21	46011468	rs465279	A	G	0.236	0.764	G	A	0.3689	0.6311	A	G	0.4435	0.5565	G	A	0.2714	0.7286	G	A	0.4029	0.5971
21	46032087	rs111668637	A	G	0.143	0.857	A	G	0.01441	0.98559	A	G	0	1	A	G	0	1	A	G	0	1
21	46032094	rs411254	A	G	0.06203	0.93797	A	G	0.04755	0.95245	A	G	0.1815	0.8185	A	G	0.0497	0.9503	A	G	0.1503	0.8497
21	46047857	rs9980129	C	T	0.4924	0.5076	C	T	0.3256	0.6744	C	T	0.3512	0.6488	C	T	0.2932	0.7068	C	T	0.3517	0.6483
21	46057806	rs4818950	A	G	0.4342	0.5658	G	A	0.3329	0.6671	G	A	0.3333	0.6667	G	A	0.3072	0.6928	G	A	0.3701	0.6299
21	46057950	rs7280841	A	G	0.07791	0.92209	A	G	0.008646	0.991354	A	G	0	1	A	G	0	1	A	G	0.003067	0.996933
21	46117792	rs34302939	G	A	0.3767	0.6233	A	G	0.1383	0.8617	A	G	0.1151	0.8849	A	G	0.1392	0.8608	A	G	0.3722	0.6278
21	46117823	rs61745911	A	G	0.3389	0.6611	A	G	0.06628	0.93372	A	G	0.02778	0.97222	A	G	0.04076	0.95924	A	G	0.2055	0.7945
22	44079728	rs6006438	C	A	0.03026	0.96974	C	A	0.004323	0.995677	C	A	0	1	C	A	0	1	C	A	0	1