A BIOSOCIAL STUDY EXPLORING SELF-REPORTED VS PREDICTED ANCESTRY USING THE VEROGEN FORENSEQ™ SIGNATURE PREP KIT

A thesis presented to the faculty of the Graduate School of Western Carolina University in partial fulfillment of the requirements for the degree of Master of Science in Biology.

By

Xykiera Charde' Sims

Director: Dr. Beverly Collins
Professor
Department of Biology

Committee Members: Dr. Frankie West, Department of Forensic Science
Brittania Bintz, Department of Forensic Science
Maureen Hickman, Department of Biology

July 2021

ACKNOWLEDGMENTS

First, I would like to humbly thank God for granting me this opportunity and using this journey as an unexpected blessing in preparation for my next chapter in life. I would like to thank my committee members, director, and my peers in the biology department for their assistance, commitment, and encouragement. In particular, Dr. Frankie West, Brittania Bintz, and Maureen Hickman. Thank you Dr. Frankie West for all your positivity which kept me motivated, your infinite wisdom and constant problem solving. Thank you Brittania Bintz for your magnitude of scientific knowledge and eagerness to see me thrive. Thank you Maureen Hickman for your willingness and constant excitement about my project.

I also extend sincere thanks to my family and friends, without whom this thesis would not have been possible: Rhonda and Ramon Sims, Otis Hutcherson, TaKiyah Sims, Dr. Timothy (Baba) and Dr. Rita (Nani) Turner, Kaileigh Naylor, and Erica McCurdy. Thank you Mom and Dad for always being by my side and pushing me to become the best version of myself, in which I was destined to be. Thank you Otis for being a proud brother and supporting all of my endeavors, past and present. Thank you TaKiyah for your endless hugs and being my calmness through the storm. Thank you Baba for your mentorship as a fellow scientist and perfectly timed check ins. Thank you Nani for your infinite love and support. Thanks to my best friends for being my passengers on this long road trip. Thank you all for being an amazing support system, a listening ear and believing in my capabilities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIMs - Ancestry Informative Markers

aiSNPs - Ancestry Informative Single Nucleotide Polymorphisms

BP - Base Pairs

CCD - Charge-Coupled Device

CODIS - Combined DNA Index System

C_T - Cycle Threshold

ddNTP - Dideoxynucleotide Triphosphates

DI - Degradation Index

DNA - Deoxyribonucleic Acid

dNTP - Deoxyribonucleotide Triphosphate

dsDNA - Double-Stranded Deoxyribonucleic Acid

FBI - Federal Bureau of Investigation

FSP - ForenSeq Sample Plate

gDNA - Human Genomic DNA

IRB - Institutional Review Board

MCMC - Markov Chain Monte Carlo

MLR - Multinomial Logistic Regression

Mol - Mole

MPS - Massively Parallel Sequencing

mtDNA - Mitochondrial DNA

NDIS - National DNA Index System

ng - Nanogram

NGS - Next Generation Sequencing

ng/mL - Nanogram per Milliliter

ng/µL - Nanogram per Microliter

nM - Nanomolar

NLP - Normalization Library Plate

NTC - No Template Control

PCA - Principal Component Analysis

PCR - Polymerase Chain Reaction

piSNPs - Phenotypic Informative Single Nucleotide Polymorphisms

PLP - Purification Library Plate

Pg - Picogram

pM - Picomolar

qPCR - Quantitative Polymerase Chain Reaction

RB - Reagent Blank

RNA - Ribonucleic Acid

SBS - Sequencing-By-Synthesis

SNP - Single Nucleotide Polymorphism

SPB - Sample Purification Beads

ssDNA - Single-Stranded Deoxyribonucleic Acid

STR - Short Tandem Repeat

UAS - Universal Analysis Software

µL - Microliter(s)

WCU - Western Carolina University

ABSTRACT

A BIOSOCIAL STUDY EXPLORING SELF-REPORTED VS PREDICTED ANCESTRY USING THE VEROGEN FORENSEQ™ SIGNATURE PREP KIT

Xykiera Charde' Sims, M.S.

Western Carolina University (July 2021)

Director: Beverly Collins, Ph.D.

Since the 1900's forensic scientists have generally relied upon short tandem repeats (STRs) as a DNA typing method used for positive identification or exclusion of suspects in crimes. STR typing, while well established, requires comparison to a reference sample for positive identification. New technological advances in massively parallel sequencing (MPS) have expanded forensic DNA analysis beyond traditional STR profiles to include additional markers for ancestry and phenotype estimates which are informative in the absence of reference samples. The Verogen ForenSeq™ Signature Prep Kit, one of two commercially available MPS forensic DNA kits, is used to simultaneously generate results for identity, ancestry, and phenotypic informative single nucleotide polymorphism (SNP) markers in addition to standard STRs. The ancestry SNPs included provide an estimation of biogeographic origin of the sample donor, but little research has been done on how these compare to self-reported assessments of ancestry. This study examined the correlation between self-reported ancestry and ancestry estimations generated using the ForenSeq™ Signature Prep Kit in a group of 12 Western Carolina University students between the ages of 18-24. Participants were asked to complete a detailed demographic survey and submitted a DNA sample for analysis. This research study explored the correspondence of and discrepancies between genetic data and more nuanced concepts of self-

identification and biosocial ancestry. The results of this research may add to the understanding of the interaction between self-identification and the use of ancestry predictions generated by a commonly used forensic DNA kit. Also, the results can inform applications and limitations of the use of the ForenSeq™ panel in a diverse U.S. population.

CHAPTER ONE: INTRODUCTION

The field of forensic biology focuses on the identification and analysis of biological evidence related to medicolegal investigations. Sources of biological evidence can include, but are not limited to, blood and epithelial cells (including those from fingerprints), saliva, hair, and semen, all of which are commonly recovered crime scene sample types and serve as sources for deoxyribonucleic acid (DNA) analysis. DNA analysis is critical to the field of forensics as it can be used for human identification and can provide critical information in criminal investigations.

DNA Analysis Overview

Forensic DNA analysis is a multistep process that starts with isolation of DNA from cellular material through extraction. Extraction is a process in which cell membranes and nuclei are lysed to release genetic material, which is then purified to remove unwanted proteins, macromolecules, ribonucleic acid (RNA), lipids, and cellular debris that can affect downstream analysis. DNA extraction practices include organic and inorganic methods such as phenol chloroform and proteinase K, as well as solid-phase and magnetic bead-based methods (Tan, S. C., 2009). After DNA is extracted, it is quantified to determine the quality and quantity recovered from the sample, as a specific amount of DNA is required for optimal downstream applications. Following quantification, polymerase chain reaction (PCR), is used to amplify DNA targets so that exponential copies of a specific region of interest are made. Following amplification, PCR products can be sequenced using Sanger sequencing, MPS, or fragment analysis can be performed. Resulting data is then analyzed.

Sanger Sequencing

Sanger Sequencing is the standard sequencing method where amplified DNA fragments are combined with a DNA polymerase, a primer, four deoxyribonucleotide triphosphates (dNTP)

and four fluorescently labeled dideoxynucleotide triphosphates (ddNTP) that correspond to the four base pairs (A, T, G, and C). The mixture is heated which allows the double-stranded DNA fragments to denature into single strands. Once denatured, the mixture is cooled, allowing the primer to bind to the template strand. The mixture is heated again, to promote extension of the primer by DNA polymerase. DNA polymerase synthesizes the strand by adding one nucleotide at a time until a dideoxy nucleotide is randomly incorporated. Dideoxy nucleotides lack the 3' -OH group on the nucleotide complex which prevents addition of subsequent nucleotides to the growing DNA chain. This process is repeated for several cycles and the products are separated by size using automated capillary electrophoresis. Smaller fragments travel faster through the capillary, whereas longer fragments travel more slowly. As each fragment travels through the capillary, the fluorescent dye that is bound to the terminating ddNTP is excited by a laser. Emitted light that is characteristic of the fluorophore bound to each ddNTP is captured by a camera and converted to a basecall, compiling a DNA sequence one nucleotide at a time. The Sanger sequencing technique is advantageous because of its accuracy, easy workflow, and cost-effectiveness when sequencing a small number of target regions (Heather, J. M., 2016). Limitations to this technique include inability to sequence in high throughput, low sensitivity, and high sample input requirement.

To date, DNA analysts have commonly used STRs as markers of interest for DNA analysis in criminal investigations. STRs are repetitive sequences that are 2-6 base pairs in length. Due to normal human variation, the number of repeats vary between individuals. Probing multiple STR loci can be highly discriminatory ultimately leading to the irrefutable identification of an individual (Lim, S., 2015). STR fragment sizes are estimated using automated capillary electrophoresis. Initially, a fluorescent dye is attached to one primer in a pair that flanks the

targeted region of interest and the dye is incorporated during PCR. Labelled PCR products are combined with a size standard and are placed on a capillary electrophoresis instrument. During this process, negatively charged DNA fragments travel through a sieving medium inside a capillary towards a positively charged electrode in the presence of an electric current. Smaller DNA fragments travel faster than larger fragments and before reaching the detection window, fragments are separated by size. As fragments pass through the window, a laser beam excites the fluorophore that was previously incorporated during PCR. The fluorescence is color-coded based on the different targets and fluoresce at different emission wavelengths. A charge-coupled device (CCD) camera detects the fluorescence emission and digitally displays the fluorescence as peaks in an electropherogram. In the resulting electropherogram, the peaks indicate the number of consecutive repeats at a specific STR locus. The collection of data for several STR markers is referred to as an STR profile which is then compared to other profiles for positive identification.

In 1997, the U.S. Federal Bureau of Investigation (FBI) established 13 STR markers that encompass the core loci of the combined DNA index system (CODIS) (Lim, S., 2015). These markers include: D3S1358, D5S818, D7S820, D8S1179, D16S539, D13S317, D18S51, D21S11, CSF1PO, FGA, TH01, TPOX and VWA (Lim, S., 2015). In 2017, seven more STR markers were added to establish the current 20 core loci. These additional marker include: D1S1656, D2S1338, D2S441, D10S1248, D12S391, D19S433, D22S1045. From crime scene samples, analysts generate an STR profile in which can be compared to reference profiles and/or database profiles such as those found in the National DNA Index System (NDIS). If no matching profile is found, the usefulness of STR profiling is limited. In these cases, there are no investigative leads generated from the genetic data, since STR markers offer little information about the physical

appearance or biogeographic origin of the donor of the questioned sample (for exceptions, see Hughes, Algee-Hewitt, and Konigsberg 2019; West & Algee-Hewitt, 2020).

Next Generation Sequencing

Recently, new approaches to forensic DNA analysis, such as MPS, or Next Generation Sequencing (NGS), have emerged as important technological tools that can expand the capabilities for forensic biology beyond traditional STR typing. MPS is a high-throughput methodology that enables rapid sequencing of thousands to millions of DNA templates in parallel. MPS methods are categorized into two groups, sequencing by hybridization and sequencing by synthesis (SBS). The sequencing by hybridization method uses arrayed DNA oligonucleotides of known sequences to label the sequencing DNA through a series of hybridizing and washing. Sequencing by hybridization is beneficial in a diagnostic setting using specific probes for identification of disease-related SNPs (Slatko, 2018). The Verogen SBS method uses reversible fluorescently labeled dNTPs that are added all at once, imaged and then cleaved to allow incorporation for the next base (Slatko, 2018). Advantages of MPS include high throughput, the ability to produce both length-based and sequence-based information, low cost per base sequenced, and small amount of input DNA (Heather, J. M., 2016). This technique also allows for simultaneous sequencing of STRs, ancestry informative single nucleotide polymorphisms (aiSNPs) and phenotypic informative single nucleotide polymorphisms (piSNPs) (England & Harbison, 2019). Disadvantages of MPS include short reads, increased error rate compared to sanger sequencing, and higher cost for equipment.

The MiSeq FGx™ Forensic Genomics System is an MPS platform originally released in 2015 by Illumina© and subsequently acquired by Verogen. The Verogen ForenSeq™ Signature Prep Kit (hereafter, the ForenSeq™ Kit) allows for the simultaneous amplification of up to 58

STR loci and 172 SNPs in a single reaction. The 56 aiSNPs included in the ForenSeq™ Panel are analyzed using a principal component analysis (PCA) model (Zheng et al., 2012) within the ForenSeq™ Universal Analysis Software (UAS) with four built-in reference populations (African, European, East Asian, Admixed American) that are clustered into a biplot. The sample tested is plotted and its location in relation to the reference populations is used to estimate the biogeographical ancestry (England & Harbison, 2019). The 24 piSNPs incorporated in the ForenSeq™ panel are used to estimate eye and hair color using a multinomial logistic regression model, HIrisPlex (Walsh, 2013). The donor of a sample is given the probability of having each of the three eye colors (blue, brown, and intermediate) and each of the four hair colors (brown, red, black, and blonde) estimated by the HIrisPlex model.

While there have been several validation studies done using the ForenSeq™ Kit, few studies have addressed how predictions of biogeographic ancestry relate to self-identification of ancestry/race/ethnicity, especially in cases of mixed ancestry in the U.S.. Additional research must be done on the performance of the ancestry and piSNPs across more diverse populations (England & Harbison, 2019) since the idea of diversity in the human population has undergone a dramatic shift since the 20th century (Brown & Armelagos, 2001; Kittles & Weiss, 2003). Genetic research now suggests that populations cannot be neatly and easily divided since factors such as admixture, culture, and the way in which we define ancestry must all be considered (Koenig, Lee, & Richardson, 2008). In efforts to provide additional information about the performance of these markers, this thesis project was developed to explore the relationship between individual self-identifications of ancestry and phenotype and predicted biogeographic ancestry and phenotype estimates using the ForenSeq™ Kit.

Research Overview

A total of 12 Western Carolina University (WCU) undergraduate students between the ages of 18-24 were recruited as a research cohort. This research was approved by the WCU Institutional Review Board (IRB). The participants completed a detailed self-identification survey – including questions regarding ancestry and phenotype using the Qualtrics online survey software. After obtaining informed consent, the participants were asked to provide a non-invasive buccal swab DNA sample. Samples were anonymized and linked to the matching survey data for tracking purposes. Following DNA extraction and quantification, MPS libraries were prepared using the ForenSeq™ Kit. Samples were run on the MiSeq FGx™ instrument but the initial sequencing run resulted in an error. Several steps were taken to troubleshoot but samples failed to generate any data for the purpose of this thesis following various troubleshooting measures. Results will be generated Fall 2021.

Following sequencing, the analysis of biogeographic ancestry and phenotypic estimates will be compared to the self-identifications submitted by the participants in the online survey. Discrepancies between genetic analysis and self-identification will be analyzed and further interrogated as inconsistencies may present major challenges in making positive identifications in cases of missing persons and crime scene samples.

Significance

This research seeks to expand knowledge and application in forensic biology by determining the correspondence between self-reported and genetic prediction of biogeographical ancestry and phenotype using the ForenSeq™ Kit. Additionally, the research will highlight situations of discordance between the two sources of data, and ultimately lead to additional discussions and questions regarding the relationship between these two areas. Inferring

biogeographical ancestry can be helpful in criminal investigations by narrowing down a large pool of suspects (Walsh, 2013). Phenotypic characteristics can be used to predict the physical appearance of a suspect, both of which help to guide investigations when they are at a dead end.

According to a Scripps Howard News Service study of the FBI's Uniform Crime Report, nearly 185,000 cases of homicide and non-negligent manslaughter went unsolved from 1980 to 2019 (Hargrove, 2019). Gaining knowledge about the performance of ancestry and piSNPs in light of self-identification can improve our understanding of the capabilities and limitations of the ForenSeq™ Kit as it becomes routinely used in casework. The ForenSeq™ Kit has the potential to enhance the amount of information provided from forensic DNA samples and therefore can potentially help reduce unsolved cases.

CHAPTER TWO: BACKGROUND

Human diversity is characterized by genetic variation where the majority of variation, 80-90%, is between individuals, and only 10-20% is due to continental population differences (Shriver, 1997). This small margin of genetic variation between populations is due to the rather recent divergence of the human species into continental groups (Shriver, 1997). The 10-20% of genetic variation is largely accounted for in SNPs as they are the most common type of variation. SNPs refer to positions within the human genome that exist in at least two variant forms (alleles) at a frequency of 1% or more (Brookes, 1999). SNP differences are critical for characterizing continental populations as one SNP variant may be commonly found in one population, but rarely in another.

These SNP variants account for physical and physiological differences which reflect continuing adaptions to environmental conditions, genetic drift, and sexual selection (Shriver, 1997). According to Chakraborty et al. (1991), "unique alleles" are SNP variants that are only found in one population. Forensic analysts focus on these types of alleles as they present the largest allele-frequency differences among populations (Chakraborty R., 1991). Currently, over ten million reference SNPs are available in the public database of the Human Genome Project. From this database, 38% of SNPs are very rare, 32% occur at a frequency of 1-5%, 17% at 5-20%, and 13% at 20-50%. The range of SNP frequencies allows for detection of potential differentiation based on evolutionary population histories (Stephens et al., 2001; Rosenberg 2003; Kidd, 2011).

Advantages in Forensics

SNPs offer several advantages in forensic DNA analysis making them potentially more applicable for degraded samples than conventional STRs (Lee, 2017). First, SNPs are abundant,

distributed across the genome at an average of 1 SNP per 1,000 base pairs (bp) (Wang et al. 1998). Abundance allows for a larger search pool of potential testing markers as well as the elimination of poor performing markers (Butler, 2007). SNPs have a relatively low mutation rate, approximately 10^{-8} (Li, 2018), suggesting SNP inheritance is more stable than STRs as their mutation rate is approximately 100 thousand times lower (Butler, 2007). SNPs enable recovery of more information from small target regions, with potential targets as low as 60-80bp. Short amplicons are beneficial when dealing with highly compromised samples in forensically relevant situations resulting in analysis of degraded DNA. SNP markers also play an important role in increasing the power of kinship analyses and family relationship analysis for unidentified remains. SNPs are found to be more suitable for applications such as mitochondrial DNA (mtDNA) testing, ancestry prediction, phenotype prediction, and the use of Y- chromosomal SNPs as lineage markers (Butler, 2007). Additionally, identity-informative SNPs, provide the same function as forensically selected STR loci, as they are used for individualization. These SNPs provide information to genetically differentiate individuals, excluding them from being the source of an evidentiary sample (Budowle, 2008). These applications can be used to gain additional information which may be pertinent for criminal investigations.

Development of Ancestry Informative SNPs

The allele frequency of an aiSNP is defined by a database but differs depending on the population group. Markers, specifically ancestry informative markers (AIMs), are sets of polymorphisms that have been selected based on their inherent informativeness. These markers help define population groups and can be used to construct a panel of aiSNPs, that accurately and confidently predicts biogeographical ancestry of a sample donor. Over the years, there have been several proposed panels of SNPs to predict ancestry but most lack specificity and efficiency for

routine casework. The challenge has been to create a global panel in which a small number of efficient and robust SNPs are analyzed. Historically, one of the most commonly used panels was developed by Kidd and colleagues, composed of 55 aiSNPs tested on 73 known populations with global distinction of seven to eight biogeographical regions (Kidd, 2014).

Kidd Panel

For the development of the aiSNP panel, several sources were used to identify potential targets. The Applied Biosystems database of allele frequencies containing four populations, Japanese, Chinese, European, and African American, the HGDP-CEPH panel, the Kidd lab database, and data collected from an additional 1300 individuals (Kidd, 2014). Candidate SNPs were selected based on pairwise absolute allele frequency differences representing genetic differentiation or F_{st} values, selecting those with the greatest difference between continental populations. To ensure robustness and help with identifying regional distinctions, SNP candidates were balanced, and the unsupervised clustering program STRUCTURE was used (Pritchard, 2000). A finalized panel of 55 aiSNPs was developed and was tested using 73 known populations with resulting allele frequencies published (Kidd, 2014).

STRUCTURE and principal component analysis. STRUCTURE, a program developed in 2000 by Pritchard et. al., analyzes the distribution of genetic differences amongst populations using a Bayesian iterative algorithm (Pritchard, 2000). This algorithm uses a systematic clustering approach applying *Markov Chain Monte Carlo* (MCMC) estimation (Porras-Hurtado, 2013). This process starts by randomly assigning individuals into a predetermined number of groups, then estimating variant frequencies in each group. Following, individuals are re-assigned based on those frequency estimates. This process is repeated several times (in a process known as burn-in) until reliable allele frequency estimates are met for each

population and membership probabilities of individuals to a population are obtained. STRUCTURE performs individual analysis for each assumed population number ranging from one to K. K can be pre-selected by the user and represents a reasonable number of assumed populations based on the sampling regime. During STRUCTURE analysis, membership coefficients, equaling one, are assigned to each individual within the group. These coefficients represent probability of belonging and the sample with the highest coefficient can be considered a member of the group, if admixture is not a factor. Analyses can be run with or without admixture. When admixture is present, membership coefficients are distributed across multiple clusters.

PCA is a non-parametric linear algorithm that enables dimensionality reduction into classes or clusters. PCA is used as a tool for making predictive models and visualizing distance in genetic variation between populations. The XLSTAT Software uses PCA to evaluate effectiveness of these SNPs for distinguishing among populations and to determine the major factors accounting for the population frequencies.

Development of Phenotypic-informative SNPs

To aid criminal investigations, Walsh et al. developed the HIrisPlex system to simultaneously predict eye and hair color of the doner of a DNA sample (Walsh, 2013). The HIrisPlex assay can produce complete profiles with as little as 63 picogram (pg) of input DNA. Walsh reported that HIrisPlex could produce complete profiles in 88% of cases when tested on a variety of simulated forensic casework samples (Walsh, 2013). The HIrisPlex assay includes 23 SNPs and 1 insertion/deletion polymorphism from 11 genes: MC1R, HERC2, OCA2, SLC24A4, SLC45A2, IRF4, EXOC2, TRYP1, TYR, KITLG, and PIGU/ASIP (Walsh, 2013). The HIrisPlex assay uses the IrisPlex model, previously validated, as a prediction tool. Of the 24 total

DNA variants included in the assay, six (HERC2, OCA2, SLC24A4, SLC45A2, TYR, IRF4), are used solely for eye color.

Eye and Hair Color

As previously mentioned, HIrisPlex uses the previously validated IrisPlex model, previously validated, as a prediction tool for eye color (Walsh, 2011). The IrisPlex assay is a single multiplex genotyping system that uses eye color informative SNPs to predict human eye color (specifically blue and brown). Brown eye color is assumed to be reflective of the ancestral human state and prominent everywhere in the world, whereas non-brown eye colors are assumed to be of European origin (Walsh, 2011). Of the six genes used to predict eye color, the HERC2 and OCA2 genes harbor most of the blue and brown eye color genetic variation information (Walsh, 2011). The IrisPlex model predicts eye color using a formula based on a Multinomial Logistic Regression (MLR) developed by Liu et al. (2009). Using the formula, probabilities of an individual being brown, blue, or intermediate are calculated. That individual is then classified as being brown, blue, or intermediate based on the highest probability value. For worldwide distribution, a minimum threshold of 0.7 is used for categorization (Walsh, 2011). Individuals are characterized as undefined if the highest probability is less than 0.7.

Like eye color, the HIrisPlex assay also uses a MLR to predict hair color using 22 DNA variables (Walsh, 2013) . The model groups individuals into four hair color categories, blonde, brown, red, and black with the highest probability value being the hair color indicator. The minor allele of each DNA variant is input and applied to the MLR where alpha and beta values are generated. The HIrisPlex assay also has the capability of predicting density of the hair color, light or dark for blonde and black hair colors, respectively.

The aiSNPs and piSNPs found in the ForenSeq™ Signature Prep Kit by Verogen utilize PCA to estimate biogeographical ancestry and the HIrisPlex model to estimate hair and eye color. Using the 1000 Genomes data, the PCA model was trained on European, East Asian, and African super populations (Verogen, 2018). Based on the unknown sample's aiSNP genotype calls, the ForenSeq™ kit incorporates an Ad-Mixed American projection. The ForenSeq™ Kit uses all 55 aiSNPs found in the Kidd panel (Kidd, 2014) and includes one additional SNP, rs1919550. The ForenSeq™ Kit uses 22 piSNPs found in Walsh's HIrisPlex model (Walsh, 2013). The HIrisPlex model produce eye color probabilities for brown, blue, and intermediate. Additionally, the model produce hair color probabilities for brown, red, black, and blonde.

CHAPTER THREE: MATERIALS AND METHODS

This study entails investigating the association between self-identification and predicted biogeographic ancestry and phenotype using Verogen's ForenSeq™ Signature Prep Kit. A research cohort completed a self-identification survey and submitted a DNA sample to be sequenced. Sequencing data will be analyzed using the ForenSeq™ UAS and predictions will be compared to survey responses to determine discordance.

Research Involving Human Subjects

This study was reviewed by Western Carolina University's IRB via an expedited review application. In preparation, the thesis candidate and all members of the thesis committee completed the IRB training in Biomedical and Social/Behavioral Research. On October 26, 2020, the IRB approved the project: *A Biosocial Study Exploring Self-reported vs Predicted Ancestry using the Verogen ForenSeq™ Signature Prep Kit.* IRB training certificates and approval email can be found in Appendix A.

Recruitment

Following IRB approval, recruitment flyers (Figure B1) were posted around the campus of Western Carolina University. The flyers highlighted the purpose of the study, eligibility, and participation requirements. The flyer instructed prospective participants to contact Brittania Bintz, the project's neutral mediator. This study included a neutral mediator in an effort to eliminate bias and reduce chances of participant identification. Once the prospective participant reached out to the mediator and confirmed their involvement in the study, they were forwarded an instructional email explaining steps for participation. Participants were instructed to pick up a participation packet from the bin labeled "Outgoing" outside the Forensic Science laboratory. Each packet was marked with a unique ID# and included submission instructions. The packet included two consent forms, one to be marked and returned and one for the participant to keep.

Names and signatures were not collected on the consent form. The consent form (Appendix B) was created using the IRB consent form template provided by the University and required participants to acknowledge their consent with a marked "X" instead of their signature to maintain anonymity. The packet also included a QR code to access the online self-identification survey, two buccal swabs, and two buccal swab boxes. Participants were instructed to complete the online survey substituting their name for their unique ID# as well as swab the inside of their cheeks for 30 seconds using the swabs provided. These requirements could be fulfilled by the participant on their own time and once completed, the participant was instructed to seal and return the packet to the same location, placing the packet in the bin labeled "Incoming." The returned packet needed to include the marked consent form and both buccal swabs.

The finalized research cohort included a total of 12 participants with the following unique ID#'s: MDZ743, TSB269, NLY688, FEI489, JUM775, IBX426, FDE636, XDW222, ELW787, CGL584, DDE742, ULF815.

Extraction and Quantification

DNA was extracted from 12 buccal swabs, one per participant, using the PrepFiler® BTA Forensic DNA Extraction Kit and protocol supplied by Applied Biosystems (Applied Biosystems, 2012). This kit utilizes magnetic particles to optimize DNA yield and remove PCR inhibitors. This study followed the protocol for the "Body fluids on swabs" sample type which called for a lysis incubation time of 40 minutes. A control sample, a reagent blank (RB), was added to the experimental batch. The RB is used to monitor contamination which may be introduced during extraction and downstream.

Extracted DNA was quantified using quantitative PCR (qPCR) with the Quantifiler™ Trio DNA Quantification Kit and protocol (Appendix D) also supplied by Applied Biosystems (Applied Biosystems, 2017). This kit enables analysts to obtain a quantitative and qualitative

assessment of the total amount of amplifiable human DNA in each sample. The Quantifiler™ Trio Kit is advantageous as it is highly accurate, allowing for detection of both total human and male DNA in a ratio. Additionally, the kit is highly sensitive with quantification of concentrations as low as 0.005 nanograms per microliter (ng/µL) and generates results in approximately 1 hour (Applied Biosystems, 2017). The qPCR assay was run on the Applied Biosystems® 7500 Real-Time PCR system and included a no template control (NTC) sample for detection of contamination introduced during qPCR. An internal positive control (IPC), included in the kit's reaction mix, was used as a control for the detection of PCR inhibitors. Quantification results for each sample can be found in Table 3.

Quantification Dilutions

The ForenSeq™ protocol recommends a total input of 1 ng genomic DNA (gDNA) in a total volume of 5 µL. To comply with this recommendation, samples were diluted to 0.2 ng/µL based on their total concentration determined during quantification using the following formula: $M_1V_1 = M_2V_2$. M_1 equals the starting concentration in ng/µL for each sample, M_2 was the desired concentration of 0.2 ng/µL, and V_2 was the desired final stock volume. For dilution and to mitigate pipetting error, all samples were prepared with a desired final volume of 100 µL, with the exception of sample NLY658. NLY658 was prepared with a final stock volume of 150 µL so a minimum of 2 µL of DNA extract was pipetted for dilution to reduce pipetting error. V_1 was calculated for each sample and was then subtracted from the total desired volume of the diluted sample to determine the volume of nuclease-free water needed to reach a final concentration of 0.2 ng/µL. Dilution calculations can be found in Table 4.

ForenSeq™ DNA Signature Prep Kit

Amplify and Tag Targets

Once DNA was extracted, quantified, and diluted, a total of 1 ng of gDNA (5 µL of 0.2 ng/µL stock) from each sample was amplified via PCR in a PCR plate (ForenSeq Sample Plate, FSP) using ForenSeq™ Primer Mix B included in the ForenSeq™ Kit. Primer Mix B is a multiplex that contains primer pairs to simultaneously amplify 27 autosomal STRs, 7 xSTRs, 24 ySTRs, Amelogenin, 94 identity-informative SNPs, 56 aiSNPs and 22 piSNPs (2 ancestry-informative SNPs are also used for phenotype prediction) (Verogen, 2018a). Each PCR primer has been designed to contain a 3' target binding site and a 5' adapter sequence that later serves as the Verogen sequencing primer region (Figure 1). Primer pairs are complementary to flanking DNA sequences located upstream and downstream of each target locus (i.e., STR or SNP). Control DNA (2800M) was added as another experimental sample to serve as a positive template control. The "Amplify and Tag Targets" step of the protocol was followed as is adhering to the PCR1 thermal cycler settings. The purpose of this step was to increase the sensitivity of the method and ensure that the resulting DNA fragments were the appropriate length for sequencing (England and Harbison, 2019).

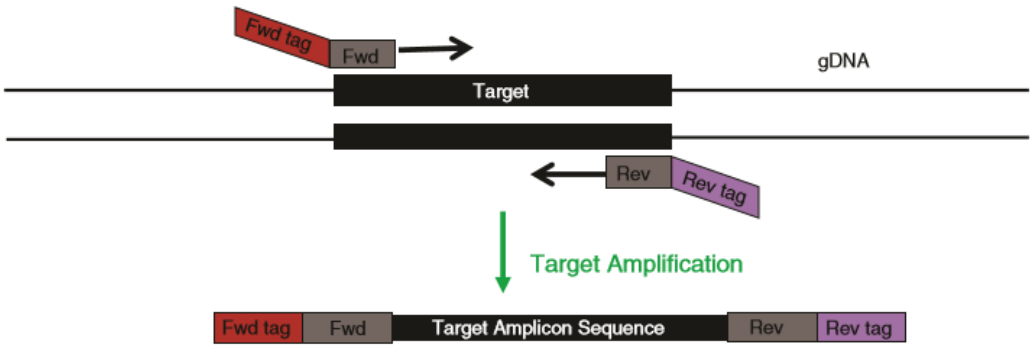


Figure 1. Figure representation of target amplification (PCR1), (England R, Harbison S. (2019, June 17). A review of the method and validation of the MiSeq FGx™ Forensic Genomics Solution. WIREs Forensic Sci. 2019;e1351.https://doi.org /10.1002/wfs2.1351).

Enrich Targets

Table 1: Sequence adapters i5 and i7 added to each sample during PCR2.

Sample (Unique ID#)	i5 Adapter	i7 Adapter
MDZ743	A507 – TAAGTTCC	R711 – GGCTACAT
TSB269	A502 – TGCTAAGT	R711 – GGCTACAT
NLY688	A508 – TAGACCTA	R711 – GGCTACAT
FEI989	A504 – TAAGACAC	R712 – CTTGTAAT
JUM775	A506 – CTAGAACA	R711 – GGCTACAT
IBX426	A501 – TGAACCTT	R712 – CTTGTAAT
FDE636	A502 – TGCTAAGT	R712 – CTTGTAAT
XDW222	A503 – TGTTCTCT	R712 – CTTGTAAT
ELW787	A504 – TAAGACAC	R711 – GGCTACAT
CGL548	A501 – TGAACCTT	R711 – GGCTACAT
DDE742	A505 – CTAATCGA	R711 – GGCTACAT
ULF815	A503 – TGTTCTCT	R711 – GGCTACAT
PC 2800M	A506 – CTAGAACA	R712 – CTTGTAAT
NTC	A505 – CTAATCGA	R712 – CTTGTAAT

Sample specific sequence adapters were added to the PCR1 amplicons in a 15-cycle PCR reaction, known as PCR2 (Verogen, 2018a). The adapters are composed of a 3’ region that is complementary to the forward and reverse tags that were incorporated during PCR1, a barcoding index, and a flow cell adapter. The complementary region binds to products generated during PCR1 and serves as a primer during PCR2 (Figure 2). The i5 and i7 indices serve as barcodes for

respective samples, allowing data from each tagged library to be bioinformatically parsed after sequencing (Table 1). The ForenSeq™ Kit includes eight different i5 index sequences and 12 different i7 sequences, which offers a unique barcoding strategy for multiplexing up to 96 individual samples, 20 of which can be sequenced in a single run using primer set B with positive and negative controls. The P5 and P7 adapters allow the enriched targets to bind to the surface of a flow cell where cluster generation and sequencing take place.

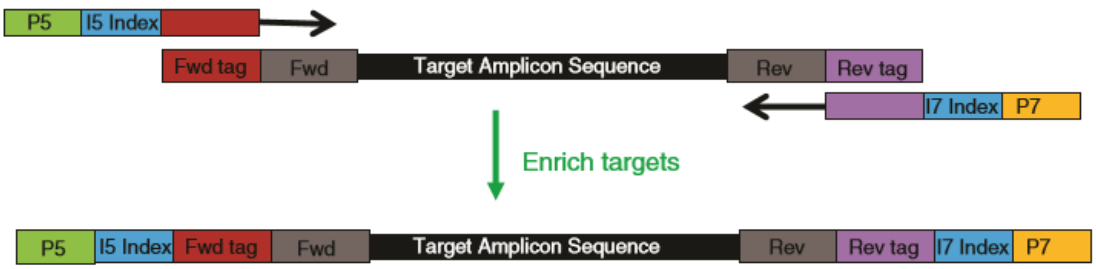


Figure 2: Figure representation of target enrichment (PCR2), (England R, Harbison S. (2019, June 17). A review of the method and validation of the MiSeq FGx™ Forensic Genomics Solution. WIREs Forensic Sci. 2019;e1351.https://doi.org /10.1002/wfs2.1351).

Purification and Bead-based Normalization of Libraries

Libraries were purified in a PCR plate (Purified Library Plate, PLP) using Sample Purification Beads (SPB) and several washes as per the manufacturer's instructions (Verogen, 2018a). During this process, DNA molecules bind to the surface of the SPBs while excess reaction components and small DNA fragments are removed, optimally selecting for fragments between 200-600 base pairs (Jäger et al., 2017). Each library is then normalized in a new PCR plate (Normalized Library Plate, NLP) to an equal molarity, ensuring an equal representation while sequencing. This step was important for generating the same amount of sequencing clusters per individual library. This minimizes the chance of allele and locus drop out, both of which could impact DNA profile interpretations (England and Harbison, 2019).

Library Pooling, Denaturing and Dilution

Once normalized, 5 µL of each library and positive/negative controls were pooled (England R, Harbison S, 2019). Using a combination of heat and 2N sodium hydroxide, the libraries were denatured to single-stranded DNA (ssDNA) to enable binding to the flow cell. Following, libraries were diluted with hybridization buffer and loaded into the sequencing cartridge.

MiSeq FGx™ Instrument

Cluster Generation

The pooled library (600 µL total) was loaded onto the MiSeq FGx™ reagent cartridge. The flow cell was cleaned and loaded into the MiSeq FGx™ instrument. The flow cell is an optically transparent glass slide with an etched fluidic lane containing a lawn of short oligonucleotides bound to its surface. These oligonucleotides have sequences that are complementary to the P5 and P7 library adapters incorporated to the experimental samples during enrichment. During the initial stages of sequencing, the ssDNA library is washed over the flow cell and library fragments hybridized to the oligonucleotide anchors. A complementary strand of each fragment is synthesized with the hybridized oligonucleotide acting as a primer. The duplexed DNA is then denatured, and the original strand is washed away leaving the copied strand covalently bound to the flow cell surface. The ssDNA folds over and hybridizes to an adjacent oligonucleotide and a complementary strand is synthesized. This process is known as "bridge amplification" and repeats for several cycles until a small clonal cluster is formed in a process known as "cluster generation" (England R & Harbison S, 2019). After the cluster is formed, all the reverse stands are removed leaving only the forward strands for sequencing.

Sequencing by Synthesis

The MiSeq FGx™ uses a SBS technique in which each of the clusters are sequenced base-by-base. To begin, a sequencing primer hybridizes to the 3' end of the P5 adapter of the

forward strand. A polymerase and all four modified nucleotide bases are then washed over the flow cell. These modified nucleotides contain a base-specific fluorescent tag and a reversible blocking group on the 3' hydroxyl to avoid binding of multiple nucleotides in any given cycle. After incorporation, the blocking group and fluorophore are removed from the incorporated nucleotide, the flow cell is washed and a fluorescence signal for each cluster is captured by a camera. This process continues for several cycles and as each base is added, the light they emit is recorded. Once the desired cycles have been reached, the product strand is removed, and the process is repeated with the reverse strand.

Troubleshooting

Following initial sequencing, an error message appeared on the MiSeq FGx™ instrument indicating a failed sequencing run. The error message received was " Best focus is too near edge of range" indicating that there was insufficient cluster density for the instrument to focus, resulting in the run failure. Technical support was sought from Verogen and a series of troubleshooting steps were taken.

Bioanalyzer System

First, library sample were run on an Agilent 2100 Bioanalyzer system using the Agilent DNA 1000 Kit and protocol. A total of 1 µL of each sample from the NLP was run on the Bioanalyzer and no DNA was detected. Following, 1 µL of each sample from the PLP, the previous step, were run to determine if any amplification could be detected before normalization and positive amplification results were observed, demonstrating that the issue was with the normalization step. Bead-based normalization was attempted again and the samples run on the Bioanalyzer; however, no DNA was detected.

Manual Normalization

Since the bead-based normalization step was observed to be problematic, the next step was to proceed with manual normalization. For this step, we followed a protocol obtained from Verogen's technical support team (Appendix D). To begin the manual normalization, the concentration of each sample prior to normalization was calculated, utilizing stock of each sample following purification. Total concentrations were found using the Qubit® Fluorometer instrument and Qubit® dsDNA High Sensitivity Assay Kit and protocol. A total of 1 µL of sample was used and concentration results can be found in Table 5. After quantifying, 2nM dilutions were made for each sample. Equal volumes (5 µL) of each sample were pooled together. Following the MiSeq FGx™ guidelines, libraries were denatured and diluted. Two samples which had low concentrations (NLY688 and ELW787) were discarded from the study and the remaining ten were pooled at a concentration of 10 picomolar (pM). A total of 600 µL of the pooled library was loaded onto the Verogen MiSeq FGx™ sequencing cartridge and run on a Micro flow cell which accommodates sequencing of 10 samples and positive and negative controls. However, the second run attempt produced the same error as the initial run, indicating insufficient cluster production. Further trouble shooting will be used to determine the source of the issues. One notable issue is the low concentration of the positive control (2800M), which may indicate an issue with the initial amplification steps in the protocol. The ForenSeq™ Kits underwent a thaw event caused by a broken freezer, which may account for a reduction in amplification efficiency. Additional attempts will be made to amplify the samples in the fall.

CHAPTER FOUR: RESULTS

Self-identification Survey

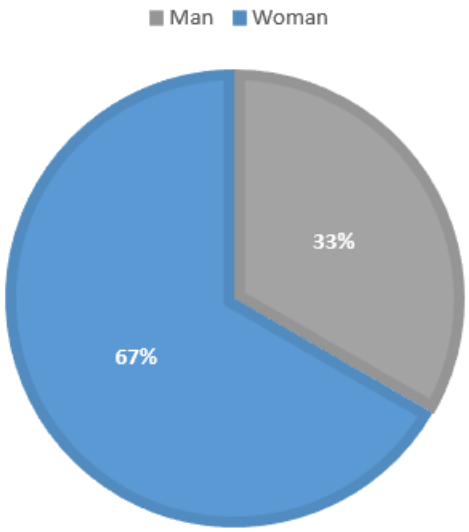


Figure 3. Responses to the gender identification question from the Self-identification survey.

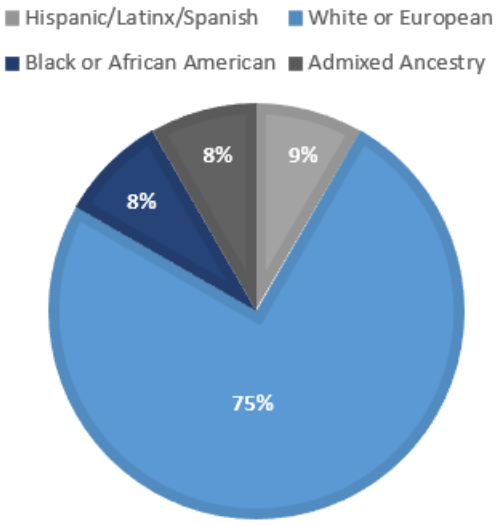


Figure 4. Responses to the Ancestry identification question from the Self-identification survey.

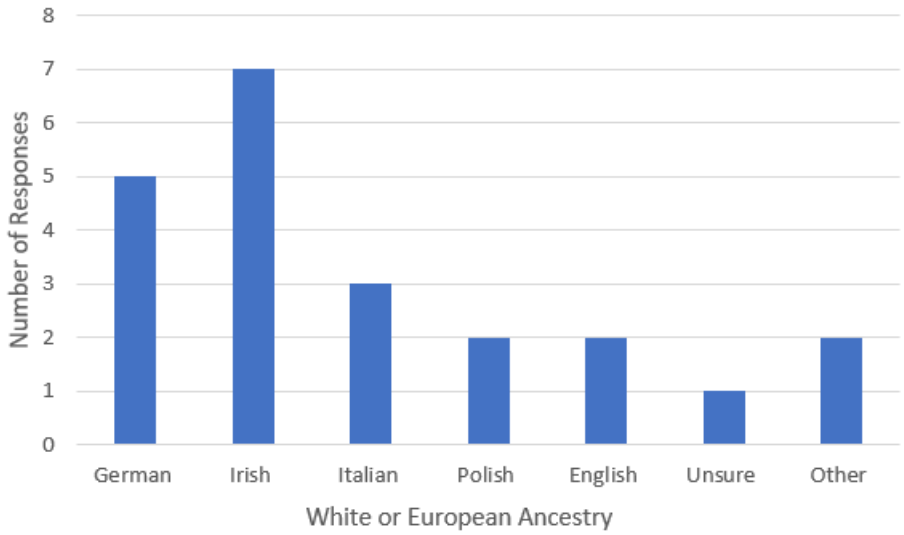


Figure 5. Population designations within the White or European population.

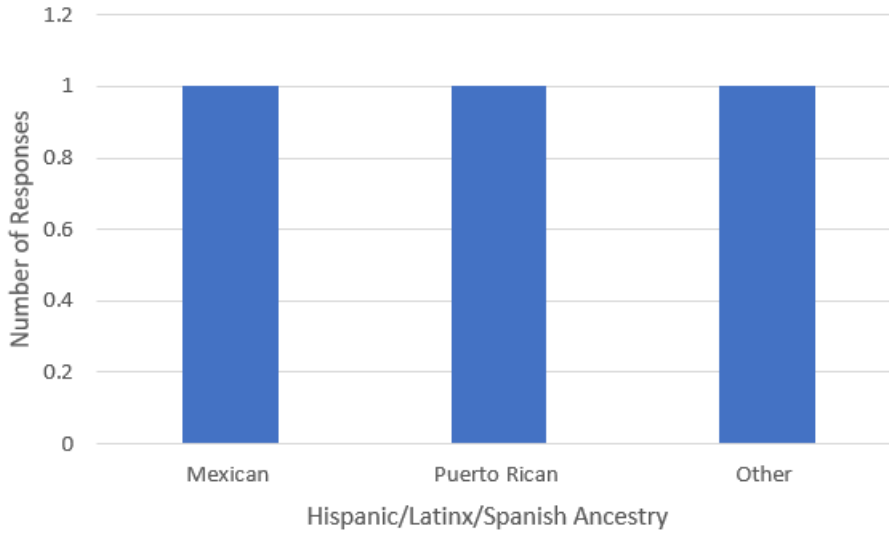


Figure 6. Population designations within the Hispanic/Latinx/Spanish population.

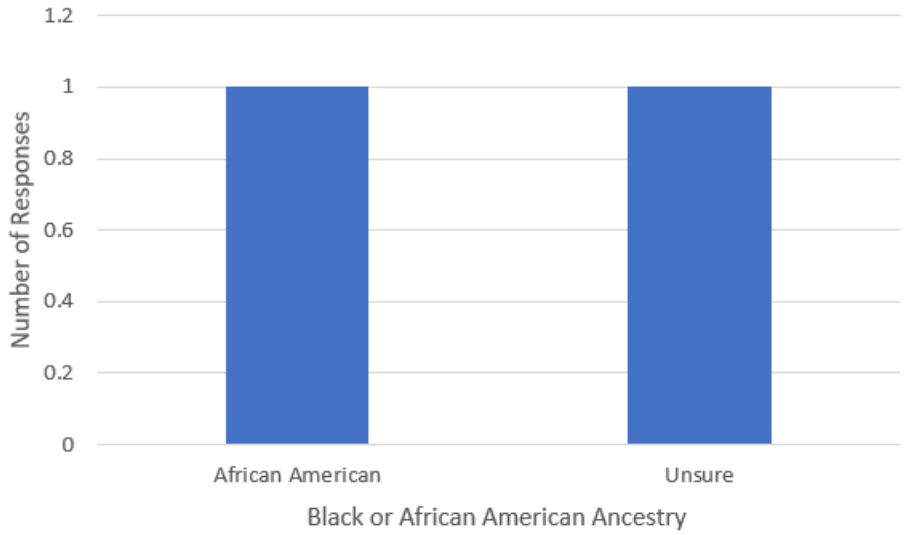


Figure 7. Population designations within the Black or African American population.

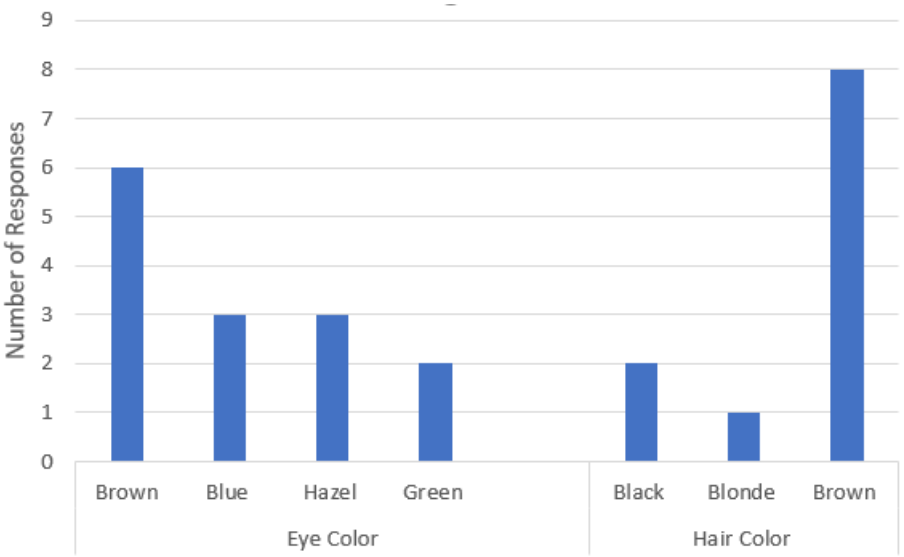


Figure 8. Responses to the eye and hair color identification questions from the Self-identification survey.

The self-identification survey was completed online through the Qualtrics online survey software in its entirety by all participants. The self-identification survey instrument and self-identification survey responses can be found in Appendix C. Of the 12 participants in the research cohort, 67% identified as women and 33% identified as men (Figure 3). Choices also included a write in choice and preference to not report gender identity. A total of 75% of the participants

identified as White or European, 9% as Hispanic/Latinx/Spanish, 8% as Black or African American, and 8% with Admixed Ancestry, identifying with more than one continental population. The 75% White or European participants can be further divided into the following populations based on selected ancestry: German, Irish, Italian, Polish, English and Other (Figure 5). Population "Other" can be classified as Scottish and Dutch, both of which were written responses from participants. The 9% Hispanic/Latinx/Spanish participants can be further divided into the following populations based on selected ancestry: Mexican, Puerto Rican, and Other (Figure 6). For the single "Other" response, the individual reported as Guatemalan. The 8% of participants who selected Black or African American ancestry can be grouped into African American as a subpopulation within the total group (Figure 7). One participant, making up 8% of the research cohort, selected more than one continental population as Admixed. In this case, the participant selected Hispanic/Latinx/Spanish and Black or African American. Figure 8 categorizes responses to questions relevant to self-identifying phenotypic characteristics with brown hair and eye color at the highest frequency.

DNA Quantification

Table 2. Results of the Standard Curve for PCR plate containing experimental samples.

Quantifiler™ Trio Targets	**Slope**	**Y-intercept**	**R^2**
Large Autosomal	-3.616	23.721	1.0
Small Autosomal	-3.502	25.156	0.999
Male (T.Y)	-3.564	25.262	0.999

Table 3. Quantifiler™ Trio results of the small autosomal target for the experimental samples.

Sample (Unique ID#)	**Quantity mean (ng/µL)**	**C_T mean**	**Degradation index (DI)**
MDZ743	5.74	22.57	0.8096
TSB269	6.90	22.22	1.2020
NLY688	13.49	21.23	1.4906
FEI989	2.24	23.93	0.9912

JUM775	3.90	23.22	0.8904
IBX426	3.39	23.32	0.8921
FDE636	2.58	23.72	1.0795
XDW222	1.42	24.62	0.7030
ELW787	1.91	24.27	1.0106
CGL548	2.52	30.40	0.4065
DDE742	3.80	23.27	1.0497
ULF815	3.20	23.40	1.5166

Table 4. Quantification dilutions for the experimental samples.

Sample (Unique ID#)	**M_1 – Starting concentration (ng/µL)**	**V_1 – amount of solution required (µL)**	**Amount of nuclease-free water required (µL)**	**M_2 – Desired concentration (ng/µL)**	**V_2 – Desired final volume (100-150 µL)**
MDZ743	5.74	3.48	96.52	0.2	100
TSB269	6.90	2.90	97.10	0.2	100
NLY688	13.49	2.22	147.78	0.2	150
FEI989	2.24	8.93	91.07	0.2	100
JUM775	3.90	5.13	94.87	0.2	100
IBX426	3.39	5.90	94.10	0.2	100
FDE636	2.58	7.75	92.25	0.2	100
XDW222	1.42	14.08	85.92	0.2	100
ELW787	1.91	10.47	89.53	0.2	100
CGL548	2.52	7.94	92.06	0.2	100
DDE742	3.80	5.26	94.74	0.2	100
ULF815	3.20	6.25	93.75	0.2	100
NTC	0.00	100	-	-	100

The success of DNA extraction using the PrepFiler BTA™ Forensic DNA Extraction Kit was determined based on DNA yields measured using the Quantifiler™ Trio DNA Quantification Kit (Table 3). The characteristics of the standard curve, including slope, Y-intercept, and R^2, can be found in Table 2. The slope is an indicator of PCR amplification efficiency with a slope of -3.3 indicating 100% efficiency (Applied Biosystems, 2017). The manufacturers provides an acceptable range of standard curve slope values for each target; large autosomal -3.1 to -3.7, small autosomal -3.0 to -3.6, and Male Y target -3.0 to -3.6. The slopes

produced from the PCR plate containing the experimental samples fell within the acceptable ranges provided. The Y-intercept indicates the expected cycle threshold (C_T) value for a sample with Qty = 1 (Applied Biosystems, 2017). The Y-intercept can be used to directly compare the C_T mean value for each sample at that target. R^2 measures the line of best fit between the standard curve regression line and the individual C_T points of the standard reaction (Applied Biosystems, 2017). Manufacturers recommend an R^2 >0.98 which was obtained during the run (Table 2).

Experimental samples contained concentrations of DNA ranging from 1.42 to 13.49 ng/µL for the small autosomal target (Table 3). C_T means ranged from 21.23 to 32.73 as compared to the small autosomal Y-intercept expected C_T value, 25.156 (Table 3). C_T indicates the cycle at which the fluorescence surpassed the background fluorescence. The degradation index (DI) is used to assess the quality of the DNA by comparing the performance of large DNA fragments relative to small DNA fragments (Applied Biosystems, 2017). DI is calculated using the following formula:

$$\frac{\text{Small Autosomal Target DNA Concentration (ng/µL)}}{\text{Large Autosomal Target DNA Concentration (ng/µL)}}$$

Manufactures provide the following interpretation: DI <1 indicates no degradation, DI 1 to 10 indicates moderate degradation, and DI >10 indicates significant degradation (Applied Biosystems, 2017). DI calculated for the experimental samples did not exceed moderate degradation (Table 3). As expected, the RB and NTC produced no indication of the presence of DNA in the samples. This is an indicator no contaminant DNA was introduced into the samples during extraction and quantification. The IPC C_T values produced no flags indicating normal

amplification efficiency and no detection of PCR inhibitors. Post quantification, samples were diluted to 0.2 ng/µL for a total of 1 ng input for the ForenSeq™ Kit (Table 4).

ForenSeq™ DNA Signature Prep Kit

This study failed to produce any sequencing results using the ForenSeq™ Kit and MiSeq FGx™ instrument. After the initial failed sequencing run, samples were normalized a second time with the addition of four successfully sequenced samples from a previous study used as positive controls.

Troubleshooting

Bioanalyzer Analysis

Following the second ForenSeq™ run, all 12 experimental samples, four control samples, plus negative and positive controls failed to produce results using the Agilent DNA 1000 Kit and Agilent 2100 Bioanalyzer system. A total of 1 µL of sample from the NLP was run on the Bioanalyzer and no DNA was detected. Samples from the PLP were run to determine if DNA could be detected prior to normalization and positive amplification results were observed, demonstrating that the issue was with the normalization step. An additional normalization step was performed and the samples run on the Bioanalyzer; however, no DNA was detected.

Manual Normalization

Table 5. Concentration of pre-normalized samples using the Qubit® 2.0 Fluorometer.

Sample (Unique ID#)	Concentration (ng/mL)	Concentration (ng/µL)
MDZ743	1.39	0.00139
TSB269	2.62	0.00262
NLY688	Sample too low <0.50	-
FEI989	5.19	0.00519
JUM775	2.79	0.00279
IBX426	0.86	0.00086
FDE636	5.58	0.00558
XDW222	5.41	0.00541
ELW787	Sample too low <0.50	-
CGL548	0.88	0.00088

DDE742	2.17	0.00217
ULF815	3.79	0.00379
NTC	Sample too low <0.50	-
PC2800M	Sample too low <0.50	-

Table 6. Concentration dilutions of pre-normalized samples.

Sample (Unique ID#)	**M_1 – Starting concentration in nM**	**V_1 – amount of PLP solution required (µL)**	**Amount of nuclease-free water required (µL)**	**M_2 – Desired concentration (nM)**	**V_2 – Desired final volume (5 µL)**
MDZ743	1.5717	-	-	2	5
TSB269	2.9625	3.38	1.62	2	5
FEI989	5.8684	1.704	3.296	2	5
JUM775	3.17	3.17	1.83	2	5
IBX426	0.9724	-	-	2	5
FDE636	6.3094	1.585	3.415	2	5
XDW222	6.1171	1.635	3.365	2	5
CGL548	0.9950	-	-	2	5
DDE742	2.4536	4.0	1.0	2	5
ULF815	4.2854	2.3	2.7	2	5
NTC	-	5	-	-	5
PC2800M	-	5	-	2	5

To begin the second phase of troubleshooting, pre-normalized samples were quantified using the Qubit® Fluorometer instrument for total DNA concentration (Table 5). Two experimental samples plus negative and positive controls produced concentration readings that were "too low," less than 0.50 ng/mL. Due to the uncertainty of the exact concentration of the two experimental samples, they were removed from downstream analysis. It was expected that the negative control would produce low results as no DNA was expected to be present in this sample. The positive control was unexpectedly low and many indicate an issue with the amplification efficiency of the kit.

Qubit® concentration readings were measured in ng/mL. To comply with the manual normalization protocol, concentrations were converted to ng/µL (Table 5). Concentrations in

ng/µL were used in the following formula to calculate concentration in nM where the average library size is 268 as per Verogen's instructions:

$$\frac{\text{(Concentration in ng/µL)}}{\text{(660 mol * Average Library Size)}} \times 10^6 = \text{Concentration in nM}$$

Dilutions were made using the following formula: $M_1V_1 = M_2V_2$ (Table 6). M_1 equals the starting concentration in nM for each sample, M_2 was the desired concentration of 2 nM, and V_2 was the desired final stock volume, 5 µL. Samples were normalized and pooled together. The two samples which had low concentrations (NLY688 AND ELW787) were not included. Pooled samples at a concentration of 10 pM in 600 µL were loaded onto the Verogen MiSeq FGx™ sequencing cartridge and run on a Micro flow cell which accommodates sequencing of 10 samples plus positive and negative controls. However, the second run attempt produced the same error as the initial run, indicating insufficient cluster production. Further trouble shooting will be used to determine the source of the issues. One notable issue is the low concentration of the positive control, which may indicate an issue with the initial amplification steps in the protocol. The ForenSeq™ Kits underwent a thaw event caused by a broken freezer, which may account for a reduction in amplification efficiency. Additional attempts will be made to amplify the samples in the fall.

CHAPTER FIVE: CONCLUSION

The goal of this study was to explore the association between self-identification and predicted biogeographic ancestry and phenotype using Verogen's ForenSeq™ Signature Prep Kit. Due to the lack of DNA sequencing results, this assessment will be continued in the Fall of 2021. As part of future research, the results will provide additional information about the performance of the aiSNPs and piSNPs found in the ForenSeq™ Kit across a diverse Generation Z population group. Specifically, discrepancies between self-identification and genetic ancestry and phenotypic predictions will be interrogated. If discrepancies occur, at what rate do they occur and where do they occur? To achieve this goal, 12 WCU students ages 18-24 completed a self-identification survey and submitted a DNA sample to be sequenced and analyzed.

Discussion

Survey

All 12 participants completed the self-identification survey. In terms of diversity, the research cohort lacked individuals with Asian, Middle Eastern, Native Hawaiian/Pacific Islander, Native American/American Indian/Alaskan Native ancestry, likely due to low participation numbers. The lack of participation was largely due to the Covid-19 pandemic and the disruption in normal student schedules and access to campus. Scottish, Dutch, and Guatemalan are ancestral populations not listed on the self-identification survey but filled in by participants. Two out of 12 participants, 16.66%, listed "unsure" about their ancestral subpopulation indicating a fairly high awareness rate (83.34%) of ancestry and culture for those included in this research cohort. This information can be used to investigate Generation Z's awareness of ancestry, culture, and familial origins.

When assessing phenotypic self-identification, three participants listed their eye color as hazel and two participants listed their eye color as green. In terms of predicted estimations using the ForenSeq™ kit, eye color can only be categorized as brown, blue, and intermediate as per the Walsh iris color prediction system (Walsh, 2013). For positive concordance, these participants' predictions would correspond to intermediate for prediction purposes. There was high prevalence in both brown hair and eye color for this research cohort so specific attention should be paid to the performance of the piSNPs used to predict brown eye and hair color.

Quantification

DNA yields produced during quantification indicate successful DNA extraction. Standard curves for all three targets for the qPCR fell within the acceptable range with a slope close to -3.3 indicating near 100% amplification efficiency. Samples produced a wide range of DNA concentrations with minimal degradation. Those samples with slightly higher degradation indices could have be produced by environmental degradation caused by room temperature incubation prior to intake and refrigeration. Time from initial submission to DNA extraction could also contribute to the DI. The RB and NTC samples produced no quantification results indicating no introduced contamination during extraction and quantification.

ForenSeq™ Workflow

It is to be concluded that the failed sequencing run presenting the error message "Best focus is too near edge of range" may be due to issues within the ForenSeq™ Kit. The ten extracts chosen for sequencing in the second run were determined to have the requisite amount of DNA following quantification. Troubleshooting indicated failure within the normalization step as Qubit® results confirmed viable samples post-purification but pre-normalized. Both attempts at normalization (bead-based and manual) resulted in failed sequencing runs. It can be

deduced that although there were amplicons produced, there were not enough amplicons for sufficient cluster generation. In addition, the four previously sequenced samples and the positive control provided with the ForenSeq™ Kit failed to produce any sequencing results, indicating an issue with the bead-based normalization step. Further investigations will assess in more detail issues with the kit and possible options for data generation.

Future Directions

In Fall 2021, samples will be re-sequenced using a newly purchased ForenSeq™ Kit if the existing kit is not viable. Following sequencing, data will be analyzed and compared to the self-identification survey responses to investigate the relationship between the two and determine discordance, if any. It can be hypothesized that discrepancies may occur in individuals with admixed ancestry. This study may prompt additional scrutiny regarding the intersection of culture (biosocial identity) as it applies to forensic biology. In particular, how often are discrepancies encountered between forensic genetic ancestry predictions and cultural designations of race? The results of this study can also provide information pertinent for use of this kit in routine forensic casework.

Applications

DNA intelligence, a new frontier approach to DNA profiling, can be used as an investigative tool for law enforcement when STR profiling leads to a dead end. Specifically, aiSNPs can help narrow down a large suspect pool with the prediction of biogeographical ancestry. Combining aiSNPs with piSNPs, hair and eye color can help create a preliminary sketch and add a visual aspect to a traditional DNA profile.

To expand upon this thesis, a replication study can be done with an increased sample size in an effort to create a more diverse research cohort. Sampling from a diverse population allows

for the reporting of new information about populations that may not be characterized in the UAS algorithm. A separate study can be conducted exploring ForenSeq™ results and interpretation by law enforcement and nonscientific personnel. A simplified and uniform way of interpretating results from the ForenSeq™ Kit regarding ancestry and phenotype predictions should be established across all laboratories and law enforcement jurisdictions. Predicting ancestry and phenotype using the ForenSeq™ Kit may add significantly to the existing tools of forensic biologists, especially when coupled with studies which interrogate the application of such analyses to diverse populations in the United States.

REFERENCES

Agilent Technologies. (2016). Agilent DNA 1000 Kit Guide (G2938-90014 Rev. C).

Algee-Hewitt, B. B., Edge, M., Kim, J., Li, J., & Rosenberg, N. (2016). Individual identifiability predicts population identifiability in forensic microsatellite markers. *Current Biology, 26*(7), 935-942. https://doi.org/10.1016/j.cub.2016.01.065

Applied Biosystems. (2012). PrepFiler® and PrepFiler® BTA Forensic DNA Extraction Kits (Publication No. 4468426 Rev. B).

Applied Biosystems. (2017). Quantifiler™ HP and Trio DNA Quantification Kits User Guide (Publication No. 4485354 Rev. G).

Brookes, A. J. (1999). *The essence of SNPs*. Elsevier B.V. https://doi.org/10.1016/S0378-1119(99)00219-X

Brown, R. A., & Armelagos, G. J. (2001). Apportionment of racial diversity: A review. *Evolutionary Anthropology, 10*(1), 34-40. https://doi.org/10.1002/1520-6505(2001)10:1<34::AID-EVAN1011>3.0.CO;2-P

Budowle, Bruce & Daal, Angela. (2008). Forensically relevant SNP classes. *BioTechniques*. 44. 603-8, 610. 10.2144/000112806

Butler, J. M., Coble, M. D., & Vallone, P. M. (2007). STRs vs. SNPs: Thoughts on the future of forensic DNA testing. *Forensic Science, Medicine, and Pathology, 3*(3), 200-205. https://doi.org/10.1007/s12024-007-0018-1

Chakraborty R, Kamboh M, Ferrell R. (1991). "'Unique' alleles in admixed populations: a strategy for determining 'hereditary' population differences of disease frequencies." *Ethn Dis.* Summer;1(3):245-56. PMID: 1842537.

England R, Harbison S. (2019, June 17). A review of the method and validation of the MiSeq FGx™ Forensic Genomics Solution. WIREs Forensic Sci. 2019;e1351.https://doi.org /10.1002/wfs2.1351

Hargrove, T. (2019). Cold Case Homicide Stats. Retrieved from https://www.projectcoldcase.org/cold-case-homicide-stats/.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Jäger, A. C., Alvarez, M. L., Davis, C. P., Guzmán, E., Han, Y., Way, L., … Holt, C. L. (2017). Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories. Forensic Science International: Genetics, 28, 52-70. http://doi.org/10.1016/j.fsigen.2017.01.011

Kidd, J. R., Friedlaender, F. R., Speed, W. C., Pakstis, A. J., De La Vega, Francisco M, & Kidd, K. K. (2011). Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investigative Genetics, 2*(1), 1-1. https://doi.org/10.1186/2041-2223-2-1

Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F. R., & Kidd, J. R. (2014). Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Science International : Genetics, 10*, 23-32. https://doi.org/10.1016/j.fsigen.2014.01.002

Kittles, R. A., & Weiss, K. M. (2003). Race, Ancestry, and Genes: Implications for defining disease risk. *Annual Review of Genomics and Human Genetics, 4*(1), 33-67. https://doi.org/10.1146/annurev.genom.4.070802.110356

Koenig, B. A., Lee, S. S., & Richardson, S. S. (2008). *Revisiting race in a genomic age*. Rutgers University Press.

Lee, H., Lee, J. W., Jeong, S. J., & Park, M. (2017). How many single nucleotide polymorphisms (SNPs) are needed to replace short tandem repeats (STRs) in forensic applications? *International Journal of Legal Medicine, 131*(5), 1203-1210. https://doi.org/10.1007/s00414-017-1564-z

Li, Chengtao. "Forensic Genetics." *Forensic sciences research*, vol. 3,2 103-104. 18 Jul. 2018, doi:10.1080/20961790.2018.1489445

Lim, S., Youn, J.P., Moon, S.O. *et al.* Characterization of human short tandem repeats (STRs) for individual identification using the Ion Torrent. *BioChip J* **9,** 164–172, 2015. https://doi.org/10.1007/s13206-015-9210-7

Liu, F., van Duijn, K., Vingerling, J. R., Hofman, A., Uitterlinden, A. G., Janssens, A. C. J. W., & Kayser, M. (2009). Eye color and the prediction of complex phenotypes from genotypes. Current Biology, 19(5), R192-R193. doi:10.1016/j.cub.2009.01.027

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., & Lareu, M. V. (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in genetics, 4*, 98. https://doi.org/10.3389/fgene.2013.00098

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics (Austin), 155*(2), 945-959.

Rosenberg, N. A., Li, L. M., Ward, R., & Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics, 73*(6), 1402-1422. https://doi.org/10.1086/380416

Shriver, M.D., et al. "Ethnic-Affiliation Estimation by use of Population-Specific DNA Markers." *American Journal of Human Genetics*, vol. 60, no. 4, 1997, pp. 957-964.

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current protocols in molecular biology*, *122*(1), e59. https://doi.org/10.1002/cpmb.59

Stephens, J. C., et al. "Haplotype Variation and Linkage Disequilibrium in 313 Human Genes." *Science,* vol. 293, no 5529, pp. 489-493.

Tan, S. C., & Yiap, B. C. (2009). DNA, RNA, and protein extraction: the past and the present. *Journal of biomedicine & biotechnology*, *2009*, 574398. https://doi.org/10.1155/2009/574398

Thermo Fisher Scientific Inc. (2015). Qubit® dsDNA HS Assay Kits (Catalog nos. Q32851, Q32854).

Verogen. (2018a). ForenSeq™ DNA Signature Prep Reference Guide (Document No. VD2018005 Rev. A).

Verogen. (2018). ForenSeq™ Universal Analysis Software Guide (Document No. VD2018007 Rev. A).

Walsh, S., Lindenbergh, A., Zuniga, S. B., Sijen, T., de Knijff, P., Kayser, M., & Ballantyne, K. N. (2011). Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence. *Forensic Science International: Genetics*, 5(5), 464-471. doi:10.1016/j.fsigen.2010.09.008

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., … Kayser, M. (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Science International: Genetics, 7(1), 98–115.

Wang, David G., et al "Large-Scale Identification, Mapping and Genotyping of Single Nucleotide Polymorphisms in the Human Genome," *Science*, vol. 280, no. 5366, 1998, pp. 1077-1082

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics, 28(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

APPENDIX A: IRB MATERIALS

IRB Training Certificates

Figure A1. Human Research, Biomedical Research 1 – Basic Course IRB Training Certificate

Figure A2. Human Research, Social/Behavioral Research 1 – Basic Course IRB Training Certificate

Approval Email

From: Jamie Carson [mailto:no-reply@irbnet.org]
Date: Monday, October 26, 2020 11:15 AM
Subject: IRBNet Board Action

Please note that Western Carolina University IRB has taken the following action on IRBNet:

Project Title: [1653330-1] A Biosocial Study Exploring Self-reported vs. Predicted Ancestry using the Verogen ForenSeq™ Signature Prep Kit Principal Investigator: Frankie West, PhD

Submission Type: New Project
Date Submitted: September 2, 2020

Action: APPROVED
Effective Date: October 26, 2020
Review Type: Expedited Review

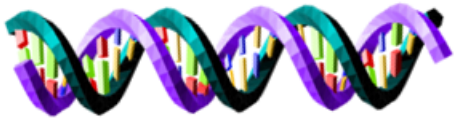
Should you have any questions you may contact Jamie Carson at jcarson@wcu.edu.

Thank you,
The IRBNet Support Team

www.irbnet.org

APPENDIX B: RECRUITMENT MATERIALS

Recruitment Flyer



Interested in Helping with DNA Research?

I am conducting a study about self-identification of race/ancestry and forensic genetics for my Master's thesis project.
This study will investigate the relationship between self-identity and genetic predictions of ancestry made by a commonly used forensic DNA kit.

How to participate:

- **Contact Brittania Bintz at bbintz@wcu.edu**
- Complete an online survey
- You will be asked to donate 2 cheek swabs for DNA analysis
- This study will require about 10 minutes of your time
- Your survey & genetic data will be anonymized

Eligibility:

- You must be a Western Carolina Student
- You must be between 18-24 years old

*There is no compensation for participating and you will not receive any genetic testing results. However, your participation may help contribute to a better understanding of forensic genetic ancestry estimation and biosocial identity.

Figure B1. Thesis project recruitment flyer

Consent Form

Western Carolina University Consent Form to Participate in a Research Study

Project Title: *A Biosocial Study Exploring Self-reported vs. Predicted Ancestry using the Verogen ForenSeq™ Signature Prep Kit*

This study is being conducted by: Xykiera Sims (Graduate Student, Department of Biology), Dr. Frankie West (Assistant Professor of Forensic Science, Department of Chemistry & Physics) & Brittania Bintz, MS (Research Scientist, Forensic Science Program)

Description and Purpose of the Research: You are invited to participate in a research study about genetics and ancestry because you are a Western Carolina University student between the ages of 18-24. Through your participation in this study we hope to learn more about how people identify using self-reported ancestry compared to forensic DNA ancestry predictions used in forensic genetics labs.

What you will be asked to do: You will be asked to take two buccal swab samples from the inside of your cheeks. This is done by taking the provided swab and brushing it against the inside of your cheek for 30 seconds and involves no pain or discomfort. Next, the swabs will be placed in the provided boxes, marked with anonymized numbers. In the lab, the DNA from these swabs will be used to generate genetic ancestry and hair and eye color predictions using the Verogen ForenSeq™ Signature Prep Kit. Your DNA will be destroyed after we complete this test and will not be used for any other purpose. The results from the DNA analysis cannot be linked to you through any identifying information.

Following collection of your buccal swabs, you will be asked to take an online Qualtrics survey. You will then enter the same number from your buccal swab box into Qualtrics survey and answer the questions in the online survey. This should take approximately 10 minutes. You will not receive any data from the Qualtrics survey or your genetic testing results. The total amount of time required is approximately 15 minutes.

In the event that Western Carolina University is closed due to the COVID-19 pandemic, participants will be mailed a sample collection kit with anonymous identifiers. Participants will also be provided the access link to the Qualtrics online self-identification survey via email. Email addresses and IP addresses will not be collected within the survey. All identifying material from the samples submitted by mail will be collected by the study staff, Brittania Bintz. All identifying material from the samples will be destroyed and not linked to DNA samples nor survey data.

The study team will not generate any clinically relevant information and will not return clinically relevant information to you.

Risks and Discomforts: We anticipate that your participation in this survey presents no greater risk than everyday use of the Internet. All of your answers will be anonymized and cannot be linked back to you, the participant.

If any of the questions asked as part of this study may make you feel uncomfortable, you may refuse to answer any of the questions, take a break or stop your participation in this study at any time.

The buccal swab collection process should present no physical discomfort. The genetic data produced from your buccal swab will be anonymized and cannot be linked back to you, the participant. All data will be stored on a secured server in the Forensic Genetics laboratory and can only be accessed by study personnel.

Benefits: There are no direct benefits to you for participating in this research study. Your participation will help contribute to a better understanding of how a diverse group of Generation Z individuals self-identify compared to forensic genetic ancestry estimations, which will contribute to our understanding of the utility of these tests in forensic contexts.

Privacy/Confidentiality/Data Security: The data collected in this study are anonymous. This means that not even the research team can match you to your data. All data collected over the course of this study will be anonymized and kept confidential, stored on password protected computers in locked facilities.

If you give the research team permission to quote you directly, the researchers will give you a pseudonym and will generalize your quote to remove any information that could be personally identifying. This study will inform the Master's thesis research of a graduate student in Biology, the results from which will be published in an academic peer-reviewed journal.

Biospecimens (buccal swabs) collected for this study will become property of Western Carolina University. You will not share in any commercial value or receive compensation if any commercial products are developed using the biospecimens. Your information and biospecimens will not be used or distributed for future research studies.

Voluntary Participation: Participation is voluntary and you have the right to withdraw your consent or discontinue participation at any time without penalty. If you choose not to participate or decide to withdraw, there will be no impact on your grades, academic standing, or student employment.

Compensation for Participation: There is no payment, extra credit, or direct compensation for participating in this project. However, your participation will help contribute to a better understanding of how a diverse group of Generation Z individuals self-identify compared to forensic genetic ancestry estimations, which will contribute to our understanding of the use of these tests in forensic contexts.

Contact Information: For questions about this study, please contact Brittania Bintz at bbintz@wcu.edu or (828)-227-3680.

If you have questions or concerns about your treatment as a participant in this study, you may contact the Western Carolina University Institutional Review Board through the Office of Research Administration by calling 828-227-7212 or emailing irb@wcu.edu. All reports or correspondence will be kept confidential to the extent possible.

You will be given a copy of this information to keep for your records.

I understand what is expected of me if I participate in this research study. I have been given the opportunity to ask questions, and understand that participation is voluntary. By making a mark on the line, I show that I agree to participate and am at least 18 years old and that I agree to have investigators quote me directly in research and publications.

Participant Acknowledgement: Check here: ____

Anonymous Participant ID: _____

Date: _____

I do ___ or do not ___ give my permission to the investigators to quote me directly in their research.

Participant Acknowledgment: Check here: ____

APPENDIX C: SELF-IDENTIFICATION QUALTRICS SURVEY

Survey Questions

C1. Directions: Please answer each question as accurately as possible. If you are unsure about a question, fill in the space provided with "N/A."

Please enter your unique ID#: *ID# can be found on the provided buccal swab box*
Free response

C2. Age
Free response

C3. I identify as a:
Multiple Choice:
Man
Woman
Other (open ended)
Prefer not to say

C4. Where were you born? (City, State, Country if not the US)
Free response

C5. Where did you grow up? (City, State, Country if not the US)
Free response

The next section will be used to focus more closely on how you identify in terms of race/ancestry/ethnicity.

C6. How do you self-identify in terms of race/ancestry/ethnicity? Please include all ways you identify, if more than one.
Free response

C7. I identify as (Choose all that apply):
Multiple Answer:
White or European
Black or African
Black or African American
Hispanic/Latinx/Spanish
Asian
Middle Eastern
Native Hawaiian or Other Pacific Islander
Native American/American Indian/Alaskan Native
Other (open ended)

C8. White or European

Multiple Answer:

German
Italian
Irish
Polish
English
French
Spanish
Other (open ended)
Unsure

C9. Hispanic/Latinx/Spanish

Multiple Answer:

Mexican
Salvadoran
Puerto Rican
Spanish
Dominican
Cuban
Colombian
Haitian
Other (open ended)
Unsure

C10. Black or African – Black or African American

Multiple Answer:

African American
Nigerian
Jamaican
Ethiopian
Haitian
Somali
Other (open ended)
Unsure

C11. Asian

Multiple Answer:

Chinese
Vietnamese
Filipino
Korean
Asian Indian
Thai
Japanese
Other (open ended)

Unsure

C12. Native Hawaiian or Other Pacific Islander
Multiple Answer:
Native Hawaiian
Tongan
Samoan
Fijian
Chamorro
Marshallese
Other (open ended)
Unsure

C13. Middle Eastern
Multiple Answer:
Lebanese
Syrian
Iranian
Moroccan
Egyptian
Israeli
Other (open ended)
Unsure

C14. Native American/American Indian/Alaskan Native – Please fill in tribal affiliation(s).
Free response

C15. Where was your mother born? (City, State, Country if not the US)
Free response

C16. Where was your father born? (City, State, Country if not the US)
Free response

C17. What is your eye color?
Free response

C18. What is your natural hair color?
Free response

C19. I am interested in my ancestry because? (Mark all that apply)
Multiple Answer:
I'm adopted and interest in more information
School project/course requirement
Social trend
It relates to my health
Other (open ended)

C20. My ancestry is relevant to me because I think it is also related to: (Mark all that apply)
Multiple Answer:
Health
Genetics
Physical features
Culture and traditions
My social identity
Other (open ended)

The following tables summarize the response data for the self-identification survey. Survey questions 8-14 (C8-C14) were not required to be answered for each participant. These questions are sub-questions to question 7 (C7) and would only be displayed on the self-identification survey if the participant chose the prior answer choice that corresponds to each sub-question.

Survey Responses

Table C1. Question 1 Responses (Please enter your unique ID#: *ID# can be found on the provided buccal swab box*)

Participant	Question 1 Response
1	FDE636
2	IBX426
3	ULF815
4	TSB269
5	FEI989
6	JUM775
7	MDZ743
8	DDE742
9	ELW787
10	CGL584
11	NLY688
12	XDW222

Table C2. Question 2 Responses (Age)

Unique ID#	Question 2 Response
FDE636	21
IBX426	19
ULF815	22
TSB269	24

FEI989	20
JUM775	19
MDZ743	21
DDE742	24
ELW787	22
CGL584	22
NLY688	19
XDW222	22

Table C3. Question 3 Responses (I identify as a:)

Unique ID#	**Question 3 Response**
FDE636	Woman
IBX426	Woman
ULF815	Man
TSB269	Man
FEI989	Woman
JUM775	Woman
MDZ743	Woman
DDE742	Man
ELW787	Man
CGL584	Woman
NLY688	Woman
XDW222	Woman

Table C4. Question 4 Responses (Where were you born? (City, State, Country if not the US))

Unique ID#	**Question 4 Response**
FDE636	Mount Airy, NC
IBX426	Durham, NC
ULF815	Matthews, NC
TSB269	Newport News, VA
FEI989	Stanley, NC
JUM775	Jesup, GA
MDZ743	Fort Bragg, NC
DDE742	Bayamon, Puerto Rico
ELW787	Murphy, GA
CGL584	Waynesville, NC
NLY688	Wilmington, NC
XDW222	Raleigh, NC

Table C5. Question 5 Responses (Where did you grow up? (City, State, Country if not the US))

Unique ID#	**Question 5 Response**

FDE636	Mount Airy, NC
IBX426	Raleigh, NC
ULF815	Monroe, NC
TSB269	Newport News, VA
FEI989	Stanley, NC
JUM775	Spartanburg, SC
MDZ743	Jonesboro, GA
DDE742	Thomasville, NC
ELW787	Bryson City, NC
CGL584	Waynesville, NC
NLY688	Wilmington, NC
XDW222	Raleigh, NC

Table C6. Question 6 Responses (How do you self-identify in terms of race/ancestry/ethnicity? Please include all ways you identify, if more than one.)

Unique ID#	**Question 6 Response**
FDE636	American Mexican
IBX426	European descent with German, Scottish and Irish being the most prevalent. It's been long discussed that we have Native American blood and can see it in some of my family members but that doesn't confirm
ULF815	White
TSB269	White, some Italian and Polish
FEI989	White/Caucasian
JUM775	White
MDZ743	White, half European
DDE742	Afro-Caribbean
ELW787	White with some Irish ancestry
CGL584	White or European
NLY688	Caucasian
XDW222	Black

Table C7. Question 7 Responses (I identify as (Choose all that apply):)

Unique ID#	**Question 7 Response**
FDE636	Hispanic/Latinx/Spanish
IBX426	White or European
ULF815	White or European
TSB269	White or European
FEI989	White or European
JUM775	White or European

MDZ743	White or European
DDE742	Black or African American, Hispanic/Latinx/Spanish
ELW787	White or European
CGL584	White or European
NLY688	White or European
XDW222	Black or African American

Table C8. Question 8 Responses (White or European)

Unique ID#	**Question 8 Response**
IBX426	German, Irish, Other (Scottish)
ULF815	German, Irish
TSB269	Italian, Irish, Polish
FEI989	German, Polish, Other (Dutch)
JUM775	Unsure
MDZ743	German, Italian, Irish
ELW787	Irish
CGL584	German, Italian, Irish, English
NLY688	Irish, English

Table C9. Question 9 Responses (Hispanic/Latinx/Spanish)

Unique ID#	**Question 9 Response**
FDE636	Mexican
DDE742	Puerto Rican
GLY239	Other - Guatemalan

Table C10. Question 10 Responses (Black or African – Black or African American)

Unique ID#	**Question 10 Response**
DDE742	Unsure
XDW222	African American

Table C11. Question 15 Responses (Where was your mother born? (City, State, Country if not the US))

Unique ID#	**Question 15 Response**
FDE636	San Fernando Tamaulipas, Mexico
IBX426	Durham, NC
ULF815	Jacksonville, FL
TSB269	Annapolis, MD

FEI989	Charlotte, NC
JUM775	Hope Valley, RI
MDZ743	Heidelberg, Germany
DDE742	Catano, Puerto Rico
ELW787	Murphy, GA
CGL584	Fort Lauderdale, FL
NLY688	Wilmington, NC
XDW222	Morganton, NC

Table C12. Question 16 Responses (Where was your father born? (City, State, Country if not the US)

Unique ID#	**Question 16 Response**
FDE636	Pancho Villa Tamaulipas, Mexico
IBX426	Scottsdale Arizona
ULF815	Halifax, Canada
TSB269	Baltimore, MD
FEI989	Hickory, NC
JUM775	Detroit, MI
MDZ743	Parkersburg, WV
DDE742	Carolina, Puerto Rico
ELW787	Bryson City, NC
CGL584	Fort Lauderdale, FL
NLY688	Wilmington, NC
XDW222	Mocksville, NC

Table C13. Question 17 Responses (What is your eye color?)

Unique ID#	**Question 17 Response**
FDE636	Brown
IBX426	Light blue
ULF815	Hazel
TSB269	Brown
FEI989	Hazel/Blue/Green
JUM775	Brown
MDZ743	Blue
DDE742	Dark Brown
ELW787	Hazel
CGL584	Green
NLY688	Brown
XDW222	Brown

Table C14. Question 18 Responses (What is your natural hair color?)

Unique ID#	Question 18 Response
FDE636	Black
IBX426	Blonde/dirty blonde
ULF815	Brown
TSB269	Brown
FEI989	Light Brown
JUM775	Brown
MDZ743	Brown
DDE742	Black
ELW787	Brown
CGL584	Light Brown
NLY688	Brown
XDW222	Black

Table C15. Question 19 Responses (I am interested in my ancestry because? (Mark all that apply))

Unique ID#	Question 19 Response
FDE636	"Curiosity. My mother believes she had Native American blood. My father believes he has Spaniard blood. Neither really know give their circumstances."
IBX426	It relates to my health, "I've just always been interested! There's a lot that goes into why, it's cool to know your roots and pay respects to your ancestors. Knowing health risks is also something I want to look out for."
ULF815	Always been curious
TSB269	It relates to my health
FEI989	School project/course requirement
JUM775	Other - To help
MDZ743	Social trend
DDE742	Social trend
ELW787	Other – Most people's ancestors live short, hard lives to get them here. The least I could do is know where they come from
CGL584	School project/course requirement, Social trend
NLY688	Other - I'm just curious and fascinated by genetics
XDW222	School project/course requirement, social trend, It relates to my health

Table C16. Question 20 Responses (My ancestry is relevant to me because I think it is also related to: (Mark all that apply))

Unique ID#	Question 20 Response
FDE636	Health, Genetics, Culture and traditions, My social identity
IBX426	Health, Genetics, Physical features, Culture and traditions, "Seeing where my family stems from and how certain influences during history affected me currently. Figuratively and literally."
ULF815	Genetics, Physical features
TSB269	Health
FEI989	Health, Genetics, Physical features, Culture and traditions, my social identity
JUM775	Health, Genetics, Physical Features
MDZ743	Genetics, Culture and traditions, My social identity
DDE742	Health, Genetics, Physical features
ELW787	Health, Genetics, Physical features
CGL584	Health, Genetics, Physical features
NLY688	Genetics, Physical features
XDW222	Health, Genetics, Physical features, Culture and traditions, My social identity

APPENDIX D: PROTOCOL

Forensic Manual Normalization method – non official method.

- Stop the protocol at page 17 at the end of the library purification step and do not proceed to normalisation.
- Quantify each library using the Qubit HS kit (you may want to do this in duplicate to ensure no pipetting errors). Use the qubit to calculate the ng/ul concentration.
- Calculate the nM using the following equation, where the average library size is 236bp for Primer set A and 268 for primer set B.

  (concentration in ng/µl)
  (660 mol * average library size) x 10^6 = concentration in nM
- Make 2nM (or your desired concentration) dilutions for each sample.
- Pool equal volumes (eg 5ul) for each sample into a single Eppendorf so you have a 2nM pool.
- Follow the MiSeq RUO guidelines for denaturing and diluting the libraries on page 5 using protocol A do not spike in phix.
- Make a final 600ul aliquot of library to 10pM (this may need to be adjusted but it's a good starting point)
- Prepare the HSC in the same way as specified on page 22 of the Forenseq signature prep guide. Spike in 2ul of diluted and denatured HSC to your 600ul of diluted and denatured library prepared in the steps above.
- Load 600ul into the reagent cartridge and proceed as normal with the run set up.

Figure D1. Screenshot of manual normalization protocol provided by Verogen's technical support team.