THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF VETERINARY AND BIOMEDICAL SCIENCES

VARIATION IN AMPLICONIC GENE COPY NUMBER AMONG HUMAN Y CHROMOSOMAL HAPLOGROUPS

DANLING YE
SPRING 2017

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Veterinary and Biomedical Sciences
with honors in Veterinary and Biomedical Sciences

Reviewed and approved* by the following:

Kateryna Makova
Professor of Biology
Thesis Supervisor

Robert Van Saun
Professor of Veterinary Science
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Between the two sex chromosomes, the male-specific Y chromosome remains understudied compared to the X chromosome because it is highly repetitive in nature and difficult to sequence. Nevertheless, studies focusing on the Y chromosome are important as they provide insights on sexually dimorphic traits, as well as on the evolutionary relationships among individuals and their ancestry along the paternal lineage. Approximately 10.2 megabases (Mb) of the Y chromosome is contained in the ampliconic region, which consists of nine multi-copy gene families implicated in male-specific reproductive processes (Bhowmick et al. 2007; Bachtrog 2013). This study analyzes the ampliconic gene copy number variation of sixty-one males representing seventeen different Y chromosomal haplogroups from all over the world. Overall, the results show that copy numbers of ampliconic genes vary among different haplogroups and even among different individuals in the same haplogroup. Higher levels of variation were observed for the *RBMY* and *TSPY* gene families, while lower for the *BPY2*, *CDY*, *DAZ*, *HSFY*, *PRY*, *XKRY*, and *VCY* gene families. These research results pave the way for studying potential associations between ampliconic gene copy number and sexually dimorphic traits in humans in future studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This project would not have been possible without the help of many individuals. Thank you, Dr. Kateryna Makova for letting me join your lab during my freshmen year and for your mentorship and help throughout the years. To Dr. Marta Tomaszkiewicz, thank you for your supervision, teaching me protocols, and answering my questions; I would not have been able to complete this project without your constant guidance.

To Dr. Van Saun: thank you for advising me, editing my thesis, and making sure I met the requirements to graduate as a Schreyer Scholar.

Thank you to the many collaborators who made this project possible: Dr. Mark Shriver (Department of Anthropology) and Corey Liebowitz and Brooke Mattern from his lab, Dr. David A. Puts (Department of Anthropology), and Dr. Michael DeGiorgio (Department of Biology). Also, thank you to Dr. Arslan Zaidi who guided the statistical analysis for this project.

To my friends and family: thank you for making me into the scientist I am today and for all your encouragement and support throughout the years.

Funding for this project was provided by the Seed Grant from the Center for Human Evolution and Diversity (CHED) by the Tobacco Settlement Grant.

Introduction

Why study the Y Chromosome and Especially the Human Y Chromosome?

Throughout recent years, the number of scientific studies focusing on genetics has increased. Sequencing technology and bioinformatics tools have improved, making it easier to analyze thousands of genes and to elucidate their functions. Currently, the genome sequences of over 260,000 organisms are publicly available on GenBank, and approximately 12% of the sequences available are of human origin (Benson et al. 2007). Despite the array of DNA sequence information available, the Y chromosome remains relatively understudied. The vast majority of mammalian genome projects choose to study females because of the diploid nature of their sex chromosomes (XX), as a result of which the X is easier to sequence and assemble compared to the Y chromosome, which is highly repetitive (Tomaszkiewicz et al. 2017). Studying the Y chromosome provides insights into biological differences between the two sexes and leads to a more complete understanding of evolutionary history that cannot be gained from genetic studies focusing on females alone (Skaletsky et al. 2003; Oven et al. 2013). Thus, Y chromosomal studies are essential to our understanding of human genetics. Several studies have shown that males and females have sex-specific differences in evolutionary patterns (Hammer et al. 2008). Dispersal patterns and mating patterns between the two sexes differ, resulting in sexually dimorphic patterns of gene flow and genetic drift (Hammer et al. 2008). In human populations, the mating system is moderately polygynous, in which fewer males mating with multiple females have offspring (higher variance in reproductive success), leading to a smaller

male effective population size in comparison to females (Hammer et al. 2008). Polygyny in the human population leads to higher relative levels of genetic variation on the X chromosome in comparison to autosomes, demonstrating the importance of considering sex-specific factors in understanding genetic patterns and evolution (Hammer et al. 2008).

Functions of the Y chromosome

In humans, the Y chromosome is important for sex determination and evaluating sex-specific disease risks. The Y chromosome contains genes that override female development, the biological default (Zhao 2016). Specifically, the sex determining region Y (*SRY*) gene produces a transcription factor that regulates the development of male reproductive organs (testis) and prevents the development of female reproductive organs (uterus and fallopian tubes) (Harley et al. 1992).

The male-specific region of the Y chromosome (MSY) has been implicated in processes such as skeletal growth, germ-cell tumorigenesis, and graft rejection (Skaletsky et al. 2003). The genes contained on the Y chromosome are essential for male reproduction; deletions in the MSY are the most common known cause of spermatogenic failure in humans (Skaletsky et al. 2003). For males with low sperm concentrations, the molecular diagnosis of Y-chromosomal microdeletions can aid in understanding the genetic basis of the infertility phenotype (Krausz et al. 2013). Knowledge about the function of specific Y-linked genes allows for more detailed genetic counseling aimed at determining the risk of passing on undesired traits, such as impaired spermatogenesis, to future generations (Krausz et al. 2013).

Studying the Y chromosome is important not only for informing male-specific diseases but also for providing insights on the differences in disease progression and risk between the two sexes. In the past decades, biologists oftentimes thought that the genetic dimorphism between females and males had limited functional significance because of X-inactivation in females and limited understanding about the expression of Y chromosomal genes (Skaletsky et al. 2003). However, recently obtained information demonstrating that MSY genes are expressed throughout the body in males and that some X chromosomal genes can be expressed from both of the X chromosomes in females (Carrel and Willard 2005) have led to a renewed interest in studying both sex chromosomes, with implications to understanding differences between the sexes in anatomy, behavior, and disease risk (Skaletsky et al. 2003).

For example, Turner syndrome, which is associated with a 45 X (XO) karyotype in approximately half of the females affected, impedes normal development (NIH 2017). Females with the disease often have a short stature and loss of ovarian function; as a result, most females with the syndrome are unable to conceive (NIH 2017). It is hypothesized that haploinsufficiency in genes found on both the X and Y is implicated in the phenotypic characteristics of the disease (Skaletsky et al. 2003). Turner syndrome suggests that there is a specific dosage requirement for sex-linked genes in humans; less than 1% of monosomy X fetuses survive to term (Bellot et al. 2014)

Additionally, sex hormone production plays a role in non-reproductive diseases, such as autoimmune diseases (Ortona et al. 2016). Hormones such as estrogen and prolactin are found in higher concentrations in women and are thought to stimulate the immune response, whereas progesterone and testosterone have an immunosuppressive effect, making men less prone to autoimmune diseases compared to women (Ortona et al. 2016). Sex-specific differences in

hormone secretion are influenced by the presence of the Y chromosome; thus, Y chromosome genetics has functional impacts in not only reproductive but also other diseases.

The Y chromosome also plays an important role in human evolution. Since it recombines only over two relatively short pseudoautosomal regions (PARs), Y chromosomal single nucleotide polymorphisms (SNPs) can be used to develop phylogenies that display important information about human population genetics and evolutionary relationships (Oven et al. 2013).

Structure of the Y chromosome

The human Y chromosome is 59 million base pairs (Mb) in length and comprises approximately 2% of the total DNA in male cells (Skaletsky et al. 2003). Recombination occurs along the length of the chromosome; Y-Y gene conversion allows recombination in the MSY and in the pseudoautosomal region, X-Y crossingover frequently occurs during meiosis (Skaletsky et al. 2003).

The MSY, which is approximately 95% of the chromosome's length, is unique to the Y chromosome (Skaletsky et al. 2003). It contains a mosaic of heterochromatic sequences and euchromatic sequences: X-transposed, X-degenerate, and ampliconic (Skaletsky et al. 2003). The heterochromatic block is approximately 40 Mb long (Bachtrog 2013). In total, the euchromatic region is approximately 23 Mb long; 8 Mb of the euchromatic sequence is found on the short arm of the Y chromosome (Yp) and 14.5 Mb on the long arm (Yq) (Skaletsky et al. 2003). It contains 78 protein-coding genes that can produce 27 different proteins (Skaletsky et al. 2003). The 3.4 Mb long X-transposed region is human-specific and contains two genes and has 99% identity to the X chromosome; it resulted from an X-to-Y transposition that occurred around 3-4 million

years ago (Skaletsky et al. 2003). The X-degenerate region contains 16 single-copy genes that have homologues on the X chromosome (Skaletsky et al. 2003). The ampliconic region is a highly repetitive sequence spanning 10.2 Mb; large regions of sequence pairs in the ampliconic region show greater than 99.9% identity due to frequent gene conversion (Bachtrog 2013).

In addition to the MSY, the Y chromosome contains pseudoautosomal regions (PARs), which are found on both sides of the MSY (Skaletsky et al. 2003). PAR1, which is located at the terminal region of the short arm (Yp) is 2.6 Mb long and PAR2, which is located at the tip of the long arm (Yq), is 0.32 kb long (Quintana-Murci and Fellous, 2001). During meiosis, the PARs, especially PAR1, exchange genetic material with the PAR of the X chromosome, thus genes found in the PARs are inherited like autosomal genes (Quintana-Murci and Fellous, 2001).

Y haplogroups

In 2002, the Y Chromosome Consortium published a haplogroup phylogenetic tree using 245 mutational events on the non-recombining region of the Y chromosome (NRY) and resulted in 153 haplogroups (The Y Chromosome Consortium 2002). In 2008, Karafet and colleagues published a revised Y chromosome phylogenetic tree using approximately 600 binary markers. Their tree used the rules established by the Y Chromosome Consortium and contained 311 haplogroups (Karafet et al. 2008).

Identifying Y chromosomal haplogroups is important for better understanding human male phenotypes. Different human haplogroups have different phenotypic characteristics and accurately determining the haplotype of individuals has applications in fields such as forensics and medical genetics (The Y Chromosome Consortium 2002). Certain haplotypes might be at a

higher risk for certain diseases than others. Many phenotypic characteristics are determined by single nucleotide polymorphisms (SNPs) in autosomes; likewise, Y haplogroups may be the cause of certain phenotypes, including disease phenotypes (Lu et al. 2016). Individuals in certain Y chromosomal haplogroups may have autosomal variants that influence phenotypic characteristics, such as spermatogenic impairment (Lu et al. 2016). Thus combining Y haplogroup information with information about SNPs in autosomes may prove to be a fruitful method in understanding genetic causes of spermatogenic impairment (Lu et al. 2016).

Research has shown that males belonging to some Y haplogroups, such as haplogroups K and Q1, are at a higher risk of being infertile (Lu et al. 2016). Other haplogroups may have a protective effect against spermatogenic impairment; members of Y-hg O3e* had a significant lower incidence of non-obstructive azoospermia (NOA) (Lu et al. 2016). Y-chromosomal haplogroups can interact with autosomal variants; the study found synergistic and antagonistic interactions between Y-chromosomal haplogroups and certain autosomal SNPs (Lu et al. 2016).

Ampliconic Genes

For our project, we focused on studying the ampliconic regions, which make up 10.2 Mb of the human MSY (Bachtrog 2013). The Y chromosome contains nine Y-specific ampliconic gene families: basic protein Y2 (*BPY2* or *BPY*), chromodomain Y (*CDY*), deleted in azoospermia (*DAZ*), heat-shock transcription factor Y (*HSFY*), PTP-BL related Y (*PRY*), RNA-binding motif Y (*RBMY*), testis-specific Y (*TSPY*), X Kell blood-related Y (*XKRY*), and variable charge (*VCY*) (Bhowmick et al. 2007).

All of the ampliconic gene families are contained within palindromes (P1, P2, P3, P4, P5, P8), and an inverted repeat (IR2), with the exception of *TSPY*, which is arrayed in tandem outside of the palindromes (Bhowmick et al. 2007). Approximately 5.7 Mb of the ampliconic region is located in palindromes that have greater than 99.9% identity between the arms (Skaletsky et al. 2003). The *TSPY* array is part of a 20.4 kb repeat unit that encodes *TSPY* on one strand and a transcription unit, *CYorf16*, on the opposite strand (Skaletsky et al. 2003). Additionally, there is a *TSPY* repeat unit located more distally on the short arm of the Y chromosome (Yp); the *TSPY* repeat unit has a 3% divergence from the consensus *TSPY* sequence (Skaletsky et al. 2003). The *TSPY* array is the largest tandem protein-coding gene array present in the human genome (Skaletsky et al. 2003).

Ampliconic genes evolved through a variety of molecular mechanisms. Through the course of evolutionary history, the Y chromosome acquired and amplified testis-expressed gene families, allowing this chromosome to become specialized for male reproduction (Bellot et al. 2014). The genes on the Y chromosome have evolved from their X homologues for male-specific roles in reproductive functions, such as spermatogenesis (Bellot et al. 2014). *VCY* and *RBMY* originated from the common ancestor located on the autosome that gave rise to the X and Y chromosomes (Skaletsky et al. 2003). *CDY* arose by retroposition of a processed messenger RNA (Skaletsky et al. 2003). *DAZ* originated via transposition and then amplification of the autosomal transcription unit, *DAZL*, and a series of autosomal transpositions and subsequent amplifications accounts for the presence of several ampliconic sequences (Skaletsky et al. 2003). Though the ampliconic genes arose from several different mechanisms, they have evolved similar patterns of tissue expression, as they are predominantly expressed in testis; thus, they

have male-specific functions due to convergent evolution (Skaletsky et al. 2003). The ampliconic genes *SRY*, *RBMY*, *TSPY*, and *HSFY* have diverged in function from their X homologues, *SOX3*, *RBMX*, *TSPX* and *HSFX* (Bellot et al. 2014).

The ampliconic genes encode proteins necessary for reproduction; for example, deletion of *DAZ* often results in spermatogenic failure and *HSFY* is involved in azoospermia and oligospermia (Bhowmick et al. 2007).

Seven of the nine families are implicated in spermatogenesis or sperm production, and the ampliconic gene families are expressed predominantly or exclusively in testes (Table 1; Bhowmick et al. 2007). Since the Y chromosome contains multiple copies of these ampliconic genes, it is hypothesized that throughout evolutionary history, ampliconic genes accumulated that could enhance male reproductive fitness (Bellot et al. 2014).

Table 1 . Y Chromosomal Ampliconic Gene Families. The ampliconic genes found on the Y chromosome play an important role in spermatogenesis; many of these genes have been implicated in subfertility or infertility.

Ampliconic Gene	Importance/Function (NCBI)
basic protein Y2 (*BPY2*)	-located in the nonrecombining portion of the Y chromosome -expressed specifically in testis -encoded protein interacts with ubiquitin protein ligase E3A and may be involved in male germ cell development and male infertility (https://www.ncbi.nlm.nih.gov/gene/9083)
chromodomain Y (*CDY*)	-gene encodes a protein containing a chromodomain and a histone acetyltransferase catalytic domain

	-chromodomain proteins are components of heterochromatin-like complexes and can act as gene repressors -this protein is localized to the nucleus of late spermatids where histone hyperacetylation takes place (https://www.ncbi.nlm.nih.gov/gene/9085)
deleted in azoospermia (*DAZ*)	-expressed in premeiotic germ cells, particularly in spermatogonia -encodes an RNA-binding protein important for spermatogenesis (https://www.ncbi.nlm.nih.gov/gene/1617)
heat-shock transcription factor Y (*HSFY*)	-encodes member of heat shock factor family; heat shock factors are transcriptional activators for heat shock proteins -the gene is proposed to be responsible for azoospermia as it is found in a region of the Y chromosome that is sometimes deleted in infertile males (https://www.ncbi.nlm.nih.gov/gene/86614)
PTP-BL related Y (*PRY*)	-on the nonrecombining portion of the Y chromosome -expressed specifically in testis (https://www.ncbi.nlm.nih.gov/gene/9081)
RNA-binding motif Y (*RBMY*)	-encodes protein implicated as a splicing regulator during spermatogenesis (https://www.ncbi.nlm.nih.gov/gene/5940)
testis-specific Y (*TSPY*)	-expressed specifically in testis; may be involved in spermatogenesis (https://www.ncbi.nlm.nih.gov/gene/7258)
X Kell blood-related Y (*XKRY*)	-expressed specifically in testis, encodes protein similar to the putative member transport protein XK (X-linked Kell blood

	group precursor) (https://www.ncbi.nlm.nih.gov/gene/9082)
variable charge (*VCY*)	-expressed exclusively in male germ cells -encodes small positively charged protein, possibly a nuclear protein (https://www.ncbi.nlm.nih.gov/gene/9084)

Hypotheses

The Y chromosome's male-specific lineage makes it useful when constructing phylogenies and in understanding selective pressures on males throughout evolutionary history. We predict that ampliconic gene copy number will be statistically different between different haplogroups and can be used as a method to elucidate the ancestry of individuals. We predict that the study will allow us to gain a greater understanding of expected copy number values for individuals from different haplogroups.

Additionally, Y chromosome haplogroups can also be correlated with phenotypic traits. We predict that some sexually dimorphic phenotypic and behavioral traits are impacted by ampliconic gene copy number. For example, anthropometric phenotypes such as voice acoustics, hand strength, height, weight, skin pigmentation, scalp hair width, and facial shape, as well as behavioral traits such as sociosexual orientation and physical aggression, are hypothesized to be linked, at least in part, to the Y chromosome.

In order to obtain a complete picture of human biology, understanding the genetics of sex differences in human anatomical and behavioral traits is important. Recent NIH announcements have emphasized the importance of understanding sex differences in preclinical studies (Clayton

and Collins 2014). We predict that genes affecting patterns and levels of sexual dimorphism will likely affect other sex-dependent traits, including disease risks.

Droplet Digital PCR: history, advantages, disadvantages, and current uses

Droplet Digital PCR (ddPCR) was chosen as the experimental method for the experiment for several reasons. It is a method of PCR that builds upon its predecessors, polymerase chain reaction and real-time quantitative PCR, which allows for effective detection of copy number variation.

Polymerase Chain Reaction (PCR)

Polymerase chain reaction has been used for many decades to amplify a target nucleic acid sequence in a sample or detect the presence of certain sequences, which is essential in many biological experiments and in clinical settings (Hindson et al. 2011).

For PCR to work, several components must be present. Primers, DNA polymerase, nucleotides, and DNA template, buffer, and water are present in the reaction mix [PCR (NCBI 2014)]. In PCR, primers are designed to complement the target sequence [PCR (NCBI 2014)]. DNA polymerase is used to synthesize new DNA strands that are complementary to the target sequence. Oftentimes, Taq DNA polymerase is used; however, there are other polymerases available, such as Pfu DNA polymerase [PCR (NCBI 2014)]. Nucleotides (dNTPs) are also added to the reaction and are used to synthesize new stands of the target sequence as the reaction

progresses [PCR (NCBI 2014)]. A buffer is used to provide optimal conditions for the activity of the DNA polymerase enzyme (Thermo Fisher Scientific 2017).

In PCR, a thermocycler heats the reaction until the DNA denatures (NHGRI 2017). Then, the reaction temperature is lowered to allow the primers to anneal complementary to the DNA segment of interest and in the extension step, the DNA polymerase builds new strands of DNA complementary to the template strands (NHGRI 2017). The cycle repeats approximately 30 to 40 times, and in each cycle, the amount of target DNA doubles; thus at the end of the reaction, billions of copies of the DNA segment of interest are present, which can be used for analysis (NHGRI 2017). After completion of the PCR reaction, gel electrophoresis is used to determine the presence of the DNA segment of interest (Hindson et al. 2011). Despite its utility, conventional PCR is limited in its ability to quantify PCR products [PCR (NCBI 2014)].

Real-time Quantitative PCR (qRT-PCR)

Real-Time PCR allowed for better quantification of PCR products. The method uses fluorescent probes to measure how much DNA is amplified after each PCR cycle (Hindson et al. 2011). The cycle threshold, a point in which the fluorescence crosses an intensity threshold, is used to determine the concentration of the target (Hindson et al. 2011; Pinheiro et al. 2012).

Though this method is useful for quantification of nucleic acids, it requires standard curves (Hindson et al. 2011; Pinheiro et al. 2012). Suboptimal amplification efficiency influences cycle threshold values, which can ultimately result in inaccurate quantification of the target (Hindson et al. 2011).

Digital PCR (dPCR)

In 1992, digital PCR was developed and allowed for improved quantification of DNA targets. In dPCR, DNA molecules are divided into many reactions; thus, target DNA is distributed across multiple replicates, with some sub-reactions having no template and others having multiple templates present (Hindson et al. 2011; Vossen and White 2016). Initially, separate tubes or wells were used for each sub-reaction (Vossen and White 2016). dPCR uses end-point measurements and Poisson statistics to quantify nucleic acids and does not need standard curves (Hindson et al. 2011; Pinheiro et al. 2012).

Droplet Digital Polymerase Reaction (ddPCR)

For our experiments, we used Droplet Digital PCR (ddPCR) to determine ampliconic gene copy number. ddPCR is a form of polymerase chain reaction that was developed relatively recently. ddPCR was chosen because of its accuracy and precision, especially for copy number variation research (Vossen and White 2016). ddPCR used water-emulsion oil technology to divide a sample into approximately 20,000 nanoliter-sized droplets [ddPCR(Bio-Rad 2017)].

Each droplet serves as an individual reaction and can contain zero to multiple DNA molecules due to the partitioning process (Vossen and White 2016). Droplets that contain a target DNA product can be detected by fluorescence after amplification (Vossen and White 2016). Each droplet then serves as an individual data point, providing a larger sample size for analysis than traditional methods [ddPCR(Bio-Rad 2017)].

The reagents used for ddPCR are similar to those used in traditional PCR: primers, DNA template, and water are needed for both. However, for our ddPCR reactions, EvaGreen supermix

and *HindIII* are added to the reaction mixture. EvaGreen supermix contains a non-specific double-stranded DNA binding dye, which enables the detection of amplified DNA targets [EvaGreen (Bio-Rad 2017)]. *HindIII* is a restriction enzyme, and in ddPCR, restriction digestion is useful because it separates tandem gene copies, reduces sample viscosity, and makes the DNA temples more accessible [Probes (Bio-Rad)].

ddPCR target detection

There are two methods used to detect ddPCR products. The first method uses dual-labeled hydrolysis probes (Taqman chemistry) (Vossen and White 2016). There are several advantages to this method; first, the probes are extremely specific to the target sequence (Geoffrey et al. 2013; Rebolledo-Jaramillo et al. 2014). For each target, researchers must design unique probes and primers; this specificity is useful for applications such as rare single nucleotide polymorphism detection (Geoffrey et al. 2013). Additionally, the dual nature of Taqman allows researchers to measure more than one target in a single reaction by using two probes with different spectral wavelengths; dyes such as FAM, VIC, and HEX can be used for this application (Vossen and White 2016). Since two different targets can be measured in the same reaction, an unknown and control can be run in the same reaction, minimizing pipetting error that could happen if the control and unknown were run separately (Vossen and White 2016). However, a major disadvantage of this system is the cost: designing unique probes and primers increases both the cost and complexity of the experiments, and for many experiments, the specificity afforded by Taqman does not justify the costs (Geoffrey et al. 2013).

A second method is to use a non-specific DNA-binding/intercalating dye, EvaGreen, which emits a fluorescent signal upon binding to double-stranded DNA (Geoffrey et al. 2013; Tomaszkiewicz et al. 2016). The fluorescent emission is directly proportional to the amount of DNA present (Geoffrey et al. 2013). Using a double-stranded DNA-binding system has several benefits; they only require the design and synthesis of new primers and available primers can oftentimes be used (Geoffrey et al. 2013; Vossen and White 2016). Additionally, these dyes can be incorporated into already optimized protocols and are cheaper than TaqMan based chemistry while offering comparable precision and dynamic range (Geoffrey et al. 2013)

Disadvantages to using double-stranded binding dyes include non-specificity. Since only one target can be measured in each reaction, it has the disadvantage of splitting up the control and unknown into separate reactions. However, for our experiments, we only needed one control (the *SRY* gene) for multiple unknowns (the ampliconic genes under investigation); thus the number of reactions needed did not increase too drastically using EvaGreen compared to using TaqMan (Vossen and White 2016).

Advantages

There are many advantages to using ddPCR as a method for DNA quantification. Firstly, the system is able to partition the sample into millions of droplets in a 96-well plate, effectively increasing the number of data points generated and the accuracy of the results [ddPCR(Bio-Rad 2017)]. The large number of partitions is especially useful in our project because we are measuring copy number variation [ddPCR(Bio-Rad 2017)]. ddPCR is a precise method that can detect small fold differences in copy number, for example, accurately determining the difference

between a copy number of two versus three [ddPCR(Bio-Rad 2017)]. The method allows absolute quantification of the target DNA copies without having to run a standard curve; for our project, this allows accurate measurements of the target DNA [ddPCR(Bio-Rad 2017)]. Additionally, the ddPCR method allows for easier quantification of the target, as neither calibration standards or a reference is needed, unlike other methods [ddPCR(Bio-Rad 2017)].

Disadvantages

Despite its many advantages, there are some disadvantages to using the ddPCR system. Firstly, the cost of acquiring the reagents and machines to run ddPCR assays is significant. Secondly, the quantification of nucleic acid depends on classifying droplets as either positive or negative. However, some droplets emit a fluorescence signal that cannot be clearly classified as either positive or negative (Jones et al. 2014). The "rain" droplets found between clearly positive droplets and clearly negative droplets could be positive droplets emitting a reduced fluorescent signal, or negative droplets with an increased background fluorescence (Jones et al. 2014). This assignment can be especially problematic when dealing with low copy numbers (Jones et al. 2014).

QuantaSoft (Bio-Rad program) automatically draws a threshold between positive and negative droplets (Jones et al. 2014). QuantaSoft determines positive or negative droplets either by manual or automatic gating (Dean et al. 2016). Manual gating has the disadvantages of subjectivity and non-reproducibility (Dean et al. 2016). For automatic gating, the algorithm for QuantaSoft's droplet characterization is not publically available (Jones et al. 2014; Dean et al. 2016). Results from the system can produce poor results, for example when using formalin-fixed

paraffin-embedded (FFPE) samples (Dean et al. 2016). For example, in an experiment that attempted to quantify wild-type (double-positive) and mutant (FAM-positive) alleles of the *BRAF* gene, the QuantaSoft software assigned all droplets with a high FAM signal to a single cluster, failing to differentiate between double-positive and FAM-positive droplets (Dean et al. 2016).

Current Uses and Applications

ddPCR is useful for a variety of genetics studies. It can be used to detect rare DNA targets, accurately determine copy number variation, and measure gene expression [ddPCR(Bio-Rad 2017)]. It can be extremely useful in studying disease processes; for example, ddPCR was used to study preferential allelic imbalance (PAI) in cancer and was an accurate measurement of tumor PAI (Smith et al. 2015). It was also useful in determining somatic copy number alternations at loci with a non-significant trend towards preferential selection of the risk allele (Smith et al. 2015). Recently this method was also used for copy number evaluation of ampliconic Y chromosome genes (Tomaszkiewicz et al. 2016) and for validating low-frequency heteroplasmy variants in mitochondrial DNA (Rebolledo-Jaramillo et al. 2014). In addition, a literature screening of studies using ddPCR for HIV quantification showed that ddPCR was more accurate and precise compared to qPCR, though the two methods had similar sensitivities (Trypsteen et al. 2016). Thus, ddPCR is a useful method in basic research and in studying a wide variety of pathological processes.

Materials and Methods

In this study, we elucidated the variation in ampliconic gene copy number in 61 human males representing the major Y-chromosomal haplogroups. The DNA samples extracted from saliva represented the following seventeen Y-chromosome haplogroups: C3, E1b1a, E1b1b1, E1b1b1a, G2, I2a1b, L1, O1, O2, O3, Q1, R1b1a2a1a2c (R1-1), R1b1a2a1a2b (R1-2), R1b1a2a1a1 (R1-3), R1a1a1 (R1-4), and T.

For each DNA sample of interest, we performed ddPCR copy number assay of the nine ampliconic gene of interest and *SRY*, a single-copy gene as a reference. Each sample was run in triplicate. We used DNA samples whose donors already had phenotypic information available from previous studies.

We used ddPCR to quantify copy number variation because of its precision and accuracy in absolute quantification. Unlike other methods used to determine copy number variation, such as qPCR, ddPCR does not require a standard curve (Bio-Rad 2017). Additionally, ddPCR does not require many replicates in order to accurately determine copy number (Bio-Rad 2017).

ddPCR protocol

Using Droplet Digital PCR (ddPCR) copy number assays, we analyzed the ampliconic gene copy number of five males for each of the seventeen haplogroups included in the study. DdPCR Master Mix was prepared for three reactions in a 8-strip PCR tube for each assay and was composed of primers, template DNA, Evagreen supermix, *HindIII*, and water (Bio-Rad 2017). Four 8-strip PCR tubes were used for each run, with a total of 32 different samples

analyzed per run. Each reaction was run in triplicates, so after all the reagents were combined, each of the reactions in the 8-strip PCR tube was divided into three 22 ul samples placed in a 96-well ddPCR plate.

Table 2. ddPCR reagents.

Component in final reaction	**vol (ul) per reaction**	**vol (ul) per 3.5 rxns**
water	5.8	20.3
Evagreen supermix (2x)	11.0	38.5
Primer F (2.2 uM)	1	3.5
Primer R (2.2 uM)	1	3.5
Diluted *HindIII* (2.2 u./ul)	1	3.5
Template DNA (5 ng/ul)	2.2	7.7
Total	22.0	77.0

Afterwards, samples are placed in the Droplet Generator and each sample is partitioned into 20,000 droplets in a volume of 20 ul prior to amplification (Bio-Rad 2017). Following the ddPCR workflow, we used PCR amplification after the droplet generation step. Afterwards, the samples were placed in the droplet reader machine. The fluorescence in each droplet was measured and a threshold was drawn. Droplets above the threshold were counted as positive, and those below were counted as negative. The concentrations of the ampliconic genes of interest were each divided by the concentration of the reference, *SRY*. Since *SRY* is a single-copy gene in a human male genome, this is an effective way to determine the copy number for the ampliconic genes of interest (Tomaszkiewicz et al. 2016).

Primer design

We used the same primers for this study as the ones designed for a previous study that also used ddPCR to evaluate Y-chromosome ampliconic gene copy number variation in humans (Tomaszkiewicz et al. 2016). The primers were designed with Primer3Plus (v2.3.6) using parameters recommended in the Droplet Digital PCR Applications Guide (Bio-Rad). They were human-specific primers and designed using the latest annotation of the human Y chromosome (GCF_000001405.26 GRCh38/hg38) (Tomaszkiewicz et al. 2016). The general parameter settings were set as follows: product size range between 60-150 bp; primer size between 15-30 nt with an optimum size of 22 nt; primer melting temperature (T_m) between 58°C–65°C with an optimum temperature of 62°C; primer GC content between 50%–60% with an optimum GC content of 55% (Tomaszkiewicz et al. 2016). The primers were specific for functional ampliconic genes, targeting functional ampliconic genes excluding those found in pseudogenes for all the ampliconic genes of interest except for *TSPY* (Tomaszkiewicz et al. 2016). For *TSPY*, a section of the functional ampliconic genes was present in some of the pseudogenes, so primers were designed to hit the smallest number of pseudogenes (Tomaszkiewicz et al. 2016).

ddPCR assay steps

The first step of ddPCR is droplet generation. The droplets are identical in size and volume; thus, each droplet serves as an independent reaction during the PCR amplification step that follows (Bio-Rad 2017).

The second step of the ddPCR protocol is PCR amplification of the droplets. A thermal cycler was used to perform PCR (Table 3). The PCR conditions were optimized for the primers

used in the study (Bio-Rad 2017). Previously, gradient PCR was performed using the primers targeting the ampliconic genes of interest to determine the optimal annealing temperature for each of the primer pairs. For all of the primer pairs, the optimal annealing temperature was 59°C with the exception of *DAZ, PRY,* and *VCY,* whose optimal annealing temperatures were 63°C respectively. Thus for each DNA sample included in the study, the reference gene *SRY*, and the ampliconic genes, *BPY, CDY, HSFY, TSPY,* and *XKRY* were amplified with an annealing temperature of 59°C in one plate. The reference gene *SRY* and ampliconic genes *DAZ, PRY,* and *VCY* were amplified with an annealing temperature of 63°C in one plate.

Table 3. PCR protocol.

Step Number	**Temperature (in Celsius)**	**Duration (minutes: seconds)**
1	95°C	5:00
2	95°C	0:30
3	59°C or 63°C	1:00
4	go to step 2, repeat 45 x	
5	4°C	5:00
6	90°C	5:00
7	4°C	infinite hold

After amplification, the plate is placed in the ddPCR Droplet Reader machine (Bio-Rad 2017). The machine analyzes each droplet individually for fluorescence intensity (Bio-Rad 2017). A threshold is drawn and used to determine the number of positive droplets, those above the threshold intensity, and negative droplets, or those below the threshold intensity (Bio-Rad 2017). The data follows a Poisson distribution that is used to determine the initial concentration of the sample of interest (Bio-Rad 2017).

Data Analysis

The data was analyzed using RStudio and Plink software.

Principle Component analysis was performed on the ampliconic gene copy number data and SNP data in RStudio to determine whether there were trends by major haplogroups. Analysis of Variance (ANOVA) tests were performed for each ampliconic gene in RStudio to determine whether ampliconic gene copy number differed significantly between haplogroups.

Results

Between-haplogroup and within-haplogroup variability in ampliconic gene copy numbers

A total of 61 samples from nine major haplogroups (five from each haplogroup will eventually be analyzed) were used in the initial analysis of ampliconic gene copy number data. The haplogroups came from different geographic locations (Table 4).

Table 4. Major Haplogroups and Major Geographic Locations.

Major Haplogroup Name	Major Geographic Location
C	Asia
E	Africa
G	Africa
I	Europe
L	Asia
O	Asia
Q	Asia
R	Europe
T	Eurasian

Trends in Copy Number Variation Data

For all of these samples, ampliconic gene copy number data were generated for each of the nine ampliconic gene families (Table 5).

Table 5. Average Ampliconic Gene Copy Number Per Haplogroup. The average ampliconic gene copy number per haplogroup and the ranges for each haplogroup are displayed**.**

Ampliconic gene families/ Y haplogroup	*BPY*	*CDY*	*DAZ*	*HSFY*	*PRY*	*RBMY*	*TSPY*	*VCY*	*XKRY*
North Atlantic R1b1a2a1a2c (N=3)	3 (3-4)	5 (4-5)	5 (4-5)	2 (2-3)	2 (2-3)	12 (11-13)	30 (22-34)	3 (3-3)	2 (2-2)
Italo-Gaulish R1b1a2a1a2b (N=4)	3 (2-4)	4 (3-4)	4 (4-5)	2 (2-2)	2 (2-2)	11 (10-13)	25 (22-28)	3 (2-3)	2 (2-2)
Proto-Germanic R1b1a2a1a1 (N=5)	3 (3-4)	4 (3-6)	5 (3-6)	2 (2-3)	2 (2-3)	12 (10-14)	29 (25-33)	2 (2-3)	2 (2-3)
West and Northeast European R1a1a1 (N=3)	3 (3-3)	4 (4-5)	4 (3-5)	2 (2-2)	2 (1-2)	14 (11-15)	27 (24-30)	2 (2-3)	2 (2-2)
Native American Q1 (N=3)	3 (1-3)	4 (3-4)	3 (2-4)	2 (2-2)	2 (2-2)	8 (8-8)	28 (26-29)	2 (2-3)	2 (2-2)
Eurasian T (N=3)	4 (2-6)	4 (3-6)	5 (4-6)	2 (2-2)	2 (2-2)	12 (9-16)	36 (33-38)	2 (2-3)	2 (2-2)
African E1b1a (N=4)	3 (2-4)	4 (3-5)	4 (3-5)	2 (2-2)	2 (2-3)	11 (9-12)	34 (32-36)	2 (2-3)	2 (2-2)
Mediterranean and North African E1b1b1a (N=4)	3 (2-4)	4 (3-6)	4 (2-5)	2 (2-3)	2 (2-3)	7 (5-9)	32 (26-27)	2 (2-3)	2 (2-2)
Siberia C3 (N=3)	2 (1-3)	3 (3-4)	3 (2-4)	2 (2-2)	2 (1-2)	11 (8-14)	29 (25-32)	2 (2-3)	2 (2-2)
Africa E1b1a1a1g1a (N=4)	3 (3-4)	4 (3-5)	4 (4-5)	2 (2-2)	2 (2-2)	9 (7-11)	35 (32-40)	2 (2-3)	2 (2-2)
Northeast Africa/Near East E1b1b1	3 (3-4)	4 (4-4)	4 (4-4)	2 (2-2)	2 (2-3)	9 (8-10)	30 (25-38)	2 (2-3)	2 (2-2)

(N=4)									
West Asia G2 (N=4)	3 (3-4)	4 (3-5)	5 (4-6)	2 (1-3)	2 (2-3)	9 (8-11)	33 (28-40)	3 (2-4)	2 (1-2)
Slavic countries I2a1b (N=5)	3 (2-4)	4 (3-5)	4 (4-5)	2 (2-2)	2 (2-2)	9 (8-10)	22 (16-26)	2 (2-3)	2 (2-2)
India, Southern Pakistan and Sri Lanka L1 (n=2)	4 (3-6)	5 (4-6)	6 (4-9)	2 (2-2)	2 (2-3)	10 (8-11)	31 (27-35)	2 (2-3)	2 (2-2)
East Asia O1 (n=2)	3 (2-3)	4 (3-4)	4 (4-4)	2 (2-2)	2 (2-2)	9 (8-11)	29 (28-30)	3 (3-3)	2 (2-2)
China O2 (n=5)	3 (2-4)	4 (3-5)	4 (2-5)	2 (2-3)	2 (2-3)	9 (8-11)	31 (27-35)	2 (2-3)	2 (2-2)
East and Southeast Asian O3 (n = 3)	5 (3-8)	4 (3-6)	4 (4-6)	2 (2-3)	2 (2-2)	11 (11-12)	36 (35-37)	2 (2-3)	2 (2-3)

The average copy number for *BPY* was either 3 or 4 for every haplogroup, with the exception of C3, where the average was 2 and O3, where the average was 5. For all the samples analyzed, male humans had between 1 and 8 copies of *BPY*.

The average copy number for *CDY* per haplogroup was either 4 or 5, with the exception of C3, where the average was 3. The copy number for *CDY* was between 3 and 6 for all the individuals included in the study.

The average copy number per haplogroup for *DAZ* was between 3 and 6. Individuals in the study had a *DAZ* copy number between 2 and 9.

HSFY average copy number per haplogroup was consistently 2; all individuals had a *HSPY* copy number between 1 and 3.

The average copy number for *PRY* for each haplogroup was 2, and individuals had *PRY* copy numbers between 1 and 3.

RBMY ampliconic gene copy number varied between individuals and haplogroups. Individuals had an *RMBY* copy number between 5 and 16, and average copy numbers per haplogroup ranged from 7 to 16. The lowest *RBMY* average copy number (CN) was found in E1b1b1a (CN = 7; range: 5-9), an African haplogroup and the highest in R1a1a1 (CN= 14; range: 11-15), an European haplogroup.

TSPY showed the greatest variation out of all the gene families analyzed in the study. Individuals had a *TSPY* copy number between 16 and 40. The average *TSPY* copy number per haplogroup was between 22 and 36. The lowest *TSPY* average copy number was found in I2a1b (CN = 22; range: 16-26), a European haplogroup, and the highest in haplogroup O3 (CN= 36; range: 35-37), an Asian haplogroup.

The average copy number for *VCY* was between 2 and 3 for each haplogroup, and individuals had a *VCY* copy number ranging from 2 to 3.

XKRY copy number ranged from 1 to 3, with an average copy number of 2 per haplogroup.

AVOVA analysis of ampliconic gene copy numbers

To determine whether ampliconic gene copy number is significantly different among haplogroups, analysis of variance (ANOVA) was performed individually for each gene family (Table 6). Haplogroups that are evolutionarily close to each other were grouped into a 'major haplogroup' category to increase the statistical power of the test. For example, individuals from the O1, O2, and O3 haplogroups were grouped into the 'O' major haplogroup category. Therefore, major haplogroups listed in Table 4 were compared. Ampliconic gene copy number (CN) was not significantly different among major haplogroups for the genes *BPY, CDY, HSFY, XKRY, PRY* and *VCY* (Table 6). However, there was significant variation in ampliconic gene copy number for the genes *RBMY* ($P < 0.001$), *TSPY* ($P < 0.001$), and *DAZ* ($P < 0.01$) among major haplogroups. The distribution of ampliconic gene copy numbers per gene family is shown in Figures 1-9.

Table 6. Analysis of Variance of ampliconic gene copy number data. ANOVA was performed to determine which ampliconic gene copy numbers vary significantly between major haplogroups. Major_haplo values = variation between haplogroups; residuals =variation within haplogroups; Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Gene		Degrees of Freedom	Sum of Squares	Mean of Squares	F value	Pr (>F)	Significance
BPY	major_haplo	8	13.828	1.7286	1.6058	0.1459	
BPY	Residuals	52	55.974	1.0764			
CDY	major_haplo	8	5.846	0.73077	1.1412	0.3523	
CDY	Residuals	52	33.299	0.64037			
HSPY	major_haplo	8	0.1339	0.016736	0.2038	0.989	
HSPY	Residuals	52	4.2712	0.082139			
RBMY	major_haplo	8	124.07	15.5089	5.156	9.28E-05	***
RBMY	Residuals	52	156.41	3.0079			
TSPY	major_haplo	8	690.02	86.253	5.0408	0.000116	***
TSPY	Residuals	52	889.77	17.111			
XKRY	major_haplo	8	0.2226	0.02782	0.3955	0.9181	
XKRY	Residuals	52	3.6581	0.070348			
DAZ	major_haplo	8	24.91	3.11381	3.1643	0.005385	**
DAZ	Residuals	52	51.17	0.98405			
PRY	major_haplo	8	0.9658	0.12073	1.0983	0.3795	
PRY	Residuals	52	5.7158	0.10992			
VCY	major_haplo	8	0.6848	0.085603	0.3803	0.9263	
VCY	Residuals	52	11.7039	0.225076			

BPY

As seen from the graph (Figure 1), the medium *BPY* copy number was around 3 with the exception of haplogroups C, which had a lower medium copy number, and haplogroup L, which had a higher medium copy number. Both of these haplogroups are Asian haplogroups.

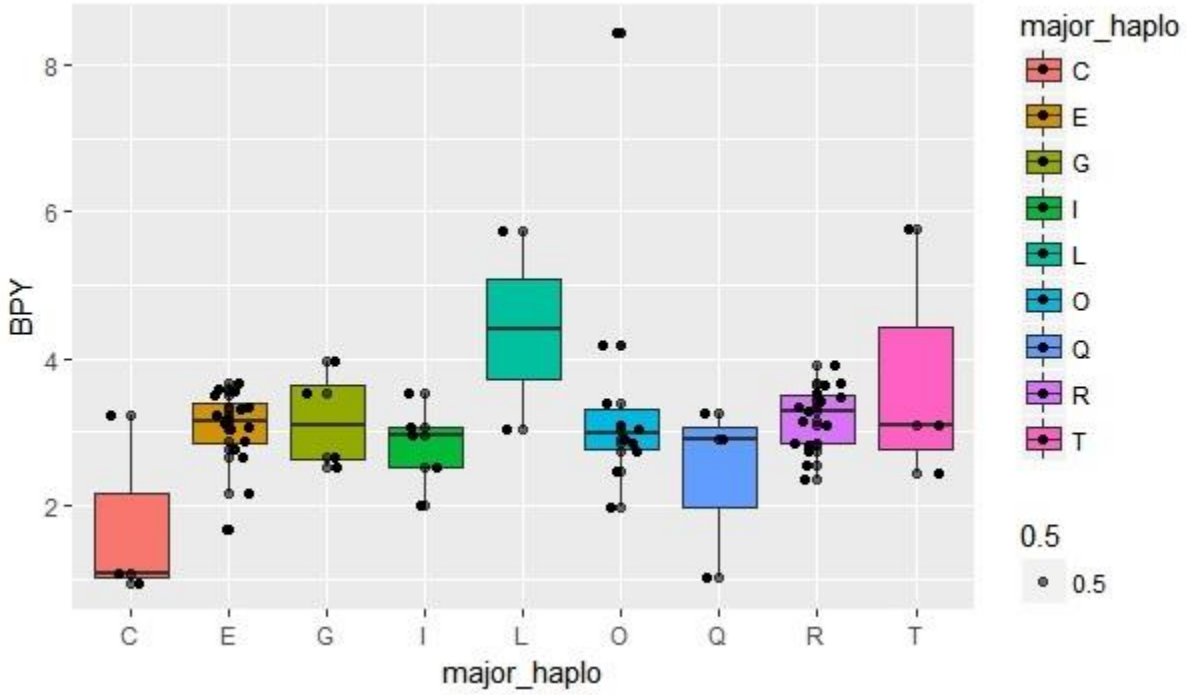

Figure 1. Ampliconic gene copy number for *BPY* based on major haplogroup (graph generated in RStudio).

CDY

There was some variation in CDY ampliconic gene copy number between major haplogroups, Haplogroup C, an Asian haplogroup, has the smallest median ampliconic gene copy number and haplogroup L, which is also an Asian haplogroup, has the largest median ampliconic gene copy number (Figure 2).

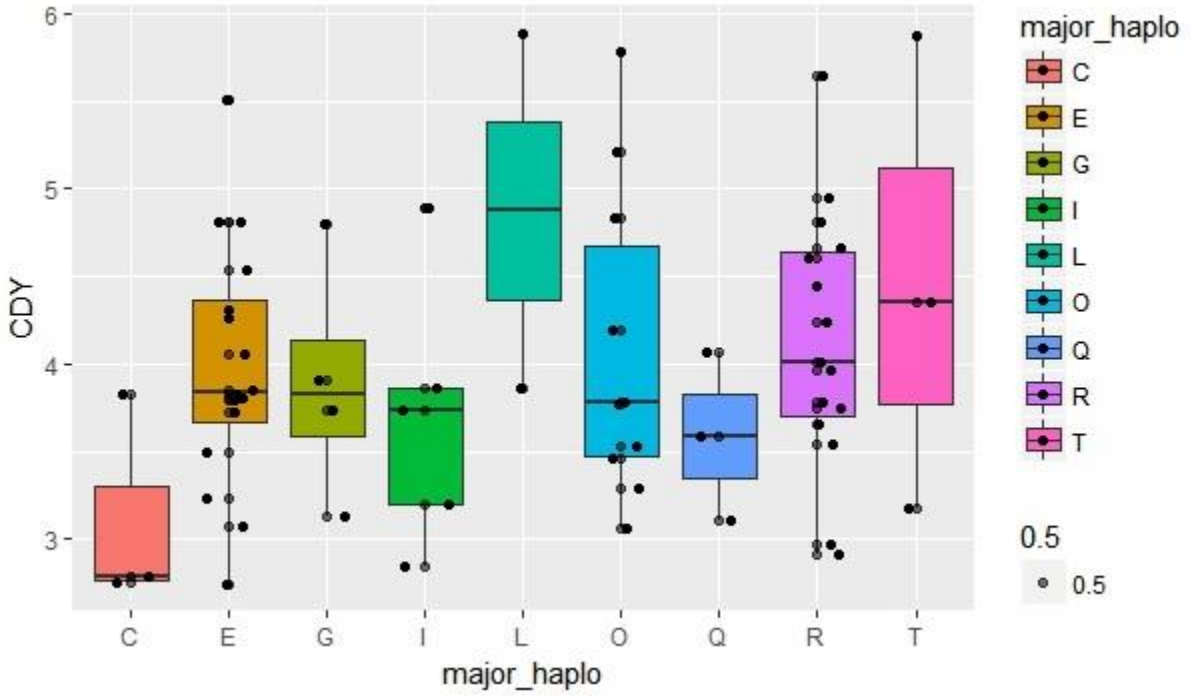

Figure 2. Ampliconic gene copy number for *CDY* based on major haplogroup (graph generated in RStudio).

DAZ

Lower DAZ ampliconic gene copy numbers were found in haplogroup C, an Asian haplogroup, whereas highest copy numbers were found in L, which is also an Asian haplogroup (Figure 3).

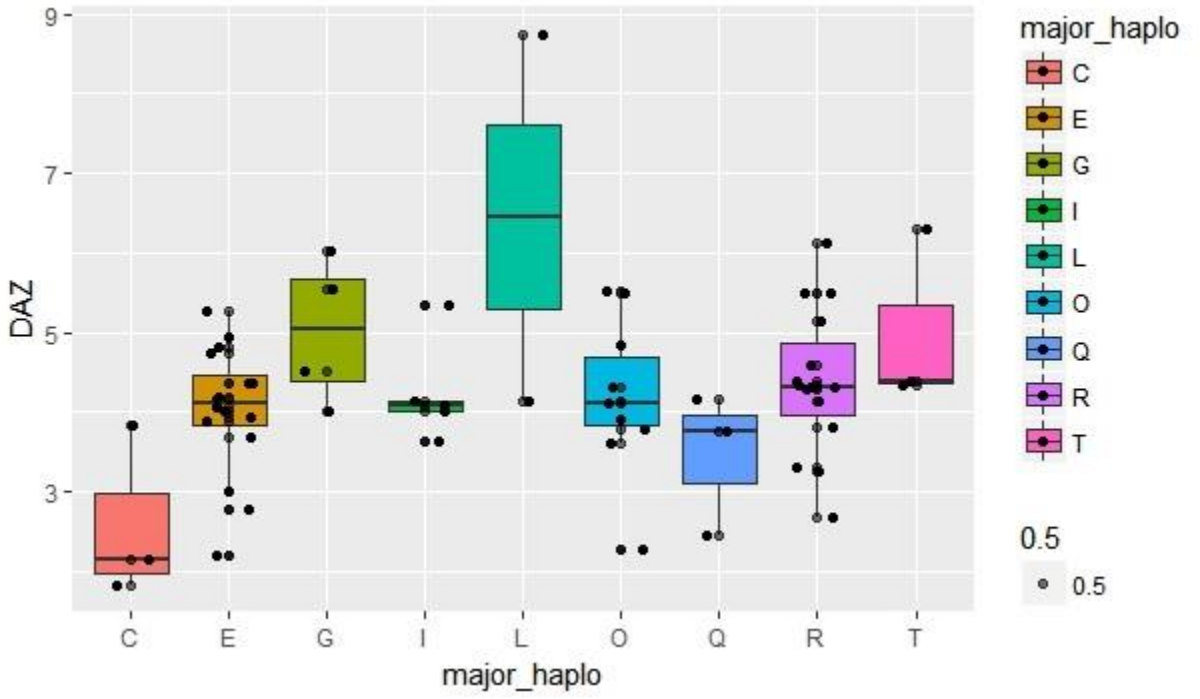

Figure 3. Ampliconic gene copy number for DAZ based on major haplogroup (graph generated in RStudio).

HSFY

HSPY copy number does not vary much between different individuals or haplogroups; the range of *HSPY* copy numbers is 1-3 with an average of 2 for each haplogroup (Figure 4).

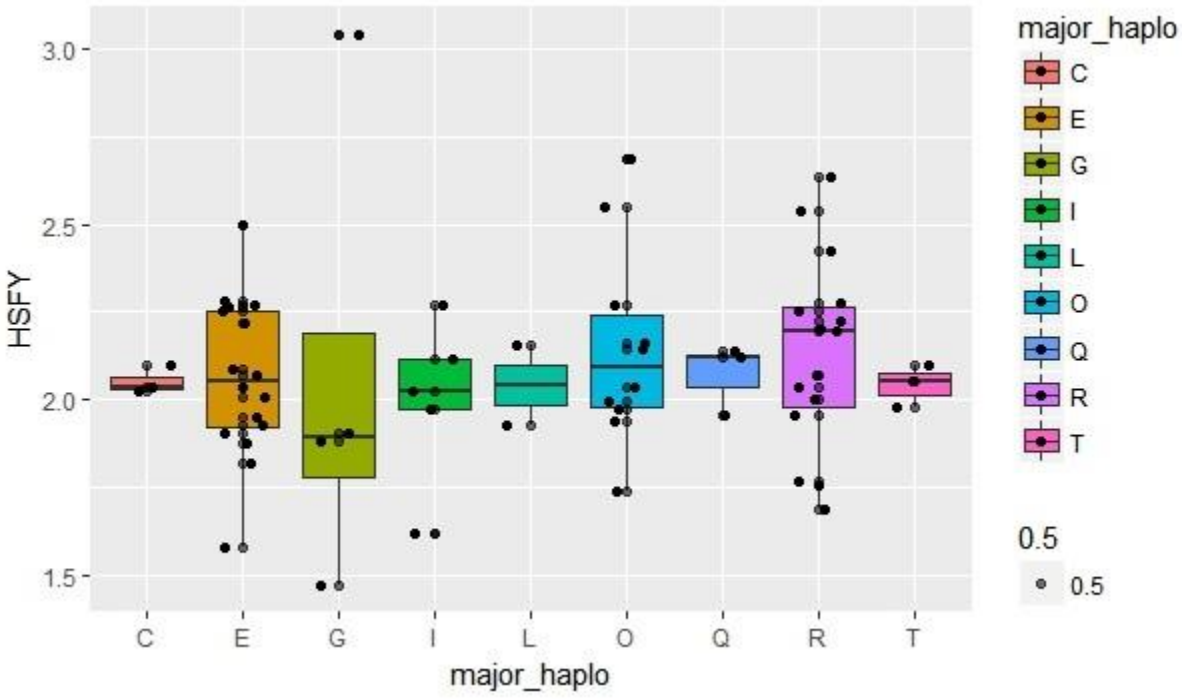

Figure 4. Ampliconic gene copy number for *HSFY* based on major haplogroup (graph generated in RStudio).

PRY

The median copy number for *PRY* for each major haplogroup was 2, and individuals had *PRY* copy numbers between 1-3 (Figure 5).

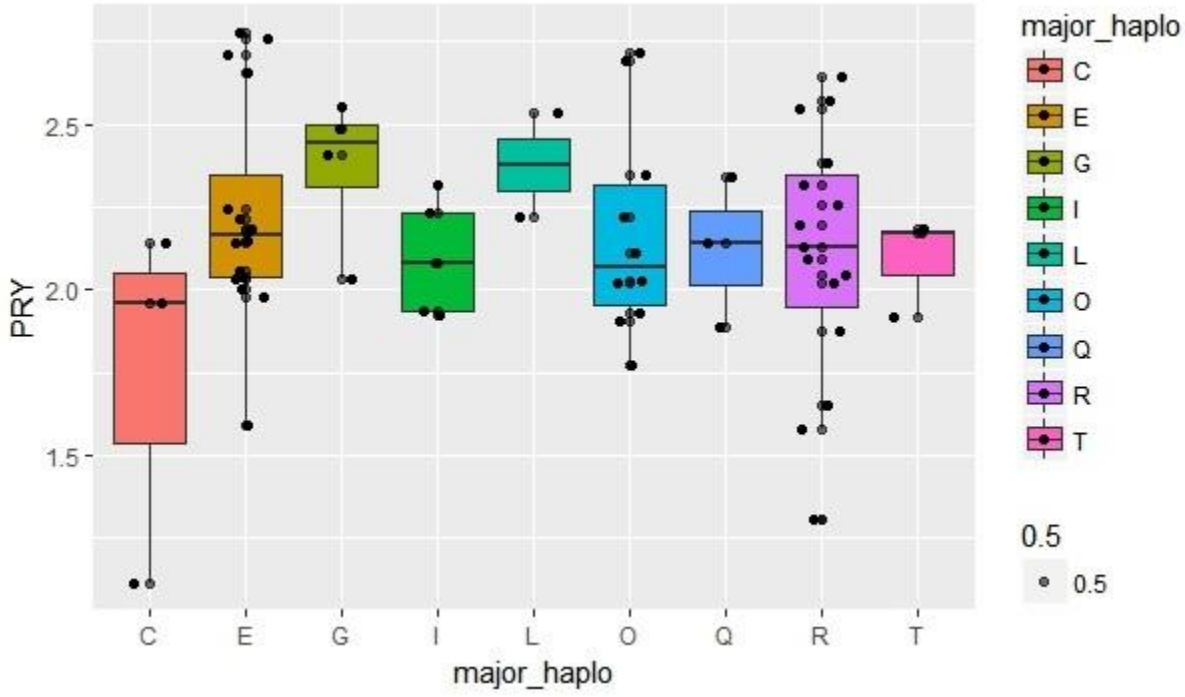

Figure 5. Ampliconic gene copy number for *PRY* based on major haplogroup (graph generated in Rstudio).

RBMY

There was variation in *RBMY* copy number between different haplogroups. The highest RBMY copy numbers were found in R, a European haplogroup, while lower copy numbers were found in Q, an Asian haplogroup (Figure 6).

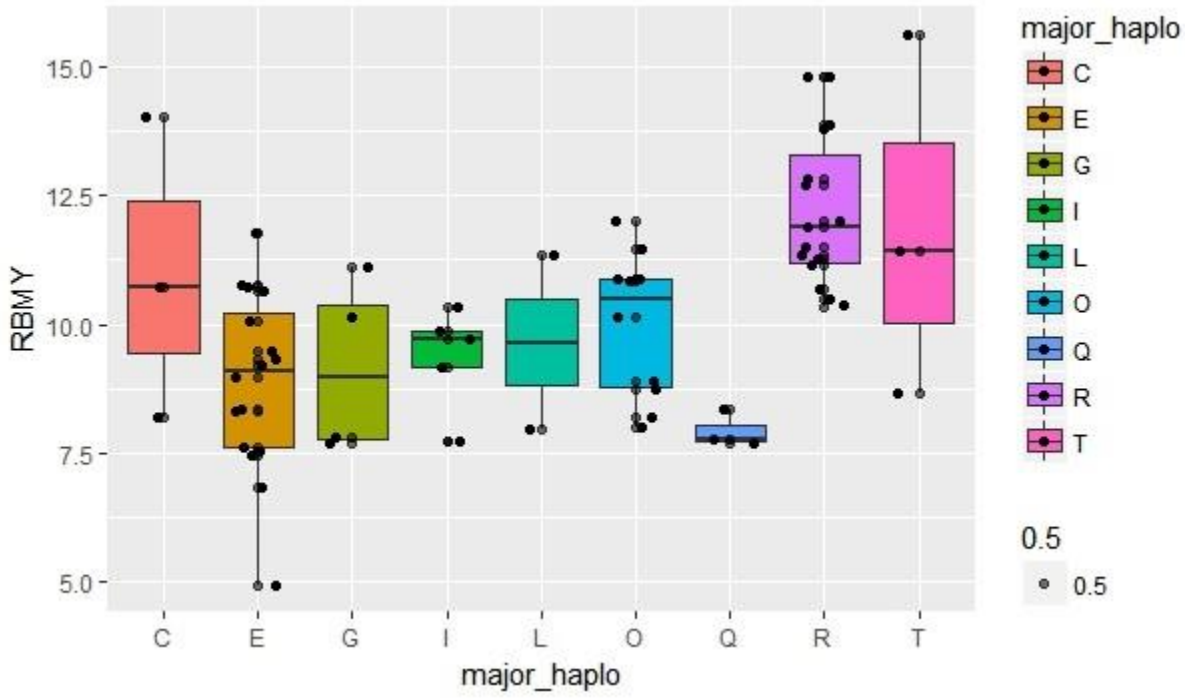

Figure 6. Ampliconic gene copy number for *RBMY* based on major haplogroup (graph generated in RStudio).

TSPY

Haplogroup I and E, which are the European haplogroups, had lower *TSPY* copy numbers than other haplogroups, such as the African haplogroups E and G and haplogroup T, a Eurasian haplogroup (Figure 7).

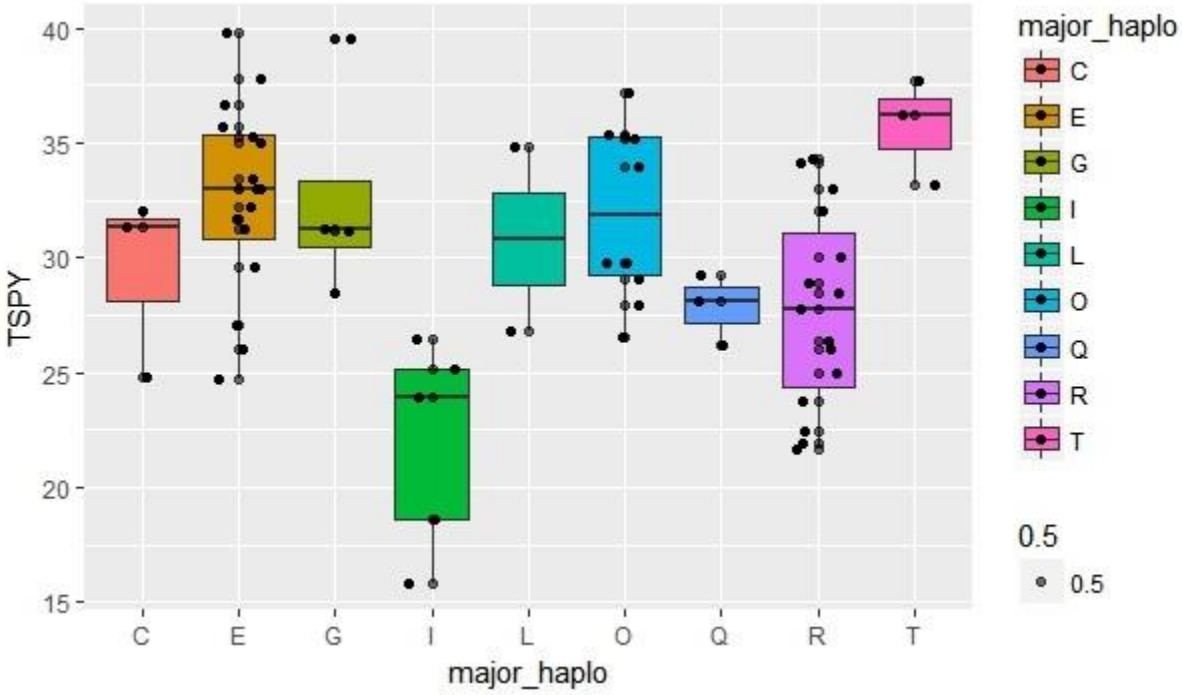

Figure 7. Ampliconic gene copy number for *TSPY* based on major haplogroup (graph generated in RStudio).

VCY

VCY median ampliconic gene copy numbers were relatively consistent among different major haplogroups (Figure 8). An individual from haplogroup G, an African haplogroup, had unusually high copy number.

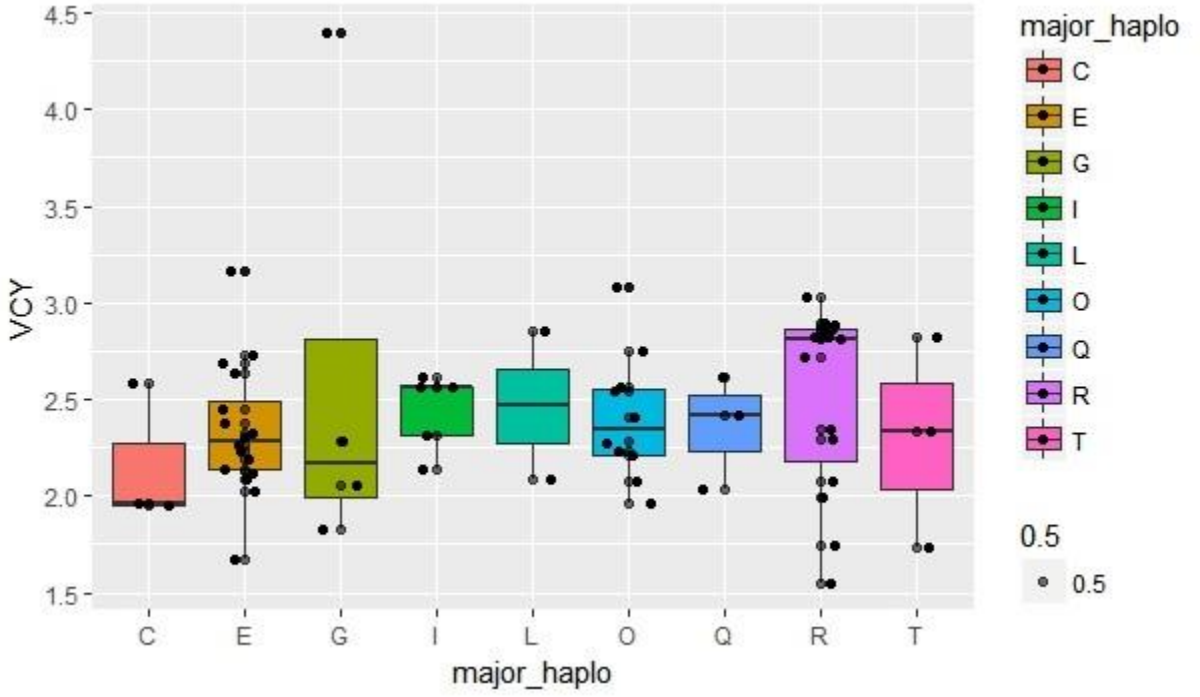

Figure 8. Ampliconic gene copy number for *VCY* based on major haplogroup (graph generated in RStudio).

XKRY

There was little variation in *XKRY* ampliconic gene copy number among the major haplogroups (Figure 9).

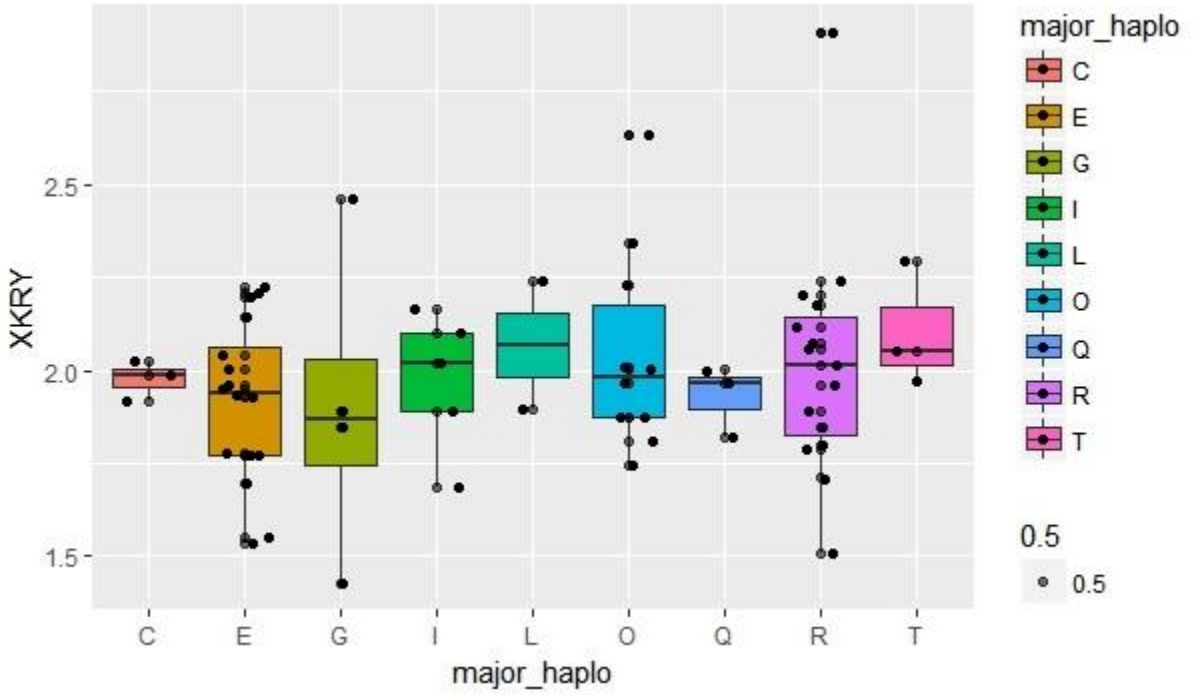

Figure 9. Ampliconic gene copy number for *XKRY* based on major haplogroup (graph generated in RStudio).

Investigating correlations among gene copy numbers using PCA

Investigation as to whether the variation in ampliconic gene copy number is able to separate major haplogroups from each other was undertaken. To test whether this is the case, principal components analysis (PCA) on the copy numbers of all genes was performed. Together, the first and second PCs explain more than 90% of the variation in gene copy numbers (Figures 10 and 11).

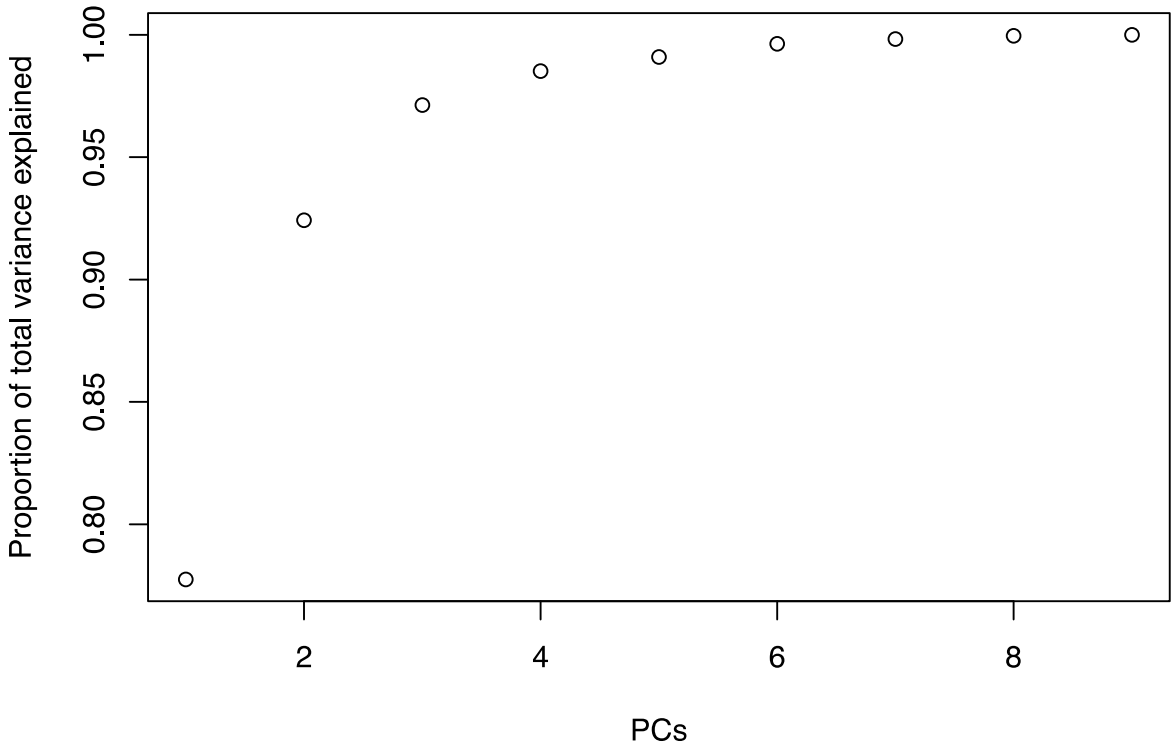

Figure 10. Proportion of total ampliconic gene variance explained by each PC (graph generated in RStudio).

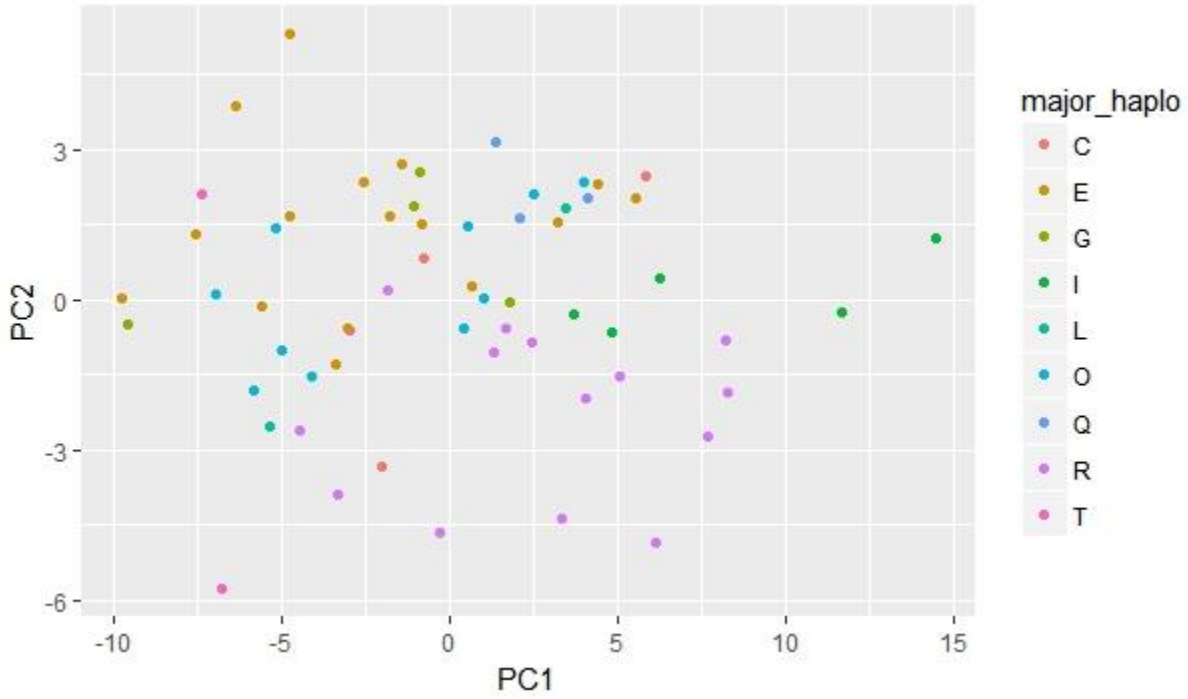

Figure 11. Principle Component Analysis of ampliconic gene copy number. Clustering by haplogroups can be observed for principle component 1 (PC1) and principle component 2 (PC2). On PC1, haplogroup I, which is most prevalent in Europeans, seems to cluster away from the rest. On PC2, Haplogroup R, which is also commonly found in Europeans cluster away from the rest (graph generated in RStudio).

The principle component analysis shows that there are trends in ampliconic gene copy number data that correspond to the major haplogroups of interest. The first two principle components, PC1 and PC2, separate copy number variation based on major haplogroup.

Median ampliconic gene copy number and variance

Analysis to determine if the median ampliconic gene copy number for each ampliconic gene correlated with the variance in copy number for that gene in the samples included in the study was performed (Figures 12 and 13).

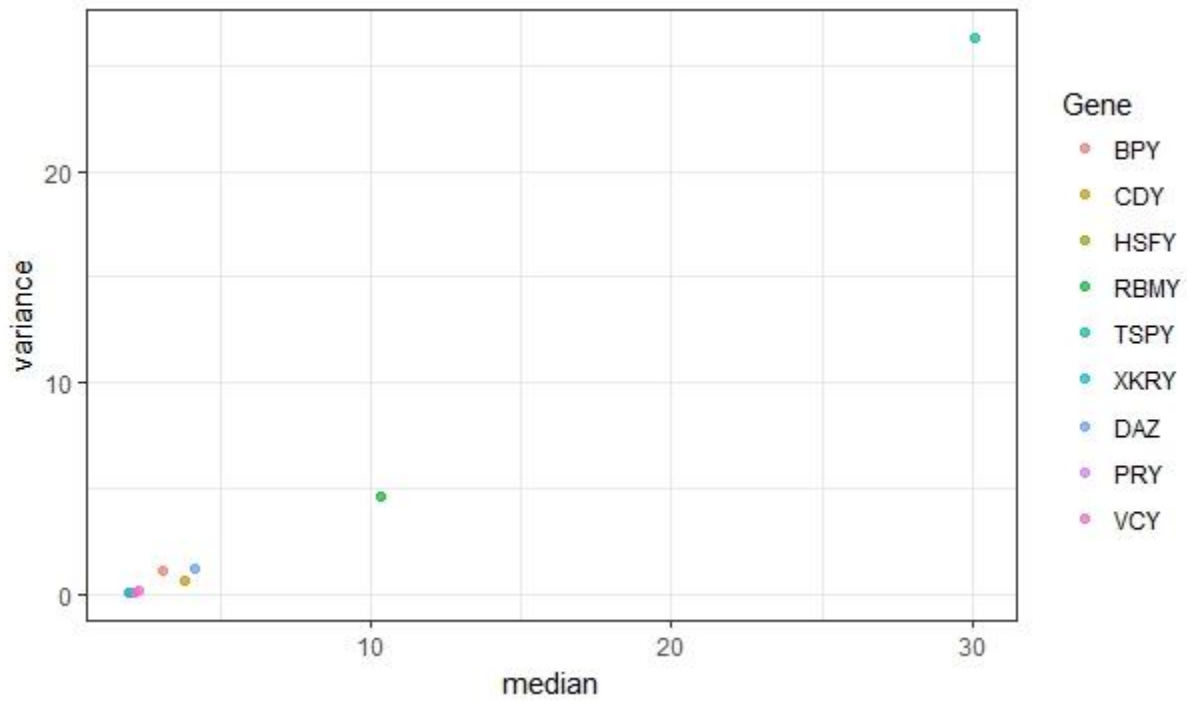

Figure 12. Median ampliconic gene number and variance in ampliconic gene copy number for samples included in the study (graph generated in RStudio).

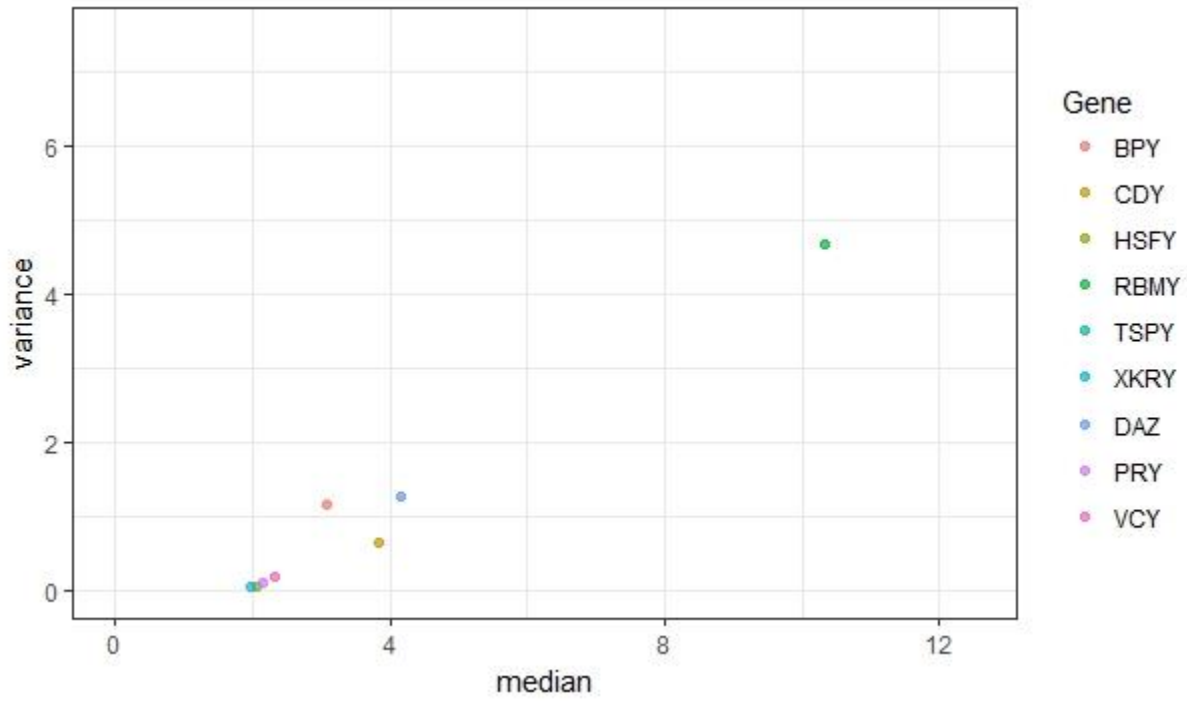

Figure 13. Median ampliconic gene number and variance in ampliconic gene copy number for all ampliconic genes except *TSPY* for samples included in the study (graph generated in RStudio).

There is a positive correlation between median ampliconic gene copy number and variance (Figures 12 and 13). *TSPY* had the highest ampliconic gene CN median (median = 30) and the highest variance in CN (variance = 26.3) among individuals included in the study, followed by *RBMY* (median = 10; variance = 4.7) (Figure 12). *DAZ* also followed the trend (median = 4, variance = 1.3). *CDY* also had a median CN of four, but a lower variance than *DAZ* (variance = 0.65). *BPY* had a median CN of three (variance = 0.65). *VCY*, *PRY*, *HSFY*, and *XKRY* all had a median CN of two and the lowest variance between individuals (variance =0.2, 0.11, 0.07, 0.06 respectively). Thus, ampliconic genes with higher CN values tended to vary more among individuals than ampliconic genes with lower CN values.

Comparing structure in ampliconic gene copy numbers with SNP genotypes

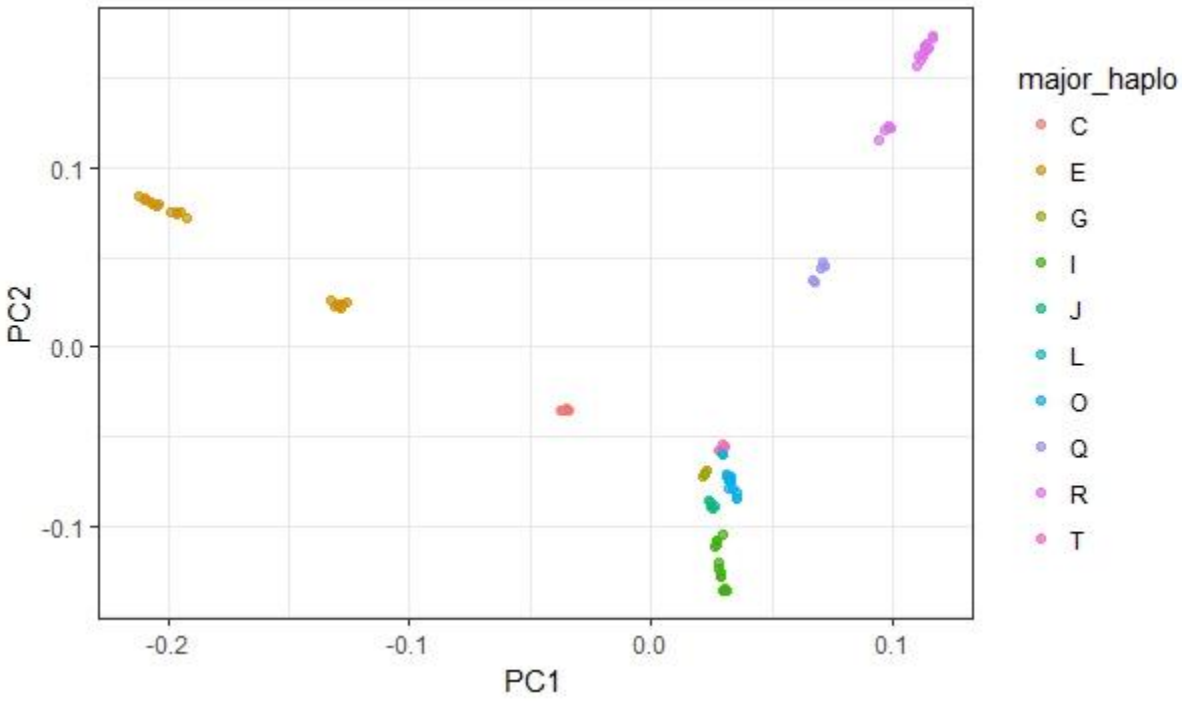

Figure 14. Principle Component Analysis of SNP data (PC1, PC2). As expected, the major haplogroups cluster according to PC1 and PC2 of the SNP 23andme data. Y-haplogroups are classified according to SNP data, and the principle component analysis reflects the association between the haplogroups and SNP data (graph generated in RStudio).

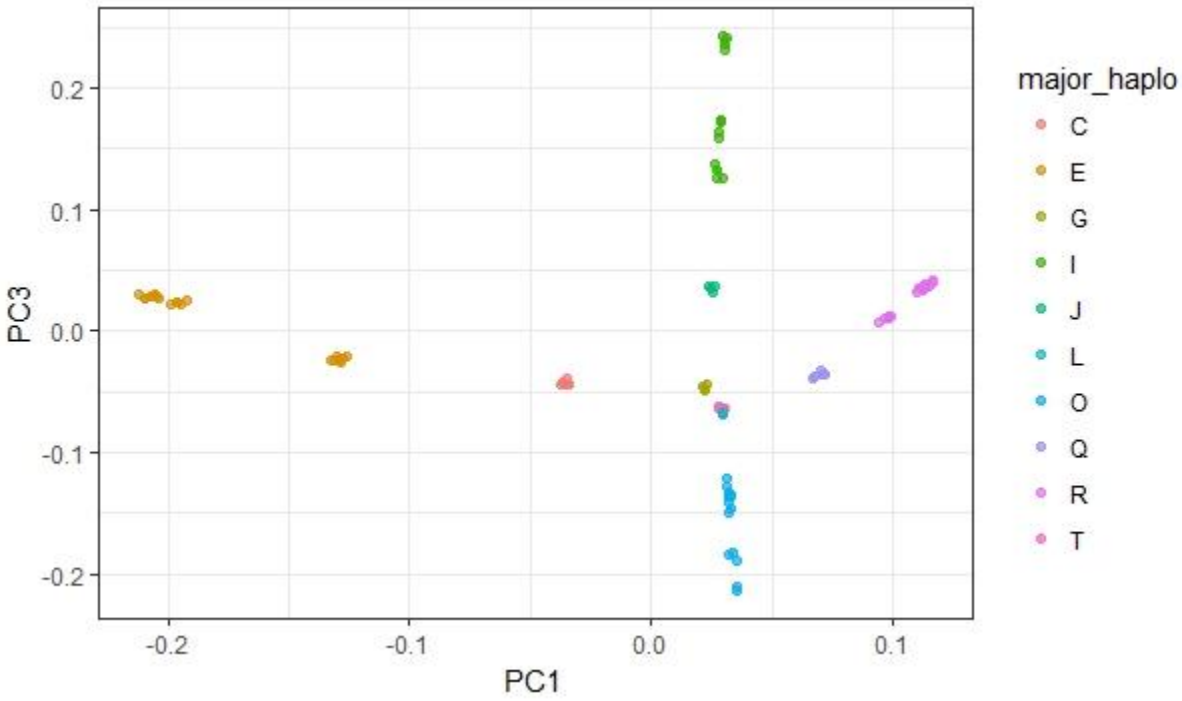

Figure 15. Principle Component Analysis of SNP data (PC1, PC3). As expected, the major haplogroups cluster according to PC1 and PC3 of the SNP 23andme data. Y-haplogroups are classified according to SNP data, and the principle component analysis reflects the association between the haplogroups and SNP data (graph generated in RStudio).

The SNP data plots (Figure 14 and 15), generated from the complete dataset of 100 males, show clusters by major haplogroups when looking at the principle components (PC1-PC3), indicating that individuals in the same haplogroup have similar values for the principle components (PC1-PC3) under investigation. This is to be expected, as SNPs are used to determine haplogroups. The separation of clustering for SNP data is more pronounced than separation of ampliconic gene copy number data for each haplogroup.

Discussion

The analysis of ampliconic gene copy numbers in 61 human males, representatives of nine major Y haplogroups, provides a new insight in the evolution of fertility genes across different human populations. Generally moderately polygynous (Hammer et al., 2008) human populations can vary in mating patterns and this might impact the evolution of the Y chromosome and its genes. The sex chromosomes evolved from an ordinary pair of autosomes, with the Y chromosome shrinking in size over time (Lahn et al. 2001). Since the Y chromosome recombines minimally with the X chromosome and is inherited differently than autosomes, its evolutionary pattern is different from that of autosomes. Unlike the autosomes, which are diploid and biparentally inherited, the Y chromosome is haploid and paternally inherited, making it more strongly subjected to the effects of genetic drift (Ghenu et al. 2016). Additionally, the neutral effective population size for the MSY is 25% that of autosomes, which again increases the role that genetic drift plays in Y chromosome evolution, as a smaller population size increases the likelihood of changes in allele frequency in subsequent generations (Ghenu et al. 2016). The ANOVA analysis of ampliconic gene copy numbers among different haplogroups showed distinct patterns for different haplogroups. This indicates that there were forces in effect throughout evolutionary history of the Y chromosome and its genes that have impacted ampliconic gene copy number in such a way that the copy number differs in human populations from different haplogroups. Selection or genetic drift probably played a role in the evolutionary history of ampliconic genes, and specific mechanisms impacting their evolution should be investigated further. Mating patterns and sperm competition may also influence evolution of ampliconic fertility genes; in mating systems with more sperm competition, it is likely that there

is more of a selective pressure to increase the copy number of certain ampliconic genes beneficial to reproduction than in strictly monogamous mating systems.

Effects of evolutionary mechanisms on the Y chromosome can be seen when analyzing ampliconic gene copy number. All ampliconic gene families are contained within palindromes (P1, P2, P3, P4, P5, P8), and an inverted repeat (IR2), with the exception of *TSPY* (Bhowmick et al. 2007). The arrangement of the genes explains some of the variability in copy number seen in the study. Gene conversion, or non-reciprocal transfer of DNA from one duplex to another, occurs frequently in palindrome arms and allows the fixation of near identical copies of the ampliconic genes found in the palindromes (Rozen et al. 2003). However, since *TSPY* is arrayed outside of the palindromes, it is not subjected to the same levels of gene conversion and has a higher variability in copy number.

There was a positive correlation between median ampliconic gene copy number and variance; generally, genes with greater ampliconic gene CN had greater variation in CN among individuals included in the study. In this study, *TSPY* had the greatest level of variability of all the ampliconic genes; additionally, it displays variability between different haplogroups. It could have been under different selective pressures or influenced by genetic drift, accounting for the variability in copy number. *TSPY* had lower copy numbers in European haplogroups and higher copy numbers in African haplogroups.

The European haplogroup R had higher *RBMY* copy numbers than other haplogroups studied, which could be due to different selective pressures on the population after migration out of Africa. Additionally, *DAZ* showed significant differences in ampliconic gene copy number among haplogroups. Human males usually have four copies of *DAZ*, and lower copy numbers of the gene often results in infertility or subfertility (Jan et al. 2002). For the other ampliconic

genes, *BPY*, *CDY*, *HSFY*, *PRY*, *VCY*, and *XKRY*, there was no significant difference in copy number among different haplogroups, and minimal variation in copy number between different individuals.

When comparing principle component analysis of SNP data to ampliconic gene data, the separation of clustering for SNP data is more pronounced than separation of ampliconic gene copy number data for each haplogroup. Over 300 SNPs were used in the analysis, whereas only nine ampliconic genes were used. The larger number of SNPs available allowed SNP data to be a better predictor of haplogroup. Additionally, there might have been biological reasons for the differences in clustering; ampliconic genes may have evolved at a slower rate than SNPs, and balancing selective and gene conversion could have occurred, allowing the fixation of certain copy numbers of ampliconic genes.

Conclusion and Future Work

The study of ampliconic gene copy number from males in different geographic locations provides a better understanding of Y chromosome evolution and how ampliconic gene copy number differs between humans from different haplogroups. Ampliconic gene copy number variation between different haplogroups was statistically significant among *RBMY*, *TSPY*, and *DAZ* but not for the other ampliconic genes in the study. *RBMY* had a copy number between 5 and 16 and *TSPY* had the biggest range of observed ampliconic gene copy numbers, between 16 and 40.

The trends observed indicate that evolutionary forces, possibly selection and/or genetic drift, played an important effect on the amplification of genes found on the Y chromosome. Genetic drift likely plays a more important role in Y chromosome evolution than autosomal chromosome evolution because of the Y chromosome's smaller effective population size. Additionally, since the ampliconic genes have been implicated in spermatogenesis and male-specific reproductive functions, selective pressures probably played a role in fixing copy numbers that gave a reproductive advantage. The trends in ampliconic gene copy number will be further evaluated to determine their correlation, or lack thereof, with a phylogenetic tree based on the SNP data.

We will also investigate the possible mechanisms and causes of statistically significant variation in ampliconic copy number between different haplogroups as well as possible correlations between sexually dimorphic phenotypic characteristics and ampliconic gene copy number.

Because of the differences in effective population size and inheritance patterns between the Y chromosome and autosomes, further comparisons and study of differences between Y chromosomal evolution and autosomal evolution can be done using the data generated, leading to a greater understanding of the differences between evolution of autosomes and evolution of the Y chromosome. For example, previous studies have compared phylogeny based on Y-SNPs to phylogeny mitochondrial DNA; our lab could compare a similar phylogeny that includes ampliconic gene copy number information to shed more light onto the differences in evolutionary patterns between autosomes and sex chromosomes.

Future directions include testing whether the variation in gene copy numbers among populations is more than that expected under genetic drift. If it is, it could indicate the role of selection in driving differences in gene copy numbers among population.

Overall, the work provides a better understanding of human genetics by focusing on the relatively understudied Y chromosome and its evolution.

Supplementary Information: Rstudio and Plink Analysis Scripts

#data from results (used in R for initial analysis) to generate ampliconic gene copy number graphs

```
library(ggplot2)
dat<-read.table('clipboard',sep="\t",header=TRUE)
View(dat)
pcs<-prcomp(dat[,-c(1:5)],center=T)
names(pcs)
pcs.x<-pcs$x
head(pcs.x)
dim(pcs.x)
pcs.x<-as.data.frame(pcs.x)
pcs.x$major_haplo<-as.character(dat$major_haplo)
View(dat)
View(pcs.x)
ggplot(dat,aes(major_haplo,gene,fill=major_haplo))+geom_boxplot()+geom_point(major_haplo
gene,color=major_haplo)
#where gene = BPY, CDY, HSPY, RBMY, TSPY, XKRY, DAZ, PRY, VCY
```

#script for PCA test for ampliconic gene copy number

```
ggplot(pcs.x,aes(PC1,PC2,color=major_haplo))+geom_point()
ggplot(pcs.x,aes(PC1,PC3,color=major_haplo))+geom_point()
```

#plot of median ampliconic gene copy number and variance

```
library(plyr)
library(reshape2)
mdat<-melt(dat[,-c(2:5)],id.vars="ID")
colnames(mdat)<-c("ID","Gene","CN")
ddat<-ddply(mdat,.(Gene),summarize,median=median(CN),variance=var(CN))
```

#ANOVA analysis of ampliconic gene copy number data

```
(anova(lm(data=dat,gene~major_haplo)))
#where gene = BPY, CDY, HSPY, RBMY, TSPY, XKRY, DAZ, PRY, VCY
```

#SNP data was converted and formatted using Plink

```
open .map file in text editor
copy and paste into excel
replace chr column (first) with 2
save as .map file (poly2.map)
copy and save .ped file with the same name (poly2.ped)
plink --file poly2 --pca 10 --out poly2
plink --file ched_polysnps2 --pca 10 --out poly_snps2
```

results in two files:
poly2.eigenvec and poly2.eigenval
copy and paste poly2.eigenvec into a spreadsheet
add haplogroup, major_haplogroup, geographic information to this sheet

#Rstudio scripts to analyze SNP data

```
pcs<-read.table('clipboard',sep="\t",header=TRUE)
library(ggplot2)
ggplot(pcs,aes(PCx,PCy,color=major_haplo))+geom_point(alpha=0.7)+theme_bw()+labs(title="
PC1 vs PC2 based on Y-SNPS")
#where PCx and PCy are combinations such as PC1, PC2
```

BIBLIOGRAPHY

1. Bachtrog D. 2013. Y-chromosome Evolution: Emerging Insights into Processes of Y-chromosome Degeneration. Nature Reviews Genetics. 14(2): 113-24.
2. Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho T, Koutseva N, Zaghlul S, Graves T, Rock S, Kremitzki C, Fulton RS, Dugan S, Ding Y, Morton D, Khan Z, Lewis L, Buhay C, Wang Q, Watt J, Holder M, Lee S, Nazareth L, Alföldi J, Rozen S, Muzny DM, Warren WC, Gibbs RA, Wilson RK, Page DC. 2014. Mammalian Y Chromosomes Retain Widely Expressed Dosage-sensitive Regulators. Nature. 508 (7520): 494-499.
3. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Wheeler DL. 2007. GenBank. Nucleic Acids Research. 36 (suppl_1): D25-D30.
4. Bhowmick BK, Satta Y, Takahata N. 2007. The Origin and Evolution of Human Ampliconic Gene Families and Ampliconic Structure. Genome Research. 17 (4) : 441-50.
5. Bio-Rad. 2017. Digital PCR for Copy Number Variation Analysis. <http://www.bio-rad.com/en-us/applications-technologies/digital-pcr-for-copy-number-variation-analysis#2> Accessed 2017 Feb 27.
6. Bio-Rad. 2017. Droplet Digital™ PCR (ddPCR™) Technology. <http://www.bio-rad.com/en-us/applications-technologies/droplet-digital-pcr-ddpcr-technology#2> Accessed 2017 Feb 27.

7. Bio-Rad. 2017. Droplet Digital™ PCR (ddPCR™) Technology. <http://www.bio-rad.com/en-us/applications-technologies/droplet-digital-pcr-ddpcr-technology>. Accessed 2017 Jan 3.

8. Bio-Rad. 2017. QX200™ DdPCR™ EvaGreen Supermix #1864034. <http://www.bio-rad.com/en-jp/sku/1864034-qx200-ddpcr-evagreen-supermix>. Accessed 2017 Jan 2.

9. Bio-Rad. ddPCRTM supermix for probes. <http://www.bio-rad.com/webroot/web/pdf/lsr/literature/10026235.pdf>. Accessed 2017 Jan 2.

10. Carrel L, Willard HF. 2005. X-inactivation Profile Reveals Extensive Variability in X-linked Gene Expression in Females. Nature. 434(7031): 400-04.

11. Clayton JA, Collins FC. 2014. Policy: NIH to Balance Sex in Cell and Animal Studies. Nature 509 (7500): 282-83.

12. Dean A, Bidshahri R, Haynes C, Bryan J. 2016. Ddpcr: An R Package and Web Application for Analysis of Droplet Digital PCR Data. F1000Research. 5:1411.

13. Geoffrey MP, Do D, Litterst CM, Maar D, Hindson CM, Steenblock ER, Legler TC, Jouvenot Y, Marrs SH, Bemis A, Shah P, Wong J, Wang S, Sally D, Javier L, Dinio T, Han C, Brackbill TP, Hodges SP, Ling Y, Klitgord N, Carman GJ, Berman JR, Koehler RT, Hiddessen AL, Walse P, Bousse L,Tzonev S, Hefner E, Hindson BJ, Cauly TH, Hamby K, Patel VP, Regan JF, Wyatt PW, Karlin-Neumann GA, Stumbo DP, Lowe AJ. 2013. Multiplexed Target Detection Using DNA-Binding Dye Chemistry in Droplet Digital PCR. Analytical Chemistry. 85 (23): 11619-1627.

14. Ghenu A, Bolker BM, Melnick DJ, Evans BJ. 2016. Multicopy Gene Family Evolution on Primate Y Chromosomes. BMC Genomics. 17 (1): 157.

15. Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. 2008. Sex-Biased Evolutionary Forces Shape Genomic Patterns of Human Diversity. PLoS Genetics. 4(9): e1000202.

16. Harley VR, Jackson DI. DNA Binding Activity of Recombinant SRY from Normal Males and XY Females. 1992. Trends in Cell Biology. 2 (4): 97.

17. Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, Bright IJ, Lucero MY, Hiddessen AL, Legler TC, Kitano TK, Hodel MR, Petersen JF, Wyatt PW, Steenblock ER, Shah PH, Bousse LJ, Troup CB, Mellen JC, Wittmann DK, Erndt NG, Cauley TH, Koehler RT, So AP, Dube S, Rose KA, Montesclaros L, Wang S, Stumbo DP, Hodges SP, Romine S, Milanovich FP, White HE, Regan JF, Karlin-Neumann GA, Hindson CM, Saxonov S, Colston BW. 2011. High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. Analytical Chemistry. 83(22): 8604-610.

18. International Society of Genetic Genealogy. 2011. Y-DNA Haplogroup Tree 2011, Version: 12.89. <http://www.isogg.org/tree/> Accessed 2017 Jan 5.

19. Jones M, Williams J, Gärtner K, Phillips R, Hurst J, Frater J. 2014. Low Copy Target Detection by Droplet Digital PCR through Application of a Novel Open Access Bioinformatic Pipeline, ‘definetherain’. Journal of Virological Methods. 202 (100):46-53.

20. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New Binary Polymorphisms Reshape and Increase Resolution of the Human Y Chromosomal Haplogroup Tree. Genome Research. 18 (5): 830-38.

21. Krausz CL, Hoefsloot SM, Tüttelmann F. 2013. EAA/EMQN Best Practice Guidelines for Molecular Diagnosis of Y-chromosomal Microdeletions: State-of-the-art 2013. Andrology 2(1): 5-19.

22. Lahn BT, Pearson NM, Jegalian K. 2001. The Human Y Chromosome, in the Light of Evolution. Nature Reviews Genetics. 2(3): 207-16.

23. Lu C, Wen Y, Hu W, Lu F, Qin Y, Wang Y, L Si A, Yang S, Lin Y, Wang C, Jin L, Shen H, Sha J, Wang X, Hu Z, Xia Y. 2016. Y Chromosome Haplogroups Based Genome-wide Association Study Pinpoints Revelation for Interactions on Non-obstructive Azoospermia. Scientific Reports. 6 : 33363.

24. Miotto E, Saccenti E, Lupini L, Callegari E, Negrini M, Ferracin M. 2014. Quantification of Circulating MiRNAs by Droplet Digital PCR: Comparison of EvaGreen- and TaqMan-Based Chemistries. Cancer Epidemiology Biomarkers & Prevention. 23 (12): 2638-642.

25. [NCBI] National Center for Biotechnology Information. 2017 Feb 5. BPY2 Basic Charge, Y-linked, 2 [Homo Sapiens (human)]. <https://www.ncbi.nlm.nih.gov/gene/9083>. Accessed 2017 Feb 25.

26. [NCBI] National Center for Biotechnology Information. 2017 Feb 20. CDY1 Chromodomain Y-linked 1 [Homo Sapiens (human)]. < https://www.ncbi.nlm.nih.gov/gene/9085>. Accessed 2017 Feb 25.

27. [NCBI] National Center for Biotechnology Information. 2017 Feb 20. DAZ1 Deleted in Azoospermia 1 [Homo Sapiens (human)]. <https://www.ncbi.nlm.nih.gov/gene/1617>. Accessed 2017 Feb 25.

28. [NCBI] National Center for Biotechnology Information. 2017 Feb 17. HSFY1 Heat Shock Transcription Factor, Y-linked 1 [Homo Sapiens (human)]. < https://www.ncbi.nlm.nih.gov/gene/86614>. Accessed 2017 Feb 25.

29. [NCBI] National Center for Biotechnology Information. 2014 Sept 09. Polymerase Chain Reaction (PCR). <https://www.ncbi.nlm.nih.gov/probe/docs/techpcr/>.Accessed 2017 Jan 2.

30. [NCBI] National Center for Biotechnology Information. 2017 Feb 20. PRY PTPN13-like, Y-linked [Homo sapiens (human)]. <https://www.ncbi.nlm.nih.gov/gene/9081>. Accessed 2017 Feb 25.

31. [NCBI] National Center for Biotechnology Information. 2017 Feb 20. RBMY1A1 RNA Binding Motif Protein, Y-linked, Family 1, Member A1 [Homo Sapiens (human)]. <https://www.ncbi.nlm.nih.gov/gene/5940>. Accessed 2017 Feb 25.

32. [NCBI] National Center for Biotechnology Information. 2017 Feb 25. SRY Sex Determining Region Y [Homo Sapiens (human)]. <https://www.ncbi.nlm.nih.gov/gene/6736>. Accessed 2017 Feb 27.

33. [NCBI] National Center for Biotechnology Information. 2017 Feb 20. TSPY1 testis specific protein, Y-linked 1 [Homo sapiens (human)]. <https://www.ncbi.nlm.nih.gov/gene/7258>. Accessed 2017 Feb 25.

34. [NCBI] National Center for Biotechnology Information. 2017 Feb 16. VCY variable charge, Y-linked [Homo sapiens (human)]. <https://www.ncbi.nlm.nih.gov/gene/9084>. Accessed 2017 Feb 25.

35. [NCBI] National Center for Biotechnology Information. 2017 Feb 20. XKRY XK related, Y-linked [Homo sapiens (human)]. <https://www.ncbi.nlm.nih.gov/gene/9082>. Accessed 2017 Feb 25.

36. [NHGRI] National Human Genome Research Institute. 2015 June 16. Polymerase Chain Reaction (PCR) Fact Sheet. <https://www.genome.gov/10000207/polymerase-chain-reaction-pcr-fact-sheet/>. Accessed 2017 Jan 2.

37. [NIH] National Institutes of Health. 2017 Feb 28. Turner Syndrome - Genetics Home Reference. <https://ghr.nlm.nih.gov/condition/turner-syndrome#genes>. Accessed 2017 March 4.

38. [NIH] National Institutes of Health. 2017 Feb 28. Y Chromosome - Genetics Home Reference. <https://ghr.nlm.nih.gov/chromosome/Y>. Accessed 2017 March 4.

39. Ortona E , Pierdominic M, Maselli A, Veroni C, Aloisi F, Shoenfeld Y. 2016. Sex-based Differences in Autoimmune Diseases. Annali. 52 (2): 205-12.

40. Oven MV, Geystelen AV, Kayser M, Decorte R, Larmuseau MH. 2013. Seeing the Wood for the Trees: A Minimal Reference Phylogeny for the Human Y Chromosome. Human Mutation. 35(2): 187-91.

41. Pinheiro LB, Coleman VA, Hindson CM, Herrmann J, Hindson BJ, Bhat S, Emslie KR. 2012. Evaluation of a Droplet Digital Polymerase Chain Reaction Format for DNA Copy Number Quantification. Analytical Chemistry. 84(2): 1003-011.

42. Purcell, S. Package: PLINK (v.1.9) <http://pngu.mgh.harvard.edu/purcell/plink/> Accessed 2017 March.

43. Quintana-Murci L, Fellous M. 2001. The Human Y Chromosome: The Biological Role of a Functional Wasteland. Journal of Biomedicine and Biotechnology. 1(1): 18-24.

44. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. 2003. Abundant Gene Conversion between Arms of Palindromes in Human and Ape Y Chromosomes. Nature. 423(6942): 873-76.

45. RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA <http://www.rstudio.com/> Accessed 2017 March.

46. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou S, Latrielle P, Leonard S, Mardis E, Maupin R, Mcpherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang S, Waterston RH, Wilson RK, Rozen S, Page DC. 2003. The Male-specific Region of the Human Y Chromosome Is a Mosaic of Discrete Sequence Classes. Nature 423 (6942) : 825-37.

47. Smith AJ. De, Walsh KM, Hansen HM, Endicott AA, Wiencke JK, Metayer C, Wiemels JL. 2015. Somatic Mutation Allelic Ratio Test Using DdPCR (SMART-ddPCR): An Accurate Method for Assessment of Preferential Allelic Imbalance in Tumor DNA. Plos One. 10 (11): e0143343.

48. The Y Chromosome Consortium. 2002. A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups. Genome Research. 12(2): 339-48.

49. Thermo Fisher Scientific. 2017. PCR Buffers. <https://www.thermofisher.com/us/en/home/life-science/pcr/pcr-reagents/pcr-buffers.html>. Accessed 2017 Jan 2.

50. Tomaszkiewicz M, Medvedev P, Makova KD. 2017. Y and W Chromosome Assemblies: Approaches and Discoveries. Trends in Genetics. In Press.

51. Tomaszkiewicz M, Rangavittal S, Cechova M, Sanchez RC, Fescemyer HW, Harris R, Ye D, O'brien PM, Chikhi R, Ryder OA, Ferguson-Smith MA, Medvedev P, Makova KD. 2016. A Time- and Cost-effective Strategy to Sequence Mammalian Y Chromosomes: An Application to the De Novo Assembly of Gorilla Y. Genome Research 26 (4): 530-40.

52. Trypsteen W, Kiselinova M, Vandekerckhove L, & Spiegelaere WD. 2016. Diagnostic utility of droplet digital PCR for HIV reservoir quantification. Journal of Virus Eradication. 2(3):162–169.

53. U.S. National Library of Medicine. National Institutes of Health. 14 March 2017. SRY Gene - Genetics Home Reference. Accessed 2017 March 17.

54. Vossen, MR, White SJ. 2016. Quantitative DNA Analysis Using Droplet Digital PCR. Methods in Molecular Biology Genotyping. 1492: 167-77.

55. Vries J, Hoffer J, Repping S, Hoovers JM, Leschot NJ, Veen FV. 2002. Reduced Copy Number of DAZ Genes in Subfertile and Infertile Men. Fertility and Sterility 77(1): 68-75.

56. Wei W, Fitzgerald TW, Ayub Q, Massaia A, Smith BH, Dominiczak AF, Morris AD, Porteous DJ, Hurles ME, Tyler-Smith X, Xue Y. 2015. Copy Number Variation in the Human Y Chromosome in the UK Population. Human Genetics 134 (7): 789-90.

57. Zhao RF. 2006. The Y Chromosome: Beyond Gender Determination. < https://www.genome.gov/27557513/the-y-chromosome-beyond-gender-determination/>. Accessed 2016.

ACADEMIC VITA

Danling Ye

Email: dxy5063@psu.edu
Phone: (484) 885-9841

EDUCATION

The Pennsylvania State University Schreyer Honors College, *University Park, PA*
Class of 2017

Major: Veterinary and Biomedical Sciences
Minor: French and Francophone Studies

PUBLICATION

Tomaszkiewicz M, Rangavittal S, Cechova M, Sanchez R, Fescemyer H, Harris R, **Ye D**, O'Brien P, Chikhi R, Ryder A, Ferguson-Smith M, Medvedev P, Makova K. A time- and cost-effective strategy to sequence mammalian Y Chromosomes: an application to the de novo assembly of gorilla Y. *Genome Research.* 26.4: 530-40. 2016 March.

RESEARCH EXPERIENCES

Dr. Makova's Lab, *University Park, PA*
Undergraduate Researcher: *January 2014-present*

- Performs research to elucidate the sequence and evolution of hominine Y chromosomes
- Studies DNA samples using gel electrophoresis, Qubit fluorometer, and PCR machines
- Participates in project meetings and discussions of scientific publications

GlaxoSmithKline, *Collegeville, PA*
Scientists Student Intern: Summer 2016

- Uses bioinformatics to analyze expression of oncogenes
- Discovers trends in RNA sequencing data and microarray data to screen potential cancer therapeutic targets
- Generates graphs, figures, and summaries of findings

OTHER EXPERIENCES

West Chester Veterinary Medical Center, *West Chester, PA* — **Shadow**: *Summer 2014*
Philadelphia Zoo, *Philadelphia, PA* — **Conservation Steward**: *Summer 2014*
Hope Veterinary Specialists, *Malvern, PA* — **Volunteer**: *Summer 2014*
Main Line Animal Rescue, *Phoenixville, PA* — **Volunteer**: *January 2012-August 2014*
Penn State Dairy Barns, *University Park, PA* — **Employee**: *Summer 2016*

LEADERSHIP POSITIONS

Circle K, International Service Organization — **Secretary**: *May 2015-present*
Penn State Alternative Breaks — **Site Director**: *August 2015-present*

HONORS AND AWARDS

Provost Award | Student Leader Scholarship | Dean's List, multiple semesters