Genetic Epidemiology of Smoking Behaviors in Samoans

by

Ramya Janani Saravanan

BS, University of Pittsburgh, 2020

Submitted to the Graduate Faculty of the

Human Genetics Department

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Public Health

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This essay is submitted

by

Ramya Janani Saravanan

on

April 29, 2021

and approved by

Essay Advisor: Candace M. Kammerer, PhD, Associate Professor, Human Genetics, Graduate School of Public Health, University of Pittsburgh

Essay Reader: Iva Miljkovic, MD, PhD, FAHA, Associate Professor, Epidemiology, Graduate School of Public Health, University of Pittsburgh

Essay Reader: Ryan L. Minster, PhD, MSIS, Associate Professor, Human Genetics, Graduate School of Public Health, University of Pittsburgh

Genetic Epidemiology of Smoking Behaviors in Samoans

Ramya Janani Saravanan, MPH

University of Pittsburgh, 2021

Abstract

Smoking tobacco is one of the leading causes of death around the world, and it is also a critical risk factor for many chronic diseases, including lung cancer and heart disease. Although smoking rates have declined in many countries, tobacco use continues to be a problem in other regions, such as the Pacific Islands, where 30–50% of the population are smokers. Researchers have reported that smoking behaviors are influenced by both environmental and genetic factors, but the specific loci involved are mostly unknown. In addition, the majority of genetic studies have been conducted in Caucasian populations, so the results may not be applicable to other ethnic groups. I investigated whether 20 genetic variants reported to influence smoking behaviors had a similar effect in Samoans, using genotype and demographic data available for 3,476 Samoan participants from the Soifua Manuia (Good Health) study. I also compared the allele frequencies (AFs) of these 20 loci in Samoans to those in other ethnic groups using principal component analysis. As expected, the Samoan population clustered most closely with East Asian populations, yet they had unique AFs for many loci. Regarding smoking phenotypes, males were more likely to be current smokers (CS) ($p < 2.2\times10^{-16}$), and they smoked more cigarettes per day (CPD) than females ($p = 2.0\times10^{-12}$). The association analyses revealed that the G allele of locus rs848353 was associated with increased odds of being a current smoker (OR = 2.635 [2.433–35.63]; $p = 0.001$),

whereas the G allele of rs1329650 corresponded to a significant increase in CPD ($\beta = 0.146 \pm 0.044$; $p = 0.001$); both variants are in noncoding regions of the genome. Additional genetic studies in Polynesians are needed to confirm these findings and help us understand the function of these noncoding regions in the etiology of addictive disorders. This knowledge may help development of targeted interventions for smoking cessation, eventually leading to a reduction in chronic disease morbidity and mortality and overall improving public health in Samoa.

Table of Contents

List of Tables

List of Figures

Preface

We would like to thank the Samoan participants and local village authorities and research assistants over the years. We acknowledge the support of our research collaboration with the Samoa Ministry of Health; the Samoa Bureau of Statistics; and the Samoa Ministry of Women, Community and Social Development. The Samoan Obesity, Lifestyle, and Genetics Adaptations (OLaGA) Study Group investigators are Ranjan Deka, Jenna Carlson, Kima Fa‘asalele-Savusa, Nicola L. Hawley, Vaimoana Lupematisila, Stephen T. McGarvey, Ryan L. Minster, Leausa Toleafoa Take Naseri, Muagututi‘a Sefuiva Reupena, Melania Selu, John Tuitele, Asiata Satupa‘itea Viali and Daniel E. Weeks.

1.0 Introduction

1.1 Smoking: A Public Health Issue

Tobacco use has long been one of the world's biggest health issues, and smoking is one of the leading causes of premature death. In 2017 alone, smoking accounted for more than 7 million deaths globally (Stanaway et al., 2018). In addition, tobacco use is the biggest external risk factor for many chronic diseases, including lung cancer, heart disease, and respiratory diseases, which are also among the leading causes of death around the world (Jha & Peto, 2014). Thanks to public health interventions, smoking rates have decreased in many countries, but tobacco use continues to be a problem in certain regions of the world.

Rates of smoking often correlate with a country's level of economic development. An article by Jha and Peto in the New England Journal of Medicine described the shifting trends in tobacco use around the world. They noted that high-income, developed countries had high smoking rates throughout the 20th century. However, there has since been a dramatic decline in the prevalence of smoking due to persistent public health efforts, and cigarette consumption per adult has been nearly halved in some countries (Jha & Peto, 2014). Unfortunately, smoking prevalence has shifted to low and middle-income countries (defined by World Bank gross national income standards), where smoking cessation is still uncommon. Bilano and colleagues modeled global trends in smoking prevalence from 2010 to 2025 in men and women; they reported that tobacco use among men will likely increase in low and middle-income countries in eastern Europe, the western Pacific, and Africa. Tobacco control may not be a top priority in these countries due to limited resources and the need to address other pressing health concerns (Bilano et al., 2015).

Future global health interventions (e.g., awareness, prevention efforts, etc.) should be focused on these low-to-middle income countries to help them bring down the rates of smoking in their populations.

1.2 Prevalence of Smoking in Samoa

In the Pacific Islands, a region of the world that is already experiencing a rise in non-communicable diseases due to globalization and changing diets/behaviors, tobacco smoking adds to the burden by increasing the risk of developing these diseases. Southeast Asia and the Pacific region have some of the highest rates of smoking in the world, particularly among men, ranging from 30–50% of the population (Adia, Hawley, Naseri, Reupena, & McGarvey, 2019). This is substantially higher than the global smoking rate estimated by the World Health Organization: 20.2% of individuals aged ≥ 15 are current smokers (Commar, Prasad, d'Espaignet, & Wolfenden, 2018).

Tobacco use has steadily declined in Samoa over the years, from nearly 64.1% of men and 21.0% of women smoking tobacco in 1991, to 39.5% and 16.8%, respectively, in 2013 (Linhart et al., 2017). Linhart and colleagues analyzed population-based surveys of Samoans aged 25–64 and found that daily tobacco smoking also substantially dropped from 1978 (75.8% in men; 26.5% in women) to 2013 (36.3% in men; 14.9% in women). Despite the progressive decrease, the percentage of male smokers still remains fairly high. The prevalence of tobacco use among younger age groups is worrisome as well, with 32.2% of Samoan adolescents smoking cigarettes, which is higher than in other low and middle-income countries (Adia et al., 2019). These percentages illustrate the extent to which smoking is a significant public health concern in Samoa.

1.2.1 Social Factors & Trends in Smoking Initiation

Tobacco use around the world is often associated with men of lower levels of education and socioeconomic status (Adia et al., 2019). These trends were also reflected in the Samoan population. A 2005 cross-sectional study of 1,834 Samoan adults found that current smokers were more likely to be male, younger (18–34 years), single, and have lower incomes (< $20,000/year) (Mishra, Osann, & Luce, 2005). The 2014 Demographic and Health Survey conducted by the Samoan government noted that among males aged 15–49, those in the highest wealth quintile and with more than secondary education generally had lower rates of tobacco use (Adia et al., 2019).

Smoking is well entrenched in Samoan society and is seen as a way of socializing with friends and sometimes family. Based on results from focus groups conducted in Apia, Samoa's capital city, Tanielu et al. found that most individuals had their first smoking experience in a social gathering with peers. This first experience commonly occurred in the late teens and early twenties. Many individuals stated that they started out of curiosity or picked up the habit along with drinking. Results from the focus groups also revealed some widely held misconceptions that smoking is beneficial in increasing energy, relieving indigestion, and accelerating the effects of alcohol. Female smokers viewed it as a relief from the stress and pressures of their daily lives (Tanielu, McCool, Umali, & Whittaker, 2018).

Despite citing the perceived benefits of smoking, the majority of Samoans acknowledge that tobacco use is generally harmful to health. Moreover, 81% of participants in the focus groups expressed that they intended to quit smoking and two-thirds of the group had previously tried to quit. The primary reasons for giving up the habit were out of concern for family and tensions at home. The cost of cigarettes took a toll on finances, and smokers often found it difficult to prioritize the needs of the family over the desire to satisfy their addiction (Tanielu et al., 2018).

1.3 Genetic Factors

1.3.1 Evolutionary History of Samoans

The region of Oceania was first inhabited at least 50,000 years ago by the ancestors of Aboriginal Australians and Papuans. A second wave of migration occurred 5,000 years ago when the region was colonized by peoples from the islands of Southeast Asia who spoke Austronesian languages. Genetic analyses show that there was initially minimal admixture between the Austronesians and the Papuans. However, as expected, admixture increased over time, and modern-day speakers of Austronesian languages in the Pacific have 70–80% Austronesian and 20–30% Papuan ancestry (Harris et al., 2020).

The island of Samoa was founded around 2,800 years ago. Researchers hypothesize that early inhabitants initially lived in small, isolated groups that later formed into a system of complex chiefdoms. Unfortunately, the island's sparse early archaeological record has made it difficult to decipher its population history (Harris et al., 2020). Population genetics may help answer some of these questions which is why it is crucial to include these groups in genetic studies; yet Oceanic populations, and Polynesian populations in particular, continue to be undersampled in research. The phenomenon of genetic drift due to founder effects, small/isolated populations, and population bottlenecks has led to unique variants and allele frequencies (AFs) in Samoans (Minster et al., 2016). Comparing the frequencies of these variants with those in other populations can provide important insight into the genetics behind various phenotypic traits.

1.3.2 Genetics of Nicotine Dependence

Nicotine is the primary addictive compound in tobacco products, and tobacco use is specifically motivated by nicotine dependence (ND). Understanding the factors that contribute to ND is necessary in order to decrease the prevalence of smoking. Addiction is not commonly regarded in a hereditary context, yet numerous studies have shown that there is a significant genetic component to many addictive behaviors, including ND. Multiple quantitative genetic studies conducted on monozygotic (identical), dizygotic (fraternal) twins, and adopted children from many ethnic groups reported that heritability of ND and the number of cigarettes smoked per day (CPD) was consistently between 50–60%, after accounting for variation due to environmental factors (Gorwood, Le Strat, & Ramoz, 2017; Mackillop, Obasi, Amlung, McGeary, & Knopik, 2010; Yoon et al., 2012). However, these studies do not provide insight into the specific genes associated with increased risk of addiction.

To identify possible susceptibility genes, one approach is to assess variation in genes involved in the molecular pathways for nicotine metabolism. Genes associated with differences in nicotine pharmacokinetics, which is a person's metabolic capacity, are most likely implicated in ND (Mackillop et al., 2010). Individuals that metabolize nicotine faster are more dependent on it and exhibit greater ND. On the other hand, people with slower nicotine metabolism tend to have lower CPD counts. They also respond better to nicotine patch therapy and thus have a higher probability of successful smoking cessation (Claw et al., 2020).

Nicotine metabolism is largely determined by CYP2A6, the primary enzyme that catalyzes the conversion of approximately 75% of the nicotine that enters the body into cotinine. The enzyme is encoded by the highly polymorphic *CYP2A6* gene; some genetic variants substantially reduce enzymatic metabolism of nicotine, while others increase nicotine metabolism. Individuals with a

loss-of-function mutation in just one allele have ~50% decrease in enzymatic activity and slower metabolism compared to wild-type. The CYP2B6 enzyme also metabolizes nicotine in the liver but does so less efficiently than CYP2A6. A recent study found that *CYP2A6* and *CYP2B6* interact and balance each other out if there is a deficiency in one of the enzymes (Mackillop et al., 2010).

The receptors in this pathway also play an important role in mediating nicotine metabolism. Nicotine acts as an agonist for nicotinic acetylcholine receptors (nAChRs), a family of ligand-gated ion channels that are expressed in the brain and various other tissues (Conlon & Bewick, 2011). The *CHRNA5-CHRNA3-CHRNB4* gene cluster on the long arm of chromosome 15q25 encodes the α5, α3, and β4 subunits of the receptors, respectively, and genetic variants in this region have been associated with ND and other smoking-related traits. In particular, the nonsynonymous variant rs16969968 in *CHRNA5* has consistently yielded the strongest association with ND, as well as lung cancer, across many studies. Other variants in the cluster may also contribute to the effect on ND by regulating gene expression. Assigning causality is difficult due to the high linkage disequilibrium (LD) in this region (S.-H. Lee, Ahn, Seweryn, & Sadee, 2018).

In addition to the key genes involved in nicotine metabolism, genes involved in other pathways may also play a role in ND. Genome-wide association studies (GWAS) can assess millions of single-nucleotide variants (SNVs) for correlations with different smoking phenotypes, such as CPD, ND, and smoking cessation. Unfortunately, the majority of GWAS have been conducted in individuals of European descent. Given the distinct differences in LD between ethnic groups, applying results from these studies to the Samoan population is difficult. Once again, this highlights the importance of focusing on more diverse groups in these large-scale studies (Yoon et al., 2012).

A few GWAS have been conducted in East Asian populations, which are more similar to Samoans due to their evolutionary past. Yoon et al. conducted the first GWAS on smoking behaviors in an Asian population. They assessed 352,228 SNVs in a cohort of 8,842 Korean individuals for associations between these genetic variants and the smoking-related phenotypes, ND and smoking initiation (SI). Ordinal categories of CPD were used as a measure of ND, while SI was defined as a binary trait comparing individuals who never smoked to those who have had regular smoking experiences in their lifetime. Because an independent sample of Asian smokers was not available, they replicated their findings in European-Americans and African-Americans. The researchers discovered two SNVs that were significantly associated with SI: rs7747583 (p_{meta} = 6.40×10^{-6}) and rs2349433 (p_{meta} = 5.57×10^{-6}). Both are located on chromosome 6 in the *RGS17* gene, which regulates G-protein signaling, and are in complete LD. They also identified two SNVs that were associated with ND. One of these, rs4424567 (p_{meta} = 2.30×10^{-6}), was on chromosome 10 in the intronic region of the *FRMD4A* gene, which encodes a scaffolding protein. The other SNV, rs848353 (p_{meta} = 9.16×10^{-8}), was on chromosome 7 at the q31.1 locus. The team hypothesized a few biologically plausible interpretations that support these results, but further research into the function of these genes is needed (Yoon et al., 2012).

Matoba et al. conducted a GWAS for smoking-related traits (SI, smoking cessation, age of smoking initiation, and CPD) in a sample of 165,436 Japanese individuals and tested 6,108,828 imputed SNVs. They identified three new loci associated with CPD (*EPHX2–CLU*, *RET*, and *CUX2–ALDH2*) and three loci associated with SI (*DLC1*, *CXCL12–TMEM72-AS1*, and *GALR1–SALL3*). In particular, the lead SNV (highest association significance) of *DLC1*, rs117036946, is close to rs289519 which was found to be associated with ND in a previous GWAS in African Americans. The association of rs117036946 with SI remained significant even after

accounting for LD between the two SNVs ($p = 4.3\times10^{-8}$), suggesting that they are two independent signals. Furthermore, several of these loci were identified in a sex-stratified GWAS, indicating that the genetics of SI significantly varies not only by ethnicity, but also by sex (Matoba et al., 2019).

The difficulty with both of the above GWA studies is the lack of proper replication samples to confirm their findings. Unfortunately, most studies on smoking initiation and ND have only been done in Caucasian populations, although tobacco use is a significant health concern plaguing Asia and the Pacific Islands. This situation makes it even more imperative to conduct research in other populations to gain insights regarding smoking addiction and its underlying risk factors.

1.4 Specific Aims

The goal of this study is to determine if gene variants associated with smoking in other populations have similar effects in Samoans. Specifically, I will:

1. Analyze demographic data and distribution of smoking behaviors among the Samoan population.
2. Identify genes and SNVs reported to be associated with SI and ND in the literature.
3. Compare the frequencies of these risk alleles in Samoans to the AFs in other ethnic groups.
4. Perform association analyses between these SNVs and smoking phenotypes in the Samoan population.

2.0 Methods

2.1 Samples & Genotyping

The SNV and covariate data on Samoans came from the Soifua Manuia (Good Health) study which focused particularly on the issue of obesity and other related cardiometabolic disorders in Samoa. A total of 3,476 participants across 33 villages in Samoa were surveyed using a questionnaire about their health, family history, and various health behaviors.

All participants in this study gave written informed consent in Samoan language consent forms. The research in Samoa was reviewed and approved by the institutional review boards of Miriam Hospital, Providence, RI; Brown University; University of Cincinnati; and University of Pittsburgh. Research in Samoa was also reviewed and approved by the Health Research Committee of the Samoan Ministry of Health.

Of the participants, 3,072 individuals were genotyped for 659,492 markers using an Affymetrix 6.0 chip (Minster et al., 2016). Only individuals whose maternal and paternal grandparents were of Samoan ancestry were eligible for the study. Many variants in the genes involved in nicotine metabolism (*CYP2A6*, *CYP2B6*, *CHRNA5*, etc.) were not included on the chip, and I did not have access to imputed data which was a major limitation in my analysis.

2.2 Smoking-Related Phenotypes

2.2.1 Current Smoking Status

Current smoking status (CS) was defined as a binary trait, comparing individuals who currently smoke to those who do not. A better assessment of smoking status would have been to compare individuals who were smokers at any point in their lifetime (current or former) to those who have never smoked. However, there were many discrepancies in participants reporting smoking cessation, so I was unable to accurately determine which individuals were former smokers and include them in my analyses.

2.2.2 Cigarettes Per Day

The number of cigarettes smoked per day (CPD) was defined as a continuous trait and used as an indicator of nicotine dependence (ND). I was unable to develop other measures of ND that are commonly used in the literature, such as the Fagerström Test for Nicotine Dependence (FTND), because this information was not asked in the questionnaire. The primary form of tobacco use in Samoa was manufactured cigarettes, but the survey also included questions about alternate forms of tobacco use, and individuals reported using hand-rolled cigarettes, pipes, and cigars. Because many participants used alternate forms of tobacco, I converted these products to the equivalent number of cigarettes based on nicotine content (Table 1). Nicotine levels in cigarettes vary by manufacturer, so I set the mean nicotine content (1.45 mg) as the standard for one cigarette. Thus, for example, one pipe has approximately the same amount of nicotine as four cigarettes. To correct for the many outliers (likely a result of inaccurate reporting), I plotted the $\log_{10}$ of the total

CPD and excluded any values greater than three standard deviations below/above the mean. I used the log-transformed data in subsequent regression analyses to adjust for the highly skewed distribution of CPD (Figure 3).

Table 1 Nicotine content in tobacco products

Tobacco product	Nicotine content (mg)	Equivalent # of cigarettes
Cigarette	1.1–1.8	1
Pipe	5.2	4
Cigar	13.3	9

(Essenmacher, 2012)

2.3 Candidate Genes & SNVs

After conducting an extensive literature review (Section 1.3.2) and comparing the results to genomic databases, such as dbSNP and the GWAS Catalog, I compiled a list of genes and SNVs reported to be significantly associated with smoking status and CPD. Most genomic databases do not contain information on Samoans, or Polynesians in general, and very few genetic studies have been conducted in these populations. Therefore, I used studies and data on East Asian (EAS) populations as the standard for comparison. Unfortunately, many of the variants previously associated with risk of smoking were not included on the chip used for the Soifua Manuia study. Moreover, several risk variants were too rare in Samoa (MAF > 0.01) to obtain meaningful results and were excluded from further analysis. Table 2 below lists the SNVs I used for association testing, along with their location in the genome, alleles, and minor allele frequency (MAF) in the Samoan population.

Table 2 Candidate SNVs associated with smoking behaviors

Chromosome	Gene	Consequence	SNV	A1	A2 (minor)	MAF	Reference study
2	LOC105374754	ncRNA (Intron)	rs6718569	T	C	0.1799	(Matoba et al., 2019)
6	RGS17	Intron	rs2349433	G	A	0.4227	(Yoon et al., 2012)
7	–	Intergenic	rs848353*	A	G	0.1267	(Yoon et al., 2012)
7	–	Intergenic	rs1404697*	G	C	0.1210	(Yoon et al., 2012)
10	FRMD4A	Intron	rs4424567	A	G	0.4841	(Yoon et al., 2012)
10	HECTD2-AS1 (–)	ncRNA (Intron)	rs1329650	T	G	0.2417	(Furberg et al., 2010)
11	BDNF-AS (+)	ncRNA (Intron)	rs1013442	A	T	0.4755	(Furberg et al., 2010)
		ncRNA (Intron)	rs4923460*	T	G	0.4430	
		ncRNA (Intron)	rs879048*	C	A	0.4189	
11	BDNF (–)	Missense	rs6265	T	C	0.4314	(Furberg et al., 2010)
15	HYKK	Intron	rs9788682	G	A	0.1473	(Furberg et al., 2010)
		Intron	rs7163730	A	G	0.2314	(Li et al., 2010) (Furberg et al., 2010)
15	CHRNA3	Intron	rs6495308	C	T	0.4904	(Liu et al., 2010) (Li et al., 2010)
15	CHRNA5	Intron	rs16969948	A	G	0.1644	(Li et al., 2010)
15	CHRNB4	Upstream Variant	rs7166158	A	T	0.2263	(Li et al., 2010)
15	LOC112268142	ncRNA (Intron)	rs8043123	T	C	0.4151	(Li et al., 2010)
15	–	ncRNA (Intron)	rs11072793	G	A	0.3873	(Li et al., 2010)
17	CAMKK1	Intron	rs758642	G	A	0.1478	(Caporaso et al., 2009)
19	EGLN2	Intron	rs3733829	G	A	0.4352	(Furberg et al., 2010)
19	CYP2B6	Missense	rs3745274	G	T	0.3852	(Zanger et al., 2007)

A2 refers to the minor allele in Samoans. MAF is the frequency of this allele in the Samoan population.

***SNVs in high LD: rs848353 and rs1404697 (r^2 = 1); rs4923460 and rs879048 (r^2 = 0.83)**

2.4 Statistical Testing

2.4.1 Population Characteristics

The difference in mean age of male and female participants was compared using a two-sample *t* test with pooled variance. Differing trends in smoking behaviors were analyzed with a two-sample *Z* test (CS) or two-sample *t* test with unequal variances (CPD). The threshold *p*-value for all the tests was 0.05.

2.4.2 Principal Component Analysis

Principal component analysis (PCA) was used to compare allele frequencies (AFs) of the 20 SNVs of interest (Table 2) in Samoans to those in other populations. I obtained the AFs of the candidate SNVs from the dbSNP database and used the populations defined by the 1000 Genomes Project. I performed the PCA and plotted the results using the prcomp() function in the *stats* R package and the autoplot() function in *ggplot2*.

2.4.3 Association Testing

In order to determine if the SNVs found in the literature were associated with the smoking-related phenotypes in the Samoan population, I used the glm() function in the *stats* R package to perform logistic and linear regression analyses, adjusting for sex as a covariate due to the significant difference in smoking patterns between males and females. A logistic regression model was fit for the binary phenotype CS, and a linear regression model was fit for CPD. For both

phenotypes, I coded the alleles additively (number of copies of the minor allele) to determine if there was a dosage effect, as well as by genotype. The major allele or homozygous major allele genotype, respectively, was designated as the reference. I calculated odds ratios (OR) and 95% confidence intervals using the confint() function in R.

A threshold *p*-value of 0.003 was established after correcting for multiple testing and accounting for LD. I used the NIH's SNPclip tool to determine the number of SNVs with R^2 less than 0.8, which I considered to be independent SNVs; only AFs from East Asian populations were included because they share the closest genetic similarity to Samoans. Then, I divided the nominal significance level of 0.05 by the number of independent SNVs (18) to obtain the adjusted significance threshold ($p \leq 0.003$).

3.0 Results

3.1 Demographics

More females (58.7%) than males (41.3%) participated in this study (Table 3). The ages of the study participants ranged from 22 to 70, with a mean age of ~45, and the difference in mean age between males and females was not significant. Because the original target age range of the study was 25 to 65, there were very few participants between the ages of 20 to 25, but the number of individuals in the other age groups was equally distributed (Figure 1).

Table 3 Demographic variables

Characteristic	**Total (n = 3476)**	**Males (n = 1436)(41.3%)**	**Females (n = 2040)(58.7%)**	***p*-value (between genders)**
Age (mean ± SD)	44.8 ± 11.2	45.1 ± 11.4	44.6 ± 11.1	0.2
Current smoker	1181 (34.0%)	735 (51.2%)	446 (21.9%)	$< 2.2\times10^{-16}$
CPD (mean ± SD)	11.3 ± 10.1	12.7 ± 11.5	8.9 ± 6.6	2.0×10^{-12}

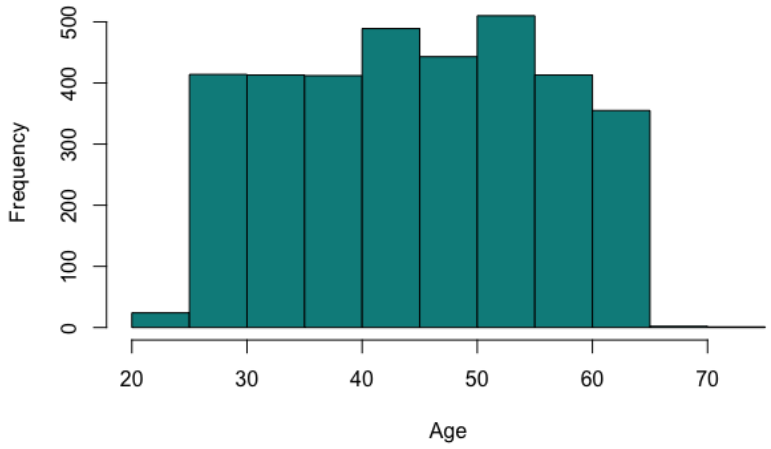

Figure 1 Age distribution of study participants

3.2 Smoking Trends & Distribution

The distribution of smoking status was analyzed by sex to see if there were any differences between the groups. Table 3 shows that 34% of the study sample were current smokers. However, the smoking frequency in males was significantly higher than in females (51% versus 22% respectively; $p < 2.2\times10^{-16}$); more than half of all male participants in the study were smokers.

The majority of participants started smoking in their late teens and early twenties (Figure 2). As described in the Methods, all tobacco products were converted into the equivalent number of cigarettes based on nicotine content. The mean CPD among the Samoan population was 11.3. Yet, just as smoking frequency differed by gender, mean CPD also differed significantly by gender (Table 3). The mean CPD in males was 12.7, whereas in females it was 8.9 ($p = 2.0\times10^{-12}$). Furthermore, as can be seen by the skewed distribution in Figure 3, 94.8% of participants smoked less than or equal to the nicotine content in a pack of cigarettes per day (~20 cigarettes), and 68.4% smoked less than or equal to half a pack per day. However, 2% of individuals smoked approximately two packs a day, and a few individuals smoked nearly 4–5 packs per day.

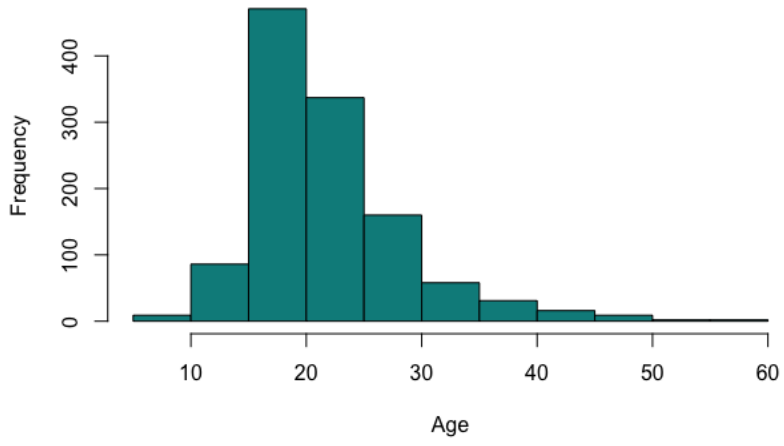

Figure 2 Age of smoking initiation

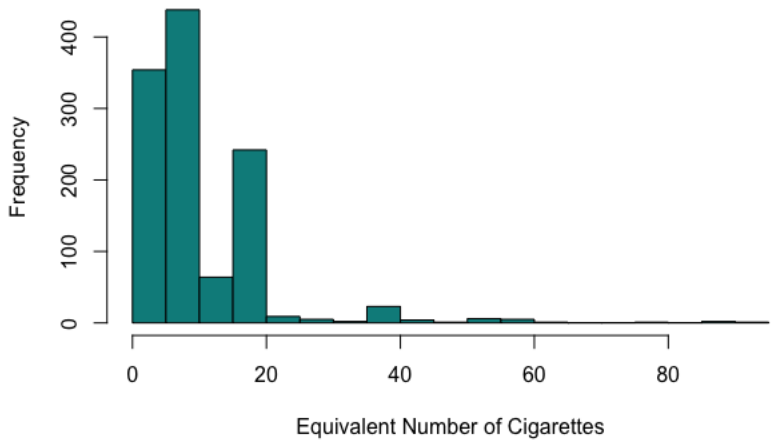

Figure 3 Cigarettes smoked per day

3.3 Allele Frequencies

Table 4 presents SNV allele frequencies (AFs) for Samoans and other "superpopulation" groups, as defined by the 1000 Genomes Project (Appendix Table 1). The frequencies of several alleles differed between Samoans and other populations. For example, the AF of rs3733829 [A] is 0.4352 in Samoans but ranges from 0.5060 to 0.9758 in the other populations; thus, the minor and major alleles are switched for this SNV.

To assess the overall relationships among the populations, I conducted principal component analysis (PCA) using the AFs of the 20 SNVs from each population (Table 4) and plotted the first and second principal components. These two components accounted for 45.71% and 28.61% of the variance, respectively, among AFs across the populations (Figure 4). The PCA plot displays the similarity between the Samoan and East Asian AFs by clustering them together. The South Asian, American (admixed Latino), and European populations loosely formed another cluster, whereas the African (sub-Saharan) populations were least similar to the other groups.

Table 4 Allele frequencies by population

Chromosome	SNV	Samoan	East Asian	American	South Asian	European	African
2	rs6718569 [C]	0.1799	0.1101	0.2450	0.0600	0.0666	0.2617
6	rs2349433 [A]	0.4227	0.3671	0.6240	0.4790	0.7048	0.4561
7	rs848353 [G]	0.1267	0.1429	0.0940	0.0710	0.0696	0.3593
7	rs1404697 [C]	0.1210	0.1429	0.0620	0.0690	0.0596	0.0378
10	rs4424567 [G]	0.4841	0.5536	0.4350	0.3940	0.2763	0.1452
10	rs1329650 [G]	0.2417	0.2619	0.6010	0.3850	0.7187	0.9380
11	rs1013442 [T]	0.4755	0.1498	0.2870	0.2720	0.2485	0.0280
11	rs4923460 [G]	0.4430	0.5556	0.8210	0.7290	0.7803	0.8381
11	rs879048 [A]	0.4189	0.5139	0.8070	0.7380	0.7813	0.6014
11	rs6265 [C]	0.4314	0.5119	0.8470	0.7980	0.8032	0.9894
15	rs9788682 [A]	0.1473	0.2887	0.4710	0.3920	0.2207	0.1747
15	rs7163730 [G]	0.2314	0.4256	0.5240	0.4030	0.2356	0.2458
15	rs6495308 [T]	0.4904	0.2669	0.4500	0.5300	0.7594	0.6762
15	rs16969948 [G]	0.1644	0.0585	0.0140	0.0020	0.0020	0.1437
15	rs7166158 [T]	0.2263	0.2500	0.1250	0.0700	0.0775	0.1498
15	rs8043123 [C]	0.4151	0.5308	0.5040	0.5340	0.7624	0.8222
15	rs11072793 [A]	0.3873	0.4573	0.6240	0.5260	0.7465	0.2005
17	rs758642 [A]	0.1478	0.1349	0.3570	0.2920	0.3459	0.3623
19	rs3733829 [A]	0.4352	0.6419	0.5060	0.5860	0.6372	0.9758
19	rs3745274 [T]	0.3852	0.2153	0.3730	0.3810	0.2356	0.3744

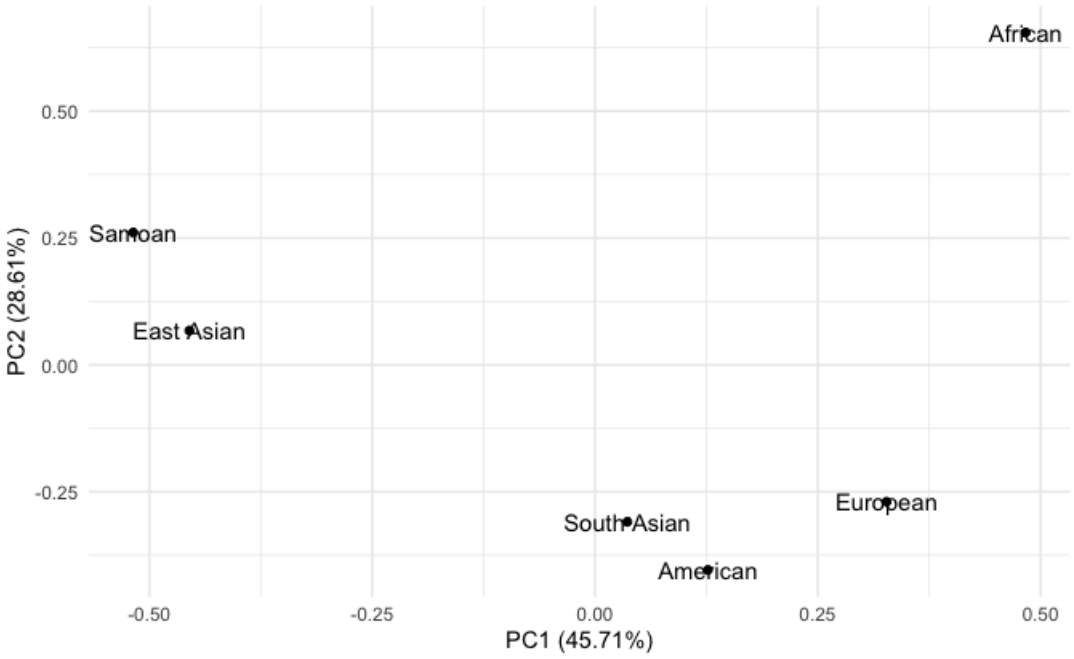

Figure 4 PCA plot of population allele frequencies

3.4 Association Analyses

The effect of the SNVs on CS and CPD in Samoans was analyzed using logistic and linear regression models with sex as a covariate. I coded the alleles additively (number of copies of the minor allele) to determine if there was a dosage effect, and I also analyzed them by genotype. Tables 5 and 6 contain the SNVs with suggestive ($p \leq 0.01$) and significant results ($p \leq 0.003$) from these tests. The complete tables are provided in the Appendix (Appendix Tables 1 and 2).

Homozygotes for the minor allele [G] for SNV rs848353 had significantly higher odds of being current smokers than heterozygotes or homozygotes for the major allele [A] (OR = 2.635 [2.433–35.631]; $p = 0.001$). In other words, two copies of the minor allele [G] more than doubled the odds of being a smoker. Figure 5 shows that ~50% of individuals with the genotype GG were

current smokers, which was higher than the proportion of smokers with the other genotypes. In contrast, homozygotes for the minor allele [T] for SNV rs3745274 were less likely to be smokers, although this effect did not meet the threshold for significance (OR = 0.710 [0.253–0.809]; p = 0.008).

For CPD, homozygotes for the minor allele [G] for SNV rs1329650 smoked significantly more cigarettes per day than heterozygotes or homozygotes for the more frequent allele [T] (β = 0.146 ± 0.044; p = 0.001). This meant that having two copies of the risk allele [G] corresponded to an increase in smoking quantity of 0.146 CPD compared to baseline. Figure 6 shows that the distribution of the $\log_{10}$ of CPD for the genotype GG only contained values of CPD at the higher end of the range and was clearly different from the distributions for the other genotypes.

Table 5 SNVs associated with smoking phenotypes (additive model)

SNV	A1	A2 (minor)	CS		CPD	
			OR [95% CI]	*p*-value	β ± SE	*p*-value
rs3745274	G	T	0.696 [0.532–0.911]	0.008		
rs1329650	T	G			0.049 ± 0.017	0.004

Table 6 SNVs associated with smoking phenotypes (by genotype)

SNV	Genotype	CS		CPD	
		OR [95% CI]	*p*-value	β ± SE	*p*-value
rs848353	AA	Reference (1.000)			
	AG	0.960 [0.586–1.407]	0.673		
	GG	2.635 [2.433–35.631]	*0.001		
rs3745274	GG	Reference (1.000)			
	GT	0.887 [0.514–1.119]	0.163		
	TT	0.710 [0.253–0.809]	0.008		
rs1329650	TT			Reference (0.000)	
	TG			0.025 ± 0.022	0.266
	GG			0.146 ± 0.044	*0.001

***Below significance threshold ($p \leq 0.003$)**

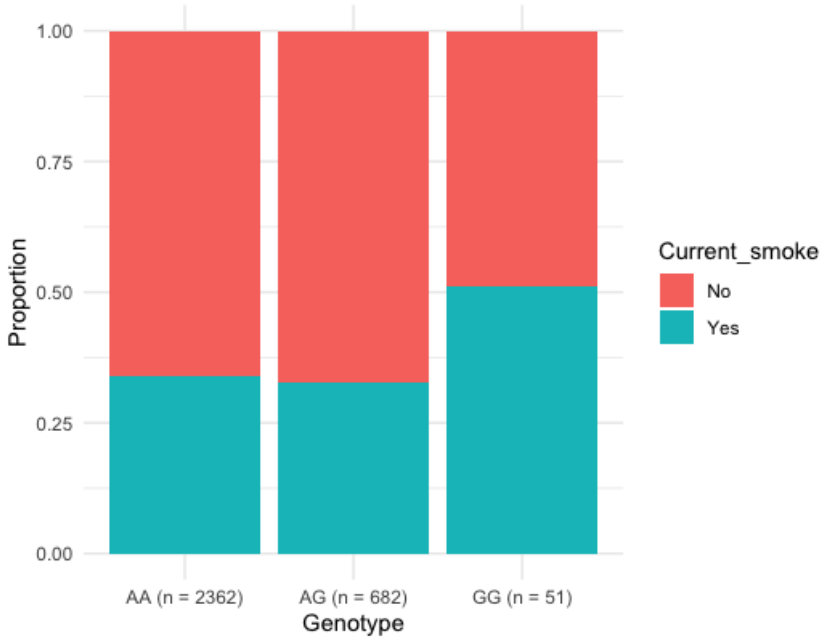

Figure 5 Proportion of smokers by genotype for rs848353

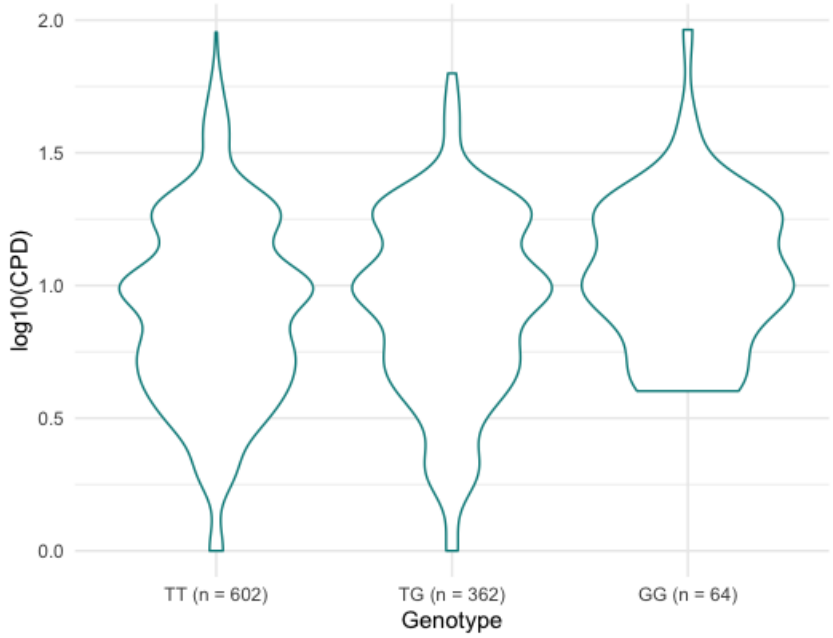

Figure 6 Distribution of CPD by genotype for rs1329650

4.0 Discussion

Smoking tobacco is a risk factor for many chronic conditions, including lung cancer, heart disease, and respiratory diseases, which are among the leading causes of death around the world. Although tobacco use has decreased in the past decades, Southeast Asia and the Pacific Islands continue to have some of the highest rates of smoking in the world, ranging from 30–50% of the population (Adia et al., 2019). In Samoa, in particular, 39.5% of men and 16.8% of women were smokers (Linhart et al., 2017). However, a majority of Samoans reported a desire to quit smoking, but nicotine dependence makes smoking cessation difficult. In my analysis, I used demographic and genotype data from the Soifua Manuia (Good Health) study to assess whether risk alleles for smoking susceptibility in other populations were associated with smoking behaviors in Samoans.

The percentages of current smokers among participants in the Soifua Manuia study (51.2% of males; 21.9% of females) were higher than those in a previous study, particularly the percentage of male smokers. Linhart and colleagues reported that 39.5% of males and 16.8% of females in Samoa were smokers. The difference between the two studies may be attributed to sample size; Linhart et al. used population-based surveys from a total of 9,223 individuals, whereas only 3,476 individuals participated in the Sofiua Manuia study.

However, the distribution of age of smoking initiation was similar to trends seen in other populations. Most individuals start smoking in their late teens and early twenties, most likely in a social setting in which their peers are also smoking (Tanielu et al., 2018). The same result was observed in this study sample; the majority of participants (71.7%) began smoking between the ages of 15 to 25.

Comparing patterns pertaining to CPD with studies in other populations is difficult because each study measured the variable differently. Many studies transformed the continuous variable into ordinal categories, and most did not indicate that they included alternate forms of tobacco use. Adia et al. reported that the mean CPD among Samoan men and women was 10.9 and 8.7, respectively, whereas the mean CPD for the Soifua Manuia study participants was 12.7 for males and 8.9 for females. The higher mean may be due to the integration of alternate forms of tobacco use (hand-rolled cigarettes, pipes, and cigars) into my analyses. Unfortunately, few studies have quantitatively analyzed smoking behaviors among Samoans; thus, it is difficult to make any overall conclusions regarding smoking rates and CPD in Samoa until additional research is conducted.

The logistic regression analyses of current smoking status (CS) and general linear model methods for CPD identified SNVs that have a significant effect on both traits. I found that homozygotes for the minor allele for SNV rs848353 [G] were 2.6 times more likely to be smokers than homozygotes for the major allele [A] (OR = 2.635 [2.433–35.631]; p = 0.001). The large confidence interval for the odds ratio is due to the small sample size of individuals that were homozygous for the minor allele (n = 51). Yoon et al. also reported an association between SNV rs848353 and increased smoking behaviors. They found that the minor allele was significantly associated with higher CPD (OR = 1.37; $p = 5.43\times10^{-6}$), which they coded as an ordinal phenotype comprising five levels. They conducted the GWAS in a Korean population and replicated their findings in European American and African American populations. Although this SNV affected different smoking phenotypes in Samoans, Koreans, European Americans, and African Americans, the minor allele was associated with an increase in the risk of smoking susceptibility overall. SNV rs848353 is located on the long arm of chromosome 7q31.1 and is an intergenic locus that has been linked to various disorders related to ulcerative colitis, cognitive performance, and

serum metabolites. However, its role in the etiology of these complex phenotypes, as well as smoking-related traits, is still unknown (Yoon et al., 2012).

The results also suggested a correlation between SNV rs3745274 in the *CYP2B6* gene and CS, although the association did not meet the threshold for significance; a larger sample size of individuals is needed to confirm this finding. The minor allele [T] had a protective effect and homozygotes were less likely to be smokers (OR = 0.710 [0.253–0.809]; $p = 0.008$). The gene's role in nicotine metabolism makes it a potential target for pharmacotherapies. Bupropion, a drug widely used to help with smoking cessation, is metabolized into hydroxybupropion by the CYP2B6 enzyme (Zanger et al., 2007). A previous clinical study reported that among individuals who were heterozygous or homozygous for SNV rs3745274 [T], bupropion produced significantly higher rates of smoking cessation than the placebo (32.5% versus 14.3% respectively; $p = 0.01$) (A. M. Lee et al., 2007). These findings have useful implications for selecting the best candidates for bupropion treatment. Moreover, the variant allele [T] has the highest frequency in Papua New Guineans (62%), so bupropion therapy could be particularly beneficial in Polynesians, who have 20-30% Papuan ancestry (Zanger et al., 2007).

I detected a significant association between SNV rs1329650 and CPD. Homozygotes for the minor allele [G] had an increase of 0.146 CPD compared to homozygotes for the major allele [T] ($\beta = 0.146 \pm 0.044$; $p = 0.001$). The variant seemed to have a dominance effect with respect to CPD; the major allele was dominant, the minor allele was recessive, and heterozygotes [TG] had similar CPD counts to dominant homozygotes [TT] ($\beta = 0.025 \pm 0.022$; $p = 0.266$). Furberg et al. conducted meta-analyses of smoking behaviors across three different consortia (Tobacco and Genetics Consortium, European Network of Genetic and Genomic Epidemiology, and Oxford-GlaxoSmithKline) and reported similar results. They coded CPD as a continuous variable and

found that “each additional copy of the risk allele [G] corresponded to an increase in smoking quantity of ~0.5 CPD” ($\beta = 0.367 \pm 0.059$; $p = 5.7\times10^{-10}$) (Furberg et al., 2010). SNV rs1329650 is located on chromosome 10q23.32 in a noncoding RNA (ncRNA), the antisense strand of the *HECTD2* gene. This region has not previously been implicated in nicotine dependence or other substance use disorders, and little is known about its function, as is the case with many noncoding regions.

Although some SNVs associated with smoking behaviors in other populations had similar effects in Samoans, additional risk loci specific to Samoans may exist. Samoans are a unique population, distinguished by their evolutionary history and relatively isolated geographic location in the middle of the Pacific Ocean. Their genetic history is reflected in their allele frequencies (AFs) for various SNVs which differed significantly from AFs in other populations. Although I included only 20 SNVs in my analysis, and these variants were not definitive markers of ancestry, the clustering of populations by AFs was clearly seen in the PCA plot (Figure 4). Overall, the frequencies of these 20 loci in Samoans were most similar to those in East Asians which is likely attributable to their ancestral and evolutionary history. I would have liked to compare the Samoan AFs to those in other Polynesian populations (e.g., Māori, Tongan, Hawaiian, etc.) to assess how all the Polynesians clustered in relation to East Asians. However, none of the genomic databases contained information on Polynesians because they have not been included in any major sequencing projects, and few GWAS studies have been conducted in these populations, highlighting the lack of diversity in genetics research.

The results of this study emphasized the extent to which smoking is a major public health concern in Samoa. Despite significant improvements in smoking rates over the past decades, the problem persists. More than half of all male participants in this study were smokers, indicating

that further intervention is necessary. By researching the social and genetic factors influencing tobacco use and cessation, we can develop more effective strategies for mitigating its prevalence, eventually leading to a decrease in chronic disease morbidity and mortality and overall improving public health in Samoa.

4.1 Limitations

The key limitations in this study were related to the variants on the genotyping chip, sample size, inaccurate reporting, and lack of comparable research. As discussed in the Background and Methods, variants in many genes known to be involved in nicotine metabolism and dependence were not assayed on the genotyping chip used for the Soifua Manuia study. Thus, I was not able to include these risk loci in my analyses and measure their effect on smoking behaviors in Samoans. Sample size was another limitation; although the overall sample size used for the association testing was adequate (n = 3072), few individuals were homozygous for the minor allele. This result is to be expected because, by definition, the minor allele is present at a lower frequency in the population. The issue of sample size was exacerbated because I had to exclude some participants from the analyses due to inaccurate reporting of CPD. Another way to adjust for the outliers and skewed distribution of CPD would have been to transform the continuous variable into ordinal categories (≤10 CPD, 11–20 CPD, etc.); this approach will be used in any future studies. As mentioned in the Methods, I could not include former smokers in my assessment of smoking status due to missing data on smoking cessation. The lack of comparable research presented another problem; few genetic studies have been conducted in Samoans, or Polynesians in general, and none of these studies analyzed smoking behaviors. Furthermore, previous studies

of smoking in other populations defined different smoking phenotypes, using measures I could not assess, such as smoking cessation and the Fagerström Test for Nicotine Dependence (FTND). These studies also did not include alternate forms of tobacco use. In converting these tobacco products into cigarette equivalents, I made the assumption that nicotine was absorbed in the same way. However, nicotine absorption is much higher when smoking cigarettes because the smoke is inhaled into the lungs. Overall, these limitations highlighted the need for additional research to replicate my findings and confirm whether the identified loci are truly associated with increased risk of smoking in Samoans.

4.2 Future Research

A GWAS of smoking behaviors in Samoans should be conducted to follow-up the results from this study; imputed genotypes should be used so that all of the known genetic variants associated with nicotine metabolism and dependence can be analyzed. Identification of additional risk loci may enable development of targeted interventions for smoking cessation. After further investigation into current cessation therapies used in Samoa, we can determine whether commonly used drugs are effective in this population and develop new treatments using a pharmacogenetics-based approach.

This study highlighted many gaps in our current understanding of noncoding regions of the genome and their implication in complex phenotypes, such as nicotine dependence. Both of the SNVs found to be significantly associated with smoking behaviors in this study were located in noncoding regions, and extensive research is needed in order to gain a thorough understanding of their function. This study also emphasized the lack of research being conducted in diverse

populations. Unfortunately, the scientific community is neglecting important parts of human evolutionary history; we need to include individuals from all ethnic groups if we want to decipher the puzzle that is the human genome and develop interventions to improve population health.

Appendix

Appendix Table 1 Populations in the 1000 Genomes Project

Superpopulation	Population description
East Asian	Han Chinese in Beijing, China Japanese in Tokyo, Japan Southern Han Chinese Chinese Dai in Xishuangbanna, China Kinh in Ho Chi Minh City, Vietnam
Admixed American	Mexican Ancestry in Los Angeles, California Puerto Ricans in Puerto Rico Colombians in Medellin, Colombia Peruvians in Lima, Peru
South Asian	Gujarati Indians in Houston, Texas Punjabis in Lahore, Pakistan Bengalis in Bangladesh Sri Lankan Tamils in UK Indian Telugus in UK
European	Northern and Western European Ancestry in Utah (CEPH) Tuscans in Italy Finnish in Finland British in England and Scotland Iberian Population in Spain
African	Yoruba in Ibadan, Nigeria Luhya in Webuye, Kenya Gambians in Western Divisions in Gambia Mende in Sierra Leone Esan in Nigeria African Ancestry in SW USA African Caribbeans in Barbados

("Which populations are part of your study?,")

Appendix Table 2 Results of association testing (additive model)

SNV	A1	A2 (minor)	CS		CPD	
			OR [95% CI]	*p*-value	β ± SE	*p*-value
rs6718569	T	C	0.945 [0.675–1.317]	0.739	0.005 ± 0.020	0.795
rs2349433	G	A	0.911 [0.702–1.180]	0.480	-0.010 ± 0.015	0.505
rs848353	A	G	1.270 [0.867–1.854]	0.218	-0.006 ± 0.021	0.786
rs1404697	G	C	1.111 [0.741–1.659]	0.609	-0.008 ± 0.023	0.732
rs4424567	A	G	1.026 [0.867–1.854]	0.840	0.014 ± 0.015	0.328
rs1329650	T	G	0.936 [0.696–1.258]	0.663	0.049 ± 0.017	*0.004

rs1013442	A	T	0.880 [0.680–1.137]	0.327	-0.001 ± 0.015	0.927
rs4923460	T	G	1.051 [0.814–1.358]	0.702	0.006 ± 0.015	0.717
rs879048	C	A	1.061 [0.821–1.371]	0.649	0.002 ± 0.015	0.889
rs6265	T	C	1.053 [0.814–1.361]	0.695	0.005 ± 0.015	0.752
rs9788682	G	A	0.660 [0.449–0.965]	0.033	0.025 ± 0.023	0.277
rs7163730	A	G	0.832 [0.613–1.127]	0.237	0.025 ± 0.018	0.158
rs6495308	C	T	1.215 [0.944–1.565]	0.130	-0.029 ± 0.015	0.052
rs16969948	A	G	0.899 [0.638–1.261]	0.537	-0.008 ± 0.020	0.687
rs7166158	A	T	1.092 [0.806–1.477]	0.569	0.008 ± 0.017	0.658
rs8043123	T	C	0.993 [0.766–1.286]	0.956	-0.030 ± 0.015	0.048
rs11072793	G	A	1.018 [0.785–1.319]	0.893	-0.034 ± 0.015	0.022
rs758642	G	A	0.972 [0.676–1.393]	0.877	-0.012 ± 0.021	0.562
rs3733829	G	A	1.072 [0.825–1.392]	0.604	0.014 ± 0.015	0.371
rs3745274	G	T	0.696 [0.532–0.911]	*0.008	0.020 ± 0.016	0.211

***Suggestive association ($p \leq 0.01$)**

Appendix Table 3 Results of association testing (by genotype)

SNV	Genotype	CS		CPD	
		OR [95% CI]	*p*-value	β ± SE	*p*-value
rs6718569	TT	Reference (1.000)		Reference (0.000)	
	TC	1.013 [0.692–1.529]	0.886	0.013 ± 0.023	0.561
	CC	0.824 [0.211–1.833]	0.416	-0.028 ± 0.067	0.679
rs2349433	GG				
	GA	0.994 [0.660–1.476]	0.946	-0.016 ± 0.023	0.483
	AA	0.909 [0.469–1.369]	0.422	-0.018 ± 0.032	0.578
rs848353	AA				
	AG	0.960 [0.586–1.407]	0.673	-0.009 ± 0.026	0.725
	GG	2.635 [2.433–35.631]	**0.001	0.000 ± 0.067	0.997
rs1404697	GG				
	GC	0.969 [0.599–1.437]	0.744	-0.058 ± 0.030	0.333
	CC	2.301 [1.135–38.965]	0.032	0.109 ± 0.094	0.250
rs4424567	AA				
	AG	1.030 [0.698–1.649]	0.753	-0.014 ± 0.025	0.575
	GG	1.021 [0.636–1.734]	0.849	0.031 ± 0.029	0.290
rs1329650	TT				
	TG	0.949 [0.604–1.297]	0.534	0.025 ± 0.022	0.266
	GG	0.992 [0.451–2.098]	0.963	0.146 ± 0.044	**0.001
rs1013442	AA				
	AT	1.082 [0.783–1.839]	0.406	0.020 ± 0.025	0.414
	TT	0.879 [0.441–1.251]	0.265	-0.006 ± 0.031	0.835
rs4923460	TT				

	TG	1.162 [0.934–2.140]	0.103	0.029 ± 0.024	0.238
	GG	1.012 [0.608–1.732]	0.919	0.005 ± 0.031	0.864
	CC				
rs879048	CA	1.078 [0.794–1.782]	0.402	0.020 ± 0.024	0.401
	AA	1.034 [0.637–1.827]	0.772	-0.002 ± 0.031	0.947
	TT				
rs6265	TC	1.190 [0.993–2.251]	0.054	0.028 ± 0.024	0.237
	CC	0.997 [0.583–1.683]	0.977	0.002 ± 0.031	0.949
	GG				
rs9788682	GA	0.848 [0.446–1.042]	0.078	0.027 ± 0.025	0.284
	AA	0.624 [0.064–1.524]	0.176	0.033 ± 0.098	0.740
	AA				
rs7163730	AG	0.904 [0.540–1.163]	0.237	0.016 ± 0.023	0.477
	GG	0.896 [0.336–1.750]	0.547	0.073 ± 0.049	0.136
	CC				
rs6495308	CT	1.200 [0.981–2.363]	0.062	-0.040 ± 0.026	0.125
	TT	1.183 [0.886–2.452]	0.135	-0.058 ± 0.030	0.051
	AA				
rs16969948	AG	0.927 [0.555–1.264]	0.402	-0.017 ± 0.024	0.472
	GG	1.011 [0.353–2.862]	0.964	0.015 ± 0.061	0.801
	AA				
rs7166158	AT	0.973 [0.639–1.378]	0.748	0.006 ± 0.023	0.784
	TT	1.255 [0.747–3.757]	0.203	0.018 ± 0.045	0.684
	TT				
rs8043123	TC	1.014 [0.692–1.546]	0.872	-0.022 ± 0.023	0.355
	CC	0.986 [0.566–1.652]	0.909	-0.064 ± 0.031	0.042
	GG				
rs11072793	GA	0.937 [0.580–1.275]	0.452	-0.024 ± 0.023	0.288
	AA	1.058 [0.661–1.950]	0.640	-0.074 ± 0.031	0.019
	GG				
rs758642	GA	0.963 [0.602–1.391]	0.683	-0.015 ± 0.024	0.553
	AA	1.103 [0.351–4.243]	0.721	-0.014 ± 0.075	0.855
	GG				
rs3733829	GA	0.952 [0.591–1.349]	0.589	0.038 ± 0.024	0.118
	AA	1.088 [0.715–2.059]	0.471	0.021 ± 0.031	0.502
	GG				
rs3745274	GT	0.887 [0.514–1.119]	0.163	0.023 ± 0.023	0.313
	TT	0.710 [0.253–0.809]	*0.008	0.037 ± 0.034	0.281

***Suggestive association ($p \leq 0.01$)**

****Below significance threshold ($p \leq 0.003$)**

Bibliography

Adia, A. C., Hawley, N. L., Naseri, T., Reupena, M. S., & McGarvey, S. T. (2019). Tobacco smoking patterns in Samoa in 2010: Implications for interventions. *Tob Prev Cessat, 5*, 50. doi:10.18332/tpc/114093

Bilano, V., Gilmour, S., Moffiet, T., d'Espaignet, E. T., Stevens, G. A., Commar, A., . . . Shibuya, K. (2015). Global trends and projections for tobacco use, 1990–2025: an analysis of smoking indicators from the WHO Comprehensive Information Systems for Tobacco Control. *The Lancet, 385*(9972), 966-976. doi:https://doi.org/10.1016/S0140-6736(15)60264-1

Caporaso, N., Gu, F., Chatterjee, N., Sheng-Chih, J., Yu, K., Yeager, M., . . . Bergen, A. W. (2009). Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS One, 4*(2), e4653. doi:10.1371/journal.pone.0004653

Claw, K. G., Beans, J. A., Lee, S. B., Avey, J. P., Stapleton, P. A., Scherer, S. E., . . . Robinson, R. F. (2020). Pharmacogenomics of Nicotine Metabolism: Novel CYP2A6 and CYP2B6 Genetic Variation Patterns in Alaska Native and American Indian Populations. *Nicotine Tob Res, 22*(6), 910-918. doi:10.1093/ntr/ntz105

Commar, A., Prasad, V., d'Espaignet, E. T., & Wolfenden, L. (2018). *WHO global report on trends in prevalence of tobacco smoking 2000-2025*. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/272694/9789241514170-eng.pdf?ua=1

Conlon, M. S., & Bewick, M. A. (2011). Single Nucleotide Polymorphisms in CHRNA5 rs16969968, CHRNA3 rs578776, and LOC123688 rs8034191 Are Associated With Heaviness of Smoking in Women in Northeastern Ontario, Canada. *Nicotine & Tobacco Research, 13*(11), 1076-1083. doi:10.1093/ntr/ntr140

Essenmacher. (2012). Nicotine Content in Tobacco Products. Retrieved from https://sntc.medicine.ufl.edu/Content/Webinars/SupportingDocs/3031-Essenmacher_-_Handout_1.pdf

Furberg, H., Kim, Y., Dackor, J., Boerwinkle, E., Franceschini, N., Ardissino, D., . . . Genetics, C. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics, 42*(5), 441-447. doi:10.1038/ng.571

Gorwood, P., Le Strat, Y., & Ramoz, N. (2017). Genetics of addictive behavior: the example of nicotine dependence. *Dialogues Clin Neurosci, 19*(3), 237-245. doi:10.31887/DCNS.2017.19.3/pgorwood

Harris, D. N., Kessler, M. D., Shetty, A. C., Weeks, D. E., Minster, R. L., Browning, S., . . . O'Connor, T. D. (2020). Evolutionary history of modern Samoans. *Proc Natl Acad Sci U S A, 117*(17), 9458-9465. doi:10.1073/pnas.1913157117

Jha, P., & Peto, R. (2014). Global Effects of Smoking, of Quitting, and of Taxing Tobacco. *New England Journal of Medicine, 370*(1), 60-68. doi:10.1056/NEJMra1308383

Lee, A. M., Jepson, C., Hoffmann, E., Epstein, L., Hawk, L. W., Lerman, C., & Tyndale, R. F. (2007). CYP2B6 genotype alters abstinence rates in a bupropion smoking cessation trial. *Biol Psychiatry, 62*(6), 635-641. doi:10.1016/j.biopsych.2006.10.005

Lee, S.-H., Ahn, W.-Y., Seweryn, M., & Sadee, W. (2018). Combined genetic influence of the nicotinic receptor gene cluster CHRNA5/A3/B4 on nicotine dependence. *BMC Genomics, 19*(1), 826. doi:10.1186/s12864-018-5219-3

Li, M. D., Yoon, D., Lee, J. Y., Han, B. G., Niu, T., Payne, T. J., . . . Park, T. (2010). Associations of variants in CHRNA5/A3/B4 gene cluster with smoking behaviors in a Korean population. *PLoS One, 5*(8), e12183. doi:10.1371/journal.pone.0012183

Linhart, C., Naseri, T., Lin, S., Taylor, R., Morrell, S., McGarvey, S. T., . . . Zimmet, P. (2017). Tobacco smoking trends in Samoa over four decades: can continued globalization rectify that which it has wrought? *Globalization and Health, 13*(1), 31. doi:10.1186/s12992-017-0256-2

Liu, J. Z., Tozzi, F., Waterworth, D. M., Pillai, S. G., Muglia, P., Middleton, L., . . . Marchini, J. (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet, 42*(5), 436-440. doi:10.1038/ng.572

Mackillop, J., Obasi, E., Amlung, M. T., McGeary, J. E., & Knopik, V. S. (2010). The Role of Genetics in Nicotine Dependence: Mapping the Pathways from Genome to Syndrome. *Curr Cardiovasc Risk Rep, 4*(6), 446-453. doi:10.1007/s12170-010-0132-6

Matoba, N., Akiyama, M., Ishigaki, K., Kanai, M., Takahashi, A., Momozawa, Y., . . . Kamatani, Y. (2019). GWAS of smoking behaviour in 165,436 Japanese people reveals seven new loci and shared genetic architecture. *Nature Human Behaviour, 3*(5), 471-477. doi:10.1038/s41562-019-0557-y

Minster, R. L., Hawley, N. L., Su, C. T., Sun, G., Kershaw, E. E., Cheng, H., . . . McGarvey, S. T. (2016). A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nat Genet, 48*(9), 1049-1054. doi:10.1038/ng.3620

Mishra, S. I., Osann, K., & Luce, P. H. (2005). Prevalence and predictors of smoking behavior among Samoans in three geographical regions. *Ethn Dis, 15*(2), 305-315.

Stanaway, J. D., Afshin, A., Gakidou, E., Lim, S. S., Abate, D., Abate, K. H., . . . Murray, C. J. L. (2018). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet, 392*(10159), 1923-1994. doi:https://doi.org/10.1016/S0140-6736(18)32225-6

Tanielu, H., McCool, J., Umali, E., & Whittaker, R. (2018). Samoan Smokers Talk About Smoking and Quitting: A Focus Group Study. *Nicotine Tob Res, 20*(9), 1132-1137. doi:10.1093/ntr/ntx152

Which populations are part of your study? (2021). Retrieved from https://www.internationalgenome.org/faq/which-populations-are-part-your-study/

Yoon, D., Kim, Y. J., Cui, W. Y., Van der Vaart, A., Cho, Y. S., Lee, J. Y., . . . Park, T. (2012). Large-scale genome-wide association study of Asian population reveals genetic factors in FRMD4A and other loci influencing smoking initiation and nicotine dependence. *Hum Genet, 131*(6), 1009-1021. doi:10.1007/s00439-011-1102-x

Zanger, U. M., Klein, K., Saussele, T., Blievernicht, J., Hofmann, M. H., & Schwab, M. (2007). Polymorphic CYP2B6: molecular mechanisms and emerging clinical significance. *Pharmacogenomics, 8*(7), 743-759. doi:10.2217/14622416.8.7.743