Polygenic Risk Score Analysis of E-Cigarette Smoking:
Genetic Association of E-cigarette and Traditional Cigarette Behaviors

Yueh Yueh Ling

Institute of Behavioral Genetics, Department of Molecular, Cellular, Developmental Biology

University of Colorado at Boulder

Defense Date: April 10th, 2017

Committee:

Advisor: Dr. Scott Vrieze, Department of Psychology and Neuroscience, Institute of Behavioral Genetics

Dr. Soo Rhee, Department of Psychology and Neuroscience, Institute of Behavioral Genetics

Dr. Brian DeDecker, Department of Molecular, Cellular, and Developmental Biology

Abstract

Background and Aims

Although the genetic etiology of traditional cigarette smoking is well studied, the genetic etiology of e-cigarette smoking is not fully elucidated. This study utilizes polygenic risk scores to examine the potential genetic overlap of traditional and e-cigarette behaviors. Each score reflects the combined effect of selected risk alleles for smoking.

Method

A polygenic risk score analysis was calculated using the Genes for Good (*n*=258) sample and cigarettes per day (CPD) regression weight from UK Biobank (*n*=19,357). Scores were generated at six *p*-value thresholds, with $5e^{-8}$ being the most stringent. Scores were then correlated to cigarettes per day (CPD) and frequency of e-cigarette use.

Results

One correlation between CPD and risk scores at a *p*-value threshold of 5e-7 approached significance (p~0.05). No other correlations were significant.

Conclusions

This is the first study, to the knowledge of the investigators, that examines the genetic etiology of e-cigarettes. However, due to the limited sample size in this study, significant associations were not detected. As the Genes for Good study continues to grow, further investigation conducted with sufficient power would be necessary to better elucidate the genetic etiology of e-cigarette behaviors.

Keywords: polygenic score, genetic association, e-cigarettes, traditional, substance use

Table of Contents

Introduction

Cigarette smoking continues to have a substantial cost on the United States population, both societally and personally. Currently, about 16 million Americans are living with a disease caused by smoking, and 443,000 U.S. adults die from smoking-related illnesses each year ("Current cigarette smoking among adults — United States, 2011," 2012). Smoking is also estimated to cost the U.S. about $96 billion in direct medical expenses and $97 billion in lost productivity annually ("Current Cigarette Smoking Among Adults," 2012). Fortunately, smoking rates in the US have been nearly halved since the first Surgeon General's report released in 1964, which also marked the beginning of increased tobacco control efforts (US Department of Health and Human Services, 2014). Despite increased efforts to reduce cigarette use (i.e., increasing tobacco prices, antitobacco media campaigns, and implementing smoke-free laws), the decline in smoking rates has tapered off in recent years (Jamal et al., 2016).

The introduction of Electronic Nicotine Delivery Systems in 2003, colloquially known as electronic cigarettes (e-cigarettes), has also affected the decline of cigarette smoking. According to a survey conducted in 2013, 73% of the US population is aware of e-cigarettes and its use occurs in both nondaily smokers and heavy smokers (Adkison et al., 2013). Awareness of e-cigarettes has doubled from 16.5% in 2009 to 32.2% in 2010, and rates of ever using e-cigarettes quadrupled from 2009 (0.6%) to 2010 (2.7%) (Regan, Promoff, Dube, & Arrazola, 2013). Although research on the use of e-cigarettes as a cessation aid is not extensive, there is some evidence that, when used as an aid, use of e-cigarettes can reduce the number of cigarettes smoked in current smokers and increase the success of quitting (Siegel et al., 2011). Present research shows some evidence that the short-term effects of e-cigarettes on cardiovascular and respiratory functions are exponentially less harmful than smoking traditional cigarettes (albeit still with some transient negative effects); however, the long-term effects of e-cigarettes are still unknown (Farsalinos & Polosa, 2014; Hajek, Etter, Benowitz, Eissenberg, & McRobbie, 2014).

While rates of cigarette smoking are decreasing, the rates of first time e-cigarette users continue to increase. E-cigarettes are a relatively new development, and the relationship between cigarette smoking and e-cigarette smoking are still not well delineated. However, it has become evident that cigarette smoking as a trait arises due to both environmental and genetic factors. The collection of family, adoption, and twin studies strongly support the heritability of cigarettes in tobacco initiation, maintenance of cigarette use, and nicotine dependence (Sullivan & Kendler, 1999a). For example, studies have indicated that about 60% of the variance in liability to start smoking is due to genetic factors (Maes et al., 2004). In addition to the indirect evidence found in these studies for the genetic contribution to smoking, there have also been attempts to locate the specific genes that contribute to smoking behaviors (Caporaso et al., 2009; Sullivan & Kendler, 1999b; Vink et al., 2014).

However, for cigarette smoking and other complex traits, it is important to note that there is not usually one single gene that contributes to smoking behavior. Complex traits like cigarette smoking appear to be highly polygenic, and analysis using quantitative methods like the genome-wide association study (GWAS) and the polygenic risk score (PRS) are becoming increasingly prevalent. A GWAS scans all common genetic variation to determine which variants are associated with the phenotype of interest. Recent GWAS results have been able to identify single-polynucleotide polymorphisms (SNPs) that are highly associated with smoking initiation, cigarettes smoked per day, smoking cessation, and other smoking behaviors (Furberg & Sullivan, 2010). From GWAS results, which identify single variants associated with a trait, one can construct polygenic risk scores to evaluate the extent to which hundreds or thousands of variants, aggregated together, are associated with the phenotype of interest. Both GWAS and PRS methods have been used successfully in studies of complex traits ranging from mental illnesses like schizophrenia to physical traits like height and body mass index (Dudbridge, Whittaker, Iorio, Balding, & Lange, 2013).

Despite extensive research on the genetic etiology of cigarette smoking, there has been little to no research conducted on the genetic etiology of the use of e-cigarettes. Although there are many behavioral similarities between cigarette smoking and the use of e-cigarettes, only about 12-14% of electronic cigarette smokers become daily users, indicating that satisfaction may not be as high as traditional cigarette smoking (Hajek et al., 2014). The purpose of the current study is to assess whether or not there is a common genetic etiology between traditional cigarette smoking and electronic cigarette smoking.

To do this, we constructed polygenic risk scores for smoking based on a in-house GWAS of cigarettes smoked per day among current smokers in the UK BioBank, a large study of 150,000 subjects dedicated to investigating the genetic and nongenetic determinants of disease (Sudlow et al., 2015). This GWAS identified variants in the genome that are associated with cigarettes smoked per day. The risk alleles and weights calculated from the UK BioBank GWAS were then used to generate polygenic risk scores for each participant in Genes for Good, a study unrelated to the UK Biobank that measured cigarette and e-cigarette use in thousands of participants. These scores indicated the predicted genetic cumulative effect of all the variants in each individual that contributed to cigarettes smoked per day. By correlating these scores with a measure of e-cigarette use, we were then able to compare the genetic etiology of cigarette and e-cigarette smoking.

Methods

Discovery Sample from UK Biobank

The UK Biobank is a large ongoing collection of diagnostic healthcare information from over 500,000 volunteers across the UK. The phenotype utilized in this study was cigarettes per day (CPD) (*n*=19,357) which was defined as average number of cigarettes smoked per day, either as a current or former smoker. These quantitative measures were then binned into five categories (1: 1-5 cigarettes, 2: 6-15 cigarettes, 3: 16-25 cigarettes, 4: 26-35 cigarettes, 5: 36+ cigarettes). Those who either never smoked or whom there is no available data were set to missing.

The UK Biobank is genotyped on 2 arrays, the UK BiLEVE array and the UK Biobank array. However, only those genotyped on the BiLEVE array were utilized in this study. The BiLEVE subsample consists only of heavy smokers and never smokers, but only heavy smokers were utilized in this study (Wain et al., 2015).

Genotypes were imputed to a reference panel including the UK10K and 1000 Genomes whole genome sequence datasets. Only those of European ancestry were included in the analysis. Summary statistics of the GWA meta-analysis of CPD were obtained through the GWAS and Sequencing Consortium of Alcohol and Nicotine Use (GSCAN), a large consortium utilizing the UK Biobank data.

Target Sample from Genes For Good

The target sample consisted of subjects from Genes for Good, a genetic research study whose primary recruitment method is as an application on the popular social media website Facebook. When individuals over 18 consent to the study, they are asked to fill out various health-related surveys. After the participant fills out 18 surveys, they then have the opportunity to be genotyped via a mail in saliva kit, thus contributing data to the study while also receiving personal ancestry information and their raw genotypic data. At the time of writing, over 26,026 people have participated in the study and over 6,610 DNA samples have been genotyped.

The current study utilized phenotypic data specifically from the Tobacco Use survey in Genes for Good, which consisted of 11,945 participants. However, only participants who also had genotypic information and were e-cigarette users (n=258) were included in the polygenic risk score analysis.

All DNA extraction and genotyping was conducted at the University of Michigan Sequencing Core. Genotyping was done using the Illumina Human Core Exome array, which genotypes 500,000 variants. Genotypes underwent standard quality control, were phased by software SHAPEIT (Delaneau, Marchini, & Zagury, 2011), and then imputed to the 1000 Genome phase 3 whole genome reference panel using Minimac3 (Das et al., 2016). In this sample, a genetic principal component analysis (PCA) was constructed and superimposed onto 1000 Genomes PCAs for comparison. Those of European ancestry were then selected out for analysis. Additionally, variants were removed if they were in linkage disequilibrium with a more significant variant. We considered the most significant SNP and removed all variants in linkage disequilibrium $r^2 > 0.1$ within 500 kilobases.

The e-cigarette phenotype was developed around the GFG survey question "How often do you smoke e-cigarettes?" with answer options as less than once a day, once a day, a few times per day, and all day long. The cigarettes per day (CPD) phenotype (n=1287) was a combination of two GFG survey questions involving former and current smokers. When asked how many cigarettes participants formerly or currently smoked, they were given the option of selecting 1 through 40+ in a drop-down menu.

Polygenic risk scores and statistical analysis

A polygenic risk score is a weighted sum across variants in an individual, with each weight being equal to the effect size for that variant in the UK Biobank GWAS for CPD. The primary function of a polygenic risk score analysis is to reflect the combined effect of selected risk alleles into one quantifiable score. For individual *i*, the polygenic risk score for variants *j* through *n* is more formally depicted below:

$$\mathrm{PRS_i} = \sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{UK\ Biobank}\ \beta_{\mathrm{ij}} \times (\mathrm{no.\ of\ risk\ alleles_{ij}})$$

The coding of the risk alleles and the beta were generated with the in-house UK Biobank GWAS for CPD. After computing the polygenic risk score, we tested for an association between the score and the two constructed phenotypes (i.e., frequency of e-cigarette use and CPD) from Genes for Good. We first considered the polygenic risk scores of all of the genome-wide-significant variants that had a p value of 5e^{-8} (number of SNPs = 22) in the UK Biobank GWAS. This threshold was gradually made less stringent, and the polygenic risk scores for the following thresholds were also considered: 5e^{-7} (n SNPs = 26), 5e^{-6} (n SNPs = 51), 5e^{-5} (n SNPs=294), 5e^{-4} (n SNPs=2,054), and 5e^{-3} (n SNPs=17,864) with a total of 20,311 SNPs analyzed. An association between a polygenic risk score and an outcome variable was considered significant if $P < 0.05$.

Results

Descriptive statistics for the Genes for Good sample are displayed below in Table 1. Most participants in the Tobacco Use survey were primarily from 22 to 30 years old with a large majority being female (Table 1). This mimics the demographic of Facebook users overall. Figure 2 below displays a histogram for the cigarettes per day measure in Genes for Good (M=12.53, Mdn=10, SD=7.34) (Figure 2). Most participants smoked around 10 cigarettes per day or 20 cigarettes per day. For the e-cigarette measure in Genes for Good, most participants reported that they smoked less than once a day and were primarily females (Table 2).

Figure 1 displays a summary of the traditional cigarette and e-cigarette usage in the Tobacco Use survey. Non-smokers make up the largest percentage of this survey, with about 27% of the sample being solely traditional cigarette smokers. There are more individuals that smoke both cigarettes and e-cigarettes than those who smoke e-cigarettes alone.

Lastly, Figure 3 displays the results from the polygenic risk score analysis. The figure shows the correlation between the phenotypes and the risk score at each p-value threshold. The only polygenic risk score association approaching significance was the correlation of CPD with the risk score at the p-value threshold of 5e-^{7} (p ~ 0.05) with r=0.11. No correlations at any other threshold approached significance.

Discussion

Although e-cigarette use is becoming more popular, from our sample it is evident that a large portion of the population is still solely smoking traditional cigarettes. Despite this fact, e-cigarette rates of use continue to increase due to its role as a safer and healthier alternative to traditional cigarette smoking (Regan et al., 2013). This recent development raises questions about the genetic etiology of e-cigarette use and whether or not it is shared with traditional cigarette use.

The aim of this study was to investigate the genetic overlap between e-cigarette and traditional smoking behaviors. By using the association results of cigarettes per day from the UK Biobank GWAS, we were able to calculate a polygenic risk score for each individual in the Genes for Good sample. These scores were then correlated with two different phenotypes, cigarettes per day and frequency of e-cigarette use. From the results of this study there is no evidence to suggest that cigarette use and e-cigarette use share the same genetic etiology, with small to negative correlations between the frequency of e-cigarette use and risk scores.

However, these results must be viewed in light of the study's limitations. Firstly, the study lacked power. In order to detect a correlation of 0.1 with a standard beta of 0.2, the study would have required a sample size of at least 646 participants, at least twice the amount in the current study, in order to have 80% power to detect a polygenic correlation. Furthermore, the polygenic risk score correlation between CPD of Genes for Good and CPD of UK Biobank was conducted as a baseline for comparison to the actual phenotype of interest, frequency of e-cigarette use. However, even CPD was not statistically significantly predicted by the CPD-based polygenic risk score generated from the UK Biobank. Only one PRS approached significance at a PRS p threshold of 5e^{-7} with r=0.11. This indicates that the power in this study is not sufficient to confidently conclude that e-cigarette and cigarette behaviors do not have similar genetic etiologies.

Another contribution to the lack of power could potentially be the size of the discovery sample. In a polygenic risk score, power increases with the size of the discovery sample (Vink et al., 2014). The subset of data used from UK Biobank only had a sample size of around 20,000 smokers. This is considered small in many studies of quantitative genetics, with some samples reaching up to 70,000 participants before power is sufficient to detect significance. Additionally, the discovery sample in this study consisted only of heavy smokers, reducing chances of detecting a significant correlation due to a decreased range of behaviors (Wain et al., 2015).

The lack of correlation may also be due to the nature of the Genes for Good study. Because of its mode of delivery through the popular Facebook App, it is highly available and easy to access. This availability allows it to reach more participants; however, it also provides an opportunity for users to answer haphazardly, incentivized by the promised complimentary ancestry analysis and raw genotypes after completing the necessary number of surveys. This effect may be magnified due to the self-report nature and perceived anonymity of the application, since participants do not have to meet in-person to fill out information in the surveys.

Although the conclusion of the current study is limited, this is the first study, to the knowledge of the investigators, to examine the genetic etiology of e-cigarette behaviors, as well as examine the genetic overlap between e-cigarette use and traditional cigarette use. We expect that the sample size of Genes for Good will grow with time, as it is an ongoing study. Further investigation conducted with sufficient power would be necessary to better elucidate the genetic etiology of e-cigarette behaviors.

Literature Cited

Adkison, S. E., O'Connor, R. J., Bansal-Travers, M., Hyland, A., Borland, R., Yong, H.-H., … Fong, G. T. (2013). Electronic Nicotine Delivery Systems: International Tobacco Control Four-Country Survey. *American Journal of Preventive Medicine*, *44*(3), 207–215. https://doi.org/10.1016/j.amepre.2012.10.018

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. https://doi.org/10.1371/journal.pone.0004653

Current Cigarette Smoking Among Adults. (2012). *Centers for Disease Control and Prevention*, *61*(44), 889–894.

Current cigarette smoking among adults — United States, 2011. (2012). *MMWR. Morbidity and Mortality Weekly Report*.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Delaneau, O., Marchini, J., & Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. *Nature Methods*, *9*(2), 179–181. https://doi.org/10.1038/nmeth.1785

Dudbridge, F., Whittaker, J., Iorio, M. De, Balding, D., & Lange, K. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, *9*(3), e1003348. https://doi.org/10.1371/journal.pgen.1003348

Farsalinos, K. E., & Polosa, R. (2014). Safety evaluation and risk assessment of electronic cigarettes as tobacco cigarette substitutes: a systematic review. *Therapeutic Advances in Drug Safety*, *5*(2), 67–86. https://doi.org/10.1177/2042098614524430

Furberg, H., & Sullivan, P. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*, (42), 441–447. https://doi.org/10.1038/ng.571

Hajek, P., Etter, J.-F., Benowitz, N., Eissenberg, T., & McRobbie, H. (2014). Electronic cigarettes: review of use, content, safety, effects on smokers and potential for harm and benefit. *Addiction*, *109*(11), 1801–1810. https://doi.org/10.1111/add.12659

Jamal, A., King, B. A., Neff, L. J., Whitmill, J., Babb, S. D., & Graffunder, C. M. (2016). Current Cigarette Smoking Among Adults — United States, 2005–2015. *MMWR. Morbidity and Mortality Weekly Report*, *65*(44), 1205–1211. https://doi.org/10.15585/mmwr.mm6544a2

Maes, H., Sullivan, P., Bulik, C., Neale, M., Prescott, C., Eaves, L., & Kendler, K. (2004). A twin study of genetic and environmental influences on tobacco initiation, regular tobacco use and nicotine dependence. *Psychological Medicine*, (34), 1251–1261. https://doi.org/10.1017/S0033291704002405

Regan, A. K., Promoff, G., Dube, S. R., & Arrazola, R. (2013). Electronic nicotine delivery systems: adult use and awareness of the "e-cigarette" in the USA. *Tobacco Control*, *22*(1), 19–23. https://doi.org/10.1136/tobaccocontrol-2011-050044

Siegel, M., Tanwar, K., & Wood, K. (2011). Electronic Cigarettes As a Smoking-Cessation Tool : Results from an Online Survey. *American Journal of Preventive Medicine*, *40*(4), 472–475. https://doi.org/10.1016/j.amepre.2010.12.006

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., … Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Sullivan, P., & Kendler, K. (1999a). The genetic epidemiology of smoking. *Nicotine & Tobacco Research*, *1*(2), S51–S57. https://doi.org/10.1080/14622299050011811

Sullivan, P., & Kendler, K. (1999b). The genetic epidemiology of smoking. *Nicotine & Tobacco Research*, *1*(1), 51–57. https://doi.org/10.1080/14622299050011811

US Department of Health and Human Services. (2014). *2014 Surgeon General's Report: The Health Consequences of Smoking--50 Years of Progress* (Surgeon General Reports).

Vink, J. M., Hottenga, J. J., de Geus, E. J. C., Willemsen, G., Neale, M. C., Furberg, H., & Boomsma, D. I. (2014). Polygenic risk scores for smoking: predictors for alcohol and cannabis use? *Addiction*, *109*(7), 1141–1151. https://doi.org/10.1111/add.12491

Wain, L. V, Shrine, N., Miller, S., Jackson, V. E., Ntalla, I., Artigas, M. S., … Hall, I. P. (2015). Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet Respiratory Medicine*, *3*(10), 769–781. https://doi.org/10.1016/S2213-2600(15)00283-0

Acknowledgements

I would like to thank my primary advisor, Dr. Scott Vrieze, for his guidance during this project. His expertise, his patience, and the energy he has put into my learning both encouraged and inspired me. Additionally, I would like to thank my committee for being accommodating and supportive along the way.

I would also like to thank everyone else in the lab for their infinite patience and investment into my humble project. Specifically, I would like to thank Mengzhen Liu for being a human being and teaching me so much through her patient explanations, specifically about emulsions, legless lizards, and indoor skiing. Her endless supply of gummy bears and countless hours of brainstorming and problem solving with me made this project possible. I would also like to thank Hannah Young for being my non-competitive lab mate and supporting me throughout the months. Her work ethic and drive for learning inspired me throughout the entire process.

All of these people have showed me what passion in the field of science looks like and were essential in the path to helping me develop my own. They pushed me to step outside of my comfort zone and were constantly challenging me. I will always be grateful for their impact on my journey.

Lastly, I would like to thank my family for always loving me and planting seeds of passion for education early in my life. I would not be here without their support.

The funding sources for this project are the University of Michigan Genomics Initiative and NIDA award DA037904. In addition, thank you to the Biological Scholars Initiative and its staff for their support and encouragement in providing the funding to do my independent project.

Thank you!

Tables and Figures

Table 1

Age Range in Genes for Good Target Sample (n=11,303)

Age Range	N (%)	Male	Female	% Male
18 – 21	680 (6.01%)	215	465	32.0%
22 – 30	3708 (32.81%)	1177	2530	31.8%
31 – 40	2680 (23.71%)	742	1938	27.7%
41 – 50	1708 (15.11%)	405	1303	23.7%
51 – 60	1416 (12.53%)	356	1059	25.2%
61 – 70	882 (7.80%)	251	631	28.5%
71 – 99	229 (2.03%)	72	157	31.4%
Total	11,303	3,218	8,083	28.5%

Table 2

Distribution of e-cigarette variable in the Genes for Good (GFG) Target Sample (n=238)

Categories	N (%)	Male	Female	% Male
Less than once a day	114 (47.9%)	49	65	42.9%
Once a day	6 (2.52%)	3	3	50.0%
A few times per day	67 (28.15%)	30	37	44.8%
All day long	51 (21.43%)	23	28	45.1%
Total	238	105	133	44.1%

Figure 1

Summary of Cigarette and E-Cigarette Behaviors

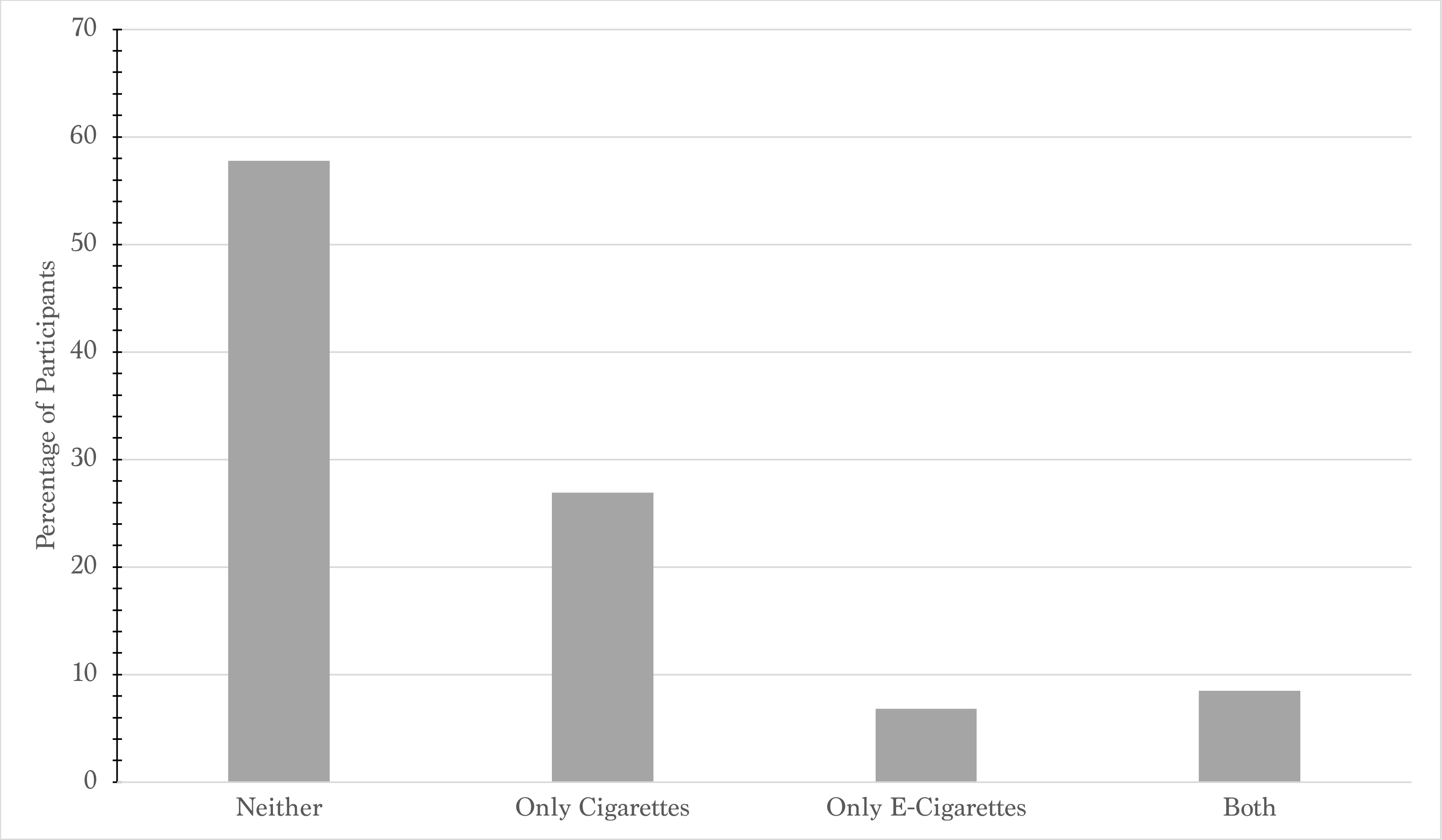


Figure 2

Histogram of CPD in Former and Current Smokers in Genes for Good Sample (n=1287)

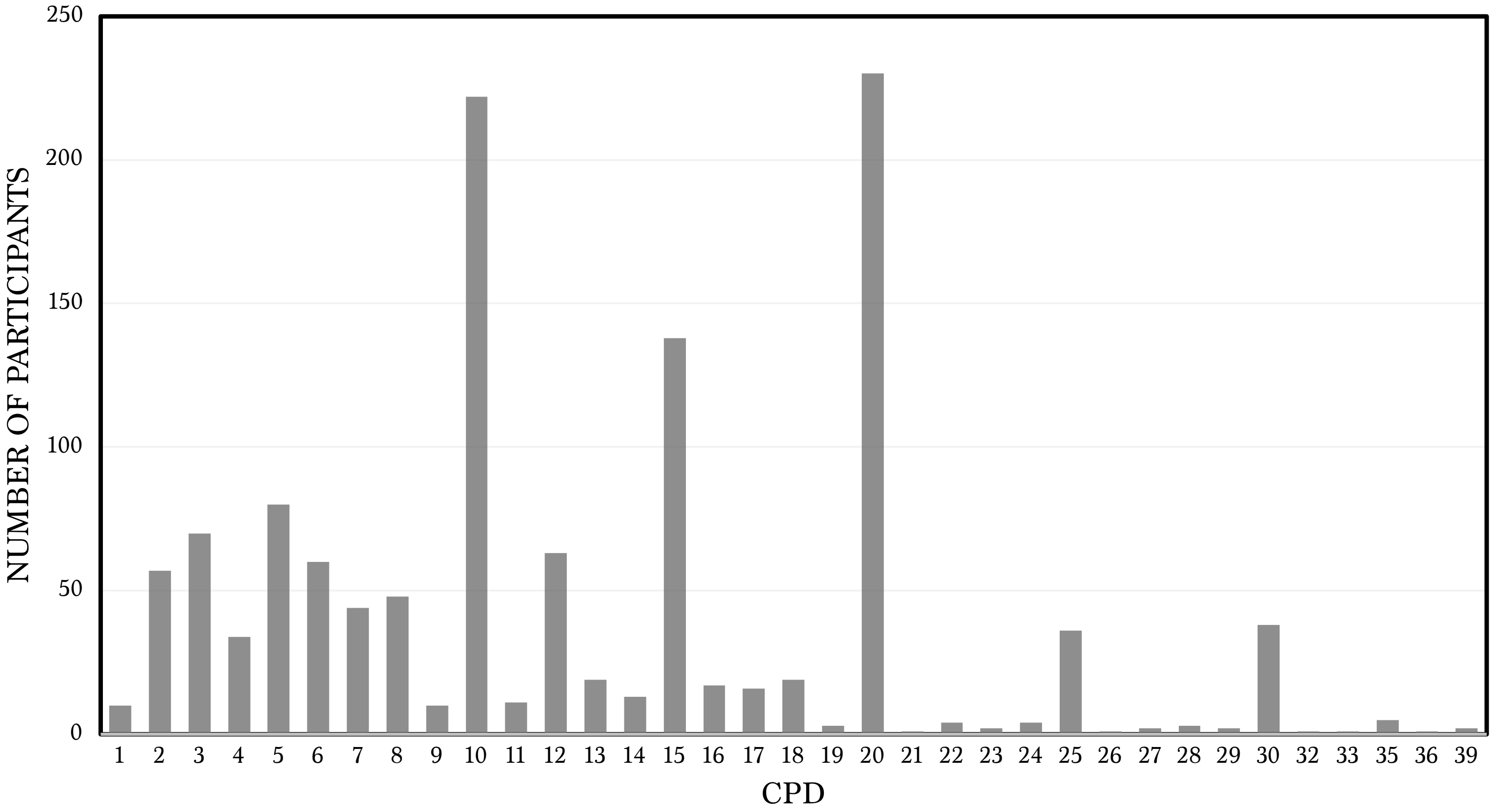
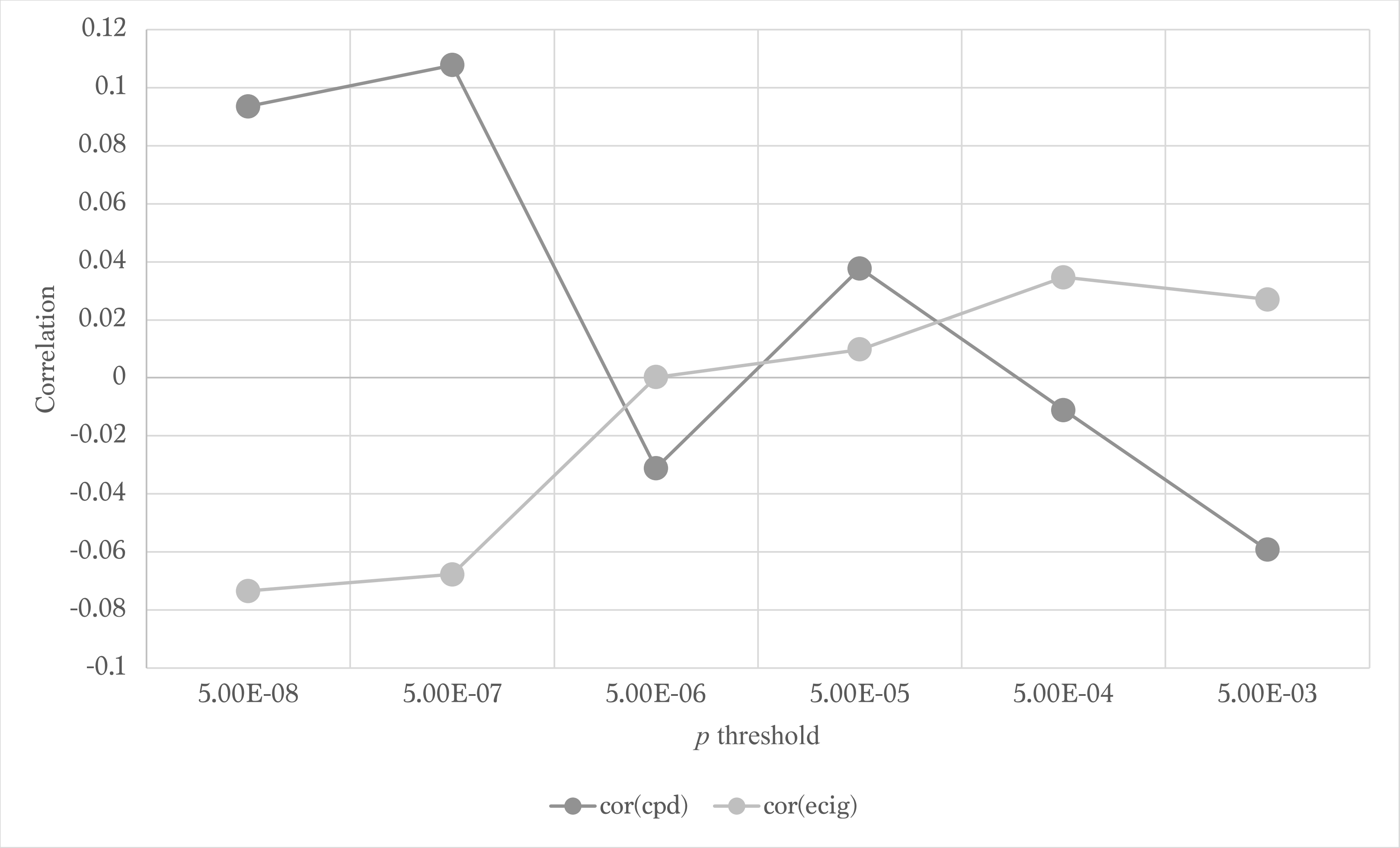


Figure 3

Correlations of PRS and E-Cigarette Frequency and CPD in Genes for Good



cor(cpd) = correlation of cigarettes per day (CPD) and weighted genetic score

cor(ecig) = correlation of frequency of e-cigarette (e-cig) use and weighted genetic score