Loci on Chromosome 20 Interact with rs16969968 to Influence Cigarettes per Day in European Ancestry Individuals

Marissa Ehringer (✉ marissa.ehringer@colorado.edu)
University of Colorado

Pamela Romero Villela
University of Colorado

Teemu Palviainen
Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki https://orcid.org/0000-0002-7847-8384

Luke Evans
University of Colorado Boulder https://orcid.org/0000-0002-7458-1720

Richard Border

Jaakko Kaprio
University of Helsinki https://orcid.org/0000-0002-3716-2455

Rohan Palmer
Emory University https://orcid.org/0000-0002-6809-9962

Matthew Keller
University of Colorado https://orcid.org/0000-0002-6075-9882

Article

Keywords:

Posted Date: August 30th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1981574/v1

Abstract

Our understanding of the molecular genetic contributions to smoking are limited to the additive effects of individual single nucleotide polymorphisms (SNPs), but the underlying genetic risk is likely to also include dominance, epistatic, and gene-environment interactions. To begin to address this complexity, this study attempted to identify potential genetic interactions between rs16969968, the most replicated SNP associated with smoking quantity, and all SNPs and genes across the genome. Using the UK Biobank, we found one gene, *PCNA*, that showed a genome-wide significant interaction with rs16969968 for smoking behaviors in a sample of 116 442 smokers of European ancestry. We replicated this finding in a meta-analysis of five Finnish samples (n = 40 140): FinHealth, FINRISK, Finnish Twin Cohort, GeneRISK, and Health-2000-2011. To our knowledge, this represents the first reliable epistatic effect between measured genetic variants for smoking behaviors and provides a novel direction for possible future functional studies related to this interaction. Furthermore, this work demonstrates the feasibility of these analyses, which may be applied to other top SNPs for smoking and/or other phenotypes.

Introduction

Smoking cigarettes is the leading cause of preventable death in the United States [1]. In fact, one in five deaths in the United States can be attributed to smoking [1]. Smoking also burdens the economy; smoking-related health costs are around $300 billion per year in the United States alone [2]. While 68% of smokers report wanting to quit [3], only 8% of these successfully do so [1], reflecting the tremendous addictive potential of nicotine, one of the most addictive psychoactive drugs [4]. Previous work has demonstrated a substantial genetic component to smoking behaviors, and twin studies estimate the heritability of smoking quantity and nicotine dependence to be between 40% and 75% [5, 6] in adults across multiple ancestries. Recent genome-wide association studies (GWAS) have identified several hundred individual variants associated with various smoking-related behaviors [7], but these loci explain only 4–8% of the estimated heritability and demonstrate the polygenic nature of these traits [8, 9]. A variety of factors are believed to contribute to the mismatch between twin and genome-wide SNP-heritability estimates, most notably, the influence of functional variants that are poorly tagged by SNPs on modern arrays. Moreover, the variants within these genes and their regulatory elements are likely to influence a complex trait via a minute perturbation across a complex, non-linear set of physiological networks (i.e., transcriptional, neuronal, and developmental) [10]. The physiological intricacy in which complex traits, such as smoking, develop suggests that interactions between loci or whole genes (i.e., epistasis) are likely, since there are numerous ways and stages at which these interactions could arise. Furthermore, while current evidence of epistatic effects in humans has been limited, work on model organisms further suggests that epistatic effects are common [11] and may be particularly important for predicting an individual's genetic risk to disease such as nicotine dependence [12].

SNP rs16969968 in the *CHRNA5/A3/B4* gene cluster of neuronal nicotinic receptor genes is the most widely replicated genetic variant associated with smoking behaviors [13–15], emerging from early GWAS studies of lung cancer and smoking behaviors [16–18]. Nicotine is an agonist for neuronal nicotinic acetylcholine receptors (*CHRN* genes) and repeated nicotine use leads to their upregulation [19]. rs16969968 was the original top SNP identified in the *CHRNA5* gene and has been the major focus of further study because it changes an amino acid (aspartate to asparagine; D398N) and has been shown to confer functional effects using cell culture methods *in vitro* [20] and behavioral effects in a mouse genetic model [21–23]. However, careful investigation of statistically independent SNPs within the *CHRNA5/A3/B4* gene cluster revealed high complexity of the underlying genetic structure. In a meta-analysis of smoking quantity led by Saccone et al., the authors identified at least two signals within the region that are statistically independent of rs16969968, tagged by rs578776 and rs588765 [24]. The major (risk) allele of rs578776 is in phase with the minor (risk) allele of rs16969968. In the case of rs16969968, the minor allele increases risk for nicotine dependence, but for rs578776 the minor allele is protective against it. Consequently, although the risk loci are correlated with each other, the minor alleles are out of phase, and when controlling for rs16969968, rs578776 is no longer genome-wide significant. Moreover, when controlling for rs16969968, the significance and direction of effect for smoking risk for rs588765 is adjusted; rs588765 reaches genome-wide significance and its minor allele is associated with an increased risk for smoking quantity, whereas its minor allele was associated with a protective effect in the single locus model[24]. In short, previous studies have demonstrated that controlling for rs16969968 has the potential to uncover new associations with smoking behaviors and provide further nuance to previously discovered ones.

Based on current evidence, rs16969968 is not associated with early or subjective smoking behaviors, but rather an increased risk for mature smoking behaviors. To illustrate, a meta-analysis of ‘age of tobacco initiation’ and ‘age of onset of regular smoking’ [25] conducted by Stephens et al. found that rs16969968 was not associated with age of tobacco initiation; this was replicated in the GSCAN GWAS of 1.2 million individuals [7]. In addition, this study also failed to find a significant association between rs16969968 and age of regular smoking. A separate study demonstrated another non-significant finding for rs16969968 in relation to a subjective response phenotype, “dizziness”, suggesting that rs16969968 is not associated with subjective responses elicited with nicotine use [26]. In contrast, rs16969968 has been continuously associated with heaviness of smoking and nicotine dependence. On average, individuals homozygous for the risk allele in rs16969868 smoke up to 1.5 more cigarettes per day and are exposed to higher levels of tobacco smoke, as measured by their cotinine levels, than those with the GG non-risk allele [24, 27]. Absent balanced cross-over interactions, interaction effects are likely to be associated with at least some additive effect, and at one or the other locus, making rs16969968 a reasonable *a priori* candidate locus to study as a moderator. The number of cigarettes per day an individual consumes is simple to assay and readily available in many large biobanks, while still being a reasonable proxy for nicotine dependence [28]. In addition, this SNP is relatively common allele in some ancestral groups, found in around 37–43% of individuals of European and Middle Eastern descent, respectively. Although it is less common in other ancestral groups such as East Asian and African, at 2% and 7% respectively [20], the SNP has been associated with smoking behaviors in trans-ancestry analyses [29, 30]. In sum, since cigarettes per day is a widely used phenotype in many biobanks and rs16969968 is a highly replicated and common signal of large effect, we hypothesized that G×GWAS investigations using rs16969968 would be better powered than other SNPs in our search for epistatic effects influencing nicotine dependence. Moreover, given that the number of pairwise interactions between all genome-wide SNPs would be computationally infeasible to investigate, we concentrated our efforts in genome-wide interactions with rs16969968. To our knowledge, no study has explored potential interaction effects between rs16969968 and genome-wide loci influencing smoking behaviors. We aimed to determine whether rs16969968 interacts with other genome-wide loci at the SNP or the gene level to influence smoking quantity.

Subjects And Methods

Discovery Sample: UK Biobank Smokers of European Ancestry

We conducted our primary analyses in the UK Biobank [31], a biorepository with approximately 500 000 individuals. All unrelated participants of European ancestry reported currently or formerly smoking and had genotype data that passed quality controls were used (N = 116 442). We controlled for the two different SNP chips (Bioleve and Axiom) by controlling for batch and center. Participants were 40 years of age or older. Around 46% of our sample of unrelated smokers were female. We limited our analyses to only unrelated participants of European descent to minimize confounding factors such as population stratification and shared environmental influences. To identify individuals of European ancestry, we performed principal component analysis and retained those whose top scores on the first four principal components fell within the range of European ancestry previously determined by the UK Biobank (field 22006).

All data analysis and cleaning were performed using PLINK2 [32]. We first removed 849 individuals whose self-reported sex differed from their chromosomal sex determination (UKB data fields 31 and 22001) due to their increased probability of being a sample mix-up, 46 people with irregularly high inbreeding coefficients ($|F_{het}| > 0.02$), and 159 individuals who requested their information be redacted from the UKB, as well as any individuals whose genetic data did not pass quality controls identified by Affymetrix and the UK Biobank (fields 220010 and 22051). Then, we used MAF- and LD-pruned array markers (plink2 command: –maf 0.01 –hwe $1x10^{-8}$ –indep-pairwise 50 5 0.2) to randomly select individuals among related European smokers. For our analyses, we used the HRC-imputed dosage data provided by the UK Biobank's full release, which used the HRC reference panel v.1.1 [33] and an information score greater or equal to 0.3. We filtered MAF > 1% and tested ~ 10M SNPs across the 22 autosomal chromosomes.

Measures

Smoking quantity was measured by the average number of cigarettes smoked per day (CPD), including both current and former smokers (UKB field IDs 2887, 3456, and 6183; average = 18.22, median = 20, range 1-140, inclusive). Overall, most users tend to underestimate the amount they smoke, and this is particularly pronounced in former smokers in whom telescoping can partly explain why our measure of smoking quantity was right-skewed [34] (Fig. S1A). To assess whether changes to scale influence the tests of the interactions, we also investigated log10 transformed CPD (Fig. S1B).

All models tested employed the following covariates: sex (field 31), age at time of assessment (field 21003), Townsend Deprivation Index (proxy for socioeconomic status in the UK, field 21003), educational attainment (qualification, categorical, field 6138), genotyping batch (field 22000), assessment center (field 54), and the first 10 genetic principal components to control for ancestry. To calculate these 10 genetic PCs, we used common array markers with a minor allele frequency (MAF) equal of greater than 1% and filtered genetic markers by linkage disequilibrium using 50 variants at a time with a step size of 5 if the correlation between the variants surpassed 0.2 (plink2 command: –maf 0.01 –hwe $1\text{x}10^{-8}$ –indep-pairwise 50 5 0.2). Next, we used flashpca [35] on these LD-pruned array markers to calculate 10 genetic PCs to control for ancestry and population stratification. To reduce collinearity in our covariates, we ran principal component analysis using the prcomp function in R [36] to remove the axes that explained trivial variance, resulting in one dropped axis.

Replication Sample: Finnish Samples

To replicate any significant interactions, we chose five Finnish sub-samples with genetic and cigarette use data available as a replication sample. These five sub-samples include: FinHealth 2017 study (FinHealth) [37], FINRISK [38], Finnish Twin Cohort (FTC) [39, 40], GeneRISK [41], and the Health-2000-2011 (T2000-2011) [42]. These datasets varied in sample size (ranging from around 994 smoking individuals in GeneRISK up to 26 751 in FINRISK) and the granularity of the cigarette use outcome (i.e., cigarettes per day versus binned cigarettes per day). For more information on these samples, please see *Supplementary Methods*. The minor allele (A) frequency for rs16969968 in Finland is 0.33 and does not differ from other European populations. We confirmed our Finnish sample was an appropriate replication sample by comparing the Finnish linkage disequilibrium patterns of any gene regions of interest to those in our original UKB European sample (Fig.S6).

We performed a rs16969968×SNP interaction analysis for any replication regions in each of the five Finnish samples as described previously for the genome-wide interaction analyses. Then, we aggregated the interaction signals from our replication region to genes using MAGMA v.1.09 as performed in our previous UK Biobank analyses. Lastly, we meta-analyzed the results from the five Finnish subsets for both the rs16969968×SNP and gene-level analyses using METAL's inverse variance weighing model [43].

Genome-wide Interaction Study of rs16969968

We used PLINK2 to run a linear regression model (plink2 command: –linear interaction) to estimate SNP-by-rs16969968 interaction associations with CPD. We included all rs16969968×covariate and SNP_j×covariate interactions to avoid potential confounding [44]. Because covariate scales varied widely, all covariates and their products were further standardized (plink2 command: –covar-variancestandardize). We used a standard GWAS threshold of $5x10^{-8}$ for this analysis. Our regression model took the following form:

$$\mathrm{CPD} = \beta_0 + \beta_1 \mathrm{G} + \beta_2 \mathrm{Z_j} + \beta_3 \mathrm{G} * \mathrm{Z_j} + \sum_{p=1}^{q} \beta_p \mathrm{X_p} + \sum_{p=1}^{q} \beta'_p X_p * G + \sum_{p=1}^{q} \beta''_p X_p * Z_j + \varepsilon \text{ (Equation 1)}$$

Where X_p indicates the 1...q covariates, G indicates the number of risk alleles at rs16969968, Z_j indicates the j^{th} SNP in the G×GWAS, ϵ denotes environmental noise and measurement error.

In the Finnish samples, we defined a replication region as all SNPs within 250kb of the lead SNP in a significant gene interaction from the UKB analyses. This ensured that all SNPs in common between the Finnish and UKB samples in our region of interest plus any new SNPs that were likely to be in linkage disequilibrium with our SNPs of interest would also be included. We performed a rs16969968xSNP interaction analysis for any replication regions in each of the five Finnish samples as described previously for the genome-wide interaction analyses. We meta-analyzed the results from the rs16969968xSNP analyses across only the Finnish sub-samples (labelled Fin_Meta-analysis) and across both Finnish and UKB samples (labelled All_Meta-analysis) using METAL's inverse variance weighing model [43]. To determine the number of independent tests conducted, we performed a principal component analysis using the UKB on all the SNPs within any replication regions of interest using R.

MAGMA Gene-Level Analyses

To investigate rs16969968 interactions with gene level effects, we fed the resulting rs16969968-by-SNP_j *p*-values into the multi-marker analysis of genomic annotation (MAGMA) [45] v.1.09 to test gene-level interaction associations for CPD and log10-transformed CPD. Using MAGMA, one can employ either the "SNP-wise mean" or the "SNP-wise top" model to aggregate genome-wide signals at the gene level. The SNP-wise mean model is more powerful when several SNPs within a gene show a moderate association with the outcome of interest; the SNP-wise top model, on the other hand, is more powerful when a single SNP is strongly associated with the trait [46, 47]. To ensure our analyses would be sensitive to varying unknown genetic architectures, we used both MAGMA's top and mean *p*-value models separately (MAGMA commands –model SNP-wise top and –model SNP-wise mean, respectively). To our knowledge, this was the first time MAGMA has been used to perform GxGWAS interaction analyses. We investigated the likelihood of getting spurious results from using MAGMA in this novel fashion by simulating a random phenotype and running our rs16969968×SNP and subsequently our rs16969968×Gene analyses genome wide (See *Supplementary Methods*). While we did see inflation of the *p*-values across both the SNP-wise mean and SNP-wise top models, no genes were significant after controlling for multiple testing via a Bonferroni correction (Fig. S7).

In all the MAGMA analyses, variants were annotated to genes using a 25Kb window around the start and end of a gene. SNPs were successfully mapped onto a total of 18,573 genes using genome build 38. We used the SNP x rs16969968 interaction *p*-values for each SNP from the original GWAS, which accounted for the appropriate main effects, covariates and covariate interactions as described above, and included MAGMA's default covariates in the analysis (gene size, density, inverse minor allele count, per-gene sample size, plus the log value of each). We used a Bonferroni multiple testing correction significance threshold based on the number of genes tested ($p = 0.05/18\ 573 = 2.70\text{x}10^{-6}$), which is conservative given LD structure and overlap of gene regions.

In the Finnish samples, we aggregated the interaction signals from any replication regions to genes using MAGMA v.1.09 as performed in our previous UK Biobank analyses. We meta-analyzed the results from the rs16969968xSNP analyses across Finnish sub-samples (labelled Fin_Meta-analysis) using METAL's inverse variance weighing model.

Characterizing Significant Interactions

For any statistically significant genes from the gene-level MAGMA analysis ($p = 2.70\text{x}10^{-6}$), we followed up by inspecting the linkage disequilibrium patterns and performing additional conditional analyses on any gene regions of interest. Gene regions of interest include all the SNPs within a significant gene with a suggestive significance of $p < 1\text{x}10^{-5}$. We used HaploView [48] as well as LocusZoom [49] to visualize the linkage disequilibrium pattern of the gene regions of interest for any genes that reached statistical significance. To test whether a significant gene contained a single or multiple signals, we conducted interaction analyses on the all the SNPs within a gene region of interest while conditioning on the top SNP of that gene region. Our multiple testing correction threshold for these conditional analyses was defined by the number of SNPs in that region. To test the interactive effect of other SNPs in the gene with rs16969968 while conditioning on the top SNP, we exported the additive coding of all SNPs in the gene within MAGMA's 25kb window using PLINK (plink flag: –recode A), included the interaction between the top SNP and rs16969968 as well as the main effect of the top SNP and its interaction with the rest of our covariates. The conditional analysis followed the following regression model:

$$\text{CPD} = \beta_0 + \beta_1 \text{G} + \beta_2 \text{Z}_{\text{top}} + \beta_3 \text{Z}_{\text{j within gene region}} + \ \beta_4\, G * \text{Z}_{\text{top}} + \beta_5\, G * \text{Z}_{\text{j within gene region}} + \sum_{p=1}^{q} \beta_p\, X_p + \sum_{p=1}^{q} \beta'_{\ p}\, X_p * G + \ \sum_{p=1}^{q} \beta''_{\ p}\, X_p * \text{Z}_{\text{top}} \text{ (Equation 2)}$$

rs16969968 x Nicotinic Receptor Genes

Following the basic gene-level analysis, we ran a competitive gene set analysis using MAGMA v. 1.09 and the results from both our top and mean models in the UKB. We tested whether a nicotinic gene set previously curated by Melroy-Greif et al. (2017) interacted with rs16969968 more than other genes in the genome. This nicotinic gene set included 107 genes curated through a literature search that have been previously found to be involved in the function, processing, upregulation, or downstream effects of nicotinic receptors [50].

Results

Genome-wide Interaction Study of rs16969968

We found no genome-wide significant interactions of SNPs with rs16969968 for either CPD or log-transformed CPD (Fig. 1A & Fig.S2, respectively). However, 11 SNPs on chromosome 20 were suggestively significant (Fig. 1A). For example, rs17178947 and rs73586411 ($p = 6.60\text{x}10^{-8}$, $p = 7.49\text{x}10^{-8}$ respectively) are located within the *CDS2* gene and close to *TMEM* and *PCNA*. Most of the rs16969968xSNP interactions nearing significance were within this region, highlighting it as a potential epistatic region. To visualize this potential interaction, we plotted the average number of cigarettes per day smoked across genotypes for rs16969968 and rs73586411 (Fig.S3A). The MAF for the T allele tagging the interaction is 0.092, which explains our modest case counts.

We also tested this region in our Finnish replication sample. When meta-analyzing across both the UKB and the Finnish sub-sets, the SNP-level interaction for rs16969968×rs73586411 reached genome-wide significance (Table S1, $p = 2.31\text{x}10^{-8}$), though we note that this statistic suffers from winner's curse bias in that we meta-analyzed the SNP×SNP results because of the significant SNP×Gene interaction. Figure 3 shows the estimated effect sizes for this interaction within individual samples and across all samples. When meta-analyzing only across the Finnish samples, we first determined that the number of independent tests in our region of interest equaled 4 through principal component analysis (See *Methods*). Across the Finnish sub-samples, nine SNPs were nominally significant ($p < 0.04$; Table S1), but no SNP reached significance after adjusting for multiple testing ($p < 0.05/4 = 1.25\text{x}10^{-2}$).

rs16969968xGene Analyses

We used MAGMA to aggregate the resulting *p*-values from the rs16969968×SNP analysis by gene to detect any potential gene-level interactions with rs16969968. In both the SNP-wise Mean and Top models, we found the *PCNA* gene to significantly interact with rs16969968 when raw CPD was our outcome measure (Fig. 1B, $p = 8.02\text{x}10^{-7}$; Fig. 1C, $p = 3.67\text{x}10^{-7}$, respectively). However, no genes reached genome-wide significance for $\log_{10}$CPD in either model (Fig. S4A, $p = 2.71\text{x}10^{-5}$; Fig. S4B, $p = 2.21\text{x}10^{-5}$). We went back to the rs16969968×SNP results to investigate the SNP-level signals driving the significant interaction with the *PCNA* gene. Seeing that the SNP signals driving this significant gene interaction were located within the *CDS2* gene but were part of MAGMA's gene analyses for PCNA, CDS2, and TMEM230, we followed up on all three of these genes in our Finnish replication study, underscoring that the signal is shared across these three genes. In our rs16969968×Gene meta-analysis of the Finnish samples, all three genes (*CDS2*, *TMEM230*, and *PCNA*) were significant after multiple-testing correction ($p < 1.67\text{x}10^{-2}$)

across both the SNP-wise Mean and SNP-wise Top models (Table 1A and 1B, respectively); thereby successfully replicating our results for the *PCNA* gene in the UKB.

Exploring the suggestive interaction of rs16969968 x rs73586411

We used LocusZoom and HaploView to visualize the pattern of associations as a function of their linkage disequilibrium with the lead SNP (rs73586411) in the *PCNA* gene. All our suggestively significant interactions ($p < 1x10^{-5}$) from the rs16969968×SNP analyses for the *PCNA* gene were highly correlated with one another (Fig. 2) and aggregated in a single LD block, block 3 (Fig. S4).

To confirm whether this was a single signal, we conducted rs16969968×SNP interaction analyses for the SNPs within *PCNA*, conditioning on the rs16969968×rs7586411 interaction, the interaction with the lowest *p*-value in the *PCNA* gene. No SNPs were significant after controlling for multiple testing (p = 0.05/61 SNPs in the region = 0.00082).

rs16969968 x Nicotinic Receptor Genes

We performed competitive gene set analyses on CPD and $\log_{10}$-transformed CPD using the nicotinic gene set curated by Melroy-Greif et al. (2017) to test whether these genes significantly interacted with rs16969968 more strongly than other genes in the genome. Neither the SNP-wise mean (p = 0.29) nor the SNP-wise top (p = 0.10) competitive tests were significant, meaning that nicotinic gene set genes curated by Melroy-Greif et al (2017) do not more strongly interact with rs16969968 to influence CPD or its log10 transform compared to other genes in the genome.

Discussion

We conducted an exploratory study of SNP and gene interactions with the SNP rs16969968 on daily cigarette consumption. In the single SNP G×GWAS interaction analysis, none of the individual SNPs reached genome-wide significance. Notably, in the gene-level analysis, one gene, *PCNA*, did achieve genome-wide significance when aggregating our rs16969968×SNP *p*-values at the gene-level. This result was consistent with the individual SNP analysis, where some SNPs in the same region (tagged by rs73586411) had *p*-values approaching significance. Importantly, we replicated this gene-level finding in an independent dataset of five Finnish samples by specifically testing for an interaction between rs16969968 and three genes and meta-analyzing the results. Collectively, this replication sample confirmed our novel finding for all three genes, with *p*-values ranging from 0.0017 to $3.67x10^{-7}$, depending on the model used. The fact that all three of these genes were statistically significant in our replication analyses using the Finnish samples supports our conclusion that a region tagged by lead SNP rs73586411 and shared across these three genes significantly modulates the effect of the risk allele of rs16969968 and its effects on daily cigarette consumption.

A caveat is that both the SNP and gene level interactions for log10-transformed cigarettes per day were insignificant. At the SNP level using log10-transformed CPD, the *p*-value for rs73586411 was 9.66x10^{-5} compared to 7.50x10^{-8} for raw CPD. However, at the gene level, the interaction between rs16969968 and *PCNA* for log10-CPD was suggestively significant (*p* = 2.71x10^{-5} for SNP-wise mean, *p* = 2.21x10^{-5} for SNP-wise top model). Therefore, while there is some evidence to suggest that the interaction disappears on the multiplicative scale, we believe that our replication using an independent sample supports our initial findings of a significant interaction between rs16969968 and one or more SNPs found near the *PCNA* gene.

We explored the LD structure of the SNPs in the *PCNA* gene and conducted conditional analyses to determine that this is a single signal coming from an LD block containing 11 SNPs. We note that since we used a 25kb window, all these 11 nominally significant SNPs driving the interaction with *PCNA* also span part of the *CDS2* and *TMEM230* genes [51]. It is likely that the reason why *PCNA* resulted as statistically significant in our UKB analyses while not *CDS2* nor *TMEM230* was because the *PCNA* gene boundary used contained 48 SNPs, whereas the *CDS2* and *TMEM230* gene region boundaries contained 221 and 67, respectively. Therefore, we hypothesize that the higher number of SNPs in *CDS2* and *TMEM230* genes diluted the interaction signal between rs16969968 and rs73586411. None of the SNPs in high linkage disequilibrium are located within coding regions of any of the three genes. Most are located within intronic regions of *CDS2*, but there is no evidence for functional impact based on current information available for possible epigenetic areas or other known gene regulatory elements. In sum, we emphasize that this interaction is due to a single signal within the *PCNA, CDS2*, and *TMEM230* region of chromosome 20, but prioritization of possible functional SNPs cannot be identified in our analysis.

PCNA encodes for proliferating cell nuclear antigen, which is widely expressed across many tissues and involved in leading strand synthesis of DNA during replication. According to the GWAS catalog [52], height is the only phenotype with evidence of association with *PCNA* [53]. In contrast to GWAS, animal and transcriptomic studies have linked *PCNA* to smoking. For example, animal studies have linked nicotine exposure to *PCNA* damage in lung and kidney cell cultures in a dose-dependent fashion [54]. Interestingly, *PCNA* expression levels were higher in hepatic and pancreatic cells of rats exposed to both ethanol and tobacco compared to tobacco alone [55]. According to GeneWeaver [56], in humans, *PCNA* has been previously linked to tobacco smoke pollution, as well as having a couple of publications linking *PCNA* to nicotine according to the Comparative Toxicogenomics Database. *CDS2* codes for CDP-diacylglycerol synthase 2, which is an enzyme that regulates levels of phosphatidylinositol and is therefore involved in second messenger signaling for regulating cell growth, calcium metabolism, and protein kinase C activity. Notably, there are two genes that code for this enzyme, the other of which is located on chromosome 4q21. *CDS2* has emerged in four GWAS reports: two studies of height [57, 58], one on Ebbinghaus illusion, an inability to contextualize relative size perception [59], and most relevant to the present study, another identifying gene-gene interactions with pathological hallmarks of Alzheimer’s disease [60]. *TMEM230*, transmembrane 230, is expressed in neurons, as well as many other tissues, and may be involved in synaptic vesicle trafficking and recycling. It was identified in a GWAS study of acute myeloid leukemia [61], another with hair morphology [62], and there is ongoing debate about whether it may be associated with Parkinson's Disease [63]. In short, of the three genes encompassing our epistatic region of interest, to our knowledge, *PCNA* is the only one previously linked to smoking behaviors.

Our two-step approach of conducting a genome-wide interaction study and later aggregating these signals within genes successfully increased our power to detect genome-wide interactions while keeping our type I error rate low when evaluating unlinked SNPs; we recognize that LD among interacting SNPs can lead to false positive tests of epistasis [64, 65]. Moreover, it provided the flexibility to increase power while also allowing for follow-up of identified SNP×SNP results for further examination. The approach developed here will be useful for other researchers in the field attempting to discover genome-wide interactions with a wide range of complex traits. We used a 25kb window around the start and end of each gene, but there is no clear standard in the field for this. When using genes discovered in model organisms associated with nicotine consumption, Palmer et al. found that heritability for human nicotine consumption was enriched in genomic regions surrounding the genes compared to the protein-coding regions of these genes. In addition, after comparing 5, 10, 25, and 35kb gene windows, they found that enrichment began decreasing after 10kb [66]. These findings suggest that it is beneficial to use a gene window, although the best size of the window still merits further investigation and could vary across traits and across genes. In general, we recommend pooling data from multiple datasets to increase sample size, limiting SNP×SNP epistatic analyses to common variants, and using a 10kb-25kb upstream and downstream gene window when aggregating SNP×SNP results at the gene-level. These results serve as a guide for others in the field as they also attempt to study epistatic interactions at the SNP level.

In summary, this is the first study to report an interaction between rs16969968 and any genome-wide loci influencing cigarette consumption. Five of our nominally significant SNPs, such as rs73586411 and rs6053152, previously failed to reach significance for cigarettes per day in GSCAN, with sample sizes roughly 3–10 times the size used here [7]. This highlights the power of interaction studies to detect novel variants that would not be found otherwise. Future work could expand on our current pipeline to investigate interactions between rs16969968 and genome-wide loci for other smoking behaviors such as smoking cessation. In addition, one could apply our two-stage pipeline to SNP hits from large scale meta-analyses such as GSCAN to investigate other potential genome-wide interactions influencing smoking behaviors. These findings will help inform the work of basic scientists who are working on characterizing epistatic effects influencing smoking behaviors using animal models. Understanding how well-established risk variants such as rs16969968 alter risk for smoking behaviors in conjunction with the rest of the genome is increasingly important with the rise of precision medicine.

Declarations

ACKNOWLEDGEMENTS

This research was done using the UK-Biobank under project ref. 16651; We thank the participants of the UK Biobank and studies from which we utilized summary statistics. This work was supported by the

National Institutes of Health (grant numbers R01 AG046938-06 to Chandra A. Reynolds, R01 DA044283-01 to Scott I. Vrieze, and R01 MH100141-06 to Matthew C. Keller) and the University of Colorado Boulder Institute for Behavioral Genetics. This work utilized the Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University. Access to the Finnish samples was made possible thanks to Dr. Palmer through his NIH grant ref. DA042742. We also would like to thank Michael Stallings, Ph.D. and Chelsie Benca-Bachman, Ph.D. for their feedback on this manuscript. JK acknowledges support of the Sigrid Juselius Foundation and the Academy of Finland.

CONFLICT OF INTEREST

Authors report no personal conflicts of interest.

DATA AVAILABILITY

The UK Biobank data was accessed under application number 16651. These data cannot be shared with other investigators.

The Finnish datasets (FinHealth, FINRISK, FTC, GeneRisk, and T2000-2011) were accessed by an application to the Finnish Institute for Health Welfare Biobank by Drs. Palmer and Kaprio. These data should be applied for from the Biobank, for details see: https://thl.fi/en/web/thl-biobank/for-researchers.

CODE AVAILABILITY

All analyses were performed using open-source software. Genotypic and phenotypic data were collected by the UK Biobank. Code for the rs16969968xSNP linear interaction analyses (PLINK), or for the rs16969968xGene interaction analyses (MAGMA) can be provided upon request.

Supplementary information is available at MP’s website

References

1. Alberg AJ, Shopland DR, Cummings KM. The 2014 Surgeon General’s Report: Commemorating the 50th Anniversary of the 1964 Report of the Advisory Committee to the US Surgeon General and Updating the Evidence on the Health Consequences of Cigarette Smoking. Am J Epidemiol. 2014;179:403–12.
2. 2..S. Federal Trade Commission (FTC). Federal Trade Smokeless Tobacco Report for 2019. 2019.
3. Creamer MR, Wang TW, Babb S, Cullen KA, Day H, Willis G, et al. Tobacco Product Use and Cessation Indicators Among Adults — United States, 2018. MMWR Morb Mortal Wkly Rep [Internet]. Centers for Disease Control MMWR Office; 2019 [cited 2022 Jul 5];68:1013–9. Available from: https://www.cdc.gov/mmwr/volumes/68/wr/mm6845a2.htm

4. Kenny PJ, Markou A. Nicotine Self-Administration Acutely Activates Brain Reward Systems and Induces a Long-Lasting Increase in Reward Sensitivity. Neuropsychopharmacology [Internet]. 2006 [cited 2021 Nov 5];31:1203–11. Available from: http://www.acnp.org/citations/
5. Kaprio J. Genetic Epidemiology of Smoking Behavior and Nicotine Dependence. COPD J Chronic Obstr Pulm Dis. 2009;6:304–6.
6. Lessov-Schlaggar CN, Pang Z, Swan GE, Guo Q, Wang S, Cao W, et al. Heritability of cigarette smoking and alcohol use in Chinese male twins: the Qingdao twin registry. Int Epidemiol Assoc Int J Epidemiol [Internet]. Oxford University Press on; 2006 [cited 2021 Jun 18];35:1278–85. Available from: https://academic.oup.com/ije/article/35/5/1278/762228
7. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet. 2019;51:237–44.
8. Evans LM, Jang S, Hancock DB, Ehringer MA, Otto JM, Vrieze SI, et al. Genetic architecture of four smoking behaviors using partitioned SNP heritability. Addiction [Internet]. John Wiley & Sons, Ltd; 2021 [cited 2022 Feb 28];116:2498–508. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/add.15450
9. Quach BC, Bray MJ, Gaddis NC, Liu M, Palviainen T, Minica CC, et al. Expanding the genetic architecture of nicotine dependence and its shared genetics with multiple traits. Nat Commun 2020 111 [Internet]. Nature Publishing Group; 2020 [cited 2022 Feb 27];11:1–13. Available from: https://www.nature.com/articles/s41467-020-19265-z
10. Kauffman S. The Origins of Order. Oxford: Oxford University Press; 1993.
11. Mackay TFC. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nat Rev Genet 2013 151 [Internet]. Nature Publishing Group; 2013 [cited 2022 Feb 28];15:22–33. Available from: https://www.nature.com/articles/nrg3627
12. Mackay TFC, Moore JH. Why epistasis is important for tackling complex human disease genetics. Genome Med 2014 66 [Internet]. BioMed Central; 2014 [cited 2022 Feb 28];6:1–3. Available from: https://link.springer.com/articles/10.1186/gm561
13. Picciotto MR, Kenny PJ. Mechanisms of Nicotine Addiction. Cold Spring Harb Perspect Med [Internet]. Cold Spring Harbor Laboratory Press; 2021 [cited 2022 May 30];11:a039610. Available from: http://perspectivesinmedicine.cshlp.org/content/11/5/a039610.full
14. Wen L, Jiang K, Yuan W, Cui W, Li MD. Contribution of Variants in CHRNA5/A3/B4 Gene Cluster on Chromosome 15 to Tobacco Smoking: From Genetic Association to Mechanism. Mol Neurobiol. 2016;53:472–84.
15. Chen LS, Horton A, Bierut L. Pathways to precision medicine in smoking cessation treatments. Neurosci Lett. Elsevier; 2018;669:83–92.
16. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 2008 405 [Internet]. Nature

Publishing Group; 2008 [cited 2022 May 31];40:616–22. Available from: https://www.nature.com/articles/ng.109

17. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature [Internet]. 2008 [cited 2017 Jul 25];452:638–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18385739
18. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature [Internet]. 2008 [cited 2017 Jul 25];452:633–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18385738
19. Fowler CD, Turner JR, Imad Damaj M. Molecular Mechanisms Associated with Nicotine Pharmacology and Dependence. Handb Exp Pharmacol [Internet]. Handb Exp Pharmacol; 2020 [cited 2022 May 31];258:373–93. Available from: https://pubmed.ncbi.nlm.nih.gov/31267166/
20. Bierut LJ, Stitzel JA, Wang JC, Hinrichs AL, Grucza RA, Xuei X, et al. Variants in Nicotinic Receptors and Risk for Nicotine Dependence. Am J Psychiatry. American Psychiatric Association; 2008;165:1163–71.
21. Koukouli F, Rooy M, Tziotis D, Sailor KA, O'Neill HC, Levenga J, et al. Nicotine reverses hypofrontality in animal models of addiction and schizophrenia. Nat Med. Nature Publishing Group; 2017;23:347–54.
22. O'Neill HC, Wageman CR, Sherman SE, Grady SR, Marks MJ, Stitzel JA. The interaction of the Chrna5 D398N variant with developmental nicotine exposure. Genes, Brain Behav [Internet]. John Wiley & Sons, Ltd; 2018 [cited 2022 May 31];17:e12474. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/gbb.12474
23. Buck JM, O'Neill HC, Stitzel JA. The Intergenerational Transmission of Developmental Nicotine Exposure-Induced Neurodevelopmental Disorder-Like Phenotypes is Modulated by the Chrna5 D397N Polymorphism in Adolescent Mice. Behav Genet [Internet]. Springer; 2021 [cited 2022 May 31];51:665–84. Available from: https://link.springer.com/article/10.1007/s10519-021-10071-x
24. Saccone NL, Culverhouse RC, Schwantes-An TH, Cannon DS, Chen X, Cichon S, et al. Multiple independent loci at chromosome 15q25.1 affect smoking quantity: A meta-analysis and comparison with lung cancer and COPD. PLoS Genet. 2010;6.
25. Stephens SH, Hartz SM, Hoft NR, Saccone NL, Corley RC, Hewitt JK, et al. Distinct loci in the CHRNA5/CHRNA3/CHRNB4 gene cluster are associated with onset of regular smoking. Genet Epidemiol. 2013;37:846–59.
26. Ehringer MA, McQueen MB, Hoft NR, Saccone NL, Stitzel JA, Wang JC, et al. Association of CHRN genes with “dizziness” to tobacco. Am J Med Genet Part B Neuropsychiatr Genet [Internet]. John Wiley & Sons, Ltd; 2010 [cited 2020 Apr 14];153B:600–9. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ajmg.b.31027
27. Munafò MR, Timofeeva MN, Morris RW, Prieto-Merino D, Sattar N, Brennan P, et al. Association Between Genetic Variants on Chromosome 15q25 Locus and Objective Measures of Tobacco

Exposure. JNCI J Natl Cancer Inst [Internet]. Oxford Academic; 2012 [cited 2022 Mar 17];104:740–8. Available from: https://academic.oup.com/jnci/article/104/10/740/929306

28. Sanchez-Roige S, Cox NJ, Johnson EO, Hancock DB, Davis LK. Alcohol and cigarette smoking consumption as genetic proxies for alcohol misuse and nicotine dependence. Drug Alcohol Depend. Elsevier; 2021;221:108612.
29. Olfson E, Saccone NL, Johnson EO, Chen LS, Culverhouse R, Doheny K, et al. Rare, low frequency and common coding variants in CHRNA5 and their contribution to nicotine dependence in European and African Americans. Mol Psychiatry 2016 215 [Internet]. Nature Publishing Group; 2015 [cited 2022 Apr 21];21:601–7. Available from: https://www.nature.com/articles/mp2015105
30. Adjangba C, Border R, Romero Villela PN, Ehringer MA, Evans LM. Little Evidence of Modified Genetic Effect of rs16969968 on Heavy Smoking Based on Age of Onset of Smoking. Nicotine Tob Res [Internet]. Oxford University Press; 2021 [cited 2022 Jul 5];23:1055. Available from: /pmc/articles/PMC8150133/
31. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med [Internet]. Public Library of Science; 2015 [cited 2022 Mar 23];12:1001779. Available from: /pmc/articles/PMC4380465/
32. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience [Internet]. BioMed Central Ltd.; 2015 [cited 2022 Mar 10];4:7. Available from: https://academic.oup.com/gigascience/article/4/1/s13742-015-0047-8/2707533
33. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016 4810 [Internet]. Nature Publishing Group; 2016 [cited 2022 Apr 6];48:1279–83. Available from: https://www.nature.com/articles/ng.3643
34. Krall EA, Valadian I, Dwyer JT, Gardner J. Accuracy of Recalled Smoking Data. Public Health. 1989;
35. Abraham G, Inouye M. Fast Principal Component Analysis of Large-Scale Genome-Wide Data. PLoS One [Internet]. Public Library of Science; 2014 [cited 2022 Jan 11];9:e93766. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0093766
36. R Foundation for Statistical Computing. R: A language and environment for statistical computing. Vienna;
37. National FinHealth Study - THL [Internet]. [cited 2022 Mar 16]. Available from: https://thl.fi/en/web/thlfi-en/research-and-development/research-and-projects/national-finhealth-study
38. The National FINRISK Study - THL [Internet]. [cited 2022 Mar 16]. Available from: https://thl.fi/en/web/thlfi-en/research-and-development/research-and-projects/the-national-finrisk-study

39. Kaidesoja M, Aaltonen S, Bogl LH, Heikkilä K, Kaartinen S, Kujala UM, et al. FinnTwin16: A Longitudinal Study from Age 16 of a Population-Based Finnish Twin Cohort. Twin Res Hum Genet [Internet]. Cambridge University Press; 2019 [cited 2022 Mar 16];22:530–9. Available from: https://www.cambridge.org/core/journals/twin-research-and-human-genetics/article/finntwin16-a-longitudinal-study-from-age-16-of-a-populationbased-finnish-twin-cohort/066CB09103674DBA0210715C3AECC560

40. Kaprio J, Bollepalli S, Buchwald J, Iso-Markku P, Korhonen T, Kovanen V, et al. The Older Finnish Twin Cohort — 45 Years of Follow-up. Twin Res Hum Genet [Internet]. Cambridge University Press; 2019 [cited 2022 Mar 16];22:240–54. Available from: https://www.cambridge.org/core/journals/twin-research-and-human-genetics/article/older-finnish-twin-cohort-45-years-of-followup/94760871D9703AF3A974BB8E32B0180F

41. Widén E, Junna N, Ruotsalainen S, Surakka I, Mars N, Ripatti P, et al. How Communicating Polygenic and Clinical Risk for Atherosclerotic Cardiovascular Disease Impacts Health Behavior: an Observational Follow-up Study. Circ Genomic Precis Med [Internet]. Circ Genom Precis Med; 2022 [cited 2022 Mar 17]; Available from: https://pubmed.ncbi.nlm.nih.gov/35130028/

42. Health-2000-2011 - THL [Internet]. [cited 2022 Mar 16]. Available from: https://thl.fi/en/web/thlfi-en/research-and-development/research-and-projects/health-2000-2011

43. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinforma Appl NOTE [Internet]. 2010 [cited 2022 Feb 1];26:2190–1. Available from: http://www.sph.umich.edu/csg/abecasis/metal/

44. Keller MC. Gene × environment interaction studies have not properly controlled for potential confounders: The problem and the (simple) solution. Biol Psychiatry [Internet]. Elsevier; 2014;75:18–24. Available from: http://dx.doi.org/10.1016/j.biopsych.2013.09.006

45. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput Biol. 2015;11.

46. de Leeuw CA, Stringer S, Dekkers IA, Heskes T, Posthuma D. Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure. Nat Commun [Internet]. Nature Publishing Group; 2018 [cited 2020 Oct 6];9:3768. Available from: http://www.nature.com/articles/s41467-018-06022-6

47. de Leeuw CA, Neale BM, Heskes T, Posthuma D, Christiaan A. de Leeuw, Benjamin M. Neale TH and DP. The statistical properties of gene-set analysis. Nat Rev Genet [Internet]. Nature Publishing Group; 2016 [cited 2020 May 28];17:353–64. Available from: https://www.nature.com/articles/nrg.2016.29.pdf

48. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.

49. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics [Internet]. Oxford Academic;

2010 [cited 2022 May 31];26:2336–7. Available from: https://academic.oup.com/bioinformatics/article/26/18/2336/208507

50. Melroy-Greif WE, Simonson Phd MA, Corley RP, Lutz SM, Hokanson JE, Phd MAE, et al. Examination of the Involvement of Cholinergic-Associated Genes in Nicotine Behaviors in European and African Americans. Nicotine Tob Res [Internet]. 2017 [cited 2020 Dec 2];19:417–25. Available from: http://pngu.mgh.harvard.edu/~purcell/plink/index.
51. A K, S. S. No Title. Single Nucleotide Polymorph. Database Nucleotide Seq. Var. p. Chapter 5.
52. GWAS Catalog [Internet]. [cited 2022 Mar 16]. Available from: https://www.ebi.ac.uk/gwas/home
53. Barton AR, Sherman MA, Mukamel RE, Loh PR. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. Nat Genet 2021 538 [Internet]. Nature Publishing Group; 2021 [cited 2022 Mar 16];53:1260–9. Available from: https://www.nature.com/articles/s41588-021-00892-1
54. Salama SA, Arab HH, Omar HA, Maghrabi IA, Snapka RM. Nicotine mediates hypochlorous acid-induced nuclear protein damage in mammalian cells. Inflammation [Internet]. Springer New York LLC; 2014 [cited 2022 Jul 11];37:785–92. Available from: https://link.springer.com/article/10.1007/s10753-013-9797-6
55. Wang YY, Liu Y, Ni XY, Bai ZH, Chen QY, Zhang YE, et al. Nicotine promotes cell proliferation and induces resistance to cisplatin by α7 nicotinic acetylcholine receptor-mediated activation in Raw264.7 and El4 cells. Oncol Rep [Internet]. Spandidos Publications; 2014 [cited 2022 Jul 11];31:1480–8. Available from: http://www.spandidos-publications.com/10.3892/or.2013.2962/abstract
56. Baker EJ, Jay JJ, Bubier JA, Langston MA, Chesler EJ. GeneWeaver: a web-based system for integrative functional genomics. Nucleic Acids Res [Internet]. Oxford Academic; 2012 [cited 2022 Jul 11];40:D1067–76. Available from: https://academic.oup.com/nar/article/40/D1/D1067/2903626
57. Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am J Hum Genet [Internet]. Am J Hum Genet; 2019 [cited 2022 Mar 16];104:65–75. Available from: https://pubmed.ncbi.nlm.nih.gov/30595370/
58. Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshiba S, et al. A cross-population atlas of genetic associations for 220 human phenotypes. Nat Genet [Internet]. Nat Genet; 2021 [cited 2022 Mar 16];53:1415–24. Available from: https://pubmed.ncbi.nlm.nih.gov/34594039/
59. Zhu Z, Chen B, Na R, Fang W, Zhang W, Zhou Q, et al. A genome-wide association study reveals a substantial genetic basis underlying the Ebbinghaus illusion. J Hum Genet 2020 663 [Internet]. Nature Publishing Group; 2020 [cited 2022 Mar 16];66:261–71. Available from: https://www.nature.com/articles/s10038-020-00827-4
60. Wang H, Yang J, Schneider JA, De Jager PL, Bennett DA, Zhang HY. Genome-wide interaction analysis of pathological hallmarks in Alzheimer’s disease. Neurobiol Aging. Elsevier; 2020;93:61–8.
61. Lv H, Zhang M, Shang Z, Li J, Zhang S, Lian D, et al. Genome-wide haplotype association study identify the FGFR2 gene as a risk gene for Acute Myeloid Leukemia. Oncotarget [Internet]. Impact

Journals; 2016 [cited 2022 Mar 16];8:7891–9. Available from: https://www.oncotarget.com/article/13631/text/

62. Medland SE, Nyholt DR, Painter JN, McEvoy BP, McRae AF, Zhu G, et al. Common Variants in the Trichohyalin Gene Are Associated with Straight Hair in Europeans. Am J Hum Genet [Internet]. Elsevier; 2009 [cited 2022 Mar 16];85:750–5. Available from: http://www.cell.com/article/S0002929709004649/fulltext
63. Wang X, Whelan E, Liu Z, Liu CF, Smith WW. Controversy of TMEM230 Associated with Parkinson's Disease. Neuroscience. Pergamon; 2021;453:280–6.
64. Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, McRae AF, et al. Retraction Note: Detection and replication of epistasis influencing transcription in humans. Nat 2021 5967871 [Internet]. Nature Publishing Group; 2021 [cited 2022 Jul 7];596:306–306. Available from: https://www.nature.com/articles/s41586-021-03766-y
65. de los Campos G, Sorensen DA, Toro MA. Imperfect Linkage Disequilibrium Generates Phantom Epistasis (& Perils of Big Data). G3 Genes|Genomes|Genetics [Internet]. Oxford Academic; 2019 [cited 2022 Jul 7];9:1429–36. Available from: https://academic.oup.com/g3journal/article/9/5/1429/6026428
66. Palmer RHC, Benca-Bachman CE, Huggett SB, Bubier JA, McGeary JE, Ramgiri N, et al. Multi-omic and multi-species meta-analyses of nicotine consumption. Transl Psychiatry 2021 111 [Internet]. Nature Publishing Group; 2021 [cited 2022 Jul 11];11:1–10. Available from: https://www.nature.com/articles/s41398-021-01231-y

Figures

Figure 1

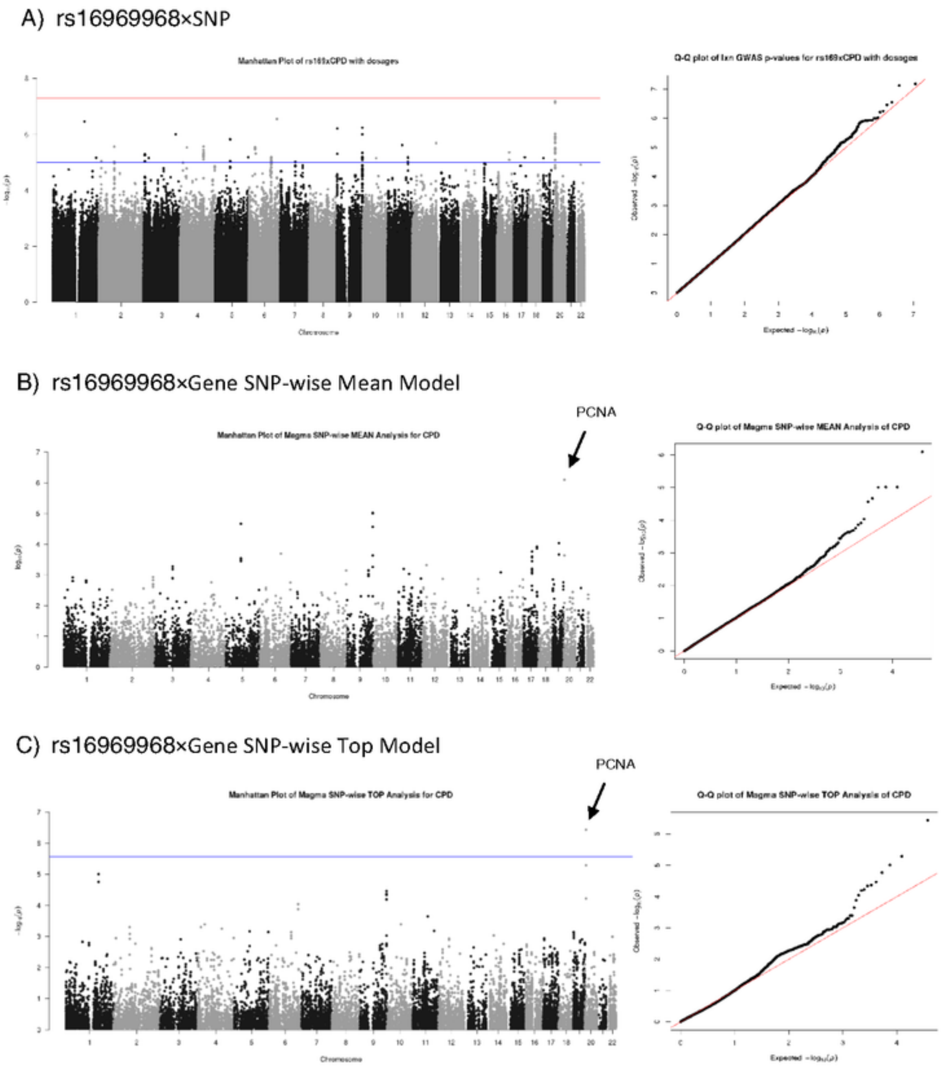

Figure 1

(**A**) Manhattan plot of associations with cigarettes per day for interactions between rs16969968 and genome-wide SNPs. Red line denotes genome-wide significance ($p < 5 \times 10^{-8}$), while the blue line denotes suggestive significance ($p < 1 \times 10^{-5}$).

(**B**) *PCNA* gene significantly interacts with rs16969968 to influence raw CPD when aggregating the rs16969968×SNP signals within genes and investigating the average association within genes. Blue line on the MAGMA results denotes genome-wide significance after correcting for the number of genes tested ($p < 2.70\text{x}10^{-6}$).

(**C**) *PCNA* gene again showed a significant association with cigarettes per day when aggregating the rs16969968×SNP results within genes and testing the top association of each gene.

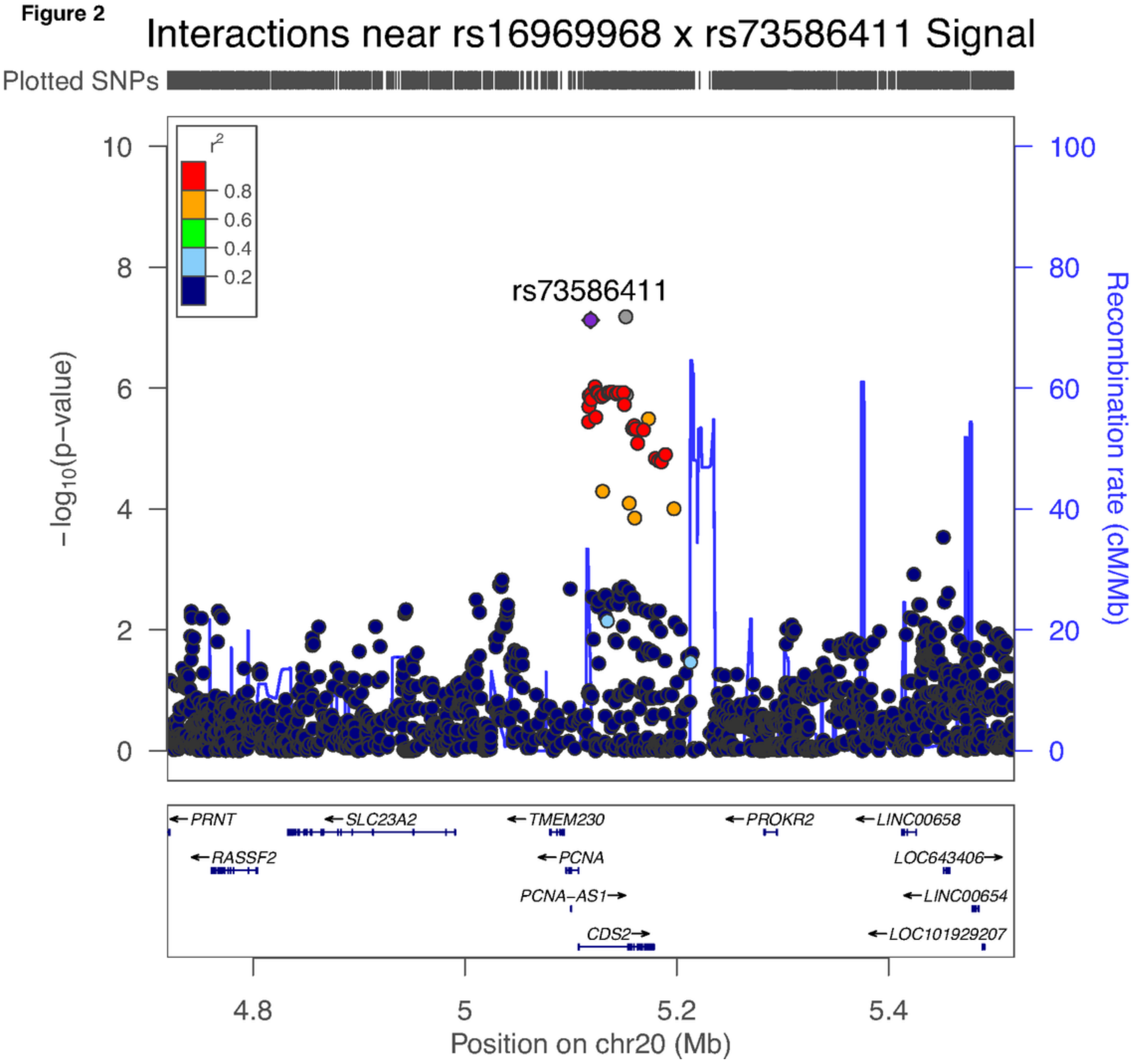

Figure 2

Locus Zoom plot for region of interest, tagged by rs73586411.

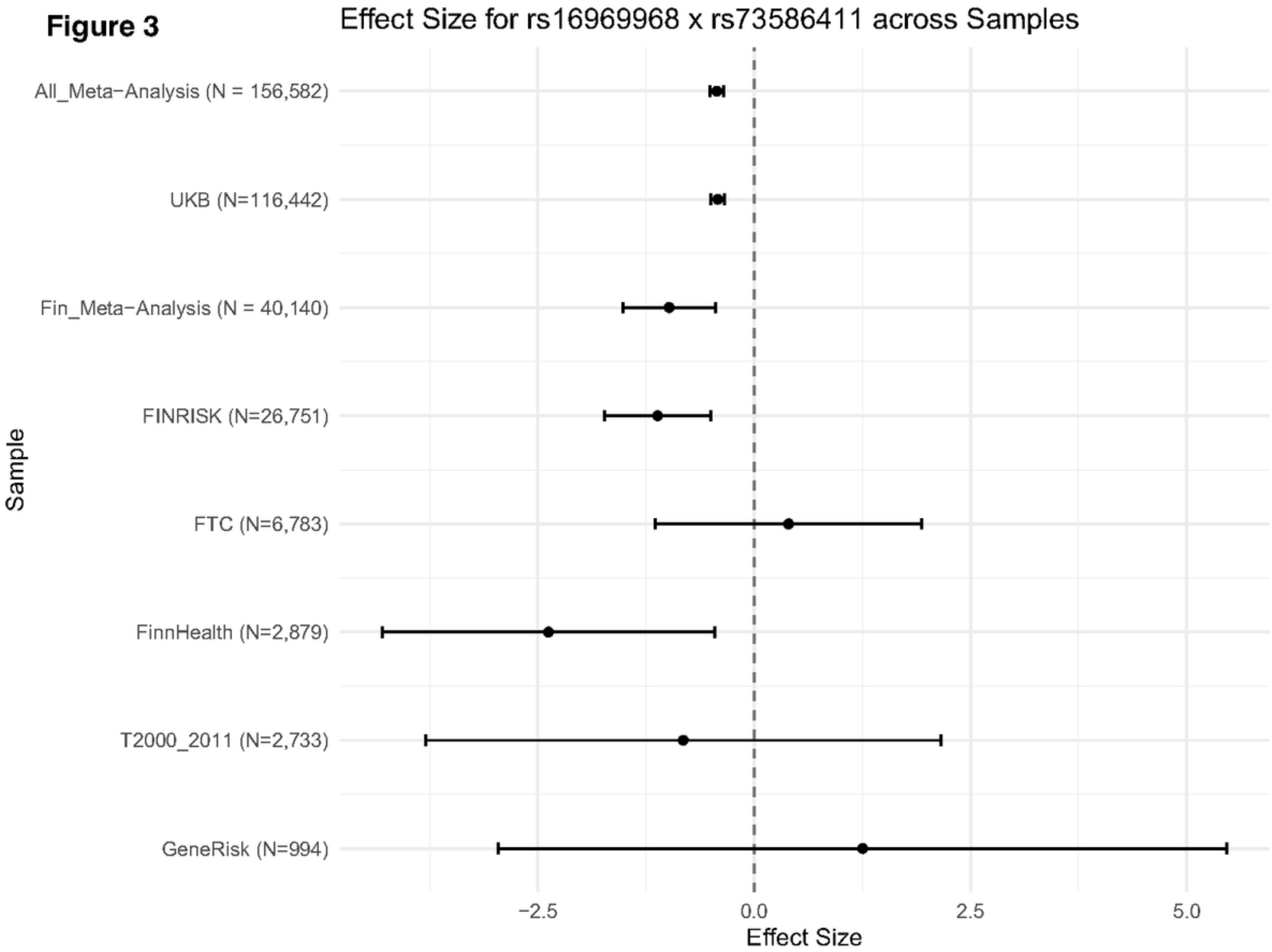

Figure 3

Estimated rs16969968×rs735864111effect sizes, alongside their standard error for those estimates across samples. The sample size of each sample is denoted in parentheses; samples are ordered according to decreasing sample size.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RomeroVillelaSupplementaryTable.docx

- RomeroVillelaSupplementaryFigures.docx
- RomeroVillelaSupplementaryMethodResults.docx