Exploring Genetic Variants Associated with Smoking Initiation Using GWAS Summary Statistics

Yuning Shu

Abstract

This study investigates the genetic basis of smoking initiation by analyzing summary statistics from a large genome-wide association study (GWAS) dataset (ieu-b-4877) by Liu et al. (2019). We aim to identify independent loci significantly associated with the likelihood of initiating smoking and annotate them to uncover potential biological mechanisms. We conduct functional mapping using Ensembl and perform linkage disequilibrium (LD) clumping to filter top-associated variants.

Our findings reveal multiple genome-wide significant loci, including previously implicated genes such as CHRNA5 and BDNF, as well as potentially novel variants. These results reveal the relevance of potential signaling pathways in smoking behaviors and offer candidate targets for future behavioral or pharmacological interventions.

Introduction

Smoking is a major risk factor for numerous diseases, including lung cancer, upper aerodigestive cancer, and chronic obstructive pulmonary disease (COPD) (Ezzati et al., 2002). Previous studies have shown that while environmental and social factors influence smoking behavior, genetic factors also play a significant role in the determination of smoking initiation (SmkInit) (Batra et al., 2003).

While large-scale genome-wide association studies have uncovered hundreds of loci associated with smoking-related phenotypes, many signals remain uncharacterized at the functional level. The availability of public GWAS summary statistics enables post-GWAS analyses to explore the genetic architecture underlying these traits.

A recent GWAS meta-analysis by Liu et al. (2019) involving over 1.2 million individuals identified hundreds of loci related to various smoking-related phenotypes, including initiation. In this study, we focus on post-GWAS analyses of the SmkInit phenotype using summary statistics from the ieu-b-4877 dataset derived from Liu et al.'s study. To identify and interpret the most significant associations, a series of post-GWAS computational steps is applied. First, statistical visualizations such as Manhattan, QQ, and volcano plots are used to assess the strength and distribution of association signals. Then, Functional annotation is conducted using the Ensembl Variant Effect Predictor to assess variant consequences. Also, LD clumping is performed to isolate independent signals, followed by biological interpretation. Together, these methods provide both statistical and mechanistic insight into the genetic basis of smoking initiation.

Methods

1. GWAS Summary Statistics Acquisition

Summary-level association statistics for the phenotype smoking initiation (SmkInit) are downloaded from the IEU OpenGWAS Project (dataset ID: ieu-b-4877). This dataset is obtained from Liu et al. (2019), who performed a meta-analysis of smoking phenotypes in over 1.2 million individuals of predominantly European ancestry.

2. Quality Control and SNP Filtering

To ensure robust downstream analysis, the following filters are applied. First, SNPs with missing or null values in effect size, p-value, or allele information are excluded. Then, variants with minor allele frequency (MAF) < 0.01 are removed to focus on common variants. SNPs are sorted based on ascending p-values.

3. Visualization of Results

Three key plots are generated using Python to examine statistical properties. Manhattan plot plots $-\log_{10}(p)$ across genomic coordinates to identify significant loci. QQ plot compares observed and expected p-values under the null. Volcano plot displays $-\log_{10}(p)$ against effect size.

4. Functional Annotation

Variant-level functional annotation is performed using the Ensembl Variant Effect Predictor (VEP), which classifies SNPs based on their predicted impact, such as intronic, intergenic, missense, synonymous, or regulatory. Annotation is based on Ensembl gene models (GRCh37). Variants with potential biological relevance can be identified.

5. Linkage disequilibrium (LD) clumping

LD clumping is then conducted using PLINK 1.9, with a primary p-value threshold of 5e-8, a secondary threshold of 1.0, an r^2 threshold of 0.001, and a clumping window of 10,000 kb. The European population panel from the 1000 Genomes Phase 3 dataset serves as the LD reference. This step reduces redundancy by retaining only the lead variant from each cluster of correlated SNPs.

Results and Discussion

1. Genome-wide Significance and Distribution of Association Signals

The Manhattan plot (Figure 1) shows clear peaks on several chromosomes, including 2, 6, 11, and 15. The peak on chromosome 15 corresponds to the CHRNA5 gene, which encodes neuronal acetylcholine receptor subunit alpha-5 and has been found associated with nicotine dependence

and smoking behavior (Chen et al., 2015). The peak on chromosome 11 aligns with the brain-derived neurotrophic factor (BDNF) locus, which has been found to be important to learning and memory (Noble et al., 2011).

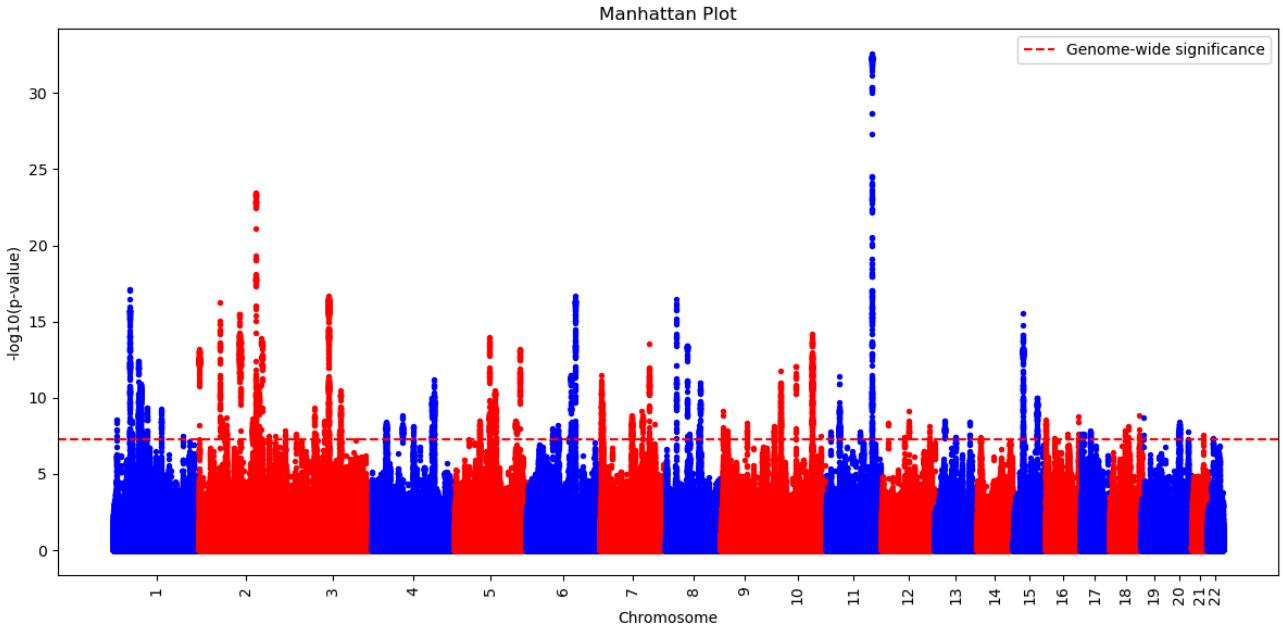

Figure 1. *Manhattan plot of genome-wide association results for smoking initiation. Each point represents an SNP, with chromosomal position on the x-axis and $-\log_{10}$(p-value) on the y-axis. Genome-wide significance threshold is marked at $p = 5 \times 10^{-8}$.*

The QQ plot (Figure 2) shows that the observed inflation reflects a polygenic signal in the data. This pattern of enrichment is also consistent with Liu et al.'s 2019 study's reported SNP-based heritability and polygenic model for smoking initiation.

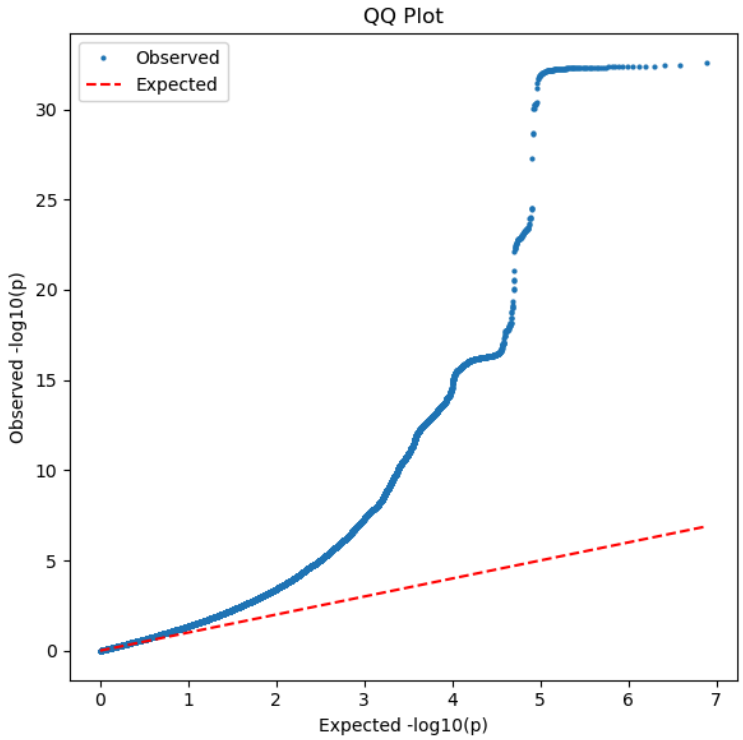

Figure 2. *Quantile-Quantile (QQ) plot of observed versus expected p-values. Deviation from the diagonal at the tail indicates polygenic enrichment and potential true associations.*

2. SNP Effect Sizes and Statistical Significance

In the volcano plot (Figure 3), a cluster of SNPs with strong significance ($-\log_{10}(p) > 15$) also shows moderate-to-large effect sizes ($|\beta| > 0.05$). These variants may serve as candidate causal loci or proxies in LD with functional mutations. This reflects similar observations in Liu et al. (2019), where multiple loci were found to have both genome-wide significance and functional relevance.

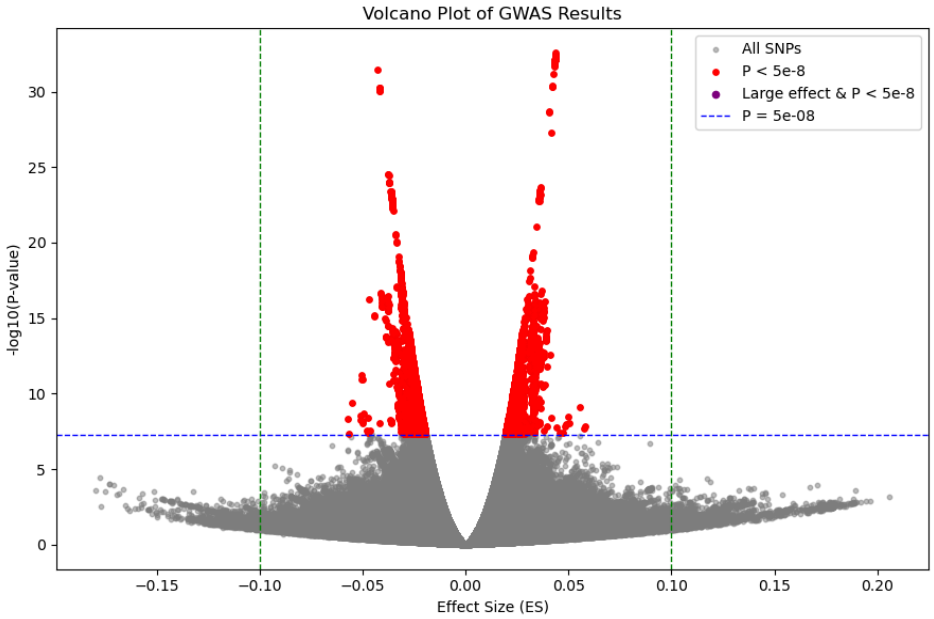

Figure 3. *Volcano plot of SNP effect sizes versus $-\log_{10}$(p-values). SNPs with both strong effect sizes and statistical significance are highlighted in the upper corners.*

3. Functional Consequences of Variants

The functional consequence pie chart (Figure 4) shows that the majority of significant SNPs (50%) are intron variants, and another 30% are non-coding transcript variants. A small fraction (~2%) falls in intergenic regions, suggesting proximity to enhancers or long-range regulatory elements. Surprisingly, 0% were in annotated regulatory regions or UTRs, likely due to annotation granularity or limitations in current gene models. Among coding variants (Figure 5), 60% are missense variants, with the remainder being synonymous. This finding supports the model proposed by Liu et al. (2019), who noted that regulatory variants account for the bulk of genetic influence on smoking traits, though notable protein-coding mutations such as rs16969968 in CHRNA5 play critical roles.

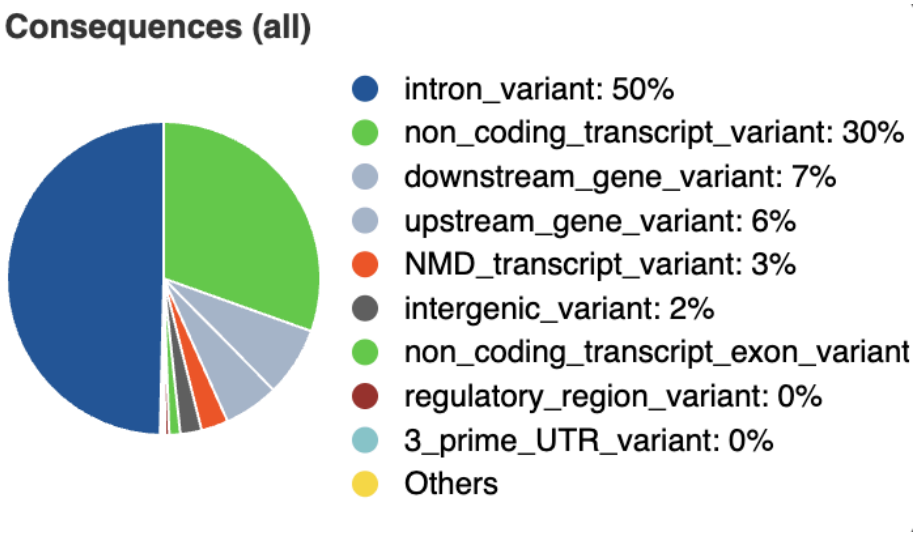

Figure 4. *Distribution of functional consequences for all annotated SNPs based on Ensembl VEP.*
Most variants are intronic or located in non-coding transcript regions, consistent with regulatory influence.

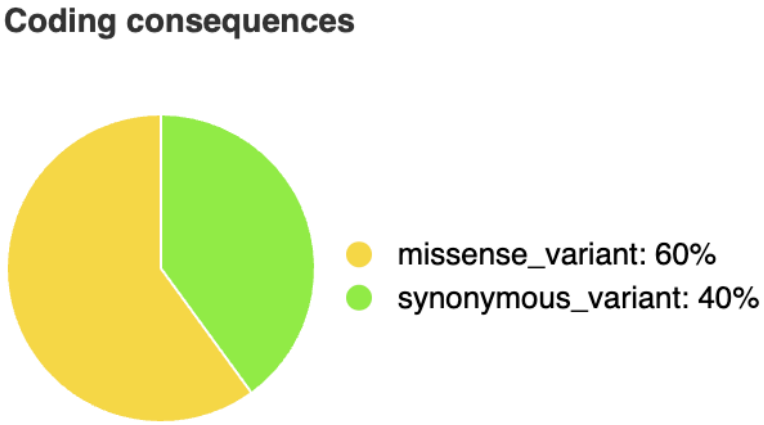

Figure 5. *Distribution of coding consequences among protein-altering SNPs.*
Missense variants dominate among the exonic SNPs, suggesting potential effects on protein function.

4. LD clumping

After LD clumping, 157 independent genome-wide significant SNPs remain. Chromosome 2 shows the highest concentration and contains 21 of the lead variants. The most statistically significant SNP is rs7938812 on chromosome 11 ($p = 2.71 \times 10^{-33}$). These variants represent a core set of candidate loci for future functional studies, genetic modeling, and cross-phenotype comparison in addiction biology.

Conclusions

In this project, we perform a post-GWAS analysis to explore the genetic architecture of smoking initiation using summary statistics. By applying statistical visualization, functional annotation through Ensembl VEP, and LD clumping, we identify 157 independent SNPs with genome-wide significance, enriched in brain-expressed genes and key neurobiological pathways, providing insights into future work on personalized smoking cessation strategies or preventive interventions targeting high-risk individuals.

Supplementary Materials

The GWAS summary statistics and the reference data for LD clumping are publicly available. The codes for data filtering, visualization, and LD clumping can be found at https://github.com/yuningshu/gwas-smoking.

References

Batra, Vikas, et al. "The genetic determinants of smoking." Chest, vol. 123, no. 5, May 2003, pp. 1730–1739, https://doi.org/10.1378/chest.123.5.1730.

Chen, Li-Shiun, et al. "CHRNA5 risk variant predicts delayed smoking cessation and earlier lung cancer diagnosis—a meta-analysis." *JNCI: Journal of the National Cancer Institute*, vol. 107, no. 5, 13 Apr. 2015, https://doi.org/10.1093/jnci/djv100.

Ezzati, Majid, et al. "Selected major risk factors and global and regional burden of disease." The Lancet, vol. 360, no. 9343, Nov. 2002, pp. 1347–1360, https://doi.org/10.1016/s0140-6736(02)11403-6.

Liu, Mengzhen, et al. "Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use." Nature Genetics, vol. 51, no. 2, 14 Jan. 2019, pp. 237–244, https://doi.org/10.1038/s41588-018-0307-5.

Noble, Emily E., et al. "The lighter side of BDNF." *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 300, no. 5, May 2011, https://doi.org/10.1152/ajpregu.00776.2010.