A meta-analysis of genome-wide associations with COPD

by

Guojun Ma
B.Sc., University of Alberta, 2020

Supervisory Committee

Dr. Yu-Ting Chen, (Co-)Supervisor
(Department of Mathematics and Statistics)

Dr. Xuekui Zhang, (Co-)Supervisor
(Department of Mathematics and Statistics)

ABSTRACT

This paper presents a comprehensive review of methodologies and software tools for conducting meta-analyses in the context of genome-wide association studies (GWAS). In addition, we conduct a meta-analysis of genetic associations with Chronic Obstructive Pulmonary Disease (COPD) using ancestrally diverse samples from the Global Biobank Meta-analysis Initiative (GBMI). Our analysis identifies 28 COPD-associated genes, including 18 novel loci, with several exhibiting ancestry-specific heterogeneous genetic effects. Gene-set enrichment analysis reveals significant overlap between COPD-associated genes and pathways implicated in other inflammatory conditions, particularly inflammatory bowel diseases, suggesting shared mechanisms that merit further functional investigation.

Table of Contents

Chapter 1

Review of the meta-analysis methods in GWAS study

1.0.1 Introduction

The Genome-Wide Association Study (GWAS) is an observational research method that explores the relationship between genetic variations and susceptibility to diseases or traits across the entire genome. This approach primarily focuses on analyzing variations in single-nucleotide polymorphisms(SNPs), the most common type of genetic variation in the genome between individuals. Over the past decade, GWAS has become a cornerstone of human genetic research. As of 2017, researchers have successfully identified approximately $55,000$ unique loci in the genome associated with almost $5,000$ all diseases and traits [MacArthur et al., 2017]. The widespread adoption of GWAS has been facilitated by advancements in sequencing technology, resulting in a significant reduction in the cost of sequencing the entire genome. The results of GWAS not only assist researchers in gaining insight into the underlying biology, but also contribute to advancements in disease prediction and personal health treatment.

In scientific research, it is common to encounter discrepancies when multiple studies investigate the same research question using varied methodologies and study populations. Such inconsistencies can lead to conflicting conclusions and hinder the formation of consensus. To address this problem, researchers often conduct meta-analysis, a statistical technique that quantitatively integrates the results from multiple independent

studies. By pooling data, meta-analyses improve statistical power, identify patterns across studies, and yield more precise and objective estimates of effect sizes. Accordingly, meta-analyses are regarded as the highest level of evidence in the hierarchy of evidence-based medicine [Herrera Ortiz et al., 2022].

Meta-analysis is often an essential procedure in genome-wide association studies (GWAS). Because many genetic associations have small effect sizes, very large sample sizes are required to achieve sufficient statistical power for detection. Although advances in sequencing technology have made genotyping more accessible, collecting millions of individual-level samples remains logistically and financially infeasible. Moreover, due to privacy concerns, many biobanks and research studies do not publicly release individual-level genetic data; instead, they provide only summary statistics. Meta-analysis provides a practical solution by allowing researchers to leverage this publicly available information to detect novel variants. Additionally, genetic architecture can vary substantially across populations. Meta-analytic frameworks also support the investigation of such heterogeneity by modelling variation due to ancestry, environmental exposures, and other modifying factors, making them particularly valuable in multi-ethnic genetic research.

1.1 Methods

1.1.1 Study design

There are generally two types of meta-analysis that are commonly performed in practice. The first type is aggregate-based meta-analysis, which combines summary data such as means, standard deviations, and sample sizes from individual studies to estimate the overall effect. The second type is patient-based meta-analysis, which pools individual-level data from multiple studies instead of just the summary data. In general, patient-based meta-analysis is often more time-consuming and cumbersome than aggregate-based meta-analysis since it requires researchers to access, manage, and analyze large sets of data from different sources.

In the context of GWAS, Meta-analysis commonly follows the following basic principles:

- In the first stage, the summary statistics from each study are collected and cleaned. Convert each one to the same genomic build(the newest genomic build is GRCh38 and the older one is GRCh37). Remove poor-quality and rare SNPs. Ensure the measuring unit in each study is described consistently.

- For each SNP, calculate the combined effect estimate as the weighted average of the effects estimated in each study:

$$\hat{\theta} = \frac{\sum Y_i W_i}{\sum W_i},$$

  where Y_i is the effect estimated in the i-th study, W_i is the weight given to the i-th study, and the summation is across all studies.

- Consider whether the true effect is the same or varies across different studies. If it is assumed the true effect is the same, it is common to conduct a fixed-effect meta-analysis. Alternatively, a random-effects meta-analysis or meta-regression can be performed to model heterogeneity in true effect across studies.

- Assess whether there is heterogeneity among the results of the separate studies and investigate the source of variations.

1.1.2 Fixed-effect meta-analysis

It is common the assume the observed effect for each study follows the additive model:

$$y_i = \theta + e_i, \tag{1.1}$$

where y_i denotes the observed effect in the i-th study, θ denotes the corresponding unknown true effect and e_i denotes the noise for the i-th study with mean 0 and variance s_i^2. One way to estimate the unknown true effect is the inverse-variance

estimator, which is

$$\hat{\theta} = \frac{\sum y_i(1/s_i^2)}{\sum(1/s_i^2)}, \tag{1.2}$$

where the summation is over each study. Studies with larger sample sizes typically have smaller variances s_i^2 , resulting in a more significant weight being assigned. In the case of unknown variance s_i^2, one can first estimate it and plug it into the above formula. The standard error of the pooled effect estimate is

$$SE(\hat{\theta}) = \sqrt{\frac{1}{\sum(1/s_i^2)}}.$$

The confidence interval is given as $\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$, where $z_{\alpha/2}$ is the critical value from a standard normal distribution corresponding to a given significance level α.

1.1.3 Random-effect meta-analysis

The inverse variance method can be biased when there is heterogeneity in the true effect θ across studies. Variations could arise from differences in population structure, sequencing technology, age, gender, etc. One way to model such heterogeneity is to assume the true effects θ follow a probability distribution with mean μ and variance σ^2. The model (1.1) can be written as

$$y_i = \mu + d_i + e_i, \tag{1.3}$$

where d_i is a random variable with mean 0 and variance σ^2 represents the between-study heterogeneity. In this case, it is common to employ a two-step approach to obtain the estimator for the average true effect μ. First, we estimate the variance $\hat{\sigma}^2$, using a method such as the Dersimonian and Laird (DL) estimator. Then, estimate μ using the inverse-variance method

$$\hat{\mu} = \sum \hat{w}_i y_i / \sum \hat{w}_i,$$

where $\hat{w}_i = (\hat{\sigma}^2 + s_i^2)^{-1}$.

1.1.4 Meta-regression

Although random-effect meta-analysis accounts for heterogeneity by assuming the true effects vary between studies, it does not identify the sources of this variability. Meta-regression is a method that investigates whether certain study-level characteristics explain variations in effect sizes across studies. This approach is beneficial when researchers want to explore whether differences in study design, population or interventions contribute to the observed variation. The mathematical framework of meta-regression as follows:

$$y_i = \mu + \beta x_i + e_i, \tag{1.4}$$

where μ is the true effect, x_i is the covariate for the study i, β is the coefficient of covariate, and e_i is the noise term. For example, the covariate could be the average age of a specific study group. A non-zero coefficient β implies there is a correlation between age and the effect size. More variables can be included in this model as well, allowing researchers to ask more complex questions.

1.1.5 Measuring heterogeneity

During a meta-analysis, variation among effect estimates from different studies is often observed. This variation can arise from various factors, such as differences in group characteristics or changes in study design. For instance, when examining the effectiveness of the COVID-19 vaccine, it may be found that the vaccine has a more pronounced impact on reducing mortality rates among older individuals compared to younger ones. Researchers would be interested to know if such variations arose from chance or due to different groups' characteristics.

Various methods can be utilized to measure the degree of heterogeneity in meta-analysis. One effective method is the forest plot, which displays the effect estimate, standard error, and confidence interval for each study. If the confidence intervals of individual studies do not overlap, it indicates the presence of heterogeneity. Additionally, statistical tests like the χ^2 test statistic can be used to evaluate heterogeneity,

defined as

$$Q = \sum_{i=1}^{k} \frac{(\theta_i - \hat{\theta})^2}{s_i^2},$$

where k the number of studies, θ_i the effect estimate for the study i, $\hat{\theta}$ is the inverse-variance weighted average, and s_i is the standard error for the study i. Under the null hypothesis that there is no heterogeneity, Q follow the χ^2 distribution with $k-1$ degrees of freedom. The p-value can be obtained by referring to χ^2 distribution.

Another way to measure the degree of heterogeneity is by using the I -squared statistic. It is defined as:

$$I^2 = \frac{Q - df}{Q} \times 100\%,$$

where Q is the χ^2 statistics and df is the degrees of freedom. I^2 describes the percentage of the variability in effect estimates that is due to heterogeneity. According to the Cochrane Handbook [**?**], some rough guidelines for interpreting I-squared are:

- 0 % to 40 % : might not be important;
- 30 % to 60 % : moderate heterogeneity;
- 50 % to 90 % : substantial heterogeneity;
- 75 % to 100 % : Considerable heterogeneity.

When a meta-analysis shows significant variation, researchers use different strategies to identify and explore the underlying sources of this variability. Meta-regression involves adding covariates to the meta-analysis and testing whether specific covariates are associated with effect size differences. Subgroup analysis, which compares the effect estimates within different subgroups, is another valuable approach. Sensitivity analyses can be used to check for outliers by excluding studies to see if the overall finding remains consistent.

1.2 Software

There are multiple specialized software programs for conducting meta-analyses in the GWAS studies 1.1, which includes specific functions to address challenges encountered in GWAS research. One of the unique challenges is the inconsistent coding of SNPs across various datasets, often referred to as the 'strand' issue. For instance, if a SNP has alleles A and T, one study may designate A as the reference allele, while another study may designate T as the reference allele. This inconsistency may lead to reversing the SNP's effect direction. Most specialized software has the feature to detect and correct such errors.

Another challenge is population stratification, which can arise from differences in allele frequencies or differences in the pattern of linkage disequilibrium(LD) between subpopulations. This can introduce confounding factors in the association between SNPs and traits. One could utilize a meta-regression approach implemented in MR-MEGA, which generates the genetic principal components from the summary statistics and uses them as covariates in the meta-analysis model. One could also use the Bayesian approach implemented in MANTRA. Alternatively, genomic control correction methods are included in many software programs.

Furthermore, GWAS data files are commonly available in various formats and are often substantial in size. Therefore, an ideal software tool for GWAS should be able to handle multiple file formats, have efficient memory management, and be fast in computing.

Table 1.1: Common software used for GWAS meta-analysis

Software	**METAL**	**GWAMA**	**PLINK**	**MANTRA**	**MR-MEGA**
Fix-effect	✓	✓	✓		
Random-effect		✓	✓		
meta-regression					✓
Bayesian model				✓	
Flexible File format	✓	✓		✓	✓
Resolve strand issue	✓	✓	✓	✓	✓
Genomic control correction	✓	✓	✓	✓	✓

Chapter 2

Multi-ancestry genome-wide association meta-analysis of COPD

2.1 Background

Chronic obstructive pulmonary disease (COPD) is a respiratory disorder characterized by persistent airflow limitation and breathing difficulties. Common symptoms include cough, difficulty breathing, wheezing and tiredness. The primary causes are prolonged exposure to noxious substances such as cigarette smoke and air pollutants. In some cases, COPD can also arise from a genetic condition called alpha-1 antitrypsin deficiency, which predisposes individuals to early-onset disease [Talamo, 1975]. According to the World Health Organization (WHO), COPD is the fourth leading cause of death globally, responsible for 3.5 million deaths in 2021, accounting for nearly 5% of all global deaths. Notably, the burden of premature COPD mortality, defined as deaths before age 70, is disproportionately concentrated in low- and middle-income countries.

Over the past decade, genome-wide association studies (GWAS) have been instrumental in uncovering the genetic underpinning of COPD and related lung function traits. Early GWAS efforts successfully identified several loci associated with COPD susceptibility, such as those in the CHRNA3/5, HHIP and FAM13A regions, which advanced insights into disease mechanisms and pathways [Pillai et al., 2009, Cho et al., 2010, Cho et al., 2012, Cho et al., 2014]. However, like many early GWAS,

these studies were predominantly focused on European-ancestry populations, limiting the generalizability of their finding. To address this limitation, the Global Biobank Meta-analysis Initiative (GBMI), established in 2019, integrates genomic and phenotypic data from more than 20 biobanks worldwide, encompassing representative individuals from each part of the world [Zhou et al., 2021]. By leveraging this ancestrally diverse cohort, the present study aims to (1) identify novel genetic variants associated with COPD across multiple populations, (2) characterize heterogeneity in genetic effects driven by ancestry differences, and (3) improve disease risk prediction and therapeutic insights through a more globally representative framework.

2.2 Materials and Methods

2.2.1 Data

We obtained the GWAS summary statistics from the Global Biobank Meta-analysis Initiative (GBMI). For the current study focusing on chronic obstructive pulmonary disease (COPD), detailed information for each ancestry group is provided in the table **??**. For each ancestry-specific cohort, an inverse variance weighted meta-analysis of its sub-cohorts was conducted in advance by members of the GBMI. SNPs with low imputation quality (INFO score < 0.3) or minor allele frequency (MAF < 0.001) were excluded from the analysis. After filtering, a total of $5,001,334$ SNPs were consistently represented across all cohorts and subsequently included in the meta-analysis. The Genomic inflation factor(λ) was calculated for each cohort to assess potential population stratification or other bias. All cohorts showed values close to 1, indicating minimal inflation of p-value. No additional genomic control was applied either before or after the meta-analysis.

2.2.2 Meta-analysis

To further combine results across different ancestry groups, we employed the meta-regression approach implemented in the software MR-MEGA [Mägi et al.,]. MR-

MEGA first generates genetic principal components for each cohort based on the summary statistics, and then incorporates these components as covariates in a meta-regression model. This approach enables the separation of different sources of heterogeneity, particularly the detection of ancestry-related heterogeneity in genetic effects. We include one principal component in the meta-regression model. To control for false positives arising from multiple testing, we applied a Bonferroni-adjusted significance threshold of $p = 1 \times 10^{-8}$.

The meta-analysis results were annotated using ANNOVAR through the FUMA platform [Watanabe et al.,]. The Linkage disequilibrium estimates (r^2) was calculated based on the 1000 Genomes reference panel, including all ancestries. We identifies near-independent significant SNPs by clustering all genome-wide significant variants and retaining those with pairwise $r^2 < 0.6$ as independent signals. To map genetic association to potential functional elements, we performed positional mapping, assigning significant SNPs to the nearest genes based on their physical proximity within the genome. To identify potentially novel loci and genes associated with COPD, we compared our results with those reported in the previous meta-analysis [Liu et al.,] as well as with entries curated in the GWAS catalog. Variants and genes not previously associated with COPD were considered novel candidates for further investigation.

2.2.3 Gene set enrichment analysis

From the list of genes identified in the meta-analysis, we researched their biological functions from the GeneCards database. We also performed gene-set enrichment analysis and tissue specificity testing using the GENE2FUNC module within the FUMA platform [Watanabe et al.,]. Detailed descriptions of the statistical methods and underlying datasets used in these analyses are available on the FUMA tutorial page (https://fuma.ctglab.nl/tutorialgene2func). For the gene-set enrichment analysis, hypergeometric tests were conducted to assess whether the identified genes were significantly overrepresented in predefined gene sets and pathways, including those from MSigDB, WikiPathways, and previously reported genes curated in the GWAS catalogue. To control for multiple testing within each category, the Benjamini-Hochberg

Table 2.1: Cohort description

Cohort Population	European	East Asian	Latino	African	All
Cases	58,559	19,044	1,503	1,978	81,084
Controls	995,917	310,689	13,583	27,704	1,347,893
Total	1,054,476	329,733	15,086	29,682	1,428,977
Number of subcohorts	12	4	3	6	16
Number of SNPS	26,338,574	9,984,861	10,374,687	19,514,941	5,001,443
λ	1.0025	1.008	1.003	1.0039	1.102

procedure was applied to control the false discovery rate (FDR). Gene sets with an adjusted p-value less than 0.05 were considered statistically significant.

Tissue-specific expression analysis was conducted based on expression profiles from the GTEx v8 reference database. This analysis tested whether the identified genes showed preferential expression in particular tissues, providing biological insights into the associated loci. Significant tissue enrichments were determined using Bonferroni-corrected p-value threshold of $p < 0.05$.

2.3 Results

We identified a total of 76 near-independent SNPs that reached genome-wide significance ($p < 1 \times 10^{-8}$), as summarized in Table2.2 and Figure 2.1. Of these, 50 variants had odds ratios greater than 1, indicating risk-increasing effects; while 26 had odds ratios less than 1, indicating protective effects. A significant proportion of these SNPS are clustered near a locus in Chromosome 15 2.2. Most of the SNPS are located in the intergenic or intronic regions in the genome, suggesting possible regulatory rules. Based on positional mapping, these variants correspond to 28 genes.

By comparing our findings with those reported in the latest COPD meta-analysis by [Liu et al.,] and entries in the GWAS catalog, we identify 18 potentially novel genes associated with COPD risk. These include RP11-25K21.6, RP11-152C17.1, RNU6-699P, RP11-361D14.2, STAG3, PILRA, ASZ1, CFTR, PBX3, C11orf30, MAPKAPK5, SMAD3, PSMA4, CHRNB4, ADAMTS7, GSDMB, CTC-490E21.13 and LILRB2. We also successfully replicated results at 10 previously reported genes, including AFAP1, FAM13A, KRT18P51, C5orf56, IREB2, HYKK, CHRNA3, CHRNA4, CHRNA5 and

THRA. Multiple intronic variants in FAM13A (rs2464520, rs2464522, rs1585258) demonstrated modest risk increases (OR = 1.04), consistent with prior findings. Similarly, the loci IREB2, HYKK, CHRNA3, CHRNA4, and CHRNA5 cluster on chromosome 15 showed a large number of highly significant signals (e.g., rs12231881, $p = 1.4710^{-39}$; rs12231890, $p = 7.3310^{-55}$; rs12231899, $p = 5.5810^{-59}$). We were unable to replicate some of the SNPs reported by [Liu et al.,], who identified 32 SNPs near 25 genes.

Among the significant SNPs, we identified 19 variants exhibiting significant ancestral heterogeneity($p_{anc-het} < 0.05$), and no significant residual heterogeneity($p_{res-het} > 0.05$). This suggests that the observed differences in genetic effects across populations are primarily driven by population-specific genetic architecture, rather than unaccounted confounders. These 19 variants map to 14 distinct loci, including RNU6-699P, KRT18P51, STAG3:GATS:GATS, PILRA, C11orf30, SMAD3, IREB2, CHRNA5, CHRNA3, CHRNB4, CHRNB4:RP11-160C18.2, THRA, CTC-490E21.13 and LILRB2. We illustrated a subset of these variants using forest plots **??**. Notably, some variants display opposite effects across ancestries. For instance, C11orf30 rs7130588 shows a protective effect in Africans (OR = 0.90), but a risk-increasing effect in both East Asians (OR = 1.06) and Europeans(OR = 1.04).

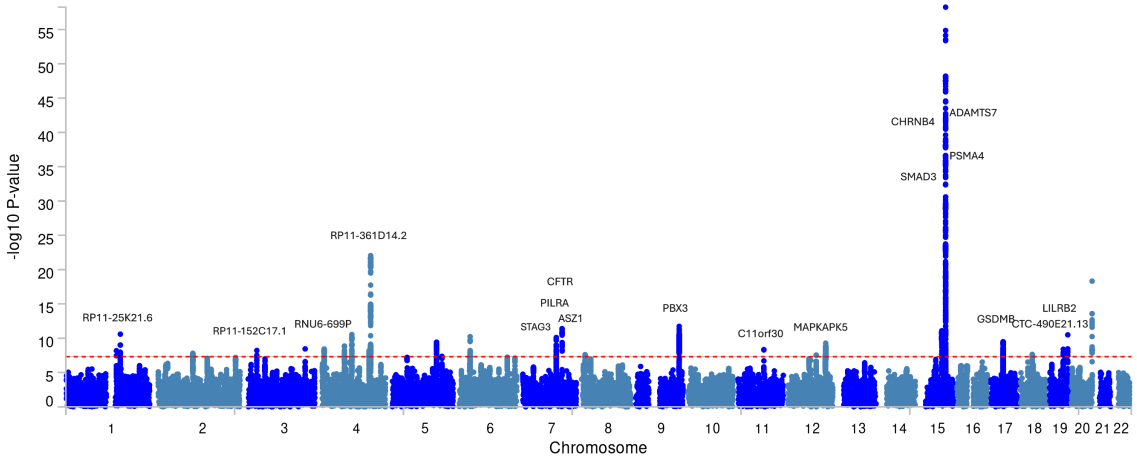


Figure 2.1: Manhattan plot
The Manhattan plot shows the significance level across the whole genome. The x-axis denotes the chromosome and location of SNPs. The y-axis denotes the p-value on a logarithmic scale of 10. The dashed line indicates the significant threshold $p = 1 \times 10^{-8}$. The novel genes are annotated in the plot.

Table 2.2: Independent significant SNPs

Gene	SNP ID	Chr:Pos:A1:A2	Function	OR	95% CI	GWAS p-value	P.value_ancestry_het	P.value_residual_het	Novel Gene
RP11-25K21.6	rs2099684	1:161500130:A:G	ncRNA_intronic	1.03	1.01-1.06	2.56e-11	1.01e-01	1.64e-01	Yes
RP11-152C17.1	rs879394	3:168709843:G:T	intergenic	1.04	1.03-1.04	3.58e-09	2.79e-01	9.81e-01	Yes
AFAP1	rs4488938	4:7850903:A:G	intronic	1.07	1.05-1.09	4.45e-09	5.43e-01	7.62e-01	No
AFAP1	rs73073939	4:7890828:C:T	intronic	1.05	1.03-1.08	3.73e-09	7.81e-01	3.39e-01	No
RNU6-699P	rs35631117	4:67804304:G:T	intergenic	0.95	0.93-0.97	1.38e-09	1.42e-02	1.87e-01	Yes
FAM13A	rs2464520	4:89851019:C:G	intronic	1.04	1.01-1.08	1.21e-10	2.59e-01	3.45e-02	No
FAM13A	rs2464522	4:89860843:A:G	intronic	1.04	1.01-1.07	2.92e-11	3.67e-01	7.13e-02	No
FAM13A	rs1585258	4:89879196:G:T	intronic	1.04	1-1.07	8.60e-11	6.53e-02	4.11e-02	No
RP11-361D14.2	rs4465995	4:145291206:C:T	intergenic	0.95	0.95-0.96	7.41e-10	1.52e-01	8.33e-01	Yes
RP11-361D14.2	rs7666298	4:145401950:A:C	intergenic	0.95	0.92-0.97	1.17e-14	4.20e-01	7.66e-02	Yes
RP11-361D14.2	rs11727583	4:145433002:A:G	intergenic	1.05	1.03-1.06	1.72e-09	3.76e-01	5.42e-01	Yes
RP11-361D14.2	rs1032295	4:145434584:G:T	intergenic	0.95	0.93-0.97	8.51e-15	9.47e-01	1.50e-01	Yes
KRT18P51	rs7663578	4:145468791:A:G	intergenic	0.95	0.93-0.96	1.73e-15	1.43e-01	4.42e-01	No
KRT18P51	rs6831503	4:145471945:A:G	intergenic	0.95	0.92-0.97	1.45e-21	3.76e-01	4.45e-02	No
KRT18P51	rs13141641	4:145506456:C:T	intergenic	0.94	0.92-0.96	9.01e-23	7.05e-01	1.07e-01	No
KRT18P51	rs11407396	4:145516737:C:CA	intergenic	0.94	0.93-0.95	1.43e-12	3.90e-02	7.29e-01	No
C5orf56	rs10066308	5:131790616:A:G	intronic	1.05	1.04-1.06	2.53e-09	1.22e-01	5.92e-01	No
C5orf56	rs3749833	5:131799626:C:T	intronic	1.06	1.02-1.09	6.03e-10	1.43e-02	1.90e-02	No
C5orf56	rs10077785	5:131801158:C:T	intronic	1.05	1.03-1.07	3.80e-10	9.15e-01	2.65e-01	No
STAG3:GATS:GATS	rs35358918	7:99815152:C:CAA	ncRNA_intronic	0.97	0.97-0.97	1.81e-10	2.85e-02	9.99e-01	Yes
PILRA	rs2405442	7:99971313:C:T	exonic	0.98	0.97-0.99	7.88e-11	3.32e-03	7.29e-01	Yes
ASZ1	rs73480763	7:117041941:A:G	intronic	1.10	1.08-1.12	3.99e-12	3.08e-01	7.96e-01	Yes
CFTR	rs34159932	7:117175143:A:G	intronic	1.07	1.05-1.09	8.01e-10	3.93e-01	5.06e-01	Yes
CFTR	rs1025342	7:117217725:C:T	intronic	1.07	1.03-1.1	1.80e-09	9.98e-02	1.68e-01	Yes
PBX3	rs6478715	9:128648473:C:G	intronic	1.06	1.05-1.06	3.46e-11	2.25e-01	8.24e-01	Yes
PBX3	rs143238353	9:128649078:T:TACAACTTGA	intronic	1.06	1.04-1.07	1.86e-12	5.88e-01	6.81e-01	Yes
PBX3	rs10987055	9:128707568:C:T	intronic	1.05	1.04-1.06	8.56e-10	5.17e-01	6.28e-01	Yes
C11orf30	rs7130588	11:76270683:A:G	intergenic	1.02	1-1.05	4.55e-09	6.06e-03	1.47e-01	Yes
MAPKAPK5	rs12231873	12:112331915:C:T	downstream	1.07	0.93-1.22	5.25e-10	5.23e-01	1.46e-02	Yes
SMAD3	rs12231874	15:67470208:A:G	intronic	0.97	0.95-1	1.57e-10	3.77e-02	1.45e-01	Yes
SMAD3	rs12231875	15:67475083:C:T	intronic	0.97	0.96-0.99	1.02e-09	2.43e-02	3.19e-01	Yes
SMAD3	rs12231876	15:67487549:C:G	downstream	0.96	0.94-0.98	8.28e-12	3.32e-01	3.23e-01	Yes
IREB2	rs12231877	15:78711803:G:T	intergenic	1.06	1.02-1.1	5.31e-10	7.79e-02	1.67e-02	No
IREB2	rs12231878	15:78712119:A:T	intergenic	1.08	1.04-1.12	6.61e-10	1.29e-02	7.29e-02	No
IREB2	rs12231879	15:78724256:A:G	intergenic	1.07	1.03-1.12	1.02e-15	1.81e-02	2.54e-03	No
IREB2	rs12231880	15:78724469:C:T	intergenic	1.09	1.04-1.13	2.51e-24	6.94e-01	1.23e-02	No
IREB2	rs12231881	15:78750549:A:G	intronic	1.10	1.04-1.15	1.47e-39	5.74e-02	1.30e-05	No

Table 2.2: Independent significant SNPs *(continued)*

Gene	SNP ID	Chr:Pos:A1:A2	Function	OR	95% CI	GWAS p-value	P.value_ancestry_het	P.value_residual_het	Novel Gene
IREB2	rs12231882	15:78752114:C:T	intronic	1.08	1.05-1.11	1.84e-24	2.45e-01	9.92e-02	No
IREB2	rs12231883	15:78754102:A:G	intronic	1.10	1.05-1.17	2.19e-39	6.31e-03	1.69e-05	No
IREB2	rs12231884	15:78779384:C:T	intronic	1.08	1.06-1.1	1.70e-34	9.82e-01	1.60e-01	No
HYKK	rs12231885	15:78802869:C:T	intronic	0.91	0.88-0.94	1.26e-46	9.87e-01	1.70e-02	No
HYKK	rs12231886	15:78814046:A:G	intronic	0.91	0.87-0.96	1.67e-42	9.52e-01	7.19e-05	No
HYKK	rs12231887	15:78826948:C:T	intronic	0.92	0.9-0.94	5.31e-37	7.62e-01	1.21e-01	No
HYKK	rs12231888	15:78828086:G:T	intronic	1.10	1.1-1.11	6.40e-49	4.51e-01	9.45e-01	No
PSMA4	rs12231889	15:78832349:C:T	upstream	1.09	1.07-1.11	2.86e-23	5.92e-01	8.03e-01	Yes
CHRNA5	rs12231890	15:78849914:C:T	intergenic	1.05	1.05-1.06	1.17e-10	1.02e-07	8.77e-01	No
CHRNA5	rs12231891	15:78867482:A:C	intronic	1.11	1.1-1.12	7.33e-55	1.31e-01	6.52e-01	No
CHRNA3	rs12231892	15:78892661:C:T	intronic	1.09	1.04-1.14	1.45e-30	7.04e-01	4.82e-04	No
CHRNA3	rs12231893	15:78896129:A:G	intronic	1.11	1.07-1.15	3.57e-45	1.04e-01	7.01e-03	No
CHRNA3	rs12231894	15:78897865:C:CT	intronic	0.94	0.92-0.97	1.82e-12	4.08e-13	1.04e-01	No
CHRNA3	rs12231895	15:78901173:G:GA	intronic	1.10	1.1-1.1	1.90e-27	9.07e-01	9.72e-01	No
CHRNA3	rs12231896	15:78909480:A:G	intronic	1.04	0.99-1.09	2.48e-09	1.13e-09	4.33e-05	No
CHRNA3	rs12231897	15:78910258:C:T	intronic	0.91	0.89-0.92	1.88e-43	5.44e-01	4.41e-01	No
CHRNA3	rs12231898	15:78910267:A:C	intronic	0.90	0.88-0.93	2.51e-31	3.24e-02	8.07e-02	No
CHRNA3	rs12231899	15:78911181:C:T	exonic	1.11	1.09-1.14	5.58e-59	4.07e-02	1.18e-01	No
CHRNA3	rs12231900	15:78912472:A:C	intronic	0.89	0.89-0.89	4.41e-17	4.10e-02	9.99e-01	No
CHRNB4	rs12231901	15:78915864:A:G	downstream	1.10	1.06-1.14	6.74e-47	3.32e-01	2.01e-03	Yes
CHRNB4	rs12231902	15:78928399:A:G	intronic	1.09	1.05-1.12	1.12e-22	6.91e-02	5.87e-02	Yes
CHRNB4	rs12231903	15:78934318:C:G	intronic	1.09	1.07-1.12	2.68e-31	5.17e-01	4.22e-01	Yes
CHRNB4	rs12231904	15:78934551:A:G	intronic	1.10	1.07-1.13	1.26e-26	4.96e-02	1.71e-01	Yes
CHRNB4	rs12231905	15:78946633:C:T	intronic	0.93	0.91-0.95	1.62e-13	2.91e-02	4.11e-01	Yes
CHRNB4:RP11-160C18.2	rs12231906	15:78960529:A:C	ncRNA_intronic	1.07	1.05-1.1	4.95e-23	8.86e-02	1.08e-01	Yes
CHRNB4:RP11-160C18.2	rs12231907	15:78961421:A:G	ncRNA_intronic	0.91	0.87-0.95	3.81e-09	1.25e-01	1.34e-01	Yes
CHRNB4:RP11-160C18.2	rs12231908	15:78973356:C:T	ncRNA_intronic	0.90	0.87-0.93	5.77e-10	6.73e-02	3.49e-01	Yes
CHRNB4:RP11-160C18.2	rs12231909	15:78992025:C:T	ncRNA_intronic	1.07	1.03-1.1	3.15e-15	4.71e-02	4.27e-02	Yes
CHRNB4:RP11-160C18.2	rs12231910	15:79007950:C:T	ncRNA_intronic	1.07	1.06-1.09	6.11e-20	6.98e-04	2.96e-01	Yes
ADAMTS7	rs12231911	15:79053284:C:T	intronic	1.07	1.06-1.09	1.11e-17	7.96e-01	6.24e-01	Yes
ADAMTS7	rs12231912	15:79074253:A:G	intronic	1.06	1.02-1.1	6.25e-14	1.74e-01	7.75e-03	Yes
GSDMB	rs12231913	17:38064876:A:G	intronic	0.96	0.94-0.98	3.20e-10	9.01e-01	2.88e-01	Yes
GSDMB	rs12231914	17:38075426:A:G	upstream	1.03	1.02-1.05	1.72e-09	4.77e-01	5.82e-01	Yes
THRA	rs12231915	17:38218773:A:G	UTR5	1.05	1.05-1.06	1.02e-09	3.20e-03	9.64e-01	No
CTC-490E21.13	rs12231916	19:41412185:A:G	intergenic	1.02	1.01-1.04	4.07e-09	1.50e-02	5.40e-01	Yes
LILRB2	rs12231917	19:54783521:C:T	intronic	0.96	0.95-0.98	3.24e-11	3.99e-02	4.64e-01	Yes
CHRNA4	rs12231918	20:61984317:C:T	intronic	0.93	0.92-0.94	4.83e-19	5.21e-01	5.35e-01	No
CHRNA4	rs12231919	20:61986949:C:T	splicing	1.06	1.04-1.09	2.74e-14	9.30e-01	2.30e-01	No

Table 2.2: Independent significant SNPs *(continued)*

Gene	SNP ID	Chr:Pos:A1:A2	Function	OR	95% CI	GWAS p-value	P.value_ancestry_het	P.value_residual_het	Novel Gene
CHRNA4	rs12231920	20:61990878:C:T	intronic	1.07	1.02-1.12	1.82e-09	7.79e-01	4.68e-02	No

Gene: The nearest gene to the SNP; SNP ID: SNP ID; CHR:POS:A1:A2: Chromosome, position, reference allele and alternative allele for the SNP(based on the hr38 build); Function: The functional impact of this SNP on the gene; OR:Odds ratio for COPD risk comparing the reference allele to the alternative allele; CI: 95%confidence interval for the odds ratio; GWAS p-value: two-sided p-value of association from meta-analysis; P.value ancestry het: two-sided p-value of heterogeneity test for ancestral heterogeneity, chi-square test with degrees of freedom of 1; P.value residual het : two-sided p-value of heterogeneity test for residual heterogeneity, chi-square test with degrees of freedom of 2.

To explore the functional relevance of the 18 novel genes associated with COPD risk, we queried the Human Gene Database and summarized the information in Table 2.3. Many of these genes are protein-coding and are implicated in diverse biological processes and disease phenotypes. Several genes are involved in the lung disease, such as CFTR and PSMA4, which are both linked to cystic fibrosis, while GSDMB has linked to childhood-onset asthma. A subset of genes plays roles in immune or inflammatory pathways. For example, PILRA is associated with herpes simplex and inflammatory bowel disease. Four of the novel loci—RNU6-699P, RP11-25K21.6, RP11-152C17.1, and RP11-361D14.2—are non-coding RNAs with little or no functional annotation. Their genome-wide associations with COPD suggest that regulatory mechanisms may contribute to disease risk.

The heatmap 2.5 shows the expression level of novel genes across 30 tissues. We observed no significant tissue-specific expression enrichment after FDR correction 2.7 We didn't find that the discovered genes show significant expression level enrichment in any tissues after Bonferroni correction 2.6. Gene-set enrichment analysis further demonstrates that our COPD genes overlap significantly with the gene set associated with inflammatory bowel disease, encompassing both Crohn's disease and Ulcerative colitis 2.7. We also found significant enrichment in gene sets associated with smoking behaviour, FEV1 variation, asthma, type 2 diabetes, coronary artery disease and other pulmonary diseases 2.6, underscoring the pleiotropic nature of these COPD risk loci.

2.4 Discussion

In this study, we leveraged the ancestrally diverse data from the Global Biobank Meta-analysis Initiative (GBMI) to identify 76 near-independent genome-wide significant SNPs mapping to 28 genes, including 18 novel genes. We employed a meta-regression framework capable of disentangling ancestry-specific heterogeneity from residual confounding, a substantial improvement over conventional random-effect models. This approach revealed 19 SNPs with significant ancestry-dependent genetic effects. Our findings underscore the value of expanding genetic research beyond European-ancestry

Gene	Type	Associated diseases	Gene Ontology(GO)
CFTR	Protein coding	Cystic Fibrosis and Vas Deferens, Congenital Bilateral Aplasia Of	enzyme binding and PDZ domain binding
PSMA4	Protein Coding	Cystic Fibrosis and Parkinson's Disease	endopeptidase activity and threonine-type endopeptidase activity
STAG3	Protein Coding	Premature Ovarian Failure 8 and Spermatogenic Failure 61	binding
GSDMB	Protein Coding	Childhood-Onset Asthma and Oral Squamous Cell Carcinoma	-
PILRA	Protein Coding	Herpes Simplex and Inflammatory Bowel Disease 25	-
MAPKAPK5	Protein Coding	Neurocardiofaciodigital Syndrome and Pulmonary Immaturity	transferase activity, transferring phosphorus-containing groups and protein tyrosine kinase activity
CHRNB4	Protein Coding	Frontotemporal Dementia 1	extracellular ligand-gated monoatomic ion channel activity and ligand-gated monoatomic ion channel activity
LILRB2	Protein Coding	Malaria and Leukemia, Acute Myeloid	signaling receptor activity and protein phosphatase 1 binding
ADAMTS7	Protein coding	Arthritis and Weill-Marchesani Syndrome	peptidase activity and metallopeptidase activity
ASZ1	Protein Coding	Hepatitis E and Male Infertility With Azoospermia Or Oligozoospermia Due To Single Gene Mutation	obsolete signal transducer activity
SMAD3	Protein Coding	Loeys-Dietz Syndrome 3 and Familial Thoracic Aortic Aneurysm And Dissection	DNA-binding transcription factor activity and sequence-specific DNA binding
PBX3	Protein Coding	Myoepithelioma and Leukemia	DNA-binding transcription factor activity and sequence-specific DNA binding
C11orf30	Protein Coding	childhood asthma, polysensitization	-
CTC-490E21.13	RNA Gene (lncRNA)	-	-
RNU6-699P	RNA Gene (lncRNA)	-	-
RP11-25K21.6	RNA Gene (lncRNA)	-	-
RP11-152C17.1	RNA Gene (lncRNA)	-	-
RP11-361D14.2	RNA Gene(lncRNA)	-	-

Table 2.3: Functional summary of novel genes

populations to refine our understanding of disease etiology and to promote equitable advances in precision medicine.

Among our novel findings, several protein-coding genes have clear relevance to pulmonary biology. The CFTR protein, critical for regulating salt and water balance in the body, is directly related to cystic fibrosis. PSMA4, which encodes an -subunit of the 20S proteasome core, is essential for intracellular proteostasis and has also been implicated in cystic fibrosis pathways. C11orf30 (EMSY) participates in DNA damage repair and has been linked to childhood-onset asthma, suggesting a role in airway epithelial integrity. Immune-regulatory genes such as PILRA and LILRB2 highlight the contribution of innate and adaptive signalling pathways to COPD risk. Finally, the identification of non-coding RNAs (e.g., RNU6-699P), suggests the potential importance of transcriptional regulatory mechanisms in COPD etiology and warrants targeted functional follow-up.

Gene-set enrichment analysis revealed that COPD-associated genes are significantly enriched in pathways implicated in other inflammatory disorders, most notably inflammatory bowel diseases, highlighting shared immune-mediated mechanisms that warrant detailed functional characterization. Clinically, incorporating ancestry-specific effect estimates into polygenic risk scores (PRS) substantially enhances risk prediction across diverse populations, thereby mitigating the Eurocentric biases of conventional PRS models. Finally, analyzing drug-target at novel loci (e.g., SMAD3, a TGF- signaling mediator) may inform therapeutic development; however, population-specific efficacy must be validated.

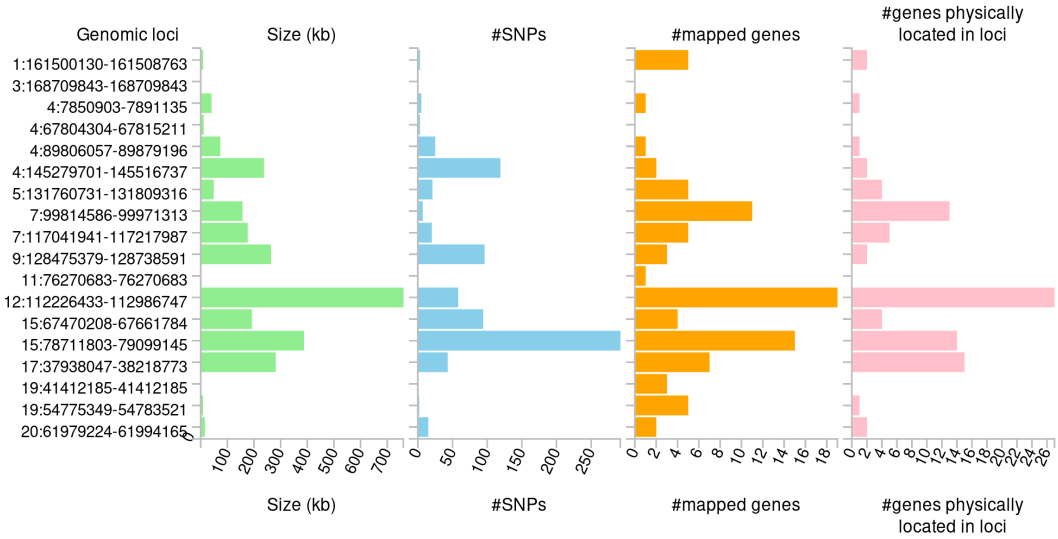
Despite leveraging a multi-ancestry cohort, individuals of European descent still comprise the majority of our sample. In particular, the African and Latin American subgroups remain underpowered, precluding the discovery of genome-wide significant variants within these populations. As biobank resources expand and the cost of sequencing continues to decline, future investigations should aim to recruit larger, more balanced multi-ethnic cohorts. Such efforts will be crucial to uncover population-specific risk loci that may inform tailored prevention and treatment strategies.

In addition, the functional mechanisms through which the identified variants in-

fluence COPD risk remain largely unresolved. To bridge this gap, integrative functional genomics approaches are essential. For example, Expression quantitative trait locus (eQTL) mapping can help link non-coding variants to gene expression changes in disease-relevant tissues, while chromatin conformation assays (e.g., Hi-C, Capture-C) can reveal long-range regulatory interactions. Importantly, given the strong role of environmental exposures in COPD etiology, such as cigarette smoke and air pollution, it is essential to incorporate environmental factors into genetic analyses. Modelling gene–environment interactions may uncover context-dependent genetic effects that are otherwise undetectable in standard GWAS.

2.5 Supplementary Figure

Figure 2.2: Summary per genomic locus



The graph shows the size, number of SNPs, mapped genes, and genes physically located in each genomic locus. The Genomic loci is chromosome: beginning - ending base pair position. There are a total of 18 genomic loci .

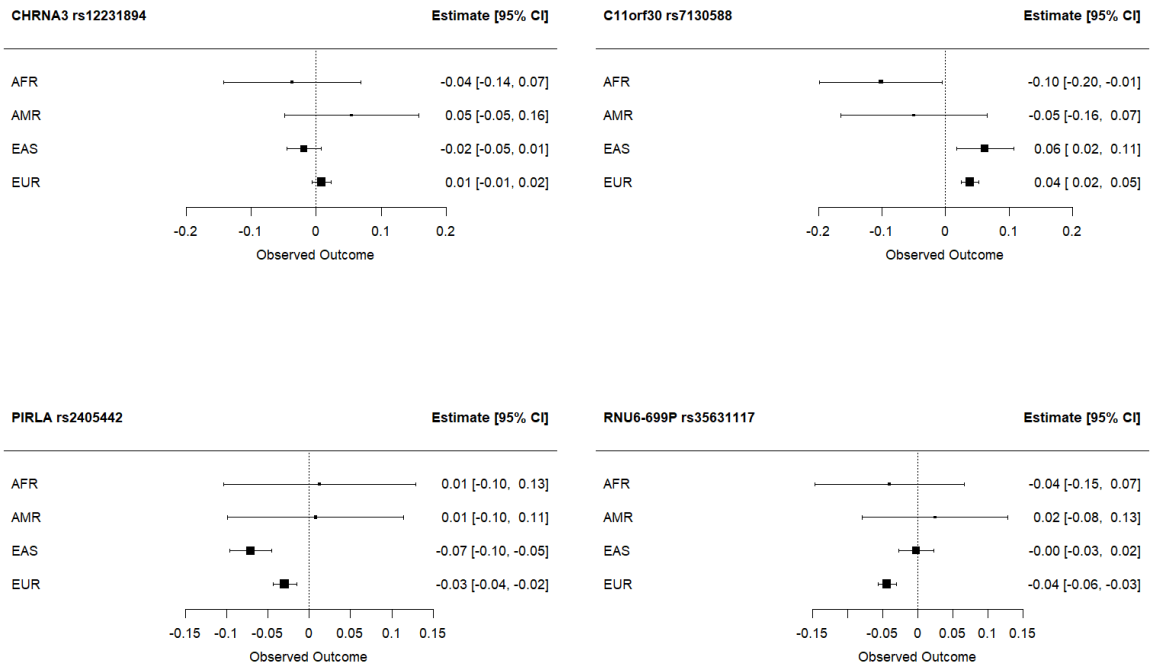
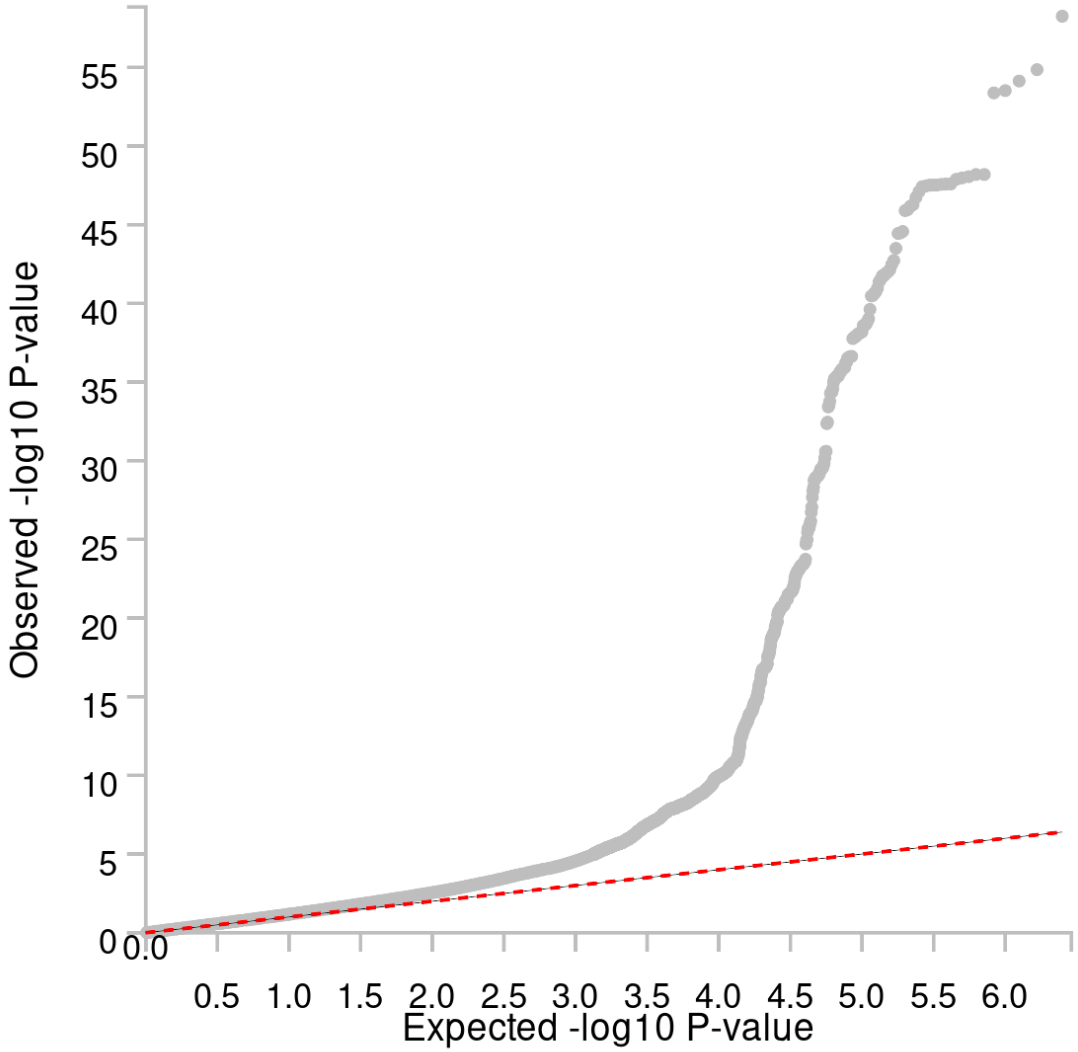


Figure 2.3: The forest plots display the effect estimate (the unit is increase in log-odds) and confidence interval of each subpopulation

Figure 2.4: QQ plot



The x-axis shows the expected p-values under the null hypothesis of no association. and the y-axis shows the observed p-values for each SNP. The p-values are plotted on a log 10 scale

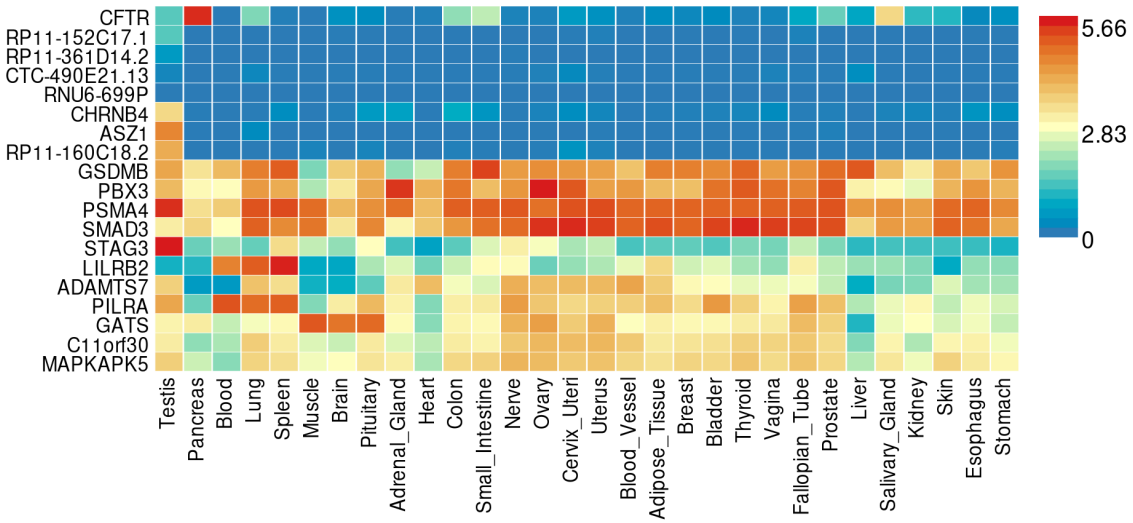


Figure 2.5: Gene expression heatmap
The table shows the expression level of 19 genes across 30 general tissue types according to the Genotype-Tissue Expression (GTEx) database. The colour gradient indicates the average expression per label (log2 transformed).

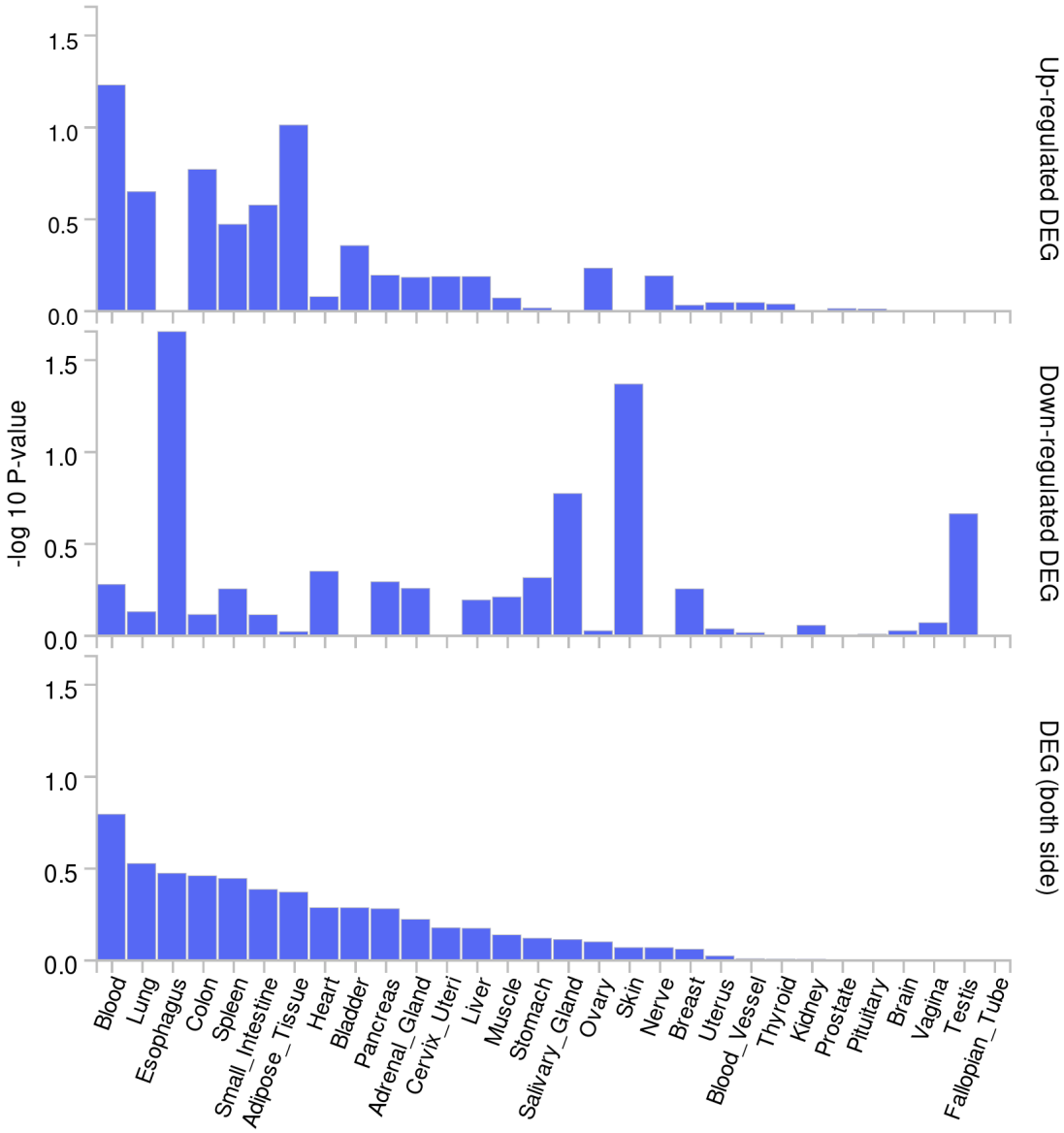


Figure 2.6: Tissue specificity test
The x-axis shows the different tissue types, and the y-axis shows the Bonferroni-corrected p-value in log10 scale from the tissue specificity test.

Figure 2.7: Gene set enrichment analysis: gwas catalog reported gene sets

Bibliography

[Cho et al., 2010] Cho, M. H., Boutaoui, N., Klanderman, B. J., Sylvia, J. S., Ziniti, J. P., Hersh, C. P., DeMeo, D. L., Hunninghake, G. M., Litonjua, A. A., Sparrow, D., et al. (2010). Variants in fam13a are associated with chronic obstructive pulmonary disease. *Nature genetics*, 42(3):200–202.

[Cho et al., 2012] Cho, M. H., Castaldi, P. J., Wan, E. S., Siedlinski, M., Hersh, C. P., Demeo, D. L., Himes, B. E., Sylvia, J. S., Klanderman, B. J., Ziniti, J. P., et al. (2012). A genome-wide association study of copd identifies a susceptibility locus on chromosome 19q13. *Human molecular genetics*, 21(4):947–957.

[Cho et al., 2014] Cho, M. H., McDonald, M.-L. N., Zhou, X., Mattheisen, M., Castaldi, P. J., Hersh, C. P., DeMeo, D. L., Sylvia, J. S., Ziniti, J., Laird, N. M., et al. (2014). Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *The lancet Respiratory medicine*, 2(3):214–225.

[Herrera Ortiz et al., 2022] Herrera Ortiz, A. F., Cadavid Camacho, E., Cubillos Rojas, J., Cadavid Camacho, T., Zoe Guevara, S., Tatiana Rincón Cuenca, N., Vásquez Perdomo, A., Del Castillo Herazo, V., and Giraldo Malo, R. (2022). A practical guide to perform a systematic literature review and meta-analysis. *Principles and Practice of Clinical Research*, 7(4):47–57.

[Liu et al.,] Liu, C., Ran, R., Li, X., Liu, G., Xie, X., and Li, J. Genetic Variants Associated with Chronic Obstructive Pulmonary Disease Risk: Cumulative Epidemiological Evidence from Meta-Analyses and Genome-Wide Association Studies. 2022:3982335.

[MacArthur et al., 2017] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901.

[Mägi et al.,] Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M. I., COGENT-Kidney Consortium, T.-G. C., and Morris, A. P. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. 26(18):3639–3650.

[Pillai et al., 2009] Pillai, S. G., Ge, D., Zhu, G., Kong, X., Shianna, K. V., Need, A. C., Feng, S., Hersh, C. P., Bakke, P., Gulsvik, A., et al. (2009). A genome-wide association study in chronic obstructive pulmonary disease (copd): identification of two major susceptibility loci. *PLoS genetics*, 5(3):e1000421.

[Talamo, 1975] Talamo, R. C. (1975). Basic and clinical aspects of the alpha1-antitrypsin. *Pediatrics*, 56(1):91–99.

[Watanabe et al.,] Watanabe, K., Taskesen, E., family=Bochoven, given=Arjen, p. u., and Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. 8(1):1826.

[Zhou et al., 2021] Zhou, W., Kanai, M., Wu, K., Humaira, R., Tsuo, K., Hirbo, J., Wang, Y., Bhattacharya, A., Zhao, H., Namba, S., et al. (2021). Global biobank metaanalysis initiative: powering genetic discovery across human diseases. medrxiv preprint.