参赛学生姓名：Matthew Tunan Jiang

中　　学：南京外国语学校

省　　份：江苏

国 家 / 地 区：中国

指导老师姓名：周翊峰

指导老师单位：南京医科大学 江苏省人民医院

论 文 题 目：**Association Study of Single Nucleotide Polymorphisms in X-chromosome Inactivation Escape Genes with Susceptibility to Immune Mediated Diseases in Female Populations**

Association Study of Single Nucleotide Polymorphisms in X-Chromosome Inactivation Escape Genes with Susceptibility to Immune-mediated Diseases in Female Populations

Matthew T. Jiang

Nanjing Foreign Language School

Jiangsu, China

September 15, 2025

Name of Participant: Matthew Tunan Jiang

School: Nanjing Foreign Language School

Province: Jiangsu

Country: China

Supervisor: Yifeng Zhou

Supervisor's Affiliation: Jiangsu Provincial Hospital, Nanjing Medical University

Title of the Paper: Association Study of Single Nucleotide Polymorphisms in X-Chromosome Inactivation Escape Genes with Susceptibility to Immune-mediated Diseases in Female Populations

Contents

Abstract

Objective: Immune-mediated diseases are a major global health concern with complex mechanisms involving genetic factors. Gender disparities suggest critical roles for sex-linked genetics. Lyon's hypothesis suggests that about 12% of X-chromosome inactivation escape (XIE) genes escape inactivation in females. Genetic variations in these escape genes may significantly influence immune-mediated susceptibility in females.

Methods: We analyzed UK Biobank genetic data for 20 immune-mediated diseases, including 63,886 female cases and 185,279 female controls. To identify female-specific susceptibility loci, we used logistic regression models to evaluate the association between single nucleotide polymorphisms (SNPs) in XIE genes and immune-mediated diseases. We functionally annotated these loci using expression quantitative trait loci (eQTL) from GTEx database and gene expression data from GEO database. Co-expression and pathway enrichment analyses were conducted to explore the pathogenic pathways associated with the susceptibility genes. We conducted immune infiltration analysis with the CIBERSORT software.

Results: The prevalence of females with at least one immune-mediated disease is 34.48%. In asthma, the SNP rs58199603 at Xp11.23 was associated with increased risk in females (OR = 1.23 [1.15–1.30]). The eQTL analysis indicated that the rs58199603-G allele reduced *CCDC120* expression ($P = 3.36 \times 10^{-5}$), consistent with the decreased *CCDC120* expression observed in asthma samples from GEO. Additionally, rs859595-G in Xq26.1 increased the risk of allergic rhinitis (AR) in females (OR = 1.80 [1.57-2.07]), with *AIFM1* identified as a susceptibility gene($P = 1.17 \times 10^{-4}$) and highly expressed in tissue lesions. Co-expression and pathway enrichment analyses revealed that *CCDC120* and *AIFM1* co-expressed genes were enriched in immune pathways. We found that *CCDC120* expression was positively correlated with the abundance of resting dendritic cells (Cor $r = 0.42$, $P = 6.46 \times 10^{-4}$) and resting NK cells (Cor $r = 0.28$, $P = 2.85 \times 10^{-3}$). While *AIFM1* expression was positively correlated with the abundance of resting dendritic cells (Cor $r = 0.52$, $P =$

1.81×10^{-5}) and Macrophages M0 (Cor $r = 0.39$, $P = 8.12 \times 10^{-4}$).

Conclusions: Our study reveals that XIE-located SNPs and susceptibility genes may influence female-specific risks of asthma and AR, providing insights into sex-based disparities in immune-mediated disease incidence.

Keywords: Female; immune-mediated diseases; Inactivation escape gene; X-inactivation escape; Asthma; Allergic rhinitis; Susceptibility

1. Background

The immune system is indispensable not only for defending the body against infections but also for regulating a wide spectrum of diseases and medical conditions, including cancer, trauma, and organ transplantation. It also plays a central role in the pathogenesis and progression of hypersensitivity disorders such as asthma, dermatitis, and other allergies, as well as systemic and organ-specific immune-mediated diseases (IMDs) such as multiple sclerosis, systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), and diabetes. According to the widely accepted modern view, IMDs encompass a broad spectrum of disorders driven by immune dysregulation. These include: (i) primary or inherited and secondary or acquired immunodeficiencies, and (ii) immunoproliferative disorders, such as immune system malignancies (e.g., multiple myeloma, lymphoma, leukemia), autoimmune diseases (e.g., RA), and immune hypersensitivities (e.g., allergies)[1].

Over the past three decades, the incidence of IMDs has risen sharply, with striking geographic and ethnic variations in prevalence[2]. Globally, up to 10% of the population is affected[3-5]. These conditions manifest through diverse symptoms—including pain, chronic fatigue, dermatological lesions, organ dysfunction, and malignancies—that significantly impair patients' quality of life[3, 6, 7]. Beyond the individual level, IMDs often lead to chronic disability, imposing considerable socioeconomic burdens by reducing productivity and increasing healthcare costs[8].

Epidemiological studies show that while the incidence of IMDs differs across populations, a recurring pattern is their higher prevalence among females[9]. For example, most immune-mediated diseases, despite the wide variability in age of onset, clinical settings, and drug responses, share a common characteristic: the prevalence of female sex[9]. Sex-based differences in immune responses are now recognized as critical determinants of disease onset, clinical course, and therapeutic outcomes. Females display greater susceptibility to IMDs such as Sjögren's syndrome (SS), SLE, hypothyroidism (HT), RA, and Graves' disease (GD)[9-11]. The female-to-male prevalence ratios range from 2:1 to 3:1

in conditions such as multiple sclerosis, RA, and scleroderma, with SLE showing the most pronounced disparity at 9:1[12]. This striking sex bias underscores the urgency of elucidating the biological mechanisms underlying female vulnerability to IMDs.

The etiology of IMDs is multifactorial, involving genetic predisposition, epigenetic regulation (e.g., miRNA activity, DNA methylation, histone modification), environmental triggers (e.g., infections, toxins, dietary factors), hormonal influences, microbiome dysbiosis, and other immune-activating mechanisms[13]. Higher female susceptibility has been appreciated for decades, though its precise foundation is unclear. Genetic factors are central, in light of the fundamental chromosomal distinction between sexes—the female is XX, male is XY[14]. In females, one X chromosome undergoes inactivation during early embryonic development to balance gene dosage; however, this process is incomplete. Approximately 12% of X-linked genes escape inactivation, and another ~15% show variable inactivation depending on individual, tissue, or cell type[15]. These genes that escape X inactivation are termed as X-inactivation escape (XIE) genes[16]. Multipie studies have indicated that these genes may critically shape sex differences in disease susceptibility, with mounting evidence linking genetic and epigenetic alterations in XIE genes to female-biased IMDs[17-19].

Further supporting a genetic basis, IMDs often cluster in families, and multiple immune disorders may co-occur within the same individual[10, 20, 21]. Genome-wide association studies (GWAS) have identified numerous susceptibility loci, many shared across distinct IMDs[22, 23]. For instance, the major histocompatibility complex (MHC) is strongly associated with most IMDs[24], while the CTLA4 locus contributes to RA, type 1 diabetes, and other autoimmune diseases[25, 26]. Importantly, the X chromosome encodes many immune-related genes, including TLR7, TASL, CXCR3, and CD40LG, which are frequently overexpressed in autoimmune conditions[27-29]. Single nucleotide polymorphisms (SNPs) within XIE regions may further modulate disease risk in females. Integrating GWAS data[30] with functional annotation approaches such as expression

quantitative trait loci (eQTL) mapping[31] holds promise for identifying causal variants and clarifying how X-linked factors drive female-biased susceptibility.

Taken together, IMDs impose profound physical, psychological, and socioeconomic burdens worldwide, disproportionately affecting women. Deciphering the mechanisms of female-biased susceptibility is therefore essential for developing targeted prevention and therapeutic strategies, ultimately reducing the global impact of these debilitating conditions.

Objectives

To identify SNPs in the XIE region and female-specific susceptibility loci for immune-mediated diseases by comparing female patients with healthy controls using UK Biobank genetic data.

To functionally annotate susceptibility loci and map these loci to susceptibility genes by integrating eQTL and differential expression data.

To explore how susceptibility loci regulate target genes and influence the pathogenesis of immune-mediated diseases in females using bioinformatics approaches.

Hypothesis

Genetic variations in the XIE region can influence the onset of immune-mediated diseases in female populations by regulating the expression of disease-associated target genes.

2. Methods

2.1 Study population and overall design

All participants included in this study were derived from the UK Biobank (UKB) cohort, and the study design is available online[32]. The UKB is a large-scale, population-based prospective cohort study in the United Kingdom, which recruited over 500,000 participants during its initial phase from 2006 to 2010. The UKB aims to comprehensively and reliably assess the associations between genetic factors, environmental factors, lifestyle, and their combinations with various health outcomes. As an open-access resource, the UKB provides researchers with the opportunity to explore these complex relationships. It focuses on investigating the etiology of various complex diseases in middle-aged and older adults, thereby facilitating improvements in disease prevention, diagnosis, and treatment. For this specific study, information was obtained from 502,507 participants within the UKB cohort.

Our overall study design is summarized in **Figure 1**. In short, we first conducted quality control on the genetic and phenotypic data from the UK Biobank database. We extracted phenotypic and genetic data based on ICD-10. Then, based on logistic model and female population, we evaluate the effect of SNPs in XIE genes between patients and healthy controls. In order to functionally annotate the identified candidate SNPs & genes, we integrated evidence from eQTL analysis, differential gene expression analysis, pathway enrichment analysis, and Immune cell infiltration analysis.

In this study, participants were excluded based on the following criteria: unclear history of IMD diagnosis, unclear sex information, and lack of corresponding genetic testing information. Given that previous research has shown that cancer patients may be more susceptible to immune-related diseases, we also excluded participants with a history of cancer from the analysis to avoid potential bias[3, 7]. Ultimately, a total of 249,165 female participants from the UKB cohort were included (**Figure 2**). Based on the 48 immune-mediated diseases identified in a previous study [3], we excluded those with a sample size of <400, ultimately including 20 immune-mediated diseases in this study (**Figure 2**).

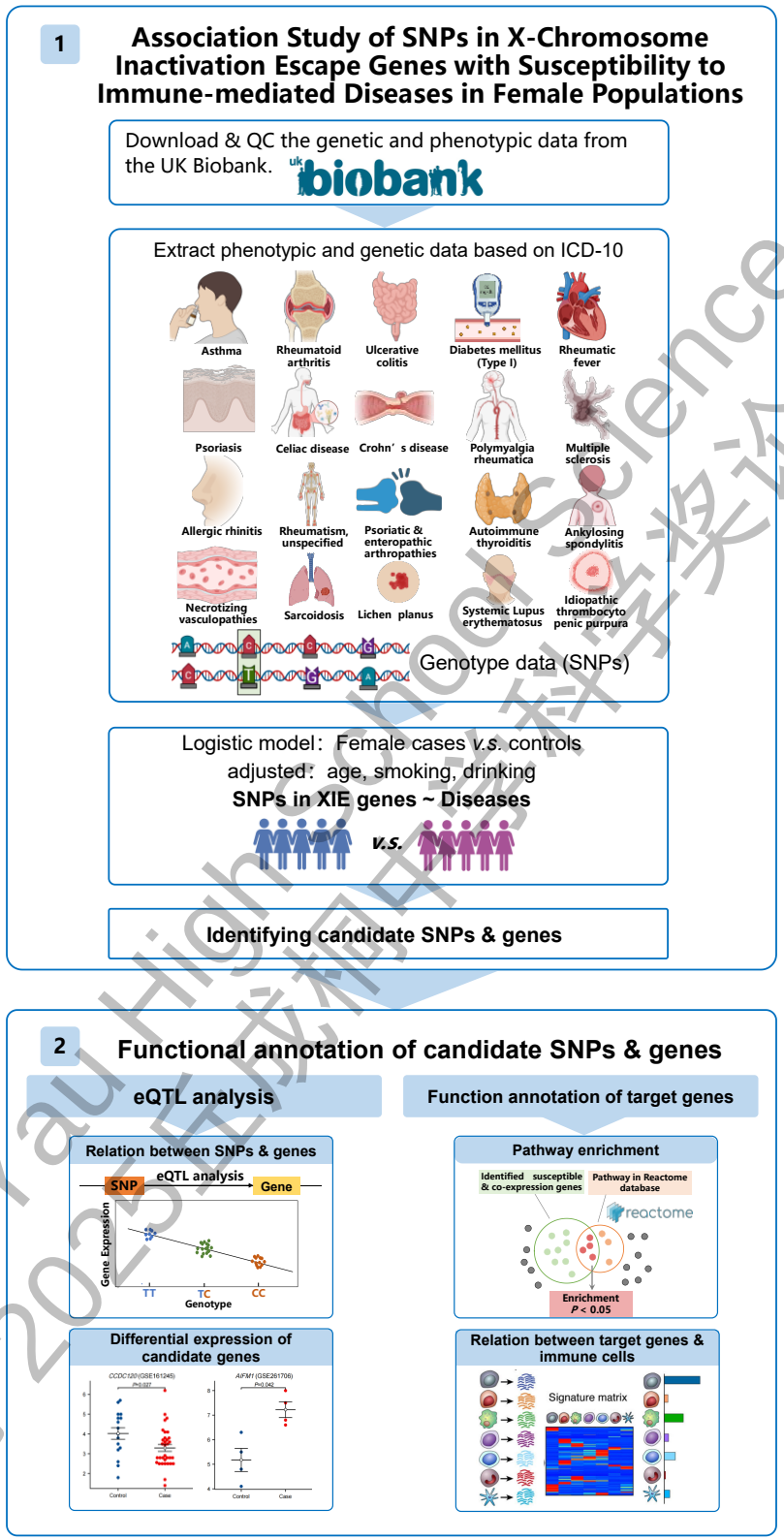


Figure 1. The overall design of this study

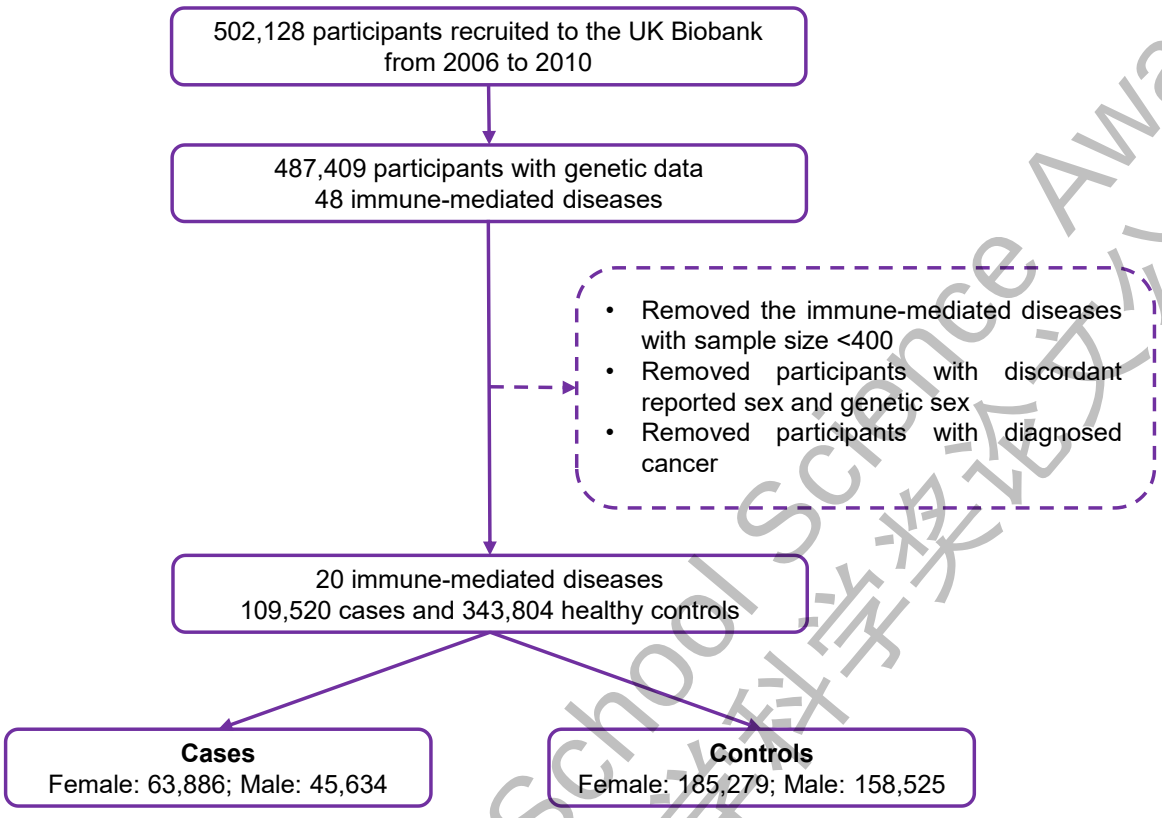


Figure 2. The flowchart of participants inclusion

2.2 Ethics statement

The UK Biobank research had approval from the North West Multi-Centre Research Ethical Committee (REC reference: 21/NW/0157), and all participants provided informed written consent.

2.3 Data extraction

In the UKB cohort, a total of 48 IMDs were identified in the previous study[3]. Utilizing the UK Biobank Research Analysis Platform (https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform), we ultimately included 20 immune-mediated diseases with a sample size of at least 400 in this study (**Figure 2**). Genotyping data and general demographic information corresponding to these diseases were extracted from the database using the 10th revision of the International Classification of Diseases (ICD-10) codes (**Table 1**).

Table 1. Immune-mediated diseases and ICD-10 codes

No	Diseases	ICD-10
1	Asthma	J45, J46
2	Rheumatoid arthritis	M05, M06, M08
3	Ulcerative colitis	K51
4	Diabetes mellitus (Type I)	E10
5	Rheumatic fever / rheumatic heart diseases	I00-I02, I05-I09
6	Psoriasis	L40
7	Celiac disease	K90
8	Crohn's disease	K50
9	Polymyalgia rheumatica	M353
10	Multiple sclerosis	G35
11	Allergic rhinitis	J301-304
12	Rheumatism, unspecified	M790
13	Psoriatic and enteropathic arthropathies	M07
14	Graves' disease / Autoimmune thyroiditis	E05.0, E06.3
15	Ankylosing spondylitis	M45
16	Necrotizing vasculopathies	M31
17	Sarcoidosis	D86
18	Lichen planus	L43
19	Systemic Lupus erythematosus	L93, M32
20	Idiopathic thrombocytopenic purpura	D693

ICD-10, International Classification of Diseases.

2.4 List of X-inactivation escape (XIE) genes

We compiled a list of 108 X-chromosome inactivation escape genes based on previous literature (**Table 2**)[27, 30]. To comprehensively evaluate the genetic variations of these genes, we expanded the regions around each gene by 100 kb upstream and downstream from their start and end positions, according to the hg19 reference genome. This defined the analysis region for each gene. Subsequently, we extracted all single nucleotide polymorphisms (SNPs) within these expanded regions from the UK Biobank database. SNPs, the most common type of genetic variation in the human genome, are widely used in disease association studies.

Table 2. List of collected 108 XIE genes

Gene	Start	End	Type	Source（PMID）
CCDC120	48911101	48927509	protein-coding	26719789
MSL3	11776278	11793870	protein-coding	34426515
ZFX	24167290	24234372	protein-coding	26719789；34426515
ARSD	2822011	2847392	protein-coding	26719789
PNPLA4	7866288	7895780	protein-coding	26719789；34426515
KAL1	8496915	8700227	protein-coding	26719789
WAS	48534985	48549818	protein-coding	34426515
GEMIN8	14026398	14048012	protein-coding	26719789；34426515
OFD1	13752832	13787480	protein-coding	26719789；34426515
GPM6B	13789150	13956757	protein-coding	26719789
CTPS2	16606126	16731059	protein-coding	26719789
GYG2	2746829	2800859	protein-coding	26719789
OTUD5	48779305	48815648	protein-coding	34426515
FUNDC1	44382885	44402247	protein-coding	26719789；34426515
P2RY10	78200829	78217451	protein-coding	34426515
MXRA5	3226606	3264682	protein-coding	26719789
STS	7137497	7272851	protein-coding	26719789；34426515
AIFM1	110909043	111003877	protein-coding	34426515
NAA10	153194695	153200676	protein-coding	34426515
PQBP1	48755195	48760420	protein-coding	34426515
EMD	153607557	153609883	protein-coding	34426515
CDK16	47077259	47089396	protein-coding	26719789；34426515
VGLL1	135614311	135638966	protein-coding	26719789
TAF7L	100523241	100548059	protein-coding	26719789
SASH3	128913955	128929177	protein-coding	34426515
RAB9A	13707244	13728625	protein-coding	26719789
IQSEC2	53262058	53350522	protein-coding	26719789
XG	2670091	2734539	protein-coding	26719789
ZDHHC9	128937264	128977885	protein-coding	26719789
POF1B	84532402	84634748	protein-coding	34426515
USP9X	40944888	41095832	protein-coding	26719789
KDM5C	53220503	53254604	protein-coding	26719789
NXF5	101087085	101112549	protein-coding	26719789
HDHD1	6966961	7066231	protein-coding	26719789
ACE2	15579156	15620271	protein-coding	26719789
EIF2S3	24072833	24096088	protein-coding	26719789；34426515
UBA1	47050260	47074527	protein-coding	26719789；34426515
NLGN4X	5758678	6146904	protein-coding	26719789

KDM6A	44732757	44971847	protein-coding	26719789；34426515
GPR174	78426469	78427726	protein-coding	34426515
IL2RG	70327254	70331958	protein-coding	34426515
HTR2C	113818551	114144624	protein-coding	26719789
HSD17B10	53458206	53461320	protein-coding	34426515
HUWE1	53559057	53713673	protein-coding	26719789；34426515
KANTR	53123327	53196196	protein-coding	26719789；34426515
RIBC1	53449639	53458059	protein-coding	34426515
SMC1A	53401070	53449677	protein-coding	34426515
CLIC2	154505500	154563966	protein-coding	34426515
GPR112	135383122	135519215	protein-coding	26719789
FUNDC2	154254255	154288578	protein-coding	34426515
PPP2R3B	294698	347690	protein-coding	26719789；34426515
DHRSX	2137557	2420846	protein-coding	26719789；34426515
ASMTL	1522032	1572655	protein-coding	26719789；34426515
SLC25A6	1505045	1511617	protein-coding	26719789；34426515
CA5B	15706953	15805747	protein-coding	26719789
ZRSR2	15808595	15841383	protein-coding	26719789
IL1RAPL1	28605516	29974840	protein-coding	34426515
SYAP1	16737755	16783459	protein-coding	26719789
S100G	16668281	16672793	protein-coding	26719789
HCFC1	153213004	153237258	protein-coding	34426515
EIF1AX	20142636	20159962	protein-coding	26719789
TCEANC	13671225	13700083	protein-coding	26719789
GTPBP6	220025	230886	protein-coding	26719789；34426515
FAM133A	92929012	92967273	protein-coding	34426515
MED14	40507558	40595110	protein-coding	34426515
P2RY8	1581465	1656000	protein-coding	26719789；34426515
AP1S2	15843929	15873054	protein-coding	26719789
PLCXD1	192989	220023	protein-coding	26719789；34426515
PRKX	3522411	3631649	protein-coding	26719789；34426515
MAP7D2	20024831	20135035	protein-coding	34426515
IL3RA	1455509	1501578	protein-coding	26719789；34426515
CXorf38	40488285	40506819	protein-coding	26719789；34426515
SHOX	585079	620146	protein-coding	26719789
NAP1L3	92925929	92928567	protein-coding	26719789
CA5BP1	15693055	15721847	pseudogene	26719789；34426515
FAM9C	13053737	13062801	protein-coding	26719789
ZCCHC16	111697727	111700473	protein-coding	26719789
CHM	85116185	85302566	protein-coding	34426515
ASMT	1733894	1761974	protein-coding	26719789

TRAPPC2	13730363	13752754	protein-coding	26719789；34426515
TLR7	12885202	12908499	protein-coding	34426515
MAGEA6	151867214	151870825	protein-coding	34426515
AKAP17A	1710486	1721407	protein-coding	26719789；34426515
RPS4X	71475529	71497150	protein-coding	26719789；34426515
CSF2RA	1387693	1429274	protein-coding	26719789
L1CAM	153126969	153174677	protein-coding	26719789
DMD	31115794	33357558	protein-coding	34426515
TMSB4X	12993227	12995346	protein-coding	34426515
RP11-706O15.5	3809479	3838787	lincRNA	34426515
RP11-706O15.1	3735569	3761898	protein-coding	34426515
ARSH	2924654	2951612	protein-coding	26719789
CRLF2	1314890	1331616	protein-coding	26719789
ZBED1	2404455	2419008	protein-coding	26719789；34426515
DDX3X	41192651	41223725	protein-coding	26719789；34426515
MAGEA3	151934652	151938240	protein-coding	34426515
CD99P1	2527389	2575270	pseudogene	26719789
INE1	47064320	47065264	protein-coding	26719789
JPX	73164159	73290243	lincRNA	26719789；34426515
CXorf28	3189861	3202694	protein-coding	26719789
AK026512	103172005	103174131	antisense	26719789
EOLA2	149097745	149107029	protein-coding	34426515
PPP2R3B-AS1	281725	282586	antisense	26719789
PUDP	6966961	7066231	protein-coding	34426515
RBBP7	16857406	16888537	protein-coding	34426515
TXLNG	16804550	16862642	protein-coding	34426515
SEPTIN6	118749687	118827333	protein-coding	34426515
STK26	131157293	131209971	protein-coding	34426515
TASL	30576941	30595961	protein-coding	34426515

Note: lincRNA, long intergenic non-coding RNA

2.5 Genotyping, imputation and quality control

All GWAS samples were genotyped by high-throughput SNP genotyping arrays (Affymetrix Inc, Santa Clara, CA, USA; or Illumina Inc., San Diego, CA, USA within individual samples) according to the manufacturers' protocols. Quality control (QC) within each sample was implemented at both the individual and SNP levels. At the individual level, sex compatibility was checked by imputing sex from X-chromosome genotype data with

PLINK. The UK Biobank team has utilized the SHAPEIT3 software to determine haplotypes, imputed the genotype data with IMPUTE2 software. Our analysis will include individuals with complete genotyping data and excluded those with inconsistent sex verification results[33].

2.6 Association study

After quality control and imputation, we conducted association analyses using the PLINK (v1.9) software. We calculated the odds ratio (OR) and its 95% confidence interval (95% CI) for the effect of each SNP on disease risk under an additive model. The association analysis was performed using logistic regression, with the imputed genotypes as the independent variables and the case-control status as the dependent variable within female participants. The case group consisted of female patients with at least one of the diagnosed immune-mediated diseases, while the controls include healthy female participants. During the association analysis, we adjusted for the following covariates: age, smoking status and drinking status. We considered a genome-wide corrected significance level of $p < 5\times10^{-8}$ as the threshold for statistical significance.

2.7 Functional annotation of target genes

The GTEx project (https://www.gtexportal.org/) integrates quantitative trait data from multiple molecular phenotypes, including expression quantitative trait loci (eQTL). For the identified susceptibility loci, we utilized GTEx data to annotate genes whose expression levels were associated with genotyping (eQTL genes). This allowed us to identify potential functional candidates at the identified loci by linking genetic variation to gene expression across multiple tissues (**Figure 3**). To further evaluate the functional relevance of the candidate genes, we integrated tissue transcriptome sequencing data from the Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/geo/). Specifically, we assessed the expression levels of candidate genes in corresponding lesion tissues and adjacent normal tissues. We employed the Wilcoxon rank-sum test to statistically evaluate the differences in expression levels between lesion and normal tissues, thereby identifying

genes with significant expression changes that may be implicated in disease pathogenesis.

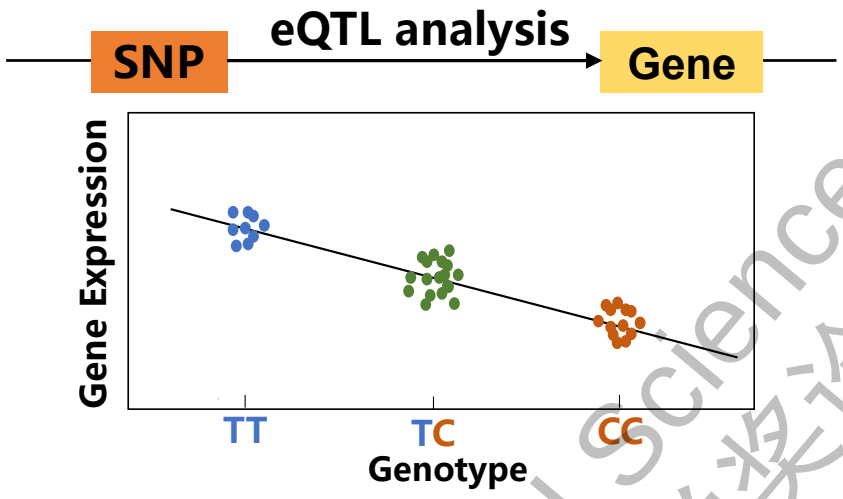


Figure 3. The Schematic diagram of eQTL analysis

2.8 Co-expression and pathway enrichment analysis

To further elucidate the functions and pathways involving the candidate genes, we conducted a co-expression analysis. Specifically, we applied the Spearman correlation test to evaluate the correlation between the expression levels of each candidate gene and those of other protein-coding genes in two datasets: GSE161245 (asthma) and GSE261706 (allergic rhinitis) from the GEO database (Gene Expression Omnibus) (https://www.ncbi.nlm.nih.gov/geo/). This analysis was conducted separately for each dataset to identify co-expressed genes within the specific contexts of these respiratory conditions.

For each candidate gene, we calculated the Spearman correlation coefficient (rho) and the corresponding *p*-value. Genes with a correlation coefficient of rho ≥ 0.60 and a significance level of $p < 0.05$ were considered as co-expressed genes. These co-expressed genes were then subjected to pathway enrichment analysis using the Reactome database (https://reactome.org/). In short, the enrichment algorithm compared the target genes appearance frequency against background genes with that of pathway genes. If target genes

appear more frequent in certain pathway genes than random, we could conclude that target genes are enriched in this pathway (**Figure 4**). This approach enabled us to identify potential biological pathways and functional networks in which the candidate genes might be involved, thereby providing insights into their roles in asthma and allergic rhinitis.

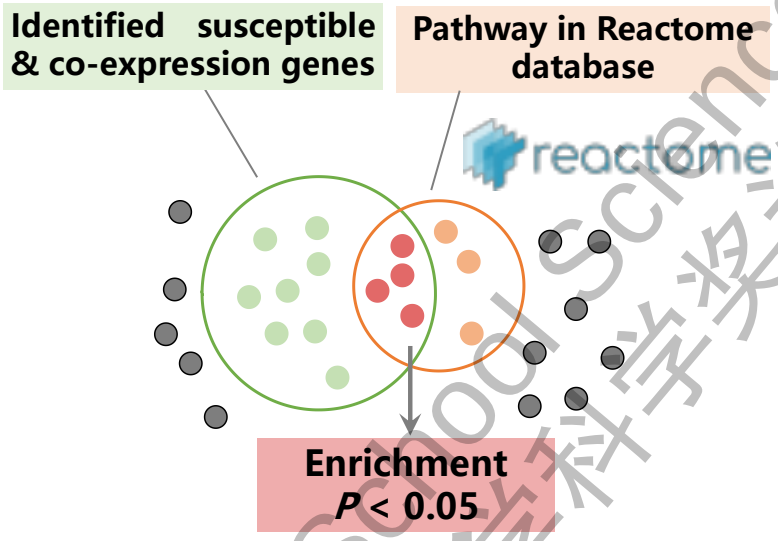


Figure 4. The Schematic diagram of enrichment analysis

2.9 The estimation of immune cell infiltration in lesion tissues

Since current research showed that immune-mediated diseases are often correlated with inflammation and immune system inappropriate response[34], we would like to explore the correlation between the immune cell clusters abundance and target genes expression. In order to determine the immune cell infiltrating levels, we used the CIBERSORT algorithm with 22 types of infiltrating immune cells[35]. the CIBERSORT workflow is based on the deconvolution algorithm, using the 'CIBERSORT' R package in calculation (CIBERSORT R script v1.03; http://cibersort.stanford.edu/, **Figure 5**)[36]..

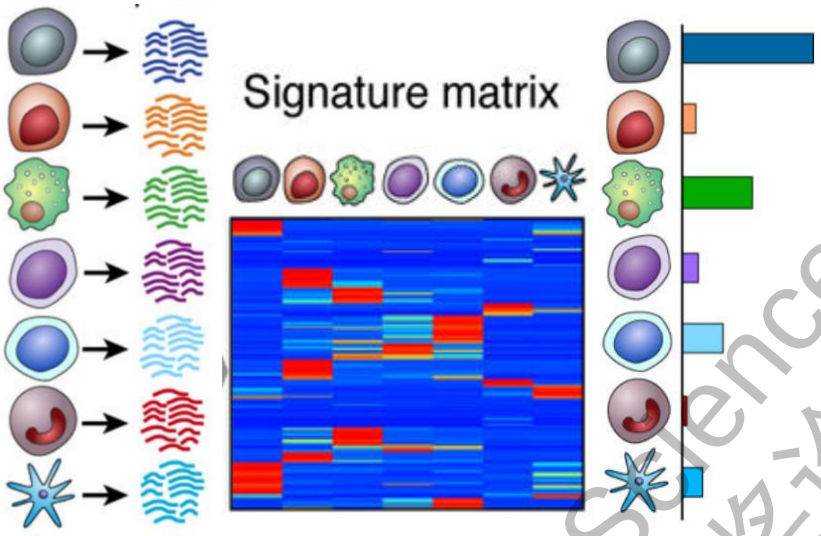


Figure 5. The Schematic diagram of immune infiltration analysis

2.10 Statistical analysis

For continuous variables, the Wilcoxon rank-sum test was used to compare groups. Spearman rank correlation analysis was employed to estimate the correlation between expression levels. The chi-square test and Fisher's exact test were used to assess differences in categorical data between groups. Logistic regression analysis was performed to calculate odds ratios (ORs) for relevant SNPs associated with the disease. PLINK 1.9 and R 3.6.2 were the primary software tools utilized for the analysis. We utilized the LocusZoom plotting tool to analyze the susceptible regions. Regional plots for SNPs within the candidate gene locus, and eQTL box plots were all generated using the ggplot2 package in R. Enrichment analysis was conducted using the clusterProfiler R package[37]. To estimate the correlation between immune cell infiltration in lesion tissues and candidate genes expressions, Spearman correlation test was conducted. The significance level for all statistical tests was set at $\alpha = 0.05$.

3. Results

3.1 The prevalence of immune-mediated diseases in UKB

For the 20 immune-mediated diseases (with sample size ≥400), we included 109,520 cases (Female: 63,886; Male: 45,634) and 343,804 healthy controls (Female: 185,279; Male: 158,525). Sample sizes of included disease are listed in **Table 3**. We found that the total prevalence of immune-mediated diseases is 24.16%, while female (25.65%) suffered more than that in male (22.35%). We also found that asthma has the highest prevalence in both male and female population. Some of the representative prevalence of immune-mediated diseases were showed in **Figure 6**.

Table 3. The prevalence of immune-mediated diseases among the study subjects

Diseases	Female			Male			Total		
	No.of case	No.of control	Prevalence	No.of case	No.of control	Prevalence	No.of case	No.of control	Prevalence
Asthma	27,126	185,279	10.89%	18,299	158,525	8.96%	45425	343,804	10.02%
Rheumatoid arthritis	6,349	185,279	2.55%	2,964	158,525	1.45%	9313	343,804	2.05%
Ulcerative colitis	2,517	185,279	1.01%	2,697	158,525	1.32%	5214	343,804	1.15%
Diabetes mellitus (Type I)	1,989	185,279	0.80%	2,693	158,525	1.32%	4682	343,804	1.03%
Rheumatic fever	3,994	185,279	1.60%	5,005	158,525	2.45%	8999	343,804	1.99%
Psoriasis	2,485	185,279	1.00%	2,806	158,525	1.37%	5291	343,804	1.17%
Celiac disease	2,439	185,279	0.98%	1,308	158,525	0.64%	3747	343,804	0.83%
Crohn's disease	1,541	185,279	0.62%	1,270	158,525	0.62%	2811	343,804	0.62%
Polymyalgia rheumatica	2,427	185,279	0.97%	1,388	158,525	0.68%	3815	343,804	0.84%
Multiple sclerosis	1,434	185,279	0.58%	563	158,525	0.28%	1997	343,804	0.44%
Allergic rhinitis	2,420	185,279	0.97%	2,117	158,525	1.04%	4537	343,804	1.00%
Rheumatism	1,082	185,279	0.43%	172	158,525	0.08%	1254	343,804	0.28%
Psoriatic and enteropathic arthropathies	770	185,279	0.31%	677	158,525	0.33%	1447	343,804	0.32%
Graves' disease / Autoimmune thyroiditis	3,388	185,279	1.36%	966	158,525	0.47%	4354	343,804	0.96%
Ankylosing spondylitis	423	185,279	0.17%	728	158,525	0.36%	1151	343,804	0.25%
Necrotizing vasculopathies	961	185,279	0.39%	522	158,525	0.26%	1483	343,804	0.33%
Sarcoidosis	616	185,279	0.25%	560	158,525	0.27%	1176	343,804	0.26%
Lichen planus	778	185,279	0.31%	366	158,525	0.18%	1144	343,804	0.25%
Systemic Lupus erythematosus	723	185,279	0.29%	145	158,525	0.20%	868	343,804	0.19%
Idiopathic thrombocytopenic purpura	424	185,279	0.17%	388	158,525	0.19%	812	343,804	0.18%

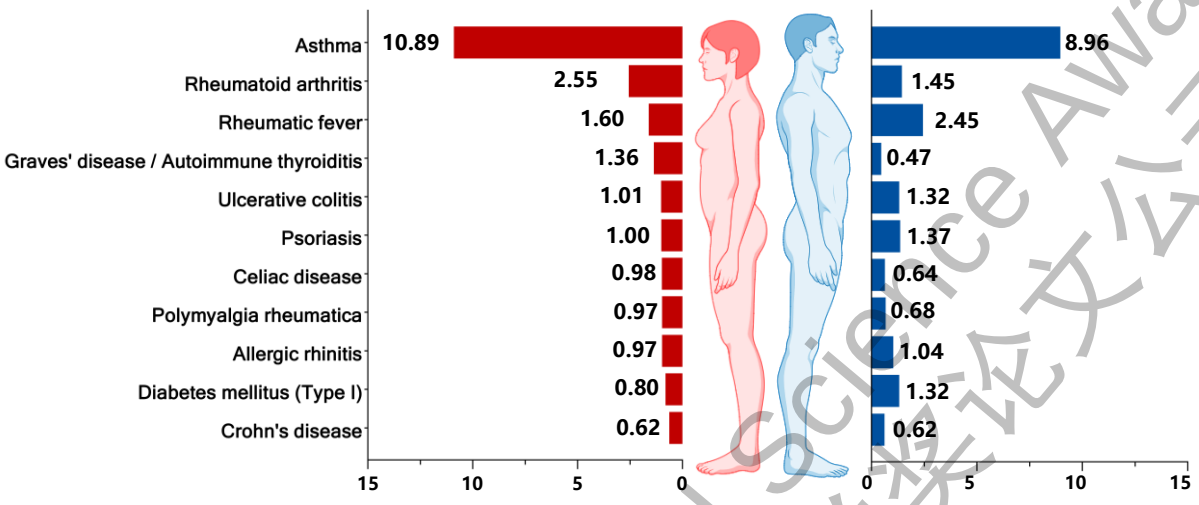
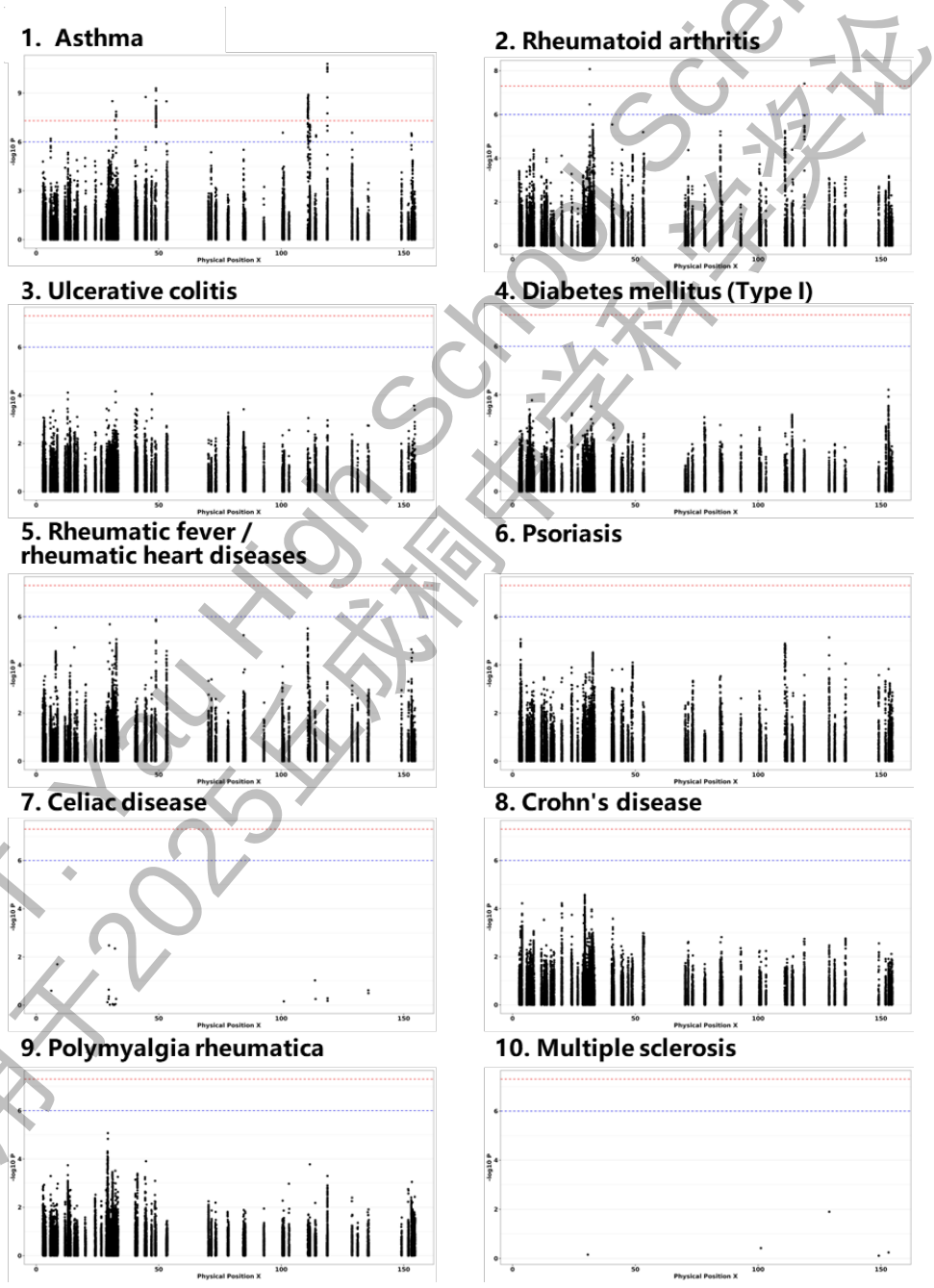


Figure 6 Remarkable differences in the prevalence of immune-mediated diseases between male and female population in the UK Biobank database

3.2 Association results between SNPs and diseases

We extracted the SNPs in the XIE ranges of 20 immune-mediated diseases with each sample sizes ⩾400, and used an additive effect model to assess the correlation between variants and diseases. We found that associations in asthma and allergic rhinitis achieved genome-wide significance ($P < 5 \times 10^{-8}$, **Figure 7**), while no statistical significance was found for the remaining 18 immune-mediated diseases.



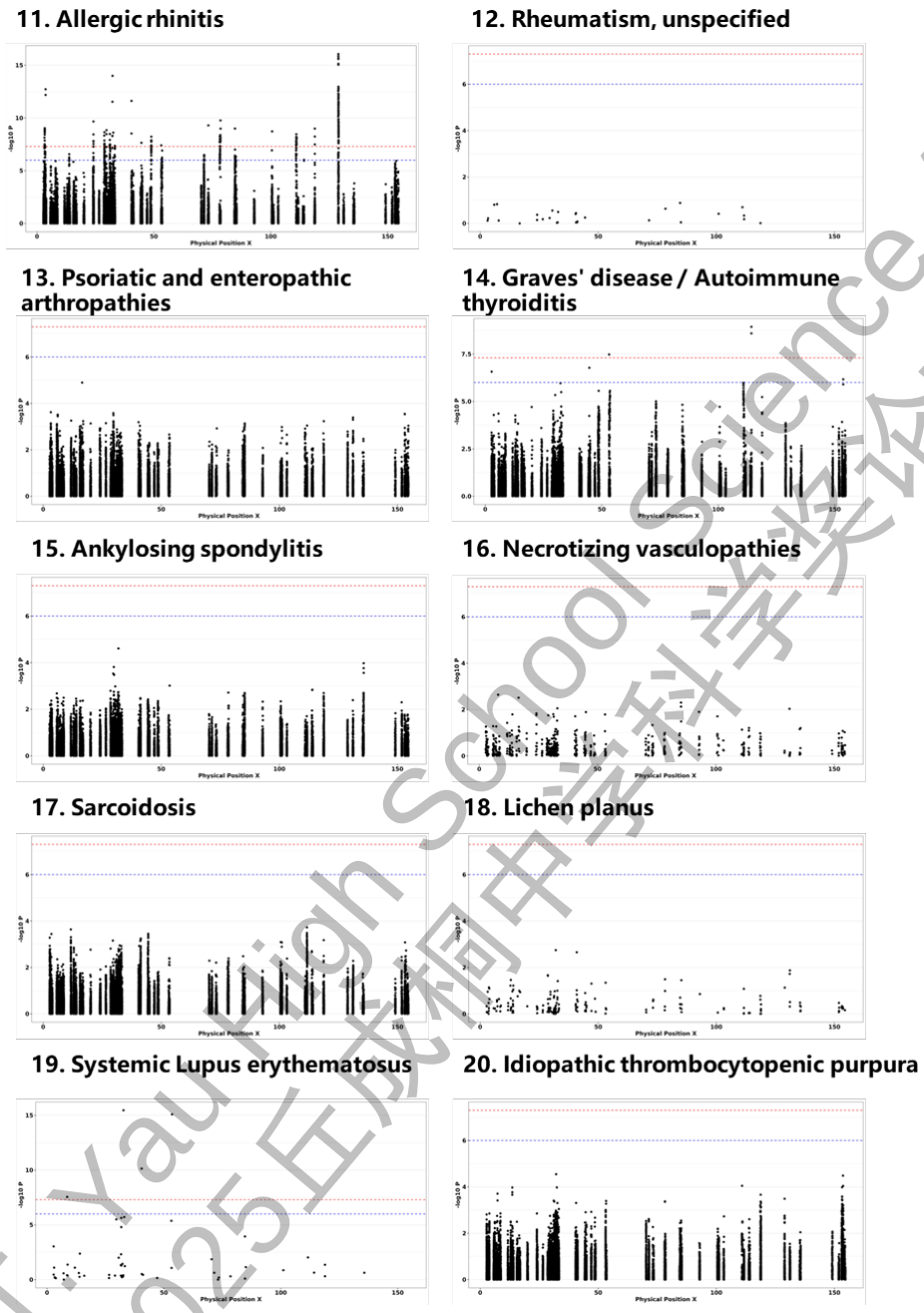


Figure 7. The association results of genetic polymorphisms in the XIE region of 20 immune-mediated diseases with disease risk

Note: Additive model was calculated the effect of genetic polymorphisms in the XIE region of the X chromosome in female patients and healthy controls for 20 immune-mediated diseases. X-axis: X chromosome coordinate (Mb). Y-axis: -log10 *P*-value of the association model.

3.3 Functional annotations of candidate genes

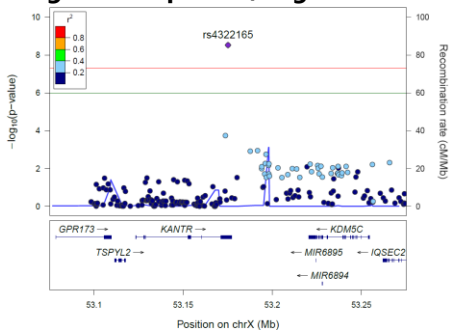
Next, we analyzed the susceptible regions using locus zoom plots to pinpoint the susceptible genes. We found 7 regions in asthma reached genome significance, and the two lowest *P*-value tag SNPs are rs2873179 (OR = 1.25, 95%CI: 1.17-1.33, $P = 1.60\times10^{-11}$, **Figure 8, Table 4**) and rs58199603 (OR = 1.23, 95%CI: 1.15-1.31, $P = 4.73\times10^{-10}$, **Figure 8, Table 4**). In allergic rhinitis, 16 regions reached genome significance, and the two lowest *P*-value tag SNPs were rs859595 (OR= 1.80, 95%CI: 1.57- 2.07, $P= 8.16\times10^{-17}$, **Figure 9, Table 5**) and rs6418643 (OR = 1.74, 95%CI = 1.52-2.01, $P = 9.30\times10^{-15}$, **Figure 9, Table 5**).

Table 4. Association of tag SNPs in each region with asthma in females

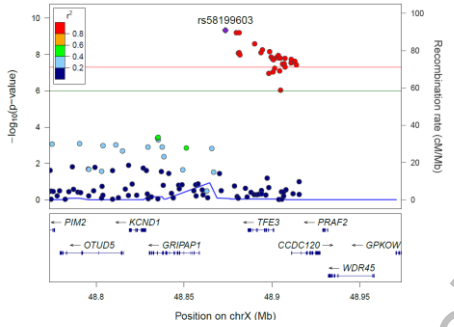
Tag SNP	Position	A1	A2	Cytoband	Gene	Gene type	MAF_{Case}	$MAF_{Control}$	OR (95%CI)	*P* value
rs4322165	53175182	T	C	Xp11.22	*KANTR*	ncRNA_exonic	0.015	0.011	1.22(1.14-1.30)	3.01E-09
rs58199603	**48873505**	**G**	**T**	**Xp11.23**	***CCDC120***	**intergenic**	**0.014**	**0.011**	**1.23(1.15-1.31)**	**4.73E-10**
rs12012959	44672725	T	C	Xp11.3	*DUSP21*	intergenic	0.018	0.014	1.20(1.13-1.28)	1.76E-09
rs5972616	32620898	G	A	Xp21.1	*DMD*	intronic	0.017	0.013	1.18(1.12-1.25)	1.28E-08
rs73617042	31084936	A	G	Xp21.2	*FTHL17*	intergenic	0.014	0.011	1.24(1.16-1.33)	2.89E-09
rs5985357	111037392	C	T	Xq23	*TRPC5*	intronic	0.028	0.023	1.14(1.10-1.20)	1.16E-09
rs2873179	**118841335**	**T**	**C**	**Xq24**	***45906***	**intergenic**	**0.016**	**0.012**	**1.25(1.17-1.33)**	**1.60E-11**

Note: OR: odds ratio; 95%CI: 95% confidence interval; ncRNA: non-coding RNA

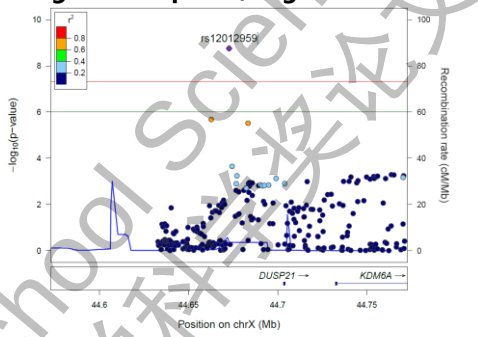
Region 1: Xp11.22, Tag: rs4322165



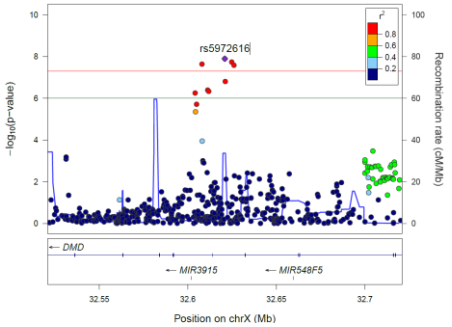
Region 2: Xp11.23, Tag: rs58199603



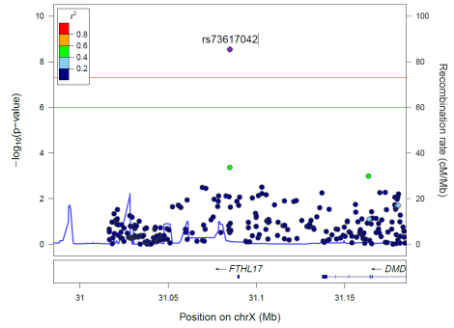
Region 3: Xp11.3, Tag: rs12012959



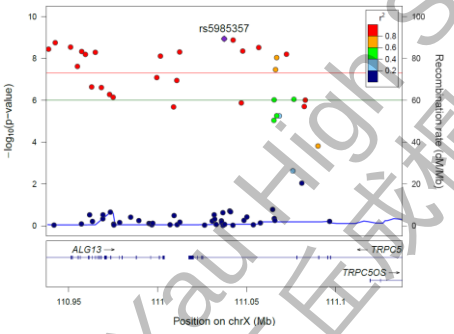
Region 4: Xp21.1, Tag: rs5972616



Region 5: Xp21.2, Tag: rs73617042



Region 6: Xq23, Tag: rs5985357



Region 7: Xq24, Tag: rs2873179

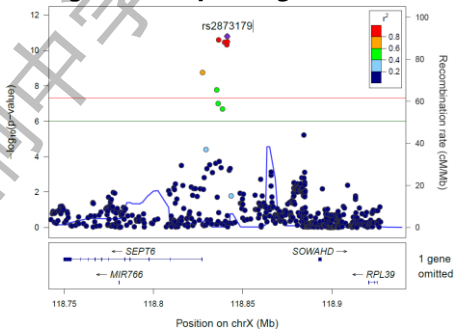


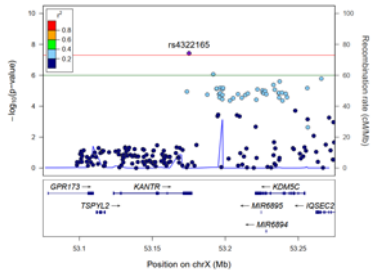
Figure 8. The regional plot of association results in asthma

Table 5. Association of tag SNPs in each region with allergic rhinitis in females

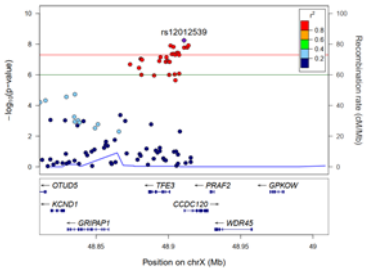
Tag SNP	Position	A1	A2	Cytoband	Gene	Gene type	MAF_{Case}	$MAF_{Control}$	OR (95%CI)	*P* value
rs4322165	53175182	T	C	Xp11.22	*KANTR*	ncRNA_exonic	0.021	0.011	1.60(1.35-1.89)	3.79E-08
rs12012539	48910840	G	T	Xp11.23	*CCDC120*	upstream	0.021	0.011	1.60(1.37-1.88)	5.66E-09
rs12012959	44672725	T	C	Xp11.3	*DUSP21*	intergenic	0.024	0.014	1.57(1.34-1.84)	2.22E-08
rs6610423	40415869	G	A	Xp11.4	*ATP6AP2*	intergenic	0.023	0.013	1.72(1.44-2.05)	3.04E-09
rs6418643	**32316546**	**T**	**C**	**Xp21.1**	***DMD***	**intronic**	**0.036**	**0.020**	**1.74(1.52-2.01)**	**9.30E-15**
rs7887863	29736687	C	T	Xp21.2	*IL1RAPL1*	intronic	0.048	0.032	1.46(1.29-1.65)	1.42E-09
rs1482953	28853998	T	C	Xp21.3	*IL1RAPL1*	intronic	0.045	0.030	1.43(1.28-1.61)	2.11E-09
rs7877712	24145844	G	A	Xp22.11	*ZFX-AS1*	intergenic	0.055	0.037	1.46(1.30-1.64)	2.18E-10
rs5961722	3655543	A	G	Xp22.33	*PRKX*	intergenic	0.025	0.012	1.80(1.54-2.11)	1.78E-13
rs12688161	73304834	T	C	Xq13.2	*FTX*	ncRNA_intronic	0.022	0.012	1.68(1.42-1.97)	5.09E-10
rs6619628	78480587	T	C	Xq21.1	*GPR174*	intergenic	0.040	0.025	1.56(1.36-1.79)	1.84E-10
rs181047651	84681838	A	C	Xq21.2	*POF1B*	intergenic	0.028	0.016	1.61(1.38-1.88)	9.32E-10
rs202021509	100591862	AT	A	Xq22.1	-	-	0.021	0.011	1.73(1.45-2.08)	2.00E-09
rs59107836	110938661	G	T	Xq23	*ALG13*	intronic	0.037	0.022	1.42(1.26-1.59)	3.32E-09
rs2873171	118829479	A	G	Xq24	*SEPT6*	intergenic	0.034	0.020	1.54(1.34-1.77)	9.88E-10
rs859595	**128907796**	**G**	**A**	**Xq26.1**	***XPNPEP2***	**intergenic**	**0.032**	**0.015**	**1.80(1.57-2.07)**	**8.16E-17**

Note: OR: odds ratio; 95%CI: 95% confidence interval; ncRNA: non-coding RNA

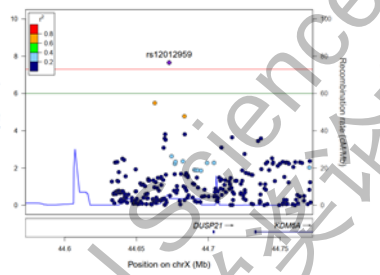
Region 1: Xp11.22, Tag: rs4322165



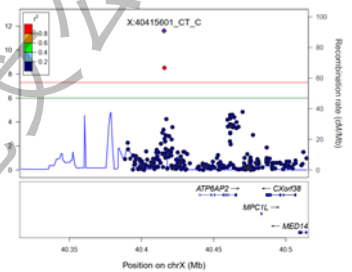
Region 2: Xp11.23, Tag: rs12012539



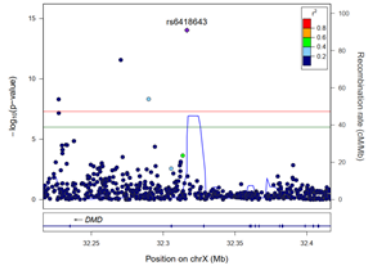
Region 3: Xp11.3, Tag: rs12012959



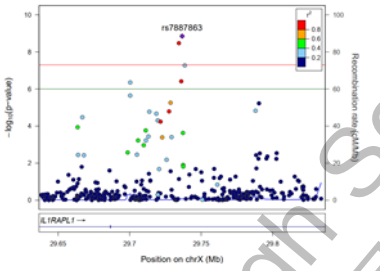
Region 4: Xp11.4, Tag: rs34565123



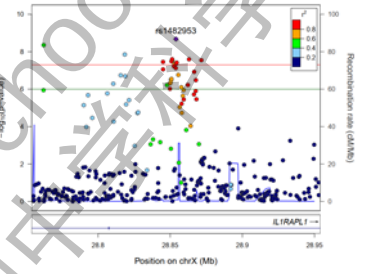
Region 5: Xp21.1, Tag: rs6418643



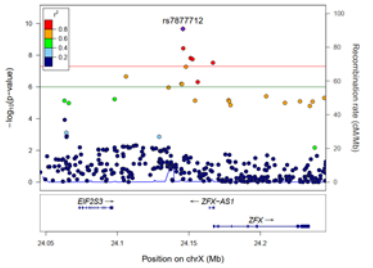
Region 6: Xp21.2, Tag: rs7887863



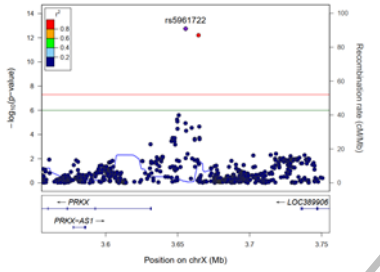
Region 7: Xp21.3, Tag: rs1482953



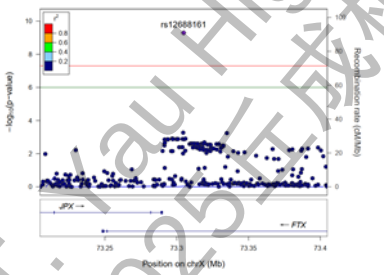
Region 8: Xp22.11, Tag: rs7877712



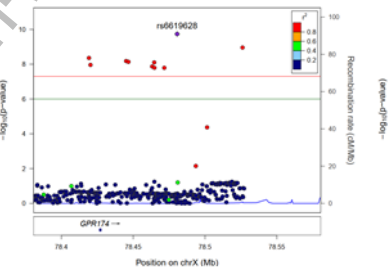
Region 9: Xp22.33, Tag: rs5961722



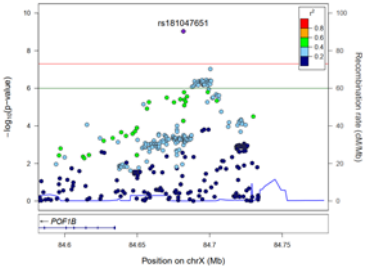
Region 10: Xq13.2, Tag: rs12688161



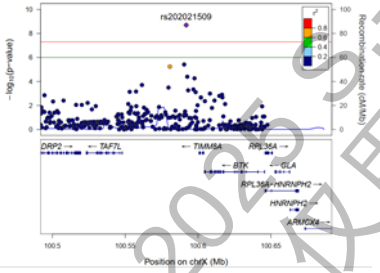
Region 11: Xq21.1, Tag: rs6619628



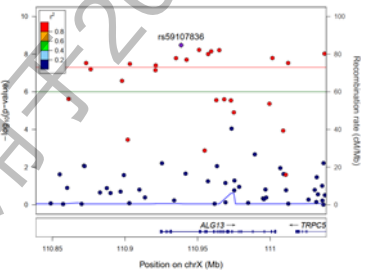
Region 12: Xq21.2, Tag: rs181047651



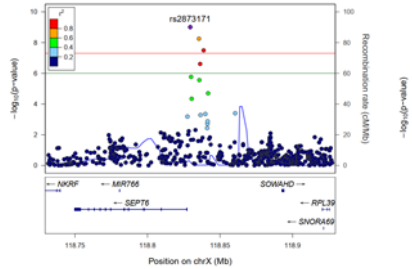
Region 13: Xq22.1, Tag: rs202021509



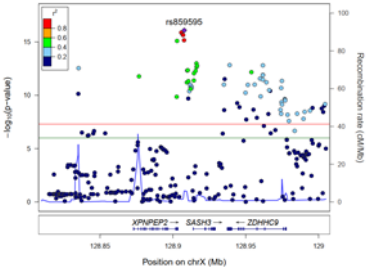
Region 14: Xp23, Tag: rs59107836



Region 15: Xp24, Tag: rs2873171



Region 16: Xq26.1, Tag: rs859595



3.4 eQTL and differential expression genes

In order to further elucidate the function of susceptible SNPs, we seek to integrate evidence from eQTL (GTEx database) & differential expressed analysis (GEO datasets). We identified risk allele G of rs58199603 in asthma was related with *CCDC120* expression decrease (β = -0.30, $P = 3.36\times10^{-5}$, **Figure 10A**) , while risk allele of rs859595-G in allergic rhinitis was related with *AIFM1* expression decrease (β = -0.15, $P = 1.17\times10^{-4}$, **Figure 10B**). The differential expression analysis results show that, the expression level of *CCDC120* is decreased in asthma tissues (P = 0.027, **Figure 10C**), while *AIFM1* was increased in allergic rhinitis tissues (P = 0.042, **Figure 10D**).

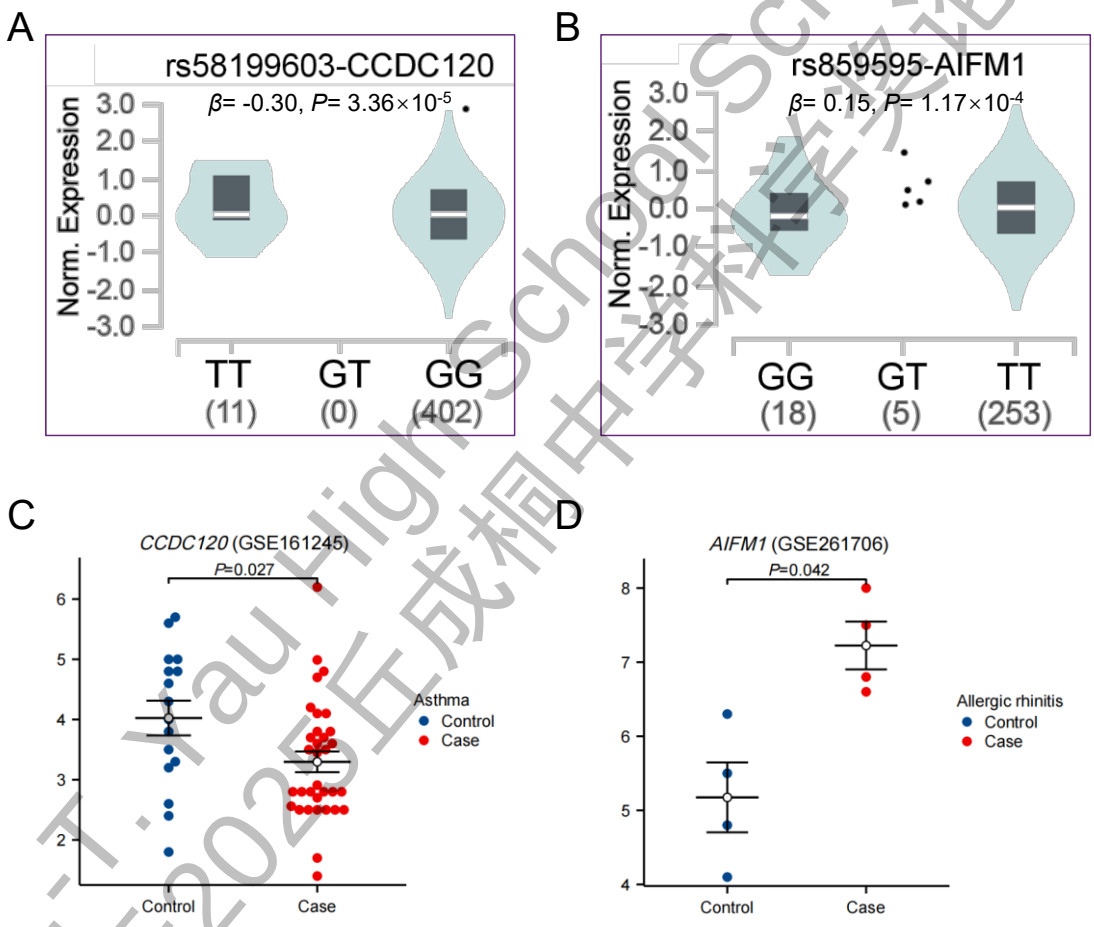


Figure 10. eQTL and differential expression genes of candidate genes

(A) eQTL analysis of rs58199603 in asthma;

(B) eQTL analysis of rs859595 in allergic rhinitis;

(C) Differential expression analysis of *CCDC120* in asthma GSE161245 datasets;

(D) Differential expression analysis of *AIFM1* in allergic rhinitis GSE261706 datasets;

3.5 Co-expression and Pathway enrichment of susceptible genes

Next, we perform co-expression and pathway enrichment analysis to study which important intracellular pathways the candidate gene is involved in. We found that *CCDC120* related genes were enriched in immune-related pathway like cytokine signaling in immune system, innate immune system and interleukin-10 system (**Figure11A**), while co-expressed genes with *AIFM1* were enriched in regulation of cell growth and cell adhesion, such as regulation of cell growth, collagen formation and anti-apoptosis (**Figure11B**).

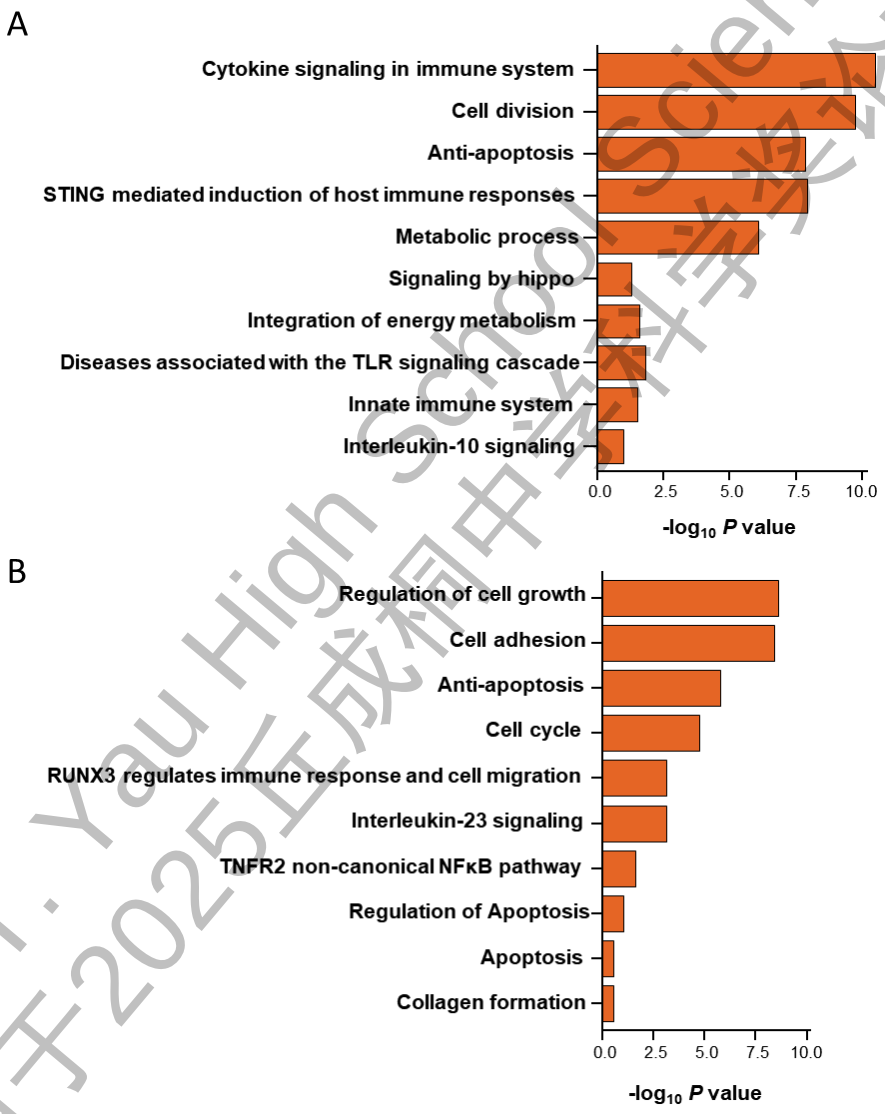


Figure 11. Pathway enrichment analysis of candidate genes

(A) Pathway enrichment analysis of *CCDC120* in asthma.

(B) Pathway enrichment analysis of *AIFM1* in allergic rhinitis.

3.6 Relationship between target genes and immune cells

Since immune-mediated diseases are often originated from dysfunctional immune cells abundance. We then used CIBERSORT, a deconvolution algorithm-based tool, to infer immune cell composition from bulk RNA-seq data in asthma (GSE161245, with 17 mild asthma cases, 5 moderate asthma cases, 17 severe asthma cases) and allergic rhinitis (GSE261706, with 4 cases). The average proportions of inferred immune infiltration were estimated in asthma (**Figure 12A**) and allergic rhinitis (**Figure 12B**) respectively. Furthermore, in order to identify the immune cell clusters which may related with candidate susceptible genes, we calculate the correlation between asthma candidate gene *CCDC120* (**Figure 13A**), allergic rhinitis gene *AIFM1* (**Figure 13B**) and immune cell infiltration levels respectively. We found that asthma candidate gene *CCDC120* expression level was positive correlated with the abundance of resting dendritic cells (Cor $r = 0.42$, $P = 6.46 \times 10^{-4}$) and resting NK cells (Cor $r = 0.28$, $P = 2.85 \times 10^{-3}$). We also found that allergic rhinitis susceptible gene *AIFM1* expression was positive correlated with the abundance of resting dendritic cells (Cor $r = 0.52$, $P = 1.81 \times 10^{-5}$) and Macrophages M0 (Cor $r = 0.39$, $P = 8.12 \times 10^{-4}$).

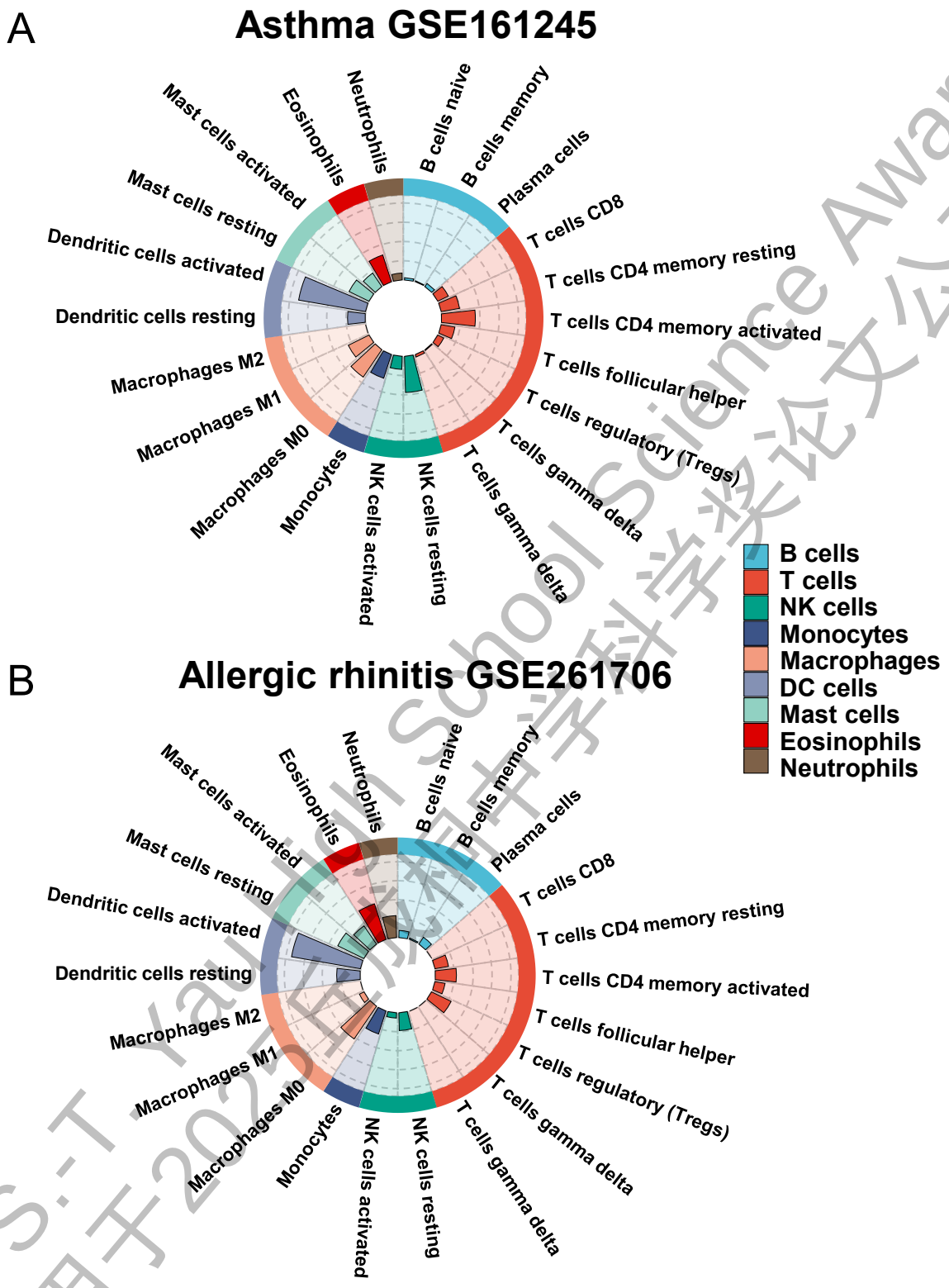


Figure 12 Inferred average immune infiltration abundance in asthma and allergic rhinitis samples by CIBERSORT algorithm

(A) Inferred average immune infiltration abundance in asthma

(B) Inferred average immune infiltration abundance in allergic rhinitis

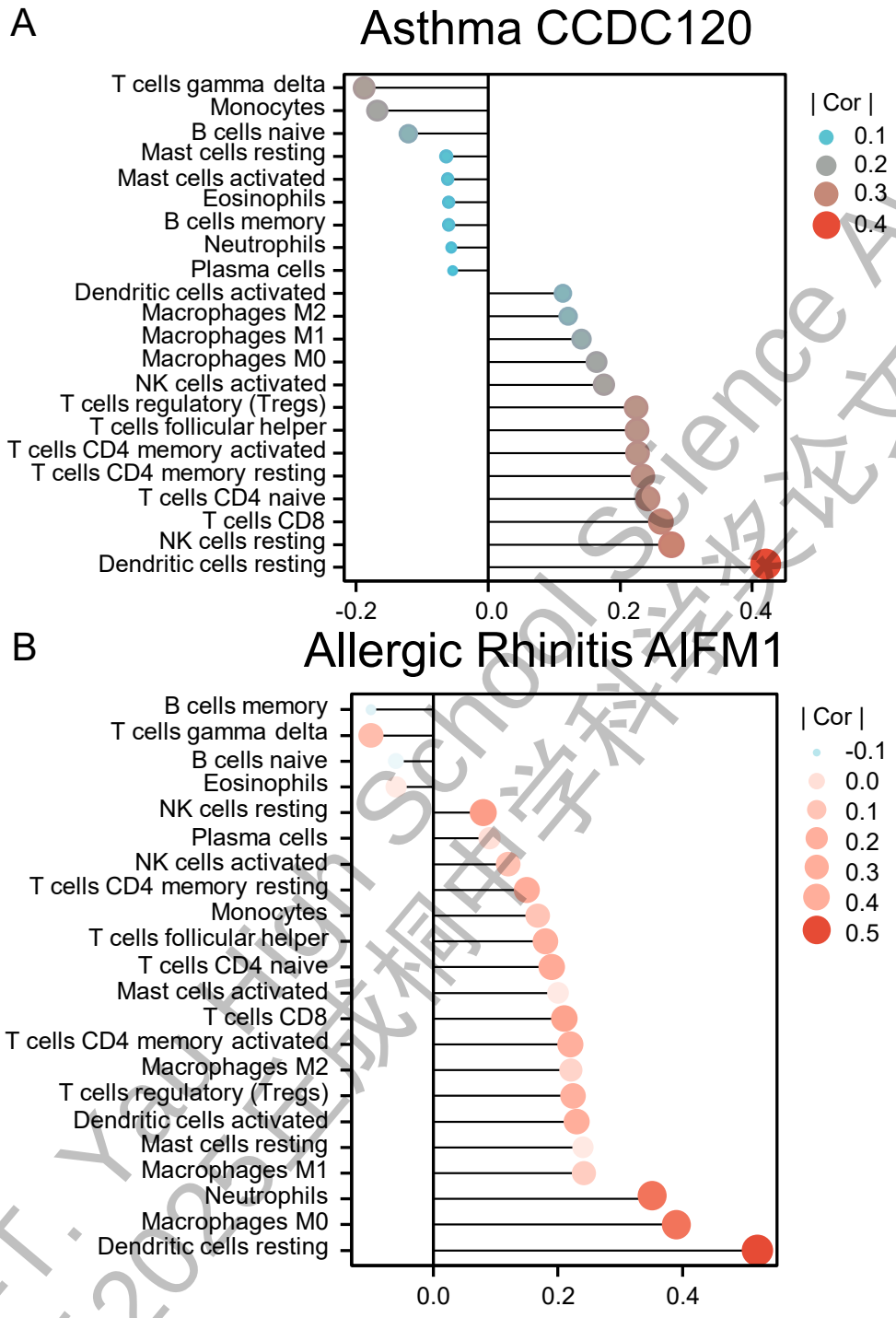


Figure 13 Correlation between susceptible genes expressions and inferred immune infiltration abundance in asthma and allergic rhinitis samples

(A) Correlation between expression of *CCDC120* and inferred immune infiltration abundance in asthma

(B) Correlation between expression of *AIFM1* and inferred immune infiltration abundance in allergic rhinitis

4. Discussion

Although the XIE has been implicated in immune regulation, the contribution of its gene expression changes to sex-specific susceptibility to IMDs remains insufficiently understood. In this study, we conducted an association analysis using UK Biobank data to systematically evaluate the relationships between XIE genes and 20 IMDs. Our findings suggest that two genes—CCDC120 and AIFM1—may be linked to asthma and allergic rhinitis in females, respectively, with both showing correlations to immune cell infiltration. These results provide new genetic and immunological insights into the mechanisms underlying sex-based differences in IMD susceptibility.

In this study, we found candidate gene *CCDC120* plays a role in asthma. Previous literature has shown that the protein encoded by *CCDC120* contains a coiled-coil domain[38]. This protein is widely expressed in various tissues, including skin and esophagus, and may be involved in cell signal transduction and protein complex formation. Additionally, *CCDC120* is associated with cell maintenance and neurite outgrowth. *CCDC120* functions as a centrosome-associated protein and participate in the hierarchical assembly of subdistal appendages of centrioles, thereby contributing to process of cell division and ciliogenesis[38]. Another study indicated that *CCDC120* can bind to the Arf6 guanine nucleotide exchange factor cytohesin-2 (*CYTH2*) and transport along neurites, thereby promoting neurite outgrowth[39]. This funding suggests that the gene is closely associated with intracellular transport and cell morphology regulation. In a study on the osteoporosis spectrum in Turkey, Tuysuz et al. identified *CCDC120* as a potential candidate pathogenic gene for this disease[40], supporting the view that *CCDC120* may play a potential role in the occurrence of human diseases.

Asthma is a chronic inflammatory airway disease, and its pathological process is closely related to the infiltration and activation of immune cells such as Th2 cells, mast cells, and eosinophils, as well as the inflammatory responses mediated by these cells[41]. Our study further analyzed the correlation between *CCDC120* and immune cells, revealing a positive correlation between *CCDC120* and dendritic cells and resting NK cells. The results indicate that *CCDC120* may influence the local immune infiltration in the airways by modulating immune cell infiltration or the release of inflammatory factors, thereby promoting the pathological progression of asthma. These findings provide new evidence from a gene–immune interaction perspective regarding the role of *CCDC120* in asthma, further supporting its potential function in the establishment

and maintenance of the airway inflammatory microenvironment.

We also found that elevated expression of candidate gene *AIFM1* may related to allergic rhinitis. *AIFM1* is a key regulator involved in apoptosis and primarily encodes a mitochondrial flavoprotein[42]. This protein typically resides in the mitochondrial intermembrane space and participates in energy metabolism and redox regulation. However, if the cell is under stress or damaged, the mitochondrial membrane potential decreases, leading to the release of *AIFM1* from mitochondria and translocation into the nucleus. Within the nucleus, *AIFM1* directly binds to DNA, triggering large-scale fragmentation and ultimately inducing apoptosis[43]. Apart from apoptosis regulation, the abnormal expression or mutation of AIFM1 may also impair mitochondrial function, thereby affecting the physiological processes of cells[44]. Previous studies have shown that *AIFM1* mutations can lead to a broad spectrum of clinical manifestations, involving degeneration of various components of the central and peripheral nervous systems, and may cause various diseases such as hearing loss, cerebellar ataxia, and Cowchock syndrome[45]. Simultaneously, dysregulated expression of *AIFM1* is also closely associated with poor outcomes in cancers. For example, *AIFM1* can promote lung cancer progression by enhancing mitochondrial respiration and oxidative phosphorylation, and its high expression is correlated with poor prognosis in patients with non-small cell lung cancer[46]. Wang et al. also reported that *AIFM1* is highly expressed in uterine carcinosarcoma and significantly associated with unfavorable prognosis and enhanced immune cell infiltration[47], suggesting its potential role in influencing disease progression through the regulation of the tumor microenvironment. In the correlation analysis between *AIFM1* and immune cells, we found that *AIFM1* is associated with resting dendritic cells and Macrophages M0. The inflammatory immune responses represent the core pathogenic mechanism of allergic rhinitis, and the initiation and persistence of this disease involve the coordinated action of multiple immune cells. Specifically, Th2 cells and their cytokines enhance epithelial barrier permeability and promote eosinophil infiltration into the nasal mucosa, while the eosinophils recruited to the nasal mucosa can release potent inflammatory substances, becoming the key factor driving the inflammatory response[48, 49]. Therefore, we speculate that *AIFM1* may participate in regulating this inflammatory process by influencing immune cell function, apoptotic processes, or mitochondrial metabolism. The precise regulatory mechanisms involved remain to be fully elucidated.

5. Conclusion

Our study identified susceptibility loci and genes in X-chromosome inactivation escape region, which may potentially influence the risk of asthma and allergic rhinitis in female populations, and could help to reveal the differences in immune-mediated disease onset between males and females (**Figure 14**).

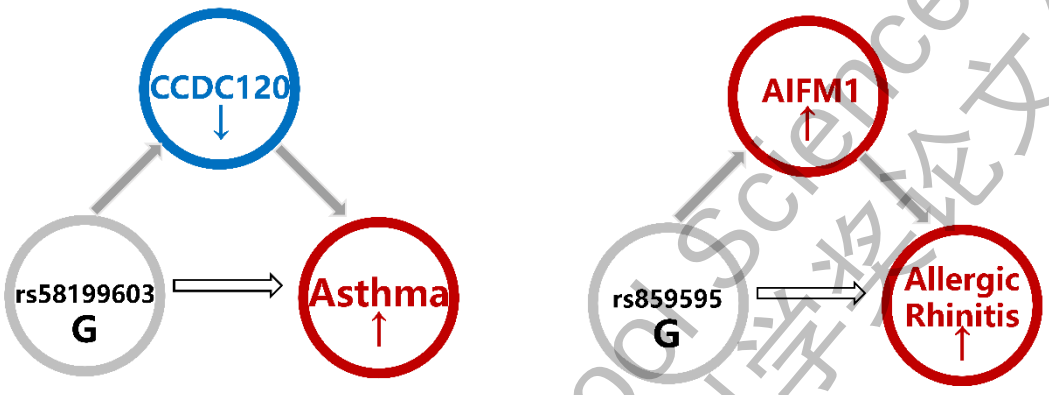


Figure 14. Hypothesis to explain the biological mechanisms

6. Summary

Immune-mediated diseases (IMDs) are characterized by high relapse rates, incurability, and complex etiology, imposing a substantial burden on public health and socioeconomic systems. Notably, most IMDs exhibit a significantly higher prevalence in females than in males. However, the biological mechanisms underlying this sex-based disparity remain unclear, and the potential role of XIE in female susceptibility to IMDs has not been systematically investigated. Therefore, this study explores the association between XIE regions and susceptibility to IMDs in females to advance the understanding of sex-specific differences in IMDs pathogenesis.

This study was primarily conducted using the UK Biobank database. After quality control, 20 IMDs were included for analysis. Logistic regression analysis was performed between 108 reported XIE genes and these 20 diseases, and significant loci were functionally annotated using eQTL data from the GTEx database. The results revealed that rs58199603 in the Xp11.23 region was associated with asthma risk in females, and its G allele correlated with reduced expression of *CCDC120*. Meanwhile, the rs859595-G allele in the Xq26.1 region was significantly associated with an increased risk of allergic rhinitis in females. This variant is associated with the downregulation of *AIFM1* expression. These findings suggest a potential link between XIE and the development of IMDs.

To further investigate the signaling pathways and immune cell infiltration involving the susceptibility genes, co-expression analysis, pathway enrichment analysis, and immune infiltration analysis were conducted. We found that genes associated with *CCDC120* were significantly enriched in multiple immune-related pathways, including cytokine signaling and innate immune regulation. Moreover, the expression level of *CCDC120* was positively correlated with the abundance of resting dendritic cells and resting NK cells. In contrast, genes co-expressed with *AIFM1* were primarily enriched in pathways such as cell growth regulation and anti-apoptosis, and its expression level showed a positive correlation with the infiltration of resting dendritic cells and macrophages M0. This part of the study preliminarily reveals the potential mechanisms by which XIE genes contribute to the pathogenesis of IMDs in females from both functional pathway and immune microenvironment perspectives.

In summary, this study systematically investigated the role of XIE regions in susceptibility to immune-mediated diseases in females, providing new insights into this

field. The findings underscore the importance of developing targeted prevention and treatment strategies to reduce the disease risk in the female population.

7. Limitations and future work

However, our study has several limitations. First, although the genetic association analysis employed in this study effectively reduces the possibility of reverse causality to some extent, we still cannot completely rule out the potential influence of unmeasured confounding factors on the results. Second, the functional annotation of candidate genes relies primarily on eQTL-based methods and lacks experimental validation through functional assays to elucidate the specific mechanisms by which these genes regulate disease development and immune cell infiltration. Third, the genetic data and association results are predominantly derived from the UK Biobank database, which consists mainly of individuals of European ancestry. Therefore, the conclusions require further validation in populations of diverse racial backgrounds.

Our study provides important evidence for early screening and risk assessment of IMDs in females, while also offering new research directions for subsequent mechanistic exploration and clinical practice. Future work should incorporate both in vitro and in vivo functional experiments - such as CRISPR screening, dual-luciferase reporter assays, and animal models - to validate the roles of candidate functional genes in disease progression and to elucidate the signaling pathways involved in regulating the immune microenvironment. This will help uncover the complete molecular biological pathway from genetic variation to disease phenotype. Furthermore, the relationships between these genetic markers and disease severity, progression rate, and treatment response should be evaluated to explore their potential clinical utility as prognostic tools or diagnostic biomarkers.

8. References

1. Shurin MR, Smolkin YS. Immune-Mediated Diseases II Congress: summary. **J Immunotoxicol 2008**, 5(2)**:** 159-162.
2. Collaborators GI. Global, regional, and national incidence of six major immune-mediated inflammatory diseases: findings from the global burden of disease study 2019. **EClinicalMedicine 2023**, 64**:** 102193.
3. He MM, Lo CH, Wang K, Polychronidis G, Wang L, Zhong R*, et al.* Immune-Mediated Diseases Associated With Cancer Risks. **JAMA Oncol 2022**, 8(2)**:** 209-219.
4. El-Gabalawy H, Guenther LC, Bernstein CN. Epidemiology of immune-mediated inflammatory diseases: incidence, prevalence, natural history, and comorbidities. **J Rheumatol Suppl 2010**, 85**:** 2-10.
5. Agrawal M, Shah S, Patel A, Pinotti R, Colombel JF, Burisch J. Changing epidemiology of immune-mediated inflammatory diseases in immigrants: A systematic review of population-based studies. **J Autoimmun 2019**, 105**:** 102303.
6. Russell AS, Gulliver WP, Irvine EJ, Albani S, Dutz JP. Quality of life in patients with immune-mediated inflammatory diseases. **J Rheumatol Suppl 2011**, 88**:** 7-19.
7. Stewart DR. Insights Into Immune-Mediated Disease and Cancer Risk-Delivering on the Promise of UK Biobank Big Data. **JAMA Oncol 2022**, 8(2)**:** 219-220.
8. Whittaker PG, Morgan MR, Dean PD, Cameron EH, Lind T. Serum equilin, oestrone, and oestradiol levels in postmenopausal women receiving conjugated equine oestrogens ('Premarin'). **Lancet 1980**, 1(8158)**:** 14-16.
9. Cooper GS, Stroehla BC. The epidemiology of autoimmune diseases. **Autoimmun Rev 2003**, 2(3)**:** 119-125.
10. Cooper GS, Bynum ML, Somers EC. Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. **J Autoimmun 2009**, 33(3-4)**:** 197-207.
11. Feng Z, Liao M, Zhang L. Sex differences in disease: sex chromosome and immunity. **J Transl Med 2024**, 22(1)**:** 1150.
12. Ngo ST, Steyn FJ, McCombe PA. Gender differences in autoimmune disease. **Front Neuroendocrinol 2014**, 35(3)**:** 347-369.
13. Lahita RG. Sex and gender influence on immunity and autoimmunity. **Front Immunol 2023**, 14**:** 1142723.
14. Wilkinson NM, Chen HC, Lechner MG, Su MA. Sex Differences in Immunity. **Annu Rev Immunol 2022**, 40**:** 75-94.
15. Navarro-Cobos MJ, Balaton BP, Brown CJ. Genes that escape from X-chromosome inactivation: Potential contributors to Klinefelter syndrome. **Am J Med Genet C Semin Med Genet 2020**, 184(2)**:** 226-238.
16. Flaquer A, Rappold GA, Wienker TF, Fischer C. The human pseudoautosomal regions: a review for genetic epidemiologists. **Eur J Hum Genet 2008**, 16(7)**:** 771-779.
17. Haupt S, Caramia F, Herschtal A, Soussi T, Lozano G, Chen H*, et al.* Identification of cancer sex-

disparity in the functional integrity of p53 and its X chromosome network. **Nat Commun 2019**, 10(1)**:** 5385.

18. Li CH, Prokopec SD, Sun RX, Yousif F, Schmitz N, Subtypes PT*, et al.* Sex differences in oncogenic mutational processes. **Nat Commun 2020**, 11(1)**:** 4330.
19. Rubin JB, Lagas JS, Broestl L, Sponagel J, Rockwell N, Rhee G*, et al.* Sex differences in cancer mechanisms. **Biol Sex Differ 2020**, 11(1)**:** 17.
20. Bao YK, Weide LG, Ganesan VC, Jakhar I, McGill JB, Sahil S*, et al.* High prevalence of comorbid autoimmune diseases in adults with type 1 diabetes from the HealthFacts database. **J Diabetes 2019**, 11(4)**:** 273-279.
21. Bogdanos DP, Smyk DS, Rigopoulou EI, Mytilinaiou MG, Heneghan MA, Selmi C*, et al.* Twin studies in autoimmune disease: genetics, gender and environment. **J Autoimmun 2012**, 38(2-3)**:** J156-169.
22. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C*, et al.* Pervasive sharing of genetic effects in autoimmune disease. **PLoS Genet 2011**, 7(8)**:** e1002254.
23. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S*, et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. **Nature 2015**, 518(7539)**:** 337-343.
24. Matzaraki V, Kumar V, Wijmenga C, Zhernakova A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. **Genome Biol 2017**, 18(1)**:** 76.
25. Chiou J, Geusz RJ, Okino ML, Han JY, Miller M, Melton R*, et al.* Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. **Nature 2021**, 594(7863)**:** 398-402.
26. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K*, et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. **Nature 2014**, 506(7488)**:** 376-381.
27. Sauteraud R, Stahl JM, James J, Englebright M, Chen F, Zhan X*, et al.* Inferring genes that escape X-Chromosome inactivation reveals important contribution of variable escape genes to sex-biased diseases. **Genome Res 2021**, 31(9)**:** 1629-1637.
28. Souyris M, Mejia JE, Chaumeil J, Guery JC. Female predisposition to TLR7-driven autoimmunity: gene dosage and the escape from X chromosome inactivation. **Semin Immunopathol 2019**, 41(2)**:** 153-164.
29. Youness A, Miquel CH, Guery JC. Escape from X Chromosome Inactivation and the Female Predominance in Autoimmune Diseases. **Int J Mol Sci 2021**, 22(3).
30. Balaton BP, Cotton AM, Brown CJ. Derivation of consensus inactivation status for X-linked genes from genome-wide studies. **Biol Sex Differ 2015**, 6**:** 35.
31. Ishigaki K. Beyond GWAS: from simple associations to functional insights. **Semin Immunopathol 2022**, 44(1)**:** 3-14.
32. Allen NE, Lacey B, Lawlor DA, Pell JP, Gallacher J, Smeeth L*, et al.* Prospective study design and data analysis in UK Biobank. **Sci Transl Med 2024**, 16(729)**:** eadf4428.
33. Wang MH, Cordell HJ, Van Steen K. Statistical methods for genome-wide association studies. **Semin Cancer Biol 2019**, 55**:** 53-60.
34. Ota M, Fujio K. Multi-omics approach to precision medicine for immune-mediated diseases. **Inflamm Regen 2021**, 41(1)**:** 23.

35. Gu Y, Niu X, Yin L, Wang Y, Yang Y, Yang X*, et al.* Enhancing Fatty Acid Catabolism of Macrophages Within Aberrant Breast Cancer Tumor Microenvironment Can Re-establish Antitumor Function. **Front Cell Dev Biol 2021**, 9**:** 665869.
36. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. **Methods Mol Biol 2018**, 1711**:** 243-259.
37. Xu S, Hu E, Cai Y, Xie Z, Luo X, Zhan L*, et al.* Using clusterProfiler to characterize multiomics data. **Nat Protoc 2024**, 19(11)**:** 3292-3320.
38. Huang N, Xia Y, Zhang D, Wang S, Bao Y, He R*, et al.* Hierarchical assembly of centriole subdistal appendages via centrosome binding proteins CCDC120 and CCDC68. **Nat Commun 2017**, 8**:** 15057.
39. Torii T, Miyamoto Y, Tago K, Sango K, Nakamura K, Sanbe A*, et al.* Arf6 guanine nucleotide exchange factor cytohesin-2 binds to CCDC120 and is transported along neurites to mediate neurite growth. **J Biol Chem 2014**, 289(49)**:** 33887-33903.
40. Tuysuz B, Usluer E, Uludag Alkaya D, Ocak S, Saygili S, Seker A*, et al*. The molecular spectrum of Turkish osteopetrosis and related osteoclast disorders with natural history, including a candidate gene, CCDC120. **Bone 2023**, 177**:** 116897.
41. Hammad H, Lambrecht BN. The basic immunology of asthma. **Cell 2021**, 184(6)**:** 1469-1485.
42. Wischhof L, Scifo E, Ehninger D, Bano D. AIFM1 beyond cell death: An overview of this OXPHOS-inducing factor in mitochondrial diseases. **EBioMedicine 2022**, 83**:** 104231.
43. Rinaldi C, Grunseich C, Sevrioukova IF, Schindler A, Horkayne-Szakaly I, Lamperti C*, et al.* Cowchock syndrome is associated with a mutation in apoptosis-inducing factor. **Am J Hum Genet 2012**, 91(6)**:** 1095-1102.
44. Joza N, Susin SA, Daugas E, Stanford WL, Cho SK, Li CY*, et al.* Essential role of the mitochondrial apoptosis-inducing factor in programmed cell death. **Nature 2001**, 410(6828)**:** 549-554.
45. Zambon AA, Ghezzi D, Baldoli C, Cutillo G, Fontana K, Sofia V*, et al.* Expanding the spectrum of neonatal-onset AIFM1-associated disorders. **Ann Clin Transl Neurol 2023**, 10(10)**:** 1844-1853.
46. Rao S, Mondragon L, Pranjic B, Hanada T, Stoll G, Kocher T*, et al.* AIF-regulated oxidative phosphorylation supports lung cancer development. **Cell Res 2019**, 29(7)**:** 579-591.
47. Wang L, Xu H, Xu D. AIFM1 is a prognostic biomarker for uterine carcinosarcoma and is associated with immune infiltration. **Asian J Surg 2024**, Sep 21:S1015-9584(24)02022-0.
48. Eifan AO, Durham SR. Pathogenesis of rhinitis. **Clin Exp Allergy 2016**, 46(9)**:** 1139-1151.
49. He Y, Chen Y, Xu S, Luo Y, Qin F, Hu W. Pathogenesis and Key Cells in Allergic Rhinitis. **Int Arch Allergy Immunol 2025**, 186(5)**:** 418-429.

9. Contributions

This study was independently carried out by Matthew Tunan Jiang under the supervision of Professor Yifeng Zhou at Jiangsu Provincial Hospital, Nanjing Medical University. The research topic was originally conceived by Matthew Tunan Jiang, and the relevant concepts and study design were developed through extensive discussion between Professor Zhou and Matthew Tunan Jiang. The overall methodological framework was jointly designed by Matthew Tunan Jiang and Professor Zhou. Code implementation, data organization, quality control, and analysis were conducted by Matthew Tunan Jiang with assistance from Professor Zhou's doctoral students. The full manuscript was drafted by Matthew Tunan Jiang, while final review and revision were completed by Professor Zhou.

10. Acknowledgements

Over the past three years, I have gained so many insights into the many challenges, obstacles and opportunities that are present within the field of science. Meanwhile, my heart overflows with immense joy and thanks, for I have found boundless kindness and guidance along the way.

First and foremost, I would like to express my greatest appreciation to my parents. My father, who is a practicing physician, has greatly influenced my outlook in both his career and his interest in biomedical research. He first instilled in me an early interest in the life sciences and medicine, and his near three-decade-long bedside practice has taught me the paramount importance that basic research has for drug development and therapy. This conviction was taught early and has guided my career in science.

I particularly owe a debt to my mentor, Professor Yifeng Zhou. The life sciences demand both rigorous data analysis and hands-on operation experience—skills that I lacked especially when I first started. Knowing about these lacks, Professor Zhou gave me patient and irreplaceable mentorship. From scanning literature and data downloading to learning R for handling large data, each step involved investing hours from ground zero. PhDs in Prof. Zhou's lab also helped to train me on a regular basis. From writing R scripts to implementing statistical models, each lesson was a demonstration of their concern and high expectations. On important occasions in data processing, quality control, and data analysis, Professor Zhou and his team kept encouraging me.

Of course, the research presented in this paper is a success that I cherish. Over the past few years of exposure to research, I've also experienced countless failures and setbacks, both on this project and on many others. The lab room was a world of repeated challenges. I remember fragile gel bands after electrophoresis, repeated mistakes at plasmid transfection, and intimidating pages of unfamiliar protocols. I once managed agar without gloves, spilled buffer while extracting plasmid, and clumsily botched bottles and tubes that I knew little about—too embarrassed sometimes to ask for help.

I remember Professor Zhou once shared an interesting tale from his years ago as a visiting scholar in the U.S.: when people would casually ask, "How's it going?" he would reply quite literally, "I'm going to the lab." Just like me, at the beginning of his science career, he was similarly gauche in the lab. He assured me that mistakes are never terminal—the mortal peril is being afraid to ask. Emboldened by that, little by

little, I overcame my shyness and learned to ask openly for help, even if my program was littered with mistakes or my data were less than perfect.

I owe just as great a debt to the rich resources and supporting environment provided by Professor Zhou's laboratory. There, aside from hands-on technical experience, I learned to think for myself and work out problems on my own. Through the gentle tutelage of so many people in his lab, during my winter and summer research periods, I progressed a long way—being changed and transformed from a passive observer to an active participant, who could carry out independent research and draw relevant results.

As a parting comment, I would like to thank Professor Shing-Tung Yau and the organizers of the S.-T. Yau High School Science Award. I recount that Professor Yau once said in an interview—copying other people's successes is useless; truly scholarly work demands freedom and independence of mind. Yau Award has provided me with a perfect platform on which to prove myself, expand my horizons, and develop my passion and confidence for a research career in science.

Finally, I would like to say, my biggest thanks are to each and every individual who has guided and supported me along the way. And I'll always remember, stay hungry, stay foolish. Keeping a humble mind and a curious heart, I believe firmly that the path forward will be full of prospects and hope.

11. Participant Resume

Name: Matthew Tunan Jiang

School: Nanjing Foreign Language School, Class of 2026

Education

- **Sept 2023 – Present**: Nanjing Foreign Language School (High School)
- **Sept 2020 – Jun 2023**: Nanjing Foreign Language School (Middle School)

Academic & Research Experience

- Initiated biological experimentation in Grade Eight, cultivating a strong interest in biology and medicine. Studied molecular biology and bioinformatics, and actively engaged in research and practical projects.
- **Jul 2024**: Completed the Engineering Summer Academy at Penn (Biotechnology Track), University of Pennsylvania; acquired advanced molecular biology laboratory skills. GPA: 4.0/4.0.
- **Jun – Jul 2025**: Completed the UChicago Summer Session – Biomolecules of Life, University of Chicago; pursued advanced coursework in biology. GPA: 4.0/4.0.

Competitions & Awards

- **United States of America Biology Olympiad**: Gold Medal
- **British Biology Olympiad**: Gold Medal
- **International Genetically Engineered Machine Competition:**
  - High School Division: Global Gold Medal
  - High School Division: Global Top Ten

12. Mentor Biography

Name: Yifeng Zhou, Ph.D.

Gender: Male

Title & Position: Professor

Affiliation: Jiangsu Provincial Hospital, Nanjing Medical University

Honors & Distinctions

- Recipient of the **National Science Fund for Distinguished Young Scholars**
- Recipient of the Jiangsu Provincial Distinguished Young Scholar Award
- Vice Chair, Cancer Etiology Committee, Chinese Anti-Cancer Association

Academic Contributions

- Published 30+ peer-reviewed articles as corresponding author in leading international journals, including *Gastroenterology*, *The EMBO Journal*, *Journal of Experimental Medicine*, *Cancer Research*, among others.
- 2 articles indexed as ESI Highly Cited Papers in the field of *Molecular Biology & Genetics*.
- Total citations: over 7,000; H-index: 45.

Editorial & Peer-Review Roles

- Academic Editor, *Cancer Medicine*, *Gene*
- Ad hoc reviewer for journals including *Nature*, *Gastroenterology*, *Gut*, *Nature Communications*, among others.