Gene-Gene Interactions Contributing To Asthma Susceptibility, An Exploration Using Uk Biobank Dataset

Yuyuan Lin
yuyuan_lin@icloud.com

Follow this and additional works at: https://elischolar.library.yale.edu/ysphtdl

Part of the Public Health Commons

Recommended Citation
https://elischolar.library.yale.edu/ysphtdl/2070

This Open Access Thesis is brought to you for free and open access by the School of Public Health at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Public Health Theses by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.
Gene-gene Interactions Contributing to Asthma Susceptibility

An Exploration using UK Biobank Dataset

by

Yuyuan Lin

A thesis presented for the degree of

Master of Public Health

in

Chronic Disease Epidemiology

Primary Advisor: Andrew Thomas DeWan, PhD, MPH

Secondary Advisor: Yasmmyn Salinas, PhD, MPH

Yale School of Public Health

Spring 2021
Abstract

**Background:** Previous studies have attempted to identify gene-gene interactions for asthma. However, most of these studies suffered from lack of replication or insufficient statistical power. In this study, we aimed to explore the gene-gene interactions affecting asthma susceptibility and tried to replicate the results. Both the discovery and replication datasets were nested in UK Biobank data.

**Methods:** Nested case-control design was used. In the discovery analysis (N= 306,859, cases = 35,483, controls = 271,376), Univariate genome-wide association analysis was performed to prioritize loci for the interaction analysis. Pairwise search for epistasis was conducted among 5,389 SNPs. Replication of the top interactions was then conducted (N = 40,945, cases = 5,623, controls = 35,322).

**Results:** Two interactions met statistical significance (rs1496042 x rs6674451, rs10793149 x rs1939469) after Bonferroni correction. In the replication analysis, one of the interactions detected (rs10793149 x rs1939469) achieved statistical significance. Both interactions showed consistent effect size and direction in the replication dataset.

**Conclusion:** In this study, we identified two interactions associated with asthma susceptibility and successfully replicated one of them. In the interaction that replicated, rs1939469 is located in or near EMSY, which has previously been reported associated with asthma. This result provided suggestive evidence that the interaction detected has possible biological mechanisms behind it.
Acknowledgement

I have received a great deal of support during my thesis project and the writing of this thesis. First and foremost, I would like to express my sincere gratitude to my thesis advisor Professor Andrew DeWan, who provided invaluable assistance during every stage of my thesis project. Without his expertise, I would never have been able to complete this thesis. Second, I would like to thank Professor Yasmmyn Salinas for her insightful comments and suggestions, which sharpened my thinking and made my thesis more comprehensive. I would also like to acknowledge Megan Cahill and Zihan Dong for their hard work on data cleaning and QC. Finally, I would like to thank my parents for their unwavering support during my two years at Yale and Yijun Yang for her love and support during this very special time.
# TABLE OF CONTENTS

1 Background 6

2 Methods 6
   2.1 Study Population 6
   2.2 Quality Control Steps 7
   2.3 Statistical Analysis 9

3 Results 10
   3.1 Univariate GWA Analysis 10
   3.2 Pairwise Epistasis Analysis 11
   3.3 Sensitivity Analysis 12
   3.4 Replication 12

4 Discussion 13

Reference 15
LIST OF TABLES

Table 1. Number of subjects/SNPs removed during different stages of QC procedure 8
Table 2. Significant gene-gene interactions 11
Table 3. Top Two gene-gene interactions detected in the sensitivity analysis 12
Table 4. Replication results 12

LIST OF FIGURES

Figure 1. Patterns for the significant interactions in the primary analysis. 11
Figure 2. Patterns for the significant interactions in the replication analysis 13
1 Background

Asthma is a well-studied chronic disease in genetic epidemiology. Several genetic loci have been identified to influence individual susceptibility to asthma such as ADAM33 (Van Eerdewegh et al., 2002), DPP10 (Allen et al., 2003), NPSR1 (Laitinen et al., 2004), PHF11 (Y. Zhang et al., 2003), and ORMDL3 (Halapi et al., 2010; Lluis et al., 2011; Moffatt et al., 2007). However, one limitation is that single genetic variants cannot explain the complete heritability of asthma, known as “missing heritability” (Eichler et al., 2010; Manolio et al., 2009). To have a better understanding of the genetic etiology of asthma, gene-gene interactions (or epistasis) need to be considered when learning asthma genetics. Although there has been long-standing interest in exploring the effect gene-gene interactions in asthma, studies looking for epistasis are usually faced up with issues such as lack of statistical power and lack of replication (Murk et al., 2015). Despite these limitations, there have been some successes identifying gene-gene interactions for asthma. Most of these studies used candidate gene approaches (Howard et al., 2002; Lee et al., 2004; Wang et al., 2004) and one did a genome-wide search (Murk & DeWan, 2016).

The UK Biobank study, which contains genetic information of more than 500,000 people, offers a unique opportunity to search for gene-gene interactions in asthma susceptibility. The cohort contained 488,282 subjects with asthma information at baseline, the size of which is larger than other typical genome databases and provides sufficient statistical power for interaction analysis. In this study, we explored the epistasis of asthma using a subset of UK Biobank data. To maintain the power of detecting interactions, a two-stage approach was used to limit the search space. In addition, we also tried to replicate the most significant interactions detected in an independent subset of subjects from the UK Biobank.

2 Methods

2.1 Study Population

This study comprised two stages: discovery and replication. Both of the two datasets were nested within the original UK Biobank cohort. UK Biobank is a large, population-based prospective study with more than 500,000
participants aged 40-69 years recruited at 22 centers across the UK between 2006 and 2010. Blood samples for genotyping and blood chemistry assays, self-reported health questionnaires and electronic health record (EHR) data were included. (Sudlow et al., 2015)

For the discovery dataset, we selected 306,859 subjects (35,483 asthma cases and 271,376 controls), who were (1) genetically determined White British European ancestry (2) unrelated pairwise identity by descent (IBD) estimate < 0.1875, which is halfway between the expected IBD for third- and second-degree relatives. (Anderson et al., 2010) (3) with asthma information at baseline. Prevalent asthma status was determined according to the self-reported medical history and EHR data. The detailed definition of cases and controls are listed as below:

Cases:
  - At baseline visit, self-reported a diagnosis of asthma from a doctor (field 6152) or ICD10 code recorded for asthma or severe asthma (J45 or J46 for field 41270).

Controls:
  - Self-reported never to have received a diagnosis of asthma from a doctor (field 6152, any time point)
  - No ICD10 codes recorded for asthma
  - No ICD10 codes recorded for an autoimmune disease
  - Do not self-report autoimmune disease (broadly, in field 20002, or self-report of sarcoidosis diagnosis by doctor in field 22133)

For the replication dataset, a similar criterion was applied but genetically confirmed non-British White European subjects were selected. In total, 40,945 subjects were selected (5,623 asthma cases and 35,322 controls).

2.2 Quality Control Steps

Quality control (QC) steps were performed before the association analysis. The QC procedure can be divided into three steps: initial SNP QC, participant QC and a second stage of SNP QC.

In the initial SNP QC, loci were included if all of the following criteria were met: (1) being autosomal variants; (2) being covered by both arrays; (3) passing Batch-level QC (4) being SNPs only (indels were removed).
Next, we did QC on the participant level. Subjects were excluded if any of the following were true: (1) genetic sex and reported sex did not match; (2) sex chromosomes were not XX or XY; (3) being outliers for heterozygosity/missing rate; (4) individual call rate was less than 99%; (5) being flagged for removal due to relatedness with other subjects; (6) genetic ethnic group being Caucasian for the discovery dataset, and non-Caucasian for the replication dataset.

Finally, the second stage of SNP QC was carried out in the unrelated subjects who were selected in the second step. SNPs were excluded if any of the following were true: (1) call rate was less than 99%; (2) the test for deviation from Hardy-Weinberg Equilibrium (HWE) had p-value less than $5 \times 10^{-8}$; (3) having minor allele frequency (MAF) less than the threshold. We chose 5% as the MAF threshold for the discovery dataset, which is higher than the typical choice (1%) because we wanted to limit the search space for interactions to increase statistical power in the analysis.

| Table 1. Number of subjects/SNPs removed during different stages of QC procedure |
|---------------------------------|----------------|----------------|
|                                  | SNPs removed | Individuals removed |
| **Initial SNP QC**              |              |                  |
| Not autosomal variant           | 21,170       | -               |
| Not covered by both arrays      | 50,934       | -               |
| Batch-level QC                  | 46,318       | -               |
| Not SNPs                        | 12,515       | -               |
| **Participant QC**              |              |                  |
| Genetic and reported sex not match | -           | 372             |
| Sex chromosomes not XX/XY      | -            | 470             |
| Outliers in heterozygosity/missing rate | -        | 963             |
| non Caucasian                   | -            | 78,291          |
| Individual call rate < 99%     | -            | 41,434          |
| Related with other subjects     | -            | 59,493          |
| **Second Stage SNP QC**         |              |                  |
| Call rate < 99%                 | 34,627       | -               |
| HWE (p value < 5e-8)            | 29,843       | -               |
| MAF < 5%                        | 289,825      | -               |
| **Final Number for Analysis**   | 320,194      | 307,259         |

The QC procedures for discovery dataset are shown in Table 1. For the replication dataset, we used quality-controlled subjects, but limited our sample to 44,161 non-British white European subjects. We then removed
1,549 of them due to their relatedness to other subjects. The final number for analysis in the replication dataset was 42,612. For SNPs, we took all the loci that were involved in the significant interactions detected in the discovery dataset.

2.3 Statistical Analysis

Since the power to detect the effect of gene-gene interactions can be limited even when the sample size is large, we used an analytical approach to limit the search space for epistasis in the discovery phase. First, to control for the population stratification, a principal component analysis (PCA) was done using genotype data pruned for linkage disequilibrium (LD) and the first ten principal components (PCs) were extracted. The data pruning was done in PLINK using the independent pairwise command with the following parameters: window size in SNPs was 50; the number of SNPs to shift the window at each step was 5; the R squared threshold used was 0.2. (Purcell et al., 2007) The PCA was performed using EIGENSOFT. (Price et al., 2006) A univariate genome-wide association (GWA) analysis was performed to prioritize SNPs associated with asthma. Logistic regression models were performed using the logistic command in PLINK to estimate the effect of single loci on asthma. Sex, age and the first 10 PCs were included as covariates in the model.

Following the GWA analysis, a pairwise epistasis analysis was conducted to search for epistasis among the loci identified in the first step. SNPs were included in the pairwise search if all of the following criteria were met: (1) univariate p-value less than 0.005; (2) not located in HLA regions (not located between gene HLA-F and KIFC1, i.e., the location of SNP is not between 29691117 and 33377699 based on Genome Reference Consortium Human Build 37 (GRCh37) (Shiina et al., 2004)); (3) not in LD with other SNPs (pruned using the independent pairwise command in PLINK, the parameters were set as 50, 5, and 0.2). The effects of gene-gene interactions were estimated using logistic regressions with an interaction term, which were also fit in PLINK, using the epistasis command. Due to PLINK not being able to handle covariates directly in the epistasis command, we were unable to adjust for covariates in the primary epistasis analysis. The model can be described as: logit\[P(\text{asthma} = 1)] = \beta_1SNP_1 + \beta_2SNP_2 + \beta_3SNP_1 \times SNP_2.\] The effect size was estimated using odds ratios (OR) from the logistic regression model (log \(\beta_3\)).
To account for the potential confounding, a sensitivity analysis was performed for the epistasis analysis which adjusted for covariates (sex, age, body mass index and first ten PCs) using a two-stage regression strategy (i.e. residual method (Kipnis et al., 1997)). The first-stage regression was a logistic regression using the phenotype (asthma) as the outcome and all covariates as predictors. Then the second-stage regression was used to evaluate interactions effect using the Pearson’s residuals from the first-stage regression as the outcome. The assumption behind this approach is that the residual from the first stage regression is an estimate of the susceptibility of asthma uncorrelated with the covariates. The sensitivity analysis was also performed among the SNPs prioritized in the univariate GWA analysis.

To replicate the results, the significant interactions detected in the discovery dataset were tested on the replication dataset using the logistic regression model, which was the same as that in primary epistasis analysis.

To account for the multiple interactions tested in the discovery analyses, the Bonferroni correction was used ($p < 0.05 / [\# \text{interactions}]$). The alpha level was set as 0.05 for replication analyses. All the analyses were conducted on the high-performance cluster (Farnam). R (3.6.2) and the tidyverse package was also used to manage the data as well as extract and visualize results (R Development Core Team, 2019; Wickham et al., 2019).

### 3 Results

#### 3.1 Univariate GWA Analysis

In the univariate GWA analysis, 320,194 SNPs were tested. Of these, 958 SNPs passed the genome-wide significance threshold ($1.56 \times 10^{-7}$). Most of the marginally significant SNPs were located on the Chromosome 6 ($n = 610$), primarily in the HLA region. The majority of remaining significant SNPs were located on chromosomes 2, 5 and 17 ($n = 39, 62, \text{ and } 52$ respectively). This result matched previous GWA analysis results for asthma using the UK biobank data, and no additional significant SNPs were detected (Salinas et al., 2021).

Among all the SNPs tested, 6,983 had p-values less than 0.005. After pruning for LD and removing SNPs from the HLA region, 5,389 SNPs were included in the pairwise search for epistasis.
3.2 Pairwise Epistasis Analysis

Pairwise epistasis analysis was performed among the SNPs prioritized in the first step. Two interactions passed the threshold after Bonferroni correction ($p < 0.05 / (5,292,631) = 9.44e-09$; Table 2).

**Table 2.** Significant gene-gene interactions

<table>
<thead>
<tr>
<th>SNP 1</th>
<th>SNP 2</th>
<th>Interaction P value*</th>
<th>Interaction OR**</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSID</td>
<td>CHR</td>
<td>Gene</td>
<td>P value*</td>
</tr>
<tr>
<td>rs1496042</td>
<td>1</td>
<td>LOC112268240</td>
<td>1.007e-06</td>
</tr>
<tr>
<td>rs10793149</td>
<td>11</td>
<td>LOC105369395</td>
<td>2.432e-03</td>
</tr>
</tbody>
</table>

* Univariate $p$-values were estimated using logistic models adjusting for sex, age and first 10 PCs in the first step univariate GWA analysis.

** Interaction ORs and $p$-values were estimated using logistic models containing the main effect of the loci and their interaction term.

Both interactions had a negative beta indicating an antagonistic effect on the susceptibility of asthma. The patterns of the interactions are illustrated in Figure 1. It is clear that in both interactions, the effect of one genetic locus was affected by the genotype of the other locus. For the second interaction, the effect of the rs1939469 appears to be non-linear when rs10793149 has the genotype C/T or C/C. This phenomenon might be a result of the wide standard error of the estimates caused by the limited number of subjects carrying the minor allele.

![Figure 1.a](image1.png)  ![Figure 1.b](image2.png)

Figure 1. Patterns for the significant interactions in the primary analysis. Figure 1.a is for the first interaction, and Figure 2.b is for the second one. All the ORs and corresponding standard errors in plots were calculated using contingency tables and not adjusted for covariates. The reference level was set as the major allele for both variants in the interactions.
3.3 Sensitivity Analysis

After adjusting for covariates (sex, age, BMI, and first 10 PCs), the results of sensitivity analysis are very close to the primary analysis. The top two interactions remain the same as the primary analysis (Table 3). In this analysis, a two-stage regression method was used and the second stage regression is a linear regression model, so the effect size of interaction was estimated using beta values (i.e. estimated linear regression coefficient).

Table 3. Top Two gene-gene interactions detected in the sensitivity analysis

<table>
<thead>
<tr>
<th>SNP 1 RSID</th>
<th>SNP 1 CHR</th>
<th>SNP 1 Gene</th>
<th>SNP 1 P value*</th>
<th>SNP 2 RSID</th>
<th>SNP 2 CHR</th>
<th>SNP 2 Gene</th>
<th>SNP 2 P value</th>
<th>Interaction P value</th>
<th>Interaction OR*</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1496042</td>
<td>1</td>
<td>LOC112268240</td>
<td>1.007e-06</td>
<td>rs6674451</td>
<td>1</td>
<td></td>
<td></td>
<td>0.182</td>
<td>0.960616</td>
</tr>
<tr>
<td>rs10793149</td>
<td>11</td>
<td>LOC105369395</td>
<td>2.432e-03</td>
<td>rs1939469</td>
<td>11</td>
<td>EMSY</td>
<td>9.866e-09</td>
<td>0.005166</td>
<td>0.870408</td>
</tr>
</tbody>
</table>

* Univariate p-values were estimated using logistic models adjusting for sex, age and first 10 PCs in the first step univariate GWA analysis.
** Interaction betas and p-values were estimated using linear models containing the main effect of the loci and their interaction term.

After Bonferroni correction (corrected alpha level: 9.45e-09), the first interaction remained significant. However, the second interaction was marginally insignificant although it is close to the statistical significance threshold based on the corrected alpha level. This is not unexpected and can be explained by the p-value fluctuation after adjusting for covariates.

3.4 Replication

We also tested the same two interactions in the replication dataset (Table 4).

Table 4. Replication results

<table>
<thead>
<tr>
<th>SNP 1 RSID</th>
<th>SNP 1 CHR</th>
<th>SNP 1 Gene</th>
<th>SNP 1 P value*</th>
<th>SNP 2 RSID</th>
<th>SNP 2 CHR</th>
<th>SNP 2 Gene</th>
<th>SNP 2 P value</th>
<th>Interaction P value</th>
<th>Interaction OR*</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1496042</td>
<td>1</td>
<td>LOC112268240</td>
<td></td>
<td>rs6674451</td>
<td>1</td>
<td></td>
<td></td>
<td>0.182</td>
<td>0.960616</td>
</tr>
<tr>
<td>rs10793149</td>
<td>11</td>
<td>LOC105369395</td>
<td></td>
<td>rs1939469</td>
<td>11</td>
<td>EMSY</td>
<td></td>
<td>0.005166</td>
<td>0.870408</td>
</tr>
</tbody>
</table>

* Interaction ORs and p-values were estimated using logistic models containing the main effect of the loci and their interaction term.

The second interaction was significant in the replication dataset. The first interaction, however, did not achieve statistical significance. The direction and effect size of both two interaction were consistent with the results from the primary analysis. The patterns of the interactions in the replication dataset are shown in Figure 2.
Figure 2. Patterns for the significant interactions in the replication analysis. Figure 1.a is for the first interaction, and Figure 2.b is for the second one. All the ORs and corresponding standard errors in plots were calculated using contingency tables and not adjusted for covariates. The reference level was set as the major allele for both variants in the interactions.

Compared to the results given by primary analysis, the interaction patterns of the replication have broader standard error due to the relatively small sample size. While the patterns are generally consistent with those in the primary analysis.

4 Discussion

In this study, we detected two significant gene-gene interactions associated with the susceptibility of asthma and successfully replicated one of them on an independent set of subjects. The other one, which did not achieve statistical significance in the replication test, also showed consistent direction and size of effect on the replication dataset. Previous studies exploring epistasis have been plagued by the issue of lacking replication. In our study, two different datasets from the same original cohort (UK Biobank) were used to decrease heterogeneity which likely contributed to the successful replication.

Compared to previous studies, which tended to search for interactions among candidate genes, one strength of our study is that our study benefited from the large sample size of UK Biobank, which allowed us to do a more thorough search based on the results of GWA analysis, instead of narrowing the genetic space to a few candidate genes in which to search for interactions.
Among the loci in the two significant gene-gene interactions, rs1939469 is located on gene EMSY, which was reported to be associated with asthma susceptibility (M. Zhang et al., 2020). Both loci on chromosome 1 were reported to be associated with the expression of several genes in lung tissue according to GTEX data (Carithers et al., 2015). Unfortunately, the functional data of the detected loci is quite limited. Further effort should be spent on exploring the mechanism of how these loci and their interactions affect the asthma susceptibility.

There are some limitations of this study. For statistical and computational reasons, we used analytical approaches to limit the search space for interactions, instead of doing a genome-wide search. Some signals might be missed by using this approach where the loci did not reach the threshold we chose for the initial screening. Nevertheless, we included as many loci as possible in the pairwise search by choosing a decent screening threshold in the univariate GWA analysis step. Moreover, the epistasis model used in our study was the logistic regression model with an interaction term, and we only searched for interactions involving two loci. We did not take higher-order interactions into consideration due to the difficulty of detecting them. However, higher-order interactions may have an important genetic effect on the etiology of a complex disease like asthma (Taylor & Ehrenreich, 2015). There have been statistical approaches such as Multifactor-Dimensionality Reduction (MDR) having reasonable power to identify interactions among two or more loci (Ritchie et al., 2001), which can be applied in our further analysis.

In conclusion, our work has shown that the interaction between SNPs rs10793149 and rs1939469 is associated with the development of asthma using the UK Biobank dataset. Although there is still a lot of work that needs to be done before this result can have an impact on clinical practice, our results provide insight into how these genetic interactions could contribute to the etiology of asthma.
References


Taylor, M. B., & Ehrenreich, I. M. (2015). Higher-order genetic interactions and their contribution to complex...


