

# Statistical and Computational Methods to Dissect Ancestry-Biased Germline Effects in Lung Cancer

by

Assel Ismoldayeva

Bachelor of Science in Computer Science and Engineering  
Massachusetts Institute of Technology (2021)

Submitted to the Department of Electrical Engineering and Computer  
Science

in partial fulfillment of the requirements for the degree of  
Master of Engineering in Electrical Engineering and Computer Science  
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2023

© Massachusetts Institute of Technology 2023. All rights reserved.

Author .....  
Department of Electrical Engineering and Computer Science  
January 30, 2023

Certified by .....  
Manolis Kellis  
Professor  
Thesis Supervisor

Certified by .....  
Yosuke Tanigawa  
Postdoctoral Associate  
Thesis Supervisor

Accepted by .....  
Katrina LaCurts  
Chair, Master of Engineering Thesis Committee



# Statistical and Computational Methods to Dissect Ancestry-Biased Germline Effects in Lung Cancer

by

Assel Ismoldayeva

Submitted to the Department of Electrical Engineering and Computer Science  
on January 30, 2023, in partial fulfillment of the  
requirements for the degree of  
Master of Engineering in Electrical Engineering and Computer Science

## **Abstract**

Lung cancer is a complex disease influenced by a variety of genetic and environmental factors. The germline mutations associated with the disease vary greatly between the East Asian and the European populations. We explore these differences by analyzing genome-wide association study summary statistics from European and Japanese biobanks. Using stratified linkage disequilibrium regression in conjunction with gene expression-based and epigenetic annotations, we derive cell-types and biological processes associated with lung cancer and smoking in both populations.

Thesis Supervisor: Manolis Kellis  
Title: Professor

Thesis Supervisor: Yosuke Tanigawa  
Title: Postdoctoral Associate



## Acknowledgments

I would like to thank my supervisor Yosuke Tanigawa for his indispensable help with selecting a project topic, thoughtfully guiding me through all of the stages of my first research endeavor in computational biology and being a patient and approachable mentor. I thank my research advisor Manolis Kellis for his generous guidance on the research direction of my project, inspiring me to explore my newfound interest in computational biology and insightful and engaging conversations during our meetings and group lunches. I am also grateful to all the members of the Kellis Lab for their valuable ideas and help with setting up various tools and datasets, especially Alexandra Berg, Benjamin James, Jackie Yang, Lei Hou, Xikun Han and Zunpeng Liu. I will miss our group lunches and group meetings!

I would also like to thank everyone in the research community who was involved in collecting the data and developing the methods that I based my analysis on.

Finally, I want to thank my family, my friends and my living group for their support and boundless love. I would like to especially thank my parents for always listening to my worries and for giving me great advice any time I need it. I love you! My sincerest gratitude also goes to Andrew for his genuine words of encouragement, thrilling conversations about various fields of biology and being the greatest friend I could ever imagine having.

THIS PAGE INTENTIONALLY LEFT BLANK

# Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Introduction</b>  | <b>15</b> |
| 1.1      | Objectives for this thesis . . . . .   | 16        |
| 1.2      | Thesis outline . . . . .   | 17        |
| <b>2</b> | <b>Background</b>  | <b>19</b> |
| 2.1      | Datasets Used . . . . .  | 19        |
| 2.1.1    | GWAS Datasets . . . . .  | 19        |
| 2.1.2    | Annotations Datasets . . . . .   | 20        |
| 2.2      | Differences between European and East Asian GWAS loci . . . . .                                | 20        |
| <b>3</b> | <b>Methods</b>   | <b>23</b> |
| <b>4</b> | <b>Results and Discussion</b>  | <b>25</b> |
| 4.1      | Heritability enrichment with gene expression data and stratified LD score regression . . . . . | 25        |
| 4.1.1    | Lung cancer . . . . .  | 25        |
| 4.1.2    | Smoking . . . . .  | 28        |
| 4.1.3    | Discussion . . . . .   | 30        |
| 4.2      | Heritability enrichment with histone marks and stratified LD score regression . . . . .        | 32        |
| 4.2.1    | Lung cancer . . . . .  | 32        |
| 4.2.2    | Smoking . . . . .  | 34        |
| 4.2.3    | Discussion . . . . .   | 36        |

|          |   |           |
|----------|---|-----------|
| 4.3      | Heritability enrichment with EpiMap DHS data and stratified LD score regression . . . . . | 38        |
| 4.3.1    | Lung cancer . . . . .   | 38        |
| 4.3.2    | Smoking . . . . .   | 45        |
| 4.3.3    | Discussion . . . . .  | 53        |
| <b>5</b> | <b>Conclusion</b>   | <b>55</b> |
| 5.1      | Future work . . . . .   | 55        |
| 5.2      | Summary . . . . .   | 57        |

# List of Figures

|      |  |    |
|------|--|----|
| 2-1  | Manhattan plot of the European lung cancer GWAS from FinnGen <sup>1</sup> .                    | 21 |
| 2-2  | Manhattan plot of the East Asian lung cancer GWAS from BioBank Japan <sup>2</sup> . . . . .    | 21 |
| 2-3  | GWAS meta-analysis heterogeneity test between the European and East Asian population . . . . . | 22 |
| 4-1  | Heritability enrichment in lung cancer for the European population .                           | 26 |
| 4-2  | p-value of heritability enrichment in lung cancer for the European population . . . . .        | 26 |
| 4-3  | Heritability enrichment in lung cancer for the East Asian population .                         | 27 |
| 4-4  | p-value of heritability enrichment in lung cancer for the East Asian population . . . . .      | 27 |
| 4-5  | Heritability enrichment in smoking for the European population . . .                           | 28 |
| 4-6  | p-value of heritability enrichment in smoking for the European population                      | 29 |
| 4-7  | Heritability enrichment in smoking for the East Asian population . .                           | 29 |
| 4-8  | p-value of heritability enrichment in smoking for the East Asian population . . . . .          | 30 |
| 4-9  | Heritability enrichment in lung cancer for the European population .                           | 32 |
| 4-10 | p-value of heritability enrichment in lung cancer for the European population . . . . .        | 33 |
| 4-11 | Heritability enrichment in lung cancer for the East Asian population .                         | 33 |
| 4-12 | p-value of heritability enrichment in lung cancer for the East Asian population . . . . .      | 34 |

|      |   |    |
|------|---|----|
| 4-13 | Heritability enrichment in smoking for the European population . . .                        | 35 |
| 4-14 | p-value of heritability enrichment in smoking for the European population                   | 35 |
| 4-15 | Heritability enrichment in smoking for the East Asian population . . .                      | 36 |
| 4-16 | p-value of heritability enrichment in smoking for the East Asian population . . . . .       | 36 |
| 4-17 | Heritability enrichment in lung cancer for the European population .                        | 39 |
| 4-18 | Heritability enrichment in lung cancer for the East Asian population .                      | 40 |
| 4-19 | p-value of heritability enrichment in lung cancer for the European population . . . . .     | 41 |
| 4-20 | p-value of heritability enrichment in lung cancer for the East Asian population . . . . .   | 41 |
| 4-21 | Regression coefficient in lung cancer for the European population . .                       | 42 |
| 4-22 | Regression coefficient in lung cancer for the East Asian population . .                     | 43 |
| 4-23 | p-value of regression coefficient in lung cancer for the European population . . . . .      | 44 |
| 4-24 | p-value of regression coefficient in lung cancer for the East Asian population . . . . .    | 44 |
| 4-25 | Heritability enrichment in smoking for the European population . . .                        | 45 |
| 4-26 | Heritability enrichment in smoking for the East Asian population . .                        | 46 |
| 4-27 | p-value of heritability enrichment in smoking for the European population                   | 47 |
| 4-28 | p-value of heritability enrichment in smoking for the East Asian population . . . . .       | 47 |
| 4-29 | Regression coefficient in smoking for the European population . . . .                       | 49 |
| 4-30 | Regression coefficient in smoking for the East Asian population . . .                       | 50 |
| 4-31 | p-value of regression coefficient in smoking for the European population                    | 51 |
| 4-32 | p-value of regression coefficient in smoking for the East Asian population                  | 51 |
| 4-33 | Most significantly enriched GO terms in lung cancer for the European population . . . . .   | 52 |
| 4-34 | Most significantly enriched GO terms in lung cancer for the East Asian population . . . . . | 52 |

|      |   |    |
|------|---|----|
| 4-35 | Most significantly enriched GO terms in smoking for the European population . . . . .           | 53 |
| 4-36 | Most significantly enriched GO terms in smoking for the East Asian population . . . . .         | 53 |
| 5-1  | Colocalizing the significant loci between the European smoking and lung cancer GWAS . . . . .   | 56 |
| 5-2  | Colocalizing the significant loci between the East Asian smoking and lung cancer GWAS . . . . . | 56 |

THIS PAGE INTENTIONALLY LEFT BLANK

# List of Tables

|     |  |    |
|-----|--|----|
| 2.1 | Number of samples in the lung cancer GWAS summary statistics . . . | 19 |
| 2.2 | Number of samples in the smoking GWAS summary statistics . . . .   | 20 |

THIS PAGE INTENTIONALLY LEFT BLANK

# Chapter 1

## Introduction

Lung cancer poses a significant public risk. Lung cancer is the deadliest of all cancers—it has caused 1.8 million deaths in 2020, more than any other type of cancer<sup>3</sup>. This is partially due to most lung cancer patients presenting with already advanced stages of cancer<sup>4</sup>. Furthermore, some types of lung cancer, such as non-small-cell lung carcinoma (which accounts for 85% of all lung cancers<sup>5</sup>), do not respond well to conventional chemotherapy<sup>5</sup>.

There has been a lot of work done to understand the genetic underpinnings of lung cancer<sup>6,7,8</sup>. So far, most of these studies have been focused on populations of single ancestry (especially European ancestry). Some work has also been done in examining the common ways lung cancer arises in different ancestries and it was able to find that mutations that promote endogenous DNA damage can frequently be responsible for lung cancer in these populations<sup>9</sup>. However, there are some differences between how lung cancer develops in people of different ancestry.

In terms of lung cancer risk, in the US, the odds ratio for lung cancer in smokers relative to non-smokers is about 10 times larger than the odds ratio in Japan<sup>10</sup>. Furthermore, it has been shown that individuals of Asian ancestry have different outcomes and toxicity in lung cancer compared to individuals of European ancestry<sup>11</sup>.

When comparing across ancestries, it has been shown that genetic alterations associated with lung cancer differ based of ancestry. For example, East Asian patients have a much higher prevalence of epidermal growth factor receptor (EGFR) mutation

(approximately 30% vs. 7%, predominantly among patients with adenocarcinoma and never-smokers) as well a lower prevalence of K-Ras mutation (less than 10% vs. 18%, predominantly among patients with adenocarcinoma and smokers)<sup>12</sup>. In a study from Brazilian lung cancer patients<sup>13</sup>, it was found that EGFR mutations were associated with high-Asian ancestry, whereas KRAS mutations were associated with non-Asian ancestry. Finally, correlation between ancestry and specific somatic alterations, including driver mutations in EGFR and KRAS, has been reported in a study that considered admixed Latin American populations<sup>14</sup>. Additionally, their analysis suggests that germline mutations in the Native American population are correlated with the somatic mutations. All of this points to differences in mechanisms that govern how lung cancer develops in people of different backgrounds.

Understanding these differences could improve the choice of treatments. Coupled with not fully knowing how genetic variants affect smoking, which in turn affects the risk of development of lung cancer, it highlights the need to do research of populations of different ancestries.

## 1.1 Objectives for this thesis

In this work, I use genome-wide association study (GWAS) summary statistics from two populations, European and East Asian, to find how the two populations differ in their germline genetic variations associated with lung cancer and smoking (due to its strong connection to lung cancer).

The analysis contains three main directions:

- identifying relevant cell-types and pathways associated with lung cancer for both populations
- identifying relevant cell-types and pathways associated with smoking for both populations
- comparing the smoking and lung cancer GWAS, as well as associated cell-types and pathways, between the European and East Asian population

## 1.2 Thesis outline

The rest of this thesis is outlined as follows. Chapter 2 introduces the datasets used and presents the differences between the Manhattan plots for the lung cancer and smoking GWASes in the two populations. Chapter 3 lays out the methods that were used to analyze the data. Chapter 4 presents the results of this thesis and provides a discussion of the interpretation of these results. Chapter 5 summarizes my work and proposes future directions for this research.

THIS PAGE INTENTIONALLY LEFT BLANK

# Chapter 2

## Background

### 2.1 Datasets Used

#### 2.1.1 GWAS Datasets

I have used publicly available GWAS summary statistics<sup>15,16,17</sup> that use data from UK Biobank<sup>18</sup>, FinnGen<sup>1</sup> and BioBank Japan<sup>2</sup>. These biobanks contain the genotypes and phenotypes from hundreds of thousands of people. In particular, Table 2.1 and Table 2.2 contain the number of samples in each of the two populations for the two phenotypes used (lung cancer and smoking/cigarettes per day). Note that the cigarettes per day GWAS are continuous, therefore there is no case/control separation for the samples. Also note that data from FinnGen and UK Biobank is merged into one single European GWAS summary statistic.

| Population | Total samples | Cases | Controls |
|------------|---------------|-------|----------|
| European   | 492,803       | 3,791 | 489,012  |
| East Asian | 178,726       | 4,444 | 174,282  |

Table 2.1: Number of samples in the lung cancer GWAS summary statistics

| Population | Total samples |
|------------|---------------|
| European   | 128,434       |
| East Asian | 74,893        |

Table 2.2: Number of samples in the smoking GWAS summary statistics

### 2.1.2 Annotations Datasets

For the annotations used to calculate the heritability enrichment in LD score regression, I used multiple datasets. Various types of annotations (histone modifications, DNase hypersensitivity peaks, gene expression) are used to derive insights on the biological meaning behind the GWAS associations.

My initial analysis uses gene expression annotations and epigenomic annotations in the form of LD scores from the cell-type specific analysis performed by Funicane et al<sup>19</sup>. The gene expression data was from GTEx<sup>20,21</sup> and another dataset from Lude Franke’s lab<sup>22</sup> that contains gene expression data from human, mouse and rat samples, for a total of 205 cell-type- or tissue-specific gene expression-based annotations. The epigenomic annotations contain narrow peaks from the Roadmap Epigenomics consortium for DNase I hypersensitivity (DHS) and five activating histone marks (H3K27ac, H3K4me3, H3K4me1, H3K9ac and H3K36me3). Each of these six features was present in a subset of the 88 primary cell types or tissues, for a total of 397 cell-type- or tissue-specific epigenomic annotations.

In addition to the above datasets, I have also used annotations from EpiMap<sup>23</sup>. It was built using multiple histone mark annotations and chromatin accessibility regions. In particular, I used the EpiMap modules which were derived by clustering the enhancers into 300 distinct modules.

## 2.2 Differences between European and East Asian GWAS loci

Figure 2-1 and Figure 2-2 show the Manhattan plots for lung cancer GWAS in the Finnish and East Asian population. We can see that the distributions of associations

is greatly different and the most significant SNPs are differing between the two populations. This could suggest differences in causal variants, which would motivate a need for different treatment and testing strategies.

Figure 2-1: Manhattan plot of the European lung cancer GWAS from FinnGen<sup>1</sup>

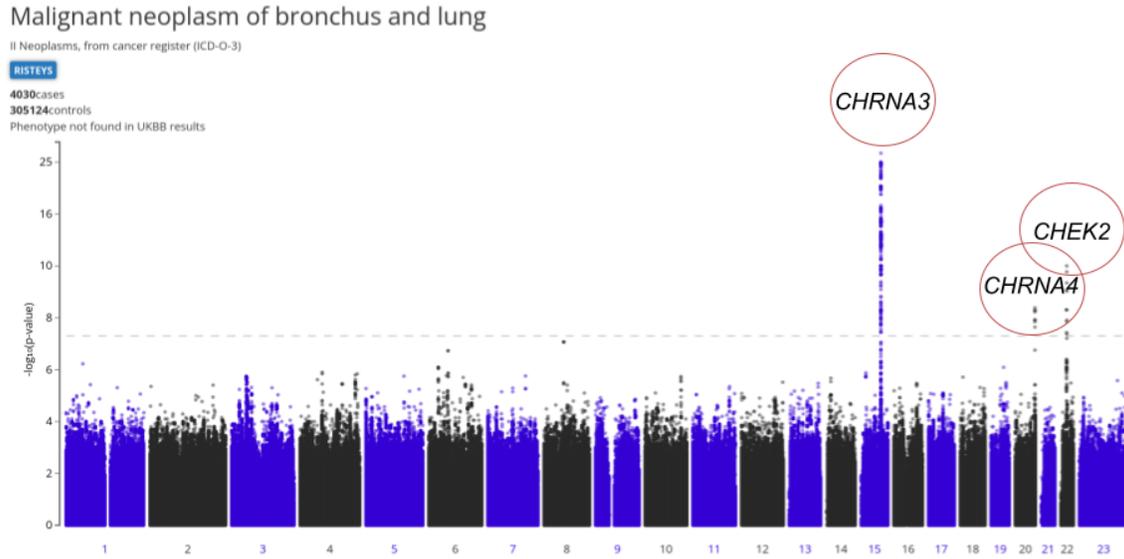
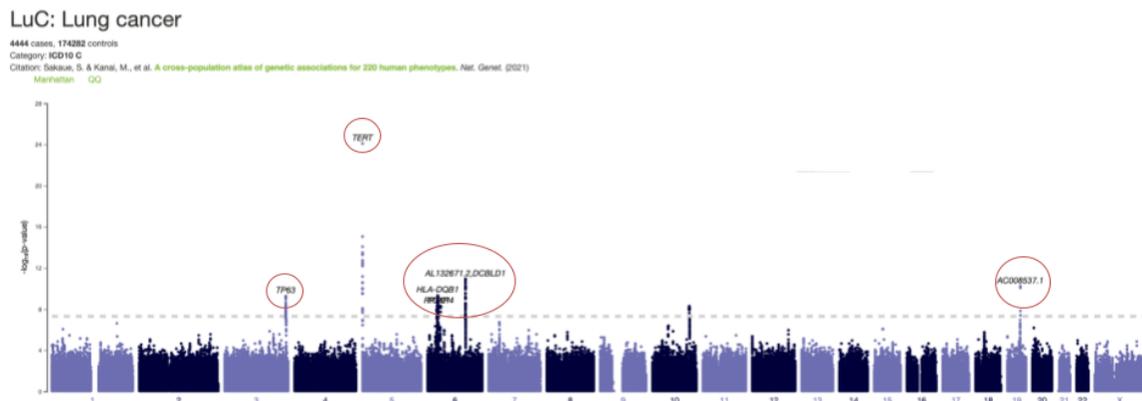
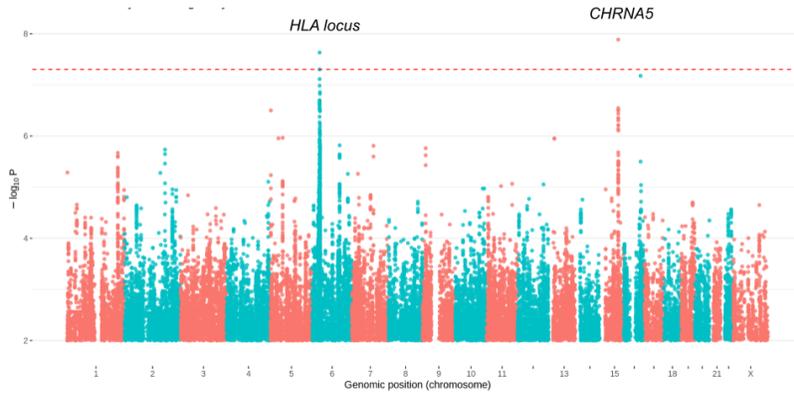


Figure 2-2: Manhattan plot of the East Asian lung cancer GWAS from BioBank Japan<sup>2</sup>



We have also performed a heterogeneity test to verify our observation. It shows that there are significant differences between the East Asian and European (UK Biobank and FinnGen combined) GWASs.

Figure 2-3: GWAS meta-analysis heterogeneity test between the European and East Asian population



# Chapter 3

## Methods

The genome has a natural structure where loci near each other are highly correlated with each other in linkage disequilibrium blocks. Due to this, GWAS plots show multiple correlated loci all with a high association with the phenotype, which makes it hard to deduce the causal variant.

Linkage disequilibrium (LD) score regression is a method first introduced by Bulik-Sullivan et al.<sup>24</sup> to help understand whether the inflation of test statistics is due to polygenicity of the studied phenotypes or confounding biases in the population.

We can use this method to help us identify the key cell types and pathways from the GWAS data. By partitioning the SNPs into categories (that could correspond to conserved regions of the genome or gene regions expressed in a given tissue, for example), we can ask whether SNPs that belong to that category have a higher heritability enrichment than expected. In particular, stratified LD score regression<sup>25</sup> uses LD scores to compute how much of the heritability of a given SNP can be contributed to each of the categories. These heritability enrichment estimates can be used to find relevant tissue and cell-types by selecting annotations that cover sets of loci associated with various tissue and cell-types<sup>19</sup>.

I used stratified LD score regression in conjunction with various sources of annotation data to analyze the GWAS summary statistics. I used gene expression data<sup>20,21</sup>, as well as epigenetic data<sup>23,26</sup> to find differences and similarities between the relevant cell-types and pathways that could explain the differences in rates of lung cancer and

smoking across the two populations.

I also used the EpiMap enhancer modules data<sup>23</sup> to relate the most relevant categories to gene ontology terms to get insight on biological processes associated with the modules with high heritability enrichment. The gene ontology enrichments were conducted using GREAT v3.0.0 for the biological process, cellular component and molecular function ontologies.

# Chapter 4

## Results and Discussion

In the next three sections, all p-value plots (such as Figure 4-2 and Figure 4-4) display two vertical lines: one at the significance level of  $p = 0.05$  and one at the Bonferroni corrected significance level. On all plots in the next three sections, the annotations on the y-axis are always ordered in ascending order of heritability enrichment p values.

### 4.1 Heritability enrichment with gene expression data and stratified LD score regression

For this analysis, the annotation data used was multi-tissue gene expression LD scores data from the cell-type specific LD score analysis performed by Funicane et al<sup>19</sup>. I performed stratified LD score regression as described in Chapter 4 on lung cancer and smoking GWAS summary statistics data from the European and East Asian population.

#### 4.1.1 Lung cancer

In both populations, none of the heritability enrichment p-values were significant after multiple hypothesis testing correction.

We can still see a very different ordering between the top gene expression annotations between the East Asian and European populations. We also see a large

heritability enrichment for various cell-type annotations in both of the populations in Figure 4-1 and Figure 4-3.

Figure 4-1: Heritability enrichment in lung cancer for the European population

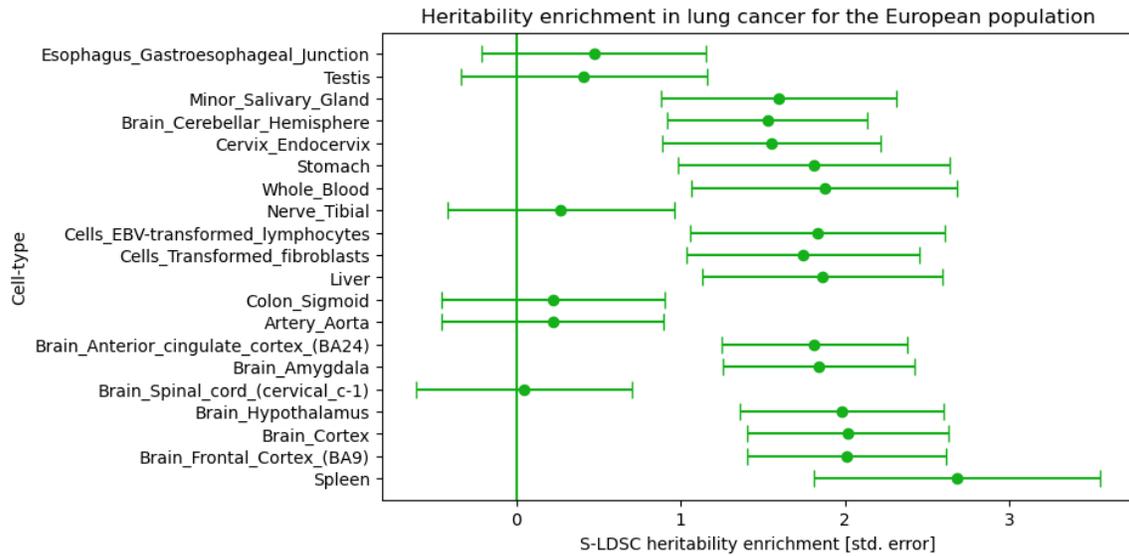


Figure 4-2: p-value of heritability enrichment in lung cancer for the European population

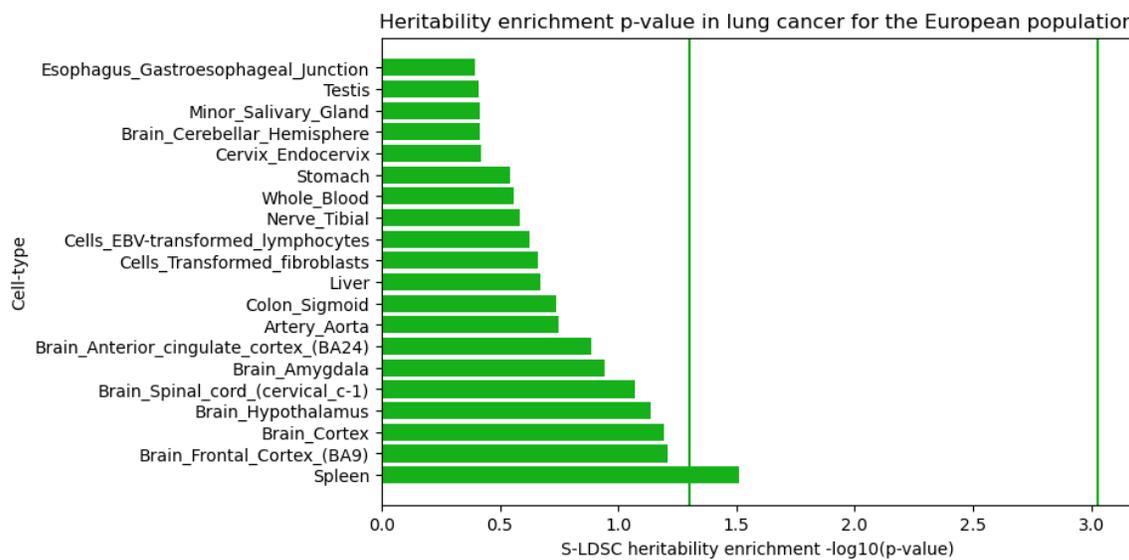


Figure 4-3: Heritability enrichment in lung cancer for the East Asian population

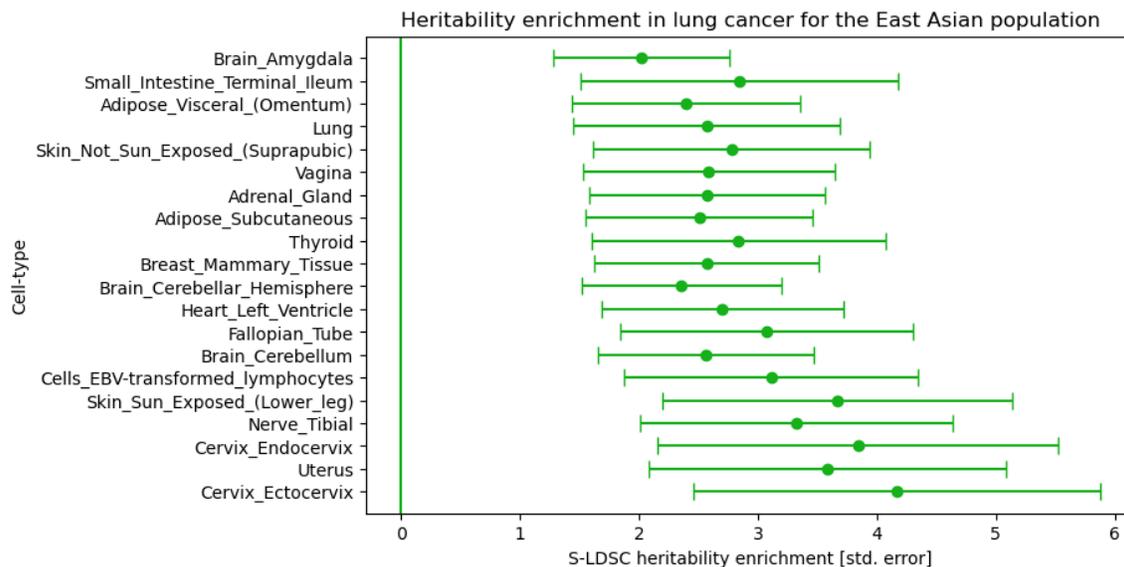
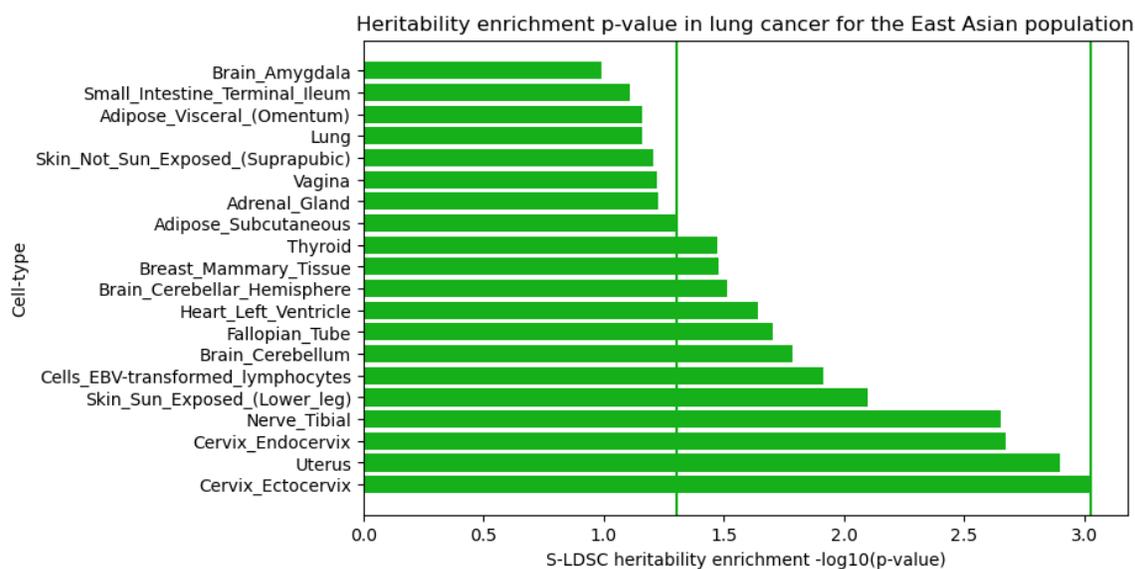


Figure 4-4: p-value of heritability enrichment in lung cancer for the East Asian population



## 4.1.2 Smoking

In the European population, we see a few cell-type annotations pass the significance threshold after multiple hypothesis testing correction but in the East Asian population none of the annotations show significance after the Bonferroni correction. Again, we can see a very different ordering between the top gene expression annotations between the East Asian and European populations. We also see a large heritability enrichment for various cell-type annotations in both of the populations in Figure 4-5 and Figure 4-7. The average heritability enrichment is higher for lung cancer compared to smoking across both populations.

Figure 4-5: Heritability enrichment in smoking for the European population

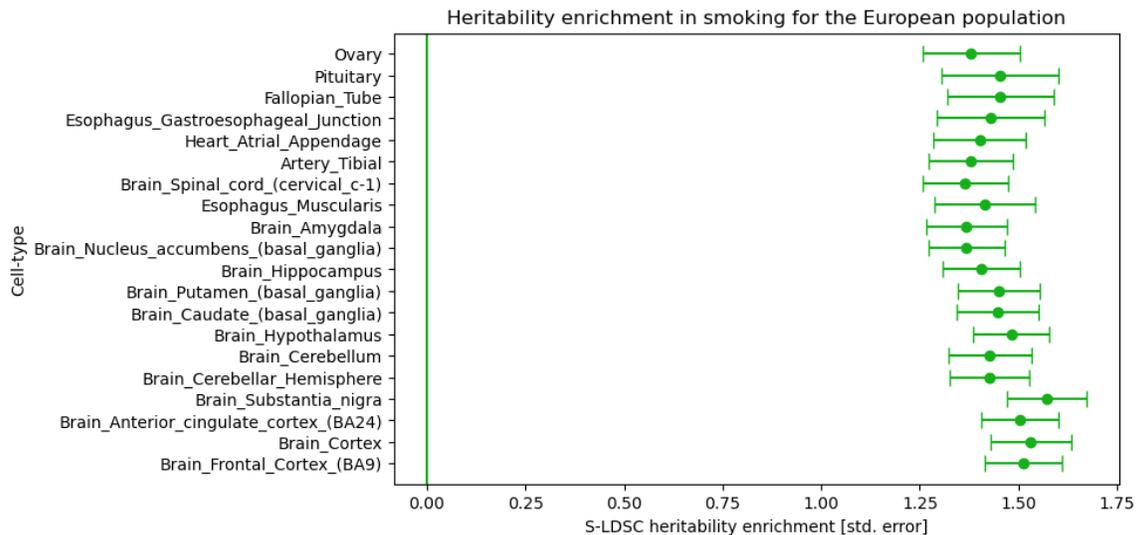


Figure 4-6: p-value of heritability enrichment in smoking for the European population

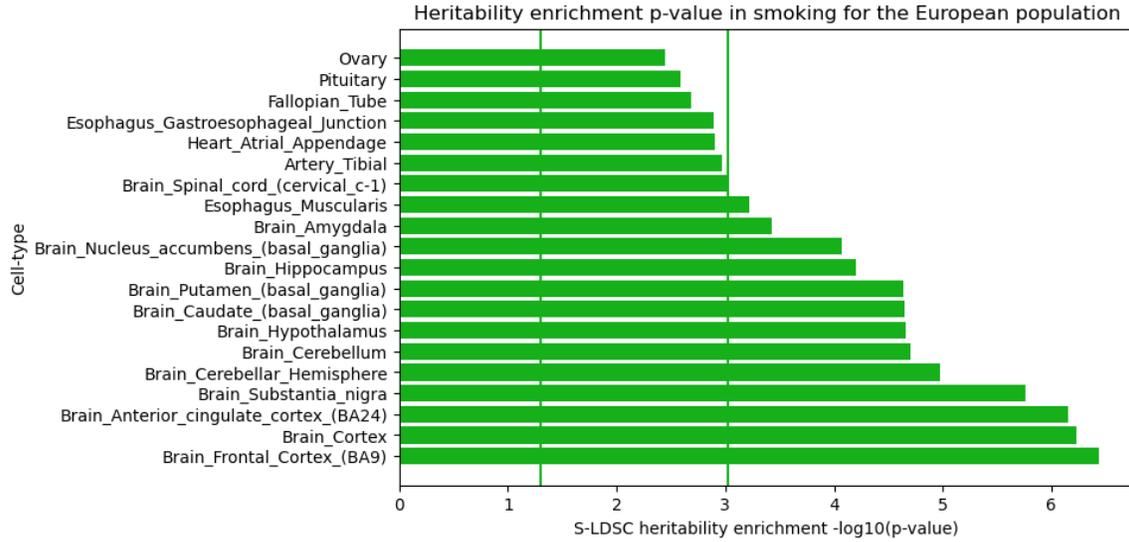


Figure 4-7: Heritability enrichment in smoking for the East Asian population

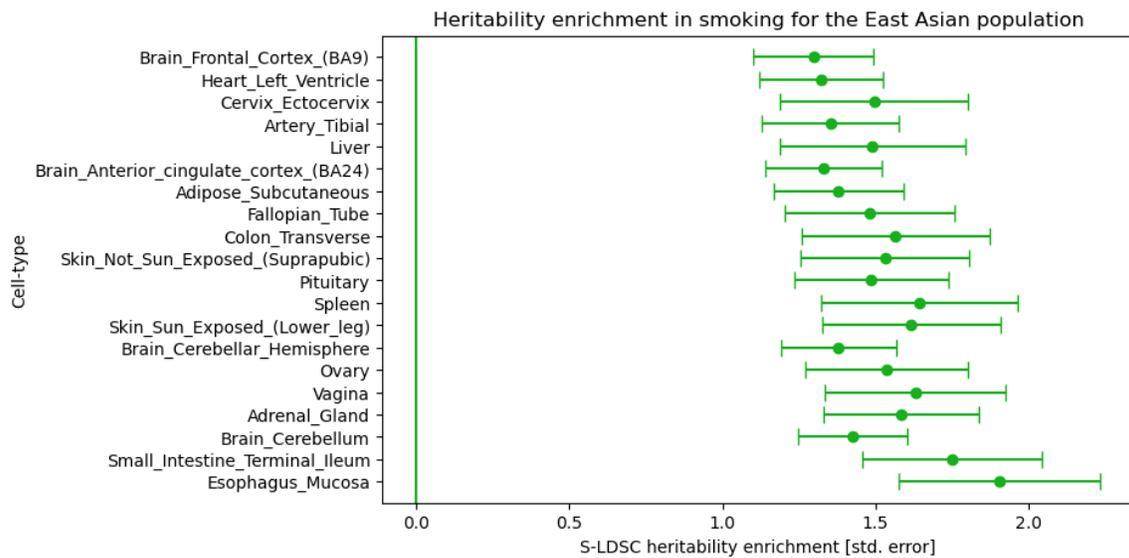
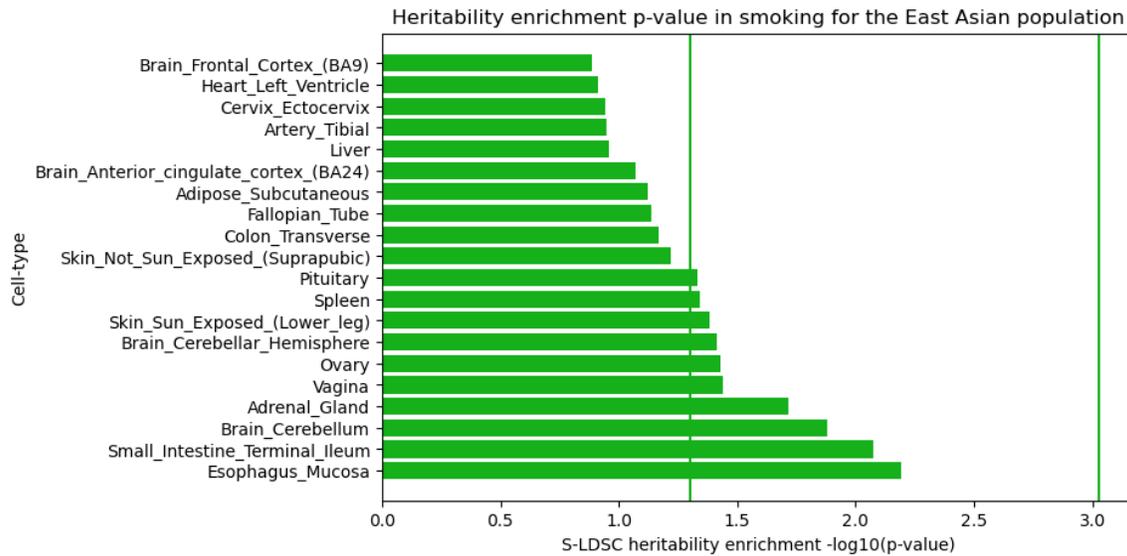


Figure 4-8: p-value of heritability enrichment in smoking for the East Asian population



### 4.1.3 Discussion

There are a few interesting observations that we can make from this data.

First, from the analysis on lung cancer in the European population, we see the highest significance in heritability enrichment in brain-related cell-types and immune-related cell-types, followed by some epithelium-related cell-types. On the other hand, in the East Asian population, we see the highest significance in heritability enrichment in epithelium-related cell-types, followed by a few of brain-related cell-types and immune-related cell-types. This suggests that either:

- individuals of European ancestry are more frequently affected by germline mutations in brain-related cell-types and immune-related cell-types compared to individuals of East Asian ancestry or
- given the same frequency of germline mutations in brain-related cell-types and immune-related cell-types, individuals of European ancestry are more likely to develop lung cancer

Second, from the analysis on smoking in the European population, we see the dominant highest significance in heritability enrichment in brain-related cell-types, followed by some epithelium-related cell-types. On the other hand, in the East Asian population, we see the highest significance in heritability enrichment in epithelium-related cell-types, followed by a few brain-related cell-types. This suggests that either:

- individuals of European ancestry are more frequently affected by germline mutations in brain-related cell-types compared to individuals of East Asian ancestry or
- given the same frequency of germline mutations in brain-related cell-types, individuals of European ancestry are more likely to develop smoking habits. This hypothesis could imply differences in smoking addiction mechanisms between the two populations.

Finally, from comparing Figure 4-1 and Figure 4-3 to Figure 4-5 and Figure 4-7, we see that among the top associated cell-types, the average heritability enrichment for lung cancer is about twice as large that for smoking across the two populations. This could be interpreted as smoking being more dependent on environmental effects as opposed to germline effects when compared with lung cancer, which would make sense since smoking is a behavior enforced by the surrounding community.

## 4.2 Heritability enrichment with histone marks and stratified LD score regression

For this analysis, the chromatin annotations used were from multi-tissue DNase I hypersensitivity (DHS) sites and five activating histone marks (H3K27ac, H3K4me3, H3K4me1, H3K9ac and H3K36me3) from Roadmap Epigenomics<sup>26</sup>. The LD score data from these chromatin annotations is from the cell-type specific LD score analysis performed by Funicane et al<sup>19</sup>. I performed stratified LD score regression as described in Chapter 4 on lung cancer and smoking GWAS summary statistics data from the European and East Asian population.

### 4.2.1 Lung cancer

In both populations, none of the heritability enrichment p-values were significant after multiple hypothesis testing correction.

Like in the previous section, we see a different ordering between the top chromatin annotations between the East Asian and European populations. We also see a large heritability enrichment for various cell-type annotations in both of the populations in Figure 4-9 and Figure 4-11.

Figure 4-9: Heritability enrichment in lung cancer for the European population

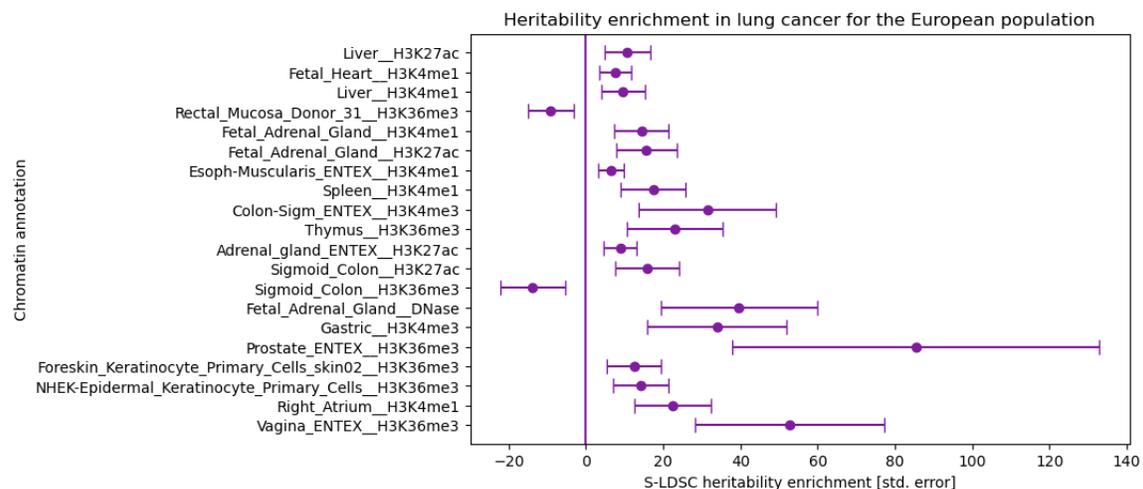


Figure 4-10: p-value of heritability enrichment in lung cancer for the European population

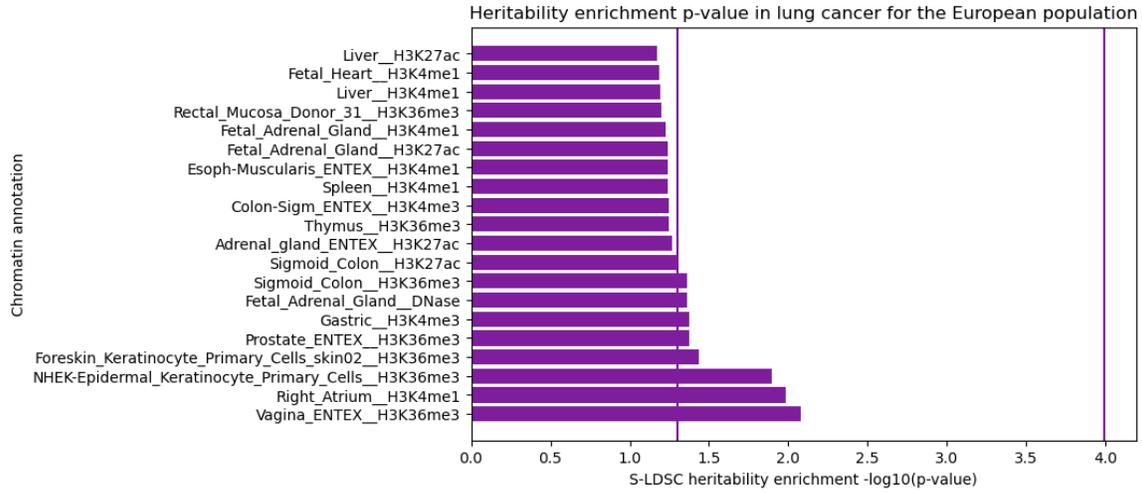


Figure 4-11: Heritability enrichment in lung cancer for the East Asian population

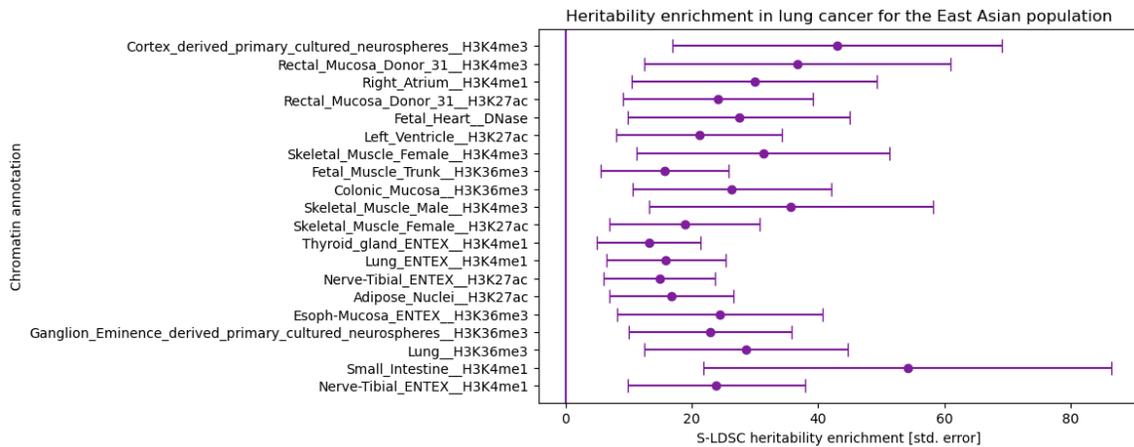
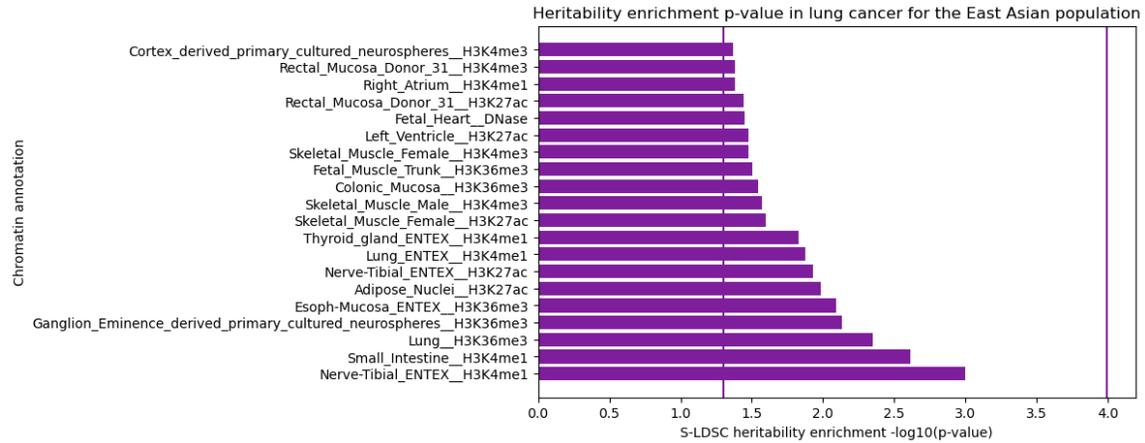


Figure 4-12: p-value of heritability enrichment in lung cancer for the East Asian population



## 4.2.2 Smoking

In the European population, we see a few chromatin annotation pass the significance threshold after multiple hypothesis testing correction but, again, in the East Asian population none of the annotations show significance after the Bonferroni correction. There is also an extremely different ordering between the annotations between the East Asian and European populations and a large heritability enrichment for various chromatin annotations in both of the populations in Figure 4-13 and Figure 4-15. The average heritability enrichment is higher for lung cancer compared to smoking across both populations.

Figure 4-13: Heritability enrichment in smoking for the European population

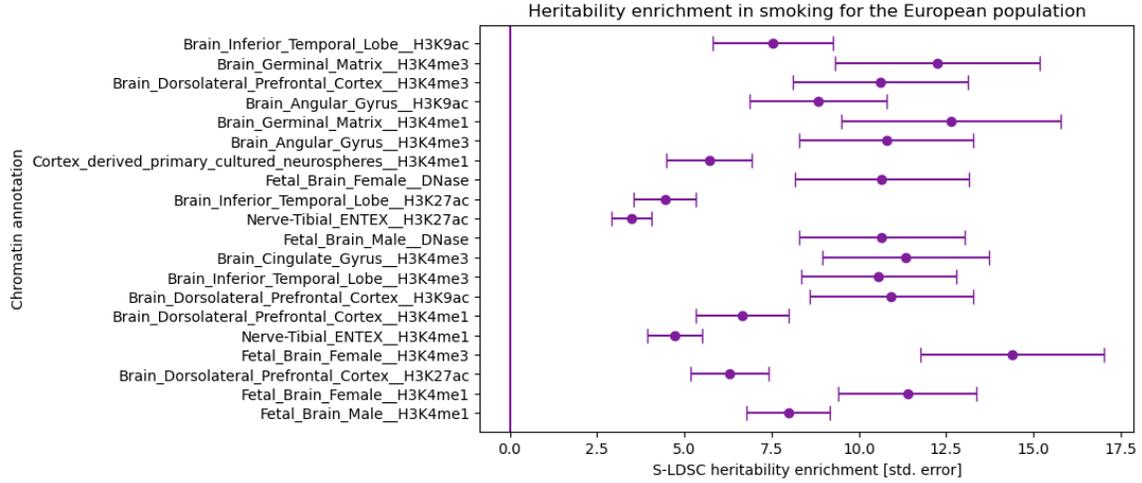


Figure 4-14: p-value of heritability enrichment in smoking for the European population

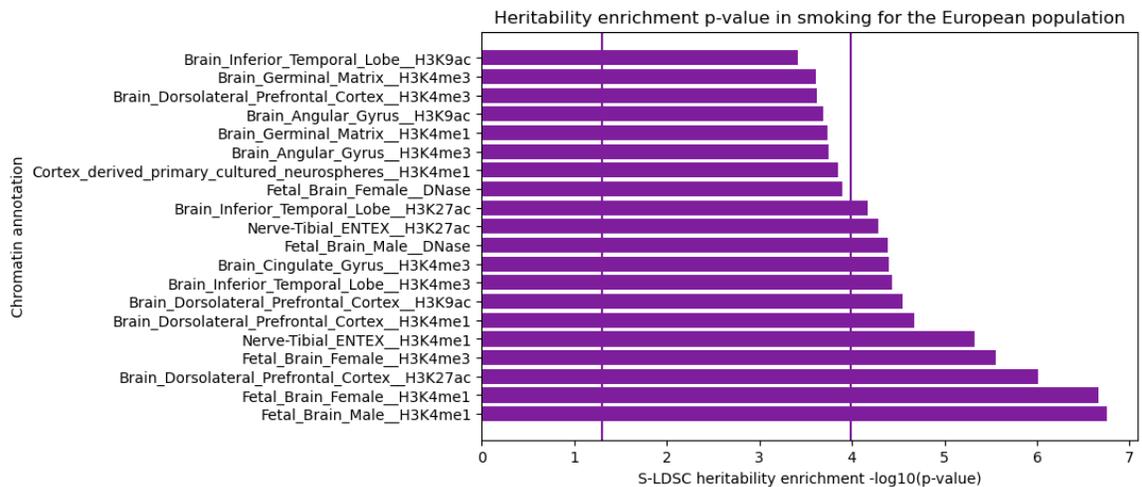


Figure 4-15: Heritability enrichment in smoking for the East Asian population

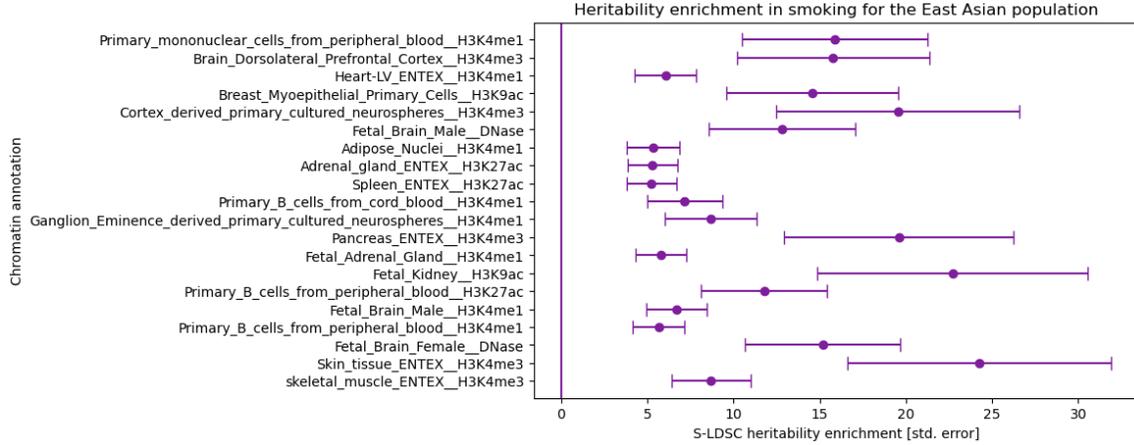
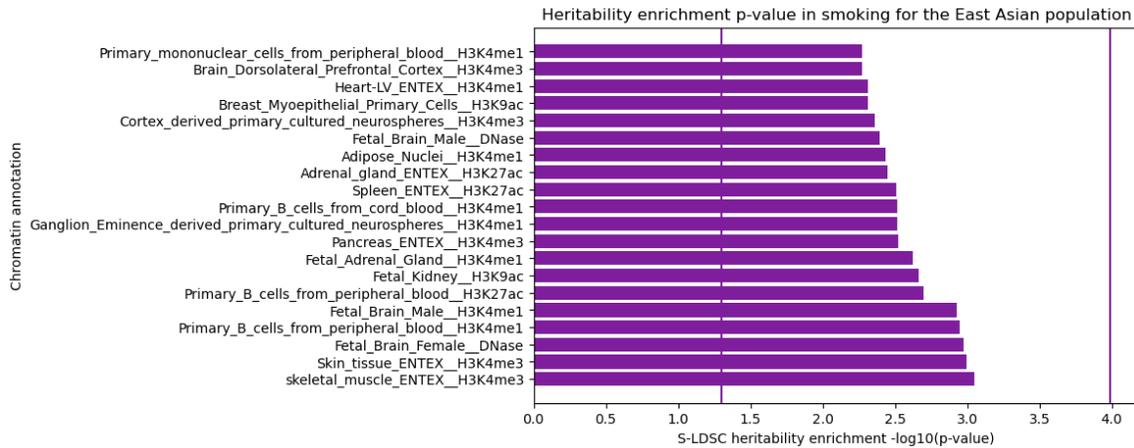


Figure 4-16: p-value of heritability enrichment in smoking for the East Asian population



### 4.2.3 Discussion

There are a few interesting observations that we can make from this data.

First, from the analysis on lung cancer, we see that the top annotation lists are more similar between the two populations than they were when we used gene expression annotations. In both populations, we see the highest significance in heritability enrichment in epithelium-related active regions, followed by some immune-related ac-

tive regions in the European population and nervous-system-related active regions in the East Asian population.

Second, from the analysis on smoking in the European population, we see the dominant highest significance in heritability enrichment in brain-related active regions (just like in the previous section). On the other hand, in the East Asian population, we see the highest significance in heritability enrichment in immune-related and nervous-system-related active regions with a few epithelial-related active regions. This again suggests that either:

- individuals of European ancestry are more frequently affected by germline mutations in brain-related cell-types compared to individuals of East Asian ancestry or
- given the same frequency of germline mutations in brain-related cell-types, individuals of European ancestry are more likely to develop smoking habits. This hypothesis could imply differences in smoking addiction mechanisms between the two populations.

It is also worth noticing that different recruitment strategies between the two biobanks could be influencing the differences we see between the two populations. In general, participants in the UK BioBank are on average healthier than the average population<sup>27</sup>.

Finally, from comparing Figure 4-9 and Figure 4-11 to Figure 4-13 and Figure 4-15, we see that among the top associated annotations, the average heritability enrichment for lung cancer is about twice as large that for smoking across the two populations. Again, like in the previous section, this could be interpreted as smoking being more dependent on environmental effects as opposed to germline effects when compared with lung cancer.

## 4.3 Heritability enrichment with EpiMap DHS data and stratified LD score regression

For this analysis, the epigenetic annotations used were enhancer modules from EpiMap<sup>23</sup>. There is a total number of 300 EpiMap modules that were defined by k-centroids clustering of active enhancers. The LD score data from the EpiMap annotations was computed using the software package LDSC<sup>28</sup> by Brendan Bulik-Sullivan and Hilary Funicane. I performed stratified LD score regression as described in Chapter 4 on lung cancer and smoking GWAS summary statistics data from the European and East Asian population.

### 4.3.1 Lung cancer

In both populations, while none of the heritability enrichment p-values were significant after multiple hypothesis testing correction, we see some modules showing signs of a significant heritability enrichment.

Like in the previous section, we see a very different set of modules showing the highest heritability enrichment between the East Asian and European population. We also see a large heritability enrichment for various cell-type annotations in the European populations in Figure 4-17. However, for the East Asian population, we see very small (centred around 0) enrichment values in Figure 4-18.

Figure 4-17: Heritability enrichment in lung cancer for the European population

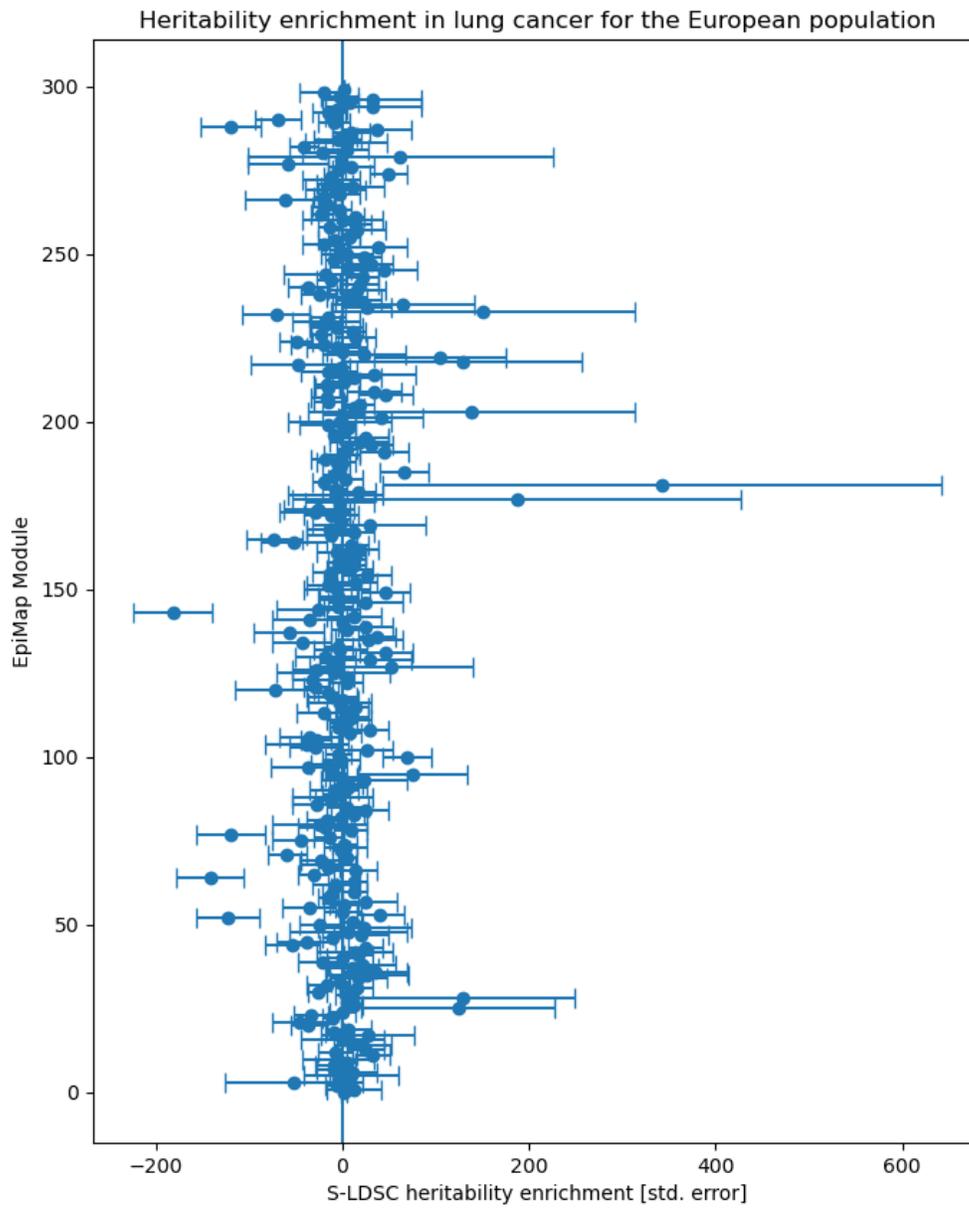


Figure 4-18: Heritability enrichment in lung cancer for the East Asian population

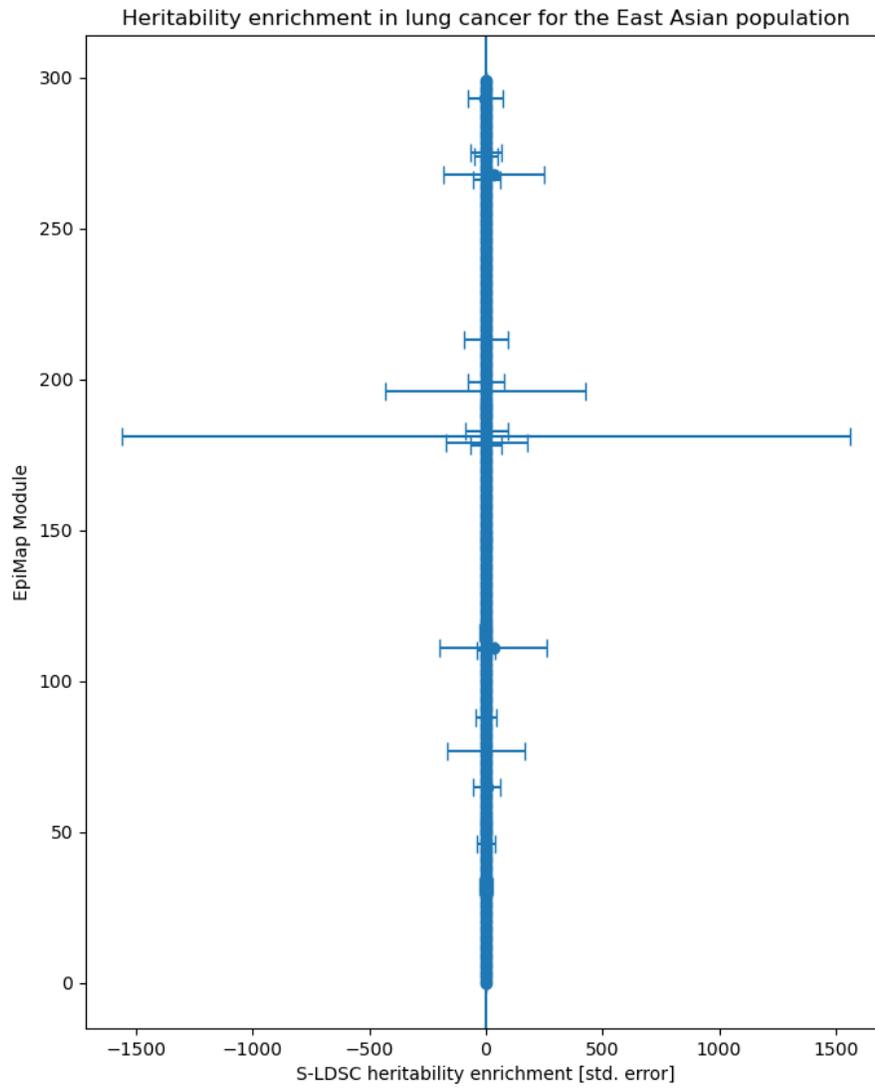


Figure 4-19: p-value of heritability enrichment in lung cancer for the European population

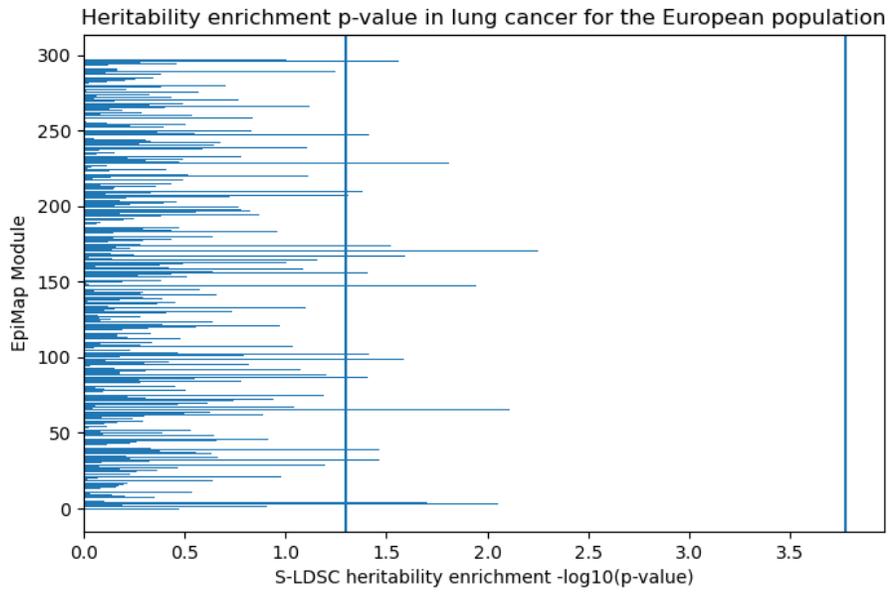
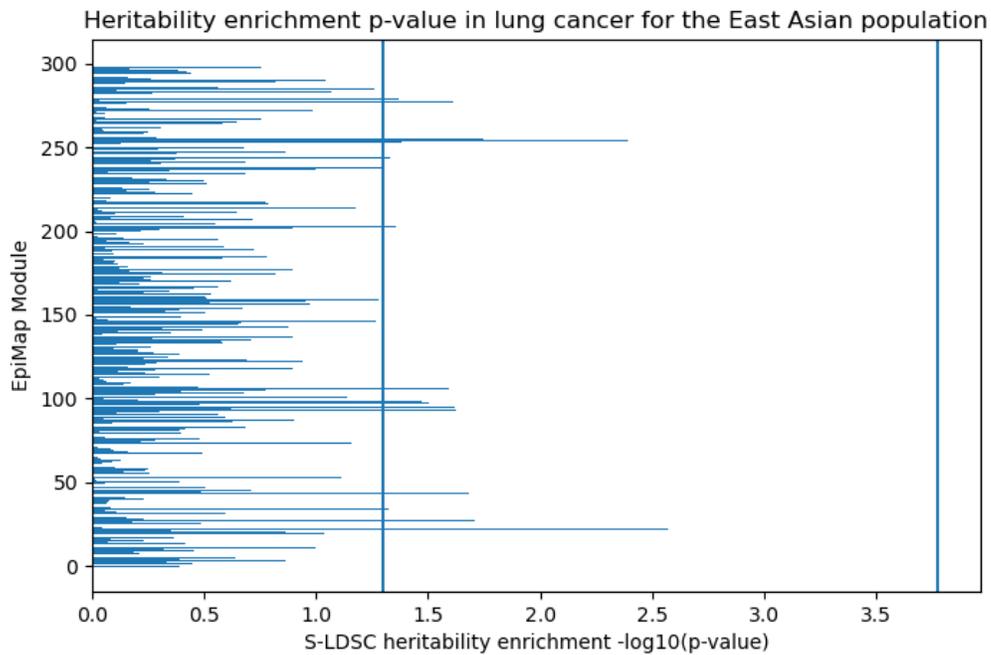


Figure 4-20: p-value of heritability enrichment in lung cancer for the East Asian population



For this reason, I have decided to also plot the error bars for the regression coefficients (they correspond to how much heritability is enhanced if a SNP is part of that annotation module relative to the baseline model<sup>25</sup>). Figure 4-22 shows a distribution of regression coefficients that isn't entirely centered around 0 compared to Figure 4-18. This lets us indirectly compare the enrichment distribution among the two populations.

Figure 4-21: Regression coefficient in lung cancer for the European population

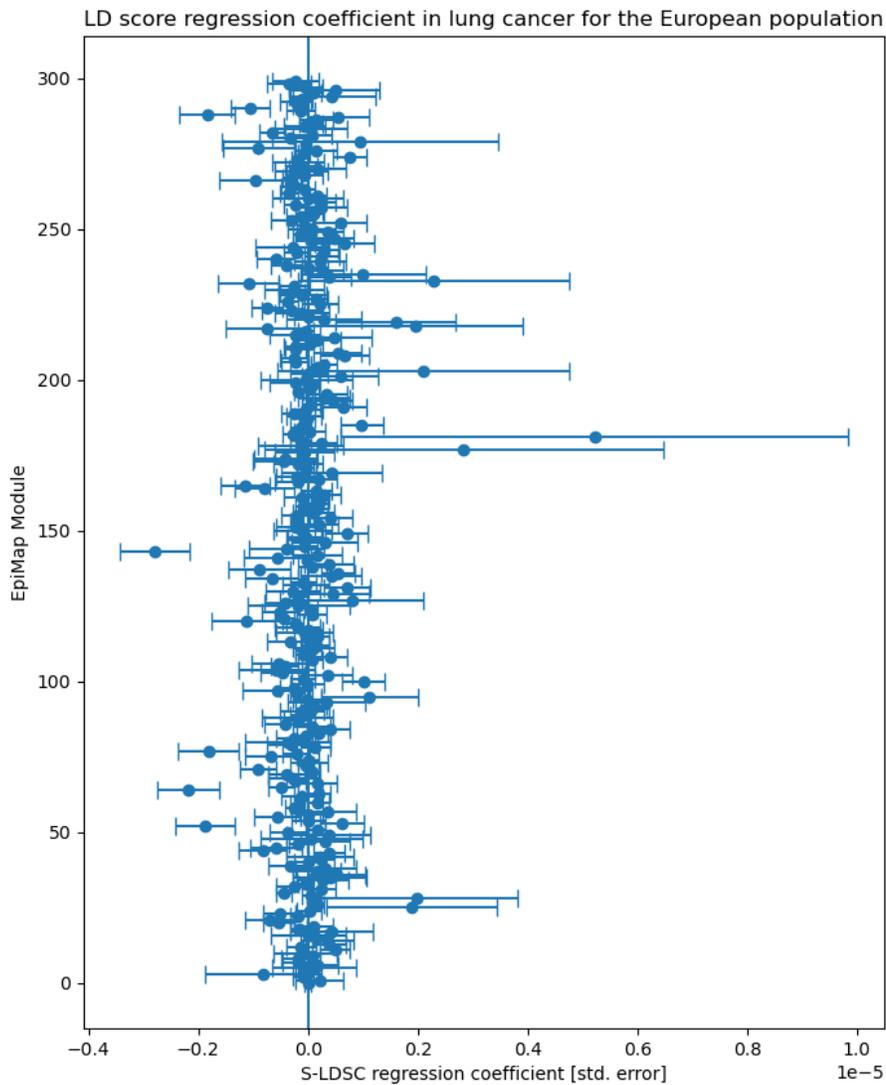


Figure 4-22: Regression coefficient in lung cancer for the East Asian population

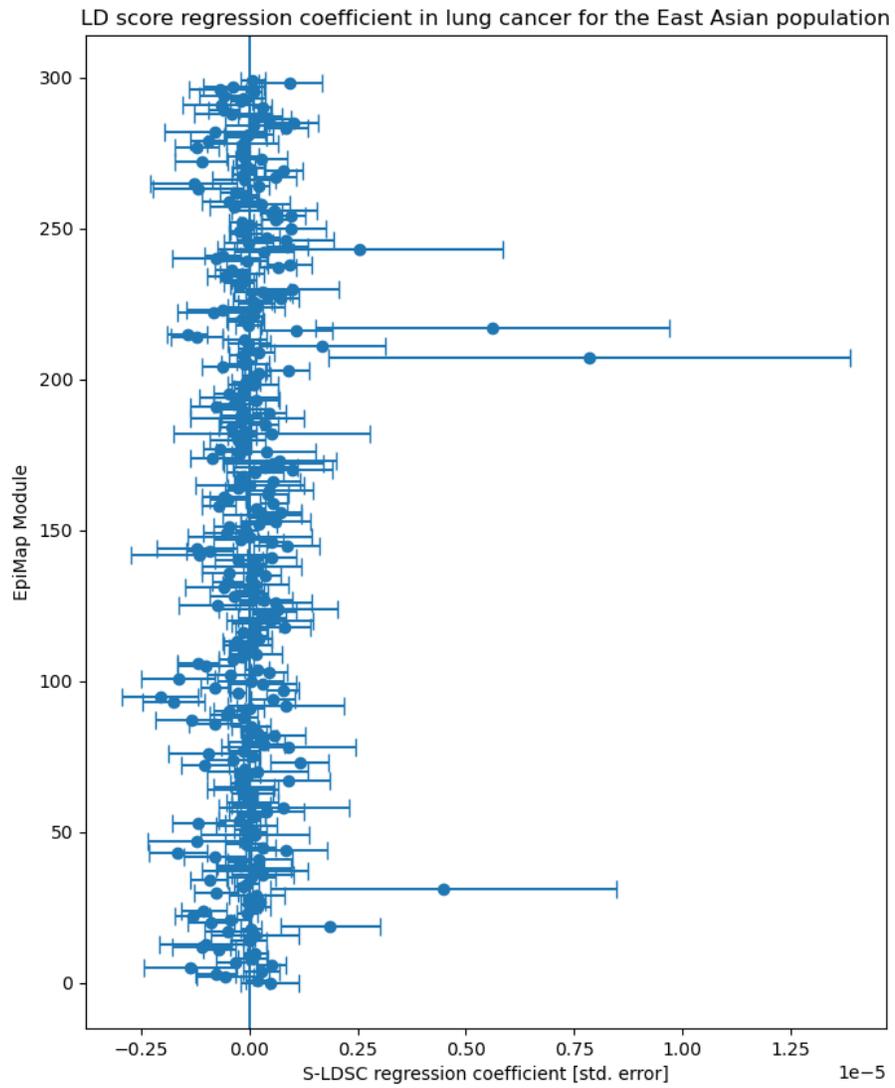


Figure 4-23: p-value of regression coefficient in lung cancer for the European population

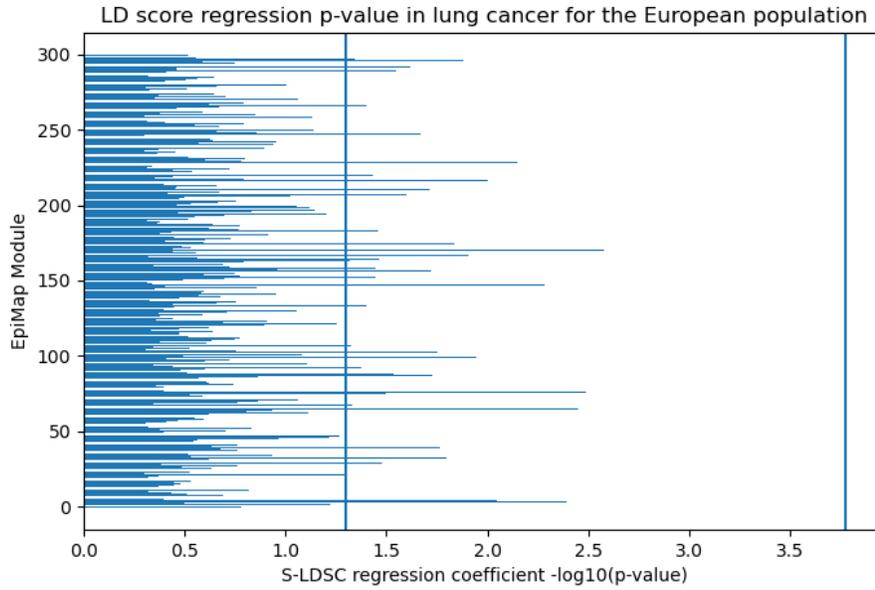
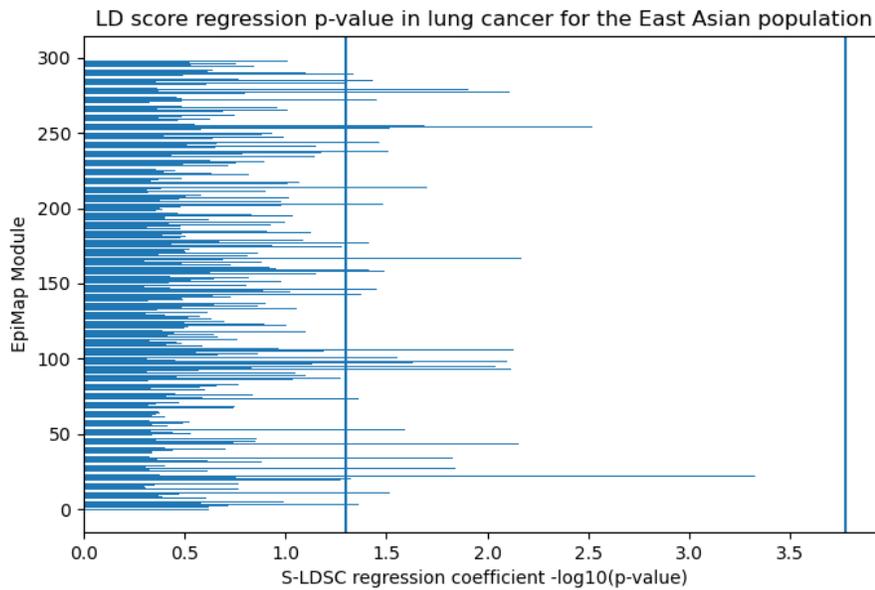


Figure 4-24: p-value of regression coefficient in lung cancer for the East Asian population



### 4.3.2 Smoking

Figure 4-25: Heritability enrichment in smoking for the European population

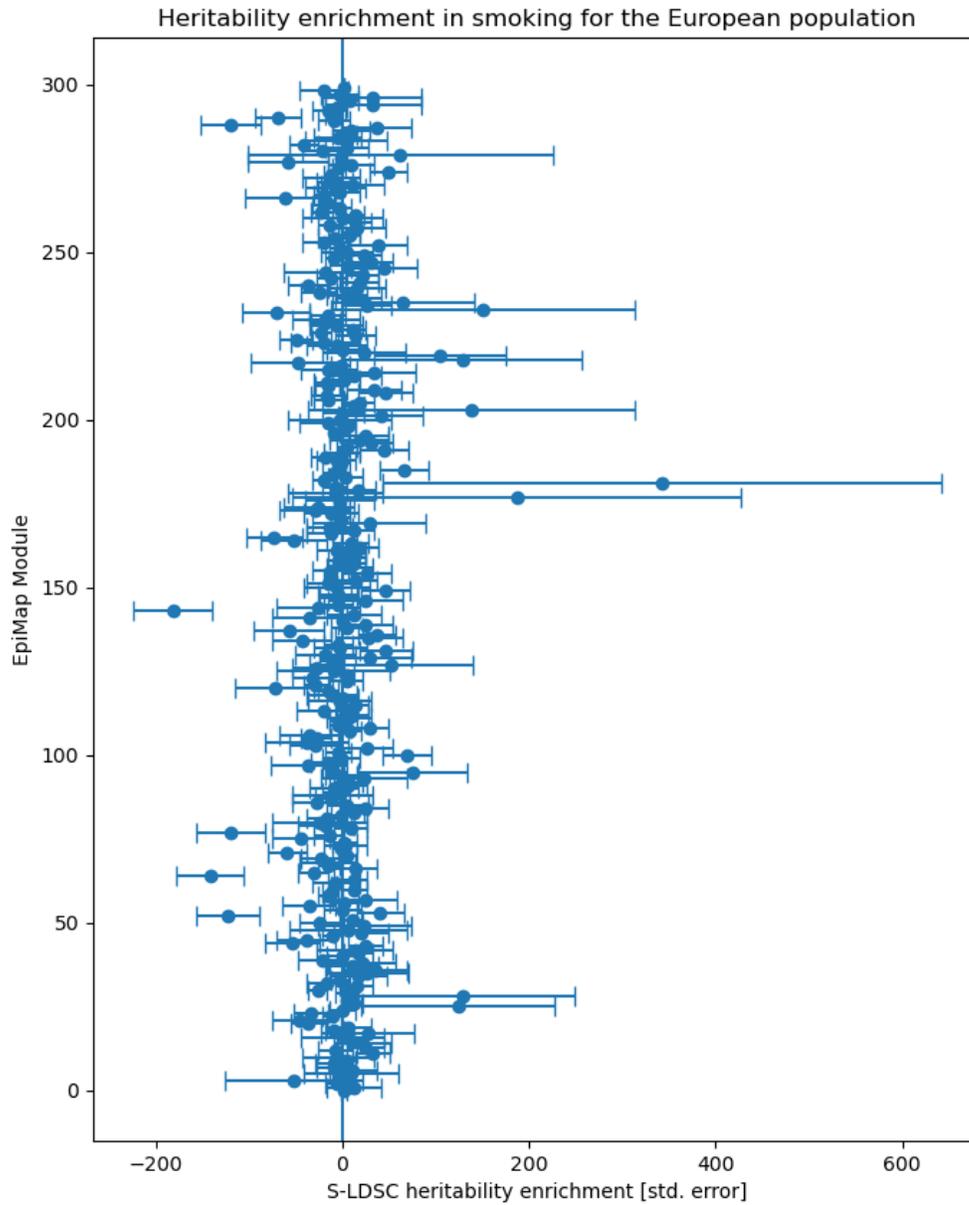




Figure 4-27: p-value of heritability enrichment in smoking for the European population

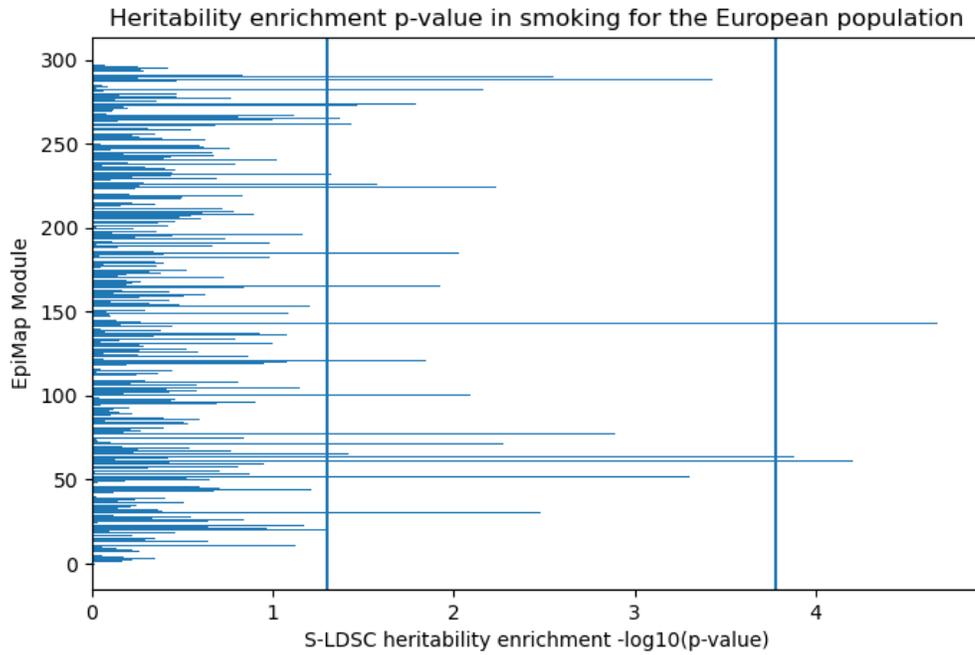
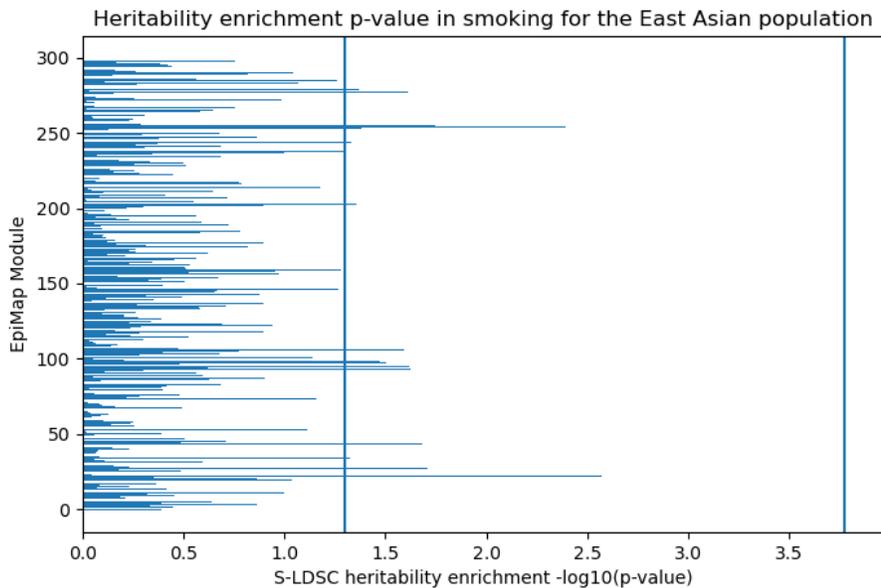


Figure 4-28: p-value of heritability enrichment in smoking for the East Asian population



In the European population, we see a few modules pass the significance threshold after multiple hypothesis testing correction but, again, in the East Asian population none of the annotations show significance after the Bonferroni correction. There is also an extremely different distribution of enrichment values and p-values between the annotations between the East Asian and European populations. We also see a large heritability enrichment for multiple modules in the European populations in Figure 4-25. However, for the East Asian population, we see very small (centred around 0) enrichment values in Figure 4-26.

Similarly to what I did for the lung cancer analysis, I have plotted the error bars for the regression coefficients (corresponding to how much heritability is enhanced if a given SNP is part of that annotation module relative to the baseline model<sup>25</sup>). Figure 4-30 shows a distribution of regression coefficients that isn't entirely centered around 0 compared to Figure 4-26. This lets us indirectly compare the enrichment distribution among the two populations.

Figure 4-29: Regression coefficient in smoking for the European population

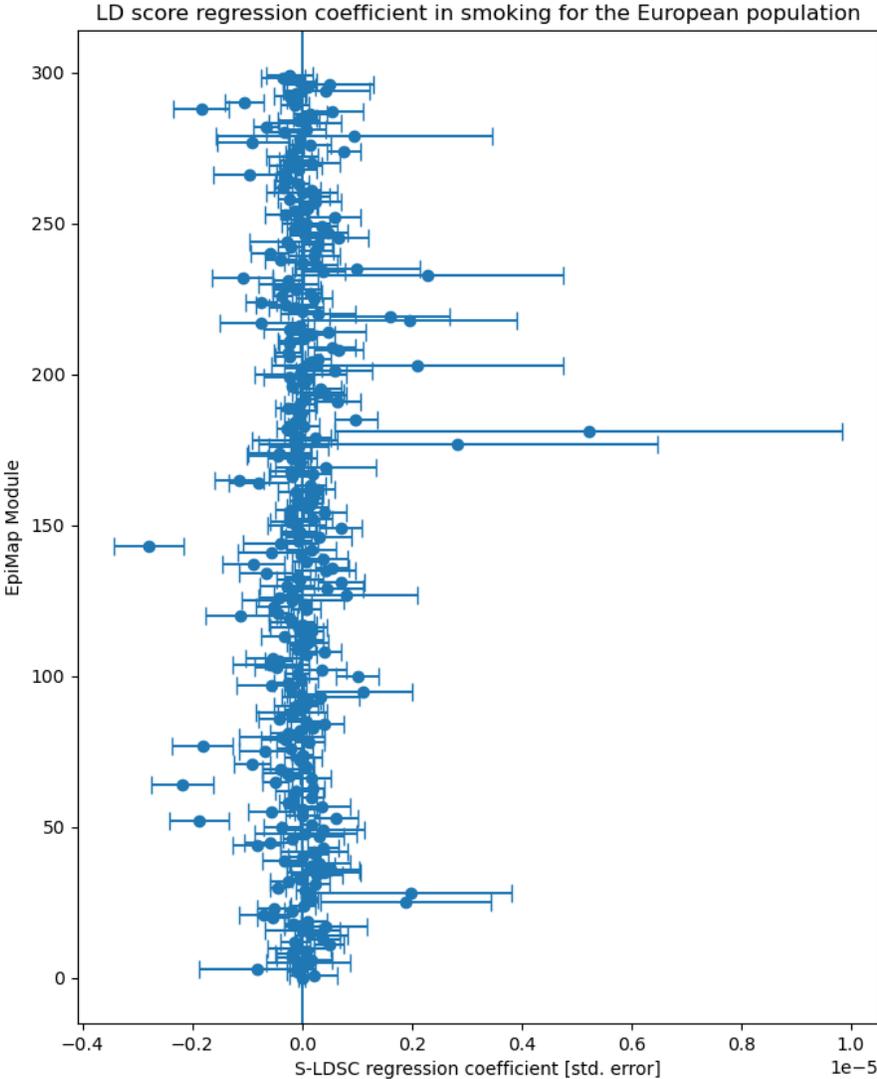


Figure 4-30: Regression coefficient in smoking for the East Asian population

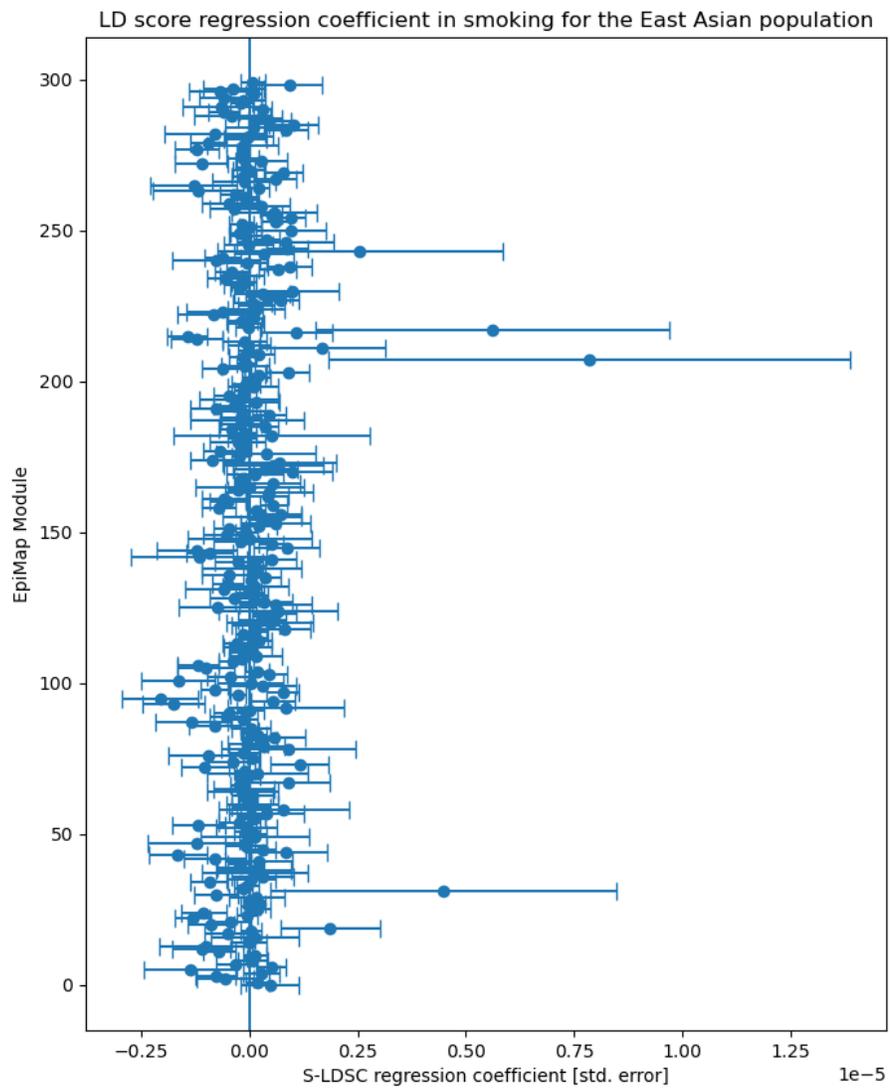


Figure 4-31: p-value of regression coefficient in smoking for the European population

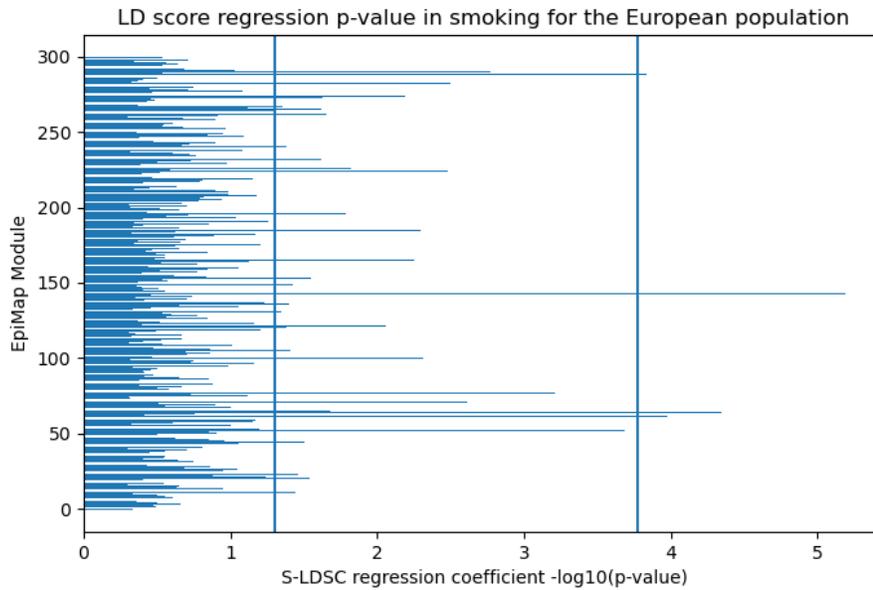
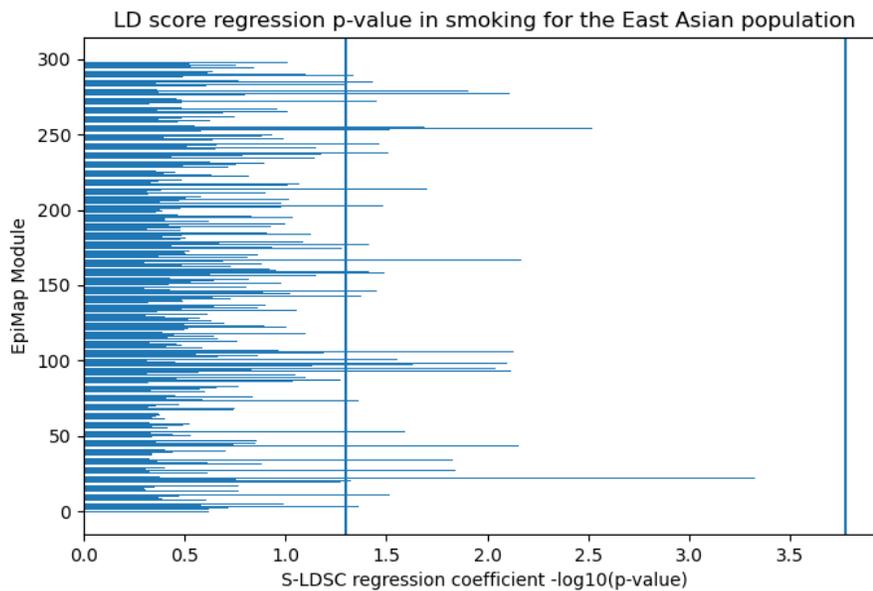


Figure 4-32: p-value of regression coefficient in smoking for the East Asian population



After obtaining the most enriched modules, I have used the gene ontology enrichment data from EpiMap<sup>23</sup> for these modules to understand more about the function and biological meaning of these highly associated modules.

Figure 4-33: Most significantly enriched GO terms in lung cancer for the European population

| EpiMap module | GO enrichment                                     | GO enrichment                                      | GO enrichment                                     | GO enrichment                                     | GO enrichment                                     |
|---------------|---|--|---|---|---|
| 170           | negative regulation of macromolecule metabolic... | cytosol  | modification of morphology or physiology of ot... | modification of morphology or physiology of ot... | natural killer cell activation                    |
| 76            | embryo development                                | embryo development ending in birth or egg hatching | chordate embryonic development                    | tube formation                                    | morphogenesis of embryonic epithelium             |
| 65            | anatomical structure morphogenesis                | tissue development                                 | cell development                                  | cardiovascular system development                 | regulation of cellular component movement         |
| 3             | respiratory tube development                      | lung development                                   | respiratory system development                    | organ morphogenesis                               | inner ear development                             |
| 147           | regulation of cell development                    | postsynaptic density                               | regulation of skeletal muscle fiber development   | positive regulation of developmental process      | negative regulation of transcription, DNA-depe... |
| 228           | synapse part                                      | dendrite   | postsynaptic membrane                             | regulation of neuron differentiation              | telencephalon development                         |
| 4             | anatomical structure morphogenesis                | cell development                                   | cell morphogenesis involved in differentiation    | neuron projection morphogenesis                   | neurogenesis                                      |
| 167           | anatomical structure morphogenesis                | cardiovascular system development                  | vasculature development                           | blood vessel development                          | regulation of locomotion                          |
| 99            | anatomical structure morphogenesis                | tissue morphogenesis                               | vasculature development                           | tissue development                                | heart development                                 |
| 216           | regulation of multicellular organismal develop... | central nervous system development                 | neuron differentiation                            | generation of neurons                             | chemotaxis  |

Figure 4-34: Most significantly enriched GO terms in lung cancer for the East Asian population

| EpiMap module | GO enrichment                                | GO enrichment                                     | GO enrichment                                     | GO enrichment                                     | GO enrichment                                     |
|---------------|--|---|---|---|---|
| 124           | anatomical structure morphogenesis           | skeletal system development                       | blood vessel development                          | vasculature development                           | cardiovascular system development                 |
| 139           | renal system development                     | metanephros development                           | urogenital system development                     | kidney development                                | renal tubule development                          |
| 181           | lipid binding                                | response to growth factor stimulus                | positive regulation of intracellular protein k... | endocytosis                                       | negative regulation of cysteine-type endopepti... |
| 16            | anatomical structure morphogenesis           | tissue morphogenesis                              | cardiovascular system development                 | tissue development                                | cell development                                  |
| 174           | sequence-specific DNA binding                | axon part   | cellular macromolecule biosynthetic process       | regulation of RNA metabolic process               | energy reserve metabolic process                  |
| 67            | vasculature development                      | blood vessel development                          | regulation of locomotion                          | cardiovascular system development                 | proteinaceous extracellular matrix                |
| 117           | regulation of mesenchymal cell proliferation | cell leading edge                                 | phospholipid binding                              | regulation of locomotion                          | regulation of cell migration                      |
| 89            | organic acid metabolic process               | oxoacid metabolic process                         | enzyme linked receptor protein signaling pathway  | transmembrane receptor protein tyrosine kinase... | positive regulation of MAPK cascade               |
| 5             | mRNA metabolic process                       | nuclear part                                      | RNA binding                                       | nuclear lumen                                     | nucleoplasm                                       |
| 45            | anatomical structure morphogenesis           | anatomical structure formation involved in mor... | tissue development                                | cardiovascular system development                 | regulation of multicellular organismal develop... |

Figure 4-35: Most significantly enriched GO terms in smoking for the European population

| EpiMap module | GO enrichment                                       | GO enrichment                                      | GO enrichment                                | GO enrichment   | GO enrichment  |
|---------------|---|--|--|---|--|
| 143           | positive regulation of transcription, DNA-dependent | positive regulation of gene expression             | positive regulation of RNA metabolic process | positive regulation of macromolecule biosynthetic process | positive regulation of transcription from RNA polymerase II promoter |
| 61            | neuron projection morphogenesis                     | cell projection organization                       | cell development                             | cell morphogenesis involved in neuron differentiation     | axonogenesis   |
| 64            | multicellular organismal signaling                  | transmission of nerve impulse                      | synaptic transmission                        | neurogenesis  | neuron differentiation   |
| 288           | blood vessel morphogenesis                          | blood vessel development                           | angiogenesis                                 | vasculature development                                   | anatomical structure morphogenesis                                   |
| 52            | tissue development                                  | regulation of locomotion                           | regulation of cell motility                  | regulation of cell migration                              | regulation of cellular component movement                            |
| 77            | cardiovascular system development                   | blood vessel development                           | vasculature development                      | anatomical structure morphogenesis                        | blood vessel morphogenesis   |
| 290           | anatomical structure morphogenesis                  | cardiovascular system development                  | embryo development                           | organ morphogenesis                                       | tube development   |
| 30            | positive regulation of gene expression              | embryo development ending in birth or egg hatching | chordate embryonic development               | in utero embryonic development                            | positive regulation of RNA metabolic process                         |
| 71            | immune system process                               | regulation of immune system process                | regulation of immune response                | immune response   | positive regulation of immune system process                         |
| 224           | myofibril   | contractile fiber part                             | contractile fiber                            | sarcomere   | I band   |

Figure 4-36: Most significantly enriched GO terms in smoking for the East Asian population

| EpiMap module | GO enrichment                      | GO enrichment                      | GO enrichment                            | GO enrichment   | GO enrichment  |
|---------------|------------------------------------|------------------------------------|--|---|--|
| 22            | extracellular matrix               | proteinaceous extracellular matrix | cell projection morphogenesis            | regulation of nucleotide metabolic process                  | anatomical structure formation involved in morphogenesis |
| 254           | cell development                   | neuron differentiation             | anatomical structure morphogenesis       | generation of neurons                                       | neurogenesis   |
| 215           | response to xenobiotic stimulus    | xenobiotic metabolic process       | cellular response to xenobiotic stimulus | regulation of transcription from RNA polymerase II promoter | negative regulation of response to stimulus              |
| 255           | generation of neurons              | neuron differentiation             | neuron development                       | transmission of nerve impulse                               | neuron projection morphogenesis                          |
| 27            | anatomical structure morphogenesis | neurogenesis                       | generation of neurons                    | central nervous system development                          | organ morphogenesis                                      |
| 43            | cardiovascular system development  | tissue development                 | vasculature development                  | negative regulation of cellular biosynthetic process        | heart development  |
| 93            | blood vessel development           | vasculature development            | blood vessel morphogenesis               | lipid binding   | regulation of locomotion                                 |
| 95            | anatomical structure morphogenesis | tissue development                 | embryonic morphogenesis                  | regulation of multicellular organismal development          | positive regulation of cell differentiation              |
| 277           | organic acid metabolic process     | oxoacid metabolic process          | carboxylic acid metabolic process        | sterol metabolic process                                    | steroid metabolic process                                |
| 106           | myofibril                          | contractile fiber                  | positive regulation of gene expression   | sarcomere   | contractile fiber part                                   |

### 4.3.3 Discussion

Like in the previous sections, we see a stark difference between the distributions of highly associated annotations between the European and the East Asian population. In addition, we can see that the GO enrichment terms differ between the two populations in a similar pattern as in the last two sections. For lung cancer, in the European population the top enrichments indicate morphogenesis, tissue development and nervous-system-related associations, while in the East Asian population the top enrichments indicate morphogenesis and epithelial tissue development. For smoking,

we see immune-related enrichment present in European associations but not as much in the East Asian ones.

# Chapter 5

## Conclusion

### 5.1 Future work

There are a few interesting avenues that can be pursued in the future to extend the work from this thesis.

The first future direction can be to perform colocalization between smoking and lung cancer GWAS data. We might find that a specific SNP is highly associated with both lung cancer and smoking in one population but in the other population it is only highly associated with smoking and not lung cancer. One aspect to consider in this analysis is that due to linkage disequilibrium multiple nearby loci will be clustered together due to their correlation to each other.

I have performed preliminary analysis that highlights the difference between the two plots visually: Figure 5-1 from European ancestry is showing a few loci with high association in both lung cancer and smoking while Figure 5-2 from East Asian ancestry one does not. I hypothesize that this might be the cause of East Asian smokers having a lower increase in risk of lung cancer than European smokers. Perhaps some of the associated loci are related to addiction (and thus might not be directly associated with lung cancer) and others are associated with both lung cancer and smoking through common pathways.

Figure 5-1: Colocalizing the significant loci between the European smoking and lung cancer GWAS

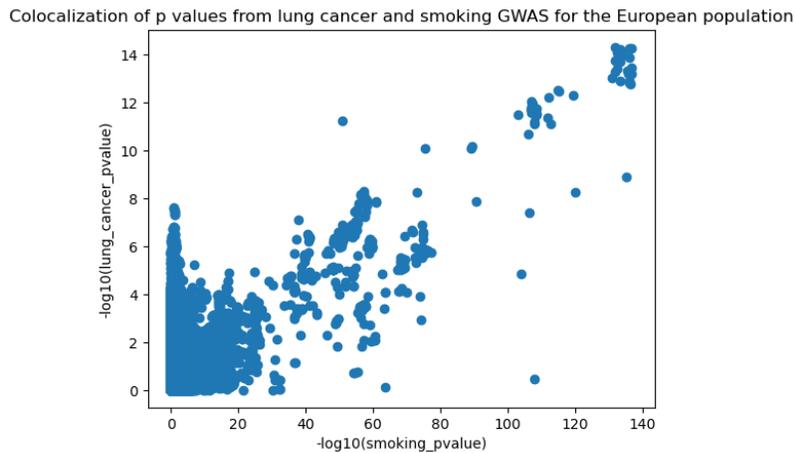
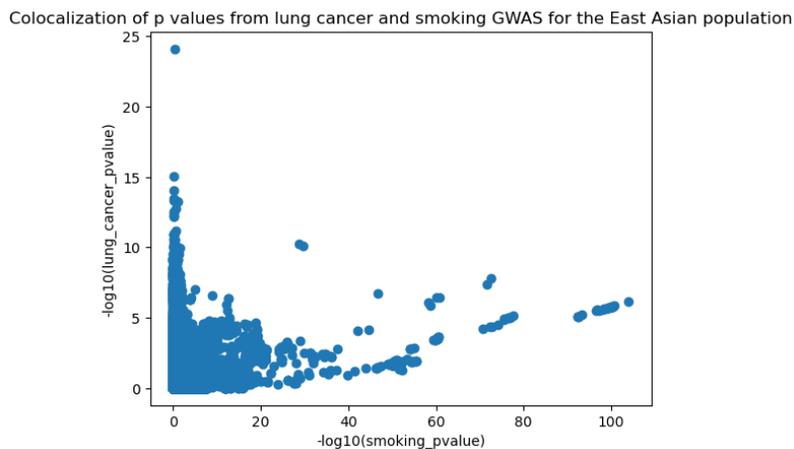


Figure 5-2: Colocalizing the significant loci between the East Asian smoking and lung cancer GWAS



Another direction can be generating polygenic scores. It would be interesting to see if explicitly taking into account the ancestry information or the significant SNPs associated with only one ancestry could be of help in improving the polygenic risk scores for lung cancer and even potentially for smoking. So far, the performance of polygenic scores on lung cancer biobank data has not had great predictive power<sup>29</sup>.

We can also compare the results from analyzing the smoking GWAS to GWAS of other addictive behaviors. We could try to verify if there is a common thread of addiction related loci only present in one population and not the other.

Finally, it can be helpful to perform colocalization of smoking and lung cancer GWAS data with eQTL data. This can be done using eCAVIAR<sup>30</sup>, for example. Buyn et al. have done this for SNPs that are common across the ancestries that they considered<sup>9</sup> but it would be helpful to perform this analysis for loci that are not common across ancestries to uncover differences between them.

## 5.2 Summary

This work highlights the large differences between germline variation in lung cancer patients from East Asian and European background, suggesting different causal variants and mechanisms for the disease. We analyze GWAS data from three biobanks across European and East Asian ancestry and use gene expression and epigenetic annotations to uncover differences in associated cell-types and biological processes. More data from different ancestries is needed to fully understand disease heterogeneity across the populations and extensive experimental work is needed to verify the loci and pathways that are found to be associated with the disease in studies similar to this one.

THIS PAGE INTENTIONALLY LEFT BLANK

# Bibliography

- [1] Kurki, M. *et al.* FinnGen: Unique genetic insights from combining isolated population and national health register data. *medRxiv* (2022).
- [2] Nagai, A., Hirata, M., Kamatani, Y. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol.* **27**, S2–S8 (2017).
- [3] World Health Organization cancer fact-sheet. <https://www.who.int/news-room/fact-sheets/detail/cancer> (2022).
- [4] Spiro, S. & Silvestri, G. One hundred years of lung cancer. *American Journal of Respiratory and Critical Care Medicine* **172**, 523–529 (2005).
- [5] Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. Non-small cell lung cancer: Epidemiology, risk factors, treatment, and survivorship. *Mayo Clinic Proceedings* **83**, 584–594 (2008).
- [6] Sekido, Y., Fong, K. & Minna, J. Molecular genetics of lung cancer. *Annual Review of Medicine* **54**, 73–87 (2003).
- [7] Brennan, P., Hainaut, P. & Boffetta, P. Genetics of lung-cancer susceptibility. *The Lancet Oncology* **12**, 399–408 (2011).
- [8] Risch, A. & Plass, C. Lung cancer epigenetics and genetics. *International Journal of Cancer* **123**, 1–7 (2008).
- [9] Byun, J., Han, Y., Li, Y. *et al.* Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. *Nature Genetics* **54**, 1167–1177 (2022).
- [10] Stellman, S., Takezaki, T., Wang, L. *et al.* Smoking and lung cancer risk in American and Japanese men: an international case-control study. *Cancer Epidemiology Biomarkers Prev.* **10**, 1193–1199 (2001).
- [11] Soo, R. *et al.* Differences in outcome and toxicity between Asian and caucasian patients with lung cancer treated with systemic therapy. *Future Oncology* **8**, 451–462 (2012).
- [12] Zhou, W. & Christiani, D. East meets West: ethnic differences in epidemiology and clinical behaviors of lung cancer between East Asians and Caucasians. *Chinese Journal of Cancer* **30**, 287–292 (2011).

- [13] Leal, L., de Paula, F., Marchi, P. D. *et al.* Mutational profile of brazilian lung adenocarcinoma unveils association of EGFR mutations with high asian ancestry and independent prognostic role of KRAS mutations. *Scientific Reports volume 9*, 3209 (2019).
- [14] J. Carrot-Zhang, N. P., G. Soca-Chafre *et al.* Genetic ancestry contributes to somatic mutations in lung cancers from admixed latin american populations. *Cancer Discovery 11*, 591–598 (2021).
- [15] Sakaue, S., Kanai, M., Tanigawa, Y. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics 53*, 1415–1424 (2021).
- [16] Pan-UK Biobank. <https://pan.ukbb.broadinstitute.org> (2020).
- [17] Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv* (2021).
- [18] Sudlow, C. *et al.* UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine 12*, 1–10 (2015).
- [19] Finucane, H., Reshef, Y., Anttila, V. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics 50*, 621–629 (2018).
- [20] Lonsdale, J., Thomas, J., Salvatore, M. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics 45*, 580–585 (2013).
- [21] The GTEx Consortium, Aguet, F., Anand, S., Ardlie, K. *et al.* The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science 369*, 1318–1330 (2020). URL <https://www.science.org/doi/abs/10.1126/science.aaz1776>.
- [22] Fehrmann, R., Karjalainen, J., Krajewska, M. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics 47*, 115–125 (2015).
- [23] Boix, C., James, B., Park, Y. *et al.* Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature 590*, 300–307 (2021).
- [24] Bulik-Sullivan, B., Loh, P., Finucane, H. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics 47*, 291–295 (2015).
- [25] Finucane, H., Bulik-Sullivan, B., Gusev, A. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics 47*, 1228–1235 (2015).

- [26] Bernstein, B., Stamatoyannopoulos, J., Costello, J. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology* **28**, 1045–1048 (2010).
- [27] Fry, A., Littlejohns, T., Sudlow, C. *et al.* Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American Journal of Epidemiology* **186**, 1026–1034 (2017).
- [28] Bulik-Sullivan, B. & Finucane, H. LDSC (LD SCore). <https://github.com/bulik/ldsc> (2020).
- [29] Tanigawa, Y. *et al.* Significant sparse polygenic risk scores across 813 traits in UK Biobank. *PLOS Genetics* **18**, 1–21 (2022).
- [30] Hormozdiari, F., van de Bunt, M., Segrè, A. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *American Journal of Human Genetics* **99**, 1245–1260 (2016).