University of Alberta

Identification of Shared and Distinct Gene-disease Associations Among Multiple Related Diseases and Multiple Subtypes of a Disease

by

Conrado Franco-Villalobos

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Science in Epidemiology

School of Public Health

Fall 2013
Edmonton, Alberta

Abstract

Genome-wide association study (GWAS) is an approach with high-throughput genotyping to uncover genetic susceptibilities of complex diseases. However, the genetic susceptibilities discovered usually carry very small risk increments. Additionally, the current approach to assess whether these genetic associations are shared among a group of diseases relies mainly on statistical significance alone, ignoring biologically relevant information such as magnitude and direction of the associations.

The methodology proposed takes into account not only strength and direction of the associations but also the resemblance of the biological mechanism by using logic regression to generate a graphical representation of the similarity of the associations. We found evidence that 149 genetic associations have certain degree of uniqueness with Crohn's Disease, Rheumatoid Arthritis, and Type I Diabetes while 11 were shared between at least 2 diseases. Additionally, the gene-level analysis of TB cases stratified by age, strain, and lineage identified 3 new susceptibility genes (*ZFHX1B*, *FER*, and *FAM77*) associated with different TB subgroups.

Table of Contents

1 Introduction **1**
- 1.1 Thesis Organization . . . 1
- 1.2 Rationale . . . 1
- 1.3 Research Questions . . . 3
- 1.4 Hypotheses . . . 3

2 Identification of shared and distinct genes: Crohns Disease, Type I Diabetes & Rheumatoid Arthritis in WTCCC data **7**
- 2.1 Introduction . . . 7
- 2.2 Materials and Methods . . . 9
  - 2.2.1 Logic Regression Gene-level Association Analysis . . . 9
  - 2.2.2 Shared vs. Distinct Gene Analysis using Logic Regression 11
- 2.3 Results . . . 13
- 2.4 Discussion . . . 18

3 Novel Tuberculosis Susceptibility Genes Discovered by Logic Regression: a Stratified Analysis in Thai Population **22**
- 3.1 Introduction . . . 22
- 3.2 Materials and Methods . . . 23
- 3.3 Results . . . 25
  - 3.3.1 Age-stratified analysis . . . 26
  - 3.3.2 Strain-stratified analysis . . . 29
  - 3.3.3 Lineage-stratified analysis . . . 31
- 3.4 Discussion . . . 35

4 Conclusions **40**
- 4.1 Review of Hypotheses . . . 40
- 4.2 Discussion . . . 41
- 4.3 Future Work . . . 43
- 4.4 Conclusions . . . 44

A R codes **46**
A.1 Logic Regression – WTCCC 46
A.2 Logic Regression – TB 48

B MATLAB codes **51**
B.1 2D distance error calculation 51
B.2 3D distance error calculation 52
B.3 2D Unconstrained Nonlinear Optimization 54
B.4 3D Unconstrained Nonlinear Optimization 54

List of Tables

2.1 **Logic structures, frequencies, and associated Crohn's Disease odds ratios of the HLA-DRA gene** 15
2.2 **Logic structures, frequencies, and associated Type I Diabetes odds ratios of the HLA-DRA gene** 15
2.3 **Logic structures, frequencies, and associated Rheumatoid Arthritis odds ratios of the HLA-DRA gene** 16
2.4 **Logic structures, frequencies, and associated Crohn's Disease odds ratios of the PTPN22 gene** 17
2.5 **Logic structures, frequencies, and associated Type I Diabetes odds ratios of the PTPN22 gene** 17
2.6 **Logic structures, frequencies, and associated Rheumatoid Arthritis odds ratios of the PTPN22 gene** 17

3.1 **Number of old TB cases by lineage and strain.** 26
3.2 **Number of young TB cases by lineage and strain.** . . . 26
3.3 **Genes with the strongest evidence of association with either young or old TB risk with chromosomal locations and approximate p-values.** 27
3.4 **Logic structures, frequencies, and associated old-TB odds ratios of the ZFHX1B gene** 27
3.5 **Logic structures, frequencies, and associated young-TB odds ratios of the ZFHX1B gene** 27
3.6 **Logic structures, frequencies, and associated old-TB odds ratios of the FER gene** 28
3.7 **Logic structures, frequencies, and associated young-TB odds ratios of the FER gene** 28
3.8 **Genes with the strongest evidence of association with either ancient or modern TB strain risk with chromosomal locations and approximate p-values.** 30
3.9 **Logic structures, frequencies, and associated ancient TB strain odds ratios of the ZFHX1B gene** 30
3.10 **Logic structures, frequencies, and associated modern TB strain odds ratios of the ZFHX1B gene** 31

3.11 **Logic structures, frequencies, and associated ancient TB strain odds ratios of the FER gene** 31
3.12 **Logic structures, frequencies, and associated modern TB strain odds ratios of the FER gene** 31
3.13 **Genes with the strongest evidence of association with TB risk in Beijing, EAI, and *other* lineages with chromosomal locations and approximate p-values.** 32
3.14 **Logic structures, frequencies, and associated EAI-lineage TB odds ratios of the ZFHX1B gene** 33
3.15 **Logic structures, frequencies, and associated Beijing-lineage TB odds ratios of the ZFHX1B gene** 33
3.16 **Logic structures, frequencies, and associated *other*-lineage TB odds ratios of the ZFHX1B gene** 34
3.17 **Logic structures, frequencies, and associated EAI-lineage TB odds ratios of the FER gene** 34
3.18 **Logic structures, frequencies, and associated Beijing-lineage TB odds ratios of the FER gene** 34
3.19 **Logic structures, frequencies, and associated *other*-lineage TB odds ratios of the FER gene** 34

List of Figures

2.1 Distance plot of gene *HLA-DRA*. 15
2.2 Distance plot of gene *PTPN22*. 16

3.1 Distance plots of genes *ZFHX1B* and *FER* for age-stratified TB subgroups. 28
3.2 Distance plots of genes *ZFHX1B* and *FER* for strain-stratified TB subgroups. 30
3.3 Distance plots of genes *ZFHX1B* and *FER* for lineage-stratified TB subgroups. 33

List of Abbreviations

GWAS	*Genome-Wide Association Study*
SNP	*Single-Nucleotide Polymorphism*
CD	*Crohn's Disease*
T1D	*Type I Diabetes*
RA	*Rheumatoid Arthritis*
TB	*Tuberculosis*
WTCCC	*Wellcome Trust Case Control Consortium*
EAI	*East African Indian*
GVS	*Genetic Variation Score*
BF	*Bayes Factor*

Chapter 1

Introduction

1.1 Thesis Organization

This paper-based thesis was prepared in accordance to the Faculty of Graduate Studies and Research (FGSR) of the University of Alberta guidelines. The thesis is organized as follows:

Chapter 2 - First manuscript

Identification of shared and distinct genes: Crohns Disease, Type I Diabetes & Rheumatoid Arthritis in WTCCC data

Chapter 3 - Second manuscript

Novel Tuberculosis Susceptibility Genes Discovered by Logic Regression: a Stratified Analysis in Thai Population

Chapter 4 - Summary and Conclusions

1.2 Rationale

Autoimmune diseases are chronic conditions that involve an inappropriate response of the body to non-harmful substances and tissues in the body. This type of disorders are thought to arise due to a combination of genetic and environmental factors. The familial clustering of autoimmune diseases as well as

association of multiple disorders in single individuals suggest that there might be common genetic susceptibility factors shared among this type of diseases [1].

Genome-wide association studies (GWAS) is a high-throughput approach currently used to uncover genetic susceptibilities of complex diseases by examining many genetic variants of individuals to assess if any of them is associated with a trait [2]. GWAS has also been used to assess shared and/or unique genetic susceptibilities of multiple diseases that are hypothesized to be biologically related [3] [4] [5]. However, the current approach widely used in GWAS to assess whether genetic variants are commonly or uniquely associated with multiple diseases overlooks relevant information by focusing largely or sorely on statistical significance.

To our knowledge, the current approach to assess whether genetic associations are shared among a group of diseases fails to take into consideration relevant information such as the magnitude and direction of the genetic effect which could potentially lead to inaccurate inferences about the sharedness of the associations. Chapter 2 of this thesis proposes a method that aims to provide a better insight of the genetic associations with a group of diseases by not only taking into account statistical significance, but also strength, direction, and similarity of the biological association. This Chapter also presents a test of the methodology using the Wellcome Trust Case Control Consortium (WTCCC) GWAS data of Crohn's Disease, Rheumatoid Arthritis, and Type I Diabetes [6].

GWASs have identified thousands of genetic variants associated with complex diseases and traits providing a better understanding of their genetic etiology. However, most of these genetic variants carry small risk increments which can only explain a small proportion of the clustering observed in family studies leading to a phenomenon called *missing heritability* [7]. Most human geneticists hypothesize that additional variants that have not been discovered can provide the explanation of this phenomenon. Specifically, many quantitative geneticists and biologists recognize that interactions might be responsible

for the missing heritability phenomenon since they can greatly affect the heritability calculations and they are rarely investigated in GWASs [8].

Chapter 3 investigates whether genetic interactions can uncover novel susceptibility loci for tuberculosis (TB) in Thai. The analysis was performed on different subgroups stratified by age (> 45 years, ≤ 45 years) [9], TB strain (ancient, modern), or TB lineage (Beijing, EAI, *other*) with shared controls. An additional analysis was performed to assess whether the associations uncovered by the stratified GWASs were shared with other subgroups.

1.3 Research Questions

Chapter 2

1. Do p-values provide enough information to assess whether a genetic associations is shared among a group of diseases?
2. Does taking into account strength, direction, and similarity of the biological association provide better insight of the sharedness of genetic associations among a group of diseases?

Chapter 3

1. Does stratified analysis of SNP-SNP interactions uncover new genetic TB susceptibilities in Thai?
2. Are the newly discovered genetic susceptibilities shared among different TB subgroups?

1.4 Hypotheses

Chapter 2

- P-values alone do not provide enough information to assess whether a genetic association is shared among a group of diseases and can lead to inaccurate inferences.
- Taking into account strength, direction, and similarity of the biological association provide better insight and a stricter definition of the sharedness of genetic associations among a group of diseases.

Chapter 3

- SNP-SNP interactions are responsible for a proportion of the TB susceptibility and explain to a greater extent the TB genetics.
- Some of the genetic susceptibilities are shared while others are unique for certain TB subgroups.

Bibliography

[1] Kevin G Becker, Richard M Simon, Joan E Bailey-Wilson, Boris Freidlin, William E Biddison, Henry F McFarland, and Jeffrey M Trent. Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases. *Proceedings of the National Academy of Sciences*, 95(17):9979–9984, 1998.

[2] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[3] Alexandra Zhernakova, Eli A Stahl, Gosia Trynka, Soumya Raychaudhuri, Eleanora A Festen, Lude Franke, Harm-Jan Westra, Rudolf SN Fehrmann, Fina AS Kurreeman, Brian Thomson, et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-hla shared loci. *PLoS genetics*, 7(2):e1002004, 2011.

[4] Paula S Ramos, Lindsey A Criswell, Kathy L Moser, Mary E Comeau, Adrienne H Williams, Nicholas M Pajewski, Sharon A Chung, Robert R Graham, Raphael Zidovetzki, Jennifer A Kelly, et al. A comprehensive analysis of shared loci between systemic lupus erythematosus (sle) and sixteen autoimmune diseases reveals limited genetic overlap. *PLoS genetics*, 7(12):e1002406, 2011.

[5] Christian C Abnet, Neal D Freedman, Nan Hu, Zhaoming Wang, Kai Yu, Xiao-Ou Shu, Jian-Min Yuan, Wei Zheng, Sanford M Dawsey, Linda M Dong, et al. A shared susceptibility locus in plce1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nature genetics*, 42(9):764–767, 2010.

[6] P.R. Burton, D.G. Clayton, L.R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D.P. Kwiatkowski, M.I. McCarthy, W.H. Ouwehand, N.J. Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[7] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[8] O. Zuk, E. Hechter, S.R. Sunyaev, and E.S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[9] Surakameth Mahasirimongkol, Hideki Yanai, Taisei Mushiroda, Watoo Promphittayarat, Sukanya Wattanapokayakit, Jurairat Phromjai, Rika Yuliwulandari, Nuanjun Wichukchinda, Amara Yowang, Norio Yamada, et al. Genome-wide association studies of tuberculosis in asians identify distinct at-risk locus for young tuberculosis. *Journal of human genetics*, 57(6):363–367, 2012.

Chapter 2

Identification of shared and distinct genes: Crohns Disease, Type I Diabetes & Rheumatoid Arthritis in WTCCC data

2.1 Introduction

Autoimmune diseases, such as Crohn's Disease (CD), Type I Diabetes (T1D), and Rheumatoid Arthritis (RA), are hypothesized to share similar genetic factors given the disease mechanisms and the observed clustering of diseases within families [1]. Crohn's disease is a type of inflammatory bowel disease characterized by the presence of abdominal pain, fever, and bowel obstruction or diarrhea with passage of blood, mocus or both [2]. Type I Diabetes results from an autoimmune destruction of β-cells in the pancreas which causes defects in insulin secretion, action, or both [3]. Rheumatoid Arthritis is an inflammatory disease that affects mainly the joints of the hands and feet as the result of over-expressed degradative enzymes which destroy articular tissues [4].

Although it has been suggested that CD, T1D, and RA share similar genetic etiology due to their autoimmune nature and a number of articles has been published addressing this scientific hypothesis, most of the approaches taken to identify common susceptibility loci has been rather simplistic and sta-

tistically oriented with a lack of biological relevance [5] [6]. The most common type of study involves performing a systematic review of previously reported susceptibility loci in genome-wide association studies (GWAS) for each individual disease and compare the findings across the diseases. If a loci was found to be significantly associated with two or more diseases, the association is considered "shared", if it was significantly associated with a single disease, it is considered "distinct". For example, it has been reported that gene *PTPN22* is commonly associated with CD, T1D, and RA [7]. The main limitation about this type of studies is that they don't take into consideration the strength and direction of the associations as well as the commonality of the SNPs encompassed in the reported genes.

Another approach, similar to the one discussed previously, has gone further in understanding the genetic etiology of autoimmune diseases. It has been reported that SNPs/genes *rs917997/IL18RAP* and *rs1738074/TAGAP* are commonly associated with Type I Diabetes and Celiac Disease by analyzing the association of SNP-disease independently of the other disease. The authors go further in the discussion section by noting that the minor allele in both SNPs is negatively associated with Type I Diabetes but positively associated with Celiac Disease suggesting opposite biological effects [8]. However, this approach still lacks an analysis of the strength of the associations which might indicate a minor/major role in the underlying biological mechanism depending on the disease.

Novel approaches that have tried to address the issues of incorporating the strength and direction of the associations include the calculation of a *Genetic Variation Score (GVS)* for each disease-SNP pair given by $GVS[d, s] = sign(log(OR[d, s])) \times log(p\text{-}value[d, s])$ where d, s, and OR represent a specific disease, a specific SNP, and the odds ratio of the association, respectively. After calculating the GVS vector for each disease, a correlation coefficient is calculated as an estimate of the degree of genetic concordance between pairs of diseases. In the formula, the term $sign(log(OR[d, s]))$ captures the direc-

tion of the association while the $log(p\text{-}value[d, s])$ term is supposed to capture the strength of the association [9]. Although the direction of the association is effectively captured by this score, there is a statistical misconception that *p-value* indicates the strength of an association while it is a function of the sample size and can be changed by the sample size. Furthermore, the calculation of a correlation coefficient between pairs of vectors just gives an idea of the overall tendency of the statistical significance of the SNP-disease associations.

Even though the methodologies mentioned above fail to address some important issues in the analysis of shared and distinct genetic variants among a group of diseases, they have the advantage that they can be performed based solely on summary statistics (i.e., odds ratios and *p-values*) without requiring the raw data. Additionally, the computational requirements are relatively low which make the analyses easy to perform.

2.2 Materials and Methods

2.2.1 Logic Regression Gene-level Association Analysis

The proposed method involves gene-level association analysis by incorporating specific forms of SNP-SNP interactions that are biologically meaningful. Specifically, we investigated two forms of interactions: the first is *SNP-SNP intersection* which states that multiple SNPs need to have their respective high-risk genotypes in order for the disease-risk to be elevated (i.e., *SNP-A and SNP-B*). The second form of interaction is *SNP-SNP union* which states that any of the SNPs needs to have their respective high-risk genotype in order for the disease-risk to be elevated (i.e., *SNP-A or SNP-B*). To incorporate these interaction into our analysis, we used logic regression to explore the best set of SNP-SNP interactions that are associated with the phenotype of interest [10]. Logic regression is a technique used to model an outcome (e.g., phenotype) using *intersections* and *unions* of potential binary predictors, such as SNP genotypes (i.e., indicator of the minor allele homozygous)

as potential predictors. In the context of set theory, *intersection* and *union* are called Boolean operations because they act on binary variables. The logic regression model has the form shown in Eq. 2.1 where Y is the binary phenotype (i.e., disease and controls), $\beta_0, ..., \beta_p$ are the parameters, and $L_0, ..., L_p$ are Boolean combinations of SNP genotype indicators which are also called *logic trees*. Logic regression has been succesfully applied to SNP data analysis with selected candidate genes as well as GWAS to explain to a greater extent the disease genetics of highly heritable diseases [11] [12] [13].

$$logit\left(E\left[Y\right]\right) = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + ... + \beta_p L_p \tag{2.1}$$

Logic regression was performed on each gene independently 20 times varying the seed for the random number generator at the beginning of the stochastic search of logic regression. The seed variation allows us to search more broadly the solution space and diminish the probability of converging to a local optimum.

To evaluate the evidence of association, we also perform logic regression using the same genotype data but with 20 sets of permuted phenotype labels; each of the 20 is fit 20 times varying the starting random seed. This procedure allows us to perform a statistical significance test by obtaining an approximate distribution of the test statistic under the null hypothesis by comparing the likelihood of the original vs. the phenotype-permuted label models. These comparisons yield an approximate Bayes Factor (BF) for each gene. BF can be used as a measure of statistical evidence. Here we use it merely as a test statistic and for calculating a *p-value*. Specifically, the *p-value* for each gene is calculated as the proportion of all permuted BF values of all genes larger than the gene's observed BF. This calculation takes into account properly the multiple testing.

2.2.2 Shared vs. Distinct Gene Analysis using Logic Regression

The method is composed of two stages. In the first stage, we perform logic regression on each gene for each of the three diseases independently as described in Subsection 2.2.1. After determining which genes are strongly associated with each individual disease, we proceed with the next step but restricting the analysis to genes that achieved statistical significance in at least one disease. The second step involves performing several logic regression analysis by merging data from diseases and controls in order to evaluate to what degree the gene can differentiate between groups of diseases and controls.

For each of the genes that were found to be strongly associated with at least one disease, we perform additional logic regression analyses by combining any of the disease groups and the control group to create two new groups (e.g., Disease 1 & Disease 2 vs. Control) or taking any two of the groups (e.g., Disease 1 vs. Disease 2). The newly created groups could be interpreted as a fictional cluster of subjects with characteristics that averages both groups involved. The purpose of these extra analyses is to use the deviance of the logic regression model as a measure of whether the combining of the two groups was biologically appropriate. Two diseases truly sharing the association can be combined and should give a similar association as when each disease was evaluated for the association. Multiple diseases can show individual significance for the association but when combined may show no association if the biological association underlying the statistical association is not identical across the diseases.

Since the deviance is a function of the sample size and different phenotipic groups are merged, there will be an innate bias towards higher deviance values among those comparisons involving more than 2 groups (e.g., Disease 1 & Disease 2 vs. Controls). To address this issue, we perform a robust standardization of the deviances to make them comparable given by:

$$std\_dev = (dev\_ori - MEDIAN[perm\_dev])/IQR[perm\_dev]$$

where *dev_ori* is the deviance calculated with the original data and *perm_dev* is a vector of the 20 deviances calculated with the randomly permuted phenotype labels. At the end of this analysis and assuming 3 diseases under study (D1, D2, D3) with shared controls (CL), we obtain a total of 18 standardized deviances coming from all the 6 pairwise comparisons among those 4 groups (3 diseases + controls), as well as 12 comparisons of those 4 groups and selected new phenotypic groups created by merging group pairs (i.e., D1 & D2, D1 & D3, D2 & D3, D1 & CL, D2 & CL, D3 & CL) in such a way that no group appears twice in a comparison (e.g., D1 & D2 vs. D1 is not performed, but D1 & D2 vs. D3 is). A comprehensive list of the source of the 18 deviances can be found in the last code-comment section of Appendix A.1.

The standardized deviances can be interpreted as "distances" of biological similarity of the gene-disease association among the different groups. The smaller the standardized deviance is between 2 groups, the more we suspect the groups are biologically similar with respect to the gene-disease association. The distance (standardized deviance) given by the logic regression models involving merged data (e.g., Disease 1 & Disease 2 vs. Controls) can be interpreted as the distance between the midpoint of the two merged groups (e.g., Disease 1 & Disease 2) and the third one (e.g., Controls).

After calculating the distances between groups, we perform an *unconstrained nonlinear optimization* [14] to estimate the best set of coordinates on a 3-dimensional euclidean space for highly significant gene-disease associations on at least 1 disease. The best solution is the set of coordinates that minimizes the sum of the errors between the estimated distances based on the coordinates and the calculated distances based on the standardized deviances. We opted for minimizing the sum of the errors rather than the vector of errors because the later gives higher fitting priority to longer distances. The errors whose sum we want to minimize are given by:

$$\epsilon = dist([G1_x, G1_y, G1_z]\&[G2_x, G2_y, G2_z]) - dist_{empirical}$$

where $dist([G1_x, G1_y, G1_z]\&[G2_x, G2_y, G2_z])$ is a function that returns the Euclidean distance between two sets of coordinates corresponding to groups $G1$ and $G2$ and $dist_{empirical}$ is the distance (standardized deviance) from the experiments mentioned previously.

Once we get the best estimate of the set of coordinates for each group, we proceed to plot these points, as well as pairwise distances, to get a visual representation of the degree of sharedness of the gene-disease associations. If two disease groups are close to each other, we expect the association to be shared. If two disease groups are separated from each other, we expect that the underlying biological mechanism acts in a different way for each disease, even though the gene-disease association might be highly significant for both diseases.

The method was tested on the Wellcome Trust Case Control Consortium (WTCCC) dataset of Crohn's Disease (1748 subjects), Type I Diabetes (1963 subjects), and Rheumatoid Arthritis (1860 subjects) and shared controls (2936 subjects) [15]. The list of candidate genes were obtained based on the work of Sharaf Eldin *et al.* entitled *Within-Gene Interactions in GWAS Identifies Novel Susceptibility Loci - WTCCC Revisited* as well as [11] and consisted of 158 genes which showed strong evidence of association with at least one of the diseases under study. These studies also limited the number of logic trees to 2 and the number of SNPs interacting in the model to 5 due to the high computational requirement of the stochastic search for the optimum solution in a high-dimensional space of the logic regression.

2.3 Results

In this section, we present some key examples to illustrate the method and the different scenarios that were found followed by a summary of the findings by

categorizing genes based on the potential sharedness of the association among the 3 diseases.

We analyzed gene *HLA-DRA* in *Chromosome 6* which was strongly associated with each of CD, T1D, and RA. The distance plot for this gene is shown in Figure 2.1. We can observe that all disease groups are far away from each other and CD is close to the Control group. This suggests that the strength of the association differs greatly between CD and T1D & RA. We can also tell by the distances and locations that the three diseases are strongly associated with *HLA-DRA* but with a different underlying biological mechanism among them. Additionally, we can analyze the individual logic regression models to get a better understanding of the SNP interactions that might be ocurring.

Table 2.1 shows a small effect of gene *HLA-DRA* on Crohn's disease odds ratios which range from 0.63 to 1.61. On the other hand, Tables 2.2 and 2.3 show a stronger effect on the Type I Diabetes and Rheumatoid Arthritis odds ratios which range from 1.00 to 32.28 and 1.00 to 6.24, respectively. We can also note that most of the SNPs that appeared in the model for each individual disease are unique for the particular disease and the strength of the associations vary broadly.

Although it has been reported by several studies that *HLA* loci associations are shared among autoimmune diseases [5] [16] [17], these results suggest that, even though all diseases were found to be associated with gene *HLA-DRA*, each of them follows its own biological mechanism so it would be inaccurate to call this association "shared".

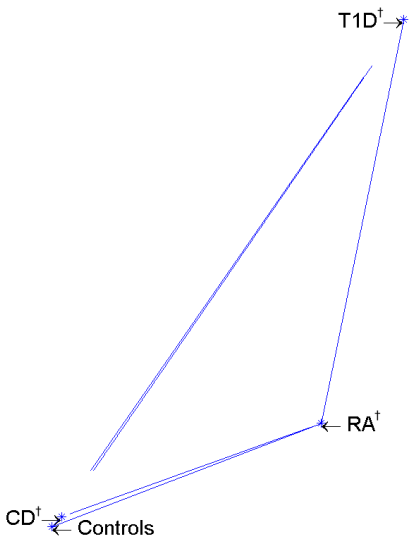

Figure 2.1: Distance plot of gene *HLA-DRA*.
† indicates statistical significance.

Table 2.1: **Logic structures, frequencies, and associated Crohn's Disease odds ratios of the HLA-DRA gene**

SNP Genotype	rs9268831 TT	rs9268877 AG or GG	rs7194 AG or GG	rs9268862 AA	rs3135393 AG or GG	Logic-based Risk Groups		
Cases	420 (24.03%)	1521 (87.01%)	1057 (60.47%)	950 (54.35%)	535 (30.61%)			
Controls	717 (24.42%)	2526 (86.04%)	1879 (64.00%)	1633 (55.62%)	976 (33.24%)			
Logic 1	((AND) OR) AND					Frequency		OR
Logic 2						Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO				Logic 2 = NO	677	862	1.61
	Logic 1 = YES				Logic 2 = NO	536	1098	1.00
	Logic 1 = NO				Logic 2 = YES	531	963	1.13
	Logic 1 = YES				Logic 2 = YES	4	13	0.63

Table 2.2: **Logic structures, frequencies, and associated Type I Diabetes odds ratios of the HLA-DRA gene**

SNP Genotype	rs9268831 TT	rs3129877 AA or AG	rs9268645 CG or GG	rs5000563 GG	rs9268877 GG	Logic-based Risk Groups		
Cases	638 (32.50%)	1273 (64.85%)	1638 (83.44%)	226 (11.51%)	1233 (62.81%)			
Controls	717 (24.42%)	1460 (49.73%)	1777 (60.52%)	263 (8.96%)	1111 (37.84%)			
Logic 1	(OR) AND (OR)					Frequency		OR
Logic 2						Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO				Logic 2 = NO	48	958	1.00
	Logic 1 = YES				Logic 2 = NO	682	867	15.71
	Logic 1 = NO				Logic 2 = YES	104	413	5.03
	Logic 1 = YES				Logic 2 = YES	1129	698	32.28

Table 2.3: **Logic structures, frequencies, and associated Rheumatoid Arthritis odds ratios of the HLA-DRA gene**

SNP Genotype	rs9268853 TT	rs9268645 CC	rs3177928 GG	rs3129877 GG	rs9268853 CT or TT	Logic-based Risk Groups		
Cases	437 (23.49%)	442 (23.76%)	1248 (67.10%)	967 (51.99%)	1341 (72.10%)			
Controls	1180 (40.19%)	1159 (39.48 %)	2120 (72.21%)	1476 (50.27%)	2544 (86.65%)			
Logic 1	(OR)					Frequency		OR
Logic 2			((OR) AND)			Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO		Logic 2 = NO			744	489	6.24
	Logic 1 = YES		Logic 2 = NO			318	555	2.35
	Logic 1 = NO		Logic 2 = YES			546	858	2.61
	Logic 1 = YES		Logic 2 = YES			252	1034	1.00

Gene *PTPN22* illustrates a different possible scenario: the gene was found to be strongly associated with RA & T1D but not CD. It can be seen from Figure 2.2 that T1D and RA are clustered together and far from the CD and Control group. This suggests that the biological mechanism of this gene is the same for RA and T1D so the association could be shared among them, but distinct from CD. We can obtain a better insight of the SNP interactions by looking at the logic trees of the individual models.

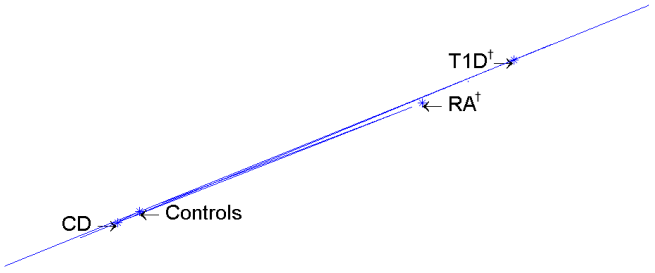

Figure 2.2: Distance plot of gene *PTPN22*.
† indicates statistical significance.

Table 2.4: **Logic structures, frequencies, and associated Crohn's Disease odds ratios of the PTPN22 gene**

SNP Genotype	rs1217414 GG	rs2488457 CC or CG	rs2488457 CC	rs1217414 AA	Logic-based Risk Groups		
Cases	922 (52.75%)	1693 (96.85%)	1150 (68.79%)	160 (9.15%)			
Controls	1573 (53.58%)	2814 (95.84%)	1869 (63.66%)	212 (7.22%)			
Logic 1	((AND) OR)				Frequency		OR
Logic 2					Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO			Logic 2 = NO	219	448	0.81
	Logic 1 = YES			Logic 2 = NO	1369	2276	1.00
	Logic 1 = NO			Logic 2 = YES	0	0	—
	Logic 1 = YES			Logic 2 = YES	160	212	1.25

Table 2.5: **Logic structures, frequencies, and associated Type I Diabetes odds ratios of the PTPN22 gene**

SNP Genotype	rs2488457 CC or CG	rs1217414 GG	rs3789609 CC	rs2488457 CC	Logic-based Risk Groups		
Cases	1818 (92.61%)	1091 (55.58%)	1023 (52.11%)	1062 (54.10%)			
Controls	2814 (95.84%)	1573 (53.58%)	1431 (48.74%)	1869 (63.66%)			
Logic 1	((AND) OR)				Frequency		OR
Logic 2					Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO			Logic 2 = NO	146	122	2.23
	Logic 1 = YES			Logic 2 = NO	755	945	1.49
	Logic 1 = NO			Logic 2 = YES	304	456	1.24
	Logic 1 = YES			Logic 2 = YES	758	1413	1.00

Table 2.6: **Logic structures, frequencies, and associated Rheumatoid Arthritis odds ratios of the PTPN22 gene**

SNP Genotype	rs2488457 CG or GG	rs3789609 TT	rs1217414 AA or AG	rs2488457 CC or CG	Logic-based Risk Groups		
Cases	844 (45.38%)	137 (7.37%)	806 (43.33%)	1739 (93.49%)			
Controls	1067 (36.34%)	255 (8.69%)	1363 (46.42%)	2814 (95.84%)			
Logic 1	(OR (AND))				Frequency		OR
Logic 2					Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO			Logic 2 = NO	0	0	—
	Logic 1 = YES			Logic 2 = NO	121	122	1.82
	Logic 1 = NO			Logic 2 = YES	1014	1869	1.00
	Logic 1 = YES			Logic 2 = YES	725	945	1.41

The logic tables for gene *PTPN22* and the diseases show that T1D and RA might share the association since the SNPs and the risk-increasing alleles involved in the models match while CD does not show a strong association with the gene. Although, given that only 4 SNPs were in the gene, high similarity of the models is expected. The difference in the distance from T1D and RA to

the Control group could be attributed to the strength of the association since *PTPN22* is showing bigger odds ratios while also being further away from the Control group compared to RA.

Based on our analyses, out of 158 genes, we have evidence supporting that 3 gene-disease associations are shared among all 3 diseases, 4 between T1D & RA, 3 between CD & RA, and 1 between CD & T1D . It can be noticed that the vast majority of the associations are distinct for a specific individual disease. These results suggest that most of the traditional and novel approaches overestimate the degree of sharedness of the genetic etiology of diseases since they are based only on significance without taking into account strength and direction of assocation.

2.4 Discussion

The methodology proposed here is based on a strict definition of "shared" association. By using this approach, we take into account not only single-disease significance, direction, and strength of the gene-level associations when making inferences, which are measures relevant to single SNPs analysis, but also incorporate the idea of how a set of SNPs of a gene are associated with the diseases. This strict definition of "shared" association was used because even if 2+ diseases are associated with the loci with the same OR direction, it does not mean they share the biological association (e.g., through the same biological pathway).

The higher accuracy of the method requires a significant amount of computational time due to the neccesity of fitting additional logic regression models on top of the high computational time required by the gene-level logic regression GWAS analysis. We reduced the computational demand of the logic regression by limiting the number of SNPs interacting in the model which makes the search not comprehensive and more complex interactions will not be discovered.

Despite the limitation imposed by fixing the number of SNPs interacting in the logic regression models, we were able to demonstrate the potentially erroneous inferences that can arise when assessing whether a genetic association is shared or distinct among a group of diseases if only limited information such as *p-values* and odds ratios are used. For most of the associations studied in this paper, the results suggest some degree of uniqueness in the gene-disease associations which indicate a potentially different biological role of the genes on the etiology of each disease.

The methodology was designed to be applied to GWASs using the same genotyping platform which might present a limitation when working with cross-platforms studies. In case the platforms differ among disease groups, we could adapt the method to allow the analysis to be done. One way to overcome this limitation is by using just the matching SNPs across platforms which could significantly reduce the number of available SNPs to work with and the results might not represent the single-disease analysis results. Another alternative could be to select proxies for the non-matching SNPs based on proximity (if any) to compensate for the mismatch. A third more elaborate approach would involve an imputation process before the single-disease analysis [18].

There has been an increasing attention paid to pathway analysis in GWAS [19]. The methodology proposed could be extended to candidate pathway-level analysis with some adaptations. The main challenge with pathway-level analysis is the computational demand and convergence since the number of SNPs involved would be high and the solution space to explore would be big increasing the chances of convergence to a local optimum if the space is not explored appropriately.

Future work involves quantifying the variance of the association strength and direction possibly by using a random-effects model. This would allow us to obtain disease-specific estimates of effect and statistically test their significance making able to quantify the sharedness of the gene-disease associations.

Bibliography

[1] Jing-Ping Lin, Joseph M Cash, Sharon Z Doyle, Sandra Peden, Keith Kanik, Chris I Amos, Sherri J Bale, and Ronald L Wilder. Familial clustering of rheumatoid arthritis with other autoimmune diseases. *Human genetics*, 103(4):475–482, 1998.

[2] Daniel C Baumgart and William J Sandborn. Crohn's disease. *The Lancet*, 2012.

[3] Diabetes Mellitus. Diagnosis and classification of diabetes mellitus. *Diabetes care*, 29:S43, 2006.

[4] Gary S Firestein. Evolving concepts of rheumatoid arthritis. *Nature*, 423(6937):356–361, 2003.

[5] Angélica Delgado-Vega, Elena Sánchez, Sara Löfgren, Casimiro Castillejo-López, and Marta E Alarcón-Riquelme. Recent findings on genetics of systemic autoimmune diseases. *Current opinion in immunology*, 22(6):698–705, 2010.

[6] Guillaume Lettre and John D Rioux. Autoimmune diseases: insights from genome-wide association studies. *Human molecular genetics*, 17(R2):R116–R121, 2008.

[7] Lauren A Zenewicz, Clara Abraham, Richard A Flavell, and Judy H Cho. Unraveling the genetics of autoimmunity. *Cell*, 140(6):791–797, 2010.

[8] D.J. Smyth, V. Plagnol, N.M. Walker, J.D. Cooper, K. Downes, J.H.M. Yang, J.M.M. Howson, H. Stevens, R. McManus, C. Wijmenga, et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. *New England Journal of Medicine*, 359(26):2767–2777, 2008.

[9] Marina Sirota, Marc A Schaub, Serafim Batzoglou, William H Robinson, and Atul J Butte. Autoimmune disease classification by inverse association with snp alleles. *PLoS genetics*, 5(12):e1000792, 2009.

[10] I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

[11] Irina Dinu, Surakameth Mahasirimongkol, Qi Liu, Hideki Yanai, Noha Sharaf Eldin, Erin Kreiter, Xuan Wu, Shahab Jabbari, Katsushi Tokunaga, and Yutaka Yasui. Snp-snp interactions discovered by logic regression explain crohn's disease genetics. *PloS one*, 7(10):e43035, 2012.

[12] Yutaka Suehiro, Chi Wai Wong, Lucian R Chirieac, Yutaka Kondo, Lanlan Shen, C Renee Webb, Yee Wai Chan, Annie SY Chan, Tsun Leung Chan, Tsung-Teh Wu, et al. Epigenetic-genetic interactions in the apc/wnt, ras/raf, and p53 pathways in colorectal carcinoma. *Clinical Cancer Research*, 14(9):2560–2569, 2008.

[13] Christina Justenhoven, Ute Hamann, Falk Schubert, Marc Zapatka, Christiane B Pierl, Sylvia Rabstein, Silvia Selinski, Tina Mueller, Katja Ickstadt, Michael Gilbert, et al. Breast cancer: a candidate gene approach across the estrogen metabolic pathway. *Breast cancer research and treatment*, 108(1):137–149, 2008.

[14] J John E Dennis and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16. Siam, 1983.

[15] P.R. Burton, D.G. Clayton, L.R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D.P. Kwiatkowski, M.I. McCarthy, W.H. Ouwehand, N.J. Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[16] Hariklia Eleftherohorinou, Victoria Wright, Clive Hoggart, Anna-Liisa Hartikainen, Marjo-Riitta Jarvelin, David Balding, Lachlan Coin, and Michael Levin. Pathway analysis of gwas provides new insights into genetic susceptibility to 3 inflammatory diseases. *PloS one*, 4(11):e8068, 2009.

[17] Oliver Brand, Stephen Gough, and Joanne Heward. Hla, ctla-4 and ptpn22: the shared genetic master-key to autoimmunity. *Expert Rev Mol Med*, 7(23):1–15, 2005.

[18] Sharon R Browning. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human genetics*, 124(5):439–450, 2008.

[19] Kai Wang, Haitao Zhang, Subra Kugathasan, Vito Annese, Jonathan P Bradfield, Richard K Russell, Patrick Sleiman, Marcin Imielinski, Joseph Glessner, Cuiping Hou, et al. Diverse genome-wide association studies associate the il12/il23 pathway with crohn disease. *The American Journal of Human Genetics*, 84(3):399–405, 2009.

Chapter 3

Novel Tuberculosis Susceptibility Genes Discovered by Logic Regression: a Stratified Analysis in Thai Population

3.1 Introduction

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* and it is one of the leading causes of death in developing countries. The infection is generally transmitted by droplets generated during abrupt respiratory movements such as coughing. Tuberculosis infection can be either active or latent, depending on the presence or absence, respectively, of symptoms, which usually involve cough, chest pain, shortness of breath, fatigue, weight loss, fever and night sweats. Latently infected people cannot transmit the disease to other people but remain at risk of becoming actively infected [1].

While environmental factors play major roles in TB infection, the role of genetics in the active infection of tuberculosis has been acknowledged by several family studies [2] [3] [4]. Further genome-wide association studies (GWAS) have discovered susceptibility loci such as single-nucleotide polymorphisms (SNPs) *rs3024505* and *rs9373180* of genes *IL10* and *IFNGR1*, respectively [5].

However, these findings are still far from explaining the heritability of tuberculosis as it is common for most GWASs of complex diseases [6].

Recently, a new GWAS uncovered a new susceptibility locus by empirically stratifying the TB case group as young (≤ 45 years old) and old patients (> 45 years old) in Thai and Japanese populations. SNP *rs6071980* of the *HSPEP1-MAFB* region was found to be significantly associated with TB with odds ratios of 1.82 and 1.81 for young Thai and Japanese populations, respectively [7]. Given the success of novel approaches to explore SNP-SNP interactions such as logic regression to explain complex disease genetics in GWASs [8], we hypothesize novel suceptibility genes could be discovered by following a similar approach on stratified TB cases.

To explore SNP-SNP interactions in our TB GWAS analysis, we propose using logic regression which incorporates SNP intersection, union and combinations of them to assess whether a group of SNPs are jointly associated with a stratified phenotype [9]. As mentioned before, logic regression has been succesfully applied to groups of SNPs but mostly based on candidate genes due to the high computational demand [10] [11] [12]. To further analyze whether the gene-disease associations are shared among the stratified groups, we propose to do a shared vs. distinct analysis as described in Chapter 2.

3.2 Materials and Methods

The method we proposed is a gene-level analysis that is performed by incorporating two specific forms of SNP-SNP interactions which are motivated by biological knowledge. One form of interaction is *SNP-SNP intersection*, which requires that multiple SNPs need to have their high-risk genotypes for the disease-risk to be elevated (i.e. *SNP-A and SNP-B*). The second form of interaction is *SNP-SNP union*, which requires the high-risk genotype of at least one of the two SNPs for the disease-risk to be elevated (i.e. *SNP-A or SNP-B*). Logic regression incorporates these two types of SNP-SNP interactions

into the models by exploring the optimum set of SNP-SNP interactions that are associated with the disease or trait of interest [9]. By using logic regression, we model the outcome (i.e. TB case or control status) as *intersections* and *unions* of binary SNP genotype indicators. The logic regression model used has the specific form shown in Eq. 3.1 where Y is the binary phenotype (i.e. TB case and controls), $\beta_0, ..., \beta_p$ are the model parameters, and $L_0, ..., L_p$ are combinations of SNP genotype indicators which are also referred to as *logic trees*. Logic regression has been succesfully applied to GWAS to explain to a greater extent the disease genetics of Crohn's Disease, a highly heritable disease [8].

$$logit\left(E\left[Y\right]\right) = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + ... + \beta_p L_p \tag{3.1}$$

The logic regression SNP-SNP interaction analysis we performed used only those SNPs in such a way that no pair of them within a gene were in linkage disequilibrium ($r^2 \geq 0.8$). The maximum number of *logic trees* we allowed in our models was two (L_1, L_2) with at most five interacting SNPs in total. These restrictions were specified in our models due to the large number of possible interactions to explore and the exponencially increasing computational cost associated with each additional tree or leaf. Since the search of the solution space is done stochastically by means of a simulated annealing algorithm, we fit the logic regression 20 times varying the initial random seed of the stochastic search. At the end of the 20 fitting processes, we keep the model with the lowest deviance among them. This process is also repeated with 20 sets of case-control randomly permuted labels.

The deviance of the best model in addition to the deviances of the best model with case-control labels randomly permuted allow a statistical significance test. A Bayes Factor (BF) can be obtained from this comparison, which can later be used to calculate a p-value by ordering all the BF values and then calculate the fraction of all permuted BF values smaller than the gene's observed BF. This p-value calculation incorporates the multiple testing cor-

rection and takes into account the potential for genes with more SNPs to overfit. Further analysis to assess whether the common associations for both age groups are shared or not was also carried out to determine if each gene is potentially affecting the same biological mechanism.

The samples for the genome-wide genotyping included 613 TB patients and 727 healthy controls; 206 cases were less than 45 years old, our empirical age threshold to identify young TB subjects based on the distribution of age at onset of TB in Thailand [13]. Statification by TB strain was also carried out with a total of 182 and 184 subjects with ancient and modern TB strains, respectively as well as stratification by TB lineage with a total of 140, 181, and 45 Beijing, East African-Indian (EAI), and "other" subjects, respectively. The patients were recruited from Chian Rai, Lampang, and Bangkok provinces of Thailand due to the high similarity of their populations [14]. All cases were human immunodeciency virus-seronegative when TB was diagnosed and later confirmed by microscopic identification or mycobacterial culture. The genotyping was performed using Illumina Hapmap 610 chip (Illumina, San Diego, CA, USA). For each gene, we excluded subjects with missing genotype values for any of the SNPs within it since the logic regression package cannot deal with missing values. Standard quality control was performed including Hardy-Weinburg equilibrium cutoff at $p\text{-}value < 10^{-5}$ and minimum allele frequency of 0.05. Multidimensional scaling of pairwise identity by state statistics was carried out using GenABEL package [15] and indicated three outlier samples, which were excluded. The genomic inflation factor (λ) was calculated from trend test p-values; at $\lambda = 1.02$ the level of population stratification was acceptable.

3.3 Results

The number of old and young TB cases is shown in Table 3.1 and Table 3.2, respectively, stratified by strain and lineage for those cases with available

relevant information. Lineage and strain information was not available for 247 cases. It can be observed that all Beijing cases and most "other" lineage cases were modern TB strain holders while all EAI cases were ancient TB strain holders.

Table 3.1: **Number of old TB cases by lineage and strain.**

	Ancient strain	Modern strain
Beijing	0	76
EAI	135	0
Other	0	24

Table 3.2: **Number of young TB cases by lineage and strain.**

	Ancient strain	Modern strain
Beijing	0	64
EAI	44	0
Other	1	20

3.3.1 Age-stratified analysis

We examined 18,278 genes. Out of these, 6 were found to have a strong assocation with young TB cases and 3 with old TB cases with an overlap of 2 genes among those. A summary of the findings can be found in Table 3.3. Genes *ZFHX1B* and *FER* showed a strong association with both age groups so we decided to investigate further whether the association is shared or distinct among them. Further analysis showed that the association with each gene is highly likely to be shared as shown in the Figure 3.1 since both age groups seem to be equally distant from control group and close to each other.

The shared gene *ZFHX1B* encodes *zinc finger E-box-binding homeobox 2* proteins and has been associated with several congenital neural disorders at different levels such as Mowat-Wilson syndrome, congenital heart disease, hypospadias, and renal tract anomalies [16]. Gene emphFER encodes *proto-oncogene tyrosine kinase* protein which participates in intracellular signalling or differentiation processes [17].

Table 3.3: **Genes with the strongest evidence of association with either young or old TB risk with chromosomal locations and approximate p-values.**

Gene	Location	Young TB p-value	Old TB p-value
ZFHX1B	2q22.3	$< 2.73 \times 10^{-6}$	$< 2.73 \times 10^{-6}$
FER	5q21.3	$< 2.73 \times 10^{-6}$	$< 2.73 \times 10^{-6}$
DAB2IP	9q33.2	3.56×10^{-5}	0.808
GNAQ	9q21.2	6.57×10^{-5}	0.364
C8orf48	8p22	7.93×10^{-5}	0.155
C11orf16	11p15.4	8.21×10^{-5}	0.616
SH3MD2	4q32.3	9.03×10^{-5}	0.195

Table 3.4: **Logic structures, frequencies, and associated old-TB odds ratios of the ZFHX1B gene**

SNP	rs2052807	rs7568133	rs7565134	rs2162571	rs7565134	Logic-based		
Genotype	AA	AA	AA	AA	AA or AG	Risk Groups		
Cases	73 (18.11%)	50 (12.41%)	119 (29.53%)	266 (66.00%)	223 (55.33%)			
Controls	114 (15.90%)	76 (10.60%)	399 (55.65%)	489 (68.20%)	522 (72.80%)			
Logic 1	(AND)					Frequency		OR
Logic 2			((OR) AND)			Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO		Logic 2 = NO			267	282	3.33
	Logic 1 = YES		Logic 2 = NO			9	1	31.61
	Logic 1 = NO		Logic 2 = YES			123	482	1.00
	Logic 1 = YES		Logic 2 = YES			4	2	7.02

Table 3.5: **Logic structures, frequencies, and associated young-TB odds ratios of the ZFHX1B gene**

SNP	rs17738837	rs12691693	rs3770305	rs6738630	rs7565134	Logic-based		
Genotype	AA	AG or GG	AG or GG	AA	AG or GG	Risk Groups		
Cases	182 (88.78%)	155 (75.61%)	195 (95.12%)	312 (19.02%)	147 (71.71%)			
Controls	609 (84.94%)	515 (71.83%)	678 (94.56%)	118 (16.46%)	318 (44.35%)			
Logic 1	((OR) AND) OR)					Frequency		OR
Logic 2						Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO				Logic 2 = NO	0	32	0.00
	Logic 1 = YES				Logic 2 = NO	58	367	1.00
	Logic 1 = NO				Logic 2 = YES	0	17	0.00
	Logic 1 = YES				Logic 2 = YES	147	301	3.09

Table 3.6: **Logic structures, frequencies, and associated old-TB odds ratios of the FER gene**

SNP Genotype	rs4957798 GG	rs9326759 AA or AG	rs4957798 AG or GG	rs9326761 AG or GG	rs17391678 AC or CC	Logic-based Risk Groups		
Cases	247 (61.44%)	74 (18.41%)	358 (89.05%)	114 (28.36%)	37 (9.20%)			
Controls	292 (40.44%)	139 (19.25%)	541 (74.93%)	260 (36.01%)	105 (14.54%)			
Logic 1	((OR) AND)					Frequency		OR
Logic 2				(OR)		Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO			Logic 2 = NO		78	220	1.00
	Logic 1 = YES			Logic 2 = NO		200	198	2.85
	Logic 1 = NO			Logic 2 = YES		16	119	0.38
	Logic 1 = YES			Logic 2 = YES		108	185	1.65

Table 3.7: **Logic structures, frequencies, and associated young-TB odds ratios of the FER gene**

SNP Genotype	rs11952637 GG	rs9326745 GG	rs4957798 AA or AG	rs12657495 AG or GG	rs6875865 GG	Logic-based Risk Groups		
Cases	113 (54.33%)	81 (38.94%)	87 (41.83%)	42 (20.19%)	86 (41.35%)			
Controls	446 (61.77%)	333 (46.12%)	430 (59.56%)	210 (29.09%)	304 (42.11%)			
Logic 1	((OR) AND)					Frequency		OR
Logic 2				(AND)		Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO			Logic 2 = NO		168	365	1.00
	Logic 1 = YES			Logic 2 = NO		32	277	0.25
	Logic 1 = NO			Logic 2 = YES		6	51	0.26
	Logic 1 = YES			Logic 2 = YES		2	29	0.15

(a) ZFHX1B (b) FER

Figure 3.1: Distance plots of genes *ZFHX1B* and *FER* for age-stratified TB subgroups.
† indicates statistical significance.

3.3.2 Strain-stratified analysis

We found 7 genes strongly associated with either modern or ancient strains of TB although only gene *ZFHX1B* achieved statistical significance in the ancient strain group. A summary of the results can be found in Table 3.8. The consistent appearance of genes *ZFHX1B* and *FER* as top genes between the age-stratified and strain-stratified analyses suggests these genes play an important role in the disease etiology and a potential common biological mechanism acting on these subgroups.

Figure 3.2 presents the results of the analysis performed to assess whether the *ZFHX1B* and *FER* associations are shared between the two subgroups based on the strain stratification. The strain-stratified analysis shows a potential common susceptibility between ancient and modern TB strains as illustrated by the distance to the control group from both disease subgroups and the relatively small separation between the two disease subgroups.

To further analyze the *ZFHX1B* and *FER* gene-disease associations, we constructed the logic tables of each individual subgroup model. The odds ratios range from 0 to 26.89 with p-value of 0.130 and from 2.97 to 18.75 with p-value of 1.09×10^{-5} associated to *FER* gene in the modern and ancient TB strains, respectively. The odds ratios range from 0.18 to 0.68 with p-value of 0.157 and from 2.77 to 8.95 with p-value of $< 3.8 \times 10^{-6}$ associated to *ZFHX1B* gene in the modern and ancient TB strains, respectively.

Table 3.8: **Genes with the strongest evidence of association with either ancient or modern TB strain risk with chromosomal locations and approximate p-values.**

Gene	Location	Modern TB p-value	Ancient TB p-value
ZFHX1B	2q22.3	0.157	$< 3.8 \times 10^{-6}$
FER	5q21.3	0.130	1.09×10^{-5}
LOC646024	6q25.1	0.168	1.37×10^{-5}
LOC387720	10q26.2	8.21×10^{-6}	0.822
FAM77C	1p35.2	2.19×10^{-5}	0.960
LOC646952	1p21.2	2.19×10^{-5}	0.265
SALF	2p16.3	6.84×10^{-5}	0.462

(a) ZFHX1B strain-stratified

(b) FER strain-stratified

Figure 3.2: Distance plots of genes *ZFHX1B* and *FER* for strain-stratified TB subgroups.
† indicates statistical significance.

Table 3.9: **Logic structures, frequencies, and associated ancient TB strain odds ratios of the ZFHX1B gene**

SNP Genotype	rs2052807 AA or AC	rs13002663 GG	rs12691693 AA	rs13413446 GG	rs7565134 AG or GG	Logic-based Risk Groups		
Cases	118 (65.56%)	62 (34.44%)	61 (33.89%)	10 (5.56%)	130 (72.22%)			
Controls	439 (61.4%)	229 (32.03%)	202 (28.25%)	30 (4.20%)	316 (44.20%)			
Logic 1	(OR) AND (OR)					Frequency		OR
Logic 2						Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO				Logic 2 = NO	34	341	1.00
	Logic 1 = YES				Logic 2 = NO	16	58	2.77
	Logic 1 = NO				Logic 2 = YES	97	279	3.49
	Logic 1 = YES				Logic 2 = YES	33	37	8.95

Table 3.10: **Logic structures, frequencies, and associated modern TB strain odds ratios of the ZFHX1B gene**

SNP Genotype	rs13413446 AG or GG	rs4662223 AA or AG	rs7599224 AA or AC	rs7565134 AG or GG	rs1365778 AG or GG	Logic-based Risk Groups		
Cases	73 (40.33%)	66 (36.46%)	145 (80.11%)	110 (60.77%)	130 (71.82%)			
Controls	274 (38.82%)	260 (36.36%)	534 (74.69%)	316 (44.20%)	462 (64.42%)			
Logic 1	((OR) AND) OR					Frequency		OR
Logic 2						Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO				Logic 2 = NO	5	73	0.18
	Logic 1 = YES				Logic 2 = NO	46	180	0.68
	Logic 1 = NO				Logic 2 = YES	18	162	0.30
	Logic 1 = YES				Logic 2 = YES	112	300	1.00

Table 3.11: **Logic structures, frequencies, and associated ancient TB strain odds ratios of the FER gene**

SNP Genotype	rs4616948 GG	rs17161562 AA	rs17473831 AC or CC	rs4957798 GG	rs9326761 AA	Logic-based Risk Groups		
Cases	147 (81.22%)	147 (81.22%)	178 (98.34%)	110 (60.77%)	126 (69.61%)			
Controls	602 (83.61%)	612 (85.00%)	720 (100.00%)	291 (40.42%)	461 (64.03%)			
Logic 1	((OR) AND)					Frequency		OR
Logic 2				(AND)		Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO			Logic 2 = NO		11	4	18.75
	Logic 1 = YES			Logic 2 = NO		72	491	1.00
	Logic 1 = NO			Logic 2 = YES		1	0	Inf
	Logic 1 = YES			Logic 2 = YES		97	225	2.94

Table 3.12: **Logic structures, frequencies, and associated modern TB strain odds ratios of the FER gene**

SNP Genotype	rs10477929 AA	rs12657495 AG or GG	rs9326758 AA	rs3797838 AA or AG	rs9326758 AA	Logic-based Risk Groups		
Cases	115 (62.84%)	44 (24.04%)	7 (3.83%)	60 (32.79%)	7 (3.83%)			
Controls	491 (68.19%)	210 (29.17%)	10 (1.39%)	224 (31.11%)	10 (1.39%)			
Logic 1	AND ((OR) AND)					Frequency		OR
Logic 2						Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO				Logic 2 = NO	176	676	1.00
	Logic 1 = YES				Logic 2 = NO	0	34	0.00
	Logic 1 = NO				Logic 2 = YES	7	1	26.89
	Logic 1 = YES				Logic 2 = YES	0	9	0.00

3.3.3 Lineage-stratified analysis

We found 5 genes strongly associated with either Beijing, EAI or *other* lineages of TB. Two genes achieved statistical significance in the EAI lineage subgroup, one gene in the Beijing lineage subgroup, and no gene in the *other* lineage subgroup. A summary of the results can be found in Table 3.13. The consistent appearance of genes *ZFHX1B* and *FER* as top genes among

the age-stratified, strain-stratified, and lineage-stratified analyses keeps adding evidence that these genes play an important role in the disease etiology and a potential common biological mechanism acting on these subgroups.

Figure 3.3 presents the results of the analysis performed to assess whether the *ZFHX1B* and *FER* associations are shared between the three subgroups based on the lineage stratification. The lineage-stratified analysis shows a potential common susceptibility between Beijing and *other* TB lineages but different than the susceptibility to EAI lineage as illustrated by the close distance between Beijing and *other* subgroups and the significant separation between these two subgroups and the EAI lineage.

To further analyze the *ZFHX1B* and *FER* gene-disease associations, we constructed the logic tables of each individual subgroup model. The odds ratios range from 0 to 4.81 with p-value of $< 3.8 \times 10^{-6}$, from 2.97 to 17.05 with p-value of 0.127, and from 0.06 to 2.15 with p-value of 0.171 associated to *FER* gene in the EAI, Beijing, and *other* TB lineages, respectively. The odds ratios range from 0.21 to 1.68 with p-value of $< 3.8 \times 10^{-6}$, from 0.63 to 3.81 with p-value of 0.058, and from 0.00 to 15.79 with p-value of 0.165 associated to *ZFHX1B* gene in the EAI, Beijing, and *other* TB lineages, respectively.

Table 3.13: **Genes with the strongest evidence of association with TB risk in Beijing, EAI, and *other* lineages with chromosomal locations and approximate p-values.**

Gene	Location	Beijing lineage TB p-value	EAI lineage TB p-value	*Other* lineage TB p-value
ZFHX1B	2q22.3	0.058	$< 3.8 \times 10^{-6}$	0.165
FER	5q21.3	0.127	$< 3.8 \times 10^{-6}$	0.171
LOC646024	6q25.1	0.141	5.47×10^{-6}	0.327
FAM77C	1p35.2	$< 3.8 \times 10^{-6}$	0.971	0.090
RGS6	14q24.2	0.428	0.976	6.29×10^{-5}

(a) ZFHX1B lineage-stratified (b) FER lineage-stratified

Figure 3.3: Distance plots of genes *ZFHX1B* and *FER* for lineage-stratified TB subgroups.
† indicates statistical significance.

Table 3.14: **Logic structures, frequencies, and associated EAI-lineage TB odds ratios of the ZFHX1B gene**

SNP	rs7565134	rs1365778	rs10185359	rs7599224	rs12691693	Logic-based		
Genotype	AA	AA or AG	AA or AG	AA or AC	AG or GG	Risk Groups		
Cases	50 (27.93%)	143 (79.89%)	75 (41.90%)	128 (71.51%)	118 (65.92%)			
Controls	399 (55.80%)	606 (84.76%)	293 (40.98%)	534 (74.69%)	513 (71.75%)			
Logic 1	(AND (OR))					Frequency		OR
Logic 2				(AND)		Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO			Logic 2 = NO		69	126	1.68
	Logic 1 = YES			Logic 2 = NO		25	180	0.43
	Logic 1 = NO			Logic 2 = YES		72	221	1.00
	Logic 1 = YES			Logic 2 = YES		13	188	0.21

Table 3.15: **Logic structures, frequencies, and associated Beijing-lineage TB odds ratios of the ZFHX1B gene**

SNP	rs7600752	rs7565134	rs7568133	rs10185359	rs10196335	Logic-based		
Genotype	AA or AG	GG	GG	AA or AG	AA	Risk Groups		
Cases	115 (83.94%)	59 (43.07%)	68 (49.64%)	56 (40.88%)	15 (10.95%)			
Controls	549 (76.78%)	193 (26.99%)	329 (46.01%)	293 (40.98%)	115 (16.06%)			
Logic 1	((OR) AND) OR					Frequency		OR
Logic 2						Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO				Logic 2 = NO	79	525	1.00
	Logic 1 = YES				Logic 2 = NO	43	75	3.81
	Logic 1 = NO				Logic 2 = YES	8	84	0.63
	Logic 1 = YES				Logic 2 = YES	7	31	1.50

Table 3.16: **Logic structures, frequencies, and associated *other*-lineage TB odds ratios of the ZFHX1B gene**

SNP Genotype	rs13413446 AG or GG	rs1035822 AA or AG	rs3928425 AA	rs7599224 CC	rs7565134 AA	Logic-based Risk Groups		
Cases	23 (51.11%)	20 (22.22%)	31 (68.89%)	6 (13.33%)	18 (40.00%)			
Controls	274 (38.32%)	73 (10.21%)	405 (56.64%)	181 (25.31%)	399 (55.80%)			
Logic 1	((AND) AND)					Frequency		OR
Logic 2				(AND)		Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO			Logic 2 = NO		38	600	1.00
	Logic 1 = YES			Logic 2 = NO		7	7	15.79
	Logic 1 = NO			Logic 2 = YES		0	103	0.00
	Logic 1 = YES			Logic 2 = YES		0	5	0.00

Table 3.17: **Logic structures, frequencies, and associated EAI-lineage TB odds ratios of the FER gene**

SNP Genotype	rs4616948 GG	rs17161562 AA	rs17473831 AC or CC	rs4957798 GG	rs9326761 AA	Logic-based Risk Groups		
Cases	147 (81.67%)	146 (81.11%)	177 (98.33%)	110 (61.11%)	126 (70.00%)			
Controls	602 (83.61%)	612 (85.00%)	720 (100.00%)	291 (40.42%)	461 (64.03%)			
Logic 1	((AND) OR)					Frequency		OR
Logic 2				(OR)		Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO			Logic 2 = NO		10	4	17.05
	Logic 1 = YES			Logic 2 = NO		72	491	1.00
	Logic 1 = NO			Logic 2 = YES		1	0	Inf
	Logic 1 = YES			Logic 2 = YES		97	225	2.94

Table 3.18: **Logic structures, frequencies, and associated Beijing-lineage TB odds ratios of the FER gene**

SNP Genotype	rs9326745 GG	rs4365877 GG	rs10477929 GG	rs4957798 AA or AG	rs9326758 AG or GG	Logic-based Risk Groups		
Cases	56 (40.29%)	37 (26.62%)	4 (2.88%)	67 (48.2%)	132 (94.96%)			
Controls	333 (46.25%)	243 (33.75%)	13 (1.81%)	429 (59.58%)	710 (98.61%)			
Logic 1	((OR) AND) OR					Frequency		OR
Logic 2						Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO				Logic 2 = NO	7	7	4.81
	Logic 1 = YES				Logic 2 = NO	0	3	0.00
	Logic 1 = NO				Logic 2 = YES	132	635	1.00
	Logic 1 = YES				Logic 2 = YES	0	75	0.00

Table 3.19: **Logic structures, frequencies, and associated *other*-lineage TB odds ratios of the FER gene**

SNP Genotype	rs7710223 AA or AG	rs4957798 AG or GG	rs7710223 GG	rs7737443 AA or AC	rs7715208 AA	Logic-based Risk Groups		
Cases	32 (71.11%)	41 (91.11%)	13 (28.89%)	34 (75.56%)	38 (84.44%)			
Controls	554 (76.94%)	539 (74.86%)	166 (23.06%)	554 (76.94%)	572 (79.44%)			
Logic 1	((OR) OR)					Frequency		OR
Logic 2				(OR)		Cases	Controls	
Logic-based Risk Groups	Logic 1 = NO			Logic 2 = NO		1	229	0.06
	Logic 1 = YES			Logic 2 = NO		31	411	1.00
	Logic 1 = NO			Logic 2 = YES		13	80	2.15
	Logic 1 = YES			Logic 2 = YES		0	0	—

3.4 Discussion

The results of these gene-level analyses illustrate the power of logic-regression to uncover multiple-SNP interactions that could potentially explain the genetics of complex traits. We found strong evidence of association of 13 newly identified genes with different strains, lineages, and age groups of TB that traditional single-SNP analysis have not been able to uncover, explaining to a greater extent the genetics of TB.

Genes *FER* and *ZFHX1B* were found to be consistently associated with different age groups as well as specific strains and lineages. Gene *FER* has been found to encode a member of the FPS/FES protein-tyrosine kinase family. It is involved in the regulation of cell-cell adhesion as well as the mediation of signaling from the cell surface to the cytoskeleton [18]. Gene *ZFHX1B* encodes protein zinc finger E-box-binding homeobox 2 and mutations of this gene has been associated with Mowat-Wilson syndrome which is characterized by a number of defects such as microcephaly, mental retardation, and epilepsy, among others [19].

For our analyses, we reduced the computational intensity by limiting the case-control label permutations to 20, number of SNPs interacting to a maximum of 5, and number of logic trees to 2. These limitations make the search not comprehensive as there might be higher-order interactions that could explain the genetic TB-risk. Nevertheless, the structure permits an approximate of more complex interaction structures, far closer to them than the single-SNP analysis could approximate.

False positive results are a common concern in GWASs due to the large number of tests performed. Findings should be further validated in order to rule-out spurious associations due to population stratification or genotyping errors [20]. An alternative to reduce the computational demand of logic regression applied to data-driven approaches such as GWAS is to use a candidate-genes analysis by just analyzing the highly significant genes in the primary

GWAS. Adequate phenotyping is also a major concern in this type of studies to be able to discover new genetic associations. Special attention should be paid to GWAS of infectious diseases due to the possibility of controls becoming cases later on as well as the definition used as the findings can be highly sensitive to these factors [21].

Increasing attention has been paid to pathway analysis in GWAS [22]. This approach is biologically appealing because it would incorporate gene-interactions that are not currently being captured by our approach. Logic regression has the potential to be extended to pathway-level analysis but the computational requirements would be too demanding given the dimensionality of the solution space and the necessity of minimizing the possibility of converging to a local optimum. There would also be a need to incorporate biological knowledge to the pathway analysis to understand how genes interact within the pathway in order to model the interactions accurately [23].

Bibliography

[1] Anastasios Konstantinos et al. Testing for tuberculosis. *Australian Prescriber*, 33(1):12–18, 2010.

[2] Richard Bellamy, Nulda Beyers, Keith PWJ McAdam, Cyril Ruwende, Robert Gie, Priscilla Samaai, Danite Bester, Mandy Meyer, Tumani Corrah, Matthew Collin, et al. Genetic susceptibility to tuberculosis in africans: a genome-wide scan. *Proceedings of the National Academy of Sciences*, 97(14):8005–8009, 2000.

[3] Annette Jepson, Amanda Fowler, Winston Banya, Mahavir Singh, Steve Bennett, Hilton Whittle, and Adrian VS Hill. Genetic regulation of acquired immune responses to antigens ofmycobacterium tuberculosis: a study of twins in west africa. *Infection and immunity*, 69(6):3989–3994, 2001.

[4] Jamila El Baghdadi, Marianna Orlova, Andrea Alter, Brigitte Ranque, Mohamed Chentoufi, Faouzia Lazrak, Moulay Idriss Archane, Jean-Laurent Casanova, Abdellah Benslimane, Erwin Schurr, et al. An autosomal dominant major gene confers predisposition to pulmonary tuberculosis in adults. *The Journal of experimental medicine*, 203(7):1679–1684, 2006.

[5] Thorsten Thye, Fredrik O Vannberg, Sunny H Wong, Ellis Owusu-Dabo, Ivy Osei, John Gyapong, Giorgio Sirugo, Fatou Sisay-Joof, Anthony Enimil, Margaret A Chinbuah, et al. Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11. 2. *Nature genetics*, 42(9):739–741, 2010.

[6] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[7] Surakameth Mahasirimongkol, Hideki Yanai, Taisei Mushiroda, Watoo Promphittayarat, Sukanya Wattanapokayakit, Jurairat Phromjai, Rika Yuliwulandari, Nuanjun Wichukchinda, Amara Yowang, Norio Yamada, et al. Genome-wide association studies of tuberculosis in asians identify distinct at-risk locus for young tuberculosis. *Journal of human genetics*, 57(6):363–367, 2012.

[8] Irina Dinu, Surakameth Mahasirimongkol, Qi Liu, Hideki Yanai, Noha Sharaf Eldin, Erin Kreiter, Xuan Wu, Shahab Jabbari, Katsushi Tokunaga, and Yutaka Yasui. Snp-snp interactions discovered by logic regression explain crohn's disease genetics. *PloS one*, 7(10):e43035, 2012.

[9] I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

[10] Charles Kooperberg and Ingo Ruczinski. Identifying interacting snps using monte carlo logic regression. *Genetic epidemiology*, 28(2):157–170, 2005.

[11] Yutaka Suehiro, Chi Wai Wong, Lucian R Chirieac, Yutaka Kondo, Lanlan Shen, C Renee Webb, Yee Wai Chan, Annie SY Chan, Tsun Leung Chan, Tsung-Teh Wu, et al. Epigenetic-genetic interactions in the apc/wnt, ras/raf, and p53 pathways in colorectal carcinoma. *Clinical Cancer Research*, 14(9):2560–2569, 2008.

[12] Ingo Ruczinski, Charles Kooperberg, and Michael L LeBlanc. Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications. *Journal of Multivariate Analysis*, 90(1):178–195, 2004.

[13] Potjaman Siriarayapon, Hideki Yanai, Judith R Glynn, Somboonsak Yanpaisarn, and Wat Uthaivoravit. The evolving epidemiology of hiv infection and tuberculosis in northern thailand. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 31(1):80–89, 2002.

[14] Surakameth Mahasirimongkol, Wasun Chantratita, Somying Promso, Ekawat Pasomsab, Natini Jinawath, Wallaya Jongjaroenprasert, Viraphong Lulitanond, Phanida Krittayapoositpot, Sissades Tongsima, Pathom Sawanpanyalert, et al. Similarity of the allele frequency and linkage disequilibrium pattern of single nucleotide polymorphisms in drug-related gene loci between thai and northern east asian populations: implications for tagging snp selection in thais. *Journal of human genetics*, 51(10):896–904, 2006.

[15] Yurii S Aulchenko, Stephan Ripke, Aaron Isaacs, and Cornelia M van Duijn. Genabel: an r library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296, 2007.

[16] Meredith Wilson, David Mowat, Dastot-Le Moal, Valère Cacheux, Helena Kääriäinen, Danny Cass, Dian Donnai, Jill Clayton-Smith, Sharron Townshend, Cynthia Curry, et al. Further delineation of the phenotype associated with heterozygous mutations in zfhx1b. *American Journal of Medical Genetics Part A*, 119(3):257–265, 2003.

[17] C Morris, N Heisterkamp, QL Hao, JR Testa, and J Groffen. The human tyrosine kinase gene (fer) maps to chromosome 5 and is deleted in myeloid leukemias with a del (5q). *Cytogenetic and Genome Research*, 53(4):196–200, 2008.

[18] T Pawson, K Letwin, T Lee, QL Hao, N Heisterkamp, and J Groffen. The fer gene is evolutionarily conserved and encodes a widely expressed member of the fps/fes protein-tyrosine kinase family. *Molecular and cellular biology*, 9(12):5722–5725, 1989.

[19] Wolfram Heinritz, Christiane Zweier, Ursula G Froster, Sibylle Strenge, Annegret Kujat, Steffen Syrbe, Anita Rauch, and Volker Schuster. A missense mutation in the zfhx1b gene associated with an atypical mowat–wilson syndrome phenotype. *American Journal of Medical Genetics Part A*, 140(11):1223–1227, 2006.

[20] Huixiao Hong, Leming Shi, James C Fuscoe, Federico Goodsaid, Donna Mendrick, and Weida Tong. Potential sources of spurious associations and batch effects in genome-wide association studies. *Batch effects and noise in microarray experiments: Sources and solutions*, pages 191–201, 2009.

[21] Marlo Möller, Erika De Wit, and Eileen G Hoal. Past, present and future directions in human genetic susceptibility to tuberculosis. *FEMS Immunology & Medical Microbiology*, 58(1):3–26, 2010.

[22] Tom HM Ottenhoff. New pathways of protective and pathological host defense to mycobacteria. *Trends in microbiology*, 2012.

[23] Angus J Clarke and David N Cooper. Gwas: heritability missing in action&quest. *European Journal of Human Genetics*, 18(8):859–861, 2010.

Chapter 4

Conclusions

4.1 Review of Hypotheses

Chapter 2

- P-values alone do not provide enough information to assess whether a genetic association is shared among a group of diseases and can lead to inaccurate inferences. – **We found evidence that inferences based solely on p-values tend to overestimate the degree of sharedness of genetic associations among a group of diseases since the association might be in opposite direction, in different magnitude, and/or biologically different.**
- Taking into account strength, direction, and similarity of the biological association provide better insight and a stricter definition of the sharedness of genetic associations among a group of diseases. – **We found evidence that by taking into account strength, direction, and similarity of the biological association we were able to detect some disimilarities among commonly reported "shared" genetic susceptibilities. Furthermore, we found evidence that some shared associations might exist among groups of diseases that the traditional approach**

wouldn't have been able to uncover if the association is not found statistically significant in one of the diseases.

Chapter 3

- SNP-SNP interactions are responsible for a proportion of the TB susceptibility and explain to a greater extent the TB genetics. – **We found evidence of 3 genes statistically significantly associated with different TB subgroups that, to our knowledge, haven't been reported before. Additionally, we found evidence of an additional 11 genes which were strongly associated with different TB subgroups but did not achieve statistical significance. These findings contribute to the understanding of TB genetic susceptibility.**
- Some of the genetic susceptibilities are shared while others are unique for certain TB subgroups. – **We found evidence that the two most statistically significant gene-disease associations, *ZFHX1B* and *FER*, were shared between > 45 and ≤ 45 TB age subgroups as well as ancient and modern TB strains while these associations seem to be different among TB lineage subgroups.**

4.2 Discussion

As mentioned in Chapter 2, our methodology proposed is based on a strict definition of "shared" association since we take into account not only direction and strength of the gene-disease associations, but also incorporate the structure of how a set of SNPs of a gene are associated with the diseases. This strict definition of "shared" association was used because even if a gene is statistically significantly associated with 2+ diseases, with the same odds

ratio magnitude and direction, it does not mean they share the same biological association.

The stricter definition of a "shared" association used in this thesis uncovered that the traditional approach which focuses mainly on p-values overestimate the degree of sharedness. Our analysis showed that only 11 out of 158 genes that achieved statistical significance in at least one out of the three (CD, RA, T1D) GWASs could be stated as "shared" associations between 2 or more diseases. All the other genes showed hints of a certain degree of uniqueness in the association either because of the strength, direction or the SNP-SNP interactions within the gene.

The methodology has certain limitations when compared to other approaches such as the need of using the same genotyping platform and access to the raw data. The simplest way to perform the analysis in this scenario is to only use matching SNPs across the different platforms, but this might substancially reduce the number of available SNPs in the analysis. Another way to deal with different genotyping platforms is to perform an imputation process although it is computationally intensive procedure and requires specific knowledge of how to perform it.

In Chapter 3 we found that genes *FER* and *ZFHX1B* were consistently associated with different age groups as well as specific strains and lineages. To our knowledge, these genes have not been reported to be associated with TB before. The role of gene *FER* has been found to be encoding a member of the FPS/FES protein-tyrosine kinase family and it is involved in the regulation of cell-cell adhesion as well as the mediation of signaling from the cell surface to the cytoskeleton [1]. Gene*ZFHX1B* encodes protein zinc finger E-box-binding homeobox 2 and mutations of this gene has been associated with Mowat-Wilson syndrome which is characterized by a number of defects such as microcephaly, mental retardation, and epilepsy, among others [2].

The discovery of these two novel gene susceptibilities for TB keeps adding evidence of the power of logic regression to uncover new gene-disease associ-

ations and help explain to a greater extent the genetics of complex diseases that traditional SNP-level GWAS cannot achieve. For our analyses, we reduced the computational demand by permuting the case-control label only 20 times, fixing the number of SNPs interacting to a maximum of 5 as well as the number of logic trees to 2. These restrictions make the random search not comprehensive since there might be more interactions that could better explain the genetic TB-risk. Still, the model specifications used can be used as an approximate of more complex structures.

Regarding the findings in Chapter 3, false positive results are a common concern in GWASs due to the large number of tests performed. Since the analysis is not hypothesis-driven, these findings should be validated in order to rule-out the possibility of spurious associations due to population stratification or genotyping errors. Additionally, GWAS of infectious diseases are prone to the possibility of controls getting infected and become cases later in time so the findings can be biased because of these factors as well as the protocol used to identify cases [3].

4.3 Future Work

Further research aims to extend the shared gene analysis to other groups of diseases which are more challenging and time consuming due to the different platforms used for genotyping the subjects. Additionally, we need to perform a replication study to confirm our TB findings in order to rule out the possibility of spurious association due to population stratification and other factors.

Specifically, some of the research goals that arose from the work presented in this thesis are:

- To enable to quantify the degree of sharedness of the gene-disease associations and perform the appropriate statistical hypothesis tests
- Develop methods for handling multiple diseases GWASs with different genotyping platforms

- Replicate the TB study using new independent GWAS dataset

4.4 Conclusions

In conclusion, traditional analysis performed to assess whether genetic associations are shared or distinct among a group of diseases tend to overestimate the degree of sharedness because they rely mostly on p-values and statistical significance. Our approach considers a stricter definition of a "shared association" since it takes into account not only strength and direction of the association, but also a complete biological similarity. We were able to uncover some interesting patterns in the gene-disease associations the traditional and recent novel approaches are not able to do. This provides a better insight of the biological mechanism acting on each disease.

Additionally, we were able to demonstrate the power of logic regression to uncover new genetic susceptibilities and explain to a greater extent the genetics of complex diseases. We found 3 newly identified genes that were statistically significant associated with different subgroups of TB: *ZFHX1B*, *FER*, and *FAM77*. We additionally identified 11 genes which were strongly associated with different subgroups of TB but did not achieve statistical significance at the GWAS level.

Bibliography

[1] T Pawson, K Letwin, T Lee, QL Hao, N Heisterkamp, and J Groffen. The fer gene is evolutionarily conserved and encodes a widely expressed member of the fps/fes protein-tyrosine kinase family. *Molecular and cellular biology*, 9(12):5722–5725, 1989.

[2] Wolfram Heinritz, Christiane Zweier, Ursula G Froster, Sibylle Strenge, Annegret Kujat, Steffen Syrbe, Anita Rauch, and Volker Schuster. A missense mutation in the zfhx1b gene associated with an atypical mowat–wilson syndrome phenotype. *American Journal of Medical Genetics Part A*, 140(11):1223–1227, 2006.

[3] Marlo Möller, Erika De Wit, and Eileen G Hoal. Past, present and future directions in human genetic susceptibility to tuberculosis. *FEMS Immunology & Medical Microbiology*, 58(1):3–26, 2010.

Appendix A

R codes

A.1 Logic Regression – WTCCC

```
args=commandArgs(trailingOnly = TRUE);
ChrNum=as.numeric(args[1]);
GeneName=as.character(args[2]);

#Load the library
library('LogicReg')

#DATA LOADING AND GENE SELECTION#

#Load the genotype file of all 3 diseases
infile=paste('A01Chr',ChrNum,'.txt',sep='')
gene=read.table(infile,sep='\t',header=TRUE)
gene=gene[as.character(gene[,4])==GeneName,]

#Total phenotype vector matching the genotype file
respTotal=rbind(as.matrix(rep('Control',2936)),
   as.matrix(rep('CD',1748),as.matrix(rep('T1D',1963)),
   as.matrix(rep('RA',1860)))

#LOGIC REGRESSION SETUP#

#Logic regression parameters
nsnp=length(unique(geneCD[,1]))
leaf=min(nsnp,5)
trees=2

#Deviances calculation function of the
#phenotype vector and genotype matrix
devianceCalc = function(resp, gene, trees, leaf){
   set.seed(1)
   dev=rep(1000000,21)

   #Deviances for the original response model
```

```
   for (i in 1:20){
      myfit = logreg(resp=resp,bin=t(gene[,-c(1:4)]),
      type=3,select=1,ntrees=trees,nleaves=leaf)
      if (myfit$model$score<dev[1]){
         dev[1]=myfit$model$score
      }
   }

   #Deviances for the permuted response models
   for (i in 1:20){
      respPerm=sample(resp,length(resp),replace=FALSE)
      for (j in 1:20){
         myfit = logreg(resp=respPerm,bin=t(gene[,-c(1:4)]),
         type=3,select=1,ntrees=trees,nleaves=leaf)
         if (myfit$model$score<dev[i+1]){
            dev[i+1]=myfit$model$score
         }
      }
   }

   return(dev)
}

#THE CODE CALLS THE FUNCTION ABOVE 18 TIMES WITH
#DIFFERENT DATA SUBSETS AND PHENOTYPE VECTORS:
#dev1.  CD & T1D vs.  Controls
#dev2..  CD & RA vs.  Controls
#dev3.  T1D & RA vs.  Controls
#dev4.......  CD vs.  T1D & Controls
#dev5.......  CD vs.  RA & Controls
#dev6......  T1D vs.  CD & Controls
#dev7......  T1D vs.  RA & Controls
#dev8.......  RA vs.  CD & Controls
#dev9.......  RA vs.  T1D & Controls
#dev10......  CD vs.  T1D & RA
#dev11.....  T1D vs.  CD & RA
#dev12......  RA vs.  CD & T1D
#dev13......  CD vs.  T1D
#dev14......  CD vs.  RA
#dev15.....  T1D vs.  RA
#dev16......  CD vs.  Controls
#dev17.....  T1D vs.  Controls
#dev18......  RA vs.  Controls

deviance=cbind(dev1,dev2,dev3,dev4,dev5,dev6,dev7,dev8,
   dev9,dev10,dev11,dev12,dev13,dev14,dev15,dev16,dev17,dev18)

outfile=paste('dev_',GeneName,'.txt',sep='')

write.table(deviance,outfile,sep='\t',
   col.names=FALSE, row.names=FALSE, quote=FALSE)
```

A.2 Logic Regression – TB

```
args=commandArgs(trailingOnly = TRUE);
ChrNum=as.numeric(args[1]);
GeneName=as.character(args[2]);

#Load the library
library('LogicReg')

#DATA LOADING AND GENE SELECTION#

#Load the genotype file of all TB cases and controls
infile=paste('A01Chr',ChrNum,'.txt',sep='')
gene=read.table(infile,sep='\t',header=TRUE)
gene=gene[as.character(gene[,4])==GeneName,]

#Load the phenotype file of all TB cases and controls
pheno=read.table('pheno extended.txt',sep='\t',
   header=TRUE, colClasses = 'character')

#Eliminate samples with missing values
keep=which(colSums(gene=='\00')==0)
gene=gene[,keep]
pheno=pheno[keep[5:length(keep)]-4,]

#Logic regression parameters
leaf=min(5,dim(gene)[1]/2)
bin=as.matrix(t(gene[,5:dim(gene)[2]]))
class(bin)='numeric'

#Deviances calculation function of the
#phenotype vector and genotype matrix
devianceCalc = function(resp, gene, trees, leaf){
   set.seed(1)
   dev=rep(1000000,21)

   #Deviances for the original response model
   for (i in 1:20){
      myfit = logreg(resp=resp,bin=t(gene[,-c(1:4)]),
      type=3,select=1,ntrees=trees,nleaves=leaf)
      if (myfit$model$score<dev[1]){
         dev[1]=myfit$model$score
      }
   }

   #Deviances for the permuted response models
   for (i in 1:20){
      respPerm=sample(resp,length(resp),replace=FALSE)
      for (j in 1:20){
         myfit = logreg(resp=respPerm,bin=t(gene[,-c(1:4)]),
         type=3,select=1,ntrees=trees,nleaves=leaf)
```

```
            if (myfit$model$score<dev[i+1]){
               dev[i+1]=myfit$model$score
            }
         }
      }

      return(dev)
}

#THE CODE CALLS THE FUNCTION ABOVE 6+6+18 TIMES WITH
#DIFFERENT TB DATA SUBSETS AND PHENOTYPE VECTORS

#For the age-stratified analysis:
#dev1.  ≤ 45 & >45 vs.  Controls
#dev2.......  ≤ 45 vs.  >45 & Controls
#dev3........  >45 vs.  ≤ 45 & Controls
#dev4.......  ≤ 45 vs.  >45
#dev5.......  ≤ 45 vs.  Controls
#dev6........  >45 vs.  Controls

dev_age=cbind(dev_age1,dev_age2,dev_age3,dev_age4,dev_age5,dev_age6)

outfile=paste('devAge_',GeneName,'.txt',sep='')

write.table(dev_age,outfile,sep='\t',
   col.names=FALSE, row.names=FALSE, quote=FALSE)

#For the strain-stratified analysis:
#dev1.  modern & ancient vs.  Controls
#dev2...........  modern vs.  ancient & Controls
#dev3..........  ancient vs.  modern & Controls
#dev4...........  modern vs.  ancient
#dev5...........  modern vs.  Controls
#dev6..........  ancient vs.  Controls

dev_strain=cbind(dev_strain1,dev_strain2,dev_strain3,
   dev_strain4,dev_strain5,dev_strain6)

outfile=paste('devStrain_',GeneName,'.txt',sep='')

write.table(dev_strain,outfile,sep='\t',
   col.names=FALSE, row.names=FALSE, quote=FALSE)

#For the lineage-stratified analysis:
#dev1....  Beijing & EAI vs.  Controls
#dev2..  Beijing & Other vs.  Controls
#dev3......  EAI & Other vs.  Controls
#dev4..........  Beijing vs.  EAI & Controls
#dev5..........  Beijing vs.  Other & Controls
#dev6..............  EAI vs.  Beijing & Controls
#dev7..............  EAI vs.  Other & Controls
```

```
#dev8.............  Other vs.  Beijing & Controls
#dev9.............  Other vs.  EAI & Controls
#dev10.........  Beijing vs.  EAI & Other
#dev11..............  EAI vs.  Beijing & Other
#dev12............  Other vs.  Beijing & EAI
#dev13.........  Beijing vs.  EAI
#dev14.........  Beijing vs.  Other
#dev15..............  EAI vs.  Other
#dev16.........  Beijing vs.  Controls
#dev17..............  EAI vs.  Controls
#dev18............  Other vs.  Controls

dev_lin=cbind(dev_lin1,dev_lin2,dev_lin3,dev_lin4,dev_lin5,
   dev_lin6,dev_lin7,dev_lin8,dev_lin9,dev_lin10,dev_lin11,
   dev_lin12,dev_lin13,dev_lin14,dev_lin15,dev_lin16,
   dev_lin17,dev_lin18)

outfile=paste('devLineage_',GeneName,'.txt',sep='')

write.table(dev_lin,outfile,sep='\t',
   col.names=FALSE, row.names=FALSE, quote=FALSE)
```

Appendix B

MATLAB codes

B.1 2D distance error calculation

```
function f = distance2D(x)
%Function receives as argument a vector size 3
%corresponding to the 3 non-fixed coordinates
%3 coordinates have to be fixed (to 0) to converge

%Declaration of dist as a global variable
%containing the standardized deviances
global dist
x(1) = x(1); %x-coordinate Disease #1
x(2) = x(2); %y-coordinate Disease #1
x(3) = x(3); %x-coordinate Disease #2
x(4) = 0; %y-coordinate Disease #2
x(5) = 0; %x-coordinate Controls
x(6) = 0; %y-coordinate Controls

%Disease #1 & Disease #2 vs Controls
f(1) = abs(sqrt(((x(1)+x(3))/2-x(5))^2 + ...
   ((x(2)+x(4))/2-x(6))^2) - dist(1));
%Disease #1 vs Disease #2 & Controls
f(2) = abs(sqrt(((x(3)+x(5))/2-x(1))^2 + ...
   ((x(4)+x(6))/2-x(2))^2) - dist(2));
%Disease #2 vs Disease #1 & Controls
f(3) = abs(sqrt(((x(1)+x(5))/2-x(3))^2 + ...
   ((x(2)+x(6))/2-x(4))^2) - dist(3));
%Disease #1 vs Disease #2
f(4) = abs(sqrt((x(1)-x(3))^2 + (x(2)-x(4))^2) - dist(4));
%Disease #1 vs Controls
f(5) = abs(sqrt((x(1)-x(5))^2 + (x(2)-x(6))^2) - dist(5));
%Disease #2 vs Controls
f(6) = abs(sqrt((x(3)-x(5))^2 + (x(4)-x(6))^2) - dist(6));

%Return the sum of the vector of errors
f=sum(abs(f));
```

B.2 3D distance error calculation

```
function f = distance3D(x)
%Function receives as argument a vector size 7
%corresponding to the 7 non-fixed coordinates
%5 coordinates have to be fixed (to 0) to converge

%Declaration of dist as a global variable
%containing the standardized deviances
global dist

x(1) = x(1); %x-coordinate Disease #1
x(2) = x(2); %y-coordinate Disease #1
x(3) = x(3); %z-coordinate Disease #1
x(4) = x(4); %x-coordinate Disease #2
x(5) = x(5); %y-coordinate Disease #2
x(6) = x(6); %z-coordinate Disease #2
x(7) = x(7); %x-coordinate Disease #3
x(8) = 0; %y-coordinate Disease #3
x(9) = 0; %z-coordinate Disease #3
x(10) = 0; %x-coordinate Controls
x(11) = 0; %y-coordinate Controls
x(12) = 0; %z-coordinate Controls

%Disease #1 & Disease #2 vs Controls
f(1) = pdist([x(1)/2+x(4)/2,x(2)/2+x(5)/2,x(3)/2+x(6)/2; ...
   x(10),x(11),x(12)],'euclidean') - dist(1));
%Disease #1 vs Disease #2 & Controls
f(2) = pdist([x(10)/2+x(4)/2,x(11)/2+x(5)/2,x(12)/2+x(6)/2; ...
   x(1),x(2),x(3)],'euclidean') - dist(2));
%Disease #2 vs Disease #1 & Controls
f(3) = pdist([x(1)/2+x(10)/2,x(2)/2+x(11)/2,x(3)/2+x(12)/2; ...
   x(4),x(5),x(6)],'euclidean') - dist(3));
%Disease #1 & Disease #3 vs Controls
f(4) = pdist([x(1)/2+x(7)/2,x(2)/2+x(8)/2,x(3)/2+x(9)/2; ...
   x(10),x(11),x(12)],'euclidean') - dist(4));
%Disease #1 vs Disease #3 & Controls
f(5) = pdist([x(7)/2+x(10)/2,x(8)/2+x(11)/2,x(9)/2+x(12)/2; ...
   x(1),x(2),x(3)],'euclidean') - dist(5));
%Disease #3 vs Disease #1 & Controls
f(6) = pdist([x(1)/2+x(10)/2,x(2)/2+x(11)/2,x(3)/2+x(12)/2; ...
   x(7),x(8),x(9)],'euclidean') - dist(6));
%Disease #2 & Disease #3 vs Controls
f(7) = pdist([x(4)/2+x(7)/2,x(5)/2+x(8)/2,x(6)/2+x(9)/2; ...
   x(10),x(11),x(12)],'euclidean') - dist(7));
%Disease #2 vs Disease #3 & Controls
f(8) = pdist([x(7)/2+x(10)/2,x(8)/2+x(11)/2,x(9)/2+x(12)/2; ...
   x(4),x(5),x(6)],'euclidean') - dist(8));
%Disease #3 vs Disease #2 & Controls
f(9) = pdist([x(4)/2+x(10)/2,x(5)/2+x(11)/2,x(6)/2+x(12)/2; ...
   x(7),x(8),x(9)],'euclidean') - dist(9));
```

```
%Disease #1 vs Disease #2 & Disease #3
f(10) = pdist([x(4)/2+x(7)/2,x(5)/2+x(8)/2,x(6)/2+x(9)/2; ...
   x(1),x(2),x(3)],'euclidean') - dist(10));
%Disease #2 vs Disease #1 & Disease #3
f(11) = pdist([x(1)/2+x(7)/2,x(2)/2+x(8)/2,x(3)/2+x(9)/2; ...
   x(4),x(5),x(6)],'euclidean') - dist(11));
%Disease #3 vs Disease #1 & Disease #2
f(12) = pdist([x(1)/2+x(4)/2,x(2)/2+x(5)/2,x(3)/2+x(6)/2; ...
   x(7),x(8),x(9)],'euclidean') - dist(12));
%Disease #1 vs Disease #2
f(13) = pdist([x(1),x(2),x(3);x(4),x(5),x(6)], ...
   'euclidean') - dist(13));
%Disease #1 vs Disease #3
f(14) = pdist([x(1),x(2),x(3);x(7),x(8),x(9)], ...
   'euclidean') - dist(14));
%Disease #2 vs Disease #3
f(15) = pdist([x(4),x(5),x(6);x(7),x(8),x(9)], ...
   'euclidean') - dist(15));
%Disease #1 vs Controls
f(16) = pdist([x(1),x(2),x(3);x(10),x(11),x(12)], ...
   'euclidean') - dist(16));
%Disease #2 vs Controls
f(17) = pdist([x(4),x(5),x(6);x(10),x(11),x(12)], ...
   'euclidean') - dist(17));
%Disease #3 vs Controls
f(18) = pdist([x(7),x(8),x(9);x(10),x(11),x(12)], ...
   'euclidean') - dist(18));

%Return the sum of the vector of errors
f=sum(abs(f));
```

B.3 2D Unconstrained Nonlinear Optimization

```
%Declaration of dist as a global variable
%containing the standardized deviances
global dist

%Initial solution guess set to 1's
x0=ones(1,3);

%Change sign of standardized deviances and make
%the remaining negatives ones to 0 + small cap
dist=max(-dist,zeros(1,6)+.01);

%Set maximum iterations to a big number and hide iteration details
options=optimset('MaxFunEvals',1000000,'Display','off');

%Get the solution
[x,fval] = fminunc(@distance2D,x0,options);
x=[x, zeros(1,3)];
```

B.4 3D Unconstrained Nonlinear Optimization

```
%Declaration of dist as a global variable
%containing the standardized deviances
global dist

%Initial solution guess set to 1's
x0=ones(1,7);

%Change sign of standardized deviances and make
%the remaining negatives ones to 0 + small cap
dist=max(-dist,zeros(1,18)+.01);

%Set maximum iterations to a big number and hide iteration details
options=optimset('MaxFunEvals',1000000,'Display','off');

%Get the solution
[x,fval] = fminunc(@distance2D,x0,options);
x=[x, zeros(1,5)];
```