GENE-ENVIRONMENT INTERACTIONS IN NON-HODGKIN LYMPHOMA: A STATISTICAL ANALYSIS

by

Maria de los Angeles Santiago Jimenez

B.Sc., Universidad de las Américas-Puebla, 2002

A Project submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2013

APPROVAL

Name: Maria de los Angeles Santiago Jimenez

Degree: Master of Science

Title of Project: Gene-environment interactions in non-Hodgkin lymphoma: a statistical analysis

Examining Committee:

Dr. Tim Swartz
Professor
Chair

Dr. Jinko Graham
Senior Supervisor
Associate Professor

Dr. John Spinelli
Co-Supervisor
Adjunct Professor

Dr. Brad McNeney
External Examiner
Associate Professor

Date Approved: _____

Partial Copyright Licence

Abstract

An emerging focus of cancer epidemiology is the role of the environment together with genes in determining risk, often referred to as gene-environment interaction. For non-Hodgkin lymphoma (NHL), environmental exposures such as organochlorines are important risk factors. On the other hand, familial clustering of NHL suggests that genetics also plays a role. In this project, we analyze data from a BC population-based case-control study of NHL, to evaluate gene-environment interactions between the organochlorine oxychlordane and single-nucleotide polymorphisms (SNPs) that tag genes involved in the elimination of foreign compounds from the body. A statistically significant interaction between oxychlordane and an intronic SNP within the ABCC4 gene was identified at false-discovery rate level 10%. The same intronic region of ABCC4 produced the four most significant interactions. These results may be viewed in the context of recent work connecting intronic SNPs to regulation of gene expression and the development of cancer.

Keywords: gene-environment interaction; Benjamini-Hochberg procedure; LRT statistic; logistic regression; ABCC4; alternative splicing

To my beloved family. A Chibilita, Don groño, mi Manta, Teti, mi Samoso y mis futuros samositos.

"Education is for improving the lives of others and for leaving your community and world better than you found it."

— *Marian Wright Edelman,* The Measure of Our Success: A letter to My Children & Yours, *Boston, Beacon Press, 1992*

Acknowledgments

I will always keep in my heart a deep gratitude for all the wonderful people I had the chance to meet in SFU and Vancouver because without each of them, I would have not accomplished this dream. I see this degree as a privilege because back in Mexico, only 2 out of 100 kids who start school get to this point.

Many thanks to all my professors: Dr. Tom Loughin, Dr. Joan Hu, Dr. Rachel Altman, Dr. Richard Lockhart, Dr. Carl Schwartz, Dr. Tim Swartz, Dr. Brad McNeney and Profr. Ian Bercovitz; committed, brilliant and extraordinary human beings. Also, my special thanks to Robin Insley, Dr. Dave Campbell, Dr. Steve Thompson and Dr. Derek Bingham for the opportunity they gave me to work as their TA and learn from them.

My deep gratitute to my supervisor Dr. Jinko Graham who was always interested in my academic success and whose guidance helped me to identify my areas of opportunity. Thank you very much for your support for the completion of this thesis. Also, many thanks to my co-supervisor Dr. John Spinelli and my committee member, Dr. Brad McNeney for their support, guidance and valuable suggestions for improving this work.

Sadika, Kelly and Charlene, you made my student life much more easier. I appreciate your kindness, your smiles and constant support in the statistics department.

Also, I would like to thank ALL my friends in IRMACS, K9501 and also all my new friends in P9309. Each of you helped me in your own unique way, each of you are awesome and each of you are special to me. I wish you all great success in life.

Finally, thank you God for all your blessings.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Understanding the risk of complex diseases in a population involves the study of a variety of factors. In particular, genetic information and environmental exposures interact with each other and affect disease risk. Gene-environment interactions (GE) are understood as the joint effect of one or more genes with one or more environmental factors that cannot be fully explained by their separate marginal effects. Traditionally, the null hypothesis of no interaction is based on a multiplicative model in which the odds of disease in subjects with both the genetic and environmental risk factors is the product of the odds of disease in subjects with each risk factor.

In this investigation, we analyze data from a BC population-based case-control study of non-Hodgkin lymphoma (NHL) (Spinelli *et al.*, 2007), to evaluate gene-environment interactions between the organochlorine oxychlordane and single-nucleotide polymorphisms (SNPs) that tag variation in genes involved in the elimination of foreign compounds from the body. Previous studies have found evidence that environmental exposures such as organochlorines are important risk factors and familial clustering of NHL suggests that genetics also plays a role.

This project is organized in the following way. After a general description of relevant genetic and environmental risk factors in non-Hodgkin lymphoma, in Chapter 2 we present descriptive statistics and features of the genotype and environmental data from our study. Chapter 3 presents the statistical methods used including logistic regression models and the Benjamini-Hochberg multiple testing procedure as well as power assessment under specific scenarios. A summary of the results obtained in our statistical analysis is shown in Chapter 4. Finally, in Chapter 5 we discuss our findings and their possible biological context as well

as potential improvements and directions for future work.

1.1 Overview of non-Hodgkin lymphoma (NHL)

NHL occurs in *lymphocytes*, a type of white blood cell of small size that constitutes one quarter to one third of the total number of white blood cells in humans. Lymphocytes help the immune system by identifying foreign substances and microorganisms and removing them from the body. These cells are developed in the spleen, bone marrow, lymph nodes and thymus and when mature they become part of the lymph and blood. There are three types of lymphocytes: T-cell, B-cell and NK-cells.

Lymphomas are cancers that occur when lymphocytes develop an abnormal size or multiply without control or when older lymphocytes no longer perform their functions adequately. Most cases of NHL have been observed in B-cells with *diffuse large B-cell lymphoma* and *follicular lymphoma* being the most frequent. The categorization of lymphoma as Hodgkin and non-Hodgkin has been used since the 19th century, but currently the World Health Organization (WHO) classifies them in 4 groups and 80 sub-types. Nonetheless, the former classification is still in use and more than 30 sub-types of NHL have been identified based on their specific morphology, immunophenotype and somatic genetics.

In 2008, the WHO estimated 356,000 new cases of NHL and 192,000 deaths from this disease in the world. Such numbers account for nearly 5% of all cancer cases and 2.7% of all cancer deaths. Worldwide, NHL is the eighth most common cancer type in men and the eleventh in women (Bofetta, 2011). The highest incidence rates are observed in developed countries such as US, UK, Canada and Australia. The rates have been rising for the last 40 years but during the last decade they have become stable. NHL is the fifth most frequent type of cancer diagnosed in men and the sixth in women in Canada. The Canadian Cancer Society reported 7,800 new cases and 2,800 NHL deaths in 2012. With respect to ethnicity, higher incidence rates have been observed in the US for whites than for African American or Asian American individuals. It is also known that the risk of NHL increases with age and cases in men are more frequent than in women. In the US National Cancer Institute's Surveillance Epidemiology and End Results database (`http://seer.cancer.gov`), the age-adjusted incidence rate reported in 2010 for non-hispanic white individuals was 17.16 for females and 24.9 for males whereas the overall 2010 incidence rate was 19.5.

Other risk factors are weakened or deficient immune system, certain infections and autoimmune diseases. Studies have suggested that exposure to radiation and toxic chemicals such as benzene and certain herbicides and insecticides are associated to higher risk of NHL (Kramer *et al.*, 2012). Spinelli *et al.* (2007) provided evidence that organochlorine chemicals contribute to NHL, finding the strongest association in oxychlordane, a metabolite of the pesticide chlordane.

1.2 Xenobiotic metabolism pathway

Metabolic pathways involve sequences of chemical reactions occurring within a cell which allow the organism to convert an initial molecule into another product for immediate, posterior use or elimination. Molecules of drugs, poisons and environmental pollutants are identified as *xenobiotics* or strange compounds in the body where a network of pathways metabolizes them. Metabolic functions are coded in the DNA and recent studies suggest that the genetic variations in these pathways may affect NHL risk. Wang *et al.* (2010) presents a summary of NHL research in GE interactions and emphasized that identifying an important pathway is not an easy task; for instance, it is still not clear whether exposure to organochlorine compounds affect NHL risk through an immune mechanism (Colt *et al.*, 2009) or through a DNA repair mechanism.

1.3 Research question

To avoid overlap with other research projects that are allied with the parent study, this project will focus on the statistical analysis of gene-environment interactions between individual SNPs in the xenobiotic metabolic pathway and the organochlorine oxychlordane. SNPs are changes in the DNA sequence occurring at a single base pair. Earlier independent analyses of the NHL data have examined only main effects of SNPs as well as main effects of organochlorines (e.g. Liu, 2012; Qu, 2009) and for this reason we have conducted a preliminary analysis to gain insight into the power of the GE interaction tests, restricting our analysis to individuals of white ethnicity.

Chapter 2

Data

In this chapter we describe the main features of the NHL data (Spinelli *et al.*, 2007). As mentioned, these data have been analyzed with a focus on environmental risk factors only. Qu (2009) found that several organochlorines were associated with NHL, chief among them, oxychlordane. The data have also been analyzed with a focus on genetic risk factors only. Liu (2012) followed up preliminary results on a gene in the histone pathway. He looked at all available SNPs from candidate genes in this pathway and found the group of all SNPs to be associated with NHL.

2.1 Data description

2.1.1 Case-control study

The NHL study was conducted between March 2000 and February 2004 in British Columbia and Victoria, Canada. There were 828 cases and 848 controls who participated in the study and that were frequency matched to cases based on age, gender and residential location in a 1:1 ratio approximately. A subset of 881 individuals (422 cases and 459 controls) had organoclorines measures after excluding cases in which individuals reported weight loss above 10% during the last 12 months prior to the analysis as well as cases treated with chemotherapy before blood collection. Special care was taken to select cases that were not infected with HIV. Based on this parent study, in this project we focused on all subjects of white European background who had oxychlordane measures and genotype data, resulting in a subgroup of 653 subjects.

2.1.2 Matching covariates

There are 341 individuals in the control group and 312 NHL cases. Of the 653 subjects, 350 are men and 303 women. Table 2.1 presents the number of cases and matched controls by sex and age group. Following Spinelli *et al.* (2007) we used a categorical coding for individual's age, with the following groups: 20-49, 50-59, 60-69, and 70+ years.

	Cases (%)		Controls (%)		Total
Gender					
Female	140	(21%)	163	(25%)	303
Male	172	(26%)	178	(27%)	350
Age-group (years)					
20-49	46	(7%)	56	(9%)	102
50-59	75	(11%)	83	(13%)	158
60-69	81	(12%)	82	(13%)	163
70+	110	(17%)	120	(18%)	230
Total	312	(48%)	341	(52%)	653

Table 2.1: Summary of case-control samples

2.1.3 Genotype data

Sporadically missing genotypes in the subgroup of 653 subjects have been imputed with the genotype imputation program BEAGLEv3.3 (Browning & Browning, 2009) as described previously (Liu, 2012). A total of 243 SNPs in 43 genes within 18 chromosomes were considered for our analysis. A summary of the number of the SNPs within each gene is shown in Table 2.2 and names of the 243 SNPs are shown in Appendix A. All of these SNPs correspond to genes in the xenobiotic metabolism pathway, excluding the AHR gene and CYP family of genes. The exclusions of AHR and CYP genes were done in order to avoid overlap with concurrent analyses. Figure 2.1 presents the "allele frequency spectrum" or the distribution of the minor allele frequencies (MAFs), both in cases and in controls. Similar spectra are observed for both cases and controls.

Chromosome	Gene	No.	Chromosome	Gene	No.
1	GSTM3	2	4	ABCG2	9
1	ARNT	7	5	AHRR	12
1	NR1I3	3	6	GSTA2	2
1	EPHX1	3	6	SOD2	1
2	ABCB11	6	7	ABCB1	15
2	NFE2L2	2	7	PON1	14
2	UGT1A9	2	8	NAT1	11
2	UGT1A7	1	8	NAT2	6
2	UGT1A6	4	8	GSR	3
2	UGT1A7	1	9	ABCA1	14
2	UGT1A6	1	10	ABCC2	6
2	UGT1A9	1	11	SLC22A8	4
2	UGT1A3	3	11	GSTP1	1
2	UGT1A1	1	13	ABCC4	34
2	UGT1A9	1	14	ESR2	9
2	UGT1A1	1	16	ABCC1	15
2	UGT1A7	1	16	NQO1	4
2	UGT1A1	7	17	ABCC3	9
2	UGT1A8	5	19	SULT2A1	4
3	NR1I2	6	22	COMT	5
4	UGT2B7	1	23	HPRT1	2
4	UGT2B4	4	Total		243

Table 2.2: Number of SNPs analyzed by gene

Oxychlordane levels	Cases (%)		Controls (%)		Total
≤ 6.07	74	(11%)	51	(8%)	125
$> 6.07 \leq 9.76$	92	(14%)	69	(11%)	161
$> 9.76 \leq 13.7$	85	(13%)	71	(11%)	156
$> 13.7 \leq 58.21$	90	(14%)	121	(19%)	211
Total	341	(52%)	312	(48%)	653

Table 2.3: Frequency of oxychlordane levels in cases and controls

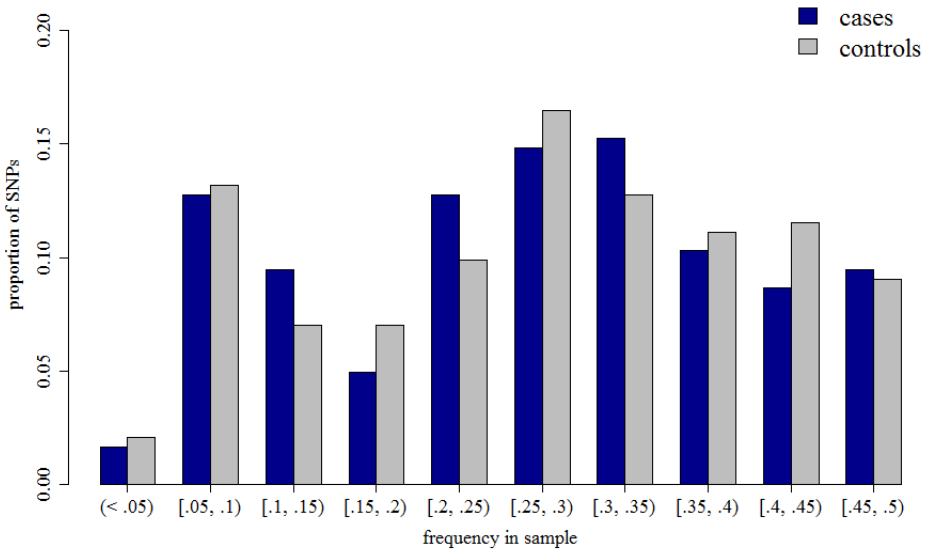

Figure 2.1: Distribution of the minor allele frequencies

2.1.4 Environmental data

Following Spinelli *et al.* (2007), 23 organochlorine measurements taken from blood samples were analyzed and recoded according to their empirical quartiles, so that all measurements in the same quartile are assigned the quartile-specific median value. For our analysis, we will use oxychlordane, which was the most significantly associated organochlorine with NHL (Qu, 2009). Table 2.3 presents the frequency of oxychlordane values (ng/g) in both case and control samples. The four categories are ≤ 6.07, $> 6.07 \leq 9.76$, $> 9.76 \leq 13.7$, $> 13.7 \leq 58.21$ with quartile-specific median values $4.56, 8.01, 11.48$ and 17.67 respectively. These values represent the $12.5th, 37.5th, 62.5th$ and $87.5th$ percentiles.

Figure 2.2 presents the number of NHL cases classified by oxychlordane level and age group. The majority of individuals correspond to the oldest group and the highest level of oxychlordane whereas younger individuals show lower oxychlordane levels in general. A similar distribution was also observed in controls (results not shown) and may reflect the fact that older individuals were exposed to chlordane for longer periods than younger subjects and also, the fact that the metabolism of these foreign substances changes with age.

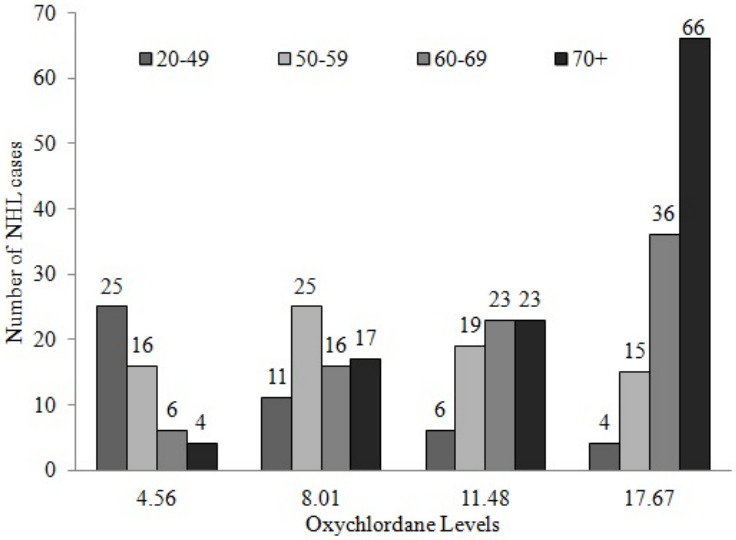

Figure 2.2: Oxychlordane levels by age group in NHL cases

2.2 Data processing

Genotype and subject information for white Europeans were merged with the 23 organochlorine measurements. Unfortunately, many subjects with genetic data had to be dropped (463/1116) because they had no organochlorine measurements. As mentioned, the sporadically missing genotypes were imputed by Beagle v3.3 (Browning & Browning, 2009), a program that applies *haplotype-phase-inference* methods to determine the allele dosages of missing SNPs.

We started with 1079 SNPs in 653 subjects, including SNPs outside the xenobiotic metabolism pathway. Out of a total of 346 SNPs in the xenobiotic metabolism pathway, we removed 6 of the SNPs because their minor allele frequency was less than 1% in the 653 subjects (i.e. they were essentially monomorphic). We then removed 97 SNPs within the AHR gene or the CYP family of genes. These 97 SNPs were removed from the pathway because their corresponding genes have been or are currently the focus of separate investigations. After pruning, we were left with a total of 243 SNPs to be tested for individual statistical interaction with oxychlordane. Each SNP in the data is coded based on the number of copies of the population minor allele as 0, 1 or 2. We coded oxychlordane levels as the numeric value representing the mid-points of the empirical quartiles, i.e., $4.56, 8.01, 11.48$ and 17.67.

Chapter 3

Methods

Case-control studies are based on retrospective sampling designs where a set of covariates are observed for each subject conditional on the response variable. Logistic regression models may be used to estimate odds ratios and effects in such studies (Prentice & Pyke, 1979). The major aim of most case-control studies is to test for main effects and the chosen study sample size should guarantee the predetermined power for detecting such effects. In the original NHL study (Spinelli *et al.*, 2007), interest was in the environmental main effects (organochlorines) and the study was powered accordingly. However, in this investigation we are interested in GE interactions; therefore, we conducted a preliminary analysis to gain insight into the power of the tests.

This chapter is organized as follows. In the first section we describe the characteristics of the proposed logistic regression model that includes an interaction term between a single SNP (G) and the environmental exposure (E) and then we present the characteristics of the Wald and Likelihood Ratio Test (LRT) statistics. The second section includes methodological considerations after conducting multiple tests for interaction effects and describes the Benjamini and Hochberg procedure (1995) used to adjust the LRT p-values and control the family-wise error rate. In addition, we include the methodology used to determine the power of the tests and, finally, we describe a pairwise linkage disequilibrium estimator to assess the potential association between SNPs.

3.1 Single SNP

3.1.1 Logistic regression model

Consider Y as the binary response variable indicating disease status where $Y = 1(Y = 0)$ stands for an affected (unaffected) individual. For a k-dimensional covariate X, the model for the conditional probability of disease given the covariates $\pi(x) = P(Y = 1|X = x) = E(Y = 1|X = x)$ for individual i is:

$$\pi(x) = \frac{exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}$$

or equivalently,

$$logit[\pi(x)] = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

where the parameter β_j refers to the effect of covariate x_j on the log odds that $Y = 1$, while controlling the other covariates.

For m subjects and provided that the Y_i's are independent binary variables, the data follow a Bernoulli distribution,

$$Y_i|X_i = x_i \sim Bernoulli(\pi(x_i))$$

therefore, the likelihood function for these data is

$$\mathcal{L}(\beta) = \prod_{i=1}^{m} \pi(x_i)^{Y_i}(1 - \pi(x_i))^{1-Y_i}.$$

The likelihood equations result from setting $\partial\mathcal{L}(\beta)/\partial\beta = 0$. These equations are nonlinear and require numerical methods to solve. The algorithm used to obtain the MLE $\hat{\beta}$ is called iteratively reweighted least squares (Green, 1984).

Notation

In a case-control study with m subjects each having n SNPs genotyped in the xenobiotic metabolic pathway, let Y denote disease status of an individual, where

$$Y = \begin{cases} 1 & \text{if NHL case} \\ 0 & \text{if control} \end{cases}$$

Let G denote the SNP genotype, coded as the number of copies (0, 1 or 2) of the index allele and E denote the environmental exposure or oxychlordane level (4.56, 8.01, 11.48 or 17.67). Also, let $\mathbf{A}$ represent the set of adjustment covariates:

- Gender (with value 0 for males and 1 for females)
- Age group (with categories: 20-49, 50-59, 60-69 and 70+ years). We will use the first group as the baseline category.

The adjustment covariates $\mathbf{A}$ are coded as binary vectors; for example, [1 0 0 0] represents the vector of a 21 year-old woman, [0 0 0 1] is the vector of a 72 year-old man and [1 1 0 0] represents the covariates of a 53 year-old woman.

Model

The logistic regression model to test for gene-environment (GE) interaction can be written as:

$$\pi(x) = \frac{exp(\beta_0 + \beta_G G + \beta_E E + \beta_{GE} GE + \boldsymbol{\beta}_{\boldsymbol{A}}^T \boldsymbol{A})}{1 + exp(\beta_0 + \beta_G G + \beta_E E + \beta_{GE} GE + \boldsymbol{\beta}_{\boldsymbol{A}}^T \boldsymbol{A})}$$

where β_0 represents the intercept term, β_G, β_E and β_{GE} are the fixed regression coefficients for the SNP genotype G, the environmental exposure E and the interaction term GE respectively. Also, $\boldsymbol{\beta_A}$ is the vector of regression coefficients for the adjustment covariates gender and age group.

The marginal effect of the genetic component on the risk of NHL for individuals of the same gender and age group depends on the oxychlordane measurement observed and is equal to $exp(\beta_G + \beta_{GE} E)$. Similarly, $exp(\beta_E + \beta_{GE} G)$ represents the marginal effect of oxychlordane level on the risk of NHL for individuals of the same gender and age group for a given SNP genotype. The term $exp(\beta_{GE})$ represents the GE interaction effect on the risk of NHL for individuals of the same gender and age group.

3.1.2 Statistical tests of GE interaction

Two asymptotically equivalent tests may be used to examine gene-environment interaction: Wald test and Likelihood Ratio Test (LRT).

The Wald test considers the following hypotheses

$$H_0 : \beta_{GE} = 0 \text{ vs } H_1 : \beta_{GE} \neq 0$$

to test GE interaction in the logistic regression model. The corresponding Wald test statistic has the form:

$$z = \frac{\hat{\beta}_{GE}}{se(\hat{\beta}_{GE})}$$

where $\hat{\beta}_{GE}$ is the MLE of β_{GE} and $se(\hat{\beta}_{GE})$ is the standard error of $\hat{\beta}_{GE}$. Under H_0, z^2 is asymptotically χ^2_1. It is important to mention that the Wald test may have lower power than LRT when $|\beta_{GE}|$ is relatively large and can show unstable behavior under some circumstances (Hauck & Donner, 1977).

The Likelihood ratio test (LRT) considers information of both the log-likelihood at the null value $\beta_{GE} = 0$ and at $\hat{\beta}_{GE}$. For small or moderate sample sizes, simulation studies have shown that the LRT is more reliable than the Wald test and it is considered more versatile (Agresti, 2012). This test statistic has the form:

$$-2(\hat{L}_0 - \hat{L}_1)$$

where $\hat{L}_1$ is the log-likelihood of the fitted model with GE interaction (M_1) and $\hat{L}_0$ is the log-likelihood of the fitted model without GE interaction (M_0). For large sample sizes, this model comparison statistic has an approximate chi-squared null distribution. Since in our model we consider the case where G and E are treated as continuous variables, the difference in degrees of freedom based on (M_1) and (M_0) is 1. Therefore, for testing GE interaction we assume that $-2(\hat{L}_0 - \hat{L}_1) \sim \chi^2_1$ under H_0.

3.2 Multiple SNPs

3.2.1 Q-Q plot analysis

Q-Q plots on the $-log10$ scale are commonly used to compare LRT p-values versus their null expectation. The use of Q-Q plots in this setting relies on the fact that under the null hypothesis, the expected p-values are i.i.d. observations from a standard uniform distribution and their k-th order statistic follows a beta distribution. A justification of the beta distribution based on these premises is provided in Appendix B.

Because of the linkage disequilibrium structure of SNPs in our data, we caution that dependence amongst test statistics may exist and the use of this plot may not be fully justified except as an exploratory tool.

3.2.2 Multiple testing of GE interactions

Since many GE interactions need to be tested, a multiple testing procedure has to be used to control the family-wise error rate. One of the most common methods is the Benjamini-Hochberg (BH) procedure (1995) that controls the false-discovery rate (FDR), which is the expected proportion of false rejections (rejected true null hypotheses) amongst all rejected hypotheses.

Let n be the total number of null hypotheses and n_0 the true null hypotheses. Based on a specific *rule*, we reject R hypotheses in which $V \subset R$, $V \leq n_0$ correspond to true null hypotheses. The false-discovery rate is defined as:

$$FDR = \begin{cases} E\left[\frac{V}{R}\right] & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

The BH procedure defines a rejection *rule* in the following way:

- Consider the ordered p-values $P_{(1)} < ... < P_{(n)}$
- Let $P_{(T)}$ be the BH rejection threshold, where T is defined as:

$$T = max\left\{ i : P_{(i)} < \frac{i}{n}\alpha \right\}$$

- Then, the set of rejected hypotheses is:

$$R = \left\{ H_{0j} : P_{(j)} \leq P_{(T)} \right\}$$

The FDR-adjusted p-values $\tilde{P}_{(i)}$ can be interpreted as the smallest nominal FDR at which we would reject the set of null hypotheses, given the values of the test statistics. These adjusted p-values are defined as follows:

$$\tilde{P}_{(i)} = \begin{cases} P_{(n)} & \text{for } i = n \\ min(\tilde{P}_{(i+1)}, \frac{n}{i}\tilde{P}_{(i)}) & \text{for } i = n-1, ..., 1 \end{cases}$$

For this project, we will declare a result interesting for potential follow up if $\tilde{P}_{(i)}$ is significant at false discovery rate level 10%.

3.2.3 Power analysis

In order to gain insight into the power of the tests of interaction between a SNP (G) and oxychlordane (E) we conducted a preliminary power analysis following the methodology described by Gauderman (2002) which is based on the computation of the expected values of the LRT statistic. For tractability, we assumed a simple logistic regression model for an unmatched case-control study involving G and E that are measured without error. This model ignores the heterogeneity in NHL risk due to age and gender; hence, the power calculations are expected to be optimistic.

We assume the true model is:

$$log\left[\frac{P(D=1|G=g,E=e)}{P(D=0|G=g,E=e)}\right] = \beta_0 + \beta_G * g + \beta_E * e + \beta_{GE} * g * e.$$

Let

$$\begin{aligned} odds_0(e) &= \frac{P(D=1|G=g,E=e)}{P(D=0|G=g,E=e)} \\ &= exp\big[\beta_0 + \beta_G * (g+0) + \beta_E * e + \beta_{GE} * (g+0) * e\big] \end{aligned}$$

$$\begin{aligned} odds_1(e) &= \frac{P(D=1|G=(g+1),E=e)}{P(D=0|G=(g+1),E=e)} \\ &= exp\big[\beta_0 + \beta_G * (g+1) + \beta_E * e + \beta_{GE} * (g+1) * e\big] \end{aligned}$$

The genotype odds ratio (OR_g) which describes the genetic effect at a given value of e is determined as:

$$OR_g(e) = \frac{odds_1(e)}{odds_0(e)} = exp\big[\beta_G + \beta_{GE} * e\big]$$

To obtain plausible values of β_{GE} we can contrast genotype odds ratios at different values of E that are one standard deviation (s.d.) apart. We know that the reported oxychlordane levels correspond to the $12.5th$, $37.5th$, $62.5th$ and $87.5th$ percentiles, which are the mid-points of the quartiles. We assumed a normally distributed E and used the value of the $87.5th$ percentile to approximate the $84th$ percentile, which roughly corresponds to the mean plus 1 s.d. σ_e. Then we can determine σ_e as:

$$\begin{aligned} \sigma_e &\approx 84th \text{ percentile}_e - \text{mean}_e \\ &\approx 87.5th \text{ percentile}_e - \frac{1}{2}(37.5th \text{ percentile}_e + 62.5th \text{ percentile}_e) \\ &\approx 17.67 - \frac{8.01 + 11.5}{2} = 7.915 \approx 7.5 \end{aligned}$$

Therefore, the contrast f in genotype odds ratios for values e_1 and e_2 of E such that $e_2 - e_1 = \sigma_e$ is determined as:

$$f = \frac{OR_g(e_2)}{OR_g(e_1)} = exp\big(\beta_{GE} * \sigma_e\big) \approx exp\big(\beta_{GE} * 7.5\big)$$

Table 3.1 presents OR contrast values f and the corresponding plausible values of the parameter β_{GE}. For instance, a contrast f value of 1.4 (i.e. $\beta_{GE} = 0.05$) would indicate that the genotypic odds ratio at the upper extreme of E is approximately four times larger than at the lower extreme (4 s.d. apart). Also, for $f = 2.9$ the genotypic odds ratio at the upper extreme of E is more than 60 times larger than at the lower extreme.

f	β_{GE}	$exp(\beta_{GE})$
1.4	0.05	1.05
1.7	0.07	1.07
1.9	0.09	1.09
2.2	0.10	1.11
2.5	0.12	1.13
2.9	0.14	1.15

Table 3.1: OR contrasts values f

Power calculations were conducted using the software QUANTO v1.2.4 (Gauderman, 2002), specifying an unmatched case-control study with a sample size of 312 cases and one control per case. To account for the multiple testing of 243 SNPs, we used a significance level for a 2-sided alternative of $0.1/243 = 0.0004$. We assumed that only one SNP interacts with E under the alternative hypothesis and that all SNPs have the same minor allele frequency q_A that ranged from 0.05 to 0.5. The standard deviation of the oxychlordane measurements σ_e was taken to be 7.5. G and E main effects were specified by odds ratios $exp(\beta_G) = exp(\beta_E) = 1.1$, based on contrasts of one unit increase in G and E. The use of an odds ratio of 1.1 for one unit increase in E roughly corresponds to a 2.7 odds ratio for a 13-unit increase in E (highest vs lowest quartiles), which is the reported effect size estimate in the study (Spinelli *et al.*, 2007). Other main effect sizes may be considered for the power analysis. The software also requires users to specify the disease risk K_p over the study period that we estimated based on the NHL incidence for white ethnicity published by the US National Cancer Institute in their Surveillance Epidemiology and End Results (SEER) database. We

considered an approximate incidence of 21/100000 (for both males and females) and, for a four-year study, the NHL risk was estimated as $K_p = 4(21)/100000 = 84/100000$.

3.2.4 Linkage disequilibrium estimator

Linkage disequilibrium (LD) reflects the occurrence of specific combinations of alleles at a frequency that is different than the one expected based on their independent individual frequencies. Generally speaking, neighboring SNPs are likely to be in linkage disequilibrium and genetic association studies can exploit these associations to facilitate the identification of potential functional SNPs by tagging the most representative ones within a region.

First, suppose A_1 and a_1 are the possible alleles at SNP 1 and A_2, a_2 are the alleles at SNP 2. Let $p_{A_1}, p_{a_1}, p_{A_2}$ and p_{a_2} be the population frequencies for each allele and let $p_{A_1A_2}, p_{A_1a_2}, p_{a_1A_2}, p_{a_1a_2}$ be the joint probability of each allele pair. If we consider the distribution of alleles for m subjects, there will be $2m$ base pairs to analyze in the sample.

The most common pairwise LD measure is the r^2 statistic which corresponds to the squared Pearson correlation between two binary variables, in this setting: SNP 1 and SNP 2:

$$r^2 = \frac{(p_{A_1A_2} - p_{A_1}p_{A_2})^2}{p_{A_1}p_{a_1}p_{A_2}p_{a_2}}.$$

This measure ranges between 0 and 1 where a value close to 1 indicates a high level of LD. A high level of LD means that both SNPs would provide nearly the same information in the context of association studies.

Chapter 4

Results

In this chapter we present the results of our statistical analysis starting with the power analysis based on the LRT statistic. Next, we analyze plots of the observed p-values versus their null expectation and the false-discovery rates after multiple testing adjustment. Finally, based on our findings, we present the estimated GE interaction effects.

4.1 Power analysis

After specifying the plausible interaction effect sizes in table 3.1, in Figure 4.1 we present the power curves from scenario 1 where we specified fixed G and E main effects by odds ratios of $exp(\beta_G) = exp(\beta_E) = 1.1$.We also considered no genetic main effect for scenario 2 with similar results. We computed the power considering a 2-sided alternative with significance level of 0.0004 that takes into account the multiple testing adjustment. A decent power ($> 80\%$) is observed when the minor allele frequencies are around 0.25 and β_{GE} is in an approximate range of 0.1 to 0.13. Power decreases when the minor allele frequencies are near 0.05 or 0.5. This behavior was observed in both scenarios.

4.2 LRT p-values and false-discovery rates

For each of the 243 logistic regression models based on SNPs in the xenobiotic metabolism pathway, the raw p-values based on Wald and LRT statistics were obtained; however, only the results of LRT were considered for further analysis. We constructed a Q-Q plot (Figure 4.2) on the $-log10$ scale based on LRT p-values versus their null expectation. Also, we

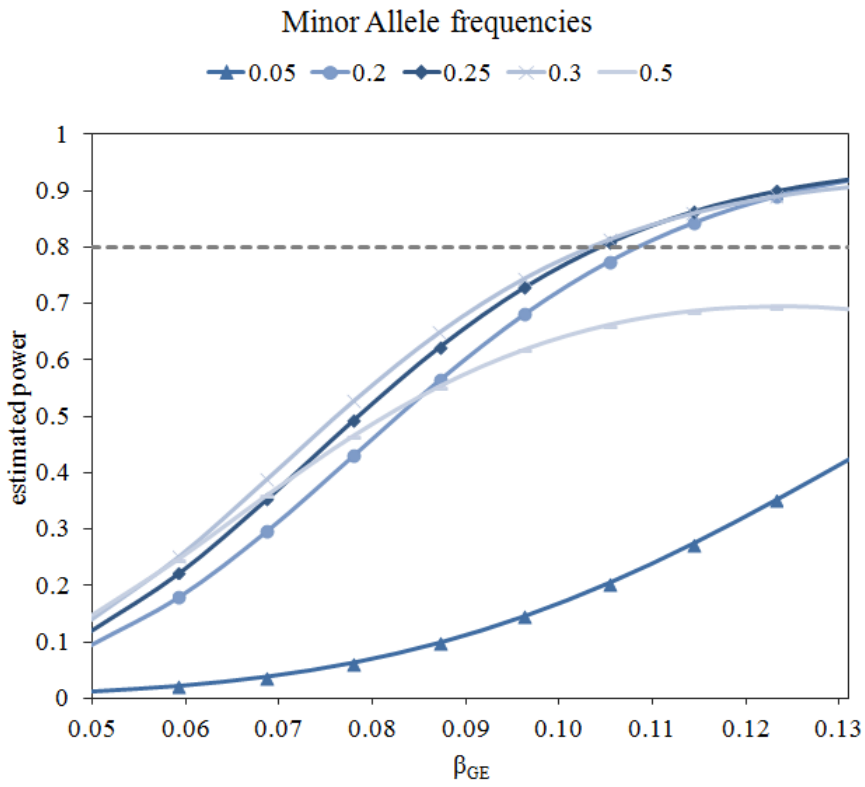

Figure 4.1: Statistical power under scenario 1

included 90% prediction intervals that were constructed based on the beta distribution with parameters $\alpha = i, \beta = 244 - i, i = 1, ..., 243$ corresponding to the 243 analyzed SNPs. From the plot we do not observe clear evidence of an interaction "signal" because all observed p-values lie within the 90% prediction intervals indicating no deviation from the null p-values. However, SNPs in gene *ABCC4* showed the largest deviations from the expected p-values.

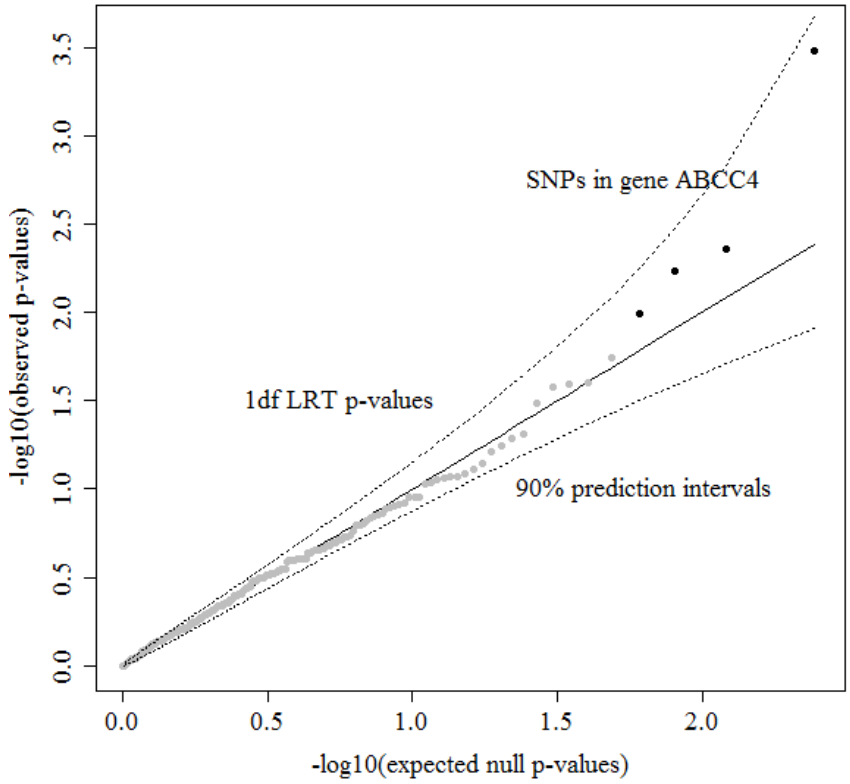

Figure 4.2: Q-Q plot of LRT p-values versus their null expectation

Figure 4.3 shows the p-values adjusted for multiple testing based on the BH procedure. The four most significant interactions were produced by SNPs in gene *ABCC4* which corresponds to chromosome 13. SNPs *rs1189465*,*rs9561773* and *rs1618738* have FDR-adjusted p-values (i.e. q-values) less than 50% (Table 4.1). As can be observed, only SNP *rs1189465* was identified to have a q-value less than 10%.

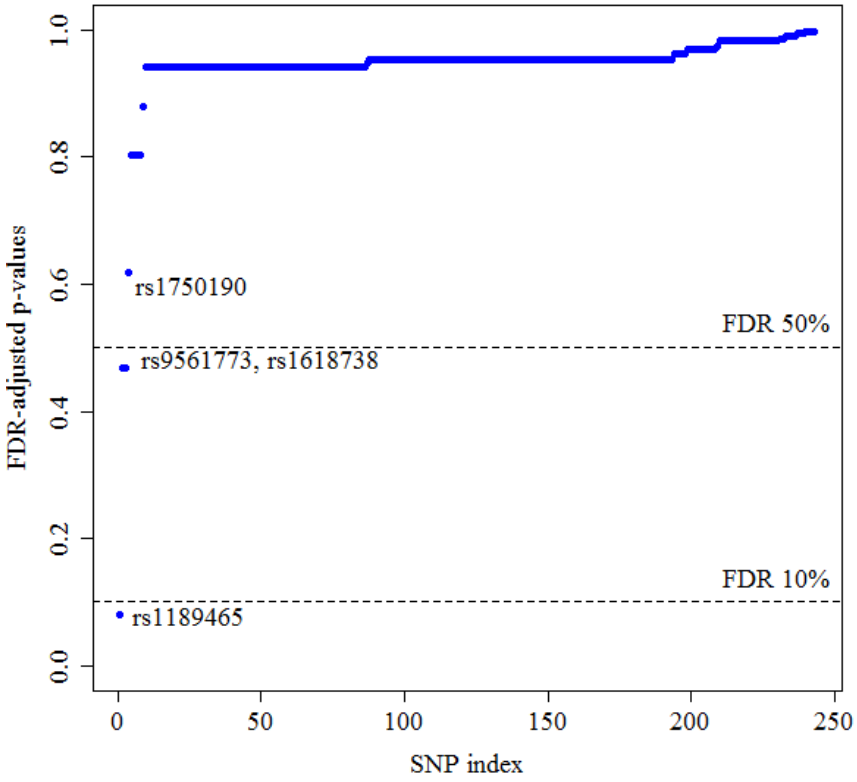

Figure 4.3: FDR-adjusted p-values for GE interaction

SNP	Gene	Chr.	P-value	
			nominal	adjusted
rs1189465	ABCC4	13	0.00033	0.08
rs9561773	ABCC4	13	0.00578	0.47
rs1618738	ABCC4	13	0.00434	0.47
rs1750190	ABCC4	13	0.01018	0.62

Table 4.1: Most significant SNPs

4.3 Linkage disequilibrium of significant SNPs

As mentioned, the four most significant SNPs correspond to gene ABCC4. Table 4.2 shows the value of r^2 for pairwise linkage disequilibrium. These values show some degree of LD amongst the SNPs.

SNPs	rs1189465	rs9561773	rs1618738	rs1750190
rs1189465	-	0.35	0.26	0.60
rs9561773	-	-	0.25	0.45
rs1618738	-	-	-	0.06
rs1750190	-	-	-	-

Table 4.2: Linkage disequilibrium r^2 of SNPs in gene ABCC4

4.4 GE interaction model for SNP *rs1189465*

We fit the logistic regression model that includes SNP *rs1189465*; Table 4.3 presents the corresponding parameter estimates. The *z-value* is the Wald test statistic.

Coefficients	Estimate	Std. Error	z-value	$Pr(>\|z\|)$
(Intercept)	- 1.54	0.43	- 3.564	0.00037
Gender: Female	- 0.18	0.16	- 1.100	0.27120
Age Group 50-59	- 0.01	0.27	- 0.025	0.98030
Age Group 60-69	- 0.11	0.28	- 0.397	0.69133
Age Group 70+	- 0.33	0.28	- 1.176	0.23971
E: Oxychlordane	0.16	0.03	4.861	0.00000
G: rs1189465	0.79	0.29	2.714	0.00666
GE interaction	- 0.08	0.02	- 3.524	0.00043

Table 4.3: Parameter estimates for GE interaction model for SNP *rs1189465*

Based on this model, Table 4.4 presents the estimated odds ratios and 95% confidence intervals (CI) of subjects of a given genotype and oxychlordane level with respect to genotype CC and the lowest oxychlordane level (4.56). Figure 4.4 shows the estimated odds of NHL for males in the oldest age group. The pattern observed in the plots was similar for all subgroups. We can observe that the most important characteristic of the GE interaction

effect is that high levels of oxychlordane are associated with higher risk of NHL in the case of individuals with genotypes CC and CT whereas little change in susceptibility with oxychlordane levels is observed in individuals with genotype TT.

Genotype	Oxychlordane level			
	4.56	8.01	11.48	17.67
CC	1.00(————)	1.76(1.40, 2.21)	3.10(1.97, 4.90)	8.53(3.60, 20.25)
CT	1.51(1.02, 2.22)	1.98(1.29, 3.04)	2.61(1.59, 4.31)	4.27(2.17, 8.41)
TT	2.27(1.05, 4.93)	2.24(1.14, 4.41)	2.20(1.18, 4.12)	2.14(1.09, 4.18)

Table 4.4: Estimated odds ratios (95%CI) with respect to genotype CC and lowest oxychlordane level, for individuals of the same age and gender

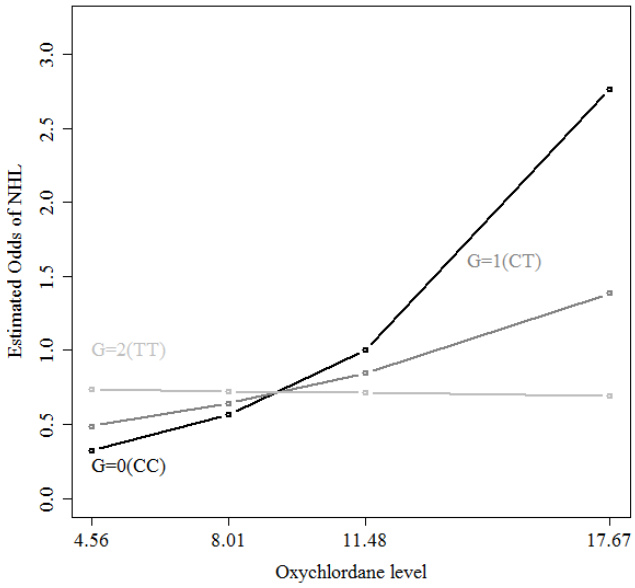

Figure 4.4: Estimated odds of NHL for males in the oldest age group for *rs1189465*

Chapter 5

Discussion

This project presented a statistical analysis of gene-environment interactions between individual SNPs in the xenobiotic metabolism pathway and the organochlorine oxychlordane. Earlier independent analyses of the NHL data examined only main effects of SNPs as well as main effects of organochlorines and for this reason we have conducted a preliminary analysis to gain insight into the power of the GE interaction tests. Most data from the parent study correspond to individuals of white European background and therefore we restricted our analysis to this subpopulation.

The analysis and interpretation of observed GE interactions will be influenced by features of the study design such as the sample size and the accuracy of the genotyping technology. Assuming measurements without error, testing for interaction effects in case-control studies requires considerable sample sizes to preserve power of the tests. Smith *et al.* (1984) emphasized that the sample size needed to test for the effect of interactions of a given magnitude has to be at least four times larger than the sample size required to test for main effects of comparable magnitude. Other aspects influencing power are the population minor allele frequencies and the GE interaction effect size under the alternative hypothesis. In this project we explored the power under two scenarios where the main genetic effect is either present or absent. A decent power was reached when the minor allele frequencies are around 0.25 and GE interaction effect sizes $exp(\beta_{GE})$ under the alternative hypothesis are in an approximate range of 1.11 to 1.15. These behaviors were observed in both scenarios we examined.

The proposed logistic regression models were based on an underlying additive genetic model and included *gender* and *age group* as covariates to reflect the frequency matching

of the study design. The underlying genetic model lead us to code G as numeric and consequently only 1 degree of freedom was required to test GE interactions because the environmental factor was assumed to be continuous.

Four neighboring SNPs in gene ABCC4 (Chr. 13) produced the most significant interactions. SNPs *rs1189465, rs9561773* and *rs1618738* were identified to have false-discovery rate less than 50%. The specific function of the protein encoded by gene ABCC4 has not yet been determined but it is known to be involved in multi-drug resistance and possibly in cellular detoxification (`http://www.genecards.org`).

SNP *rs1189465*	
Functional class:	intron
Wild type nucleotide:	C
Variant nucleotide:	T
Chromosome strand orientation:	forward
Chromosome variant position:	94524073
Protein mRNA accession:	NM_005845

Table 5.1: Information of SNP *rs1189465* (dbSNP database)

Table 4.1 provides information about SNP *rs1189465*. The UCSC Gene Browser (`http://genome.ucsc.edu`) and information of the ENCODE project (`http://www.genome.gov`) show that the 3 SNPs correspond to an intronic region which has been found to be functionally relevant in several ways. In particular, the Single Nucleotide Polymorphism database (dbSNP) (`http://www.ncbi.nlm.nih.gov/SNP`) predicts that variants in these region, including *rs1189465* may change the structure of the transcript. This mechanism is known as alternative splicing. Several studies have identified connections between aberrant splicing patterns, the behavior of splicing factors and tumor drug resistance (He *et al.*, 2009; Renshaw *et al.*, 2004; Sampath *et al.*, 2003; Vegran *et al.*, 2006). Alternatively, the SNPs we have identified may be tagging an unobserved functional variant.

The most important characteristic of the GE interaction effect for both genders and all age groups is that high levels of oxychlordane are associated with higher risk of NHL in the case of individuals with genotypes CC and CT whereas little variation in susceptibility was observed in individuals with genotype TT. Needless to say, these findings still have to be replicated in future studies. Based on the available data, slightly higher odds estimates are

observed in the males group than in females. It is worth mentioning that higher variability in the odds estimates are observed for younger ages (20-49 years old) and more precise estimates were obtained in the 70+ years old group; this is reasonable if we consider that in the study, the latter group is twice as large as the former group.

Since 25 organochlorine measurements were included in the parent study, we also verified potential GE interactions of *PCB180*, an organochlorine of potential biological interest (Spinelli, personal communication). However, none of these interactions were significant at false-discovery rate level 10% (results not shown).

Future work for this project may include the assessment of power under other scenarios and comparison of alternative methodologies to identify GE interactions such as the ones proposed by Yoo *et al.* (2012) which include classification trees and random forests.

Appendix A

Names of analyzed SNPs

Chr.	Gene	SNPs
1	GSTM3	rs2234696, rs2234696
	ARNT	rs3768015, rs10305695, rs10305710, rs3738483, rs10305724, rs10847, rs10847
	NR1I3	rs3003596, rs2307424, rs2307424
	EPHX1	rs1051740, rs2671272, rs2671272
2	ABCB11	rs17267869, rs4148797, rs853774, rs494874, rs17540154, rs17540154
	NFE2L2	rs13005431, rs13005431
	UGT1A9	rs7587916, rs7587916
	UGT1A7	rs12476197
	UGT1A6	rs12466747, rs6751673, rs7592624, rs7592624
	UGT1A7	rs12988520
	UGT1A6	rs17863787
	UGT1A9	rs12463641
	UGT1A3	rs4663965, rs11891311, rs11891311
	UGT1A1	rs10178992
	UGT1A9	rs4124874
	UGT1A1	rs10929302
	UGT1A7	rs4399719
	UGT1A1	rs4148325, rs6742078, rs17864705, rs28946889, rs17868341, rs12052787, rs12052787
	UGT1A8	rs10929303, rs11888492, rs11563251, rs2302538, rs2302538
3	NR1I2	rs3732360, rs11929668, rs6771638, rs2461822, rs4234666, rs4234666
4	UGT2B7	rs4356975
	UGT2B4	rs17671289, rs1845558, rs1826690, rs1826690
	ABCG2	rs9999111, rs2622624, rs6857600, rs2725256, rs1481012, rs2622621, rs12505410, rs2231164, rs2231164
5	AHRR	rs2672777, rs2466287, rs4956936, rs11746079, rs11742006, rs2672746, rs11740668, rs4956935, rs11742957,rs11133994, rs2672737, rs2672737
6	GSTA2	rs6577, rs6577
	SOD2	rs5746151
7	ABCB1	rs13233308, rs1858923, rs17327624, rs1211152, rs13226726, rs1989831, rs2235015, rs2235023, rs868755, rs2235033, rs1922242, rs12720066, rs2032583, rs17064, rs17064
	PON1	rs2237583, rs2299262, rs854568, rs2049649, rs3917490, rs854566, rs2272365, rs2074351, rs2299257, rs662, rs854555, rs854552, rs854550, rs854550
8	NAT1	rs4986782, rs8190845, rs7003890, rs6586714, rs17693103, rs13253389, rs4921580, rs11203943, rs4298522, rs7017402, rs7017402
	NAT2	rs1799930, rs1799929, rs1801280, rs2410556, rs13277605, rs13277605
	GSR	rs2978663, rs2253409, rs2253409
9	ABCA1	rs2487049, rs2472510, rs2437811, rs10820743, rs3905001, rs4743764, rs2515602, rs2065412, rs2254884,rs2740479, rs2740484, rs2482433, rs4149338, rs4149338
10	ABCC2	rs4148398, rs2002042, rs11190291, rs2804397, rs2756105, rs2756105
11	SLC22A8	rs2187383, rs4149182, rs2276299, rs2276299
	GSTP1	rs1138272
13	ABCC4	rs9524885, rs7324283, rs9524873, rs8001475, rs4148434, rs4148437, rs4283094, rs9516546, rs9634642, rs4773856, rs4773844, rs1751025, rs1751022, rs1678384, rs4148481, rs1611822, rs1564352, rs9561797,rs997777, rs1564355, rs1729775, rs4148527, rs1678354, rs4148530, rs1618738, rs1189465, rs1750190, rs9561773, rs1189446, rs1189449, rs10508024, rs4148546, rs4148549, rs4148549
14	ESR2	rs1952586, rs1269056, rs1273196, rs7154455, rs12435857, rs7157428, rs8017441, rs1256064, rs1256064
16	ABCC1	rs2283512, rs3887893, rs11864374, rs4148359, rs2889517, rs10852377, rs35626, rs246226, rs924135, rs1967120, rs152022, rs152023, rs215049, rs7190484, rs7190484
	NQO1	rs2917670, rs689453, rs1800566, rs1800566
17	ABCC3	rs2240802, rs11658264, rs8075406, rs739923, rs17562516, rs4793666, rs12051822, rs739922, rs739922
19	SULT2A1	rs2547231, rs2547238, rs2910393, rs2910393
22	COMT	rs4646316, rs1544325, rs933271, rs737865, rs737865
23	HPRT1	rs17324671, rs17324671

Table A.1: Names of analyzed SNPs in xenobiotic metabolism pathway

Appendix B

Null distribution of p-values

Distribution of null p-values

Given the test statistic T with continuous distribution $F(T)$, w.l.o.g., we can define the p-value P as $P = F(T)$. We can write the distribution of P as

$$Pr[P < p] = Pr[F(T) < p] = Pr[T < F^{-1}(p)] = F[F^{-1}(p)] = p$$

which corresponds to a standard uniform cdf.

Distribution of null p-values k-th order statistic

In general, for $X_1, X_2, ..., X_n$ i.i.d. continuous random variables with pdf f and cdf F we can determine the distribution function of the k-th order statistic $X_{(k)}$ by defining $N_x = \sum_{i=1}^{n} 1(X_i < x)$, the number of observations that are less or equal to $x \in \mathbb{R}$. We can observe that $\{X_{(k)} \leq x\} \Leftrightarrow \{N_x \geq k\}$ which indicates the event that at least k observations are smaller than x. Therefore

$$F_k(x) = Pr[X_{(k)} \leq x] = Pr[N_x \geq k] = \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j}$$

. Now we can formally derive the pdf $f_k(x) = F_k'(x)$.

$$\begin{aligned}
F_k'(x) &= \sum_{j=k}^{n} \binom{n}{j} \frac{d}{dx} \Big\{ [F(x)]^j [1-F(x)]^{n-j} \Big\} \\
&= \sum_{j=k}^{n} \binom{n}{j} \Big\{ j[F(x)]^{j-1} f(x)[1-F(x)]^{n-j} - (n-j)[1-F(x)]^{n-j-1} f(x) F(x)^j \Big\} \\
&= f(x) \Big\{ \sum_{j=k}^{n} j \binom{n}{j} [F(x)]^{j-1} [1-F(x)]^{n-j} - \sum_{j=k}^{n} (n-j) \binom{n}{j} [1-F(x)]^{n-j-1} F(x)^j \Big\} \\
&= f(x) \Big\{ \sum_{j=k}^{n} n \binom{n-1}{j-1} [F(x)]^{j-1} [1-F(x)]^{n-j} - \sum_{j=k}^{n} n \binom{n-1}{j} [1-F(x)]^{n-j-1} F(x)^j \Big\} \\
&= nf(x) \Big\{ \binom{n-1}{k-1} F(x)^{k-1} [1-F(x)]^{n-k} \Big\} \\
&= f(x) \frac{n!}{(n-k)!(k-1)!} F(x)^{k-1} [1-F(x)]^{n-k}
\end{aligned}$$

In the case of n i.i.d. observations from a standard uniform distribution, we have $f(x) = 1$ and $F(x) = x = p$. Therefore,

$$\begin{aligned}
f_k(p) &= \frac{n!}{(n-k)!(k-1)!} p^{k-1} [1-p]^{n-k} \\
&= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} p^{k-1} [1-p]^{n-k}
\end{aligned}$$

Which is the beta distribution with shape parameter k and rate parameter $n - k + 1$.

Bibliography

[1] Agresti, A. (2012). "Categorical Data Analysis", 3rd edition, *Wiley*.

[2] Benjamini, Y. and Hochberg, Y. (1995),"Controlling the false discovery rate: a practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society*, No. 57, 289300.

[3] Boffetta P. (2011). "Epidemiology of adult non-Hodgkin lymphoma" *Ann Oncol*, 22: iv27iv31.

[4] Browning, B.L. and Browning, S. R. (2009). "A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals", *Am J Hum Genetics*, 84:210-223.

[5] Canadian Cancer Society. (2013)."Non-Hodgkin lymphoma statistics". Available from: `http://www.cancer.ca`. Accessed May 2013.

[6] Colt JS, Rothman N, Severson RK et al. (2009). "Organochlorine exposure, immune gene variation, and risk of non-Hodgkin lymphoma", *Blood*, 113(9), 18991905.

[7] ENCODE Project Consortium, Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E et al. (2011). "A user's guide to the encyclopedia of DNA elements (ENCODE)", *PLoS Biol*,9(4):e1001046. Epub 2011 Apr 19. PMID: 21526222; PMCID: PMC3079585

[8] Foulkes, A.S. (2009)." Applied Statistical Genetics with R: For Population Based Association Studies". Springer Science: New York.

[9] Gauderman, WJ. (2002). "Sample size requirements for matched case-control studies of gene-environment interaction", *Stat Med*, 21:35-50.

[10] Green, P. J. (1984). "Iteratively Reweighted Least-Squares for Maximum Likelihood Estimation and Some Robust and Resistant Alternatives" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 46, 149-192.

[11] Hauck WW, Dormer A. (1977). "Walds test as applied to hypotheses in logit analysis". *J Am Stat Assoc*, 72: 851-853.

[12] He, C, et al.“ A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis”. (2009). *PLoS One*, 4(3):e4732

[13] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. (2002). “The human genome browser at UCSC” *Genome Res*, 12(6):996-1006.

[14] Kramer, Shira; Hikel, Stephanie Moller; Adams, Kristen; Hinds, David; Moon, Katherine. (2012). “Current Status of the Epidemiologic Evidence Linking Polychlorinated Biphenyls and Non-Hodgkin Lymphoma, and the Role of Immune Dysregulation”. *Environmental Health Perspectives*, 120 (8): 106775.

[15] Liu, Jie (2012). “A global test of association between Non-hodgkin lymphoma and SNPs in Histone-pathway genes”, (MSc Thesis, Department of Statistics and Actuarial Science, *Simon Fraser University*).

[16] Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, and Kent WJ. (2012). “The UCSC Genome Browser database: extensions and updates 2013”, *Nucleic Acids Res*, [Epub ahead of print]

[17] National Cancer Institute (2013). “Non-Hodgkin Lymphoma”. Available from: `http://www.cancer.gov/cancertopics/types/non-hodgkin`. Accessed April 2013.

[18] Piegorsch WW, Weinberg CR, Taylor J. (1994). “Non hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies” *Stat in Med*,1994;13:153-162.

[19] Prentice, R., and R. Pyke. (1979). “Logistic disease incidence models and case-control studies”, *Biometrika*,66: 403-412.

[20] Qu, Conghui (2009), “Multiple hypothesis testing procedures with applications to epidemiologic studies”, (MSc Thesis, Department of Statistics and Actuarial Science, *Simon Fraser University*).

[21] Renshaw, J., Orr, R. M., Walton, M. I., Te Poele, R., Williams, R. D., Wancewicz, E.V., et al. (2004). “Disruption ofWT1gene expression and exon 5 splicing following cytotoxic drug treatment: Antisense down-regulation of exon 5 alters target gene expression and inhibits cell survival”, *Molecular Cancer Therapeutics*, 3(11), 1467-1484.

[22] Sampath, J., Long, P. R., Shepard, R. L., Xia, X., Devanarayan, V., Sandusky, G. E., et al. (2003). “Human SPF45, a splicing factor, has limited expression in normal tissues, is overexpressed in many tumors, and can confer a multidrug-resistant phenotype to cells”, *The American Journal of Pathology*, 163(5), 17811790.

[23] Smith, P. G.and Day, N.E. (1984). "The design of case-control studies: The influence of confounding and interaction effects", *Int Journal of Epidemiology*, 3(3), 356-364.

[24] Spinelli, J. J., Ng, C. H., Weber, J.-P., Connors, J. M., Gascoyne, R. D., Lai, A. S., Brooks-Wilson, A. R., Le, N. D., Berry, B. R., & Gallagher, R. P. (2007). "Organochlorines and risk of non-hodgkin lymphoma", *Int J Cancer*, 121(12), 2767-2775.

[25] US National Cancer Institute. (2013). "Surveillance Epidemiology and End Results database", Available from: `http://www.http://seer.cancer.gov`. Accessed May 2013.

[26] Vegran, F., Boidot, R., Oudin, C., Riedinger, J. M., Bonnetain, F., & Lizard-Nacol, S. (2006). "Overexpression of caspase-3s splice variant in locally advanced breast carcinoma is associated with poor response to neoadjuvant chemotherapy", *Clinical Cancer Research*, 12(19), 57945800.

[27] Walter W. Hauck, Jr. , Donner A. (1977). "Wald's Test as Applied to Hypotheses in Logit Analysis", *Journal of the American Statistical Association*, Vol. 72, No. 360, pp. 851-853.

[28] Wang S, Nieters A. (2010). "Unraveling the interactions between environmental factors and genetic polymorphisms in non-Hodgkin lymphoma risk" *Expert Reviews: Anticancer Therapy*. 10(3), 403-413.

[29] Weizmann Institute of Science. (2013). "GeneCards, the human gene compendium", Avaliable from: `http://www.genecards.org`. Accessed May 2013.

[30] Yoo Wonsuk, Ference Brian A., Cote Michele L, and Schwartz Ann. (2012) "A Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions". *Int J Appl Sci Technol*. 2(7): 268.

[31] Zheng G, Yang Y, Zhu X, and Elston RC (2012). "Analysis of Genetic Association Studies", Springer, New York (414).