Comparison of Statistical Methods of Haplotype Reconstruction and Logistic Regression for Association Studies

by

Karey Shumansky

B.Sc., University of Alberta, 2003.

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department
of
Statistics and Actuarial Sciences

SIMON FRASER UNIVERSITY
Summer 2005

APPROVAL

Name: Karey Shumansky

Degree: Master of Science

Title of project: Comparison of Statistical Methods of Haplotype Reconstruction and Logistic Regression for Association Studies

Examining Committee: Dr. R. Lockhart
Chair

____

Dr. J. Graham
Senior Supervisor
Simon Fraser University

____

Dr. J. Spinelli
Supervisory Committee
BC Cancer Agency and Simon Fraser University

____

Dr. B. McNeney
External Examiner
Simon Fraser University

Date of Defense: June 23rd, 2005

Abstract

Investigating association between disease and single nucleotide polymorphisms (SNPs) has been an approach for genetic association studies and more recently investigating association between disease and haplotypes has become another accepted method. Haplotypes are physically linked combinations of alleles from a stretch of DNA and can serve to increase power of finding an association due to interactions between inclusive SNPs and the increased area of chromosome that is taken into consideration.

Determining haplotypes experimentally or by family studies is a costly and time-inefficient method, so haplotype reconstruction by statistical methods has become an adopted practice. The problem with computational methods is the extra source of error from ambiguous haplotypes that has to be included in statistical analysis. This paper investigates methods of error management with three different logistic regression packages, two of which are specific to analysis of genetic data. Methods are applied to simulated data and a data set looking for genetic risk factors for non-Hodgkin Lymphoma.

Dedication

I dedicate this to my parents who always supported and encouraged me in so many ways along this long road.

Acknowledgements

Special thanks to my supervisor Dr. John Spinelli for all the helpful advice along the way, giving me the great experience of working at the BC Cancer Agency, sticking with me when the going got tough in a few spots and endlessly rereading my project. Thanks to Dr. Jinko Graham and Dr. Brad McNeney for offering a great course in analysis of genetic data which increased my knowledge in the area of genetics, and for answering my many questions. Thanks to the great group of students who I had the opportunity to share K9501 with during my time at SFU, especially Maria and Amy, I don't know if I could have done it without you guys. Thanks to Mike for being there and surviving the last 2 weeks.

Contents

Approval ii

Abstract iii

Dedication iv

Acknowledgements v

Contents vi

List of Tables viii

1 Introduction 1

2 Non-Hodgkin Lymphoma (NHL) Study 4

2.1 Non-Hodgkin Lymphoma 4

2.2 Candidate Genes 6

2.2.1 Thl/Th2 cytokines 7

2.2.2 Break Repair and Checkpoint Genes 7

2.3 NHL Study Data 8

2.4 Methods 9

3 Haplotype Reconstruction 11

3.1 Genetic Background 11

3.2 Testing Hardy-Weinberg Equilibrium 12

3.3 Haplotypes 15

3.4 Haplotype reconstruction 16

3.5 Reconstruction and Logistic Regression Methods under Investigation . . . 18
3.5.1 PHASE Reconstruction and Logistic Regression . . . 18
3.5.2 Hapassoc Package for the R Programming Environment 21
3.5.3 Haplo.stats Package for the R Programming Environment . . . 24
3.6 Statistical Issues with Genotype Analysis . . . 26
4 Analysis of the BC Cancer Agency NHL Data . . . 29
4.1 Tests of HWE . . . 29
4.2 SNP Analysis . . . 31
4.2.1 Penetrance Analysis . . . 31
4.3 Haplotype Analysis . . . 36
5 Simulation Study . . . 48
5.1 Haplotype Data Generation . . . 48
5.2 Analysis of Simulated Data . . . 50
6 Summary and Conclusion . . . 61
6.1 Non-Hodgkin Lymphoma Data Analysis . . . 61
6.2 Simulation . . . 63
6.3 Conclusions . . . 64

List of Tables

2.1 Table of Candidate Genes. 10

4.1 Tests of Hardy-Weinberg Equilibrium 30

4.2 Univariate SNP analysis 32

4.3 Multivariate SNP analysis 33

4.4 Table of significant SNP penetrances 34

4.5 Table of output from BC Cancer Agency SNP data, H2AX, 3 loci . . 38

4.6 Table of output from BC Cancer Agency SNP data, NBS1 39

4.7 Table of output from BC Cancer Agency SNP data, IL10 40

4.8 Table of output from BC Cancer Agency SNP data, H2AX, 2 loci . . 41

5.1 Table of Haplotype Frequencies. 49

5.2 Initial Estimates of Logistic Regression Coefficients, 5 Covariates . . . 51

5.3 PHASE LR of 50,000 individuals, 5 Covariates 52

5.4 Haplo.stats LR of 50,000 individuals, 5 Covariates 53

5.5 PHASE 100 LR, 500 individuals, 5 Covariates 54

5.6 Hapassoc 100 LR, 500 individuals, 5 Covariates 55

5.7 Haplo.stats 100 LR, 500 individuals, 5 Covariates 56

5.8 PHASE Reconstruction and Regression Output for Pooled Haplotype Variable, with Outliers Excluded 58

Chapter 1

Introduction

Single nucleotide polymorphisms (SNPs) are the most common form of DNA sequence variation and account for 90% of human variation. They are useful genetic markers to investigate genes related to susceptibility to diseases or genes related to drug responsiveness. Scientists use SNPs in association studies and more recently have been using SNPs in combination along regions of chromosomes that have not been broken up by recombination. SNPs in combination along a stretch of a chromosome are called haplotypes. Haplotypes can potentially increase the power of an association because they take a larger area of the chromosome into consideration and can be used to investigate SNP interaction. Using methods such as family studies and direct physical determination of alleles on a gene to determine haplotypes can be a costly and time-consuming process, so statistical methods have gained popularity in the reconstruction of haplotypes for use in association studies.

There are many different statistical methods for haplotype reconstruction, including a parsimony algorithm, an expectation-maximization (EM) algorithm and a Bayesian approach that uses coalescent theory to estimate haplotypes. After haplotypes have been reconstructed from genotype data, a generalized linear model can be used to analyze the haplotype data in an association analysis. However, due to extra ambiguity incurred by using statistical methods to assign haplotypes additional error should be taken into account by inflating the standard errors of regression coefficients obtained from logistic regression. The ambiguity of the haplotype reconstruction can also be taken into account by doing a weighted logistic regression on haplotype data properly weighted according to how likely the haplotype assignment would be, given the genotype data.

This project investigated the effect on estimates and significance testing when using different methods of haplotype reconstruction and regression. Statistical computing packages that do an iterative weighted logistic regression by method of weights adequately inflating standard errors of regression coefficients obtained by regressing reconstructed haplotypes were compared to weighted logistic regression that doesn't inflate standard errors. Also, reconstructed haplotype data was analyzed using computing programs that do a weighted logistic regression compared to regular logistic regression. Three haplotype reconstruction packages were compared.

Chapter 2 starts with an introduction to non-Hodgkin lymphoma (NHL) and outlines some of the causes. Section 2.2 describes how and why the SNPs used in analysis were chosen, as well as some background on the SNPs. Section 2.3 is a brief description of the BC Cancer Agency NHL study data and how it was obtained. 2.4 is a short outline of the methods of analysis that are used.

Chapter 3 begins with some genetic background that will explain genetic terminology and section 3.2 is a background of the use of the Hardy-Weinberg model to test Hardy-Weinberg Equilibrium (HWE) for individual SNPs. Section 3.3 explains haplotype use in association studies and section 3.4 outlines some statistical haplotype reconstruction methods. Section 3.5 enumerates haplotype reconstruction and logistic regression methods being used in the project and a brief sketch of the statistical theory used by each one. To conclude the chapter, statistical issues regarding haplotype reconstruction will be outlined in section 3.6.

Chapter 4 is an analysis of the BC Cancer Agency NHL data. Section 4.1 tests Hardy-Weinberg equilibrium for SNPs in the BC Cancer Agency data set. The next section, 4.2, is a univariate and multivariate case control analysis of the individual SNPs, followed by section 4.3, analysis of combinations of SNPs using three methods of haplotype reconstruction and seven methods of logistic regression. The methods of haplotype reconstruction and logistic regression are contrasted with and without adjustment for non-SNP variables age group, sex, ethnicity and region of residence.

Chapter 5 describes a simulation study used to investigate differences in estimates and standard errors of regression coefficients obtained from logistic regression of reconstructed haplotypes. Section 5.1 gives a description of the generation of haplotype data. A haplotype is chosen to be the "affected" haplotype and outcome data is generated for each individual based on the number of "affected" haplotypes they carry. To compare regression coefficients from reconstructed haplotypes, in section 5.2 initial estimates for original "known" haplotype data is obtained using logistic regression, followed by a complete investigation of the three reconstruction and seven logistic regression methods.

Finally, Chapter 6 is a summary of the preceding BCCA and simulated data analysis sections with some conclusions.

Chapter 2

Non-Hodgkin Lymphoma (NHL) Study

2.1 Non-Hodgkin Lymphoma

The lymphatic system is the body's inner immune system that helps filter out infection and disease using a network of tube-like vessels that branch into tissues throughout the body. It is where certain white blood cells and antibodies are produced, and it is also important for the distribution of fluids and nutrients in the body. Along the network of vessels are lymph nodes, small pea-sized organs grouped along the route of large blood vessels in the neck, underarms, groin, abdomen and pelvis. Other parts of the lymphatic system are in the spleen, bone marrow and tonsils, lymphatic tissue is also found in stomach and skin.

Lymphomas are cancers of the lymphatic system. They arise when white blood cells become cancerous, dividing out of control, not undergoing normal cell death. They accumulate, crowding out other functioning white blood cells and other nearby normal cells within affected organs. If cells keep dividing, when not needed, this can create an extra mass of tissue that can turn into a tumor.

There are two groups of lymphomas: Hodgkin lymphoma and non-Hodgkin lymphoma (NHL). Hodgkin lymphoma accounts for less than 1% of all new cancers diagnosed in the USA, it is more common in males than females, and has an increase in incidence for young adults from their teens and peaking around age 25 and those over 55 years of age. Age-adjusted survival rates have increased in the period 1986-90 and can be attributed to better available treatments [28]. Non-Hodgkin lymphoma is less predictable and is more likely to spread beyond the lymphatic system to other parts of the body. It is a heterogeneous disease, which has many different subtypes and disease entities, numbering around 30. NHL subtypes can be grouped into low grade, intermediate grade, high grade and miscellaneous lymphomas [30], with the chance of survival depending partially on grade and stage of cancer. Low grade or indolent NHL can be hard to treat because some forms, such as follicular lymphoma, which include approximately 70% of indolent lymphomas, tend to be resistant to treatments that induce cell death. Patients with low grade or indolent lymphomas can live many years because the cancers can be very slow growing. Intermediate and high grade NHL are more aggressive forms of lymphoma but are more likely to be cured with chemotherapy. Also, the more aggressive the lymphoma, the more frequently it is localized and depending on the site of the disease, can be more easily treated.

There are many factors that have been associated with an increased risk of NHL. These range from factors such as inherited conditions causing immunodeficiency, therapies that artificially suppress the immune system following transplant surgeries to viruses such as HIV, which acts to depress the immune system, and a herpes virus called the Epstein Barr Virus [33]. A link between NHL and chemicals like pesticides [17], solvents [39], and those in hair dyes [5] has been suggested. The rate of NHL increases with age and it is more common in men than women. NHL occurs in different ethnicities with differing rates; in an Israeli study Jews were found to have a higher incidence rate of NHL relative to non-Jews [29]. Since the 70's NHL has been on the rise in Canada [31] and other developed countries [6][20], with increases seen in all age groups [14], in men and women, and in different ethnicities [34]. Some of this increase has been linked to the rise of AIDS and better diagnosis of the elderly where rate increases have been the largest [13], but it cannot all be explained by these factors [14].

It has been shown that some patients with NHL have family histories revealing that blood relatives have similar types of immunodeficiency disorders more than one would expect by coincidence [7]. Given that NHL risk is associated with conditions that alter the immune system, primary immunodeficiency diseases, acquired immunodeficiency diseases, autoimmune diseases, and patients immunosuppressed following transplantation [2], it is a logical step to investigate genes coding for Cytokines, a group of secreted proteins that mediate immune reactions by influencing the growth and differentiation of lymphocytes [1], making them prime candidates for genetic susceptibility. Cytokines comprise Interleukins, which induce fever and inflammation and activation, differentiation and proliferation of B and T cells, Lymphokines, that act as chemical messengers activating immune reactions, Monokines, that mediate immune responses, and Chemokines, that activate and attract leukocytes to infection sites, playing a major role in acute inflammation.

2.2 Candidate Genes

Immunological candidate genes were chosen from different pathways and genotyped; Th1 and Th2 cytokines, DNA break repair histones and S-PHASE checkpoint/DNA crosslink response genes. The Th1 and Th2 cytokine SNPs chosen were based on previous studies and have an allele frequency greater than 5% in the general population. The break repair and checkpoint SNPs were chosen using experimental means of SNP discovery, sequencing individuals, concentrating on coding regions of the gene and extracting polymorphic regions if they occurred with a high enough frequency. The SNPs that were finally chosen had less than 5% missing data when all individuals were genotyped, so that accurate analysis could be done on the SNP data.

2.2.1 Thl/Th2 cytokines

Helper T (Thl/Th2) cells have two important functions: to stimulate cellular immunity and inflammation, and to stimulate B cells to produce antibodies. These two functionally distinct groups of cytokines promote different activities, regulating the activities of the other. IL-4, IL-10, IL-4RA cytokines regulate Thl/Th2 responses. IL-10 stimulates antibody production by B cells and inhibits both production of proinflammatory cytokines and assisting functions of macrophages in T cell activation.

The immunological gene IL-10 was chosen for genotyping at the BCCA since the gene is involved in immunity, and immune suppression for a variety of reasons (congenital or suppression for transplants) has been found to be a risk factor in NHL. (SNP background information from discussion with Dr. J. Spinelli).

2.2.2 Break Repair and Checkpoint Genes

NBS1 is a gene responsible for the Nijmegen breakage syndrome (NBS), a rare autosomal recessive condition of chromosomal instability and homozygosity of the major mutation of the NBS1 gene has been linked to a number of disorders such as microcephaly, immunodeficiency, congenital heart disease and chromosomal instability. NBS shares a number of features with ataxia-telangiectasia (AT), a degenerative disease in the brain that leads to a lack of muscle control, immunodeficiency and a predisposition to malignancies of the blood system such as lymphoma and leukemia. Functional interactions between ataxia-telangiectasia mutated and NBS1 genes were studied [24] and observations linked ATM and NBS1 in a common signaling pathway, explaining the similarities between AT and NBS and providing the idea to study the NBS1 gene for association with NHL.

Chen et al. [9] reported that the NBS1 protein and histone gamma-H2AX, which associate with irradiation-induced DNA double strand breaks (DSBs), are also found at sites of variable, diversity, joining recombination-induced DSBs. Conclusions of a study suggested that surveillance of T-cell receptor recombination intermediates by NBS1 and gamma-H2AX may be important for preventing translocations that contribute to cancer formation. Another study of mice that were homozygous for an H2AX deletion were radiation sensitive, growth retarded, and immune deficient, and mutant males were infertile [8]. These phenotypes were associated with chromosomal instability, repair defects, and impaired recruitment of NBS1. The conclusion was that H2AX is critical for facilitating the assembly of specific DNA-repair complexes on damaged DNA.

2.3 NHL Study Data

The data being used in this analysis was obtained as part of a case control study of non-Hodgkin lymphoma. All non-Hodgkin lymphoma cases age 20-79 diagnosed during the period March 2000-Febraury 2004 and living in the Greater Vancouver Regional District (GVRD) and the Capitol Regional District (CRD; Greater Victoria) were ascertained from the British Columbia Cancer Registry. Each case was contacted by letter and requested to participate in the study, potential subjects who had not replied within a certain time frame after the initial contact letter were telephoned and asked if they would be willing to participate. Subjects taking part in the study were asked to complete a phone interview and provide either a blood or mouthwash sample. Exclusions included those subjects not able to give informed consent or complete the questionnaire, due to language, illness or death.

The control data used in the analysis was collected from the Client Registry of the BC Ministry of Health. The Registry includes almost all (98%) of residents of BC as it is the central list of subscribers to the provincial health plan. Exclusions were primarily people who had lived in the province for less than 3 months. The control subjects were chosen randomly and were frequency matched to the NHL case subjects by age (within 5-year age group), sex and region of residence (GVRD or CRD). The control subjects were also asked to complete a telephone interview and provide either a blood or mouthwash sample.

Control subjects were frequency matched on age, sex and region of residence, so these variables were adjusted for in the case-control analyses. Since the incidence of NHL varies with respect to ethnic origin, a variable for ethnic group was also used as an adjustment variable in any analysis. The variability of NHL incidence in different ethnic groups was accounted for because there was a risk of it being a confounding factor due to genetic variability between the groups.

H2AX, NBS1 and IL10 were analyzed for association analysis with NHL. H2AX gene has three SNPs of interest that were used for association analysis as SNPs and reconstructed as haplotypes. The NBS1 gene has five SNPs of interest but two of the SNPs are in high linkage disequilibrium so only one of those two were included in analysis, resulting in four SNPs that were analyzed as SNPs and reconstructed as haplotypes for association analysis. IL10 gene has two SNPs of interest that were reconstructed as haplotypes for analysis.

Table 2.1 shows a table of the three candidate genes and the SNPs that were investigated for association with NHL. For the first NBS1 SNP, the possible alleles listed in the table are WT and del(WT). WT stands for "wild type", which is the allele found in the majority of the wild population, usually the normally functioning allele. The first NBS1 SNP has a wild type that corresponds to an allele sequence of AGTA, del(WT) stands for a deletion of the standard allele where the genetic information ATGA is missing in the particular spot on the chromosome. The deletion mutations are slightly different than SNPs but can be analyzed the same way as a SNP.

2.4 Methods

All subjects in the study were genotyped using a TaqMan fluorogenic 5' nuclease assay, validated by an assay design, optimization and validation service, then genotyped using Polymerase Chain Reaction (PCR) machines followed by reading of fluorescent products in a PCR instrument that detects and quantitates nucleic acid sequences.

Pathway	Gene	Allele	Genotype Frequency
Th1/Th2 Cytokines	IL10	A(1082)G	A/A=0.422, A/G=0.458, G/G=0.12
		T(-3575)A	T/T=0.348, T/A=0.489, A/A=0.162
DNA repair Histones	H2AX	G(-417)A	A/A=0.263, A/G=0.468, G/G=0.269
		C(1057)T	C/C=0.389, C/T=0.440, T/T=0.171
		A(1528)G	A/A=0.916, A/G=0.075, G/G=0.008
S-PHASE checkpoint/ DNA crosslink responses	NBS1	WT(-352)del(WT)	del/del=0.004, WT/del=0.081, WT/WT=0.915
		G(102)A	A/A =0.112, A/G =0.446, G/G =0.442
		G(553)C	C/C =0.115, G/C=0.473, G/G=0.412
		A(2016)G	A/A=0.423, A/G=0.474, G/G=0.103

Table 2.1: Table of Candidate Genes.

Chapter 3

Haplotype Reconstruction

3.1 Genetic Background

Critical to the understanding of the genetic basis for complex diseases is the modeling of human variation. The vast majority (about 99.9%) of genetic sequences are identical and the remaining 0.1% of variation can be characterized by single nucleotide polymorphisms (SNPs), which are mutations at a single nucleotide position. Single nucleotide polymorphisms occur when a single nucleotide in a genetic sequence is altered (e.g. A replaces T or C replaces G), such as genetic sequence ATTA being altered to AATA. To be considered a SNP, the alteration must occur in at least 1% of the population. SNPs make up 90% of the genetic variation in the human population, occurring in coding and non-coding regions of the genome. SNPs occurring in coding sequences are of interest because researchers believe that some of these genetic variations have protective or susceptibility implications for cancer and other diseases, as well as for response to therapeutic drugs. Even if a SNP isn't directly responsible for a disease or response to treatment, it is possible to find genes that influence such traits using a nearby or closely-linked SNP. They are relatively stable genetically and they may be used as markers for harmful or positive mutations. SNP markers can help unearth mutations and accelerate efforts to find therapeutic drugs. An important tool in the study of SNPs is an association study to investigate the extent to which a mutation is associated with the occurrence of disease.

3.2 Testing Hardy-Weinberg Equilibrium

In association studies it is important to check for association between two alleles at a SNP locus because association may indicate a population substructure, biased sampling of individuals or genotyping error, any of which would render a positive association in further analysis false. A preliminary check of population equilibrium for the individuals in a study would show whether there was a chance of some error in sampling or genotyping or even some substructure to the data.

The Hardy-Weinberg equilibrium model [18][38] describes and predicts genotype and allele frequencies in a non-evolving population. The model has some basic assumptions, specifically:

- the population is large
- there is no gene flow between populations
- mutations are negligible
- individuals are mating randomly
- natural selection is not operating on the population.

Given these assumptions, a population's genotype and allele frequencies will remain unchanged over successive generations, and the population is said to be in Hardy-Weinberg equilibrium. The Hardy-Weinberg model equations can be applied to the genotype frequency of a single locus.

As an example, say that we have a diallelic locus with alleles A and a, A signifies the dominant allele and a is the recessive allele. If allele frequencies for the locus are

p (frequency of a dominant allele A) and q (frequency of a recessive a), then for the whole population we would have:

$$p + q = 1.$$

Using Mendelian theory the homozygous genotype AA, heterozygous genotype Aa and the homozygous genotype aa would have proportions $p^2 : 2pq : q^2$, which can be derived using a Punnett square (used in simple mating examples to calculate proportions of offspring genotypes). This can be expressed for the population as:

$$p^2 + 2pq + q^2 = 1.$$

The Hardy-Weinberg equations enable us to compare a population's actual genetic structure over time with the genetic structure we would expect if the population were in Hardy-Weinberg equilibrium (i.e., not evolving). If genotype frequencies differ from those we would expect under equilibrium, it may be assumed that the alleles within individuals are associated. This could prompt an investigator to check if there is an error in the data and if the error could not be fixed, association analysis may not be done on alleles of a locus not under HWE.

It is important to test for Hardy-Weinberg equilibrium (HWE) in a sample of genetic data so there will be a level of confidence in the association analysis of the chosen loci. For large samples, a test of HWE is a chi-square goodness of fit test, but sometimes with genetic data even if the sample size is large an allele may have a small expected count which can lead to misleading results. The large sample χ^2 test statistic is as follows, with O=observed counts and E=expected counts for the k=3 genotypes for a diallelic locus:

$$\chi^2 = \sum_{i}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Let $\tilde{P}_{MM}$ be the observed probability of a homozygous major allele genotype, $\tilde{P}_{Mm}$ is the observed probability of a heterozygous genotype, $\tilde{P}_{mm}$ the observed probability of an observed homozygous minor allele gentotype, $\tilde{p}_M$ is the observed probability of the major allele frequency and $\tilde{p}_m$ the observed probability of the minor allele frequency for a data set. Explicitly, a large sample χ^2 Hardy-Weinberg test statistic to test the null hypothesis $H_o = HWE$ is:

$$\chi^2_{HW} = N\left[\frac{(\tilde{P}_{MM}-\tilde{p}_M^2)^2}{\tilde{p}_M^2}+\frac{(\tilde{P}_{Mm}-2\tilde{p}_M\tilde{p}_m)^2}{2\tilde{p}_M\tilde{p}_m}+\frac{(\tilde{P}_{mm}-\tilde{p}_m^2)^2}{\tilde{p}_m^2}\right].$$

Another tool for testing HWE is a permutation test that evaluates χ^2 for all possible sets of genotypic counts consistent with the observed allelic counts in the data set. Hardy-Weinberg disequilibrium statistics D, D' and r are computed, a bootstrap confidence interval is computed for the statistics, then a p-value for the permutation test is found by calculating the proportion of χ^2 values that are as large as or larger than the χ^2 observed value. Using bootstrapping for the confidence interval and simulation for the p-value avoids reliance on the assumptions of the χ^2 approximation. This is important when some allele pairs have small counts because they won't fit with the χ^2 large sample assumption, making the χ^2 test an incorrect approach to testing goodness-of-fit.

The disequilibrium statistics are D, which is defined as the half of the raw difference in frequency between the observed number of heterozygotes and the expected number, D' which is defined as D rescaled to span the range [-1,1] (D/D_{MAX}) and r which is the correlation coefficient between two alleles.

For a rare disease the HWE test may be performed on control data, because the control data approximates a sample from the general population and should fit the Hardy-Weinberg model. Case data has been specifically chosen for a reason such as having a certain disease and Hardy-Weinberg equilibrium cannot be assumed in this population. Admixture of the two populations (i.e. doing a test of HWE on the whole population) could result in a type I error when the HWE test is carried out.

3.3 Haplotypes

By studying stretches of DNA where a SNP or many SNPs in combination mark a harmful mutation, researchers may locate disease-causing genes. SNPs in combination along a stretch of the chromosome are called haplotypes. Theoretically, there could be many combinations of SNPs in a haplotype. If there are 10 SNPs in a haplotype, there are 2^{10} possible haplotypes associated with these SNPs but only a few of these haplotypes will be frequently-occurring enough to warrant inclusion in analysis as a separate variable. It is common practice to combine subjects with rare haplotypes into a pooled category, since there are typically not enough rare haplotype individuals in a study to investigate the association of a rare haplotype with a disease or drug response.

Haplotypes are also relatively stable genetically, occurring in genetic sequences that are the same in many individuals. These sequences are in sections of the chromosome that haven't been shuffled by genetic recombination, and are separated by sections that have been altered by genetic recombination. Many dozen kilobases long, haplotype blocks make up greater than 65% of the human genome.

Gene mutations on the same haplotype block marked by SNPs can interact with one another, producing effects in combination that would be difficult to evaluate by looking at one SNP marker at a time. A genetic mutation may also be located in or near the stretch of chromosome that is marked by the SNPs defining the haplotype. Essentially, examination of haplotypes versus single SNPs potentially increase the power of finding an association because of interactions between SNPs and the increased area of a chromosome that is taken into consideration when more SNPs are included in haplotyping.

If there is no family information available for an individual who has been genotyped at certain loci, it is not possible to correctly recreate the haplotype unless the individual is homozygous for all loci or all less-one loci. Since DNA is divided into two strands, the alleles for each locus that has been genotyped can be arranged in two different ways, also with more heterozygous SNPs there are more possible haplotypes.

There are many available methods of haplotype determination including genotyping relatives of each individual included in a study and using the relationship data, if it is known, to determine the phase of the markers and the haplotype; direct physical determination of which allele is on the same DNA molecule as another using various processes; and by means of a statistical method used to infer phase at linked loci from genotypes and thus reconstruct haplotypes. Genotyping relatives of each individual and direct physical determination of alleles are time-consuming and costly processes to determine haplotypes for all individuals, making statistical methods a superior choice for investigators with time and budget constraints.

3.4 Haplotype reconstruction

Computing algorithms to construct haplotypes include maximum likelihood using a parsimony algorithm created by Clark [11], Bayesian methods that uses a priori expectations to estimate haplotypes [37][32], and an expectation-maximization (EM) algorithm [25][15][19].

The parsimony method by Clark [11] is an algorithm that infers haplotypes from samples of genotyped individuals. It starts by identifying all genotypes that are homozygotes or single-site heterozygotes, and then determining whether any of the ambiguous ($>$ 1 heterozygous site) haplotypes could be explained by the already-resolved haplotypes (if not, then stop; otherwise continue). Each time a previously-observed haplotype is identified as one of the possible haplotypes in an ambiguous case, the complementary haplotype is added to the list of previously observed haplotypes. The algorithm keeps running until as many genotypes are determined as possible. This method is done many times on different orderings of the data. Drawbacks of the method are that it is possible that the algorithm won't start if there are no homozygotes or single-site heterozygotes and that it depends on the ordering of the data. The method would also have to be modified to allow for the case of missing data due to individuals who had loci that weren't able to be genotyped.

Stephens et al. [37] describe a Bayesian method to evaluate the conditional distribution of haplotypes, given genotype data. The method employs Gibbs sampling (a Markov chain-Monte Carlo algorithm) to create a sample from the posterior distribution of haplotypes, given genotypes. The algorithm is given a starting value of $H^{(0)}$ for H, the set of haplotype pairs corresponding to G, the set of genotype data. An individual is repeatedly chosen at random and haplotypes are estimated under the assumption that all other haplotypes are correctly estimated, this process is repeated enough times to obtain an approximate sample for $Pr(H|G)$. This model is difficult to apply in theory but results in a simpler algorithm that gives similar output as an EM algorithm. This method can create both a most probable haplotype reconstruction that assigns each person the most probable haplotype given their genotype information or a haplotype reconstruction that assigns each person all possible haplotypes given their genotype and the associated weights for each haplotype.

The EM algorithm is an iterative optimization method to estimate some unknown parameters, given some known data. It is a method of finding maximum likelihood estimates of model parameters that may not be obtained easily by conventional means. The EM algorithm can be used to estimate population haplotype probabilities via maximum likelihood estimation; finding the values of the haplotype probabilities which optimize the probability of the observed data [25]. The maximum likelihood estimates of the haplotype probabilities are obtained by maximization of the likelihood. The log-likelihood of the haplotype model is

$$lnL = \sum_{i=1}^{N} lnPr(P_i)$$

where $Pr(P_i)$ is the probability of the i^{th} person's phenotype (i.e. unphased genetic data). $Pr(P_i)$ is calculated by summing up the probabilities of all genotypes (i.e. haplotype pairs) that can express the phenotype, based on the assumption of Hardy Weinberg equilibrium for a haplotype:

$$P(h_k h_l) = \begin{cases} p_k^2 & \text{if } k{=}l, \\ 2p_k p_l & \text{if } k \neq l \end{cases}$$

The expectation step calculates the expected numbers of copies that an individual contributes to the overall expected count of haplotypes given by their phenotype P_i. The expectation looks like:

$$E[n_{abc}|P_i] = \frac{2f_{abc}\sum_{H_{a^*b^*c^*}} f_{a^*b^*c^*}}{Pr(P_i)}$$

f_{abc} is the frequency of a three-locus haplotype H_{abc}, and $f_{a^*b^*c^*}$ the frequency of another haplotype $H_{a^*b^*c^*}$, that can combine with H_{abc} to form P_i. The total expected number of each haplotype in the data set after each iteration is taken over 2×(the number of individuals in the data set) to update the haplotype probabilities. The maximization step updates the haplotype frequencies until the log-likelihood stabilizes. A simple haplotype assignment can be done by choosing the most probable haplotype assignment given genotype data and haplotype probabilities obtained from the EM algorithm.

Hapassoc [3][4][12] and Haplo.stats [23][12] haplotype reconstruction and logistic regression packages use an extension of the maximum likelihood estimation of haplotype frequencies. Both packages implement an EM-based logistic regression for binary response. The maximum likelihood approach is used in jointly estimating the haplotype and non-SNP risk parameters and the haplotype frequencies on the basis of case/control status, non-SNP variables and diallelic SNP data.

3.5 Reconstruction and Logistic Regression Methods under Investigation

3.5.1 PHASE Reconstruction and Logistic Regression

The PHASE program [37] is an implementation of the Bayesian method of haplotype reconstruction, allowing use of a priori expectations to correctly assign haplotypes to individuals in a dataset. Starting with a sample of n diploid (receiving a chromosome from each parent) individuals from a population, we have known genotypes $G = (G_1, \cdots, G_n)$, corresponding unknown haplotype pairs $H = (H_1, \cdots, H_N)$, a set of unknown population haplotype frequencies $F = (F_1, \cdots, F_M)$ and a set of unknown sample haplotype frequencies $f = (f_1, \cdots, f_M)$, M denoting the number of possible haplotypes for the sample. The M possible haplotypes are arbitrarily labeled $1, \cdots, M$. The PHASE method regards the unknown haplotypes as unobserved random quantities and aims to evaluate their conditional distribution given the information that can be obtained from the known genotype data. PHASE uses a Gibb's sampling algorithm, a type of Markov chain-Monte Carlo (MCMC) algorithm to obtain an approximate sample from the posterior distribution of $Pr(H|G)$.

We start with an initial guess for the resolved haplotype information for all individuals, $H^{(0)}$, maybe just the known genotype information arranged into haplotype form, for example. We want to obtain $H^{(t+1)}$ from $H^{(t)}$ for $t = 0, 1, 2, \cdots$. The algorithm is as follows:

1. An individual is chosen at random from all ambiguous individuals (those individuals who have more than one possible haplotype, given their genotype)

2. A subset S of the ambiguous (heterozygous) loci from individual i is chosen to be updated. Let H(S) denote the haplotype information for the individual i at the ambiguous loci S and let $H(-S)$ denote the complement of $H(S)$, the haplotype information from all other individuals as well as the homozygous locus information within the i^{th} individual. Sample $H^{(t+1)}(S)$ from $Pr[H(S)|G, H^{(t)}(-S)]$.

3. Set $H^{t+1}(-S) = H^{(t)}(-S)$.

At each iteration it is necessary to update $Pr[H(S)|G, H^{(t)}(-S)]$. For $H(S)$ consistent with genotype information G, the conditional distribution is:

$$\begin{aligned} Pr[H(S)|G, H^{(t)}(-S)] &\propto Pr(H_i|H_{-i}) && (3.1)\\ &\propto \pi(h_{i1}|H_{-i})\pi(h_{i2}|H_{-i}, h_{i1}). && (3.2) \end{aligned}$$

This is equivalent to the conditional distribution for the haplotype pair $H_i = (h_{i1}, h_{12})$, consistent with genotypes G_i where $\pi(\cdot|H)$ is the conditional distribution of a future-sampled haplotype, given a set H of previously sampled haplotypes. H_{-i} is the set of haplotypes excluding individual i. Stephens and Donnelly [36] suggest an approximation to the unknown $\pi(\cdot|H)$ as

$$\pi(h|H) \;=\; \sum_{\propto \epsilon E} \sum_{s=0}^{\infty} \frac{r_{\propto}}{r} \left(\frac{\theta}{r+\theta} \right)^{s} \frac{r}{r+\theta} (P^{s})_{\propto h}, \tag{3.3}$$

where E is the set of haplotypes for a general mutation model and P^s is a reversible mutation (transition) matrix that describes the probabilities of haplotype α transforming to the next sampled haplotype h. $r_{\propto}$ is the number of haplotypes of type $\propto$ in the set H, r is the total number of haplotypes in H, and θ is a scaled mutation rate. The authors simplify this further by specifying that this corresponds to the next sampled haplotype, h, being obtained by applying a random number of mutations, s, to a randomly chosen existing haplotype, $\propto$. s is sampled from a randomly generated geometrically-distributed population. Equation (3.3) should be substituted into equation (3.2) for $\pi(h_{i1}|H_{-i})$ to complete the second step of the algorithm. $\pi(h_{i2}|H_{-i}, h_{i1})$ can be resolved because if information is known for $hi1$ then $hi2$ is the compliment and can be resolved easily if we know information on one half of the haplotype pair. θ must also be estimated to complete the calculation, a possible choice suggested is to use $\theta = S^{*}/log(2n)$, S^{*} corresponds to the number of loci that are used in the reconstruction.

The key to the logic of the PHASE algorithm is that unresolved haplotypes tend to be similar to known haplotypes, and the way in which the a priori expectation is calculated by using coalescent theory and other theories in population genetics. PHASE outputs include files that specify the number of each possible haplotype generated from the dataset and for each individual, give the most likely haplotype pair and all possible haplotype pairs, with corresponding probabilities. The haplotype reconstruction output files can be merged back with outcome data to allow logistic regression using another statistical programming package.

3.5.2 Hapassoc Package for the R Programming Environment

The Hapassoc package [3][4][12] for the R programming environment uses an EM algorithm by the method of weights to estimate regression parameters. The expectation step involves computing the conditional expected log likelihood of the complete data (x, y) given the observed data (x_{obs}, y) and the current parameter estimates. The maximization step maximizes the resulting function. The program also augments standard errors to account for ambiguity in reconstructed haplotype data using a formula by Louis[26].

Hapassoc starts the algorithm by generating "pseudo-individuals" for all individuals that have genotypes that result in multiple possible haplotypes. These "pseudo-individuals" represent all possible haplotype configurations for the ambiguous genotype and have a weight associated that can be calculated using Bayes' rule.

The conditional expected log likelihood [21] is a function of haplotype counts and is as follows:

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \sum_{i=1}^{n} E[l_{y|x}(\theta|x_i, y_i)|x_{obs,i}, y_i, \theta^{(t)}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{|S|} w_{ij}(\theta^{(t)}) l_{y|x}(\theta|x^{(j)}, y_i) + \sum_{i=1}^{n}\sum_{j=1}^{|S|} w_{ij}(\theta^{(t)}) l_x(\theta, x^{(j)}). \end{aligned}$$

$l_{y|x}$ is the log-likelihood for the regression model and l_x is the log-likelihood for the parameters of the covariate model, $\theta^{(t)} = (\beta^{(t)}, \gamma^{(t)})$ corresponds to the current parameter estimates, β is the regression parameters and γ is the covariate model parameters. The log-likelihood for β doesn't involve the parameters of the distribution of covariates γ so the maximizations can be done separately. Assuming independence of the non-genetic ("environmental") factors, x_e, and the genetic factors, x_g, the covariate vector can be partitioned into two separate components, x_e and x_g, each depending on parameters γ_e and γ_g, respectively. The log likelihood of the known parameters can be sectioned as genetic and non-genetic components $l_x(\gamma) = l_{x_g}(\gamma_g) +$

$l_{x_e}(\gamma_e)$, coupled with the assumption of completely observed environmenal covariates means that γ_e doesn't have to be estimated.

Weights w_{ij} for the pseudo-individuals with $x^{(j)}$ being the jth covariate vector are calculated using Bayes' rule:

$$\begin{aligned} w_{ij}^{(t)} &= Pr\{x_i = x^{(j)}|x_{obs,i}, y_i, \theta^{(t)}\} \\ &= \begin{cases} 0, & \text{if } x^{(j)} \text{ is not compatible with } x_{obs,i} \\ \frac{Pr(y_i|x^{(j)},\theta^{(t)})Pr(x^{(j)}|\theta^{(t)})}{\sum Pr(y_i|x^{(k)},\theta^{(t)})Pr(x^{(k)}|\theta^{(t)})}, & \text{if } x^{(j)} \text{ is compatible with } x_{obs,i} \end{cases} \\ &= \begin{cases} 0, & \text{if } x^{(j)} \text{ is not compatible with } x_{obs,i} \\ \frac{Pr(y_i|x^{(j)},\theta^{(t)})Pr(x_g^{(j)}|\theta^{(t)})Pr(x_e^{(j)}|\theta^{(t)})}{\sum Pr(y_i|x^{(k)},\theta^{(t)})Pr(x_g^{(k)}|\theta^{(t)})Pr(x_e^{(k)}|\theta^{(t)})}, & \text{if } x^{(j)} \text{ is compatible with } x_{obs,i} \end{cases} \end{aligned}$$

The summation over $x^{(k)}$ is all haplotypes that can derived from the $x_{obs,i}$ genotype. Since pseudo-individuals representing all possible haplotype configurations for an individual's genotype have been generated and added to the dataset, the weights w_{ij} for each can be represented as a_i for simplification of the double subscript. There will then be n + (# of pseudo-individuals added to the dataset)= M.

An assumption of complete non-genetic covariate information is made, which eliminates the need for estimation of γ_e. Assuming Hardy-Weinberg equilibrium further simplifies the covariate distribution, so the number of covariate parameters will be a maximum of $r-1$, where r is the number of haplotypes. The covariate model parameters will be the probabilities of each of the $r-1$ haplotypes. Pseudo-individuals homozygous for a haplotype k will contribute $2log\gamma_k$ to the complete-data log-likelihood for γ since their weight is 1, and those heterozygous for haplotypes k and l will contribute $a_i^{(t)}log\gamma_k\gamma_l = a_i^{(t)}log\gamma_k + a_i^{(t)}log\gamma_l$.

To update regression parameters, the weighted log-likelihood given by:

$$\sum_{i=1}^{M} a_i \left\{ y_i log(\frac{p_i}{1-p_i}) + log(1-p_i) \right\},$$

is maximized, where p_i is the probability that the ith person has the disease. The $t+1^{th}$ estimate of the regression coefficients are found by solving $\beta^{t+1} = \beta^t -$

$[I(\beta)^{-1}S(\beta)]|_{\beta=\beta^{(t)}}$ via Newton-Raphson optimization, $I(\beta)$ being the Hessian and $S(\beta)$ being the score function of the log-likelihood.

Standard errors calculated according to maximum likelihood theory for weighted logistic regression with known weights are not correct because of extra error caused by ambiguous haplotype information. The observed data likelihood is broken up into a sum of two terms: $logPr(X = x_{obs}, y|\theta) = logPr(x_e|\gamma_e)+log(\sum Pr(y|x,\beta)Pr(x_g,\gamma_g))$, assuming independence of (β, γ_g) and γ_e. We are only estimating standard errors of the regression parameters which we can obtain by inverting the submatrix $\mathrm{I}(\beta, \gamma_g)$.

Using a formula by Louis[26]:

$$\begin{aligned} I(\theta) &= E_\theta[I_c(\theta)|x_{obs}, y] - (E_\theta[S_c(\theta)S_c(\theta)^T|x_{obs}, y] - E_\theta[S_c(\theta)|x_{obs}, y]E_\theta[S_c(\theta)^T|x_{obs}, y]) \\ &= E_\theta[I_c(\theta)|x_{obs}, y] - cov[S_c(\theta)|x_{obs}, y] \end{aligned}$$

where $I(\theta)$ is the negative Hessian of the observed data log-likelihood, and $I_c(\theta)$ and $S_c(\theta)$ are the negative Hessian and score of the complete-data log-likelihood function, respectively. The second term of the expression for the information matrix may be viewed as a correction to account for the ambiguity of the haplotype phase. The Fisher Information matrix for $I(\beta, \gamma_g)$ is as follows and can be used to estimate the variance of regression parameters:

$$\begin{bmatrix} X^TWVX & 0 \\ 0 & NG + n_r/(1 - \sum_{i=1}^{r-1} r_i)^2 J \end{bmatrix} - S^T(W - WBW)S$$

A diagonal matrix of weights W estimated from the EM algorithm and necessary parameters are substituted into the Information matrix and it is inverted to find variances associated with regression parameters. S is a matrix whose rows are complete-data score vectors $S_{c,ij}$ for pseudo-individuals with rows arranged so that extra pseudo-individuals for subject i are arranged in consecutive order. B is a block diagonal matrix of 1's with the number of rows and columns of each block equal to the number of haplotypes compatible with a given subject's observed covariates. Each matrix has n^* rows corresponding to the number of pseudo-individuals with ambiguous haplotype phase.

$N_{(r-1)\times(r-1)}$ is a diagonal matrix whose elements are the sums over all individuals of the weighted counts of the first $r-1$ haplotypes, n_r is the sum of the weighted numbers of the rth haplotype, $G_{(r-1)\times(r-1)} = diag(1/\gamma_k^2)$ and $J_{(r-1)\times(r-1)}$ is a matrix of 1's.

3.5.3 Haplo.stats Package for the R Programming Environment

Haplo.stats package by Schaid [35][23][12] does a haplotype reconstruction and risk estimation in a similar manner to that of Burkett et al. [3][4][12], iteratively estimating haplotype frequencies conditional on observed data and the current estimate of regression parameters. Maximum likelihood estimation of the regression parameters is computed with the EM algorithm, computing the conditional expectation of the complete data log-likelihood given the observed data and current estimates of the parameters, and maximizing the resulting function.

For the case of haplotype/non-genetic interactions the vector of covariates $z = (x_e|x_g|x_{ge})$, x_e denotes non-genetic (environmental) covariates, x_g denotes genetic covariates and x_{ge} denotes the interaction covariate terms of non-genetic and genetic covariates, and β is a vector of associated regression coefficients, $\beta = (\beta_e|\beta_g|\beta_{ge})$. The likelihood for the genetic data is a function of haplotype frequencies, $\theta = (\theta_1, \cdots, \theta_J)$, θ_j being the frequency for the jth haplotype, $j = 1, \cdots, J$. h is a vector of haplotype frequencies for an individual, with the jth component being equal to the number of h_j haplotypes that the individual possesses. The likelihood is a function of haplotype frequencies and assuming Hardy-Weinberg equilibrium the probability of genotype g or, equivalently, haplotype h is: $Pr_\theta(g) = Pr_\theta(h) = \prod_{j=1}^{J} \binom{2}{h_j} \theta_j^{h_j}$.

For θ reparameterized as $\varphi_j = \frac{\theta_j}{1-\sum_{j=1}^{J-1}\theta_j}$ according to the constraint $\sum_{j=1}^{J}\theta_j = 1$, the probability of h can be expressed as:

$$Pr_\varphi(h) = \binom{2}{h_J} \prod_{j=1}^{J-1} \binom{2}{h_j} \varphi_J^{h_j} \left(1 + \sum_{j=1}^{j-1} \varphi_j\right)^{-2}.$$

If G_i is the set of haplotype pairs consistent with the observed phenotype k_i for the ith individual and $\Phi = (\beta, \varphi)$, the likelihood contribution under independence of x_g and x_e is:

$$L_i(\Phi) \quad = \quad \sum_{g \epsilon G_i} \{f_\beta(y|z)\} Pr_\varphi(g).$$

Adopting the EM by the method of weights (Ibrahim, 1990) with a generalized linear model, the density of y can be expressed as:

$$f_\beta(y|z) \quad = \quad exp\left\{\frac{y\eta - b(\eta)}{\alpha(\psi)} + c(y, \psi)\right\}.$$

Assuming canonical link function $\eta = z^T\beta$ the likelihood is a function of haplotype frequencies the complete data log-likelihood for the ith subject is

$$\begin{aligned} logL_i^{(c)}(\Phi) &= log\{f_\beta(y_i|z_i)\} + log\{Pr_\varphi(g_i)\} \\ &\propto \frac{y_i z_i^T\beta - b(z_i^T\beta)}{\alpha(\psi)} + \sum_{j=1}^{J-1} h_{ij} log\varphi_j - 2log(1 + \sum_{j=1}^{J-1} \varphi_j) \end{aligned}$$

The E-step of the EM algorithm involves taking the conditional expectation of the complete data log-likelihood given the observed data and is a function of the conditional probability of the haplotype counts $Pr(h_{ij}, h_{ij'}|d_i^{(obs)})$ for $i = 1, \cdots, N$ and $j, j' = 1, \cdots, J$ given the observed data and $d_i^{(obs)}$. The general form of the joint conditional probability distribution of haplotype counts under independence of x_g and x_e is

$$Pr(h_{ij}, h_{ij'}|d_i^{(obs)}) \quad = \quad \frac{f_\beta(y_i|z_i)Pr_\varphi(g_i)}{\sum_{g\epsilon G} f_\beta(y_i|z_i)Pr_\varphi(g)}$$

The M-step of the EM algorithm involves maximization of the conditional expectation of the complete data log-likelihood, the model parameters φ estimated from the kth iteration of the EM algorithm. The regression parameters can be easily estimated with a weighted regression where the weights are the conditional probabilities of the subjects' haplotype data, $w_{gi} = Pr(h_{ij}, h_{ij'}|d_i^{(obs)})$.

Again, due to extra ambiguity caused by uncertain haplotype phase the standard errors of the regression coefficients are augmented in the logistic regression. Instead of using the observed information matrix of the observed data computed using Louis' [26] formula, the observed information matrix is approximated by the empirical observed information matrix

$$I_e(\hat{\Phi}; y, m, X_e) = \sum_{i=1}^{N} s_i(\Phi) s_i^T(\Phi)|_{\Phi=\hat{\Phi}}.$$

where $s_i(\Phi)$ is the score function from the observed data likelihood for the ith individual [27].

3.6 Statistical Issues with Genotype Analysis

The complete association analysis of genetic data involves analysis of individual SNPs and if there is two or more SNPs present from a single gene in linkage disequilibrium they can be analyzed as haplotypes. Before SNP analysis Hardy-Weinberg equilibrium (HWE) should be tested on all SNPs to detect any genotyping error, underlying population substructure or biased sampling. Those not in HWE may not be the most reliable SNPs and further analysis may not include those SNPs. Hardy-Weinberg equilibrium was tested on the BCCA non-Hodgkin lymphoma study genes (H2AX, NBS1, and IL10) that had been chosen for SNP analysis. SNPs in Hardy-Weinberg equilibrium were then analyzed as independent variables in univariate and multivariate logistic regression models, adjusted for frequency-matched and confounding variables.

Penetrance analysis of SNPs involves comparison of different penetrance models to see if a simpler one is a better fit. Penetrance models include the most complex model, the codominant model, which is where each genotype is tested independently (e.g. for a SNP with alleles A and G, G being the major allele, the codominant model is $logit(y) = \beta_0 + \beta_1 \times \mathrm{G/A} + \beta_2 \times \mathrm{A/A}$), a recessive model, where the homozygous recessive allele is tested for significance (e.g. $logit(y) = \beta_0 + \beta_1 \times (\mathrm{A/A})$), a dominant model where an increase in the number of recessive alleles doesn't affect the relative risk (e.g $logit(y) = \beta_0 + \beta_1 \times (\mathrm{G/A} + \mathrm{A/A})$) and a multiplicative model, where an increase in the number of causal alleles a individual possesses also increases the risk of a disease to that factor (e.g. $logit(y) = \beta_0 + \beta_1 \times$ (a), a=0, 1, 2 copies of A). A test of different models is done using an analysis of variance (ANOVA) to see if the simpler model is significantly different than the baseline codominant model. If the simpler model isn't significantly different than the more complex model, an investigator would choose a simpler model of penetrance to describe the SNP. Significant BCCA SNPs were analyzed for underlying penetrance models to investigate if a recessive allele was significantly associated with NHL or if the risk of developing NHL multiplied by a factor of the number of causal alleles that an individual possessed.

Association studies of haplotypes rely on reconstruction methods to assign haplotypes to individuals based on their phenotype. If a computational method of haplotype reconstruction is used that assigns the "best" possible haplotype pair to a subject out of all possible haplotype pairs, an additional source of ambiguity is added to modeling of haplotype associations. The additional source of ambiguity results from those subjects who have several possible haplotype pairs, given a phenotype with heterozygous loci or missing genotype information. Using logistic regression to model the "best" possible haplotype against a response variable would result in regression parameters and standard errors where extra ambiguity isn't reflected accurately by the analysis, i.e. the error would be underestimated and regression coefficients wouldn't be properly weighted. To do a correct association analysis of reconstructed haplotypes, a weighted logistic regression should be used and the variance of the regression parameters computed would have to be inflated to account for haplotype ambiguity.

Also, if there is missing genotype data in the original data set and the haplotypes are reconstructed for all individuals, a large amount of missing data could result in regression coefficients biased towards the null. Investigators should be aware of both possibilities and account for them by correcting the standard errors and recording the fraction of missing data in a data set when reporting results of a study analysis.

Three methods of haplotype reconstruction and subsequent logistic regression were contrasted to explore the error associated with resulting regression parameters. The packages include Hapassoc [3][12], Haplo.stats [23][12] and PHASE version 2.1 [37].

Hapassoc and Haplo.stats both use weighted logistic regression in the M-step and methods that inflate standard errors of the regression coefficients; the packages use slightly different calculations to compute the standard errors. PHASE outputs all possible haplotype assignments and corresponding weights so these were used to do a weighted logistic regression. Standard errors calculated from the PHASE regression outputs were compared with Hapassoc and Haplo.stats for comparison with the inflated standard errors using those methods. The three methods were applied to BCCA study data where the three genes (H2AX, NBS1, and IL10) had been chosen for SNP analysis, having between two and four loci each. A simulated data set with three SNP loci was also used to contrast regression coefficients and standard errors between the methods.

Reconstructed haplotypes were analyzed as continuous variables with the input haplotype data having three values, 0, 1 and 2, the number of each haplotype that an individual possessed. This is the default method of analysis for Hapassoc and the only way to analyze haplotype data with Haplo.stats. PHASE reconstructed haplotype data was also analyzed as continuous variables in order to have consistent, comparable regression output from all methods. The resulting model has a maximum of $r-1$ covariates, the number of haplotypes present in the data set minus the baseline comparison haplotype, h_i is the haplotype covariate, $i = 1, \cdots, r-1$:

$$logit(y) \ = \ \beta_0 + \sum_{i=1} \beta_i * h_i, \text{ for } i = 1, \cdots, r-1.$$

Changing default pooling and zero tolerance (smallest frequency which a haplotype must have to be considered present in the data set) in the Hapassoc and Haplo.stats packages allows the user to specify the haplotypes that will be analyzed as individual variables. Haplotypes that occur with frequency below the zero tolerance weren't included in analysis and haplotypes that occur below the pooling tolerance but above the zero tolerance were grouped into a pooled haplotype variable.

Chapter 4

Analysis of the BC Cancer Agency NHL Data

4.1 Tests of HWE

Permutation tests of Hardy-Weinberg equilibrium were performed for all loci of interest in the BCCA data set because some loci have small expected values for allele frequencies, which would lead to questionable test results if large-sample test theory was to be used. The control data for all loci were tested using the permutation test of HWE; results are shown in table 4.1.

At a significance value of 0.05, all loci in the BCCA data set have Hardy-Weinberg equilibrium except for H2AX-12. However, the evidence against Hardy-Weinberg proportions for H2AX-12 is not significant at the 5% level after Bonferroni adjustment for multiple testing of 9 SNPs. For completeness, association analysis of the H2AX-12 was still performed, although if it did happen to achieve significance, the result may be viewed with a critical eye. Further analysis is shown with and without the H2AX-12 SNP, i.e. haplotype analysis was done for H2AX with and without H2AX-12.

SNP	Genotype	Proportion	Observed	Expected	P-value
H2AX-8	G/G	0.263	126	120	
	G/A	0.468	224	239	
	A/A	0.269	129	120	p=0.166
H2AX-11	C/C	0.389	185	177	
	C/T	0.44	209	223	
	T/T	0.171	81	72	p=0.104
H2AX-12	A/A	0.916	437	430	
	G/A	0.075	36	39	
	G/G	0.008	4	1	p=0.016
NBS1-11	WT/WT	0.004	2	1	
	WT/del(WT)	0.081	39	37	
	del(WT)/del(WT)	0.915	442	445	p=0.071
NBS1-12	G/G	0.112	53	55	
	G/A	0.446	211	212	
	A/A	0.442	209	206	p=1.000
NBS1-14	C/C	0.115	55	59	
	C/G	0.473	226	217	
	G/G	0.412	197	202	p=.435
NBS1-16	A/A	0.423	180	186	
	G/A	0.474	202	191	
	G/G	0.103	44	49	p=0.280
il10-20	A/A	0.422	186	186	
	A/G	0.458	202	201	
	G/G	0.12	53	54	p=0.770
il10-21	T/T	0.348	146	146	
	A/T	0.489	205	203	
	A/A	0.162	68	70	p=0.916

Table 4.1: Tests of Hardy-Weinberg Equilibrium

4.2 SNP Analysis

The outcome of interest was the case/control status of the NHL data set. Analysis of the SNP genotypes was carried out as a case-control analysis. Cases were individuals who had NHL and controls were patients who didn't have NHL. Case control analysis of the SNPs compared the variant allele frequencies between case and control subjects. Univariate analysis was carried out on all SNPs. Multivariate analysis was done with adjustment for non-SNP covariates that the case control data was matched on, sex, age group, region of residence and ethnicity.

Results of univariate analysis and multivariate analysis were similar, shown in table 4.2 and table 4.3. The only SNP that reached significance at the α=0.05 level was H2AX-8, examining the p-values computed for the global test of genotypic association (the first entry for each SNP in the column of p-values). Even if the significance level is corrected using the Bonferroni correction for multiple tests (α=0.05/9 if H2AX-12 included and α=0.05/8 if H2AX-12 not included in the multiple tests of significance), H2AX-8 is still borderline significant, having a p-value $\approx$ 0.005. The relative risks for H2AX-8 indicate that with fewer copies of the G allele, the risk of developing NHL is less or that more copies of the A allele has a protective effect.

4.2.1 Penetrance Analysis

The codominant model for SNP analysis was used for analysis of BCCA SNPs. The codominant model of penetrance was compared, using ANOVA, in turn to a recessive model of penetrance to test if the homozygosity of the recessive allele was a significant predictor of NHL and a multiplicative model of penetrance to test whether an increase in the number of recessive or dominant alleles increased or decreased the risk of disease by a factor equal to the number of those alleles present. The ANOVA p-value is the output p-value for a likelihood ratio test of model difference.

H2AX-8 SNP was significant in SNP analysis so a penetrance analysis was done to see if H2AX-8 follows a dominant, recessive or multiplicative model. Table 4.4 shows

				95% CI		
SNP Variant	Genotype	N	RR	Lower	Upper	p-value
H2AX-8	G/G	272				0.005
	A/G	446	0.894	0.661	1.209	0.467
	A/A	204	0.558	0.386	0.808	0.002
		922				
H2AX-11	C/C	363				0.513
	T/C	400	0.950	0.715	1.262	0.723
	T/T	143	0.796	0.539	1.174	0.250
		906				
H2AX-12	A/A	838				0.496
	G/A	67	0.938	0.570	1.546	0.803
	G/G	5	0.272	0.030	2.448	0.246
		910				
NBS1-11	WT/WT	815				0.657
	WT/del(WT)	65	0.790	0.472	1.322	0.370
	del(WT)/del(WT)	4	1.185	0.166	8.453	0.866
		884				
NBS1-12	G/G	387				0.950
	G/A	390	0.996	0.752	1.321	0.978
	A/A	95	0.930	0.592	1.462	0.754
		872				
NBS1-14	G/G	390				0.485
	C/G	441	0.971	0.739	1.128	0.833
	C/C	96	0.761	0.485	1.194	0.235
		927				
NBS1-16	A/A	357				0.253
	G/A	383	0.911	0.683	1.216	0.528
	G/G	72	0.647	0.386	1.086	0.099
		812				
il10-20	A/A	267				0.138
	G/A	384	1.054	0.770	1.441	0.744
	G/G	151	1.473	0.986	2.199	0.058
		802				
il10-21	T/T	346				0.285
	A/T	388	1.070	0.801	1.431	0.646
	A/A	117	1.404	0.922	2.138	0.114
		851				

Table 4.2: Univariate SNP analysis

			95% CI		
SNP Variant	Genotype	RR	Lower	Upper	p-value
H2AX-8	G/G				0.006
	A/G	0.903	0.665	1.226	0.513
	A/A	0.557	0.381	0.815	0.003
H2AX-11	C/C				0.582
	T/C	0.965	0.723	1.287	0.807
	C/C	0.813	0.548	1.207	0.305
H2AX-12	A/A				0.414
	G/A	0.954	0.573	1.588	0.856
	G/G	0.223	0.024	2.067	0.186
NBS1-11	WT/WT				0.539
	WT/del(WT)	0.748	0.444	1.258	0.274
	del(WT)/del(WT)	1.188	0.165	8.537	0.864
NBS1-12	G/G				0.991
	G/A	1.018	0.765	1.353	0.905
	A/A	0.997	0.629	1.580	0.989
NBS1-14	G/G				0.617
	C/G	0.997	0.757	1.314	0.985
	C/C	0.804	0.509	1.271	0.350
NBS1-16	A/A				0.296
	G/A	0.928	0.693	1.242	0.615
	A/A	0.658	0.389	1.113	0.119
il10-20	A/A				0.076
	G/A	1.168	0.828	1.646	0.377
	A/A	1.639	1.064	2.524	0.025
il10-21	T/T				0.285
	C/T	1.107	0.815	1.503	0.516
	C/C	1.450	0.941	2.234	0.092

Table 4.3: Multivariate SNP analysis

H2AX-8										
	Unadjusted					Adjusted				
Genotype	RR	Lower	Upper	p-val	ANOVA p-val	RR	Lower	Upper	p-val	ANOVA p-val
AG	0.894	0.661	1.209	0.467		0.903	0.665	1.226	0.513	
AA	0.558	0.386	0.808	0.002		0.557	0.381	0.815	0.003	
AG+AA	0.773	0.582	1.027	0.076	0.010	0.785	0.588	1.048	0.101	0.010
AA	0.599	0.436	0.823	0.002	0.470	0.595	0.429	0.824	0.002	0.510
nA	0.758	0.632	0.910	0.003	0.180	0.759	0.629	0.915	0.004	0.160
IL10-20										
	Unadjusted					Adjusted				
Genotype	RR	Lower	Upper	p-val	ANOVA p-val	RR	Lower	Upper	p-val	ANOVA p-val
GA	1.054	0.770	1.441	0.744		1.168	0.828	1.646	0.377	
GG	1.473	0.986	2.199	0.059		1.639	1.064	2.524	0.025	
GA+GG	1.158	0.863	1.555	0.329	0.080	1.278	0.920	1.774	0.143	0.080
GG	1.428	1.001	2.038	0.050	0.740	1.477	1.026	2.126	0.036	0.380
nG	1.190	0.978	1.449	0.083	0.330	1.235	1.025	1.575	0.029	0.540

Table 4.4: Table of significant SNP penetrances

that, as there was no rejection of the null hypothesis of a codominant model being different from a recessive model (p-val=0.47) or a multiplicative one (p-val=0.18), they were improvements over the codominant model or a dominant model. Inspection of ANOVA p-value for significant differences from the codominant model showed that the p-value was larger for the recessive model than for the multiplicative model. We chose a recessive model of penetrance since it had a larger p-value than the multiplicative model. Therefore, only if an individual has two copies of the A allele they are at a decreased risk of developing NHL.

A comment on SNPs with interesting results: the SNP not in HWE, H2AX-12, didn't reach significance and SNP IL10-20 is borderline significant when adjusted for non-SNP covariates with the A allele having a protective effect or alternatively, the G allele increasing risk of NHL. Using the Bonferroni correction, however, IL10-20 is no longer significant. Penetrance analysis of IL10-20, summarized in table 4.4 indicates that the SNP follows a multiplicative model, where an increase in the number of G alleles corresponds to a linear increase in the log risk of developing NHL.

Of the SNPs analyzed for association with NHL, only H2AX-8 had a borderline significant association with NHL after a Bonferroni correction for multiple comparisons of the SNPs. Before the correction, the IL10-20 SNP had a borderline significant association (p=0.076). Additional studies with a larger number of individuals would help increase the power of seeing an association, and the IL10-20 may produce a significant result.

H2AX-8 is a new polymorphism being studied for association with NHL and with no other prior published research with which to compare the results. After publication, other researchers will need to confirm the positive association with NHL with new studies.

4.3 Haplotype Analysis

Using three methods of haplotype reconstruction and logistic regression (Hapassoc, Haplo.stats, PHASE haplotype reconstruction), the SNP data from the BC Cancer Agency was analyzed. There are seven different ways to do an association analysis of the data with the three methods. The seven different ways to do the analysis are adjusted for the non-SNP variables of age, sex, ethnicity and region of residence. Additionally, an unadjusted analysis could be done.

Input for the reconstruction programs Hapassoc and Haplo.stats include SNP and non-SNP variables. SNP and outcome data were input into Hapassoc and Haplo.stats with the non-SNP variables of age, sex, region of residence and ethnic group for haplotype reconstruction of the SNP variables and EM weighted logistic regressions. It was also of interest to see the difference in standard errors and regression coefficients between Hapassoc or Haplo.stats EM weighted regression that inflates standard errors and a regular weighted regression with no inflation of standard errors using the reconstructed haplotype information from Hapassoc and Haplo.stats. Hapassoc does an initial step of finding starting points for the weights associated with haplotypes by performing a reconstruction with SNP data only. It uses initial weights in the EM step of calculating the regression coefficients and standard errors and updates the associated weights when it updates the regression estimates. A comparison was possible between weighted regression output calculated from initial weights output from the Hapassoc procedure and the weighted regression output calculated from final weights output from the Hapassoc procedure. We were able to see the impact of using slightly different weights in weighted regression.

The comparisons able to be done with the PHASE program were different than for Hapassoc and Haplo.stats since PHASE uses only the information in SNP input data for reconstruction; it doesn't use information from other non-SNP variables in its haplotype reconstruction or associated weight calculation. PHASE outputs files have a "best" haplotype assignment for individuals that assigns each individual a single haplotype pair, and a weighted reconstruction weighting all possible haplotype assignments for an individual by the probability that it is the correct haplotype assignment. Using PHASE output a logistic regression using the "best" haplotype assignment was done as well as a weighted logistic regression with the output haplotypes and associated weights.

The seven different ways of doing the association analysis are:

1. PHASE "best" haplotype assignment with regular logistic regression, labelled PHASE "best" LR in tables
2. PHASE assigns all possible haplotypes to each individual with associated weights analyzed with a weighted logistic regression, labelled PHASE weighted LR in tables
3. Hapassoc with full EM, labelled Hapassoc EM in tables
4. Hapassoc with weighted logistic regression using INITIAL weights, labelled Hapassoc INITIAL LR in tables
5. Hapassoc with weighted logistic regression using FINAL weights, labelled Hapassoc FINAL LR in tables
6. Haplo.stats with full EM, labelled Haplo.stats EM in tables
7. Haplo.stats with weighted logistic regression using weights output from Haplo.stats EM, labelled Haplo.stats weighted LR in tables

The association analysis variations were applied to the BC Cancer Agency SNP data. As discussed there were three genes analyzed, IL10 with two SNPs, and H2AX and NBS1, with three and four SNPs respectively. The multi-SNP haplotypes were reconstructed and the regression model for each gene was chosen using a pooling frequency of 0.01, which pools haplotypes that have population frequency less than 1%. So the regression models consisted of haplotype covariates that have frequency greater than 1% and a pooled haplotype covariate (if needed).

H2AX, 3 loci							
PHASE		"best" LR		Weighted LR			
Haplotype	Freq	β	SE(β)	β	SE(β)		
GCA	0.498						
GCG	0.038	-0.324	0.234	-0.324	0.231		
ACA	0.073	-0.567	0.186	-0.571	0.178		
ATA	0.386	-0.231	0.098	-0.221	0.097		
Pooled	0.004	0.678	0.721	0.379	0.561		
Hapassoc		EM		INITIAL LR		FINAL LR	
Haplotype	Freq	β	SE(β)	β	SE(β)	β	SE(β)
GCA	0.495						
GCG	0.040	-0.424	0.251	-0.339	0.236	-0.424	0.239
ACA	0.081	-0.659	0.195	-0.591	0.185	-0.659	0.187
ATA	0.376	-0.205	0.101	-0.200	0.100	-0.205	0.100
Pooled	0.008	0.759	0.692	0.378	0.571	0.759	0.594
Haplo.stats		EM		Weighted LR			
Haplotype	Freq	β	SE(β)	β	SE(β)		
GCA	0.495						
GCG	0.039	-0.413	0.250	-0.413	0.233		
ACA	0.082	-0.674	0.195	-0.674	0.180		
ATA	0.377	-0.235	0.100	-0.235	0.097		
Pooled	0.008	0.780	0.718	0.780	0.574		

Table 4.5: Table of output from BC Cancer Agency SNP data, H2AX, 3 loci

NBS1							
PHASE		"best" LR		Weighted LR			
Haplotype	Freq	β	SE(β)	β	SE(β)		
WTGGA	0.616						
WTAGA	0.002	0.848	1.247	0.828	1.221		
WTACG	0.346	-0.124	0.102	-0.098	0.100		
DELGGA	0.037	-0.306	0.239	-0.210	0.231		
Hapassoc		EM		INITIAL LR		FINAL LR	
Haplotype	Freq	β	SE(β)	β	SE(β)	β	SE(β)
WTGGA	0.618						
WTAGA	0.002	0.843	1.250	0.836	1.243	0.843	1.245
WTACG	0.338	-0.089	0.105	-0.088	0.105	-0.089	0.105
DELGGA	0.043	-0.230	0.243	-0.214	0.233	-0.230	0.234
Haplo.stats		EM		Weighted LR			
Haplotype	Freq	β	SE(β)	β	SE(β)		
WTGGA	0.618						
WTAGA	0.002	0.825	0.008	0.825	1.196		
WTACG	0.339	-0.102	0.103	-0.102	0.100		
DELGGA	0.041	-0.201	0.239	-0.201	0.226		

Table 4.6: Table of output from BC Cancer Agency SNP data, NBS1

IL10							
PHASE		"best" LR		Weighted LR			
Haplotype	Freq	β	SE(β)	β	SE(β)		
TA	0.552						
TG	0.068	0.418	0.174	0.313	0.152		
AA	0.017	1.120	0.390	0.641	0.274		
AG	0.363	0.160	0.102	0.160	0.101		
Hapassoc		EM		INITIAL LR		FINAL LR	
Haplotype	Freq	β	SE(β)	β	SE(β)	β	SE(β)
TA	0.539						
TG	0.095	0.459	0.178	0.354	0.158	0.459	0.160
AA	0.026	1.128	0.392	0.755	0.301	1.128	0.327
AG	0.340	0.173	0.107	0.178	0.104	0.173	0.105
Haplo.stats		EM		Weighted LR			
Haplotype	Freq	β	SE(β)	β	SE(β)		
TA	0.539						
TG	0.095	0.445	0.177	0.445	0.155		
AA	0.026	1.119	0.391	1.119	0.316		
AG	0.340	0.165	0.107	0.165	0.101		

Table 4.7: Table of output from BC Cancer Agency SNP data, IL10

H2AX, 2 loci							
PHASE		"best" LR		Weighted LR			
Haplotype	Freq	β	SE(β)	β	SE(β)		
GC	0.536						
GT	0.004	0.898	0.849	0.872	0.834		
AC	0.074	-0.544	0.184	-0.529	0.175		
AT	0.386	-0.202	0.096	-0.196	0.095		
Hapassoc		EM		INITIAL LR		FINAL LR	
Haplotype	Freq	β	SE(β)	β	SE(β)	β	SE(β)
GC	0.534						
GT	0.004	0.868	0.852	0.758	0.768	0.868	0.783
AC	0.085	-0.593	0.186	-0.558	0.178	-0.593	0.179
AT	0.377	-0.204	0.097	-0.205	0.097	-0.204	0.097
Haplo.stats		EM		Weighted LR			
Haplotype	Freq	β	SE(β)	β	SE(β)		
GC	0.534						
GT	0.004	0.899	0.851	0.899	0.767		
AC	0.085	-0.586	0.186	-0.586	0.175		
AT	0.377	-0.204	0.097	-0.204	0.095		

Table 4.8: Table of output from BC Cancer Agency SNP data, H2AX, 2 loci

- H2AX, 3 loci
  - Regression Coefficients

    Overall, when comparing Hapassoc and Haplo.stats EM methods in table 4.5, they are relatively consistent in their computation of regression coefficients. The absolute values of the regression coefficients are generally larger for both Hapassoc EM and Haplo.stats EM than PHASE weighted logistic regression. One exception was the haplotype ATA in the Hapassoc model, which was likely due to the slight difference in frequencies of the haplotypes as well as the slightly different weights computed. Slightly different weights are computed by PHASE because PHASE only takes the SNP information into account when computing weights, this affects the regression coefficients to a larger degree than was initially expected. The same can also be seen with the Hapassoc initial weighted regression coefficients, which are smaller in absolute value than the properly weighted Hapassoc and Haplo.stats EM and Hapassoc final weighted regression coefficients. The frequencies of Hapassoc EM and Haplo.stats EM haplotype output are comparable with the largest difference being 0.001, PHASE frequencies are all slightly smaller, with the exception of ATA, which could account for the larger coefficient.
  - Standard Errors

    The standard errors are consistently larger for Hapassoc EM and Haplo.stats EM than for PHASE weighted and "best" regressions for the smaller frequency haplotypes, the large frequency haplotype ATA didn't have standard errors that varied much between the methods. The inflation of standard errors don't make a large difference in the p-values between the methods because the smaller regression coefficients output by PHASE and the smaller standard errors allow us to arrive at similar p-values as the larger regression coefficients and standard errors of the Hapassoc EM and Haplo.stats EM packages.

Comparing Hapassoc and Haplo.stats EM regressions with weighted logistic regressions on haplotype reconstruction output from the packages shows a negligible difference between the packages. The standard errors of regression coefficients are larger for Hapassoc and Haplo.stats than their corresponding weighted logistic regressions, which was as expected. For example, the coefficient for GCG for both methods had a larger standard error (Hapassoc=0.251, Haplo.stats=0.250) than for their corresponding weighted regression output with no inflation (Hapassoc=0.236, 0.239, Haplo.stats=0.233). PHASE had a standard error of 0.234 for the GCG coefficient with a "best" haplotype regular logistic regression and 0.231 for the weighted regression, both of which are smaller than the inflated standard errors from Hapassoc and Haplo.stats.

  - Interpretation

    For H2AX, both haplotypes ATA and ACA appear to exhibit a protective effect against non-Hodgkin lymphoma. If we were trying to choose a model at the 0.05 level the same outcome would be reached if any of the seven regression methods were used. Even though a correct model would be chosen for the PHASE "best" regression, the model conclusions would be found by faulty methodology. The correct conclusion is reached with this set of data but for another set of data there is the chance that using faulty methodology would lead to incorrect conclusions when testing significance of model variables or calculating confidence intervals.

- NBS1

  - Regression coefficients

    The PHASE "best" regression output calculates the largest regression coefficients of any of the methods as seen in table 4.6, particularly WTAGA and DELGGA. The regression coefficients are otherwise similar for all methods.

  - Standard Errors

    When doing similar comparisons for NBS1 as for H2AX, the same conclusions can be reached. Standard errors are appropriately inflated for the Hapassoc and Haplo.stats EM regressions and if the either package were being used for model selection the same models would be specified. There is an interesting outcome in the standard error for the Haplo.stats logistic regression, the standard error for WTAGA haplotype is very small (0.0008 for Haplo.stats regression with inflated standard error). After doing testing to explore the cause, the only resulting hypothesis is that there is an error in the Haplo.stats code that under certain conditions (one of them being an extremely small frequency of the resulting haplotype) causes Haplo.stats to output an incorrect standard error. Even though the WTAGA haplotype had a small frequency, it was still included in analysis since there was only 4 possible haplotypes present for NBS1 and haplotype DELGGA had a frequency above the pooling tolerance. There were no other haplotypes in the "pooled" group so the WTAGA haplotype was analyzed on its own.

  - Interpretation

    The same model conclusions would be reached for all methods of regression on NBS1 haplotypes - no haplotypes had significance at the 0.05 level.

- IL10

  - Regression coefficients

    The regression coefficients for the small frequency haplotypes are larger for the Hapassoc EM, Hapassoc final weighted regressions and Haplo.stats EM and weighted logistic regressions than for the PHASE weighted regressions and Hapassoc initial weighted regressions in table 4.7. The difference may be due to the slightly different weights computed by the reconstructions with non-SNP information and the small frequencies of the haplotypes. The interesting logistic regression output is the PHASE "best" logistic regression with estimates that are comparable to the Hapassoc EM and

Haplo.stats EM logistic regressions, possibly because of different frequencies calculated for the resulting haplotypes, almost a 3% difference for the TG haplotype.

- Standard Errors

  The IL10 SNPs were the most interesting, since there was an fairly substantial inflation of standard errors for the Hapassoc and Haplo.stats EM logistic regressions, compared with weighted regressions. For example, looking at haplotype TG, the corresponding standard error for the Hapassoc regression is 0.178, for the Haplo.stats regression is 0.177, while the regular weighted regressions using the output from these methods only shows standard errors of 0.158 (Hapassoc initial weights), 0.160 (Hapassoc final weights) and 0.155 (Haplo.stats weighted). The IL10 SNPs were the best example of the packages adjusting coefficient standard error, as there were the maximum possible different haplotypes present and even though most of the haplotypes were characterized in two groups (TA and AG), there was still a third haplotype group (TG) that contained almost 10% of the total haplotypes.

- Interpretation

  There were two haplotypes, TG and AA, which appear to significantly increase the risk of developing NHL. Due to the small frequency of the AA haplotype, we have less confidence in only saying that AA haplotype definitely increases the risk of developing NHL.

  Compared with the individual SNP analysis of the IL10 SNPs which shows only that the SNP il10-20 is borderline significant at the α=0.05 level, it would appear that haplotype analysis is a necessary tool with which to analyze the SNP data in this instance. The haplotype analysis gives more clues as to where a possible mutation may occur because a larger area was taken into consideration when the SNPs were analyzed as a unit. Interaction of the two IL10 SNPs is another possible cause for the significance of the association with NHL. A study with a larger number of people may give a better idea of whether these haplotypes are indeed associated with the disease. The small number of people with these haplotypes makes us cautious to declare significance with much confidence.

- H2AX, 2 loci

  - Regression Coefficients

    The regression coefficients in table 4.8 are comparable to the regression coefficients in table 4.5, the GT variable being comparable to the pooled variable in the H2AX table with three loci.

  - Standard Errors

    As with the three-locus haplotypes, the coefficient standard errors were larger when comparing the Hapassoc and Haplo.stats EM weighted regression standard errors to the PHASE "best" and weighted regressions in table 4.8. The same haplotypes would be kept in the final model for either method. Conclusions don't differ whether or not the corrected standard errors are used.

  - Interpretation

    The same model conclusions would be reached for all methods of logistic regression. There is not enough variability in the SNP input data set to inflate the standard errors significantly enough to come to different conclusions with regards to model selection or confidence interval calculation.

    When using the most frequent haplotype as a baseline for comparison, there were two haplotypes (AT and AC) significant at the α=0.05 level for the H2AX SNPs for all methods of reconstruction and regression. These two haplotypes corresponded to the first two loci of the two significant three-locus H2AX haplotypes, indicating that we could simplify the haplotypes into two locus haplotypes and get similar results. This also cuts down on the area of the genome where a possible disease-causing mutation may occur. If we want to simplify the area of possible disease-causation even more, we may turn to the analysis of the H2AX SNPs. SNP analysis has shown that H2AX-8 is significant in the prediction of NHL, but H2AX-11 is not significant in the prediction of NHL. Penetrance model selection and allele analysis has shown that it is most likely a larger number of copies of the G allele of the H2AX-8 that is associated with increased risk of NHL and that a larger number of copies of the A allele is associated with a protective effect with regards to NHL.

Chapter 5

Simulation Study

To investigate differences in the coefficients and standard errors of logistic regressions between haplotype reconstruction and logistic regression methods, a simulation study was used. Regression estimates and standard errors were studied for each of the three methods of reconstruction (Hapassoc, Haplo.stats, PHASE) and logistic regression under differing percentages of missing data.

5.1 Haplotype Data Generation

Using information gained from the small comparison exercise of Hapassoc EM regression and a weighted logistic regression with haplotype assignments and weights generated from Hapassoc, haplotype frequencies and model specifications were chosen to be the same as those in the Hapassoc example data. The frequencies that were used resulted in the potential haplotype ambiguity being increased and the distribution of haplotype frequencies being more equal and not mostly allocated to one or two haplotypes.

Using the R program, an effective population of 50,000 individuals was generated. The eight possible haplotypes used in the simulation and corresponding frequencies for each are shown in table 5.1.

Haplotype	Locus	Frequency
	1 2 3	
1	0 0 0	0.2517911
2	0 0 1	0.2605418
3	0 1 0	0.23606001
4	0 1 1	0.0916067
5	1 0 0	0.10133627
6	1 0 1	0.02636844
7	1 1 0	0.0108126
8	1 1 1	0.02148268

Table 5.1: Table of Haplotype Frequencies.

Next, to prepare data for logistic regression, outcomes had to be generated for the haplotype data. The haplotype that was the second-most frequent was chosen to be the ”affected” haplotype, meaning that a carrier of this haplotype was at increased risk of developing a disease. Overall, in the sample, the probability of having disease was fixed at 0.5 to make the simulation as much representative of a case/control study as possible. The risk increase for the having of a single affected haplotype from having no affected haplotypes was arbitrarily fixed as 1.5 and the risk increase of having two affected haplotypes from having no affected haplotypes was fixed at 3. The equation

$$Pr(\mathrm{D}) = \sum_{i=0}^{2} Pr(h_i) * Pr(\mathrm{D}|h_i)$$

where h_i, the number of ”affected” haplotypes, was filled in with known information, such as haplotype probabilities from the generated data set and the pre-specified beta coefficients. $Prob(\mathrm{D})$ is the probability of having the disease and h_i is the number of ”affected” haplotypes that an individual is carrying, h_1 means one affected haplotype is present, etc. The equation:

$$\begin{aligned} 0.5 &= Pr(h_0) * Pr(\mathrm{D}|h_0) + Pr(h_1) * Pr(\mathrm{D}|h_1) + Pr(h_2) * Pr(\mathrm{D}|h_2) \\ &= Pr(h_0) * \frac{\exp(\beta_0)}{1+\exp(\beta_0)} + Pr(h_1) * \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} + Pr(h_2) * \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)} \end{aligned}$$

was solved for β_0, since all other information was known. Probabilities of the number of "affected" haplotypes was easily calculated, β_1 was already specified to be ln(1.5) and β_2 was ln(3). The probabilities of being affected given 0, 1 or 2 "affected" haplotypes were then calculated. Binary outcomes were generated for each number of "affected" haplotypes that an individual was carrying using the calculated probabilities and assigned to the corresponding haplotype data.

Missing genotype data was randomly generated for 5% and 10% of subjects. For example, if the percentage of missing data was set at 5% the number of people who would have missing data was calculated as $0.05 * N$, then split between the 3 loci, so each locus would have 1/3 of the data re-assigned as missing. The same was then done for 10% missing data.

5.2 Analysis of Simulated Data

The model used all haplotypes with frequency greater than 5% and a pooled variable for those less than 5%. The model specified is:

$$logit(y) = \beta_0 + \beta_1 * h000 + \beta_2 * h010 + \beta_3 * h011 + \beta_4 * h100 + \beta_5 * pooled$$

where the pooled variable is a category created from those haplotypes that have a probability less than 5 % , haplotypes 6, 7 and 8. Base comparison haplotype is haplotype 2 (001). A logistic regression on known haplotypes was done to produce a basis for comparison for the three haplotype reconstruction and seven logistic regression methods.

Table 5.2 has the output of the logistic regression using the true haplotypes for the larger model with five covariates, table 5.3 has PHASE regression output for 50,000 individuals, table 5.4 has Haplo.stats regression output for 50,000 individuals, table 5.5 has PHASE regression output for 100 logistic regressions of 500 subjects each, table 5.6 has Hapassoc regression output for 100 logistic regressions of 500 subjects each and table 5.7 has Haplo.stats regression output for 100 logistic regressions of 500 subjects each.

50,000 indiv.			
Haplotype	β	SE(β)	
000	0.489	0.018	
010	0.008	0.018	
011	-0.007	0.024	
100	0.011	0.024	
Pooled	0.040	0.029	
100 LR, 500 indiv.			
Haplotype	Mean(β)	Mean(SE(β))	SD(β)
000	0.497	0.183	0.183
010	0.007	0.183	0.199
011	-0.010	0.248	0.250
100	0.012	0.239	0.263
Pooled	0.043	0.294	0.287

Table 5.2: Initial Estimates of Logistic Regression Coefficients, 5 Covariates

- PHASE and Haplo.stats, 50,000 individuals

  Looking at table 5.3 there isn't a large difference between the regression coefficients of the PHASE "best" and PHASE weighted regressions, other than PHASE weighted coefficients being slightly smaller. Both logistic regression methods approximate the estimates from logistic regression using the true haplotypes in table 5.2 rather well, with PHASE weighted regression coefficients being slightly biased. The standard errors are also similar to those calculated in the logistic regression using the true haplotypes.

  Haplo.stats output in table 5.4 shows a similar outcome, with regression coefficients being well approximated by both methods, compared to the logistic regression using the true haplotypes in table 5.2. There is some slight variation in the standard errors, but we can attribute that to random variation. In general, the logistic regressions of 50,000 individuals didn't lead to any interesting conclusions about the PHASE and Haplo.stats methods.

- PHASE, 100 logistic regressions 500 individuals each

  The PHASE methods in table 5.5 appear to estimate the regression coefficients

50,000 indiv.		PHASE "best" LR		Weighted LR	
Haplotype	Freq	β	SE(β)	β	SE(β)
Complete Data					
001	0.289				
000	0.213	0.465	0.018	0.404	0.018
010	0.257	0.009	0.018	0.004	0.018
011	0.081	0.017	0.025	-0.101	0.024
100	0.125	0.042	0.022	0.024	0.024
Pooled	0.035	0.063	0.061	-0.079	0.029
5% Missing Data					
001	0.289				
000	0.219	0.452	0.018	0.386	0.018
010	0.256	0.009	0.018	0.006	0.018
011	0.078	0.013	0.026	-0.112	0.024
100	0.123	0.036	0.022	0.020	0.024
Pooled	0.034	0.078	0.062	-0.092	0.028
10% Missing Data					
001	0.296				
000	0.218	0.441	0.018	0.380	0.018
010	0.256	0.009	0.018	0.000	0.018
011	0.075	-0.002	0.026	-0.110	0.024
100	0.122	0.039	0.022	-0.026	0.023
Pooled	0.033	0.026	0.062	-0.092	0.029

Table 5.3: PHASE LR of 50,000 individuals, 5 Covariates

50,000 indiv.		Haplo.stats EM		Weighted LR	
Haplotype	Freq	β	SE(β)	β	SE(β)
Complete					
001	0.260				
000	0.252	0.522	0.018	0.522	0.018
010	0.236	0.004	0.019	0.004	0.018
011	0.093	0.040	0.031	0.040	0.024
100	0.100	0.019	0.026	0.019	0.024
Pooled	0.060	0.064	0.034	0.064	0.029
5% Missing Data					
001	0.256				
000	0.252	0.521	0.022	0.521	0.018
010	0.236	0.006	0.019	-0.024	0.018
011	0.093	0.035	0.032	0.035	0.024
100	0.100	0.014	0.026	0.014	0.024
Pooled	0.059	0.063	0.035	0.063	0.029
10% Missing Data					
001	0.260				
000	0.251	0.517	0.022	0.517	0.018
010	0.235	-0.001	0.019	-0.001	0.018
011	0.093	0.027	0.032	0.027	0.024
100	0.101	0.020	0.026	0.020	0.024
Pooled	0.060	0.049	0.035	0.049	0.029

Table 5.4: Haplo.stats LR of 50,000 individuals, 5 Covariates

100 LR, 500 indiv.	PHASE "best" LR			Weighted LR		
Haplotype	Mean(β)	Mean(SE(β))	SD(β)	Mean(β)	Mean(SE(β))	SD(β)
Complete Data						
000	0.473	0.183	0.181	0.414	0.221	0.223
010	0.008	0.180	0.178	0.007	0.222	0.225
011	0.017	0.259	0.266	-0.105	0.300	0.281
100	0.041	0.220	0.233	0.029	0.290	0.292
Pooled	-0.047	2.245	1.380	-0.073	0.353	0.277
5% Missing Data						
000	0.460	0.183	0.183	0.393	0.230	0.207
010	0.009	0.180	0.182	0.007	0.233	0.200
011	0.012	0.262	0.272	-0.118	0.311	0.270
100	0.036	0.221	0.238	0.018	0.308	0.280
Pooled	-0.026	2.250	1.384	-0.101	0.359	0.266
10% Missing Data						
000	0.448	0.184	0.182	0.382	0.240	0.222
010	0.009	0.180	0.183	-0.006	0.242	0.251
011	-0.006	0.265	0.274	-0.124	0.327	0.269
100	0.039	0.222	0.216	0.022	0.316	0.291
Pooled	-0.202	4.236	1.896	-0.105	0.386	0.298

Table 5.5: PHASE 100 LR, 500 individuals, 5 Covariates

100 LR, 500 indiv.	Hapassoc EM			Weighted LR-INITIAL			Weighted LR-FINAL		
Haplotype	Mean(β)	Mean(SE(β))	SD(β)	Mean(β)	Mean(SE(β))	SD(β)	Mean(β)	Mean(SE(β))	SD(β)
Complete									
000	0.537	0.217	0.226	0.412	0.182	0.166	0.537	0.184	0.226
010	0.004	0.188	0.200	0.004	0.183	0.190	0.004	0.184	0.200
011	0.046	0.321	0.314	-0.101	0.247	0.197	0.046	0.249	0.314
100	0.016	0.264	0.255	0.025	0.238	0.234	0.016	0.240	0.255
Pooled	0.074	0.354	0.368	-0.078	0.294	0.276	0.074	0.296	0.368
5% Missing									
000	0.536	0.220	0.234	0.398	0.182	0.164	0.536	0.184	0.234
010	0.007	0.190	0.204	0.002	0.183	0.188	0.007	0.184	0.205
011	0.040	0.327	0.318	-0.100	0.247	0.195	0.040	0.249	0.318
100	0.011	0.266	0.265	0.025	0.238	0.230	0.011	0.240	0.265
Pooled	0.074	0.360	0.382	-0.081	0.295	0.271	0.074	0.297	0.382
10% Missing									
000	0.532	0.224	0.231	0.386	0.182	0.153	0.532	0.184	0.231
010	0.000	0.191	0.206	0.004	0.183	0.186	0.000	0.184	0.206
011	0.033	0.332	0.313	-0.102	0.247	0.188	0.033	0.250	0.313
100	0.019	0.269	0.266	0.025	0.239	0.229	0.019	0.240	0.266
Pooled	0.063	0.365	0.379	-0.078	0.295	0.275	0.063	0.297	0.379

Table 5.6: Hapassoc 100 LR, 500 individuals, 5 Covariates

100 LR, 500 indiv.	Haplo.stats EM			Weighted LR		
Haplotype	Mean(β)	Mean(SE(β))	SD(β)	Mean(β)	Mean(SE(β))	SD(β)
Complete Data						
000	0.537	0.217	0.226	0.537	0.184	0.226
010	0.004	0.188	0.200	0.004	0.184	0.200
011	0.046	0.321	0.314	0.046	0.249	0.314
100	0.016	0.264	0.255	0.016	0.240	0.255
Pooled	0.075	0.354	0.368	0.075	0.296	0.368
5% Missing Data						
000	0.537	0.221	0.234	0.537	0.184	0.234
010	0.007	0.190	0.204	0.007	0.184	0.204
011	0.041	0.327	0.319	0.041	0.249	0.319
100	0.011	0.266	0.265	0.011	0.240	0.265
Pooled	0.075	0.360	0.383	0.075	0.297	0.383
10% Missing Data						
000	0.532	0.224	0.231	0.532	0.184	0.231
010	0.000	0.191	0.206	0.000	0.184	0.206
011	0.034	0.332	0.313	0.034	0.248	0.313
100	0.018	0.270	0.265	0.018	0.240	0.265
Pooled	0.062	0.365	0.378	0.062	0.297	0.378

Table 5.7: Haplo.stats 100 LR, 500 individuals, 5 Covariates

rather well compared to the logistic regressions using the true haplotypes obtained in table 5.2. The PHASE "best" method even appears to approximate the regression coefficients better than the PHASE weighted regressions, as some of the estimates for the weighted regressions appear to be slightly biased. The weighted regression was biased because weights output by the PHASE haplotype reconstruction aren't computed utilizing information from the outcome variable, they only use information from SNP input data. The standard errors for both PHASE methods adequately approximate the computed standard deviations and the standard deviations computed from the logistic regression using the true haplotypes.

However, it seems that the PHASE "best" regression method has some trouble with rare haplotypes. One of the most striking results is for PHASE "best" analysis of 100 logistic regressions of 500 subjects each (table 5.5) is that there is a rather large mean standard error for all of the coefficients for the pooled haplotype in the PHASE "best" reconstruction and unweighted logistic regression of 100 regressions of 500 individuals (standard errors=2.245, 2.250, 4.236, second column of table 5.5). Inspection of the list of regression covariates shows that there is two groups of 500 individuals where the regression coefficient and associated standard error is extremely large and affects the means and standard deviations. It appears that the logistic regressions didn't converge for these samples and this is likely due to a sparse data problem. When the two simulations with a very large βs and standard errors were deleted for each of the outcome data sets, the standard errors for the pooled variable (table 5.8) are more comparable to the population standard deviations. The problem may be that since the pooled haplotypes are rare, there were very few pooled haplotypes in the one sample of 500 individuals that was calculated to have the large standard error. An alternative to recording mean estimates of regression coefficients is to record median estimates. Median estimates for the regression coefficients and standard errrors are given in the table 5.8 for comparison.

	PHASE "best" LR				
Haplotype	Mean(β)	Median(β)	Mean(SE(β))	Median(SE(β))	SD(β)
Complete Data, 100 LR, 500 subjects					
Pooled	-0.030	0.039	0.688	0.657	0.723
5% Missing Data, 100 LR, 500 subjects					
Pooled	0.092	0.020	0.693	0.661	0.726
10% Missing Data, 100 LR, 500 subjects					
Pooled	0.048	0.041	0.694	0.662	0.700

Table 5.8: PHASE Reconstruction and Regression Output for Pooled Haplotype Variable, with Outliers Excluded

- Hapassoc, 100 logistic regressions 500 individuals each

  There is a large difference in the regression coefficients for the Hapassoc full EM method as compared to using the Hapassoc INITIAL weights with a weighted logistic regression, which can be seen in table 5.6 when comparing the mean(β) columns for Hapassoc EM and weighted regression using INITIAL weights. The reason is that Hapassoc does a reconstruction of the haplotypes and associated weights for an input data set in an initial step without non-SNP information before the EM regression. Even though initial weights vary only slightly from the updated weights calculated in the EM regression with weights (the largest difference between the initial and final weights is approximately 0.036), they underestimate some of the regression coefficients in a noticeable way when compared with the estimates from logistic regression using the true haplotypes in table 5.2. The output for the weighted regression using the correct updated Hapassoc weights is in FINAL logistic regression columns of table 5.6 and was quite different, with estimates being similar to those given by the Hapassoc full EM method. The Hapassoc EM method outputs regression coefficients slightly larger than the coefficients given by logistic regression using the true haplotypes.

  The mean(SE(β)) for the Hapassoc EM weighted regressions were larger overall and closer approximated the SD(β), the population standard error for the regression coefficients than comparable mean(SE(β)) for the regular weighted regressions in the FINAL weights columns of table 5.6. This shows that the inflation of the standard errors of the Hapassoc EM correct the bias in the unadjusted standard errors.

- Haplo.stats, 100 logistic regressions 500 individuals each

  In table 5.7 the standard errors between the Haplo.stats methods were different, as was expected, with the Haplo.stats EM weighted regression mean(SE(β)) being larger than the mean(SE(β)) of the regular weighted regression and a better approximation to the population standard deviations.

  The regression coefficients were the same between methods and approximate the estimates computed by logistic regression using the true haplotypes in table 5.2 to a reasonable degree.

Overall, the methods of regression that adjust the standard errors (Hapassoc and Haplo.stats EM) are more conservative and would result in the correct number of positive conclusions in tests of significance.

Chapter 6

Summary and Conclusion

6.1 Non-Hodgkin Lymphoma Data Analysis

The SNP analysis found only two SNPs that were statistically significantly associated with the incidence of NHL. The H2AX-8 and IL10-20 SNPs were significant at the 0.05 level. Penetrance analysis revealed that the H2AX-8 SNP followed a recessive model of penetrance with individuals homozygous for the A allele having a lower relative risk (RR=0.592) of developing NHL compared with having any G alleles present. The IL10-20 SNP also followed a recessive model with homozygosity of the G allele increasing the risk of NHL (RR=1.470), compared to having an A allele present. When adjustment for multiple tests is considered using a Bonferroni correction, only the H2AX-8 SNP is significant in predicting NHL. The haplotypes were reconstructed from H2AX and IL10 SNPs using three different reconstruction methods, then analyzed using seven logistic regression methods, and adjusted for non-SNP variables of age, sex, region of residence and ethnicity for each regression. All regression methods indicated that the three-locus H2AX haplotypes ATA and ACA were significantly associated with NHL (table 4.6), both having a protective effect with a relative risk approximately 0.5 and 0.82, respectively, compared with the baseline haplotype GCA. All methods also indicated that the IL10 haplotypes TG and AA were significant in increasing risk of NHL, each with relative risk approximately 1.5 and 2.9, respectively (table 4.8).

An additional analysis in which the third locus of the H2AX SNPs, H2AX-12, was dropped from haplotype analysis because it wasn't in HWE was done. Also, there were very few individuals who had a G allele in the third position and the haplotypes with a G in the third position failed to have a significant association with NHL, compared with the baseline haplotype. The two-locus H2AX haplotypes AT and AC were significant with relative risks approximately 0.58 and 0.82 (table 4.9). We may hypothesize further and say that since the second SNP locus, H2AX-11 wasn't significant in SNP analysis that it is H2AX-8 that is the main SNP associated with NHL through close linkage with a disease-causing mutation. The A allele is the ancestral allele for the H2AX-8 locus; when the chimpanzee genome was investigated there was an A allele in the same position. It could be argued that G arose sometime during evolution between chimp and human and is in linkage disequilibrium with something that is causing NHL since the H2AX-8 SNP is non-coding and doesn't appear to have any properties that would make it causal.

Comparison of standard errors after using regular weighted logistic regression methods and Hapassoc or Haplo.stats EM regression which inflates standard errors showed small increases in standard errors of small-frequency haplotypes for the EM regressions.

Regression coefficients were similar for both Haplo.stats and Hapassoc EM weighted logistic regressions and Haplo.stats and Hapassoc final weighted logistic regressions. Convergence criterion for the Hapassoc package has been made more strict and is currently available in a newer version of the package, a result of this convergence criterion update is that the Hapassoc regression estimates will closer reflect those calculated by the Haplo.stats package. Regression coefficients were biased for the PHASE weighted regression and the Hapassoc initial weighted regression, because weights were used that were calculated only using SNP data.

6.2 Simulation

A data set of haplotypes was generated for 50,000 individuals with three-locus haplotypes. The haplotype frequencies were distributed over the eight possible haplotypes. Outcome data was generated for the data set and the second-most frequent haplotype was deemed the "affected" haplotype.

Only the PHASE reconstructed method and Haplo.stats method were able to be compared for a data set of 50,000 individuals, since Hapassoc had memory restrictions. The PHASE weighted regressions showed a slight bias to the coefficients, generally underestimating the coefficients given by logistic regression using the true haplotypes. The PHASE "best" and both Haplo.stats methods approximated the true haplotype regression coefficients and the standard errors adequately. The standard errors were all the same to three decimal places.

In the 100 regressions of 500 individuals, all three methods were able to be compared. The regression coefficients for the Hapassoc initial weighted regressions and the PHASE weighted regressions were underestimated. The weights for both methods were calculated without incorporating the case-control information for the subjects and this appeared to have a bias effect on the estimates.

Both Hapassoc and Haplo.stats EM weighted packages had more accurate standard errors than PHASE weighted regression, Hapassoc weighted regression or Haplo.stats weighted regression. The EM weighted regressions had mean(SE(β)) columns that more closely approximated the SD(β) columns, while the regular weighted regressions had smaller standard errors and didn't approximate the population standard deviations as well. The standard error of the pooled variable in the PHASE "best" regression (table 5.3) is extremely large, which can be explained by the small frequency with which it occurs. Overall, PHASE "best" regression handles small frequency haplotypes poorly and it fails to take all haplotype information for an individual into account.

6.3 Conclusions

Haplotype research is becoming an important tool in genetic analysis. In this paper I reviewed the methodology for haplotype reconstruction and explored different methods of analysis using real and simulated data.

Analysis of NHL data from the BC Cancer Agency offered little insight into the implications of not accounting for haplotype ambiguity when doing a case control analysis of haplotypes. The amount of ambiguous and missing data present in the data set did little to affect the outcomes when investigating the differences in estimates and associated standard errors. Different methods of reconstruction and regression didn't affect significance testing outcomes, although a data set with more ambiguity might inflate standard errors enough to affect whether a test of significance has a different outcome using different methods.

Using simulated data it was found that an implication of not inflating standard errors in association studies of haplotypes to account for extra ambiguity of computationally reconstructed haplotypes is the risk of declaring a false positive significant result, and having incorrectly calculated confidence intervals. This is most apparent when there are many possible small-frequency haplotypes, a significant amount of ambiguous genetic data and a significant amount of missing genetic data. Hapassoc and Haplo.stats EM weighted regressions inflate the errors the most under these conditions and it appears that under these conditions that mistakes in significance testing would be the most likely.

The Hapassoc and Haplo.stats EM packages most closely approximate the population standard deviations for regression coefficients of haplotypes. Using these packages would result in the best analysis outcome. The PHASE reconstruction method is commonly used by researchers today, but statistical analysis done with PHASE weighted output can result in biased regression coefficients and underestimated standard errors and PHASE "best" regression handles rare frequency haplotypes poorly. It is advisable for association studies to use statistical packages that inflate standard errors in order to take genetic ambiguity into account to avoid false positive results and incorrectly calculated statistics. Even though there doesn't appear to be a large impact for standard errors for data that doesn't have as much ambiguity, it is recommended that proper methods be used at all times to analyze genetic data.

Bibliography

[1] Abbas A.K., Lichtman A.H., Pober J.S. 1994. Cellular and Molecular Immunology. 2nd Edition. W.B. Saunders Company.

[2] Baris D., Zahm S.H. 2000. Epidemiology of lymphomas. *Curr. Opin. Oncol.* 12(5):383-94.

[3] Burkett K. 2000. Logistic Regression with Missing Haplotypes. *Master's project*

[4] Burkett K., McNeney B., Graham J. 2004. A note on inference of trait associations with SNP haplotypes and other attributes in generalized linear models. *Human Heredity* 57(4):200-6.

[5] Cantor KP, Blair A, Everett G, et al. 1988. Hair-dye use and risk of leukemia and lymphoma. *Am J Public Health* 78:570-571.

[6] Cartwright, R. 1992. Changes in the descriptive epidemiology of non-Hodgkin's lymphoma: a review. *Cancer Res.* 52(suppl): 5441s-5442s

[7] Cartwright R., http://www.lymphoma.org.uk/support/Factfiles/CausesofnonHodgkinlymphoma.htm

[8] Celeste A., Petersen S., Romanienko P.J., Fernandez-Capetillo O., Chen H.T., Sedelnikova O.A., Reina-San-Martin B., Coppola V., Meffre E., Difilippantonio M.J., Redon C., Pilch D. R., et al. 2002. Genomic instability in mice lacking histone H2AX. *Science* 296: 922-927.

[9] Chen H.T., Bhandoola A., Difilippantonio M.J., Zhu J., Brown M. J., Tai X., Rogakou E.P., Brotz T.M., Bonner W.M., Ried T., Nussenzweig A. 2000. Response to RAG-mediated V(D)J Cleavage by NBS1 and gamma-H2AX. *Science* 290: 1962-1964.

[10] Christopherson K.W. II, Hromas R. 2004. Chemokine Regulation of Normal and Pathologic Immune Responses. *Stem Cells* 19(5):388-396.

[11] Clark AG. 1990. Inference of Haplotypes from PCR-amplified Samples of Diploid Populations. *Mol. Biol. Evol.*, 7:111-122.

[12] http://cran.r-project.org/

[13] Devesa SS, Silverman DT, Young JL, et al. 1987. Cancer incidence and mortality trends among whites in the United States. *J Natl Cancer Inst* 79:701-770

[14] Devesa S., Fears T. 1992. Non-Hodgkin's Lymphoma Time Trends: United States and International Data. *Cancer Res.* 52(suppl): 5432s-5440s.

[15] Excoffier L. and Slatkin M. 1995. Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population. *Mol. Biol. Evol.*, 12(5):921-927.

[16] Fallin D. and Schork N.J. 2000. Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data. *Am. J. Hum. Genet.* 67: 947-959.

[17] Hardell L, Eriksson, Lenner P, et al. 1981. Malignant lymphoma and exposure to chemicals, especially organic solvents, chlorophenols and phenoxy acids: a case-control study. *Br J Cancer* 43:169-176

[18] Hardy, G. 1908. Mendelian Proportions in a Mixed Population. *Science* 28:49-50

[19] Hawley M. and Kidd K. 1995. HAPLO: A Program Using the EM Algorithm to Estimate the Frequencies of Multi-site Haplotypes. *J. Hered.* 86: 409-411.

[20] Hjalgrim H., Frisch M., Begtrup K., et al. 1996. Recent Increase in the Incidence of Non-Hodgkin's Lymphoma Among Young Men and Women in Denmark. *Br. J. Cancer.* 73: 951-954.

[21] Horton N.J., and Laird N.M. 1999. Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research* 8: 37-50.

[22] Ibrahim J.G. 1990. Incomplete Data in Generalized Linear Models. *Journal of the American Statistical Association.* 85: 765-769.

[23] Lake S.L., Lyon H., Tantisira K., Silverman E.K., Weiss S.T., Laird N.M., Schaid D.J. 2003. Estimation and Tests of Haplotype-Environment Interaction when Linkage Phase Is Ambiguous. *Hum. Hered.* 55:56-65.

[24] Lim D.S., Kim S.T., Xu B., Maser R.S., Lin J., Petrini J.H.J., Kastan M. B. 2000. ATM phosphorylates p95/nbs1 in an S-phase Checkpoint Pathway. *Nature* 404: 613-617.

[25] Long J., Williams R. and Urbanek M. 1995. An E-M Algorithm and Testing Strategy for Multiple-Locus Haplotypes. *Am. J. Hum. Genet.* 56:799-810.

[26] Louis T.A. 1982. Finding the Observed Information Matrix when Using the EM Algorithm. *J. R. Statist. Soc. B* 44(2):226-233.

[27] Meilijson, I. 1989. A Fast Improvement to the EM Algorithm on its Own Terms. *J. R. Statist. Soc. B* 51(2):127138.

[28] Miller BA, Ries LAG, Hankey BF et al (eds.) 1993. SEER Cancer Statistics Review: 1973-1990. National Cancer Institute. NIH Publication No. 93-2789.

[29] Muir C, Waterhouse J, Mack T, et al. 1987. Cancer Incidence in Five Continents Vol. V IARC Scientific Publication No. 88. Lyon, International Agency for Research on Cancer.

[30] National Cancer Institute (NCI) 1982. Non-Hodgkin's lymphoma pathologic classification project writing group: National Cancer Institute sponsored study of classifications of non-Hodgkin's lymphomas. Summary and description of a working formation for clinical usage. *Cancer* 49:2112-2135.

[31] National Cancer Institute of Canada. 2001 *Canadian Cancer Statistics. Toronto.*

[32] Niu T., Qin Z.S., Xu X., Liu J.S., 2002. Bayesian Haplotype Inference for Multiple Linked Single-nucleotide Polymorphisms. *Am. J. Hum. Genet.* 70:157-169.

[33] Pallesen G, Hamilton-Dutoit SJ, Zhou X, 1993. The association of Epstein-Barr Virus (EBV) with T cell lymphoproliferations and Hodgkin's disease: Two new developments in the EBV field. *Adv Cancer Res.* 62:179-239.

[34] Ries L., Eisner M., Kosary C., et al. 2001. SEER Cancer Statistics Review, 1973-1998. National Cancer Institute. Bethesda, MD, http:seer.cancer.gov/Publications/CSR1973_1998.

[35] Schaid D.J., Rowland C.M., Times D.E., Jacobson R.M. and Poland G.A. 2002. Score Tests for Association Between Traits and Haplotypes when Linkage Phase is Ambiguous. *Am. J. Hum. Genet.* 70:425-434.

[36] Stephens M., Donnely P. 2000. Inference in Molecular Population Genetics. *J. R. Stat. Soc. B* 62:605-655.

[37] Stephens M., Smith N.J. and Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68: 978-979.

[38] Weinberg, W. 1908. On the Demonstration of Heredity in Man. *Naturkunde in Wurttemberg, Stuttgart* 64:368-382.

[39] Zoloth SR, Michaels DM, Villalbi JR, et al. 1986. Patterns of mortality among commercial pressmen. *J Natl Cancer Inst* 76:1047-1051