

12-2011

Development of a Bayesian Joint Logistic Model to Better Study the Association between Haplotypes and Disease

Anthony M. D'Amelio Jr

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations



Part of the [Biostatistics Commons](#), [Statistical Models Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

D'Amelio, Anthony M. Jr, "Development of a Bayesian Joint Logistic Model to Better Study the Association between Haplotypes and Disease" (2011). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 197.
https://digitalcommons.library.tmc.edu/utgsbs_dissertations/197

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

DEVELOPMENT OF A BAYESIAN JOINT LOGISTIC MODEL
TO BETTER STUDY THE ASSOCIATION
BETWEEN HAPLOTYPES AND DISEASE

by

Anthony M. D'Amelio Jr., B.S.E.

APPROVED:

Carol J. Etzel, PhD, Supervisory Professor

Sanjay Shete, PhD

Randa El-Zein, MD, PhD

Christopher Amos, PhD

Wayne Newhauser, PhD

APPROVED:

Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

DEVELOPMENT OF A BAYESIAN JOINT LOGISTIC MODEL
TO BETTER STUDY THE ASSOCIATION
BETWEEN HAPLOTYPES AND DISEASE

A DISSERTATION

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

M. D. Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Anthony M. D'Amelio Jr., B.S.E.

Houston, Texas

December 2011

Dedication

To my family, especially my mom and dad, who supported my intellectual and personal journey to become a better scientist and become Dr. D'Amelio.

Acknowledgements

I would like to thank my mentor, Dr. Carol J. Etzel for her strong support in my development as a biostatistician, scientist, and critical thinker. When I first arrived in her laboratory in September 2006 on a part-time basis, she introduced me to projects which dramatically increased my ability in biostatistics and programming. It was her studies in risk modeling which interested me enough to become a full-time graduate research assistant in her lab in June 2007. Specifically, it was her idea for me to study admixture mapping that initially started my interest in genetic analysis. After a six-month short term fellowship in admixture, I was able to construct a dissertation topic, and now this thesis will reflect what I have learned and studied in my tenure with Dr. Etzel's laboratory. Also, I would like to thank her for her support in more personal matters. When I was struggling to prepare for my candidacy exam, and I had a death in my family, she gave me the support that I needed to both handle this personal issue and still prepare for my candidacy exam. Without her support during that time, I probably would have not passed the candidacy exam.

I would also like to thank the rest of my supervisory committee members as well. I would like to thank Dr. Christopher Amos for helping me to better develop the thesis so that it is a completely new approach to modeling and genetics. I would like to thank Dr. Sanjay Shete for his insightful questions and critical examinations of my work which helped to make my scientific contributions and presentation style better. I would like to thank Dr. Randa El-Zein for her support throughout my graduate school career, and also access to the Hodgkin disease dataset. I would like to thank Dr. Wayne Newhauser for his comments in making presentations, which helped me make my presentations more accessible to a wider audience, as well as giving suggestions to improve both my thesis and my overall development as a scientist.

I would like to also thank my laboratory group as a whole. They have helped me become a better statistician due to their experience in the field, and making me think critically during my development. They have helped me become a better-rounded scientist and I will always be grateful for their advice. Also, they have helped me think more outside of the box as I am someone who tends to get locked into a specific way of thinking (both professionally and personally) and their guidance has made me more introspective, which has helped my growth as a scientist and a person.

Many individuals outside of academia have helped me become a better person and scientist throughout the years as a grad student, especially my immediate family and friends. I would like to thank my immediate family and friends for supporting my move to Texas to become a graduate student, and also being there to talk to when I was struggling both professionally and personally, which happened numerous times during this journey. Finally, I would like to thank the Lummis family, specifically Marilyn Lummis, for her generosity through the Marilyn and Frederick R. Lummis Jr, M.D., fellowship in Biomedical Sciences, which has funded all research in this dissertation since January 1st, 2011.

Abstract

DEVELOPMENT OF A BAYESIAN JOINT LOGISTIC MODEL TO BETTER STUDY THE ASSOCIATION BETWEEN HAPLOTYPES AND DISEASE

Publication No. _____

Anthony M. D'Amelio Jr., B.S.E

Supervisory Professor: Carol J. Etzel, PhD.

In 2011, there will be an estimated 1,596,670 new cancer cases and 571,950 cancer-related deaths in the US. With the ever-increasing applications of cancer genetics in epidemiology, there is great potential to identify genetic risk factors that would help identify individuals with increased genetic susceptibility to cancer, which could be used to develop interventions or targeted therapies that could hopefully reduce cancer risk and mortality.

In this dissertation, I propose to develop a new statistical method to evaluate the role of haplotypes in cancer susceptibility and development. This model will be flexible enough to handle not only haplotypes of any size, but also a variety of covariates. I will then apply this method to three cancer-related data sets (Hodgkin Disease, Glioma, and Lung Cancer). **I hypothesize that there is substantial improvement in the estimation of association between haplotypes and disease, with the use of a Bayesian mathematical method to infer haplotypes that uses prior information from known genetics sources.**

Analysis based on haplotypes using information from publically available genetic sources generally show increased odds ratios and smaller p-values in both the Hodgkin, Glioma, and Lung data sets. For instance, the Bayesian Joint Logistic Model (BJLM) inferred haplotype TC had a substantially higher estimated effect size (OR=12.16, 95% CI = 2.47-90.1 vs. 9.24, 95% CI

= 1.81-47.2) and more significant p-value (0.00044 vs. 0.008) for Hodgkin Disease compared to a traditional logistic regression approach. Also, the effect sizes of haplotypes modeled with recessive genetic effects were higher (and had more significant p-values) when analyzed with the BJLM. Full genetic models with haplotype information developed with the BJLM resulted in significantly higher discriminatory power and a significantly higher Net Reclassification Index compared to those developed with *haplo.stats* for lung cancer.

Future analysis for this work could be to incorporate the 1000 Genomes project, which offers a larger selection of SNPs can be incorporated into the information from known genetic sources as well. Other future analysis include testing non-binary outcomes, like the levels of biomarkers that are present in lung cancer (NNK), and extending this analysis to full GWAS studies.

Table of Contents

Dedication	iii
Acknowledgements	iv
Abstract	vi
List of Tables	xiv
List of Figures	xvii
Chapter 1: Introduction	1
1.1. What is Risk Modeling in the Clinical Construct?	3
1.1.1 Well-known Cardiovascular Risk Models	3
1.1.2 Risk Modeling in Cancer: An Overview	5
1.1.3 Risk Modeling in Cancer: Evolution of Lung Cancer Modeling	6
1.2. Genetic Analysis: Introduction to Haplotypes	11
1.2.1 Microsatellites to Single Nucleotide Polymorphisms (SNPs): The Beginning of Statistical Genetic Analysis	11
1.2.2 Introduction to Linkage Disequilibrium and Haplotypes	15
1.2.3 Introduction for Haplotype Analysis	18
1.3. Overview of Dissertation Thesis	20
Chapter 2: Validation and Calibration of the Spitz Lung Cancer Risk Model	22
2.1. Introduction to the NELSON data set	22
2.1.1 Relationship between NELSON and NLST	22
2.1.2 Testing the Spitz Model with NELSON Data	23
2.2. Methods for Conducting Spitz Model Validation with NELSON Data	24
2.2.1 NELSON Study Population	24
2.2.2 Determination of 5-year Absolute Risk for Lung Cancer	25

2.2.3. Estimation of Calibration and Discriminatory Power for NELSON analysis	27
2.2.4. Estimation of clinical utility for NELSON analysis	28
2.3. NELSON Data Results	28
2.4. Discussion of NELSON Results	34
Chapter 3: Expansion of Existing Risk Models Using Genetic Factors	39
3.1. Incorporating SNPs into Cancer Risk Models: A Primer	39
3.2. Materials and Methods to add SNPs to Risk Models	41
3.2.1. Study Population for MD Anderson Lung Cancer Study	41
3.2.2. Incorporating Genetic Information into Risk Modeling	41
3.3. Risk Model Results for Adding Genetic Information to Spitz models	42
3.3.1. Results: 1 st Model: Top SNP + Spitz Original Model	42
3.3.2. Results: 2 nd Model: SNPs from Chromosomes 15, 5, and 6 + Spitz Original Model	44
3.3.3. Results: 3 rd Model: Haplotypes from Chromosome 5 and 15 + Original Spitz Model	45
3.3.4. Results: Comparison of three genetic model extensions	47
3.4. Discussion: Adding Haplotypes and SNPs to the Spitz Lung Cancer Risk Model	50
3.5. Why we need the BJLM	52
Chapter 4: Development of the Bayesian Joint Logistic Method	53
4.1. Relationship between the BJLM and Haplotype Analysis	53
4.2. Development of the framework for the BJLM	56

4.2.1. Extracting information from HapMap for BJLM use	56
4.2.2. Determining Haplotypes from Full Data Using Information from HapMap	61
4.2.3. Using a Modified Forward-Backward Algorithm to Infer Haplotypes for Individuals with Missing SNP Data	63
4.2.4. Development of the Bayesian Logistic Model for Haplotype Association	67
4.3. Haplotype Simulations to Test Ability of BJLM to infer Haplotypes	70
4.3.1. Procedure for Haplotype Simulations	70
4.3.2. Haplotype (2 SNP) Simulation Results	72
4.3.3. Haplotype (3 SNP) Simulation Results	74
4.3.4. Haplotype (12 SNP) Simulation Results	77
4.4. BJLM Creation Conclusion and Application to Real-World Data Sets	80
Chapter 5: Application One of the BJLM: Using Haplotype Analysis To Elucidate Significant Associations between Genes and Hodgkin Disease	82
5.1. Introduction to Hodgkin Disease	82
5.2. Materials and Methods in Hodgkin Disease Study	84
5.2.1. Study Population for Hodgkin Disease Study	84
5.2.2. SNP Selection for Hodgkin Disease Study	84
5.2.3. Haplotype Analysis for Hodgkin Disease Study	85
5.2.4. Developing an Experimental Risk Model with the BJLM	86
5.3. Hodgkin Disease Results	88
5.3.1. Haplotypes and Haplotype Frequencies for Hodgkin Cases and Controls	88

5.3.2. Determining Associations between Haplotypes and Hodgkin Disease	91
5.3.3. Incorporation of the BJLM for Hodgkin Disease Study	92
5.4. Hodgkin Disease Study Discussion	94
Chapter 6: Application Two of the BJLM: Elucidating Significant Haplotype Associations between Genes and Lung Cancer	98
6.1. Introduction to Expanding Spitz Lung Cancer Risk Model with Haplotypes	98
6.2. Methods to Conduct Expansion of Spitz Model with Top SNPs from Lung meta-analysis	99
6.2.1. Study Population and Selection of SNPs for Expansion of Spitz Model with Haplotypes	99
6.2.2. Selecting the Haplotype Blocks for Future Analysis	100
6.3. Results for Expanding Spitz Models Using SNPs derived from the Top 200 SNP list in the Lung Cancer Meta-Analysis	106
6.3.1. Univariate Haplotype Block Analysis to Extend Spitz Model	106
6.3.2. Developing New Spitz Lung Cancer Risk Models	109
6.3.3. Comparing the Updated Spitz Lung Cancer Risk Models	111
6.4. Discussion for Expanding Spitz Models Using SNPs derived from the Top 200 SNP list in the Lung Cancer Meta-Analysis	113
Chapter 7: Application Three of the BJLM: Using Haplotype Analysis and the BJLM to Determine Significant Genetic Risk Factors for Glioma	117
7.1. Introducing Glioma for Genetic Analysis	118
7.2. Methods for Glioma Analysis	119
7.2.1. Study population and Selection of SNPs for Glioma Analysis	119

7.2.2.	Determining top SNPs for Model Analysis	120
7.2.3.	Determining top Haplotypes for Glioma Model Analysis	121
7.2.4.	Combining both SNP and Haplotype Data	122
7.3.	Results for Glioma Analysis	123
7.3.1.	Single SNP model development	123
7.3.2.	Haplotype model development for Glioma Data Set	127
7.3.3.	Model development with both SNPs and Haplotypes for Glioma	130
7.3.3.1.	Best SNPs + Best Haplo.stats Inferred Haplotypes for Glioma	130
7.3.3.2.	Best SNPs + Best BJLM Inferred Haplotypes for Glioma	133
7.4.	Discussion of Glioma Modeling	136
Chapter 8:	Summary and Future Analysis	140
8.1.	Summary	140
8.1.1.	Summary of Spitz Lung Cancer Extension	140
8.1.2.	Summary of Why We Need to Better Incorporate Haplotypes into Risk Models	141
8.1.3.	Summary of Developing Risk Models with BJLM Inferred Haplotypes	142
8.2.	Future Analysis	142
8.2.1.	Incorporate 1000 Genomes as External Data into the BJLM	144
8.2.2.	Incorporate Non-Binary Outcomes	147
8.2.3.	Extend Analysis to Full GWAS studies and Final Conclusion	147

References	149
Appendix 1: Incidence rates and mortality rates used to calculate absolute risk in the validation of the Spitz, LLP, and Bach models for Chapter 2	179
Appendix 2: Simulating Missing Data Results for all Haplotypes > 5% Frequency in the BJLM Simulation Study for Haplotypes of Length 4 to 11	180
Appendix 3: Chapter 7 (Glioma Development)	196
Vita	201

LIST OF TABLES

Table 1.1: Populations and List of populations available for analysis in HapMap	14
Table 2.1: Lifestyle Variables Used in Original Spitz Model and in Validation of Spitz Model with NELSON data	27
Table 2.2: Demographic characteristics of the NELSON population	29
Table 2.3: Calibration Details for Each Risk Level	30
Table 3.1: Multivariate analysis of Spitz Lung Cancer Model Risk Factors plus SNP rs8034191	43
Table 3.2: Multivariate analysis of Spitz Lung Cancer Model Risk Factors plus three SNPs: rs2736100, rs7718456, and rs1051730	44
Table 3.3: Haplotype Frequencies for Top SNPs in GWAS that are in High Linkage Disequilibrium with Each Other According to Texas lung GWAS Study	45
Table 3.4: Multivariate analysis of Spitz Lung Cancer Model Risk Factors plus haplotypes from Chromosome 15 and Chromosome 5	46
Table 3.5: Discriminatory power for the three genetic models and the Original Spitz Model	48
Table 3.6: Net Reclassification Index results for comparing the three extensions of the Spitz risk models with the Original Spitz Model	49
Table 4.1: Marker Information for the 12 SNPs in the Simulation Analysis	70
Table 4.2: Haplotype Simulation results for 2 SNP haplotypes	72
Table 4.3: Summary of Simulated haplotypes of 3 SNP length	75
Table 4.4: Summary of Simulated haplotypes of 12 SNP length	78
Table 5.1: Location and rs number for each Single Nucleotide Polymorphism	85

Table 5.2: Estimated frequencies for haplotypes on chromosomes 1-3, 5-7, 10, 16, and 19	89
Table 5.3: Determining associations between haplotypes and HL using the joint effect logistic model controlling for age, sex, race, and smoking status	91
Table 5.4: Haplotype model results using haplotypes inferred with both haplo.stats and BJLM assuming an additive genetic model	93
Table 5.5: Estimation of Odds Ratios for haplotypes using both haplo.stats and the BJLM	94
Table 6.1: Chromosome and Base-Pair Information for all SNPs included in the expansion of the Spitz Model	102
Table 6.2: Individual haplotype block analysis using haplo.stats and the BJLM for former smokers	107
Table 6.3: Individual haplotype block analysis using haplo.stats and the BJLM for former smokers	108
Table 6.4: Full Updated Spitz Lung Cancer Risk Model Development Using Haplotypes inferred with Haplo.stats and the BJLM	110
Table 6.5: Discriminatory Power Results for the Original Spitz Lung Model, the Spitz model with Haplo.stats inferred haplotypes, and the Spitz Model with BJLM inferred haplotypes	112
Table 6.6: Net Reclassification Index results for comparing the Original Spitz Lung Model, the Spitz model with Haplo.stats inferred haplotypes, and the Spitz Model with BJLM inferred haplotypes	113
Table 7.1: Full Single SNP model using Inflammation SNPs for Glioma	126

Table 7.2: Best Haplotypes within Glioma Inflammation Dataset as Calculated by Haploview	127
Table 7.3: Multivariate logistic regression for inferred haplotypes from both BJLM and Haplo.stats	129
Table 7.4: Best SNPs plus Haplo.stats inferred haplotypes for full genetic Glioma model	131
Table 7.5: Complete Genetic Risk Model with SNPs and Haplotypes modeled together using WINBUGS	134

LIST OF FIGURES

Figure 1.1: Example of a microsatellite with four ACG repeats on DNA	11
Figure 1.2: Single Nucleotide Polymorphisms from 12 areas on the chromosome	13
Figure 1.3: Actual Haplotypes from five SNPs associated with lung cancer in the study at MD Anderson Cancer Center	17
Figure 1.4: Using Haploview to determine a haplotype from Chromosome 1 on the genome	19
Figure 2.1: Calibration Results with 95% CI's for each median risk level	30
Figure 2.2: Clinical Utility Results in Graphical Form for the Spitz model	32
Figure 2.3: Expanded Clinical Utility for 5-year Absolute Risks from 1% to 5%	33
Figure 3.1: Gold Plot from Haploview for the SNPs in HapMap between rs2736100 and rs401681	51
Figure 4.1: Flow-Chart of the BJLM from the 1 st step: HapMap to the Final Step: Creation of a Bayesian Logistic Model	55
Figure 4.2: Sample section from HapMap genotype data dump	57
Figure 4.3: Output Example with both Missing SNP Data and without Missing SNP Data	59
Figure 4.4: Algorithm that determines the Haplotypes inferred from HapMap individuals that is used for the BJLM	60
Figure 4.5: Determining Missing Haplotype Pair from individuals with missing SNP data with the Forward-Backward Algorithm	64
Figure 4.6: Graphical representations of the haplotype frequency for simulations of 2 SNP Haplotypes	73

Figure 4.7: Graphical representations of the haplotype frequency for simulations of 3 SNP Haplotypes	76
Figure 4.8: Graphical representations of the haplotype frequency for simulations of 12 SNP Haplotypes	79
Figure 5.1: Sample Code for WINBUGS for conducting Haplotype Analysis	87
Figure 5.2: Chromosome 1 and Chromosome 16 Gold Plots from Hodgkin Study	90
Figure 6.1: Haploview Gold Plot showing a haplotype block from HapMap data that contains three SNPs from both the top 200 SNP list and the Texas lung GWAS from MD Anderson	101
Figure 6.2: Example WinBugs Code for the Development of a Bayesian Logistic Risk Model Incorporating Haplotypes inferred from the BJLM	105
Figure 7.1: Manhattan Plot showing the 4647 SNPs examined for Glioma Analysis	124
Figure 8.1: Process for Extracting 1000 Genomes Data for Future Analysis	146

Chapter 1: Introduction

The objective of this thesis is to develop valid risk models combining both genetic and non-genetic information to better model associations with disease, increase discriminatory power, and increase accuracy and clinical utility. Discriminatory power is the ability of the model to correctly differentiate between case and control, and with improved risk modeling, discriminatory power will increase. Clinical utility refers to the effectiveness of the model's use in a clinical setting. A more powerful model, in terms of discriminatory power, will identify more cases correctly labeled as affected compared to a true control incorrectly labeled affected. Genetic information in this thesis will consist of linked sets of single nucleotide polymorphisms (SNPs), which are specific locations within the genetic sequence of an individual where differences can arise from one person to another. Linked sets of SNPs on the same chromosome are also called haplotypes.

Haplotype analysis, defined as the study of a set of linked alleles occurring on the same chromosome has been used to discover sets of linked markers associated with specific diseases. With haplotype analysis, more power can be obtained in discovering a link between haplotypes and disease compared to the case of just examining the relationship between an individual SNP and disease. After determining haplotypes in the data set, we can then determine the associations between haplotypes and disease, with the use of a logistic regression model. Each haplotype is compared to the most frequent haplotype, as the most frequent haplotypes represent what would be more often seen in nature. This is also called joint logistic analysis.

Haplotypes have the potential to be viable markers for the early prediction of cancer development, and if not directly observed, can be inferred from sets of linked single nucleotide polymorphisms (SNPs) by either using frequentist mathematical methods or by using Bayesian

mathematical methods. Bayesian mathematical methods for haplotype inference differ from frequentist methods because in Bayesian methods, priors (either informative or non-informative) can be used to determine the haplotype frequency. Currently there these exist two prominent frequentist methods for risk model development with inferred haplotypes (Joint Logistic Analysis, Separate-Effects Model)^{62,65}; however, no Bayesian method currently exists that not only infers haplotypes, but also can directly elucidate risk models that can be used to develop genetic risk profiles for disease.

I propose to develop a new statistical method using Bayesian theory to evaluate the role of haplotypes in cancer susceptibility and development. This model will be flexible enough to handle not only haplotypes of any size, but also a variety of covariates; whether they are binary, continuous, or categorical. I will then apply this method to cancer-related data sets to evaluate genetic susceptibility. Haplotype analysis will be improved by directly incorporating previous published genetic information from HapMap, which is a large online collection of over 1 million SNPs from 11 different populations. **I hypothesize that there is substantial improvement in the association between haplotypes and disease, with the use of a Bayesian mathematical method to infer haplotypes using priors developed from publicly available genetics sources.** In order to account for haplotypes in disease association studies and to maximize the full potential of haplotypes in such studies, I propose three specific aims: **1) to develop and test a novel Bayesian-based method, namely the Bayesian Joint Logistic Model (BJLM) to improve power to detect association between SNPs haplotypes constructed from single nucleotide polymorphisms (SNPs); 2) to incorporate the BJLM to develop risk models for a variety of diseases; 3) and, re-incorporate SNPs into risk models with the haplotypes inferred from the BJLM to maximize the effect of genetics.**

All specific aims were accomplished with a Hodgkin data set from Dr. Randa El-Zein, a Glioma data set from Dr. Melissa Bondy, and a Lung Cancer GWAS set from Dr. Christopher Amos. All three of these data sets were developed at M.D. Anderson Cancer Center. The end result was a new Bayesian logistic model that evaluates the relationship between haplotypes and disease in the presence of possible covariates.

This thesis is innovative, and has five specific innovations for the scientific community. First, I expand genetic analysis to include possible associations for disease from haplotypes instead of just hits from individual SNPs. Second, I develop and implement a forward-backward algorithm to infer haplotypes in the presence of missing SNP information (all or partial) by directly using the haplotypes that have been inferred with complete SNP data. Third, I develop priors for Bayesian Analysis from HapMap data with the use of a Expectation Maximization Algorithm. Fourth, I create the Bayesian Joint Logistic Method to allow for development of risk models with haplotypes. Finally, I incorporate different sets of priors for different types of covariates (binary, categorical, continuous) in the development of risk models with haplotypes.

Before developing the BJLM, I will review some necessary background concepts such as risk modeling, genetic analysis and especially the use of haplotypes in genetic analysis.

1.1. What is Risk Modeling in the Clinical Construct?

1.1.1 Well-known Cardiovascular Risk Models

Risk models have been used to estimate risk in a wide variety of complex diseases, from cardiovascular disease to lung cancer. For instance, the Reynolds risk score¹, is a cardiovascular disease risk model for women developed at Harvard University. With this risk model, 24,558 healthy women (with no immediate signs of cardiovascular disease) were followed for a median of 10.2 years for specific cardiovascular diseases such as coronary revascularization, myocardial

infarction, ischemic stroke, and cardiovascular death¹. Development of the Reynolds risk score involved use of two-thirds of the study population (n = 16,400) with validation occurring in the other third of the study population (n = 8158)¹. The Reynolds risk score contains the risk variables of age, systolic blood pressure, high sensitivity C-reactive protein, total cholesterol, high-density lipoprotein cholesterol, current smoking status, and family history of heart disease for those under the age of 60¹. The calculation of 10 year cardiovascular risk, according to this model, is¹:

$$R (\%) = 100 * (1 - 0.98634(e^{B-22.325})) \quad (1.1)$$

where B is a factor that aggregates the effects of the risk variables listed above¹. In the next few paragraphs, I will discuss some prominent examples of risk modeling.

The most well-known example of risk modeling for clinical purposes are the cardiovascular disease risk models developed from the Framingham Health Study². Initial enrollment for this study began in the time period between 1948 and 1950 with inclusion of 5,127 men and women aged 30-62 with no coronary artery disease. A second and third generation cohort was introduced in 1971 and 2002². The first risk model produced from the Framingham cohort data was published in 1998, and it calculated the 10-year probability of coronary heart disease⁴. With the coronary heart disease model, both a simple model based on a point system for risks, and a much more complicated model based on relative risks, were developed to predict 10-year risk³. Other risk models using the Framingham Heart Study data studied cardiovascular diseases like atria fibrillation⁴, congestive heart failure⁵, general cardiovascular disease⁶, hard coronary heart disease⁷, and stroke⁸.

1.1.2. Risk Modeling in Cancer: An Overview

Risk modeling has become an ever increasing tool in cancer epidemiology and cancer prevention. The Gail model for breast cancer was first introduced in 1989, and contains variables such as age (as defined as those before or older than 50 years old), age at menarche, previous breast biopsies, age for 1st born child, and family history of breast cancer⁹. The risk for breast cancer are calculated below in equation 1.2⁹,

$$\begin{aligned} \log(\text{Breast Cancer odds}) &= -0.74948 + 0.09401(\text{Agemen}) + 0.52926(\text{Nbiops}) \\ &+ 0.218643 (\text{Ageflb}) + 0.95830 (\text{Numrel}) + 0.01081 (\text{Agecat}) \\ &- 0.28804(\text{Nbiops} * \text{Agecat}) - 0.19081(\text{Ageflb} * \text{Numrel}) \end{aligned} \quad (1.2)$$

where Agemen = Menarche age, Nbiops = Previous breast biopsies, Ageflb = age for 1st born child, Numrel = family history of breast cancer, and Agecat = aged greater than 50.

This model was validated in Caucasian females, and the Gail model performed well within specific risk factor strata and for individuals over the age of 60¹⁰. With the Gail model, high-risk individuals for breast cancer were recommended for more extensive screening^{11,12}. Gail then created a new breast cancer model for African-Americans named CARE, based from the Women's Contraceptive and Reproductive Experiences study, because of differences in risk profiles between Caucasian and African-American women¹³. Then, in 2005, a variable that elucidates mammographic breast density was added to the Gail model by Jeffrey A. Tice's group, and its discriminatory power increased modestly from 67% (95% CI = 0.65-0.68) to 68% (95% CI = 0.66-0.70) using data from the San Francisco Mammography Registry¹⁴. With the success of the Gail model in estimating risk for breast cancer, risk models were created for other cancers, such as colorectal^{15,16}, bladder¹⁷, ovarian¹⁸, and melanoma^{19,20}.

1.1.3. Risk Modeling in Cancer: Evolution of Lung Cancer Modeling

Lung cancer risk modeling has become a viable tool to study potential risk for lung cancer, and several models have been developed. The first models for lung cancer were developed in 2000 by Dr. Graham Colditz at Harvard and it was based on different scores for an individual's environment, smoking intensity, family history of smoking, occupational history, and diet²¹. His research group included experts in the risk factors listed above, and they estimated risk for lung cancer based on consensus from experts in lung cancer oncology. The point guide for all non-occupational risk factors is listed in Table 2 of Colditz et al.²¹.

For each individual, the risk points for each applicable risk factor are added, added with risk points from occupational exposures, and then divided by the total US population risk point average to obtain a cancer risk score. This result is then given a Surveillance, Epidemiology, and End Results (SEER)²² multiplier which represents the increase or decrease in potential risk compared to the "average" individual risk profile, and then that multiplier is multiplied by the average 10-year absolute risk for lung cancer in the United States at a specified age (as determined by SEER) to determine an individual's 10-year lung cancer risk²¹. For example, if a 60-year old male individual has 50 risk points, and the US population risk point average is 20, they have a cancer risk score of 2.5. This cancer risk score of 2.5 corresponds to a SEER multiplier of three²¹, and a 60-year old male has a 0.547% chance of obtaining lung cancer in the next ten years according to SEER²¹. Therefore, this 60-year old male has a 1.641% chance of lung cancer in the next ten years according to the Colditz risk model.

The second lung cancer risk model was published by Peter Bach in 2003. Unlike the previous lung cancer model developed by the Colditz group which was developed as a consensus, this model was developed based on the placebo arm of the Carotene and Retinol

Efficacy Trial cohort (CARET)²³, which was a large randomized cohort in which 18172 individuals who were heavy smokers and aged 50-75, were tested for effect of beta-carotene on lung cancer. The CARET trial was stopped in July 1996 after an interim report showed that beta-carotene increased lung cancer risk in heavy smokers^{24,25}. The Bach model, created 1-year proportional hazard models for both incidence of lung cancer allowing for mortality from all causes other than lung cancer²⁵. These models allowed for competing causes, so absolute lung cancer risks were reduced according to the probability of non-lung cancer mortality. Key variables in the Bach models included cigarettes per day, smoking duration, quitting time (in years), age, asbestos exposure, and sex²⁵. One advantage of the Bach model is the easy calculation of a multiple of absolute lung cancer risks as the one year models only have to be added for determination of multi-year lung cancer risk prediction. Validation of the Bach model from Kronin et al. showed that the Bach model has moderate discriminatory power, varying from 0.57 (95% CI = 0.49-0.67) for those aged 65-69 to 0.77 (95% CI = 0.70-0.84) for those aged 50-54²⁶.

The third lung cancer risk model developed to examine absolute risk of lung cancer incidence is the Spitz model, which was constructed from a matched case-control study of 1851 cases and 2001 controls enrolled at MD Anderson Cancer Center from 1995-2006. Cases and controls were matched on age (± 5 years), sex, smoking status and race²⁷. The matching on smoking status restricted the model to be stratified for never, former, and current smokers. Below are the never, former, and current smoker, relative risk equations (equations 1.3 to 1.5), respectively, for Caucasian individuals, which were created for the Spitz lung cancer model²⁷,

Never Smokers:

$$\log_e(\text{Relative Risk}) = -0.8806 + (0.5874 * Ets) + (0.6954 * fh) \quad (1.3)$$

Former Smokers:

$$\begin{aligned} \log_e(\text{Relative Risk}) \\ = -0.7606 + 0.9734 * Emph + 0.4654 * Dust + 0.4636 * fh + 0.2130 \\ * quit1 + 0.4080 * quit2 - 0.3711 * Hf \quad (1.4) \end{aligned}$$

Current Smokers:

$$\begin{aligned} \log_e(\text{Relative Risk}) \\ = -0.7173 + 0.7561 * Emph + 0.3067 * Dust + 0.3859 * sfh + 0.2219 \\ * pk1 + 0.3747 * pk2 + 0.6151 * pk3 + 0.4109 * asb - 0.4047 * Hf \quad (1.5) \end{aligned}$$

where Ets = environmental tobacco smoke, fh = two or more family members with cancer, Emph = self-reported emphysema, Dust = exposure to dust variables, quit1 = quit smoking between ages 42-53, quit2 = quit smoking at age 54 or older, sfh = one or more family members with a smoking related cancer, pk1 = pack-years between 28 and 42, pk2 = pack-years between 42 and 57.5, pk3 = pack-years greater than 57.5, and asb = exposure to asbestos²⁷.

After obtaining the relative risks, the absolute-risk of lung cancer are determined by using data from the Surveillance, Epidemiology, and End Results (SEER) on lung cancer incidence, and data from the National Center for Health Statistics (NCHS) for non-lung cancer mortality, combined with the relative risks results^{22,28}. The concordance statistic (measure of discriminatory power) for the Spitz model varies from 0.59 (95% CI = 0.51-0.67) in never smokers to 0.65 (95% CI = 0.60-0.69) for current smokers. Later analysis led to the creation of lung cancer models for African-Americans and genetic extensions of the original Spitz model^{29,30}.

The Liverpool Lung Project (LLP) group in Liverpool, England developed the LLP model from a population-based case-control data set, with individuals aged 20-80, and matched by one case to two controls with both controls having the same gender and within 2 years of age with the case³¹. Unlike the Spitz model, no matching exists for smoking, hence smoking intensity is the most dominant risk factor in the model, with four different risk classifications³¹. This model differentiates between early family history of lung cancer (family members obtaining lung cancer before age 60), and later family history of lung cancer (family members obtaining lung cancer after age 60)³¹. Other risk factors include self-reported exposure to pneumonia, asbestos exposure, and previously having malignant tumors. Age and sex are incorporated with the use of five-year Liverpool incidence rates for cancer. These incidence rates allow for automatic calculation of a five-year absolute risk model for lung cancer, unlike the Bach and Spitz models, which can predict up to ten-year absolute risk calculations. With the case-control data set for which the LLP model was derived, discriminatory power results were moderately good (AUC = 0.71), and these results were similar to those found in a 10-fold cross validation analysis (AUC = 0.70)³¹.

The most recent lung cancer model is the Prostate, Lung, Colorectal, and Ovarian (PLCO) model, which was developed from 70,962 subjects from the non-screening arm section of the PLCO trial, and was validated in the screening arm section of the PLCO trial (n = 38,254)³². The PLCO model consisted of the variables age, socioeconomic status, recent chest x-ray, COPD, smoking status, smoking duration, pack-years smoked, body mass index, and family history of lung cancer. Unlike previous lung cancer models, some of the variables, like pack-years smoked and age, are modeled with restricted cubic splines (RCS). RCS allows for variables to be separated into k breakpoints, with the variable being linear below and above the

maximum and minimum breakpoints, and also have k-2 cubic variables which are determined for all values above the breakpoints in between the maximum and minimum breakpoints^{33,34}.

Discriminatory power results were excellent in both the non-screening arm (AUC = 0.859, 95% CI = 0.848-0.871) and the screening arm (AUC = 0.841, 95% CI = 0.813-0.870)³². Future analysis of this model will need to be conducted to see whether these results are validated in external datasets.

These risk models need to be examined for their usefulness in a clinical setting. Typically, this is accomplished by determining discriminatory power, accuracy, calibration, and clinical utility. First, the discriminatory power of a risk model is its ability to differentiate between a clinical case and a control. Discriminatory power can be calculated by estimating the area under a curve (AUC) from a receiver operator curve (ROC) which is determined by comparing the sensitivity and 1-specificity at each calculated risk value³⁵. A good risk model will have discriminatory power values substantially higher than chance (50%). Second, accuracy is determined by measuring the positive predictive value (PPV) and the negative predictive value (NPV) which are calculated in equations 1.6 and 1.7³⁶,

$$PPV = \frac{\text{Number of Individuals with Disease}}{\text{Number of Individuals Predicted to be Positive According to Model}} \quad (1.6)$$

$$NPV = \frac{\text{Number of Individuals without Disease}}{\text{Number of Individuals Predicted to be Negative According to Model}} \quad (1.7)$$

Third, calibration examines whether the observed absolute risk for a disease calculated by the risk model is similar to the ratio between cases and total individuals for a defined set of risk groups. Most calibrations are done with 10 fold cross-validations, or separating the cohort data into 10 groups based on observed absolute risk³⁷⁻³⁹. Finally, clinical utility is the measurement

of the percentage of individuals correctly calculated as a case (based on a selected absolute risk cut-off value) compared to the percentage of individuals incorrectly selected by the model to be a case (even though the “case” is truly a control) at specific risk intervals.

$$\text{Clinical Utility} = \frac{\% \text{ of Individuals Correctly Calculated as a Case}}{\% \text{ of Individuals Incorrectly Selected to be a Case}} \quad (1.8)$$

1.2. Genetic Analysis: Introduction to Haplotypes

1.2.1. Microsatellites to Single Nucleotide Polymorphisms (SNPs): The Beginning of Statistical Genetic Analysis

Differences in the genome among individuals have been explored extensively. One of the first methods to explore these differences in the genome is the use of microsatellites. Microsatellites are also known as short tandem repeats, and typically involve many repeats of a small set of genetic information (1 to 6 base pairs)⁴⁰. They have mutation values higher than that of other sections of the genome; hence they have been used as markers for basic population analysis, and have been very popular in the studies of non-human animals⁴¹. Below is a simple hypothetical example of a microsatellite on the two strands of DNA (Figure 1.1).

Figure 1.1: Example of a microsatellite with four ACG repeats on DNA

A	C	G	A	C	G	A	C	G	A	C	G
T	G	C	T	G	C	T	G	C	T	G	C

Figure 1.1: A hypothetical section of 12 base pairs of DNA are examined. The ovals show that these 12 base pairs consist of 4 ACG repeats on the forward strand of DNA. The backward or second strand, of DNA, shows 4 repeats of TGC, which occurs when the forward strand of DNA has ACG repeats.

However, microsatellites can have varying amounts of amplification, so that many microsatellites appear to be homozygous, but instead are really heterozygous; which can lead to incorrect evaluations about disease associations.

After the discovery of microsatellites, geneticists began to study specific base pairs on the genome to see whether a group of individual base pairs affect disease risk. Locations on the genome where they are single nucleotide differences in alleles, which are variant forms of the same gene at a specific base-pair, between one person and another are called SNPs⁴². SNP studies have become very popular because of their specificity and the sheer numbers of SNPs throughout the genome; currently, roughly 11 million SNPs have been identified⁴³. Below is an hypothetical example of SNPs on a set of chromosomes (Figure 1.2).

Figure 1.2: Single Nucleotide Polymorphisms from 12 areas on the chromosome

Chromosome 1	T	C	G	A	C	G	A	C	C	A	C	G
Chromosome 2	T	G	G	A	C	G	T	C	C	A	C	C
Chromosome 3	T	C	G	A	C	G	A	C	C	A	C	G
Chromosome 4	T	G	G	A	C	G	T	C	C	A	C	C
Chromosome 5	T	C	G	A	C	G	A	C	C	A	C	G
Chromosome 6	T	G	G	A	C	G	T	C	C	A	C	C
Chromosome 7	T	C	G	A	C	G	A	C	C	A	C	G
Chromosome 8	T	G	G	A	C	G	T	C	C	A	C	C
Chromosome 9	T	C	G	A	C	G	A	C	C	A	C	G
Chromosome 10	T	G	G	A	C	G	T	C	C	A	C	C

Figure 1.2: Ten chromosomes have been listed in the above figure, and twelve areas on each chromosome are being examined. Nine of these twelve areas have the same allele for all ten chromosomes, while three of these areas have at least two different alleles in the set of 10 chromosomes. These areas are marked by the red ovals, and since these areas have at least two different alleles, they are listed as SNPs.

The two largest databases of SNPs currently available are from HapMap⁴⁴ and the 1000 genomes project⁴³. HapMap has SNP data from 11 populations from various locations around the world (Table 1.1) that can be used to develop a haplotype map (hence HapMap) of the human genome which could be used to potentially detect patterns of genetic variation.

Table 1.1: Populations and List of populations available for analysis in HapMap:

The 1st column is the population identifier for each population from the HapMap Genome Browser release #29, the second column has the location of each population, and the final column contains the number of individuals available for analysis.

Population Identifier	Location of Population	Number of Individuals Available for Analysis
ASW	African ancestry in Southwest United States	83
CEU	European Ancestry from Utah Residents	174
CHB	Han Chinese from Beijing China	86
CHD	Chinese in Denver, Colorado	85
GIH	Gujarati Indians in Houston, Texas	88
JPT	Japanese in Tokyo, Japan	89
LWK	Luhya in Kenya	90
MEX	Mexican Ancestry from Los Angeles	77
MKK	Maasai in Kenya	171
TSI	Tuscan in Italy	88
YRI	Yourban individuals from Nigeria	176

Also, the development of HapMap has been used to attempt to find tagSNPs, which are estimated at around 300,000-600,000 SNPs out of the roughly 11,000,000 SNP currently detected⁴⁴. With the 1000 Genome Project⁴³, the goal is to attempt to find more common variants that have minor allele frequency of 1% by the process of gene sequencing. Next generation sequencing is being conducted in which short segments are randomly amplified and sequenced, then realigned to a consensus of the underlying genomic sequence. The depth of converge relates to the average number of sequences per individual per location. In the current available results from 1000 genomes project, sequencing has been conducted at four fold coverage depths. Then, for 1000 specific regions of the genome, sequencing is conducted at a much higher rate (50x). Since sequencing can be a very expensive process, not as many individuals are sequenced in the 1000 genomes project compared to HapMap per population listed in Table 1.1, but there are more SNPs available. For instance, 174 individuals from the

CEU population are available for analysis in HapMap, but only 92 individuals are available from the 1000 genomes project from the same CEU population.

It was theorized that with SNPs, scientists and genetics could find regions of the genome with substantial differences from one person to another, especially with disease research. This hypothesis, known as the Common Disease, Common Variant (CDCV) hypothesis became the basis for genome wide association studies (GWAS), which have been successful at finding links between SNPs and disease. For instance, three lung cancer studies showed some moderate links between SNPs on chromosome five or 15 and lung cancer⁴⁵⁻⁴⁷. However, it was soon discovered that increases in discriminatory power were modest for most diseases³⁰. Also, these SNPs only had odds ratios between 1.2 and 1.4, so thousands of individuals were needed to power the analysis that found the modest increases in discriminatory power. The search for greater power in population and disease studies will be a major theme of this thesis.

1.2.2. Introduction to Linkage Disequilibrium and Haplotypes

Throughout the genome, there are roughly 2.9 billion base pairs⁴⁸ and 11 million SNPs⁴³, and for substantial sections of the genome, these SNPs and base pairs are not completely independent. For instance, the allele structure at SNP rs1801131 could be dependent on the allele structure of SNP rs1801133. These SNPs are located within 2 kb of each other on the MTHFR gene, and can be a factor in modulating non-Hodgkin lymphoma risk^{49,50}. The dependency between rs1801131 and rs1801133 is referred to as linkage disequilibrium (LD), which is defined as having an association between different alleles⁵¹. Statistically, LD is calculated with both the D' and the r-squared statistic, and they are dependent on the allele frequency of both the individual SNPs and the possible combination of alleles. Below are these equations⁵¹:

$$D = P(A_1B_1) - P(A_1)P(B_1) \quad (1.9)$$

$$D' = \frac{D}{\min(P(A_1)P(B_2), P(A_2)P(B_1))} \quad \text{if } D > 0 \quad (1.10)$$

$$\text{or } D' = \frac{D}{-(1) * \min(P(A_1)P(B_1), P(A_2)P(B_2))} \quad \text{if } D < 0 \quad (1.11)$$

$$r^2 = \frac{(D)^2}{(P(A_1)P(B_1)P(A_2)P(B_2))} \quad (1.12)$$

where $P(A_1)P(A_2)$, $P(B_1)$, and $P(B_2)$ are the frequency of the two alleles at each base pair respectively, and $P(A_1B_1)$, $P(A_2B_2)$, $P(A_1B_2)$, and $P(A_2B_1)$ are the probability of having this combination of alleles. In cases of low LD, D' and r^2 are close to zero, but when LD is high, D' and r^2 are close to one.

Areas of the genome with high levels of LD are commonly referred to as haplotype blocks, and can vary from two to many SNPs within a relatively short distance in the genome (< 100 kb). An illustrative example is shown on Figure 1.3.

Figure 1.3: Actual Haplotypes from five SNPs associated with lung cancer in the study at MD Anderson Cancer Center⁴⁵

	rs8034191	rs3885951	rs2036534	rs6495306	rs680244
Haplotype 1	A	A	A	A	G
Haplotype 2	A	A	A	A	A
Haplotype 3	G	A	A	A	G
Haplotype 4	G	G	A	A	G

Figure 1.3: Four inferred haplotypes from the lung cancer GWAS conducted at MD Anderson are shown for five SNPs in strong LD with each SNP. This set of haplotypes that begin with SNP rs8034191 (Chromosome 15 at base pair location 76593078) and end at SNP rs680244 (Chromosome 15 at base pair location 76681394).

Larger size haplotype blocks typically do not occur in nature due to the high level of allele recombination throughout the genome. According to Greenwood et. Al⁵², by using haplotype block data collected by Gabriel's group in 2002⁵³, and genome recombination rates⁵⁴, there was a strong negative correlation between recombination rate and haplotype block size. With sex-averaged recombination rates of greater than 4 cM/Mb, the average haplotype block size decreased to nearly zero, while much small recombination rates lead to much larger haplotype block sizes. Mathematically the relationship between distance and recombination rates can be expressed as:

$$\rho = \frac{4Nc}{d} (1.13)$$

where ρ is the recombination rate, N is the Effective Diploid Population Size, c is the probability of recombination per generation between two consecutive markers or SNP, and d is the distance between two consecutive markers⁵⁵. Gabriel's group also had a strict definition for alleles to

form haplotypes, and the 95% CI of D' must contain the value one. This definition is built into the program Haploview⁵⁶, and all haplotypes discussed in this thesis will use this definition.

1.2.3. Introduction for Haplotype Analysis

As a predictive marker for disease, haplotypes can have higher power to detect genetic associations for disease over a single SNP⁵⁷⁻⁵⁹. The increase in power exists because haplotypes incorporate the linkage disequilibrium aspect of a genome section being studied. Because of this proposed increased power, haplotypes have been increasingly used as a disease risk predictor. Studies based on haplotypes are known as haplotype analysis, which is defined as the study of a set of linked alleles occurring on the same chromosome⁶⁰. With haplotype analysis, a geneticist can examine if the same SNPs that could have some association with disease also contain haplotypes that associate with the outcome of interest.

Some programs that infer haplotypes are PHASE^{55,61}, Haploview⁵⁶, and Haplo.stats^{62,63}. These programs use different methods to infer haplotypes, from a Bayesian method (PHASE), to haplotypes linked directly by strong regions of linkage disequilibrium (Haploview), and finally using an Expectation Maximization method (haplo.stats). With PHASE, haplotype frequencies for each individual are evaluated with the use of a Markov Chain Monte Carlo (MCMC) algorithm which incorporates the distance between each locus (in base pairs) along a chromosome. According to Stephens and Scheet⁵⁵, haplotypes tend to group together in clusters along a chromosome, so to obtain accurate estimates of haplotypes at a particular region, linkage disequilibrium must be taken into account, and that can be accomplished by stating the base pair location of each genetic location at which haplotypes could occur in nature. With Haploview, an open source code program written completely in JAVA, a multitude of haplotype analyses, from examination of Linkage Disequilibrium (LD) within a set of genetic markers located on a

specific chromosome, to single-marker and multi-marker association tests can be conducted. Haploview uses either familial data sets (data sets containing members of the same family), or case-control sets like that of the Puerto Rican data set. An example of Haploview output is shown on Figure 1.4.

Figure 1.4: Using Haploview to determine a haplotype from Chromosome 1 on the genome

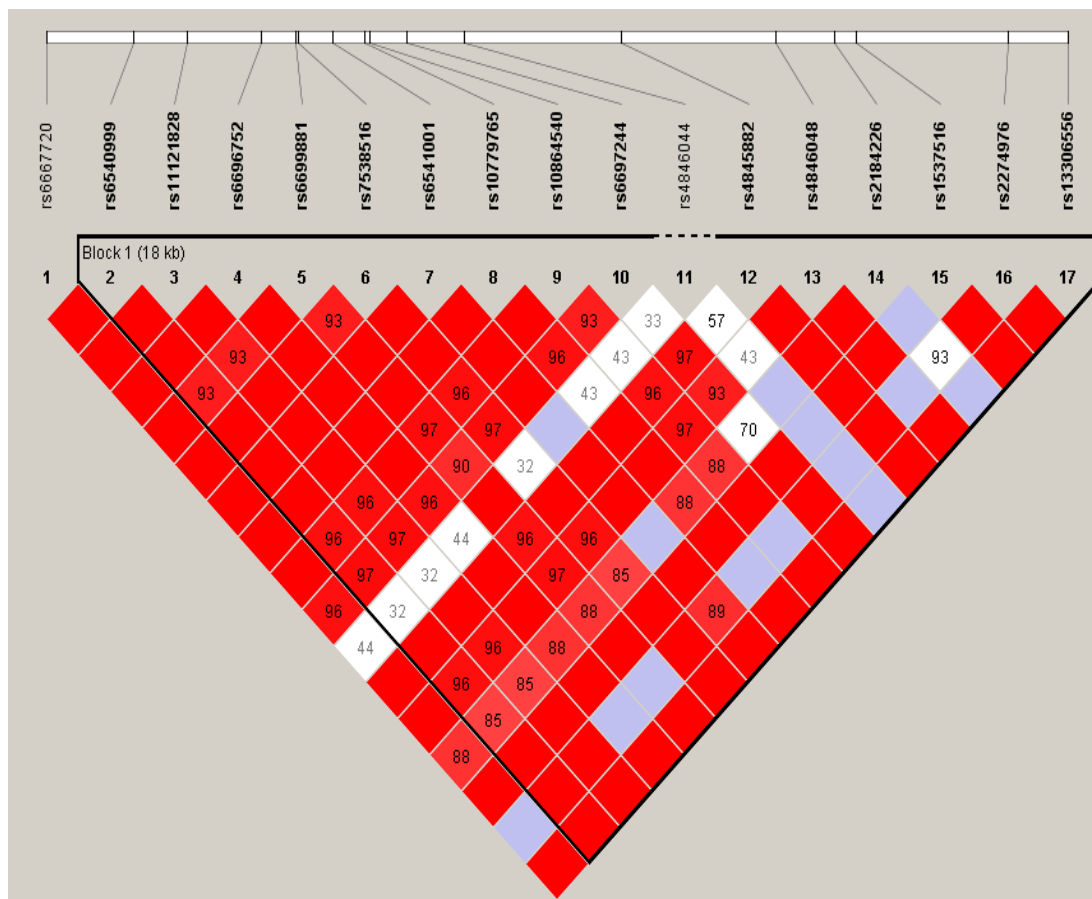


Figure 1.4: This sample haplotype is derived from a 20 kb set of SNPs in the HapMap database that begin at rs6667720 on chromosome 1 at base-pair location 11,754,202 and end at rs13306556 on chromosome 1 at base-pair location 11,774,697. Sixteen of the 17 SNPs listed in this block are linked together as one single haplotype.

Haplo.stats is a program developed in the programming language R which uses a generalized linear model to first determine the frequency of haplotypes for each individual, and then determine which haplotypes show significant differences between cases and controls with the use of a haplotype-specific score⁶². Haplo.stats also allows for the estimation of odds ratios to represent the association between either a haplotype or a covariate and disease with the use of a joint-effect linear model developed by the Lake group in which each haplotype is compared to the most frequent haplotype^{62,63}. With the assumption that the most frequent haplotype represents the “normal” haplotype status, one can examine whether variant haplotypes are associated with disease. Another popular linear model incorporated into the Haplo.stats is the separate-effects linear model, in which each haplotype is compared to a combination of all other possible haplotype (all other haplotypes are pooled together to form the reference group)⁶⁴⁻⁶⁶.

1.3. Overview of Dissertation Thesis

In this thesis, the ideas from risk modeling, haplotype analysis, and Bayesian inference, will be examined in the construction of a new Bayesian mathematical model, the Bayesian Joint Logistic Model (BJLM) that will be used to elucidate disease risk in three separate data sets. In Chapter 2, I will conduct a calibration of the Spitz lung cancer risk model using data from the Netherlands-Leuven Longkanker Screenings ONderzoek cohort (NELSON) or also known as the Dutch-Belgium randomized lung cancer screening trial. This analysis expands on the original validation of the Spitz lung cancer risk model, which was conducted using patients and controls from a Harvard case-control study (PI: David Christiani)⁶⁷. In Chapter 3, I will explore the value of including haplotypes to estimate genetic susceptibility to lung cancer and extend the Spitz model using an individual SNP approach. Chapter 4 contains the development of the BJLM including a simulation experiment to test its effectiveness with inferring haplotypes.

Chapter 5 contains an application of the BJLM to a Hodgkin data set to develop a novel risk model for Hodgkin disease consisting of haplotypes inferred with the BJLM. In Chapter 6, I use the BJLM to construct a haplotype-based extension of the Spitz model and show that this model has higher discriminatory power compared to previous model extensions (including the single SNP approach). Finally, in Chapter 7, I also highlight a modification of the BJLM in which I re-introduce SNPs as well as the inferred haplotypes from the BJLM to construct a genetics-only risk model for Glioma. I then conclude this thesis in Chapter 8 whereby I summarize the results observed within this thesis and pose possible research plans for the future.

Chapter 2: Validation and Calibration of the Spitz Lung Cancer Risk Model

2.1. Introduction to the NELSON data set

2.1.1. Relationship between NELSON and NLST

Risk modeling has been shown to be an important tool in early detection and prevention of disease. In the past few years, there have been substantial developments in risk model development for cancer; especially for breast cancer with the Gail model⁹. As stated in Chapter 1, validation of this model has been conducted in Caucasian females, and the Gail model performed well within specific risk factor strata and with individuals over the age of 60¹⁰. With the Gail model, high-risk individuals for breast cancer have been placed into screening trials for future analysis^{11,12}. Screening trials have been recently shown to decrease mortality in cancer, especially lung cancer. In late 2010, the National Lung Screening Trial (NLST), which was established in 2002 to determine whether screening participants at high risk for lung cancer with spiral CT or X-ray would reduce lung cancer mortality rates, has recently observed a 20% increase in 5-year survival rates within the CT group compared to the standard x-ray group⁶⁸. Colorectal risk models have been used as important indicators to develop more effective cost-benefit screening trials for this disease⁶⁹⁻⁷².

Creating valid and reliable risk models for lung cancer becomes especially important when one considers that over 80% of lung cancer cases are directly related to smoking, but only an estimated 11% of female smokers and 17% of male smokers will be diagnosed with lung cancer in their lifetimes^{73,74}. To address this challenge in estimating lung cancer risk, absolute risk models have been developed to identify high-risk individuals for lung cancer, and refine selection of individuals for screening trials^{67,75,76}.

2.1.2. Testing the Spitz Model with NELSON Data

One risk model is the Spitz lung cancer risk prediction model^{27,67}. The Spitz model was developed from an ongoing case-control study at M.D. Anderson Cancer Center with data available from 1995 to 2006. All cases had histological confirmed lung cancer²⁷. Controls were frequency matched by age (± 5 years), gender, race, and smoking status and were recruited from the largest multispecialty physician practice in Houston. To attempt to be as comprehensive as possible in lung cancer risk prediction, potential lung cancer risk factors including exposures, and co-morbidities such as hay fever and emphysema were included in the models. Standard, well established risk factors in the Spitz model included family history of smoking, passive smoking, age at quitting smoking for former smokers, pack-years for current smoking, asbestos and dusts exposures. Since the case-control study design included matching on smoking status, separate models were developed for never, former, and current smokers. The Spitz model has comparable discriminatory power and clinical utility with other lung cancer risk models (Bach and Liverpool Lung Project (LLP)⁷⁷ lung cancer models)⁶⁷. In particular, the Spitz model had excellent clinical utility, which is the ratio of correctly identified cases compared to incorrectly identified cases that are truly controls, at defined absolute risk values⁶⁷. Although the Spitz model has good discriminatory power and clinical utility, these measures have only been evaluated using case-control data, not cohort data, and model calibration has not been evaluated.

The purpose of this study was to determine the calibration, discriminatory power, and clinical utility of the Spitz model using prospective data from the Dutch-Belgian randomized lung cancer screening trial, “Nederlands-Leuvens Longkanker Screenings ONderzoek” (NELSON)⁷⁸⁻⁸³. The NELSON trial is especially important since it will be useful in determining future screening policy for lung cancer, like the recently announced results of the NLST trial^{68,84}.

In the NELSON trial, the lung cancer mortality of those who undergo computer tomography (CT-Scan) screening offered by the NELSON group (screening arm), are compared to those who do not undergo CT-screening by the NELSON group (control arm)⁷⁷. A previous study has shown that contamination, defined as lung cancer screening in the control arm, was limited⁸⁰. Information on cancer status is currently known for the screening arm, but not for the control arm, of the NELSON trial. So, the validation of the Spitz model will be conducted with the screening arm participants only.

2.2. Methods for Conducting Spitz Model Validation with NELSON Data

2.2.1. NELSON Study Population

The NELSON group distributed a questionnaire regarding health and smoking history to 548,489 individuals residing in the Netherlands and Belgium between the ages of 50 and 75⁸⁰. Those individuals who had smoked more than 15 cigarettes per day for > 25 years, or > 10 cigarettes per day for > 30 years, and were still smoking or had quit for 10 years or less, were invited into the NELSON trial⁸⁰. Exclusion criteria for this study included: Those with self-reported moderate or poor health status in combination with an inability to climb two sets of stairs, those with a recent history of lung cancer (<5 years), those with a reported chest CT-Scan in the year before study recruitment, a body weight of more than 308 pounds (140 kg), and those with a history of renal, melanoma, or breast cancer⁸⁰. After exclusion, 15822 individuals were selected for the NELSON trial, and there was a 1:1 random draw to determine those who would get CT-screening. Seven thousand nine hundred and fifteen individuals were selected for CT-screening, while 7907 individuals were selected for the control arm. Out of 7915 individuals in the screening, there were 196 individuals with confirmed lung cancer at the end of Phase 3 of the NELSON trial. One hundred eighty eight of these individuals had their lung cancer originally

detected by CT-Screening as part of the screening process, while another eight individuals were discovered to have lung cancer by linking the screen population with the cancer registry.

Those individuals with missing data from the set of risk factors associated with the Spitz model were excluded from analysis. After this exclusion, 109 individuals with lung cancer, and 4622 individuals without lung cancer, were included for validation analysis. The NELSON trial was approved by the Ethical Boards of all participating centers. The Minister of Health of the Netherlands approved the NELSON trial after positive advice from the Dutch Health Council according to the Dutch Screening Act.

2.2.2. Determination of 5-year Absolute Risk for Lung Cancer

Calculation of the 5-year absolute risk of lung cancer using the Spitz model was determined by first obtaining the relative risk profile of each individual in this analysis. Exact risk calculations using the Spitz risk model have been outlined in detail in the original published manuscript for the Spitz model²⁷, in which separate risk models were developed for former and current smokers. The model for former smokers incorporated the following variables: Quitting age; physician-diagnosed emphysema; dust exposure, prior self-reported hay fever and family history of cancer in first degree relatives. The model for current smokers included the following variables: pack-years, physician-diagnosed emphysema, dusts exposure, prior self-reported hay fever, family history of a smoking related cancer and asbestos exposure. The relative risk was then calculated by multiplying the log odds from the risk components of the logistic model for former or current smokers.

Absolute risk calculations were determined by first obtaining the age- and gender-specific incidence rates from the Netherlands cancer registry⁸⁵, and all-cause mortality (excluding lung cancer) rates from the Netherlands cancer registry⁸⁶ (Appendix 1: Table 1). To

account for the fact that Spitz model is stratified by smoking status, and to adjust for the NELSON population containing ever smokers only in this analysis, an adjustment factor for the incidence rate was used (Appendix 1: Table 2)²⁷. The age and gender adjusted incidence rates were multiplied by this adjustment factor, in which the percentage of ever-smokers developing lung cancer (stratified for sex) was divided by the percentage of either former smokers or current smokers in the general Dutch population (stratified for sex)⁸⁷⁻⁸⁹. For instance, for a 61 year-old male current smoker, the constant adjustment factor is derived from the ratio of the proportion of all lung cancer cases in ever-smoking men (0.964) divided by the proportion of male current smokers in the population at risk (0.322), i.e., $ac_{13} = 0.964/0.322 = 2.99$. At age 61, the male incidence rate for cancer is 192.4 individuals per 100000. Hence, this individual's adjusted incidence rate is 2.99×192.4 per 100000, or 0.00575. Finally, the relative risks, the incidence rates, and the mortality rates are combined by using the Dupont-Plummer equation for absolute risks⁹⁰ (Equation 2.1).

$$p(s, a, R) = R \int_a^{a+s} \lambda(t) \left(\exp \left[- \int_a^{a+s} R\lambda(t) + \mu(t) dt \right] \right) dt \quad (2.1)$$

where R = relative risk calculated by the Spitz model, a = current age, s = number of years to calculate absolute risk, $\lambda(t)$ = annual age and sex specific incidence rate, $\mu(t)$ = annual age and sex specific mortality rates from all other causes than the disease being examined, and $p(s, a, R)$ = s year absolute risk for disease

For this validation, the variables hay fever, dusts exposure, and family history were not available in the NELSON trial database. A list of all variables used to construct absolute risk is summarized in Table 2.1.

Table 2.1: Lifestyle Variables Used in Original Spitz Model and in Validation of Spitz Model with NELSON data

Variables	Original Spitz variables	Validation variables
Pack-Years	Yes	Yes
Age Stopped Smoking	Yes	Yes
Age	Used for LC incidence rate and LC-free mortality rate	Used for LC incidence rate and LC-free mortality rate
Sex	Used for LC-specific incidence rate and LC-free mortality rate	Used for LC-specific incidence rate and LC-free mortality rate
Family History	Yes	No
Asbestos Exposure	Yes	Yes
Dusts Exposure	Yes	No
Emphysema	Yes	Yes
Hay Fever	Yes	No
LC Incidence Rate	Yes (SEER rate)	Yes (Netherlands rate)
LC-free Mortality Rate	Yes (NCHS rate)	Yes (Netherlands rate)

Smokers were defined as those who had smoked at least 100 cigarettes in their lifetime, while former smokers were those who had quit smoking at least 1 year before filling out the initial questionnaire. Age was defined at the 1st questionnaire time point. Asbestos exposure was determined by a separate questionnaire of 17 items relating to work in fields that have exposure to asbestos. If an individual answered positively to any of those items, and had worked in an asbestos-related industry, they were considered as having been exposed to asbestos. Self-reported physician-diagnosed emphysema at any time before entering the NELSON trial was listed as positive for an individual to have emphysema.

2.2.3. Estimation of Calibration and Discriminatory Power for NELSON analysis

Calibration of the Spitz model within the Nelson trial data was conducted by comparing the ratio of cases to total individuals, and the 95% Confidence interval of each level of absolute risk, which was defined at the mid-point for each of the ten possible risk levels. Good calibration

was evident if the ratio of cases to total individuals were within the 95% Confidence Interval for the observed absolute risk of lung cancer. This analysis was similar to the analysis from Figure 2 in Bach et al³¹. With the absolute risk lung cancer calculation, we calculated discriminatory power by obtaining receiver operator characteristic (ROC) curves and estimating the area under the curve (AUC) (empirical method) with SPSS 17.0™ statistical software (SPSS, Kaysville, UT). Discriminatory power was calculated for the entire NELSON data set, and also for the former and current smoker subsets of the NELSON data set.

2.2.4. Estimation of clinical utility for NELSON analysis

Clinical utility for the Spitz model in the Nelson data set for all individuals was evaluated using scaled rectangle diagrams as developed in the Search Partition Analysis (SPAN, Auckland, New Zealand) program^{91,92}. With scaled rectangle diagrams, a graphical presentation of model discrimination is obtained by displaying the risk for disease for a specific risk model and true disease status. With these diagrams, the white rectangle represents all individuals, the green rectangle represents all cases, and the blue, purple, and red rectangles represent individuals with three increasing levels of risk for lung cancer (1.0%, 2.0%, and 3.0%, respectively). Better clinical utility is defined as having a large ratio of cases correctly inferred by the model and having fewer individuals incorrectly inferred as cases. Clinical utility was also represented using a varying absolute risk rate of lung cancer from 1.0% to 5.0%, and this analysis was conducted with Matlab.

2.3. NELSON Data Results

The epidemiologic and lifestyle information for the 109 individuals with lung cancer and 4622 controls are presented in Table 2.2.

Table 2.2: Demographic characteristics of the NELSON population

	Cases (N = 109)	Controls (N = 4622)	P-Value
Age (years): mean \pm s.e.	61.7 \pm 5.5	58.1 \pm 5.4	< 0.001
Sex: No. (%)			
Male	107 (98.2%)	4520 (97.8%)	0.794
Female	2 (1.8%)	102 (2.2%)	
Smoking Status: (%)			
Current	76 (69.7%)	2878 (62.3%)	0.133
Former	33 (30.3%)	1744 (37.7%)	
Current Smokers			
Pack-Years: mean \pm s.e.	45.2 \pm 18.8	40.5 \pm 16.9	0.016
Emphysema No. (%)			
Yes	6 (7.9%)	85 (3.0%)	0.028
No	70 (92.1%)	2793 (97.0%)	
Asbestos Exposure No. (%)			
Yes	14 (18.4%)	506 (17.6%)	0.850
No	62 (81.6%)	2372 (82.4%)	
Former Smokers			
Age Stopped Smoking: mean \pm s.e.	55.1 \pm 9.9	52.5 \pm 8.4	0.078
Emphysema No. (%)			
Yes	1 (3.0%)	103 (5.9%)	0.718
No	32 (97.0%)	1641 (94.1%)	

Cases (mean age, 61.7 years) were on average three and a half years older than controls (mean age, 58.1 years; $P < 0.001$). Almost all the cases and controls (98%) were male. Higher percentages of controls (37.7%) were former smokers compared to cases (30.3%), and the reverse was true for current smokers (62.3% vs. 69.7%) but these results were not statistically significant ($P = 0.133$). Current smokers who were cases reported significantly higher pack-years (45.2) compared to controls (40.5) ($P = 0.016$) and were more likely to report prior history of emphysema ($P = 0.028$). However, such a difference was not evident in former smokers ($P = 0.718$). For the other variables in the analysis, there were no significant differences in either former or current smokers. Calibration results are shown in Figure 2.1 with the defined risk levels listed in Table 2.3.

Table 2.3: Calibration Details for Each Risk Level:

Ten levels of risk are developed to test calibration and these levels are based on the calculated 5-year absolute risk range. For instance, level of risk 2 corresponds to those subjects in the 10 to 20 percentile of absolute risks in the NELSON population.

Levels of Risk	Number of Individuals	Absolute risk range (%)	Median (%)
1	474	0.46-0.75	0.67
2	473	0.75-0.90	0.80
3	473	0.90-1.05	1.00
4	473	1.05-1.25	1.14
5	473	1.25-1.50	1.36
6	473	1.50-1.81	1.65
7	473	1.81-2.17	1.97
8	473	2.17-2.67	2.19
9	473	2.68-3.62	3.09
10	473	3.62-11.49	4.15

Figure 2.1: Calibration Results with 95% CI's for each median risk level

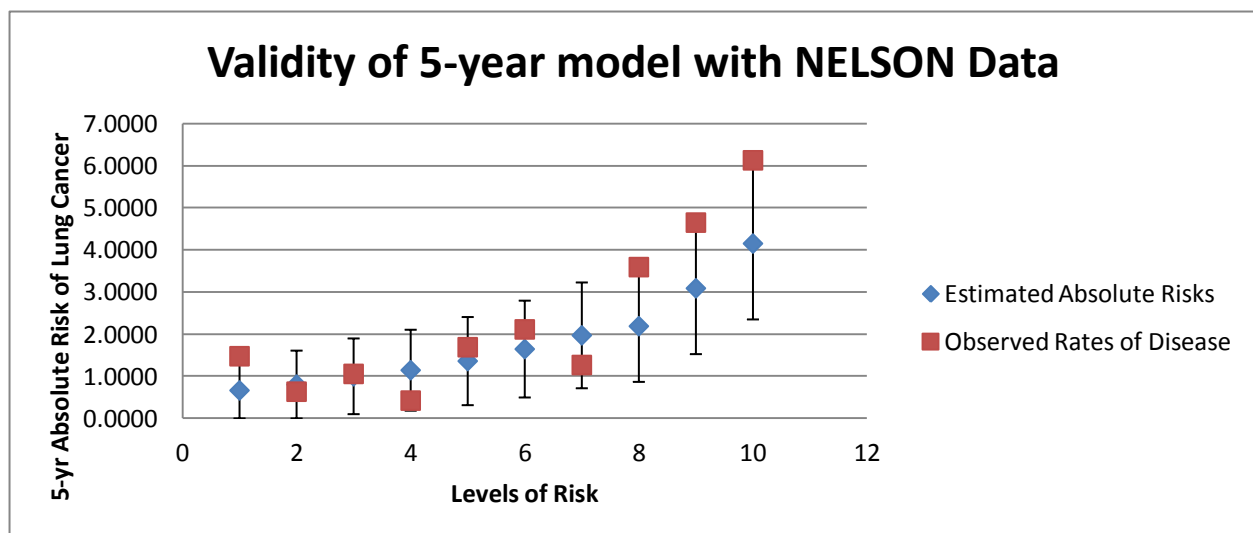


Figure 2.1: Calibration results for the Spitz Model with NELSON data. Blue diamonds represent the midpoint for each estimated absolute risk level. Red squares represent the observed rates of disease, while the black lines are the 95% CI of the midpoint of the estimated absolute risks for each risk level.

As the level risk increased, the range of absolute risks within the risk level increased until the highest risk level, which had a much larger range of absolute lung cancer risks. In all risk strata except for the three highest and the lowest, the observed ratio between cases and total individuals were well within the 95% CI of the calculated 5 yr risk. Four of the ten risk levels had an almost exact match between the observed cases to the total individual ratio and the calculated absolute risk, while two of the ten risk levels had their observed cases to the total individual ratio in close proximity to the calculated absolute risk. Only for the highest risk individuals, and the lowest risk individuals, did there appear to be some separation between observed lung cancer ratios and calculated absolute risks for lung cancer.

We next tested the discriminatory power of the Spitz model in the NELSON data set. The discriminatory power of the Spitz model for ever smokers in the NELSON data set was 0.69 (95% CI = 0.64-0.75). When stratified by smoking status, the discriminatory power for current smokers was 0.74 (95% CI = 0.67-0.80), and for former smokers, the discriminatory power was 0.61 (95% CI = 0.52-0.71).

Clinical utility results are graphically shown on Figure 2.2 using SPAN at three specific absolute risk values: 1.0% (shaded in blue), 2.0% (shaded in purple), and 3.0% (shaded in red).

Figure 2.2: Clinical Utility Results in Graphical Form for the Spitz model

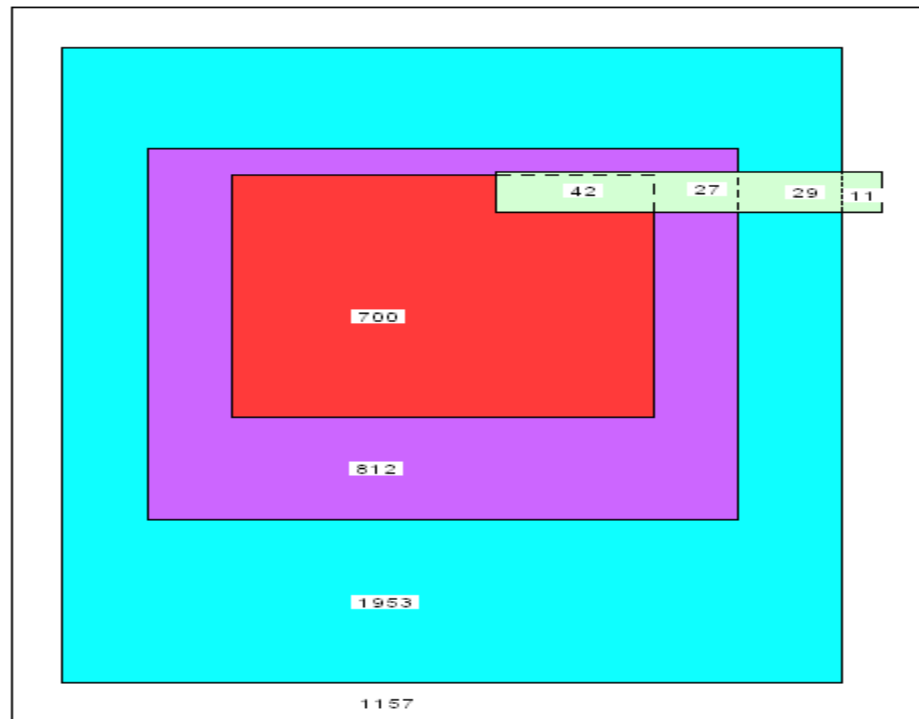


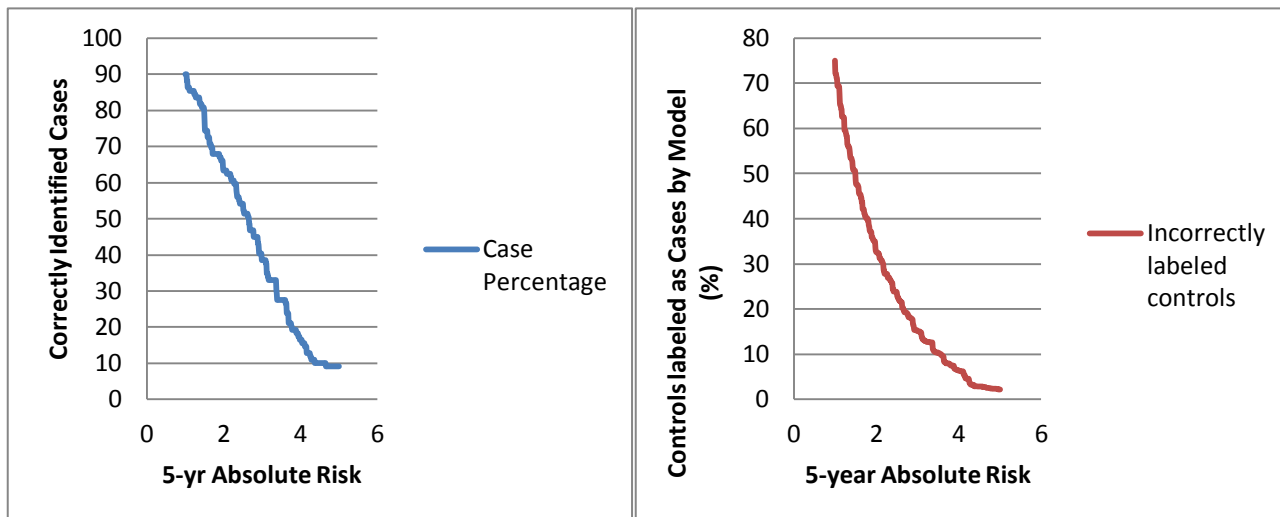
Figure 2.2: Clinical utility of the Spitz models. Scaled rectangle diagrams for the Spitz Model at defined levels of lung cancer risk. For each color of the diagram: white equals all controls with < 1.0% risk, and light green equals all cases. Blue represents all individuals with at least 1.0% risk, but less than 2.0% risk. Purple represents all individuals with at least 2.0% risk, but less than 3.0% risk. Red represents all individuals with at least 3.0% risk. All of the numbers represent the number of cases or controls within a specific risk level. For example, there are 1157 controls with risk less than 1.0%.

At these risk values, the percentage of cases correctly identified with increasing risk values listed above are 89.9%, 63.3%, and 38.5%, respectively, while the percentage of controls incorrectly identified as cases were 75.0%, 32.7%, and 15.1%. Next, the analysis was expanded to include 5-year absolute risks from 1% to 5%, and this analysis is graphically shown in Figure 2.3.

Figure 2.3: Expanded Clinical Utility for 5-year Absolute Risks from 1% to 5%

A) Correct % of Cases Identified

B) Incorrect % of Controls identified as Cases



C) Ratio of Correctly Identified Cases to Incorrectly Identified Controls

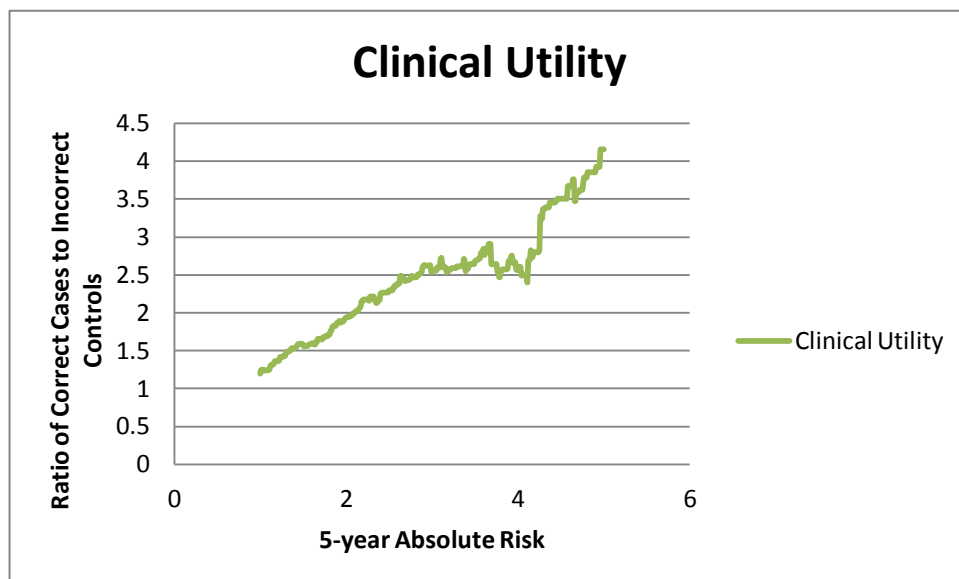


Figure 2.3: Expanded Clinical utility of the Spitz model without dusts in the NELSON data set for a wide region of 5-year absolute risks. Components of clinical utility, correctly identified % of cases, and incorrectly identified % of identified as cases by the model, are shown in Figures 2.4A and Figures 2.4B, respectively. Figure 2.4C shows the ratio of correct identified % of cases to incorrect % of controls identified as cases for 5-year absolute risk range from 1% to 5%.

In Figure 2.3A, we show that the percentage of cases drops from 89.9% with absolute risk 1% or greater to about 9.2% for those with 5% absolute risk or greater. In Figure 2.3B, we show that for percentages of incorrectly identified cases that are truly controls, this percentage dropped from about 75.0% with absolute risk 1% or greater to about 2.2% for those with 5% absolute risk or greater. The ratio of correct cases to incorrect controls steadily increase from 1.2 correct cases for every incorrectly labeled true control at 1% absolute risk to over 4 correct cases for every incorrectly labeled true control at 5% absolute risk (Figure 2.3C).

2.4. Discussion of NELSON Results

The purpose of this analysis was to validate the Spitz model with data from NELSON, a lung cancer screening trial in the Netherlands and Belgium. The analysis showed good calibration with the five year absolute risk calculations. These calibration results are the first in the lung cancer literature to show that five year absolute risk models can be calibrated just as well as a ten year absolute risk model for lung cancer.

Further analysis of the calibration results suggested that the 5 year absolute risk calculations only faltered slightly at the highest risk levels. There could be two reasons for the discrepancy between observed risk for cancer measured by the total numbers of cases divided by the total number of individuals and the 95% CI of the calculated 5 year absolute risk for lung cancer. First, more stringent methods to measure some of the effects of the non-smoking variables, especially emphysema, may be needed. Emphysema has been shown in numerous studies to be a statistically significant factor for lung cancer^{27,93-95}, so if the risk of emphysema is being understated, if patient self-reported emphysema is not clinically validated, or if the discriminatory effect of emphysema is low within a data set, absolute risk values may be biased toward the null for these individuals, leading to less accurate validity results. Also, the matching

of smoking status within the Spitz model could result in lower absolute risks, and hence, the full effect of smoking on lung cancer may not be measured for those with higher lung cancer risks despite the use of adjustment factors to account for smoking status.

Discriminatory power from this analysis matches up well with the overall results from that we previously reported using Harvard case-control data^{67,96-99}. We show that the Spitz models currently have about a 69% discriminatory power to separate cases and controls. However, differences do exist between former and current smokers in the two studies. For former smokers, discriminatory power was higher with the Harvard data set (70%) compared to the NELSON data set (61%), while for current smokers; discriminatory power was higher for the NELSON data set (73%) compared to the Harvard data set (68%)⁶⁷. These differences, especially in former smokers, can be explained by the lack of family history information and the fact that emphysema, in the NELSON data set did not differ between cases and controls. Emphysema is the most significant factor in the Spitz model for former smokers, and this was especially true in the Harvard dataset. In the Harvard dataset, there were some suggestions that the presence of emphysema (a self-reported variable) could lead to recall bias, but because of the structure of the Harvard study, recall bias was limited, and there was a possibility that the AUC results were conservative^{67,99-101}. The lack of information about emphysema with NELSON data drove its discriminatory power towards the null, so therefore it is possible that for both studies, the discriminatory power of the Spitz model was actually conservative.

The discriminatory power results does compare favorably to models developed for melanoma (0.62) and breast cancer (0.58-0.77)^{14,16,102-105}. However, models for colorectal cancer have higher discriminatory power (0.84-0.86) compared to the Spitz model¹⁹. A future goal could be to improve the Spitz model such that discriminatory power increases to 0.75, which is

what is suggested for screening those with increased disease risk^{107, 108}. One approach to improving the discriminatory power could be to expand the model by incorporating risk biomarkers, especially SNPs identified from genome-wide association studies. Unfortunately, strong discriminatory factors that can increase overall discriminatory power do not currently exist. Others have shown that adding one or more top hit SNP's from GWAS does not substantially increased discriminatory ability of the models^{108,109}.

The clinical utility of the Spitz model's performance improved from a ratio of roughly one correctly labeled case per misclassified case at lower absolute risk values to a ratio of four correctly labeled cases for every misclassified case, at higher absolute risk values. These clinical utility results were similar to the Harvard study, as both studies had absolute risk values in which four times as many cases were labeled correctly compared to misclassifieds⁶⁷. However, the Scaled rectangle diagram shows that work still needs to be done in developing models for lung cancer risk prediction as all of the risk boxes only include some of the cases in the NELSON trial.

One explanation for the lower detection of actual cancers is the very low ratio of cases to controls in this study. For instance, at the 2.5% 5-year absolute risk listed in this analysis, there were 1148 individuals with at least a 2.5% absolute risk or greater for ever smokers, and 591 and 557 individuals, for current and former smokers respectively. For those with 2.5% or more absolute risk, there were only 59 individuals overall, with 16 being former smokers, and 43 current smokers, who had a diagnosis of lung cancer. For diseases with lower prevalence, or for studies with a preponderance of controls, these results are not that unusual. Systemic lupus erythematosus has a prevalence of only 33 in 100000 individuals worldwide, and when an antinuclear antibody test was designed for this disease, it had a sensitivity of 94% and a

specificity of 97%¹¹⁰⁻¹¹². However, when this test was applied to the entire population to test its accuracy, NPV values were close to 100% and PPV values were around 1%, which were due to having many more true negative and false positive results compared to false negative and true positive results, respectively¹¹⁰.

This validation study of the Spitz model had some limitations. First, there was no information in the NELSON data set for hay fever, dusts exposure, and family history of cancer (both smoking related and non-smoking related). So, it is possible that the true discriminatory power and the clinical utility could have been understated. However, after creating a dusts variable based on an asbestos questionnaire that was given to all accepted members of the NELSON trial to determine asbestos exposure, there were decreases in both discriminatory power and clinical utility (data not shown). The dusts variable was created using the same questionnaire that was used for the creation of the asbestos variable, so since the asbestos variable did not show strong discrimination, the dusts variable also showed very weak discrimination, and hence, lower discriminatory power and clinical utility. Second, there were few lung cancer cases relative to controls, and the cohort trial is still ongoing, so this discriminatory power, clinical utility and calibration values could change as more members of the NELSON study are diagnosed with lung cancer.

Despite these limitations, the Spitz model had respectable discriminatory power, clinical utility, and calibration, results in a completely independent, longitudinal cohort. Since 1.35 million individuals are diagnosed with lung cancer every-year worldwide and about 30% of all lung cancers occur despite reducing the prevalence of the major risk factors, it is essential that a risk model can differentiate between lung cancer patients and controls¹¹³⁻¹¹⁴. These promising results show that no matter if the study population is a random draw of individuals of heavy

smokers from the general populations of Belgium and the Netherlands, or a case-control study with hospital-based participants⁶⁷, the Spitz model is a useful model for predicting risk.

Recently, the NLST trial showed an improved 5-year survival rates among high-risk smokers screened via CT versus standard x-ray. The NLST defined high risk smokers as those aged 55-74 with smoking intensity of thirty or more pack-years, and are either current smokers or former smokers who have quit within the past 15 years⁸⁸. Well-calibrated and validated lung cancer risk prediction models, such as the Spitz model, can be used in conjunction with NLST-like screening trials as a cost-effective way to better identify those high-risk individuals who would most benefit from screening and hence decrease mortality rates as well as increase screening efficiency. Such risk models can also be used to educate smokers to their personalized risk, which have been shown in preliminary studies to show promise in improving smoking cessation¹¹⁵⁻¹¹⁷. Now, the Spitz lung risk model will be extended using genetic factors in the next chapter as an attempt to improve its discriminatory power.

Chapter 3: Expansion of Existing Risk Models Using Genetic Factors

In the previous chapter, the Spitz model has been validated and calibrated with the NELSON data set originating from Erasmus MC, which was an extension of the validation conducted with the Harvard lung cancer data set. These validations determined the applicability of models consisting of non-genetic risk factors like self-reported health co-morbidities like hay-fever and emphysema, pack-years, family history of cancer, and also occupation-related exposures like asbestos and dusts. Discriminatory power results varied from 0.61 in former smokers with the NELSON data set, to 0.74 in current smokers. In this chapter, I begin to explore the applicability of SNPs as risk factors that can better discriminate between cases and controls.

3.1. Incorporating SNPs into Cancer Risk Models: A Primer

With the development of genome-wide association studies, specific areas on the genome have been shown to be linked to increased risks of a number of diseases. Specifically, in recent genome-wide association studies, the objective has been to find SNPs, or genes with variant base pairs, that show increased risk for disease¹¹⁸. However, when SNPs are added to already existing models for cancer, the increase in the models ability to discriminate between cases and controls may be modest at best^{119,120}. For instance, when seven SNPs that were associated with an increased risk of breast cancer were found in two separate genome-wide association studies^{121,122} were added to the National Cancer Institute's Breast Cancer Risk Assessment Tool (BCRAT), they increased the discriminatory power less than just the addition of mammographic density to the BCRAT¹¹⁹.

Even more recent studies have suggested that the increase for breast cancer is questionable. A recent simulation study based on a meta-analysis of the breast cancer literature

suggested that 41 SNPs that showed significant associations for breast cancer can lead to a risk model with discriminatory power of 0.67¹²³. Addition of 50 additional variants with varying odds ratios of 1.2 to 1.5 will lead to discriminatory power between 0.70-0.80¹²³. However, in a commentary titled, “Predicting the Future of Genetic Risk Prediction”, there are concerns that these results may be too generous¹²⁴. According to Chatterjee’s group, 27.9% of the known heritability of breast cancer would need to be discovered to obtain a discriminatory power of 0.67¹²⁵, but the 18 SNPs found in GWAS associated with the meta-analysis only found 7.9% of known heritability¹²⁴. Do the same concerns about the effectiveness of SNPs in breast cancer translate to lung cancer?

In 2007, Dr. Margaret Spitz and her colleagues at the Department of Epidemiology at M.D. Anderson Cancer Center proposed a lung cancer risk prediction model²⁷. This model has been extended for former and current smokers to include biomarkers of DNA repair capacity and mutagen sensitivity¹²⁶. Obtaining these biomarkers is time-consuming and requires some level of technical expertise. Therefore while feasible in a controlled academic setting, they are not applicable for widespread population-based implementation. For this analysis, the original Spitz model will be expanded by incorporating four SNPs; two from chromosome 5 (rs2736100, and rs401681), and two from chromosome 15 (rs1051730 and rs8034191); that have been found to be associated with lung cancer risk^{45-47,127-129}. This will be an expansion of the results published previously in which SNPs rs1051730, rs8034191 and rs401681 added to the original Spitz model, led to a modest increase in discriminatory power of 0.012, from 0.661 to 0.673³⁰.

3.2. Materials and Methods to add SNPs to Risk Models

3.2.1. Study Population for MD Anderson Lung Cancer Study

A total of 3852 lung cancer patients and controls were accrued for this study. Lung cancer patients (N=1851) were enrolled from the Thoracic Center at the University of Texas MD Anderson Cancer Center starting in July of 1995 and ending in May 2006²⁷. All lung cancers were histologically confirmed, were newly diagnosed, and had no treatment (chemo or radiation) for lung cancer²⁷. Controls (N=2001) were recruited from the Kelsey-Seybold clinics, and these individuals were lung cancer-free individuals with no prior history of cancer (except for nonmelanoma skin cancer). These controls were frequency matched by age (± 5 years), sex, ethnicity, and smoking status²⁷. The risk factors in the original Spitz study are listed in **Table 2.1 in Chapter 2 on page 27**. Only ever smokers (individuals that have only smoked > 100 cigarettes in their lifetimes) are included in this analysis.

3.2.2. Incorporating Genetic Information into Risk Modeling

To test whether top hits from GWAS do indeed increase discriminatory power in lung cancer risk modeling, I will develop and examine three risk models with the addition of top SNPs from the Texas lung cancer GWAS conducted by Dr. Christopher Amos's group at the University of Texas MD Anderson Cancer Center⁴⁵, and from epidemiological data extracted from the Original Spitz lung cancer risk study. First, I will add the top SNP from the lung cancer GWAS, rs8034191, which is located on chromosome 15. Second, I will add three top SNPs, with two top SNPs from chromosome 5 (rs2736100 and rs401681), and a third SNP which is in strong linkage disequilibrium (LD) with SNP rs8034191 (rs1051730). Finally, I will develop risk modeling using haplotypes based on the SNPs in LD that were top hits in the GWAS study

from chromosome 5 (rs2736100 and rs401681) and chromosome 15 (rs1051730 and rs8034191), and genotype information will be available for 2291 individuals (1154 cases and 1137 controls).

For the development of the risk model with haplotypes, all haplotypes were inferred with Haplo.stats⁶²⁻⁶³, and all haplotypes were compared to the most frequent haplotype for both chromosome 15 and chromosome 5. After inferring of haplotypes, the risk of the non-common haplotypes in chromosomes 15 and 5 will be determined with the rest of the covariates from the Spitz model.

All risk models will have its discriminatory power calculated. These models will be compared to each other, and to the original Spitz model. Discriminatory power results will be calculated for all individuals, and former and current smokers, separately. Also, risk models, including the original Spitz model, will be compared to each other by using the Net Reclassification Index (NRI)¹³⁰ which quantifies the improvement of classification due to changes in the predictive value of the extended model compared to the original model. NRI results will be calculated overall and separately for cases and controls.

3.3. Risk Model Results for Adding Genetic Information to Spitz models

3.3.1. Results: 1st Model: Top SNP + Spitz Original Model

The first model to be examined includes all variables in the original Spitz model plus SNP rs8034191. The risks for each variable in this model are listed in Table 3.1.

Table 3.1: Multivariate analysis of Spitz Lung Cancer Model Risk Factors plus SNP rs8034191

New risk models incorporating all relevant Spitz risk factors²⁷ for Former and Current Smokers plus SNP rs8034191 are developed using multivariate logistic regression

Risk Factor	Former Smokers (584 cases / 654 controls)		Current Smokers (533 cases / 481 controls)	
	Regression Coefficient	OR (95% CI)	Regression Coefficient	OR (95% CI)
Constant	-1.202	N/A	-1.204	N/A
Emphysema	0.933	2.543 (1.786-3.620)	0.982	2.669 (1.839-3.875)
Dust Exposure	0.470	1.600 (1.254-2.040)	0.189	1.208 (0.919-1.589)
Family history (≥ 2)	0.485	1.625 (1.263-2.091)	N/A	N/A
Age Stopped Smoking 42-53 years	0.081	1.084 (0.804-1.461)	N/A	N/A
Age Stopped Smoking ≥ 54 years	0.414	1.513 (1.127-2.031)	N/A	N/A
Hay Fever (No vs. Yes)	0.375	1.455 (1.079-1.963)	0.391	1.478 (1.048-2.084)
Asbestos Exposure	N/A	N/A	0.503	1.654 (1.117-2.451)
Smoking Family History (≥ 1)	N/A	N/A	0.546	1.726 (1.279-2.328)
Pack-Years (28-41.9)	N/A	N/A	0.226	1.254 (0.836-1.881)
Pack-Years (42-57.4)	N/A	N/A	0.321	1.378 (0.927-2.048)
Pack-Years (≥ 57.5)	N/A	N/A	0.653	1.922 (1.311-2.818)
rs8034191 (G vs. A)	0.210	1.234 (1.039-1.464)	0.227	1.254 (1.037-1.518)

With this extended model, the discriminatory power increases slightly from 0.660 (95% CI = 0.637-0.681) in the original Spitz model to 0.668 (95% CI = 0.646-0.690), and this increase is significant ($p = 0.019$). Also, the new model has superior ability to classify case/control status in all individuals compared to the original Spitz model according to the NRI value, (0.182, p -value = < 0.001). This NRI increase is more prevalent with true cases (NRI = 0.348, p -value = < 0.001), compared to cases (NRI = -0.167, p -value = < 0.001).

3.3.2. Results: 2nd Model: SNPs from Chromosomes 15, 5, and 6 + Spitz Original Model

The second model to be examined includes all variables in the original Spitz model plus SNPs rs2736100, rs401681, and rs1051730. The risks for each variable in this model are listed in Table 3.2.

Table 3.2: Multivariate analysis of Spitz Lung Cancer Model Risk Factors plus three SNPs: rs2736100, rs401681, and rs1051730

New risk models incorporating all relevant Spitz risk factors²⁷ for Former and Current Smokers plus SNPs rs2736100, rs401681, and rs1051730 are developed using multivariate logistic regression

Risk Factor	Former Smokers (583 cases / 654 controls)		Current Smokers (534 cases / 481 controls)	
	Regression Coefficient	OR (95% CI)	Regression Coefficient	OR (95% CI)
Constant	-1.475	N/A	-1.752	N/A
Emphysema	0.930	2.534 (1.777-3.611)	0.992	2.695 (1.851-3.926)
Dust Exposure	0.468	1.596 (1.251-2.037)	0.198	1.218 (0.924-1.606)
Family history (≥ 2)	0.486	1.625 (1.262-2.093)	N/A	N/A
Age Stopped Smoking 42-53 years	0.074	1.077 (0.798-1.452)	N/A	N/A
Age Stopped Smoking ≥ 54 years	0.417	1.518 (1.130-2.039)	N/A	N/A
Hay Fever (No vs. Yes)	0.386	1.472 (1.089-1.988)	0.388	1.474 (1.042-2.085)
Asbestos Exposure	N/A	N/A	0.532	1.702 (1.145-2.530)
Smoking Family History (≥ 1)	N/A	N/A	0.571	1.770 (1.307-2.395)
Pack-Years (28-41.9)	N/A	N/A	0.285	1.330 (0.883-2.004)
Pack-Years (42-57.4)	N/A	N/A	0.368	1.445 (0.969-2.155)
Pack-Years (≥ 57.5)	N/A	N/A	0.679	1.973 (1.341-2.901)
rs2736100 (C vs. A)	0.143	1.154 (0.975-1.365)	0.303	1.354 (1.118-1.641)
rs1051730 (A vs. G)	0.229	1.257 (1.060-1.491)	0.239	1.270 (1.048-1.540)
rs401681 (C vs. T)	0.091	1.095 (0.919-1.304)	0.106	1.173 (0.967-1.423)

With this extended model, the discriminatory power increases slightly from 0.659 (95% CI = 0.636-0.681) in the original Spitz model to 0.674 (95% CI = 0.652-0.696), and this increase is significant ($p = 0.004$). Also, the new model has much better ability to classify case/control status in all individuals compared to the original Spitz model according to the NRI value, (0.268, p -value = < 0.001). This NRI increase is more prevalent with true cases (NRI = 0.357, p -value = < 0.001), compared to controls (NRI = -0.089, p -value = 0.003).

3.3.3. Results: 3rd Model: Haplotypes from Chromosome 5 and 15 + Original Spitz Model

The third model to be examined includes all variables in the original Spitz model plus haplotypes from chromosome 15 (rs8034191 and rs1051730) and chromosome 5 (rs2736100 and rs401681). The haplotypes and their frequencies are listed on Table 3.3.

Table 3.3: Haplotype Frequencies for Top SNPs in GWAS that are in High Linkage Disequilibrium with Each Other According to Texas lung GWAS Study⁴⁵

Chromosome 15 Haplotypes			
SNPs	bp	Haplotypes	Frequencies
rs8034191 rs1051730	76,593,078 76,681,394	AG	0.62096
		AA	0.01150
		GG	0.01281
		GA	0.35472
Chromosome 5 Haplotypes			
rs2736100 rs401681	1,339,516 1,375,087	AT	0.24978
		AC	0.23538
		CT	0.17601
		CC	0.33883

From the haplotype frequency results, this model will be constructed from the AA, GG, and GA haplotypes from chromosome 15, and the AT, AC, and CT haplotypes from

chromosome 5. Also, all of the variables from the Spitz model will be included in this model.

The risks for each variable are listed in Table 3.4.

Table 3.4: Multivariate analysis of Spitz Lung Cancer Model Risk Factors plus haplotypes from Chromosome 15 and Chromosome 5

New risk models incorporating all relevant Spitz risk factors²⁷ for Former and Current Smokers plus the haplotypes from Chromosomes 5 and 15.

Risk Factor	Former Smokers (584 cases / 654 controls)		Current Smokers (534 cases / 481 controls)	
	Regression Coefficient	OR (95% CI)	Regression Coefficient	OR (95% CI)
Constant	-1.000	N/A	-0.862	N/A
Emphysema	0.927	2.526 (1.771-3.604)	0.996	2.707 (1.857-3.945)
Dust Exposure	0.471	1.601 (1.254-2.044)	0.201	1.223 (0.927-1.613)
Family history (≥ 2)	0.491	1.635 (1.269-2.106)	N/A	N/A
Age Stopped Smoking 42-53 years	0.073	1.076 (0.798-1.452)	N/A	N/A
Age Stopped Smoking ≥ 54 years	0.422	1.525 (1.135-2.048)	N/A	N/A
Hay Fever(No vs. Yes)	0.377	1.459 (1.080-1.970)	0.385	1.469 (1.039-2.078)
Asbestos Exposure	N/A	N/A	0.529	1.697 (1.141-2.524)
Smoking Family History (≥ 1)	N/A	N/A	0.571	1.769 (1.306-2.397)
Pack-Years (28-41.9)	N/A	N/A	0.280	1.323 (0.877-1.994)
Pack-Years (42-57.4)	N/A	N/A	0.369	1.446 (0.970-2.157)
Pack-Years (≥ 57.5)	N/A	N/A	0.679	1.972 (1.340-2.902)
Chromosome 15 Haplotype				
AA vs. AG	0.328	1.388 (0.631-3.054)	-0.068	0.934 (0.402-2.170)
GG vs. AG	-0.001	0.999 (0.458-2.181)	-0.219	0.803 (0.357-1.806)
GA vs. AG	0.224	1.251 (1.052-1.488)	0.239	1.270 (1.044-1.544)
Chromosome 5 Haplotype:				
AT vs. CC	-0.236	0.790 (0.630-0.990)	-0.469	0.625 (0.489-0.800)
AC vs. CC	-0.145	0.865 (0.661-1.131)	-0.238	0.788 (0.581-1.069)
CT vs. CC	-0.087	0.916 (0.674-1.246)	-0.088	0.916 (0.655-1.279)

With this new model, the discriminatory power increases slightly from 0.659 (95% CI = 0.636-0.681) in the original Spitz model to 0.675 (95% CI = 0.653-0.697), and this increase is significant ($p = 0.003$). Also, the new model has much better ability to classify case/control status in all individuals compared to the original Spitz model according to the NRI value, (0.254, $p\text{-value} = < 0.001$). This NRI increase is greater with true cases (NRI = 0.349, $p\text{-value} = < 0.001$), compared to controls (NRI = -0.094, $p\text{-value} = 0.001$).

3.3.4. Results: Comparison of three genetic model extensions

When comparing these three models created in sections 3.3.1 through 3.3.3 and the original Spitz model, the discriminatory power (with the AUC) and the NRI were calculated. Discriminatory power results between each model (and the 95% CI's) are listed in Table 3.5.

Table 3.5: Discriminatory power for the three genetic models and the Original Spitz Model

All discriminatory power calculations are based on relative risk calculations, and P-values were determined with NCSS/PASS software. For the p-value columns:

(1) = Comparing the Original Spitz lung cancer risk models to the three new risk models defined in this chapter;

(2) = Comparing the Spitz models with one SNP added to the other two genetic models; and

(3) = Comparing the Spitz models with 3 SNPs to the Spitz models with Haplotypes

Discriminatory Power in All Individuals (1116 cases and 1135 controls)				
Model	AUC (95% CI)	p-value		
		(1)	(2)	(3)
Original Spitz	0.659 (0.636-0.681)	---	---	---
Original Spitz + rs8034191	0.668 (0.646-0.690)	0.0183	---	---
Original Spitz + 3 SNPs	0.675 (0.652-0.696)	0.0036	0.103	---
Original Spitz + Haplotypes	0.676 (0.653-0.697)	0.0024	0.075	0.383
Discriminatory Power for Former Smokers (583 cases and 654 controls)				
Original Spitz	0.641 (0.609-0.671)	---	---	---
Original Spitz + rs8034191	0.650 (0.619-0.680)	0.073	---	---
Original Spitz + 3 SNPs	0.655 (0.623-0.684)	0.036	0.290	---
Original Spitz + Haplotypes	0.655 (0.623-0.684)	0.039	0.324	0.842
Discriminatory Power for Current Smokers (533 cases and 481 controls)				
Original Spitz	0.673 (0.639-0.705)	---	---	---
Original Spitz + rs8034191	0.683 (0.649-0.714)	0.106	---	---
Original Spitz + 3 SNPs	0.692 (0.658-0.722)	0.034	0.208	---
Original Spitz + Haplotypes	0.693 (0.660-0.724)	0.020	0.131	0.276

With these results, for all individuals and current smokers, the models with either three SNPs or two haplotypes are significantly superior to the Original Spitz models. However, there are no significant differences (at the 5% level) among the genetic models. To test the ability of the genetic models to improve the risk profiles of each individual compared to the original Spitz Model, the NRI was calculated for each genetic model compared to the Spitz model, and for each genetic model against each other (Table 3.6).

Table 3.6: Net Reclassification Index results for comparing the three extensions of the Spitz risk models with the Original Spitz Model

All Net Reclassification Index results (NRI) and P-values were determined with MATLAB. For the p-value columns:

- (1) = Comparing the Original Spitz lung cancer risk models to the three new risk models defined in this chapter;
(2) = Comparing the Spitz models with one SNP added to the other two genetic models; and
(3) = Comparing the Spitz models with 3 SNPs to the Spitz models with Haplotypes

NRI for All Individuals						
Model	(1)		(2)		(3)	
	NRI	p-value	NRI	p-value	NRI	p-value
Original Spitz	---	---	---	---	---	---
Original Spitz + rs8034191	0.183	< 0.001	---	---	---	---
Original Spitz + 3 SNPs	0.269	< 0.001	0.167	< 0.001	---	---
Original Spitz + Haplotypes	0.255	< 0.001	0.159	< 0.001	0.099	0.018
NRI for Cases						
Original Spitz	---	---	---	---	---	---
Original Spitz + rs8034191	0.350	< 0.001	---	---	---	---
Original Spitz + 3 SNPs	0.358	< 0.001	0.005	0.858	---	---
Original Spitz + Haplotypes	0.349	< 0.001	-0.020	0.510	0.201	< 0.001
NRI for Controls						
Original Spitz	---	---	---	---	---	---
Original Spitz + rs8034191	-0.167	< 0.001	---	---	---	---
Original Spitz + 3 SNPs	-0.089	0.003	0.161	< 0.001	---	---
Original Spitz + Haplotypes	-0.094	0.001	0.179	< 0.001	-0.103	< 0.001

For the NRI results, the genetic models with SNPs from the GWAS and the genetic model with haplotypes improve the risk profile results for both cases and controls with the 3 SNP model having significant superior values in NRI for all individuals (NRI = 0.269, p-value = < 0.001) and cases (NRI = 0.358, p-value = < 0.001) compared to the original Spitz model. Interestingly, not one genetic model improves discrimination in controls as all of them are significantly worse than the original Spitz model. The genetic model with haplotypes does worse in discrimination with controls compared to the model with three SNPs (NRI = -0.103, p-

value = < 0.001), but the haplotype model measures discrimination in cases better than the three SNP genetic model (NRI = 0.201, p-value = < 0.001).

3.4. Discussion: Adding Haplotypes and SNPs to the Spitz Lung Cancer Risk Model

These results show that adding SNPs or haplotypes can increase the discriminatory power compared to the original Spitz model. The largest increases in discriminatory power occur when the two haplotypes on chromosomes 15 and 5 are added to the original Spitz model. However, the NRI gives contradictory evidence that the haplotype model outperforms the three SNP model. When compared directly to each other, there is significant improvement for the haplotype model in terms of NRI for all individuals, but the three SNP model improves discrimination between cases and controls the most compared to the original Spitz model. Also, for all individuals, former smokers, and current smokers, there are no significant differences in AUC between the haplotype model and the three SNP model.

A potential reason for the non-significant increase in discriminatory power with the haplotype model compared to the model with 3 SNPs is that the haplotypes may not be generally true haplotypes. Even though they are listed as areas of strong LD in the Texas lung GWAS manuscript⁴⁵, data from HapMap in the Caucasian (CEU) population uploaded into Haploview suggest quite the opposite for the Chromosome 5 haplotype containing SNPs rs2736100 and rs401681 (Figure 3.1)

Figure 3.1: Gold Plot from Haploview for the SNPs in HapMap between rs2736100 and rs401681.

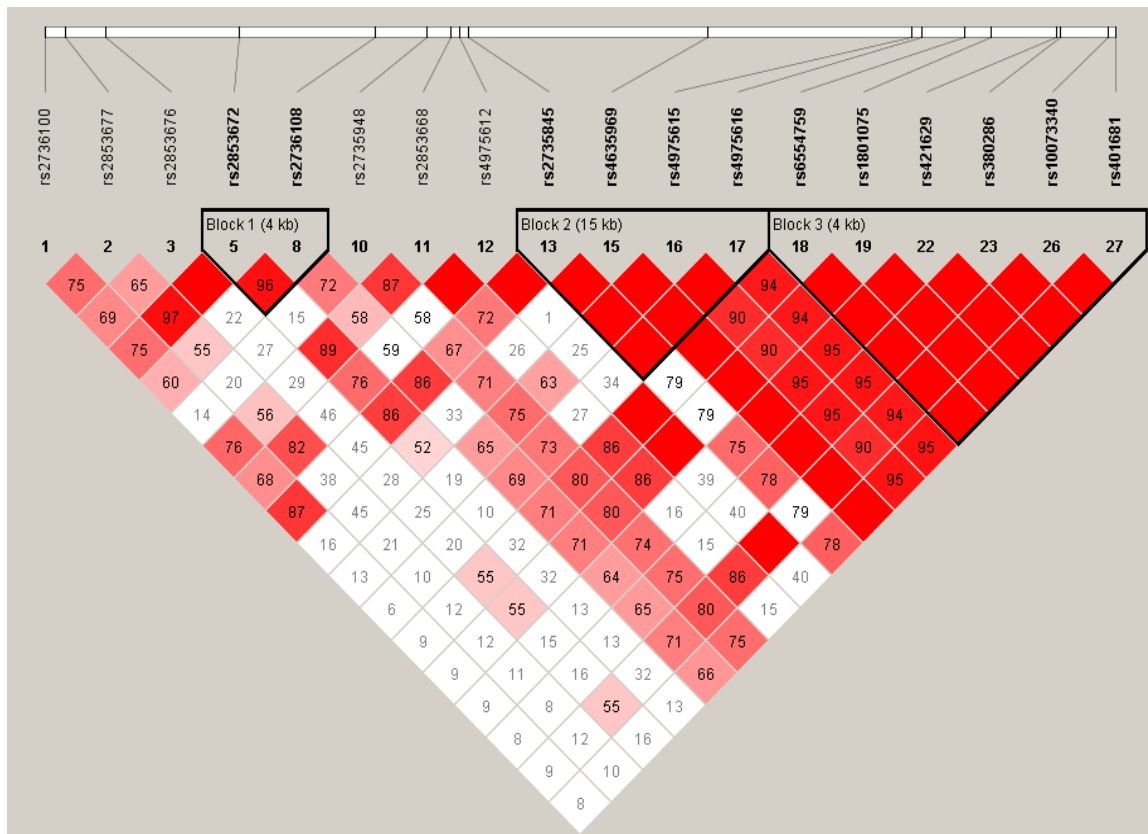


Figure 3.1: This figure highlights three haplotype blocks inferred via the Gabriel et. al method⁵³. SNP rs401681 is located within one haplotype block, while SNP rs2736100 is not located within the same haplotype block. Also, and of higher importance, the LD between SNPs rs2736100 and rs401681 is weak, as expressed by the white diamond with the value 8 in the middle at the bottom of the plot. That value represents the D' for these two SNPs (0.08), and the 95% CI of D' for these two SNPs are from 0.00 to 0.23.

With the weak LD between these two SNPs, it could be argued that these SNPs should have been modeled separately instead of a haplotype. In this case, modeling these SNPs as a haplotype may have lead to weaker modeling, and a non-significant increase in discriminatory power compared to the model with 3 SNPs and no haplotypes. If haplotypes in stronger LD are found in the full Texas lung GWAS dataset, there is a possibility that these haplotypes can lead to

model development with higher discriminatory power compared to models with either weak LD haplotypes or with SNPs only.

3.5. Why we need the BJLM

In the previous three chapters, risk modeling and haplotype analysis have been introduced. Chapter 2 showed that risk models with basic epidemiological factors like smoking status, self-reported health co-morbidities, and family history for disease can moderately discriminate between cases and controls, and occasionally show good clinical utility. The highest discriminatory power recorded was 0.74 from the current smoker set of the NELSON trial, but ever smokers recorded a discriminatory power of only 0.69 in both of the validation data sets. Excellent risk models have discriminatory power of 0.80, while good and very good risk models generally have discriminatory power values of 0.70 and 0.75, respectively¹³¹.

In an attempt to increase the discriminatory power of the Spitz lung cancer model, three new models were created with 1 SNP, 3 SNPs, and 2 haplotypes added to the original Spitz lung cancer model in Chapter 3. These models showed a general increase in discriminatory power, but this increase was small and non-significant when trying to add haplotypes to the original Spitz model compared to adding SNPs to the original Spitz model. According to numerous manuscripts, haplotypes are supposed to have greater power compared to individual SNPs, but the results from this analysis may suggest otherwise. Can modeling and selection of haplotypes be improved such that increased discriminatory power will be achieved in risk models that could be use to select individuals for screening trials or to identify individuals at high-risk for incidence of disease? This will be attempted in the next chapter with the introduction of the BJLM.

Chapter 4: Development of the Bayesian Joint Logistic Method

4.1. Relationship between the BJLM and Haplotype Analysis

To attempt to improve the effectiveness of using haplotypes into risk modeling, I will develop the Bayesian Joint Logistic Method (BJLM). This chapter will contain the full development of the BJLM with both simulation results for development of haplotypes and also results from constructing a risk model for the Hodgkin dataset. Haplotype analysis is the study of a pattern of descent of a set of linked alleles occurring on the same chromosome⁶⁰, and this analysis has been used to discover sets of linked markers associated with specific diseases. With haplotype analysis, more power can be obtained in discovering a link between haplotypes and disease compared to the case of just examining the relationship between a SNP and disease¹³². Haplotype analysis has been used to discover new genetic risk factors in a wide variety of cancers, including breast, pancreatic, and Hodgkin lymphoma¹³³⁻¹³⁵.

Some programs that infer haplotypes for haplotype analysis are PHASE⁵⁵, Haploview⁵⁶, and Haplo.stats⁶², with Haplo.stats having the ability to construct frequentist models of haplotype data. Haplo.stats allows for the estimation of odds ratios to represent the association between either a haplotype or a covariate and disease with the use of either a joint-effect linear model or a separate-effect linear model. Currently there exist two prominent frequentist methods for risk model development with inferred haplotypes (Joint Logistic Analysis, Separate-Effects Model)^{62,65}; however, no Bayesian method currently exists that not only infers haplotypes, but also can directly elucidate risk models that can be used to develop genetic risk profiles for disease.

Linkage disequilibrium (LD) studies have accelerated in recent years with the development and implementation of a haplotype map of the human genome (HapMap) beginning

in 2003, and with the newest version, HapMap 3, fully updated in May 2010¹³⁶⁻¹³⁹. Currently, there are over 1,000,000 QC+ SNPs genotyped and non-redundant from 11 different populations throughout the world. Extracting this information could be very useful with inferring haplotypes more accurately which could lead to stronger associations with disease.

I hypothesize that there could be substantial improvement in the association between haplotypes and disease, with the use of a Bayesian mathematical method to infer haplotypes that uses priors developed from known genetics sources (HapMap). The development of this model, the Bayesian Joint Logistic Model (BJLM), will be completed in four stages; 1) extraction of the haplotypes and their counts from HapMap data 2) the framework, in this case a Hidden Markov Chain Monte Carlo (MCMC) model with mixing, with associated Dirichlet prior from HapMap data to begin the estimation of haplotype frequencies, 3) a method to infer haplotypes that contains missing SNP data, and 4) the logistic model to estimate the risk for each haplotype reconstructed by the BJLM. The complete developmental flow-chart for the BJLM is shown below in Figure 4.1.

**Figure 4.1: Flow-Chart of the BJLM from the 1st step: HapMap to the Final Step:
Creation of a Bayesian Logistic Model**

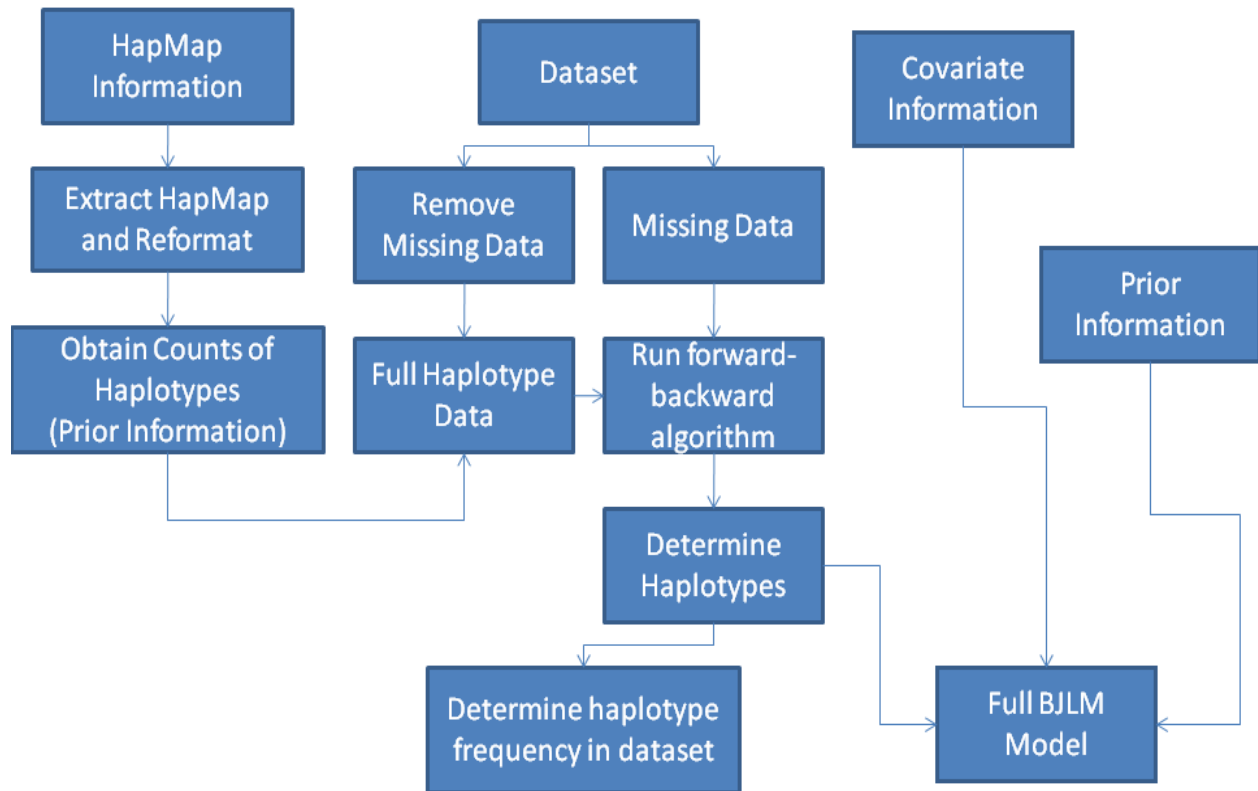


Figure 4.1: In this flow chart, the 1st section is to extract the SNP data from HapMap and infer haplotypes using a frequentist method in which the haplotype frequencies of all haplotypes in a haplotype block are updated until the likelihood stays within 0.001 of the previous iteration. Then, this information is used as prior information to determine the haplotypes with full SNP data for each haplotype block. Then missing haplotypes (those with partial SNP data) are inferred using a forward-backward algorithm to determine all haplotypes for all individuals. Finally, this haplotype information is added to the covariate information, and a Bayesian logistic risk model is created in WinBugs. Full details are expressed in the next section.

4.2. Development of the framework for the BJLM

4.2.1. Extracting information from HapMap for BJLM use

Programs like Haploview⁵⁶ have been created to read HapMap data in order to generate LD statistics, develop Haplotype blocks⁵³, determine Hardy-Weinberg equilibrium¹⁴⁰ and determine which SNPs are tagSNPs¹⁴¹. Unfortunately, the newest version of HapMap has formatting which is not fully compatible with SNP data from the non-CEU (or Caucasians currently living in UTAH) population (version 4.2). It is also imperative a standalone program exist in order to extract the newest HapMap data. I created a MatLab-based program called HapExtract that accepts HapMap dumped genotype data from the Human HapMap website (<http://www.hapmap.org>), and the program is flexible enough to accept genotype data from each of the 11 populations in HapMap3. Downloaded genotype data from HapMap is bounded by the base-pair information of the SNPs at each end of the portion of the genome to be examined. For example, the input chr1:11777000..11778970 for the HapMap website will extract all genotypes from that region of the chromosome. With these data from HapMap, all of the genotypes can be converted into more user-friendly data, or a list of SNPs can be inputted into HapExtract that will only extract the HapMap data from the inputted SNP list.

A portion of the HapMap genotype dump file is shown below in Figure 4.2.

Figure 4.2: Sample section from HapMap genotype data dump

```
#Tue Mar 23 10:41:25 2010: HapMap genotype data dump, SNPs genotyped in population CEU
on chr1:11777000..11778970
#For details on file format, see http://www.hapmap.org/genotypes/
rs# alleles chrom pos strand assembly# center protLSID assayLSID panelLSID QCcode NA06984 ..
rs4846051 A/G chr1 11777044 + ncbi_b36 perlegen
urn:lsid:perlegen.hapmap.org:Protocol:Genotyping_1.0.0:2
urn:lsid:perlegen.hapmap.org:Assay:25769.987725:1 urn:lsid:dcc.hapmap.org:Panel:CEPH-30-
trios:1 QC+ NN ...
rs1801131 G/T chr1 11777063 + ncbi_b36 broad
urn:LSID:affymetrix.hapmap.org:Protocol:GenomeWideSNP_6.0:3
urn:LSID:broad.hapmap.org:Assay:SNP_A-8699092:3 urn:lsid:dcc.hapmap.org:Panel:CEPH-60-
trios:3 QC+ TT ...
rs12121543 A/C chr1 11777258 + ncbi_b36 broad
urn:LSID:affymetrix.hapmap.org:Protocol:GenomeWideSNP_6.0:3
urn:LSID:broad.hapmap.org:Assay:SNP_A-8302865:3 urn:lsid:dcc.hapmap.org:Panel:CEPH-60-
trios:3 QC+ CC...
rs6541003 A/G chr1 11778454 + ncbi_b36 sanger
urn:LSID:illumina.hapmap.org:Protocol:Human_1M_BeadChip:3
urn:LSID:sanger.hapmap.org:Assay:H1Mrs6541003:3 urn:lsid:dcc.hapmap.org:Panel:CEPH-60-
trios:3 QC+ AG ...
rs1801133 A/G chr1 11778965 + ncbi_b36 sanger
urn:LSID:illumina.hapmap.org:Protocol:Human_1M_BeadChip:3
urn:LSID:sanger.hapmap.org:Assay:H1Mrs1801133:3 urn:lsid:dcc.hapmap.org:Panel:CEPH-60-
trios:3 QC+ GG ...
```

Figure 4.2: The first three lines are the standard printout for any HapMap genotype dump, regardless of the population dataset being extracted. The data between the blue lines list the components of the file, and the ids of all individuals that have genotype data. All data below the blue lines are the genotypes and rs numbers which are collected by the program, and all other information between the genotypes and the rs numbers are discarded. Finally, the bolded SNPs represent the information that the user wants saved in the HapMap file for future analysis.

After selecting the SNPs used for future analysis, the genotype information for each individual in the HapMap data set in allele form (A, T, C, G). Upon obtaining the HapMap data set in allele form, HapExtract converts the information into a Haploview ready format with two files; the first file contains the marker name and the base-pair information, and the second file contains the allele information for each marker in Haploview format (A=1,C=2,G=3,T=4). Next,

the user can translate the allele information to a haplo.stats ready format if applicable. Then, the allele information can be translated into the number of variant alleles for each SNP. This format could be used for haplotype analysis with the program HapReg¹⁴², or for future haplotype analysis in the form of calculating haplotype frequencies from the HapMap data.

Within each HapMap population, there are some pockets of missing SNP data, which is represented by the allele NN, and this information is removed from future use. Output examples from the HapMap extraction program are shown in Figure 4.3.

Figure 4.3: Output Example with both Missing SNP Data and without Missing SNP Data

A) HapMap Format

Individual	rs1801131	rs1801133
1	TT	GG
2	GT	GG
3	GT	AG
4	TT	AG
5	NN	NN
6	TT	AG
7	GG	GG
8	TT	GG
...
171	GT	AG
172	TT	AA
173	GT	AG
174	GG	GG

B) Variant Allele Format

Individual	rs1801131	rs1801133
1	0	0
2	1	0
3	1	1
4	0	1
6	0	1
7	2	0
8	0	0
...
171	1	1
172	0	2
173	1	1
174	2	0

Figure 4.3: Figure 4.3.A is the formatted HapMap genotype data from HapMap for all 174 individuals in the CEU population, and figure 4.3.B. contains the number of variant alleles for each CEU individual in the SNPs rs1801131 and rs1801133. In this population, CEU individual 5 has missing genotype data, coded by NN in figure 4.3A. For determining the set of haplotypes that will serve as prior data for inferring haplotypes, this individual will be removed from the analysis.

After obtaining the number of variant alleles for each individual, the haplotypes from HapMap are inferred by using a simple algorithm that continues until the likelihood changes by less than 0.001 (Figure 4.4). Below is the calculation for the likelihood

$$L(haplotypes) = \left(\sum_{i=1}^M \ln(\theta_i) \right) + \ln(\theta_{begin}) \quad (4.1)$$

where $i = 1, 2, \dots, M$ and M = number of haplotypes in the HapMap sample, θ_i = haplotype frequency for each haplotype in the HapMap sample, and θ_{begin} = initial set of haplotype frequencies. For the initial set of haplotype frequencies, these frequencies represent an equal percentage for each possible haplotype.

Figure 4.4: Algorithm that determines the Haplotypes inferred from HapMap individuals that is used for the BJLM

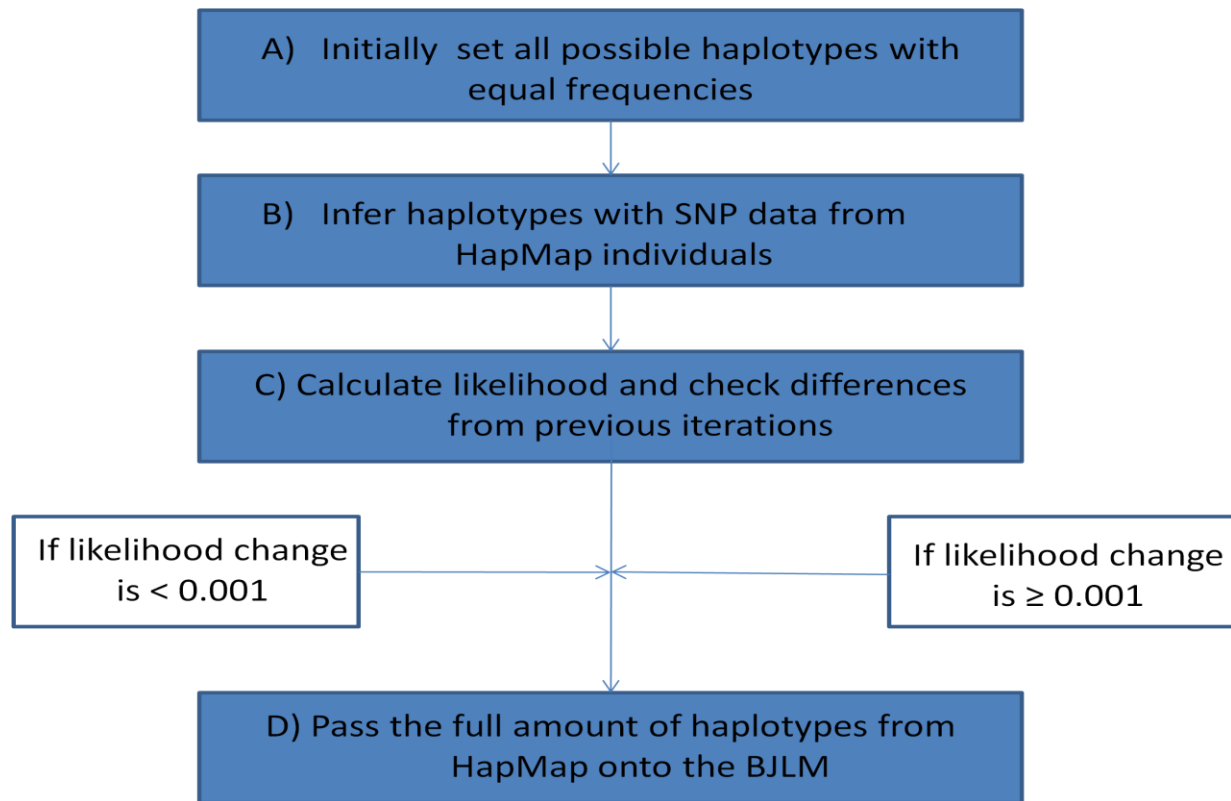


Figure 4.4: All potential haplotypes are inferred at an equal frequency for the 1st step. Then, with these frequencies and the possible haplotypes for every individual in the HapMap dataset according to the available variant allele data, new haplotype frequencies for each potential haplotypes are calculated. These frequencies are inserted into equation 4.1, and the likelihood is calculated. Steps B and C are conducted until the change in haplotype from one iteration to another are less than 0.001. Finally, once this algorithm is completed, and the set of haplotypes and their counts determined, they are passed onto the BJLM for analysis with the dataset to be examined.

4.2.2. Determining Haplotypes from Full Data Using Information from HapMap

In the previous section, haplotypes inferred from HapMap SNP data was inferred using a likelihood algorithm as expressed in Figure 4.4. In this section, the second component of the BJLM includes the framework for the determination of haplotypes from a disease dataset to be examined, which will be a Hidden MCMC model. MCMC models can be used to extract information for haplotypes because of their ability to sample quickly through genotypes to produce haplotypes without making unnecessary assumptions like phase determination⁶¹. In the literature, some of the more well-known examples of using Bayesian MCMC models to infer haplotypes are from Xing et. al¹⁴³, Stephens et. al.^{55,61}, Niu et.al¹⁴⁴, and Bansal et. al¹⁴⁵. Gibbs sampling¹⁴⁶ will be used to sample out the haplotypes in question. With Gibbs sampling, one can extrapolate unknown haplotypes, if one can determine the frequencies of alleles for each loci, the genotypes for each individual, and the prior distribution in question, $g(\theta_1, \theta_2, \dots, \theta_h)^{147}$.

The prior to be used for the BJLM will be previous haplotype data extracted from HapMap, and the procedure for this was discussed in **section 4.2.1**. With this prior, haplotype information can be examined from a known, verified source, and incorporated into the analysis for all individuals with complete SNP data. For calculation of the haplotypes with complete SNP data from the disease dataset, the haplotypes from HapMap will be modeled from a Dirichlet distribution. This distribution made sense to use as the prior distribution because the haplotype frequencies estimated from HapMap equaled one, and the computation simplicity of the Dirichlet distribution. The Dirichlet distribution is the conjugate prior for multinomial distribution, and that is how the genotypic counts are modeled, hence the computational simplicity of this distribution for Bayesian haplotype analysis (Equation 4.2).

$$f(\theta_1, \dots, \theta_n) = \frac{\Gamma(\beta_1 + \dots + \beta_n)}{\Gamma(\beta_1) * \dots * \Gamma(\beta_n)} \theta_1^{\beta_1-1} * \dots * \theta_n^{\beta_n-1}, \text{ where } \sum \theta_i = 1, \theta \geq 0 \quad (4.2)$$

where θ_i is the haplotype frequency, and β_1 is the hyper parameter value for the haplotype based on the racial composition of the group.

With the prior information from HapMap, one can determine the probability of a previously unknown haplotype with the associated genotypes and haplotype frequencies $P(Y, Z, \Theta)$ according to¹⁴⁴:

$$P(Y, Z, \Theta) \propto \sum_{i=1}^n \theta_{z_{i_1}} \theta_{z_{i_2}} \sum_{g=1}^M \theta_g^{\beta_g-1} \quad (4.3)$$

with Y is the genotypes, Z is the haplotypes, Θ is the set of population haplotype frequencies, $\theta_{z_{i_1}}$ is the frequency for assigned haplotypes, n is the number of individuals, M is the number of possible haplotypes, $\theta_g^{\beta_g-1}$ is the maximum likelihood estimate at each haplotype (g) according to the parameters of the Dirichlet distribution (β)¹⁴⁴.

The iterations of the above parameters, can update the haplotype frequencies given the genotypes and haplotypes, hence constituting a Gibbs sampling algorithm¹⁴⁴. Eleven hundred iterations are conducted to determine haplotype frequency, and this occurs in three separate processes. The first 100 iterations to determine haplotype frequency will assume no prior information to remove non compatible haplotypes between HapMap data and the dataset to be analyzed. The second set of 100 iterations incorporate the HapMap prior information into the haplotype draw, and are used as a burn-in set for which the chain of frequency values for each haplotype will become stable again. Finally, 900 iterations will be conducted that will determine the haplotype draws for each individual, the mean frequency for each haplotype, and the 95% credible interval for each haplotype. The haplotype information for the dataset with all SNP

information will be used to then determine the missing haplotypes for those with incomplete SNP information (if applicable) with the forward-backward algorithm for the 3rd part of the BJLM.

4.2.3. Using a Modified Forward-Backward Algorithm to Infer Haplotypes for Individuals with Missing SNP Data

Since real data sets sometimes contain missing genotype data, whether due to genotypic error or human error, the third step in the Bayesian framework is to impute the missing haplotype pair, if applicable. Estimation of missing data will occur in cases and controls separately since the frequency of haplotypes will be different in cases and controls. This will be accomplished with the use of a forward-backward algorithm, which can estimate the missing haplotype pair by taking into account the probability of haplotypes as determined in **section 4.2.2** both before and after the individual with missing SNP data¹⁴⁸⁻¹⁵⁰. A flowchart for the modified forward-backward algorithm and its application to missing haplotype data is shown in Figure 4.5.

Figure 4.5: Determining Missing Haplotype Pair from individuals with missing SNP data with the Forward-Backward Algorithm

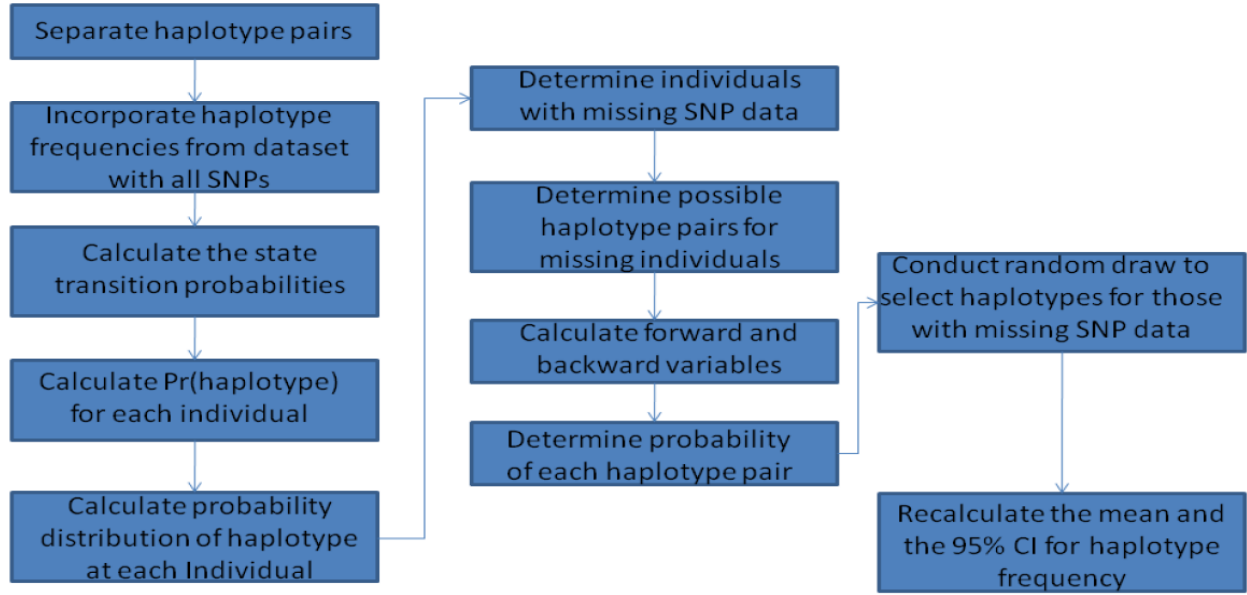


Figure 4.5: First, the forward-backward algorithm will be conducted by separating the haplotype pair so that two sets exist which contain the 1st haplotype and the 2nd haplotype for each missing individual separately. Second, haplotype frequencies from the whole dataset will be calculated for both haplotype sets. Third, the state transition probabilities (a_{ij}) will be calculated by multiplying the haplotype frequencies calculated in the previous step (Equation 4.4).

$$a_{ij} = P(q_t = \theta_j | q_{t-1} = \theta_i) = P(\theta_j) \quad (4.4)$$

where θ_i and θ_j = haplotypes that exist in the haplotype sets, i and $j = 1, \dots, N$, where N = number of haplotypes available for each haplotype set, and q_t = actual haplotype at individual t . Fourth, for each individual, the haplotype frequencies for each individual (π_i) based on the 900 iterations of the BJLM conducted in **section 4.2.2** is calculated (Equation 4.5).

$$\pi_i = \frac{(\sum_{n=1}^{iter} P[q_t = \theta_i])}{iter} \quad (4.5)$$

Fifth, the probability distribution of a haplotype based for each individual $b_i(O_t)$ is calculated by using the *mnrnd* function in Matlab, which selects random values from a

multinomial distributions for as many times as the user desires given a set of multinomial probabilities. In this case, the *mnrnd* function incorporates the haplotype probabilities for all possible haplotypes and conducts a random haplotype draw for every individual for as many times as the user desires. This probability distribution assumes that no information on the actual haplotypes for each individual with missing SNP data is known, and that the probability is just based on the haplotype frequencies for the dataset with full SNP information. In this case, the number of random draws equal the number of cases or controls in the study (depending on whether the program is determining haplotypes for missing cases or controls) (Equation 4.6).

$$b_i(O_t) = \frac{(\sum_{n=1}^{\#cases \text{ or } \#controls} P[r_t = \theta_i])}{cases \text{ or } \#controls} \quad (4.6)$$

where r_t = random draw of haplotype i at individual t .

Sixth and seventh, the individuals with missing SNP data are selected, and then the potential haplotype pairs for those with missing data are determined by the haplotypes inferred with the full data set. This ensures that the only haplotypes that can be selected for an individual will match the haplotypes already in the sample.

The next step of the forward-backward algorithm involves the calculation of the forward and backward variables. For the forward portion of the algorithm, the forward variable is initially estimated at the 1st individual in the data set by determining the probability that the state of Markov chain equals the actual state (probability that the haplotype exists for the individual) for the 1st individual times the probability distribution of an observation at the 1st individual (which is the probability of the haplotype in each part of the pair based on the full dataset)¹⁴⁹. Then, the forward variable is updated by taking into account all of the probabilities for each haplotype based on the full dataset, and also the probability of a haplotype at the next individual.

This continues until the individual is reached with the missing haplotype. The equations for the forward portion of the algorithm are¹⁴⁸:

$$\alpha_1(i) = \pi_i b_i(O_1), 1 \leq i \leq N \quad (4.7)$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(O_{t+1}), 1 \leq t \leq T-1, 1 \leq i \leq N \quad (4.8)$$

where t = individual up to missing individual (T), $t = 1, 2, \dots, T-1$, $b_i(O_t)$ = probability distribution of a haplotype based on the full dataset at individual t , N = number of haplotypes in the full SNP dataset, q_t = actual haplotype at individual t , where θ_i and θ_j = haplotypes that exist in the haplotype sets, i and $j = 1, \dots, N$, a_{ij} = set of transition probabilities from one haplotype to another, and $\alpha_t(i)$ = forward variable used in calculation of the missing haplotype¹⁴⁹.

The backward portion of the algorithm begins at either the individual before the next missing individual, or the end of the disease dataset if no more missing individuals exist. This portion is initialized with the backward variable equaling one at the first individual, and then inducted by taking into account the backward variable value at individual t , probability of haplotype at individual t , and probability of change from one haplotype to another while the program heads backward toward the individual with the missing haplotypes. The equations for the backward portion of the algorithm are¹⁴⁸:

$$\beta_Q(i) = 1, 1 \leq i \leq N \quad (4.9)$$

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), t = Q-1, \dots, 1, 1 \leq i \leq N \quad (4.10)$$

where Q = next missing individual (or end of disease dataset if no more missing individuals exist), $\beta_t(i)$ = backward variable at individual t .

After determining the estimates of the forward and backward variables, we can then determine the probability of each haplotype at the individual with missing haplotype data, $\gamma_t(i)$.

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^N (\alpha_t(i)\beta_t(i))}, i = 1, \dots, N \quad (4.11)$$

We then determine the missing haplotype pair by multiplying the probabilities of the haplotypes for each separate haplotype pair and conducting a random draw to select the haplotype pair for each missing individual. This random draw is conducted for each iteration of the BJLM that was used to determine the initial haplotypes and their frequencies in **section 4.2.2**. Finally, the haplotypes frequencies are calculated once again with the 95% credible intervals, and the full haplotype pairs for all individuals are passed onto the modeling section of the BJLM.

4.2.4. Development of the Bayesian Logistic Model for Haplotype Association

The third component of the BJLM model is the incorporation of a Bayesian binary logistic model to link the case-control status of a disease (dependent variable) and the haplotypes constructed in the earlier components of the BJLM (independent variables). Logistic models have been used extensively throughout epidemiology, especially in developing models to estimate probability for a range of cancers^{27,30,31,151,152}. With the BJLM model, the intention is to introduce the use of a Bayesian binary logistic model for haplotype analysis to both estimate associations between haplotypes and disease and also to increase the power to detect the association, when compared to classical logistic models. Bayesian logistic models incorporate the logistic model with a prior multiplied to the exposure as shown in equations 4.12 and 4.13¹⁵³.

$$p(y = +1|\boldsymbol{\beta}, \mathbf{x}_i) = \psi\left(\sum_t \beta_t \chi_{i,t}\right) \quad (4.12)$$

$$p(\beta_t|\tau_t) = N(0, \tau_t) = \frac{1}{\sqrt{2\pi\tau_t}} \exp\left(\frac{-\beta_t^2}{2\tau_t}\right), t = 1, \dots, d \quad (4.13)$$

where $\boldsymbol{\beta}$ is the vector of Gaussian priors for each individual t (β_t), \mathbf{x}_i is the vector of value of exposed risks (i) to each individual (t) $\chi_{i,t}$. ψ is the logistic risk function, τ_t is the variance of β_t , and d is the number of individuals in analysis. When construction of the Bayesian logistic model is complete, the BJLM model can begin to be used for preliminary analysis of data with the use of an MCMC method¹⁵⁴⁻¹⁵⁶.

To conclude the construction of the BJLM model for covariates, methods to incorporate binary, categorical, and continuous covariates must be taken into account. For any binary covariates, the same procedure used to incorporate haplotypes into the BJLM (see eqs. 4.12 and 4.13) will be used to incorporate binary covariates into the model.

However, for categorical and continuous covariates, more complicated Bayesian methods must be used. Categorical or polychotomous variables have been incorporated into Bayesian analysis previously¹⁵⁷⁻¹⁵⁹, and the approach that will be used for the BJLM model will be to separate the categorical variables into $m-1$ separate binary variables, where m is the number of categories for each categorical variable. Then, the same procedure to incorporate haplotypes into the BJLM can be used to incorporate the $m-1$ separate binary variables for each categorical variable.

Continuous covariates can be included into the Bayesian logistic model framework by incorporating smoothing priors to estimate these covariates nonparametrically¹⁶⁰. An example of a smoothing prior is a second-order Gaussian random walk prior, which allows for both

flexibility, but also lessens the impact of extreme values for the covariate¹⁵⁶. An example of such a prior is given by¹⁵⁶:

$$p(f|\tau_f^2) \propto \exp\left(-\frac{\tau_f^2}{2} \sum_{t=3}^T (f_t - 2f_{t-1} + f_{t-2})^2\right) \quad (4.14)$$

where $p(f|\tau_f^2)$ is the prior function, τ_f^2 is the variance for the continuous covariate in the sample, T is the number of individuals in the sample, and f is the entire continuous covariate value set ranked from smallest (f_1) to largest (f_T). With this smoothing prior that can be used to estimate the contribution of the continuous covariates, the BJLM can now incorporate all continuous variables without concern of variables with extreme values¹⁵⁶.

All modeling discussed in section 4.2.4. will be done using WINBUGS. For the non-informative Gaussian priors discussed earlier, all priors will have means of zero and variances of 100 so that the beta(s) are both non-informative and WINBUGS can run without any trap issues. Trap issues can occur when a prior beta value is selected that cannot be resolved by the WINBUGS program during an iteration. For each haplotype and covariate, an empirical Bayes p-value will be constructed that calculates the frequency that a beta goes above and below zero during the non-burn in iterations. Below is that equation¹⁶¹:

$$\text{Empirical Bayesian P-Value} = 2 * \min(P(\text{betas} > 0), (P(\text{betas} < 0))) \quad (4.15)$$

Generally, in Bayesian analysis, p-values are not generally used as a method to determine whether variables should or should not be in a model, but in this case, the Empirical Bayesian p-value will allow for a direct comparison between models elucidated using frequentist techniques, and models elucidated using Bayesian techniques. WINBUGS can calculate Bayesian “p-values” with the use of step functions¹⁶¹.

4.3. Haplotype Simulations to Test Ability of BJLM to infer Haplotypes

4.3.1. Procedure for Haplotype Simulations

To conduct simulations of the BJLM and its ability to infer haplotypes, genotypes from Chromosome 1 from HapMap between base pair values 11754200 and 11774700 with the CEU population as a “control” population, and the TSI population as the “case” population were extracted with the BJLM using HapExtract **in section 4.2.1**. Then, using Haploview, both populations were examined to test whether the SNPs in the base pair range stated above contained haplotypes inferred from the same SNPs in both populations. After discovering that a haplotype of 12 SNPs did exist for both populations, and these 12 SNPs were exactly the same, the genotypes for these 12 SNPs were extracted from HapExtract (Table 4.1).

Table 4.1: Marker Information for the 12 SNPs in the Simulation Analysis:

For these 12 SNPs extracted from the CEU and TSI populations, the rs number, chromosome, base-pair, and major/minor allele data is listed.

rs Number	Major/Minor Allele	Chromosome	Base-pair
rs6667720	T/C	1	11754202
rs6540999	G/A	1	11755953
rs11121828	A/G	1	11757041
rs6696752	C/T	1	11758522
rs6699881	C/T	1	11759215
rs7538516	T/C	1	11759269
rs6541001	C/T	1	11759954
rs10779765	C/T	1	11760598
rs4846048	A/G	1	11768839
rs2184226	T/C	1	11770023
rs1537516	G/A	1	11770448
rs13306556	C/T	1	11774697

Full haplotype analysis was conducted in accordance with **section 4.2.1 of the BJLM**, so that the haplotypes for the CEU and TSI populations were extracted.

For both populations, 500 haplotypes will be simulated based on the extracted haplotypes from the HapMap datasets with HapSim, which is a haplotype simulator program in the R language¹⁶². HapSim expands on previous simulation methods^{163,164} by including the patterns of linkage disequilibrium and pre-specified allele frequencies to simulate large numbers of haplotypes effectively and quickly. The functions *haplodata* and *haplosim* in the HapSim program extract the basic information from the haplotype files created with the BJLM, and then simulate the haplotypes for future analysis. Since each individual has two possible haplotypes, the two haplotypes are combined so that the numbers of variant alleles per individual are determined, and in this format, the BJLM can determine both haplotypes and haplotype frequency.

Since HapSim creates haplotypes with no missing data, the 1st run of the program will include information for all individuals. To determine the effectiveness of the forward-backward algorithm, missing data is randomly added to each SNP in the haplotype with defined frequencies of 1%, 2%, 3%, 4%, or 5%. The final cutoff for missing SNP data was selected at 5% because I assumed that SNPs would not be included in genetic analysis if the call rate was less than 95%. Then the forward-backward algorithm is applied to the missing data using the frequencies calculated from the full haplotype set in the 1st run of the program. One hundred replicates of missing data are conducted to ensure that a full spectrum of haplotype results is obtained from this analysis. Simulations will be conducted with haplotypes varying from 2 to 12 SNPs, so there will be 11 sets of analysis. In this chapter, haplotype simulations with 2, 3, and 12 SNPs will be analyzed. Appendix 2 contains the results for haplotype simulations including 4-11 SNPs.

4.3.2. Haplotype (2 SNP) Simulation Results

The two SNP haplotype simulation results are presented in Table 4.2 and Figure 4.6.

Results are shown for the full set of haplotypes with frequency > 5 %.

Table 4.2: Haplotype Simulation results for 2 SNP haplotypes.

The haplotype index number is listed in the far left column, with the associated haplotype in the next column to the right. The haplotype frequency assuming no missing data is listed in the 3rd column from the left, and the average frequency for each haplotype from the 100 replicates are listed in the column next to the no missing data column. Finally, the 95% credible interval (CI) for these frequencies from the 100 replicates is listed in the final two columns.

Haplotype #	Haplotype	No Missing Data	Median Frequency	2.5% CI	97.5% CI
1% Missing SNP Data					
1	TG	0.564	0.568	0.540	0.590
4	CA	0.263	0.262	0.241	0.286
2	TA	0.124	0.123	0.105	0.141
2 % Missing SNP Data					
1	TG	0.564	0.565	0.534	0.590
4	CA	0.263	0.266	0.238	0.290
2	TA	0.124	0.121	0.100	0.140
3% Missing SNP Data					
1	TG	0.564	0.569	0.541	0.590
4	CA	0.263	0.267	0.244	0.287
2	TA	0.124	0.121	0.105	0.138
4% Missing SNP Data					
1	TG	0.564	0.571	0.554	0.595
4	CA	0.263	0.271	0.241	0.289
2	TA	0.124	0.118	0.097	0.137
5% Missing SNP Data					
1	TG	0.564	0.571	0.544	0.592
4	CA	0.263	0.271	0.249	0.297
2	TA	0.124	0.118	0.100	0.137

Figure 4.6: Graphical representations of the haplotype frequency for simulations of 2 SNP
Haplotypes

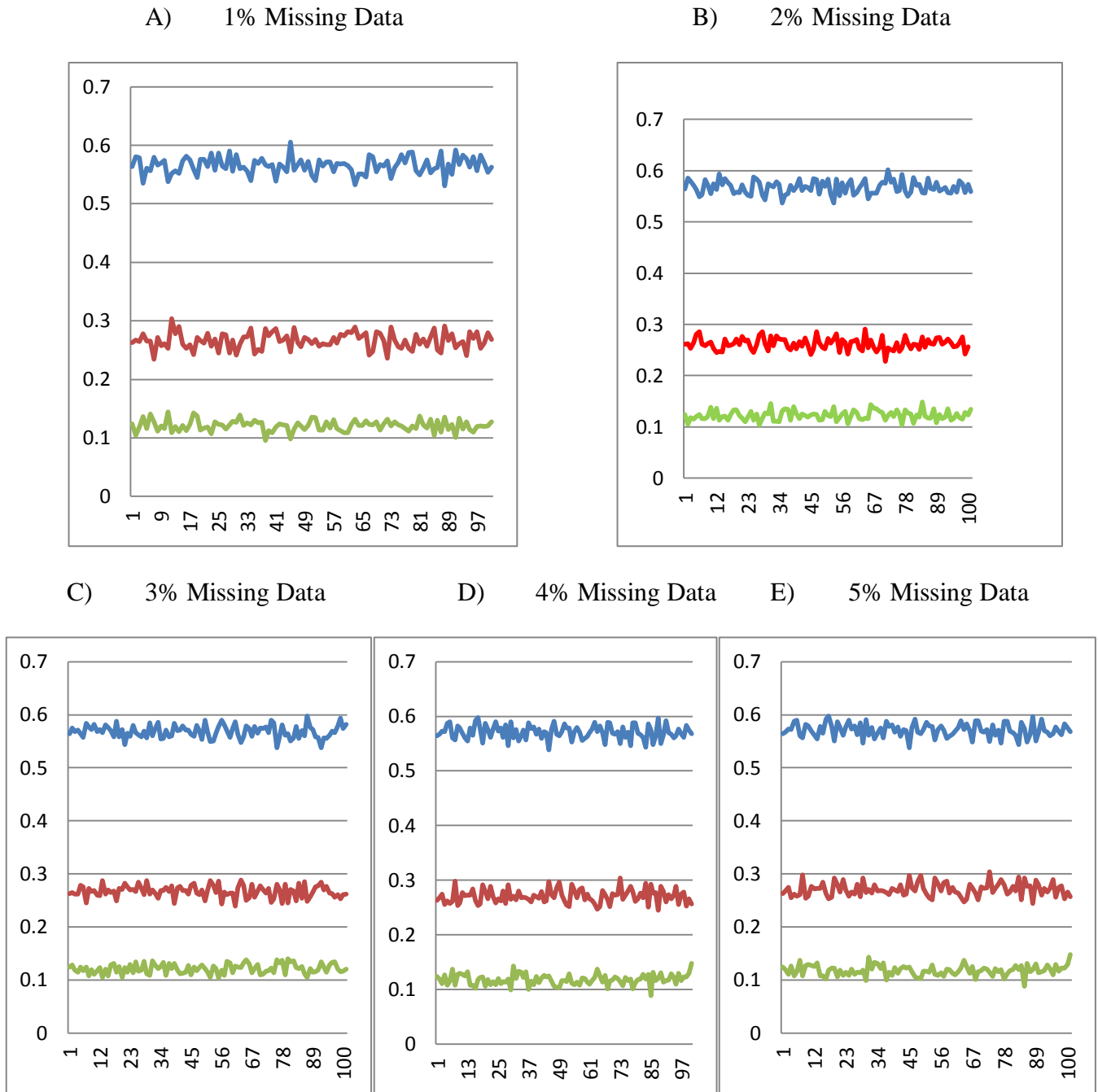


Figure 4.6: The X axis is the simulation run number, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = Haplotype TG, Red Line = Haplotype CA, and Green Line = Haplotype TA.

With these results, the missing data results are similar to the results with no missing data, as none of the haplotypes have more than a 1% difference in frequency between those with and without missing data. Also, as the percentage of missing data increases, the 95% CI only increases slightly in length. This suggests that the BJLM is very stable for 2 SNP haplotypes when missing data is 5% or less.

4.3.3. Haplotype (3 SNP) Simulation Results

The three SNP haplotype simulation results are presented in Tables 4.3 and Figures 4.7. Results are shown for the full set of haplotypes with frequency $> 5\%$.

Table 4.3: Summary of Simulated haplotypes of 3 SNP length.

The haplotype index number is listed in the far left column, with the associated haplotype in the next column to the right. The haplotype frequency assuming no missing data is listed in the 3rd column from the left, and the average frequency for each haplotype from the 100 replicates are listed in the column next to the no missing data column. Finally, the 95% credible interval (CI) for these frequencies from the 100 replicates is listed in the final two columns.

Haplotype #	Haplotype	No Missing Data	Median Frequency	2.5% CI	97.5% CI
1% Missing SNP Data					
1	TGA	0.322	0.323	0.294	0.351
2	TGG	0.242	0.246	0.219	0.275
7	CAA	0.171	0.177	0.154	0.198
8	CAG	0.092	0.084	0.069	0.099
3	TAA	0.073	0.073	0.057	0.089
4	TAG	0.051	0.051	0.038	0.064
2 % Missing SNP Data					
1	TGA	0.322	0.322	0.296	0.352
2	TGG	0.242	0.243	0.216	0.270
7	CAA	0.171	0.180	0.155	0.204
8	CAG	0.092	0.086	0.067	0.104
3	TAA	0.073	0.074	0.055	0.091
4	TAG	0.051	0.049	0.036	0.065
3% Missing SNP Data					
1	TGA	0.322	0.324	0.290	0.348
2	TGG	0.242	0.245	0.218	0.265
7	CAA	0.171	0.180	0.157	0.197
8	CAG	0.092	0.087	0.071	0.104
3	TAA	0.073	0.073	0.059	0.088
4	TAG	0.051	0.049	0.035	0.064
4% Missing SNP Data					
1	TGA	0.322	0.322	0.294	0.350
2	TGG	0.242	0.249	0.227	0.271
7	CAA	0.171	0.181	0.159	0.204
8	CAG	0.092	0.088	0.068	0.112
3	TAA	0.073	0.074	0.058	0.091
4	TAG	0.051	0.046	0.035	0.062
5% Missing SNP Data					
1	TGA	0.322	0.327	0.303	0.358
2	TGG	0.242	0.243	0.215	0.272
7	CAA	0.171	0.184	0.161	0.213
8	CAG	0.092	0.090	0.072	0.106
3	TAA	0.073	0.074	0.057	0.091
4	TAG	0.051	0.041	0.030	0.056

Figure 4.7: Graphical representations of the haplotype frequency for simulations of 3 SNP
Haplotypes

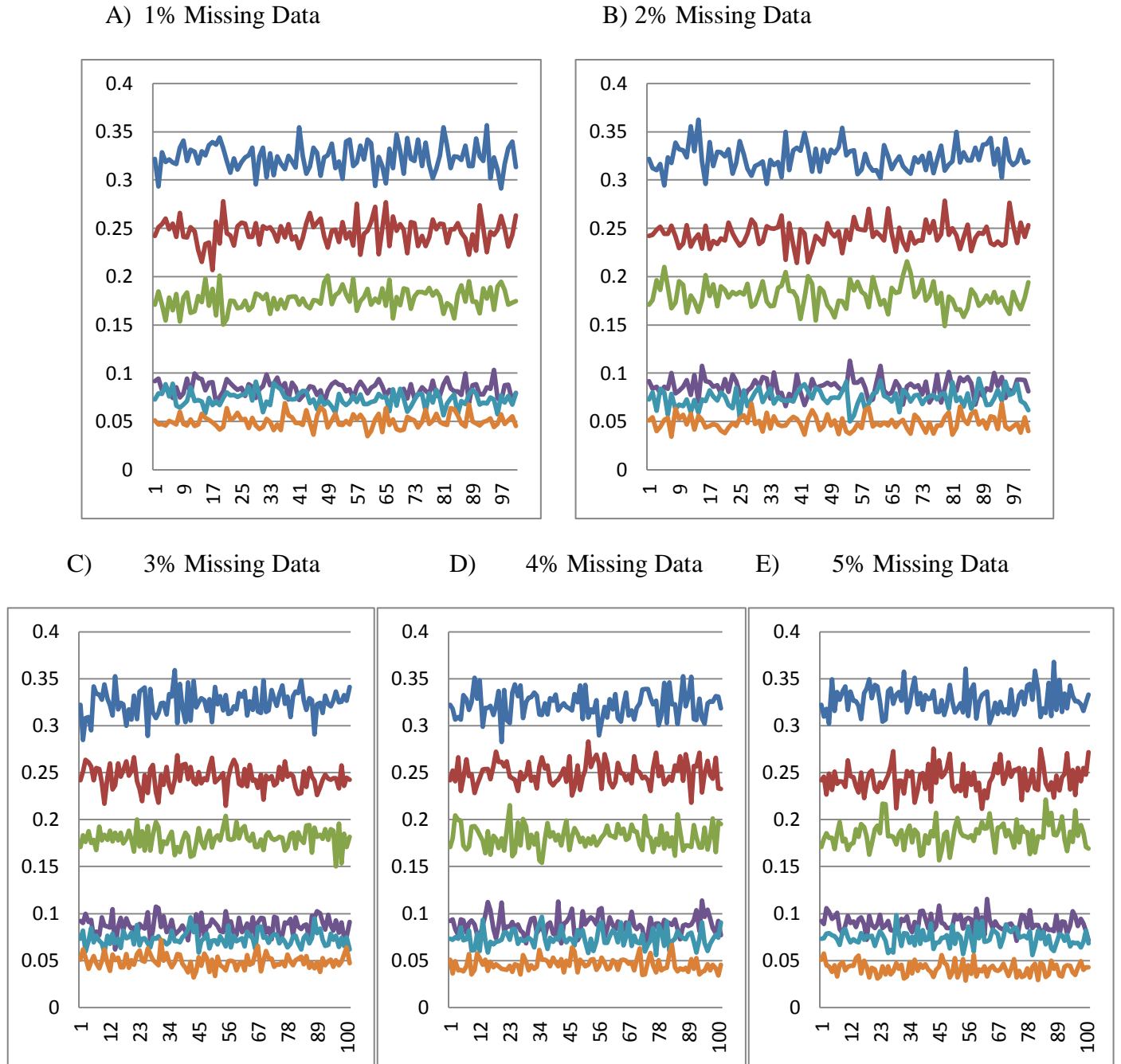


Figure 4.7: The X axis is the simulation run number, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = Haplotype TGA, Red Line = Haplotype TGG, Green Line = Haplotype CAA, Purple Line = Haplotype CAG, Light Blue Line = Haplotype TAA, and the Orange Line = TAG.

With these results, the missing data results are similar to the results with no missing data for four of the six haplotypes. Haplotypes CAA and TAG do have differences of at least 1% in haplotype frequency when they are examined with three and five percent missing data compared to the situation where no missing data exists. As the percentage of missing data increases, the 95% CI only increases slightly in length. This suggests the BJLM is stable for 3 SNP haplotypes when missing data is 5% or less.

4.3.4. Haplotype (12 SNP) Simulation Results

The 12 SNP haplotype simulation results are presented in Tables 4.4 and Figures 4.8. Results are shown for the full set of haplotypes with frequency $> 5\%$.

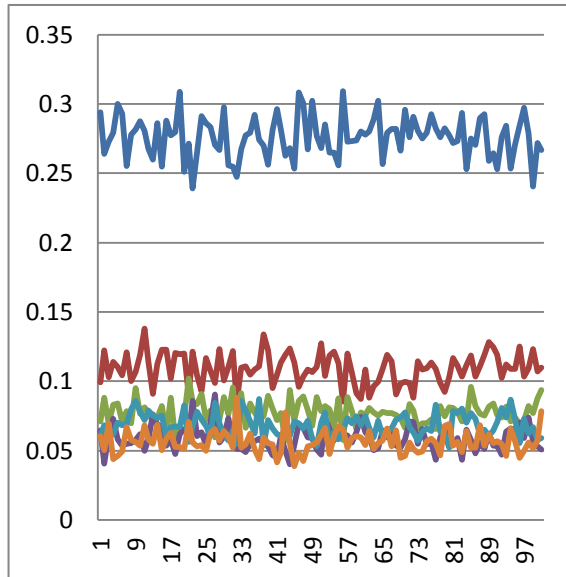
Table 4.4: Summary of Simulated haplotypes of 12 SNP length.

The haplotype index number is listed in the far left column, with the associated haplotype in the next column to the right. The haplotype frequency assuming no missing data is listed in the 3rd column from the left, and the average frequency for each haplotype from the 100 replicates are listed in the column next to the no missing data column. Finally, the 95% credible interval (CI) for these frequencies from the 100 replicates is listed in the final two columns.

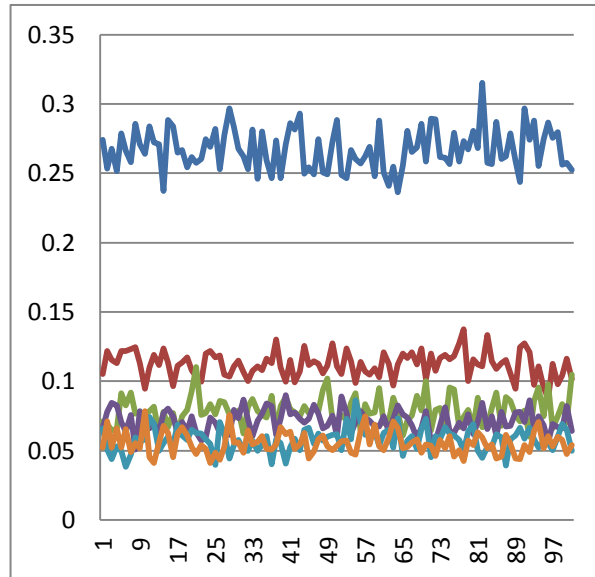
Haplotype #	Haplotype	No Missing Data	Median Frequency	2.5%	97.5%
1% Missing SNP Data					
1	TGACCTCCATGC	0.271	0.276	0.244	0.305
14	TGGTTTCTGTGC	0.117	0.109	0.086	0.128
33	CAACCCCCATGC	0.069	0.078	0.059	0.096
46	CAGTTCCTGTGC	0.067	0.068	0.050	0.087
9	TGGCCTCCATGC	0.057	0.059	0.042	0.075
17	TAACCCCCATGC	0.056	0.057	0.042	0.074
2 % Missing SNP Data					
1	TGACCTCCATGC	0.271	0.267	0.239	0.295
14	TGGTTTCTGTGC	0.117	0.112	0.095	0.129
33	CAACCCCCATGC	0.069	0.079	0.061	0.101
46	CAGTTCCTGTGC	0.067	0.071	0.056	0.087
9	TGGCCTCCATGC	0.057	0.058	0.040	0.073
17	TAACCCCCATGC	0.056	0.056	0.042	0.073
3% Missing SNP Data					
1	TGACCTCCATGC	0.271	0.267	0.239	0.293
14	TGGTTTCTGTGC	0.117	0.115	0.091	0.136
33	CAACCCCCATGC	0.069	0.078	0.059	0.096
46	CAGTTCCTGTGC	0.067	0.069	0.050	0.087
17	TAACCCCCATGC	0.057	0.059	0.041	0.072
9	TGGCCTCCATGC	0.056	0.056	0.043	0.074
4% Missing SNP Data					
1	TGACCTCCATGC	0.271	0.274	0.250	0.301
14	TGGTTTCTGTGC	0.117	0.114	0.091	0.140
33	CAACCCCCATGC	0.069	0.080	0.065	0.100
46	CAGTTCCTGTGC	0.067	0.067	0.051	0.091
9	TGGCCTCCATGC	0.057	0.054	0.038	0.067
17	TAACCCCCATGC	0.056	0.051	0.033	0.067
5% Missing SNP Data					
1	TGACCTCCATGC	0.271	0.271	0.247	0.299
14	TGGTTTCTGTGC	0.117	0.116	0.094	0.139
33	CAACCCCCATGC	0.069	0.082	0.064	0.101
46	CAGTTCCTGTGC	0.067	0.062	0.045	0.078
9	TGGCCTCCATGC	0.057	0.055	0.039	0.071
17	TAACCCCCATGC	0.056	0.053	0.039	0.074

Figure 4.8: Graphical representations of the haplotype frequency for simulations of 12 SNP
Haplotypes

A) 1% Missing Data



B) 2% Missing Data



C) 3% Missing Data

D) 4% Missing Data

E) 5% Missing Data

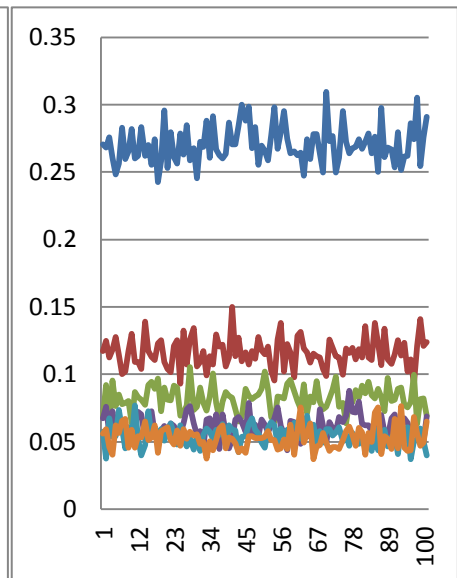
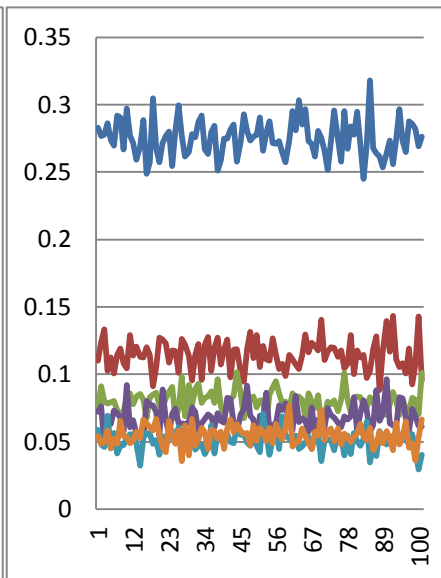
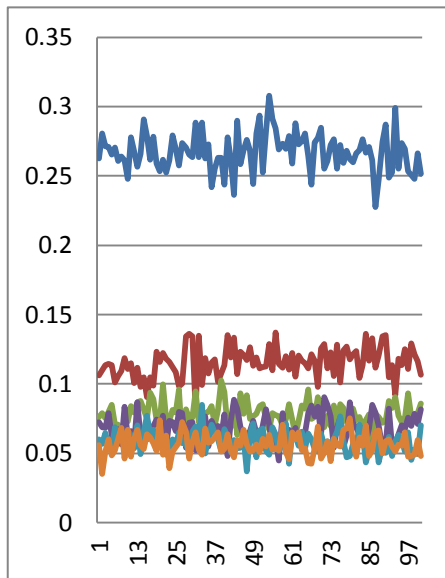


Figure 4.8: The X axis is the simulation run number, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = Haplotype TGACCTCCATGC, Red Line = Haplotype TGGTTTCTGTGC, Green Line = Haplotype CAACCCCATGC, Purple Line = Haplotype TGGCCTCCATGC, Light Blue Line = Haplotype CAGTTCCTGTGC, and the Orange Line = Haplotype TAACCCCATGC.

With these results, the missing data results are similar to the results with no missing data for five of the six haplotypes. Haplotype CAACCCCATGC did have differences between% 0.9 and 1.3% when missing data existed compared to no missing data. As the percentage of missing data increases, the 95% CI only increases slightly in length. This suggests the BJLM is stable for 12 SNP haplotypes when missing data is 5% or less.

4.4. BJLM Creation Conclusion and Application to Real-World Data Sets

With the BJLM, a program which can simultaneously infer haplotypes with or without missing data, I can process HapMap information into many other formats used for haplotype analysis, and can setup haplotypes for risk model analysis has been created. The modified forward-backward algorithm allows for stable estimation of haplotypes from 2 to 12 SNPs, with missing data up to 5%. Therefore, assuming that the call rate of all SNPs used to infer haplotypes in the analysis are at least 95%, stable estimation of haplotype from 2 to 12 SNPs occur with the BJLM.

In the next three chapters, the BJLM will be applied to three separate data sets. First, the BJLM will be tested in a Hodgkin data set with 200 cases and 220 controls that include 62 SNPs derived from a candidate gene analysis¹³³. A risk model including potential haplotypes from these 62 SNPs and covariate information including sex, age, and smoking status will be created using the BJLM for the Caucasian population in the Hodgkin dataset (n = 358, 163 cases and 195 controls). Second, an updated Spitz lung cancer risk model will be created that will include non-genetic risk variables from the original Spitz lung cancer risk model plus inferred haplotypes that include the top SNPs listed in a meta-analysis of 10 lung cancer GWAS¹⁶⁵. Finally, for the first time, a Glioma risk model will be created that include the top haplotypes from an inflammation

pathway analysis plus top SNPs from the inflammation pathway that are not in LD with any of the top haplotypes in the inflammation pathway.

Chapter 5: Application One of the BJLM: Using Haplotype Analysis to Elucidate Significant Associations between Genes and Hodgkin Disease

With increasing number of genes available for study through high throughput platforms, more opportunities are available to discover key associations between genetic factors and disease. Candidate gene studies have successfully determined genetic associations for risk in diseases and possible therapeutic targets, thus highlighting the continuing importance of candidate gene analysis. In this study, the association between the inferred haplotypes in the inflammation (*IL1B*, *IL4*, *IL4R*, *IL10*), DNA repair (*MGMT*, *XPC*) and folate (*MTHFR*) pathways and overall risk of Hodgkin lymphoma were estimated. Genetic and epidemiological data was obtained from a Hodgkin lymphoma cohort conducted at The University of Texas MD Anderson Cancer Center between 1987 and 1992 consisting of 200 cases and 220 controls matched by age, gender, and race/ethnicity.

5.1. Introduction to Hodgkin Disease

In 2010, approximately 8,490 individuals were diagnosed in the United States (US) with Hodgkin lymphoma (HL) with about 1320 deaths¹⁶⁶. Incidence of HL increases to 4 cases per 100,000 in individuals between the ages of 20 and 30, decreases slightly through ages 40 and 60, and then increases again to 4 cases per 100,000 individuals between the ages of 70 and 80^{167,168}. Evidence suggests that genetic factors could be responsible for both the peak in incidence for young adults and also the peak in older adults^{133,169,170}.

Epidemiological studies have shown associations between genetic polymorphisms in genes and cancer risk¹⁷¹⁻¹⁷⁴, and hundreds of single nucleotide polymorphisms (SNPs) in multiple pathways have been identified. Several studies have suggested that many of these polymorphisms result in amino acid substitutions which may alter wild-type protein function and

lead to substantial changes in protein levels, thereby contributing to cancer susceptibility. This is especially true with modifications in the DNA repair, folate, and inflammation pathways^{49,50,175-183}. However, the single-SNP based strategy, in which each SNP is analyzed individually, may be insufficient to explain the risk of developing cancer, as complex diseases most likely result from genetic variants in multiple genes in different pathways. Recently, in accordance with this hypothesis, several studies have shown that epistatic interactions of multiple SNPs are necessary to contribute to cancer susceptibility¹⁸⁴⁻¹⁸⁷.

Haplotype analysis is the study of a pattern of descent of a set of linked alleles occurring on the same chromosome. Several programs that infer haplotypes are Haploview⁵⁶, PHASE^{55,61}, and Haplo.stats⁶². Haplo.stats also uses a joint-effects logistic model to estimate an odds ratio for the association between either a haplotype or a covariate with disease. In a joint-effects model, each haplotype is compared to the most frequent haplotype under the assumption that the most frequent haplotype represents the “normal” haplotype status⁶¹. In two previous studies, it has been shown using logistic regression analysis that variants among SNPs and combinations of haplotypes containing two SNPs can modulate HL risk^{133,188,189}. For this study, it is hypothesized that gene-gene interactions between candidate SNPs in the DNA repair (*MGMT*, *XPC*), folate pathways (*MTHFR*), and inflammation (*IL1B*, *IL4*, *IL4R*, *IL10*) may contribute to HL susceptibility. To test this hypothesis, haplotype analysis was used to discover sets of linked SNPs that could potentially elucidate the association of multiple polymorphisms in these three pathways and the risk of HL. Also, the BJLM from **Chapter 4** was incorporated to develop an experimental risk model from the non-Hispanic Caucasian section of the Hodgkin dataset.

5.2. Materials and Methods in Hodgkin Disease Study

5.2.1. Study Population for Hodgkin Disease Study

Details of the study population were published elsewhere¹³³. Briefly, the study included 200 histological confirmed adult cases registered at the University of Texas MD Anderson Cancer Center and 220 frequency matched controls with respect to age- (± 5 years), sex-, and race/ethnicity. Interviews were conducted by trained interviewers and demographic and clinical characteristics included: age at diagnosis, sex, race/ethnicity, family history of cancer, HL histological subtypes, disease stage and presence of B symptoms. Demographic analysis showed that controls were on average two years older than cases, and the majority of the participants were male (54.8%), and the vast majority of participants (85.2%) were non-Hispanic whites¹³³. For non-Hispanic whites, 163 out of the 358 individuals were cases (45.5%), while for those who were not non-Hispanic whites, 37 out of the 62 individuals were cases (59.7%). Fifty two percent (52%) of the cases (N=103) were diagnosed with stage II HL and 14% of the cases (N=27) were diagnosed with Stage IV HL¹³³. The institutional review board at the University of Texas MD Anderson Cancer Center approved the study.

5.2.2. SNP Selection for Hodgkin Disease Study

Sixty-two SNPs in DNA repair, inflammation, and folate pathway genes were genotyped using Taq-Man-based methods as detailed elsewhere¹³³. The location (chromosome and base-pair information), and the reference SNP (rs) number for each SNP in this study is listed in Table 16. For quality control and to ensure proper genotyping of samples, 5% of all samples were selected randomly for repeat analysis.

Table 5.1: Location and rs number for each Single Nucleotide Polymorphism.

This table contains the genetic information of all 62 SNPs in this study

Gene	Rs number	Chm.	Base pair location	Gene	Rs number	Chm.	Base pair location
<i>MTHFR</i>	rs1801131	1	11788742	<i>IL6</i>	rs1800796	7	22539486
<i>MTHFR</i>	rs1801133	1	11790644	<i>IL6</i>	rs1800795	7	22539885
<i>COX2</i>	rs20417	1	183381978	<i>PolB</i>	rs3136794	8	42345711
<i>IL10</i>	rs1800872	1	203334802	<i>NBN</i>	rs1805794	8	91059655
<i>IL10</i>	rs1800871	1	203335029	<i>TLR4</i>	rs2737191	9	117542269
<i>IL10</i>	rs1800896	1	203335292	<i>TLR4</i>	rs12377632	9	117552284
<i>PARP1</i>	rs1136410	1	222862037	<i>TLR4</i>	rs1554973	9	117560366
<i>MTR</i>	rs1805087	1	233374541	<i>MGMT</i>	rs12917	10	131396273
<i>IL1B</i>	rs1143634	2	113306621	<i>MGMT</i>	rs2308321	10	131455054
<i>IL1B</i>	rs1143627	2	113310618	<i>MGMT</i>	rs2308327	10	131455160
<i>IL1B</i>	rs16944	2	113311098	<i>MS4A2</i>	rs535630	11	59618108
<i>OGG1</i>	rs1052133	3	9773773	<i>MS4A2</i>	rs569108	11	59619680
<i>PPARG</i>	rs1801282	3	12368125	<i>IL18</i>	rs187238	11	111540198
<i>XPC</i>	rs2228001	3	14162450	<i>XPG</i>	rs17655	13	102326003
<i>XPC</i>	rs2228000	3	14174889	<i>APEX1</i>	rs3136819	14	19994929
<i>IL8</i>	rs4073	4	74971059	<i>XRCC1</i>	rs861539	14	103235506
<i>NFKB1</i>	rs1020759	4	103867704	<i>IL4R</i>	rs1805011	16	27281373
<i>IL2</i>	rs2069762	4	123735585	<i>IL4R</i>	rs1805012	16	27281465
<i>TLR3</i>	rs3775291	4	187379223	<i>IL4R</i>	rs1805015	16	27281691
<i>MTRR</i>	rs1801394	5	7923973	<i>IL4R</i>	rs1801275	16	27281901
<i>CCNH</i>	rs2266690	5	86731030	<i>IL4R</i>	rs1805016	16	27282428
<i>IL13</i>	rs1800925	5	132020708	<i>XRCC3</i>	rs25487	19	48747566
<i>IL13</i>	rs20541	5	132023863	<i>XRCC1</i>	rs1799782	19	48749414
<i>IL4</i>	rs2243250	5	132037053	<i>XPB</i>	rs13181	19	50546759
<i>IL4</i>	rs2070874	5	132037609	<i>XPB</i>	rs238406	19	50560149
<i>TNF</i>	rs1799964	6	31650287	<i>ERCC1</i>	rs3212986	19	50604576
<i>TNF</i>	rs1800630	6	31650455	<i>LIG1</i>	rs20580	19	53346365
<i>TNF</i>	rs1799724	6	31650461	<i>LIG1</i>	rs20579	19	53360642
<i>TNF</i>	rs1800629	6	31651010	<i>TLR</i>	rs179008	23	12663316
<i>TNF</i>	rs361525	6	31651080	<i>TLR8</i>	rs5744077	23	12696844
<i>IL6</i>	rs1800797	7	22539461	<i>TLR4</i>	rs2179356	23	141783842

5.2.3. Haplotype Analysis for Hodgkin Disease Study

Inferred haplotype construction was completed using Haploview⁵⁶. Gold plots were constructed based on the SNPs for each gene and those SNPs with strong linkage disequilibrium

(LD), as determined by the confidence interval method. This method states that a block of SNPs are significant if 95% of all informative comparisons of the SNPs in questions show strong LD⁵³. The PHASE package was used to compare frequencies of inferred haplotypes between cases and controls with significance determined by the method of Li and Stephens¹⁹⁰.

To determine whether the haplotypes inferred from each linked gene showed significant association for HL, these haplotypes were modeled accounting for age, sex, race/ethnicity, and smoking status with the Haplo.stats program incorporated into the software environment R^{62,63}. The haplotype analyses were based on the regression-based method for binary (case-control) responses as described by Schaid et al. 2002⁶² and Lake et al. 2003⁶³. Haplotypes larger than a frequency of 0.05 were included in the analyses. For each chromosome and linked gene, the joint-effects model was developed in which each haplotype is compared to the most frequent haplotype (which was used as the reference group)^{62,63}. Analyses were conducted on these linked genes assuming both additive and genetic effects.

5.2.4. Developing an Experimental Risk Model with the BJLM

As a parallel analysis to the haplotype analysis conducted in section 5.2.3, haplotype models were developed to determine the effectiveness of combining the significant haplotypes from section 5.2.3 into multivariate logistic models. Haplotypes were inferred using both the BJLM and haplo.stats, and the effectiveness of these genetic models created with haplotypes from these two programs will be compared to each other. For haplo.stats, the haplotype results are collected from the program, and a risk model is created using multivariate logistic regression in SPSS.

For analysis with the BJLM, the prior information was obtained by determining the counts of all possible haplotypes inferred with HapMap data¹³⁷⁻¹⁴⁰. These counts were then

incorporated into the inferring of haplotypes from the Hodgkin disease dataset. Two hundred burn-in runs of the BJLM were conducted to remove haplotypes with no counts in the Hodgkin disease dataset. Then, 900 runs of the BJLM were incorporated to extract the relevant haplotype information for each individual in the study. Haplotype frequencies for each haplotype block, except for the most frequent haplotype, was saved and then presented to WINBUGS (version 1.4.3) for further analysis. Sample code for WINBUGS is shown below in Figure 5.1:

Figure 5.1: Sample Code for WINBUGS for conducting Haplotype Analysis

```
model
{
  for( i in 1 : N )
  {
    casecntl[i] ~ dbin(p[i],1)
    logit(p[i]) <- alpha +beta1*Haplo2[i]
  }
  # Priors for logit model
  alpha ~ dnorm(0.0,1.0E-2)
  beta1 ~ dnorm(0.0,1.0E-2)
  # Determine p-value in Bayesian Context
  beta1pabove <- step(beta1-0)
  beta1pbelow <- 1-step(beta1-0)
}
```

Figure 5.1: With this code, the case/control status was modeled as a binomial variable with size one, and the logit function contained all haplotypes that were being compared to the most frequent haplotype. For the prior information in dealing with the effects of each haplotypes, all alphas and betas were modeled as normal distribution with mean zero and variance 100, hence leading to non-informative priors. Also, the frequencies of beta being above and below zero are calculated with the step functions, and this leads to an empirical p-value which is two times the minimum frequency among beta being above or below zero.

This analysis is conducted for all three genetic models for each haplotype block, and haplotype results are collected after 50,000 iterations of the WINBUGS code (1st 5000 iterations

are burn-in, and are not counted in the analysis). All haplotypes with empirical p-values less than 0.05 are presented for future analysis.

When examining the genetic model for each haplotype whether using the Haplo.stats or the BJLM, the most significant haplotype out of the additive and dominant genetic models are selected for multivariate logistic regression analysis. For haplotypes constructed in Haplo.stats, multivariate logistic regression analysis is conducted with SPSS, and for haplotypes constructed in the BJLM, WINBUGS is then used again with the same construction as in Figure 6.2. Finally, the discriminatory power for both haplotype models are calculated in SPSS using the relative risk profiles calculated by the multivariate logistic risk models developed in either SPSS (frequentist) using haplotypes inferred from haplo.stats or the haplotypes inferred by the BJLM (Bayesian).

To remove possible confounding by race, all non-Caucasians were removed from the analysis, and also to examine more rare haplotypes, all haplotypes were examined that had frequencies of at 1%. This led to an analysis with 358 individuals, with 163 cases and 195 controls, and models will be developed from SNPs from both significant SNPs and significant haplotypes, assuming additive and dominant genetic effects. These haplotypes will also be modeled in the presence of the three matching risk variables in this study: Age, Sex, and Smoking Status. Due to the small sample size, recessive genetic effects will not be modeled.

5.3. Hodgkin Disease Results

5.3.1. Haplotypes and Haplotype Frequencies for Hodgkin Cases and Controls

Comparisons of haplotype frequencies are summarized in Table 5.2.

Table 5.2: Estimated frequencies for haplotypes on chromosomes 1-3, 5-7, 10, 16, and 19.

Gene (Chr)	SNPs	bp	Haplo.	Haplotype Frequency: Cases (%)	Haplotype Frequency: Controls (%)	p-value
<i>MTHFR</i> (1)	rs1801131	11788742	AC	43.38	37.05	0.09
	rs1801133	11790644	AT	30.80	32.63	
			CC	25.66	29.65	
<i>IL10</i> (1)	rs1800872	203334802	ACA	31.38	30.46	0.38
	rs1800871	203335029	ACG	41.75	44.09	
	rs1800896	203335292	CTA	26.37	25.45	
<i>IL1B</i> (2)	rs1143627	113310618	CT	36.16	36.12	0.99
	rs16944	113311098	TC	62.79	62.88	
<i>XPC</i> (3)	rs2228001	14162450	CC	34.11	37.20	0.08
	rs2228000	14174889	CT	27.38	20.76	
			AC	38.49	42.04	
<i>IL4</i> (5)	rs2243250	132037053	CC	82.50	79.08	0.45
	rs2070874	132037609	TT	14.50	16.82	
<i>MGMT</i> (10)	rs2308321	131455054	AA	89.50	90.45	0.61
	rs2308327	131455160	GG	10.50	9.55	
<i>IL4R</i> (16)	rs1805012	27281465	TT	85.85	82.02	0.21
	rs1805015	27281681	TC	5.82	8.89	
			CC	8.33	9.09	
<i>IL4R</i> (16)	rs1801275	27281901	AT	14.53	17.82	0.09
	rs1805016	27282428	AG	5.56	8.18	
			GT	79.90	74.00	

Chr = Chromosome

bp = base pair location

Haplo = Haplotypes

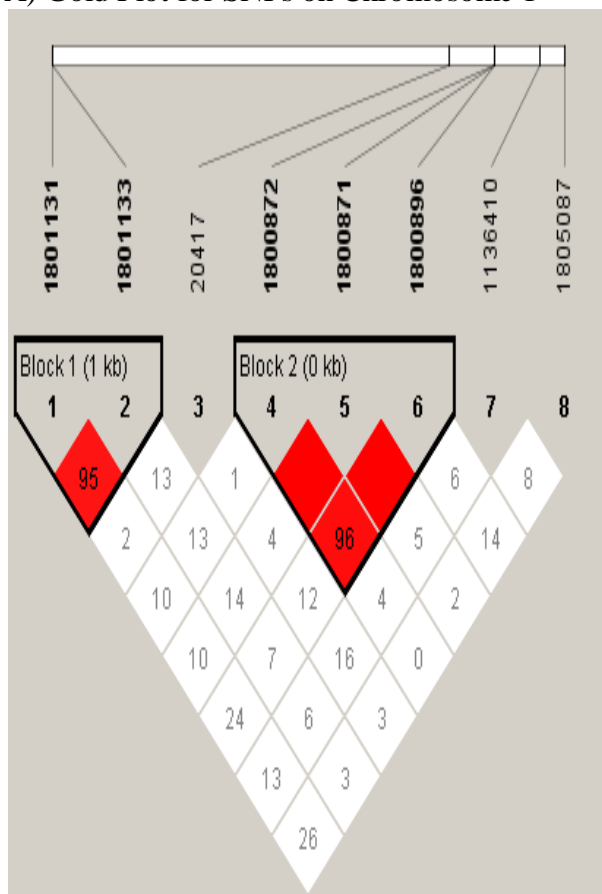
Haplotypes with frequencies greater than 0.05 are shown.

Only three inferred haplotypes resulted in significance at the 10% level, namely, the *MTHFR* haplotype (rs1801131 and rs1801133) on chromosome 1 ($p = 0.09$), the *XPC* haplotype (rs2228001 and rs2228000) on chromosome 3 ($p=0.08$), and a haplotype with the *IL4R* gene (rs1801275 and rs1805016) on chromosome 16 ($p=0.09$). On chromosome 1, the largest differences in haplotype frequency between cases and controls were observed with haplotype AC; on chromosome 3, the largest differences in frequency were observed with haplotype CT;

and haplotype GT on chromosome 16. Figure 5.2 shows the haplotypes on chromosomes 1 and 16, as determined by the gold plot.

Figure 5.2: Chromosome 1 and Chromosome 16 Gold Plots from Hodgkin Study

A) Gold Plot for SNPs on Chromosome 1



B) Gold Plot for SNPs on Chromosome 16

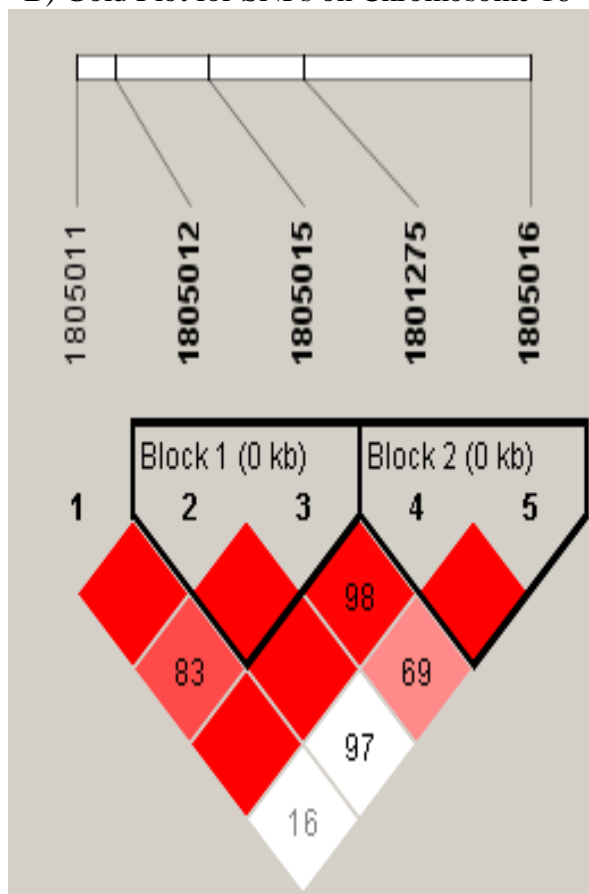


Figure 5.2: Gold Plots constructed from Single Nucleotide Polymorphisms on Chromosome 1 (1a) and Chromosome 16 (1b). The rs numbers for each SNP are listed at the top of each figure, and those SNPs bolded are the SNPs which forms a haplotype. Haplotypes are constructed with the Gabriel method, and are shown as blocks on the gold plot with the length of these haplotypes in kilobases. Diamond values represent the D' for the 2 SNPs being examined. For example, the D' value for SNPs rs1801131 and rs1801133 is 0.95.

For example, using the SNPs on chromosome 1 that were genotyped in association with this HL study, the first two SNPs, and the 4th through 6th SNPs are linked together in separate

blocks. For the *XPD* gene, SNPs were not linked strongly enough according to the Gabriel confidence interval method⁵³ to conduct haplotype analysis.

5.3.2. Determining Associations between Haplotypes and Hodgkin Disease

Associations between inferred haplotypes for each set of linked SNPs and HL are summarized in Table 5.3 assuming both additive and dominant genetic effects.

Table 5.3: Determining associations between haplotypes and HL using the joint effect logistic model controlling for age, sex, race, and smoking status

		Genetic Effects			
		Additive		Dominant	
Gene (Chromosome)	Haplotype	OR (95% CI)	p-value	OR (95% CI)	p-value
<i>MTHFR</i> (1)	AC	1.00	NA	1.00	NA
	AT	0.87(0.64-1.20)	0.414	0.78(0.51-1.20)	0.226
	CC	0.73(0.52-1.04)	0.079	0.80(0.53-1.22)	0.300
<i>IL10</i> (1)	ACA	1.09(0.77-1.54)	0.617	1.20(0.79-1.82)	0.385
	ACG	1.00	NA	1.00	NA
	CTA	1.08(0.76-1.52)	0.675	1.10(0.73-1.66)	0.643
<i>IL1B</i> (2)	CT	1.00 (0.76-1.32)	0.992	0.97(0.65-1.45)	0.879
	TC	1.00	NA	1.00	NA
<i>XPC</i> (3)	CC	0.98(0.72-1.32)	0.873	0.91(0.60-1.38)	0.671
	CT	1.49(1.03-2.16)	0.037	1.75(1.15-2.65)	0.009
	AC	1.00	NA	1.00	NA
<i>IL4</i> (5)	CC	1.00	NA	1.00	NA
	TT	0.80(0.55-1.16)	0.244	0.83(0.54-1.30)	0.422
<i>MGMT</i> (10)	AA	1.00	NA	1.00	NA
	GG	1.19(0.75-1.89)	0.448	1.14(0.69-1.87)	0.607
<i>IL4R</i> (16)	TT	1.00	NA	1.00	NA
	TC	0.63(0.34-1.00)	0.052	0.48(0.26-0.88)	0.017
	CC	0.93(0.57-1.51)	0.757	0.88(0.52-1.50)	0.647
<i>IL4R</i> (16)	AT	0.72(0.50-1.05)	0.085	0.70(0.46-1.09)	0.117
	AG	0.57(0.33-0.99)	0.045	0.58(0.32-1.06)	0.076
	GT	1.00	NA	1.00	NA

NA = Not Applicable

Significant associations ($p < 0.05$) were observed on chromosomes 3 and 16, and borderline significant associations ($p < 0.10$) were observed on chromosomes 1 and 16. A significantly increased risk of HL was observed with haplotype CT, inferred from SNPs rs2228001 and rs2228000 on the XPC gene, assuming either dominant genetic effects (OR = 1.75, 95% CI = 1.15-2.65, p -value = 0.009) or additive genetic effects (OR = 1.49, 95% CI = 1.03-2.16, p -value = 0.037). A significant protective association between inferred haplotypes and HL was calculated for chromosome 16 with haplotype TC, inferred from SNPs rs1805012 and rs1805015 on the IL4R gene, assuming dominant genetic effects (OR = 0.48, 95% CI = 0.26-0.88, p -value = 0.017). However with this same haplotype, the protective association was borderline significant when assuming additive genetic effects (OR = 0.63, 95% CI = 0.34-1.00, p -value = 0.052). Another protective association was calculated with haplotype AG from the IL4R gene, inferred from SNPs rs1801275 and rs1805016, assuming additive genetic effects (OR = 0.57, 95% CI = 0.33-0.99, p -value = 0.045) but the protective association was only borderline significant when assuming dominant genetic effects (OR = 0.58, 95% CI = 0.32-1.06, p -value = 0.076). Another borderline protective haplotype was calculated with haplotype CC from the MTHFR gene, inferred from SNPs rs1801131 and rs1801133, assuming additive genetic effects (OR = 0.73, 95% CI = 0.52-1.04) and no association was found assuming dominant genetic effects.

5.3.3. Incorporation of the BJLM for Hodgkin Disease Study

Both haplo.stats and the BJLM resulted in four haplotypes that showed some significance ($p < 0.20$) with Hodgkin disease in a multivariable logistic model assuming additive genetic effects, and these results are listed in Table 5.4.

Table 5.4: Haplotype model results using haplotypes inferred with both haplo.stats and BJLM assuming an additive genetic model

For the analysis with the BJLM, an empirical p-value is estimated using the WINBUGS code from Figure 5.1

		Genetic Model Program			
		Haplo.stats		BJLM	
Gene (Chromosome)	Haplotype	OR (95% Confidence Interval)	p-value	OR (95% Credible Interval)	p-value
<i>XPC</i> (3)	CT	1.33 (0.92-1.94)	0.134	1.41 (0.93-2.15)	0.107
<i>IL4</i> (5)	TC	10.8 (1.98-58.3)	0.006	17.4 (3.18-145.2)	0.001
<i>IL4R</i> (16)	TT	1.93 (0.86-4.34)	0.111	2.19 (0.93-5.36)	0.073
<i>IL4R</i> (16)	AT	0.50 (0.27-0.95)	0.035	0.43 (0.20-0.88)	0.019

Using these four haplotypes inferred from the BJLM, results for CT haplotype from the *XPC* gene, the TC haplotype from the *IL4* gene, the CC haplotype on the 1st haplotype block from the *IL4R* gene, and the AT haplotype from the 2nd haplotype block from the *IL4R* gene all had odds ratios that were further away from one, and also had more significant p-values. However, when comparing these models against each other, the genetic model with haplotypes inferred from haplo.stats (AUC = 0.632, 95% CI = 0.567-0.690) and the genetic model with haplotypes inferred from the BJLM (AUC = 0.634, 95% CI = 0.569-0.691) are not significantly different in discriminatory power (p-value = 0.1562).

Both haplo.stats and the BJLM resulted in four haplotypes that showed some significance ($p < 0.20$) with Hodgkin disease in a multivariable logistic model assuming dominant genetic effects, and these results are listed in Table 5.5.

Table 5.5: Estimation of Odds Ratios for haplotypes using both haplo.stats and the BJLM

For the analysis with the BJLM, an empirical p-value is estimated using the WINBUGS code from Figure 5.1

Gene (Chr)	SNPs	Haplo.	Haplo.stats OR (95% Confidence Interval)	BJLM OR (95% Credible Interval)
XPC (3)	rs2228001 rs2228000	CT	1.51 (0.97-2.36) p-value = 0.069	1.50 (0.95-2.33) p-value = 0.061
IL4 (5)	rs2243250 rs2070874	TC	9.24 (1.81-47.2) p-value = 0.008	12.16 (2.47-90.1) p-value = 0.00044
IL4R (16)	rs1805012 rs1805015	CC	1.94 (0.79-4.76) p-value = 0.149	2.05 (0.84-5.18) p-value = 0.118
IL4R (16)	rs1801275 rs8105016	AT	0.46 (0.21-0.98) p-value = 0.044	0.43 (0.19-0.90) p-value = 0.028

Just like with the additive genetic effects, the BJLM showed a much stronger association for the TC haplotype on the IL4 gene, although with the dominant effects, the association was almost identical for the CT haplotype on the XPC gene when either haplo.stats or the BJLM are used to infer haplotypes. When examining the discriminatory power of genetic models inferred using dominant genetic effects, the genetic model with haplotypes inferred from haplo.stats (AUC = 0.643, 95% CI = 0.580-0.699) and the genetic model with haplotypes inferred from the BJLM (AUC = 0.644, 95% CI = 0.581-0.699) did not have significantly different results (p-value = 0.3477).

5.4. Hodgkin Disease Study Discussion

The purpose of this study was to use haplotype analysis to elucidate associations in HL that could not be obtained solely with single SNP based strategies. Although differences in frequency between case and control haplotypes were only borderline significant in the overall HL set (i.e., *MTHFR* gene on chromosome 1, *XPC* gene on chromosome 3, and *IL4R* gene on chromosome 16), there were some haplotypes which showed either susceptibility to HL

development or protection from HL development. Two sets of linked SNPs on IL4R: rs1805012 and rs1805015, as well as rs1801275 and rs1805016, showed significant protective associations. A set of linked SNPs on *XPC* from chromosome 3 (rs2228001 and rs2228000) showed strong associations with HL. When attempting to develop risk models for the study of Hodgkin disease, the BJLM with the Hodgkin data did estimate stronger associations with more significant p-values compared to the haplo.stats inferred haplotypes for individual haplotypes. But, when the haplotypes were combined to form full genetic models, models with haplotypes inferred with the BJLM did not perform significantly better than models with haplotypes inferred with haplo.stats even though the discriminatory power was slightly higher with BJLM inferred haplotype models.

In previous studies, variant alleles on the *MTHFR* gene have been shown to be protective for non HL risk⁴⁹ and relapse cancer events when both variant alleles are present for MTHFR 677 C>T SNP (rs1801133) and the MTHFR 1298 A>C SNP (rs1801131)¹⁹¹. These variant alleles result in decreased *MTHFR* gene activity, which leads to rises in methyl donors that limit possible breaks in cell mitosis, and possible cancer proliferation¹⁹¹. A variation of genes that catalyze the conversion of 5-10-methyl-tetrahydrofolate (THF) into 5-methyl-THF, the more commonly circulated form of folate throughout the body, was shown to affect the risk of non-HL development¹⁴⁴. Further the alteration of the Methylenetetrahydrofolate reductase (*MTHFR*) gene has been shown to decrease in the enzyme needed to catalyze the conversion into the more common form of folate by 30 to 60 percent^{192,193}. In this study, the decrease in risk of HL was only borderline significant in the presence of haplotypes inferred from these two SNPs for both differences in haplotype frequency while assuming an additive genetic mode, but when assuming a dominant genetic model, the effect of MTHFR is minimal. For non-Hodgkin lymphoma, the effect of folate has been studied extensively, and the results have been mixed^{194,195}. More studies

will be needed to elucidate the relationship between genes that regulate folate, and Hodgkin lymphoma.

SNPs in DNA repair genes, such as *XPC* and *XPB* genes, have been linked to many cancers^{133,196,197}. DNA damage and subsequent repair are critical for maintaining genomic integrity and stability. Modifications with SNPs in DNA repair genes may modulate the DNA repair phenotype, particularly when these SNPs are located within coding or regulating regions, leading to alterations in protein expression and in functional properties of repair enzymes^{176,177}. In our study, study subjects with both variant alleles for the *XPC* haplotype had an increased risk for HL with the joint effect haplotype model. Interestingly enough, haplotypes with only the 1st *XPC* SNP as a variant allele showed little association with HL, but the combination of the 2nd variant allele with the 2nd *XPC* SNP caused an increase in the cancer risk. This finding also highlights the importance of haplotype analysis. If only the SNP rs2228001 would have been tested, the association between DNA damage modification on the *XPC* gene and HL would have been missed.

In a hospital-based case control study of 322 lung cancer patients and 326 healthy controls conducted in a Chinese population, haplotypes of *XPC* genes consisting of variant alleles of SNPs rs2228001 (*XPC* 499 A>C) and rs2228000 (*XPC* 939 C>T) showed an increased risk associated with lung cancer, especially haplotype CT (OR = 2.37, 95% CI = 1.33-4.21)¹⁹⁸. Increased susceptibility of cancer in the presence of variant alleles associated with the *XPC* gene are due to decreased DNA repair capacity, possibly resulting from a transfer in amino acid production from lysine to glutamine¹⁹⁷⁻¹⁹⁹.

In chronic inflammation, the cell death process is stopped, and cell growth is uncontrolled¹⁷⁸. With HL, the inflammatory response is initiated by the presence of the

malignant Hodgkin and Reed Sternberg cells¹⁷⁹ and is characterized by the massive presence of reactive inflammatory cells in response to the malignant Hodgkin and Reed Sternberg cells, thus 99% of the tumor mass consists of the inflammatory cells^{182,183}. Inflammation genes have been strongly associated with cancer, including stomach, breast, and HL^{178-181,188}. In this study, haplotypes consisting of variant alleles from SNP rs1801275 and SNP rs1805016 from gene *IL4R* gene show decreased risk of HL. A plausible explanation for these results is that interleukin 4 receptors, IL4R α and IL-13R α 1, are expressed in Hodgkin tumor cells^{200,201}. Therefore the common polymorphisms in promoter regions of genes related to the pro- and anti-inflammatory response may contribute to susceptibility to HL and serve as plausible candidates for further study.

One of the study limitations is the relatively small sample size, with 420 individuals available for analysis. Despite the small sample size, this study demonstrated the use of haplotype analysis to analyze the effects of linked SNPs associated with HL. Larger studies are warranted to further elucidate our results. In conclusion, the interactions between specific SNPs in both the inflammation and the DNA repair pathway have been shown to play an important role in possible HL development. Haplotype analysis may be useful to detect interactions that could not be detected through single SNP analyses, and this will be further studied using the BJLM in the next two chapters.

Chapter 6: Application Two of the BJLM: Elucidating Significant Haplotype Associations between Genes and Lung Cancer

6.1. Introduction to Expanding Spitz Lung Cancer Risk Model with Haplotypes

In Chapter 3, the first attempt at extending the Spitz model with haplotypes was conducted using areas from chromosome 15 (rs8034191 and rs1051730) and chromosome 5 (rs2736100 and rs401681). This model did improve discriminatory power with the risk model substantially ($p < 0.001$) in a subset of 2253 individuals (1118 cases/1135 controls) from 0.659 (95% CI = 0.636-0.681) in the original Spitz model to 0.675 (95% CI = 0.653-0.697), and this increase is significant ($p = 0.003$). Also, the new model has much better ability to classify case/control status in all individuals compared to the original Spitz model according to the NRI value, (0.254, p -value = < 0.001). This NRI increase is more prevalent with true cases (NRI = 0.349, p -value = < 0.001), compared to controls (NRI = -0.094, p -value = 0.001). However, when this risk model was compared to a risk model with 3 SNPs (rs2736100, rs401681, and rs1051730) this model and the haplotype model were similar in discriminatory power (0.675 for three SNP vs. 0.676 for haplotype) and the improvement in the NRI compared to the Original Spitz model was superior for the 3 SNP model (NRI = 0.268, p -value = < 0.001) compared to the improvement with the haplotype model. These results strongly suggest that haplotypes need to be selected more carefully to fully extract their potential to improve discriminatory power.

In 2009, the NCI conducted an extensive lung GWAS with 5,739 lung cancer cases and 5848 controls to examine specific SNPs that could be associated with specific histology's of lung cancer^{165,202-205}. They also conducted a meta-analysis from ten additional lung GWAS, which included the Texas lung GWAS⁴⁵, and this added 7561 cases and 13818 controls to the analysis. With a fixed effect model²⁰⁶, the estimates for per-allele odds ratios and their standard error was

determined with the meta-analysis, and the top 200 SNPs with the lowest p-values were summarized in a supplementary table. The SNPs available in the Texas lung GWAS that match the top 200 SNPs from the meta-analysis will be tested to see if these SNPs can form haplotypes from each other, and whether a better Spitz risk model with haplotypes can be constructed.

6.2. Methods to Conduct Expansion of Spitz Model with Top SNPs from Lung meta-analysis

6.2.1. Study Population and Selection of SNPs for Expansion of Spitz Model with Haplotypes

A total of 2291 lung cancer patients and controls from the original Spitz lung cancer study and also the Texas Lung Cancer GWAS were accrued for this study^{27,45}. Lung cancer patients (N=1154) were enrolled from the Thoracic Center at the University of Texas MD Anderson Cancer Center starting in July of 1995 and ending in May 2006²⁷. All lung cancers were histologically confirmed, were newly diagnosed, and had no treatment (chemo or radiation) for lung cancer²⁷. Controls (N=1137) were recruited from the Kelsey-Seybold clinics, and these individuals were lung cancer-free individuals with no prior history of cancer (except for nonmelanoma skin cancer). These controls were frequency matched by age (± 5 years), sex, ethnicity, and smoking status²⁷. The risk factors in the original Spitz study are listed in **Table 2.1 in Chapter 2 on page 27**. Only ever smokers (individuals that have only smoked > 100 cigarettes in their lifetimes) are included in this analysis. All cases (N = 1154) and controls (N = 1137) were genotyped with the Illumina HumanHap300 v.1.1 BeadChips, and 317,498 tagSNPs were obtained for potential analysis⁴⁵. These SNPs were then matched with a list of 200 top SNPs from the lung GWAS meta-analysis paper¹⁶⁵, and 157 top SNPs from the meta-analysis paper exist in the Texas lung GWAS study from MD Anderson.

6.2.2. Selecting the Haplotype Blocks for Future Analysis

To extract the potential haplotype blocks, chromosome and base pair information for all 157 SNPs was extracted from Supplemental Table 2 in the Lung Cancer Meta-Analysis⁴⁵. For these 157 SNPs, all SNPs were loaded into Haploview⁵⁶ using data from the Texas Lung GWAS to determine the potential haplotypes that could be examined for addition into the Spitz model. An example of the potential haplotype blocks from Chromosome 15 to be examined in the extension of the Spitz lung cancer risk model was shown in Figure 6.1.

Figure 6.1: Haploview Gold Plot showing two haplotype blocks from Texas Lung GWAS data that contains SNPs within the top 200 SNP list from the lung cancer GWAS meta-analysis

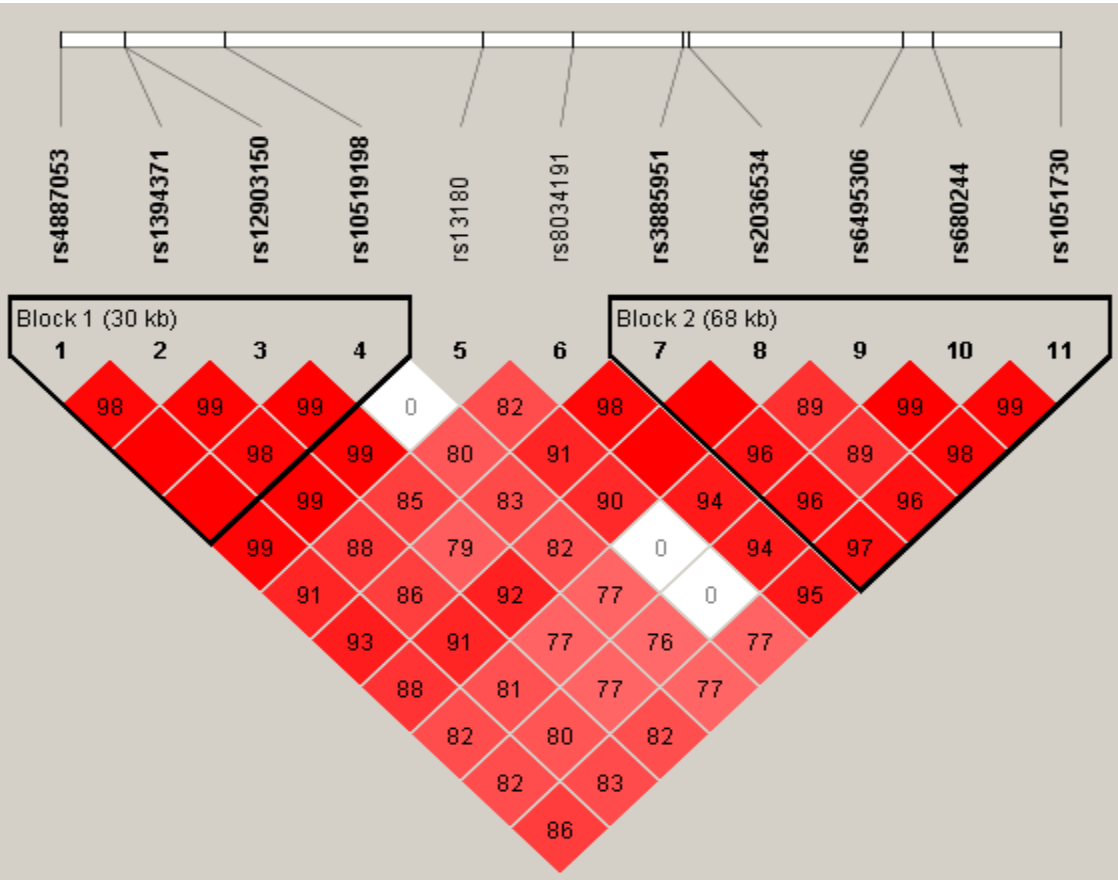


Figure 6.1: This section of SNPs from chromosome 15 encompasses two haplotype blocks that is inferred from 4 SNPs (rs4887053, rs1339471, rs12903150, and rs10519198) and 5 SNPs (rs3885951, rs2036534, rs6495306, rs680244, and rs1051730) respectively, and these haplotype block was formed with the Gabriel’s et.al.⁵³ confidence interval method. These haplotype blocks contain haplotypes that could be used to extend the Spitz lung cancer risk model in former smokers and ever smokers respectively.

From these 157 SNPs, 20 haplotype blocks were formed that contained a combined 55 SNPs, and the information for these blocks and SNPs are listed in Table 6.1.

Table 6.1: Chromosome and Base-Pair Information for all SNPs included in the expansion of the Spitz Model.

These SNPs will form 9 haplotype blocks that will be used to extract potential haplotypes for use in the genetic section of the updated Spitz model

Haplotype Block	Rs Number	Chromosome	Base-pair	Used for Model
1	rs4635969	5	1361552	Former, Current
	rs4975616	5	1368660	
2	rs2256543	6	30045812	Former, Current
	rs2523946	6	30049922	
3	rs2844773	6	30315474	Former, Current
	rs3094073	6	30339203	
	rs3130380	6	30387109	
	rs3130350	6	30435818	
4	rs3132610	6	30652380	Current
	rs9262143	6	30760760	
	rs3094127	6	30805426	
	rs2535319	6	30822458	
5	rs4887053	15	76499754	Former, Current
	rs1394371	15	76511524	
	rs12903150	15	76511700	
	rs10519198	15	76529809	
6	rs3885951	15	76612972	Former, Current
	rs2036534	15	76614003	
	rs6495306	15	76652948	
	rs680244	15	76658343	
	rs1051730	15	76681394	
7	rs6495309	15	76702300	Former, Current
	rs1948	15	76704454	
8	rs12594247	15	76733688	Current
	rs6495314	15	76747584	
	rs8038920	15	76761600	
	rs11638372	15	76770614	
9	rs2277547	15	76869486	Current
	rs3743057	15	76876062	

These 9 haplotype blocks were then further analyzed with both Haplo.stats⁶² and the BJLM (from Chapter 4). Haplo.stats allows for the estimation of odds ratios to represent the association between either a haplotype or a covariate and disease with the use of a joint-effect linear model

developed by the Lake group in which each haplotype is compared to the most frequent haplotype^{62,63}. With the assumption that the most frequent haplotype represents the “normal” haplotype status, one can examine whether variant haplotypes are associated with disease. This analysis is conducted for all three genetic models for each haplotype block. All haplotypes with empirical p-values less than 0.05 and are the most significant assuming additive, dominant, or genetic effects in each haplotype block are presented for future analysis.

For analysis with the BJLM, the same procedure as listed in Chapter 5 is conducted, but briefly, the prior information was obtained by determining the counts of all possible haplotypes inferred with HapMap data¹³⁷⁻¹⁴⁰. These counts were then incorporated into the inferring of haplotypes from the Texas Lung GWAS dataset. Two hundred burn-in runs of the BJLM were conducted to remove haplotypes with no counts in the Texas Lung GWAS dataset. Then, 900 runs of the BJLM were incorporated to extract the relevant haplotype information for each individual in the study. Haplotype frequencies for each haplotype block, except for the most frequent haplotype, was saved and then presented to WINBUGS (version 1.4.3) for further analysis. Sample code for WINBUGS is shown in **Figure 5.1**. This analysis is conducted for all three genetic models for each haplotype block, and haplotype results are collected after 50,000 iterations of the WINBUGS code (1st 5000 iterations are burn-in, and are not counted in the analysis). All haplotypes with empirical p-values less than 0.05 and are the most significant among the three risk models in each haplotype block are presented for future analysis.

When developing updated Spitz lung cancer risk models that incorporate haplotypes, the most significant haplotype that was modeled assuming additive, dominant, or recessive genetic traits were selected for multivariate logistic regression analysis. These haplotypes are added to the original Spitz lung cancer risk variables for both former and current smokers, and those

haplotypes with p-values less than 0.05 are added to the Spitz models. For haplotypes constructed in Haplo.stats, multivariate logistic regression analysis with the original Spitz variables is conducted with SPSS, and for haplotypes constructed in the BJLM, WINBUGS is then used again with the construction listed below (Figure 6.2):

Figure 6.2: Example WinBugs Code for the Development of a Bayesian Logistic Risk

Model Incorporating Haplotypes inferred from the BJLM

```
model
{
  for( i in 1 : N ) {
    casecntl[i] ~ dbin(p[i],1)
    logit(p[i]) <- alpha + (beta[1]*dusts[i]) + (beta[2]*emphys[i])
    + (beta[3]*hayfev[i]) + (beta[4]*newfamcan[i]) + (beta[5]*agequit1[i])
    + (beta[6]*agequit2[i]) + (beta[7]*Hap12Chm5_Dom[i])
  }

  # Priors for logit model
  alpha ~ dnorm(0.0,1.0E-2)
  beta[1] ~ dnorm(0.0,1.0E-2)
  beta[2] ~ dnorm(0.0,1.0E-2)
  beta[3] ~ dnorm(0.0,1.0E-2)
  beta[4] ~ dnorm(0.0,1.0E-2)
  beta[5] ~ dnorm(0.0,1.0E-2)
  beta[6] ~ dnorm(0.0,1.0E-2)
  beta[7] ~ dnorm(0.0,1.0E-2)

  # Determine p-value in Bayesian Context
  for( o in 1:P)
  {
    betapabove[o] <- step(beta[o]-0)
    betapbelow[o] <- 1-step(beta[o]-0)
  }
}

# Insert Data

Inits1 list(alpha = 0, beta = c(0,0,0,0,0,0,0))
```

Figure 6.2: With this code, the original Spitz lung cancer risk variables for former smokers are listed first²⁷, and the significant haplotype for former smokers is listed next. The case/control status was modeled as a binomial variable with size one, and the logit function contained all haplotypes that were being compared to the most frequent haplotype. For the prior information in dealing with the effects of each haplotypes, all alphas and betas were modeled as normal distribution with mean zero and variance 100, hence leading to non-informative priors. Also, the frequencies of beta being above and below zero are calculated, and this leads to an empirical p-value which is two times the minimum frequency among beta being above or below zero.

Finally, the discriminatory power for both haplotype models are calculated in NCSS/PASS using the relative risk profiles calculated by the multivariate logistic risk models developed in either SPSS (frequentist) using haplotypes inferred from haplo.stats or the haplotypes inferred by the BJLM (Bayesian).

6.3. Results for Expanding Spitz Models Using SNPs derived from the Top 200 SNP list in the Lung Cancer Meta-Analysis

6.3.1 Univariate Haplotype Block Analysis to Extend Spitz Model

Joint-effect logistic model analysis is then conducted with the haplotypes blocks listed in Table 6.1, and results assuming an additive, dominant, and recessive genetic trait are generated. Genetic models that lead to the most significant results for each haplotype with both Haplo.stats and the BJLM are displayed in Table 6.2 for former smokers and Table 6.3 for current smokers

Table 6.2: Individual haplotype block analysis using Haplo.stats and the BJLM for former smokers

Haplotype information is listed on the 1st 4 columns on the left, and the results from Haplo.stats and BJLM are listed in the last 4 columns on the right. Significant associations from each haplotype block are highlighted in this table.

Former Smokers							
Chm	Haplo.	Gene	Genetic Model	Haplo.stats		BJLM	
				OR (Confidence Interval) (2.5% to 97.5%)	P-Value	OR (Credible Interval) (2.5% to 97.5%)	Empirical P-Value
5	AG	TERT	Dominant	0.737 (0.579-0.938)	0.0133	0.734 (0.576-0.934)	0.0140
6	AG	HCG4P4	Additive	1.246 (1.054-1.473)	0.0098	1.247 (1.056-1.473)	0.0090
6	AAGC	HLA-L	Additive	1.585 (1.048-2.396)	0.0293	1.559 (1.037-2.361)	0.0322
15	AGAC	IREB2	Additive	0.777 (0.615-0.981)	0.0342	0.774 (0.612-0.981)	0.0337
15	CGGA	IREB2	Additive	0.765 (0.624-0.938)	0.0101	0.763 (0.621-0.935)	0.0092
15	CGAC	IREB2	Recessive	6.56 (0.787-54.66)	0.0857	9.85 (1.25-218.1)	0.0250
15	AAAGA	PSMA4	Additive	1.305 (1.064-1.599)	0.0104	1.307 (1.070-1.602)	0.0094
15	GA	CHRNA 3	Dominant	0.723 (0.573-0.911)	0.0060	0.722 (0.573-0.906)	0.0045

Table 6.3: Individual haplotype block analysis using Haplo.stats and the BJLM for current smokers

Haplotype information is listed on the 1st 4 columns on the left, and the results from Haplo.stats and BJLM are listed in the last 4 columns on the right. Significant associations from each haplotype block are highlighted in this table.

Current Smokers							
				Haplo.stats		BJLM	
Chm	Haplo.	Gene	Genetic Model	OR (Confidence Interval) (2.5% to 97.5%)	P-Value	OR (Credible Interval) (2.5% to 97.5%)	Empirical P-Value
5	AG	TERT	Dominant	0.754 (0.580-0.980)	0.0357	0.753 (0.580-0.979)	0.0344
6	AAGC	HLA-L	Additive	0.518 (0.329-0.815)	0.0046	0.527 (0.333-0.822)	0.0040
6	AGAG	GNL1	Additive	1.269 (1.005-1.603)	0.0458	1.269 (1.005-1.605)	0.0454
6	AGGG	GNL1	Recessive	0.175 (0.020-1.503)	0.1126	0.115 (0.005-0.959)	0.0450
15	AGAC	IREB2	Additive	0.746 (0.574-0.971)	0.0295	0.744 (0.571-0.969)	0.0284
15	CGGA	IREB2	Additive	0.793 (0.632-0.995)	0.0463	0.790 (0.626-0.991)	0.0416
15	AAAGG	PSMA4	Additive	0.444 (0.249-0.792)	0.0061	0.428 (0.236-0.758)	0.0031
15	AGAGG	PSMA4	Dominant	0.726 (0.550-0.959)	0.0224	0.731 (0.553-0.966)	0.0269

With these results, assuming dominant genetic traits leads to the most significant joint-effect logistic model results for five haplotypes, while for additive and recessive genetic traits, leads to the most significant univariate logistic regression results for thirteen and three haplotypes respectively. The most susceptible haplotype for lung cancer development is haplotype CGAC on the IREB gene using either Haplo.stats (OR = 6.56, 95 % CI = 0.787-54.66) or the BJLM (OR = 9.85, 95% CI = 1.25-218.2) in former smokers, and the most significant protective haplotype against lung cancer development is haplotype AAAGG on the PSMA4 gene using either

Haplo.stats (OR = 0.444, 95% CI =0.249-0.792, p-value = 0.0061) or the BJLM (OR = 0.428, 95% CI = 0.236-0.758, p-value = 0.0031).

6.3.2 Developing New Spitz Lung Cancer Risk Models

The results from section 6.3.1 were then incorporated into full Spitz lung cancer risk models for former and current smokers. Each haplotype listed in section 6.3.1 was added to the original Spitz lung cancer risk factors. Haplo.stats results were collected within the Haplo.stats program and modeled within a multivariate logistic regression regime with SPSS, and BJLM results were modeled with non-informative priors in WINBUGS. Results are shown in Table 6.4.

Table 6.4: Full Updated Spitz Lung Cancer Risk Model Development Using Haplotypes inferred with Haplo.stats and the BJLM.

Results are generated for former and current smokers separately since in the original Spitz lung cancer risk models, smoking status was a matching variable.

Former Smokers						
Chm. (Gene)	Haplo.	Genetic Trait	Haplo.stats results		BJLM results	
			OR (95% Confidence Interval)	P-value	OR (95% Credible Interval)	Empirical P-value
5 (TERT)	AG	Dominant	0.72 (0.56-0.91)	0.007	0.71 (0.55-0.91)	0.006
6 (HCG4P4)	AG	Additive	1.21 (1.02-1.43)	0.026	1.21 (1.02-1.43)	0.024
6 (HLA-L)	AAGC	Additive	1.82 (1.18-2.81)	0.007	1.81 (1.17-2.80)	0.008
15 (IREB2)	CGAC	Recessive	N/A	N/A	11.5 (1.31-249)	0.025
Non-Genetic Risk Factors						
Dusts (Yes vs. No)			1.58 (1.24-2.02)	< 0.001	1.60 (1.25-2.04)	< 0.001
Emphysema (Yes vs. No)			2.60 (1.82-3.71)	< 0.001	2.63 (1.85-3.77)	< 0.001
Hay Fever (Yes vs. No)			0.67 (0.49-0.90)	0.008	0.65 (0.48-0.88)	0.006
Family History (≥ 2 vs. < 2)			1.63 (1.27-2.10)	< 0.001	1.64 (1.27-2.12)	< 0.001
Quitting Age (42-53 yrs)			1.07 (0.80-1.45)	0.639	1.10 (0.81-1.48)	0.554
Quitting Age (> 54 yrs)			1.51 (1.13-2.03)	0.006	1.53 (1.13-2.06)	0.006
Constant			0.470	< 0.001	0.462	N/A
Current Smokers						
Chm. (Gene)	Haplotype	Genetic Trait	OR (95% Confidence Interval)	P-value	OR (95% Credible Interval)	P-value
5 (TERT)	AG	Dominant	0.69 (0.52-0.91)	0.009	0.68 (0.51-0.90)	0.008
6 (HLA-L)	AAGC	Additive	0.52 (0.32-0.85)	0.009	0.53 (0.32-0.86)	0.009
15 (PSMA4)	AAAGG	Additive	0.41 (0.22-0.77)	0.006	0.38 (0.20-0.71)	0.002
15 (PSMA4)	AAAGA	Recessive	3.03 (1.58-5.82)	0.001	3.12 (1.63-6.20)	< 0.001
Non-Genetic Risk Factors						
Asbestos (Yes vs. No)			1.70 (1.14-2.53)	0.010	1.71 (1.14-2.56)	0.008
Dusts (Yes vs. No)			1.24 (0.94-1.64)	0.129	1.25 (0.94-1.65)	0.122
Emphysema (Yes vs. No)			2.66 (1.83-3.89)	< 0.001	2.71 (1.86-3.98)	< 0.001
Hay Fever (Yes vs. No)			0.67 (0.47-0.95)	0.025	0.66 (0.47-0.95)	0.022
Smoking Family History of Cancer			1.75 (1.29-2.37)	< 0.001	1.76 (1.30-2.39)	< 0.001
Pack-Years (28-42 vs. < 28)			1.22 (0.81-1.84)	0.350	1.22 (0.81-1.84)	0.342
Pack-Years (42-57.5 vs. < 28)			1.41 (0.94-2.10)	0.094	1.41 (0.95-2.10)	0.093
Pack-Years (> 57.5 vs. < 28)			1.92 (1.30-2.82)	0.001	1.93 (1.32-2.85)	0.001
Constant			0.617	0.006	0.615	N/A

The most susceptible haplotype in the updated Spitz lung cancer risk model for lung cancer development is haplotype CGAC on the IREB gene using the BJLM (OR = 11.5, 95% CI = 1.31-249) for former smokers, and this haplotype is notable because it is significant with the BJLM only. The haplotype AG on the TERT gene located on the chromosome five, and haplotype AAGC on the HLA-L gene on chromosome six are the only haplotype present in the former smoker and current smoker models. Emphysema is the most significant non-genetic risk variable in the updated models.

6.3.3. Comparing the Updated Spitz Lung Cancer Risk Models

To test the effectiveness of the full genetic models (with BJLM inferred haplotypes, and haplo.stats inferred haplotypes respectively) compared to models with only the original Spitz lung cancer risk model, the discriminatory power (Table 6.5) and NRI (Table 6.6) will be calculated and examined for all three types of models.

Table 6.5: Discriminatory Power Results for the Original Spitz Lung Model, the Spitz model with Haplo.stats inferred haplotypes, and the Spitz Model with BJLM inferred haplotypes.

(1) = Comparison of the Original Spitz models to the Haplotype extended Spitz models
(2) = Comparison of the haplotype extended Spitz models

Discriminatory Power in All Individuals (1154 cases and 1137 controls)			
Model	AUC (95% CI)	p-value	
		(1)	(2)
Original Spitz	0.659 (0.636-0.681)	---	---
Original Spitz + BJLM	0.687 (0.665-0.708)	< 0.0001	---
Original Spitz + Haplo.stats	0.684 (0.662-0.705)	< 0.0001	0.0328
Discriminatory Power for Former Smokers (604 cases and 656 controls)			
Original Spitz	0.641 (0.610-0.671)	---	---
Original Spitz + BJLM	0.665 (0.634-0.694)	0.0047	---
Original Spitz + Haplo.stats	0.659 (0.628-0.689)	0.0213	0.0440
Discriminatory Power for Current Smokers (550 cases and 481 controls)			
Original Spitz	0.672 (0.638-0.704)	---	---
Original Spitz + BJLM	0.708 (0.675-0.738)	0.0003	---
Original Spitz + Haplo.stats	0.707 (0.674-0.737)	0.0004	0.3593

With the full SNP and haplotype model with haplotypes modeled by haplo.stats and the BJLM, discriminatory power is significantly improved (p-value < 0.0001 for both models) compared to the original Spitz model for all individuals. However, for all individuals, these models constructed with haplotypes inferred from the BJLM are significantly superior to models constructed with haplo.stats inferred haplotypes (p = 0.0328) from each other, and there are significant differences between these models when separated by smoking status for former smokers (p-value = 0.0440), but not for current smokers (p-value = 0.3593).

Table 6.6: Net Reclassification Index results for comparing the Original Spitz Lung Model, the Spitz model with Haplo.stats inferred haplotypes, and the Spitz Model with BJLM inferred haplotypes

- (1) Comparison of the Original Spitz models to the Haplotype extended Spitz models
(2) Comparison of the haplotype extended Spitz models

NRI for All Individuals				
Model	(1)		(2)	
	NRI	p-value	NRI	p-value
Original Spitz	---	---	---	---
Original Spitz + Haplo.stats	0.2661	< 0.0001	---	---
Original Spitz + BJLM	0.3011	< 0.0001	0.3820	< 0.0001
NRI for Cases				
Original Spitz	---	---	---	---
Original Spitz + Haplo.stats	0.4132	< 0.0001	---	---
Original Spitz + BJLM	0.3953	< 0.0001	0.0930	0.0018
NRI for Controls				
Original Spitz	---	---	---	---
Original Spitz + Haplo.stats	-0.1471	< 0.0001	---	---
Original Spitz + BJLM	-0.0943	0.0014	0.2890	0.0001

The increase in overall NRI is higher and more significant when examining the full genetic model (for both haplo.stats and the BJLM) compared to the original Spitz lung cancer models; however, for controls only, the original Spitz lung cancer risk models outperform the models with haplotypes added to the original models. These results confirm the discriminatory power results because the model with BJLM inferred haplotypes is seen as the superior model (NRI = 0.3820, p-value < 0.0001).

6.4. Discussion for Expanding Spitz Models Using SNPs derived from the Top 200 SNP list in the Lung Cancer Meta-Analysis

The purpose of this analysis was to extend the Spitz lung cancer risk models by incorporating haplotypes to elucidate associations in lung cancer that could not be obtained solely with single SNP based strategies. This study showed that haplotypes can increase the

discriminatory power and the correct estimation of risk in cases compared to the original Spitz lung cancer risk models. For instance, the TERT, HCG4P4, HLA-L, IREB2, and PSMA4 genes all contained haplotypes in the extended Spitz models. The strongest genetic effect that showed susceptibility with Lung Cancer development occurred with the inferred haplotype CGAC from the IREB2 gene on chromosome 15. Also, the recessive genetic effect for this haplotype existed only in the Bayesian model, which suggests that incorporating prior data to rare haplotypes might lead to increased odds ratios for susceptible rare haplotypes.

Three aspects of the results from section 6.3 deserve future inquiry. First, when comparing the models developed with haplotypes inferred from haplo.stats and the BJLM, the model with BJLM inferred haplotypes was superior. Discriminatory power analysis suggests that the model BJLM inferred haplotypes perform better compare to the model with haplo.stats inferred haplotypes, although these results are not significant for current smokers. However, when examining the NRI, the model with BJLM inferred haplotypes have better results compared to the model with haplo.stats inferred haplotypes (NRI = 0.3820, p-values < 0.0001). The NRI showed that the lung cancer risk was increased in cases when measured with the BJLM model, and also showed that the lung cancer risk was decreased in controls with the BJLM, and this elucidates the notion that when measured by the NRI, BJLM modeling was superior to haplo.stats in the Texas Lung GWAS dataset.

The second aspect of the results that requires further inquiry was that both haplotype extended Spitz models, according to the NRI, were significantly worse in determining the risk profile of controls. This suggests that there could be better haplotype blocks that were not modeled in this analysis that could improve the risk profile of controls. A future analysis that could attempt to solve this problem would be to search for haplotypes throughout the Texas

Lung GWAS, or to infer SNPs in the top 200 lung cancer meta-analysis SNP list that are not in the Texas Lung GWAS but are in LD with SNPs that form haplotypes in the Texas Lung GWAS study.

The final aspect of these results that require further inquiry was the decreased risk of pack-years for those with pack-year values between 42 and 57.5. The odds ratio for those that have pack-years from 42 to 57.5 decreased from 1.45 (95% CI = 1.05-2.01, p-value = 0.024) in the original Spitz lung cancer risk model, to 1.41 (95% CI = 0.94-2.10, p-value = 0.094) in the Spitz model with haplo.stats inferred haplotypes, and the odds ratio decreased as well in the Spitz model with BJLM inferred haplotypes to 1.41 (95% CI = 0.95-2.10, p-value = 0.093). There is evidence that suggests that the genes located on the 15q25 section of the genome (LOC123688 and CHRNA5) have predisposition to nicotine dependence²⁰⁷. In a study of 15,000 European individuals, those who smoke more than 15 cigarettes per day, and especially those who smoke 25 cigarettes per day are strongly linked with SNPs and Haplotypes located in the 15q25 section of the genome²⁰⁷. One SNP in particular, rs1317286, is located on the same haplotype block in HapMap data as the AAAGG haplotype for gene PSMA4 that was protective for lung cancer, and in the study of 15,000 European individuals, SNP rs1317286 was strongly associated with those that smoke at least 25 cigarettes per day ($p = 0.0000026$)²⁰⁷. Eighty eight of the 256 individuals in the Texas Lung GWAS that have pack-years between 42 and 57.5 smoke at least 25 cigarettes per day. The presence of haplotype GAGA in this study which is protective could be decreasing the effect of smoking intensity for those in the pack-year group (42-57.5), a group that showed significance ($p < 0.05$), but not strong significance ($p < 0.01$) in the original Spitz lung cancer risk study.

In this analysis, the Spitz lung cancer risk models have been expanded with haplotypes from the Texas Lung GWAS that were inferred from SNPs that match both the Top 200 SNP in the lung cancer meta-analysis manuscript and the Texas Lung GWAS. Improved discriminatory power was noted in current smokers and ever smokers. However, to further improve the effectiveness of risk models with haplotypes, SNPs that are highly significant but not linked to any haplotypes can be added to these models in an attempt to increase discriminatory power. In the final application of the BJLM, the inflammation pathway of a Glioma GWAS will be studied and genetic risk models will be created for the first time with top SNPs and top Haplotypes.

Chapter 7: Application Three of the BJLM: Using Haplotype Analysis and the BJLM to Determine Significant Genetic Risk Factors for Glioma

To further understand and elucidate the genetic risk factors for Glioma will substantially improve both cancer prevention and treatment options for this insidious disease. Genetic analysis, both single SNP and haplotype, has been very beneficial in finding genetic associations for many cancers, like breast and lung. In this analysis, a complete risk model with both single SNPs and haplotypes will be developed to better determine the genetic risk profile of Glioma.

Cases consists of 1224 histological-confirmed Caucasian Glioma cases from MD Anderson Cancer Center in Houston, Texas, and controls consists of 2224 individuals from the Cancer Genetic Markers of Susceptibility (CGEMS) group. Genetic information from both populations was available for 4647 SNPs from 204 genes. Plink and Haploview were used to determine the top 20 SNPs and top 20 haplotype blocks respectively. Then, univariate logistic regression was used to determine the best genetic model for the top SNPs. Joint logistic risk model regression was used to determine haplotype risk for each individual haplotype block. Finally, multivariate logistic regression was used to develop complete risk model using SNPs only, Haplotypes only, and SNPs plus haplotypes.

When incorporating haplotypes into risk models, the discriminatory power increased. The SNP only risk model contained 17 SNPs, and had discriminatory power of 63.7% (95% CI = 0.617-0.657), while a risk model from the top 20 Haplotypes blocks had a discriminatory power of 64.1% (95% CI = 0.622-0.661). Finally, incorporating 15 top SNPs and 10 top Haplotypes which does not have any of the top 15 SNPs have a discriminatory power of 65.4% (95% CI = 0.634-0.673). Also, adding top haplotypes lead to a significant increase in risk modeling power as calculated by the net reclassification index (NRI = 0.1541, p-value = <0.0001).

7.1. Introducing Glioma for Genetic Analysis

According to the latest 2011 Cancer Facts and Figures report, an estimated 22,340 American individuals will be diagnosed with brain cancer this year, and 13,110 American individuals will die from this disease²⁰⁸. Roughly 77% of all brain cancers are gliomas, even though only 42% of brain tumors are Glioma, and certain types of Gliomas like Astrocytomas and Oligodendrogliomas are the most dangerous because they typically cannot be removed by surgery^{209,210}. Non-genetic risk factors for Glioma have been studied extensively, but many of the hypothesized risk factors (smoking, hazards from certain occupations, heavy cell phone use, and radiation exposure) show results that are weak and sometimes contradictory²¹¹⁻²¹⁸.

To find some genetic risk factors for Glioma, there seems to be some promise for those that focus on inflammation genes. Some studies have found single nucleotide polymorphisms that can both protect and be susceptible to Glioma, and also have found biomarkers, like Immunoglobulin E levels, that can show protection against Glioma formation^{209, 219-224}. Recently, a study examining interactions among pairs of SNPs has shown promise with finding new risk factors for Glioma, as the MAP3K7 and the CRADD genes have about 14% of the significant interactions among SNPs in individuals with Glioma²²⁴. This information suggests that risk models from genetic risk factors in the inflammation pathway can be a viable option in both cancer prevention and cancer recognition for Glioma.

Risk models can be developed using both frequentist and Bayesian methods. With frequentist methods, models are created with the information readily available for that specific study and with no mathematical assumptions using prior information. For instance, in the original development of the Spitz model, risk models for never, former, and current smokers were developed based on the available information of 1851 cases and 2001 controls with

smoking intensity, family history of lung cancer, self-reported emphysema, and other risk factors²⁷. With Bayesian methods, one can add prior information to the dataset being examined with the hopes of strengthen associations or non-associations with disease. For instance, HapMap⁴⁴ and 1000 genome information⁴³ can be used to better infer haplotypes in individuals in which the SNP information is known or unknown, and this is the basis for the BJLM introduced in Chapter 4 of this thesis. In this analysis, genetic risk factors will be analyzed using SNP information available from American Glioma cases at MD Anderson Cancer Center and controls from the CGEMS studies²²⁵⁻²²⁷. Both top haplotypes and top SNPs will be extracted in separate analysis to determine their risk profiles. Afterward, logistic regression analysis will be conducted with use of both SPSS (for a frequentist prospective) and the BJLM to determine the best possible model for haplotypes. Finally, both Bayesian and frequentist methods will be incorporated to determine the best risk model combining both top SNPs and top Haplotypes.

7.2. Methods for Glioma Analysis

7.2.1. Study population and Selection of SNPs for Glioma Analysis

A study population of 3448 individuals (1224 cases and 2224 controls) was accrued for this analysis. Study population information for the cases and controls have been discussed previously, but briefly, Glioma cases (N = 1224) were ascertained from MD Anderson cancer center from 1990 to 2008, and all individuals were Caucasian adults²²⁴. A small aliquot of blood (20-ml) was collected from each case for genetic analysis²²⁴, and genotyping was conducting using Illumina HumanHap 610 SNP Chip, which contained 575,837 SNPs for possible analysis²²⁵. The controls for this experiment (N = 2224) were ascertained from CGEMS, which was launched in 2005 to identify breast and prostate cancer variants²²⁸, and all of these samples were of Caucasian descent. CGEMS initially contained 1,142 breast cancer

controls and 1,101 prostate cancer controls which were genotyped with the Illumina 550K Bead-Chip²²⁵⁻²²⁷. After removal of individuals who had low call rates or of non-European descent, 2224 controls were available for analysis. The study received institutional review board approval from MD Anderson Cancer Center.

The selection of SNPs in the cases has been discussed previously²²⁴, but briefly, an all encompassing list of key signaling pathways for inflammation was extracted using both the Biocarta pathway maps website (<http://www.biocarta.com/genes/allpathways.asp>) and the Anatomy project for Cancer Genomes project (<http://cgap.nci.nih.gov/Pathways>) website. Twenty eight inflammation pathways containing 204 genes were found for possible analysis, and in these 204 genes, there were 5304 SNPs in the case dataset. Out of these 5304 SNPs, 4647 SNPs were also available in the CGEMS control data set, and all of these SNPs had call rates > 95%, a minor allele frequency greater than >1% in the controls, and HWE equilibrium p-values of > 1×10^{-5} in the controls.

7.2.2. Determining top SNPs for Model Analysis

Genetic information from section 7.2.1 was transformed into a .ped file, which is a standard file format for genetic analysis, and ran in PLINK²²⁹ using the basic case/control association analysis option (`plink -file mydata -assoc`). With this option, PLINK runs a simple chi-square test for each SNP in the study. After completion of this analysis, the 20 SNPs with the lowest p-values are extracted. With these top 20 SNPs, univariate logistic regression in SPSS (version PASW Statistics 17.0) is conducted on each SNP using these three genetic traits: Additive, Dominant, and Recessive, to determine which genetic trait has the lowest p-value in the analysis. Finally, the top 20 SNPs with the best genetic model for each SNPs are examined together with multivariate logistic regression in SPSS to determine which SNPs have the best

association with Glioma, and also to develop a risk model with SNPs as the only risk factors. Finally, the discriminatory power for this specific model was calculated in SPSS.

7.2.3. Determining top Haplotypes for Glioma Model Analysis

To extract the best haplotypes within the top 20 haplotype blocks, chromosome and base pair information for all 4647 SNPs was determined from a genetic map SNP site associated with the Conway Institute Bioinformatics Service (<http://integrin.ucd.ie/cgi-bin/rs2cm.cgi>). For each chromosome, all SNPs were loaded into Haploview⁵⁶, and all results from the Association tab and Haplotypes sub-tab were extracted. All haplotypes are formed with the Gabriel's et.al.⁵³ confidence interval method. Then, all haplotype blocks that contained haplotypes with p-values of less than 0.01 were collected from all chromosomes. After collection of these haplotype blocks, the top 20 haplotypes blocks were then selected for further analysis.

All haplotypes within the 20 top haplotype blocks were then further analyzed with both Haplo.stats⁶⁰ and the BJLM (from Chapter 4). The procedure to conduct this analysis was discussed in **section 5.2.4**, but briefly, each haplotype block will be analyzed with Haplo.stats first to determine the odds ratio and p-value of the significant haplotypes that will be used to begin construction of a haplotype only Glioma model. All haplotypes with empirical p-values less than 0.05 and are the most significant among the three risk models in each haplotype block are presented for future analysis. This analysis is conducted for assuming additive, dominant, and recessive genetic traits for each haplotype block.

The procedure for adding haplotypes to risk models when using the BJLM was also discussed in **section 5.2.4.**, but briefly, the prior information was obtained by determining the counts of all possible haplotypes inferred with HapMap data¹³⁷⁻¹⁴⁰. These counts were then incorporated into the inferring of haplotypes from the Glioma dataset. After running the BJLM,

the haplotype frequencies for each haplotype block, except for the most frequent haplotype, was saved and then presented to WINBUGS (version 1.4.3) for analysis with each haplotype block. This analysis is conducted for all three genetic models for each haplotype block, and haplotype results are collected after 50,000 iterations of the WINBUGS code (1st 5000 iterations are burn-in, and are not counted in the analysis). All haplotypes with empirical p-values less than 0.05 and are the most significant among the three risk models in each haplotype block are presented for future analysis.

For determining the haplotypes used in development of a Glioma risk model, the same procedure to incorporate the results from each haplotype block was also used to expand the Spitz lung cancer risk model. The most significant haplotype out of the three genetic models are selected for multivariate logistic regression analysis in both haplo.stats and the BJLM. The discriminatory power for both haplotype models are calculated in SPSS using the relative risk profiles calculated by the multivariate logistic risk models developed in either SPSS (frequentist) using haplotypes inferred from haplo.stats or the haplotypes inferred by the BJLM (Bayesian).

7.2.4. Combining both SNP and Haplotype Data

To develop a model based on the best SNPs and best Haplotypes (using the BJLM) from the previous two parts, it is imperative to make sure that there is no inclusion of a top 20 SNP that is also included in the same model as a haplotype containing the most significant SNPs. In this case, all haplotypes that are not inferred with a top 20 significant SNP will be added to the model consisting of single SNPs. The list of haplotypes that are missing the most significant SNPs are listed in the 1st Table of Appendix 3 (Appendix 3: Table 1).

The top 20 SNPs are then added to the haplotypes from Appendix 3: Table 1 to construct a more complete model for Glioma using multivariate logistic regression. Three models will be

constructed and analyzed. The SNPs included in the single SNP model will be modeled with the haplotypes determined by Haplo.stats in one model, and with haplotypes determined by the BJLM for the other two models. From the two models with BJLM inferred haplotypes, one model will contain non-informative priors that are normal distributions with mean zero and precision of 0.01 for the SNPs and haplotypes. The other model will incorporate informative priors with normal prior distributions based on the means and precisions of the results from the univariate logistical regression and joint-effect logistic analysis for SNPs and haplotypes respectively. WINBUGS code for this model will be listed in Appendix 3: Figure 1. Models including haplotype information from haplo.stats will be examined with SPSS, while those models with BJLM inferred haplotypes will be examined with WINBUGS. Discriminatory power will be calculated with all three models with NCSS/PASS, and also the Net Reclassification Index (NRI)²³⁰ will be calculated with these three models separately compared to the models with SNPs only and haplotypes only.

7.3. Results for Glioma Analysis

7.3.1. Single SNP model development

After running PLINK for all 4647 inflammation SNPs, the top 20 most significant SNPs are listed in Appendix 3: Table 2 and highlighted in the Manhattan plot below (Figure 7.1).

Figure 7.1: Manhattan Plot showing the 4647 SNPs examined for Glioma Analysis

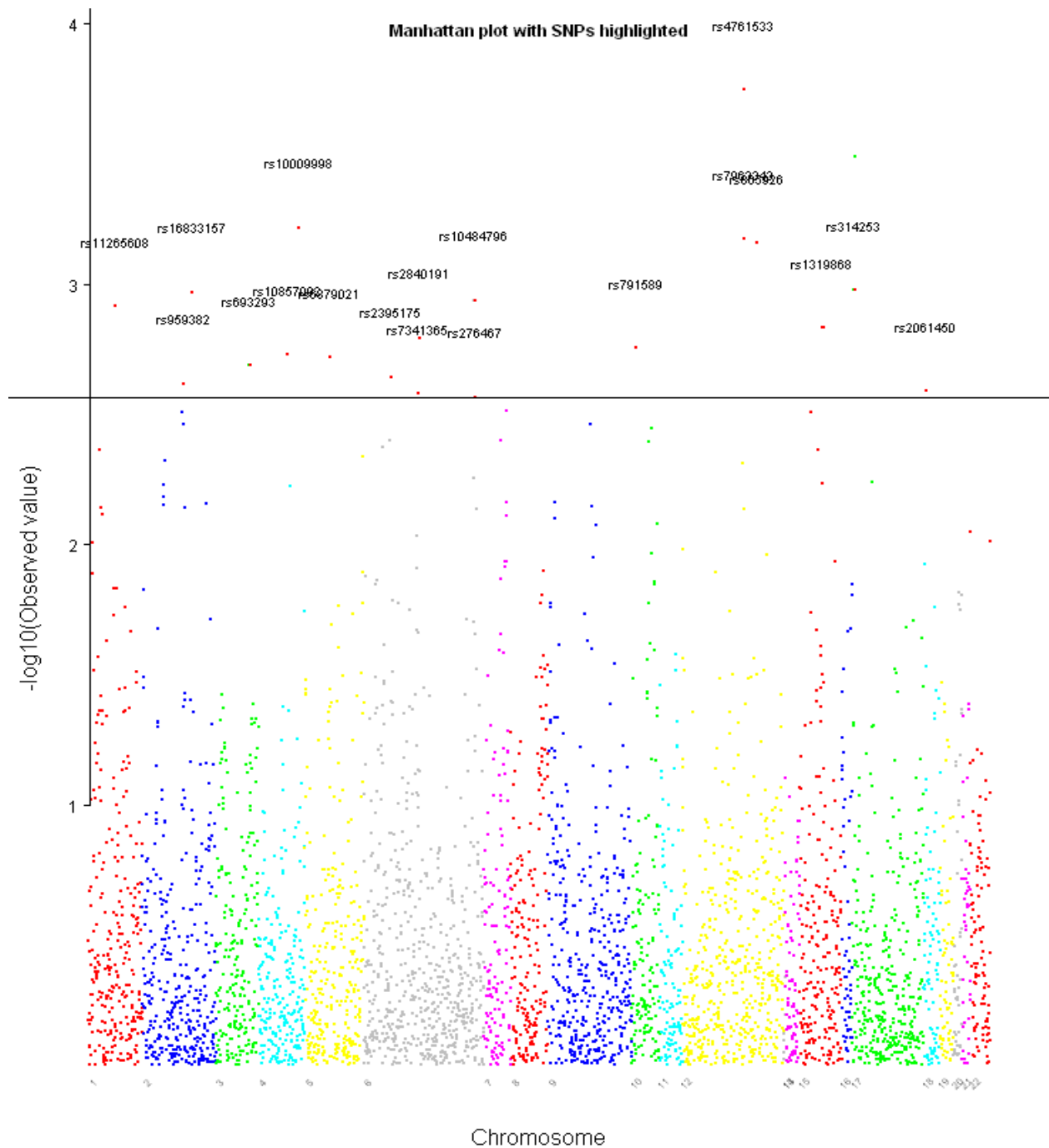


Figure 7.1: Chromosome six contains the most significant inflammation SNPs with five, while chromosomes 2, 4, 12, and 17 contain at least two significant inflammation SNPs. All SNPs have p-values less than 0.003, which suggests that these SNPs could be viable markers for Glioma detection. Five SNPs in particular, rs4761533, rs8079544, rs10009998, rs7963343, and rs865926, have p-values less than 0.001. Twenty SNPs have p-values smaller than 0.002712, which is transformed by negative log 10.

Univariate logistic regression analysis is then conducted with the SNPs listed in Appendix 3: Table 2, and results assuming an additive, dominant, and recessive genetic trait are generated. Genetic models that lead to the most significant results for each SNP are displayed in the Appendix 3: Table 3. With these results, assuming dominant genetic traits leads to the most significant univariate logistic regression results for 11 SNPs, while for additive and recessive genetic traits, leads to the most significant univariate logistic regression results for eight and one SNP model. The most susceptible SNP for Glioma development are SNP rs11265208 assuming an additive genetic trait (OR = 1.303, 95% CI = 1.111-1.529), and SNP rs8079544 assuming a dominant genetic trait (OR = 1.499, 95% CI = 1.212-1.854). No significant susceptible SNPs are available assuming a recessive genetic trait; however, the most protective SNP, rs10009998, against Glioma is assumed from a recessive genetic trait (OR = 0.109, 95% CI = 0.034-0.350).

Multivariate logistic regression was conducted for the 20 SNPs with the best genetic model for each SNP as determined with Appendix 3: Table 3, and the results are displayed in Table 7.1.

Table 7.1: Full Single SNP model using Inflammation SNPs for Glioma

The rs number for each SNP is listed on the 1st column from the left, and the genetic trait modeled for each SNP is listed on the 2nd column from the left. The magnitude of susceptibility to Glioma for each SNP is defined in terms of betas and odds ratio in columns 3 and 4 respectively with its associated p-value in the final column.

SNP	Genetic Trait	Beta	OR (95% CI)	P-Value
rs4761533	Dominant	0.346	1.413 (1.186-1.685)	< 0.001
rs8079544	Dominant	0.442	1.556 (1.242-1.949)	< 0.001
rs10009998	Recessive	-2.256	0.105 (0.032-0.340)	< 0.001
rs865926	Dominant	0.352	1.421 (1.104-1.830)	0.006
rs314253	Additive	0.190	1.210 (1.085-1.349)	0.001
rs16833157	Additive	-0.419	0.658 (0.504-0.859)	0.002
rs10484796	Additive	-0.121	0.886 (0.797-0.985)	0.025
rs11265608	Additive	0.318	1.375 (1.162-1.627)	< 0.001
rs1319868	Dominant	0.286	1.332 (1.120-1.582)	0.001
rs791589	Additive	0.227	1.255 (1.078-1.462)	0.003
rs10857092	Dominant	0.235	1.264 (1.023-1.563)	0.030
rs6879021	Additive	0.146	1.157 (1.043-1.283)	0.006
rs693293	Dominant	-0.333	0.717 (0.591-0.869)	0.001
rs2395175	Additive	0.176	1.193 (1.035-1.375)	0.015
rs959382	Dominant	0.196	1.217 (1.040-1.423)	0.014
rs2061450	Additive	-0.176	0.839 (0.741-0.948)	0.005
rs276467	Dominant	0.199	1.221 (1.049-1.421)	0.010
Constant		-1.169	0.311	<0.001

With these results, the SNP with the most susceptibility toward Glioma development and the most protective SNP against Glioma are the same as shown in the univariate logistic regression analysis. Three SNPs, rs7963343, rs2840191, and rs7341365 lose their significance when examined with the rest of the SNPs that compose the single SNP model. Finally, almost half of the SNPs in this full SNP model have p-values of less than or equal to 0.001, which shows potential strong factors for Glioma susceptibility or protection.

7.3.2. Haplotype model development for Glioma Data Set

In terms of most significant p-value, the top 20 haplotypes as calculated by Haploview are shown in Table 7.2.

Table 7.2: Best Haplotypes within Glioma Inflammation Dataset as Calculated by Haploview

Haplotype information is listed on the 1st 2 columns on the left, and the frequency of each significant haplotype with its associated Chi-Square and P-Values are listed in the last three columns. All results are generated by testing the frequency differences between cases and controls in Haploview.

Chromosome (Gene)	Haplotype	Frequency	Chi-Square	P-Value
12 (CRADD)	GCAGA	0.116	14.414	0.0001
1 (IL6R)	GGAA	0.103	11.186	0.0008
1 (CD247)	AGG	0.012	11.209	0.0008
17 (DLG4)	AG	0.345	10.455	0.0012
6 (HLA-DRA)	ATT	0.166	10.181	0.0014
12 (IGF1)	AAT	0.054	10.025	0.0015
15 (IGF1R)	AT	0.119	10.117	0.0015
6 (IL22RA2)	CCTGA	0.524	9.869	0.0017
3 (PIK3CB)	TGCGACC	0.100	9.777	0.0018
6 (MAP3K7)	GT	0.272	9.624	0.0019
12 (CRADD)	AGCGG	0.083	9.69	0.0019
15 (SMAD3)	GTTTA	0.231	9.334	0.0022
16 (ITGAX)	CGA	0.433	9.307	0.0023
6 (HLA-DRA)	GCGACCAGTAC	0.156	9.047	0.0026
1 (CD247)	TA	0.496	8.993	0.0027
2 (STAT)	AGA	0.249	8.789	0.003
6 (MAP3K7)	AC	0.724	8.623	0.0033
10 (PRF1)	CGT	0.669	8.446	0.0037
7 (GNAI1)	TTTCA	0.226	8.325	0.0039
6 (HLA-DRA)	ACGACCGGGGC	0.133	8.194	0.0042

From these top 20 haplotypes, 18 of these haplotypes were on unique haplotype blocks, with two top haplotypes inferred on the same haplotype block in genes HLA-DRA and MAP3K7. All of the top haplotypes were highly significant with P-values less than 0.005. One haplotype from the CRADD, IL6R, and CD247 genes had p-values less than 0.001. Finally, haplotype length played

no role in association results, as haplotypes of length 2 to 5 SNPs, 7 SNPs, and 11 SNPs all showed significant associations with Glioma.

Joint-effect logistic model analysis is then conducted with the haplotypes listed in Table 26, and results assuming an additive, dominant, and recessive genetic trait are generated. Genetic models that lead to the most significant results for each haplotype with both Haplo.stats and the BJLM are displayed in the Appendix 3: Table 4. With these results, assuming dominant genetic traits leads to the most significant joint-effect logistic model results for 10 haplotypes, while for additive and recessive genetic traits, leads to the most significant univariate logistic regression results for eight and two haplotypes respectively. The most susceptible haplotype for Glioma development is haplotype AGG on the CD247 gene using either Haplo.stats (OR = 2.10, 95 %CI = 1.35-3.25) or the BJLM (OR = 2.1, 95% CI = 1.36-3.27), and the most protective haplotype against Glioma development is haplotype GCTCA on the SMAD3 gene using either Haplo.stats (OR = 0.26, 95% CI = 0.08-0.86) or the BJLM (OR = 0.22, 95% CI = 0.05-0.70). With the haplotype GCTCA from the SMAD3 gene, the p-value was much more significant when using the BJLM (p-value = 0.00677) compared to using Haplo.stats (p-value = 0.0277) to infer the haplotypes.

These results were then incorporated into a full haplotype model analysis in which each haplotype listed above were tested together. Haplo.stats results were collected within the Haplo.stats program and modeled within a multivariate logistic regression regime with SPSS, and BJLM results were modeled with non-informative priors in WINBUGS. Results are shown in Table 7.3.

Table 7.3: Multivariate logistic regression for inferred haplotypes from both BJLM and Haplo.stats

Haplotype information is listed on the 1st 3 columns on the left, and the results from Haplo.stats and BJLM are listed in the last 4 columns on the right. Significant associations from each haplotype block are highlighted in this table.

(1) = Empirical Bayes P-Value as calculated in Chapter 4 of this thesis

Chm. (Gene)	Haplotype	Genetic Trait	Haplo.stats results		BJLM results	
			OR (95% CI)	P-value	OR (95% CI)	P-value ⁽¹⁾
12 (CRADD)	GCAGA	Dominant	1.39 (1.17-1.65)	< 0.001	1.39 (1.17-1.65)	<0.001
1 (IL6R)	GGAA	Additive	1.29 (1.09-1.52)	0.003	1.29 (1.09-1.51)	0.002
1 (CD247)	AGG	Dominant	1.96 (1.24-3.09)	0.004	1.99 (1.26-3.15)	0.003
17 (DLG4)	AG	Additive	1.22 (1.09-1.35)	< 0.001	1.21 (1.09-1.35)	< 0.001
6 (HLA-DRA)	ATT	Additive	0.81 (0.70-0.93)	0.004	0.81 (0.70-0.93)	0.004
12 (IGF1)	AAT	Dominant	1.44 (1.15-1.81)	0.002	1.44 (1.15-1.81)	0.003
15 (IGF1R)	AG	Dominant	1.26 (1.08-1.48)	0.004	1.26 (1.08-1.48)	0.003
15 (IGF1R)	AT	Dominant	1.41 (1.19-1.68)	< 0.001	1.42 (1.19-1.69)	< 0.001
6 (IL22RA2)	TCGGG	Additive	0.86 (0.76-0.97)	0.017	0.86 (0.76-0.97)	0.016
6 (IL22RA2)	TTGAA	Dominant	0.76 (0.63-0.93)	0.008	0.76 (0.62-0.93)	0.007
3 (PIK3CB)	TGCGAC C	Dominant	0.72 (0.60-0.88)	0.001	0.73 (0.60-0.88)	0.001
6 (MAP3K7)	GT	Dominant	1.33 (1.15-1.54)	<0.001	1.33 (1.15-1.54)	<0.001
12 (CRADD)	AGCGG	Recessive	2.48 (1.00-6.13)	0.049	2.53 (1.03-6.62)	0.042
15 (SMAD3)	GCTCA	Recessive	0.22 (0.06-0.75)	0.016	0.20 (0.04-0.62)	0.003
15 (SMAD3)	GTTTA	Additive	0.83 (0.73-0.93)	0.002	0.82 (0.73-0.93)	0.002
16 (ITGAX)	AGG	Dominant	1.25 (1.08-1.44)	0.003	1.25 (1.08-1.45)	0.002
6 (HLA-DRA)	GCGACC AGTAC	Additive	1.17 (1.02-1.35)	0.026	1.18 (1.02-1.35)	0.026
1 (CD247)	TG	Additive	1.15 (1.03-1.30)	0.016	1.14 (1.02-1.28)	0.027
1 (CD247)	GA	Additive	1.19 (1.03-1.37)	0.016	1.17 (1.02-1.34)	0.031
2 (STAT)	AGA	Additive	0.85 (0.76-0.96)	0.008	0.85 (0.75-0.96)	0.007
10 (PRF1)	TGC	Additive	1.14 (1.02-1.28)	0.020	1.15 (1.02-1.28)	0.020
7 (GNAI1)	TTTCA	Dominant	1.26 (1.09-1.46)	0.002	1.27 (1.09-1.47)	0.001
Constant			0.251 (N/A)	< 0.001	0.253 (N/A)	< 0.001

With these results, the GCTCA haplotype on gene SMAD3 was still the most protective haplotype against Glioma using either Haplo.stats to infer the haplotypes (OR = 0.22, 95% CI = 0.06-0.75) or the BJLM (OR = 0.20, 95% CI = 0.04-0.62). However, the AGCGG haplotype on

the CRADD gene was now the most susceptible haplotype toward Glioma development using either haplotypes inferred from Haplo.stats (OR = 2.48, 95% CI = 1.00-6.13) or the BJLM (OR = 2.53, 95% CI = 1.03-6.62). In this model, the AGCGG haplotype on the CRADD gene was modeled assuming a recessive genetic trait because this haplotype failed significance assuming both additive and dominant genetic traits in the multivariate logistic regression.

7.3.3. Model development with both SNPs and Haplotypes for Glioma

7.3.3.1. Best SNPs + Best Haplo.stats Inferred Haplotypes for Glioma

A full genetic model which incorporates both the single SNP model results from section 7.3.1, and the haplotype results from section 7.3.2 are developed with both SPSS and the BJLM. Full SNP and haplotype data is available for 1194 cases and 2056 controls. Results using haplotypes from Haplo.stats and SNP data are listed in Table 7.4.

Table 7.4: Best SNPs plus Haplo.stats inferred haplotypes for full genetic Glioma model

Best SNPs				
SNP	Genetic Trait	Beta	OR (95% CI)	P-Value
rs4761533	Dominant	0.311	1.365 (1.141-1.633)	0.001
rs8079544	Dominant	0.456	1.577 (1.256-1.981)	< 0.001
rs10009998	Recessive	-2.268	0.104 (0.032-0.338)	< 0.001
rs314253	Additive	0.208	1.231 (1.102-1.375)	< 0.001
rs16833157	Additive	-0.425	0.654 (0.500-0.855)	0.002
rs10484796	Additive	-0.140	0.870 (0.781-0.968)	0.011
rs11265608	Additive	0.316	1.372 (1.158-1.627)	< 0.001
rs1319868	Dominant	0.265	1.303 (1.095-1.551)	0.003
rs791589	Additive	0.221	1.248 (1.071-1.454)	0.005
rs10857092	Dominant	0.248	1.281 (1.035-1.585)	0.023
rs6879021	Additive	0.144	1.155 (1.041-1.282)	0.007
rs693293	Dominant	-0.328	0.720 (0.593-0.875)	0.001
rs959382	Dominant	0.238	1.268 (1.083-1.486)	0.003
rs2061450	Additive	-0.184	0.832 (0.734-0.942)	0.004
rs276467	Dominant	0.198	1.219 (1.046-1.421)	0.011
Best Haplotypes without Best SNPs				
Chm. (Gene)	Haplotype	Genetic Trait	OR (95% CI)	P-Value
1 (CD247)	AGG	Dominant	2.146 (1.336-3.448)	0.002
6 (HLA-DRA)	ATT	Additive	0.787 (0.679-0.911)	0.001
12 (CRADD)	AGCGG	Recessive	3.291 (1.242-8.725)	0.017
15 (SMAD3)	GCTCA	Additive	0.826 (0.728-0.938)	0.003
15 (SMAD3)	GTTTA	Recessive	0.236 (0.068-0.818)	0.023
16 (ITGAX)	AGG	Dominant	1.251 (1.076-1.454)	0.004
1 (CD247)	TG	Additive	1.149 (1.020-1.295)	0.023
1 (CD247)	GA	Additive	1.166 (1.008-1.348)	0.039
10 (PRF1)	TGC	Additive	1.136 (1.012-1.277)	0.031
7 (GNAI1)	TTTCA	Dominant	1.259 (1.083-1.464)	0.003
		Beta	Odds Ratio	P-Value
Constant		-1.384	0.251	<0.001

In this model, the strongest variable toward Glioma development is the inferred haplotype

AGCGG on the CRADD gene (OR = 3.291, 95% CI = 1.242-8.725), while the most protective variable against Glioma exists with SNP rs10009998 (OR = 0.104, 95% CI = 0.032-0.338).

Also, both of these genetic risk factors were modeled assuming recessive genetic traits. None of the haplotypes had p-values less than 0.001, while 4 SNPs had p-values less than 0.001. The

most susceptible genetic risk variable assuming either an additive or dominant genetic trait was the haplotype AGG on gene CD247 assuming a dominant genetic trait (OR = 2.146, 95% CI = 1.336-3.448), and the most protective genetic risk variable that was not calculated assuming recessive genetics was SNP rs16833157 when assuming an additive genetic trait (OR = 0.654, 95% CI = 0.500-0.855).

To test the effectiveness of the full genetic model (with Haplo.stats inferred haplotypes) compared to models with just haplotypes and SNPs only, the discriminatory power and NRI will be calculated and examined for all three types of models. With the full SNP and haplotype model with haplotypes modeled by Haplo.stats, discriminatory power is significantly improved (p-value = 0.0112) with the full genetic model (AUC = 0.654, 95% CI = 0.634-0.673) compared to the SNP only model (AUC = 0.637, 95% CI = 0.617-0.656). The full genetic model did have improved discriminatory power compared to the haplotype only model (AUC = 0.642, 95% CI = 0.622-0.661), but these results are not significant (p-value = 0.1295). The increase in overall NRI is higher and more significant when examining the full genetic model compared to the SNP only model (NRI = 0.1541, p-value = <0.0001), then examining the full genetic model compared to the haplotype only model (NRI = 0.0766, p-value = 0.0335). More improvement is shown for cases when SNPs are added to a haplotype only model (NRI = 0.1752, p-value = <0.0001) compared to when haplotypes are added to a SNP only model (NRI = 0.0545, p-value = 0.0595). Finally, adding haplotypes to a SNP only model substantially improves the results for the controls (NRI = 0.0996, p-value = < 0.0001) compared to adding SNPs to a Haplotype only model (NRI = -0.0986, p-value = < 0.0001).

7.3.3.2. Best SNPs + Best BJLM Inferred Haplotypes for Glioma

The next step is to conduct this same analysis using WINBUGS in which all SNPs and Haplotypes from the previous section are modeled with non-informative priors and with 5000 burn-in and 45000 non-burnin iterations. Models will be constructed using both non-informative priors and informative priors for each SNP and haplotype. Results using haplotypes from Haplo.stats and SNP data are listed in Table 7.5.

Table 7.5: Complete Genetic Risk Model with SNPs and Haplotypes modeled together
using WINBUGS

Non-Inform priors refers to Priors as defined in **Figure 6.2 on page 105**, where the distribution is normal with mean zero and variance of 100. Informative priors refer to the Priors from **Appendix 3: Figure One on page 196 and page 197**

	Non-Inform Priors		Informative Priors	
SNP	OR (95% CI)	P-Value	OR (95% CI)	P-Value
rs4761533	1.37 (1.14-1.64)	0.000711	1.39 (1.23-1.57)	$< 2.1 \times 10^{-5}$
rs8079544	1.58 (1.26-1.98)	0.000089	1.54 (1.31-1.79)	$< 2.1 \times 10^{-5}$
rs10009998	0.09 (0.02-0.27)	$< 2.1 \times 10^{-5}$	0.10 (0.04-0.23)	$< 2.1 \times 10^{-5}$
rs314253	1.23 (1.10-1.38)	0.000178	1.21 (1.12-1.30)	$< 2.1 \times 10^{-5}$
rs16833157	0.65 (0.49-0.85)	0.001422	0.66 (0.55-0.79)	$< 2.1 \times 10^{-5}$
rs10484796	0.87 (0.78-0.97)	0.011200	0.86 (0.80-0.93)	0.000089
rs11265608	1.38 (1.16-1.63)	0.000222	1.34 (1.19-1.50)	$< 2.1 \times 10^{-5}$
rs1319868	1.31 (1.10-1.55)	0.005066	1.33 (1.18-1.49)	$< 2.1 \times 10^{-5}$
rs791589	1.25 (1.07-1.46)	0.005066	1.25 (1.13-1.39)	$< 2.1 \times 10^{-5}$
rs10857092	1.28 (1.03-1.59)	0.023340	1.33 (1.15-1.54)	0.000133
rs6879021	1.16 (1.04-1.28)	0.007556	1.16 (1.08-1.25)	$< 2.1 \times 10^{-5}$
rs693293	0.72 (0.59-0.87)	0.000711	0.73 (0.63-0.83)	$< 2.1 \times 10^{-5}$
rs959382	1.27 (1.08-1.49)	0.003734	1.29 (1.16-1.44)	$< 2.1 \times 10^{-5}$
rs2061450	0.83 (0.73-0.94)	0.002666	0.83 (0.77-0.91)	$< 2.1 \times 10^{-5}$
rs276467	1.22 (1.05-1.42)	0.010978	1.24 (1.12-1.38)	0.000044
Haplotype (Gene)	OR (95% CI)	P-Value	OR (95% CI)	P-Value
AGG (CD247)	2.16 (1.34-3.51)	0.001244	2.12 (1.54-2.91)	$< 2.1 \times 10^{-5}$
ATT (HLA-DRA)	0.79 (0.68-0.91)	0.001467	0.79 (0.71-0.87)	$< 2.1 \times 10^{-5}$
AGCGG (CRADD)	3.43 (1.30-9.78)	0.011734	3.30 (1.73-6.34)	0.000178
GCTCA (SMAD3)	0.82 (0.73-0.94)	0.003288	0.83 (0.75-0.91)	0.00044
GTTTA (SMAD3)	0.20 (0.05-0.67)	0.006222	0.24 (0.10-0.55)	0.000311
AGG (ITGAX)	1.25 (1.08-1.46)	0.003422	1.26 (1.14-1.40)	0.00044
TG (CD247)	1.15 (1.02-1.30)	0.020940	1.15 (1.07-1.25)	0.000711
GA (CD247)	1.17 (1.01-1.35)	0.036660	1.17 (1.07-1.30)	0.001289
TGC (PRF1)	1.14 (1.01-1.28)	0.029560	1.15 (1.06-1.25)	0.000400
TTTCA (GNAI1)	1.26 (1.08-1.47)	0.002800	1.23 (1.11-1.37)	$< 2.1 \times 10^{-5}$

With these results, no matter if the association results for the genetic risk factors were calculated with non-informative priors or informative priors, the most susceptible risk factors and the most protective risk factors for Glioma are the same as the results using Haplo.stats haplotypes and the best SNPs from the previous analysis. The odds ratios with the more informative priors are not as extreme for the GTTTA haplotype on the SMAD 3 gene (OR = 0.20 with non-informative priors vs. OR = 0.24 with informative priors), and also for the AGCGG haplotype for the CRADD gene (OR = 3.43 with non-informative priors vs. OR = 3.30 with informative priors).

However, the more informative priors lead to much stronger associations in terms of smaller empirical p-values. With the informative priors, 15 out of the 25 genetic risk factors had no iterations where the odds ratio was either below one for those associations with mean odds ratios greater than one, or the odds ratio was above one for those associations with mean odds ratios less than one. The discriminatory power for models with BJLM inferred haplotypes and non-informative priors is 0.654 (95% CI = 0.634-0.673), and this model is significantly better compared to the SNP only model (p-value = 0.0110), but not significantly better compared to the haplotype only model (p-value = 0.1283). For models with BJLM inferred haplotypes plus SNPs and informative priors, the discriminatory power is 0.653 (95% CI = 0.633-0.673). This model is significantly better compared to the SNP only model (p-value = 0.0143), but not significantly better compared to the haplotype only model (p-value = 0.1543).

To further test the effectiveness of the full genetic model (with BJLM inferred haplotypes) compared to models with just haplotypes and SNPs only, the NRI will be calculated and examined for both BJLM based models. The increase in overall NRI is higher and more significant when examining the full genetic model compared to the SNP only model with both non-informative (NRI = 0.1594, p-value = <0.0001) and informative (NRI = 0.1494, p-value =

<0.0001) priors, then examining the full genetic model compared to the haplotype only model with both non-informative (NRI = 0.0818, p-value = 0.0246) and informative (NRI = 0.0780, p-value = 0.0320) priors.

More improvement is shown for cases when SNPs are added to a haplotype only model with non-informative (NRI = 0.1625, p-value = <0.0001) and informative (NRI = 0.1558, p-value = <0.0001) priors compared to when haplotypes are added to a SNP only model with either non-informative (NRI = 0.0486, p-value = 0.0932) or informative (NRI = 0.0532, p-value = 0.0640) priors. Finally, adding haplotypes to a SNP only model substantially improves the results for the controls with either non-informative (NRI = 0.1108, p-value = < 0.0001) or informative (NRI = 0.0958, p-value = < 0.0001) priors compared to adding SNPs to a Haplotype only model with either non-informative (NRI = -0.0807, p-value = 0.0003) or informative (NRI = -0.0778, p-value = 0.0004) priors.

7.4. Discussion of Glioma Modeling

The purpose of this study was to develop genetic risk modeled that incorporated haplotype analysis to elucidate associations in Glioma that could not be obtained solely with single SNP based strategies. This study showed that haplotypes can increase the discriminatory power and the correct estimation of risk in cases and controls, and that even in the presence of very significant SNPs, 10 haplotypes did show significant association with Glioma development. For instance, the SMAD3 and CD247 gene were well represented, with five out of 10 haplotypes in the full genetic model. The strongest genetic effect that showed susceptibility with Glioma development occurred with the inferred haplotype AGCGG from the CRADD gene on chromosome 12. Also, the genetic effect of the AGCGG haplotype on the CRADD gene jumped almost forty percent in the Bayesian model, from OR = 2.53 to OR = 3.43 when SNPs

were included in the analysis, which suggests that adding SNPs to haplotype only model may be very beneficial in finding strong, but rare associations with haplotypes.

Variant alleles on the *CD247* gene have been shown to be important factors in designating risk for diseases such as lymphocytic leukemia and systemic sclerosis^{230,231}. *CD247* is known to interact with key chronic lymphocytic leukemia biomarker ZAP70, with an increase of activity of roughly 1.8 times that seen in normal individuals ($p\text{-value} = 7.1 \times 10^{-8}$)²³⁰. However, for systemic sclerosis, SNP rs2056626, a SNP on the *CD247* gene, showed a slight protective effect against systemic sclerosis²³¹. *CD247* encodes subunits of the T-cell receptor zeta complex, participates in immune response regulation, and has some local expression in the adult brain²³¹⁻²³³. Since the local expression of the gene occurs in the brain for adults, it could be hypothesized that altered expression of this gene could lead to increased risk for Glioma.

Another gene which showed multiple associations with Glioma was *SMAD3*, or mothers against decapentaplegic homolog 3 gene located on Chromosome 15. However, unlike the *CD247* gene, variants in this gene were highly protective against Glioma development. *SMAD3* is involved with in modulating signals with the protein activin and transforming growth factor beta, and can limit expression the expression of human telomerase reverse transcription²³⁴. Depressed expression of this gene could be a risk factor for cancer, especially with gastric cancer, where 3/8th of the cases have no detective levels of Gastric cancer²³⁴. *SMAD3* has shown decreased expression generally in stage 4 glioblastoma and other forms of Glioma²³⁴⁻²³⁶. In this analysis, the up-regulation of haplotypes along the *SMAD3* strongly protect individuals from Glioma. Haplotypes consisting of variant alleles from the *SMAD3* gene could eventually be a great target for possible gene or drug therapy.

Genomic risk factors with large effects on either Glioma development or protection against Glioma have been found on the death domain-containing protein (CRADD) gene located on chromosome 12, and the SNP rs10009998, which is located on chromosome 4. First, the CRADD gene has been shown to interact with the caspase-2 pathway, which is a pathway that helps to regulate apoptosis via the mitochondria²³⁷⁻²⁴¹. By initiating variants on this gene, the effect of the CRADD gene on apoptosis could be lessened, which will lead to increased cell growth, and this could be a reason for the increased risk of Glioma development. In a separate SNP-SNP analysis in which the interactions of specific SNPs were examined, the number of interactions involving the CRADD gene was the third highest among all the inflammation genes, which suggests that no matter the analysis, whether single SNP based or haplotype based, the CRADD gene is a substantial target for possible Glioma treatment²²⁴. Second, the SNP rs10009998 exists on the IL15 gene which is located on chromosome 4. IL-15 is a cell death regulator that allows for creation of the natural killer 92 cells (NK-92)²⁴². Up-regulation of IL-15 leads to a marked increase in expression of NK-92 cells which can increase the rate of apoptosis²⁴². With increased apoptosis rates, cell growth is substantially lessened, and this could be why variant alleles on the SNP rs10009998 could be crucial in dramatically reducing Glioma risk.

Despite showing very good results with the genetic risk models, there are some limitations with this study. First, there are no covariates with the controls from CGEMS that can be examined for further analysis, or to improve the risk models even more. As stated earlier in this chapter, most of the potential non-genetic risk factors show weak and contradictory evidence for association with Glioma. However, having age and sex as covariates would have allowed for potential absolute risk calculations that could have stated the risk of Glioma for a

determined time period. Second, the CGEMS controls were used initially for prostate and breast cancer and not Glioma. However, the CGEMS controls have been used successfully for previous analysis, including an earlier Glioma study²²⁵, so CGEMS controls have been shown to be viable in genetic research.

This study has shown the effectiveness of jointly using SNPs and haplotypes in a genetic risk model. In a recent study, simulating 41 and 96 genetic variants for breast cancer with odds ratios of 1.3-1.5 are needed to for the discriminatory power to be 0.67 and 0.68 respectively¹²³, and there are questions whether these values are valid, or have been discovered by chance¹²⁴. In this analysis, 25 genetic variants can lead to discriminatory power of 0.654 for Glioma, and also possibly find new areas of association for disease. By using haplotypes and mixing traits of both SNPs and haplotypes, one can maximize the increase in discriminatory power and NRI, and begin to develop genetic risk models that can be crucial in discovering new areas of research for treatment of Glioma.

Chapter 8: Summary and Future Research Directions

8.1. Summary

8.1.1. Summary of Spitz Lung Cancer Validation

In this dissertation thesis, the Spitz lung cancer model was validated with data from the NELSON group in the Netherlands (PI: Dr. Rob J van Klavaran). Discriminatory power results from the NELSON dataset demonstrated that the Spitz model successfully predicted case/control status at 69% (95% CI = 64-75%), which is higher than the internal validation results²⁷. In the Harvard data set⁶⁷, discriminatory power was superior for those who were former smokers compared to current smokers (0.70 vs. 0.68); however, the opposite is true in the NELSON data set (0.74 for current smokers vs. 0.61 for former smokers). When compared to two other lung cancer risk models, the Spitz model's discriminatory power was comparable to both the Liverpool Lung Project's risk model and the Bach lung cancer risk model.

In addition, the results of the calibration of the Spitz model, using the NELSON data, was good. In the lowest risk group and the three highest risk groups, the observed rates of disease were within the bounds of the 95% CI, but in the higher end of the CI range. There could be two reasons for the discrepancy between observed risk for cancer measured by the total numbers of cases divided by the total number of individuals and the 95% CI of the calculated 5 year absolute risk for lung cancer. First, more stringent methods to measure some of the effects of the non-smoking variables, especially emphysema, may be needed. Emphysema has been shown in numerous studies to be a statistically significant factor for lung cancer^{27,93-95}, so if the risk of emphysema is being understated, if patient self-reported emphysema is not clinically validated, or if the discriminatory effect of emphysema is low within a data set, absolute risk values may be biased toward the null for these individuals, leading to less accurate validity results. Second, the

matching of smoking status within the Spitz model could result in lower absolute risks, and hence, the full effect of smoking on lung cancer may not be measured for those with higher lung cancer risks despite the use of adjustment factors to account for smoking status.

8.1.2. Summary of Why We Need to Better Incorporate Haplotypes into Risk

Models

In Chapter 3, the 1st attempt to extend the Spitz lung cancer risk models using haplotypes was conducted using two areas of linkage disequilibrium which was reported by Amos et al⁴⁵. These haplotypes were inferred from SNPs rs2736100 and rs401681 from chromosome 5, and SNPs rs1051730 and rs8034191 on chromosome 15. However, models with inferred haplotypes did not improve the discriminatory power compared to models with top SNPs from the Texas Lung GWAS. Therefore, the question was asked at the end of chapter three: Can modeling and selection of haplotypes be improved such that increased discriminatory power will be achieved in risk models that could be use to select individuals for screening trials or to identify individuals at high-risk for incidence of disease? This would lead to the creation of the BJLM in Chapter 4.

With the BJLM, data from HapMap was utilized to infer haplotypes. The BJLM was designed to seamlessly incorporate the genetic information from all 11 populations in the HapMap data (Table 1.1). Haplotypes were inferred with the use of a Bayesian MCMC model (section 4.2.2) for individuals with full genetic SNP data. Individuals with missing SNP data had their haplotypes inferred using the set of haplotypes determined with full genetic SNP data by the incorporation of a modified forward-backward algorithm (section 4.2.3). Then, these haplotype results were collected assuming additive, dominant, and recessive genetic effects, and were then automatically formatted into WINBUGS code from MATLAB so that the effect of each inferred haplotype (frequency > 0.01) can be estimated. Finally, simulations with

haplotypes ranging from two to twelve SNPs were created with HapSim, a haplotype simulator program written in R¹⁶². Genetic data for these simulations were extracted from both the CEU and TSI populations in HapMap from the base-pair region 11,754,200 and 11,774,700 in chromosome 1. Also, in the simulation analysis, the missing data ranged from 1% to 5% for each SNP to test the robustness of the BJLM. For all haplotype lengths, increasing amounts of missing data did not substantially increase the 95% Credible Intervals of each haplotype inferred, therefore the BJLM is stable for inferring haplotypes of length 2 to 12 SNPs.

8.1.3. Summary of Developing Risk Models with BJLM Inferred Haplotypes

In Chapters 5, 6, and 7, risk models were constructed using haplotypes from the BJLM and haplo.stats with a Hodgkin disease dataset, a Texas Lung GWAS dataset, and a Glioma dataset, respectively. For the Hodgkin data set, discriminatory power was increased for models with haplotypes inferred from the BJLM compared to haplo.stats assuming both additive (AUC = 0.634, 95% CI = 0.569-0.691 vs. AUC = 0.632, 95% CI = 0.567-0.690) and dominant genetic effects (AUC = 0.644, 95% CI = 0.581-0.699 vs. AUC = 0.643, 95% CI = 0.580-0.699), but this increase was not significant for either genetic model (p-value = 0.1562, and p-value = 0.3477 for additive and dominant genetic models, respectively).

In Chapter six, models with haplotypes inferred from the BJLM had significantly increased discriminatory power compared to the original Spitz lung cancer risk model²⁷, and also the model with haplotypes inferred from haplo.stats. In fact, the BJLM classified cases 9% better than haplo.stats, and classified controls 29% better than haplo.stats according to the Net Reclassification Index.

For the Glioma dataset discussed in Chapter 7, to attempt to improve the genetic models by incorporating more genetic information, models with both top SNPs and best Haplotypes

were developed. It was imperative to make sure that there was no inclusion of a top 20 SNP that was also included in the same model as a haplotype containing the most significant SNPs. In this case, all haplotypes that were not inferred with a top 20 significant SNP was added to the model consisting of single SNPs. With models that included top SNPs and top haplotypes inferred from haplo.stats, there was significant improvement in terms of discriminatory power (0.654 vs. 0.637, $p = 0.0112$), and the NRI (NRI = 0.1541, $p\text{-value} = < 0.0001$) compared to models with just SNPs. These results for discriminatory power (0.654 vs. 0.637, $p = 0.0110$) and NRI (NRI = 0.1594, $p\text{-value} = < 0.0001$) are similar for haplotypes inferred with the BJLM.

The improvement with the BJLM compared to other haplotype inferring methods like haplo.stats come from estimating more precise associations in terms of p -values from both haplotypes with frequencies $< 5\%$ and recessive modeled haplotypes. For instance, with haplotype TC on the IL4 gene on Chromosome 5 for the Hodgkin dataset, the overall frequency of the haplotype was less than 5%, and the odds ratio as determined by haplo.stats is 9.24 (95% CI = 1.81-47.2, $p\text{-value} = 0.006$). However, the odds ratio as estimated by the BJLM was 12.16 (95% CI = 2.47-90.1, $p\text{-value} = 0.00044$). For haplotypes inferred assuming recessive traits, p -values were more significant with haplotypes inferred using BJLM compared to haplo.stats in all five instances where recessive genetic effects were modeled in the Texas Lung GWAS and Glioma datasets. The better determination of haplotype associations assuming recessive genetic effects was especially important for extending the Spitz former smoker model using Texas Lung GWAS data as haplotype CGAC on the IREB2 gene from Chromosome 15 was significantly associated with lung cancer with the BJLM (OR = 9.85, 95% CI = 1.28-218.1, $p = 0.0250$), but not in haplo.stats (OR = 6.56, 95% CI = 0.787-54.66, $p\text{-value} = 0.0857$). Improvement with estimating association with low-frequency haplotypes and recessive haplotypes in the BJLM is

due directly to the use of genetic information from HapMap. With the genetic information from HapMap, the control haplotypes assessed in the population being studied are better inferred, so the differences in cases can be more precisely extracted compared to programs that use no outside information to infer their haplotypes, like haplo.stats.

8.2. Future Research Directions

In this dissertation, a novel Bayesian method was constructed to incorporate genetic information from HapMap to estimate haplotypes to study any type of disease. This involved a binary outcome of either disease or no disease as the dependent variable, and the inferred haplotypes and other covariates as the independent variables. Data sets that were examined consisted of candidate genes (Hodgkin), inflammation pathways (Glioma), or a selected sub-set of SNPs acquired from a meta-analysis (Lung). Therefore, three potential areas of future analysis would be to incorporate genetic information from sources other than HapMap, extend the dependent variable to examine non-binary outcomes, and to extend the analysis to full genome wide association studies.

8.2.1. Incorporate 1000 Genomes as External Data into the BJLM

The goal of the 1000 genomes project is to attempt to find more common variants that have minor allele frequency of 1% by the process of gene sequencing, which breaks a subject's DNA into smaller sections. In the 1000 genomes project, sequencing was first conducted four times each for each subject (4x). Then, for 1000 specific regions of the genome, sequencing was conducted at a much higher rate (50x). Data for the 1000 genome project is located on the 1000 genome website (<http://browser.1000genomes.org/index.html>) and can be extracted into Haploview ready format consistently for a small base pair range (10000-25000 bp). The 1000 genome data has some advantages over HapMap data in that there are actually more SNPs

available for genetic analysis; however, it generally has less samples per population than HapMap data because of the expense that was involved with sequencing.

To incorporate 1000 genome data seamlessly into the BJLM, the genetic SNP data would have to be filtered similarly to the data from HapMap. This process could be done in a spreadsheet relatively easily, but would need to be coded for use by the BJLM. The process needed to extract the 1000 genomes data from the 1000 genomes website for a specific population is shown schematically (Figure 8.1).

Figure 8.1: Process for Extracting 1000 Genomes Data for Future Analysis

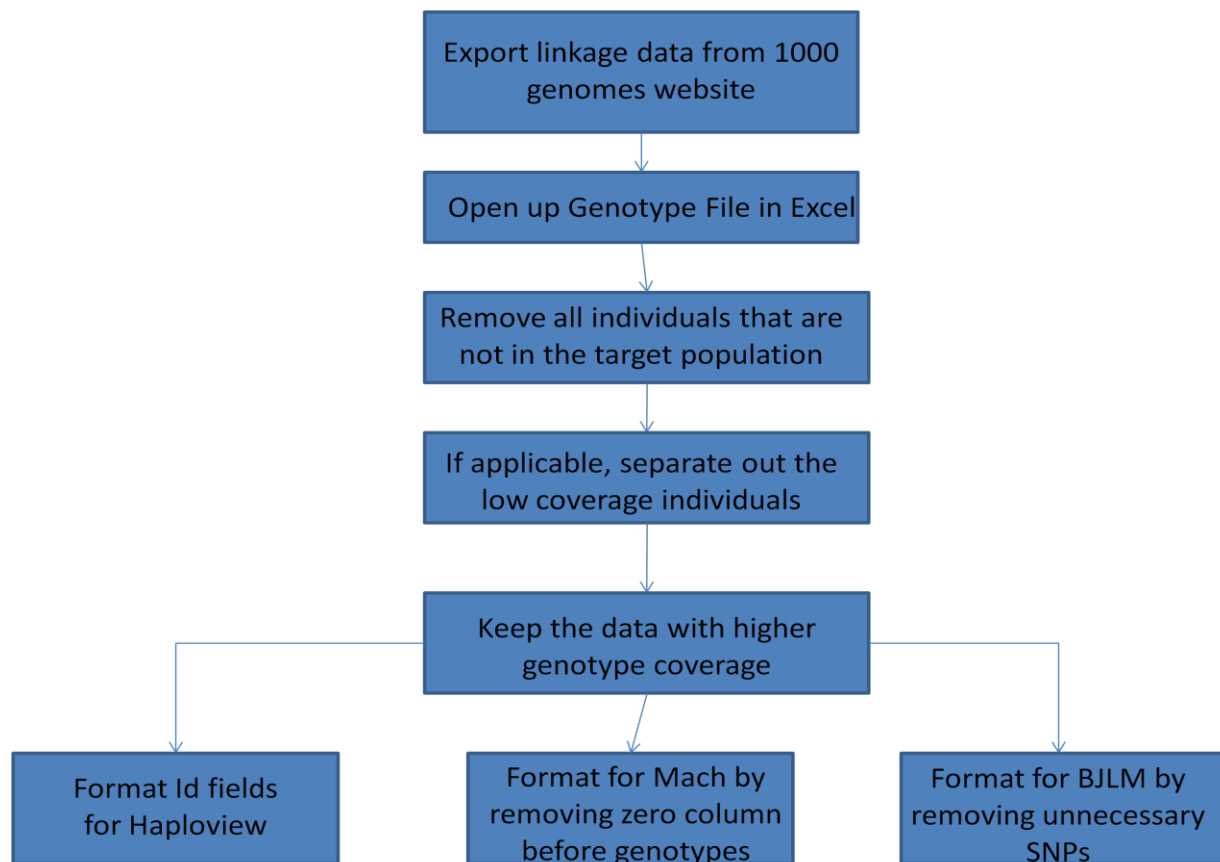


Figure 8.1: With 1000 genome data, genetic information from all populations studied in the 1000 genomes project is loaded as a genotype file for further examination. Therefore, only those individuals that occur in the same population as the dataset to be examined would have to be extracted from the 1000 genome database. For example, if studying the Texas Lung GWAS population from Chapter 6, only Caucasians from the 1000 genomes data can be used to make inferences on the genetic data, and this occurs in the 1000 genome data from those with id numbers NA06984 to NA12890. Also, in the 1000 genome data, multiple sets of genotype information are available for the same individual, and this has to be filtered out so that only one set of genotypes exist for one individual.

All of this could be conducted with a spreadsheet program, but for perfect integration into the BJLM, code would have to be developed in MATLAB, or a similar language, to perform this task automatically.

8.2.2. Incorporate Non-Binary Outcomes

In this dissertation, the dependent variable was disease status; however, there are many cases in which the dependent variable may be non-binary. For example, one may want to test the association between genetic information like SNPs and haplotypes and inflammatory biomarkers. For lung cancer, an important host susceptibility biomarker of increasing importance is the smokers' sensitivity to the nicotine-derived nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), which is a strong pulmonary carcinogen²⁴³⁻²⁴⁶. Identifying important genes involved in the metabolism and/or repair of damage induced by such a potent carcinogen may allow better identification of high risk individuals. With programs like WINBUGS, models can be constructed that can examine the association between important areas of the genome for lung cancer susceptibility and protectively and the mechanism of NNK carcinogenicity. Better elucidation of the relationship between haplotypes and the NNK could lead to better drug targets to treat lung damage.

8.2.3. Extend Analysis to Full GWAS studies and Final Conclusion

In two of the three analyses conducted with the BJLM, genetic data was obtained from sub-sections of large GWAS studies, whether that was for the extension of the Spitz lung cancer risk models with haplotypes or for the creation of a Glioma risk model. The Spitz lung cancer risk model was extended with haplotypes, and this led to higher discriminatory power increases than with previously published extension of the Spitz lung cancer risk model with just SNPs³⁰. Also, with just 15 SNPs and 10 haplotypes, a purely genetic Glioma model was created from genetic information in the inflammation pathway, and discriminatory power was 65.4%. Furthermore, this model had significantly increased discriminatory power compared to a model with just top SNPs from the inflammation pathway (0.654 vs. 0.637, $p = 0.0110$). These

promising results suggest that genetic models that incorporate the information from a full GWAS analysis will lead to even better modeling of disease. Therefore, I propose conducting analysis throughout the genome to find the absolute best haplotype and SNP associations with lung cancer and Glioma (p-value of 10^{-7} or less only), and developing risk models that could be used in cancer prevention studies, screening trials, personalized medicine clinics, or even drug development. Finding the genetic associations that lead to better modeling of disease causation will lead to better treatment of disease, and the BJLM may become a vital instrument in finding these important genetic associations with disease.

REFERENCES

1. Ridker PM, Buring JE, Rifai N, Cook NR (2007) Development and Validation of Improved Algorithms for the Assessment of Global Cardiovascular Risk in Women. *JAMA* **297**:611-619
2. Framingham Health Study. <http://www.framinghamheartstudy.org/> (Accessed January 10, 2011)
3. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* **97**:1837-1847
4. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, Sr., Newton-Cheh C, Yamamoto JF, Magnani JW, Tadros TM, Kannel WB, Wang TJ, Ellinor PT, Wolf PA, Vasan RS, Benjamin EJ (2009) Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort. *Lancet* **373**:739-745
5. Kannel WB, D'Agostino RB, Silbershatz H, Belanger AJ, Wilson PW, Levy D (1999) Profile for estimating risk of heart failure. *Arch Intern Med* **159**:197-204
6. D'Agostino Sr. RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB (2008) General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **117**:743-753
7. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (2001) Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA* **285**:2486-2497

8. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB (1994) Stroke risk profile: adjustment for antihypertensive medication: The Framingham Study. *Stroke* **25**:40-43
9. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Shairer C, Mulvihill JJ (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* **81**:1879-1886
10. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA (2001) Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *J Natl Cancer Inst* **93**:358-66.
11. Hoskins KF, Zwaagstra A, Ranz M (2006) Validation of a tool for identifying women at high risk for hereditary breast cancer in population-based screening. *CANCER* **107**: 1769-1776.
12. Capasso I, Esposito E, Montella M, Crispo A, Grimaldi M, D'Aiuto M, Beneduce G, Esposito G, D'Aiuto G (2009) Gail's model as first step for early diagnosis: National Cancer Institute of Naples experience. *Breast Cancer Research* **11(Suppl 1)**: S9 (doi:10.1186/bcr2270)
13. Gail MH, Costantino JP, Pee D, Bondy M, Newman L, Selvan M, Anderson GL, Malone KE, Marchbanks PA, McCaskill-Stevens W, Norman SA, Simon MS, Spirtas R, Ursin G, and Bernstein L (2007) Projecting Individualized Absolute Invasive Breast Cancer Risk in African American Women. *J Natl Cancer Inst* **99**:1782-1792.
14. Tice JA, Cummings SR, Ziv E, Kerlikowske K (2005) Mammographic breast density and the Gail model for breast cancer risk prediction in a Screening Population. *Breast Cancer Res Treat* **94**:115–122

15. Imperiale T, Wagner D, Lin CY, Larkin GN, Rogge JD, Ransohoff DF (2000) Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal cancer findings. *N Engl J Med* **343**:169-174
16. Selvachandran SN, Hodder RJ, Ballal MS, Jones P, Cade D (2002) Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study. *Lancet* **360**:278-283
17. Wu X, Lin J, Grossman HB, Huang M, Gu J, Etzel CJ, Amos CI, Dinney CP, Spitz MR (2007) Projecting Individualized Probabilities of Developing Bladder Cancer in White Individuals. *J Clin Oncol* **25**:4974-4981
18. Hartge P, Whittemore AS, Itnyre J, McGowan L, Cramer D (1994) Rates and risks of ovarian cancer in subgroups of white women in the United States. *Obstet Gynecol* **84**:760-764
19. Cho E, Rosner BA, Feskanich D, Colditz GA (2005) Risk factors and individual probabilities of melanoma for whites. *J Clin Oncol* **23**:2669-2675
20. Fears TR, Guerry D 4th, Pfeiffer RM, Sagebiel RW, Elder DE, Halpern A, Holly EA, Hartge P, Tucker MA (2006) Identifying individuals at high risk of melanoma: A practical predictor of absolute risk. *J Clin Oncol* **24**:3590-3596
21. Colditz GA, Atwood KA, Emmons K, Monson RR, Willett WC, Trichopoulos D, Hunter DJ (2000) Harvard Report on Cancer Prevention. Volume 4: Harvard Cancer Risk Index Working Group, Harvard Center for Cancer Prevention. *Cancer Causes Control* **11**:477-488
22. Surveillance and End Results (SEER) Lung Cancer Incidence for Surveillance and End Results (SEER). U.S. National Institutes for Health. <http://www.seer.cancer.gov> [Last accessed: August 10,2011]

23. Omenn GS, Goodman G, Thornquist M, Grizzle J, Rosenstock L, Barnhart S, et al (1994)
The beta-carotene and retinol efficacy trial (CARET) for chemoprevention of lung cancer in high risk populations: smokers and asbestosexposed workers. *Cancer Res* **54**:2038s–2043s
24. Omenn GS, Goodman GE, Thornquist MD, Balmes J, Cullen MR, Glass A, et al (1996)
Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *New Engl J Med* **334**:1150–1155
25. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, Hsieh LJ, Begg CB (2003) Variations in Lung Cancer Risk Among Smokers. *J Natl Cancer Inst* **95**: 470-478
26. Cronin KA, Gail MH, Zou Z, Bach PB, Virtamo J, Albanes D (2006) Validation of a model of lung cancer risk prediction among smokers. *J Natl Cancer Inst* **98**:637-640
27. Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Qiong D, Shete S, Etzel CJ (2007) A Risk Model for Prediction of Lung Cancer. *J Natl Cancer Inst* **99**:715-726
28. National Center for Health Statistics (2003) Worktable 210R. Death rates for 113 selected causes, alcohol-induced causes, drug-induced causes and injury by firearms, by 5-year age groups, race, and sex.
29. Etzel CJ, Kachroo S, Liu M, D'Amelio A, Dong Q, Cote ML, Wenzlaff AS, Hong WK, Greisinger AJ, Schwartz AG, Spitz MR (2008) Development and validation of a lung cancer risk prediction model for African-Americans *Cancer Prev Res* **1**:255-265
30. Spitz MR, Amos CI, D'Amelio A Jr, Dong Q, Etzel C (2009) Re: Discriminatory Accuracy from Single-Nucleotide Polymorphisms in Models to Predict Breast Cancer Risk. *J Natl Cancer Inst* **101**:1731-1732.

31. Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, Field JK (2008)
The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer*
98:270-276
32. Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, Riley TL,
Commins J, Oken MM, Berg CD, Prorok PC (2011) Lung Cancer Risk Prediction: Prostate,
Lung, Colorectal, and Ovarian Cancer Screening Trial Models and Validation. *J Natl Cancer*
Inst **103**:1058-1068
33. Marrie RA, Dawson NV, Garland A (2009) Quantile regression and restricted cubic splines
are useful for exploring relationships between continuous variables. *Journal of Clinical*
Epidemiology **62**: 511-517
34. Harrell Jr. FE (2001) Regression modeling strategies: With Applications to Linear Models,
Logistic Regression, and Survival Analysis. New York: Springer
35. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating
curve (ROC) curve. *Radiology* **143**:29-36
36. Altman DG, Bland JM (1994) Statistics Notes: Diagnostic tests 2: predictive values. *BMJ*
309:102
37. Harrell FE, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing
models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*
15:361–387
38. Efron B, Gong G (1983) A leisurely look at the bootstrap, the jack knife, and cross-
validation. *Am Stat* **37**:36–48
39. Schumacher M, Hollander N, Sauerbrei W (1997) Resampling and crossvalidation
techniques: a tool to reduce bias caused by model building? *Stat Med* **16**:2813–2827

40. Turnpenny P, Ellard S (2005) Emery's Elements of Medical Genetics, 12th. ed. Elsevier, London.
41. Santana QC, Coetzee MPA, Steenkamp ET, Mlonyeni GNA, Wingfield MJ, Wingfield BD (2009) *Bio Techniques* **46**:217-223
42. The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**:928-933
43. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**:1061-1073
44. Thorisson GA, Smith AV, Krishnan L, Stein LD (2005) The International HapMap Project Web Site. *Genome Res* **15**:1592-1593
45. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai YY, Chen WV, Shete S, Spitz MR, Houlston RS (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1 *Nat Genet* **40**:616-622
46. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al (2008) A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**:633-637
47. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al (2008) A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**:638-642
48. International Human Genome Sequencing Consortium (2004) Fishing the euchromatic sequence of the human genome. *Nature* **431**:931-945

49. Kim KN, Lee IK, Kim YK, Tran HT, Yang DH, Lee JJ, Shin MH, Park KS, Shin MG, Choi JS, Kim HJ (2008) Association between folate-metabolizing pathway polymorphism and non-Hodgkin lymphoma. *Br J Haematol.* **140**:287-294.
50. Frosst P, Blom HJ, Milos R, Goyette P, Sheppard CA, Matthews RG, et al (1995) A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nat Genet.* **10**:111–113
51. Weir BS (1996) Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sinauer Associates Inc., Publishers. Sunderland, Massachusetts.
52. Greenwood TA, Rana BK, Schork NJ (2004) Human Haplotype Block Sizes are Negatively Correlated with Recombination Rates. *Genome Res* **14**:1358-1361
53. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. (2002) The Structure of Haplotype Blocks in the Human Genome. *Science* **296**:2225-2229
54. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. (2002) A high-resolution recombination map of the human genome. *Nat. Genet* **31**:241-247
55. Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing data imputation. *American Journal of Human Genetics* **76**:449-462
56. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**:263-265
57. Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, et al (2000) SNPping away at complex diseases. *Am J Hum Genet* **67**:383-394.
58. Escamilla MA, McInnes LA, Spesny M, Reus VI, Service SK, Shimayoshi N, et al. (1999) Assessing the feasibility of linkage disequilibrium methods for mapping complex traits: an initial screen for bipolar disorder loci on chromosome 18. *Am J Hum Genet* **64**:1670-1678.

59. Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics* **9**:291-300
60. Glossary of Nature Review Genetics. October 2002. Volume 3. No. 10
61. Stephens M, Smith N, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*. **68**:978-989
62. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet*. **70**:425-434
63. Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ (2003) Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered*. **55**:56-65
64. Lin DY, Deng K, Millikan R (2005) Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genet Epidemiol*. **29**:299-312
65. Zeng D, Lin DY (2005) Estimating haplotype-disease associations with pooled genotype data. *Genet Epidemiol*. **28**:70-82
66. Zeng D, Lin DY, Avery CL, North KE, Bray MS (2006) Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics* **7**:486-502
67. D'Amelio Jr. AM, Cassidy A, Asomaning K, Raji OY, Duffy SW, Field JK, Spitz MR, Christiani D, Etzel CJ (2010). Comparison of Discriminatory Power and Accuracy of Three Lung Cancer Risk Models. *Br J Cancer*. **103**:423-429
68. National Lung Screening Trial Research Team (2011) The National Lung Cancer Screening Trial: Overview and Study Trial. *Radiology* **258**:243-253

69. Habbema JD, van Oortmarssen GJ, Lubbe JT, van der Maas PJ (1985) The MISCAN simulation program for the evaluation of screening for disease. *Comput Methods Programs Biomed* **20**:79-93
70. Loeve F, Boer R, van Oortmarssen GJ, van Ballegooijen M, Habbema JD (1999) The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. *Comput Biomed Res* **32**:13-33
71. Loeve F, Brown ML, Boer R, van Ballegooijen M, van Oortmarssen GJ, Habbema JD (2000) Endoscopic colorectal cancer screening: a cost-saving analysis. *J Natl Cancer Inst.* **92**:557–563.
72. Vijan S, Hwang I, Inadomi J, Wong RK, Choi JR, Napierkowski J, Koff JM, Pickhardt PJ (2007) The cost-effectiveness of CT colonography in screening for colorectal neoplasia. *Am J Gastroenterol* **102**:380-390.
73. Samet, JM, Wiggins CL, Humble CG, Pathak DR (1988). Cigarette smoking and lung cancer in New Mexico. *American Review of Respiratory Disease* **137**:1110–1113.
74. Villeneuve, PJ, Mao Y (1994). Lifetime probability of developing lung cancer, by smoking status, Canada. *Canadian Journal of Public Health* **85**:385–388.
75. Duffy SW, Raji OY, Agbaje OF, Allgood PC, Cassidy A, Field JK (2009) Use of lung cancer risk models in planning research and service programmes in CT screening for lung cancer. *Expert Rev Anticancer Ther* **9**:1467-1472
76. Vickers AJ, Kramer BS, Baker SG (2006) Selecting patients for randomized trials: a systematic approach. *Trials* **7**:30, doi:10.1186/1745-6215-7-30
77. Cassidy A, Duffy SW, Myles JP, Liloglou T, Field JK (2007) Lung cancer risk prediction: a tool for early detection. *Int J Cancer* **120**:1–6

78. van Iersel CA, de Koning HJ, Draisma G, Mali WPTM, Scholten ET, Nackaerts K, Prokop M, Habbema JDikF, Oudkerk M, van Klaveren RJ (2006) Risk-based selection from the general population in a screening trial: Selection criteria, recruitment and power for the Dutch-Belgian randomized lung cancer multi-slice CT screening trial (NELSON). *Int J. Cancer* **120**:868-874
79. van Klaveren RJ, Oudkerk M, Prokop M, Scholten ET, Nackaerts K, Vernhout R, van Iersel CA, van den Bergh KAM, van't Westeinde S, et al (2009) Management of Lung Nodules Detected by Volume CT Scanning. *N Engl J Med.* **361**:2221-2229
80. Baecke E, de Koning HJ, Otto SJ, van Iersel CA, van Klaveren RJ (2010) Limited contamination in the Dutch-Belgian randomized lung cancer screening trial (NELSON). *Lung Cancer.* **69**:66-70
81. van den Bergh KAM, Essink-Bot ML, Borsboom GJJM, Scholten ET, Prokop M, de Koning HJ, van Klaveren RJ (2010) Short-term health-related quality of life consequences in a lung cancer CT screening trial (NELSON) *Br J Cancer.* **102**:27-34
82. van den Bergh KAM, Essink-Bot ML, Bunge EM, Scholten ET, Prokop M, van Iersel CA, van Klaveren RJ, de Koning HJ (2008) Impact of Computed Tomography Screening for Lung Cancer on Participants in a Randomized Controlled Trial (NELSON Trial). *Cancer* **113**:396-404.
83. van den Bergh KAM, Essink-Bot ML, van Klaveren RJ, de Koning HJ (2009) Informed participation in a randomized controlled trial of computed tomography screening for lung cancer. *Eur Respir J* **34**:711-720
84. Black WC, Baron JA (2007). CT Screening for Lung Cancer: Spiraling into Confusion? *JAMA.* **297**:995-997.

85. Association of Comprehensive Cancer Centres, the Netherlands. Netherlands Cancer Registry. Table A4a: Age specific incidence rates (per 100,000) for invasive tumours in males according to site in 2007 and Table A4b: Age specific incidence rates (per 100,000) for invasive tumours in females according to site in 2007.
http://www.ikcnet.nl/page.php?id=1522&nav_id=97 [Last Accessed: July 15th, 2010]
86. Association of Comprehensive Cancer Centres, the Netherlands. Netherlands Cancer Registry. Table B4a: Age specific mortality rates per 100,000 in males according to site in 2007 and Table B4b: Age specific mortality rates per 100,000 in females according to site in 2007. http://www.ikcnet.nl/page.php?id=1523&nav_id=97 [Last Accessed: July 15th, 2010]
87. Hogervorst JGF, Schouten LJ, Konings EJM, Goldbohm RA, van den Brandt PA (2009) Lung Cancer Risk in Relation to Dietary Acrylamide Intake. *J Natl Cancer Inst* **101**:651-662
88. European Health for All Database. World Health Organization: Regional Office for Europe. Update: January 2011. <http://data.euro.who.int/hfad/> [Last Accessed: January 30, 2011]
89. Agyemang C, Stronks K, Tromp N, Bhopal R, Zaninotto P, Unwin N, Nazroo J, Kunst AE (2010) A cross-national comparative study of smoking prevalence and cessation between English and Dutch South Asian and African origin populations: the role of national context. *Nicotine and Tobacco Research*. **12**:557-566
90. Dupont WD, Plummer WD Jr (1996) Understanding the relationship between relative and absolute risk. *Cancer* **77**:2193-2199
91. Marshall RJ (2001) Displaying categorical data relationships by scaled rectangle diagrams. *Stat Med* **20**:1077-1088
92. Marshall RJ (2005) Scaled rectangle diagrams can be used to visualize clinical and epidemiological data. *J Clin Epidemiol* **58**:974-981

93. Mayne ST, Buenconsejo J, Janerich DT (1999) Previous lung disease and risk of lung cancer among men and women nonsmokers. *Am J Epidemiol* **149**:13-20
94. Brenner AV, Wang Z, Kleinerman RA, Wang L, Zhang S, Metayer C, et al (2001) Previous pulmonary diseases and risk of lung cancer in Gansu Province, China. *Int J Epidemiol* **1**:118-124.
95. Schabath MB, Delclos GL , Martynowicz MM , Greisinger AJ , Lu C , Wu X , et al (2005) Opposing effects of emphysema, hay fever, and select genetic variants for lung cancer risk . *Am J Epidemiol.* **161**:412-422
96. Garcia-Closas M, Kelsey KT, Wiencke JK, Xu X, Wain JC, Christiani DC (1997) A case–control study of cytochrome P450 1A1, glutathione S-transferase M1, cigarette smoking and lung cancer susceptibility *Cancer Causes Control* **8**:544-553
97. Xu X, Kelsey KT, Wiencke JK, Wain JC, Christiani DC (1996) Cytochrome P450 CYP1A1 *MspI* polymorphism and lung cancer susceptibility. *Cancer Epidemiol Biomarkers Prev* **5**:687-692
98. Wang LI, Miller DP, Sai Y, Liu G, Su L, Wain JC, Lynch TJ, Christiani DC (2001) Manganese Superoxide Dismutase Alanine-to-Valine Polymorphism at Codon 16 and Lung Cancer Risk. *J Natl Cancer Inst* **93**:1818-1821
99. Asomaning K, Miller DP, Liu G, Wain JC, Lynch TJ, Su L, Christiani DC (2008) Second hand smoke, age of exposure and lung cancer risk. *Lung Cancer* **61**:13-20
100. Miller DP, Neuberg D, de Vivo I, Wain JC, Lynch TJ, Su L, Christiani DC (2003) Smoking and the Risk of Lung Cancer: Susceptibility with GSTP1 Polymorphisms. *Epidemiology* **14**:545-551

101. Greenland S, Lash TL (2008). Bias analysis. In Modern Epidemiology, 3rd ed. Rothman KJ, Greenland S and Lash TL (eds) pp 345–380. Lippincott–Williams–Wilkins: Philadelphia.
102. Tyrer J, Duffy SW, Cuzick J (2004) A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med* **23**:1111-1130
103. Rockhill B, Byrne C, Rosner B, Louie MM, Colditz G (2003) Breast cancer risk prediction with a log-incidence model: evaluation of accuracy. *J. Clin. Epidemiol* **56**:856-861
104. Antoniou AC, Hardy R, Walker L, et al (2008) Predicting the likelihood of carrying a BRCA1 or BRCA2 mutation: validation of BOADICEA, BRCAPRO, IBIS, Myriad and the Manchester scoring system using data from UK genetics clinics. *J Med Genet.* **45**:425–431.
105. Gail MH, Mai PL (2010) Comparing Breast Cancer Risk Assessment Models. *J Natl Cancer Inst.* **102**:665-668
106. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM (2006) Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* **8**:395-400
107. Janssens AC, Moonesinghe R, Yang Q, Steyerberg EW, van Duijn CM, Khoury MJ (2007) The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genet Med* **9**:528-535
108. Kraft P, Hunter DJ (2009) Genetic risk prediction--are we there yet? *N Engl J Med* **360**:1701-1703
109. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet* **5**:e1000337, doi:10.1371/journal.pgen.1000337

110. Johnson AE, Gordon C, Palmer RG, Bacon PA (1995) The prevalence and incidence of SLE in Birmingham, England. Relationship to ethnicity and country of birth. *Arthritis Rheum* **38**:551-558.
111. Boey ML. Systemic lupus erythematosus (1992) *Singapore Med J* **33**:291-293.
112. Loong T-W (2003) Understanding sensitivity and specificity with the right side of the brain. *BMJ*. **327**:716-719
113. Parkin DM, Bray F, Ferlay J, Pisani P (2005) Global cancer statistics, 2002. *CA Cancer J Clin* **55**:74-108
114. Danaei G, Vander Hoorn S, Lopez AD, Murray CJL, Ezzati M (2005) Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *Lancet* **366**:1784-1793
115. McBride CM, Bepler G, Lipkus IM, Lyna P, Samsa G, Albright J, Datta S, Rimer BK (2002) Incorporating Genetic Susceptibility Feedback into a Smoking Cessation Program for African-American Smokers with Low Income. *Cancer Epidemiol Biomarkers Prev* **11**:521-528
116. Carpenter MJ, Strange C, Jones Y, Dickson MR, Carter C, Moseley MA, Gilbert GE (2007) Does genetic testing result in behavior health change? Change in smoking behavior following testing for alpha-1 antitrypsin deficiency. *Ann Behav Med* **33**:22-28
117. Young RP, Whittington CF, Hopkins RJ, Hay BA, Epton MJ, Black PN, Gamble GD (2010) Chromosome 4q31 locus in COPD is also associated with lung cancer. *Eur Respir J* **36**:1375-1382

118. Gail MH, Pfeiffer RM, Wheeler W, Pee D (2008) Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. *Biostatistics*. **9**:201–215
119. Gail MH (2008) Discriminatory Accuracy From Single-Nucleotide Polymorphisms in Models to Predict Breast Cancer Risk. *J Natl Cancer Inst* **100**:1037-1041
120. Pharoah PDP, Antoniou AC, Easton DF, Ponder BAJ (2008) Polygenes, Risk Prediction, and Targeted Prevention of Breast Cancer. *N Engl J Med*. **358**:2796-2803.
121. Stacey SN, Manolescu A, Sulem P, Rafner T, Gudmundsson J, Gudjonsson SA, et al. (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet*. **39**:865-869
122. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*. **447**:1087-1095.
123. van Zitteren M, van der Net JB, Kundu S, Freedman AN, van Duijn CM, Janssens CJW (2011) Genome-Based Prediction of Breast Cancer Risk in the General Population: A Modeling Study Based on Meta-Analyses of Genetic Associations. *Cancer Epidemiol Biomarkers Prev*. **20**:9-22
124. Chatterjee N, Park J-H, Caporaso N, Gail MH (2011) Predicting the Future of Genetic Risk Prediction. *Cancer Epidemiol Biomarkers Prev*. **20**:3-8
125. Pharoah PDP, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet*. **31**:33-36.

126. Spitz MR, Etzel CJ, Dong Q, Amos CI, Wei Q, Wu X, Hong WK (2008) An Expanded Risk Prediction Model for Lung Cancer. *Cancer Prev Res.* **1**:250-254
127. McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, et al (2008) Lung cancer susceptibility locus at 5p15.33. *Nat Genet.* **40**:1404-1406.
128. Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, et al (2008) Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet.* **40**:1407-1409
129. Kang JU, Koo SH, Kwon KC, Park JW, Kim JM (2008) Gain at chromosomal region 5p15.33, containing TERT, is the most frequent genetic event in early stages of non-small cell lung cancer. *Cancer Genet Cytogenet.* **182**: 1-11
130. Pencina MJ, D'Agostino Sr. RB, D'Agostino Jr, RB, Vasan RS (2008) Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statist Med.* **27**:157-172
131. Messaoudi N, De Cocker J, Stockman BA, Bossaert LL, Rodrigus IER (2009) Is EuroSCORE useful in the prediction of extended intensive care unit stay after cardiac surgery? *European Journal of Cardio-thoracic Surgery.* **36**:35-39
132. Judson R, Stephens JC (2001) Notes from the SNP vs. haplotype front. *Pharmacogenomics.* **2**:7-10.
133. El-Zein R, Monroy CM, Etzel CJ, Cortes AC, Xing Y, Collier AL, Strom SS (2009) Genetic Polymorphisms in DNA Repair Genes as Modulators of Hodgkin Disease Risk Cancer. **115**:1651-1659.

134. Yaspan BL, Breyer JP, Cai Q, Elmore JB, Amundson I, Bradley KM, Shu X-O, Gao Y-T, Dupont WD, Zeing W, Smith JR (2007) Haplotype Analysis of CYP11A1 Identifies Promoter Variants Associated with Breast Cancer Risk. *Cancer Res.* **67**:5673-5682
135. Naccarati A, Pardini B, Polakova V, Smerhovsky Z, Vodickova L, Soucek P, Vrana D, Holcatova I, Ryska M, Vodicka P (2010) Genotype and haplotype analysis of TP53 gene and the risk of pancreatic cancer: an association study in the Czech Republic. *Carcinogenesis.* **31**:666-670.
136. The International HapMap Consortium (2003) The International HapMap Project. *Nature.* **18**:789–796.
137. Smith EM, Wang X, Littrell J, Eckert J, Cole R, Kissebah AH, Olivier M. (2006) Comparison of linkage disequilibrium patterns between the HapMap CEPH samples and a family-based cohort of Northern European descent. *Genomics.* **88**:407-414.
138. Neale BM (2010) Introduction to Linkage Disequilibrium, the HapMap, and Imputation. *Cold Spring Harbor Protocol.* doi:10.1101/pdb.top74
139. Andiappan AK, Anantharaman R, Nilkanth PP, Wang DY, Chew FT (2010) Evaluating the transferability of Hapmap SNPs to a Singapore Chinese population. *BMC Genetics.* **11**:36 doi: 10.1186/1471-2156-11-36
140. Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet.* **76**:887-893.
141. de Bakker PIW, Yelensky R, Pe'er, I, Gabriel SB, Daly MJ, Altshuler D (2005) Efficiency and power in genetic association studies. *Nat Genet.* **37**:1217-1223.

142. Chen YH, Chatterjee N, Carroll RJ (2008) Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. *Biostatistics*. **9**:81-99
143. Xing EP, Jordan MI, Sharan R (2007) Bayesian Haplotype Inference via the Dirichlet Process. *J Comput Biol*. **14**:267-284
144. Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms. *Am J Hum Genet*. **70**:157-169
145. Bansal V, Halpern AL, Axelrod N, Bafna V (2008) An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Res*. **18**:1336-1346.
146. Gilks WR, Richardson S, Spiegelhalter DJ (Eds.) (1996) Markov chain Monte Carlo in Practice. Chapman & Hill. London, England
147. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6**:721-741.
148. Rabiner LR (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*. **77**:257-286.
149. Baker JK (1975) The dragon system – An overview. *Acoust. Speech Signal Processing*. **ASSP-23**:24-29
150. Jelinek F (1969) A fast sequential decoding algorithm using a stack. *IBM J. Res. Develop*. **13**:675-685
151. Schutter EM, Sohn C, Kristen P, Möbus V, Crombach G, Kaufmann M, Caffier Hm Kreienberg R, Verstraeten AA, Kenemans P (1998) Estimation of probability of malignancy using a logistic model combining physical examination, ultrasound, serum CA 125, and

- serum CA 72-4 in postmenopausal women with a pelvic mass: an international multicenter study. *Gynecol Oncol.* **69**:56-63
152. Virtanen A, Gomari M, Kranse R, Stenman UH (1999) Estimation of prostate cancer probability by logistic regression: free and total prostate-specific antigen, digital rectal examination, and heredity are significant variables. *Clin Chem.* **45**:987-994
 153. Genkin A, Lewis DD, Madigan D (2007) Large-scale Bayesian logistic regression for text categorization. *Technometrics.* **49**:291–304.
 154. Fahrmeir L, Lang S (2001) Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Journal of the Royal Statistical Society C (Applied Statistics).* **50**:201-220.
 155. Tutz G (2004) Generalized semiparametrically structured mixed models. *Comput Statist Data Anal.* **46**:777-800.
 156. Fahrmeir L, Kneib Th, Lang S (2004) Penalized additive regression for space-time data: a Bayesian perspective. *Statistica Sinica.* **14**:731-761.
 157. Dey D, Ghosh SK, Mallick BK (eds.) (1999) Generalized Linear Models: A Bayesian Perspective. New York: Marcel Dekker.
 158. Albert J, Chib S (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association.* **88**:669-679.
 159. Holmes CC, Held L (2006) Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis.* **1**:145-168.
 160. Kazembe LN, Chirwa TF, Simbeye JS, Namangale JJ (2008) Applications of Bayesian approach in modeling risk of malaria-related hospital mortality. *BMC Med Res Methodol* **8**:6
doi:10.1186/1471-2288-8-6

161. Ntzoufras I (2009) (ed) Bayesian Modeling Using WinBUGS. John Wiley & Sons.
Hoboken, New Jersey
162. Montana G (2005) HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics*. **21**:4309-4311.
163. Excoffier L, Novembre J, Schneider S (2000) SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J Hered*. **91**:506–509.
164. Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*. **18**:337–338.
165. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, et al (2009) A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome 5p15 associated with Risk for Adenocarcinoma. *Am J Hum Genet*. **85**:679-691.
166. American Cancer Society. *Cancer Facts & Figures 2010*. Atlanta: American Cancer Society. (2010). URL:
<http://www.cancer.org/acs/groups/content/@epidemiologysurveillance/documents/document/acspc-026238.pdf> _[Accessed February 17th, 2011].
167. Horner MJ, Ries LAG, Krapcho M, Neyman N, Aminou R, Howlander N, Altekruse SF, Feuer EJ, Huang L, Mariotto A, Miller BA, Lewis DR, Eisner MP, Stinchcomb DG, Edwards BK (eds) (2009) SEER Cancer Statistics Review, 1975-2006, National Cancer Institute. Bethesda, MD, [http://seer.cancer.gov/csr/1975_2006/], based on November 2008 SEER data submission, posted to the SEER web site [Last Accessed: September 9th, 2009]

168. The Lymphomas: Hodgkin Lymphoma and Non-Hodgkin Lymphoma, URL:
http://www.leukemia-lymphoma.org/attachments/National/br_1161891669.pdf [Last
Accessed: September 9th, 2009].
169. Diehl V, Tesch H (1996) Hodgkin's Disease – Environment or Genetic? *N Engl J Med.*
332:461-462.
170. Kamel OW, van de Rijn M, Weiss LM, Del Zoppo GJ, Hench PK, Robbins BA,
Montgomery PG, Warnke RA, Dorfman RF (1993) Reversible lymphomas associated with
Epstein-Barr virus occurring during methotrexate therapy for rheumatoid arthritis and
dermatomyositis *N Engl J Med.* **328**:1317-1321.
171. Berwick M, Vineis P (2000) Markers of DNA repair and susceptibility to cancer in
humans: an epidemiologic review. *J Natl Cancer Inst.* **92**:874-897.
172. Shen M, Purdue MP, Krickler A (2007) Polymorphisms in DNA repair genes and risk of
non-Hodgkin's lymphoma in New South Wales, Australia. *Haematologica.* **92**:1180-1185
173. Pakakasama S, Sirirat T, Kanchanachumpol S, Udomsubpayakul U, Mahasirimongkol S,
Kitpoka P, Thithapandha A, Hongeng S (2007) Genetic polymorphisms and haplotypes of
DNA repair genes in childhood acute lymphoblastic leukemia. *Pediatr Blood Cancer.*
48:16-20.
174. Brenner AV, Butler MA, Wang SS, Ruder AM, Rothman N, Schulte PA, Chanock SJ,
Fine HA, Linet MS, Inskip PD (2007) Single-nucleotide polymorphisms in selected cytokine
genes and risk of adult glioma. *Carcinogenesis.* **28**:2543-2547.
175. Shen H, Sturgis EM, Khan SG, Qiao Y, Shahnavi T, Eicher SA, Su Y, Wang X, Strom
SS, Spitz MR, Kraemer KH, Wei Q (2001) An intronic Poly(AT) polymorphism of the DNA

- repair gene XPC and risk of squamous cell carcinoma of the head and neck: a case-control study. *Cancer Res.* **61**:3321-3325
176. Manuguerra M, Saletta F, Karagas MR, Berwick M, Vegila F, Vineis P, Matullo G (2006) XRCC3 and XPD/ERCC2 single nucleotide polymorphisms and the risk of cancer: a HuGE review. *Am J Epidemiol* **164**:297-302
177. Park DJ, Stoecklacher J, Zhang W, Tsao-Wei D, Groshen S, Lenz HJ (2001) A Xeroderma pigmentosum group D gene polymorphism predicts clinical outcome to platinum-based chemotherapy in patients with advanced colorectal cancer. *Cancer Res.* **61**:8654–8658
178. Coussen LM, Werb S (2002) Inflammation and Cancer. *Nature.* **420**:860-867
179. Ernst PB, Gold BD (2000) The disease spectrum of *Helicobacter pylori*: the immunopathogenesis of gastroduodenal ulcer and gastric cancer. *Annu Rev Microbiol.* **54**:615-640.
180. Khan G (2006) Epstein-Barr virus, cytokines, and inflammation: A cocktail for the pathogenesis of Hodgkin's lymphoma? *Exp Hematol.* **34**:399–406.
181. Balkwill F, Coussens LM (2004) Cancer: An inflammatory link. *Nature.* **431**:405-406.
182. Skinnider BF, Mak TW (2002) The role of cytokines in classical Hodgkin lymphoma. *Blood.* **99**:4283-4287
183. Yung L, Linch D (2003) Hodgkin's lymphoma. *Lancet.* **36**:943-951.
184. Vineis P, Anttila S, Benhamou S, Spinola M, Hirvonen A, Kiyohara C, Garte SJ, Puntoni R, Rannug A, Strange RC, Taioli E (2007) Evidence of gene gene interactions in lung carcinogenesis in a large pooled analysis. *Carcinogenesis.* **28**:1902-1905.

185. Galvan A, Loannidis LP, and Dragani TA (2010) Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends Genet.* **26**:132-141.
186. Ribas G, Gonzalez-Neira A, Salas A, Milne RL, Vega A, Carracedo B, et al. (2006) Evaluating Hap- Map SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. *Human Genet.* **118**:669-679.
187. Hartman M, Loy EY, Ku CS, Chia KS (2010) Molecular epidemiology and its current clinical use in cancer management. *Lancet Oncol.* **11**:383-390.
188. Monroy CM, Cortes AC, Lopez MS, D'Amelio AM, Etzel CJ, Younes A, Strom SS, El-Zein R (2011) Hodgkin disease risk: role of genetic polymorphisms and gene-gene interactions in inflammation pathway genes. *Mol Carcinog.* **50**:36-46
189. Monroy CM, Cortes AC, Lopez M, Rourke E, Etzel CJ, Younes A, Strom SS, El-Zein R (2011) Hodgkin lymphoma risk: Role of genetic polymorphisms and gene-gene interactions in DNA repair pathways. *Mol Carcinog* **50**:825-834
190. Li N, Stephens M (2003) Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics.* **165**:2213-2233.
191. Kuzelicki NK, Milek M, Jazbec J, Mlinaric-Rascan I (2009) 5,10-Methylenetetrahydrofolate reductase (MTHFR) low activity genotypes reduce the risk of relapse-related acute lymphoblastic leukemia (ALL). *Leuk Res.* **33**:1344-1348.
192. van der Put NM, Gabreëls F, Stevens EM, Smeitink JA, Trijbels FJ, Eskes TK, van den Heuvel LP, Blom HJ (1998) A second common mutation in the methylenetetrahydrofolate

- reductase gene: an additional risk factor for neural-tube defects? *Am J Hum Genet.* **62**:1044–1051.
193. Lucock M (2000) Folic acid: nutritional biochemistry, molecular biology, and role in disease processes. *Mol Genet Metab.* **71**:121–138.
 194. Zhang SM, Hunter DJ, Rosner BA, Giovannucci EL, Colditz GA, Speizer FE, Willett WC (2000) Intakes of Fruits, Vegetables, and Related Nutrients and the Risk of Non-Hodgkin's Lymphoma among Women. *Cancer Epidemiol Biomarkers Prev* **9**:477-485
 195. Lim U, Schenk M, Kelemen LE, Davis S, Cozen W, Hartge P, Ward MH, Stolzenberg-Solomon R (2005) Dietary Determinants of One-Carbon Metabolism and the Risk of Non-Hodgkin's Lymphoma: NCI-SEER Case-Control Study, 1998–2000. *Am J Epidemiol* **162**:953–964
 196. Vogel U, Overvad K, Wallin H, Tjønneland A, Nexø BA, Raaschou-Nielsen O (2005) Combinations of polymorphisms in XPD, XPC and XPA in relation to risk of lung cancer, *Cancer Lett.* **222**:67-74
 197. Hu Z, Wang Y, Wang X, Liang G, Miao X, Xu Y, Tan W, Wei Q, Lin D, Shen H (2005) DNA repair gene XPC genotypes/haplotypes and risk of lung cancer in a Chinese population. *Int J Cancer.* **115**:478-483.
 198. Khan SG, Metter EJ, Tarone RE, Bohr VA, Grossman L, Hedayati M, Bale SJ, Emmert S, Kraemer KH (2000) A new xeroderma pigmentosum group C poly(AT) insertion/deletion polymorphism, *Carcinogenesis.* **21**:1821-1825.
 199. Vodicka P, Kumar R, Stetina R, Sanyal S, Soucek P, Haufroid V, et al. (2004) Genetic polymorphisms in DNA repair genes and possible links with DNA repair rates, chromosomal aberrations and single-strand breaks in DNA. *Carcinogenesis.* **25**:757-763.

200. Kawakami M, Kawakami K, Kioi M, Leland P, Puri RK (2005) Hodgkin lymphoma: therapy with interleukin-4 receptor-directed cytotoxin in an infiltrating animal model. *Blood*. **105**:3707-3713.
201. Murata T, Obiri NI, Puri RK (1998) Structure of and signal transduction through interleukin-4 and interleukin-13 receptors. *Int J Mol Med*. **1**:551-557.
202. Landi MT, Consonni D, Rotunno M, Bergen AW, Goldstein AM, Lubin JH, Goldin L, Alavanja M, Morgan G, Subar AF, et al. (2008). Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population based case-control study of lung cancer. *BMC Public Health*. **8**:e203.
203. The ATBC Cancer Prevention Study Group. (1994). The alphas-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. *Ann. Epidemiol*. **4**:1–10.
204. Hayes RB, Sigurdson A, Moore L, Peters U, Huang WY, Pinsky P, Reding D, Gelmann, EP, Rothman N, Pfeiffer RM, et al. (2005). Methods for etiologic and early marker investigations in the PLCO trial. *Mutat. Res*. **592**:147–154.
205. Calle EE, Rodriguez C, Jacobs EJ, Almon ML, Chao A, McCullough ML, Feigelson HS, Thun MJ (2002). The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer*. **94**:2490–2501.
206. Higgins, JP, Thompson SG (2002). Quantifying heterogeneity in a meta-analysis. *Stat. Med*. **21**:1539–1558.
207. Berrettini W, Yuan X, Tozzi F, Song K, Francks C, Chilcoat H, Waterworth D, Muglia P, Mooser V (2008) α -5/ α -3 nicotinic receptor subunit alleles increase risk for heavy smoking *Mol Psychiatry*. **13**: 368–373.

208. American Cancer Society (2011) Cancer Facts & Figures 2011. Atlanta: American Cancer Society.
209. Schwartzbaum JA, Fisher JL, Aldape KD, Wrensch M (2006) Epidemiology and molecular pathology of Glioma. *Nat Clin Pract Neurol* **2**:494-503.
210. Central Brain Tumor Registry of the United States [<http://www.CBTRUS.org>]
211. Hardell L, Carlberg M, Söderguist F, Mild KH, Morgan LL (2007) Long-term use of cellular phones and brain tumours: Increased risk associated with use for ≥ 10 years. *Occup Environ Med.* **64**:626–632
212. Hepworth SJ, Schoemaker MJ, Muir KR, Swerdlow AJ, van Tongeren MJ, McKinney PA (2006) Mobile phone use and risk of glioma in adults: case–control study. *BMJ (Clinical research ed)* **332**:883–887
213. Hu J, Little J, Xu T, Zhao X, Guo L, Jia X, Huang G, Bi D, Liu R (1999) Risk factors for meningioma in adults: a case–control study in northeast China. *Int J Cancer* **83**:299–304
214. Inskip PD, Møller M, Gridley G, Olsen JH (1998) Incidence of intracranial tumors following hospitalization for head injuries (Denmark). *Cancer Causes and Control* **9**: 109–116
215. Lambe M, Coogan P, Baron J (1997) Reproductive factors and the risk of brain tumors: a population-based study in Sweden. *Int J Cancer* **72**:389–393
216. Navas-Acién A, Pollán M, Gustavsson P, Plato N (2002) Occupation, exposure to chemicals and risk of gliomas and meningiomas in Sweden. *Am J Ind Med* **42**:214–227
217. Benson VS, Pirie K, Green J, Casabonne D, Beral V (2008) Lifestyle factors and primary Glioma and meningioma tumours in the Million Women Study cohort. *Br J Cancer.* **99**:185-190

218. World Health Organization International Agency for Research on Cancer Monograph Working Group (2011) Carcinogenicity of radiofrequency electromagnetic fields. *Lancet Oncol.* **12**:624-626
219. Schwartzbaum JA, Ahlbom A, Lonn S, Malmer B, Wigertz A, Auvinen A, et al (2007) An international case-control study of interleukin-4Ralpha, interleukin-13, and cyclooxygenase-2 polymorphisms and glioblastoma risk. *Cancer Epidemiol Biomarkers Prev* **16**:2448-2454.
220. Schwartzbaum J, Ahlbom A, Malmer B, Lonn S, Brookes AJ, Doss H, et al. (2005) Polymorphisms associated with asthma are inversely related to glioblastoma multiforme. *Cancer Res.* **65**:6459-6465.
221. Bondy ML, Scheurer ME, Malmer B, Barnholtz-Sloan JS, Davis FG, Il'yasova D, et al. (2008) Brain tumor epidemiology: consensus from the Brain Tumor Epidemiology Consortium. *Cancer.* **113**:1953-1968.
222. Schoemaker MJ, Swerdlow AJ, Hepworth SJ, McKinney PA, van TM, Muir KR (2006) History of allergies and risk of glioma in adults. *Int J Cancer* **119**:2165-2172.
223. Brenner AV, Butler MA, Wang SS, Ruder AM, Rothman N, Schulte PA, Chanock SJ, Fine HA, Linet MS, Inskip PD (2007) Single-nucleotide polymorphisms in selected cytokine genes and risk of adult glioma. *Carcinogenesis.* **28**:2543-2547.
224. Amirian ES, Scheurer ME, Liu Y, D'Amelio Jr AM, Houlston RS, Etzel CJ, Shete S, Swerdlow AJ, Schoemaker MJ, McKinney PA, Fleming SJ, Muir KR, Lophatananon A, Bondy ML (2011) A Novel Approach to Exploring Potential Interactions among Single-Nucleotide Polymorphisms of Inflammation Genes in Gliomagenesis: An Exploratory Case-Only Study. *Cancer Epidemiol Biomarkers Prev.* **20**:1683-1689

225. Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B, Simon M, Marie Y, Boisselier B, et al. (2009) Genome-wide association study identifies five susceptibility loci for Glioma. *Nat Genet.* **41**:899-904. doi:10.1038/ng.407
226. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager MA, Hankinson SE, Wacholder S, Wang Z, Welch R, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet.* **39**:870-874.
227. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, et al. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**:645-649.
228. National Cancer Institute. Executive Summary: Cancer Genetic Markers of Susceptibility (CGEMS) Project. National Cancer Institute. National Institutes of Health. Bethesda, Maryland. http://cgems.cancer.gov/executive_summary.html [Accessed June 23, 2011]
229. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet.* **81**:559-575
230. Zhang J, Xiang Y, Ding L, Keen-Circle K, Borlawsky TB, Ozer HG, Jin R, Payne P, Huang K (2010) Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. *BMC Bioinformatics.* **11**:Suppl 9:S5
231. Radstake TRDJ, Gorlova O, Rueda B, Martian J-E, Alizadeh BZ, Palomino-Morales R, Coenen MJ, Vonk MC, Voskuyl AE, et al (2010) Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. *Nat Genet.* **42**:426-429.

232. Call ME, Wucherpfennig KW (2004) Molecular mechanisms for the assembly of the T cell receptor-CD3 complex. *Mol. Immunol.* **40**:1295–1305
233. Barreca A, Lasorsa E, Riera L, Machiorlatti R, Piva R, Ponzoni M, Kwee I, Bertoni F, Piccaluga PP (2011) Anaplastic Lymphoma Kinase in Human Cancer. *J Mol Endocrinol.* **47**:R11-R23
234. Kjellman C, Olofsson SP, Hansson O, Von Schantz T, Lindvall M, Nilsson I, Salford LG, Sjögren HO, Widegren B (2000) Expression of TGF-beta isoforms, TGF-beta receptors, and SMAD molecules at different stages of human Glioma. *Int J Cancer* **89**:251-258
235. Li H, Liu JP (2007) Mechanisms of action of TGF-beta in cancer: evidence for SMAD3 as a repressor of the hTERT gene. *Ann N Y Acad Sci.* **1114**:56-68
236. Han SU, Kim HT, Seong DH, Kim YS, Park YS, Bang YJ, Yang HK, Kim SJ (2004) Loss of the Smad3 expression increases susceptibility to tumorigenicity in human gastric cancer. *Oncogene.* **23**:1333-1341
237. Lennon G, Auffray C, Polymeropoulos M, Soares MB (1996) The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33**:151–152.
238. Horvat S, Medrano JF (1999) A 500-kb YAC and BAC contig encompassing the high-growth deletion in mouse chromosome 10 and identification of the murine Raidd/Cradd gene in the candidate region. *Genomics.* **54**:159–164.
239. Shearwin-Whyatt LM, Harvey NL, Kumar S (2000) Subcellular localization and CARD-dependent oligomerization of the death adaptor RAIDD. *Cell Death Differ.* **7**:155–165.
240. Hasegawa H, Yamada Y, Tsukasaki K, Mori N, Tsuruda K, Sasaki D, Usui T, Osaka A, Atogami S, et al (2011) LBH589, a deacetylase inhibitor, induces apoptosis in adult T-cell

leukemia/lymphoma cells via activation of a novel RAIDD-caspase-2 pathway. *Leukemia*. **25**:575-587

241. Guo Y, Srinivasula SM, Druilhe A, Fernandes-Alnemri T, Alnemri ES (2002) Caspase-2 induces apoptosis by releasing proapoptotic proteins from mitochondria. *J Biol Chem*. **277**:13430–13437
242. Zhang Jian, Sun R, Wei H, Zhang Jianhua, Tian Z (2004) Characterization of interleukin-15 gene-modified human natural killer cells: implications for adoptive cellular immunotherapy. *Haematologica*. **89**:338-347
243. Thun MJ, Lally CA, Flannery Jr JT (1997) Cigarette smoking and changes in the histopathology of lung cancer. *J Natl Cancer Inst* **89**:1580–1586.
244. Fenech M (2002) Chromosomal biomarkers of genomic instability relevant to cancer. *Drug Discov Today* **7**:1128-1137
245. El-Zein RA, Schabath MB, Etzel CJ, Lopez MS, Franklin JD, Spitz MR (2006) Cytokinesis-Blocked Micronucleus Assay as a Novel Biomarker for Lung Cancer Risk. *Cancer Res* **66**:6449-6656
246. El-Zein R, Vral A, Etzel CJ (2011) Cytokinesis-blocked micronucleus assay and cancer risk assessment. *Mutagenesis* **26**:101-106

Appendix 1: Incidence rates and mortality rates used to calculate absolute risk in the validation of the Spitz lung cancer risk model in Chapter 2

Table 1: Lung cancer and mortality rates per 100,000 (excluding lung cancer) by age and sex for residents of the Netherlands

Age (years)	Men		Women	
	Incidence	Mortality	Incidence	Mortality
50–54	61.4	43.9	75.0	49.4
55–59	115.8	87.0	90.6	69.3
60–64	192.4	151.7	129.9	104.8
65–69	325.1	258.4	142.3	114.9
70–74	474.2	432.8	174.4	166.7

Table 2: Adjustment constants* (ac_{ji}) to estimate smoking status–specific incidence rates in the NELSON Data Set

Sex	Never Smokers	Former Smokers	Current Smokers
Male	0.12	2.55	2.99†
Female	0.42	3.08	3.17

* Adjustment constants (ac_{ji} , $j = 1$ male, $j = 2$ female; $i = 1$ never smoker, $i = 2$ former smoker, $i = 3$ current smoker) computed based on the following prevalence estimates: According to data from the Netherlands Cohort Study on Diet and Cancer, 96.4% of all male lung cancer cases and 79.3% of all female lung cancer cases occur in ever smokers^{1a}; 32.2% of men (aged 15+) and 25.0% of women (aged 15+) were current smokers^{2a}, 30.0% of men and 49.7% of women were never smokers^{3a}, and therefore 37.8% of men and 25.7% of women were former smokers^{3a}.

† Therefore, for a male current smoker, the constant is derived from the ratio of the proportion of all lung cancer cases in ever-smoking men (0.964) to the proportion of male current smokers in the population at risk (0.322), i.e., $ac_{13} = 0.964/0.322 = 2.99$

Appendix 2: Simulating Missing Data Results for all Haplotypes > 5% Frequency in the BJLM Simulation Study for Haplotypes of Length 4 to 11

Table 1: Haplotype Frequencies for 4 SNP Haplotypes. CI = Credible Interval

Haplotype #	Haplotype	Median Frequency	2.5% CI	97.5% CI
1% Missing SNP Data				
1	TGAC	0.307	0.284	0.334
11	CAAC	0.156	0.137	0.181
4	TGGT	0.140	0.118	0.161
3	TGGC	0.102	0.086	0.122
14	CAGT	0.081	0.061	0.098
5	TAAC	0.075	0.057	0.089
2% Missing SNP Data				
1	TGAC	0.306	0.281	0.331
11	CAAC	0.161	0.141	0.181
4	TGGT	0.142	0.124	0.159
3	TGGC	0.102	0.084	0.121
14	CAGT	0.083	0.065	0.104
5	TAAC	0.075	0.059	0.089
3% Missing SNP Data				
1	TGAC	0.309	0.286	0.335
11	CAAC	0.159	0.138	0.184
4	TGGT	0.141	0.122	0.161
3	TGGC	0.101	0.080	0.122
14	CAGT	0.085	0.069	0.102
5	TAAC	0.078	0.058	0.092
4% Missing SNP Data				
1	TGAC	0.314	0.283	0.338
11	CAAC	0.164	0.137	0.189
4	TGGT	0.144	0.125	0.167
3	TGGC	0.098	0.084	0.115
14	CAGT	0.084	0.068	0.104
5	TAAC	0.075	0.063	0.091
5% Missing SNP Data				
1	TGAC	0.304	0.271	0.327
11	CAAC	0.167	0.145	0.187
4	TGGT	0.152	0.135	0.175
3	TGGC	0.104	0.077	0.127
14	CAGT	0.084	0.067	0.104
5	TAAC	0.071	0.054	0.088

Figure 1: Graphical Representations for 4 SNP Haplotypes

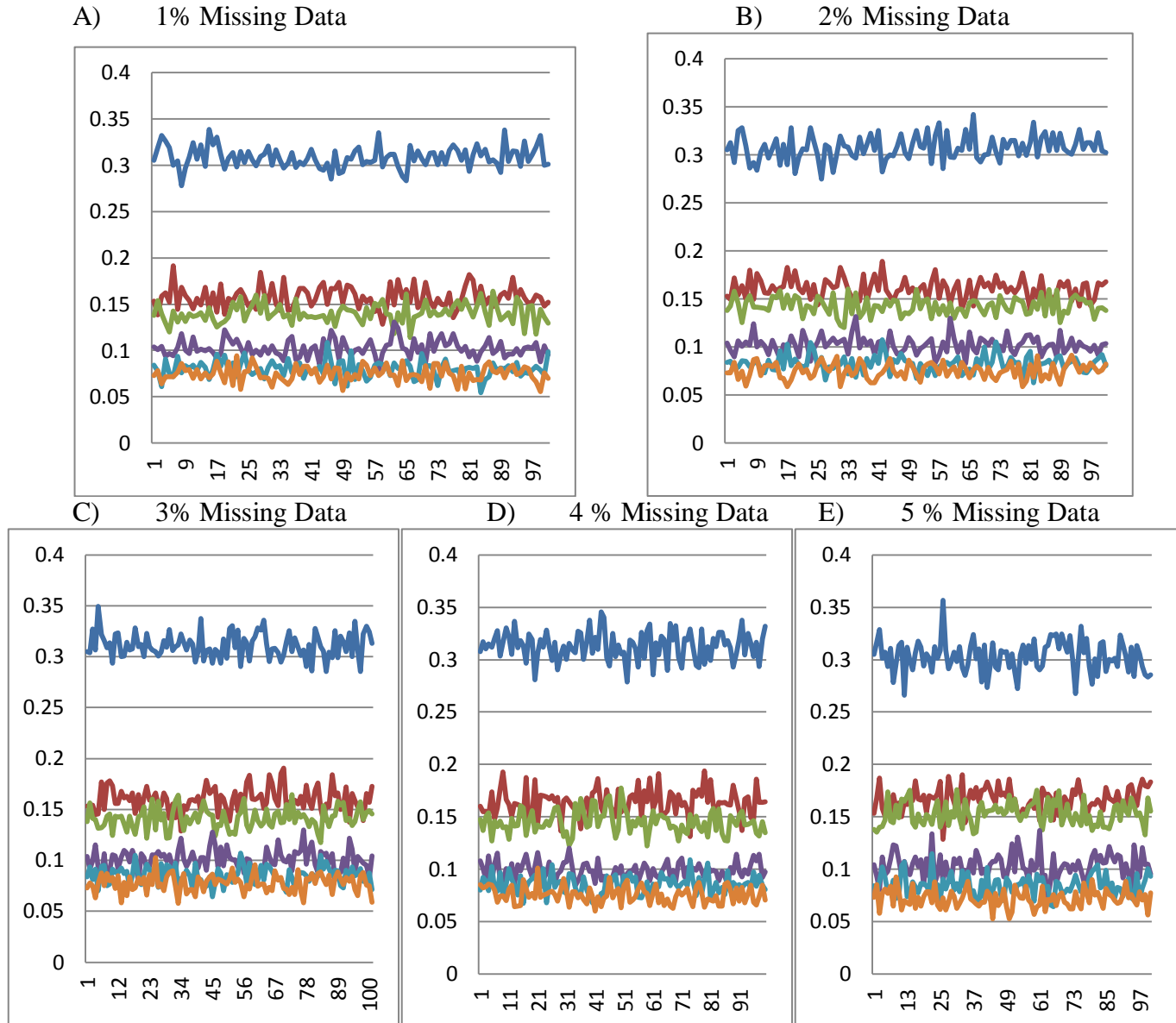


Figure 1: The X axis is the simulation run, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = TGAC, Red Line = CAAC, Green Line = TGGT, Purple Line = TGGC, Light Blue Line = CAGT, and Orange Line = TAAC.

Table 2: Haplotype Frequencies for 5 SNP Haplotypes. CI = Credible Interval

Haplotype #	Haplotype	Median Frequency	2.5% CI	97.5% CI
1% Missing SNP Data				
1	TGACC	0.305	0.274	0.329
13	CAACC	0.157	0.129	0.178
6	TGGTT	0.140	0.122	0.161
4	TGGCC	0.102	0.089	0.121
18	CAGTT	0.079	0.062	0.093
7	TAACC	0.074	0.060	0.088
2% Missing SNP Data				
1	TGACC	0.304	0.283	0.327
13	CAACC	0.162	0.131	0.180
6	TGGTT	0.143	0.122	0.160
4	TGGCC	0.100	0.084	0.119
18	CAGTT	0.081	0.061	0.098
7	TAACC	0.075	0.060	0.094
3% Missing SNP Data				
1	TGACC	0.306	0.284	0.332
13	CAACC	0.161	0.138	0.179
6	TGGTT	0.141	0.120	0.163
4	TGGCC	0.100	0.078	0.117
18	CAGTT	0.083	0.064	0.103
7	TAACC	0.080	0.063	0.097
4% Missing SNP Data				
1	TGACC	0.297	0.268	0.328
13	CAACC	0.161	0.137	0.184
6	TGGTT	0.147	0.132	0.168
4	TGGCC	0.102	0.083	0.121
18	CAGTT	0.086	0.068	0.105
7	TAACC	0.077	0.059	0.091
5% Missing SNP Data				
1	TGACC	0.298	0.269	0.330
13	CAACC	0.165	0.139	0.190
6	TGGTT	0.151	0.124	0.167
4	TGGCC	0.097	0.077	0.117
18	CAGTT	0.084	0.061	0.100
7	TAACC	0.079	0.062	0.098

Figure 2: Graphical Representations for 5 SNP Haplotypes

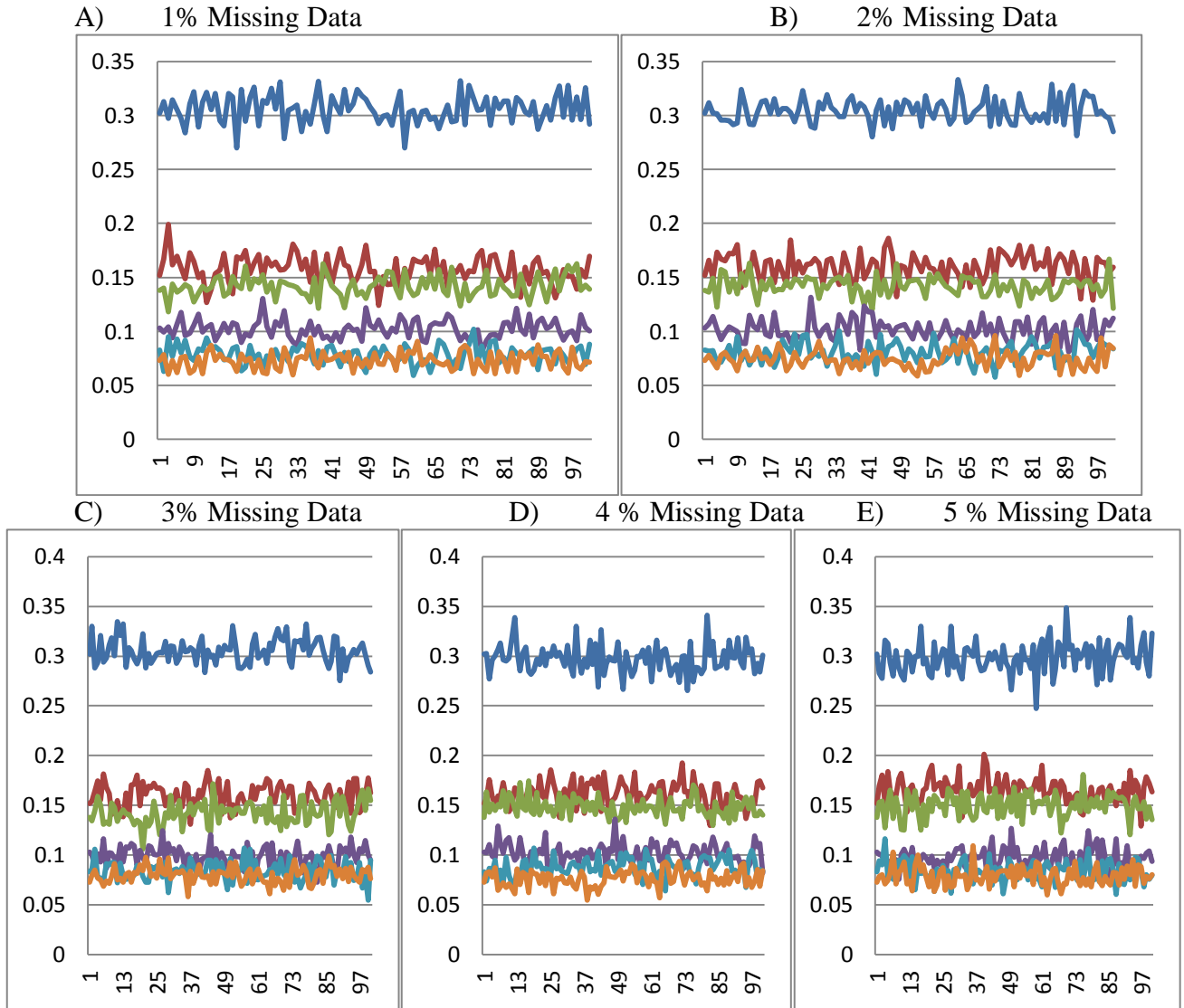


Figure 2: The X axis is the simulation run, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = TGACC, Red Line = CAACC, Green Line = TGGTT, Purple Line = TGGCC, Light Blue Line = CAGTT, and Orange Line = TAACC.

Table 3: Haplotype Frequencies for 6 SNP Haplotypes. CI = Credible Interval

Haplotype #	Haplotype	Median Frequency	2.5% CI	97.5% CI
1% Missing SNP Data				
1	TGACCT	0.303	0.281	0.327
13	CAACCC	0.156	0.135	0.175
6	TGGTTT	0.141	0.123	0.162
4	TGGCCT	0.099	0.082	0.118
18	CAGTTC	0.081	0.059	0.101
7	TAACCC	0.075	0.057	0.087
2% Missing SNP Data				
1	TGACCT	0.300	0.274	0.323
13	CAACCC	0.158	0.137	0.183
6	TGGTTT	0.142	0.120	0.165
4	TGGCCT	0.101	0.085	0.123
18	CAGTTC	0.082	0.064	0.100
7	TAACCC	0.079	0.065	0.093
3% Missing SNP Data				
1	TGACCT	0.299	0.272	0.322
13	CAACCC	0.162	0.138	0.181
6	TGGTTT	0.145	0.126	0.164
4	TGGCCT	0.101	0.083	0.120
18	CAGTTC	0.083	0.067	0.100
7	TAACCC	0.079	0.060	0.101
4% Missing SNP Data				
1	TGACCT	0.293	0.264	0.318
13	CAACCC	0.159	0.138	0.189
6	TGGTTT	0.148	0.124	0.167
4	TGGCCT	0.100	0.082	0.117
18	CAGTTC	0.084	0.071	0.103
7	TAACCC	0.084	0.066	0.102
5% Missing SNP Data				
1	TGACCT	0.291	0.265	0.324
13	CAACCC	0.164	0.139	0.188
6	TGGTTT	0.146	0.126	0.173
4	TGGCCT	0.100	0.084	0.120
18	CAGTTC	0.085	0.067	0.104
7	TAACCC	0.082	0.059	0.100

Figure 3: Graphical Representations for 6 SNP Haplotypes

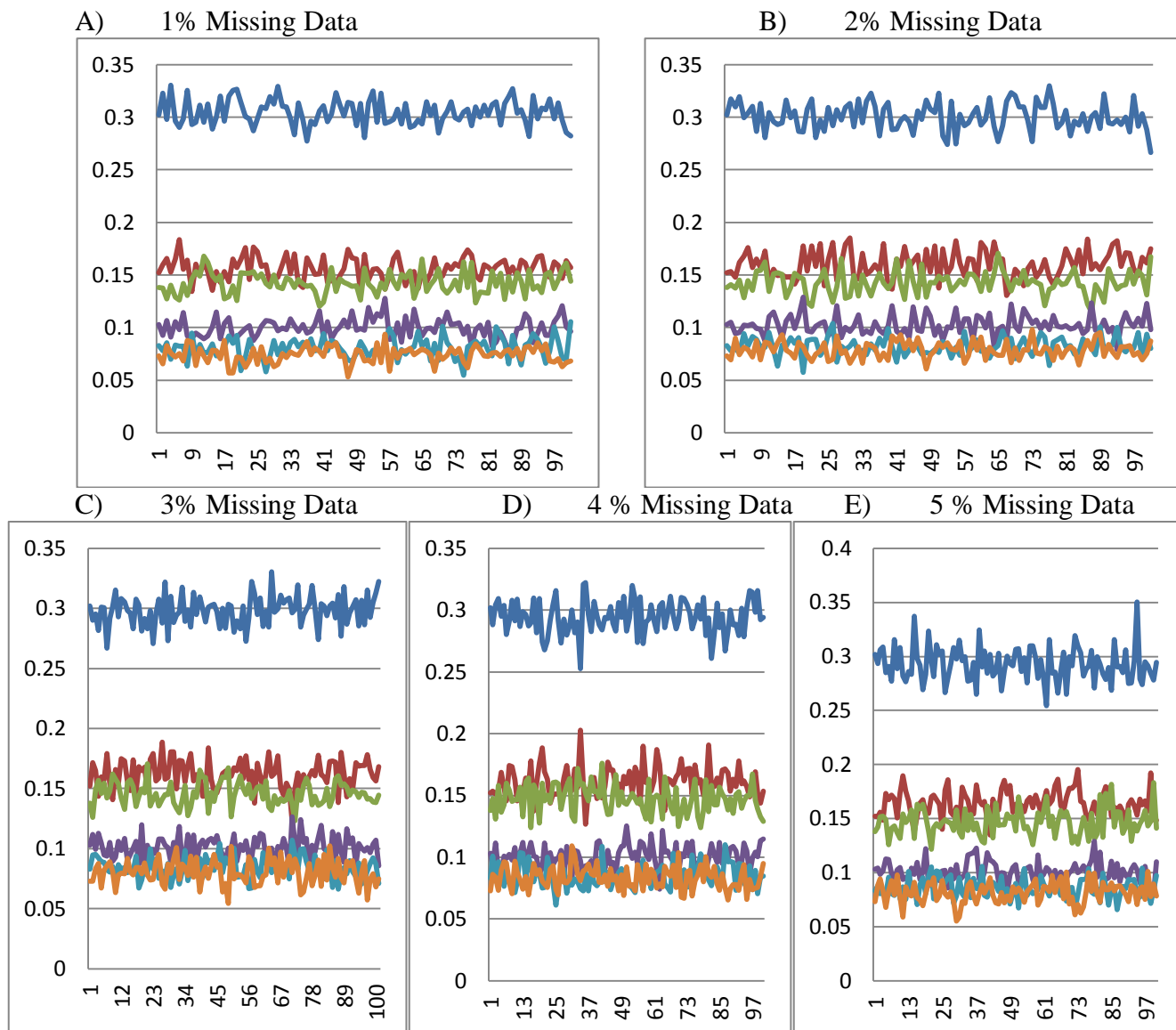


Figure 3: The X axis is the simulation run, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = TGACCT, Red Line = CAACCC, Green Line = TGGTTT, Purple Line = TGGCCT, Light Blue Line = CAGTTC, and Orange Line = TAACCC.

Table 4: Haplotype Frequencies for 7 SNP Haplotypes. CI = Credible Interval

Haplotype #	Haplotype	Median Frequency	2.5% CI	97.5% CI
1% Missing SNP Data				
1	TGACCTC	0.292	0.258	0.315
19	CAACCCC	0.153	0.129	0.175
8	TGGTTTC	0.113	0.088	0.128
24	CAGTTCC	0.081	0.065	0.104
10	TAACCCC	0.072	0.058	0.086
5	TGGCCTC	0.061	0.046	0.073
2% Missing SNP Data				
1	TGACCTC	0.288	0.261	0.311
19	CAACCCC	0.159	0.137	0.181
8	TGGTTTC	0.112	0.089	0.132
24	CAGTTCC	0.084	0.067	0.101
10	TAACCCC	0.072	0.055	0.087
5	TGGCCTC	0.059	0.044	0.074
3% Missing SNP Data				
1	TGACCTC	0.281	0.256	0.308
19	CAACCCC	0.161	0.137	0.180
8	TGGTTTC	0.116	0.098	0.132
24	CAGTTCC	0.084	0.068	0.100
10	TAACCCC	0.073	0.060	0.094
5	TGGCCTC	0.060	0.045	0.076
4% Missing SNP Data				
1	TGACCTC	0.281	0.251	0.313
19	CAACCCC	0.162	0.140	0.184
8	TGGTTTC	0.115	0.098	0.133
24	CAGTTCC	0.085	0.068	0.104
10	TAACCCC	0.076	0.060	0.091
5	TGGCCTC	0.057	0.044	0.078
5% Missing SNP Data				
1	TGACCTC	0.284	0.259	0.306
19	CAACCCC	0.165	0.141	0.189
8	TGGTTTC	0.115	0.094	0.130
24	CAGTTCC	0.086	0.067	0.101
10	TAACCCC	0.076	0.057	0.093
5	TGGCCTC	0.059	0.043	0.075

Figure 4: Graphical Representation for 7 SNP Examinations

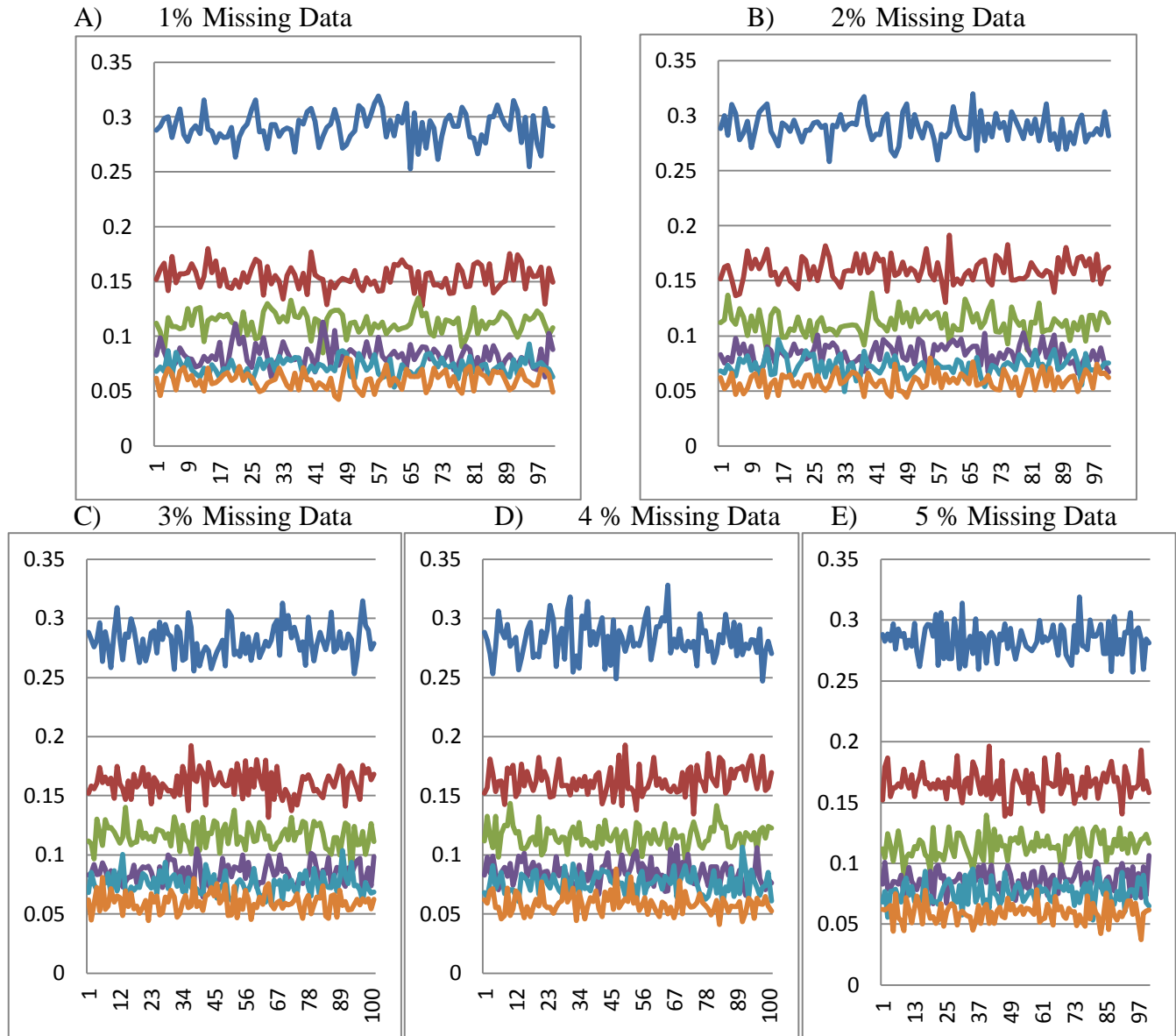


Figure 4: The X axis is the simulation run, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = TGACCTC, Red Line = CAACCCC, Green Line = TGGTTTC, Purple Line = CAGTTCC, Light Blue Line = TAACCCC, and Orange Line = TGGCCTC.

Table 5: Haplotype Frequencies for 8 SNP Haplotypes. CI = Credible Interval

Haplotype #	Haplotype	Median Frequency	2.5% CI	97.5% CI
1% Missing SNP Data				
1	TGACCTCC	0.288	0.263	0.319
19	CAACCCCC	0.160	0.132	0.180
8	TGGTTTCT	0.110	0.091	0.126
24	CAGTTCCT	0.081	0.063	0.101
10	TAACCCCC	0.071	0.056	0.082
5	TGGCCTCC	0.061	0.049	0.075
2% Missing SNP Data				
1	TGACCTCC	0.287	0.268	0.309
19	CAACCCCC	0.159	0.140	0.182
8	TGGTTTCT	0.115	0.095	0.137
24	CAGTTCCT	0.082	0.067	0.097
10	TAACCCCC	0.075	0.057	0.094
5	TGGCCTCC	0.059	0.045	0.073
3% Missing SNP Data				
1	TGACCTCC	0.278	0.254	0.304
19	CAACCCCC	0.167	0.143	0.191
8	TGGTTTCT	0.121	0.098	0.139
24	CAGTTCCT	0.077	0.059	0.094
10	TAACCCCC	0.074	0.061	0.090
5	TGGCCTCC	0.058	0.046	0.073
4% Missing SNP Data				
1	TGACCTCC	0.279	0.251	0.308
19	CAACCCCC	0.166	0.142	0.188
8	TGGTTTCT	0.119	0.097	0.139
24	CAGTTCCT	0.084	0.064	0.102
10	TAACCCCC	0.075	0.058	0.092
5	TGGCCTCC	0.060	0.045	0.077
5% Missing SNP Data				
1	TGACCTCC	0.278	0.250	0.301
19	CAACCCCC	0.171	0.145	0.195
8	TGGTTTCT	0.123	0.100	0.150
24	CAGTTCCT	0.080	0.060	0.097
10	TAACCCCC	0.079	0.062	0.095
5	TGGCCTCC	0.058	0.045	0.074

Figure 5: Graphical Representation for 8 SNP Examinations

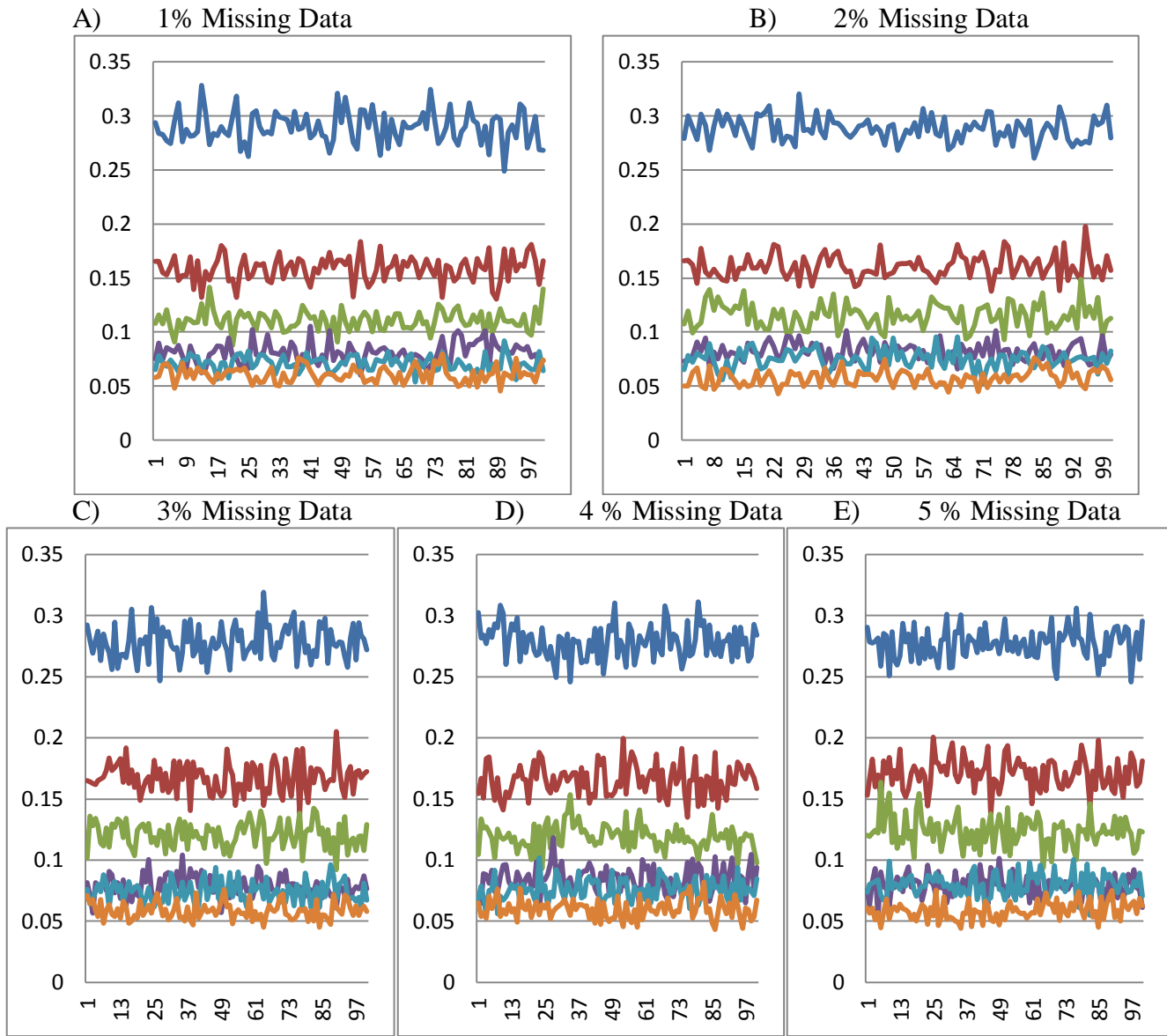


Figure 5: The X axis is the simulation run, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = TGACCTCC, Red Line = CAACCCCC, Green Line = TGGTTTCT, Purple Line = CAGTTCCT, Light Blue Line = TAACCCCC, and Orange Line = TGGCCTCC.

Table 6: Haplotype Frequencies for 9 SNP Haplotypes. CI = Credible Interval

Haplotype #	Haplotype	Median Frequency	2.5% CI	97.5% CI
1% Missing SNP Data				
1	TGACCTCCA	0.285	0.257	0.310
22	CAACCCCA	0.158	0.137	0.178
9	TGGTTTCTG	0.112	0.097	0.128
28	CAGTTCCTG	0.079	0.063	0.094
11	TAACCCCA	0.074	0.059	0.093
6	TGGCCTCCA	0.060	0.046	0.072
2% Missing SNP Data				
1	TGACCTCCA	0.286	0.261	0.318
22	CAACCCCA	0.157	0.133	0.186
9	TGGTTTCTG	0.116	0.094	0.134
28	CAGTTCCTG	0.084	0.065	0.101
11	TAACCCCA	0.074	0.057	0.090
6	TGGCCTCCA	0.059	0.044	0.076
3% Missing SNP Data				
1	TGACCTCCA	0.281	0.255	0.311
22	CAACCCCA	0.160	0.138	0.182
9	TGGTTTCTG	0.117	0.092	0.140
28	CAGTTCCTG	0.085	0.071	0.101
11	TAACCCCA	0.076	0.057	0.093
6	TGGCCTCCA	0.061	0.046	0.078
4% Missing SNP Data				
1	TGACCTCCA	0.277	0.250	0.303
22	CAACCCCA	0.161	0.137	0.191
9	TGGTTTCTG	0.120	0.103	0.139
28	CAGTTCCTG	0.083	0.068	0.103
11	TAACCCCA	0.078	0.064	0.094
6	TGGCCTCCA	0.059	0.043	0.072
5% Missing SNP Data				
1	TGACCTCCA	0.285	0.254	0.312
22	CAACCCCA	0.165	0.143	0.187
9	TGGTTTCTG	0.115	0.099	0.139
28	CAGTTCCTG	0.084	0.065	0.099
11	TAACCCCA	0.079	0.058	0.098
6	TGGCCTCCA	0.056	0.041	0.072

Figure 6: Graphical Representation for 9 SNP Examinations

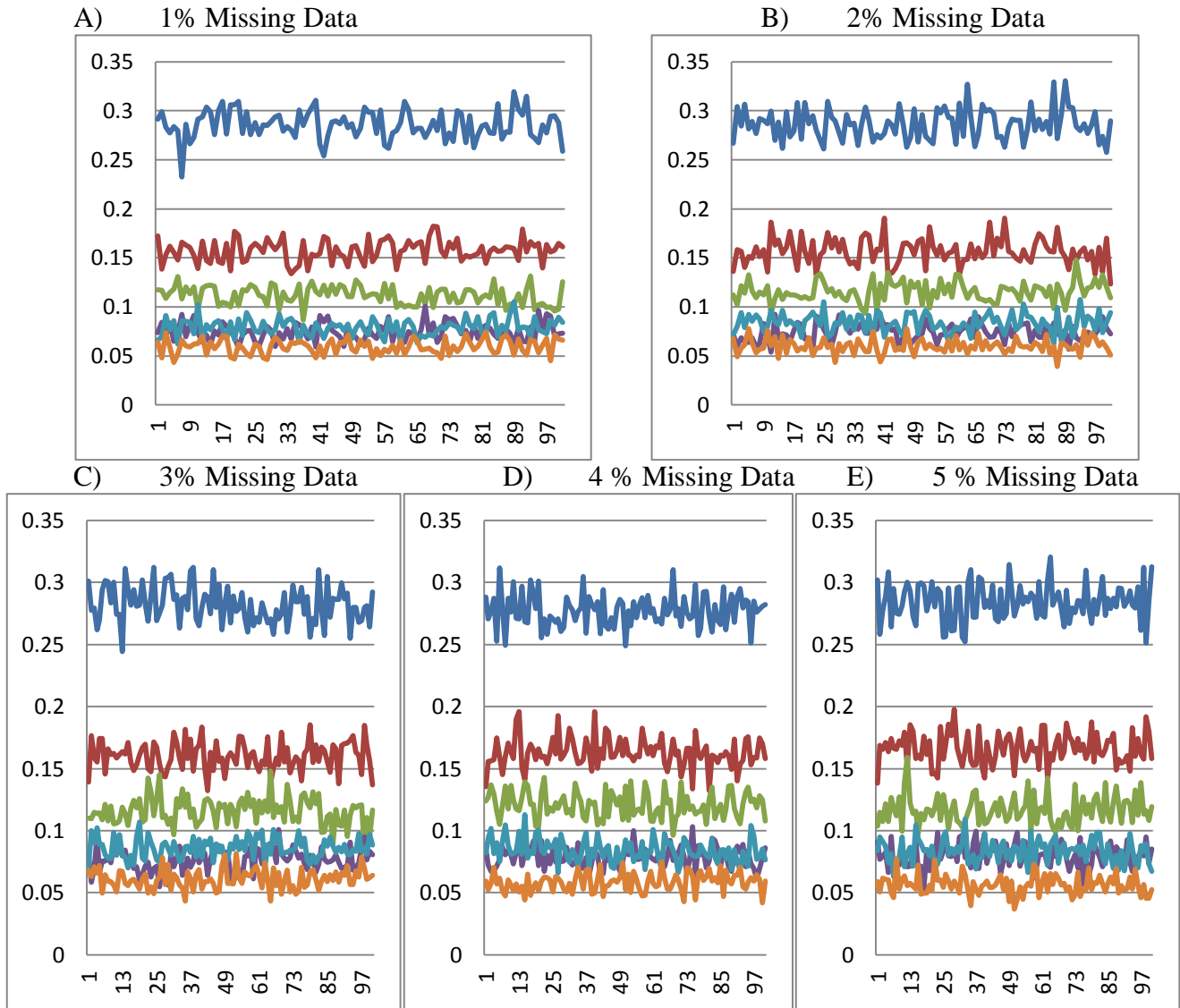


Figure 6: The X axis is the simulation run, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = TGACCTCCA, Red Line = CAACCCCA, Green Line = TGGTTTCTG, Purple Line = CAGTTCCTG, Light Blue Line = TAACCCCA, and Orange Line = TGGCCTCCA.

Table 7: Haplotype Frequencies for 10 SNP Haplotypes. CI = Credible Interval

Haplotype #	Haplotype	Median Frequency	2.5% CI	97.5% CI
1% Missing SNP Data				
1	TGACCTCCAT	0.286	0.258	0.315
9	TGGTTTCTGT	0.112	0.092	0.130
24	CAACCCCCAT	0.103	0.079	0.122
11	TAACCCCCAT	0.076	0.058	0.094
34	CAGTTCCTGT	0.066	0.050	0.084
6	TGGCCTCCAT	0.059	0.047	0.074
2% Missing SNP Data				
1	TGACCTCCAT	0.281	0.251	0.308
9	TGGTTTCTGT	0.113	0.090	0.134
24	CAACCCCCAT	0.105	0.084	0.121
11	TAACCCCCAT	0.074	0.061	0.089
34	CAGTTCCTGT	0.073	0.054	0.090
6	TGGCCTCCAT	0.061	0.044	0.073
3% Missing SNP Data				
1	TGACCTCCAT	0.275	0.249	0.302
9	TGGTTTCTGT	0.116	0.093	0.136
24	CAACCCCCAT	0.104	0.087	0.125
11	TAACCCCCAT	0.076	0.056	0.094
34	CAGTTCCTGT	0.073	0.057	0.090
6	TGGCCTCCAT	0.062	0.046	0.076
4% Missing SNP Data				
1	TGACCTCCAT	0.278	0.241	0.306
9	TGGTTTCTGT	0.120	0.099	0.139
24	CAACCCCCAT	0.107	0.084	0.128
11	TAACCCCCAT	0.081	0.061	0.102
34	CAGTTCCTGT	0.069	0.051	0.081
6	TGGCCTCCAT	0.059	0.043	0.075
5% Missing SNP Data				
1	TGACCTCCAT	0.281	0.253	0.308
9	TGGTTTCTGT	0.123	0.098	0.147
24	CAACCCCCAT	0.109	0.090	0.129
11	TAACCCCCAT	0.077	0.057	0.093
34	CAGTTCCTGT	0.071	0.051	0.093
6	TGGCCTCCAT	0.060	0.045	0.077

Figure 7: Graphical Representation for 10 SNP Examinations

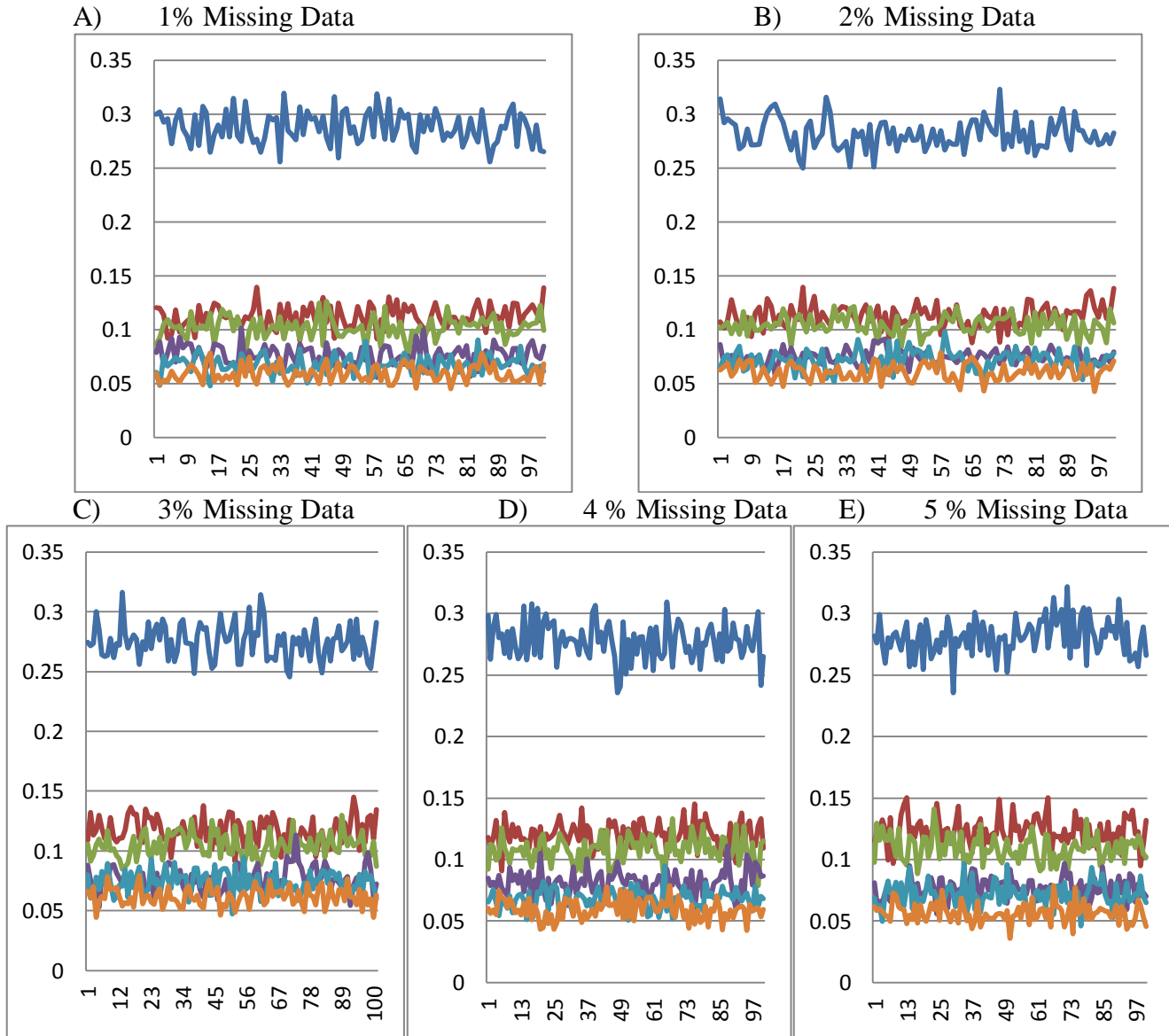


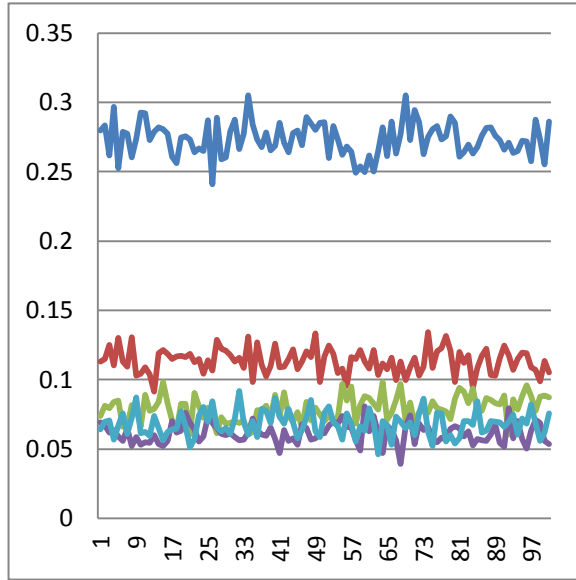
Figure 7: The X axis is the simulation run, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = TGACCTCCAT, Red Line = TGGTTTCTGT, Green Line = CAACCCCAT, Purple Line = TAACCCCAT, Light Blue Line = CAGTTCCTGT, and Orange Line = TGGCCTCCAT.

Table 8: Haplotype Frequencies for 11 SNP Haplotypes. CI = Credible Interval

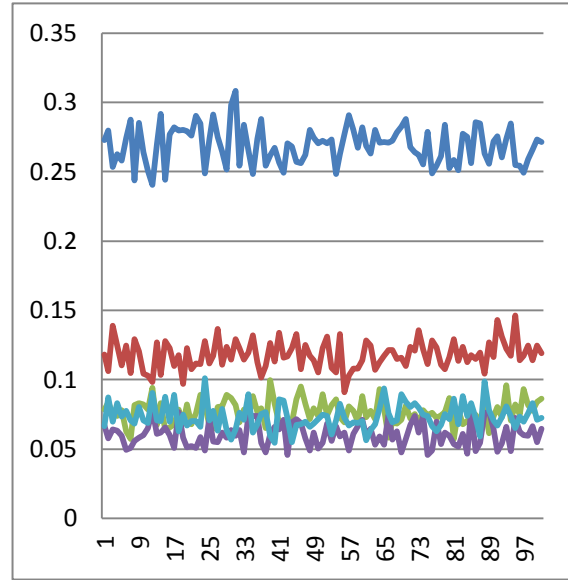
Haplotype #	Haplotype	Median Frequency	2.5% CI	97.5% CI
1% Missing SNP Data				
1	TGACCTCCATG	0.273	0.249	0.296
12	TGGTTTCTGTG	0.114	0.095	0.131
29	CAACCCCCATG	0.079	0.061	0.097
42	CAGTTCCTGTG	0.068	0.052	0.086
8	TGGCCTCCATG	0.061	0.047	0.075
14	TAACCCCCATG	0.055	0.044	0.070
2 % Missing SNP Data				
1	TGACCTCCATG	0.270	0.244	0.292
12	TGGTTTCTGTG	0.117	0.098	0.138
29	CAACCCCCATG	0.076	0.058	0.095
42	CAGTTCCTGTG	0.071	0.056	0.092
8	TGGCCTCCATG	0.060	0.046	0.077
14	TAACCCCCATG	0.053	0.040	0.069
3% Missing SNP Data				
1	TGACCTCCATG	0.267	0.230	0.300
12	TGGTTTCTGTG	0.116	0.097	0.138
29	CAACCCCCATG	0.078	0.059	0.096
42	CAGTTCCTGTG	0.073	0.052	0.091
8	TGGCCTCCATG	0.060	0.048	0.075
14	TAACCCCCATG	0.055	0.040	0.068
4% Missing SNP Data				
1	TGACCTCCATG	0.273	0.243	0.298
12	TGGTTTCTGTG	0.117	0.098	0.135
29	CAACCCCCATG	0.082	0.062	0.096
42	CAGTTCCTGTG	0.069	0.052	0.085
8	TGGCCTCCATG	0.059	0.041	0.072
14	TAACCCCCATG	0.053	0.038	0.075
5% Missing SNP Data				
1	TGACCTCCATG	0.269	0.237	0.305
12	TGGTTTCTGTG	0.123	0.098	0.146
29	CAACCCCCATG	0.081	0.059	0.100
42	CAGTTCCTGTG	0.064	0.046	0.080
8	TGGCCTCCATG	0.061	0.043	0.074
14	TAACCCCCATG	0.052	0.034	0.066

Figure 8: Graphical Representations for 11 SNP Haplotypes

A) 1% Missing Data



B) 2% Missing Data



C) 3% Missing Data

D) 4% Missing Data

E) 5% Missing Data

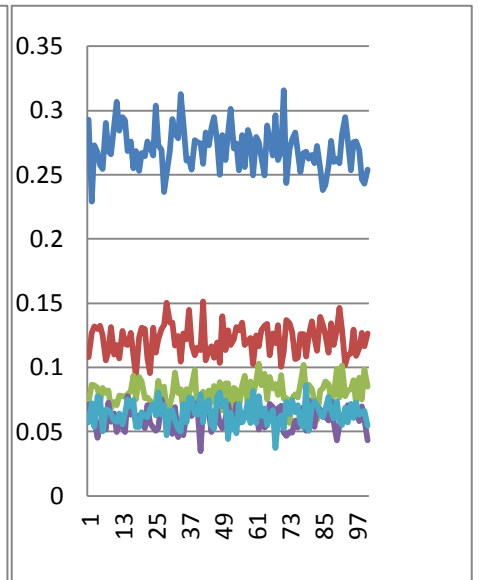
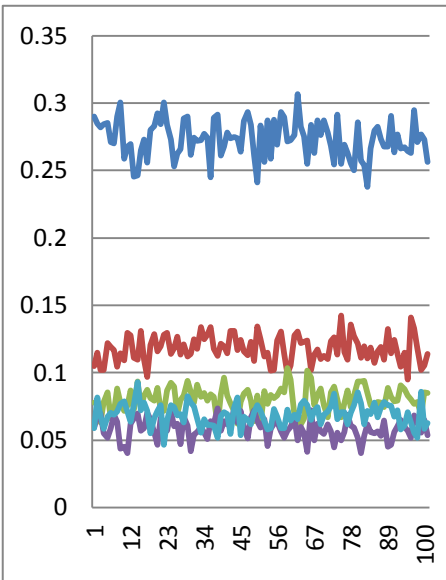
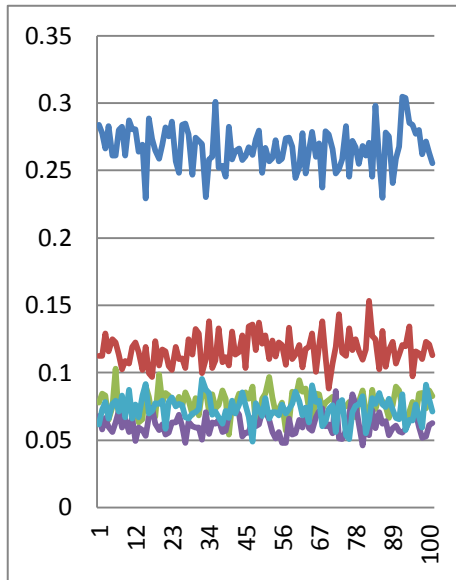


Figure 8: The X axis is the simulation run, and the Y axis is the haplotype frequency. The lines are as follows: Blue Line = Haplotype TGACCTCCATG, Red Line = Haplotype TGGTTTCTGTG, Green Line = Haplotype CAACCCCATG, Purple Line = Haplotype TGGCCTCCATG, and the Light Blue Line = Haplotype CAGTTCCTGTG.

Appendix 3: Chapter 7 (Glioma Development)

Figure 1: Sample WINBUGS Code for Analysis with Informative Priors

```
model
{
  for( i in 1 : N ) {
    casecntl[i] ~ dbin(p[i],1)
    logit(p[i]) <- alpha + (beta[1]*rs4761533[i]) + (beta[2]*rs8079544[i])
    + (beta[3]*rs10009998[i]) + (beta[4]*rs314253[i]) + (beta[5]*rs16833157[i])
    + (beta[6]*rs10484796[i]) + (beta[7]*rs11265608[i]) + (beta[8]*rs1319868[i])
    + (beta[9]*rs791589[i]) + (beta[10]*rs10857092[i]) + (beta[11]*rs6879021[i])
    + (beta[12]*rs693293[i]) + (beta[13]*rs959382[i]) + (beta[14]*rs2061450[i])
    + (beta[15]*rs276467[i]) + (beta[16]*Haplo2CD247[i])
    + (beta[17]*Haplo2HLADRA[i]) + (beta[18]*Haplo8CRADD2[i])
    + (beta[19]*Haplo8SMAD3[i]) + (beta[20]*Haplo5SMAD3[i])
    + (beta[21]*Haplo5ITGAX[i]) + (beta[22]*Haplo2CD2472[i])
    +(beta[23]*Haplo3CD2472[i]) + inprod(d[,Genetics[i,])
  }
  # Inprod functions represents a matrix of the final two Haplotypes to be examined
  # (PRF1 and GNAI1)
  # Priors for logit model
  alpha ~ dnorm(0.0,1.0E-2)
  beta[1] ~ dnorm(0.341,138.4083)
  beta[2] ~ dnorm(0.405,84.168)
  beta[3] ~ dnorm(-2.219,2.815189)
  beta[4] ~ dnorm(0.173,355.9986)
  beta[5] ~ dnorm(-0.414,60.09254)
  beta[6] ~ dnorm(-0.165,384.4675)
  beta[7] ~ dnorm(0.265,152.4158)
  beta[8] ~ dnorm(0.296,145.1589)
  beta[9] ~ dnorm(0.230,182.615)
  beta[10] ~ dnorm(0.319,94.25959)
  beta[11] ~ dnorm(0.157,384.4675)
  beta[12] ~ dnorm(-0.311,113.1734)
  beta[13] ~ dnorm(0.270,173.1302)
  beta[14] ~ dnorm(-0.178,277.7778)
  beta[15] ~ dnorm(0.237,192.9012)
  beta[16] ~ dnorm(0.7401,20.03704)
  beta[17] ~ dnorm(-0.23413,180.1721)
  beta[18] ~ dnorm(1.1938,5.07264)
  beta[19] ~ dnorm(-0.1866,191.834)
  beta[20] ~ dnorm(-1.3624,2.614936)
  beta[21] ~ dnorm(0.2366,183.6062)
  beta[22] ~ dnorm(0.144,313.2587)
  beta[23] ~ dnorm(0.172,203.4998)
```

```

d[1]~dnorm(0.1568,309.9583)
d[2]~dnorm(0.1897,172.6755)
# Determine Empirical p-value in Bayesian Context (P = 23, K = 2)
for (o in 1:P) {
  betapabove[o] <- step(beta[o]-0)
  betapbelow[o] <- 1-step(beta[o]-0)
}
for (l in 1:K)
{
  dpabove[l] <- step(d[l]-0)
  dpbelow[l] <- 1-step(d[l]-0)
}
}

#INSERT DATA HERE

Inits1 list(alpha = 0, beta = c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0), d = c(0,0))

```

Table 1: Location of Haplotypes that do not overlap with each other

Haplotypes	Chromosome	Genes
AGG	1	CD247
ATT	6	HLA-DRA
AAT	12	IGF1
AGCGG	12	CRADD
GCTCA	15	SMAD3
GTTTA	15	SMAD3
AGG	16	ITGAX
TG	1	CD247
GA	1	CD247
AGA	2	STAT
TGC	10	PRF1
TTTCA	7	GNAI1

Table 2: Most Significant SNPs for the Glioma Inflammation Set

SNP (rs Number)	Chromosome	Base-Pair	P-Value
rs4761533	12	92826085	0.000179
rs8079544	17	7520777	0.000325
rs10009998	4	142981726	0.000607
rs7963343	12	92819753	0.000669
rs865926	12	101469341	0.000696
rs314253	17	7032374	0.001055
rs16833157	2	191570643	0.001073
rs10484796	6	137477360	0.001151
rs11265608	1	152630764	0.001218
rs1319868	15	97004502	0.001469
rs2840191	6	91543934	0.001611
rs791589	10	6129577	0.00176
rs10857092	4	123608669	0.001864
rs6879021	5	86530513	0.001916
rs693293	3	139891866	0.00205
rs2395175	6	32513004	0.002267
rs959382	2	181915047	0.002414
rs2061450	18	58911792	0.002572
rs7341365	6	91537489	0.002641
rs276467	6	137505911	0.002712

Table 3: Associations between Best SNPs and Glioma assuming best genetic model for each SNP

Chromosome	SNP	Base-Pair	Genetic Model	OR (95% CI)	P-Value
12	rs4761533	92826085	Dominant	1.406 (1.191-1.660)	< 0.001
17	rs8079544	7520777	Dominant	1.499 (1.212-1.854)	< 0.001
4	rs10009998	142981726	Recessive	0.109 (0.034-0.350)	< 0.001
12	rs7963343	92819753	Dominant	1.326 (1.141-1.540)	< 0.001
12	rs865926	101469341	Dominant	1.523 (1.197-1.938)	0.001
17	rs314253	7032374	Additive	1.189 (1.072-1.318)	0.001
2	rs16833157	191570643	Additive	0.661 (0.513-0.851)	0.001
6	rs10484796	137477360	Additive	0.848 (0.768-0.937)	0.001
1	rs11265608	152630764	Additive	1.303 (1.111-1.529)	0.001
15	rs1319868	97004502	Dominant	1.344 (1.141-1.583)	< 0.001
6	rs2840191	91543934	Dominant	1.341 (1.166-1.542)	< 0.001
10	rs791589	6129577	Additive	1.259 (1.089-1.455)	0.002
4	rs10857092	123608669	Dominant	1.376 (1.125-1.683)	0.002
5	rs6879021	86530513	Additive	1.170 (1.059-1.292)	0.002
3	rs693293	139891866	Dominant	0.733 (0.610-0.880)	0.001
6	rs2395175	32513004	Additive	1.231 (1.077-1.407)	0.002
2	rs959382	181915047	Dominant	1.310 (1.129-1.521)	< 0.001
18	rs2061450	58911792	Additive	0.837 (0.744-0.941)	0.003
6	rs7341365	91537489	Dominant	1.325 (1.152-1.524)	< 0.001
6	rs276467	137505911	Dominant	1.268 (1.100-1.461)	0.001

Table 4: Best association results from the 18 unique haplotypes blocks, and the best genetic model for which the results were determined, are listed below:

Chm. (Gene)	Haplotype	Genetic Model	Haplo.stats results		BJLM results	
			OR (95% CI)	P-value	OR (95% CI)	P-value ¹
12 (CRADD)	GCAGA	Dominant	1.45(1.22-1.71)	0.000021	1.44(1.22-1.71)	0.000044
1 (IL6R)	GGAA	Additive	1.37(1.15-1.62)	0.000303	1.36(1.15-1.62)	0.000178
1 (CD247)	AGG	Dominant	2.10(1.35-3.25)	0.000932	2.11(1.36-3.27)	0.000756
17 (DLG4)	AG	Additive	1.16(1.04-1.30)	0.00841	1.16(1.04-1.30)	0.009022
6 (HLA-DRA)	ATT	Additive	0.79(0.68-0.92)	0.00170	0.79(0.68-0.91)	0.00187
12 (IGF1)	AAT	Dominant	1.45(1.16-1.82)	0.00112	1.45(1.16-1.81)	0.00111
15 (IGF1R)	AG	Dominant	1.27(1.09-1.48)	0.00243	1.27(1.09-1.48)	0.00218
15 (IGF1R)	AT	Dominant	1.42(1.20-1.69)	0.00038	1.42(1.20-1.69)	0.000133
6 (IL22RA2)	TCGGG	Additive	0.85(0.75-0.97)	0.0121	0.85(0.75-0.96)	0.0106
6 (IL22RA2)	TTGAA	Dominant	0.75(0.61-0.91)	0.00359	0.75(0.61-0.91)	0.00267
3 (PIK3CB)	TGCGAC C	Dominant	0.72(0.60-0.87)	0.000704	0.72(0.60-0.87)	0.000444
6 (MAP3K7)	GT	Dominant	1.34(1.16-1.54)	0.000048	1.34(1.17-1.54)	0.000088
12 (CRADD)	AGCGG	Additive	1.35(1.13-1.61)	0.000889	1.36(1.14-1.62)	0.000978
15 (SMAD3)	GCTCA	Recessive	0.26(0.08-0.86)	0.0277	0.22(0.05-0.70)	0.00671
15 (SMAD3)	GTTTA	Additive	0.83(0.72-0.96)	0.00982	0.83(0.72-0.96)	0.00947
16 (ITGAX)	AGG	Dominant	1.27(1.10-1.46)	0.00136	1.30(1.12-1.51)	0.000578
6 (HLA-DRA)	GCGACC AGTAC	Additive	1.25(1.07-1.49)	0.00575	1.25(1.07-1.47)	0.00538
6 (HLA-DRA)	ACGACC GGGGC	Recessive	0.52(0.27-0.99)	0.0461	0.51(0.25-0.95)	0.0336
1 (CD247)	TG	Additive	1.15(1.03-1.29)	0.0111	1.14(1.02-1.27)	0.0241
1 (CD247)	GA	Additive	1.18(1.04-1.36)	0.0142	1.16(1.01-1.33)	0.0359
2 (STAT) ²	AGA	Additive	0.86(0.76-0.96)	0.0116	0.85(0.76-0.96)	0.0115
10 (PRF1)	TGC	Additive	1.17(1.05-1.31)	0.00602	1.17(1.05-1.30)	0.00644
7 (GNAI1)	TTTCA	Dominant	1.21(1.04-1.40)	0.0127	1.21(1.05-1.41)	0.0112

¹ P-Value in this case is $2 \cdot \min(\text{freq}(\text{coefficient above zero}), \text{freq}(\text{coefficient below zero}))$ for each non burn-in run of the Bayesian Logistic model

² Actually is a haplotype that begins on the STAT1 gene and ends on the STAT4 gene

Vita

Anthony M. D'Amelio Jr. was born in Philadelphia, PA, on May 12, 1982. He received a Bachelor of Science of Engineering in the field of Biomedical Engineering from Duke University in May of 2004. Anthony worked as a youth fellow and research assistant with the Hepatitis B Foundation in Doylestown, PA, for a year from August 2004 to August 2005. During his time at the Hepatitis B Foundation, Anthony's contributions helped to publish two manuscripts in which he was a contributor. They are, "GP73, a resident Golgi glycoprotein, is a novel serum marker for hepatocellular carcinoma" in the Journal of Hepatology, and "N-linked glycosylation of the liver cancer biomarker GP73" in the Journal of Cell Biochemistry. Anthony also worked as a research assistant in the Quality Control department at Discovery Laboratories in Doylestown, PA, from October 2005 to April 2006. After deciding that Quality Control was not his life calling, he entered the Graduate School of Biomedical Sciences at University of Texas-Houston as a PhD student in August 2006. He entered the laboratory of Dr. Carol J. Etzel as a rotational student (20 hours a week for 10 weeks) from September 2006 to November 2006. After two more laboratory rotations, Anthony joined Dr. Carol J. Etzel's lab full time in June 2007. His research interest includes risk model development and validation, genetic analysis, and mathematical methods to improve discriminatory power. His research topic for his dissertation was the development of a Bayesian mathematical model to examine the associations between haplotypes (a linked set of single nucleotide polymorphisms) and disease.