A GLOBAL TEST OF ASSOCIATION BETWEEN NON-HODGKIN LYMPHOMA AND SNPS IN HISTONE-PATHWAY GENES

by

Jie Liu
B.Sc. in Actuarial Science, Simon Fraser University, 2010

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

Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2012

APPROVAL

Name: Jie Liu

Degree: Master of Science

Title of Project: A Global Test of Association between Non-Hodgkin Lymphoma and SNPs in Histone-pathway Genes

Examining Committee: Dr. Tim Swartz
Chair
Professor

Dr. Jinko Graham
Senior Supervisor
Associate Professor

Dr. John Spinelli
Supervisor
Adjunct Professor/SFU

Dr. Brad McNeney
External Examiner
Associate Professor

Date Approved:

Abstract

Identifying biological pathways affecting cancer susceptibility can provide insight into prevention. In tumour cells of non-Hodgkin lymphoma (NHL), histone-pathway genes are frequent targets of somatic mutation. We analyze data from a population-based case-control study to test whether NHL is associated with SNPs in histone-pathway genes. When individual SNP associations are minor, the standard approach of testing SNPs one-at-a-time and then correcting for multiple testing has low power. Our global testing approach avoids the multiple-testing penalty by modelling random SNP effects and testing variance. We show how an approximate score statistic may be derived by writing the likelihood as an expected conditional likelihood given latent genetic values for each individual and then applying a Taylor-series approximation. The resulting statistic is applied to the NHL data and its statistical significance is evaluated using a permutation-based procedure. Our results add to growing evidence that the histone pathway plays a role in NHL.

Keywords: global test; joint effect; score statistic; Taylor-series approximation; permutation

To my family.

" The journey of a thousand miles begins with one step."

— Lao Tzu

Acknowledgments

I would like to offer sincere gratitude to my supervisor, Dr. Jinko Graham, who has supported and guided me throughout the completion of this thesis with patience and knowledge. I also offer special gratitude to my co-supervisor, Dr. John Spinelli, for his knowledge, advice, and kind encouragement when I needed it most. I would also like to express my sincere gratitude to my committee member, Dr. Brad McNeney, for his valuable suggestions for my thesis. Thank you all for your suggestions and encouragement during the thesis-writing process and my graduate study at Simon Fraser University.

I would also like to acknowledge the faculty and staff in the Department of Statistics and Actuarial Science who have been extremely helpful to me in numerous ways throughout the course of my study.

I would like to give very special thanks to my friends and colleagues, who have given me strength in completing this project through their constant support and friendship. I thank Ji-Hyung (Jean) Shin and Jeong-Eun Min for sharing their knowledge and invaluable assistance.

Finally, I would like to thank my family for your understanding and endless love through the duration of my study.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Background

1.1 Introduction to non-Hodgkin lymphoma (NHL)

Non-Hodgkin lymphoma (NHL) is a type of cancer that occurs from lymphocytes, special blood cells that form part of the immune system. NHL happens when the white blood cells divide out of control or can not undergo normal cell death. The cause of the phenomenon can be found in the nuclei of the cancer cells, where the genome malfunctions. Therefore, NHL, the same as other cancers, is a complex disease which is caused by genetic changes within cells. Single Nucleotide Polymorphisms (SNPs) are the most common genetic change that can occur within a person's DNA sequence. SNPs account for 90% of DNA sequence variation, making them useful genetic markers for disease association studies.

NHL can occur in both B and T lymphocytes that mature in the bone marrow and thymus respectively. NHL develops when one of the lymphocytes, either a B-cell or T-cell, becomes abnormal. B-cell lymphomas are more common than T-cell lymphomas. The two most common types of B-cell lymphomas are diffuse large B-cell lymphoma and follicular lymphoma. There are many other forms of NHL. We will not consider NHL subtypes in this analysis.

While the incidence of most cancers are constant, that of NHL is steadily increasing with an annual rate of $1-2\%$ worldwide. NHL is the fifth most common cancer and the sixth leading cause of all cancer deaths in Canada according to Canadian Cancer Statistics 2011 published by Canadian Cancer Society. Estimated new cases and deaths from NHL in the United States in 2012 are $70,130$ and $18,940$ (National Cancer Institute). The statistics show that the highest incidence for non-Hodgkin lymphoma in the US occurs in whites and men have higher incidence rates than women (Centers for Disease Control and Prevention). Similar to other types of cancer, incidence rates increase with age (Fisher & Fisher, 2004).

1.2 Research question

The causes of NHL have been studied in the past twenty years. Varied environmental and genetic factors have been considered.

Some environment agents have been suspected as possible risk factors for NHL. One of the examples is organochlorine pesticides. Spinelli et al. (2007) investigated polychlorinated biphenyls (PCBs) and organochlorine pesticides and risk of NHL in a population-based case-control study in British Columbia, Canada. Congeners of PCBs and pesticides or pesticide metabolites were measured in plasma of 422 pretreatment cases and 460 control subjects. Several PCB congeners were associated with increased risk of NHL. Six pesticide analytes also showed a significant association with NHL. The study results provide evidence that organochlorines may contribute to NHL risk.

Some genes have been found to be associated with NHL. From the same study, Novik et al. (2007) showed association with a SNP (rs2509049) in gene *H2AFX*, which encodes a histone involved in signalling the presence of double stranded breaks. Morin et al. (2011) discussed the frequent mutation of histone-modifying genes in NHL and linked somatic mutations in the *MLL2* gene to B-cell NHL. Somatic mutations in genes with roles in histone modification were found in the two most common NHLs, Follicular lymphoma (FL) and diffuse large B-cell lymphoma (DLBCL). For example, 32% of DLBCL and 89% of FL cases had somatic mutations in *MLL2*, which encodes a histone transferase enzyme. Somatic mutation is defined as a change in the genetic structure that is neither inherited nor passed to offspring; however, it may provide leads to where to look for germline differences.

The histone pathway is our interest based on the previous studies which show that genes in the histone pathway have been associated with NHL. The histone pathway involves genes regulating the histones, which are proteins in the cell nucleus that package and order the DNA into structural units, playing a role in gene regulation. We are interested in whether the histone pathway as a whole is associated with risk of NHL.

In our study, SNPs in 6 candidate genes from the histone pathway have been genotyped. We test whether the SNPs jointly exert an effect on NHL risk. A standard analysis which tests the SNPs individually rather than jointly is expected to lack power when each SNP in the pathway has a small effect on NHL. As we are interested in the joint effect of the histone SNPs on the risk of NHL, a global test of association is appropriate for our study.

1.3 Comparison of data

The Novik et al manuscript used data from a case-control study of NHL conducted in British Columbia, Canada. All NHL cases aged 20 to 79 diagnosed in British Columbia during the period March 2000 to February 2004 and residing in the greater Vancouver (Greater Vancouver Regional District) and greater Victoria (Capital Regional District) metropolitan areas were invited to participate. HIV-positive cases and cases with prior transplant were excluded. Population controls were identified from the Client Registry of the British Columbia Ministry of Health, a list virtually including all British Columbia residents. The controls were frequency matched to cases by age (within 5-year age group), gender, and residence within the Greater Vancouver Regional District or Capital Regional District. All participants gave written informed consent and provided the information on the ethnicity of his or her four grandparents, demographic information, and medical history.

We used data from the same study; however, our data differs in the following aspects.

- When the Novik et al analysis was conducted, the case-control survey was still in process. Therefore, Novik et al manuscript didn't used the full data set. After the survey completed, the raw data were cleaned. Some SNPs and samples of substandard quality were removed (Schuetz et al., 2012). We used the cleaned data for our study.

- Novik et al investigated 3 SNPs in the gene *H2AFX*, while we investigated 38 SNPs in the histone pathway, including 15 SNPs in the *H2AFX* gene. Novik et al selected 7 SNPs in *H2AFX* based on the variation found in 95 NHL cases after complete resequencing of the gene. Four of the SNPs were discarded for quality reasons. The remaining three SNPs were genotyped for all the cases and controls available at the time. In our data, we have 38 SNPs in the histone pathway among which there are 15 SNPs in the *H2AFX* gene. Only one SNP included in Novik et al. (2007) was included in our cleaned data set. Table 1.1 provides a summary of the SNPs available for our analysis. In our data, there are 6 genes in the histone pathway: *H2AFZ* (chromosome 4), *H2AFX* (chromosome 11), *PRMT5* (chromosome 14), *YY1* (chromosome 14), *RPA1* (chromosome 17), *ZFX* (X chromosome).

– In addition, Novik et al. analyzed all ethnicities dividing the subjects into 4 groups: White Caucasian, Asian, south Asian, and mixed/other/unknown. By contrast, we focus on white subjects only because white is the biggest ethnic group in our data set (about 80% of the samples collected) and no other ethnic group composed more than 10% of the samples. Although we study whites only, we have more subjects than Novik et al. We have 1116 whites, while Novik et al. have 1018 subjects from all ethnic groups.

Both the analysis in Novik et al and ours adjusted for gender and age, as is standard for potential confounding variables. Though controls are frequency-matched to cases on gender and age, we only used about 80% of the samples collected and so gender and age are not perfectly balanced.

Table 1.1: Summary of histone-pathway SNPs in the cleaned data

		Number of SNPs considered	
Gene	Chromosome	In current analysis	In Novik et al analysis
H2AFZ	4	3	0
H2AFX	11	15	1
PRMT5	14	1	0
YY1	14	2	0
RPA1	17	11	0
ZFX	X	6	0
Total		38	1

Chapter 2

Methods

2.1 Motivation

The purpose of our study that is to test whether the group of SNPs in the histone pathway as a whole are associated with NHL. Novik et al. (2007) found a SNP in the histone pathway is significantly associated with NHL. Another study, Morin et al. (2011) discussed the linkage between the mutations in the histone modifying gene and NHL. We wish to test the genetic effects of the SNPs in the histone pathway simultaneously, rather than testing individual SNPs one at a time.

The global test (Goeman et al., 2004) may be applied to test whether a group of SNPs is associated with a clinical outcome. A p-value for the group is obtained instead of a p-value for a single SNP. If the test is significant, the conclusion is that the SNPs in the group are jointly associated with the clinical outcome.

The global test is preferred over multiple testing of marginal SNP effects for several reasons. First, power is improved by avoiding the multiple testing adjustments that are necessary when working with marginal associations of one SNP at a time. Second, the global test has been shown to have optimal power in a local neighborhood of the null hypothesis (Goeman et al., 2006). Third, as a score test, the global test does not require estimation of model parameters that pertain to the alternative hypothesis.

We use an empirical Bayes model for the effects of the m SNPs in the histone pathway (Goeman et al., 2006). An empirical Bayes model is a two-stage hierarchical model. In the model, the observed data are assumed to be generated from a distribution defined by a set of parameters. This set of parameters is considered to be a sample from a prior distribution defined by parameters called hyperparameters. Hyperparameters may be estimated using the observed data or set to certain values based on prior knowledge.

The empirical Bayes model is useful for modelling SNP effects. Under the alternative hypothesis of a pathway associated with NHL, we believe that most SNP effects will be small or negligible (i.e. around zero). Our belief is expressed in a prior distribution for the SNPs with mean 0 and some variance τ^2. Let β be an m-vector of SNPs effects. Then the prior distribution under the alternative hypothesis can be written as,

$$
\begin{aligned}
&E(\beta) = \underline{0} \\
&cov(\beta) = E(\beta\beta^{'}) = \tau^2\Sigma,
\end{aligned}
$$

where $\underline{0}$ is a vector of zeroes, and Σ is the $m \times m$ variance-covariance matrix of the random SNP effects. In general, realistic restriction of the possibilities for the alternative hypothesis leads to gains in power. Imposing this restriction and considering the SNP effects as random with mean $\underline{0}$ leads to

$$
\begin{aligned}
&H_0 : \tau^2 = 0 \\
&H_1 : \tau^2 > 0.
\end{aligned}
$$

The score test of the variance has one degree of freedom while the simultaneous testing of m fixed SNP effects has m degrees of freedom. The reduction in the degrees of freedom should lead to an increase in power, so long as our prior beliefs about the alternative hypothesis are realistic.

2.2 Hypothesis and model

2.2.1 Notation

Since there are n samples and m SNPs in the data set, the variables and coefficients in the model are:

– Y, an $n-$vector of responses (NHL case or control status).

– X, an $n \times m$ design matrix for the SNPs, with x_{ij} being the dosage (number of copies) of the index allele of SNP j for sample i (values $0, 1, 2$).

- Z, an $n \times p$ design matrix for the intercept and the adjustment covariates, Gender and Age, where p is the number of columns of the design matrix for the intercept, Gender, and Age.

- α, the vector of fixed but unknown regression coefficients for the intercept, Gender, and Age.

- β, the vector of random regression coefficients for the m SNPs, which have mean $\underline{0}$ and variance covariance $\tau^2\Sigma$.

2.2.2 Logistic regression model

Two potential confounding variables, Gender and Age, are adjusted for in the model. Let

$$E\left(Y \mid \alpha, \beta\right) \quad = \quad logit^{-1}\left(Z\alpha + X\beta\right),$$

where Age is grouped into four categories (20–49, 50–59, 60–69 and 70+ years), which are approximate quartiles of age following Novik et al. (2007). The 20–49 year age group is taken as the baseline group.

For sample i $(i = 1, ..., n)$,

$$E\left(Y_i|\alpha, \beta\right) = logit^{-1}\left(\sum_{l=1}^{p} z_{il}\alpha_l + \sum_{j=1}^{m} x_{ij}\beta_j\right).$$

The intercept is denoted by α_1. Since Gender and Age are adjustment covariates, their coefficients, α_2 for Gender Female and $\alpha_3, \alpha_4, \alpha_5$ for Age group 2, Age group 3, and Age group 4, are assumed to be non-random, fixed parameters. In the model, the elements of the vector β are assumed to be random effects with mean 0 and variance τ^2. The elements of Y, Y_i, are conditionally independent given the random effects β.

2.2.3 Random linear predictor

Let $r_i = \sum_j x_{ij}\beta_j, i = 1, \ldots, n \;\; j = 1, \ldots, m$ where $\beta_1, \ldots, \beta_m$ are SNP effects that are assumed to be generated from a prior distribution with unknown shape. In classical quantitative genetics, this linear predictor r_i is known as the *genetic value* (Falconer & Mackay, 1996). In an additive quantitative-genetics model of a trait, subject i's trait value is $Y_i = \mu + g_i + \epsilon_i$, where μ is the population mean trait value, $g_i = \sum_j x_{ij}\beta_j$, the β_j's are random effects, and the ϵ_i's are independent normal errors. The random g_i is the genetic value of subject i. A key idea of the classical model is that, under the alternative hypothesis of a genetic association, subjects with similar values of X should have trait values that are more correlated.

With the random linear predictors r_i, the model can be rephrased as:

$$E\left(Y_i|r_i,\alpha\right) = logit^{-1}\left(Z_i\alpha + r_i\right). \tag{2.1}$$

Let r be the vector of r_i's. Since $E(\beta) = \underline{0}$ under H_1 and $r_i = \sum_j x_{ij}\beta_j$, we get $E\left(r_i\right) = 0$; that is, $E\left(r\right) = \underline{0}$. Under H_1, we also have $cov(\beta) = E(\beta\beta') = \tau^2\Sigma$. Thus

$$\begin{aligned} cov(r) &= cov\left(X\beta\right) \\ &= E\left[X\beta\left(X\beta\right)'\right] - E\left(X\beta\right)E\left(X\beta\right)' \\ &= E\left(X\beta\beta'X'\right) - \underline{0} \\ &= XE\left(\beta\beta'\right)X' \\ &= X\tau^2\Sigma X'. \end{aligned} \tag{2.2}$$

In what follows, we assume that $\Sigma = I$ where I is an $m \times m$ identity matrix; i.e., the random effects for the SNPs are uncorrelated with common variance. In fact, linkage disequilirium (LD) amongst the SNPs could call this assumption into question. LD refers to correlation between SNPs. However, the LD is small since tag SNPs are selected. Moreover, we use the assumption only to motivate the form of the test statistic. A permutation test based on the test statistic will be valid, regardless of the form of the alternative hypothesis (or the sample size).

2.2.4 Likelihood

In our modeling, Y is an observed random variable, but the linear predictor r is not observed. The observed-data likelihood is obtained by integrating out the random r_i values:

$$\begin{aligned} L\left(\tau^2,\alpha\right) &= f\left(y|\tau^2,\alpha\right) \\ &= \int f\left(y,r|\tau^2,\alpha\right)dr \\ &= \int f\left(y|r,\tau^2,\alpha\right)g\left(r|\tau^2\right)dr \\ &= \int f\left(y|r,\alpha\right)g\left(r|\tau^2\right)dr \end{aligned}$$

where $g\left(r|\tau^2\right)$ is an unspecified distribution for the random r_i's such that

$$E\left(r|\tau^2=0\right)=0.$$

Since the Y_i's are conditionally independent given the r_i's, we obtain

$$\begin{aligned}
L\left(\tau^2,\alpha\right) &= \int_r f\left(Y|r,\alpha\right) g\left(r|\tau^2\right) dr \\
&= \int_r \prod_i f\left(Y_i \mid r_i,\alpha\right) g\left(r \mid \tau^2\right) dr \\
&= E_r\left[\prod_i f\left(Y_i \mid r_i,\alpha\right)\right],
\end{aligned}$$

where $f\left(Y_i \mid r_i,\alpha\right)$ is the conditional probability mass function (pmf) of Y_i given r_i, $g\left(r \mid \tau^2\right)$ is the density function of r and E_r is the expectation with respect to the marginal density of r.

Let $f\left(Y_i \mid r_i,\alpha\right)=f_i\left(r_i\right)$. The second order Taylor expansion for $L\left(\tau^2,\alpha\right)$ with respect to the vector r at $r=0$ yields

$$\begin{aligned}
L\left(\tau^2,\alpha\right) &= E_r\left[\prod_i f_i\left(r_i\right)\right] \\
&\approx E_r\left[\prod_i f_i\left(0\right)+\sum_i r_i\frac{\partial f_i\left(0\right)}{\partial r_i}\prod_{j\neq i} f_j\left(0\right)\right. \\
&\quad \left.+\frac{1}{2}\left(\sum_i r_i^2\frac{\partial^2 f_i\left(0\right)}{\partial r_i^2}\prod_{j\neq i} f_j\left(0\right)+\sum_i\sum_{j\neq i} r_i r_j\frac{\partial f_i\left(0\right)}{\partial r_i}\frac{\partial f_j\left(0\right)}{\partial r_j}\prod_{k\neq i,j} f_k\left(0\right)\right)\right] \quad (2.3)
\end{aligned}$$

Taking the expectation of equation (2.3), and then substituting several expressions such as the first and second derivatives, $\frac{\partial f_i(r_i)}{\partial r_i}$ and $\frac{\partial^2 f_i(r_i)}{\partial r_i^2}$, of the likelihood function, yields the final expression for $L\left(\tau^2,\alpha\right)$ at the end of the section, derived as follows.

Let $R=\frac{1}{m}XX'$, an $n\times n$ matrix that is proportional to the variance-covariance matrix of r when the β_i's are uncorrelated with common variance because

$$\begin{aligned}
R &= \frac{1}{m}XX' \\
&= \frac{1}{m}\frac{1}{\tau^2}\tau^2 XX' \\
&= \frac{1}{m\tau^2}cov\left(r\right).
\end{aligned}$$

The following expressions are used for the derivation of the final expression for $L\left(\tau^2, \alpha\right)$. Since $E\left(r\right) = \underline{0}$ and $cov\left(r\right) = \tau^2 XX' = \tau^2 mR$, we have $E_r\left(r_i\right) = 0$, $E_r\left(r_i^2\right) = Var\left(r_i\right) = \tau^2 mR_{ii}$, and $E_r\left(r_i r_j\right) = cov(r_i, r_j) = \tau^2 mR_{ij}$. Let $l_i\left(r_i\right) = log\left[f_i\left(r_i\right)\right]$. The approximation to $L\left(\tau^2, \alpha\right)$ can be simplified by substituting expressions for the first and second derivatives of the function, $f_i\left(r_i\right)$, in terms of the derivatives of the log-likelihood function, $l_i\left(r_i\right)$. Since the derivative of the log-likelihood function is

$$\frac{\partial l_i\left(r_i\right)}{\partial r_i} = \frac{1}{f_i\left(r_i\right)} \frac{\partial f_i\left(r_i\right)}{\partial r_i},$$

we get that

$$\frac{\partial f_i\left(r_i\right)}{\partial r_i} = f_i\left(r_i\right) \frac{\partial l_i\left(r_i\right)}{\partial r_i}, \tag{2.4}$$

and

$$\begin{aligned} \frac{\partial^2 f_i\left(r_i\right)}{\partial r_i^2} &= \frac{\partial f_i\left(r_i\right)}{\partial r_i} \frac{\partial l_i\left(r_i\right)}{\partial r_i} + f_i\left(r_i\right) \frac{\partial^2 l_i\left(r_i\right)}{\partial r_i^2} \\ &= f_i\left(r_i\right) \left[\left(\frac{\partial l_i\left(r_i\right)}{\partial r_i}\right)^2 + \frac{\partial^2 l_i\left(r_i\right)}{\partial r_i^2}\right]. \end{aligned} \tag{2.5}$$

The derivatives of the log-likelihood function, $\frac{\partial l_i(r_i)}{\partial r_i}$ and $\frac{\partial^2 l_i(r_i)}{\partial r_i^2}$, are derived in Appendix A.

Taking the expectation of equation (2.3), and then substituting expressions (2.4) and

(2.5), yields

$$
\begin{aligned}
L\left(\tau^2,\alpha\right) \approx\ & \prod_i f_i\left(0\right) + \sum_i E_r\left(r_i\right)\frac{\partial f_i\left(0\right)}{\partial r_i}\prod_{j\neq i} f_i\left(0\right) \\
& + \frac{1}{2}\left[\sum_i E_r(r_i^2)\frac{\partial^2 f_i\left(0\right)}{\partial r_i^2}\prod_{j\neq i} f_j\left(0\right) + \sum_i\sum_{j\neq i} E_r(r_i r_j)\frac{\partial f_i\left(0\right)}{\partial r_i}\frac{\partial f_j\left(0\right)}{\partial r_j}\prod_{k\neq i,j} f_k\left(0\right)\right] \\
=\ & \prod_i f_i\left(0\right) + 0 \\
& + \frac{1}{2}\left[\sum_i \tau^2 m R_{ii}\frac{\partial^2 f_i\left(0\right)}{\partial r_i^2}\prod_{j\neq i} f_j\left(0\right) + \sum_i\sum_{j\neq i}\tau^2 m R_{ij}\frac{\partial f_i\left(0\right)}{\partial r_i}\frac{\partial f_j\left(0\right)}{\partial r_j}\prod_{k\neq i,j} f_k\left(0\right)\right] \\
=\ & \prod_i f_i\left(0\right) + \frac{1}{2}\left[\sum_i \tau^2 m R_{ii}\left\{f_i\left(0\right)\left[\left(\frac{\partial l_i\left(0\right)}{\partial r_i}\right)^2 + \frac{\partial^2 l_i\left(0\right)}{\partial r_i^2}\right]\right\}\prod_{j\neq i} f_j\left(0\right)\right. \\
& \left. + \sum_i\sum_{j\neq i}\tau^2 m R_{ij} f_i\left(0\right)\frac{\partial l_i\left(0\right)}{\partial r_i} f_j\left(0\right)\frac{\partial l_j\left(0\right)}{\partial r_j}\prod_{k\neq i,j} f_k\left(0\right)\right] \\
=\ & \prod_i f_i\left(0\right) + \frac{1}{2}\left[\sum_i \tau^2 m R_{ii}\left[\left(\frac{\partial l_i\left(0\right)}{\partial r_i}\right)^2 + \frac{\partial^2 l_i\left(0\right)}{\partial r_i^2}\right]\prod_j f_j\left(0\right)\right. \\
& \left. + \sum_i\sum_{j\neq i}\tau^2 m R_{ij}\frac{\partial l_i\left(0\right)}{\partial r_i}\frac{\partial l_j\left(0\right)}{\partial r_j}\prod_k f_k\left(0\right)\right] \\
=\ & \prod_i f_i\left(0\right)\left(1 + \frac{1}{2}\tau^2\sum_i m R_{ii}\left[\frac{\partial^2 l_i\left(0\right)}{\partial r_i^2} + \left(\frac{\partial l_i\left(0\right)}{\partial r_i}\right)^2\right]\right. \\
& \left. + \frac{1}{2}\tau^2\sum_i\sum_{j\neq i} m R_{ij}\frac{\partial l_i\left(0\right)}{\partial r_i}\frac{\partial l_j\left(0\right)}{\partial r_j}\right).
\end{aligned} \tag{2.6}
$$

This is the final expression for $L(\tau^2,\alpha)$ which will be used for derivation of global test statistics in the next section.

2.3 Test statistics and derivation

2.3.1 Notation

Notations for the expectation and higher moments of Y_i are defined as follows.

– Let the first conditional moment of Y_i be

$$\begin{aligned}\mu_{1i}(r_i) &= E(Y_i|r_i,\alpha)\\ &= logit^{-1}(Z_i\alpha + r_i)\end{aligned}$$

where α is fixed, Z_i is observed, and r_i is random. Under H_0, we have $r_i = 0$, then

$$\begin{aligned}\mu_{1i}(r_i = 0) &= logit^{-1}(Z_i\alpha + 0)\\ &= logit^{-1}(Z_i\alpha)\\ &= Pr(Y_i = 1|H_0)\\ &= \frac{e^{Z_i\alpha}}{1 + e^{Z_i\alpha}}.\end{aligned}$$

– Let $\mu_{1i} = \mu_{1i}(r_i = 0)$ and μ_1 be the $n-$vector of μ_{1i}'s.

– Let the second, conditional, *centred* moment of Y_i be

$$\mu_{2i}(r_i) = E\left[(Y_i - \mu_{1i}(r_i))^2 |r_i\right].$$

– Under H_0, we have $r_i = 0$, so that

$$\begin{aligned}\mu_{2i}(r_i = 0) &= E\left[(Y_i - \mu_{1i}(r_i))^2 |r_i = 0\right]\\ &= \mu_{1i}(1 - \mu_{1i}).\end{aligned}$$

– Let $\mu_{2i} = \mu_{2i}(r_i = 0)$.

– Similarly, let μ_{ji} be the jth centred moment of Y_i under H_0, for $j = 2, 3, \ldots$.

A score test statistic T was derived for H_0 and then an equivalent test statistic Q was found (Goeman et al., 2004). Since Q is simple in expression, the global test package in R uses it as the test statistic.

2.3.2 The original test statistic

Assuming that the random effects of the SNPs are uncorrelated with equal variances, the score test statistics T turns out to be,

$$T = \frac{(Y - \mu_1)' R (Y - \mu_1) - \text{trace}(RV)}{\left[2\sum_i\sum_j R_{ij}^2\mu_{2i}\mu_{2j} + \sum_i R_{ii}^2\left(\mu_{4i} - 3\mu_{2i}^2\right)\right]^{\frac{1}{2}}}, \tag{2.7}$$

where V is the diagonal matrix with $V_{ii} = \mu_{2i}$. When the SNP effects are uncorrelated with equal variances, the last equality follows from Equation (2.2) with $\Sigma = I$. Under H_0, T is asymptotically normally distributed (Goeman et al., 2004), although we do not rely on these asymptotic results in this project.

2.3.3 Derivation of original test statistic

The derivation follows that given in Le Cessie & Van Houwelingen (1995) and Houwing-Duistermaat et al. (1995). Throughout, we treat the approximation to $L\left(\tau^2, \alpha\right)$ in equation (2.6) as an equality. We assume α is known for the moment and write the likelihood as $L\left(\tau^2\right)$. We come back to this point when discussing the equivalent test statistic in Section 2.3.5.

For a regular score test of the hypothesis

$$H_0 : \tau^2 = 0$$
$$H_1 : \tau^2 > 0$$

in the empirical Bayes model (2.1), the test statistic is in the form

$$T = \left[\frac{\partial logL\left(0\right)}{\partial \tau^2}\right] / \left\{E\left(\left[\frac{\partial logL(0)}{\partial \tau^2}\right]^2\right)\right\}^{\frac{1}{2}},$$

where $L\left(0\right)$ is the likelihood function when $\tau^2 = 0$; i.e. $L(0) = L\left(\tau^2\right)|_{\tau^2=0}$. An overview of the derivation is:

(i) get the numerator, $\frac{\partial logL(0)}{\partial \tau^2}$, of T, using the expression (2.6) for $L\left(\tau^2\right)$.

(ii) get the denominator, $\left\{E\left(\left[\frac{\partial logL(0)}{\partial \tau^2}\right]^2\right)\right\}^{\frac{1}{2}}$, of T, using the expression for the numerator.

Numerator of test statistic

We will show that the numerator of T is:

$$\frac{\partial logL(0)}{\partial \tau^2} = \frac{m}{2}\left[\left(Y - \mu_1\right)' R\left(Y - \mu_1\right) - \text{trace}\left(RV\right)\right]. \tag{2.8}$$

As shown in Appendix A, the first and second partial derivatives of the log-likelihood for a Bernoulli distribution are

$$\frac{\partial l_i(r_i)}{\partial r_i} = Y_i - \mu_{1i}(r_i) \tag{2.9}$$

$$\frac{\partial^2 l_i(r_i)}{\partial r_i^2} = -\mu_{2i}(r_i) \tag{2.10}$$

When $r_i = 0$, we obtain

$$\begin{aligned} \frac{\partial l_i(0)}{\partial r_i} &= Y_i - \mu_{1i} \qquad \text{and} \\ \frac{\partial^2 l_i(0)}{\partial r_i^2} &= -\mu_{2i}, \end{aligned}$$

since we defined $\mu_{1i} = \mu_{1i}(r_i = 0)$ and $\mu_{2i} = \mu_{2i}(r_i = 0)$.

Returning to equation (2.6) for $L(\tau^2)$ and taking the first partial derivatives of $logL\left(\tau^2\right)$ with respect to τ^2 at $\tau^2 = 0$, and then plugging in the derivatives above, yields:

$$\begin{aligned} \frac{\partial logL(0)}{\partial \tau^2} &= \frac{1}{L(\tau^2)} \frac{\partial L(\tau^2)}{\partial \tau^2} \mid_{\tau^2=0} \\ &= \frac{1}{\prod_i f_i(0)} \prod_i f_i(0) \left(\frac{1}{2} \sum_i m R_{ii} \left[\frac{\partial^2 l_i(0)}{\partial r_i^2} + \left(\frac{\partial l_i(0)}{\partial r_i} \right)^2 \right] \right. \\ &\quad \left. + \frac{1}{2} \sum_i \sum_{j \neq i} m R_{ij} \frac{\partial l_i(0)}{\partial r_i} \frac{\partial l_j(0)}{\partial r_j} \right) \\ &= \frac{1}{2} \left\{ \sum_i m R_{ii} \frac{\partial^2 l_i(0)}{\partial r_i^2} + \sum_i \sum_j m R_{ij} \frac{\partial l_i(0)}{\partial r_i} \frac{\partial l_j(0)}{\partial r_j} \right\} \\ &= \frac{m}{2} \left\{ \sum_i R_{ii}(-\mu_{2i}) + \sum_i \sum_j R_{ij}(Y_i - \mu_{1i})(Y_j - \mu_{1j}) \right\} \\ &= \frac{m}{2} \left[\sum_i \sum_j R_{ij}(Y_i - \mu_{1i})(Y_j - \mu_{1j}) - \sum_i R_{ii}\mu_{2i} \right] \\ &= \frac{m}{2} \left[(Y - \mu_1)' R (Y - \mu_1) - \operatorname{trace}(RV) \right], \end{aligned}$$

the numerator of T, the same as expressed in equation (2.8).

Denominator of test statistic

In the numerator of the test statistic given in equation (2.8), the only random quantity is Y. Taking the expectation of the squared numerator of T yields the denominator of T:

$$E\left\{\left[\frac{\partial logL(0)}{\partial\tau^2}\right]^2\right\}^{\frac{1}{2}} = \frac{m}{2}\left[2\sum_i\sum_j R_{ij}^2\mu_{2i}\mu_{2j} + \sum_i R_{ii}^2\left(\mu_{4i} - 3\mu_{2i}^2\right)\right]^{\frac{1}{2}}. \quad (2.11)$$

The derivation of the expression (2.11) is as follows. Let $Q = (Y-\mu_1)^{'} R(Y-\mu_1)$. Since $E\left([Q-E(Q)]^2\right) = E\left(Q^2\right) - [E(Q)]^2$ and $E(Q) = \text{trace}(RV)$, shown later, we have

$$\begin{aligned}
E\left[\left(\frac{\partial logL(0)}{\partial\tau^2}\right)^2\right] &= E\left(\left(\frac{m}{2}\right)^2[Q - \text{trace}(RV)]^2\right) \\
&= \left(\frac{m}{2}\right)^2\left(E\left(Q^2\right) - [\text{trace}(RV)]^2\right) \\
&= \left(\frac{m}{2}\right)^2\left(E\left\{\left[\sum_i\sum_j R_{ij}(Y_i-\mu_{1i})(Y_j-\mu_{1j})\right]^2\right\}\right. \\
&\quad \left. - \left(\sum_i R_{ii}\mu_{2i}\right)^2\right) \\
&= \left(\frac{m}{2}\right)^2\left[E\left\{\left(W^{'}RW\right)^2\right\} - \left(\sum_i R_{ii}\mu_{2i}\right)^2\right], \quad (2.12)
\end{aligned}$$

where $W = (Y_1-\mu_{11}, Y_2-\mu_{12}, ..., Y_n-\mu_{1n})^{'}$ and $W_i = Y_i - \mu_{1i}$.

Under $H_0: \tau^2 = 0$, we have $E(Y_i) = \mu_{1i}$ so that $E(W_i) = 0$. Also,

$$E\left((W_i^k\right) = E\left((Y_i-\mu_{1i})^k\right) = \mu_{ki}$$

where $k = 2, 3, \ldots$. The W_i's are independent since Y_i's are independent. Since $E(W_i) = 0$ and the W_i's are independent, we have

$$E(W_iW_jW_kW_l) = \begin{cases} \mu_{4i}, & \text{if } i=j=k=l \\ \mu_{2i}\mu_{2k}, & \text{if } i=j, k=l, i\neq k \\ \mu_{2i}\mu_{2j}, & \text{if } i=k, j=l \text{ or } i=l, j=k, i\neq j \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, in equation (2.12),

$$\begin{aligned}
E\left[\left(W'RW\right)^2\right] &= E\left(\sum_i\sum_j\sum_k\sum_l R_{ij}R_{kl}W_iW_jW_kW_l\right) \\
&= \sum_i R_{ii}^2\mu_{4i} + \sum_i\sum_{k\neq i} R_{ii}R_{kk}\mu_{2i}\mu_{2k} + 2\sum_i\sum_{j\neq i} R_{ij}^2\mu_{2i}\mu_{2j} \\
&= \sum_i R_{ii}^2\mu_{4i} + \sum_i\sum_k R_{ii}R_{kk}\mu_{2i}\mu_{2k} + 2\sum_i\sum_j R_{ij}^2\mu_{2i}\mu_{2j} - 3\sum_i R_{ii}^2\mu_{2i}^2 \\
&= \sum_i R_{ii}^2\left(\mu_{4i} - 3\mu_{2i}^2\right) + \left(\sum_i R_{ii}\mu_{2i}\right)^2 + 2\sum_i\sum_j R_{ij}^2\mu_{2i}\mu_{2j} \qquad (2.13)
\end{aligned}$$

Substituting equation (2.13) into equation (2.12) yields

$$E\left\{\left[\frac{\partial logL(0)}{\partial\tau^2}\right]^2\right\}^{\frac{1}{2}} = \frac{m}{2}\left[\sum_i R_{ii}^2\left(\mu_{4i} - 3\mu_{2i}^2\right) + 2\sum_i\sum_j R_{ij}^2\mu_{2i}\mu_{2j}\right]^{\frac{1}{2}}.$$

This is the denominator of statistic T as shown in expression (2.11).

Based on expressions (2.8) and (2.11),

$$T = \frac{(Y-\mu_1)'R\,(Y-\mu_1) - \text{trace}\,(RV)}{\left[2\sum_i\sum_j R_{ij}^2\mu_{2i}\mu_{2j} + \sum_i R_{ii}^2\left(\mu_{4i} - 3\mu_{2i}^2\right)\right]^{\frac{1}{2}}},$$

as claimed in equation (2.7).

2.3.4 An equivalent test statistic

The expression for T is complex, making it difficult to calculate. An equivalent test statistic Q in simple form is desirable. The equivalent test statistic is

$$Q = \left(Y-\mu_1\right)' R\left(Y-\mu_1\right).$$

We will show that

$$T = \frac{Q - E\left(Q\right)}{sd\left(Q\right)}$$

by establishing that,

$$E\left(Q\right) = \text{trace}\left(RV\right)$$

and

$$sd\left(Q\right) = \left[2\sum_i\sum_j R_{ij}^2\mu_{2i}\mu_{2j} + \sum_i R_{ii}^2\left(\mu_{4i} - 3\mu_{2i}^2\right)\right]^{\frac{1}{2}} \text{ (the denominator of } T\text{).}$$

As a score statistic, T is asymptotically normally distributed if H_0 is true and hence so is Q. However, we don't use the asymptotic distribution in our study.

To show that $E\left(Q\right) = \text{trace}\left(RV\right)$, we use the fact that, under regularity conditions, scores have mean zero. Hence, under H_0,

$$0 = E\left[\frac{\partial logL(\tau^2)}{\partial\tau^2}\;|_{\tau^2=0}\right] = E\left(\frac{m}{2}\left[Q - \text{trace}\left(RV\right)\right]\right)$$

$$\text{or} \quad E\left(Q\right) = \text{trace}\left(RV\right)$$

The proof for $sd\left(Q\right) =$ the denominator of T is based on the definition of T. From the derivation of T, we know that

$$\begin{aligned} T &= \frac{score}{sd\left(score\right)} \\ &= \left[\frac{\partial logL\left(0\right)}{\partial\tau^2}\right] / \left\{E\left(\left[\frac{\partial logL(0)}{\partial\tau^2}\right]^2\right)\right\}^{\frac{1}{2}} \\ &= \frac{\frac{m}{2}\left[(Y-\mu_1)'R(Y-\mu_1) - \text{trace}(RV)\right]}{\frac{m}{2}\left[2\sum_i\sum_j R_{ij}^2\mu_{2i}\mu_{2j} + \sum_i R_{ii}^2\left(\mu_{4i} - 3\mu_{2i}^2\right)\right]^{\frac{1}{2}}}, \end{aligned}$$

where

$$sd\left(score\right) = \frac{m}{2}\left[2\sum_i\sum_j R_{ij}^2\mu_{2i}\mu_{2j} + \sum_i R_{ii}^2\left(\mu_{4i} - 3\mu_{2i}^2\right)\right]^{\frac{1}{2}}$$

and

$$\begin{aligned} score &= \frac{m}{2}\left[\left(Y-\mu_1\right)' R\left(Y-\mu_1\right) - \text{trace}\left(RV\right)\right] \\ &= \frac{m}{2}\left[Q - E\left(Q\right)\right]. \end{aligned}$$

Taking the standard deviation of the score, we get

$$\begin{aligned} sd\,(score) &= sd\left(\frac{m}{2}\left[Q - E\,(Q)\right]\right) \\ &= \frac{m}{2} sd\left[Q - E\,(Q)\right] \\ &= \frac{m}{2} sd\,(Q) \\ &= \frac{m}{2}\left[2\sum_i\sum_j R_{ij}^2 \mu_{2i}\mu_{2j} + \sum_i R_{ii}^2\left(\mu_{4i} - 3\mu_{2i}^2\right)\right]^{\frac{1}{2}}. \end{aligned}$$

Therefore,

$$sd\,(Q) = \left[2\sum_i\sum_j R_{ij}^2 \mu_{2i}\mu_{2j} + \sum_i R_{ii}^2\left(\mu_{4i} - 3\mu_{2i}^2\right)\right]^{\frac{1}{2}}.$$

2.3.5 Calculation of test statistic

The derivation of T and hence Q assumes that the regression coefficient α is known, and therefore the mean $\mu_1 = logit^{-1}\,(Z\alpha)$ under H_0 is known. In practice, α is estimated from the observed data.

Let $\hat{\alpha}$ be the maximum likelihood estimate for the model under $H_0 : \tau^2 = 0$. Then the estimate of the expectation of Y under H_0 is:

$$\hat{\mu}_1 = logit^{-1}\,(Z\hat{\alpha})\,.$$

With the insertion of the estimate, $\hat{\mu}_1$, the test statistic is

$$\begin{aligned} Q_{obs} &= \left(Y - \hat{\mu}_1\right)' R\,(Y - \hat{\mu}_1) \\ &= \frac{\left(Y - logit^{-1}\,(Z\hat{\alpha})\right)' XX' \left(Y - logit^{-1}\,(Z\hat{\alpha})\right)}{m} \end{aligned}$$

2.4 Permutation null distribution

Goeman et al. (2004) used Q as the test statistic and evaluated the empirical Bayes model null hypothesis using asymptotic distribution of Q. However, we prefer the permutation test over the asymptotic test because the finite sample properties of the asymptotic test have not been established (Goeman et al., 2006).

Permuting outcomes is consistent with the null hypothesis that none of the SNPs are associated with NHL. Under the null hypothesis, the probability of NHL only depends on the adjustment covariates, Age and Gender. Therefore, within strata, the outcomes are exchangeable under the null hypothesis. By randomly shuffling the outcomes we can make up as many permuted data sets as we like. Thus, the accuracy of the p-value is only restricted by the computing time. In our context, we stratify on the adjustment covariates, Age and Gender. There are eight combinations of these two covariates as summarized in Table 2.1. The random shuffling of outcome values is conducted within each combination. Permuting outcome within a stratum of age and gender preserves the dependence amongst the SNPs (i.e. the LD). Within strata, the outcomes are exchangeable under the null hypothesis of no association with any of the SNPs.

After repeating this algorithm many times, the resulting Q's will form a reference distribution close to the true permutation distribution of the test statistic. The p-value is obtained from this distribution. Specifically, the p-value of the test is calculated as the proportion of Q's based on the permutation data that are greater than or equal to the Q based on the original data.

Before we conduct the permutation test, we need to impute the missing genotypes which is discussed in the next chapter.

Table 2.1: Frequency of samples by age and gender

	Age Group (years)				
	(20-49)	(50-59)	(60-69)	(70+)	Total
Males	120	129	183	175	607
Females	86	131	134	158	509
Total	206	260	317	333	1116

Chapter 3

Genotype Imputation

3.1 Data

As stated in chapter one, in our cleaned data there are 1116 whites and 1286 SNPs including 38 SNPs in the histone pathway. The study samples in the cleaned data are summarized in Table 3.1. From the table, we see that the data are well balanced between cases and controls for different gender and age groups. As some of the genotypes are missing for the 38 SNPs in the histone pathway, we impute the genotype data, as described next.

Table 3.1: Summary of the case-control samples

	Cases (%)	Controls (%)	Total
Gender			
Female	241(47%)	268(53%)	509
Male	328(54%)	279(46%)	607
Age group (years)			
20-49	87(42%)	119(58%)	206
50-59	138(53%)	122(47%)	260
60-69	165(52%)	152(48%)	317
70+	179(54%)	154(46%)	333
Total	569(51%)	547(49%)	1116

3.2 Motivation

The global test is a joint analysis requiring complete data for all SNPs. The global test R package can process an incomplete data set by removing all the samples with missing genotypes. If the cleaned data with 1116 whites and the 38 SNPs are tested directly by the R package, 203 samples will be lost, including the one found by Novik et al. (2007).

To preserve power, we used Beagle V3.3 (Browning & Browning, 2009) to impute sporadically missing genotypes. The approach incorporates the information from surrounding markers to increase the quality of the imputation by making use of the concept of a haplotype. A haplotype is a set of SNPs on a single chromosome of a chromosome pair. The alleles of the SNPs are statistically associated in the population. These associations and the identification of a few alleles of a haplotype block can help identify other unknown alleles in the genomic region. The Beagle imputation process jointly models the observed genotype data to infer missing genotypes. Very briefly, a hidden Markov model (HMM) is applied, in which the haplotype phase is the hidden state and the observed genotypes are the observed data. The expectation-maximization (EM) algorithm is used to fit the HMM parameters by maximizing the likelihood (Browning & Browning, 2009).

We chose Beagle because it is among the most accurate programs for genotype imputation. Marchini & Howie (2010) compared the computational performance and error rate of the most popular imputation methods including IMPUTE, MACH, fastPHASE, and Beagle, and found Beagle to be comparable to the others.

3.3 Illustration of basic idea

In this section, we show how SNPs in low to moderate LD can provide information about a missing genotype.

In our cleaned data set, we randomly selected a SNP with one missing genotype and examined the relationship between the imputed posterior probability of its missing genotype and the estimated population haplotype frequencies for surrounding SNPs. A SNP rs1845558 in gene *UGT2B4* on chromosome 4 was selected. The genotype of this SNP was missing for sample 04-1981.

The SNP and 4 neighbor SNPs are on chromosome 4 in genes *UGT2B7* and *UGT2B4* as shown in Table 3.2. These SNPs are in low LD ($R^2 < 0.5$). That means these SNPs are slightly correlated. For sample 04-1981, the genotypes of these 5 SNPs were complete except the SNP rs1845558 in the middle (See Table 3.3). This person must have haplotypes GA_AG and GA_CG.

Based on the cleaned data, we found that other than the SNP rs1845558 missing for sample 04-1981, the SNP rs1826690 has 3 samples missing (sample 03-1295, sample 03-1333, sample 03-1359W). In total, there are 4 incomplete samples for the 5 SNPs. The Beagle output provides phased haplotypes for all 1116 samples based on the haplotype pair with maximum posterior probability for the individual. To obtain the population haplotype frequencies shown in Table 3.4, I used the haplotype phasing with maximum posterior probability in the 1112 complete samples. Specifically, I treated these maximum *a posteriori* haplotypes as known and estimated their relative frequencies by the appropriate proportions.

In Table 3.4, there are two possible haplotypes for GA_AG, which are $GACAG$ and $GAGAG$, and they account for an estimated 16.1% and 5.3% of population haplotypes respectively. There is only one possible haplotype for GA_CG, which is $GACCG$. Therefore, it is highly likely (with chance $\frac{16.1\%}{16.1\%+5.3\%} = 75.2\%$) that the missing genotype for rs1845558 would be CC. Table 3.5 shows the Beagle output on the posterior probability of genotypes for the SNP. The posterior probability of being CC is high (75.89%), which is consistent with the above analysis. Thus, even though these 5 SNPs are in low LD, taken together as a haplotype, they can still provide good information about imputing a missing genotype (i.e. reasonable imputation certainty).

Table 3.2: Tag SNPs selected in low LD (allelic $R^2 < 0.5$)

SNP	Chromosome	Position	Assignment	Gene
rs4356975	4	70007052	A/G	*UGT2B7*
rs1826690	4	70386855	A/G	*UGT2B4*
rs1845558	4	70388122	C/G	*UGT2B4*
rs17671289	4	70393434	A/C	*UGT2B4*
rs13145834	4	70393657	A/G	*UGT2B4*

In summary, the basic idea of imputation using Beagle is that the genotypes of missing SNPs are inferred based on the correlation pattern (LD) of the surrounding markers in the population. Even tag SNPs in relatively low LD, when taken together as a haplotype, can provide sufficient information to impute a missing genotype with reasonable certainty.

Table 3.3: Genotypes for the 5 SNPs of Table 3.2 in sample 04-1981

rs4356975	rs1826690	rs1845558*	rs17671289	rs13145834
G	A	NA	A	G
G	A	NA	C	G

* This sample was missing a genotype at rs1845558.

Table 3.4: Estimated distribution of haplotypes

Haplotype	Number	Frequency (%)
GGGAG	471	21.2
GACAG	359	16.1
AAGAA	341	15.3
GACCG	321	14.4
GAGAA	231	10.4
AACCG	188	8.5
GAGAG	117	5.3
AGGAG	84	3.8
AACAG	76	3.4
AAGAG	19	0.9
AGGCG	15	0.7
AGCAG	1	0.0
GGGCG	1	0.0
Total	2224	100

3.4 Implementation

Although we need the complete data for the 38 SNPs in the histone pathway, we impute the cleaned data set with all 1288 SNPs first and then select the SNPs in the histone pathway from the imputation output. The cleaned data set with all SNPs provides a better reference for the missing genotypes of the 38 SNPs than a data set consisting of only the 38 SNPs, resulting in better imputation quality. Therefore, the imputation is based on 1116 whites and 1288 SNPs. Imputation was conducted for each chromosome separately. In total, 23 chromosomes were considered including the X chromosome. Imputation for SNPs on the X chromosome is a special case and we will return to this point later. The Beagle output files from different chromosomes were combined to obtain the complete data set for downstream analysis of SNPs in the histone pathway.

Our single imputation data were based on "expectation-substitution"(Jiao et al., 2011) wherein the missing genotype value is filled in with the estimated posterior mean value

Table 3.5: Posterior probabilities for the missing genotype at rs1845558 in sample 04-1981

Genotype	CC	CG	GG
Probability (%)	75.89	24.11	0.01

of the dosage at the SNP, given the data at other SNPs. For example, let the estimated posterior genotype probabilities for an individual with a missing genotype be P(AA), P(AB) and P(BB) where B is the index allele. Then their estimated posterior mean genotype value of the dosage at the SNP is $0 \times P(AA) + 1 \times P(AB) + 2 \times P(BB) = P(AB) + 2P(BB)$. For our example with sample 04-1981, the imputed genotype dosage value by expectation substitution is $P(AB) + 2P(BB) = 2 \times 0.2411 + 0.01 = 0.4911$.

We also generate 5 multiple imputation data sets. In a multiple imputation, a missing genotype value is filled in by sampling from possible genotypes based on the estimated posterior probabilities. For example, using an individual's estimated posterior genotype probabilities P(AA), P(AB) and P(BB) for a missing genotype, we would sample the value 0 with probability P(AA), the value 1 with probability P(AB), and the value 2 with probability P(BB).

The imputation quality of a SNP is measured by allelic R^2. The allelic R^2 value of a SNP is the estimated squared correlation between the allele dosage with the highest posterior probability and the true allele dosage for the marker. The true allele dosage is not observed but Beagle gives the estimated posterior probabilities for the true genotype. For one particular SNP, let X be the unobserved true genotypes for all samples, coded as 0 for AA, 1 for AB, and 2 for BB, and Z be the genotypes with highest posterior probability, coded as for X. The expression for the allelic R^2 is

$$R^2 = \frac{Cov\left(X,Z\right)^2}{Var\left(X\right)Var\left(Z\right)}. \tag{3.1}$$

Larger values indicate more accurate genotype imputation (Browning & Browning, 2009). The high correlation between the most likely genotype and the expected genotype suggests that the maximum posterior probability is very large (i.e. close to 1), making the most likely genotype to be the true genotype with high certainty. To minimize the variability in our results due to imputation, we only retained SNPs with allelic R^2 value of 90% or higher.

The X chromosome is different from autosomal chromosomes. Males carry only one copy of the X chromosome (that is, they are hemizygous), in contrast to the two copies carried by females. Beagle 3.3 does not automatically impute sporadic missing genotypes

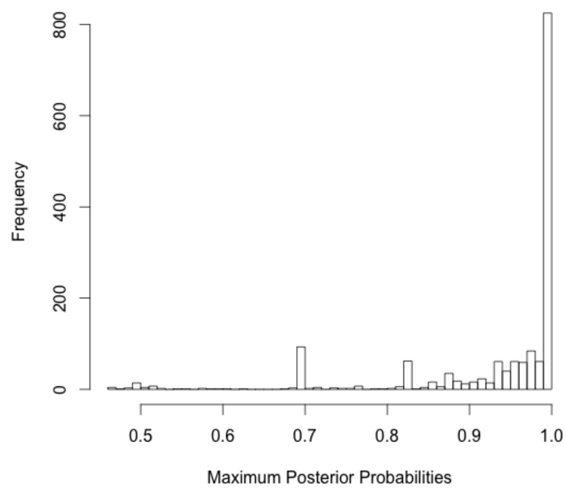

Figure 3.1: Histogram of 1,568 maximum posterior probabilities

on the X chromosome due to the special condition on males. We discuss our method of X chromosome imputation in Appendix B.

Table 3.6: Summary of the distribution of 1,568 maximum posterior probabilities

Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
0.4609	0.9325	0.9930	0.9362	0.9994	1.0000

3.5 Imputation results

We retained 35 of the 38 SNPs in the histone pathway, and all 1116 subjects. Table 3.8 summarizes the results.

Two autosomal SNPs (rs28990980 and rs1042897) with low imputation certainty (allelic $R^2 < 90\%$) were discarded. A third X-chromosome SNP (rs5949211) with allelic R^2 value of 96.4% was also discarded because it had males with high posterior probability of being heterozygous. The X chromosome SNP rs5949211 had 98 males with missing genotypes. As shown in Figure 3.2, five of these 98 males has a high posterior probability (0.4997) of being

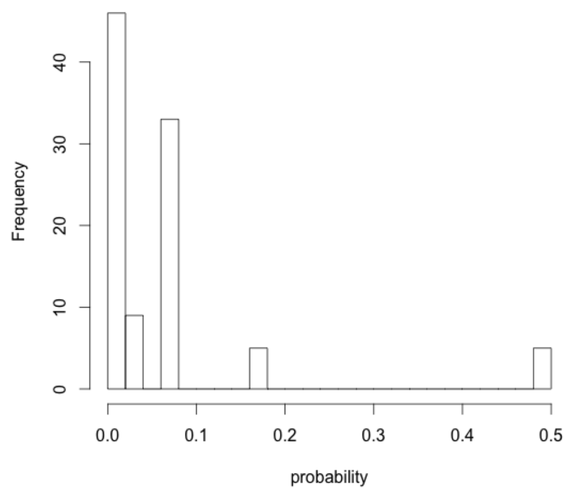

Figure 3.2: Distribution of the estimated posterior probabilities of being heterozygous at rs5949211 in males

a heterozygote. As summarized in Table 3.7, no other X chromosome SNP had males with such high posterior probabilities for being a heterozygote. We therefore decided to remove rs5949211 from the analysis. Thus, the imputation based on Beagle not only kept all the subjects in our dataset but also kept the loss of SNPs to a minimum. This helps to improve the power of our association analysis.

To verify that Beagle provides a reasonable imputation of our data, we check the proportion of missing data and the histogram of the maximum posterior probabilities. There are only 3.7% of genotypes missing. Figure 3.1 and Table 3.6 for 38 SNPs show that most of the posterior genotype probabilities are larger than 80%. The high proportion of imputed genotypes with high certainty and the low proportion of missing data suggest that single-imputation is reasonable for our analysis.

However, we investigate the variability in results due to imputation in the next chapter.

Table 3.7: Estimated posterior probabilities of being heterozygous for X-chromosome SNPs in males with missing genotypes

	Marker	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.	n
1	rs6526373	0.0908	0.0908	0.0908	0.0908	0.0908	0.0908	2
2	rs2704849	NA	NA	NA	NA	NA	NA	0
3	rs17312136	0.0369	0.03705	0.0372	0.0372	0.03735	0.0375	2
4	rs5949211	0	0.0018	0.03885	0.06175	0.0664	0.4997	98
5	rs5990013	NA	NA	NA	NA	NA	NA	0
6	rs6629824	NA	NA	NA	NA	NA	NA	0
7	rs17324671	0.1826	0.1826	0.1826	0.1826	0.1826	0.1826	2
8	rs6634993	0.1929	0.1929	0.1929	0.1929	0.1929	0.1929	1

Table 3.8: Summary of SNPs in the histone pathway

No.	SNP	Chr.	Position	Gene	Missing.Rate	Allelic.R2	Kept
1	rs3756087	4	101090954	H2AFZ	0.0054	0.994	TRUE
2	H2AZ-IVS4(-74)-AG	4	101093998	H2AFZ	0.0000	1.000	TRUE
3	H2AZ-DWN(+4)-CT	4	101094509	H2AFZ	0.0009	0.999	TRUE
4	rs673768	11	118425663	H2AFX	0.0036	0.995	TRUE
5	rs1804690	11	118427410	H2AFX	0.0000	1.000	TRUE
6	rs3825061	11	118449885	H2AFX	0.0009	0.999	TRUE
7	rs494048	11	118466441	H2AFX	0.1496	0.979	TRUE
8	rs28990986	11	118468501	H2AFX	0.0018	0.997	TRUE
9	rs640603	11	118469540	H2AFX	0.1496	0.976	TRUE
10	rs28990980	11	118471332	H2AFX	0.1487	0.842	FALSE
11	rs2509049	11	118471731	H2AFX	0.1487	0.998	TRUE
12	rs7759	11	118472501	H2AFX	0.1487	1.000	TRUE
13	rs8551	11	118472734	H2AFX	0.0036	1.000	TRUE
14	rs643788	11	118472968	H2AFX	0.0009	1.000	TRUE
15	rs604714	11	118475906	H2AFX	0.0009	1.000	TRUE
16	rs603826	11	118476088	H2AFX	0.0108	0.998	TRUE
17	rs649870	11	118476461	H2AFX	0.0000	1.000	TRUE
18	rs571445	11	118493181	H2AFX	0.0009	1.000	TRUE
19	rs8007089	14	22461753	PRMT5	0.0000	1.000	TRUE
20	rs4905941	14	99795191	YY1	0.0000	1.000	TRUE
21	rs1042897	14	99818376	YY1	0.1487	0.810	FALSE
22	rs2287321	17	1703101	RPA1	0.0018	1.000	TRUE
23	rs2287320	17	1703387	RPA1	0.0009	0.997	TRUE
24	rs36088524	17	1716950	RPA1	0.0009	1.000	TRUE
25	rs2277694	17	1730562	RPA1	0.0000	1.000	TRUE
26	rs17338990	17	1733770	RPA1	0.0000	1.000	TRUE
27	rs2270412	17	1738924	RPA1	0.0018	0.998	TRUE
28	rs2230931	17	1741930	RPA1	0.1505	0.967	TRUE
29	rs7406062	17	1746967	RPA1	0.0018	1.000	TRUE
30	rs1131636	17	1747939	RPA1	0.0036	0.999	TRUE
31	rs17339382	17	1748622	RPA1	0.0000	1.000	TRUE
32	rs17339395	17	1749251	RPA1	0.1523	0.999	TRUE
33	rs6526373	X	24076118	ZFX	0.0072	0.998	TRUE
34	rs2704849	X	24085448	ZFX	0.0000	1.000	TRUE
35	rs17312136	X	24100898	ZFX	0.0018	1.000	TRUE
36	rs5949211	X	24143025	ZFX	0.1550	0.964	FALSE
37	rs5990013	X	24144295	ZFX	0.0018	0.999	TRUE
38	rs6629824	X	24146559	ZFX	0.0018	0.999	TRUE

Chapter 4

Results and Conclusion

4.1 Permutation test

We permuted the imputed data as described in section 4 of chapter 2. Then we applied the permutation test to find the p-value using the global test statistic Q. We also examined the variability of test results due to imputation.

The p-value for a global test of the association between NHL and SNPs in histone-pathway genes is 0.0154. Hence, the group of SNPs in the histone pathway genes is significantly associated with NHL. Figure 4.1 shows the approximate permutation distribution of the test statistic Q based on $10,000$ permutation replicates, with the value of Q observed for our data set imputed by expectation substitution marked on the horizontal axis.

We then looked at test statistic values and p-values across 5 multiply-imputed data sets. Table 4.1 shows that the test statistics are very similar across the multiple imputations. Therefore, single imputation by expectation substitution appears to be reliable in our study, as the test statistics and p-values do not vary greatly across the imputed data sets.

Table 4.1: Test statistics and p-values for permutation test

	ES*	Rep1	Rep2	Rep3	Rep4	Rep5
Test statistic	283.46	295.29	298.21	296.86	294.19	292.40
P-value	0.0154	0.0171	0.0162	0.0165	0.0176	0.0178

* ES: expectation substitution

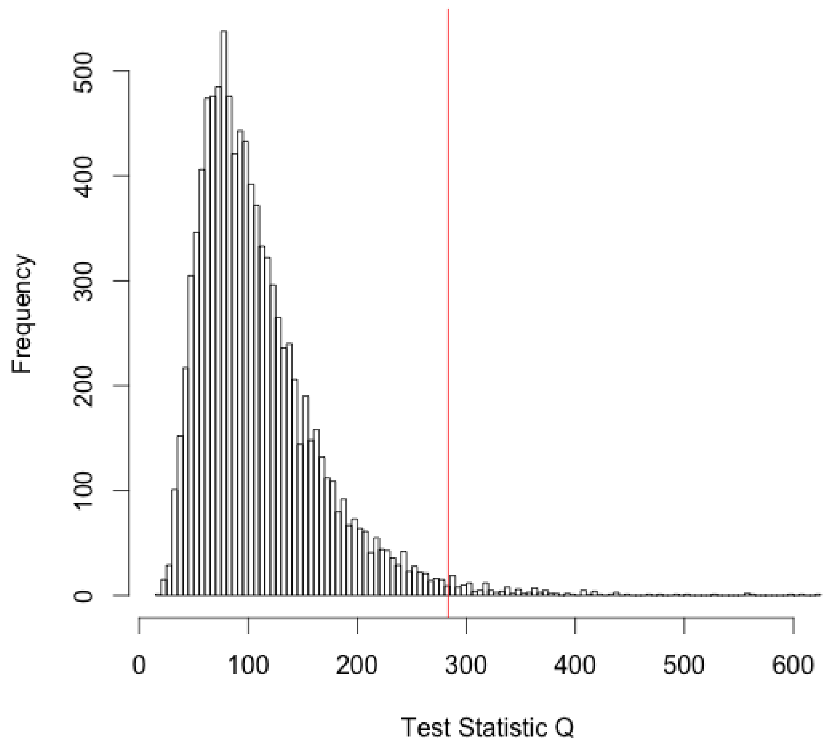

Figure 4.1: Permutation distribution for the test statistic Q

4.2 Conclusions and future work

Single SNPs or genes in the histone pathway have been found to be associated with NHL. In this project, our interest was in whether the SNPs in histone-pathway genes are, as a group, associated with NHL. Because we were interested in a group association, we used a global test. We modelled the effects of SNPs from genes in the histone pathway as random and then tested their variance with a score test. As a score test, the global test is the locally most powerful test. When the SNP effects are close to zero, we therefore expect the global test to have improved power. In particular, we expect improved power over the standard approach which tests SNP associations one-at-a-time. The standard approach requires larger effect sizes to overcome the multiple-testing penalty.

We derive the score statistic from the likelihood, by writing the likelihood as an expected value of a conditional likelihood given latent random genetic values for each individual. Calculation of the score statistic is enabled by approximating the conditional likelihood with a Taylor series expansion in a neighborhood of the null hypothesis. This score statistic is then re-expressed in a simpler form that is more practical to compute. A permutation-based procedure is applied to assess the statistical significance of the association between NHL and SNPs in genes of the histone pathway. To preserve power, sporadically missing genotypes are imputed once by substituting their posterior expected value (expectation substitution) then five times by randomly sampling from their posterior distributions.

Our permutation test applied to the data set imputed with posterior expected values was significant (p-value= 0.0154). Moreover, test results were similar across five multiply-imputed data sets. The similarity of test results across the multiple-imputation data sets indicates low imputation uncertainty. Hence, our results based on the data imputed by expectation substitution are reliable.

Novik et al. (2007) found an association between NHL and a SNP in the gene $H2AFX$ of the histone pathway. In future work, it would be of interest to see if any of the SNPs in $H2AFX$ contribute to the histone pathway association we have found with NHL. There are software tools within the global test R package that would enable such an investigation. The global test has the flexibility to test groups of SNPs of different sizes. This flexibility allows the R package to decompose the data set to investigate the contribution to association of different subsets of SNPs, including single SNPs.

The global test R package also offers a function to find the subjects that have an overly large influence on the test result. The test result can be displayed in a plot that illustrates the relative contributions of the subjects. In future work, it would be of interest to apply these tools to obtain a better idea of how much each SNP contributes to the pathway association with NHL, and also which subjects (e.g. from a particular subtype of NHL) contribute.

Appendix A

Derivatives of log-likelihood function

This Appendix is to derive the derivatives of log-likelihood function with respect to r_i. These derivatives are used in Section 2.2.4.

We have $E\left(Y_i|r_i\right) = \mu_{1i}\left(r_i\right) = logit^{-1}\left(Z_i\alpha + r_i\right) = \frac{e^{z_i\alpha+r_i}}{1+e^{z_i\alpha+r_i}}$. Since $Y_i|r_i$ has a Bernoulli distribution with mean $\mu_{1i}\left(r_i\right)$, we obtain

$$
\begin{aligned}
l_i\left(r_i\right) &= logf\left(Y_i|r_i,\tau^2\right) \\
&= log\left(\left[\mu_{1i}\left(r_i\right)\right]^{Y_i}\left[1-\mu_{1i}\left(r_i\right)\right]^{1-Y_i}\right) \\
&= Y_i log\left(\frac{\mu_{1i}\left(r_i\right)}{1-\mu_{1i}\left(r_i\right)}\right) + log\left(1-\mu_{1i}\left(r_i\right)\right) \\
&= Y_i\left(Z_i\alpha + r_i\right) - log\left(1+e^{z_i\alpha+r_i}\right).
\end{aligned}
$$

Hence

$$
\begin{aligned}
\frac{\partial l_i\left(r_i\right)}{\partial r_i} &= Y_i - \frac{1}{1+e^{Z_i\alpha+r_i}}e^{Z_i\alpha+r_i} \\
&= Y_i - \mu_{1i}\left(r_i\right)
\end{aligned}
$$

and

$$\begin{aligned}
\frac{\partial^2 l_i\left(r_i\right)}{\partial {r_i}^2} &= -\frac{\partial}{\partial r_i}\mu_{1i}\left(r_i\right) \\
&= -\left[\frac{1}{1+e^{Z_i\alpha+r_i}}\frac{\partial}{\partial r_i}e^{Z_i\alpha+r_i}+e^{Z_i\alpha+r_i}\frac{\partial}{\partial r_i}\left(\frac{1}{1+e^{Z_i\alpha+r_i}}\right)\right] \\
&= -\left[\frac{1}{1+e^{Z_i\alpha+r_i}}e^{Z_i\alpha+r_i}+e^{Z_i\alpha+r_i}\left(-1\right)\left(\frac{1}{1+e^{Z_i\alpha+r_i}}\right)^2 e^{Z_i\alpha+r_i}\right] \\
&= -\left[\mu_{1i}\left(r_i\right)-\left(\mu_{1i}\left(r_i\right)\right)^2\right] \\
&= -\mu_{1i}\left(r_i\right)\left[1-\mu_{1i}\left(r_i\right)\right] \\
&= -var\left(y_i|r_i\right) \\
&= -\mu_{2i}\left(r_i\right)
\end{aligned}$$

Appendix B

X chromosome imputation

Beagle 3.3 does not automatically impute sporadic missing genotypes on the X chromosome. Since females have two X chromosomes while males have only one X chromosome (i.e. are hemizygous), we have to impute males and females separately. Beagle can process two files for each gender in a single run.

The output of Beagle for autosomal chromosomes includes a log file, a phased file, a genotype probabilities file, a genotype dosage file, and an allelic R^2 file. The phased file gives the haplotype pair with the highest posterior probability for each sample conditional upon the genotypes for the sample and the haplotype frequency model. The genotype probabilities file gives three columns for each sample indicating the estimated posterior probabilities, P(AA), P(AB), and P(BB), that the true genotype is AA, AB, or BB, respectively, where B is the index allele. The genotype dosage file provides the estimated B-allele dosage $(0 \times P(AA) + 1 \times P(AB) + 2 \times P(BB))$ for each SNP for all samples. The genotype dosage is used in our logistic regression model for the association study.

When running Beagle, we can either specify the males as haploid or as diploid. If we specify males as haploid, the dosage file is the only Beagle output file for males. However, to evaluate the imputation quality of the SNPs, we need the posterior probability file to obtain Z in equation (3.1) for the allelic R^2. Females have all the output files as for imputation on autosomal chromosomes when males are specified as haploid. In the contrast, if we specify the males as diploid, both females and males have all the files, such as dosage file, posterior probability file, allelic R^2 file, etc. However, some males with missing genotype data may be imputed as heterozygous which is impossible.

As recommended (Brian Browning, personal communication), we used the output files for females when males are specified as haploid and the output files for males when males are specified as diploid. When males are treated as haploid, the imputation results should be more reliable than when they are treated as diploid. Unfortunately, at this time it is not possible to get the necessary output files for males when they are treated as haploid. Hence we are forced to treat the males as diploid to get the imputation results for males.

Next, we show how to find the combined allelic R^2 for each X-chromosome SNP across all samples as in equation (3.1).

Let Y be the vector of estimated posterior genotype probabilities for a random sample, i.e.

$$Y = (P(AA), P(AB), P(BB)).$$

We approximate the variance of Z and X in formula (3.1) by using the sample mean, i.e.

$$\begin{aligned} Var\left(Z\right) &= E\left[Z^2\right] - \left(E\left[Z\right]^2\right) \\ &\approx \frac{1}{n}\sum_i Z_i^2 - \frac{1}{n^2}\left(\sum_i Z_i\right)^2 \end{aligned}$$

and

$$\begin{aligned} Var\left(X\right) &= E\left[X^2\right] - \left(E\left[X\right]^2\right) \\ &= E\left[E\left[X^2|Y\right]\right] - \left(E\left[E\left[X|Y\right]\right]\right)^2 \\ &\approx \frac{1}{n}\sum_i E\left[X_i^2|Y_i\right] - \frac{1}{n^2}\left(\sum_i E\left[X_i|Y_i\right]\right)^2, \end{aligned}$$

where $i = 1, ..., n$ for sample i. Similarly, we can estimate covariance of X and Z as

$$\begin{aligned} Cov\left(X, Z\right) &= E\left[XZ\right] - E\left[X\right]E\left[Z\right] \\ &= E\left[E\left[XZ|Y\right]\right] - E\left[E\left[X|Y\right]\right]E\left[Z\right] \\ &\approx \frac{1}{n}\sum_i \left(E\left[X_i|Y_i\right]Z_i\right) - \frac{1}{n^2}\left(\sum_i E\left[X_i|Y_i\right]\right)\sum_i Z_i. \end{aligned}$$

In females, the first and second moments of X conditional on Y are estimated as follows:

$$\begin{aligned} E\left[X|Y\right] &= 0 \times P\left(AA\right) + 1 \times P\left(AB\right) + 2 \times P\left(BB\right) = P\left(AB\right) + 2P\left(BB\right), \\ E\left[X^2|Y\right] &= 0 \times P\left(AA\right) + 1^2 \times P\left(AB\right) + 2^2 \times P\left(BB\right) = P\left(AB\right) + 4P\left(BB\right). \end{aligned}$$

Moreover,

$$Z = \begin{cases} 0, & \text{if } max\left(Y\right) = P\left(AA\right) \\ 1, & \text{if } max\left(Y\right) = P\left(AB\right) \\ 2, & \text{if } max\left(Y\right) = P\left(BB\right). \end{cases}$$

In males, the calculation is not as straight forward, because males must have either an A only or a B only "genotype"for any SNPs on the X chromosome. We approximate the posterior probabilities for males as

$$\begin{aligned} P(A) &= P(AA) + \frac{P(AB)}{2} \\ P(B) &= P(BB) + \frac{P(AB)}{2} \end{aligned}$$

and obtain

$$\begin{aligned} E\left[X|Y\right] &= 0 \times P\left(A\right) + 1 \times P\left(B\right) = P\left(B\right) \\ E\left[X^2|Y\right] &= 0 \times P\left(A\right) + 1^2 \times P\left(B\right) = P\left(B\right). \end{aligned}$$

We set

$$Z = \begin{cases} 0, & \text{if } P\left(A\right) > P\left(B\right) \\ 1, & \text{if } P\left(A\right) < P\left(B\right). \end{cases}$$

Substituting these approximate conditional moments and the values of Z into the expressions above for $Var(Z)$, $Var(X)$, and $cov(X,Z)$, and then applying equation (3.1), we obtain an approximation to the allelic R^2 for each SNP across all samples.

The dosage of X-chromosome SNPs in males is different from the autosomal SNPs becauese it ranges from 0 to 1, rather than 0 to 2. We chose to double the allele dosages of SNPs on the X chromosome for males in order to keep them in the range from 0 to 2. Then we merged the final dosage file for SNPs on the X chromosome with the dosage files for the other SNPs. The dosage data are used for later analysis.

Bibliography

Browning, B. L. & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, **84**(2), 210 – 223.

Centers for Disease Control and Prevention (2012). *Cancer Data and Statistics.* Available from: `http://www.cdc.gov/cancer/dcpc/data/`. Accessed July 2012.

Falconer, D. S. & Mackay, T. (1996). *Introduction to Quantitative Genetics. Fourth edition.* Addison Wesley Longman, Harlow, Essex, UK.

Fisher, S. G. & Fisher, R. I. (2004). The epidemiology of non-hodgkin's lymphoma. *Oncogene*, **23**(38), 6524–6534.

Goeman, J. J., van de Geer, S. A., de Kort, F., & van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**(1), 93–99.

Goeman, J. J., Van De Geer, S. A., & Van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(3), 477–493.

Houwing-Duistermaat, J. J., Derkx, B. H., Rosendaal, F. R., & Van Houwelingen, H. C. (1995). Testing familial aggregation. *Biometrics*, **51**(4), 1292–1301.

Jiao, S., Hsu, L., Hutter, C. M., & Peters, U. (2011). The use of imputed values in the meta-analysis of genome-wide association studies. *Genet Epidemiol*, **35**(7), 597–605.

Le Cessie, S. & Van Houwelingen, H. C. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics*, **51**(2), 600–614.

Marchini, J. & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet*, **11**(7), 499–511.

Morin, R. D., Mendez-Lago, M., Mungall, A. J., Goya, R., Mungall, K. L., Corbett, R. D., Johnson, N. A., Severson, T. M., Chiu, R., Field, M., Jackman, S., Krzywinski, M., Scott, D. W., Trinh, D. L., Tamura-Wells, J., Li, S., Firme, M. R., Rogic, S., Griffith, M., Chan, S., Yakovenko, O., Meyer, I. M., Zhao, E. Y., Smailus, D., Moksa, M., Chittaranjan, S., Rimsza, L., Brooks-Wilson, A., Spinelli, J. J., Ben-Neriah, S., Meissner, B., Woolcock, B., Boyle, M., McDonald, H., Tam, A., Zhao, Y., Delaney, A., Zeng, T., Tse, K., Butterfield, Y., Birol, I., Holt, R., Schein, J., Horsman, D. E., Moore, R., Jones, S. J. M., Connors, J. M., Hirst, M., Gascoyne, R. D., & Marra, M. A. (2011). Frequent mutation of histone-modifying genes in non-hodgkin lymphoma. *Nature*, **476**(7360), 298–303.

National Cancer Institute (2012). *Non-Hodgkin Lymphoma*. Available from: `http://www.cancer.gov/cancertopics/types/non-hodgkin`. Accessed July 2012.

Novik, K. L., Spinelli, J. J., MacArthur, A. C., Shumansky, K., Sipahimalani, P., Leach, S., Lai, A., Connors, J. M., Gascoyne, R. D., Gallagher, R. P., & Brooks-Wilson, A. R. (2007). Genetic variation in h2afx contributes to risk of nonhodgkin lymphoma. *Cancer Epidemiology Biomarkers and Prevention*, **16**(6), 1098–1106.

Schuetz, J. M., Daley, D., Graham, J., Berry, B. R., Gallagher, R. P., Connors, J. M., Gascoyne, R. D., Spinelli, J. J., & Brooks-Wilson, A. R. (2012). Genetic variation in cell death genes and risk of non-hodgkin lymphoma. *PLoS ONE*, **7**(2), e31560.

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