University of Alberta

Gene-Environmental Interaction Assessment in Genome Wide Association Study

by

Wei Liu

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

Master of Science

in

Epidemiology

Department of Public Health

Fall 2011

Edmonton, Alberta, Canada

Examining Committee

Dr. Yutaka Yasui, School of Public health, University of Alberta

Dr. Irina Dinu, School of Public health, University of Alberta

Dr. Yan Yuan, School of Public health, University of Alberta

Dr. Smita Bhatia, Department of Population Health Science, City of Hope, California, USA

Abstract

This thesis aimed to investigate congestive heart failure (CHF) associated with anthracycline exposures among the childhood cancer survivors, trying to identify differential anthracycline effects by genotype of certain genes. We had 130 childhood cancer survivors who developed CHF as cases and 269 individually-matched controls. We assessed each single nucleotide polymorphism's (SNP) interaction with anthracycline dose on the case-control status of CHF. Besides the single SNP analysis, we developed a method that considered SNPs from the same gene as a set and assessed the set's effect on anthracycline toxicity by extending a gene-set analysis method developed previously for gene expression studies. Some SNPs had interactions with anthracycline exposure on the development of CHF. Moreover, there was stronger evidence from the SNP-set analysis indicated that the interaction between certain sets of SNPs (genes) and anthracycline exist.

Acknowledgements

I owe my deepest gratitude to my supervisor Dr. Yutaka Yasui for his excellent guidelines and generous support. I would not complete the analysis and writing of this thesis without his continuous assistance and encouragement. During the development of this thesis, I learned lots of specialty knowledge from him, which will benefit my career in the future. His attitude and enthusiasm to research impressed me very much. Dr. Yutaka Yasui did not only influence me academically, his kindness to people also gave me indication for my daily life.

My sincere thanks also go to Dr. Smita Bhatia who initiated the project of my thesis. It's a precious experience for me to be involved in such a great project. She also gave me lots of help as a member of my committee. I also wanted to show my gratitude to other committee members, Dr. Irina Dinu and Dr. Yan Yuan who provided constructive feedbacks in my thesis writing.

I am grateful for the assistance of every member of our research group. It's so lucky to study together with so many excellent and kind people over the last two and half years.

Finally, I want to express my heartfelt gratitude to my family members. I am grateful to my parents and sisters who always support me. I would like reserve my special thanks to my two cute nephews who bring so much happiness to the whole family.

Table of contents

List of Tables

List of Figures

Chapter 1 Introduction

More than 80% of childhood cancer patients become long-term survivors due to remarkable improvements in cancer treatment in the last several decades (1,2). Anthracycline is one of the most widely used chemotherapy drugs in childhood cancer therapy and is highly effective against certain types of cancer, such as leukemia and lymphoma (3). However, anthracycline exposure can cause serious health problems many years after the treatment, one of which is the elevated risk of congestive heart failure (CHF).

CHF risk following childhood cancer treatment is known to be strongly associated with cumulative doses of anthracycline in cancer treatment (4). Studies have shown that anthracycline's cardiotoxicity is generally not apparent under the cumulative dose below 300 mg/m^2 (5). For childhood cancer survivors who had received cumulative doses of anthracycline exceeding 300 mg/m^2, more than 10% developed cardiac diseases within 20 years following the cancer therapy (6). Moreover, the observation that toxicity increases greatly at cumulative doses greater than 550 mg/m^2 has made 550mg/m^2 an upper limit in practice (2,7).

However, observations showed that some childhood cancer survivors exposed to low doses of anthracycline could experience CHF, while other survivors tolerated high doses of anthracycline (e.g., excess of 1000 mg/m^2) . A potential explanation for this heterogeneity in CHF risk may be the genetic variations among cancer patients that modify anthracycline toxicities.

Although genetic factors can elevate disease risks directly, many diseases are induced by the joint effect of genetic variations and environmental exposures. Epidemiological studies have shown that many diseases and human traits, such as diabetes, mental disorders, and cancers, are the results of complex interactions between genetic factors and environment exposures (8-10). Similarly, late effects of treatment are a result of the joint effect of genetic variations and the treatment, where genetic variations alone may not cause the late effect but influence susceptibility to treatment-related risk. Small differences in genetic codes could make different responses to the same environmental/treatment exposures. Understanding this complex interaction of genes and environment/treatment can help us to setup specific approaches for preventing disease/late effects in patients with different genotypes (11).

As a distinct adverse effect of anthracycline, cardiotoxicity has been studied and known for a long time. A few studies have examined the role of biomarkers in anthracycline-induced cardiotoxicity. A review which summarized 14 studies showed that the patients with elevated blood proteins B-type natriuretic peptide, N-terminal pro-BNP and cardiac troponit T after or within anthracycline therapy were prone to develop cardiac dysfunction; it suggested that these biomarkers could be used in the early detection of elevated cardiac disease risk among childhood cancer patients receiving anthracycline treatment (12). Nevertheless, the mechanism for genetic variants in the development of anthracycline-associated cardiac disease is not well understood. The current study aimed to fill

this gap by examining genetic variants across the whole genome in relation to the risk of anthracycline-related CHF among childhood cancer survivors.

This study was a genome-wide association study (GWAS) which scanned the single nucleotide polymorphisms (SNPs) across a spectrum of hypothesized cardiovascular-disease related genes in the whole genome. The specific objective of the GWAS analysis conducted in this thesis work was to explore genetic susceptibilities defined by *a set of SNPs* to anthracycline-associated CHF. First, we conducted a single-SNP analysis, which is the typical approach in GWAS analysis, to identify single SNPs associated with anthracycline-induced CHF risk. Second, we extended an SNP-set analysis method developed for gene expression studies to GWAS (genetic) analysis, by incorporating the SNPs in the same gene as a SNP set to examine their combined effect and to identify functional genes for modifying anthracycline toxicity in the development of CHF.

This thesis consists of six chapters. Following this introduction, there are five other chapters. Chapter 2 explains some basic concepts related to molecular biology, as a basis for its subsequent discussions. In Chapter 3, two important and common statistical issues in high-dimensional data analysis for genetic study, namely, multiple testing and over-fitting are discussed. In Chapter 4, we describe the statistical methods for the main theme of this thesis: the extension of an analysis approach from genomic studies to genetic studies. In this chapter, concepts related to GWAS and the standard analysis procedures in GWAS are described. We then summarize gene expression analysis methods, including Significance Analysis of Microarray (SAM) for single-gene expression analysis

and SAM-GS for gene-set expression analysis. Then, we extended SAM-GS for gene-set expression analysis to SNP-set analysis of genetic data. This is followed by the description of the main tool for this thesis, a SNP-set gene-environment interaction analysis. Chapter 5, an application chapter, returns to the analysis of childhood cancer survivor data and presents all aspects of the anthracycline related CHF in childhood cancer survivors, including the interaction of SNPs and anthracycline dose. In Chapter 6, strengths and limitations of our study are discussed. Finally, we conclude with recommendations for possible future work based on this study.

Chapter 2 A Brief Summary of Relevant Molecular Biology

2.1 Gene and gene expression

2.1.1 DNA and gene

Deoxyribo Nucleic Acid, which is known as DNA, stores human genetic information (13-15). In the nucleus of human body cell, the double DNA chain is twisted into chromosome. There are 23 pairs of chromosomes, 22 pairs of which are numbered chromosomes, called autosomes, and the other pair consists of two sex chromosomes (14). For each pair of chromosomes, one comes from the mother and the other comes from the father. Four different types of nucleotides are the basic units of constructing the DNA molecule, Adenine, Thymine, Cytosine, and Guanine expressed by A, T, C and G, respectively (13,15). Each nucleotide is paired to a complementary nucleotide following the "pair rule", which is, A and T are paired, and C and G are paired. By this rule, two single DNA chains are combined into a double-stranded DNA molecule.

Genes are the functional fragments of the double chain DNA, determined by the sequence of nucleotides. There are approximately 30,000 genes in human DNA (16). They are the codes for recording and passing down biological information, controlling the birth, growth, disease, aging and death, and all life phenomena of organisms. Genes may also interact with environmental factors in contributing to these biological functions.

2.1.2 Single nucleotide polymorphism (SNP)

While more than 99.9% DNA sequences are identical in a population, the less than 0.1% of genetic differences (variation) lead to observed phenotypic differences within the population (17,18). The most common type of genetic variation is called *single nucleotide polymorphism* (SNP), the genotype variation of one nucleotide, where the proportion of the rarer genotype is greater than 1% in a population(17-19). Each SNP represents a difference at a DNA location. As shown in figure 3.1 (20), the replacement of the nucleotide pair (C, G) with the nucleotide pair (T, A) at a certain locus of DNA chain forms a SNP.

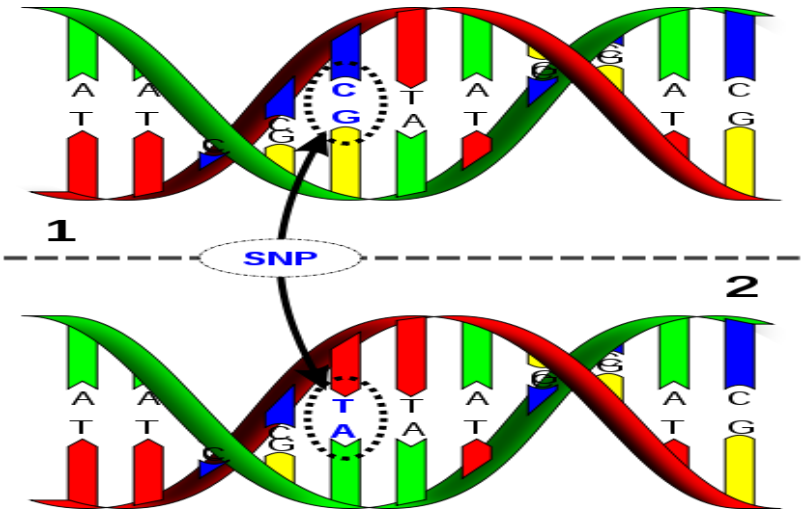

Figure 2.1: DNA molecule 1 differs from DNA molecule 2 at a single base-pair location

In human DNA, there is one SNP in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome (21,22). Since SNPs represent the basic variation in genes, some changes in SNPs could cause differences in their genes' functions. Therefore, SNPs can be regarded as

biological markers that could help studies of the association between genes and a phenotype of interest (19). For example, SNPs can be used to explain the family history of some diseases with heritable components, such as cancers (23). As a biological marker, SNPs are also helpful in predicting an individual's response to certain drugs and the risk of developing some particular diseases (24).

2.1.3 Gene expression

Gene expression involves complex processes in which the information of DNA sequence is expressed under certain conditions; and its quantification measures to what extent the gene is simulated to synthesize its products (25). In the process of gene expression, after the activation of a gene (trigged by internal stimuli and/or external environment), the DNA produces messenger RNA (mRNA), which could "translate" genetic message to biological characteristics through synthesized functional proteins (26).

The step from DNA to mRNA, which takes place in the cell nucleus, is called transcription. The step from RNA to protein is called translation, which takes place outside the nucleus (27). In the transcription process, the sequence of a strand of mRNA is based on the sequence of the complementary strand of DNA (27,28).

Since gene expression is the intermediate step at which a genotype gives rise to a protein that may determine the phenotype, gene expression is also helpful in examining the association between a phenotype (disease) and genes (26,29,30).

Research on gene expression assists scientists towards understanding the functions of genes.

2.2 Microarray technology

2.2.1 Microarray for SNPs

Microarray-based gene expression analyses have been introduced with the biotechnological advances in the last two decades (31). This new technology permits thousands of gene expression measurements in a single assay, providing a massive amount of biological information on study samples. Similarly, SNPs could be genotyped by a SNP chip, which is a type of DNA microarray designed to identify genetic variants associated with a phenotype of interest(32,33). There are thousands of probes attached on the surface of a chip, which represent the nucleotide sequences of the single-stranded DNA chain (34,35). Although there are different microarray chips for SNP genotyping using different technologies, the "pair rule" applies everywhere. During genotyping, the DNA samples are separated into two single stranded fragments and both are labelled by fluorescent substance (19,34). Then, the DNA fragments will be attached onto the microchip and hybridized with the synthetic sequences on the chip following the "pair rule". After hybridization, specialized computer equipment will be used to measure the fluorescent signal intensity contained in each probe (36). Genotypes for the alleles of a locus can be inferred from the fluorescent signals.

2.2.2 Tag SNPs

There are about 10 million SNPs in human genome (37). Although SNP genotyping could infer massive genetic information, measuring all 10 million SNPs for each individual is currently not feasible in large-scale studies due to cost and time it takes. However, SNPs which are close to each other tend to be transmitted together from generation to generation. A set of SNPs that are highly likely to be transmitted together is called haplotype (38). The linkage among SNPs is referred to as linkage disequilibrium (LD), which can be measured by correlation between SNPs (39,40). High correlations among SNPs enable one of them to be served as proxies for the other(s) (41). Figure 2.2 shows a haplotype which contains five SNPs (SNP1- SNP4) and true disease-causing SNP which is unknown (42). Because of the high correlations among the five SNPs in the haplotype block, the disease-causing SNP can be presented by the other 4 SNPs which can be called as “Tag SNPs” or “Marker SNPs”.

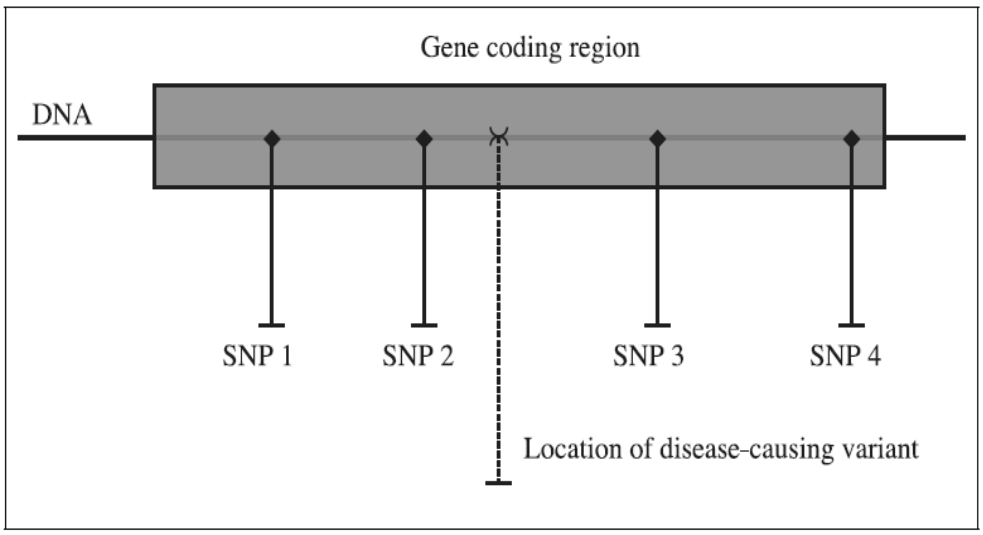

Figure 2.2: Tag SNPs in a haplotype

2.2.3 Microarray for gene expression

Gene expression of a gene refers to the amount of mRNAs synthesized (transcribed) from the gene based on its DNA sequence. Microarray technology for the SNP chip described above and that for gene expression are based on the nucleotide pair rule and, thus, similar. While the SNP microarray is to examine the nucleotide sequence in the DNA chain, the microarray for gene expression is to measure the amount of gene products (mRNA) under a certain set of circumstances.

When measuring the level of gene expression, researchers actually measure the amount of complementary DNA (cDNA) to the mRNA (43,44). cDNA is the product of reverse transcription which copies mRNA into DNA, pairing the RNA bases(A, T, G and C) to their corresponding DNA counterparts (T, A, C and G) (44). The reason that researchers measure cDNA rather than mRNA is because

RNAs are inherently less stable than DNA and techniques for routinely amplifying and purifying individual RNA molecules do not exist (45).

To measure gene expression, researchers firstly need to collect the mRNA expressed under a certain set of circumstances. Then reverse transcriptase enzymes would generate a complementary DNA (cDNA) to the mRNA, through which researchers can label and quantify the mRNA with fluorescent nucleotides attached to the cDNA (44,46). Generally, if a gene is highly active, it produces more mRNA and more corresponding cDNA than genes that are less active. Based on the pair rule, the fluorescent-labelled cDNAs which represent mRNAs of the gene will match to their synthetic complementary DNAs on the microarray slide (19,46). Researchers can use a special scanner to measure the fluorescent intensity for each gene to quantify their level of expression.

Figure 2.3: Microarray process for gene expression

Note that the information in genes (DNA) is determined by their nucleotides and generally unchanged throughout the lifetime. The study of this gene/DNA information is called genetics. On the other hand, the study of gene expression is concerned about mRNA copies which change constantly to adapt/respond to the circumstances the host and the cell are surrounded by.

Chapter 3 Statistical Issues in High-dimensional Data Analysis

3.1 Multiple comparison problem

The multiple comparison problem is a challenge in high dimensional biological data analyses such as those of gene expressions and GWAS based on the microarray technology (47,48). In statistical analysis, a p-value is the probability of observing a test statistic value as extreme as, or more extreme than, the observed test statistic, assuming that the null hypothesis is true (49,50). It is a common practice to reject or accept a hypothesis based on a p-value, and p-value of 0.05 is usually considered as a threshold of making this decision. If the p-value of a test is equal to or less than 0.05, the null hypothesis is normally rejected. This also means that the probability of rejecting the null hypothesis when it is actually true is 5%, which is referred as Type I error (or false positive) (49).When a set of tests are conducted simultaneously, even if the probability of committing Type I error for each test is set at 5%, it would lead to an overall Type I error probability that is larger than 5% (51). In a GWAS study, for example, there are 1000 SNPs that need to be investigated at the same time and 0.05 is set as the cut-off for each SNP, then 50 null SNPs that do not have associations with the outcome would be expected to be found positive by chance, which is to say, 50 null SNPs will be falsely regarded to be significantly associated with the outcome. Thus, in GWAS studies which may scan millions of SNPs, the multiple comparison needs to be adjusted properly.

3.2 Multiple comparison adjustment procedures

3.2.1 Bonferroni adjustment

Suppose there are N hypotheses that need to be tested, and for each single test, α_1 is set as the cut-off for rejecting the null hypothesis, which means α_1 is the Type I error probability for each single test. If we do more tests, the chance of making an error will increase. For N independent tests, the probability of no Type I error equals to $(1-\alpha_1)^N$ and the probability of having at least one false positive is $1-(1-\alpha_1)^N$ which approximate to $N\alpha_1$ when α_1 is very small. We should decrease this probability to maintain the overall Type I error (51,52).

The Bonferroni adjustment is a straightforward correction to the problem of multiple comparisons. It applies $\alpha' = \alpha_1 / N$ in place of α_1 as the significant level of each single test (53). Then, for all N tests, the overall Type I error would be controlled to be no greater than α_1. Suppose, for example, we conduct 1000 tests to check 1000 genes in association with the disease of interest and we want to control Type I error of this study at an overall level of α_1=0.05. We would expect 50 genes to be false positive by chance if we set α_1=0.05 as the p-value threshold. Bonferroni adjustment applies α'=0.05/1000=0.00005 as the significant level for each test, and then the chance of making a false discovery for the whole 1000 tests will be controlled as 0.05.

Although the Bonferroni adjustment is effective for controlling Type I error for the multiple testing problem, it has some limitations. Firstly, when Bonferroni

adjustment is applied, it is assumed that all the tests are independent, which is untenable in any of the high-dimensional studies (42,54). As illustrated above, one individual SNP is frequently correlated with the SNPs nearby. Secondly, Bonferroni adjustment is a conservative adjustment method where the adjusted Type I error probability is less than the target and it could lead to omissions of the truly significant associations.

3.2.2 False discovery rate

3.2.2.1 Definition of FDR

False discovery rate (FDR) is another statistical method which is widely used in addressing the multiple comparison problem. FDR is defined as the expected proportion of hypotheses which are truly nulls among those declared to be significant (42). While Bonferroni correction controls Type I error by adjusting the overall threshold of p-value for all the tests, FDR aims to estimate the error rate for tests which are declared significant (55,56).

To discuss FDR, consider a summary of *m* multiple test results shown in Table 3.1 (42). Of the *m* tests, F is the number of false positives, T is the number of true positives, and R is the total number of tests declared significant.

Table 3.1: Property of multiple hypothesis tests

		Test		
		Significant	Non-significant	
Truth	H_0	F	V	m_0
	H_A	T	S	$m-m_0$
		R	m-R	m

FDR is an expected proportion of false positives among all the tests called significant and is given by:

$$FDR = E(\frac{F}{R}) \tag{3.1}$$

If we use a p-value for deciding to reject or accept a hypothesis, when there are thousands of tests that are conducted simultaneously, it would result in a large number of false positives. FDR can be applied to control the number of false positives, because an FDR estimate for a set of hypotheses that are declared significant estimates the chance of them being false positives. A p-value of 0.05 for each test would result in 50 false positives among 1000 tests assuming all hypotheses are null. This is true no matter how many tests are truly non-null. Using FDR, however, we do not have to consider "what if all the hypotheses are null." We know if there are *R* tests that are declared significant and the

corresponding FDR estimate was *f*, *Rf* is the estimated number of false discoveries. FDR is adaptive to control the number of truly significant tests.

3.2.2.2 Estimation of FDR

When setting a significant p-value threshold at θ, FDR could be donated as

$$\hat{FDR} = E(\frac{F}{R}) \approx \frac{E[F(\theta)]}{E[R(\theta)]} \tag{3.2}$$

where $E[R(\theta)]$ could be estimated by the observed number of significant tests which is $R(\theta)$ and an estimate for $E[F(\theta)]$ is $m_0\theta$. Then we need to estimate m_0 which could be expressed as $m \cdot \pi_0$, where m is the number of total tests and π_0 is the proportion of truly null tests. Under the null, p-values are uniformly distributed in [0, 1], whereas under the non-null, p-values tend to be small. Thus, we could estimate π_0 by checking the proportion of "the flat part" in the density of the p-value distribution.

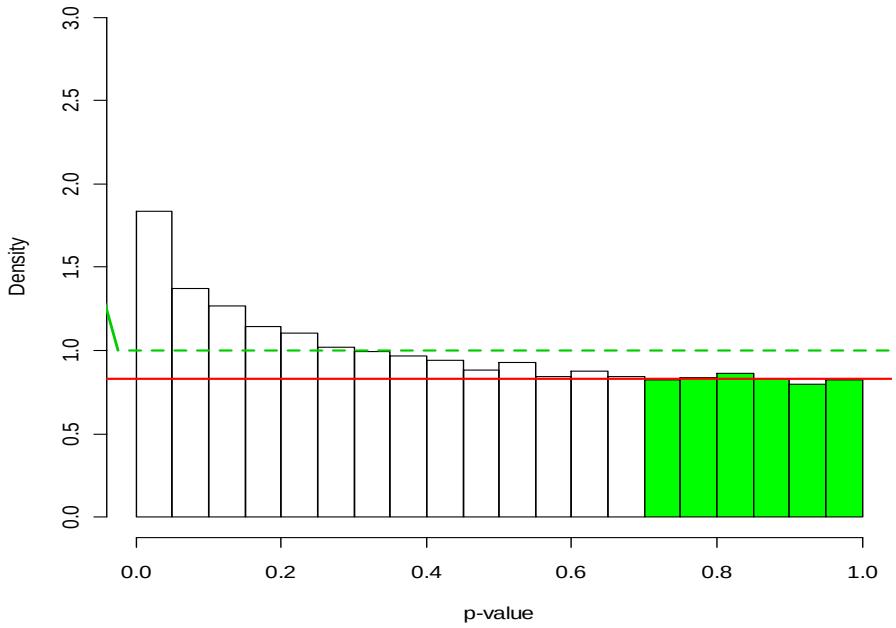

Figure 3.1: Distribution of p-values

As shown in Figure 3.1, if all the tests are null, we expect the p-value distribute flatly as the green dashed line. The region below the red line can be considered as the proportion of truly null tests (π_0); and the part with p-value larger than λ=0.7, which is indicated as green bars, is the "flat part" that can be used to estimate the red line and the number of truly null tests m_0 (57).

Let λ=0.7, the number of tests with p-value$\geq \lambda$ is given by:

$$\#\{p > \lambda\} \approx \hat{m}_0 \cdot P\{p > \lambda\} = \hat{m}_0(1-\lambda) \tag{3.3}$$

Then we can estimate π_0 as:

$$\hat{\pi}_0 = \frac{\#\{p > \lambda\}}{m(1-\lambda)} \tag{3.4}$$

and:

$$\hat{FDR}(\theta) = \frac{\hat{\pi}_0 \, m \cdot \theta}{R(\theta)} \tag{3.5}$$

While FDR is a useful measurement of the overall error rate for a set of tests declared significant, q-value is a measure given to each single test. The q-value of an individual hypothesis test is an estimated measure of the probability of false discovery when this test is declared significant (58). If we declare a test is significant with a p-value of t, the FDR under the p-value level of t can be estimated: this is the q-value for this test. Similar to the p-value cut-off, we can set a threshold of α for q-values: all the tests with q-values smaller than α are expected to yield the false positive proportion less than α.

3.2 Over-fitting

3.2.1 Over-fitting problem

Statistical modeling is an effective approach to describe the association between an outcome of interest and independent variables. Based on a statistical model, we can also predict new observations. Given a sample, statistical modeling estimates the underlying data-generation mechanism to be generalized to new data from the same population. A model may provide an adequate explanation for the sample, which is often referred to as *training data*, but may fail to demonstrate a valid

prediction for unseen data, which is often referred to as *testing data*. A statistical model could fail in the generalization to new data in a number of ways, including incorrect model assumptions, insufficiency/bias of training data, training and testing samples being from different populations, and a phenomenon called *over-fitting*. Over-fitting means the statistical model tend to describe random error or noise instead of the underlying relationship. Over-fitting can lead to an overestimate of the performance of a model, misleading the users of the model with respect to its performance capacity.

When an over-fitting happens, the statistical model describes random error, or "noise", in the training data as a systematic pattern of data generation. This occurs more frequently when the sample size is small. When the study sample is small, it is difficult or impossible to estimate the true underlying distribution of the population, separating it from random error. The model from the training data would then be prone to describe the noise as part of the systematic pattern of the data. Second, if a model is allowed to be excessively complex, such as having too many parameters or predictors relative to the number of observations, over-fitting is more likely to occur. More predictors can improve the performance of the model in the training data. However, a model is made more complex in order to maximize its performance on the specific training data at hand without considering its generalizability, the model will be ineffective with respect to predicting new data in a testing dataset.

In human microarray studies, which involve high dimensional data analysis, over-fitting is a potential problem. Given that thousands of genes or SNPs are considered in a microarray study, it is prone to consider a highly complex model trying to fit the given data perfectly. Moreover, due to the large cost of genotyping for genes and SNPs, the sample size may be relatively small. When the study samples come from a limited subset of the population to be inferred, the statistical model would not be well generalized. This may be a reason for a commonly encountered phenomenon where microarray results are not replicable in other follow-up studies.

3.2.2 Cross-validation and the use of validation sets

There are several methods for preventing over-fitting, among which cross-validation is widely used. Cross-validation is a method to evaluate how a statistical model can generalize to a new set of observations (59). A *k*-fold cross-validation divides the sample data into *k* parts. In one of *k* steps of the cross-validation, the *k* parts are divided into two groups: one group, consisting of (*k*-1) of the *k* parts, is training data which are used to build a model; and the other group, consisting of just one of the *k* parts, is testing data which are used to measure the performance of the model from the first group. This single step is repeated *k* times until all *k* parts were considered as the testing data. By summarizing model performance across *k* testing datasets, we can evaluate the overall model performance and assess the over-fitting problem.

For example, in Figure 3.2 (59), a 10-fold cross-validation is illustrated. The sample data are equally divided into 10 folds and the step is repeated 10 times.

For each step, 90% of the data (yellow colored) are used to form the training data, and the other 10% of the data (blue coloured) are used to estimate the prediction error.

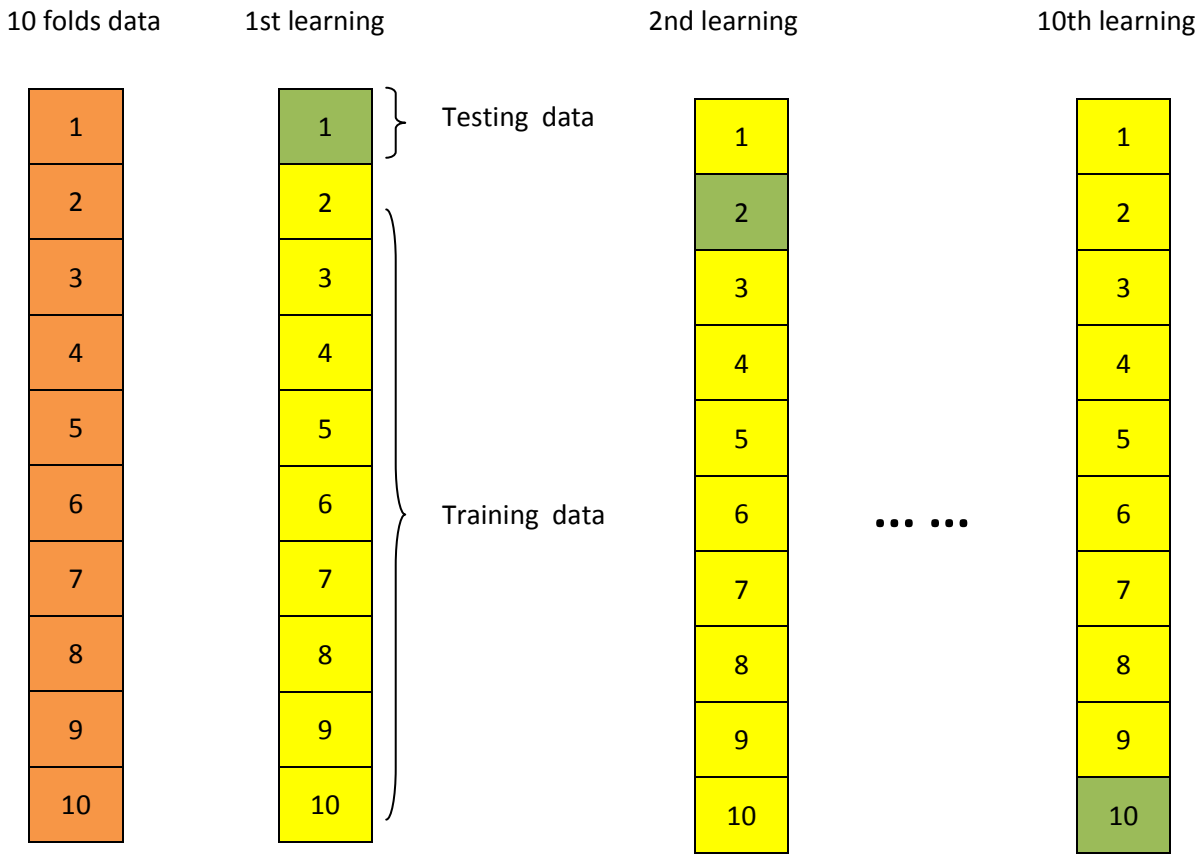

Figure 3.2: Procedure of the 10-fold cross-validation

After *k* (10) steps, the error of the model would be estimated by the summary of the 10 predicted errors.

Chapter 4 Extending gene-set analysis from genomics to genetics

4.1 Genome wide association studies (GWAS)

Genome wide association studies (GWAS) involve the scanning of the entire human genome to identify genetic factors which are associated with the traits of interest (60,61). If certain genetic variants are more frequent in people with a disease, the variants are considered to be associated with the disease. Thus, a common GWAS design compares the frequency differences of genetic variants between case and control groups (62,63).

Given the recent advances in microarray technologies, more than one million SNPs can be genotyped in a high-throughput assay for a large number of subjects in epidemiological studies. Moreover, the Human Genome Project which has mapped nearly all human DNA sequences provided the basis for scanning genetic variants across the whole genome and identifying potential susceptibilities which would contribute to the disease of interest (64,65). Based on these remarkable technological advances, GWAS has gained great popularity and has resulted in important findings in recent years. For example, the Wellcome Trust Case-Control Consortium (WTCCC) conducted a series of large GWASs and successfully identified many novel genetic loci associated with several common human diseases, such as coronary artery disease, Crohn's disease and diabetes (62,63).

Unlike candidate polymorphism or candidate gene studies which examine specific genetic markers hypothesized to be associated with the disease of interest based on prior knowledge, GWAS is able to check up to one million of tag SNPs in a single study (64). Candidate SNP/gene studies and replication studies are often conducted to validate the findings from GWAS.

4.2 Quality control methods for GWAS

In GWAS, we need to perform data quality assessment to avoid bias and to increase the power of the study. There are different methods to control the data quality in experimental and analyses procedure. They are applied to both SNP genotype data quality and sample quality assessments. We discuss here four widely applied measurements for quality control.

4.2.1 Genotyping call rate

The call rate of a SNP is the proportion of non-missing genotype in the sample, while the call rate of a subject is the percentage of SNPs which are successfully genotyped in that subject. Call rates are the quality control measurements to measure the missing frequency and indicate the accuracy of SNP genotyping (60,66). It is applied to the selection of both SNPs and subjects samples for subsequent analyses. High call rates do not only represent the low missing values, but high accuracy of genotyping, thereby leading to higher precision and power of subsequent statistical analysis.

The accuracy of genotyping is determined by the technology of microarray as well as the quality of the sample materials (e.g., blood samples). The higher the

call rate is, the better the data quality is. GWAS researchers often set 90% as the minimum acceptable cut-off for call rates for quality control. If the call rate for a SNP or a subject is less than 90%, it would be excluded from the sample. Many GWAS studies used a call rate threshold of >95% for quality control (67,68).

4.2.2 Minor allele frequency

At every DNA locus, a SNP is usually consisted of two alleles. Minor allele frequency (MAF) for a SNP refers to the frequency of the less frequent allele in a given population. Polymorphism is usually defined by MAF greater than 1% (sometimes higher): the range of MAF for a SNP is, therefore, from 1% to 50% (42). As it is difficult to detect the association of a rare variant with the disease of interest, the SNPs with too small MAFs are usually deleted from GWAS: it requires high statistical power to detect rare variants' association with the disease. MAF of 1%, 5% or 10% is usually applied to filter the SNPs in many GWAS studies (60,69).

4.2.3 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium (HWE) refers to the independence of two alleles at a single DNA locus (70). HWE implies that the probability of an allele of a SNP does not depend on its complementary allele. For example, for a SNP which includes genotypes of AA, AB and BB, denoting the allele frequencies of A and B by $p(A)$ and $p(B)$, respectively, Table 4.1 shows their relationship:

Table 4.1: Allele frequencies and genotypes

		A	B
		$p(A)$	$p(B)$
A	$p(A)$	AA	AB
B	$p(B)$	AB	BB

According to the HWE, the probabilities of A and B alleles are independent in a population; the probabilities for genotypes of AA, AB and BB should be equal to $p(AA) = p(A)^2$, $p(AB) = 2p(A)p(B)$, and $p(BB) = p(B)^2$, respectively. The HWE law hold when a series of assumptions such as no genotyping error, no mutation, and random mating to hold. The failure of the HWE law to hold in a SNP indicates the violation of any of these assumptions (e.g., genotyping errors) (71). SNPs violating the HWE law with high statistical significance are excluded from the subsequent statistical analyses.

HWE check should be conducted for the control group in case that the violation of HWE is caused by different allele frequencies between case and control groups. Testing of HWE includes a chi-square test and Fisher's exact test, both of which can assess statistical evidence of the SNPs' deviation from HWE (72,73). Because the tests of HWE involve multiple comparisons, one could use Bonferroni correction to adjust the p-values and select a threshold which is determined by the number of tests/SNPs (60). The SNPs with a p-value smaller than the chosen cut-off are excluded from the study. FDR can also be used for this multiple testing problem.

4.2.4 Correction for population stratification

Differences in genotype frequencies between case and control samples can be not due to the disease, but caused by population substructures, for example, by different ancestry or non-random mating (74). This problem is referred to as *population stratification*. Population stratification would affect the findings in GWAS study, leading to false discoveries (and false negatives) of genetic markers (75,76). Therefore, it is important to assess population-structure heterogeneity and correct for the population stratification.

Of the methods for population stratification correction, Genomic control and Eigenstrat are widely used. Genomic control applies an inflation factor to control the inflation of an association-test statistics due to population stratification (75,77). A Q-Q plot can be used to show the degree of inflation of a test statistics. If there is significant difference in the population structure between the case and control samples, the test-statistics distribution from association tests would be more dispersed than the theoretical null distribution. With an estimated inflation factor λ, the actual association test statistic is adjusted by divided it with the estimated λ for a population stratification correction (78).

Another popular method used for correcting population stratification is Eigenstrat which uses principal component analysis to detect and correct population stratification. Eigenstrat applies principal component analysis to genotype data to investigate continuous axes of major genetic variation (79). The principal component axes can capture the major ancestry variability of the data. For

example, an axis of variation can be interpreted as ancestry difference in geography. The sample subjects could be divided into subpopulations using this method.

4.3 Single SNP association analysis

The standard approach to GWAS analysis has been single-SNP association analysis which refers to analyzing SNPs in the whole genome one by one, for identifying SNPs which are associated to the trait of interest. Each SNP would be treated individually without considering the other SNPs. Case-control study design is the typical study design for GWAS. Single SNP analysis investigates an individual SNP by comparing the genotype frequency difference at this locus between case and control groups. Chi-square test is usually used for single SNP analysis.

Given that hundreds of thousands of SNPs in the whole genome are investigated, multiple testing problems should be taken into consideration in single SNP analysis. The SNPs can be ranked based on their p-values first. Then Bonferroni correction or FDR can be applied to examine the SNPs with strong evidence against the null hypothesis of no association.

Multivariable regression could be applied to adjust for potential effects of non genetic factors and the interaction of individual SNPs and environmental factors could be examnied. If available, replication studies and candidate gene testing can be performed to validate the results from single-SNP analysis.

4.4 Important differences between gene and gene expression

Genes are the functional portion of the double DNA chain. The sequences of the nucleotides (A, T, C and G) in a gene are genetic codes which determine the gene's genetic information. These genetic codes are stable because the nucleotide sequence of a gene for a person is largely fixed in the whole life except mutations.

Gene expression is the process by which information from a gene is used in the synthesis of a gene product, RNAs. The product of gene expression is Messenger RNA (mRNA) and transfer RNA (tRNA) which carry the gene information to synthesize proteins. Gene expression is the middle process in which genotypes give rise to the organism's phenotypes.

While a gene specifies the nucleotide sequence in DNA chain, the gene expression is the process that determines the amount of the corresponding protein to be made. In other words, the DNA sequence spells out the code for producing a specific protein, whereas the expression level is the number of copies and amount that will be produced. Although the genetic code of a gene for a person is largely fixed, the gene expression level of this gene would change from time to time. The objective of genetic studies, such as GWAS, is to identify the genetic factors (SNPs and genes) associated with the trait of interest. On the other hand, gene expression studies, which are also referred to as genomic studies, aim to investigate differences of gene expression levels associated with the trait of interest.

4.5 Gene expression analysis and SAM

Gene expression studies analyze gene expression levels of a gene under certain conditions. In studies of gene expression, the basic measure used in analysis is the amount of gene products, which is typically a real-valued positive number. By comparing the gene expression levels between targeted groups, such as patients with cancer and normal subjects, one can identify the association between gene expression levels and the disease.

Single-gene analysis is the standard for gene expression studies using the measures of gene expression levels for one individual gene at a time. Many microarray studies target at the discovery of individual genes whose expressions are associated with a phenotype of interest. A popular statistical method for single-gene analysis is Significance Analysis of Microarray (SAM) proposed by Tusher, Tibshirani and Chu in 2001 (80). SAM identifies statistically significant genes by carrying out gene-specific t-like tests and computes a statistic $d(i)$ for each gene i, which measures the strength of the relationship between gene expression and a phenotype (e.g., cases vs. controls): while SAM is used for a wide variety of phenotypes, we focus on the binary phenotype here. The statistic $d(i)$ measuring the relative difference in gene expression is:

$$d(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0} \tag{4.1}$$

where $\bar{x}_1(i)$ is the average level of expression for gene i in the case group, while $\bar{x}_2(i)$ is the average expression level for gene i in the control group. The salient feature of SAM is the use of $s(i)$, a pooled standard deviation over the two groups of genotype:

$$s(i) = \sqrt{a\{\sum[x_1(i) - \bar{x}_1(i)]^2 + \sum[x_2(i) - \bar{x}_2(i)]^2\}} \qquad (4.2)$$

where $a = (1/n_1 + 1/n_2)/(n_1 + n_2 - 2)$, n_1 and n_2 are the numbers of cases and controls, respectively; s_0 is a small positive constant that adjust for the "small variability problem" in microarray measurements. The reason to add s_0 to the denominator is to make the variance of $d(i)$ independent of the mean level of gene expression: at lower expression levels, values of $d(i)$ could become very high due to very small values of $s(i)$ (the small variability problem); adding a small positive constant s_0 to the denominator ensures that the variance of $d(i)$ is independent of the mean level of gene expression.

Permutation of case-control labels is used to calculate the p-value for each gene i. Samples in the case group are exchanged randomly with the samples in the control group to get the permuted test statistic $d'(i)$. Through comparing the original $d(i)$ with the permuted set of test statistic $d'(i)$'s, we can get the p-value for gene i, permutation test.

4.6 Gene-set expression analysis and SAM-GS

Besides single-gene analysis, gene-set analysis has been conducted for gene expression studies. A gene set is a group of genes which are *a priori* defined according to their biological properties (81). Genes in the same gene set may contribute to certain biological functions, and their sets are also referred to biological pathways. Gene-set analysis evaluates the expression levels of genes in the set, or biological pathway, as a group together, rather than one single gene at a time, in association with a phenotype (81,82).

Extending SAM for single-gene analysis, Dinu *et al.* proposed a gene-set analysis in 2007, called SAM-G(83,84). SAM-GS tests a null hypothesis that the mean vector of gene expressions in a gene set does not differ by the phenotype of interest (83). While SAM evaluates the single gene expression difference between cases and controls, SAM-GS aims to examine the gene set expression difference between cases and controls.

SAM-GS steps are as follows:

1) For each gene i in the gene set under consideration, the same *t*-like statistic $d(i)$ as in SAM is calculated:

$$d(i) = \frac{\overline{x}_1(i) - \overline{x}_2(i)}{s(i) + s_0} \tag{4.3}$$

where $\bar{x}_1(i)$ is the average level of expression for gene i in the case group while $\bar{x}_2(i)$ is the average expression level for gene i in the control group, $s(i)$ is the pooled standard deviation of gene expression over the two groups of phenotype, s_0 is the small positive constant that adjusts for the small variability encountered in microarray data (83-85).

2) Calculate the following SAM-GS test statistic for the gene set:

$$SAMGS = \sum_{i \in s} d^2(i) \qquad (4.4)$$

where *S* is number of gene in a gene-set.

3) Permute the phenotype labels and compute the permutation statistic $d'(i)$ for each gene i in the gene set, then get the test statistic value of $SAMGS'$ for the permuted data.

4) Repeat the permutation for a large number of times to get a permutation distribution for $SAMGS'$, and then compare the observed value of $SAMGS$ with its permutation distribution to get the statistical significance level (p-value) for the gene set.

SAM-GS is an extension of SAM for single gene analysis to gene-set analysis. Compared to single gene analysis, gene-set analysis has multiple advantages. First, gene-set analysis can detect some important effects which may be missed by single gene analysis. Because some single genes may have subtle effects which

would be missed by single gene analysis; the combination of their effects, however, could make magnificent difference on phenotype and can be detect by gene-set analysis. Second, gene-set analysis can reduce the number of hypothesis tested, which will help reduce problems rising from multiple hypothesis testing (81,82). Moreover, based on the *a priori* biological expert knowledge, gene-set analysis can promote biological interpretation of gene expression data.

Another method proposed by Wang and Dinu considers the gene expression level of a gene-set as a linear combination of single genes' expression (86). Then, we can test the association between gene-set and a phenotype by comparing the combination means of case and control groups.

This method supposes the expression of K genes in a gene set is expressed by $(X_1, X_2 .....X_K)$ and $Z(\beta) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_K X_K$ is a linear combination of the genes. Then, the gene expressions of gene set can be expressed by a univariate model:

$$Z_{ij}(\beta) = \mu_i + e_{ij} \qquad (4.5)$$

where μ_1 and μ_2 are the mean gene expressions for case and control groups respectively, and $e_{ij} \sim N(0, \sigma^2)$.

Through this transformation, this test becomes a classical two-sample comparison problem. What we need to do is to find the premium vector β^* to maximize the two sample t-test statistic

$$T^2(\beta^*) = \frac{(\bar{Z}_1 - \bar{Z}_2)^2}{S^2_{Z_1 Z_2}(\frac{1}{n_1} + \frac{1}{n_2})} \quad (4.6)$$

Where $\bar{Z}_1 = \beta^T \bar{X}_1$, $\bar{Z}_2 = \beta^T \bar{X}_2$, $S^2_{Z_1 Z_2} = \beta^T \hat{\Omega} \beta$. $\bar{X}_1$ and $\bar{X}_2$ are the K-sample average of gene expressions of case and control groups. $\hat{\Omega}$ is a pooled covariance matrix over the two groups and would be estimated with a shrinkage covariance matrix proposed by Schafer and Strimmer to deal with the singular problem (87).

4.7 SNP-set analysis: An extension of SAM-GS to genetic studies

GWAS aims to identify SNPs associated with an outcome. There are more than 10 million SNPs in the whole genome of a person, some of which may work jointly in influencing the outcome. For many outcomes, genotype difference at a single locus may not dominant enough to make the difference of the phenotype. As genes in the same biological pathway presumably have related biological functions, the SNPs in the same gene (or pathway) may be associated with the outcome (through the related biological functions).

By viewing a SNP set as a gene set, we attempted extending the method of SAM-GS applied in gene expression studies to genetics studies. Specifically, for a given gene i, suppose that there are N genotyped SNPs. When we perform single-SNP analysis for the association of each SNP with a disease of interest, we could get a chi-square statistic donated by x_j for each of these N SNPs (j =1, 2, 3 ... ... N) for the association with the disease phenotype. Then a statistic for the SNP set i

(e.g, gene) can be constructed by combining the statistic x_j in this SNP set (gene). Let Z_i be the summary test statistic for SNP set (gene) i, then, following the idea of SAM-GS, Z_i could be expressed as:

$$Z_i = \sum_{j=1}^{N} x_j \qquad (4.7)$$

In order to test how extreme the observed test statistic Z_i is, we apply permutation to form a null distribution of the test statistic for SNP set (gene) i. Permutation test is a type of randomization test, in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values, or a random sample from them, of the test statistic under rearrangements of the observed case and control samples (88,89). For example, if we want to perform k permutations, in every one of k permutations, the observed cases and controls are changed randomly to form a new set of cases and that of controls. From the new permuted case-control data, we perform single SNP analysis for each individual SNP j in SNP-set (gene) i to get a new test statistic x_{jk}. A new permuted test statistic for the k-th permutation can be constructed by:

$$Z_{ik} = \sum_{j=1}^{N} x_{jk} \qquad (4.8)$$

From repeated permutations, a distribution of the test statistic under the null hypothesis is formed. If we set k=10,000, then we could get a null distribution containing 10,000 test statistic values for SNP set (gene) i, which can be donated

by ($Z_{i1}, Z_{i2}, Z_{i3}, ...Z_{i10,000}$). Though comparing the observed test statistic value Z_i with the permuted null distribution of Z_{ik} , we could get a p-value for the SNP set (gene) i .

Chapter 5 An Application of GWAS SNP-Set Analysis

In this chapter, GWAS is applied to identify the genetic susceptibility in the development of therapy-related congestive heart failure (CHF) after childhood cancer. We start with describing the project background, study design, sample and data collection. Then different analyses are described exploring the association between genetic variants and CHF. Besides single SNP analysis which is the standard approach in GWAS studies, we also applied the SNP-set analysis. These include examinations of a SNP-treatment and SNP-set-treatment interaction in developing CHF after cancer. Following the description of the analysis methods, results of the analyses are presented and discussed in this chapter.

5.1 Background

Over the past several decades, survival rates for childhood cancers have improved significantly, reflected in the 5-year relative survival rates for major childhood cancers (90,91). More than 80% of children who are treated for childhood cancer can be cured and survive as least 5 years after diagnosis (92,93). The marked improvement in childhood cancer survival rates has been attributed to several factors: development of multimodal treatment protocols; and the centralization of care and support services (94).

Although childhood cancer patients can survive for a long time, their quality of life after cancer treatment is an important issue for the patients, their family and healthcare providers. Some diseases of cancer survivors are caused by late effects

of cancer treatment that includes radiation and chemotherapy drugs. The treatment can kill not only cancerous cells but also normal cells, subsequently leading to diseases. Anthracycline has been widely used in the treatment for childhood cancers, such as leukemia and lymphoma. But the toxicity of anthracycline is a potential cause for many late effects, among which cardio toxicity is one of the most frequent and serious adverse effects for childhood cancer survivors. Specifically, anthracycline induced CHF occurs in a substantial subgroup of childhood cancer survivors; up to 10% of childhood cancer survivors treated with 300mg/m^2 have been reported to develop anthracycline-associated CHF (95,96).

The risk of CHF is highly related to the dosage of anthracycline during cancer treatment (97). With the cumulative dose of anthracycline, the risk of CHF increases. Although there is a dose-response relationship between anthracycline and CHF, some patients can tolerate high doses of anthracycline without experiencing any adverse event. On the contrary, cardiac disease has been recorded in some patients exposed to very low doses of anthracycline. The reason underlies this difference could be explained by genetic susceptibility, which has not been fully explored. This project aims to examine genetic risk factors associated with the development of anthracycline-caused CHF by examining genetic loci which are relevant to cardiovascular metabolic and inflammatory syndromes.

5.2 Objective and hypothesis

The overall objective of this study was to identify therapeutic exposures and genetic risk factors associated with the development of CHF by studying doses of anthracycline, demographic factors and genetic variants among cases and controls. The specific aims of this study included:

1) To characterize survivors with the adverse event (cases), in comparison with individuals from the same pool of cancer survivors without the outcome (controls), to examine the role of genetic susceptibility in the development of CHF.

2) To assess the modification effect of genetic factors on anthracycline toxicity in inducing CHF.

The main research hypothesis to be tested in this study is:

Among the childhood cancer survivors, the toxic effect of anthracyclines in the development of CHF is modified by a subset of genetic variants.

5.3 Study population and study design

This study was initiated by Dr. Smita Bhatia from City of Hope Comprehensive Cancer Center in California, USA. The study population of this project is the childhood cancer survivors who developed cancer before age 21 in USA. Because the incidence rate of childhood cancer is low, in order to conduct a study with a sufficient size, it is necessary to establish a resource pool which can share the information and allow for a larger, multi-organizational study. The answer to this

question was the Children's Oncology Group (COG) which is a worldwide clinical trial cooperative group supported by the National Cancer Institute, with the mission of studying issues related to childhood cancers (98). COG follows up a cohort of all childhood cancer patients since the time of developing cancer.

A matched case-control design was used for this study based on the established childhood cancer survivor cohort. Cases were selected from study participants developing CHF and the childhood survivors who had not developed CHF served as controls.

5.4 Data collection

5.4.1 Case definition

The adverse event of interest in this study is CHF. If a clinical document (i.e., medical records) indicated that the patient had a diagnosis of CHF with echocardiagraphic confirmation, then the patient was labelled as a case with CHF.

5.4.2 Selection of cases and controls

Childhood cancer patients with CHF were recruited based on eligibility criteria. After obtaining informed consents from those patients (or legal guardians), they were registered via the remote data entry (RDE) system into the COG Cancer Registry, with completed on-line Eligibility Worksheet and Event Reporting Worksheet, as well as the clinical document and echocardiogram used to verify CHF.

For each case enrolled, 1 to 4 matched controls were randomly selected by the COG statistical center from the cohort. Upon each new case registration in COG Operations Office, controls were identified using the following criteria to match with the case: 1) same primary childhood cancer diagnosis; 2) same date of diagnosis; 3) at least the same length of follow-up time since diagnosis of primary cancer; 4) and same race/ethnicity. The subject sample was comprised of 399 childhood cancer survivors from COG, of whom 130 were CHF cases and 269 were controls.

5.4.3 Study procedures

5.4.3.1 Data availability

For each registered subject, case or control, data abstraction forms such as the therapy summary form and radiation treatment form were prepared by the institution that treated the patient, and submitted via electronic RDE within 3 months of the patient's study enrolment. Details regarding chemotherapy included date of treatment initiation, total number of cycles, protocols/regimens, and cumulative doses of key therapeutic agents per square meter. Details on radiation therapy included dates, total lifetime dose, field, fractions, dose per fraction, and copy of institutional radiation therapy summary report. Anthracycline cardio toxicity score was calculated by multiplying cumulative dose by a factor that reflects the cardio toxic potential of each drug (doxorubicin=1, daunorubicin=0.75, idarubicin and mitoxantrine=3).

A self-administered questionnaire was administered for each registered case or control. This questionnaire comprised of sections concerning personal medication-use history and health-related behaviours such as use of cancer screening tests. Some demographic variables were also collected through the questionnaire, such as sex, age and smoking status. For the participants who were too young to give the information themselves, their parents were requested to offer the information and help them to complete the questionnaire.

5.4.3.2 Genotyping

Blood sample was collected from all the participants to perform genotyping with rigorous quality control (QC) requirements. Genotyping was performed on the Illumina IBCv2 BeadChip array which contained approximately 50K gene-centric SNPs that were specially designed to assess relevant loci across a spectrum of cardiovascular, metabolic, and inflammatory syndromes. Although the IBCv2 array contains only 50K SNPs, based on haplotype information, it can cover genetic diversity of a broader set of genetic loci using data from the HapMap and SeatleSNPs projects by the tag SNP approach (99). In order to avoid bias, personnel performing genotyping were blinded regarding the case-control status of the participants. A total of 48,742 SNPs were successfully genotyped for each participant.

5.5 Quality control

In order to increase the power and accuracy of the analyses for this study, several quality control methods were applied to the sample subjects and genotype data.

We conducted quality control using R package "GenABEL" which is a specialized R package for GWAS.

First, we applied call rate to remove the sample subjects and SNPs with low quality which were caused by inaccurate genotyping. If the proportion of SNPs successfully genotyped for a subject was less than 95%, the subject was removed from the sample. For a SNP, if the successful genotyping proportion among all sample subjects was below 95%, then it was deleted.

Second, in order to increase the power of the analysis and reduce the number of multiple testing, we applied 10% as the MAF threshold. The SNPs with MAF less than 10% were removed from the SNP set to be analyzed.

Thirdly, we excluded the SNPs which deviated from the Hardy-Weinberg Equilibrium (HWE). The multiple comparison problem was also taken into account. In the quality control with respect to HWE, we applied the method of false discovery rate (FDR) with an FDR of <0.2 as a threshold. If a SNP violated HWE with q-value smaller than 0.2, then it was deleted.

Because race was one of the matched variables, we didn't check population structure in the analysis.

5.6 Association analysis

After quality control, there were 25,572 SNPs. A total of 399 sample subjects were eligible for further analyses, of whom 130 subjects were cases and 269 subjects were matched controls. For coding of 3 genotypes (AA, AB and BB)

with A and B representing the two alleles of an SNP at a single locus, we considered two different ways: 1) AA=0, AB or BB=1, which took AA as the reference group; 2) BB=0, AB or AA=1, which took BB as the reference group. These two codes corresponded to the notion of dominant and recessive effects of alleles.

5.6.1 Analysis of single SNP

5.6.1.1 Main effect of SNPs

Main effect models (main effects of single SNPs) without any other covariate were considered to examine the associations between CHF and each single SNP. This analysis assessed whether any single SNP was associated with the risk of CHF. Conditional logistic regression models were applied based on the matched case-control study design. We conducted two conditional logistic regressions based on the two codings for genotype: if genotype AA was the reference group (AA=0, AB or BB=1), we were assessing the effect of allele B for the SNP; and if genotype BB was set as reference (BB=0, AB or AA=1) then we were examining the effect of allele A. The main effect model was of the form:

$$\log\left(\frac{E[Y]}{1-E[Y]}\right) = \alpha_k + \beta_1 SNP \qquad (5.1)$$

where α_k was the baseline risk within each matched sample and β_1 was the effect of SNP genotype.

5.6.1.2 Analysis of single-SNP-treatment interaction

After checking the direct effect of individual SNPs in developing CHF, interaction between single SNPs and therapies applied to cancer treatment was studied. To answer the question whether the effect of anthracycline on the development of CHF was modified by genotypes, we assessed the interaction of single SNPs and anthracycline. Since heart radiation was also known to be associated with CHF risk, heart radiation status was added to this model as a potential confounder to be adjusted for. Anthracycline dose was taken as a continuous variable and heart radiation was an indicator variable. The model for the interaction assessment took the following form:

$$\log\left(\frac{E[Y]}{1-E[Y]}\right) = \alpha_k + \beta_1 SNP + \beta_2 Athracycline + \beta_3 Radiation + \beta_4 Athracycline * SNP \qquad (5.2)$$

where α_k was the baseline risk of CHF within each set of matched childhood cancer samples, β_1 was the effect of SNP genotype, β_2 was the effect of anthracycline dose, β_3 was the effect of heart radiation and β_4 was the coefficient of interaction between anthracycline and SNPs.

Because this study was a matched case-control study, the p-value and power depend on the number of discordant pairs. If the number of discordant pairs for a SNP was small, the power was low. Thus, we selected the SNPs which contained more than 35 discordant pairs for statistical analysis and ran 35,769 (two coding ways together) regression models in total.

5.7 Simulation

In order to interpret and validate the analysis results above, we conducted simulations based on the real data which contained 35,769 SNPs (two SNP coding ways together) with 35 or more discordant pairs (the numbers of discordant pairs for simulated SNPs were kept exactly the same as the real observed SNPs). To simplify the simulation, we simulated 130 cases, and for each case, we simulated one matched control (the simulation data was 1:1 matched). We assumed an odds ratio (OR) equal to 1 for the null SNPs; while for the SNPs associated with CHF, we simulated with an OR equal to 1.5.

We conducted the simulation assuming varying proportions (1%, 3%, 5%, 10% and 15% respectively) of SNPs having an association with the disease with OR=1.5 (the rest of the SNPs have no association with OR=1). The aim of assuming different proportions of associated SNPs in simulations was so that we can identify the approximate proportion which was most similar to our true data.

5.8 SNP-set analysis

We performed the SNP-set interaction analysis, which took all the SNPs in the same gene as a SNP-set, to assess the gene-treatment interaction in association with anthracycline-induced CHF. IBCv2 array contains some SNPs that do not belong to any genes; some of these SNPs are marked as “near to” some genes and the rest has no gene information. If a SNP is marked as near to a gene, then we mapped it to that gene. We successfully mapped 24,224 SNPs, which accounted

for about 95% of the entire set of the 25,572 SNPs with good quality, into 2,991 genes.

We performed 10,000 permutations to get a p-value for gene-anthracycline interaction for each gene. Specifically, we considered the interaction between each gene and anthracycline dose. A statistic for the interaction between a gene and anthracycline was constructed by summing the chi-square statistic of testing the interaction between anthracycline dose and each individual SNP over all SNPs in the gene. The chi-square statistic came from a conditional logistic model which contained SNP, anthracycline dose, heart radiation and the interaction between SNP and anthracycline. Because we coded each SNP in two ways, we had two SNP-anthracycline interaction chi-square statistic values for each SNP. We chose the larger of the two as the chi-square statistic for the SNP-anthracycline interaction.

A permutation test was then applied to get a p-value for each gene's interaction with anthracycline dose. The procedure of the permutation test was as follows. We randomly shuffled the labels of each case and its matched controls within the matched case-control set. Then the same procedure as the original case-control data was followed to calculate a permuted test statistic for each permutation. That is, for each permuted dataset, we calculated each SNP's chi-square statistic for SNP-anthracycline interaction and summed these chi-square statistic values over each gene to obtain the test statistic for the interaction of gene and anthracycline dose. After 10,000 permutations, there were of 10,000 test statistic values for the interaction of each gene and anthracycline dose which formed a null distribution

of the test statistic. By comparing the test statistic from the original data with the null statistic distribution from the permutations, we calculated a p-value for the interaction of each gene and anthracycline dose.

5.9 Result

5.9.1 Single-SNP analysis results

5.9.1.1 Single-SNP main effect

The SNP main effect was examined by a conditional logistic regression model including only one single SNP. This was repeated for 51,144 times for the two ways of coding for 25,572 SNPs that passed the quality control (QC). The p-value distribution of all the 51,144 tests is shown in Figure 5.1.

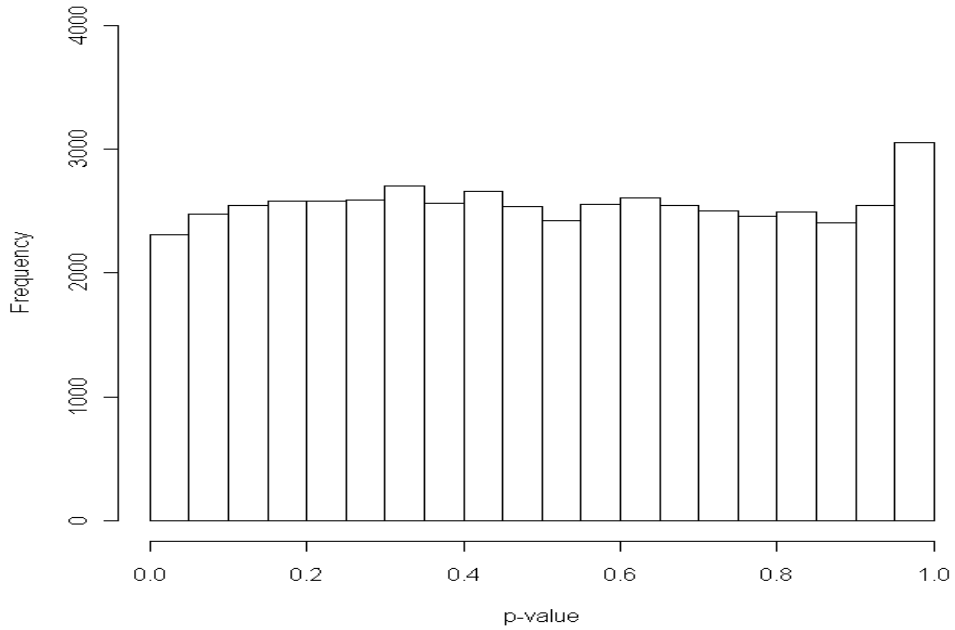

Figure 5.1: Distribution of p-values of SNP

The flat histogram of p-values indicated that the p-value distribution was consistent to the null hypothesis and the p-values were uniformly distributed between 0 and 1. Note that, although there were some individual SNPs with p-values<0.05, their number was as expected under the null hypothesis. For example, after taking into account of the multiple comparison problem by FDR, there was no evidence against the null hypothesis of no association. In summary, the main effect analysis of single SNPs indicated that there was no evidence that any individual SNP alone is associated with the risk of developing CHF.

5.9.1.2 Interaction of SNP and anthracycline treatment

Figure 5.2 shows the results of the SNP-treatment interaction analysis which assessed the modification effect of each individual SNP for anthracycline toxicity on CHF, with the heart radiation exposure indicator as a covariate in the model. As discussed previously, based on the consideration of statistical power, only SNPs which had more than 35 discordant pairs were examined. A total of 35,769 SNP anthracycline interactions (two coding ways together) were examined, among which, 3,288 interactions (accounting for 9.2%) had a p-value less than or equal to 0.05.

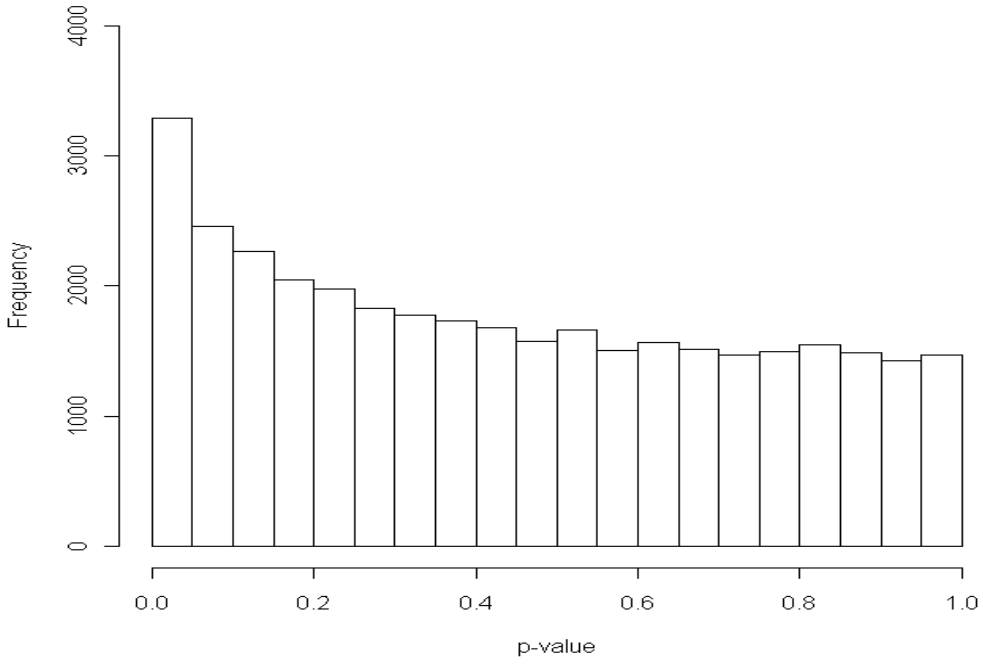

Figure 5.2: Distribution of p-values of the interaction between single SNP and anthracycline dose

The p-value histogram peaked towards small p-values, which is a typical p-value distribution showing the existence of true associations when a large number of potential associations were tested. Thus, this result indicated that the effect of anthracycline dose on the development of CHF was modified by certain SNP genotypes.

Table 5.1 and table 5.2 showed the analysis results of SNPs with the smallest p-values for the interaction. Q-values for these 10 interactions were all 0.35 indicating that the probability of each of these most significant associations being a false discovery was 35%.

Odds ratios (ORs) associated with 100 mg/m2 increase in anthrocycline were also shown. For example, rs2956529 in gene *RAP1B* is the SNP which has the most statistically significant interaction with anthracycline dose with a p-value of 0.00003. The odds ratio for this interaction shows that (a) if a cancer patient has genotype AA in rs2956529, the odds of developing CHF would decrease by a factor of 0.92 with every 100 mg/m^2 increase of anthracycline dose after controlling for heart radiation; and (b) if the genotype is AB or BB, the odds of developing CHF would increase by a factor of 2.34 times with every 100 mg/m^2 increase of anthracycline dose after controlling for heart radiation.

Table 5.1: Ten SNPs with smallest p-values for the interaction of SNP and anthracycline dose

SNPs	Chr position	Gene	OR for 100 mg/m^2 anth (Reference)	OR for 100 mg/m^2 anth (Non-Reference)	p-value	q-value
rs2956529	12q14	RAP1B	0.92	2.34	3.00E-05	0.35
rs2232228	12q22.1	HAS3	3.86	1.41	6.00E-05	0.35
rs2284136	12p13	KCNA5	0.98	2.28	7.00E-05	0.35
rs140615	5q33-q34	SGCHF	1.13	2.43	7.00E-05	0.35
rs17138476	17cen-q21.3	HNF1B	1.44	3.62	9.00E-05	0.35
rs11667974	19q13.4	HAS1	1.46	3.83	1.00E-04	0.35
rs2649747	3p22	N/A	1.13	2.35	1.00E-04	0.35
rs566552	11q22	N/A	2.32	1.21	2.00E-04	0.35
rs2975734	8p23.1	MSRA	1.29	2.53	2.00E-04	0.35
rs11689738	2q31.3	ITGA4	2.72	1.43	2.00E-04	0.35

5.9.2 Simulation result

We also conducted simulations under different assumptions to validate and interpret the above results. The histograms shown in Figures 5.3 and 5.4 are simulated p-value distributions for the 35,769 SNPs (two coding ways together) with underlying OR=1.0 and OR=1.5, respectively.

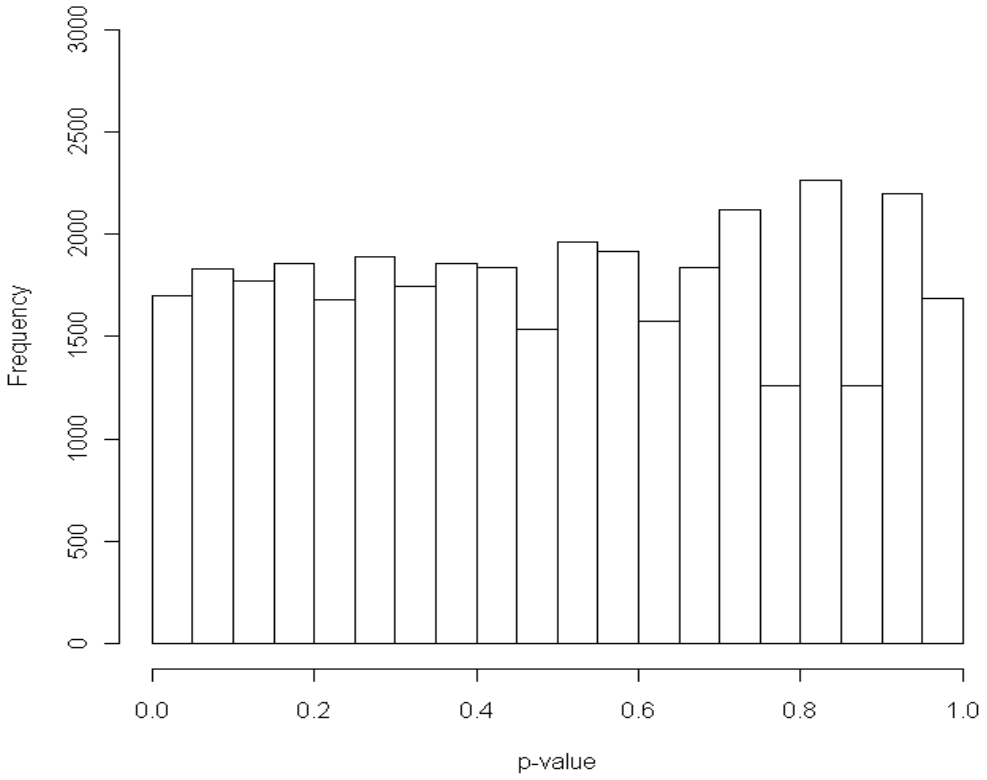

Figure 5.3: Distribution of 35,769 p-values of simulated SNP data with underlying OR=1.0

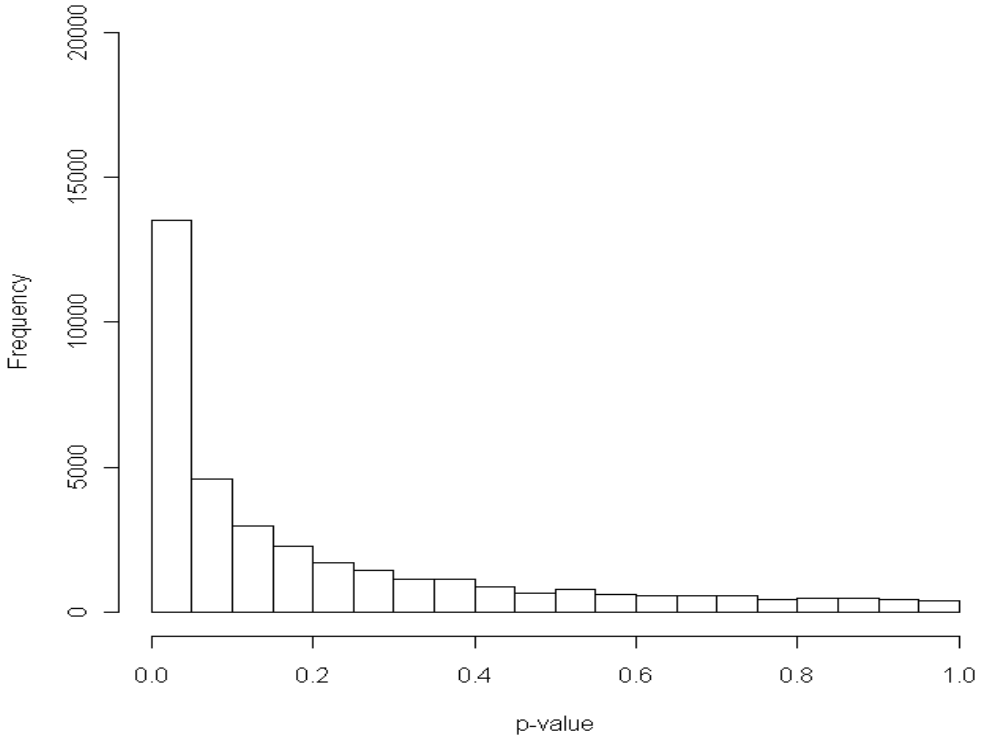

Figure 5.4: Distribution of 35,769 p-values of simulated SNPs with estimated OR=1.5

The p-value density from the simulation with OR=1, shown in Figure 5.3, was flat, consistent to the null hypothesis of no association, and similar to the p-value distribution of the main effect analysis of single SNPs shown in Figure 5.1. Conversely, when the estimated OR was set to be 1.5 which means there were true associations, the p-value distribution was completely different and peaked towards 0. These further confirm that the single-SNP main effect analysis result was consistent to the notion that no single-SNP had a main effect on CHF risk.

Figure 5.5 represents the results of simulations with varying proportions (1%, 3%, 5%, 10% and 15% respectively) of SNPs having associationa with the disease with OR=1.5, with the remaining SNPs having no association with OR=1.0.

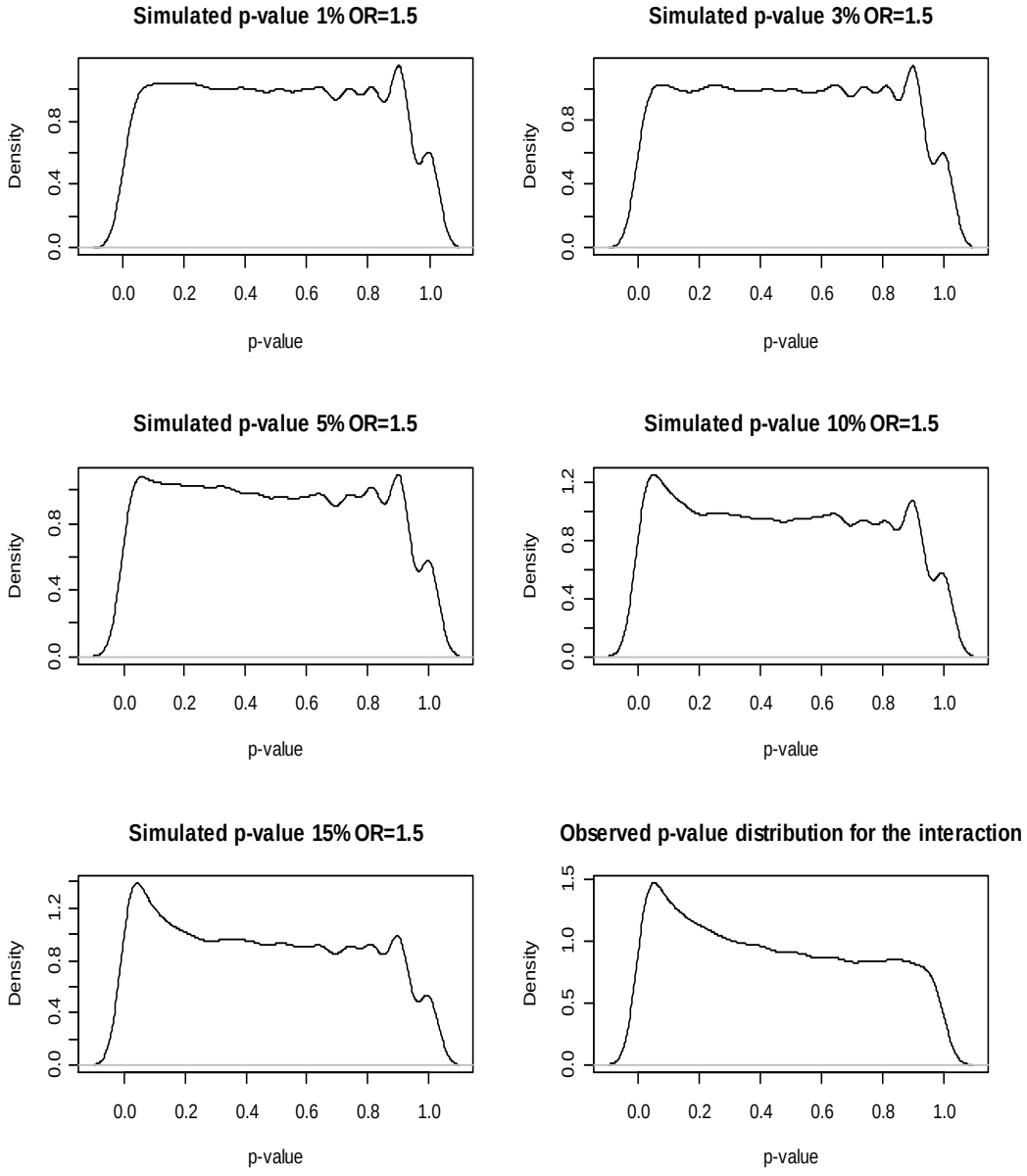

Figure 5.5: Distributions of 35,769 p-values of simulated SNP data with varying OR=1.5 proportions and the observed p-value distribution for the interaction between single SNPs and anthracycline dose

From the observed p-value distribution of the interaction with 35+ discordant pairs, shown in the last panel of Figure 5.5, we observe the consistency of this distribution to the one that has 15% SNPs with OR=1.5. The simulation indicated that there may be approximately 15% of SNPs which may modify the effect of anthracycline dose on CHF risk.

5.9.3 Results of SNP-set analysis

For the 2,991 genes, we assessed the interactions between their SNP-sets and anthracycline. There were 399 interactions with a p-value less than or equal to 0.05, accounting for 13.3% of the genes examined. The distribution of 2,991 p-values is shown in Figure 5.6.

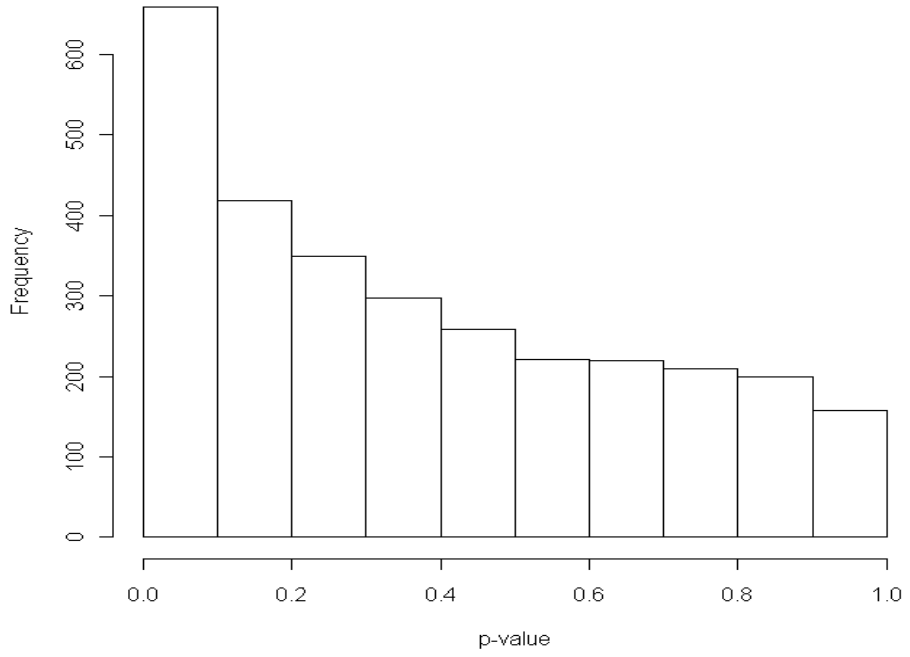

Figure 5.6: Distribution of p-values of interactions between genes and anthracycline dose

This graph indicates there are strong interaction effects between genes and anthracycline dose on CHF risk. Comparing with Figure 5.2 which represented the result of the interaction analysis between single SNPs and anthracycline dose, the distribution of the p-values in Figure 5.6 are much more peaked towards 0. And the proportion of significant genes is much larger than that of significant single SNPs (13.3% vs 9.2%). It shows that the SNP-set interaction analysis resulted in more statistically significant discoveries than the single-SNP interaction analysis.

P-values for SNP-set anthracycline interaction analysis were obtained from 10,000 permutations. The smallest p-value was < 0.0001 as shown in Table 5.3. Both p-values and q-values for the top genes were smaller than the top SNPs from the single-SNP interaction analysis. Among the statistically significant genes, different genes contained varying numbers of SNPs. Some of these SNPs were statistically significant in the single-SNP interaction analysis, while others were not. The reason might be that some SNPs' effects on anthracycline modification are not so significant that would be tested in single SNP analyses. However, their combined effects, identified by SNP-set analyses, do make differences in modifying anthracycline toxicity. In this SNP-set analysis, we identified 5,154 SNPs in significant genes, which accounts about 21.3% of the SNPs examined.

Table 5.2: Top genes with smallest p-values for the interaction of a gene (SNP-set) and anthracycline dose

Gene	Number of SNPs	Gene p-value	Gene q-value	SNP	SNP p-value	SNP q-value	Discordant pairs
CCR8	4	<0.0001	<0.0001				
				rs2649747	1.09E-04	5.51E-01	51
				rs1909560	1.17E-03	5.51E-01	87
				rs13077620	4.39E-03	5.51E-01	85
				rs1909554	3.06E-01	8.29E-01	18
HBEGF	7	<0.0001	<0.0001				
				rs4150196	4.21E-04	5.51E-01	64
				rs7268	9.50E-04	5.51E-01	77
				rs2074613	1.01E-03	5.51E-01	78
				rs13385	1.40E-02	5.77E-01	87
				rs3776089	7.42E-02	6.80E-01	31
				rs4912711	1.11E-01	7.15E-01	69
				rs1862176	3.46E-01	8.47E-01	35
AVPR1A	3	0.0001	0.0551				
				rs7298346	3.22E-04	5.51E-01	20
				rs3803107	4.34E-04	5.51E-01	79
				rs1042615	2.23E-01	7.84E-01	76
HAS1	9	0.0004	0.1201				
				rs11667974	1.10E-04	5.51E-01	78
				rs3829646	1.01E-03	5.51E-01	79
				rs3764545	1.44E-03	5.51E-01	91
				rs4802856	3.92E-02	6.35E-01	30
				rs11672222	1.36E-01	7.34E-01	21
				rs8112223	2.32E-01	7.89E-01	46
				rs10401482	2.51E-01	8.00E-01	63
				rs11084109	3.87E-01	8.66E-01	30
				rs3794977	5.29E-01	9.16E-01	87
KCNA5	7	0.0004	0.1201				
				rs2284136	7.83E-05	5.51E-01	40
				rs11063479	8.97E-04	5.51E-01	83
				rs9788217	1.03E-02	5.70E-01	73
				rs3741930	1.30E-02	5.75E-01	45
				rs1056468	1.45E-01	7.39E-01	91
				rs7961013	2.10E-01	7.77E-01	69
				rs7973471	4.67E-01	8.94E-01	61
CALR	4	0.0006	0.1201				
				rs2293683	5.21E-04	5.51E-01	84

				rs1049481	7.50E-04	5.51E-01	75
				rs1010222	1.29E-03	5.51E-01	83
				rs2974754	8.52E-02	6.89E-01	38
LLGL1	4	0.0007	0.1201				
				rs7498	1.31E-03	5.51E-01	91
				rs2746028	4.06E-03	5.51E-01	51
				rs2245430	1.07E-02	5.73E-01	95
				rs2290507	4.40E-01	8.84E-01	68
BNIP2	4	0.0008	0.1201				
				rs2307275	8.70E-04	5.51E-01	86
				rs6151589	9.66E-04	5.51E-01	84
				rs11854063	1.50E-03	5.51E-01	82
				rs6151592	8.43E-03	5.67E-01	74
NCAPH	1	0.0008	0.1201				
				rs17633463	1.22E-03	5.51E-01	89
RAP1B	4	0.0008	0.1201				
				rs2956529	2.53E-05	5.51E-01	37
				rs2439759	1.43E-02	5.77E-01	88
				rs11177315	1.61E-01	7.50E-01	89
				rs2468430	6.31E-01	9.48E-01	59

Chapter 6 Discussion and Conclusion

6.1 Validity of the current study

This study was designed by Dr. Bhatia in cooperation with the Children's Oncology Group (COG). Children's Oncology Group (COG) is the world's largest, cooperative children's cancer research entity and it establishes a large cohort for studying childhood cancer survivors (98). The underlying cohort of cancer survivors and the sampling of cases and controls from it alleviated potential selection bias.

The study was a matched case-control study which is a common method in epidemiological studies. Since matched sampling can control for the effects of the matching variables in the design stage, it is an effective way for eliminating potential confounding and the need to adjust for it in the analysis stage. This is especially advantageous when the sample size is not large.

There were 3,288 SNPs that had p-values less than or equal to 0.05 for the SNP-anthracycline-dose interaction on CHF risk, which accounted for 9.2% of all SNPs examined in the single-SNP interaction analysis. Moreover, 399 genes, i.e., 13.3% of all genes examined in the SNP-set interaction analysis, had p-values less than or equal to 0.05 for the gene-anthracycline-dose interaction. After considering the multiple comparison problem by FDR estimates (q-values), there were statistically significant interactions between some genes and anthracycline dose on CHF risk.

A further literature review exploring underlying biological mechanisms for several reported SNPs and genes was conducted to substantiate the biological validity. In our analysis results, gene KCNA5 (p=0.0004, q=0.12) was one of the most statistically significant genes reported in Table 5.3. There are studies show that mutation in gene KCNA5 could cause cardiac arrhythmia (100-102). HAS1 (p=0.0004, q=0.12) and HAS3 (p=0.0015, q=0.12) were also among the significant findings from the SNP-set analysis. Hyaluronan (HA) is an obligatory component in the ventricular myocardium and plays an important role in experimental myocardial infarction (103). HA is expressed by three different hyaluronan synthases which are encoded in genes HAS1, HAS2 and HAS3(103,104). HAS2 has been proven to be associated with cardiovascular morphogenesis (104,105). A study conducted by Urban *et al.* (2007) also showed that there was an association of genes HAS1 and HAS2 with cardiac hypertrophy (105), which is in agreement with our finding on HAS1. Although we are not aware of any direct evidence to support HAS3's involvement with anthracycline-associated CHF, our finding may suggest a new direction to examine the role of HAS3 in modifying anthracycline's toxicity in the development of CHF. Some other genes which are known to have an association with cardiac disease, such as AVPR1A (p=0.0001, q=0.06) (106,107), were also validated in our study.

6.2 Impacts of the study

To our knowledge, this study led by Dr. Bhatia is the first comprehensive GWAS study which attempted to explore genetic susceptibility for anthracycline-induced CHF among childhood cancer survivors. There was a candidate gene study

examined whether common polymorphisms in gene CBR3 and NQO1 had an impact on the risk of anthracycline-related CHF (108). But these two candidate genes were selected based on the prior knowledge of their biological functions on heart functions. Our study explored a much larger number of genes over the whole genome, which is one of GWAS studies' advantages over candidate gene studies. Our approach may lead to unexpected, but important genetic findings.

The main aim of this study was to explore interactions between genetic factors and cancer treatment in developing CHF. Family studies have shown genetic factors play an important role in cardiac diseases (109). Environmental effect and life style habits are also highly related to CHF. Thus, studying CHF with multiple genetic factors, environmental factors, and their interaction is important for improving accuracy and precision in the assessment of both genetic and environmental influences.

As a promising direction of GWAS-based genetic research, SNP-set analysis, which grouped SNPs in a gene into a SNP set and tested the joint effect of the SNPs in the set, was proposed in this study. In GWAS, single-SNP analysis is the standard mode of analysis, but there are several problems/limitations with it.

One of the problems of single SNP analysis is that it could miss the SNPs which are truly associated with the disease due to the multiple comparison problem. Because a large number of SNPs is tested in a typical GWAS study and most of them are expected to have no association with the phenotype of interest. If we apply Bonferroni adjustment to control Type I errors, the threshold for

significance would be very stringent. For example, if there are 50,000 SNPs and the overall target of Type I error is 5%, the cut-off of p-value will be set to 10^{-6}. True causal SNPs with p-values those are not as extremely significant as this stringent threshold would not be discovered. Given the large number of null SNPs which account for the dominant majority, when FDR is applied, the q-value of the truly causal SNPs will be high and they will not be discovered.

The SNP-set analysis is advantageous in controlling for Type I errors (i.e., false positive discoveries) over the single-SNP analysis because it has a much smaller number of tests to be performed. In our study, gene-set analysis grouped 24,224 individual SNPs that were examined by the single SNP analysis into 2,991 SNP-sets.

Secondly, the single-SNP analysis would not be powerful enough to detect the SNPs with moderate effects that are truly associated with the outcome. Some disease phenotypes may be caused by the interactions of SNPs which have moderate effect at different loci (110,111). When we check these SNPs with small effects individually, they may be missed due to low power, especially when the sample size is not enough. The SNP-set analysis which considers the joint of effect of the SNPs in the same gene can be capable of capturing the truly associated SNPs. Thus, SNP-set analysis can improve the statistical power of GWAS.

Thirdly, from the biological point of view, a single SNP may not have appreciable effects compared to the gene it belongs to on the development of the disease.

Therefore, the joint effect of a set of SNPs based on prior biological criteria (i.e., SNPs in the same gene or SNPs of genes with similar biological functions) is more meaningful, interpretable, and likely to show underlying associations if they exist.

In view of the limitations of the single-SNP analysis above, the SNP-set analysis which included the SNPs in the same gene can be complementary and/or advantageous in overcoming these restrictions. Moreover, it's easier to interpret the result of the SNP-set analysis biologically than that of the single-SNP analysis due to grouping of multiple SNPs under a biological concept.

6.3 Limitations

There are several limitations in this project. First, the sample size of our study may be not large enough to assess SNPs or genes if their interaction with anthracycline dose on CHF is moderate. Since GWAS scan thousands of SNPs or SNP sets in the genome, in order to sustain statistical significance even after a multiple-comparison correction, the associations must achieve a very high level of statistical significance for which a large number of subjects are needed. For example, in the GWAS studies conducted by WTCCC, more than 2,000 case subjects and 3,000 control subjects were enrolled (63). In our study, the sample was composed of 130 cases and 266 matched controls which past the quality control process: this may not have produced adequate statistical power.

There is a key assumption in the method we applied in the SNP-set analysis, which is a limitation of our method. In the SNP-set analysis, we constructed a test

statistic for the interaction between a SNP-set and anthracycline dose by summing up the chi-square test statistic of the interaction of each single SNP in the same gene and anthracycline dose. The underlying assumption for this method is that the effect of a gene on CHF is equal to the sum of the effects of its single SNPs on CHF. For this to hold, the causal effect must be observable for single SNPs: if the effect is caused by SNP-SNP interactions, for example, this may not hold. In addition, this assumption does not reflect the underlying mechanism properly when some SNPs modify anthracycline effects on CHF, but others do not. This possibility of heterogeneity in a SNP set with respect to the effect of interest was not considered in our approach and could make our approach inefficient (less powerful).

6.4 Conclusion

In conclusion, there was strong evidence that the associations of anthracycline dose with CHF risk among childhood cancer survivors were modified according to genotypes of SNPs; and the SNP-set interaction analysis showed interactions between certain genes and therapeutic exposure of anthracycline exist. Results from our study should aid the understanding of the mechanism of genetic factors and anthracycline exposure interaction on the development of CHF after surviving childhood cancer.

6.5 Future work

The work in this study is an initial step in exploring the interplaying mechanism of gene and cancer treatments in developing anthracycline-induced CHF for

childhood cancer survivors. The following possible directions may be considered for further work.

In order to confirm the current findings of our study, a replication study based on an independent set of cases and controls could be conducted. Permutation was applied to obtain the null distribution of the test statistic for an interaction between gene and anthracycline dose. However, a recent study suggested that exact permutation test for gene-gene and gene-environment interaction may be difficult to obtain in GWAS study (112). This may limit our power for finding the truly important interaction. Therefore, future research needs to assess this issue and clarify the validity and efficiency of our approach.

It may be valuable to examine gene expression levels of the significant genes from our study in relation to CHF. Assessing gene expression, an intermediate biological process between DNA and the disease, by measuring mRNA levels of the genes we found associated with CHF would give additional insights into the molecular mechanism. Thus a gene expression study could provide functional biological evidence in validating our study findings.

References

(1) Barr RD, Greenberg ML, Shaw AK, Mery L. The Canadian Childhood Cancer Surveillance and Control Program (CCCSCP): a status report. Pediatr Blood Cancer 2008 Feb;50(2 Suppl):518-519.

(2) Bountioukos M, Doorduijn JK, Roelandt JR, Vourvouri EC, Bax JJ, Schinkel AF, et al. Repetitive dobutamine stress echocardiography for the prediction of anthracycline cardiotoxicity. Eur J Echocardiogr 2003 Dec;4(4):300-305.

(3) Valcarcel D, Montesinos P, Sanchez-Ortega I, Brunet S, Esteve J, Martinez-Cuadron D, et al. A scoring system to predict the risk of death during induction with anthracycline plus cytarabine-based chemotherapy in patients with de novo acute myeloid leukemia. Cancer 2011 Jun 29.

(4) Iarussi D, Indolfi P, Casale F, Martino V, Di Tullio MT, Calabro R. Anthracycline-induced cardiotoxicity in children with cancer: strategies for prevention and management. Paediatr Drugs 2005;7(2):67-76.

(5) Childhood Acute Lymphoblastic Leukaemia Collaborative Group (CALLCG). Beneficial and harmful effects of anthracyclines in the treatment of childhood acute lymphoblastic leukaemia: a systematic review and meta-analysis. Br J Haematol 2009 May;145(3):376-388.

(6) Sieswerda E, Kremer LC, Vidmar S, De Bruin ML, Smibert E, Sjoberg G, et al. Exercise echocardiography in asymptomatic survivors of childhood cancer treated with anthracyclines: a prospective follow-up study. Pediatr Blood Cancer 2010 Apr;54(4):579-584.

(7) O'Shaughnessy J. Liposomal anthracyclines for breast cancer: overview. Oncologist 2003;8 Suppl 2:1-2.

(8) Dempfle A, Scherag A, Hein R, Beckmann L, Chang-Claude J, Schafer H. Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur J Hum Genet 2008 Oct;16(10):1164-1172.

(9) Srivastava K, Srivastava A, Sharma KL, Mittal B. Candidate gene studies in gallbladder cancer: A systematic review and meta-analysis. Mutat Res 2011 Jun 25.

(10) Dick DM. Gene-environment interaction in psychological traits and disorders. Annu Rev Clin Psychol 2011 Apr;7:383-409.

(11) David J.Hunter. Gene-Environment Interactions in Human Diseases. April 2005;6:287-288-298.

(12) Mavinkurve-Groothuis AM, Kapusta L, Nir A, Groot-Loonen J. The role of biomarkers in the early detection of anthracycline-induced cardiotoxicity in children: a review of the literature. Pediatr Hematol Oncol 2008 Sep;25(7):655-664.

(13) National Human Genome Research Institute (NHGRI). Deoxyribonucleic Acid (DNA). Available at: http://www.genome.gov/25520880. Accessed 09/15, 2010.

(14) National DNA Data Bank (NDDB). Available at: http://www.nddb-bndg.org/main_e.htm. Accessed 09/15, 2010.

(15) Genetics Home Reference. Available at: http://ghr.nlm.nih.gov/handbook/basics/dna.

(16) National Human Genome Research Institute (NHGRI). Available at: http://www.genome.gov/10001772. Accessed 09/15, 2010.

(17) Shastry BS. SNP alleles in human disease and evolution. J Hum Genet 2002;47(11):561-566.

(18) Li J, Pan YC, Li YX, Shi TL. Analysis and application of SNP and haplotype in the human genome. Yi Chuan Xue Bao 2005 Aug;32(8):879-889.

(19) Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998 May 15;280(5366):1077-1082.

(20) Wikipedia. Single-nucleotide polymorphism. Available at: http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism. Accessed 11/11, 2010.

(21) Genetics Home Reference. Available at: http://ghr.nlm.nih.gov/handbook/genomicresearch/snp. Accessed 09/15, 2010.

(22) Syvanen AC. Toward genome-wide SNP genotyping. Nat Genet 2005 Jun;37 Suppl:S5-10.

(23) Hemminki K, Bermejo JL. Relationships between familial risks of cancer and the effects of heritable genes and their SNP variants. Mutat Res 2005 Dec 30;592(1-2):6-17.

(24) Gura T. Genetics. SNP-ing drugs to size. Science 2001 Jul 27;293(5530):595.

(25) Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature 2011 May 19;473(7347):337-342.

(26) Shu L, Lou Q, Ma C, Ding W, Zhou J, Wu J, et al. Genetic, proteomic, and metabolic analysis of the regulation of energy storage in rice seedlings in response to drought. Proteomics 2011 Aug 5.

(27) Shu Y, Hong-Hui L. Transcription, translation, degradation, and circadian clock. Biochem Biophys Res Commun 2004 Aug 13;321(1):1-6.

(28) Zenklusen D, Stutz F. Nuclear export of mRNA. FEBS Lett 2001 Jun 8;498(2-3):150-156.

(29) Guo L, Liu Y, Bai Y, Sun Y, Xiao F, Guo Y. Gene expression profiling of drug-resistant small cell lung cancer cells by combining microRNA and cDNA expression analysis. Eur J Cancer 2010 Jun;46(9):1692-1702.

(30) Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 2001 Dec;25(4):402-408.

(31) Ballman KV. Genetics and genomics: gene expression microarrays. Circulation 2008 Oct 7;118(15):1593-1597.

(32) Zhang Z, Zeng D, Ma H, Feng G, Hu J, He L, et al. A DNA-Origami chip platform for label-free SNP genotyping using toehold-mediated strand displacement. Small 2010 Sep 6;6(17):1854-1858.

(33) Gunderson KL, Steemers FJ, Ren H, Ng P, Zhou L, Tsan C, et al. Whole-genome genotyping. Methods Enzymol 2006;410:359-376.

(34) LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res 2009 Jul;37(13):4181-4193.

(35) Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 2005 May;37(5):549-554.

(36) Kerstens HH, Crooijmans RP, Veenendaal A, Dibbits BW, Chin-A-Woeng TF, den Dunnen JT, et al. Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. BMC Genomics 2009 Oct 16;10:479.

(37) National Human Genome Research Institute (NHGRI). About the International HapMap Project. Available at: http://www.genome.gov/11511175. Accessed 09/20, 2010.

(38) International HapMap Consortium. A haplotype map of the human genome. Nature 2005 Oct 27;437(7063):1299-1320.

(39) Maniatis N. Linkage disequilibrium maps and disease-association mapping. Methods Mol Biol 2007;376:109-121.

(40) Morton NE. Linkage disequilibrium maps and association mapping. J Clin Invest 2005 Jun;115(6):1425-1430.

(41) Gunderson KL, Kuhn KM, Steemers FJ, Ng P, Murray SS, Shen R. Whole-genome genotyping of haplotype tag single nucleotide polymorphisms. Pharmacogenomics 2006 Jun;7(4):641-648.

(42) Foulkes AS. Applied statistical genetics with R: For Population-based Association Studies, Use R. : Springer Science+Business Media LLC; 2009.

(43) Tenenbaum SA, Carson CC, Lager PJ, Keene JD. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. Proc Natl Acad Sci U S A 2000 Dec 19;97(26):14085-14090.

(44) National Human Genome Research Institute (NHGRI). Available at: http://www.genome.gov/10000533. Accessed 09/25, 2010.

(45) Max Animations. cDNA. Available at: http://www.maxanim.com/genetics/cDNA/cDNA.htm. Accessed 09/25, 2010.

(46) Sealfon SC, Chu TT. RNA and DNA microarrays. Methods Mol Biol 2011;671:3-34.

(47) Mancheron A, Uricaru R, Rivals E. An alternative approach to multiple genome comparison. Nucleic Acids Res 2011 Jun 6.

(48) Rosenberg PS, Che A, Chen BE. Multiple hypothesis testing strategies for genetic case-control association studies. Stat Med 2006 Sep 30;25(18):3134-3149.

(49) Richard J.Larsen, Morris L.Marx. An Introduction to Mathematical Statistics and Its Applications. ; 2006.

(50) Kulldorff M, Graubard B, Velie E. The P-value and P-value function. Epidemiology 1999 May;10(3):345-347.

(51) Moskvina V, Schmidt KM. On multiple-testing correction in genome-wide association studies. Genet Epidemiol 2008 Sep;32(6):567-573.

(52) Morgan JF. p Value fetishism and use of the Bonferroni adjustment. Evid Based Ment Health 2007 May;10(2):34-35.

(53) Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998 Apr 18;316(7139):1236-1238.

(54) Gao X, Becker LC, Becker DM, Starmer JD, Province MA. Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol 2010 Jan;34(1):100-105.

(55) Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res 2001 Nov 1;125(1-2):279-284.

(56) Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics 2005 Jul 1;21(13):3017-3024.

(57) Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003 Aug 5;100(16):9440-9445.

(58) Storey JD. The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. The Annals of Statistics 2003 Dec.;31(6):pp. 2013-2035.

(59) Overfitting Prevention with Cross-Validation. ; January; ; 2007.

(60) Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biom J 2008 Feb;50(1):8-28.

(61) Yan WL. Genome-wide association study on complex diseases: genetic statistical issues. Yi Chuan 2008 May;30(5):543-549.

(62) Seng KC, Seng CK. The success of the genome-wide association approach: a brief story of a long struggle. Eur J Hum Genet 2008 May;16(5):554-564.

(63) Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007 Jun 7;447(7145):661-678.

(64) Tu X, Shi LS, Wang F, Wang Q. Genomewide association study: advances, challenges and deliberation. Sheng Li Ke Xue Jin Zhan 2010 Apr;41(2):87-94.

(65) Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA 2008 Mar 19;299(11):1335-1344.

(66) Teo YY. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. Curr Opin Lipidol 2008 Apr;19(2):133-143.

(67) Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 2007 Jun 1;316(5829):1336-1341.

(68) Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 2007 Jun 1;316(5829):1341-1345.

(69) Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 2009 May;41(5):585-590.

(70) Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 1992 Jun;48(2):361-372.

(71) Li M, Li C. Assessing departure from Hardy-Weinberg equilibrium in the presence of disease association. Genet Epidemiol 2008 Nov;32(7):589-599.

(72) Application of chi-square test and exact test in Hardy-Weinberg equilibrium testing. Fa Yi Xue Za Zhi 2004;20(2):116-119.

(73) Goddard KA, Ziegler A, Wellek S. Adapting the logical basis of tests for Hardy-Weinberg Equilibrium to the real needs of association studies in human and medical genetics. Genet Epidemiol 2009 Nov;33(7):569-580.

(74) Enoch MA, Shen PH, Xu K, Hodgkinson C, Goldman D. Using ancestry-informative markers to define populations and detect population stratification. J Psychopharmacol 2006 Jul;20(4 Suppl):19-26.

(75) Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 2005 Nov;37(11):1243-1246.

(76) Wang K. Testing for genetic association in the presence of population stratification in genome-wide association studies. Genet Epidemiol 2009 Nov;33(7):637-645.

(77) Tebbutt SJ, He JQ, Burkett KM, Ruan J, Opushnyev IV, Tripp BW, et al. Microarray genotyping resource to determine population stratification in genetic association studies of complex disease. BioTechniques 2004 Dec;37(6):977-985.

(78) Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999 Dec;55(4):997-1004.

(79) Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006 Aug;38(8):904-909.

(80) Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001 Apr 24;98(9):5116-5121.

(81) Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005 Oct 25;102(43):15545-15550.

(82) Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics 2009 May 27;10:161.

(83) Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, et al. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics 2007 Jul 5;8:242.

(84) Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y. Comparative evaluation of gene-set analysis methods. BMC Bioinformatics 2007 Nov 7;8:431.

(85) Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, et al. Gene-set analysis and reduction. Brief Bioinform 2009 Jan;10(1):24-34.

(86) Wang X, Dinu I, Liu W, Yasui Y. Linear combination test for hierarchical gene set analysis. Stat Appl Genet Mol Biol 2011;10(1):Article 13.

(87) Schafer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 2005;4:Article32.

(88) Lage-Castellanos A, Martinez-Montes E, Hernandez-Cabrera JA, Galan L. False discovery rate and permutation test: an evaluation in ERP data analysis. Stat Med 2010 Jan 15;29(1):63-74.

(89) Bliss RY, Weinberg J, Vieira V, Ozonoff A, Webster TF. Power of permutation tests using generalized additive models with bivariate smoothers. J Biom Biostat 2010 Sep 12;1(104):1000104.

(90) Taub JW. Factors in improved survival from paediatric cancer. Drugs 1998 Nov;56(5):757-765.

(91) Mott MG. Neoplasia in childhood--25 years of progress. Ann Oncol 1995;6 Suppl 1:3-8; discussion 8-9.

(92) Edelstein K, D'agostino N, Bernstein LJ, Nathan PC, Greenberg ML, Hodgson DC, et al. Long-term Neurocognitive Outcomes in Young Adult Survivors of Childhood Acute Lymphoblastic Leukemia. J Pediatr Hematol Oncol 2011 Aug;33(6):450-458.

(93) von der Weid NX. Adult life after surviving lymphoma in childhood. Support Care Cancer 2008 Apr;16(4):339-345.

(94) Canadian Cancer Society. Canadian Cancer Statistics 2008 Childhood Cancer in Canada : Fast Facts. 2008.

(95) van Dalen EC, van der Pal HJ, Kok WE, Caron HN, Kremer LC. Clinical heart failure in a cohort of children treated with anthracyclines: a long-term follow-up study. Eur J Cancer 2006 Dec;42(18):3191-3198.

(96) Kremer LC, van Dalen EC, Offringa M, Ottenkamp J, Voute PA. Anthracycline-induced clinical heart failure in a cohort of 607 children: long-term follow-up study. J Clin Oncol 2001 Jan 1;19(1):191-196.

(97) Iarussi D, Indolfi P, Galderisi M, Bossone E. Cardiac toxicity after anthracycline chemotherapy in childhood. Herz 2000 Nov;25(7):676-688.

(98) Children's Oncology Group (COG) Available at: http://www.curesearch.org/ArticleView2.aspx?id=8917. Accessed 11/15, 2010.

(99) Keating BJ, Tischfield S, Murray SS, Bhangale T, Price TS, Glessner JT, et al. Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS One 2008;3(10):e3583.

(100) Nielsen NH, Winkel BG, Kanters JK, Schmitt N, Hofman-Bang J, Jensen HS, et al. Mutations in the Kv1.5 channel gene KCNA5 in cardiac arrest patients. Biochem Biophys Res Commun 2007 Mar 16;354(3):776-782.

(101) Remillard CV, Tigno DD, Platoshyn O, Burg ED, Brevnova EE, Conger D, et al. Function of Kv1.5 channels and genetic variations of KCNA5 in patients

with idiopathic pulmonary arterial hypertension. Am J Physiol Cell Physiol 2007 May;292(5):C1837-53.

(102) Platoshyn O, Brevnova EE, Burg ED, Yu Y, Remillard CV, Yuan JX. Acute hypoxia selectively inhibits KCNA5 channels in pulmonary artery smooth muscle cells. Am J Physiol Cell Physiol 2006 Mar;290(3):C907-16.

(103) Adamia S, Maxwell CA, Pilarski LM. Hyaluronan and hyaluronan synthases: potential therapeutic targets in cancer. Curr Drug Targets Cardiovasc Haematol Disord 2005 Feb;5(1):3-14.

(104) Camenisch TD, Spicer AP, Brehm-Gibson T, Biesterfeldt J, Augustine ML, Calabro A,Jr, et al. Disruption of hyaluronan synthase-2 abrogates normal cardiac morphogenesis and hyaluronan-mediated transformation of epithelium to mesenchyme. J Clin Invest 2000 Aug;106(3):349-360.

(105) Hellman U, Hellstrom M, Morner S, Engstrom-Laurent A, Aberg AM, Oliviero P, et al. Parallel up-regulation of FGF-2 and hyaluronan during development of cardiac hypertrophy in rat. Cell Tissue Res 2008 Apr;332(1):49-56.

(106) Li X, Chan TO, Myers V, Chowdhury I, Zhang XQ, Song J, et al. Controlled and Cardiac-Restricted Overexpression of the Arginine Vasopressin V1A Receptor Causes Reversible Left Ventricular Dysfunction Through G{alpha}q-Mediated Cell Signaling. Circulation 2011 Aug 2;124(5):572-581.

(107) Hiroyama M, Wang S, Aoyagi T, Oikawa R, Sanbe A, Takeo S, et al. Vasopressin promotes cardiomyocyte hypertrophy via the vasopressin V1A receptor in neonatal mice. Eur J Pharmacol 2007 Mar 22;559(2-3):89-97.

(108) Blanco JG, Leisenring WM, Gonzalez-Covarrubias VM, Kawashima TI, Davies SM, Relling MV, et al. Genetic polymorphisms in the carbonyl reductase 3 gene CBR3 and the NAD(P)H:quinone oxidoreductase 1 gene NQO1 in patients who developed anthracycline-related congestive heart failure after childhood cancer. Cancer 2008 Jun 15;112(12):2789-2795.

(109) Yoshitama T, Nakao S, Takenaka T, Teraguchi H, Sasaki T, Kodama C, et al. Molecular genetic, biochemical, and clinical studies in three families with cardiac Fabry's disease. Am J Cardiol 2001 Jan 1;87(1):71-75.

(110) Suh Y, Vijg J. SNP discovery in associating genetic variation with human disease phenotypes. Mutat Res 2005 Jun 3;573(1-2):41-53.

(111) Karchin R, Kelly L, Sali A. Improving functional annotation of non-synonomous SNPs with information theory. Pac Symp Biocomput 2005:397-408.

(112) Buzkova P, Lumley T, Rice K. Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. Ann Hum Genet 2011 Jan;75(1):36-45.

Appendix 1: R code for simulation

```
library(survival)

## input the SNPs data to get the number of SNPs ##

## --aa as the reference --##

ct=read.table("Z:\\Common\\wei\\thesis
project\\send2dr_batia\\CD\\interaction\\aa.txt",header=T,sep="\t")

n1=17123

############################################################

## --bb as the reference --##

#ct=read.table("Z:\\Common\\wei\\thesis
#project\\send2dr_batia\\CD\\interaction\\bb.txt",header=T,sep="\t")

#n1=18646

############################################################

count=c(ct[1:n1,1])

n2=length(count)

###### “pi” which represents estimated odds ratio ###########

pi=1.5

num=130

pvalue1=rep(0,n2)

for (j in 1:n2)

    {

     x=count[j]

     y=rbinom(x,1,pi/(1+pi))

     set1_1=cbind(c(1:x),y,rep(1,x))
```

```
y=1-y
set1_2=cbind(c(1:x),y,rep(0,x))
set1=rbind(set1_1,set1_2)
p=runif(1,0,1)
s1=rbinom(130-x,1,p)
set2_1=cbind(c((x+1):num),rep(1,(num-x)),s1)
set2_2=cbind(c((x+1):num),rep(0,(num-x)),s1)
set2=rbind(set2_1,set2_2)
set=rbind(set1,set2)
set=as.data.frame(set)
colnames(set)=c("match_set","status","snp")
reg=clogit(status~snp+strata(match_set),data=set)
sum=summary(reg)
sum1=sum$coefficients
## get the simulated p-values ##
pvalue1[j]=sum1[,5]
} ############end of j
```

Appendix 2: R code for permutation

```
library(survival)

### input clinical data ###

clin1=read.table()

### input genotype data ###

genetype1=read.table()


genetype2=genetype1[,-c(401:405)]

genetype3=rbind(as.matrix(genetype2),as.matrix(clin1))

genetype4=t(genetype3)


names=genetype4[1,]

colnames(genetype4)=names

set=as.data.frame(genetype4[-1,])

c1=c(1:dim(set)[1])

set1=cbind(set[,24216:24221],c1)

case1=set[set$status==1,]

match_set=case1$match_set

pair1=as.data.frame(match_set)

pair2=pair1[,1]

pair3=as.character(pair2)

pair4=as.numeric(pair3)
```

```
### Make a matrix of 130 rows where each row contains C1 values of
set1$match_set==k ###

c1list=matrix(NA,130,10)

for(d in 1:130){

    k=pair4[d]

    junk=set1[set1$match_set==k,]$c1

    c1list[d,1:length(junk)]=junk

    }


### Define a ranodm sampling function ###

sample1=function(x)

            {

              sample(x[!is.na(x)],1)

                 }


##--times of permutation--##

nperm=10000

##--number of SNPs--##

nsnp=24242

###################################

stat=matrix(0,nsnp,nperm)

for ( i in 1:nperm)

   {

        perm1=apply(c1list,1,sample1)     #### apply sample function #####

#-----newcase dataset------#
```

```
permcase1=set[perm1,]
#-----newcontrol dataset----#
perm2=c1[-perm1]
permcontrol1=set[perm2,]

pcase=permcase1[,c(24217: 24221)]
pcontrol=permcontrol1[,c(24217: 24221)]

###--prepare the variables needed for regression--###
dat1=rbind(pcase,pcontrol)
disease=c(rep(1,130),rep(0,269))
match_set=dat1$match_set
anth1=as.character(dat1$totalanth)
anth=as.numeric(anth1)
rd=dat1$rd
rd=as.numeric(as.character(rd))
##########################################
for (p in 1:nsnp)
   {
     case=permcase1[,p]
     control=permcontrol1[,p]
     snp=c(case,control)
     ## some of the regression don't converge and the loop would stop, ###
     ## insert this program to make the permutation continue          ###
     aa = try({
```

```
        reg=clogit(disease ~ snp+anth+rd+snp*anth+strata(match_set))
          })### end of try


if  (class(aa)[1]=="try-error") stat[p,i]=NA
   else{
        sum1=summary(aa)$coefficients
        stat[p,i]=(sum1[4,4])^2
              } ## end of if
            }   ## end of p
}## end of i
```