國立交通大學

統計學研究所

碩士論文

相加模型下藉由單獨的單一核甘酸多形性關係探測其交互作用的趨勢

Detecting Interaction Patterns Based on Single SNP Association Under Additive Model

研 究 生：許庭瑋

指導教授：盧鴻興　教授

中 華 民 國 九 十 八 年 六 月

Detecting Interaction Patterns Based on Single SNP Association Under Additive Model

Student: Ting-Wei Hsu
Advisor: Prof. Henry Horng-Shing Lu

A Thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics

Institute of Statistics, National Chiao Tung University
Hsinchu, Taiwan, Republic of China

June, 2009

相加模型下藉由單獨的單一核甘酸多形性關係探測其交互作用的趨勢

研 究 生： 許庭瑋

指導教授： 盧鴻興 教授

國立交通大學統計學研究所

摘要

此篇論文包含了兩個部分，針對相加模型下藉由單獨的單一核甘酸多形性(SNP)關係探測其交互作用的趨勢，而我們方法的重點在於檢定力的損失與節省運算時間的權衡。

在 GWAS 探討交互作用關係的運算時間是相當驚人的，我們首先找出單獨 SNP 關係與配對 SNP 關係的關聯，希望透過損失一些檢定力，使得運算時間能大幅降低。研究中的第二部分是利用條件最大期望值 (ECM)來估計在實際資料中的 λ_{AB}(基因型 AB 的相對外顯率)、f_A(對偶基因 A 的頻率) 、f_B(對偶基因 B 的頻率)，並且可藉由估計值來計算檢定力的損失。

型一誤差(α)與型二誤差(β)之拉扯乃統計假設檢定中著名的問題，然而，在 GWAS 中做多重檢定，5×10^7 或1×10^5 這類的型一誤差是相當常見的，如此一來檢定力($1-\beta$)由於型二誤差很大而變得非常差。換句話說，當使用很小的型一誤差時，會使得假設檢定的結果過於保守。

利用此方法來分析 WTCCC 所提供之高血壓的資料，我們偵測到已有文獻提及與高血壓有關的一些基因或 SNP，諸如 CHRM2 (rs7800093), KCNB2 (rs11782342), HTR3B (rs17116117), rs2820037, GAB1 (rs300916, rs300915, rs300913), BCAT1 (rs7961152, rs11613673, rs12424348), MYBPC1 (rs11110912)。然而也有一些是至今尚未發現的，如 rs825148, rs1553460, LOC100129858 (rs6840033), rs4131463, RPL18P4 (rs1528356), rs17797701, OTOG (rs11024327), rs10843660, CHST11 (rs11112069), SIP1 (rs8011855), RHOJ (rs1957779)這些值得將來繼續深入研究的基因或 SNP。

關鍵字：Loss of power, expectation-conditional maximization, genome-wide association study, single nucleotide polymorphism, additive model, hypertension

Detecting Interaction Patterns Based on Single SNP Association Under Additive Model

Student: Ting-Wei Hsu
Advisor: Prof. Henry Horng-Shing Lu

Institute of Statistics
National Chiao Tung University

Abstract

This thesis consists of two main parts for detecting interaction patterns based on single nucleotide polymorphism (SNP) association under additive model. Our approach is focused on the trade-off between loss of power and the reduction in computation time.

The computation time for interaction association in genome-wide association study (GWAS) is usually tremendous. Our first task is to find the relation between single SNP association and paired SNPs association such that computation time could be greatly reduced through some loss of power.

In the second research area, expectation-conditional maximization (ECM) algorithm is used to estimate λ_{AB} (relative penetrance rate for genotype AB), f_A (allele frequency A), f_B (allele frequency B) in real genome-wide association study, and consequently provide reasonable parameters for estimating the loss of power.

The trade-off for α (type I error) and β (type II error) is well-known in statistical hypothesis testing. However, a small α such as $5 \times 10^{-7}, 1 \times 10^{-5}$ are used often in case-control association study since in multiple testing, the power $(1-\beta)$ will be badly weakened due to large β. In other words, a small α makes hypothesis testing over-conservative.

Analyzing data with this approach, which imitates WTCCC of hypertension, we have detected parts of known genes or SNPs, such as CHRM2 (rs7800093), KCNB2 (rs11782342), HTR3B (rs17116117), rs2820037, GAB1 (rs300916, rs300915, rs300913), BCAT1 (rs7961152, rs11613673, rs12424348), MYBPC1 (rs11110912). Nevertheless, we have also detected unknowns, such as rs825148, rs1553460, LOC100129858 (rs6840033), rs4131463, RPL18P4 (rs1528356), rs17797701, OTOG (rs11024327), rs10843660, CHST11 (rs11112069), SIP1 (rs8011855), RHOJ (rs1957779) which are worthy of digging for statistical replication and biological experiments in the future.

Keywords: Loss of power, expectation-conditional maximization, genome-wide association study, single nucleotide polymorphism, additive model, hypertension.

誌　　謝

研究所兩年的時間匆匆呼嘯而過，夾雜著書香味、客運味以及最重要的人情味。感謝爸爸、媽媽、姊姊所給予我的一切，你們永遠是我精神上的最大支柱，每當卸下一身疲憊回到花蓮，總是感覺花蓮好可愛，家裡好溫暖；感謝雅云從大學一路陪伴我到現在研究所畢業，這兩年因為分隔兩地，加上我時常忙於課業，真是辛苦妳了，幾乎每個禮拜的高雄行，即使舟車但卻不勞頓，妳總是能鼓勵我，讓我調整好步伐繼續出發，也要感謝雅云的家人，把我當成一家人，包容我不時的叨擾。

感謝盧鴻興老師與楊照崑老師毫不吝嗇所給予的指導與機會，從你們身上我所學習到的不只是學問、態度，更值得歡喜的則是看事物的不同角度，越往高處爬就越覺得自己的渺小，虛心學習任何新的事物。

感謝研究室的各位同學，讓我這半吊子常常麻煩你們，甚麼時候才能再找時間一起玩樂呢？香菸、下巴、阿北、阿木、人夫、丁丁、澄竹、小豬、卿卿、慧潔、飛飛等，好多好多人要感謝，謝謝你們陪我度過這難忘的兩年時光。感謝火哥、清大的志偉學長、雪芳學姊、立欣學姊，兩年來一起打球的歡樂時光。感謝可愛的學弟妹們辦的送舊，讓我們既興奮又感傷。

最後要感謝交通大學統計所帶給我的一切，老師們的無私付出、郭姐的關懷與包容、小陳的球棒支援，有你們在所上真好。

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The trade-off for α (type I error) and β (type II error) is well-known in statistical hypothesis testing. However, a small α such as $5 \times 10^{-7}, 1 \times 10^{-5}$ are used often in case-control association study because of multiple testing. Thus, the power $(1-\beta)$ will be badly weakened. In other words, a small α usually makes hypothesis testing over-conservative. Multiple comparisons are the primary concern in many previous studies. Our approach is focused on the loss of power and the reduction in computation time.

First of all, our approach attempts to suggest a reasonable threshold (such as $\xi_1 = 2.7(\alpha = 0.1)$ in single gene tests) for reducing the effort in finding interaction association based on low loss of power. Second, our results provide a quantitative assessment between the loss of power and the gain of computation time (reduce 99.59% in this study). In addition, expectation-conditional maximization (ECM) is used to estimate λ_{AB} (relative penetrance rate for genotype AB), f_A (frequency A), f_B (frequency B) in order to provide parameters for further calculating power loss.

Replication of the Wellcome Trust genome-wide association study of hypertension by this approach, we detected some SNPs or genes are significantly associated with hypertension risk. Some of them are known, such as CHRM2 (rs7800093), KCNB2 (rs11782342), HTR3B (rs17116117), rs2820037, GAB1 (rs300916, rs300915, rs300913), BCAT1 (rs7961152, rs11613673, rs12424348), MYBPC1 (rs11110912), LOC100132798 (rs2398162), MAGI1 (rs2091244, rs2177686, rs17073046). However, those other unknowns, such as rs825148, rs1553460, LOC100129858 (rs6840033), rs4131463, RPL18P4 (rs1528356), rs17797701, OTOG (rs11024327), rs10843660, CHST11 (rs11112069), SIP1 (rs8011855), RHOJ (rs1957779) are worthy of digging for statistical replication and biological explanation in the future. We know that statistical significance is not equivalent to biological significance. Hence, We hope that the results in this study can provide information in multiple SNPs association.

Chapter 2

Literature Review

2.1 Association study

Association study between genetic marker and phenotype has been used widely to identify regions of the genome and genes that affect phenotype in genetics. Restriction fragment length polymorphism (RFLP), minisatellite, microsatellite, and single nucleotide polymorphism (SNP) can be biomarkers. Phenotypes can be hair color, drug response, disease status, etc. We may know the association between biomarkers and disease through case-control association study. If the association is significant, either there is a linkage between the biomarkers and real gene which controls the phenotype or the biomarkers is exactly situated on real gene.

The detection of genetic factors is often used in complex disease study, such as hypertension, schizophrenia, cancer, and diabetes, which are affected by multiple genetic and environmental factors. In many situation, genomic association study has more power than linkage analysis to identify the putative genes since numerous multiple effects are too complex for linkage study [Risch and Merikangas, 1996].

2.2 Single nucleotide polymorphism (SNP)

A single nucleotide polymorphism (SNP) is a kind of widespread DNA sequence variation that occur when a single nucleotide (A, T, C, or G) in the genome sequence is changed, namely, there are two or more alleles on specific locus. In the past, we called "mutation" when the minor allele frequency is less than or equal to 1%, otherwise regarded it as "SNP", but the definition is no longer necessary (SNPs with minor allele frequency are less than or equal to 1% included in dbSNP).

SNP is often regarded as genetic marker in studies, owing to the high frequency of about 0.1% in humans, however, not all of SNPs have real clinical meaning. The following are four types of SNP:

- non-coding SNP:

The locus of SNP is on untranslated region, such as promoter.

- coding SNP (cSNP):
  The antonym of non-coding SNP, it may alter the structure or function of protein.

- synonymous SNP:
  The SNP belongs to cSNP, but does not alter the translated protein product.

- non-synonymous SNP:
  The antonym of synonymous SNP, it will result different amino acids which may alter the function.

Researchers can find out disease susceptibility locus of SNP, and design personalized medicine by SNP related to drug metabolism. Previous studies had interesting discoveries, for instance, APOE with Alzheimer's disease, TCF7L2 with type 2 diabetes, and HTR2A with schizophrenia.

2.3 Multiple comparisons

The densely spaced biomarkers are the source of multiple comparisons in genome-wide association study (GWAS). In GWAS, testing a great amount of hypothesis simultaneously is a prerequisite. As the first paragraph mentioned in introduction, the trade-off for α and β will be a topic in this case. Numerous researchers and approaches, such as Bonferroni procedure [Bonferroni, 1936], Sidak procedure, Holm procedure [Holm, 1979], Hochberg procedure [Hochberg, 1988], and Benjamini & Hochberg procedure [Benjamini and Hochberg, 1995], contribute on this issue before bio-technology has been rapidly elevated recent years. The traditional Bonferroni procedure is frequently used, but it is well-known that this procedure is over-conservative. To increase the power by Bonferroni procedure, we consider the generalized family-wise error rate (gFWER) and the false discovery rate (FDR).

2.4 Data quality control

By quality control, reliability for further study can be promoted such that the result is more meaningful. Genetic markers and samples are two targets to be filtered out in GWAS. The Genotyping Facility at the Wellcome Trust Sanger Institute (WTSI) high-throughput genotyping quality control includes SNP call rate, minor allele frequency (MAF), and Hardy-Weinberg equilibrium (HWE) for each genetic marker, sample call rate, heterozygosity, and cryptic relatedness for each sample.

2.4.1 SNP call rate

Low SNP call rate occurs when there are too many missing data (probe intensity value doesn't pass the detection filter score) on automated SNP calling algorithm. Its definition is the proportion of non-missing data over whole sample. Exclusion criteria is often SNP call rate $\leq 95\%$.

2.4.2 Minor allele frequency (MAF)

The allele frequency is the proportion of the allele over whole sample. SNPs are usually biallelic. The minor allele is the less frequence allele at a locus that is observed in a specific population. SNPs would usually be excluded if MAF $\leq 1\%$.

2.4.3 Hardy-Weinberg equilibrium (HWE)

The Hardy-Weinberg equilibrium indicates that allele frequencies in a population remain constant from generation to generation unless specific external force, such as non-random mating (includes inbreeding, assortative mating, genetic drift), selection, and mutation. Thus, deviation from HWE would be checked, SNPs will often be excluded with p-value $\leq 10^{-5}$ in HWE testing.

2.4.4 Sample call rate

Low sample call rate occurs when there are too many missing data (probe intensity value does not pass the detection filter score) on automated SNP calling algorithm. Its definition is the proportion of non-missing data per sample. The exclusion criteria is generally sample call rate $\leq 97\%$.

2.4.5 Heterozygosity

The genotypes AA, aa are homozygous and the genotype Aa is heterozygous for a biallelic SNP, which has allele A, and a. By definition, heterozygosity per individual is the proportion of SNPs that are heterozygous within whole typed SNPs. If heterozygosity $\leq 22.5\%$ or $\geq 30\%$, the individual would be filtered out owing to low heterozygosity can result in more heterozygote genotypes being no called and excess heterozygosity may indicate contamination by foreign DNA.

2.4.6 Cryptic relatedness

In many statistical techniques, we usually assume independent property, the approach we use is no exception. However, real relationship for consanguinity is sometimes unascertainable. The identity-by-state (IBS, sum of the number of identical-by-state alleles at

each locus divided by twice the number of loci) is possible to assess the unknown relationships within sample population and to avoid non-trivial degrees of relatedness, which may violate the assumption. Average IBS between each pair of individuals can be a measurement to determine the individual is excluded or not. The individual could be suspect with IBS $\geq 86\%$ or IBS $\geq 99\%$.

Chapter 3

Methodology

3.1 Loss of Power

The detection of interaction for complex human diseases is usually important, but the tremendous computation time is primary problem in the genetic study. Minimizing the loss of power in hypothesis may be a proper direction by setting a reasonable threshold on single SNP testing to avoid further testing of interaction of this gene.

3.1.1 Algorithm

Table 3.1: Single SNP Allele

	Allele		
Group	A	a	Total
Disease	N_{11}	N_{12}	n_D
Control	N_{21}	N_{22}	n_C
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	$n_{\cdot}$

Table 3.2: Single SNP Genotype

	Genotype			
Group	A/A	A/a	a/a	Total
Disease	N_{1AA}	N_{1Aa}	N_{1aa}	N_1
Control	N_{2AA}	N_{2Aa}	N_{2aa}	N_2
Total	$N_{\cdot AA}$	$N_{\cdot Aa}$	$N_{\cdot aa}$	$N_{\cdot}$

In an additive model, table 3.1 is condensed from table 3.2, and

$$\begin{aligned} n_D &= 2N_1,\ N_{11} = 2N_{1AA} + N_{1Aa},\ N_{12} = N_{1Aa} + 2N_{1aa} \\ n_C &= 2N_2,\ N_{21} = 2N_{2AA} + N_{2Aa},\ N_{22} = N_{2Aa} + 2N_{2aa}. \end{aligned}$$

Table 3.3: Interaction SNP Allele

	Allele				
Group	AB	Ab	aB	ab	Total
Disease	n_{11}	n_{12}	n_{13}	n_{14}	n_D
Control	n_{21}	n_{22}	n_{23}	n_{24}	n_C
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$n_{\cdot 4}$	$n_{\cdot}$

Also table 3.3 is condensed from table 3.4, and

$$\begin{aligned}
n_{11} &= 2n_{1ABAB} + n_{1ABAb} + n_{1ABaB} + n_{1ABab} \\
n_{12} &= n_{1ABAb} + 2n_{1AbAb} + n_{1AbaB} + n_{1Abab} \\
n_{13} &= n_{1ABaB} + n_{1AbaB} + 2n_{1aBaB} + n_{1aBab} \\
n_{14} &= n_{1ABab} + n_{1Abab} + n_{1aBab} + 2n_{1abab} \\
n_{21} &= 2n_{2ABAB} + n_{2ABAb} + n_{2ABaB} + n_{2ABab} \\
n_{22} &= n_{2ABAb} + 2n_{2AbAb} + n_{2AbaB} + n_{2Abab} \\
n_{23} &= n_{2ABaB} + n_{2AbaB} + 2n_{2aBaB} + n_{2aBab} \\
n_{24} &= n_{2ABab} + n_{2Abab} + n_{2aBab} + 2n_{2abab}.
\end{aligned}$$

Table 3.4: Interaction SNP Genotype

	Genotype										
Group	AB/AB	AB/Ab	AB/aB	AB/ab	Ab/Ab	Ab/aB	Ab/ab	aB/aB	aB/ab	ab/ab	Total
Disease	n_{1ABAB}	n_{1ABAb}	n_{1ABaB}	n_{1ABab}	n_{1AbAb}	n_{1AbaB}	n_{1Abab}	n_{1aBaB}	n_{1aBab}	n_{1abab}	N_1
Control	n_{2ABAB}	n_{2ABAb}	n_{2ABaB}	n_{2ABab}	n_{2AbAb}	n_{2AbaB}	n_{2Abab}	n_{2aBaB}	n_{2aBab}	n_{2abab}	N_2
Total	$n_{\cdot ABAB}$	$n_{\cdot ABAb}$	$n_{\cdot ABaB}$	$n_{\cdot ABab}$	$n_{\cdot AbAb}$	$n_{\cdot AbaB}$	$n_{\cdot Abab}$	$n_{\cdot aBaB}$	$n_{\cdot aBab}$	$n_{\cdot abab}$	$N_{\cdot}$

Let the allele frequency for A and B be f_A and f_B respectively. In addition, it is assumed that

$$\begin{aligned}
&P(D|g = AB/AB) : P(D|g = AB/*) : P(D|g = */*) \\
&= \lambda_{AB}^2 : \lambda_{AB} : 1,
\end{aligned}$$

where g means genotype, D means disease, $*$ means not AB, and λ_{AB} represents the relative penetrance rate.

Hence, in the disease population,

$$
\begin{aligned}
P(g=AB/AB|D) &= \frac{P(D|g=AB/AB)P(g=AB/AB)}{P(D)} \\
&= \frac{\frac{P(D|g=AB/AB)}{P(D|g=*/*)}P(g=AB/AB)}{\frac{P(D)}{P(D|g=*/*)}} = \frac{\lambda_{AB}^2 \times f_A^2 f_B^2}{\frac{P(D)}{P(D|g=*/*)}} \\
P(g=AB/Ab|D) &= \frac{P(D|g=AB/Ab)P(g=AB/Ab)}{P(D)} \\
&= \frac{\frac{P(D|g=AB/Ab)}{P(D|g=*/*)}P(g=AB/Ab)}{\frac{P(D)}{P(D|g=*/*)}} = \frac{\lambda_{AB} \times 2f_A^2 f_B f_b}{\frac{P(D)}{P(D|g=*/*)}} \\
P(g=AB/aB|D) &= \frac{P(D|g=AB/aB)P(g=AB/aB)}{P(D)} \\
&= \frac{\frac{P(D|g=AB/aB)}{P(D|g=*/*)}P(g=AB/aB)}{\frac{P(D)}{P(D|g=*/*)}} = \frac{\lambda_{AB} \times 2f_A f_a f_B^2}{\frac{P(D)}{P(D|g=*/*)}} \\
P(g=AB/ab|D) &= \frac{P(D|g=AB/ab)P(g=AB/ab)}{P(D)} \\
&= \frac{\frac{P(D|g=AB/ab)}{P(D|g=*/*)}P(g=AB/ab)}{\frac{P(D)}{P(D|g=*/*)}} = \frac{\lambda_{AB} \times 2f_A f_a f_B f_b}{\frac{P(D)}{P(D|g=*/*)}} \\
P(g=Ab/Ab|D) &= \frac{P(D|g=Ab/Ab)P(g=Ab/Ab)}{P(D)} \\
&= \frac{\frac{P(D|g=Ab/Ab)}{P(D|g=*/*)}P(g=Ab/Ab)}{\frac{P(D)}{P(D|g=*/*)}} = \frac{f_A^2 f_b^2}{\frac{P(D)}{P(D|g=*/*)}} \\
P(g=Ab/aB|D) &= \frac{P(D|g=Ab/aB)P(g=Ab/aB)}{P(D)} \\
&= \frac{\frac{P(D|g=Ab/aB)}{P(D|g=*/*)}P(g=Ab/aB)}{\frac{P(D)}{P(D|g=*/*)}} = \frac{2f_A f_a f_B f_b}{\frac{P(D)}{P(D|g=*/*)}} \\
P(g=Ab/ab|D) &= \frac{P(D|g=Ab/ab)P(g=Ab/ab)}{P(D)} \\
&= \frac{\frac{P(D|g=Ab/ab)}{P(D|g=*/*)}P(g=Ab/ab)}{\frac{P(D)}{P(D|g=*/*)}} = \frac{2f_A f_a f_b^2}{\frac{P(D)}{P(D|g=*/*)}}
\end{aligned}
$$

$$
\begin{aligned}
P(g = aB/aB|D) &= \frac{P(D|g = aB/aB)P(g = aB/aB)}{P(D)} \\
&= \frac{\frac{P(D|g = aB/aB)}{P(D|g = */*)}P(g = aB/aB)}{\frac{P(D)}{P(D|g = */*)}} = \frac{f_a^2 f_B^2}{\frac{P(D)}{P(D|g = */*)}} \\
P(g = aB/ab|D) &= \frac{P(D|g = aB/ab)P(g = aB/ab)}{P(D)} \\
&= \frac{\frac{P(D|g = aB/ab)}{P(D|g = */*)}P(g = aB/ab)}{\frac{P(D)}{P(D|g = */*)}} = \frac{2f_a^2 f_B f_b}{\frac{P(D)}{P(D|g = */*)}} \\
P(g = ab/ab|D) &= \frac{P(D|g = ab/ab)P(g = ab/ab)}{P(D)} \\
&= \frac{\frac{P(D|g = ab/ab)}{P(D|g = */*)}P(g = ab/ab)}{\frac{P(D)}{P(D|g = */*)}} = \frac{f_a^2 f_b^2}{\frac{P(D)}{P(D|g = */*)}}, \\
\frac{P(D)}{P(D|g = */*)} &= \text{sum of the 10 numerators above.}
\end{aligned}
$$

The probability of AB, Ab, aB, ab in table 3.3 disease row are,

$$
\begin{aligned}
p_{AB|D} &= P(g = AB/AB|D) + 0.5\left[P(g = AB/Ab|D) + P(g = AB/aB|D) + P(g = AB/ab|D)\right] \\
p_{Ab|D} &= P(g = Ab/Ab|D) + 0.5\left[P(g = AB/Ab|D) + P(g = Ab/aB|D) + P(g = Ab/ab|D)\right] \\
p_{aB|D} &= P(g = aB/aB|D) + 0.5\left[P(g = AB/aB|D) + P(g = Ab/aB|D) + P(g = aB/ab|D)\right] \\
p_{ab|D} &= P(g = ab/ab|D) + 0.5\left[P(g = AB/ab|D) + P(g = Ab/ab|D) + P(g = aB/ab|D)\right].
\end{aligned}
$$

Similarly, $p_{AB|C}, p_{Ab|C}, p_{aB|C}, p_{ab|C}$ in table 3.3 control row are computed by the same formulas with $\lambda_{AB} = 1$, and we know

$$
\begin{aligned}
(n_{11}, n_{12}, n_{13}, n_{14}) &\sim \text{Multinomial}\left(n_D; p_{AB|D}, p_{Ab|D}, p_{aB|D}, p_{ab|D}\right) \\
(n_{21}, n_{22}, n_{23}, n_{24}) &\sim \text{Multinomial}\left(n_C; p_{AB|C}, p_{Ab|C}, p_{aB|C}, p_{ab|C}\right),
\end{aligned}
$$

proposed approach used to simulate the contingency table 3.3 at the moment that given $\lambda_{AB}, f_A, f_B, n_D = 4000, n_C = 6000$, construct hypothesis testing,

$$
\begin{aligned}
H_0 &: \lambda_{AB} = \lambda_{Ab} = \lambda_{aB} = \lambda_{ab} = 1 \\
H_1 &: \lambda_{AB} > \lambda_{Ab} = \lambda_{aB} = \lambda_{ab} = 1,
\end{aligned}
$$

and find out the loss of power what we concern,

$$
\frac{P(Q_2 > \xi_2|H_1) - P(Q_2 > \xi_2 \text{ and } Q_1 > \xi_1|H_1)}{P(Q_2 > \xi_2|H_1)}.
$$

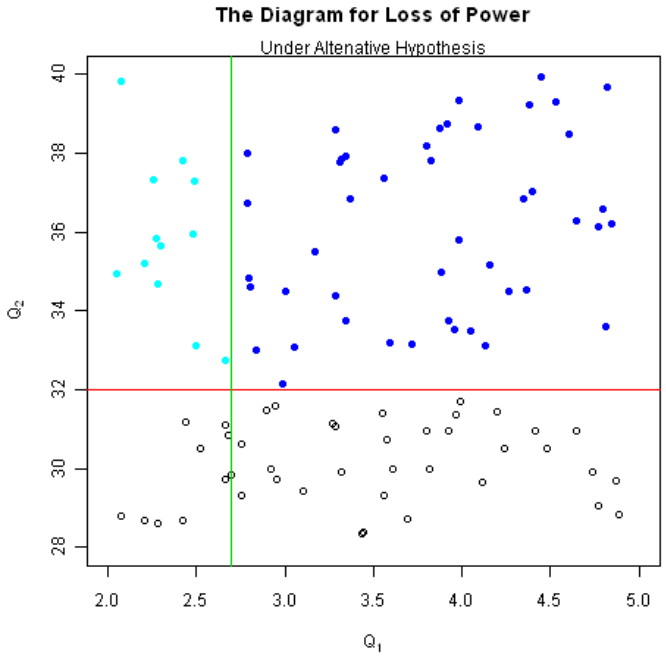

Figure 3.1: The Hypothetical Diagram for Loss of Power

where Q_1 is the test statistic (Cochran-Armitage trend test [Armitage, 1955], Fisher's exact test [Fisher, 1922], or Chi-square test [Pearson, 1900]) from table 3.1 or table 3.2, Q_2 is the test statistic (Chi-square test or Fisher's exact test) from table 3.3, and ξ_1, ξ_2 are the thresholds respectively. From figure 3.1, the loss of power defines as $\frac{\#\text{ of sky blue dots}}{\#\text{ of sky blue and aquanarine dots}}$.

$$
Q_1 = \begin{cases} \dfrac{N_{\cdot}\left[N_{\cdot}(N_{1Aa}+2N_{1aa})-N_1N_{\cdot Aa}+2N_{\cdot aa})\right]^2}{N_2N_1\left[N_{\cdot}(N_{\cdot Aa}+4N_{\cdot aa})-(N_{\cdot Aa}+2N_{\cdot aa})^2\right]} & \text{, Cochran-Armitage trend test} \\ \text{Sum of all P-values which are} \\ \qquad \leq P_{\text{cutoff}} = \dfrac{(n_D!n_C!)(N_{\cdot 1}!N_{\cdot 2}!)}{n_{\cdot}!(N_{11}!N_{12}!N_{21}!N_{22}!)} & \text{, Fisher's exact test} \\ \displaystyle\sum_{i=1}^{2}\left[\frac{(N_{1i}-\frac{n_DN_{\cdot i}}{n_{\cdot}})^2}{\frac{n_DN_{\cdot i}}{n_{\cdot}}}+\frac{(N_{2i}-\frac{n_CN_{\cdot i}}{n_{\cdot}})^2}{\frac{n_CN_{\cdot i}}{n_{\cdot}}}\right] & \text{, Chi-square test} \end{cases}
$$

$$
\sim \mathcal{X}^2(1)
$$

$$
Q_2 = \begin{cases} \displaystyle\sum_{i=1}^{4}\left[\frac{(n_{1i}-\frac{n_Dn_{\cdot i}}{n_{\cdot}})^2}{\frac{n_Dn_{\cdot i}}{n_{\cdot}}}+\frac{(n_{2i}-\frac{n_Cn_{\cdot i}}{n_{\cdot}})^2}{\frac{n_Cn_{\cdot i}}{n_{\cdot}}}\right] & \text{, Chi-square test} \\ \text{Sum of all P-values which are} \\ \qquad \leq P_{\text{cutoff}} = \dfrac{(n_D!n_C!)(n_{\cdot 1}!n_{\cdot 2}!n_{\cdot 3}!n_{\cdot 4}!)}{n_{\cdot}!(n_{11}!n_{12}!n_{13}!n_{14}!n_{21}!n_{22}!n_{23}!n_{24}!)} & \text{, Fisher's exact test} \end{cases}
$$

$$
\sim \mathcal{X}^2(3)
$$

3.1.2 Simulation

Table 3.5 and table 3.6 below show the simulation results when thresholds are

$$
\xi_1 = 2.7\ (\alpha = 0.1) \text{ or } 3.17\ (\alpha = 0.075),\ \text{and } \xi_2 = 32\ (\alpha = 5 \times 10^{-7}),
$$

we could set a threshold for single association (ξ_1) depends on these reference tables, such that both of reduced computation time and loss of power are tolerable for us.

Table 3.5: Loss of Power by Simulation when $\xi_1 = 2.7(\alpha = 0.1)$, $\xi_2 = 32(\alpha = 5 \times 10^{-7})$

λ_{AB}	f_A	f_B	Original Power [a]	Absolute LOP	Relative LOP (%)	λ_{AB}	f_A	f_B	Original Power	Absolute LOP	Relative LOP (%)
		0.1	0.00000	0.00000	0.0000			0.1	0.01447	0.00698	48.2377
	0.1	0.2	0.00003	0.00001	33.3333		0.1	0.2	0.57907	0.00518	0.8945
		0.3	0.00131	0.00006	4.5802			0.3	0.97942	0.00001	0.0010
		0.1	0.00000	0.00000	0.0000			0.1	0.58140	0.14230	24.4754
1.50	0.2	0.2	0.01840	0.00232	12.6087	2.00	0.2	0.2	0.99977	0.00004	0.0040
		0.3	0.26578	0.00100	0.3763			0.3	1.00000	0.00000	0.0000
		0.1	0.00104	0.00065	62.5000			0.1	0.97795	0.13650	13.9578
	0.3	0.2	0.26421	0.02107	7.9747		0.3	0.2	1.00000	0.00000	0.0000
		0.3	0.85797	0.00070	0.0816			0.3	1.00000	0.00000	0.0000
		0.1	0.00007	0.00003	42.8571			0.1	0.98602	0.01444	1.4645
	0.1	0.2	0.04221	0.00329	7.7944		0.1	0.2	1.00000	0.00000	0.0000
		0.3	0.39722	0.00052	0.1309			0.3	1.00000	0.00000	0.0000
		0.1	0.03980	0.02035	51.1307			0.1	1.00000	0.00000	0.0110
1.75	0.2	0.2	0.83047	0.00683	0.8224	3.00	0.2	0.2	1.00000	0.00000	0.0000
		0.3	0.99834	0.00000	0.0000			0.3	1.00000	0.00000	0.0000
		0.1	0.39718	0.16843	42.4065			0.1	1.00000	0.00000	0.0010
	0.3	0.2	0.99842	0.00101	0.1012		0.3	0.2	1.00000	0.00000	0.0000
		0.3	1.00000	0.00000	0.0000			0.3	1.00000	0.00000	0.0000
		0.1	0.00000	0.00000	0.0000			0.1	0.02068	0.01481	71.6151
	0.1	0.2	0.00000	0.00000	0.0000		0.1	0.2	0.85486	0.03506	4.1013
		0.3	0.00000	0.00000	0.0000			0.3	0.99974	0.00002	0.0020
		0.1	0.00000	0.00000	0.0000			0.1	0.85408	0.37084	43.4198
0.75	0.2	0.2	0.00000	0.00000	0.0000	0.25	0.2	0.2	1.00000	0.00028	0.0280
		0.3	0.00000	0.00000	0.0000			0.3	1.00000	0.00000	0.0000
		0.1	0.00000	0.00000	0.0000			0.1	0.99965	0.26981	26.9904
	0.3	0.2	0.00000	0.00000	0.0000		0.3	0.2	1.00000	0.00001	0.0010
		0.3	0.00039	0.00009	23.0769			0.3	1.00000	0.00000	0.0000
		0.1	0.00000	0.00000	0.0000			0.1	0.76622	0.38398	50.1135
	0.1	0.2	0.00216	0.00076	35.1852		0.1	0.2	1.00000	0.00146	0.1460
		0.3	0.07424	0.00226	3.0442			0.3	1.00000	0.00000	0.0000
		0.1	0.00228	0.00165	72.3684			0.1	1.00000	0.15986	15.9860
0.50	0.2	0.2	0.39539	0.03595	9.0923	0.05	0.2	0.2	1.00000	0.00000	0.0000
		0.3	0.96237	0.00037	0.0384			0.3	1.00000	0.00000	0.0000
		0.1	0.07241	0.04819	66.5516			0.1	1.00000	0.05470	5.4700
	0.3	0.2	0.96015	0.02989	3.1131		0.3	0.2	1.00000	0.00000	0.0000
		0.3	1.00000	0.00000	0.0000			0.3	1.00000	0.00000	0.0000

[a] Calculates by 100000 simulations.

Table 3.6: Loss of Power by Simulation when $\xi_1 = 3.17(\alpha = 0.075)$, $\xi_2 = 32(\alpha = 5 \times 10^{-7})$

λ_{AB}	f_A	f_B	Original Power [a]	Absolute LOP	Relative LOP (%)	λ_{AB}	f_A	f_B	Original Power	Absolute LOP	Relative LOP (%)
		0.1	0.00000	0.00000	0.0000			0.1	0.01447	0.00815	56.3463
	0.1	0.2	0.00003	0.00001	33.3333		0.1	0.2	0.57907	0.00970	1.6747
		0.3	0.00131	0.00016	12.1951			0.3	0.97942	0.00002	0.0020
		0.1	0.00000	0.00000	0.0000			0.1	0.58140	0.19737	33.9466
1.50	0.2	0.2	0.01840	0.00378	20.5184	2.00	0.2	0.2	0.99977	0.00013	0.0130
		0.3	0.26578	0.00209	0.7868			0.3	1.00000	0.00000	0.0000
		0.1	0.00104	0.00082	78.9474			0.1	0.97795	0.20778	21.2464
	0.3	0.2	0.26421	0.03404	12.8852		0.3	0.2	1.00000	0.00000	0.0000
		0.3	0.85797	0.00184	0.2149			0.3	1.00000	0.00000	0.0000
		0.1	0.00007	0.00004	61.5385			0.1	0.98602	0.02639	2.6762
	0.1	0.2	0.04221	0.00539	12.7639		0.1	0.2	1.00000	0.00000	0.0000
		0.3	0.39722	0.00094	0.2355			0.3	1.00000	0.00000	0.0000
		0.1	0.03980	0.02385	59.9231			0.1	1.00000	0.00016	0.0160
1.75	0.2	0.2	0.83047	0.01405	1.6915	3.00	0.2	0.2	1.00000	0.00000	0.0000
		0.3	0.99834	0.00000	0.0000			0.3	1.00000	0.00000	0.0000
		0.1	0.39718	0.21232	53.4560			0.1	1.00000	0.00010	0.0010
	0.3	0.2	0.99842	0.00295	0.2955		0.3	0.2	1.00000	0.00000	0.0000
		0.3	1.00000	0.00000	0.0000			0.3	1.00000	0.00000	0.0000
		0.1	0.00000	0.00000	0.0000			0.1	0.02068	0.01813	87.6725
	0.1	0.2	0.00000	0.00000	0.0000		0.1	0.2	0.85486	0.12434	14.5446
		0.3	0.00000	0.00000	0.0000			0.3	0.99974	0.00040	0.0400
		0.1	0.00000	0.00000	0.0000			0.1	0.85408	0.60546	70.8907
0.75	0.2	0.2	0.00000	0.00000	0.0000	0.25	0.2	0.2	1.00000	0.00436	0.4360
		0.3	0.00000	0.00000	0.0000			0.3	1.00000	0.00000	0.0000
		0.1	0.00000	0.00000	0.0000			0.1	0.99965	0.53804	53.8228
	0.3	0.2	0.00000	0.00000	0.0000		0.3	0.2	1.00000	0.00016	0.0160
		0.3	0.00039	0.00010	27.6490			0.3	1.00000	0.00000	0.0000
		0.1	0.00000	0.00000	0.0000			0.1	0.76622	0.57763	75.3870
	0.1	0.2	0.00216	0.00124	57.3333		0.1	0.2	1.00000	0.01131	1.1310
		0.3	0.07424	0.00827	11.1422			0.3	1.00000	0.00000	0.0000
		0.1	0.00228	0.00205	89.7561			0.1	1.00000	0.38617	38.6170
0.50	0.2	0.2	0.39539	0.10093	25.5268	0.05	0.2	0.2	1.00000	0.00000	0.0000
		0.3	0.96237	0.00394	0.4091			0.3	1.00000	0.00000	0.0000
		0.1	0.07241	0.06199	85.6122			0.1	1.00000	0.18605	18.6050
	0.3	0.2	0.96015	0.11882	12.3755		0.3	0.2	1.00000	0.00000	0.0000
		0.3	1.00000	0.00012	0.0120			0.3	1.00000	0.00000	0.0000

[a] Calculates by 100000 simulations.

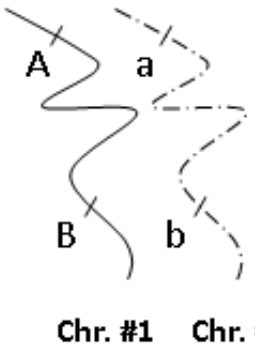

Figure 3.2: Genotype: AB/ab

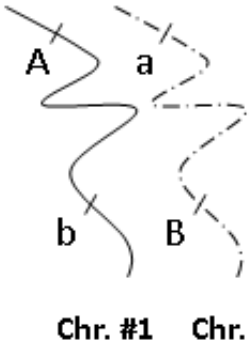

Figure 3.3: Genotype: Ab/aB

3.2 Expectation-Conditional Maximization (ECM)

For the approach above, we consider the additive model, nevertheless, there is an ambiguity for (AB/ab) and (aB/Ab) in real data analysis. The following ECM algorithm [Meng and Rubin, 1993] not only assigns the frequencies for (AB/ab) and (aB/Ab) but also estimates $\lambda_{AB}, f_A,$ andf_B.

Table 3.7: Observed Incomplete Data

	Genotype									
Group	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb	Total
Disease	y_{1D}	y_{2D}	y_{3D}	y_{4D}	y_{5D}	y_{6D}	y_{7D}	y_{8D}	y_{9D}	n_D
Control	y_{1C}	y_{2C}	y_{3C}	y_{4C}	y_{5C}	y_{6C}	y_{7C}	y_{8C}	y_{9C}	n_C

Table 3.8: Unobserved complete Data

	Genotype										
Group	AB/AB	AB/Ab	AB/aB	AB/ab	Ab/Ab	Ab/aB	Ab/ab	aB/aB	aB/ab	ab/ab	Total
Disease	x_{1D}	x_{2D}	x_{3D}	x_{4D}	x_{5D}	x_{6D}	x_{7D}	x_{8D}	x_{9D}	x_{10D}	n_D
Control	x_{1C}	x_{2C}	x_{3C}	x_{4C}	x_{5C}	x_{6C}	x_{7C}	x_{8C}	x_{9C}	x_{10C}	n_C

Firstly, we have three parameters in ECM,

$$\boldsymbol{\theta} = (\lambda_{AB}, f_A, f_B)\,,$$

incomplete data $\boldsymbol{Y} = (\boldsymbol{Y}_D, \boldsymbol{Y}_C) = (Y_{1D}, Y_{2D}, \ldots, Y_{9D}, Y_{1C}, Y_{2C}, \ldots, Y_{9C})\,,$

$$\boldsymbol{Y} \sim \text{Multinomial}\,(n_D; \boldsymbol{P}_{YD}) \times \text{Multinomial}\,(n_C; \boldsymbol{P}_{YC})\,, \tag{3.1}$$

where

$$\begin{aligned} n_D &= Y_{1D} + Y_{2D} + \cdots + Y_{9D},\ n_C = Y_{1C} + Y_{2C} + \cdots + Y_{9C}, \\ \boldsymbol{P}_{YD} &= (p_{1D}, p_{2D}, \ldots, p_{9D}),\ \boldsymbol{P}_{YC} = (p_{1C}, p_{2C}, \ldots, p_{9C}), \end{aligned}$$

and complete data $\boldsymbol{X} = (\boldsymbol{X}_D, \boldsymbol{X}_C) = (X_{1D}, X_{2D}, \ldots, X_{10D}, X_{1C}, X_{2C}, \ldots, X_{10C})$,

$$\boldsymbol{X} \sim \text{Multinomial}\,(n_D; \boldsymbol{P}_{XD}) \times \text{Multinomial}\,(n_C; \boldsymbol{P}_{XC}), \tag{3.2}$$

where

$$\begin{aligned} n_D &= X_{1D} + X_{2D} + \cdots + X_{10D},\ n_C = X_{1C} + X_{2C} + \cdots + X_{10C}, \\ \boldsymbol{P}_{XD} &= (p_{1D}, p_{2D}, \ldots, p_{10D}),\ \boldsymbol{P}_{XC} = (p_{1C}, p_{2C}, \ldots, p_{10C}), \end{aligned}$$

Therefore, by equation (3.1), the likelihood function on incomplete space is,

$$\begin{aligned} L^{in}(\boldsymbol{\theta}|\boldsymbol{y}) &= g(\boldsymbol{y}|\boldsymbol{\theta}) \\ &= \frac{n_D!}{y_{1D}! \times y_{2D}! \times \cdots \times y_{9D}!} \left[\frac{\lambda_{AB}^2 f_A^2 f_B^2}{P^*(D)}\right]^{y_{1D}} \left[\frac{\lambda_{AB} \times 2 f_A^2 f_B (1-f_B)}{P^*(D)}\right]^{y_{2D}} \\ &\quad \left[\frac{f_A^2 (1-f_B)^2}{P^*(D)}\right]^{y_{3D}} \left[\frac{\lambda_{AB} \times 2 f_A (1-f_A) f_B^2)}{P^*(D)}\right]^{y_{4D}} \\ &\quad \left[\frac{\lambda_{AB} \times 2 f_A (1-f_A) f_B (1-f_B) + 2 f_A (1-f_A) f_B (1-f_B)}{P^*(D)}\right]^{y_{5D}} \\ &\quad \left[\frac{2 f_A (1-f_A)(1-f_B)^2}{P^*(D)}\right]^{y_{6D}} \left[\frac{(1-f_A)^2 f_B^2}{P^*(D)}\right]^{y_{7D}} \\ &\quad \left[\frac{2(1-f_A)^2 f_B (1-f_B)}{P^*(D)}\right]^{y_{8D}} \left[\frac{(1-f_A)^2 (1-f_B)^2}{P^*(D)}\right]^{y_{9D}} \\ &\quad \frac{n_C!}{y_{1C}! \times y_{2C}! \times \cdots \times y_{9C}!} \left[\frac{\lambda_{AB}^2 f_A^2 f_B^2}{P^*(C)}\right]^{y_{1C}} \left[\frac{\lambda_{AB} \times 2 f_A^2 f_B (1-f_B)}{P^*(C)}\right]^{y_{2C}} \\ &\quad \left[\frac{f_A^2 (1-f_B)^2}{P^*(C)}\right]^{y_{3C}} \left[\frac{\lambda_{AB} \times 2 f_A (1-f_A) f_B^2)}{P^*(C)}\right]^{y_{4C}} \\ &\quad \left[\frac{\lambda_{AB} \times 2 f_A (1-f_A) f_B (1-f_B) + 2 f_A (1-f_A) f_B (1-f_B)}{P^*(C)}\right]^{y_{5C}} \\ &\quad \left[\frac{2 f_A (1-f_A)(1-f_B)^2}{P^*(C)}\right]^{y_{6C}} \left[\frac{(1-f_A)^2 f_B^2}{P^*(C)}\right]^{y_{7C}} \\ &\quad \left[\frac{2(1-f_A)^2 f_B (1-f_B)}{P^*(C)}\right]^{y_{8C}} \left[\frac{(1-f_A)^2 (1-f_B)^2}{P^*(C)}\right]^{y_{9C}}. \end{aligned} \tag{3.3}$$

(3.4)

where

$$\begin{aligned} P^*(D) &= \frac{P(D)}{P(D|g = */*)} \\ P^*(C) &= \frac{P(C)}{P(C|g = */*)} \end{aligned}$$

By equation (3.2), the likelihood function on complete space is,

$$
\begin{aligned}
L^c(\boldsymbol{\theta}|\boldsymbol{x}) &= f(\boldsymbol{x}|\boldsymbol{\theta}) \\
&= \frac{n_D!}{x_{1D}! \times x_{2D}! \times \cdots \times x_{10D}!} \left[\frac{\lambda_{AB}^2 f_A^2 f_B^2}{P^*(D)}\right]^{x_{1D}} \left[\frac{\lambda_{AB} \times 2 f_A^2 f_B (1-f_B)}{P^*(D)}\right]^{x_{2D}} \\
&\quad \left[\frac{f_A^2 (1-f_B)^2}{P^*(D)}\right]^{x_{3D}} \left[\frac{\lambda_{AB} \times 2 f_A (1-f_A) f_B^2)}{P^*(D)}\right]^{x_{4D}} \\
&\quad \left[\frac{\lambda_{AB} \times 2 f_A (1-f_A) f_B (1-f_B)}{P^*(D)}\right]^{x_{5D}} \left[\frac{2 f_A (1-f_A) f_B (1-f_B)}{P^*(D)}\right]^{x_{6D}} \\
&\quad \left[\frac{2 f_A (1-f_A) (1-f_B)^2}{P^*(D)}\right]^{x_{7D}} \left[\frac{(1-f_A)^2 f_B^2}{P^*(D)}\right]^{x_{8D}} \\
&\quad \left[\frac{2 (1-f_A)^2 f_B (1-f_B)}{P^*(D)}\right]^{x_{9D}} \left[\frac{(1-f_A)^2 (1-f_B)^2}{P^*(D)}\right]^{x_{10D}} \\
&\quad \frac{n_C!}{x_{1C}! \times x_{2C}! \times \cdots \times x_{10C}!} \left[\frac{\lambda_{AB}^2 f_A^2 f_B^2}{P^*(C)}\right]^{x_{1C}} \left[\frac{\lambda_{AB} \times 2 f_A^2 f_B (1-f_B)}{P^*(C)}\right]^{x_{2C}} \\
&\quad \left[\frac{f_A^2 (1-f_B)^2}{P^*(C)}\right]^{x_{3C}} \left[\frac{\lambda_{AB} \times 2 f_A (1-f_A) f_B^2)}{P^*(C)}\right]^{x_{4C}} \\
&\quad \left[\frac{\lambda_{AB} \times 2 f_A (1-f_A) f_B (1-f_B)}{P^*(C)}\right]^{x_{5C}} \left[\frac{2 f_A (1-f_A) f_B (1-f_B)}{P^*(C)}\right]^{x_{6C}} \\
&\quad \left[\frac{2 f_A (1-f_A) (1-f_B)^2}{P^*(C)}\right]^{x_{7C}} \left[\frac{(1-f_A)^2 f_B^2}{P^*(C)}\right]^{x_{8C}} \\
&\quad \left[\frac{2 (1-f_A)^2 f_B (1-f_B)}{P^*(C)}\right]^{x_{9C}} \left[\frac{(1-f_A)^2 (1-f_B)^2}{P^*(C)}\right]^{x_{10C}} .
\end{aligned}
\tag{3.5}
$$

Consequently, we can obtain conditional pdf by (3.3) and (3.5),

$$
\begin{aligned}
k\left(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}\right) &= \frac{f\left(\boldsymbol{x}|\boldsymbol{\theta}\right)}{g\left(\boldsymbol{y}|\boldsymbol{\theta}\right)} \\
&= \frac{y_{5D}!}{x_{5D}!\times x_{6D}!}
\left[\frac{\lambda_{AB}\times 2f_A(1-f_A)f_B(1-f_B)}{\lambda_{AB}\times 2f_A(1-f_A)f_B(1-f_B)+2f_A(1-f_A)f_B(1-f_B)}\right]^{x_{5D}} \\
&\quad \left[\frac{2f_A(1-f_A)f_B(1-f_B)}{\lambda_{AB}\times 2f_A(1-f_A)f_B(1-f_B)+2f_A(1-f_A)f_B(1-f_B)}\right]^{x_{5D}} \\
&\quad \frac{y_{5C}!}{x_{5C}!\times x_{6C}!}
\left[\frac{\lambda_{AB}\times 2f_A(1-f_A)f_B(1-f_B)}{\lambda_{AB}\times 2f_A(1-f_A)f_B(1-f_B)+2f_A(1-f_A)f_B(1-f_B)}\right]^{x_{5C}} \\
&\quad \left[\frac{2f_A(1-f_A)f_B(1-f_B)}{\lambda_{AB}\times 2f_A(1-f_A)f_B(1-f_B)+2f_A(1-f_A)f_B(1-f_B)}\right]^{x_{6C}} \\
&= \frac{y_{5D}!}{x_{5D}!\times x_{6D}!}\left(\frac{\lambda_{AB}}{\lambda_{AB}+1}\right)^{x_{5D}}\left(\frac{1}{\lambda_{AB}+1}\right)^{x_{6D}} \\
&\quad \frac{y_{5C}!}{x_{5C}!\times x_{6C}!}\left(\frac{1}{1+1}\right)^{x_{5C}}\left(\frac{1}{1+1}\right)^{x_{6C}} \\
&= \frac{y_{5D}!}{x_{5D}!\times x_{6D}!}\left(\frac{\lambda_{AB}}{\lambda_{AB}+1}\right)^{x_{5D}}\left(\frac{1}{\lambda_{AB}+1}\right)^{x_{6D}} \\
&\quad \frac{y_{5C}!}{x_{5C}!\times x_{6C}!}\left(\frac{1}{2}\right)^{x_{5C}}\left(\frac{1}{2}\right)^{x_{6C}}. \\
\therefore X_{5D}, X_{5C}|\boldsymbol{Y},\boldsymbol{\theta} &\sim \text{Bin}\left(y_{5D}, \frac{\lambda_{AB}}{\lambda_{AB}+1}\right)\times \text{Bin}\left(y_{5C}, \frac{1}{2}\right) \\
X_{6D}, X_{6C}|\boldsymbol{Y},\boldsymbol{\theta} &\sim \text{Bin}\left(y_{5D}, \frac{1}{\lambda_{AB}+1}\right)\times \text{Bin}\left(y_{5C}, \frac{1}{2}\right).
\end{aligned}
\tag{3.6}
$$

3.2.1 Expectation

We can obtain $Q\left(\boldsymbol{\theta}^{'}|\boldsymbol{\theta},\boldsymbol{y}\right)$ by the result from (3.6),

$$
\begin{aligned}
Q\left(\boldsymbol{\theta}^{'}|\boldsymbol{\theta},\boldsymbol{y}\right) &= E\left[\ln L^{c}\left(\boldsymbol{\theta}^{'}|\boldsymbol{x}\right)|\boldsymbol{\theta},\boldsymbol{y}\right] \\
&= E\Bigg\{x_{1D}\ln\left(\frac{\lambda_{AB}^{'}{}^{2}f_{A}^{'}{}^{2}f_{B}^{'}{}^{2}}{P^{*}(D)}\right)+x_{2D}\ln\left(\frac{\lambda_{AB}^{'}\times 2f_{A}^{'}{}^{2}f_{B}^{'}(1-f_{B}^{'})}{P^{*}(D)}\right) \\
&\quad +x_{3D}\ln\left(\frac{f_{A}^{'}{}^{2}(1-f_{B}^{'})^{2}}{P^{*}(D)}\right)+x_{4D}\ln\left(\frac{\lambda_{AB}^{'}\times 2f_{A}^{'}(1-f_{A}^{'})f_{B}^{'}{}^{2}}{P^{*}(D)}\right) \\
&\quad +x_{5D}\ln\left(\frac{\lambda_{AB}^{'}\times 2f_{A}^{'}(1-f_{A}^{'})f_{B}^{'}(1-f_{B}^{'})}{P^{*}(D)}\right) \\
&\quad +x_{6D}\ln\left(\frac{2f_{A}^{'}(1-f_{A}^{'})f_{B}^{'}(1-f_{B}^{'})}{P^{*}(D)}\right)+x_{7D}\ln\left(\frac{2f_{A}^{'}(1-f_{A}^{'})(1-f_{B}^{'})^{2}}{P^{*}(D)}\right) \\
&\quad +x_{8D}\ln\left(\frac{2f_{A}^{'}(1-f_{A}^{'})(1-f_{B}^{'})^{2}}{P^{*}(D)}\right)+x_{9D}\ln\left(\frac{2(1-f_{A}^{'})^{2}f_{B}^{'}(1-f_{B}^{'})}{P^{*}(D)}\right) \\
&\quad +x_{10D}\ln\left(\frac{(1-f_{A}^{'})^{2}(1-f_{B}^{'})^{2}}{P^{*}(D)}\right) \\
&\quad +x_{1C}\ln\left(\frac{f_{A}^{'}{}^{2}f_{B}^{'}{}^{2}}{P^{*}(C)}\right)+x_{2C}\ln\left(\frac{2f_{A}^{'}{}^{2}f_{B}^{'}(1-f_{B}^{'})}{P^{*}(C)}\right) \\
&\quad +x_{3C}\ln\left(\frac{f_{A}^{'}{}^{2}(1-f_{B}^{'})^{2}}{P^{*}(C)}\right)+x_{4C}\ln\left(\frac{2f_{A}^{'}(1-f_{A}^{'})f_{B}^{'}{}^{2}}{P^{*}(C)}\right) \\
&\quad +x_{5C}\ln\left(\frac{2f_{A}^{'}(1-f_{A}^{'})f_{B}^{'}(1-f_{B}^{'})}{P^{*}(C)}\right) \\
&\quad +x_{6C}\ln\left(\frac{2f_{A}^{'}(1-f_{A}^{'})f_{B}^{'}(1-f_{B}^{'})}{P^{*}(C)}\right)+x_{7C}\ln\left(\frac{2f_{A}^{'}(1-f_{A}^{'})(1-f_{B}^{'})^{2}}{P^{*}(C)}\right) \\
&\quad +x_{8C}\ln\left(\frac{2f_{A}^{'}(1-f_{A}^{'})(1-f_{B}^{'})^{2}}{P^{*}(C)}\right)+x_{9C}\ln\left(\frac{2(1-f_{A}^{'})^{2}f_{B}^{'}(1-f_{B}^{'})}{P^{*}(C)}\right) \\
&\quad +x_{10C}\ln\left(\frac{(1-f_{A}^{'})^{2}(1-f_{B}^{'})^{2}}{P^{*}(C)}\right)+c|\boldsymbol{\theta},\boldsymbol{y}\Bigg\} \\
&= A\ln\left(\lambda_{AB}^{'}\right)+B\ln\left(f_{A}^{'}\right)+C\ln\left(f_{B}^{'}\right)+D\ln\left(1-f_{A}^{'}\right)+E\ln\left(1-f_{B}^{'}\right) \\
&\quad +F\ln 2-n_{D}\ln P^{*}(D)+c
\end{aligned}
\tag{3.7}
$$

where

$$
\begin{aligned}
A &= 2y_{1D} + y_{2D} + y_{4D} + y_{5D}\frac{\lambda_{AB}}{\lambda_{AB}+1} \\
B &= 2y_{1D} + 2y_{2D} + 2y_{3D} + y_{4D} + y_{5D} + y_{6D} + 2y_{1C} + 2y_{2C} + 2y_{3C} + y_{4C} + y_{5C} + y_{6C} \\
C &= 2y_{1D} + y_{2D} + 2y_{4D} + y_{5D} + 2y_{7D} + y_{8D} + 2y_{1C} + y_{2C} + 2y_{4C} + y_{5C} + 2y_{7C} + y_{8C} \\
D &= y_{4D} + y_{5D} + y_{6D} + 2y_{7D} + 2y_{8D} + 2y_{9D} + y_{4C} + y_{5C} + y_{6C} + 2y_{7C} + 2y_{8C} + 2y_{9C} \\
E &= y_{2D} + 2y_{3D} + y_{5D} + 2y_{6D} + y_{8D} + 2y_{9D} + y_{2C} + 2y_{3C} + y_{5C} + 2y_{6C} + y_{8C} + 2y_{9C} \\
F &= y_{2D} + y_{4D} + y_{5D} + y_{6D} + y_{8D} + y_{2C} + y_{4C} + y_{5C} + y_{6C} + y_{8C} \\
P^*(D) &= 1 - 2\lambda'_{AB} {f'_A}^2 {f'_B}^2 + {\lambda'_{AB}}^2 {f'_A}^2 {f'_B}^2 + {f'_A}^2 {f'_B}^2 - 2f'_A f'_B + 2\lambda'_{AB} f'_A f'_B \\
P^*(C) &= 1
\end{aligned}
$$

3.2.2 Conditional maximization

By partial differentiation,

$$
\begin{aligned}
\frac{\partial Q\left(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{y}\right)}{\partial \lambda'_{AB}} &= \frac{A}{\lambda'_{AB}} - n_D\left(\frac{-2{f'_A}^2{f'_B}^2 + 2\lambda'_{AB}{f'_A}^2{f'_B}^2 + 2f'_A f'_B}{1 - 2\lambda'_{AB}{f'_A}^2{f'_B}^2 + {\lambda'_{AB}}^2{f'_A}^2{f'_B}^2 + {f'_A}^2{f'_B}^2 - 2f'_A f'_B + 2\lambda'_{AB}f'_A f'_B}\right) \\
&= \frac{A}{\lambda'_{AB}} - n_D\left(\frac{2f'_A f'_B}{1 - f'_A f'_B + \lambda'_{AB} f'_A f'_B}\right) = 0
\end{aligned}
$$

Obviously, the estimator of relative penetrance rate λ'_{AB} which maximized likelihood is,

$$
\lambda'_{AB} = \frac{(f'_A f'_B - 1)A}{f'_A f'_B (A - 2n_D)} \tag{3.8}
$$

The estimators of allele frequency f'_A, f'_B which maximized likelihood are,

$$
\frac{\partial Q\left(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{y}\right)}{\partial f'_A} = \frac{B}{f'_A} - \frac{D}{1-f'_A} - n_D\left(\frac{2(\lambda'_{AB}-1)f'_B}{1 - f'_A f'_B + \lambda'_{AB} f'_A f'_B}\right) = 0 \Rightarrow P{f'_A}^2 + Qf'_A + B = 0
$$

$$
\frac{\partial Q\left(\boldsymbol{\theta}'|\boldsymbol{\theta},\boldsymbol{y}\right)}{\partial f'_B} = \frac{C}{f'_B} - \frac{E}{1-f'_B} - n_D\left(\frac{2(\lambda'_{AB}-1)f'_A}{1 - f'_A f'_B + \lambda'_{AB} f'_A f'_B}\right) = 0 \Rightarrow R{f'_B}^2 + Sf'_B + C = 0
$$

where

$$
\begin{aligned}
P &= \left[(f'_B - \lambda'_{AB}f'_B)(B+D) + 2n_D(\lambda'_{AB}-1)f'_B\right] \\
Q &= \left[(-f'_B + \lambda'_{AB}f'_B - 1)B - D - 2n_D(\lambda'_{AB}-1)f'_B\right] \\
R &= \left[(f'_A - \lambda'_{AB}f'_A)(C+E) + 2n_D(\lambda'_{AB}-1)f'_A\right] \\
S &= \left[(-f'_A + \lambda'_{AB}f'_A - 1)C - E - 2n_D(\lambda'_{AB}-1)f'_A\right]
\end{aligned}
$$

since both of equations above are parabolic,

$$
f'_A = \frac{-Q \pm \sqrt{Q^2 - 4PB}}{2P} \tag{3.9}
$$

$$
f'_B = \frac{-S \pm \sqrt{S^2 - 4RC}}{2R} \tag{3.10}
$$

3.2.3 Simulation

The simulation follows the following algorithm,

1. Set the initial values, $\lambda_{AB} = 1$, f_A, f_B are the method of moment estimate.
2. Update $\boldsymbol{\theta} = (\lambda_{AB}, f_A, f_B)$ by equation (3.8), equation (3.9), and equation (3.10) by sequence such that likelihood elevate.
   - If the estimate is out of reasonable range or not real number root, the old one remains unchanged.
   - If two solutions for f_A or f_B both are reasonable, choose the one with higher likelihood function.
3. Repeat step 2 until $L^{in}\left(\boldsymbol{\theta}^{new}|\boldsymbol{y}\right) - L^{in}\left(\boldsymbol{\theta}^{old}|\boldsymbol{y}\right) < 1 \times 10^{-30}$.

The brief simulation result is displayed in table 3.9 below.

Table 3.9: Expectation-Conditional Maximization by Simulation

λ_{AB} $(\hat{\lambda}_{AB})$	f_A $(\hat{f}_A)$	f_B $(\hat{f}_B)$
1.500(1.455±0.245)	0.100(0.100±0.004)	0.100(0.101±0.004)
1.500(1.483±0.162)	0.100(0.100±0.003)	0.200(0.200±0.005)
1.500(1.502±0.114)	0.100(0.100±0.004)	0.300(0.301±0.005)
1.500(1.503±0.093)	0.100(0.100±0.004)	0.400(0.400±0.006)
1.500(1.494±0.170)	0.200(0.201±0.005)	0.100(0.100±0.004)
1.500(1.489±0.114)	0.200(0.200±0.006)	0.200(0.200±0.005)
1.500(1.511±0.090)	0.200(0.200±0.005)	0.300(0.300±0.006)
1.500(1.487±0.072)	0.200(0.200±0.004)	0.400(0.400±0.007)
1.500(1.508±0.151)	0.300(0.300±0.006)	0.100(0.099±0.004)
1.500(1.502±0.103)	0.300(0.300±0.006)	0.200(0.200±0.005)
1.500(1.495±0.070)	0.300(0.300±0.006)	0.300(0.300±0.005)
1.500(1.506±0.062)	0.300(0.300±0.006)	0.400(0.401±0.006)
1.500(1.510±0.117)	0.400(0.400±0.006)	0.100(0.100±0.004)
1.500(1.506±0.067)	0.400(0.401±0.007)	0.200(0.200±0.005)
1.500(1.505±0.072)	0.400(0.400±0.006)	0.300(0.299±0.005)
1.500(1.504±0.053)	0.400(0.401±0.006)	0.400(0.401±0.006)
2.000(1.969±0.277)	0.100(0.100±0.004)	0.100(0.100±0.004)
2.000(2.001±0.189)	0.100(0.100±0.003)	0.200(0.200±0.005)
2.000(2.012±0.126)	0.100(0.100±0.004)	0.300(0.301±0.006)
2.000(2.008±0.101)	0.100(0.100±0.004)	0.400(0.400±0.006)
2.000(2.015±0.208)	0.200(0.200±0.005)	0.100(0.100±0.003)
2.000(1.997±0.119)	0.200(0.200±0.005)	0.200(0.200±0.004)

Table 3.9: Expectation-Conditional Maximization by Simulation

λ_{AB} ($\hat{\lambda}_{AB}$)	f_A ($\hat{f}_A$)	f_B ($\hat{f}_B$)
2.000(2.000±0.109)	0.200(0.200±0.005)	0.300(0.299±0.006)
2.000(1.999±0.082)	0.200(0.200±0.005)	0.400(0.400±0.006)
2.000(2.022±0.170)	0.300(0.299±0.006)	0.100(0.100±0.004)
2.000(2.028±0.113)	0.300(0.299±0.007)	0.200(0.199±0.005)
2.000(2.014±0.086)	0.300(0.300±0.005)	0.300(0.299±0.005)
2.000(2.006±0.072)	0.300(0.300±0.005)	0.400(0.400±0.005)
2.000(2.013±0.140)	0.400(0.400±0.006)	0.100(0.100±0.003)
2.000(1.981±0.094)	0.400(0.400±0.006)	0.200(0.201±0.005)
2.000(2.002±0.067)	0.400(0.401±0.006)	0.300(0.300±0.005)
2.000(1.993±0.064)	0.400(0.399±0.006)	0.400(0.399±0.006)
2.500(2.537±0.281)	0.100(0.100±0.003)	0.100(0.100±0.004)
2.500(2.469±0.177)	0.100(0.100±0.004)	0.200(0.200±0.005)
2.500(2.469±0.143)	0.100(0.100±0.004)	0.300(0.301±0.006)
2.500(2.496±0.113)	0.100(0.100±0.004)	0.400(0.399±0.006)
2.500(2.473±0.214)	0.200(0.200±0.005)	0.100(0.100±0.004)
2.500(2.497±0.157)	0.200(0.199±0.005)	0.200(0.200±0.005)
2.500(2.494±0.112)	0.200(0.200±0.004)	0.300(0.299±0.006)
2.500(2.504±0.100)	0.200(0.200±0.004)	0.400(0.399±0.006)
2.500(2.500±0.156)	0.300(0.301±0.006)	0.100(0.100±0.004)
2.500(2.497±0.129)	0.300(0.300±0.005)	0.200(0.200±0.004)
2.500(2.500±0.096)	0.300(0.301±0.006)	0.300(0.300±0.006)
2.500(2.506±0.091)	0.300(0.301±0.005)	0.400(0.399±0.006)
2.500(2.521±0.143)	0.400(0.400±0.006)	0.100(0.100±0.004)
2.500(2.480±0.094)	0.400(0.401±0.006)	0.200(0.200±0.005)
2.500(2.499±0.084)	0.400(0.401±0.006)	0.300(0.301±0.006)
2.500(2.494±0.070)	0.400(0.400±0.006)	0.400(0.400±0.005)
3.000(3.022±0.338)	0.100(0.100±0.004)	0.100(0.100±0.004)
3.000(2.990±0.224)	0.100(0.100±0.003)	0.200(0.200±0.005)
3.000(2.988±0.161)	0.100(0.101±0.003)	0.300(0.300±0.005)
3.000(2.995±0.143)	0.100(0.100±0.003)	0.400(0.401±0.006)
3.000(2.979±0.231)	0.200(0.200±0.005)	0.100(0.100±0.004)
3.000(3.001±0.160)	0.200(0.200±0.005)	0.200(0.201±0.005)
3.000(3.009±0.130)	0.200(0.200±0.004)	0.300(0.299±0.005)
3.000(3.011±0.104)	0.200(0.201±0.004)	0.400(0.400±0.006)
3.000(2.980±0.205)	0.300(0.299±0.006)	0.100(0.100±0.004)
3.000(3.014±0.147)	0.300(0.301±0.006)	0.200(0.200±0.005)
3.000(2.978±0.115)	0.300(0.301±0.005)	0.300(0.300±0.005)

Table 3.9: Expectation-Conditional Maximization by Simulation

λ_{AB} ($\hat{\lambda}_{AB}$)	f_A ($\hat{f}_A$)	f_B ($\hat{f}_B$)
3.000(3.008±0.085)	0.300(0.300±0.005)	0.400(0.400±0.005)
3.000(2.995±0.173)	0.400(0.399±0.006)	0.100(0.100±0.004)
3.000(3.013±0.121)	0.400(0.400±0.005)	0.200(0.200±0.005)
3.000(2.998±0.089)	0.400(0.400±0.006)	0.300(0.300±0.005)
3.000(3.011±0.080)	0.400(0.401±0.006)	0.400(0.400±0.005)

Chapter 4

Analysis of the Data from WTCCC

The Wellcome Trust Case Control Consortium (WTCCC), a research group funded by the Wellcome Trust in UK and engages to designing and analyzing genome-wide association study (GWAS). In phase one study, the WTCCC identified genetic variants which affect susceptibility to 7 common complex diseases, bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes by the Affymetrix 500K and Illumina 550K chips, and the results published in Nature [The Wellcome Trust Case Control Consortium, 2007]. Moreover, additional 5 diseases, ankylosing spondylitis, autoimmune thyroid disease, multiple sclerosis, breast cancer, and tuberculosis have been studied. In phase two study, the WTCCC will perform association studies with other 13 diseases, ankylosing spondylitis, Barrett's oesophagus and oesophageal adenocarcinoma, glaucoma, ischaemic stroke, multiple sclerosis, pre-eclampsia, Parkinson's disease, psychosis endophenotypes, psoriasis, schizophrenia, ulcerative colitis and visceral leishmaniasis by the Affymetrix v6.0 and Illumina 1M chips.

4.1 Hypertension

4.1.1 Data source

The data for hypertension study comprised 1504 controls from the 1958 British Birth Cohort (58C), 1500 controls from the UK Blood Service Control Group (NBS), and 2001 cases from the WTCCC Hypertension Group (HT). The data is called by Chiamo which is developed by the WTCCC instead of the standard algorithm, BRLMM by Affymetrix.

4.1.2 Quality control

The quality control follows the WTCCC's procedures as the description in literature review mentions. Furthermore, there is no missing data by Chiamo calling algorithm,

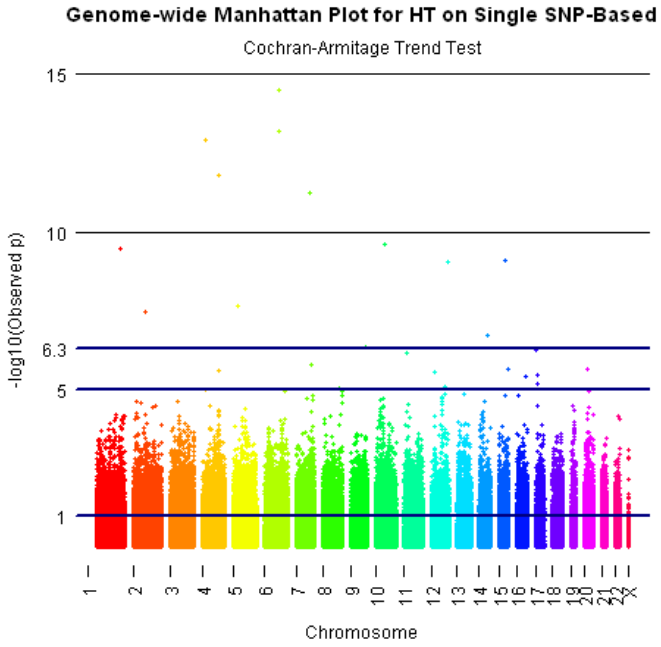

Figure 4.1: Genome-wide Manhattan Plot for Hypertension on Single SNP-Based by Cochran-Armitage Trend Test

therefore there is no excluded SNP and individual by excluding SNP and individual if SNP call rate is $\leq 95\%$ and sample call rate is $\leq 97\%$, respectively.

For exclusion of SNPs, we filtered out 62701 SNPs by minor allele frequency (MAF) is $< 1\%$, and the other 31779 SNPs by Hardy-Weinberg equilibrium (HWE).

For exclusion of samples, we sieved out 23 samples by heterozygosity per individual is $< 22.5\%$ or $> 30\%$, and the other 37 samples by cryptic relatedness.

Finally, the raw data we used reduced to 406088 (500568 - 94480) SNPs and 4945 samples after data quality control above.

4.1.3 Test of association

Single SNP association

First of all, for the processed data above, we adopt Fisher's exact test and Cochran-Armitage trend test for the genotypic test and the allele test respectively. Subjects are SNPs whose p-value is less than 5×10^{-7} (the strongest association, see table 4.1 and 4.2) or greater than 5×10^{-7} and less than 1×10^{-5} (moderate association, see table 4.3 and 4.4) for either the genotypic test or the allele test. Even though we found that the genetic variants evaluated the strongest and moderate associated with hypertension risk, some associated SNPs do not identify known genes or the relevance to hypertension.

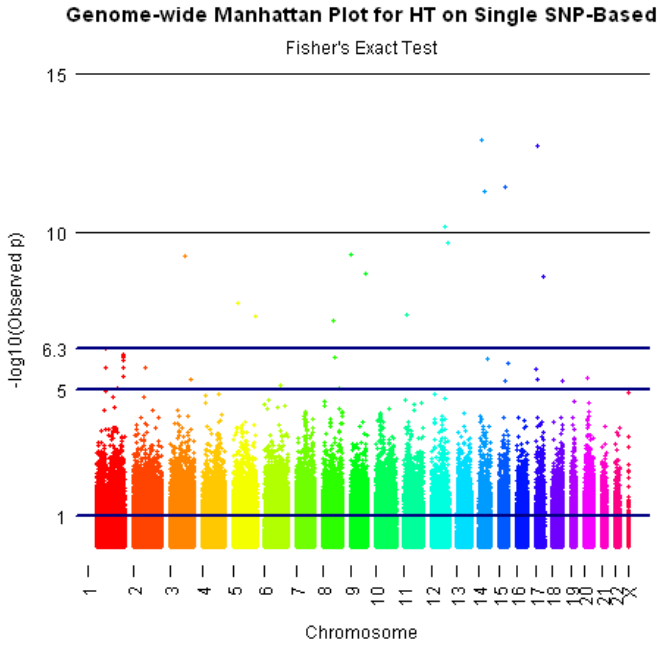

Figure 4.2: Genome-wide Manhattan Plot for Hypertension on Single SNP-Based by Fisher's Exact Test

CHRM2 (cholinergic receptor, muscarinic 2) belongs to a larger family of G protein-coupled receptors. The muscarinic cholinergic receptor 2 is involved in mediation of bradycardia and a decrease in cardiac contractility [Hautala et al., 2009]; [Zhang et al., 2008]. Carriers of the variant G of CHRM2 (rs7800093) has a significantly lower or higher risk of hypertension compared with individuals with the common homozygote genotype: odds ratio [95% CI] for heterozygotes 0.02 [0.00-0.11] and for homozygotes 53.00 [12.98-216.38].

KCNB2 (potassium voltage-gated channel, Shab-related subfamily, member 2), the diverse functions of the protein include regulating neurotransmitter release, heart rate, insulin secretion, neuronal excitability, epithelial electrolyte transport, smooth muscle contraction, and cell volume. KCNB2 (rs11782342) has a significant increase in risk among homozygote variants: odds ratio [95% CI] = 1.98[1.56-2.53]. The association between KCNB2 and cardiovascular disease risk has been found in the previous study [Vasan et al., 2007].

HTR3B (5-hydroxytryptamine (serotonin) receptor 3B) encodes subunit B of the type 3 receptor for 5-hydroxytryptamine (serotonin), a biogenic hormone that functions as a neurotransmitter, a hormone, and a mitogen. It is a known gene affecting the heart rate [Silva et al., 2007]. The variant allele G in HTR3B (rs17116117) shows significantly increase risk compared with common homozygote genotype, especially among heterozygote variants: odds ratio [95% CI] = 3.76[3.13-4.52].

Table 4.1: Genes of the Genome Showing the Strongest Association

Gene	Chromosome	dbSNP ID	Function	Trend P-value	Genotypic P-value
	1	rs10494787		4.69E-02	3.65E-22
	1	rs825148		3.25E-10	2.50E-101
	2	rs1870340		3.30E-08	4.79E-36
	3	rs804980		1.43E-03	5.64E-10
	4	rs16837871		3.27E-26	1.80E-41
	4	rs1553460		1.22E-13	1.29E-62
LOC100129858	4	rs6840033	Intron	1.64E-12	8.94E-23
	5	rs4867173		2.28E-08	1.72E-08
	5	SNP_A-2171701		2.67E-02	4.46E-08
	6	rs4131463		6.25E-14	4.90E-89
	6	rs10499044		3.01E-15	5.47E-24
	7	rs193837		2.97E-04	4.09E-27
RPL18P4	7	rs1528356	Intron	5.81E-12	2.96E-133
CHRM2*	7	rs7800093	Intron	1.59E-06	6.25E-44
KCNB2*	8	rs11782342	Intron	9.20E-04	6.59E-08
	9	rs7864098		9.20E-01	5.12E-10
	9	rs17797701		1.07E-03	2.48E-52
	9	rs488101		4.50E-07	2.19E-09
	10	rs11005510		2.36E-10	3.65E-23
OTOG	11	rs11024327	Intron	6.61E-07	4.36E-08
HTR3B*	11	rs17116117	Intron	5.07E-49	2.70E-48
	12	rs10843660		1.90E-32	1.04E-69
CHST11	12	rs11112069	Intron	4.54E-03	6.70E-11
	12	rs4765066		8.52E-10	2.18E-10
	13	rs17667894		5.41E-21	3.70E-40
SIP1	14	rs8011855	Intron	3.35E-03	1.23E-13
RHOJ	14	rs1957779	nearGene-5	2.34E-05	5.39E-12
	14	rs6574988		2.00E-07	1.03E-06
	15	rs2865199		8.24E-10	3.68E-12
	16	rs16955238		3.88E-06	3.61E-41
	17	SNP_A-1948953		6.31E-06	1.81E-13
	17	rs7217721		3.80E-04	2.47E-09

* Denotes the gene or SNP has been found in published document.

The variant in rs2820037 is significantly associated with hypertension as the previous study described [The Wellcome Trust Case Control Consortium, 2007], [Ehret et al., 2008]. The SNP rs11782342 has a significant increase in risk among heterozygote variants: odds ratio [95% CI] = 1.41[1.24-1.60].

GAB1 (GRB2-associated binding protein 1) encodes the protein which is a member of the IRS1-like multisubstrate docking protein family. The protein is an important mediator of branching tubulogenesis and plays a central role in cellular growth response,

Table 4.2: Detection of SNPs with the Strongest Association

dbSNP ID	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
rs10494787	G	0.69[0.57-0.83]	14.09[6.45-30.78]	0.068	0.079
rs825148	C	0.05[0.02-0.10]	Inf[NaN-Inf]	0.041	0.078
rs1870340	G	0.31[0.20-0.49]	114.32[15.88-822.97]	0.021	0.044
rs804980	A	0.91[0.80-1.03]	2.02[1.60-2.54]	0.217	0.246
rs16837871	A	0.36[0.31-0.42]	0.79[0.58-1.08]	0.183	0.101
rs1553460	T	0.61[0.53-0.69]	2.82[2.37-3.34]	0.291	0.369
rs6840033	T	0.52[0.45-0.59]	0.88[0.69-1.12]	0.236	0.174
rs4867173	T	1.48[1.30-1.68]	1.22[0.75-1.98]	0.132	0.171
SNP_A-2171701	T	0.93[0.80-1.07]	3.18[2.09-4.84]	0.117	0.132
rs4131463	C	0.09[0.05-0.16]	116.72[28.89-471.54]	0.037	0.081
rs10499044	C	0.44[0.37-0.51]	0.97[0.67-1.42]	0.134	0.081
rs193837	C	0.74[0.62-0.88]	10.38[5.90-18.24]	0.084	0.107
rs1528356	G	0.00[0.00-0.03]	27.77[15.09-51.08]	0.057	0.104
rs7800093	G	0.02[0.00-0.11]	53.00[12.98-216.38]	0.017	0.036
rs11782342	A	0.97[0.86-1.10]	1.98[1.56-2.53]	0.226	0.255
rs7864098	A	0.75[0.64-0.88]	3.62[2.20-5.97]	0.090	0.091
rs17797701	G	0.01[0.00-0.08]	28.04[10.24-76.79]	0.024	0.038
rs488101	C	0.68[0.60-0.77]	0.74[0.62-0.88]	0.384	0.334
rs11005510	A	0.01[0.00-0.10]	Inf[NaN-Inf]	0.017	0.003
rs11024327	A	1.44[1.27-1.63]	1.14[0.81-1.59]	0.172	0.212
rs17116117	G	3.76[3.13-4.52]	1.77[0.11-28.34]	0.032	0.101
rs10843660	T	0.31[0.27-0.35]	0.53[0.45-0.62]	0.430	0.303
rs11112069	A	0.88[0.77-1.00]	2.21[1.71-2.85]	0.183	0.207
rs4765066	A	1.55[1.36-1.76]	1.22[0.78-1.92]	0.129	0.173
rs17667894	G	0.02[0.01-0.07]	1.62[0.58-4.46]	0.035	0.005
rs8011855	A	0.88[0.74-1.05]	8.53[4.34-16.77]	0.069	0.086
rs1957779	A	1.69[1.46-1.96]	1.44[1.21-1.72]	0.474	0.515
rs6574988	T	1.45[1.26-1.67]	1.63[0.83-3.20]	0.090	0.122
rs2865199	C	0.21[0.12-0.35]	Inf[NaN-Inf]	0.019	0.005
rs16955238	C	0.22[0.13-0.35]	Inf[NaN-Inf]	0.022	0.042
SNP_A-1948953	A	0.99[0.88-1.12]	0.35[0.26-0.48]	0.302	0.262
rs7217721	C	1.05[0.84-1.30]	15.88[4.85-52.01]	0.037	0.053

transformation and apoptosis. Carriers of the variant T of GAB1 (rs300916) has a significantly lower risk of hypertension compared with individuals with the common homozygote genotype: odds ratio [95% CI] for heterozygotes 0.81 [0.72-0.92] and for homozygotes 0.67 [0.56-0.80]. Nakaoka has proved that the relationship between GAB1 and hypertrophic cardiomyopathy [Nakaoka et al., 2003], and hypertension can result in hypertrophic cardiomyopathy.

BCAT1 (branched chain aminotransferase 1, cytosolic) encodes the cytosolic form of the enzyme branched-chain amino acid transaminase. This enzyme catalyzes the reversible transamination of branched-chain alpha-keto acids to branched-chain L-amino acids es-

Table 4.3: Genes of the Genome Showing Moderate Association

Gene	Chromosome	dbSNP ID	Function	Trend P-value	Genotypic P-value
NEGR1	1	rs10889923	Intron	1.13E-01	2.03E-06
	1	rs1896250		3.84E-04	5.08E-07
	1	rs12729977		6.25E-01	9.05E-06
	1	rs2820026		6.70E-05	3.96E-06
	1	rs9428826		1.21E-04	1.95E-06
	1	rs2790622		7.96E-05	8.58E-07
	1	rs2820037*		8.10E-05	7.78E-07
	1	rs2820038		7.25E-05	9.26E-07
	1	rs2820046		8.35E-05	1.12E-06
CREG2	2	rs4850969	Intron	1.50E-01	2.00E-06
PRKCI	3	rs2140825	Intron	4.93E-02	5.01E-06
GAB1*	4	rs300916	Intron	2.49E-06	1.45E-05
LOC100128588	6	rs1935683	Intron	9.33E-05	7.29E-06
CNBD1	8	rs7825717	Intron	9.36E-01	9.28E-07
ZHX2	8	rs10095188	Intron	1.27E-02	9.48E-06
	8	rs4242382		8.96E-06	3.86E-05
	8	rs11166882		9.58E-06	5.03E-05
BCAT1*	12	rs7961152	Intron	2.86E-06	1.41E-05
MYBPC1*	12	rs11110912	Intron	8.12E-06	1.84E-05
	15	rs921535		1.63E-05	5.47E-06
LOC100132798*	15	rs2398162	Intron	2.13E-06	1.44E-06
YWHAE	17	rs16945811	Intron	5.54E-07	2.24E-06
	17	rs17201619		3.58E-06	4.69E-06
ZNF236	18	rs4890866	Intron	2.04E-02	5.34E-06
SEC23B	20	rs1022684	nearGene-5	2.36E-06	4.19E-06

* Denotes the gene or SNP has been found in published document.

sential for cell growth. Hypertension can cause atherosclerosis, furthermore, BCAT has been implicated in the pathogenesis of atherosclerosis [Coles et al., 2009]. Carriers of the variant A of BCAT1 (rs7961152) has a significantly higher risk of hypertension compared with individuals with the common homozygote genotype: odds ratio [95% CI] for heterozygotes 1.17 [1.03-1.34] and for homozygotes 1.49 [1.26-1.76] [The Wellcome Trust Case Control Consortium, 2007].

MYBPC1 (rs11110912). Carriers of the variant G of MYBPC1 (rs11110912) has a significantly higher risk of hypertension compared with individuals with the common homozygote genotype: odds ratio [95% CI] for heterozygotes 1.33 [1.18-1.51] and for homozygotes 1.34 [0.97-1.86] [The Wellcome Trust Case Control Consortium, 2007]. In the previous study, MYBPC1 is also related to hypertrophic cardiomyopathy [Konno et al., 2003].

LOC100132798 is similar to hCG1774772. Carriers of the variant G of LOC100132798 (rs2398162) has a significantly higher or lower risk of hypertension compared with individ-

uals with the common homozygote genotype: odds ratio [95% CI] for heterozygotes 24.33 [3.22-183.63] and for homozygotes 0.75 [0.59-0.95] [The Wellcome Trust Case Control Consortium, 2007].

SEC23B (Sec23 homolog B (S. cerevisiae)) encodes the protein which is a member of the SEC23 subfamily of the SEC23/SEC24 family. The encoded protein has similarity to yeast Sec23p component of COPII. COPII is the coat protein complex responsible for vesicle budding from the ER. The function of this gene product has been implicated in cargo selection and concentration. Subjects with the variant T of SEC23B (rs1022684) shows significantly reduced risk compared with common homozygote genotype: odds ratio [95% CI] for heterozygotes 0.70 [0.58-0.83] and for homozygotes 0.21 [0.06-0.69].

Table 4.4: Detection of SNPs with Moderate Association

dbSNP ID	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
rs10889923	C	1.18[1.04-1.34]	0.77[0.64-0.92]	0.410	0.394
rs1896250	A	1.41[1.24-1.60]	1.21[1.01-1.45]	0.379	0.414
rs12729977	C	1.22[1.08-1.39]	0.83[0.69-1.00]	0.402	0.397
rs2820026	T	1.39[1.22-1.58]	0.97[0.65-1.44]	0.138	0.167
rs9428826	T	1.40[1.23-1.59]	0.93[0.64-1.35]	0.140	0.168
rs2790622	C	1.41[1.24-1.60]	0.90[0.61-1.33]	0.141	0.170
rs2820037	T	1.41[1.24-1.60]	0.89[0.60-1.32]	0.141	0.170
rs2820038	T	1.41[1.24-1.60]	0.90[0.61-1.34]	0.141	0.170
rs2820046	A	1.40[1.23-1.60]	0.90[0.61-1.33]	0.141	0.170
rs4850969	T	1.02[0.89-1.18]	0.08[0.02-0.32]	0.113	0.104
rs2140825	C	1.12[0.99-1.27]	0.71[0.59-0.87]	0.399	0.381
rs300916	T	0.81[0.72-0.92]	0.67[0.56-0.80]	0.406	0.359
rs1935683	C	0.73[0.65-0.83]	0.95[0.69-1.31]	0.198	0.167
rs7825717	C	1.14[0.97-1.33]	0.00[0.00-NaN]	0.082	0.081
rs10095188	C	1.02[0.90-1.16]	0.45[0.31-0.63]	0.185	0.165
rs4242382	A	0.73[0.63-0.84]	0.64[0.35-1.18]	0.125	0.097
rs11166882	T	0.64[0.35-1.18]	0.68[0.54-0.85]	0.285	0.244
rs7961152	A	1.17[1.03-1.34]	1.49[1.26-1.76]	0.413	0.461
rs11110912	G	1.33[1.18-1.51]	1.34[0.97-1.86]	0.165	0.200
rs921535	C	1.38[1.21-1.57]	1.07[0.70-1.63]	0.141	0.173
rs2398162	G	24.33[3.22-183.63]	0.75[0.59-0.95]	0.260	0.218
rs16945811	A	1.48[1.27-1.72]	1.50[0.70-3.19]	0.074	0.102
rs17201619	A	0.71[0.60-0.85]	0.19[0.06-0.63]	0.079	0.055
rs4890866	G	1.07[0.95-1.20]	0.61[0.49-0.77]	0.322	0.300
rs1022684	T	0.70[0.58-0.83]	0.21[0.06-0.69]	0.078	0.054

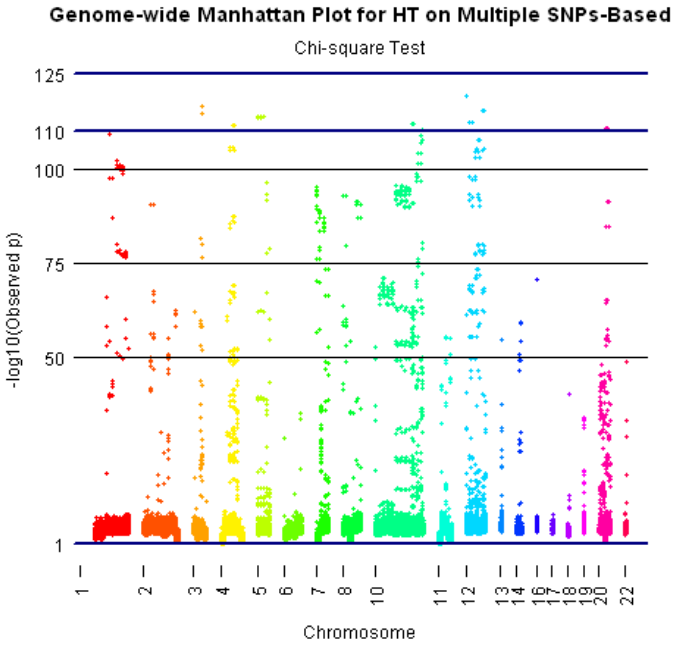

Figure 4.3: Genome-wide Manhattan Plot for Hypertension on Multiple SNPs-Based by Chi-square Test

Multiple SNPs association

According to the interactions of SNPs within the strongest and moderate association, side effects are also siginificant if main effects are associated with disease. Consequently, we do not focus on known and obvious interactions, we are interested in SNPs that are usually ignored, namely, we focus on the interactions of SNPs without single SNP associations we found before. In addition to this, we can apply filterable method as mentioned in chapter 3, setting $\lambda_{AB} = 1.75$, $f_A = 0.2$, $f_B = 0.2$ by conservative rule due to the estimate $\hat{\lambda}_{AB}$ in interactions of SNPs within the strongest and moderate association are pretty high (even $\hat{\lambda}_{AB} = 6$). Thus we can reduce computation time about $(1 - \frac{C_2^{26108}}{C_2^{406088}}) = 99.59\%$ by p-value is higher than 1×10^{-1} in single association, i.e. we set $\xi_1 = 2.7\,(\alpha = 0.1)$ due to our tolerable loss of power is under 1%. Of course, adjusting the threshold ξ_1 repeatedly for the methodology as mentioned in chapter 3 can find the threshold ξ_1 as exact as possible. Consequently, the computation time would be improved as possible.

In the beginning, we narrowed down the target SNPs for less computation time by p-value between 1×10^{-4} and 1×10^{-5} in single association. By figure 4.3, we listed interactions within chromosome at table 4.5 with $1 \times 10^{-110} \leq$ p-value $\leq 1 \times 10^{-125}$, and figure 4.4 shows the relation of p-value between single SNP and paired SNPs association.

The SNPs rs2091244, rs2177686, rs17073046 all locate on the gene MAGI1. MAGI1

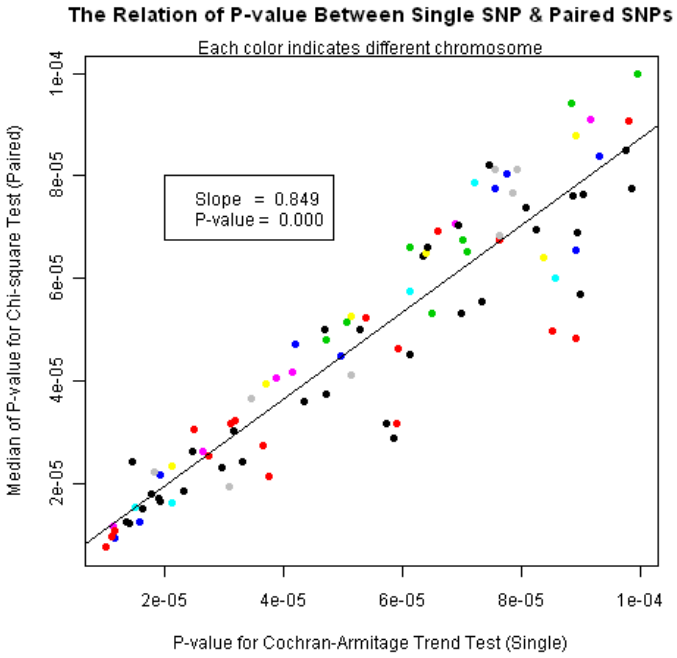

Figure 4.4: The Relation of P-value Between Single SNP & Paired SNPs Association for Hypertension

(membrane associated guanylate kinase, WW and PDZ domain containing 1) encodes the protein which is a member of the membrane-associated guanylate kinase homologue (MAGUK) family.The product of this gene may play a role as scaffolding protein at cell-cell junctions. To date, we just know that MAGI1 is important for vascular endothelial-cadherin-dependent Rap1 activation upon cell-cell contact [Sakurai et al., 2006], however, we cannot connect it with hypertension.

GAB1 and BCAT1 not only have been found in the single SNP association we mentioned before but also have been proved by previous study. However, some interactions on genes C10orf72, C10orf128, LOC728883 or not identify genes have not yet been proposed and proven from the biological aspect.

Table 4.5: Detection of Multiple SNPs-Based Association

Chromosome	dbSNP ID 1 (Gene 1)	dbSNP ID 2 (Gene 2)	Trend P-value	Trend P-value 1	Trend P-value 2	Relative Penetrance Rate
3	rs2091244 (MAGI1*)	rs2177686 (MAGI1*)	1.47E-115	9.80E-05	1.77E-04	6.55
3	rs2091244 (MAGI1*)	rs17073046 (MAGI1*)	3.11E-117	9.80E-05	1.22E-04	6.58
4	rs300915 (GAB1*)	rs300913 (GAB1*)	4.00E-112	5.06E-05	4.71E-05	6.44
5	rs1490800	rs1490796	3.09E-114	9.94E-05	7.55E-05	5.95
5	rs1490800	rs1490795	9.17E-115	9.94E-05	7.75E-05	5.95
5	rs1490796	rs1490795	1.06E-114	7.55E-05	7.75E-05	5.96
10	rs12269023 (C10orf72)	rs7097933 (C10orf72)	1.54E-112	3.72E-05	3.46E-05	6.77
10	rs2725181 (C10orf128)	rs2725190 (LOC728883)	5.47E-111	7.86E-05	1.58E-04	8.67
12	rs11613673 (BCAT1*)	rs12424348 (BCAT1*)	4.83E-120	6.95E-05	1.49E-04	10.28
12	rs7300456	rs1452237	3.97E-113	1.65E-05	1.91E-05	6.92
12	rs4761100	rs4761102	5.44E-116	2.97E-05	2.33E-05	7.66
20	rs2424430	rs431904	2.53E-111	1.18E-05	3.65E-05	8.15

* Denotes the gene or SNP has been found in published document.

Chapter 5

Conclusion

According to the results in table 3.5, table 3.6, and the real data, the loss of power is reasonable and tolerable when λ_{AB} is large enough or the allele frequency is not too small. Each pair of SNPs association has an unknown λ_{AB} originally, but estimate all λ_{AB} is unusable because our major work is to find out a reasonable threshold by only one λ_{AB} and other parameters. We found that $\hat{\lambda}_{AB}$ within the strongest or moderate associations are quite large, such as 6.0 or 7.6, but we cannot promise that λ_{AB} for all existing associations are large, too. That is the reason why we use more conservative and robust rule as $\lambda_{AB} = 1.75$ in this study. We can reduce computation time about

- $99.04\% = (1 - \frac{C_2^{39762}}{C_2^{406088}})$, loss of power = 0.2612%, when $\xi_1 = 2.07\,(\alpha = 0.15)$
- $99.59\% = (1 - \frac{C_2^{26108}}{C_2^{406088}})$, loss of power = 0.8224%, when $\xi_1 = 2.7\,(\alpha = 0.1)$
- $99.77\% = (1 - \frac{C_2^{19424}}{C_2^{406088}})$, loss of power = 1.6915%, when $\xi_1 = 3.17\,(\alpha = 0.075)$

Analyzing the data with this approach, which imitates WTCCC of hypertension, we have detected parts of known genes or SNPs, such as CHRM2 (rs7800093), KCNB2 (rs11782342), HTR3B (rs17116117), rs2820037, GAB1 (rs300916, rs300915, rs300913), BCAT1 (rs7961152, rs11613673, rs12424348), MYBPC1 (rs11110912), LOC100132798 (rs2398162), MAGI1 (rs2091244, rs2177686, rs17073046). Nevertheless, those other unknowns, such as rs825148, rs1553460, LOC100129858 (rs6840033), rs4131463, RPL18P4 (rs1528356), rs17797701, OTOG (rs11024327), rs10843660, CHST11 (rs11112069), SIP1 (rs8011855), RHOJ (rs1957779) are worthy of digging for statistical replication and biological explanation in the future. Furthermore, the associations of higher order are also our ultimate goal for finding the susceptibility for complex human diseases, for instance, hypertension and type 2 diabetes.

Originally, for convenience and custom, the approach for loss of power and ECM algorithm both are based on additive model. However, compare figure 4.1 with figure 4.2,

some SNPs' associations are clearly quite different in these two figures. Thus the extension for no model assumption may be more accurate and informative (single association test uses genotypic test instead of trend test).

We have not considered the dominant or recessive model in the method and analysis. In general, the models for most of SNPs are still unknown, integrate information (consider the dominant or recessive model additionally) from every models and revise our method is a part of future work. Using this method to calculate the loss of power and use ECM algorithm to find suitable parameters may provide a good guidance to threshold selection.

Bibliography

N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273:1516–1517, 1996.

C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilit‘a. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802, 1988.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57: 289–300, 1995.

P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11: 375–386, 1955.

R. A. Fisher. On the Interpretation of q2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85:87–94, 1922.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to hove arisen from random sampling. *Phios. Mag. Ser.*, 50:157–175, 1900.

X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278, 1993.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

A. J. Hautala, M. P. Tulppo, A. M. Kiviniemi, T. Rankinen, C. Bouchard, T. H. M'akikallio, and H. V. Huikuri. Acetylcholine receptor m2 gene variants, heart rate recovery, and risk of cardiac death after an acute myocardial infarction. *Annals of Medicine*, 41:197–207, 2009.

L. Zhang, A. Hu, H. Yuan, L. Cui, G. Miao, X. Yang, L. Wang, J. Liu, X. Liu, S. Wang, Z. Zhang, L. Liu, R. Zhao, and Y. Shen. A missense mutation in the chrm2 gene is associated with familial dilated cardiomyopathy. *Circulation Research*, 102:1426–1432, 2008.

R. S. Vasan, M. G. Larson, J. Aragam, T. J. Wang, G. F. Mitchell, S. Kathiresan, C. Newton-Cheh, J. A. Vita, M. J. Keyes, C. J. O'Donnell, D. Levy, and E. J. Benjamin. Genome-wide association of echocardiographic dimensions, brachial artery endothelial function and treadmill exercise responses in the Framingham Heart Study. *BMC Medical Genetics*, 8 Suppl 1:S2, 2007.

Gustavo J. J. Silva, Alexandre C. Pereira, Eduardo M. Krieger, and Jos'e E. Krieger. Genetic mapping of a new heart rate QTL on chromosome 8 of spontaneously hypertensive rats. *BMC Medical Genetics*, 8:17, 2007.

G. B. Ehret, A. C. Morrison, A. A. O'Connor, M. L. Grove, L. Baird, K. Schwander, A. Weder, R. S. Cooper, D. C. Rao, S. C. Hunt, E. Boerwinkle, and A. Chakravarti. Replication of the Wellcome Trust genome-wide association study of essential hypertension: the Family Blood Pressure Program. *European Journal of Human Genetics*, 16:1507–1511, 2008.

Y. Nakaoka, K. Nishida, Y. Fujio, M. Izumi, K. Terai, Y. Oshima, S. Sugiyama, S. Matsuda, S. Koyasu, K. Yamauchi-Takihara, T. Hirano, I. Kawase, and H. Hirota. Activation of gp130 transduces hypertrophic signal through interaction of scaffolding/docking protein Gab1 with tyrosine phosphatase SHP2 in cardiomyocytes. *Circulation Research*, 93:221–229, 2003.

S. J. Coles, P. Easton, H. Sharrod, S. M. Hutson, J. Hancock, V. B. Patel, and M. E. Conway. S-Nitrosoglutathione inactivation of the mitochondrial and cytosolic BCAT proteins: S-nitrosation and S-thiolation. *Biochemistry*, 48:645–656, 2009.

T. Konno, M. Shimizu, H. Ino, T. Matsuyama, M. Yamaguchi, H. Terai, K. Hayashi, M. Mabuchi, T. amd Kiyama, K. Sakata, T. Hayashi, M. Inoue, T. Kaneda, and H. Mabuchi. A novel missense mutation in the myosin binding protein-C gene is responsible for hypertrophic cardiomyopathy with left ventricular dysfunction and dilation in elderly patients. *Journal of the American College of Cardiology*, 41:781–786, 2003.

A. Sakurai, S. Fukuhara, A. Yamagishi, K. Sako, Y. Kamioka, M. Masuda, Y. Nakaoka, and N. Mochizuki. MAGI-1 is required for Rap1 activation upon cell-cell contact and for enhancement of vascular endothelial cadherin-mediated cell adhesion. *Molecular Biology of the Cell*, 17:966–976, 2006.

T. Becker and M. Knapp. Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet Epidemiol*, 27:21–32, 2004.

D. Sasieni. From genotypes to genes: doubling the sample size. *Biometrics*, 53:1253–1261, 1997.

Mark C. K. Yang. *Introduction to Statistical Methods in Modern Genetics*. Gordon and Breach Science Publishers, 2000.

K. Hao, X. Xu, N. Laird, X. Wang, and X. Xu. Power Estimation of Multiple SNP Association Test of Case-Control Study and Application. *Genetic Epidemiology*, 26: 22–30, 2004.

National Center for Biotechnology Information, National Library of Medicine, and National Institutes of Health. Entrez gene. `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene`, June 2009.