國立交通大學

統計學研究所

碩士論文

一套關於全基因相關性分析
的標準流程

A standard Flow Path of Making a Genome-wide Association (GWA) Study Analysis

研究生：陳彥銘

指導教授：黃冠華 博士

中華民國九十七年六月

一套關於全基因相關性分析
的標準流程

A standard Flow Path of Making a Genome-wide Association (GWA) Study Analysis

研究生：陳彥銘　　Student：Yan-Ming Chen

指導教授：黃冠華　　Advisor：Dr. Guan-Hua Huang

國立交通大學
統計學研究所
碩士論文

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
June 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年六月

A standard Flow Path of Making a Genome-wide Association (GWA) Study Analysis

Student :Yan-Ming Chen　Advisor:Guan-Hua Huang

Institute of Statistics
National Chiao Tung University

ABSTRACT

There are increasing evidences that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases [1]. Many studies had successfully performed the GWA study to identify novel susceptible loci. However, there is a lack of agreement about what constitutes an adequate analytic procedure. In this study, we review existing genome-wide association studies to identify such a procedure and implement the built procedure to real datasets from the Wellcome Trust case-control Consortium. Our procedure includes four steps: data management, preliminary analysis, association testing and result visualization. In order to get the true association between disease and SNP, we execute 2 preliminary processes, the quality control (QC) and population stratification. Furthermore, we can plot the quantile-quantile (Q-Q) plot and Manhattan plot to visualize association analysis results. At the end of the study, we have successfully (1) identified the necessary and important analyses for GWA, (2) identified currently available software for these analyses, (3) performed the analysis on the Wellcome Trust case-control Consortium data, and (4) provided general guidelines for performing GWA.

Key words: genome-wide association study, GWA study, Manhattan plot, PLINK

一套關於全基因相關性分析
的標準流程

研究生：陳彥銘　　　　　　　　　指導教授：黃冠華 博士

國立交通大學統計學研究所

摘要

有越來越多的證據顯示，全基因相關性研究在找出與常見人類疾病相關的基因是有效的方法。很多研究成功的利用全基因相關性研究找出新的致病位置。然而，目前並沒有一套一致性的分析程序。在這個研究中，我們檢閱目前的全基因相關性研究去整理出一套標準流程並利用 WTCCC 的真實資料去驗證。我們的流程包括四個部分：資料處理、預處理分析、相關性檢定、以及視覺化結果呈現。為了得到疾病與單體核苷酸多樣性的真實關係，我們做了兩個預處理程序，品質控制與母體分層。除此之外，我們可以畫 Q-Q 圖及 Manhattan 圖去視覺化我們的相關性分析結果。在研究的最後，我們成功的(1)確認必要且重要的分析程序(2)確認目前可用來做這些分析的軟體(3)利用 WTCCC 的資料完成了分析(4)提供了執行全基因相關性研究的一般指導方針。

*關鍵字：*全基因相關性研究 、*Manhattan圖*、*PLINK*

誌謝

兩年的研究生生活，真的是學了很多。從一開始懵懵懂懂的考上了研究所，一無所知，完全不知道未來要做什麼、走什麼方向，到後來目標漸漸的明確，成長了不少，不管是從老師們身上，或是從同學身上，學到的遠比自己從書上得到的多。很開心在兩年後的現在，終於把論文完成了。

在此，要先感謝黃冠華老師，每次的meeting都是一次大的收穫。從原本不太了解什麼是SNP，到後來漸漸能夠處理，老師幫的忙真的是很多很多。我的英文不是很好，每次看paper都一知半解的，多虧老師都會犧牲自己的時間再看一次paper，讓我對於paper的內容有更進一步的了解；也多虧老師耐心的申請data，使我的論文能夠有實際資料的加持。在此，在一次感謝黃冠華老師。

其次，所上的師長們、學長姐、及同學們也給予我很大的幫助，不管是課業上、生活上、或是感情上，都使我能夠快樂又充實的過完這兩年。感謝師長們的教導，使我在處理事情上，又多了許多不同的視野。感謝學長姐對於課業或程式上的幫助。感謝同學們在平常的時候，不管是吃飯、玩樂、作業、考試，都能夠一起討論，讓我不必孤軍奮戰，也使得平時的壓力都得以宣洩。

再來，感謝家人及女朋友，不管是在我回家，或者回台中的時候，都能夠讓我有地方放鬆心情，並且傾聽我的牢騷，真的非常感謝。

最後，謹以此篇論文，獻給我的家人、師長及朋友們。

陳彥銘　　　2008.07.02

Contents

Tables and Figures Content

1 Introduction

In genetic epidemiology, a genome-wide association study (GWAS) is an examination of genetic variation across the human genome, designed to identify genetic associations with observable traits, or why some people get a disease or condition. There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. If genetic variations are more frequent in people with the disease, the variations are said to be "associated" with the disease. The associated genetic variations are then considered pointers to the region of the human genome where the disease-causing problem resides. We attempt to construct a standard GWA path flow for those who want to make a GWA study analysis.

By reviewing a series of literatures (The Wellcome Trust Case Control Consortium (2007), Douglas F. Easton, et al (2007), Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research: Richa Saxena, et al (2007)), we identify four main procedures for performing a GWA study; they are data management, preliminary analysis, association testing, and result visualization. After the recruitment, these individual will be hybridized to the Affymetrix 500K chip. For the chip, we can use a standard genotyping algorithm, BRLMM, developed by Affymetrix, to call the genotype from the chip. Another calling algorithm, CHIAMO, developed by Wellcome Trust Case Control Consortium (WTCCC), a collaboration of 24 leading human geneticists, who will analyze thousands of DNA samples from patients suffering with different diseases to identify common genetic variations for each condition, is applied to simultaneously call the genotypes from all individuals. Cross-platform comparison showed CHIAMO to outperform BRLMM by having an error rate under 0.2%, and comparison of 108 duplicate genotypes in WTCCC study data gave a discordance rate of 0.12%. So our data is called by CHIAMO. We take data from WTCCC and convert them by a c++ program to our analysis file format.

For a case-control GWA study, there are two parts for the preliminary process before tests of association. First part is quality control (QC), and the second part is population stratification. For quality control, it can raise the DNA quality and reduce contamination. For population stratification, since some relatedness among samples may be cryptic, we may identify and exclude individuals whose GWA data reveal substantial differences in genetic background, and adjust for residual stratification. After preprocessing, data will be robust and reliable for the tests of association. With this, we can eventually find the true relatedness between SNPs and disease. For quality control, it contains Single Nucleotide Polymorphism (SNPs) call rate, sample call rate, minor allele frequency (M.A.F.) for SNPs, Hardy-Weinberg equilibrium for SNPs, heterozygosity for individuals, and cryptic relatedness for individuals. For population stratification, multidimensional scaling (MDS) and genome control (GC) are used to identify and exclude individuals whose GWA data reveal substantial differences in genetic background, and adjust for residual stratification.

After quality control and population stratification, we take serial tests of association. We take allele-count test and genotype-count test for single SNP and haplotype-based test for multiple SNPs. We can verify our results by previously robustly replicated loci. In addition to show our results by tables, we can make a visualization display for plotting the quantile-quantile (Q-Q) plot and Genome-wide Manhattan plots to see the pattern.

We use a real data, which is collected by WTCCC, to complement the 4 step procedure. This data is collect to study the association between SNP and CAD disease.

2 Literature Review

2.1 Background of genome-wide association study (GWAS)

(http://en.wikipedia.org/wiki/Genome-wide_association_study)

In genetic epidemiology, a genome-wide association study (GWAS) is an examination of genetic variation across the human genome, designed to identify genetic associations with observable traits, such as blood pressure or weight, or why some people get a disease or condition.

The completion of the Human Genome Project in 2003 made it possible to find the genetic contributions to common diseases and analyze whole-genome samples for genetic variations that contribute to their onset.

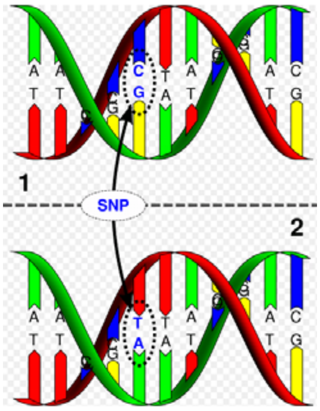
These studies require two groups of participants: people with the disease and similar people without. After obtaining samples from each participant, the set of markers such as SNPs are scanned into computers. The computers survey each participant's genome for markers of genetic variation.

If genetic variations are more frequent in people with the disease, the variations are said to be "associated" with the disease. The associated genetic variations are then considered pointers to the region of the human genome where the disease-causing problem resides.

2.1.1 Single nucleotide polymorphism (SNP)

(http://en.wikipedia.org/wiki/Single_nucleotide_polymorphism)

A single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from



different individuals, AAGC***C***TA to AAGC***T***TA, contain a difference in a single nucleotide. In this case we say that there are two alleles: C and T. Almost all common SNPs have only two alleles. For a variation to be considered a SNP, it must occur in at least 1% of the population.

Variations in the DNA sequences of humans can affect how humans develop diseases and respond to pathogens, chemicals, drugs, etc. Technologies from Affymetrix and Illumina allow for genotyping hundreds of thousands of SNPs for typically under $1,000.00 in a couple of days.

2.1.2 Analysis Software- PLINK, R, and Haploview

PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml)

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

The focus of PLINK is purely on analysis of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype calls from raw data).

PLINK (one syllable) is being developed by Shaun Purcell at the Center for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard & MIT, with the support of others.

HaploView (http://www.broad.mit.edu/mpg/haploview/)

Haploview is designed to simplify and expedite the process of haplotype analysis by providing a common interface to several tasks relating to such analyses. Haploview currently supports the following functionalities:

- LD & haplotype block analysis
- haplotype population frequency estimation

- single SNP and haplotype association tests
- permutation testing for association significance
- implementation of Paul de Bakker's Tagger tag SNP selection algorithm.
- automatic download of phased genotype data from HapMap
- visualization and plotting of PLINK whole genome association results including advanced filtering options

Haploview is fully compatible with data dumps from the HapMap project and the Perlegen Genotype Browser. It can analyze thousands of SNPs (tens of thousands in command line mode) in thousands of individuals.

R (http://www.r-project.org/)

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. It is convenient for statistic analysis with R. It contains distributions, tests, plots and other about statistic.

2.2 Wellcome Trust Case Control Consortium (WTCCC)

(http://www.wtccc.org.uk/)

The Wellcome Trust Case Control Consortium (WTCCC) is a collaboration of 24 leading human geneticists, who will analyze thousands of DNA samples from patients suffering with different diseases to identify common genetic variations for each condition. It is hoped that by identifying these genetic signposts, researchers will be able to understand which people are most at risk, and also produce more effective treatments.

The WTCCC has now searched for the genetic variation associated with tuberculosis, coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension. The research was conducted at a number of institutes throughout the UK, including the Wellcome Trust Sanger Institute, Cambridge

University and Oxford University.

Researchers will have analyzed over 19,000 DNA samples - two thousand patients for each disease and three thousand control samples - searching for important genetic differences between people who do and don't have each disease.

2.2.1 CHIAMO vs. BRLMM

CHIAMO is a program for calling genotypes from the Affymetrix 500K Mapping chip. The program allows for multiple cohorts which have potentially different intensity characteristics that can lead to elevated false-positive rates in genome-wide studies. The underlying model has a hierarchical structure that allows for correlation between the parameters of each cohort. CHIAMO is developed by WTCCC to replace BRLMM, the standard genotype calling algorithm, to calling genotype accuracy. The large number of misclassification will reduce the power of analysis (See Figure 1).

2.3 Study Population

The CAD individuals recruited by WTCCC and control from 1958 British Birth Cohort (58C) and UK Blood Services (UKBS) are hybridized to Affymetrix 500K chip subsequently. They are living within England, Scotland and Wales ('Great Britain') and the vast majority had self-identified themselves as white Europeans.The standard algorithm BRLMM developed by Affymetrix is used to called the genotype from the chip. WTCCC developed another algorithm, CHIAMO, to call the chip. Cross-platform comparison showed CHIAMO to outperform BRLMM by having an error rate under 0.2%, and comparison of 108 duplicate genotypes in WTCCC study data gave a discordance rate of 0.12%. Our data is called by CHIAMO.

2.3.1 Data conversion

When we get the genotype raw data, we convert the data by a c++ program to our data format, pedigree and map format. Since the number of SNP is large, we can't convert them all at a time. We convert the data chromosome by chromosome and merge them by software PLINK to single one file. For the .ped file, the pedigree format file, each row represent a individual and each column represent Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype, SNP1, SNP2, ... in turn. For the .map file, the map format file, each row represent a SNP and each column represent chromosome, rs# or SNP identifier, genetic distance, base-pair position in turn.

2.4 Quality control

For a raw data, if we analysis it directly without remove low DNA quality SNPs or individuals, then the analysis results won't be robust and reliable. For a restrained analysis, we must do the following procedures. For each SNP, we check for call rate (or missingness), minor allele frequency (M.A.F.), and Hardy-Weinberg disequilibrium. For each individual, we check for call rate (or missingness), heterozygosity, and cryptic relatedness.

2.4.1 Call rate

For each individual, call rate is the proportion of non-missing SNPs per sample. For each SNP, call rate is the proportion of non-missing data over all samples. The missing data rate per sample acts as an indicator of low DNA quality.

2.4.2 Minor allele frequency (M.A.F.)

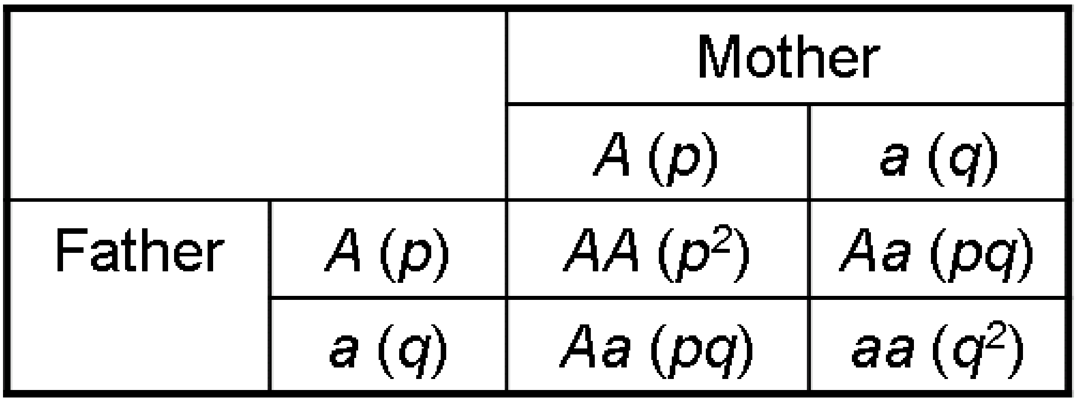
For introducing the minor allele frequency (M.A.F.), we may speak of allele frequency first. For each SNP, there are two alleles. For each allele, allele frequency is the proportion of this allele in this SNP over all samples. And for each SNP, the smaller one allele frequency is called minor allele frequency (M.A.F.).

2.4.3 Heterozygosity

For introducing the heterozygosity, we may speak of homozygous and heterozygous. For a SNP, each allele may be P or p, so the genotype is PP, Pp or pp. Homozygous represent genotype PP or pp, and heterozygous represent genotype Pp. For each individual, heterozygosity is the proportion of SNPs that are heterozygous or are a heterozygote (i.e., SNPs with different alleles in the homologous chromosome pair) among all typed SNPs. Excess heterozygosity may indicate contamination. Low heterozygosity can result in the lack of the mechanism that maintains polymorphism and helps to explain some kinds of genetic variability. For each SNP, heterozygosity $=1-\sum_{i=1}^{n} p_i^2$, where p_i is the frequency of the ith allele, and n=2 is the total number of alleles. The higher the value, the more polymorphic the SNP is.

2.4.4 Hardy-Weinberg equilibrium (HWE)

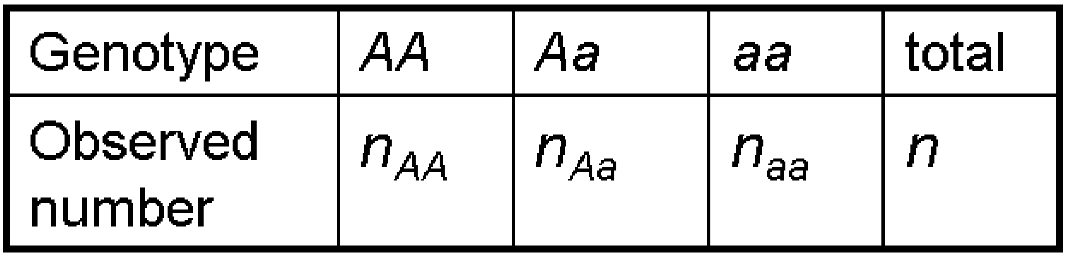
For combined control samples, we check the Hardy-Weinberg equilibrium. HWE holds at a locus in a population when the two alleles are not statistically associated. Deviations from HWE can be due to inbreeding, population stratification, selection, deletion polymorphism, or a segmental duplication that could be important in disease causation. So far, researchers have tested for HWE primarily as a data quality check and have discarded SNPs, for example, deviate from HWE among controls at certain significance level α (e.g., $=10^{-3} or 10^{-4}$). For HWE testing, suppose that parents have the following inheritable rule for passing their features to their offspring.



		Mother	
		$A\ (p)$	$a\ (q)$
Father	$A\ (p)$	$AA\ (p^2)$	$Aa\ (pq)$
	$a\ (q)$	$Aa\ (pq)$	$aa\ (q^2)$

The final three possible genotypic frequencies in the offspring become: $\Pr(AA) = p^2, \Pr(Aa) = 2pq, \Pr(aa) = q^2$.

For n samples, if the frequency of the observed genotype is the following.



Genotype	AA	Aa	aa	total
Observed number	n_{AA}	n_{Aa}	n_{aa}	n

From which allele frequencies can be estimated as:

$$\widehat{p} = \frac{2n_{AA} + n_{Aa}}{2(n_{AA} + n_{Aa} + n_{aa})}, \quad \widehat{q} = 1 - \widehat{p}$$

So the Hardy-Weinberg expectation is:

$$E(AA) = \widehat{p}^2 n, E(Aa) = 2\widehat{p}\widehat{q}n, E(aa) = \widehat{q}^2 n$$

So the Pearson's chi-square test statistic is:

$$\chi^2 = \sum_{AA,Aa,aa} \frac{(O-E)^2}{E} \overset{H_o:HWE}{\sim} \chi^2 \quad with \quad d.f. = 1$$

When there are low genotype count, and it is better to use a Fisher exact test.

2.4.5 Cryptic relatedness

For identify cryptic relatedness, first we may speak of identical-by-state (IBS). The

IBS is sum of the number of IBS alleles at each locus divided by twice the number of loci. For example, two unrelated individuals each with blood group AB share two alleles IBS. There is Evidence that, despite allowance for known family relationships, individuals in the study sample have residual, non-trivial degrees of relatedness, which can violate the independence assumptions of standard statistical techniques. So we select a set of SNPs, within which no pair was correlated with $r^2 > 0.2$. This can be done by compute pairwise r^2 for 50 SNPs each other per time and delete SNPs until no one pairwise $r^2 > 0.2$ and shift 5 SNPs for the next time and go on. Note that two SNPs with different chromosome will not be computed. For this set of nearly independent SNPs, we computed genome-wide average identity by state (IBS) between each pair of individuals. Individuals with too much IBS sharing will be exclude, likely duplicates (IBS>99%) or relatives (IBS>86%).

2.5 Population stratification

The presence of population stratification may result by different ancestral and demographic histories in the study samples. If cases and controls differ with respect to these features, markers that are informative for them might be confounded with disease status and lead to spurious associations. Cryptic population structure that is not recognized by investigators is potentially more problematic. If there is population stratification to exist, we may identify and exclude individuals whose GWA data reveal substantial differences in genetic background, and adjust for residual stratification.

2.5.1 Genome control (GC)

For genome control, recall that the Armitage-test statistics $\chi_G^2 \sim \chi^2(1)$ under H_0 (a test for the single SNP association, will discuss later). At the first, we require a number (preferably >100) of widely spaced null SNPs (i.e., it is unlikely that any one SNP is tightly linked to a disease-susceptibility gene) that have been genotyped in cases and controls in addition to the candidate SNPs. When there exists population stratification, χ_G^2 will no longer follow a chi-squared distribution under H0, but instead follow a scaled chi-squared distribution, i.e., $\chi_G^2 \sim \lambda\chi^2(1)$ under H_0, where λ is a constant termed "variance inflation factor". The estimation of variance inflation factor can be made by the following step: Genotype a number of null SNPs, and then the following can serve as an estimate ofλ. For theλ, there are two ways to estimate it. One is compute the mean of the Armitage-test statistics across these null SNPs (recall that $E\left\{\chi^2(1)\right\}=1$). And another one is to find the median of the Armitage-test statistics across these null SNPs, divided by the predicted median for the $\chi^2(1)$ distributions (i.e., = 0.456). Estimated variance inflation factor λ by median is robust than mean. In a GWA study, it is difficult to identify the null SNPs. However, because the bulk of the tested loci in a GWA will naturally be null, it provided that a robust estimator is chosen. Therefore, the SNPs used in making estimates of λ are those that pass the filter for quality control. Then the Armitage test is applied at the candidate SNPs, and if $\hat{\lambda} > 1$ the test statistics are divided by $\hat{\lambda}$ (i.e., $\chi_G^2 / \hat{\lambda}$).

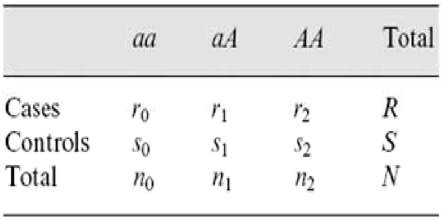
2.5.2 Multidimensional scaling (MDS)

We use MDS for detecting individuals with different ancestry. MDS is a set of related statistical techniques often used in data visualisation for exploring similarities or dissimilarities in data. An MDS algorithm starts with a matrix of item-item similarities, and then assigns a location of each item in a low-dimensional space, suitable for graphing or 3D visualisation. For detecting individuals with different ancestry using MDS, at the first we select a set of SNPs, within which no pair were correlated with $r^2 > 0.2$. This can be done by compute pairwise r^2 for 50 SNPs each other per time and delete SNPs until no one pairwise $r^2 > 0.2$ and shift 5 SNPs for the next time and go on. Note that two SNPs with different chromosome will not be computed. For this set of nearly independent SNPs, we computed genome-wide average identity by state (IBS) between each pair of individuals along with the 270 HapMap samples. Convert these IBS-relationships to distances by subtracting them from 1, and the matrix is used as input to MDS. The projection onto the two multi-dimensional scaling axes is shown. Since the 270 HapMap samples are composed of three races, so we can clearly identify those who with different cryptic ancestral. When we make Armitage-test, the statistic $\chi_G^2 \sim \chi^2(1)$ under H_0.When there exist population stratification, χ_G^2 will no longer follow a chi-squared distribution under H_0, but instead follow a scaled chi-squared distribution, i.e. $\chi_G^2 \sim \lambda\chi^2(1)$ under H_0, where λ is a constant termed "variance inflation factor". For estimate of λ, both the mean of the Armitage-test statistics across these null SNPs (recall that $E\{\chi^2(1)\} = 1$ and the median of the Armitage-test statistics across these null SNPs, divided by the predicted median for the $\chi^2(1)$ distributions (i.e., = 0.456) can make it, and the later is more robust. And then the statistic can be divided by λ.

2.6 Tests of association

There are several tests for association between SNP and disease. For single SNP, we take genotype-count and allele-count tests. Analyzing SNPs one at a time can neglect information in their joint distribution, so we take the multiple SNPs test. For multiple SNPs, haplotype-based method is used.

2.6.1 Genotype-count test



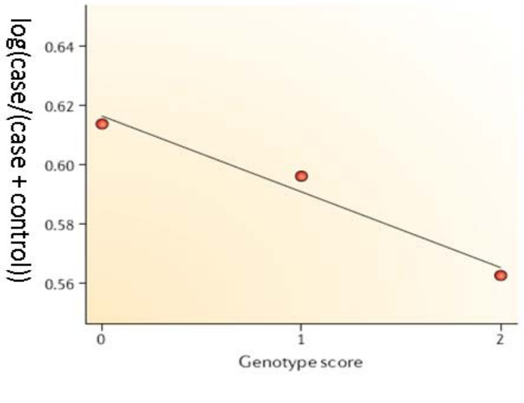
	aa	aA	AA	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

For genotype-count test, the most natural analysis of SNP genotypes and case-control status at a single SNP is to test the null hypothesis of no association between rows and columns of the 2 × 3 matrix that contains the counts of the three genotypes (the two homozygotes and the heterozygote) among cases and controls. Users have a choice between, among others, a Pearson test (2 d.f.) or a Fisher exact test. For lower count number, Fisher exact test performs better. If we consider a dominant / recessive model, if A is dominant, one can assign genotypes (aa,aA,AA) with score x=(0,1,1), and then test for association between case-control status and x; if A is recessive, one can assign genotypes (aa,aA,AA) with score x=(0,0,1), and then test for association between case-control status and x. And we can make the test similar to genotype-count test for dominant / recessive model.

2.6.2 Allele-count test

For complex traits, it is widely thought that contributions to disease risk from individual SNPs will often be roughly additive — that is, the heterozygote risk will be intermediate between the two homozygote risks. One way to improve power to detect additive risks is to count alleles rather than genotypes so that each individual contributes twice to a 2 × 2 table and a Pearson 1-df test can be applied. However, this procedure is not recommended because it requires an assumption of HWE in cases and controls combined and does not lead to interpretable risk estimates. The Cochran-Armitage test (also known as just the Armitage test and called within R the proportion trend test) is similar to the allele-count test. It is more conservative and does not rely on an assumption of HWE.

The dots indicate the proportion of



cases, among cases and controls combined, at each of three SNP genotypes (coded as 0, 1 and 2), together with their least-squares line. The Armitage test corresponds to testing the hypothesis that the line has zero slopes.

The Cochran-Armitage test: $\chi_G^2 = \dfrac{N\left(N\sum r_i x_i - R\sum n_i x_i\right)^2}{R\left(N-R\right)\left\{N\sum n_i x_i^2 - \left(\sum n_i x_i\right)^2\right\}}$

, where $\mathrm{x}_i = \mathrm{i},\ \mathrm{i} = 0,1,2$ under H$_0$: no association, $\chi_G^2 \sim \chi^2(d.f.=1)$

The Cochran-Armitage test is equivalent to the score test for testing $H_0:\beta=0$ in the logistic regression model (assuming additive effect)

$\Pr\left(D=d \mid x_i;\beta_0,\beta\right)=\dfrac{\exp d\left(\beta_0+\beta x_i\right)}{1+\exp\left(\beta_0+\beta x_i\right)}$, where $\mathrm{d}=0(control),\ 1(\mathrm{case})$; x_i=i, i=0, 1, 2

There is no generally accepted answer to the question of which single-SNP test to use. An intermediate choice is to take the maximum test statistic from those designed for additive, dominant or recessive effects.

2.6.3 Haplotype-based method

A popular strategy, suggested by the block like structure of the human genome, is to use haplotypes to try to capture the correlation structure of SNPs in regions of little recombination. This approach can lead to analyses with fewer degrees of freedom, but this benefit is minimized when SNPs are ascertained through a tagging strategy. Perhaps more importantly, haplotypes can capture the combined effects of tightly linked cis-acting causal variants.

An immediate problem is that haplotypes are not observed; instead, they must be inferred. It can be hard to account for the uncertainty that arises in phase inference when assessing the overall significance of any finding. However, when LD between markers is high, the level of uncertainty is usually low. Given haplotype assignments, the simplest analysis involves testing for independence of rows and columns in a 2 × k contingency table, where k denotes the number of distinct haplotypes. Alternative approaches can be based on the estimated haplotype proportions among cases and controls, without an explicit haplotype assignment for individuals (Schaid 2004).

One problem with both these approaches is reliance on assumptions of HWE and of near-additive disease risk. Including rare haplotypes in analyses can lead to loss of power because there are too many degrees of freedom. One common but unsatisfactory solution is to combine all haplotypes that are rare among controls into a “dustbin” category.

Another problem with defining haplotypes is that block boundaries can vary according to the population sampled, the sample size, the SNP density and the block definition. In software Haploview, the haplotype block can be defined by the software and we can take them to make test of association. But in software PLINK; we may compute all block size 2, 3, 4, 5 haplotype for making test of association.

2.7 Visualization display and previous evidence

After all procedures are finished, in addition to show our result by tables, we can even make a visualization display for plotting the quantile-quantile (Q-Q) plot and Genome-wide Manhattan plots. Q-Q plot provide a visual summary of the distribution of the observed test statistics generated by a GWA study. Typically, a single test statistic (for case–control studies, a chi-squared (χ^2) comparison of absolute genotype counts) is calculated for each variant passing quality control. And Manhattan plots display GWA findings with respect to their genomic positions, highlighting signals of particular interest. This can help us to see the pattern of our result clearly. We can construct a table for previous robustly replicated loci and verify our analysis.

3 Method

3.1 Review of 3 GWA studies

3.1.1 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

WTCCC made GWA studies on British population by 2,000 individuals for each of 7 major diseases and a shared set of 3,000 controls. These diseases are bipolar disorder (BD), coronary artery disease (CAD), Crohn’s disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). They are human diseases of major public health importance. Controls are composed of 1958 Birth Cohort Controls (58BC) and UK Blood Services Controls (UKBS). People in study were living within England, Scotland and Wales (‘Great Britain’) and the vast majority had self-identified themselves as white Europeans. These individuals were genotyped 500568 SNPs.

They found it necessary to normalize the Affymetrix probe intensity data to minimize chip-to-chip variability. A C++ program, CHIAMO, was written to carry out this normalization efficiently. CHIAMO is a new genotype calling algorithm, implemented in C++. It uses a hierarchical statistical model, which allows it to simultaneously call genotypes at all data samples.

For these 500,568 SNPs and 17000 individuals, they do quality control filter for (1) SNP call rate < 97% (missingness), (2) Heterozygosity > 30% or < 23% across all SNPs, (3) External discordance with genotype or phenotype data, (4) Individuals identified as having recent non-European ancestry by the Multidimensional Scaling analysis, (5) Duplicates, and (6) Individuals with too much IBS sharing (>86%); likely relatives. There are 16179 individuals and 469557 SNPs (93.8%) pass QC filter.

For the remaining SNPs and individuals, they do association assess for classical and bayesian statistical approaches. They performed trend tests (1 degree of freedom) and general genotype tests (2 degrees of freedom) between each case collection and the pooled controls, and calculated analogous Bayes factors. They also do sex-differentiated test. Sex-differentiated test is sensitive to associations of a different magnitude and/or direction in the two sexes. They also did the combined diseases association test with potential aetiological overlap, and multilocus method by simulate, or impute, genotype data at 2,193,483 HapMap SNPs not on the Affymetrix chip and then tested for association. For test of association, they used snpMatrix and SNPTEST. Both quantitative and qualitative phenotypes can be analyzed using snpMatrix and flexible association testing functions are provided that control for potential confounding by quantitative and qualitative covariates. SNPTEST is a standalone C++ program that implements both frequentist tests and bayesian analysis of association and allows the user to include quantitative or qualitative covariates.

3.1.2 Genome-wide association study identifies novel breast cancer susceptibility loci

Breast cancer is about twice as common in the first-degree relatives of women with the disease as in the general population. In 1990s, two major susceptibility genes for cancer, BRCA1 and BRCA2, were identified. Large case-control association studies have identified variants in the DNA repair genes CHEK2, ATM, BRIP1 and PALB2 that confer an approximately twofold risk of breast cancer, but these variants are rare in the population. A recent study has shown that a common coding variant in CASP8 is associated with a moderate reduction in breast cancer risk. After accounting for all the known breast cancer loci, more than 75% of the familial risk of the disease remains unexplained. They perform a three-stage association study.

At stage Ⅰ, they recruited 408 cases (family history score ≥ 2, diagnosed under age 60, excluded BRCA1 & BRCA2 cases) and 400 controls (age ≥50, free of cancer at entry) and genotyped 266722 SNPs (m.a.f. ≥ 5%). SNPs and individuals were check if call rate ≤

80%. For first stage, there were 390 cases and 364 controls, 227876 SNPs left. These SNPs did the stage Ⅰ filter, (1) call rate ≤ 90%, (2) HWE with p-value < 10^{-5}, and there were 205568 SNPs left. From stage Ⅰ, 12711 (about 5%) SNPs selected on the basis of significance of the difference in genotype frequency between cases and controls (P-trend < 0.052 or weighted P-trend < 0.01 or P < 0.01 under dominant/recessive model), then genotyped in 3990 cases and 3916 controls from the SEARCH study, using a custom-designed oligonucleotide array. SNPs were check if call rate ≤ 80% and filter. These remain SNPs did the stage Ⅱ filter, (1) call rate ≤ 95 %, (2) HWE with p-value < 10^{-5}, and there were 10405 SNPs left. For stage Ⅲ, 22714 cases of invasive breast cancer and 1020 cases of carcinoma in situ (CIS) and 23369 controls from 22 case-control study are collected. These individuals were genotyped on 10405 SNPs and check if call rate ≤ 80% and there were 21860 cases of invasive breast cancer and 988 cases of carcinoma in situ (CIS) and 22578 controls left for stage Ⅲ. They tested 31 of the most significant SNPs (P trend of P(2d.f.)< 0.00002) on these individuals. Those test statistics for stage Ⅰ and stage Ⅱ were adjusted by genome control.

For tests of association, they performed Cochran-Armitage trend test (1 degree of freedom) for single SNP and stratified Cochran-Armitage trend test (1 degrees of freedom) where stage 1 was given a weight of 4 for stage1+2 combined SNPs. For stage 3, each study was treated as a separate stratum. P-value < 10^{-7} level has been proposed as appropriate for genome-wide studies. And they performed fine-scale mapping for the region significance SNP located by tag SNPs which r^2>c, then use Haplotype analysis to find the possible causable allele. For significance SNPs, perform a multiple logistic regression analysis of these variates to find the odds ratios and confidence intervals. And the databases are from dbSNP, HapMAp, Perlegen.

3.1.3 Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and

Triglyceride Levels

Type 2 diabetes, obesity, and cardiovascular risk factors are caused by a combination of genetic susceptibility, environment, behavior, and chance. Whole-genome association studies (WGAS) offer a new approach to gene discovery unbiased with regard to presumed functions or locations of causal variants. New strategies for prevention and treatment of type 2 diabetes (T2D) require improved insight into disease etiology. Patients with T2D, geographically matched controls, and discordant sib-ships were selected from Finland and Sweden. To avoid admixture with type 1 diabetes, patients had an age at onset > 35 years and no detectable glutamic acid decarboxylase antibodies (GAD Ab). Members from families with carriers of mutations causing maturity onset diabetes of the young (MODY; HNF4A, GCK, TCF1, IPF1, TCF2) were excluded, except for Skara where no screening for MODY mutations had been performed. Control subjects were defined as normal glucose tolerant. They recruited 1,022 cases and 1,075 controls for unrelated matched population and 442 cases and 392 controls for discordant sib-ships data. And 10,850 individuals (European ancestry) were used to replicate original T2D findings in this study. These samples were genotyped 500,568 SNPs.

For individual inclusion criteria, they do the following check: (1) passing the fingerprint quality checks, (2) Genotyping call rates ≥ 95%, (3) Gender call from X chromosome genotype data was discrepant with the gender obtained from medical records were excluded from the analysis, and (4) in order to verify the existing known familial relationships identity-by-descent (IBD) analysis was performed using the PLINK analysis software package. After these processes, there are 2,931 individuals left. For SNP quality control, they do the following check: (1) did not map to multiple locations in the genome (3,605 markers excluded), (2) showed a >95% genotype call rate (34,532 markers excluded) and a >90% genotype call rate in both population and familial subsets of data (229 markers excluded), (3) MAF >1% 2,931 individuals (66,787 markers excluded) and >1% in both population and familial subsets of the data (2,909 markers excluded), and (4) demonstrated Hardy Weinberg equilibrium with a P > 10^{-6} in controls (5,775 markers excluded). After these processes, there are 386,731 SNPs left. EIGENSTRAT was used to evaluate population structure in the samples. And they also adjusted population structure by using genome control to estimate a genomic inflation factor based on the median chi-squared test in the matched population-based case/control sample.

To extend the set of putative causal alleles tested for association, we developed 284,968 additional multimarker (haplotype) tests based on these SNP genotypes. The 671,699 allelic tests capture (correlation coefficient r2 ≥ 0.8) 78% of common SNPs in HapMap CEU. Each SNP and haplotype test was assessed for association to T2D and each of 18 traits with the software package PLINK. For T2D, a weighted meta-analysis was used to combine results for the population-based and family-based subsamples. For quantitative traits, multivariable linear or logistic regression with or without covariates was performed. To perform association testing in the populaton sample, we performed a Cochran-Mantel-Haenszel (CMH) stratified test. To perform association testing in the familial sample, we used the DFAM procedure in PLINK.

For replication data, 107 SNPs was tested. SNPs were tested for association using a simple Chi-square analysis in each of the three T2D replication samples. Combined analyses of replication samples or of all DGI samples was performed using Mantel Haenzel meta-analysis of the odds ratio. For this study, they use EIGENSTRAT to evaluate population structure and PLINK to test association.

3.2 Summarized Procedures for a GWA Study

For above three literatures, we arrange the procedures for a GWA study. Our procedure includes four steps: data management, preliminary analysis, association testing and result visualization.

3.2.1 Data Management

3.2.1.1 Genotype Calling

Instead of use BRLMM, WTCCC develop a new algorithm, CHIAMO to call the Signal intensity on the raw chip and turn it into genotype data. CHIAMO can simultaneously call the genotypes from all individuals. Cross-platform comparison showed CHIAMO to outperform BRLMM by having an error rate under 0.2%, and comparison of 108 duplicate genotypes in WTCCC study data gave a discordance rate of 0.12%.

3.2.1.2 Data Conversion

CAD case control genotype data obtained from WTCCC were converted by a c++ program to our study format, pedigree and map format (See Figure2, Figure 3). Since the number of SNP is large, we can't convert them all at a time. We convert the data chromosome by chromosome and merge them by software PLINK to single one file. For the .ped file, the pedigree format file, each row represent a individual and each column represent Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype, SNP1, SNP2, ... in turn. For the .map file, the map format file, each row represents a SNP and each column represent chromosome, rs# or SNP identifier, genetic distance, and base-pair position in turn.

3.2.2 Quality Control

There are several steps for data quality control. For each individual, we may check for call rate (or missingness), heterozygosity, and cryptic relatedness. For each Single Nucleotide Polymorphism (SNP), we may check for call rate (or missingness), minor allele frequency (M.A.F.), and Hardy-Weinberg disequilibrium.

3.2.2.1 Call rate

For each individual, sample call rate is the proportion of non-missing SNPs per sample. We exclude individuals with Call rate ≤ 97% (or missingness ≥ 3%). For each SNP, SNP call rate is the proportion of non-missing data over all samples. We exclude individuals with Call rate ≤ 95% (or missingness ≥ 5%).

3.2.2.2 Minor allele frequency (M.A.F.)

For each SNP, the smaller one allele frequency is called minor allele frequency (M.A.F.). We exclude SNPs with M.A.F. < 1%, and we exclude 68444 SNPs from our data.

3.2.2.3 Heterozygosity

For each individual, genome-wide heterozygosity is the proportion of SNPs that are heterozygous or are a heterozygote (i.e., SNPs with different alleles in the homologous chromosome pair) among all typed SNPs. For each SNP, heterozygosity $=1-\sum_{i=1}^{n} p_i^2$ where p_i is the frequency of the ith allele, and n=2 is the total number of alleles. We exclude SNPs with genome-wide heterozygosity ≤ 30% or genome-wide heterozygosity ≥35%.

3.2.2.4 Hardy-Weinberg equilibrium

For combined control samples, we check the Hardy-Weinberg equilibrium (HWE). HWE holds at a locus in a population when the two alleles are not statistically associated. When there are low genotype count, and it is better to use a Fisher exact test. We exclude SNPs with HWE testing p-value threshold $5.7*10^{-7}$.

3.2.2.5 Cryptic relatedness

The IBS is sum of the number of IBS alleles at each locus divided by twice the number of loci. We select a set of SNPs, within which no pair were correlated with $r^2 > 0.2$. This can be done by compute pairwise r^2 for 50 SNPs each other per time and delete SNPs until no one pairwise $r^2 > 0.2$ and shift 5 SNPs for the next time and go on. For this set of nearly independent SNPs, we computed genome-wide average identity by state (IBS) between each pair of individuals. Individuals with too much IBS sharing will be exclude, likely duplicates (IBS>99%) or relatives (IBS 86-99%).

3.2.3 Population stratification

For the SNPs passing quality control, they are check for population stratification. Population stratification is the presence in study samples of individuals with different ancestral and demographic histories. If cases and controls differ with respect to these features, markers that are informative for them might be confounded with disease status and lead to spurious associations. We should identify and exclude individuals whose GWA data reveal substantial differences in genetic background, and adjust for residual stratification. We attempt to identify population stratification by genome control (GC) and multidimensional scaling (MDS).

3.2.3.1 Genome control (GC)

The SNPs used in making estimates of λ are those that pass the filter for quality control. We can estimate λ by find the median of the Armitage-test statistics across these null SNPs, divided by the predicted median for the $\chi^2(1)$ distributions (i.e., = 0.456). The adjusted test statistics are divided by $\hat{\lambda}$.

3.2.3.2 Multidimensional scaling (MDS)

At the first we select a set of SNPs, within which no pair were correlated with $r^2 > 0.2$. This can be done by compute pairwise r^2 for 50 SNPs each other per time and delete SNPs until no one pairwise $r^2 > 0.2$ and shift 5 SNPs for the next time and go on. For this set of nearly independent SNPs, we computed genome-wide average identity by state (IBS) between each pair of individuals along with the 270 HapMap samples. Convert these IBS-relationships to distances by subtracting them from 1, and the matrix is used as input to MDS. The projection onto the two multi-dimensional scaling axes is shown.

3.2.4 Test of association

We check the association by test for single SNP and multiple SNPs. For single SNP, we test the association for using genotype-count test and allele-count test. For multiple SNPs, we just talk about the haplotype-based method.

3.2.4.1 Single SNP

Genotype count test

The most natural analysis of SNP genotypes and case-control status at a single SNP is to test the null hypothesis of no association between rows and columns of the 2 × 3 matrix that contains the counts of the three genotypes (the two homozygotes and the heterozygote) among cases and controls. Users have a choice between, among others, a Pearson test (2 d.f.) or a Fisher exact test.

Dominant / recessive model

If we consider a dominant / recessive model, if A is dominant, one can assign genotypes (aa,aA,AA) with score x=(0,1,1), and then test for association between case-control status and x; if A is recessive, one can assign genotypes (aa,aA,AA) with score x=(0,0,1), and then test for association between case-control status and x. And we can make the test similar to genotype-count test for dominant / recessive model.

Allele count test

We count alleles rather than genotypes so that each individual contributes twice to a 2 × 2 table and a Pearson 1-df test can be applied.

Cochran-Armitage test

The Cochran-Armitage test: $\chi_G^2 = \dfrac{N\left(N\sum r_i x_i - R\sum n_i x_i\right)^2}{R\left(N-R\right)\left\{N\sum n_i x_i^2 - \left(\sum n_i x_i\right)^2\right\}}$

where $\mathrm{x}_i = \mathrm{i},\ \mathrm{i} = 0,1,2$ Under H_0 : no association, $\chi_G^2 \sim \chi^2(d.f.=1)$

There is no generally accepted answer to the question of which single-SNP test to use. An intermediate choice is to take the maximum test statistic from those designed for additive, dominant or recessive effects.

3.2.4.2 Multiple SNPs

The procedure of haplotype-based method is that given haplotype assignments, the simplest analysis involves testing for independence of rows and columns in a 2 × k contingency table, where k denotes the number of distinct haplotypes. Alternative approaches can be based on the estimated haplotype proportions among cases and controls, without an explicit haplotype assignment for individuals (Schaid 2004).

We can define the haplotype block by the software Haploview and take them to make test of association. But this software has SNP number limitation. So we can perform the haplotype tests in software PLINK. For PLINK, we can’t define the haplotype block; instead, we take all block size 2, 3, 4, 5 haplotypes for making test of association. For example, a size 2 haplotype for two SNPs with genotype Aa, Pp may be AP, aP, Ap, and ap. And we take all 4 haplotype for haplotype test.

3.2.5 Visualization display and previous evidence

We can construct a table for previous robustly replicated loci and verify our analysis. In addition, we can even make a visualization display for plotting the quantile-quantile (Q-Q) plot and Genome-wide Manhattan plots. Q-Q plot provide a visual summary of the distribution of the observed test statistics generated by a GWA study. And Manhattan plots display GWA findings with respect to their genomic positions, highlighting signals of particular interest. This can help us to see the pattern of our result clearly. Q-Q can be done by software R, and Manhattan plot can be done by software haploview.

4 Analysis of the CAD data from WTCCC

4.1 Study Population

1988 CAD individuals included in the Wellcome Trust Case Control Consortium (WTCCC) study were living within England, Scotland and Wales (‘Great Britain’) and the vast majority had self-identified themselves as white Europeans. Coronary Artery Disease (CAD) is common familial disease of major public health importance both in the UK and globally, and for which suitable nationally representative sample sets were available. The control individuals came from two sources: 1,504 individuals from the 1958 British Birth Cohort (58C) and 1,500 individuals selected from blood donors recruited as part of this project (UK Blood Services (UKBS) controls). All 4992 samples were genotyped with the GeneChip 500K Mapping Array Set (Affymetrix 500K chip), which comprises 490032 SNPs.

4.2 Data management

Instead of use BRLMM, WTCCC develop a new algorithm, CHIAMO to call the Signal intensity on the raw chip and turn it into genotype data. CHIAMO can simultaneously call the genotypes from all individuals. Cross-platform comparison showed CHIAMO to outperform BRLMM by having an error rate under 0.2%, and comparison of 108 duplicate genotypes in WTCCC study data gave a discordance rate of 0.12%. CAD case control genotype data obtained from WTCCC were converted by a c++ program to our study format, pedigree and map format (See Figure2, Figure 3). For the .ped file, the pedigree format file, each row represent a individual and each column represent Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype, SNP1, SNP2, ... in turn. For the .map file, the map format file, each row represents a SNP and each column represent chromosome, rs# or SNP identifier, genetic distance, and base-pair position in turn.

4.3 Quality control

For sample call rate we exclude individual if call rate ≤ 97%, and we exclude 0 individual since our data may be imputed. For SNP call rate we exclude SNP if call rate ≤ 95%, and we exclude 0 SNP since our data may be imputed.

For Genome-wide heterozygosity we exclude individual if genome-wide heterozygosity > 35% or < 30%, and we exclude 29 individuals from the analysis (See Figure 4).

For cryptic relatedness we sieve out 82,686 SNPs for calculating the genome-wide average IBS, then we exclude individual with IBS > 99% (duplicated) or IBS 86-99% (relatives). Finally, we exclude 16 (duplicated) + 43 (relatives) individuals from the analysis.

For Minor allele frequency (M.A.F.) we exclude SNP with M.A.F. < 1% and we exclude 68,444 SNPs from the analysis.

For Hardy-Weinberg equilibrium (HWE) we exclude SNP for HWE testing with p-value threshold $5.7*10^{-7}$ and we exclude 5,915 SNPs from the analysis.

For quality control, we show it by figure (See Figure 5) for a clear display. It shows the criteria for each check and exclusion number for this check.

4.4 Population stratification

For the population stratification, we compute $\lambda = 1.087280702$ by the median of Armitage-test statistics. This can be used to adjust the Armitage test statistic for the deviation of null hypothesis.

We sieve out 82,686 nearly independent SNPs for Multidimensional scaling (MDS), and plot MDS for the first two dimensions. For a threshold for first dimension axes value > 0.5, we exclude 23 individuals from MDS (See Figure 6).

So far, we exclude 111 individuals and 74359 SNPs from quality control and MDS from our analysis data, and we have 4881 individuals and 415673 SNPs left.

4.5 Tests of Association

For single SNP, we do the tests of allele count test, genotype count test, dominant and recessive model, and Armitage trend test. For multiple SNPs, we do the haplotype-based tests for haplotype block size 2, 3, 4, and 5. For these tests, we set a p-value threshold by $5*10^{-7}$. The results are showed by Q-Q plots and Manhattan plots. We also show the result by table for previous robust replication loci. (See Figure 7 — Figure 14, Table. 1).

5 Conclusion and Discussion

5.1 Conclusion

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We attempt to construct a standard GWA flow path. First stage, we may arrange our data to our analysis format. Then we do the quality control to ensure high DNA quality. The third stage is to identify and exclude individuals whose GWA data reveal substantial differences in genetic background, and adjust for residual stratification. We can do genome control and MDS for this stage. The data passing our preliminary processes can be used to test the associations. For testing the association between SNP and disease, there are single SNP method and multiple SNPs method. And we built up an overall flow path for GWA study (See Figure 15, Table 2).

5.2 Discussion

Although the tests of association have large number of significant SNPs, most of them are significant just because they are highly correlated with disease SNPs. The next step is to do the fine mapping to find the causal SNP and verify it by biological explanation. And since the number of individuals is smaller than the number of SNPs, so the power may be an issue for us to resolve. There are increasing researchers devoted to GWA study. Thousand of method and theory are develop to resolve the problem they meet at the study. Our purpose is to construct a standard flow path such that everyone who wants to do the GWA study has accidence for knowing what GWA study is. Of course, there are many method and theory we do not mention, anyone who make the GWA study can extend it by search information from the internet.

Reference

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

2. Balding DJ (2007). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7, 781-91.

3. McCarthy MI, et al (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 9, 356-369.

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

5. Pritchard JK, et al (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945-959.

6. Schaid DJ (2004). Evaluating associations of haplotypes with traits. *Genet. Epidemiol.* 27, 348–364.

7. Douglas F. Easton, et al (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447, 1087-1093

8. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research: Richa Saxena, et al (2007). Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. *Science* 316, 1331-1336.

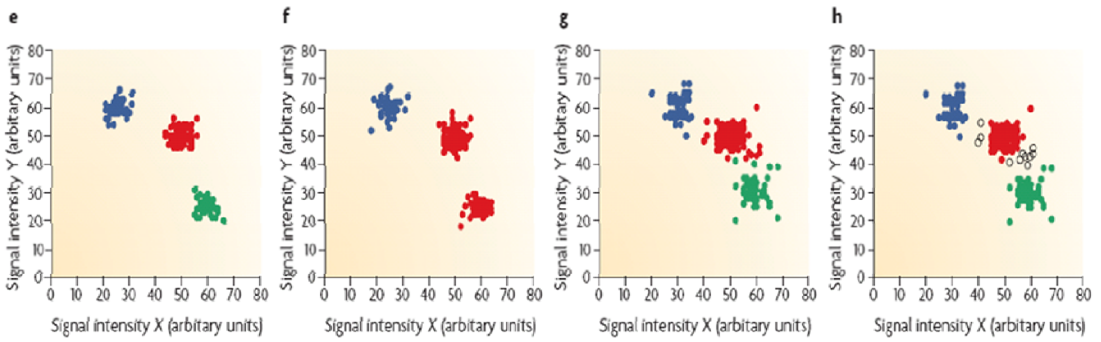
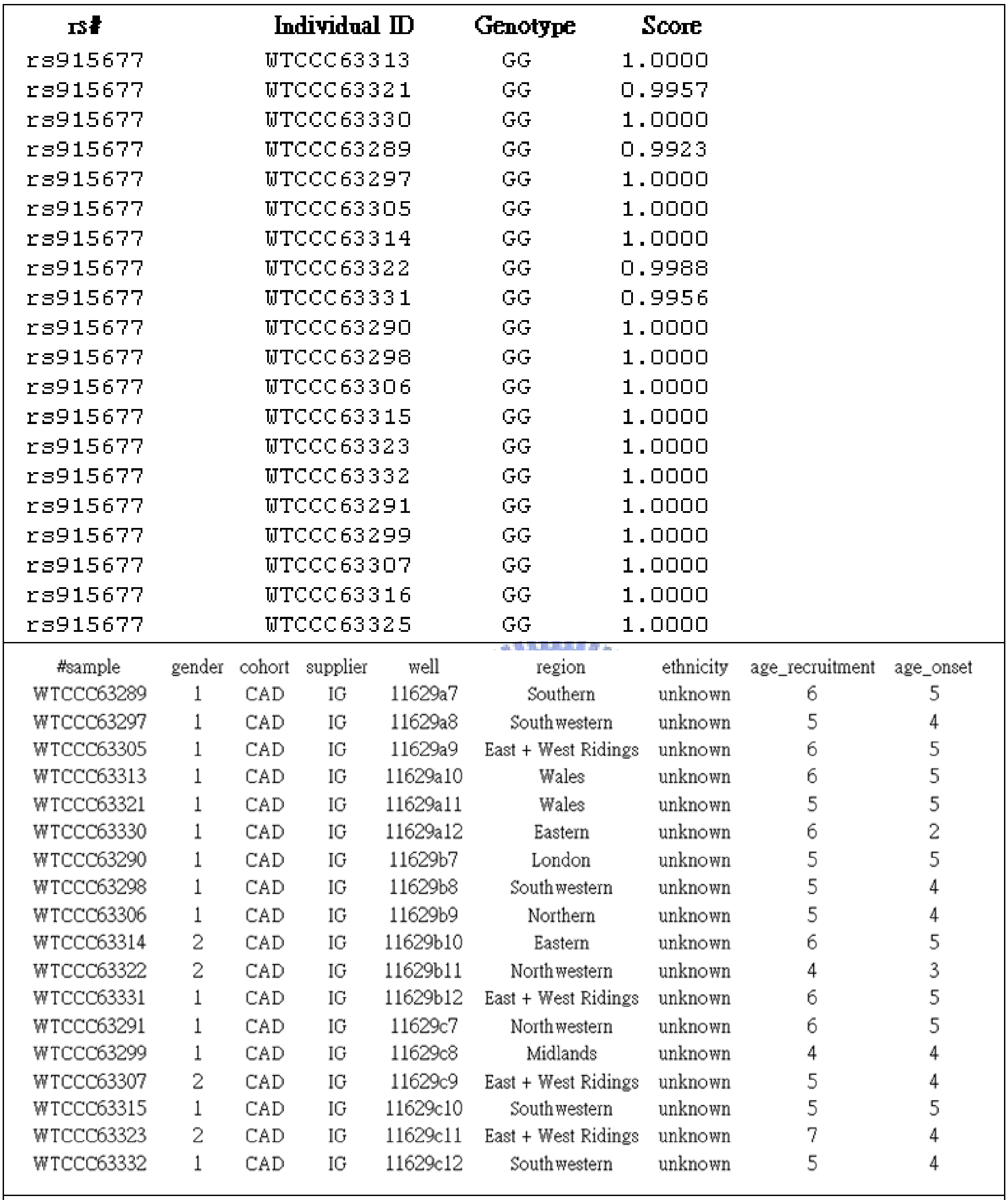


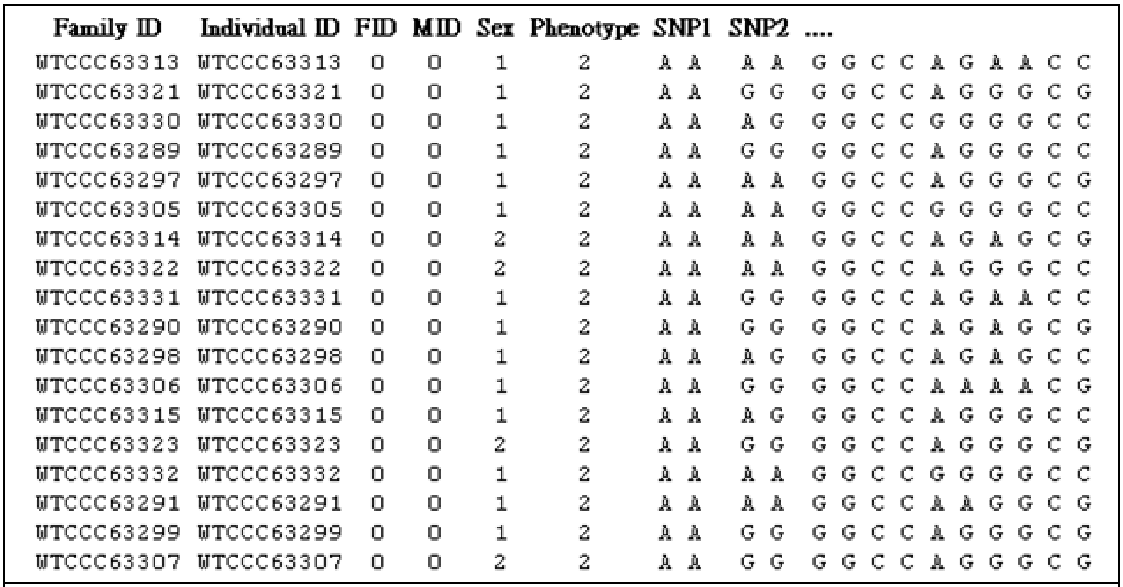
Figure 1. Calling algorithm influence Genotype accuracy. Three colors represent different genotypes. The ideal genotype cluster will be separated clearly (panel e). Panel f has a good cluster, but a mistake calling algorithm. In panel g and h, significant overlap between clusters is likely to result in failure to call certain genotypes



rs#	Individual ID	Genotype	Score
rs915677	WTCCC63313	GG	1.0000
rs915677	WTCCC63321	GG	0.9957
rs915677	WTCCC63330	GG	1.0000
rs915677	WTCCC63289	GG	0.9923
rs915677	WTCCC63297	GG	1.0000
rs915677	WTCCC63305	GG	1.0000
rs915677	WTCCC63314	GG	1.0000
rs915677	WTCCC63322	GG	0.9988
rs915677	WTCCC63331	GG	0.9956
rs915677	WTCCC63290	GG	1.0000
rs915677	WTCCC63298	GG	1.0000
rs915677	WTCCC63306	GG	1.0000
rs915677	WTCCC63315	GG	1.0000
rs915677	WTCCC63323	GG	1.0000
rs915677	WTCCC63332	GG	1.0000
rs915677	WTCCC63291	GG	1.0000
rs915677	WTCCC63299	GG	1.0000
rs915677	WTCCC63307	GG	1.0000
rs915677	WTCCC63316	GG	1.0000
rs915677	WTCCC63325	GG	1.0000

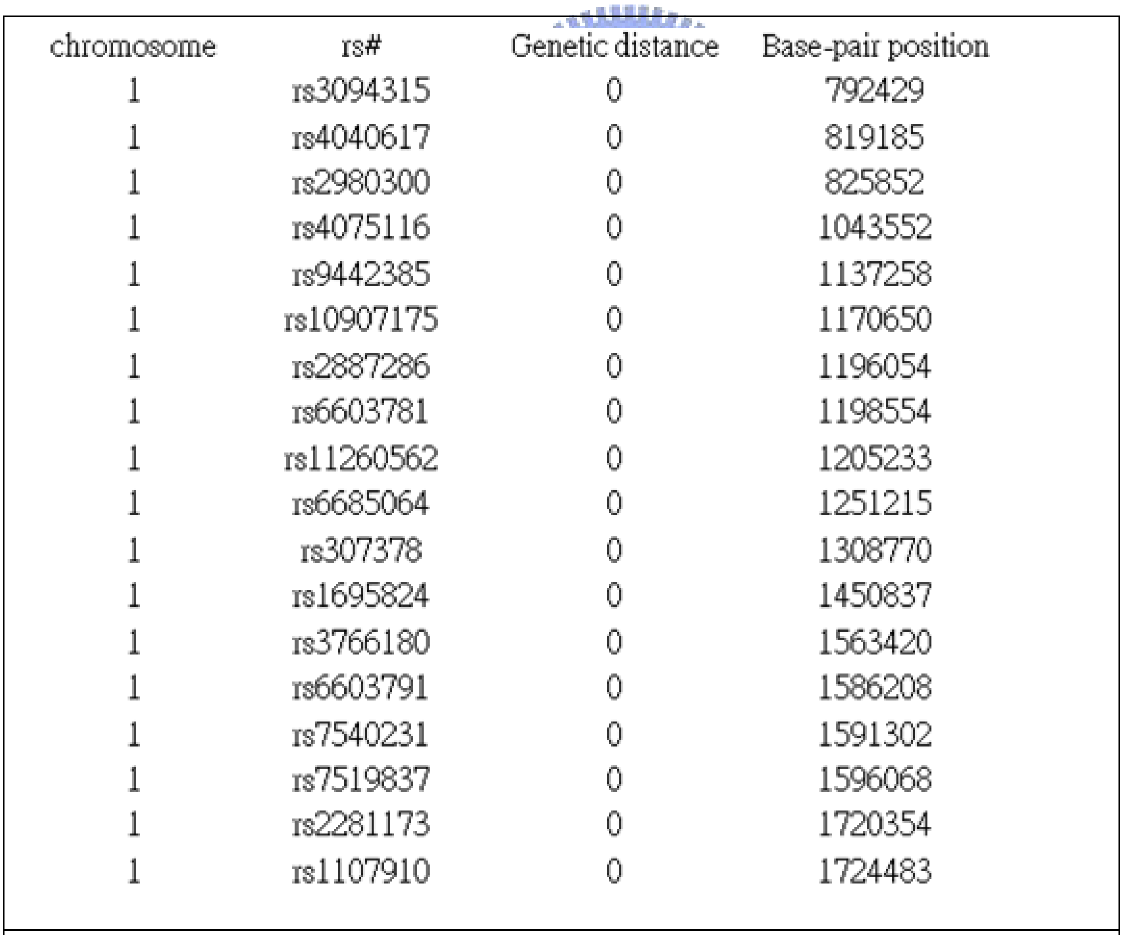
#sample	gender	cohort	supplier	well	region	ethnicity	age_recruitment	age_onset
WTCCC63289	1	CAD	IG	11629a7	Southern	unknown	6	5
WTCCC63297	1	CAD	IG	11629a8	Southwestern	unknown	5	4
WTCCC63305	1	CAD	IG	11629a9	East + West Ridings	unknown	6	5
WTCCC63313	1	CAD	IG	11629a10	Wales	unknown	6	5
WTCCC63321	1	CAD	IG	11629a11	Wales	unknown	5	5
WTCCC63330	1	CAD	IG	11629a12	Eastern	unknown	6	2
WTCCC63290	1	CAD	IG	11629b7	London	unknown	5	5
WTCCC63298	1	CAD	IG	11629b8	Southwestern	unknown	5	4
WTCCC63306	1	CAD	IG	11629b9	Northern	unknown	5	4
WTCCC63314	2	CAD	IG	11629b10	Eastern	unknown	6	5
WTCCC63322	2	CAD	IG	11629b11	Northwestern	unknown	4	3
WTCCC63331	1	CAD	IG	11629b12	East + West Ridings	unknown	6	5
WTCCC63291	1	CAD	IG	11629c7	Northwestern	unknown	6	5
WTCCC63299	1	CAD	IG	11629c8	Midlands	unknown	4	4
WTCCC63307	2	CAD	IG	11629c9	East + West Ridings	unknown	5	4
WTCCC63315	1	CAD	IG	11629c10	Southwestern	unknown	5	5
WTCCC63323	2	CAD	IG	11629c11	East + West Ridings	unknown	7	4
WTCCC63332	1	CAD	IG	11629c12	Southwestern	unknown	5	4

Figure 2. Raw data format before convert



Family ID	Individual ID	FID	MID	Sex	Phenotype	SNP1	SNP2	....
WTCCC63313	WTCCC63313	0	0	1	2	A A	A A	G G C C A G A A C C
WTCCC63321	WTCCC63321	0	0	1	2	A A	G G	G G C C A G G G C G
WTCCC63330	WTCCC63330	0	0	1	2	A A	A G	G G C C G G G G C C
WTCCC63289	WTCCC63289	0	0	1	2	A A	G G	G G C C A G G G C C
WTCCC63297	WTCCC63297	0	0	1	2	A A	A A	G G C C A G G G C G
WTCCC63305	WTCCC63305	0	0	1	2	A A	A A	G G C C G G G G C C
WTCCC63314	WTCCC63314	0	0	2	2	A A	A A	G G C C A G A G C G
WTCCC63322	WTCCC63322	0	0	2	2	A A	A A	G G C C A G G G C C
WTCCC63331	WTCCC63331	0	0	1	2	A A	G G	G G C C A G A A C C
WTCCC63290	WTCCC63290	0	0	1	2	A A	G G	G G C C A G A G C G
WTCCC63298	WTCCC63298	0	0	1	2	A A	A G	G G C C A G A G C C
WTCCC63306	WTCCC63306	0	0	1	2	A A	G G	G G C C A A A A C G
WTCCC63315	WTCCC63315	0	0	1	2	A A	A G	G G C C A G G G C C
WTCCC63323	WTCCC63323	0	0	2	2	A A	G G	G G C C A G G G C G
WTCCC63332	WTCCC63332	0	0	1	2	A A	A A	G G C C G G G G C C
WTCCC63291	WTCCC63291	0	0	1	2	A A	A A	G G C C A A G G C G
WTCCC63299	WTCCC63299	0	0	1	2	A A	G G	G G C C A G G G C G
WTCCC63307	WTCCC63307	0	0	2	2	A A	G G	G G C C A G G G C G

Figure 3.a. PLINK format data-ped format



chromosome	rs#	Genetic distance	Base-pair position
1	rs3094315	0	792429
1	rs4040617	0	819185
1	rs2980300	0	825852
1	rs4075116	0	1043552
1	rs9442385	0	1137258
1	rs10907175	0	1170650
1	rs2887286	0	1196054
1	rs6603781	0	1198554
1	rs11260562	0	1205233
1	rs6685064	0	1251215
1	rs307378	0	1308770
1	rs1695824	0	1450837
1	rs3766180	0	1563420
1	rs6603791	0	1586208
1	rs7540231	0	1591302
1	rs7519837	0	1596068
1	rs2281173	0	1720354
1	rs1107910	0	1724483

Figure 3.b. PLINK format data-map format

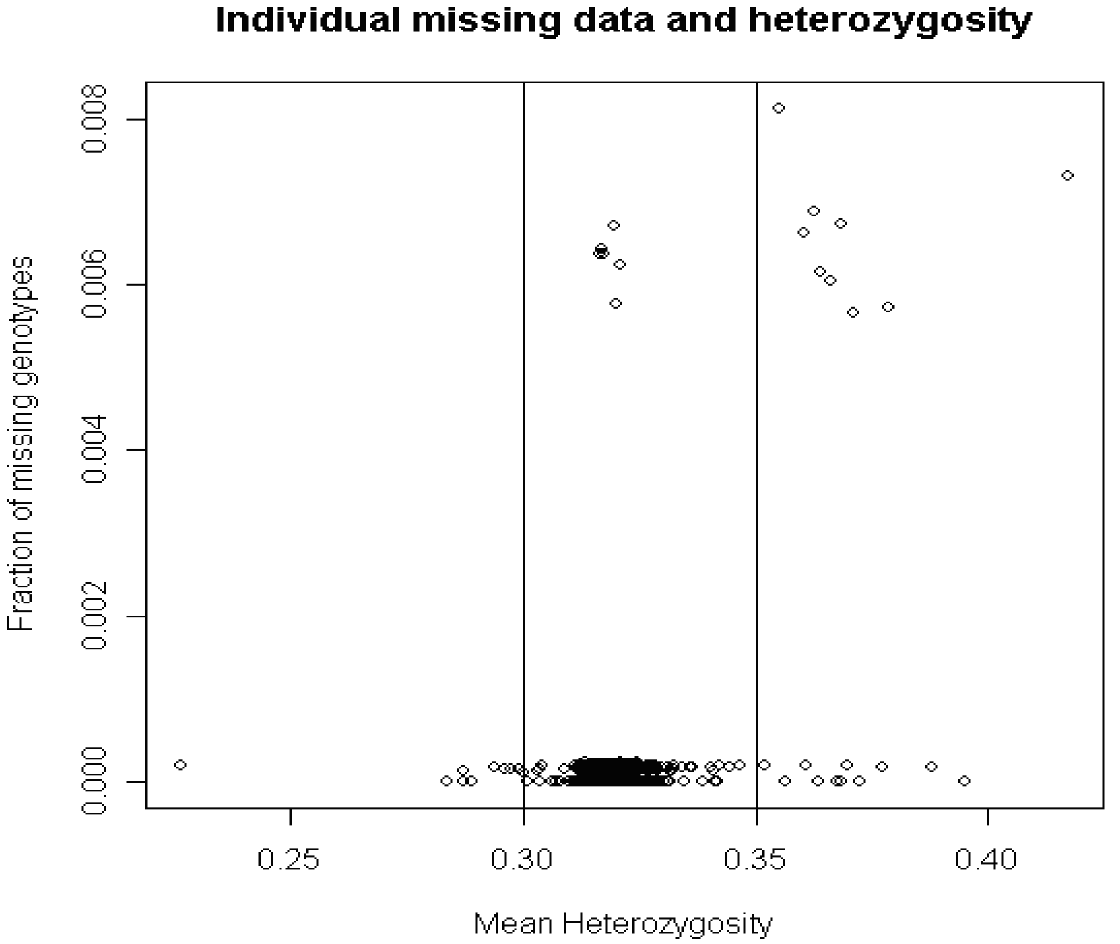


Figure 4. Scatter plot for mean heterozygosity and fraction of missing genotype for individuals. We exclude individual by heterozygosity >0.35 or heterozygosity < 0.3.

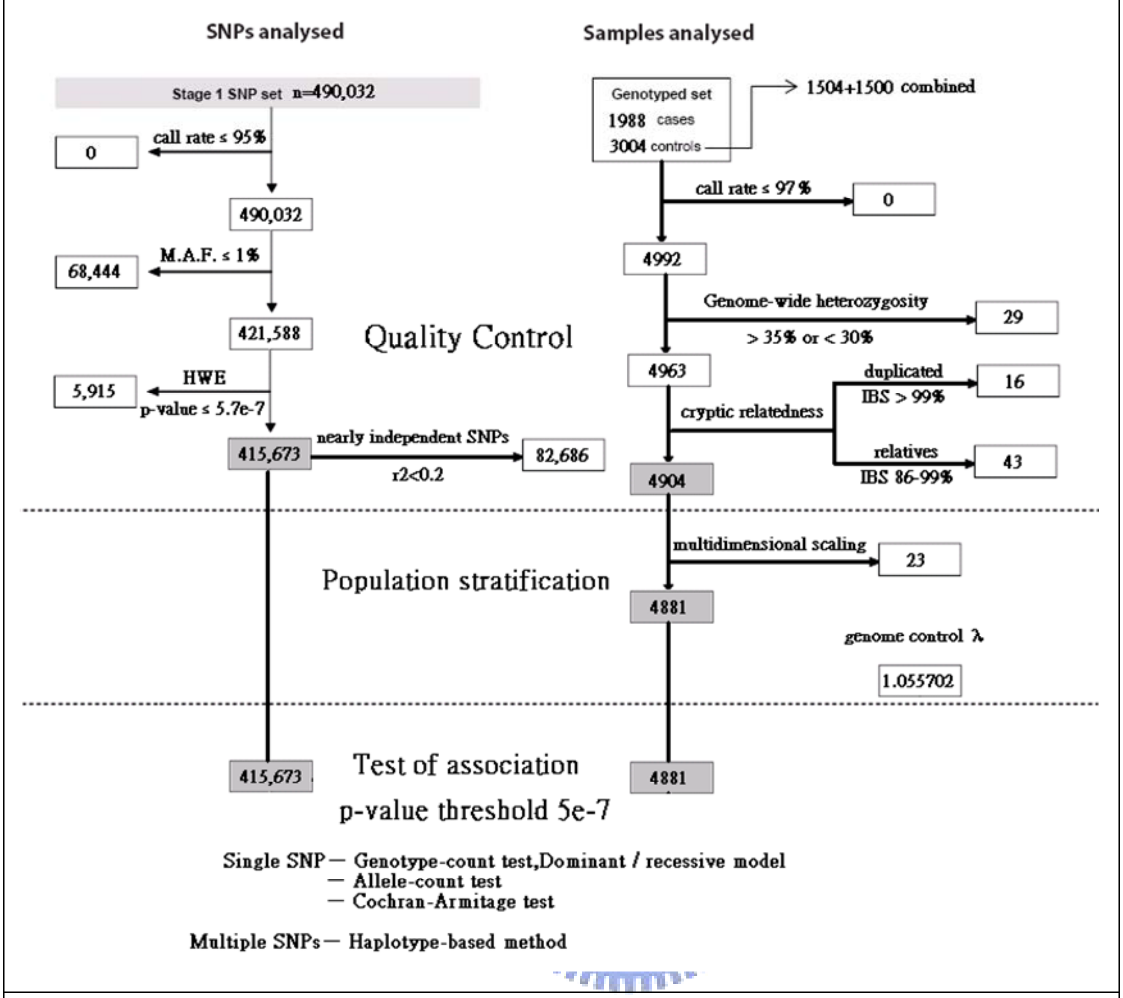


Figure 5. Quality control flow path

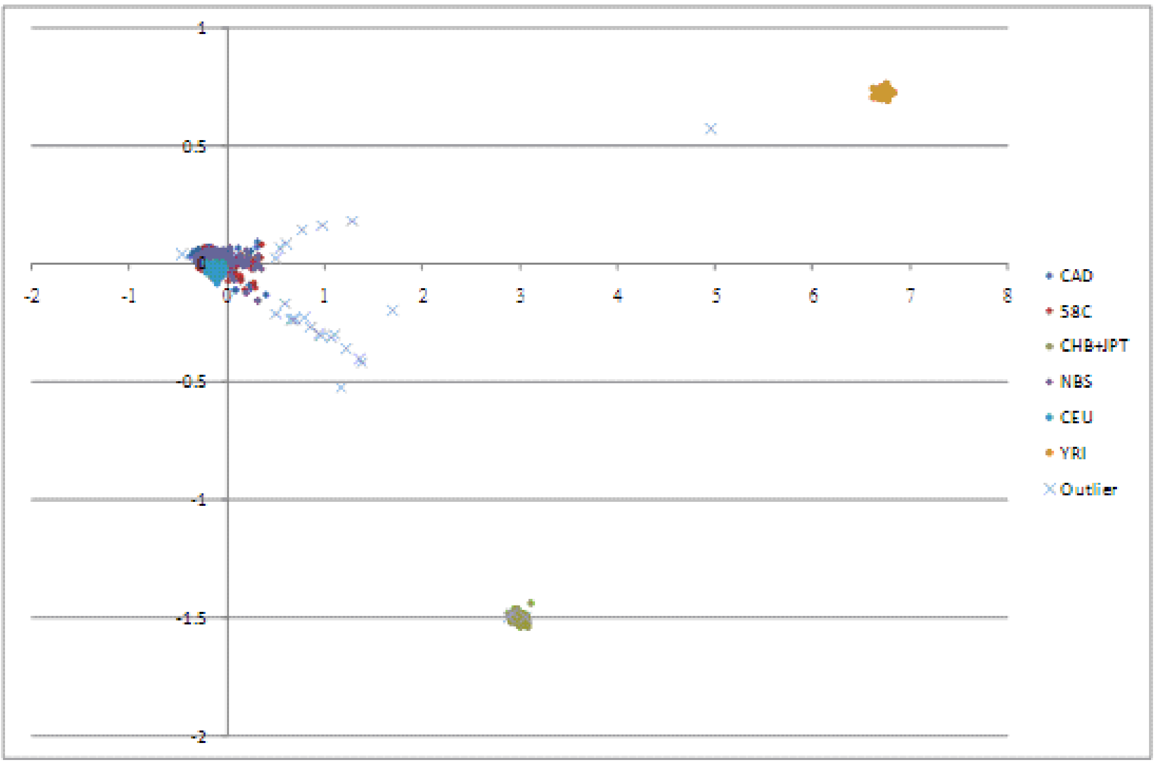


Figure 6. MDS plot for CAD case and combined control. We exclude individuals by points with x-axis value>0.5.

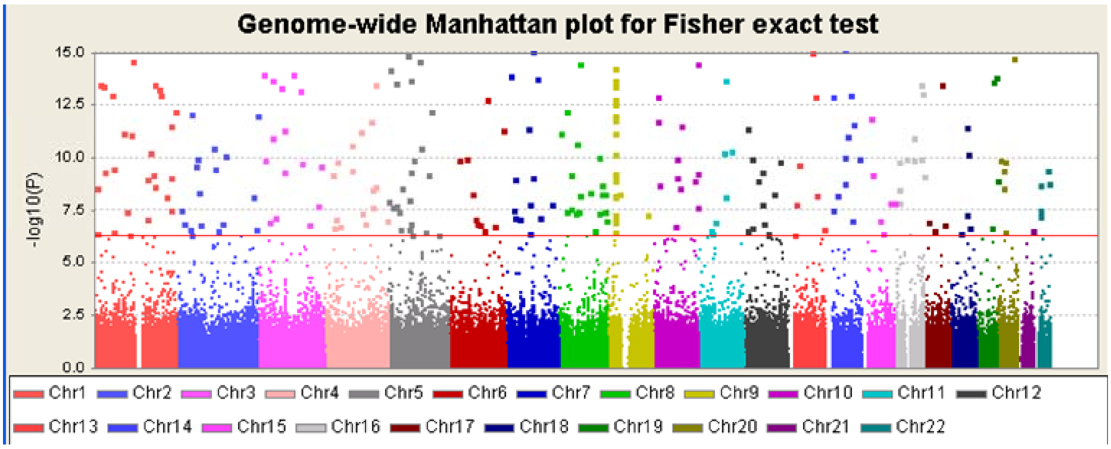


Figure 7. Genome-wide Manhattan plot for fisher's exact test

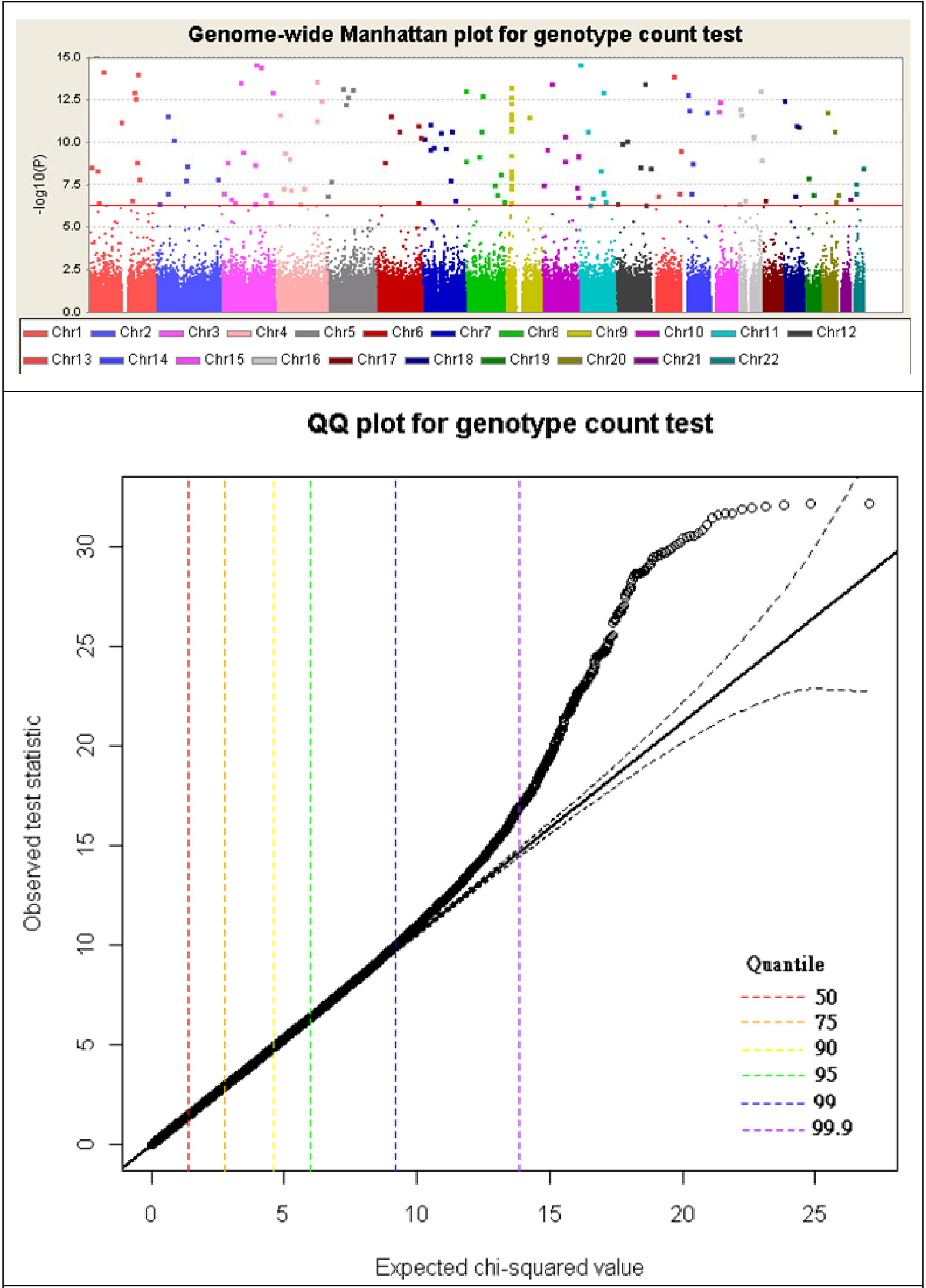


Figure 8. QQ plot and Genome-wide Manhattan plot for genotype test

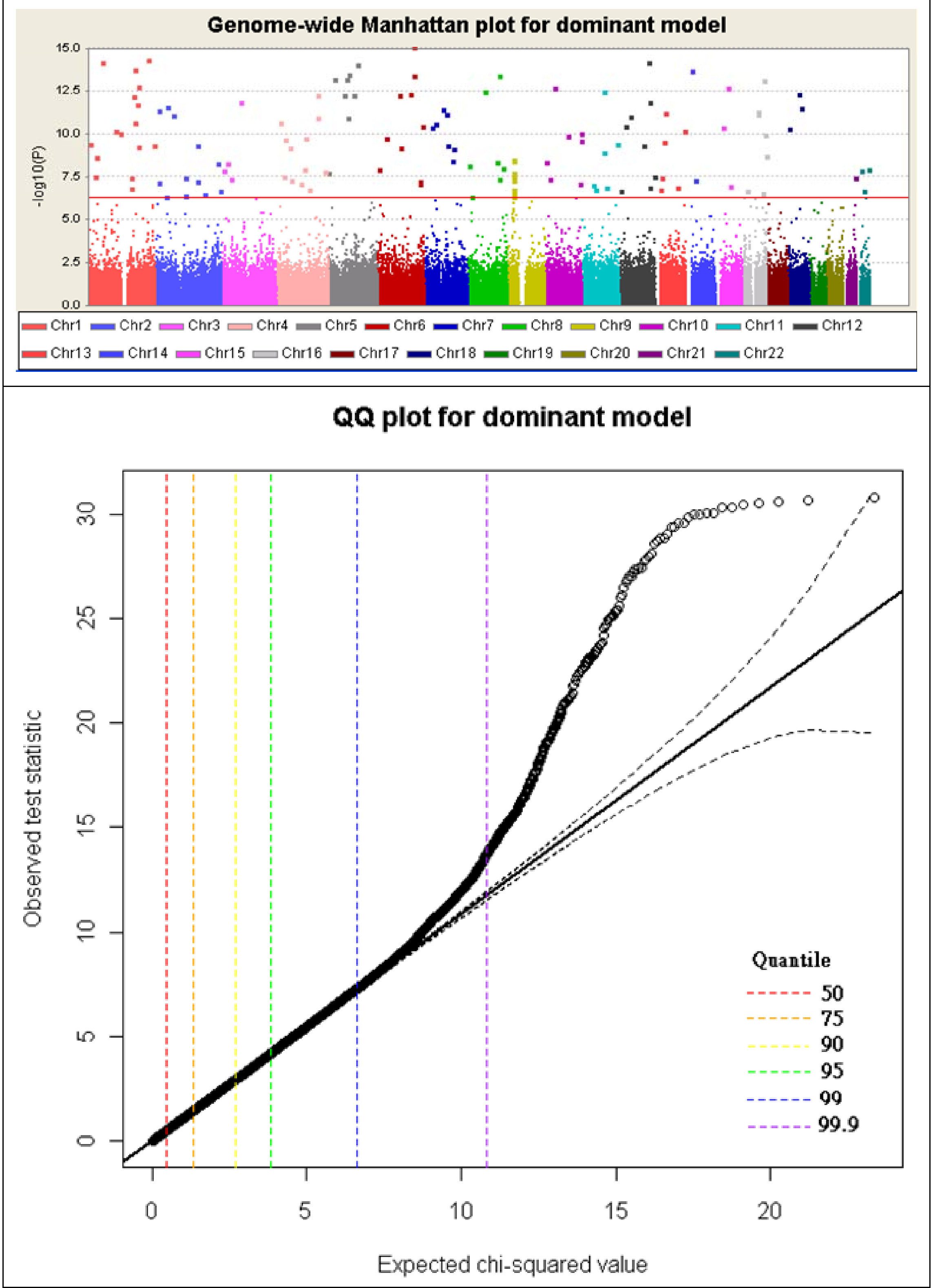


Figure 9. QQ plot and Genome-wide Manhattan plot for dominant model

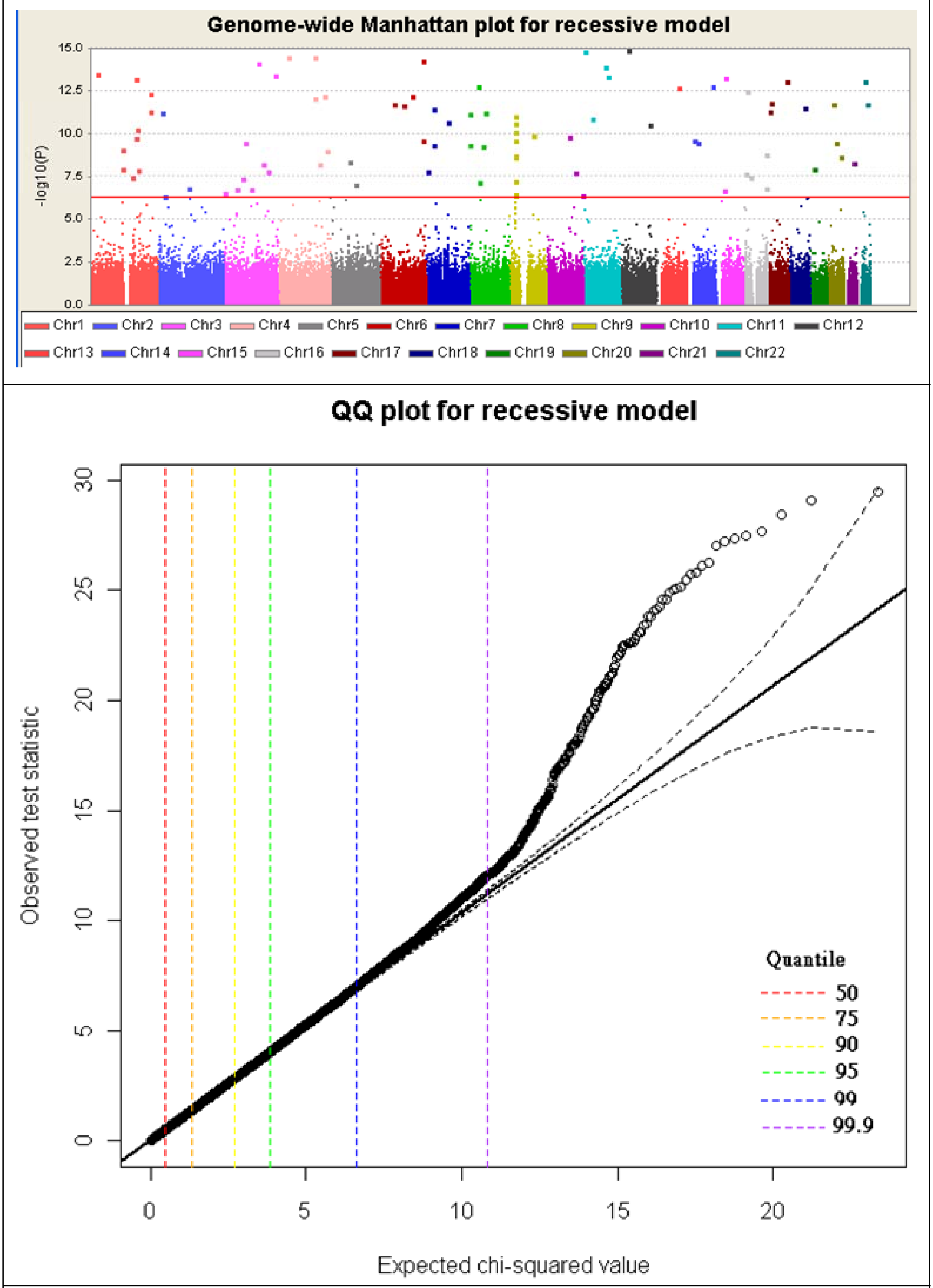


Figure 10. QQ plot and Genome-wide Manhattan plot for recessive model

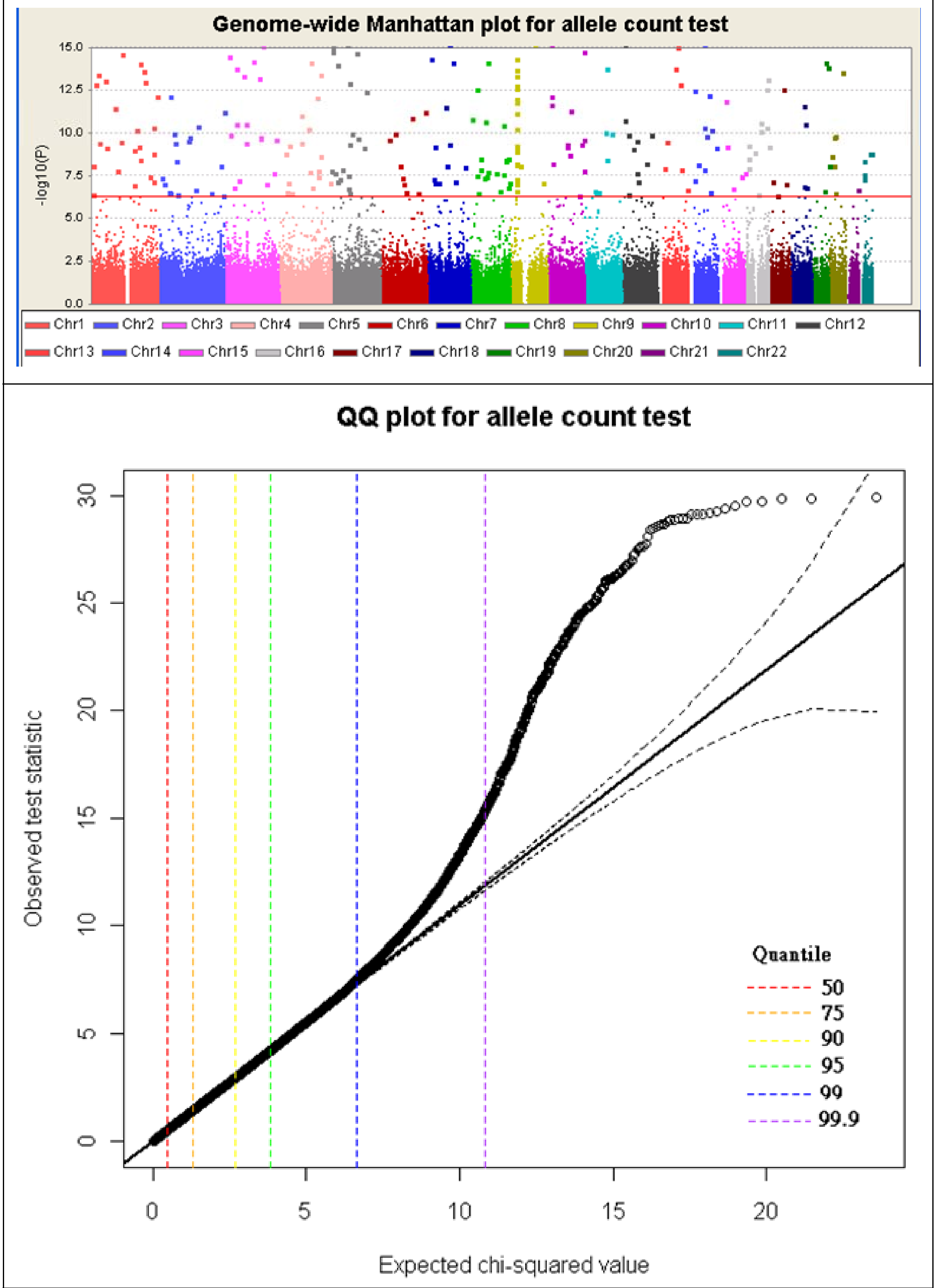


Figure 11. QQ plot and Genome-wide Manhattan plot for allele count test

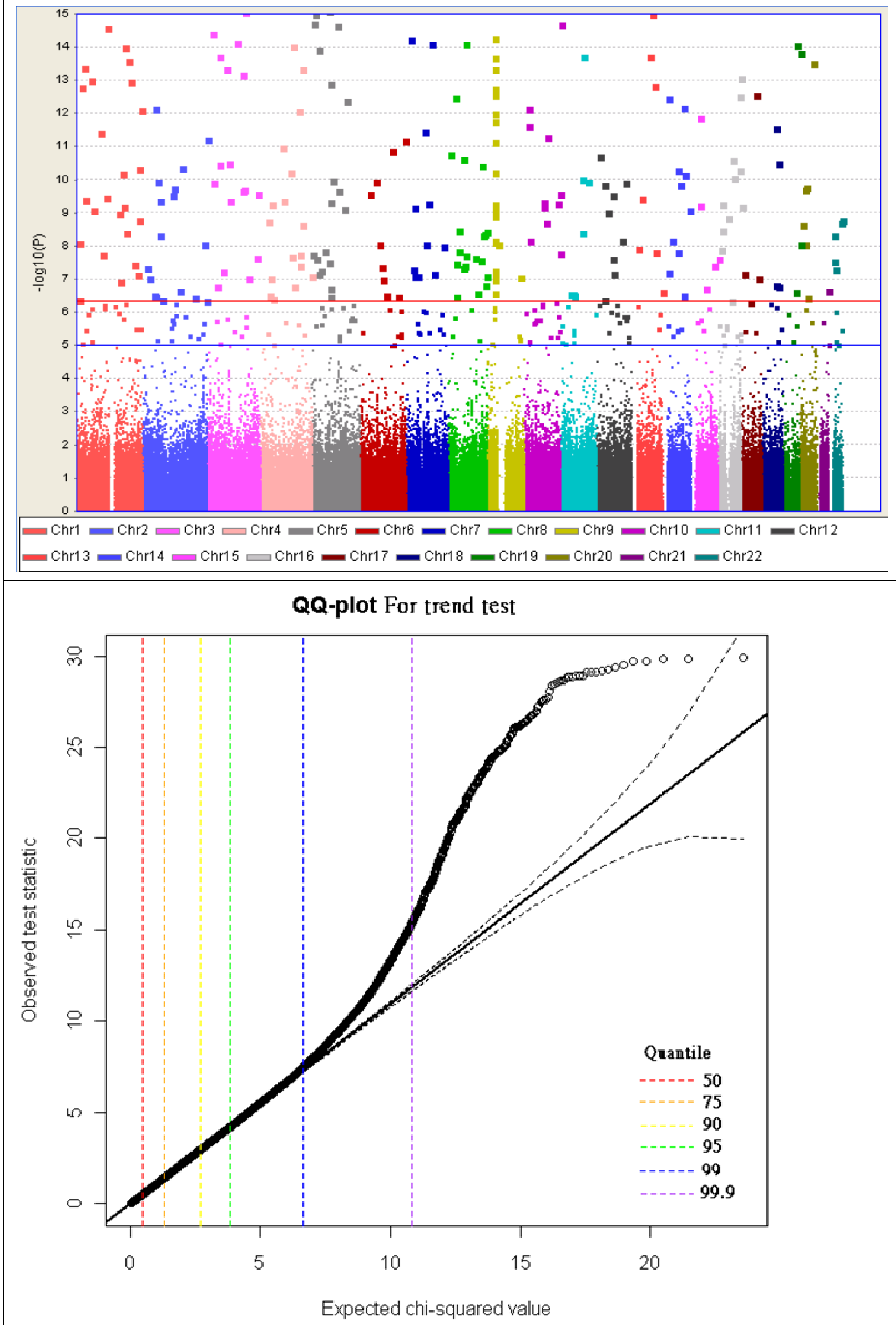


Figure 12. QQ plot and Genome-wide Manhattan plot for Cochran-Armitage trend test

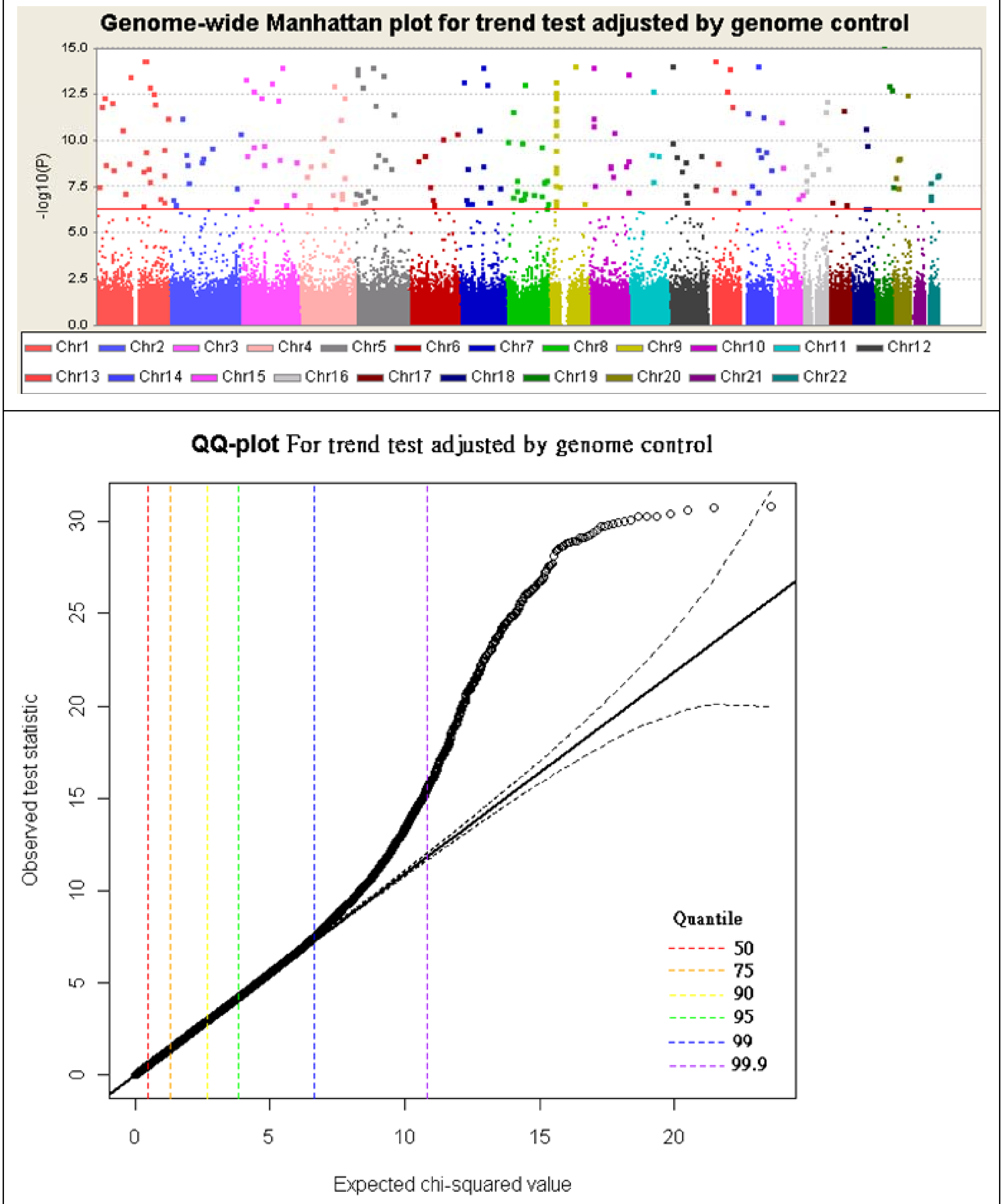


Figure 13. QQ plot and Genome-wide Manhattan plot for Cochran-Armitage trend test adjusted by GC

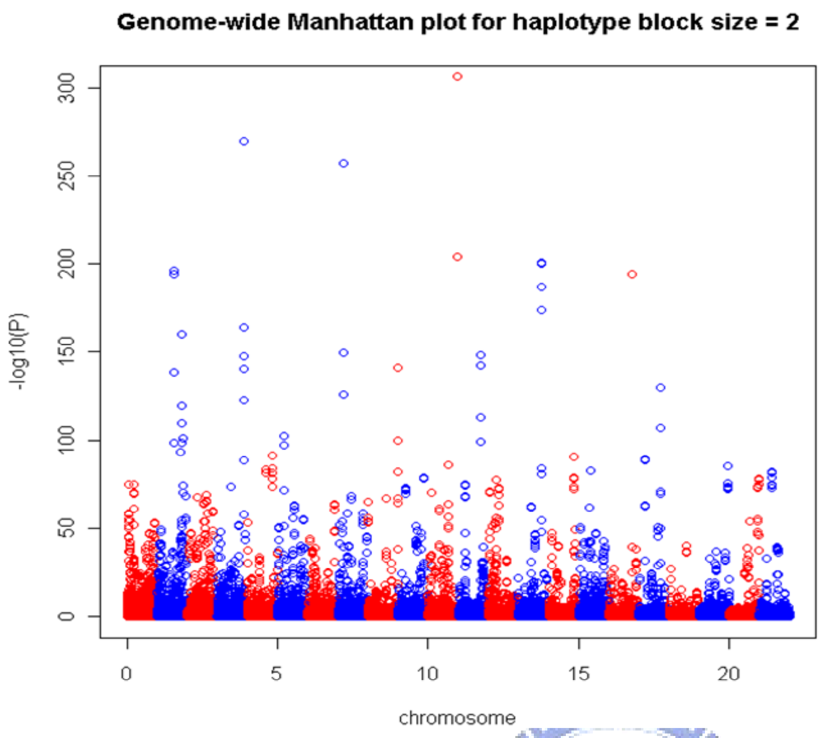


Figure 14.a. Genome-wide Manhattan plot for haplotype test block size = 2 Chromosomes are divided by colors.

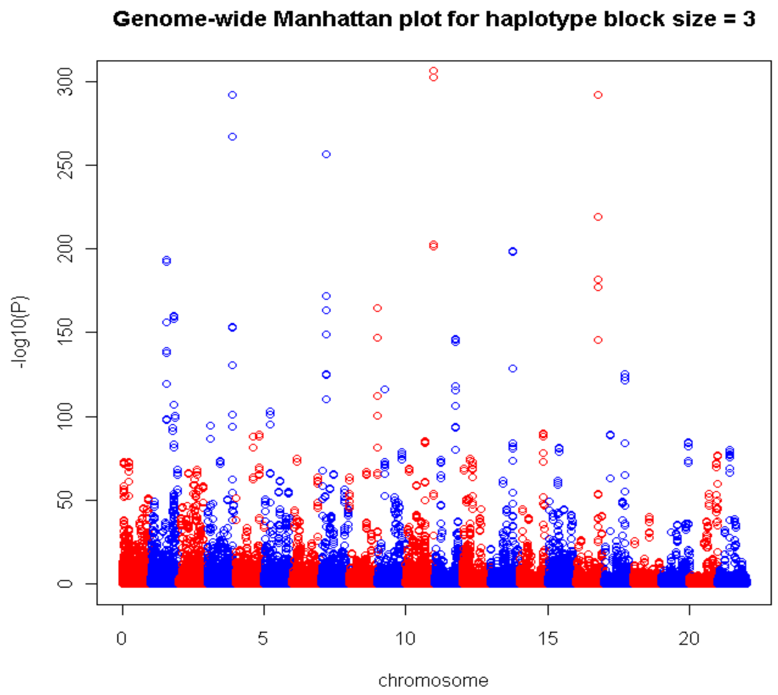


Figure 14.b. Genome-wide Manhattan plot for haplotype test block size = 3 Chromosomes are divided by colors.

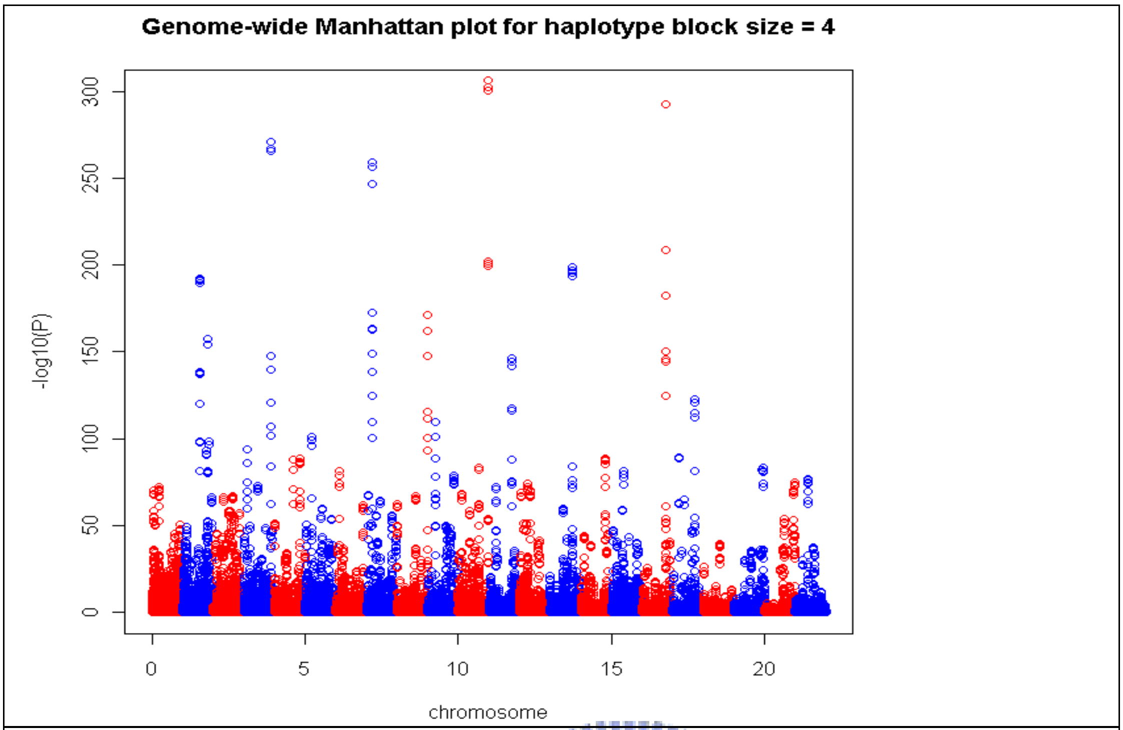


Figure 14.c. Genome-wide Manhattan plot for haplotype test block size = 4 Chromosomes are divided by colors.

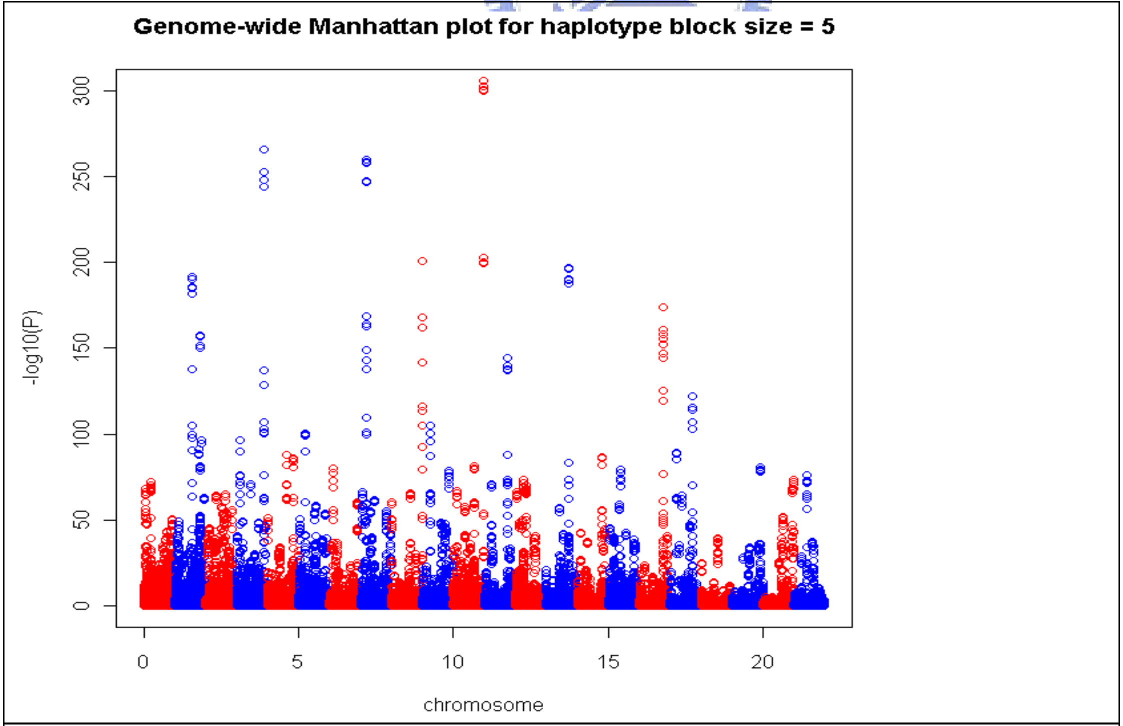


Figure 14.d. Genome-wide Manhattan plot for haplotype test block size = 5 Chromosomes are divided by colors.

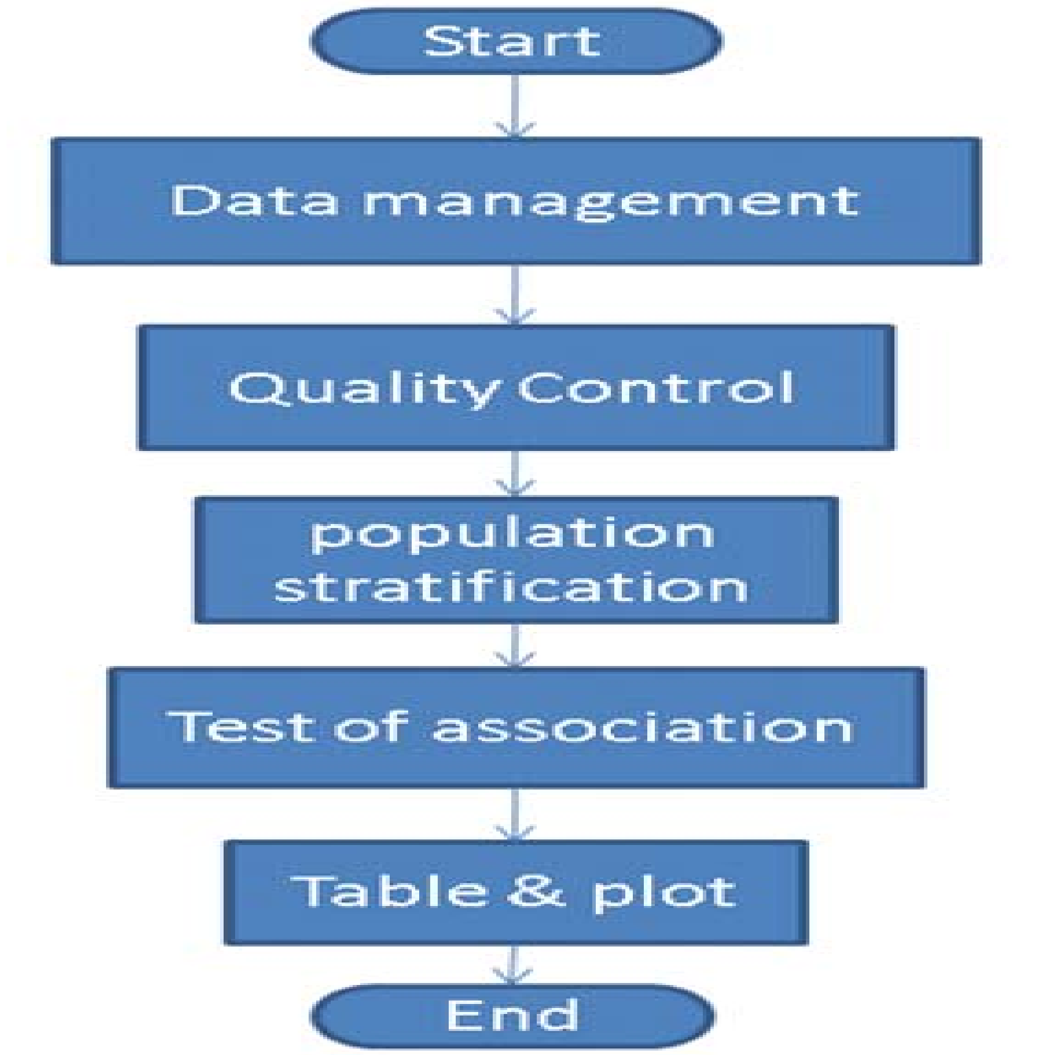
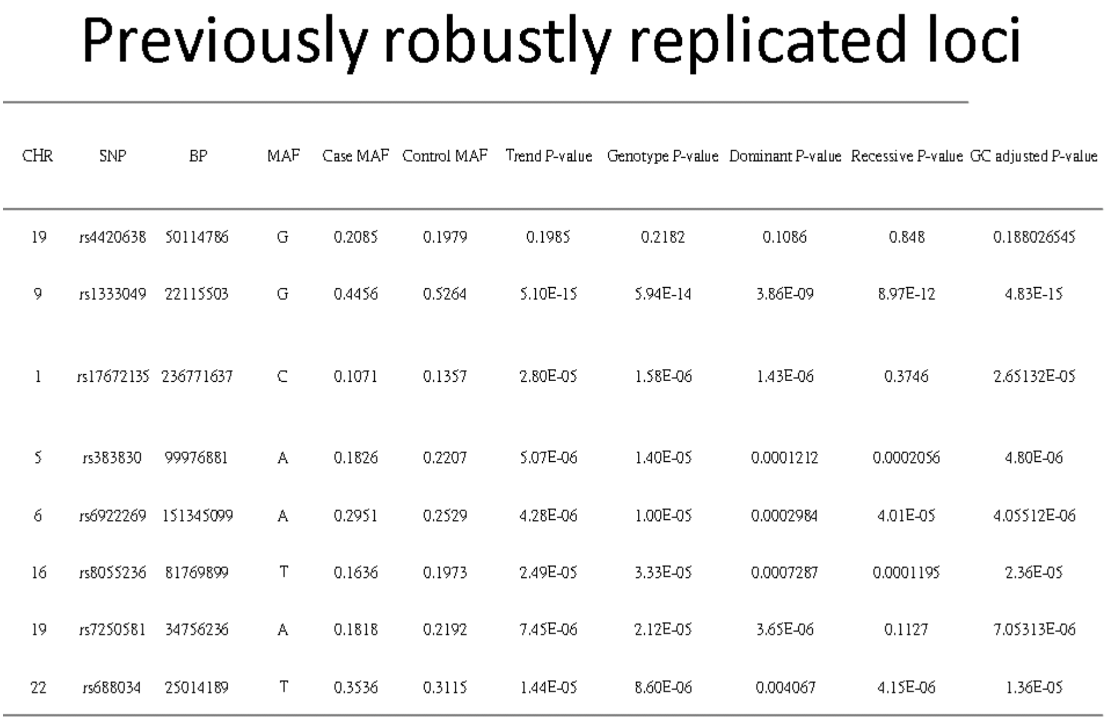


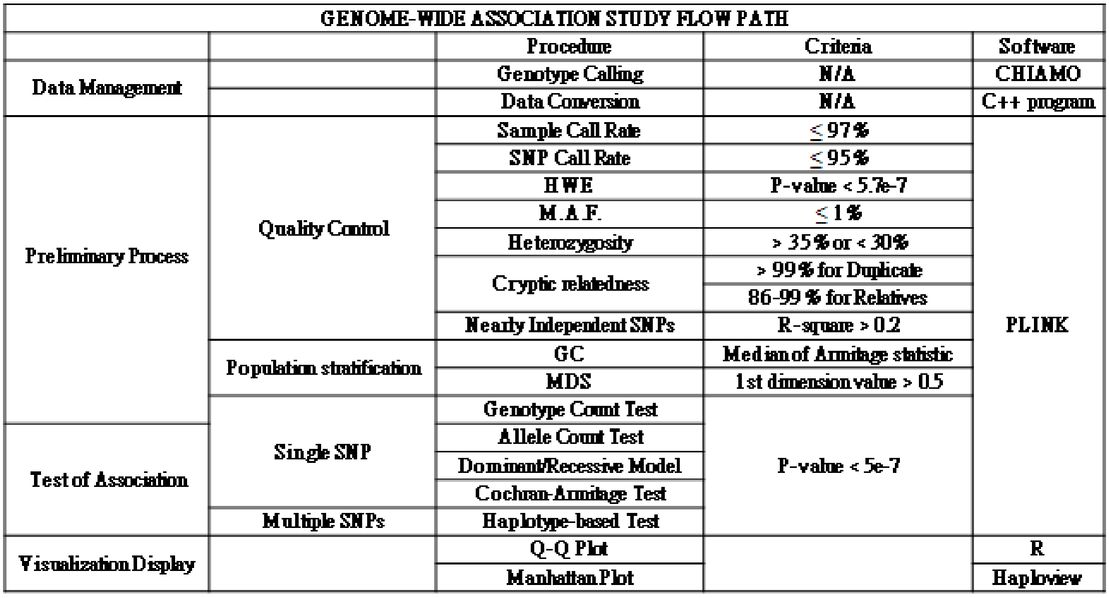
Figure 15. GWA study flow chart



Previously robustly replicated loci

CHR	SNP	BP	MAF	Case MAF	Control MAF	Trend P-value	Genotype P-value	Dominant P-value	Recessive P-value	GC adjusted P-value
19	rs4420638	50114786	G	0.2085	0.1979	0.1985	0.2182	0.1086	0.848	0.188026545
9	rs1333049	22115503	G	0.4456	0.5264	5.10E-15	5.94E-14	3.86E-09	8.97E-12	4.83E-15
1	rs17672135	236771637	C	0.1071	0.1357	2.80E-05	1.58E-06	1.43E-06	0.3746	2.65132E-05
5	rs383830	99976881	A	0.1826	0.2207	5.07E-06	1.40E-05	0.0001212	0.0002056	4.80E-06
6	rs6922269	151345099	A	0.2951	0.2529	4.28E-06	1.00E-05	0.0002984	4.01E-05	4.05512E-06
16	rs8055236	81769899	T	0.1636	0.1973	2.49E-05	3.33E-05	0.0007287	0.0001195	2.36E-05
19	rs7250581	34756236	A	0.1818	0.2192	7.45E-06	2.12E-05	3.65E-06	0.1127	7.05313E-06
22	rs688034	25014189	T	0.3536	0.3115	1.44E-05	8.60E-06	0.004067	4.15E-06	1.36E-05

Table 1. Previously robustly replicated loci



GENOME-WIDE ASSOCIATION STUDY FLOW PATH

		Procedure	Criteria	Software
Data Management		Genotype Calling	N/A	CHIAMO
		Data Conversion	N/A	C++ program
Preliminary Process	Quality Control	Sample Call Rate	≤ 97%	PLINK
		SNP Call Rate	≤ 95%	
		HWE	P-value < 5.7e-7	
		M.A.F.	≤ 1%	
		Heterozygosity	> 35% or < 30%	
		Cryptic relatedness	> 99% for Duplicate 86-99 % for Relatives	
		Nearly Independent SNPs	R-square > 0.2	
	Population stratification	GC	Median of Armitage statistic	
		MDS	1st dimension value > 0.5	
	Single SNP	Genotype Count Test	P-value < 5e-7	
Test of Association		Allele Count Test		
		Dominant/Recessive Model		
		Cochran-Armitage Test		
	Multiple SNPs	Haplotype-based Test		
Visualization Display		Q-Q Plot		R
		Manhattan Plot		Haploview

Table 2. GWA procedures and software used in the paper