GENETIC VARIATIONS AND THEIR EFFECTS ON CORONARY HEART DISEASE AND CERVICAL CANCER

By

Yalu Wen

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Epidemiology

2012

ABSTRACT

GENETIC VARIATIONS AND THEIR EFFECTS ON CORONARY HEART DISEASE AND CERVICAL CANCER

By

Yalu Wen

Benefiting from high throughput technologies, significant progress has been made by genome wide association studies (GWASs) and gene expression profiling to map genetic susceptibility region for complex human diseases. Evidences show that the current findings can only explain part of the genetic etiologies, which could be partially due to the noise and artifacts introduced by high throughput technologies and the deficiencies in the current available analytical tools. To reduce the effects of noise and artifacts and facilitate the genetic studies, I develop three statistical methods which aim at 1) providing accurate genotype calls by modeling the underlying hybridization process of microarray with the consideration of batch effect; 2) reducing false positive and false negative findings for differentially expressed gene identification by incorporating the variability of data preprocessing into the differentially expressed gene detection algorithm and 3) studying the genetic etiologies contributing to comorbidity between complex human diseases by proposing a multivariate Mann-Whitney method built based upon a U-statistic with forward selection algorithm. Through simulations, analyses of the Latin Square Data, and the HapMap data, I show that the three proposed methods outperform the current existing methods and are robust under various experimental conditions and disease models. I further apply these methods to datasets obtained from *Wellcome Trust Case Control Consortium* to identify the genetic susceptibility loci predisposing to coronary heart disease and to the comorbidity between coronary heart disease and Type II diabetes. With these newly developed

methods, the loci identified for coronary heart disease are consistent with the findings by various technologies, which indicates the proposed method could provide accurate genotype calls and benefit the downstream analysis. No loci have been selected to be associated with the co-morbidity of coronary heart disease and type II diabetes which may be due to the study design and the candidate gene approach used in this research. Further studies are needed to investigate the comorbidity between coronary heart disease and type II diabetes. I also apply my method to identify differentially expressed genes for a cervical cancer study. The findings replicate most of the original discoveries. In addition, several other genes, which potentially play an important role in the cervical cancer development, have also been identified.

DEDICATION

To

my father Jingyu Wen and my mother Limin Yao

ACKNOWLEDGEMENTS

I am greatly indebted to my advisor, Dr. Wenjiang Fu who has provided me with tremendous help for every step of my graduate studies. Your support and guidance have kept me in the right track towards the completion of my dissertation and achievement of my doctoral degree. I would also like to express my sincerest gratitude to my thesis committee, Dr. Qing Lu, Dr. Ellen Velie, and Dr. Donna Wang for their professional and academic guidance during various aspects of my research.

I am also thankful to Dr. Katherine Alaimo, Dr. Karin Pfeiffer and Dr. Gretchen Birbeck for providing financial support and opportunities to broaden my knowledge. I appreciated all the assistance and help provided by my colleague Ming Li. I gratefully acknowledge all the staff in the Department of Epidemiology for their assistance through my academic training.

Finally, I would like to thank my family and friends for their continuous support and encouragement throughout the entire academic training.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CAD	Coronary artery disease
CHD	Coronary heart disease
COM	Composite phenotype method
CVD	Cardiovascular disease
DEG	Differentially expressed genes
GPDNN	Generalized positional-dependent-nearest-neighbor model
GWAS	Genome-wide association study
NBS	UK Blood Service Control
PDNN	Positional-dependent-nearest-neighbor model
PICR	Probe intensity composite representation model
SNP	Single nucleotide polymorphisms
T2D	Type II diabetes
MMW	Multivariate Mann-Whitney
WTCCC	Wellcome Trust Case-Control Consortium

CHAPTER 1

BACKGROUND AND OBJECTIVES

1.1 Background and Significance

The genetic etiology of complex human diseases is of great interest to clinicians, researchers as well as the general public. Mapping the genetic susceptibility loci onto the genome relies on accurate and efficient algorithm to evaluate the effect of target loci. So far, many computational approaches have been proposed to improve the accuracy of genotype calls, the power of association and the sensitivity and specificity of the detection of differentially expressed gene [1-17]. The ultimate goals of these genetic studies are to identify population at high risk, to promote new diagnostics and therapeutics, and to develop personalized medicine to treat patients. In this dissertation research, I focus on identifying genetic underlying mechanisms for two common complex human diseases, coronary heart disease and cervical cancer.

1.1.1 Coronary Heart Disease

Coronary heart disease (CHD) is one of the leading causes of death and disability worldwide, especially in the developed countries [18]. In the United States, though the death rate declines significantly during the past few decades possibly due to medical advances in disease prevention and treatment[19], the CHD is still responsible for about 30% of all deaths in people over 35 years old[20, 21] and is one of the largest killer for both men and women. Evidences from the Behavioral Risk Factor Surveillance System (BRFSS) of the Centers for Disease Control and Prevention (CDC) show that 4% of the respondents had a history of myocardial infarction (MI) and 4.4% of the respondents had a history of CHD. The prevalence of CHD increases significantly with age, and men have a significantly higher prevalence than women (5.5% vs.

3.4%). Evidences suggest the CHD prevalence is negatively associated with education level, and among all racial groups the American Indian/Alaska Native has the highest prevalence [22]. The lifetime risks of CHD for people over 40 years old are 49% and 32% for men and women , respectively[23], and the CHD accounts for more than 50% of all cardiovascular disease for men and women over 75 years[24]. CHD incidence in women lags behind men by 10 years and 20 years, respectively for total CHD and for serious CHD events such as sudden death[24]. The risk factors of CHD can be classified as modifiable risk factors and non-modifiable risk factors. Age and gender are two well known non-modifiable risk factors for CHD [22-24]. High blood cholesterol, cigarette smoking, hypertension, diabetes mellitus, obesity and overweight, physical inactivity, alcohol overconsumption, and diet low in antioxidants are well known modifiable risk factors for CHD, and the effects of these risk factors are consistent among different racial/ethnicity groups and across varied geographic regions[25-35].

Family history of CHD is related to each stage of the disease[36]: it elevates risk factors[37], subclinical atherosclerosis[38], and clinical manifestation of CHD[39]. The increased susceptibility for people with family history of CHD may be due to the shared culture, lifestyle and environmental factors as well as multiple susceptibility genes [40-42]. For example in the Framingham Offspring Study, the odds ratio for men and women with family history of cardiovascular disease (CVD) is 2.6, and 2.3, respectively. With adjustment of traditional risk factors, the family history of CVD is still a significantly factor that contributes to the development of CVD [43], which indicates that the CVDs are heritable traits.

Microarray technology allows for the simultaneous exploration of thousands of genes, and the knowledge of the new sequence variations of the genome shed light on the new causal biologic pathways of CHD which may lead to the improvement in the treatment and prevention

of CHD. The completion of the International HapMap Project and the Human Genome Project [15, 44] allows for genome-scale screening to detect the common genome variations contributing to the disease. Recent genome-wide linkage and association studies have successfully identified several genetic loci that are related to CHD[45]. For example through genetic linkage studies genes LDLR[46], APOB and PCSK9 [47, 48] have been identified to be associated with Mendelian lipid disorders which are well established risk factors contributing to CHD. Genome-wide association studies have also made considerable progress in detecting the risk factors for CHD. For example, the Ottawa Heart Study [49], the deCODE Genetics [50] , and the Wellcome Trust Case-Control Consortium (WTCCC) [51] have identified a significant locus on chromosome 9p21. By the time of publications, no prior genetic studies have pointed out the identified genetic region and the region does not relate to any well known risk factors. Subsequent studies confirmed the association of the locus on chromosome 9p21 with MI and other type of vascular disease[45]. In addition to the locus on chromosome 9p21, the WTCCC study identified several other genetic risk loci, such as rs646776 on chromosome1p13, rs17465637 on chromosome 1q41, rs1746048 on chromosome 10q11, which later have been replicated by other studies[45, 52]. The first GWAS study for plasma lipid concentrations was conducted by the Diabetes Genetics Initiative, and the SNP for low-density lipoprotein cholesterol near the APOE gene, the SNP for high-density lipoprotein cholesterol near the CETP, and the SNP for triglycerides in an intron of GCKR gene have been successfully identified[53]. The subsequent GWASs on plasma lipid concentrations have identified loci that are located near the well-known lipid regulators [45, 51]. A number of loci associated with other risk factors of CHD also have been identified by GWASs [54-57], and recently GWAS has been applied to several emerging risk factors for CHD, such as fibrinogen, inflammatory biomarkers [58, 59].

Recently substantial evidences from both clinical and epidemiological studies suggest a considerable amount of comorbidity exist between cardiovascular disease and type II diabetes. For example, Martin *et al*. reported that in a German cohort, at the time of diagnosis of type II diabetes 22% of patients had coronary heart disease present [60]. According to the American Diabetes Association, an estimates of two third of diabetic people died from cardiovascular disease, and adults with diabetes are at least twice as likely to have heart disease or stroke than those without diabetes. Indeed, the American Heart Association recommends treating diabetes as one of the major controllable risk factors for cardiovascular disease[61, 62]. Various factors can determine the co-occurrences of the two diseases, ranging from genetic predisposition and lifestyle of individual to the general health policy on the public. According to the 13 co-morbidity models proposed by Neale and Kendler, co-morbidity between coronary heart disease and type II diabetes may be due to the fact that one of the co-morbid conditions is the cause or consequence of the other [63-65]. It is also possible that the two diseases share the same or correlated risk factors, such as obesity, physical inactivity, and insulin resistance, making the co-morbid conditions more likely to occur simultaneously [25, 34, 64, 66-73].

Though a large proportion of coronary heart disease cases and type II diabetes cases can be explained by environmental factors, genetic factors also play an important role in predisposing to both diseases. It has been reported that single nucleotide polymorphism (SNP) rs1801282, which is located in gene PPARG, is associated with both cardiovascular disease and type II diabetes [54, 74, 75]. G allele of SNP rs4420638, which is located 14kb away from ApoC1 gene and co-inherited with ApoE, increases the risk of coronary heart disease as well as type II diabetes[53, 76]. The chromosome 9p21 region has also been identified to be associated

with both type 2 diabetes and cardiovascular disease, though different SNPs were reported to each disease in different studies[22, 77].

Though remarkable progress has been made in the past few decades, the current findings of genetic susceptibility genes can only explain a small part of the heredity of CHD and the co-morbidity between type II diabetes, highlighting the need of further exploration of CHD etiology and mapping the susceptibility loci on the genome.

1.1.2 Cervical Cancer

Cervical cancer is the second most common type of cancer [78] and is one of the leading causes of cancer-related death in women worldwide[79], especially in the low-income developing countries. It is estimated that 500,000 new cases arise every year and 80% occurs in developing countries [80-82]. In the US, based on the data from Surveillance Epidemiology and End Results (SEER) [83] , the age-adjusted incidence rate was 8.1 per 100,000 and the age-adjusted death rate was 2.4 per 100,000 in women for all racial groups. Hispanic women had the highest incidence rate (12 per 100,000), followed by Black, White, American Indian/Alaska Native and Asian/Pacific Islander, while the Black had the highest death rate (4.4 per 100,000), followed by American Indian/Alaska Native, Hispanic, White, and Asian/Pacific Islander. Both death rate and incidence of cervical cancer decrease over years, which are largely due to the regular Pap smear screening [83, 84]. From 2003 to 2007, the median diagnostic age of cervical cancer was 48 years old and the median death age of cervical cancer was 57 years old. The 5-year relative survival rates were inversely related to the stage of the disease, with only 17% 5-year relative survival for women with cervical cancer that has metastasized [83].

It has been generally accepted that Human papillomavirus (HPV) infection with high-risk types is a necessary factor to cause cervical cancer [82, 85]. HPV infection is a very common infection in the US, and more than 6 millions of new HPV infections occur each year. To date more than 150 types of HPV have been identified, and among them types 16 and 18 contribute to 70% of cervical cancer cases [85]. Numerous epidemiological studies consistently suggest that sexual activity including the number of sexual patterns, age of the first sexual activity and sexual activity of the partner is highly associated with cervical cancer [86-90]. In addition to sexual activity, several studies have shown that cigarette smoking [91], the number of live births[92], and diet are factors[93-96] that are associated with cervical cancer as well. Evidences suggest that women who smoke double the chance of developing cervical cancer. Diet high in vegetable and fruits are associated with a 54% decreased risk of persistence of HPV, which leads to a decreased risk of developing cervical cancer. Studies also have shown that a high intake in vitamin C and beta carotene may reduce the risk of cervical cancer, and diet might be one of the factors that explain the between-country difference in cervical cancer incidence rate [87].

Though radical breakthrough has been made to understand risk factors of cervical cancer, the etiology at molecular level largely remains unknown. Evidences suggest that copy number increases on chromosome 20 at 20q11.2 and 20q13.1, and it is highly related to the stage of cervical cancer [97]. The survival rate of cervical cancer at advanced stage is significantly lower than the localized stage[84], and the failure to the treatment of advanced stage cervical cancer is partly due to the lack of understanding of the etiology of cervical cancer at the molecular level [97]. Further investigation of the mechanisms of cervical cancer at molecular level may help to improve treatment, which may reduce death rate of cervical cancer, especially for advanced stage cases.

1.1.3 Limitation of the Current Methods

With microarray technology, extensive studies have been conducted to understand the genetic etiology contributing to coronary heart disease and cervical cancer, but only a small portion of the cases can be explained by the identified risk loci or genetic regions. The artifacts and noise in measured intensity of microarray may attenuate the effect of risk loci/biomarker or increase the effect of false-positive loci/biomarker, which leads to insufficient or invalid association results. Current methods in microarray data analysis mostly focus on developing association tests without careful consideration of the uncertainty introduced by the microarray experiment, but extensive evidences suggest that experimental conditions such as batch size and microarray probe sequence design can largely confound the association studies. A considerable amount of risk loci contributing to either of the type II diabetes or coronary heart disease have been identified, but the underlying mechanism leading to the co-morbidity remains largely unknown. To the best of my knowledge, currently there is no statistical method that is capable of identifying risk factors contributing to co-morbid diseases with consideration of joint gene-gene actions. Identification of predisposing genetic variants and environmental factors common to the co-occurrence of diseases, unique to each co-morbid condition is of great importance to clinicians, researchers as well as the general public, as it helps elucidate the causes of co-morbidity and promotes new diagnostic and therapeutic strategies for the diseases.

1.2 Objectives

The fast development of biotechnologies enables us to scan the entire genome and profile thousands of genes simultaneously in large scale epidemiological studies[98]. Marvelous progress has been made in identifying genetic susceptibility loci, differentially expressed genes

and teasing apart biological pathways in the past few decades through both candidate gene and genome-wide association study (GWAS) approaches[22, 99-104]. However, the current findings can only explain part of the disease susceptibility, and it is still a great challenge to understand the etiology of complex human diseases. The objective of this research is to develop statistical methods that aim at detecting various genetic susceptibility loci and differentially expressed genes for complex human diseases of high impact, including cardiovascular diseases and cervical cancer. This research holds great promise for identifying mechanism of disease and promoting the development of new interventions. The specific aims are:

1. Develop a Statistical Method for Accurate Genotype Calling and Apply the Proposed Method to a Coronary Heart Disease Study

The first aim of this research is to develop an accurate genotype calling method based on Probe Intensity Composition Representation (referred to as PICR[105]) with Empirical Bayesian and Normal Mixture models, and applies it to genome wide association study to detect risk loci for coronary artery disease (CAD) which may help to unravel more susceptibility loci and elucidate the genetic etiology of CAD.

2. Develop a Statistical Method for Detecting Genetic Risk Factors for Co-morbidity between Cardiovascular Disease and Type II Diabetes

The second aim of this research is to develop a statistical method that can detect genetic architecture between co-morbid diseases with consideration of high order gene-gene interaction effects. The new proposed method is applied to datasets from *Wellcome Trust Case-Control Consortium* (WTCCC) to investigate the co-morbidity between CAD and type II diabetes (T2D), which may shed light on new therapeutic strategies that are effective for both diseases.

3. Develop a Statistical Method for Detecting Differential Expressed Genes and Apply the Proposed Method to a Cervical Cancer Study

The third aim of this research is to develop a method that can preprocess and analyze microarray data simultaneously with consideration of the variability introduced in data preprocessing step. The proposed method uses Positional Dependent Nearest Neighbor (referred to as PDNN) model and Empirical Bayesian model and is applied to a gene profiling study of cervical cancer to unravel more differentially expressed genes, which may help to explain the different survival rate for cervical cancer at different stages.

REFERENCES

REFERENCES

1. Irizarry, R.A., Z. Wu, and H.A. Jaffee, *Comparison of Affymetrix GeneChip expression measures.* Bioinformatics, 2006. **22**(7): p. 789-94.

2. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 31-6.

3. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.

4. Wu, Z.J., et al., *A model-based background adjustment for oligonucleotide expression arrays.* Journal of the American Statistical Association, 2004. **99**(468): p. 909-917.

5. Zhang, L., M.F. Miles, and K.D. Aldape, *A model of molecular interactions on short oligonucleotide microarrays.* Nat Biotechnol, 2003. **21**(7): p. 818-21.

6. Astrand, M., *Contrast normalization of oligonucleotide arrays.* J Comput Biol, 2003. **10**(1): p. 95-102.

7. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.

8. Baldi, P. and A.D. Long, *A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes.* Bioinformatics, 2001. **17**(6): p. 509-19.

9. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.

10. Sartor, M.A., et al., *Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments.* BMC Bioinformatics, 2006. **7**: p. 538.

11. Kadota, K., Y. Nakai, and K. Shimizu, *A weighted average difference method for detecting differentially expressed genes from microarray data.* Algorithms Mol Biol, 2008. **3**: p. 8.

12. Breitling, R., et al., *Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.* FEBS Lett, 2004. **573**(1-3): p. 83-92.

13. Carvalho, B., et al., *Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data.* Biostatistics, 2007. **8**(2): p. 485-99.

14. Carvalho, B.S., T.A. Louis, and R.A. Irizarry, *Quantifying uncertainty in genotype calls.* Bioinformatics, 2010. **26**(2): p. 242-9.

15. Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.

16. Affymetrix, *Birdseed Algorithm – Affymetrix Genotyping Console Software 2.0*. 2007, Affymetrix, Inc: Santa Clara, CA.

17. Korn, J.M., et al., *Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs.* Nature genetics, 2008. **40**(10): p. 1253-60.

18. Riala, K., et al., *Teenage smoking and substance use as predictors of severe alcohol problems in late adolescence and in young adulthood.* J Adolesc Health, 2004. **35**(3): p. 245-54.

19. Fox, C.S., et al., *Temporal trends in coronary heart disease mortality and sudden cardiac death from 1950 to 1999: the Framingham Heart Study.* Circulation, 2004. **110**(5): p. 522-7.

20. Lloyd-Jones, D., et al., *Executive summary: heart disease and stroke statistics--2010 update: a report from the American Heart Association.* Circulation, 2010. **121**(7): p. 948-54.

21. Rosamond, W., et al., *Heart disease and stroke statistics--2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee.* Circulation, 2008. **117**(4): p. e25-146.

22. Wellcome Trust Case Control Consortium, *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.

23. Lloyd-Jones, D.M., et al., *Lifetime risk of developing coronary heart disease.* Lancet, 1999. **353**(9147): p. 89-92.

24. Thom, T., et al., *Cardiovascular disease in the Unite States and preventive approaches*, in *Hurst's The Heart, Arteries and Veins.*, V. Fuster, et al., Editors. 2001, McGraw-Hill: New York, Y.

25. Stamler, J., *Established major risk factors.*, in *Coronary heart disease epidemiology*, M. Marmot and P. Elliot, Editors. 1992, Oxford Univeristy Press: New York, NY.

26. Balady, G.J., et al., *Usefulness of exercise testing in the prediction of coronary disease risk among asymptomatic persons as a function of the Framingham risk score.* Circulation, 2004. **110**(14): p. 1920-5.

27. Danesh, J., R. Collins, and R. Peto, *Lipoprotein(a) and coronary heart disease. Meta-analysis of prospective studies.* Circulation, 2000. **102**(10): p. 1082-5.

28. Greenland, P., et al., *Major risk factors as antecedents of fatal and nonfatal coronary heart disease events.* JAMA, 2003. **290**(7): p. 891-7.

29. Kannel, W.B., *New perspectives on cardiovascular risk factors.* Am Heart J, 1987. **114**(1 Pt 2): p. 213-9.

30. Kuller, L.H., et al., *10-year follow-up of subclinical cardiovascular disease and risk of coronary heart disease in the Cardiovascular Health Study.* Arch Intern Med, 2006. **166**(1): p. 71-8.

31. Law, M.R., N.J. Wald, and S.G. Thompson, *By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease?* BMJ, 1994. **308**(6925): p. 367-72.

32. Schreiner, P.J., et al., *Race and gender differences in the association of Lp(a) with carotid artery wall thickness. The Atherosclerosis Risk in Communities (ARIC) Study.* Arterioscler Thromb Vasc Biol, 1996. **16**(3): p. 471-8.

33. Smolders, B., R. Lemmens, and V. Thijs, *Lipoprotein (a) and stroke: a meta-analysis of observational studies.* Stroke, 2007. **38**(6): p. 1959-66.

34. Vasan, R.S., et al., *Relative importance of borderline and elevated levels of coronary heart disease risk factors.* Ann Intern Med, 2005. **142**(6): p. 393-402.

35. Yusuf, S., et al., *Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study.* Lancet, 2004. **364**(9438): p. 937-52.

36. Tyroler, H.A., *Coronary heart disease epidemiology in the 21st century.* Epidemiol Rev, 2000. **22**(1): p. 7-13.

37. Hamsten, A. and U. de Faire, *Risk factors for coronary artery disease in families of young men with myocardial infarction.* Am J Cardiol, 1987. **59**(1): p. 14-9.

38. Bensen, J.T., et al., *Family history of coronary heart disease and pre-clinical carotid artery atherosclerosis in African-Americans and whites: the ARIC study: Atherosclerosis Risk in Communities.* Genet Epidemiol, 1999. **16**(2): p. 165-78.

39. Myers, R.H., et al., *Parental history is an independent risk factor for coronary artery disease: the Framingham Study.* Am Heart J, 1990. **120**(4): p. 963-9.

40. Gliksman, M.D., et al., *Childhood socioeconomic status and risk of cardiovascular disease in middle aged US women: a prospective study.* J Epidemiol Community Health, 1995. **49**(1): p. 10-5.

41. Perusse, L., et al., *Genetic and environmental influences on level of habitual physical activity and exercise participation.* Am J Epidemiol, 1989. **129**(5): p. 1012-22.

42. Wannamethee, S.G., et al., *Influence of fathers' social class on cardiovascular disease in middle-aged men.* Lancet, 1996. **348**(9037): p. 1259-63.

43. Lloyd-Jones, D.M., et al., *Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring.* JAMA, 2004. **291**(18): p. 2204-11.

44. *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.

45. Musunuru, K. and S. Kathiresan, *Genetics of coronary artery disease.* Annu Rev Genomics Hum Genet, 2010. **11**: p. 91-108.

46. Goldstein, J.L. and M.S. Brown, *The LDL receptor.* Arterioscler Thromb Vasc Biol, 2009. **29**(4): p. 431-8.

47. Abifadel, M., et al., *Mutations in PCSK9 cause autosomal dominant hypercholesterolemia.* Nature genetics, 2003. **34**(2): p. 154-6.

48. Soria, L.F., et al., *Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100.* Proc Natl Acad Sci U S A, 1989. **86**(2): p. 587-91.

49. McPherson, R., et al., *A common allele on chromosome 9 associated with coronary heart disease.* Science, 2007. **316**(5830): p. 1488-91.

50. Helgadottir, A., et al., *A common variant on chromosome 9p21 affects the risk of myocardial infarction.* Science, 2007. **316**(5830): p. 1491-3.

51. Samani, N.J., et al., *Genomewide association analysis of coronary artery disease.* N Engl J Med, 2007. **357**(5): p. 443-53.

52. Kathiresan, S., et al., *Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants.* Nature genetics, 2009. **41**(3): p. 334-41.

53. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science, 2007. **316**(5829): p. 1331-6.

54. Zeggini, E., et al., *Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.* Science, 2007. **316**(5829): p. 1336-41.

55. Unoki, H., et al., *SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations.* Nature genetics, 2008. **40**(9): p. 1098-102.

56. Yasuda, K., et al., *Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus.* Nature genetics, 2008. **40**(9): p. 1092-7.

57. Zeggini, E., et al., *Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.* Nature genetics, 2008. **40**(5): p. 638-45.

58. Danik, J.S., et al., *Novel loci, including those related to Crohn disease, psoriasis, and inflammation, identified in a genome-wide association study of fibrinogen in 17 686 women: the Women's Genome Health Study.* Circ Cardiovasc Genet, 2009. **2**(2): p. 134-41.

59. Dehghan, A., et al., *Association of novel genetic Loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts.* Circ Cardiovasc Genet, 2009. **2**(2): p. 125-33.

60. Martin, S., et al., *Epidemiology of complications and total treatment costs from diagnosis of Type 2 diabetes in Germany (ROSSO 4).* Exp Clin Endocrinol Diabetes, 2007. **115**(8): p. 495-501.

61. American Diabetes Association, *Consensus Development Conference on the Diagnosis of Coronary Heart Disease in People with Diabetes.* Diabetes Care, 1988. **21**: p. 1551–1559.

62. American Heart Association. Available from: http://www.heart.org/HEARTORG/Conditions/Diabetes/WhyDiabetesMatters/Cardiovascular-Disease-Diabetes_UCM_313865_Article.jsp.

63. de Groot, V., et al., *How to measure comorbidity. a critical review of available methods.* J Clin Epidemiol, 2003. **56**(3): p. 221-9.

64. Neale, M.C. and K.S. Kendler, *Models of comorbidity for multifactorial disorders.* Am J Hum Genet, 1995. **57**(4): p. 935-53.

65. Simonoff, E., *Extracting meaning from comorbidity: genetic analyses that make sense.* J Child Psychol Psychiatry, 2000. **41**(5): p. 667-74.

66. Youngstrom, E.A., L.E. Arnold, and T.W. Frazier, *Bipolar and ADHD Comorbidity: Both Artifact and Outgrowth of Shared Mechanisms.* Clin Psychol (New York), 2010. **17**(4): p. 350-359.

67. Lind, P.A., et al., *A genomewide association study of nicotine and alcohol dependence in Australian and Dutch populations.* Twin Res Hum Genet, 2010. **13**(1): p. 10-29.

68. Travers, M.E. and M.I. McCarthy, *Type 2 diabetes and obesity: genomics and the clinic.* Hum Genet, 2011. **130**(1): p. 41-58.

69. McCarthy, M.I., *Genomics, type 2 diabetes, and obesity.* N Engl J Med, 2010. **363**(24): p. 2339-50.

70. Johnson, R.L., S.M. Williams, and I.J. Spruill, *Genomics, nutrition, obesity, and diabetes.* J Nurs Scholarsh, 2006. **38**(1): p. 11-8.

71. Reaven, G.M. and Y.D. Chen, *Insulin resistance, its consequences, and coronary heart disease. Must we choose one culprit?* Circulation, 1996. **93**(10): p. 1780-3.

72. Schernthaner, G., *[Hypertension, insulin resistance and diabetes mellitus: pathophysiological interactions and therapeutic consequences]*. Wien Klin Wochenschr, 1990. **102**(24): p. 707-12.

73. Smit, J.W. and J.A. Romijn, *Acute insulin resistance in myocardial ischemia: causes and consequences.* Semin Cardiothorac Vasc Anesth, 2006. **10**(3): p. 215-9.

74. Bego, T., et al., *Association of PPARG and LPIN1 gene polymorphisms with metabolic syndrome and type 2 diabetes.* Med Glas Ljek komore Zenicko-doboj kantona, 2011. **8**(1): p. 76-83.

75. Regieli, J.J., et al., *PPAR gamma variant influences angiographic outcome and 10-year cardiovascular risk in male symptomatic coronary artery disease patients.* Diabetes Care, 2009. **32**(5): p. 839-44.

76. Willer, C.J., et al., *Newly identified loci that influence lipid concentrations and risk of coronary artery disease.* Nature genetics, 2008. **40**(2): p. 161-9.

77. Broadbent, H.M., et al., *Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p.* Hum Mol Genet, 2008. **17**(6): p. 806-14.

78. Armstrong, E.P., *Prophylaxis of cervical cancer and related cervical disease: a review of the cost-effectiveness of vaccination against oncogenic HPV types.* J Manag Care Pharm, 2010. **16**(3): p. 217-30.

79. World Health Organization. *Fact sheet No. 297: Cancer.* [cited 2011 Jan 26]; Available from: http://www.who.int/mediacentre/factsheets/fs297/en/index.html.

80. Clifford, G.M., *Global access to HPV vaccination: what are we waiting for?* Lancet, 2009. **374**(9706): p. 1948-9.

81. Kent, A., *HPV Vaccination and Testing.* Rev Obstet Gynecol, 2010. **3**(1): p. 33-4.

82. Parkin, D.M., *The global health burden of infection-associated cancers in the year 2002.* Int J Cancer, 2006. **118**(12): p. 3030-44.

83. Howlader, N., et al., *SEER Cancer Statistics Review, 1975-2008,* B. Edwards, Editor. 2011, National Cancer Institute.

84. Canavan, T.P. and N.R. Doshi, *Cervical cancer.* Am Fam Physician, 2000. **61**(5): p. 1369-76.

85. Walboomers, J.M., et al., *Human papillomavirus is a necessary cause of invasive cervical cancer worldwide.* J Pathol, 1999. **189**(1): p. 12-9.

86. Franco, E.L., et al., *Correlation patterns of cancer relative frequencies with some socioeconomic and demographic indicators in Brazil: an ecologic study.* Int J Cancer, 1988. **41**(1): p. 24-9.

87. Franco, E.L., E. Duarte-Franco, and A. Ferenczy, *Cervical cancer: epidemiology, prevention and the role of human papillomavirus infection.* CMAJ, 2001. **164**(7): p. 1017-25.

88. Graham, S., et al., *Genital cancer in wives of penile cancer patients.* Cancer, 1979. **44**(5): p. 1870-4.

89. Li, J.Y., et al., *Correlation between cancers of the uterine cervix and penis in China.* J Natl Cancer Inst, 1982. **69**(5): p. 1063-5.

90. Schiffman, M.H. and L.A. Brinton, *The epidemiology of cervical carcinogenesis.* Cancer, 1995. **76**(10 Suppl): p. 1888-901.

91. Winkelstein, W., Jr., *Smoking and cervical cancer--current status: a review.* Am J Epidemiol, 1990. **131**(6): p. 945-57; discussion 958-60.

92. Brinton, L.A., et al., *Sexual and reproductive risk factors for invasive squamous cell cervical cancer.* J Natl Cancer Inst, 1987. **79**(1): p. 23-30.

93. Giuliano, A.R., et al., *Dietary intake and risk of persistent human papillomavirus (HPV) infection: the Ludwig-McGill HPV Natural History Study.* J Infect Dis, 2003. **188**(10): p. 1508-16.

94. Herrero, R., et al., *A case-control study of nutrient status and invasive cervical cancer. I. Dietary indicators.* Am J Epidemiol, 1991. **134**(11): p. 1335-46.

95. Sedjo, R.L., et al., *Human papillomavirus persistence and nutrients involved in the methylation pathway among a cohort of young women.* Cancer Epidemiol Biomarkers Prev, 2002. **11**(4): p. 353-9.

96. Yeo, A.S., et al., *Serum micronutrients and cervical dysplasia in Southwestern American Indian women.* Nutr Cancer, 2000. **38**(2): p. 141-50.

97. Scotto, L., et al., *Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression.* Genes Chromosomes Cancer, 2008. **47**(9): p. 755-65.

98. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nat Methods, 2008. **5**(1): p. 16-8.

99. DeRisi, J.L., V.R. Iyer, and P.O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 1997. **278**(5338): p. 680-6.

100. Lee, C.K., et al., *Gene expression profile of aging and its retardation by caloric restriction.* Science, 1999. **285**(5432): p. 1390-3.

101. Ly, D.H., et al., *Mitotic misregulation and human aging.* Science, 2000. **287**(5462): p. 2486-92.

102. Barrett, J.C., et al., *Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.* Nature genetics, 2008. **40**(8): p. 955-62.

103. Frayling, T.M., *Genome-wide association studies provide new insights into type 2 diabetes aetiology.* Nature reviews. Genetics, 2007. **8**(9): p. 657-62.

104. Adeyemo, A., et al., *A genome-wide association study of hypertension and blood pressure in African Americans.* PLoS genetics, 2009. **5**(7): p. e1000564.

105. Wan, L., et al., *Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation.* Nucleic Acids Res, 2009. **37**(17): p. e117.

CHAPTER 2

MA-SNP — A NEW GENOTYPE CALLING METHOD FOR OLIGONUCLEOTIDE SNP ARRAYS MODELING THE BATCH EFFECT WITH A NORMAL MIXTURE MODEL

2.1 Abstract

Genome-wide association studies hold great promise in identifying disease-susceptibility variants and understanding the genetic etiology of complex diseases. Microarray technology enables the genotyping of millions of single nucleotide polymorphisms. Many factors in microarray studies, such as probe selection, sample quality, and experimental procedure and batch, have substantial effect on the genotype calling accuracy, which is crucial for downstream analyses. Failure to account for the variability of these sources may lead to inaccurate genotype calls and false positive and false negative findings. In this study, we develop a SNP-specific genotype calling algorithm based on the probe intensity composite representation (PICR) model, while using a normal mixture model to account for the variability of batch effect on the genotype calls. We demonstrate our method with SNP array data in a few studies, including the HapMap project, the coronary heart disease and the UK Blood Service Control studies by the Wellcome Trust Case-Control Consortium, and a methylation profiling study. Our single array based approach outperforms PICR, which is also a single array based genotype calling algorithm, and is comparable to the best multi-array genotype calling methods.

Keywords: Affymetrix, genotyping, hierarchical model, hybridization, normal mixture model

2.2 Introduction

Genome-wide association study (GWAS) holds great promises in identifying genetic regions and variants that contribute to complex human diseases and understanding of disease etiology. Recent GWASs have identified numerous novel disease-susceptibility loci for common diseases, such as coronary artery disease [1], Crohn's disease [1, 2], hypertension [1, 3], and Type 2 diabetes [1, 4]. Microarray technology allows simultaneous genotyping of millions of single nucleotide polymorphisms (SNPs) across the entire genome [5, 6]. Affymetrix SNP arrays are among the most widely used array platforms in GWAS studies [6]. In general, two alleles are observed for each SNP (typically referred to as allele A and allele B). Genotype calling algorithms provide an estimate of the SNP genotype (AA, AB or BB assuming no copy number variation) and a corresponding confidence measure. Several statistical algorithms have been proposed and achieved relatively high accuracy in genotype calling, including RLMM[7], BRLMM[8], CRLMM[9, 10], CHIAMO [11], BIRDSEED [12, 13], BEAGLE [14]. Most of the calling methods need preprocessing of the array raw intensities and require multiple samples [15, 16] depending on the between-array normalization procedure used. In addition, little attention has been paid to the underlying mechanisms of the hybridization process of microarray which may affect the accuracy of genotype calls. Extensive studies have shown that probe intensities of oligonucleotide arrays depend on not only the concentration of the target sequence but also the binding affinity of the probes[17-20]. For example, RLMM, BRLMM and CRLMM first preprocess the observed raw intensities with quantile normalization, and then fit a robust linear model to the normalized intensities before the final clustering procedure using the Mahalanobis distance. Both BRLMM and CRLMM apply a Bayesian approach to account for the variability introduced by low minor allele frequencies. In contrast to RLMM and BRLMM, the CRLMM

utilizes the GC content and DNA fragment length information, which has been shown to be highly important in microarray hybridization process, to remove artifacts and achieves higher accuracy [21]. It highly suggests that genotype calling accuracy can be improved by modeling the mechanism of the array underlying hybridization process.

In a recent work, we studied a single array genotyping approach - the probe intensity composite representation (PICR) model[18], herein referred to as the PICR. The PICR makes genotype calls by decomposing the observed probe intensities into allelic target concentration obtained from specific binding signals, and the SNP-specific background obtained from nonspecific binding signals. It utilizes the probe sequence information, which remains the same and is independent of samples and laboratories, and models the physico-chemical properties of probe binding through probe sequence structure. The PICR yields accurate genotype calls consistently across samples, experiments and array platforms, and performs well in comparison to existing multi-array based methods, such as BRLMM and CRLMM. The single array approach is well suited to small studies. Though the physico-chemical properties of probe binding between the probe sequence and target sequence can be largely modeled with the probe sequence structure, the complicated hybridization process may also be influenced by certain unknown factors. This is indicated by the observation that different SNPs behave slightly differently [7, 9]. Because PICR applies a universal genotype calling criteria to all SNPs, it may fail to take into account SNP-specific factors that influence the intensities. Moreover, PICR does not provide a confidence measure for each genotype calls, which is needed to identify inefficient or invalid annotation that may further lead to invalid association findings [10, 22]. Indeed a valid confidence measure for the genotype uncertainty is of great importance for further analysis [10, 22]. During the last two years, numerous studies have reported the effect of batch size and

composition on the genotype calling accuracies [10, 23, 24]. Though the random non-biological signals can be largely removed by SNP-specific background according to PICR, the batch specific factors may affect the hybridization process systematically, and lead to biased results. Failure to consider the potential batch effect may lower the genotype call rate and the calling accuracies. In addition, when batch is confounded with the study outcome, statistical associations may be artifacts of the experimental design/processing [25, 26].

In this research, we develop a novel SNP-specific genotype calling algorithm and quality control criteria for each SNP based on the PICR model (referred to as MA-SNP). The MA-SNP has the following advantages: It 1) makes genotype calls based on only individual data; 2) applies to small sample studies and is robust across samples; 3) provides an efficient method to correct for batch effect; 4) yields standardized target sequence concentration that can directly be used for copy number variation studies. The rest of this chapter is arranged as follows. In Section 2.3, we first give a brief review of PICR and the related statistical problems, and then outline a SNP-specific genotype model and parameter training procedure. We further construct a quality measure for assigned genotype calls, and develop a model for the potential batch effect correction. In Section 2.4, we illustrate our method with microarray data in three studies, including the data set from the HapMap project [27], the coronary artery disease (CAD) and the UK Blood Service Control (NBS) data sets from the Wellcome Trust Case Control study [1], and the data set from methylation profiling study [28]. In the last section of this chapter, we discuss and summarize our findings.

2.3 The Model

2.3.1 Decomposing raw intensities into biological signals and noise

Microarray data consist of biological signals of research interest, noise in the probe intensity measurement, and artifacts due to experimental procedure and design of the technology. For accurate genotype calls, it is imperative that the observed probe intensities be decomposed into biological signals and non-biological components so that the genotype estimation is not severely biased by the noise and artifacts. Our former PICR model, which can be applied to Affymetrix Mapping 100K Array and Mapping 500K Arrays, estimates the biological signals by studying the underlying mechanism of hybridization between the probe sequence and target sequence through the calculation of the binding free energy with the generalized positional-dependent-nearest-neighbor (GPDNN) models[18]. The allelic target concentrations (denoted by N_A and N_B) are then estimated by PICR decomposition of probe intensity through equation (1),

$$\begin{cases} \vdots \\ I_{PA} = b + N_A f^{\{S^{PA}, S^{TA}\}} + N_B f^{\{S^{PA}, S^{TB}\}} + \varepsilon_{PA} \\ I_{PB} = b + N_A f^{\{S^{PB}, S^{TA}\}} + N_B f^{\{S^{PB}, S^{TB}\}} + \varepsilon_{PB} \\ I_{MA} = b + N_A f^{\{S^{MA}, S^{TA}\}} + N_B f^{\{S^{MA}, S^{TB}\}} + \varepsilon_{MA} \\ I_{MB} = b + N_A f^{\{S^{MB}, S^{TA}\}} + N_B f^{\{S^{MB}, S^{TB}\}} + \varepsilon_{MB} \\ \vdots \end{cases} \quad (1)$$

where *I*s are the observed raw probe intensities of a given SNP; *b* is the SNP- specific baseline, and *f* is a function of binding free energy of a perfect match probe or mismatch probe (denoted

by S^P and S^M, respectively) and its corresponding target (denoted by S^T) for a given SNP probe set. The parameters representing allelic target sequence concentrations (N_A and N_B) and SNP-specific baseline (b) are estimated through linear regression. Since biological signal is contained in the target sequence concentrations only[18], only the allelic target concentrations are used for subsequent analysis, including genotyping, association test and copy number variation studies.

The above PICR model not only removes the artifacts of unequal footing of the array intensity across the samples as discussed in Wan *et al.* [18], but also removes the genomic wave artifacts[29]. The former is usually taken care of by a between-array normalization procedure, while the latter requires special techniques and procedures [29-31]. Unlike other models such as CRLMM and BRLMM, PICR makes genotype calling based on allelic target concentrations (i.e. N_A,N_B) instead of normalized intensities (i.e. *I*), which makes PICR genotype calling procedure robust as the majority of the artifact has been taken care of by the SNP-specific baseline (i.e. *b*).

Although PICR provides a single array based genotype calling algorithm, confidence scores that quantify the quality of genotype calls across SNPs are not available. While the physico-chemical properties of probe sequence binding can be largely modeled by probe sequence, the complicated hybridization process may also be affected by other unknown factors, which may potentially lead to biased estimation of the allelic concentrations. The estimated allelic concentrations may deviate slightly from their expected values (Figure 2.1). We are thus motivated to take the SNP specific factors into consideration aiming to improve the genotype calling accuracy. Compared to the universal genotype calling criteria to all SNPs across all batches and labs by the PICR (i.e. PICR applies 1/3 and 2/3 to N_A/N_B ratio as genotype calling

criteria for all SNPs, and this is equivalent to -0.5 and 0.5 when the genotype is made based on (N_A-N_B)/ (N_A+N_B) ratio), our new approach also considers the batch effect to further improve the genotype calling accuracy and genotype call rate.

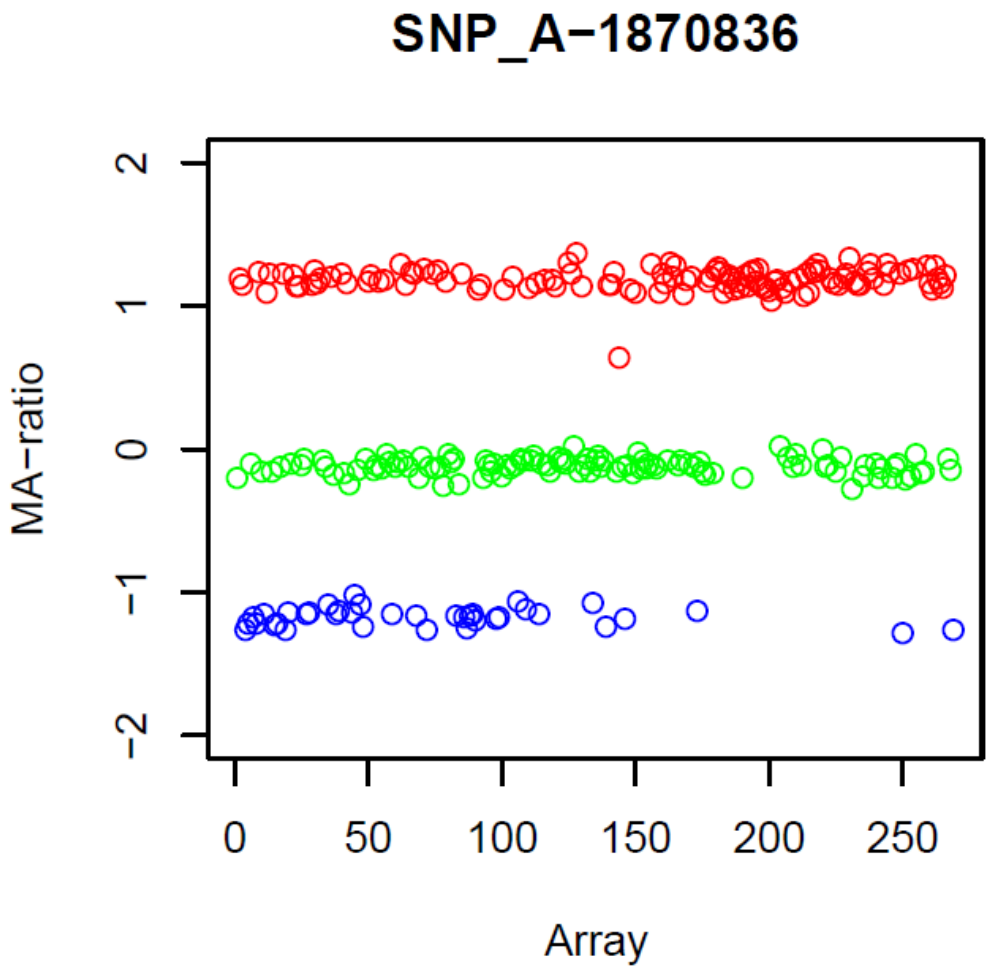


Figure 2.1 Plot of MA-ratio of four SNPs illustrates the SNP to SNP variability for genotype cluster centers using HapMap data. (AA: red; AB: green; BB: blue. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation).

2.3.2 The hierarchical model

The allelic copy numbers (N_A, N_B) estimated from PICR are standardized by their variance-covariance matrix, and the standardized allelic copy numbers ($N_{A_{sd}}$, $N_{B_{sd}}$) are expected to be normally distributed and independent of each other. We then define the MA-ratio, a quantity measuring the signal ratio between the two alleles of the SNP:

$$R = \frac{N_{A_{sd}} - N_{B_{sd}}}{N_{A_{sd}} + N_{B_{sd}}} \quad (2)$$

If no factors other than the probe sequence affect the hybridization process, the MA-ratios are expected to be 1, 0, and -1, for genotypes AA, AB, and BB respectively. This is based on the rationale that the expected value for $N_{A_{sd}}$ and $N_{B_{sd}}$ is 0 for the genotype BB and AA, respectively. In addition, the expected values of $N_{A_{sd}}$ and $N_{B_{sd}}$ would be close to each other, if not the same, for a heterozygous SNP of the genotype AB. However, given the complex hybridization process, other factors besides probe sequence may also likely influence the physico-chemical process of the sequence binding. Therefore, we construct the following hierarchical model to estimate the SNP-specific genotype cluster center to allow for SNP-specific deviation from $\boldsymbol{\mu} = (\mu_{AA}, \mu_{AB}, \mu_{BB})' = (1, 0, -1)'$.

$$(R_{i,j} \mid G_{i,j} = k, \mu_k, m_{i,k}) \sim N(\mu_k + m_{i,k}, \sigma^2_{i,k}) \qquad (3)$$
$$\mathbf{m_i} \sim N(\mathbf{0}, \mathbf{V})$$

where $R_{i,j}$ is the MA-ratio for SNP *i* of sample *j*; *k*=AA, AB or BB, represents the genotype;

$\mathbf{m_i} = (\mathrm{m_{i,AA}}, \mathrm{m_{i,AB}}, \mathrm{m_{i,BB}})'$ represents the SNP-specific deviation from the expected value;

$G_{i,j}$ is the true genotype for SNP *i* of sample *j*, and $\sigma^2_{i,k}$ is the variance of the MA-ratio for SNP *i* with genotype *k*.

We take the HapMap [27] genotype annotation as the gold standard for the parameter training, as the genotype calls from the HapMap project are based on consistent results confirmed by various technologies. Due to the low minor allele frequencies, the availability of all 3 types AA, AB and BB for a given SNP in the HapMap project varies largely from SNP to SNP. To obtain accurate parameter estimates, we follow the parameter training procedure proposed in the CRLMM method, and employ an empirical Bayes approach [9, 10, 32]. An inverse Gamma prior is assumed for the SNP-specific variation.

$$\sigma^2_{i,k} \propto \frac{1}{d_{0,k} s^2_{0,k}} \chi^2_{d_0,k} \qquad (4)$$

The SNP-specific shifts and variances have the following closed forms in equation (5).

$$\hat{m}_{i,k} = \sum_{j \in J_{i,k}} (R_{i,j} - \mu_{i,k}) \Big/ N_{i,k}$$
$$\hat{\sigma}^2_{i,k} = \sum_{j \in J_{i,k}} (R_{i,j} - \mu_{i,k} - \hat{m}_{i,k})^2 \Big/ (N_{i,k} - 1) \qquad (5)$$

where $J_{i,k}$ is the set of samples whose genotype for SNP *i* is *k*; $N_{i,k}$ is the number of samples in the set $J_{i,k}$.

As some genotypes for a given SNP may have very few observations because of low minor allele frequency, we borrow strength across SNPs as in the CRLMM [10], and compute the shrinkage estimates of $\hat{m}_{i,k}$ and $\hat{\sigma}^2_{i,k}$

$$\tilde{\mathbf{m}}_{\mathbf{i}} = (\mathbf{V}^{-\mathbf{1}} + \mathbf{N}_{\mathbf{i}} \Sigma^{-\mathbf{1}})^{-\mathbf{1}} \mathbf{N}_{\mathbf{i}} \Sigma^{-\mathbf{1}} \hat{\mathbf{m}}_{\mathbf{i}}$$
$$\tilde{\sigma}^2_{i,k} = \frac{(N_{i,k} - 1)\hat{\sigma}^2_{i,k} + d_{0,k} s^2_{0,k}}{N_{i,k} - 1 + d_{0,k}} \qquad (6)$$

where ***V*** is estimated with the sample variance-covariance matrix of

$$\mathbf{m}_{\mathbf{i}} = (\mathrm{m}_{\mathrm{i,AA}}, \mathrm{m}_{\mathrm{i,AB}}, \mathrm{m}_{\mathrm{i,BB}})',$$

$$\mathbf{N_i} = (\mathrm{N_{i,AA}}, \mathrm{N_{i,AB}}, \mathrm{N_{i,BB}})' \text{ and } \Sigma = \begin{bmatrix} s^2_{0,AA} & 0 & 0 \\ 0 & s^2_{0,AB} & 0 \\ 0 & 0 & s^2_{0,BB} \end{bmatrix}$$

The hyperparameters $d_{0,k}$ and $s^2_{0,k}$ are estimated following the algorithm proposed by Smyth [32]. The SNP-specific cluster center $\boldsymbol{M_i}$ for genotyping is defined for each SNP *i* by equation (7)

$$\mathbf{M_i} = (\mathrm{m_{i,AA}}, \mathrm{m_{i,AB}}, \mathrm{m_{i,BB}})' + (1,0,-1)' \quad (7)$$

The parameters for the binding free energy in equation (1) remain the same as in the PICR, in which they were trained with a single HapMap array[18]. The SNP-specific cluster center $\boldsymbol{M_i}$ and variances are trained based on 180 HapMap samples, consisting of 60 CEU, 60 YRI, and 60 CHB+JPI samples. The parameters remain the same for the testing data.

2.3.3 SNP-specific genotype calling criteria and quality score for genotype call

The genotype of a given SNP would be assigned *k*, if *k* minimizes $|(\delta_{i,j} | G_{i,j} = k)|$ for *k*=AA, *k*=AB, or *k*=BB, where $\delta_{i,j}$ represents the deviation of $R_{i,j}$ from the cluster center of a given genotype and is defined in equation (8)

$$(\delta_{i,j} | G_{i,j} = k) = (R_{i,j} - M_{i,k} | G_{i,j} = k) \quad (8)$$

For a given genotype call, a large deviation suggests uncertainty of the estimate, and hence the quality score (QS) of SNP *i* in sample *j* with assigned genotype *k* is given by

$$(QS_{i,j} \mid G_{i,j} = k) = \frac{\delta_{i,j} \mid G_{i,j} = k}{\sqrt{\text{var}(\delta_{i,j} \mid G_{i,j} = k)}}$$

$$\begin{aligned} E(\delta_{i,j} \mid G_{i,j} = k) &= E(R_{i,j} - M_{i,k} \mid G_{i,j} = k) \\ &= E_{M_{i,g}}[E(R_{i,j} - M_{i,k} \mid G_{i,j} = k, M_{i,g})] = 0 \end{aligned} \tag{9}$$

$$\begin{aligned} \text{var}(\delta_{i,j} \mid G_{i,j} = k) &= \text{var}([E(R_{i,j} - M_{i,k} \mid G_{i,j} = k, M_{i,g})]) + E[\text{var}(R_{i,j} - M_{i,k} \mid G_{i,j} = k, M_{i,g})] \\ &= \frac{\sigma^2_{i,k}}{N_{i,k}} + \sigma^2_{i,k} \end{aligned}$$

where *k* is the assigned genotype call. Assuming no batch effect, the quality score is symmetric and centered about zero, and a large deviation from zero indicates poor quality of genotype calls. The QS criteria could be set depending on the needs of further analysis, and based on our training data we recommend that the SNPs with $|QS_{i,j}| > 3.56$ should be set aside.

2.3.4 Batch effect

Assuming no batch effect, the quality score is symmetric about zero, and a large systematic deviation from zero for a given SNP may indicate batch effect. The cluster center may shift due to the batch effect as shown in Figure 2.2, which results in low call rate or low accuracy in SNP genotyping. We propose a mixture normal model to update the cluster centers and further correct the potential batch effect. We assume that the MA-ratios for a SNP in a given batch, as defined

below, follow a mixture normal distribution, and model them using a normal prior on the mean and an inverse gamma prior on the variance, as suggested [33].

$$R_{i,j} \mid batch \sim \sum_{k \in \{AA, AB, BB\}} p_k N(M_{i,k,b}, \sigma^2_{i,k,b}), \qquad \sum_{k \in \{AA, AB, BB\}} p_k = 1$$

$$M_{i,k,b} \mid \sigma^2_{i,k,b} \sim N(M_{i,k}, \frac{\sigma^2_{i,k,b}}{\kappa_i}) \tag{10}$$

$$\sigma^2_{i,k,b} \sim inverseGamma(\frac{\upsilon_{i,k}}{2}, \frac{\sigma^2_{i,k}}{2})$$

where $M_{i,k,b}$ and $\sigma^2_{i,k,b}$ are the batch-specific cluster center and variance for SNP_i with genotype *k*, respectively. Specifically, we assign $\upsilon_{i,k} = 3$ and $\kappa_i = 0.1$ for the procedure.

The posterior mean and variance are given by equation (11)[33].

$$\hat{M}_{i,k,b} = \frac{n_k \bar{R}_{i,k} + \kappa_i M_{i,k}}{\kappa_i + n_k}$$

$$\hat{\sigma}^2_{i,k,b} = \frac{\sigma^2_{i,k} + \kappa_i n_k (\bar{R}_{i,k} - M_{i,k})^2 \Big/ (\kappa_i + n_k) + \sum_{j=1}^{n} z_{jk} (R_{i,j} - \bar{R}_{i,k})^2}{\upsilon_{i,k} + n_k + 3} \tag{11}$$

where *n* is the total number of samples; z_{jk} is the conditional probability that SNP *i* of sample *j* belongs to genotype *k*; $n_k = \sum_{j=1}^{n} z_{jk}$ and $\bar{R}_{i,k} = \sum_{j=1}^{n} z_{jk} R_{i,j} / n_k$.

With updated SNP-specific genotype cluster center, the genotyping and quality score calculation follow the same strategy as mentioned in section 2.3.3.

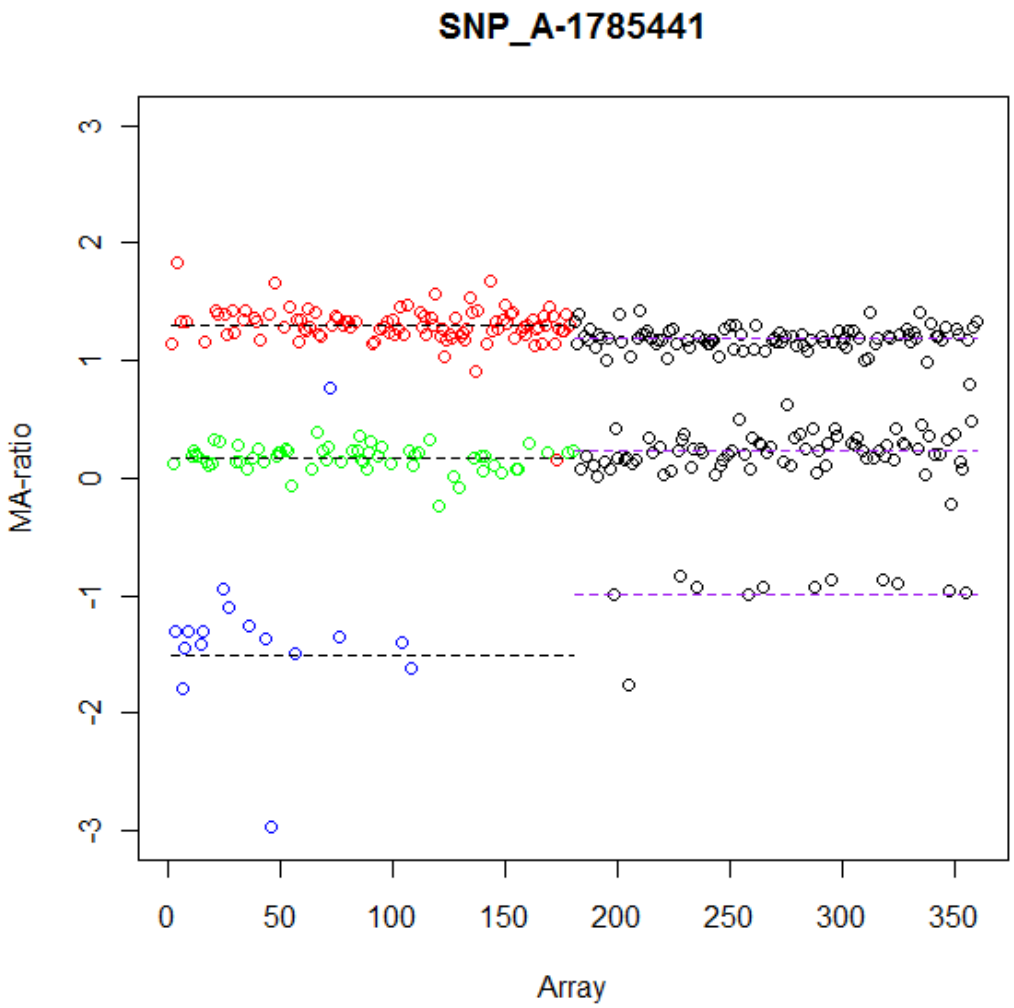


Figure 2.2 Plot of MA-ratio in HapMap samples and Methylation profiling samples of one specific SNP. Red dots: AA genotype HapMap samples; Green dots: AB genotype HapMap samples; Blue dots: BB genotype HapMap samples; Black: samples in Methylation Profiling Study data. Black dashed line: SNP-specific cluster center for the training data; Purple dashed line: SNP-specific cluster center for the Methylation Profiling Study data. The shift of the cluster

center of the MA-ratio of each genotype between the two study data sets, which is the largest among the samples of BB genotype, indicates the between study batch effect.

2.4 Data

Data set I

The HapMap data[27]: This data set contains 270 samples of Affymetrix Mapping 250K Nsp Array and Mapping 250K Sty Array. It was downloaded from ftp.ncbi.nih.gov. The annotations of the corresponding Mapping 500K Array Set were downloaded from Affymetrix Inc website http://www.affymetrix.com/support/support_result.affx. The genotype annotations of the HapMap Project 2009-02_phaseII+III were downloaded from ftp.ncbi.nih.gov.

Data set II

Wellcome Trust Case Control Study data[1]: The Wellcome Trust Case Control Study data sets were acquired from the Wellcome Trust Case-Control Consortium. We use the coronary artery disease (CAD) and the UK Blood Service Control (NBS) from the WTCCC study. The CAD study data set we obtained consists of about 1991 individuals, and the NBS has 1500 control individuals. The two chosen data sets were genotyped on the Affymetrix GeneChip Human Mapping 500K Array Set by the WTCCC.

Data set III

Methylation profiling by Affymetrix SNP array[28]: The data set was downloaded from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20123. The study includes 30 individuals, and the samples consist of DNA from blood, tumor, and normal tissues of the same

individual. The samples with and without Hpa II digestion were annotated using Affymetrix 500K SNP arrays.

2.5 Result

2.5.1 SNP quality control and genotype calling accuracy

Among the 270 HapMap samples, we randomly select 60 samples from each of CEU, YRI, CHB+JPI to serve as the training data, and the remaining 90 samples serve as testing data. We use the HapMap samples for testing because the gold-standard genotype for most of the SNPs allows for the assessment of the accuracy of our genotype calling method. Table 2.1 and Figure 2.3 display the summary statistics of the genotype calling accuracy on the NSP and STY arrays. With 100% call rate, the genotype calling accuracies of the MA-SNP on the 90 testing samples of the NSP array and STY array are 99.34% (SD=0.0081) and 99.30% (SD=0.0130), respectively. We further filter out the SNPs with the quality score |QS|> 3.56. With an average call rate of 99% (SD=0.0144 for NSP array, SD=0.0114 for STY array) the genotyping accuracies for 90 testing samples of the NSP array and STY array are 99.61% (SD=0.0075) and 99.74% (SD=0.0029), respectively. In addition, we compare the MA-SNP with PICR and other genotype calling methods, in particular with the CRLMM method, which has been reported to perform better than the RLMM and BRLMM[21]. We compare the MA-SNP with the CRLMM under two scenarios, and to make a fair comparison between CRLMM and our single array-based MA-SNP, under scenario one, we split the 90 testing arrays into 30 groups with 3 arrays per group. Under scenario two, we use all 90 testing arrays as input for CRLMM. As reported in Table 2.1, MA-SNP outperforms PICR and CRLMM with 3 arrays per run by 1%. With 100% call rate, the accuracy of MA-SNP is slightly lower than CRLMM with all 90 testing arrays.

However, for the SNPs that pass our quality control criteria, the accuracy of MA-SNP is similar to CRLMM with 90 arrays per run.

Table 2.1. Comparison of MA-SNP with PICR and CRLMM in genotype-calling accuracy against the HapMap gold-standard annotation

Samples	Genotype-calling method	Mean	SD[a]	5th	50th	95th
90 Nsp arrays	MA-SNP[b]	0.9961	0.0075	0.9938	0.9977	0.9986
	MA-SNP[c]	0.9934	0.0130	0.9904	0.9957	0.9968
	PICR	0.9856	0.0133	0.9757	0.9885	0.9937
	CRLMM[d]	0.9834	0.0052	0.9730	0.9839	0.9898
	CRLMM[e]	0.9980	0.0013	0.9961	0.9984	0.9990
90 Sty arrays	MA-SNP[b]	0.9974	0.0029	0.9921	0.9984	0.9991
	MA-SNP[c]	0.9952	0.0046	0.9869	0.9968	0.9975
	PICR	0.9902	0.0062	0.9783	0.9919	0.9955
	CRLMM[d]	0.9869	0.0060	0.9781	0.9885	0.9928
	CRLMM[e]	0.9982	0.0021	0.9955	0.9988	0.9991

a: Standard deviation
b: The absolute value of quality score is less than 3.56
c: 100% call rate
d: CRLMM with 3 arrays as input per run
e: CRLMM with 90 arrays as input

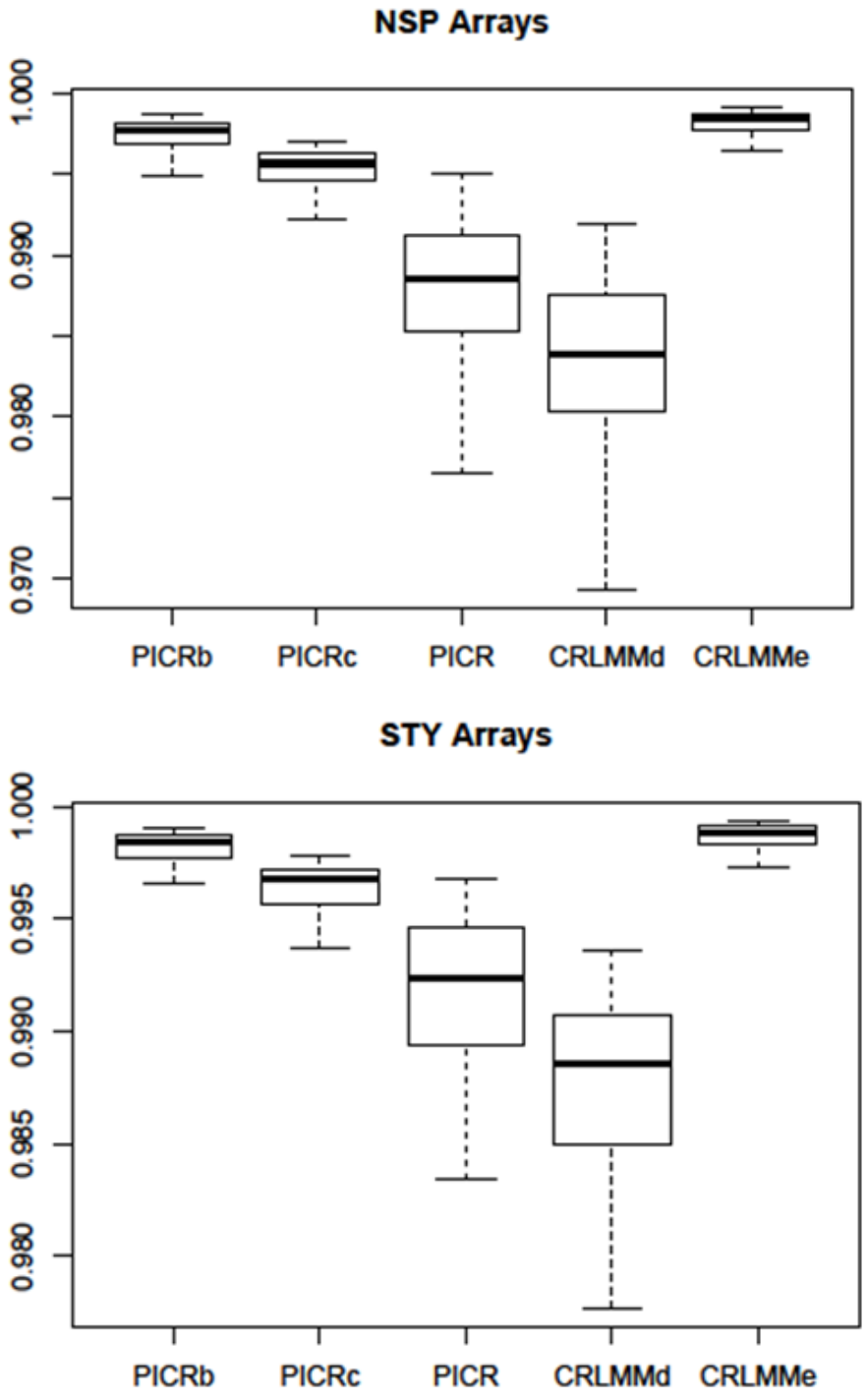
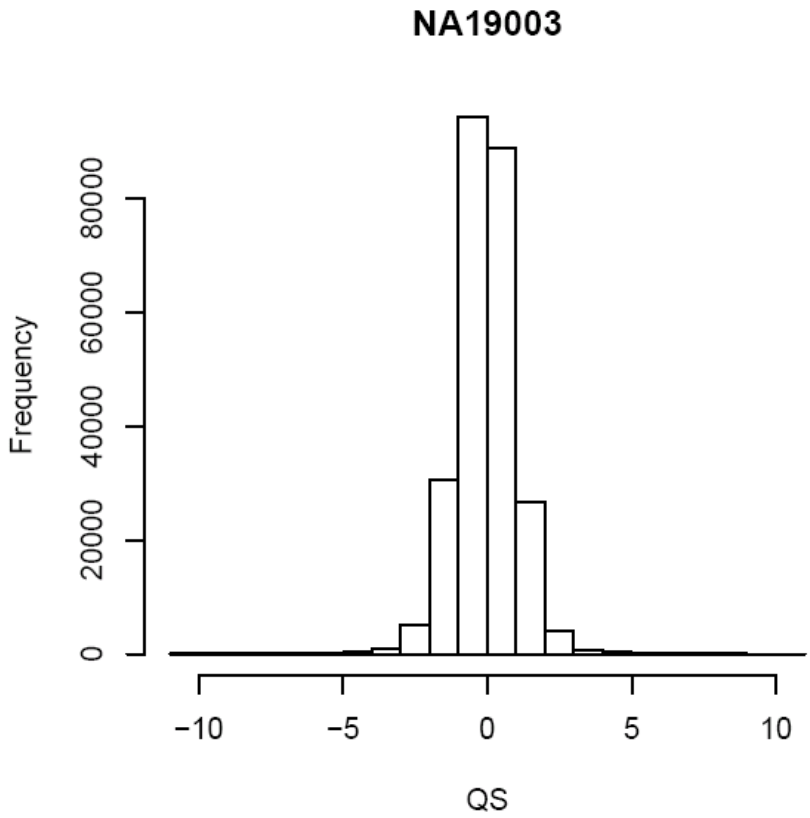


Figure 2.3: Boxplot of genotype-calling accuracy. PICRb: MA-SNP method with |QS|<3.56; PICRc: MA-SNP method with 100% call rate; PICR: PICR method; CRLMMd: CRLMM with 3 arrays as input per run; CRLMMe: CRLMM with 90 arrays as input.

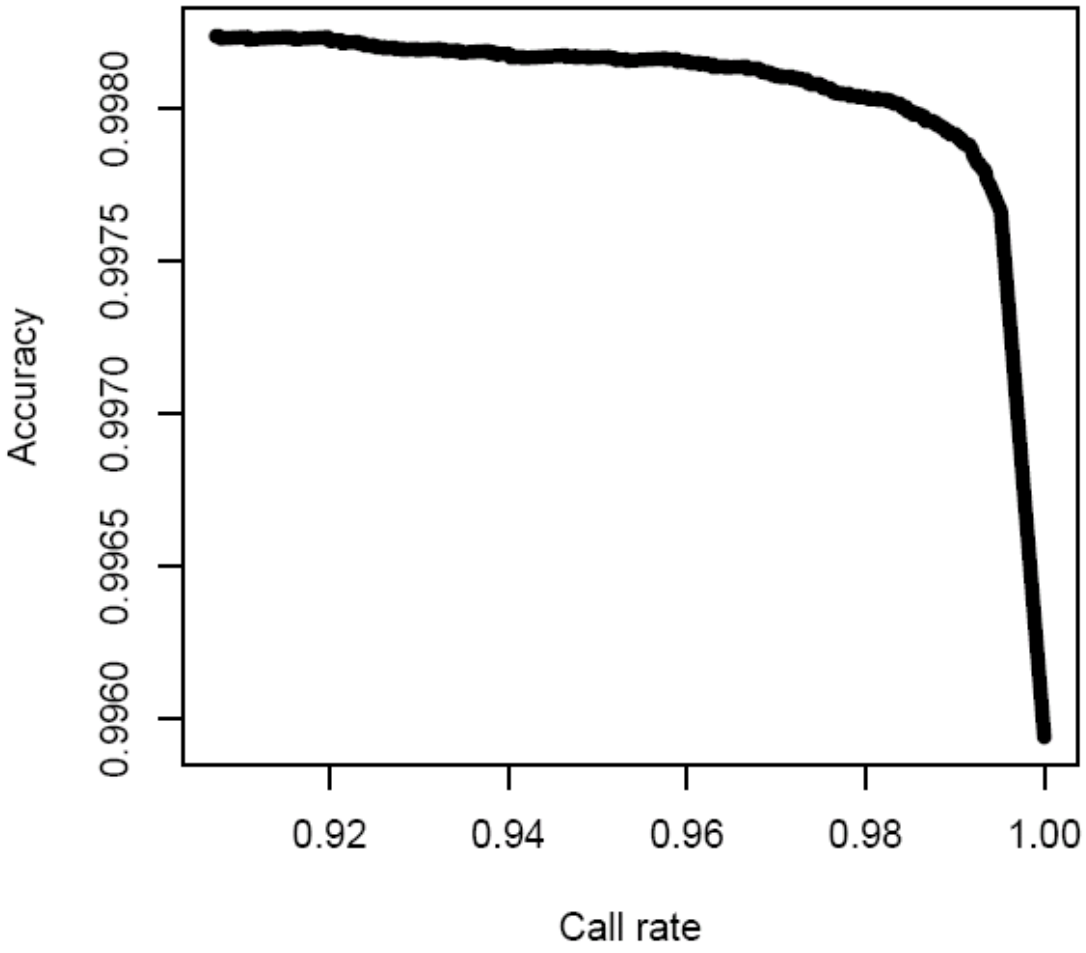
To assess the performance of the quality measure introduced in MA-SNP, we randomly select a testing array and plot the distribution of quality score, the accuracy against call rate, and the call rate against quality score, as shown in Figure 2.4. Table 2.2 displays summary statistics of the accuracy and call rate under different quality score threshold for the 90 testing arrays. We have observed that there is a tradeoff between call rate and accuracy. On one hand the more stringent quality score threshold, the higher the genotyping accuracy one can achieve. On the other hand, the more stringent quality score threshold, the lower the call rate one can achieve. Based on this testing study, we found the suggested quality control criterion |QS| > 3.56 achieves relatively high accuracy and low no-call rate.



a: Distribution of quality score.

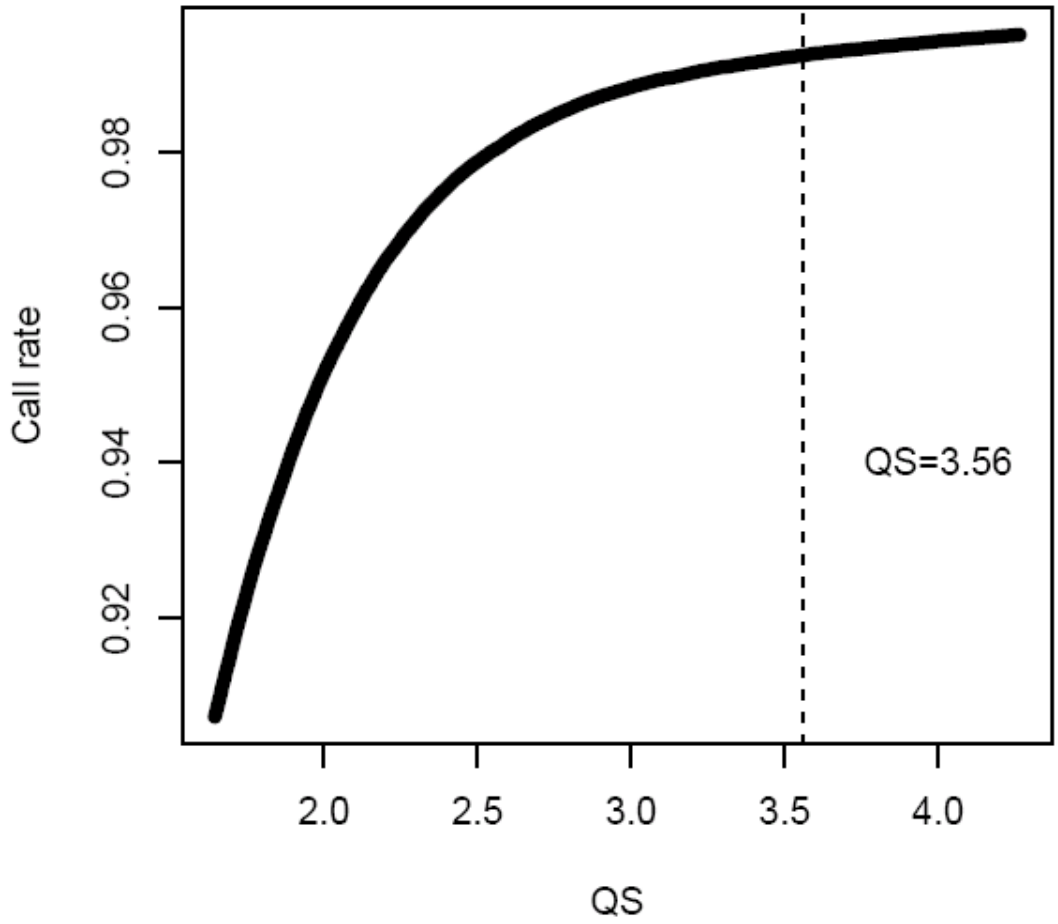
Figure 2.4 Plots of the distribution of quality score, the accuracy against call rate and the call rate against quality score for a randomly selected array

Figure 2.4 *cont'd*



b: The accuracy against call rate.

Figure 2.4 *cont'd*



c: The call rate against quality score.

Table 2.2 Quality score threshold criteria and genotype calling accuracy of the HapMap samples

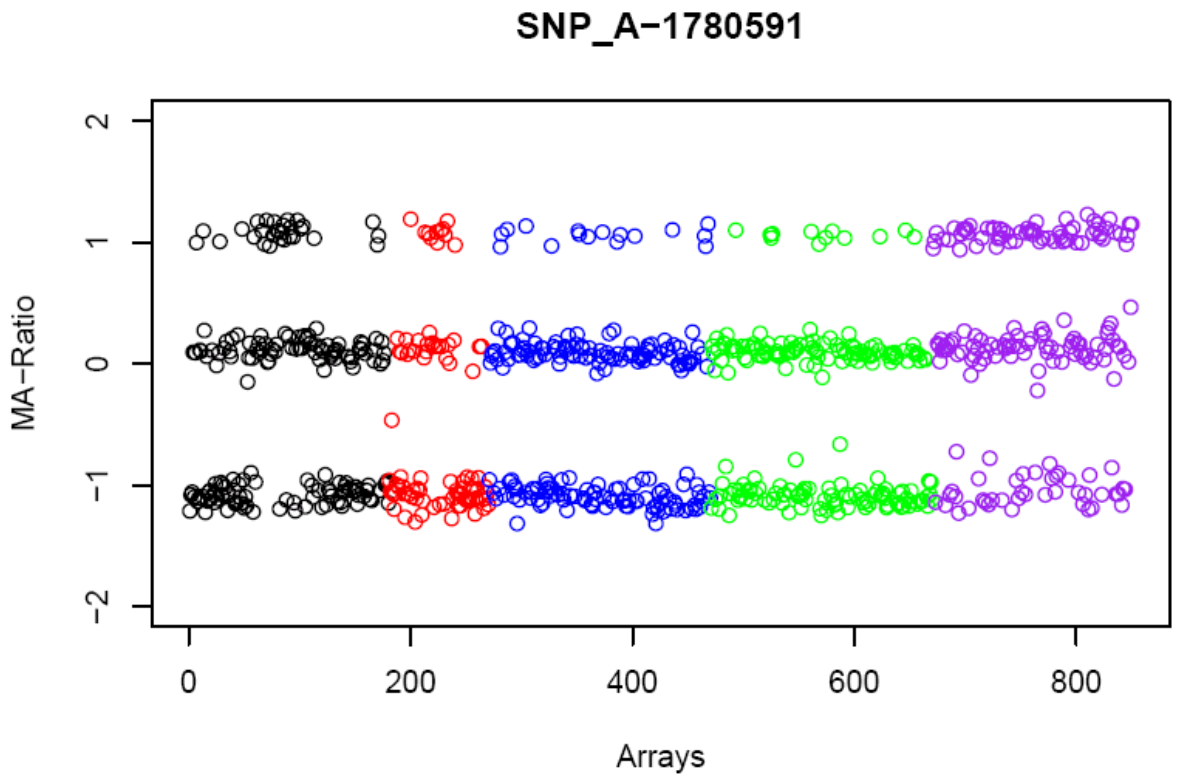
Samples	|QS| Threshold	Mean Call Rate (SD[a])	Mean Accuracy (SD[a])	5th	50th	95th
90 Nsp arrays	1.64	90.28% (0.0406)	0.9975 (0.0045)	0.9961	0.9984	0.9990
	2.33	96.17% (0.0241)	0.9970 (0.0055)	0.9953	0.9982	0.9989
	2.77	97.56% (0.0191)	0.9967 (0.0063)	0.9947	0.9980	0.9988
	3.30	98.38% (0.0159)	0.9963 (0.0070)	0.9942	0.9978	0.9987
	3.56	98.71% (0.0144)	0.9961 (0.0075)	0.9938	0.9977	0.9986
	5.26	99.39% (0.0103)	0.9952 (0.0093)	0.9926	0.9972	0.9983
	7.00	99.66% (0.0077)	0.9946 (0.0106)	0.9919	0.9968	0.9980
	14.00	99.88% (0.0043)	0.9934 (0.0118)	0.9908	0.9960	0.9971
90 Sty arrays	1.64	89.48% (0.0516)	0.9984 (0.0016)	0.9956	0.9989	0.9992
	2.33	95.82% (0.0294)	0.9981 (0.0020)	0.9944	0.9987	0.9991
	2.77	97.37% (0.0208)	0.9979 (0.0023)	0.9935	0.9986	0.9991
	3.30	98.31% (0.0142)	0.9976 (0.0026)	0.9926	0.9985	0.9991
	3.56	98.68% (0.0114)	0.9974 (0.0029)	0.9921	0.9984	0.9991
	5.26	99.46% (0.0051)	0.9968 (0.0037)	0.9890	0.9981	0.9987
	7.00	99.74% (0.0027)	0.9964 (0.0042)	0.9885	0.9978	0.9985
	14.00	99.92% (0.0010)	0.9956 (0.0045)	0.9873	0.9972	0.9979

a: Standard deviation

2.5.2 Robustness of the methods and batch effect

We apply MA-SNP genotype calling method to the samples in the CAD and NBS data in the WTCCC study and samples in the methylation profiling study. We notice that our clustering decision rules are robust for most of the SNPs across laboratories even without batch effect correction (Figure 2.5). This is not surprising because the new genotype calling method and quality scores are constructed based on allelic target concentrations, and the batch effect is mostly captured by the SNP-specific background. However, based on the density distribution of MA-ratios, we have observed a slight shift for each data set from the HapMap training data. The batch effect correction procedure is carried out independently for each of the CAD, NBS, and the Methylation Profiling Study data sets. With quality score threshold being 3.56, the call rate with / without batch effect correction and concordance rates are shown in Table 2.3. The concordance rates are above 99.5% for all the study data sets, which suggests that our method is robust across labs, and can be applied to the study with relatively small sample size. In addition, we notice that with batch effect correction the call rates increased by approximately 2.4% with an average of 99.2%. The distributions of quality score for 2 randomly selected arrays are shown in Figure 2.6. As expected the distribution of quality score without batch effect correction usually has a heavier tail and the call rate is lower than those with batch effect correction. This is because a small shift from the cluster center of the training data usually does not influence the genotype calling but affects the quality score. The distribution of quality score is no longer centered at zero if batch effect exists, which leads to a large quality score and a large portion of missing genotype call given a pre-specified quality threshold. With batch effect correction, the quality score is centered at zero and the confidence of genotype call increases, which leads to a higher call rate. For example, for SNP-A-1785441 shown in Figure 2.2, the call rate increases from 92.22% without

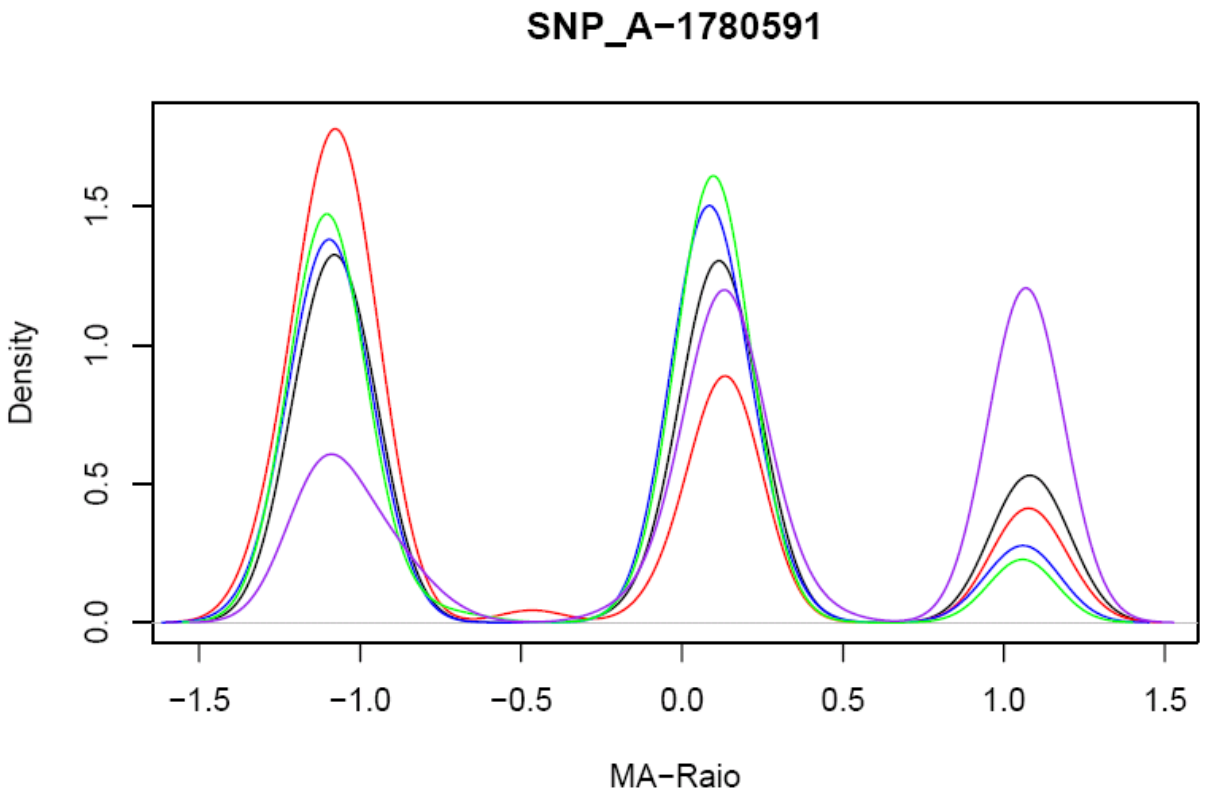
batch effect correction to 97.22% with batch effect correction, mainly because the BB type SNPs can be called with the updated cluster center.



a: Genotyping clusters.

Figure 2.5 Robustness of the MA-ratio clustering across samples in different studies with 2 randomly selected SNPs. Black: the HapMap training data; Red: the HapMap testing data; Blue: 200 randomly selected samples of CAD data; Green: 200 randomly selected samples of NBS data; Purple: methylation profiling data.

Figure 2.5 *cont'd*



b: Density of MA-ratios.

Table 2.3 Comparison between call rate with batch effect correction and without correction

Array Platform	Samples	Call rate[a] (SD[c])	Call rate[b] (SD[c])	Concordance rate(SD[c])
Affymetrix Mapping 250K Nsp Array	CAD[d]	98.1% (0.019)	99.3% (0.015)	99.7% (0.001)
	NBS[e]	98.3% (0.012)	99.4% (0.008)	99.8% (0.001)
	Cancer[f]	95.9% (0.032)	99.0% (0.009)	99.6% (0.002)
Affymetrix Mapping 250K Sty Array	CAD[d]	96.6% (0.029)	99.2% (0.016)	99.5% (0.003)
	NBS[e]	96.4% (0.023)	99.3% (0.006)	99.6% (0.003)
	Cancer[f]	95.4% (0.035)	98.9% (0.010)	99.7% (0.002)

a: Without batch effect correction, the percentage of SNPs with |QS|<3.56
b: With batch effect correction, the percentage of SNPs with |QS|<3.56
c: Standard deviation
d: Coronary artery disease from Wellcome Trust Case Control Study data set
e: UK Blood Service control from Wellcome Trust Case Control Study data set
f: From data set III

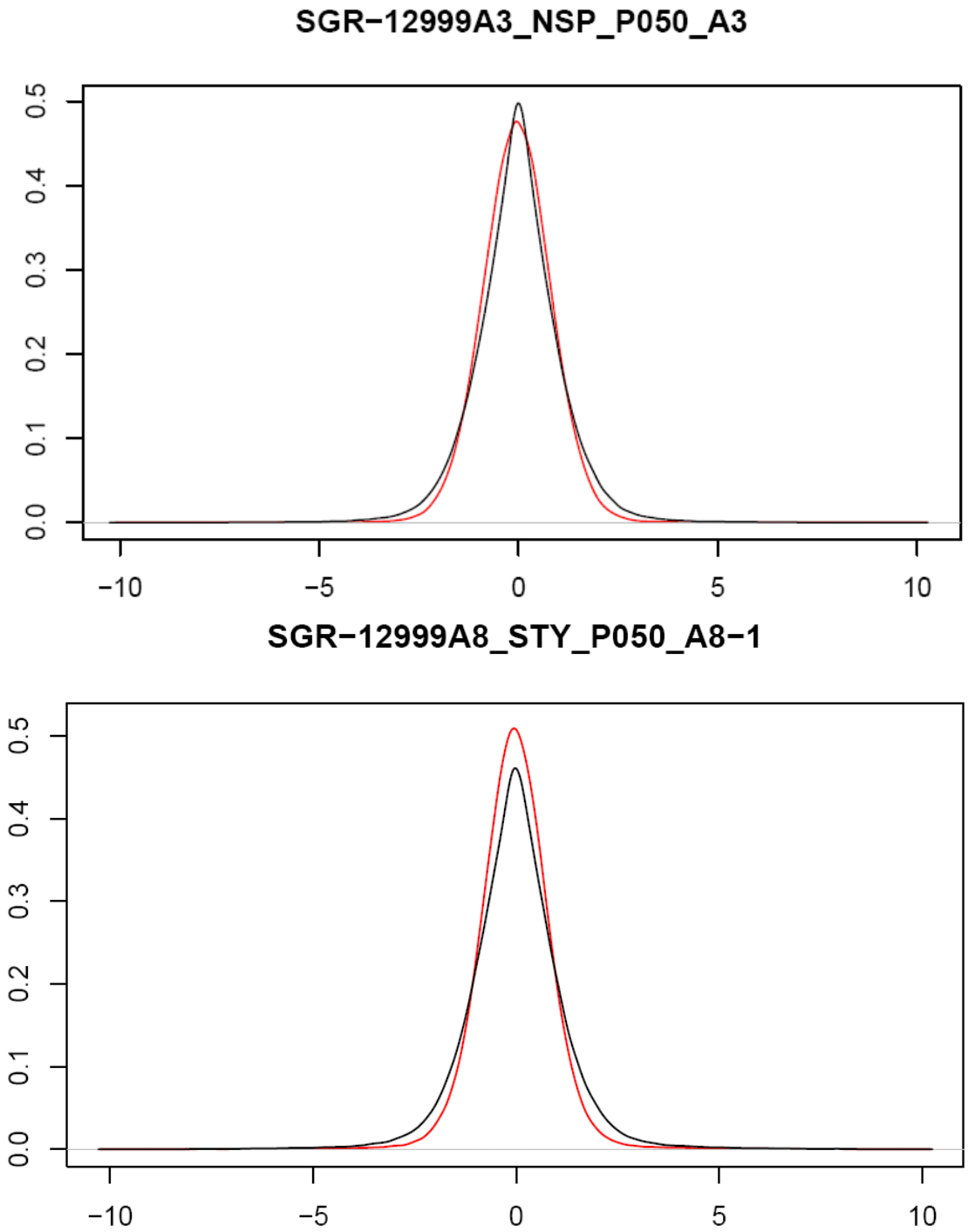


Figure 2.6 The distribution of quality score of 2 randomly selected samples with batch effect correction (red curve) and without batch effect correction (black curve).

To further demonstrate the MA-SNP genotyping method, we scanned the whole genome of chromosome 9 with the CAD and NBS data from the Wellcome Trust Case-Control Study. All

the SNPs that have been identified by a single locus association analysis with p-value<10^{-7} are located at 9p21.3. Figure 2.7 displays the genotyping clusters for a randomly selected two of these seven SNPs. All SNPs (rs1333042, rs1333048, rs1333049, rs2891168, rs4977574, and rs6475606) except for one rs9632884 have been reported with strong association by studies using various technologies [34-42], which indicates that our method does not generate many false positive findings.

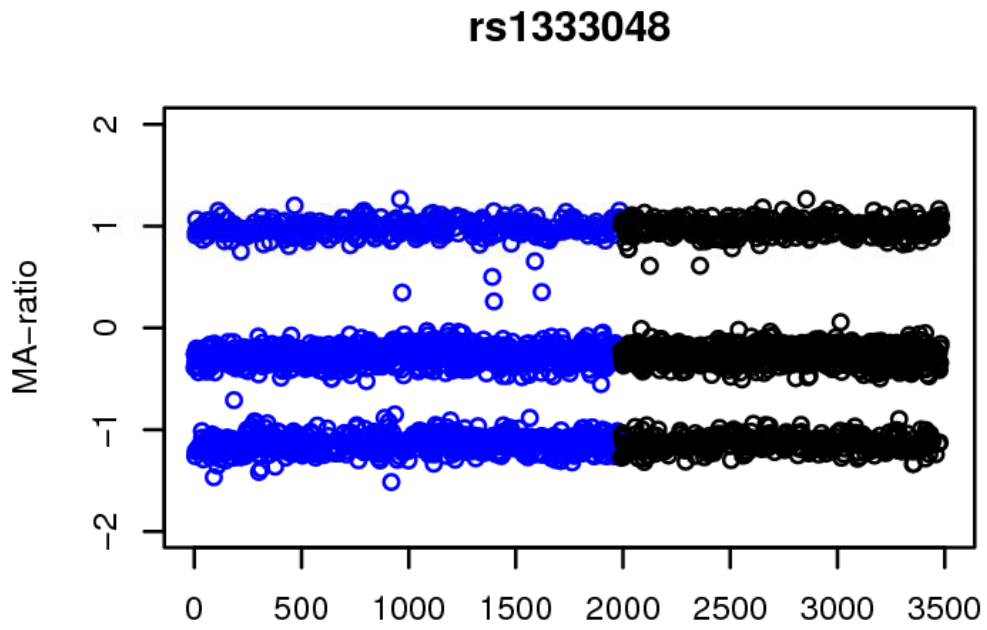
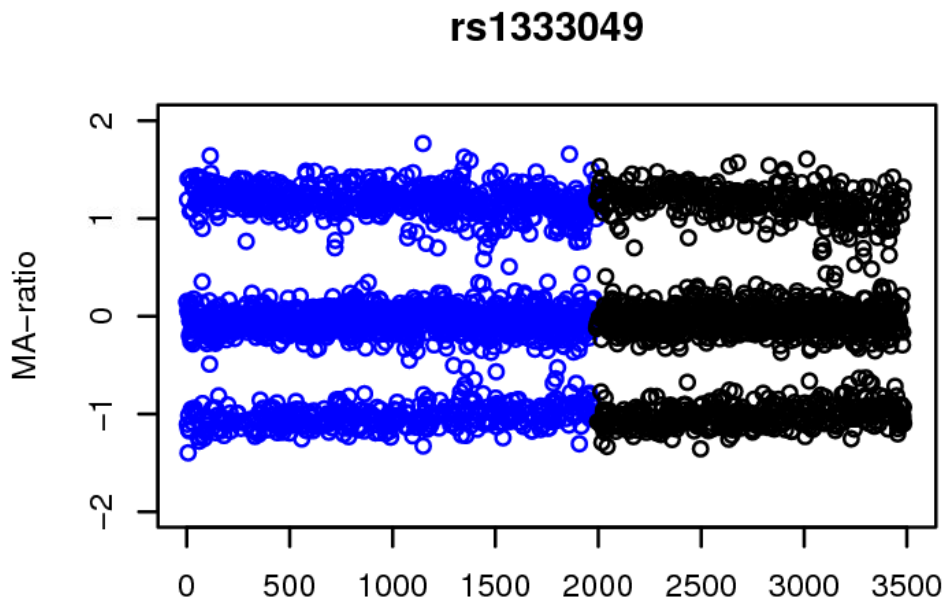


Figure 2.7 MA-ratio clustering of the 2 SNPs on chromosome 9 that showed strong association with the CAD at the significance level 10^{-7}. Blue: MA-ratios of CAD data; Black: MA-ratios of NBS data.

Figure 2.7 *cont'd*



2.6 Discussion

SNP genotype calling is susceptible to artifacts of the technology and experimental conditions that contribute to batch effects. The removal of batch effects usually requires additional computational effort. Our previous single array PICR method provides a universal criterion for genotyping SNPs. By modeling the array hybridization process, PICR removes simultaneously a number of artifacts, including the array unequal footing[18] and the genomic wave[29], and achieves consistent high accuracy across arrays and across array platforms. Although microarray hybridization can be largely modeled with the probe sequence structure, as evidenced by the PICR model[18], other factors may also affect the hybridization process, and may, if not properly adjusted for, lead to systematic bias of the estimated allelic copy number. Furthermore, the PICR neither provides a quality measure of the genotype calls, nor addresses the batch effect. In the current study, we adopted a normal mixture model to incorporate the SNP-specific

features, and developed a new SNP genotype calling method, the MA-SNP. The MA-SNP 1) estimates the allelic copy numbers of each SNP by the PICR model, 2) calculates the MA-ratio so that SNPs of different genotypes potentially cluster around different values of the MA-ratio, and 3) models the density of the MA-ratio using a normal mixture model allowing for SNP-specific features. It not only largely improves the genotype calling accuracy, but also provides quality measure for the genotyping and assessment of the batch effect through the mixture model.

Across-array normalization has been widely used for preprocessing the raw intensities to remove the artifacts that may confound the biological signals. However, such normalization procedure may introduce variability from unrelated individuals, which may potentially pose problems to downstream or subsequent analysis. Our proposed model depends on the allelic copy number estimated from PICR, a single array approach free of across-array normalization. Consequently, MA-SNP inherits the advantages of single array approach, and generates genotype calls fully determined by the individual's data, which makes it possible to conduct the genotype call for very small sample studies, even with one sample.

Although normal mixture model has been used previously in SNP genotype calling algorithm, such as in the CRLMM [9, 10], our MA-SNP method differs largely from the CRLMM. As discussed above, the CRLMM is a multiple array genotype calling method, while our MA-SNP is a single array method and thus has its unique features and properties in genotype calling. It provides quality measure for each SNP in a single sample, which could be used to set aside the problematic SNPs and improve the accuracy of downstream analysis. It also corrects for batch effect for multiple samples, which is highly recommended but difficult to implement with a general approach that fits all platforms. These features of MA-SNP make the approach applicable to both small sample studies and cross-laboratory large sample studies. The R-

package is available upon request. It can run on a laptop computer with 3GB memory and only takes approximately 50s to genotype each array.

Our model can be further improved in several ways. First, we have noticed that copy number variations may have an effect on genotyping accuracy for heterozygous SNP while homozygous SNPs are not affected. This is because the MA-ratio for homozygous SNP with copy number variation remains the same which indicates that the genotype calling procedure won't be affected. On the other hand, for heterozygous SNPs the MA-ratio increases with the ratio of allelic copy number if the ratio of allelic copy number is positive while MA-ratio decreases with the ratio of allelic copy number if the ratio of allelic copy number is negative. Second, the batch effect correction procedure can be extended to consider the possible effect of copy number variation. For example, rather than using Gaussian mixture model with 3 components we may employ a non-parametric method, such as kernel density estimation, to estimate the probability density function of the MA-ratio to account for new clusters for copy number variations at heterozygous SNPs. Third, our current model treats the batch effect as a fixed effect, and a random effect model that borrows strength across batchs might be useful especially for the SNPs with low minor allele frequencies and small batches/samples coming from different labs[10]. We will further explore methods in this direction.

In conclusion, we have presented a powerful tool to provide relatively accurate genotype call for a single array. It corrects for the batch effect and improves the call rate and genotyping accuracy. Our results provide strong evidence that modeling the underlying mechanism of array sequence hybridization can remove non-biological signals even without across-array normalization. The systematic non-biological signals can be modeled through the training data,

and it remains the same regardless of the sources of the samples, either from the same lab or from different labs.

REFERENCES

REFERENCES

1. Wellcome Trust Case Control Consortium, *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.

2. Barrett, J.C., et al., *Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.* Nature genetics, 2008. **40**(8): p. 955-62.

3. Adeyemo, A., et al., *A genome-wide association study of hypertension and blood pressure in African Americans.* PLoS genetics, 2009. **5**(7): p. e1000564.

4. Frayling, T.M., *Genome-wide association studies provide new insights into type 2 diabetes aetiology.* Nature reviews. Genetics, 2007. **8**(9): p. 657-62.

5. Dong, S., et al., *Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation.* Genome research, 2001. **11**(8): p. 1418-24.

6. Kennedy, G.C., et al., *Large-scale genotyping of complex DNA.* Nature biotechnology, 2003. **21**(10): p. 1233-7.

7. Rabbee, N. and T.P. Speed, *A genotype calling algorithm for affymetrix SNP arrays.* Bioinformatics, 2006. **22**(1): p. 7-12.

8. Affymetrix, *BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500K Array Set*, in *Technical Report, White Paper*. 2006, Affymetrix, Inc: Santa Clara, CA.

9. Carvalho, B., et al., *Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data.* Biostatistics, 2007. **8**(2): p. 485-99.

10. Carvalho, B.S., T.A. Louis, and R.A. Irizarry, *Quantifying uncertainty in genotype calls.* Bioinformatics, 2010. **26**(2): p. 242-9.

11. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nature Genetics, 2007. **39**(7): p. 906-913.

12. Affymetrix, *Birdseed Algorithm – Affymetrix Genotyping Console Software 2.0*. 2007, Affymetrix, Inc: Santa Clara, CA.

13. Korn, J.M., et al., *Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs.* Nature genetics, 2008. **40**(10): p. 1253-60.

14. Browning, B.L. and S.R. Browning, *A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.* American journal of human genetics, 2009. **84**(2): p. 210-23.

15. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(1): p. 31-6.

16. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.

17. Held, G.A., G. Grinstein, and Y. Tu, *Modeling of DNA microarray data by using physical properties of hybridization.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(13): p. 7575-80.

18. Wan, L., et al., *Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation.* Nucleic acids research, 2009. **37**(17): p. e117.

19. Zhang, L., M.F. Miles, and K.D. Aldape, *A model of molecular interactions on short oligonucleotide microarrays.* Nature biotechnology, 2003. **21**(7): p. 818-21.

20. Zhang, L., et al., *Free energy of DNA duplex formation on short oligonucleotide microarrays.* Nucleic acids research, 2007. **35**(3): p. e18.

21. Lin, S., et al., *Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays.* Genome biology, 2008. **9**(4): p. R63.

22. Miclaus, K., et al., *Variability in GWAS analysis: the impact of genotype calling algorithm inconsistencies.* The pharmacogenomics journal, 2010. **10**(4): p. 324-35.

23. Hong, H., et al., *Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples.* BMC bioinformatics, 2008. **9 Suppl 9**: p. S17.

24. Miclaus, K., et al., *Batch effects in the BRLMM genotype calling algorithm influence GWAS results for the Affymetrix 500K array.* The pharmacogenomics journal, 2010. **10**(4): p. 336-46.

25. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.

26. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nature reviews. Genetics, 2010. **11**(10): p. 733-9.

27. The International HapMap Consortium, *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

28. Yang, H.H., et al., *Influence of genetic background and tissue types on global DNA methylation patterns.* PloS one, 2010. **5**(2): p. e9355.

29. Wen, Y., M. Li, and W.J. Fu, *Catching the genomic wave in oligonucleotide SNP arrays by modeling sequence binding.* Bioinformatics (submitted), 2011.

30. Diskin, S.J., et al., *Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms.* Nucleic acids research, 2008. **36**(19): p. e126.

31. Marioni, J.C., et al., *Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization.* Genome biology, 2007. **8**(10): p. R228.

32. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Statistical applications in genetics and molecular biology, 2004. **3**: p. Article3.

33. Fraley, C. and A.E. Raftery, *Bayesian regularization for normal mixture estimation and model-based clustering*, in *Technical Report*. 2005, Department of Statistics, University of Washington.

34. Cluett, C., et al., *The 9p21 myocardial infarction risk allele increases risk of peripheral artery disease in older people.* Circulation. Cardiovascular genetics, 2009. **2**(4): p. 347-53.

35. Ellis, K.L., et al., *A common variant at chromosome 9P21.3 is associated with age of onset of coronary disease but not subsequent mortality.* Circulation. Cardiovascular genetics, 2010. **3**(3): p. 286-93.

36. Lanktree, M., J. Oh, and R.A. Hegele, *Genetic testing for atherosclerosis risk: inevitability or pipe dream?* The Canadian journal of cardiology, 2008. **24**(11): p. 851-4.

37. Preuss, M., et al., *Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls.* Circulation. Cardiovascular genetics, 2010. **3**(5): p. 475-83.

38. Qi, L., et al., *Genetic risk score and risk of myocardial infarction in Hispanics.* Circulation, 2011. **123**(4): p. 374-80.

39. Saleheen, D., et al., *Association of the 9p21.3 locus with risk of first-ever myocardial infarction in Pakistanis: case-control study in South Asia and updated meta-analysis of Europeans.* Arteriosclerosis, thrombosis, and vascular biology, 2010. **30**(7): p. 1467-73.

40. Schaefer, A.S., et al., *Identification of a shared genetic susceptibility locus for coronary heart disease and periodontitis.* PLoS genetics, 2009. **5**(2): p. e1000378.

41. Silander, K., et al., *Worldwide patterns of haplotype diversity at 9p21.3, a locus associated with type 2 diabetes and coronary heart disease.* Genome medicine, 2009. **1**(5): p. 51.

42. Yamada, Y., S. Ichihara, and T. Nishida, *Molecular genetics of myocardial infarction.* Genomic medicine, 2008. **2**(1-2): p. 7-22.

CHAPTER 3

A NEW MULTIVARIATE MANN-WHITNEY APPROACH TO STUDY THE COMORBIDITY BETWEEN CORONARY HEART DISEASE AND TYPE II DIABETES

3.1 Abstract

Co-morbidity among complex human diseases is well documented and converging evidences suggest that the interplay among multiple genetic variants contributes to disease co-morbidity. The discovery of common genetic variants and their interactions likely shed light on etiology, as well as promote effective prevention and treatment for co-morbidity conditions. Despite its potential importance, co-morbidity of complex diseases has been under-studied and its associated analytic tools have been much less developed. A common practice to investigate co-morbidity is through a composite phenotype method (COM). However, the method does not take into account individuals with only one of the co-morbid conditions and thus could be subject to decreased power, especially when co-morbidity rate is low. Moreover, COM only identifies common genetic variants predisposing to co-morbidity, but not those unique to each disease outcome. To address these issues, we propose a multivariate Mann-Whitney (MMW) approach to unravel common genetic variants and interactions contributing to disease co-morbidity, as well as those unique to each co-morbid condition. Through simulations, we find MMW outperforms COM in a variety of underlying disease and correlation models between two co-morbid conditions. Finally, we apply our method to datasets from the *Wellcome Trust Case Control Consortium* to investigate the co-morbidity between coronary artery disease (CAD) and type II diabetes (T2D). The co-morbidity analysis using MMW identified 3-locus and 5-locus models for CAD and T2D respectively, but no loci contributing to both diseases have been found from the datasets.

Key Words: Co-morbidity, Forward selection, High-order interaction, CAD, T2D

3.2 Introduction

The concept of "co-morbidity" was first introduced in 1970s by Feinstein. It stands for the scenario where "a distinct clinical entity" occurred together with a specific disease under study [1-3]. Recently, multi-comorbidity has been introduced, referring to the scenario where multiple medical conditions occur in one person without an emphasis on the presence of a specific disease [4, 5]. Both co-morbidity and multi-comorbidity are used in the domains of clinical care, epidemiological studies, and health service policies [1-7]. In the rest of my dissertation, we use comorbidity to refer to both co-morbidity and multi-comorbidity.

The relation between co-morbidity conditions is complex and presents in various forms. To describe the underlying mechanisms leading to disease co-morbidity, Neale and Kendler have proposed thirteen theoretical co-morbidity models [8, 9]. The simplest scenario is that the co-morbidity conditions are independent of each other and occur together simply by chance or due to a third distinct disease[8]. Co-morbidity can also be the cause or consequence of one of the co-morbid conditions with possible reciprocal causality [3, 8, 10]. Another common scenario is that co-morbidity conditions share the same or correlated risk factors, which makes the co-morbid conditions more likely occur together [8, 11, 12]. In certain circumstance, co-morbidity may also reflect the fact that the co-morbid conditions are alternative manifestations of a single liability[8].

The radical breakthrough in biotechnologies has made it possible to rapidly and accurately genotype millions of single nucleotide polymorphism (SNP) with affordable cost. Benefiting from these high throughout technologies and the HapMap project[13], significant progress has been made in genome-wide association studies to discover novel genetic variants that contribute to complex human diseases[14-25]. With the increase of genetic findings,

cumulative evidence revealed that the same genetic variants could be associated with multiple related disease outcomes. For example, recent studies have provided evidence that neuronal nicotinic acetylcholine receptors (*nAChRs*) subunit genes may play an important role in the common pathophysiological pathway of nicotine dependence (ND) and alcohol dependence (AD) [26-28]. Similarly, clinical and epidemiological studies suggested a high-degree of co-morbidity between bipolar disorder and migraine, which could be partially explained by a shared genetic component [29-33]. In addition, co-morbidity between coronary heart disease and type II diabetes is well documented in literature[34-36], and it may reflect the fact that one of the co-morbid conditions is the cause or consequence of the other [3, 8, 10]. It is also possible that the two diseases share the same or correlated risk factors, such as obesity, physical inactivity, and insulin resistance, making the co-morbid conditions more likely to occur simultaneously [8, 11, 12, 37-44]. Despite these findings, the pathophysiology and etiology of disease co-morbidity remain largely unknown. [32]. Identification of genetic variants and environmental determinants common to disease co-morbidity, as well as unique to each condition, is of great importance, as it helps elucidate the causes of co-morbidity and promote new diagnostic and therapeutic strategies for both diseases.

A common practice to study co-morbidity is through a composite phenotype (COM) method, in which the "cases" are defined as individuals with all co-morbid conditions while the "controls" are defined as individuals with none of the co-morbid conditions [12, 19, 32]. Though easy to implement, such method does not take into account individuals with only one of the co-morbid conditions. As a consequence, it may lack the power to catch pathophysiological pathway underlying the disease due to reduction in sample size. In addition, COM is designed to identify common genetic variants leading to co-morbidity, but not unique genetic variants for

each disease outcome. To address these limitations, we propose a multivariate Mann-Whitney (MMW) approach for co-morbidity analysis. The proposed method utilizes the entire sample, and is capable of capturing shared genetic variants and their possible interactions, contributing to disease co-morbidity, as well as unique genetic variants for each disease outcome. In the following sections, we first lay out the details of the MMW approach, and then evaluate the performance of the proposed method with simulations. We further apply the new method to datasets obtained from Wellcome Trust Case Control Consortium (WTCCC), where we focus on the co-occurrence of coronary artery disease (CAD) and type II diabetes (T2D). In the last section, we summarize and discuss our findings.

3.3 The Model

Consider a co-morbidity study of *N* unrelated individuals and *G* genetic markers, where we are interested in identifying shared and unique genetic susceptibility markers contributing to co-morbid conditions. Without loss of generality, we illustrate the MMW method using two co-morbid conditions. Let Y be the response measurement and $Z = (Z_1, Z_2, \ldots, Z_G)$ be the measurement of *G* makers, where Y =0 for controls and Y =*k* (*k*=1,2) for cases with *k* co-morbid condition. The MMW method first applies a Mann-Whitney-based forward selection algorithm [45] to search for genetic variants predisposing to each of the two conditions. The algorithm starts with a null model without any genetic markers, and then gradually selects disease-susceptibility markers into the model, considering its interaction with other selected loci. In step one, it searches all *G* genetic markers for a marker most strongly associated with the given co-morbid condition. In step two, it searches for the second marker that is most related to the condition, considering its possible interaction with the selected marker. The whole process

continues until it reaches a full model, and *K*-fold cross-validation is then used to choose the most parsimonious model.

By applying Mann-Whitney-based forward selection to each of two co-morbid conditions, we identify two sets of disease-susceptibility markers, $X_1 = (Z_{p_1}, Z_{p_2}, \ldots, Z_{p_M})$, $M \le G$ and $X_2 = (Z_{q_1}, Z_{q_2}, \ldots\ldots, Z_{q_S})$, $S \le G$, for co-morbid conditions 1 and 2, respectively. Let $X_C = X_1 \bigcap X_2$ be the common set of markers shared by both co-morbid diseases, $X_{U_1} = X_1 \bigcap \overline{X_C}$ be the subset of markers unique for disease 1, and $X_{U_2} = X_2 \bigcap \overline{X_C}$ be the subset of makers unique for disease 2. A multivariate likelihood ratio can be obtained based on individuals' genotypes to measure their risks of the co-morbid conditions,

$$LR^M = LR(X_C) \prod_{k=1}^{2} LR(X_{U_k}), \qquad (1)$$

where $LR(X_C) = \frac{P(X_C \mid Y \neq 0)}{P(X_C \mid Y = 0)}$ and $LR(X_{U_k}) = \frac{P(X_{U_k} \mid X_C, Y = k)}{P(X_{U_k} \mid X_C, Y = 0)}$, *k*=1, 2. In the cases when a null set occurs (i.e., no marker has been selected), we define

$$\begin{cases} LR(X_{U_k}) = \dfrac{P(X_{U_k} \mid Y = k)}{P(X_{U_k} \mid Y = 0)}; \ LR(X_C) = 1 & X_C = \phi \\ LR(X_{U_k}) = 1 & X_{U_k} = \phi \end{cases}.$$

Given the multivariate likelihood ratio values, we can form a MMW statistic to assess the joint association of disease-susceptibility markers with the co-morbid conditions, allowing for interaction,

$$U_{MMW} = \sum_{i=1}^{N_{Y\neq 0}} \sum_{j=1}^{N_{Y=0}} \psi(LR_i^M, LR_j^M), \quad (2)$$

where $N_{Y\neq 0}$ and $N_{Y=0}$ are the number of individuals with at least one of two co-morbid condition and the number of non-disease individuals, respectively. The kernel function ψ equals 1 if LR_i^M is greater than LR_j^M, 0.5 if equal, and 0 if less. Asymptotically,

$U_{MMW} \sim N(\frac{N_{Y\neq 0}N_{Y=0}}{2}, (S_D + S_{\bar{D}})^2)$. The hypothesis testing can then be conducted to assess the significance of the joint association,

$$\frac{U_{MMW} - \frac{N_{Y\neq 0}N_{Y=0}}{2}}{S_D + S_{\bar{D}}}, \quad (3)$$

where $S_D = \sum_{i=1}^{N_{Y\neq 0}} (\sum_{j=1}^{N_{Y=0}} \psi(LR_i^M, LR_j^M) - \frac{U_{MMW}}{N_{Y\neq 0}})^2$ and

$S_{\bar{D}} = \sum_{j=1}^{N_{Y=0}} (\sum_{i=1}^{N_{Y\neq 0}} \psi(LR_i^M, LR_j^M) - \frac{U_{MMW}}{N_{Y=0}})^2$.

3.4 Results

3.4.1 Simulation I

In the first set of simulations, we compare the performance of MMW and COM under a variety of co-morbidity correlation models. We simulate two co-morbid diseases and consider 1) a model where two diseases are unrelated, 2) a model where two diseases share one single nucleotide polymorphism (SNP), 3) a model where two diseases share a two-locus interaction, and 4) a model where two diseases are associated with exact the same disease susceptibility loci (Table 3.1). Each co-morbid disease is associated with two SNPS with interaction effect and an independent SNP, where we assume the two-locus interaction follow a multiplicative-interaction model or a threshold-interaction model [46], and the independent SNP is additive. All genetic variants are simulated under the Hardy-Weinberg Equilibrium (HWE) assumption with minor allele frequencies ranged from 0.3 to 0.4. In addition to disease-susceptibility loci, we also introduce 5 non-disease associated SNPs for each disease, and randomly assign their minor allele frequencies from a uniform distribution ranged from 0.1 to 0.5. For each underlying correlation model, 1000 replicates are simulated, and each comprised of 1000 control individuals and 1000 affected individuals with at least one of the co-morbid conditions. We analyze each replicate by using the proposed MMW method, as well as the COM method. With COM method, Mann-Whitney-based forward selection algorithm [45] is applied to search for genetic variants predisposing to both diseases among controls and individuals with both co-morbid conditions. Permutation test is then used to assess significance level of MMW and COM, adjusting for the inflated Type I error due to the use of model selection. For this purpose, the empirical null distribution is formed based on 1000 permutations, and the empirical p-value is then obtained by

comparing the observed statistic to the empirical null distribution. Type I error and power for each co-morbidity model are summarized in Table 3.2.

Table 3.1. Summary of the Simulation I settings

	Disease models		Gene-Gene Interaction			Single Locus			
	Disease1	Disease2	OR [b]	OR [c]	MAF [d]	OR [b]	OR [c]	MAF [f]	MAF [g]
Simulation I[a]	A+B*C	D+E*F	1.45	1.45	[0.3,0.4] [e]	1.45	1.45	0.40	0.40
	A[i]+B*C	***A***+D*E	1.45	1.45	[0.3,0.4]	1.45	1.45	0.40	0.40
	A+***B*C***	D+***B*C***	1.45	1.45	[0.3,0.4]	1.45	1.45	0.40	0.40
	A+B*C	***A+B*C***	1.45	1.45	[0.3,0.4]	1.45	1.45	0.40	0.40
Simulation I[h]	A+B*C	D+E*F	1.70	1.70	[0.3,0.4]	1.70	1.65	0.40	0.35
	A+B*C	***A***+D*E	1.70	1.70	[0.3,0.4]	1.70	1.65	0.40	0.35
	A+***B*C***	D+***B*C***	1.70	1.70	[0.3,0.4]	1.70	1.65	0.40	0.35
	A+B*C	***A+B*C***	1.70	1.70	[0.3,0.4]	1.70	1.65	0.40	0.35

a: A multiplicative-interaction model;
b: Odds Ratio for co-morbid condition 1;
c: Odds Ratio for co-morbid condition 2;
d: Minor Allele Frequency
e: Minor Allele Frequency simulated in the model ranged from 0.3 to 0.4.
f: Minor Allele Frequency for co-morbid condition 1.
g: Minor Allele Frequency for co-morbid condition 2.
h: A threshold-interaction model;
i: Bold and Italic letter represents the shared locus.

The results show that the type I errors from both approaches are well controlled at the level of 0.05. We also have observed that, the power of COM increases with the increase of shared genetic components. In the extreme case, when the two co-morbid conditions share the same genetic loci, the power of COM attains its highest value, which can be largely explained by the increasing number of individuals with both co-morbid conditions. Nevertheless, when two co-morbid conditions are independent and the simultaneous manifestation of both diseases occur only by chance, the power of COM is significantly reduced. Compared with COM, MMW attains higher or at least equivalent power under all models. The performance of MMW is also less affected by the relationship between co-morbid conditions, remaining almost the same across all models. While we expect that COM has no power under the model where two diseases are independent with no shared loci, the result shows that COM obtains power of 0.530 and 0.561 under the multiplicative-interaction model and the threshold-interaction model, respectively. As we demonstrate in the later simulation (Simulation III), the power of COM can also be partially explained by loci unique to each condition (i.e., if a locus is strongly associated with one of the co-morbid conditions, it could have an effect on subset of individuals with two conditions as well). However, the drawback of COM is that it cannot distinguish the shared and unique disease-susceptibility loci, while MMW has the capacity to correctly infer shared or unique loci.

Table 3.2 Type I error and power comparison of MMW and COM under different correlation models

Models		Multiplicative-interaction		Threshold-interaction	
		MMW	COM	MMW	COM
Without shared loci					
Disease1:A+B*C	Power	0.970	0.530	0.880	0.561
Disease2:D+E*F	Type I error	0.040	0.050	0.041	0.050
Shared one locus					
Disease1:***A***[a]+B*C	Power	0.930	0.650	0.931	0.570
Disease2:***A***+D*E	Type I error	0.055	0.010	0.049	0.041
Shared two loci					
Disease1:A+***B*C***	Power	0.969	0.826	0.950	0.680
Disease2:D+***B*C***	Type I error	0.049	0.050	0.058	0.060
Shared all loci					
Disease1:***A***+***B*C***	Power	0.989	0.929	0.989	0.809
Disease2:***A***+***B*C***	Type I error	0.055	0.041	0.054	0.031

a: Bold and Italic letters represent the shared loci between two diseases

3.4.2 Simulation II

In this set of simulations, we vary both underlying disease models and relations between two co-morbid conditions, and evaluate their impact on two approaches. We start with a simple model with a two loci of interaction effect and two independent loci, and then consider a more complex model involving a high-order interaction (i.e., a three-locus interaction) and a model involving more than one interaction (i.e., two two-locus interactions). The common disease-susceptibility loci contributed to both diseases are assumed to be 1) the interacting loci, 2) the interacting loci and one independent locus, and 3) two independent loci. Two types of interaction models, a multiplicative-effect interaction model and a threshold-effect model [46], are considered in the simulation. The details of simulation settings are summarized in Table 3.3.

The type I errors are well controlled at the level of 0.05 for both approaches (Table 3.4). Similar as observed in Simulation I, the performance of COM highly depends on the number of shared loci (i.e. as the number of shared loci increases, the power of COM approach increases a

lot). Compared with COM, MMW is robust to a variety of relations between two co-morbid conditions, and has higher power under all kinds of underlying disease models, regardless of the complexity of disease models and different types of interaction models.

Table 3.3 Summary of the Simulation II settings

Models		OR[c]		MAF[d]	
Disease1	Disease2	Disease1	Disease	Disease1	Disease2
Two-locus interaction models					
AB*+***C***+D[a]	***A***B*+***C***+E[a]	1.4;1.4;1.35	1.4;1.4;1.35	0.4;0.35;0.4;0.3	0.4;0.35;0.4;0.3
AB*+C+D[a]	***A***B*+E+F[a]	1.4;1.4;1.35	1.4;1.4;1.35	0.4;0.35;0.4;0.3	0.4;0.35;0.4;0.3
A*B+***C***+***D***[a]	E*F+***C***+***D***[a]	1.4;1.4;1.35	1.4;1.4;1.35	0.4;0.35;0.4;0.3	0.4;0.35;0.4;0.3
AB*+***C***+D[b]	***A***B*+***C***+E[b]	1.5;1.4;1.4	1.5;1.4;1.4	0.4;0.35;0.4;0.3	0.4;0.35;0.4;0.3
AB*+C+D[b]	***A***B*+E+F[b]	1.5;1.4;1.4	1.5;1.4;1.4	0.4;0.35;0.4;0.3	0.4;0.35;0.4;0.3
A*B+***C***+***D***[b]	E*F+***C***+***D***[b]	1.5;1.4;1.4	1.5;1.4;1.4	0.4;0.35;0.4;0.3	0.4;0.35;0.4;0.3
Three-locus interaction models					
AB*C*+***D***+E[a]	***A***B*C*+***D***+F[a]	1.3;1.3;1.25	1.29;1.28;1.23	0.3;0.35;0.4;0.35;0.35	0.3;0.35;0.4;0.35;0.35
AB*C*+D+E[a]	***A***B*C*+F+G[a]	1.3;1.3;1.25	1.29;1.28;1.23	0.3;0.35;0.4;0.35;0.35	0.3;0.35;0.4;0.35;0.35
A*B*C+***D***+***E***[a]	F*G*H+***D***+***E***[a]	1.3;1.3;1.25	1.29;1.28;1.23	0.3;0.35;0.4;0.35;0.35	0.3;0.35;0.4;0.35;0.35
AB*C*+***D***+E[b]	***A***B*C*+***D***+F[b]	1.38;1.3;1.3	1.4;1.3;1.3	0.3;0.35;0.4;0.35;0.35	0.3;0.35;0.4;0.35;0.35
AB*C*+D+E[b]	***A***B*C*+F+G[b]	1.38;1.3;1.3	1.4;1.3;1.3	0.3;0.35;0.4;0.35;0.35	0.3;0.35;0.4;0.35;0.35
A*B*C+***D***+***E***[b]	F*G*H+***D***+***E***[b]	1.38;1.3;1.3	1.4;1.3;1.3	0.3;0.35;0.4;0.35;0.35	0.3;0.35;0.4;0.35;0.35
Two two-locus interaction models					
AB*+***C***D*+E[a]	***A***B*+***C***D*+F[a]	1.35;1.35;1.3	1.35;1.35;1.3	0.3,0.35,0.4,0.3,0.35	0.3,0.35,0.4,0.3,0.35
AB*+C*D+***E***[a]	***A***B*+F*G +***E***[a]	1.35;1.35;1.3	1.35;1.35;1.3	0.3,0.35,0.4,0.3,0.35	0.3,0.35,0.4,0.3,0.35
A*B+C*D+***E***[a]	F*G+H*I+***E***[a]	1.35;1.35;1.3	1.35;1.35;1.3	0.3,0.35,0.4,0.3,0.35	0.3,0.35,0.4,0.3,0.35
AB*+***C***D*+E[b]	***A***B*+***C***D*+F[b]	1.5;1.5;1.6	1.5;1.5;1.6	0.3,0.35,0.4,0.3,0.35	0.3,0.35,0.4,0.3,0.35
AB*+C*D+***E***[b]	***A***B*+F*G +***E***[b]	1.5;1.5;1.6	1.5;1.5;1.6	0.3,0.35,0.4,0.3,0.35	0.3,0.35,0.4,0.3,0.35
A*B+C*D+***E***[b]	F*G+H*I+***E***[b]	1.5;1.5;1.6	1.5;1.5;1.6	0.3,0.35,0.4,0.3,0.35	0.3,0.35,0.4,0.3,0.35

a: The interaction effect follows multiplicative-interaction model
b: The interaction effect follows threshold-interaction model

c: Odds ratio
d: Minor allele frequency

Table 3.4 Type I error and power comparison of MMW and COM under different interaction models

Models		Multiplicative-interaction		Threshold-interaction	
		MMW	COM	MMW	COM
Two-locus interaction models					
Disease1:***A*******B***+***C***+D	Power	0.991	0.840	0.887	0.828
Disease2:***A*******B***+***C***+E	Type I error	0.050	0.051	0.047	0.066
Disease1:***A*******B***+C+D	Power	0.977	0.649	0.887	0.588
Disease2:***A*******B***+E+F	Type I error	0.043	0.049	0.065	0.048
Disease1:A*B+***C***+***D***	Power	0.932	0.407	0.844	0.573
Disease2:E*F+***C***+***D***	Type I error	0.052	0.053	0.040	0.054
Three-locus interaction models					
Disease1:***A*******B*******C***+***D***+E	Power	0.931	0.793	0.928	0.847
Disease2:***A*******B*******C***+***D***+F	Type I error	0.053	0.045	0.050	0.066
Disease1:***A*******B*******C***+D+E	Power	0.900	0.729	0.901	0.820
Disease2:***A*******B*******C***+F+G	Type I error	0.045	0.045	0.039	0.053
Disease1:A*B*C+***D***+***E***	Power	0.822	0.359	0.828	0.486
Disease2:F*G*H+***D***+***E***	Type I error	0.053	0.040	0.041	0.069
Two two-locus interaction models					
Disease1:***A*******B***+***C*******D***+E	Power	0.998	0.876	0.975	0.964
Disease2:***A*******B***+***C*******D***+F	Type I error	0.044	0.067	0.048	0.059
Disease1:***A*******B***+C*D+***E***	Power	0.998	0.874	0.946	0.885
Disease2:***A*******B***+F*G +***E***	Type I error	0.054	0.048	0.050	0.060
Disease1:A*B+C*D+***E***	Power	0.908	0.427	0.898	0.713
Disease2: F*G+H*I+***E***	Type I error	0.039	0.064	0.052	0.043

3.4.3 Simulation III

One of the unique features of MMW is that it can distinguish unique loci predisposing each co-morbid condition from common loci contributing to co-morbidity. To demonstrate this feature, a simple disease model is simulated where each of the two co-morbid diseases is associated with a common two-locus interaction and a unique locus. We vary the ratio of the effect size of the two-locus interaction to that of the independent loci, and calculate the probability of misclassifying a unique locus as a shared locus. Both multiplicative-interaction and threshold-interaction models are considered in the simulation. The details of the model settings and the results are summarized in Table 3.5.

Table 3.5 Misclassification rate of unique loci to common loci by MMW and MMW

Disease model	Odds ratio	Multiplicative-interaction		Threshold-interaction	
		MMW	COM	MMW	COM
Disease1: ***A****B**+C Disease2: ***A****B**+D	1.4a; 1.9b; 1.9c	0.115	0.749	0.146	0.903
	1.4a; 1.8b; 1.8c	0.106	0.679	0.151	0.835
	1.4a; 1.7b; 1.7c	0.105	0.587	0.147	0.826
	1.4a; 1.6b; 1.6c	0.089	0.500	0.116	0.808
	1.4a; 1.5b; 1.5c	0.090	0.414	0.141	0.713
	1.4a; 1.4b; 1.4c	0.079	0.319	0.160	0.659
	1.4a; 1.3b; 1.3c	0.075	0.333	0.158	0.567
	1.4a; 1.2b; 1.2c	0.083	0.256	0.149	0.494
	1.4a; 1.1b; 1.1c	0.112	0.261	0.148	0.500

a: Odds ratio for the common risk loci;
b: Odds ratio for the risk locus unique to disease 1.
c: Odds ratio for the risk locus unique to disease 2.

As shown in Table 3.5, as the effect size of risk loci unique to each disease increases, the COM approach is more likely to misclassify them as common risk loci. COM treats all the selected loci as common loci associated with co-morbidity, and thus does not differentiate unique and shared loci. When the effect size of loci is large, regardless of whether the risk loci are

common or unique loci, they are more likely to be selected as the common risk factors by COM. In contrary to COM, MMW considers only loci selected for both conditions as shared loci, and the remaining loci associated with one of the condition as unique loci. Thus it has the capacity of differentiating unique and shared loci. From Table 3.5, regardless of effect size of the unique loci, MMW remains a low and stable misclassification rate.

3.4.4 Application to Coronary Heart Disease and Type II Diabetes

Substantial evidences from both clinical and epidemiological studies suggest a considerable amount of comorbidity exist between cardiovascular disease and type II diabetes. For example, Martin *et al.* reported that in a German cohort, at the time of diagnosis of type II diabetes 22% of patients had coronary heart disease present [34]. According to the American Diabetes Association, an estimates of two third of diabetic people died from cardiovascular disease, and adults with diabetes are at least twice as likely to have heart disease or stroke than those without diabetes. Indeed, the American Heart Association recommends treating diabetes as one of the major controllable risk factors for cardiovascular disease [35, 36]. Various factors can determine the co-occurrences of the two diseases, ranging from genetic predisposition and lifestyle of individual to the general health policy on the public. According to the 13 co-morbidity models proposed by Neale and Kendler, co-morbidity between coronary heart disease and type II diabetes may due to the fact that one of the co-morbid conditions is the cause or consequence of the other [3, 8, 10]. It is also possible that the two diseases share the same or correlated risk factors, such as obesity, physical inactivity, and insulin resistance, making the co-morbid conditions more likely to occur simultaneously [8, 11, 12, 37-44]. Though a large proportion of coronary heart disease cases and type II diabetes cases can be explained by environmental factors, genetics factors also play an important role in predisposing to both diseases. It has been reported

that rs1801282, which is located in gene *PPARG*, is associated with both cardiovascular disease and type II diabetes [47-49]. G allele of SNP rs4420638, which is located 14kb away from *ApoC1* gene and co-inherited with *ApoE*, increases the risk of coronary heart disease as well as type II diabetes [50, 51]. The chromosome 9p21 region has also been identified to be associated with both type II diabetes and cardiovascular disease, though different SNPs were reported to each disease in different studies [25, 52].

In this research, we apply the proposed method to study the co-morbidity between CAD and T2D with datasets from WTCCC. The CAD study data set consists of 1991 individuals and T2D data set contains 1810 subjects. The controls used in this study are obtained from National Blood Service (NBS) and it consists of 1440 individuals. All the three chosen datasets were genotyped on Affymetrix GeneChip Human Mapping 500K Array Set, and called by the algorithm we develop in Chapter 2 of the dissertation. We filter the SNPs that were of poor quality and we further correct for the batch effect. In this analysis, cases are defined as individual with either CAD or T2D, while controls are defined as individuals who have never met the diagnosis criteria of both CAD and T2D. From previous literature, we collected 21 SNPs and 35 SNPs that were available on Affymetrix GeneChip Human Mapping 500K Array Set and had been reported for potential association with CAD and T2D, respectively.

We initiated the analysis by applying the MMW to WTCCC to search for potential joint gene-gene interactions among 56 known CAD and T2D associated SNPs and then permutation test is used to assess the significance of identified associations. The identified SNPs are jointly associated with CAD or T2D and obtain a p-value less than 0. 001. The new method identifies a 3-locus model and 5-locus model for CAD and T2D, respectively. The summary of identified loci by MMW is shown in Table 3.6. It is shown that the model does not identify any SNPs that

are associated with both diseases. The stepwise results for joint gene action for T2D and CAD are shown in Table 3.7 and Table 3.8, respectively. Logistic regression analyses are conducted to further evaluate the effect of each locus contributing to CAD and T2D. All 3-ways and pair-wise interaction effects are evaluated for CAD, and there is no significant evidence to suggest interactions among the three selected loci. The effect estimates obtained from logistic regression with main effect only are shown in Table 3.9. The genotypes GG of SNPs rs2891168 and rs7250581 as well as genotype TT of rs688034 significantly increase the risk of coronary heart disease. For T2D, a logistic regression model with all possible interactions among the 5 selected loci is fitted and a forward stepwise model selection procedure is applied to identify the most parsimonious model, which is shown in Table 3.10. Controlling for high-order interaction effects, genotypes AT and TT of rs4506565, GG of rs1495377, and AA and AC of rs8050136 have significantly increased the risk of T2D.

Table 3.6 Summary of SNPs identified in from WTCCC data sets

SNPs	Allele	Risk Group	Chromosome	Position	Gene	Disease
rs2891168	A/G	{AA;AG} **{GG}**[a]	9	22088619	*CDKN2BAS*	CAD
rs7250581	A/G	{AA;AG} **{GG}**	19	34756236	*POP4*	CAD
rs688034	C/T	{CC;CT} **{TT}**	22	25019635	*SEZ6L*	CAD
rs4506565	A/T	{AA} **{AT;TT}**	10	114746031	*TCF7L2*	T2D
rs9472138	C/T	{CC} **{CT;TT}**	6	43919740	*VEGFA*	T2D
rs7659604	A/G	{AA;AG} **{GG}**	4	122884964	*ANXA5*	T2D
rs1495377	C/G	{CC;CG} **{GG}**	12	69863368	*TSPAN8*	T2D
rs8050136	A/C	**{AA;AC}** {CC}	16	52373776	*FTO*	T2D

a: Bold letter represents the genotype groups that increase the risk of each disease.

Table 3.7 Stepwise result for joint association analysis for T2D

Steps	Selected SNPs	P-values
1	rs4506565	9.341e-09
2	rs4506565, rs9472138	1.345e-12
3	rs4506565, rs9472138, rs7659604	1.220e-17
4	rs4506565, rs9472138, rs7659604, rs1495377	1.024e-23
5 [a]	rs4506565, rs9472138, rs7659604, rs1495377, rs8050136	3.207e-31

a: The most parsimonious model for T2D identified by the MMW approach.

Table 3.8 Stepwise result for joint association analysis for CAD

Steps	Selected SNPs	P-values
1	rs2891168	4.256e-08
2	rs2891168, rs7250581	1.540e-11
3 [a]	rs2891168, rs7250581, rs688034	4.789e-16

a: The most parsimonious model for CAD identified by the MMW approach.

Table 3.9 Logistic regression result for CAD

SNPs	Effect Estimates	Standard Errors	P-values
rs2891168 (GG)[a]	0.4226	0.0795	1.04e-07****
rs7250581 (GG)	0.2745	0.0722	1.43e-04****
rs688034 (TT)	0.4903	0.1150	2.00e-05****

a: Modeled genotypes
**** P-value<0.001

We also realize that in the case where we only know the phenotypic information of one disease, COM method cannot be implemented as there is no "case". Although MMW method does not identify common risk factors contributing to the co-morbidity between CAD and T2D, it still provides an opportunity to identify risk factors that are unique to each disease.

Table 3.10 Logistic regression result for T2D

SNPs	Effect Estimates	Standard Errors	P-values
rs4506565 (AA)[a]	-0.3345	0.1158	0.0039***
rs9472138 (CC)	-0.0668	0.1212	0.5817
rs7659604 (GG)	0.1069	0.1617	0.5087
rs1495377 (GG)	0.4193	0.1853	0.0237**
rs8050136 (CC)	-0.5446	0.1672	0.0010****
rs4506565:rs7659604	-0.3081	0.1532	0.0444**
rs4506565:rs1495377	0.2255	0.2010	0.2620
rs4506565:rs8050136	0.0789	0.1824	0.6654
rs9472138:rs7659604	-0.496	0.2041	0.0150**
rs9472138:rs1495377	-0.504	0.2094	0.0162**
rs9472138:rs8050136	0.0733	0.1940	0.7056
rs7659604:rs1495377	-0.5881	0.2717	0.0304**
rs7659604:rs8050136	0.4428	0.2442	0.0698*
rs1495377:rs8050136	0.8307	0.2791	0.0029***
rs4506565:rs1495377:rs8050136	-0.7194	0.3690	0.0513*
rs9472138:rs7659604:rs1495377	1.095	0.3439	0.0015***
rs9472138:rs7659604:rs8050136	-0.5579	0.3284	0.0893*
rs7659604:rs1495377:rs8050136	-0.7501	0.3754	0.0457**

a: Modeled genotypes
*P-value<0.05
* *P-value<0.05
*** P-value<0.01
**** P-value<0.001

3.5 Discussion

Co-morbidity among complex human diseases is believed to be caused by interplay among multiple genetic variants and environmental determinants. The identification of genetic and environment risk predictors contributing to co-morbidity will promote better understanding of disease etiology and new diagnostic and therapeutic strategies [1-3]. The findings from the discovery process can be enhanced by adopting novel statistical approaches, as demonstrated here. A multivariate joint association approach allowing for gene-gene interactions can facilitate

the detection of genetic variants and gene-gene interaction contributing to co-morbid conditions. To the best of our knowledge, the proposed method, MMW, is one of the first methods, developed for the identification of genetic variants contributing to co-morbid conditions, with the consideration of high-order interactions. Similar to other Mann-Whitney based methods [45], it is a non-parametric approach, which does not assume model of inheritance, and is free from the issues of increasing number of parameters. MMW adopts a forward selection algorithm, which substantially reduces the searching space of interaction combinations and allows for high-order interactions. These features make MMW more appealing for co-morbidity analysis of complex disease with the consideration of possible interactions.

Through simulation, we have shown that MMW attains high power than COM under a variety of disease models, and is robust than COM under different correlation models between co-morbid conditions. We consider this important as our knowledge of disease co-morbidity is limited and the underlying correlation among co-morbid diseases could vary from case to case. Compared with COM, MMW allows for the identification of genetic risk variants common to co-morbid conditions, as well as unique to each co-morbid condition, which leads to a better understanding of relationship among co-morbid diseases. In addition, MMW makes use of the entire sample, which potentially increases the power to indentify genetic variants associated with co-morbid conditions, especially when the co-morbidity rate is low or when few co-morbidity individuals are in the data. In an extreme case, where each dataset is designed to study one of the co-morbid conditions and the information regarding the other disease statuses is not measured, COM is not applicable as there is no case. However, MMW is still has the capacity to identify risk loci common and unique to co-morbidity conditions, as it selects risk loci for each disease and then builds an overall test to assess the association.

The co-morbidity between CAD and T2D is well documented, however, the genetic etiology contributing to the co-morbidity remains largely unknown. To search for the risk factors predisposing to both CAD and T2D, unique to each of CAD and T2D, we apply the proposed method to the data sets obtained from WTCCC. We have identified a 3-locus model for CAD by MMW, and logistic regression analysis suggests no interaction effects present among the selected loci. A 5-locus model for T2D has also been identified by MMW, and among the identified markers 3-way interaction effect and 2-way interaction effect are present. Unfortunately we are unable to find any loci that contribute to both diseases. There are several possible reasons leading to this negative finding. First the proposed method is a candidate gene based approach and the markers used for this comorbidity analysis are those that have been targeted by Affymetrix GeneChip Human Mapping 500K Array Set and reported to be associated with either CAD or T2D. The markers that have not been targeted by Affymetrix 500K SNP array but potentially associated with both T2D and CAD are not included in this analysis, which might be one of the reasons why there is no genetic components found leading to comorbidity between CAD and T2D. Imputations might be needed to enlarge the marker pools. In addition, due to the candidate gene approach, it overlooked those markers that may be associated with both diseases but not have been reported. A modification of the current algorithm is needed to allow for whole genome wide scan. Second, environmental factors such as obesity, physical inactivity and family history, play an important role in both T2D and CAD. However, in the first stage of WTCCC study such environmental factors are not measured. As a consequence, in our analysis none of the environmental factors have been controlled, which may attenuate the genetic effects and the power of detecting loci contributing to both diseases. Third, phenotypic heterogeneity in both T2D and CAD may be another reason for the negative finding, as strong

genetic variants predisposing to specific subset of each disease in a small homogeneous population might be negligible in the whole population[53, 54], which leads to lack of power to identify genetic variants associated with both diseases. A refinement of disease definition, which could be based upon age onset, severity of disease and progression profile, may help to identify genetic variants predisposing to a subset of both diseases.

In conclusion, we have proposed a powerful tool to reveal common genetic variants and interactions contributing to disease co-morbidity, as well as those unique to each co-morbid condition. Through simulations we have demonstrated that our method attains more power compared with the common practice (i.e. the composite phenotype analysis) in a variety of underlying disease models and correlation models between two co-morbid conditions, especially in the case when co-morbidity rate is low. Though we do not identify any loci contributing to both CAD and T2D, further analyses and studies are needed.

REFERENCES

REFERENCES

1. Maj, M., *"Psychiatric comorbidity": an artefact of current diagnostic systems?* Br J Psychiatry, 2005. **186**: p. 182-4.

2. Feinstein, A., *The pre-therapeutic classficaitoin of co-morbidity in chronic disease.* J Chronic Dis, 1970(23): p. 455-68.

3. de Groot, V., et al., *How to measure comorbidity. a critical review of available methods.* J Clin Epidemiol, 2003. **56**(3): p. 221-9.

4. Valderas, J.M., et al., *Defining comorbidity: implications for understanding health and health services.* Ann Fam Med, 2009. **7**(4): p. 357-63.

5. Bayliss, E.A., et al., *Processes of care desired by elderly patients with multimorbidities.* Fam Pract, 2008. **25**(4): p. 287-93.

6. Gijsen, R., et al., *Causes and consequences of comorbidity: a review.* J Clin Epidemiol, 2001. **54**(7): p. 661-74.

7. Campbell-Scherer, D., *Multimorbidity: a challenge for evidence-based medicine.* Evid Based Med, 2010. **15**(6): p. 165-6.

8. Neale, M.C. and K.S. Kendler, *Models of comorbidity for multifactorial disorders.* Am J Hum Genet, 1995. **57**(4): p. 935-53.

9. Rhee, S.H., et al., *The validity of the Neale and Kendler model-fitting approach in examining the etiology of comorbidity.* Behav Genet, 2004. **34**(3): p. 251-65.

10. Simonoff, E., *Extracting meaning from comorbidity: genetic analyses that make sense.* J Child Psychol Psychiatry, 2000. **41**(5): p. 667-74.

11. Youngstrom, E.A., L.E. Arnold, and T.W. Frazier, *Bipolar and ADHD Comorbidity: Both Artifact and Outgrowth of Shared Mechanisms.* Clin Psychol (New York), 2010. **17**(4): p. 350-359.

12. Lind, P.A., et al., *A genomewide association study of nicotine and alcohol dependence in Australian and Dutch populations.* Twin Res Hum Genet, 2010. **13**(1): p. 10-29.

13. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

14. Schlaepfer, I.R., N.R. Hoft, and M.A. Ehringer, *The genetic components of alcohol and nicotine co-addiction: from genes to behavior.* Curr Drug Abuse Rev, 2008. **1**(2): p. 124-34.

15. Bierut, L.J., et al., *A genome-wide association study of alcohol dependence.* Proc Natl Acad Sci U S A, 2010. **107**(11): p. 5082-7.

16. Treutlein, J., et al., *Genome-wide association study of alcohol dependence.* Arch Gen Psychiatry, 2009. **66**(7): p. 773-84.

17. Caporaso, N., et al., *Genome-wide and candidate gene association study of cigarette smoking behaviors.* PLoS One, 2009. **4**(2): p. e4653.

18. Wang, K.S., et al., *A meta-analysis of two genome-wide association studies identifies 3 new loci for alcohol dependence.* J Psychiatr Res, 2011.

19. Bierut, L.J., et al., *A genomic scan for habitual smoking in families of alcoholics: common and specific genetic factors in substance dependence.* Am J Med Genet A, 2004. **124A**(1): p. 19-27.

20. Baum, A.E., et al., *Meta-analysis of two genome-wide association studies of bipolar disorder reveals important points of agreement.* Mol Psychiatry, 2008. **13**(5): p. 466-7.

21. Moskvina, V., et al., *Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk.* Mol Psychiatry, 2009. **14**(3): p. 252-60.

22. Ferreira, M.A., et al., *Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder.* Nat Genet, 2008. **40**(9): p. 1056-8.

23. Sklar, P., et al., *Whole-genome association study of bipolar disorder.* Mol Psychiatry, 2008. **13**(6): p. 558-69.

24. Barrett, J.C., et al., *Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.* Nature genetics, 2008. **40**(8): p. 955-62.

25. Wellcome Trust Case Control Consortium, *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.

26. John, U., et al., *Probabilities of alcohol high-risk drinking, abuse or dependence estimated on grounds of tobacco smoking and nicotine dependence.* Addiction, 2003. **98**(6): p. 805-14.

27. Riala, K., et al., *Teenage smoking and substance use as predictors of severe alcohol problems in late adolescence and in young adulthood.* J Adolesc Health, 2004. **35**(3): p. 245-54.

28. Clark, A., et al., *Chronic infusion of nicotine can increase operant self-administration of alcohol.* Neuropharmacology, 2001. **41**(1): p. 108-17.

29. Dilsaver, S.C., et al., *Migraine Headache in Affectively Ill Latino Adults of Mexican American Origin Is Associated With Bipolarity.* Prim Care Companion J Clin Psychiatry, 2009. **11**(6): p. 302-306.

30. Dilsaver, S.C., et al., *Is a family history of bipolar disorder a risk factor for migraine among affectively ill patients?* Psychopathology, 2009. **42**(2): p. 119-23.

31. Bowden, C.L., et al., *A randomized, placebo-controlled 12-month trial of divalproex and lithium in treatment of outpatients with bipolar I disorder. Divalproex Maintenance Study Group.* Arch Gen Psychiatry, 2000. **57**(5): p. 481-9.

32. Oedegaard, K.J., et al., *A genome-wide association study of bipolar disorder and comorbid migraine.* Genes Brain Behav, 2010. **9**(7): p. 673-80.

33. Oedegaard, K.J., et al., *A genome-wide linkage study of bipolar disorder and co-morbid migraine: replication of migraine linkage on chromosome 4q24, and suggestion of an overlapping susceptibility region for both disorders on chromosome 20p11.* J Affect Disord, 2010. **122**(1-2): p. 14-26.

34. Martin, S., et al., *Epidemiology of complications and total treatment costs from diagnosis of Type 2 diabetes in Germany (ROSSO 4).* Exp Clin Endocrinol Diabetes, 2007. **115**(8): p. 495-501.

35. American Diabetes Association, *Consensus Development Conference on the Diagnosis of Coronary Heart Disease in People with Diabetes.* Diabetes Care, 1988. **21**: p. 1551–1559.

36. American Heart Association. Available from: http://www.heart.org/HEARTORG/Conditions/Diabetes/WhyDiabetesMatters/Cardiovascular-Disease-Diabetes_UCM_313865_Article.jsp.

37. Travers, M.E. and M.I. McCarthy, *Type 2 diabetes and obesity: genomics and the clinic.* Hum Genet, 2011. **130**(1): p. 41-58.

38. McCarthy, M.I., *Genomics, type 2 diabetes, and obesity.* N Engl J Med, 2010. **363**(24): p. 2339-50.

39. Johnson, R.L., S.M. Williams, and I.J. Spruill, *Genomics, nutrition, obesity, and diabetes.* J Nurs Scholarsh, 2006. **38**(1): p. 11-8.

40. Stamler, J., *Established major risk factors.*, in *Coronary heart disease epidemiology*, M. Marmot and P. Elliot, Editors. 1992, Oxford Univeristy Press: New York, NY.

41. Vasan, R.S., et al., *Relative importance of borderline and elevated levels of coronary heart disease risk factors.* Ann Intern Med, 2005. **142**(6): p. 393-402.

42. Reaven, G.M. and Y.D. Chen, *Insulin resistance, its consequences, and coronary heart disease. Must we choose one culprit?* Circulation, 1996. **93**(10): p. 1780-3.

43. Schernthaner, G., *[Hypertension, insulin resistance and diabetes mellitus: pathophysiological interactions and therapeutic consequences]*. Wien Klin Wochenschr, 1990. **102**(24): p. 707-12.

44. Smit, J.W. and J.A. Romijn, *Acute insulin resistance in myocardial ischemia: causes and consequences.* Semin Cardiothorac Vasc Anesth, 2006. **10**(3): p. 215-9.

45. Lu, Q., et al., *A Likelihood Ratio based Mann-Whitney Approach Finds Novel Replicable Joint Gene Action for Type 2 Diabetes.* Submitted to Genetic Epidemiology, 2011.

46. Marchini, J., P. Donnelly, and L.R. Cardon, *Genome-wide strategies for detecting multiple loci that influence complex diseases.* Nat Genet, 2005. **37**(4): p. 413-7.

47. Bego, T., et al., *Association of PPARG and LPIN1 gene polymorphisms with metabolic syndrome and type 2 diabetes.* Med Glas Ljek komore Zenicko-doboj kantona, 2011. **8**(1): p. 76-83.

48. Regieli, J.J., et al., *PPAR gamma variant influences angiographic outcome and 10-year cardiovascular risk in male symptomatic coronary artery disease patients.* Diabetes Care, 2009. **32**(5): p. 839-44.

49. Zeggini, E., et al., *Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.* Science, 2007. **316**(5829): p. 1336-41.

50. Willer, C.J., et al., *Newly identified loci that influence lipid concentrations and risk of coronary artery disease.* Nature genetics, 2008. **40**(2): p. 161-9.

51. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science, 2007. **316**(5829): p. 1331-6.

52. Broadbent, H.M., et al., *Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p.* Hum Mol Genet, 2008. **17**(6): p. 806-14.

53. Morris, A.P., et al., *A powerful approach to sub-phenotype analysis in population-based genetic association studies.* Genet Epidemiol, 2010. **34**(4): p. 335-43.

54. Warde-Farley, D., et al. *Mixture model for sub-phenotyping in GWAS*. 2011 [cited 2011 Nov.15]; Available from: http://psb.stanford.edu/psb-online/proceedings/psb12/warde-farley.pdf.

CHAPTER 4

A TWO STAGE MODEL FOR DETECTING DIFFERENTIALLY EXPRESSED GENES ACCOUNTING FOR THE VARIABILITY IN THE DATA PREPROCESSING STEP

4.1 Abstract

Since 1990s, gene expression profiling has been widely used in biological research to study pathways and differentially expression genes (DEG) contributing to complex human diseases. Typically, the DEG detection algorithm involves both data preprocessing and DEG analysis, and the DEG analysis is conducted based on the summarized gene expression level estimated from data preprocessing. Various methods have been proposed for each of the DEG analysis and microarray data preprocessing, but the variability in the estimated gene expression level obtained from data preprocessing has been overlooked, which may lead to false positive and false negative findings. In this study, we develop a two-stage model based on LIMMA (referred to as two-stage LIMMA), while incorporating the variability in the estimated gene expression level obtained from a positional-dependent-nearest-neighbor (PDNN) model. We demonstrate the utility of our method through simulations and the analyses of the Affymetrix Latin Square Data, both of which show that our method outperforms LIMMA. Finally, we apply the two-stage LIMMA method to detect the DEGs from a cervical cancer study. Our findings replicate most of the original findings, and identify new genes which are potentially associated with cervical cancer.

Keywords: Affymetrix, hierarchical model, hybridization, LIMMA, two-stage model.

4.2. Introduction

Since the introduction of microarray technology in the mid 1990s[1], gene expression profiling has been used extensively in biological research to study biological pathways and identify genes and gene sets associated with complex human diseases [2-4]. Typically, investigators seek to obtain a list of differentially expressed genes (DEGs) under different treatment conditions from the gene expression microarray[5],and for technical reasons the analysis usually involves both data preprocessing and DEG analysis based on the preprocessed data. The data preprocessing usually comprises of background correction, normalization and summarization[6] to remove noise and improve the correlation between biological effect and measured intensities. Numerous preprocessing algorithms have been developed during the past few years, such as MAS5.0, dChip[7], RMA[8], GCRMA[9], PerfectMatch[10], ZAM[11], and combined relatively well with the downstream analyses, such as SAM[12], iBMT[13], and LIMMA[14]. However, little attention has been paid to the underlying mechanisms of the hybridization process which may bias the expression summary, and consequently the DEG analysis can be affected leading to false positive and false negative findings. Extensive studies have shown that probe intensities depend on not only the concentration of the target genes but also the binding affinity of the probes ([15-18]). For example, both RMA[8] and GCRMA[9] normalize data with quantile normalization and summarize the intensities with a robust linear model fit, while GCRMA generally performs better than RMA because GCRMA models the probe sequence information and corrects background accordingly instead of using a global background correction as RMA does. It highly suggests that probe-set expression level summary can be improved by modeling the mechanism of the array underlying hybridization process.

In the past few years, various approaches have been proposed to identify differentially expressed genes, and many of them are modification of t-statistic based methods, as the estimates of variance based on classical t-statistic are not accurate due to the small sample size (e.g. 3 treatments vs. 3 controls) and modification of estimated variance is one of the key issues for the DEG analysis. For example, SAM [12] evaluated the significance of gene expression level based on empirical null distribution, in which the estimated variance is adjusted by a fudge factor. Regularized t-test[19] assumed an empirical relationship between the average gene expression level and variance, and estimated the variance with hierarchical Bayesian models. Variance estimation of moderated-t test proposed by Smyth [14] is based on empirical Bayesian approach, and the DEG hypothesis can be tested within the traditional linear models framework. Similar to the moderated t-test, iBMT[13] employed an empirical Bayesian approach to estimate the variance but it also accounts for the dependence of variance on gene expression level. In addition to t-statistic based method, weighted average difference[20], fold change, rank based methods[21] are also widely used. Most of the current methods [12-14]on DEG analysis highly depend on the expression summary obtained from data preprocessing method, but the variability in the estimated gene expression summary has been overlooked, which may lead to inefficient and invalid results for the DEG analysis as the association results based on estimated gene expression level with large variability in the data preprocessing step are lack of confidence.

Zhang *et al.*[10, 18] developed a positional-dependent-nearest-neighbor (PDNN) model in which they decomposed the observed probe intensities into specific binding signals, non-specific binding signals and array background. It utilizes probe sequence information and models the probe binding with a weighted pair-wise stack energy approximation and Langmuir-like adsorption model. The PDNN model mimics the mechanism of hybridization process and the

estimated gene-expression level is in general consistent with the biological signals, but the PDNN model, just as the other preprocessing algorithms, does not provide uncertainty in the estimated gene-expression level. The uncertainty in the estimated expression level should be taken into account and traditional DEG analysis algorithms need to be adapted accordingly.

In this paper, we incorporate the uncertainty introduced in data preprocessing step into DEG analysis by using a two-stage model. The proposed method has the following advantages: It 1) estimates the gene expression level summary based solely on individual's data; 2) provides a measure of the uncertainty of the preprocessing algorithm; 3) incorporates the uncertainty measure into the DEG analysis. The rest of the paper is organized as follows. In Section 2, we first give a brief introduction to the PDNN model, and then we outline our procedure to estimate the uncertainty of gene-expression level obtained from PDNN model. Furthermore, we propose a hierarchical two-stage model in which we take into account the variability of estimated gene expression level obtained from data preprocessing. In Section 3, the proposed method is compared with other methods based on simulated data and real microarray data in two studies, including a spike-in dataset obtained from Affymetrix Inc. and a cervical cancer data set. In the last section, we discuss and summarize our findings.

4.3. Method

4.3.1 The PDNN model

We use PDNN[10] as our data preprocessing model, which is constructed based on the rationale that the measured probe intensities can be decomposed into three parts: the specific binding intensity, the non-specific binding and the array background intensities, as shown in equation (1). The specific binding intensities come from the binding between the probe and a target transcript

with the exact complementary sequence, whereas the non-specific binding intensities result from the binding between probe and transcripts with many mismatches

$$I_{ij} = \frac{N_j}{1+\exp(E_{ij})} + \frac{N^*}{1+\exp(E^*_{ij})} + b + \varepsilon_{ij} \quad (1)$$

where I_{ij} is observed intensity for probe *i* and gene *j*, N_j is the number of target concentration of gene *j*, N^* is the total number of transcripts for the non-specific binding, *b* is the array background, and E_{ij} is the binding free energy of the specific binding, and E^*_{ij} is the average binding free energy for non-specific binding.

The E_{ij} and E^*_{ij} are estimated by weighted sums of stacking energies, and it can be calculated through equation (2)

$$E_{ij} = \sum \omega_k \sigma(b_k, b_{k+1}), \quad E^*_{ij} = \sum \omega^*_k \sigma^*(b_k, b_{k+1}) \quad (2)$$

where $(b_1, b_2, b_3 ..., b_{25})$ is the probe sequence; ω_k and ω^*_k are weighted factors depending on the position of the nucleotide along the probe, and $\sigma(b_k, b_{k+1})$ and $\sigma^*(b_k, b_{k+1})$ are pair-wise stacking energy for specific binding and non-specific binding, respectively.

With the energy parameters the gene expression level can be calculated as

$$N_j = \frac{\sum_i \left\{ [I_{ij} - b - N^* / (1+\exp(E^*_{ij}))] / \lambda_{ij} \right\}}{\sum_i \left[1/(1+\exp(E_{ij})) / \lambda_{ij} \right]} \quad (3)$$

where $\lambda_{ij} = \sqrt{I_{ij}(1+\exp(E_{ij}))}$

4.3.2 Two-Stage Model

Data Preprocessing-First Stage Model

As the PDNN algorithm models the underlying hybridization process, the binding free energy should be independent of labs and samples, and remain the same for the same array platform, which have been shown by Wan *et al*[16]. Given the parameters trained through a Monte Carlo simulation procedure provided by Zhang *et al.* [10], the target concentrations as well as overall non-specific binding molecular concentration can be estimated through a simple linear regression shown in equation (4).

$$I_{mkjp} = N_{mkj}\phi(E_{jp}) + N^*\phi(E^*_{jp}) + b + \varepsilon_{jp},\ \phi(x) = \frac{1}{\exp(x)+1} \quad (4)$$

where I_{mkjp} is the observed intensities for probe *p* of gene *j* on array *k* under experimental condition *m*, *j*=1, 2, …, *n*, *k*=1,2,…m_K, *m*=1,2,…*M*; N_{mkj} is the true target concentration of gene *j* on array *k* under condition *m*; E_{jp} is the binding free energy of the specific binding of probe *p* of gene *j*, and E^*_{jp} is the average binding free energy for non-specific binding of probe *p* of gene *j*.

Let

$$\phi(E_j^*) = (\phi(E_{j1}^*), \phi(E_{j2}^*), ..., \phi(E_{jn_j}^*))^T$$

$$\phi(E^*) = (\phi(E_1^*), \phi(E_2^*), ..., \phi(E_n^*))^T$$

$$\phi(E_j) = (\phi(E_{j1}), \phi(E_{j2}), ..., \phi(E_{jn_j}))^T$$

$$X_1 = (J_N, \phi(E^*))$$

$$X_2 = diag(\phi(E_j))$$

$$X = (X_1, X_2)_{N \times (n+1)}$$

where n denotes the total number of genes on the array; n_j denotes the total number of probes for gene j; $N = \sum_{j=1}^{n} n_j$ denotes the total number of probes and J_N is a $N \times 1$ matrix with all elements equal to 1.

As the dimension of the design matrix X is large, we apply the AS274 algorithm[22] implemented in the BIGLM of R package, to estimate the gene expression level as well as the variance of the estimates. In general, the AS274 algorithm constructs a design matrix based on part of the data, and gradually adds the rest of data into the design matrix to improve the efficiency of the algorithm. It first diagonalizes the design matrix with Cholesky factorization, and then it adopts a series of planar rotation to annihilate the added row and keep the design matrix diagonal, which is critical to improve the efficiency of the algorithm. As X^TX is a positive definite symmetric matrix, and X^TX can be written as $X^TX=R^TR$ by Cholesky factorization, where R is the unique upper triangular matrix. For example, let $X^TX = A = \begin{bmatrix} 4 & 2 & -2 \\ 2 & 10 & 2 \\ -2 & 2 & 5 \end{bmatrix}$,

then the Cholesky factorization for this matrix would be

$A = \begin{bmatrix} 4 & 2 & -2 \\ 2 & 10 & 2 \\ -2 & 2 & 5 \end{bmatrix} = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3 & 1 \\ 0 & 0 & \sqrt{3} \end{bmatrix}^T \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3 & 1 \\ 0 & 0 & \sqrt{3} \end{bmatrix} = R^T R$. Consider the case when one new line of observation is added, the upper triangular factorization can be updated by a sequence of planar rotation on the left to annihilate the elements of the added row. For example, consider a simple example when one new observation is added, we have $\begin{bmatrix} R \\ x \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{21} & r_{22} \\ 0 & 0 & r_{23} \\ x_1 & x_2 & x_3 \end{bmatrix}$. We first apply a planar rotation to annihilate x_1, x_2 and x_3 are changed to be x'_2 and x'_3. The next rotation is to annihilate x'_2, and x'_3 is replaced by x''_3. Finally, another rotation is conducted to annihilate x''_3. In our settings, due to the large dimension of the design matrix we divide the design matrix into small blocks, roughly 1000 genes per block, and updated the estimated gene expression level ($\hat{N}_{mkj}$) and the estimated variances (s^2_{mkj}) of the estimates accordingly.

DEG detection-Second Stage Model

Let $\hat{Y}^T_{mj} = (\hat{N}_{m1j}, \hat{N}_{m2j}, ..., \hat{N}_{mm_Kj})$, $\hat{Y}^T_j = (\hat{Y}^T_{1j}, \hat{Y}^T_{2j}, ..., \hat{Y}^T_{Mj})$ and

$\sum_{mj} = \text{var}(\hat{Y}^T_{mj}) = diag(\sigma^2_{mkj})$. We assume that the summarized gene expression level obtained from data preprocessing algorithm follows a normal distribution with mean equal to the true

target concentrations of each sample (i.e. $\hat{Y}_j^T \sim N(Y_j^T, \Sigma)$, where Y_j^T is the true target concentration vector for gene *j* and $\Sigma = \begin{vmatrix} \Sigma_{1j} & & & & & \\ & \Sigma_{2j} & & & & \\ & & . & & & \\ & & & . & & \\ & & & & . & \\ & & & & & \Sigma_{M-1j} \\ & & & & & & \Sigma_{Mj} \end{vmatrix}$

We further assume $Y_j = X\alpha_j + \delta, \delta \sim N(0, \sigma_j^2)$, where *X* is a full rank matrix and α_j is a coefficient vector. Suppose certain contrast of the coefficient is of biological interest and is defined by $\beta_j = C^T \alpha_j$. We assume that it is of interest to test whether β_j is equal to zero (i.e. $C^T \alpha_j = 0$). Similar to LIMMA, a linear model is fitted to the target concentration for each gene to obtain coefficient estimator $\hat{\alpha}_j$, and therefore we have

$$E(\hat{\beta}_j) = E(C^T \hat{\alpha}_j) = C^T E(\hat{\alpha}_j) = C^T E(E(\hat{\alpha}_j \mid \hat{Y}_j)) = C^T E((X^T X)^{-1} X^T \hat{Y}_j)$$
$$= C^T (X^T X)^{-1} X^T E(Y_j) = C^T (X^T X)^{-1} X^T X \alpha_j = C^T \alpha_j$$

and

$$\text{var}(\hat{\beta}_j) = \text{var}(C^T \hat{\alpha}_j) = C^T \text{ var}(\hat{\alpha}_j) C = C^T \left[E(\text{var}(\hat{\alpha}_j \mid \hat{Y}_j)) + \text{var}(E(\hat{\alpha}_j \mid \hat{Y}_j)) \right] C$$
$$= C^T \left[(X^T X)^{-1} X^T \text{ var}(\hat{Y}_j) X (X^T X)^{-1} + (X^T X)^{-1} \sigma_j^2 \right] C$$
$$= C^T \left[(X^T X)^{-1} X^T \Sigma X (X^T X)^{-1} + (X^T X)^{-1} \sigma_j^2 \right] C = C^T \left[UVU + U\sigma_j^2 \right] C$$

where $U = (X^T X)^{-1}, V = X^T \sum X$

The variance of estimated gene expression level (i.e. σ^2_{mkj}) is estimated by s^2_{mkj}, and it highly depends on the data preprocessing algorithm. Therefore, here we introduce a scale parameter ρ, which depends on both the chosen data preprocessing algorithm and the goals of biological research, to adjust for the estimated effects of data preprocessing. Through both simulations and Affymetrix spike-in experiment data sets, ρ can be set in the range of 0.10 to 0.15 to obtain optimal DEG detection. In practice, we recommend to set ρ at 0.10. The parameters σ_j^2 are estimated by the same procedure as LIMMA. We define the test statistic as

$t = \dfrac{\hat{\beta}_j - C^T \alpha_j}{\sqrt{C^T \left[UVU + U\sigma_j^2 \right] C}}$. The degree of freedom taken by the data preprocessing algorithm is large, and therefore approximately $t \sim N(0, C^T \left[UVU + U\sigma_j^2 \right] C)$.

4.4. Results

4.4. 1 Simulations

We conduct a series of simulations to assess the performance of our two-stage LIMMA method and LIMMA under different model settings. We assume that the summarized intensities for each gene N_{mkj} follow a Gamma distribution with an exchangeable Gamma (0.8, 15) prior on rate parameter and a constant shape parameters equal to 5. We assume 12 probes are used to annotate each gene and each raw observed probe intensity I_{mkjp} is simulated from a normal distribution

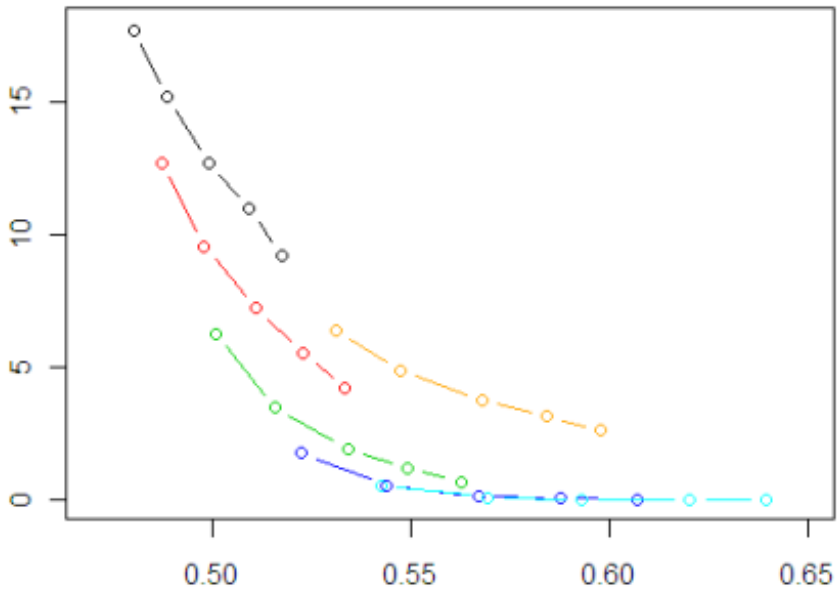
with mean N_{mkj} and variance σ^2_{mkj}, where *p*=1,2,…12. . σ^2_{mkj} is set to be proportional to σ^2_j, with $\sigma^2_{mkj} / \sigma^2_j$ ratio ranging from 0.5 to 1.5, to account for different level of uncertainty in the first stage model. The parameters in the above simulations are set in the proximity of real data. We fix the total number of genes to be 10000, the number of samples per group to be three, and we vary the percentage of differentially expressed genes from 1% to 10%. For each model and condition, we generate 1000 datasets and compare the false positive and false negative findings between the two-stage LIMMA and LIMMA.

The results of the simulations are summarized in Table 4.1. We have observed that when ρ sets to be 0.10, compared with LIMMA two-stage model has a considerable lower false positive, while keeping the false negative comparable. In addition, as expected we have observed that as the $\sigma^2_{mkj} / \sigma^2_j$ ratio increases, the false negative rates in both two-stage LIMMA and LIMMA increase. However, our method still has the capacity to well control the false positive findings, as the false positive findings by LIMMA is more than twice of the positive findings by our method with $\rho = 0.1$ and $\sigma^2_{mkj} / \sigma^2_j = 1.5$.

Table 4.1 The comparison false positive (FP) and false negative rate (FNR) between LIMMA and Two-Stage LIMMA with GG model (FNR, FP)

No. of DEG genes	Methods	ρ	$\sigma^2_{mkj} / \sigma^2_j$ ratios				
			0.5	0.75	1	1.25	1.5
100	LIMMA		(0.53, 6.40)	(0.55, 4.88)	(0.57, 3.78)	(0.58, 3.14)	(0.60, 2.63)
	Two-stage	0.05	(0.49, 12.71)	(0.50, 9.56)	(0.51, 7.22)	(0.52, 5.54)	(0.53, 4.21)
		0.10	(0.50, 6.29)	(0.52, 3.48)	(0.53, 1.93)	(0.55, 1.21)	(0.56, 0.68)
		0.25	(0.53, 0.95)	(0.56, 0.24)	(0.58, 0.09)	(0.60, 0.02)	(0.62, 0.01)
200	LIMMA		(0.50, 15.34)	(0.51, 12.71)	(0.53, 10.39)	(0.54, 8.84)	(0.55, 7.68)
	Two-stage	0.05	(0.46, 26.81)	(0.47, 20.99)	(0.48, 16.52)	(0.49, 13.01)	(0.50, 10.26)
		0.10	(0.47, 14.98)	(0.49, 9.23)	(0.51, 5.69)	(0.52, 3.65)	(0.53, 2.23)
		0.25	(0.50, 3.18)	(0.53, 1.06)	(0.55, 0.35)	(0.57, 0.10)	(0.60, 0.04)
500	LIMMA		(0.44, 48.79)	(0.46, 43.01)	(0.47, 37.65)	(0.48, 33.93)	(0.49, 29.66)
	Two-stage	0.05	(0.42, 69.74)	(0.43, 58.54)	(0.44, 48.66)	(0.45, 40.77)	(0.46, 33.48)
		0.10	(0.43, 45.30)	(0.45, 31.34)	(0.46, 21.49)	(0.47, 15.32)	(0.49, 10.38)
		0.25	(0.46, 13.94)	(0.49, 5.64)	(0.51, 2.47)	(0.53, 1.02)	(0.55, 0.45)
800	LIMMA		(0.42, 86.40)	(0.43, 78.14)	(0.44, 69.86)	(0.45, 63.67)	(0.46, 58.54)
	Two-stage	0.05	(0.39, 114.02)	(0.40, 97.43)	(0.41, 83.03)	(0.42, 69.99)	(0.43, 60.18)
		0.10	(0.40, 79.45)	(0.42, 57.75)	(0.43, 41.69)	(0.45, 30.40)	(0.46, 22.27)
		0.25	(0.44, 28.75)	(0.46, 13.47)	(0.48, 6.36)	(0.50, 3.02)	(0.53, 1.52)
1000	LIMMA		(0.40, 110.85)	(0.41,102.54)	(0.42, 92.49)	(0.43, 85.38)	(0.44, 79.28)
	Two-stage	0.05	(0.38, 141.87)	(0.39,123.47)	(0.40,105.46)	(0.41, 91.57)	(0.42, 78.84)
		0.10	(0.39, 101.78)	(0.41, 76.34)	(0.42, 56.32)	(0.43, 42.56)	(0.44, 31.50)
		0.25	(0.42, 40.86)	(0.45, 19.50)	(0.47, 10.12)	(0.49, 4.99)	(0.51, 2.50)

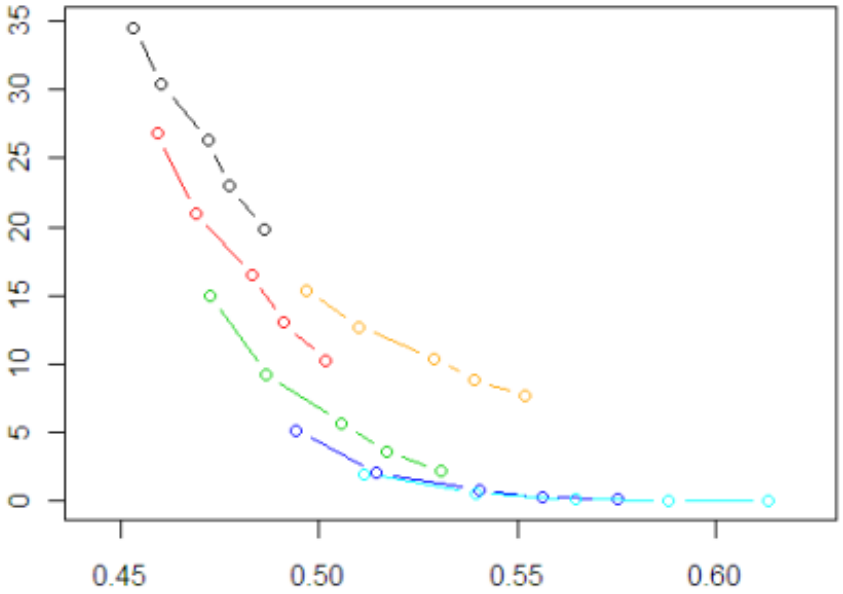
To investigate the optimal range of ρ, we calculate the false positive and false negative at different ρ values. As shown in Figure 4.1, there is a trade-off between false positive and false negative. In general, when ρ is small, we tend to have more false positive and fewer false negative. In contrary, when ρ is large, the false positive rate is well controlled while we tend to have more false negative. In practice, we recommend to set ρ in the range of 0.1 to 0.15 to achieve an optimal performance of the proposed method.



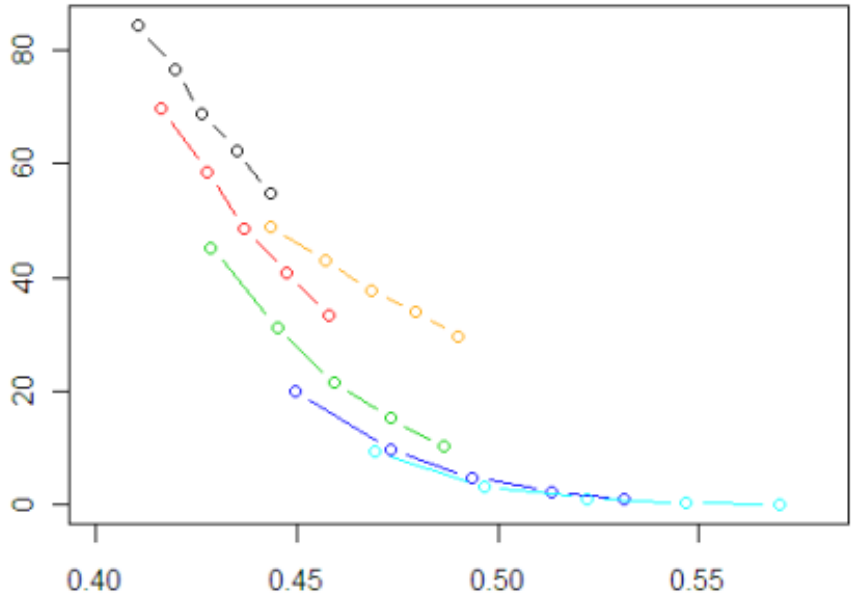
a. Total Number of DEG=100.

Figure 4.1 False positive rate and false negative with different ρ values for GG models. Cyan: 2-stage Limma with $\rho = 0.3$; Blue: 2-stage Limma with $\rho = 0.2$; Green: 2-stage Limma with $\rho = 0.1$; Red: 2-stage Limma with $\rho = 0.05$; Black: 2-stage Limma with $\rho = 0.03$; and Orange: Limma.

Figure 4.1 *cont'd*

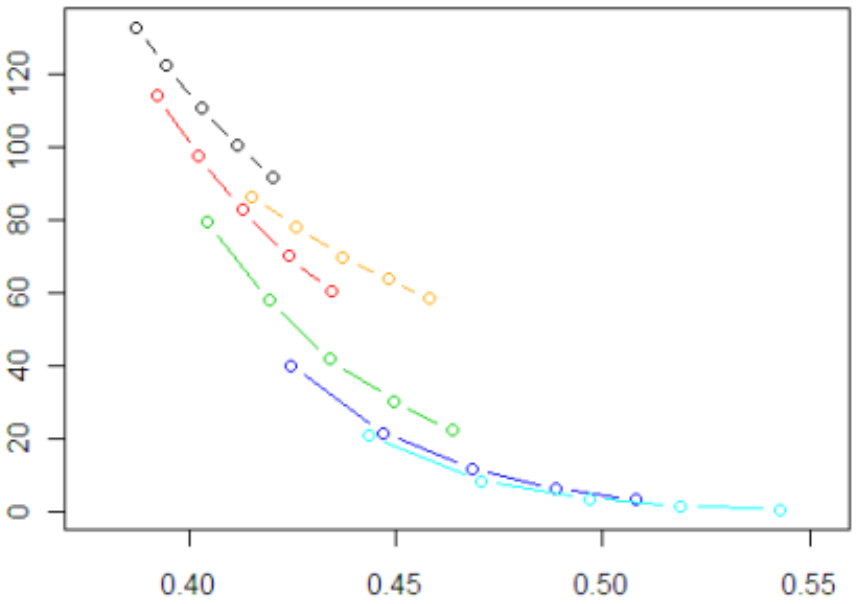


b. Total Number of DEG=200.

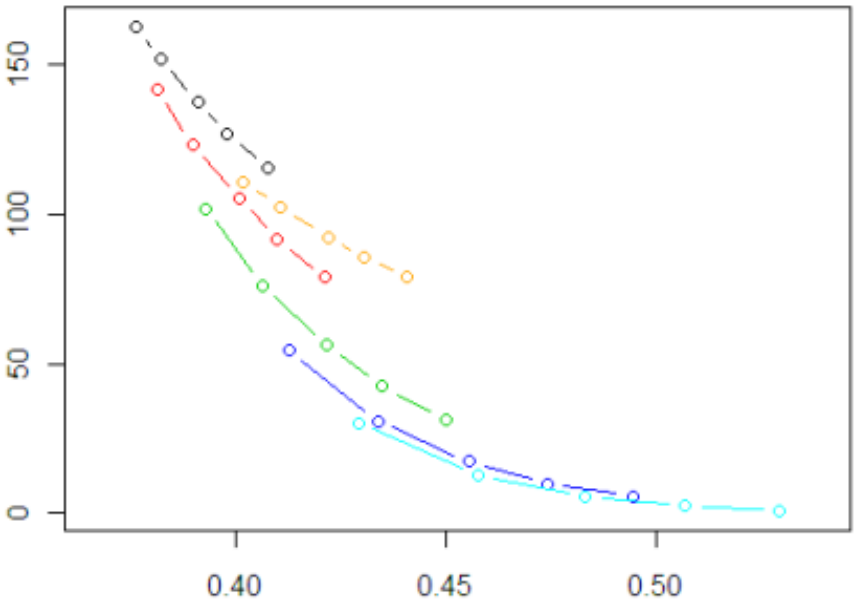


c. Total Number of DEG=500.

Figure 4.1 *cont'd*



d. Total Number of DEG=800.



e. Total Number of DEG=1000.

4.4.2 Affymetrix spike-in experiment

The Human Genome U133A Spike-in data set was downloaded from http://www.affymetrix.com/support/technical/sample_data/datasets.affx. This data set consists of 3 replicates per experimental conditions. A total of 42 spiked-in genes at concentrations ranging from 0.125pM to 512pM were grouped into 14 groups. In addition to the 42 spiked-in genes, Sheffler *et al.* and Lo *et al.* claim that another 20 genes should be treated as spiked-in genes. In addition, 3 genes with probe sequence exactly matching those of spiked-in genes were also considered to be spiked-in genes. Same as Lo *et al.*, a total of 65 genes are treated as spiked-in genes in our analyses.

To evaluate the performance of our two-stage LIMMA method and LIMMA, all 14 different experimental groups are compared with each other and a total of 91 (a combination of C_2^{14}) comparisons were made. Each comparison consists of 3 replicates from each of the experimental conditions. We use PDNN model to preprocess the data for our method and LIMMA, and we also used RMA to preprocess the data for LIMMA. As shown in Table 4.2, with the same preprocessing method, our method significantly outperforms LIMMA. Even with the default preprocessing algorithm for LIMMA, our method performs similar if not better than LIMMA for all comparisons. For some of the scenarios (asterisked in Table 4.2), our method performs significantly better than LIMMA. For example, the maximum number of false positive for our method is 45, whereas the maximum number of false positive for LIMMA is 1443. Among the 91 comparisons, there are 9 times that LIMMA obtained more than 100 false positive findings.

Table 4.2 Comparison between LIMMA and Two-Stage LIMMA with Latin Square Data

Sample 1	Sample 2	LIMMA with RMA		LIMMA with PDNN		Two-Stage LIMMA	
		FNR	FP	FNR	FP	FNR	FP
1	2	0.31	0	0.31	8	0.34	3
1	3	0.15	1	0.11	44	0.17	10
1	4	0.08	6	0.03	110	0.08	13
1	5	0.05	6	0.02	262	0.05	18
1	6	0.03	0	0.00	24	0.00	14
1	7	0.00	2	0.00	141	0.00	12
1	8	0.00	8	0.00	375	0.00	37
1	9	0.02	24	0.00	342	0.00	23
1	10	0.02	14	0.00	234	0.00	26
1	11	0.06	5	0.03	219	0.05	27
1	12	0.05	9	0.03	152	0.06	14
1	13	0.09	4	0.08	71	0.12	8
1	14	0.20	7	0.20	23	0.26	9
2	3	0.49	0	0.54	13	0.55	6
2	4	0.09	1	0.08	15	0.08	11
2	5	0.18	1	0.17	30	0.25	11
2	6	0.12	0	0.17	6	0.11	9
2	7	0.02	0	0.00	23	0.00	13
2	8	0.00	20	0.00	92	0.00	23
2	9	0.02	3	0.00	47	0.00	19
2	10	0.02	6	0.00	50	0.00	22
2	11	0.03	6	0.02	35	0.02	16
2	12	0.05	7	0.03	25	0.03	14
2	13	0.05	4	0.06	30	0.08	9
2	14	0.15	7	0.14	17	0.14	9
3	4	0.09	1	0.06	33	0.15	3
3	5	0.29	3	0.15	109	0.40	6
3	6	0.25	0	0.23	13	0.23	9
3	7	0.00	2	0.00	68	0.00	11
3	8	0.00	93	0.00	625	0.00	26
3	9	0.00	7	0.00	142	0.00	15
3	10	0.00	11	0.00	130	0.00	24
3	11	0.00	18	0.00	209	0.00	23
3	12	0.02	6	0.00	51	0.02	12
3	13	0.00	13	0.00	122	0.03	13
3	14	0.05	13	0.05	61	0.08	10
4	5	0.17	11	0.06	117	0.25	3
4	6	0.38	0	0.37	5	0.35	4
4	7	0.18	20	0.00	1427	0.18	12
4	8	0.00	1443*	0.00	5427	0.00	45

Table 4.2 *(cont'd)*

Sample 1	Sample 2	LIMMA with RMA		LIMMA with PDNN		Two-Stage LIMMA	
		FNR	FP	FNR	FP	FNR	FP
4	9	0.00	247	0.00	1486	0.00	24
4	10	0.02	54	0.00	1360	0.00	37
4	11	0.00	253	0.00	6299	0.00	30
4	12	0.00	120	0.00	828	0.00	19
4	13	0.00	47	0.00	621	0.00	12
4	14	0.02	22	0.02	248	0.03	12
5	6	0.43	0	0.42	3	0.43	2
5	7	0.08	0	0.00	354	0.06	8
5	8	0.00	1097*	0.00	3292	0.00	26
5	9	0.00	167	0.00	1151	0.00	23
5	10	0.02	37	0.00	627	0.00	29
5	11	0.00	83	0.00	2674	0.00	27
5	12	0.00	27	0.00	578	0.00	16
5	13	0.00	32	0.00	316	0.00	16
5	14	0.00	21	0.00	236	0.02	16
6	7	0.31	0	0.28	10	0.29	4
6	8	0.11	1	0.11	8	0.09	9
6	9	0.06	1	0.05	28	0.05	17
6	10	0.05	4	0.03	23	0.02	16
6	11	0.03	3	0.02	15	0.00	13
6	12	0.02	3	0.00	20	0.00	16
6	13	0.02	6	0.00	29	0.00	16
6	14	0.00	5	0.00	22	0.00	20
7	8	0.12	0	0.11	22	0.14	1
7	9	0.06	23	0.06	342	0.11	14
7	10	0.06	12	0.06	76	0.06	18
7	11	0.02	5	0.02	48	0.02	13
7	12	0.02	5	0.00	59	0.00	15
7	13	0.02	5	0.00	111	0.00	15
7	14	0.02	6	0.00	50	0.00	18
8	9	0.15	1224*	0.14	2453	0.26	18
8	10	0.14	23	0.12	146	0.14	19
8	11	0.08	4	0.03	51	0.08	18
8	12	0.03	23	0.00	180	0.02	27
8	13	0.00	103	0.00	755	0.00	29
8	14	0.00	507*	0.00	733	0.00	31
9	10	0.32	3	0.22	20	0.34	3
9	11	0.12	10	0.06	540	0.12	12
9	12	0.08	9	0.06	79	0.08	17
9	13	0.08	8	0.02	186	0.06	18
9	14	0.02	6	0.00	66	0.02	21

Table 4.2 *(cont'd)*

Sample 1	Sample 2	LIMMA with RMA		LIMMA with PDNN		Two-Stage LIMMA	
		FNR	FP	FNR	FP	FNR	FP
10	11	0.18	6	0.20	18	0.22	9
10	12	0.11	4	0.12	54	0.14	19
10	13	0.09	6	0.05	63	0.09	21
10	14	0.03	8	0.00	61	0.02	25
11	12	0.26	2	0.25	28	0.29	10
11	13	0.17	12	0.15	90	0.17	17
11	14	0.11	10	0.12	68	0.12	22
12	13	0.25	4	0.23	32	0.31	7
12	14	0.12	7	0.09	32	0.15	9
13	14	0.22	4	0.18	18	0.26	7
Mean		0.08	66.49	0.07	412.48	0.09	15.97

4.4.3 Cervical Cancer Data

The cervical cancer data was downloaded from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9750. This data set consists of 33 primary tumors and 24 normal cervical epithelium, and they are profiled with Affymetrix U133A oligo-nucleotide microarray. It has been reported that chromosome 20 plays an important role in the tumor pathophysiology. To further demonstrate the utility of the proposed method, we analyze the gene expression profiles on chromosome 20. A total of 571 probe sets are present on this Gene Chip, and they represent 4.6% of the genome (3% on 20q and 1.6% on 20p). With FDR adjusted p-value less than 0.01, a total of 106 probes and 146 probes are found to be differentially expressed by our Two-stage LIMMA method and LIMMA (Table 4.3 and Table 4.4), respectively. There are 95 probes that have been identified by both our method and LIMMA. Among the 106 identified differentially expressed probes, 13 probes are down-regulated and the rest are up-regulated compared with normal samples. The details of the results are summarized in

Table 4.3. Of the 93 over-expressed probes 22 were mapped to 20p, and the remaining is mapped to 20q. Consistent with the findings from Scotto *et al.*, most of the over-expressed genes located on chromosome 20 belong to several functional groups. For instance, AURKA, DSN1, CDC25B, SYCP2, UBE2C, AURKA, TPX2, PCNA, MYBL2, KIF3B, and E2F1 are associated with cell cycle regulation; CSE1L, SOX12, CSE1L, ARFGAP1, RALY, SNRPB, CTNNBL1, PSMA7, SNRPB, SNRPB2, DDX27 and CSE1L are associated with nuclear function; RPN2, RPN2, B4GALT5, B4GALT5, RPN2 are associated with transferase; LAMA5 and PSMA7 are associated with viral replication; AHCY and C20orf20 are associated with methylation and chromatin remodeling, and MMP9 and WFDC2 are associated with endopeptidase activity. In addition to the genes identified by Scotto *et al.*, we also find some up-regulated genes that are possibly associated with cervical cancer. For example, we find TNFRSF6B (fold change=2.0), which belongs to the tumor necrosis factor receptor superfamily and over-expression of this gene is present for gastrointestinal tract tumors. GMEB2 (fold change=1.4), which is a member of KDWK gene family, is one of the essential factors for parvovirus DNA replication. STAU1 (fold change=1.2) is involved in the transportation of mRNA to different organs. We also identify CEBPB (fold change=1.2), KCNS1 (fold change=1.3) and RGS19 (Fold change=1.2) that are associated with cell cycle regulation [23-28].

Among the 13 down-regulated probes located on 9 genes, 2 of them are mapped to 20p, while the rest are mapped to 20q. Though currently there is no evidence to suggest these genes are responsible for cervical cancer, they may contribute to the etiology of the tumor. For example, the downstream of WISP2 (fold change=0.54) in WNT1 signaling pathway has been shown to be related to malignant transformation, which may also associated with cervical cancer[29]. Further studies are needed to investigate the effects of these genes on cervical cancer progression.

Table 4.3 Differentially expressed genes identified by Two-stage model.

Probe.Set.ID	Chromosomal.Location	Gene.Symbol	Fold Change (Tumor/Normal)
206102_at	chr20p11.21	GINS1	2.53a
208079_s_at	chr20q13.2-q13.3	AURKA	2.49a
203936_s_at	chr20q11.2-q13.1	MMP9	2.11a
206467_x_at	chr20q13.3	TNFRSF6B	2.01c
219512_at	chr20q11.23	DSN1	1.93a
201853_s_at	chr20p13	CDC25B	1.88a
206546_at	chr20q13.33	SYCP2	1.82a
202954_at	chr20q13.12	UBE2C	1.81a
204092_s_at	chr20q13.2-q13.3	AURKA	1.80a
210052_s_at	chr20q11.2	TPX2	1.80a
202946_s_at	chr20p12.2	BTBD3	1.78a
204639_at	chr20q12-q13.11	ADA	1.75a
203892_at	chr20q12-q13.2	WFDC2	1.75a
218586_at	chr20q13.33	C20orf20	1.75a
212898_at	chr20q11.23	KIAA0406	1.73a
219888_at	chr20q11.21	SPAG4	1.69a
201202_at	chr20pter-p12	PCNA	1.62a
210042_s_at	chr20q13	CTSZ	1.61
201710_at	chr20q13.1	MYBL2	1.58a
203943_at	chr20q11.21	KIF3B	1.52a
204947_at	chr20q11.2	E2F1	1.49a
216705_s_at	chr20q12-q13.11	ADA	1.46a
201204_s_at	chr20p12	RRBP1	1.45a
206567_s_at	chr20q11.22-q11.23	PHF20	1.44
206656_s_at	chr20p11.22-p11.21	C20orf3	1.43a
211678_s_at	chr20q13.13	RNF114	1.42
216548_x_at	chr20q11.22	HMGB3L1	1.41a
201704_at	chr20p11.2-p11.22	ENTPD6	1.41a
200867_at	chr20q13.13	RNF114	1.41
200875_s_at	chr20p13	NOP56	1.40
213140_s_at	chr20q13.3	SS18L1	1.40a
210766_s_at	chr20q13	CSE1L	1.39a
222251_s_at	chr20q13.33	GMEB2	1.39
204432_at	chr20p13	SOX12	1.38a
213090_s_at	chr20q13.33	TAF4	1.37
201558_at	chr20q13.31	RAE1	1.37
207366_at	chr20q12	KCNS1	1.35c
201111_at	chr20q13	CSE1L	1.34a

Table 4.3 *(cont'd)*

Probe.Set.ID	Chromosomal.Location	Gene.Symbol	Fold Change (Tumor/Normal)
217888_s_at	chr20q13.33	ARFGAP1	1.33a
202071_at	chr20q12	SDC4	1.33
201271_s_at	chr20q11.21-q11.23	RALY	1.32a
221827_at	chr20p13	RBCK1	1.32
212430_at	chr20q13.31	RBM38	1.32a
201415_at	chr20q11.2	GSS	1.31a
213399_x_at	chr20q12-q13.1	RPN2	1.31a
200903_s_at	chr20cen-q13.1	AHCY	1.31a
206918_s_at	chr20q11.22	CPNE1	1.31a
213491_x_at	chr20q12-q13.1	RPN2	1.31a
208821_at	chr20p13	SNRPB	1.31a
218282_at	chr20q11.22	EDEM2	1.31
209049_s_at	chr20q13.12	ZMYND8	1.29
221021_s_at	chr20q11.23-q12	CTNNBL1	1.29a
212062_at	chr20q13.2	ATP9A	1.28a
221484_at	chr20q13.1-q13.2	B4GALT5	1.28a
209684_at	chr20p11.22	RIN2	1.28
221485_at	chr20q13.1-q13.2	B4GALT5	1.27a
221741_s_at	chr20q13.33	YTHDF1	1.27a
203459_s_at	chr20p13-p12	VPS16	1.27
211630_s_at	chr20q11.2	GSS	1.27a
201281_at	chr20q13.33	ADRM1	1.27a
216262_s_at	chr20q11.2-q12	TGIF2	1.26
212864_at	chr20p13	CDS2	1.26a
218708_at	chr20p12-p11.2	NXT1	1.26
210150_s_at	chr20q13.2-q13.3	LAMA5	1.26b
201032_at	chr20q11.2-q12	BLCAP	1.26
221499_s_at	chr20q13.32	STX16	1.25a
201114_x_at	chr20q13.33	PSMA7	1.25a
215852_x_at	chr20q11.23	C20orf117	1.24a
213175_s_at	chr20p13	SNRPB	1.24a
204336_s_at	chr20q13.33	RGS19	1.24
208689_s_at	chr20q12-q13.1	RPN2	1.24a
211318_s_at	chr20q13.31	RAE1	1.24
208743_s_at	chr20q13.1	YWHAB	1.24c
217286_s_at	chr20q11.21-q11.23	NDRG3	1.23
208948_s_at	chr20q13.1	STAU1	1.23
218315_s_at	chr20pter-q11.23	CDK5RAP1	1.22a
201112_s_at	chr20q13	CSE1L	1.22a

Table 4.3 *(cont'd)*

Probe.Set.ID	Chromosomal.Location	Gene.Symbol	Fold Change (Tumor/Normal)
213037_x_at	chr20q13.1	STAU1	1.22
202505_at	chr20p12.2-p11.22	SNRPB2	1.22a
221500_s_at	chr20q13.32	STX16	1.21a
44146_at	chr20q13.33	GMEB2	1.21
215693_x_at	chr20q13.13	DDX27	1.20a
221780_s_at	chr20q13.13	DDX27	1.20a
217792_at	chr20p11	SNX5	1.20
201206_s_at	chr20p12	RRBP1	1.18a
212501_at	chr20q13.1	CEBPB	1.18c
207320_x_at	chr20q13.1	STAU1	1.17
218159_at	chr20p13	DDRGK1	1.17c
217770_at	chr20q12-q13.12	PIGT	1.16a
209171_at	chr20p	ITPA	1.14c
218559_s_at	chr20q11.2-q13.1	MAFB	0.81a
221528_s_at	chr20q13	ELMO2	0.80
220363_s_at	chr20q13	ELMO2	0.70
217154_s_at	chr20q13.2-q13.3	EDN3	0.67a
55692_at	chr20q13	ELMO2	0.66
205792_at	chr20q12-q13.1	WISP2	0.54
206482_at	chr20q13.3	PTK6	0.52a
219090_at	chr20p13	SLC24A3	0.51a
57588_at	chr20p13	SLC24A3	0.48a
206004_at	chr20q11.2	TGM3	0.34a
220022_at	chr20q13.12	ZNF334	0.29
220388_at	chr20q11.22	FER1L4	0.28c
208399_s_at	chr20q13.2-q13.3	EDN3	0.05

a: The gene has been identified by our method ,LIMMA and Scotto *et al*.
b: The gene has been identified by our method and Scotto *et al*., but not by LIMMA.
c: The gene has been identified by our method , but not by LIMMA.
d: by ours and LIMMA (no footnote).

Table 4.4 Differentially expressed genes identified by LIMMA only.

Probe.Set.ID	Chromosomal.Location	Gene.Symbol
201021_s_at	chr20p12.1	DSTN
201022_s_at	chr20p12.1	DSTN
201053_s_at	chr20p13	PSMF1
202190_at	chr20q13.2	CSTF1[a]
202924_s_at	chr20q11.21	PLAGL2
203650_at	chr20q11.2	PROCR

Table 4.4 (*cont'd*)

Probe.Set.ID	Chromosomal.Location	Gene.Symbol
203691_at	chr20q12-q13	PI3a
204869_at	chr20p11.2	PCSK2
205243_at	chr20q12-q13.1	SLC13A3[a]
205286_at	chr20q13.2	TFAP2C
205287_s_at	chr20q13.2	TFAP2C
205296_at	chr20q11.2	RBL1
205557_at	chr20q11.23-q12	BPI
208725_at	chr20pter-q12	EIF2S2
208726_s_at	chr20pter-q12	EIF2S2
209221_s_at	chr20q13.3	OSBPL2
210702_s_at	chr20q13.13	PTGIS
210720_s_at	chr20q11.22	NECAB3
211085_s_at	chr20q11.2-q13.2	STK4
212234_at	chr20q11.1	ASXL1
212349_at	chr20q11	POFUT1[a]
212437_at	chr20p13	CENPB
213799_s_at	chr20p13	PTPRA
214498_at	chr20q11.2-q12	ASIP
215346_at	chr20q12-q13.2	CD40
215707_s_at	chr20p13	PRNP
215822_x_at	chr20q13.33	MYT1
215927_at	chr20q13.13	ARFGEF2
216505_x_at	chr20p13	RPS10P5
217024_x_at	chr20p13	SIRPA
218010_x_at	chr20q13.33	PPDPF
218081_at	chr20p13	C20orf27
218325_s_at	chr20q13.33	DIDO1
218448_at	chr20q13.33	C20orf11
218579_s_at	chr20q11.22-q12	DHX35[a]
218968_s_at	chr20q13.2	ZFP64
219536_s_at	chr20q13.2	ZFP64
220668_s_at	chr20q11.2	DNMT3B
221209_s_at	chr20p12.1-p11.23	OTOR[a]
221890_at	chr20q11.21-q13.12	ZNF335
222044_at	chr20q13.12	PCIF1
222106_at	chr20pter-p12	PRND
222259_s_at	chr20q13.2-q13.3	SPO11
32723_at	chr20q13.2	CSTF1[a]
41469_at	chr20q12-q13	PI3[a]
78047_s_at	chr20q11.22	LOC729580

a: The gene has been identified by LIMMA and Scotto *et al*.

4.5 Discussion

Microarray gene expression profiling has been used in biological research to search for differentially expressed genes and pathways responsible for complex human diseases. Due to the noise and artifacts in the microarray data, data preprocessing has to be conducted to improve correlation between the biological signals and measured probe intensities. By applying the PDNN model to our data preprocessing step, we explicitly model the underlying mechanisms of array hybridization and take the probe binding affinity into account, which has been shown to be related to probe intensities. Although gene expression level estimated by the PDNN model could represent the true target concentration, the model, just as the other preprocessing algorithms, does not provide uncertainty measure in the estimated gene-expression level. In the current study, we adopt a two stage model to incorporate the uncertainty in data preprocessing into the DEG detection algorithm, and develop a new DEG detection model, the two-stage-LIMMA. The two-stage-LIMMA 1) first applies the statistical algorithm AS274 to estimate the variance of the estimated gene expression level based on the PDNN model, 2) uses LIMMA to estimate the variance associated with the DEG detection algorithm, and 3) builds a new test statistic which considers both the variance associated with data preprocessing as well as the variance associated with DEG detection algorithm. The new method not only provides the confidence measure for the data preprocessing algorithm, but also improves the sensitivity and specificity of the DEG detection algorithm, which helps to identify genes and gene combinations that contribute to complex human diseases. The findings extracted from microarray can be enhanced by adopting novel statistical approaches, as demonstrated here. Through both simulations and Affymetrix Latin Square data we have shown that by incorporating the uncertainty in the data preprocessing step our method can reduce the number of false positives significantly while keep the false

negative comparable to LIMMA. In addition, our method provides a powerful and flexible framework to trade-off between sensitivity and specificity by specifying different weight values to suit for various needs for biological research.

Across-array normalization has been widely used for data preprocessing to remove noise, artifacts of microarray data. However, such normalization procedure may introduce correlations from unrelated samples, which may pose problems when the DEG detection algorithm is applied to the preprocessed data. Our proposed method depends on the gene expression level estimated from PDNN, a single array approach free of cross-array normalization. Consequently, two-stage-LIMMA inherits the advantage of the PDNN model, and identifies differentially expressed genes fully determined by the individual's sample, which is free from the issues of cross-subject dependence introduced by data preprocessing algorithms.

Although the DEG detection (i.e. second stage model) of our method depends on LIMMA, our Two-stage-LIMMA method differs largely from the LIMMA. As discussed above, the LIMMA method takes the estimated gene expression level from data preprocessing step as the raw data and it does not take into consideration the variance of the estimated expression level and the possible correlation between samples introduced by the data preprocessing algorithm, while our new method adopts a single array based data preprocessing algorithm and explicitly takes the variability of the data preprocessing step into account. As shown through both simulations and Affymetrix Latin Square data, our method outperforms LIMMA in most scenarios, especially in the case when there is a large variability in the observed probe intensities. In addition, our method is easily adapted to meet different needs of biologists by specifying different values of ρ. For example, if the false positive is a big concern for the researchers, ρ could be set at a larger value to control for the false positive. On the other hand, if the false

negative is of great concern in the study, ρ could be set at a smaller value to achieve more findings at the expense of increasing false positive slightly. Currently the second stage model is built based on LIMMA, but it could be easily adapted to adopt other moderated-t statistics based DEG detection algorithm by modifying the variance of the second stage model.

In conclusion, we have presented a powerful tool for DEG detection. It uses PDNN to preprocess the raw probe intensities and incorporates both the variance of DEG detection algorithm and the variance of estimated gene expression level into the two-stage model. It provides flexible framework for trade-off between sensitivity and specificity to suit for different needs for biological research. Our results provide strong evidence that by modeling the variance introduced in data preprocessing step can reduce the false positive findings.

REFERENCES

REFERENCES

1. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.

2. DeRisi, J.L., V.R. Iyer, and P.O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 1997. **278**(5338): p. 680-6.

3. Lee, C.K., et al., *Gene expression profile of aging and its retardation by caloric restriction.* Science, 1999. **285**(5432): p. 1390-3.

4. Ly, D.H., et al., *Mitotic misregulation and human aging.* Science, 2000. **287**(5462): p. 2486-92.

5. Lemieux, S., *Probe-level linear model fitting and mixture modeling results in high accuracy detection of differential gene expression.* BMC Bioinformatics, 2006. **7**: p. 391.

6. Irizarry, R.A., Z. Wu, and H.A. Jaffee, *Comparison of Affymetrix GeneChip expression measures.* Bioinformatics, 2006. **22**(7): p. 789-94.

7. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 31-6.

8. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.

9. Wu, Z.J., et al., *A model-based background adjustment for oligonucleotide expression arrays.* Journal of the American Statistical Association, 2004. **99**(468): p. 909-917.

10. Zhang, L., M.F. Miles, and K.D. Aldape, *A model of molecular interactions on short oligonucleotide microarrays.* Nat Biotechnol, 2003. **21**(7): p. 818-21.

11. Astrand, M., *Contrast normalization of oligonucleotide arrays.* J Comput Biol, 2003. **10**(1): p. 95-102.

12. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.

13. Sartor, M.A., et al., *Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments.* BMC Bioinformatics, 2006. **7**: p. 538.

14. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.

15. Held, G.A., G. Grinstein, and Y. Tu, *Modeling of DNA microarray data by using physical properties of hybridization.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(13): p. 7575-80.

16. Wan, L., et al., *Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation.* Nucleic acids research, 2009. **37**(17): p. e117.

17. Zhang, L., M.F. Miles, and K.D. Aldape, *A model of molecular interactions on short oligonucleotide microarrays.* Nature biotechnology, 2003. **21**(7): p. 818-21.

18. Zhang, L., et al., *Free energy of DNA duplex formation on short oligonucleotide microarrays.* Nucleic acids research, 2007. **35**(3): p. e18.

19. Baldi, P. and A.D. Long, *A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes.* Bioinformatics, 2001. **17**(6): p. 509-19.

20. Kadota, K., Y. Nakai, and K. Shimizu, *A weighted average difference method for detecting differentially expressed genes from microarray data.* Algorithms Mol Biol, 2008. **3**: p. 8.

21. Breitling, R., et al., *Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.* FEBS Lett, 2004. **573**(1-3): p. 83-92.

22. Miller, A.J., *Algorithm AS 274.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1992. **41**(2): p. 458-478.

23. NCBI. *TNFRSF6B tumor necrosis factor receptor superfamily, member 6b, decoy [Homo sapiens]* Available from: http://www.ncbi.nlm.nih.gov/gene/8771.

24. NCBI. *GMEB2 glucocorticoid modulatory element binding protein 2 [Homo sapiens]* Available from: http://www.ncbi.nlm.nih.gov/gene/26205.

25. NCBI. *STAU1 staufen, RNA binding protein, homolog 1 (Drosophila) [Homo sapiens]* Available from: http://www.ncbi.nlm.nih.gov/gene/6780.

26. NCBI. *CEBPB CCAAT/enhancer binding protein (C/EBP), beta [Homo sapiens]* Available from: http://www.ncbi.nlm.nih.gov/gene/1051.

27. NCBI. *KCNS1 potassium voltage-gated channel, delayed-rectifier, subfamily S, member 1 [Homo sapiens]* Available from: http://www.ncbi.nlm.nih.gov/gene/3787.

28. NCBI. *RGS19 regulator of G-protein signaling 19 [Homo sapiens]* Available from: http://www.ncbi.nlm.nih.gov/gene/10287.

29. NCBI. *WISP2 WNT1 inducible signaling pathway protein 2 [Homo sapiens]* Available from: http://www.ncbi.nlm.nih.gov/gene/8839.

CHAPTER 5

CONCLUSION AND DISCUSSION

The radical breakthrough in biotechnologies has made it possible to simultaneously genotype millions of single nucleotide polymorphisms or profile thousands of genes at an affordable cost. Benefiting from the high throughput technologies and the completion of the HapMap project[1], significant progress has been made in identifying associated genetic and environmental determinants, which is of great importance to clinicians, researchers as well as the general public. However, the genetic etiologies of complex human diseases, such as coronary heart disease, type II diabetes and cervical cancer, remain largely unknown. Findings from studies with high-dimensional data can be further enhanced by adopting computationally efficient and powerful analytic approaches. The development and application of new statistical approaches will help to achieve a better understanding of genetic etiology of complex human diseases, and eventually promote new diagnostic and therapeutic strategies. In this research, I have proposed three statistical methods which are designed to provide accurate genotype calls, to improve the understanding of pathophysiological and etiological pathways of co-morbid diseases, and to facilitate the identification of differentially expressed genes.

MA-SNP — A NEW GENOTYPE CALLING METHOD.

Microarray technologies are subject to many noise and artifacts, such as cross hybridization[2], batch effect [3, 4] and genomic wave[5]. Extensive studies have shown that probe intensities of oligonucleotide arrays depend on not only the concentration of the target sequence but also the binding affinity of the probes[2, 6-8]. Though the physico-chemical properties of probe binding between the probe sequence and target sequence can be largely modeled with the probe sequence

structure as demonstrated by PICR[2], the complicated hybridization process may also be influenced by certain unknown factors, which may, if not properly adjusted for, lead to systematic bias of the estimated allelic copy number. The universal genotype criterion used by PICR fails to take SNP-specific factors that influence the observed intensities into account, and therefore it may lack the capacity to provide accurate genotype calls for all SNPs targeted on the array. In addition, the PICR model does not provide confidence measure of the genotype calls, which is of great importance for downstream analysis. Without setting aside inaccurate genotype calls one may get insufficient or invalid association result. Recently, substantial evidences suggest that effect of batch size and composition also affects the genotype calling accuracy and lowers the genotype call rate.

To address these issues, I have developed a new genotype calling algorithm, which is built based upon PICR model. Similar to PICR, the new MA-SNP model first estimates allelic target concentration through linear regression without pre-processing observed intensities (i.e. a cross-array normalization), and makes the inference of genotype calls solely using the estimated allelic target concentration based on each individual's data. Instead of applying a universal criterion for all SNPs on the array as PICR, the new MA-SNP method adopts an empirical model to estimate the SNP-specific genotype calling criteria, and provides a confidence measure of the given genotype calls, which benefits the downstream analysis. Moreover, the proposed method models the density of the MA-ratios using a normal mixture model to account for the potential batch effect, and by explicitly correcting the potential batch effect the MA-SNP method improves the genotype calling accuracy and the genotype call rate. Though the proposed method achieves relatively high genotyping accuracy, it can be further improved by introducing a

random effect model to correct for the batch effect, especially for studies with small sample size. Further investigation is needed to explore the performance of the model.

To further demonstrate the utility of the proposed method, I apply the proposed method to datasets obtained from Wellcome Trust Case Control study to investigate the genetic susceptibility loci for coronary heart disease on chromosome 9. The SNPs that have been identified by a single locus association analysis with p-value$<10^{-7}$ are located at 9p21.3. All SNPs (rs1333042, rs1333048, rs1333049, rs2891168, rs4977574, and rs6475606) except for one rs9632884 have been reported with strong association by studies using various techniques [9-17], which indicates that the proposed method does not generate many false positive findings.

THE COMORBIDITY STUDY

With the increase in genetic findings, converging evidence has revealed that the same genetic susceptibility loci could be associated with multiple disease outcomes. For example, both clinical and epidemiological studies have reported a high-degree of co-morbidity between bipolar disorders and migraine, which could be partially due to the shared genetic variants [18-22]. Identifying genetic susceptibility loci contributing to co-morbidity, as well as those loci that are unique risk factors to each co-morbid condition is of great importance, as it helps elucidate the causes of co-morbidity and promotes new prevention/treatment to co-morbid conditions. The relation between co-morbid diseases varies from case to case. To the best of my knowledge, there are no statistical methods except for composite phenotype method that are aimed at studying co-morbidity.

To investigate the co-morbidity between complex human diseases, I have proposed a multivariate Mann-Whitney approach for co-morbidity analysis. The proposed method utilizes

the entire sample, and is capable of capturing shared genetic variants and their possible interactions contributing to disease co-morbidity, as well as unique genetic variants for each disease outcome. Similar to other Mann-Whitney based methods, it is a non-parametric approach, which does not assume a model of inheritance, and is free of the issues of an increasing number of parameters. It adopts a forward selection algorithm, which substantially reduces the searching space of interaction combinations and allows for high-order interactions. Through simulations, I have shown that the multivariate Mann-Whitney attains a higher or equivalent power to the composite phenotype analysis under a variety of disease models, and is more robust under different correlation models simulated between comorbid diseases. This feature is important, as the current knowledge of disease co-morbidity is limited. Though MMW method achieves higher power than the commonly adopted method, it still lacks the capacity of dealing with whole genome wide data due to the computational burden. Further studies are needed to propose a more computationally efficient method that can deal with millions of SNPs simultaneously.

To investigate the co-morbidity between coronary heart disease and type II diabetes, I apply the proposed method to the data sets obtained from WTCCC. Three-locus and five-locus models have been identified for coronary heart disease and type II diabetes, respectively. However, no loci are identified for the comorbidity between the two diseases. There are several possible reasons leading to this negative finding. First, this study is a candidate gene based analysis, and the markers that have not been reported to be associated with either of the diseases are not included in this analysis. In addition, no statistical imputation has been conducted and any loci that are not targeted by Affymetrix GeneChip Human Mapping 500K Array Set are not included in this study. Second, environmental factors such as obesity, physical inactivity and family history, play an important role in both T2D and CAD, but they are not measured in the

first stage of WTCCC study. As a consequence, the power of detecting loci contributing to both diseases may be attenuated.

A TWO STAGE MODEL FOR DIFFERENTIALLY EXPRESSED GENE DETECTION

Since the introduction of microarray technology in the mid 1990s, gene expression profiling has been used in numerous biological researches to tease apart biological pathways and identify genes and joint gene actions contributing to complex human diseases [23-25]. The analysis for gene expression microarray usually involves both data preprocessing, which often comprises of background correction, normalization and summarization to improve the correlation between biological effect and measured intensities[26], and DEG analysis based on the preprocessed data. Numerous data preprocessing algorithms have been proposed, but little attention has been paid to the underlying array hybridization process, which has been shown to be of great importance to obtain accurate gene expression level summary [2, 6-8]. In addition, most of the current DEG detection algorithms are built based upon the gene expression summary obtained from data preprocessing step, however, the uncertainty in the estimated gene expression level has not been carefully addressed, which may lead to false positive/negative findings.

To address these issues, I adopt a two stage model to incorporate the uncertainty in data preprocessing into the DEG detection algorithm, and develop a new DEG detection model, the 2-stage-Limma. The new method first applies the statistical algorithm AS274[27] to estimate the gene expression level and the corresponding variance of each gene, and then uses Limma [28] to estimate the variance associated with the DEG detection algorithm. Finally it builds a new test statistics which takes both the variability in the data preprocessing step and the variability associated with DEG algorithm into account. The new proposed method not only provides the

confidence measure for the data preprocessing step, but also improves the sensitivity and specificity for DEG detection. It is also easily adapted to suit different goals of biological research by tuning one scale parameter. Through both simulations and Affymetrix Latin Square data, I have shown that by incorporating the uncertainty in the data preprocessing step our method can significantly reduce the false positive findings while keeping the false negative comparable to Limma. Currently the new method can only use PDNN model as its data preprocessing algorithm, and further studies are needed to evaluate the performance of the new model with different data preprocessing algorithms, such as RMA and GCRMA.

I further apply the proposed method to study the gene expression profiling for cervical cancer on chromosome 20[29]. Out of 571 targeted genes on HU133A array platforms, I have identified 106 probes (83 genes) that are differentially expressed when we compared the tumor cell with the normal tissues. Among these probes, 13 probes have decreased expression and 93 probes have increased gene expression level. Most of the identified differentially expressed genes belong to several functional groups, such as cell cycle regulation, nuclear function, transferase, viral replication, methylation and chromatin remodeling, and endopeptidase activity[29]. The findings may help understand the genetic etiology of cervical cancer, and facilitate the biomarker identification and eventually lead to new therapeutic strategies for cervical cancer to improve the survival rate especially the advanced stage cases.

REFERENCES

REFERENCES

1. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

2. Wan, L., et al., *Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation.* Nucleic acids research, 2009. **37**(17): p. e117.

3. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.

4. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nature reviews. Genetics, 2010. **11**(10): p. 733-9.

5. Wen, Y., M. Li, and W.J. Fu, *Catching the genomic wave in oligonucleotide SNP arrays by modeling sequence binding.* Bioinformatics (submitted), 2011.

6. Held, G.A., G. Grinstein, and Y. Tu, *Modeling of DNA microarray data by using physical properties of hybridization.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(13): p. 7575-80.

7. Zhang, L., M.F. Miles, and K.D. Aldape, *A model of molecular interactions on short oligonucleotide microarrays.* Nature biotechnology, 2003. **21**(7): p. 818-21.

8. Zhang, L., et al., *Free energy of DNA duplex formation on short oligonucleotide microarrays.* Nucleic acids research, 2007. **35**(3): p. e18.

9. Cluett, C., et al., *The 9p21 myocardial infarction risk allele increases risk of peripheral artery disease in older people.* Circulation. Cardiovascular genetics, 2009. **2**(4): p. 347-53.

10. Ellis, K.L., et al., *A common variant at chromosome 9P21.3 is associated with age of onset of coronary disease but not subsequent mortality.* Circulation. Cardiovascular genetics, 2010. **3**(3): p. 286-93.

11. Lanktree, M., J. Oh, and R.A. Hegele, *Genetic testing for atherosclerosis risk: inevitability or pipe dream?* The Canadian journal of cardiology, 2008. **24**(11): p. 851-4.

12. Preuss, M., et al., *Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls.* Circulation. Cardiovascular genetics, 2010. **3**(5): p. 475-83.

13. Qi, L., et al., *Genetic risk score and risk of myocardial infarction in Hispanics.* Circulation, 2011. **123**(4): p. 374-80.

14. Saleheen, D., et al., *Association of the 9p21.3 locus with risk of first-ever myocardial infarction in Pakistanis: case-control study in South Asia and updated meta-analysis of Europeans.* Arteriosclerosis, thrombosis, and vascular biology, 2010. **30**(7): p. 1467-73.

15. Schaefer, A.S., et al., *Identification of a shared genetic susceptibility locus for coronary heart disease and periodontitis.* PLoS genetics, 2009. **5**(2): p. e1000378.

16. Silander, K., et al., *Worldwide patterns of haplotype diversity at 9p21.3, a locus associated with type 2 diabetes and coronary heart disease.* Genome medicine, 2009. **1**(5): p. 51.

17. Yamada, Y., S. Ichihara, and T. Nishida, *Molecular genetics of myocardial infarction.* Genomic medicine, 2008. **2**(1-2): p. 7-22.

18. Dilsaver, S.C., et al., *Migraine Headache in Affectively Ill Latino Adults of Mexican American Origin Is Associated With Bipolarity.* Prim Care Companion J Clin Psychiatry, 2009. **11**(6): p. 302-306.

19. Dilsaver, S.C., et al., *Is a family history of bipolar disorder a risk factor for migraine among affectively ill patients?* Psychopathology, 2009. **42**(2): p. 119-23.

20. Bowden, C.L., et al., *A randomized, placebo-controlled 12-month trial of divalproex and lithium in treatment of outpatients with bipolar I disorder. Divalproex Maintenance Study Group.* Arch Gen Psychiatry, 2000. **57**(5): p. 481-9.

21. Oedegaard, K.J., et al., *A genome-wide association study of bipolar disorder and comorbid migraine.* Genes Brain Behav, 2010. **9**(7): p. 673-80.

22. Oedegaard, K.J., et al., *A genome-wide linkage study of bipolar disorder and co-morbid migraine: replication of migraine linkage on chromosome 4q24, and suggestion of an overlapping susceptibility region for both disorders on chromosome 20p11.* J Affect Disord, 2010. **122**(1-2): p. 14-26.

23. DeRisi, J.L., V.R. Iyer, and P.O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 1997. **278**(5338): p. 680-6.

24. Lee, C.K., et al., *Gene expression profile of aging and its retardation by caloric restriction.* Science, 1999. **285**(5432): p. 1390-3.

25. Ly, D.H., et al., *Mitotic misregulation and human aging.* Science, 2000. **287**(5462): p. 2486-92.

26. Irizarry, R.A., Z. Wu, and H.A. Jaffee, *Comparison of Affymetrix GeneChip expression measures.* Bioinformatics, 2006. **22**(7): p. 789-94.

27. Miller, A.J., *Algorithm AS 274.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1992. **41**(2): p. 458-478.

28. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.

29. Scotto, L., et al., *Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression.* Genes Chromosomes Cancer, 2008. **47**(9): p. 755-65.